Genomics of Salmonella sevovars Mbandaka, Typhimurium and its monophasic variant in milk and pork food sectors

Madeleine de Sousa Violante

To cite this version:

Madeleine de Sousa Violante. Genomics of Salmonella sevovars Mbandaka, Typhimurium and its monophasic variant in milk and pork food sectors. Bioinformatics [q-bio.QM]. Université Paris-Est, 2022. English. NNT : . tel-04176972

HAL Id: tel-04176972
https://hal.inrae.fr/tel-04176972v1

Submitted on 3 Aug 2023

THESE DE DOCTORAT

pour obtenir le grade de

Docteur de l'Université Paris-Est Sup

Spécialité : Bioinformatique et Génomique bactérienne

École doctorale n°581
Agriculture, alimentation, biologie, environnement et santé (ABIES)

par

Madeleine DE SOUSA VIOLANTE

Genomics of *Salmonella* sevovars Mbandaka, Typhimurium and its monophasic variant in milk and pork food sectors

Directeur de thèse : **Michel-Yves MISTOU**
Co-directeur de thèse : **Nicolas RADOMSKI**
Co-encadrant de thèse : **Ludovic MALLET**
Co-encadrante de thèse : **Valérie MICHEL**
Co-encadrante de thèse : **Carole FEURER**

Thèse présentée et soutenue à Paris, le 17 novembre 2022

Composition du jury :

Céline SCORNAVACCA, Directrice de recherche, CNRS (Montpellier)	Présidente
Guy PERRIERE, Directeur de recherche, CNRS (Lyon)	Rapporteur & Examinateur
Benoit DOUBLET, Directeur de recherche, INRAE (Nouzilly)	Rapporteur & Examinateur
Alexandra CALTEAU, Chercheure, CEA (Paris-Saclay)	Examinatrice
Marie TOUCHON, Chargée de recherche, CNRS (Institut Pasteur)	Examinatrice
Michel-Yves MISTOU, Directeur de recherche, INRAE (Jouy-en-Josas)	Directeur de thèse
Nicolas RADOMSKI, Expert, IZSAM (Teramo, Italie)	Co-Directeur de thèse
Ludovic MALLET, PhD, Institut Claudius Regaud (Toulouse)	Invité
Valerie MICHEL, PhD, ACTALIA (La Roche-sur-Foron)	Invitée
Carole FEURER, PhD, IFIP-Institut du porc (Le Rheu)	Invitée

StatInfOmics - MaIAGE
Domaine de Vilvert 78350 JOUY-EN-JOSAS - FRANCE

Acknowledgements

First of all, I would like to express my sincere gratitude to ACTALIA and IFIP for giving me the opportunity to conduct this thesis research.

I would like to thank my thesis supervisors Michel-Yves Mistou and Nicolas Radomski. Thank you very much for all the time, the knowledge and ongoing support and guidance throughout this project. Michel-Yves, thank you very much for maintaining supervision and support remotely throughout my three years of this PhD journey. Nicolas, thank you for joining my thesis along the way and for having accepted to mentor me in difficult times.

I'm also grateful to Ludovic Mallet, who was the first to give me the opportunity to work on this fascinating subject, first as an internship and then as a PhD student. Thank you so much for your guidance, your time, your passion and your inspiring motivation on this research topic, I learned a lot from you.

A special thought for Valérie Michel et Carole Feurer. Thank you both proposing me this thesis, as well as for the help that you provided during the orientation of my work, and especially for all the technical knowledge you provided me. Thank you also for the enriching opportunity to visit the sampling sites.

I would like to thank the jury: my reviewers Guy Perriere and Benoit Doublet, but also my examiners Céline Scornavacca, Alexandra Calteau and Marie Touchon. I'm obliged by you having accepted to evaluate my thesis report. I filled with gratitude for the members of my thesis committee, Hélène Chiapello, Sabrina Cadel-Six, Raluca Uricaru, Isabelle Virgoleux-Payant and Sandra Blaise-Boisseau. Thanks for your precious advices and for monitoring my progress of my thesis.

Furthermore, I would like to extend my sincere thanks to the members of the UMT-ASIICS, Renaud Lailler, Hélène Amar, Laurent Laloux, Didier Majou, Choreh Farrokh, as well as SCA-Dublin members for their follow-up of my projects, and the enriching discussions we were able to have.

I also thank all LSAL of Anses staff for allowing me to realize this thesis in a great working environment. I would like to thank the open space community that I learned to know during my thesis: Donatien, Mai Lan, Sandrine, Sandra, Arnaud, all the trainees I met, and all the people I could interact with during my thesis. Thanks for the SEL team for your continuous support, and thanks again to Sabrina for your dedication helping with me to recover strains, and all your support on the genomics analysis. I would like to thank the former GAMeR team, Pierre-Emannuel Douarre, Arnaud Felten, Pauline Barbet, Meryl Vila Nova for their help and all the scientific exhanges.

I would like to express gratitude to the MaIAGE team at INRAE Jouy-en-Josas. Thanks for your warm welcome and your stimulating collective dynamics. Thanks to the the 1st floor team for making the final rush of my thesis much more enjoyable: Samantha, Julie, Romane, Solène, Cédric, Irène, Mariya, Stephane and all the trainees I met. Thank you for all the cookies you gave me while I was writing. I really appreciated the lunches and the moving dialogs, thanks Mahendra, Cyprien, Sandra, Anne-Laure, Arnaud, Louise.

This thesis has been particularly difficult with the COVID context. So I also need to thank all my friends, IRL or online, who helped me to bear this situation. Thanks to my JDR friends, Olivier and Axel, but also to my former classmates. Thanks to my guildmates, and to the PhD students of the PhD server, with whom I spent a lot of time chatting online.

Finally, I would like to thank my family, and especially my parents for the English revision of this section, for their warm support during these three years of thesis. Thanks to my grandfather that has been a model for me. And the last but not the least, I would like to warmly thank my fiancé for his confidence, his support, his composure and understanding, especially in this end of thesis.

Contents

List of Figures

List of Tables

List of abbreviations

ANSES	French Agency for Food, Environmental and Occupational Health & Safety
AMR	Antimicrobial resistance
APC	Aerobic plate count
ARS	Agences Régionales de Santé
ASIICS	Action for Surveillance, Investigation and Interventions in Sanitary Crisis
ASSuT	Ampicillin, streptomycin, sulfonamides, and tetracycline
CDC	The Centers for Disease Control and Prevention
CFU	Colony forming units
CGE	Center for Genomic Epidemiology
cgMLST	Coregenome MLST
CNR	Centre National de Référence
CNV	Copy number variations
DDPP	Direction départementale de la protection des populations
DGAL	Direction générale de l'Alimentation
DGCCRF	Direction générale de la Concurrence, de la Consommation et de la Répression des fraudes
DGS	Direction Générale de la Santé
DO	Déclaration obligatoire
EC	European Commission
ECDC	The European Centre for Disease Prevention and Control
EFSA	The European Food Safety Authority
ELFA	Enzyme-linked fluorescent assay
ELISA	Enzyme-linked immunosorbent assay
GFF	General feature format
GWAS	Genome-Wide Association Study
HACCP	Hazard Analysis Critical Control Point
IMS	Immuno-magnetic separation
HGT	Horizontal Gene Transfer
HTS	High-throughput sequencing
ICM	Institute for Brain and Spinal Cord
IFIP	The French Pork and Pig Institute

LNR	Laboratoire National de Référence
LPS	Lipopolysaccharide
MGE	Mobile genetic element
ML	Maximum likelihood
MLST	Multilocus sequence typing
MLVA	Multiple Loci VNTR Analysis
MPF	Mate-pair formation
MST	Minimum spanning tree
MUS	Mission des urgences sanitaires
NGS	Next generation sequencing
NTS	Non-typhoidal *Salmonella*
PCR	Polymerase Chain reaction
PCR-DGGE	PCR gradient gel electrophoresis
PFGE	Pulsed-field gel electrophoresis
pgSNP	Pan-genome SNPs
RASFF	Rapid Alert System for Food and Feed
RT-PCR	Reverse transcription PCR
SCA	Surveillance de la Chaine Alimentaire
SGI	*Salmonella* Genomic Island
SNPs	Single nucleotide polymorphisms
SPC	Standard plate count
SpF	Santé publique France
SPI	*Salmonella* Pathogenicity Island
ST	Sequence type
SV	Structural variants
TIAC	Toxi-infections alimentaires collectives
T1F	Type 1 fimbriae
T1SS	Type 1 secretory systems
T3SS	Type 3 secretion systems
TMV	Typhimurium monophasic variant
UMT	Unité Mixte Technologique
WHO	World Health Organization
wgMLST	Whole Genome MLST
WGS	Whole Genome Sequencing

Chapter 1

Introduction

This thesis was written in order to obtain the doctoral degree from the Université Paris-Est Sup via the doctoral school n°581 Agriculture, Food, Biology, Environment and Health (ABIES). This work was carried out as part of a CIFRE (Conventions industrielles de formation par la recherche) thesis, financed by ACTALIA and IFIP-Institut du porc, and hosted first at the ANSES (French Agency for Food, Environmental and Occupational Health & Safety) laboratory within the GAMeR (Genome Analysis Modelling and Risk) mission, and then at the INRAE laboratory within research unit 1404 MaIAGE (Mathematics and Informatics Applied to the Genome and Environment) in the StatInfOmics team.

This thesis was also part of the CasDAR-RT (Compte d'affection Spécial au Développement Agricole et Rural) project n°1710 EMISSAGE (Epidemiology of Salmonella in the animal sector by genomic approach), and supervised within the UMT ASIICS (Action pour la Surveillance, l'Investigation et l'Intervention dans les Crises Sanitaires) of which ACTALIA is the lead organization. The objective of the CasDAR-RT was to improve the surveillance and characterization of *Salmonella* in different food sectors.

Salmonella is a major global bacterial pathogen, highly polymorphic in its diversity of host range, clinical manifestation and outcome. Its impact on public health and its economical burden have continuously been driving efforts to understand the epidemiological situation or reduce its dissemination, historically by leveraging the most suitable typing methods available at the time.

The genomics era brought a valuable aid in the investigation and characterization of pathogenenic bacteria for public health. While pathogenic mechanisms and determinants have been characterised, the determinants of host tropism are still elusive and the toll on food safety and public health continues to frequently hit the headlines in the global news.

This thesis project was designed to address the limitations of current methods focusing on *Salmonella* genomics and to transfer the resolutive power of tailored genomics to the understanding of *Salmonella* adaptive paths driving, in a context of food-safety control, the host tropism (food, herd, contamination), the persistence, the resistance and some discriminant markers. A focus was made in this thesis on two *Salmonella* prevalent serovars in dairy, pig and pork food sectors : *Salmonella* Mbandaka, *Salmonella* Typhimurium and its monophasic variant.

The chapter 2 of this thesis presents how *Salmonella* is unique and diverse, and what is the state of knowledge after a century of research. A special emphasize was made on the genomics methods and their increasing usage in both fundamental and applied research related to food safety. The chapter 3 presents the developments implemented to increase the resolution of bioinformatics methods in genome comparison, notably thanks to an innovative method based on the pangenome. Finally, we will show in the chapter 4 the application of bioinformatics methods in the genomic research of *Salmonella* in the dairy and swine sectors, especially in understanding the diversity and dissemination of these strains.

Chapter 2

State of the art

2.1 Public health and food safety

2.1.1 Brief history of food safety

The recognition and avoidance of inherently poisonous foods may have marked the beginning of the history of food safety, which is almost as old as the history of humanity itself [1]. Numerous scientific and technological advancements brought many discoveries that we still benefit today for the quality and safety of food. While most of food hazard affected only a small part of the population, as human diets, habits and foods changed, food safety became more standardised. Starting with specific applications of laws for certain products (the Assize of Bread - 1202 [2]), states and countries started to define sanitary criteria to address issues related to health and adulterated products (first U.S. Food Safety Act - 1785 [3]).

In the early 1900s, foodborne diseases (typhoid fever, tuberculosis bovis, botulism and scarlet fever, etc.) were at the highest incidence and prevalence, with the highest mortality rates in all over the world [1, 4]. As a result, in the 19th and 20th centuries, laws governing food safety and sanitation regulations as well as research into potentially hazardous microorganisms were widely established. Some of the pathogens we know today are linked to the names of microbiologists who discovered them. For example, David E. Salmon, who worked on the hog cholera, identified the bacterial genus *Salmonella*, which bears his name and is well-known for being a serious threat to food safety [5].

Food safety concerns persist today despite thousands of years of experience, 150 years of food microbiology research and appearance of the latest molecular biology techniques, particularly given that epidemiological surveillance has shown a constant increase in the prevalence of foodborne illness [6]. Indeed, devastating outbreaks occurred recently, particularly salmonellosis [7], listeriosis [8], enterohaemorrhagic *Escherichia coli* infections [9], hepatitis A [10] and other diseases in both developed and developing countries. The risk of foodborne illness has increased dramatically, due to biological and chemical contamination of the areas where food is produced, processed and consumed [6].

With regard to hygiene monitoring, Hazard Analysis Critical Control Point (HACCP) programs were created to monitor critical control points for potential contamination during food processing [11]. Rapid analytical assays were necessary in this quality control scenario due to the stringent compliance to sanitary procedures in a food-processing environment [12]. Because of this dogma, the public health community gave a high importance to analytical assays that are affordable and easy to use in the framework of quality control in food-processing environment.

To assure the quality of the detection of foodborne hazard, microbiological methods has been develop to quickly characterise pathogens. Traditionally, detection of viable bacteria is performed by cultivating and monitoring the growth of individual microorganisms. Several tens of commonly used bacteriological media in the food industry have their own purposes for monitoring of microbiological contamination and/or detecting of pathogenic bacteria [13]. For instance, the use of routine non-selective media such as trypticase soy agar or standard methods agar, known as the aerobic plate count (APC) or standard plate count (SPC), is worthy of discussion in terms of the numerous alternative techniques that have been designed to improve upon it.

Viable and culturable cells can be grown on solid media and produce colonies assumed to be clones from each isolated single cell (e.g. non-fastidious bacteria on standard media, fastidious bacteria on enriched media, specific bacterial taxa on differential agar, specific bacterial taxa on selective agar). Biological culture dependent methods demand divers apparatus and require laborious work. Usually they involve a long process of sample collection, serial dilution, plating on selective and suitable media and waiting for appropriate incubation time to get visible colonies. This quantification of bacteria in each sample is frequently reported as total number of colony forming units (CFUs). It is recognised that only cells cultivable under control conditions can be counted with these methods (i.e. incubation temperature, incubation time, selective media and oxygen availability) [14].

Therefore, despite the advantages of these usual culture methods for viable bacteria detection (i.e. easy of use and low cost), their sensitivity levels is still relatively low compared with alternative methods [15].

Immunological and nucleic-acid sequence based detection techniques are supposed to be the most robust technologies in microbiology. Nucleic-acid sequence-based approaches like PCR (Polymerase Chain reaction) can detect microbial cells and specific genetic elements (e.g. toxins), while immunological approaches are not so specific for the detection of microorganisms because the transcription of detected proteins is affected by the cell environment [15].

Thus, different methods based on several PCR targets were created to reach accurate robust identification and microorganisms detection while combining PCR along with capillary electrophoresis [16], multiplex PCR based detection (which involves the simultaneous detection of multiple targets in a single reaction well, with multiple pairs of primers to each target [17, 18], and denaturing gradient gel electrophoresis (PCR-DGGE) (electrophoresis that uses chemical gradient to denature the nucleic acids as it moves across an acrylamide gel). Usually it separates genes of the same size based on denaturing ability and length heterogeneity.

Concerning these PCR-based techniques aiming at detecting quickly bacterial contaminants in food with high sensitivity, the enrichment step is crucial to increase bacterial cell numbers, prior to nucleic acid extraction and primer-specific amplification. Nevertheless, chimeric sequences induced by PCR-based techniques and associated electrophoresis technical issues limit the application of such approaches [19].

Immunological methods are based on the specific binding of antibody and antigen. More precisely, a targeted part of antigen (i.e. epitope) binds with the available antibody. Widely

Disease	Number of confirmed human cases	Hospitalisation		Deaths	
		Reported hospitalised cases	Proportion hospitalised (%)	Reported deaths	Case fatality (%)
Campylobacteriosis	120,946	8,605	21.0	45	0.05
Salmonellosis	52,702	6,149	29.9	57	0.19
Yersiniosis	5,668	353	29.1	2	0.07
STEC infections	4,446	652	40.9	13	0.42
Listeriosis	1,876	780	97.1	167	13.0
Tularaemia	641	64	52.0	0	0
Echinococcosis	488	44	60.3	0	0
Q fever	523	NA	NA	5	2.1
West Nile virus infection a	322	219	91.6	39	12.1
Brucellosis	128	36	64.3	2	3.6
Trichinellosis	117	16	72.7	0	0
Rabies	0	NA	NA	NA	NA

Table 2.1: Reported hospitalisations and case fatalities due to zoonoses in confirmed human cases in the EU from EFSA [37]

used immunological approaches in food investigation are enzyme-linked immunosorbent assay (known as ELISA, which uses antibodies and identify the substance by color change), enzyme-linked fluorescent assay (known as ELFA, is an immunological-based method, which is similar to ELISA, but a more sensitive biochemical test) and immuno-magnetic separation (known as IMS, is a laboratory tool that can effectively isolate cells from various food, blood fecal samples, or other body fluids) [20, 21, 22, 23]. Major foodborne diagnostics are currently characterized by commercially available ELISA kits [24].

Finally, as sequencing technology has advanced, microbiology has begun to give way to methods of genomic characterization of strains. Whole genome sequencing (WGS) has been widely used to provide detailed characterization of foodborne pathogens (detailed in section 2.2). These genomes of diverse species including *Salmonella*, *Escherichia coli*, *Listeria*, *Campylobacter* and *Vibrio* have provided a better understanding of the genetic composition of these pathogens [25]. Using WGS approaches, numerous government agencies, industry and academic institutions have created novel applications for food safety, such as outbreak detection [26, 27, 28] and characterization [9, 29], source tracking [30], determining the root cause of a contamination event [31], profiling of virulence and pathogenicity attributes [32, 25, 33], antimicrobial resistance monitoring [34], quality assurance for microbiology testing [35, 36], as well as many others. The future looks bright for additional applications that arise from new technologies and tools in genomics and metagenomics.

2.1.2 Foodborne pathogen cases

In 2020, the reporting of foodborne outbreaks in the EU was affected by the COVID-19 pandemic, with a decrease in the number of total outbreaks, human cases, hospitalisations and deaths (Figure 2.1) [37]. Even if the number of foodborne cases decreased, campylobacteriosis was the most commonly reported zoonosis, as it has been since 2005 (Table 2.1). It accounted for more than 60% of all the reported cases in 2020. Reported cases are the cases for which microbiological detection have been carried out and confirmed the presence of the pathogenic agent. It was followed by other bacterial diseases frequently reported, such like salmonellosis, yersiniosis and STEC infections. A total of 3086 outbreaks caused by foodborne agents was declared in EU in 2020, where *Salmonella* was the principal agent accounting for 22.5 % (N=694). *Salmonella* is also responsible for 19% of foodborne outbreaks in USA in 2017 [38]. Listeriosis and West Nile virus infection were the two most severe diseases with the highest mortality and hospitalisation rates, with 13% and 12.1% of a fatal outcome.

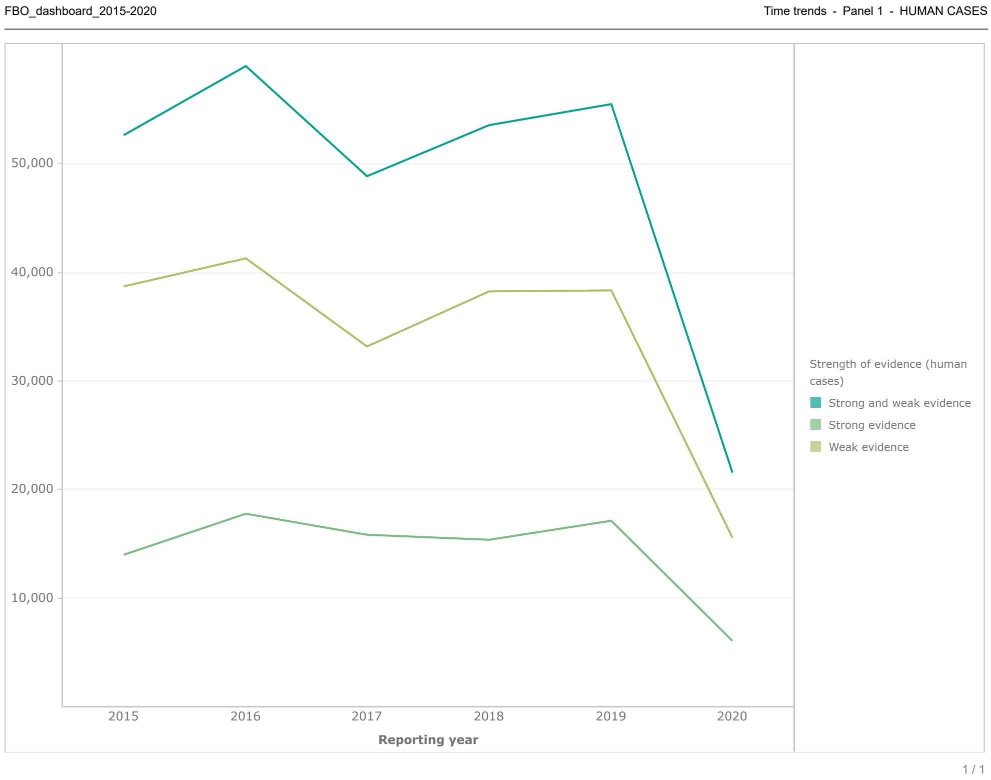


Figure 2.1: Number of human cases caused by foodborne agent between 2015 and 2020. Figure made from *https://www.efsa.europa.eu/en/microstrategy/FBO-dashboard* [37]

Most of outbreaks concern public catering and restaurants, pubs, street vendors, takeaway and canteens [39, 40, 38], where the main food vectors are Crustaceans, shellfish, molluscs and products thereof [37]. However, similar number of outbreaks were reported from domestic settings in EU [37], underlying the importance of proper HACCP implementation in public catering. It is also important to take into account that domestic settings outbreaks are not always reported, and therefore difficult to estimate accurate numbers.
In France, grouped cases, defined by the apparition of at least 2 similar cases, generally gastrointestinal, where the cause of which can be traced to the same food origin, are called TIAC (Toxi-infections alimentaires collectives). 1,783 TIAC have been declared in France in 2019, affecting more than 15,641 people, of whom 609 (4%) were admitted to hospital and 12 (0.08%) died (Table 2.2) [41]. The most frequently confirmed pathogen was *Salmonella* with 139 TIAC (36% of TIAC with confirmed agent), followed by *Bacillus cereus* (16%) and *Campylobacter* (14%). 54% of TIAC declared in France are detected in domestic settings. As in Europe, the consumption of shellfish is the most responsible of TIAC (13%), but the suspected foods are multiple and do not allow to suspect a particular food category.

The research activities of the present PhD thesis focus mainly on *Salmonella* which is a relevant pathogenic bacteria in France (Table 2.2) and Europe (Table 2.1).

	Declared cases		Sick cases		Hospitalised cases	
Agent	N	%	N	%	N	%
Total confirmed agents	**390**	**22**	**4,577**	**29**	**281**	**46**
Salmonella	139	36	807	18	161	57
Campylobacter	55	14	241	5	23	8
Bacillus cereus	62	16	988	22	37	13
Staphylococcus aureus	17	4	107	2	8	3
Clostridium perfringens	39	10	957	21	10	4
Norovirus	49	13	1342	29	18	6
Histamine	8	2	41	1	4	1
Diarrheic Shellfish Poison	1	0	3	0	0	0
Other pathogens	20	5	91	2	20	7
Total suspected agents	**1,102**	**62**	**8,789**	**56**	**227**	**37**
Total undetermined agents	**291**	**16**	**2,275**	**15**	**101**	**17**
Total	**1,783**	**100%**	**15,641**	**100%**	**609**	**100%**

Table 2.2: Detailed report of TIAC declared in France in 2019 from Santé Publique France [41]

2.1.3 Surveillance of clinical and foodborne cases of *Salmonella* in France

French surveillance of human cases is carried out by the National Reference Center (CNR for "Centre National de Référence") for *Salmonella* located at the Institut Pasteur in Paris, while French surveillance for food contamination is carried out by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) [42]. The CNR-Institut Pasteur Paris analyzes and serotypes strains sent by hospitals and medical analysis laboratories, and collects information on strains to follow the evolution of the number of *Salmonella* strains isolated in humans, and to detect clustered cases. The CNR receives nearly 10,000 strains of *Salmonella* every year, including around 400-600 from babies under the age of one.

For surveillance of non-human strains, the ANSES developed a food safety network providing passive monitoring throughout the food chain called the *Salmonella* network [43]. It is the counterpart to the human *Salmonella* network, and thus participates in the food safety system. The network was officially created in 1997 and today includes nearly 150 public and private veterinary laboratories in 94 departments across France. It is coordinated by ANSES's Maisons-Alfort Laboratory for Food Safety, which receives *Salmonella* strains of non-human origin from the various partner laboratories for serological typing (serotyping and/or molecular serotyping), and collects epidemiological informations on *Salmonella* strains that have undergone serotyping by the partner laboratories. One of the main objectives of the network is to collect and characterise *Salmonella* serotypes of non-human origin isolated from the food chain nationwide, in order to analyse geographical changes and differences over time. Each year, the *Salmonella* Network collects and centralizes epidemiological information (date isolation, pathway, matrix, isolation context, etc.) on approximately 12,000 isolates of *Salmonella*. Also, the *Salmonella* Network provides a source of information on rare serotypes and those not covered by regulations.

These surveillance data are reported to Santé publique France (SpF) to monitor the salmonelosis incidence, as well as emerging serovars and outbreaks. In addition, there is an active surveillance of TIAC cases, which can be declared by doctors, managers of collective or social catering establishments or consumers who are aware of an episode that may be a TIAC. Since 1987, TIAC have been subjected to mandatory declaration. SpF centralizes the mandatory declarations of TIACs notified to the departmental health authorities from departmental surveillance agencies (DDPPs - Direction départementale de la protection des populations), as well as to the regional health agencies (ARS - *Agences Régionales de Santé*). Concerning the food monitoring, the DGAL is responsible for the surveillance of zoonoses in the food chain and of foodborne diseases. It draws up the regulations that govern its core tasks and verifies their proper application, working through decentralised services in France's departments and regions. In order to elaborate surveillance plans and control plans the DGAL collaborates with the Direction Générale de la Santé (DGS), Direction générale de la Concurrence, de la Consommation et de la Répression des fraudes (DGCCRF), SpF, ANSES and national reference laboratory (LNR for "Laboratoire naltional de référence") for *Salmonella*. In the event of a health alert, the *Mission des urgences sanitaires* (MUS) plays a national leadership role and coordinates health alerts at the national level by receiving information on product non-compliances and human case reports from the various actors presented above.

In France, TIAC and grouped cases are detected on the basis of the mandatory reporting system (DO - *déclaration obligatoire*), and in parallel by the CNR system surveillance (about 2/3 of *Salmonella* samples detected in humans have been detected at CNR [44]). In case of a foodborne outbreak, SpF decides to investigate or not (depending on the epidemiological context), in connection with the MUS regarding the alert elements in the food chain. SpF also contacts the ANSES and the *Salmonella* network to investigate a potential food contamination as the origin of the TIAC. If some food isolates are suspected to be linked to the foodborne outbreak, SpF centralizes the strains, and sequence them (section 2.2) to check their linkage. In parallel, SpF conducts a survey investigation to identify the most likely source of contamination. When the contaminated food product is identified, it is removed from the market, or recalled if it has already been sold. Then, SpF reports the investigation to the field actors with the DDPP to investigate the cause of contamination and the suppliers of the product.

Finally, other platforms are also being developed in parallel, such as the SCA (*Surveillance de la Chaine Alimentaire*) platform, which brings together the main players in health monitoring of the food chain, both public and private, with the aim of providing methodological and operational support for the design, deployment, animation, promotion and evaluation of health monitoring systems.

2.1.4 Surveillance of *Salmonella* at the European Level

In addition to the national surveillance systems, a European surveillance system has been implemented in response to the expansion of worldwide trading favoring inter-country exchanges of contaminated food.

Between 52,702 (in 2020) and 94,425 (in 2016) salmonellosis cases are reported each year [37], estimating an overall economic burden of human salmonellosis as high as 3 billion euros per year. To protect consumers from *Salmonella* and other pathogenic bacteria, the EU has adopted an integrated approach for food safety from farm to fork. The approach consists of both risk assessment and risk management measures involving all key actors: EU Member States, the European Commission, the European Parliament, the European Food Safety Authority (EFSA) and the European Centre for Disease Prevention and Control (ECDC). The approach is supported by timely and effective risk communication activities, and helped to reduce human cases of salmonellosis in the EU by almost one-half over five years (2005-2009). In 2003, the EU set up an extended control programme for zoonotic diseases, with *Salmonella* as a priority. European surveillance of *Salmonella* is mainly controlled by ECDC and EFSA. While ECDC analyse and interpret data from EU countries on 52 communicable diseases and conditions, EFSA is focused on diseases and emerging risks associated with the food chain. EFSA's findings are used by risk managers in the EU and the Member States to monitor the situation, to define control measures and to set or review reduction targets for *Salmonella* in the food chain. They are also used by risk assessors such as EFSA's Panel on Biological Hazards to provide risk estimates. EFSA evaluates the food safety risks of *Salmonella* and provides scientific advice about control options in response to requests from risk managers or its own initiative, especially with the EFSA report which proposes an overview of important pathogens in Europe every year [37]. It also assesses the impact of setting new EU-wide reduction targets for *Salmonella* in various animals. Countries participating in EFSA monitoring report their cases to make global estimation of foodborne and zoonotic pathogen prevalence in Europe.

EFSA and ECDC cooperate and coordinate their work in accordance with their respective mandates under their Founding Regulations and other relevant legal acts. The benefits of this cooperation are in the areas of food safety, control of communicable diseases, infectious diseases prevention and emergency response. For instance, EFSA and ECDC assessed a multi-country outbreak of *Salmonella* Enteritidis infections linked to eggs in EU countries [45]. More recently, EFSA and ECDC collaboration was able to identify the cause and the extent of contamination of chocolate products that generated a multi-country *Salmonella* outbreak [46].

Finally, to connect all actors together, network like RASFF (Rapid Alert System for Food and Feed), created in 1979, make it able to share information between different EU members and organisation or commissions, to provide a round-the-clock service to ensure that urgent notifications are sent, received and responded to collectively and efficiently [47, 48]. Thanks to RASFF, many food safety risks had been averted before they could have been harmful to European consumers. Vital information exchanged through RASFF can lead to products being recalled from the market. A robust system, which has matured over the years, RASFF continues to show its value to ensure food safety in the EU and beyond.

2.1.5 *Salmonella* control strategies

In Europe, the top five *Salmonella* serovars involved in human infections were distributed as follows : S.Enteritidis (48.7%), S.Typhimurium (12.4%), monophasic variant of S.Typhimurium (1,4,[5],12:i:- described in section 2.3.1) (11.1%), *S.* Infantis (2.5%) and *S.* Derby (1.2%) [37]. In 91% to 95% of cases, *Salmonella* is transmitted to human by food consumption [49]. In France, the products mostly frequently associated with food poisoning by *Salmonella* are eggs (20%), meat (10%), dairy products (7%) and poultry (7%) [44]. The three most commonly food vectors involved in strong-evidence foodborne salmonellosis outbreaks were "eggs and egg products" followed by "pig meat and products thereof" and "bakery products" [37]. In developed countries, *Salmonella* was rarely reported officially in water-borne outbreaks despite it was frequently detected in surface waters including recreational waters and waters used for irrigation or as a drinking water source [50]. The top five major sources responsible for human infections are broilers, cattle, turkeys, laying hens and pigs, isolated from both animals and food, with a panel of 17,877 serotyped isolates from food and food-producing animals isolated in 2020. Compared to humans, broilers are most of the time asymptomatic, thus representing a major difficulty to detect the infection at the slaughter step. [51].

As *Salmonella* is a risk for human health, contaminated products with *Salmonella* enforces product withdrawals as defined by the European Parliament Regulation (EC) directive n°2073/2005 and an obligation to take corrective measures to control and reduce risks for the producer. These measures have a significant economic cost for the food industry and distributors. Control measures are framed by the zoonosis directive of the EC No 2160/2003 and of the Council of 17 November 2003 which aims to ensure the detection and control of *Salmonella* at every stage, particularly during all stages of the food chain and in animal feed, to reduce its prevalence and the risk to public health [52, 53]. It establishes a coordinated approach with the consultation of EU country to propose regulation for each animal group (breeding flocks of chickens in 2007, laying hens in 2008). The purpose of these checks is to guarantee the microbiological safety of products dedicated to the human consumption. For information purposes, EU 2073/2005 imposes the absence of *Salmonella* in 25g for a large number of products intended for human consumption, except for minced meat and mechanically separated meat where 10g is imposed.

EC regulation does not exist in pig farming. In France, the EU 2073/2005 and DGAL/SDSSA/ 2014-860 imply ISO standards related to the sampling of pig carcass at the slaughterhouse. However, a detailed guide of pigs measures "Analysis of the costs and benefits of setting a target for the reduction of *Salmonella* in breeding pigs" [54] was published by the FCC Consortium for the European Commission SANCO/2008/E2/056 in 16 march 2011. This study has estimated that the cost of human salmonellosis attributed to the pig and pork sectors was 86.1 million euros per annum across the EU (in 2011). This corresponded to 600 euros per human case. It proposed several control measures with different scenarios models, especially in feed sourcing and monitoring, to replace contaminated breeding pigs by *Salmonella* free pigs from a trustworthy supplier (which provides a *Salmonella* -free certification), quarantine procedure for new arrival livestock, regulation veterinary checks and vaccination, bio-security measures as cleaning and disinfection practices. The measures proposed here benefit both human health and pig production, and ensures a lower cost than the management of a *Salmonella* contamination crisis over the long term.

In the French dairy industry, EU 853/2004 imposes the elimination of milk from sick animals with clinical symptoms such as fever, diarrhoea and for females having aborted until the end of vulvar discharge [55]. In a case of a sale, the seller must declare on the health certificate if the cattle comes from a herd where there have been cases of clinical salmonellosis. The regulation describes that the case of salmonellosis must not have occurred over a period of 2 months or less, or less than 6 months when at least 2 cattle cases with clinical salmonellosis were detected. Also, as part of the *Brucellosis* surveillance, abortions must been declared to the sanitary police. Finally, if dairy foods are found in human cases, local authorities may take any action to help maintain public safety. As with pork, regulatory and monitoring practices depend on the country.

To tackle the *Salmonella* issue and reinforce *Salmonella* regulation, surveillance platforms have been developed to identify the main sources of contamination, to improve source attribution speed, and also limit the contamination spread in case of outbreaks.

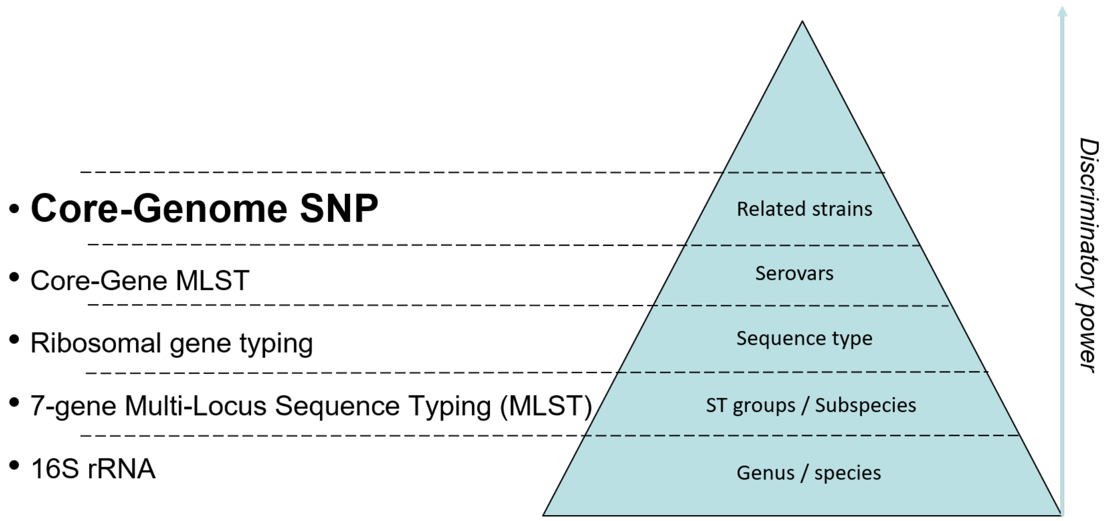


Figure 2.2: Genetic markers for taxonomic purposes. Description of WGS methods and their resolution level. Adapted from Ranieri et al. [59].

2.2 Comparative genomics approaches in food safety

Faced with the persistence and risk of *Salmonella* (section 2.1.4), whole genome sequencing approaches, able to pin down mutations at the nucleotide level, have gained interest while high-throughput sequencing and bioinformatics have enabled popularization.

The first DNA sequencing method (which made it possible to determine the exact base pair sequence of a DNA fragment) was developed by Maxam-Gilbert and Sanger as early as 1977. It did not initially find broad application in microbial typing as it had a low discriminatory power at the genomics scale and was classified as a specialized and expensive method [56]. However, with the development of high-throughput sequencing (HTS) technologies in the 2000s, also called next generation sequencing (NGS), the whole genome sequencing became more affordable and rapid for detection and typing of foodborne pathogens.

Different sequencing and analytical methods were developed to increase discriminatory power and decrease cost (Figure 2.2). Firstly, bacterial sequencing methods were able to identify the genus and/or the species based on 16S rRNA. Then, one of the most successful sequence-based typing approaches, multilocus sequence typing (MLST) [57], has been able to define subspecies or sequence type groups of bacteria (see more description in the section 2.2.1). Other methodological approaches, like ribosomal gene typing [58], have been able to increase the discriminatory power with a more precise typing through the indexing of variations across 53 genes encoding the bacterial ribosomal protein subunits.

Nowadays, core-gene and whole-gene MLST along with coregenome SNP methods have been proposed as methods of choice to characterize *Salmonella* strains [59]. In addition, NGS sequencing has been implemented in different countries as a routine method (especially in high income countries) [60], allowing a constant increasing of outcome accuracy and *Salmonella* genomes available in international archives. Nowadays, WGS is a helpful tool for *Salmonella* phylogenomic analysis, detection of antimicrobial resistance (AMR) genes and virulence gene predictions [61].

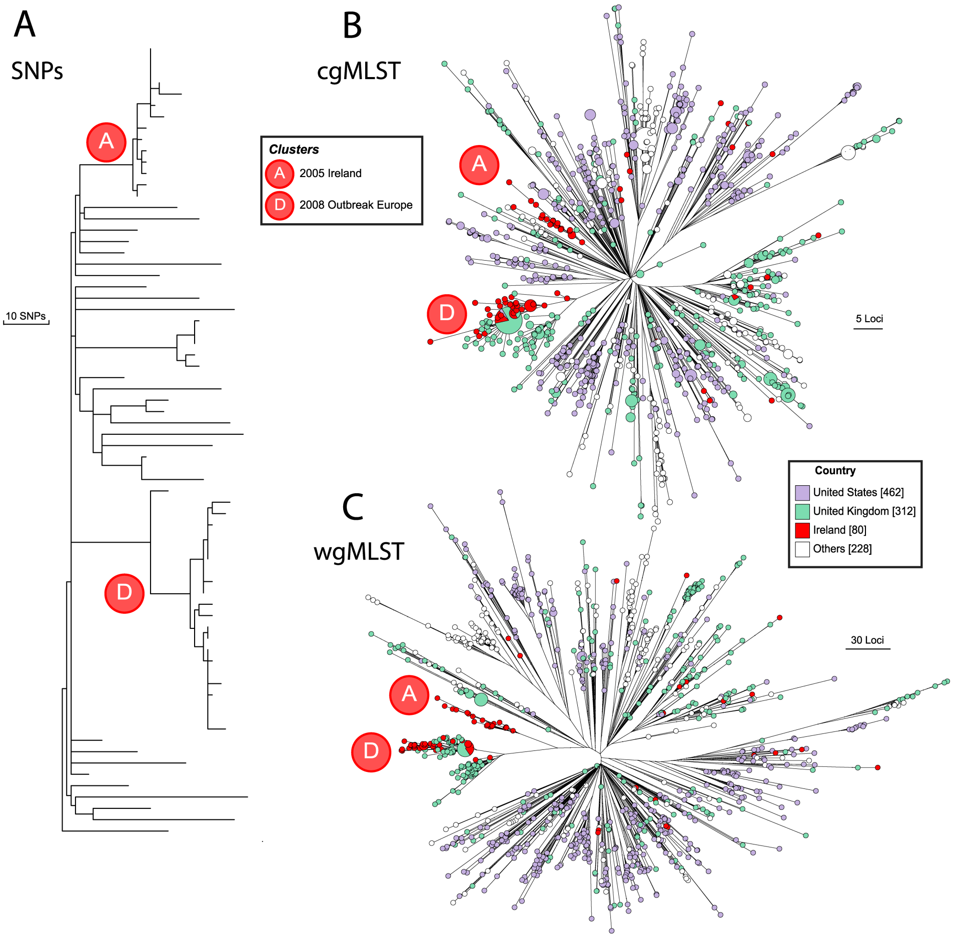


Figure 2.3: Comparison of coregenome SNPs (A), cgMLST (B) and wgMLST (C) methods from Alikhan et al. [62] . (A) Maximum parsimony genealogy of 73 genomes of serovar Agona based on 846 core SNPs. (B) GrapeTree [63] of cgMLST (3002 loci) from 1082 Agona genomes in EnteroBase. (C) GrapeTree of wgMLST (21065 loci) of the same genomes as in part B.

Here, we will describe the latest methods used in characterization of *Salmonella*.

2.2.1 Multilocus sequence typing

MLST was first described by Maiden et al. across a few loci within a *Neisseria meningitidis* dataset, and proposed this technique for molecular typing and characterization of bacterial species in a context of inter-country surveillance system [64]. The characterization is based on the sequences of internal fragments of (usually) seven house-keeping genes. For each house-keeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST). Different schemes have been proposed, and the two of the most popular are the pubmlst scheme [65] and Enterobase scheme [66]. An advantage of this approach is that, as with MLST, loci used in the schemes are readily maintained and easily shared among laboratories using the same or similar online databases.
Today, the MLST method together with precision provided by WGS and development of new algorithms provide powerful tools for international surveillance of pathogens [67]. Indeed, bioinformatics tools have been developed using MLST scheme at the core and whole genome scales.

The coregenome (cg) MLST is a MLST approach including loci that are present in all isolates of a given population or a subset thereof [57]. Also, whole genome (wg) MLST has been developed to take into account a selection of accessory loci as well. This resulted in a deeper analysis than the historical MLST performed on few loci. Results are then analysed in a phylogenomic tree, or in a minimal spanning tree representing the diversity of the dataset. Phylogenomic trees are trees representing the evolutionary relationships among various biological species. An example is represented in Figure 2.3 (A). The branch length represents the evolutionary time between two nodes (or isolates). The evolutionary distance between two nodes is computed as the sum of all the horizontal branch lengths (in an horizontal tree). A minimum spanning tree is a subset of weighted edges that connect all isolates in a graph. Only edges that are the total sum of edge weights that connect the entire graph are displayed. Examples are displayed in (Figure 2.3 (B) and (C)).

2.2.2 k-mer

In genomics, *k*-mers are substrings of specific length *k* contained within a biological sequence. Analysis of *k*-mers is frequent and widely used to estimate the genome size [68], copy number and repetitive sequences [69], short read assemblies with De Brujin graphs [70] and genomic distance estimation [71, 72, 73]. The advantage is that the genomes are analysed without resource consuming alignment-based algorithms. Comparative genomics methods using *k*-mers (Figure 2.4) relies first on *k*-mers counting for each genomes, where the length of *k* was determined before, and *k*-mers sets for each genomes are compared to calculate pairwise distances. These distances can be used as such, but also clustered to visually determine the distances between genomes. *k*-mers methods has significant advantages, such as easy addition of a new genome into a comparative analysis [72], and and allow *k*-mer profiles to be processed across genomes, taking into account accessory loci [74]. For the record, a kmer voting approach was recently used to build a cg/wgMLST workflow [75].

2.2.3 Coregenome SNPs

One way to exploit WGS data is the identification of single nucleotide polymorphisms (SNPs) that vary among isolates. Single nucleotide polymorphisms, frequently called SNPs, are the most common type of genetic variation among isolates. Others variations, like CNV (Copy number variations) or SV (Structural variants) define large genomic alterations, and are not as precise as SNPs. Each SNP represents a difference compared to a reference in a single DNA building block, called a nucleotide. SNPs can be highly informative markers, which are capable of revealing evolutionary history of homogeneous segments [77, 78], as well as detecting and tracing outbreaks [27, 79, 80, 81]. InDels (insertions or deletion) are also detected along with SNPs, but are known to be more challenging to call with accuracy than SNPs [**redelings_incorporating_2007** , 82, 84]. It has been showed that some serovars have a higher mutation rate than other. For example, the mean mutation rates were estimated at 2.2 × 10-7 substitutions (1.01 SNPs per genome per year) for *Salmonella* Enteritidis lineages [85] or 9.3×10–8 per nucleotide/year for the accumulation of core SNPs (or 0.44 SNPs per genome/year) for *S.* Agona [86].

It is difficult to extrapolate these results because *Salmonella* is under pressure depending on its environment. It was shown that increasing number of SNP differences was observed when *Salmonella* is under pressure, and the genetic diversity within a *Salmonella* serovar depends upon sevorar [87]. Also, *Salmonella* have an adaptation tool to its vector host thanks to horizontal transfers. Some strains can also acquire new genomics content that other lineages

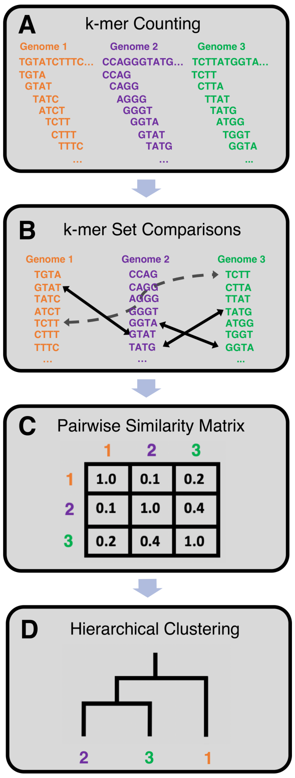


Figure 2.4: Example of workflow for alignment-free genome comparisons using *k*-mers detection. Figure from Bussi et al. [76]. (A) Genomes are processed from left to right with a sliding window of fixed length resulting in *k*-mer databases. (B) *k*-mer set comparisons are then computed for pairs of genomes. Arrows indicate the *k*-mers shared between *k*-mer sets. (C) From the set comparisons, similarity scores are calculated resulting in a pairwise similarity matrix. (D) Hierarchical clustering of the similarity matrix yields a tree.

do not have, called accessory genome. Today, SNP-based methods focus on core SNPs because this analytical methods were originally developed to perform on genomes harboring less horizontal gene transfer (HGT) [88] and also due to some limits discussed in section 2.2.8.

Coregenome SNP tools were mainly developed as their discriminatory power is higher than cgMLST [89], and different pipeline emerged during the time [90, 91, 92, 27, 93]. Coregenome SNP tools relies on variant calling analysis [94, 95] between raw data from sequencing and a reference genome, and then comparing the SNPs between all isolates. Raw data from sequencing are mapped against a reference genome [96, 97, 98], and variants are detected for each position across the genome. Then, using variants, comparison between all isolates are visualized within a phylogenomic tree (Figure 2.3 (C)). SNPs are also analysed for AMR (Section 2.3.5.5) or host adaption [98].

2.2.4 Pangenome genes

Workflows were developed to extract all genes from the pangenome taking into account core and accessory genes. Core genes correspond to genes shared by all strains in a dataset, while accessory genes correspond to unique genes or genes shared by some strains. For example, Roary [99], or more recently Panaroo [100] allow an exploration of the content of core and accessory genes. Roary performs a hierarchical clustering based on a accessory gene patterns and provide aligned coregenes for downstream character-based phylogenomic reconstruction (i.e. based in SNPs). Contrary to Roary which does not take into account non-coding part of the genome playing an important role in regulation [78, 101], Panaroo is able to deal with intergenic regions [102].

2.2.5 Pangenome SNPs

Pandora [102] uses graph-based reference on genes or intergenic regions to build a pangenome, and can identify variants across the full bacterial pangenome. Here, the pangenome corresponds to the core and the accessory genomes (coding and non-coding). Pandora is efficient on distanced related samples, where a single reference can not represent the whole panel. Also, it can recover more rare SNPs, in the core panel or in the accessory panel. Unfortunately, variants detected by Pandora are in the form of a graph, and cannot be used for phylogeny, especially since some accessory variants have a different evolution rate that must be taken into account.

2.2.6 Genomic comparison using phylogenomic methods

Phylogenomic methods that explore the relationships between microbial genomes have been used to study the emergence [30], geographical diffusion [103] and transmission of infections [104]. Phylogenomic topologies can also be informative in terms of source attribution during outbreak investigations [30].

Given a collection of aligned genomes or genes, a phylogenomic tree can be built based on different tools. Some fast and simple tools, such as fasttree, are able to estimate approximate maximum likelihood trees, however they may have limited accuracy [105]. Maximum likelihood (ML) phylogenomic tree providing large number of evolutionary models and boostrap setting can be inferred with RAxML [106] or IQ-TREE [107]. Phylogenomic tree based on Bayesian methods (Markov chain Monte Carlo method) can be inferred with MrBayes [108], or also other tools like BEAST [109] which focus on timescaled trees. Another tool, ClonalFrameML [110] allows the phylogenomic inference of recombinant bacterial species while mitigating the effect of horizontal sequence transfer on phylogenomic reconstructions. This tool is also used on *Salmonella* to remove recombinant SNP, as phylogenomic reconstruction based on nucleotide substitution models are supposed to only take into account points mutations [111].

Finally, if some alignments are segmented, or focused on different locus, reconciliations phylogenomic methods has been developed. Supermatrices refers to the concatenation of multiple sequence alignments from different genes/sequences. Unavailable genes/sequences constitute the so-called missing data in the supermatrix. Nowadays, these methods are implemented in Bayesian and ML methods, and can take into account the variation of evolutionary rates all over the sequences. To account for different evolutionary scenarios, partition models were introduced to allow application of different substitution models to different genes substitution models [112]. In opposite, supertree methods, combining inferred trees (called "source trees" or "subtrees") into one “supertree” have been developed to tackle the time consuming method of supermatrices due to large and potentially heterogeneous datasets. Divided into overlapping smaller subsets, trees are estimated on each subset, and then combined into a single tree representing the full dataset using a supertree method [113]. These methods have also been recently applied on *Salmonella* dataset [114].

2.2.7 Using WGS to develop fast tracking methods

If WGS methods have been widely used in outbreaks and epidemiological investigation, this method is still expensive (a bit less than 100 euros/sample in routine analysis laboratories [115]) and time consuming (up to 2 to 3 days for DNA extraction, sequencing and then analyse the results) for routine source attribution. PCR and other molecular typing techniques are still used because they are fast and allow development of new PCR markers targeting sources without having to explore systematically the entire genome with WGS. But with the new advances of WGS, methods that can develop new primers have grown. Genomes are now explored to find patterns for source attribution, and these patterns are identified to develop fast, cheap and simple markers to trace back the infection reservoirs (i.e. regions or sources). Primer-BLAST [116] is one of the most useful software to design target-specific primer for PCR. It allows the user to design PCR from a sequence of interest and proposes specific forward and reverse flanking primers. Also, experimental properties are taken into account to ensure the matching with the sequence of interest. A good requirement is that primers should have :

- Length of 18-24 bases
- 40-60% G/C content
- Start and end with 1-2 G/C pairs
- Melting temperature (Tm) of 50-60°C
- Primer pairs should have a Tm within 5°C of each other
- Primer pairs should not have complementary regions

Additional requirements may also be applied in certain cases. For example, to avoid unwanted amplification of genomic DNA in reverse transcription PCR (RT-PCR), it is recommended that a primer pair span an intron or that one of the primers be located at an exon-exon junction. Another concern is the possible impact of SNPs in the primer regions. Since SNPs may act as a mismatch in some cases, one should consider picking primers outside of such regions.

To select these regions, different strategies have been developed. Some strategy which rely on conserved sequences from a set of target isolates against a set of exception isolates [117]. Some other strategies are based on kmer [118], micro satellites or low variable regions [119, 120] to ensure the stability of the primers. Finally, primers are most of the time designed from genes of interest (virulence, host adaptation, AMR) [121], or serovar under strong regulation in food chain [122, 123, 124]. In the latter, multiplex PCR were developed focusing on flagellar and lipopolysaccharide genes to quickly determine known serovars of a particular host. Unfortunately, multiplex PCR are under stringent selection as incorporating more than five to six primer pairs in a single reaction becomes a challenge due to the increasing difficulty in optimising PCR conditions and difficulty in differentiating PCR product sizes by agarose gel electrophoresis [125]. Hence, multiplex PCR are limited, and primers should be selected carefully with only one source target to avoid confusion in results.

2.2.8 Limits

While WGS is more sensitive and specific than conventional typing methods (MLVA), there are still barriers to uptake for genomic surveillance around complexity of reporting of WGS results, timeliness, acceptability, and stability [126]. Sequencing can contain many errors, including contamination [127], which must be detected. These errors can influence the identification of

SNPs. Also, in term of bioinformatics methods, it was observed that cgMLST present a high discriminatory power and is able to distinguish outbreak and non-outbreak strains in agreement with the epidemiological data for some *Salmonella* serovars which are heterogeneous enough. However, it cannot discriminate outbreak and non-outbreak strains for homogeneous strains (*Salmonella* Typhimurium for example) [27, 79].

SNP typing can be extremely powerful, but the current existence of more than 20,000 genomes of each of Typhimurium or Enteritidis in EnteroBase makes computationally difficult a such approach. While SNP-based methods present the highest resolution, cgMLST is less time consuming and can be easily used on outbreak detection with high sensitivity [62]. Many public health agencies rely on these techniques, with a risk of letting out some true positives samples.

Finally, even the most resolutive approach (SNPs from coregenome or pangenes) displays limitations. These limitations are associated with phylogeny-based methods only handling variations that are present in all of the considered samples, due to transition models lacking the frequencies to the unknown and absent states. This restricts variation analyses to the minimal common sequences shared by all compared samples. Subsequently, the higher the number of samples, the higher the probability to shrink down the "shared by all" base of comparison, lowering resolution and depriving downstream analyses such as source attribution or Genome-Wide Association Studies (GWAS) from a substantial amount of candidates. On the other hand, pangene SNP approaches relies only on presence/absence of genes, letting out of the analysis SNPs. Pandora, the only tools taking into account accessory SNPs, rely on multiple alignment, which is time consuming, and pangenome variation output is not compatible with phylogenomic inference methods [102]. In addition, it does not solve the issue about missing data during phylogenomic inference.

To date, the WGS methods in production for conducting foodborne outbreaks investigations and source attribution rely on coregene analysis or coregenome SNPs and allowed to improve the detection of outbreaks [79, 128]. These methods also refined epidemiologic investigations to improve source identification and enable identification of more outbreaks at earlier stages. However, this means that a part of the genomic diversity signal is not used and weakens any conclusion drawn.

2.3 The many faces of *Salmonella*

Salmonella is a major global bacterial pathogen, highly polymorphic in its diversity of host range, clinical manifestation and outcome. Its impact on public health and its economical burden have continuously been driving efforts to understand the situation or reduce its consequences, historically by leveraging the most suitable methods available at the time. This part outlines the uniqueness and diversity of *Salmonella*, and state of knowledge after a century of research.

2.3.1 The taxomomy of *Salmonella*

The genus *Salmonella* is a gram-negative Enterobacteria and a common foodborne pathogen with global public health concerns, leading to 52,702 cases of gastroenteritis in Europe in 2020 [37, 129]. The genus name *Salmonella* had been adopted in the honour of Dr. D.E. Salmon, who discovered *Salmonella* from the intestine of a pig in 1885 infected by the "hog cholera bacillus", considered to be the causative agent of swine plague [130]. It was later on found to

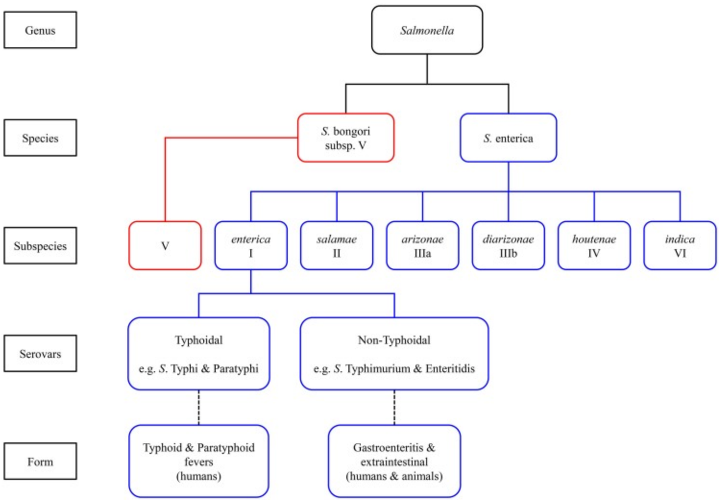


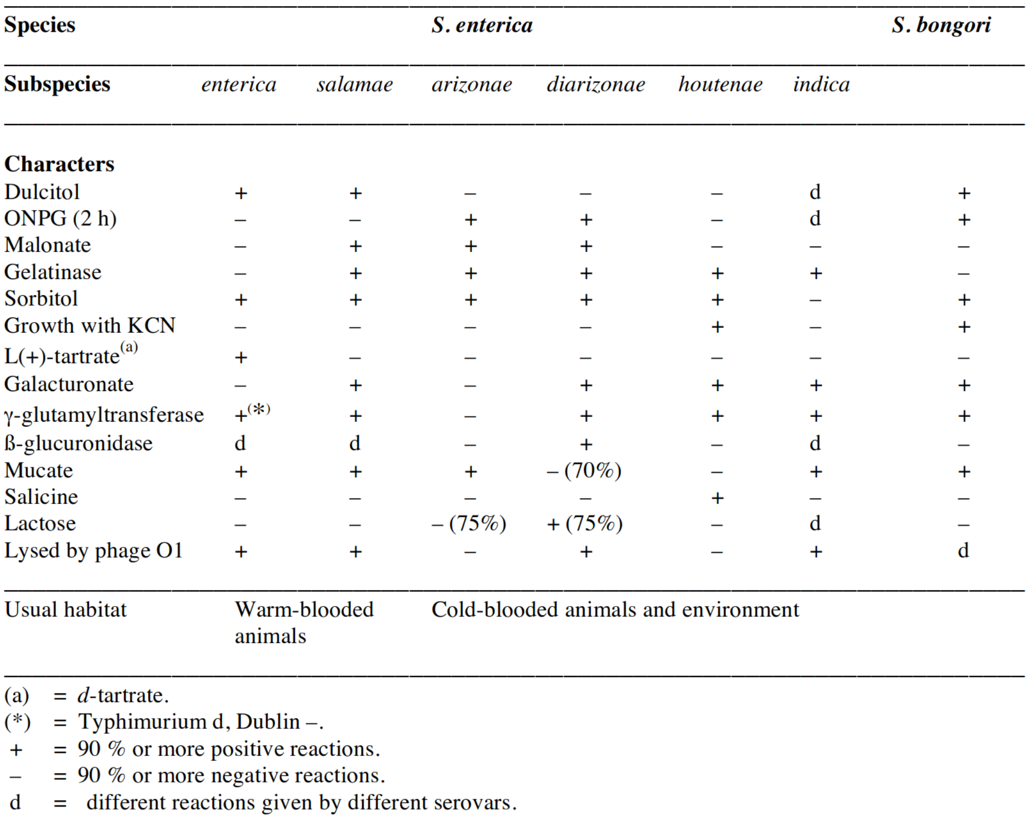
Figure 2.5: Classification of *Salmonella* species and subspecies from Hurley et al. [133].

be only a secondary invader and was named as *S.* Choleraesuis. Eversince, *Salmonella* strains have been isolated from a wide host range, causing infectious diseases in both humans and animals [131] with distinct syndromes [132].

Molecular methods based on 16S RNA genes sequences have shown that the genus *Salmonella* consists of two species, *S. enterica* and *S. bongori* (also referred to as subsp. V) [134]. *Salmonella enterica* is divided into the following six subspecies : *S. enterica* subsp. *enterica*, *S. enterica* subsp.*salamae*, *S. enterica* subsp. *arizonae*, *S. enterica* subsp. *diarizonae*, *S. enterica* subsp. *houtenae* and *S. enterica* subsp. *indica* (Figure 2.5). *S. enterica* subsp. *enterica* is the most well-represented species, accounting for approximately 60% of all serovars identified and greater than 95% of *Salmonella* isolates obtained from humans and domestic mammals [135]. *S. enterica* subsp. *enterica* is biochemically differentiated into serovars based on the composition of their carbohydrate, flagellar, and lipopolysaccharide (LPS) structures. All *Salmonella* serovars can be designated by an antigenic formula proposed by Kauffmann, based on somatic (O) and flagellar (H) antigens in addition to capsular (Vi) antigens. *Salmonella* includes more than 2,600 serovars, which differs in host adaptation and virulence [136]. Some serovars are host-specific, meaning they can only cause disease in one species. For example, the serovar Gallinarum is specific to poultry. Other serovars can be host restricted, which means they are predominantly associated with one host species, but can cause disease in other species as well. The serovar Dublin, for example, is generally associated with severe systemic disease in cattle but may also infrequently causes disease in humans [137]. Finally, some serovars are generalists, meaning they can infect and cause disease in a broad range of hosts. For instance, the host range of the serovar Typhimurium includes humans, livestock, domestic fowl, rodents, and birds [138, 136]. The *Salmonella* host range and adaptation can be driven by different genetic determinants related to the feeding environment of the animal, livestock diets, environmental stimuli, physiological properties of the animal, and work habits for livestock health protection [98].

2.3.2 Morphology and biochemical properties

Salmonella is a gram-negative rod belonging to the Enterobacteriaceae family. The size of the bacteria varies with a cell diameter between about 0.7 and 1.5 µm and length from 2 to 5 µm. They are aerobic and facultative anaerobes, and show predominantly peritrichous motility



Species	***S. enterica***						***S. bongori***
Subspecies	*enterica*	*salamae*	*arizonae*	*diarizonae*	*houtenae*	*indica*	
Characters							
Dulcitol	+	+	–	–	–	d	+
ONPG (2 h)	–	–	+	+	–	d	+
Malonate	–	+	+	+	–	–	–
Gelatinase	–	+	+	+	+	+	–
Sorbitol	+	+	+	+	+	–	+
Growth with KCN	–	–	–	–	+	–	+
L(+)-tartrate(a)	+	–	–	–	–	–	–
Galacturonate	–	+	–	+	+	+	+
γ-glutamyltransferase	+(*)	+	–	+	+	+	+
ß-glucuronidase	d	d	–	+	–	d	–
Mucate	+	+	+	– (70%)	–	+	+
Salicine	–	–	–	–	+	–	–
Lactose	–	–	– (75%)	+ (75%)	–	d	–
Lysed by phage O1	+	+	–	+	–	+	d
Usual habitat	Warm-blooded animals	Cold-blooded animals and environment					

(a) = *d*-tartrate.
(*) = Typhimurium d, Dublin –.
+ = 90 % or more positive reactions.
– = 90 % or more negative reactions.
d = different reactions given by different serovars.

Figure 2.6: Different biochemical characters of *Salmonella* species and sub species from Grimont and Weil [5].

(except for *S. e. G*allinarum and Pollorum) [139].
They grow radially, on simple media, over a range of pH between 3.8 and 9.5 and temperature between 5 to 50 °C (optimum pH=7 and 37 °C) [140]. Most of *Salmonella* perish after being heated to 60 °C for 3 min [141], although if inoculated in high fat, high liquid substances like peanut butter, they gain heat resistance and can survive up to 90 °C for 30 min [142]. The bacteria ferments glucose and can reduce nitrate. Some of the biochemical properties of most *Salmonella* spp. can help distinguish subspecies and serovar. These biochemical properties rely on the utilization of citrate, production of hydrogen sulphide from inorganic sulphur, decarboxylation of ornithine and lysine (Figure 2.6).

2.3.3 Characterization of *Salmonella* by typing methods

The characterization and typing of *Salmonella* has steadily increased along with the technology sweep. In the context of food safety control or food poisoning investigations, characterization of *Salmonella* serovars or sequence types is mandatory for source attribution. By analyzing and comparing how often a given pathogen occurs in food and comparing it to those isolated subtypes of humans and animals and/or animals production, it may be possible to make inferences about the sources of human infections [143]. Here, we will described the typing methods used in characterization of *Salmonella*.

2.3.3.1 Subtyping by slide agglutination

Salmonella serotyping method is based on the Kauffmann-White-Le Minor scheme [5]. The first publication of the Kauffmann-White scheme in 1934 listed 44 serovars [144]. The scheme then got several changes, containing 958 serovars following Kauffmann's retirement in 1964, 2,267 serovars following L. Le Minor's retirement and following Popoff's retirement there were 2555 serovars. Since L. Le Minor described most of the presently known serovars, it was proposed to change the name of Kauffmann-White scheme to White-Kauffmann-Le Minor scheme [145].

Serotyping is based on the agglutination of bacteria with specific sera to identify variants of the somatic (O) and flagellar (H) antigens [146]. The O antigen is the saccharidic component of the lipopolysaccharide exposed on the bacterial surface [147], encoded by *rfb* genes. Its reacting toward specific antisera forms the basis of the *Salmonella* serotyping scheme. Several O antigens may be expressed together at the surface of a single cell. The flagellar (H) is composed of two phage flagellar antigens, phase 1 and phase 2 encoded by *fliC* and *fljB* genes. Some serovars have the property to express only one flagellar phase, like TMV serovar. Isolates are assigned to serovar depending on the agglutination as a reaction against the antisera prepared against these antigens. It is still the method of first choice in routine and monitoring of *Salmonella* serovars [148], and also in public health investigation.

2.3.3.2 Antibody microarrays

Some developments aiming at identifying serovars were based on microarray method [149, 150]. These researches were motivated by parallel detection of multiple serovars to reduce analysis time compared to traditional serotyping. The main advantages of this method over traditional serotyping are the standardized agglutination detection, and simultaneous detection of the O and H antigens. In addition, the production and quality control of the hundreds of antisera required to generate more than 2,500 serotypes is difficult and time-consuming. At this time, the method is used on specific serovars for strains with significant interest in food safety. This method needs further development upon its successful validation on a much larger number of serovars.

2.3.3.3 Phage typing

Other methods based on phage identification has been developed to discriminate serovars of *Salmonella* [151]. The characterization is based on the phage lysis patterns between strains, and is highly valuable in epidemiological studies [152]. Phage typing is a rapid and low cost approach for the epidemiological surveillance and outbreak investigation for highly studied *Salmonella* serovar like Typhimurium [151, 152] or Enteritidis [153]. The advantage of phage typing resides in the simplicity of its implementation, which requires only basic laboratory equipment, but the method is also limited by the number of available phages and the fact that some serovars cannot be discriminated by this method.

2.3.3.4 Molecular Typing techniques

2.3.3.4.1 Pulsed-field gel electrophoresis

Pulsed-field gel electrophoresis (PFGE) was until recently the method of chose for the molecular characterization of bacterial pathogenic strains [154]. The method is based on enzyme restriction of bacterial DNA. Restricted DNA bands are then separated using a pulsed-field electrophoresis [155, 156]. Standard PFGE protocols have been developed for typing of different bacterial species including *Salmonella*, leading it to be one of the most widely used methods for phylogenomic studies, food safety surveillance, infection control and outbreak investigations. Potential drawbacks of this technique such as low throughput, time consuming and labor-intensive caused researchers to think about replacing it with WGS methods due to their higher resolution [157]. In addition, PFGE displays some limits in discriminating clonal serovars such as *S.* Derby [158].

2.3.3.4.2 Molecular typing based on genes and genomics markers

As somatic and flagellar antigens are encoded in DNA, it is possible to develop multiplex PCRs which can distinguish serovars based on the length of the amplified DNA fragments. Some studies focused on phase 1 and phase 2 H antigens where the higher detection score was 84.6% of 500 routines samples isolates [18, 159, 160]. Others studies focused on lipopolysaccharide of the O antigen *rfb* gene displaying promising results (94.3% O-antigen group correctly detected) with reduced time compared to traditional serotyping [161]. But it has been observed that the global diversity of *Salmonella* evolves independently in subspecies and serovars, and therefore markers have been developed focusing on others genes to determine *Salmonella* serovar. Multiplex PCR using gene marker focuses mainly on very persistent and abundant serovars to reduce the identification time of the most monitored serovars in public health security. These detection methods mainly focus on the presence of serovar specific genes to discriminate samples quickly [124, 123, 162].

2.3.3.5 Limits

Identification of *Salmonella* at the species level poses few problems. This is because most members of this genus share common biochemical profiles and a high level of genetic similarity [163]. However, further identification at the species and serovar levels requires expertise and extra resources. Indeed, the complete *Salmonella* nomenclature represents a very complex scheme, and scientists use different methods to describe and communicate about this genus, which can make the results confusing [164]. It should also be noted that some serovars share the same antigenic formula and require additional testing for unambiguous identification, e.g. the clinically important serovar *S.* Chloeraesuis shares its antigenic formula (6,7:c:1,5) with serovars S. Paratyphi C and *S.* Typhisuis [35].

Despite the utility of serotyping, problems associated with antiserum production and isolates for which the serovar antigen cannot be detected have led to take advantage of molecular approaches. With such assumptions, the discrimination power limits of serotyping by agglutination or pulsed-Field-Gel Electrophoresis (PFGE) are often challenged, especially on frequently isolated serovars like Derby or Dublin [59, 165, 166]. Furthermore, traditional serotyping is often not sensitive enough to provide the level of discrimination needed for foodborne illness outbreak investigations, and it cannot be used to infer phylogenomic relationships [59].

2.3.4 Physiopathology

Salmonella enterica is an invasive pathogen able to colonize the gut lumen from its host and multiply in it [167]. *Salmonella* colonizes the small and large intestine, causing gastroenteritis. Following ingestion and passage through the stomach, *Salmonella* encounters the intestinal

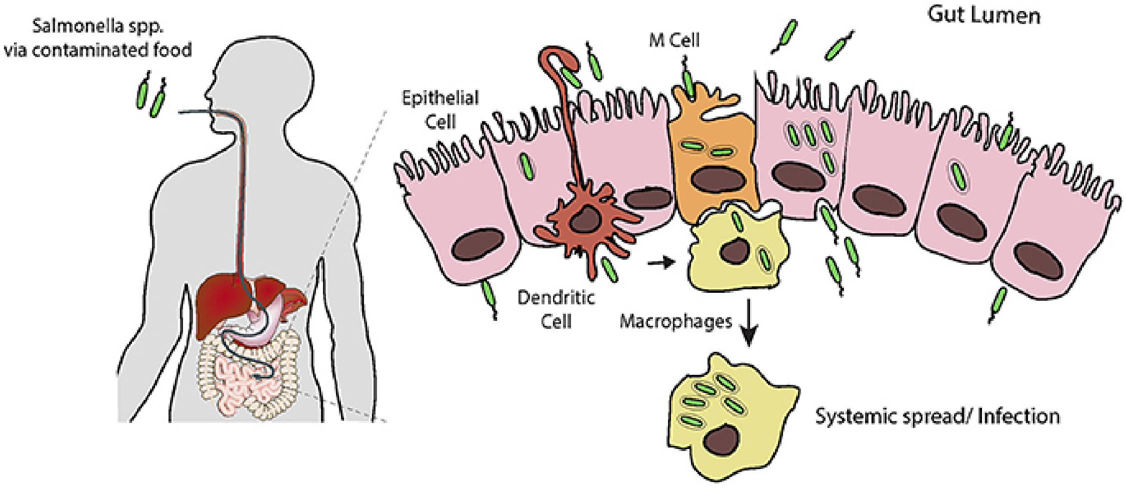


Figure 2.7: *Salmonella* infection described in Hume et al. [167]

epithelial layer (Figure 2.7). The epithelial monolayer underlying the mucus layer contains distinct cell types with different roles. M cells, the specialised antigen-sampling cells of the mucosal immune system, are capable of transporting luminal antigens to the underlying lymphoid tissues. These cells are exploited by *Salmonella* as the preferred route to invade the host [168]. The genetics underlying this strategy is found in *Salmonella* pathogenicity islands (SPIs) with gene clusters located at the large chromosomal DNA region and encoding for the structures involved in the invasion process [169] (section 2.3.5). The bacterial effectors then activate the signal transduction pathway and trigger the reconstruction of host cell's actin cytoskeleton, resulting in the outward extension of the epithelial cell membrane to engulf the bacteria [130].

Once the bacteria invade the intestinal mucosa, they replicate, releasing newly formed bacteria into the gut. *Salmonella* can survive and replicate long enough to allow systemic spread through the reticuloendothelial system. In more severe forms in immunocompromised individuals, the bacteria can invade further and spread to the bloodstream or the peripheral organs [170]. The evolution of the disease then varies according to the serovar.
Based on the clinical patterns in human salmonellosis, *Salmonella* strains can be grouped into typhoid *Salmonella* and non-typhoidal *Salmonella* (NTS).

2.3.4.1 Enteric fever

Salmonella Typhi is the causative agent of typhoid fever, while paratyphoid fever is caused by *S.* Paratyphi A, B and C [171]. Typhoid and paratyphoid fever present most often clinically similar illnesses and are indistinguishable without accurate diagnosis relying on laboratory confirmation [172]. Both *S.* Typhi and *S.* Paratyphi fever are called "enteric fever", and are referred to typhoid *Salmonella*.

Salmonella Typhi and *Salmonella* Paratyphi are adapted to human hosts and are considered as host-restricted serovars [173]. Humans can become ill by consuming contaminated food or water but also by direct transmission. They can carry the bacteria in their gut for very long periods (chronic carriers), and classical symptoms include gradual onset of sustained fever, chills, hepatosplenomegaly and abdominal pain. In some cases, patients experience rash, nausea, anorexia, diarrhea or constipation, headache, bradycardia and reduced level of consciousness [174].

While the number of cases declined thanks to the improvement of prevention and control strategies such as vaccination measures [171], it is estimated that between 11 and 20 million of typhoid and paratyphoid fevers cases occurred every year (WHO data, and 14.3 million cases in 2017 [175]). It was estimated that 135,000 deaths from typhoid and paratyphoid fever occurred globally in 2017 [175]. Low income countries are the most affected, particularly in Asia and sub-Saharan Africa.

2.3.4.2 Non-typhoidal *Salmonella*

Salmonella strains other than *S.* Typhi and *S.* Paratyphi are referred as NTS, and are predominantly found in animal reservoirs. *Salmonella* exhibits highly variable host range among the mammals and colonize the gut of various livestock animals such as poultry, pigs or cows. It can invade a broad range of hosts causing both acute and chronic infections. Infection with non-typhoidal *Salmonella* manifests itself through a broad range of clinical symptoms and can result in either asymptomatic carriage or gastroenteritis and in severe cases, death [176].
Most people with *Salmonella* infection presents symptoms such as diarrhea, fever, and stomach cramps but *Salmonella* strains can sometimes cause infection in urine, blood, bones, joints, or the nervous system (spinal fluid and brain) of immunocompromised hosts. Symptoms usually begin six hours to six days after infection and last for four to seven days. However, some people do not develop symptoms for several weeks after infection and others may experience symptoms for several weeks.

2.3.4.3 Host specificity

Salmonella speciation happened 25 to 40 million years ago, diverging from *Escherichia coli* lineage. Then, *Salmonella* most likely evolved in three phases [177]. during the first phase the microorganism acquired, through horizontal gene transfer, the SPI-1 which conferred it pathogenicity determinants. During the phase 2, two distinct species of *Salmonella* diverged : *Salmonella enterica* by acquiring SPI-2 to help surviving within host cells, and *Salmonella bongori*. Finally, the third phase is related to the division of *Salmonella enterica* to subspecies, and the adaptation of subspecies. Subspecies *enterica* has adapted to warm-blooded vertebrates (mammals, birds) [137], increasing the range of hosts, while the remaining subspecies are mainly adapted to cold-blooded vertebrates [178].
Among the different serovars of *Salmonella enterica* subsp. *enterica*, there is observed differences in host range. All animal species are susceptible to carry different *Salmonella* serovars with different clinical range. Most of the serovars belonging to subspecies *enterica* adopt a generalist behavior in clinical symptoms, but the susceptibility to host adaptation is variable, from serovar with single or restraints hosts to a greater number of hosts. For example, *Salmonella* Dublin and *Salmonella* Choleraesuis cause systemic disease in cattle and pigs, respectively, while other species or human could be infected accidentally, or through these vectors. However, in this example, infection is frequently asymptomatic, making them healthy carrier animals and shedders in the environment. A smaller number of serovars have only one narrower host range, generally causing more serious pathologies than a simple gastroenteritis [177] (Figure 2.8).

2.3.5 Pathogenomics

Salmonella is able to displays different virulence and resistance capacities due to its genomic adaptation [179, 180]. The colonization of hosts by *Salmonella enterica* depends on the

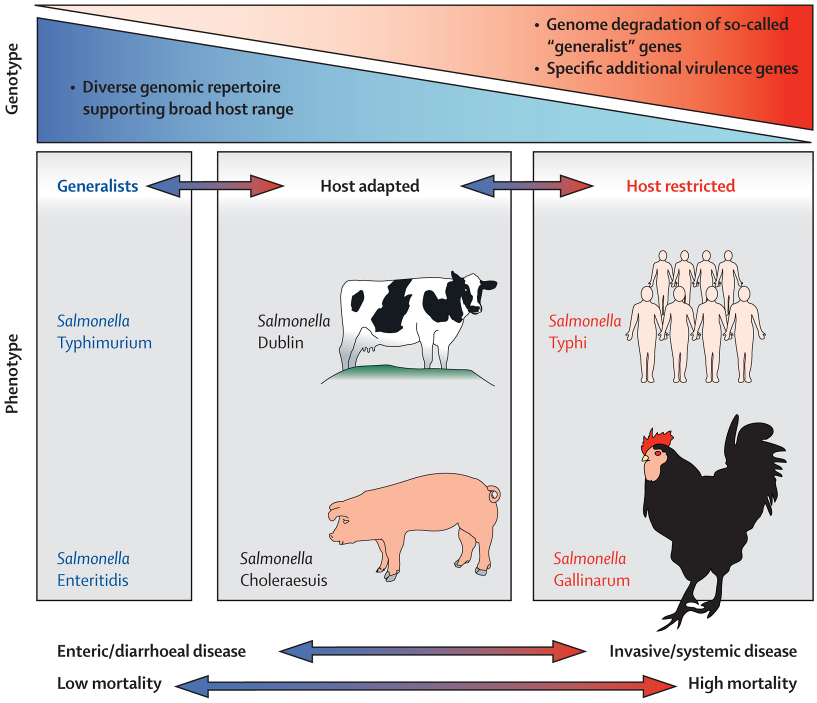


Figure 2.8: Host adaptation of *Salmonella enterica* subspecies *enterica* serovars from Feasey et al. [176].

function of a large number of virulence determinants [181]. Identification of these virulence determinants has been approached by various methods such as screening mutants and analysis of the genes content that contribute to the virulence traits at the molecular and cellular levels. Different virulence factor, including flagella, capsule, plasmids, adhesion systems, and type 3 secretion systems (T3SS) are most of the time encoded within *Salmonella* Pathogenicity Islands (SPI) [182, 183].

SPIs are gene clusters located in certain areas of the *Salmonella* chromosome and encodes various virulence factors (adhesion, invasion, toxin genes, etc.). These factors allow *Salmonella* to colonize its host through attaching, invading, surviving, and bypassing the host's defense mechanisms such as the gastric acidity, gastrointestinal proteases and defensins as well as aggressins of the intestinal microbiome [184, 185]. For example, *Salmonella* encodes a type III secretion system (T3SS) within SPI-1 (the SPI-1 T3SS), which is necessary for bacteria-mediated endocytosis and epithelial invasion in the intestine [186]. So far, a total of 17 [186] or 18 [187] SPIs are recognized.
Here, we will described only SPIs that have an important role in this thesis.

2.3.5.1 *Salmonella* Pathogenicity Island 1

SPI-1 is a 40kb locus that encodes virulence genes for promoting the invasion of host cells and induction of macrophage apoptosis (Figure 2.9). The T3SS is assembled from the proteins encoded by SPI-1 and is termed the needle complex. The needle complex spans the bacterial envelope, and a needle-like extension protrudes from the bacterial inner and outer membranes to the host cell membranes. Different proteins, as described in Figure 2.9 are required to encode the protein complex, such as *invA*, *SpaP*, *SpaQ*, *SpaR*, and *SpaS* that form the complex base [190], or *invG*, *PrgH* and *prgK* that form the rest of the structure [191]. Finally, *SipB*

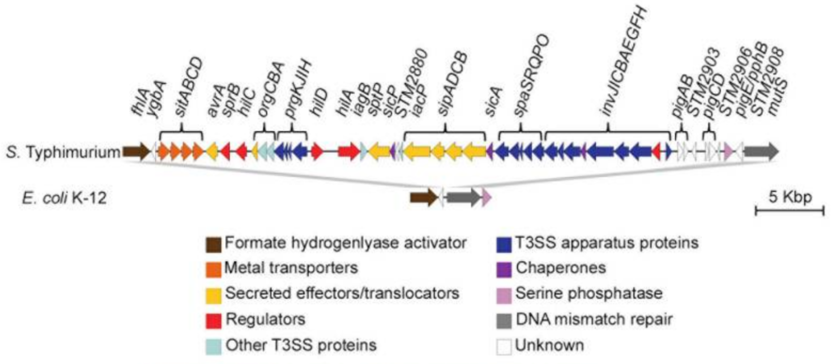


Figure 2.9: Comparative genomic analysis of *Salmonella* Pathogenicity Island 1 (SPI-1) from Lerminiaux et al. [188]. Alignment of *Salmonella enterica* serovar Typhimurium LT2 SPI-1 to the same locus in Escherichia coli K12. Genes are coloured by function based on annotations in Genbank and SalCom.

is a *Salmonella* translocon protein that is inserted into host membranes to form a channel associated with *SipD* at the needle tip, through which T3SS effectors are translocated into the host cell [192]. The schema of the interaction of *Salmonella* and the host cell using T3SS is displayed in Figure 2.10.

Among serovars, SPI-1 can have different functional genes as some genes coding effectors can be missing. For example, for *Salmonella* Typhi, a partial insertion sequence and transposase are present at the end of the locus, compared to *Salmonella* Typhimurium [187].

2.3.5.2 Others *Salmonella* Pathogenicity Islands

Here, we will describe on other SPIs identified in this thesis.
SPI-2 is a 40kb locus that encodes for a second T3SS, which is involved in intracellular survival, in both in intestinal and disseminated infection, and is required for growth within cells of different hosts [193]. The full repertoire of SPI-2 T3SS effectors is not present in all *Salmonella* serovars. However, loss of function of the SPI-2 T3SS in different serovars invariably causes a strong virulence defect, and when tested, this is usually associated with an intracellular growth defect, regardless of host cell type [194].

SPI-3 is a 17kb pathogenicity island involved in survival in macrophages and also required for growth of *Salmonella* in low-magnesium environments [195]. The distribution of SPI-3 sequences varies among the salmonellae. The virulence gene *mgtC*, which confers the survival, is conserved in all eight subspecies of *Salmonella*. However, some parts of the sequence are variable, suggesting a different incorporation process in the evolution of *Salmonella* subspecies.

SPI-4 is a 27kb region which harbors genes responsible for toxin secretion and apoptosis as well as intramacrophage survival, thanks to the contribution of *siiD*, *siiE* and *siiF* genes [196]. It has also been demonstrated that SPI-4 has a major role in influencing intestinal colonization of mammalian species [197]. But some studies still contradict each others as the exact function and the contribution of the virulence genes contained in SPI-4 had not previously been conclusively shown to be required for pathogenicity in vivo.

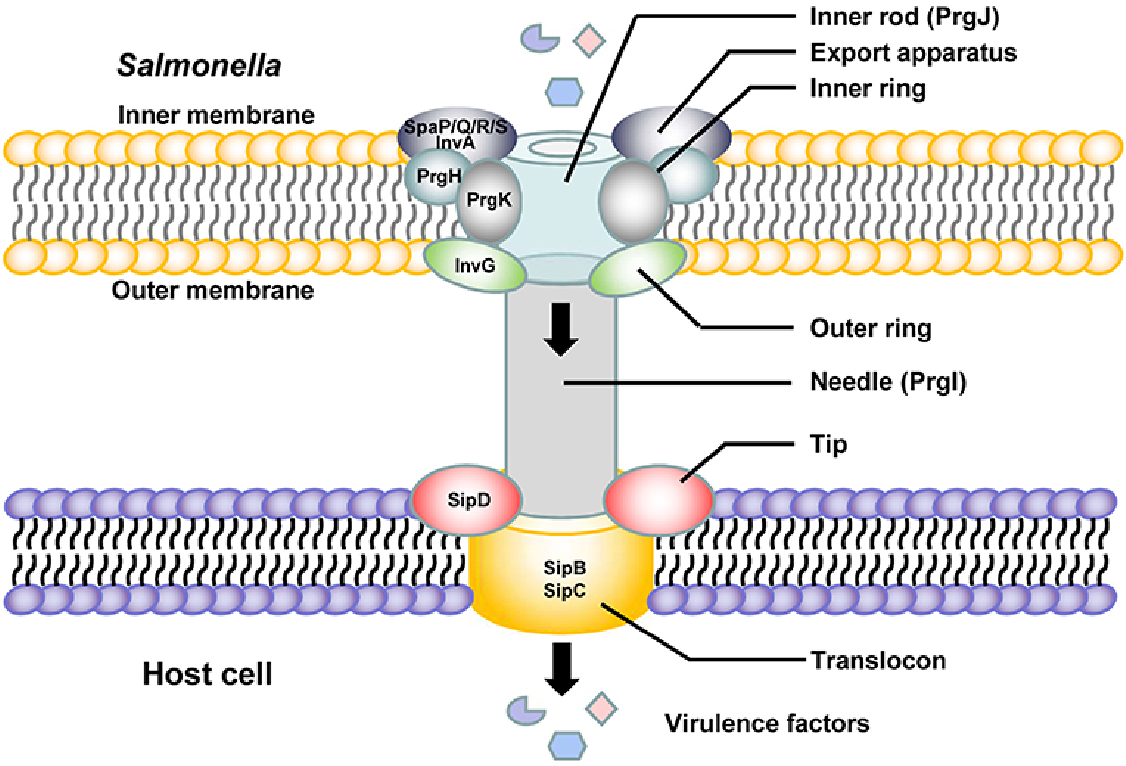


Figure 2.10: Schematic diagram of the SPI-1-related T3SS needle apparatus in contact with a host cell from Lou et al. [189].

SPI-5 is an island <8kb in size and consists of five genes (*pipA*, *pipB*, *pipC*, *sopB*, and *pipD* [198]). SPI-5 plays a vital role in enteropathogenicity and encodes effectors of SPI-1 and SPI-2.

SPI-9 is a 16kb harboring 3 ORFs that encodes a type I secretory apparatus (T1SS) and a single, large RTX-like protein. It encodes for virulence factors of type I secretion system [186, 199].

SPI-12 is a 15kb region that encompasses tRNA gene [186] and encodes a remnant phage known to contribute to bacterial virulence [200]. This SPI is well known in *Salmonella* Choleraesuis but also found in *Salmonella* Typhimurium.

SPI-13 is a 19kb locus carrying 18 functionally genes, mostly found in *Salmonella* Gallinarum and *Salmonella* Typhimurium. Theses genes contributes to the virulence of *Salmonella*, involved in the bacterial metabolism, and are likely linked to nutritional virulence of this pathogen [201].

SPI-14 corresponds to a 9kb region encoding for a transcriptional regulator and others unknown function [202]. SPI-14 was studied in *Salmonella* Enteriditis and displays a contribution to systemic infections in chickens [203]. This SPI was also identified in *Salmonella* Typhimurium isolates, with hypothesis of a contribution in the virulence of this serovar [204].

SPI-16 is a 4.5kb long island encoding for genes that are responsible for lipopolysaccharides modification and linked glucose translocases [205].

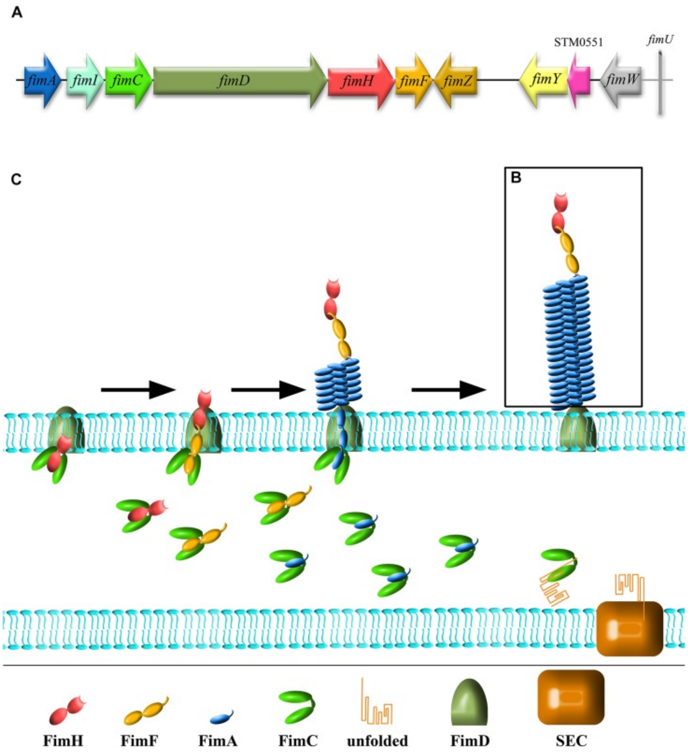


Figure 2.11: Schematic representation of *Salmonella* type 1 fimbriae (T1F) from Kolenda et al. [206]. (A) fim gene cluster organization, (B) structure of fimbrium, and (C) biogenesis by the chaperon-usher pathway.

2.3.5.3 Fimbriae

Fimbriae or pili play an important role in the pathogenesis of *Salmonella*, as they represent the most common adhesion system to host cells, a crucial process for the pathogenicity of the bacterium. Fimbriae represents proteinous structures found on the surface of the bacteria that can mediate interaction with cells. They mediate adhesion of *Salmonella* to different surfaces (hosts' cells, food, stainless steel, etc.) and have been implicated in a variety of other roles such as biofilm formation, cellular invasion [206], and macrophage interactions [207].

Duguid et al. classified fimbriae in seven types (types 1–6 and F) according to the morphology and haemagglutination patterns [208]. However, another classification, based on serology, better predicted genetic relatedness of fimbrial antigens. Fimbriae can also be classified based on the mechanism by which these appendages are assembled on the bacterial surface. This classification has gained popularity because members of an assembly class can be readily identified by the sequence homology of their fimbrial biosynthesis genes [209].

Among dozens of different bacterial fimbriae, type 1 fimbriae (T1F) are one of the most common adhesive organelles in the members of the Enterobacteriaceae family, including *Salmonella*, and are important virulence factors [206]. Figure 2.11 displays the structure of fimbriae proteins *fimAICDHF* and three independently transcribed regulatory genes *fimZYW*. It has been showed that there is a high degree of allelic variation in adhesins that promote host colonization *fimH* in different serovar, likely to contribute to pathoadaption to diverse hosts of *Salmonella* serovar [210]. This adaptation is important to mention because most studies of host adaptation and pathogenicity only rely on the analysis of specific virulence genes, not undertaking the function and the consequence on the protein structure of the fimbriae. Also, the production of the fimbriae proteins is important for the pathogenicity of the bacteria, because it was shown that deletion of any one of *fimA*, *fimF* or *fimH* results in an absence of fimbriae production [211].

2.3.5.4 Others important virulence factors

Salmonella virulence factors can also be located on extra-chromosomal genetic elements or in segments inserted within the chromosome that originate from other genomes. The acquisition of a new gene may occur by genetic transformation, but when virulence genes are located on plasmids (self-replicating double-stranded circles of DNA) they can be mobilised by conjugative transfer [212]. Often, the genes carried in plasmids provide bacteria with genetic advantages, such as antibiotic resistance.

Some *Salmonella* serovars possess specific virulence plasmids. In these plasmids, there is a *Salmonella* plasmid virulence (*spv*) locus, whose expression in *Salmonella* organisms has been reported to be important for intramacrophage survival and multiplication of *Salmonella* within the reticuloendothelial system including liver cells and the spleen [213, 214, 215, 187]. It has also been described that some virulence plasmids confer advantage to certain host. *Salmonella* Kentucky isolates have been identified containing an IncF virulence plasmid acquired from an avian pathogenic *Escherichia coli* strain that may confer an advantage in an avian host [216].

2.3.5.5 Antimicrobial resistance

The modern era of antibiotics started with the discovery of penicillin by Sir Alexander Fleming in 1928 [217] and since then, they have been used for combating pathogenic bacterial agents in both human and animal. Unfortunately, the extensive use of the antimicrobial agents has led to the evolutionary emergence of resistance to one or more of the antibiotics used against pathogenic bacteria. Shortly after the prescription to serious infections by using penicillin, penicillin resistance bacteria emerged, became a substantial clinical problem, so that, by the 1950s, many of the advances of the prior decade were threatened [218]. Since then, resistance has been seen to nearly all antibiotics that have been developed (Figure 2.12).

The Centers for Disease Control and Prevention (CDC) predicted that 16% of non-typhoidal infections in the United States between 2015 and 2018 were resistant to one antibiotic or more [219]. The World Health Organization (WHO) published a survey on antibiotic-resistant bacteria that present a serious threat to public health [220]. *Salmonella spp* is reported on this document as a priority pathogen because of the emergence of Fluoroquinolone resistant *Salmonella*. Fluoroquinolone resistance in *Salmonella* seriously compromises treatment options, especially for invasive salmonellosis, and different mecanisms have evolved to increase the survival of the bacteria [221, 222]. In *Salmonella* Mbandaka, a total of five quinolone resistance genes (*qnrB1*, *qnrB19*, *qnrB6*, *qnrB9*, and *qnrS1*) have been already identified in the genome, but without phenotype confirmation [103]. In *Salmonella* Typhimurium, *qnrB5* and *qnrS1* has been described as plasmid-mediated quinolone resistance [223]. Finally, in monophasic variant of *Salmonella* Typhimurium, *oqxAB*, *qnrB* and *qnrS* have been identified to drive quinolone resistance [224].

One of the most important factors influencing the spread of AMR is horizontal gene transfer. *Salmonella* is able to resort to different mechanisms (DNA incorporation, integration of plasmid content) which allows it to acquire resistance and effective virulence for better survival in its environment. Mobile genetic elements, including plasmids, phage, and transposons, can facilitate HGT via conjugation, transduction, and transformation, respectively [225, 226].

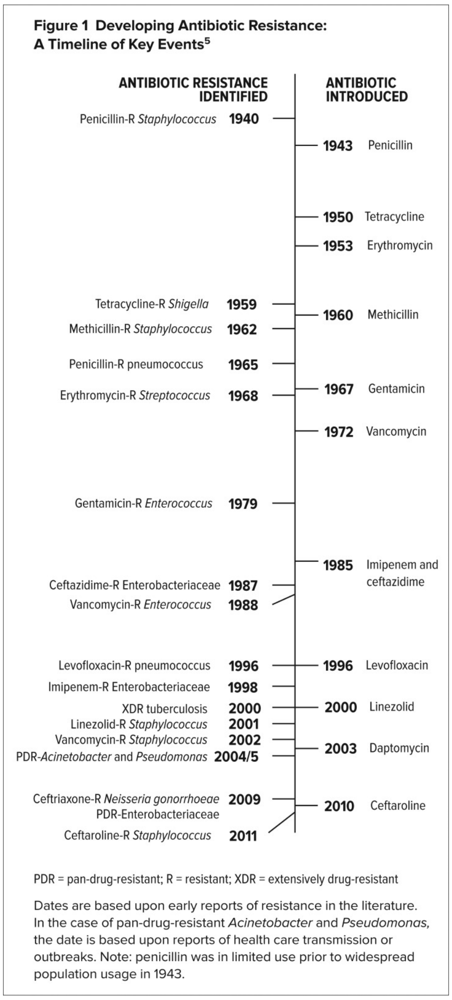


Figure 1 Developing Antibiotic Resistance: A Timeline of Key Events[5]

ANTIBIOTIC RESISTANCE IDENTIFIED
ANTIBIOTIC INTRODUCED
Penicillin-R Staphylococcus 1940
1943 Penicillin
1950 Tetracycline
1953 Erythromycin
Tetracycline-R Shigella 1959
1960 Methicillin
Methicillin-R Staphylococcus 1962
Penicillin-R pneumococcus 1965
Erythromycin-R Streptococcus 1968
1967 Gentamicin
1972 Vancomycin
Gentamicin-R Enterococcus 1979
1985 Imipenem and ceftazidime
Ceftazidime-R Enterobacteriaceae 1987
Vancomycin-R Enterococcus 1988
Levofloxacin-R pneumococcus 1996
1996 Levofloxacin
Imipenem-R Enterobacteriaceae 1998
XDR tuberculosis 2000
2000 Linezolid
Linezolid-R Staphylococcus 2001
Vancomycin-R Staphylococcus 2002
2003 Daptomycin
PDR-Acinetobacter and Pseudomonas 2004/5
Ceftriaxone-R Neisseria gonorrhoeae 2009
PDR-Enterobacteriaceae
2010 Ceftaroline
Ceftaroline-R Staphylococcus 2011

PDR = pan-drug-resistant; R = resistant; XDR = extensively drug-resistant

Dates are based upon early reports of resistance in the literature. In the case of pan-drug-resistant *Acinetobacter* and *Pseudomonas*, the date is based upon reports of health care transmission or outbreaks. Note: penicillin was in limited use prior to widespread population usage in 1943.

Figure 2.12: Timeline of Key Events of developing Antibiotic and Antibiotic resistance reports from Ventola CL [217]. Dates are based upon early reports of resistance in the literature.

Plasmids, self-replicating mobile DNA elements transferable from one bacterium to another, are an important vector for these genes of resistance [226]. Resistance genes on plasmids are usually located on mobile DNA elements (transposons) that can recombine in plasmids or in the bacterial chromosome via insertion sequences that allow them to integrate into host DNA. Conjugative plasmids are self-transmissible, giving them the potential to increase the spread of antimicrobial resistance (AMR) genes. A conjugative plasmid is composed of *oriT* genes which prepare the transfer of the DNA content, MOB genes which process the DNA being replicated and transferred to the new cell, and finally mate-pair formation (MPF) genes that form the channel between the two cells where the DNA can travel. Figure 2.13 describe the process of the conjugation. Mobilizable plasmids carry the genes *oriT* and MOB genes for plasmid transfer but generally lack the MPF genes needed for pilus formation. Mobilizable plasmids can use the help of other coresident plasmids in the same cell to complete the conjugation process [227]. Some plasmids are called non-mobilizable because they are neither conjugative nor mobilizable. They spread by natural transformation or by transduction. Hence, plasmids can be classified into three categories according to mobility: conjugative, mobilizable, and non-mobilizable. Studies found that there are 15% conjugative, 24% mobilizable, and 61% non transmissible plasmids in prokaryotes. [227].

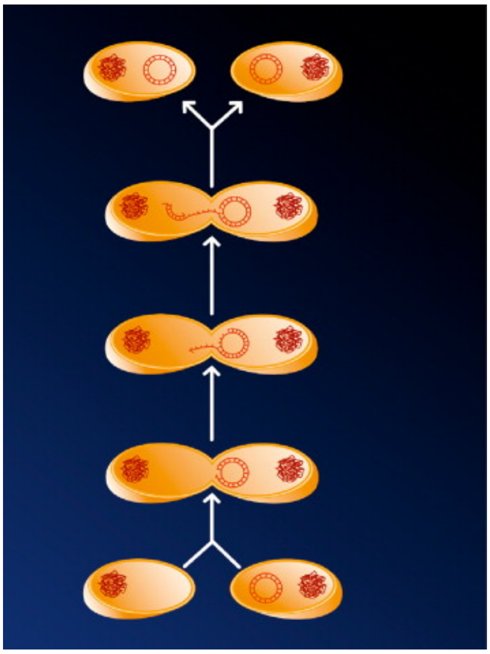
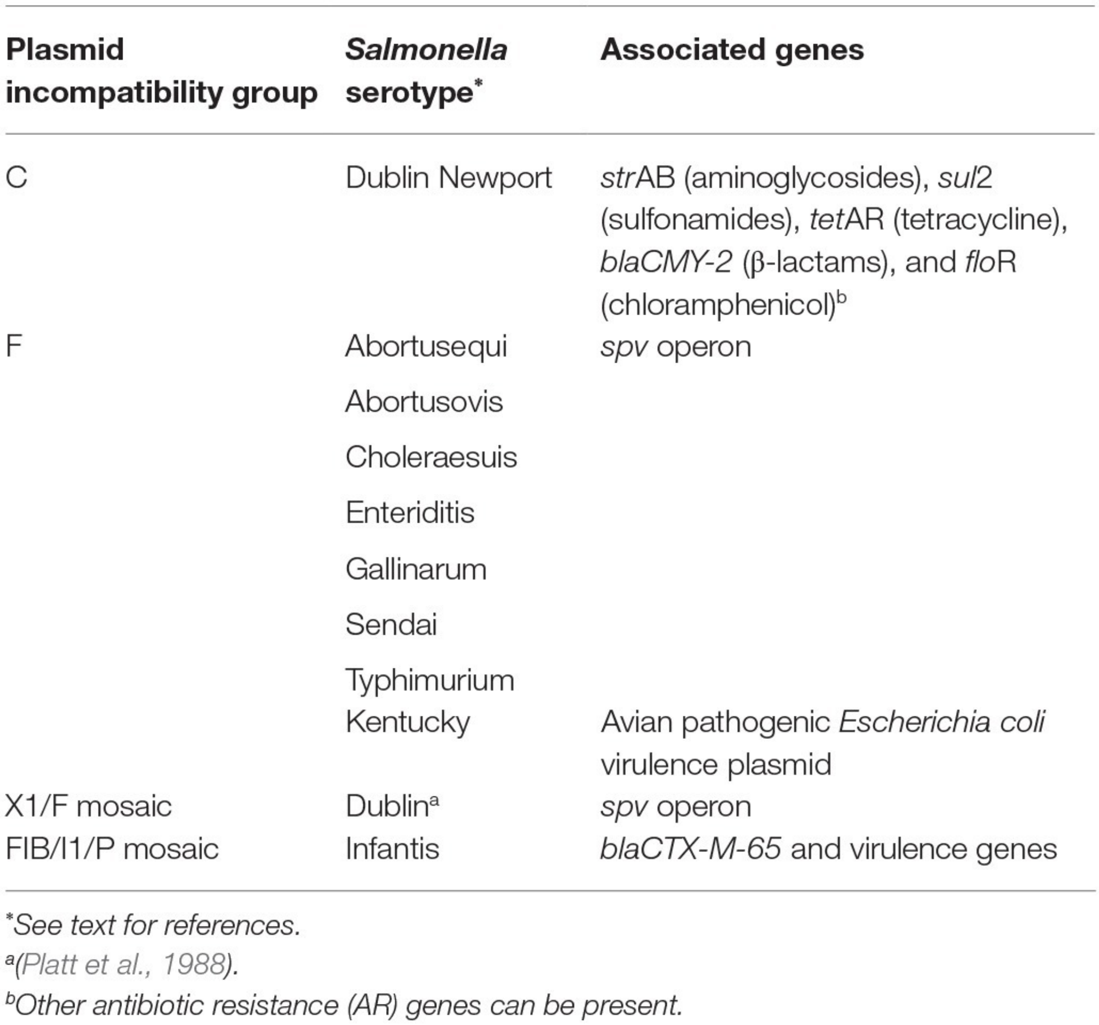


Figure 2.13: Schematic representation of plasmid conjugation from Périchon et al. [228]. Bottom right, donor bacterium; bottom left, recipient bacterium. The chromosomes are represented in a condensed state. Plasmid is represented as a ring. After a single nick on one of the two complementary DNA strands of the plasmid, one strand is transferred from the donor to the recipient. During this process, the complementary strand of the remaining DNA strand in the donor is synthesized while the complementary strand of the incoming DNA is synthesized in the recipient. After transfer, each bacterium contains a copy of the plasmid (top) and can therefore act, in turn, as a donor.

Plasmids are typed based on incompatibility with other plasmids, which are defined as the inability of two plasmids to be maintained together in the same cell line [229]. Twenty-eight incompatibility groups have been isolated in *Enterobacteriaceae* and differ by the AMR genes they can carry. Despite the identification of the plasmids in an *Enterobacteriaceae* and a *Salmonella*, the plasmid can evolve and gain or lost functions, making not possible the prediction of AMR of a *Salmonella* based on the presence of a plasmid [230]. Thus, even though *Salmonella* contain plasmids that can be found in many organisms, serotypes of *Salmonella* differ in the frequencies of plasmids they contain, and new plasmids can emerge within any serotype. While some serotypes are more likely to contain certain plasmids than others, these plasmids can exist in any serotype, and no incompatibility group is confined to a single serotype. Presented examples of serotypes that can contain certain plasmids more frequently than others (Figure 2.14).

Typing plasmids is important to understand the AMR circulating in strains. The cmy-2 gene was detected in a plasmid distributed in *Salmonella* Typhimurium strains [231], which confers resistance to ceftriaxone, a highly used drug treatment of children with invasive *Salmonella* infections. IncHI2 plasmids have been identified in *Salmonella* Mbandaka strains [103], mediating for quinolone resistance in Mbandaka, but also β-lactams resistance from food and clinical strains of *Salmonella* Typhimurium [232].

Finally, other vectors as phages and Genomic Island can mediate the AMR and the multiple



Plasmid incompatibility group	*Salmonella* serotype*	Associated genes
C	Dublin Newport	*str*AB (aminoglycosides), *sul*2 (sulfonamides), *tet*AR (tetracycline), *blaCMY-2* (β-lactams), and *flo*R (chloramphenicol)^{b}
F	Abortusequi Abortusovis Choleraesuis Enteriditis Gallinarum Sendai Typhimurium	*spv* operon
	Kentucky	Avian pathogenic *Escherichia coli* virulence plasmid
X1/F mosaic	Dublin^{a}	*spv* operon
FIB/I1/P mosaic	Infantis	*blaCTX-M-65* and virulence genes

**See text for references.*

^{a}(Platt et al., 1988).

^{b}Other antibiotic resistance (AR) genes can be present.

Figure 2.14: Examples of plasmids groups associated with some *Salmonella* serotypes from McMillan et al. [226]. Virulence or antimicrobial resistance genes added by the plasmid is displayed.

AMR of *Salmonella*. While *Salmonella* Pathogenicity Islands (SPI) displays high set of virulence genes, *Salmonella* Genomic Islands (SGI) tend to carry antibiotic resistance genes. SGI-1 (*Salmonella* Genomic Island 1) is a chromosomal locus identified in *Salmonella* Typhimurium and widely distributed in other serovars. SGI-1 carries different resistance genes, with variable profiles according to the serovar [233]. Resistance to ampicillin, chloramphenicol, florfenicol, streptomycin, spectinomycin, sulfonamides, and tetracycline is a well-studied combination of resistance genes found in SGI-1 [234, 233].

2.3.5.6 Biocides resistance

Salmonella is in constant mutation leading to more persistent and more resistant strains to the different sanitation products used in industries [44], threatening general hygiene and safety. The determinants driving persistence, transmission and their evolutionary frame are not fully known. Biocide type used depend on the *Salmonella* serovar identified [235], as some cleaning product are no longer efficient due to the emergence of multi-resistant *Salmonella*.

The main targets of biocides action are the lipoproteins and phospholipids present in the outer membrane of bacteria. To ensure its survival, *Salmonella* uses strategies, such as changing the type of its outer membrane to resist biocides [236, 237]. Other genes determinants, like *acrB* has displayed a biocide resistance phenotype for *Salmonella* Typhimurium [238]. Finally, SPI can carry resistance genes like *PipB* that is associated with membranes and play a critical role in biocide resistance [239].

In addition, to tackle *Salmonella* dissemination, biocides play an essential role in limiting the spread of infectious disease. The food industry is dependent on these agents, and their increasing use is a matter for concern. *Salmonella* with increased tolerance to biocides emerged, endangering the food industry. Some resistance to biocides (benzalkonium chloride, chlorhexidine, triclosan) has been identified in *Salmonella* serovars Enteritidis and Typhimurium [240]. Furthermore, over-exposures of biocides contribute to the development of microbial resistance mechanisms, highlighting that inappropriate use of biocides in situations where they is unnecessary are dangerous. Other studies showed a strong resistance of some *Salmonella* strains. *Salmonella* isolated from 39 hen eggshell surfaces displayed tolerance to benzalkonium chloride (7.7%), cetrimide (7.7%), hexadecylpyridinium chloride (10.3%), bisphenols triclosan (17.9%), hexachlorophene (30.8%) and finally P3-oxonia (25.6%) [241]. *Salmonella* strains with higher tolerance would be expected to also have greater possibilities for survival to disinfection processes. Also, the study pointed out that diluting biocides below their effective concentrations would play a role in survival and then in the acquisition of resistance genes [242], and also in cases of biofilms, biocide-tolerant strains could play an important role of *Salmonella* persistance in the environment.

While partial explanatory factors were characterized within the genome and transcriptional regulation realms [243], it is clear that the complete elucidation of biocides resistance is far beyond the boundaries of the routine analyses. To add to the pressure resulting from sanitation, co-evolution with the host drives the pathoadaptation and the clinical manifestations, impacting for example the transience of *Salmonella* presence in milk. Being most essentially a gastric pathogen, *Salmonella* coexists with the gut microbiome which restrains its possibilities in term of adaptation. All these constraints forced all sides – the pathogen, the host and the flora – to adapt their defense mechanisms, leading to a rapidly obsolete knowledge of systems and arguing for continuous research, survey and updates.

2.3.6 Monitoring *Salmonella* using genomics

WGS can provide more insight in outbreaks investigations [244], thus some public health agencies, especially in developed countries, have developed WGS methods to overcome the lack of precision of usual *Salmonella* typing methods [26, 245, 246]. Conventional typing results do not always have sufficient granularity or robustness to define strains unequivocally, and sufficient epidemiological data are not always available to establish links between patients and the environment. The most advantages reported by public health agencies using WGS is the database of genomes used for the monitoring of the overall number of *Salmonella* isolated at frontline laboratories and the number of isolates referred to the laboratories [247]. Different kind of data can be saved (ST, SNPs, virulence, AMR) and compared to provide more insights of a new outbreaks. Monitoring *Salmonella* using WGS was able to make possible the detection of new epidemic cases and new linked samples on an epidemic case [126].

WGS have been demonstrated to provide improved and exhaustive data related to pathogen genotypic characteristics and can allow the identification of virulence determinants, AMR genes and serotypes for *Salmonella*, allowing a real-time monitoring of emerging resistances [245]. It can also evaluate the evolution of strains during an outbreak [248].

Studies displayed that both SNP calling and cg/wg MLST accurately distinguishes outbreak-related isolates from non-outbreak isolates [249]. The identification of genetically closely related isolates using cgSNPs or wgMLST are highly concordant with epidemiological data and provides more resolution than cgMLST. But it was also reported that cgMLST was sufficient for bacterial surveillance [128], especially due to clustering through minimum spanning tree (MST) [63].

At the Public Health of England, the SNP detected in genomes is now utilized by epidemiologists and microbiologists as the primary method for identifying microbiological clusters of gastrointestinal infections and detecting potential outbreak events. Case isolates that fall within a 5-SNP single linkage cluster are considered likely to be exposed to a common source of contamination. The 5-SNP threshold is variable depending on the serovar, but it was estimated to be a stringent and sufficient threshold for common serovar isolated in England [250, 251]. In Australia, the SNP cutoff is $<$ 8 SNPs for *Salmonella* Typhimurium, also defined from observed SNPs in known outbreaks. Using this threshold, WGS data combined with epidemiological data link an additional 9% of isolates to at least one other outbreak isolate (compared to MLVA), and 19% more isolates compared to epidemiological data alone [126]. Finally, some approaches based on dynamically increased SNP cut-offs are used to generate outbreak investigation clusters, but these methods are still under study [252].

In France, WGS is not systematically implemented as the main typing tool for *Salmonella* in foodborne outbreak investigation. CNR Pasteur, described in section 2.1.3, has implemented WGS as its primary typing tool [44], but it is not routinely implemented at ANSES. It is therefore difficult to trace back outbreaks. Combined with epidemiological data, investigators can track back the dissemination of strains at the regional scale and point-out exchanges of strains between places and the origin of their contamination.

2.3.7 *Salmonella* in the pig and pork industry

2.3.7.1 Serovar prevalence in pig and pork

Salmonella Choleraesuis was the first *Salmonella* serotype isolated from pigs [253], only 2 years after the first isolation ever of *Salmonella*, performed by Gaffky in 1884 [254]. During 1950s and 1960s, *Salmonella* Choleraesuis, including variant Kunzendorf, was the predominant serotype isolated from pigs worldwide [255]. From all the *Salmonella* serotypes, the more important ones causing clinical disease in pigs are *Salmonella* Choleraesuis and *Salmonella* Typhimurium. In recent years, *S.* Typhimurium and its monophasic variant have replaced *S.* Choleraesuis as the predominant serovar (described later) in many countries [256, 33, 257].

Pork is the most significant source of meat responsible for human transmission of *Salmonella* [259]. It is well known that pork is a strong vector of *Salmonella* [260, 261]. Pigs can be either asymptotic or display strong inflammatory response leading to salmonellosis and sometimes death. In French human food sector, 37 different serovars were isolated in 2019 from pork food, monophasic variant of Typhimurium serovar being the most prevalent followed by serovar Derby and Typhimurium, which together account for 56.5% of serovars detected [262].

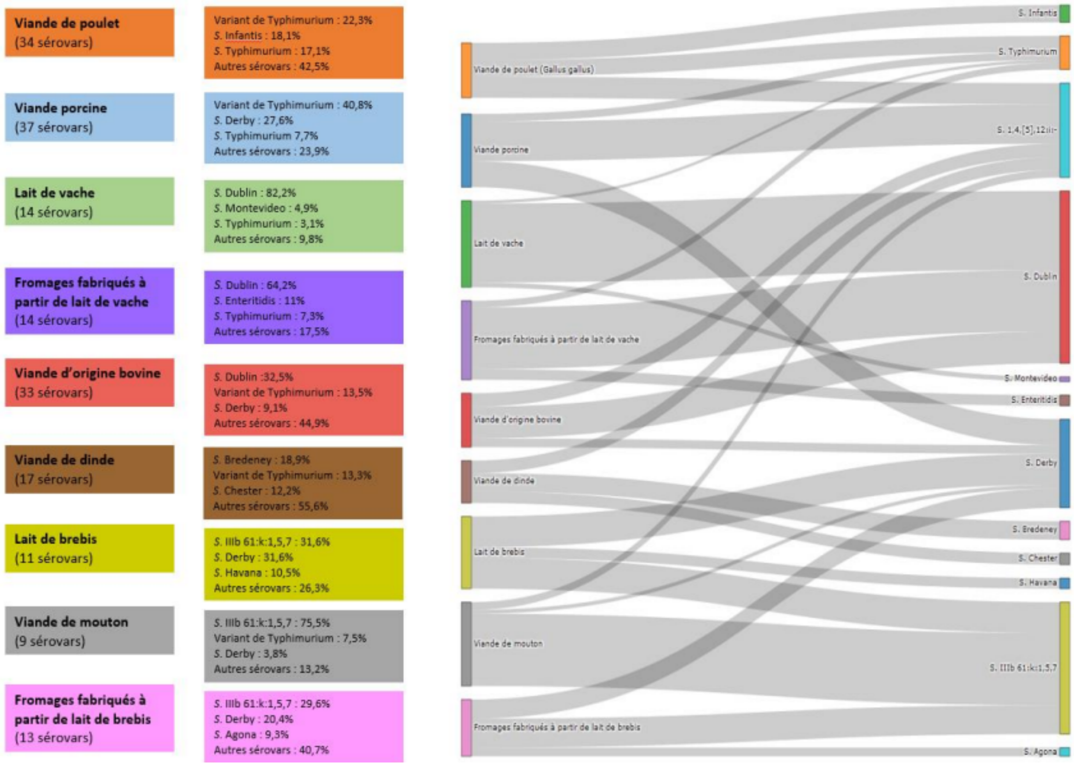


Figure 2.15: Distribution of the three main serovars for the main sectors of the human food sector from Leclerc et al. [258].

Together with *S.* Enteritidis, *S.* Typhimurium and its monophasic variant, these three serovars represented 77.6% of the confirmed human cases acquired in the EU in 2020 [37]. S.Typhimurium was mainly linked with broiler and pig sources. The monophasic variant of Typhimurium was related mainly to pig and secondly to broiler sources. *Salmonella* Enteritidis is primarily related to broiler source, but rarely found in pork [37].

In the pig and pork industry, non-typhoidal *Salmonella* infections are mostly related to serovars *S.* Typhimurium and its monophasic variant [259], representing 17.9% and 43.7% of all *Salmonella* detected in pork meat in 2015 [43]. They are 2 of the 3 main serovars detected in the human food sector (Figure 2.15), and have caused outbreaks especially in dry sausage in 2010 [263].

The monophasic variant of Typhimurium (TMV) first appeared in Europe in the mid-1990s and the threat dramatically increased between 2005 and 2008, where it became one of the top three isolated serovars in human health and has kept its place today [44, 44]. This serovar is characterized by a high prevalence of resistance to ampicillin, streptomycin, sulfonamides, and tetracyclin [264, 265], representing a public health issue.

2.3.7.2 Clinical and prevalence of *Salmonella* in pigs

Prevalence of *Salmonella* is hard to estimate at the pig farming level due to the asymptomatic carriage of pigs. But, according to studies, the prevalence of *Salmonella* in pigs and pork products can vary between 25% [266] to 50% [267] in some country, with different prevalence according to the health protocols of certain farms. Reporting salmonellosis cases in pig is mandatory, but asymptomatic cases do not allow a complete follow-up of the number of cases. In slaughterhouse and processing industries, controls are mandatory on raw materials and finished products, and allow a good vision on the distribution of serovars in this part of the food chain. Between 2009 and 2015, *Salmonella* Derby was mainly found in slaughterhouse carcasses (28% of isolated serovars), while *S.* Typhimurium was more prevalent in food processing and pork products (31% of isolated serovars) [268]). In 2015, the mean prevalence of *Salmonella* in pig carcass was between 1.8% and 9.5% depending on the slaughterhouse category [256]. *Salmonella* might occur at any age, but is more frequent in growing pigs of 8 weeks or older. Infection with *Salmonella* Typhimurium has been associated with a strong inflammatory response and activation of immune mechanism in pig [269]. Although infections in pigs by ubiquitous *Salmonella* serovars could result in enteric and even fatal disease, infected animals frequently and asymptomatically carry *Salmonella* in the tonsils, gut and gut-associated lymphoid tissue without seeing clinical effect. Without the implementation of specific measures during slaughtering, such as carcasses bagging or specific singeing of fecal stains, a risk of cross contamination of the carcasses from the faeces during evisceration or removal of the tonsils exists resulting in the contamination of the meat, thus being a potential risk for human health [270, 259].

Pigs can be subclinical carriers of *Salmonella* for long periods of time because the organism survives in the mesenteric lymph nodes located in the intestine. In addition, the bacteria survives well in feces and infecting pigs can shed after 24h of the infection, increasing the time of exposure to the disease [271]. It was also demonstrated that a few pigs shedding low levels of *Salmonella* organisms before slaughter can result in rapid transmission and subsequent shedding by many swine [271]. Many of these carriers do not excrete the bacteria in faeces, unless they are under stressing conditions like transport to the slaughterhouse. Transmission of *Salmonella* between pigs occurs mainly via the faecal–oral route [261] although some studies have demonstrated that the upper respiratory tract and lungs could be portals of entry as well [272].

In addition, contamination of pig carcasses can be linked to cross-contamination from other carcasses and the presence of *Salmonella* in the environment [259]. The disease depends on the strain and the dosage. To limit humans cases due to contaminated pigs, various treatments, such as probiotics, prebiotics [273], and vaccination have been developed [274, 275].

2.3.7.3 *Salmonella* Typhimurium, the most prevalent serovar in pig industry...

Salmonella Typhimurium represents 17.9 % of *Salmonella* serovar isolated in pork industry in 2015 [43], and 7.7% in 2019 [262]. While this serovar is multi-host, it is well implemented in this industry, causing serious health hazards. Irish data from the EU baseline survey (2016) on the prevalence of *Salmonella* in slaughter pigs showed that 57% of *Salmonella* isolated was *Salmonella* Typhimurium [267]. In UK, 43.9% of *Salmonella* isolated in pig in 2020 were *Salmonella* Typhimurium [276]. Although this serovar is also common in Africa [176, 277] and Asia [278], it is impossible to compare these data since they are muddled with TMV isolates (described in 2.3.7.4).

Studies have been conducted to understand the prevalence of this serovar in some hosts, and also their evolution pathways that emerged in different ST. For example, a study suggested that passerine-adapted *S.* Typhimurium from USA and UK countries have emerged in recent decades, shared a common ancestor that may be spread by gulls and terns, and formed lineages distinct from the major *S.* Typhimurium lineages originating from humans and domestic animals [279].

Different sequence types have been described for this serovar. Some ST are more prevalent in some continents or countries. For example, in Africa, ST313 is the major ST, followed by ST19 and ST394 [280]. In Brazil, ST313 and ST19 were identified, while ST313 is rare outside of sub-Saharan Africa, displaying the ability of the *Salmonella* to adapt in different environment. In Europe, ST19 is most commonly associated with *Salmonella* Typhimurium in pork [79, 281]. Others ST (ST328, ST34) have also been isolated from this host.

2.3.7.4 ...becoming less prevalent than its monophasic variant

Salmonella serovars can produce two types of flagellar proteins and switch from one type to the other by the expression regulation of *fliC*, *fljBA*, and *hin* genes [282]. *Salmonella* Typhimurium is bi-phasic, meaning it can express the two type of flagellar porteins : phase 1 flagellin (*FliC*)and phase 2 flagellin (*FljB*). However, since the late 1990s, a new variant, lacking expression of the phage 2 flagella has increased among human cases of salmonellosis. It was detected in Spain (10 strains from 1993 to 1996), but started to spread as the "Spanish clone," which emerged rapidly since 1997 [283].

This lack of mutation can be explained by a various deletion of *FljBA*, or other mutations and deletions of *fljA* or *hin*, or even deletions directly on promoters that control the expression of *fljB* and *fliC* [284, 285].

Nowadays, it became the most prevalent serovar in different countries such as France [37, 44]. It represents 43.7% of *Salmonella* serovars detected in pork in 2015 [43] and 25% of all human cases of non-Typhoidal Salmonellosis. It has taken a prominent place among *Salmonella* isolates in France, and has been increasing steadily from 2008, mainly due to the international spread of the multi-resistant antibiotic clones [286]. In 2020 in UK, for the first time since 2014, TMV was less present than *Salmonella* Typhimurium in pigs (34.4% TMV, 43.9% *Salmonella* Typhimurium), but this event is rare, as TMV is the 1st isolated serovar in previous years, ahead of *Salmonella* Typhimurium by between 3% to 27%.

While there is no clear explanation about its prevalence, TMV showed a large panel of heavy metal gene resistance like copper, silver or mecury [287]. This panel of resistance can be explained with the description of genomics island encoding for these genes [180]. Several studies targeted its genomic and phenotypic singularities, for instance several insertion [180, 288] or deletions [289] in intergenic regions specific to *Salmonella* Typhimurium were found. None of them however explained which determinant makes this variant persistent neither it is known how it spread around the world.

With an evolutionary point of view, TMV clones are spread word-wildly. Studies demonstrated that UK isolates associated with many animal species and human clinical infections in the United Kingdom arose recently, clustering with previously described North America and Spain isolates [180].

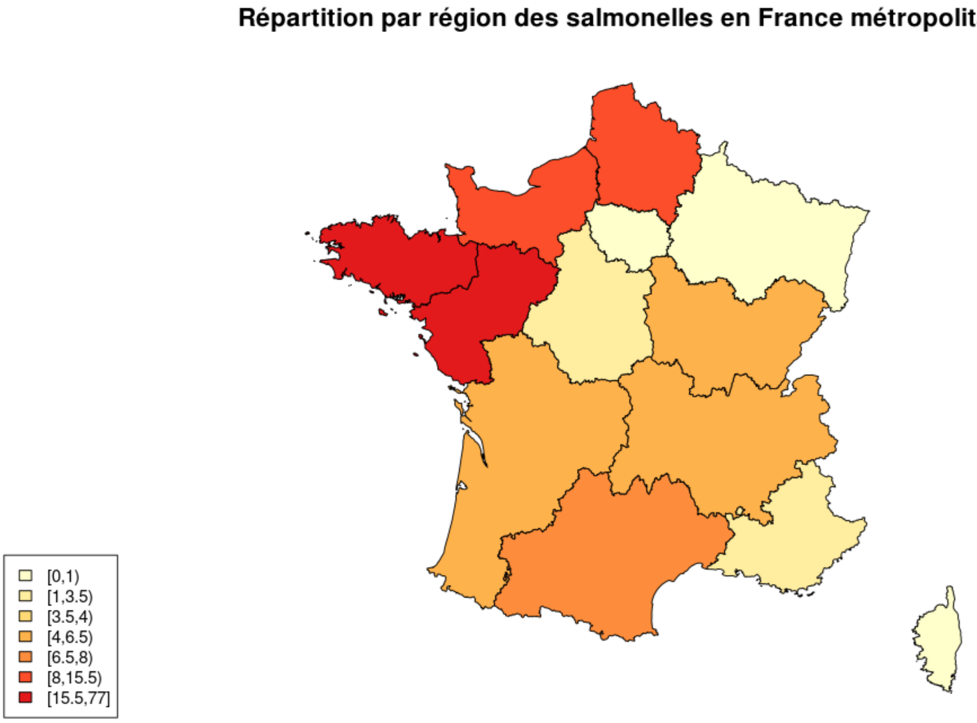

Répartition par région des salmonelles en France métropolitaine
[0,1)
[1,3.5)
[3.5,4)
[4,6.5)
[6.5,8)
[8,15.5)
[15.5,77]

Figure 2.16: Geographical dissemination of monophasic variant of *Salmonella* in France. Map inferred by *Salmonella* Network data [43] with personal data isolated from pig sector.

2.3.7.5 Notes about specific genomics investigation of *Salmonella* Typhimurium and its monophasic variant

Salmonella Typhimurium and its monophasic variant have been well described due to their prevalence in humans and animal reservoirs. AMR has been studied, *in silico* and *in vitro* [228, 290]. In USA, the most frequently observed antibiotic resistance patterns found in *S.* Typhimurium were tetra-resistant pattern ASSuT (ampicillin, streptomycin, sulfonamides, and tetracycline) and the penta-resistant pattern ACSSuT (ampicillin, chloramphenicol, streptomycin, sulfonamides, and tetracycline [291]). In Asia [224], tetracycline, ampicillin, sulfisoxazole, and streptomycin resistance were found in more than 75% of the isolates. Other resistances were less prevalent for ciprofloxacin, cefotaxime, ceftriaxone, cefepime, ceftazidime, and colistin, but of great concern in terms of their current clinical importance [220]. Studies at the pig herd level revealed that most clonal groups of *Salmonella* Typhimurium and TMV were highly drug resistant due to the presence of multiple AMR genes [292], with evidence of recent on-farm plasmid-mediated acquisition of additional AMR genes (pSTM and IncHI2 [293, 232]).

Also, phage content was strongly analysed as it provides rapid, accurate, and a cheap method of investigating *Salmonella* Typhimurium and TMV strains for epidemiological reasons. Phage type U288 is particularly associated with pigs and is the most commonly isolated phage type of *S.* Typhimurium in 2020 in UK, followed by DT193 [276]. DT204, phage resistant against sulfonamides and tetracycline were one of the most frequently isolated phage from pigs in *Salmonella* Typhimurium, but recently, TMV clones DT104 [294], virulent and drug-resistant isolates, are spreading all around the world [295].

Finally, to analyze *Salmonella* genome variants, most of the time a pipeline based on reference-mapping is implemented. To increase the quality and accuracy of the downstream analysis, the choice of a good quality and a well representative of the dataset reference genome is crucial [296]. In *Salmonella* Typhimurium, the most common used reference is *Salmonella* Typhimurium LT2 [297]. *Salmonella* Typhimurium LT2 is one of the most studied lineage since the 1950s, with a major focus on its genetic and biochemical content [298].

2.3.8 *Salmonella* in cattle

2.3.8.1 Serovar prevalence in cattle

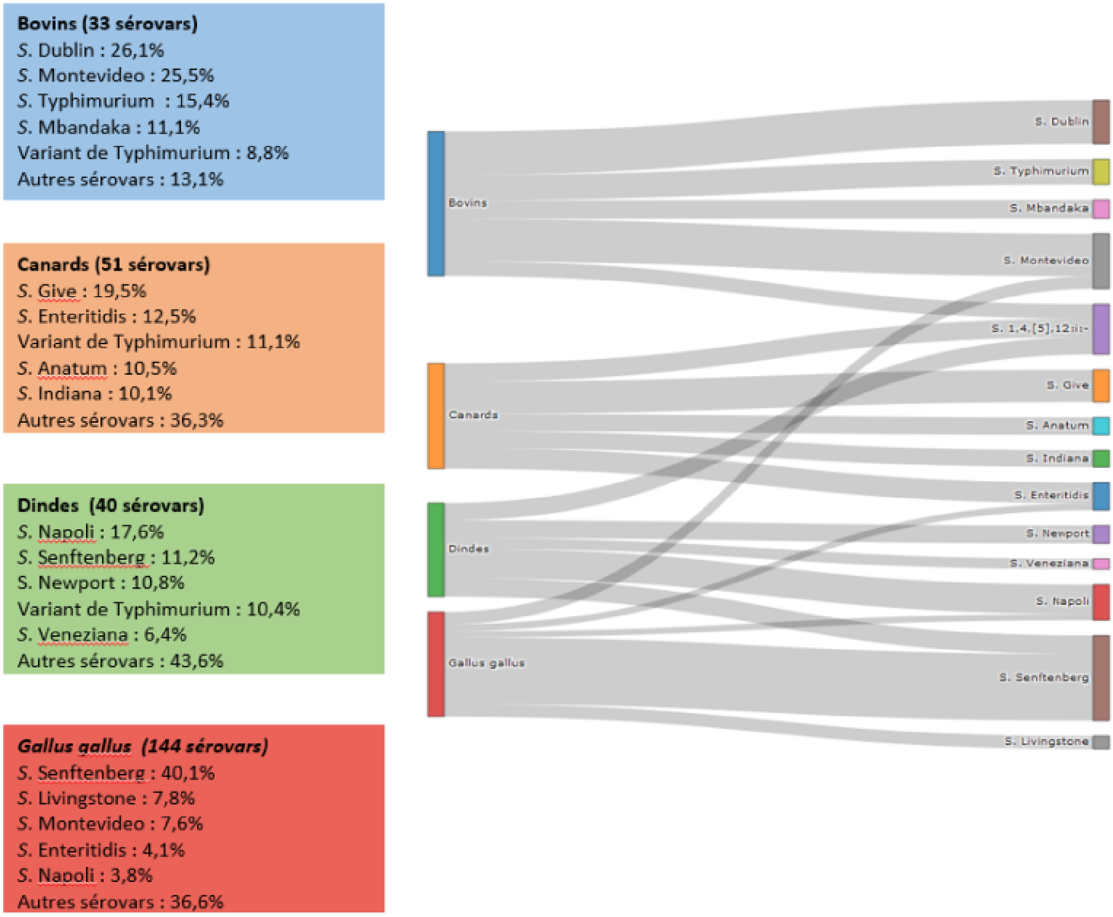
Overall, the 10 most frequently reported serotypes between 2000 and 2017 in cattle are Montevideo, Typhimurium, Kentucky, Meleagridis, Anatum, Cerro, Mbandaka, Muenster, Newport, and Senftenberg [299]. According to this study, these serovar accounts for 65% of the isolates. *Salmonella* Montevideo and *S.* Dublin are the most frequently reported serotypes in North America and Europe, respectively, while *S.* Typhimurium is the most frequent in Africa and Asia. In UK, *Salmonella* Dublin remained the most commonly reported serovar accounting for 57.7% of total cattle isolations. It was also observed that the prevalences of *Salmonella* serovar have little variations, and the most prevalent serovars are persistent. The second most common serovar in 2020 in UK was *S.* Mbandaka (16.2% of total cattle isolations) [276].

In France in 2019, in animal health and production, the sectors mostly affected by *Salmonella* were the poultry and bovine sectors, and the most frequently isolated serovars from bovine sectors were Dublin, Montevideo, Typhimurium and Mbandaka (Figure 2.17), displaying a concordance with data from others European country [258, 262, 300].
Focusing on milk industry, the two most detected *Salmonella* serovars are *S.* Montevideo (26.3%), *S.* Mbandaka (21.1%) and *S.* Dublin (17.3%) in France [43]. Several outbreaks of food poisoning occurred in the dairy industry associated to different raw milk cheeses [301, 302, 303] for *S.* Montevideo and *S.* Dublin, demonstrating that the identified strains are able to be transferred all along the chain to infect humans [258]. While others serovars seem to be adapted to cattle industry from all over the country, *Salmonella* Mbandaka is isolated mainly in northern France and Brittany regions (Figure 2.18).

2.3.8.2 Clinical and prevalence of *Salmonella* in cattle

According to a systematic review based on studies published between 2000 and 2017 [299], pooled prevalence of *Salmonella* in cattle is around 9% in the world. Significantly high heterogeneity was observed within continents, where the prevalence varied from 2% (Europe) to 16% (North America). In addition, prevalence varies among all studies, because some studies have focused on farms with very high prevalence for the purpose of investigation. For example, a study in USA estimated a prevalence from 27 to 31% of dairy herds are colonized by *Salmonella*, taking into account environmental samples [304]. In France, prevalence of *Salmonella* in cattle is at the same level as for pork, around 9% [300]. As for pork, reporting salmonellosis cases in cattle and dairy farms are not mandatory.

Beyond the considerable economic losses, contaminated raw milk or finished products infected by carrier cows can cause severe infections in dairy cattle [305]. Although diarrhoea is a common consequence of *Salmonella* infections in cattle, the consequences of others serovars like *S.* Dublin commonly reach respiratory syndromes in calves or abortion in gravid cattle



Bovins (33 sérovars)
S. Dublin : 26,1%
S. Montevideo : 25,5%
S. Typhimurium : 15,4%
S. Mbandaka : 11,1%
Variant de Typhimurium : 8,8%
Autres sérovars : 13,1%

Canards (51 sérovars)
S. Give : 19,5%
S. Enteritidis : 12,5%
Variant de Typhimurium : 11,1%
S. Anatum : 10,5%
S. Indiana : 10,1%
Autres sérovars : 36,3%

Dindes (40 sérovars)
S. Napoli : 17,6%
S. Senftenberg : 11,2%
S. Newport : 10,8%
Variant de Typhimurium : 10,4%
S. Veneziana : 6,4%
Autres sérovars : 43,6%

***Gallus gallus* (144 sérovars)**
S. Senftenberg : 40,1%
S. Livingstone : 7,8%
S. Montevideo : 7,6%
S. Enteritidis : 4,1%
S. Napoli : 3,8%
Autres sérovars : 36,6%

Figure 2.17: Distribution of the five main serovars for each of the four main sectors of the animal health and production sector (2019) from Leclerc et al. [262].

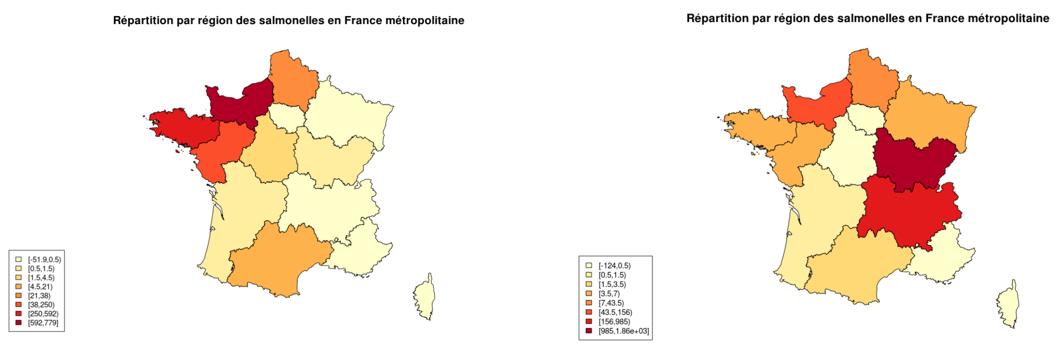


Figure 2.18: Geographical dissemination of *Salmonella* Mbandaka and *Salmonella* Dublin in France between 1999 and 2021. Map inferred by *Salmonella* Network data [43] with personal data isolated from bovine sector. Left : map of *Salmonella* Mbandaka isolated from bovine sector in France. Right : map of *Salmonella* Dublin isolated from bovine sector.

[306, 307]. Calves aged two to six weeks are most commonly affected often following purchase from a market although such trade in dairy calves is now much less common. The clinical signs depend on age and the presence of passively derived immunity. Some studies report high morbidity (14% to 60%), and mortality could vary from 0% to 14% in some cases of adult cattle [308]. When salmonellosis occurs on a farm, large numbers of animals can become very sick in a short period of time. Consequently, this disease can be extremely costly. *Salmonella* infections in cattle can produce long-term asymptomatic carriers that can periodically shed bacteria in the environment, contributing to the propagation within herds [309], or to humans through direct contact or consumption of contaminated products. Also, a seasonal trend has been observed for *Salmonella* contamination in calves, with a peak seen in the autumn, which probably reflects the tendency for many dairy herds to calve later in the year [276]. The prevalence of *Salmonella* in cattle could be explained by a diversity of factors: the bacterial ability to survive in the environment, the asymptomatic carriage of individuals, the intermixing of cattle and their exchanges between farms, contaminated food and other factors [310].

2.3.8.3 *Salmonella* Mbandaka, a multi-host serovar

Salmonella enterica subsp. *enterica* serovar Mbandaka was firstly isolated from human salmonellosis in the Belgiun Congo in 1948. Soon after, this serovar has become a global problem, with occurrence of salmonellosis and strain isolation in Sweden, Belgium, the Netherlands and Germany [311]. Finally, S.Mbandaka became widespread globally, being currently classified as one of the top-10 serovars responsible for salmonellosis cases in humans in the EU [37]. It was detected in intermittent outbreaks in Michigan, or in a multistate outbreak infection linked to Kellog's cereals [312]. In Poland, this serovar has been associated with in feed and poultry [313].

Salmonella Mbandaka is a multi-host serovar, often found in cattle and poultry. Contrary to *Salmonella* Dublin, the most prevalent serovar in cattle salmonellosis, *Salmonella* Mbandaka causes asymptomatic carriage and fecal shedding but can be deadly in some cases [300, 299]. In France, *Salmonella* Dublin is the most prevalent serovar but there is a strong prevalence of *Salmonella* Mbandaka in North France. This serovar represents 21.1% of *Salmonella* serovars detected in milk in France in 2015 [43]. It is also the third serovar isolated in Uk in 2018 [276]. Despite a high prevalence, *Salmonella* has not be well studied in genomics, mostly because human cases are rare, and cases of illness among bovines are not as violent as others serovar like *Salmonella* Dublin. But some outbreaks have been studied where the number of *Salmonella* Mbandaka found in livestock or human cases where exceptional [314, 315].

2.3.8.4 Mbandaka is poorly studied in genomics

In genomics, *Salmonella* Mbandaka has been poorly studied, mainly because of the lack of human cases. Most of the time, paper reporting *Salmonella* Mbandaka genomics results are from paper with multi-serovar analysis [315]. For example, the first sequencing of *S.* Mbandaka strains from cattle was achieved in 2013 [315] and displayed some annotation on the genome, like GC content, genome length and number of genes predicted. Another study from Timme et al. [316] associated a *Salmonella* Mbandaka strain with a sublineage close to *Salmonella* Typhi. Finally, a study [87] calculated the genetic changes of *Salmonella* Mbandaka after short timed heat treatment in a low water activity and high fat matrix. These strains displayed 19 randomly appeared SNPs in the genome after 10 heat treatment cycles, and were responsible for a population of diverse isolates. These results highlighted the importance of stress conditions of the *Salmonella* in source tracking investigations.

Finally, one thesis [317] focused on genomics aspects of *Salmonella* Mbandaka as AMR or host attribution. The phylogenomic structure of *S.* Mbandaka was studied, incorporating enormous amount of genomes from different country and sources. It also identified virulence and AMR gene repertoire of *Salmonella* Mbandaka, and revealed the capability of this serovar as a potential threat to public health (streptomycin resistance and aminoglycoside resistance genes for example). It also identified the major sequence types of *Salmonella* Mbandaka : ST413, followed by ST1602, ST2238, ST2404 and ST2444.

2.4 Problem statement

To improve the surveillance and the characterization of *Salmonella* in different food sectors, the UMT (Unité Mixte Technologique, Joint Technological Unit) ASIICS (Action for Surveillance, Investigation and Interventions in Sanitary Crisis) was created, teaming up three institutes involved in food safety in different food sectors: the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) and two agro-industrial technical centers, the French Pig and Pork Institute (Ifip) and the French Food Safety and Dairy Products Institute (ACTALIA). They are aiming at sharing their resources, knowledge, data, workforce, experience and expertise to gather targeted material, spread and transfer the genomic approaches and methods to food safety actors, perform research, genomics analyses and ultimately edit safety guidelines and outreach by publication the spread of the scientific findings.

In French pig sector, the dissemination of *Salmonella* Typhimurium and its monophasic variant is not yet clearly understood, especially concerning the prevalence of the TMV over *Salmonella* Typhimurium. Even if some studies have been made at the European scale [286], there is no study about the geographic diversity of these serovars in France in a specific host. The persistence of these strains in pig farms and slaughterhouses is not explained, and we do not know if there are local adaptations that explain it. There is a need to characterize their diversity over the country, and compare and contextualize this diversity with the worldwide diversity. Also, there is still some genomic investigation left to understand the reason for their breakthrough in agri-food lines in spite of the current sanitary and safety levels.

Concerning the dairy sector, the problematic is a bit different. *Salmonella* Mbandaka has never been fully investigated at a genomic scale due to the rarity of outbreaks in humans, but it remains very present and persistent in cattle farm, especially in the north-western France, without any clues about this specific geographical location. Some hypothesis about this persistence targets food products or environment contamination, but research has not been carried out. In addition, no fully investigation and genomics approach on this serovar has been performed in France, and very little precise genomic investigation in relation to its persistence and diversity has been carried out in other countries. We also do not know what factors could explain its adaptation to cattle, and whether the strains have health risks for humans.

The serovars of our project, which albeit sharing the greater part of their genome, highly differ in their global genomic diversity level, host spectrum and sanitary outcomes. Thus, two kind of limitation of the state of the art methodology appear to be challenging the breadth of the investigative efforts, calling for development of inclusive methodology.

Previous genome investigations on *Salmonella* Typhimurium, and especially TMV, demonstrated a highly conserved genome. These serovars do not harbor much SNP differences, which makes it challenging to analyze the diversity of strains. Delineated isolates within homogeneous samples is delicate, especially in the context of sanitary investigation, where attribution of source and identification of contamination is essential. With the highest resolution, coregenome SNPs, there are still limits that prevent to analyse fully the accessory genome to fully characterize the diversity at the SNP level.

The first bioinformatics limit is the requirement of a reference genome for coregenome investigation. In coregenome SNP approaches, the selection of a reference genome is a crucial step and depends on the analysed dataset. Mapping reads against a reference genome which is not well selected can bias amount of called variants and even report false positive variants from homologous recombination events and repeated regions, and then on the phylogeny [318, 319]. In addition, reads coming from a sample region which does not exist in the reference genome will not map and the related variants will not be called. This will affect subsequent analysis, especially for dataset with isolates of genetically diverse origins. Some coregenome information can be left out when analysis new emerging serovar, or a highly clonal dataset where the reference selected is not close enough.

The second bioinformatics limit is that accessory genome is left out of analysis. The accessory genome, coding and non-coding, has been shown to be responsible for key evolutionary elements in the bacterium [101, 320, 246], and therefore ignoring these data could skew epidemiological interpretations. Accessory genome could explain the diversity of a clonal cluster, or reconnect missing links between strains of the same outbreak that cannot be explained by the coregenome. News methods are taking into account pangenes [99, 100] when analysing genomics content, but these results remain focused on the presence or absence of genes and do not take into account the SNPs. While the last pangenomic up-to-date pipeline is able to take into account variant on accessory genes and intergenic regions, it does not take into account non-coding SNPs in phylogenomic tree, and no other visual representation has been proposed to display this new information.

2.5 Aims and objectives

This project calls for the holistic analysis at the genomic variant level of three serovars which recently became of notorious concerns in food chains, including the characterization of their respective diversity over the country, their comparison and contextualization with the worldwide diversity through public international data and the investigation of the reason for their breakthrough in agrifood lines in spite of the current sanitary and safety levels.

For the dairy sector, the main questions that will be addressed in this thesis are:
1/ what is the extent of the biodiversity of *Salmonella* Mbandaka in the different reservoirs (environment, feed, herd, cow, milk, cheese)?
2/ Is the *Salmonella* diversity comparable between the reservoirs and is it possible to trace back the origin of the observed diversity (such as cow feed, animal trade, environmental contamination . . .)?
3/ Are there markers distinguishing host reservoir in *Salmonella* Mbandaka genomes?
4/ What are the main genetic factors favoring the *Salmonella* persistence in livestock?

For the pork industry, the main investigation will be focus on:
1/ depict the genomic diversity of *S.* Typhimurium and its monophasic variant SI 4,[5], 12:i:-
2/ assess the link between the geographical distribution of farms and the phylogeny of the

strains
3/ compare the French diversity to worldwide diversity.

2.6 Solution approach

Fundamental work will tend to 1) develop a holistic genomic analysis method and 2) unravel the evolutionary patterns of *Salmonella* and its diversity while applied endeavors will advance toward high throughput diagnostic marker probes for sanitary control and database population from perceptive sampling. Although the thesis project is structured in two independent parts for each serovar according to practical issues (sampling, history and organization of the production chain, etc.) the concept is however common: genomics for sanitary control, assessment of *Salmonella* biodiversity in an animal production sector, origin of contamination, selection of variants and the bioinformatics methods used.

Concerning the bioinformatics analysis, a more discriminant comparative genomics method is need according to the literature. In this project, we aim at developping a whole-genome approach, on coding and non-coding regions, core and accessory, to take into account all available SNPs. After identifying this new accessory information, it also requires to use the accessory parts of the genome to increase the phylogenomic signal. The objective is to display the importance of the accessory genome in epidemiological investigations. This new method developed in this thesis will be presented in chapter 3 of the manuscript.

Concerning the pig sector, the focus is on the geographic diversity of *Salmonella* strains. Sample from waiting rooms, processing premises and pork carcasses at the slaughterhouse will be analysed to understand the diversity all along the food chain. A special attention will be given on isolates from pigs herds, as the contamination can be disseminated from herds to slaughterhouses. The genomic analysis will also focus on multi- and extended-AMR of the TMV to understand the prevalence. Finally, worldwide genomic diversity of the TMV spread will be assessed, from raw-reads material made publicly available.

For the dairy sector, we focused on the characterization of the genomic diversity all along the steps of the cheese production chain. Sampling from cattle farms (feed, water, environment, animals), transport chain, dairy and cheese production plants and distribution steps will be analysed. To compare to another host, the collection will be complemented with strains coming from poultry sector. These results will be compared to outcomes of a *Salmonella* Dublin study I also performed during this thesis, where the diversity has been characterized at a geographic scale.

For both serovar, the host or the geographic diversity will be able to serve to conceive cheaper and quicker surveillance assays, for example based on PCR primers or hybridization chip probes for high-throughput screening to give industrial workers the possibility of quickly tracing their source of contamination. All genomic analysis of these three serovars will be presented in chapter 4 of the manuscript.

Finally, results will be discussed in chapter 5. Annexe and Supplementary data will be presented in chapter 6 and 7.

Chapter 3

Improving food safety with bioinformatics development

3.1 Introduction of the Chapter

Distinguishing isolates within homogeneous samples can be challenging, especially in the context of sanitary investigation, where the identification of the origin of contamination is essential. In this chapter, I will present the strategies and developments I set up to reach a higher-resolution method in genomics-based phylogeny (i.e. named phylogenomic method in the present manuscript) [59, 146]. To improve the resolution in phylogenomic approaches, a new method has been developed called "pgSNP". The pgSNP's purpose is to take into account all pangenomic information, core and accessory [321], coding and non-coding [322, 102], in order to integrate the whole genomics variation to distinguish isolates.

The study described below is the main works of my thesis. In this chapter, the pgSNP pipeline that contributed in the characterization of bacterial samples in sanitary investigation will be presented. This pangenomic pipeline has been validated on several robust epidemiological datasets [79, 323, 324], each time consistent with expected epidemiological clustering, emphasising the importance to take into account information from the core and accessory genome at the same time to perform a phylogenomic reconstruction. These improvements are also discussed in this chapter, along with additional project proposals for even finer strain analysis.

3.2 Material and methods

3.2.1 The pgSNP pipeline

The strategy of the pangenome analysis we designed can be summarized in two main steps: 1/ defining the reference pangenome on which we can compare all sequences present in at least four samples among all those of a set and 2/ characterizing sample using phylogenomic approaches under an evolutionary model. In brief, we summarize the genomic content present across all the samples into a reference pangenome : a repertoire of unique sequences, where redundant genomic elements are merged. This reference pangenome is then used together with variant calls for each sample to reconstruct a phylogenomic tree depicting the sample phylogeny. To address the phylogenomic tree reconstruction of samples that do not share all of their sequences due to the inclusion of the accessory part, i.e. when a sample do not have an accessory segment, we resort to a two step approach: First, we generate multiple trees, one for each segment of continuous homogeneous set of samples. Second, we reconcile the phylogenomic information from all the segment trees into a super tree. In the following section, I will present the details of the methods we selected, developed and combined to implement pgSNP.

3.2.1.1 Building a reference pangenome

We decided to build the reference pangenome from contigs gathered from all the paired-end reads involved in the sample dataset of interest in order to identify the entire set of pangenomic variants, namely single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels), during downstream variant calling analyses.

In order to produce a reference pangenome, paired-end reads are assembled into contigs by a *de novo* assembly workflow called ARTwork [98] for each sample in the dataset. First, reads are filtered based on quality control (QC) and normalized [325]). Reads are then trimmed by Trimmomatic [326] to remove technical sequences such as adapters or polymerase chain reaction primers. Finally, contigs are created by Spades [70] which perform a *de novo* assembly based on a de Bruijn Graph. At this step, we have as many genome assemblies as there are samples in the dataset.

To produce a pangenome, we summarize the genomic content present across all the samples into a more unique repertoire of sequences, as a union function where redundant genomic elements are merged. Based on *de novo* assemblies of samples, we created an in-house script based on BLASTN [327] in order to build a file which contains all genomic variability existing across the dataset. First, contigs built previously are sorted based on their length, and the largest contig is used as a reference. Then, contigs or fragments of contigs which do not match with a defined BLASTN parameter set against the reference pangenome are added into this last. Iteratively, all contigs in all samples are processed through this BLAST-based process. BLASTN parameters were selected to keep a high quality and amount of aligned reads, and described at section 3.3.2. This allows us to build a dataset-specific (reference) pangenome to identify all the variants in the following variant calling analyses. The obtained reference pangenome represents all sequences of the dataset.

3.2.1.2 Pan reads alignments and variant detection

In the present study, the genomic variability between all samples of a dataset refers to all core and accessory variants identified across paired-end reads mapped against the reference pangenome. In the strategy implemented here, the short paired-reads are aligned against the reference pangenome (Bowtie2 [328]) and variants are identified through variant calling analysis (i.e. FreeBayes [95]) based on programs implemented into the Snippy workflow [329]. Heterozygous SNPs are then filtered out to only retain relevant variants and stored in a VCF (Variant Call Format) file.

3.2.1.3 Contig alignment

Based on positions of variants in comparison to the reference pangenome from VCF files, I developed an in-house script to produce multi-FASTA files harboring variants for each sample and sites containing the same character for all samples (i.e. invariant sites) for each contigs of the reference pangenome. This multi-FASTA file is required to estimate tree topology and branch lengths during phylogenomic reconstruction. More precisely, variants are modified directly in a version of the reference pangenome. An option for integrating insertion and deletion is added but was not used in the present study (discussed in section 3.4.1.2). Then, the alignments of each contigs are produced and samples without corresponding contigs are discarded from this alignment and will not be present in the contig tree. As the reference genome is not included into the multi-FASTA files, we produced as many alignments as there are contigs present in the reference pangenome. In summary, we obtained associated alignments of each contig including only variants of strains whose paired-end reads aligned to these contigs.

3.2.1.4 Phylogenomic tree inference

3.2.1.4.1 Source trees

A phylogenomic tree is a diagram that represents evolutionary relationships among organisms. To represent the relationships among bacterial samples for a fragment of DNA information, a phylogenomic tree was built for each contig alignment. To this end, we used IQ-TREE [107] which is an well-known maximum likelihood method for phylogenomic inference. It is widely used in the computation of huge trees and can also provide accurate results on small datasets [107, 330]. It is also able to propose a model which fit best on data based on the log-likelihoods of an initial parsimony tree [331].

3.2.1.4.2 Supertrees

Finally, to obtain one final tree which takes into account all core and accessory variants identified in the bacterial dataset, supertrees methods were implemented in this pipeline to reconcile subtrees (contigs tree) into a single species tree. By abuse of language, supertree methods was developed to compute species trees. Here, for our examples, it is always the same species, so the final tree inferred by supertree method will be called "pangenome tree" in this pipeline. FastRFS [113] was selected for this part. It divides source trees into quartet trees using ASTRAL-II [332] and resolve the Robinson-Foulds Supertree Problem (find a species tree that have the minimum RF distance with the input source trees). The selection of the supertree method is described in section 3.2.4.2.

3.2.1.4.3 Branch lengths

Branch lengths are an important attribute of phylogenomic trees, providing essential information to understand the evolutionary time. However, supertree methods focuses on the topology or structure and does not estimate the branch lengths. To estimate branch lengths, we used ERaBLE [333]. ERaBLE estimates the branch lengths from a phylogenomic tree and sub-distance distance matrices using a weighted least squares (WLS) branch length estimation. WLS fits the branch lengths of a tree making the distances between its leaves as close as possible to the input distances. In pgSNP pipeline, distance matrix from subtrees (section 3.2.1.4.1) are used to calculate the branch length and then applied to the supertree (section 3.2.1.4.2).

3.2.2 Quality assessment

To ensure that the quality of the pangenome, we designed a stringent set of rules and filters as follow.

3.2.2.1 *De novo* assemblies quality

Quality control was systematically performed and subsequent assemblies failing to meet a set of highly stringent rules were discarded. All datasets of raw-reads were assembled using the same ARtWORK workflow ([98]). More precisely, Trimmomatic [326] was used for the trimming step. The applied quality rules for the raw-reads are: (1) length of reads higher or equal to 50 base pairs (bp) otherwise excluded, (2) Phred score per base higher or equal to 30X, and (3) filter away adapters based on an internal database with Illumina adapters. FastQC version 0.11.5 was used to control the read quality. Then, SPAdes [70] was used to perform *de novo* assembly. For all datasets, I rejected samples presenting high number of assembled bases unaligned to the reference of the species or a high number of InDels per 100kb computed by QUAST [334]. Potential inter- and intra-genus contamination was detected using Confindr [127] based on assembly metrics and blast, respectively. Samples with inter- or intra-genus contamination according to the default Confindr parameters (samples with multiple genera found in the Mash screen step or more than two single nucleotide variants (SNVs) in ribosomal genes) were discarded from the study.

3.2.2.2 Sample set quality

Concerning *Salmonella* datasets, sample serotyping was finally performed *in silico* based on the assembled genomes using SeqSero [335]. SeqSero determines the *Salmonella* serotype based on a curated database of *Salmonella* serotype determinants (i.e genes of antigen and flagels). This workflow consists of an alignment of genes in genome sequences using BLAST and determine O and H antigens. Then, serotype is predicted according to White-Kauffman-Le minor scheme. Unless conflicting or with reasonable doubt on the error source (metadata, low coverage, etc.), lab-typed and predicted serotypes other than the studied *Salmonella* serovar [79] have been discarded. The *Salmonella* Typhimurium serovar was previously confirmed by glass slide agglutination according to the White-Kauffmann-Le Minor scheme [5] and PCRs following EFSA recommendations concerning monophasic variants of serovar Typhimurium [336, 337]. I checked carefully that the strains had similar genomics patterns and they did not present high dissimilarity that could impact the final phylogeny (i.e. long branches).

3.2.2.3 Pangenome quality criteria

Using 57 genomes and raw-reads of *Salmonella* Typhimurium, a dataset was build as test dataset to access parameters of the pgSNP pipeline. Several pangenomes were computed with different length (from the less stringent to the most stringent in term of contig length), and raw reads were aligned against all of them. At the same time, raw reads were also aligned on a single reference genome (here, Typhimurium LT2). Samtools [338] was used to compute the number of reads mapped against each alignment file from each sample. Only reads that are paired-end aligned against a reference were taken into account, since only these reads are taken into account into downstream analysis.

In addition, the read mapping quality was assessed using the Shannon entropy calculation [339]. Entropy is a measure of uncertainty. In our context, it means to quantify the sequence variability at a particular site. It is used in genomics to calculate the local variability between genomes, or within the same genome by comparing all sites. This metric is most of the time applied on virus analysis [340, 341] or in protein analysis [342], but it is also used in bacterial genomics [343, 344]. Shannon entropy is calculated for each site following the formula :

$$-\sum_{i=1}^{n} P_i log(P_i)$$

where P_i is the probability of the event i.
The lower the entropy score is, the lower the variability is. In other words, reads aligned on a position present less SNPs, and are consequently better aligned.

3.2.3 Coregenome pipeline

PgSNP was compared to a coregenome workflow, iVARCall2 [345], which detect SNPs and small InDels following the best practices proposed by the Genome Analysis ToolKit (GATK [346]). More precisely, reads are mapped on *Salmonella* Typhimurium LT2 (NCBI NC_003197.1) genome reference, and duplications were excluded before variant calling analysis via local *de novo* assembly of haplotypes in active regions. The matrices of pairwise SNP differences and pseudogenomes were computed using in-house Python scripts called 'VCFtoMATRIX' and 'VCFtoPseudoGenome', respectively [345].

3.2.4 Benchmarking of tools

3.2.4.1 Selection of the variant calling method

We wanted to check on the impact of the variant caller on the final phylogenomic tree. From the literature, some variant caller have better performance and sensibility than other [94]. In the pgSNP pipeline, the variant caller Snippy [329] was selected for several reasons. Snippy can detect variants on contigs, is really easy to use, presents additional features compared to Freebayes [95], and finally has a high performance compared to other variant callers [94].

To ensure the robustness of Snippy, the program was compared to GATK [347], also one of the most used variant calling in bacterial samples. GATK 4.0 has a high performance and is widely used because of its best practices guides which facilitate the installation and the interpretation of the results [348]. The same reads quality cleaning and the BWA [328] aligner parameters were applied on both variant callers.

3.2.4.2 Supertree method selection

Supertree methods are able to take into account all source trees (subtrees) to build a "species" tree which represents all subtrees [349, 350]. Supertree methods are not very often used on bacterial datasets. These under-representations of supertree methods comes from the improvement of sequencing and small-scale size of bacterial genomes. Supertree methods are used most of the time on datasets with partially sequenced genomes and specific data from genomes where only genes are selected, or when the alignments are way too large to be managed simply by ML or Bayesian phylogenomic methods [351, 352, 353, 354]. Before phylogenomic tree methods could manage colossal alignments in a lower computational time, super-alignment methods were trending and in competition against supertree methods [355, 352, 349].

These two methods are gradually being abandoned, but studies still try to improve their accuracy from large datasets such as mammalian genomes. Here, our principal interest in supertree methods is the management of contigs resulting from the creation of the reference pangenome, and taking into account missing data due to the accessory genome.

We compared three supertree methods [113, 332, 356] with a coregenome tree [345]. ASTRAL was proposed [332] because at it was the most accurate supertree method with its third version [357] at the beginning of this PhD.

ASTRAL finds the species tree that agrees with the largest number of quartet trees induced by the set of gene tree, also called "Maximum Quartet Support Species Tree" problem. One of the advantages of ASTRAL is its branch length in coalescence units, because all supertree methods does not calculate the branch length. Branch lengths calculated by ASTRAL are on all branch except leafs. Distance of a set of samples will appear on the tree, but the individual supplementary distance will be left out from the analysis.

Secondly, ASTRID [356] was selected, as it is the only tool published and tested on a *Salmonella* dataset to build a supertree [358]. ASTRID's method is based on the calculation of an average distance matrix and then using neighbor joining method to produce the species tree. ASTRID is less accurate than ASTRAL according to the ASTRAL-III paper, but can handle large amounts of data in a little time.

Finally, a new method developed by ASTRID author emerged, call fastRFS [113]. FastRFS divides the source trees into quartet partitions using ASTRAL-II and resolves the Robinson-Foulds (RF) supertree problem (i.e. find a species tree that has the minimum RF distance with the input source trees). FastRFS is more accurate than ASTRID and ASTRAL-II, but was not compared to ASTRAL-III. Notably, fastRFS was validated on coding and non-coding dataset.

To select the best supertree method for our pgSNP method, the selection was based on few criteria :

- First, the computation time.
- Second, compatibility with IQ-TREE.
- Third consistency of the supertrees with epidemiological data and coregenome phylogenomic inferences.

To select the method according to these criteria, supertree methods were compared on a *Salmonella* Typhimurium and monophasic variant of Typhimurium (TMV) datasets to check that the results are consistent with the serovars results. This dataset is described in detail in the section 3.3.4.1.

To compare which supertree method fit the dataset best with our criteria described above, we compared them using the Robison-Foulds distance [359], which calculate the size of the symmetric difference in splits between a tree T_1 and a tree T_2. This distance calculated using the formula :

$$d(T1, T2) = i(T_1) + i(T_2) - 2vs(T_1, T_2)$$

where $i(T_1)$ and $i(T_2)$ corresponds to the number of internal edges, and

$$2vs(T_1, T2_)$$

correspond to the number of internal splits shared by the two trees.

3.2.5 Datasets

To evaluate the behaviour and results of pgSNP, we selected bacterial outbreak datasets published or available along with epidemiological metadata and initial interpretation based on core-SNP or MLST methods. The goal being to check whether the results are epidemically consistent, and to analyze the impact of the accessory genome on the phylogenomic relationships evaluation of strains.

3.2.5.1 *Salmonella* Typhimurium and its monophasic variant outbreak dataset

The studied *Salmonella* serovars in this PhD are *Salmonella* Typhimurium and monophasic variant of Typhimurium (TMV) from the pork industry, and *Salmonella* Mbandaka from the cattle industry. To evaluate and justify the pgSNP pipeline, a well described epidemiological genome dataset of *Salmonella* Typhimurium and its monophasic variant was analysed first [79].

This paper is composed of 4 outbreaks, two *Salmonella* Typhimurium outbreaks, and two TMV outbreaks that occurred in France between 2010 and 2014. The collection of strains included 63 samples from outbreaks (8 from outbreak 1, 12 from outbreak 2, 21 from outbreak 3 and 22 from outbreak 4), with different food source of contamination (Pork, eggs and diary products). To characterize the diversity and the strain dynamics, 129 non-outbreak controls presenting the same PFGE patterns (defined in section 2.3.3.4.1) as outbreaks samples were added to the dataset. The author of the paper compared samples using different Whole Genome Sequencing approach such as coregenome SNPs, genes, kmer, cgMLST and wgMLST. Phylogenomic tree was inferred on the coregenome SNPs using iVARCall2 pipeline [345], and the authors splitted the sample dataset in two (one with all *Salmonella* Typhimurium, and one with all TMV), and made a phylogenomic tree for each using RAxML.

In this study [79], 2013LSAL03045, a sample isolated 2 years after outbreak 4 in a different geographical area was observed to be closely related to the outbreak 4 while 11CEB5591SAL, a sample supposed to be link to outbreak 4 from epidemiological data had a genomic distance too large to be related. Other outbreak samples were clustered together as predicted, and the pairwise SNP distances between intra- and clusters supported these epidemiological and phylogenomic-based clusters. Some sporadic samples are near outbreak samples, but not supported enough by non-parametric tests between pairwise SNP distances to be considered as linked to the studied outbreaks.

In addition, *Salmonella* Typhimurium samples (n=57) have been used in the validation of the pgSNP pipeline.

3.2.5.2 *Escherichia coli* outbreak dataset

Then, to explore the possibility of using pgSNP on another bacteria dataset, I selected a dataset of a species well-studied, as well with high quality metadata for outbreaks and sporadics samples. *Escherichia coli* infection was the third most reported zoonosis in humans, which increased from 2015 to 2019 [360]. Since the complete genome of *Escherichia coli* O15:H7 has been assembled [361], I decided to study Verotoxigenic *Escherichia coli* (VTEC) O157:H7 strains based on a dataset presenting well-characterized epidemiological data [323]. The other advantages of this dataset are the short isolation period and the short branch lengths displays on the phylogenomic trees. This would allow us to more easily demonstrate the impact of the accessory on the phylogeny, as differences in the accessory genome could have a considerable impact on the topology of the phylogeny and the links between the strains.

The authors collected 209 samples defined in eight outbreaks from 14 multi jurisdictional states identified between 2011 to 2013 in Canada [323]. They also added 41 non-outbreaks isolates occurring within 60 days of the selected outbreaks, making a collection of 250 samples. IIt must be noted that there was a duplication of samples whose raw-reads had different QUAST scores. We proceded to included these supplementary raw-reads in the analysis as an independent sample (n=251). In comparison to a traditional subtyping method, the goal of the study was to demonstrate the increased genetic resolution of WGS for cluster detection during VTEC O157:H7 outbreak investigation in Canada. The authors compared an in-house pipeline called SNVPhyl against wgMLST method, and demonstrated the high level of congruence with respect to the typologies generated by the two methods.

According to the study conducted by the authors on this dataset, *Escherichia coli* samples were demonstrated as highly clonal, and the strains have very few pairwise coregenome differences. In this PhD study, pgSNP was compared against the single nucleotide variant (SNV) phylogenomic tree provided by the authors, to analyze the topology and reconciliation differences between the two methods.

3.2.5.3 *Neisseria meningitidis* dataset

In the two previous datasets, the new links identified by pgSNP can not be proven because the sporadic strains do not have enough metadata information (time of isolation, geographical data, possibility to be related to another outbreak sample). Consequently, we searched databases and the literature for datasets that would allow us a more accurate assessment of the pgSNP pipeline. We found a dataset composed of validated *Neisseria meningitidis* outbreaks strains that could be relevant for that purpose [324]. *Neisseria meningitidis* is a highly recombinant species responsible of meningitis, with a high flexibility on its chromosomal structure to adapt to human upper respiratory tract invasion and other local bacteria [362, 363]. An important criteria was met with this dataset, namely that sporadic samples are annotated with robust metadata that could help determine whether they are linked to an outbreak or not.

This dataset is composed of 15 epidemiologically defined *Neisseria meningitidis* outbreaks from 2009 to 2015 in the US. It is composed of 201 samples in total, with 84 outbreak samples and 117 sporadic samples from 10 states. Samples are separated in two serogroup, serogroup B (Outbreak = 32, Sporadic = 61) and serogroup C (Outbreak = 52, Sporadic = 56). Samples related to outbreak are annotated with the outbreak number (OB1, OB2 etc), and sporadic samples which were collected at the same time and region of an occurred outbreak are annotated with the number of the outbreak (SP1, SP2, etc). The original study aimed to compare WGS methods in order to define the suitable procedure for the outbreak investigation of meningococcal disease cases [324].

The study analysed five genome analysis pipelines. The first one used the Snippy phylogenomic pipeline based on aligned core-SNPs (reference-based short read mapping) corrected for homologous recombination event using ClonalFrameML [110]. The second one used kSNP, a phylogenonic pipeline based on aligned core-SNPs [364]. The third one proposes a core-gene MLST phylogenomic inference [65] based on aligned core-gene. The fourth one is the Parsnp phylogenomic pipeline [365] using the whole genome core alignment generated by Parsnp with

SNP corrected from recombination using CloneFrameML. Finally, a Roary [99] phylogenomic pipeline was used based on the aligned core-gene protein sequence inferred by Roary and concatened by PRANK [366], with SNP corrected from recombination using ClonalFrameML.

This study concluded that most of the US meningococcal outbreaks were caused by clonal strains with a low rate of genetic variation, except for two meningococcal outbreaks (OB7 and OB8) where outbreak were caused by divergent strains. The authors observed that most of the sporadic isolates were not phylogenomically related to the outbreak isolates. Some sporadic strains, SP7 and SP14, were clustered with samples from OB7 and OB14, respectively. These sporadic isolates were collected in the same timeframe than the outbreak, suggesting they were in fact linked to the outbreak but not determined at the time of the investigation. The authors also observed that some sporadic isolates were strongly linked to outbreak strains, although they were not sampled in the same time frame.

3.3 Results: Implementation and validation of the pgSNP pipeline

3.3.1 Description of the pgSNP pipeline

The development of this pipeline was divided into three distinct tasks.

- 1 - Build a reference pangenome
- 2 - Mapping and variant identification
- 3 - Infer phylogenomic tree

For the first point, a reference pangenome was built using a cumulative iterative BLAST approach [327] (described in section 3.2.1.1). Then for the second point, using the reference pangenome, raw-reads were mapped and variants were identified across isolates to reconstruct contigs from each sample [97, 95] (described in section 3.2.1.2). Finally, for the last point, trees were inferred from alignments of contigs [107] and used to build the pangenome tree based on core and accessory variants [113] (described in section 3.2.1.4). All steps and methods of the pgSNP pipeline is described in Figure 3.1.

This pipeline was elaborated during the first two yearS of my PhD. The expectation was to develop a pangenomic pipeline, taking into account as much as information contained in samples, and also in a reasonable computation time. In order to be able to deploy the pipeline on the laboratory site and in open-source, the computation time needs to be fast to meet the needs of health investigations. The computation time depends on the dataset length, for example in a 400 samples dataset, the final phylogenomic tree is built within less than 24 hours on a 4x48 CPUs cluster. To provide a comparison, the implemented usual coregenome based-pipeline called iVARCall2 [345] used by the laboratory take to two days to analyse a 400 samples dataset. In term of effectiveness, only the variant calling step and the subtrees step are parallelized on a cluster, where one sample (for variant calling step) and one contig alignment (for subtrees step) are dispersed into one cluster job. The other steps were not optimised because of their very short computational time, e.g. steps related to the pangenome building and read alignments, which take two hours and fifteen minutes, respectively.

Now, the next section will describe the choice of parameters and methods selected for the pgSNP pipeline, and the robustness assertions established to ensure the veracity of the results.

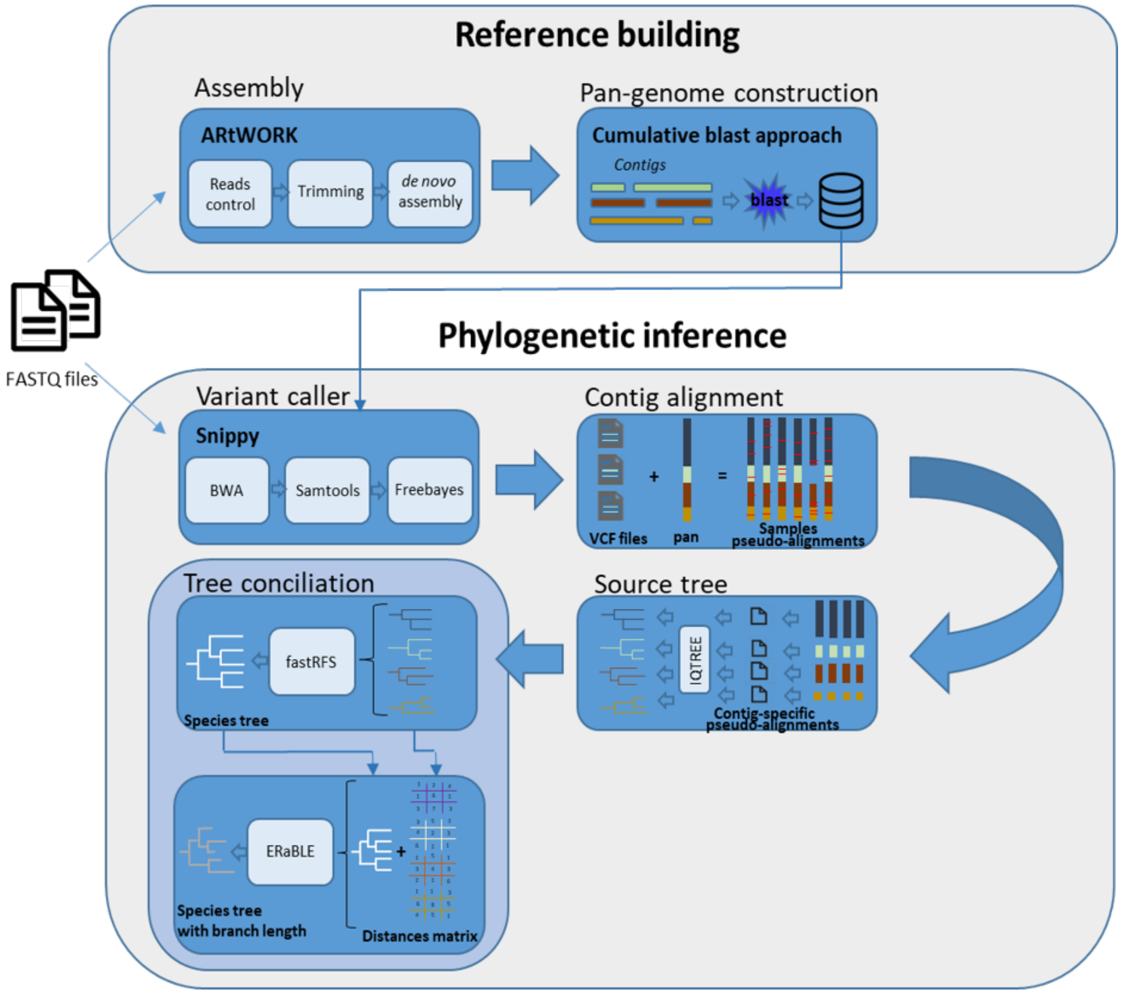


Figure 3.1: Pangenome SNP workflow description

3.3.2 pgSNP pipeline parameters

In this section we describe the set of parameters used to build the reference pangenome. Our goal was to find the best BLASTN identity parameters, and the minimum length of a DNA sequences that can be added to the reference pangenome. The BLASTN identity threshold corresponds to the maximum identity between the contigs and reference pangenome. If the BLASTN identity is set at 80%, then only sequences with average nucleotide identity below 80% will be added to the pangenome reference. The minimum length of a DNA sequence corresponds to the minimum length of contigs or non-hits sequence that we can add in the reference pangenome. If the minimum length is 500bp, then only sequences with more than 500bp are added to the reference pangenome.

In practice, the identity value quantify the redundancy in the reference pangenome, and the minimum length quantify the quality of assembly. These assessments were done to provide default setting recommendations about BLASTN identify (i.e. redundancy) and contig length (i.e. assembly quality) during the reference pangenome building.

To validate the quality of the built reference pangenome based on an iterative BLASTN method, we assessed two criteria:

- The first one is the quantity of information gained during read mapping in comparison to a usual single reference-based workflow.
- The second one is the quality of read mapping in comparison to a usual single reference-based workflow.

In this part, assessment of the pangenome quality and choices of default BLASTN parameters will be described.

3.3.2.1 Reference pangenome allows a higher read mapping

First, the objective of the reference pangenome is that it represents the diversity of the strains of the dataset, i.e that there is a maximum of reads aligned on this reference. The big disadvantage of a single reference is that some parts may not be covered, and reads that do not align with them are removed from the analysis. Here, I wanted to show that the reference pangenome makes it possible to have a greater coverage of aligned reads.

To achieve this, 57 *Salmonella* Typhimurium raw-reads (collected from the dataset described in section 3.2.5.1) were mapped on 7 reference pangenomes, including a single reference genome used in coregenome analyses, named LT2 reference genome. For the reference pangenome, BLASTN identity was set to 80% and the minimum contig length was varied to analyse the impact of the length of the reference pangenome on read mapping.

In Figure 3.2, the purple line represent all reads mapped against the LT2 reference genome. In total, 203,654,876 reads are correctly mapped against the LT2 reference genome. The number of reads was converted in percentage to get a proper graphical representation. As we demonstrated in section 3.3.2.2, intermediate values between 750 bp and 10 kb are not interesting parameters for the reference pangenome, and thus not explored in this part. The observation here displays a big gap of reads lost between pangenomes which incorporates contigs >10kb and and contigs >750 bp. Lower than this value, pangenomes that contain smaller contigs slightly increase the number of reads that align with the genome, reaching a level of about 2% additional mapped reads (corresponding to 4M reads), so more information is computed. These reads mainly correspond to accessory parts of genomes, because of the low variability of the dataset, and the fact that the reference is very close to the selected strains. Despite this, a high number of reads are aligned against the reference pangenome and will be taken into account in downstream analysis. Including others serovars, more samples or a different sequence type (ST) should increase the variability of the dataset, and thus increase the number of reads corresponding to accessory part.

From this simulation, we concluded that the reference pangenomes built with contigs below a minimum length of 750bp allow for the inclusion of a higher amount of reads and improve the quality of mapping as compared to *Salmonella* Typhimurium reference LT2 genome. Using a reference pangenome increases the number of reads aligned on it (4M reads here), showing that more information can be explored after on variant calling.

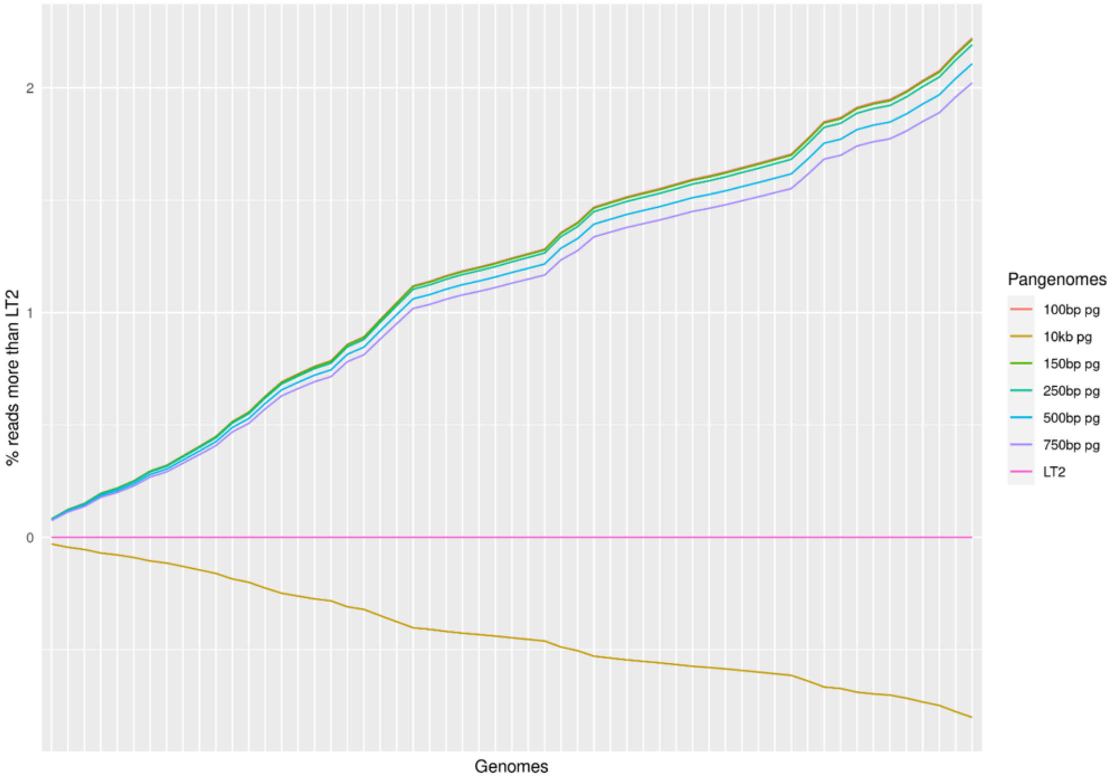


Figure 3.2: Cumulative plot of the percentage of supplementary reads which aligned on a reference. X = Genome number. Y = % of reads added to the reference. The 0 line is determined based on the number of reads aligned on LT2 reference genome. Calculated on 57 genomes

3.3.2.2 Reference pangenome allows a higher quality of read mapping

3.3.2.2.1 Global entropy computed on all reference pangenomes

Even if more reads are aligned thanks to the reference pangenome, it is necessary to ensure that the alignment of the reads is correct, and if the reference allows better alignment of reads. We computed pangenomes with different parameters, and compared the mean entropy of these to a single reference (LT2 reference genome), in order to showcase the advantage of using a reference pangenome. To avoid wrong entropy estimation due to different number of reads mapped against pangenomes, only reads that aligned onto the LT2 reference genome were selected.

This selection was made by retrieving reads which map on LT2 reference genome and use BamTools [367] and BBMap [325] filterbyname.sh to create new raw reads containing only reads which map against LT2 reference genome. By this method, all pangenomes are based on the same dataset of reads, and a bias related to the number of aligned reads aligned on the pangenome is avoided.

The entropy is computed for each pangenome, and normalized by the number of reads that align on the pangenome. The main concern here was to find a pangenome having a lower entropy than the others because less reads were mapped on it. For example, as discussed in section 3.3.2.1 and displayed in Figure 3.2, the stringent pangenome that accept only contigs

80	6.42522	6.418995	6.396853	6.39147	6.396567	6.401563	6.507905	6.695316
90	6.370421	6.356472	6.339352	6.336123	6.338286	6.342758	6.4333	6.604887
95	6.31184	6.297994	6.281942	6.27745	6.282369	6.284922	6.410159	6.531734
99	6.254178	6.248966	6.221627	6.219032	6.21915	6.220883	6.299971	6.356559
99.5	6.245479	6.231302	6.214109	6.214356	6.214767	6.215864	6.302092	6.350428
99.9	6.1715321	6.165166	6.151652	6.157276	6.153789	6.1563	6.232711	6.275764
id/contig	**100**	**150**	**250**	**500**	**750**	**1000**	**5000**	**10000**

Table 3.1: Entropy calculated for different pangenome parameters, normalized by number of reads. Score are multiplied by $e{-}11$. LT2 entropy is 6.418278$e{-}11$

longer than 10kb harbors less mapped reads. In that case, the entropy can be lowered due to a smaller sequence coverage and thus distort the comparison.

The results displayed in Table 3.1 show that decreasing identity stringency increases entropy. The lower the entropy is, the lower the variability is, meaning the reads are better aligned. Decreasing the identity causes the pangenome to have fewer contigs. This result display that reads have the ability to align correctly when they have more potential multiple position. In addition, this means that forcing reads to align on a low quality reference genome increases the entropy in some sites, emphasising misalignment events.

Furthermore, the results indicate that a shorter contig length results in a lower entropy, and so as with BLASTN identity, reads align better when they have more choices. An entropy plateau is observed between 250bp and 500bp. When small contigs (100bp 150bp) are added in the pangenome, the entropy slightly increase.

Compared to LT2 entropy, the maximum BLASTN parameters that displays a lower entropy than LT2 are BLASTN parameters = 80% identity bound with maximum contig length of 1000bp. Even the least stringent reference has a lower entropy than LT2 reference genome, highlighting a questioning about the wide use of single references used during health crises and outbreak investigations.

We showed that the use of a pangenome reduces the mean entropy, meaning that reads are properly aligned against pangenomes in contrast to LT2 reference genome.

3.3.2.2.2 Understanding the impact of pangenome reconstruction parameters on the entropy

To understand if the decrease in entropy observed above impacts a high number of positions in the genome or not, we compared the least stringent pangenome (10kb) with various pangenome (from 150 to 750bp) pangenome, and calculated the entropy for each position shared by them.

Results are plotted in Figure 3.3. On the Y axis, we observed that the number of bases where the entropy is lower than in 10kb reference (meaning there is less variability in this position) is increasing when contigs accepted in the reference were smaller. A level is reached between 250 and 500 bases where the count of difference does not seems to change significantly.

However, a slight decreasing of number of position with lower entropy is observed in smaller contigs (between 250 and 100bp). Between 10kb stringent reference and 250bp pangenome

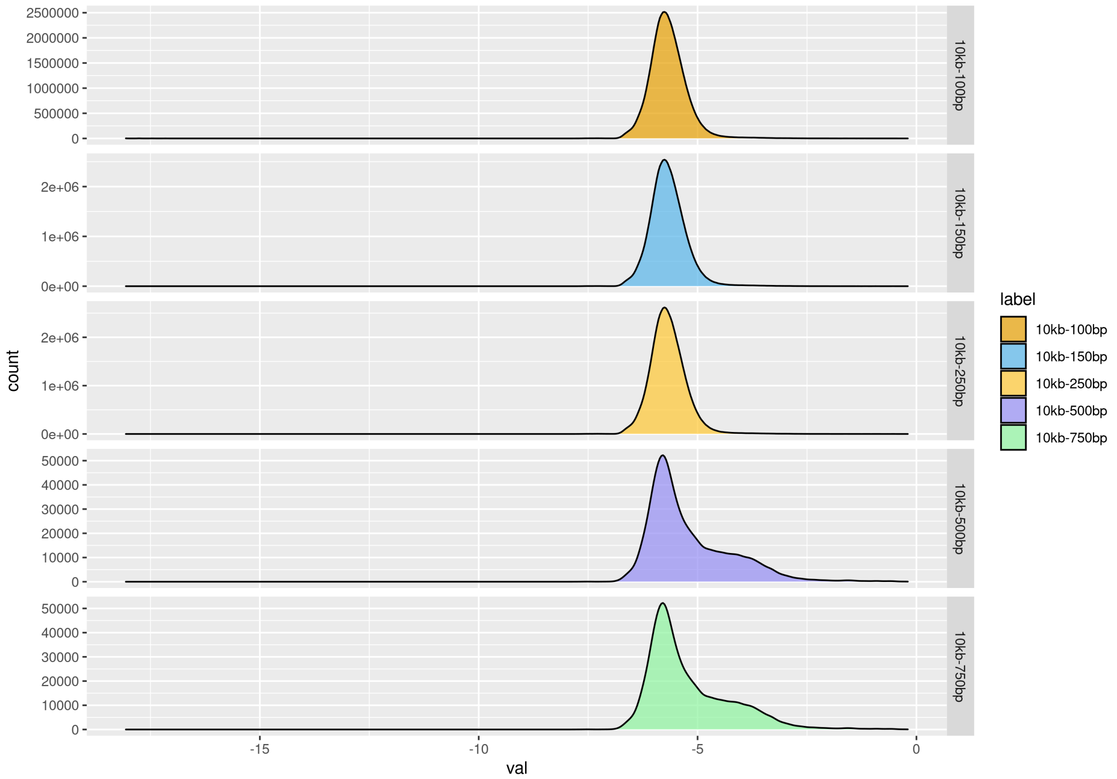


Figure 3.3: Density of the difference of entropy score calculated on each bases position, compared between a stringent reference (10kb) and various pangenome (from 150 to 750bp). The Y axis represents the number of bases in the labelled pangenome where the entropy is lower than the 10kb pangenome. The X axis represents the density of the difference of entropy score for each bases between the labelled pangenome and the 10kb pangenome. The difference is calculated between the entropy score at the position X in 10kb pangenome and the entropy score at the position X in another pangenome

reference, 2,287,281 positions exhibits lower entropy, while 2,204,359 positions exhibits lower entropy between 10kb stringent reference and 100bp pangenome reference. We hypotheses that reads with many alignment possibilities are prompt to misalignment, or else reads are truly dispatched on too many sites, which create a higher entropy than observe if all reads where mapped at the same position. These results agree with the observations on Table 3.1 as discussed previously.

In conclusion, these observations showed that the impact of pangenome parameters on read alignment is important on the quality of read mapping, and also affects a large part of the genome.

3.3.2.2.3 Example of comparison between a pangenome reference and LT2 reference genome

Finally, to provide an example, we specifically compared the LT2 reference genome versus a pangenome with 95% identity and 250bp minimum contig length. The two reference were aligned using Mauve aligner, and reads were mapped against aligned blocs for each reference. This way, the entropy score are compared on the same site for the two references.

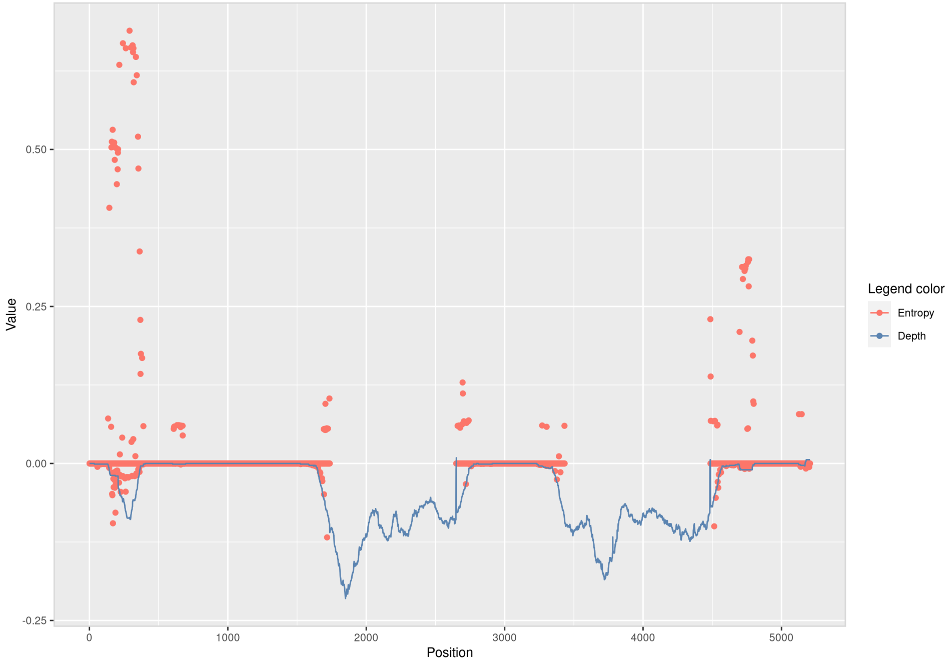


Figure 3.4: Example of entropy and depth comparison between LT2 reference genome and a pangenome reference (identity = 95%, minimum contig length = 500bp) on bloc 38. Positive values means a lower entropy or a deeper depth coverage for the reference pangenome. Negative values means a lower entropy or a deeper depth coverage for the single reference. X axis corresponds to the position on both genomes

The Figure 3.4 displays the difference of entropy and depth between a bloc aligned to reference pangenome and LT2 reference genome. In this specific bloc, there is two patterns. The first one is from site 0 to site 500 where the entropy of the reference pangenome is lower that this of the LT2 reference genome, together with a small increase of depth for the LT2 reference genome. The second one is from 1,800 to 2,800 and 3,200 to 4,600, where no reads are mapped against the pangenome, resulting in non-estimated entropy at these positions. Using reads identifier, reads which mapped against LT2 reference genome but not against the reference pangenome were located in two different unique blocs of the pangenome.

Using this information, the whole entropy of this alignment was compared between the reference pangenome and LT2 reference genome. Comparing the density of entropy for all the reads mapped against the LT2 reference genome and reference pangenome as displayed in Figure 3.5, reads aligned against the reference pangenome present higher quality scores of alignment and exhibit a lower entropy than if there are mapped on LT2 reference genome. The two peaks of entropy distribution have the same density, but the one of the single reference has higher entropy values. Also, the distribution shows that the very high entropy values in the reference single are lower in in the reference pangenome. The mean entropy in the pangenome reference is 0.013, while the mean entropy in the single reference is 0.028, displaying the advantage of using a pangenome reference.

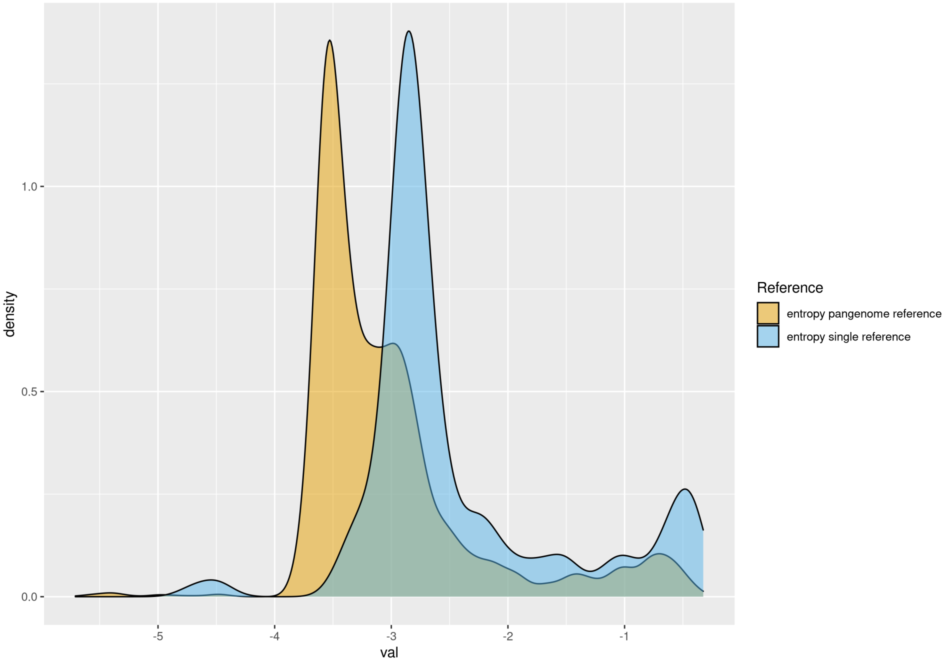


Figure 3.5: Log entropy distribution between the LT2 reference genome and reference pangenome. X axis corresponds to the log of all entropy values calculated on bloc 38 for the LT2 reference genome and reference pangenome (identity=95%, minimum contig length = 500 bp).

3.3.2.2.4 Taking into account the alignment length

Finally, one of the crucial steps of the pgSNP pipeline is the alignment step. In the alignment step, aligned genomes depend on the used pangenome, and trees are inferred on each contig. While this method allows to take into account the accessory genome, contigs where less than 4 isolates aligned against it were excluded from the phylogenomic reconstruction. So unique contigs cannot be inferred in the final tree. This element made us review the reference pangenome, because being too stringent on the identity will create unique contigs, and therefore the additional information taken into account by the reference pangenome will not be included into the phylogenomic reconstruction. To prevent the loss of information, the length of all aligned contigs which will be inferred through phylogenomic trees is also calculated and taken into account in the final assessment of the entropy score.

80	4718276	4717007	4716572	4714644	4711467	4708500	4632906	4516636
90	4736375	4736450	4734722	4730367	4728271	4725723	4669101	4565573
95	4743618	4743121	4742320	4740308	4736498	4734067	4677975	4596856
99	4681864	4681373	4680834	4679193	4677262	4675656	4670228	4643122
99.5	4596214	4595670	4595511	4595442	4597042	4598563	4597991	4587784
99.9	3720871	3719893	3719304	3719634	3721112	3723171	3747591	3786391
id/contig	**100**	**150**	**250**	**500**	**750**	**1000**	**5000**	**10000**

Table 3.2: Alignment length of all contigs inferred by phylogenomic trees

The table 3.2 displays the length of each pangenomic alignment used for tree inference. Using a very stringent reference with 99.9% identity, the alignment length strongly decreases due to unique contigs distinguishing strains. With this very stringent identity, sequences that displays 1 SNP every 1,000 bases are considered as two contigs. Looking at all the score, the alignment length increases from 80% identity to 95% identity, displaying additional accessory genome segments into the alignment. From 95% identity to 99.9% identity, the decreasing of the alignment length highlights the distinguishing of contigs, which means that the information is separated in multiple unique contigs. This therefore impacts the total alignment which is taken into account in the phylogenomic trees. Based on these observations, I hypothesise that all accessory genome has been added with 95% identity, and having much stringent genome does not add any additional information to the reference pangenome, but distinguish contigs more stringently.

3.3.2.2.5 pgSNP default parameters selection

In this section, the pangenome parameters were analysed and highlighted the advantage of a reference pangenome compared to a circular reference genome accessible in the international archives (e.g. the LT2 reference genome). The use of a pangenome as a reference increases the number of aligned reads and hence includes more diversity. In addition, the reference pangenome improves the quality of read mapping. In regard to the observations above, the final entropy score was normalized by the alignment length (see Appendix table 7.1). The best combination that displays the lowest entropy with a large alignment size is 95% BLASTN identity and a minimal contig length of 500bp and used as default settings. Users will have the ability to change this setting and adapt it to the properties of their dataset. For example, in species with fewer repeat sequences in the genome, it might be worthwhile to decrease the BLASTN identity to avoid mapping errors.

3.3.3 Selection of pgSNP pipeline tools

In this part, I will justify the choice of tools selected for variant calling and phylogenomic inference. For each part of the pipeline, tools has been reviewed, on one hand by literature, and one the other hand, when several tools had similar results or have not been compared on data sets similar to our problematic.

3.3.3.1 Variant calling comparison

The two pipelines were tested on the same 57 strains dataset described above (section 3.2.5.1). GATK identified very few different variants, most of the time in the coregenome. Some differences occurs in the accessory part, rather on the beginnings of sequences, most likely due to the different scores the two pipelines use to denote whether or not a SNP occurs. Overall, as described in 3.6, the differences are very small and have very little impact on the phylogenomic tree. The new reconciliations are made between the strains which have few differences on the core and the accessory genome. The RF distance is 30, which is negligible in view of the dataset sizes I worked with. In addition, GATK displayed slow computation time compared to Freebayes in some datasets [368].

With regard to the above results, the choice of the variant calling method does not seem to have a significant impact on the final phylogenomic tree, especially when both caller variants perform very well and have almost no difference in terms of genome clustering. In the final pipeline, Snippy was chosen due to its speed of execution and its ease of use. An option was added in the pipeline to allow the user to choose GATK if desired.

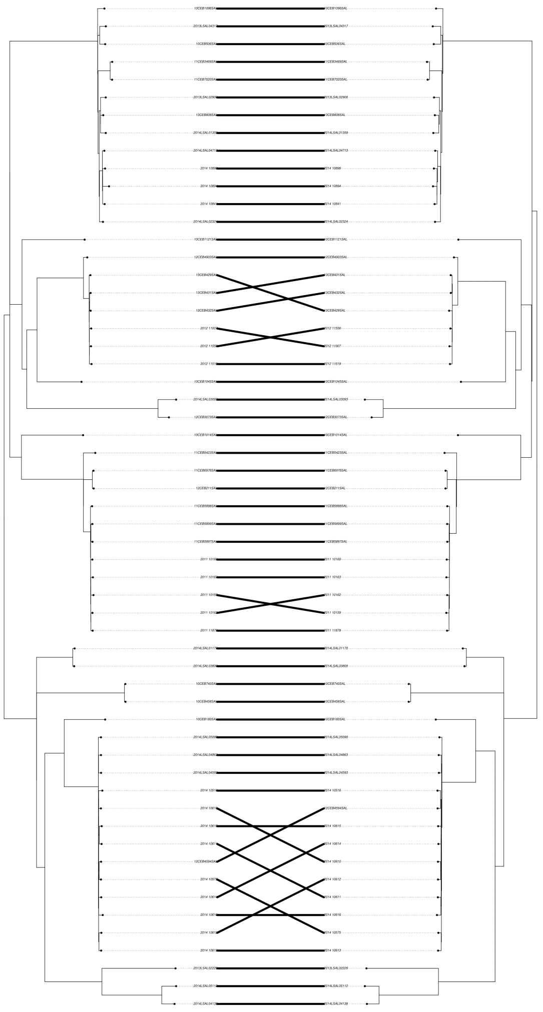


Figure 3.6: Variant calling methods impact on the final phylogenomic tree. Left tree is phylogenomic tree inferred using Snippy variant detection, right tree is phylogenomic tree inferred using GATK variant detection.

3.3.3.2 Supertree selection

As previously discussed before in section 3.2.1.4, we compared three supertree methods [113, 332, 356] compared to a coregenome tree [345]. Only ASTRID [356] has been used on a *Salmonella* dataset, while the others two focused on mammalian datasets. There was therefore a need to evaluate these methods with *Salmonella* datasets.

In the Figure 3.7, 3.8 and 3.9, all supertree methods manage to distinguish serovars as expected. We can notice that ASTRID and ASTRAL could distinguish *Salmonella* Typhimurium samples from TMV samples, while these methods were not able to recognize TMV outbreaks. With regard to RF distances between reference pangenomes and a coregenome trees (RF=308), ASTRID seems to present higher topology differences compared to the two other supertree methods. While ASTRAL and fastRFS present the same RF distances (ASTRAL : RF=228, fastRFS : RF=228), ASTRAL is not able to cluster samples from outbreaks 3 and 4. FastRFS is the method displaying the more consistency between reference pangenome and coregenome trees, and is able to cluster samples from outbreaks as expected in agreement with epidemiological data.

fastRFS was the most consistent phylogenomic tree for each outbreak compared to the other two supertree method. According to these results, we consequently selected fastRFS as the supertree method to implement into the developed pgSNP pipeline.

3.3.4 Pipeline validation on different outbreaks

In this part, I will present pgSNP results on the three different datasets presented in section 3.2.5. The objective here is to show that pgSNP produces consistent results with epidemiological data (section 3.3.4.1), and also that there are advantages of using a reference pangenome over a coregenome (section 3.3.4.2) and taking into account the accessory genome (section 3.3.4.3).

3.3.4.1 Study of *Salmonella* Typhimurium outbreaks

First, the pgSNP pipeline has been tested on a *Salmonella* dataset [79], to ensure its robustness and its consistency with epidemiological data. To compare the coregenome phylogenomic tree from the study and developed pgSNP-based reference pangenome trees, the coregemome trees were re-inferred using iVARCall2 [345] and IQ-TREE [107] methods. The final tree in 3.10 presents the same clustering as displayed by the author in the study. Three trees were inferred using pgSNP : one tree composed of all samples (n=186) was inferred to overview the variability of the dataset. The two coregenome trees from each serovar were inferred, splitting samples as presented in the study [79] and including all samples together.

3.3.4.1.1 pgSNP impact on the dataset

A reference pangenome tree was built using the default parameters previously optimized in section 3.3.2. The pgSNP pipeline took one day to create the reference, identify variants and build phylogenomic trees. The highest calculation time was taken by the phylogenomic trees (i.e. 12 hours on 4x48 CPUs).
The alignment length is 6,432,965 base long, so pgSNP adds about 1.6Mb of genetic information compared to coregenome-LT2 based analysis. This pangenome tree was calculated on 127 subtrees, where 77 are coregenome trees representing 4,859,394 bases. 157,764 supplementary bases are analysed in more than 50% of the sample, and 1,415,807 bases positions are present in less than 50% of the sample in the dataset. These rare contigs can correspond to specific adaptation to a stress of some isolates, and/or highlight acquisition of genetic material from other bacteria through horizontal transfers.

Looking at the phylogenomic tree inferred by pgSNP in Figure 3.11, the results highlight the consistency between epidemiological data and the pangenomics clustering. The *Salmonella* Typhimurium and TMV samples form two clusters on each side of the tree, with limited inconsistencies between the sub-typing annotation and the genomic clustering. Indeed, 12CEB3073SAL, 12CEB3073SAL and 12CEB4594SAL, annotated as TMV, are clustered with *Salmonella* Typhimurium samples with two long branches. In addition, 10CEB1178SAL, 10CEB1178SAL, 10CEB01160SAL, 12CEB70SAL and 13CEB1205SAL are 5 samples annotated as *Salmonella* Typhimurium but are clustered with TMV samples. These inconsistencies are a bit visible on the coregenome phylogenomic trees of the article, but are not explored

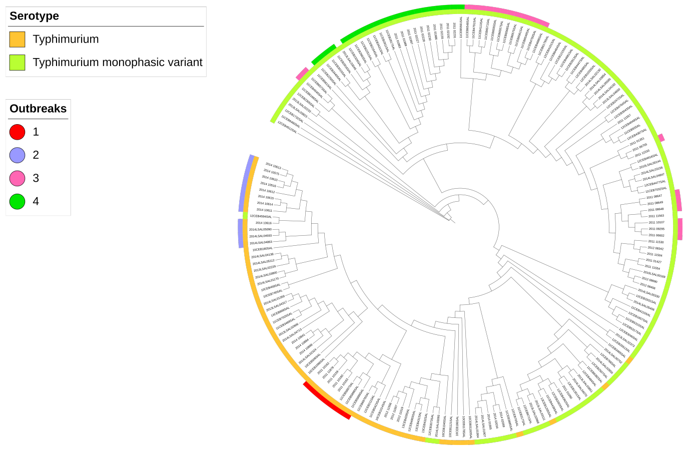


Figure 3.7: ASTRID supertree inferred on *Salmonella* Typhimurium outbreak [79]. Pangenome reference parameters : identity = 95%, minimum contig length = 500 bp. The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

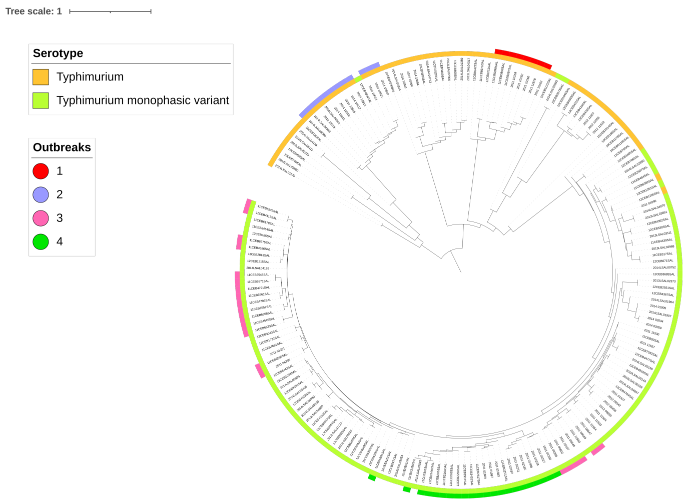


Figure 3.8: ASTRAL supertree inferred on *Salmonella* Typhimurium outbreak [79]. The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

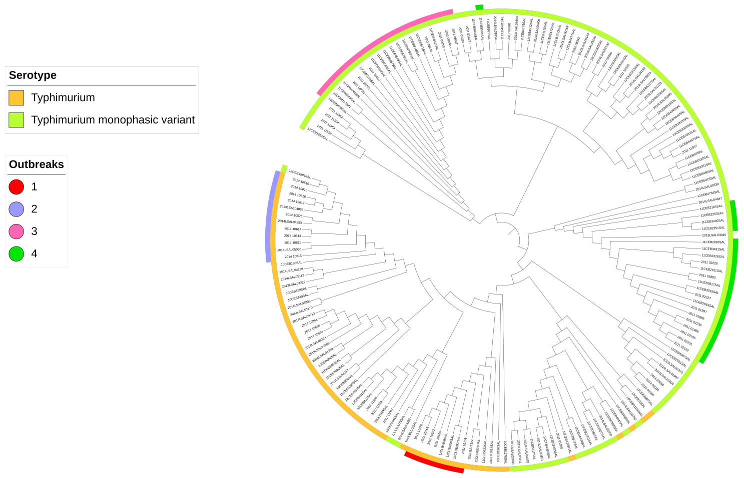


Figure 3.9: fastRFS supertree inferred on *Salmonella* Typhimurium outbreak [79]. The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

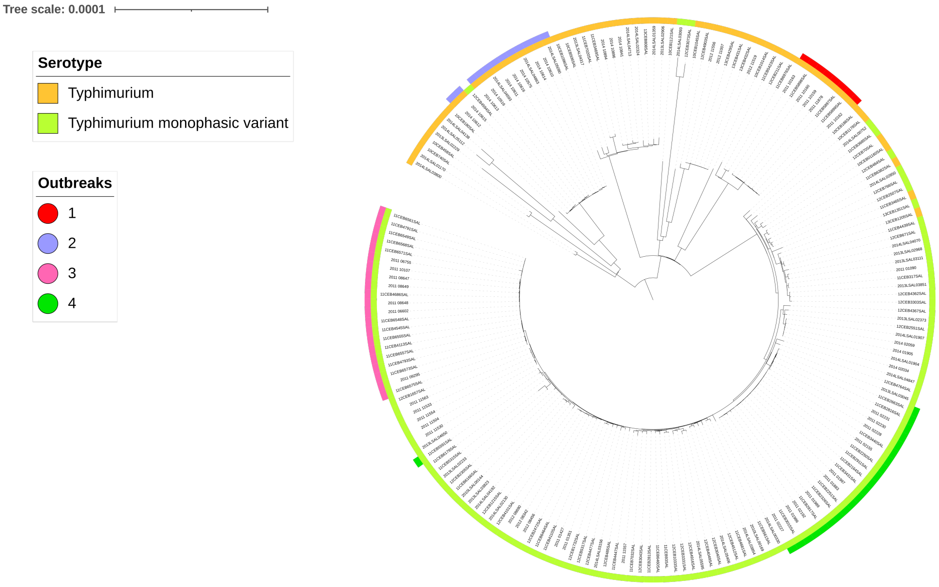


Figure 3.10: Coregenome iVARCall2 tree from the paper, computed with all dataset in one tree (n=186). The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

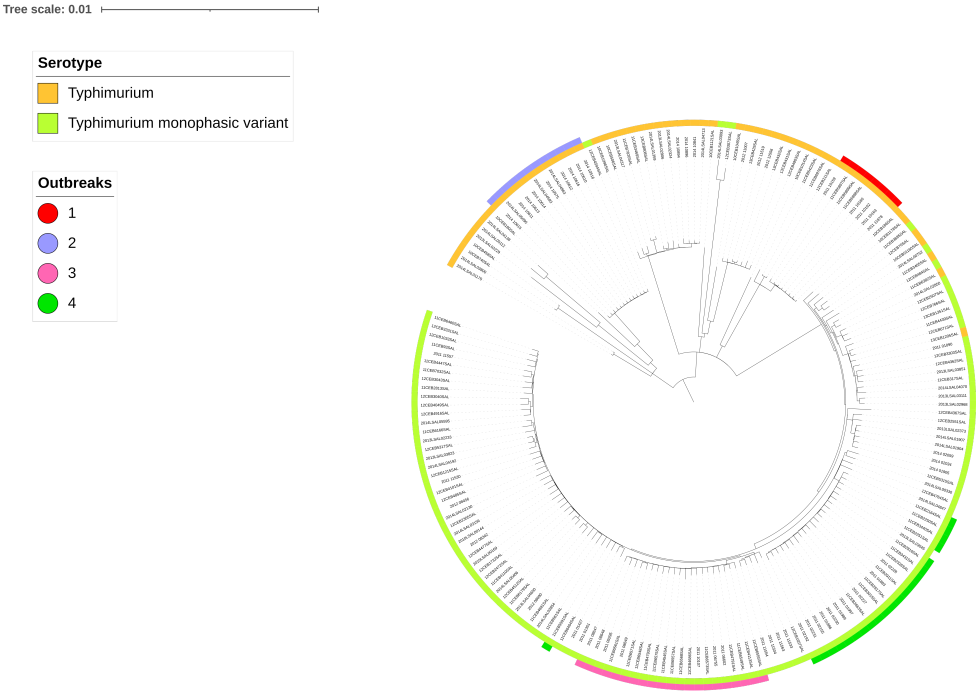


Figure 3.11: Pangenome tree inferred by pgSNP on *Salmonella* Typhimurium outbreaks from [79]. Parameters : 95% identity, contig minimum length 500bp. The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

by the authors [79]. As the annotation was made by sub-typing, we cannot conclude in a potential annotation error. Genomic can show a different genomic pattern that pushes towards a different evolution for the *Salmonella*, but this does not automatically has an impact on the expression of the phenotype. As the main interest here is the genomic content of the bacteria and impact of the dispensable genome in epidemiological investigation, serovars will be estimated according to genomics interpretation in downstream analysis.

Concerning the outbreaks, samples are clustered together, except for three. The first two are the false positive (11CEB5591SAL) and negative (2013LSAL03045) samples described previously in the paper [79]. The third one is a false negative sample not previously described (12CEB4594SAL) which is a sporadic sample clustered with samples from outbreak 2 . As the sample was sub-typed as a TMV, this result was not discovered by the authors. While discussing the results with the authors, they linked this sample with another *Salmonella* Typhimurium outbreak 2 in another analysis, reinforcing the assumptions that the strain is well connected to the outbreak [286]. However, the metadata of this sample show a high geographic distance and temporal distance with the samples from outbreak 2. Outbreak 2 occurred in 2014 from contaminated eggs in two neighbouring French departments. 12CEB4594SAL has the same ST than Typhimurium samples, so the genomics link is possible, but this sample was isolated 400km away. Also, this sample was isolated in October 2012, so the time lapse does not match a short-term outbreak. Another sample from the same region at the same date was sub-typed as a TMV, and was not clustered at all with samples from the outbreak 2.

There are two hypotheses for this sample (12CEB4594SAL). The first one is that the metadata are wrong and that this sample comes from the same department than the outbreak 2 samples. This hypothesis is proposed because the metadata are the same as the 12CEB4916SAL sample except on the ST prediction. The second hypothesis is that this sample is in fact linked to the outbreak, as the matrix isolates are the same than the source contamination of the outbreak. The transport network in poultry industry is really complex, but maybe a link between the two herds can be established. The hypothesis of the contamination through network is likely to be possible as this kind of transport is well implemented in France [369]. In addition, the 12CEB4916SAL sample was isolated from a broiler, not eggs or hens, so the two sample can come from the same parent flocks, and the contamination could have happened earlier in the production chain [336]. Finally, hens and broilers are subject to contamination from the environment, so the contamination through a vector like wild birds is possible. It was described previously [370, 371] that infected bird droppings contaminate food or water and thus transmit the disease to the farm. It could also explain the two year span between the samples.

Differences of topology were not observed between pgSNP and iVARCALL2. In both trees, outbreaks are clustered together. The few differences come from the new reconciliations intra-cluster with longer branch lengths in pgSNP phylogenomic tree. At leaf level, isolates presents differences in the accessory genome as observed in the subtrees and pgSNP tree, even in outbreak clusters. The topology difference of the two trees was assessed using cophyloplot [372].

There are two profiles of differences in the plot 3.12. In the left, *Salmonella* Typhimurium samples have a topology recombination by cluster, meaning that the reconciliation only change for inner branches, not for the leaves. On the other hand, from the middle to the right of the tree, TMV samples have reconciliation on the leafs. From the observation, it seems that the *Salmonella* Typhimurium topology difference is impacted by the TMV dataset more

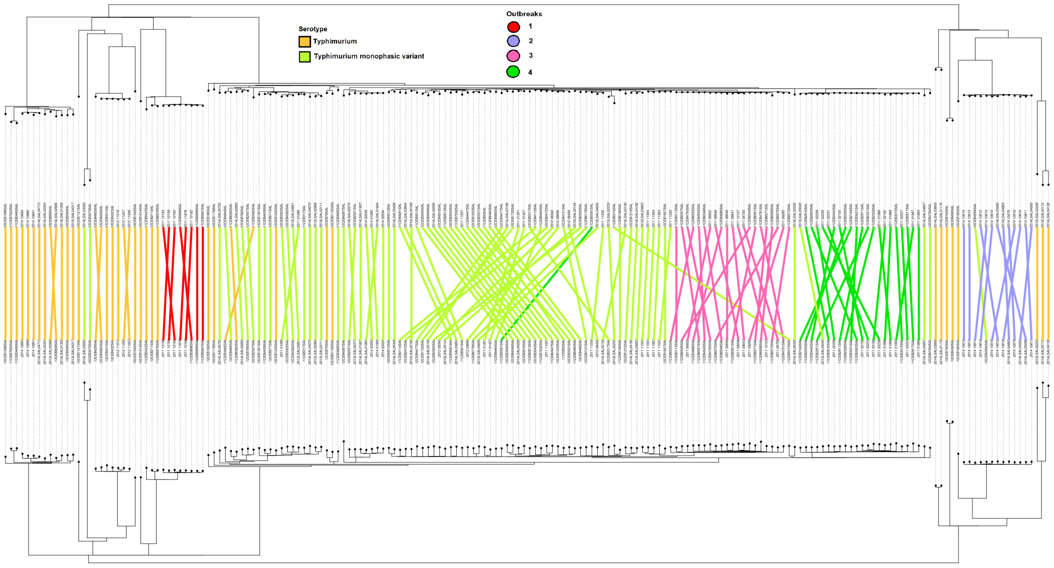


Figure 3.12: Cophylo plot of tree comparison between pgSNP and iVARCall2. Bottom is pgSNP, up is iVARCall2 tree. Color : green : TMV annotated sample, orange : *Salmonella* Typhimurium annotated sample. Outbreak color : 1 in red, 2 in light blue, 3 in pink, 4 in dark green

than others *Salmonella* Typhimurium strains themselves. Meanwhile, for the TMV dataset, reconciliations between samples seems to happen more often on the leafs, even for outbreak samples. Overall, the RF distance between the two trees is 228, and even if the outbreak clusters are respected, adding accessory genome create slightly more distances between these samples, but not enough distance to really split them completely.

To visualise the results more precisely and to understand the different reconciliations, *Salmonella* Typhimurium dataset (57 samples) was separated from the TMV dataset (123) for a more precise comparison between the two trees. The annotation of *Salmonella* Typhimurium or TMV isolated was redone using the genomic results, and the samples where the serotype was not determined were excluded from the analysis.

While looking at the comparison of the two trees in Figure 3.13, we observe that the topology of the *Salmonella* Typhimurium dataset is more preserved in both trees concerning the topology of the TMV dataset. In addition, the RF distance between pgSNP tree and iVARCall2 tree is 186 for the TMV dataset, and 58 for the *Salmonella* Typhimurium dataset, showing much more topology differences for the TMV dataset compared to the *Salmonella* Typhimurium dataset. One of the hypotheses is that the distance is bigger between *Salmonella* Typhimurium samples than between TMV samples. Because of this higher distance between the samples, adding accessory information does not create news reconciliations between samples. The new reconciliations only affect samples which are very close, for instance outbreak samples or non-outbreak samples with small branch length in the tree. On the other hand, for the TMV tree, all the isolates are very close genetically to each other (few SNPs distance), so the slightest information added on the alignment has a great impact in the topology of the tree. As we can see on the figure (Figure 3.13), the clusters are preserved, but the reconciliation between the samples is different, and some isolates or group of isolates (e.g. outbreak 4 in dark green) harbor alternate topologies.

3.3.4.1.2 Additional information taken into account

To add perspective on these topology differences, the genome alignment taken into account to create the subtrees was also studied. Looking at the *Salmonella* Typhimurium alignment, the supertree was inferred from 5,227,660 bases. As a reminder, the LT2 reference genome was inferred tree from 4,857,450 bases, so *Salmonella* Typhimurium seems to harbor a vast accessory genomes. Looking at the alignment of the subtrees, 123 subtrees correspond to coregenome alignment, in a total of 4,754,134 analyzed bases. The rest (i.e. 86 subtrees, 473,526 bases) is considered as accessory as it is not present in all samples. This result is interesting because even if we only analysed only 57 *Salmonella* Typhimurium samples, the wide accessory genome provides itself much informative.

For TMV dataset sample, the alignment length is 5689954 bases. The coregenome of TMV is 4876166 bases distributed on 127 subtrees, making 813788 bases considered as accessory genome (i.e. 73 subtrees). Putting it in perspective with the number of samples (i.e. 123 for TMV isolates, 57 for *Salmonella* Typhimurium isolates), *Salmonella* Typhimurium seems to contains more accessory genome than TMV samples.

But the main results here is that some sample coregenome fragments not existing in LT2 reference genome are taken into account by pgSNP. This result emphasizes that the pipeline take also into account coregenome fragments which would not be recognise as such by methods using a reference genome which do not harbor these last.

Finally, most of the accessory fragment in TMV are present in a very small panel of strains, hypothesising short evolution span. In addition, *Salmonella* Typhimurium accessory presence is more distributed in the panel. Looking at the accessory genome, the evolution of TMV seems to be more clonal, as described in chapter 4 section 4.3.4.1.

On a side note, the accessory genome alignment content seems to be larger for the entire dataset compared to the sum of accessory in the TMV dataset and the *Salmonella* Typhimurium dataset, considered separately. In the largest dataset, there are new contigs added that were not taken into account in single-serovar datasets because there were not enough strains that shared this contig to reach our threshold of sequence inclusion (four). For example, we have contigs which are shared by 2 *Salmonella* Typhimurium and 2 TMV samples that are added as accessory genome. Also, there are 6 supplementary samples in the largest dataset, that were not taken into account in single datasets due to the contradiction between the *in vitro* serotyping result and the *in sillico* genomic serotyping prediction between *Salmonella* Typhimurium or TMV of these strains.

3.3.4.1.3 Content of the reference pangenome

To analyse precisely the difference between our reference pangenome and the LT2 reference genome, we aligned them using Mauve aligner [373], to retrieve DNA segments specific to each reference. Mauve aligns genomes using local alignment to identify genome rearrangements. Each local alignment detected by Mauve are separated into blocs. Parts of DNA which does not align are considered as unique blocs. Mauve analysis showed that 1,749,298 bases of the pangenome reference are absent from the LT2 reference genome. Some blocs

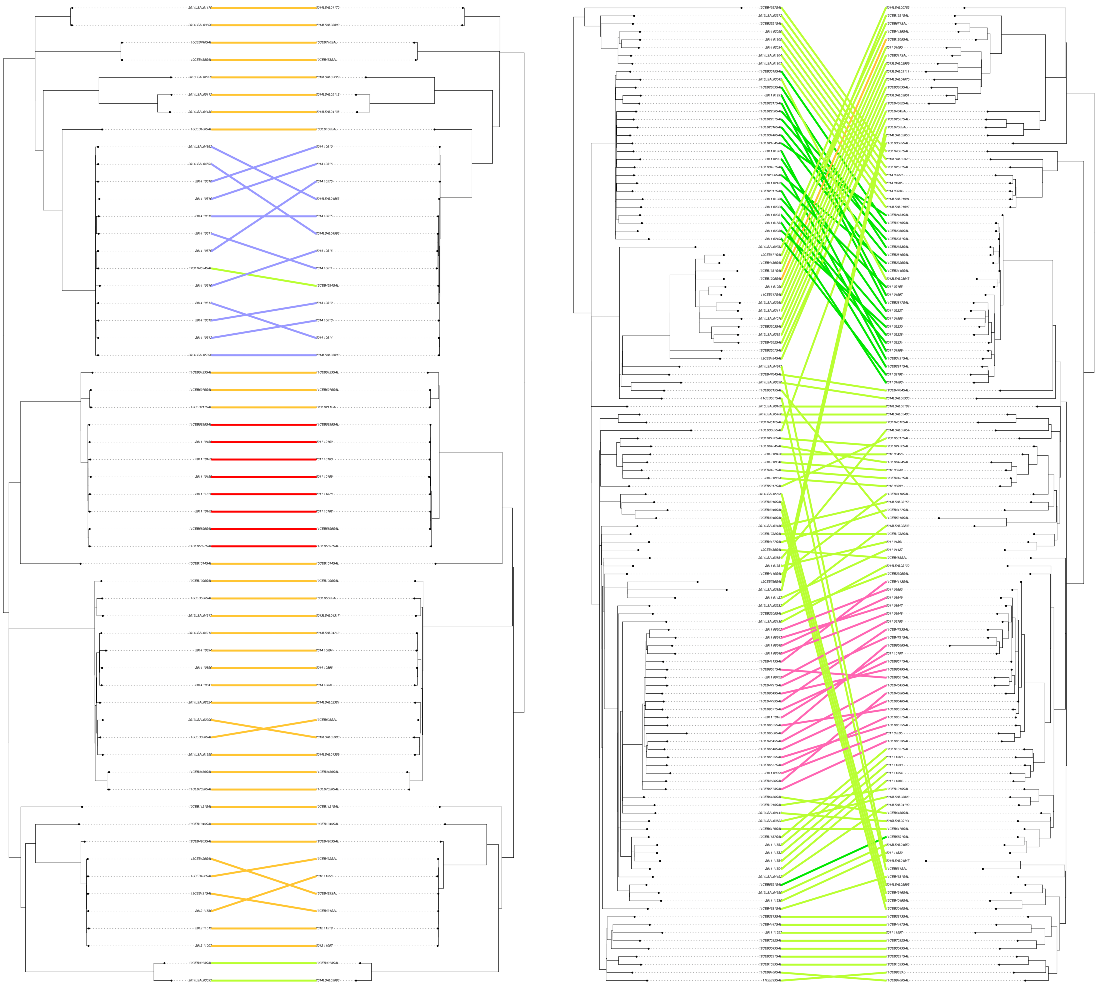


Figure 3.13: Cophylo plot of tree comparison between pgSNP and iVARCall2. Lefts tree corresponds to pgSNP, rights tree corresponds to iVARCall2. Left comparison is *Salmonella* Typhimurium dataset. Right comparison is monophasic variant of Typhimurium dataset. Color : green : monophasic variant of Typhimurium annotated sample, orange : *Salmonella* Typhimurium annotated sample. Outbreak color : 1 in red, 2 in light blue, 3 in pink, 4 in dark green

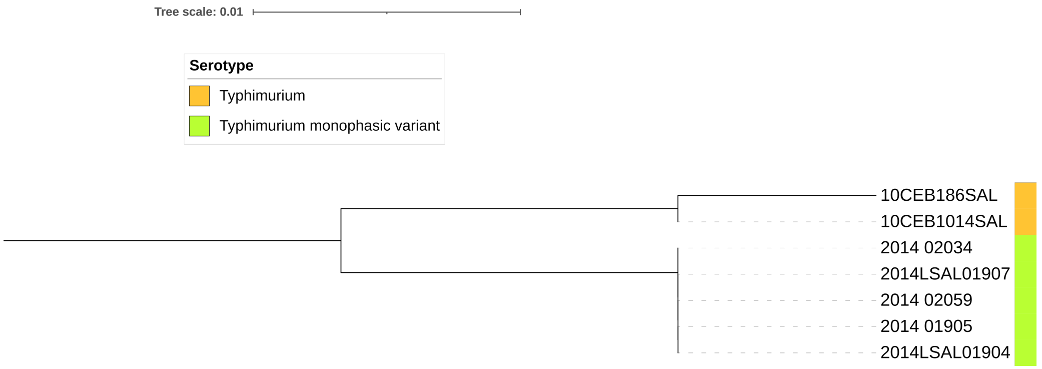


Figure 3.14: Example of a contig subtree containing plasmid pEQ2 contig. Seven samples contain this contig, and present structural and SNPs variation

were annotated using BLASTN [327] to understand the origin and functions of these sequences.

Some contigs come from plasmids or phages. For instance, in a bloc of 314kb (nine contigs), one of the contigs corresponds to *Escherichia coli* plasmid pEQ2. This plasmid displays structural variation among strains, and therefore create different reconciliations on the contig tree, as shown in Figure 3.14. One strain contains all the plasmids contigs (10CEB186SAL), while others contains only some genes (10CEB1014SAL), or only one gene (five samples) coding for a IS1 transposase gene *insB* [293, 374]. Insertion sequences (IS) are the most abundant mobile genetic elements in the *Enterobacteria*, and a study suggests that the IS1 was recently transferred between *Escherichia coli* and *Salmonella* Typhimurium [375]. This could be a perspective to understand why this small cluster of samples harbors this accessory genome, and how these samples are epidemiologically related. In the five samples, two of them come from sausage samples isolated in April 2014 in Loire region, and the other three come from three human samples isolated in March 2014 in Haut-de-Seine region. The time interval being very small, we can imagine that these strains are related. Looking at the coregenome tree, these samples are already related, by adding accessory data and detecting DNA contents that are only found in these strains emphasizes the fact that there was an outbreak that definitely occurred in this time frame, that passed under surveillance radars.

Another interesting plasmid was identified in almost all *Salmonella* Typhimurium samples except for three genomes (10CEB1014SAL, 2011_10160 and 2013LSAL02229). This plasmid is identified as the virulence plasmid pSTV-MU1, which can carry multiple *Salmonella* virulence factor genes [376] including the *Salmonella* plasmid virulence (*spv*) locus and the plasmid encoded fimbriae (*pef*) locus, along with an extensive array of IncF-associated genes [377]. Otherwise, this plasmid was detected in other *Salmonella* Typhimurium with different annotation (E40V, PNCS014854). Interestingly, this plasmid seems to be detected in some TMV samples by BLASTN, but is completely absent in all of our TMV samples. This plasmid does not present any variation in its structure, but some SNPs which are in concordance with *Salmonella* Typhimurium clusters found in the final tree, as shown in Figure 3.15. For example, *Salmonella* Typhimurium clusters of non-sporadic samples are identical to those from the corresponding plasmid tree, highlighting the existence of sporadic strains that could be related to an outbreak, undercover so far.

Finally, we detected the presence of phages in the reference pangenome. Using PHASTER [378], we detected a total of six phages incorporated in the pangenome of *Salmonella* Ty-

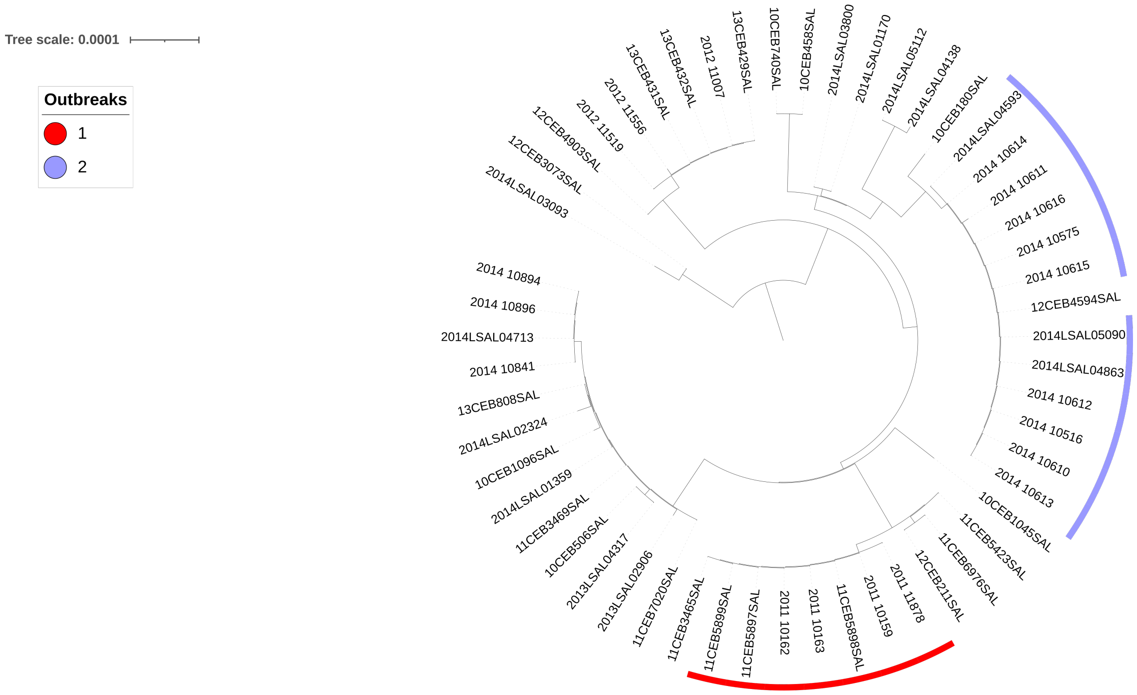


Figure 3.15: Example of a contig subtree containing plasmid pSTV-MU. 54 samples contain this contig, and present SNPs variation

phimurium and TMV in this dataset. Two phages are also identified in the reference coregenome of *Salmonella* Typhimurium LT2, but the four others are not present in the latter. In the five phages, two are present in all isolates (SfI, GF_2, Phi20), one is present in 171 isolates (118970_sal3), and finally one is present in 14 isolates (fiAA91). The mean length of the phages is 20 kb, and some phages present SNPs and structure differences between samples. For example, in phage 118970_sal3, in 46 isolates, one tail protein and some hypothetical protein are not present. This phage is present in almost all *Salmonella* Typhimurium, so this phage seems to be serovar specific. These SNPs along the phage, which are unique to *Salmonella* Typhimurium outbreak 2, strengthen the connection to epidemiological data.

3.3.4.1.4 Accessory genome distinguishes between *Salmonella* Typhimurium samples and TMV samples

To deeper understand the accessory genome distribution trough the strains, a phylogenomic tree was inferred, containing only the accessory subtrees (i.e. the subtrees which do not contain all the strains contained in the dataset). The accessory genome harbored sufficient coverage in the subtrees, so the final supertree contained the whole dataset (n=186) as shown in 3.16. Even if the accessory genome is not evenly distributed between all samples (1.4Mb present in less than 50% samples described in section 3.3.4.1.1), there is a big difference in genome content between *Salmonella* Typhimurium and TMV which distinguishes them into two clusters. This result is surprising, because there are few contigs that are serovar specific. This shows that there is enough difference in SNPs or smaller contigs in the accessory genome to distinguish strain types. On the other hand, the outbreaks samples are not clustered together, displaying the effect of coregenome on these strains, but also pointing that the accessory content can be quite different between two strains of the same outbreak, calling for a questioning about the threshold to consider whether two strains are related or not.

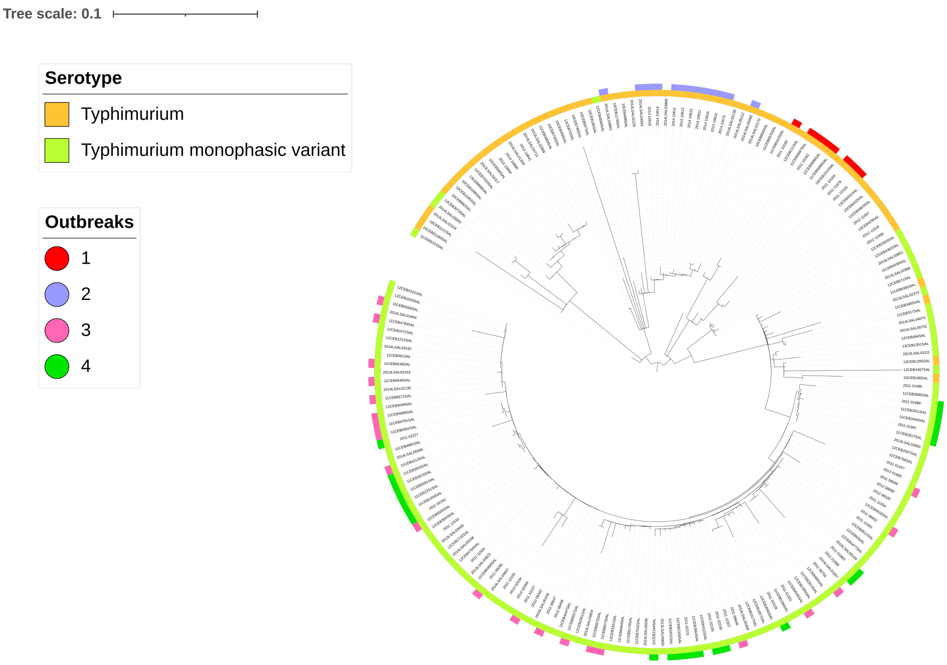


Figure 3.16: Phylogenomic tree inferred using accessory subtrees on *Salmonella* Typhimurium and TMV dataset. The inner ring corresponds to serovar annotation; outer ring corresponds to outbreaks annotation.

3.3.4.1.5 Conclusion

In conclusion, using pgSNP with this dataset does not add new reconciliations with outbreaks, but provides a better resolution when investigating these samples. The addition of the accessory genome brings a better consistency, and hypothetical links between non-outbreak strains when they share a lot of dispensable genomes. In the study, the authors conclude that the *Salmonella* Typhimurium and TMV datasets present a similar pangenome construction, but as some sample considered as TMV were serotyped as *Salmonella* Typhimurium and vice versa, the pangenome analysis was a bit distorted. More precisely, we could see that the accessory genome is different in the two datasets when annotating the samples from the phylogenomic analysis. In addition, in this dataset, the TMV samples have a fewer variability than the *Salmonella* Typhimurium samples, so the impact of the accessory genome on reconciliation has a greater consequence on branches reconciliation.

3.3.4.2 Analysis of *Escherichia coli* outbreaks

The pgSNP pipeline has been tested on a *Escherichia coli* dataset [323], to ensure its robustness and its consistency with epidemiological data for another bacterial study. A supertree of 251 genomes has been inferred using pgSNP (Figure 3.18), and compared to *Escherichia coli* tree inferred by SNVPhyl (described in section 3.2.5.2).
[323]

3.3.4.2.1 Coregenome analysis

As described by the authors and displayed 3.17, all the outbreaks are well clustered regarding their epidemiological data, except for outbreaks 3 and 6 which are scattered in the phylogenomic tree. Most of the sporadic data are clustered together on the right of the tree, close to the samples from outbreak 3 and 6. Intra-cluster distances from outbreak samples present very low SNV differences (18 maximum SNVs for each cluster). Overall, the *Escherichia coli* phylogenomic tree suggests a clonal evolution except for outbreak 1 and 2 which are distant from the rest of the tree.

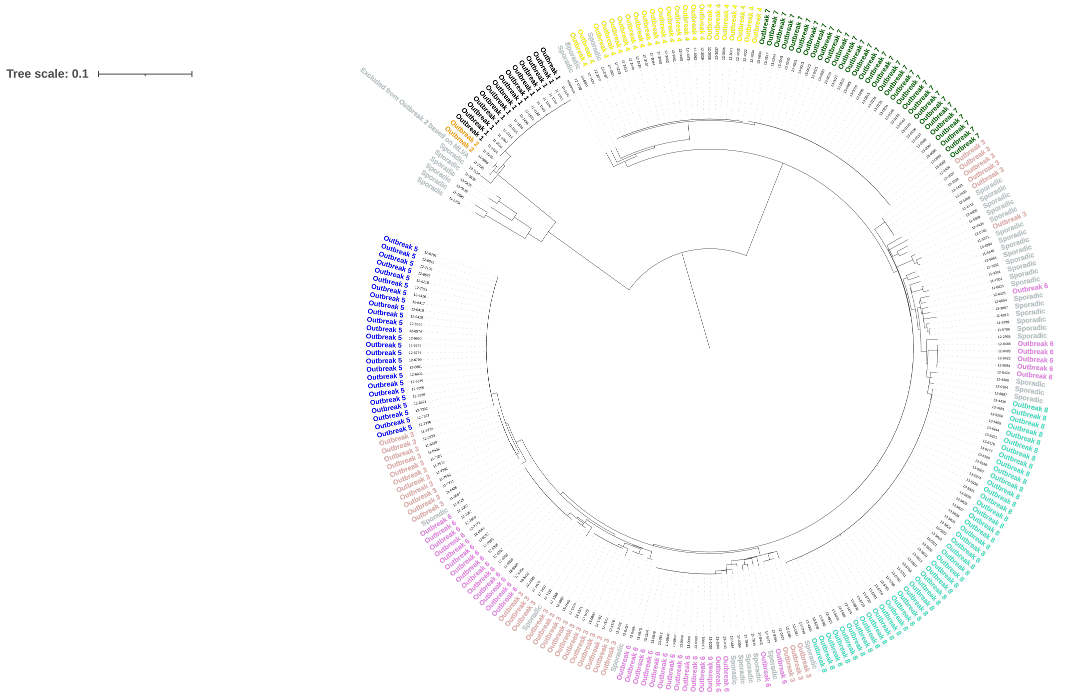


Figure 3.17: *Escherichia coli* phylogenomic tree from the study. Tree inferred on SNV pipeline described in the paper. The outer ring corresponds to outbreaks and sporadic annotation.

3.3.4.2.2 The addition of accessory genome provides new reconciliations

Using pgSNP, the topology of the phylogenomic tree (Figure 3.18) is preserved for almost all the outbreaks and is in good agreement with the epidemiological clusters. The only difference is located at the right side of the tree, where the sporadic strains close to the outbreak 3 and 6 samples observed in the coregenome are located. Concerning the SNVPhyl tree, one the right of the tree, sample 12-0745 from outbreak 3 is clustered with 9 spodaric samples between outbreak 3 and outbreak 6 samples. On the pgSNP tree, this sample and the 9 sporadic samples are clustered at the bottom on the tree. The sample 12-0745 is closer to 3 others outbreak 3, showing a consistency with the epidemiological data that did not appear on the study tree.

While investigating this difference, we observed that the accessory genome was quite small compared to *Salmonella*. Indeed, the reference pangenome alignment is 6.0 Mb long, while the reference genome NC_002695 is 5.4 Mb long. This is a low addition of accessory genomes compared to the other investigated datasets, even more for a collection of 251 samples. The

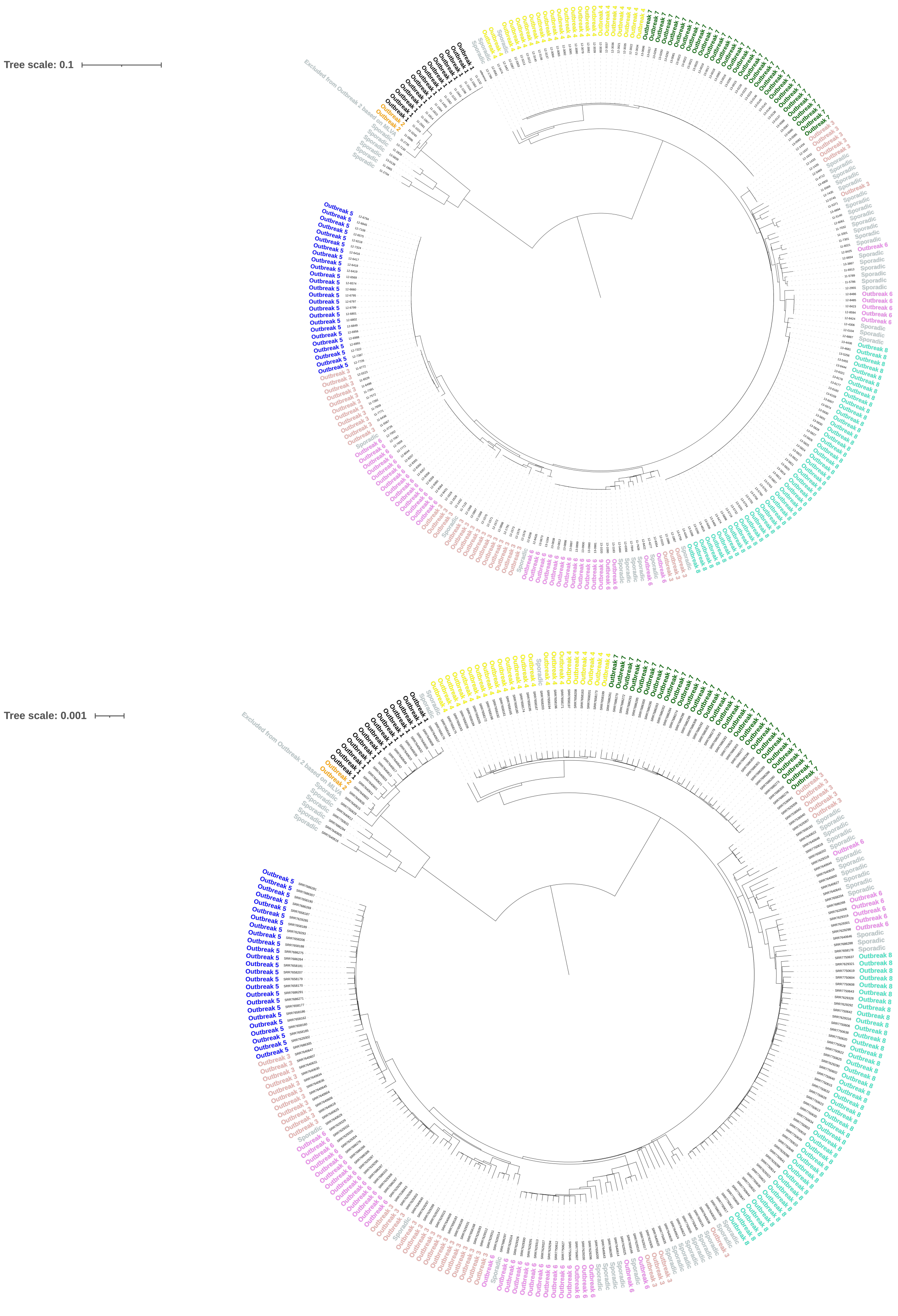


Figure 3.18: *Escherichia coli* phylogenomic tree. Left : Phylogenomic tree from the study. Tree inferred on SNV pipeline described in the paper. The outer ring corresponds to outbreaks and sporadic annotation. Right : pgSNP phylogenomic tree inferred on the dataset. Blast parameters : identity=95%, contig maximum length = 500bp. The outer ring corresponds to outbreaks and sporadic annotation.

length of the alignment highlights the low genomic variability of this dataset.

As adding the accessory genome in the phylogenomic tree did not have such a high impact on leafs of the tree, we hypothesise that the new reconciliation of these 10 samples could come from the coregenome impacted by the reference pangenome.

To confirm this hypothesis, we investigated differences between the reference pangenome and the NC_002695 reference genome of *Escherichia coli* O157:H7 used to infer the tree presented in the studied paper [323]. All the coregenome contigs were retrieved from the reference pangenome, and were aligned against the *Escherichia coli* O157:H7 reference genome by using MauveAligner. Using Biopython, the blocs of the pangenome reference which does not align on the reference genome were collected. Finally, 206kb that do not align on the *Escherichia coli* reference genome were retrieved. In other words, 206kb of DNA segments present in all isolates did not exist in the *Escherichia coli* reference genome. This is an interesting result because it highlights that the use of a reference genome that does not contain all the genomics variability can distort the phylogenomic reconstruction.

To prove that the 200k supplementary bases had the effect on the tree topology observed in the pangenome tree (3.18), a tree containing only coregenome subtrees (meaning that all samples are present in each subtrees) was inferred.

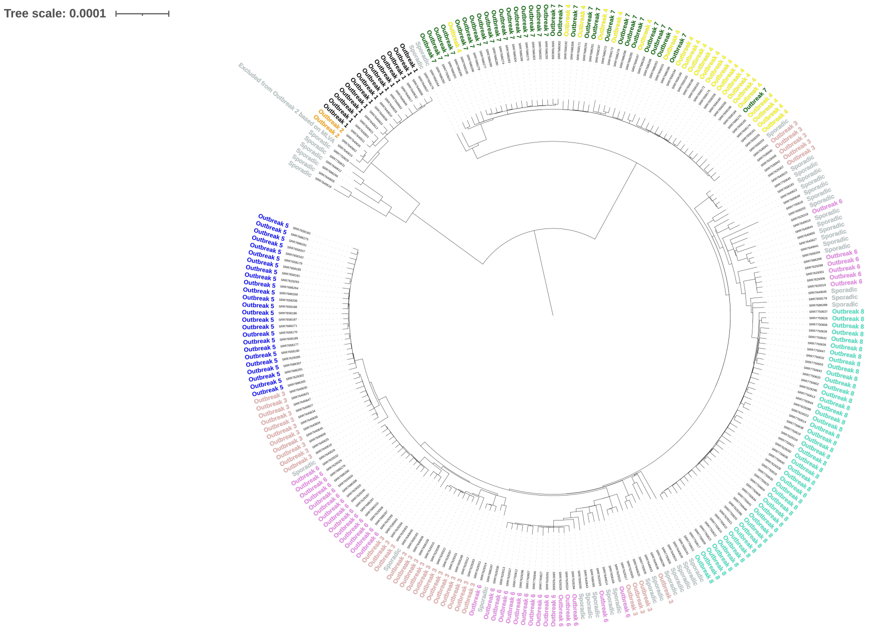


Figure 3.19: E.coli core-subtrees pgSNP inferred on the dataset. Blast parameters : identity=95%, contig maximum length = 500. The outer ring corresponds to outbreaks and sporadic annotation

Using only the core subtrees, the topology displayed in Figure 3.19 is the same than the pgSNP tree (Figure 3.18), meaning the reconciliation of news clusters does come from the core-subtrees identified using pgSNP. In Figure 3.19, the outbreaks 4 and 7 are not clustered together due the lack of a discriminating contig that segregates these strains. Based on the outcome of pgSNP, this contig is considered as accessory, thus not taken into account in this core-tree. The core-tree with this supplementary contig is displayed in supplementary Figure

7.1, and shows a reconciliation identical to that shown with the whole pangenome.

3.3.4.2.3 Conclusion

On this dataset, pgSNP provided a new strain clustering that seems to better correspond to available epidemiological information. Because the reference pangenome contains pretty much all the core and accessory SNPs of the dataset, the likelihood to miss a fragment of core DNA during the analysis is very low. In this example, the new reconciliation of the 10 samples is due to a fraction of core DNA that is not contained in this *Escherichia coli* reference genome. This result highlights the impact of the reference in coregenome analyses, and also demonstrates the benefit of using pgSNP. The accessory genome provides a higher resolution on this dataset but does not add new reconciliations. Nevertheless, the accessory genome highlights that this pipeline works very well on a bacterial dataset other than *Salmonella*, and also on a low variability and low accessory content dataset.

3.3.4.3 Testing pgSNP on *Neisseria meningitidis* outbreaks

In the two previous datasets, the new links identified by pgSNP can not be proven because the sporadic strains do not have enough metadata information (time of isolation, geographical data, possibility to be related to another outbreak sample). Here, we applied pgSNP on 201 samples from *Neisseria meningitidis* [324] with robust metadata for sporadic samples (described in section 3.2.5.3), to be able to fully determine new links highlighted by the pangenome tree. Compared to *Escherichia coli*, *Neisseria meningitidis* is highly recombinant, and has a much smaller genome size, and therefore allows the pipeline to be tested on a completely different genomic type. In this study, pgSNP has been compared to the Snippy pipeline proposed in the study (described in section 3.2.5.3). Trees inferred by the study are displayed in Figure 3.20.

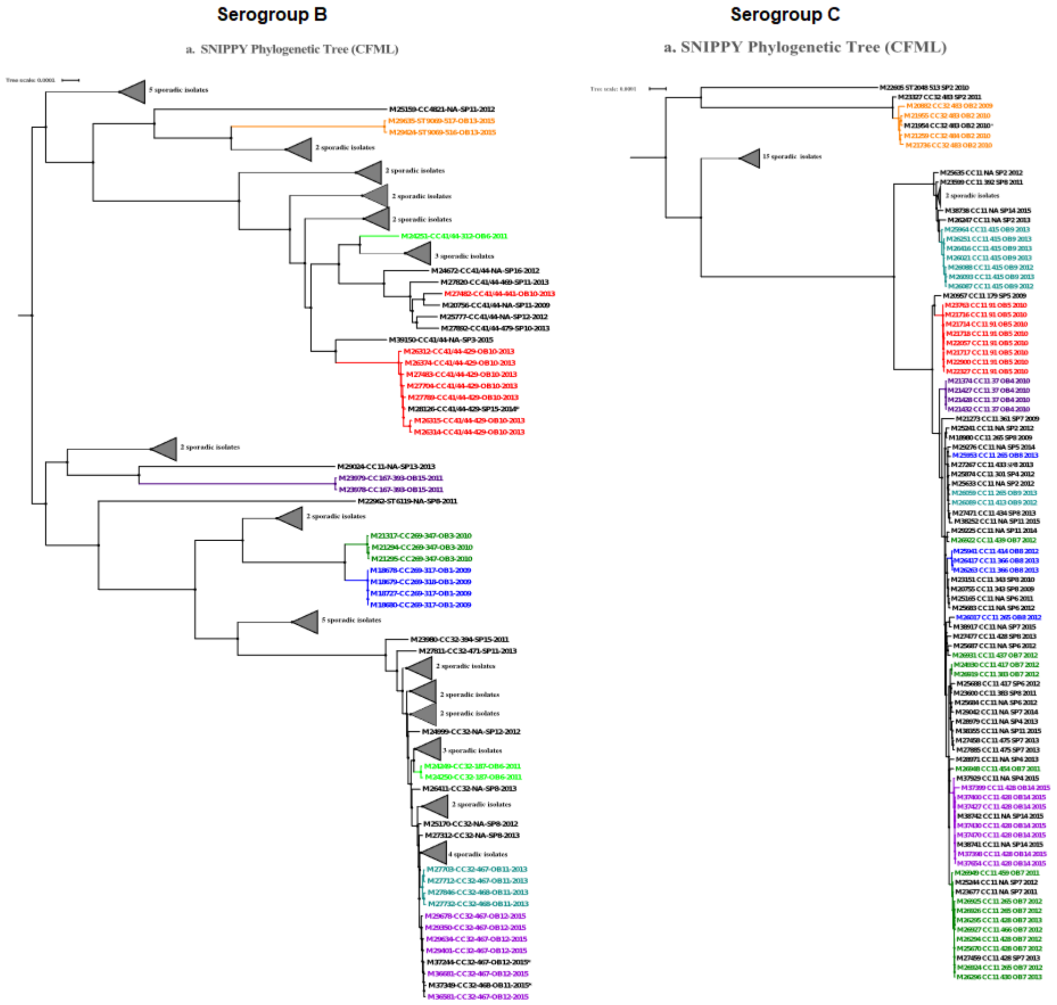
3.3.4.3.1 Results of the pgSNP analysis

In our pgSNP analysis of the *Neisseria meningitidis* dataset, we set the parameters at 95% identity and 500 bp contig minimum length. A tree with 193 *Neisseria meningitidis* samples was obtained, as eight *Neisseria meningitidis* samples had contamination or low-quality reads that did not allow proper assembly. First, the tree was inferred with the whole dataset to overview the variability and behavior of reconciliations taking into account the accessory genome.

The pgSNP tree is displayed in Figure 3.21. Outbreaks are defined in two groups : the serogroup B group and the serogroup C group. As sporadic SP8 samples were not in the same serogroup as the OB8 outbreaks samples, and because I did not want to miss potential new reconciliations via pgSNP, it has been decided to include all isolates first. Because a high number of outbreaks is assessed, the result of the analyses is reported below by serogroup for the sake of clarity.

3.3.4.3.2 Analysis of the serogroup B

Looking at the serogroup B, the outbreaks 1, 3, 6, 10, 12, 13 and 15 display the same topology clustering as described in the paper [324]. Nevertheless, the outbreak 11 (OB11) samples are not as close to each other as described in the paper. In addition, the OB 11 M27846 sample is not clustered with the other outbreak samples, and a sporadic sample (M25166) is also



Serogroup B

a. SNIPPY Phylogenetic Tree (CFML)

Serogroup C

a. SNIPPY Phylogenetic Tree (CFML)

Figure 3.20: *Neisseria meningitidis* outbreaks tree from the study. Phylogenomic trees has been inferred using SNIPPY pipeline described in 3.2.5.3. Left is the phylogenomic tree inferred on segroup B isolates. Right is the phylogenomic tree inferred on serogroup C isolates. Samples are colored by clusters.

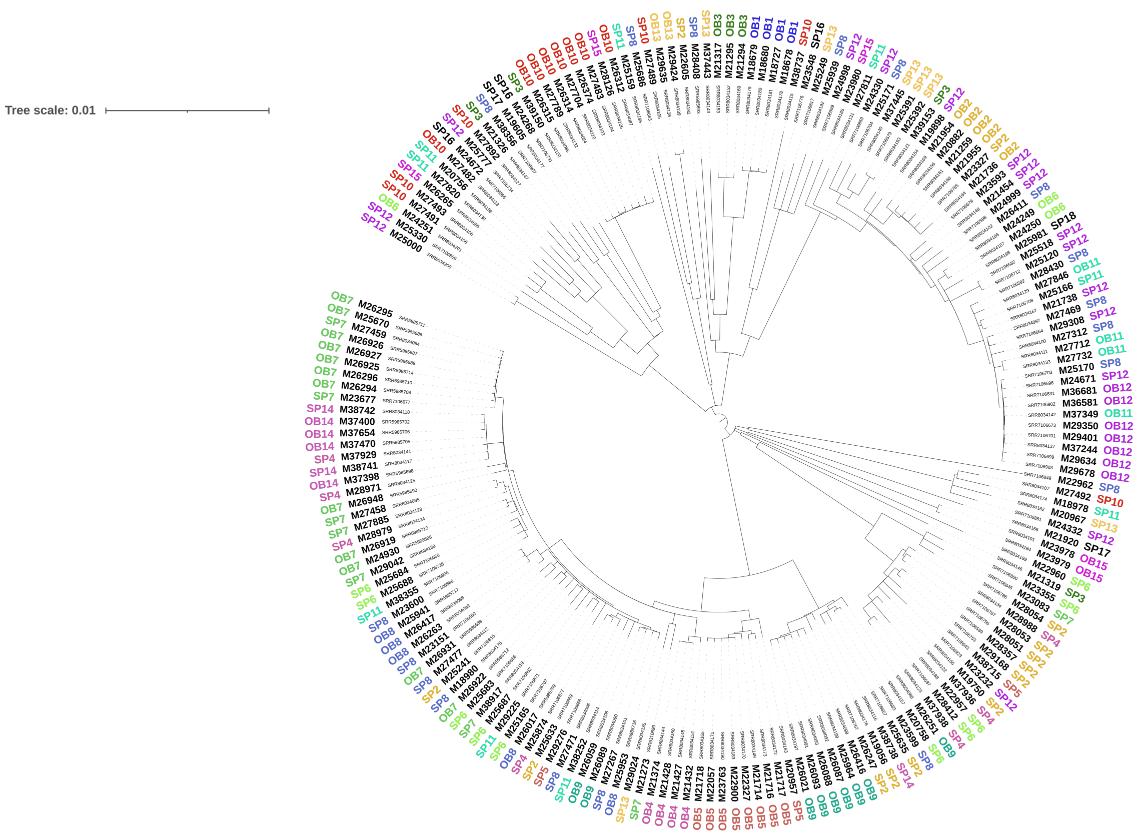


Figure 3.21: pgSNP on *Neisseria meningitidis* outbreaks. The inner ring corresponds to sample name annotation as described on the paper; outer ring corresponds to outbreaks and sporadic annotation

closer to this isolate (OB 11 M27846) compared to others from the outbreak (OB11 M27312 and OB11 M2773). One sample from OB11 is missing in the dataset because it presented contamination, so we have not been able to check the position of this isolate in the tree. To get higher discriminatory power, I decided to reconstruct the phylogenomic relationships within the serogroup B samples only.

The supertree in Figure 3.24 is inferred on a total of 3,246,450 bases, while *Neisseria meningitidis* reference genome length is approximately between 2.1 and 2.2 Mb. The Genome size common to all samples is 1,787,956 bases long. Adding bases from the 90% coregenome (contained in 90% samples), we obtain a reference pangenome of 2.2Mb. 425 kb more are carried by 50% of samples, meaning that the accessory genome with low prevalence is huge (637kb) and can have an impact on the topology of outbreak 11.

To compare the new reconciliations made with accessory genome, we focused on five relevant samples. OB 11 M27846 clustered with OB 11 M27712 and OB 11 M27732 samples in the study tree, while it is clustered with the sporadic sample SP11 M25166 in the pgSNP analysis. To understand why OB 11 M27846 is closer to the sporadic sample in pgSNP instead of other outbreak 11 samples, SNP distance and subtrees comparison was made on a total of 9 samples from cluster where the strains of interest are 3.22 and Table 3.3 . OB 11 M27712 and OB 11 M27732 samples are clustered with SP12 M29308, SP8 M27469 and SP8 M27312, while OB11 M27846 is clustered with SP11 M25166, SP12 M21738 and SP8 M28430. The pange-

id	**SP8 M27312**	**SP12 M29308**	**SP8 M27469**	**OB11 M27732**	**OB11 M27732**	**SP11 M25166**	**SP8 M28430**	**OB11 M27846**	**SP12 M21738**
SP8 M27312	0	72116	78197	56107	50156	158140	158131	173616	158960
SP12 M29308	72116	0	78746	81713	75922	170662	171324	183420	168223
SP8 M27469	78197	78746	0	82283	87221	174532	175398	184416	172097
OB11 M27732	56107	81713	82283	0	58884	170125	169874	**156883**	169817
OB11 M27732	50156	75922	87221	58884	0	165373	165504	**176691**	164618
SP11 M25166	158140	170662	174532	170125	165373	0	20054	**41721**	21280
SP8 M28430	158131	171324	175398	169874	165504	20054	0	45196	24483
OB11 M27846	173616	183420	184416	156883	176691	41721	45196	0	40348
SP12 M21738	158960	168223	172097	169817	164618	21280	24483	40348	0

Table 3.3: Snp distance calculate by snp-dists with option -a to calculate all difference on pangenome alignment of M27312, M29308, M27469, M27732, M27732, M25166, M28430, M27846 and M21738

nomic alignment is analysed on 2,614,710 bases, where 2,440,390 is considered as "core trees" (length of the genome of the subtrees shared by the 9 samples), and 177,197 as "accessory genome" (length of the genome of the subtrees shared between 1 or 8 samples).

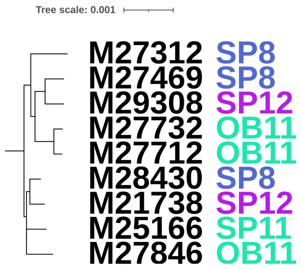


Figure 3.22: Prunned tree from OB 11 outbreak cluster.

To ensure that the topology differences come from the accessory genome and not from a coregenome fragment which would not exist in the *Neisseria meningitidis* reference genome, a tree was inferred with only coregenome subtrees in figure 3.23. In this tree, we observe that the two outbreak samples OB11 M27712 and OB11 M27732 are clustered together with outbreak sample OB 11 M27846, as described in the study.

Consequently, I hypothesized that accessory genome has an impact on the topology of these strains. pgSNP produced 96 of accessory trees for this alignment, but the presence absence matrix of trees show a higher similarity between sporadic sample SP11 M25166 and outbreak sample OB11 M27846 compared to outbreak sample OB11 M27846 and outbreak sample OB11 M27712, justifying the distance between the two outbreak isolates. If we analyse the entire pangenome alignment of all the samples and compute the distance between all of them taking into account gaps, we observe (see table 3.3) that outbreak sample OB11 M27846 and sporadic sample SP11 M25166 are indeed closer than outbreak sample OB11 M27846 to the two other OB11 samples.

Compared to coregenome, the accessory genome impacts the topology locally, and displays that outbreak sample OB 11 M27846 is not so close of other OB11 strains due to its divergence in the accessory genome. OB 11 outbreaks already displays high diversity in the study due to OB 11 M37349 sample not clustered with other, creating a maximum SNP distance of 836 SNPs (calculated by Snippy method described in 3.2.5.3). OB 11 outbreak happened in 2013, while sporadic samples in the same cluster was isolated between 2012 and 2014, which corresponds to the timeline. Unfortunately, only SP 11 M25166 can really show proximity to OB11 M27846 because the metadata show that they are isolated in the same geographical area. For the other sporadic strains, we do not have the information. Given that *Neisseria meningitidis* is recombinant [362], we can think that the strains have a possibility to be derived from the same ancestor.

Overall, pgSNP adds genetic distance between outbreak samples, especially due to the mobile genetic elements and genetic flexibility of *Neisseria meningitidis*, that can make outbreak investigation hazardous [379]. However, outbreak samples with close coregenome and accessory genome (for example, outbreak 13 samples) are still clustered together in the phylogenomic tree, demonstrating the importance of accessory genome in epidemiological investigation. This also opens up a discussion of the weight of the accessory genome in a phylogenomic tree and health surveys, addressed in the section 3.4.1.3.

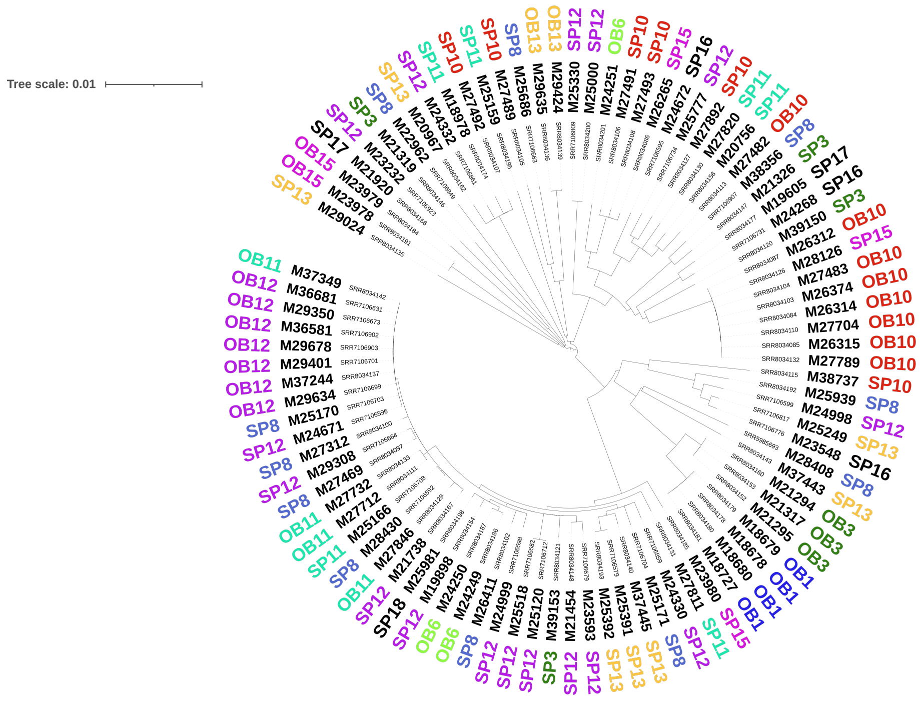


Figure 3.23: Only coregenome subtrees from pgSNP on serogroup B dataset. The inner ring corresponds to sample name annotation as described on the paper; outer ring corresponds to outbreaks and sporadic annotation

Finally concerning the outbreak 12 and analysing the tree including only the serogroup B outbreak, we observe that OB12 samples are still clustered together, but divided in two subgroups : M37244, M28634, M29678 and M29401 together, and others OB12 isolates together. Looking

id	OB8 M25941	OB8 M26417	OB8 M26263	SP8 M23151	SP6 M25165	SP6 M25683
OB8 M25941	0	20275	19523	33327	158103	168713
OB8 M26417	20275	0	17793	33352	159700	170730
OB8 M26263	19523	17793	0	33424	159267	170686
SP8 M23151	33327	33352	33424	0	146751	158926
SP6 M25165	158103	159700	159267	146751	0	79105
SP6 M25683	168713	170730	170686	158926	79105	0

Table 3.4: Snp distance calculate by snp-dists with option -a to calculate all difference on pangenome alignment of OB8 M25941, M26417 and M26263, SP8 M23151 and two sporadic sample M25165 and M25683

at the accessory genome, there is at least 7 kb of difference between those two sets of samples, associated to contigs of 5kb and 1 kb. These contigs are present in one of the datasets, making them closer intra-cluster. This difference was not seen on the whole dataset tree (Figure 3.21) because of the scale, but is visible on the tree of the serogroup (Figure 3.23), adding divergence and therefore question epidemiological clusters. This divergence can be due to homologous recombination, which can divide the cluster in two. This kind of phenomenon has already been observed in other species, when outbreak sample diverged due to homologous recombination events across 20% of genomes [380].

3.3.4.3.3 Analysis of the segroup C

For the serogroup C, the samples from outbreak 2, 4, 8 and 14 display the same clustering as previously described in the paper [324]. However, a difference appears on three samples from outbreak 8 (OB8 M25941, OB8 M26417 and OB8 M26263). They are clustered together in the original paper, while pgSNP also links the sample SP8 M23151 to these three samples. Looking at the accessory genome, sporadic sample SP8 M23151 harbors a lot of contigs in common with the three isolates of the outbreak OB8. At least 2.2Mb are shared by these four samples (OB8 M25941, OB8 M26417, OB8 M26263 and SP8 M23151). In conclusion, we have strong presumptions that the sporadic sample SP8 M23151 belongs to outbreak OB8. This would need more metadata information to conclude on these results.

Finally, we noticed that the M26251 sample (outbreak OB9) did not cluster within the same branch together with other OB9 strains in the pgSNP analysis. This isolate seems closer to other sporadic isolates (SP6 M20758 and SP14 M38738), but as the non-sporadic samples were not named in the study tree [324], no new reconciliations could be proven on these strains.

3.3.4.3.4 Conclusion

We were able to show the advantage of pgSNP using a new non-foodborne outbreak dataset, involving a bacterial species whose genome length and genome evolution dynamics is different from *Salmonella*. Using *Neisseria meningitidis* outbreaks, the ability of pgSNP to manage small length genome with high recombination rate was emphasized. pgSNP shows great consistency with the epidemiological data and the results obtained on the coregenome based on species with variable genomic characteristics. The reference pangenome increases the amount of gathered genomic information between strains and we observed that the accessory genome has an impact on the clustering of specific strains. We were able to demonstrate that one sample from outbreak 11 had a completely different dispensable genome compared to other outbreak 11 samples, and we also established a new link between outbreak SP8 samples and one sporadic sample that was not observed previously. In conclusion, this dataset analysis confirms

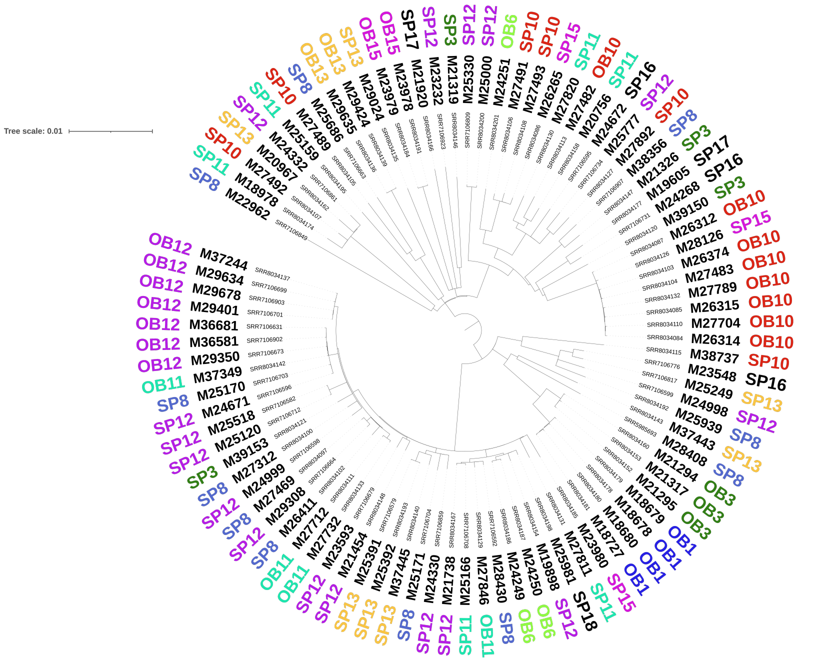


Figure 3.24: pgSNP on segroup B from *Neisseria meningitidis* outbreak. The inner ring corresponds to sample name annotation as described on the paper; outer ring corresponds to outbreaks and sporadic annotation.

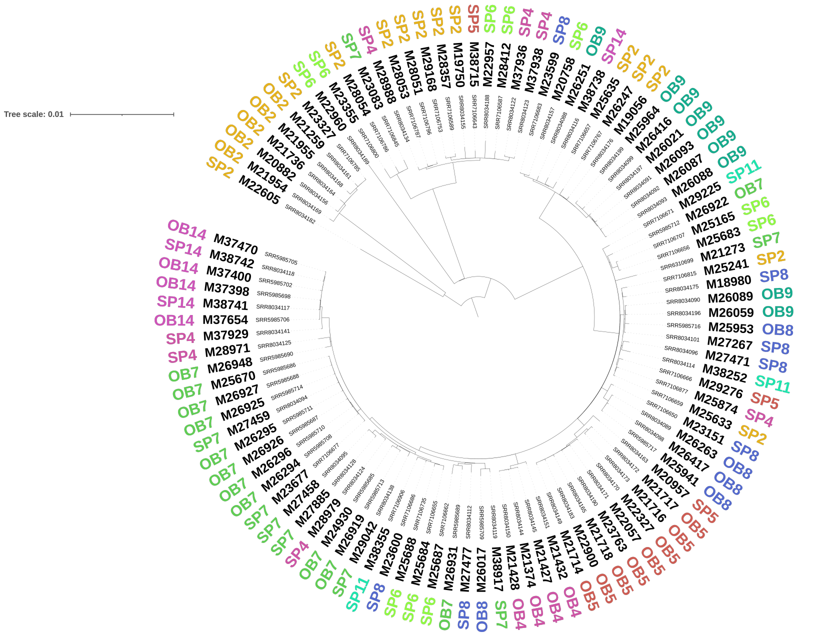


Figure 3.25: pgSNP on segroup C from *Neisseria meningitidis* outbreak. The inner ring corresponds to sample name annotation as described on the paper; outer ring corresponds to outbreaks and sporadic annotation.

the interest of using the information associated to the pangenome for outbreak investigation.

3.4 Discussion

3.4.1 How to improve pgSNP

3.4.1.1 How to improve the pangenome reference

In this section of the PhD thesis, I developed a reference pangenome with the objective of inferring a phylogenomic tree. This idea was always kept in mind when developing this method, and therefore the reference pangenome was optimised as well as possible with this objective. The reference pangenome represents a biological reality of the variability contained in a dataset. We did not create a pangenome with consensus contigs, because we wanted to ensure that raw-reads would align to the contigs. A consensus could distort the reference pangenome, and lead to a potential loss of reads. In contrast, the management of repeated regions remains difficult, and thus advanced management of these repeats would be needed to make a consensus pangenome.

The reference pangenome described as a genomic sequence is an innovative method, which has already been explored by C. Jandrasits a PhD student from the Robert Koch Institute (Berlin, GER). Indeed, Christine Jandrasits developed a pangenome reference method called Seq-seq-pan [381], and then a pairwise variant calculation called PANPASCO [382]. Seq-seq-pan is a method to build a pangenome reference which is similar to pgSNP, but using Mauve aligner instead of BLAST, and which creates one reference pangenome constituted of a unique contig. In further development, it could be interesting to compare results of the present PhD thesis with outcomes of the other method. However, the Seq-seq-pan reference pangenome constituted of a unique contig, would increase the computing time and be based on a unique nucleotide substitution model in comparison with the pgSNP pipeline.

These aspects were not taken into account by the authors [382], because they only used their reference pangenome to detect SNPs. Accessory and core SNPs are mixed because their goal was to add information to compare very close samples. Also, the authors started with the analysis based on SNP differences rather than phylogenomic inferences. In the current PhD thesis, we answer to a more general problematic: taking into account accessory genome to delineate homogeneous isolates but also try to find if some sporadic samples are linked to outbreak samples to support official investigations. Therefore, we considered accessory SNPs as well as their different evolutionary processes depending on the contig in which these SNPs are found.

A new trending way to analyze the pangenome is to use a pangenome graph as a reference. The idea of a pangenome graph was previously proposed, but still at this time, pangenome graph can hardly be used as reference genome for mapping. For example, Minigraph [383] creates graphs that detect SVs (structural variants) within samples, but depends on a reference genome to build the graph, as described in many studies [384, 385, 386, 387, 388, 389]. DNA information will be missed and will not be taken into account into the phylogenomic reconstruction. We did not want to rely on a reference genome to build our reference pangenome, so we did not use those kinds of methods.

Most of the pangenome graph developed nowadays are used for variant genotyping, or to overview large dataset of genomes.These methods are very recent and are still in progress, but show very encouraging results. The construction of pgSNP reference pangenome is linear, but a pangenome graph may improve the raw-reads alignment and also the variant calling detection. Pandora [102] was published at the beginning of my third year, and was the first method to build graphs without a reference genome, and detect SNPs from core and accessory, genes and intergenic regions. Pandora is efficient with distantly related samples, where a usual reference genome can not represent the whole diversity. Despite the limitation, Pandora is the most advanced tool for pangenomic graph analysis. We are currently in the preliminary stage of these new methods. Pangenome graphs really started to be implemented in 2020, so we can anticipate a new wave of development for the next 5-10 years that would be focused on pangenomes graphs.

Our method can build a reference pangenome without using a predefined reference genome, and we wanted to use a method which would be able to work with already implemented downstream analysis, due to the conceptual simplicity of identifying variants with a linear reference genome, and the mature tool chains developed during decades. The reference pangenome in fasta format is right in the middle of new developments nowadays. Graphs have to build their own methods to detect SNPs which were not developed at the beginning of my thesis.

Finally, the selected parameters were the optimal combination for the lowest entropy, the high supplementary number of aligned reads and also the highest alignment length. While this combination is display a good mapping quality (3.3.2.2), the pangenome built with a 95% identity could insert sequencing error or assembly error. In pgSNP pipeline, the pangenome contigs that do not harbor sufficient aligned reads are discarded from the multiple-alignments. In addition, as the phylogenomic tree needs to have at least 4 samples in the alignment, contigs which do not meet this requirements are also discarded. Assembly errors would generate unique contigs, and thus be discarded in our analysis. At the end, a high number of contigs can be not taken into account in the downstream analysis, but it ensures us that the error rate is low, and the results are robust even if the assembly may present errors.

To improve the mapping quality, we should have taken the best entropy parameter. As the number of samples by alignment is the sticking point, the contig alignment length has to be taken into account. If the method proposed a solution about the integration of unique DNA or DNA contained in less than four samples, the lowest entropy score would have be retained. At this stage of development, it is the best compromise that we have found.

3.4.1.2 Variant detection

Detecting variant on accessory genome is easily done by Snippy, GATK or whatever variant caller if reads of coverage is sufficient for variant calling. In addition, the variant calling only identifies variants to then recreate the genomics alignments correctly. Variants uniquely identified due to variation into the reference are consequently not taken into account in the phylogenomic inference.

In this PhD thesis, we only investigate SNPs because of the different level of complexity we had to control in the pipeline. SNP and presence/absence of contigs was already difficult to handle because of the issues related to missing data management during phylogenomic inferences. Nevertheless, an option about the identification of InDel was added. InDel is still not well managed by phylogenomic methods and may have different evolutionary speed [82, 83, 84], so we did not include it during phylogenomic inferences. But adding them and analyse the impact of InDel is the next step to improve the resolution.

3.4.1.3 How to improve the substrees

We designed the pipeline with subtrees due to a main problematic in phylogenomic inference: phylogenomic tree methods have difficulties in handling missing data. Some phylogenomic tree methods count gaps as a "5th base", and eliminate columns in the alignment when missing data are too much present. In the context of handling missing data, supermatrices methods have been developed, and are now incorporated into phylogenomic tree tools like IQ-TREE [112]. But these methods are time consuming, and it also brings a new reflection on the management of the different evolution parts of the genome (presented in Annexe section 6.1.2). Indeed, *Salmonella* is prone to genomics content exchange through horizontal transfers, and incorporates several new elements like plasmids, phages, SPI, ICE and other accessory elements (described in section 2.3.5). Instead of aligning DNA segments together to infer a phylogenomic tree, the reference pangenomic contigs are handled in separate alignments, and a phylogenomic tree is built for each alignment. With this method, two parts of the accessory genome are taken into account :

- First, the accessory can have a different evolutionary model and evolutionary rate. As each tree is computed independently, the evolution rate can be different between two trees if their genetic content is different.

- Second, each accessory genome is found in some strains, but not in all. Using a tree for each alignment, only samples which have the same accessory genome will be found in the subtrees.

With this method, subtrees can be analysed independently and ease the analysis of accessory genomes. Also, accessory trees can be easily be analysed to search for presence and absence of DNA segments or SNPs on elements of interest. Merging all contig alignment into one super-alignment was not possible, because of the difficulties to connect all contigs and clusters them into one single alignment. Here, by separating all contigs alignments, we are in a divide-to-conquer strategy which was used for a long time on large datasets, before maximum likelihood (ML) phylogeny methods were able to find solutions for large datasets in a timely manner or other ressource-wise.

A great innovation in this pipeline is the use of subtrees and then a supertree method to infer a phylogenomic tree that would take into account core and accessory genome at the same time. This method is robust, as demonstrated previously, but depends a lot on the contig selection during reference pangenome building. In addition, all substrees are taken into account in fastRFS with the same weight. This method works well given the concordance of the results with epidemiological data, and without knowledge about contig exactness, it would not be wiser to give different weight to subtrees formed by contigs. One of the most promising developments would focus on contig cutting to improve the resolution and to more robustly define the evolutionary speed and evolutionary model for each DNA segment [390, 391, 392].

Methods which derive genomes into accessory segments are often based on a multi-alignment with windows to measure the impact of a DNA content on another window [391, 392, 393]. We wanted to build a method independent of a multi-alignment and keep the ability to analyse the genome content (% GC for example) and create a windows where the length varies according to the nucleotide content. This might reveal windows with diverse genetic compositions, which might indicate a different evolutionary process, such as recombinations or horizontal genes transfer. Unfortunately, no published method which fit this problematic was found. Alternatively, we might attempt establishing a new window by first aligning the readings to the reference and then cutting the alignments when we notice a region with a reduced depth of coverage compared to the prior window. This method was not a priority because the pgSNP outcomes were most of the time consistent with the epidemiological data. Nevertheless, concerning the contigs from plasmids or phages, it would have be judicious to split these related contigs because these elements are known to be impacted by recombination events which bring together DNA fragments with theirs own evolution history and may integrate into the chromosome [111, 394]. A script was developed which cut a contig window based on the presence or absence of a contig, and the percentage of SNP detected in the window. For example, if a window has a percentage of SNP higher than another window, it could mean that this window is not subject to the same selection pressure, thus resulting in identification of a recombination event or another foreign gene. Nevertheless, because of the lack of time, this method was not properly investigated. I consider that this step is the most important step to improve in a near future. This development would also solve the issue observed in the *Neisseria meningitidis* dataset when outbreaks were split due to accessory genome and the creation of long intermediate branch and final leafs in the phylogenomic tree.

3.4.1.4 How to improve the branch length

Another problematic step in pgSNP was the branch lengths. Supertree methods do not calculate branch lengths, so we had to find a way to add this important information into the tree. Some papers propose to calculate branch lengths based on pairwise-distance, or derive the branch length from a concatenated alignment [114].The fact that ASTRAL employs coalescence units allows computing of some branch lengths, but it does not calculate the final branch lengths, which is crucial for understanding the divergence of the strains. We choose to use ERaBLE since it addresses our problem. ERaBLE is based on distance matrix from all subtrees, so individual genetic distance can not be taken into account. However, ERaBLE displays some negative branch lengths, which is a unsolved issue. The authors advised to delete negative branch lengths (as it is advised for negative branch lengths from Neighbour Joining trees), but this solution is not the most suitable for our dataset. The use of partitioning data in the phylogenomic inference was an idea developed to counter this branch length problem. But the computation time is too expensive to invest more time in this method (discussed in Annexe section 6.1.2).

Even though it is imperfect, the only option we now have is the one in place. To get a higher resolution and also take into account the unique accessory fragments, we could develop a ML-like method which can calculate the differences from alignments and overlay the results on the pangenome tree. Nevertheless, we would encounter once again difficulties from missing data.

3.4.1.5 Perspectives of pgSNP

This study advances pangenome analysis, and can be extended on different pathways : a valuation of tools described here with several benchmarks ; a methodological development on evolutionary events such as InDels, but also genome fraction (integration of plasmid, of genes, prophage etc) ; suggest a reference pangenome for all serovars, in order to improve analysis resolution and develop a common method, and publishing variants instead of sequences.

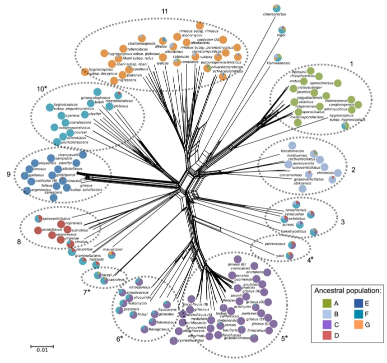


Figure 3.26: A phylogenomic network for *Streptomyces* generated using SplitsTree to investigate recombination events. Figure from [397]

Establishing a pangenome reference database could be an interesting development to highlight the current state of knowledge of strain sequences in public databases. In order to develop a pangenome reference database, a selection of the most analysed samples in genomics is required. Then, a pangenome could be built on all the reference genome existing for a species, and this reference would be available to biologists and bioinformaticians to perform their research with an increased resolution.

Finally, these references could be used in other downstream analyses, for example accessory analysis to quickly target phages, plasmids or other accessory elements in order to better understand the genomics variability of a strain or to create markers. Otherwise, instead of publishing pangenome, a pangenome variant database could be also developed. It would work like the pangenome database, but with pangenome variants to ensure the anonymity of the provenance of the strains. Variants databases already exists for SARS-CoV [395] in the idea of tracking the evolution of a pathogen, and therefore it would be possible to repeat the same idea but with pangenomic variants.

3.4.2 Is a phylogenomic tree the best view to understand genomics variability?

A great reflection was carried out during the thesis about phylogenomic trees. Most of the genomic pipelines developed nowadays depends on detecting variants them infer a phylogenomic tree to compare genomics samples between them, analyse their genomics evolution, and try to identify outbreaks. Adding accessory genome, we were not convinced that it was a good idea to mix all the data together and put them into 2D dimensions represented by a tree.

Because an accessory SNP does not have the same evolution pattern and mutation rate, the question arises whether to mix this data with coregenome data. We explored a bit 3D possibility, but the research subject should have been completely based on it, in addition to taking into account accessory genome. Split Tree [396] is a phylogenomic method considering that a unique tree is not enough to represent the variability of a dataset, especially when there is conflicting phylogenomic signal. Instead of a phylogenomic tree, the dataset is represented as a tree-like network, and it is used in genomics content analysis, or for example in recombination events investigation in a dataset, as represented in 3.26. This type of graph is an interesting method to explore, with the possibility of making accessory genome links and missing data.

We also explored graph-like representation, as described in 6.1.1, but the results were not fully consistent with epidemiological data when there were too much missing data. Finally, subtrees seem to be the easiest option to explore accessory genome, and supertree was the logical continuation. We could also try to use subtrees and create a graph which concatenates all distances, but we did not have time to explore this option.

Putting aside graph options, we thought of a statistical method in the earliest steps of the PhD. The goal was to develop a method that would analyse each SNPs and DNA content to seek correlation and difference between samples. This innovative method could give a correlation score between strains while taking into account certain original aspects of these parameters such as the rarity of SNPs or accessory contents. This score could also define if two samples are linked together, without the use of thresholds of pairwise mutation differences, but a threshold on the first order risk - i.e. calling two strains not linked when they truely are - based on the probability knowning from the databases the frequencies of each character common or different between 2 samples. This score could also be plotted as a graph for a visual representation. Also, this method could resolve partially the horizontal gene transfer problematic, as some non-parametric approaches are less impacted by recombination events than some phylogenomic method [79, 398]. As the implementation of new statistical measures appeared difficult and time-consuming, we had kept this idea in mind as a side project. This kind of method seems very promising by combining it with pangenomic graph, because at the time of my thesis, tools were not to available to represent the differences between strains based on graphs and SNPs identified from reference pangenome.

3.4.3 How to define an outbreak in genomics

When outbreak samples harbor differences from accessory genome, it becomes difficult to determine if the strains are well linked or not. In pgSNP, the question often asked is how significant these differences are that the strains are no longer related. For example, in section 3.3.4.3, we showed that two strains which were related in coregenome were distant in accessory genome. But, as the distance is also not very large compared to the dataset, it is an utopia to define a threshold. Also, the genomic difference between two isolates can be highly due to the presence or the absence of accessory genome, adding complexity to the threshold delimitation.

With pgSNP phylogenomic trees, we could see visually the clusters in the tree by using the topology. But, when outbreak samples have small variations in their accessory genome, how to take it into account to define epidemiological links between isolates? Also, if one isolate has a stranger element which induces a resistance or virulence factor that could explain an outbreak, does an isolate which does not possess this characteristic must be left out of the analysis?

Ten years ago, the cost of whole genome sequencing was too high to prioritize developments of SNP-based approaches [399]. But today, even when whole genome sequencing pipelines has been developed as a routine implementation [26, 245, 246], sub-typing and PFGE is still used to define outbreaks. Cg and wgMLST are also still widely used due to their ability to share sequence type between laboratories, although these methods have difficulty to segregate related samples of TMV [79] and are highly dependant of the *de novo* assembly quality of samples [67].

Defining a threshold in each method is an utopia given the heterogeneity of the serovars [79, 400]. While some serovar can be linked by a threshold of less than 5 SNPs according to other authors (monophasic variant of Typhimurium samples [80], *Salmonella* Dublin sample [401]), some others, like *Salmonella* Mbandaka can not be linked with this information. For example in a *Salmonella* Mbandaka outbreak study [314], the authors defined an outbreak of *Salmonella* Mbandaka human cases with SNP differences between 10 and 25 SNPs (Figure 3.27). When looking at another potential cluster with sporadic samples linked to outbreak samples, the SNP difference varies between 12 and 82. Besides, SNPs cutoff also depends on the genomics method used, which adds an additional problem.

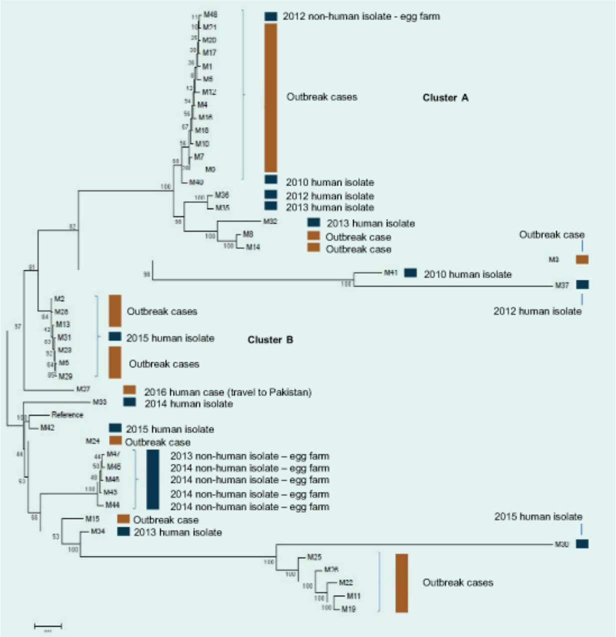


Figure 3.27: Phylogenomic tree generated from whole-genome SNPs for outbreak and non-outbreak cases of *S.* Mbandaka in New South Wales. From [314]

We wanted to develop a cut-off like method which could take into account the scarcity of some SNPs or accessory genome content to add more weight on subtrees in the final supertree reconstitution. But these kinds of methods are complex to implement, and the parameters settings would take a long time, so we did not considered it.

3.5 Conclusion

During this chapter, we were able to set up an innovative pipeline called pgSNP, which takes into account the accessory, coding and non-coding genome, and makes it possible to infer these results on a phylogenomic tree. This pipeline contributes to the identification of the variability in the accessory genome of different samples, to understand its prevalence and distribution, the persistence and the hazard in food safety, which is not detected and oblivious in coregenome analyses.

During this chapter, we display the advantages and the news results inferred by pgSNP. pgSNP is able to find consistent results with a coregenome SNPs approach, but also to add more resolution on phylogenomic tree analyses. In this study, we displayed news reconciliations due to the use of the reference pangenome in *Escherichia coli* dataset, but also news reconciliations due to the use of accessory genome in *Neisseria meningitidis* dataset. In *Salmonella* Typhimurium and monophasic variant of Typhimurium outbreaks, pgSNP was able to show the impact of the accessory genome on very similar strains, and the benefit of using this tool to further investigate the accessory genome.

Finally, we demonstrated the advantages and the limits of the pipeline, with suggestions for script improvements. We are very aware that pangenomic analyzes will greatly improve in the coming years, and we believe that this study will have enabled progress in this area.

Chapter 4

Analysing the genomic diversity of *Salmonella* in the context of food safety

4.1 *Salmonella* issues in milk and pork food sectors

In this chapter, I will present the genomic analysis of the two main serovars investigated in my thesis : *Salmonella* Mbandaka, *Salmonella* Typhimurium and its monophasic variant. A study of *Salmonella* Dublin will also be presented in this chapter, with the objective of understanding the dissemination of this serovar in bovine strains from two French regions. Other serovar specific questions are presented in 2.4. Although the thesis project is structured in independent sections dedicated to each serovar specific issues (e.g. sampling, history and organization of the production chain, etc.), the conceptual framework is however common: genomics for sanitary control, assessment of *Salmonella* biodiversity in an animal production sector, origin of contamination, selection of variants and the bioinformatics methods used.

First, I will present the selection of samples to answer each important serovar specific questions. For *Salmonella* Typhimurium and its monophasic variant, the dataset represents the geographical diversity of these serovars in pig and pork sectors. For *Salmonella* Mbandaka, the database includes strains of dairy origin from the north of France and another host (i.e. poultry). For *Salmonella* Dublin, the dataset should represent the diversity of strains detected in the cattle and dairy sector of two regions.

Concerning the pig and pork industry, *Salmonella* Typhimurium and its monophasic variant are the predominant serovars encountered in this sector. Although monophasic variants of *Salmonella* Typhimurium (TMV) are now more prevalent than *Salmonella* Typhimurium, they are still distinguished only based on phenotypic traits and therefore it is important to fully analyze this serovar at the genomic scale. The objectives of this section are to understand the diversity of this serovar in the pig and pork industry, from pig herds to the finished product, and also its dissemination in France by comparing its genomics diversity to its geographical diversity. Finally, the genomic contents will be analysed comparing French TMV strains to worldwide strains in order to understand the links that may exist between the diversity of strains from various countries.

For the bovine industry, I will first characterize the biodiversity of *Salmonella* Mbandaka in

	Subtyping			**SeqSero2 prediction**		
Project	Typhimurium	TMV	N/A	Typhimurium	TMV	N/A
Ifip	45	136	3	28	143	13
GAMER	15	123	0	14	124	0

Table 4.1: Summary of all available data.

	Year		**Matrices**		
	<2016	$\geq$ 2016	Slaughterhouses	Pig herds	Processing plants
Samples	73	249	182	17	123

Table 4.2: Metadata summary of the dataset (n=322).

the different reservoirs (i.e. environment, feed, herd, cow, milk, cheese). The capacity of adaptation to the host of *Salmonella* Mbandaka will be investigated in the poultry sector of the north-western France. A complete genomic review of *Salmonella* Mbandaka will be presented, as we have little global knowledge in genomics for this serovar.

Finally, I will present results of the track of environmental and geographical persistence of *Salmonella* Dublin in two regions. As this serovar is well-studied compared to *Salmonella* Mbandaka, I will rely on this work to then compare with the genomic diversity of *Salmonella* Mbandaka.

4.2 Material and Methods

4.2.1 *Salmonella* Typhimurium and its monophasic variant dataset

4.2.1.1 Slaughterhouse and processing plant dataset

In order to investigate the genomic diversity of these *Salmonella* Typhimurium and its monophasic variant in France, a dataset has been built using available data from swine. This dataset comes from two sources. The first one was collected by the IFIP (The French Pork and Pig Institute) and is constituted of samples from internal collaborations (i.e. samples collected from partner companies) and a monitoring plan managed by the DGAl (Direction Générale de l'alimentation) between 2016 and 2018. The second source is an in-house sequencing database in ANSES (GAMeRDB) where paired-end reads from different projects were deposited for pre-processing and further analyses. Metadata of these sequences come from Acteolab (ACTEOLab-Salmonella), a national network for the epidemiological surveillance of *Salmonella* strains of non-human origin.

185 *Salmonella* Typhimurium and its monophasic variant genomic sequences were provided by IFIP and were isolated from swine strains. One sample was predicted as *Salmonella* Derby, thus withdrawn from the analysis. 248 *Salmonella* Typhimurium and its monophasic variant genomic sequences were isolated in France and came from GAMeRDB and Acteolab. Looking more precisely, 138 samples isolated from swine data were selected for the study. In total, 322 samples were analysed. These dataset collections are described in Table 4.1.

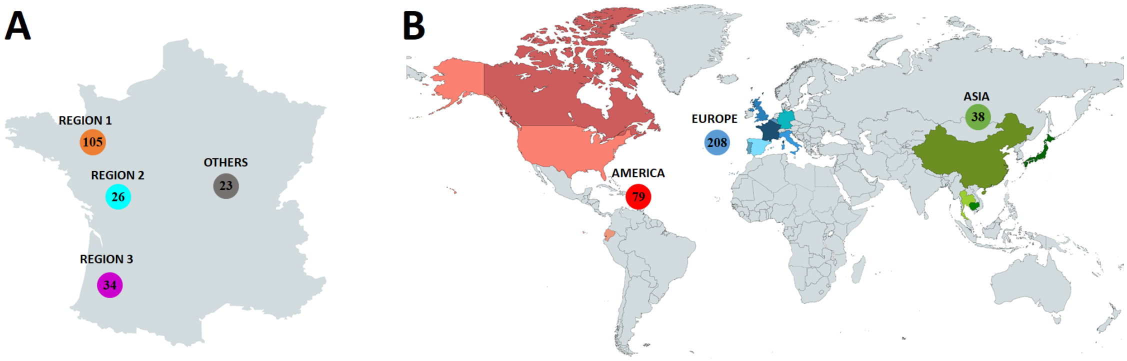


Figure 4.1: Dataset selection of *Salmonella* Typhimurium and its monophasic variant. A : Data selected from France from pig herds. B : Data selected from the world

	Extracted	Assembled	Total
IFIP	138	138	138
ANSES	51	50	50

Table 4.3: Dataset selection of *Salmonella* Typhimurium and its monophasic variant from different partners of this study

4.2.1.2 Geographical dataset

To study in detail the genomic diversity of *Salmonella* Typhimurium and its monophasic variant, I collected genomes harboring information related to departments of herds. The selection was delicate, as the western French regions produced roughly 78% of the French pig production [402]. We selected samples from 3 regions in order to be able to analyse the genomic variability across geographical origin. The region 1 corresponded to northern-west of France (Brittany region) where most of pigs are produced. The region 2 corresponded to the mid-west of France (Pays-de-la-Loire) which corresponded to the 2nd most producing region. The region 3 corresponded to the south west of France. Other samples from other regions were labelled as "others" in the corresponded study. Samples were collected from different sources: IFIP, ANSES Maisons-Alfort (i.e. internal network and *Salmonella* network) and ANSES Fougères (i.e. monitoring plan).

In brief, 138 samples from IFIP had the department of origin of a pig found in a slaughterhouse. As most samples have been isolated from region 1, 51 samples from region 2 and region 3 have been added to the dataset by ANSES. While all samples have been isolated, sequenced and assembled without any contamination or quality problems, 1 sample has been discarded due to metadata incongruences. A total of 50 isolates has been added to the dataset.

In total, 188 samples were assembled and have passed quality checks to be used in this study (Table 4.3). The amount of samples selected for each region is illustrated in Figure 4.1-A.

4.2.1.3 Worldwide dataset

With the objective to access the global diversity, I also selected TMV samples from other countries available in Enterobase and published studies [286]. Only ST 34 samples from swine were selected in this study. Using data previously analysed, I selected 132 TMV samples from France where *in vivo* and *in vitro* investigations identified monophasic variant of Typhimurium

Europe						
Belgium	**France**	**Germany**	**Italy**	**Portugal**	**Spain**	**UK**
3	132	24	29	1	1	18
Asia						
Cambodia	**China**	**Japan**	**Thailand**			
2	1	30	5			
America						
Canada	**Ecuador**	**USA**				
23	5	51				

Table 4.4: Description of samples selected from other countries using Enterobase

and ST 34.

Concerning the other countries, 38 samples have been selected from the COMPARE project [286] where all samples have been isolated from pigs from Europeans countries. Using Enterobase, only ST 34 samples were selected. From 21,191 available samples, I retrieved those presenting an expected SeqSero2 identification (i.e. TMV) isolated after 2012 with available Illumina raw reads. Finally, 964 samples were selected including 858 samples isolated in USA. Further criteria were added from USA samples, such as the subtyping prediction and the geographical metadata available to downsize the dataset to 50 samples. Finally, with 5 samples discarded due to poor raw reads quality or poor assembly, only 51 samples have been selected from USA. All samples used from other countries are described in Table 4.4 and plotted in Figure 4.1-B. At last, 38 samples were selected from study [286], 132 from France, and 155 from Enterobase for a total of 325 genomes.

4.2.2 *Salmonella* Mbandaka datasets

4.2.2.1 Bovine data

S. Mbandaka is highly prevalent in north-western France, while there is no explanation for its persistence in this region (as described in section 2.3.8.1 and Figure 2.18). To first describe the diversity of *S.* Mbandaka, a collection of *S.* Mbandaka isolates from bovine was collected from this region of interest. In total, 148 samples have been collected from Normandy, thanks to ACTALIA and industrial partners (FGIE) and Caen University. From the 148 samples, 143 have been sequenced with Illumina NextSeq technology at ICM (Institute for Brain and Spinal Cord) and produced proper paired-end raw reads that could be assembled with a SPAdes-based pipeline [70, 345] which is described in section 4.2.4. All assemblies were subjected to a quality aassessment as described in the section 4.2.5. Among these samples, three were identified as serovar other than Mbandaka and discarded from the downstream analysis. All other *Salmonella* presents the same GC% (52.10%), a alignment length superior to 4.6Mb (*S.* Mbandaka reference is 4.8Mb, *S.* reference selected here is described below), no more than 500kb of miss assembled bases, and InDels per 100kb is around 2 for all genomes. Number of genomes from each source is described in 4.5. Finally, 140 proper draft genomes from Normandy were available for this study.

As *Salmonella* Mbandaka is monitored in the dairy sector in this study, we selected samples presenting metadata related to bovine (i.e. faeces, manure) and dairy origins (i.e. milk, milk filter). Also, samples from cheese and feed products were added to the dataset to monitor whether contamination was due to matrix-specific clusters or occurred through all the produc-

	FGIE	Caen University
With metadata	100	48
Assembled	99	44
Mbandaka serovar	97	43

Table 4.5: Summary of data source

Cow and its environment	Milk	Cheese	Feed products
55	70	6	9

Table 4.6: Summary of data matrices

tion chain. Strains from feed products were also added to check whether contamination in farms occurred. The summary of the data matrices is displayed in Table 4.6.

4.2.2.2 Poultry data

We collected 170 poultry samples from ANSES and collaborators (ANSES HQPAP, ANSES Fougère and *Salmonella* Network [43]) between 2016 and 2020. Unlike cattle, the major production basin of poultry is not Normandy. I selected the maximum of poultry isolates from north-western France (70%) as described described in Table 4.7. In a nutshell, 51 poultry strains were collected elsewhere in France such as in Hauts-de-France, Auvergne-Rhône-Alpes, Grand Est or Occitanie regions. One strain of duck and five strains of turkey isolated in north-western France were also included in the panel.

All samples has been serotyped through glass slide agglutination, according to the White-Kauffmann-Le Minor scheme [5] as described previously. Sequencing, assembly and quality criteria of selected samples are the same that those described in the section 4.2.2.1. One sample had incomplete metadata and had been taken out from the analysis. From the 168 poultry samples correctly assembled, 1 genome was predicted as *Salmonella* Tennessee, and 4 genomes belonged to the ST3016. As the prevalent ST from our dataset was ST413, we withdrew these samples from our study. In total, we have gathered 164 poultry samples. One strain *Salmonella* Mbandaka 100727 (NCBI ID SRR6860551) belonging to the ST506 has been added to the dataset to root the phylogenomic tree.

4.2.2.3 Reference selection

Looking at a previously published genomic analysis of *Salmonella* Mbandaka, two reference genomes are commonly used : CP022489 and CP019183 [403]. To select the best reference genome, all raw reads from *Salmonella* Mbandaka bovine (n=140) has been mapped against both references with BWA [328]. Then, the breadth of coverage of genomes has been calculated against the reference with Samtools [338]. Breadth of coverage corresponds to the proportion of nucleotides from the aligned reads according to the reference sequence length. The mean breadth coverage of CP022489 was 98%, while the mean breadth coverage of CP019183 was

Pays-de-la-Loire	Bretagne	Normandie	Others
47	34	32	51

Table 4.7: Summary of region of isolation of poultry dataset

89% (Figure 7.14). This results showed that the CP022489 reference fits better to the *S.* Mbandaka raw reads compared to CP019183. For downstream analysis, CP022489 will be used as single reference.

4.2.2.4 Wild bird data

To explore the hypothesis of a possible cross-contamination between wild and farm animals, we extracted, from the open access *Salmonella* Enterobase database [62], 2,465 raw reads (on October 2021) whose serovar was confirmed as *S.* Mbandaka, the MLST profile of ST413 to keep the same profile for all strains of the dataset, and available epidemiological information such as host and country of isolation. Among these reads, 42 were retained because they were isolated from wild animals with 10 genomes from "Avian", 2 from "Canine", 2 from "Marsupial", 4 from "Reptile" and 23 "No-determined", respectively. Among these 42 genomes, no one was isolated in French or Europe. We still decided to go through the analysis, as some wild birds might be responsible of inter-continent contamination [404].

Finally, we chose 9 of the 10 genomes from strains isolated from wild birds (i.e. "Avian"). One genome was excluded due to read errors. The selected 9 genomes came from the American East coast: 6 from the United States, 2 from Canada and 1 from Mexico.

4.2.3 *Salmonella* Dublin dataset

During this thesis, I also had the opportunity to work on *Salmonella* Dublin serovar which is extremely prevalent in the bovine sector. In this work I performed a retrospective study of *Salmonella* Dublin outbreak that took place in France between 2015 and 2017. Two regions producing a raw milk cheese linked to human cases were specifically targeted. I investigated the diversity and the circulation of *Salmonella* Dublin strains in 2 regions in cow herd, milk and cheese environments with the aim of understanding the routes of contamination.

This study required a collaboration between four private and public laboratories to collect 2,249 strains within the time range of the outbreak. Beyond the results provided by the study, the objectives of the project was to demonstrate that WGS can provide more insight during outbreak investigations for the actors of this sector. In order to demonstrate the advantage of WGS, additional samples from linked cases of salmonellosis under local epidemiological and microbiological investigations were added in the study, called selection A,C,D,E (n=104). Description of the targeted selection samples is available in Table4.8.

Selections	**A**	**C**	**D**	**E**
Description	Single dairy farm where strong clinical signs of salmonellosis cases in cows were observed over the years	Contamination from cattle to cheese from limited geographic area	Contamination from milk in short period of time in restricted geographic area	Samples from cattle on different farms in restricted area

Table 4.8: Summary of targeted selection metadata

Collection of metadata and samples are described more precisely in the article presented in the

Data origin		Isolates	Metadata	Gower	Assembly ok	QC ok
This study	**Random**	2249	2101	398	331	315
	Targeted	104	104	104	70	70
From ANSES		77	77	77	59	58
From Pasteur		371	109	109	109	37
Tally		**2697**	**2478**	**917**	**569**	**480**

Table 4.9: Contingency table of sample sources, filtering and selection. RND: Random. Targeted corresponds to SELEC A + SELEC C + SELEC D + SELEC E. An exhaustive filtering description is presented in methods.

section 4.5. Overall, a collection of 480 samples selected from private and public institutes is displayed in Table 4.9. The selection of samples for the study was performed based on a subsampling algorithm based on the Gower distance [405] and 2,249 strains collected at the time of the epidemic outbreak. In addition, ANSES provided 77 samples from its collections of strains and the Pasteur Institute provided humans samples [301] in order to bring genomic background to the outbreak, and also propose a efficient surveillance plan that would cover the genomics diversity of *Salmonella* Dublin in the two incriminated regions.

4.2.4 Assembly

The ARTwork pipeline filtered reads based on quality control and normalization by estimating the coverage of reads with bbmap [325], normalized the reads with bbnorm [406] and controled the quality of the reads with fastqc (https://github.com/s-andrews/FastQC). Reads were then trimmed by Trimmomatic [326] to remove technical sequences such as adapters or polymerase chain reaction primers. Contigs were produced by Spades [70] which perform a *de novo* assembly based on a de Bruijn Graph. Sequence type (ST) of isolates were detected by MSLT based on PubMLST scheme (https://github.com/tseemann/mlst). Scaffolding was performed with Medusa [407] using the closest reference detected with Mash [408]. Finally, gap filling was done with GapCloser [409] and contigs were trimmed with Biopython [410].

4.2.5 Quality assessment

Quality assessment has been described in the section 3.2.2.
Rules for analysis was carried out with QUAST are described as:

- more than 1,000,000 assembled bases unaligned to the reference
- less than 4,000,000 assembled bases aligned to the reference
- more than 2 InDels per 100kbp
- less than 80% of assembled bases with 30X coverage
- absence of the genome fraction estimation computed by QUAST
- assembly fragmented into more than 200 contigs
- contamination detected in the assembly

4.2.6 Characterization of serovars and sequence type

Sample serotyping was performed *in silico* based on the assembled genomes using SeqSero2 [411] described in the section 3.2.2.

In addition, all genomes were characterized by *in silico* MLST using the 7 housekeeping gene sequences (*aroC*, *dnaN*, *hemD*, *hisD*, *purE*, *sucA*, and *thrA*) described in the PubMLST database [58]. The ST of each genome was obtained with MLST tseeman tool (described in 4.2.4).

4.2.7 Coregenome analysis and phylogenomic inference

Coregenome analysis was performed by the iVARCall2 workflow described in the section 3.2.3. Recombination tracks were identified using ClonalFrameML [110] with the following parameters set to true: -em, -guess_initial_m, -use_incompatible_sites, -reconstruct_invariant_sites, -output_filtered. The parameter -emsim was set to 20 and other parameters were kept to their default values.
Phylogenomic tree was inferred by IQTREE [107]. IQTree was subsequently used with -m TEST model selection on alignments with and excluding variants from homologous recombination detected by ClonalFrameML [110]. Robustness was tested with IQTree parameters -alrt 1000 and -bb 1000.

4.2.8 Identification of virulence factors and resistance genes

Genomes were screened for the presence/absence of genes mediating resistance and virulence using Abricate (https://github.com/tseemann/abricate). The Blast-based Abricate application was used in combination with the VFDB [412] database available from the Institute of Pathogen Biology, the MEGAResV2 database [413], the ResFinder [414] and SPIFinder [415] databases available at the Center for Genomic Epidemiology (CGE) (Denmark). The Abricate outputs show only the genes found on at least one genome of the analysed panel. The threshold was set at a 90% identity over at least 3/5 of the length of the gene or genomic region.

4.2.9 Plasmid identification

Assembly were further analysed to characterise the presence of mobile genetic element (MGE). MOB-suite tool [416] was used to identify plasmids. MOB-suite identifies plasmids contents in genomes by predicting the mobility that is based on the presence of relaxase, mate-pair formation and oriT sequences. Putative plasmids were blasted against the NCBI nucleotide archive to identify the closest plasmid neighbor.

4.2.10 Pan-genes analysis

Genes content detected in the core and accessory genome was performed by Panaroo [100]. Panaroo is a graph-based pangenome clustering tool that is able to account for many of the sources of error introduced during the annotation of prokaryotic genome assemblies. Using GFF (General feature format) files from each genomes produced by Prokka [417], Panaroo detect and clusters genes and produces a pangenome (i.e. statistics about the gene content and alignment of genes). The identity threshold for the gene clustering was increased to 90% to minimize divergence between targeted markers.

4.2.11 Markers exploration

In order to identify robust combinations of phenotype specific markers, I then developed a Python3 tool called MarkerFindr. MarkerFindr calculates the combination of maximum 3 genes or variants with the best discrimination accuracy score according to a phenotypic criterion, so called the host origin in the present study. The two MarkerFindr input files are a file with the list of genes or variants and a file compiled with the corresponding phenotype. MarkerFindr output compiles the best combination of maximum 3 genes or variants found according to associated accuracy scores: (TP + TN) / (TP + TN + FP + FN) with TP corresponding to "true positive result", TN to "true negative", FP to "false positive" and FN to "false negative".

4.3 Results: Characterisation of the *Salmonella* Typhimurium and its monophasic variant geographical diversity in the pig and pork production

In this section of the thesis, we will characterize the diversity of *Salmonella* Typhimurium and its monophasic variant in the pig and pork sector in France. Using these results, we will try to understand the link between the genomic diversity of *Salmonella* Typhimurium and its monophasic variant and geographical distribution of farms. Finally, we will characterize the diversity of French strains compared to worldwide strains, and compare the diversity with the aim of producing useful tools for monitoring the dissemination of TMV strains.

4.3.1 Depicting the genomic diversity of *S.* Typhimurium and its monophasic variant

To decipher the diversity, we used a coregenome SNP approach using iVARCall2 [345] with *Salmonella* LT2 reference on 322 samples (section 4.2.1.1), as described in 3.2.3. To measure the impact of recombination in this dataset, we investigated homologous recombination events and inferred two trees : one using all positions from isolate alignments, and one with variants from homologous recombination events excluded from alignments, using ClonalFrameML [110] as previously described 4.2.7.

In total, 37 homologous recombination events have been detected, where 24 were located in leaves, and 13 in internal nodes. More homologous recombination events have been detected on this dataset compared to *Salmonella* Dublin (described in section 4.5.2 and *Salmonella* Mbandaka dataset (described in section 4.4.1). Looking at the samples impacted directly by homologous recombination events, 9 were harbored by 15 *Salmonella* Typhimurium genomes, and 9 by TMV genomes. In internal nodes, 1 homologous recombination event was located in the node splitting *Salmonella* Typhimurium and the TMV samples, and another on the node splitting TMV samples and all other genomes at the east of the tree (Figure 4.2). Other homologous events in internal nodes were located in *Salmonella* Typhimurium internal nodes (4) and TMV internal nodes (7). This result is interesting because it demonstrates different behaviors between *S.* Typhimurium and TMV strains. Even if the dataset size of TMV strains was larger than *S.* Typhimurium one, much more homologous recombination events were identified in TMV genomes than *Salmonella* Typhimurium genomes.

We inferred both phylogenomic trees using IQ-TREE [107]. Comparison of the two trees is displayed in Appendix Figure 7.3. After exclusion of the variants located in homologous recombination segments, 4,762 out of 4,893 SNPs remained. The topology of the coregenome

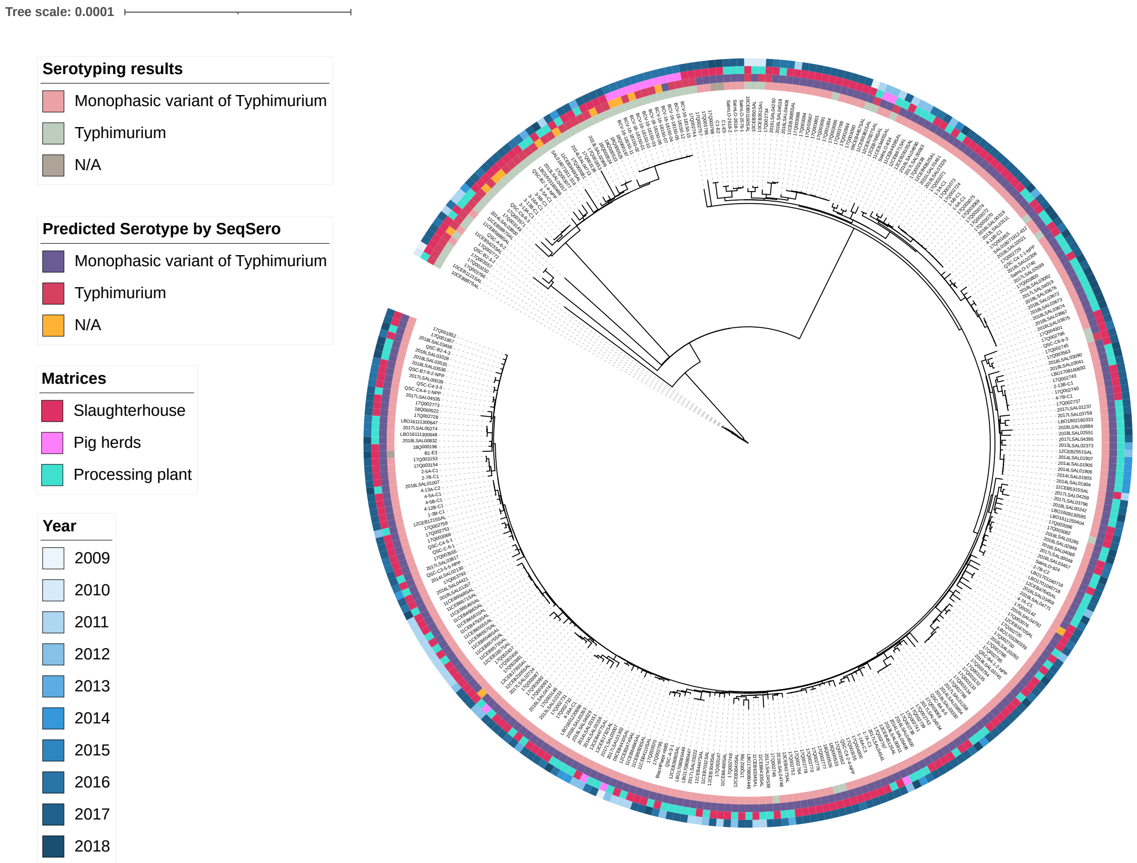


Figure 4.2: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Inner ring corresponds to serovar annotated by sub-typing. Second ring corresponds to the serovar predicted by SeqSero2. Third ring corresponds to the matrix of isolation. Outer ring corresponds to the year of isolation.

SNP-based maximum likelihood (ML) inference was slightly impacted by the removal of homologous recombination variants, with only few minor differences observed between nodes. Most of the differences relied on TMV samples (Appendix Figure 7.2). To characterize the diversity of this dataset, the phylogenomic tree of samples excluding homologous recombination variants has been inferred in Figure 4.2. The tree was obtained after convergence at 102 iterations with an optimal log-likehood of -6796629 and follows an evolutionary model TVM+F+I.

First, the global diversity of these serovars was analysed. Looking at Figure 4.2, *Salmonella* Typhimurium samples are clustered together at the top left of the tree, while TMV samples are clustered together on the right. *Salmonella* Typhimurium genomes at the top left of the tree were all annotated as ST 19 predicted by MLST (https://github.com/tseemann/mlst), while TMV samples were annotated as ST 34. Samples annotated as N/A were *Salmonella* isolates with ambiguous phase 2 flagellin (*FljB*) (i.e. doutable agglutination results).

There were few differences between SeqSero prediction and the serotyping identification, but most of the tree was clustered according to serovar. Six *Salmonella* Typhimurium according to SeqSero are clustered in the monophasic variant clade, against 15 *S.* Typhimurium according to serotype identification. The presence of TMV clustered with *Salmonella* Typhimurium raises different hypothesizes. The simplest hypothesis is that these strains were Typhimurium whose agglutination serotyping did not work well. False-positive reactions may occur as a result of weak, non-specific agglutination [418]. Another hypothesis is that these strains present some mutations on the genes or promoters of phase 2 flagellin which means that this phenotype does not appear during agglutination. Finally, the emergence of TMV is still under assumptions, as some authors described that TMV does not represent a single lineage, but rather a diversity of lineages originating from a few common ancestors [419]. Given the quite different length of the branches of these TMVs compared to the other ST19 *Salmonella* on the left of the tree, it is possible that these strains are indeed TMVs.

While the mean SNP difference of this dataset was 216 SNPs, it was shown that *Salmonella* Typhimurium had a higher diversity in terms of SNPs than TMV. Indeed, the mean SNPs difference between TMV genomes was 57 SNPs (Max = 117 SNPs), while the mean SNPs difference between *Salmonella* Typhimurium was 356 (Max = 780 SNPs). The high diversity of this dataset is due to *Salmonella* Typhimurium which displayed a higher divergence compared to TMV genomes which seemed to be more clonal. This observation allowed us to conclude that ST 19 had a higher diversity than ST 34, and therefore *Salmonella* Typhimurium strains had a higher diversity than TMV strains. This could be also explained by the fact that TMV strains emerged in the 1990s [283], and therefore may have had less time to diverge compared to *S.* Typhimurium.

Source or year of isolation did not explain or showcase *Salmonella* Typhimurium and its monophasic variant variability (Figure 4.2). Samples were most of the time clustered with other samples from the same sampling date and department, but source clade did not appear in phylogenomic trees. Since most of the samples originated from 2017, the molecular evolution analysis of strains by year was challenging due to the short time period. Sample position must be studied case-by-case to highlight contamination or a proximity of strains. For example, the monophasic variant found in a breeding in 2016 was found in different carcass samples in a nearby department in 2017, maybe due to a contamination at the slaughterhouse, during transport, or linked to the purchase of equipment between farms [420]. Looking more precisely, isolates from different sources were disseminated all around the tree. For example, a sample isolated in the slaughterhouse (17Q003071) had 5 SNPs differences with a sample isolated from pork meat in processing plant (2018LSAL03329) with 1 year difference. Otherwise, a sample isolated in a pig herds (12CEB4512SAL) has been linked to a sample detected in a pork meat preparation processing plant (2014LSAL05406) with only 8 SNPs differences (Supplementary figure 7.3). Similarly, 6 SNPs differences have been detected between an isolate from pig herds (12CEB1732SAL) and one from a pig carcass (11CEB4110SAL) in a slaughterhouse one year earlier. Last isolates described were in a small cluster including samples from different years (from 2009 to 2017) were clustered together with a low SNP difference, which suggested a continuous contamination between these pigs herds, the slaughterhouse and the processing plant. All these comparisons and dissemination of genomes with different sources all around the tree suggested that the strains contaminated the whole production chain, without showing any adaptation to a specific source.

4.3.2 Assessing the link between the geographical distribution of farms and the phylogenomic reconstruction

The section just above showed a continuous exchange of strains between sources over the years. Unfortunately, it was not possible to show whether the diversity was related to geography or not, due to the fact that most of the strains available came from slaughterhouses and few samples were isolated directly from the farm. Here, news samples were selected for this study to focus on the geographical diversity of *S.* Typhimurium and its monophasic variant.

4.3.2.1 Phylogenomic analysis revealed an undiversified dissemination of TMV in France

To explore the geographical diversity of *Salmonella* Typhimurium and TMV, 188 samples has been selected (section 4.2.1.2). Coregenome phylogenomic tree was inferred by IQ-TREE [107] using SNPs identified by iVARCall2 [345] pipeline, as described before 3.2.3. The tree followed an evolutionary model of TVM+F+I, and converged with a commensurate negative likelihood (-6794043.972) after 108 iterations. As displayed in Appendix 7.5, most of the nodes were supported by high bootstrap values. The tree was inferred on 4,247 SNPs. Using ClonalFrameML [110], 132 SNPs have been identified in 25 homologous recombination events in a total of 3,668 bp (14 in leaves, 11 in internal nodes). Two samples presented 2 homologous recombination events detected in their leaves. Looking more precisely at homologous recombination events detected in leaves, only 3 have been detected for monophasic variant of Typhimurium (TMV), while 11 have been detected in *Salmonella* Typhimurium leaves. On internal nodes, 4 homologous recombination events corresponded exclusively to TMV genomes. Typhimurium with longest branches (top mid of Figure 4.3) had multiple homologous recombination events (3): 1 event was detected including all *Salmonella* Typhimurium, and 1 other was detected in the branch that splitted *Salmonella* Typhimurium from monophasic variants.

Comparing the trees including and excluding homologous recombination events (model= TVM+F+I, log-likelihood=-6787102.637, iteration=102), the phylogenomic topology was slightly impacted (RF=114) (Appendix Figure 7.4) with some new reconciliations in very few branches. Overall, only few minor differences were observed between nodes.
The left of the produced tree is composed of *Salmonella* Typhimurium isolates, 3 TMV genomes and 2 genomes with undetermined serovar (4.3). As previously discussed in the section 4.3.1, all these genomes were predicted as ST 19. From the top right of the tree to the bottom left, most of the TMV samples were clustered together. This cluster was constituted of ST 34 samples, except for 2 samples which were predicted as ST 5239. The difference

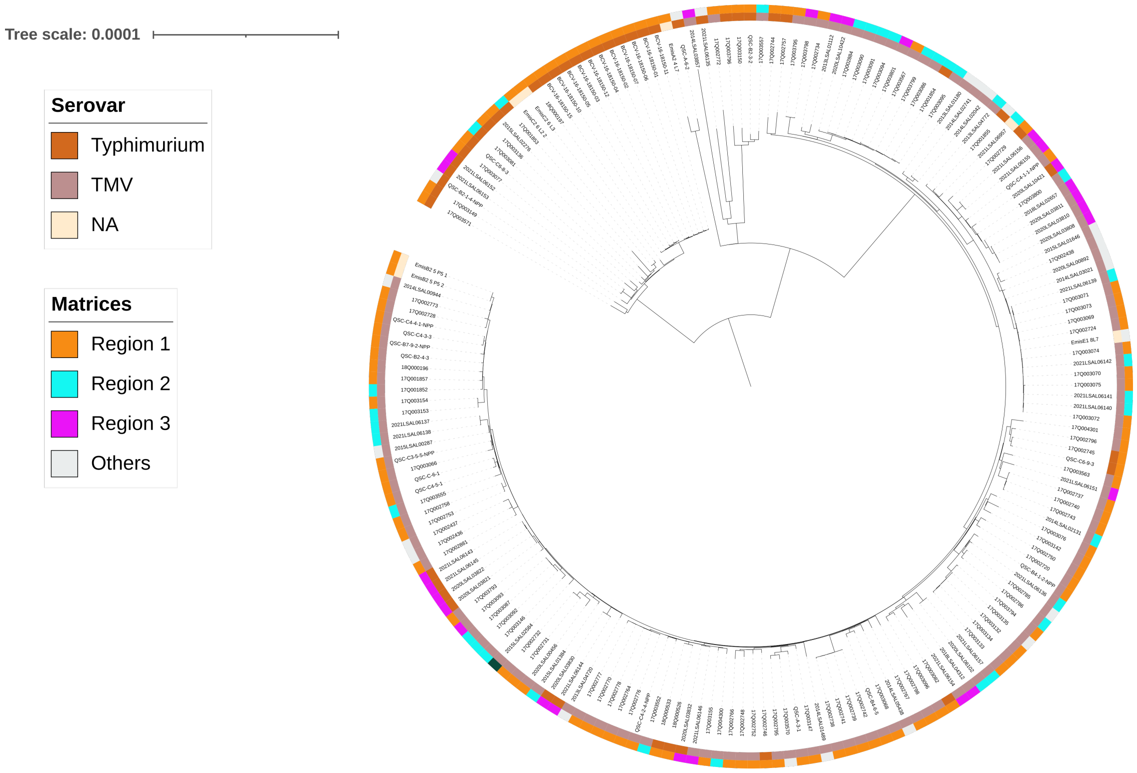


Figure 4.3: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Inner ring corresponds to serovar annotated by sub-typing. Outer ring corresponds to the region of isolation of the genome.

between ST 34 and ST 5239 is located on the loci *sucA*, where the allele profile predicted for ST 34 is 9, while the allele profile predicted for ST 5239 is 826. At the genomic level, these two strains have a longer common branch than the other ST 34s, but were clustered together at the south of the tree in the middle of the ST 34s. Overall, TMV clusters presented a lower diversity than *Salmonella* Typhimurium samples. This result was described previously in others studies [421, 422], where TMV genomes were described as less heterogeneous compared with *Salmonella* Typhimurium genomes.

Looking at the geographical metadata, we observed that the main genomic association came from the ST, and not from the geographical distance. The strains of TMV were clustered together without any geographical link. All regions were scattered around the tree. Region 1 was the most present with respect to the number of strains, but several strains from different regions were clustered together. This was the case of the genomes located at the north-east of the tree, where the difference between samples from 3 different regions is lower than 6 SNPs (3 SNPs between 17Q003567 from Region 1 and 17Q003094 from Region 2, 6 SNPs between 17Q003567 from Region 1 and 17Q003801 from Region 3).

Because the genomic difference between samples was really low, we hypothesized that only one clone was disseminated in France. Looking more precisely, the mean SNP difference of the 152 TMV samples from the north-east to the west of the tree was 64 SNPs. Topologically, a inner node splitted the 152 TMV into two clusters of 104 and 48 TMV samples, with an intra cluster mean of 49 and 51 SNPs, respectively. In addition, it seemed that TMV genomes did not present a geographic speciation, and therefore would not have region-specific environmental adaptation factors. To better understand if other adaptation factors exists between regions that would discriminate the diversity of TMV in pigs herds, we also analysed the genomic content of all strains.

4.3.2.2 Genomic analysis of *Salmonella* Typhimurium and TMV reveals a complete arsenal to adapt to the swine environment

4.3.2.2.1 Virulome analysis

In total, 126 virulence genes and 10 different SPIs have been detected in the 188 *Salmonella* Typhimurium and TMV dataset. The SPI-2, 3, 5, 9, 13, 14 have been identified in all isolates. SPI-1 has been detected in all samples except 2 (17Q003133 and 17Q004300). Looking more precisely at the virulence genes, the *sipD* gene involved in a type III secretion system has not been detected on 17Q003133, while all others genes coding for the type III secretion system SPI-1 were present, such as *inv*, *prg*, *org*, *sic*, *sip*, *ssp*, *spa*, *sopE*, *sopE2* and *sptP* [189]. In addition, genes coding for the type III secretion system SPI-1 were also present in 17Q004300. Because the most important genes encoding T3SS-1 were present, we assumed the presence of SPI-1 in these two samples. The lack of detection by SPIFinder can possibly be due to mutations localised in intergenic regions, and/or fragmented SPI sequences induced by poor assembly.

In the same way, the SPI-2 genes coding for the T3SS-2 were detected in all genomes, such as *ssaJ* to *ssaU* genes, *ssa*, *ssc* and *pip*. These genes are the main component of T3SS-2, and have been deeply analysed in *Salmonella* Typhimurium.

The *MgtB* and *MgtC* genes coding for Mg2+ transporter and membrane protein were present in all genomes from the dataset and located on the SPI-3 which has been described as a requirement for intramacrophage survival, virulence in mice and growth in low-Mg2+ media [195]. In addition, the *misL* gene described an extracellular matrix adhesin involved in intestinal colonization has also been identified in all *Salmonella* Typhimurium and TMV [423, 424].

Interestingly, the SPI-4 has been detected in all samples except 19. Looking at the figure 7.6, the absence of SPI-4 did not matched a pattern, except for 4 genomes clustered together at the bottom of the tree (17Q002738,17Q002741,17Q002739 and 17Q002742).

The *pipB* and *sopB* genes localised in the SPI-5 were detected in all genomes [198]. These genes code for effector proteins that alter host cell physiology and promote bacterial survival in host tissues, but are not necessary to invade the host.

The SPI-12 has been identified in 3 genomes (2014LSAL03857, 2021LSAL06139, EmisE1_8L7), in one *Salmonella* Typhimurium sample and two TMV samples. This SPI-12 has already been identified in *Salmonella* Typhimurium [200] and contributes to the bacterial virulence, but the fact that this SPI was rare in our dataset showed that this SPI is not mandatory for *Salmonella* Typhimurium invasion and pathogenicity.

Finally, the SPI-13, SPI-14 and SPI-16 carried virulence genes which were not identified by VFDB, most certainly because the genes are not present in the database. The SPI-13 and SPI-14 were identified in all genomes, however the SPI-16 was absent in 2 *Salmonella* Typhimurium samples (BCV-16-18150-12 and 17Q002757) not clustered together.

As detected previously with other *Salmonella* serovars, gene clusters of *csg* [425], *fim* [426], *shdA*, *bcf* [427] and *lpf* involved in Curli fibers and fimbriae have also been identified in all genomes. The *ratB* effector gene coding fimbriae has been identified in different serovars [428]. The *pef* [429] and *grvA* genes, coding for fimbriae and antivirulence gene [430] have been identified only in *Salmonella* Typhimurium samples (Appendix Figure 7.8). The *grvA* was annotated as antivirulence gene because it has been demonstrated that this gene acts to decrease *Salmonella* Typhimurium virulence in mice. This result is interesting, because this gene is located in the prophage Gifsy-2. The *sodCI* gene also located in the same phage was detected in all genomes, and was described as a positive virulence factor [431]. Overall, these results suggested the Gifsy-2 phage is present in all studied *Salmonella*, and proposed that TMV genomes lost the *grvA* likely to keep virulent abilities.

Finally, the *spv* and *rck* genes have also been identified only in *Salmonella* Typhimurium genomes (Appendix Figure 7.8). The *spv* encodes for a toxin [432], while *rck* expression mimics the natural host cell ligands and triggers engulfment of the bacterium by interacting with the epidermal growth factor receptor [433]. The *rck* expression has been demonstrated to be linked with the *pef* expression [433], therefore, detecting them together was expected.

4.3.2.2.2 Biocides and heavy metal resistance analysis

Five genes involved in biocide resistance activity were identified. Four of them had been detected in *Salmonella* Mbandaka in section 4.4.3.2. More precisely, these genes are involved in paraquat herbicide (*yddG*), peroxide (*sodA*), hydrogen peroxide – monochloramine (*rpoS*) and cation biocide resistance (*smvA*) [434, 435, 436, 437, 438]. In addition, the *nmpC* gene was identified in all genomes as a paraquat resistance. As *yddG* and *sodA*, the *nmpC* gene codes for a porin genes and is encoded near the *smvA* gene [434, 439]. All genes has been identified in all genomes, as described in Appendix Table 7.2.

As detected in *Salmonella* Mbandaka, metals resistance genes have enabled strains to cope with metals like magnesium and cobalt (*cor* and *mgtA*) [440, 441, 442], copper (*CUEP* [443]), gold (*golS*) or arsenic (*PSTB*), compounds found in food and water [444]. Concerning *Salmonella* Typhimurium and TMV in pigs, copper resistance has been identified previously, suggesting that the success of this serovar may have been driven by the extensive use of Copper as a growth promotor in pig rearing [180, 445]. In addition, the *sil*, *pco* and *ars* genes were identified in the 150 ST 34 samples (Appendix Figure 7.9), and have been reported as gene conferring resistance to silver [446], copper [443] and arsenic [447], respectfully. The *pco*, *sil* and *ars* genes are carried by the SPI-4 and have been associated with a strong enhanced resistance to Copper in anaerobic conditions, environment encountered by *Salmonella* in the host intestinal tract [448]. The *mer* genes conferring resistance to arsenic [449] were detected on 80 ST 34 samples, but in different phylogenomic clusters (Appendix Figure 7.9).

Finally, the same 8 genes involved in multi-compound resistance were detected (Appendix Table 7.3) as described in *Salmonella* Mbandaka (section 4.4.3.2). All *Salmonella* Typhimurium and TMV exhibited copper resistance due to the *cuiD* gene or iron resistance due to the *prmG* gene, both previously identified during a metagenomic study related to co-occurrence of antibiotics, biocide and metal resistance genes in pigs [450].

4.3.2.2.3 Plasmids and antibiotics resistance analysis

In total, 33 mobilizable plasmids and 18 conjugative plasmids were detected. Otherwise, 180 non-mobilizable plasmids have been identified, meaning they are missing relaxase and oriT, but can be mobile by other process, e.g transduction, natural transformation or cointegration in mobile plasmids [227]. Conjugative and mobilizable plasmids were only found in few genomes, between 1 and 30. Some plasmids were unique to a serovar, but only in small number. All plasmids considered, none of them were able to distinguish regions of *Salmonella* isolation.

More antimicrobial resistance (AMR) genes (MEGAResV2 [413] and Resfindr) were detected in *Salmonella* Typhimurium and TMV compared to *Salmonella* Mbandaka. TMV is known to harbor multiple AMR genes related to the antibiotics tetracycline, ampicillin, sulfisoxazole, and streptomycin [224]. In our study, we detected AMR genes against aminoglycoside (*aph*, *aac*, *aph*), becta-lactam (*blaTEM-1B_1*, *tem*, *carB*), phenicol (*floR*), sulphonamide (*sul*, *folP*), tetracycline (*tet*) and trimethoprim (*dhfr*).

We observed that beta-lactam resistance was carried by most TMV strains, whereas resistance to phenicol was carried by *S.* Typhimurium strains (Supplementary figure 7.7 and 7.9) through the *floR* gene. The phenicol resistance carried by *floR* in *Salmonella* Typhimurium [295, 451] and beta-lactam resistance associated to TMV isolates [224] have been previously observed.

Otherwise, trimethoprim resistance (*dhfr*) was carried by 15 genomes, disseminated all around the tree. Looking at the plasmid data, this gene did not seem to be carried by a plasmid, but it was studied before in *Salmonella* Typhimurium and others *Enterobacteria* [452]. Further analysis would be needed to understand its involvement in the diversity of *Salmonella*.

4.3.2.3 Genes and variants diversity highlights the low diversity of *Salmonella* Typhimurium and TMV in France

The objective was to examine whether it is possible to find geographical markers that could be used to develop rapid (PCR) methods to identify the origin of strains. To reach this objective, genomic analysis of genes and variants contents has been made as described in section 4.4.2.2 using variants from coregenome SNP analysis [345] and gene contents from a strict pangenomic analysis [100].

In total 7,350 genes were detected in the pangenome (Panaroo), from where 4,069 were identified as coregenes. More precisely, 2,406 genes were considered as cloud genes (contained in less than 15% of strains), displaying a consequent accessory genome. On the other hand 4,590 variants have been identified in the reference-based pangenome (iVARCall2), from where 343 were identified as core variants thus withdrawn from the analysis. As observed for *Salmonella* Mbandaka, we did not detect a single gene or variant that could distinguish the regions of strain isolation.

In order to explore combination of markers, we used the MarkerFindr script developed initially for *Salmonella* Mbandaka analysis (section 4.4.2.2). More precisely, the best combination of maximum 3 genes 4.10 or variants 4.11 were compiled, in term of associated accuracy scores. For each combination, the accuracy, the sensitivity and the specificity [453] of *Salmonella* Typhimurium and TMV isolates (n=188) were tested, as well as TMV isolates only (n=152). For genes (Table 4.10 and variants 4.11), the highest accuracy scores were able to distinguish samples isolated in region 1 from samples isolated in regions 2 and 3.

	Name	Accuracy	Sensitivity	Specificity
Typhimurium + TMV	group_3078, hin_2~hin, group_2530	0.76	97%	48%
TMV	tufB~tufB_2~tufA_2, group_3092, group_0238	0.77	83%	70%

Table 4.10: Combination of 3 genes to find marker for *Salmonella* Typhimurium and its monophasic variant

	Name	Accuracy	Sensitivity	Specificity
Typhimurium + TMV	9724_A_G, 178926_C_T, 4626125_A_G	0.74	72%	77%
TMV	2082_C_T, 178926_C_T, 4626125_A_G	0.82	87%	77%

Table 4.11: Combination of 3 variants to find marker for *Salmonella* Typhimurium and its monophasic variant

As the combination of 2 genes did not show accuracy greater than 0.60, combinations of 3 genes or 3 variants were investigated. For the combinations of 3 genes, an accuracy of 0.76 was proposed to discriminate Region 1 sample from Region 2 and Region 3 sample, using all data (Table 4.10). Using only the TMV isolates, the accuracy increased to 0.77, with 83 % sensitivity (67/78 true positive) and 70 % specificity (22 false positive) (Table 4.20). Compared to the accuracy using all samples, the accuracy did not change much, but the specificity increased from 48% to 70%, which showed that the selected genes were much more discriminating on strains from other regions. Overall, as the accuracy is inferior to 90%, this result tends to hypothesize that there is no discrimination between samples from the different regions. Genes annotation in Table 4.12 indicates that the combination using the whole dataset (n=188, *Salmonella* Typhimurium and TMV isolates) relies on accessory genome. 2 genes identified in the combination correspond to proteins identified on *Salmonella* phage 118970_sal3. The last one is gene identified is a *hin* gene, a DNA-invertase required for the inversion of the *fljB* controlling region [282], the phase 2 flagellar of *Salmonella* Typhimurium. This discovery is intriguing since it should be able to distinguish *Salmonella* Typhimurium from TMV samples more effectively than a specific region. A study using only TMV strains was also conducted for this reason. Looking only at monophasic variant of Typhimurium strains, genes selected seems to be less accessory, except for an ORF identified in *Salmonella* phage ST64T. Other genes are identified as elongator factor described in *Escherichia coli* [454], or a protein produced by an insertion sequence [455]. Overall, these genes seem coming from horizontal transfers or mobile elements, which would explain their selection as the discriminator.

Using variants, an accuracy of 0.74 was obtained for a combination of 3 variants using the whole dataset. This result is surprising, because it is smaller than the accuracy found using genes as marker. On the *Salmonella* Mbandaka dataset, variants were far more discriminating

Panaroo name	Annotation	Definition	Sensitivity	Comment
group_3078	118970sal3 (00123)	Putative transcriptional activator	249	found in *Salmonella* phage 118980sal3
hin_2~hin	hin	DNA-invertase hin	540	
group_2530	118970sal3 (00130)	Putative holin	387	found in *Salmonella* phage 118980sal3
tufB~tufB_2~tufA_2	tufB	Translation elongation factor Tu	300	
group_3092	orf-81	ORF from phage	255	found in *Salmonella* phage ST64T
group_0238	Unamed	Protein p-43	387	

Table 4.12: Annotation of genes found as markers from Table 4.10 for *Salmonella* Typhimurium and TMV isolates.

than genes. However, using the combination of 3 variants using exclusively TMV strains, the accuracy increase to 0.82, with 87% sensitivity and 77% specificity, which is much higher than the gene's accuracy. The three variants identified are located on genes (Table 4.13). One mutation located in position 2082 on gene *thrA* can have impact on bacteria's pathways, such as thialysine resistance as studied before [456]. These mutations are interesting, but does not allow to discriminate geographical areas due to low accuracy.

4.3.2.4 Applying pgSNP on the dataset to explore accessory diversity

In this section, we will present pgSNP applied to the problematic of the variability of *Salmonella* Typhimurium and TMV accessory genome within pig herds in France. As no markers were

Position	Gene name	CDS begin	CDS end	Comment
9724	yaaH	9376	9950	Putative regulatory protein
178926	aceF	178918	180807	Acetyltransferase (component of pyruvate dehydrogenase complex)
4626125	ulaA	4625215	4626645	Ascorbate-specific (PTS system EIIC component)
2082	thrA	337	2799	Bifunctional aspartokinase homoserine dehydrogenase

Table 4.13: Annotation of genes found as markers from Table 4.11 for *Salmonella* Typhimurium and TMV isolates.

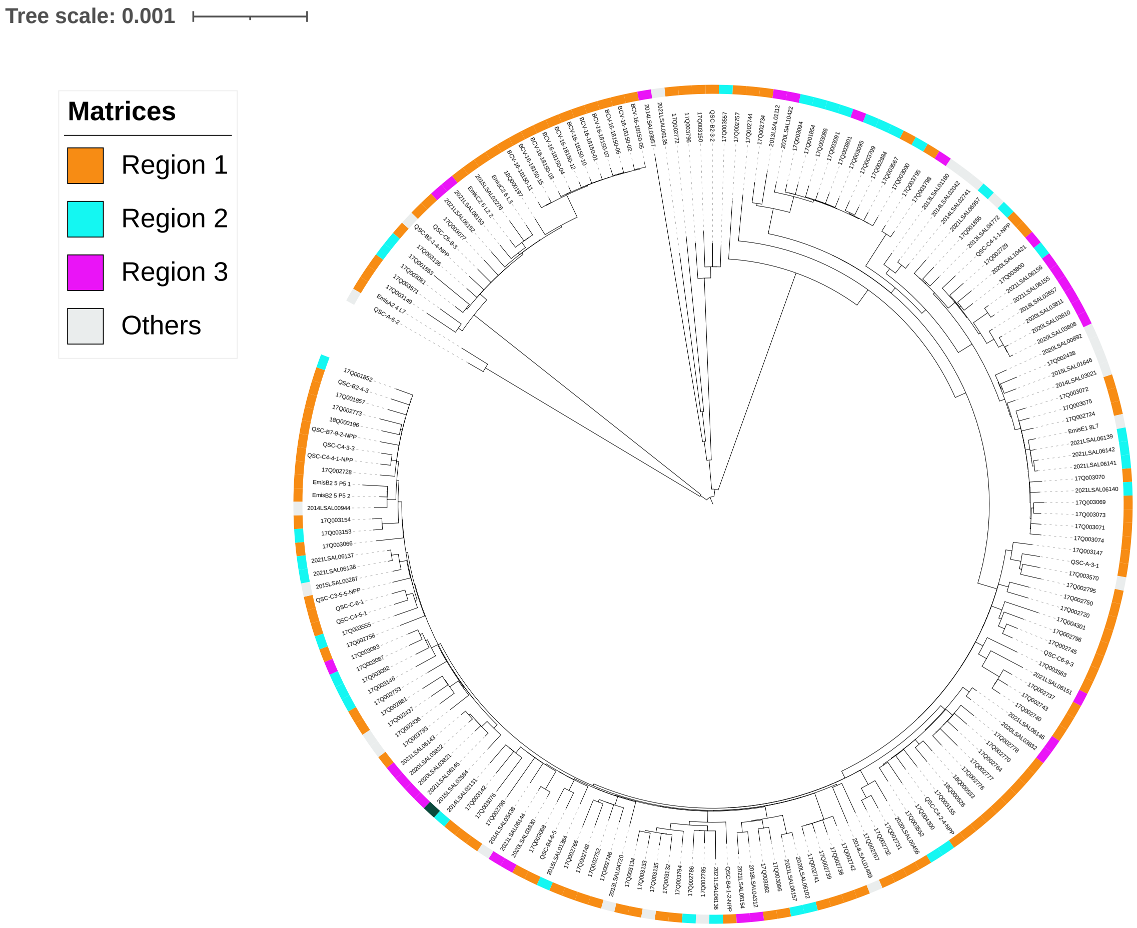


Figure 4.4: pgSNP on *Salmonella* Typhimurium TMV dataset from pig herds (n=188). Pangenome reference parameters : identity = 95%, minimum contig length = 500 bp. The outer ring corresponds to the sample department origin.

identified in the coregenome, and given the coregemone proximity of the TMV strains, a hypothesis arose that geographical markers could be found in the accessory genome.

4.3.2.4.1 pgSNP analysis supports the similar diversity between regions

To understand if the low diversity between regions is only significant in the coregenome, accessory genome was explored using pgSNP (section 3.2.1). A tree was inferred on 188 samples with department data from *Salmonella* Typhimurium and TMV dataset described in section 4.2.1.2. From the 4.4, it does not seems that there is any new geographical reconciliations. *Salmonella* Typhimurium are still on the left of the tree, and TMV samples are clustered together, without any geographical clustering. There are longer branch lengths on the leaf, but the scale is still small. The total reference pangenome alignment length (6064743 bases) is smaller than this of *Salmonella* Typhimurium and TMV from 3.3.4.1, while the datasets are approximately the same size (188 *vs* 192 samples). The time scale is wider on this dataset, but the matrices are wider on the outbreak dataset (strains from pork, eggs, humans and dairy products), that could explain a greater genomic variability in terms of analyzed base number.

Compared to the coregenomic tree 7.12, we observed a similar clustering for the strains of *Salmonella* Typhimurium in the center of the pangenomic tree and for TMV strains on the left of the figure. Overall, the RF distance is 194, which is high but less than *Salmonella* Typhimurium and TMV outbreak dataset (RF=228 between pgSNP phylogenomic tree and coregenome tree inferred by iVARCall2 [345]), for this dataset the addition of accessory genome does not provide additional information allowing to differentiate strains. Looking the department data, there is no new reconciliation that could explain a potential prevalence of one type of strain per region. We find some reconciliation between samples from the same department, showing some consistency in the fact that strains that come from the same breeding will most likely be related.

4.3.2.4.2 pgSNP allows a rapid identification of accessory content of *Salmonella* Typhimurium and TMV

Using reference pangenome, it is extremely easy to analyse the core and accessory genome present in dataset sample. Each contig was analysed to identify some elements which could explain a prevalence of the strains in herds. We focused our attention on mobile element content like plasmids using MOB-suite [416] and Blast. These accessory pieces are key elements of the adaptability of *Salmonella*, and as we have studied them before, it allows us to compare results and displays the importance of taking into account SNPs and structural variation, instead of only presence and absence of plasmids. We identified 15 plasmids or plasmids fragment mobilizable or conjugative (defined in section 2.3.5.5) in the whole dataset that are displayed in table 4.14 and figure 4.5. There is not a single plasmid contained in all strains and four plasmids are contained in more than 80 isolates. Compared to plasmids identified using assembly in section 4.3.2.2.3, it was difficult to make the link between the two sections, because the annotation of the closest plasmid may have duplicates due to the low identity (80% and 90%) [416] of the plasmids elements (oriT, mpf, described in section 2.3.5.5). But overall, plasmids present in a large number of strains are well retrieved using reference assemblies and pangenome, supporting the use of pgSNP to rapidly identify the presence and SNPs of plasmids.

What is interesting with pgSNP is that it is possible to identify SNPs in plasmid. For example, the biggest conjugative plasmid AR_0116 found in *Citrobacter freundii* (230kb), is detected as a complete sequence in 4 isolates, but one transposase of this plasmid is contained in 153 other samples. These genes can be found in other plasmids as well, therefore it is not possible to hypothesize on the link of integration of this gene in the chromosome of these strains. This plasmid is also found in *Escherichia coli*, *Shigella*, and some *Salmonella* serovars Manhattan, Senftenberg and Choleraesuis, often found in pigs. SNPs are present on this plasmids, and the 4 strains that contain the entire plasmids are not clustered together, which can lead us to think that the integration of this plasmid was done through independent events. Analysing the alignment, 9 SNPs are located all along the sequence, but also 17kb of gaps appears on one isolate, which shows a variation in structure. This part of the plasmid is not annotated, so the phenotypic repercussion remains uncertain, but this pattern is taken into account phylogenomically (Supplementary figure 7.13).

Phage content was also investigated in *Salmonella* Typhimurium and TMV samples (Table 4.15). The advantage of pgSNP is to be able to use only the reference pangenome for annotation. Nine phages have been identified in the dataset. The presence of these phages is displayed in figure 4.5. Gifsy-2, 118970_sal3 and SPN9CC contained in 188 isolates, Gifsy_1 contained in 183 isolates, SfII contained in 128 isolates, GF_2 contained in 26 isolates, Entero_lambda and pro483 contained in 8 isolates, and finally SSU5 contained in 7 isolates. Gifsy-2 was previously mentioned in section 4.3.2.2.1 as virulome genes of this phage has been

Plasmids name	Number of isolates
pETEC_6	138
pExPB5-59-1	131
pST1120	81
pKPHS4	36
pKPHS4_2	36
p10-3184.2	33
p11-0813.1	28
p2CFSAN000752	20
p3.8k	11
pO26_2	5
L725 plasmid unnamed4	5
FDAARGOS_647	4
pSH14-009_2	4

Table 4.14: Mobilizable and conjugative plasmid presence in *Salmonella* Typhimurium and TMV dataset

Phage name	Number of isolates
Gifsy-2	188
118970_sal3	188
SPN9CC	188
Gifsy_1	183
SfII	128
GF_2	26
Entero_lambda	8
pro483	8
SSU5	7

Table 4.15: Phage presence in *Salmonella* Typhimurium and TMV dataset

identified in all genomes. Phages also present structural variation, such as pro483. One isolate has a Phage-like supplementary protein in the beginning of the sequence which is taken into account in the subtree corresponding to this phage, but as these strains are not clustered together in the final tree, it is difficult to prove an effect on the phylogenomic tree. Others phage such as GF_2 exhibits some SNPs between samples, but overall all the sequence of phages is conserved in the strains.

4.3.2.4.3 Impact of the accessory diversity contribution to the analysis

The pgSNP phylogenomic tree shows that there are new reconciliations thanks to the accessory genome, but not associated with a geographical prevalence by department. Using the reference pangenome, we could easily observe the content of accessory genome and horizontal gene transfer by the presence of plasmids and phages. We showed that the presence of a plasmid or a phage could be highly variable across the strains, and that there were structural differences which are most certainly due to an integration of these elements at different evolution time, or through environmental exchanges with different other organisms. As demonstrated in 3.3.4.1, the accessory genome impact more TMV than *Salmonella* Typhimurium samples, mostly due to the low diversity of TMV in the coregenome. Even with the new reconciliations, this supports

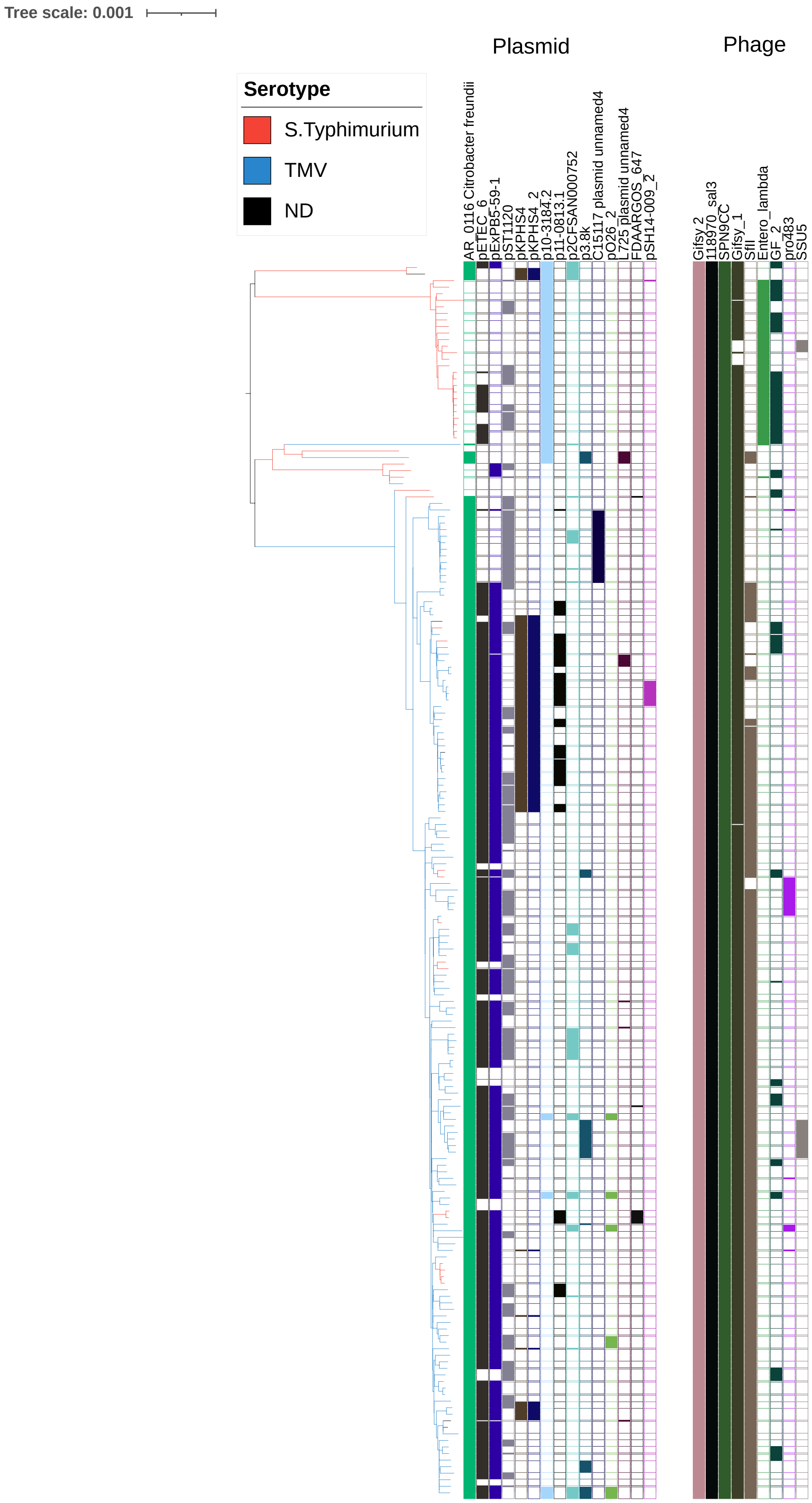


Figure 4.5: Mobilizable and conjugative plasmids, and phages presence in *Salmonella* Typhimurium and TMV pig herd dataset. Branches are colored according to the serotype. The first column group corresponds to plasmids present in the dataset. The second column group corresponds to phage detected in the dataset.

the hypothesis of a low diversity of TMV circulating in France in pig farms, which evolves and could adapt to its environment thanks to its accessory genome. In conclusion, pgSNP has brought resolution to the issue of genomic variability of *Salmonella* Typhimurium and monophasic variant of Typhimurium in France.

4.3.3 Compare the French diversity to the worldwide diversity

As we previously showed that the diversity of TMV in France seems to be clonal, a new question arose : is the low diversity in France unique or is this diversity shared by all country? In this part, I will compare the diversity found previously in France and the diversity from worldwide strains available online.

4.3.3.1 Phylogenomic analysis revealed a geographical diversity of TMV

325 samples from France and others countries were selected as described in section 4.2.1.3. Coregenome phylogenomic tree was inferred by IQ-TREE [107] using SNPs identified by iVARCall2 [345] pipeline, as described before 3.2.3.

The tree followed an evolutionary model of TVM+F+I, and converged with a commensurate negative likelihood (-6795270.421) after 140 iterations. Bootstraps values are displayed in Appendix Figure 7.10. As described previously, the impact of recombination events on the phylogenomic tree was measured using ClonalFrameML [110] in Appendix Figure 7.11. The phylogenomic tree with homologous recombination events was inferred using 3825 SNPs, while 114 SNPs were detected in 34 homologous events (10 on internal nodes, 24 on leaves). 17 homologous recombination events on leaves were located on other isolates than French ones. As previsouly shown, the homologous recombination events did not have a high impact on the topology of the tree 7.11, as only few minor differences were observed between nodes.

Analysing the phylogenomic tree of monophasic variant of Typhimurium, long branches were observed on some genomes (SRR11901838 from Canada, ERR3415697 from China, ERR5443072 from Belgium), but overall the scale was small compared to *Salmonella* Typhimurium or *Salmonella* Mbandaka phylogenomic tree. The mean SNP difference between all samples from the tree was 68 SNPs, which can be explained by the fact that the dataset was larger, with a higher diversity in geographical areas. Also, long branches created further distance between samples, as the maximum SNP difference was 186 (SRR11901838 from Canada - ERR5443072 from Belgium). Overall, the phylogenic tree presented topological structure, with samples with low SNPs difference between them, and others with a unique diversity displayed by long branches.

Looking at the global topology of the phylogenomic tree, we observed that French genomes were disseminated in two groups. This difference is illustrated in Figure 4.7, where both groups were topologically visible on the *Salmonella* Typhimurium and TMV from French tree. We observed that in the group 1 (blue), genomes from Italy were clustered with genomes from France, with fairly large branch distances. A small cluster of 6 samples from America was close to two French samples (17Q003798 and 17Q003798), but these two strains had a very large internal node which demonstrated a high genetic distance between them and the American strains. Looking at the group 2 (green), French samples were clustered with 7 strains from other European countries (Germany, Italy), and one sample from Thailand. Overall, French genomes seemed to cluster together, with some exceptions of isolates from bordering countries.

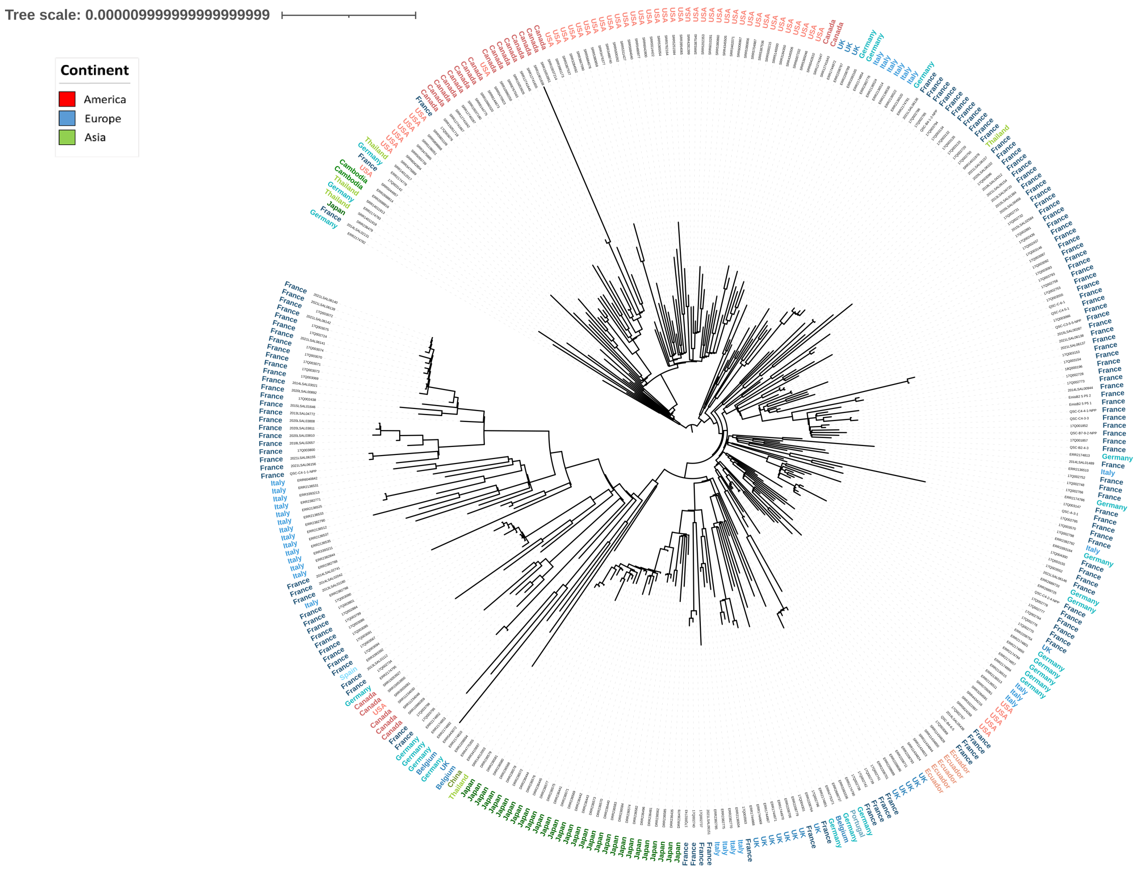


Figure 4.6: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of Monophasic variant of Typhimurium isolated from pigs (n=325). Outer ring corresponds to the country origin of the strains, colored by continent.

This observation highlighted the genomic specificity of monophasic variants in France.

Looking at genomes from other countries, Japanese strains clustered together with low SNPs difference. All Japanese genomes came from the same laboratory, with a time scale of two years (2016 and 2017), which could explain this low diversity. Thailand strains were disseminated all around the tree, while strains had all been isolated by the same laboratory in 2019. A previous investigation of strains in Thailand displayed different antimicrobial resistance pattern [457], but further analysis would be necessary, especially focusing on trading market as Thailand exports and imports pigs and pork products [458].

Focusing on European countries, Germany samples were disseminated all around the tree, with Italian and UK samples, while most of UK samples were clustered together at the bottom of the tree. Overall, European countries seemed to share close genomes, certainly due to the spread of ST 34 in all European countries [286] in all sectors from humans to farm and environment. Finally, all American strains were clustered at the top of the tree, except for Ecuador samples which were clustered at the bottom of the tree. Canada and USA shared some strains that clustered together, highlighting the geographical proximity linked with the genomics proximity. This result is interesting, because most of both countries strains shared an internal node, that then cut in two clusters, one for each country. This displayed a genomics distinction between these two countries. Further analysis focusing on these strains would be

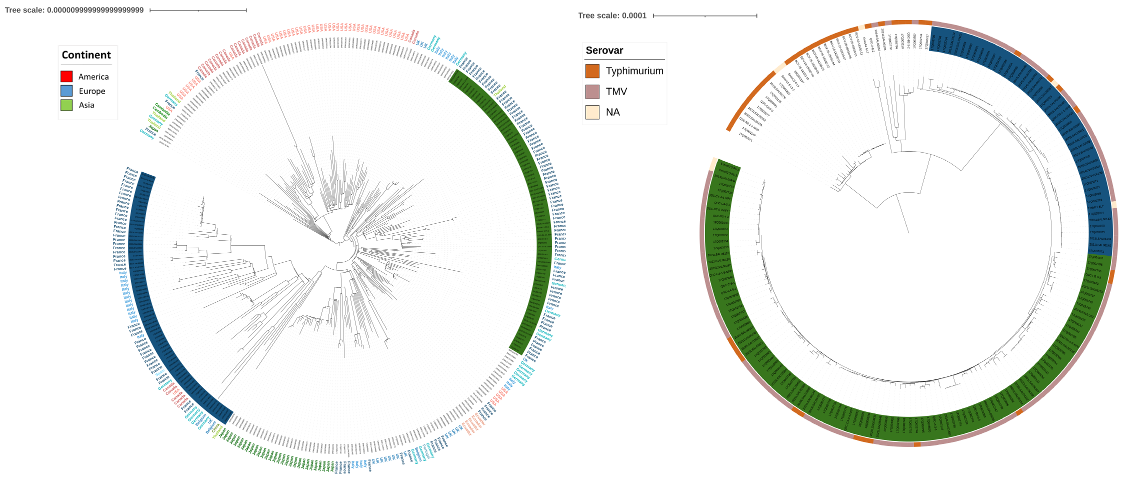


Figure 4.7: Comparison of the topology of Monophasic variant of Typhimurium in France and in the world. Left : coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of Monophasic variant of Typhimurium isolated from pigs (n=325) in the world. Right : coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant in France (n=188). Outer ring is described by the caption to the right of each tree. French TMV isolates are colored in blue (group 1) or green (group 2) according to topological groups.

interesting to conclude on these hypotheses.

With this information we can estimate that there may be two clones in France, one of which is strongly shared with Italian strains and the other which seems more unique.

4.3.3.2 Can a diversity pattern be attributed to French samples?

Finally, to understand if the diversity is really discriminating between European countries and worldwide countries, or between France and worldwide countries, the genes and the variants contents was studied, combined with MarkerFinder to see if there were any discriminating combinations. 9,671 genes have been identified with Panaroo, where 3,458 were considered as core-genes. Meanwhile, iVARCALL2 identified 4,148 variants, where 3,825 were considered as SNPs. As the diversity between TMV samples is low, no unique gene or unique variant could discriminate France samples from worldwide samples.

Using the combination, European strains could not be discriminate from worldwide strains, given that the combination accuracy values were low. However, focusing on France samples against worldwide samples, a combination of 3 genes was found with 0.86% accuracy, and a combination of 3 variants with 0.81% accuracy (Table 4.16). The fact that the accuracy score for variants combinations was lower than the accuracy score for genes combinations was surprising because the score of the variants was always higher in the previous tests on *Salmonella* Typhimurium and TMV dataset in France, and *Salmonella* Mbandaka. However, this difference can easily be explained by the fact that the variants detected by iVARCall2 were variants on coregenome positions, so all the accessory part was not taken into account. On the other hand, the screening of gene combination will prioritize the accessory genes via gene presence/absence detection from Panaroo. Due to the fact that TMV genomes do not have much diversity, it is very important to take into account the accessory genes to add discrimina-

	Name	Accuracy	Sensitivity	Specificity
Gene	group_5096, group_2046, napA_2∼nap_A_1	0.86	86%	86%
Variant	201482_G_A, 658777_G_A, 680767_A_T	0.82	55%	99%

Table 4.16: Table of combination of variants and genes with the highest accuracy score.

Panaroo name	Annotation	Definition	Sensitivity	Comment
group_5096	STM3521	Putative ribonucleoprotein related-protein	168	operon in a RNA polymerase-holoenzyme subunit
group_2046	tnpA1	Transposase for insertion sequence element IS200	459	identified in multi-resistant bacteria
napA_2∼nap_A_1	napA_1	Nitrate reductase	231	

Table 4.17: Annotation of genes found as markers, discriminating France TMV samples from worldwide TMV samples

tion. Yet, the variant combination was really interesting, due to the exclusivity score at 99%. Indeed, 192/193 genomes from world dataset was not targeted by this variant combination, but only 73/132 genomes from France were targeted. In any case, this combination could ensure that a strain detected from a pig did not come from another pig coming from abroad.

However, the combinations of 3 genes seemed more stable, and further analysis would be needed to improve the accuracy, in particular by analysing whether there are any accessory variants on these genes.

Finally, genes and variants detected as combination markers to discriminate French samples from worldwide samples were annotated, using the method previously described 4.4.2.2. In Table 4.17, one gene is annotated as a transposase, previously described in multi-resistant bacteria [459]. Others genes have been annotated as an operon in a RNA polymerase holoenzyme subunit [460], or a nitrate reductase [461]. These genes are important in the bacterial growth, notably *napA_1* which is mainly used by *Salmonella* under anaerobic growth conditions in the presence of low nitrate concentrations. Variants markers have been identified on genes (Table 4.18) essential for growth of the bacteria (*yadF*[462]), or to mediate most reactions of

Position	Gene name	CDS begin	CDS end	Comment
201482	yadF	201410	202088	Carbonic anhydrase
658777	entE	657443	659053	dihydroxybenzoate-AMP ligase
680767	rna	679992	680798	RNase I

Table 4.18: Annotation of variants found as markers, discriminating France TMV samples from worldwide TMV samples

RNA metabolism [463]. The impact of the variant has not been investigated, but it would be interesting to understand why these variants discriminate against worldwide genomes.

4.3.4 Discussion and limits of the study

4.3.4.1 Low divergence of ST 34 strains

Low divergence of monophasic variant of *Salmonella* Typhimurium has been demonstrated in this study, with a hypothesis of the dissemination of one or two clones in France in all the farms. While this conclusion is based on our observations and our dataset, it was previously described in Ireland in pigs [292]. The author pointed out that TMV isolated colonised recently pig herds, and undergone limited sequence divergence. Low divergence was observed between 9 farm investigated in 3 provinces (between 0 and 12 SNPs), and in some cases identical or near identical strains were isolated from more than one farm or feed mill, suggesting a common source of contamination. As we observed in this study, this kind of pattern suggest either contamination of multiple farms from a common source, or direct transmission between the farms. Further analysis would be needed, as integrating information regarding movement of animals and all risks of contamination related to these movements, such as livestock exchanges, contamination by transport from farm to slaughterhouse, or also contamination by the human vector (personnel). Also, as discussed in *Salmonella* Mbandaka, implication of feed as an important source of *Salmonella* on farms should be investigated. A PhD thesis demonstrated that 2 TMV samples isolated from feed mill samples has been associated to 2 farms, using molecular typing [464]. These data provided evidence that feed had a possible role to play in transmission of *Salmonella* to pigs [465, 466]. Using WGS, the analyzes could be much more precise and could show the link in a more robust way. The few isolates from feed of our pork dataset did not allow us to conclude on this hypothesis, but offered a new perspective to this work.

4.3.4.2 Persistence of strains

In this study, I was able to display biocides resistance from different genes involved in paraquat herbicide or compounds in other detergents. These compounds are most of the time used in agricultural for weed and grass control. Hydrogen peroxide resistance can also be linked to detergent [467]. However, resistance genes to the main compound of detergents like quaternary ammonium has not been detected. In MEGAResV2, only *emeA*, *galE* and *bcr* genes have been identified as quaternary ammonium resistance genes [468]. In the dataset, I identified *qacL* gene in 3 genomes, which has been described as a quaternary ammonium compound resistance protein [469].

It was demonstrated that *Salmonella* is able to develop resistance were exposed to a stressful environment [470, 471]. A study pointed out that cleaning and disinfection protocols were not sufficient to eliminate *Salmonella* in slaughterhouse, that thus were able to contaminate the carcasses [472]. If these protocols are not sufficiently bactericidal, they can help the *Salmonella* to adapt and develop new resistances to detergent compounds. The study proposed that to avoid cross-contamination in slaughterhouses, *Salmonella* status of the herds should be determined closer to the slaughter date, to adapt cleaning and disinfection protocols especially for critical machinery and better hygienic designed equipment. However, *Salmonella* shredding is intermittent in pigs, and can be accelerated by stress, such as transportation to a slaughterhouse [473, 474]. It is therefore difficult to calculate when the status of salt should be done.

It was also discussed previously that ST34 samples are circulating in Europe and are present in all sectors from farm and environment to humans, but remain predominantly associated with the pig reservoir and does not persist long-term at individual farm level in other food animal species as humans or poultry ([286]. Further analysis comparing genomes from different animal hosts could also highlight resistance genes which would be included only in pork *Salmonella* strains and would explain this persistence in this host.

Finally, studies are needed to understand the modes of dissemination of *Salmonella* Typhimurium and its monophasic variant. It should be noted that clinical salmonellosis is more often caused by *Salmonella* Typhimurium (37.5% in 2008) contamination and not TMV (1% in 2008) [475], while pigs more often shed TMV (85% serovars detected in 4 farms [476], 20% in 100 farms between 2020 and 2022 [477]).

4.3.4.3 Limits

The limit of this part of the study is the lack of balance in the dataset. This is mainly due to the difficulty of finding other strains outside of Region 1, for *Salmonella* Typhimurium but also for TMV strains. Region 1 accounts for 78% of pig production in France [402]. To compensate, strains from slaughterhouses with known geographical data of the pig herd have been added. However, a pig that is slaughtered while contaminated can spread the bacteria to the rest of the chain resulting in cross contaminations. For example, in a study carried out on the analysis of strains in slaughterhouses, the authors have shown that a recycled water used in dehairing machine was identified as a important source for *Salmonella* contamination [478]. This means that there is a possibility that strains found on a pig may come from another infected pig, which would distort the results. However, in France, most of the herds are located near slaughterhouses, with a mean distance between the herd and the slaughterhouse of 120km (see Figure 4.8). While these data should not change significantly for Regions 1 and 3, it is quite possible that slaughterhouses in Region 2 can accommodate pigs from Region 1. Furthermore, trade agreements between producer groups and slaughterhouses can distort the results, with animals being transported between different regions.

Despite, I was not able to find genomics markers that would be able to discriminate samples between Region 3 and Region 1 while lefting out Region 2 samples from the analysis. So even if there is the possibility of having cross-contamination in the slaughterhouse, the conclusion remains the same. In any case, there is a lack of direct monitoring in pig farms, and the lack of data has been filled by slaughterhouses data. Overall, additional data could provide better robustness in the event of contamination of the same clone, or displays the diversity spread over the whole country.

Another limit to the study was the lack of methodology to take into account accessory variants. On TMV, I showed that the genetic markers were rather on the accessory, as these genomes showed little divergence on the coregenome. Here, I displayed the importance of taking into account accessory genome, especially for serovar like TMV. In the next chapter, methodology that took into account dispensable genome will be described, but the lack of time did not allow me to use these accessory SNPs as markers in the dataset. Given the scores determined with the accessory genes, we can hypothesize that adding a variant would increase the accuracy score.

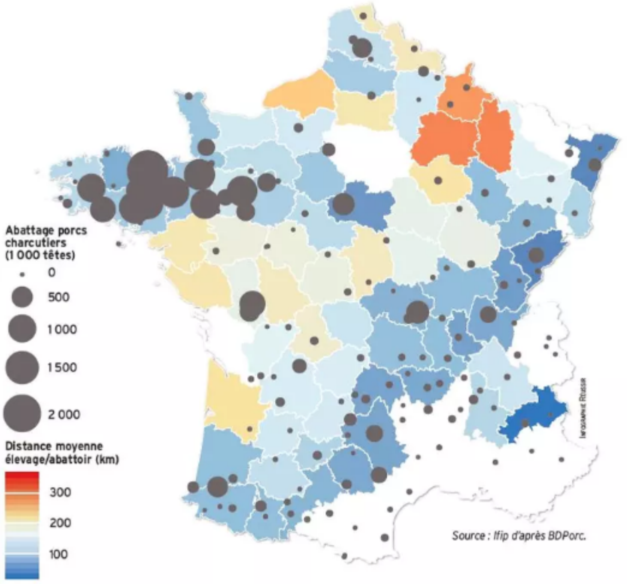

Abattage porcs charcutiers (1 000 têtes)
0
500
1 000
1 500
2 000
Distance moyenne élevage/abattoir (km)
300
200
100
Source : Ifip d'après BDPorc.

Figure 4.8: Slaughtering and distances by department of origin of farms. The size of circles corresponds to the number of pigs slaughtered, from small (less than 500) to large (more than 2000). Colors corresponds to the mean distance between the herd and the slaughterhouse, from blue to red. Map from IFIP.

4.3.5 *Salmonella* Typhimurium and its monophasic variant in pigs: conclusions

Geographical diversity of *Salmonella* Typhimurium and its monophasic variant has been studied in this part. By comparing strains, I highlighted the low diversity between the strains from the 3 different regions of the dataset. I also highlighted that this diversity was not linked to an origin/source in the chain production chain and, despite the year of isolation, strains from different origins/sources were found clustered together. In France the strains are disseminated all along the chain, without any particular adaptation to the source. This suggests a continuous contamination between the pigs herds, the slaughterhouse and the processing plant (section 4.3.1). I also characterized the French diversity of monophasic variant of Typhumurium compared to the diversity existing in other European and worldwide countries (section 4.3.2.1 and 4.3.3). Comparison with monophasic variant isolates from other countries highlighted the genomic specificity of monophasic variants in France, with some exceptions of isolates from bordering countries. We observed a discriminative signal between two groups of TMV from France, where one seemed to be shared by Italian samples, and the other one seemed unique. Bordering countries sharing the same diversity could be due to pigs trades, or contamination by a vector such as food products, trucks or food mill.

Looking at the genomic content, the high number of resistance genes to antibiotics, heavy metals and biocides explained the prevalence of these two serovars in pig farms. Considering my results, genomic diversity did not appear to be related to geographic diversity, as the analysis of variants or genes did not pointed out a discrimination between samples from different regions. This results supported the hypothesis of the dissemination of the same TMV clone in the country. For *Salmonella* Typhimurium, the lack of strains did not allow us to support the same conclusion as for TMV.

Finally, I also was able to characterize the diversity using a combination of genes and variants, and pointed out the possibility to trace back a French human infection to pigs herds using genomic markers. Further work would be needed to verify that these markers are viable for potential PCR, and also further analysis of accessory variants may reveal better accuracy. But looking at the scores, the hypothesis that there is not enough regional genomic diversity between the French strains to separate them seems the most logical.

4.4 Results: Investigation of *Salmonella* Mbandaka in bovine and poultry industry

Salmonella Mbandaka is poorly studied serovar at the genomic scale, but highly prevalent in north-western France in cattle herds and production chain (discussed in section 2.3.8.1). In order to understand its dissemination and its diversity, three main axes presented in the introduction will be explored.

- First, we will analyse the diversity of *Salmonella* Mbandaka in France, in different reservoirs.
- Second, the adaptation of *Salmonella* Mbandaka to its host will be investigated to find genomics markers that would explain the dissemination of *Salmonella* Mbandaka in bovine.
- Third, main genetic factors will be explore with a major review on the genomics of *Salmonella* Mbandaka which was poorly studied before.

4.4.1 Analysing the extent of the biodiversity of *Salmonella* Mbandaka in the different reservoirs

To understand the global diversity of this serovar in bovine industry in north-western France, we inferred a phylogenomic tree using 140 genomes. All samples were defined as ST 314 by MLST [479]. Phylogenomic tree was inferred by IQ-TREE [107] using SNPs identified by iVARCall2 [345] pipeline, as described in section 3.2.3. The tree followed an evolutionary model of GTR+F+I, and converged with a commensurate negative likelihood (-6663601.9824) after 121 iterations. Most of the nodes were supported by high bootstrap values.

Altogether, 1,062 SNPs has been detected in the 140 bovines genomes, which is roughly the same number of SNPs observed in *Salmonella* Dublin (section 4.5.2), except that the dataset is more than 3 times smaller than *Salmonella* Dublin dataset. Using ClonalFrameML [110], 48 SNPs has been identified in 8 homologous recombination events in 285 bp (4 on leaves, 4 on internal nodes). Excluding the 48 variants located in homologous recombination segments, a tree was inferred to measure the impact of the homologous recombination segments in the topology. The tree without variants from homologous recombination also converged with a commensurate negative likelihood (-6662273.430) after 200 iterations, following TVM+F+I model. Comparing the two phylogenomic trees (Figure 7.17), the topology was slightly impacted by the removal of homologous recombination variants. Two neighboring clusters are switched in the phylogenomic tree without homologous recombination variants, but overall with only few minor differences observed between nodes.

The phylogenomic reconstruction indicated that the isolates do not tend to group by years or matrices in Figure 4.9. Some isolates from cheese and milk are clustered together on the left

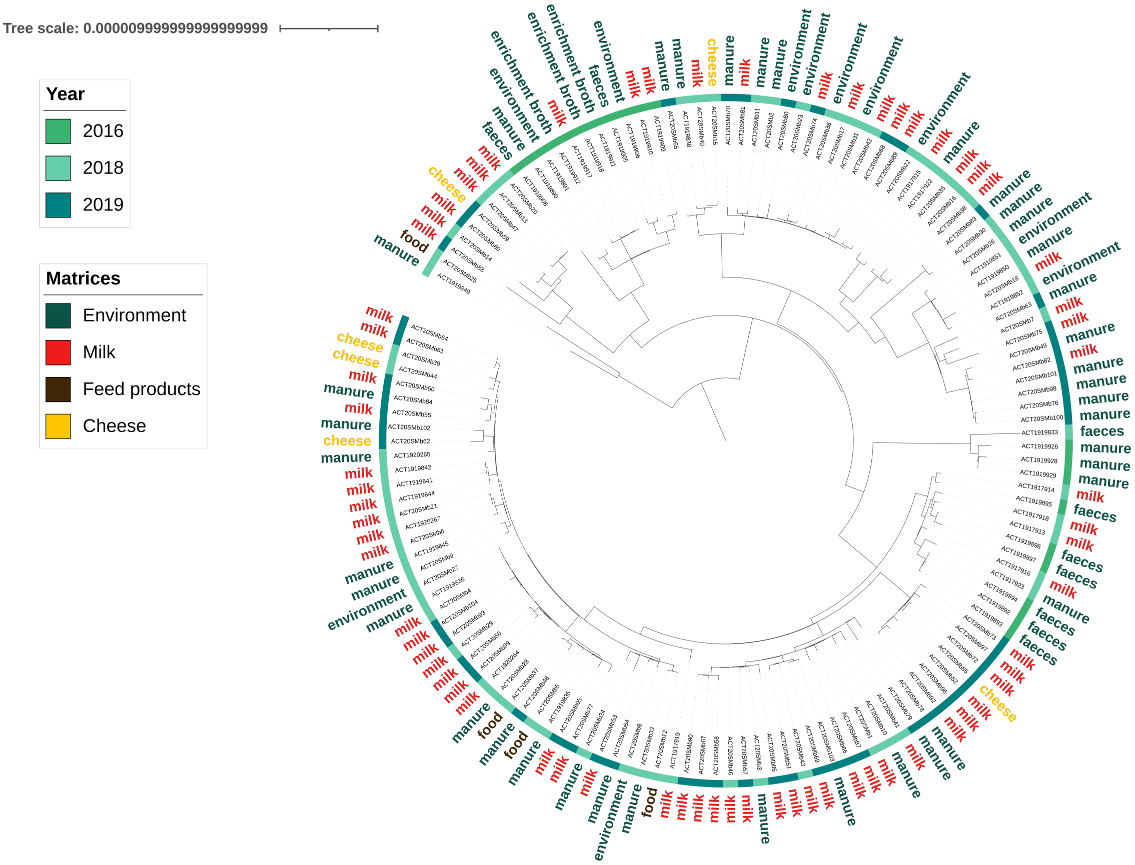


Figure 4.9: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine between 2016 and 2019, with homologous recombination events excluded. Inner ring describe isolation year. Outer ring display the isolation matrix.

of the tree, suggesting a likely contamination through the food chain production. Also, it is interesting to note that feed products are all around the tree, which suggests a potential source of contamination in cattle by this vector, itself heterogenous. Looking at isolate sources in Figure 7.15, we observed that most of the sample isolated from the same farm are clustered together. However, some samples isolated from raw milk from different farms (ACT20SMb64 and ACT20SMb61) are identical at the coregenome scale, with 0 SNPs differences while some samples from geographic area "V2" are disseminated all around the tree. All isolates from "V2" are milk isolates, suggesting a contamination through different bovine farms. Moreover, the "Ind" isolates corresponding to the cheese matrices are scattered all around the tree, which indicates a potential contamination by different milks. The samples isolated from the "B" farm at the same year displayed a 4 SNPs difference. This number of SNPs between the two strain can come from two different *Salmonella* populations introduced at different time [478]. Overall, the max difference of SNPs between 2 strains is 197 SNPs (ACT1919833 and ACT20SMb25), and the mean difference is 82 SNPs (median = 102). This finding demonstrates the great diversity of *Salmonella* Mbandaka, as well as the contamination along the production chain, but also between herds.

4.4.2 Is *Salmonella* Mbandaka adapted to its host?

As observed previously, *S.* Mbandaka is described by a high number of SNPs on a small geographic scale. To understand if this genomics diversity is specific to the dairy industry, or if this serovar has an overall high diversity, bovine samples have been compared from poultry samples, as this serovar is well associated to poultry industry in France [43].

4.4.2.1 Phylogenomic analysis of the diversity of *Salmonella* Mbandaka in bovine and poultry

Phylogenomic tree was inferred by IQ-TREE [107] using SNPs identified by iVARCall2 [345] pipeline, as previously described. The tree followed an evolutionary model of TVM+F+I detected with ModelFinder, and converged with a commensurate negative likelihood (-6705306.4880) after 196 iterations. Most of the nodes were supported by high bootstrap values.
Altogether, 3,567 SNPs has been detected in the 304 genomes, of which 64 lay within 13 homologous recombination events (7 on leaves, 5 on internal nodes) among a total of 1460 bp. Host specific clusters were not identified according to homologous recombination events. A tree was inferred excluding the 64 variants located in homologous recombination segments. The tree without homologous recombination variants converged with also a commensurate negative likelihood (-6702439.362) after 110 iterations, following TVM+F+I model. As observed previously, the comparison of the topology with and without homologous recombination events displayed a slightly difference (Figure 7.19), impacted only by bovine samples.

The phylogenomic analysis in Figure 4.10 revealed clusters of bovine and poultry disseminated all around the tree. Some genomes (n=11) are clustered with samples from other matrices, but overall we mainly observed genomics clusters distinguishing between the strains isolated from cattle and poultry. More precisely, 6 poultry samples were clustered in bovine groups, while 5 bovine samples belonged to poultry groups. We investigated these samples focusing on more precise matrices of isolation displayed in Figure 7.18. The 6 poultry samples and the 5 bovine samples clustered within other hosts were all isolated in the same geographic region (i.e. Normandy).

Of these six poultry samples, two (S17LNR0564, S19LNR1916) were genetically closed to bovine samples from milk, displaying a mean of 37 SNPs with the 5 bovine closed to the poultry sample. Two other poultry samples (S17LNR0975, S16LNR1497) were clustered with a environmental bovine sample from faeces with 38 SNPs apart. One poultry sample (S17LNR0496) was close (20 SNPs differences) to 4 bovines samples isolated in raw milk and enrichment broth. Finally, one poultry sample (S18LNR1821) was clustered with one bovine sample from milk, exhibiting 27 SNPs difference.

Of these five bovine samples, four (ACT1919926, ACT1919928, ACT1919929, ACT1919833) were isolated from manure and faeces, and therefore rather come from an environmental matrix. ACT1919929, only harbors 22 SNPs difference with a broiler isolate (S20LNR0837). Other bovine displayed 24, 29 and 45 SNPs difference with the nearest poultry strains.

The last bovine sample (ACT20SMb25) was isolated from cattle feed, and displayed only 10 SNPs differences with the 2 other poultry samples in the same internal branch. This observation is interesting, as some food products are used to both feed cattle and chickens [480], and therefore allows us to hypothesize a potential source of contamination by food products.

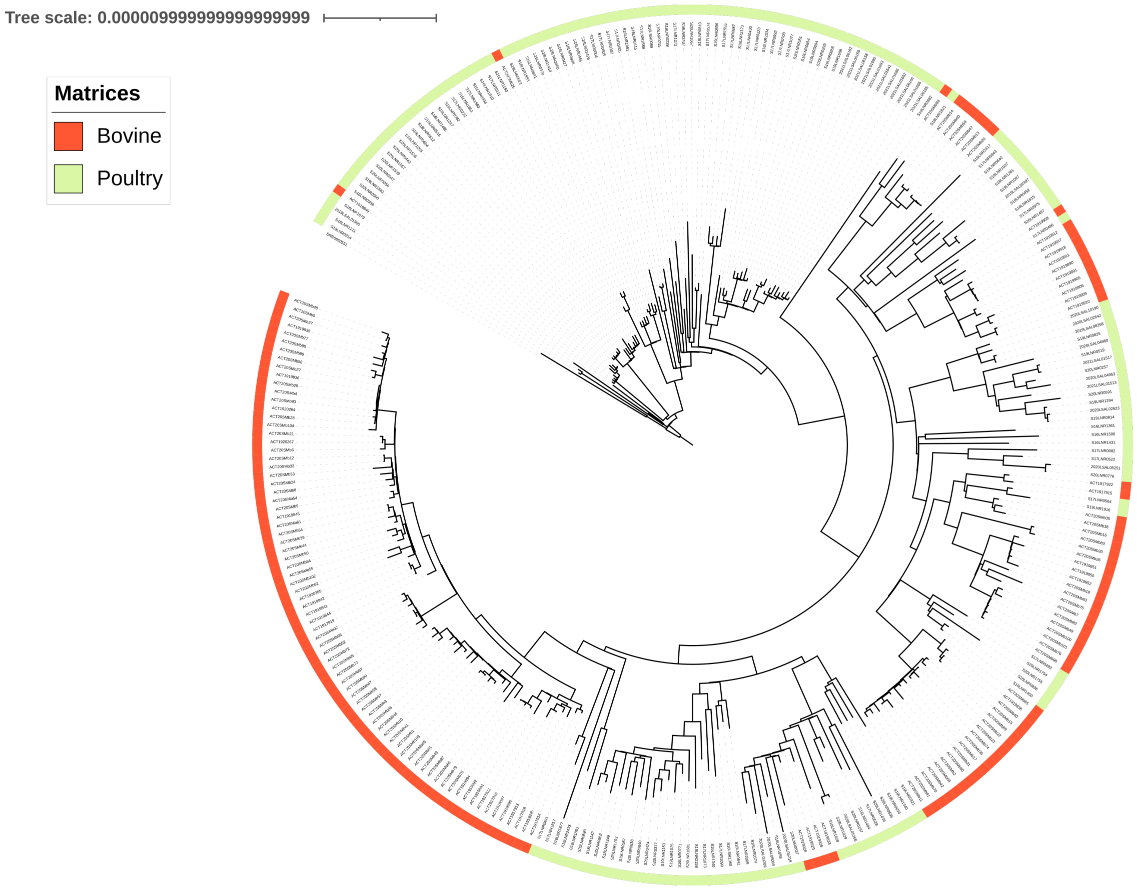


Figure 4.10: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Outer ring corresponds to the matrix of isolation.

Overall, the results observed on this tree for bovine groups are the same as those previously observed based on the cattle tree: the strains are not clustered by matrices and bovine milk strains are clustered with bovine cheese strains, demonstrating a continuous exchange of strains between matrices over the year, and a possibility of transmission of the bacteria throughout the production line. Interestingly, both poultry samples isolated from layer and broiler hens also cluster together. Finally, turkey samples are disseminated all around the tree, but with unique long branch length, and were consequently identified as singletons.

The geographical metadata in Figure 7.18 reveals that most of the time, poultry sample cluster with isolates from the same region (east of the tree), or from the same bordering regions (south and north-east clusters). The geographical data are not precise enough to suggest that the geographical distance is a major factor in genomic divergence, but with these minimal metadata we can hypothesize that the geographical and genetic distances are correlated.

However, in the upper part of the phylogenomic tree, strains from different regions are clustered together, without any geographical proximity links between the regions. This low diversity between strains from different geographical sources almost exclusively affects poultry isolates, which allows us to hypothesize that a common breeding source may exist. Indeed, parenting flocks are in farms and take care of the births of the hens, which are then sorted into broilers

	Name	Accuracy	Sensitivity	Specificity
Genes	group_1062, ydiO	0.80	72%	91%
Variants	1567699_G_T, 3213391_T_C	0.91	87%	96%

Table 4.19: Duo combination of markers

	Name	Accuracy	Sensitivity	Specificity
Genes	group_716, group_2947, group_2909	0.92	91%	96%
Variants	199204_C_T, 1567699_G_T, 3213391_T_C	0.95	95%	97%

Table 4.20: Trio combination of markers

and layers hens. These hens are then redistributed in farms, and can spread the disease from the same source. Unfortunately, the lack of metadata does not allow us to explore this hypothesis.

4.4.2.2 Screening for genomic markers of *Salmonella* Mbadandaka linked to host discrimination

As the phylogenomic tree revealed different host specific clusters, we hypothesised that genomic determinants specific to one of the hosts may exist. Indeed, the main objective of this study was to find markers in order to discriminate poultry samples from bovine samples. To achieve this objective, genomic analysis of genes and variant content has been performed. We retrieved variants from the coregenome SNP analysis with iVARCall2 [345]. Gene content was inferred by using Panaroo [100], a graph-based pangenome clustering tool, with a threshold increased to 90% of cluster identity to avoid finding markers that would cluster with other genes.

First, the content of variants and genes were analysed. Panaroo detected 9488 genes, where 3655 were identified as core genes (present in all samples). 3922 variants have been identified by iVARCall2, where 342 were identified as core variant. Looking at accessory genes and accessory genome, no unique gene or variant that could discriminate bovine samples from poultry samples was found, so combinations of genes or variants has been explored.

A combination of 2 genes has been proposed to discriminate samples using MarkerFindr (section 4.2.11) from bovine or poultry host, with an accuracy of 0.80 (Table 4.19). Using a combination of 3 genes (Table 4.19), the accuracy increase to 0.92, with 91 % sensibility (127/140 bovines identified) and 96 % specificity (11 false positive poultry) (Table 4.20). Genes identified annotation reported in Table 4.21 displays that the identified genes corresponds to accessory DNA (phage gene) except *ydiO*. This gene has been predicted as a acyl-CoA dehydrogenase and a modification methylase subunit [481] to protect the DNA from cleavage by the BsuMI endonuclease [482]. It has been previously identified in *Salmonella* Typhimurium as a secondary β-oxidation pathways which can be used by the bacteria to growth on lipids in anaerobic conditions [177].
Using the combination of 2 variants, the accuracy increase to 0.91, which is higher than the

Panaroo name	Annotation by Uniprot	Definition	Gene length (bp)	Comment
group_1062	Not found	tyrosine-type recombinase/ integrase	1071	integrase of phage 29485 found in *Salmonella*
ydiO	ydiO	putative BsuMI modification methylase subunit YdiO	1197	
group_716	Not found	phage tail protein	486	found in Klebsiella phage 4LV2017 and *Salmonella* phage RE2010
group_2947	mEp235_051	portal protein/ Uncharacterized protein	270	found in Enterobacteria phage mEp235
group_2909	AMBK_57	DNA replication protein O	279	found in *Salmonella* phage vB_SosS_Oslo

Table 4.21: Annotation of genes found as markers

Position	Gene name	CDS begin	CDS end	Comment
1567699	STM1546 (99% identity)	1566277	1567785	Serves to protect the cell against DNA damage by alkyl hydroperoxides
3213391	tsr	3212787	3214448	Serine sensor receptor / Methyl-accepting chemotaxis protein I
199204	prgH	198059	199237	Type III secretion system inner membrane ring protein PrgH

Table 4.22: Annotation of variants found as markers

combination of 2 genes (Table 4.19). The combination of 3 variants displays a accuracy of 0.95, with 95% sensitivity and 97% specificity (Table 4.20). The 3 variants identified are in position 199204, 1567699 and 3213391 in the genome of *Salmonella* Mbandaka SA20026234. All variants are located in genes not impacted with homologous recombination events. The position of these variants can have a significant impact on the phenotype of the bacteria (Table 4.22), such as the variant at position 199204 located on a type III secretion system protein, impacting directly the invasion of the bacteria [193]. These mutations can be used to investigate the host source of a *S.* Mbandaka genome. Overall, the criteria of sensitivity and specificity are not high enough to use these genes or variants as markers in the field, but they can very well be used after sequencing.

4.4.2.3 Making topological groups to propose diversity traceability

In order to analyse more precisely these host groups, clusters based on the topology of the phylogenomic tree (branch with robust bootstrap) were proposed, containing more than 5 strains. As SNP thresholds are not present in the literature for *Salmonella* Mbandaka, I

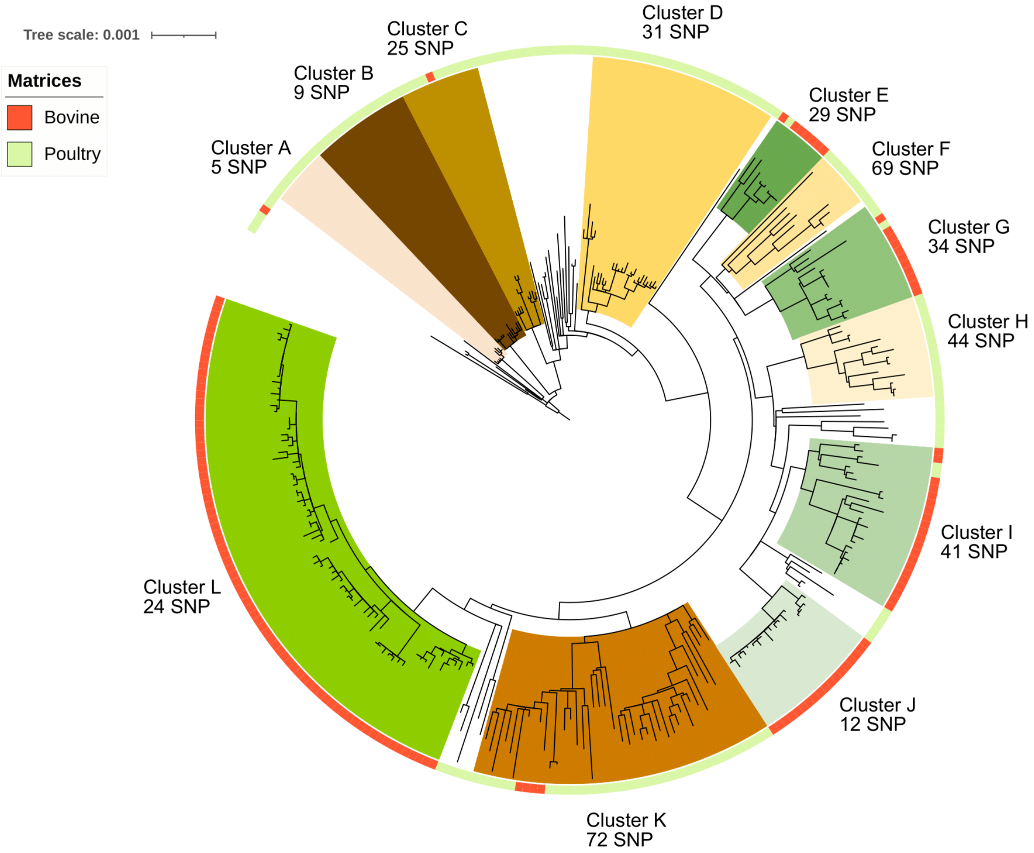


Figure 4.11: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Inner ring corresponds to the matrices. Clusters are colored in green for major bovine clusters, and brown/yellow for major poultry clusters

focused on topology to identify host clusters. Using this methodology, 12 major groups were selected (Figure 4.11). Among these 12 groups, 5 contained 134 bovine genomes over the total 140 bovine genomes analyzed (95%). These 5 groups were characterized by SNP differences comprised between 12 and 34 SNPs. The resting 7 groups contained 118 poultry genomes over the total 164 analyzed poultry genomes (72%). These 7 groups were characterized by SNP differences between 5 and 68 SNPs (Figure 4.11).

Among poultry samples, 40 were not clustered in any groups, while among bovine samples only 1 genome was not clustered (i.e. these unclustered genomes will be called "singleton" below). The maximum SNP difference was 163 SNPs between bovine samples, while it was 214 SNPs between poultry. This result is not surprising, as poultry samples were collected from a larger geographic area than bovine (Figure 7.18).

Using these clusters, it was proposed to identify cluster specific genomic markers in order to propose an alternative to host markers and allow actors in the field to follow the evolution of strains in their sector (a farm or a dairy for example). Using the gene content (Table 4.23), groups A,B,C,E can be identified by one gene, with 100% sensitivity and 100% specificity. The

		Source		Markers for the cluster	Sensitivity	Specificity
Clusters		Bovine	Poultry			
A	Poultry	0	8	group_2040	100%	100%
B	Poultry	0	14	rrrD_1	100%	100%
C	Poultry	1	10	intQ	100%	100%
D	Poultry	0	26	/	ND	ND
E	Bovine	7	1	group_2923	100%	100%
F	Poultry	0	8	group_5614	100%	99.7% (FP 1 Poultry)
G	Bovine	11	3	livJ_1	78% (11/14)	100%
H	Poultry	0	14	/	ND	ND
I	Bovine	21	2	/	ND	ND
J	Bovine	18	0	group_261	100%	99.7% (FP 1 Poultry)
K	Poultry	4	38	/	ND	ND
L	Poultry	77	0	/	ND	ND

Table 4.23: Genes marker analysis of each clusters

groups F and J can be identified with one gene, with 100% sensitivity and 99.7% specificity, due to false positive isolates. The group G can be identified with one gene with 78% sensitivity and 100% specificity, due to 3 true negative isolates. Other groups could not be identified by one unique gene. However, each group can identified by a variant which is unique to the group (Table 4.24). With this result, it is possible to find the origin of the cluster of a strain by looking at a precise position of the genome. This tool can be used for monitoring diversity within a farm (introduction of a new strain), or monitoring in the high risk areas in agri-food sectors.

4.4.2.4 Exploration of wild bird contamination

In this PhD thesis, several hypotheses about contamination sources have been put forward for *Salmonella* Mbandaka in cattle. The first hypothesis concerned cross-contaminations due to the feed products that would be passed on to the bovine farms. This hypothesis was assessed partially, due to the lack of feed strains and inaccessibility to information related to the production and distribution of feed products. The other hypothesis from the local actor in the bovine sector is that the persistence of *S.* Mbandaka in the north-western of France would come from an animal vector which would have transmitted the bacteria to cattle. It is known that *Salmonella* can easily adapt to wild bird [483], and can have a major role in contamination of outdoor herds [370, 484].

The variant calling (SNPs) phylogenomic analysis was carried out as described above with the 9 American genomes and 215 French genomes (section 4.2.2.4). The 215 French genomes were selected to be representative of the different groups identified by the phylogenomic analysis of bovine and avian strains from France (n=100 bovine, n=114 poultry) in section 4.4.2.1.

Phylogenomic reconstruction is displayed in Figure 4.12. American genomes isolated from wild birds clustered with French genomes. One sample (SRR14570238) clustered with a poultry sample isolated in a farm next to the sea (S16LNR3059) and a sample isolated from bovine

		Source		Markers for the cluster	Sensitivity	Specificity
Cluster		Bovine	Poultry			
A	Poultry	0	8	1140003_C_A	100%	100%
B	Poultry	0	14	1125159_C_T	100%	100%
C	Poultry	1	10	679466_A_T	100%	100%
D	Poultry	0	26	243237_G_T	100%	100%
E	Bovine	7	1	713565_C_T	100%	100%
F	Poultry	0	8	149101_T_G	100%	100%
G	Bovine	11	3	887092_A_G	100%	100%
H	Poultry	0	14	706527_T_C	100%	100%
I	Bovine	21	2	87065_C_A	100%	100%
J	Bovine	18	0	35615_C_T	100%	100%
K	Poultry	4	38	423715_C_T	100%	100%
L	Bovine	77	0	41677_T_A	100%	100%

Table 4.24: Variant marker analysis of each clusters

manure (ACT1919846); another sample (SRR13620966) clustered with a poultry sample isolated further inland (S18LNR1879); and 5 wild bird genomes clustered together but very close to poultry strains from the French coast. Among the 5 poultry samples clustered with the 9 wild bird genomes, 2 are from the coast (S16LNR0359, S18LNR1211) and 2 are further inland (S18LNR1879, S18LNR0214) but close to a near river that connects directly to the Atlantic Ocean (S18LNR1879), and last one (2019LSAL01500) where only the department crossed by a river was addressed. The lowest SNP difference (29 SNPs) is between the sample isolated from bovine manure (ACT1919849) and a wild bird sample isolated from an avian from United States (SRR1515029).

Given the geographical distance between the French strains and wild bird, we could expect a large genomic distance or a cluster composed exclusively of wild bird strains. In fact, the genomic distances are high between American wild bird strains, but similar to the genomics distances between poultry strains. Moreover, wild bird strains cluster with genomes isolated in departments bordering the ocean, which reinforces the hypothesis of a epidemiological link between these strains. This hypothesis deserves further investigation but would require access to isolated samples of French birds.

4.4.3 What are the main genetic factors favoring the *Salmonella* persistence in livestock?

In this section, we analysed more precisely the genomic content of *Salmonella* Mbandaka, like antimicrobial agents, virulence genes and others mobile genetic elements which have never been studied in *Salmonella* Mbandaka in France previously. Virulome and accessory genome were analysed to better characterise this serovar and then to identify possible genomic differences between strains isolated from bovine and poultry sectors. This approach explores the presence/absence of virulence and resistance genes compiled in open access databases targeting known virulence factors, *Salmonella* Pathogenicity Island (SPI) genes, multi-drug resistance genes and genes conferring heavy metal or biocide resistances.

Because we have not able to identify markers that could discriminate bovine from poultry sample, probably explained by highly divergent genes, this section will focus on genomic elements

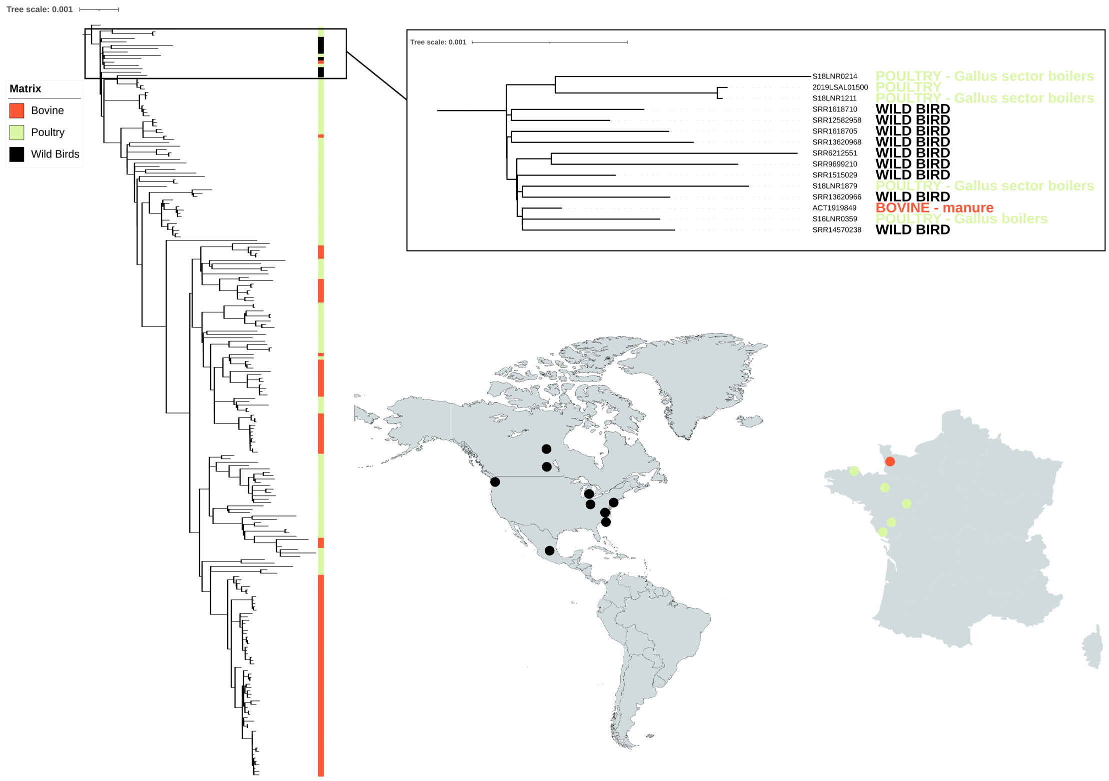


Figure 4.12: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine, poultry and wild birds. Inner ring corresponds to the region of isolation

of interest potentially explaining the persistence of *S.* Mbandaka strains in cattle and poultry.

4.4.3.1 Virulome analysis

The infection of cells by *Salmonella* requires different steps involving a arsenal of virulence factors. After moving to reach targeted host cells through its flagella (e.g. *fliC* and *fljB* genes coding for z10 and e,n,z15 flagellar types), adhesion factors (e.g. fimbriae and adhesins present on its surface) allow solid attachment to the surface of host cells. Among the considered samples, 60 virulence genes have been detected. Among them, gene clusters *csg* [485], *fim*, *bcf* and *lpf* [429], coding for both Curli and chaperonne/placier type fimbriae, were detected in all considered French strains. In addition, *misL* [486] and *mgtCB* genes coding for adhesins were also observed in the bovine and poultry samples.

The SPI-1, SPI-2, SPI-4 and SPI-9 coding for type III (T3SS-1 and T3SS-2) and type I secretory apparatus (T1SS) responsible for survival and proliferation in various intracellular environments, were identified in the genomes (Figure 7.20).

SPI-1 genes, encoding for T3SS-1 apparatus required for invasion of intestinal epithelium cells, such as *inv*, *prg*, *org* gene clusters and *sicA*, *sipABCD*, *sspABC*, *sopB*, *sopE*, *sopE2* regulator/effector genes, were identified in all genomes. In the same way, SPI-2 genes encoding components of the T3SS-2 apparatus, required for bacterial virulence and proliferation in macrophages, such as *ssa*, *ssc*, *sse* gene clusters and *pipB*, *pipB2*, *soxS* regulator genes, were also identified within all French genomes. Interestingly, *siiA/B/C/D/E/F* SPI-4 gene cluster, encoding for a type I secretion apparatus that contributes to the colonization of the bovine intestines, was also observed in the dataset with the exception of 5 out the 164 poultry genomes.

Even if SPI-9 was identified by SPIFinder database blast analysis in all genomes, important operon such as STY2876, STY2877, STY2878 and STY2875 were not displayed by VFDB blast analysis. This incoherence could be explained by the fact that the open reading frames STY2876, STY2877 and STY2878 present 98% identity with type 1 secretory apparatus (T1SS), and therefore not identified by blast. Further analyses would valid or invalid this hypothesis.

4.4.3.2 Biocide and heavy metal resistance analysis

Six genes implied in biocide resistance activity were identified. Among these 6 genes, 3 (*yddG*, *baeR* and *baeS* [435]), are described in the literature as involved in methyl viologen dichloride hydrate (paraquat herbicide) resistance [436] (See Appendix 7.4). The other 3 genes (*sodA*, *rpoS* and *smvA*) code for peroxide [437], hydrogen peroxide – monochloramine and cation biocide resistance [438], respectively. These components can be used in bovine industry to rinse the claw sleeves or that clean dairy cow's udder [487, 488], or in water disinfection [489]. Metal resistance genes detected in all samples allow strain survival in present of metals like copper (*CUEP*), gold (*golS/T*) or arsenic (*PSTB*), components found in cattle and poultry farm, especially in food [490, 491] and water [492] (See Appendix 7.4). Also, 2 of the metal resistance genes are annotated as multi-resistance. The *mgtA* gene induces cyclohexane resistance and mediates magnesium influx to the cytosol [442], while the *corA/B/C/D* gene lead to magnesium and cobalt resistance [440, 441]. Finally, 8 genes involved in multi-compound resistance were identified, like *cuiD* which exhibits a copper sensitive phenotype as well as a normal resistance to other metal ions [493], or *AcrD* which is involved in heavy metal efflux and encodes components of a resistance-nodulation-division family transporter that actively transport aminoglycoside drugs out of the bacterial cell.

4.4.3.3 Plasmid and antibiotic resistance analysis

In total, 36 mobilizable plasmids and 23 conjugative plasmids were identified. Only 15 plasmids are contained in more than 4 samples. 159 non-mobilizable plasmids have also been identified. Overall, no plasmids has been identified in all samples from one host or one group. Most of plasmids has been identified in one or two genomes. The plasmid p15ODMR present in ten poultry strains of the genomic group D carring multi drug resistance genes (*blaTEM-1b*), sulfonamide (*sul2*), streptomycin (*aph(6)-Id*), tetracycline (*tet(A)*) and trimethoprim (*dfrA14*), was characterized as non-mobilisable. Interestingly, the non-mobilizable plasmid p-F219, previously associated with epidemic multi-drug-resistance strains of the *S.* Infantis isolated from a small farm in the southern region of Peru in 2017 [494], has been identified in 285 genomes of our panel (122 bovines and 163 poultry) and the conjugative plasmid pSLU-1913 previously described in the *S.* Montevideo strains has been identified in 27 strains of our panel [495].

In parallel, genes involved in antimicrobial resistance (AMR) genes were identified in all samples, including antibiotic families belonging to fluoroquinolone (*gyrB*, *parC/E*) and rifampin (*rposB*). Additional AMR genes related to sulfonamide (*sul2*), beta-lactam (*tem*) and tetracycline (*tetA/B*) have also been identified for a cluster of samples at the top of the phylogenomic tree in Figure 7.21, which corresponds to the presence of the p150DMR plasmid.

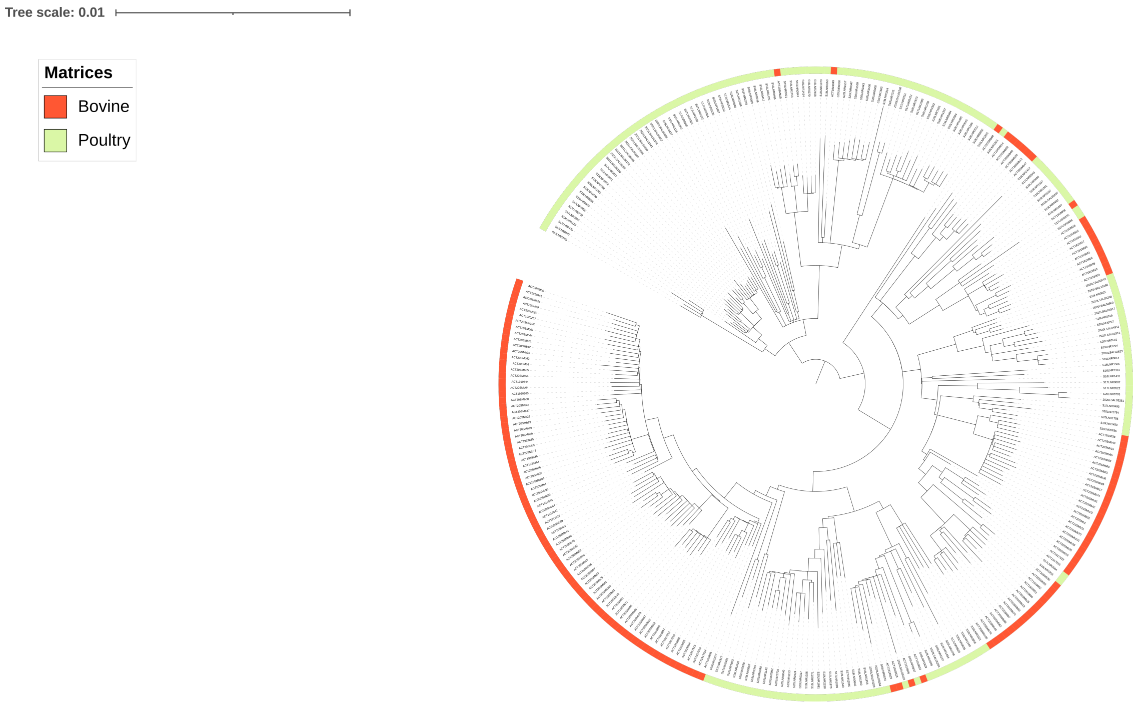


Figure 4.13: pgSNP on *Salmonella* Mbandaka dataset from bovine and poultry samples (n=304). The outer ring corresponds to the sample host.

4.4.4 Does the accessory genome have an impact on the topology and links between strains?

As described above (paragraph 4.4.2), *Salmonella* Mbandaka presents a quite different genetic diversity compared to TMV and *Salmonella* Dublin (section 4.5.2), with longer branch lengths and many genomics differences between strains from the same source. Despite this high genomics diversity, host specific of some topological clusters has been observed and consequently may allow identification of clustered host specific genomics markers. However, even if some specific genomics diversity has been observed in some clusters, the accessory genome has not been taken into account and could play a significant role in the host adaptation [496, 497]. To fully understand the genomic diversity of *Salmonella* Mbandaka and related host adaptation (i.e. bovine and poultry), the developed pgSNP pipeline with default parameters allowed analysis of the phylogenomic effect of the accessory genome.

4.4.4.1 *Salmonella* Mbandaka genomics differences are even more frequent in accessory sequences

The pangenomic tree in figure 4.13 displayed the same clusters than the coregenome tree 4.4.2.3. The branch length is surprising, because other pangenomes phylogenomic inferences did not harbored such long branches so far. The total reference pangenome alignment length is 6170413 bases, where 4.7Mb corresponds to coregenome, which roughly corresponds to the size of the Mbandaka reference genome used in the coregenome-based inferences. Around 1.2Mb corresponds to accessory genome shared by less than 50% of strains, suggesting that most of accessory genome has been acquired recently due to local adaptation to specific stress or environment.

In comparison to the coregenome SNP tree (Figure 4.14), the addition of accessory genome did

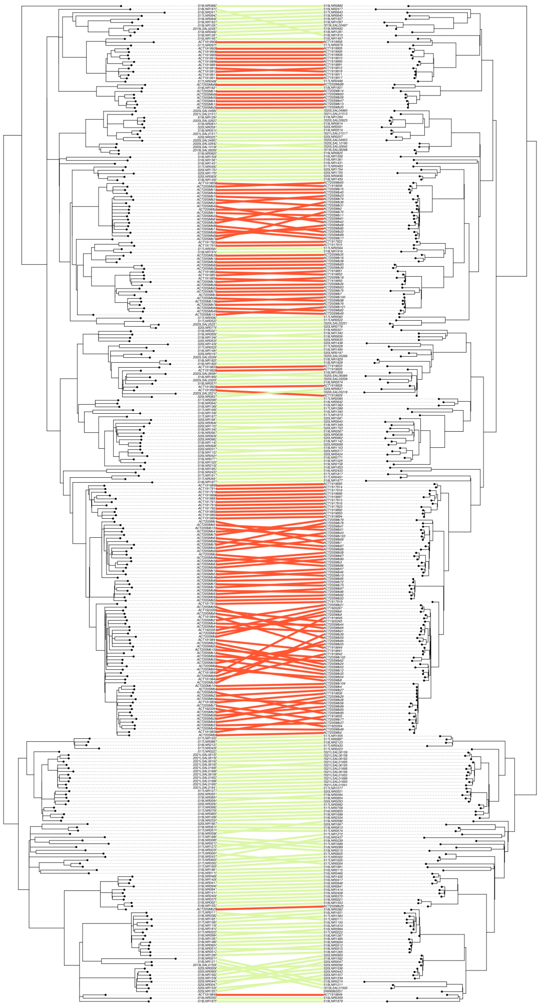


Figure 4.14: Comparison of pgSNP tree and iVARCALL2 tree on *Salmonella* Mbandaka dataset. Left tree is pgSNP tree, right tree is coregenome SNP tree. Branches are colored according to the host sample.

not impact substantially the topology. As discussed previously, *Salmonella* Mbandaka presented high coregenome divergence between samples, and adding the accessory genome increased branch distances, but did not produce news reconciliation of branches. We only observed few new reconciliations between bovines samples that harbored low pairwise coregenome differences.

4.4.4.2 Using pgSNP for genomic and annotation analysis

We investigated the content of accessory genomes and more particularly elements which can lead to a prevalence of *Salmonella* Mbandaka in farms, such as phages and plasmids [103, 317]. Some phages and plasmids identified in the considered *S.* Mbandaka genome collection have not been previously identified in this serovar, in favor of other serovars isolated from the same hosts.

For example, a unnamed plasmid gene has been identified in 55 samples of *Salmonella* Mbandaka bovine and poultry that carries unnamed genes previously identified in *Salmonella* Kentucky, serovar often isolated from poultry host, and others *Salmonella* from poultry indus-

Blocks	Number of samples
0kb-17kb	59
17kb-20kb	1
20kb-23kb	52
23kb-29kb	4
29kb-35kb	1
35kb-53kb	66
53kb-68kb	67
68kb-71kb	18
71kb-96kb	59

Table 4.25: Description of the plasmid p12-4374

try [498, 499]. This plasmid is contained in *Salmonella* Mbandaka from bovine and poultry, which could support the hypothesis that this plasmid is not host dependent, even if it has been identified rather in poultry serovars. Also, in a more global way, it could also suggest that *S.* Mbandaka samples in poultry are passed on to cattle.

One surprising results is the identification of a 96kb plasmid in 67 isolates that present structural variation all along the sequence. This plasmid is annotated as *Salmonella* Heidelberg p12-4374 plasmid [500] from poultry with 92% coverage. Structure of this plasmid among the 70 genomes is interesting and separated in blocs. Only one isolate harbored the entire plasmid, while others blocks were missing in some samples. Blocs are simply described in table 4.25 (and described more precisely in appendix section 7.24).
Interestingly, the 8% coverage not found in p12-4374 plasmid corresponds to 2 small blocks from 29kb-35kb and 68kb-71kb, blocks with low coverage in the plasmid alignment (in Supplementary Figure 7.24). The phylogenomic subtree did not reflect any host pattern (in Supplementary Figure 7.23). Otherwise, some clusters of isolates in the subtrees were also in concordance with the main pgSNP tree, for example 13 bovines samples are clustered together in both trees and displayed few differences.

Some large plasmids like pSTM2 (97kb, contained in 70 samples) and pGD27-62 (50kb) were detected in some samples which were localized all around the tree without any specific SNP pattern. On the other hand, smaller plasmids were also detected. For example a small 2kb mobilizable plasmid was found in 61 isolates, and another of 1.5kb was identified in 7 isolates. These two small plasmids carred genes predicted as replication protein which plays an important role in DNA replication and homologous recombination [501]. Overall, 17 contigs were identified as mobilizable or conjugative plasmids in between 4 to 70 isolates, showing an important exchange between *Salmonella* Mbandaka and its surrounding microbiota.

8 phages were identified among the *S.* Mbandaka genomes of out panel (Supplementary figure 7.22). One of them, the *Salmonella* phage PSP3 is detected on all samples. This phage belonging to the P2-like phage family and is widely spread within other *Salmonella* serovars [502]. The others phages identified were displayed at table 4.26.

All these phages are known among *Salmonella* serovars, but some of them such as phiV10 have a high number of genes in common with *Escherichia coli* specific phages [503]. Interestingly, the three phages phiV10, SEN34 and ENT39118 were observed only in the genomes of strains isolated from poultry sector with the exception of a strain (ACT20SMb25 presenting the phage

Phage name	Number of samples
PSP3	304
Fels2b	207
pro483	161
ECP1	148
4LV2017	41
phiV10	29
SEN34	14
ENT39118	11

Table 4.26: Phage presence in *Salmonella* Mbandaka dataset

ENT39118) that has been isolated from animal feed in a cattle farm. We have not observed a geographic repartition of the phages within the French genome analysed however some of the groups identified were characterised by specific phage profiles such as the groups A, D and E characterised by the phage PSP3 and Fels2, the group B characterised by the phages PSP3, ECP1, phiV10 and SEN34 or the group F characterised by the phages PSP3, Fels2 and pro483.

The structure of some of these phages were conserved all along the contig across all samples, like PSP3 and ENT39118. In contrast, some phages were partially conserved in some isolates. For example, 13 samples harbored the entire ECP1 phage. The structure of the phage was described in Figure 7.16. All bovines samples except one only harbored a potential phage tail site, while some poultry samples had the tail site, the fiber protein, some hypothetical protein and portal proteins. Otherwise, 29 poultry samples had additional coat protein or hypothetical protein, and 13 of them have the terminase site. Some of phages studied in Mbandaka [504] were not detected in our dataset. This could be explained by the difference of annotation between some authors. In contrast, we identify the P2 phage in agreement with [317].

4.4.4.3 What including accessory sequences brings to *Salmonella* Mbandaka genomics studies.

In conclusion, the analysis of a small dataset of Mbandaka strains isolated from a restricted geographical area showed that this serovar displayed a high genomics diversity. The analysis conducted with pgSNP led to a higher resolution than the coregenome-based inference. Thanks to accessory genome, it has been demonstrated that this serovar harbor diversified genomic patterns from what is known in other bovine serovars. For example, *Salmonella* Dublin is highly clonal and the geographical distance is a major factor in the genomic divergence for this serovar [401]. In addition, *Salmonella* Mbandaka harbored long branch length, often linked to higher mutation rate and/or many homologous or non-homologous horizontal transfers. This diversity is surprising given the small studied geographical area, and would require further genomics study to understand the high prevalence of this serovar in north-western France.

4.4.5 Discussion and limits of the study of *Salmonella* Mbandaka

4.4.5.1 Segregation of *Salmonella* samples from bovine and poultry

In this section, a comparison between *Salmonella* Mbandaka from bovine and poultry host was performed. To further assess the hypothesis about the genomic distinction between poultry and bovine samples, it would be necessary to take also into account another host, like pork, where *Salmonella* Mbandaka is less prevalent [267, 505]. In addition, geographical origins must also be taken into account to be able to prove that hosts are associated with the genonic background. In a *Salmonella* recent paper [103], the authors described that they could not find any specific pattern in the *Salmonella* Mbandaka population structure in relation to either geographical origin or isolation source, disproving the hypothesis that host specific clades may be emerging in this serotype, based on a dataset of 403 samples from different sources isolated in different countries. Looking at the thesis from the author [317], some sub-clusters presented geographical pattern (i.e. samples coming from the same country in different hosts), or matrix pattern (i.e. samples coming from the same matrices of isolation). In one cluster, samples from food commodities of Asia, North America and Europe were clustering with human samples. SNP difference was not comparable with our dataset because the authors only selected coregene variations from the concatenated and aligned housekeeping genes. If the theory that *Salmonella* is persistent in its environment and that it spreads all along the chain, then this kind of clusters could be explained by the globalization of products and contamination in humans. Also, some clusters displayed porcine samples linked to animal feed, as hypothesised in our analysis. Another study performed with a larger dataset demonstrated that *S.* Mbandaka strains clustered very closely by geographic source and host, producing fairly even clusters of isolates from cows, plants, dairy, and chicken farms, in agreement with our results on bovine strains and poultry [506]. It also displayed that almost all clades were country-specific, but each country were dispatched in multiple micro-clades. This results has been produced with cgMLST pipeline, so SNP analysis would therefore be necessary to corroborate our results. Overall, further analysis with geographical data may prove some potential links between herds or explain some links between bovine and poultry strains.

In our study, we used isolates from wild birds available on Enterobase. The clustering of the wild birds within poultry and bovine samples, which come from a geographical area close to the sea strongly, validated the hypothesis of terrestrial animal contamination through this avian vector. If the hypothesis that there is no adaptation to the host in *S.* Mbandaka is true, then this hypothesis about contamination with different vectors is even stronger. However, wild bird strains were still clustered among poultry strains, as the bovine strain from the environment did not show a real link with the bovine host. This result could partially support the hypothesis of host discrimination. The complete validation of this hypothesis would require to validate these links with wild bird strains from France. Combined with data from other countries that share the same wild bird diversity (UK, Belgium, Netherlands) it would be possible to demonstrate that wild birds can be a vector of contamination.

4.4.5.2 Persistence and diversity of *Salmonella* Mbandaka in herds

It is very difficult to determine the whole diversity of *S.* Mbandaka in cattle, but compared to the dataset of *Salmonella* Dublin (section 4.5.2), *S.* Mbandaka seems to have a higher diversity (1062 SNPs) even if it the study was restricted to only one region. However, the whole dataset does not allow us to conclude about the overall diversity of the serovar. Matrix specific clusters were not observed, suggesting a continuous contamination all along the food chain, from bovine to cheese. Feed samples are disseminated all around the phylogenomic tree, suggesting a potential way of contamination through food. Finally, isolates from the same geographical area which are disseminate all around the tree were isolated most of the time from milk, suggesting that down-the-line transformation steps are more likely associated to host genomics diversity.

Among 9,488 genes including 3,655 coregenome, genes involved in AMR, virulence and heavy metals resistance did not distinguish geographical or host origins. Most strains displayed the same resistance profile, whether in poultry and cattle, with an exception for 10 multi-resistant poultry strains. The discriminatory power of small variation was higher than genes because higher accuracy was estimated from variants than genes.

Genomics content analysis can provide genetic hypothesis explaining the persistence of *Salmonella* Mbandaka within herds, but remains to be proven with *in vivo* phenotype tests. Indeed, the presence of many known resistance genes have already been tested with other *Salmonella* species, but mutations in promotors and/or regulators, as well as epigenetic siganls (i.e. adenine methylation) [507], can prevent the translation of this gene.

It is worth mentioning that *S.* Mbandaka is more prevalent than *S.* Dublin in the north-western of France, and may have may have been subject to undiscovered specific adaptations to the environment (e.g. climate). Our observations about the wild birds and the arsenal of resistance common to all isolates may explain the persistence because the same *S.* Mbandaka genomes may continuously contaminate the environment.

4.4.5.3 Lack of *Salmonella* Mbandaka genomic information

The main difficulty of the present study was to work with a serovar which has poorly been studied at the genomic scale. In spite of the high prevalence of this serovar in bovine [300, 299], the lack genomics studies may be explained by the rareness of human outbreaks related to *Salmonella* Mbandaka. It was pointed out that *Salmonella* Mbandaka seems to be commonly shed to the environment by livestock rather than being a primary human pathogen [506]. One study focused on human cases of salmonellosis caused by *Salmonella* Mbandaka in Australia, and described two humans clusters where the SNP difference varied between 12 and 82 for one case, and between 10 and 25 SNPs for the other case. This SNP difference is very high considering the threshold proposed by different studies to identified related strain of serovars isolated from cattle like *Salmonella* Dublin [401] or other serovars [252, 508, 509, 510] (between 5 to 15 SNPs). In the present study, the SNP differences were high considering the scale of the study (i.e. same geographic area, low time scale), but in agreement with the high diversity of other genetic elements (i.e. genes and homologous recombinaison events). It seems to be difficult to explain this diversity with the lack of precise geographical metadata. More likely, the bacterium may have been (or may be) under high stress pressures, and developed quickly mutations to adapt to its environment, as it was shown before that genetic changes has been observed on *Salmonella* Mbandaka after the exposure to heat treatment (up to 19 SNPs after 10 cycles in [87]). This high number of SNPs could also been explained by a higher mutation rate, as it was observed that some serovar have a higher mutation rate than others [85, 86].

4.4.6 Conclusion of *Salmonella* Mbandaka in bovine sector

In conclusion, this study has brought to light news aspects of *S.* Mbandaka genomic biodiversity and provided a first overview of the epidemiological dynamic of this serovar in the farms of northern-west France. Phylogenomic clusters were not associated with matrices of isolations, suggesting a continuous contamination all along the food chain, from bovine to cheese, and in poultry herds. Despite this heterogeneity, our results revealed a degree of adaptation to bovine and avian hosts, with clusters more adapted to one or the other host. In addition, we have shown that there was a high probability of contamination between hosts, whether through environmental bovine strains that cluster with poultry, or also through another vector such as wild birds.

In this study, we investigated genomics markers to distinguish hosts of *S.* Mbandaka. I was able to develop a pipeline to consider genes and variants combination and also identify unique markers from the studied strain clusters. This opened up a possibility of monitoring the evolution of strains within a farm and between farms in the same department. As the markers studied are conserved in all sequence, it is possible to set up PCR amplification or sequencing protocols [120, 116, 118, 123]. The identified markers can be useful for monitoring in the agri-food sectors, but also in case of TIAC to trace back quickly the pathogenic agents. There is today a need to characterize strains more and more precisely for monitoring throughout the chain. With additional metadata, it would have been possible to study more precisely the hypothesis about geographic persistence, and also the track of contamination in foods.

Analyses the genomic content of *Salmonella* Mbandaka demonstrated that the bacteria can adapt to its environment, explaining its persistence in herds. *Salmonella* Mbandaka presented an important arsenal of virulence genes which may provide with a higher pathogenicity and adhesion to its host. At least 4 SPI was identified, with lingering doubts for the presence of SPI-9. These SPIs are responsible for virulence and pathogenicity through attaching, invading, surviving, and bypassing the host's defense mechanisms. Biocides resistance was identified in all genes, which may provide bacterium with abilities to survive longer in its environment. Heavy metal resistance screening displayed also the bacteria's ability to adapt to its environment. A large panel of AMR genes was detected, including fluoroquinolone resistant genes that may compromise human treatment options [220]. A multi AMR plasmid was identified in poultry strains that carried important AMR genes. Many AMR genes were carried by plasmids, which appeared a key element of the AMR spreading between bacterial lineages. I also investigated *FimH* gene to analysis whether a host-specific adaptation exists in the flagellar of *Salmonella* Mbandaka (Section Annexe 6.2.1), and I showed that this flagellar did not distinguish host origins of *Salmonella* Mbandaka.

Overall, this study is the first characterization study of *Salmonella* Mbandaka strains from the dairy and poultry sectors in France, and provides new hypotheses and an overview of the diversity of this serovar.

4.5 Results: Assessing the genomic diversity of *Salmonella* Dublin in France in bovine industry

4.5.1 Context of the study

In January 2016, FNRC-ESS and Pasteur Institute in Paris, reported to Santé publique France, the national public health agency, an increase of *Salmonella* Dublin infections across the country. Epidemiological and microbiological investigations were published [301] pointing two types of raw-milk cheeses as vehicles of the *S.* Dublin outbreak at that time, but the causes and sources of contamination had not been identified. To better understand the dissemination and circulation of strains in the two incriminated regions, a working group was formed to perform a retrospective study using samples isolated from different matrices of the two regions (i.e. cow, milk, cheese, processing plant and human).

The goal of this retrospective genomics analysis of 480 *Salmonella* Dublin samples (section

4.2.3) was to characterise the overall genomics diversity of this serovar in two regions, and contextualised with geographical and temporal links to actual outbreak events. Phylogeographic and coregenome approaches were developed in this study, with an extra attention given on the dataset due to the sensitive data.

With this study, I displayed the precision brought by WGS methods for the identification of different clusters and unknowmn links between samples. I also demonstrated that the geographical distance is a major factor in the genomic divergence for *Salmonella* Dublin concerning the early stages of the production processes (i.e. animals, farms), while downstream transformation steps are more likely to harbour genomic diversity. This study was a good pillar to understand bacterial genomics in a targeted region, and also allowed me to develop coregenome SNP and phylogeographic methods that can be applied to other serovars of *Salmonella*. These findings also pave the way toward the development of news comparative tools integrating others sources of variation as a discriminative metric along with SNPs, such as InDels, structural variations, mobilome and accessory genome contingency. On a side note, this study was also a demonstration of the possibility of using WGS in outbreaks and genomics analyses of *Salmonella* for food safety actors. Altogether, the present results brought an insight on regional genomic diversity of highly related genomes involved in foodborne outbreaks, underlining the necessity to drive investigations toward the most resolutive comparative genomics methods.

4.5.2 Journal article

https://doi.org/10.1093/nargab/lqac047

A retrospective and regional approach assessing the genomic diversity of *Salmonella* Dublin

Madeleine De Sousa Violante[1,2], Gaëtan Podeur[1], Valérie Michel[1], Laurent Guillier[3], Nicolas Radomski[4], Renaud Lailler[3], Simon Le Hello[5], François-Xavier Weill[6], Michel-Yves Mistou[2] and Ludovic Mallet[7,*]

[1]Actalia, 419 route des champs laitiers, CS 50030, 74801 La Roche sur Foron, France, [2]INRAE, MaIAGE, Université Paris-Saclay, F-78352 Jouy-en-Josas, France, [3]ANSES, 14 Rue Pierre et Marie Curie, 94700 Maisons-Alfort, France, [4]Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise 'Giuseppe Caporale' (IZSAM), via Campo Boario, 64100 Teramo, TE, Italy, [5]UNICAEN, Groupe de Recherche sur l'Adaptation Microbienne, GRAM 2.0, EA2656, University of Caen Normandy, Caen, France, [6]Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Centre National de Référence des Escherichia coli, Shigella et Salmonella, Paris, France and [7]Institut Claudius Regaud, 1 avenue Irène Joliot-Curie, 31059 Toulouse Cedex 9, France

Received September 24, 2021; Revised May 30, 2022; Editorial Decision June 07, 2022; Accepted June 13, 2022

ABSTRACT

From a historically rare serotype, *Salmonella enterica* subsp. *enterica* Dublin slowly became one of the most prevalent *Salmonella* in cattle and raw milk cheese in some regions of France. We present a retrospective genomic analysis of 480 *S.* Dublin isolates to address the context, evolutionary dynamics, local diversity and the genesis processes of regional *S.* Dublin outbreaks events between 2015 and 2017. Samples were clustered and assessed for correlation against metadata including isolation date, isolation matrices, geographical origin and epidemiological hypotheses. Significant findings can be drawn from this work. We found that the geographical distance was a major factor explaining genetic groups in the early stages of the cheese production processes (animals, farms) while down-the-line transformation steps were more likely to host genomic diversity. This supports the hypothesis of a generalised local persistence of strains from animal to finished products, with occasional migration. We also observed that the bacterial surveillance is representative of diversity, while targeted investigations without genomics evidence often included unrelated isolates. Combining both approaches in phylogeography methods allows a better representation of the dynamics, of outbreaks.

INTRODUCTION

Salmonella is one of the most common bacterial pathogens worldwide in human and animal infection (1). The most frequent *Salmonella* subspecies is *Salmonella enterica* subsp. *enterica*, which is one of the four major causes of diarrheal diseases worldwide. Gastroenteritis cases due to Non-typhoidal *Salmonella* were estimated to 153 million annually, including 56 000 deaths (1). Salmonellosis is the second most frequently reported bacteriologically related zoonosis in many European countries (2). The majority of salmonellosis cases cannot be associated with outbreaks and are classified as sporadic cases (3). *Salmonella* exhibits a highly variable host range among animals, especially in mammals and colonizes the gut of various livestock such as poultry, swine or cattle (4). The outcome of infection depends on the pathogenic genotype and *Salmonella* host, ranging from asymptomatic carriage to diverse stages of gastric disorders and in certain cases, evolving into potentially fatal pathogenic conditions.

Salmonella enterica subsp. *enterica* serotype Dublin is a host-adapted strain, found especially in cattle farms (5). *S.* Dublin was historically a rare serotype, but slowly became one of the most prevalent *Salmonella* serotype in cattle and cow's raw milk cheese (6,7). Since 2000, *S.* Dublin is often found in the top 20 of most prevalent serotypes at the French National Reference centre for *Escherichia coli*, *Shigella* and *Salmonella* (FNRC-ESS), and has been a persistent cause of human infections for 20 years (8). In France, processes involved in the production of raw milk cheese, uncooked pressed cheese, semi-cooked cheese or soft cheese may prevent the milk to reach a temperature high enough to kill *Salmonella* (9). Beyond the considerable economic losses, contaminated raw milk or finished products infected

[*]To whom correspondence should be addressed. Email: mallet.ludovic@iuct-oncopole.fr

by carrier cows can cause severe infections in humans (5) and dairy cattle (10). Although diarrhoea is a common consequence of *Salmonella* infections in cattle, the consequences of *S*. Dublin infections commonly reach respiratory syndromes in calves or abortion in gravid cattle (11). *S*. Dublin infections can produce long-term asymptomatic carriers that can periodically shed bacteria in the environment, contributing to the propagation within herds (12), or to humans through direct contact or consumption of contaminated products.

The prevalence of *S*. Dublin in cattle could be explained by a diversity of factors: the bacterial ability to survive in the environment, the symptomatic carriage of individuals, the intermixing of cattle and their exchanges between farms, contaminated food and other factors (13–15). Unfortunately, serotyping or epidemiological data are not sufficient to describe fully contamination links, especially for outbreak events that spans over several months. However, it is possible to trace links between strains at the genomic level which supports hypotheses on the spread, the routes of contamination and history of outbreaks (16).

Diagnostic of *S*. Dublin is commonly based on serotyping (17) and more recently on Whole Genome Sequencing (WGS) methods which have been implemented in many studies and laboratories to improve outbreak resolution (18–20). This method has shown an enhanced cluster detection, an improved resolution and a more accurate result in comparison to laboratory methods (PFGE, MLVA) usually applied to characterize *Salmonella* (21). When performing WGS-based investigations of outbreaks, phylogenomic history is usually reconstructed with core genome point mutations and based on evolution models (22,23). As recently resumed (24), these stochastic point mutations correspond to single nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELs) induced by replication errors or damage of DNA. Point mutation-based phylogenomic reconstructions can be biased when the bacterial genomes are impacted by recombination events (25,26), such as the replacement or inversion of similar sequences (27) (i.e. homologous recombination events), or new genetic material from exogenous genome (i.e. non-homologous recombination events). *S*. Dublin is well known to be impacted by recombination events (28,29) that can induce biases when looking at closely related isolates.

In January 2016, FNRC-ESS, Institut Pasteur, Paris, reported to Santé publique France, the national public health agency, an increase of *Salmonella* Dublin infections across the country. After an epidemiological and microbiological investigation with the help of WGS data, pointed to two types of raw-milk cheeses as vehicles of the *S*. Dublin outbreak at that time (30), a working group was formed to perform a retrospective study focusing on this serotype in the most affected regions. Thanks to the extensive epidemiological and microbiological investigations at the time, a large dataset of strains was collected between 2015 and 2017 in two regions producing the incriminated cheese. The study was designed to understand the circulation of strains and to improve the overall surveillance of *S*. Dublin through whole genome analysis.

We present a retrospective genomic analysis of 480 *S*. Dublin isolates from collections, field investigations, controls and surveillance plans spanning the production process steps from fields to products, all characterised with a set of minimal metadata and contextualised with geographical and temporal links to actual outbreak events. We crossed the phylogenetic signal with metadata to depict and characterize a region-wide *S*. Dublin diversity and strain dynamics over several years to unravel the genesis of outbreak events.

We investigated the impact of homologous recombination events on phylogenomic reconstructions accuracy (i), the correlation of phylogenomic clustering with isolation date, isolation matrices, geographical origin and epidemiological hypotheses (ii), as well as the potential epidemiological relationships with a WGS-based phylogeographic analysis (iii). We highlighted the benefit of WGS approach on closely related isolates and demonstrated how the results from large datasets can help to understand the contamination dynamics of strains, with the objective of supporting sanitary and health authorities to design appropriate safety policies.

MATERIALS AND METHODS

Data availability, privacy and anonymization

To comply with statistical, geographical and data privacy, metadata presented in the manuscript were transparently anonymized. Geographical data were anonymized in order to avoid disclosing identity from sparse point distribution, through jittering coordinates with uniform distributions computed separately for longitude and latitude and with amplitudes ranging from −0.2 to +0.2. To improve readability in high-density area, an additional uniform distribution jitter was added, proportional to the point density computed on all coordinates with an amplitude of up to +0.2. All raw sequencing materials used in the manuscript are publicly available (Bioproject: PRJNA737646).

Collection of samples and related metadata

In this study, four laboratories from the public and private sectors collected 2249 strains of *S*. Dublin between 2015 and 2017 (31). Samples were collected all along the production stages from cows to finished products. After harmonization of metadata, 2101 strains of *S*. Dublin were sub-sampled using the Gower Algorithm (31) in order to build a representative collection of samples as described below. The laboratory for food safety in Maisons-Alfort from ANSES provided 77 samples from its collections of strains, through the French food-chain surveillance (SCA) platform and the biennial 'National Salmonella Surveillance plan' 2015 and 2017 (www.plateforme-sca.fr). One sample was collected in December 2014, and was added due to temporality proximity with the dataset.

Targeted samples (selections A, C, D, E) linked to cases of salmonellosis under local epidemiological and microbiological investigations were added in the study. The samples of the selection A are restricted to a single dairy farm where strong clinical signs of salmonellosis cases in cows were repeatedly observed over the years. One sample from 2009 and one sample from 2010 were added to investigate the persistence of the strain within the herd. The samples of the selection C come from a limited geographic area, where contamination from cattle to cheese was found. The samples of the selection D correspond to a period in a restricted geographic area with a large number of milk samples contaminated with *S*. Dublin. Finally, the samples of the selection E were isolated from cattle in a restricted area.

Strains selection

In order to downsample the dataset of isolates representing the *S*. Dublin diversity, we used a previously described method (31) leveraging available metadata (i.e. year sample, region, production stage and type). Based on Gower coefficient (GC), the distance between two units is the sum of all the variable-specific distances associated to the metadata, whose attributes have a mixed of categorical and numerical values. Each variable can have a weight and consequently change the importance of each metadata class. First, dissimilarity matrix between samples is computed using Gower's distance. Then, hierarchical clustering is applied on the dissimilarity matrix to cluster samples. Finally, the 'silhouette' plot is displayed, measuring how close each point in one cluster is to the points in the neighbouring cluster (Supplementary Figure S1-A). The script is available on https://github.com/lguillier/LISTADAPT/tree/master/metadata2assocation. Out of the 2101 samples, 398 samples were drawn from a random selection weighted to balance Gower clusters representation and maximize diversity of sampling (Supplementary Figure S1-B). Finally, 104-targeted samples from selections A, C, D, E were added to the dataset.

DNA extraction and sequencing

S. Dublin strains were isolated and grown on *Salmonella*-selective media (XLD or BHI) and the genomic DNA was extracted using the 'KingFisherTM Duo Prime' protocol. Then, the quantity, purity and integrity of DNA samples were assessed using a Qubit, a Nanodrop and electrophoresis migrations on agarose gels, respectively. Next generation sequencing (NGS) was performed by the 'Institut du Cerveau et de la Moelle Épinière' (ICM – Hôpital Pitié Salpêtrière, Paris). More precisely, the NGS libraries were prepared using the Nextera XT DNA Library Prep Kit and paired-end sequenced (2 × 150 bases) with an Illumina NextSeq500 sequencer.

Other studied genomes

Human samples raw reads from a previous study (30) were downloaded from the Sequence Read Archive (SRA) (32) (Bioproject: PRJEB28817).

Sequence assembly

The assembly was performed with ARTwork, a freely available workflow developed by the team GAMeR at ANSES (https://github.com/afelten-Anses/ARtWORK). In summary, ARTwork estimates the coverage of reads dependently of the LT2 reference genome (bbmap (33)), normalizes the reads (bbnorm (34)), controls the quality of the reads (fastqc (https://github.com/s-andrews/FastQC)) and trims them (35). Then, *de novo* assembly is performed with SPAdes (36), PubMLST scheme is detected by MSLT (https://github.com/tseemann/mlst) and closest reference is retrieved using Mash (37). Finally, scaffolding is performed with Medusa (38), gap filling is done with GapCloser (39), contigs are trimmed with Biopython (40) and an assembly synthesis is carried out with QUAST (41). Three samples displayed sequencing errors and could not be assembled.

Quality assessment

Quality control was systematically performed and subsequent assemblies failing to meet a set of highly stringent rules were discarded. We rejected samples matching any of the following criteria: more than 1 000 000 assembled bases unaligned to the reference, less than 4 000 000 assembled bases aligned to the reference, >2 INDELs per 100 kb, <80% of assembled bases with 30× coverage, absence of the genome fraction estimation computed by QUAST, or assembly fragmented into >200 contigs. Potential inter- and intra-genus contaminations were detected using Confindr (42) based on assembly metrics and blast respectively. Samples with inter- or intra-genus contamination according to the default Confindr parameters (samples with multiple genera found in the Mash screen step or more than two single nucleotide variants (SNVs) in ribosomal genes) were discarded from the study. Finally, sample serotyping was performed *in silico* based on the assembled genomes using SeqSero (43). Unless conflicting or with reasonable doubt on the error source (metadata, low coverage, etc.), lab-typed and predicted serotypes other than *Salmonella enterica* subsp. *enterica* serotype Dublin have been discarded.

To extend the quality check beyond the sample-specific metrics, we enforced a dataset-wide two-factor (breadth of coverage × depth of coverage) criterion. In a large-dataset context, each genome with its own depth of coverage variation along its sequence exhibits segments with low coverage, thus locally preventing a sound SNP calling. When adding up samples in a dataset, the number of regions where the minimal depth of coverage is not met for at least one of the samples steadily increases. This figure rises sharply when poor quality samples are included in the data set, drastically reducing the breadth of sequence actually used for the phylogeny reconstruction and isolate discrimination. To ensure data consistency and maximize callable SNP positions in a core genome SNPs (cgSNP) analysis context, we rejected high core-losing samples as identified by iteratively comparing depth of coverage drawn from N-samples to that drawn from N-1 samples. Depth of coverage was calculated for each isolate using samtools depth (44). We implemented this method using an in-house filter that keeps each sample containing more than 4.0M positions covered at more than 30×, altogether resulting in a dataset-wide >30×, with a core genome size estimated at 3.8 Mb.

cgSNP caller and phylogenetic inference

The cgSNP were detected using iVARCall2 (45) which maps (BWA (46)) trimmed reads (Trimmomatic (35)) on the *Salmonella* Dublin CT_02021853 reference, sort reads (Samtools), remove duplicates of mapped reads (Picard

(47)), and realign reads around INDELs before detecting variants with HaplotypeCaller from the Genome Analysis ToolKit (GATK) (48). Pseudogenomes have been reconstituted as previously described (45,47). Variants from homologous recombination events were excluded using ClonalFrameML (27). Phylogenetic inferences for both trees were performed by IQ-TREE (49). Core genome SNP-based phylogenomic excluding SNPs from homologous recombination events is unrooted, following an evolutionary model K3Pu + F + I. The consensus tree displayed in Figure 1 was obtained after convergence at 103 iterations with an optimal log-likehood of −6724901.

Homologous recombination filtering

Recombination tracks were identified using ClonalFrameML (27) with the following parameters set to true: -em, -guess_initial_m, -use_incompatible_sites, -reconstruct_invariant_sites, -output_filtered. The parameter -emsim was set to 20 and other parameters were kept to their default values. Required inputs were constituted by a multiple sequence alignment and a sample tree produced as follow: a reconstructed pseudogenome sequence was generated individually for each sample by mapping the sequencing reads against the *Salmonella* Dublin CT_02021853 reference genome, calling consensus variants and reporting them back onto the reference sequence. Pseudo-sequences from all the samples were piled up to yield the pseudo multiple sequence alignment. IQTree was subsequently used with default parameters on this multiple alignment to build the primary sample phylogenetic tree. Robustness was tested with IQTree parameters -alrt 1000 and -bb 1000. In order to comply with ClonalFrameML, the primary tree was rooted using a midpoint method as implemented in FigTree (https://github.com/rambaut/figtree/). Following evidences sustaining that phylogenetic inferences relying on a Markov chain model of nucleotide substitutions should only take into account points mutations (22). Although filtering of SNPs from homologous recombination events might induce partial loss of information (45,50), we decided to characterize the impact of recombination filtering (Supplementary Figure S2) and subsequently to discuss the results excluding recombinant variants (Figure 1).

Clustering

rPinecone (51) was used in order to cluster samples, based on a root-to-tip approach with SNP distance relative to ancestral nodes. Given the observed phylogeny, a SNP-scaled tree was generated with pyjar from the rPinecone's main analysis (52), and then a 5 SNP threshold was selected for clustering. A five SNPs threshold is fairly conservative and allows strong assumptions on the links established within clusters, favouring specificity regarding investigated scenarii of reconstruction.

RESULTS

Construction of the genomic dataset

After asserting presence of compulsory metadata and harmonizing values, 2 101 samples of *S.* Dublin were subsampled using the Gower Algorithm down to 398 (Supplementary Table S1, Random, 'RND' throughout the manuscript) (Supplementary Table S1, Gower). Additional samples ($n = 104$) were included to resolve intricate strain detections in unexpected contexts, understand transmission routes and contribute to food production quality standards, 34 of which did not pass filters (Supplementary Table S1, Targeted, containing selections SELEC A + SELEC C + SELEC D + SELEC E). In order to contextualize the regional study with the epidemiological investigation of 2015–2017, paired-reads from Ung *et al.* (30) encompassing samples from food sources and human cases were included in the analysis as well as 77 strains from the ANSES strain collection. Most samples were obtained after 2015, from two regions, and five different matrices: cow, milk, cheese, processing plant and human (Table 1).

After all the assembly and quality assessment steps, a set of 480 *S.* Dublin genomes associated with trusted samples and metadata was constructed. The set was considered as representative of the diversity of *S.* Dublin circulating in the contaminated area over the years 2015–2017 from both the metadata granularity and the clustering/singleton distribution

Analysis of the diversity at the core genome level

All samples from the study are predicted as *Salmonella enterica* subsp. *enterica* serotype Dublin by SeqSero (43), and sequence type (ST) 10 by MLST (https://github.com/tseemann/mlst). According to prior knowledge about the clonal expansion of *S.* Dublin (53,54) and within a limited time and geographical span, mutation rate and recombination events were quite low through the genomes of interest. Altogether, 1041 SNPs were detected along the 480 genomes, of which 17 lay within 8 homologous recombination events spanning a total of 299 bp (6 on leaves, 2 on internal nodes). After exclusion of the variants located in homologous recombination segments, 1024 SNPs remained. The core genome SNP-based maximum likelihood (ML) phylogeny topology was slightly impacted by the removal of homologous recombination variants, with only few minor differences observed between nodes (Supplementary Figure S2).

Phylogenomic reconstruction highlights a regional segregation of *S.* Dublin isolates

The tree excluding the homologous recombination events converged with a commensurate negative likelihood, consistent with the fact that most of the nodes were supported by high bootstrap values (Supplementary Figure S5). Furthermore, the stability of the topology observed upon comparing trees with and without homologous recombination filtering led to establish that the reconstruction based on the cgSNP signal was robust. Two main groups of genomes were defined based on ML inference and confirmed by the sample clustering distance matrix (Supplementary Figure S4), the first one encompassing most of the isolates from the regions 1 (218 out of 272 with region metadata) while the second overall matched region 2 (134 out of 163 with region metadata) (Figure 1, inner circle).

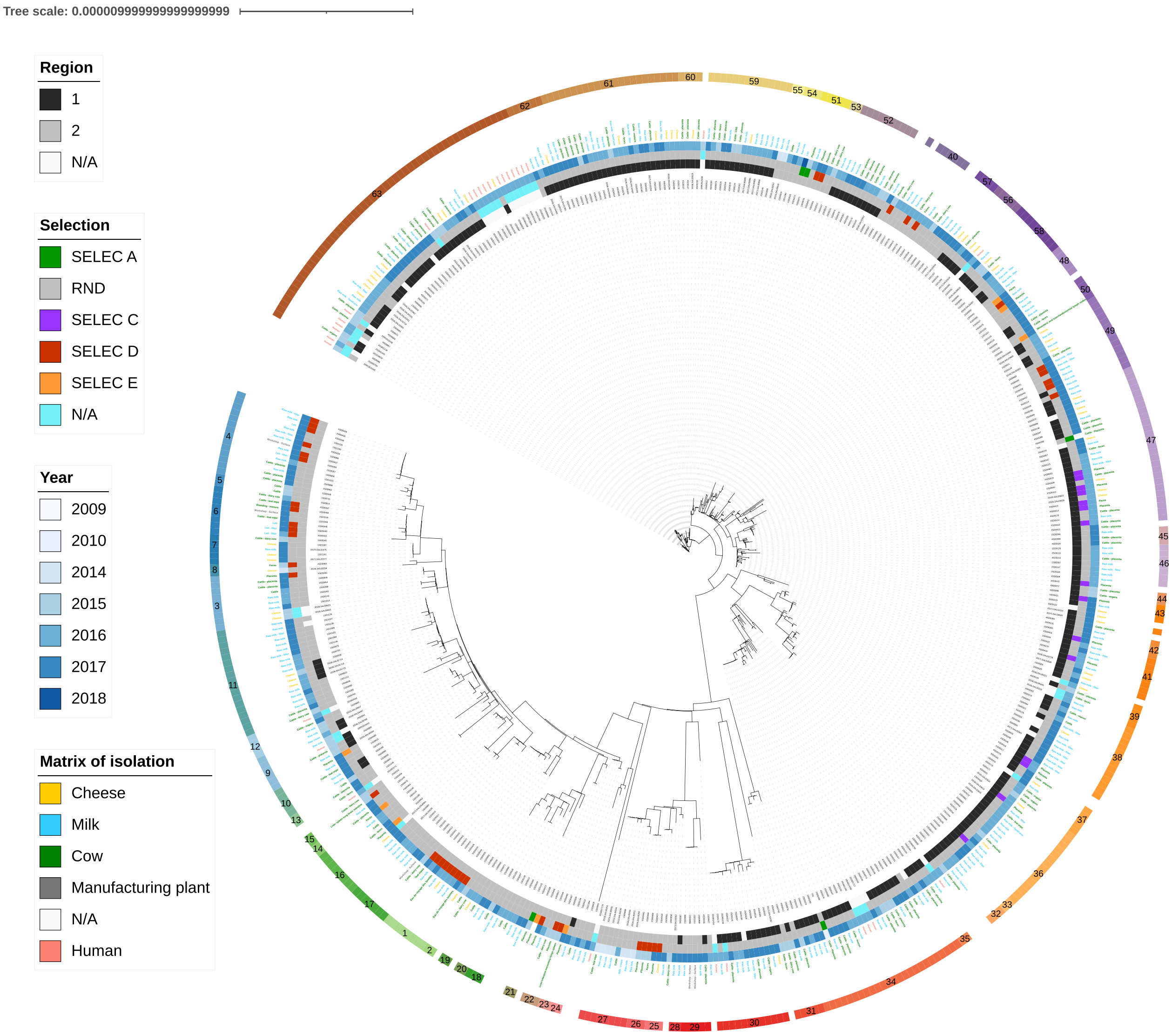


Figure 1. Core genome SNP-based phylogenomic reconstruction by Maximum Likelihood. Reconstruction excluding SNPs from homologous recombination events of *Salmonella* serotype Dublin isolated in two French regions (1 and 2) between 2009 and 2018. Outer ring represents clusters calculated by rPinecone, assigned by numbers and colours. Others rings are represented in the figure and describe isolation year and the isolation matrix. The regions 1 and 2 represent administrative districts in France. The term 'SELEC A, C, D, E' represents the four epidemiologic clusters which are investigated. RND is a random selection of samples.

Table 1. Year, region and matrices of origin of the 480 genomes of *S*. Dublin used in the study

	Year		Region			Matrix				
	≤2015	>2015	1	2	NA	Cow	Milk	Cheese	Processing plant	Human
Samples	71	409	256	190	34	167	219	58	5	31

The phylogenomic reconstruction indicated that the isolates do not tend to group by years or matrices. This suggests that within a geographical area there is a continuous exchange of strains between matrices over the years. However, the genomic similarity of a limited number of isolates from different regions indicates that geographical barriers were not completely sealed and that exchanges of strains takes place to some degree between the two regions.

WGS-based epidemiological investigation and sample clustering

Considering the stringent threshold of pairwise differences defining related genomes (i.e. <5 SNPs), 32 singletons and 63 distinct clusters were defined and encompassed between 2 and 52 samples with a median of 4 (Figure 1). In addition, most of the samples from human cases were clustered within region 1 samples, as previously observed during an investigation performed in 2015 and 2016 by SNP-based analysis and variable-number tandem repeat analysis (MLVA) (30). Finished products related to the human cases were coming from region 1. In the present analysis, we observed that isolates from human cases also clustered with samples from milk, animal and environmental origins. This result shows that strains isolated as part of surveillance plans can provide an early warning of potential future human contamination. The targeted selections (SELEC A, C, D and E) were built on the intuitions of local actors that some isolates might be related. Phylogenomic clustering showed that each of these 4 datasets is polyphyletic. For SELEC D, corresponding to milk samples in a restricted area, the strains were scattered throughout the phylogenetic tree. As a general trend, targeted epidemiological investigations were in every case not in agreement with the genomic evidence. On one hand, some of the genomes were not clustered together, and on the other hand, samples originally considered as not related to any outbreak events were clustered with outbreak strains (i.e. identified as genetically very close to defined genetic clusters) (Figure 1).

Geolocation and regional segregation of genomic diversity

In order to investigate in-depth the regional segregation, we collected accurate geographic data ($n = 261$) and built a phylogeography map of the two regions (Figure 2) according to our phylogenomic clustering (Figure 1). This phylogeographic reconstruction suggests that the geographical distance is a major factor in genomic divergence and relatedness for the early stages of the production processes (i.e. animals, farms), while downstream transformation steps are more likely to harbour genomic diversity. Some areas contain different clusters of genomes, especially the areas near frontiers. This observation is likely to reflect the diversity of origin of the samples transferred in these product hubs but also suggests that cross-contamination can occur in these locations. Most of the clusters are geographically packed, including over time, showing a persistence of *S*. Dublin sub-lineages in specific areas. The geolocation also illustrates the diversity of sources associated with most clusters (Figure 2), demonstrating the very local circulation of *S*. Dublin from animal to finished products. For example, the cluster 47 included 11 samples from bovine, 8 from milk and 7 from cheese with some strains isolated from the three matrices being virtually identical (0 SNP difference). Another example is the cluster 61 encompassing 9 samples from cattle, 9 from milk and 5 from cheese. These findings emphasize the persistence of *S*. Dublin along the production chain with highly conserved genomic characteristics.

DISCUSSION

This retrospective study demonstrates here that continuous genomic surveillance brings valuable information to understand routes of contamination and target sources of contamination faster. Both are key features bolstering investigations in an emergency context. Firstly, the sub-sampling of the strain panel was performed while balancing metadata modality and maximizing diversity representation. Singleton cluster analysis suggested that the sub-sampling covered a large fraction of the diversity and variability present in the available panel of isolates. Moreover, this considerably reduced the cost of sequencing and the computational time of bioinformatics pipelines, installing the genomic surveillance as an economically and timely efficient tool for food safety. Secondly, the core genome investigations together with epidemiological data were found resolutive and robust, allowing an easy and accurate identification of strain links at the regional scale. Finally, *S*. Dublin exhibited genetic diversity specific to its geography, which resulted in local clusters that sometimes intermix through exchange zones.

WGS analysis brings more insight into outbreak investigations

Investigations on targeted sample selections showed that epidemiological data is not enough to decipher the link between samples. This is particularly true when dealing with closely related strains, in areas foregrounding regular trading and exchanges of food, products and animals linked to the carriage and transmission of *Salmonella*. Thanks to this core genome SNPs, phylogenetic reconstruction highlighted links between strains, which were not identified from epidemiological data, revealing new potential sources of contamination. It was previously shown that WGS can provide more insight in outbreaks investigations (55), thus some public health agencies have developed WGS methods to overcome the lack of precision of *Salmonella* typing methods (19,56,57). In France, WGS is not systematically implemented as the main typing tool for Salmonella in foodborne outbreak investigation. It is therefore difficult to trace back outbreaks. Combined with epidemiological data, investigators can track back the dissemination of strains at the regional scale, and point-out exchanges of strains between places or the origin of their contamination.

Even few mutations show regional segregation of *S*. Dublin

We discovered a regional segregation of *S*. Dublin in France, which was not previously demonstrated. As previously shown (54,58), *S*. Dublin is a highly clonal serotype and harbours a highly conserved genome (13,53,54,59), which is found here with a low number of intra and inter-cluster SNPs, and from the few numbers of SNPs excluded from recombination events. This result is supported by previously published studies, which revealed low SNP differences between linked isolates and unlinked isolates from the same country (53,60,61). Even if the core-genome SNP pipelines used in these studies are different (62), the orders of magnitude between pairwise SNP differences are similar. More precisely, in the study (53) where samples were isolated between 1996 and 2016, the majority of isolates from the same geographic area clustered with a threshold of less than 10 SNPs. A comparison between French and Danish samples shows a clustering of strains by country (Supplementary Figure S3).

The investigations of the outbreak from French cheeses at a similar period (30) defined clusters and subclusters harbouring less than 15 and 5 SNP differences, respectively. In the present study, we have decided to apply a smaller threshold (five SNPs), given the shorter period of time and

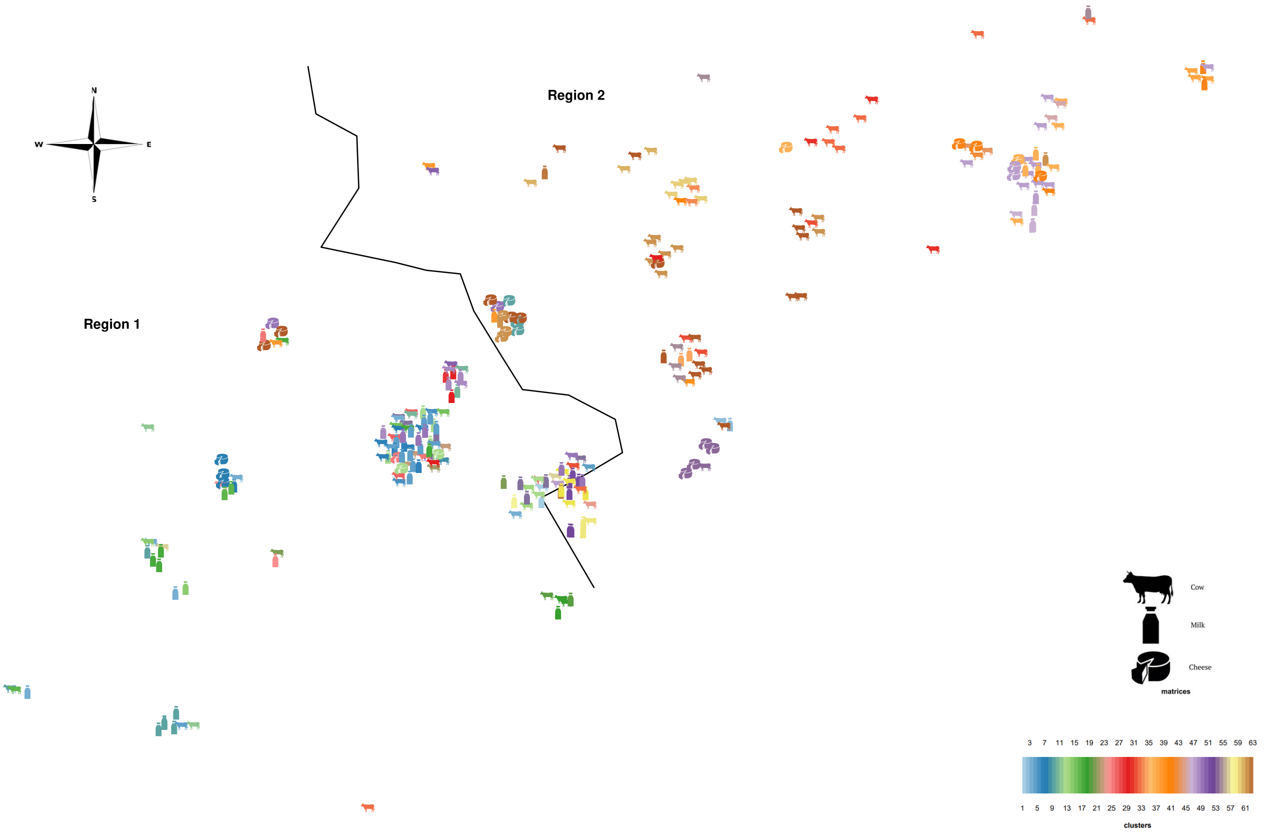


Figure 2. Jitterized geolocation of samples source and genomic clustering of *Salmonella* serotype Dublin isolated in two French regions between 2009 and 2018 ($n = 261$). The border between the two region is coarsely represented by the black delimitation. The clusters were defined with rPinecone set to target 5 or fewer SNP differences within clusters (coloured clusters). Clustering and colour scheme match those on Figure 1. Geographic location has been anonymized by adding a random variable to geographical coordinates. Pictogram of each point describes the matrix from which each isolate was sampled.

the small geographical area considered. A threshold of five SNPs is very conservative and allows identification of related samples. In addition, the drastic curation and quality assessment performed during sample selection in the present study make highly unlikely the detection of erroneous SNPs and recombination. It has been proven that the geographical partitioning impacts the core genome of *S.* Dublin (54), which supports our conclusion that strains considered to be related (i.e. that differ by five or less SNPs) belong to the same geographical area.

Geographical clusters of *S.* Dublin genomes highlight trading areas

Using the five SNPs threshold, samples were not clustered together by years or isolation matrices. On the contrary, the strains tend to cluster by geographic area, from cattle to finished products, supporting the hypothesis of persistence of the same strains infecting herds and production environments over the longer term. *S.* Dublin can be widespread in the environment and becomes an important source of infection for animals (7), which may have become latent carriers. The hypotheses about potential vertical transmission through carriers (14) would support the geographical segregation of *S.* Dublin observed in the present study. The spread of the strains in these areas could be through the purchase and contact of infected livestock. It is also possible that in these highland areas, the spread could occur through the watersheds and rivers during the rainy season.

Some geographic areas, such as cities located near the border between the two regions, correspond to towns of exchange and gathering of cattle. In this context, animals are exposed to multiple infection risk factors, such as cattle sales or agricultural forums, which promote inter-animal contamination and increase fecal-oral cycle of infection (13). In these cities, we found different strains, which clustered with different regions, showing the magnitude of bacterial exchanges taking place in the area. The region's livestock transport network and the network of farms where cheeses workshops are supplied seem also to harbour a high diversity of *S.* Dublin. Nevertheless, monitoring and investigating these networks to understand the circulation of strains remains difficult due to the amount of data and the patchy nature of sampling.

S. Dublin outbreaks had multiple origins

During the present study, we first analysed a sample set without isolates from human cases. We observed that adding samples from human cases did not change the tree topology. We also noticed that human samples are located in different clusters, showing different contamination origins as shown in Ung *et al.* (30). 4 samples from human cases are singleton, meaning that their origin remains elusive at the time. This observation suggests that the surveillance plan, despite its size and meshing, does not fully cover the diversity that exists in these two regions or that our subsampling failed to represent rare clusters from which those cases arose. The granularity of the surveillance should however not be held as a sole source of data scarcity as a large part of *Salmonella* cases are undetected or unreported (2). Although these four *S*. Dublin are included in the regional phylogenetic tree, the hypothesis of a foreign origin cannot be excluded, as genomic variability has not been studied throughout the country. The paucity of *S*. Dublin cases with available genomic resources and usable geographic metadata prevented deeper investigations.

Limits and perspectives

The surveillance, despite its size and meshing, does not seem to fully cover the diversity of these two regions up to this SNP resolution, or our subsampling failed to represent rare clusters, as singletons appear in the clustering. The sheer performance of the sampling and subsampling can nonetheless be tallied, as only 32 singletons were found out of 480 samples (6.7%) under the most stringent clustering threshold applied on *S*. Dublin. To answer those hypotheses with an in-depth analysis, we would recommend improving the monitoring plans in view of missing or ineligible metadata related to 219 samples in the present study. These observations emphasize that the implementation of a metadata nomenclature and minimal metadata sets is required for surveillance activities.

After an outbreak event in a cattle farm, all the organic material is removed and the surfaces are washed with water and detergent. A disinfectant is subsequently applied depending on the *Salmonella* species (63,64). After cleaning, measures should be taken to prevent reintroduction of the bacteria. In this study, we have a strong assumption of contamination by contact between animals from distinct facilities or through persistence within the hosts or in the environment. Indeed, heifers, calves and cows infected around the time of calving are the animals with the higher risk of becoming *S*. Dublin carriers (12,65), and environmental contamination from infected calves also plays an important role in the spread of the bacteria within calves (6). To prevent these risks of contamination and infection, biosecurity measures can be proposed. For instance, good calving area management has been associated with the probability of successful control of *Salmonella* (66). Measures, like separating calf pens by solid walls, preventing cows from calving before being moved to the designated calving pen or quarantine newly arrived animals has been proven to be effective against the spread of *Salmonella* in herds. In addition, calves are more likely to be seropositive in farms, thus monitoring the serology of all calves can predict a new outbreak within the herds (67). It could be recommended that on-farm hygiene measures be increased to limit the likelihood of transmission to cows during the production period and that milking hygiene measures be reinforced to prevent contamination of milk. Finally, milk from farms suspected of having active circulation of S. Dublin on the farm could be temporarily excluded from the production of raw milk cheeses

CONCLUSION

In this study, epidemiological and genomic data allowed the characterization of the diversity and understanding of phylogeographic location of *S*. Dublin strains. The precision brought by WGS methods bolstered the identification of different clusters and uncovered links between samples.

Our results display the geographical distance as a major factor in genomic divergence and relatedness for the early stages of the production processes (animals, farms), while down-the-line transformation steps are more likely related to host genomic diversity. The discriminative signal between samples from region 1 and region 2 from their genetic content is a precious result that can be used in the future to track back contaminations. These findings are in favour of a generalised persistence of local strains and occasional migration with a strong phylogeographic context. These findings also suggest that *S*. Dublin in those regions are geographically segregated with clusters containing different matrices potentially emphasizing spreading the bacteria over the entire food chain, and within herds. Geographic locations showing a high diversity of *Salmonella* were found to be exchange areas with several cooperatives or a large concentration of markets where different bacteria from different geographical locations meet each other. Control measures must be put in place in these exchange areas to prevent the spread of different clusters of *Salmonella* found in humans.

We appraise the benefit of a WGS cgSNP approach on closely related isolates and how the results from large datasets under proper control of the impact of breadth-of-core erosion can bring to fathom strain contamination dynamics and empower sanitary and safety authorities in designing tailored safety policies.

Altogether, the present results brought an insight on regional genomic diversity of highly related genomes involved in foodborne outbreaks, underlining the necessity to drive investigations toward the most resolutive comparative genomics methods. These findings pave the way toward the development of news comparative tools integrating others sources of variation as a discriminative metrics along with SNPs, such as INDELs, structural variations, mobilome and accessory genome contingency.

DATA AVAILABILITY

The core genome calculation of positions covered at more than 30X is available at: github/madeleinevlt. The figures and scripts are available at: github/madeleinevlt.

Raw reads of the project have been deposited in SRA under accession number PRJNA737646.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors would like to thank Choreh Farrokh for helping, mediating with the stakeholders, the operators and altogether advocating for the project and funding. Part of the computation was performed on the GAMeR HPC Cluster. The random selection group (RND) was initially labelled SELEC B and renamed for clarity. Selections A, C, D and E were however not renamed to ensure coherence with previous public and private communications.

FUNDING

French general directorate of food (DGAL); French Dairy Interbranch Organization (Cniel).

Conflict of interest statement. Valérie MICHEL is currently head of the dairy microbiology department at ACTALIA, a Technical institute that helps private enterprises to constantly improve their quality products. ACTALIA is currently accredited as an Agro-Industry Technical Institute (ITAI) by the French ministry for agriculture and food. Gaëtan Podeur and Madeleine De Sousa Violante were previously shortly employed at ACTALIA.

REFERENCES

1. Kirk,M.D., Pires,S.M., Black,R.E., Caipo,M., Crump,J.A., Devleesschauwer,B., Döpfer,D., Fazil,A., Fischer-Walker,C.L., Hald,T. *et al.* (2015) World health organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS Med.*, **12**, e1001921.
2. The European Union One Health 2019 Zoonoses report (2021) *EFSA J.*, **19**, e06406.
3. Guillier,L., Thebault,A., Fravalo,P., Mughini-Gras,L., Jourdan-da Silva,N., David,J., Kooh,P., Cadavez,V. and Gonzales-Barron,U. (2020) Risk factors for sporadic salmonellosis: a systematic review and meta-analysis. *Microb. Risk Anal.*, **17**, 100138.
4. Hoelzer,K., Moreno Switt,A.I. and Wiedmann,M. (2011) Animal contact as a source of human non-typhoidal salmonellosis. *Vet. Res.*, **42**, 34.
5. Harvey,R.R., Friedman,C.R., Crim,S.M., Judd,M., Barrett,K.A., Tolar,B., Folster,J.P., Griffin,P.M. and Brown,A.C. (2017) Epidemiology of salmonella enterica serotype dublin infections among humans, united states, 1968-2013. *Emerg. Infect. Dis.*, **23**, 1493–1501.
6. Holschbach,C.L. and Peek,S.F. (2018) Salmonella in dairy cattle. *Vet. Clin. North Am. Food Anim. Pract.*, **34**, 133–154.
7. Winfield,M.D. and Groisman,E.A. (2003) Role of nonhost environments in the lifestyles of salmonella and escherichia coli. *Appl. Environ. Microbiol.*, **69**, 3687–3694.
8. Institut Pasteur (2019) CHU Robert Debré-APHP Rapport d'activité annuel 2019 Année d'exercice 2018.
9. Bayne,H.G., Garibaldi,J.A. and Lineweaver,H. (1965) Heat resistance of Salmonella typhimurium and Salmonella senftenberg 775 w in chicken meat. *Poult. Sci.*, **44**, 1281–1284.
10. Pecoraro,H.L., Thompson,B. and Duhamel,G.E. (2017) Histopathology case definition of naturally acquired salmonella enterica serovar dublin infection in young holstein cattle in the northeastern united states. *J. Vet. Diagn. Invest.*, **29**, 860–864.
11. Mee,J.F., Jawor,P. and Stefaniak,T. (2021) Role of infection and immunity in bovine perinatal mortality: part 1. Causes and current diagnostic approaches. *Animals*, **11**, 1033.
12. Nielsen,L.R., Schukken,Y.H., Gröhn,Y.T. and Ersbøll,A.K. (2004) Salmonella dublin infection in dairy cattle: risk factors for becoming a carrier. *Prev. Vet. Med.*, **65**, 47–62.
13. McDonough,P.L., Fogelman,D., Shin,S.J., Brunner,M.A. and Lein,D.H. (1999) Salmonella enterica serotype dublin infection: an emerging infectious disease for the northeastern united states. *J. Clin. Microbiol.*, **37**, 2418–2427.
14. Wray,C., Wadsworth,Q.C., Richards,D.W. and Morgan,J.H. (1989) A three-year study of salmonella Dublin infection in a closed dairy herd. *Vet. Rec.*, **124**, 532–537.
15. Nielsen,L.R., Houe,H. and Nielsen,S.S. (2021) Narrative review comparing principles and instruments used in three active surveillance and control programmes for Non-EU-regulated diseases in the Danish cattle population. *Front. Vet. Sci.*, **8**, 685857.
16. Bonifait,L., Thépault,A., Baugé,L., Rouxel,S., Le Gall,F. and Chemaly,M. (2021) Occurrence of salmonella in the cattle production in france. *Microorganisms*, **9**, 872.
17. Persson,S., Jacobsen,T., Olsen,J.E., Olsen,K.E.P. and Hansen,F. (2012) A new real-time PCR method for the identification of salmonella dublin. *J. Appl. Microbiol.*, **113**, 615–621.
18. Deng,X., den Bakker,H.C. and Hendriksen,R.S. (2016) Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.*, **7**, 353–374.
19. Chattaway,M.A., Dallman,T.J., Larkin,L., Nair,S., McCormick,J., Mikhail,A., Hartman,H., Godbole,G., Powell,D., Day,M. *et al.* (2019) The transformation of reference microbiology methods and surveillance for salmonella with the use of whole genome sequencing in england and wales. *Front Public Health*, **7**, 317.
20. Simon,S., Trost,E., Bender,J., Fuchs,S., Malorny,B., Rabsch,W., Prager,R., Tietze,E. and Flieger,A. (2018) Evaluation of WGS based approaches for investigating a food-borne outbreak caused by salmonella enterica serovar derby in germany. *Food Microbiol.*, **71**, 46–54.
21. EFSA Panel on Biological Hazards (EFSA BIOHAZ Panel), Koutsoumanis,K., Allende,A., Alvarez-Ordóñez,A., Bolton,D., Bover-Cid,S., Chemaly,M., Davies,R., De Cesare,A., Hilbert,F. *et al.* (2019) Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. *EFSA J.*, **17**, e05898.
22. Yang,Z. and Rannala,B. (2012) Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, **13**, 303–314.
23. Radomski,N., Cadel-Six,S., Cherchame,E., Felten,A., Barbet,P., Palma,F., Mallet,L., Le Hello,S., Weill,F.-X., Guillier,L. *et al.* (2019) A simple and robust statistical method to define genetic relatedness of samples related to outbreaks at the genomic scale – application to retrospective salmonella foodborne outbreak investigations. *Front. Microbiol.*, **10**, 2413.
24. Vila Nova,M., Durimel,K., La,K., Felten,A., Bessières,P., Mistou,M.-Y., Mariadassou,M. and Radomski,N. (2019) Genetic and metabolic signatures of salmonella enterica subsp. enterica associated with animal sources at the pangenomic scale. *BMC Genomics*, **20**, 814.
25. Sheppard,S.K., Guttman,D.S. and Fitzgerald,J.R. (2018) Population genomics of bacterial host adaptation. *Nat. Rev. Genet.*, **19**, 549–565.
26. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
27. Didelot,X. and Wilson,D.J. (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.*, **11**, e1004041.
28. Didelot,X., Bowden,R., Street,T., Golubchik,T., Spencer,C., McVean,G., Sangal,V., Anjum,M.F., Achtman,M., Falush,D. *et al.* (2011) Recombination and population structure in salmonella enterica. *PLoS Genet.*, **7**, e1002191.
29. Chu,C., Feng,Y., Chien,A.-C., Hu,S., Chu,C.-H. and Chiu,C.-H. (2008) Evolution of genes on the salmonella virulence plasmid phylogeny revealed from sequencing of the virulence plasmids of s. enterica serotype dublin and comparative analysis. *Genomics*, **92**, 339–343.
30. Ung,A., Baidjoe,A.Y., Van Cauteren,D., Fawal,N., Fabre,L., Guerrisi,C., Danis,K., Morand,A., Donguy,M.-P., Lucas,E. *et al.* (2019) Disentangling a complex nationwide salmonella dublin outbreak associated with raw-milk cheese consumption, france, 2015 to 2016. *Euro Surveill.*, **24**, 1700703.
31. Listadapt (2020) D1.4-report on strain collection and strategy for selection of strains sequencing.
32. Leinonen,R., Sugawara,H. and Shumway,M., and International nucleotide sequence database collaboration (2011) the sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
33. Bushnell,B. (2014) *BBMap: A Fast, Accurate, Splice-Aware Aligner Lawrence Berkeley National Lab. (LBNL)*. Berkeley, CA .

34. Xu,S., Ackerman,M.S., Long,H., Bright,L., Spitze,K., Ramsdell,J.S., Thomas,W.K. and Lynch,M. (2015) A male-specific genetic map of the microcrustacean daphnia pulex based on single-sperm whole-genome sequencing. *Genetics*, **201**, 31–38.
35. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
36. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
37. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using minhash. *Genome. Biology.*, **17**, 132.
38. Bosi,E., Donati,B., Galardini,M., Brunetti,S., Sagot,M.-F., Lió,P., Crescenzi,P., Fani,R. and Fondi,M. (2015) MeDuSa: a multi-draft based scaffolder. *Bioinformatics*, **31**, 2443–2451.
39. Kosugi,S., Hirakawa,H. and Tabata,S. (2015) GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics*, **31**, 3733–3741.
40. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
41. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
42. Low,A.J., Koziol,A.G., Manninger,P.A., Blais,B. and Carrillo,C.D. (2019) ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ*, **7**, e6995.
43. Zhang,S., Yin,Y., Jones,M.B., Zhang,Z., Deatherage Kaiser,B.L., Dinsmore,B.A., Fitzgerald,C., Fields,P.I. and Deng,X. (2015) Salmonella serotype determination utilizing high-throughput genome sequencing data. *J. Clin. Microbiol.*, **53**, 1685–1692.
44. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R., and 1000 genome project data processing subgroup (2009) the sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
45. Felten,A., Vila Nova,M., Durimel,K., Guillier,L., Mistou,M.-Y. and Radomski,N. (2017) First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of salmonella serovars to mammalian- and avian-hosts. *BMC Microbiol.*, **17**, 222.
46. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, **26**, 589–595.
47. Ebbert,M.T.W., Wadsworth,M.E., Staley,L.A., Hoyt,K.L., Pickett,B., Miller,J., Duce,J., Kauwe,J.S.K. and Ridge,P.G., and for the Alzheimer's disease neuroimaging initiative (2016) evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, **17**, 239.
48. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., Auwera,G.A.V., Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D. *et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
49. Nguyen,L.-T., Schmidt,H.A., Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
50. Hedge,J. and Wilson,D.J. (2014) Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *Mbio*, **5**, e02158-14.
51. Wailan,A.M., Coll,F., Heinz,E., Tonkin-Hill,G., Corander,J., Feasey,N.A. and Thomson,N.R. (2019) rPinecone: define sub-lineages of a clonal expansion via a phylogenetic tree. *Microb. Genom.*, **5**, e000264.
52. Pupko,T., Pe,I., Shamir,R. and Graur,D. (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.
53. Kudirkiene,E., Sørensen,G., Torpdahl,M., de Knegt,L.V., Nielsen,L.R., Rattenborg,E., Ahmed,S. and Olsen,J.E. (2020) Epidemiology of salmonella enterica serovar dublin in cattle and humans in denmark, 1996 to 2016: a retrospective whole-genome-based study. *Appl. Environ. Microbiol.*, **86**, e01894-19.
54. Fenske,G.J., Thachil,A., McDonough,P.L., Glaser,A. and Scaria,J. (2019) Geography shapes the population genomics of salmonella enterica dublin. *Genome. Biol. Evol.*, **11**, 2220–2231.
55. Parcell,B.J., Gillespie,S.H., Pettigrew,K.A. and Holden,M.T.G. (2021) Clinical perspectives in integrating whole-genome sequencing into the investigation of healthcare and public health outbreaks – hype or help?*J. Hosp. Infect.*, **109**, 1–9.
56. Ibrahim,G.M. and Morin,P.M. (2018) Salmonella serotyping using whole genome sequencing. *Front. Microbiol.*, **9**, 2993.
57. Brown,E., Dessai,U., McGarry,S. and Gerner-Smidt,P. (2019) Use of whole-genome sequencing for food safety and public health in the united states. *Foodborne Pathog Dis.*, **16**, 441–450.
58. Srednik,M.E., Lantz,K., Hicks,J.A., Morningstar-Shaw,B.R., Mackie,T.A. and Schlater,L.K. (2021) Antimicrobial resistance and genomic characterization of salmonella dublin isolates in cattle from the united states. *PLoS One*, **16**, e0249617.
59. Paudyal,N., Pan,H., Elbediwi,M., Zhou,X., Peng,X., Li,X., Fang,W. and Yue,M. (2019) Characterization of salmonella dublin isolated from bovine and human hosts. *BMC Microbiology*, **19**, 226.
60. Ågren,E.C.C., Wahlström,H., Vesterlund-Carlson,C., Lahti,E., Melin,L. and Söderlund,R. (2016) Comparison of whole genome sequencing typing results and epidemiological contact information from outbreaks of salmonella dublin in swedish cattle herds. *Infect Ecol. Epidemiol.*, **6**, 31782.
61. Mohammed,M. and Cormican,M. (2016) Whole genome sequencing provides insights into the genetic determinants of invasiveness in salmonella dublin. *Epidemiol. Infect.*, **144**, 2430–2439.
62. Saltykova,A., Wuyts,V., Mattheus,W., Bertrand,S., Roosens,N.H.C., Marchal,K. and Keersmaecker,S.C.J.D. (2018) Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to salmonella enterica serotype typhimurium and serotype 1,4,[5],12:i:-. *PLoS One*, **13**, e0192504.
63. Beier,R.C., Callaway,T.R., Andrews,K., Poole,T.L., Crippen,T.L., Anderson,R.C. and Nisbet,D.J. (2017) Disinfectant and antimicrobial susceptibility profiles of salmonella strains from feedlot water-sprinkled cattle: hides and feces. *J. Food Chem. Nanotechnol.*, **3**, 50–59.
64. Kent,E., Okafor,C., Caldwell,M., Walker,T., Whitlock,B. and Lear,A. (2021) Control of salmonella dublin in a bovine dairy herd. *J. Vet. Intern. Med.*, **35**, 2075–2080.
65. Nielsen,L.R. (2013) Within-herd prevalence of salmonella dublin in endemically infected dairy herds. *Epidemiol. Infect.*, **141**, 2074–2082.
66. Nielsen,T.D., Vesterbæk,I.L., Kudahl,A.B., Borup,K.J. and Nielsen,L.R. (2012) Effect of management on prevention of salmonella dublin exposure of calves during a one-year control programme in 84 danish dairy herds. *Prev. Vet. Med.*, **105**, 101–109.
67. Nielsen,L.R. (2013) Review of pathogenesis and diagnostic methods of immediate relevance for epidemiology and control of salmonella dublin in cattle. *Vet. Microbiol.*, **162**, 1–9.

4.6 Discussion

4.6.1 Same WGS tools for different objectives

In this chapter, I study 3 different serovars, which displayed 3 different diversity backgrounds. Using the same WGS method tools, comparative functional genomics highlighted the whole diversity of these serovars, whether core or accessory. The geographic persistence was investigated for *Salmonella* Dublin and *Salmonella* Typhimurium and its monophasic variant, while *Salmonella* Mbandaka analyses focused on the persistence in hosts. Even focusing on the same objectives, the analyses performed with *Salmonella* Dublin and *Salmonella* Typhimurium and its monophasic variant demonstrated completely different adaptation histories. On one hand, *Salmonella* Dublin adapted and persisted in the environment. On the other hand, monophasic variant of Typhimurium had a resistance arsenal very well adapted to the pig host (copper resistance, biocide resistance), and did not seem to have developed any environmental specificity to be able to spread easily throughout the environment in France. Some studies contradict our results [286, 103] while others support them [267, 506], due to the scale of the studies. Most of the time, study that mix different hosts and/or serovars do not analyze SNP content and fail to detect the genomic patterns that we identified in the current chapter. Here, the restricted dataset without too many different vectors allowed us to explain these persistence phenomena. Thanks to all the tools used in this chapter, we have been able to analyze genomes at a core variant scale, and core and accessory gene content. Additional analysis with accessory variants would make the methodology more robust, especially for the marker screening.

4.6.2 Comparison between *Salmonella* Mbandaka and *Salmonella* Dublin

While *Salmonella* Typhimurium and its monophasic variant are well studied in genomics [511, 512, 513, 287, 180, 286], *Salmonella* Mbandaka lacks of genomic review to fully characterize the diversity of the dataset and compare it with other studies. Comparing this serovar to another widespread serovar in cattle, *S.* Dublin, the genomic dynamics are very different. Overall, 1062 SNPs has been detected in the 140 bovines genomes for *Salmonella* Mbandaka, while 1041 SNPs were detected along the 480 genomes for *Salmonella* Dublin. This is roughly the same number of SNPs observed, except that the *S.* Mbandaka dataset is more than 3 times smaller than *Salmonella* Dublin dataset. Even though these serovars have very similar clinical consequences and host range, specific environmental pressures may have shape the observed divergent genomics structures.

Despite similarities have been identified, in particular concerning the continuous contamination all along the food chain from bovine to cheese, geographical data were not precise enough to suggest that the geographical distance is a major factor explaining in genomic divergences of *S.* Mbandaka, even if low genomic diversity has been identified in the same farms (Figure 7.15).

4.6.3 Limits of the three studies

Regard this, the overall limitation that has emerged for each serovar was the lack of accurate metadata to conclude. Given the sensitivity of the information, it was difficult to obtain minimum metadata throughout my PhD, especially for serovars such as *Salmonella* Dublin which affects PDO (Protected Designation of Origin) cheeses. This lack of metadata did not allow us to conclude about important leads in each serovar. For *S.* Dublin, only half of strains could really be analyzed geographically. For *S.* Mbandaka, the precise geographical persistence of strains and links of contamination between cattle or poultry farms could not be analyzed. Finally, for *S.* Typhimurium and its monophasic variant, a link between genomic diversity and geographical diversity was not demonstrated, as only departmental data were available, and therefore it was not possible to make links between certain farms or processing plants. In addition, studies about the transport networks or the exchanges between farms would be necessary to finalize the results, but this kind of data is very difficult to obtain. With this information, we could have answer to more hypotheses, either at the scientific or industrial levels.

A methodological limitation appeared in the analysis of markers for the monophasic variant of Typhimurium, but this issue remains global for the three serovars. There is a lack of analysis of the accessory genome, although it is part of the *Salmonella* genomic background and allows an important plasticity of the genome. Concerning the gene analysis [100] core and accessory genes were clustered, but variants were not analyzed on accessory genes. Otherwise, the non-coding segments were poorly analyzed, such as the intergenic DNA, whereas a variation could have consequences on the phenotype of *Salmonella* [322, 101].
Finally, a limitation of knowledge appeared concerning *S.* Mbandaka, as this serovar is not well studied.

4.6.4 What does the accessory genome bring to the study?

Using pgSNP with the sample collections of interest, we concluded that it is highly probable that a single clone of monophasic variant of Typhimurium appeared in pig farms with slight environmental adaptation. For *Salmonella* Mbandaka, we were able to show that this serovar is very heterogeneous, and has large differences between strains from a restricted geographical area. The limited knowledge of the coregenome greatly limits the understanding of the heterogeneity of the accessory genome, and does not allow to fully understand the contribution of the phylogenomic pangenome. Overall, the main difference was displayed on the TMV dataset, but the contribution of the accessory genome did not bring new hypothesis for the dissemination of these strains in different regions of France. On the other hand, this study has highlighted the importance of SNPs in the accessory genome, and further studies could help to understand their impact, especially on important accessory elements such as phages or plasmids. These conclusions also call for an increased monitoring of the monophasic variant of Typhimurium clone in farms on the one hand, and a more in-depth investigation to understand the evolution of the *Salmonella* Mbandaka rate on the other hand.

4.7 Conclusion

During this chapter, I was able to characterize the diversity of three mains serovars detected in dairy and pork industry : *Salmonella* Dublin, *Salmonella* Mbandaka, *Salmonella* Typhimurium and its monophasic variant. This work has been developed with the aim of scientific understanding of the dissemination of strains, but also with the aim of helping industrial actors to understand the persistence of these strains.

Using coregenome analysis, I displayed the advantages of using WGS combined with epidemiological data, to better understand the persistence and the dissemination of the bacteria over a region, a country, or compared to worldwide data.

I provided a strong overview of the diversity of *Salmonella* Typhimurium and its monophasic variants in pork industry. I highlighted that the diversity of *Salmonella* Typhimurium is higher than TMV which is more clonal, and I characterized this diversity at a global scale. I also demonstrated the low diversity of TMV samples between herds, hypothesising the dissemination of a single or two clonal strains. Some samples from bordering countries shares the same diversity, highlighting the pigs trades between these countries, or contamination by a vector such as food products. The high amount of AMR genes, heavy metals and biocides explain the prevalence of these two serovars in pig farms and also reinforces the little difference between these strains. I also proposed genomics markers to detect samples from France using genes and variants screening.

Using *Salmonella* Mbandaka dataset, I highlighted that the genomic diversity reveals a degree of adaptation to bovine and avian hosts, with clusters more adapted to one or the other host. I demonstrated the absence of matrix clusters, suggesting continuous contamination all along the food chain, from bovine to cheese, and in poultry herds. I also proposed two sources of contamination that could explain the dissemination of these strains, with food products and wild birds vectors. Finally, I characterize the persistence of this serovar in cattle and poultry, analysing virulence and resistance genes, and I proposed several genomics markers to improve the surveillance of different clusters within herds.

Finally, in *Salmonella* Dublin, results display the geographical distance as a significant factor in genomic divergence and relatedness for the early stages of the production processes (animals, farms), whereas down-the-line transformation steps are more likely associated to host genomic diversity. These findings also suggest that *Salmonella* Dublin in those regions are geographically segregated with clusters containing different matrices potentially emphasising spreading the bacteria over the entire food chain, and within herds. I also demonstrated that this genomic signal can be used as a valuable tool to track back contamination.

Overall, these studies focusing on three *Salmonella* serovars demonstrate the strength of using WGS to solve different sanitary issues such as foodborne outbreaks investigations and source attribution.

Chapter 5

General discussions and conclusions

The project's aim was to investigate the genomic diversity and dissemination of *Salmonella* Mbandaka, *Salmonella* Typhimurium, and its monophasic variation in the milk and pork food sectors. These serovars recently became of notorious concern in food chains. For this purpose, the strains were characterized in two ways: first, using the methods already validated and accepted by the scientific community (chapter 4), and second, by the development of an innovative method in order to increase resolution and overcome the shortcomings of the latest genomic methods (chapter 3). In this part, I will discuss the pros and cons of phylogenomic inferences taking into account SNPs from the core and accessory genome, especially for heterogeneous serovars like those studied in this thesis. In addition, I will display the comparison of the main results for each serovar, to show the possibilities of application of WGS to various issues.

5.1 The place of the accessory genome in food safety investigation

In this thesis, a pangenomic pipeline called pgSNP was developed (chapter 3), and compared to coregenome SNPs approaches. While coregenome SNP pipeline presented robust and concrete results on the genomic diversity of strains, described in the chapter 4, I showed with pgSNP that the accessory genome left out of the analysis was too informative not to consider it. Using pgSNP, I was able to overcome 2 limitations discussed in the section 2.4.

The first limitation was the requirement of a reference genome for coregenome investigation. With coregenome SNPs approach, the selection of a reference genome for a dataset is a crucial step and poor selection can lead to loss of information, whether in variant detection, distance comparison between strains, or other downstream analysis [318, 319]. In this thesis, I proposed a straightforward BLASTN [327] implementation that might be used to quickly construct a pangenome-dependent dataset. This method demonstrated a better mapping quality and an increase in the number of mapped reads for the following analyses.

The second limitation is that accessory genome was left out of the analysis, while it represented keys to the adaptation and diversity of *Salmonella* in their environment [101, 320, 246]. Here, along with the pangenome reference, I implemented a methodology based on the pangenome reference contigs inferred in subtrees, and then concatenated in a one pangenome tree using a supertree method. Using this pipeline, I demonstrated that pgSNP brought a higher resolution to the genomic comparison of bacterial genomes, but also concordant results observed with other methodologies and epidemiological investigations.

pgSNP has been already well discussed in section 3.4.1, including opportunities for pipeline improvements to ensure robustness of results, especially in the case of food safety investigations. Overall, this tool brought a higher resolution compared to coregenome ones, but also raised new questions about the implementation of this new information in the case of an epidemic investigation. During investigations, it was important to gauge the proximity of the samples to know if they were linked to the same source of contamination or not. With the addition of the accessory genomes, the distance between strains from the same outbreak can increase. For example, if a strain has an additional mobile element but very little difference in the coregenome, do these two strains must be considered as related or unrelated? This discussion has been already observed in section 3.3.4.3, where samples from the same outbreak were divided in two subgroups due to 7kb of accessory genome from 2 contigs. While it would be possible that an outbreak related genome may be excluded from the outbreak cluster inferred with pgSNP due to high level of SNP differences induced by mobile genetic elements, it is important to note that we did not observe this phenomena with the studied dataset. While in the coregenome it has been proposed SNP thresholds to judge the proximity of the strains (section 4.5.2), it is not possible at this time to apply the same method to the accessory genome, or additional studies should be set up to gauge the contribution of the accessory genome to the characterization of outbreak strains. Nevertheless, I have shown that the accessory genome could be discriminating enough to separate 2 sequence types (ST), like *S.* Typhimurium and TMV in section 3.3.4.1.4.

In addition, I demonstrated that pgSNP allowed exploration of the accessory genome thanks to the cutting of contig and reference pangenome, which is linear, and thus easier to annotate. For example, to detect defined DNA fragments (e.g. gene presence, phages), there was no need to detect these elements with dedicated workflows (section 4.3.2.2, 4.3.2.4, 4.4.3 and 4.4.4). Indeed, I just identified these elements based on pangenome annotation. However, this strategy did not allow detection of these elements in case of segments present in less than 4 samples. In addition, if the annotation was made on the whole pangenome reference, without truncating the alignment, sequencing errors could be inserted, and thus the quality of the annotation and detection could have been weakened. In the context of food safety and outbreak investigation, the detection of these rare mobile genetic elements are important to case of emergence or reappearance of known pathogenic elements. On the other hand, this pgSNP could be very suitable for a quick screening of genes on a panel of strains, before characterizing them precisely on each assembly.

Finally, a supertree represents a variety of subtrees, thus there may be a loss of accuracy. The objective was to bring a higher resolution and accurate precision of the accessory genome using branch lengths. ERaBLE [333] was able to reflect genomic distance, and maintain the concordance that could be seen on the branches of a coregenome tree. Overall, it is possible to draw similar conclusions regarding genomic diversity from both core and pangenome SNPs.

Datasets	*Salmonella* Typhimurium and its monophasic variant	*Salmonella* Mbandaka	*Salmonella* Dublin
Number of genomes (core analysis)	322 genomes (section 4.2.1.1) 188 genomes (section 4.2.1.2) 325 genomes (section 4.2.1.3)	140 genomes (section 4.2.2.1) 304 genomes (section 4.2.2.2) 224 genomes (section 4.2.2.4)	480 genomes (section 4.2.3)
Coregenome SNPs of smallest dataset	4,247 SNPs for 188 genomes	1,062 SNPs for 140 genomes	1,041 SNPs for 480 genomes
Homologous recombination events	25	13	8
Characterisation of the diversity	Low diversity for TMV, High diversity for *Salmonella* Typhimurium	High diversity	Low diversity
Main results	- No geographical segretation in France - Large adaptation panel - Possibility of genomic specificity of TMV in France compared to other countries	- First characterisation of the genomic diversity of this serovar in France - Large adaptation panel - Possibility of host pattern - Possibility of contamination through wild birds	Regional segregation
Number of genomes (pan analysis)	118 genomes (section 4.2.1.2)	304 genomes (section 4.2.2.2)	/
% genome added to the study	(+) 30% DNA (4,7Mb ->6,2Mb)	(+) 27% (4,8Mb ->6,1Mb)	/
Contribution of the accessory genome	- Supports the low diversity - Supports the fact that there is no geographic segregation	- Support the high diversity of *Salmonella* Mbandaka - High variability of accessory genome	/

Table 5.1: Conclusion of the genomic analysis of each serovar of the thesis

5.2 Investigation of the genomic diversity of three different serovars

Three serovars have been investigated, using different methods, core and accessory ones, to characterise their diversity. It allowed me to show that genomic methodology can be applied to different issues. Table 5.1 summarizes the main results for each serovar.

For the pig and pork sector, the questions explored in this thesis were the diversity of *Salmonella* Typhimurium and its monophasic variant at different stages of the food chain (waiting rooms, processing premises and pork carcasses), the geographical diversity of these serovars in pigs herds, the comparison of this diversity at a worldwide level and finally the main genetic factors favoring the *Salmonella* persistence in pig herds.

For the dairy sector, the main questions addressed in this thesis for *Salmonella* Mbandaka were the extent of the biodiversity, the possibility of patterns related to the reservoir (food chain: bovine environment, milk, cheese; host: cattle, poultry), and the main genetic factors favoring the *Salmonella* persistence in livestock. Finally, for *Salmonella* Dublin in the dairy sector, investigations were conducted on the geographical diversity of this serovar in cattle.

Multiple datasets have been selected to answer each question. First, I displayed that the strains contaminated the whole production chain, without showing any adaptation to a specific source for the three serovars. This result was already discussed in the literature [360, 514, 259] for *Salmonella* Dublin and *Salmonella* Typhimurium and its monophasic variant.

Concerning the geographical diversity, I demonstrated that the dissemination of *Salmonella* Typhimurium and its monophasic variant was not associated with the geographical origin in France, while the geographical distance was a major factor in genomic divergence for *Salmonella*

Dublin. On the other hand, by exploring the diversity of TMV worldwide, I have shown that this diversity was quite specific to France, and partially shared by some bordering countries like Italy or Germany. For *Salmonella* Mbandaka, the hypothesis of linking the genomic diversity with the geographic diversity has not been explored in order to focus on investigating of the genomic diversity related to the host reservoir of *Salmonella* .

In this part, topological clusters linked to the host have been shown, but more in deep explorations are needed to demonstrate the existence or not of genomic patterns linked to the host, subject still under discussion by several studies [103, 506].

Regarding genetic factors favoring the *Salmonella* Typhimurium and its monophasic variant or *Salmonella* Mbandaka persistence in pig or cattle herds, I have shown in both cases that *Salmonella* could possess an arsenal of genes and genetic elements to adapt to its environment, including biocides for the farm environment, but an adaptation antibiotic resistance which can be a threat to human health [220]. The analysis of virulome also showed the possibilities of colonization of *Salmonella* within its host. These genomic analyses cannot be validated without concomitant phenotypic studies. In addition, the methodology used did not take into account SNPs found on genes and genetic elements of interest, which could invalidate the phenotype related to the presence of a gene.

Finally, the application of the accessory genome has not shown great topological differences on *S.* Mbandaka, but has been able to reinforce the hypotheses of a great diversity of this serovar. On the contrary for TMV, SNPs from the accessory genome had greater consequences because the distance in the coregenome of these strains is small. Having the opportunity to work on these two serovars allowed us to demonstrate the possibilities of using pgSNP on serovars with different genomic diversities.

These different studies allowed me to show that while TMV and *Salmonella* Dublin showed low diversity (few SNPs), *Salmonella* Typhimurium and *Salmonella* Mbandaka displayed a high diversity. This findings are supported by the high number of SNPs despite the size of the dataset, and also by the long branches on the phylogenomic reconstruction. I also observed that the number of homologous recombination events differed between serovars, with much less recombination events for *S.* Dublin.

With regards to impacts on the industrial sectors, this thesis also demonstrated the advantages of WGS for the surveillance, characterization and investigation of strains from foods. In France, WGS is not systematically implemented as the main typing tool for *Salmonella* in foodborne outbreak investigation and surveillance, issue discussed in section 4.5.2. Having the opportunity to work with industrial actors in the dairy and pork sector allowed me to have quick feedback on certain hypotheses from the field, such as the possibility of contamination by feed or transport, or possible exchanges between farms. I also had the opportunity to visit to visit dairy farms and pork slaughterhouses to have a real understanding of the foodborne risks and the sanitary protocols in place. This work has been very informative on practices, and being able to meld genomic research with direct application was an opportunity to appropriate knowledge in industries.

5.3 Conclusion

This thesis allowed the methodological development of an innovative method, and its direct application to field cases, using pangenomic approaches implemented in a tool called "pgSNP". In chapter 3, I summarized the advantages and the new results inferred by pgSNP. pgSNP was able to find consistent results with a coregenome SNPs approach, but also to provide more resolution in phylogenomic analyses. I demonstrated the possibility to apply this pipeline on different bacterial outbreak datasets to show the importance of the information gained with this method. I analysed the advantages and limits of the pipeline and suggested improvements. I had the opportunity to work on pangenomic analyses, trending methods which will be greatly improved in the coming years, and I believe that this study will have enabled progress in this area.

I also had the opportunity to finely characterize *Salmonella* serovars prevalent in the pig and pork and dairy food industries. By using comparing genomic methods, I was able to characterise the diversity of these strains under different issues. I also showed the application of my developments on serovars with different plasticity, evolution, and also on clonal or heterogeneous genomes. Even if the serovars showed very different issues and contextualization, I demonstrated that the WGS methods discussed in this thesis were efficient enough to explore the different questions raised. This thesis validated some hypotheses and proposed new ones on these prevalent serovars that cause food safety and also livestock health risks. These results opened new possibilities of studies concerning the serovars of the thesis, either in the bioinformatics or the microbiological field.

To conclude, this thesis reflects methodological research and applied research in a rapidly expanding field, and reviews the current state of research in this area, while proposing elements of answers and new topics to explore.

Chapter 6

Annexes

6.1 Others methods developed in this thesis

6.1.1 Graphs

6.1.1.1 Introduction

Graphs are the new trending way to visualise genomes, and it has been recently improved to analyse pangenomes content of a sample dataset. A graph is made of by vertices and edges to represent the relation between variable quantities. In pangenome studies, vertices are representing kmers or genomes, and edges the relation between two vertices. Pangenome graphs are built from a dataset using different algorithms to take into account all studied genomes, or a gene set. A common way to construct a basic pangenome graph is to generate a compacted de Bruijn graph (cDBG). In practice, pangenome graphs can represent all the dataset, and be useful to identify coregenome (parts of the graph were there is only one path), the part of a genome or gene set that is shared across the majority of the strains or related species in a clade.

We can also use graphs to represent the genomic variability of a dataset. These methods show an advantage to clustering methods, thanks to their precision and the fact that all distances are represented. For instance, GrapeTree [63] implemented a minimum spanning tree algorithm wich can be adapted to different kind of mutations (genes, SNPs, kmers or cg/wgMLST [515]) to infer a rapid graph representing only the minimal genetic relationships between samples. Also, SNP network analysis has been developed to investigate SNP interaction in the genome [516, 517] underlying SNPs co-evolution [518].

However methods representing genomes as vertices and SNPs distance as edges to represent the relationship of a dataset has not been fully developed and studied yet, despite the low computational time compared to ML and Bayesian phylogenomic trees. In food safety, the main goal is to understand links between samples, thus it is not necessary to trace the whole story of the strain evolution. As the phylogenomic tree is faster and faster, the developments did not focus on graphs. But in large genome alignment (for example 300 whole-genome isolates alignment) with high dissimilarity level between genomes, ML or Bayesian calculation can take up to 1 or 2 days, when speed is essential for monitoring and investigating outbreaks. Graphs could potentially help interpretation of clusters when the SNP threshold of a new serovar is not defined.

During my thesis, I worked on the development a SNP-based graph which could take into account accessory SNPs. This short study was implemented by myself and a Master 2 student (Valentin BALOCHE) whom I supervised during 1 month. To achieve this objective, I divided the project in two parts: the first one is the computation of a distance matrix, and the second one is the representation of these distances.

6.1.1.2 Distance matrix calculation

Different pairwise difference matrices calculation were tested: dist.alignment from seqinr [519], distance_calculator from Biopython [410], and home-made matrix calculation. The two first were not discriminant enough to separate some TMV strains, even using coregenome SNPs. Finally, we implemented a method with python to measure the pairwise SNP difference between isolates. For each column of the alignment, if two sample present a different base, the distance between these two samples will increase by one. This home-made script was build to control the gap score of the pairwise SNP difference matrix. Also, an implementation of a gap score that would evolve according to the constitution of the accessory genome between two strains was proposed. For example, if one genome contains an additional DNA fragment, instead of adding the distance of the length of the fragment for each genome, this addition of DNA fragment could be interpreted as a single event weighted by its SNPs to avoid under-estimation pairwise differences from accessory genome, instead of a succession of evolution in the DNA like SNPs.

We decided to select a distance calculation converting the alignment into a data frame with rows corresponding to strains and columns corresponding to positions in the sequences. The distance was then calculated referring to the most represented bases on a position. Even if this method is less resolutive by using only presence/absence of the most represented bases on a position, the discrimination is sufficient and better than that calculated by the distance calculator which uses a distance matrix for each bases difference.

6.1.1.3 Graph representation

Finally, to represent the distance into a graph, I used the Qgraph [520] package in R. Qgraph creates a graph based on R plotting methods and a distance matrix. One advantage of this method is the recalculation of isolate positions based on springs between two samples, calculated using Fruchterman Reingold algorithms, where two samples with low distance will be close into the space of the plot.

The student implemented the graph using NetworkX [521] python package and imposed a node positioning using the Kamada-Kawai algorithm. Using Fruchterman Reingold algorithm on Networkx shows inconsistent results, something I had observed myself when trying to use Networkx's layout springs.

Using the Qgraph and the simple distance calculation on coregenome SNPs alignment, we are able to quickly recognize outbreak from sporadic samples. Qgraph intensifies the color of edges based on the weight edge between two isolates, and samples which are too distant have a light color, or in some case no edges at all between them. Looking at the coregenome graph in Figure 6.1, outbreaks are clustered together on the graph, and the edge color is intensified enough to agree with epidemiological data.

NetworkX is able to discriminate TMV outbreaks samples, keeping them away from other sporadic strains. However, epidemiological clusters from TMV are clustered in the middle of sporadic samples (Figure 4 in Supplementary material 7). We hypothesise that using a distance referring to the most represented bases on a position causes the calculation to lose resolution,

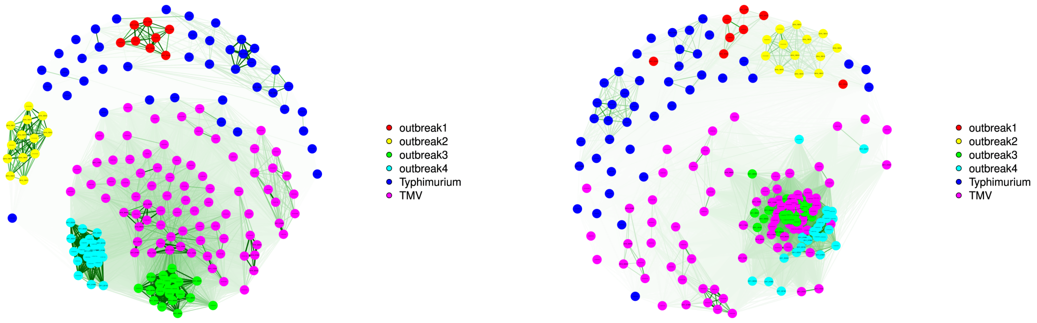


Figure 6.1: Qgraph plot of SNPs alignment for *Salmonella* Typhimurium and monophasic variant of Typhimurium dataset from [79]. Left is qgraph using core genome SNPs alignment. Right is graph using pan-genome SNPs alignment

and therefore close strains will be even closer with this simplification. Using absolute pairwise difference developed in python, it was possible to discriminate TMV outbreaks sample from sporadic sample, with higher resolution.

When adding accessory SNPs, the results are different and not really consistent with epidemiological data. In 6.1, I computed the absolute pairwise distance taking into account each bases as one single event. *Salmonella* Typhimurium outbreaks are found, except for two samples from outbreak 1. TMV samples are all clustered together, making unlikely the identification of strong outbreak relationships between samples. In this example, one accessory fragment counts for one difference into the matrix, it may be that phylogenomic core signal and consequently become in disagreement with epidemiological data. Different gap score or accessory SNPs score has been tested, but was not sufficient to distinguish outbreak samples from sporadic samples. However, these results are promising and should consequently be improved in a near future.

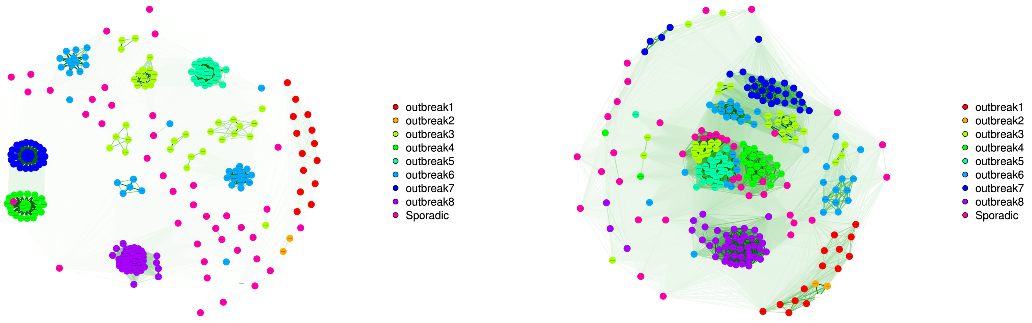


Figure 6.2: Qgraph plot of SNPs alignment for *Escherichia coli* dataset from [323]. Left is qgraph using core genome SNPs alignment. Right is graph using pan-genome SNPs alignment

This algorithm was also applied on *Escherichia coli* dataset to check that the epidemiological clusters which were not found genetically linked according to the trees (outbreak 3 and 6), could clustered in a graph. In the figure 6.2, we can see that even using difference matrix and graphs, outbreak 3 and outbreak 4 are scattered in small clusters around the graph. On the other hand, it is interesting to mention that the strains of the outbreak 1 are clustered together, but are still farther apart than expected. This results is also visible with regard to pairwise differences between two isolates from outbreak 1. For example, the 11-1024 and 11-1133 samples from outbreak 1 have 15 SNP difference, while intra-difference from others outbreak are lower (for example, 13-0081 and 13-0137 have 1 SNP difference). This result was not visible on coregenome SNP tree, while on pangenome SNP tree displays it. This could be due to the fact that the samples of outbreak 1 and 2 are very genetically different from others samples in the dataset, creating the long branch that splits the tree into two subclusters. These long branch crushes the leaves of outbreak 1 and 2, forming a rake even if the stumps are farther apart than expected. Using graphs, sample are clustered but also displays this difference. Using the pangenome SNPs data, some outbreak are well clustered together, but samples which were close in the tree (outbreak 3,4,5,6) are even closer and mixed in the graph, making the outbreak identification impossible.

6.1.1.4 Conclusion and discussion

In conclusion, this study is a preliminary work about genomics applications through graphs. Using coregenome SNPs, expected epidemiological data-based outbreaks are well identified, and allows displaying of close relationships between strains in a larger space than a phylogenomic tree. The Master 2 student study was also oriented to find thresholds on graphs to define clusters. The threshold of both, phylogenomic tree (i.e. sensitive and specific pairwise mutation differences) and graph (i.e. the intensity of edges weight) depends on the related and unrelated strains of the whole dataset. In Qgraph and Networkx, the cut value is automatically chosen from the a quantile (75th quantile for Qgraph) of all edges weight in the dataset, meaning that the cut value will change if you add only very similar strains in your dataset. The phylogenomic tree and graph thresholds can only be calculated if there are proven outbreak strains and proven non-sporadic strains in the dataset.

Concerning the outcomes of pangenome SNP graph, the poor clustering of epidemiological clusters are mainly due to the lack of time I was able to devote to the project. With hindsight, these results are promising because some outbreaks are detected and there is some consistency with the epidemiological data some cases. Also, results from the *Escherichia coli* dataset emphasizes that the clustering issue is due to the matrix rather than the graph method. The graph method allows to separate the clusters correctly, with a more realistic distance than what can be observed in a phylogenomic tree. In a near future, proper calibrations of the pairwise SNP difference matrix and graph management may provide more consistent clustering according to expected epidemiological data-based outbreak clusters.

The work made by the student Valentin Baloche who I supervised added new perspectives on the use of graphs for outbreak investigation and was also very concerned about the time factor that could benefit to approach. It also highlights the up-front work required to create a proper distance matrix that adequately represents data from the investigated sample collection. The student's report has been added to the supplementary data (Supplementary material 7).

6.1.2 Bayesian for partitioned data

6.1.2.1 Improving the downstream pipeline results

In this section, the possibility of using another strategy than the subtrees and supertree method is also discussed. One of the downsides of the pipeline is that unique accessory fragments are not taken into account, and the method for calculating branch lengths occasionally comes up with negative branch lengths that we have to replace with 0 (as proposed by the authors of ERaBLE [333]). To tackle this issue, the supermatrices-like strategy using MrBayes partition management has been tested. MrBayes [108] is able to perform Bayesian inference of phylogeny using Markov Monte Carlo (MCMC) method. It estimates the posterior distribution of model parameters and the posterior probabilities of phylogenetic trees using Bayes theorem. MrBayes is also able to divide data into partition, to use different models on the data, or also to estimate parameters separately for the individual partition.

To use this new strategy, a character set (charset) was defined on the contigs alignments. A charset defines an alignment bloc that have an individual parameters estimation. On the dataset, charset is in agreement with the pangenome contigs. MrBayes parameters were set based on a *Salmonella* publication [522] : a GTR-like model for the substitution with an invariable proportion of the sites, and 3 heated chains and one cold chain. As we had no support for using nucleotide substitution models for the accessory genome, we decided to try a small number of generation first (10k), and then increase it gradually. The selection criteria selected to stop the analysis is based on the average standard deviation of split frequencies, the Potential Scale Reduction Factor (PSRF) and the average Effective Sample Size (avgESS) [108]. If the average standard deviation is low, it means that the tree of the generation N-1 and the tree of the generation N are becoming increasingly similar. The PSRF score gives an idea of the convergence of the trees on a large dataset. Finally, the average ESS estimates how many truly independent samples of a given parameter the MCMC outcome represents [523]. All of these parameters are convergence criteria for trees constructed with MCMCs.

To investigate first the impact of partitioning data on the phylogenomic tree, the method was applied on the *Salmonella* Typhimurium and TMV dataset described in 3.3.4.1. Based on the outbreak data and serovars, the topology of two MrBayes phylogenomic tree were compared, with and without partitioning data, after 1M iterations. Based on the observation in figure 6.3 and 6.4, the partitioning data have more concordant clusters with epidemiological data. Phylogenomic tree inferred without partitioning data struggles to reconstruct epidemiological clusters of TMV outbreaks. Including the partitioning method in the dataset allows a higher reconstruction of the tree topology.

The impact of the number of generation on the topology of the tree was also investigated. After 10 000, 500 000 and 1 million generations, a phylogenomic tree is inferred and annotated with the epidemiological data. In Figure 6.5, 6.6 and 6.7, we observe that the outbreaks are gradually clustered together thanks to increased generation number. At 10 000 generations, serovars are separated on one side and the other of the tree. *Salmonella* Typhimurium outbreaks are well clustered together, but TMV outbreaks are all around the tree. At 500 000 generations, TMV outbreaks 4 isolates are clustered together, but outbreak 3 isolates are still independent of each other in the tree. Finally, at 1 million generation, we observe that all the epidemiological outbreak clusters are grouped independently as expected.

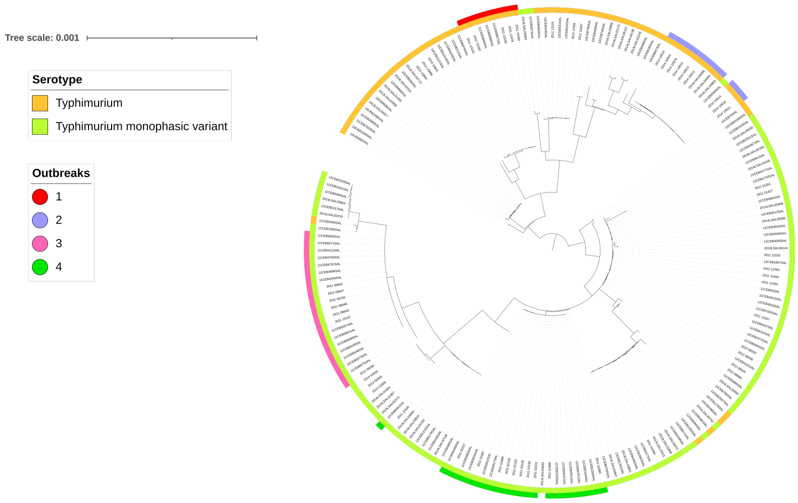


Figure 6.3: MrBayes phylogenetic tree inferred using partitioning data. Trees calculated after 1M iterations.

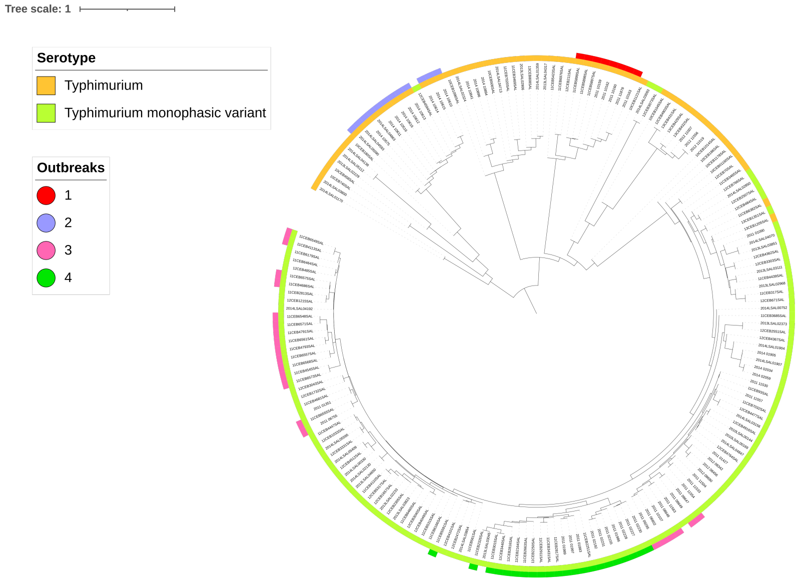


Figure 6.4: MrBayes phylogenetic tree inferred without using partitioning data. Trees calculated after 1M iterations.

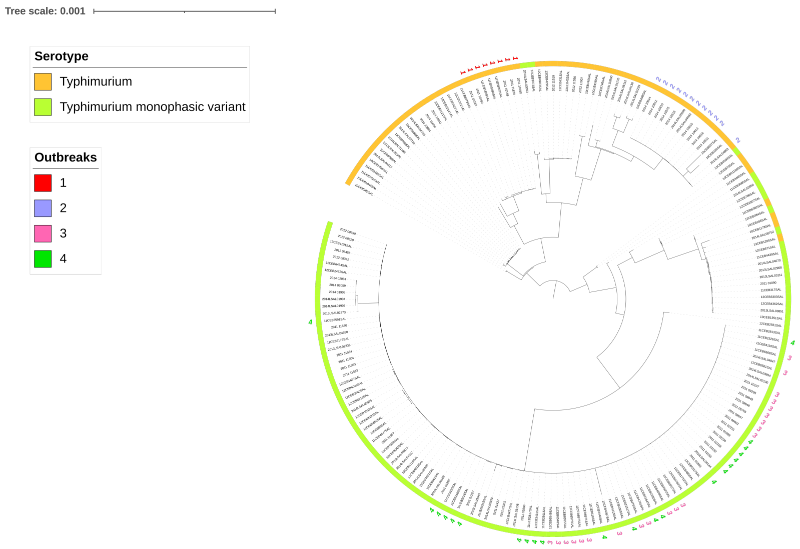


Figure 6.5: Evolution of MrBayes phylogenetic tree with partitioning data based on 10k number of generation.

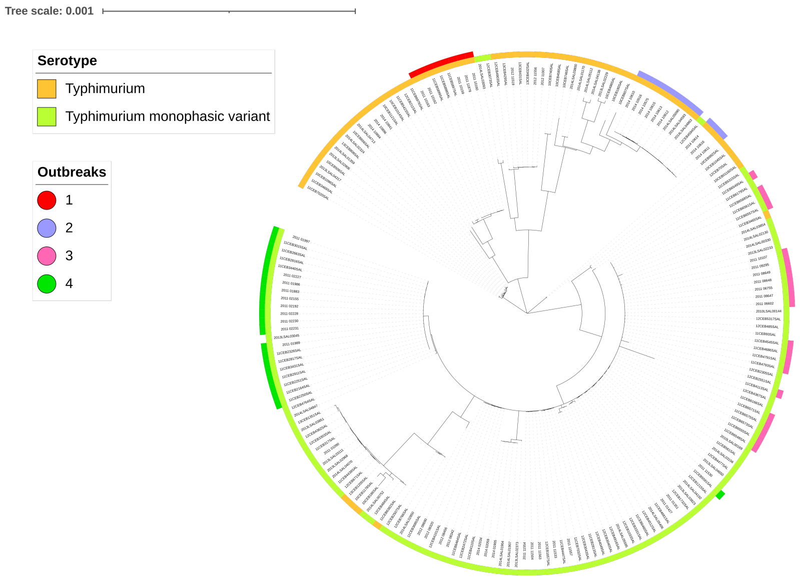


Figure 6.6: Evolution of MrBayes phylogenetic tree with partitioning data based on 500k number of generation.

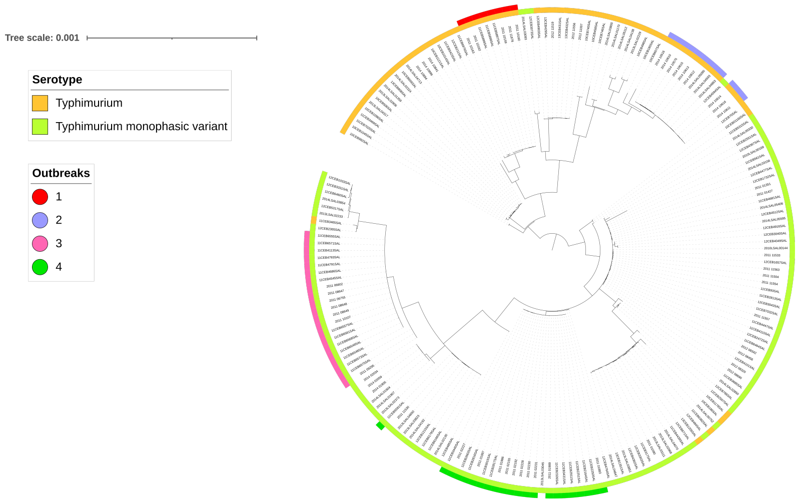


Figure 6.7: Evolution of MrBayes phylogenetic tree with partitioning data based on 1 M number of generation.

6.1.2.2 Can MrBayes be routinely used on the pipeline?

Unfortunately, even if the strains are well clustered together, the branch lengths do not reflect the genomic reality of the two serovars. Moreover, despite the large number of iterations, the results have still not converged. I built this tree after 1 week of computing. I also increased the number of generations to 2 million, and after another week the result still not converged. I estimate that it would be necessary to go up to 10 million iterations to obtain a stable phylogenomic tree, or else the basic parameters would have to be modified. But the downside of this approach is the computing time, which is too long to use this pipeline routinely. In view of the time it would have taken to develop this approach this work was put aside to focus on other approach developed in the framework of the present PhD thesis.

6.2 Others *Salmonella* Mbandaka studies

6.2.1 FimH analysis

As demonstrated by study from Min Yue et al., markers of adaptation of Salmonella has been identified on the FimH gene of the fimbriae mechanisms. The fimH gene sequence was isolated from the project PRJNA297164 from the author of the study, and detected on all *S.* Mbandaka samples using Blast. Each nucleotide sequence has been translated and compared to determine SNPs. Compared to mutations identified on *S.* Mbandaka as host markers, all proteins identified were the same, bovine and poultry included. However, a difference was determined for 1 sequences (1 bovine). Compared to *S.* Mbandaka mutations described in the study, 1 genomes displays mutations on their amino acid sequence (ACT20SMb17 -> 1 mutation), but this SNP is synonymous and thus has no impact on the nucleotide sequence. Overall, even if there is a host difference, no alteration of the FimH gene has been observed.

Bibliography

[1] Christopher J. Griffith. "Food safety: where from and where to?" In: *British Food Journal* 108.1 (Jan. 2006). Ed. by Alex von Holy and Denise Lindsay. Publisher: Emerald Group Publishing Limited, pp. 6–15. DOI: 10.1108/00070700610637599.

[2] James Davis. "Baking for the common good: a reassessment of the assize of bread in Medieval England". en. In: *The Economic History Review* 57.3 (2004), pp. 465–502. DOI: 10.1111/j.1468-0289.2004.00285.x.

[3] Sandra Hoffmann. "U.S. Food Safety Policy Enters a New Era". In: *Amber Waves* (Jan. 2011).

[4] Robert V. Tauxe. "Emerging foodborne pathogens". en. In: *International Journal of Food Microbiology*. 18th International Symposium of the International Committee on Food Microbiology and Hygeine, August 18-23, 2002, Lillehammer Norway. Necessary and Unwanted Bacteria in Food - Microbial Adaption to changing Environments 78.1 (Sept. 2002), pp. 31–41. DOI: 10.1016/S0168-1605(02)00232-5.

[5] Patrick Grimont and François-Xavier Weill. *Antigenic Formulae of the Salmonella serovars, (9th ed.) Paris: WHO Collaborating Centre for Reference and Research on Salmonella*. Jan. 2007.

[6] F. Käferstein and M. Abdussalam. "Food safety in the 21st century." In: *Bulletin of the World Health Organization* 77.4 (1999), pp. 347–351.

[7] *Milk-Borne Salmonellosis – Illinois.*

[8] Christ-Donald Kaptchouang Tchatchouang et al. "Listeriosis Outbreak in South Africa: A Comparative Analysis with Previously Reported Cases Worldwide". In: *Microorganisms* 8.1 (Jan. 2020), p. 135. DOI: 10.3390/microorganisms8010135.

[9] Alexander Mellmann et al. "Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology". en. In: *PLOS ONE* 6.7 (2011). Number: 7 Publisher: Public Library of Science, e22751. DOI: 10.1371/journal.pone.0022751.

[10] Melissa G. Collier et al. "Outbreak of hepatitis A in the USA associated with frozen pomegranate arils imported from Turkey: an epidemiological case study". eng. In: *The Lancet. Infectious Diseases* 14.10 (Oct. 2014), pp. 976–981. DOI: 10.1016/S1473-3099(14)70883-7.

[11] B.p. Quinn and N.g. Marriott. "Haccp Plan Development and Assessment: A Review". en. In: *Journal of Muscle Foods* 13.4 (2002), pp. 313–330. DOI: 10.1111/j.1745-4573.2002.tb00339.x.

[12] Marriott. *Principles of Food Sanitation by Marriott,Norman G.. [1999,4th Edition.] Hardcover*. Aspen, Jan. 1999.

[13] Wilkie F. Harrigan. *Laboratory Methods in Food Microbiology*. English. 3rd edition. San Diego: Academic Press, Oct. 1998.

[14] Catherine Davis. "Enumeration of probiotic strains: Review of culture-dependent and alternative techniques to quantify viable bacteria". en. In: *Journal of Microbiological Methods* 103 (Aug. 2014), pp. 9–17. DOI: 10.1016/j.mimet.2014.04.012.

[15] Alexander Gill. "The Importance of Bacterial Culture to Food Microbiology in the Age of Genomics". In: *Frontiers in Microbiology* 8 (May 2017), p. 777. DOI: 10.3389/fmicb.2017.00777.

[16] Monalisha Nayak et al. "Integrated sorting, concentration and real time PCR based detection system for sensitive detection of microorganisms". en. In: *Scientific Reports* 3.1 (Nov. 2013). Number: 1 Publisher: Nature Publishing Group, p. 3266. DOI: 10.1038/srep03266.

[17] M. Moriconi et al. "Multiplex PCR-based identification of Streptococcus canis, Streptococcus zooepidemicus and Streptococcus dysgalactiae subspecies from dogs". eng. In: *Comparative Immunology, Microbiology and Infectious Diseases* 50 (Feb. 2017), pp. 48–53. DOI: 10.1016/j.cimid.2016.11.011.

[18] M. Aurora Echeita et al. "Multiplex PCR-based detection and identification of the most common Salmonella second-phase flagellar antigens". eng. In: *Research in Microbiology* 153.2 (Mar. 2002), pp. 107–113. DOI: 10.1016/s0923-2508(01)01295-5.

[19] Antonio C. G. Foddai and Irene R. Grant. "Methods for detection of viable foodborne pathogens: current state-of-art and future prospects". In: *Applied Microbiology and Biotechnology* 104.10 (2020), pp. 4281–4288. DOI: 10.1007/s00253-020-10542-x.

[20] Amy L. Peace-Brewer, David W. Craft, and John L. Schmitz. "Immunologic Techniques in the Clinical Microbiology Laboratory". en. In: *Laboratory Medicine* 31.1 (Jan. 2000), pp. 24–29. DOI: 10.1309/50DX-KYEG-91UR-UQ9V.

[21] Matthias Upmann and Christine Bonaparte. "RAPID METHODS FOR FOOD HYGIENE INSPECTION". en. In: *Encyclopedia of Food Microbiology*. Ed. by Richard K. Robinson. Oxford: Elsevier, Jan. 1999, pp. 1887–1895. DOI: 10.1006/rwfm.1999.1320.

[22] R H Yolken and P J Stopa. "Enzyme-linked fluorescence assay: Ultrasensitive solid-phase assay for detection of human rotavirus." In: *Journal of Clinical Microbiology* 10.3 (Sept. 1979), pp. 317–321.

[23] M. Schloter, B. Aßmus, and A. Hartmann. "The use of immunological methods to detect and identify bacteria in the environment". en. In: *Biotechnology Advances* 13.1 (Jan. 1995), pp. 75–90. DOI: 10.1016/0734-9750(94)00023-6.

[24] Sowmya Nagaraj et al. "Development of IgY based sandwich ELISA for the detection of staphylococcal enterotoxin G (SEG), an egc toxin". eng. In: *International Journal of Food Microbiology* 237 (Nov. 2016), pp. 136–141. DOI: 10.1016/j.ijfoodmicro.2016.08.009.

[25] Marc W Allard et al. "Genomics of foodborne pathogens for microbial food safety". en. In: *Current Opinion in Biotechnology*. Food biotechnology • Plant biotechnology 49 (Feb. 2018), pp. 224–229. DOI: 10.1016/j.copbio.2017.11.002.

[26] Marie Anne Chattaway et al. "The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales". eng. In: *Frontiers in Public Health* 7 (2019), p. 317. DOI: 10.3389/fpubh.2019.00317.

[27] Manal Mohammed and Salina Thapa. "Evaluation of WGS-subtyping methods for epidemiological surveillance of foodborne salmonellosis". In: *One Health Outlook* 2.1 (July 2020), p. 13. DOI: 10.1186/s42522-020-00016-5.

[28] E. Kurt Lienau et al. "Identification of a Salmonellosis Outbreak by Means of Molecular Sequencing". In: *New England Journal of Medicine* 364.10 (Mar. 2011). Number: 10 Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMc1100443, pp. 981–982. DOI: 10.1056/NEJMc1100443.

[29] Chorong Hahm, Hae-Sun Chung, and Miae Lee. "Whole-genome sequencing for the characterization of resistance mechanisms and epidemiology of colistin-resistant Acinetobacter baumannii". en. In: *PLOS ONE* 17.3 (Mar. 2022). Publisher: Public Library of Science, e0264335. DOI: 10.1371/journal.pone.0264335.

[30] Tim Dallman et al. "Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network". In: *Microbial Genomics* 2.8 (2016). DOI: 10.1099/mgen.0.000070.

[31] Marc W. Allard et al. "Whole genome sequencing uses for foodborne contamination and compliance: Discovery of an emerging contamination event in an ice cream facility using whole genome sequencing". en. In: *Infection, Genetics and Evolution* 73 (Sept. 2019), pp. 214–220. DOI: 10.1016/j.meegid.2019.04.026.

[32] Chishih Chu et al. "Evolution of genes on the Salmonella Virulence plasmid phylogeny revealed from sequencing of the virulence plasmids of S. enterica serotype Dublin and comparative analysis". eng. In: *Genomics* 92.5 (Nov. 2008), pp. 339–343. DOI: 10.1016/j.ygeno.2008.07.010.

[33] Samantha A. Naberhaus et al. "Pathogenicity and Competitive Fitness of Salmonella enterica Serovar 4,[5],12:i:- Compared to Salmonella Typhimurium and Salmonella Derby in Swine". In: *Frontiers in Veterinary Science* 6 (2020).

[34] Mariela E. Srednik et al. "Antimicrobial resistance and genomic characterization of Salmonella Dublin isolates in cattle from the United States". en. In: *PLOS ONE* 16.9 (Sept. 2021), e0249617. DOI: 10.1371/journal.pone.0249617.

[35] Sangeeta Banerji et al. "Genome-based Salmonella serotyping as the new gold standard". en. In: *Scientific Reports* 10.1 (Mar. 2020). Number: 1 Publisher: Nature Publishing Group, p. 4333. DOI: 10.1038/s41598-020-61254-1.

[36] Katherine A. Lau et al. "Proficiency testing for bacterial whole genome sequencing in assuring the quality of microbiology diagnostics in clinical and public health laboratories". en. In: *Pathology* 53.7 (Dec. 2021), pp. 902–911. DOI: 10.1016/j.pathol.2021.03.012.

[37] European Food Safety Authority and European Centre for Disease Prevention and Control. "The European Union One Health 2020 Zoonoses Report". en. In: *EFSA Journal* 19.12 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2021.6971, e06971. DOI: 10.2903/j.efsa.2021.6971.

[38] Daniel Dewey-Mattia et al. "Surveillance for Foodborne Disease Outbreaks, United States, 2017 Annual Report". en. In: (), p. 15.

[39] Frederick J. Angulo, Timothy F. Jones, and Frederick J. Angulo. "Eating in Restaurants: A Risk Factor for Foodborne Disease?" In: *Clinical Infectious Diseases* 43.10 (Nov. 2006), pp. 1324–1328. DOI: 10.1086/508540.

[40] Andrea Osimani, Lucia Aquilanti, and Francesca Clementi. "Bacillus cereus foodborne outbreaks in mass catering". en. In: *International Journal of Hospitality Management* 72 (June 2018), pp. 145–153. DOI: 10.1016/j.ijhm.2018.01.013.

[41] SPF. *Surveillance des toxi-infections alimentaires collectives. Données de la déclaration obligatoire, 2019.* fr.

[42] J M David et al. "Structure of the French farm-to-table sur- veillance system for Salmonella". en. In: *Revue Méd. Vét.* (2011), p. 12.

[43] Vincent Leclerc et al. "The Salmonella network: a surveillance scheme for Salmonella in the food chain: 2015 results". en. In: (), p. 7.

[44] CHU Robert Debré-APHP Institut Pasteur. "Rapport d'activité annuel 2019 Année d'exercice 2018." In: ().

[45] "Multi-country outbreak of Salmonella Enteritidis infections linked to eggs, third update". en. In: (2020), p. 22.

[46] *EFSA and ECDC investigate multi-country Salmonella outbreak linked to chocolate products | EFSA*. en.

[47] Stefanie Lüth et al. "Analysis of RASFF notifications on food products contaminated with Listeria monocytogenes reveals options for improvement in the rapid alert system for food and feed". en. In: *Food Control* 96 (Feb. 2019), pp. 479–487. DOI: 10.1016/j.foodcont.2018.09.033.

[48] *RASFF - food and feed safety alerts*. en.

[49] SPF. *Estimation de la morbidité et de la mortalité liées aux infections d'origine alimentaire en France métropolitaine, 2008-2013*. fr.

[50] Caterina Levantesi et al. "Salmonella in surface and drinking water: Occurrence and water-mediated transmission". en. In: *Food Research International*. Salmonella in Foods: Evolution, Strategies and Challenges 45.2 (Mar. 2012), pp. 587–602. DOI: 10.1016/j.foodres.2011.06.037.

[51] Jean-Rémy Sadeyen et al. "Salmonella carrier state in chicken: comparison of expression of immune response genes between susceptible and resistant animals". eng. In: *Microbes and Infection* 6.14 (Nov. 2004), pp. 1278–1286. DOI: 10.1016/j.micinf.2004.07.005.

[52] European Comission. *Alert and Cooperation Network - 2021 Annual Report*. 2021.

[53] European Comission. *Control of Salmonella*. en.

[54] EUROPEAN COMMISSION, Health and Consumers Directorate-General, and SANCO/ 2008/ E2/ 056. *Analysis of the costs and benefits of setting a target for the reduction of Salmonella in breeding pigs*. Mar. 2011.

[55] Plateforme de Surveillance de la Chaîne Alimentaire. *SURVEILLER SALMONELLA SPP EN FILIÈRE BOVINE DE FABRICATION DE FROMAGES AU LAIT CRU*. 2019.

[56] Laura Uelze et al. "Typing methods based on whole genome sequencing data". In: *One Health Outlook* 2.1 (Feb. 2020), p. 3. DOI: 10.1186/s42522-020-0010-1.

[57] Martin C. J. Maiden et al. "MLST revisited: the gene-by-gene approach to bacterial genomics". eng. In: *Nature Reviews. Microbiology* 11.10 (Oct. 2013), pp. 728–736. DOI: 10.1038/nrmicro3093.

[58] Keith A. Jolley et al. "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain". en. In: *Microbiology* 158.Pt 4 (Apr. 2012). Publisher: Microbiology Society, p. 1005. DOI: 10.1099/mic.0.055459-0.

[59] Matthew L. Ranieri et al. "Comparison of Typing Methods with a New Procedure Based on Sequence Characterization for Salmonella Serovar Prediction". In: *Journal of Clinical Microbiology* 51.6 (June 2013), pp. 1786–1797. DOI: 10.1128/JCM.03201-12.

[60] Taylor Davedow et al. "PulseNet International Survey on the Implementation of Whole Genome Sequencing in Low and Middle-Income Countries for Foodborne Disease Surveillance". en. In: *Foodborne Pathogens and Disease* (May 2022). Publisher: Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA. DOI: 10.1089/fpd.2021.0110.

[61] Suchawan Pornsukarom, Arnoud H. M. van Vliet, and Siddhartha Thakur. "Whole genome sequencing analysis of multiple Salmonella serovars provides insights into phylogenetic relatedness, antimicrobial resistance, and virulence markers across humans, food animals and agriculture environmental sources". en. In: *BMC Genomics* 19 (2018). Publisher: BioMed Central. DOI: 10.1186/s12864-018-5137-4.

[62] Nabil-Fareed Alikhan et al. "A genomic overview of the population structure of Salmonella". en. In: *PLoS Genetics* 14.4 (Apr. 2018). Publisher: Public Library of Science. DOI: 10.1371/journal.pgen.1007261.

[63] Zhemin Zhou et al. "GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens". en. In: *Genome Research* 28.9 (Sept. 2018). Publisher: Cold Spring Harbor Laboratory Press, p. 1395. DOI: 10.1101/gr.232397.117.

[64] M. C. Enright and B. G. Spratt. "Multilocus sequence typing". eng. In: *Trends in Microbiology* 7.12 (Dec. 1999), pp. 482–487. DOI: 10.1016/s0966-842x(99)01609-1.

[65] Keith A. Jolley, James E. Bray, and Martin C. J. Maiden. "Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications". eng. In: *Wellcome Open Research* 3 (2018), p. 124. DOI: 10.12688/wellcomeopenres.14826.1.

[66] Zhemin Zhou et al. "The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity". eng. In: *Genome Research* 30.1 (Jan. 2020), pp. 138–152. DOI: 10.1101/gr.251678.119.

[67] Federica Palma et al. "In vitro and in silico parameters for precise cgMLST typing of Listeria monocytogenes". In: *BMC Genomics* 23.1 (Mar. 2022), p. 235. DOI: 10.1186/s12864-022-08437-4.

[68] Binghang Liu et al. *Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects*. arXiv:1308.2012 [q-bio]. Feb. 2020. DOI: 10.48550/arXiv.1308.2012.

[69] Sanzhen Liu et al. "Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement". en. In: *Scientific Reports* 7.1 (Feb. 2017). Number: 1 Publisher: Nature Publishing Group, p. 42444. DOI: 10.1038/srep42444.

[70] Anton Bankevich et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". eng. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19.5 (May 2012), pp. 455–477. DOI: 10.1089/cmb.2012.0021.

[71] Shahab Sarmashghi et al. "Skmer: assembly-free and alignment-free sample identification using genome skims". In: *Genome Biology* 20.1 (Feb. 2019), p. 34. DOI: 10.1186/s13059-019-1632-4.

[72] Maxime Déraspe et al. "Phenetic Comparison of Prokaryotic Genomes Using k-mers". In: *Molecular Biology and Evolution* 34.10 (Oct. 2017), pp. 2716–2729. DOI: 10.1093/molbev/msx200.

[73] Alon Kafri, Benny Chor, and David Horn. "Inter-chromosomal k-mer distances". In: *BMC Genomics* 22.1 (Sept. 2021), p. 644. DOI: 10.1186/s12864-021-07952-0.

[74] Ariane Bize et al. "Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history". In: *BMC Genomics* 22.1 (Mar. 2021), p. 186. DOI: 10.1186/s12864-021-07471-y.

[75] Pedro Feijao et al. "MentaLiST - A fast MLST caller for large MLST schemes". eng. In: *Microbial Genomics* 4.2 (Feb. 2018). DOI: 10.1099/mgen.0.000146.

[76] Yuval Bussi, Ruti Kapon, and Ziv Reich. "Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy". en. In: *PLOS ONE* 16.10 (Oct. 2021). Publisher: Public Library of Science, e0258693. DOI: 10.1371/journal.pone.0258693.

[77] Sophie Octavia and Ruiting Lan. "Single nucleotide polymorphism typing of global Salmonella enterica serovar Typhi isolates by use of a hairpin primer real-time PCR assay". eng. In: *Journal of Clinical Microbiology* 48.10 (Oct. 2010), pp. 3504–3509. DOI: 10.1128/JCM.00709-10.

[78] Disa L. Hammarlöf et al. "Role of a single noncoding nucleotide in the evolution of an epidemic African clade of Salmonella". en. In: *Proceedings of the National Academy of Sciences* 115.11 (Mar. 2018). ISBN: 9781714718115 Publisher: National Academy of Sciences Section: PNAS Plus, E2614–E2623. DOI: 10.1073/pnas.1714718115.

[79] Nicolas Radomski et al. "A Simple and Robust Statistical Method to Define Genetic Relatedness of Samples Related to Outbreaks at the Genomic Scale – Application to Retrospective Salmonella Foodborne Outbreak Investigations". In: *Frontiers in Microbiology* 10 (2019), p. 2413. DOI: 10.3389/fmicb.2019.02413.

[80] Assia Saltykova et al. "Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to Salmonella enterica serotype Typhimurium and serotype 1,4,[5],12:i:-". en. In: *PLOS ONE* 13.2 (2018). Publisher: Public Library of Science, e0192504. DOI: 10.1371/journal.pone.0192504.

[81] Jay Worley et al. "Salmonella enterica Phylogeny Based on Whole-Genome Sequencing Reveals Two New Clades and Novel Patterns of Horizontally Acquired Genetic Elements". eng. In: *mBio* 9.6 (Nov. 2018), e02303–18. DOI: 10.1128/mBio.02303-18.

[82] null Rokas and null Holland. "Rare genomic changes as a tool for phylogenetics". eng. In: *Trends in Ecology & Evolution* 15.11 (Nov. 2000), pp. 454–459. DOI: 10.1016/s0169-5347(00)01967-4.

[83] Benjamin D Redelings and Marc A Suchard. "Incorporating indel information into phylogeny estimation for rapidly emerging pathogens". In: *BMC Evolutionary Biology* 7 (Mar. 2007), p. 40. DOI: 10.1186/1471-2148-7-40.

[84] Adam D. Ewing and Haig H. Kazazian. "Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans". en. In: *Genome Research* 21.6 (June 2011). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 985–990. DOI: 10.1101/gr.114777.110.

[85] Xiangyu Deng et al. "Genomic Epidemiology of Salmonella enterica Serotype Enteritidis based on Population Structure of Prevalent Lineages". en. In: *Emerging Infectious Diseases* 20.9 (Sept. 2014). Publisher: Centers for Disease Control and Prevention, p. 1481. DOI: 10.3201/eid2009.131095.

[86] Zhemin Zhou et al. "Neutral Genomic Microevolution of a Recently Emerged Pathogen, Salmonella enterica Serovar Agona". en. In: *PLOS Genetics* 9.4 (2013). Publisher: Public Library of Science, e1003471. DOI: 10.1371/journal.pgen.1003471.

[87] Leen Baert et al. "Genetic changes are introduced by repeated exposure of Salmonella spiked in low water activity and high fat matrix to heat". en. In: *Scientific Reports* 11.1 (Dec. 2021). DOI: 10.1038/s41598-021-87330-8.

[88] Robert A. Power, Julian Parkhill, and Tulio de Oliveira. "Microbial genome-wide association studies: lessons from human GWAS". en. In: *Nature Reviews Genetics* 18.1 (Jan. 2017). Number: 1 Publisher: Nature Publishing Group, pp. 41–50. DOI: 10.1038/nrg.2016.132.

[89] Madison E. Pearce et al. "Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak". In: *International Journal of Food Microbiology* 274 (June 2018), pp. 1–11. DOI: 10.1016/j.ijfoodmicro.2018.02.023.

[90] Dai Yoshimura et al. "Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP". In: *Microbial Genomics* 5.5 (May 2019), e000261. DOI: 10.1099/mgen.0.000261.

[91] Tae-Ho Lee et al. "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data". In: *BMC Genomics* 15.1 (Feb. 2014), p. 162. DOI: 10.1186/1471-2164-15-162.

[92] Derek S. Sarovich and Erin P. Price. "SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets". In: *BMC Research Notes* 7.1 (Sept. 2014), p. 618. DOI: 10.1186/1756-0500-7-618.

[93] Todd J. Treangen et al. "The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes". In: *Genome Biology* 15.11 (Nov. 2014), p. 524. DOI: 10.1186/s13059-014-0524-x.

[94] Stephen J Bush et al. "Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines". In: *GigaScience* 9.2 (Feb. 2020), giaa007. DOI: 10.1093/gigascience/giaa007.

[95] Erik Garrison and Gabor Marth. *Haplotype-based variant detection from short-read sequencing*. Tech. rep. arXiv:1207.3907. arXiv:1207.3907 [q-bio] type: article. arXiv, July 2012. DOI: 10.48550/arXiv.1207.3907.

[96] U. Römling and B. Tümmler. "Bacterial genome mapping". eng. In: *Journal of Biotechnology* 35.2-3 (June 1994), pp. 155–164. DOI: 10.1016/0168-1656(94)90033-7.

[97] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". eng. In: *Bioinformatics (Oxford, England)* 25.14 (July 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[98] Meryl Vila Nova et al. "Genetic and metabolic signatures of Salmonella enterica subsp. enterica associated with animal sources at the pangenomic scale". In: *BMC Genomics* 20.1 (Nov. 2019), p. 814. DOI: 10.1186/s12864-019-6188-x.

[99] Andrew J. Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". en. In: *Bioinformatics* 31.22 (Nov. 2015), pp. 3691–3693. DOI: 10.1093/bioinformatics/btv421.

[100] Gerry Tonkin-Hill et al. "Producing polished prokaryotic pangenomes with the Panaroo pipeline". eng. In: *Genome Biology* 21.1 (July 2020), p. 180. DOI: 10.1186/s13059-020-02090-4.

[101] Hao Gong et al. "A Salmonella Small Non-Coding RNA Facilitates Bacterial Invasion and Intracellular Replication by Modulating the Expression of Virulence Factors". en. In: *PLOS Pathogens* 7.9 (Sept. 2011), e1002120. DOI: 10.1371/journal.ppat.1002120.

[102] Rachel M. Colquhoun et al. "Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs". In: *Genome Biology* 22.1 (Sept. 2021), p. 267. DOI: 10.1186/s13059-021-02473-1.

[103] Linto Antony et al. "Population structure of Salmonella enterica serotype Mbandaka reveals similar virulence potential irrespective of source and phylogenomic stratification". en. In: *F1000 Research* 9 (Sept. 2020), p. 1142. DOI: 10.12688/f1000research.25540.1.

[104] Ruth E. Timme et al. "Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance". en. In: *PeerJ* 5 (Oct. 2017), e3893. DOI: 10.7717/peerj.3893.

[105] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". en. In: *PLOS ONE* 5.3 (Mar. 2010). Publisher: Public Library of Science, e9490. DOI: 10.1371/journal.pone.0009490.

[106] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". en. In: *Bioinformatics* 30.9 (May 2014), pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033.

[107] Lam-Tung Nguyen et al. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies". en. In: *Molecular Biology and Evolution* 32.1 (Jan. 2015), pp. 268–274. DOI: 10.1093/molbev/msu300.

[108] Fredrik Ronquist et al. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space". In: *Systematic Biology* 61.3 (May 2012), pp. 539–542. DOI: 10.1093/sysbio/sys029.

[109] Remco Bouckaert et al. "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". en. In: *PLOS Computational Biology* 15.4 (2019). Publisher: Public Library of Science, e1006650. DOI: 10.1371/journal.pcbi.1006650.

[110] Xavier Didelot and Daniel J. Wilson. "ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes". en. In: *PLOS Computational Biology* 11.2 (2015), e1004041. DOI: 10.1371/journal.pcbi.1004041.

[111] Ziheng Yang and Bruce Rannala. "Molecular phylogenetics: principles and practice". en. In: *Nature Reviews Genetics* 13.5 (May 2012), pp. 303–314. DOI: 10.1038/nrg3186.

[112] Olga Chernomor, Arndt von Haeseler, and Bui Quang Minh. "Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices". In: *Systematic Biology* 65.6 (Nov. 2016), pp. 997–1008. DOI: 10.1093/sysbio/syw037.

[113] Pranjal Vachaspati and Tandy Warnow. "FastRFS: fast and accurate Robinson-Foulds Supertrees using constrained exact optimization". eng. In: *Bioinformatics (Oxford, England)* 33.5 (Mar. 2017), pp. 631–639. DOI: 10.1093/bioinformatics/btw600.

[114] Zhemin Zhou et al. "Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia". In: *Current Biology* 28.15 (Aug. 2018), 2420–2428.e10. DOI: 10.1016/j.cub.2018.05.058.

[115] Laura Ford et al. "Cost of whole genome sequencing for non-typhoidal Salmonella enterica". en. In: *PLOS ONE* 16.3 (Mar. 2021). Publisher: Public Library of Science, e0248561. DOI: 10.1371/journal.pone.0248561.

[116] Jian Ye et al. "Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction". In: *BMC Bioinformatics* 13.1 (June 2012), p. 134. DOI: 10.1186/1471-2105-13-134.

[117] Matthias Dreier et al. "SpeciesPrimer: a bioinformatics pipeline dedicated to the design of qPCR primers for the quantification of bacterial species". en. In: *PeerJ* 8 (2020). Publisher: PeerJ, Inc. DOI: 10.7717/peerj.8544.

[118] Yueni Wu et al. "ARDEP, a Rapid Degenerate Primer Design Pipeline Based on k-mers for Amplicon Microbiome Studies". In: *International Journal of Environmental Research and Public Health* 17.16 (Aug. 2020), p. 5958. DOI: 10.3390/ijerph17165958.

[119] *FullSSR*. en.

[120] Anson V. Koehler et al. "Use of a bioinformatic-assisted primer design strategy to establish a new nested PCR-based method for Cryptosporidium". In: *Parasites & Vectors* 10 (Oct. 2017), p. 509. DOI: 10.1186/s13071-017-2462-4.

[121] Greg Peterson et al. "Development of microarray and multiplex polymerase chain reaction assays for identification of serovars and virulence genes in Salmonella enterica of human or animal origin". eng. In: *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc* 22.4 (July 2010), pp. 559–569. DOI: 10.1177/104063871002200410.

[122] S.c. Ricke et al. "Molecular-based identification and detection of Salmonella in food production systems: current perspectives". en. In: *Journal of Applied Microbiology* 125.2 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jam.13888, pp. 313–327. DOI: 10.1111/jam.13888.

[123] Dan Xiong et al. "An Efficient Multiplex PCR-Based Assay as a Novel Tool for Accurate Inter-Serovar Discrimination of Salmonella Enteritidis, S. Pullorum/Gallinarum and S. Dublin". en. In: *Frontiers in Microbiology* 8 (2017). Publisher: Frontiers Media SA. DOI: 10.3389/fmicb.2017.00420.

[124] S. H. Park and S. C. Ricke. "Development of multiplex PCR assay for simultaneous detection of Salmonella genus, Salmonella subspecies I, Salm. Enteritidis, Salm. Heidelberg and Salm. Typhimurium". eng. In: *Journal of Applied Microbiology* 118.1 (Jan. 2015), pp. 152–160. DOI: 10.1111/jam.12678.

[125] Si Hong Park et al. "Current and emerging technologies for rapid detection and characterization of Salmonella in poultry and poultry products". eng. In: *Food microbiology.* 38 (Apr. 2014), pp. 250–262. DOI: 10.1016/j.fm.2013.10.002.

[126] Laura Ford et al. "Incorporating Whole-Genome Sequencing into Public Health Surveillance: Lessons from Prospective Sequencing of Salmonella Typhimurium in Australia". In: *Foodborne Pathogens and Disease* 15.3 (Jan. 2018), pp. 161–167. DOI: 10.1089/fpd.2017.2352.

[127] Andrew J. Low et al. "ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data". eng. In: *PeerJ* 7 (2019), e6995. DOI: 10.7717/peerj.6995.

[128] Alexandra Moura et al. "Real-Time Whole-Genome Sequencing for Surveillance of Listeria monocytogenes, France". In: *Emerging Infectious Diseases* 23.9 (Sept. 2017), pp. 1462–1470. DOI: 10.3201/eid2309.170336.

[129] Shannon E. Majowicz et al. "The Global Burden of Nontyphoidal Salmonella Gastroenteritis". en. In: *Clinical Infectious Diseases* 50.6 (Mar. 2010), pp. 882–889. DOI: 10.1086/650733.

[130] Shu-Kee Eng et al. "Salmonella: A review on pathogenesis, epidemiology and antibiotic resistance". In: *Frontiers in Life Science* 8.3 (July 2015). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21553769.2015.1051243, pp. 284–293. DOI: 10.1080/21553769.2015.1051243.

[131] Robert A. Kingsley and Andreas J. Bäumler. "Host adaptation and the emergence of infectious disease: the Salmonella paradigm". en. In: *Molecular Microbiology* 36.5 (2000). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.2000.01907.x, pp. 1006–1014. DOI: 10.1046/j.1365-2958.2000.01907.x.

[132] Bryan Coburn, Guntram A Grassl, and B B Finlay. "Salmonella, the host and disease: a brief review". en. In: *Immunology & Cell Biology* 85.2 (2007), pp. 112–118. DOI: 10.1038/sj.icb.7100007.

[133] Daniel Hurley et al. "Salmonella-host interactions - modulation of the host innate immune system". eng. In: *Frontiers in Immunology* 5 (2014), p. 481. DOI: 10.3389/fimmu.2014.00481.

[134] F. W. Brenner et al. "Salmonella Nomenclature". In: *Journal of Clinical Microbiology* 38.7 (July 2000), pp. 2465–2467.

[135] Keith D. MacKenzie et al. "Examining the Link between Biofilm Formation and the Ability of Pathogenic Salmonella Strains to Colonize Multiple Host Species". In: *Frontiers in Veterinary Science* 4 (Aug. 2017), p. 138. DOI: 10.3389/fvets.2017.00138.

[136] Karin Hoelzer, Andrea Isabel Moreno Switt, and Martin Wiedmann. "Animal contact as a source of human non-typhoidal salmonellosis". In: *Veterinary Research* 42.1 (Feb. 2011), p. 34. DOI: 10.1186/1297-9716-42-34.

[137] S. Uzzau et al. "Host adapted serotypes of Salmonella enterica." en. In: *Epidemiology and Infection* 125.2 (Oct. 2000), p. 229. DOI: 10.1017/s0950268899004379.

[138] Wolfgang Rabsch et al. "Salmonella enterica Serotype Typhimurium and Its Host-Adapted Variants". In: *Infection and Immunity* 70.5 (May 2002), pp. 2249–2255. DOI: 10.1128/IAI.70.5.2249-2255.2002.

[139] Anna Fàbrega and Jordi Vila. "Salmonella enterica Serovar Typhimurium Skills To Succeed in the Host: Virulence and Regulation". In: *Clinical Microbiology Reviews* 26.2 (Apr. 2013), pp. 308–341. DOI: 10.1128/CMR.00066-12.

[140] ANSES. *Fiche de description de danger biologique transmissible par les aliments : Salmonella spp.* 2021.

[141] H. G. Bayne, J. A. Garibaldi, and H. Lineweaver. "Heat resistance of Salmonella typhimurium and Salmonella senftenberg 775 W in chicken meat". eng. In: *Poultry Science* 44.5 (Sept. 1965), pp. 1281–1284. DOI: 10.3382/ps.0441281.

[142] Li Ma et al. "Thermal inactivation of Salmonella in peanut butter". eng. In: *Journal of Food Protection* 72.8 (Aug. 2009), pp. 1596–1601. DOI: 10.4315/0362-028x-72.8.1596.

[143] Rafaela G. Ferrari, Pedro H. N. Panzenhagen, and Carlos A. Conte-Junior. "Phenotypic and Genotypic Eligible Methods for Salmonella Typhimurium Source Tracking". In: *Frontiers in Microbiology* 8 (Dec. 2017), p. 2587. DOI: 10.3389/fmicb.2017.02587.

[144] "The Genus Salmonella Lignières, 1900". In: *The Journal of Hygiene* 34.3 (Oct. 1934), pp. 333–350.

[145] Nancy A. Strockbine et al. "Escherichia, Shigella, and Salmonella". en. In: *Manual of Clinical Microbiology*. John Wiley & Sons, Ltd, 2015, pp. 685–713.

[146] Pierre Wattiau, Cécile Boland, and Sophie Bertrand. "Methodologies for Salmonella enterica subsp. enterica Subtyping: Gold Standards and Alternatives". In: *Applied and Environmental Microbiology* 77.22 (Nov. 2011), pp. 7877–7885. DOI: 10.1128/AEM.05527-11.

[147] P. R. Reeves et al. "Bacterial polysaccharide synthesis and gene nomenclature". eng. In: *Trends in Microbiology* 4.12 (Dec. 1996), pp. 495–503. DOI: 10.1016/s0966-842x(97)82912-5.

[148] H. Herikstad, Y. Motarjemi, and R. V. Tauxe. "Salmonella surveillance: a global survey of public health serotyping". eng. In: *Epidemiology and Infection* 129.1 (Aug. 2002), pp. 1–8. DOI: 10.1017/s0950268802006842.

[149] H. Y. Cai et al. "Development of a Novel Protein Microarray Method for Serotyping Salmonella enterica Strains". In: *Journal of Clinical Microbiology* 43.7 (July 2005), pp. 3427–3430. DOI: 10.1128/JCM.43.7.3427-3430.2005.

[150] John R. McQuiston et al. "Molecular Determination of H Antigens of Salmonella by Use of a Microsphere-Based Liquid Array". In: *Journal of Clinical Microbiology* 49.2 (Feb. 2011). Publisher: American Society for Microbiology, pp. 565–573. DOI: 10.1128/JCM.01323-10.

[151] Wolfgang Rabsch. "Salmonella typhimurium phage typing for pathogens". eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 394 (2007), pp. 177–211. DOI: 10.1007/978-1-59745-512-1_10.

[152] Manal Mohammed. "Phage typing or CRISPR typing for epidemiological surveillance of Salmonella Typhimurium?" In: *BMC Research Notes* 10.1 (Nov. 2017), p. 578. DOI: 10.1186/s13104-017-2878-0.

[153] L. R. Ward, J. D. de Sa, and B. Rowe. "A phage-typing scheme for Salmonella enteritidis". eng. In: *Epidemiology and Infection* 99.2 (Oct. 1987), pp. 291–294. DOI: 10.1017/s0950268800067765.

[154] D. M. Olive and P. Bean. "Principles and applications of methods for DNA-based typing of microbial organisms". eng. In: *Journal of Clinical Microbiology* 37.6 (June 1999), pp. 1661–1669. DOI: 10.1128/JCM.37.6.1661-1669.1999.

[155] F. C. Tenover et al. "Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing". eng. In: *Journal of Clinical Microbiology* 33.9 (Sept. 1995), pp. 2233–2239. DOI: 10.1128/jcm.33.9.2233-2239.1995.

[156] J. Garaizar et al. "Suitability of PCR fingerprinting, infrequent-restriction-site PCR, and pulsed-field gel electrophoresis, combined with computerized gel analysis, in library typing of Salmonella enterica serovar enteritidis". eng. In: *Applied and Environmental Microbiology* 66.12 (Dec. 2000), pp. 5273–5281. DOI: 10.1128/AEM.66.12.5273-5281.2000.

[157] Efrain M. Ribot et al. "PulseNet: Entering the Age of Next-Generation Sequencing". In: *Foodborne Pathogens and Disease* 16.7 (July 2019), pp. 451–456. DOI: 10.1089/fpd.2019.2634.

[158] Annaëlle Kerouanton et al. "Genetic diversity and antimicrobial resistance profiles of Salmonella enterica serotype derby isolated from pigs, pork, and humans in France". eng. In: *Foodborne Pathogens and Disease* 10.11 (Nov. 2013), pp. 977–984. DOI: 10.1089/fpd.2013.1537.

[159] Silvia Herrera-León et al. "Blind comparison of traditional serotyping with three multiplex PCRs for the identification of Salmonella serotypes". eng. In: *Research in Microbiology* 158.2 (Mar. 2007), pp. 122–127. DOI: 10.1016/j.resmic.2006.09.009.

[160] J. R. McQuiston et al. "Sequencing and comparative analysis of flagellin genes fliC, fljB, and flpA from Salmonella". eng. In: *Journal of Clinical Microbiology* 42.5 (May 2004), pp. 1923–1932. DOI: 10.1128/JCM.42.5.1923-1932.2004.

[161] Collette Fitzgerald et al. "Multiplex, bead-based suspension array for molecular determination of common Salmonella serogroups". eng. In: *Journal of Clinical Microbiology* 45.10 (Oct. 2007), pp. 3323–3334. DOI: 10.1128/JCM.00025-07.

[162] Seonghan Kim et al. "Multiplex PCR-Based Method for Identification of Common Clinical Serotypes of Salmonella enterica subsp. enterica". en. In: *Journal of Clinical Microbiology* 44.10 (Oct. 2006). Publisher: American Society for Microbiology (ASM), p. 3608. DOI: 10.1128/JCM.00701-06.

[163] Asma Afshari et al. "Salmonella Enteritidis and Salmonella Typhimorium identification in poultry carcasses". en. In: *Iranian Journal of Microbiology* 10.1 (Feb. 2018). Publisher: Tehran University of Medical Sciences, p. 45.

[164] Caroline M. O'Hara. "Manual and Automated Instrumentation for Identification of Enterobacteriaceae and Other Aerobic Gram-Negative Bacilli". In: *Clinical Microbiology Reviews* 18.1 (Jan. 2005), pp. 147–162. DOI: 10.1128/CMR.18.1.147-162.2005.

[165] Wen Zou et al. "Prediction System for Rapid Identification of Salmonella Serotypes Based on Pulsed-Field Gel Electrophoresis Fingerprints". In: *Journal of Clinical Microbiology* 50.5 (May 2012), pp. 1524–1532. DOI: 10.1128/JCM.00111-12.

[166] Annaëlle Kérouanton et al. "Pulsed-field gel electrophoresis subtyping database for foodborne Salmonella enterica serotype discrimination". eng. In: *Foodborne Pathogens and Disease* 4.3 (2007), pp. 293–303. DOI: 10.1089/fpd.2007.0090.

[167] Peter J. Hume et al. "Swiss Army Pathogen: The Salmonella Entry Toolkit". In: *Frontiers in Cellular and Infection Microbiology* 7 (2017).

[168] Kelly Hallstrom and Beth A. McCormick. "Salmonella Interaction with and Passage through the Intestinal Mucosa: Through the Lens of the Organism". In: *Frontiers in Microbiology* 2 (Apr. 2011), p. 88. DOI: 10.3389/fmicb.2011.00088.

[169] *Impact of Salmonella enterica Type III Secretion System Effectors on the Eukaryotic Host Cell.*

[170] Jason R Devlin et al. *Salmonella enterica serovar Typhimurium chitinases modulate the intestinal glycome and promote small intestinal invasion*. en. preprint. Microbiology, Dec. 2021. DOI: 10.1101/2021.12.06.471358.

[171] John A. Crump and Eric D. Mintz. "Global trends in typhoid and paratyphoid fever". In: *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 50.2 (Jan. 2010), pp. 241–246. DOI: 10.1086/649541.

[172] John A. Crump et al. "Estimating the Incidence of Typhoid Fever and Other Febrile Illnesses in Developing Countries". In: *Emerging Infectious Diseases* 9.5 (May 2003), pp. 539–544. DOI: 10.3201/eid0905.020428.

[173] John A Crump. "Progress in Typhoid Fever Epidemiology". In: *Clinical Infectious Diseases* 68.Supplement_1 (Feb. 2019), S4–S9. DOI: 10.1093/cid/ciy846.

[174] Geoffrey C. Buckle, Christa L. Fischer Walker, and Robert E. Black. "Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010". In: *Journal of Global Health* 2.1 (June 2012), p. 010401. DOI: 10.7189/jogh.02.010401.

[175] Jeffrey D. Stanaway et al. "The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017". English. In: *The Lancet Infectious Diseases* 19.4 (Apr. 2019). Publisher: Elsevier, pp. 369–381. DOI: 10.1016/S1473-3099(18)30685-6.

[176] Nicholas A Feasey et al. "Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa". In: *Lancet (London, England)* 379.9835 (June 2012), pp. 2489–2499. DOI: 10.1016/S0140-6736(11)61752-2.

[177] Andreas J. Bäumler et al. "Evolution of Host Adaptation in Salmonella enterica". In: *Infection and Immunity* 66.10 (Oct. 1998), pp. 4579–4587.

[178] Grammato Evangelopoulou et al. "Animal salmonelloses: a brief review of "host adaptation and host specificity" of Salmonella spp." en. In: *Veterinary World* 6.10 (July 2013), pp. 703–708. DOI: 10.14202/vetworld.2013.703-708.

[179] Nobuo Arai et al. "Salmonella genomic island 3 is an integrative and conjugative element and contributes to copper and arsenic resistance of Salmonella enterica". en. In: *bioRxiv* (Mar. 2019), p. 564534. DOI: 10.1101/564534.

[180] Liljana Petrovska et al. "Microevolution of Monophasic Salmonella Typhimurium during Epidemic, United Kingdom, 2005–2010". In: *Emerging Infectious Diseases* 22.4 (Apr. 2016), pp. 617–624. DOI: 10.3201/eid2204.150531.

[181] Roman G. Gerlach and Michael Hensel. "Salmonella pathogenicity islands in host specificity, host pathogen-interactions and antibiotics resistance of Salmonella enterica". eng. In: *Berliner Und Munchener Tierarztliche Wochenschrift* 120.7-8 (Aug. 2007), pp. 317–327.

[182] France Daigle. "Typhi genes expressed during infection or involved in pathogenesis". eng. In: *Journal of Infection in Developing Countries* 2.6 (Dec. 2008), pp. 431–437. DOI: 10.3855/jidc.157.

[183] Madeleine A. Wemyss and Jaclyn S. Pearson. "Host Cell Death Responses to Non-typhoidal Salmonella Infection". en. In: *Frontiers in Immunology* 10 (2019). Publisher: Frontiers Media SA. DOI: 10.3389/fimmu.2019.01758.

[184] Camille Cavestri et al. "Salmonella enterica subsp. enterica virulence potential can be linked to higher survival within a dynamic in vitro human gastrointestinal model". en. In: *Food Microbiology* 101 (Feb. 2022), p. 103877. DOI: 10.1016/j.fm.2021.103877.

[185] Min Yue and Dieter M. Schifferli. "Allelic variation in Salmonella: an underappreciated driver of adaptation and virulence". In: *Frontiers in Microbiology* 4 (Jan. 2014), p. 419. DOI: 10.3389/fmicb.2013.00419.

[186] Sarika Kombade and Navneet Kaur. *Pathogenicity Island in <em>Salmonella</em>*. en. Publication Title: Salmonella spp. - A Global Challenge. IntechOpen, Mar. 2021. DOI: 10.5772/intechopen.96443.

[187] Sébastien C. Sabbagh et al. “So similar, yet so different: uncovering distinctive features in the genomes of Salmonella enterica serovars Typhimurium and Typhi”. eng. In: *FEMS microbiology letters* 305.1 (Apr. 2010), pp. 1–13. DOI: 10.1111/j.1574-6968.2010.01904.x.

[188] Nicole A. Lerminiaux, Keith D. MacKenzie, and Andrew D. S. Cameron. “Salmonella Pathogenicity Island 1 (SPI-1): The Evolution and Stabilization of a Core Genomic Type Three Secretion System”. In: *Microorganisms* 8.4 (Apr. 2020), p. 576. DOI: 10.3390/microorganisms8040576.

[189] Lixin Lou et al. “Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network”. In: *Frontiers in Cellular and Infection Microbiology* 9 (2019).

[190] Brianne J. Burkinshaw and Natalie C. J. Strynadka. “Assembly and structure of the T3SS”. en. In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*. Protein trafficking and secretion in bacteria 1843.8 (Aug. 2014), pp. 1649–1663. DOI: 10.1016/j.bbamcr.2014.01.035.

[191] *Interactions of the Transmembrane Polymeric Rings of the Salmonella enterica Serovar Typhimurium Type III Secretion System | mBio.*

[192] Andrew C. McShan et al. “Characterization of the Binding of Hydroxyindole, Indoleacetic acid, and Morpholinoaniline to the Salmonella Type III Secretion System Proteins SipD and SipB”. en. In: *ChemMedChem* 11.9 (2016), pp. 963–971. DOI: 10.1002/cmdc.201600065.

[193] Elliott Jennings, Teresa L. M. Thurston, and David W. Holden. “Salmonella SPI-2 Type III Secretion System Effectors: Molecular Mechanisms And Physiological Consequences”. en. In: *Cell Host & Microbe* 22.2 (Aug. 2017), pp. 217–231. DOI: 10.1016/j.chom.2017.07.009.

[194] Rita Figueira and David W.YR 2012 Holden. “Functions of the Salmonella pathogenicity island 2 (SPI-2) type III secretion system effectors”. In: *Microbiology* 158.5 (). Publisher: Microbiology Society, pp. 1147–1161. DOI: 10.1099/mic.0.058115-0.

[195] Anne-Béatrice Blanc-Potard et al. “The SPI-3 Pathogenicity Island of Salmonella enterica”. en. In: *Journal of Bacteriology* 181.3 (Feb. 1999). Publisher: American Society for Microbiology (ASM), p. 998. DOI: 10.1128/jb.181.3.998-1004.1999.

[196] Eirwen Morgan et al. “SiiE is secreted by the Salmonella enterica serovar Typhimurium pathogenicity island 4-encoded secretion system and contributes to intestinal colonization in cattle”. eng. In: *Infection and Immunity* 75.3 (Mar. 2007), pp. 1524–1533. DOI: 10.1128/IAI.01438-06.

[197] Eirwen Morgan et al. “Identification of host-specific colonization factors of Salmonella enterica serovar Typhimurium”. en. In: *Molecular Microbiology* 54.4 (2004), pp. 994–1010. DOI: 10.1111/j.1365-2958.2004.04323.x.

[198] M. W. Wood et al. “Identification of a pathogenicity island required for Salmonella enteropathogenicity”. eng. In: *Molecular Microbiology* 29.3 (Aug. 1998), pp. 883–891. DOI: 10.1046/j.1365-2958.1998.00984.x.

[199] J. Parkhill et al. “Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18”. en. In: *Nature* 413.6858 (Oct. 2001). Number: 6858 Publisher: Nature Publishing Group, pp. 848–852. DOI: 10.1038/35101607.

[200] Ana M. Tomljenovic-Berube et al. “Mapping and Regulation of Genes within Salmonella Pathogenicity Island 12 That Contribute to In Vivo Fitness of Salmonella enterica Serovar Typhimurium”. en. In: *Infection and Immunity* 81.7 (July 2013). Publisher: American Society for Microbiology (ASM), p. 2394. DOI: 10.1128/IAI.00067-13.

[201] Jacob R. Elder et al. “Genomic organization and role of SPI-13 in nutritional fitness of Salmonella”. en. In: *International Journal of Medical Microbiology* 308.8 (Dec. 2018), pp. 1043–1052. DOI: 10.1016/j.ijmm.2018.10.004.

[202] Devendra H. Shah et al. “Identification of Salmonella gallinarum virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis”. In: *Microbiology* 151.12 (). Publisher: Microbiology Society, pp. 3957–3968. DOI: 10.1099/mic.0.28126-0.

[203] Devendra H. Shah et al. "Transposon Mutagenesis of Salmonella enterica Serovar Enteritidis Identifies Genes That Contribute to Invasiveness in Human and Chicken Cells and Survival in Egg Albumen". In: *Infection and Immunity* 80.12 (Dec. 2012), pp. 4203–4215. DOI: 10.1128/IAI.00790-12.

[204] Lingyan Jiang et al. "Signal transduction pathway mediated by the novel regulator LoiA for low oxygen tension induced Salmonella Typhimurium invasion". en. In: *PLOS Pathogens* 13.6 (2017). Publisher: Public Library of Science, e1006429. DOI: 10.1371/journal.ppat.1006429.

[205] Georgios S. Vernikos and Julian Parkhill. "Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands". In: *Bioinformatics* 22.18 (Sept. 2006), pp. 2196–2203. DOI: 10.1093/bioinformatics/btl369.

[206] Rafal Kolenda, Maciej Ugorski, and Krzysztof Grzymajlo. "Everything You Always Wanted to Know About Salmonella Type 1 Fimbriae, but Were Afraid to Ask". In: *Frontiers in Microbiology* 10 (May 2019), p. 1017. DOI: 10.3389/fmicb.2019.01017.

[207] Robert A. Edwards, Dieter M. Schifferli, and Stanley R. Maloy. "A role for Salmonella fimbriae in intraperitoneal infections". In: *Proceedings of the National Academy of Sciences* 97.3 (Feb. 2000). Publisher: Proceedings of the National Academy of Sciences, pp. 1258–1262. DOI: 10.1073/pnas.97.3.1258.

[208] J. P. Duguid. "Fimbriae and adhesive properties in Klebsiella strains". eng. In: *Journal of General Microbiology* 21 (Aug. 1959), pp. 271–286. DOI: 10.1099/00221287-21-1-271.

[209] David G. Thanassi et al. "Fimbriae: Classification and Biochemistry". In: *EcoSal Plus* 2.2 (Aug. 2007). Publisher: American Society for Microbiology. DOI: 10.1128/ecosalplus.2.4.2.1.

[210] Min Yue et al. "Allelic variation contributes to bacterial host specificity". eng. In: *Nature Communications* 6 (Oct. 2015), p. 8754. DOI: 10.1038/ncomms9754.

[211] Sarah A. Zeiner, Brett E. Dwyer, and Steven Clegg. "FimA, FimF, and FimH Are Necessary for Assembly of Type 1 Fimbriae on Salmonella enterica Serovar Typhimurium". In: *Infection and Immunity* 80.9 (Sept. 2012), pp. 3289–3296. DOI: 10.1128/IAI.00331-12.

[212] Alessandra Carattoli. "Plasmid-Mediated Antimicrobial Resistance in Salmonella enterica". en. In: *Current Issues in Molecular Biology* 5.4 (Oct. 2003). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, pp. 113–122. DOI: 10.21775/cimb.005.113.

[213] R. Rotger and J. Casadesús. "The virulence plasmids of Salmonella". eng. In: *International Microbiology: The Official Journal of the Spanish Society for Microbiology* 2.3 (Sept. 1999), pp. 177–184.

[214] Brian M. M. Ahmer, Mimi Tran, and Fred Heffron. "The Virulence Plasmid of Salmonella typhimurium Is Self-Transmissible". en. In: *Journal of Bacteriology* 181.4 (Feb. 1999), pp. 1364–1368.

[215] Claudia Silva, José Luis Puente, and Edmundo Calva. "Salmonella virulence plasmid: pathogenesis and ecology". In: *Pathogens and Disease* 75.6 (Aug. 2017), ftx070. DOI: 10.1093/femspd/ftx070.

[216] Timothy J. Johnson et al. "Horizontal Gene Transfer of a ColV Plasmid Has Resulted in a Dominant Avian Clonal Type of Salmonella enterica Serovar Kentucky". en. In: *PLOS ONE* 5.12 (2010). Publisher: Public Library of Science, e15524. DOI: 10.1371/journal.pone.0015524.

[217] C. Lee Ventola. "The Antibiotic Resistance Crisis". In: *Pharmacy and Therapeutics* 40.4 (Apr. 2015), pp. 277–283.

[218] Brad Spellberg and David N. Gilbert. "The Future of Antibiotics and Resistance: A Tribute to a Career of Leadership by John Bartlett". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 59.Suppl 2 (Sept. 2014), S71–S75. DOI: 10.1093/cid/ciu392.

[219] Centers for Disease Control and Prevention (U.S.) *Antibiotic resistance threats in the United States, 2019*. en. Tech. rep. Centers for Disease Control and Prevention (U.S.), Nov. 2019. DOI: 10.15620/cdc:82532.

[220] Govindaraj V. Asokan et al. "WHO Global Priority Pathogens List: A Bibliometric Analysis of Medline-PubMed for Knowledge Mobilization to Infection Prevention and Control Practices in Bahrain". In: *Oman Medical Journal* 34.3 (May 2019), pp. 184–193. DOI: 10.5001/omj.2019.37.

[221] Jun Li et al. *Fluoroquinolone Resistance in Salmonella: Mechanisms, Fitness, and Virulence*. en. Publication Title: Salmonella - A Re-emerging Pathogen. IntechOpen, July 2018. DOI: 10.5772/intechopen.74699.

[222] Wim L. Cuypers et al. "Fluoroquinolone resistance in Salmonella: insights by whole-genome sequencing". In: *Microbial Genomics* 4.7 (July 2018), e000195. DOI: 10.1099/mgen.0.000195.

[223] Maria Sjölund-Karlsson et al. "Fluoroquinolone Susceptibility Testing of Salmonella enterica: Detection of Acquired Resistance and Selection of Zone Diameter Breakpoints for Levofloxacin and Ofloxacin". In: *Journal of Clinical Microbiology* 52.3 (Mar. 2014), pp. 877–884. DOI: 10.1128/JCM.02679-13.

[224] Xiaojie Qin et al. "Antibiotic Resistance of Salmonella Typhimurium Monophasic Variant 1,4,[5],12:i:-in China: A Systematic Review and Meta-Analysis". eng. In: *Antibiotics (Basel, Switzerland)* 11.4 (Apr. 2022), p. 532. DOI: 10.3390/antibiotics11040532.

[225] Alita R. Burmeister. "Horizontal Gene Transfer". In: *Evolution, Medicine, and Public Health* 2015.1 (Jan. 2015), pp. 193–194. DOI: 10.1093/emph/eov018.

[226] Elizabeth A. McMillan, Charlene R. Jackson, and Jonathan G. Frye. "Transferable Plasmids of Salmonella enterica Associated With Antibiotic Resistance Genes". In: *Frontiers in Microbiology* 11 (2020).

[227] Chris Smillie et al. "Mobility of Plasmids". In: *Microbiology and Molecular Biology Reviews : MMBR* 74.3 (Sept. 2010), pp. 434–452. DOI: 10.1128/MMBR.00020-10.

[228] B. Périchon and P. Courvalin. "Antibiotic Resistance". en. In: *Encyclopedia of Microbiology (Third Edition)*. Ed. by Moselio Schaechter. Oxford: Academic Press, Jan. 2009, pp. 193–204. DOI: 10.1016/B978-012373944-5.00218-2.

[229] Christopher M. Thomas. "Plasmid Incompatibility". en. In: *Molecular Life Sciences: An Encyclopedic Reference*. Ed. by Ellis Bell. New York, NY: Springer, 2021, pp. 1–3. DOI: 10.1007/978-1-4614-6436-5_565-2.

[230] Epiphanie Nyirabahizi et al. "Evaluation of Escherichia coli as an indicator for antimicrobial resistance in Salmonella recovered from the same food or animal ceca samples". en. In: *Food Control* 115 (Sept. 2020), p. 107280. DOI: 10.1016/j.foodcont.2020.107280.

[231] Magdalena Wiesner et al. "Association of virulence plasmid and antibiotic resistance determinants with chromosomal multilocus genotypes in Mexican Salmonella enterica serovar Typhimurium strains". In: *BMC Microbiology* 9.1 (July 2009), p. 131. DOI: 10.1186/1471-2180-9-131.

[232] Wenyao Chen et al. "IncHI2 Plasmids Are Predominant in Antibiotic-Resistant Salmonella Isolates". en. In: *Frontiers in Microbiology* 7 (2016). Publisher: Frontiers Media SA. DOI: 10.3389/fmicb.2016.01566.

[233] Renee S. Levings et al. "The Genomic Island SGI1, Containing the Multiple Antibiotic Resistance Region of Salmonella enterica Serovar Typhimurium DT104 or Variants of It, Is Widely Distributed in Other S. enterica Serovars". en. In: *Journal of Bacteriology* 187.13 (July 2005). Publisher: American Society for Microbiology (ASM), p. 4401. DOI: 10.1128/JB.187.13.4401-4409.2005.

[234] Steven P. Djordjevic et al. "Emergence and Evolution of Multiply Antibiotic-Resistant Salmonella enterica Serovar Paratyphi B d-Tartrate-Utilizing Strains Containing SGI1". In: *Antimicrobial Agents and Chemotherapy* 53.6 (June 2009), pp. 2319–2326. DOI: 10.1128/AAC.01532-08.

[235] Ross C. Beier et al. "Disinfectant and Antimicrobial Susceptibility Profiles of Salmonella Strains from Feedlot Water-Sprinkled Cattle: Hides and Feces". en. In: *Journal of Food Chemistry and Nanotechnology* 03.02 (2017). DOI: 10.17756/jfcn.2017-037.

[236] Amit Lahiri et al. "TolA mediates the differential detergent resistance pattern between the Salmonella enterica subsp. enterica serovars Typhi and Typhimurium". eng. In: *Microbiology (Reading, England)* 157.Pt 5 (May 2011), pp. 1402–1415. DOI: 10.1099/mic.0.046565-0.

[237] Bożena Futoma-Kołoch et al. "Outer Membrane Proteins of Salmonella as Potential Markers of Resistance to Serum, Antibiotics and Biocides". In: *Current Medicinal Chemistry* 26.11 (Apr. 2019), pp. 1960–1978. DOI: 10.2174/0929867325666181031130851.

[238] Fabrice J. C Lacroix et al. "Salmonella typhimurium acrB-like gene: identification and role in resistance to biliary salts and detergents and in murine infection". en. In: *FEMS Microbiology Letters* 135.2 (Jan. 1996), pp. 161–167. DOI: 10.1016/0378-1097(95)00443-2.

[239] Leigh A. Knodler et al. "Salmonella type III effectors PipB and PipB2 are targeted to detergent-resistant microdomains on internal host cell membranes". en. In: *Molecular Microbiology* 49.3 (2003), pp. 685–704. DOI: 10.1046/j.1365-2958.2003.03598.x.

[240] M. Braoudaki and A. C. Hilton. "Adaptive Resistance to Biocides in Salmonella enterica and Escherichia coli O157 and Cross-Resistance to Antimicrobial Agents". In: *Journal of Clinical Microbiology* 42.1 (Jan. 2004), pp. 73–78. DOI: 10.1128/JCM.42.1.73-78.2004.

[241] Maria Luisa Fernández Márquez et al. "Biocide Tolerance and Antibiotic Resistance in Salmonella Isolates from Hen Eggshells". In: *Foodborne Pathogens and Disease* 14.2 (Feb. 2017). Publisher: Mary Ann Liebert, Inc., publishers, pp. 89–95. DOI: 10.1089/fpd.2016.2182.

[242] Rebekah N. Whitehead et al. "Exposure of Salmonella enterica Serovar Typhimurium to High Level Biocide Challenge Can Select Multidrug Resistant Mutants in a Single Step". en. In: *PLOS ONE* 6.7 (2011). Publisher: Public Library of Science, e22833. DOI: 10.1371/journal.pone.0022833.

[243] Alexander J. Westermann et al. "Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions". en. In: *Nature* 529.7587 (Jan. 2016). Number: 7587 Publisher: Nature Publishing Group, pp. 496–501. DOI: 10.1038/nature16547.

[244] B. J. Parcell et al. "Clinical perspectives in integrating whole-genome sequencing into the investigation of healthcare and public health outbreaks – hype or help?" en. In: *Journal of Hospital Infection* 109 (Mar. 2021), pp. 1–9. DOI: 10.1016/j.jhin.2020.11.001.

[245] George M. Ibrahim and Paul M. Morin. "Salmonella Serotyping Using Whole Genome Sequencing". In: *Frontiers in Microbiology* 9 (2018).

[246] Eric Brown et al. "Use of Whole-Genome Sequencing for Food Safety and Public Health in the United States". In: *Foodborne Pathogens and Disease* 16.7 (July 2019), pp. 441–450. DOI: 10.1089/fpd.2019.2662.

[247] Johanne Ahrenfeldt et al. "Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods". en. In: *BMC Genomics* 18.1 (Jan. 2017), p. 19. DOI: 10.1186/s12864-016-3407-6.

[248] Carol A. Gilchrist et al. "Whole-genome sequencing in outbreak analysis". eng. In: *Clinical Microbiology Reviews* 28.3 (July 2015), pp. 541–563. DOI: 10.1128/CMR.00075-13.

[249] Dominique S. Blanc et al. "Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa". In: *Frontiers in Microbiology* 11 (July 2020), p. 1729. DOI: 10.3389/fmicb.2020.01729.

[250] Alison Waldram et al. "Epidemiological analysis of Salmonella clusters identified by whole genome sequencing, England and Wales 2014". eng. In: *Food Microbiology* 71 (May 2018), pp. 39–45. DOI: 10.1016/j.fm.2017.02.012.

[251] T. Inns et al. "Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of Salmonella Enteritidis". eng. In: *Epidemiology and Infection* 145.2 (Jan. 2017), pp. 289–298. DOI: 10.1017/S0950268816001941.

[252] Michael Payne et al. "Enhancing genomics-based outbreak detection of endemic Salmonella enterica serovar Typhimurium using dynamic thresholds". eng. In: *Microbial Genomics* 7.6 (June 2021). DOI: 10.1099/mgen.0.000310.

[253] DE Salmon and T Smith. "The bacterium of swine plague". In: *Am Month Micr J* 7 (1886), p. 204.

[254] L Le Minor. "The genus Salmonella". In: *Springer* (1994), p. 2760.

[255] P. J. Fedorka-Cray, J. T. Gray, and C. Wray. "Salmonella infections in pigs." In: *Salmonella in domestic animals.* CABI Books (Jan. 2000), pp. 191–207. DOI: 10.1079/9780851992617.0191.

[256] Sabine Itié-Hafez (1) (sabine.itie@agriculture.gouv.fr), Alain Le Roux (2), Françoise Chartier (3), Daniel Fort (3), Corinne Danan (1) and ANSES. "Surveillance of Salmonella contamination of pig carcasses through self-inspections undertaken at the slaughterhouse". fr. In: *Bulletin épidémiologique, santé animale et alimentation* 77.65 (2015).

[257] S. L. Foley, A. M. Lynne, and R. Nayak. "Salmonella challenges: Prevalence in swine and poultry and potential pathogenicity of such isolates1,2". In: *Journal of Animal Science* 86.suppl_14 (Apr. 2008), E149–E162. DOI: 10.2527/jas.2007-0464.

[258] Vincent Leclerc , Frédérique Moury, Véronique Noel, Isabelle Berta-Vanrullen, Sabrina Cadel-Six, Renaud Lailler. *Le réseau Salmonella, un dispositif de surveillance des salmonelles sur la chaîne alimentaire : bilan 2015*. 2016.

[259] S. Bonardi. "Salmonella in the pork production chain and its impact on human health in the European Union". en. In: *Epidemiology & Infection* 145.8 (June 2017). Publisher: Cambridge University Press, pp. 1513–1526. DOI: 10.1017/S095026881700036X.

[260] Grammato Evangelopoulou et al. "Pork Meat as a Potential Source of Salmonella enterica subsp. arizonae Infection in Humans". In: *Journal of Clinical Microbiology* 52.3 (Mar. 2014), pp. 741–744. DOI: 10.1128/JCM.02933-13.

[261] F. Boyen et al. "Non-typhoidal Salmonella infections in pigs: a closer look at epidemiology, pathogenesis and control". eng. In: *Veterinary Microbiology* 130.1-2 (July 2008), pp. 1–19. DOI: 10.1016/j.vetmic.2007.12.017.

[262] *[SSA] Le réseau Salmonella, un dispositif de surveillance des salmonelles de la fourche à la fourchette : bilan des données de sérotypage 2019 | Bulletin épidémiologique*.

[263] Harold Noel et al. "Épidémie nationale d'infections à Salmonella enterica subspecies enterica sérotype 4,12: i:- liée à la consommation de saucisson sec". fr. In: (), p. 3.

[264] Claudia Lucarelli et al. "Evidence for a second genomic island conferring multidrug resistance in a clonal group of strains of Salmonella enterica serovar Typhimurium and its monophasic variant circulating in Italy, Denmark, and the United Kingdom". eng. In: *Journal of Clinical Microbiology* 48.6 (June 2010), pp. 2103–2109. DOI: 10.1128/JCM.01371-09.

[265] Anna Maria Dionisi et al. "Molecular characterization of multidrug-resistant strains of Salmonella enterica serotype Typhimurium and Monophasic variant (S. 4,[5],12:i:-) isolated from human infections in Italy". eng. In: *Foodborne Pathogens and Disease* 6.6 (Aug. 2009), pp. 711–717. DOI: 10.1089/fpd.2008.0240.

[266] Zenghai Jiang et al. "Prevalence and antimicrobial resistance of Salmonella recovered from pig-borne food products in Henan, China". en. In: *Food Control* 121 (Mar. 2021), p. 107535. DOI: 10.1016/j.foodcont.2020.107535.

[267] Annette Deane et al. "Prevalence of Salmonella spp. in slaughter pigs and carcasses in Irish abattoirs and their antimicrobial resistance". eng. In: *Irish Veterinary Journal* 75.1 (Mar. 2022), p. 4. DOI: 10.1186/s13620-022-00211-y.

[268] ANSES. *Mesures de maîtrise des salmonelles en filière porcine : état des connaissances et appréciation quantitative des risques*. 2018.

[269] Juber Herrera Uribe et al. "Transcriptional analysis of porcine intestinal mucosa infected with Salmonella Typhimurium revealed a massive inflammatory response and disruption of bile acid absorption in ileum". In: *Veterinary Research* 47.1 (Jan. 2016), p. 11. DOI: 10.1186/s13567-015-0286-9.

[270] John S. Gunn et al. "Salmonella chronic carriage: epidemiology, diagnosis, and gallbladder persistence". en. In: *Trends in Microbiology* 22.11 (Nov. 2014), pp. 648–655. DOI: 10.1016/j.tim.2014.06.007.

[271] J T Gray et al. "Natural transmission of Salmonella choleraesuis in swine". In: *Applied and Environmental Microbiology* 62.1 (Jan. 1996). Publisher: American Society for Microbiology, pp. 141–146. DOI: 10.1128/aem.62.1.141-146.1996.

[272] P. J. Fedorka-Cray et al. "Alternate routes of invasion may affect pathogenesis of Salmonella typhimurium in swine". eng. In: *Infection and Immunity* 63.7 (July 1995), pp. 2658–2664. DOI: 10.1128/iai.63.7.2658-2664.1995.

[273] A Letellier et al. "Host response to various treatments to reduce Salmonella infections in swine." In: *Canadian Journal of Veterinary Research* 65.3 (July 2001), pp. 168–172.

[274] Bradley L Bearson et al. "Salmonella DIVA vaccine reduces disease, colonization and shedding due to virulent S. Typhimurium infection in swine". In: *Journal of Medical Microbiology* 66.5 (May 2017), pp. 651–661. DOI: 10.1099/jmm.0.000482.

[275] Thomas N. Denagamage et al. "Efficacy of Vaccination to Reduce Salmonella Prevalence in Live and Slaughtered Swine: A Systematic Review of Literature from 1979 to 2007". In: *Foodborne Pathogens and Disease* 4.4 (Dec. 2007). Publisher: Mary Ann Liebert, Inc., publishers, pp. 539–549. DOI: 10.1089/fpd.2007.0013.

[276] APHA. "Salmonella in Livestock Production 2020". en. In: (), p. 254.

[277] Pedro Henrique N. Panzenhagen et al. "Genetically distinct lineages of Salmonella Typhimurium ST313 and ST19 are present in Brazil". eng. In: *International journal of medical microbiology: IJMM* 308.2 (Mar. 2018), pp. 306–316. DOI: 10.1016/j.ijmm.2018.01.005.

[278] Li Bai et al. "Prevalence of Salmonella Isolates from Chicken and Pig Slaughterhouses and Emergence of Ciprofloxacin and Cefotaxime Co-Resistant S. enterica Serovar Indiana in Henan, China". en. In: *PLOS ONE* 10.12 (2015). Publisher: Public Library of Science, e0144532. DOI: 10.1371/journal.pone.0144532.

[279] Yezhi Fu et al. "Evidence for common ancestry and microevolution of passerine-adapted Salmonella enterica serovar Typhimurium in the UK and USA". In: *Microbial Genomics* 8.2 (Feb. 2022), p. 000775. DOI: 10.1099/mgen.0.000775.

[280] Chinyere K. Okoro et al. "Intra-continental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa". In: *Nature genetics* 44.11 (Nov. 2012), pp. 1215–1221. DOI: 10.1038/ng.2423.

[281] Reza Ranjbar, Parisa Elhaghi, and Leili Shokoohizadeh. "Multilocus Sequence Typing of the Clinical Isolates of Salmonella Enterica Serovar Typhimurium in Tehran Hospitals". In: *Iranian Journal of Medical Sciences* 42.5 (Sept. 2017), pp. 443–448.

[282] Heather R. Bonifield and Kelly T. Hughes. "Flagellar phase variation in Salmonella enterica is mediated by a posttranscriptional control mechanism". eng. In: *Journal of Bacteriology* 185.12 (June 2003), pp. 3567–3574. DOI: 10.1128/JB.185.12.3567-3574.2003.

[283] M. Aurora Echeita et al. "Emergence and Spread of an Atypical Salmonella enterica subsp. enterica Serotype 4,5,12:i:- Strain in Spain". EN. In: *Journal of Clinical Microbiology* (Oct. 1999). DOI: 10.1128/JCM.37.10.3425-3425.1999.

[284] Lorena Laorden et al. "Genetic Evolution of the Spanish Multidrug-Resistant Salmonella enterica 4,5,12:i:- Monophasic Variant". In: *Journal of Clinical Microbiology* 48.12 (Dec. 2010), pp. 4563–4566. DOI: 10.1128/JCM.00337-10.

[285] Honghu Sun et al. "The Epidemiology of Monophasic Salmonella Typhimurium". In: *Foodborne Pathogens and Disease* 17.2 (Feb. 2020). Publisher: Mary Ann Liebert, Inc., publishers, pp. 87–97. DOI: 10.1089/fpd.2019.2676.

[286] Sabrina Cadel-Six et al. "The Spatiotemporal Dynamics and Microevolution Events That Favored the Success of the Highly Clonal Multidrug-Resistant Monophasic Salmonella Typhimurium Circulating in Europe". English. In: *Frontiers in Microbiology* 12 (2021). Publisher: Frontiers. DOI: 10.3389/fmicb.2021.651124.

[287] Eleonora Mastrorilli et al. "A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic Salmonella Serovar 4,[5],12:i:-". English. In: *Frontiers in Microbiology* 9 (2018). DOI: 10.3389/fmicb.2018.00715.

[288] K. L. Hopkins et al. "Multiresistant Salmonella enterica serovar 4,[5],12:i:- in Europe: a new pandemic strain?" eng. In: *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 15.22 (June 2010), p. 19580.

[289] M. A. Echeita, S. Herrera, and M. A. Usera. "Atypical, fljB-negative Salmonella enterica subsp. enterica strain of serovar 4,5,12:i:- appears to be a monophasic variant of serovar Typhimurium". eng. In: *Journal of Clinical Microbiology* 39.8 (Aug. 2001), pp. 2981–2983. DOI: 10.1128/JCM.39.8.2981-2983.2001.

[290] Cui Li et al. "Antimicrobial Resistance and CRISPR Typing Among Salmonella Isolates From Poultry Farms in China". In: *Frontiers in Microbiology* 12 (Sept. 2021), p. 730046. DOI: 10.3389/fmicb.2021.730046.

[291] Xuchu Wang et al. "Antibiotic Resistance in Salmonella Typhimurium Isolates Recovered From the Food Chain Through National Antimicrobial Resistance Monitoring System Between 1996 and 2016". In: *Frontiers in Microbiology* 10 (May 2019), p. 985. DOI: 10.3389/fmicb.2019.00985.

[292] Eleonora Tassinari et al. "Microevolution of antimicrobial resistance and biofilm formation of Salmonella Typhimurium during persistence on pig farms". en. In: *Scientific Reports* 9.1 (June 2019). Number: 1 Publisher: Nature Publishing Group, p. 8832. DOI: 10.1038/s41598-019-45216-w.

[293] Monika Dolejska et al. "Complete sequences of IncHI1 plasmids carrying blaCTX-M-1 and qnrS1 in equine Escherichia coli provide new insights into plasmid evolution". In: *Journal of Antimicrobial Chemotherapy* 69.9 (Sept. 2014), pp. 2388–2393. DOI: 10.1093/jac/dku172.

[294] C Poppe et al. "Salmonella typhimurium DT104: a virulent and drug-resistant pathogen." In: *The Canadian Veterinary Journal* 39.9 (Sept. 1998), pp. 559–565.

[295] Pimlapas Leekitcharoenphon et al. "Global Genomic Epidemiology of Salmonella enterica Serovar Typhimurium DT104". In: *Applied and Environmental Microbiology* 82.8 (Apr. 2016). Publisher: American Society for Microbiology, pp. 2516–2526. DOI: 10.1128/AEM.03821-15.

[296] Sophie Schbath et al. "Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis". In: *Journal of Computational Biology* 19.6 (June 2012), pp. 796–813. DOI: 10.1089/cmb.2012.0022.

[297] M. McClelland et al. "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2". eng. In: *Nature* 413.6858 (Oct. 2001), pp. 852–856. DOI: 10.1038/35101614.

[298] K E Sanderson, A Hessel, and K E Rudd. "Genetic map of Salmonella typhimurium, edition VIII." In: *Microbiological Reviews* 59.2 (June 1995), pp. 241–303.

[299] Fanta D. Gutema et al. "Prevalence and Serotype Diversity of Salmonella in Apparently Healthy Cattle: Systematic Review and Meta-Analysis of Published Studies, 2000–2017". In: *Frontiers in Veterinary Science* 6 (2019).

[300] Laetitia Bonifait et al. "Occurrence of Salmonella in the Cattle Production in France". en. In: *Microorganisms* 9.4 (Apr. 2021). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 872. DOI: 10.3390/microorganisms9040872.

[301] Aymeric Ung et al. "Disentangling a complex nationwide Salmonella Dublin outbreak associated with raw-milk cheese consumption, France, 2015 to 2016". In: *Eurosurveillance* 24.3 (Jan. 2019), p. 1700703. DOI: 10.2807/1560-7917.ES.2019.24.3.1700703.

[302] Morgane Dominguez et al. "Outbreak of Salmonella enterica serotype Montevideo infections in France linked to consumption of cheese made from raw milk". eng. In: *Foodborne Pathogens and Disease* 6.1 (Feb. 2009), pp. 121–128. DOI: 10.1089/fpd.2008.0086.

[303] N. Adams et al. "Gastrointestinal infections caused by consumption of raw drinking milk in England & Wales, 1992–2017". In: *Epidemiology and Infection* 147 (Sept. 2019), e281. DOI: 10.1017/S095026881900164X.

[304] T. R. Callaway et al. "Fecal Prevalence and Diversity of Salmonella Species in Lactating Dairy Cattle in Four States*". en. In: *Journal of Dairy Science* 88.10 (Oct. 2005), pp. 3603–3608. DOI: 10.3168/jds.S0022-0302(05)73045-9.

[305] Heidi L. Pecoraro, Belinda Thompson, and Gerald E. Duhamel. "Histopathology case definition of naturally acquired Salmonella enterica serovar Dublin infection in young Holstein cattle in the northeastern United States". eng. In: *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc* 29.6 (Nov. 2017), pp. 860–864. DOI: 10.1177/1040638717712757.

[306] John F. Mee, Paulina Jawor, and Tadeusz Stefaniak. "Role of Infection and Immunity in Bovine Perinatal Mortality: Part 1. Causes and Current Diagnostic Approaches". en. In: *Animals* 11.4 (Apr. 2021), p. 1033. DOI: 10.3390/ani11041033.

[307] *Rebhun's Diseases of Dairy Cattle - 3rd Edition*.

[308] "Diseases of the Alimentary Tract". In: *Veterinary Medicine* (2017), pp. 175–435. DOI: 10.1016/B978-0-7020-5246-0.00007-3.

[309] L. R. Nielsen et al. "Salmonella Dublin infection in dairy cattle: risk factors for becoming a carrier". eng. In: *Preventive Veterinary Medicine* 65.1-2 (Aug. 2004), pp. 47–62. DOI: 10.1016/j.prevetmed.2004.06.010.

[310] Liza Rosenbaum Nielsen, Hans Houe, and Søren Saxmose Nielsen. "Narrative Review Comparing Principles and Instruments Used in Three Active Surveillance and Control Programmes for Non-EU-regulated Diseases in the Danish Cattle Population". In: *Frontiers in Veterinary Science* 8 (2021).

[311] "Salmonella in Livestock Production 2020". en. In: (), p. 255.

[312] William D. Nettleton. "Protracted, Intermittent Outbreak of Salmonella Mbandaka Linked to a Restaurant — Michigan, 2008–2019". en-us. In: *MMWR. Morbidity and Mortality Weekly Report* 70 (2021). DOI: 10.15585/mmwr.mm7033a1.

[313] Andrzej Hoszowski et al. "Fifteen years of successful spread of Salmonella enterica serovar Mbandaka clone ST413 in Poland and its public health consequences". en. In: *Annals of Agricultural and Environmental Medicine* 23.2 (2016), p. 5.

[314] Cassia Lindsay et al. "Retrospective use of whole genome sequencing to better understand an outbreak of Salmonella enterica serovar Mbandaka in New South Wales, Australia". In: *Western Pacific Surveillance and Response Journal : WPSAR* 9.2 (Apr. 2018), pp. 20–25. DOI: 10.5365/wpsar.2017.8.4.008.

[315] Matthew R. Hayward, Vincent AA Jansen, and Martin J. Woodward. "Comparative genomics of Salmonella entericaserovars Derby and Mbandaka, two prevalent serovars associated with different livestock species in the UK". In: *BMC Genomics* 14.1 (May 2013), p. 365. DOI: 10.1186/1471-2164-14-365.

[316] Ruth E. Timme et al. "Phylogenetic Diversity of the Enteric Pathogen Salmonella enterica subsp. enterica Inferred from Genome-Wide Reference-Free SNP Characters". In: *Genome Biology and Evolution* 5.11 (2013), pp. 2109–2123. DOI: 10.1093/gbe/evt159.

[317] Linto Antony. "Phylogenetic and Associated Phenotypic Analysis of Salmonella Enterica Serovar Mbandaka". en. In: (), p. 218.

[318] Carlos Valiente-Mullor et al. "One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads". en. In: *PLOS Computational Biology* 17.1 (Jan. 2021). Publisher: Public Library of Science, e1008678. DOI: 10.1371/journal.pcbi.1008678.

[319] *Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to Salmonella enterica serotype Typhimurium and serotype 1,4,[5],12:i:-.*

[320] Na Lyu et al. "Genomic Characterization of Salmonella enterica Isolates From Retail Meat in Beijing, China". In: *Frontiers in Microbiology* 12 (2021).

[321] Duccio Medini et al. "The microbial pan-genome". en. In: *Current Opinion in Genetics & Development*. Genomes and evolution 15.6 (Dec. 2005), pp. 589–594. DOI: `10.1016/j.gde.2005.09.006`.

[322] Nicholas Delihas. "Impact of Small Repeat Sequences on Bacterial Genome Evolution". In: *Genome Biology and Evolution* 3 (July 2011), pp. 959–973. DOI: `10.1093/gbe/evr077`.

[323] Jillian Rumore et al. "Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective". In: *BMC Genomics* 19.1 (Dec. 2018), p. 870. DOI: `10.1186/s12864-018-5243-3`.

[324] Melissa J. Whaley et al. "Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis". en. In: *Scientific Reports* 8.1 (Oct. 2018). Number: 1 Publisher: Nature Publishing Group, p. 15803. DOI: `10.1038/s41598-018-33622-5`.

[325] Brian Bushnell. "BBMap: A Fast, Accurate, Splice-Aware Aligner". In: (Mar. 2014).

[326] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". eng. In: *Bioinformatics (Oxford, England)* 30.15 (Aug. 2014), pp. 2114–2120. DOI: `10.1093/bioinformatics/btu170`.

[327] Stephen F. Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. DOI: `10.1016/S0022-2836(05)80360-2`.

[328] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows-Wheeler transform". eng. In: *Bioinformatics (Oxford, England)* 26.5 (Mar. 2010), pp. 589–595. DOI: `10.1093/bioinformatics/btp698`.

[329] Seemann Torsten. "Snippy: fast bacterial variant calling from NGS reads". In: (2015).

[330] Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt von Haeseler. "Ultrafast Approximation for Phylogenetic Bootstrap". In: *Molecular Biology and Evolution* 30.5 (May 2013), pp. 1188–1195. DOI: `10.1093/molbev/mst024`.

[331] Subha Kalyaanamoorthy et al. "ModelFinder: fast model selection for accurate phylogenetic estimates". en. In: *Nature Methods* 14.6 (June 2017). Number: 6 Publisher: Nature Publishing Group, pp. 587–589. DOI: `10.1038/nmeth.4285`.

[332] S. Mirarab et al. "ASTRAL: genome-scale coalescent-based species tree estimation". en. In: *Bioinformatics* 30.17 (Sept. 2014), pp. i541–i548. DOI: `10.1093/bioinformatics/btu462`.

[333] Manuel Binet et al. "Fast and accurate branch lengths estimation for phylogenomic trees". In: *BMC Bioinformatics* 17.1 (Jan. 2016), p. 23. DOI: `10.1186/s12859-015-0821-8`.

[334] Alexey Gurevich et al. "QUAST: quality assessment tool for genome assemblies". eng. In: *Bioinformatics (Oxford, England)* 29.8 (Apr. 2013), pp. 1072–1075. DOI: `10.1093/bioinformatics/btt086`.

[335] Shaokang Zhang et al. "Salmonella Serotype Determination Utilizing High-Throughput Genome Sequencing Data". In: *Journal of Clinical Microbiology* 53.5 (May 2015), pp. 1685–1692. DOI: `10.1128/JCM.00323-15`.

[336] EFSA. *Scientific Opinion on monitoring and assessment of the public health risk of "Salmonella Typhimurium-like" strains | EFSA*. en. Section: Scientific outputs.

[337] Sharon M. Tennant et al. "Identification by PCR of non-typhoidal Salmonella enterica serovars associated with invasive infections among febrile patients in Mali". eng. In: *PLoS neglected tropical diseases* 4.3 (Mar. 2010), e621. DOI: `10.1371/journal.pntd.0000621`.

[338] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". eng. In: *Bioinformatics (Oxford, England)* 25.16 (Aug. 2009), pp. 2078–2079. DOI: `10.1093/bioinformatics/btp352`.

[339] Claude Shannon and Warren Weaver. “The Mathematical Theory of Communication”. en. In: (), p. 132.

[340] Xiaofeng Dong et al. “Variation around the dominant viral genome sequence contributes to viral load and outcome in patients with Ebola virus disease”. In: *Genome Biology* 21.1 (Sept. 2020), p. 238. DOI: `10.1186/s13059-020-02148-3`.

[341] *ANDES: Statistical tools for the ANalyses of DEep Sequencing | BMC Research Notes | Full Text.*

[342] Iva Pritišanac et al. “Entropy and Information within Intrinsically Disordered Protein Regions”. en. In: *Entropy* 21.7 (July 2019). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 662. DOI: `10.3390/e21070662`.

[343] Sajia Akhter et al. “Applying Shannon’s information theory to bacterial and phage genomes and metagenomes”. en. In: *Scientific Reports* 3.1 (Jan. 2013). Number: 1 Publisher: Nature Publishing Group, p. 1033. DOI: `10.1038/srep01033`.

[344] Kelvin Li et al. “Analyses of the Microbial Diversity across the Human Microbiome”. In: *PLoS ONE* 7.6 (June 2012), e32118. DOI: `10.1371/journal.pone.0032118`.

[345] Arnaud Felten et al. “First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of Salmonella serovars to mammalian- and avian-hosts”. In: *BMC Microbiology* 17.1 (Nov. 2017), p. 222. DOI: `10.1186/s12866-017-1132-1`.

[346] Aaron McKenna et al. “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome Research* 20.9 (Sept. 2010), pp. 1297–1303. DOI: `10.1101/gr.107524.110`.

[347] Geraldine A. Van der Auwera and Brian D. O’Connor. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. en. Google-Books-ID: wwiCswEACAAJ. O’Reilly Media, Incorporated, 2020.

[348] Ryan Poplin et al. *Scaling accurate genetic variant discovery to tens of thousands of samples*. en. Tech. rep. Type: article. July 2018, p. 201178.

[349] Arndt von Haeseler. “Do we still need supertrees?” In: *BMC Biology* 10.1 (Feb. 2012), p. 13. DOI: `10.1186/1741-7007-10-13`.

[350] Tandy Warnow. “Supertree Construction: Opportunities and Challenges”. In: *arXiv:1805.03530 [q-bio]* (May 2018). arXiv: 1805.03530.

[351] Zhenxiang Xi, Liang Liu, and Charles C. Davis. “The Impact of Missing Data on Species Tree Estimation”. en. In: *Molecular Biology and Evolution* 33.3 (Mar. 2016), pp. 838–860. DOI: `10.1093/molbev/msv266`.

[352] Daniel A. Janies et al. “A comparison of supermatrix and supertree methods for multilocus phylogenetics using organismal datasets”. en. In: *Cladistics* 29.5 (2013), pp. 560–566. DOI: `10.1111/cla.12014`.

[353] J. Gordon Burleigh, Amy C. Driskell, and Michael J. Sanderson. “Supertree Bootstrapping Methods for Assessing Phylogenetic Variation among Genes in Genome-Scale Data Sets”. en. In: *Systematic Biology* 55.3 (June 2006), pp. 426–440. DOI: `10.1080/10635150500541722`.

[354] Rebecca T. Kimball et al. “A Phylogenomic Supertree of Birds”. en. In: *Diversity* 11.7 (July 2019). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 109. DOI: `10.3390/d11070109`.

[355] Fengrong Ren, Hiroshi Tanaka, and Ziheng Yang. “A likelihood look at the supermatrix–supertree controversy”. In: *Gene*. Phylogenomics and its Future: Devoted to Masami Hasegawa 441.1 (July 2009), pp. 119–125. DOI: `10.1016/j.gene.2008.04.002`.

[356] Pranjal Vachaspati and Tandy Warnow. “ASTRID: Accurate Species TRees from Internode Distances”. In: *BMC Genomics* 16.10 (Oct. 2015), S3. DOI: `10.1186/1471-2164-16-S10-S3`.

[357] Chao Zhang et al. “ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees”. In: *BMC Bioinformatics* 19.6 (May 2018), p. 153. DOI: 10.1186/s12859-018-2129-y.

[358] Zhemin Zhou et al. *Millennia of genomic stability within the invasive Para C Lineage of Salmonella enterica*. en. Tech. rep. Section: New Results Type: article. bioRxiv, Feb. 2017, p. 105759. DOI: 10.1101/105759.

[359] D. F. Robinson and L. R. Foulds. “Comparison of phylogenetic trees”. en. In: *Mathematical Biosciences* 53.1 (Feb. 1981), pp. 131–147. DOI: 10.1016/0025-5564(81)90043-2.

[360] “The European Union One Health 2019 Zoonoses Report”. en. In: *EFSA Journal* 19.2 (2021), e06406. DOI: 10.2903/j.efsa.2021.6406.

[361] Tetsuya Hayashi et al. “Complete Genome Sequence of Enterohemorrhagic Eschelichia coli O157:H7 and Genomic Comparison with a Laboratory Strain K-12”. In: *DNA Research* 8.1 (Jan. 2001), pp. 11–22. DOI: 10.1093/dnares/8.1.11.

[362] Neil MacAlasdair et al. “The effect of recombination on the evolution of a population of Neisseria meningitidis”. en. In: *Genome Research* (June 2021). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. DOI: 10.1101/gr.264465.120.

[363] *How clonal are Neisseria species? The epidemic clonality model revisited | PNAS*.

[364] Shea N Gardner, Tom Slezak, and Barry G. Hall. “kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome”. In: *Bioinformatics* 31.17 (Sept. 2015), pp. 2877–2878. DOI: 10.1093/bioinformatics/btv271.

[365] *The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes | Genome Biology | Full Text*.

[366] Ari Löytynoja. “Phylogeny-aware alignment with PRANK”. eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 1079 (2014), pp. 155–170. DOI: 10.1007/978-1-62703-646-7_10.

[367] Derek W. Barnett et al. “BamTools: a C++ API and toolkit for analyzing and managing BAM files”. In: *Bioinformatics* 27.12 (June 2011), pp. 1691–1692. DOI: 10.1093/bioinformatics/btr174.

[368] Sarah Sandmann et al. “Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data”. en. In: *Scientific Reports* 7.1 (Feb. 2017). Number: 1 Publisher: Nature Publishing Group, p. 43169. DOI: 10.1038/srep43169.

[369] ANSES. *OPINION of the French Agency for Food, Environmental and Occupational Health & Safety on Salmonella control measures in the pig sector: review of knowledge and quantitative risk assessment*. 2018.

[370] Alessia De Lucia et al. “Role of wild birds and environmental contamination in the epidemiology of Salmonella infection in an outdoor pig farm”. en. In: *Veterinary Microbiology* 227 (Dec. 2018), pp. 148–154. DOI: 10.1016/j.vetmic.2018.11.003.

[371] Clare McW. H. Benskin et al. “Bacterial pathogens in wild birds: a review of the frequency and effects of infection”. en. In: *Biological Reviews* 84.3 (2009), pp. 349–373. DOI: 10.1111/j.1469-185X.2008.00076.x.

[372] Emmanuel Paradis and Klaus Schliep. “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. In: *Bioinformatics* 35.3 (Feb. 2019), pp. 526–528. DOI: 10.1093/bioinformatics/bty633.

[373] Aaron E. Darling, Bob Mau, and Nicole T. Perna. “progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement”. In: *PLoS ONE* 5.6 (June 2010). DOI: 10.1371/journal.pone.0011147.

[374] Bao Ton-Hoang, Catherine Turlan, and Michael Chandler. “Functional domains of the IS1 transposase: analysis in vivo and in vitro”. eng. In: *Molecular Microbiology* 53.5 (Sept. 2004), pp. 1529–1543. DOI: 10.1111/j.1365-2958.2004.04223.x.

[375] M Biserćić and H Ochman. "The ancestry of insertion sequences common to Escherichia coli and Salmonella typhimurium". In: *Journal of Bacteriology* 175.24 (Dec. 1993). Publisher: American Society for Microbiology, pp. 7863–7868. DOI: 10.1128/jb.175.24.7863-7868.1993.

[376] *Isolation and plasmid characterization of carbapenemase (IMP-4) producing Salmonella enterica Typhimurium from cats | Scientific Reports.*

[377] Qiu-E. Yang et al. "IncF plasmid diversity in multi-drug resistant Escherichia coli strains from animals in China". In: *Frontiers in Microbiology* 6 (Sept. 2015), p. 964. DOI: 10.3389/fmicb.2015.00964.

[378] David Arndt et al. "PHASTER: a better, faster version of the PHAST phage search tool". In: *Nucleic Acids Research* 44.Web Server issue (July 2016), W16–W21. DOI: 10.1093/nar/gkw387.

[379] Christoph Schoen et al. "Genome flexibility in Neisseria meningitidis". In: *Vaccine* 27.Suppl 2 (June 2009), B103–B111. DOI: 10.1016/j.vaccine.2009.04.064.

[380] Evan S. Snitkin et al. "Genome-wide recombination drives diversification of epidemic strains of *Acinetobacter baumannii*". en. In: *Proceedings of the National Academy of Sciences* 108.33 (Aug. 2011), pp. 13758–13763. DOI: 10.1073/pnas.1104404108.

[381] Christine Jandrasits et al. "seq-seq-pan: building a computational pan-genome data structure on whole genome alignment". In: *BMC Genomics* 19.1 (Jan. 2018), p. 47. DOI: 10.1186/s12864-017-4401-3.

[382] Christine Jandrasits et al. "Computational pan-genome mapping and pairwise SNP-distance improve detection of Mycobacterium tuberculosis transmission clusters". en. In: *PLOS Computational Biology* 15.12 (2019), e1007527. DOI: 10.1371/journal.pcbi.1007527.

[383] Heng Li, Xiaowen Feng, and Chong Chu. "The design and construction of reference pangenome graphs with minigraph". In: *Genome Biology* 21.1 (Oct. 2020), p. 265. DOI: 10.1186/s13059-020-02168-z.

[384] Jouni Sirén et al. "Haplotype-aware graph indexes". In: *Bioinformatics* 36.2 (Jan. 2020), pp. 400–407. DOI: 10.1093/bioinformatics/btz575.

[385] Guillaume Gautreau et al. "PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph". en. In: *PLOS Computational Biology* 16.3 (Mar. 2020). Publisher: Public Library of Science, e1007732. DOI: 10.1371/journal.pcbi.1007732.

[386] Glenn Hickey et al. "Genotyping structural variants in pangenome graphs using the vg toolkit". In: *Genome Biology* 21.1 (Feb. 2020), p. 35. DOI: 10.1186/s13059-020-1941-7.

[387] Hongbo Li et al. "Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber". en. In: *Nature Communications* 13.1 (Feb. 2022). Number: 1 Publisher: Nature Publishing Group, p. 682. DOI: 10.1038/s41467-022-28362-0.

[388] Erik Garrison and Andrea Guarracino. *Unbiased pangenome graphs*. en. Tech. rep. Section: New Results Type: article. bioRxiv, Feb. 2022, p. 2022.02.14.480413. DOI: 10.1101/2022.02.14.480413.

[389] *pggb*. original-date: 2020-09-12T16:07:58Z. May 2022.

[390] Martin C. Frith and Risa Kawaguchi. "Split-alignment of genomes finds orthologies more accurately". In: *Genome Biology* 16.1 (May 2015), p. 106. DOI: 10.1186/s13059-015-0670-9.

[391] Yi-Pin Lai and Thomas R. Ioerger. "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility". In: *BMC Bioinformatics* 19.1 (Nov. 2018), p. 450. DOI: 10.1186/s12859-018-2456-z.

[392] Timothy M Beissinger et al. "Defining window-boundaries for genomic analyses using smoothing spline techniques". In: *Genetics, Selection, Evolution : GSE* 47.1 (Apr. 2015). DOI: 10.1186/s12711-015-0105-9.

[393] Mario P. L. Calus et al. “Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection”. In: *Genetics Selection Evolution* 48.1 (June 2016), p. 49. DOI: 10.1186/s12711-016-0225-x.

[394] Xavier Didelot et al. “Recombination and Population Structure in Salmonella enterica”. en. In: *PLOS Genetics* 7.7 (2011), e1002191. DOI: 10.1371/journal.pgen.1002191.

[395] James Hadfield et al. “Nextstrain: real-time tracking of pathogen evolution”. In: *Bioinformatics* 34.23 (Dec. 2018), pp. 4121–4123. DOI: 10.1093/bioinformatics/bty407.

[396] D. H. Huson. “SplitsTree: analyzing and visualizing evolutionary data.” en. In: *Bioinformatics* 14.1 (Jan. 1998), pp. 68–73. DOI: 10.1093/bioinformatics/14.1.68.

[397] Cheryl P. Andam et al. “Contributions of ancestral inter-species recombination to the genetic diversity of extant Streptomyces lineages”. en. In: *The ISME Journal* 10.7 (July 2016). Number: 7 Publisher: Nature Publishing Group, pp. 1731–1741. DOI: 10.1038/ismej.2015.230.

[398] Caroline M. Stott and Louis-Marie Bobay. “Impact of homologous recombination on core genome phylogenies”. In: *BMC Genomics* 21.1 (Nov. 2020), p. 829. DOI: 10.1186/s12864-020-07262-x.

[399] Esther R Robinson, Timothy M Walker, and Mark J Pallen. “Genomics and outbreak investigation: from sequence to consequence”. In: *Genome Medicine* 5.4 (Apr. 2013), p. 36. DOI: 10.1186/gm440.

[400] Sophie Octavia et al. “Genomic heterogeneity of Salmonella enterica serovar Typhimurium bacteriuria from chronic infection”. eng. In: *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 51 (July 2017), pp. 17–20. DOI: 10.1016/j.meegid.2017.03.004.

[401] Madeleine De Sousa Violante et al. “A retrospective and regional approach assessing the genomic diversity of Salmonella Dublin”. In: *NAR Genomics and Bioinformatics* 4.3 (Sept. 2022), lqac047. DOI: 10.1093/nargab/lqac047.

[402] Bérangère Lecuyer. “Des contrastes régionaux entre l’élevage et l’abattage”. In: *Réussir porc - Tech porc* (Aug. 2019).

[403] James Robertson et al. “Completed Genome Sequences of Strains from 36 Serotypes of Salmonella”. In: *Genome Announcements* 6.3 (Jan. 2018), e01472–17. DOI: 10.1128/genomeA.01472-17.

[404] V. Caliendo et al. “Transatlantic spread of highly pathogenic avian influenza H5N1 by wild birds from Europe to North America in 2021”. en. In: *Scientific Reports* 12.1 (July 2022). Number: 1 Publisher: Nature Publishing Group, p. 11729. DOI: 10.1038/s41598-022-13447-z.

[405] J. C. Gower. “A General Coefficient of Similarity and Some of Its Properties”. In: *Biometrics* 27.4 (1971), pp. 857–871. DOI: 10.2307/2528823.

[406] Sen Xu et al. “A Male-Specific Genetic Map of the Microcrustacean Daphnia pulex Based on Single-Sperm Whole-Genome Sequencing”. In: *Genetics* 201.1 (Sept. 2015), pp. 31–38. DOI: 10.1534/genetics.115.179028.

[407] Emanuele Bosi et al. “MeDuSa: a multi-draft based scaffolder”. eng. In: *Bioinformatics (Oxford, England)* 31.15 (Aug. 2015), pp. 2443–2451. DOI: 10.1093/bioinformatics/btv171.

[408] Brian D. Ondov et al. “Mash: fast genome and metagenome distance estimation using MinHash”. In: *Genome Biology* 17.1 (June 2016), p. 132. DOI: 10.1186/s13059-016-0997-x.

[409] Shunichi Kosugi, Hideki Hirakawa, and Satoshi Tabata. “GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments”. eng. In: *Bioinformatics (Oxford, England)* 31.23 (Dec. 2015), pp. 3733–3741. DOI: 10.1093/bioinformatics/btv465.

[410] Peter J. A. Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

[411] Shaokang Zhang et al. "SeqSero2: Rapid and Improved Salmonella Serotype Determination Using Whole-Genome Sequencing Data". eng. In: *Applied and Environmental Microbiology* 85.23 (Dec. 2019), e01746–19. DOI: 10.1128/AEM.01746-19.

[412] Lihong Chen et al. "VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D694–697. DOI: 10.1093/nar/gkv1239.

[413] Enrique Doster et al. "MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data". In: *Nucleic Acids Research* 48.D1 (Nov. 2019), pp. D561–D569. DOI: 10.1093/nar/gkz1010.

[414] Ea Zankari et al. "Identification of acquired antimicrobial resistance genes". eng. In: *The Journal of Antimicrobial Chemotherapy* 67.11 (Nov. 2012), pp. 2640–2644. DOI: 10.1093/jac/dks261.

[415] Louise Roer et al. "Is the Evolution of Salmonella enterica subsp. enterica Linked to Restriction-Modification Systems?" In: *mSystems* 1.3 (June 2016). Publisher: American Society for Microbiology, e00009–16. DOI: 10.1128/mSystems.00009-16.

[416] James Robertson and John H. E. Nash. "MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies". eng. In: *Microbial Genomics* 4.8 (Aug. 2018). DOI: 10.1099/mgen.0.000206.

[417] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* 30.14 (July 2014), pp. 2068–2069. DOI: 10.1093/bioinformatics/btu153.

[418] Kimmi N. Schrader et al. "Evaluation of Commercial Antisera for Salmonella Serotyping". In: *Journal of Clinical Microbiology* 46.2 (Feb. 2008), pp. 685–688. DOI: 10.1128/JCM.01808-07.

[419] Cécile Boland et al. "Extensive genetic variability linked to IS26 insertions in the fljB promoter region of atypical monophasic variants of Salmonella enterica serovar Typhimurium". eng. In: *Applied and Environmental Microbiology* 81.9 (May 2015), pp. 3169–3175. DOI: 10.1128/AEM.00270-15.

[420] R. H. Davies and M. Breslin. "Investigation of Salmonella contamination and disinfection in farm egg-packing plants". eng. In: *Journal of Applied Microbiology* 94.2 (2003), pp. 191–196. DOI: 10.1046/j.1365-2672.2003.01817.x.

[421] Lisa Barco et al. "Ascertaining the relationship between Salmonella Typhimurium and Salmonella 4,[5],12:i:- by MLVA and inferring the sources of human salmonellosis due to the two serovars in Italy". In: *Frontiers in Microbiology* 6 (Apr. 2015), p. 301. DOI: 10.3389/fmicb.2015.00301.

[422] Pernille Gymoese et al. "Investigation of Outbreaks of Salmonella enterica Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark". In: *Emerging Infectious Diseases* 23.10 (Oct. 2017), pp. 1631–1639. DOI: 10.3201/eid2310.161248.

[423] Michael Hensel et al. "Functional analysis of ssaJ and the ssaK/U operon, 13 genes encoding components of the type III secretion apparatus of Salmonella Pathogenicity Island 2". en. In: *Molecular Microbiology* 24.1 (1997), pp. 155–167. DOI: 10.1046/j.1365-2958.1997.3271699.x.

[424] Caleb W. Dorsey et al. "Salmonella enterica serotype Typhimurium MisL is an intestinal colonization factor that binds fibronectin". eng. In: *Molecular Microbiology* 57.1 (July 2005), pp. 196–211. DOI: 10.1111/j.1365-2958.2005.04666.x.

[425] Nguyen Thi Khanh Nhu et al. "Discovery of New Genes Involved in Curli Production by a Uropathogenic Escherichia coli Strain from the Highly Virulent O45:K1:H7 Lineage". In: *mBio* 9.4 (Aug. 2018). Publisher: American Society for Microbiology, e01462–18. DOI: 10.1128/mBio.01462-18.

[426] Carrie Althouse et al. "Type 1 Fimbriae of Salmonella enterica Serovar Typhimurium Bind to Enterocytes and Contribute to Colonization of Swine In Vivo". In: *Infection and Immunity* 71.11 (Nov. 2003), pp. 6446–6452. DOI: 10.1128/IAI.71.11.6446-6452.2003.

[427] Eric H. Weening et al. "The Salmonella enterica Serotype Typhimurium lpf, bcf, stb, stc, std, and sth Fimbrial Operons Are Required for Intestinal Persistence in Mice". en. In: *Infection and Immunity* 73.6 (June 2005). Publisher: American Society for Microbiology (ASM), p. 3358. DOI: 10.1128/IAI.73.6.3358-3366.2005.

[428] Sushim K. Gupta et al. "Genomic comparison of diverse Salmonella serovars isolated from swine". en. In: *PLOS ONE* 14.11 (Nov. 2019). Publisher: Public Library of Science, e0224518. DOI: 10.1371/journal.pone.0224518.

[429] Nathan A. Ledeboer et al. "Salmonella enterica serovar Typhimurium requires the Lpf, Pef, and Tafi fimbriae for biofilm formation on HEp-2 tissue culture cells and chicken intestinal epithelium". eng. In: *Infection and Immunity* 74.6 (June 2006), pp. 3156–3169. DOI: 10.1128/IAI.01428-05.

[430] Theresa D. Ho and James M. Slauch. "Characterization of grvA, an Antivirulence Gene on the Gifsy-2 Phage in Salmonella enterica Serovar Typhimurium". In: *Journal of Bacteriology* 183.2 (Jan. 2001), pp. 611–620. DOI: 10.1128/JB.183.2.611-620.2001.

[431] Avital Tidhar et al. "Periplasmic superoxide dismutase SodCI of Salmonella binds peptidoglycan to remain tethered within the periplasm". en. In: *Molecular microbiology* 97.5 (Sept. 2015). Publisher: NIH Public Access, p. 832. DOI: 10.1111/mmi.13067.

[432] M. L. Lesnick et al. "The Salmonella spvB virulence gene encodes an enzyme that ADP-ribosylates actin and destabilizes the cytoskeleton of eukaryotic cells". eng. In: *Molecular Microbiology* 39.6 (Mar. 2001), pp. 1464–1470. DOI: 10.1046/j.1365-2958.2001.02360.x.

[433] Julien Mambu et al. "An Updated View on the Rck Invasin of Salmonella: Still Much to Discover". en. In: *Frontiers in Cellular and Infection Microbiology* 7 (2017). Publisher: Frontiers Media SA. DOI: 10.3389/fcimb.2017.00500.

[434] Carlos A. Santiviago et al. "The Salmonella enterica sv. Typhimurium smvA, yddG and ompD (porin) genes are required for the efficient efflux of methyl viologen". eng. In: *Molecular Microbiology* 46.3 (Nov. 2002), pp. 687–698. DOI: 10.1046/j.1365-2958.2002.03204.x.

[435] Natalya Baranova and Hiroshi Nikaido. "The BaeSR Two-Component Regulatory System Activates Transcription of the yegMNOB (mdtABCD) Transporter Gene Cluster in Escherichia coli and Increases Its Resistance to Novobiocin and Deoxycholate". In: *Journal of Bacteriology* 184.15 (Aug. 2002), pp. 4168–4176. DOI: 10.1128/JB.184.15.4168-4176.2002.

[436] Haoyu Zhang, Yanghee Kim, and Prabir K. Dutta. "Controlled release of paraquat from surface-modified zeolite Y". en. In: *Microporous and Mesoporous Materials* 88.1 (Jan. 2006), pp. 312–318. DOI: 10.1016/j.micromeso.2005.09.026.

[437] Takashi Inaoka, Yoshinobu Matsumura, and Tetsuaki Tsuchido. "SodA and Manganese Are Essential for Resistance to Oxidative Stress in Growing and Sporulating Cells of Bacillus subtilis". In: *Journal of Bacteriology* 181.6 (Mar. 1999), pp. 1939–1943.

[438] Matthew E. Wand et al. "SmvA is an important efflux pump for cationic biocides in Klebsiella pneumoniae and other Enterobacteriaceae". eng. In: *Scientific Reports* 9.1 (Feb. 2019), p. 1344. DOI: 10.1038/s41598-018-37730-0.

[439] Hongo Etsuko et al. "The methyl viologen-resistance-encoding gene smvA of Salmonella typhimurium". en. In: *Gene* 148.1 (Oct. 1994), pp. 173–174. DOI: 10.1016/0378-1119(94)90255-0.

[440] Lateef Babatunde Salam. "Unravelling the antibiotic and heavy metal resistome of a chronically polluted soil". In: *3 Biotech* 10.6 (June 2020), p. 238. DOI: 10.1007/s13205-020-02219-z.

[441] Mingma Thundu Sherpa et al. "Distribution of antibiotic and metal resistance genes in two glaciers of North Sikkim, India". en. In: *Ecotoxicology and Environmental Safety* 203 (Oct. 2020), p. 111037. DOI: 10.1016/j.ecoenv.2020.111037.

[442] Julieta Barchiesi et al. "mgtA Expression Is Induced by Rob Overexpression and Mediates a Salmonella enterica Resistance Phenotype". In: *Journal of Bacteriology* 190.14 (July 2008), pp. 4951–4958. DOI: 10.1128/JB.00195-08.

[443] Bo-Young Yoon et al. "Structure of the periplasmic copper-binding protein CueP from Salmonella enterica serovar Typhimurium". eng. In: *Acta Crystallographica. Section D, Biological Crystallography* 69.Pt 10 (Oct. 2013), pp. 1867–1875. DOI: 10.1107/S090744491301531X.

[444] Emily V. Bushby, Louise Dye, and Lisa M. Collins. "Is Magnesium Supplementation an Effective Nutritional Method to Reduce Stress in Domestic Pigs? A Systematic Review". In: *Frontiers in Veterinary Science* 7 (2021).

[445] Joana Mourão et al. "Metal tolerance in emerging clinically relevant multidrug-resistant Salmonella enterica serotype 4,[5],12:i:- clones circulating in Europe". eng. In: *International Journal of Antimicrobial Agents* 45.6 (June 2015), pp. 610–616. DOI: 10.1016/j.ijantimicag.2015.01.013.

[446] A. Gupta et al. "Molecular basis for resistance to silver cations in Salmonella". eng. In: *Nature Medicine* 5.2 (Feb. 1999), pp. 183–188. DOI: 10.1038/5545.

[447] Ibtissem Ben Fekih et al. "Distribution of Arsenic Resistance Genes in Prokaryotes". In: *Frontiers in Microbiology* 9 (2018).

[448] Priscilla Branchu, Matt Bawn, and Robert A. Kingsley. "Genome Variation and Molecular Epidemiology of Salmonella enterica Serovar Typhimurium Pathovariants". In: *Infection and Immunity* 86.8 (July 2018). DOI: 10.1128/IAI.00079-18.

[449] Eric Boyd and Tamar Barkay. "The Mercury Resistance Operon: From an Origin in a Geothermal Environment to an Efficient Detoxification Machine". In: *Frontiers in Microbiology* 3 (2012).

[450] Xuanji Li et al. "Metagenomic evidence for co-occurrence of antibiotic, biocide and metal resistance genes in pigs". en. In: *Environment International* 158 (Jan. 2022), p. 106899. DOI: 10.1016/j.envint.2021.106899.

[451] Martine Braibant et al. "Structural and functional study of the phenicol-specific efflux pump FloR belonging to the major facilitator superfamily". eng. In: *Antimicrobial Agents and Chemotherapy* 49.7 (July 2005), pp. 2965–2971. DOI: 10.1128/AAC.49.7.2965-2971.2005.

[452] Ying Xu et al. "Moritella cold-active dihydrofolate reductase: are there natural limits to optimization of catalytic efficiency at low temperature?" eng. In: *Journal of Bacteriology* 185.18 (Sept. 2003), pp. 5519–5526. DOI: 10.1128/JB.185.18.5519-5526.2003.

[453] Jacob Yerushalmy. "Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques". In: *Public Health Reports (1896-1970)* 62.40 (1947). Publisher: Association of Schools of Public Health, pp. 1432–1449. DOI: 10.2307/4586294.

[454] Peter H. Van der Meide et al. "Elongation factor Tu isolated from *Escherichia coli* mutants altered in *tufA* and *tufB*". en. In: *Proceedings of the National Academy of Sciences* 77.7 (July 1980), pp. 3922–3926. DOI: 10.1073/pnas.77.7.3922.

[455] Mark L. V. Tizard et al. "p43, the protein product of the atypical insertion sequence IS900, is expressed in Mycobacterium paratuberculosis". In: *Microbiology* 138.8 (). Publisher: Microbiology Society, pp. 1729–1736. DOI: 10.1099/00221287-138-8-1729.

[456] Victor A. Jegede, Friederica Spencer, and Jean E. Brenchley. "Thialysine-Resistant Mutant of SALMONELLA TYPHIMURIUM with a Lesion in the thrA Gene". In: *Genetics* 83.4 (Aug. 1976), pp. 619–632.

[457] Prapas Patchanee et al. "Whole-genome characterisation of multidrug resistant monophasic variants of Salmonella Typhimurium from pig production in Thailand". en. In: *PeerJ* 8 (Aug. 2020). Publisher: PeerJ Inc., e9700. DOI: 10.7717/peerj.9700.

[458] Wichai Tantasuparuk and Annop Kunavongkrit. "PIG PRODUCTION IN THAILAND". en. In: (), p. 9.

[459] Catherine Llanes et al. "Genetic analysis of a multiresistant strain of Pseudomonas aeruginosa producing PER-1 beta-lactamase". In: *Clinical Microbiology and Infection* 12.3 (Mar. 2006). Publisher: Elsevier for the European Society of Clinical Microbiology and Infectious Diseases, pp. 270–278. DOI: 10.1111/j.1469-0691.2005.01333.x.

[460] David J. Samuels et al. "Use of a promiscuous, constitutively-active bacterial enhancer-binding protein to define the sigma-54 (RpoN) regulon of Salmonella Typhimurium LT2". en. In: *BMC Genomics* 14 (2013). Publisher: BioMed Central, p. 602. DOI: 10.1186/1471-2164-14-602.

[461] Christopher A. Lopez et al. "The Periplasmic Nitrate Reductase NapABC Supports Luminal Growth of Salmonella enterica Serovar Typhimurium during Colitis". en. In: *Infection and Immunity* 83.9 (Sept. 2015). Publisher: American Society for Microbiology (ASM), p. 3470. DOI: 10.1128/IAI.00351-15.

[462] Christophe Merlin et al. "Why Is Carbonic Anhydrase Essential to Escherichia coli?" In: *Journal of Bacteriology* 185.21 (Nov. 2003), pp. 6415–6424. DOI: 10.1128/JB.185.21.6415-6424.2003.

[463] David H. Bechhofer and Murray P. Deutscher. "Bacterial ribonucleases and their roles in RNA metabolism". In: *Critical reviews in biochemistry and molecular biology* 54.3 (June 2019), pp. 242–300. DOI: 10.1080/10409238.2019.1651816.

[464] Anne Burns et al. "Assessing the role of feed as a risk factor for Salmonella in pig production". In: Jan. 2013, pp. 125–128. DOI: 10.31274/safepork-180809-929.

[465] Peter R. Davies et al. "The role of contaminated feed in the epidemiology and control of Salmonella enterica in pork production". eng. In: *Foodborne Pathogens and Disease* 1.4 (2004), pp. 202–215. DOI: 10.1089/fpd.2004.1.202.

[466] Jan M. Sargeant et al. "Salmonella in Animal Feeds: A Scoping Review". In: *Frontiers in Veterinary Science* 8 (2021).

[467] *ITAVI : Nettoyage et désinfection d'un bâtiment.*

[468] Vikrant Dutta, Driss Elhanafi, and Sophia Kathariou. "Conservation and distribution of the benzalkonium chloride resistance cassette bcrABC in Listeria monocytogenes". eng. In: *Applied and Environmental Microbiology* 79.19 (Oct. 2013), pp. 6067–6074. DOI: 10.1128/AEM.01751-13.

[469] Arturo Rodríguez-Blanco, Manuel L. Lemos, and Carlos R. Osorio. "Integrating Conjugative Elements as Vectors of Antibiotic, Mercury, and Quaternary Ammonium Compound Resistance in Marine Aquaculture Environments". In: *Antimicrobial Agents and Chemotherapy* 56.5 (May 2012), pp. 2619–2626. DOI: 10.1128/AAC.05997-11.

[470] Silvia Guillén et al. "Impact of the Resistance Responses to Stress Conditions Encountered in Food and Food Processing Environments on the Virulence and Growth Fitness of Non-Typhoidal Salmonellae". en. In: *Foods* 10.3 (Mar. 2021). Publisher: Multidisciplinary Digital Publishing Institute (MDPI). DOI: 10.3390/foods10030617.

[471] Diana Pradhan and Vidya Devi Negi. "Stress-induced adaptations in Salmonella: A ground for shaping its pathogenesis". en. In: *Microbiological Research* 229 (Dec. 2019), p. 126311. DOI: 10.1016/j.micres.2019.126311.

[472] H. Zeng et al. "Salmonella prevalence and persistence in industrialized poultry slaughterhouses". en. In: *Poultry Science* 100.4 (Apr. 2021), p. 100991. DOI: 10.1016/j.psj.2021.01.014.

[473] T. R. Callaway et al. "Social stress increases fecal shedding of Salmonella typhimurium by early weaned piglets". eng. In: *Current Issues in Intestinal Microbiology* 7.2 (Sept. 2006), pp. 65–71.

[474] Elin Verbrugghe et al. "Stress induced Salmonella Typhimurium recrudescence in pigs coincides with cortisol induced increased intracellular proliferation in macrophages". In: *Veterinary Research* 42.1 (2011), p. 118. DOI: 10.1186/1297-9716-42-118.

[475] Isabelle CORRÉGÉ and Brice Minvielle. "Enjeux et stratégies de maîtrise de Salmonella dans la filière porcine : une analyse prospective". In: 58 (Sept. 2013), pp. 2–7.

[476] Annaëlle Kerouanton et al. *Salmonella in pig farming : excretion level, serovars and resistance to antibiotics.*

[477] M. Denis et al. *Level of Salmonella excretion of fattening pigs in alternative farm.*

[478] Hang Zeng et al. "Identification of the Source for Salmonella Contamination of Carcasses in a Large Pig Slaughterhouse". In: *Pathogens* 10.1 (Jan. 2021), p. 77. DOI: 10.3390/pathogens10010077.

[479] Torsten Seemann. *mlst (https://github.com/tseemann/mlst). Accessed June 17, 2019.* original-date: 2014-05-03T09:12:11Z. June 2019.

[480] Herbert Dei. "Soybean as a Feed Ingredient for Livestock and Poultry". In: Oct. 2011. DOI: 10.5772/17601.

[481] John W. Campbell, Rachael M. Morgan-Kiss, and John E. Cronan. "A new Escherichia coli metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway". eng. In: *Molecular Microbiology* 47.3 (Feb. 2003), pp. 793–805. DOI: 10.1046/j.1365-2958.2003.03341.x.

[482] Hideyuki Ohshima et al. "Molecular organization of intrinsic restriction and modification genes BsuM of Bacillus subtilis Marburg". eng. In: *Journal of Bacteriology* 184.2 (Jan. 2002), pp. 381–389. DOI: 10.1128/JB.184.2.381-389.2002.

[483] Yezhi Fu et al. *Salmonella enterica serovar Typhimurium from Wild Birds in the United States Represent Distinct Lineages Defined by Bird Type*. en. preprint. Microbiology, Nov. 2021. DOI: 10.1101/2021.11.26.470158.

[484] Jorge Hernandez et al. "Salmonella in Birds Migrating through Sweden". In: *Emerging Infectious Diseases* 9.6 (June 2003), pp. 753–755. DOI: 10.3201/eid0906.030072.

[485] Çagla Tükel et al. "CsgA is a pathogen-associated molecular pattern of Salmonella enterica serotype Typhimurium that is recognized by Toll-like receptor 2". en. In: *Molecular Microbiology* 58.1 (2005). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.2005.04825.x, pp. 289–304. DOI: 10.1111/j.1365-2958.2005.04825.x.

[486] Shaohui Wang et al. "Autotransporter MisL of Salmonella enterica serotype Typhimurium facilitates bacterial aggregation and biofilm formation". In: *FEMS Microbiology Letters* 365.17 (Sept. 2018), fny142. DOI: 10.1093/femsle/fny142.

[487] *La désinfection des trayons pour une bonne hygiène de traite : en 4 étapes !* fr.

[488] *Désinfection des trayons et des manchons avec Kersia*. fr.

[489] Brendan Flannery et al. "Reducing Legionella Colonization of Water Systems with Monochloramine". In: *Emerging Infectious Diseases* 12.4 (Apr. 2006), pp. 588–596. DOI: 10.3201/eid1204.051101.

[490] N. S. Bolan et al. "Distribution and bioavailability of copper in farm effluent". en. In: *Science of The Total Environment* 309.1 (June 2003), pp. 225–236. DOI: 10.1016/S0048-9697(03)00052-4.

[491] J. J. Parkins et al. "The effectiveness of copper oxide powder as a component of a sustained-release multi-trace element and vitamin rumen bolus system for cattle". en. In: *British Veterinary Journal* 150.6 (Nov. 1994), pp. 547–553. DOI: 10.1016/S0007-1935(94)80038-3.

[492] Paramita Mandal. "An insight of environmental contamination of arsenic on animal health". en. In: *Emerging Contaminants* 3.1 (Mar. 2017), pp. 17–22. DOI: 10.1016/j.emcon.2017.01.004.

[493] Sung-Young Lim et al. "CuiD is a crucial gene for survival at high copper environment in Salmonella enterica serovar Typhimurium". eng. In: *Molecules and Cells* 14.2 (Oct. 2002), pp. 177–184.

[494] Katherine Vallejos-Sánchez et al. "Whole-Genome Sequencing of a Salmonella enterica subsp. enterica Serovar Infantis Strain Isolated from Broiler Chicken in Peru". In: *Microbiology Resource Announcements* 8.43 (Oct. 2019). Publisher: American Society for Microbiology, e00826–19. DOI: 10.1128/MRA.00826-19.

[495] Marie Bugarel et al. "Complete Genome Sequences of Four Salmonella enterica Strains (Including Those of Serotypes Montevideo, Mbandaka, and Lubbock) Isolated from Peripheral Lymph Nodes of Healthy Cattle". eng. In: *Microbiology Resource Announcements* 8.2 (Jan. 2019), e01450–18. DOI: 10.1128/MRA.01450-18.

[496] Yann Sévellec et al. "Polyphyletic Nature of Salmonella enterica Serotype Derby and Lineage-Specific Host-Association Revealed by Genome-Wide Analysis". In: *Frontiers in Microbiology* 9 (May 2018), p. 891. DOI: 10.3389/fmicb.2018.00891.

[497] Arnar K. S. Sandholt et al. "Genomic signatures of host adaptation in group B Salmonella enterica ST416/ST417 from harbour porpoises". In: *Veterinary Research* 52.1 (Oct. 2021), p. 134. DOI: 10.1186/s13567-021-01001-0.

[498] Hiroaki Shigemura et al. "Food Workers as a Reservoir of Extended-Spectrum-Cephalosporin-Resistant Salmonella Strains in Japan". In: *Applied and Environmental Microbiology* 86.13 (June 2020). Publisher: American Society for Microbiology, e00072–20. DOI: 10.1128/AEM.00072-20.

[499] Emiliano Cohen et al. "Emergence of new variants of antibiotic resistance genomic islands among multidrug-resistant Salmonella enterica in poultry". eng. In: *Environmental Microbiology* 22.1 (Jan. 2020), pp. 413–432. DOI: 10.1111/1462-2920.14858.

[500] Geneviève Labbé et al. "Complete Genome and Plasmid Sequences of Three Canadian Isolates of Salmonella enterica subsp. enterica Serovar Heidelberg from Human and Food Sources". In: *Genome Announcements* 4.1 (Jan. 2016). Publisher: American Society for Microbiology, e01526–15. DOI: 10.1128/genomeA.01526-15.

[501] Cristina Iftode, Yaron Daniely, and James A. Borowiec. "Replication Protein A (RPA): The Eukaryotic SSB". In: *Critical Reviews in Biochemistry and Molecular Biology* 34.3 (Jan. 1999). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10409239991209255, pp. 141–180. DOI: 10.1080/10409239991209255.

[502] Leonard R. Bullas et al. "Salmonella phage PSP3, another member of the P2-like phage group". en. In: *Virology* 185.2 (Dec. 1991), pp. 918–921. DOI: 10.1016/0042-6822(91)90573-T.

[503] Jinwoo Kim et al. "Sensitive detection of viable Escherichia coli O157:H7 from foods using a luciferase-reporter phage phiV10lux". en. In: *International Journal of Food Microbiology* 254 (Aug. 2017), pp. 11–17. DOI: 10.1016/j.ijfoodmicro.2017.05.002.

[504] Andrea I. Moreno Switt et al. "Genomic characterization provides new insight into Salmonella phage diversity". In: *BMC Genomics* 14.1 (July 2013), p. 481. DOI: 10.1186/1471-2164-14-481.

[505] F. Martelli et al. "Abattoir-based study of Salmonella prevalence in pigs at slaughter in Great Britain". In: *Epidemiology and Infection* 149 (Sept. 2021), e218. DOI: 10.1017/S0950268821001631.

[506] Mark Achtman et al. "Genomic diversity of Salmonella enterica -The UoWUCC 10K genomes project". In: *Wellcome Open Research* 5 (Feb. 2021), p. 223. DOI: 10.12688/wellcomeopenres.16291.2.

[507] Josep Casadesús and David Low. "Epigenetic Gene Regulation in the Bacterial World". In: *Microbiology and Molecular Biology Reviews* 70.3 (Sept. 2006). Publisher: American Society for Microbiology, pp. 830–856. DOI: 10.1128/MMBR.00016-06.

[508] Claudia E. Coipan et al. "Concordance of SNP- and allele-based typing workflows in the context of a large-scale international Salmonella Enteritidis outbreak investigation". en. In: *Microbial Genomics* 6.3 (Mar. 2020). Publisher: Microbiology Society. DOI: 10.1099/mgen.0.000318.

[509] Pimlapas Leekitcharoenphon et al. "Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica". en. In: *PLOS ONE* 9.2 (2014). Publisher: Public Library of Science, e87991. DOI: 10.1371/journal.pone.0087991.

[510] Angela J. Taylor et al. "Characterization of Foodborne Outbreaks of Salmonella enterica Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection". en. In: *Journal of Clinical Microbiology* 53.10 (Oct. 2015). Publisher: American Society for Microbiology (ASM), p. 3334. DOI: 10.1128/JCM.01280-15.

[511] Nobuo Arai et al. "Phylogenetic Characterization of Salmonella enterica Serovar Typhimurium and Its Monophasic Variant Isolated from Food Animals in Japan Revealed Replacement of Major Epidemic Clones in the Last 4 Decades". en. In: *Journal of Clinical Microbiology* 56.5 (May 2018), e01758–17. DOI: `10.1128/JCM.01758-17`.

[512] Nanna Munck et al. "Four European Salmonella Typhimurium datasets collected to develop WGS-based source attribution methods". In: *Scientific Data* 7 (Dec. 2020). DOI: `10.1038/s41597-020-0417-7`.

[513] Philip M Ashton et al. "Whole Genome Sequencing for the Retrospective Investigation of an Outbreak of Salmonella Typhimurium DT 8". In: *PLoS Currents* 7 (Feb. 2015). DOI: `10.1371/currents.outbreaks.2c05a47d292f376afc5a6fcdd8a7a3b6`.

[514] Eglė Kudirkiene et al. "Epidemiology of Salmonella enterica Serovar Dublin in Cattle and Humans in Denmark, 1996 to 2016: a Retrospective Whole-Genome-Based Study". eng. In: *Applied and Environmental Microbiology* 86.3 (Jan. 2020), e01894–19. DOI: `10.1128/AEM.01894-19`.

[515] Adriano Di Pasquale et al. "SARS-CoV-2 surveillance in Italy through phylogenomic inferences based on Hamming distances derived from pan-SNPs, -MNPs and -InDels". In: *BMC Genomics* 22.1 (Oct. 2021), p. 782. DOI: `10.1186/s12864-021-08112-0`.

[516] Yang Liu, Yiu Fai Lee, and Michael K. Ng. "SNP and gene networks construction and analysis from classification of copy number variations data". In: *BMC Bioinformatics* 12.5 (July 2011), S4. DOI: `10.1186/1471-2105-12-S5-S4`.

[517] Shuzhen Sun et al. "SNP variable selection by generalized graph domination". en. In: *PLOS ONE* 14.1 (Jan. 2019). Publisher: Public Library of Science, e0203242. DOI: `10.1371/journal.pone.0203242`.

[518] Ahmad M. Alqudah et al. "Genome-wide and SNP network analyses reveal genetic control of spikelet sterility and yield-related traits in wheat". en. In: *Scientific Reports* 10.1 (Feb. 2020). Number: 1 Publisher: Nature Publishing Group, p. 2098. DOI: `10.1038/s41598-020-59004-4`.

[519] Delphine Charif and Jean R. Lobry. "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis". en. In: *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Ed. by Ugo Bastolla et al. Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer, 2007, pp. 207–232. DOI: `10.1007/978-3-540-35306-5_10`.

[520] Sacha Epskamp et al. "qgraph: Network Visualizations of Relationships in Psychometric Data". en. In: *Journal of Statistical Software* 48 (May 2012), pp. 1–18. DOI: `10.18637/jss.v048.i04`.

[521] *Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX.*

[522] Matthew R. Hayward et al. "Population structure and associated phenotypes of Salmonella enterica serovars Derby and Mbandaka overlap with host range". In: *BMC Microbiology* 16.1 (Feb. 2016), p. 15. DOI: `10.1186/s12866-016-0628-4`.

[523] Robert Lanfear, Xia Hua, and Dan L. Warren. "Estimating the Effective Sample Size of Tree Topologies from Bayesian Phylogenetic Analyses". In: *Genome Biology and Evolution* 8.8 (July 2016), pp. 2319–2332. DOI: `10.1093/gbe/evw171`.

Chapter 7

Supplementary material

80	1.361773	1.360819	1.35625	1.355663	1.357659	1.3595767	1.404713	1.482368
90	1.344999	1.342033	1.338907	1.339457	1.340508	1.342177	1.377846	1.446672
95	1.330596	1.327816	1.324656	1.324277	1.326374	1.327595	1.370285	1.4209137
99	1.335831	1.334858	1.329171	1.329082	1.329656	1.3304837	1.348964	1.369027
99.5	1.358831	1.355907	1.352213	1.352287	1.351906	1.351697	1.370619	1.384204
99.9	1.658626	1.65735	1.653979	1.655345	1.65375	1.65351	1.663125	1.657453
id/contig	**100**	**150**	**250**	**500**	**750**	**1000**	**5000**	**10000**

Table 7.1: Entropy normalized by number of reads and alignment length. Score are multiplied by $\mathrm{e}{-}17$.

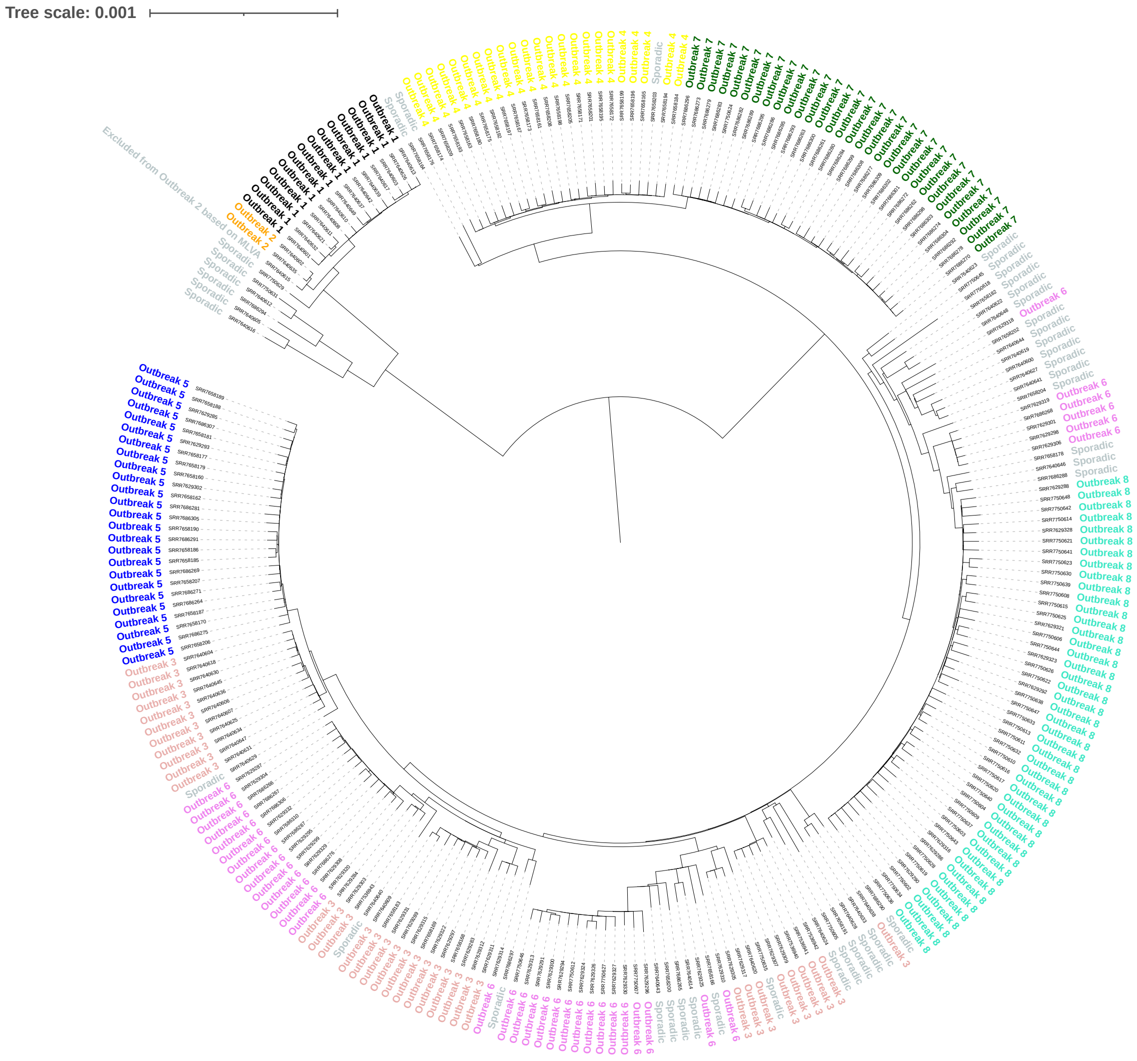


Figure 7.1: *Escherichia coli* core subtrees with discrimant contig. The outer ring corresponds to outbreaks and sporadic annotation.

Type	**Nom**	**Nb**	**Fonction**
Biocides	RPOS1	all	Multi-biocide resistance regulator
	RPOS2	5	Multi-biocide resistance regulator
	SMVA	all	Multi-biocide MFS efflux pump
	YDDG	all	Paraquat efflux pump
	SODA	**all**	**Peroxide resistance protein**
	NMPC	**all**	**Peroxide resistance protein**
Drugs	APH3/APH6	145	Aminoglycoside resistance
	CARB	22	Beta-lactam resistance
	DFRA	4	Dihydrofolate reductase
	OMPF, RAMR, SDIA	**all**	**MDR (Multi-Drug Resistance protein)**
	MLS23S	3	MLS drug resistance
	AAC6-PRIME	**all**	**Aminoglycoside resistance**
	A16S	177	Aminoglycoside resistance
	AAC3	17Q002798	Aminoglycoside resistance
	AAC3	QSC-B4-6-5	Aminoglycoside resistance
	ANT3-DPRIME	33	Aminoglycoside resistance
	LAP	2	Beta-lactam resistance
	TEM	129	Beta-lactam resistance
	TUFAB	61	Elfamycins resistance
	GYRA/B, PARC/E	**all**	**Fluoroquinolone resistance**
	QNRS	2	Fluoroquinolone resistance
	PTSL	**all**	**Fosfomycin resistance**
	CMLA	2014LSAL03857	Phenicol resistance
	FLOR	22	Phenicol resistance
	RPOB	**all**	**Rifampin resistance**
	FOLP	**all**	**Sulfonamide resistance**
	SULI	10	Sulfonamide resistance
	SULII	9	Sulfonamide resistance
	SULII	133	Sulfonamide resistance
	SULIII	2	Sulfonamide resistance
	TET	140	Tetracycline resistance
	DHFR	15	Trimethoprim resistance

Table 7.2: Table of MegaresV2 of drugs and biocides resistance results for Typhimurium dataset

Type	Name	Number of strains	Fonction
Metals	ARSA,ARSBM,ARSCM	149	Multi-metal resistance (MMR)
	CORA/B/C/D, GOLT	**all**	**MMR**
	CUEP	**all**	**Copper resistance**
	GOLS	**all**	**Gold resistance**
	MERC, MERR1,MERT	80	Mercury resistance
	MGTA	**all**	**MMR regulator**
	PCOA/B/C/D/R/S	150	Copper resistance
	PCOE	150	MMR regulator
	PSTB	**all**	**Arsenic resistance**
	SILB/C/E/F/P/S	150	MMR
	TERW/Z	4	Tellurium resistance
Multi-compound	**BAER/S**	**all**	**Drug and biocide and metal resistance**
	CUID	**all**	**Biocide and metal resistance protein**
	GESA/B/C	**all**	**Drug and biocide and metal resistance**
	MDTA/B/C/K	**all**	**Drug and biocide and metal resistance**
	ACRD	**all**	**Drug and biocide and metal resistance**
	PMRG	**all**	**Drug and metal resistance**
	QACL	3	Drug and biocide resistance
	SITA/B/C/D	**all**	**Biocide and metal resistance protein**
	SOXR	**all**	**Drug and biocide resistance**

Table 7.3: Table of MegaresV2 of metals and multi-compound resistance results for Typhimurium dataset

Type	Name	Number of strains	Fonction
Biocides	RPOS1	14	Multi-biocide resistance regulator
	RPOS2	all except S17LNR1583	Multi-biocide resistance regulator
	SMVA	all except S20LNR0591	Multi-biocide MFS efflux pump
	YDDG	all except S20LNR0591	Paraquat efflux pump
	SODA	**all**	**Peroxide resistance protein**
Drugs	DHFR	10	Dihydrofolate reductase
	FOLP	**all**	**Sulfonamide resistance**
	A16S	95	Aminoglycoside resistance
	GYRA	all except 2	Fluoroquinolone resistance
	GYRB	**all**	**Fluoroquinolone resistance**
	OMPF	**all**	**MDR (Multi-drug resistance)**
	parC/E	**all**	**Fluoroquinolone resistance**
	PTSL	all except S18LNR1829	Fosfomycin resistance
	RAMR	all except 5	MDR
	RPOB	**all**	**Rifampin resistance**
	SDIA	all except S16LNR1426	MDR
	SULI	S18LNR1211	Sulfonamide resistance
	SULII	10	Sulfonamide resistance
	TEM	9	Beta-lactam resistance
	tetA/R	11	Tetracycline resistance
	TUFAB	all except 6	EF-Tu_inhibition
Metals	**corA/B/C/D**	**all**	**MMR**
	CUEP	**all**	**Copper resistance**
	golS/T	**all**	**Gold resistance**
	merC/R1/T	S18LNR1211	Mercury resistance
	MGTA	**all**	**Multi-metal resistance protein**
	PSTB	**all**	**Arsenic resistance**
	terW/Z	S18LNR1211	Tellurium resistance
Multi-compound	baeR/S	all except S16LNR1426	Drug and biocide and metal regulator
	CUID	**all**	**Biocide and metal resistance protein**
	gesA/B/C	**all**	**Drug and biocide and metal regulator**
	mdtA/B/C/K	all except S16LNR1426	Drug and biocide and metal regulator
	ACRD	all except S18LNR1829	Drug and biocide and metal regulator
	PMRG	all except 2	Drug and metal efflux pumps
	sitA/B/C/D	**all**	**Biocide and metal efflux pumps**
	soxR/S	**all**	**Biocide and metal efflux pumps**

Table 7.4: Table of MegaresV2 results for Mbdandaka dataset

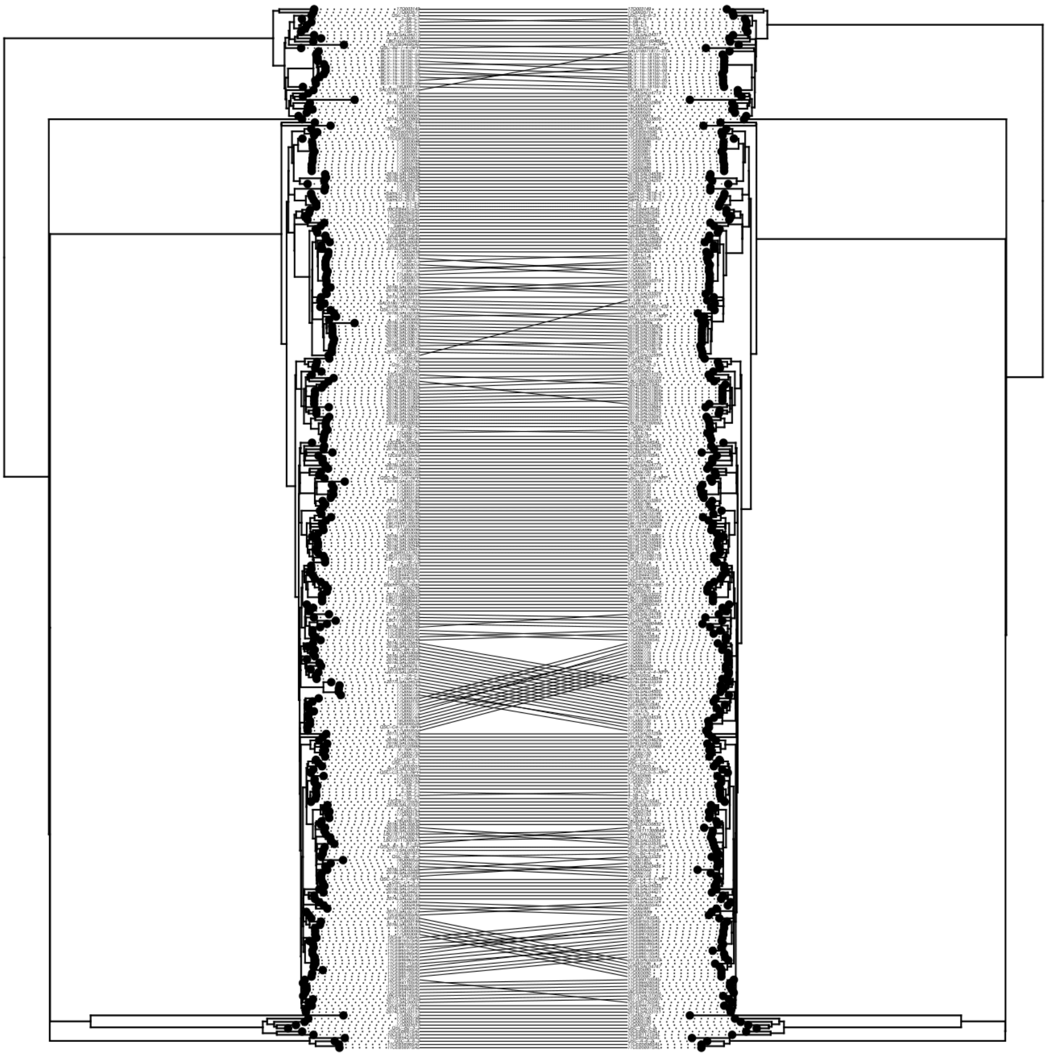


Figure 7.2: Impact of homologous recombination events on phylogenetic topology of *Salmonella* Typhimurium and its monophasic variant from France. Left: Phylogenetic tree with recombination events. Right: Phylogenetic tree with recombination events detected by ClonalFrameML and excluded. RF=196

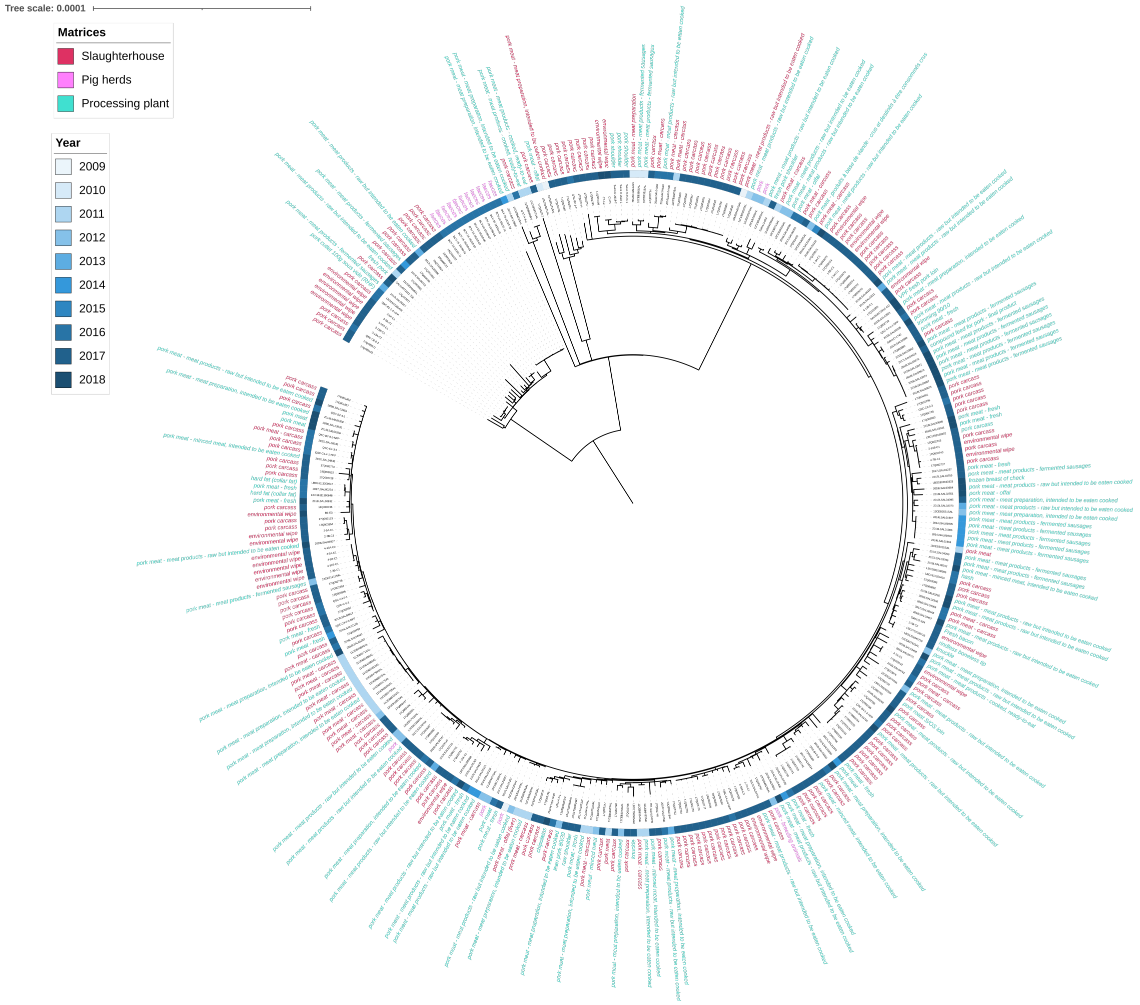


Figure 7.3: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs in France. Inner ring corresponds to the year of isolation. Outer ring corresponds to the source of isolation.

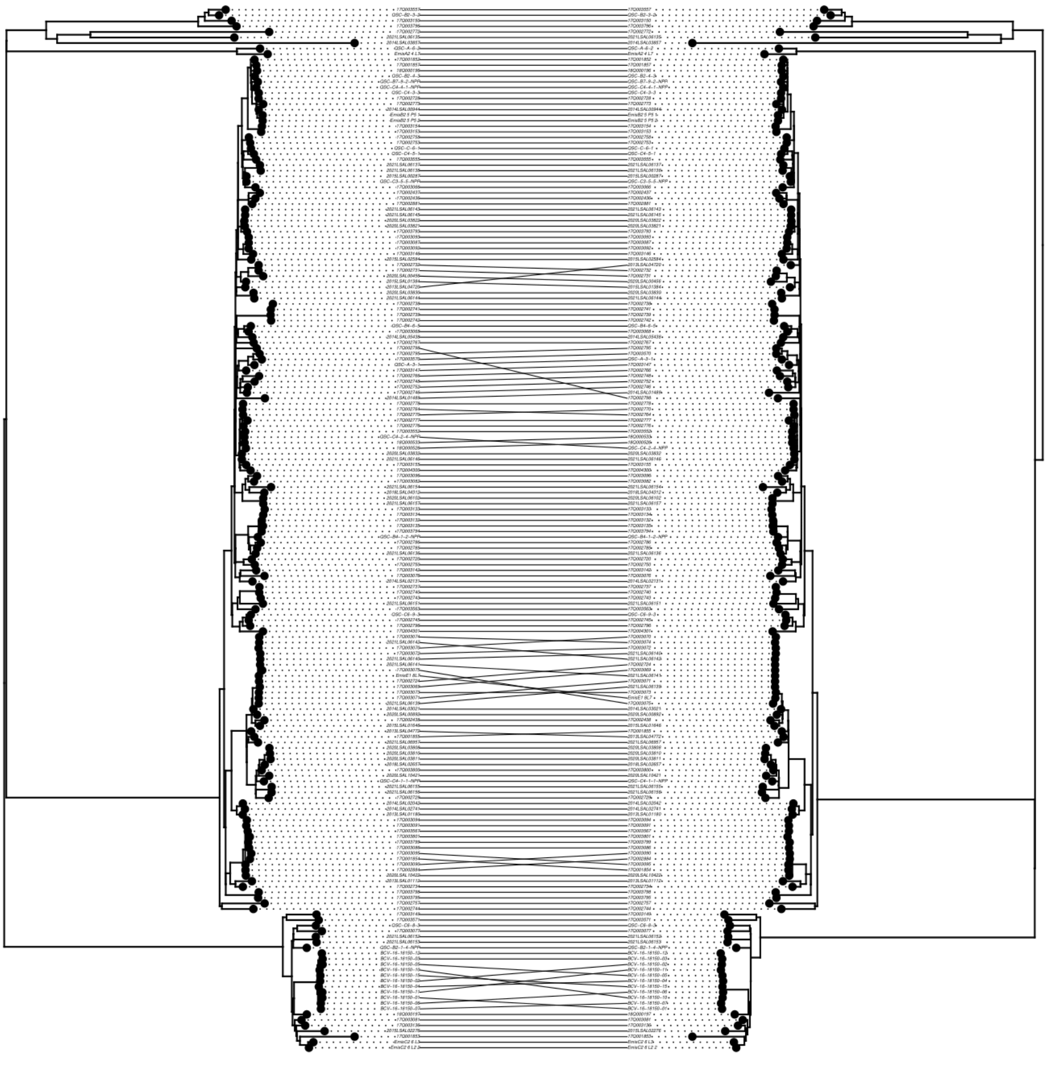


Figure 7.4: Impact of homologous recombination events on phylogenetic topology of *Salmonella* Typhimurium and its monophasic variant from pigs herds in France. Left: Phylogenetic tree with recombination events. Right: Phylogenetic tree with recombination events detected by ClonalFrameML and excluded. RF=114

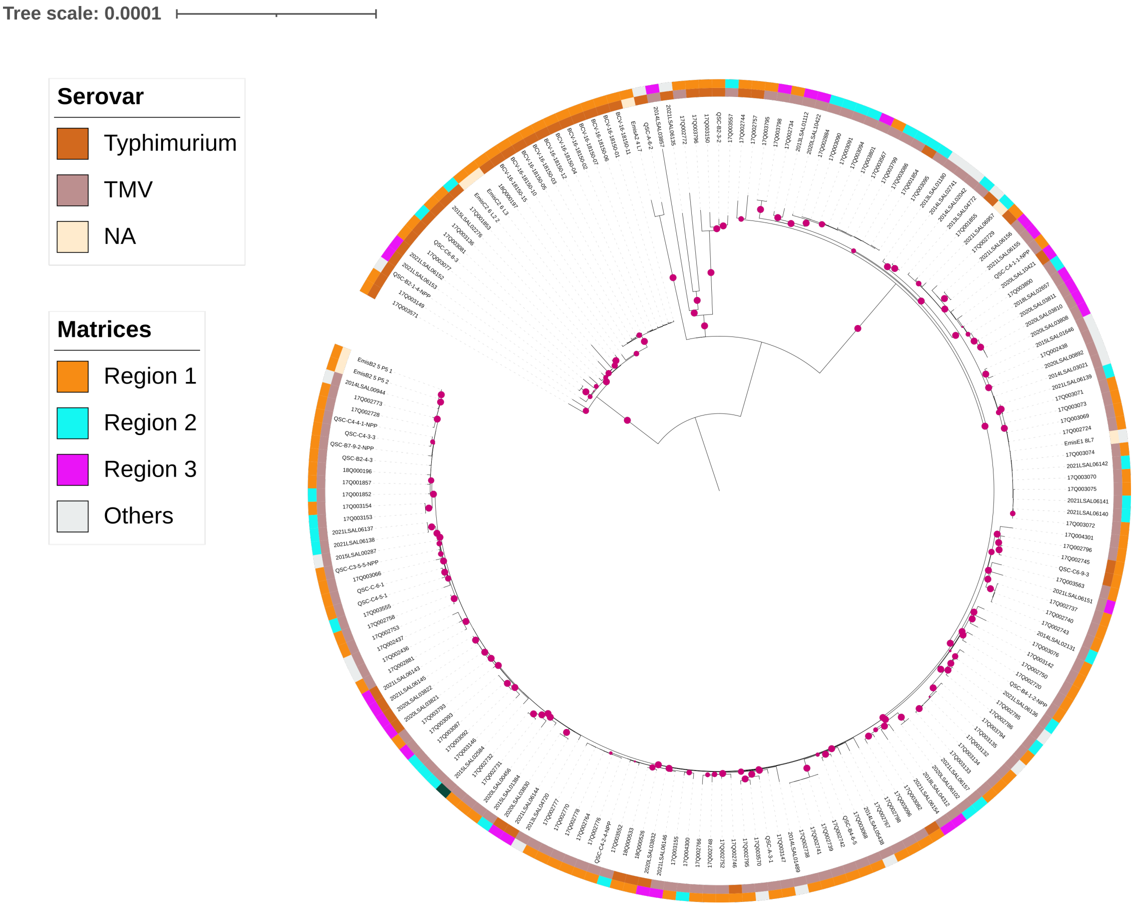


Figure 7.5: Core genome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated with pigs herds origin. Outer ring corresponds to the coding of the farms from which the strain was isolated. Branch labelled with a purple circle corresponds to branch with boostrap>90

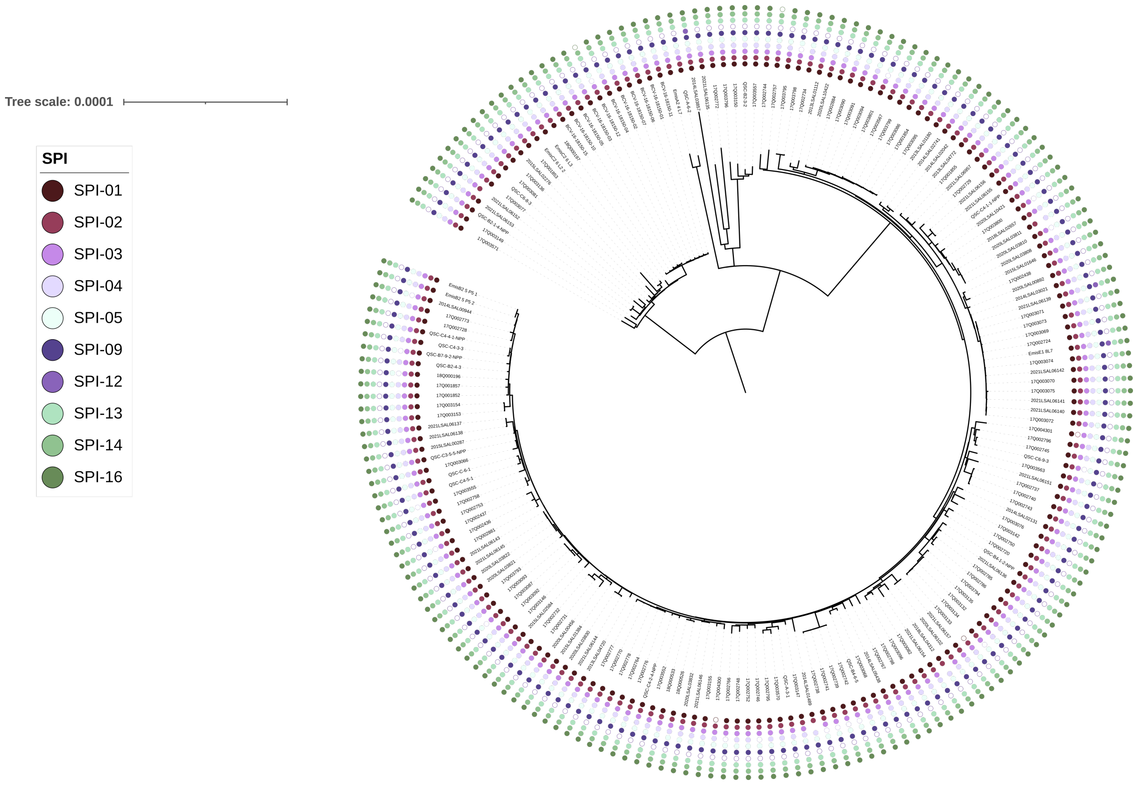


Figure 7.6: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Outer ring corresponds to the presence of SPI detected by Abricate on SPIfinder database.

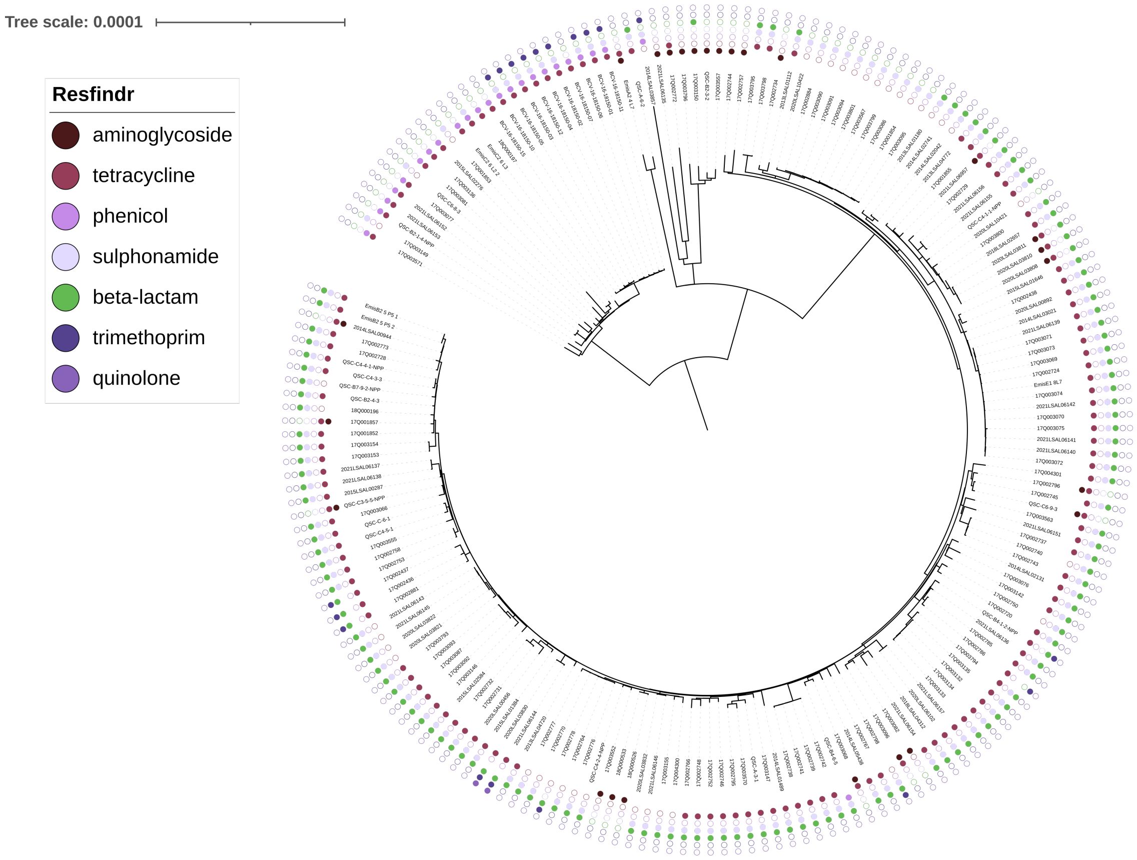


Figure 7.7: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Outer ring corresponds to the presence of antibiotic resistance detected by Abricate on ResFindr database.

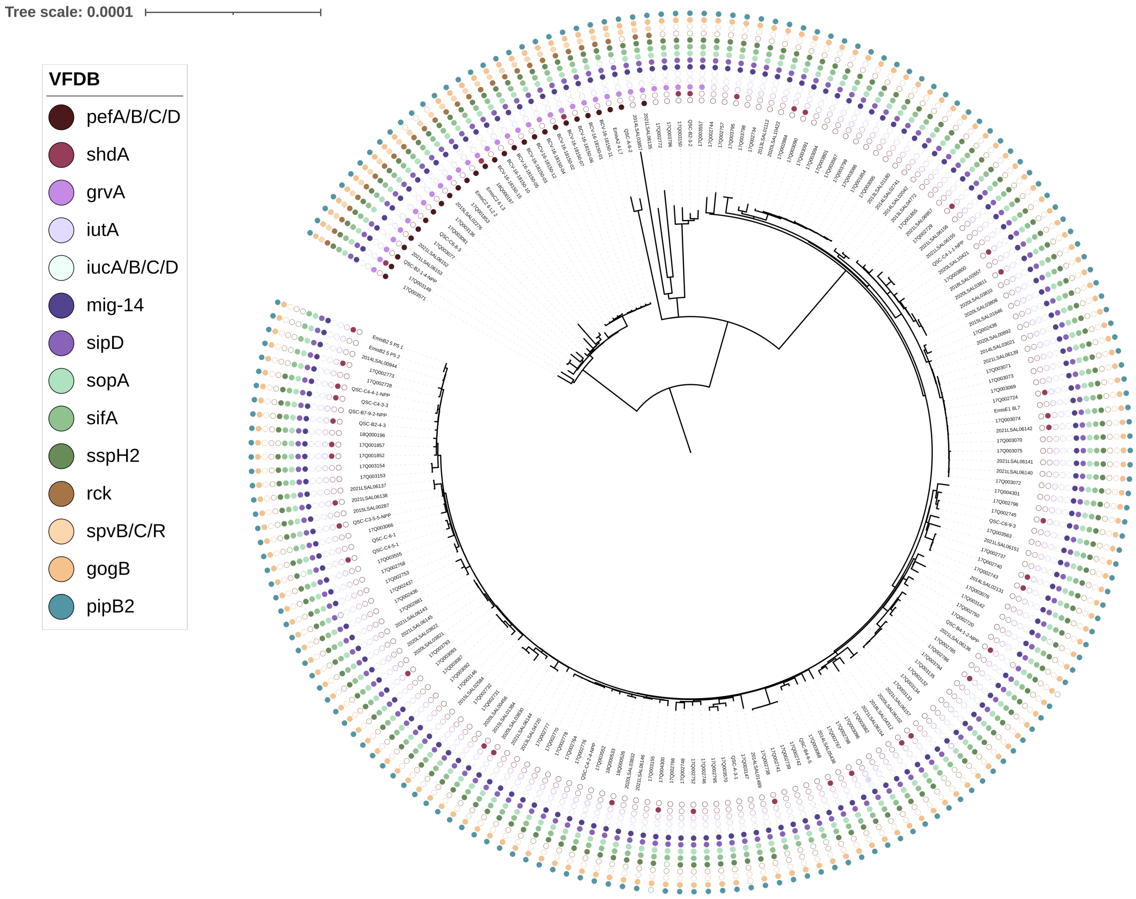


Figure 7.8: Core genome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Outer ring corresponds to the presence of virulence genes that are not detected in all genomes. Detection was inferred by Abricate on VFDB database.

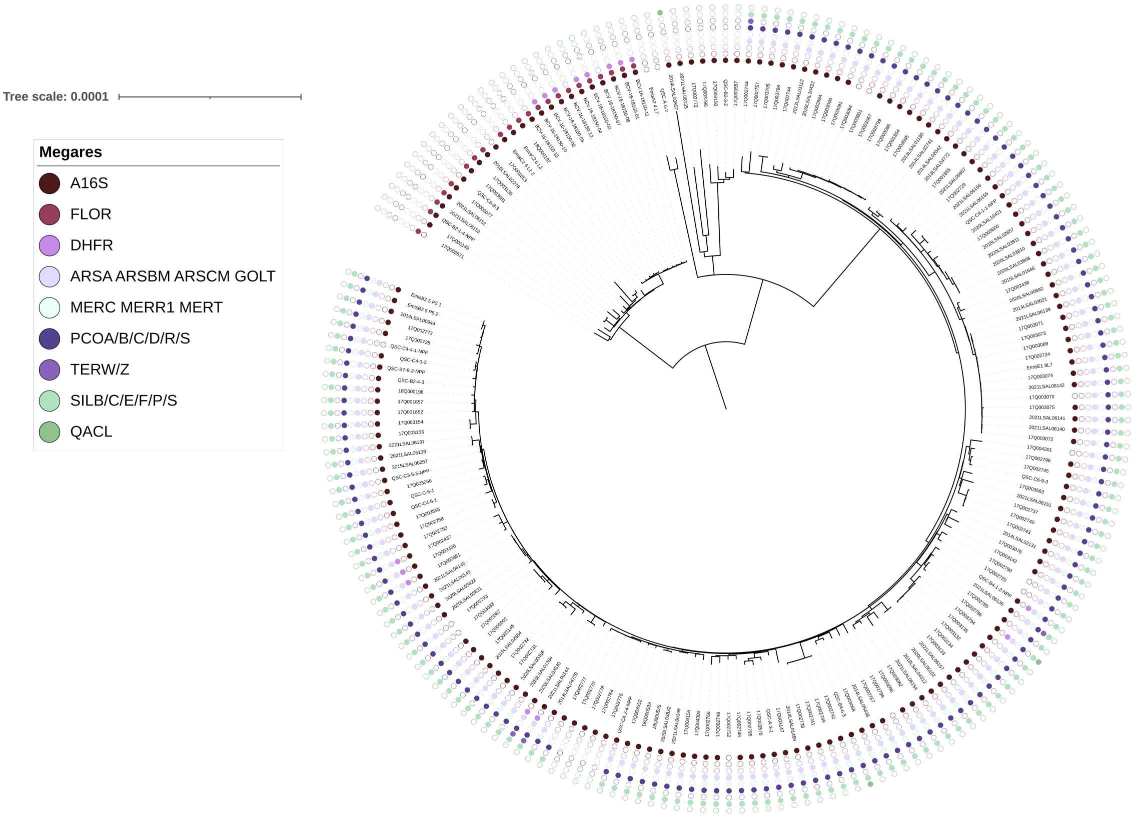


Figure 7.9: Core genome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Typhimurium and its monophasic variant isolated from pigs. Outer ring corresponds to the presence of metals, drugs, biocides and multi-compound resistance genes that are not detected in all genomes. Detection was inferred by Abricate on MegaresVS database.

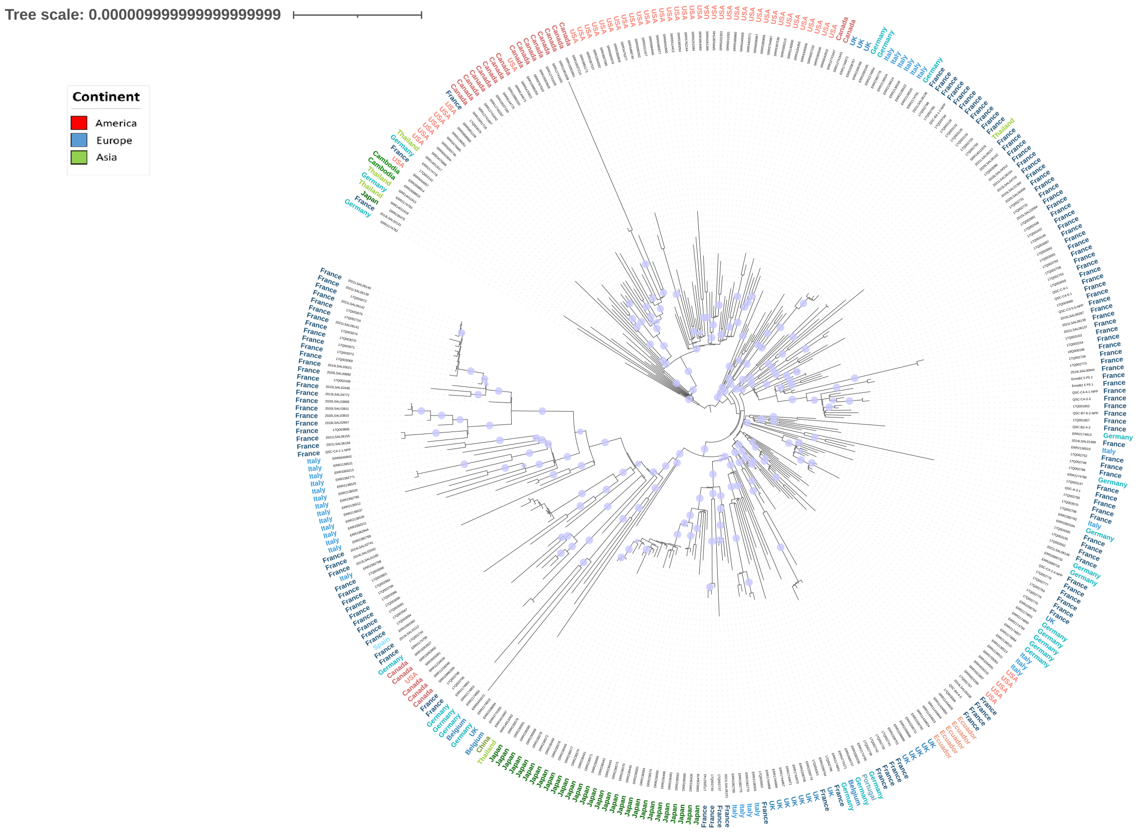


Figure 7.10: Core genome SNP-based phylogenomic reconstruction by Maximum Likelihood of monophasic variant of Typhmurium isolated from pigs from different country. Outer ring corresponds the country of selection. Blue circle corresponds to branch with boostrap value > 90.

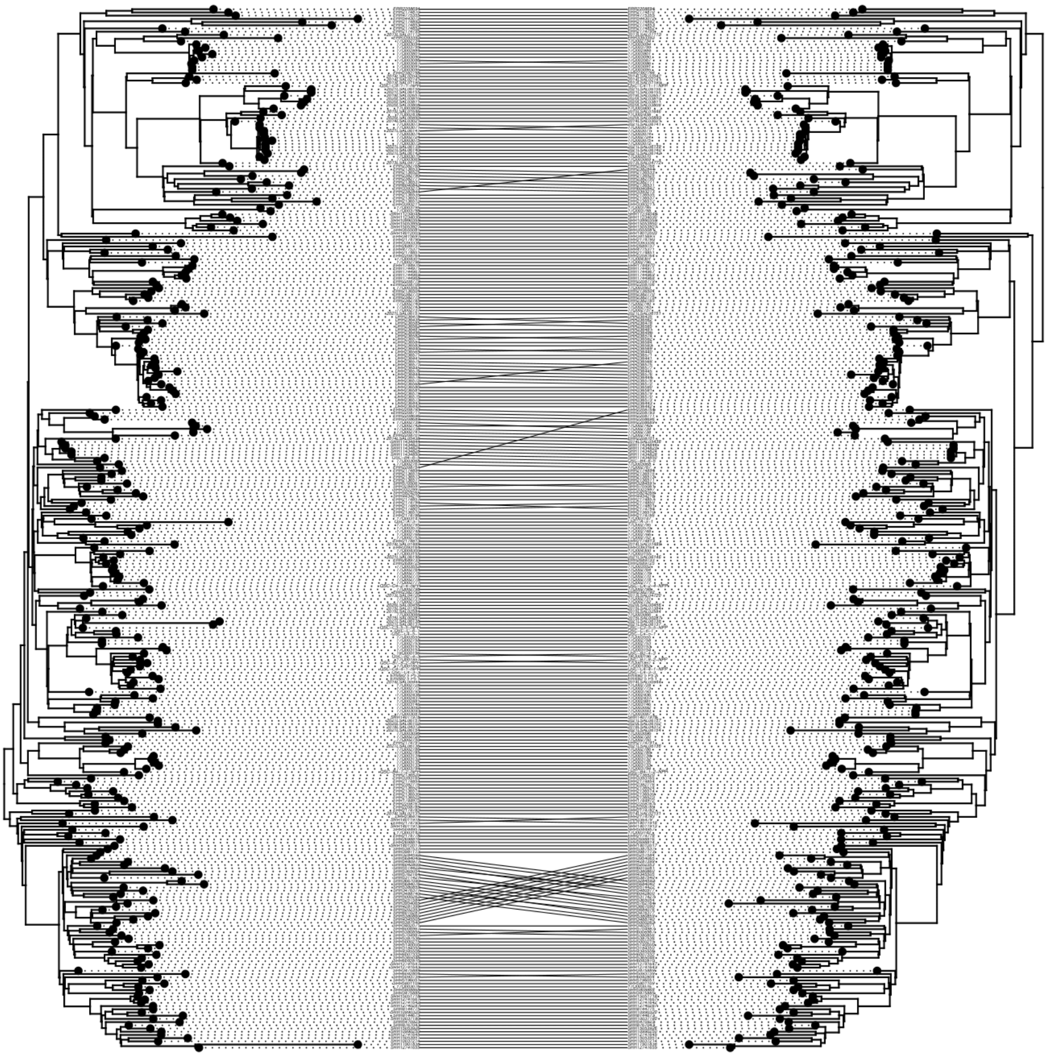


Figure 7.11: Impact of homologous recombination events on phylogenetic topology of Monophasic variant of *Salmonella* Typhimurium from pigs herds in France. Left: Phylogenetic tree with recombination events. Right: Phylogenetic tree without recombination events detected by ClonalFrameML and excluded. RF=182

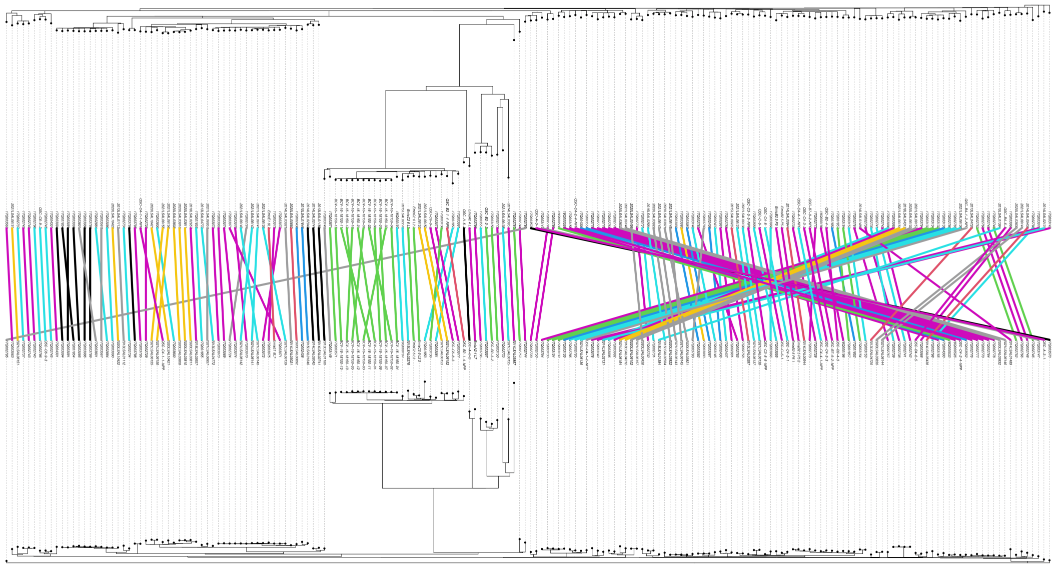


Figure 7.12: Comparison of pgSNP tree and iVARCALL2 tree on *Salmonella* Typhimurium and TMV dataset. Left tree is pgSNP tree, right tree is coregenome SNP tree. Branches are colored according to the region.

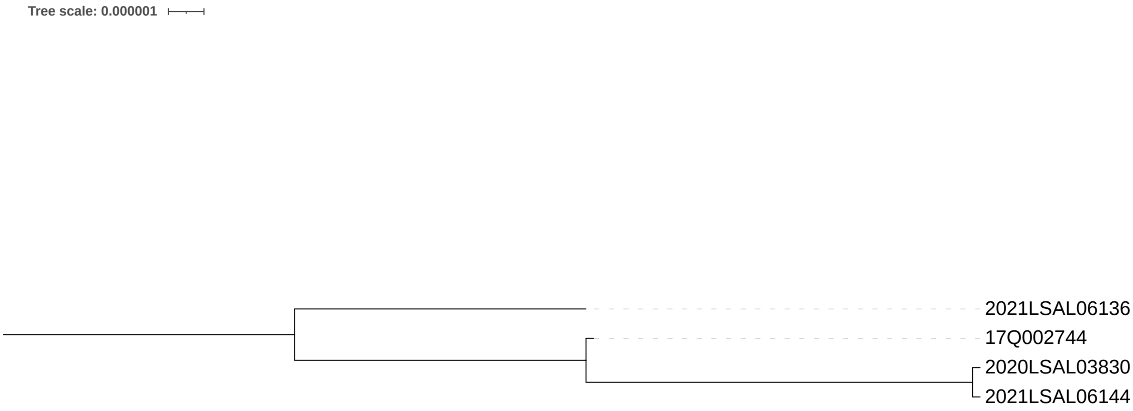


Figure 7.13: Plasmid AR_0116 subtree from pgSNP analysis

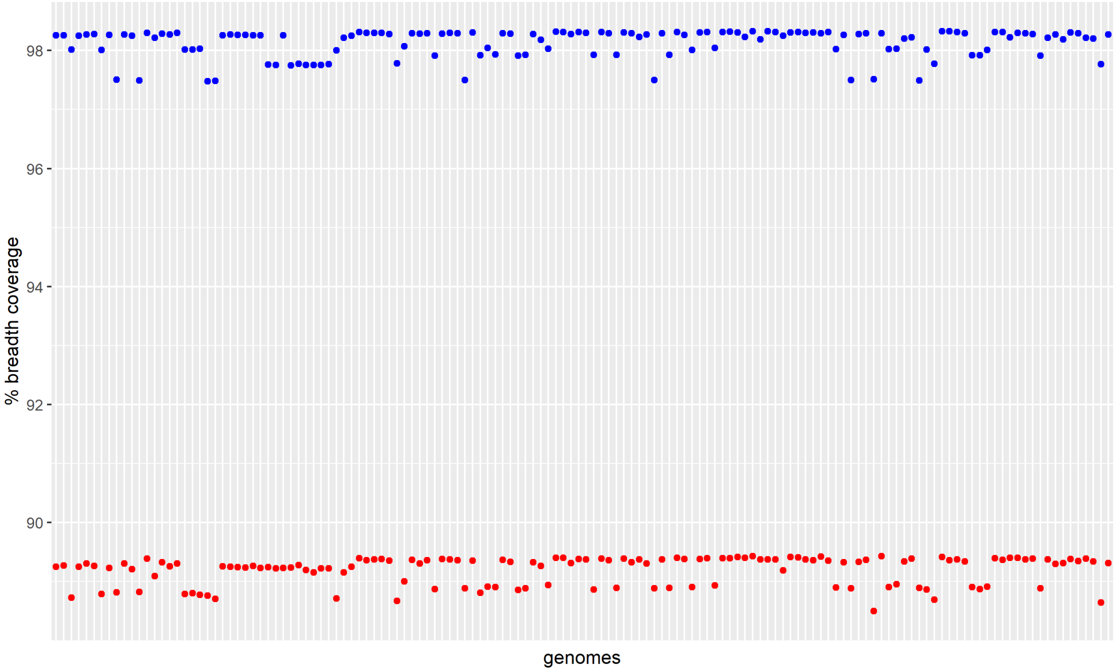


Figure 7.14: Comparison of *S.* Mbandaka breadth of coverage. In blue : reads aligned with CP022489 reference. In red : reads aligned with CP019183 reference

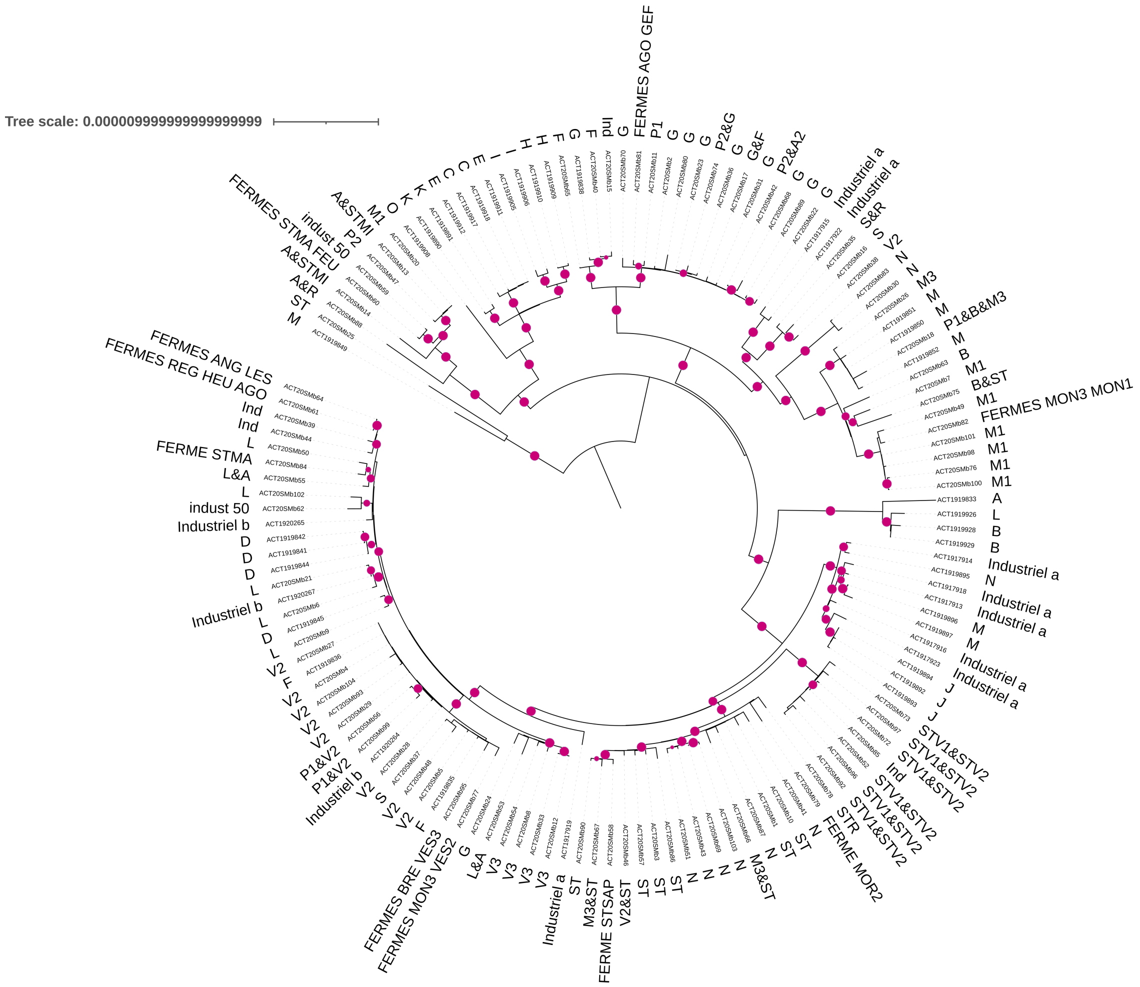


Figure 7.15: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine between 2016 and 2019. Outer ring corresponds to the coding of the farms from which the strain was isolated. Branch labelled with a purple circle corresponds to branch with boostrap>90

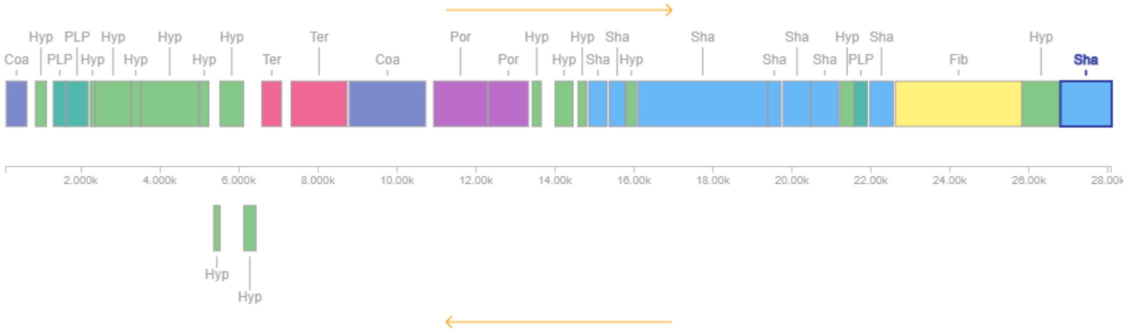

Hyp PLP Hyp Hyp Hyp Ter Por Hyp Hyp Sha Sha Sha Hyp Sha Hyp
Coa PLP Hyp Hyp Hyp Ter Coa Por Hyp Sha Hyp Sha Sha PLP Fib Sha
2.000k
4.000k
6.000k
8.000k
10.00k
12.00k
14.00k
16.00k
18.00k
20.00k
22.00k
24.00k
26.00k
28.00k
Hyp
Hyp

Figure 7.16: Plasmid ECP1 description of *Salmonella* Mbandaka

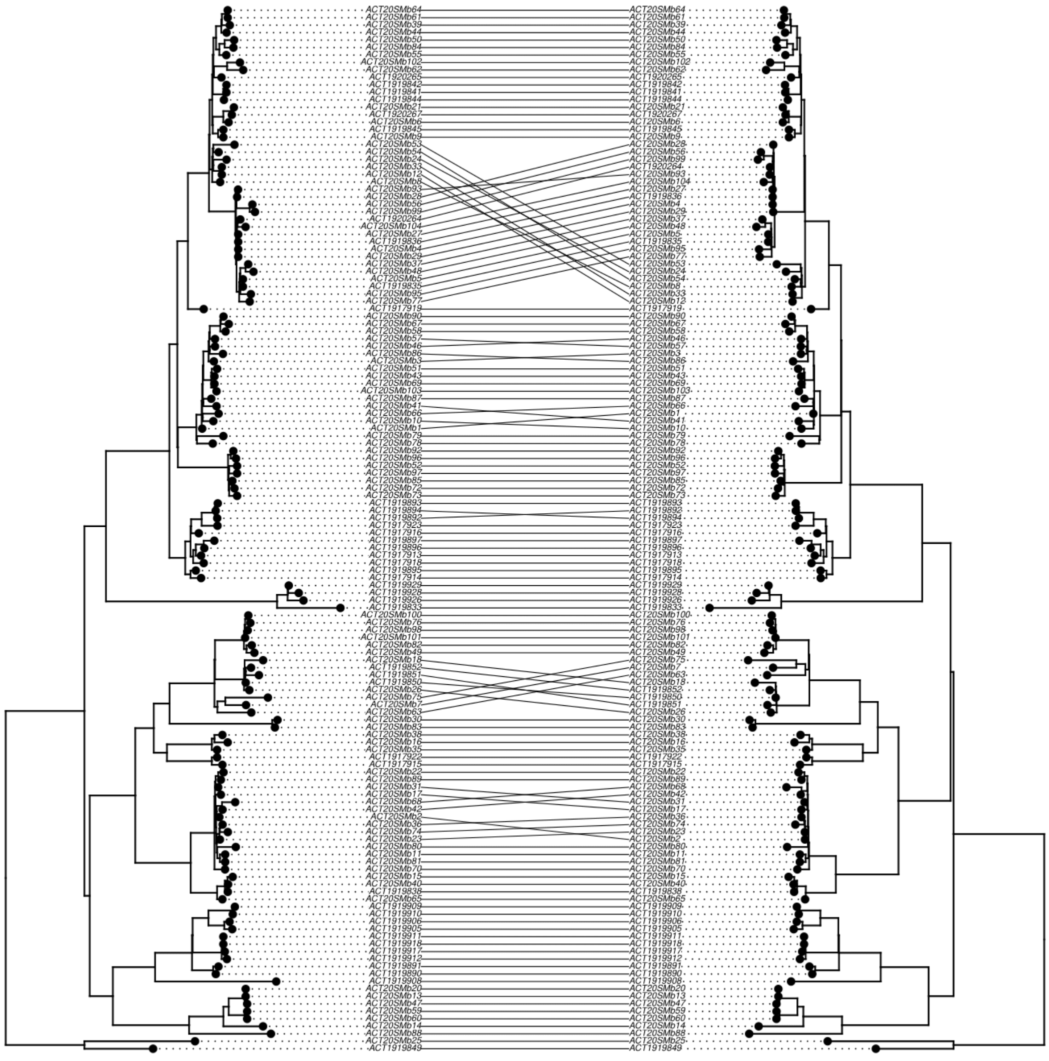


Figure 7.17: Impact of homologous recombination events on phylogenetic topology of bovine dataset. Left: Phylogenetic tree with recombination events. Right: Phylogenetic tree with recombination events detected by ClonalFrameML and excluded. RF=64

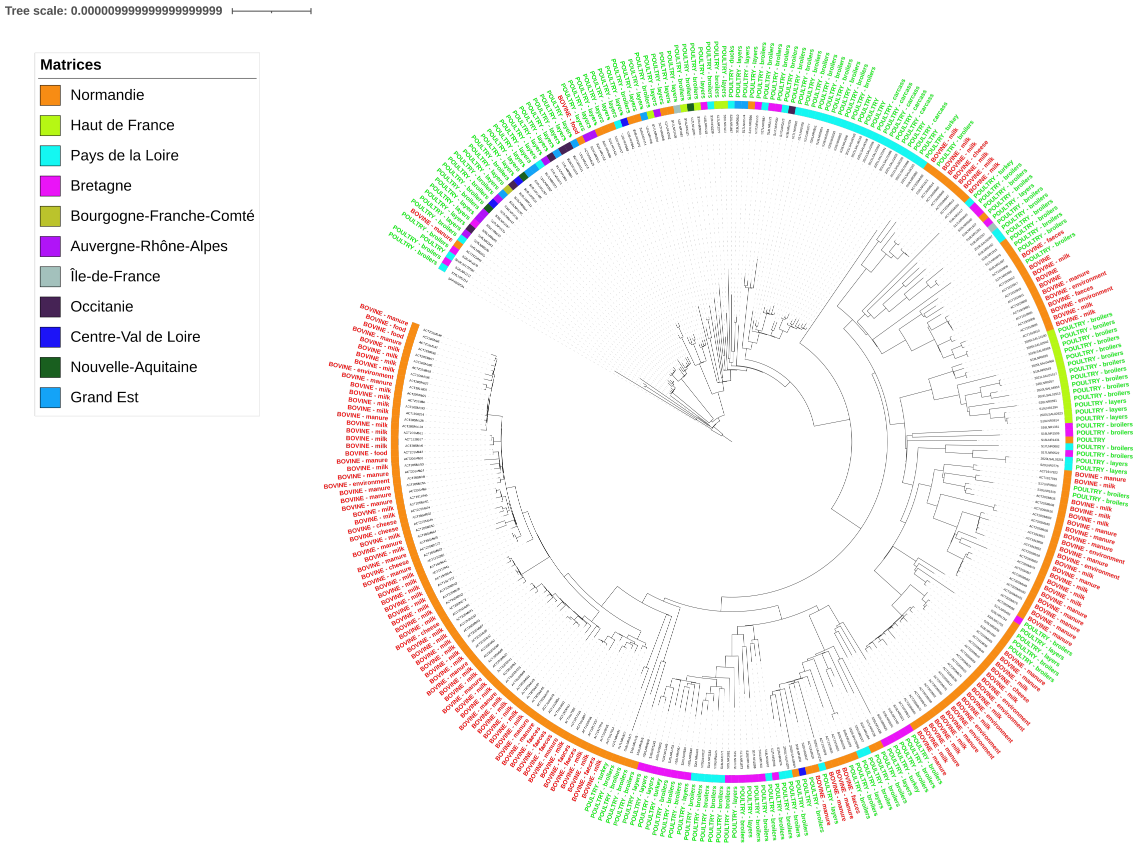


Figure 7.18: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Inner ring corresponds to the region of isolation. Outer ring corresponds to the precise matrix of isolation.

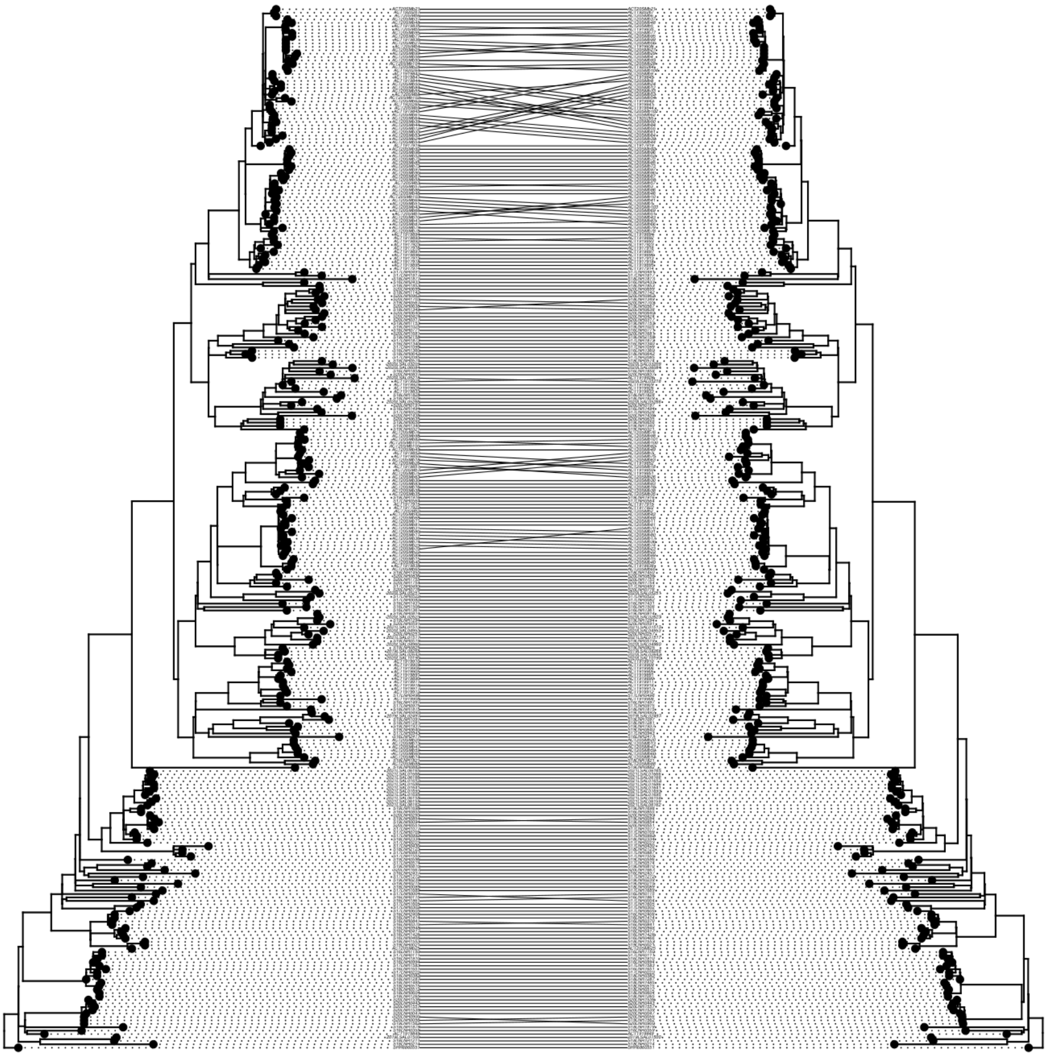


Figure 7.19: Impact of homologous recombination events on phylogenetic topology of bovine and poultry. Left: Phylogenetic tree with recombination events. Right: Phylogenetic tree with recombination events detected by ClonalFrameML and excluded. RF=134

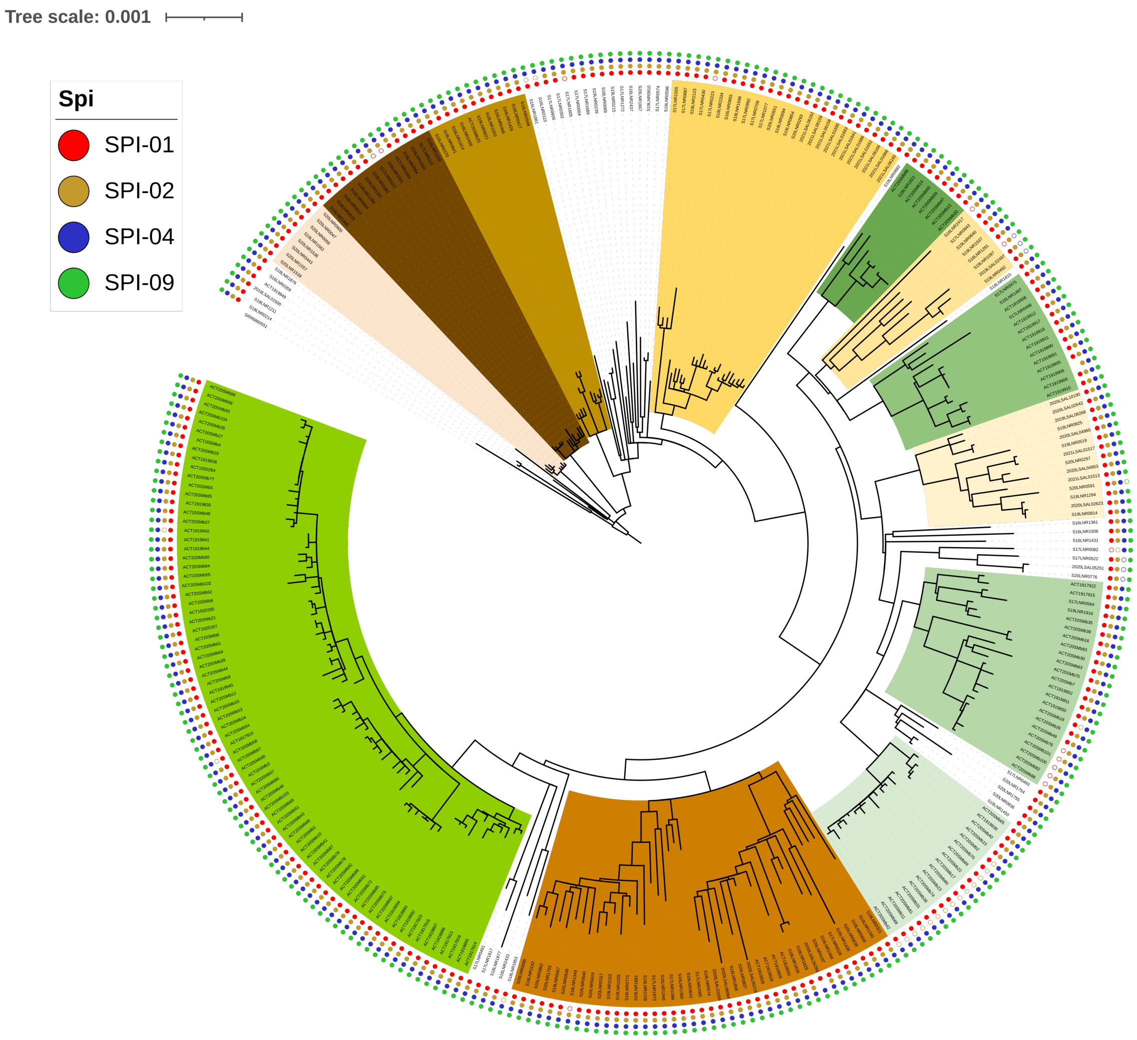


Figure 7.20: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Outer ring corresponds to the presence of different SPI in genomes

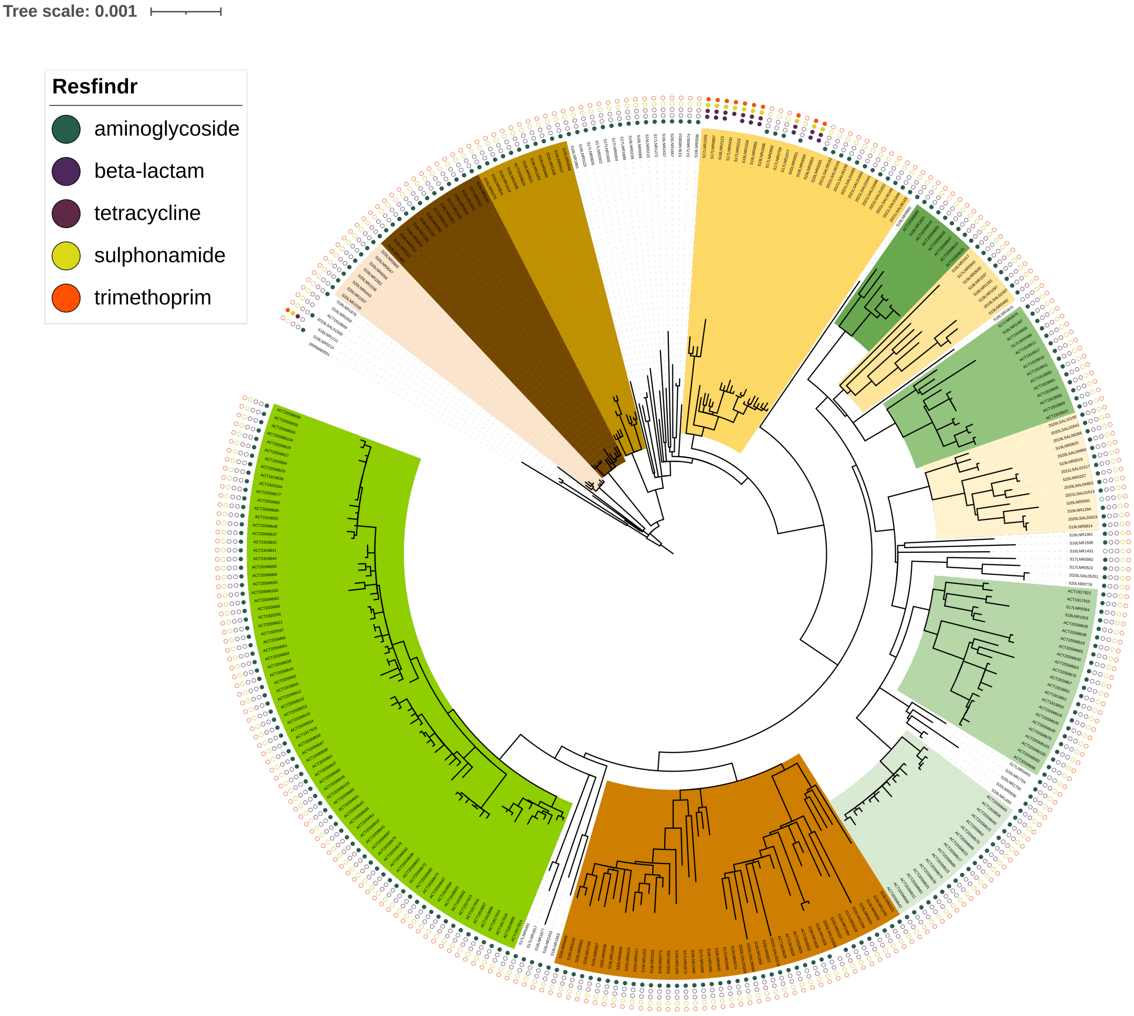


Figure 7.21: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Outer ring corresponds to the presence of different antimicrobial resistance genes found by ResFindr in genomes

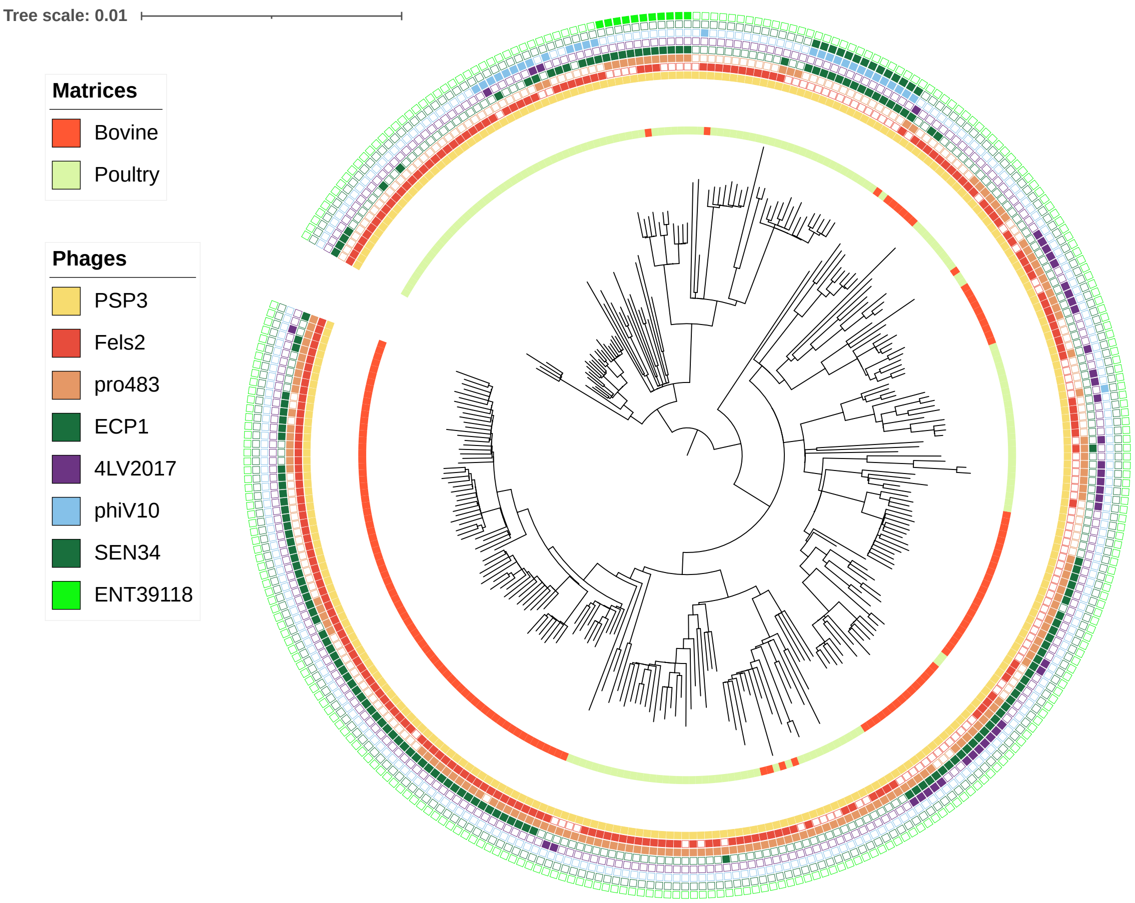


Figure 7.22: Coregenome SNP-based phylogenomic reconstruction by Maximum Likelihood of *Salmonella* Mbandaka isolated from bovine and poultry. Inner ring corresponds to the matrix of isolation. Outer rings correspond to the presence of phage.

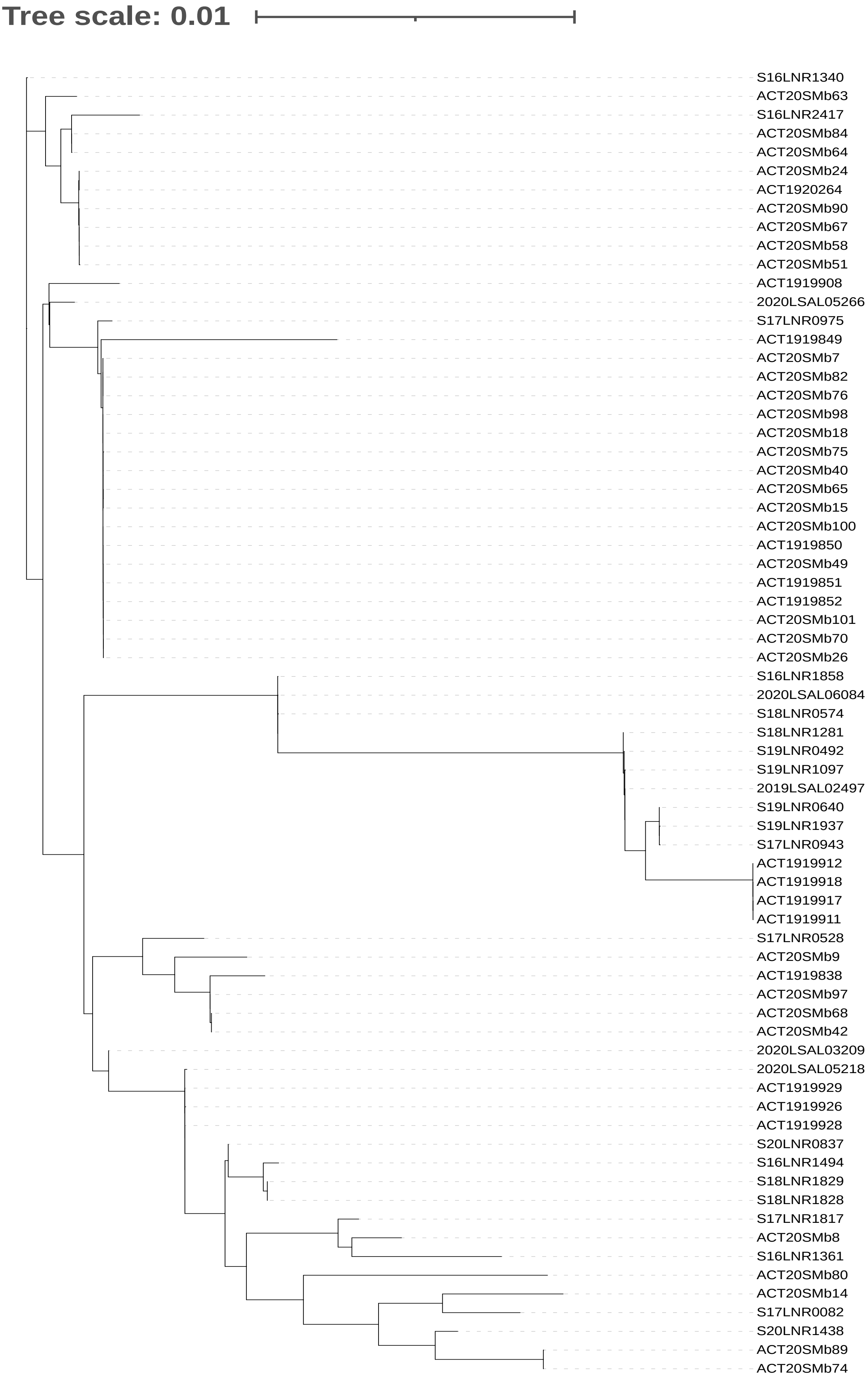
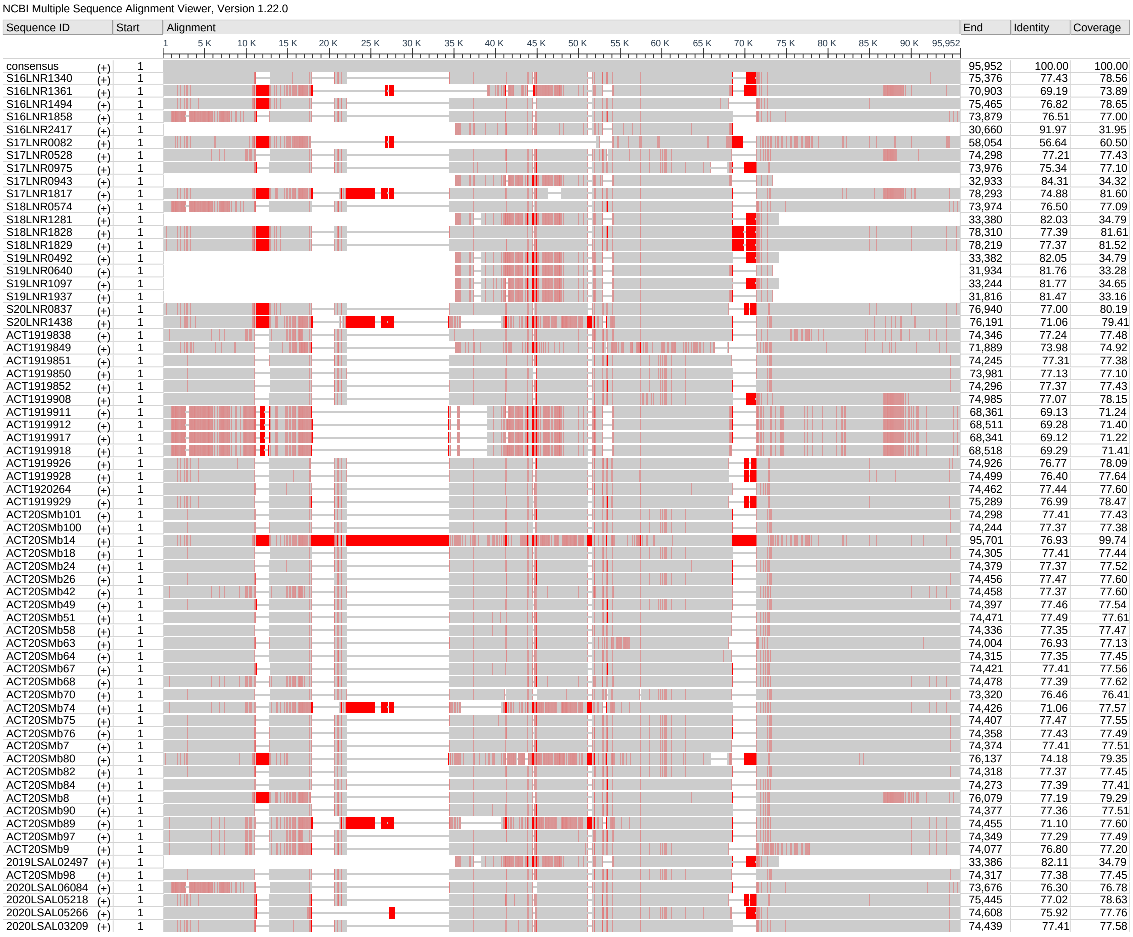


Figure 7.23: plasmid p12-4374 subtree from pgSNP analysis



NCBI Multiple Sequence Alignment Viewer, Version 1.22.0

Sequence ID		Start	Alignment	End	Identity	Coverage
consensus	(+)	1		95,952	100.00	100.00
S16LNR1340	(+)	1		75,376	77.43	78.56
S16LNR1361	(+)	1		70,903	69.19	73.89
S16LNR1494	(+)	1		75,465	76.82	78.65
S16LNR1858	(+)	1		73,879	76.51	77.00
S16LNR2417	(+)	1		30,660	91.97	31.95
S17LNR0082	(+)	1		58,054	56.64	60.50
S17LNR0528	(+)	1		74,298	77.21	77.43
S17LNR0975	(+)	1		73,976	75.34	77.10
S17LNR0943	(+)	1		32,933	84.31	34.32
S17LNR1817	(+)	1		78,293	74.88	81.60
S18LNR0574	(+)	1		73,974	76.50	77.09
S18LNR1281	(+)	1		33,380	82.03	34.79
S18LNR1828	(+)	1		78,310	77.39	81.61
S18LNR1829	(+)	1		78,219	77.37	81.52
S19LNR0492	(+)	1		33,382	82.05	34.79
S19LNR0640	(+)	1		31,934	81.76	33.28
S19LNR1097	(+)	1		33,244	81.77	34.65
S19LNR1937	(+)	1		31,816	81.47	33.16
S20LNR0837	(+)	1		76,940	77.00	80.19
S20LNR1438	(+)	1		76,191	71.06	79.41
ACT1919838	(+)	1		74,346	77.24	77.48
ACT1919849	(+)	1		71,889	73.98	74.92
ACT1919851	(+)	1		74,245	77.31	77.38
ACT1919850	(+)	1		73,981	77.13	77.10
ACT1919852	(+)	1		74,296	77.37	77.43
ACT1919908	(+)	1		74,985	77.07	78.15
ACT1919911	(+)	1		68,361	69.13	71.24
ACT1919912	(+)	1		68,511	69.28	71.40
ACT1919917	(+)	1		68,341	69.12	71.22
ACT1919918	(+)	1		68,518	69.29	71.41
ACT1919926	(+)	1		74,926	76.77	78.09
ACT1919928	(+)	1		74,499	76.40	77.64
ACT1920264	(+)	1		74,462	77.44	77.60
ACT1919929	(+)	1		75,289	76.99	78.47
ACT20SMb101	(+)	1		74,298	77.41	77.43
ACT20SMb100	(+)	1		74,244	77.37	77.38
ACT20SMb14	(+)	1		95,701	76.93	99.74
ACT20SMb18	(+)	1		74,305	77.41	77.44
ACT20SMb24	(+)	1		74,379	77.37	77.52
ACT20SMb26	(+)	1		74,456	77.47	77.60
ACT20SMb42	(+)	1		74,458	77.37	77.60
ACT20SMb49	(+)	1		74,397	77.46	77.54
ACT20SMb51	(+)	1		74,471	77.49	77.61
ACT20SMb58	(+)	1		74,336	77.35	77.47
ACT20SMb63	(+)	1		74,004	76.93	77.13
ACT20SMb64	(+)	1		74,315	77.35	77.45
ACT20SMb67	(+)	1		74,421	77.41	77.56
ACT20SMb68	(+)	1		74,478	77.39	77.62
ACT20SMb70	(+)	1		73,320	76.46	76.41
ACT20SMb74	(+)	1		74,426	71.06	77.57
ACT20SMb75	(+)	1		74,407	77.47	77.55
ACT20SMb76	(+)	1		74,358	77.43	77.49
ACT20SMb7	(+)	1		74,374	77.41	77.51
ACT20SMb80	(+)	1		76,137	74.18	79.35
ACT20SMb82	(+)	1		74,318	77.37	77.45
ACT20SMb84	(+)	1		74,273	77.39	77.41
ACT20SMb8	(+)	1		76,079	77.19	79.29
ACT20SMb90	(+)	1		74,377	77.36	77.51
ACT20SMb89	(+)	1		74,455	71.10	77.60
ACT20SMb97	(+)	1		74,349	77.29	77.49
ACT20SMb9	(+)	1		74,077	76.80	77.20
2019LSAL02497	(+)	1		33,386	82.11	34.79
ACT20SMb98	(+)	1		74,317	77.38	77.45
2020LSAL06084	(+)	1		73,676	76.30	76.78
2020LSAL05218	(+)	1		75,445	77.02	78.63
2020LSAL05266	(+)	1		74,608	75.92	77.76
2020LSAL03209	(+)	1		74,439	77.41	77.58

Figure 7.24: plasmid p12-4374 alignment on 67 genomes from pgSNP analysis

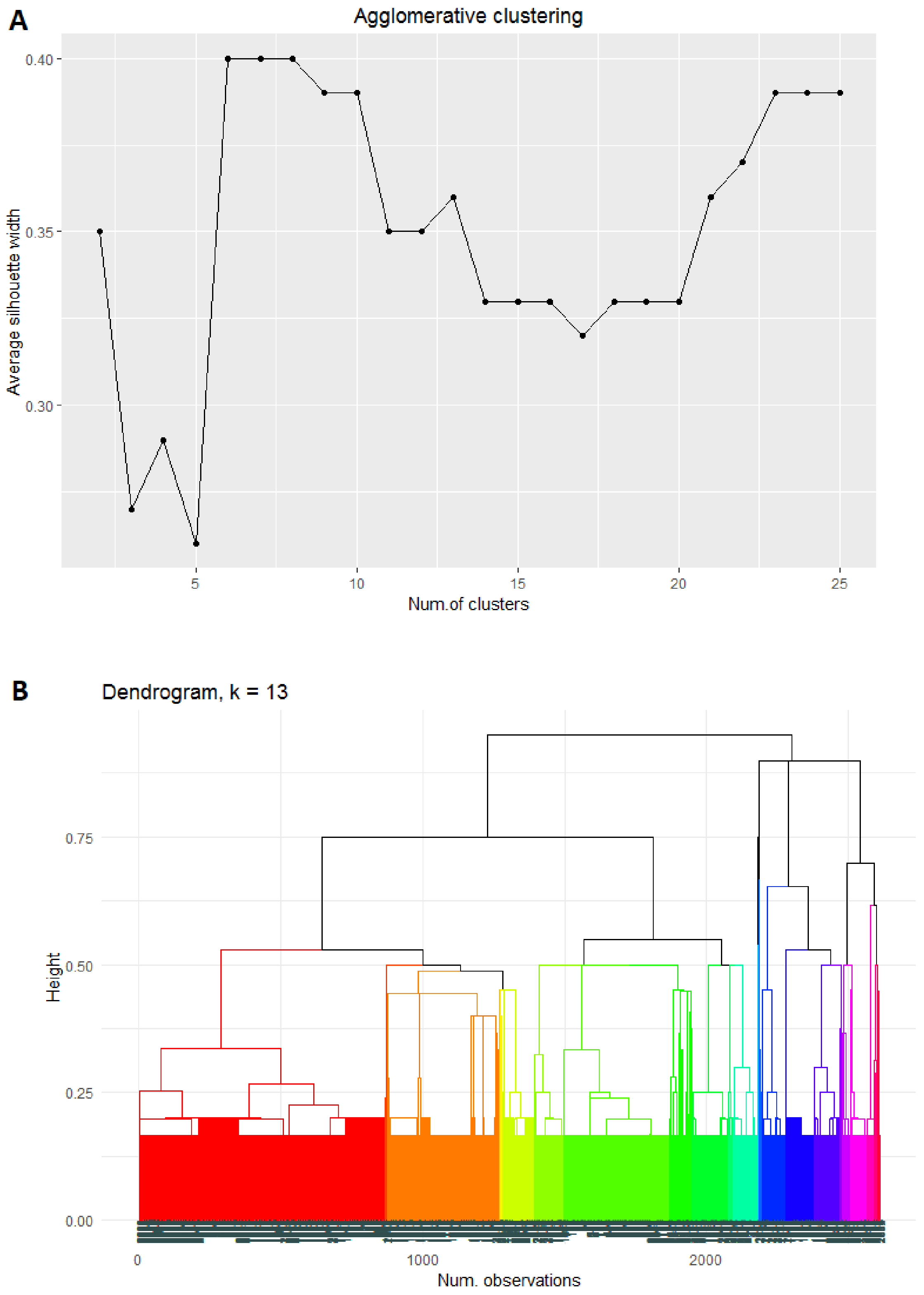


Figure 7.25: *Salmonella* Dublin Gower's results of the 398 randomly selected panel. A : Gower's distance agglomerative clustering. X axis represents the number of clusters, Y axis represents the average silhouette width. B : Dendrogram plot of Gower's distance for cluster 30. Samples are coloured by clusters.

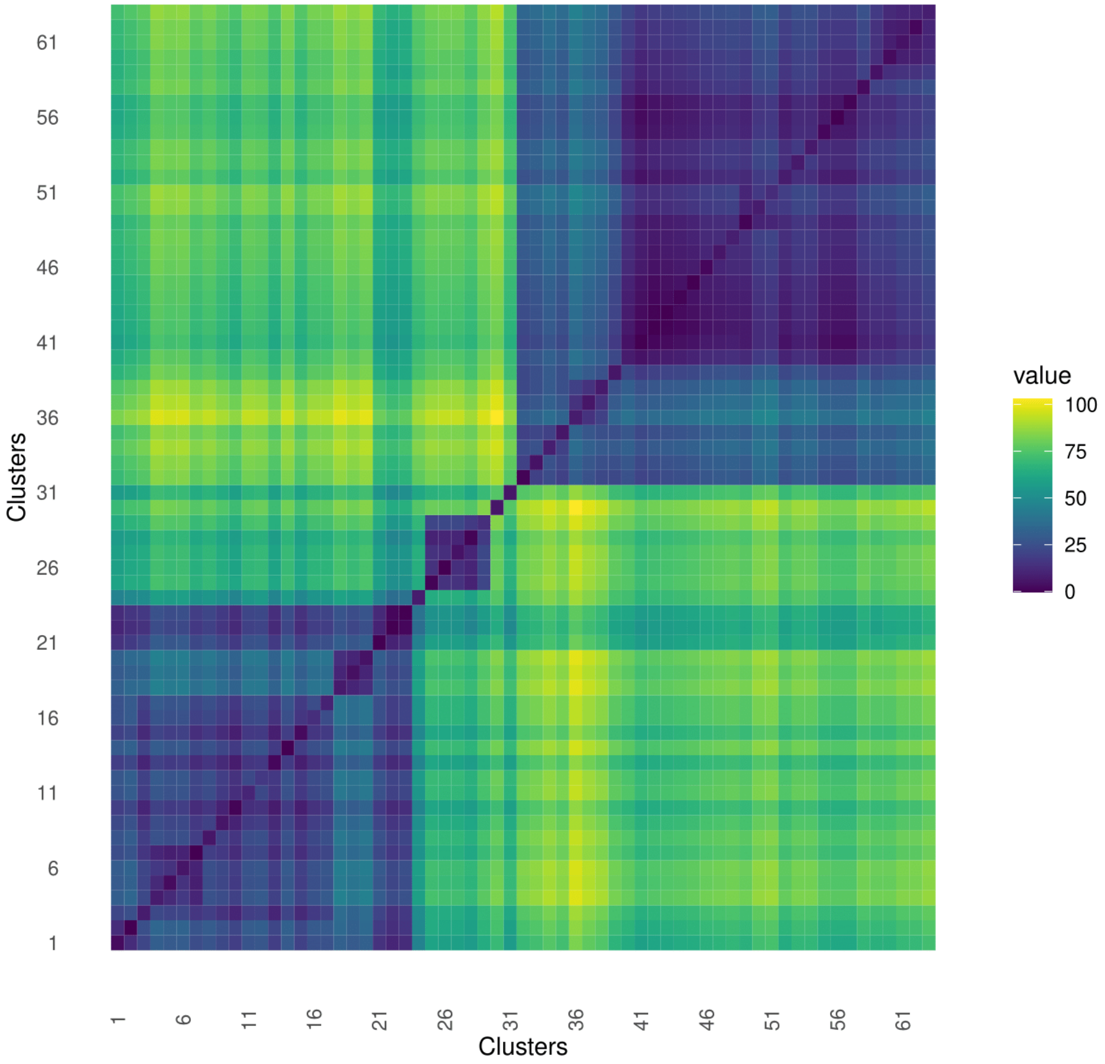


Figure 7.26: *Salmonella* Dublin heatmap of intra and inter distance of clusters selected by rPinecone. Distances are represented in SNP, from low value (blue) to high value (yellow).

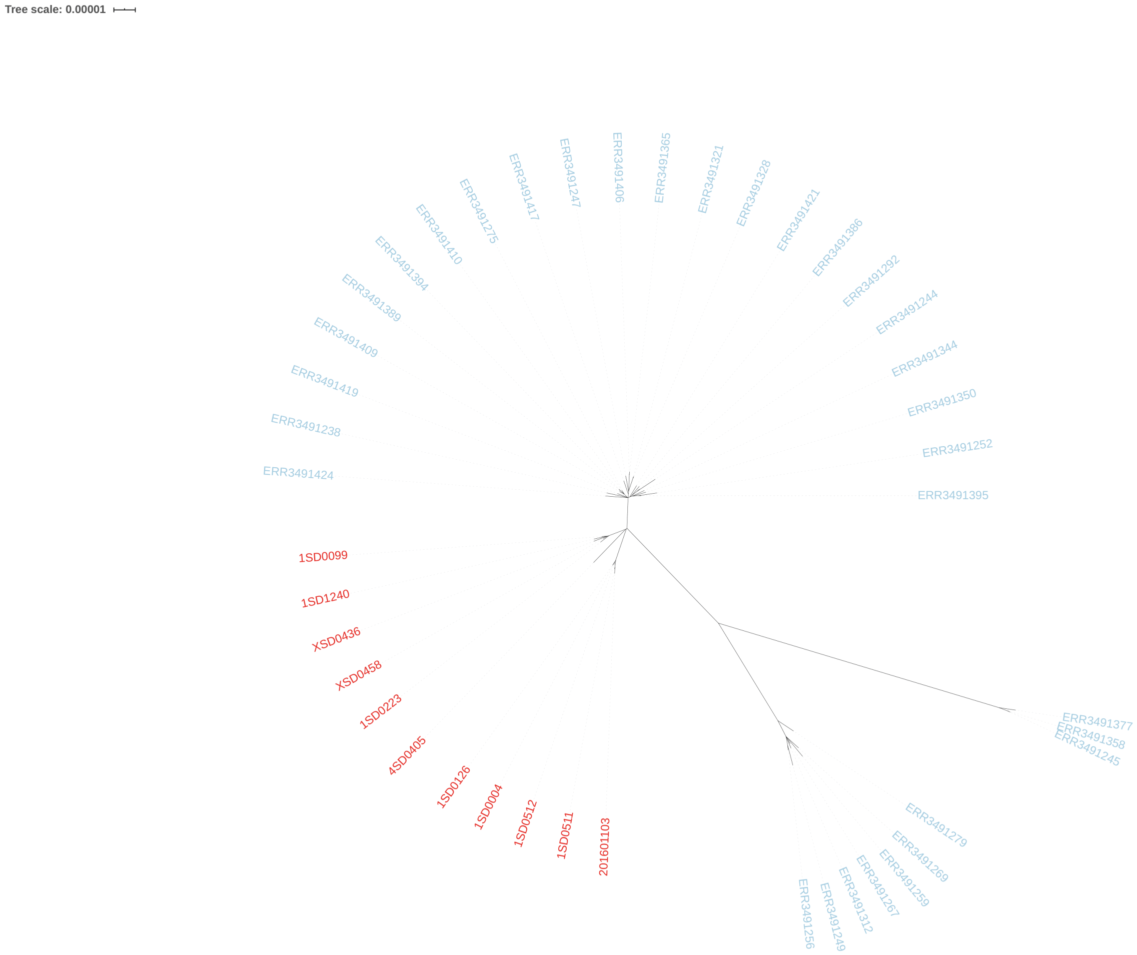


Figure 7.27: Comparison of 43 *S.* Dublin samples from France (in blue) and Denmark (in red). Phylogenetic tree is made by IQTREE with an evolutionary model K3Pu+F+I model and an optimal log-likelihood of -6728504.9209.

M2BI - Long tutored project
Supervisor: Madeleine DE SOUSA VIOLANTE
Student: Valentin BALOCHE
Notebook: https://github.com/valentinbaloche/long_project

Graph construction of SNPs data

Salmonella is one of the most prevalent bacterial pathogens in humans and animals worldwide, causing 87,923 cases of gastroenteritis in Europe in 2019 [1]. In France, Salmonella is the main pathogen confirmed in foodborne outbreaks (FBOs). During an outbreak, it is crucial for public health and regulatory agencies to have rapid, accurate, and discriminatory genomic methods to detect outbreaks and link disease cases to the source of contamination. The aim of this project is therefore to develop a method which could complete the classical phylogenetic tree construction to differentiate strains in a visual and accurate way. It is based on a visual representation of graphs constructed with strain's SNPs.

Introduction

Salmonella is among the most common foodborne pathogens worldwide, and can lead to acute gastroenteritis. The outbreaks involving this pathogen must be quickly identified. For this purpose, genome matching methods based on whole genome sequencing have been developed. The best known methods use phylogenetic trees based on the comparison of single nucleotide polymorphisms (SNPs). Today, methods that use this process involve three steps:

1. Mapping of reads on the reference genome (BWA, Bowtie, etc.)
2. Search for SNPs (GATK, Freebayes, etc.)
3. Use of these SNPs to infer a phylogenetic tree (IQTREE, RaXML, phyML, etc.)

In epidemic case studies, relationships between strains are sought. Thus, it is not necessary to trace the entire evolutionary history of the strains. Moreover, this method takes a considerable amount of computing time. The goal of this project is therefore to replace the 3rd step by the development of a graph-based method, which would save computation time and preserve the information provided by SNPs.

The main part of the development will use supervisor's data with epidemiological links. These datasets consist of concatenated SNPs isolated from different strains of S. Typhimurium or its monophasic variant S. 1,4,5,12:i:-. The term 'monophasic' characterizes the position of the flagella present on only one side of the bacterium, as opposed to the other strains which have flagella on two poles.

The second part of the study will focus on the method's effectiveness, in particular by carrying out tests on other datasets, including one with no related annotated strains.

Material and methods

Data

The dataset consists of 5 alignment files in fasta format and 2 metadata files allowing the correspondence between the IDs of the aligned strains and their type.

The first part of the project, which aims to develop and compare approaches for representing SNPs as graphs, uses the alignment files:

- 'first_alignment.fasta' (180 sequences of length 5333),
- 'SNPs_alignment_only_monophasic_variant.fasta' (123 sequences of length 826),
- 'SNPs_alignment_only_typhimurium.fasta' (57 sequences of length 4480)

(the first one being an alignment of the sequences composing the two lasts) as well as the 'metadata_first_alignment.ods'.

The second part of the project aims at evaluating the performance of the graph algorithm on more sequences of a different bacterial strain. It exploits a dataset from the literature [2] which includes 250 aligned E.coli sequences ('ecoli_SNPs.fasta') and its corresponding metadata file 'ecoli_metadata.ods'.

Finally, the last part is an exploratory study performed on an alignment of 42 Salmonella strains ('variant_mono_SNPs.fasta') predicted as monophasic variants but without additional information about their possible relatedness.

Distance matrix

The identity model was performed using the DistanceCalculator class imported from Bio.Phylo.TreeConstruction (execution time = 1 min 45). The distance matrix function was imported from scipy.spatial and used a numeric vector as input (execution time = 1.76 sec). In order to convert the DNA sequence into numeric vectors, the aligned sequences were first transferred into a dataframe with rows corresponding to the sequences and columns corresponding to each position of the sequence (execution time = 1.75 sec). After checking that each position could only take two possible forms (supplementary figure in the jupyter notebook), the majority letter was replaced by a '1' and the minority by a '0' for each position (execution time = 55 sec).

Graph generation

Graphs were generated using the Python NetworkX package [3]. The structure of the network object was generated from the distance matrices using the identifiers of the sequences as nodes and the distances as weights for the edges. Nodes were positioned using the Kamada-Kawai algorithm [4] implemented in NetworkX (execution time = 18.8 sec for the network using matrices generated with the identity method and 6.19 sec for the others). Node annotations were performed using both alignment (node name = sequence id) and metadata files when available (node color = sequence type). Finally, a gray scale was used to represent the edges as a function of the log(distance) separating the nodes, close sequences being connected by darker lines.

Visualization

Figure plots were created using Seaborn and Matplotlib packages, notably matplotlib.lines, matplotlib.colors and matplotlib.cm.

Results

Creation of the distance matrix

The first step in the elaboration of the graph is the calculation of a distance matrix between the sequences. This is a fundamental element since it will allow us to define the proximity between the points which compose the graph. The study of the distance's distribution can also help to define an objective threshold to identify the sequences belonging to the same group. I chose to perform a comparative study of 2 methods: 1) identity model proposed by biopython, 2) distance matrix proposed by scipy. The distance distributions generated by these two approaches are presented in Fig.1 and reveal important differences.

First, we can observe that the monophasic variants of S.Typhimurium (orange) are globally close to each other. We can distinguish 3 levels of proximity corresponding to the 3 peaks of the density curve. The distribution of the other S.Typhimurium (green) is much more heterogeneous and we can observe both genetically distant and close individuals.

The second important difference comes from the global aspect of density curves between both methods. While we observe for the 2nd method (Fig.1B) a distribution of distances for the total alignment (blue) that seems to be representative of the 'individual' ones (orange and green), it is clearly not the case for the 1st method (Fig.1A). For example, the distances for the monophasic variants that range from 0 to 0.12 in the 'individual' alignment are condensed around 0 for the 'total' one. This phenomenon can probably be explained by the fact that the identity method is based not only on the matches but also on the length of the sequence analyzed. Thus, if we are interested in the alignment of monophasic variants (length 826), 100 mismatches would correspond to an identity score of 88% (726/826). On the other hand, if we look at the same sequences in the 'total alignment' (length 5333), these same mismatches would correspond to an identity score of 98% (5233/5333).

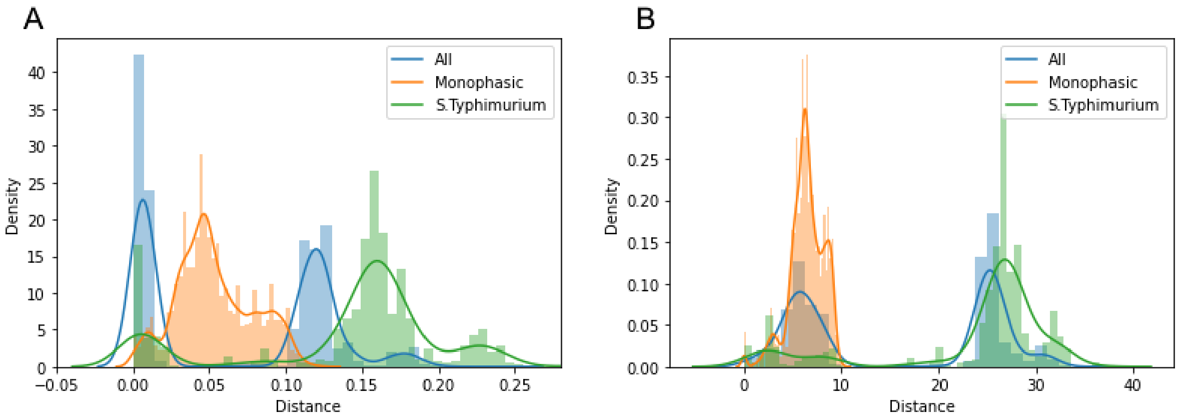


Figure 1. Comparative analysis of two distance matrix methods. *Density curves of the distances calculated by (A) identity model proposed by BioPython and (B) numeric distance matrix (Scipy.distance).*

It is important to take these differences into consideration because the large variation of the distances calculated via the identity method does not seem to be compatible with the definition of an objective and transferable threshold that allows identification of related strains (outbreaks).

Generation of the graphs

The generation of the graphs was performed using NetworkX, a Python package for the creation, manipulation and study of complex networks. It is based on the use of 'network objects' characterized by nodes, edges and potentially weights (here extracted from the distance matrices) that allow the calculation of spatial distances between nodes.

Several node positioning algorithms can be used depending on the network applications. By default, the nodes were positioned using the Fruchterman-Reingold force-directed algorithm (Fig.2A), which brings together subsets of densely connected nodes and separates different subsets from each other through repulsion (until equilibrium positions are obtained). Having a fully connected network, this algorithm wasn't relevant and generated an aesthetically-pleasing network in which all the edges were more or less equal length.

In order to generate a network that correctly represents the sequence similarities, I instead used the Kamada-Kawai algorithm for node positioning. It is also a force-directed layout method but it uses all distance values as input, and optimizes edge lengths with respect to inter-node distances [5]. Thus, it incorporates distance relationships between nodes (Fig.2B).

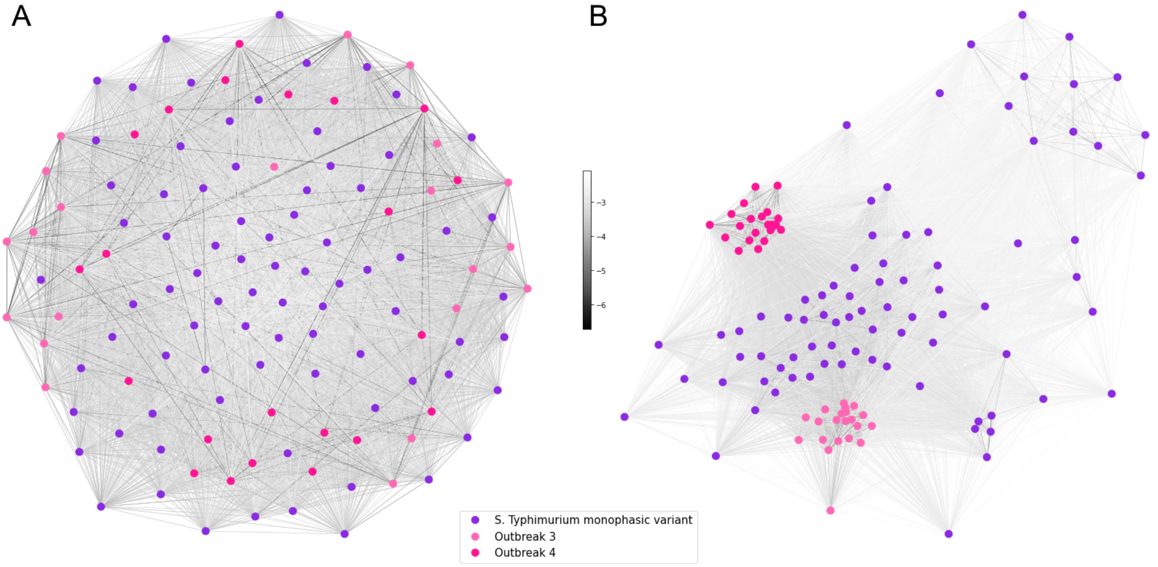


Figure 2. Graph generated using the NetworkX package. Node positions were determined using (A) Fruchterman-Reingold algorithm or (B) Kamada-Kawai algorithm. The log(distance) value is represented by a gray scale, the most similar nodes being connected by darker edges.

The networks presented in Fig.2 were constructed using the distance matrix generated by the identity method, and from the alignment of only monophasic variants (123 sequences of length 826). The positioning of nodes using the Kamada-Kawai algorithm spatially revealed the sequence proximity of samples from the same outbreaks. I therefore chose to apply the same graphical construction approach by exploiting the distance generated by Scipy as well as by exploiting the alignment of the other S.Typhimurium strains (57 sequences of length 4480) (Fig.3). In all these cases, the outbreaks appeared distinctly on the graphs. The main visible difference is a better definition of the outbreaks, which tend to form much tighter clusters, when using the identity method. However, we have seen before that with alignments mixing more heterogeneous sequences, this approach tends to bring similar sequences closer until a point where they are not really distinguishable (supplementary graph in the jupyter notebook).

Having in mind the goal of proposing an approach to identify strong related strains, transposable to different types of alignment, I chose to pursue the study using exclusively distances matrix generated with Scipy.

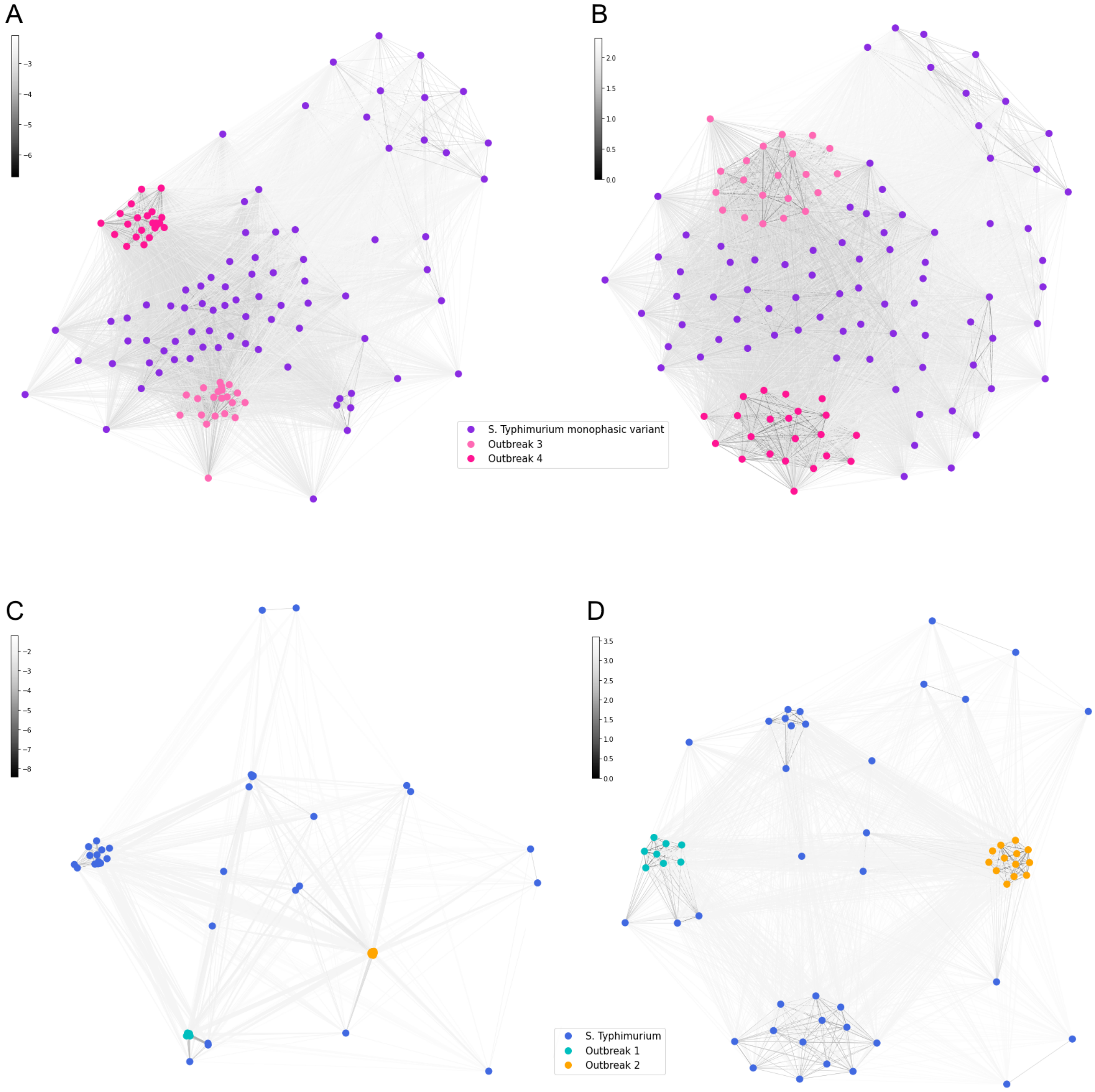


Figure 3. Comparison of graph architectures. The alignment files of both monophasic variants and other S.Typhimurium strains were used to generate graphs from (A and C) identity models and (B and D) Scipy's distance matrices.

Graph and phylogenetic tree comparison

In order to evaluate the quality of the graph construction method, I made a comparison with a phylogenetic tree made by my supervisor in her laboratory. In this type of visualization (Fig.4B), it is important to take into account the proximity of strains but especially the length of the branches. In this case, we can observe a long branch that leads to monophasic variants which are all very close genetically. Within this serovariant, there are two outbreaks: 3 and 4. This group is relatively distant from the other strains of S.Typhimurium. For these other strains, the distance between nodes is variable with some strains which are quite distant and others which form clusters. Among these clusters, we can find outbreaks 1 and

2. We can notice the presence of a monophasic variant in the outbreak 2 which was removed from the alignment files and which is therefore not present in the graph (Fig.4A).

Overall, we can observe that the graph faithfully reproduces the information present in the phylogenetic tree. Indeed, the group of monophasic variants is well defined and contains the outbreaks 3 and 4. The other strains of S.Typhimurium gravitate around with some of them organized in clusters, like the outbreaks 1 and 2.

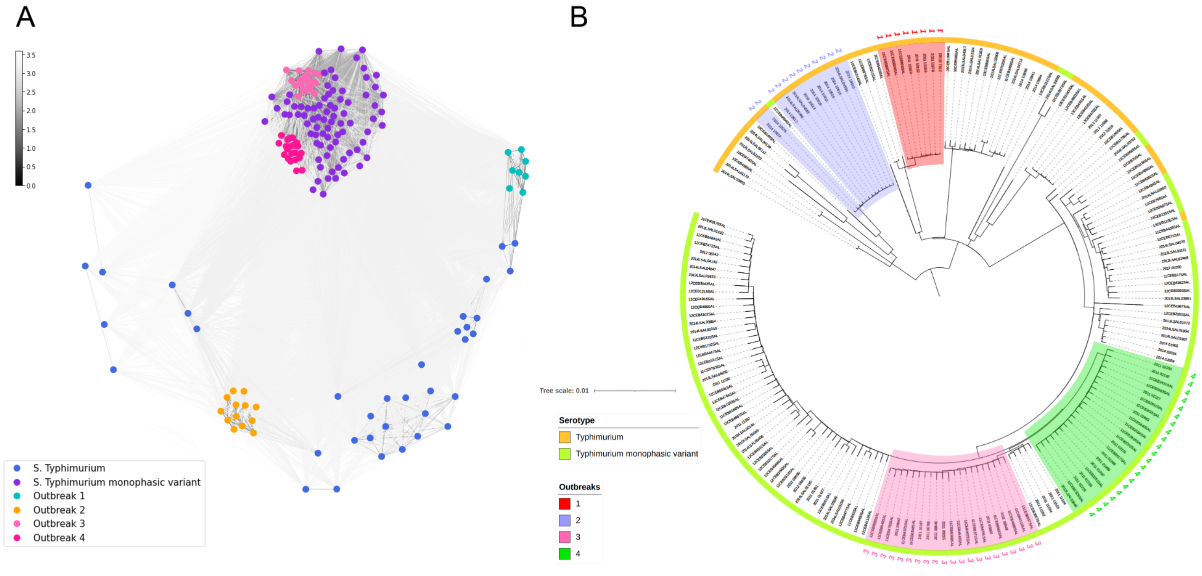


Figure 4. Graph construction method evaluation. Comparison of (A) the graph constructed from the alignment file of all the S.Typhimurium strains (monophasic or not) and (B) a phylogenetic tree constructed from the same alignment file.

Determination of a “relationship threshold”

In the context of an outbreak, it is important to identify the origin of the pathogen in order to stop its propagation. The goal is then to find a relationship between the pathogen and different collected samples. Currently, an arbitrary threshold is used to define this relationship, considering that under 5 SNPs the strains are strongly related together and come from the same place. In order to rationalize this approach, I measured the evolution of the maximum distance as a function of the alignment’s length, and according to whether or not the strains belong to the same outbreak.

As shown in Fig. 5A, we can see that the longer the alignment, the easier it is to identify strains from the same outbreak. Indeed, while the maximum distance between two strains from the same outbreak doesn’t evolve a lot, the maximum distance of unrelated strains increases almost linearly.

I used this result to define an objective distance threshold for an alignment of length 5333 which is approximately equal to 4.12. If we plot on the graph only the edges corresponding to a distance less than or equal to this threshold (Fig.5B), we can observe that all the strains constituting the outbreaks 1, 2 or 4 are interconnected, with connections limited to the group's members. However, if we look at the outbreak 3, even if the strains are all connected to each other, they are also connected to other monophasic variants. This can be partly explained by the fact that the threshold was defined using the maximum distance determined between members of an outbreak, all outbreaks combined. Furthermore, it is clear that the outbreak 3 is spatially very close to the other monophasic variants. With the method of distance calculation I used, it seems difficult to delimit this group.

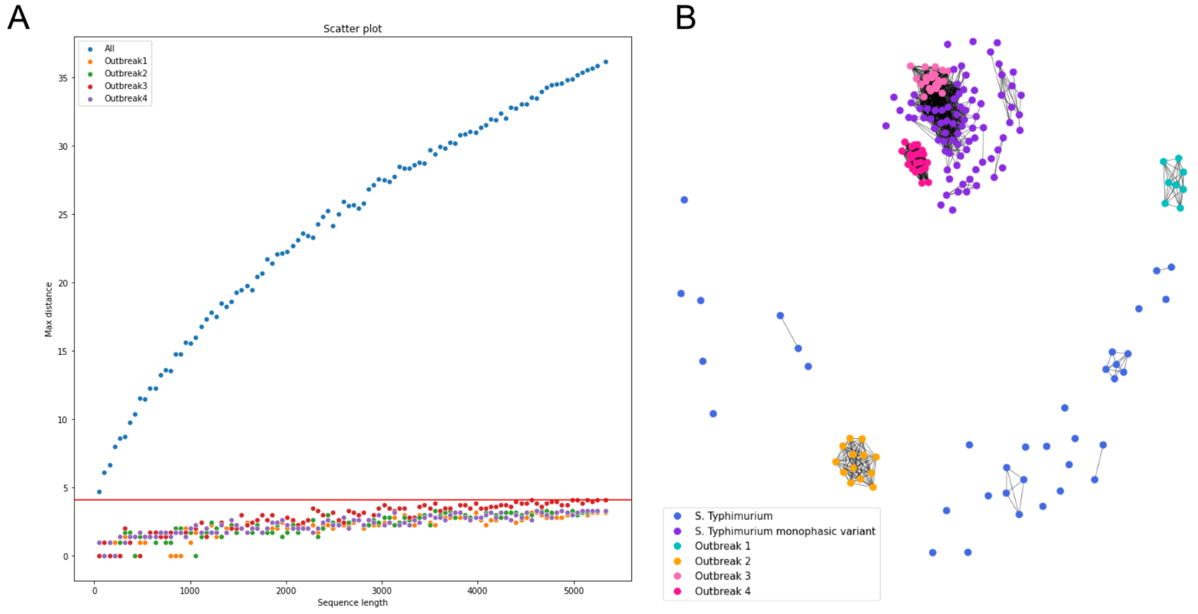


Figure 5. Study of the distances separating strains belonging to the same outbreak. (A) Sequences were generated by randomly picking positions in the alignment in order to study the average evolution of the maximum distance between two strains, depending on whether or not they belong to the same outbreak. (B) A graph was generated using the alignment file of all the S.Typhimurium strains, in which only distances less than or equal to 4.12 appear.

Testing of the construction method on a denser dataset

In order to further evaluate the effectiveness of the construction method, I decided to use the same approach on another dataset [2] that focuses on the outbreak detection of Verotoxigenic *Escherichia coli* (VTEC) O157:H7. The alignment file consisted of 250 sequences with a length of 2742. Based on the analysis shown in Fig.5A, I decided to represent edges with a value of 2.06 or less. The results obtained, presented in Fig.6, show that the method also works very well on sequences from another species. We can also see that the approach to determine an objective threshold to define species belonging to the same outbreak seems to be exploitable in several cases. However, the strains of the outbreaks 12 are not linked in this representation, which highlights once again some limitations which will be addressed in the discussion.

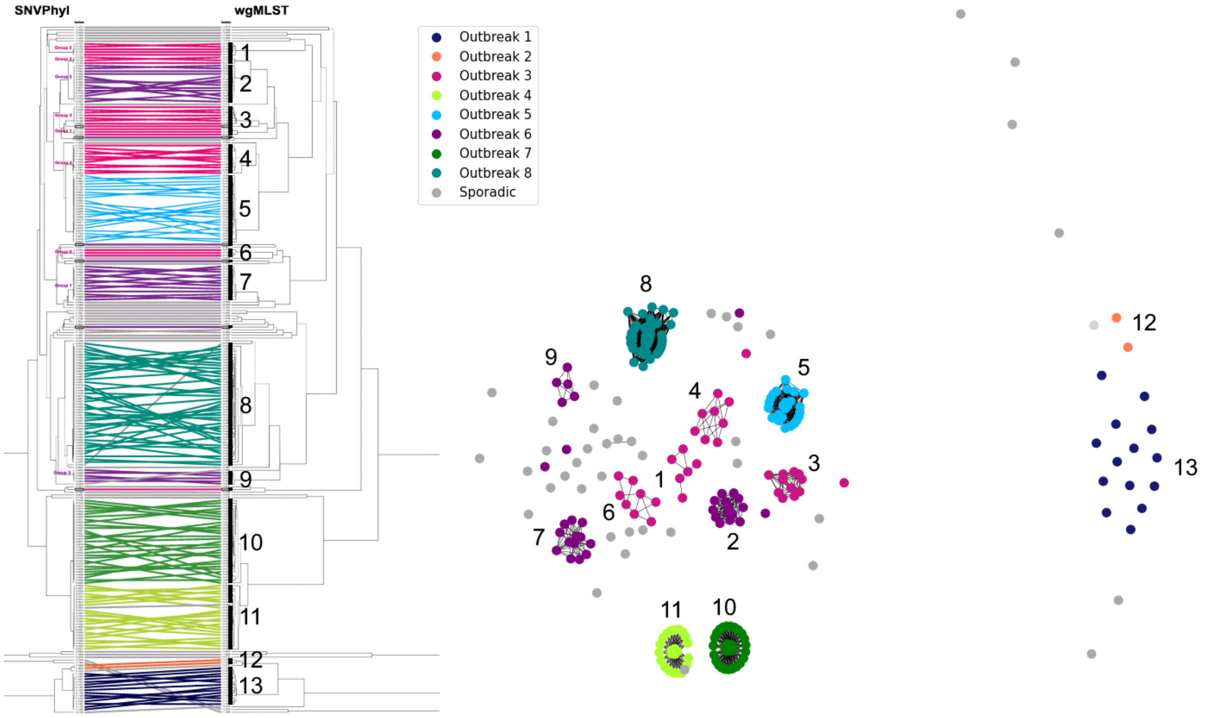


Figure 6. Graph construction method tested on a *E.coli* dataset.

Testing of the construction method on a “not-annotated” dataset

To conclude this project, I tested the construction algorithm on a dataset on which I had no information. The idea was to determine if these strains were related (coming from the same contamination site). The alignment file consisted of 42 sequences with a length of 704. Based on the analysis shown in Fig.5A, the maximum distance separating strains for an alignment of this size should have been less than 2.06. We can see that no strain in the alignment has a distance value less than or equal to this threshold (Fig.7A). This suggests that these samples all come from different sources. If we increase the threshold to 4.12 (Fig.7B) we can however notice that some strains are genetically closer than others, by noticing the presence of two clusters made of 4 interconnected strains each.

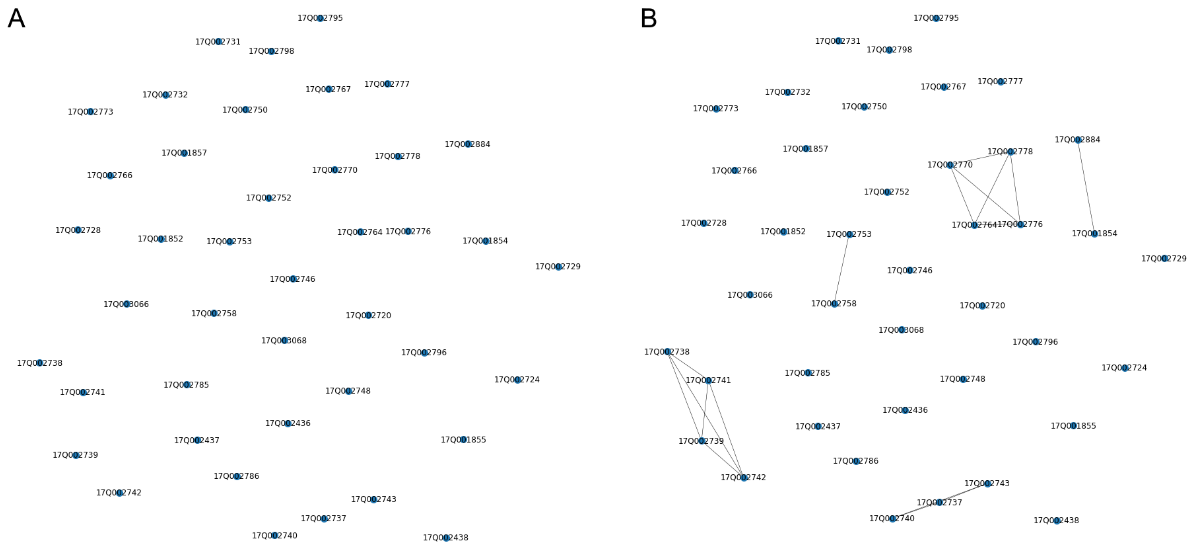


Figure 7. Graph construction method tested on a "non-annotated" dataset. The edges were represented on the graph when equal to or less than (A) 2.06 or (B) 4.12.

Discussion

The representation of SNPs alignments under a graph form has an important computational interest since it allows the visualization of genetic proximity much faster than a phylogenetic tree construction. Graphs are also much easier to understand, especially when starting to analyze large amounts of samples. This ease of analysis can facilitate communication between health authorities and actors of the food industry.

The method I developed is based on the generation of a distance matrix from aligned sequences converted into numerical vectors. We have seen that it allows to correctly transcribe the results obtained with a phylogenetic tree and that it can even allow to rationalize the determination of threshold to discriminate the strains belonging to the same outbreak. However, we have also seen that this approach does not work systematically. This is due to the fact that the distances separating the strains of the same cluster can vary. It might be possible to obtain more homogeneous values by using another method to calculate distances. Indeed, the approach proposed in this project doesn't take into account the nature of the nucleic acids. Better results could maybe have been obtained by distinguishing transitions from transversions, for example. In addition, some regions are probably more conserved than others and it might be interesting to influence the distance between two strains regarding the status of the region studied.

These questions could not be addressed in this project but could represent new working hypotheses for further studies.

References

[1] L. Bonifait, A. Thépault, L. Baugé, S. Rouxel, F. Le Gall, et M. Chemaly, « Occurrence of Salmonella in the Cattle Production in France », *Microorganisms*, vol. 9, n° 4, p. 872, avr. 2021, doi: 10.3390/microorganisms9040872.

[2] J. Rumore *et al.*, « Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective », *BMC Genomics*, vol. 19, n° 1, p. 870, déc. 2018, doi: 10.1186/s12864-018-5243-3.

[3] A. A. Hagberg, D. A. Schult, et P. J. Swart, « Exploring Network Structure, Dynamics, and Function using NetworkX », p. 5, 2008.

[4] T. Kamada et S. Kawai, « An algorithm for drawing general undirected graphs », *Information Processing Letters*, vol. 31, n° 1, p. 7-15, avr. 1989, doi: 10.1016/0020-0190(89)90102-6.

[5] A. de la Vega de León et J. Bajorath, « Design of chemical space networks incorporating compound distance relationships », *F1000Res*, vol. 5, p. Chem Inf Sci-2634, 2016, doi: 10.12688/f1000research.10021.2.

Résumé long de la thèse :

Cette thèse a été réalisée en vue d'obtenir le grade de docteur de l'Université Paris-Est Sup via l'école doctorale n°581 Agriculture, Alimentation, Biologie, Environnement et Santé (ABIES). Ce travail a été réalisé dans le cadre d'une thèse CIFRE (Conventions industrielles de formation) financée par ACTALIA et l'IFIP-Institut du porc, et a d'abord été accueilli à l'ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) au sein de la mission GAMeR (Genome Analysis Modelling and Risk), puis au laboratoire INRAE au sein de l'unité de recherche 1404 MaIAGE (Mathematics and Computer Science Applied to the Genome and the Environment) dans l'équipe StatInfOmics.
Cette thèse s'inscrit également dans le cadre du projet CasDAR-RT (Compte d'affection Spécial au Développement Agricole et Rural) n°1710 EMISSAGE (Epidémiologie des Salmonelles dans le secteur animal par une approche génomique), supervisée par l'UMT ASIICS (Action pour la Surveillance, l'Investigation et l'Intervention dans les Crises Sanitaires), dont ACTALIA est le coordinateur. L'objectif du CasDAR-RT était d'améliorer la surveillance et la caractérisation des *Salmonella* dans différents secteurs alimentaires.

Salmonella est une bactérie pathogène majeure au niveau mondial, hautement polymorphe dans sa diversité d'hôtes et de manifestations cliniques. Son impact sur la santé publique et sa charge économique ont continuellement motivé les efforts pour comprendre la situation épidémiologique ou réduire sa dissémination, historiquement en exploitant les méthodes de typage les plus appropriées disponibles. Mais malgré ces avancées, en 2020, *Salmonella* est le deuxième agent bactérien responsable d'intoxication alimentaire en Europe avec plus de 52 000 cas.

Pour lutter contre ces épidémies, la France a développé des systèmes de réponse par différents acteurs de la sécurité sanitaire. Les TIAC (Toxi-infections alimentaires collectives) et les cas groupés de *Salmonella* sont détectés sur la base du système de déclaration obligatoire (DO), et du système parallèle de surveillance du CNR (Centre National de Référence) (environ 2/3 des échantillons de *Salmonella* détectés chez l'homme l'ont été au CNR). En cas de toxi-infection alimentaire, Santé publique France (SpF) décide d'investiguer ou non en fonction du contexte épidémiologique, en lien avec la MUS (Mission des urgences sanitaires) de la DGAl (Direction Générale de l'Alimentation). SpF contacte également l'ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) et le réseau *Salmonella* pour rechercher une éventuelle contamination alimentaire comme origine de la TIAC. Si certains isolats alimentaires sont suspectés d'être liés à la TIAC, SpF centralise les souches, et les séquence pour vérifier leurs liens. En parallèle, SpF mène une enquête pour identifier la source de contamination la plus probable. Lorsque le produit alimentaire contaminé est identifié, il est retiré du marché, ou rappelé s'il a déjà été vendu. Ensuite, SpF rapporte l'enquête aux acteurs de terrain avec la DDPP (Direction départementale de la protection des populations) pour rechercher la cause de la contamination et les fournisseurs du produit incriminé.

À ces problématiques, s'ajoute la complexité du genre *Salmonella*. Les méthodes moléculaires basées sur les séquences des gènes de l'ARN 16S ont montré que le genre *Salmonella* est constitué de deux espèces, *S. enterica* et *S. bongori* (également appelée subsp. *V*). *Salmonella enterica* est divisée en six sous-espèces, dont *S. enteric*a subsp. *enterica* est la plus représentée, avec plus de 95 % des isolats de *Salmonella* obtenus chez les humains et les mammifères domestiques. *S. enterica* subsp. *enterica* est différenciée biochimiquement en sérovars sur la base de la composition de leurs structures glucidiques, flagellaires et lipopolysaccharides (LPS). Tous les sérovars de *Salmonella* peuvent être désignés par une formule antigénique proposée par Kauffmann, basée sur les antigènes somatiques (O) et flagellaires (H) en plus des antigènes capsulaires (Vi). *Salmonella* comprend plus de 2 600 sérovars, qui diffèrent par leur adaptation à l'hôte et leur virulence. Certains sérovars sont spécifiques de l'hôte, ce qui signifie qu'ils ne peuvent causer des maladies que chez une seule espèce.

Pour garantir la fiabilité de la détection des risques alimentaires liés *Salmonella*, des méthodes microbiologiques ont été développées pour caractériser rapidement les agents pathogènes.

Traditionnellement, la détection des bactéries viables est effectuée en cultivant et en surveillant la croissance des microorganismes. Plusieurs dizaines de milieux bactériologiques couramment utilisés dans l'industrie alimentaire ont leurs propres objectifs de surveillance de la contamination microbiologique et/ou de détection des bactéries pathogènes, comme par exemple, l'utilisation de milieux de routine tels que la gélose trypticase-soja ou la gélose PCA (Plate Count Agar).

La caractérisation et le typage des *Salmonella* n'ont cessé de s'améliorer au fur et à mesure des avancées technologiques. Dans le cadre du contrôle de la sécurité des aliments ou des enquêtes sur les intoxications alimentaires, la caractérisation des sérovars ou des types de séquences de *Salmonella* est obligatoire pour l'attribution de la source. La méthode de caractérisation traditionnelle des sérovars de *Salmonella* repose sur une réaction immunologique par agglutination. Cette méthode reste coûteuse et nécessite des techniciens entraînés. D'autres méthodes moléculaires ont donc été développées, notamment sur la base de l'information génétique comme l'ADN ou les phages. Néanmoins, ces méthodes présentent toutes des limites, et se font dépasser actuellement par les nouvelles méthodes en génomique.

L'ère de la génomique a apporté une aide précieuse dans l'investigation et la caractérisation des bactéries pathogènes pour la santé publique. Depuis quelques années, des méthodes d'inférence phylogénomique commencent à être implémentées dans les agences en tant que méthodes de routines, notamment sur la base du MLST (Multilocus sequence typing) aux échelles du core (cgMLST) ou pangénome (wgMLST), ou encore des mutations ponctuelles (SNP : Single nucleotide polymorphisms) à l'échelle du coregénome (cgSNP). Malgré ces identifications, ces méthodes manquent de résolution pour des sérovars avec peu de variations génétiques, et ne prennent pas en compte tout le contenu génomique des souches (c.à.d. les zones intergéniques dans le cas du cg/wgMLST et les SNPs du génome accessoire dans le cas du cgSNP).

Ce projet de thèse a donc été conçu pour répondre aux limites des méthodes génomiques actuelles en prenant comme modèle d'étude *Salmonella* et transférer le pouvoir de résolution de la génomique à la compréhension des voies adaptatives de cette espèce dans un contexte de contrôle de la sécurité des aliments, comme par exemple le tropisme de l'hôte (nourriture, troupeau, contamination), la persistance, ou la résistance et à certains marqueurs discriminants. Dans cette thèse, l'accent a été mis sur trois sérovars majeurs de *Salmonella* dans deux filières alimentaires : la filière porcine et la filière laitière.

Dans la filière porcine, *Salmonella* Choleraesuis a été le premier sérovar de *Salmonella* isolé en 1884. La viande de porc est la source carnée majeure responsable de la transmission de *Salmonella* à l'homme. Les porcs peuvent être soit asymptotiques, ou soit présenter une forte réponse inflammatoire conduisant à une salmonellose et parfois jusqu'à la mort. Dans le secteur français de l'alimentation humaine, 37 sérovars différents ont été isolés en 2019 dans des aliments à base de porc, le variant monophasique du sérovar Typhimurium étant le plus répandu, suivi des sérovars Derby et Typhimurium, qui représentent ensemble 56,5% des sérovars détectés. Le variant monophasique de Typhimurium (TMV) est apparu pour la première fois en Europe au milieu des années 1990 et sa prévalence s'est considérablement accrue entre 2005 et 2008, où il est devenu l'un des trois principaux sérovars isolés en santé humaine, et ceci jusqu'à aujourd'hui. Ce sérovar est caractérisé par une forte prévalence de la résistance à l'ampicilline, à la streptomycine, aux sulfamides et à la tétracycline, ce qui représente un problème majeur de santé publique et qui explique les réglementations qui ont été mises en place dans le cadre de la surveillance de ce sérovar.

En 2019, dans le domaine de la santé et de la production animale, les secteurs les plus touchés par *Salmonella* étaient les secteurs avicole et bovin, et les sérovars les plus fréquemment isolés dans les secteurs bovins étaient Dublin, Montevideo, Typhimurium et Mbandaka. Ces sérovars sont également très prévalents dans d'autres pays dans la même filière. Au-delà des pertes économiques considérables, le lait cru ou les produits finis contaminés par des vaches porteuses peuvent provoquer des infections graves chez les vaches laitières. Bien que la diarrhée soit une conséquence courante des infections à *Salmonella* chez les bovins, les conséquences d'autres sérovars comme *S.* Dublin sont souvent des syndromes respiratoires chez les veaux ou des avortements chez les vaches gravides.

Salmonella Mbandaka a été isolée pour la première fois d'une salmonellose humaine au Congo belge en 1948. Contrairement à *Salmonella* Dublin, le sérovar le plus prévalent dans les salmonelloses bovines, *Salmonella* Mbandaka provoque un portage et une excrétion fécale asymptomatiques, mais peut être mortelle dans certains cas. En France, *Salmonella* Dublin est le sérovar le plus répandu, mais il existe une forte prévalence de *Salmonella* Mbandaka dans le Nord de la France.

Dans cette thèse, l'accent a été mis sur trois sérovars de *Salmonella* prévalents dans les secteurs alimentaires laitiers et porcins : *Salmonella* Mbandaka, *Salmonella* Dublin, *Salmonella* Typhimurium et son variant monophasique.

Dans le secteur porcin français, la dissémination de *Salmonella* Typhimurium et de son variant monophasique n'est pas encore clairement comprise, notamment en ce qui concerne la prédominance du TMV sur *Salmonella* Typhimurium. Même si certaines études ont été réalisées à l'échelle européenne, il n'existe aucune étude sur la diversité géographique de ces sérovars en France chez un hôte spécifique. La persistance de ces souches dans les élevages porcins et les abattoirs n'a pas été expliquée, et nous ne savons pas si des adaptations locales l'expliquent. Il reste encore des recherches à mener en génomique pour comprendre la raison de leur contamination des lignes agroalimentaires malgré les niveaux sanitaires de sécurité élevés actuels.

En ce qui concerne le secteur laitier, *Salmonella* Mbandaka n'a jamais été étudiée à l'échelle génomique en raison de la rareté des foyers chez l'homme, mais elle reste très présente et persistante dans les élevages bovins, notamment dans le nord-ouest de la France, sans que l'on ait de connaissances sur cette localisation géographique spécifique. Certaines hypothèses concernant cette persistance visent les produits alimentaires ou la contamination de l'environnement, mais aucune recherche n'a été effectuée à ces sujets. De plus, aucune investigation et approche génomique complète sur ce sérovar n'a été réalisée en France, et très peu d'investigations génomiques précises en relation avec sa persistance et sa diversité ont été réalisées dans d'autres pays.

Ce projet prévoit l'analyse holistique de trois sérovars qui sont récemment devenus des préoccupations notoires dans les chaînes alimentaires, y compris la caractérisation de leur diversité respective dans le pays, la comparaison de leurs génomes et la contextualisation avec la diversité mondiale à travers des données publiques internationales, et l'investigation de la raison de leur contamination des lignes agroalimentaires malgré la présence d'un système de surveillance. Également en parallèle, ce projet appelle le développement d'une méthodologie génomique permettant de répondre à ces problématiques.

En ce qui concerne l'analyse bioinformatique, une méthode de génomique comparative plus discriminante est nécessaire selon la littérature. Dans ce projet, nous visons à développer une approche sur l'ensemble des SNPs du pangénome (coregénome et génome accessoires) en incluant à la fois les régions codantes et non-codantes. Après avoir identifié ces nouvelles informations accessoires, il a fallu également utiliser les parties accessoires du génome pour augmenter le signal phylogénomique. L'objectif est de montrer l'importance du génome accessoire dans les enquêtes épidémiologiques.

En ce qui concerne le secteur porcin, l'accent a été mis sur la diversité géographique des souches de *Salmonella*. Des échantillons provenant de salles d'attente, de carcasses de porc et de locaux de découpe à l'abattoir seront analysés pour comprendre la diversité tout au long de la chaîne alimentaire. Une attention particulière sera accordée aux isolats provenant de troupeaux de porcs, car la contamination peut être disséminée des troupeaux aux abattoirs. L'analyse génomique se concentrera également sur l'AMR (antimicrobial resistance) multiple et étendue du TMV pour décrire cette prévalence. Enfin, la diversité génomique mondiale de la propagation du TMV sera évaluée, à partir de données brutes rendues publiques.

Pour le secteur laitier, nous nous sommes concentrés sur la caractérisation de la diversité génomique tout au long des étapes de la chaîne de production du fromage. Les échantillons prélevés dans les exploitations bovines (alimentation, eau, environnement, animaux), la chaîne de transport, les usines de production de lait et de fromage et les étapes de distribution ont été analysés. Afin d'établir une comparaison avec un autre hôte, la collection sera complétée par des souches provenant du secteur de la volaille. Ces résultats seront comparés aux résultats d'une étude sur *Salmonella* Dublin que j'ai également réalisée au cours de cette thèse, où la diversité a été caractérisée à une échelle géographique.

Dans une première partie du manuscrit, les résultats sur le développement méthodologique d'une nouvelle méthode bioinformatique sont présentés. La distinction des isolats au sein d'échantillons homogènes peut s'avérer difficile, notamment dans le cadre d'enquêtes sanitaires, où l'identification de l'origine de la contamination est essentielle. Pour améliorer la résolution des approches phylogénomiques, une nouvelle méthode a été développée, appelée "pgSNP". Le but du pgSNP est de prendre en compte toutes les informations pangénomiques, core (présent dans tous les échantillons) et accessoires (présent dans au moins un échantillon), codants et non-codants, afin d'intégrer l'ensemble de la variation génomique pour distinguer les isolats.

La stratégie d'analyse du pangénome que nous avons conçu est résumée en deux étapes principales : 1/ définir le pangénome de référence sur lequel nous pouvons comparer toutes les séquences présentes dans au moins quatre échantillons parmi tous ceux d'un ensemble et 2/ caractériser l'échantillon en utilisant des approches phylogénomiques sous un modèle évolutif. En résumé, nous collectons le contenu génomique présent dans tous les échantillons en un pangénome de référence (avec BLASTN), c'est-à-dire un répertoire de séquences uniques, dans lequel les éléments génomiques redondants sont fusionnés. Puis, ce pangénome de référence est ensuite utilisé pour les alignements et appels de variants (Snippy) pour chaque échantillon afin de reconstruire un arbre phylogénomique décrivant la phylogénie de l'échantillon. Pour aborder la reconstruction de l'arbre phylogénomique des échantillons qui ne partagent pas toutes leurs séquences en raison de l'inclusion de la partie accessoire, c'est-à-dire lorsqu'un échantillon ne possède pas de segment accessoire, nous avons recours à une approche en deux étapes : premièrement, nous générons plusieurs arbres (avec IQ-TREE), un pour chaque segment d'un ensemble homogène et continu d'échantillons. Ensuite, nous réconcilions les informations phylogénomiques de tous les arbres de segments en utilisant une méthode de super arbre (avec FastRFS). Nous obtenons à la fin un arbre dit pangénomique, qui représente un maximum d'informations identifiées dans un jeu de données.

Pour sélectionner les meilleurs outils du pipeline, une analyse comparative a été réalisée sur l'étape de l'appel des variants et sur la reconstruction de l'arbre pangénomique. Pour comparer quelles méthodes correspond le mieux à des critères de performances et de cohérence avec des données épidémiologiques, nous les avons comparées en utilisant la distance Robison-Foulds (RF) qui calcule la taille de la différence symétrique des splits entre deux arbres. Les outils ont été testés sur un jeu de données de 57 *S.* Typhimurium publié par Radomski, Cadel-Six et al. en 2019. Pour l'appel des variants, il a été montré dans cette thèse que les différences sont très faibles entre deux outils (GATK et Snippy) et ont très peu d'impact sur l'arbre phylogénomique, avec une faible distance RF entre les deux arbres de 30. N'ayant pas testé la version parallélisée de GATK (Spark – GATK), nous avons donc opté pour le variant caller Freebayes dont la version parallélisée est déjà implémentée dans container de Snippy aisément installable. Pour la sélection de la méthode de super arbre, trois méthodes (ASTRID, ASTRAL et FastRFS) ont été comparées à un arbre coregénome (iVARCall2). Il a été observé que la méthode ASTRID présente des différences de topologie plus importantes (RF = 308) par rapport aux deux autres méthodes. Alors qu'ASTRAL et fastRFS présentent les mêmes distances RF (ASTRAL : RF=228, fastRFS : RF=228), ASTRAL n'étant pas en mesure de retrouver des résultats concordant avec les données épidémiologiques du jeu de données, nous avons donc opté pour fastRFS.

Le pangénome de référence a été paramétré et évalué selon deux critères : la quantité d'informations obtenues lors de l'alignement des données de séquençage par rapport à des pipelines coregénome basées sur une seule référence. Le second est la qualité de l'alignement des données de séquençage par rapport à des pipelines coregénome basées sur une seule référence. En comparant différent pangénome de référence et une référence simple (ici, référence LT2), il a été montré qu'il y a une augmentation de 2% (= 3M de reads) de donnée brute de séquençage capable de s'aligner sur le pangénome de référence. La qualité de l'alignement des données de séquençage a été évaluée à l'aide du calcul de l'entropie de Shannon, qui est la mesure de l'incertitude. Dans notre contexte, elle permet de quantifier la variabilité de la séquence sur un site particulier. Elle est utilisée en génomique pour calculer la variabilité locale entre génomes, ou au sein d'un même génome en comparant tous les sites. Il a été montré qu'un pangénome de référence avec des paramètres optimisés avait une entropie plus basse (6.15e−11) en comparaison à celle d'une référence simple (6.41e−11), et donc une meilleure qualité d'alignement des données brutes de séquençage. En prenant également en compte la taille de l'alignement final du pipeline, les paramètres avec 95% d'identités et une taille minimum de 500 paires de bases d'un morceau d'ADN ont été sélectionnés pour la construction du pangénome de référence.

Pour vérifier et valider le pipeline pgSNP, 3 jeux de données avec des données épidémiologiques ont été sélectionnés pendant cette thèse.

Le premier jeu de donnés est un jeu de données de *S.* Typhimurium et son variant monophasique qui contient 192 souches avec 4 clusters épidémiologiques, publiées par Radomski, Cadel-Six et al. en 2019. Il a été tout d'abord montré sur ce jeu de données que pgSNP ajoute environ 1,6 Mb d'information génétique par rapport à l'analyse basée sur le coregénome. Par rapport aux clusters épidémiologiques, pgSNP est capable d'identifier et de retrouver groupées les souches entre elles, à l'exception de trois souches. Deux d'entre-elles ont déjà été décrites dans le papier, tandis que la troisième provient d'une souche de TMV qui se retrouve relié à un cluster de *S.* Typhimurium dans l'arbre pangénomique. Ce résultat est appuyé par le peu de distance génétique entre les souches. Il a également été montré que pgSNP induit des différences topologiques sur les variants monophasiques, qui ont peu de différence au niveau du coregénome, par rapport aux *S.* Typhimurium dont la topologie reste préservée entre les deux méthodes. Ce résultat montre que l'ajout du génome accessoire impacte principalement des souches avec peu de variabilité sur le coregénome, même si les méthodes de coregénome évaluent la topologie sur environ 80% des données. Globalement, pgSNP obtient des résultats concordants avec les données épidémiologiques sur ce jeu de données.

Le deuxième jeu de données correspond à un jeu de données de *Escherichia coli* O157:H7 publié par Rumore et al. en 2018. Ce jeu de données contient 210 souches d'origine humaine, avec 8 clusters épidémiologiques qui ont très peu de différences sur le coregénome (< 5 SNPs). Avec pgSNP, la plupart des clusters épidémiologiques identifiés sont retrouvés en adéquation avec les résultats publiés par les auteurs. En revanche, une nouvelle réconciliation a été identifiée sur des souches du cluster 3, qui se retrouvent proche d'autres souches du cluster 3 dans l'arbre pangénome, tandis que ces souches se retrouvent proche de souches du cluster 6 dans l'arbre de l'étude. En explorant le pangénome de référence, il a été montré que 206kb d'ADN du coregénome absent dans la référence de l'étude était à l'origine de ces nouvelles réconciliations. Ces 206kb contiennent de l'ADN chromosomique, mais également 2 plasmides connus des souches O157:H7. Ce résultat souligne l'impact de la référence dans les analyses de coregénome, et démontre également l'avantage d'utiliser pgSNP. Le génome accessoire fournit une résolution plus élevée sur cet ensemble de données, mais n'ajoute pas de nouvelles réconciliations.

Enfin, pgSNP a été testé sur un jeu de données de clusters épidémiologiques de *Neisseria meningitidis* publié par Whaley et al. en 2018. Comparé aux deux autres jeux de donnée, *Neisseria meningitidis* est très recombinant, avec un génome de petite taille. Également, ce jeu de données contient des informations précises sur les souches sporadiques. Avec pgSNP, il a été montré de nouvelles réconciliations entre 5 souches du cluster 11. Si dans l'arbre de l'étude les souches du clusters 11 sont ensemble et les souches sporadiques sur une autre branche, les souches se retrouvent mélangés dans une seule branche sur l'arbre pgSNP. Les matrices de distances de SNPs corroborent ce résultat, avec des petites distances entre des souches sporadiques et épidémiologiques. Sur le cluster 8, pgSNP permet d'identifier une souche sporadique reliée à des souches épidémiques, validée par les matrices de distances et également en concordance avec les métadonnées correspondant à la souche sporadique. Dans l'ensemble, les pgSNP ajoutent une distance génétique entre les souches épidémiologiques, notamment en raison des éléments génétiques mobiles et de la flexibilité génétique de *Neisseria meningitidis*, ce qui peut rendre les enquêtes sur les foyers épidémiques difficiles. Cependant, les échantillons de foyers, dont le coregénome et le génome accessoire sont proches, sont toujours regroupés dans l'arbre pangénomique, ce qui démontre l'importance du génome accessoire dans les enquêtes épidémiologiques.

Dans cette partie de thèse, nous avons discuté des avantages de l'utilisation d'un pipeline pangénomique dans les études de cas épidémiques, mais également les difficultés d'implémentation et de prise en compte de la complexité des nouvelles informations accessoires. Certaines étapes du pipeline pourraient être améliorées, notamment le pangénome de référence à partir des travaux de Christine Jandrasits (Robert Koch Institute) qui utilise une approche linéaire basée sur un alignement Mauve, ou encore les travaux de Zamin Iqbal (EMBL) qui utilise une approche basée sur les graphes. D'autres étapes supplémentaires, comme la gestion d'insertions et délétions, mais également une gestion plus robuste des sous-alignements contribueraient à une amélioration considérable du pipeline. Ce pipeline a également ouvert de nouvelles discussions sur l'importance du génome accessoire lors des enquêtes épidémiologiques, mais aussi sur l'utilisation des arbres phylogénique dans la visualisation de ces clusters épidémiques. Enfin, plusieurs propositions d'utilisations annexes du pipeline ont été discutées, pour une meilleure portabilité des données et l'anonymisation des données publiés, notamment en se basant sur des pangénomes de référence ou des collections de pan-variants commun entre laboratoires.

En conclusion sur cette partie de thèse, nous avons pu mettre en place un pipeline innovant appelé pgSNP, qui prend en compte le génome accessoire, codant et non-codant, et permet d'inférer ces résultats sur un arbre phylogénomique. Ce pipeline contribue à l'identification de la variabilité du génome accessoire de différents échantillons, à la compréhension de sa prévalence et de sa distribution, de sa persistance et de son risque. Nous sommes très conscients que les analyses pangénomiques vont grandement s'améliorer dans les années à venir, et ainsi améliorer la qualité des analyses génomiques.

Dans cette deuxième partie, les résultats des analyses génomiques des trois principaux sérovars étudiés dans cette thèse vont être présentés : *Salmonella* Typhimurium et son variant monophasique pour la filière porcine, *Salmonella* Mbandaka et *Salmonella* Dublin pour la filière laitière.

Pour répondre aux problématiques de cette thèse, différents jeux de données ont été construits à travers différentes collaborations entre des instituts techniques (ACTALIA, FGIE, IFIP-Institut du Porc) et des instituts publics (ANSES, DGAL, Université de Caen).

Concernant *Salmonella* Typhimurium et son variant monophasique dans la filière porcine, les problématiques principales étaient dans un premier temps de comprendre la diversité de ces sérovars au sein de la chaîne de production de la matière première jusqu'au produit fini. Avec un jeu de données de 322 souches isolées sur différents maillons (élevage, abattoir et usine de transformation), il a été montré grâce à un arbre coregénome que les souches contaminent l'ensemble de la chaîne, sans montrer d'adaptation à une source spécifique. Cette approche a également souligné la grande diversité des *S.* Typhimurium en comparaison aux TMV, validé par le nombre médian de SNPs, comme observé dans la première partie de la thèse sur le jeu de données épidémiologiques.

Dans un second temps, le questionnement principal de la filière était de savoir s'il existe un lien entre la diversité géographique et la diversité génomique chez *S.* Typhimurium et TMV. Pour cela, 188 souches provenant de 3 régions productrices de porc ont été sélectionnées. Certaines souches proviennent des élevages porcins, mais la plupart sont des souches issues d'animaux prélevés à l'abattoir avec département de provenance identifié. Ce jeu de données a permis de mettre en lumière ne répartition dans toute la France de ces souches, sans d'adaptation particulière à la géographie. Comme la différence génomique entre les échantillons est vraiment faible, nous avons émis l'hypothèse qu'un seul clone était disséminé en France. En regardant plus précisément, la différence moyenne de SNP des 152 échantillons de TMV est de 64 SNP. Topologiquement, un nœud interne a divisé les 152 TMV en deux groupes de 104 et 48 échantillons de TMV, avec une moyenne intra-groupe de 49 et 51 SNPs, respectivement. Cette faible diversité des TMV peut être expliquée par l'apparition récente de ce variant, en comparaison aux souches de *S.* Typhimurium.

En examinant le contenu génomique, le nombre élevé de gènes de résistance aux antibiotiques, aux métaux lourds et aux biocides explique la prévalence de ces deux sérovars dans les élevages porcins. L'analyse des variants ou des gènes n'a pas mis en évidence de discrimination entre les échantillons provenant de différentes régions.

pgSNP a également été appliqué à ce jeu de données, afin de vérifier si l'adaptation géographique n'était pas localisée dans le génome accessoire de ces sérovars. Grâce au pipeline, le contenu du génome accessoire a été identifié, avec la présence de phages et de plasmides sur certaines souches. Les analyses ont été réalisées sur 30% de bases supplémentaires pour la reconstruction de l'arbre phylogénomique. Ces nouvelles réconciliations, obtenues grâce au génome accessoire, n'ont pas pu être reliées à la géographique.

Dans un troisième temps, la variabilité des TMV en France a été comparée à la variabilité des TMV dans d'autres pays avec l'utilisation de donnée publiée sur Enterobase. Un arbre a été inféré sur 325 souches, et a montré l'existence de deux « génotypes » de TMV qui circuleraient en France, avec quelques contaminations avec des pays frontaliers (Italie, Allemagne). Ces contaminations peuvent être dues à des échanges de matériel, lors du transport d'animaux, ou une source de nourriture commune. Le calcul de la moyenne des SNPs par pays a montré une faible diversité et comparable à celle observée en France. Cette diversité française a pu être caractérisée en utilisant une combinaison de gènes et de variants, qui est capable de discriminer les souches françaises avec une précision de 86%, soulignant la possibilité de d'identifier l'origine d'une infection humaine française jusqu'aux élevages de porcs à l'aide de marqueurs génomiques. Ces possibles marqueurs se situent sur le génome accessoire, notamment sur des séquences d'insertion (IS :insertion sequence) ou des transposases. Des travaux supplémentaires seraient nécessaires pour vérifier que ces marqueurs sont viables pour développer une potentielle méthode PCR, et une analyse plus poussée des variants accessoires pourrait également révéler une meilleure précision. Mais au regard des scores de précision, l'hypothèse selon laquelle il n'y a pas assez de diversité génomique régionale entre les souches françaises pour les séparer semble la plus probable.

Pour la filière laitière, le sérovar Mbandaka a été exploré avec l'objectif de comprendre sa diversité et sa circulation dans la région normande. Dans un premier temps, un jeu de données de 140 souches de S.Mbandaka de Normandie a été construit, avec des souches provenant de plusieurs matrices (environnement bovin, lait, fromage). La reconstruction phylogénomique a indiqué que les souches n'ont pas tendance à se regrouper par années ou par matrices, démontrant une contamination continue tout le long de la chaîne de production. Il a été également observé une plus grande diversité de *S.* Mbandaka au vu du nombre de SNPs moyen (82 SNPs) en comparaison à ce qui a pu être observé dans les autres jeux de données de cette thèse.

D'après cette observation, la filière s'est questionné sur la possibilité d'une spécificité de ce sérovar à l'hôte bovin. Pour cela, des souches de la filière volaille ont été sélectionnées de la région Normandie, et également des régions Bretagne et Pays de la Loire. L'arbre phylogénomique inférée sur les souches volailles et laitière a montré qu'il n'y a pas de spécificité à l'hôte, mais plutôt des clades aviaires intercalés par des clades bovins. Sur les différents clades d'hôtes, il n'y a pas de spécificité de matrices, que ce soit entre les échantillons de volailles isolés des poules pondeuses et des poulets à chair ou les matrices de la filière bovine. Des échantillons de dinde sont disséminés tout autour de l'arbre, mais avec une longueur de branche unique, et ont donc été identifiés comme des singletons. Les métadonnées géographiques révèlent que la plupart du temps, les échantillons de volaille sont regroupés avec des isolats provenant de la même région ou des mêmes régions limitrophes. Cependant, dans la partie supérieure de l'arbre phylogénomique, les souches de différentes régions sont regroupées ensemble, sans lien de proximité géographique entre les régions. Cette faible diversité entre les souches de différentes sources géographiques concerne presque exclusivement les isolats de volailles, ce qui nous permet d'émettre l'hypothèse de l'existence d'une source de reproduction commune. En effet, les parents se trouvent dans les élevages et s'occupent des naissances des poules, qui sont ensuite triées en poulets de chair et poules pondeuses. Ces poules sont ensuite redistribuées dans les fermes, et peuvent propager la maladie à partir de la même source, ce qui expliquerait ce clade.

PgSNP a également été appliqué à ce jeu de données, et malgré le nombre de bases supplémentaires analysées sur l'arbre phylogénomique (+ 27%), très peu de différence sur les clades d'hôtes déjà identifiés grâce au coregénome ont été identifiés.

Enfin, la piste d'une contamination commune entre les deux filières a été investiguée, sur la base des observations des instituts techniques de la région Normandie. La piste de la transmission par voie aérienne a été proposée, à partir de données publiques de souches d'oiseaux sauvages américains. La proximité de ces souches avec des souches de volailles proche des côtes, mais également d'une souche de fumier normand suggère la possible contamination par la faune sauvage. Des données supplémentaires européennes sont nécessaires afin de valider ces hypothèses. Une étude de cette filière, en lien avec des pays frontaliers pourrait valider cette piste de contamination par la faune sauvage, qui a déjà été observée chez *S.* Typhimurium avec le portage de rongeur.

S. Dublin a également été étudié dans cette thèse, dans le cadre d'une analyse rétrospective d'une épidémie de ce sérovar entre 2015 et 2016 dans la région Franche-Comté. L'objectif était également l'appropriation des outils génomiques pour l'identification et l'investigation de cas épidémiologiques de *S.* Dublin. 480 souches ont été sélectionnées de 4 laboratoires et instituts partenaires de l'étude et de la plateforme, isolées entre 2009 et 2018 et avec différentes matrices d'isolation (fromage, lait, bovin, transport et humain). Avec une étude phylogénomique et une carte anonyme, la précision apportée par les méthodes WGS pour l'identification de différents clusters et de liens inconnus entre les échantillons a été démontré. La distance géographique est un facteur majeur dans la divergence génomique pour *S.* Dublin concernant les premières étapes des processus de production (i.e. animaux, fermes), alors que les étapes de transformation en aval sont plus susceptibles d'abriter une diversité génomique. Avec peu de SNPs, il a été possible d'identifier une ségrégation nationale. Le nombre de SNPs identifiés est équivalent à ce qui a pu être observé sur *S.* Mbandaka, malgré la différence de taille des deux jeux de données. L'utilité de ces outils génomiques a été validée en interne avec l'investigation de nouveaux cas en 2019 qui ont été très rapidement identifiés à un cluster sur l'arbre.

En utilisant les mêmes outils de la méthode WGS, la génomique fonctionnelle comparative a mis en évidence toute la diversité de ces sérovars, qu'elle soit core ou accessoire. La persistance géographique a été étudiée pour *Salmonella* Dublin et *Salmonella* Typhimurium et son variant monophasique, tandis que les analyses de *S.* Mbandaka ont porté sur la persistance dans les hôtes. Sur les mêmes objectifs, les analyses réalisées avec *Salmonella* Dublin et *Salmonella* Typhimurium et TMV ont montré des adaptations complètement différentes. D'une part, *Salmonella* Dublin s'est adaptée et a persisté dans l'environnement. D'autre part, TMV a un arsenal de résistance très bien adapté à l'hôte porcin (résistance au cuivre, résistance aux biocides), et ne semblait pas avoir développé de spécificité environnementale pour pouvoir se répandre facilement dans l'environnement en France. Dans les discussions sur cette partie, il est mentionné que certaines études contredisent ces résultats alors que d'autres les confortent, en raison des échelles de ces études. Ici, l'ensemble des jeux de données restreint aux différentes problématiques des filières sans un nombre élevé de vecteur différents a permis d'expliquer ces phénomènes de persistance. Grâce à tous les outils utilisés dans ce chapitre de thèse, il a été possible d'analyser les génomes à l'échelle des variants core, ainsi que le contenu en gènes core et accessoires. Une analyse supplémentaire avec les variants accessoires rendrait la méthodologie plus robuste, notamment pour le criblage des marqueurs.

Plusieurs limites ont été rencontrées dans cette thèse, notamment liées au manque de métadonnées qui ne nous a pas permis de conclure sur des pistes importantes dans chaque sérovars. Pour

S. Dublin, seule la moitié des souches a pu être réellement analysée géographiquement. Pour *S.* Mbandaka, la persistance géographique précise des souches et les liens de contamination entre élevages bovins ou avicoles n'ont pu être démontrés. Enfin, pour *S.* Typhimurium et son variant monophasique, un lien entre diversité génomique et diversité géographique n'a pas pu être démontré, car seules des données départementales étaient disponibles, et il n'a donc pas été possible de faire des liens entre certaines fermes ou entreprises de transformation. De plus, des études sur les réseaux de transport en camion des élevages ou les échanges entre exploitations seraient nécessaires pour finaliser les résultats, mais ce type de données est très difficile à obtenir. Avec ces informations, il aurait été possible de répondre à plus d'hypothèses, que ce soit au niveau scientifique qu'industriel.

Certains biais liés au jeu de donnée ont pu être introduits pendant cette étude. Par exemple, pour la filière porcine, il y a un manque de données concernant les souches de *S.* Typhimurium, mais également de souches de TMV issues de régions hors Bretagne, dû au fait que cette région produise 78% de la viande porcine française. Pour compenser, des souches provenant d'abattoirs dont les données géographiques du cheptel porcin sont connues ont été ajoutées. Cependant, un porc abattu alors qu'il est contaminé peut transmettre la bactérie au reste de la chaîne, ce qui entraîne des contaminations croisées.

Enfin, l'ajout du génome accessoire n'a pas permis d'obtenir de nouvelles conclusions sur ces études. La connaissance limitée du coregénome empêche la compréhension de l'hétérogénéité du génome accessoire, et ne permet pas de comprendre pleinement la contribution du pangénome à l'inférence phylogénomique. Dans l'ensemble, la principale différence a été observée sur le jeu de données du TMV, mais la contribution du génome accessoire n'a pas apporté de nouvelles hypothèses pour la dissémination de ces souches dans différentes régions en France.

En conclusion de cette partie, ce travail a été développé dans le but de comprendre la dissémination des souches, mais aussi dans le but d'aider les acteurs industriels à comprendre la persistance de ces souches.

Finalement, cette thèse a permis le développement d'une nouvelle méthodologique, et son application directe à des cas de terrain, en utilisant des approches pangénomiques implémentées dans un outil appelé "pgSNP". Les avantages et les nouveaux résultats déduits par pgSNP ont été décrit dans le chapitre 3 de cette thèse. pgSNP a été capable de trouver des résultats cohérents avec une approche SNPs coregénome, mais aussi de fournir plus de résolution dans les analyses phylogénomiques. Il a été possible de démontrer la possibilité d'appliquer ce pipeline sur différents jeux de données de foyers épidémiques pour montrer l'importance des informations obtenues avec cette méthode. Les avantages et les limites du pipeline ont été analysés, et des améliorations ont été suggérées. Cette thèse a démontré l'importance des analyses pangénomiques qui seront grandement améliorées dans les années à venir, avec une possible progression dans ce domaine grâce aux différents développements de cette thèse. Des sérovars de *Salmonella* prévalents dans la filière porcine et de l'alimentation laitière ont été également caractérisés finement dans ces études. En utilisant des méthodes génomiques comparatives, la diversité de ces souches a été caractérisée sous différents enjeux. Les développements coregénomique et pangénomique ont été appliqués sur des sérovars ayant une plasticité et une évolution différentes, ainsi que sur des génomes avec peu de diversité (*S.* Dublin, TMV) ou très hétérogènes (*S.* Mbandaka). Même si les sérovars présentent des problématiques et une contextualisation très différente, les méthodes WGS présentées dans cette thèse sont suffisamment efficaces pour explorer les différentes questions soulevées. Cette thèse a permis de valider certaines hypothèses et d'en proposer de nouvelles sur ces sérovars prévalents qui posent des problèmes de sécurité des aliments et de santé animale. Ces résultats ont ouvert de nouvelles possibilités d'études concernant les sérovars étudiés dans cette thèse, que ce soit dans le domaine de la bioinformatique ou de la microbiologie. Pour conclure, cette thèse reflète une recherche méthodologique et une recherche appliquée dans un domaine en pleine expansion, et fait le point sur l'état actuel de la recherche dans ce domaine, tout en proposant des éléments de réponses et de nouveaux sujets à explorer.

Title : Genomics of Salmonella sevovars Mbandaka, Typhimurium and its monophasic variant in milk and pork food sectors

Keywords : Pangenome, Genomics, Phylogenomic, Microbiology, Methodology, Food safety

Abstract :

Salmonella Mbandaka, Typhimurium and its monophasic variant are prevalent serovars in dairy and pork food sectors. Faced with industrial limits and lack of knowledge of the mechanisms and determinants of the dissemination, the persistence and the resistance of these strains, whole genome sequencing approaches gained interest from both food sectors.

In this thesis, I have developed an innovative methodology, called "pan-genome" in order to take into account all single nucleotide polymorphisms within a considered set of *Salmonella* genomes, including the accessory and coregenome from coding and non-coding DNA fragments. In addition to have demonstrated that my developments allowed the inference of a phylogenomic tree in agreement with the epidemiological data through several datasets (*Salmonella* Typhimurium and its monophasic variant, *Escherichia coli* and *Neisseria meningitidis*), this work also revealed that the pangenomic inference produced new reconciliations between strains compared to the coregenome-based inference. These developments provided a pangenomic method with a higher discriminatory power than the usual methods based on core or accessory genomes, and consequently brought a potential solution for the improvement of outbreak investigations.

In addition, I studied in detail genomes to detect firstly host markers, and secondly geographical markers, on the French and global scales. I demonstrated that genomic clusters are harbored by *Salmonella* Mbandaka isolated from poultry and cattle. For *Salmonella* Typhimurium and its monophasic variant, I observed an absence of geographical distinction of strains isolated from pig herds, and that a single genomic profile was found dispersed in France, while a geographical segregation worldwide was observed.

Overall this research provided a solid overview of the genomic of *Salmonella* Mbandaka, Typhimurium and its monophasic variant in dairy and pig and pork food sectors, and a pangenomic method to bring further resolution in future bacterial epidemiological investigations.

Titre : Génomique des sérotypes de *Salmonella* Mbandaka, et Typhimurium et son variant monophasique, dans les secteurs alimentaires laitier et porcin.

Mots-clés : Pangénome, Génomique, Phylogénomique, Microbiologie, Méthodologie, Sécurité sanitaire

Résumé :

Salmonella Mbandaka, Typhimurium et son variant monophasique sont des sérovars de *Salmonella* très prévalents dans les secteurs alimentaires laitier et porcin. Face aux limites industrielles et la méconnaissance des mécanismes et déterminants de la dissémination, la persistance et la résistance de ces souches, les approches par séquençage du génome entier ont suscité l'intérêt des filières alimentaires en question.

J'ai développé une méthodologie innovante, dite « pangénome » afin de prendre en compte tous les polymorphismes de nucléotides simples des génomes considérés de *Salmonella*, incluant le coregénome et le génome accessoire de fragments codant et non codant. En plus d'avoir démontré que mes développements permettaient l'inférence d'un arbre phylogénétique en cohérence avec les données épidémiologiques à travers plusieurs jeux de données (*Salmonella* Typhimurium et son variant monophasique, *Escherichia coli* et *Neisseria meningitidis*), ces travaux ont aussi révélé que l'inférence pangénomique engendrait de nouvelles réconciliations entre les souches en comparaison à l'inférence basée sur le coregénome. Ces développements ont fourni une méthode pangénomique plus discriminante que les méthodes usuelles basées sur les génomes core ou accessoire, et ont par conséquent apporté une potentielle solution à l'amélioration des investigations d'épidémies.

Dans cette thèse, j'ai également étudié en détail les génomes pour rechercher dans un premier temps des marqueurs d'hôtes, et dans un deuxième temps des marqueurs géographiques, à l'échelle de la France et à l'échelle mondiale. J'ai démontré l'existence de clusters génomiques chez les *Salmonella* Mbandaka isolées de volaille et du bovin. Pour *Salmonella* Typhimurium et son variant monophasique, j'ai observé qu'il n'y avait pas de distinction géographique entre des souches isolées d'élevage porcin, et qu'un seul profil génomique se retrouvait dispersé en France, avec une ségrégation géographique à l'échelle mondiale.

Dans l'ensemble, ces travaux fournissent un aperçu solide de la génomique de *Salmonella* Mbandaka, Typhimurium et son variant monophasique, et une méthode d'analyse pangénomique pour apporter une meilleure résolution aux futures enquêtes épidémiologiques bactériennes.

PARIS-EST SUP
6-8 avenue Blaise-Pascal
Cité Descartes - Champs-sur-Marne
77455 Marne-la-Vallée, France