Quelle méthode analytique WGS choisir en vue de l'investigation d'évènements sanitaires?

Claire Yvon

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

MÉMOIRE

présenté
par

YVON Claire

pour l'obtention du Diplôme de l'École Pratique des Hautes Études

TITRE : Quelle méthode analytique WGS choisir en vue de l'investigation d'évènements sanitaires ? Comparaison d'outils bio-informatiques en vue d'une validation et accréditation de méthode

Soutenu le 15/10/2020 devant le jury suivant :

Pr. MATHIEU Laurence **- Président**
Dr. CADEL-SIX Sabrina **- Tuteur scientifique**
Pr. WIRTH Thierry **- Tuteur pédagogique**
Dr. MOURA Alexandra **- Rapporteur**
Dr. PARDOS DE LA GANDARA Maria **- Examinateur**

Mémoire préparé sous la direction de :
CADEL-SIX Sabrina
Intitulé de la structure d'accueil : ANSES – Laboratoire de sécurité des aliments – Unité SEL – Maisons-Alfort
Directeur : M. LALOUX Laurent

et de

Pr. WIRTH Thierry
Intitulé de la structure d'accueil EPHE : Muséum National d'Histoire Naturelle - UMR 7205 Institut de SYstématique, Évolution, Biodiversité (ISYEB) - Paris
EPHE (Sciences de la Vie et de la Terre)
Groupe de Recherche et d'Enseignement Thématique de l'EPHE : Évolution-Morphologie, Anthropologie, Génomique (ÉVOLUTION)

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

MÉMOIRE

présenté
par

YVON Claire

pour l'obtention du Diplôme de l'École Pratique des Hautes Études

TITRE : Quelle méthode analytique WGS choisir en vue de l'investigation d'évènements sanitaires ? Comparaison d'outils bio-informatiques en vue d'une validation et accréditation de méthode

Soutenu le 15/10/2020 devant le jury suivant :

Pr. MATHIEU Laurence **- Président**
Dr. CADEL-SIX Sabrina **- Tuteur scientifique**
Pr. WIRTH Thierry **- Tuteur pédagogique**
Dr. MOURA Alexandra **- Rapporteur**
Dr. PARDOS DE LA GANDARA Maria **- Examinateur**

Mémoire préparé sous la direction de :
CADEL-SIX Sabrina
Intitulé de la structure d'accueil : ANSES – Laboratoire de sécurité des aliments – Unité SEL – Maisons-Alfort
Directeur : M. LALOUX Laurent

et de

Pr. WIRTH Thierry
Intitulé de la structure d'accueil EPHE : Muséum National d'Histoire Naturelle - UMR 7205 Institut de SYstématique, Évolution, Biodiversité (ISYEB) - Paris
EPHE (Sciences de la Vie et de la Terre)
Groupe de Recherche et d'Enseignement Thématique de l'EPHE : Évolution-Morphologie, Anthropologie, Génomique (ÉVOLUTION)

Remerciements

Je tiens à remercier l'ensemble des personnes ayant contribuer de près ou de loin au bon déroulement de ce diplôme, au cours des deux dernières années.

Tout d'abord, M. Laurent LALOUX, Directeur du Laboratoire de Sécurité des Aliments, pour m'avoir permis de réaliser ce diplôme au sein de son laboratoire, sur un sujet d'actualité telle que la génomique au service de la santé publique. Sans oublier, Jean-Charles LEBLANC et Corinne DANAN, de m'avoir permis de mener à bien ce projet au sein de l'unité SEL, en prenant une part active aux activités annexes directement liées au sujet du présent mémoire.

M. David ALBERT, qui a été l'initiateur de ce projet et mon tuteur scientifique pendant quelques mois, grâce à qui le projet a pu prendre vie. Merci d'avoir eu confiance en moi et de m'avoir accompagnée dans cette démarche.

Le Dr. Sabrina CADEL-SIX, ma tutrice scientifique, sans qui je n'aurai pas pu mener ce projet à bien. Grâce à son implication, sa pédagogie et ses précieux conseils, j'ai eu l'occasion de développer de nouvelles compétences tout au long du diplôme.

Le Pr. Thierry WIRTH, mon tuteur pédagogique, qui m'a fait prendre conscience de nouvelles perspectives liées au projet, tout en m'accompagnant dans chaque étape clé du diplôme.

Sans l'expertise et la patience de mes collègues Emeline CHERCHAME, Maroua SAYEB, Federica PALMA, Yann SEVELLEC, Madeleine DE SOUSA VIOLANTE et Ludivine BONANNO je n'aurai pas eu l'occasion de développer autant de compétences en deux ans, je les en remercie grandement.

Je remercie également mes proches, de m'avoir accompagnée et encouragée. J'ai une pensée particulière pour mes deux fidèles relecteurs et correcteurs, Sophie et Mathieu.

Table des matières

Index des figures

Index des tableaux

Liste des abréviations

ARS : Agence Régionale de Santé

CC : *Clonal Complex* (en Français, Complexe Clonal)

CFSAN : Centre pour la sécurité alimentaire et la nutrition appliquée

CNR : Centre National de Référence

Cofrac : Comité Français d'Accréditation

DGAL : Direction Générale de l'Alimentation

DD(cs)PP : Direction Départementale (de la cohésion sociale) et de la Protection des Populations

ECDC : *European Centre for Disease Prevention and Control*

FAO : Organisation des Nations unies pour l'alimentation et l'agriculture

FDA : Food and Drug Administration (en français : Agence américaine des produits alimentaires et médicamenteux)

ISO : International Organization for Standardization (en français : Organisation Internationale de la Normalisation)

LNR : Laboratoire National de Référence

LRUE : Laboratoire de Référence de l'Union Européenne

LSAl : Laboratoire de Sécurité des Aliments

ML : *Maximum Likelihood*

MUS : Mission des Urgences Sanitaires

OIE : Organisation mondiale de la santé animale

Pb : Paire de bases

RS : Réseau *Salmonella*

SpF : Santé Publique France

ST : *Sequence Type*

TIAC : Toxi-Infection Collective

SNP : *Single Nucleotide Polymorphism*

wgSNP : *whole genome Single Nucleotide Polymorphism*

WGS : *Whole Genome Sequencing*

MLST : *MultiLocus Sequence Typing*

cgMLST : *core genome MultiLocus Sequence Typing*

wgMLST : *whole genome MultiLocus Sequence Typing*

1 Introduction

1.1 Les laboratoires français et européens impliqués dans la surveillance de *Salmonella* et *Listeria monocytogenes*

L'implication des laboratoires dans le cadre de la surveillance officielle sur les pathogènes alimentaires *Salmonella* et *Listeria monocytogenes* repose sur leur mandat de référence ou sur leur agrément ou reconnaissance délivrés par la DGAL.

En France, pour ce qui concerne les analyses issues de prélèvement humain, la surveillance sur les deux pathogènes est réalisée par le centre national de référence (CNR) de l'Institut Pasteur (Paris). A travers la collecte des souches envoyées par les laboratoires de biologie médicale et hospitaliers, le CNR est en mesure de réaliser une surveillance (évènementielle) des contaminations survenues sur le territoire et de détecter des cas groupés. Le CNR *Listeria* reçoit également les souches isolées de prélèvements non humains isolées lors d'alertes et d'investigations réalisées sur demande des autorités compétentes pour comparaison génomique avec les souches humaines.

En parallèle, les souches isolées sur la chaîne agro-alimentaires dans le cadre des contrôles programmés (plans de surveillance et de contrôle) par les autorités compétentes, sont collectées et analysées par les laboratoires nationaux de référence (LNR). Un LNR est le garant de la fiabilité des analyses effectuées par l'ensemble des laboratoires agréés et reconnus par l'état. Au niveau Européen, l'ensemble des LNR sont coordonnés par le laboratoire de référence de l'union européenne (LRUE). Celui-ci a pour mission d'harmoniser les pratiques et d'assurer la fiabilité des méthodes d'analyses utilisées par les différents Etats membres.

Au niveau international, un laboratoire peut être désigné comme laboratoire de référence OIE (LR OIE) ou centre de référence FAO (CR FAO). Ainsi, il agit en tant qu'expert sur le pathogène mandaté.

L'Agence Nationale de Sécurité Sanitaire de l'alimentation, de l'environnement et du travail (Anses) détient soixante-cinq mandats de référence nationaux, neuf mandats européens et vingt-cinq mandats internationaux.

Les CNR, LNR et LRUE sont chargés, dans le cadre de leurs mandats, d'apporter aux pouvoirs publics un appui scientifique et technique notamment en situation d'alerte sanitaire, dans le cadre d'investigations de cas groupés humains (« clusters ») ou d'épidémies. Ils participent à la recherche des sources alimentaires ou environnementales responsables des contaminations humaines. Ils effectuent la comparaison de résultats d'analyse obtenus sur des isolats bactériens provenant de matrices issues de la chaîne alimentaire, de l'environnement ou au niveau des productions primaires.

Ces demandes d'analyses émanent de différents acteurs :

- Au niveau national :
    - Santé publique France (SpF) : agence placée sous la tutelle du ministère de la santé en charge de la surveillance épidémiologique de la population humaine travaille en collaboration directe avec les CNR.
    - La mission des urgences sanitaires (MUS) de la Direction générale de l'alimentation (DGAL)
- Au niveau Européen : l'ECDC (European Center for Disease Prevention and Control), en collaboration avec l'EFSA (European Food Safety Autority) assurent une surveillance épidémiologique harmonisée respectivement chez l'homme et sur la chaine alimentaire

1.2 Le laboratoire de sécurité des aliments

L'Anses résulte de la fusion, en juillet 2010, de l'agence française de sécurité sanitaire des aliments (Afssa) et de l'Agence française de sécurité sanitaire de l'environnement et du travail (Afsset). Placée sous la tutelle des Ministères chargés de la Santé, de l'Agriculture, de l'Environnement, du Travail et de la Consommation, l'agence a pour but de contribuer à assurer la sécurité sanitaire humaine, végétale et animale, à travers des domaines multiples.

L'Anses compte neuf laboratoires en France, dont quatre sont spécialisés dans la sécurité des aliments et des eaux de consommation.

Le laboratoire de sécurité des aliments (LSAl) intervient sur les dangers biologiques et chimiques qui peuvent affecter la sécurité sanitaire et la qualité des aliments. Il participe dans son domaine de compétence à l'accomplissement des missions de référence, de recherche, de veille, de surveillance et d'expertise scientifique et technique de l'Anses.

Le LSAl se décline en plusieurs unités spécialisées dans les contaminants chimiques des aliments, dont une partie traite plus spécifiquement des produits de la pêche et de l'aquaculture, et les contaminants microbiologiques des aliments. Parmi ces unités, se situe l'unité *Salmonella* et *Listeria* (SEL). L'unité intervient donc dans l'analyse et le typage des deux pathogènes alimentaires, collectés à toutes les étapes « de la fourche à la fourchette ». Celle-ci exerce des activités dans le cadre de plusieurs mandats :

- LRUE *Listeria monocytogenes*
- LNR *Listeria monocytogenes*
- Laboratoire associé au LNR *Salmonella* (Anses de Ploufragan)

L'unité SEL anime depuis 1997 un réseau de surveillance des salmonelles d'origine non humaine, le Réseau *Salmonella* (RS), constitué d'environ 150 laboratoires français volontaires, privés et publics, permettant de constituer une souchothèque et une base de données référençant les caractéristiques biologiques données épidémiologiques des isolats détectés dans les secteurs de la santé animale, l'hygiène des aliments et l'écosystème naturel. Ce réseau et apporte un appui technique en termes de typage des souches auprès des laboratoires partenaires. L'unité collecte des souches de L. monocytogenes dans le cadre des activités de LNR et LRUE pour ce pathogène. L'animation d'un réseau national similaire au RS sur *Listeria monocytogenes* est en cours d'étude. Ainsi, l'unité collecte, caractérise et type près de 20 000 souches par an sur les deux pathogènes confondus.

1.3 Les différents types d'évènements sanitaires

Les évènements sanitaires nécessitant une étude approfondie par typage peuvent être définis en trois grands groupes : les alertes produits, les alertes sanitaires et les toxi-infections alimentaires collectives (TIAC). En France, la coordination de ces déclarations se fait entre la MuS (gérée par le DGAL), SpF et les directions départementales en charge de la protection des populations (DDPP). Selon SpF, en France, les maladies infectieuses d'origine alimentaire représentent environ deux millions de cas annuels, environ 20 000 hospitalisations et 300 décès (Van Cauteren, 2017). *Salmonella* fait partie des principaux agents responsables des cas et des hospitalisations ; les infections à *Salmonella* spp. et *Listeria monocytogenes* représentent la moitié des décès d'origine alimentaire.

En cas d'apparition d'au moins deux cas similaires d'une symptomatologie pouvant être rapportée à une même origine, on parle de Toxi-Infections Alimentaires Collectives **(TIAC)**. Elles constituent des maladies à déclaration obligatoire auprès des DDPP ou de l'Agence Régionale de Santé (ARS). Les TIAC sont souvent liées à l'utilisation de matières premières contaminées et/ou au non-respect des mesures d'hygiènes et de température (rupture de chaîne du froid ou du chaud) lors des préparations culinaires. Le plus souvent, les TIAC sont familiales ou associées aux modes de restauration commerciale ou collective. Leur signalement entraine la réalisation d'une enquête épidémiologique destinée à identifier les aliments responsables, les facteurs favorisant leur contamination et de prendre des mesures correctives locales et rapides afin d'éviter la survenue de nouveaux cas de contamination.

Une **alerte produit** correspond à la mise en évidence d'une non-conformité sur un produit à différents stades de la production : depuis la matière première jusqu'à la mise en rayon du produit fini. A diverses étapes de la chaîne alimentaire, les industriels ont obligation de vérifier la sécurité et la salubrité de leur production sur différents aspects : anomalie visuelle, odeur, contamination chimique ou contamination microbiologique. La déclaration d'une alerte produit permet de prendre des mesures de gestion afin de faire cesser l'exposition du consommateur au produit contaminé (destruction des stocks et retrait des rayons,), éviter les contaminations croisées, et informer le consommateur (mesures de rappels lorsque le produit est déjà chez le consommateur). D'après une étude anglaise, les mesures sanitaires prises en cas de déclaration d'une alerte produit par contamination en salmonelles seraient responsables d'une perte financière d'environ 600 000€ (Sockett & Roberts, 1991).

En cas d'apparition de malades en nombre inhabituel suite à l'ingestion de produits contaminés, une alerte produit peut devenir une **alerte sanitaire**. Le CNR, grâce à la collecte et l'analyse des souches humaines, permet de mettre en évidence des cas isolés ou foyers infectieux liées à l'ingestion de produits contaminés. Grâce aux enquêtes épidémiologiques il est alors parfois directement possible de suspecter un aliment. Dans d'autres cas, il est alors nécessaire d'effectuer une étude d'attribution des sources à partir des résultats des analyses microbiologiques.

Pour *Salmonella*, le RS est un dispositif majeur pour détecter l'augmentation de la fréquence d'apparition de certains sérovars par filières. En effet, compte tenu du grand nombre de laboratoires partenaires, et du nombre de souches reçues (environ 13 000 souches par an), toute recrudescence de détection d'un sérovar sur une filière peut permettre d'anticiper des alertes sanitaires.

1.4 Les microorganismes étudiés

1. Salmonella

Les salmonelles sont des bacilles à Gram négatif. La taxonomie actuelle des salmonelles est basée sur la connaissance des antigènes somatiques et flagellaires spécifiques qui ont mené à la distinction de plus de 2500 sérovars dans la classification de White-Kauffmann-Le Minor (Grimont & Weill, 2007, Tindall *et al.*, 2005). Cette classification est maintenue et actualisée par le centre collaborateur de l'Organisation Mondiale de la Santé (OMS) pour la référence et la recherche sur *Salmonella*. La méthode de référence pour la caractérisation des salmonelles repose actuellement sur une technique de sérotypage sérologique conventionnel sur lame.

Le genre *Salmonella* est représenté par deux espèces, *Salmonella enterica* et *Salmonella bongori*. *S. enterica*, représentant 95% des isolats issus de l'humain, est divisée en six sous-espèces :

- *S. enterica* subsp. *enterica* (I)
- *S. enterica* subsp. *salamae* (II)
- *S. enterica* subsp. a*rizonae* (IIIa)
- *S. enterica* subsp. *diarizonae* (IIIb)
- *S. enterica* subsp. *houtenae* (IV)
- *S. enterica* subsp. *indica* (VI)

Les salmonelles sont capables de réduire les nitrates, de fermenter le glucose mais ne possèdent pas d'oxydase. Elles peuvent utiliser le citrate comme seule source de carbone. Les sous-espèces de *S. enterica* peuvent être identifiées selon leurs caractéristiques biochimiques (Tableau 1).

Tableau 1 - Caractéristiques biochimiques des différentes espèces et sous espèces du genre Salmonella (Grimont & Weill, 2007)

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.

L. Le Minor, M. Véron, M. Popoff. *Ann. Microbiol. (Inst. Pasteur)*, 1982, **133 B**, 223-243 and 245-254.
L. Le Minor, M.Y. Popoff, B. Laurent, D. Hermant. *Ann. Inst. Pasteur/Microbiol.*, 1986, **137 B**, 211-217.

Selon le rapport conjoint de l'EFSA et de l'ECDC (EFSA and ECDC, 2018), les salmonelloses sont la 2ème cause de zoonoses entrainant plus de 91 000 cas en Europe sur l'année 2017. Les salmonelloses non typhiques touchent particulièrement les enfants, les personnes âgées ou les personnes immunodéprimées. La transmission à l'homme se fait principalement par l'ingestion d'aliments contaminés crus ou peu cuits, provoquant des symptômes de gastro-entérites. De faibles doses infectieuses sont suffisantes pour provoquer des symptômes cliniques (Majowicz *et al.*, 2010). Bien que la majorité des cas s'avèrent non critiques, une infection à *Salmonella* peut toutefois conduire à une septicémie. Après l'infection, certains patients peuvent rester porteurs et devenir des porteurs sains, contribuant à la propagation de la maladie.

La recherche de *Salmonella,* tout au long de la chaîne alimentaire, s'inscrit dans le respect de la réglementation européenne (Paquet hygiène). Les règlements (CE) N°178/2002 et N°2073/2005 définissent les responsabilités des différents acteurs de cette chaîne et les critères microbiologiques de sécurité et d'hygiène qui ciblent notamment les salmonelles dans les aliments. Dans leurs derniers avis concernant *Salmonella*, l'EFSA (2010) et l'Anses (2013) recommandent le sérotypage complet des salmonelles isolées sur la chaîne alimentaire pour fournir une information précise aux évaluateurs et gestionnaires du risque. Le paquet hygiène impose l'absence de salmonelles dans 25g dans la plupart des aliments destinés à la consommation humaine. En parallèle, en élevage avicole, en France, un plan de lutte a été défini cinq sérovars réglementés : *S.* Enteritidis, *S.* Typhimurium et son variant monophasique : *S.* 1,4,[5],12:i:, *S.* Hadar, *S.* Infantis et *S.* Virchow auxquels s'ajoute *S.* Kentucky pour lesquels la déclaration, ainsi que la mise en place de mesures de prévention, de surveillance et de lutte, sont obligatoires. Ces mesures peuvent conduire à la mise en quarantaines d'élevages, leur abattage ou encore la mise en place de contrôles renforcés, ayant un impact économique important aux différentes étapes.

Les salmonelles possèdent des génomes de taille conséquente, entre 4,4 et 5,4 Mb. La diversité génétique de *Salmonella* est principalement due à de nombreux échanges horizontaux de matériel génétique. Les salmonelles sont porteuses de nombreux éléments génétiques mobiles provenant principalement d'autres entérobactéries, tel que des transposons qui peuvent se recombiner dans des plasmides ou dans le chromosome bactérien via des séquences d'insertion qui leur permettent de s'intégrer dans l'ADN de l'hôte. L'acquisition de ces éléments permet à la bactérie d'acquérir des facteurs de virulence, une résistance aux antibiotiques ou à des caractéristiques phénotypiques particulières. Dû à des phénomènes de pressions sélectives, la résistance aux antibiotiques des salmonelles est en évolution constante, avec une observation de l'augmentation du nombre de souches multi résistantes (bactérie résistant à au moins trois familles d'antibiotiques).

Plusieurs étapes sont impliquées dans la pathogénicité de *Salmonella* :

- L'invasion de la muqueuse de l'intestin grêle
- L'adhésion et l'invasion rapides des cellules de l'épithélium associées aux follicules et des entérocytes absorbants.
- Les défenses contre les mécanismes de défense spécifiques à l'hôte qui comprennent les actions antibactériennes des cellules phagocytaires couplées à la réponse immunitaire.

L'interaction entre *Salmonella* et l'hôte est un mécanisme complexe et dépend de plusieurs ilots de pathogénicité (SPI). La plupart de ces gènes responsables de la pathogénicité de la bactérie sont situés dans cinq îlots chromosomiques hautement conservés de pathogénicité de *Salmonella* (SPI-1 à SPI-5) et dans des plasmides de virulence (par exemple pSLT).

2. *Listeria monocytogenes*

Le genre *Listeria* comporte vingt et une espèces (Quereda *et al.*, 2020) dont l'espèce la plus pathogène pour l'Homme est *Listeria monocytogenes*. Il s'agit d'un bacille à Gram positif, psychrophile et ubiquitaire. L'espèce *monocytogenes* est divisée en treize sérovars basés sur les antigènes somatiques et flagellaires (Ragon *et al.*, 2008). Depuis 2005, ces sérovars ont été remplacés par cinq génosérogroupes déterminés par PCR (Doumith *et al.*, 2005) : IIa (sérovars 1/2a et 3a), IIb (sérovars 1/2b et 3b), IIc (sérovars 1/2c et 3c), IVb (sérovars 4b, 4d et 4e) et L (autres sérovars). Parmi ceux-ci, les génosérogroupes IVb puis IIa puis IIb sont les plus reliés aux cas humains. Le typage de séquences multi locus (MLST) a permis de subdiviser les génoségroupes obtenus par PCR en Complexes Clonaux (CC). Il est actuellement considéré que tous les isolats de L. monocytogenes sont virulents, bien que certains sérotypes soient isolés plus fréquemment dans les isolats humains (Maury *et al.*, 2016).

D'après le rapport conjoint de l'EFSA et de l'ECDC (EFSA and ECDC, 2018), la listériose est la 5ème cause de zoonose, avec 2 500 cas en 2017. 99% des cas reportés ont été contaminés par des aliments. Comme pour *Salmonella*, il apparait que *Listeria* est un pathogène opportuniste, infectant principalement les populations à risques, tel que les enfants, les personnes immunodéprimées ou les femmes enceintes. La pathogénicité de *L. monocytogenes* repose principalement sur sa capacité d'invasion, plus particulièrement sur la présence de gènes de virulence. Le gène *prfA* est identifié comme étant responsable de la régulation de nombreux gènes de virulence (Zhou *et al.*, 2011).

Les génomes *Listeria* ont une taille de 2,8 à 3,2 Mb. L'évolution génomique des *Listeria* est lente, dû à une faible proportion d'acquisition ou de perte de gènes. Toutefois, *Listeria monocytogenes* apparait comme un pathogène pouvant devenir persistant à différentes étapes du processus agroalimentaire de la fourche à la fourchette. Il s'agit en effet d'un pathogène cultivant sur de grandes plages de températures (-0,5°C à 45°C), supportant un pH de 4,4 à 9,4, la dessiccation et capable de former des biofilms. En parallèle, il apparait également que *L. monocytogenes* peut acquérir des résistances spécifiques à son environnement. En effet, dans l'industrie agroalimentaire, des études ont démontré la présence de gènes responsables de la résistance à des produits d'entretiens industriels, comme le benzalkonium, ou encore des résistances aux métaux lourds (Palma, 2020). Cette résistance est due à la présence de certains gènes sélectionnés sous la pression des processus de nettoyage utilisés dans les élevages et les usines de transformations agro-alimentaires, permettant à la bactérie de développer d'autres facteurs de persistance.

1.5 Le virage vers la génomique

Historiquement, pour *Salmonella* et *Listeria*, les typages sont basés sur le sérotypage par agglutination ou moléculaire. Dans le but de mieux caractériser les souches, des techniques de sous-typage moléculaires ont été développées. Plusieurs techniques de sous-typage ont apparues :

- L'électrophorèse en champ pulsé (PFGE) (Pulsenet, 2013b, Pulsenet, 2013a)
- Le typage de séquences multi locus (MLST) (Achtman *et al.*, 2012, Bale *et al.*, 2016, Maiden *et al.*, 1998)
- L'analyse des loci à séquences variables répétées en tandem ou VNTR (en anglais, Multi Locus Variable Analysis : MLVA) (ECDC, 2016) ; (Vignaud *et al.*, 2017).

En 2001, la génomique a pris un nouveau tournant symbolique grâce au séquençage du premier génome humain complet (Lander *et al.*, 2001, Venter *et al.*, 2001). Ce premier génome complet a pu être obtenu grâce à des méthodes de séquençage dites de première génération : la technique de Sanger. Au fur et à mesure des années, les technologies de séquençage de seconde et troisième génération sont apparues, permettant un séquençage à la fois plus rapide, moins coûteux et avec moins d'erreurs (Kchouk, 2017).

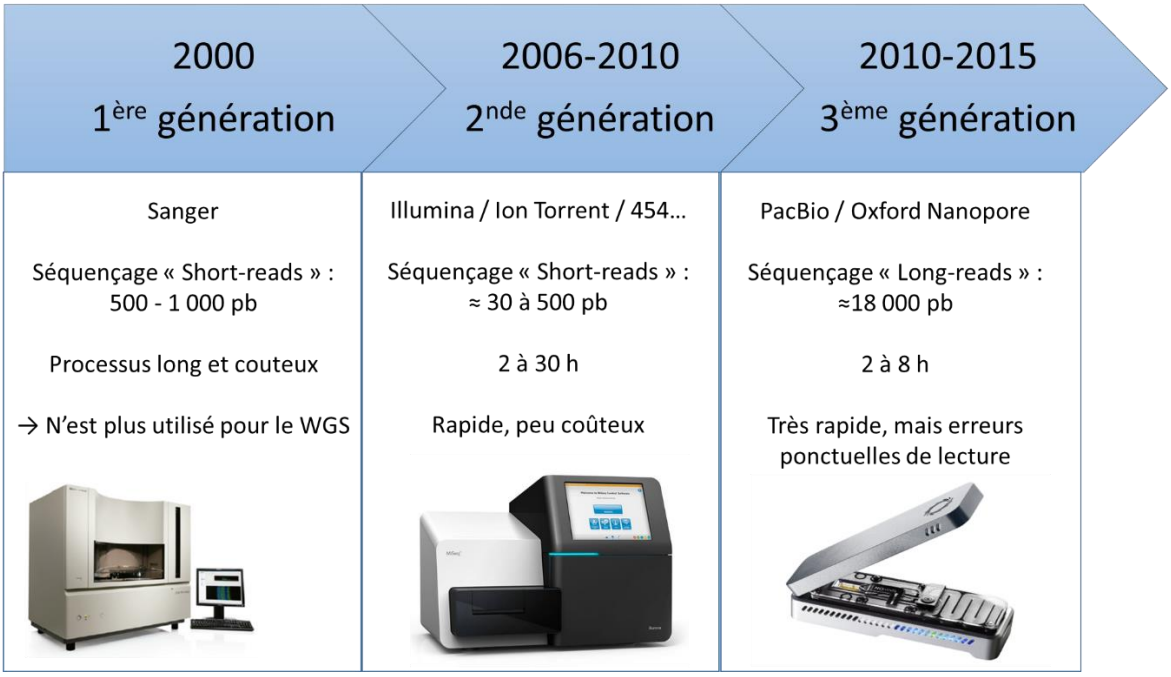


Figure 1 - Les différentes générations de séquenceurs

De ce fait, le séquençage du génome complet, appelé « Whole Genome Sequencing (WGS) », remplace progressivement les techniques historiques de biologie moléculaire permettant de comparer des génomes entiers jusqu'à la résolution d'un seul nucléotide. (Portmann *et al.*, 2018, Achtman *et al.*, 2012).

La transition étant en marche dans tous les secteurs du biomédical, le coût de génération des génomes complets est tombé en flèche. Toutefois, la diminution du coût du WGS a été accompagnée d'une augmentation conséquente de la quantité de données générées (Schwarze *et al.*, 2018, Caulfield *et al.*, 2013) et des coûts annexes associés au déploiement des systèmes d'information nécessaires à la gestion et au stockage de ces données. De plus, pour les laboratoires, il a été nécessaire de se former et de s'équiper pour employer de nouvelles compétences en vue de l'exploitation de ces nouvelles données. Toutefois, les possibilités analytiques sont quasiement infinies, permettant de combiner plusieurs analyses de typages en une seule (possibilité d'obtenir en plus des informations spécifiques au WGS, une information sur le sérotype ou encore le MLST). La plurialité des analyses possibles, ont engendré une multitude de bases de données, permettant aussi bien le partage des séquences entre différents partenaires, que l'analyse bio-informatique en ligne.

Après le séquençage du génome humain, la bactériologie ne fût pas en reste. La taille du génome bactérien, en moyenne 5 Mb, contre 3 400 Mb pour l'Humain, a permis un développement rapide de la génomique bactérienne. Il est maintenant aisé de trouver de nombreuses séquences bactériennes sur les bases de données en ligne.

La génomique, discipline initialement exploitée principalement dans le domaine de la recherche fondamentale, a trouvé sa place au sein des équipes travaillant sur la surveillance (Besser *et al.*, 2018). Les génomes sont utilisés pour identifier et caractériser directement les pathogènes isolés le long de la chaine de production alimentaire. La génomique est un outil très performant dans la gestion des alertes sanitaires. En 2018, l'Organisation Mondiale de la Santé (OMS) a publié un document présentant le WGS comme l'outil ayant la plus haute résolution dans le cadre de la surveillance et de la réponse aux alertes sanitaires (WHO World Health Organization, 2018). Par ailleurs, les souches séquencées dans le cadre de la surveillance, permettent bien souvent d'anticiper certains pics épidémiques ou de répondre dans les meilleurs délais à des alertes sanitaires (Pietzka *et al.*, 2019).

1.6 La génomique en application

Après les étapes classiques d'extraction, de préparation de librairies et de séquençage, dites étapes « *wet-lab* », arrivent les étapes d'analyses bio-informatiques dites de « *dry-lab* ». Les analyses « dry-lab » commencent par la vérification de la pureté des reads et leur normalisation.

Les *reads* sont de courtes séquences de bases nucléotidiques obtenues par le séquençage. De nombreux paramètres doivent être vérifiés à ce stade de l'analyse, comme la couverture du génome (pourcentage de *reads* assemblés qui couvrent le génome), la profondeur de séquençage (nombre moyen de *reads* couvrant une base du génome assemblé) ou encore la vérification de la pureté de la séquence obtenue par *mapping* sur un génome de référence.

La plasticité du génome bactérien offre différentes approches analytiques. En effet, celui-ci est constitué du core génome (ensemble des gènes présents chez toutes les souches d'une espèce = gènes de ménages) ainsi que du génome accessoire (gènes qui ne sont pas présents chez toutes les souches d'une même espèce), le tout formant le pan-génome bactérien (cf. figure 2).

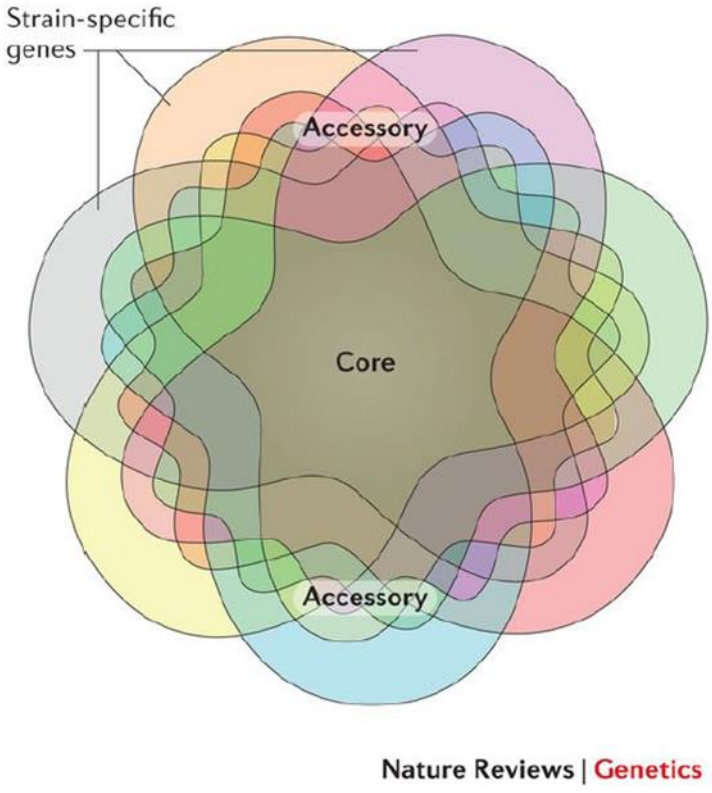


Figure 2 - Représentation graphique de la plasticité du génome bactérien. (Soucy et al., 2015)

Cette variabilité du génome bactérien nous offre la possibilité de mise en œuvre des types d'analyses suivantes :

- Core Genome MLST (cgMLST) : Analyse par séquençage multilocus des gènes du core génome
- Whole Genome MLST (wgMLST) : Basé sur le concept de variation allélique, ce qui signifie que les recombinaisons, les suppressions ou les insertions de positions multiples sont comptabilisées comme des événements évolutionnaires uniques. Cette recherche de variation allélique est effectuée sur le pangénome.
- Whole Genome Single Nucleotide Polymorphism (*SNP calling*) : analyse de l'ensemble des variations (polymorphisme) du génome, entre individus d'une même espèce ou entre isolats d'un micro-organisme.

La résolution analytique de celles-ci est croissante (cf. figure 3), mais utilise les mêmes types de données brutes de séquençage, permettant de multiplier les analyses en fonction des besoins ou de les comparer.

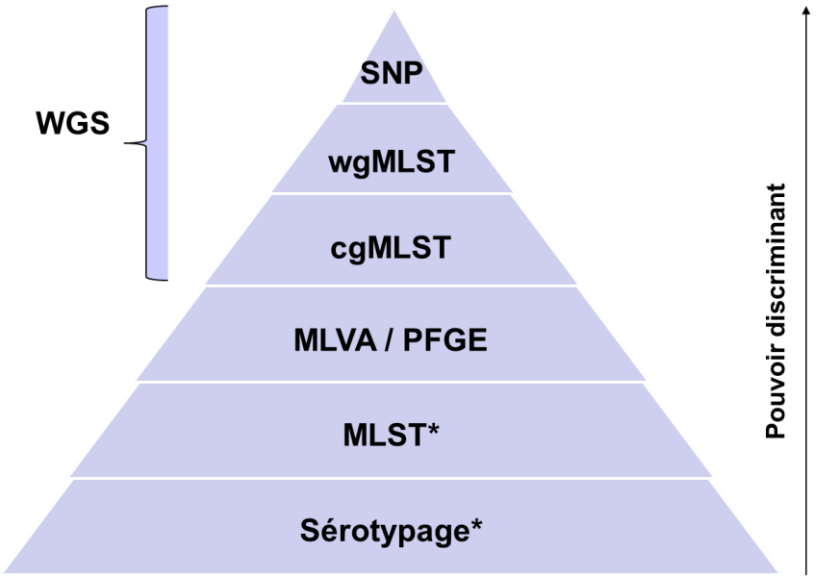


Figure 3 - Pouvoir discriminant des méthodes analytiques utilisées au sein de l'unité SEL

Pourquoi choisir une méthode analytique plutôt qu'une autre ? Les arguments peuvent être nombreux. Selon l'organisme étudié, et donc de la plasticité génomique de celui-ci, les différences de résultats inter-méthodes peuvent être flagrantes. D'autres arguments comme, par exemple, la disponibilité d'un génome de référence pour la réalisation de l'assemblage (reconstruction d'une séquence à partir de reads courts) peut avoir un impact significatif sur les sorties analytiques.

Par ailleurs, pour les méthodes cgMLST et de wgMLST, des schémas sont disponibles (https://www.cgmlst.org/ncs), référençant les gènes d'intérêts (Ruppitsch *et al.*, 2015, Alikhan *et al.*, 2018, Dangel *et al.*, 2019). L'utilisation de ces schémas, pouvant être implémentés aussi bien dans des logiciels commerciaux que dans des outils « faits maison », permettent une standardisation des pratiques inter-laboratoires. Les analyses de *SNP calling* restent plus discriminantes, permettant d'analyser avec précision les variations au sein du génome complet. Le choix du génome de référence reste une étape fondamentale dans les analyses SNP (Pightling *et al.*, 2014). Toutefois, le *SNP calling* repose principalement sur l'utilisation d'outils « faits maison », de type Linux ou Galaxy, développés par les laboratoires.

1.7 Assurance qualité

Le laboratoire a une politique d'assurance qualité qui se base sur les exigences de la norme NF EN ISO 9001 (AFNOR, 2015) et de la norme NF EN ISO 17025 (AFNOR, 2018). Le laboratoire étant accrédité Cofrac pour un certain nombre de méthodes, nous respectons également le LAB GTA 59 (Cofrac, 2019). Les projets tels que celui développé dans ce mémoire, sont soumis au respect du Manuel Qualité Recherche édité par le laboratoire (Annexe 1).

A travers mon projet, j'ai pu commencer à constituer le dossier de validation de méthode en vue d'une accréditation Cofrac des méthodes analytiques de WGS utilisées par l'unité. De ce fait, les tests des différentes méthodes analytiques ont pu mener à la rédaction de méthodes internes et modes opératoires ayant pour but d'aiguiller les bio-analystes dans leurs choix analytiques, normaliser leurs pratiques et de permettre le rendu de rapports d'analyses en bonnes et dues formes, le tout rattaché directement au système qualité de l'Agence.

2 Présentation du projet

L'unité SEL, au sein de laquelle je travaille, a une activité qui s'inscrit dans le cadre des mandats de références portés par le LSAL, sur deux pathogènes majeurs en agroalimentaire, *Salmonella et Listeria*. Dans le contexte des TIAC, des alertes produits et des alertes sanitaires, l'une de nos missions est de mener les investigations génomiques, en vue d'études d'attribution des sources. En parallèle de ces demandes officielles, le laboratoire met en œuvre son expertise en génomique bactérienne pour les besoins de clients et/ou partenaires. L'ensemble du processus analytique doit être maitrisé de la réception de la souche au rendu du rapport d'analyse envoyé aux tutelles et aux clients.

A ce jour, afin de répondre aux sollicitations de nos tutelles aucune méthodologie analytique n'est imposée. Le présent projet va donc permettre de faire un état des lieux des méthodes analytiques disponibles au laboratoire, des outils informatiques utilisés, des méthodologies associées à ces outils et de leur champ d'application. Je vais ainsi définir quelles sont les approches techniques et méthodologiques à privilégier dans le cadre d'évènements sanitaires. Il sera donc nécessaire de mener à bien et valider plusieurs étapes clés permettant de mener à bien ce projet :

Dans un premier temps je vais sélectionner un panel de souches pour chaque pathogène à l'aide des bases de données disponibles au laboratoire. Pour la comparaison de méthode, j'ai privilégié le choix de souches épidémiologiquement reliées, isolées dans le cadre d'évènements sanitaires ayant eu lieu entre 2005 et 2019. Pour *Salmonella*, j'ai choisi les panels de souches liés aux alertes sanitaires publiés respectivement en 2005 et 2018 (Espie *et al.*, 2005, Jourdan-da Silva *et al.*, 2018). Pour *Listeria*, j'ai sélectionné le panel publié par Palma *et. al* en 2020, constitué de souches persistantes présentes dans deux usines. J'ai donc utilisé ces deux panels pour effectuer la comparaison des outils bio-informatiques à notre disposition. A ces souches j'ai ajouté des souches du même sérovar/ST, non reliées épidémiologiquement, issues de la collection du laboratoire de sécurité des aliments.

A travers les premières manipulations, je vais étudier et ajuster les modes opératoires pour l'extraction d'ADN de qualité génomique et pour la préparation des librairies.

Une fois les génomes bruts obtenus, je vais appréhender les outils bio-informatiques disponibles au laboratoire et les schémas analytiques associés (*SNP calling*, cgMLST, analyse du génome accessoire).

En vue de comparer les différents outils bio-informatiques utilisés par le laboratoire je vais mener plusieurs approches comparatives et statistiques.

En vue de mieux comprendre la résurgence de souches impliquées dans des alertes ou encore la formation de biofilms, je vais rechercher les gènes de virulence, de persistance et de résistance à l'aide de bases de données internes ou décrites dans la littérature.

En parallèle des étapes décrites ci-dessus, je vais réviser ou rédiger des modes opératoires et méthodes internes liées aux analyses bio-informatiques réalisées et comparées. La rédaction de ces documents liés au système qualité permettra d'harmoniser les processus d'analyses et d'accompagner l'unité vers l'accréditation Cofrac en 2021.

3 Sélection des panels

La sélection des deux panels bactériens a été réalisée selon deux critères majeurs :

- La sélection d'un panel initial de souches reliées épidémiologiquement pouvant appartenir à une alerte produit, une alerte sanitaire ou à des souches détectées dans un même environnement à multiples reprises. Afin de vérifier la cohérence des panels sélectionnés, j'ai cherché plus spécifiquement des souches étant apparues de manière récurrentes ou sporadiques dans leur environnement. Ainsi, il sera possible de prendre en compte le taux de recombinaison des souches appartenant au même évènement sanitaire et d'apprécier la persistance de la souche.
- Une sélection annexe de souches du même sérovar ou CC, non reliées épidémiologiquement.

La prise en compte de ces deux paramètres, m'a permis de constituer un panel de *Salmonella* appartenant au sérovar Agona et un panel de *L. monocytogenes* du CC204.

3.1 *Salmonella enterica enterica* Agona

A l'aide de la base de données Acteolab gérée par le RS, j'ai pu extraire la totalité des souches de *S.* Agona collectées entre 2001 et 2019. Cela représente 424 souches réparties en sept secteurs majoritaires. La répartition des souches est illustrée en figure 4.

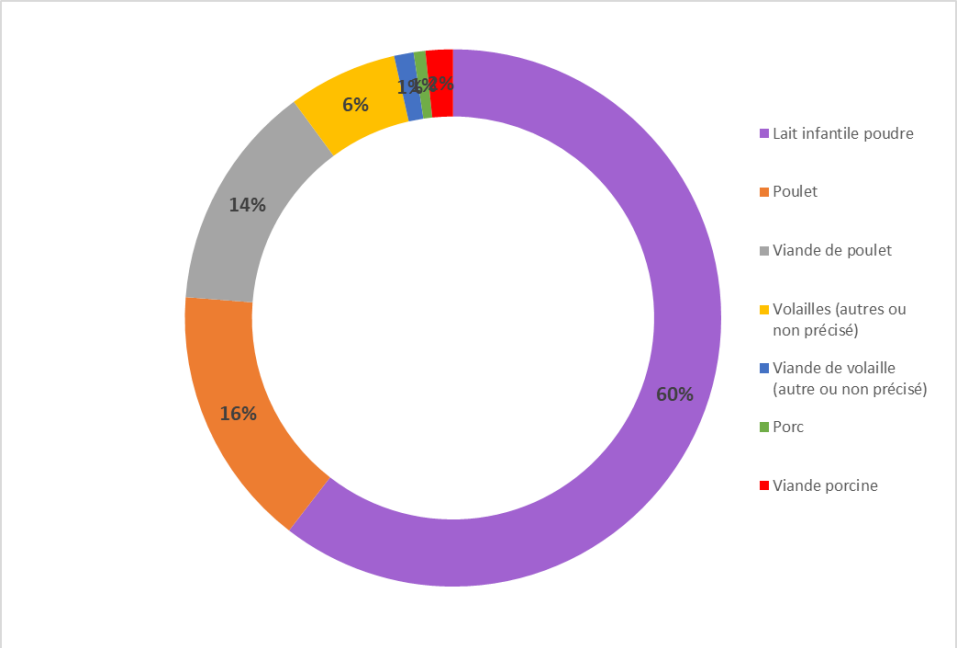


Figure 4 - Répartition par secteur des prélèvements de S. Agona collectés par le RS entre 2001 et 2019

Parmi les souches de *S.* Agona présentent dans Actéolab, 257 sont issues de poudre de laits destinés à la consommation de nourrissons, dont 256 ont été prélevées dans le cadre de deux alertes sanitaires. En mars 2005 puis en décembre 2017, deux alertes sanitaires indépendantes liées à l'ingestion, par des nourrissons, de laits en poudres contaminés par *S.* Agona, ont été reportées (Espie *et al.*, 2005, Jourdan-da Silva *et al.*, 2018). Après vérification, les productions de laits liées aux deux alertes proviennent de la même usine. Seulement huit des 256 souches d'alerte correspondent aux prélèvements réalisés lors de l'alerte de 2005. Parmi les 251 souches de l'alerte de 2017-2018, j'ai dédoublonné les prélèvements de surfaces et les prélèvements alimentaires réalisés au cours de l'alerte. Ainsi, 67 souches liées à la seconde alerte ont été sélectionnées. Une souche humaine isolée dans le cadre de cette même alerte a pu être récupérée afin d'agrémenter le panel.

Dans le but de compléter le panel, j'ai recherché sur une période de dix ans (entre 2009 et 2019) des souches non reliées épidémiologiquement à celles des alertes, issues de filières et de localisations différentes. Parmi les souches présentes dans la base de données, j'ai dédoublonné les souches (par exemple souches prélevées le même jour, dans un même endroit et sur une même matrice) et vérifié les souches disponibles en collection. Ainsi, j'ai sélectionné 31 autres souches de *S.* Agona. Le panel final comporte donc 108 souches de *S.* Agona (Tableau 1). Les données épidémiologiques sont détaillées en annexe 2 et la diversité géographique en fonction du type de matrice est illustré en figure 5.

Tableau 2 - Récapitulatif de la distribution des matrices au sein du panel de S. Agona

Matrices	Nombre de souches analysées
Aliment pour animaux domestiques	1
Crustacé	1
Dindes	1
Lait infantile en poudre	76
Matières premières industrie agroalimentaire	2
Matrice inconnue	3
Plats préparés	2
Poulet (*Gallus gallus*)	6
Selles d'origine humaine	1
Semences germées	2
Viande de canards	1
Viande de cheval	1
Viande de dinde	2
Viande de poulet (*Gallus gallus*)	7
Viande porcine	1
Viande de volaille non précisé	1
Total	108

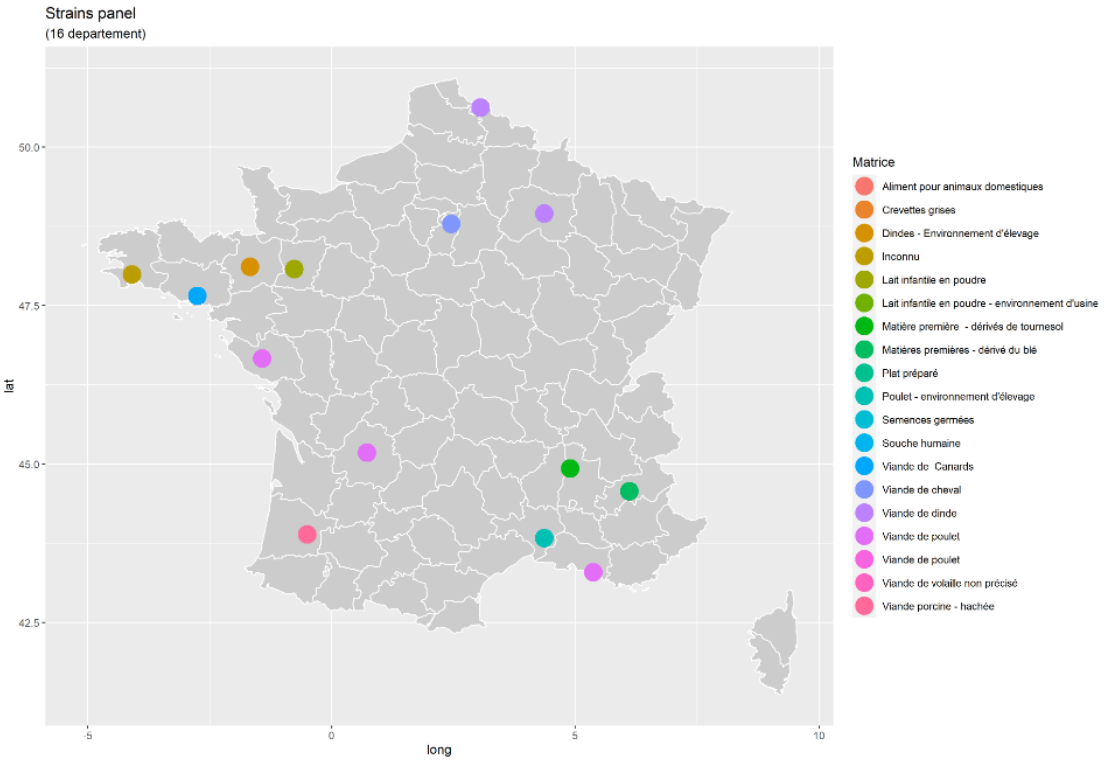


Figure 5 - Répartition géographique des souches constituant le panel S. Agona, en fonction du type de matrice

Les souches dont l'origine est inconnue ou prélevées en dehors de la métropole Française ne sont pas représentées sur cette figure.

3.2 *Listeria monocytogenes* CC204

Dans notre collection de souches de *L. monocytogenes*, collectées à partir de prélèvements réalisés « de la fourche à la fourchette », 21 CC majeurs ont pu être caractérisés par PFGE (Felix et al., 2018). Le CC204 représente en moyenne 2,2% des souches collectées, avec une forte prévalence sur les prélèvements issus de produits de la mer ou de plats cuisinés (figure 6).

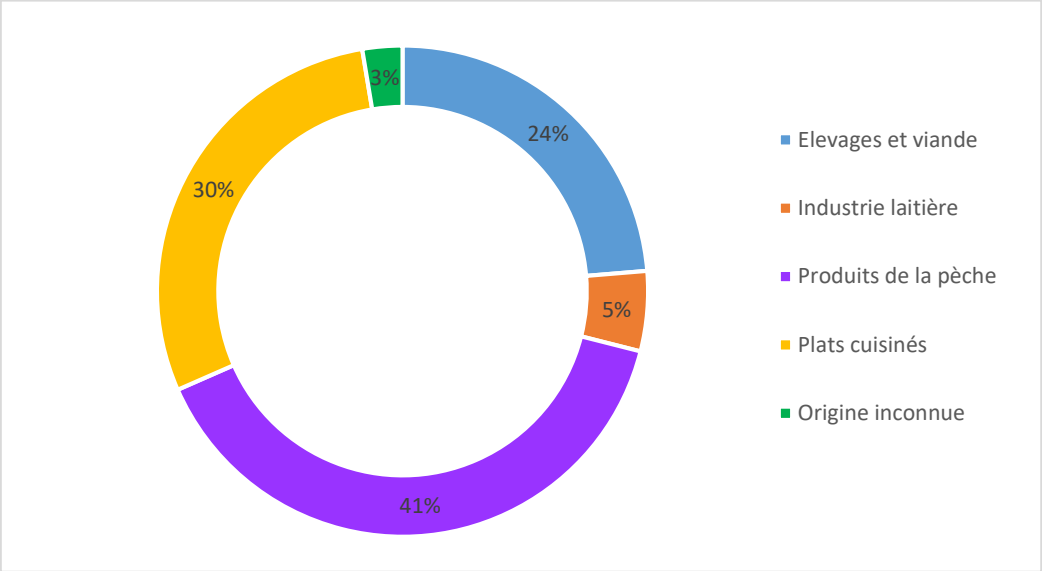


Figure 6 - Répartition par secteur des prélèvements de Listeria monocytogenes CC204 collectés pendant 20 ans par l'unité SEL

Des échantillons contaminés par *L. monocytogenes* CC204 ont été prélevés à plusieurs reprises entre 1998 et 2003 dans deux usines de produits de la mer. (Palma, 2020). La bactérie *L. monocytogenes* est considérée comme persistante si elle est retrouvée dans une même source pendant une période de six mois. Les deux usines présentent ce type de contamination persistante : l'usine A présente des prélèvements positifs entre 1999 et 2000, l'usine B entre 1998 et 2003. Dans les deux entreprises, les différentes surfaces de travail étaient principalement constituées d'acier inoxydable, de PVC ou de polyuréthane, et les opérations de nettoyage et de désinfection étaient effectuées par des prestataires de services avec des désinfectant à base d'ammonium quaternaire et/ou des solutions à base de peroxyde. Des transferts de matières premières ou de produits transformés ont été rapportés à de multiples reprises entre les deux usines.

A ces souches persistantes s'ajoutent 35 souches du CC204, non reliées aux souches des usines précédemment citées (Tableau 3). La sélection a été effectuée à partir de souches collectées par l'unité SEL sur une période de dix ans (entre 2009 et 2019). Après avoir dédoublonné les souches (une seule souche par lieu de prélèvement, date et type de matrice), j'ai sélectionné les souches de façon à représenter le mieux possible la diversité au sein du CC204 et les origines géographiques. Mon panel est donc constitué de 59 souches, dont les données épidémiologiques sont présentées en annexe 3.

Tableau 3 - Récapitulatif de la distribution par matrice des souches de L. monocytogenes (CC204)

Lieu de prélèvement	Matrice	Nombre de souches analysées
Usine A	Saumon fumé	8
	Prélèvement de surface - industrie de la pêche	3
Usine B	Saumon fumé	7
	Prélèvement de surface - industrie de la pêche	6
Aléatoire	Prélèvement de surface - industrie laitière	17
	Selles d'origine humaine	2
	Crudités	1
	Viande de canard	2
	Poisson fumé	1
	Prélèvement de surface	1
	Matrice inconnue	11
Total		59

4 Matériel et méthodes

4.1 Les étapes dites de « wet-lab »

Les étapes dites « wet-lab » correspondent à l'ensemble des activités de paillasses allant de la culture bactérienne jusqu'au séquençage (Figure 7).

L'extraction d'ADN génomique est réalisée à partir de souches bactériennes pures, cultivées sur milieux non sélectifs, dont l'appartenance au genre *Salmonella* ou à l'espèce *L. monocytogenes* a été préalablement confirmée.

J'ai extrait les ADN génomiques de ces souches selon le protocole du kit Wizard® Genomic DNA Purification (Promega, France). Après la réalisation de tests de validation de méthode, j'ai réalisé quelques ajustements à partir du protocole commercial. En effet, j'ai ajusté la densité optique des solutions, et les bouillons de culture ou tampons utilisés, en vue d'obtenir à la fois une quantité et une qualité d'ADN satisfaisante pour les deux pathogènes. Ceux-ci ont conduit à la rédaction d'une instruction technique, conforme aux exigences de notre système qualité.

La concentration d'ADN est mesurée avec un fluoromètre Qubit ® (Invitrogen, Etats-Unis) et le ratio de pureté est évalué avec un spectrophotomètre Nanodrop® (Thermo Scientific, Etats-Unis). Le ratio 260/280, indiquant la pureté en ADN doit être entre 1.6 et 2.0 (idéalement proche de 1,8). Le ratio 260/230, indiquant la présence de solvants, doit être entre 1,6 et 2.2 (idéalement proche de 2.0). Afin d'évaluer l'intégrité des ADN génomiques, on utilise des gels d'agarose de 0.8 %.

Les librairies sont réalisées à l'aide du kit Nextera XT® (Illumina). La quasi-totalité des souches analysées ont été séquencées par un prestataire, avec le kit 300 cycles High Output kit v2 cartridges (c'est-à-dire 800 millions de reads paired-end de 150 bases), sur un séquenceur NextSeq 500). Quelques souches des panels des deux pathogènes étudiés, ont été séquencées sur la plateforme Identypath de l'Anses, disposant d'un séquenceur MiSeq utilisant les mêmes réactifs que le prestataire. Ces séquençages réalisés en « interne » vont permettre à l'unité de démontrer ces capacités à gérer le flux analytique dans son ensemble, permettant de faire accréditer la méthode. Toutefois, le volume analytique n'est pas entièrement absorbable par la simple utilisation d'un séquenceur MiSeq, d'où la nécessité d'avoir recours à un prestataire. Par ailleurs, l'utilisation en direct du séquenceur disponible sur la plateforme, permet, en cas d'alerte sanitaire, de répondre dans un temps très réduit aux sollicitations de nos tutelles.

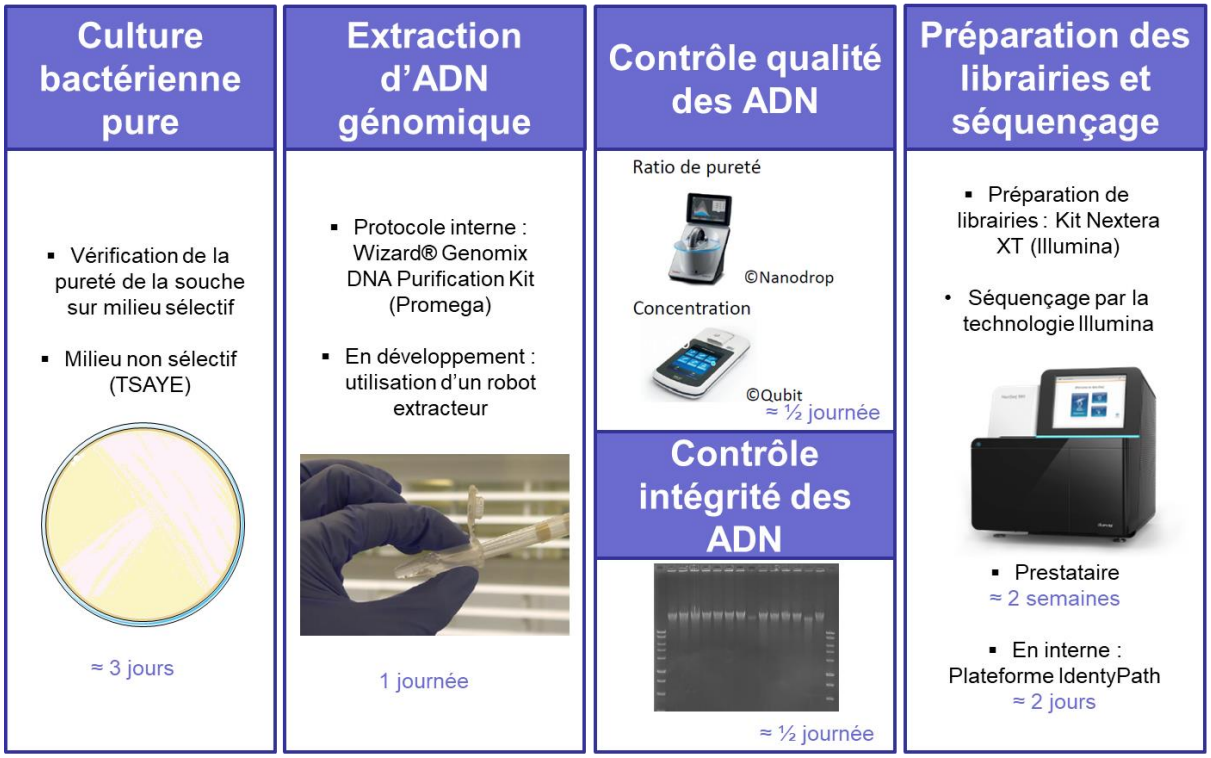


Figure 7 - Processus "wet-lab"

4.2 Analyses génomiques, dites de « dry-lab »

Avant toute analyse génomique, les reads obtenus après séquençage sont normalisés à l'aide du workflow Artwork (version 1) développé en interne par l'équipe de bio-informaticiens (figure 8) (Felten, 2017). Ce workflow vérifie la qualité des données issues du séquençage en réalisant dans un premier temps une étape de contrôle de la couverture des reads en profondeur sur génome de référence. Compte-tenu de l'implémentation constante de reads au sein de notre base de données génomiques, pour des questions de capacité de stockage, la couverture maximale des reads normalisés a été fixée à 30x. Par exemple, un fichier fastq.gz, d'environ 100-150X, occupe 3Go contre 1Go, une fois normalisé par Artwork. Toutefois, la couverture n'est abaissée qu'une fois les erreurs de séquençages contrôlés par l'étape de trimming. Dans le cas d'une couverture en profondeur trop faible, l'assemblage peut donner un résultat insatisfaisant et des régions du génome risquent de ne pas être couvertes, ce qui engendrerait un biais dans la recherche de variants et dans la construction de l'assemblage. Le workflow Artwork inclut ensuite une étape de *trimming* et *mapping*, dont les critères d'acceptabilité sont décrits dans le tableau 4.

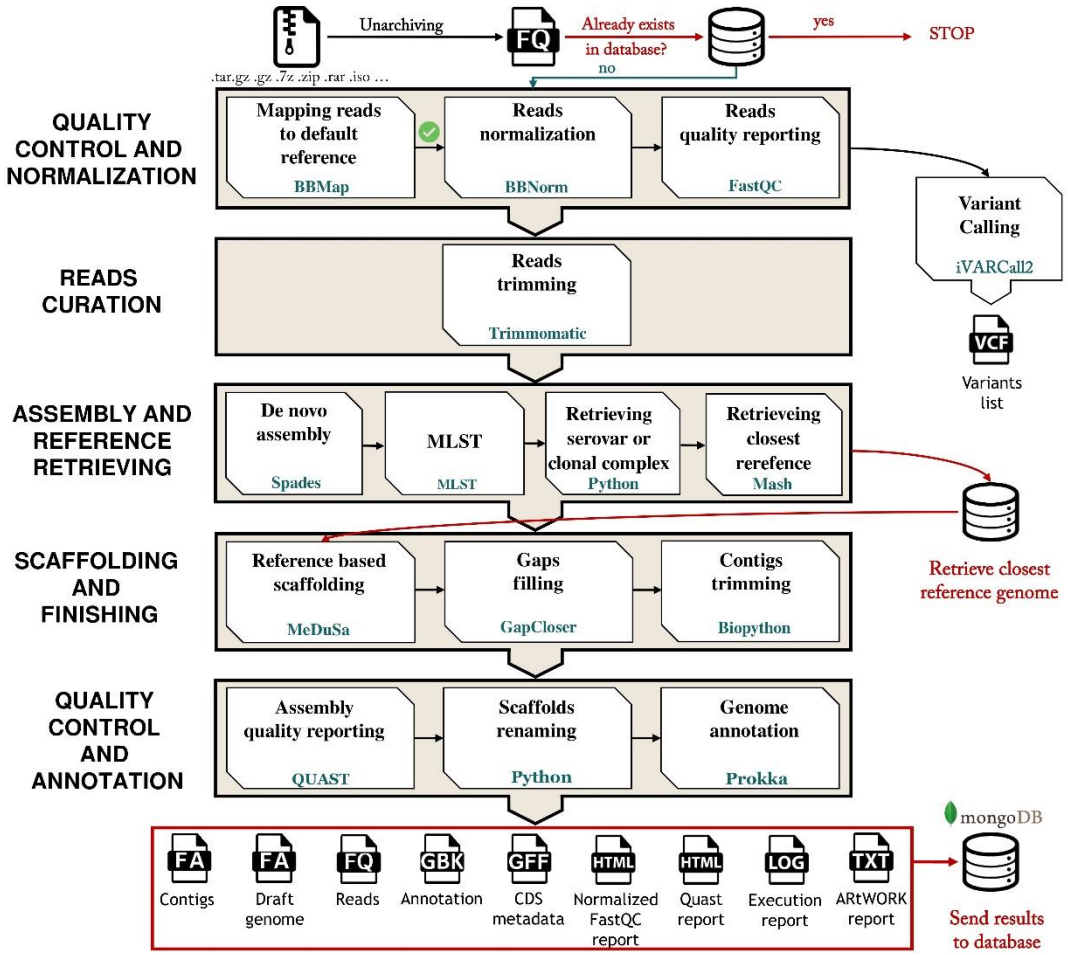


Figure 8 - Workflow ARtWORK – version 1 (source https://github.com/afelten-Anses/ARtWORK)

Tableau 4 -Workflow ARtWORK version 1 : récapitulatif des étapes de contrôle qualité et de normalisation des reads

Etapes	Version	Outils	Seuils d'acceptabilité	Références
Contrôle de la couverture des reads (profondeur)	36.14	BBMap / BBNorm	20X minimum, normalisation à 100X	(Junemann *et al.*, 2014)
Contrôle des données qualité des reads (longueur des reads, N50, %GC…)	0.11.5	FastQC	A l'appréciation	N.A.
Trimming (curation des reads de moins bonne qualité, ou trop courts)	0.33	Trimmomatic	Longueur minimale 50 pb, Phred Score <20	N.A.
Scaffolding	1.6	MeDuSa	Suppression scaffolds <200pb	N.A.
Taux de couverture (Breath coverage)	N.A.	Python	Couverture minimale 80%	N.A.
Assemblage de novo (sans référence)	3.9.1	Spades	N.A.	N.A.
Contrôle des données qualités des assemblages	4.3	Quast	A l'appréciation	N.A.
Recherche de contaminants	0.4	Confindr	≈10% maximum / vérification que l'assemblage couvre à minima 80% du génome de référence	N.A.
Confirmation du genre / espèce / sérovar / CC / ST	2.6	MLST/Python	N.A.	https://pubmlst.org/
Fermeture des gaps	1.6	GapCloser	N.A.	https://github.com/BGI-Qingdao/stLFR_GapCloser#ref

Une fois la normalisation des reads effectuée, il existe une multitude de possibilités en termes de choix d'outils bio-informatiques. Trois outils bio-informatiques sont actuellement utilisés de façon courante au sein de l'unité. Nous espérons être en mesure d'utiliser un quatrième outil, BioNumerics.

Enterobase est une base de données partagée, utilisable en ligne. Celle-ci permet l'analyse et la visualisation des variations génomiques au sein de sept espèces bactériennes dont *Salmonella (Zhou, 2019)*. Cette base permet une comparaison rapide et aisée des profils MLST de tous les génomes soumis. La comparaison des génomes intra et inter-pays est donc possible dans le cadre d'alertes. L'Anses est propriétaire de tous les résultats d'analyses réalisées sur des ressources publiques nationales ou européennes (contrôles officiels, investigation sanitaire), ou lorsqu'un accord signé préalablement avec le laboratoire fournisseur prévoit ces droits d'utilisation (cas du fonctionnement du RS). Les données de sorties d'Enterobase sont spécifiques à la base de données et sont donc non comparables aux schémas d'analyses cgMLST habituellement utilisés par notre laboratoire. Par ailleurs, aucun paramétrage des représentations phylogéniques obtenues via Enterobase n'est possible. De ce fait, je ne comparerai pas, dans le cadre de ce projet, les résultats obtenus entre Enterobase et les autres outils.

Les deux autres outils utilisés, **iVARCall2** et **SeqSphere**, vont pouvoir être étudiés au sein de ce projet, sur les deux pathogènes de notre étude. En parallèle des tests effectués, les modes opératoires correspondant à ces trois outils, ont été rédigés, conformément à notre système qualité.

Le logiciel commercial **BioNumerics**, utilisé historiquement dans l'unité pour les analyses PFGE, a rencontré des problèmes techniques dû à la délocalisation du serveur de calcul dans un *cloud* privé, en vue de sécuriser les données. Il m'a été impossible d'utiliser le serveur de calcul permettant l'analyse bio-informatique des génomes. Bien que je n'ai pas été en mesure d'utiliser le module « WGS tools » de BioNumerics, j'ai décidé de présenter ses caractéristiques afin de les évaluer au même titre que les autres outils. J'espère être en mesure de pouvoir présenter les analyses comparatives lors de ma soutenance orale, si le problème lié à la délocalisation du serveur de calcul sera résolu.

Un récapitulatif des quatre outils bio-informatiques ainsi que leurs paramètres est détaillé en tableau 5. Dans **SeqSphere+**, les paramètres de BLAST utilisés sont les suivants : pourcentage d'identité de 90%, pourcentage de couverture de 100%, taille de la séquence 11 bases et une pénalité de mésappariement. Pour **iVARCall2**, la couverture minimum des reads pour l'analyse de variants est de 30X et la taille minimum des reads pour le *trimming* est de 50 bases.

J'ai également exploité les arbres phylogénétiques générés avec une analyse des bootstraps. C'est à dire, analyse via le pourcentage de présence statistique d'une souche à un branchement donné de l'arbre phylogénétique. Plus la probabilité est forte que le branchement soit juste, selon un modèle d'évolution donné, plus le bootstraps sera proche de 100%. Les analyses ont été conduites suivant plusieurs modèles d'évolutions, en sélectionnant le « meilleur modèle » grâce à l'outil IQ-tree (Nguyen *et al.*, 2015) disponible sous environnement Linux.

Tableau 5 - Tableau des différents outils disponibles et leurs spécificités

Outil	Type d'outil	Version / Lien / GitHub	Pathogènes analysés	Types d'analyses	Type d'assemblage	Schéma MLST */ Algorithme	Phylogénie	Analyses complémentaires
Enterobase	Base de données en ligne	Version 1.1.2 http://enterobase.warwick.ac.uk/	*Salmonella*	cgMLST wgMLST	Assemblage de novo (SPAdes)	Schéma Enterobase (https://pubmlst.org/) 3002 locis Schéma Enterobase (http://enterobase.warwick.ac.uk/download_data)	Oui, Neighbor Joining (NJ) : aucune possibilité de modification d'un point vue paramétrique	Comparaison de profils inter-laboratoires Prédiction du sérovar
iVARCall2 (Independant Variant Calling 2)	Développement interne	Version 1 https://github.com/afelten-Anses/VARtools/tree/master/iVARCall2	*Salmonella et Listeria*	wgSNP + Indel	Assemblage de novo (SPAdes) + Alignement sur un génome de référence	Algorithme haplotypecaller : GATK	Oui, entièrement au choix du bioanalyste Généralement IQtree Annotation des arbres via R ou iTOL	N.A.
BioNumerics (Applied Math)	Logiciel commercial	Version 7.6.3	*Salmonella et Listeria*	cgMLST wgMLST wgSNP	Analyses sur les reads (assembly free calling) ou sur les génomes assemblés (assembly based calling).	*Salmonella* : Schéma Enterobase 3002 locis *Listeria* : Schéma Moura et al. , 2016 1748 locis (https://bigsdb.pasteur.fr/listeria/) Algorithme BioNumerics non détaillé	Oui, Maximum Parcimony (MP) ou Maximum Likelihood (ML) avec possibilité de modification d'un point de vue paramétrique	MLST (+ autres possibilités en fonction de l'installation de modules BioNumerics)
SeqSphere+ (Ridom)	Logiciel commercial	Version 6.0.2	*Salmonella et Listeria*	cgMLST	Assemblage de novo (SPAdes ou Velvet) ou assemblage sur génome de référence	*Salmonella* : Schéma Enterobase 3002 locis *Listeria* : Schéma Ruppitsch et al. , 2015 1701 locis (https://www.cgmlst.org/)	Oui, NJ avec possibilité de modification d'un point de vue paramétrique	MLST Détermination du sérovar moléculaire de *L. monocytogenes*

*Les schémas cgMLST/wgMLST implémentés dans les logiciels commerciaux sont au choix de l'utilisateur. Le tableau 5 fait objet des schémas utilisés par l'unité.

Références associées : iVARCall2 (Felten et al., 2017), SPAdes (Bankevich et al., 2012), Velvet (Zerbino & Birney, 2008)

4.3 Détection des évènements de recombinaison

Les bactéries présentent trois modes de recombinaison génétique : la conjugaison, la transformation ou la transduction.

J'ai cherché les évènements de recombinaison au sein des panels de souches sélectionnés. Il a été démontré que la plupart des bactéries subissent des évènements de recombinaison fréquents, des parties de leur génome sont remplacés par des séquences correspondantes d'autres bactéries (Smith et al., 1993, Brown et al., 2003, Jolley et al., 2005, Didelot et al., 2007).

Dans cette optique, l'outil ClonalFrameML (Didelot & Wilson, 2015), utilisant l'inférence par le maximum de vraisemblance (Maximum Likelihood (ML)) a été utilisé avec les paramètres par défaut de la commande, sous environnement Linux. Les données générées ont ensuite été traitées avec le script R « cfml_results » (https://github.com/xavierdidelot/ClonalFrameML/blob/master/src/cfml_results.R) afin d'obtenir des représentations graphiques. Cet outil permet d'intégrer les évènements de recombinaisons homologues souvent négligé avec une inférence uniquement réalisée via le ML. Les phylogénies générées sont donc corrigées, en prenant en compte les localisations des recombinaisons pour chaque branche de l'arbre. La relation entre chaque souche et le clustering des souches sont donc vérifiés afin de produire l'arbre le plus représentatif de l'évolution et de la radiation des souches.

4.4 Comparaison des méthodes

Dans le but de comparer les méthodes, plusieurs approches de comparaisons statistiques et de comparaisons d'arbres phylogéniques ont été réalisées à l'aide de RStudio.

- **Comparaison des matrices de distances générées** par les différents outils par analyse du nombre de SNP/allèles de différences entres les clusters
- Comparaison des matrices de distances obtenues par les différents outils, via le **test de Mantel**, à l'aide de R Studio et du package « ade4 ». Le test de Mantel est une approche par régression permettant d'identifier les corrélations entre deux matrices de distance. Par conséquent, si les distances génomiques obtenues à l'aide d'un outil bio-informatique par rapport à un autre sont systématiquement plus élevées mais présentent une bonne corrélation, le test de Mantel donne alors un coefficient de corrélation élevé.
- **Comparaison par appariement de deux arbres phylogéniques**. Les packages R « ade4 » pour la représentation graphique des fonctions, « ggtree » pour la manipulation d'arbres phylogéniques, « dendextend » et « phangorn » pour l'appariement de deux arbres et « phytools » sont utilisés pour visualiser et comparer les données (Henri et al., 2017).
- Une **analyse statistique** sera réalisée en vue de tester l'application d'un script R développé en interne : script « matrix2association » (Radomski *et al.*, 2019a), en vue de répondre aux alertes sanitaires. Trois tests non paramétriques : Wilcoxson, Kolgmogorov ou Kruskal sont testés en parallèle, dans le but de déterminer l'appartenance à des souches d'intérêt à un cluster épidémique, sans réaliser les étapes parfois longue et fastidieuses de génération d'arbres phylogéniques. Ce test statistique permet une analyse alternative à la classique analyse phylogénétique. Via les panels exploités, je chercherai le nombre minimal de souches reliées épidémiologiquement afin d'obtenir un résultat fiable dans le cas d'application de ces tests aux alertes sanitaires. En tenant compte du fait que le test repose sur l'implémentation de souches reliées comme base de travail, à travers les résultats obtenus, je définirai les atouts et limites de celui-ci dans le cadre de son application des contextes sanitaires majeurs. Les essais effectués par mes collègues (Radomski *et al.*, 2019b), ont démontré la nécessité d'implémenter le script d'a minima quatre souches reliées entre elles afin de déterminer si les souches analysées appartiennent au même cluster ou non.
- **Comparatif des outils à travers une étude de leurs caractéristiques**, fonctionnalités et paramétrages.

4.5 Recherche des gènes de virulence, de résistance et de persistance

Dans le cadre d'alertes sanitaires, ayant un impact important sur la population, il peut également être intéressant d'analyser le génome accessoire pour la recherche des gènes de virulence, de résistance et de persistance.

Grâce à l'utilisation d'un outil développé en interne, GENIAL (https://github.com/p-barbet/GENIAL), lui-même utilisant l'outil de screening ABRIcate (https://github.com/tseemann/abricate). Les paramètres de BLAST utilisés sont de 80 de minimum de couverture et 90% de minimum d'identité. J'ai pu effectuer un screening sur quatre bases de données :

- Recherche des 21 **îlots de pathogénicité** spécifiques à *Salmonella* (SPI : *Salmonella Pathogenicity Island*), à partir d'une **base développée en interne**, selon la littérature (Annexe 4). Les SPI sont présents dans des régions génomiques spécifiques et sont généralement acquis par transfert horizontal. L'ensemble de ces gènes de virulence permettent la colonisation de l'hôte et l'invasion systématique de *Salmonella (Marcus et al., 2000)*.
- Utilisation de la base **VFDB** permettant de rechercher 3200 gènes de **virulence** (Chen *et al.*, 2005)
- Utilisation de la base **Resfinder** permettant de rechercher 2700 gènes de **résistance aux antibiotiques** (Zankari *et al.*, 2012). Ces gènes sont d'un intérêt particulier au genre *Salmonella,* particulièrement touché par de nombreuses résistances acquises. Celles-ci sont dues au phénomène de pression de sélection résultant de l'utilisation intensive des antibiotiques.
- Utilisation de la base **BacMet** permettant de rechercher 753 gènes de **résistance aux produits d'entretiens et aux métaux lourds** (Pal *et al.*, 2014). La recherche de ces gènes, peut permettre de faire le lien de cause à effet entre la présence d'une bactérie persistante dans un environnement et l'environnement en lui-même. Ces études sont d'un intérêt majeur pour *L. monocytogenes*, dont la présence de ce type de gènes de résistance peut conduire à la formation de biofilms.

5 Résultats

5.1 Séquençage

Les ADN génomiques extraits étaient de qualité suffisante pour un séquençage. Les librairies préparées en interne étaient de bonne qualité, similaire aux exigences de notre offre de marché concernant les librairies par notre prestataire. Les séquençages réalisés en interne et en externe m'ont permis d'obtenir des génomes avec une couverture comprise entre 100 et 500X. Nous n'avons donc observé aucune différence de qualité de génomes entre les séquençages réalisés en interne et en externe.

5.2 Lancement des analyses bio-informatiques

Une fois la normalisation et le contrôle qualité des reads effectués, l'ensemble des deux panels ont été analysés successivement par les deux outils iVARCall2 et SeqSphere+.

Les analyses de *SNP calling* nécessitent l'implémentation d'un génome de référence pour les étapes de *mapping*. Pour les analyses de *SNP calling* réalisées sur les souches de S. Agona, la référence utilisée est *Salmonella enterica* subsp. *Enterica* Agona strain 24249 (CP006876.1). Pour les souches de *Listeria monocytogenes*, la référence utilisée est *Listeria monocytogenes* strain EGD-e (AL591824.1). L'assemblage des reads réalisé par SeqSphere est réalisé sur les mêmes souches de références que celles utilisé pour les analyses de *SNP calling*.

SeqSphere possède des modules de phylogénie intégrés. Pour les besoins des comparaisons de méthodes, l'analyse iVARCAll2 a été poussée jusqu'à la génération d'un arbre phylogénétique, grâce à l'outil IQ-tree. Cet outil permet de déterminer le meilleur modèle d'évolution en Maximum Likelihood en fonction du panel étudié, et de calculer les bootstraps associés.

Une fois les calculs menés, j'ai donc obtenu des sorties comparables pour l'ensemble des outils (fichiers FASTA, matrices de distances et arbres phylogénétiques au format newick).

5.2.1 *S.* Agona

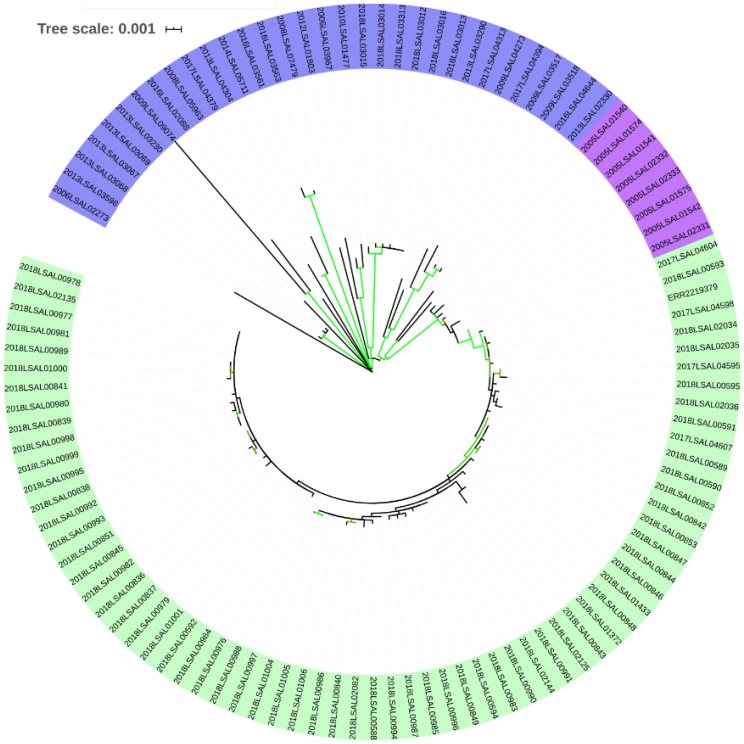


Figure 9 - Arbre phylogénétique circulaire du panel S. Agona, en ML, généré à partir des calculs iVARCall2.

L'arbre a été annoté à l'aide de iTOL. En violet, sont représentées les souches liées à l'alerte de 2005, en vert les souches liées à l'alerte de 2017 et en bleu les souches non reliées épidémiologiquement aux alertes. Les bootstraps compris entre 80 et 100% sont représentés en vert sur les branches de l'arbre. Souche de référence : Salmonella enterica subsp. *Enterica* Agona *strain 24249 (CP006876.1).*

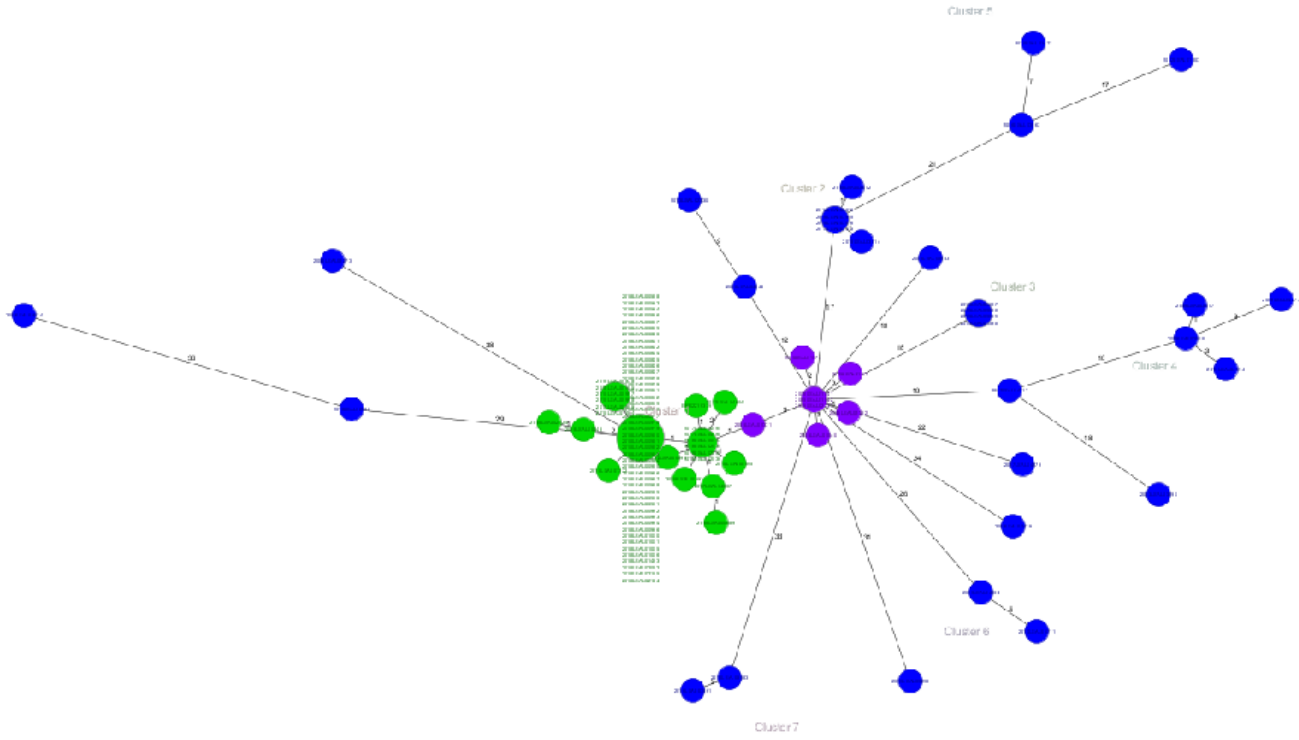


Figure 10 - Minimum spanning tree, du panel S. Agona généré avec SeqSphere (schéma cgMLST Enterobase http://enterobase.warwick.ac.uk/).

En violet, sont représentées les souches liées à l'alerte de 2005, en vert les souches les souches liées à l'alerte de 2017 et en bleu les souches non reliées épidémiologiquement aux alertes. Les longueurs de branches correspondent aux différences alléliques entre deux clades.

L'arbre phylogénétique généré à partir des calculs iVARCall2 (figure 9), permet de mettre en évidence un embranchement spécifique au cluster comportant l'ensemble des souches de *S.* Agona des deux alertes sanitaires. Au sein de cet embranchement, sont représentés deux sous-clusters, l'un en violet pour l'alerte de 2005 et le second en vert pour l'alerte de 2018. L'ensemble des souches non reliées épidémiologiquement aux alertes, constituent plusieurs embranchements de l'arbre, distinctement espacés des alertes sanitaires. L'embranchement du cluster de souches liées aux alertes présente une valeur de bootstraps de 100%. La topologie de l'arbre suggère un lien très fort entre ces souches. Les souches de 2005 enracinent la branche comportant les souches épidémiques de 2017-2018. Les analyses de *SNP calling* ont démontré que les souches des deux alertes sanitaires ont en moyenne 5 SNP d'écart entre elles, contre environ 35 SNP d'écart avec les souches non reliées épidémiologiquement. L'ensemble des souches liées aux alertes présentent en moyenne 5 SNP de différence, alors qu'elles ont été prélevées à treize ans d'écart. En prenant en compte les données obtenues grâce l'analyse *SNP calling* (tableau 6), on peut observer que les souches de l'alerte sanitaire de fin 2017, début 2018, ont en moyenne 3 SNP de différence. Ces souches ont été collectées sur une période de deux mois et issues de poudre de laits ou de prélèvements de surface au sein de l'usine. La souche humaine (ERR2219379) a un seul SNP de différence avec trois souches (2017LSAL04598, 2018LSAL00593 et 2018LSAL02034). Ces trois souches ne présentent aucun SNP de différence. De façon intéressante, les souches 2017LSAL04598 et 2018LSAL00593 ont été prélevées le même jour alors que la souche 2018LSAL002034 a été prélevée deux mois après. Seule la souche humaine a été séquencée sur la plateforme de séquençage de l'Institut Pasteur, cependant la même technologie a été utilisée : Illumina.

De la même manière, le *minimum spanning tree* généré à l'aide de SeqSphere (figure 10) met en évidence des clusters bien définis pour les souches liées à l'alerte sanitaire de 2005 et pour les souches de l'alerte de 2017. Les distances alléliques lisibles sur les branches de l'arbre nous permettent de visualiser au maximum cinq allèles de différence au sein du cluster de souches épidémiques de 2005 et au maximum six au sein du cluster de souches de 2017.

5.2.2 *L. monocytogenes* CC204

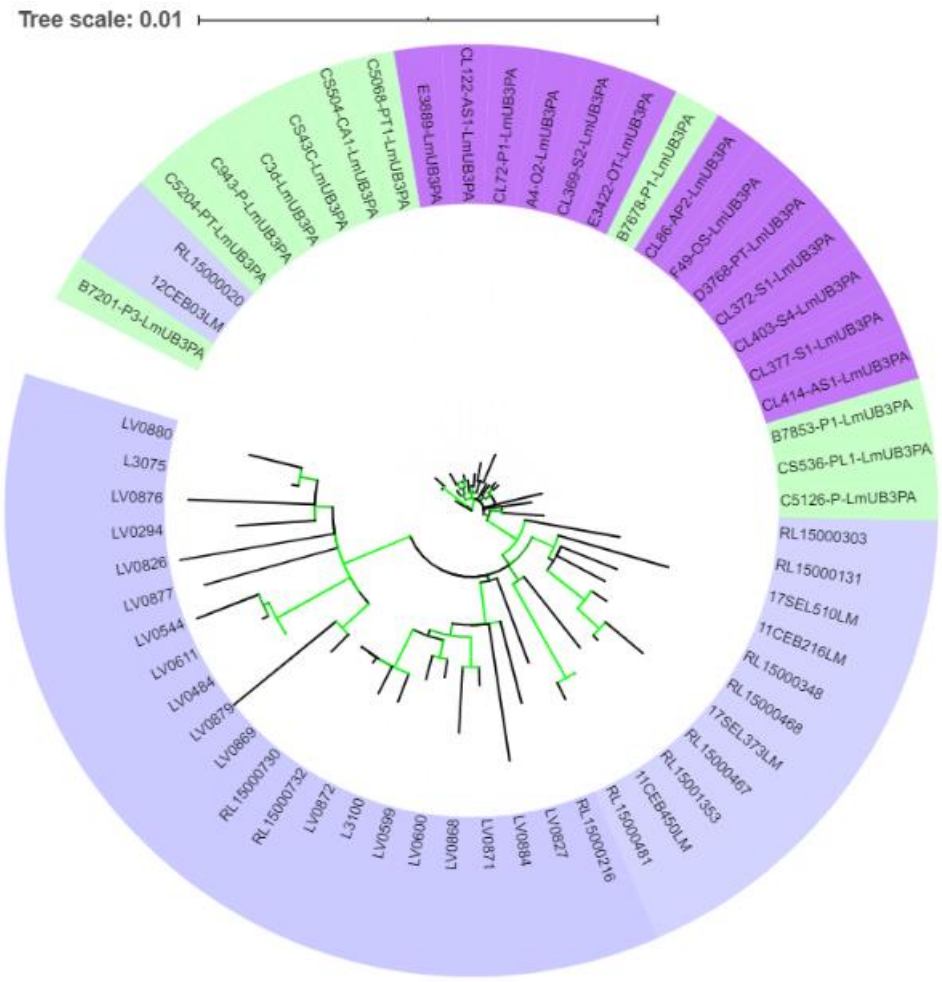


Figure 11 - Arbre phylogénétique circulaire de L. monocytogenes CC204, en ML, généré à partir des calculs iVARCall2.

L'arbre a été annoté à l'aide de iTOL. En vert sont représentées les souches persistantes provenant de l'usine A, en violet les souches provenant de l'usine B et en bleu les souches non reliées épidémiologiquement aux usines A et B. Les bootstraps compris entre 80 et 100% sont représentés en vert sur les branches de l'arbre. Souche de référence : Listeria monocytogenes strain EGD-e (AL591824.1).

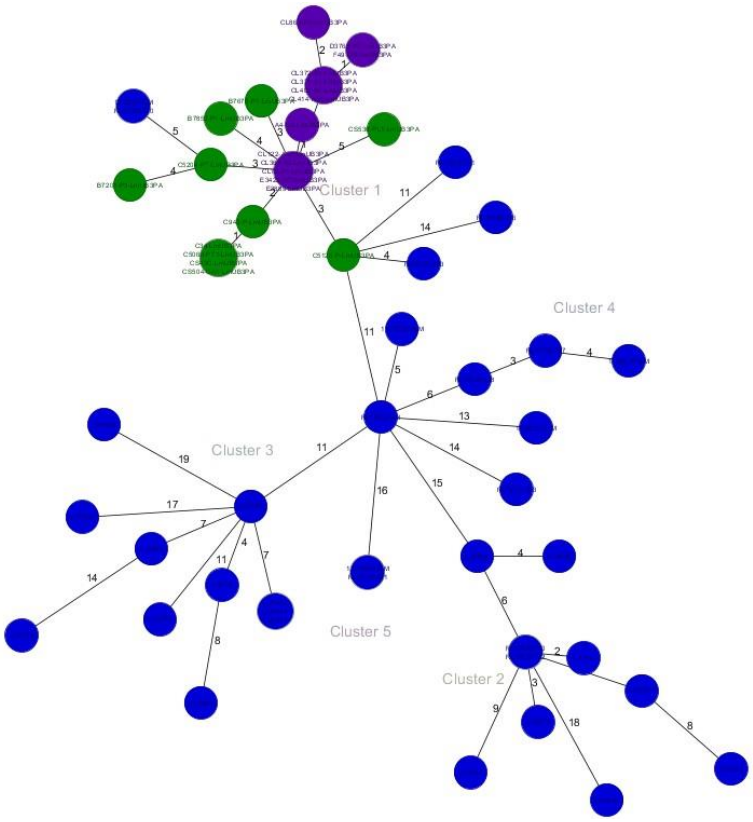


Figure 12 - Minimum spanning tree, circulaire du panel de L. monocytogenes CC204 généré avec SeqSphere (schéma cgMLST Ruppitsch et al. , 2015)

En vert sont représentées les souches persistantes provenant de l'usine A, en violet les souches provenant de l'usine B et en bleu les souches non reliées épidémiologiquement aux usines A et B. Les longueurs de branches correspondent aux différences alléliques entre deux clades.

Les deux arbres phylogénétiques des *L. monocytogenes* CC204 générés à l'aide d'iVARCall2 (figure 11) et SeqSphere (figure 12) permettent de mettre en évidence un même cluster dans lequel l'ensemble des souches des usines A et B sont regroupées. Les souches des usines A et B sont mélangées au sein du cluster, corroborant l'hypothèse d'échanges de produits agro-alimentaires entre les deux usines. Sur l'arbre généré à partir de l'analyse SNP, les bootstraps aux nœuds des clades sont suffisamment robustes (>80%). Deux autres souches Françaises,12CEB03LM et RL1500020, éloignées de 2 SNP l'une de l'autre, sont incluses au sein du cluster de souches persistantes, entre les souches B7201-P3-LmUB3PA et C5204-PT-LmUB3PA qui sont à une distance comprise entre 7 et 11 SNP d'écart. Nous n'avons pas d'information épidémiologique concernant la souche RL15000020, mais la souche 12CEB03LM est issue d'un échantillon de saumon fumé, comme les souches des usines A et B. Par ailleurs, il est intéressant de remarquer, que ces souches enracinent la totalité de l'arbre en ML. Les analyses de *SNP calling* ont démontré que les souches des usines A et B ont en moyenne 6 SNP d'écart entre elles, contre environ 32 SNP d'écart avec les souches non reliées épidémiologiquement.

5.3 Détection d'évènements de recombinaison

Afin de valider les deux panels, j'ai lancé l'outil ClonalFrameML en sortie de processus iVARCall et IQtree.

5.3.1 Le panel de *S.* Agona

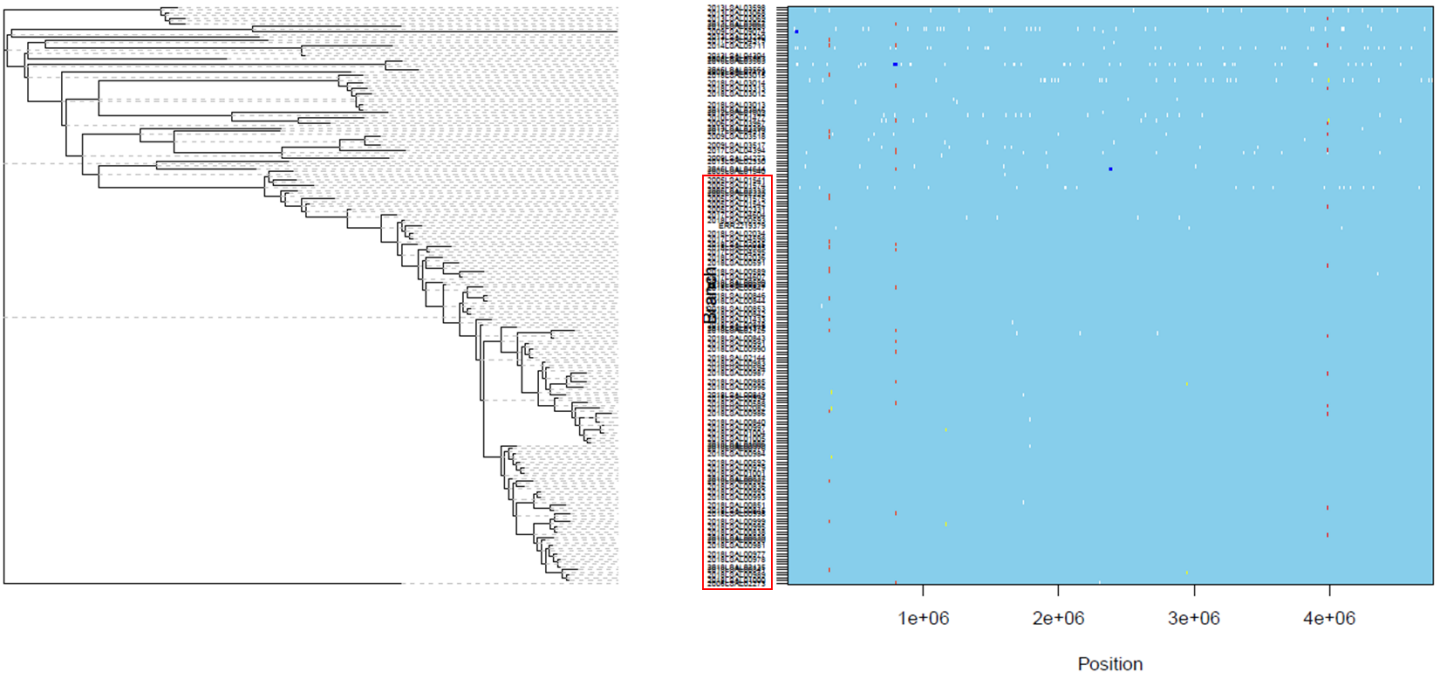


Figure 13 - Représentation graphique de la recombinaison des S. Agona*, à l'aide de l'application de ClonalFrameML.*

Les sites non polymorphiques sont indiqués en bleu clair, les régions bleues foncées représentent les recombinaisons. Les sites polymorphiques sont représentés avec un panel de couleur allant du blanc au rouge, permettant d'apprécier l'augmentation du degré d'homoplasie. Encadrées en rouges, les souches des deux alertes à *S.* Agona de 2005 et 2018.

Les résultats obtenus avec ClonalFrameML pour le panel de *S.* Agona sont illustrés dans la figure 13. J'ai observé que sur les 108 génomes étudiés, seuls trois présentent un évènement de recombinaison (en bleu foncé sur ce type de figure), sur de très courtes séquences, entre 34 et 714 bases. Aucune souche liée aux alertes sanitaire ne présente de recombinaison. Quelques évènements polymorphiques (homoplasies) sont observés. Les souches de *S.* Agona issues des alertes sanitaires, encadrées en rouge sur la figure, sont celles présentant le moins d'homoplasie. Toutefois, je remarque que celles-ci sont toujours globalement situé sur les trois mêmes positions.

La recombinaison présente sur la souche 2016LSAL04644 est une courte séquence de 34 bases étant retrouvée dans un gène régulateur de l'adhésion aux cellules par la fibronectine. La seconde recombinaison sur la souche 2017LSAL04379, d'une longueur de 714 bases correspond à une portion d'un gène codant pour une protéine, pour laquelle aucune fonction biologique n'a été décrite. Sur la souche 2009LSAL09074, j'ai observé une recombinaison de 715 bases, correspondant à un gène codant pour l'oxalacetate decarboxylase permettant la régulation des pompes Na+.

5.3.2 Le panel de *L. monocytogenes* CC204

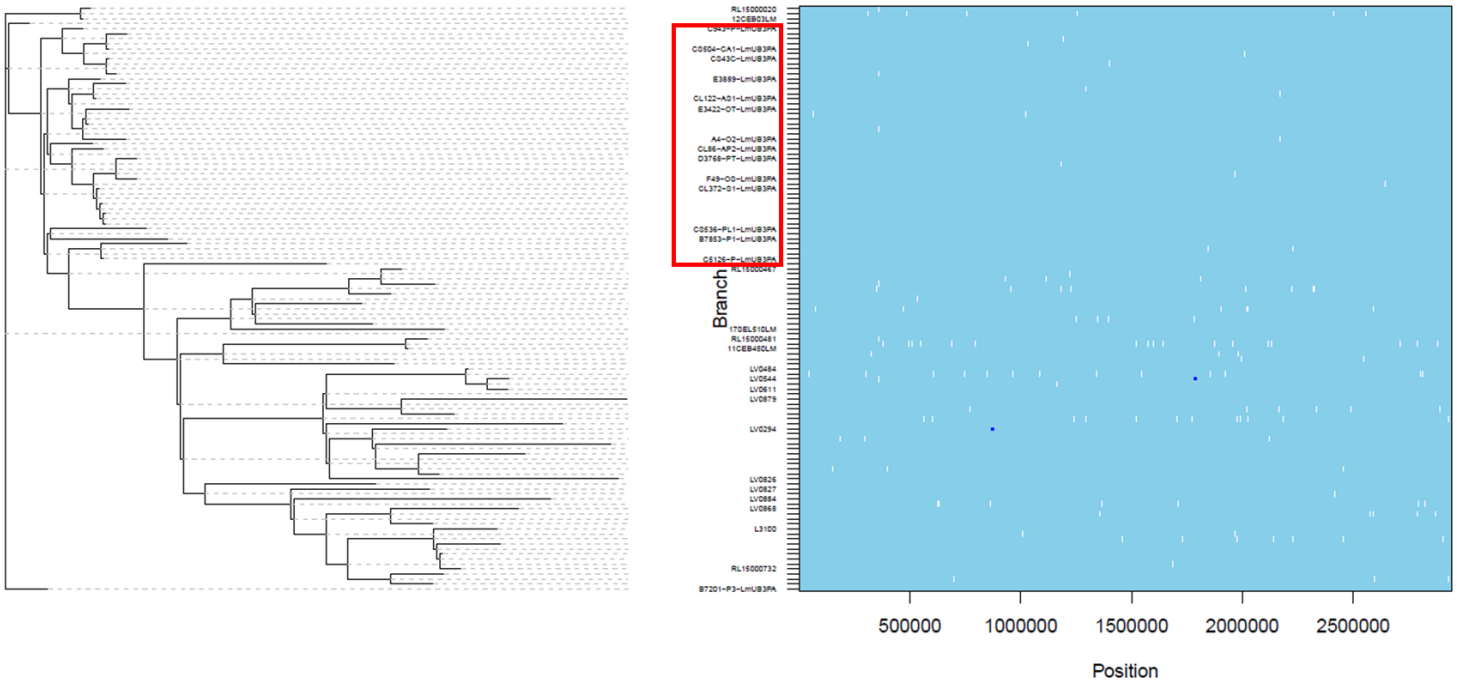


Figure 14 - Représentation graphique de la recombinaison des L. monocytogenes CC204, à l'aide de l'application de ClonalFrameML.

Les sites non polymorphiques sont indiqués en bleu clair, les régions bleues foncées représentent les recombinaisons. Les sites polymorphiques sont représentés avec un panel de couleur allant du blanc au rouge, permettant d'apprécier l'augmentation du degré d'homoplasie. Encadrées en rouges, les souches des usines A et B, d'où proviennent les souches persistantes.

Les résultats obtenus avec ClonalFrameML pour le panel de *L. monocytogenes* CC204 sont illustrés dans la figure 14. Seuls deux évènements de recombinaison sont observables sur le panel étudié, mais aucun n'apparait sur le panel de souches reliées épidémiologiquement. Comparativement aux autres souches du panel, les souches reliées épidémiologiquement présentent moins de sites polymorphiques. Les deux recombinaisons apparentes sont sur les souches LV0294 et LV0544, sur de très courtes séquences respectivement de 66 et 78 bases. La première recombinaison, sur la souche LV0294 est retrouvée au sein d'un gène codant pour le transport du calcium par les ATPases. La souche LV0544 porte quant à elle une recombinaison de 78 bases retrouvée dans un gène codant pour une oxydoréductase.

5.4 Comparaison des outils

5.4.1 Normalisation et contrôle qualité

Lors de la rédaction du Certificat de Capacité à la Recherche, rédigé en fin de première année, des comparaisons d'arbres phylogénétiques, ont permis de mettre en évidence de grosses variations entre les outils lorsque les reads sont normalisés ou non. Les différences relevées peuvent être expliquées en partie par la taille des génomes analysés : les reads normalisés n'ont pas la même taille que des reads bruts de par les étapes de *trimming* et *scaffolding* non réalisées, ou réalisés selon des paramétrages logiciels différents. Par exemple, iVARCAll2 élimine tous les reads de taille

inférieure à 50 pb alors que SeqSphere+ permet l'élimination des reads inférieurs à 200 pb (Ridom, 2019).

J'ai effectué une analyse des données FastQC des reads bruts (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Il apparait que les reads bruts ont en moyenne une couverture de 150 à 300X. Lors des étapes de normalisation, la couverture est abaissée à 100X, permettant à la fois de gagner en taille pour le stockage, mais également rendre les analyses plus rapides. Sur l'ensemble des analyses réalisées sur les reads normalisés et non normalisés, la durée de calcul nécessaire à la sortie des données a pu être comparée. Il faut, en moyenne, cinq fois plus de temps pour obtenir les données issues de reads non normalisés.

5.4.2 Comparaison des matrices de distance par analyse de distance SNP et allélique

Une première comparaison d'outils repose sur la comparaison des matrices de distances obtenues. iVARCall permet une analyse *SNP calling*, alors que SeqSphere offre une analyse cgMLST (allélique). D'après Pightling *et al.,* le seuil de définition d'un cluster par analyse de *SNP calling* est de 21 SNP de différence (Pightling *et al.*, 2018). Cette étude a été menée par le Centre pour la sécurité alimentaire et la nutrition appliquée (CFSAN), faisant partie de l'Agence américaine des produits alimentaires et médicamenteux (FDA), sur trois pathogènes *Salmonella enterica*, *Listeria monocytogenes* et *Escherichia coli*. D'après les données de l'EFSA, le seuil pour les analyses cgMLST est de 7 allèles de différence, cette valeur est également utilisée par les LRUE des deux pathogènes étudiés. Cela nous amène donc à pouvoir extrapoler un facteur trois afin d'établir une corrélation entre les distances obtenues par cgMLST et *SNP calling*. En prenant en compte les matrices de distances obtenues lors des analyses de nos deux panels, j'obtiens un ratio du même ordre grandeur. Entre les souches épidémiques de S. Agona 2018LSAL00986 et 2018LSAL00988, il y a trois SNP de différence et un allèle de différence. Ces souches sont issues de prélèvements de surfaces réalisés le même jour dans la même usine. De même, pour le panel de *Listeria*, les souches A4-02-LmUBUPA et CL369-S2-LmUB3PA présentent le même ratio d'une différence de trois SNP pour un allèle.

Je suis donc en mesure de faire une corrélation entre les résultats obtenus via les matrices de distance des deux outils.

Afin d'observer une tendance sur l'ensemble des résultats et plus particulièrement sur les souches reliées épidémiologiquement, j'ai calculé, à l'aide d'un script linux, les moyennes des nombres de SNP ou allèles de différence au sein des clusters. Il est alors possible d'apprécier la cohérence des résultats obtenus entre l'analyse *SNP calling* et celle en cgMLST.

- *Salmonella* Agona

Tableau 6 - Moyenne des SNP et allèles de différences par clusters

	iVARCall (SNP)	**SeqSphere** cgMLST (différence allélique)
Cluster épidémique 2005-2018	5,09	Estimé* 1,70
Cluster épidémique 2018	3,38	Estimé* 1,13
Panel complet	35,33	Estimé* 11,77
Cluster épidémique 2005-2018	Estimé* 5,16	1,72
Cluster épidémique 2018	Estimé* 1,92	0,64
Panel complet	Estimé* 45,27	15,09

*Par convention, il est considéré qu'il existe un facteur 3 entre les analyses SNP et cgMLST (Pightling et al., 2018)

- *Listeria monocytogenes* CC204

Tableau 7 - Moyennes des SNP et allèles de différence par clusters

	iVARCall (SNP)	**SeqSphere** cgMLST (différence allélique)
Souches persistantes (Usine A +B)	6,46	Estimé* 2,15
Panel complet	31,63	Estimé* 10,54
Souches persistantes (Usine A +B)	Estimé* 8,37	2,79
Panel complet	Estimé* 48,93	16,31

*Par convention, il est considéré qu'il existe un facteur 3 entre les analyses SNP et cgMLST (Pightling et al., 2018)

Les résultats obtenus entre les deux méthodes sont du même ordre de grandeur.

5.4.3 Comparaison des matrices de distances par le test statistique de Mantel

Pour les deux pathogènes, le test statistique de Mantel (Mantel & Fleiss, 1980) a été réalisé dans le but de comparer les matrices de distances obtenues via iVARCall2 et via SeqSphere. Le principe du test repose sur une régression. Plus le r^2 obtenu est proche de 1, plus la corrélation entre les matrices est élevée. La comparaison entre les deux matrices est jugée satisfaisante lorsque le r^2 est supérieur à 0,7, avec une P <0,05.

- *Salmonella* Agona

La comparaison des matrices iVARCall2 et SeqSphere sur le panel de *Salmonella* a permis d'obtenir un r^2 de 0,87 et P de $2,2^{e-16}$ (Figure 15). La corrélation des résultats est donc jugée satisfaisante, permettant de conclure que les matrices de distances obtenues entre les analyses cgMLST et *SNP calling* obtenues par les deux outils sont comparables.

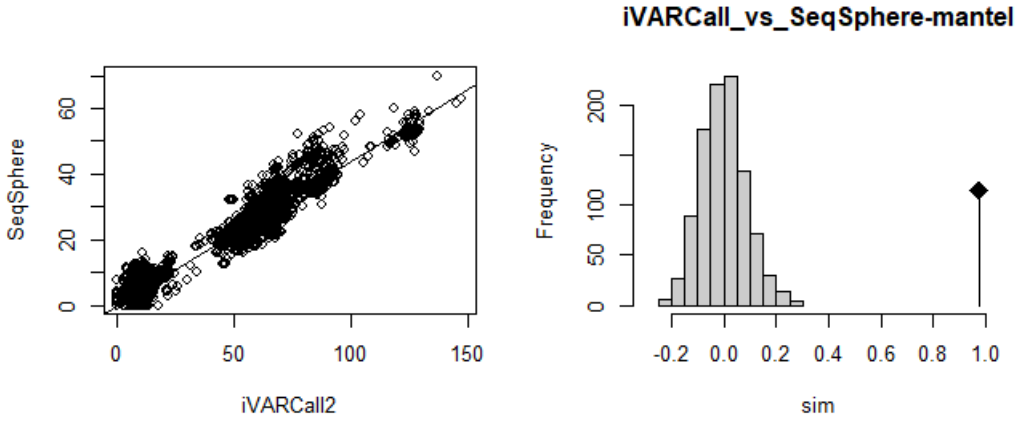


Figure 15 - Données graphiques du test de Mantel (régression linéaire et plot) pour la comparaison des matrices de distances générées par iVARCall2 et par SeqSphere pour le panel de S. Agona.

Données obtenues à l'aide de R et du package ade4

➢ *Listeria monocytogenes* CC204

La comparaison des matrices iVARCall2 et SeqSphere sur le panel de *Listeria* a permis d'obtenir un r^2 de 0,95 et P de $2{,}2^{e\text{-}16}$ (Figure 16). La corrélation des résultats est donc jugée satisfaisante, permettant de conclure que les matrices de distances obtenues entre les analyses cgMLST et *SNP calling* obtenues par les deux outils sont comparables.

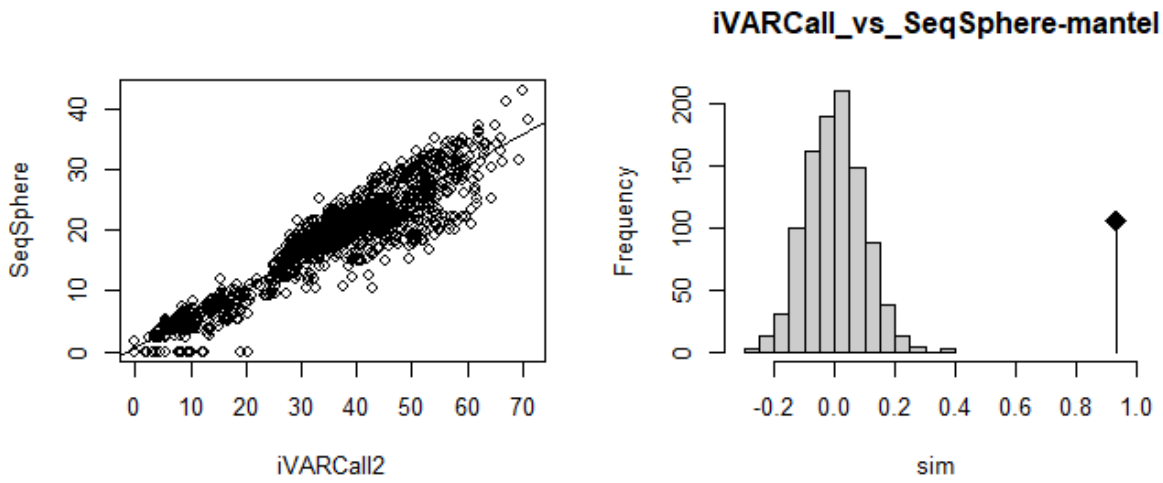


Figure 16 - Données graphiques du test de Mantel (régression linéaire et plot) pour la comparaison des matrices de distances générées par iVARCall2 et par SeqSphere pour le panel de L. monocytogenes.

Données obtenues à l'aide de R et du package ade4.

5.4.4 Comparaison des arbres phylogénétiques

L'une des comparaisons marquantes était de vérifier la cohérence entre les arbres obtenus suivant le schéma analytique conventionnel, ici iVARCall2 suivi de IQ-tree, et du même arbre retraité après intégration des données de recombinaisons via ClonalFrameML.

- *Salmonella* Agona

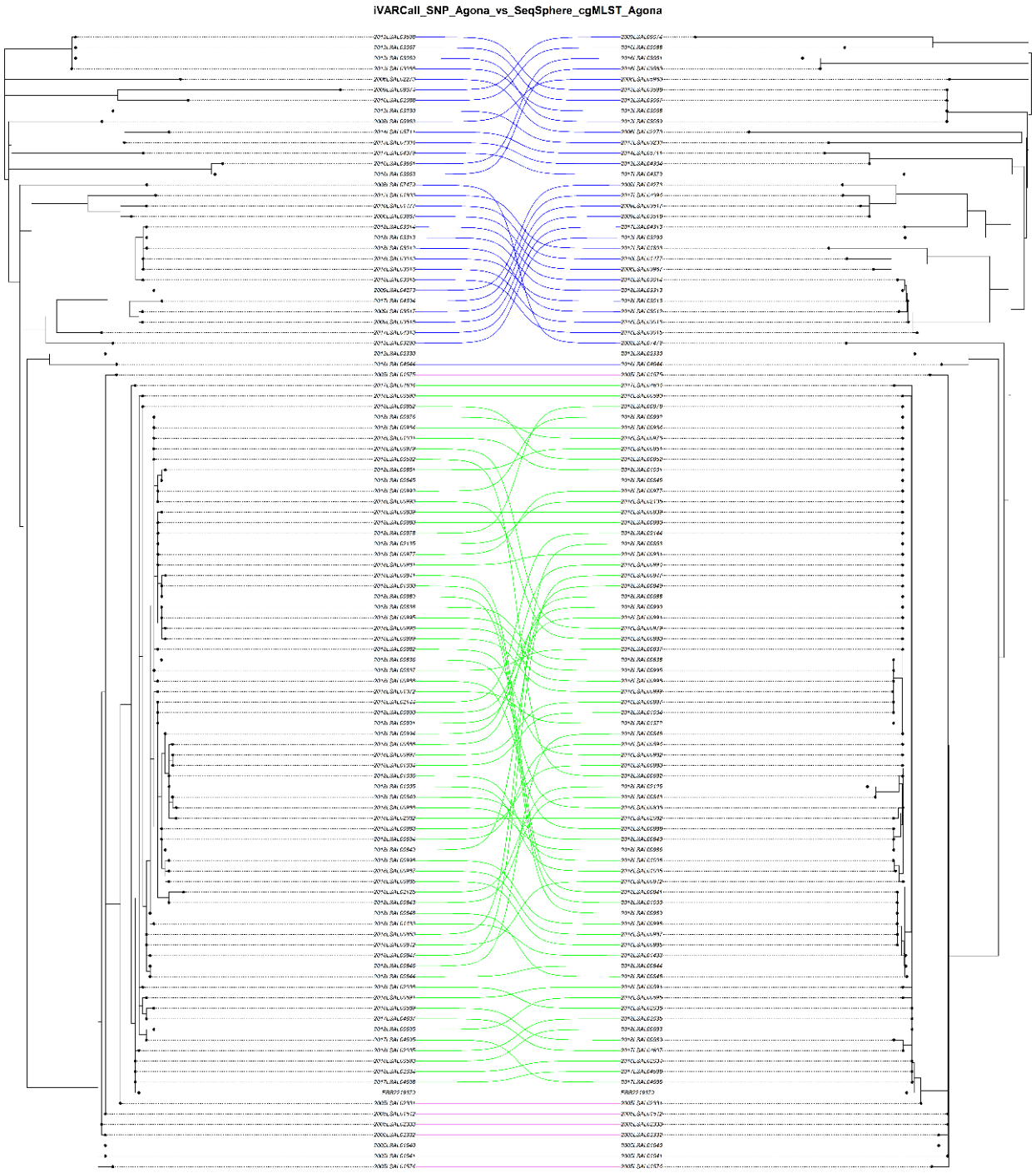


Figure 17 - Comparaison d'arbre phylogénétiques générés via iVARCall 2 et SeqSphere.

A gauche l'arbre phylogénique issu de l'analyse SNP calling par iVARCall2 puis IQ-tree, à droite l'arbre issu de l'analyse cgMLST via SeqSphere+

En bleu, les liens entre les souches non reliées épidémiologiquement aux souches des deux alertes sanitaires ; en violet, les liens entre les souches de l'alerte sanitaire de 2005 ; en vert, les liens entre les souches de l'alerte sanitaire de 2018.

Les variations de positionnement des souches sur les arbres phylogénétiques (Figure 17) sont principalement observées sur les souches de l'alerte sanitaire *S.* Agona, souches possédant en moyenne cinq SNP de différence. Dans l'arbre généré par iVARCall2, les embranchements les plus profonds de ce cluster génomique sont soutenus par une valeur de bootstrap de 100%, alors que les boostraps au sein du cluster sont compris entre 3 et 100%, du fait de la proximité génomique de ces souches (Figure 6). La position des souches au sein du cluster est aléatoire Leur position au sein de cet embranchement peut donc être négligé. J'ai remarqué que les souches très proches génétiquement sont regroupées dans le même cluster dans les deux arbres, bien que le positionnement du cluster a changé au sein de l'arbre. Les groupes soutenus par les bootstraps les plus faibles sont ceux dont la position varie entre les arbres.

➢ *Listeria monocytogenes CC204*

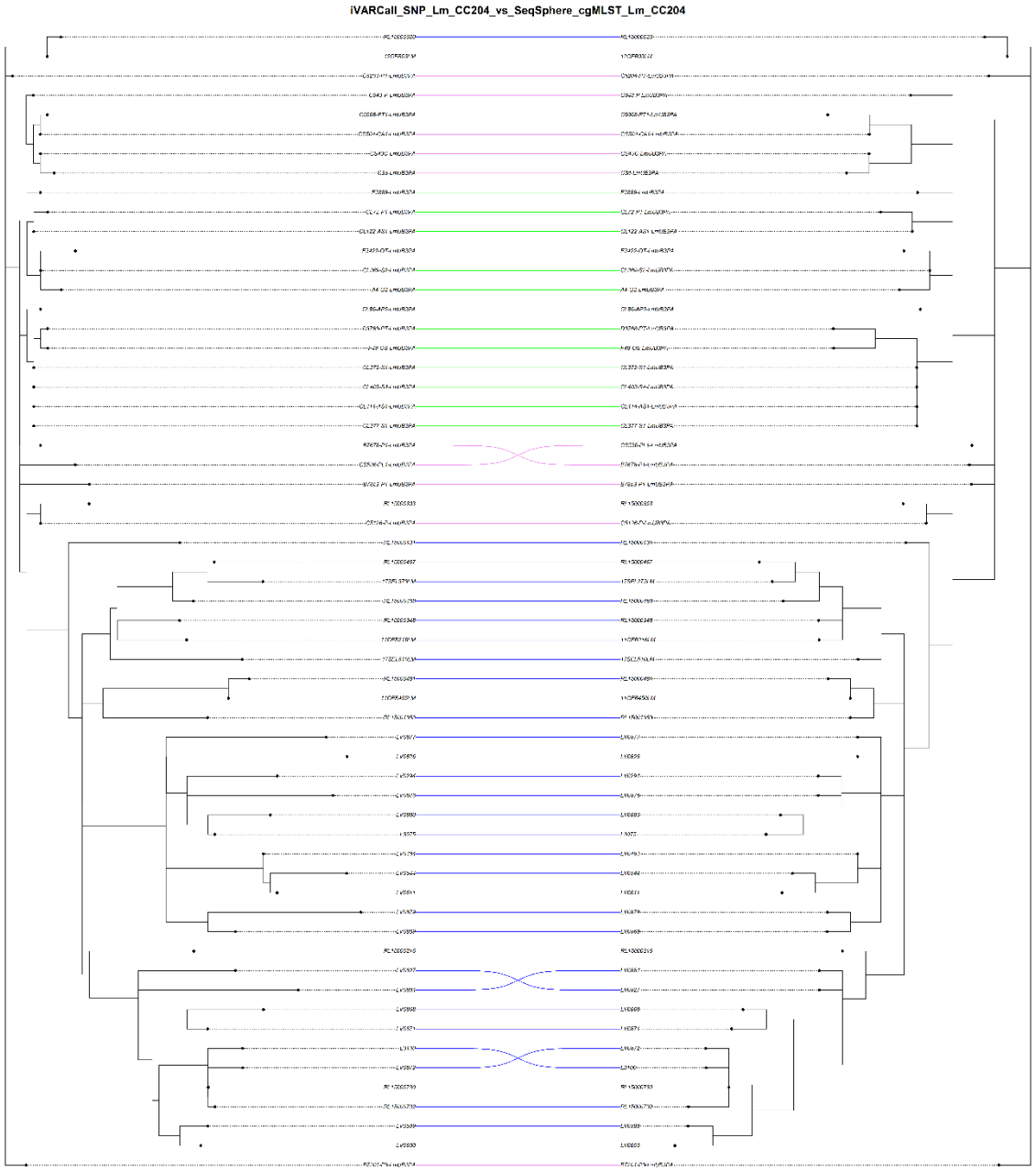


Figure 18 - Comparaison d'arbre phylogénétiques générés via iVARCall 2 et ClonalFrameML.

Pour *Listeria monocytogenes*, entre l'analyse SNP et l'analyse SeqSphere+, je remarque une conservation quasi stricte des positions et de la longueur des branches. La typologie de l'arbre est conservée rendant la phylogénie inter-comparable en fonction de la méthode analytique. Seules trois inversions sont visibles, sur des embranchements externes non supportés par des bootstraps élevé (<80%).

Une comparaison similaire, par appariement d'arbres a été réalisée entre les résultats obtenus par iVARCall2 et ClonalFrameML, en fonction des deux panels (Annexe 5). Sur les deux comparaisons, les positions sont strictement conservées. La longueur des branches reste proportionnelle entre les deux analyses, en cohérence avec le peu d'évènements recombinogènes observés au paragraphe 5.3.

5.4.5 Utilisation du script « matrix2association » permettant le clustering de souches dans le cadre d'alertes sanitaires

L'application du test « matrix2association », en contexte réel d'alerte sanitaire, peut être abordée de différentes manières. Afin de répondre au mieux à cette problématique, il est nécessaire de se mettre en situation d'alertes sanitaires. Quelles sont les données dont nous pouvons disposer et les questions scientifiques associées ? En cas de déclaration d'alerte, un certain nombre de patients sont infectés par un même sérovar/ST/CC et nous devons faire une étude d'attribution de source en vue de trouver l'aliment incriminé. Cette étude est accompagnée de données épidémiologiques (département, type de produits consommés par les malades…) et parfois par une suspicion d'un aliment en particulier grâce aux enquêtes sanitaires réalisée par la MUS ou SpF. Il est également courant que nos études révèlent une contamination chez un producteur/fabriquant lié à une ancienne alerte. Il est alors nécessaire, en vue d'éradiquer la contamination, de déterminer si la source de celle-ci est liée à un même clone bactérien persistant ou non. Ce cas peut aussi être appliqué pour une contamination croisée entre producteur ou usines de fabrication, où des échanges de matières premières ont eu lieu.

En vue de répondre à ces questions, le test statistique doit être réalisé dans les cas suivants :

- Déterminer si les souches isolées récemment appartiennent à un même cluster que des souches plus anciennes (persistance dans un environnement).
- Réaliser une étude d'attribution de source, en ayant à sa disposition les génomes épidémiques humains.

Les fichiers « associated » et « nonassociated », générés par le test permettent de visualiser facilement quelles sont appartiennent au cluster ou non.

- *Salmonella* Agona

En ne prenant que les huit souches de l'alerte de 2005 comme base pour le test « matrix2association », il n'a pas été possible de récréer un cluster contenant à la fois les souches de l'alerte de 2005 et de celle de 2018.

En sélectionnant aléatoirement au sein du panel de *S.* Agona, des souches à la fois du cluster épidémique de 2005 et de celui de 2018, j'ai essayé de reconstituer l'ensemble du cluster épidémique (78 souches) grâce au test statistique. Les tests ont été effectués à partir de quatre souches considérées comme reliées, conformément à la publication (Radomski *et al.*, 2019b). J'ai ensuite augmenté progressivement le nombre de souches considérées reliées. Afin de reconstituer environ 90% du cluster déterminé par phylogénie, à l'aide d'au moins deux des trois tests statistiques, il a été nécessaire d'implémenter neuf souches « reliées ». En prenant en considération le nombre de souches habituellement exploitées lors d'une alerte, la définition de neuf souches reliées (en prenant en compte les données épidémiologiques) semble particulièrement élevée.

Dans le but de nous mettre en situation d'alerte sanitaire, j'ai ajouté au panel des 108 souches de *S.* Agona, neuf souches humaines supplémentaires provenant de l'alerte sanitaire de 2018 (une souche humaine faisant déjà partie du panel). Le test statistique a donc été lancé en définissant les dix souches humaines comme souches « reliées ». Les résultats obtenus sont non satisfaisants, ne permettant pas de relier les souches humaines directement aux souches agro-alimentaires. De nombreuses souches non reliées aux épisodes sanitaires sont également inclus comme étant reliées par l'ensemble des trois tests lancés en parallèle.

- *Listeria monocytogenes CC204*

Le panel de *L. monocytogenes* est composé de 24 souches persistantes, prélevée sur plusieurs années, dont onze proviennent de l'usine A et treize de l'usine B. Jusqu'ici non testé sur un panel de *L. monocytogenes*, nous souhaitions lancer le test statistique dans deux cadres :

- Voir si, pour une usine distincte, en ne prenant que des souches prélevées dans le début de la campagne, nous arrivons à clustériser les dernières souches prélevées.
    - Compte-tenu du nombre de souches prélevées dans les usines (onze dans l'usine A, treize dans l'usine B) le test a été réalisé en définissant quatre à cinq souches comme étant « reliées », ces souches correspondant aux premiers prélèvements réalisés. Les résultats obtenus ont été similaires, que l'on implémente quatre ou cinq souches. Les résultats sont différents entre les trois schémas statistiques. Soient ils intègrent vingt-trois souches, soit la totalité du panel.
- Mettre en évidence les échanges de matières premières au sein des deux usines, via l'utilisation du test. Pour cela des souches des deux usines sont intégrées comme « souches reliées » afin de voir si le test est en capacité de les mettre en corrélation.

  Les essais ont été réalisés en prenant entre quatre et dix souches implémentées comme « reliées ». Les résultats varient en fonction du nombre et du type de souches sélectionnées. A partir de six souches « reliées », le cluster comprenant les deux usines apparait reconstitué à environ 90%.

5.4.6 Comparaison des outils sur le plan paramétrique et fonctionnel

Outil	Type d'analyse	Avantages	Inconvénients
iVARCall2	cgSNP wgSNP + Indel	- Utilisation gratuite sous Linux/Unix - Analyse à haut pouvoir discriminant - Besoin d'un génome de référence - Pipeline développé en interne, pour lequel les paramètres sont maitrisés, en fonction des besoins par pathogènes et selon la littérature - Outil sous Linux : système d'exploitation "open source" gratuit - Possibilité d'analyses complémentaires infinies via l'installation d'outils développés sous Linux/Unix - Outil rapide d'utilisation si installé sur un cluster de calcul conséquent	- Nécessite d'apprendre le langage en ligne de commande "bash" pour l'exploitation de scripts sous Linux - Utilisation dépendante du cluster de calcul disponibles, dont la charge de travail et le coût sont importants
BioNumerics	cgMLST wgMLST wgSNP	- Diversité de types d'analyses incorporé au package "WGS tool" important - Gestion de bases de données aisé, avec incorporation de données épidémiologiques et phénotypiques - Possibilité de création et de connexion des bases données avec d'autres organismes - Implémentation des schémas cg/wgMLST au choix - Possibilité de réalisation de la normalisation intégré au logiciel	- Logiciel commercial nécessitant l'achat de licence - Calculs dépendants du Calculation Engine (Cloud) (maintenance en cours, géré par Applied Math) - Vocabulaire utilisé dans BioNumerics propre à Applied Math : besoin d'adaptation pour comprendre et réaliser l'ensemble des paramétrages - Génération d'arbre phylogénétiques uniquement graphique : pas de fichier .newick pour analyses complémentaires
SeqSphere+	cgMLST	- Gestion de bases de données aisée, avec incorporation de données épidémiologiques et phénotypiques -Transparence du paramétrage du logiciel - Implémentation des schémas cgMLST au choix selon le microorganisme étudié - Possibilité de réalisation de la normalisation intégré au logiciel (avec choix d'un génome de référence) - Détermination du sérotype moléculaire et/ou du CC et/ou du ST intégré - Outil rapide d'utilisation si installé sur un serveur de calcul conséquent	- Logiciel commercial nécessitant l'achat de licence - Pas de calcul de bootstraps possible via la module phylogénie

5.4.7 Recherche des gènes virulence, résistance et persistance

- *Salmonella* Agona

Pour *Salmonella* Agona, j'ai pu réaliser la recherche de gènes d'intérêt dans quatre bases différentes : SPI, VFDB, Resfinder et BacMet.

La base SPI, développée en interne selon la littérature permet la recherche de 21 ilots de pathogénicité (SPI), spécifique à *Salmonella*. Seulement 6 d'entre eux ont été détectés parmi les *S.* Agona analysées, avec un pourcentage de spécificité supérieur à 90% (tableau 8).

Tableau 8 - Ilots de pathogénicité détectés, grâce à la base SPI, sur le panel de S. Agona et leurs rôles respectifs

Ilots de pathogénicité détectés	**Rôle**	**Référence**
SPI-1	**Présent chez toutes les *S. enterica***, permet la pénétration dans les cellules non phagocytaires, rôle dans la régulation des systèmes de sécrétions	(Sevellec, 2018)
SPI-2	**Présent chez toutes les *S. enterica***, code pour une protéine permettant la survie et la multiplication intracellulaire	
SPI-4	Code pour un système de sécrétion de type I permettant la colonisation de l'intestin de l'hôte	
SPI-8	Code pour deux protéines d'immunité aux bactériocines	(Singh, 2018)
SPI-9	Code pour un facteur d'adhésion aux cellules épithéliales	(Sevellec, 2018)
SPI-16	Porteur des gènes impliqués dans la modification de l'antigène O permettant aux Salmonelles d'échapper au système immunitaire	(Singh, 2018)

Parmi les ilots détectés, seuls quatre d'entre eux ont été acquis et peuvent jouer un rôle dans la spécificité de certains sérovars à être plus pathogènes que d'autres. SPI-1 et -2 sont en effet communs à toutes les *S. enterica*.

Les résultats obtenus sur les trois autres bases exploitées sont détaillés dans le tableau 9.

Tableau 9 - Gènes/SPI détectés sur les souches des deux alertes sanitaires à S. Agona de 2005 et 2018

Base de donnée	Analyse générale pour l'ensemble du panel *S.* Agona	Gènes/Ilot détectés spécifiques des souches de S. liées aux alertes sanitaires	Rôle	Mécanisme et/ou type d'acquisition	Référence
Resfinder	Profils similaires sur l'ensemble du panel, à l'exception de **2 gènes** retrouvés spécifiquement sur les souches liées aux alertes sanitaires	aac(6')	Résistance aux aminoglycosides (famille des aminosides)	Mécanisme d'acquisition de la résistance : inhibition de la synthèse protéique. Présence du gène dans un **intégron** du SPI-1	(Doublet *et al.*, 2004)
		fosA7	Résistance à la fosfomycine (famille des β-lactames)	Mécanisme d'acquisition de la résistance : inhibition de la synthèse de la paroi bactérienne. Le plus souvent, acquisition par un **plasmide** ou un transposon	(Rehman *et al.*, 2017)
VFDB	Profils similaires pour l'ensemble du panel de S. Agona. Aucun gène spécifique n'a été mis en évidence sur les souches liées aux alertes sanitaires en comparaison au reste du panel.				
BacMet	Profils similaires sur l'ensemble du panel, à l'exception de **2 gènes** retrouvés spécifiquement sur les souches liées aux alertes sanitaires	silABC	Résistance à l'Argent	Acquisition par un **plasmide**. Utilisation de produits d'entretiens contenant des ions Ag+ comme biocide	(Gupta *et al.*, 1999)
		pbrAR	Résistance au Plomb	Acquisition probable par un **plasmide** provenant de *Ralstonia metallidurans*	(Borremans *et al.*, 2001)

- *Listeria monocytogenes*

Concernant *Listeria monocytogenes* CC204, trois bases ont été testées : VFDB, Resfinder et BacMet. Les résultats d'analyses sont reportés dans le tableau 11.

Tableau 10 - Gènes détectés sur les souches provenant des usines A et B

Base de données	Analyse générale pour l'ensemble du panel *L. monocytogenes*	Gènes/Ilot détectés spécifiques aux souches persistantes des usines A et B	Rôle	Type d'acquisition	Références
Resfinder	Aucune résistance acquise détectée sur l'ensemble du panel				
VFDB	Aucun gène spécifique n'a été mis en évidence sur les souches persistantes des usines A et B par rapport au reste du panel : profils similaires. Fort pouvoir d'adhérence et d'invasion.				
BacMet	Profils similaires sur l'ensemble du panel, à l'exception de **gènes** retrouvés spécifiquement sur les souches des usines A et B	brcA brcB brcC	Résistance au chlorure de benzalkonium	Utilisation de produits d'entretiens contenant du chlorure de benzalkonium. Transmission par un **transposon** issu d'*E. coli*.	(Nishino & Yamaguchi, 2001)

5.5 Valorisation des acquis

Dans le cadre de mon diplôme, il a été question d'apprendre une toute nouvelle discipline, liée à la fois à mes désirs de maîtrise des techniques de pointe, et aux besoins analytiques de l'unité pour répondre à des missions de service public. Ma curiosité scientifique m'a permis d'apprendre, dès début 2018, la terminologie adaptée à la bio-informatique, à travers une lecture bibliographique approfondie (Pightling *et al.*, 2018), (Vincent *et al.*, 2018), etc.). A partir d'octobre 2018, une collègue bio-informaticienne m'a formée à l'utilisation des outils SeqSphere+, iVARCall2 et Enterobase. Suite à cela, fin 2018 et en 2019, j'ai réalisé les analyses bio-informatiques, et rapports d'analyses associés aux alertes sanitaires et aux demandes clients liées à *Salmonella* et *Listeria monocytogenes*. En fin d'année 2019, j'ai été nommée Responsable Technique WGS. Dans ce cadre, il m'a été confié des projets d'attribution de source, ou encore la gestion globale d'une vingtaine d'alertes sanitaires en lien direct avec les autorités compétentes.

Ma participation à des séminaires internes et externes, dont la conférence « Joint Conference Food Safety WGS » de mars 2019, m'ont permis de me tenir à jour de l'actualité bio-informatique dans le domaine de la sécurité sanitaire des aliments. Par ailleurs, j'ai eu l'opportunité de présenter mes travaux lors de réunions thématiques avec certains laboratoires partenaires ou tutelles, dont une présentation pour les DDPP d'Ile de France en janvier 2020. Deux présentations devant des collectifs de laboratoires vétérinaires sont prévus en septembre et en novembre 2020.

Forte de ces expériences, j'ai été en mesure de rédiger plusieurs documents qualités destinés au LSAl et à l'unité SEL, permettant de couvrir l'ensemble du processus analytique :

- Mode opératoire : Extraction d'ADN de qualité génomique pour *Salmonella* et *Listeria*
- Méthode interne : Analyse cgMLST via l'utilisation de Ridom SeqSphere+
- Méthode interne : Analyse SNP sur le cluster de calcul Linux : iVARCall2.
- Trame de rapport d'analyse cgMLST
- Envoi des ADN et réception des données brutes de séquençage
- Fiches de traçabilités associées

La rédaction de l'ensemble de ces documents (cf. détails en annexe 6) va permettre de normaliser les pratiques des opérateurs, tout en rentrant dans une démarche forte d'assurance qualité. Combiné aux analyses comparatives du projet, j'ai réalisé une partie importante du dossier de validation de méthode, me permettant d'accréditer la méthode lors de notre prochain audit Cofrac, courant 2021.
J'ai par ailleurs eu l'opportunité de participer en 2019 et 2020, aux essais inter laboratoires d'aptitudes, sur les deux pathogènes, par analyse *SNP calling* et cgMLST. Les résultats de ces essais sont tous satisfaisants.

6 Discussion

La réponse aux alertes produits et aux alertes sanitaires françaises et européenne est au cœur de nos missions de laboratoires de référence, mais est surtout un enjeu majeur pour la santé publique. Les évolutions technologiques et techniques de ces dix dernières années nous ont permis de donner une nouvelle dimension aux analyses réalisées dans ce cadre. Les analyses standardisées utilisées par le passé (PFGE, MLVA) sont délaissées progressivement au profit des analyses de séquençage complet du génome. De nombreuses études ont démontré l'intérêt de cette transition (Moura *et al.*, 2017) afin d'obtenir des données plus discriminantes. Cette transition, évaluée stratégiquement au sein de chaque laboratoire, est également accompagnée par les agences nationales ou européennes (ECDC-EFSA, 2019).

Accompagnée d'une équipe de bio-informaticiens chevronnés, l'unité SEL a été capable d'enclencher ce changement technique majeur, avec l'abandon progressif, courant 2019, des méthodes de typage moléculaire conventionnel. Grâce aux données et aux souches collectées dans le cadre de nos activités de surveillance, nous sommes en mesure de répondre de plus en plus rapidement aux sollicitations de nos tutelles en cas d'alerte sanitaires. En effet, la mise en routine des analyses génomique, permet la collecte de génomes utilisables pour répondre aux investigations sanitaires. Toutefois, confronté à la multitude d'analyses possibles, il a été nécessaire de faire le point sur nos possibilités analytiques.

6.1 Le « wet-lab »

Les protocoles d'extractions, de préparation de librairies et de séquençages sont efficients et répondent aux critères qualités définis par le laboratoire.

6.2 Le « dry-lab »

6.2.1 Recherche des évènements de recombinaison

J'ai recherché les évènements de recombinaison des génomes étudiés à l'aide de l'outil ClonalFrameML. Les deux panels étudiés présentent très peu de recombinaisons, uniquement sur de très courtes séquences. Les calculs effectués ne mettent pas en évidence d'impact réel sur la topologie des arbres phylogénétiques. : je n'ai pas observé de remaniement majeur des embranchements ou de la longueur des branches. Les panels étudiés semblent assez clonaux, tout en démontrant bien une large diversité génomique représentative du sérovar ou du CC dans l'environnement (nombre de SNP allant de 0 à plus de 70 pour *Listeria* et de 0 à 137 pour *Salmonella*).

Les recombinaisons présentent au sein des génomes sont plutôt retrouvées sur des portions de gènes codants pour des fonctions régulatrices, à l'exception de la recombinaison présente sur la

souche de *S.* Agona, 2016LSAL04644 dont la séquence correspond à une portion de gène codant pour l'adhésion.

6.2.2 Comparaison des matrices de distance par analyse de distance SNP et allélique

Les résultats obtenus étant du même ordre de grandeur entre les deux méthodes, cela permet de mettre en avant une cohérence entre les résultats obtenus sur les deux panels exploités. La plupart des gènes recherchés selon les schémas cgMLST sont considérés comme des gènes à évolution lente (Jagadeesan *et al.*, 2019). Ces observations sont cohérentes avec la littérature (Chen *et al.*, 2016), qui montre que le cgMLST est suffisamment discriminant, dans le cadre d'alertes sanitaires, pour définir les souches appartenant au cluster épidémique. Par contre, l'analyse de *SNP calling* permet de mieux apprécier le lien évolutif entre les souches, identifiées avec plus de précision les clusters et sous-clusters et intégrer une vision statistique sur la robustesse des branches via les bootstraps.

Pour Salmonella, j'ai calculé trois SNP de différence entre les souches 2018LSAL00986 et 2018LSAL00988 prélevées dans un laps de deux mois (alerte de 2017-2018). Cependant, j'ai observé que les souches 2017LSAL04598 et 2018LSAL00593 ne présentent aucun SNP de différence alors qu'elles ont été prélevées le même jour et séquencées sur la même flowcell. De la même façon, pour *Listeria*, j'ai retrouvé trois SNP de différence sur les souches A4-02-LmUB3PA et CL369-S2-LmUB3PA, prélevées dans la même usine à deux ans d'intervalle. Des erreurs de séquençage Illumina ont été reportées dans la littérature à hauteur de 0,1% des *reads* bruts. Les différences en SNP observées sur les souches d'un même cluster pourraient être dues à des erreurs de séquençage. Toutefois, nous ne pouvons pas exclure une réelle mutation. Par ailleurs, en faisant le parallèle avec l'analyse ClonalFrameML, j'ai mis en évidence trois homoplasies sur des régions conservées, pouvant également expliquer les variations de SNP. J'ai pu mettre en évident un lien phylogénétique très fort entre les souches des deux alertes à *S.* Agona, qui présentent seulement en moyenne 5 SNP de différence. A la lumière de ces éléments, il est possible que la souche à l'origine de la TIAC de 2005 soit persistante dans l'usine. Les souches de S. Agona ont été précédemment décrites comme capables de former des biofilms (Corcoran *et al.*, 2013, Gonzalez-Machado *et al.*, 2018). Il n'y a en moyenne que 35 SNP de différence sur l'ensemble du panel *S.* Agona alors que les souches sélectionnées sont issues de matrices très diversifiées, prélevées sur tout le territoire et sur une période de treize ans. Toutefois, certaines souches issues de pays étrangers présentent jusqu'à plus de 130 SNP d'écart avec les souches Françaises. Il serait intéressant de caractériser de façon approfondie la souche 2012LSAL01803 issue d'un prélèvement de crevettes Indiennes (commercialisées en France). Parmi les 5 530 souches analysées de notre base PFGE, 51% des souches de *S.* Agona (38/75) sont représentées par quatre pulsotypes majoritaires. Le nombre de souches de L. monocytogenes du CC204 ne représente qu'environ 2% des souches analysées par PFGE par notre laboratoire. A travers l'étude des deux panels,

je peux émettre l'hypothèse que les souches appartenant au sérovar Agona ou au CC204 de *L. monocytogenes* sont très clonales.

6.2.3 Comparaison des matrices de distance par le test statistique de Mantel

Les tests statistiques effectuées via le test de Mantel prouvent une bonne corrélation (r^2>0,7 et P <2,2^{e-16}) entre les analyses de *SNP calling* avec iVARCall2 et par cgMLST avec SeqSphere. Cela démontre bien que les deux types d'analyses sur les deux pathogènes rendent des résultats cohérents entre eux.

A travers nos résultats, j'ai pu observer que le facteur de corrélation entre les valeurs SNP et les différences alléliques, sur la totalité du panel, est plus faible pour *Listeria* que pour *Salmonella*, bien que le panel de souches de *Listeria* soit plus diversifié (pays différents). En utilisant les moyennes calculées en table 6 et 7, on obtient un facteur de 1,94 pour *Listeria* et de 2,34 pour *Salmonella*. D'ailleurs les valeurs de r^2 reflètent ces différences, le r^2 de *Listeria* étant supérieur à celui de *Salmonella*. Dans les deux cas, la pente penche vers iVARCAll2 (plus que de SNP que de différence allélique dans les matrices). Ceci est en accord avec la littérature, *Listeria* étant très clonale.

6.2.4 Comparaison des arbres phylogénétiques

Pour *S*. Agona, des différences sont observées sur le positionnement des souches épidémiques et des souches non reliées épidémiologiquement. En prenant en compte le fait que les branches au sein desquelles les inversions de positions sont observées sont soutenus par des bootstraps supérieurs à 80%, et le fait que ces souches soient très proches génétiquement, les modifications de positionnement sont alors négligeables.

Concernant *L. monocytogenes*, l'appariement entre les deux arbres phylogénétiques est révélateur d'une cohérence forte entre l'analyse cgMLST et SNP.

Afin d'expliquer les différences de résultats observées entre *Salmonella* et *Listeria*, il est possible d'avancer plusieurs hypothèses. L'analyse SNP effectuée sur le génome bactérien complet, permet la prise en compte d'éléments génétiques mobiles comme les plasmides, plus présents chez *Salmonella* que chez *Listeria*. Une autre piste est la prise en compte des gènes inclus dans le schéma cgMLST utilisé. En effet, la quasi-totalité des gènes référencés dans ces schémas sont présents dans le core-génome. Toutefois, certains gènes du génome accessoire peuvent être pris en compte, tels des gènes de résistance aux antibiotiques pour *Salmonella*.

6.2.5 Utilisation du script « matrix2association »

Cet outil n'est pas retenu dans l'analyse des évènements sanitaires. Il est nécessaire d'implémenter un nombre de souches reliées épidémiologiquement trop important pour être utilisé dans le cadre d'investigations. En effet, sans ce contexte nous ne disposons jamais de suffisamment de souches pour faire cette analyse de source attribution.

Dans tous les cas, l'utilisation de ce test, et surtout la définition des souches « reliées » semble aléatoire et dépend à la fois du pathogène analysé, de son sérovar/CC/ST ainsi que du nombre de souches sélectionnées. Aucun schéma analytique fiable ne peut être défini afin de garantir la cohérence des résultats obtenus.

Il apparait alors que cet outil semble plus approprié sur les *Salmonella* que sur les *Listeria*. En amont d'analyses phylogénétiques, cet outil peut permettre aux équipes d'avoir des indications sur la construction des clusters. Néanmoins, à ce jour, de nombreux logiciels d'inférence phylogénétique rapides ont été développés, tel que IQ-TREE utilisant le ML, rendant en moyenne un résultat cinq fois plus rapide que RAxML.

6.2.6 Recherche des gènes de virulence, de résistance et de persistance

L'analyse du génome accessoire permet de vérifier le clustering des souches liées aux alertes sanitaires, en prenant en compte des éléments génétiques mobiles. Une analyse approfondie peut permettre de mettre en évidence l'acquisition de plasmides, pouvant donner à la bactérie, par exemple, des résistances aux antibiotiques ou encore aux métaux lourds. La prise en compte de ces éléments peut permettre de soutenir les hypothèses de clustering observées précédemment ou au contraire les remettre en cause en fonction des cas. La recherche de ces gènes d'intérêts permet également de mener une étude plus pointue quant à la persistance de certaines souches dans les environnements clés des étapes de la chaine de production de « la fourche à la fourchette ». Dans ce cas, je chercherai des hypothèses permettant de donner une explication au caractère persistant des souches analysées, au sein des usines.

Pour le panel de ***S.* Agona**, sur l'ensemble des gènes détectés, seuls quatre gènes semblent spécifiques aux souches liées aux deux alertes sanitaires. Ces gènes ont probablement été acquis par transfert horizontal grâce à un plasmide. Lors de l'analyse cgMLST, les plasmides n'ont pas été pris en compte, n'influençant pas la phylogénie des souches. A l'inverse, les plasmides sont intégrés à l'analyse wgSNP pouvant expliquer une partie des différences de clustering observées entre les deux méthodes. En parallèle la résistance à l'argent et au plomb semble résulter de l'utilisation intensive de biocides dans l'usine de production, pouvant expliquer la persistance de ces souches entre 2005 et 2018. Par ailleurs, la souche 2012LSAL01803 a été identifiée comme multi-résistante (famille des aminosides, β-lactames, tétracyclines, sulfamides, triméthoprimes, diaminopyrimidines, phénicols et quinolones).

Après analyse des données épidémiologiques, il s'agit d'une souche provenant d'Inde, où de nombreuses souches multi-résistantes ont été reportées sur les dernières années. L'utilisation des antibiotiques, en filière humaine et vétérinaire, n'est pas règlementée dans ce pays. Il pourrait être intéressant de comparer cette souche avec d'autres souches disponibles sur des bases de données en libre accès (au 07/07/2020, trois génomes Indiens d'Agona sont disponibles sur Enterobase)

Concernant le panel de ***L. monocytogenes*** CC204, en comparant les données obtenues entre les souches reliées des usines A et B et les souches non reliées épidémiologiquement, un seul gène d'intérêt à retenu notre attention. Les souches persistantes des usines Françaises semblent avoir acquis, par le biais d'un transposon, des gènes de résistance au chlorure de benzalkonium (Dutta *et al.*, 2014). Il s'agit d'un composé présent dans la composition de nombreux biocides utilisés dans les industries agroalimentaires. La présence de ce gène de résistance peut être une piste afin de comprendre la persistance de ces souches sur une période de six ans dans les deux usines.

6.3 Conclusion de la discussion

Les bases de données développées et utilisées pour la recherche de gènes d'intérêts semblent fonctionnelles et en cohérence avec les besoins de compréhension de la résurgence de certaines alertes sanitaires. Cependant, ce genre d'approfondissement analytique doit être modéré et réservé aux cas d'alertes les plus complexes. Nos équipes de recherches peuvent intervenir en complément de caractérisation de certaines souches liées aux alertes sanitaires, en vue de publication, mais ces données n'ont pas aujourd'hui nécessité à apparaitre sur les rapports officiels.

A travers les résultats exploités dans ce mémoire, il apparait évident que le contrôle qualité et la normalisation des reads est un facteur clé des analyses génomiques. Compte-tenu du type d'analyse effectué, les outils bio-informatiques à notre disposition semblent cohérents entre eux. Il est donc nécessaire de dissocier nos possibilités analytiques de nos besoins. Plusieurs facteurs sont donc à prendre en compte, le pouvoir discriminant attendu dans le cadre de l'analyse, la comparabilité et/ou le partage des résultats avec d'autres organismes ainsi que la standardisation possible de l'analyse mise en œuvre. L'unité a en effet mis en routine, dans le cadre de demandes clients, une analyse cgMLST, reposant sur l'utilisation de schéma régulièrement mis à jour et téléchargeable par tous. Dans le cadre d'alerte sanitaire, le cgMLST est actuellement l'analyse la plus répandue. Elle est facilement interprétable par toutes les instances et plus facilement comparable inter-laboratoires (grâce par exemple à l'utilisation de bases de données telles qu'Enterobase). Toutefois, dans le cadre d'alertes sanitaires importantes, ayant un fort impact sur la population, il est parfois nécessaire d'avoir accès à un pouvoir discriminant plus fort, comme cela a été le cas pour l'alerte à *S*. Agona de 2018. Une des principales contraintes des analyses wgSNP est d'avoir accès à un génome de référence, à la fois complet, ayant une couverture acceptable et suffisamment proche du sérovar, CC ou ST analysé.

Le choix du type d'analyse réalisé (cgMLST, wgMLST, *SNP calling*) repose donc fondamentalement sur les besoins de l'alerte en question. Mais quels sont les outils les plus robustes et fiables pour l'analyse ? Au cours des deux années écoulées, j'ai pu prendre en main les différents outils à notre disposition (iVARCall2, SeqSphere et Enterobase). Il est évident que l'apprentissage du language « bash » et « python » est un frein à l'utilisation de l'outil iVARCall2 par le plus grand nombre. Toutefois, une fois maitrisé, nous avons grâce au système d'exploitation Linux une maitrise totale de l'ensemble du paramétrage de l'outil. Concernant SeqSphere, à l'Agence, le logiciel est relié à un serveur, permettant la fluidité des calculs. SeqSphere permet une prise en main rapide des analyses à effectuer. Son paramétrage est clair et facilement ajustable. Le logiciel offre une opportunité de réaliser une analyse de bout en bout : du contrôle qualité à la phylogénie. Dû à de fortes contraintes techniques, je n'ai pas été en mesure de tester les panels sélectionnés avec le logiciel BioNumerics dans son ensemble. Toutefois, il est évident que nous devons reconnaitre de nombreuses possibilités analytiques avec BioNumerics (cgMLST, wgMLST, wgSNP). L'implémentation de modules complémentaires à celui « WGS tools » peut permettre à l'opérateur de démultiplier les analyses. Malgré cette souplesse, il est également apparu que le langage utilisé dans BioNumerics pour le paramétrage n'est pas standardisé. Il en ressort que de nombreux utilisateurs optent pour les paramètres par défaut ou ressentent la nécessité de contacter le fabriquant à de multiples reprises. Par ailleurs, les nombreux soucis techniques rencontrés avec le serveur dématérialisé permettant les calculs, semble être un facteur récurent et limitant.

En conclusion, il semble évident qu'une importante partie de nos analyses repose sur le système d'exploitation Linux avec une vaste sélection d'outils et de paramétrages associés. Il est important de dissocier les besoins liés aux alertes sanitaires, nécessitant une normalisation des pratiques plus rigoureuse, et les besoins des activités de recherche qui sont associés aux souches d'alertes. Afin de pallier ce problème, les procédures et modes opératoires liés aux analyses d'alertes ont été rédigés et mis sous assurance qualité.

7 Conclusions et perspectives

L'une des missions principales de l'unité est de répondre aux évènements sanitaires efficacement. L'évolution des techniques bio-informatiques est rapide, démultipliant les possibilités d'analyses. Il est donc nécessaire de cadrer les attentes liées aux alertes pour y répondre synthétiquement et rapidement. La mise au point des techniques de « wet-lab » et de « dry-lab » au sein de notre laboratoire, ont permis de répondre en partie à ces besoins. La possibilité d'effectuer le séquençage sur site, a permis de diminuer drastiquement le temps analytique en vue de la réponse à une alerte sanitaire. Le laboratoire s'est récemment équipé de deux robots d'extraction utilisant la technologie de microbilles magnétiques pour automatiser la purification de l'acide nucléique et des protéines. Ces robots sont actuellement en phase de validation de méthode et pourraient permettre à l'unité d'augmenter son efficience. Le laboratoire est équipé d'un séquenceur MiSeq, technologie Illumina®, ne permettant pas un haut rendement analytique. En moyenne, en fonction de la taille du génome séquencé et de la couverture attendue, il est possible de séquencer en moyenne huit souches par cassette, pour une durée de séquençage d'environ 18 à 24h. Bien que nous ayons toujours la possibilité d'externaliser la préparation de la librairie et le séquençage, il serait peut-être utile d'envisager l'achat d'un second séquenceur Illumina ou la mise en place d'une plateforme de séquençage utilisant cette technologie. Cela permettrait à la fois d'augmenter notre débit analytique pour répondre à des alertes de grosse envergure, mais également diminuer le coût analytique lié à la cassette utilisée.

Les étapes de contrôle qualité et de normalisation reposent sur un processus commun à tout le LSAl et répond parfaitement à nos besoins et exigences. Dans le but de normaliser les sélections de génomes de référence pour les analyses de *SNP calling*, une base de données de génomes de références a été mise au point. Cependant, l'ensemble des sérovars (parfois polyphylétiques) ou des ST/CC n'ont pas de génome de référence associé. Il est donc nécessaire de les implémenter au fur et à mesure. Dans le cadre d'alerte sanitaire, nous pouvons être confrontés à des sérovars/ST rares. Compte-tenu de l'importante collection de souches bactériennes disponibles au laboratoire, il serait donc intéressant, si aucun génome circulaire n'est disponible sur les bases de données en ligne, d'être en capacité de mettre en parallèle deux technologies de séquençage, une « shorts reads » telle qu'Illumina et une « longs reads » telle que Nanopore (MinIon) ou PacBio, afin de pouvoir circulariser nous même le génome et le définir en tant que génome de référence. L'acquisition récente d'un MinIon nous permettra de tester cette approche.

Les outils mis à notre disposition permettent de couvrir nos besoins analytiques, tout en ayant une confiance dans les résultats rendus. Le fonctionnement du RS et la collecte des souches dans le cadre des demandes clients d'analyses WGS, apportent une réelle plus-value à nos investigations. En effet, les nombreuses souches disponibles en collection au laboratoire, issues de prélèvements récents permettent de rendre les études d'attribution de source dans la cadre d'alerte plus efficientes. Les

demandes clients pour des analyses WGS permettent parfois d'avoir directement à disposition les génomes normalisés pour analyse, permettant un gain de temps considérable. Bien que les échanges de données avec les CNR et avec les laboratoires européens soient opérationnels, il a été démontré via l'utilisation d'Enterobase pour les salmonelles, qu'une base de données commune peut permettre une exploitation rapide et simple des séquences générées. D'autres outils d'analyses cgMLST ou wgMLST, récemment développés, tels que chewBBACA (Silva *et al.*, 2018) ou bigsdb-Lm (https://bigsdb.pasteur.fr/listeria/), pourrait permettre, sur la base d'un outil libre de droit sous Linux ou Galaxy, la création d'une base de données allélique, offrant une harmonisation inter-laboratoires, quel que soit le pathogène étudié. BioNumerics n'ayant pas pu être testé dans le cadre de ce diplôme, il serait donc intéressant de compléter la comparaison effectuée avec BioNumerics et chewBBACA.

BioNumerics, à travers son module WGS, offre de diversité analytique importante. Il convient toutefois de noter que l'outil offre de nombreux paramétrages possibles, nécessitant une bonne prise en main de l'outil pour une utilisation optimale. Les retours d'expériences des opérateurs démontrent par ailleurs, des problématiques différentes en fonction du microorganisme étudié. Les essais menés sur cet outil devront en tenir compte.

A travers l'exploitation de mes données obtenues, j'ai pu observer une souche de *S.* Agona, issue d'un prélèvement Indien de crevettes, qui semble se détacher génomiquement du reste du panel. Il serait donc intéressant de constituer un panel annexe de *S.* Agona, ayant des origines géographiques plus vastes afin d'essayer de visualiser la diversité génomique de ce sérovar, tout en replaçant la souche Indienne au sein d'un cluster.

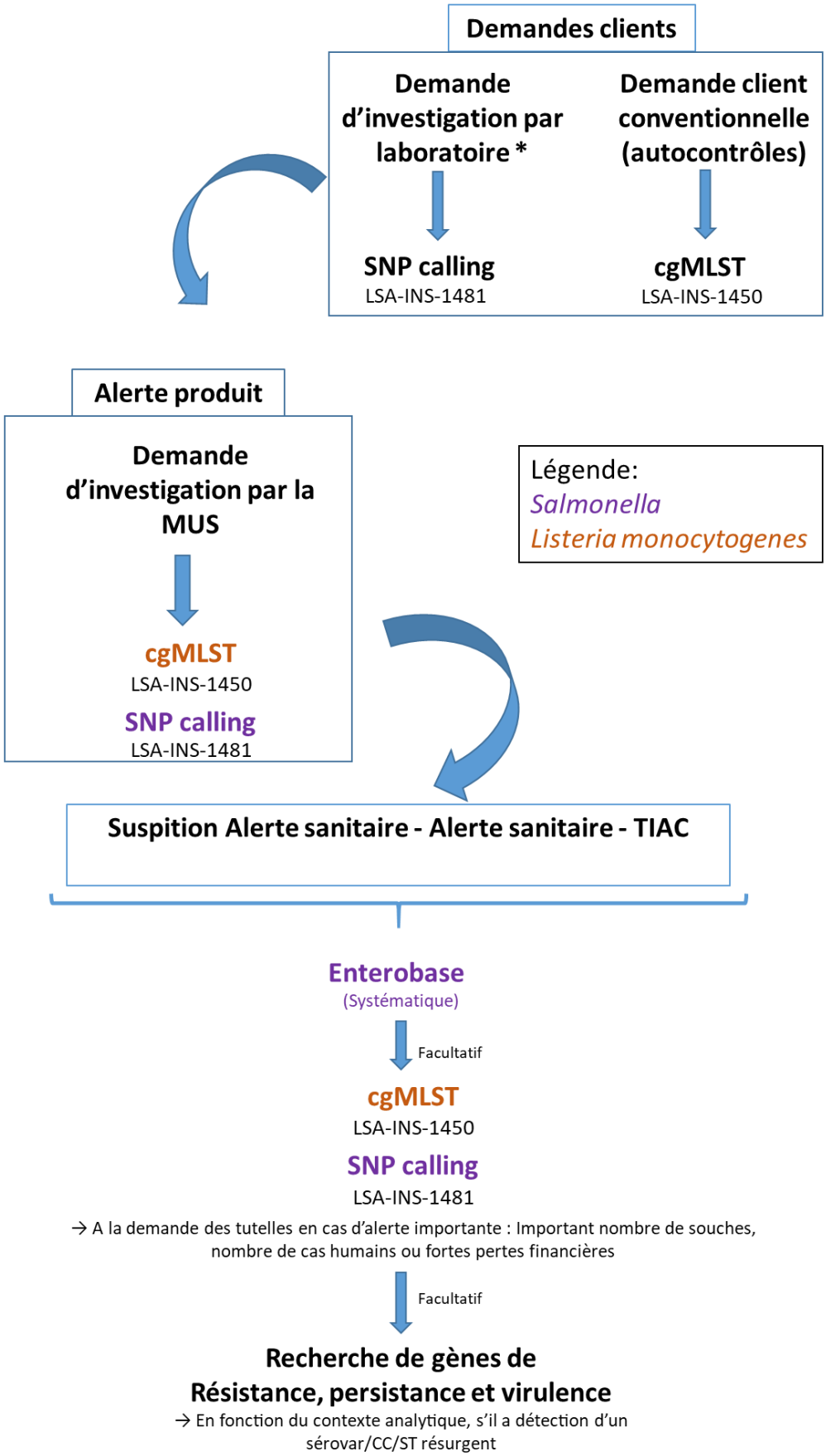
A l'aide des matrices de distances obtenues, j'ai calculé le nombre moyen de SNP de différence entre les souches. Il serait également intéressant de calculer le taux de mutation pour chaque panel. Ainsi, il sera possible d'estimer le rythme selon lequel apparait une mutation SNP au sein du génome. Cela permettra de mieux définir les clusters et d'appréhender d'une nouvelle manière les études de source attribution, en reliant par exemple deux évènements sanitaires entre eux à travers le temps.

La recherche des gènes d'intérêts effectuée permet de mettre en évidence l'intérêt de pousser la caractérisation des souches lorsqu'une alerte est liée à la persistance de souches dans un environnement. L'étude du génome accessoire peut permettre une analyse plus détaillée dans le cadre d'évènements plus complexes. Il ne semble toutefois pas nécessaire de mettre en place ces analyses pour l'ensemble des alertes sanitaires, le but principal étant de relier la souche pathogène à un aliment. Dans le cas de la présence de souches persistantes dans un environnement agroalimentaire (atelier de découpe, industries agro-alimentaires …), la recherche des gènes de virulence, de persistance et de résistance pourrait permettre une réduction drastique des pertes économiques des industries confrontées à ces problématiques.

Les tests réalisés dans le cadre du diplôme permettent d'avoir confiance dans les résultats rendus. Ces résultats sont cohérents avec d'autres études similaires réalisées sur les deux pathogènes

(Henri *et al.*, 2017, Pearce *et al.*, 2018). Ajouté à cela, les documents qualités rédigés pour ces nouvelles méthodes analytiques, permettent d'envisager sereinement une démarche d'accréditation Cofrac. Un projet de norme sur le WGS, dans le cadre de l'agroalimentaire, est actuellement en cours de rédaction par l'Organisation Internationale de la Normalisation (ISO). Différents axes permettant de constituer un dossier de validation de méthode semblent se dégager des premières réflexions, comme la comparaison des résultats obtenus entre deux laboratoires, ou obtenus par deux opérateurs différents. Nous avons à ce jour pleinement confiance dans ces résultats, déjà partiellement exploités par notre laboratoire, dans le cadre de l'organisation et de la participation à des essais inter-laboratoires d'aptitudes (EILA) pour le typage WGS de souches bactériennes.

En conclusion, il semble cohérent, dans le cadre des alertes sanitaires, de partir sur une méthodologie harmonisée telle que le cgMLST et de venir compléter les analyses, au cas par cas, selon l'ampleur et le contexte de l'alerte en question. La création d'un arbre décisionnel, présenté en figure 19, permet d'aiguiller les choix analytiques. Pour *Salmonella*, Enterobase est un outil incontournable de l'analyse génomique dans le cadre de TIAC et d'alerte sanitaire car elle permet l'inter-comparaison de données avec le CNR. L'analyse menée par Enterobase repose sur un schéma cgMLST. En cas de besoin d'analyse approfondie, il est alors judicieux de mettre en application une analyse *SNP calling* afin d'apporter un meilleur pouvoir discriminant à nos sorties analytiques. Pour *Listeria*, les résultats obtenus montrent une corrélation très forte entre les résultats obtenus sur les deux approches. Le pouvoir discriminant de l'analyse cgMLST est donc amplement suffisant pour les répondre aux investigations, tout en sollicitant une puissance de calcul moins importante. Les résultats obtenus dans cette étude peuvent être transcrits aux besoins analytiques dans le cadre de demandes clients. Ainsi, le cgMLST est l'outil préférentiellement utilisé, toutefois, dans certains contextes le *SNP calling* ayant un pouvoir discriminant plus élevé pourra être envisagé. Pour finir, les outils détaillés dans ce mémoire, semblent totalement cohérents avec les besoins du laboratoire et de nos tutelles.



* Contexte analytique lié à l'abattage d'animaux (souches règlementées), résurgence d'un sérovar au sein d'un élevage ou d'une usine, indemnisation dans le cadre de contrat d'assurance …

Figure 19 - Arbre décisionnel pour le choix analytique des données génomiques de Salmonella et Listeria monocytogenes

8 Références bibliographiques

https://www.cgmlst.org/ncs.

Achtman, M., Wain, J., Weill, F.X., Nair, S., Zhou, Z., Sangal, V., Krauland, M.G., Hale, J.L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L.H., Brisse, S., and Group, S.E.M.S. (2012) Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* **8**: e1002776.

AFNOR, (2015) NF EN ISO 9001 - Systèmes de management de la qualité - Exigences. In: . https://www.iso.org/fr/standard/62085.html, pp. .

AFNOR, (2018) NF EN ISO 17025 - Exigences générales concernant la compétence des laboratoires d'étalonnages et d'essais. In: . https://www.iso.org/fr/standard/66912.html: , pp.

Alikhan, N.F., Zhou, Z., Sergeant, M.J., and Achtman, M. (2018) A genomic overview of the population structure of *Salmonella*. *PLoS Genet* **14**: e1007261.

Bale, J., Meunier, D., Weill, F.X., dePinna, E., Peters, T., and Nair, S. (2016) Characterization of new *Salmonella* serovars by whole-genome sequencing and traditional typing techniques. *J Med Microbiol* **65**: 1074-1078.

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., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

Besser, J., Carleton, H.A., Gerner-Smidt, P., Lindsey, R.L., and Trees, E. (2018) Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* **24**: 335-341.

Borremans, B., Hobman, J.L., Provoost, A., Brown, N.L., and van Der Lelie, D. (2001) Cloning and functional analysis of the pbr lead resistance determinant of *Ralstonia metallidurans* CH34. *J Bacteriol* **183**: 5651-5658.

Brown, E.W., Mammel, M.K., LeClerc, J.E., and Cebula, T.A. (2003) Limited boundaries for extensive horizontal gene transfer among Salmonella pathogens. *Proc Natl Acad Sci U S A* **100**: 15676-15681.

Caulfield, T., Evans, J., McGuire, A., McCabe, C., Bubela, T., Cook-Deegan, R., Fishman, J., Hogarth, S., Miller, F.A., Ravitsky, V., Biesecker, B., Borry, P., Cho, M.K., Carroll, J.C., Etchegary, H., Joly, Y., Kato, K., Lee, S.S., Rothenberg, K., Sankar, P., Szego, M.J., Ossorio, P., Pullman, D., Rousseau, F., Ungar, W.J., and Wilson, B. (2013) Reflections on the cost of "low-cost" whole genome sequencing: framing the health policy debate. *PLoS Biol* **11**: e1001699.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., and Jin, Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* **33**: D325-328.

Chen, Y., Gonzalez-Escalona, N., Hammack, T.S., Allard, M.W., Strain, E.A., and Brown, E.W. (2016) Core Genome Multilocus Sequence Typing for Identification of Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of *Listeria monocytogenes*. *Appl Environ Microbiol* **82**: 6258-6272.

Cofrac, (2019) LAB GTA 59 - Analyses microbiologiques des produits et environnement agro-alimentaires. In., pp.

Corcoran, M., Morris, D., De Lappe, N., O'Connor, J., Lalor, P., Dockery, P., and Cormican, M. (2013) Salmonella enterica biofilm formation and density in the Centers for Disease Control and Prevention's biofilm reactor model is related to serovar and substratum. *J Food Prot* **76**: 662-667.

Dangel, A., Berger, A., Messelhausser, U., Konrad, R., Hormansdorfer, S., Ackermann, N., and Sing, A. (2019) Genetic diversity and delineation of *Salmonella Agona* outbreak strains by next generation sequencing, Bavaria, Germany, 1993 to 2018. *Euro Surveill* **24**.

Didelot, X., Achtman, M., Parkhill, J., Thomson, N.R., and Falush, D. (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res* **17**: 61-68.

Didelot, X., and Wilson, D.J. (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**: e1004041.

Doublet, B., Weill, F.X., Fabre, L., Chaslus-Dancla, E., and Cloeckaert, A., (2004) Variant *Salmonella* Genomic Island 1 Antibiotic Resistance Gene Cluster Containing a Novel 3'-N-Aminoglycoside Acetyltransferase Gene Cassette, aac(3)-Id, in *Salmonella enterica* Serovar Newport. In: Antimicrob Agents Chemother. pp. 3806-3812.

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

Dutta, S., Das, S., Mitra, U., Jain, P., Roy, I., Ganguly, S.S., Ray, U., Dutta, P., and Paul, D.K. (2014) Antimicrobial resistance, virulence profiles and molecular subtypes of Salmonella enterica serovars Typhi and Paratyphi A blood isolates from Kolkata, India during 2009-2013. *PLoS One* **9**: e101347.

ECDC-EFSA, V.W.I., Guerra B, Borges V, André Carriço J, Guy C, et al. (2019) EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC–EFSA molecular typing database. *EFSA Support Publication*.

ECDC, (2016) Laboratory standard operating procedure for multiple-locus variable-number tandem repeat analysis of *Salmonella enterica* serotype Enteritidis. In. ecdc.europa.eu/sites/default/files/media/en/publications/Publications/Salmonella-Enteritidis-Laboratory-standard-operating-procedure.pdf, pp.

EFSA and ECDC (2018) The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017.

Espie, E., Weill, F.X., Brouard, C., Capek, I., Delmas, G., Forgues, A.M., Grimont, F., and de Valk, H. (2005) Nationwide outbreak of *Salmonella enterica* serotype Agona infections in infants in France, linked to infant milk formula, investigations ongoing. *Euro Surveill* **10**: E050310 050311.

Felix, B., Feurer, C., Maillet, A., Guillier, L., Boscher, E., Kerouanton, A., Denis, M., and Roussel, S. (2018) Population Genetic Structure of Listeria monocytogenes Strains Isolated From the Pig and Pork Production Chain in France. *Front Microbiol* **9**: 684.

Felten, A., (2017) ARtWORK GitHub. In., pp.

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.

Gonzalez-Machado, C., Capita, R., Riesco-Pelaez, F., and Alonso-Calleja, C. (2018) Visualization and quantification of the cellular and extracellular components of Salmonella Agona biofilms at different stages of development. *PLoS One* **13**: e0200011.

Grimont, P., and Weill, F. (2007) Formules antigéniques des sérovars de *Salmonella*. *Centre Collaborateur OMS de Référence et de Recherche sur les Salmonella* **https://www.pasteur.fr/sites/default/files/vf_0.pdf**.

Gupta, A., Matsui, K., Lo, J.F., and Silver, S. (1999) Molecular basis for resistance to silver cations in *Salmonella*. *Nat Med* **5**: 183-188.

Henri, C., Leekitcharoenphon, P., Carleton, H.A., Radomski, N., Kaas, R.S., Mariet, J.F., Felten, A., Aarestrup, F.M., Gerner Smidt, P., Roussel, S., Guillier, L., Mistou, M.Y., and Hendriksen, R.S. (2017) An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes*. *Front Microbiol* **8**: 2351.

Jagadeesan, B., Baert, L., Wiedmann, M., and Orsi, R.H. (2019) Comparative Analysis of Tools and Approaches for Source Tracking *Listeria monocytogenes* in a Food Facility Using Whole-Genome Sequence Data. *Front Microbiol* **10**.

Jolley, K.A., Wilson, D.J., Kriz, P., McVean, G., and Maiden, M.C. (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol Biol Evol* **22**: 562-569.

Jourdan-da Silva, N., Fabre, L., Robinson, E., Fournet, N., Nisavanh, A., Bruyand, M., Mailles, A., Serre, E., Ravel, M., Guibert, V., Issenhuth-Jeanjean, S., Renaudat, C., Tourdjman, M., Septfons, A., de

Valk, H., and Le Hello, S. (2018) Ongoing nationwide outbreak of *Salmonella* Agona associated with internationally distributed infant milk products, France, December 2017. *Euro Surveill* **23**.

Junemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., and Harmsen, D. (2014) GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. *PLoS One* **9**: e107014.

Kchouk, M.e.a. (2017) Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M*., et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

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

Majowicz, S.E., Musto, J., Scallan, E., Angulo, F.J., Kirk, M., O'Brien, S.J., Jones, T.F., Fazil, A., Hoekstra, R.M., and Studies, f.t.I.C.o.E.D.B.o.I. (2010) The Global Burden of Nontyphoidal Salmonella Gastroenteritis. *Clinical Infectious Diseases* **50**: 882-889.

Mantel, N., and Fleiss, J.L. (1980) Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *Am J Epidemiol* **112**: 129-134.

Marcus, S.L., Brumell, J.H., Pfeifer, C.G., and Finlay, B.B. (2000) *Salmonella* pathogenicity islands: big virulence in small packages. *Microbes Infect* **2**: 145-156.

Maury, M.M., Tsai, Y.H., Charlier, C., Touchon, M., Chenal-Francisque, V., Leclercq, A., Criscuolo, A., Gaultier, C., Roussel, S., Brisabois, A., Disson, O., Rocha, E.P.C., Brisse, S., and Lecuit, M. (2016) Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat Genet* **48**: 308-313.

Moura, A., Tourdjman, M., Leclercq, A., Hamelin, E., Laurent, E., Fredriksen, N., Van Cauteren, D., Bracq-Dieye, H., Thouvenot, P., Vales, G., Tessaud-Rita, N., Maury, M.M., Alexandru, A., Criscuolo, A., Quevillon, E., Donguy, M.P., Enouf, V., de Valk, H., Brisse, S., and Lecuit, M. (2017) Real-Time Whole-Genome Sequencing for Surveillance of *Listeria monocytogenes*, France. *Emerg Infect Dis* **23**: 1462-1470.

Nguyen, L.T., Schmidt, H.A., von 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.

Nishino, K., and Yamaguchi, A. (2001) Analysis of a complete library of putative drug transporter genes in Escherichia coli. *J Bacteriol* **183**: 5803-5812.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D.G. (2014) BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res* **42**: D737-743.

Palma, F.e.a. (2020) Dynamics of mobile genetic elements of *Listeria monocytogenes* persisting in ready-toeat seafood processing plants in France. *en préparation*.

Pearce, M.E., Alikhan, N.F., Dallman, T.J., Zhou, Z., Grant, K., and Maiden, M.C.J. (2018) Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *Int J Food Microbiol* **274**: 1-11.

Pietzka, A., Allerberger, F., Murer, A., Lennkh, A., Stoger, A., Cabal Rosel, A., Huhulescu, S., Maritschnik, S., Springer, B., Lepuschitz, S., Ruppitsch, W., and Schmid, D. (2019) Whole Genome Sequencing Based Surveillance of *L. monocytogenes* for Early Detection and Investigations of Listeriosis Outbreaks. *Front Public Health* **7**: 139.

Pightling, A.W., Petronella, N., and Pagotto, F. (2014) Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS One* **9**: e104579.

Pightling, A.W., Pettengill, J.B., Luo, Y., Baugher, J.D., Rand, H., and Strain, E. (2018) Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Front Microbiol* **9**: 1482.

Portmann, A.C., Fournier, C., Gimonet, J., Ngom-Bru, C., Barretto, C., and Baert, L. (2018) A Validation Approach of an End-to-End Whole Genome Sequencing Workflow for Source Tracking of *Listeria monocytogenes* and *Salmonella enterica*. *Front Microbiol* **9**: 446.

Pulsenet (2013a) Standard operating procedure for Pulsenet PFGE of *Escherichia coli* O157:H7, *Escherichia coli* non-O157 (STEC), *Salmonella* serotypes, *Shigella sonnei* and *Shigella flexneri*.

Pulsenet (2013b) Standard operating procedure for Pulsenet PFGE of *Listeria monocytogenes*.

Quereda, J.J., Leclercq, A., Moura, A., Vales, G., Gomez-Martin, A., Garcia-Munoz, A., Thouvenot, P., Tessaud-Rita, N., Bracq-Dieye, H., and Lecuit, M. (2020) Listeria valentina sp. nov., isolated from a water trough and the faeces of healthy sheep. *Int J Syst Evol Microbiol*.

Radomski, N., Cadel-Six, S., Cherchame, E., Felten, A., Barbet, P., Mallet, L., Le Hello, S., Weill, F.-X., Guillier, L., and Mistou, M. (2019a) 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. *en cours de soumission, Frontiers in Microbiology*.

Radomski, N., Cadel-Six, S., Cherchame, E., Felten, A., Barbet, P., Palma, F., Mallet, L., Le Hello, S., Weill, F.X., Guillier, L., and Mistou, M.Y. (2019b) 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.

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

Rehman, M.A., Yin, X., Persaud-Lachhman, M.G., and Diarra, M.S. (2017) First Detection of a Fosfomycin Resistance Gene, fosA7, in *Salmonella enterica* Serovar Heidelberg Isolated from Broiler Chickens. *Antimicrob Agents Chemother* **61**.

Ridom (2019) Ridom™ SeqSphere+ Software Product sheet https://www.ridom.de/seqsphere/Ridom_SeqSphere_ProductSheet.pdf.

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

Schwarze, K., Buchanan, J., Taylor, J.C., and Wordsworth, S. (2018) Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med* **20**: 1122-1130.

Sevellec, Y., (2018) Diversité génomique de *Salmonella* Derby en France. In., pp.

Silva, M., Machado, M.P., Silva, D.N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M., and Carrico, J.A. (2018) chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* **4**.

Singh, Y.S., A ; Kumar,R ;Saxena,M., (2018) Virulence System of *Salmonella*. In: almonella - A Re-emerging Pathogen. pp.

Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**: 4384-4388.

Sockett, P.N., and Roberts, J.A. (1991) The social and economic impact of salmonellosis. A report of a national survey in England and Wales of laboratory-confirmed *Salmonella* infections. *Epidemiol Infect* **107**: 335-347.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**: 472-482.

Tindall, B.J., Grimont, P.A., Garrity, G.M., and Euzeby, J.P. (2005) Nomenclature and taxonomy of the genus *Salmonella*. *Int J Syst Evol Microbiol* **55**: 521-524.

Van Cauteren, D.L.S., Y.; Buyand, M.; Tourdjman, M.; Jourdan-Da Silva, N.; Couturier, E.; Fournet, N.; De Valk, H.; Desenclos, J-C. (2017) Estimation de la morbidité et de la mortalité liées aux infections d'origine alimentaires en France métropolitaine, 2008-2013. *BEH*.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C*., et al.* (2001) The sequence of the human genome. *Science* **291**: 1304-1351.

Vignaud, M., Cherchame, E., Marault, M., Chaing, E., Le Hello, S., Michel, V., Jourdan-Da Silva, N., Lailler, R., Brisabois, A., and Cadel-Six, S. (2017) MLVA for *Salmonella enterica* subsp. enterica Serovar Dublin: Development of a Method Suitable for Inter-Laboratory Surveillance and Application in the Context of a Raw Milk Cheese Outbreak in France in 2012. *Front Microbiol* **8**: 295.

Vincent, C., Usongo, V., Berry, C., Tremblay, D.M., Moineau, S., Yousfi, K., Doualla-Bell, F., Fournier, E., Nadon, C., Goodridge, L., and Bekal, S. (2018) Comparison of advanced whole genome sequence-based methods to distinguish strains of *Salmonella enterica* serovar Heidelberg involved in foodborne outbreaks in Quebec. *Food Microbiol* **73**: 99-110.

WHO World Health Organization (2018) Whole genome sequencing for foodborne disease surveillance. Landscape paper.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., and Larsen, M.V. (2012) Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**: 2640-2644.

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

Zhou, Q., Feng, F., Wang, L., Feng, X., Yin, X., and Luo, Q. (2011) Virulence regulator PrfA is essential for biofilm formation in *Listeria monocytogenes* but not in *Listeria innocua*. *Curr Microbiol* **63**: 186-192.

Zhou, Z.A., N.; Mohamed, K.; Agama Study Group, the ; Achtman, M. (2019) The user's guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia*.

9 Annexes

Annexe 1 : Document Assurance Qualité du projet

anses
agence nationale de sécurité sanitaire alimentation, environnement, travail

Dispositions particulières pour la traçabilité des travaux de recherche	
Laboratoire	**Laboratoire de sécurité des aliments (LSAI)**
Unité	**Salmonella et Listeria (SEL)**
Activités concernées	Diplôme EPHE Claire YVON

1. Informations à consigner

	Information	Support d'enregistrement	Commentaires
1. Identification du projet	☒ Titre, acronyme ☒ Description ☒ Hypothèses/questions de recherche ☒ Chef de projet ☐ Équipe de recherche associée ☐ Partenaires de recherche associés	cf. dépôt de projet auprès de l'EPHE (X:\DPTMIC\USEL\3_1_Equipe_Recherche\5. Stagiaire_et_thésard\2018-2020\Claire-YVON\Administratif\EPHE_Claire_2019_2020)	Changement de tuteur scientifique courant 2019 D. Albert -> S. Cadel-Six
2. Protocole	☒ Date de début et fin de l'expérimentation ☒ Indication du protocole, du choix de la méthode ☐ Modifications des protocoles standards ☒ Identification du concepteur et/ou manipulateur ☒ Calculs et méthodes de traitement des données ☐ Références des travaux/expérimentations précédents (n° de cahier, pages) ☐ Liens entre les phases de l'expérimentation	Sur le réseau (X:\DPTMIC\USEL\3_1_Equipe_Recherche\5. Stagiaire_et_thésard\2018-2020\Claire-YVON)	
3. Ressources (identification, qualification)	☒ Équipements ☒ Réactifs ☒ Échantillons ☒ Environnement (locaux)	Fiches de paillasses raccordées au système qualité	
4. Mesures	☒ Relevés de mesures ☒ Valeurs de réglages, conditions opératoires ☒ Références et localisation de documents, photos, fichiers, données liées ☒ Description du déroulement des manipulations, même si elles n'ont pas abouti (fausses pistes) ☒ Faits et observations marquants ou inhabituels	Fiches de paillasses raccordées au système qualité	
5. Analyses	☒ Interprétations et commentaires sur les résultats obtenus (confirmations, non-confirmations, controverses) ☐ Nouvelles hypothèses de travail (idées, pistes, astuces ou résultats de réflexions collectives) ☐ Critique des résultats et idées d'amélioration	Fiches de paillasses raccordées au système qualité	
6. Autres éléments	☐ ..... ☐		

2. Règles d'enregistrements

	Finalité	Moyens	Dispositions spécifiques
1	Donner confiance dans la réalité des enregistrements	☒ Faire intervenir un témoin désigné	Nomination de deux encadrants : scientifique et pédagogique
		☐ qui atteste ce qu'il a examiné	Signature de chaque page du cahier de laboratoire
		☐ à des moments définis	Périodicité définie dans le tableau ANSES/PR3/A/02-02
2	Donner confiance dans l'authenticité des enregistrements	☐ Papier : signature	L'utilisateur signe chaque page du cahier de laboratoire et tous les documents papier rattachés
		☒ Papier : règles d'écriture	- utiliser une encre indélébile. Les corrections doivent être clairement barrées afin de rester lisibles. - signaler, dater, signer tout ajout ou modification. - ne pas arracher de pages - signaler tout saut de page ou page blanche intentionnelle par un trait en travers de la page - ne pas laisser d'espaces vides (susceptibles d'être remplis après signature)
		☒ Informatique : identifiant	Cf. charte informatique, login individuel Cf. engagement de confidentialité LSA-FGE-0005
		☒ Informatique : sécurisation des contenus	Sauvegarde journalière sur réseau Cf. procédure ANSES/PG/0070 « dispositions de sauvegarde »
		☒ Dispositif d'authentification complémentaire	Utilisation d'un site de signature électronique certifiée
3	Donner confiance dans le rattachement des documents joints	☒ Rattachement des documents (papier photos, graphiques,…)	X:\DPTMIC\USEL\3_1_Equipe_Recherche\5. Stagiaire_et_thésard\2018-2020\Claire-YVON
		☒ Rattachement des documents informatiques	X:\DPTMIC\USEL\3_1_Equipe_Recherche\5. Stagiaire_et_thésard\2018-2020\Claire-YVON
4	Faciliter la lecture des enregistrements	☒ Mention des sigles et acronymes	Définis dans les documents sous assurance qualité
		☐ Utilisation de sommaires	
5	Maîtriser l'accès aux informations	☐ Consultation des données par des tiers / envoi de copies de données	
		☐ Envoi de données hors du laboratoire	
		☐ Dispositions particulières liées à la nature des données détenues	
6	Conserver les informations	☒ Conservation des enregistrements papier	- lieu de conservation : bureau 213 - durée : 5 ans avant archivage selon ANSES/NO/A/03
		☒ Conservation des enregistrements informatiques	- données machines sur les machines pendant 5 ans - autres données sur le réseau pendant 10 ans
7	Donner l'assurance de la maîtrise de la traçabilité	☐ Organisation des enregistrements	
		☐ Réalisation d'audits internes	Projet mené par le correspondante qualité de l'unité

Le – 2 JUIL. 2020 Laurent LALOUX Directeur du laboratoire	Le 30/06/2020 Corinne DANAN Chef d'unité adjoint	Le 30/06/2020 Sabrina CADEL-SIX Chef de projet / Tuteur scientifique

Annexe 2 : Collection de génomes constituant le panel analysé de Salmonella Agona

Référence	Département	Date de prélèvement	Matrice	Contexte de prélèvement	Taille du génome	N50	Breadth coverage (%)	Référence	Département	Date de prélèvement	Matrice	Contexte de prélèvement	Taille du génome	N50	Breadth coverage (%)
2005LSAL01540	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4865590	1700460	91,5	2018LSAL00844	Mayenne	21/12/2017	Lait infantile en poudre	Alerte sanitaire	4849226	4843865	91,34
2005LSAL01541	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4865168	4434062	91,48	2018LSAL00845	Mayenne	22/12/2017	Lait infantile en poudre	Alerte sanitaire	4842233	4841951	91,26
2005LSAL01542	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4689283	2406088	91,5	2018LSAL00846	Mayenne	23/12/2017	Lait infantile en poudre	Alerte sanitaire	4849299	4843938	91,33
2005LSAL01574	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4865423	2089460	91,51	2018LSAL00847	Mayenne	24/12/2017	Lait infantile en poudre	Alerte sanitaire	4846206	4840845	91,3
2005LSAL01575	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4863868	2024295	91,51	2018LSAL00848	Mayenne	25/12/2017	Lait infantile en poudre	Alerte sanitaire	4848792	4843431	91,36
2005LSAL02331	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4883428	4878067	91,5	2018LSAL00849	Mayenne	26/12/2017	Lait infantile en poudre	Alerte sanitaire	4847974	4842365	91,33
2005LSAL02332	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4861287	1235184	91,51	2018LSAL00851	Mayenne	27/12/2017	Lait infantile en poudre	Alerte sanitaire	4848883	2824344	91,35
2005LSAL02333	Mayenne	2005	Lait infantile en poudre	Alerte sanitaire	4864608	1223117	91,48	2018LSAL00852	Mayenne	28/12/2017	Lait infantile en poudre	Alerte sanitaire	4847540	2823259	91,3
2006LSAL02273	Landes	05/04/2006	Viande porcine - hachée	Autocontrôle	4926431	1737546	91,58	2018LSAL00853	Mayenne	29/12/2017	Lait infantile en poudre	Alerte sanitaire	4850979	4845618	91,29
2006LSAL03967	France	30/05/2006	Plat préparé	Autocontrôle	4755761	4717258	91,57	2018LSAL00976	Mayenne	30/12/2017	Lait infantile en poudre	Alerte sanitaire	4849747	2651250	91,33
2008LSAL05963	Dordogne	01/09/2008	Viande de poulet	Autocontrôle	4754393	1997861	91,46	2018LSAL00977	Mayenne	31/12/2017	Lait infantile en poudre	Alerte sanitaire	4847925	4842564	91,27
2008LSAL07479	Finistère	12/11/2008	Inconnu	Autocontrôle	4960585	4955620	91,46	2018LSAL00978	Mayenne	01/12/2017	Lait infantile en poudre	Alerte sanitaire	4848519	4843158	91,33
2009LSAL03517	Nord	2009	Viande de poulet	Autocontrôle	4829152	4775497	91,47	2018LSAL00979	Mayenne	02/12/2017	Lait infantile en poudre	Alerte sanitaire	4848651	4843290	91,32
2009LSAL03518	Nord	2009	Viande de dinde	Autocontrôle	4825280	3869758	91,52	2018LSAL00980	Mayenne	03/12/2017	Lait infantile en poudre	Alerte sanitaire	4849339	4843730	91,3
2009LSAL04273	Ille-Et-Vilaine	12/06/2009	Dindes - Environnement d'élevage	Autocontrôle	4922155	3957100	91,52	2018LSAL00981	Mayenne	04/12/2017	Lait infantile en poudre	Alerte sanitaire	4849331	2990801	91,37
2009LSAL09074	Val de Marne	2009	Inconnu	EILA	4862744	2103502	91,1	2018LSAL00982	Mayenne	05/12/2017	Lait infantile en poudre	Alerte sanitaire	4848381	2823567	91,34
2010LSAL01477		30/03/2010	Plat préparé	Autocontrôle	4739342	4681898	90,95	2018LSAL00983	Mayenne	06/12/2017	Lait infantile en poudre	Alerte sanitaire	4851131	2824132	91,35
2012LSAL01803		14/05/2012	Crevettes grises	Autocontrôle	5106242	1235770	90,99	2018LSAL00984	Mayenne	07/12/2017	Lait infantile en poudre	Alerte sanitaire	4848545	4842943	91,32
2013LSAL02330	Drôme	21/05/2013	Matière première - dérivés de tournesol	Autocontrôle	4771860	1209193	91,55	2018LSAL00985	Mayenne	08/12/2017	Lait infantile en poudre	Alerte sanitaire	5064244	5058883	91,35

2013LSAL03067	Bouches-Du-Rhône	23/07/2013	Viande de volaille non précisé	Plan de surveillance	4717014	2684862	91,56	2018LSAL00986	Mayenne	09/12/2017	Lait infantile en poudre	Alerte sanitaire	4849473	4602324	91,31
2013LSAL03068	Bouches-Du-Rhône	23/07/2013	Viande de poulet	Autocontrôle	4715436	2404399	91,54	2018LSAL00987	Mayenne	10/12/2017	Lait infantile en poudre	Alerte sanitaire	4848335	2824083	91,31
2013LSAL03069	Bouches-Du-Rhône	16/07/2013	Viande de poulet	Autocontrôle	4715071	2754307	91,57	2018LSAL00988	Mayenne	11/12/2017	Lait infantile en poudre	Alerte sanitaire	4848408	4843047	91,35
2013LSAL03230	Morbihan	23/07/2013	Aliment pour animaux domestiques	Autocontrôle	4810237	2004314	91,55	2018LSAL00989	Mayenne	12/12/2017	Lait infantile en poudre	Alerte sanitaire	4849043	4843682	91,31
2013LSAL03290	Réunion	03/07/2013	Viande de poulet	Autocontrôle	4781650	2406615	91,56	2018LSAL00990	Mayenne	13/12/2017	Lait infantile en poudre	Alerte sanitaire	4849118	4843757	91,31
2013LSAL03598	Bouches-Du-Rhône	13/08/2013	Viande de poulet	Autocontrôle	4714279	2951446	91,54	2018LSAL00991	Mayenne	14/12/2017	Lait infantile en poudre	Alerte sanitaire	4848929	2826603	91,33
2013LSAL04304	Vendée	24/09/2013	Viande de poulet	Autocontrôle	4821965	2630223	91,54	2018LSAL00992	Mayenne	15/12/2017	Lait infantile en poudre	Alerte sanitaire	4842555	4842555	91,36
2014LSAL05711	Mayenne	22/10/2014	Viande de poulet	Autocontrôle	4862508	4810352	91,4	2018LSAL00993	Mayenne	16/12/2017	Lait infantile en poudre	Alerte sanitaire	4842210	4842210	91,33
2016LSAL02088	Inconnu	08/07/1905	Inconnu	Autocontrôle	4874905	2862611	90,96	2018LSAL00994	Mayenne	17/12/2017	Lait infantile en poudre	Alerte sanitaire	4851307	4842121	91,34
2016LSAL03561		19/09/2016	Semences germées	Autocontrôle	4794963	2516665	91,1	2018LSAL00995	Mayenne	18/12/2017	Lait infantile en poudre	Alerte sanitaire	4849151	4843790	91,36
2016LSAL03563		19/09/2016	Semences germées	Autocontrôle	4822903	3214527	91,11	2018LSAL00996	Mayenne	19/12/2017	Lait infantile en poudre	Alerte sanitaire	4848818	2825173	91,33
2016LSAL04644	Hautes-Alpes	28/11/2016	Matières premières - dérivé du blé	Autocontrôle	4855166	1937519	91,43	2018LSAL00997	Mayenne	20/12/2017	Lait infantile en poudre	Alerte sanitaire	5031190	3008136	91,3
2017LSAL04313	Morbihan	20/11/2017	Viande de Canards	Autocontrôle	4779318	2739998	91,5	2018LSAL00998	Mayenne	21/12/2017	Lait infantile en poudre	Alerte sanitaire	4848964	2825699	91,32
2017LSAL04379	Val-De-Marne	08/11/2017	Viande de cheval	Autocontrôle	4959664	1502361	91,53	2018LSAL00999	Mayenne	22/12/2017	Lait infantile en poudre	Alerte sanitaire	4848994	2019536	91,22
2017LSAL04394	Marne	24/11/2017	Viande de dinde	Autocontrôle	6222204	1272344	93,24	2018LSAL01000	Mayenne	23/12/2017	Lait infantile en poudre	Alerte sanitaire	4849213	4843852	91,33
2017LSAL04595	Mayenne	01/12/2017	Lait infantile en poudre	Alerte sanitaire	4848852	3456329	91,34	2018LSAL01001	Mayenne	24/12/2017	Lait infantile en poudre	Alerte sanitaire	4846494	2823353	91,24
2017LSAL04598	Mayenne	02/12/2017	Lait infantile en poudre	Alerte sanitaire	4849285	2824351	91,36	2018LSAL01004	Mayenne	25/12/2017	Lait infantile en poudre	Alerte sanitaire	4851866	4846505	91,33
2017LSAL04604	Mayenne	03/12/2017	Lait infantile en poudre	Alerte sanitaire	4850030	4568373	91,34	2018LSAL01005	Mayenne	26/12/2017	Lait infantile en poudre	Alerte sanitaire	4849002	4843641	91,35
2017LSAL04607	Mayenne	04/12/2017	Lait infantile en poudre	Alerte sanitaire	4849538	4844177	91,35	2018LSAL01006	Mayenne	27/12/2017	Lait infantile en poudre	Alerte sanitaire	4848919	2823761	91,3
2018LSAL00588	Mayenne	05/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4847706	2821899	91,33	2018LSAL01372	Mayenne	28/12/2017	Lait infantile en poudre	Alerte sanitaire	4849276	4843667	91,34
2018LSAL00589	Mayenne	06/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4848189	4842580	91,32	2018LSAL01433	Mayenne	13/12/2017	Lait infantile en poudre	Alerte sanitaire	4850196	4844835	91,27

2018LSAL00590	Mayenne	07/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4851270	2823550	91,32	2018LSAL02034	Mayenne	01/02/2018	Lait infantile en poudre	Alerte sanitaire	4836497	2812085	91,34
2018LSAL00591	Mayenne	08/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4849372	2761568	91,34	2018LSAL02035	Mayenne	02/02/2018	Lait infantile en poudre	Alerte sanitaire	5096512	3069488	91,34
2018LSAL00592	Mayenne	09/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4851844	2823778	91,33	2018LSAL02036	Mayenne	03/02/2018	Lait infantile en poudre	Alerte sanitaire	4837364	4831755	91,38
2018LSAL00593	Mayenne	10/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4842738	2824303	91,34	2018LSAL02082	Mayenne	04/02/2018	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	5107734	3083347	91,35
2018LSAL00594	Mayenne	11/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	5032231	3173310	91,33	2018LSAL02125	Mayenne	12/12/2017	Lait infantile en poudre	Alerte sanitaire	4844405	2079667	91,34
2018LSAL00595	Mayenne	12/12/2017	Lait infantile en poudre - environnement d'usine	Alerte sanitaire	4849169	2825065	91,29	2018LSAL02135	Mayenne	01/01/2018	Lait infantile en poudre	Alerte sanitaire	4847607	4842246	91,31
2018LSAL00836	Mayenne	13/12/2017	Lait infantile en poudre	Alerte sanitaire	4853862	2608644	91,34	2018LSAL02144	Mayenne	02/01/2018	Lait infantile en poudre	Alerte sanitaire	4849030	2827052	91,32
2018LSAL00837	Mayenne	14/12/2017	Lait infantile en poudre	Alerte sanitaire	4849552	4844191	91,34	2018LSAL03012	Gard	01/05/2018	Poulet - environnement d'élevage	Autocontrôle	5077836	5074536	90,92
2018LSAL00838	Mayenne	15/12/2017	Lait infantile en poudre	Alerte sanitaire	4851324	4845963	91,34	2018LSAL03013	Gard	02/05/2018	Poulet - environnement d'élevage	Autocontrôle	4718762	4715462	90,91
2018LSAL00839	Mayenne	16/12/2017	Lait infantile en poudre	Alerte sanitaire	4848226	4842865	91,31	2018LSAL03014	Gard	03/05/2018	Poulet - environnement d'élevage	Autocontrôle	4714259	3060705	90,91
2018LSAL00840	Mayenne	17/12/2017	Lait infantile en poudre	Alerte sanitaire	4853063	2827059	91,33	2018LSAL03015	Gard	04/05/2018	Poulet - environnement d'élevage	Autocontrôle	4713879	4713580	90,91
2018LSAL00841	Mayenne	18/12/2017	Lait infantile en poudre	Alerte sanitaire	4848043	2113899	91,38	2018LSAL03016	Gard	05/05/2018	Poulet - environnement d'élevage	Autocontrôle	4712427	2530595	90,9
2018LSAL00842	Mayenne	19/12/2017	Lait infantile en poudre	Alerte sanitaire	4849877	4844516	91,33	2018LSAL03313	Gard	28/06/2018	Poulet - environnement d'élevage	Autocontrôle	4729625	4170248	90,94
2018LSAL00843	Mayenne	20/12/2017	Lait infantile en poudre	Alerte sanitaire	4847650	4842289	91,3	ERR2219379	Inconnu	11/12/2017	Souche humaine	Alerte sanitaire			

Annexe 3 : Collection de génomes constituant le panel analysé de Listeria monocytogenes CC204

Référence	Pays	Lieu de prélèvement	Date de prélèvement	Matrice	Contexte de prélèvement	Taille du génome	N50	Breadth coverage (%)
A4-O2-LmUB3PA	France	Usine B	21/09/1998	Prélèvement de surface	Projet de recherche - persistance	2971846	2971788	97,43
B7201-P3-LmUB3PA	France	Usine A	25/11/1999	Saumon fumé	Projet de recherche - persistance	3044060	2971846	97,42
B7678-P1-LmUB3PA	France	Usine A	09/12/1999	Prélèvement de surface	Projet de recherche - persistance	3060855	2969015	97,42
B7853-P1-LmUB3PA	France	Usine A	16/11/1999	Saumon fumé	Projet de recherche - persistance	3060284	2970038	97,45
C3d-LmUB3PA	France	Usine A	17/03/2000	Saumon fumé	Projet de recherche - persistance	3125513	2969228	97,46
C5068-PT1-LmUB3PA	France	Usine A	17/11/2000	Saumon fumé	Projet de recherche - persistance	3368887	3034696	97,45
C5126-P-LmUB3PA	France	Usine A	23/11/2000	Saumon fumé	Projet de recherche - persistance	3059704	2973156	97,44
C5204-PT-LmUB3PA	France	Usine A	30/11/2000	Saumon fumé	Projet de recherche - persistance	3059495	2968887	97,45
C943-P-LmUB3PA	France	Usine A	16/02/2000	Prélèvement de surface	Projet de recherche - persistance	3059868	2968439	97,46
CL122-AS1-LmUB3PA	France	Usine B	15/05/2000	Saumon fumé	Projet de recherche - persistance	3124140	2969051	97,42
CL369-S2-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3059770	3033323	97,45
CL372-S1-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060256	2080697	97,43
CL377-S1-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060253	2969200	97,44
CL403-S4-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060747	2731752	97,42
CL414-AS1-LmUB3PA	France	Usine B	19/09/2000	Saumon fumé	Projet de recherche - persistance	3060197	2969930	97,42
CL72-P1-LmUB3PA	France	Usine B	15/05/2000	Prélèvement de surface	Projet de recherche - persistance	3059877	2969380	97,42
CL86-AP2-LmUB3PA	France	Usine B	15/05/2000	Saumon fumé	Projet de recherche - persistance	3059745	2969060	97,43
CS43C-LmUB3PA	France	Usine A	17/03/2000	Saumon fumé	Projet de recherche - persistance	3061382	2731251	97,45
CS504-CA1-LmUB3PA	France	Usine A	09/10/2000	Saumon fumé	Projet de recherche - persistance	3125164	2970565	97,46
CS536-PL1-LmUB3PA	France	Usine A	27/11/2000	Prélèvement de surface	Projet de recherche - persistance	3123891	3034347	97,41
D3768-PT-LmUB3PA	France	Usine B	27/11/2001	Saumon fumé	Projet de recherche - persistance	3059789	3033074	94,86
E3422-OT-LmUB3PA	France	Usine B	21/11/2002	Saumon fumé	Projet de recherche - persistance	3060351	1762359	94,79
E3889-LmUB3PA	France	Usine B	24/12/2002	Saumon fumé	Projet de recherche - persistance	3060222	2969534	97,4
F49-OS-LmUB3PA	France	Usine B	09/01/2003	Saumon fumé	Projet de recherche - persistance	3059946	2969405	94,27
L3075	République Tchèque		2011	Humain	Alerte sanitaire	2936542	2969129	96,83
L3100	République Tchèque		2011	Humain	Alerte sanitaire	2947863	2941627	96,92
LV0294	République Tchèque	Usine 2	2014	Prélèvement de surface - industrie laitière	Alerte sanitaire	2945464	2919755	96,96
LV0484	République Tchèque	Usine 2	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2944908	2935269	97
LV0544	République Tchèque	Usine 2	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2950113	2932565	96,99
LV0599	République Tchèque	Usine 1	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2947845	2927870	96,93

LV0600	République Tchèque	Usine 1	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931908	2931908	97,17
LV0611	République Tchèque	Usine 2	2016	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932321	2932321	97,23
LV0826	République Tchèque	Usine 2	2007	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932242	2932242	97,2
LV0827	République Tchèque	Usine 1	2004	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932587	2932587	97,22
LV0868	République Tchèque	Usine 1	2008	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932542	2932049	97,23
LV0869	République Tchèque	Usine 2	2011	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929978	2929978	97,16
LV0871	République Tchèque	Usine 1	2010	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932168	2932168	97,22
LV0872	République Tchèque	Usine 1	2011	Prélèvement de surface - industrie laitière	Alerte sanitaire	3024376	2932170	97,23
LV0876	République Tchèque	Usine 2	2006	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931674	2931674	97,23
LV0877	République Tchèque	Usine 2	2008	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931007	2931007	97,19
LV0879	République Tchèque	Usine 2	2009	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929554	2929554	97,09
LV0880	République Tchèque	Usine 2	2010	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931974	2931974	97,24
LV0884	République Tchèque	Usine 1	2009	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929906	2929906	97,14
11CEB216LM	France		07/03/2011	Crudités	Surveillance	2973985	2970822	97,59
11CEB450LM	France		18/07/2011	Prélèvement de surface	Surveillance	3073230	2471855	97,42
12CEB03LM	France		03/01/2012	Poisson fumé	Surveillance	3061083	2970220	97,51
RL15000020	France		Inconnu	Inconnu	Projet de recherche	3057601	2966907	97,54
RL15000131	France		Inconnu	Inconnu	Projet de recherche	3058155	2637563	97,29
RL15000216	France		Inconnu	Inconnu	Projet de recherche	3050855	2968996	97,58
RL15000303	France		Inconnu	Inconnu	Projet de recherche	3219586	3012912	97,6
RL15000348	France		Inconnu	Inconnu	Projet de recherche	3110607	3019536	97,6
RL15000467	France		Inconnu	Inconnu	Projet de recherche	3027378	1749325	97,35
RL15000468	France		Inconnu	Inconnu	Projet de recherche	3026015	2244483	97,36
RL15000481	France		Inconnu	Inconnu	Projet de recherche	3072591	1708022	97,34
RL15000730	France		Inconnu	Inconnu	Projet de recherche	3024619	2197898	97,36
RL15000732	France		Inconnu	Inconnu	Projet de recherche	2933282	1657477	97,4
RL15001353	France		Inconnu	Inconnu	Projet de recherche	3025382	2849653	97,41
17SEL373LM	France		30/05/2017	Canard laqué	Surveillance	3066528	2975472	97,09
17SEL510LM	France		19/09/2017	Canard laqué	Surveillance	3062605	2971788	97,8

Référence	Pays	Lieu de prélèvement	Date de prélèvement	Matrice	Contexte de prélèvement	Taille du génome	N50	Breadth coverage (%)
A4-O2-LmUB3PA	France	Usine B	21/09/1998	Prélèvement de surface	Projet de recherche - persistance	2971846	3E+06	97,43

B7201-P3-LmUB3PA	France	Usine A	25/11/1999	Saumon fumé	Projet de recherche - persistance	3044060	3E+06	97,42
B7678-P1-LmUB3PA	France	Usine A	09/12/1999	Prélèvement de surface	Projet de recherche - persistance	3060855	3E+06	97,42
B7853-P1-LmUB3PA	France	Usine A	16/11/1999	Saumon fumé	Projet de recherche - persistance	3060284	3E+06	97,45
C3d-LmUB3PA	France	Usine A	17/03/2000	Saumon fumé	Projet de recherche - persistance	3125513	3E+06	97,46
C5068-PT1-LmUB3PA	France	Usine A	17/11/2000	Saumon fumé	Projet de recherche - persistance	3368887	3E+06	97,45
C5126-P-LmUB3PA	France	Usine A	23/11/2000	Saumon fumé	Projet de recherche - persistance	3059704	3E+06	97,44
C5204-PT-LmUB3PA	France	Usine A	30/11/2000	Saumon fumé	Projet de recherche - persistance	3059495	3E+06	97,45
C943-P-LmUB3PA	France	Usine A	16/02/2000	Prélèvement de surface	Projet de recherche - persistance	3059868	3E+06	97,46
CL122-AS1-LmUB3PA	France	Usine B	15/05/2000	Saumon fumé	Projet de recherche - persistance	3124140	3E+06	97,42
CL369-S2-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3059770	3E+06	97,45
CL372-S1-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060256	2E+06	97,43
CL377-S1-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060253	3E+06	97,44
CL403-S4-LmUB3PA	France	Usine B	19/09/2000	Prélèvement de surface	Projet de recherche - persistance	3060747	3E+06	97,42
CL414-AS1-LmUB3PA	France	Usine B	19/09/2000	Saumon fumé	Projet de recherche - persistance	3060197	3E+06	97,42
CL72-P1-LmUB3PA	France	Usine B	15/05/2000	Prélèvement de surface	Projet de recherche - persistance	3059877	3E+06	97,42
CL86-AP2-LmUB3PA	France	Usine B	15/05/2000	Saumon fumé	Projet de recherche - persistance	3059745	3E+06	97,43
CS43C-LmUB3PA	France	Usine A	17/03/2000	Saumon fumé	Projet de recherche - persistance	3061382	3E+06	97,45
CS504-CA1-LmUB3PA	France	Usine A	09/10/2000	Saumon fumé	Projet de recherche - persistance	3125164	3E+06	97,46
CS536-PL1-LmUB3PA	France	Usine A	27/11/2000	Prélèvement de surface	Projet de recherche - persistance	3123891	3E+06	97,41
D3768-PT-LmUB3PA	France	Usine B	27/11/2001	Saumon fumé	Projet de recherche - persistance	3059789	3E+06	94,86
E3422-OT-LmUB3PA	France	Usine B	21/11/2002	Saumon fumé	Projet de recherche - persistance	3060351	2E+06	94,79
E3889-LmUB3PA	France	Usine B	24/12/2002	Saumon fumé	Projet de recherche - persistance	3060222	3E+06	97,4
F49-OS-LmUB3PA	France	Usine B	09/01/2003	Saumon fumé	Projet de recherche - persistance	3059946	3E+06	94,27

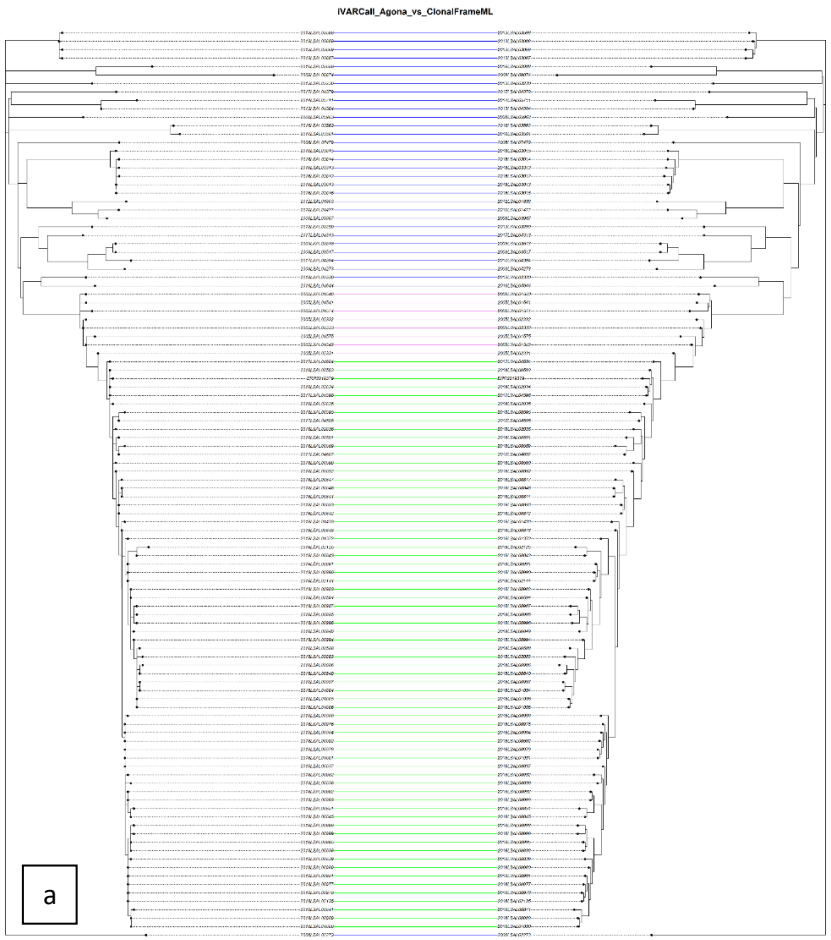
L3075	République Tchèque		2011	Humain	Alerte sanitaire	2936542	3E+06	96,83
L3100	République Tchèque		2011	Humain	Alerte sanitaire	2947863	3E+06	96,92
LV0294	République Tchèque	Usine 2	2014	Prélèvement de surface - industrie laitière	Alerte sanitaire	2945464	3E+06	96,96
LV0484	République Tchèque	Usine 2	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2944908	3E+06	97
LV0544	République Tchèque	Usine 2	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2950113	3E+06	96,99
LV0599	République Tchèque	Usine 1	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2947845	3E+06	96,93
LV0600	République Tchèque	Usine 1	2015	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931908	3E+06	97,17
LV0611	République Tchèque	Usine 2	2016	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932321	3E+06	97,23
LV0826	République Tchèque	Usine 2	2007	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932242	3E+06	97,2
LV0827	République Tchèque	Usine 1	2004	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932587	3E+06	97,22
LV0868	République Tchèque	Usine 1	2008	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932542	3E+06	97,23
LV0869	République Tchèque	Usine 2	2011	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929978	3E+06	97,16
LV0871	République Tchèque	Usine 1	2010	Prélèvement de surface - industrie laitière	Alerte sanitaire	2932168	3E+06	97,22
LV0872	République Tchèque	Usine 1	2011	Prélèvement de surface - industrie laitière	Alerte sanitaire	3024376	3E+06	97,23
LV0876	République Tchèque	Usine 2	2006	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931674	3E+06	97,23

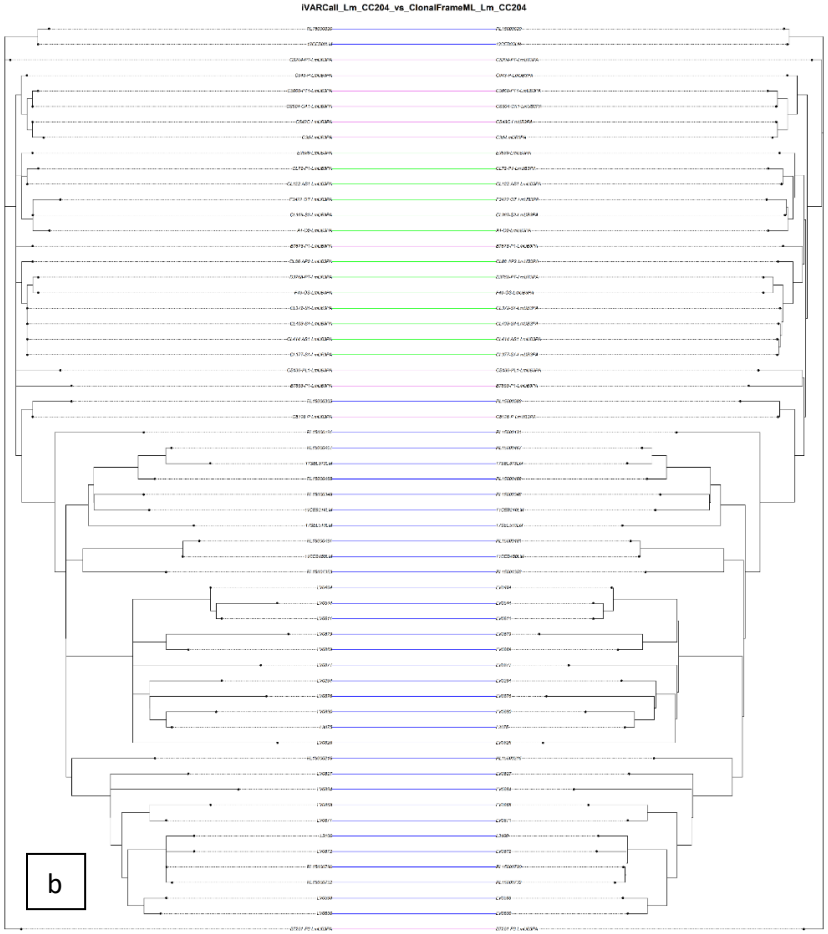
LV0877	République Tchèque	Usine 2	2008	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931007	3E+06	97,19
LV0879	République Tchèque	Usine 2	2009	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929554	3E+06	97,09
LV0880	République Tchèque	Usine 2	2010	Prélèvement de surface - industrie laitière	Alerte sanitaire	2931974	3E+06	97,24
LV0884	République Tchèque	Usine 1	2009	Prélèvement de surface - industrie laitière	Alerte sanitaire	2929906	3E+06	97,14
11CEB216LM	France		07/03/2011	Crudités	Surveillance	2973985	3E+06	97,59
11CEB450LM	France		18/07/2011	Prélèvement de surface	Surveillance	3073230	2E+06	97,42
12CEB03LM	France		03/01/2012	Poisson fumé	Surveillance	3061083	3E+06	97,51
RL15000020	France		Inconnu	Inconnu	Projet de recherche	3057601	3E+06	97,54
RL15000131	France		Inconnu	Inconnu	Projet de recherche	3058155	3E+06	97,29
RL15000216	France		Inconnu	Inconnu	Projet de recherche	3050855	3E+06	97,58
RL15000303	France		Inconnu	Inconnu	Projet de recherche	3219586	3E+06	97,6
RL15000348	France		Inconnu	Inconnu	Projet de recherche	3110607	3E+06	97,6
RL15000467	France		Inconnu	Inconnu	Projet de recherche	3027378	2E+06	97,35
RL15000468	France		Inconnu	Inconnu	Projet de recherche	3026015	2E+06	97,36
RL15000481	France		Inconnu	Inconnu	Projet de recherche	3072591	2E+06	97,34
RL15000730	France		Inconnu	Inconnu	Projet de recherche	3024619	2E+06	97,36
RL15000732	France		Inconnu	Inconnu	Projet de recherche	2933282	2E+06	97,4
RL15001353	France		Inconnu	Inconnu	Projet de recherche	3025382	3E+06	97,41
17SEL373LM	France		30/05/2017	Canard laqué	Surveillance	3066528	3E+06	97,09
17SEL510LM	France		19/09/2017	Canard laqué	Surveillance	3062605	3E+06	97,8

Annexe 4 : Références utilisées pour les différents SPI

Salmonella pathogenicity island	length (kpb)	genes nb	organism	accession number
spi1	38	39	*S.* Typhimurium	KP279311.1
spi2 common	12	10	*S.* Typhimurium	AJ224978.1
spi2 specific	25	31	*S.* Typhimurium	KP258194.1
spi3	16	10	*S.* Typhimurium	AF106566.1
spi4	27	10	*S.* Typhimurium	KP234070.1
spi5	9	8	*S.* Typhimurium	AY144492.1
spi6	47	52	*S.* Typhi	AL513382.1
spi7	134	75	*S.* Typhi	AL513382.1
spi8	8	12	*S.* Typhi	AL513382.1
spi9	16	4	*S.* Typhi	AL513382.1
spi10	33	24	*S.* Typhi	AL513382.1
spi11	14	19	*S.* Choleraesuis	AE017220.1
spi12	16	12	*S.* Choleraesuis	AE017220.1
spi13	25	19	*S.* gallinarum	AM933173.1
spi14	9	6	*S.* Typhimurium	AE006468.2
spi15	6,5	5	*S.* Typhi	AL513382.1
spi16	4,5	6	*S.* Typhimurium	AE006468.2
spi17	5	7	*S.* Typhi	AL513382.1
spi18	2	2	*S.* Typhi	AL513382.1
spi19	42	30	*S.* gallinarum	AM933173.1
spi23	36	41	*S.* Derby	LAZB00000000
CS54	25	8	*S.* Typhimurium	AE006468.2

Annexe 5 : Comparaison des arbres phylogénétiques obtenus entre iVARCall2 et ClonalFrameML





(a) A gauche l'arbre phylogénique issu de l'analyse SNP calling par iVARCall2 puis IQ-tree, à droite l'arbre tenant compte des évènements de recombinaison issu de ClonalFrameML. En bleu, les liens entre les souches non reliées épidémiologiquement aux souches des deux alertes sanitaires ; en violet, les liens entre les souches de l'alerte sanitaire de 2005 ; en vert, les liens entre les souches de l'alerte sanitaire de 2018.

(b) A gauche l'arbre phylogénique issu de l'analyse SNP calling par iVARCall2 puis IQ-tree, à droite l'arbre tenant compte des évènements de recombinaison issu de ClonalFrameML. En bleu, les liens entre les souches non reliées épidémiologiquement aux souches des deux usines d'intérêt ; en violet, les liens entre les souches provenant de l'usine A ; en vert, les liens entre les souches provenant de l'usine B.

Annexe 6 : Liste des documents qualités révisés et/ou rédigés dans le cadre du projet

L'ensemble des documents listés sont disponibles pour consultation, à la demande.

- LSA-INS-1227 « Extraction d'ADN génomique pour *Listeria* et *Salmonella* »
- LSA-INS-1450 « Analyse cgMLST via Ridom SeqSphere+ »
- LSA-INS-1481 « Analyse "SNP calling" sur le cluster de calcul Linux : iVARCall2 » - en cours de signature
- LSA-INS-0412 « Purification et vérification des souches de *Salmonella* » - en cours de signature
- LSA-INS-1177 « Programme d'habilitation – biologie moléculaire » - en cours de révision
- LSA-PS-0228 « Traitement des données brutes de séquençage »
- LSA-FGE-0393 « Modèle de rapport d'analyse SEL WGS (cgMLST) »
- LSA-FSE-1276 « Fiche de suivi analytique Microbiologie et extraction de l'ADN génomique »
- LSA-FSE-0524 « Fiche de suivi analytique : vérification de la pureté des souches de *Listeria monocytogenes* en vue de la conservation ou de la mise en analyse »
- LSA-FSE-1605 « Fichier d'enregistrement NGS_MAPA_AAAA »
- LSA-FSE-1606 « Fichier d'enregistrement NGS_externe_AAAA »
- LSA-FSE-1626 « Fichier d'enregistrement NGS_IDPA_AAAA »

ÉCOLE PRATIQUE DES HAUTES ÉTUDES
SCIENCES DE LA VIE ET DE LA TERRE

Quelle méthode analytique WGS choisir en vue de l'investigation d'évènements sanitaires ? Comparaison d'outils bio-informatiques en vue d'une validation et accréditation de méthode.

YVON Claire

Date de soutenance 15/10/2020

RÉSUMÉ

Salmonella et *Listeria monocytogenes* sont deux pathogènes majeurs de l'industrie agro-alimentaire, situés dans le top cinq des bactéries responsables de zoonoses d'après le rapport de l'EFSA/ECDC (2018). Dans le cadre de ses mandats de référence Français et Européens, l'unité SEL de l'Anses est amenée à répondre aux cas d'investigations sanitaires liés à ces deux pathogènes. Initialement menées grâces aux techniques de typage bactérien conventionnelles (PCR, MLVA, PFGE), ces investigations ont pris le virage vers la génomique. Les analyses réalisées sont aujourd'hui menées grâce au Whole Genome Sequencing (WGS), permettant d'obtenir des résultats à fort pouvoir discriminant. Les étapes dites de « wet-lab » allant de l'extraction d'ADN de qualité génomique au séquençage sont d'ores et déjà maitrisées par le laboratoire, bien que quelques ajustements aient été apporté. A travers ce projet, j'ai cherché à comparer les outils bio-informatiques à notre disposition. A l'aide de comparaison d'arbres phylogénétiques, d'analyses statistiques, d'analyses des matrices de distance j'ai pu prendre en compte les spécificités de chaque outil et de comparer les résultats obtenus. Les analyses réalisées en cgMLST et en *SNP calling* ont permis de mettre en évidence des clusterings comparables, quelques soit l'outil utilisé. Toutefois, les analyses de *SNP calling* permettent une analyse plus discriminante pouvant permettre d'apporter une réponse plus précise dans le cadre de crise sanitaire majeure. La revue des outils disponibles au laboratoire m'a également permis de tester les outils permettant la recherche de gènes de résistance, virulence et de persistance. Ces analyses permettent de compléter les analyses de source attribution dans le cadre d'alerte, afin de mieux comprendre les raisons de l'apparitions de clones émergents.

MOTS-CLÉS : Séquençage ; alertes sanitaires ; WGS ; *Listeria* ; *Salmonella*