Développement d'outils bio-informatiques décisionnels en médecine vétérinaire : application au modèle Tenacibaculum, agent pathogène de poissons marins

Sébastien Bridel

▶ **To cite this version:**

Sébastien Bridel. Développement d'outils bio-informatiques décisionnels en médecine vétérinaire : application au modèle Tenacibaculum, agent pathogène de poissons marins. Médecine vétérinaire et santé animale. Université Paris-Saclay, 2020. Français. NNT : 2020UPASV012 . tel-02971847

HAL Id: tel-02971847
https://theses.hal.science/tel-02971847

Submitted on 19 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse de doctorat

NNT : 2020UPASV012

université PARIS-SACLAY

Développement d'outils bio-informatiques décisionnels en médecine vétérinaire : application au modèle *Tenacibaculum*, agent pathogène de poissons marins

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577, Structure et dynamique des systèmes vivants, ED SDSV

Spécialité de doctorat : Sciences de la vie et de la santé

Unité de recherche : Université Paris-Saclay, UVSQ, INRAE, VIM, 78350, Jouy-en-Josas, France.

Référent : Université de Versailles -Saint-Quentin-en-Yvelines

Thèse présentée et soutenue à Jouy-en-Josas, le 29/06/2020, par

Sébastien BRIDEL

Composition du Jury

Michel-Yves MISTOU Directeur de Recherche, INRA	Président & Rapporteur
Frédérique Le ROUX Directrice de Recherche, IFREMER	Rapporteur & Examinatrice
Xavier BAILLY Ingénieur de Recherche, INRA	Examinateur
Pierre BREZELLEC Maître de Conférence, UVSQ	Examinateur
François THOMAS Chargé de Recherche, CNRS	Examinateur
Eric DUCHAUD Directeur de Recherche, INRA	Directeur de thèse
Sophie PASEK Maître de Conférence, MNHN	Co-encadrante
Pierre-Yves MOALIC Directeur Scientifique, Labofarm	Invité

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Développement d'outils bio-informatiques décisionnels en médecine vétérinaire : application au modèle Tenacibaculum, agent pathogène de poissons marins

Mots clés : *Tenacibaculum*, bioinformatique, spectrométrie de masse, aquaculture, épidémiologie

Résumé : Le genre *Tenacibaculum* (famille des *Flavobacteriaceae*, phylum *Bacteroidetes*), proposé en 2001 par Suzuki et al., comprend aujourd'hui 28 espèce valides. Ces bactéries sont exclusivement retrouvées dans les milieux marins, libres ou associées à des organismes tels que les poissons, les macroalgues et les invertébrés. Huit espèces sont pathogènes pour les poissons et essentiellement isolées dans des élevages. Ces bactéries sont à l'origine de maladies aux symptômes similaires collectivement désignées sous le nom de ténacibaculoses. L'identification basée sur des caractères phénotypiques n'est pas suffisamment discriminante et ne permet généralement pas d'identifier l'espèce responsable de la pathologie lors d'épisodes infectieux dans les élevages. L'objectif central de mon projet doctoral était de comprendre la prévalence de ces bactéries, notamment à travers le développement d'outils bioinformatiques décisionnels en médecine vétérinaire. Dans un premier temps, j'ai valorisé l'effort de séquençage important initié par l'équipe avant mon arrivée pour sur les bactéries appartenant au genre *Tenacibaculum*. En premier lieu, j'ai vérifié la taxonomie actuelle de l'ensemble du genre (espèces décrites et génomes publiés) à l'aide d'un ensemble de méthodes de phylogénomique comme l'*Average Nucleotide Identity*. Ce travail a permis de nous intéresser à un groupe d'espèces très proches, *T. dicentrarchi*, *T. finnmarkense* et *T. piscium* et a mis en évidence un exemple illustrant la théorie émise par Habib et al. (2014) soutenant l'idée d'une évolution parallèle de la pathogénicité ainsi que des acquisitions multiples de facteurs de virulence au sein du genre *Tenacibaculum*. Il semble également qu'il existe des phénomènes de convergence évolutive pour certains facteurs de virulence (gènes totalement différents mais codant pour la même fonction). J'ai ensuite développé une méthode d'identification et de typage d'isolats bactériens appartenant au genre *Tenacibaculum* par spectrométrie de masse MALDI-TOF. Cette méthode a comme principaux avantages d'être fiable, rapide et peu coûteuse. Pour développer la méthode d'identification, les spectres des références des souche types de la quasi-totalité des espèces du genre *Tenacibaculum* ont été obtenus

L'identification d'un isolat de terrain repose dès lors sur une quantification de la ressemblance entre l'empreinte spectrale de l'échantillon et les empreintes spectrales de références. Afin de développer une méthode de typage des souches à l'intérieur de l'espèce *T. maritimum*, nous avons en premier lieu utilisé les génomes complets de 25 souches (dont 22 obtenus dans le cadre de ce projet) afin d'évaluer la diversité intra-spécifique. La méthode de MALDI-typage développée repose sur l'utilisation de biomarqueurs (protéines ribosomiques polymorphes). J'ai cherché à exploiter au maximum le potentiel de la spectrométrie de masse MALDI-TOF en combinant les informations tirées des génomes avec les empreintes spectrales. J'ai choisi de définir un MALDI-Type (MT) comme étant une combinaison unique de 9 biomarqueurs en m'inspirant de la technique de MLST. Dans une collection de 130 isolats de terrain, nous avons ainsi pu identifier 20 MT qui se regroupent en 4 MALDI-groupes distincts. Enfin, j'ai développé une application web dénommée MALDIquantTypeR intégrant : la base de données de référence pour les espèces appartenant au genre *Tenacibaculum,* les outils d'identification ainsi que l'outil de typage dédié à l'espèce *T. maritimum* développés dans le cadre de ce projet. Il doit permettre à des chercheurs parfois géographiquement éloignés d'analyser directement des données brutes issues d'un spectromètre de masse. Nous pensons que la spectrométrie de masse MALDI-TOF peut être utilisée à la fois dans le cadre d'études épidémiologiques à large échelle, mais également être intégrée dans le diagnostic vétérinaire de routine. Ce projet a fait appel à différentes approches reposant essentiellement sur le séquençage et l'analyse de génomes complets, ainsi que sur la protéomique (spectrométrie de masse MALDI-TOF). Le développement des outils d'analyses a également nécessité des approches bio-informatiques ainsi que de la programmation.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

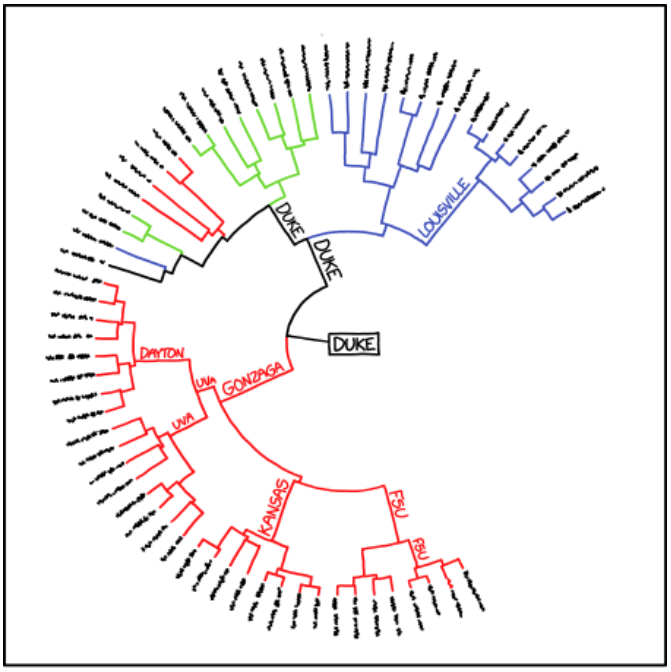
université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title: Bioinformatics tools for decisional help in veterinary medicine: fish pathogens of the genus *Tenacibaculum* as models

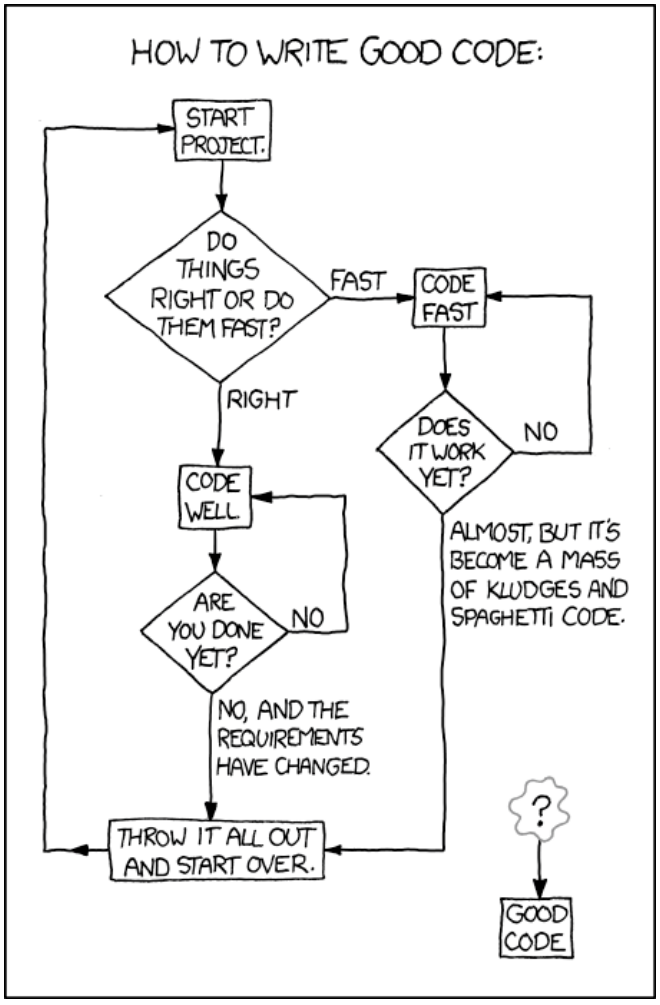
Keywords: *Tenacibaculum,* epidemiology, bioinformatics, mass spectrometry, aquaculture

Abstract: The genus *Tenacibaculum* (family *Flavobacteriaceae*, phylum *Bacteroidetes*) was proposed in 2001 by Suzuki et al. and currently includes 28 valid species. These bacteria are found exclusively in marine environments, either free or associated with organisms such as fish, macroalgae and invertebrates. Eight species are pathogenic to fish and are mainly isolated in farms. These bacteria cause diseases with similar symptoms that are collectively referred to as tenacibaculosis. Identification based on phenotypic characteristics is not sufficiently discriminating and generally does not make it possible to identify the species responsible for the pathology during infectious episodes in farms. The central objective of my doctoral project was to understand the prevalence of these bacteria, particularly through the development of bioinformatics decision-making tools in veterinary medicine. First, I developed upon the important sequencing effort initiated by the team before my arrival on bacteria belonging to the genus Tenacibaculum. I verified the current taxonomy of the entire genus (described species and published genomes) using a set of phylogenomic methods such as the Average Nucleotide Identity. This work has allowed us to focus on a group of closely related species, T. dicentrarchi, T. finnmarkense and T. piscium, and has highlighted an example illustrating the theory put forward by Habib et al (2014) supporting the idea of a parallel evolution of pathogenicity as well as multiple acquisitions of virulence factors within the genus Tenacibaculum. It also seems that evolutionary convergence phenomena have taken place for some virulence factors (i.e., totally different genes but encoding the same function). I then developed a method for identifying and typing bacterial isolates belonging to the genus Tenacibaculum using MALDI-TOF mass spectrometry. The main advantages of this method are its reliability, speed and low cost. To develop the identification method, the reference spectra of the type strains of most (24 out of 28) species of the genus *Tenacibaculum* were obtained. The identification of a field isolate is therefore based on a quantification of the similarity between the spectral footprint of the sample and the reference spectral footprints. In order to develop a typing method for strains within the *T. maritimum* species, we first used the complete genomes of 25 strains (including 22 obtained in this project) to assess intraspecific diversity. The MALDI-typing method developed is based on the use of biomarkers (polymorphic ribosomal proteins). I have sought to maximize the potential of MALDI-TOF mass spectrometry by combining information from genomes with spectral fingerprints. I chose to define a MALDI-Type (MT) as a unique combination of 9 biomarkers based on the MLST technique. In a collection of 130 field isolates, we were able to identify 20 MT that are grouped into 4 distinct MALDI-groups. Finally, I developed a web application called MALDIquantTypeR that integrates: the reference database for species belonging to the genus *Tenacibaculum*, the identification tools as well as the typing tool dedicated to the species *T. maritimum* developed as part of this project. It should allow geographically distant researchers to directly analyze raw data from any mass spectrometer. We believe that MALDI-TOF mass spectrometry can be used both in large-scale epidemiological studies and in routine veterinary diagnosis. This project used different approaches based mainly on sequencing and analysis of complete genomes, as well as proteomics (MALDI-TOF mass spectrometry). The development of analytical tools has also required bioinformatics approaches and programming.



I WAS KICKED OFF THE BIOLOGY PROJECT AFTER I SECRETLY REPLACED ALL THE PHYLOGENETIC TREES IN OUR NEW PAPER WITH MARCH MADNESS BRACKETS.

https://xkcd.com/2269/ - Phylogenetic Tree



https://xkcd.com/844/ - Good Code

“People think computers will keep them from making mistakes. They're wrong. With computers you make mistakes faster.”

Adam Osborn

“Il ne faut, à aucun prix, renoncer à l'écriture, ne serait-ce que parce que dans toutes les grandes civilisations l'écriture est liée à la calligraphie, c'est-à-dire à la beauté.”

De René Etiemble / Entretien avec Jacques Chancel - Décembre 1976

« Programmer est un monologue solitaire avec une machine froide et inhumaine. Pourtant, lorsque mes pensées finissent par produire un programme fonctionnel, je ne puis que me réjouir de la force combinée de l'étincelle créatrice de l'esprit humain et de la froideur calculatrice d'un processeur multi-cœurs. »

Sébastien Bridel

~~~~~~~~~~~~~~~~~

« Un Anneau pour les gouverner tous, un Anneau pour les trouver.

Un Anneau pour les amener tous et dans les ténèbres les lier. »

Le seigneur des anneaux : la communauté de l'anneau, J.R.R Tolkien
~~~~~~~~~~~~~~~~~

Remerciements

Je remercie chaleureusement toutes les personnes qui m'ont aidé pendant l'élaboration de ma thèse et notamment mon directeur de thèse, Monsieur Eric Duchaud, pour son intérêt et son soutien, sa grande disponibilité et ses nombreux conseils durant la rédaction de ma thèse. Sa rigueur scientifique m'a inspiré et m'a permis de réaliser ce projet en toute sérénité.

Je voudrais également remercier ma co-encadrante de thèse, Madame Sophie Pasek, pour ces précieux conseils en bioinformatique et son soutien pendant les trois années de mon projet doctoral.

Ce travail n'aurait pas été possible sans la collaboration de Labofarm du groupe Finalab, représenté par Monsieur Pierre-Yves Moalic ainsi que l'Association Nationale de la Recherche et de la Technologie, qui m'ont permis, grâce à un contrat doctoral CIFRE, de me consacrer sereinement à l'élaboration de ma thèse.

Ce travail n'aurait pu être mené à bien sans la disponibilité et l'accueil chaleureux que m'ont témoigné Jean-François Bernardet de l'équipe Immunité et Infection des Poissons pour son expertise en bactériologie et en rédaction, Frédérique Bourgeon responsable du spectromètre de masse MALDI-TOF à Bio Chêne Vert du groupe Finalab, Arnaud Marie technicien à Labofarm et maître de la culture *in vitro* des bactéries du genre *Tenacibaculum*, Eric Monvert et Joel Abusquier pour leurs précieux conseils en informatiques.

Je voudrais également remercier l'ensemble de l'équipe Immunité et Infection des Poissons, l'ensemble du personnel de l'unité Virologie et Immunologie Moléculaires ainsi que l'équipe de l'Atelier de BioInformatique qui m'accueillirent avec bienveillance dans le cadre de ce projet.

Au terme de ce parcours, je remercie enfin celles et ceux qui me sont chers et que j'ai quelque peu délaissés ces derniers mois pour achever cette thèse. Leurs attentions et encouragements m'ont accompagnée tout au long de ces années. Je suis redevable à mes parents, Élisabeth et Christian Bridel, pour leur soutien moral et matériel et leur confiance indéfectible dans mes choix. J'aimerais également remercier ma cousine Emma Fourdrignier et son entreprise 333 Studios pour avoir dessiner le logo de l'application MALDIquantTyperR.

Table des matières

Remerciements ***1***

Table des figures ***4***

Liste des abréviations ***5***

Avant-propos ***6***

I – Introduction ***7***

A – Introduction générale **7**
A1 – L'aquaculture, chiffres clés 7
A2 – Aquaculture : risques et défis 8

B – Les flavobactéries **10**

C – Les bactéries du genre *Tenacibaculum* **11**
C1 – Caractéristiques phénotypiques 11
C2 – Les espèces associées à des épisodes infectieux chez des poissons d'élevage 12
C3 – Les autres espèces associées à des pathologies 14
C4 – Les espèces isolées de l'environnement vraisemblablement non pathogènes 14

D – Taxonomie et espèce bactérienne **15**
D1 – Taxonomie bactérienne 15
D2 – Qu'est-ce qu'une espèce bactérienne ? 16
D3 – Approche polyphasique de la description d'une espèce 16

E – Identification et typage bactérien **20**
E1 – Introduction 20
E2 – Outils d'identification 21
E3 – Outils de typage 21
E4 – Conclusion. 31

F – La spectrométrie de masse MALDI-TOF **32**
F1 – Principe de fonctionnement 32
F2 – Applications de la spectrométrie de masse MALDI-TOF en microbiologie. 35
F3 – Nature des données spectrales. 40

II – Objectifs des travaux du projet doctoral ***43***

III – Résultats ***44***

Partie 1 : Étude génomique des espèces appartenant au genre *Tenacibaculum* **44**
A – Phylogénie du genre 44
B – Diversité des espèces du genre *Tenacibaculum* et *Average Nucleotide Identity* 47
C – Première publication : Étude comparée des génomes de *T. dicentrarchi*, « *T. finnmarkense* » et *T. piscium* (TNO020). 50
D – Synthèse et arbre phylogénétique complet 60

Partie 2 : Construction des spectres de référence MALDI-TOF **62**
A – Introduction 62
B – Matériel & Protocoles 63
C – Méthodes 65
D – Comparaison des approches d'identification 70

Partie 3 : Typage des isolats appartenant à l'espèce *T. maritimum* **73**
A – Introduction 73
B – Typage par classification non-supervisée des spectres MALDI-TOF 74
C – Typage par biomarqueurs, *Multi Peak Shift Typing* 79

IV – Discussion générale & Conclusion ***118***

A – Un travail collaboratif et interdisciplinaire **118**

B – L'épidémiologie des ténacibaculoses **119**

C – Le spectromètre de masse MALDI-TOF, outil de première ligne pour le diagnostic vétérinaire. 120

D – Conclusion 121

V – Annexes 123

Annexe 1 : Bootstrap classique (BP), bootstrap multi-échelle (AU) et package R *pvclust()*. 123

Annexe 2 : Illustration réalisée comme support du concours Doc'J MT180. Deuxième prix présentation orale 124

Annexe 3 : Affiche réalisée pour le congrès SFM 2019 ; 125

Cité des Sciences, Paris 125

IX – Bibliographie 126

Table des figures

Figure 1 : Production halieutiques et aquacoles mondiales (FAO 2018)..... 8
Figure 2 : Saumon atlantique, nécrose du pédoncule caudal provoquée par *Tenacibaculum dicentrarchi* (photo Carlos Sandoval)..... 9
Figure 3 : Colonie iridescente de *Tenacibaculum litopenai* sur gélose au sang (photo J.F. Bernardet)..... 12
Figure 4: Simplification des pipelines d'analyse utilisés dans le cadre de recherches académiques en santé publique pour effectuer de l'épidémiologie hospitalière..... 30
Figure 5: Principe de fonctionnement du MALDI-TOF (schéma de Mr.MORIDA)..... 34
Figure 6 : Représentation schématique d'une protéine présente plusieurs fois dans un spectre sous forme d'échos (mono-, di-, trichargée)..... 40
Figure 7: Phylogénie par maximum de vraisemblance des espèces appartenant au genre Tenacibaculum (raciné avec mid-point) ; FastTreeMP..... 46
Figure 8 : Phylogénie par maximum de vraisemblance des espèces appartenant au genre Tenacibaculum (non-raciné) ; FastTree..... 46
Figure 9 : "Genome clustering" de l'ensemble des génomes de *Tenacibaculum* intégrés dans MicroScope..... 49
Figure 10 : Arbre phylogénétique révisé représentant une proposition comprenant 4 clades au sein du genre..... 60
Figure 12: Arbre phylogénétique des 32 espèces du genre *Tenacibaculum* tenant compte des analyses d'ANI (raciné par l'outgroup *Pseudotenacibaculum*)..... 61
Figure 12: Schéma de création d'une référence "spectre entier"..... 65
Figure 13 : Schéma de l'identification d'un isolat par spectrométrie de masse MALDI-TOF. 66
Figure 14 : Evolution de la précision en fonction de la valeur seuil de décision..... 68
Figure 15 : Distribution des pourcentages d'identité obtenus sur l'ensemble du jeu de données..... 68
Figure 16: Schéma de création d'une référence basée sur des biomarqueurs ribosomiques..... 69
Figure 17: Schéma de classification non-supervisée des spectres MALDI-TOF..... 75
Figure 18 : Clustering hiérarchique, 2 clusters identifiés..... 76
Figure 19 : Clustering hiérarchique, 6 clusters identifiés (encadrés verts)..... 77
Figure 20 : Les 40 pics les plus discriminants selon une classification binaire..... 78
Figure 21: Schéma simplifié du typage d'un isolat par spectrométrie de masse MALDI-TOF..... 115
Figure 22: Schéma du flux d'analyse disponible sur l'application web MALDIquantTypeR 116

Liste des abréviations

1,5-DAN	1,5-diaminonapthalene
ADN	Acide désoxyribonucléique
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
ANIb	Average nucleotide identity BLAST
ANIm	Average nucleotide identity MUMmer
API	(galerie d'identification bactérienne commercialisée par bioMérieux) Analytical profile index
ARN	Acide ribonucléique
ARNr	Acide ribonucléique ribosomique
AST	Test de résistance aux antibiotiques
AT	Allele-type
cgMLST	Core-genome multilocus sequence typing
DDFA	Dépôt Direct avec Acide Formique
DDH	DNA-DNA hybridization
ESI	Electrospray ionization
GBDP	Genome Blast Distance Phylogeny approach
GGDC	Genome-to-Genome Distance Calculator
GWAS	Genome-wide association study
HCCA	acide α-Cyano-4-hydroxycinnamique
IIP	(équipe) Infection et Immunité des Poissons
IJSEM	International Journal of Systematic and Evolutionnary Microbiology
INRAE	Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnment (fusion INRA-IRSTEA)
INSERM	Institut National de la Santé et de la Recherche Médicale
MAD	Mean Absolute Deviation
MALDI-TOF	Matrix Assisted Laser Desorption Ionization - Time of Flight
MALDIrppa	MALDI Mass Spectrometry Data Robust Pre-Processing and Analysis
MBT-ASTRA	MALDI Biotype-Antibiotic Susceptibility Test Rapid
MBT-RESISMALDI	Biotyper-Resistance Test with Stable Isotopes
MBT-STAR-BL	MALDI Biotyper-Selective Testing of Antibiotic Resistance-Beta-lactamase
MIC	Concentration minimale inhibitrice
MICGC	MicroScope Genome Cluster
MLST	Multi Locus sequence typing
MLVA	Multiple Loci VNTR Analysis
MPST	Multi Peak Shift Typing
MS	mass spectrometer/mass spectrometry (spectromètre de masse/spectrométrie de masse)
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NCBI	National Center for Biotechnology Information
NGS	Séquençage de dernière génération
NJ	Neighbor-Joining
OMS	Organisation Mondiale de la Santé
OpenMP	Open Multi-Processing
PMF	Peptides Mass Fingerpint
PCR	Polymerase Chain Reaction
PFGE	Pulsed-Field Gel Electrophoresis
rMLST	Ribosomal MultiLocus Sequence Typing
SNIP	Statistics-sensitive Non-linear Iterative Peak-clipping algorithm
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
THAP	trihydroxyacetophenone
TIC	Total-Ion-Current
UPGMA	Unweighted Pair Group Method with Arithmetic mean
WGS	Whole Genome Sequencing

Avant-propos

J'ai toujours aimé la Biologie. Aussi loin que je me souvienne, je savais que j'étudierai cette discipline. Rien ne m'a jamais paru plus évident. Sans surprise, je me suis orienté vers une Terminale Scientifique SVT. Le premier choix que j'ai dû faire était : où réaliser mes études supérieures ? Je choisis d'intégrer l'Université Pierre et Marie Curie (UPMC). Bien que je fusse littéralement séduit par le monde universitaire, les Travaux Pratiques de biologie moléculaire furent une première déconvenue. Seuls ceux de Physiologie Végétale me plurent vraiment. Peut-être parce qu'ils ressemblaient (par rapport à mon imaginaire) à ce que devait être la Biologie : une biologie expérimentale. Par exemple, faire luire d'une lueur rouge éthérée une bouillie de chloroplastes de feuilles d'épinards sous une lampe à ultraviolet. Une vision nouvelle de la Biologie m'apparut bientôt en deuxième année de Licence. Le module d'Initiation à l'Abstraction en Biologie (introduction à la programmation) fut un véritable choc. Je venais alors de découvrir un nouvel univers : la Bioinformatique. Bien que la biologie évolutive me fascinât, j'ai décidé d'intégrer le Master de Bioinformatique et Modélisation (toujours à l'UPMC), plutôt que celui de Systématique et Évolution. Ce choix fut celui de ma passion pour le monde informatique. Durant de nombreux cours, nous entendions sans cesse combien la Recherche avait besoin de bioinformaticiens. Et maintenant, alors que je finis mon projet doctoral, un choix bien plus difficile se présente à moi.

L'expérience acquise durant ce projet fut considérable. J'ai pu suivre des formations enrichissantes dans des lieux incroyables. J'ai eu aussi de nombreuses occasions de présenter mes travaux. Ce fut une expérience enrichissante qui me permit également de présenter mon projet doctoral lors du premier congrès international sur les flavobactéries marines en 2018. J'ai ensuite réalisé le concours « Ma Thèse en 180 secondes » organisé par l'association Doc'J des doctorants et non-titulaires de Jouy-en-Josas (Annexe 3). Je suis fier du résultat obtenu et arriver à la deuxième place fut une grande joie. J'ai d'ailleurs rejoint l'équipe du bureau administratif de Doc'J quelques semaines plus tard, en tant webmaster (un des cadeaux de la bioinformatique, l'étiquette *geek* vous suit partout). Ce fut intéressant de voir les coulisses de la gestion d'une association. J'ai également présenté mes travaux aux journées de la branche française de l'European Association of Fish Pathologists en 2018 et aux Journées de la Recherche en Filière Piscicole en 2019. Durant cette dernière, je pense n'avoir jamais parlé devant autant de monde (près de 200 personnes et filmé par une caméra). J'ai aussi réalisé un poster dans le cadre du congrès de la Société Française de Microbiologie en 2019 (Annexe 4). J'ai également profité de ces 3 années pour m'essayer à l'enseignement.

En 2018 d'abord, j'ai encadré le stage de BioInformatique de Maroua Chadhil. Ce fut très instructif et également productif. Une partie de l'application web que j'ai développé (cf dernier chapitre des Résultats) est le fruit de son travail et je l'en remercie. J'ai continué sur cette lancée pour l'année universitaire 2018-2019, année pendant laquelle j'ai assuré des heures d'enseignement de Phylogénie et de Python. Bien que ce genre de mission ne soit pas toujours facile, j'ai apprécié d'avoir permis à quelques étudiants de comprendre les bases de programmation. Ce fut une grande source de satisfaction pour moi.

I – Introduction

A – Introduction générale

A1 – L'aquaculture, chiffres clés

Depuis une cinquantaine d'années, l'offre mondiale de poissons destinés à la consommation humaine a connu une croissance plus importante que celle de la population humaine. En effet, entre 1961 et 2013, elle a augmenté de 3,2% en moyenne par an, soit près du double de la croissance démographique. Cette progression s'accompagne d'une augmentation de la consommation apparente de poissons. Si elle ne représentait que 9,9 kg/habitant/an au début des années 60, celle-ci a dépassé les 20 kg/habitant/an en 2014. Les facteurs de ce changement sont multiples. Outre l'augmentation de la production, l'amélioration de la chaîne agroalimentaire a contribué à cet essor (réduction du gaspillage, amélioration des circuits de distribution). Elle a également largement été soutenue par l'évolution démographique et économique à l'échelle mondiale : croissance de la demande, accroissement de la population, augmentation des revenus et urbanisation.

La production issue de la pêche de capture est stagnante depuis la fin des années 1980. En revanche, celle issue de l'aquaculture a connu une croissance quasi ininterrompue. Une étape importante a été franchie en 2014, année durant laquelle la part de l'aquaculture à l'offre de poissons destinés à la consommation humaine a dépassé pour la première fois celle de la pêche (Figure 1). Alors qu'en 1974, cette part de l'aquaculture ne représentait que 7% de l'offre, elle a atteint 26% en 1994 puis 39% en 2004. La Chine est le premier producteur mondial, avec près de 60% de la production globale. La croissance est également soutenue par tous les autres pays producteurs, où la part de l'aquaculture a doublé. La consommation annuelle par habitant a donc nettement évolué dans les pays en développement (5,2 kg en 1961 à 18,8 kg en 2013) et dans les pays à faible revenu et à déficit vivrier (de 3,5 kg à 7,6 kg). Cependant, elle reste très inférieure à celles des régions les plus développées, même si cet écart tend à se réduire progressivement.

La progression de la consommation moyenne de poissons a conduit à une amélioration significative des régimes alimentaires partout à travers le monde, l'alimentation étant de ce fait plus variée et donc plus nutritive. En 2013, 17% des protéines animales proviennent de poissons pour l'ensemble de la population mondiale, représentant 6,7% des protéines consommées. Il est à noter également que pour plus de 3,1 milliards d'humains, le poisson représente ¼ de leur apport moyen en protéines animales. Le bénéfice nutritionnel de cet apport est indéniable. C'est une source riche en protéines de grande qualité, facile à digérer et contenant tous les acides aminés nécessaires. Le poisson est également une source d'acides gras essentiels, comme les omégas 3. Enfin, c'est un apport riche en vitamines (A, B et D) et en minéraux (calcium, iode, zinc, fer, sélénium). Grâce à ses nombreuses qualités, le poisson est une ressource précieuse, pouvant également permettre de lutter contre l'obésité (lorsqu'il est substitué à d'autre aliments).[1]

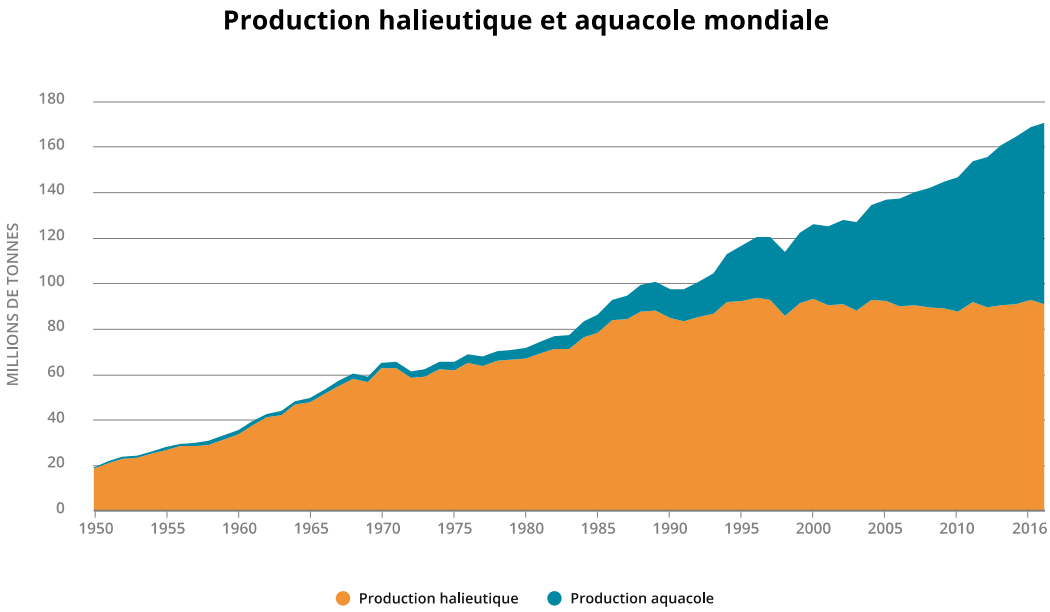


Figure 1 : Production halieutiques et aquacoles mondiales (FAO 2018)

A2 – Aquaculture : risques et défis

Face à une population mondiale toujours en expansion l'aquaculture apparaît alors comme une filière d'avenir. Cependant, et tout comme la *Green Revolution* des années 1960-1990, ce qui est désormais appelé la *Blue Revolution* doit faire face aux mêmes menaces que l'agriculture. L'aquaculture rencontre plusieurs défis majeurs. Tout d'abord, la compétition importante et toujours plus dure avec les autres utilisateurs de ressources (espace dédié, eau et aliments à destination des animaux) freine sa croissance. Face aux terribles enjeux climatiques auquel nous sommes confrontés, l'aquaculture connait d'autres enjeux directement liés à notre environnement et à l'impact des activités humaines sur celui-ci. La détérioration de la qualité des approvisionnements en eau résultant de la pollution aquatique a un impact déterminant sur les élevages. Enfin, la réponse de l'aquaculture à ce bouleversement climatique est claire : l'amélioration de la gestion environnementale est un défi inévitable pour la filière. La réduction des impacts et la prévention des risques pour la biodiversité est un des axes possibles d'un tel progrès, un autre étant l'intégration territoriale de cette aquaculture, notamment en créant de nouveau modèles d'exploitations, combinant l'aquaculture à d'autres secteurs agricoles, tels l'aquaponie. Si ces enjeux sont déjà considérables, la plus grosse menace pour un développement pérenne de l'aquaculture reste sans nul doute les pathologies, certaines connues depuis longtemps et d'autres nouvelles, véhiculées par l'apparition et la diversification des agents pathogènes découlant de l'intensification de l'aquaculture et des échanges internationaux d'œufs.[2]

Le développement d'une aquaculture de grande échelle a induit une augmentation dramatique d'épidémies graves causées par une diversité insoupçonnée d'agents pathogènes ; virus, bactéries et parasites [3–5]. Ces épisodes peuvent être tragiques pour la filière, tant au niveau économique qu'à l'échelle humaine. Le Chili a connu une période difficile en 2007, lorsque celle-ci fut frappée par une épidémie majeure d'anémie infectieuse du saumon. Cet événement fut un véritable séisme pour la filière chilienne avec une perte estimée à 1,8 milliards de dollars et près de 13000 emplois détruits [6]. Durant cet épisode, les régions rurales à faibles revenus ont été le plus lourdement touchées.

Les poissons ne sont pas les seuls animaux aquatiques pouvant être affectés par des maladies. Les crustacés et les mollusques le sont également. Par exemple, les crevettes sont sensibles à des maladies virales dévastatrices. Celles-ci sont le résultat d'une puissante pression de sélection qui est à l'œuvre dans les élevages. Certains virus furent disséminés par le transfert de larves et de géniteurs [7]. Les pathologies d'origine bactérienne peuvent être tout aussi délétères. Entre 2010 et 2014, la nécrose hepatopancréatique aigüe a causé environ 1 milliard de perte annuelle à l'industrie de l'élevage de crevettes [8]. L'impact de cette maladie est variable entre les pays, mais l'exemple de la Thaïlande est un des plus frappants. Elle a perdu près 30% de ses parts sur le marché mondial, n'atteignant plus que 10% en 2012. L'agent pathogène a été identifié comme une souche – ou un groupe de souches – appartenant à l'espèce *Vibrio parahaemolyticus* ayant acquis un plasmide contenant des facteurs de virulences (gènes codant pour des toxines) [9]. Durant l'intervalle de temps entre l'émergence de cette pathologie et l'identification du pathogène, la maladie s'est propagée dans la quasi-totalité des pays pratiquant l'élevage de crevettes [1] [8].

Une grande variété de bactéries (appartenant à plus de 25 genres) sont reconnues comme pathogènes, et ont été impliquées dans des maladies de poissons d'eau douce, d'eau saumâtre et d'eau de mer. Les flavobactéries sont l'une des premières causes de maladies dans les élevages. Ce terme désigne les bactéries appartenant à la famille des *Flavobacteriaceae,* phylum *Bacteroidetes*, réparties en environ 150 genres. Parmi ces derniers, le genre *Flavobacterium* comprend des bactéries responsables de pathologies chez les poissons d'eau douce, tandis que le genre *Tenacibaculum* comprend des espèces pathogènes de poissons marins. Les pathologies résultantes appelées ténacibaculoses regroupent un ensemble de maladies aux symptômes très proches et provoquées par des bactéries appartenant au genre *Tenacibaculum*. Celles-ci sont un problème impactant la santé et le bien-être des poissons de nombreuses espèces d'intérêt économique partout dans le monde [10, 11]. Les ténacibaculoses sont caractérisées par des lésions ulcératives, une érosion de la bouche, des nageoires effilochées et la nécrose du pédoncule caudal (exemple Figure 2). L'étude de l'ensemble de ces espèces constitue le cœur de mon projet doctoral.

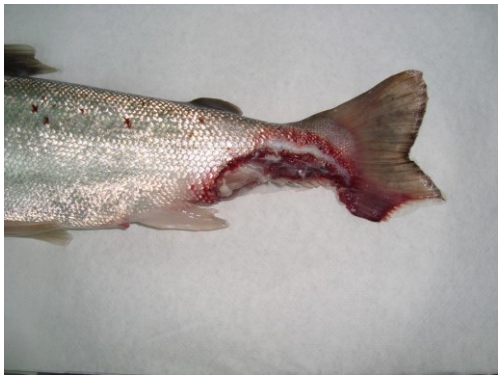


Figure 2 : Saumon atlantique, nécrose du pédoncule caudal provoquée par Tenacibaculum dicentrarchi (photo Carlos Sandoval)

B – Les flavobactéries

La famille des *Flavobacteriaceae* appartient au phylum *Bacteroidetes* (anciennement groupe *Cytophaga-Flavobacterium-Bacteroides*), à la classe des *Flavobacteriia* à l'ordre des *Flavobacteriales*.

Cette famille, la plus grande du phylum, contient au moins 150 genres ainsi que des centaines d'espèces. Les membres de cette famille sont retrouvés dans une grande variété d'habitats marins, d'eau douce et des sols ; et certains sont associés avec des animaux et des plantes.

Malgré cette diversité, certaines généralités peuvent être identifiées. La plupart des espèces sont aérobies, avec un métabolisme principalement respiratoire. Elles sont souvent retrouvées sur des algues, portées par des poissons ou des substrats organiques. Elles sont systématiquement chimioorganotrophes, et nombreuses sont celles qui peuvent digérer des polymères organiques complexes comme des protéines et des polysaccharides. Certaines espèces psychrophiles ou psychrotrophes sont isolées, par exemple, de l'eau de mer, de la glace de mer, des milieux terrestres froids comme les glaciers.

Les environnements typiques dans lesquels les membres pathogènes de la famille sont isolés varient selon les groupes taxonomiques. Ainsi, les espèces pathogènes de poissons sont communément retrouvées sur ou dans les poissons malades ou dans l'eau environnante. Les agents pathogènes des oiseaux sont associés à des foyers épidémiques de maladies chez les volailles domestiques ou les oiseaux sauvages. Les quelques flavobactéries pathogènes pour l'homme se retrouvent dans les prélèvements cliniques. [12]

C – Les bactéries du genre *Tenacibaculum*

Le genre *Tenacibaculum* (du latin *tenax*, tenace, et *baculum*, bâton) fait partie de la famille des *Flavobacteriaceae*. Le genre *Tenacibaculum* fut proposé en 2001 par Suzuki et al. [13] et soutenu par des analyses phylogénétiques basées sur les séquences de l'ARNr 16S et du gène codant pour GyrB. En effet, les auteurs constatèrent que 2 espèces alors classées dans le genre *Flexibacter* (*Flexibacter maritimus* et *Flexibacter ovolyticus*) étaient à la fois très proches de 2 nouvelles espèces isolées et en même temps très éloignées de l'espèce type du genre, *Flexibacter flexilis*. Ainsi, le genre *Tenacibaculum* nouvellement proposé comprenait à l'origine 4 espèces : *T. maritimum*, *T. ovolyticum*, *T. mesophilum* et *T. amylolyticum*. L'espèce-type du genre *Tenacibaculum* est *T. maritimum*. Depuis, il y a en moyenne une nouvelle espèce de *Tenacibaculum* décrite par an. Aujourd'hui, il y a 28 espèces décrites avec un nom valide [14]. Les bactéries appartenant au genre *Tenacibaculum* sont communément présentes en milieu marin. Elles se fixent ou s'associent habituellement à la surface d'organismes marins tels que les poissons, les macroalgues et les invertébrés.

C1 – Caractéristiques phénotypiques

Les cellules sont en forme de bâtonnets de 0,4 à 0,5 µm de diamètre et de 1,5 à 30 µm de longueur. Le pléomorphisme (fait de revêtir différentes formes sous certaines conditions) n'a pas été observé. Cependant des cellules sphériques, probablement mortes, peuvent être observées dans des cultures âgées. Les spores, les cellules en forme d'anneau et les vésicules de gaz ne sont pas observées. Les cellules ne possèdent pas de flagelles. Elles se déplacent par glissement. Ce sont des bactéries Gram-négatives. Les cellules produisent un pigment caroténoïde jaune qui est essentiellement de la zéaxanthine. Les pigments de type flexirubine (pigment majoritaire des bactéries des genres proches *Flexibacter*, *Flavobacterium*, *Chryseobacterium* et *Cytophaga)* sont absents. Les colonies présentent généralement une coloration jaune, jaune pâle ou blanchâtre. Certaines d'entre elles sont également iridescentes avec des reflets plutôt verts ou bleus (Figure 3). Ces bactéries sont aérobies et chémo-organotrophes. Elles sont également positives aux tests permettant de détecter la catalase et l'oxydase. Elles ne sont cultivables que dans des milieux contenant de l'eau de mer ou des sels marins synthétiques. La quinone respiratoire majoritaire est la ménaquinone 6. Les acides gras cellulaires majoritaires sont $C_{15:0}$, $C_{16:0}$ iso 3-OH et $C_{17:0}$ iso 3-OH. Le pourcentage de guanine + cytosine (GC%) est compris entre 30 et 35.2% [15].

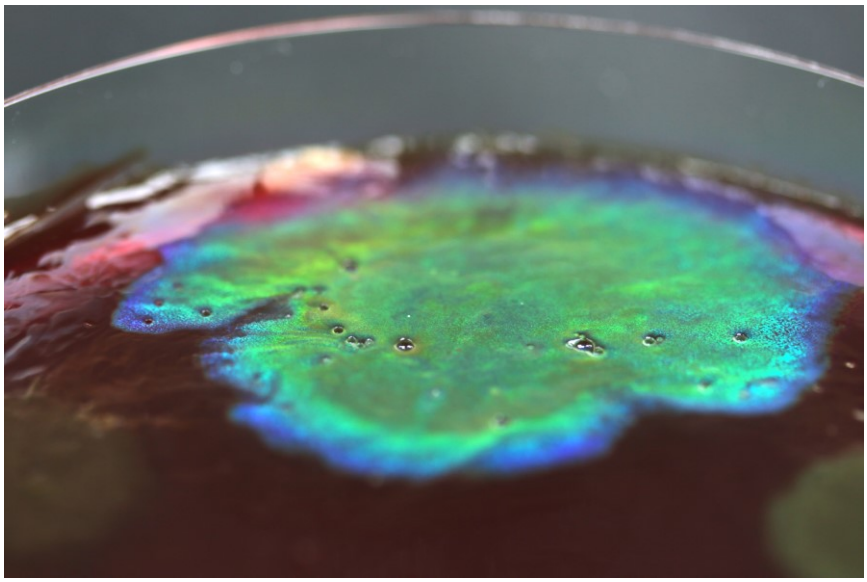


Figure 3 : Colonie iridescente de Tenacibaculum litopenai sur gélose au sang (photo J.F. Bernardet)

C2 – Les espèces associées à des épisodes infectieux chez des poissons d'élevage

Tenacibaculum maritimum fut la première espèce décrite du genre (sous le nom de *Flexibacter marinus* par Wakabayashi en 1977 [16]). Cette bactérie peut infecter un grand nombre d'espèces de poissons captifs ou d'élevage [11]. En France, elle provoque des pertes considérables dans les élevages de bar (*Dicentrachus labrax*), de daurade (*Sparus auratus*) et de turbot (*Scophthalmus maximus*). La maladie est encore peu étudiée, son apparition et son développement restent mal connus. Certaines études semblent indiquer l'importance des conditions d'élevage (densité des populations, qualité de l'eau) comme facteur possible de déclenchement des pathologies. Le réservoir (si tant est qu'il existe) de *T. maritimum* est inconnu. Certains auteurs soutiennent que cette bactérie serait un agent pathogène opportuniste causant principalement des lésions cutanées étendues et une abrasion des branchies puis, pour certaines souches plus virulentes, parfois une infection systémique. Nous verrons ici quelques exemples dans la littérature proposant cette hypothèse.

Un premier exemple est l'épisode de mortalité de saumons au Chili suite à un bloom d'algues vertes (*Pseudochattonella* spp.). Les auteurs identifièrent des pathologies sévères des branchies dans les premiers jours suivant le bloom d'algues qui révélèrent la présence d'isolats appartenant à l'espèce *T. maritimum* ainsi que *T. dicentrarchi* [17]. Un deuxième exemple est un épisode de mortalité de saumons atlantiques (*Salmo salar*) dans les îles Shetland en Ecosse. Les auteurs ont mis en évidence un lien entre un bloom de méduses (*Phialella quadrata*) et l'infection à *T. maritimum*. En effet, les premières lésions observées sur les branchies furent imputées à la présence de ces méduses microscopiques. Ensuite, ces lésions ont été colonisées par des bactéries appartenant à l'espèce *T. maritimum*. Bien que les auteurs n'aient pu déterminer la relation de causalité entre la présence des méduses et celle des bactéries dans les lésions des branchies, il semblerait qu'il existe un lien entre les deux [18]. Enfin, cette bactérie a également été isolée à partir de lésions blanchâtres observées près de la deuxième nageoire dorsale d'un requin taureau (*Carcharias taurus*) en captivité [19].

Tenacibaculum ovolyticum fut isolée pour la première fois en 1992 par Hansen et al. à partir d'épiflore d'œufs et de larves de flétan (*Hippoglossus hippoglossus*). La mortalité fut

importante dans les jours suivants l'éclosion. Les bactéries isolées avaient la capacité de dégrader à la fois le chorion et la *zona radiata* des œufs infectés [20]. L'infection expérimentale fut réalisée avec succès sur des œufs de flétan [21]. Cette espèce a également été retrouvée dans des lésions d'alevins de flétan [22] et dans l'épiflore de sardine (*Sardina pilchardus*) en Espagne [23].

Tenacibaculum soleae fut isolée pour la première fois en 2008 par Piñeiro *et al.* sur des soles sénégalaises (*Solea senegalensis*) malades en Espagne [24]. Dans la description de cette espèce, les auteurs mentionnent des tests d'infections expérimentales montrant que la souche type est virulente pour la sole et le turbot. Cette espèce fut ensuite retrouvée dans des lésions de céteau (*Dicologoglossa cuneata*) et de barbue (*Scophthalmus rhombus*) d'élevage [25]. Elle a été également retrouvée en Norvège dans des lésions caudales de labres (*Labridae sp.*) [22]. Étonnamment, elle fut également retrouvée lors d'un épisode inhabituel de mortalité d'huîtres (*Crassostrae gigas*) en Italie en 2018. Dans cette même étude, une infection expérimentale par injection dans le muscle adducteur d'huîtres induisit une mortalité significativement plus importante que dans le groupe contrôle non infecté [26].

Tenacibaculum discolor fut également découverte pour la première fois en 2008 par Piñeiro et al. à partir de reins de soles malades (*S. senegalensis)* [27]. Elle a été ensuite retrouvée durant l'année 2014 en Italie dans des reins et des lésions de la peau et des yeux de bar (*Dicentrarchus labrax)* [28]. En Italie, elle a été également retrouvée sur du bar, de la sole, du turbot et de la dorade (*Sparus aurata)*. Récemment, elle a été isolée de marais salants fertilisés avec des boues contaminées au mercure. C'était le principal taxon bactérien isolé de ce milieu. Un opéron composé de 4 gènes spécifiques de la dégradation du mercure a identifié à partir du génome de l'isolat 9A5 [29].

Tenacibaculum gallaicum fut isolée en même temps que *T. discolor* par Piñeiro *et al.* à partir de l'eau provenant d'un bassin d'élevage de turbots [27]. L'infection expérimentale fut réalisée par inoculation intrapéritonéale d'une dose unique de bactéries avec succès en 2007. L'injection de souches de *T. gallaicum* chez le turbot et la sole entraîna une mortalité de 60 à 100% [30, 31].

Tenacibaculum dicentrarchi fut d'abord isolée de bars malades en Espagne en 2012 [32]. Quatre ans plus tard, une étude fut réalisée sur la période 2010 à 2014 pendant laquelle des mortalités furent observées. Le foyer d'octobre 2010 frappa une population de 1200 individus et la mortalité atteignit alors 50-60%. Les auteurs réalisèrent une infection expérimentale par balnéation en 2016 sur le saumon atlantique et la truite arc-en-ciel (*Oncorynchus mykiss*) durant laquelle les mortalités furent respectivement de 63% et de 95%. Cependant, aucune mortalité ne fut observée pour le saumon coho [33]. Depuis, elle a été isolée d'anguilles rouges chiliennes (*Genypterus chilensis*) qui présentaient des hémorragies au niveau de la bouche et de l'opercule, des nécroses de la queue et des ulcères abdominaux. Les auteurs n'ont pas pu déterminer si les mortalités d'anguilles observées étaient dues à l'infection bactérienne. Cependant, le caractère émergent de cette bactérie doit motiver des recherches plus approfondies [34]. En 2019, *T. dicentrarchi* fut identifié comme étant l'agent pathogène à l'origine d'un foyer aigu de ténacibaculose dans un site post-smolt (saumoneau). Il a été montré que cette bactérie adhère et forme des biofilms sur des surfaces en polystyrène. Cette capacité a conduit les auteurs à penser que *T. dicentrarchi* fait partie des risques émergents majeurs pour l'aquaculture [35].

"*Tenacibaculum finnmarkense*" (les noms latins entre guillemets n'ont pas été publiés de manière valide et sont donc dépourvus de statut dans la nomenclature bactérienne officielle) fut décrite pour la première fois en 2016 à partir des lésions de peau de saumon atlantique provenant de Norvège [36]. L'infection expérimentale par balnéation sur des saumons norvégiens fut un succès [37]. En 2017, une étude a montré la co-occurrence d'un bloom de méduses (*Dipleurosoma typicum*) avec un foyer d'infection à "*T. finnmarkense*" dans un élevage de saumon mais, les auteurs n'ont pas pu confirmer que la méduse était réellement le vecteur de la bactérie. Néanmoins, les dommages causés par les nématocytes de ces méduses seraient suffisants pour favoriser l'infection par une bactérie opportuniste [38]. Cette bactérie a également été identifiée à partir de lésion de saumon au Chili [39].

C3 – Les autres espèces associées à des pathologies

Tenacibaculum litopenaei fut formellement décrite en 2007. A partir de l'eau d'un bassin d'élevage de crevettes (*Litopenaeus vannamei*) à Taiwan [40]. Il est probable que cette bactérie ait en fait été isolée pour la première fois à partir d'oursins (*Strongylocentrotus intermedius*) présentant des lésions décolorées en 1977 [41]. En effet, dans la description de *T. litopenaei,* la phylogénie obtenue à partir de l'ARNr 16S indique que le "*Flexibacter echinicida*" isolé des oursins est extrêmement proche *T. litopenaei*, suggérant qu'il s'agit d'une seule et même espèce. L'infection expérimentale d'oursins par balnéation fut un succès [42]. En revanche, aucune infection expérimentale n'a jamais été effectuée (ou du moins publiée) à partir de souches clairement identifiées comme appartenant à l'espèce *T. litopenaei.*

C4 – Les espèces isolées de l'environnement vraisemblablement non pathogènes

Les espèces suivantes ont été isolées de l'environnement et ne sont vraisemblablement pas pathogènes. Pour la plupart, un seul représentant de l'espèce a été isolé. Les informations sont résumées dans le tableau ci-dessous. Les références complètes sont dans l'ordre chronologique : [13], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62].

Espèce	Origine géographique	Echantillon	Référence (PMID)	Date de publication
Tenacibaculum amylolyticum	îles Palaos, Philippines	algue verte, *Avrainvillea riukiuensis*	11594591	2001
Tenacibaculum mesophilum	Japon	éponge, *Halichondria okadai*	11594591	2001
Tenacibaculum skagerrakense	Skagerrak, Danemark	colonne d'eau (30 m de profondeur)	15023969	2004
Tenacibaculum lutimaris	Mer Jaune, Corée du Sud	vasière de marée	15774664	2005
Tenacibaculum aestuarii	Saemankum, Corée du Sud	vasière de marée	16825632	2006
Tenacibaculum litoreum	Ganghwa, Corée du Sud	vasière de marée	16514041	2006
Tenacibaculum aiptasiae	Taiwan	anémone de mer, *Aiptasia pulchella*	18398166	2008
Tenacibaculum adriaticum	Croatie	bryozoaire, *Schizobrachiella sanguinea*	18319452	2008
Tenacibaculum crassostreae	Corée du Sud	huître du Pacifique, *Crassostrea gigas*	19542127	2009
Tenacibaculum jejuense	île de Jeju, Corée du sud	eau de mer	21460140	2012
Tenacibaculum geojense	Corée du Sud	eau de mer	21257684	2012
"*Tenacibaculum halocynthiae*"	Mer du Sud, Corée du Sud	ascidie marine, *Halocynthia roretzi*	23543245	2013
Tenacibaculum caenipelagi	Corée du Sud	vasière de marée	23733002	2013
Tenacibaculum xiamenense	Xiamen, RP Chine	eau de mer (1 à 2 m de profondeur)	23543502	2013
Tenacibaculum holothurium	Xiapu, RP Chine	holothurie, *Apostichopus japonicus*	26345588	2015
Tenacibaculum ascidiaceicola	Mer Jaune, Corée du Sud	ascidie, *Halocynthia aurantium*	26674528	2016
Tenacibaculum sediminilitoris	Mer Jaune, Corée du Sud	vasière de marée	27089227	2016
Tenacibaculum haliotis	Corée du Sud	ormeau, *Haliotis discus hannai*	28829017	2017
Tenacibaculum aestuariivivum	île de Jindo, Corée du Sud	vasière de marée	28984542	2017
Tenacibaculum agarivorans	Weihai, RP Chine	algue, *Porphhyra yezoensis*	29043952	2017
Tenacibaculum insulae	île de Jindo, Corée du Sud	vasière de marée	29148365	2017
Tenacibaculum todarodis	Mer Jaune, Corée du Sud	calamar, *Todarodes pacificus*	29521615	2018

D – Taxonomie et espèce bactérienne

D1 – Taxonomie bactérienne

La science de la classification biologique est la taxonomie (du grec *taxis*, disposition ou ordre et *nomos*, la loi ou *nemein*, distribuer ou gouverner). La taxonomie ou taxonomie, au sens large, est divisée en 3 ensembles distincts : la classification, la nomenclature et l'identification. Le terme systématique est plus général et désigne l'étude scientifique des organismes dans le but de les caractériser et de les arranger de manière ordonnée.

La classification consiste à répartir les organismes en groupes appelés taxons, selon leur similarité mutuelle. Les systèmes de classification évoluent au fil du temps, en fonction des époques et des technologies disponibles. Ainsi, un des premiers systèmes de classification fut la classification naturelle proposée par le botaniste Linné. Celui-ci était basé sur des caractères anatomiques. L'avantage de cette classification par rapport aux précédentes (qui étaient artificielles) était que la position d'un organisme dans le schéma global donnait une information sur ses propriétés intrinsèques. En microbiologie, dans les premières éditions du *Bergey's Manual of Systematic Microbiology*, la classification utilisée était strictement phénotypique, reposant sur la comparaison de caractères tels que la morphologie des cellules et des colonies, leur physiologie (préférences de milieux et de température), la composition chimique de leur membrane ou encore leurs réactions biochimiques. Les auteurs des descriptions des nouvelles espèces ont ensuite rapidement intégré la caractérisation phylogénétique dans une approche dite polyphasique. La classification phylogénétique consiste à reconstruire les liens de parenté entre les organismes sur la base de caractères. Historiquement, ces caractères étaient morphologiques (cladistique morphologique). Aujourd'hui, la phylogénie est moléculaire, c'est-à-dire basée sur les séquences de macromolécules biologiques pour calculer des distances évolutives entre les organismes. Les phylogénies basées sur les ARN ribosomiques 16S et 23S (chez les procaryotes) ou 18S et 28S (chez les eucaryotes) sont les exemples les plus courants de phylogénies moléculaires, encore utilisées aujourd'hui. Enfin, avec l'arrivée du séquençage haut-débit, la classification phylogénomique se développe. Elle consiste à comparer la similarité génomique des organismes. Elle se base sur la séquence d'un ensemble de gènes donné ou sur l'intégralité des génomes. Aujourd'hui, cette branche de la taxonomie aboutit à une nouvelle possibilité de définition d'une espèce bactérienne sur la seule base de la similarité entre plusieurs génomes bactériens. Le terme correspondant à cette définition est la génoespèce.

La nomenclature est la branche de la taxonomie consistant à nommer les groupes taxonomiques selon des règles universelles. Au contraire de la classification, la nomenclature est stable et officielle. Les noms d'espèces sont immuables : l'épithète la plus ancienne a la préséance et doit être utilisée. En revanche, un organisme peut être transféré dans un autre genre ou un nouveau si des informations nouvelles sont apportées, par exemple par l'information phylogénomique.

L'identification est le côté pratique de la taxonomie. Elle consiste à déterminer si un isolat particulier appartient à un taxon connu.

D2 – Qu'est-ce qu'une espèce bactérienne ?

La notion d'espèce bactérienne n'est pas aussi évidente à définir que pour les mammifères. Pour des raisons pratiques, il convient de suivre les règles établies par *l'International Committee on the Systematics of Prokaryotes* et le *Bacteriological Code* afin de décrire et nommer les espèces bactériennes. Le système de classification repose aujourd'hui sur une hiérarchie taxonomique. Il existe différents rangs taxonomiques qui sont dans l'ordre : Domaine, Règne, Phylum, Classe, Ordre, Famille, Genre et Espèce. L'unité taxonomique de base est l'espèce.

La définition la plus fondamentale d'une espèce de bactérie est : l'ensemble de souches qui partagent de nombreuses propriétés stables et différant de manière significative des autres groupes de souches. La souche type est la souche ayant permis de décrire l'espèce. La difficulté de la définition de l'espèce bactérienne réside dans sa souplesse. Il existe en effet différentes manières d'évaluer la ressemblance entre deux souches. Celle-ci peut être évaluée selon la morphologie, la composition chimique, les fonctions, le métabolisme ou encore le génome.

Enfin, au sein d'une espèce, il existe plusieurs manières de grouper (= typer) les souches. Les biovars (ou biotypes) sont des groupes de souches caractérisées par des différences biochimiques ou physiologiques. Les morphovars (ou morphotypes) se différencient par leur morphologie et les sérovars (ou sérotypes) ont des propriétés antigéniques distinctives. Les génomovars (ou génotypes) sont des souches qui se distinguent uniquement sur une base génomique (sans différences phénotypiques).

A l'heure actuelle, la description d'une nouvelle espèce bactérienne est cadrée par un protocole strict. Les microbiologistes parlent d'approche polyphasique car c'est l'ensemble des caractéristiques morphologiques, biochimiques et génétiques qui sont prises en compte pour toute nouvelle description d'espèce. Ces notions de base sont décrites dans le live Microbiologie de Prescott et al. paru en 2013.

D3 – Approche polyphasique de la description d'une espèce

a. *Caractérisation phénotypique*

a1. Caractérisation morphologique

Cette caractérisation est importante car ce sont les caractéristiques les plus faciles à observer. Historiquement, elle a permis de proposer une classification des bactéries basée sur la forme des cellules (e.g cocci pour des cellules rondes, bacille pour des cellules en formes de bâtonnets, etc…) et des colonies. Le résultat des tests de coloration, comme la coloration Gram sont toujours utilisés dans les descriptions d'espèces. Le type de mobilité est aussi évalué. Ces premières observations permettent généralement d'orienter l'identification vers un groupe de bactéries particulier.

a2. Caractérisation biochimique et physiologique.

Ces traits sont importants car ils sont directement en relation avec la nature des enzymes microbiennes et des protéines de transport. Ainsi, pour décrire une espèce, les microbiologistes évaluent sa capacité à métaboliser certains composés. Ses tests sont encore utilisés aujourd'hui. Les tests classiques sont par exemple : la relation à l'oxygène, le type de

métabolisme, les exigences de températures, de pH, de salinité, la dégradation de nombreux substrats (amidon, gélatine, etc), ou la capacité de réduction des nitrates.

Les limites de cette approche sont dues à la plasticité génomique. Pour un génome donné, il existe différents phénotypes. Ainsi, les conditions de culture d'une bactérie influent sur sa capacité à métaboliser les molécules présentes dans le milieu de culture. Il apparaît que les résultats de ces tests peuvent être difficiles à interpréter et pas toujours reproductibles. Néanmoins, il existe des tests standardisés et miniaturisés. Les systèmes les plus connus sont les galeries API (bioMérieux) mais il en existe bien d'autres.

a3. Caractérisation chimique : chimiotaxonomie

Généralement, les descriptions d'espèces intègrent une analyse détaillée de la composition chimique de la bactérie. Elles sont souvent focalisées sur l'étude des lipides et plus particulièrement la caractérisation des acides gras et les lipides polaires de la paroi bactérienne, ainsi des quinones respiratoires et des pigments..

b. Caractérisation moléculaire

La génétique et la possibilité de comparer l'ADN de différents isolats bactériens a été un changement majeur pour la microbiologie. L'analyse de l'ADN bactérien a permis de développer des nouvelles métriques permettant de mesurer la similarité entre deux isolats. Du séquençage du gène de l'ARNr 16S au séquençage de génomes entiers, l'idée de génoespèce émerge progressivement. L'appartenance de deux isolats à la même espèce peut être évaluée par le degré de similitude de leur génome. Dans les paragraphes suivants, nous verrons les méthodes classiques d'identification utilisant la molécule d'ADN.

b1. Du gène codant pour l'ARN ribosomique 16S (ARNr 16S)

Dans les années 1960, Dubnau et al. [63] ont observé que les séquences du gène codant pour l'ARNr 16S chez *Bacillus* spp. étaient conservées entre elles. L'utilisation généralisée de cette séquence génétique pour l'identification bactérienne a suivi un ensemble de travaux pionniers de Woese, qui en a défini les propriétés importantes. Tout d'abord, ce gène semble se comporter comme une horloge moléculaire, comme le souligne l'excellent article de synthèse de Woese [64]. Le haut degré de conservation du gène de l'ARNr 16S est supposé résulter de l'importance critique de cette molécule dans les fonctions cellulaires. En effet, le ribosome est la machinerie de synthèse de toutes les protéines constitutives de la cellule. Bien que le taux absolu de mutations de cette séquence ne soit pas connu, il marque la distance évolutive des organismes. La séquence de ce gène fait environ 1550 paires de bases (pb) et est composée de régions conservées et de régions variables. Ce gène est donc assez long, avec suffisamment de polymorphisme pour fournir une mesure statistiquement pertinente et discriminante. Les amorces universelles généralement choisies sont complémentaires des régions conservées situées au début du gène et, soit dans la région 540 pb, soit à la fin de la séquence du gène entier (région 1550 pb), la séquence variable intermédiaire étant utilisée pour la phylogénie moléculaire [65, 66]. Bien que 500 et 1500 pb soient les longueurs communément utilisées durant le séquençage, les bases de données contiennent des séquences dont la taille peut être différente. Enfin le gène ARNr 16S est universel pour les bactéries. Ainsi, la distance évolutive peut être mesurée objectivement entre toutes les bactéries [64, 67]. En général, la comparaison de séquences d'ARNr 16S permet la distinction des organismes à l'échelle du genre (pour la plupart des *phyla* bactériens) et parfois même à l'échelle de l'espèce. Les seuils de détermination de l'appartenance de deux isolats au même genre et à la même espèce sont respectivement de 95% et de 98.65% [68]. Cette méthode est devenue une méthode standard pour la description de toute nouvelle espèce

bactérienne. Toutefois, il existe quelques exceptions dans lesquelles cette approche montre ses limites. Par exemple, il peut arriver que des espèces différentes mais très proches possèdent des séquences de ce gène identiques ou presque. [69]

b2. Par hybridation ADN-ADN.

L'hybridation ADN-ADN est toujours employée dans le cadre des descriptions formelles de nouvelles espèces bactériennes, publiées dans l'*International Journal of Systematic and Evolutionnary Microbiology* (IJSEM), journal de référence pour la taxonomie bactérienne [70]. La méthode est basée sur la capacité de fragments d'ADN nucléiques monocaténaires à former des hybrides (duplexs) par appariement de bases homologues. Habituellement, l'ADN des souches à étudier est chargé sur une membrane de nitrocellulose ou de nylon, dénaturé par la chaleur puis mis en présence d'une solution d'ADN génomique monocaténaire marqué d'une souche de référence. Dans des conditions bien définies, l'étendue de formation du duplex, c'est à dire le pourcentage d'hybridation, peut servir de mesure d'homologie entre les séquences d'ADN des deux échantillons. Un certain nombre de facteurs peuvent affecter les taux d'hybridation. Un des facteurs majeurs est la température. En effet, la formation d'hybride double brin dépend directement de la température. Il est généralement recommandé de réaliser la réaction à 20-25°C sous la température de demie dénaturation (T_m). D'autres facteurs comme la concentration en sonde ADN simple brin ou le G+C% de l'ADN peuvent également influer la réaction. [71]. Le seuil recommandé pour la délimitation de deux espèces est de 70 % d'hybridation DNA-DNA (DDH).

b3. A l'ère de la génomique : *Average Nucleotide Identity, Genome sequence-based species delimitation*.

L'Average Nucleotide Identity (ANI) peut être vue comme l'équivalent *in silico* de l'hybridation ADN-ADN. Cette approche permet de vérifier l'appartenance de deux souches à la même espèce à l'aide d'une mesure du pourcentage d'identité de leurs génomes. Il existe différentes méthodes de mise en œuvre de cette méthode (BLAST, MUMMER), celle basée sur BLAST est détaillée ici. Les génomes à analyser sont découpés en fragments d'une longueur donnée (généralement 1000 bp), puis alignés les uns contre les autres. Une fois les meilleurs alignements trouvés pour chaque fragment, le pourcentage d'identité nucléique est calculé. La mesure finale correspond au pourcentage moyen d'identité sur l'ensemble des alignements. L'ANI permet de définir des unités taxonomiques (clusters de génomes regroupés selon leur similarité). En général, la valeur seuil de définition de cette unité taxonomique se situe entre 94 et 96%, selon les auteurs. Ces unités taxonomiques sont définies de telle sorte qu'elles correspondent généralement à la définition d'espèce bactérienne [72]. Bien sûr, la couverture d'alignement est également un des paramètres importants pour conclure à l'appartenance de deux génomes à la même espèce. L'ANI (et les méthodes assimilées, comme le *Genome Clustering* qui sera détaillé dans la première partie des résultats) devrait devenir une méthode standard en microbiologie. Elle serait un atout majeur notamment pour soutenir l'approche polyphasique de définition de nouvelles espèces et se substituer à l'hybridation ADN-ADN. Elle devrait également être un outil pertinent pour la gestion des bases de données génomiques en corrigeant les erreurs d'affiliation taxonomique. La collection allemande de microorganismes et de cultures cellulaires (abrégée DSMZ en allemand) a développé des approches similaires appelées *Genome-to-genome-distance-calculator* (GGDC) et la *Genome Blast Distance Phylogeny approach* (GBDP) qui permettent de calculer une probabilité que deux génomes appartiennent à la même espèce.

E – Identification et typage bactérien

E1 – Introduction

Depuis la démonstration par Koch, en 1876, que *Bacillus anthracis* est l'agent étiologique de la maladie du charbon, l'identification des microbes est devenue la priorité de la microbiologie clinique. Cependant, l'identification est insuffisante pour répondre à des questions à plus grande échelle comme comprendre les mécanismes de transmission des agents pathogènes ou encore la présence de réservoirs et l'origine de ces agents. Ces questions relèvent de l'épidémiologie, dont la base est la caractérisation fine (typage) des isolats. Le typage sert à déterminer la probabilité que deux isolats partagent une origine commune. En médecine humaine, les infections nosocomiales sont un problème de santé mondiale. L'Organisation Mondiale de la Santé estimait à 1,4 millions le nombre de patients infectés (source : site OMS) et le Ministère de la Santé enregistre environ 4000 décès par an en France pour 750 000 infections nosocomiales (source : site INSERM). Les recherches concernant ce problème doivent permettre d'identifier l'origine hospitalière ou non d'une infection. Les maladies d'origine alimentaire ou environnementale comme la listériose ou la légionellose sont aussi des problèmes majeurs en santé publique. Pour ces maladies, les différents réseaux de surveillance cherchent à identifier par des études épidémiologiques la source de la contamination afin de l'éliminer. En médecine vétérinaire, la chaîne de production alimentaire est fragile. Par exemple, en pisciculture, les écloseries (géniteurs, œufs et alevins), les zones de pré-grossissement (poissons d'environ 1 à 20 gr) et celles de grossissement sont séparées. Chaque compartiment est un maillon de cette chaîne. Dans un contexte d'épisodes infectieux ponctuels ou chroniques, le pisciculteur doit pouvoir identifier le maillon faible de la chaîne dans lequel l'infection se produit. En identifiant la période du cycle de vie du poisson et l'origine de la contamination, le pisciculteur pourrait mettre en place des mesures préventives afin de supprimer la source de la contamination. En outre, la démonstration de l'origine de l'infection a et aura de plus en plus d'implications légales, et le typage des bactéries incriminées est un des éléments judiciaires. La crainte du bioterrorisme pousse également à développer des outils performants pour identifier les germes présents dans l'environnement, pour retrouver l'origine des souches et comprendre les éventuels trafics de ces armes biologiques [73]. Pour identifier et caractériser les bactéries, il est encore souvent nécessaire d'isoler l'agent responsable en culture pure. L'élaboration de milieux de culture sélectifs reste donc une étape limitante de ces approches. Dans le cadre d'une collaboration précédente entre la société Labofarm et l'équipe Immunité et Infection des Poissons de l'INRA (projet FUI PathoTrackFish), un milieu sélectif a été développé pour les bactéries appartenant au genre *Tenacibaculum*. Le but de ce chapitre est de rendre compte, par un survol non exhaustif, de l'évolution des techniques d'identification et de typage. Cela nous permettra de comprendre les avantages et les inconvénients de ces différentes approches avant de nous intéresser à la technologie retenue dans le cadre de mon projet, la spectrométrie de masse MALDI-TOF.

E2 – Outils d'identification

Le problème de l'identification est une question commune à plusieurs disciplines : taxonomie, médecine et épidémiologie. Selon la finalité de l'identification, les contraintes temporelles et économiques ne sont pas les mêmes. Les outils d'identification sont identiques à ceux développés et utilisés en taxonomie classique. Néanmoins, les applications de « terrain » (microbiologie clinique pour le diagnostic et épidémiologie clinique) privilégient certains outils. Par exemple, l'observation à l'œil d'une colonie sur une boîte de Petri ne fournit qu'une identification présomptive (piste d'identification). Les galeries API permettent de faire une identification basée sur des réactions biochimiques. Aujourd'hui, cette approche permet d'identifier théoriquement près de 700 espèces à l'aide de près de 1000 réactions biochimiques [74]. L'identification plus précise peut, dans certains groupes bactériens, être réalisée à l'aide de tests sérologiques. Celle-ci est hautement spécifique mais constitue paradoxalement son point faible. Cette approche nécessite en effet d'avoir à disposition tous les antisérums propres à l'espèce pour identifier l'échantillon. L'identification par PCR est souvent la solution retenue (sous réserve qu'une méthode de typage pour l'espèce d'intérêt ait été élaborée). Ainsi, les laboratoires d'aujourd'hui recherchent des systèmes automatisés fournissant une identification rapide, bon marché et fiable, fournissant des résultats simples à interpréter et reproductibles. Un de ces systèmes, dont l'utilisation se généralise en médecine humaine et, dans une moindre mesure, en médecine vétérinaire, est la spectrométrie de masse MALDI-TOF. Cette technologie, fera l'objet d'un chapitre à part (Introduction, partie F).

E3 – Outils de typage

Les espèces d'importance clinique sont divisées en groupes caractérisés par une virulence variable ou une capacité de dissémination spécifique. Le typage bactérien a pour but de reconnaître des clones épidémiques et les distinguer des cas sporadiques, des clones ayant une virulence accrue ou spécifique, de comparer les isolats cliniques avec les isolats de l'environnement, suivre la propagation d'une maladie (parfois identifier un foyer infectieux, le porteur 0) afin d'anticiper son évolution, préciser la structure de la population de l'espèce responsable [75]. Le typage bactérien est utilisé aussi bien par les laboratoires cliniques où il est nécessaire d'obtenir des résultats rapides afin de pouvoir lutter contre la propagation de clones épidémiques en mettant en œuvre des mesures de contrôle des infections que par des laboratoires de recherche lors d'études épidémiologiques rétrospectives [76].

Au début du XX^{ième} siècle, l'identification bactérienne était rarement étendue au-delà de l'espèce ou la sous-espèce. Dans les cas où des identifications plus précises étaient nécessaires, le type des isolats bactériens était classiquement déterminé par la sérologie ou bien par des différences phénotypiques mesurables tels que les caractéristiques de croissance (cinétiques et substrats nécessaires) ou encore des différences de sensibilité aux antibiotiques, de résistance aux phages ou de fonctions métaboliques [77].

A partir de la fin des années 1970, l'élaboration des techniques de laboratoires pour l'analyse des protéines et des séquences d'acides nucléiques a imposé un changement de paradigme en santé publique. De nouveaux outils ont été développés pour l'identification et la caractérisation moléculaire des bactéries. Ces outils ont également permis d'améliorer la compréhension de l'écologie des maladies infectieuses, la diversité bactérienne, les dynamiques de populations, la surveillance de l'émergence de nouvelles souches ou pathotypes cliniquement important par les laboratoires de microbiologie clinique.

Durant les décennies suivantes, les progrès continus et rapides des protocoles et des techniques moléculaires ont révolutionné la pratique de la microbiologie en santé publique. Ils ont ainsi fondamentalement changé la nature, la précision et l'opportunité des données de laboratoire pour l'analyse et la gestion des épidémies.

Aujourd'hui, les méthodes de laboratoire pour l'épidémiologie moléculaire sont à un tournant majeur, avec des méthodes de transition entre l'analyse de fragments d'ADN sur gel, les techniques dérivées du séquençage haut-débit et les autres technologies de laboratoire haut-débit qui sont de plus en plus automatisées. L'utilisation de ces méthodes facilite par conséquent le typage, la caractérisation et l'analyse comparée en temps quasi-réel et à haute résolution d'isolats d'espèces pathogènes.

Cependant, l'énorme quantité de données générées par ces approches nécessite des investissements parallèles forts dans le calcul haute-performance, le stockage pérenne des données, le partage de bases de données curées et du personnel dédié, compétent et bien formé. Cette évolution technologique critique doit également intégrer des considérations particulières concernant la manière dont ces données seront organisées, intégrées et partagées. [78]

Dans cette partie, nous verrons d'abord les deux méthodes emblématiques de la caractérisation phénotypique : le sérotypage et les tests de résistance aux antibiotiques. Ensuite, nous verrons quelques techniques moléculaires de typages comme la PFGE ou le ribotypage. Enfin, nous verrons des approches basées sur le séquençage d'un ensemble restreint de gènes (par exemple la MLST) ou bien de génomes entiers.

a. Caractérisation phénotypique

a1. Par réactions sérologiques, les sérotypes.

La méthode conventionnelle de sérotypage est basée sur des réactions d'agglutinations avec un antisérum (obtenus après inoculation de l'agent pathogène à des lapins) contre chacun des épitopes [79]. Selon les groupes bactériens étudiés, le sérotypage vise soit l'antigène O du lipopolysaccharides des bactéries à Gram négatif (antigène de la membrane externe), soit l'antigène H (antigène flagellaire) ou bien l'antigène K (antigène capsulaire).

Cette technique a été intensément utilisée durant les cinquante dernières années. Elle a été normalisée par les centres de référence coordonnés par l'Organisation Mondiale de la Santé. Une des principales limitations de cette approche est le maintien des stocks de sérum (par exemple plus de 2200 antisérums différents pour caractériser les nombreux sérotypes de *Salmonella*) [80]. L'application de cette approche dans les laboratoires se heurte également au problème de réactions croisées d'un isolat à plusieurs antisérums. La préparation des antisérums sur lapins est également contraignante (à la fois en temps et en coût) et nécessite des protocoles et des contrôles qualités exigeants. Il est maintenant parfois possible de sérotyper des souches en utilisant directement les génomes complets. [81]

a2. Par sensibilité aux antibiotiques, les antibiogrammes.

Face aux bactéries résistantes, les tests de résistance aux antibiotiques (AST) sont une étape clef permettant de choisir la molécule la plus adaptée pour contrôler l'agent pathogène responsable de l'infection [82]. Les outils actuels d'AST reposent sur des techniques chronophages de culture bactérienne, suivis d'une diffusion à partir de disques chargés de différents antibiotiques [83] et/ou d'un test de susceptibilité dans un milieu gélosé contenant des concentrations variables d'antibiotique [84]. Ce processus prend souvent plusieurs jours avant que les valeurs de concentration minimale inhibitrice (MIC), mesure définissant la plus faible concentration d'antibiotique nécessaire pour prévenir la croissance des bactéries et utilisée pour déterminer si l'agent pathogène isolé est sensible ou résistant à un antibiotique, soient obtenues. [85]

b. Caractérisation moléculaire

b1. Pulsed-field gel electrophoresis (PFGE).

L'électrophorèse sur gel d'agarose est très utile pour visualiser le contenu en ADN des cellules bactériennes [86]. Elle a été une révolution en biologie moléculaire dans les années 1990. Elle trouvé une application clinique, y compris en épidémiologie moléculaire. Le principe de cette méthode repose sur le fait que les molécules d'ADN (y compris de tailles de l'ordre du Mb) peuvent être séparées par une réorientation de la migration de l'ADN en fonction de la taille des fragments d'ADN et le temps de migration, obtenue par des pulsations périodiques du champ électrique dans différentes directions. En comparant le profil de migration (pulsotypes) de souches de référence avec celui de l'échantillon à tester, cette méthode permet de mettre en évidence des différences et des ressemblances entre isolats [87].

b2. Ribotypage

Il existe différentes variantes de cette technique. Une première méthode fait appel à l'analyse du profil de restriction des gènes de l'ARN ribosomique pour distinguer les isolats bactériens. Le ribotypage implique l'isolement de l'ADN bactérien total suivi de sa digestion par des enzymes de restriction spécifiques. L'ADN digéré est séparé par électrophorèse sur un gel d'agarose et transféré sur une membrane en nylon ou nitrocellulose. Les fragments d'ADN sur la membrane sont ensuite hybridés avec des fragments d'ADN marqués complémentaires

des séquences de gènes ARNr. Chaque fragment d'ADN bactérien contenant un gène d'ARN ribosomal sera ainsi mis en évidence, créant ainsi une empreinte. Cependant, l'ADN codant pour l'ARNr est très bien conservé parmi les bactéries étroitement apparentées, ce qui rend le ribotypage moins discriminant que d'autres procédures de typage.

Une autre variante (ISR PCR) se base sur l'utilisation d'amorces spécifiques complémentaires de régions hautement conservées de l'extrémité 3' du gène codant pour l'ARNr 16S et de l'extrémité 5' de celui codant pour l'ARNr 23S. Cela permet d'amplifier par PCR l'espace intergénique des régions 16S-23S afin de détecter leurs polymorphismes [88]. La méthode peut être optimisée en termes de : pouvoir discriminant, facilité d'utilisation, vitesse et au niveau des exigences de sécurité par la sélection appropriée des amorces, la séparation électrophorétique des produits de PCR et les techniques de détection de bandes [89–91]. Le *ribotyping* par PCR est moins discriminant que la PFGE, mais plus résolutif que la PCR classique se basant sur des amorces arbitraires dans le cas de l'analyse épidémiologique des isolats de *Clostridium difficile*. Pour cette espèce, il a été montré que cette méthode est une technique de routine reproductible et robuste. En effet, celle-ci a permis des comparaisons inter-laboratoires et la mise en place de base de données dédiées [92, 93]. [94]

c. Séquençage.

La séquence d'ADN est une donnée fiable et numérique. Elle est constituée d'un nombre déterminé de valeurs discrètes et les méthodes basées sur le séquençage sont celles qui sont le plus portable. En effet, les séquences sont stockées sous forme de fichier texte ou binaire. Les données sont comparables entre elles quelle que soit la technologie de séquençage utilisée. Elle est *de facto* une connaissance facilement échangeable par internet. Pour ne citer que les deux dernières années (entre Octobre 2017 et Octobre 2019), le nombre de bases issues de WGS a triplé, passant de $2{,}3.10^{12}$ à $5{,}98.10^{12}$ bases. Le nombre de séquences WGS a doublé sur cette même période, passant de 5.10^{8} à 10.10^{18}. Depuis 1982 jusqu'à aujourd'hui, le nombre de bases présentes dans GenBank double environ tous les 18 mois.

Nous verrons dans un premier temps le *Multi-Locus Sequence Typing*, qui est une méthode essentielle marquant la transition entre l'ère moléculaire et l'ère génomique en épidémiologie. Nous détaillerons ensuite quelques applications basées sur le séquençage de génomes entiers.

c1. Multi-Locus Sequence Typing (MLST) « classique ».

Depuis son développement en 1998 [95], la MLST s'est rapidement imposée comme la technique de pointe pour le typage moléculaire bactérien. La MLST se base sur une sélection de 7 à 11 « gènes de ménage » (gènes se trouvant par définition dans toutes les bactéries et dont les polymorphismes sont probablement relativement neutres, c.-à-d., ne subissant pas de sélection adaptative fréquente). Cette sélection est appelée schéma MLST. Ces gènes sont partiellement séquencés (réactions de PCR sur les locus puis séquençage Sanger). Pour chaque gène, chaque allèle est identifié par un nombre arbitraire qui définit l'*allele-type* (AT). L'attribution allélique est basée sur la comparaison d'un échantillon par rapport aux allèles connus présents dans la base de données. Si l'allèle est inconnu, un nouveau numéro d'AT est attribué. La combinaison des ATs défini le *sequence type* (ST). Le succès de la MLST a été indéniable et repose sur un constat très simple : cette approche a apporté pour la première fois la reproductibilité et la portabilité nécessaire au développement d'une banque de données mondiale de typage d'agents pathogènes facilement accessible aux organismes de santé publique et de recherche. Le premier schéma MLST à avoir été porté sur Internet fut pour l'espèce *Neisseria meningitidis* [95], et cette tendance s'est rapidement développée pour inclure d'autres espèces bactériennes [28, 96–99]. Aujourd'hui, il existe des banques de données pour environ 79 organismes (75 bactéries, 3 champignons et 1 protozoaires) qui offrent 3 types de requête : identification et comparaison de séquences d'allèles, identification et comparaison de profils alléliques et appariement d'isolats. Cependant, cette approche se heurte à des limites : choix des locus, mise au point des conditions de PCR et de séquençage avec la méthode Sanger sur de nombreux locus, temps de main d'œuvre et coût.

c2. Génomique comparée : séquençage de génomes entiers (WGS) et technologie de séquençage de dernière génération (NGS).

i. Exemple de techniques

Le séquençage de génomes entiers apparaît aujourd'hui comme la technique la plus précise pour comparer des isolats bactériens. Les applications du « whole genome shotgun » (WGS) en microbiologie clinique et en santé publique ont déjà été testées au travers d'études réalisées rétrospectivement [100, 101]. Salipante et al. (2015) ont étudié l'utilisation du WGS comme méthode de typage de souches pour les laboratoires cliniques en utilisant un protocole universel et unique (regroupant la préparation des librairies, le séquençage et l'analyse de données) pour 3 espèces différentes : *Staphyloccocus aureus* résistante à la méthicilline, *Enterococcus faecium* résitante à la vancomycine et *Actinobacter baumannii* multirésistante aux antibiotiques. Ils ont montré que le WGS était hautement reproductible et permettait une définition fonctionnelle de la clonalité [102]. Ensuite, Kwong et al. (2016) ont comparé le WGS prospectif en routine à des méthodes conventionnelles de typage, comme la MLST et la MultiLocus Variant Analysis (MLVA), pour la surveillance épidémiologique de *Listeria monocytogenes*. Les résultats de MLST, de typage binaire et de sérotypage *in silico* déduits des données du WGS étaient très concordants (>99 %) avec les tests classiques *in vitro* correspondants effectués en laboratoire [103]. De plus, le WGS permet d'identifier des clusters distincts au sein de groupe d'isolats qui sont indiscernables en utilisant les approches classiques. Ces études montrent que le WGS apporte un plus grand niveau de discrimination que les autres méthodes conventionnelles, à la fois pour la surveillance ou pour l'inférence de lien(s) entre les différents foyers infectieux.

Différentes approches peuvent être utilisées pour traiter les données génomiques en épidémiologie. Nous en verrons ici quelques exemples afin d'illustrer l'intérêt majeur du WGS.

Tout d'abord, ces données peuvent permettre d'étudier les *single nucleotide polymorphism* (SNP) à l'échelle du génome. Un SNP est une variation d'un seul nucléotide qui apparaît dans les génomes, où chaque variation est présente dans une proportion appréciable de la population. L'approche *Genome-wide SNP typing* a été par exemple utilisée pour génotyper les souches de *Bacillus antracis* et *Bacillus cereus* [104], *Staphyloccocus suis* [105] ou *Coxiella burnetii* [106], entre autres, pour identifier les foyers d'infection [107]. Enfin, des pipelines ont été développés pour le génotypages de SNPs pour identifier les souches bactériennes ainsi que des échantillons métagénomiques [108].

D'autre part, certains auteurs ont essayé de développer des méthodes semblables à la MLST, qui présente notamment l'avantage de diminuer les effets dus à la recombinaison (plusieurs mutations au sein du même gène étant vues comme un seul changement). Nous avons vu que le génotypage par MLST est une approche intéressante pour le typage des souches bactériennes. Cependant, sa mise en œuvre peut être coûteuse, chronophage et laborieuse. Les méthodes suivantes sont donc de plus en plus souvent utilisées en remplacement de la MLST « classique » par PCR, tirant profit du séquençage haut-débit. La rMLST a été imaginée pour indexer les variations moléculaires des 53 gènes codant pour les sous-unités protéiques du ribosome [109]. La cgMLST, quant à elle, inclue tous les gènes du *core-genome* (celui-ci étant l'ensemble des gènes communs à un groupe de génomes) afin d'augmenter la résolution de l'analyse pour des populations de bactéries [110]. La wgMLST intègre quant à elle, également tous les gènes du génome accessoire. Cela peut permettre une

comparaison plus fine lorsque certains gènes ne sont présents que dans des groupes de souches particuliers. Des approches de type MLST ont aussi été développées autour de technologies particulières : c'est le cas de la HiMLST reposant sur le *Genome Sequencer Junior* de Roche [111] ou la NGMLST basée sur une PCR couplée à du séquençage PacBio [112].

Enfin les génomes peuvent être également être traités à l'aide de programmes d'alignement multiple, d'outils de reconstruction phylogénétique et d'outils de caractérisation de la dynamique de population (estimation des paramètres correspondant aux taux de recombinaison, taux de mutation, croissance de population et forces de sélection). [113]

Le séquençage de génome entier peut donc fournir des informations pertinentes pour la microbiologie clinique, allant du phénotypage *in silico* jusqu'au suivi des dynamiques épidémiologiques. Cependant, l'usage du séquençage en clinique reste encore aujourd'hui limité et privilégié notamment pour des agents pathogènes humains majeurs bénéficiant d'un important soutien financier. En effet, la diminution des coûts du séquençage brut ne s'est pas traduite par une diminution drastique des coûts totaux, qui eux se sont stabilisés. Ces coûts ne prennent pas en compte le coût du matériel informatique adapté pour traiter ces données ainsi que le coût du personnel qualifié pour effectuer ce type d'analyse. De plus, les informations de plus en plus nombreuses apportées par le WGS peuvent entrer en conflit avec les concepts microbiologiques traditionnels et les schémas de typage. Enfin, il existe déjà des pipelines d'analyse utilisés dans la recherche publique, mais ils ne sont pas toujours accessibles au clinicien classique car ils sont complexes et souvent utilisés en ligne de commande. Il y a donc une nouvelle étape à franchir avec le développement de pipelines d'analyse spécifiquement imaginés pour la microbiologie clinique.

ii. Mise en œuvre pratique du WGS clinique

Un des problèmes majeurs des analyses WGS est qu'il n'existe pour le moment que peu ou pas de références méthodologiques. Les étapes fondamentales de ces analyses pour la génomique microbienne pour ces différentes applications sont : le contrôle qualité des séquences, l'identification et la confirmation de l'isolat séquencé, la caractérisation de l'isolat (en incluant un effort de typage ainsi que la caractérisation de facteur(s) de virulence et le profil de résistances aux antibiotiques – AMR – prédictif), l'analyse épidémiologique, et enfin, le stockage des résultats. Malgré cela, la mise en œuvre de ces analyses est variable selon les espèces bactériennes étudiées et les laboratoires. Cependant, nous pouvons observer que les outils d'analyse en ligne « clés en main » se multiplient, comme https://cge.cbs.dtu.dk/services/cgMLSTFinder/. cgMLSTFinder permet de réaliser une cgMLST à partir de lectures brutes en interrogeant 6 bases de données MLST différentes (groupes d'espèces d'intérêt clinique).

Une des limitations majeures pour obtenir rapidement de l'information pertinente dans un environnement clinique est que les pipelines d'analyse préexistants pour la génomique bactérienne ont été développés pour la recherche fondamentale ou l'épidémiologie de la santé publique [114]. Cela implique généralement que ces pipelines permettent un flux d'analyses complet et sophistiqué, avec un grand nombre de modules interchangeables et d'options. Par exemple, la plateforme Galaxy possède par défaut 35 outils différents, tests et flux d'analyses pour la partie « *QC and manipulation* » de données NGS. De telles solutions sont vraiment pertinentes pour le chercheur spécialisé. En revanche, pour le clinicien cela peut paraître rédhibitoire et finalement inutilisable pour des analyses en temps réel. De plus, l'utilisateur

doit parfaitement connaître l'intérêt des outils utilisés, leurs avantages et leurs inconvénients relatifs aux méthodes employées et une compréhension fonctionnelle des paramètres critiques. Enfin, ces pipelines requièrent encore trop souvent une maîtrise des systèmes Linux et une maîtrise avancée de la ligne de commande du terminal.

Afin d'évoluer vers des résultats interprétables en temps réels pour les laboratoires cliniques, il sera nécessaire de simplifier le protocole. L'attention doit se porter sur une analyse automatisée et rapide générant des résultats clairs, fiables et non équivoques. Certaines étapes pourraient être simplement retirées pour des finalités cliniques. Par exemple, l'assemblage de génome pourrait paraître comme un goulot d'étranglement pour le WGS en temps réel pour le diagnostic. En fait, il s'avère que l'assemblage des génomes est rarement nécessaire car la caractérisation suffisante d'un isolat peut être effectuée par l'analyse des k-mers (sous-chaînes de caractères de longueur k) directement à partir des données brutes. L'identification exacte d'un isolat peut être obtenue avec des méthodes de comparaison de k-mers basées par exemple sur l'algorithme MinHash *(min-wise independent permutations locality sensitive hashing scheme*). Cet algorithme permet d'estimer rapidement la similarité de deux ensembles en utilisant la distance de Jaccard (distance binaire asymétrique). En génomique, c'est une estimation du nombre de k-mers identiques partagés entre 2 génomes. Il est inclus dans le programme Mash [115]. De plus, les éléments d'AMR peuvent aussi être identifiés à l'aide des seuls k-mers [116]. Un autre exemple d'une étape clef mais coûteuse (en temps de calcul) qui pourrait être retirée d'un pipeline type est l'inférence phylogénétique sophistiquée. La meilleure méthode pour la reconstruction d'arbres phylogénétiques pourrait inclure l'évaluation de la vraisemblance individuelle d'une grande variété d'arbres possibles, étant donné un alignement de séquences et une matrice de distances, répétée pour des milliers de répliquas de *bootstrap*. Cela aboutit à l'obtention d'un seul arbre avec la vraisemblance la plus haute, mais avec un coût élevé en temps de calcul. Un pipeline clinique pourrait utiliser des approches plus rapides et fournissant toujours des arbres phylogénétiques informatifs [117].

La figure 4 est une vision schématique d'un pipeline computationnel spécifique pour le diagnostic en microbiologie clinique proposée par Balloux *et al.*, qui illustre les aspects discutés dans ce paragraphe. Un pipeline clinique ne devrait contenir qu'un sous-ensemble des modules classiques proposés dans les pipelines de recherche académique. Ces modules devraient être dédiés à une sortie rapide et interprétable des résultats. Par exemple, la reconstruction phylogénétique pour l'inférence épidémiologique pourrait se baser sur une matrice de distance de Jaccard et la matrice obtenue pourrait être utilisée pour générer des arbres phylogénétiques à moindre coût (par exemple UPGMA ou Neighbor-Joinning). De plus, des corrélations entre la distance génétique par paires et la date d'échantillonnage peuvent être testées pour mettre en évidence un signal temporel (c'est à dire l'accumulation d'un nombre significatif de mutations au cours d'une période d'échantillonnage donnée). En présence d'un tel signal, l'utilisateur pourrait alors reconstruire une chaîne de transmission de l'agent pathogène basée sur un algorithme rapide tel que Seqtrack qui vise à reconstituer les généalogies des haplotypes ou génotypes échantillonnés pour lesquels une date de collecte est disponible [118].

Un pipeline dédié au diagnostic clinique devrait également être relié à des bases de données multi-espèces contenant des informations sur les dernières avancées concernant les schémas de typage, et contenant des facteurs cliniques d'importance comme les déterminants d'AMR. Les résultats devront être validés régulièrement, et des accréditations internationales délivrées à intervalles réguliers. Au niveau national, les organismes accréditeurs peuvent

manquer d'expertise. Il est malheureusement fréquent que des bases de données s'effondrent lorsque les financements s'arrêtent ou que la personne responsable quitte son poste. Si le WGS s'impose finalement un jour dans les laboratoires cliniques, il sera primordial de sécuriser des financements pérennes des infrastructures et des personnels dédiés pour de telles bases de données.

Le manque d'adoption global du diagnostic basé sur le WGS pourrait également être dû à une volonté compréhensible de maintenir un « *statu quo* » dans le milieu clinique. Par exemple, le milieu hospitalier souffre déjà d'un certain nombre de pressions et privilégie des systèmes établis de traitements et d'intervention. En outre, et de manière probablement plus significative, cela peut mettre en lumière une communication difficile sur les bénéfices potentiels du WGS dans la routine clinique. Les principaux défenseurs du WGS se basent souvent sur le milieu de la santé publique ou de la recherche académique et participent rarement activement à la prise de décisions cliniques. En soi, cela peut être vu comme une barrière linguistique remettant en question la mise en place d'un dialogue constructif pouvant mener à des améliorations quantifiables des systèmes existants. Enfin, la planification physique, la mise en œuvre et l'intégration des diagnostics WGS constitue un changement profond de paradigme. Il a intrinsèquement peu de chance d'aboutir sans une introduction rigoureusement planifiée et une formation continue des utilisateurs de premières lignes. Cet immense défi nécessite des investissements humains, temporels et financiers colossaux. Il est donc aujourd'hui mis en échec par l'infrastructure déjà limitée en ressources de nombreux milieux cliniques. [119]

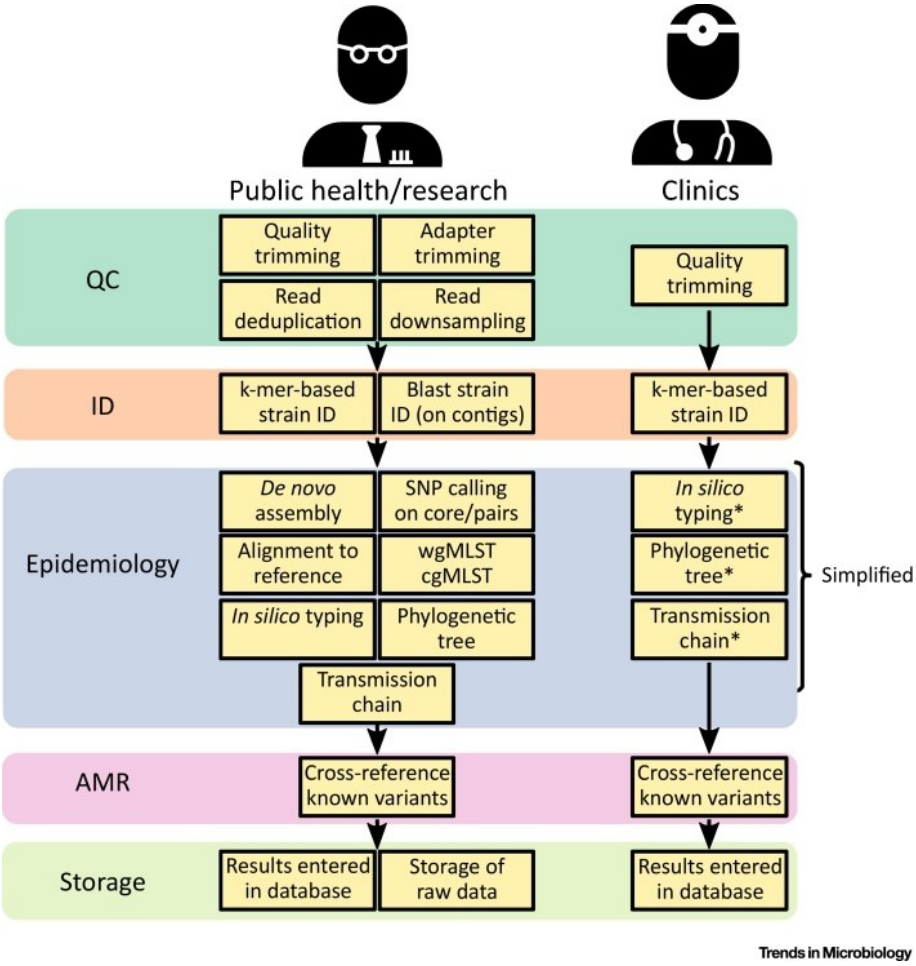


Figure 4: Simplification des pipelines d'analyse utilisés dans le cadre de recherches académiques en santé publique pour effectuer de l'épidémiologie hospitalière.

QC= *quality control; ID = identification; AMR = antimicrobial resistance*

E4 – Conclusion.

Les solutions développées pour identifier, caractériser et typer les isolats bactériens suivent l'évolution constante de la technologie. Bien que des méthodes performantes existent déjà, les problèmes restent les mêmes : coût, rapidité, fiabilité des résultats et facilité d'interprétation (conclusions non équivoques). Ces problèmes sont contraints de manière différente entre les domaines (recherche académique, santé publique, laboratoire de diagnostic clinique ou vétérinaire) et bien entendu les agents pathogènes étudiés. Par exemple, les laboratoires étudiant les agents pathogènes humains majeurs bénéficient fort heureusement d'un soutien logistique significativement plus important que ceux étudiant les agents pathogènes de poissons ! Aussi, la génomique est l'outil ultime permettant d'identifier, caractériser, typer un isolat, de prédire des fonctions, des phénotypes, des résistances mais également de suivre l'évolution des populations et reconstruire la chaîne de transmission d'un foyer épidémique en comparant les génomes des isolats. En revanche, la banalisation du séquençage haut-débit dans le domaine clinique et vétérinaire n'est pas encore une réalité tangible au regard des contraintes actuelles (économiques, logistiques, …). En attendant des conditions plus favorables pour un tel changement de paradigme, il est nécessaire de développer des solutions alternatives. Ces méthodes doivent être fiables, rapides, se prêter au haut-débit, et peu onéreuses. En outre, les résultats doivent être facilement interprétables et transposables d'un laboratoire à un autre. Peu de technologies possèdent l'ensemble de ces qualités. Parmi celles-ci la spectrométrie de masse MALDI-TOF (en anglais, *Matrix Assisted Laser Desorption Ionisation - Time of Flight*) est une technologie qui s'est progressivement imposée comme un outil d'identification intéressant, aussi bien en médecine humaine que vétérinaire. Ses avantages (par rapport aux méthodes décrites dans ce chapitre) sont considérables. En effet, elle permet de réaliser une identification peu onéreuse et très rapide, tout en étant fiable et reproductible (quelques minutes à quelques heures sont nécessaires pour l'acquisition des données et leur analyse). Cette technologie ne fournit pas une information aussi riche et précise que l'information génomique, mais présente l'avantage de pouvoir être intégrée à moindre coût dans une routine clinique. C'est un outil intéressant puisqu'il permet de capturer une partie du signal génétique exprimé et de révéler différents types d'informations. Celles-ci sont critiques pour la microbiologie clinique, et l'utilisation de cette technologie offre donc une réelle valeur ajoutée pour les laboratoires. Cependant, il faut noter qu'il est nécessaire d'investir un temps important pour constituer des pipelines d'analyses qui seront pertinents dans une routine clinique. Au travers du chapitre suivant, nous explorerons en détail la spectrométrie de masse MALDI-TOF. Nous verrons son principe de fonctionnement ainsi que la nature complexe des données qu'il génère. Nous détaillerons les applications actuelles du MALDI-TOF en microbiologie. Enfin, nous chercherons à définir la nature de l'information contenue dans les spectres, et à en comprendre le sens mais aussi les limites.

F – La spectrométrie de masse MALDI-TOF

La spectrométrie de masse (MS) est une technique analytique selon laquelle les composés chimiques sont ionisés en molécules chargées et leur rapport masse sur charge (m/z) est mesuré. Bien que découverte aux débuts des années 1900, son application se limitait à la chimie. Cependant, le développement de l'ionisation par pulvérisation électronique (ESI) et l'ionisation par désorption laser assisté par matrice (MADLI) dans les années 1980 a permis d'adapter cette approche aux grandes molécules biologiques comme les protéines. Aussi bien en ESI qu'en MALDI, les peptides (ou protéines) sont convertis en ions par addition ou perte d'un ou plusieurs protons H+. Ces deux méthodes sont basées sur une ionisation douce permettant en principe de ne pas compromettre l'intégrité de l'échantillon. Le MALDI possède certains avantages comparé à l'ESI car il ne nécessite pas d'étape préliminaire de chromatographie [120].

Le haut-débit et la vitesse d'acquisition élevée associés à une automatisation complète ont fait du MALDI-TOF MS un choix évident pour les recherches en protéomique à grande échelle [121] [122].

F1 – Principe de fonctionnement

Pour être analysé par MALDI-TOF MS, un échantillon est mélangé ou enrobé avec un composé organique appelé matrice et capable d'absorber de l'énergie. L'ensemble co-cristallise pendant le séchage à l'air libre. Cette matrice est cruciale car elle fournit les protons pour l'ionisation (qui sera initiée par un rayon laser) et sert également de support sur laquelle l'ionisation peut se produire [123]. Le laser pulsé à azote agit comme un catalyseur de la réaction d'ionisation (Figure 5, volet 1). A l'heure actuelle, il n'existe pas de modèle consensus sur le mécanisme d'ionisation. En effet, les différents modèles proposés n'expliquent pas tous les résultats observés en routine. Quoiqu'il en soit, la désorption-ionisation qui en résulte génère des ions isolés protonés à partir des composés (analytes) présents dans l'échantillon. Ces ions sont ensuite accélérés avec un potentiel donné, ce qui permet de les séparer sur la base de leur rapport masse sur charge, exprimé par m/z (Figure 5, volet 2). Les analytes chargés sont ensuite détectés et mesurés (Figure 5, volet 3) par un analyseur de temps de vol (*Time Of Flight*, TOF). Durant une acquisition MALDI-TOF, le rapport m/z est mesuré en déterminant le temps requis pour que les molécules chargées parcourent toute la longueur du tube de vol sous vide. Cette relation entre la masse et le temps de vol est explicitée par les formules ci-dessous. L'énergie potentielle (1) est convertie en énergie cinétique (2). La relation (4) montre le lien entre la vitesse, la charge et la masse de l'analyte. Enfin, en rajoutant l'équation de définition du temps (5) dans l'équation (4), nous obtenons la relation entre le temps (de vol) et la masse des analytes (6). Cette équation résume le principe fondamental du spectromètre de masse MALDI-TOF.

Certains analyseurs TOF sont également pourvus d'un miroir ionique à l'extrémité arrière du tube de vol, qui sert à réfléchir les ions dans le tube jusqu'au détecteur. Ainsi, ce miroir n'augmente pas seulement la longueur du tube, il corrige également les petites différences d'énergie entre les ions [124]. Un spectre caractéristique appelé empreinte de masses peptidiques (ou empreinte de masses protéiques, dans le cas de protéines intactes) ou PMF en anglais, est généré par les analytes contenus dans l'échantillon. Ce spectre est basé sur l'information du temps de vol. [122]

$$(1)\ E_{potentielle} = zeV$$

$$(2)\ E_{cinétique} = \frac{1}{2}mv^2$$

$$(3)\ zeV = \frac{1}{2}mv^2$$

$$(4)\ v = \sqrt{\frac{2zeV}{m}}$$

$$(5)\ t = \frac{d}{v}$$

$$(6)\ t = d\sqrt{\frac{m}{2zeV}}$$

Tableau 1: Tableau des variables

Variables	Signification
z	Charge d'un analyte (nombres d'électrons)
e	Énergie d'un électron (1.6 x 10^{-19} Coulombs)
V	Potentiel électrique de l'analyte
m	Masse de l'analyte
v	Vitesse de l'analyte
t	Temps (de vol)
d	Distance parcourue (longueur du tube sous vide)

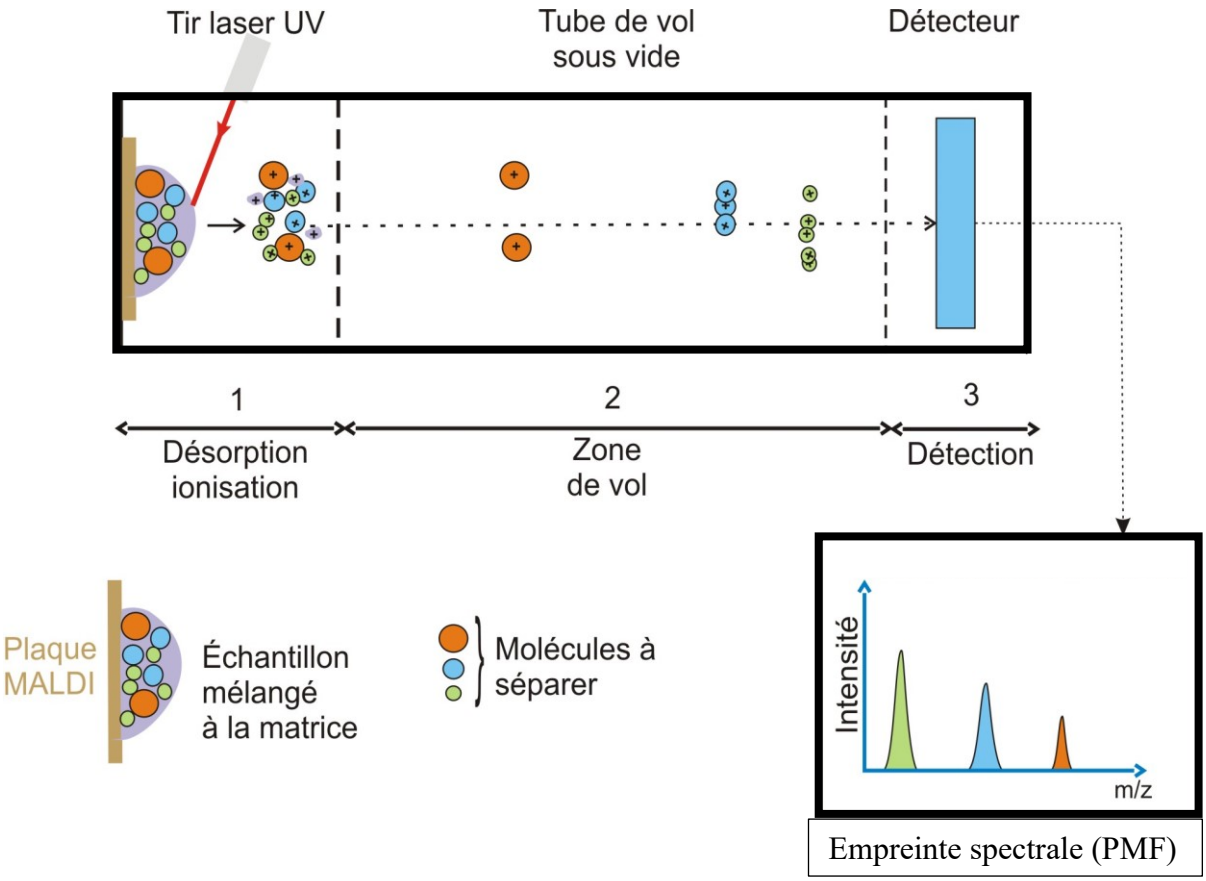


Figure 5: Principe de fonctionnement du MALDI-TOF (schéma de Mr.MORIDA)

F2 – Applications de la spectrométrie de masse MALDI-TOF en microbiologie.

a. Identification de microorganismes

L'application clinique du MALDI-TOF MS à l'identification de microorganismes est principalement réalisée à partir de cultures bactériennes sur boîtes de Pétri et transfert direct de colonies. Avec cette approche, une petite quantité de bactéries (10^4 – 10^5 bactéries) peut permettre l'identification d'une grande variété d'espèces. Une colonie est collectée manuellement depuis la culture sur gélose à l'aide d'un cure-dent stérile et déposée sur une plaque MALDI en acier rectifié, qui a une structure régulière et fine facilitant la préparation homogène des échantillons. Ces derniers sont séchés et un petit volume de matrice (généralement l'acide α-Cyano-4-hydroxycinnamique ou l'HCCA pour cette application spécifique du MALDI-TOF) est ajouté pour co-cristalliser avec l'échantillon bactérien sur la plaque MALDI. Enfin, la plaque est déposée dans le spectromètre de masse pour l'acquisition des données. Une fois les spectres acquis par l'appareil, des logiciels (souvent ceux fournis avec le spectromètre de masse) permettent de comparer l'empreinte spectrale obtenue à celles contenues dans banques de données.

L'identification est généralement réalisée à l'aide de systèmes commerciaux qui chacun possèdent une banque de données spécifique : le MALDI-TOF Biotyper (Bruker Daltonics, Bremen, Germany), le *Spectra Archive Microbial Identification System* (SARAMIS™ ; AnagnosTech, Postdam, Germany), Andromas (Andromas ; Paris, France) et le Vitek MS (bioMérieux, Marcy l'Etoile, France). Les différences entre ces différents systèmes sont : les instruments, les algorithmes de traitement des spectres, ceux liés à l'identification et la ou les banque(s) de données utilisée(s).

Des études comparées approfondies ont évalué la reproductibilité du MALDI-TOF en confrontant les résultats obtenues par des laboratoires différents et ont conclu à une haute reproductibilité du MALDI-TOF en routine d'identification bactérienne clinique [125]. La précision du MALDI-TOF MS a été évaluée à de nombreuses reprises sur divers groupes d'espèces d'intérêt clinique. En 2010, une étude a comparé le MALDI-TOF MS avec des tests biochimiques utilisés pour l'identification bactérienne de routine. Une identification correcte et fiable a été observée dans 99,1% des cas [126]. Une autre étude a évalué le taux d'identification global correct pour les staphylocoques à coagulase négative (staphylocoques commensaux de la peau humaine dont certains sont à l'origine d'infection nocosomiales, et plus particulièrement *S. epidermidis*, *S. haemolyticus*, *S. saprophyticus*, *S. capitis*, et *S. lugdunensis*)[127]. En 2011, une étude démontra que le MALDI-TOF MS est fiable (précision de 99,3%) pour l'identification d'espèces variées de staphylocoques et peut être considéré comme équivalent à la méthode standard d'identification basée sur les séquences du gène *rpoB* [128]. En 2012, une étude a montré que le taux d'identification du MALDI-TOF était de 94,9% au niveau du genre (39 genres différents) et de 83,4% au niveau de l'espèce (102 espèces), en analysant une collection de 296 isolats (aussi bien Gram-positifs que Gram-négatifs) [129]. Une autre étude a suggéré que le MALDI-TOF MS était la méthode de choix pour l'identification des espèces appartenant au genre *Campylobacter* et les microorganismes apparentés par rapport aux autres systèmes commerciaux [130]. Pour l'évaluation de la méthode sur les bactéries à Gram-négatif, en étudiant 2263 espèces significatives sur le plan clinique, Faron et al. ont montré que le MALDI-TOF MS identifie correctement 99,8% d'entre elles (2258/2263) au niveau du genre et 98,2% (2222/2263) au niveau de l'espèce pour des isolats provenant de différents établissements [131]. De plus, Garner et al. ont montré que l'identification des bactéries anaérobies à Gram-négatif était de 91,7% au niveau

de l'espèce et de 92,5% au niveau du genre [132]. Pour l'évaluation de l'identification des bactéries aérobies à Gram-positif, 1146 isolats ont été testés. Rychert et al. ont montré que l'identification était correcte à 92,8% (1063/1146) à l'échelle de l'espèce et à 95,5% (1094/1146) à l'échelle du genre [133]. Le MALDI-TOF MS a été utilisé pour l'identification rapide de bactéries à partir de bouteilles positives de cultures sanguines, après un court temps d'incubation (5,5 h) sur milieu solide, avec une précision de 82,3% [134]. Les infections urinaires, y compris les cystites (infection de la vessie et des voies urinaires inférieures) et la pyélonéphrite (infection des reins et des voies urinaires supérieures) sont parmi les types les plus courants d'infections bactériennes cliniques. Pour les cas les plus aiguës ou les plus compliqués, le diagnostic et le traitement rapide sont très importants. La réduction du temps nécessaire pour l'identification est donc cruciale et peut être effectuée en réalisant l'analyse directement à partir d'échantillons d'urine centrifugés à faible vitesse afin d'éliminer les leucocytes puis à vitesse élevée pour recueillir les bactéries. Les culots sont lavés puis déposés sur la plaque MALDI-TOF. Le taux d'identification était de 91,8% au niveau du genre et de 92,7% au niveau de l'espèce [135]. Une autre étude a révélé que le MALDI-TOF MS était capable d'identifier de manière fiable les bactéries directement dans les échantillons d'urine à des concentrations très faibles (10^3 unités formant colonies, CFU). Les mycobactéries sont responsables d'un grave problème de santé publique à l'échelle mondiale [136]. Le genre *Mycobacterium* fait partie des actinobactéries et comprend plus de 190 espèces connues. *M. tuberculosis* est l'agent causant le plus de morbidité chez l'homme. Ces dernières années, l'incidence des maladies dues aux mycobactéries non tuberculeuses a augmenté, ceci étant dû au nombre croissant de patients atteint de maladies auto-immunes et de personnes immuno-déprimées [137]. Cependant, l'identification de *Mycobacterium* est difficile en raison de la lenteur de la culture et des réactions biochimiques (de 7 à plus de 21 jours). A l'aide du MALDI-TOF, une étude a montré un taux d'identification de 100% pour *M. tuberculosis,* mais de seulement 38,5% pour les mycobactéries responsables d'infections non tuberculeuses. Le MALDI-TOF MS a donc une haute sensibilité pour *M. tuberculosis* et permet de réduire le temps d'identification à seulement 45 minutes et s'avère la méthode la plus économique et la plus rapide pour l'identification de *M. tuberculosis* [138]. Les infections à certains champignons et levures sont également des complications courantes chez les patients immunodéprimés. La spectromètrie de masse MALDI-TOF s'est dont également développée pour l'identification de ces agents, car leur identification par des méthodes classiques reste coûteuse et longue. Plus récemment, des auteurs se sont intéressés à la capacité du MALDI-TOF à caractériser des spécimens eucaryotes. Par exemple, Vega-Ruà et al. (2018) ont présenté une stratégie d'identification MALDI-TOF comme un outil innovant et alternatif pour l'identification des moustiques. Cette approche singulière pourrait avoir un impact positif spectaculaire dans le domaine de la surveillance des moustiques dans le monde [139].

Ainsi, la spectrométrie de masse MALDI-TOF est une technologie récente dont l'utilisation pour l'identification de micro-organismes s'est grandement développée. Il s'agit d'une méthode à haut-débit, à faible coût et d'une grande précision. Ceci est particulièrement important pour la microbiologie clinique de routine car les résultats peuvent venir directement d'échantillons (de sang, d'urines, ou d'autres fluides biologiques). La fiabilité et la précision du MALDI-TOF MS ont été vérifiées dans un certain nombre d'études, et la précision de l'identification repose pour une bonne partie sur le nombre d'entrées de la base de données utilisée. [122, 140]

b. Typage d'isolats bactériens par MALDI-TOF MS.

Les méthodes de typage classiquement utilisées (sérotypage, ribotypage, PFGE, MLST…) ont été présentées dans le chapitre E3 [141]. Bien que ces techniques soient largement reconnues, elles restent encore souvent chronophages, coûteuses et laborieuses. Une méthode idéale pour le typage bactérien de routine serait une méthode nécessitant une préparation minimale des échantillons à la fois rapide, automatisée et économique. Le MALDI-TOF MS présente certaines de ces qualités, et, naturellement, les microbiologistes étaient désireux de l'utiliser pour le typage bactérien. Celui-ci a donc été testé pour typer quelques de bactéries pathogènes comme *Escherichia coli* [142–156], *Staphylococcus aureus* [157–166], *Pseudomonas aeruginosa* [167], *Acinetobacter baumanii* [168], *Klebsiella pneumoniae* [169–171], *Neisseria meningitidis* [172], *Listeria monocytogenes* [173], *Haemophilus influenzae* [174], *Streptococcus pneumoniae* [175–177] et *Streptococcus pyogenes* [178, 179]. Pour certaines espèces, des groupes de souches à l'intérieur de l'espèce peuvent être séparés par MALDI-TOF. Cette capacité du MALDI-TOF à typer les souches d'une espèce bactérienne dépend de ses caractéristiques intrinsèques. Il sera difficile, voire impossible, de capturer un polymorphisme suffisant pour des espèces très homogènes et présentant très peu de diversité génétique (e.g., *Flavobacterium psychrophilum* ou *Yersinia pestis*). Spinali *et al.* ont récemment examiné en détail ces études et ont soulignés les divergences entre elles qui seraient probablement dues aux différences méthodologiques (c'est-à-dire, les différentes solutions face aux problèmes technologiques ou biologiques liés à la spectrométrie de masse, les ensembles de souches étudiées, la définition des pics spécifiques et les analyses statistiques) [180]. [181]

Un des attraits pour le typage par MALDI-TOF est que les spectres utilisés pour l'identification d'échantillons peuvent également contenir beaucoup plus d'informations qu'il n'en faut pour leur affectation fiable à une espèce donnée. Une question évidente s'impose alors : au-delà de l'identification de groupes intra-spécifiques (ou « MALDI-types »), peut-on se servir du MALDI-TOF pour obtenir des informations cliniques pertinentes?[181] Le typage peut être basé sur le spectre entier ou sur l'utilisation d'un unique pic, ou d'un sous-ensemble de pics. Afin de compléter les informations synthétisées par Sauget et al. [181], voici quelques exemples pertinents illustrant l'utilisation du MALDI-TOF comme outil épidémiologique :

- Giacometti et al. ont analysé des données spectrales d'*Arcobacter buzlori*, aussi bien en prenant en compte l'intégralité du spectre ou seulement quelques pics d'intérêt. Ils ont observé que le pouvoir discriminant de leur approche MALDI-TOF était inférieur à celui de la PFGE et de la MLST. Cependant, cette méthode pouvait attribuer correctement les isolats à des *sequence types* (ST : groupes MLST) [182].
- L'utilisation du MALDI-TOF comme outil de première ligne pour la détection d'évènements de transmission dans les foyers nosocomiaux à *Serratia marcescens* et *Citrobacter freundii*. [183]
- L'identification du ST37 et du clade MLST 4 de *Clostridium difficile* peut être réalisée par la détection de quelques pics spécifiques systématiquement présents et absents d'autres sous-groupes [184, 185].
- Pour les isolats identifiés comme appartenant à l'espèce *Staphyloccocus aureus*, la présence d'un pic à 2145 m/z semble indiquer une résistance à la méthicilline [186], mais la conclusion inverse qui serait que l'absence de ce pic indique une sensibilité n'est pas valide [187]. D'un autre côté, les complexes clonaux de *S. aureus* (« staphylocoque doré ») résistants à la méthicilline,

(MRSA en anglais) peuvent être distingués par MALDI-TOF à des fins épidémiologiques [188].

- Les entérocoques résistants à la vancomycine peuvent être distingués par MALDI-TOF MS, en accord avec les ST définis par MLST [189]. Néanmoins, les auteurs de cette étude ont expliqué que ces résultats sont peut-être spécifiques de l'environnement clinique actuel. En effet, ils n'ont pas réussi à distinguer les souches provenant de foyers infectieux des autres.
- Veeneman *et al.* ont montré que les spectres d'isolats d'*E. coli* produisant de la beta-lactamase peuvent être différenciés des autres d'une manière très concordante aux résultats obtenus avec l'AFLP. Certains MALDI-types contenaient néanmoins plusieurs ALFP-types et les auteurs ont souligné que des contrôles très drastiques étaient nécessaires pour que les données soient reproductibles [151].

Une conclusion pouvant être tirée de ces études est qu'il n'est pas possible d'établir un schéma universel de typage pour l'ensemble des espèces bactériennes, mais seulement des règles générales sur la manière d'établir et de valider un schéma [180]. Il faut également souligner que la résolution du MALDI-TOF MS a certainement des limites et qu'il n'y a aucune garantie que les objectifs seront atteints [190]. Ceux qui cherchent à utiliser le MALDI-TOF pour effectuer du typage en routine ne peuvent probablement pas éviter d'analyser en profondeur un grand nombre de souches bien caractérisées et d'évaluer avec précision la valeur prédictive des pics retenus (par exemple l'association de pic(s) avec un groupe d'isolats particuliers associés à des informations cliniques comme la sensibilité ou la résistance à un antibiotique). Mais ils seront probablement récompensés par la découverte d'informations significativement pertinentes, qui seront ensuite faciles et rapides à retrouver dans les spectres. Quelques études démontrent les possibilités qu'offre le MALDI-TOF pour distinguer les isolats bactériens ayant des degrés de virulence différents. Cependant, cette application particulière et prometteuse du MALDI-TOF n'est présentée que dans quelques rares exemples [191–193] et n'a pas été évaluée sur une application clinique réelle. [194]

c. Résistance aux antibiotiques

Comme nous l'avons vu dans la partie précédente, les tests phénotypiques de sensibilité aux antibiotiques (ASTs) prennent au minimum 18 à 24 h. Certains systèmes automatisés nécessitent 5 h afin d'avoir les premiers résultats pour certains antibiotiques et pour des espèces pathogènes à croissance rapide. C'est un désavantage pour les interventions thérapeutiques ciblées et l'application de mesures de contrôle des infections [195]. D'un autre côté, les méthodes basées sur l'amplification d'ADN, bien que plus rapides que les approches phénotypiques, sont basées sur la présence ou l'absence de gènes de résistance spécifiques, qui ne sont pas toujours corrélées avec le phénotype. De plus, les méthodes moléculaires sont seulement disponibles pour quelques cas de résistance aux antibiotiques et les nouveaux mécanismes de résistance peuvent échapper à ces méthodes [196]. L'intérêt du MALDI-TOF est qu'il pourrait combler le fossé séparant l'identification de l'espèce d'une part et le statut de résistance d'autre part. Le MALDI-TOF, comme pour l'identification précise de l'espèce, accélère significativement la détection de la résistance, en comparaison aux méthodes AST classiques. Il existe 4 méthodes pour réaliser un AST par MALDI-TOF MS [197, 198].

La première consiste à analyser les spectres de microorganismes afin de détecter un profil de pics de résistance caractéristique. Le principe de cette approche consiste simplement à identifier des différences entre les spectres d'isolats sensibles et résistants d'un microorganisme donné (caractérisé par des méthodes classiques).

La seconde consiste à analyser l'hydrolyse bactérienne d'antibiotique β-lactame (MALDI *Biotyper-Selective Testing of Antibiotic Resistance-Beta-lactamase Assay*, MBT-STAR-BL *Assay).* L'hydrolyse est détectée par l'observation d'un décalage de masse spécifique après une période d'incubation de 30 à 180 minutes de l'agent pathogène en présence de l'antibiotique testé [199–201].

La troisième consiste à détecter les acides aminés stables marqués par des isotopes non-radioactifs, qui sont incorporés dans les protéines bactériennes néo-synthétisées (MALDI *Biotyper-Resistance Test with Stable Isotopes Assay,* MBT-RESIS *Assay*). La quantité d'acides aminés marqués présents dans les protéines nouvellement synthétisées en présence de l'antibiotique permet de déterminer si l'isolat est sensible ou résistant [202].

La dernière consiste à analyser la croissance bactérienne en présence ou en absence d'antibiotiques en utilisant un contrôle standard interne (MALDI *Biotype-Antibiotic Susceptibility Test Rapid Assay*, MBT-ASTRA). La bactérie testée est incubée durant une courte période de temps, qui dépend de l'espèce et de l'antibiotique. Ensuite, les cellules sont lysées, les spectres sont enregistrés et comparés à des références internes : le manque de croissance en présence d'un antibiotique (sensibilité) conduit à une diminution des intensités des pics et une croissance normale (résistance) entraîne des pics intenses [203].

La première approche est équivalente à des analyses génotypiques tandis que les autres correspondent à des approches conventionnelles de caractérisation biochimiques. [204]

F3 – Nature des données spectrales.

Les spectres générés par le spectromètre de masse MALDI-TOF sont des données analogiques complexes. Pour les comprendre, il faut d'abord s'intéresser à la matrice choisie. Celle-ci joue un rôle critique car elle détermine la nature des analytes qui seront observables dans l'empreinte spectrale. La matrice HCCA mentionnée précédemment permet de visualiser des molécules (essentiellement des protéines) ayant en général un poids moléculaire d'au moins 2000 Da. Celle-ci est la matrice standard pour les applications en microbiologie. D'autres matrices existent, par exemple le 1,5-diaminonapthalene (1,5-DAN) utilisé notamment pour l'imagerie à haute résolution ou la trihydroxyacetophenone (THAP) utilisée pour la détection d'oligonucléotides [205].

En microbiologie clinique, un des éléments importants du spectre est la présence systématique d'échos. En effet, les protéines sont chargées avec un ou plusieurs protons H+. Ainsi, plusieurs valeurs m/z peuvent correspondre à la même protéine (Figure 6). Par exemple, une protéine ayant capté deux protons sera deux fois plus rapide et aura un rapport m/z deux fois moindre que la masse réelle attendue. Cependant, une ionisation double ou triple est moins probable que l'ionisation simple. Effectivement, l'intensité des pics correspondants à une protéine di ou tri-chargée est généralement plus faible que celle du pic correspondant à une protéine mono-chargée. Nous pouvons supposer que l'abondance de certaines protéines augmente la probabilité relative d'observer ce phénomène. En effet, les analytes observés dans l'empreinte spectrale générée en microbiologie contiennent majoritairement des protéines abondantes de la cellule. Elles proviennent essentiellement du cytoplasme, elles ont une basicité forte et sont moyennement hydrophiles.

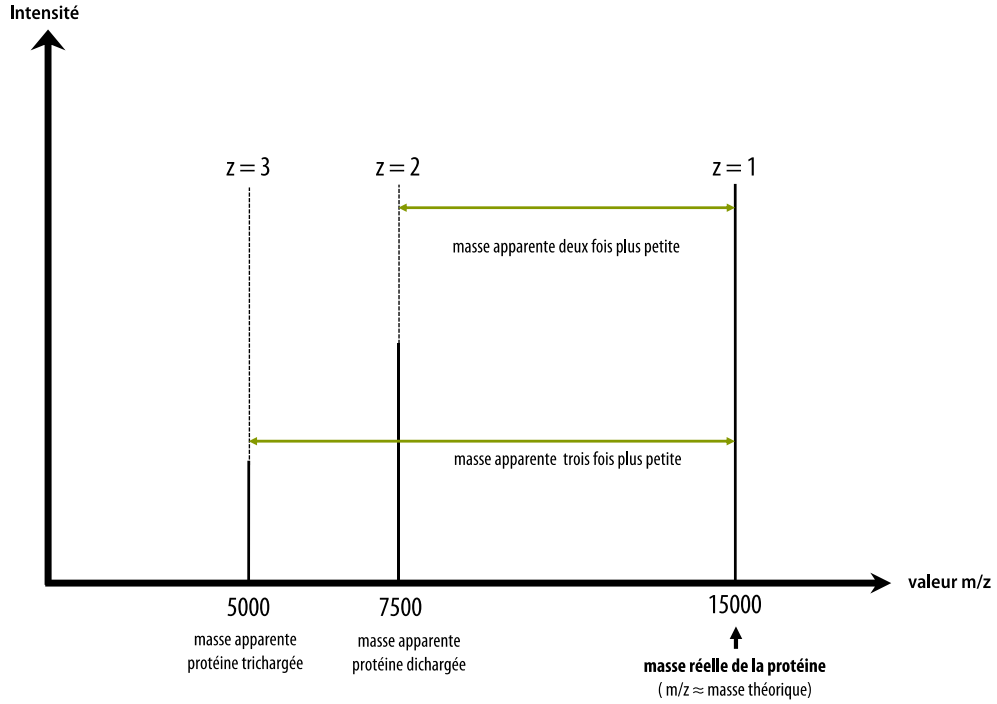


Figure 6 : Représentation schématique d'une protéine présente plusieurs fois dans un spectre sous forme d'échos (mono-, di-, trichargée)

m = masse ; z = charge

Les protéines identifiées dans ces spectres MALDI-TOF sont en en grande partie des protéines ribosomiques. Elles sont majoritaires dans les cellules en croissance (50% de la biomasse), et la plupart d'entre elles ont un faible poids moléculaire compatibles avec la gamme de masse visible en MALDI-TOF (2000 à 20000 Da). Elles possèdent également un pKi élevé (supérieur à 8) les rendant plus sensibles à l'ionisation. Parmi les autres catégories de protéines visibles, nous retrouvons celles qui se lient à l'ADN (*DNA binding protein subunit α and β*) ou encore des *cold shock proteins* (CspA, CspE, CspC) [206].

Les empreintes spectrales contiennent 3 types d'informations. La première est la simple absence/présence de pics. Cette information est particulièrement utile pour l'identification de micro-organismes en routine ou de comparaison globale de spectres. Un autre type d'information est la variation d'intensité de pics. Cette dernière est plutôt utilisée pour des analyses de résistance aux antibiotiques, car elle devient significative entre une colonie résistante et une colonie sensible. Cela est dû au fait que l'intensité observée est dépendante en partie de l'abondance des protéines, celle-ci étant elle-même dépendante de la quantité de bactéries présentes dans un échantillon. Enfin, le dernier type d'information, qui est peut-être le plus significatif, est le décalage de pic qui correspond à un changement de masse par rapport à une masse attendue (le pic décalé apparaît en aval ou en amont d'un pic de référence). Ce dernier aspect sera détaillé dans la partie IV.

Pour clore ce chapitre, il est important de comprendre que la valeur m/z lue sur une empreinte spectrale MALDI-TOF n'est qu'une caractéristique (simple) d'une protéine spécifique et devrait être considérée comme une « ombre » de cette protéine. Ainsi, un grand nombre de protéines possédant des fonctions et des séquences radicalement différentes peuvent théoriquement produire des valeurs m/z similaires voir identiques [207]. De plus, une valeur m/z propre à une protéine donnée peut changer en fonction de : son degré de modification post-traductionnelle, l'état des cofacteurs associés aux protéines ou encore le nombre de protons H^+ récupérés par la protéine durant l'étape de désorption/ionisation. Cela entraîne une variabilité de la position du pic attendu qui doit être prise en compte [180]. [180]. Le Tableau 2 ci-dessous résume les avantages et inconvénients du spectromètre de masse MALDI-TOF.

Tableau 2 : Récapitulatif des avantages et inconvénients de l'approche typage par spectrométrie MALDI-TOF

	Avantages	Inconvénients
Appareil	Coût faible d'une acquisition	Investissement de départ et entretien
Nombre d'isolats par pour une analyse	Haut débit (96 cibles)	n/a
Acquisition	Très rapide (de l'ordre de la dizaine de minutes)	Préparation de l'échantillon ; étape critique
Identification	Fiable, reproductible	Limitée par la qualité des bases de données
Typage	Reproductible	Nécessité d'identifier des biomarqueurs : cibles limitées Sensibilité moindre que d'autres approches comme la MLST
Nature des données	Biomarqueurs spécifiques	Précision Observation de masse apparente

II – Objectifs des travaux du projet doctoral

Les espèces appartenant au genre *Tenacibaculum* restent peu étudiées à ce jour, malgré leur impact croissant en pisciculture marine. Au moins 8 espèces sont pathogènes de poissons marins et toutes provoquent des symptômes similaires, rendant impossible le diagnostic visuel. Bien que l'effort de séquençage semble s'être intensifié ces 3 dernières années, les génomes de toutes les souches types du genre ne sont pas encore disponibles dans les bases de données. De fait, les outils d'identification utilisés en médecine vétérinaire restent rudimentaires et basés la plupart du temps sur une simple caractérisation morphologique des bactéries observées au microscope dans des prélèvements effectués au niveau des lésions externes. Il paraît alors évident qu'une meilleure gestion de ces maladies émergentes requiert une identification fiable, précise rapide et peu coûteuse.

Nous avons choisi la spectrométrie de masse MALDI-TOF comme outil privilégié pour ce projet. Celui-ci a donc pour but d'apporter de nouveaux outils pour l'étude des bactéries du genre en proposant des méthodes modernes d'identification et de caractérisation d'isolat. Nous avions donc pour objectif de séquencer une grande variété de génomes (l'ensemble des espèces du genre) ainsi qu'une collection de génomes appartenant à l'espèce-type du genre, *T. maritimum*. Cela permet de couvrir la diversité des espèces appartenant au genre ainsi que la diversité intra-spécifique de l'espèce *T. maritimum*. Premièrement, ces génomes sont intégrés à la plateforme d'annotation et d'analyse de génomes bactériens, MicroScope et servent de base solide pour soutenir notre étude. Ensuite, cela permet de s'assurer que les souches types de notre collection correspondent bien aux différentes espèces attendues, notamment en confrontant les séquences des rRNA 16S à celles associées aux articles décrivant ces mêmes espèces. Ces génomes sont également un atout majeur dans l'analyse des spectres MALDI-TOF ainsi que la construction de spectres de référence. En effet, nous avons constitué une base de données spectrale de référence pour l'ensemble des bactéries du genre *Tenacibaculum*. Comme c'est le cas pour de nombreux agents pathogènes responsables de maladies émergentes ou peu étudiées, il n'existait aucune base de données publique pour les espèces de ce genre. Enfin, ces génomes permettront de comprendre finement les données spectrales. L'apport des génomes pour l'analyse des empreintes spectrales est l'élément clé du projet doctoral. L'identification de la protéine associée à un biomarqueur est une étape cruciale pour la validation des méthodes de typage basées sur la spectrométrie de masse MALDI-TOF. Les données génomiques peuvent également confirmer le polymorphisme de la séquence d'acides aminés des biomarqueurs de typage entre sous-groupes d'isolats au sein d'une espèce. Le volet protéomique de ce projet est alors construit autour de méthodes de développement puis l'implémentation d'outils d'identification et de typage libre et gratuit.

Bien que le modèle choisi fût les bactéries du genre *Tenacibaculum*, les méthodes utilisées sont imaginées comme les plus génériques possibles. Elles pourront être transposées à d'autres espèces. Le dernier aspect du projet est l'accessibilité de ces nouveaux outils à la communauté scientifique. Les différents outils d'identification et de caractérisation d'isolats seront donc rendu accessible au travers d'une application web : MALDIquantTypeR. L'avantage des outils développés ici est qu'ils peuvent non seulement servir dans le cadre de recherche académique mais également servir comme méthodes innovantes pour effectuer un diagnostic de routine à partir de prélèvements issus de piscicultures.

III – Résultats

Partie 1 : Étude génomique des espèces appartenant au genre *Tenacibaculum*

Un des premiers objectifs du projet doctoral était de proposer une phylogénie solide, basée sur la totalité du génome, et intégrant la plupart des espèces du genre *Tenacibaculum*. En effet, les phylogénies publiées sont basées uniquement sur les séquences d'un gène particulier (e.g. rRNA 16S, gyrB) ou au mieux sur des données MLST (7 à 11 gènes de ménages). A ce jour, peu de génomes des espèces de *Tenacibaculum* ont été publiés ou sont disponibles dans les bases de données. Le séquençage des génomes bactériens du genre *Tenacibaculum* a débuté avant le début de ce projet mais a été renforcé dans le cadre de celui-ci. Cette analyse des génomes nous permet non seulement de proposer une phylogénie la plus pertinente possible, mais aussi – comme nous le verrons dans les deux derniers paragraphes – de vérifier l'affiliation taxonomique de l'ensemble des souches ayant des génomes publiés. Cette vérification est l'élément de base des analyses MALDI-TOF qui nous intéresseront dans les prochains chapitres. En effet, elle permet de s'assurer de l'intégrité de notre collection de souches afin de constituer une banque de données de spectres de référence pour le genre. C'est également important car ces génomes pourront servir de base à de futures analyses de génomique comparée. Tous les génomes séquencés dans le cadre de ce projet ont été assemblés puis déposer sur la plateforme d'annotation et d'analyse de génomes bactérien : MicroScope qui est disponible à l'adresse internet suivante (https://www.genoscope.cns.fr/agc/microscope/home/index.php) [208].

A – Phylogénie du genre

La phylogénie (Figures 7 et 8) proposée ici et réalisée au début de mon projet doctoral comprend 25 espèces décrites et regroupe 44 génomes, dont 40 séquencés par notre équipe. Celle-ci est basée sur le *core-genome* tel que calculé par MicroScope, en utilisant les paramètres suivants : 80% d'identité en acides aminés et 80% de couverture d'alignement. Les séquences obtenues (516 familles de gènes) ont été alignées par MUSCLE [209] à l'aide du package R *msa* [210]. Les alignements ainsi obtenus ont été vérifiés afin d'écarter ceux présentant des incohérences. Ainsi, seules les protéines codées par des gènes monocopies et dont l'alignement sur l'ensemble des 44 génomes était satisfaisant furent conservés. Au final, seuls quelques alignements ont été écartés. L'arbre fut obtenu en utilisant FastTreeMP (version 2.1.10 SSE3, OpenMP, 32 threads) [211], qui est une implémentation parallélisée de PhyML. La structure de la phylogénie est cohérente avec celle, plus partielle car comprenant moins d'espèces, qui fut obtenue il y a quelques années par Marie Touchon en utilisant 1000 protéines appartenant au core-génome (travaux non publiés). Nous pouvons également retrouver les 4 clades qu'elle identifia (clade I à IV). La totalité des nœuds en amont des nœuds correspondant aux espèces ont une valeur de fiabilité de 100% (*SH-like local support*). Ces valeurs très élevées suggèrent que la topologie de l'arbre obtenu est très probable. Nous pouvons observer différents éléments remarquables. Tout d'abord, l'espèce *T. maritimum* – l'espèce type du genre – est l'espèce la plus éloignée des autres espèces du genre. De plus, la divergence intra-clade est comparable pour les clades II à IV, mais elle est nettement plus faible dans le clade I qui est aussi celui qui regroupe le plus d'espèces. Cela pourrait être la conséquence d'un phénomène de diversification récent au sein de ce clade.

Certains génomes provenant d'espèces non identifiées ne se groupent pas avec des espèces décrites et se trouvent aux extrémités de branches de longueur notable. C'est le cas

des génomes des souches LPB0136, HZ1, MAR_2009_124 et TNO 020. Cette topologie d'arbre suggère que ces génomes pourraient appartenir à des espèces encore non décrites. Ces hypothèses formulées durant la deuxième année de mon projet sont confirmées *a posteriori* par les publications successives de 2 nouvelles descriptions d'espèces durant cette même année *T. todarodis* LPB0136 et *T. agarivorans* HZ1 [60, 62]. De plus la souche TNO020 correspond également à une souche type décrite sous le nom de *T. piscium* (A.B. Olsen, article soumis).

Nous pouvons également observer la très faible distance entre les génomes de *T. discolor* et le génome de la souche type de *T. ascidiaceicola*. Dans la partie suivante, nous compléterons ces observations pas des analyses d'*Average Nucleotide Identity*.

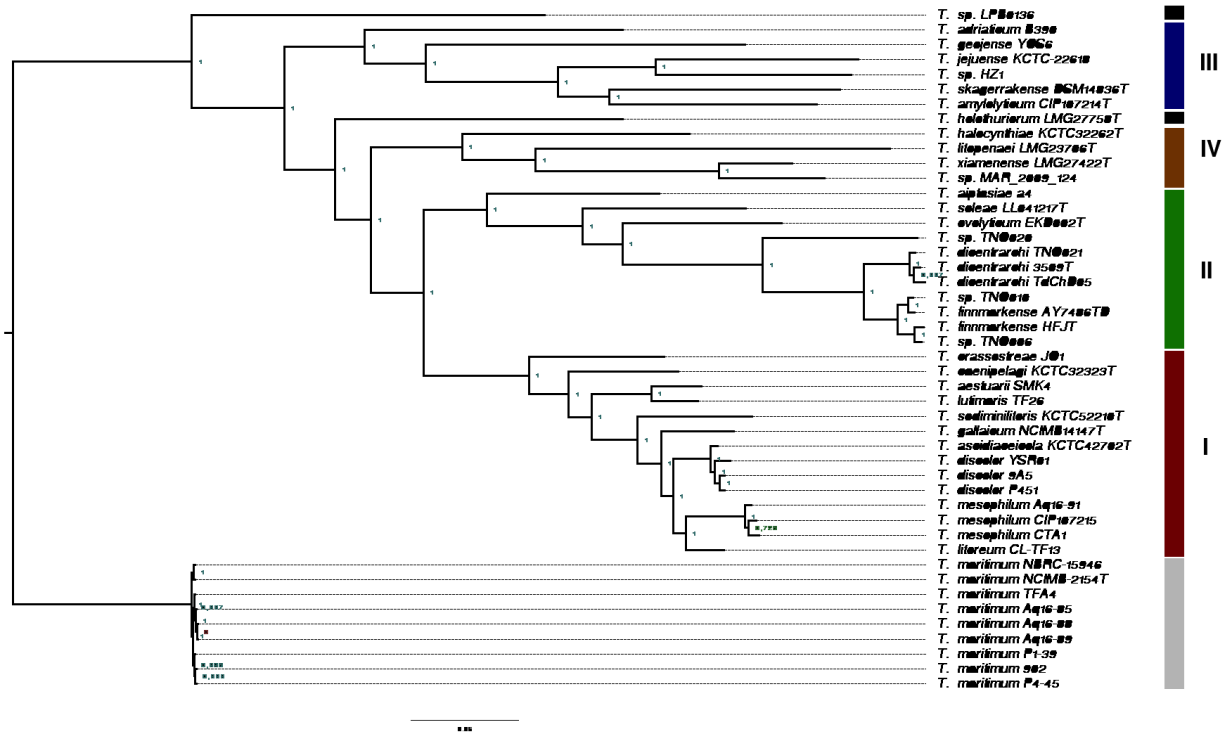


Figure 7: Phylogénie par maximum de vraisemblance des espèces appartenant au genre Tenacibaculum (raciné avec mid-point) ; FastTreeMP

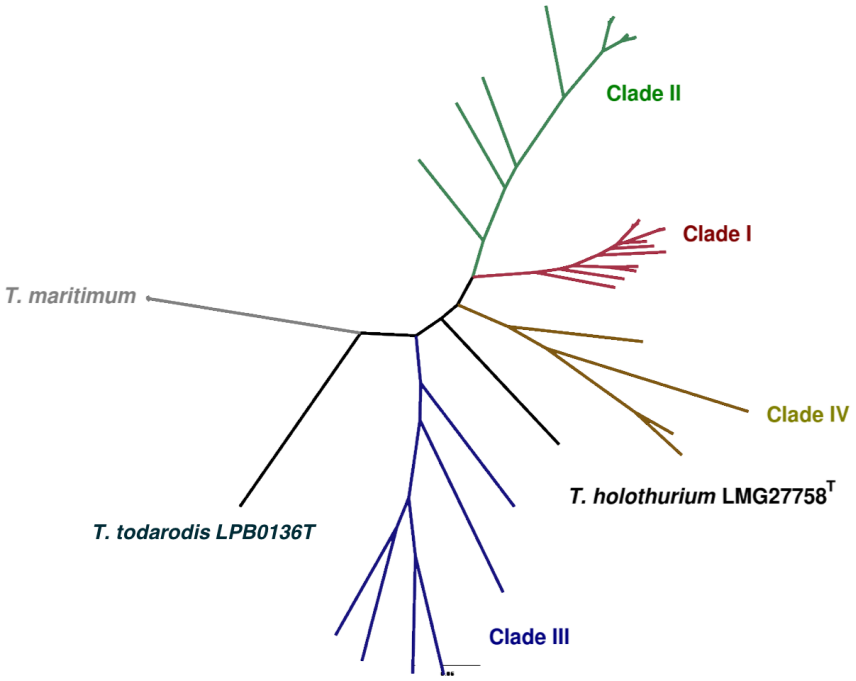


Figure 8 : Phylogénie par maximum de vraisemblance des espèces appartenant au genre Tenacibaculum (non-raciné) ; FastTree

B – Diversité des espèces du genre *Tenacibaculum* et *Average Nucleotide Identity*

Nous avons utilisé l'ANIm (MUMmer) implémentée dans pyani (résultats non présentés) ainsi que plus récemment son équivalent implémenté et disponible sur MicroScope (*Genome Clustering*) sur l'ensemble des génomes publiés sur le site du NCBI au moment de la rédaction de ce manuscrit de thèse (41 génomes) et ceux séquencés dans notre équipe, soit 82 génomes au total (Figure 9). Nous avons choisi de détailler ici les résultats obtenus par *Genome Clustering* afin de faciliter leur représentation dans ce manuscrit. Les conclusions tirées des valeurs d'ANI obtenues par pyani et celles par MicroScope sont rigoureusement identiques. Le *Genome Clustering* repose sur le calcul de distance Mash. C'est une estimation du nombre de k-mers identiques partagés entre 2 génomes. C'est également une approximation du taux de mutation [115]. La distance Mash est fortement corrélée à l'ANI ($D_{MASH} \approx$ 1-ANI). La méthode proposée par MicroScope définit des *MicroScope Genome Cluster* (MICGC) sur la base de cette distance. Un MICGC est un groupe de génomes dont la distance Mash est inférieure ou égale à 0.06. Cela équivaut approximativement à une ANI supérieure ou égale à 94%. Ainsi, un MICGC est une unité taxonomique correspondant généralement au rang de l'espèce. L'arbre de classification obtenue permet d'avoir un aperçu de la diversité présente au sein du genre.

Tout d'abord, et comme le suggérait l'arbre de la figure 7, la souche KCTC42702^T décrite comme une nouvelle espèce par Kim et al (*i.e., T. ascidiaceicola*) fait effectivement partie du cluster correspondant à l'espèce *T. discolor* (MICGC2129). De plus, les valeurs entre la souche de *T. ascidiaceicola* et les souches de *T. discolor* d'ANI (98%) et de GGDC (81.20%) sont nettement supérieurs au seuil d'appartenance à la même espèce (94-96% et 70%, respectivement) et la probabilité d'appartenir à la même espèce donnée par GGDC est de 91.44%. Ainsi, sur la base de leurs génomes, la souche type KCTC42702^T de l'espèce *T. ascidiaceicola* et celles identifiés comme appartenant à l'espèce *T. discolor* semblent appartenir à la même espèce. Ces résultats sont en désaccord avec ceux présentées dans l'article de Kim et al. Il faut néanmoins souligner que cet article contient des éléments discutables. En effet, ils montrent des caractéristiques phénotypiques différentes (*e.g.,* dégradation du Tween 80 ou l'activité de l'estérase C4) entre les espèces *T. ascidiaceicola* et *T. discolor*. Nous avons cependant vu dans le chapitre introductif que les caractéristiques phénotypiques sont insuffisantes pour identifier de manière fiable une espèce, notamment en raison de la difficulté d'interprétation et la variabilité d'expression des différentes voies métaboliques. En revanche, les auteurs ont utilisé la séquence de l'ARN 16S pour comparer la souche type de *T. ascidiaceicola* aux espèces proches et reconstruire un arbre phylogénétique. Bien que le pourcentage d'identité entre la souche type KCTC42702^T de l'espèce *T. ascidiaceicola* et la souche type de l'espèce *T. discolor* était de 99.5% (ce qui est normalement suffisant pour affirmer que les deux souches en question appartiennent à la même espèce), les auteurs ont tenu à décrire une nouvelle espèce. Le résultat majeur soutenant cette description d'espèce est le résultat de l'expérience d'hybridation ADN-ADN avec la souche KCTC42702^T : les valeurs DDH obtenues étaient de 25,2% avec *T. discolor* DSM 18842T, 17,9% avec T. *gallaicum* DSM 18841^T et 17,3% avec *T. litoreum* KCCM 42115^T. Sachant que les expériences d'hybridation ADN-ADN sont complexes, nous pouvons raisonnablement émettre l'hypothèse que ces résultats sont erronés. Cette hypothèse est la plus simple pouvant expliquer l'incompatibilité des résultats de génomiques et ceux issus des expériences d'hybridation ADN-ADN.

Nous pouvons observer un total de 32 MICGC distincts. Ces derniers correspondent aux 28 espèces formellement décrites et auquel s'ajouterait 6 espèces encore non décrites : *T.* sp. MAR_2009_124, *T.* sp. UBA9192 et UBA8975, *T.* sp. SZ-18, T. sp. Bg11-29, *T.* sp. TNO 020. Comme mentionné précédemment, la souche TNO020 corresponds à la souche type de l'espèce *T. piscium* qui est en cours de description par Olsen et al.

Nous pouvons également remarquer que le génome de la souche HSC 22, issue du NCBI (MICGC3069) et décrit comme appartenant à l'espèce *T. litoreum* (article non publié) ne fait pas partie du même cluster que la souche type TL CF13^{T} de l'espèce *T. litoreum* (MICGC2130). De plus, la valeur d'ANI calculée entre ces deux génomes est de 92.12%, légèrement en deçà du seuil de l'espèce (94-96%). Ces résultats suggèrent que les deux souches mentionnées ci-dessus pourraient appartenir à deux espèces distinctes.

Pour quatre souches, dont les drafts génomes ont été déposés au NCBI mais sans affiliation taxonomique (désignées *Tenacibaculum* sp.), et sur la base de cette analyse, nous pouvons proposer que : la souche *T.* sp. 4G03 soit affiliée à l'espèce *T. discolor*, *T.* sp. 47A_GOM-205m à l'espèce *T. lutimaris* et que les souches *T.* sp. E3R01 et *T.* sp. MAR_2010_89 soient affiliées à l'espèce *T. halocynthiae*.

Enfin, nous pouvons observer une autre contradiction dans cet arbre. En effet, les espèces *T. dicentrarchi* et *« T. finnmarkense »* font partis du même cluster (MICGC1022), ce qui suggère qu'il ne s'agit que d'une seule espèce. Pourtant, d'autres résultats ont soutenu la distinction de ces deux espèces. Ces résultats font l'objet d'un premier article qui fut publié en Janvier 2018. Dans la prochaine partie, nous discuterons en détail de cet article.

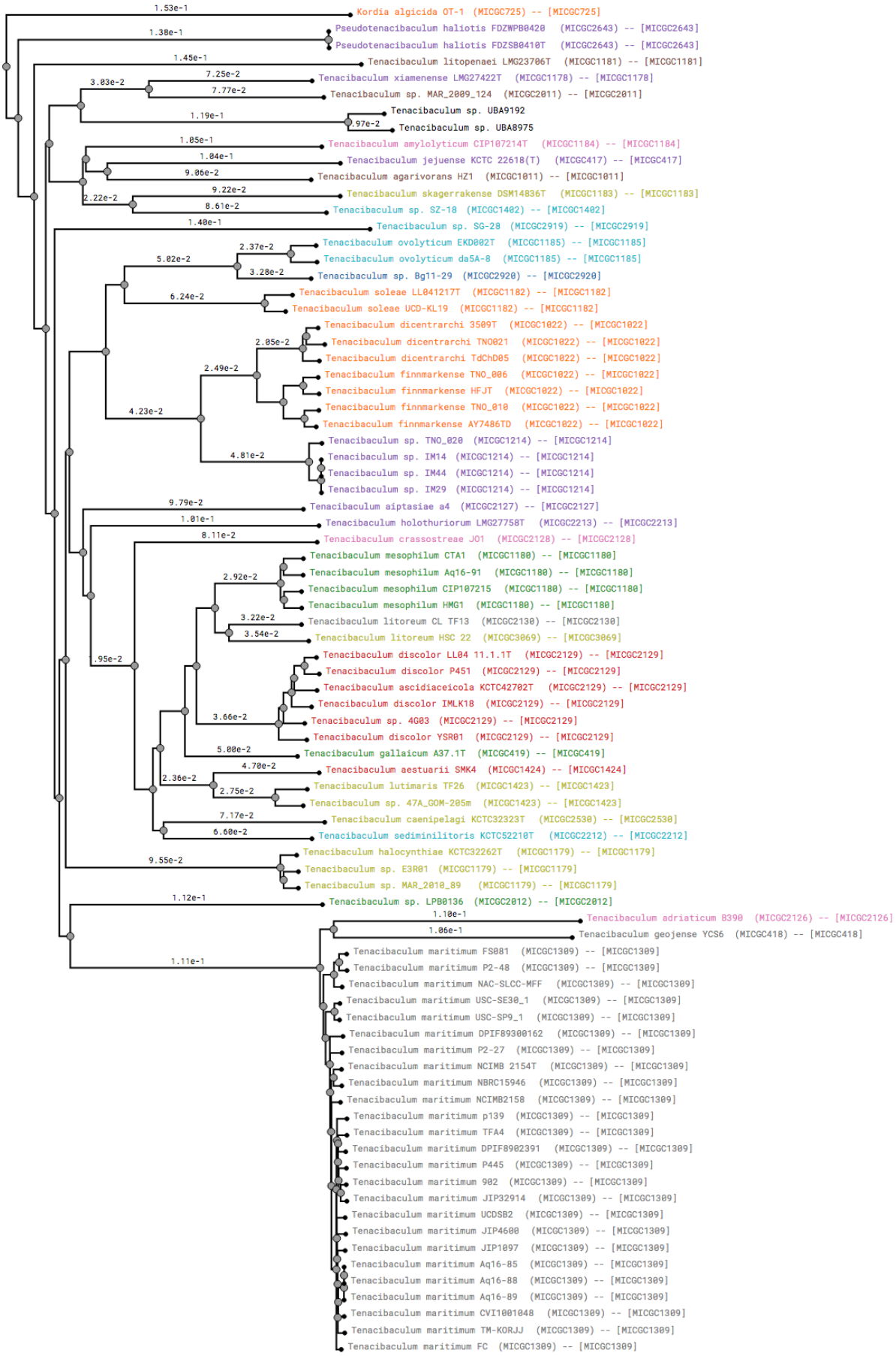


Figure 9 : "Genome clustering" de l'ensemble des génomes de Tenacibaculum intégrés dans MicroScope

C – Première publication : Étude comparée des génomes de *T. dicentrarchi*, « *T. finnmarkense* » et *T. piscium* (TNO020).

C1 – Introduction

Les espèces *T. dicentrarchi* et « *T. finnmarkense* » furent décrites en 2012 et en 2016. L'année suivante, Olsen et al. [22] ont montré une diversité insoupçonnée d'isolats appartenant au genre *Tenacibaculum* et pour la plupart appartenant à des espèces non identifiées. Ils ont notamment identifié 4 clusters importants (présentés dans la Figure 1 de l'article) correspondant à 4 espèces distinctes. Le cluster II correspond à l'espèce *T. dicentrarchi*. Le cluster I et III correspondent à 2 sous-espèces de l'espèce *« T. finnmarkense »* (description soumise pour publication). Le cluster IV correspond à l'espèce *T. piscium* (souche type TNO020). L'ensemble des isolats de cette étude (89 au total) était issus de poissons présentant des signes cliniques. Dans notre étude, nous avons réalisé une étude génomique comparée succincte. Nous avons séquencé 7 génomes dont les souches types des espèces *T. dicentrarchi*, « *T. finnmarkense* » et *T. piscium*. L'ANI (et méthodes équivalentes) permet de soutenir la distinction de ces 3 espèces malgré une distance évolutive relativement faible entre celles-ci. Nous avons également apporté un argument en faveur de l'hypothèse de Habib et al. (2014) qui suppose une évolution parallèle de la pathogénicité et une acquisition multiple des facteurs de virulences au sein du genre *Tenacibaculum*. En effet, ceux identifiés récemment chez *T. maritimum* par Pérez-Pascual et al. (2017) [212] n'ont pas été retrouvés dans les génomes de notre étude. En revanche, d'autres facteurs de virulence, codant pour des fonctions similaires, ont pu être identifiés.

C2 – Article

Comparative Genomics of *Tenacibaculum dicentrarchi* and "*Tenacibaculum finnmarkense*" Highlights Intricate Evolution of Fish-Pathogenic Species

Sébastien Bridel[1,2,3], Anne-Berit Olsen[4], Hanne Nilsen[4], Jean-François Bernardet[1], Guillaume Achaz[5], Ruben Avendaño-Herrera[6,7], and Eric Duchaud[1,*]

[1]VIM, INRA, Université Paris-Saclay, Jouy-en-Josas, France

[2]Labofarm, Finalab, Loudéac, France

[3]Université de Versailles Saint-Quentin-En-Yvelines, Montigny-Le-Bretonneux, France

[4]Norwegian Veterinary Institute, Bergen, Norway

[5]Atelier de Bioinformatique, UMR 7205 ISyEB, MNHN-UPMC-CNRS-EPHE, Muséum National d'Histoire Naturelle, Paris, France

[6]Laboratorio de Patología de Organismos Acuáticos y Biotecnología Acuícola, Departamento de Ciencias Biológicas, Facultad de Ciencias Biológicas, Universidad Andrés Bello, Viña del Mar, Chile

[7]Interdisciplinary Center for Aquaculture Research (INCAR), Concepción, Chile

*Corresponding author: E-mail: eric.duchaud@inra.fr.

Accepted: January 18, 2018

Data deposition: All genome sequencing data have been deposited at the European Nucleotide Archive under the accessions GCA_900239455, GCA_900239305, GCA_900239485, GCA_900239185, GCA_900239505, GCA_900239345, and GCA_900239495.

Abstract

The genus *Tenacibaculum* encompasses several species pathogenic for marine fish. *Tenacibaculum dicentrarchi* and "*Tenacibaculum finnmarkense*" (Quotation marks denote species that have not been validly named.) were retrieved from skin lesions of farmed fish such as European sea bass or Atlantic salmon. They cause a condition referred to as tenacibaculosis and severe outbreaks and important fish losses have been reported in Spanish, Norwegian, and Chilean marine farms. We report here the draft genomes of the *T. dicentrarchi* and "*T. finnmarkense*" type strains. These genomes were compared with draft genomes from field isolates retrieved from Chile and Norway and with previously published *Tenacibaculum* genomes. We used Average Nucleotide Identity and core genome-based phylogeny as a proxy index for species boundary delineation. This work highlights evolution of closely related fish-pathogenic species and suggests that homologous recombination likely contributes to genome evolution. It also corrects the species affiliation of strain AYD7486TD claimed by Grothusen et al. (2016).

Key words: *Tenacibaculum*, tenacibaculosis, genomes, fish pathogens, virulence, evolution.

Introduction

The rapid development of intensive aquaculture has been associated with a dramatic increase in outbreaks of infectious diseases (FAO 2016; Bayliss et al. 2017). The rapid international spread of pathogens through the trade of fish and eggs or as a response to environmental changes has been documented (Brynildsrud et al. 2014; Rahmati-Holasoo et al. 2016). In this context, the success and sustainability of aquaculture largely depend on the control of pathogens. Among those, several species of the genus *Tenacibaculum* are responsible for diseases collectively designated tenacibaculosis (Avendaño-Herrera et al. 2006; Suzuki 2015). *Tenacibaculum dicentrarchi* and "*Tenacibaculum finnmarkense*" are two among those fish-associated, recently described species. The former was first isolated from European sea bass (*Dicentrarchus labrax*) in Spain (Piñeiro-Vidal et al. 2012) and recently also identified from Atlantic salmon (*Salmo salar*) in Chile (Avendaño-Herrera et al. 2016)

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

and Norway (Olsen et al. 2017) and from red conger eel (*Genypterus chilensis*) in Chile (Irgang et al. 2017). "*Tenacibaculum finnmarkense*" was isolated from Atlantic salmon with ulcerative disease in Norway (Småge et al. 2016).

Identification of the causative agent of tenacibaculosis was first based on the isolation of bacteria from tissues of diseased fish and their characterization by phenotypic, biochemical, and serological methods (Piñeiro-Vidal, Carballas et al. 2008; Piñeiro-Vidal, Riaza et al. 2008). The use of 16 S rDNA sequencing improved the identification reliability (Cepeda et al. 2003; Fringuelli et al. 2012). However, these methods usually cannot differentiate closely related bacterial species. MLST was developed (Habib et al. 2014) and used to demonstrate the presence of *T. dicentrarchi* in Chile (Avendaño-Herrera et al. 2016) and to reveal the variety of *Tenacibaculum* spp. in a number of sea-farmed fish species in Norway (Olsen et al. 2017).

In this study, we present seven draft genomes of *Tenacibaculum* strains, including the *T. dicentrarchi* (USC 3509^T) and "*T. finnmarkense*" (HFJ^T) type strains, as well as five field isolates from Chile and Norway selected on the basis of available MLST data (Olsen et al. 2017). Comparison has been performed with available *Tenacibaculum* genomes from Genbank, including strain AYD7486TD originally described as *T. dicentrarchi* (Grothusen et al. 2016). We used Average Nucleotide Identity (ANI) to delineate species boundaries and core genome analysis to infer phylogenetic relationships between these strains. We also correct the species affiliation of strain AYD7486TD.

Materials and Methods

Bacterial Strains

The *T. dicentrarchi* type strain USC 3509^T (Piñeiro-Vidal et al. 2012) was obtained from Dr Y. Santos (University of Santiago de Compostela, Spain) and the "*T. finnmarkense*" strain HFJ^T (Småge et al. 2016) was obtained from Dr H. Duesund (Cermaq Group AS, Bergen, Norway). Strains TNO006, TNO010, and TNO020 were isolated from skin ulcers of Atlantic salmon in Norway whereas strain TNO021 was isolated from mouth ulcer of a corkwing-wrasse (*Symphodus melops*) also in Norway (Habib et al. 2014; Olsen et al. 2017). Strain TdChD05 was retrieved from external lesion of an Atlantic salmon in Chile (Avendaño-Herrera et al. 2016). All strains were routinely grown on marine agar 2216 (Difco) and in the corresponding broth at 170 rpm and 15 °C (TNO006) or 22 °C (all other strains).

Genome Sequencing and Annotation

Genomic DNA was extracted with the Wizard Genomic DNA Purification Kit (PROMEGA). All strains were sequenced with Illumina (HiSeq 2x100 pair-end reads with 300 bp insert size). Sequencing reads were assembled using Velvet (Zerbino and Birney 2008) or SPAdes (Bankevich et al. 2012). Genome annotation, including manual curation, and genome comparison including core genome computation were performed using the MicroScope platform (Médigue et al. 2017 and references therein).

ANI and Phylogenetic Reconstruction

ANIs analyses were performed using the ANIm method described by Richter and Rosselló-Móra (2009) and implemented in the Python module Pyani (https://github.com/widdowquinn/pyani). Digital DNA–DNA hybridization was performed using the GGDC website (http://ggdc.dsmz.de/) and Formula 2 (Auch et al. 2010). Logistic regression was used for reporting the probabilities that DDH is $\geq$70% and thus accounting for bacteria belonging to the same species. Pairwise alignments were computed using the MUMer software (Kurtz et al. 2004). For phylogenetic reconstruction, comparison of the gene content between strains was done by pairwise proteome similarity search using BlastP Bidirectional Best Hit and the MicroScope default parameters (i.e., >80% protein identity, >80% coverage). A set of 895 groups of orthologous proteins was retained and multiple alignments on individual orthologous proteins were performed using MUSCLE (Edgar 2004) implemented in the msa R package (Bioconductor). The resulting alignments were manually checked and concatenated for tree reconstruction using UGENE and PhyML with default parameters (Okonechnikov et al. 2012). Tree rendering was achieved using the Figtree software (http://tree.bio.ed.ac.uk/software/figtree/). The neighbor-net analysis in the Splits Tree 4 software (http://splitstree.org/; Huson and Bryant 2006) provided a phylogenetic network representing possible evolutionary relationships between the concatenated sequences of core genome genes. Minimal recombination breakpoints were identified using the four-gamete test (Hudson and Kaplan 1985). Putative recombination events were indicated as pairwise homoplasy index (PHI; Bruen et al. 2006) calculated by Splits Tree 4.

Results and Discussion

General Genome Features

A summary of the genomes analyzed in this study is presented in supplementary table 1, Supplementary Material online. Strikingly, the sizes of the genomes reported here are the smallest among those available to date for the genus *Tenacibaculum* (e.g., *T. maritimum*: 3.4 Mb, *T. soleae*: 3.0 Mb, *T. ovolyticum*: 4.1 Mb). The average genome size is 2.7 Mb and 2.9 Mb for *T. dicentrarchi* and "*T. finnmarkense*," respectively; at 2.4 Mb, strain TNO020 has the smallest genome. All strains studied are devoid of plasmid.

Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

ANI Delineates *T. dicentrarchi* and "*T. finnmarkense*" and Allocates Strain AY7486TD to "*T. finnmarkense*"

ANI was reported to be an accurate and practical method for species delineation and a 95–96% identity was proposed as the threshold (Rodriguez-R and Konstantinidis 2014). ANI comparisons were computed using *T. ovolyticum* (Suzuki et al. 2001) as an outgroup and the results were plotted as a heatmap (supplementary fig. 1A and 1B, Supplementary Material online). Strains TdChD05 from Chile and TNO021 from Norway, both previously allocated to the species *T. dicentrarchi* (Piñeiro-Vidal et al. 2012; Olsen et al. 2017; Avendaño-Herrera et al. 2016), indeed form a highly cohesive group with *T. dicentrarchi* USC 3509^T (98% ANI and ≥88% alignment coverage). In contrast, strain AY7486TD, also originally described as *T. dicentrarchi* (Grothusen et al. 2016), displays an ANI value of only 93% with the type strain of this species and consequently does not fall within the *T. dicentrarchi* cluster. Instead, strain AYD7486TD forms a cluster with strains TNO006, TNO010 and "*T. finnmarkense*" HFJ^T (≥96% ANI and ≥85% alignment coverage, above the species delineation threshold). ANI values delineate two subclusters, one grouping strains TNO010 and AYD7486TD (99% ANI) and the other grouping strain TNO006 and "*T. finnmarkense*" HFJ^T (98% ANI). Supplementary figure 1, Supplementary Material online, also shows that although *T. dicentrarchi* and "*T. finnmarkense*" are distinct species they obviously display significant proximity in terms of sequences identity (93–94% ANI) and fraction of shared genomes (77–88%). Strain TNO020 does not belong to any of the previously defined clusters (88–89% ANI; 55–68% alignment coverage) and therefore likely belongs to a yet undescribed *Tenacibaculum* species as previously suggested (Habib et al. 2014; Olsen et al. 2017). The same conclusions (supplementary fig. 1C, Supplementary Material online) were drawn using genome-to-genome distance calculator (Auch et al. 2010). As expected, *T. ovolyticum* DSM 18103^T behaves as an outgroup displaying lower sequence identity (85%) and poor alignment coverage (17–34%) with all other strains studied. Using 895 core genome-encoding protein sequences, we constructed a phylogenetic tree from the concatenation of each individual alignment (fig. 1). Bootstrap values strongly support the division between *T. dicentrarchi* and "*T. finnmarkense*" and the core genome-based phylogenetic tree perfectly matches the ANI dendogram. Furthermore, a correlation between the MLST clades defined by Olsen et al. (2017) and the clusters observed in figure 1 is obvious: "*T. finnmarkense*" HFJ^T and strain TNO006 belong to clade III, strains TNO010 and AYD7486TD to clade I, all three *T. dicentrarchi* strains to clade II and strain TNO020 to clade IV. Using Splits Tree analysis, a reticulated network structure between the four clades, indicative of within and between species recombination events (fig. 2), is observed. The dense network joining clade I and clade III is in good agreement with the grouping of "*T. finnmarkense*" strains in two connected subclusters as previously observed in the MLST data set of Olsen et al. (2017). Whatever the group of strains considered (i.e., clades I and III strains, clades I, II, and III strains as well as clades I, II, III, and IV strains), the PHI test *P*-value is 0.0, indicating significant evidence of recombination. In addition, the high number of recombination breakpoints found in *T. dicentrarchi* and "*T. finnmarkense*" genomes (10 per kilobase in average) is another clue suggesting recombination events in these species.

Comparative Genomics Highlights Intricate Relationships between *T. dicentrarchi* and "*T. finnmarkense*"

The average number of predicted CDS for *T. dicentrarchi* and "*T. finnmarkense*" strains is 2,381 and 2,536, respectively, which is in good agreement with the observed genome sizes. The core genome is composed of 2,013 and 1,947 CDS for *T. dicentrarchi* and "*T. finnmarkense*," respectively (supplementary fig. 2A and B, Supplementary Material online). These small gene sets, about half those of *Tenacibaculum agarivorans* HZ1^T (Xu et al. 2017) and *Tenacibaculum jejuense* KCTC 22618^T (Ficko-Blean 2017), seem related to a deficient biopolymer-degrading ability of *T. dicentrarchi* and "*T. finnmarkense*." Indeed, these genomes lack the pathways encoding for the degradation of marine carbohydrates (e.g., sulfatase, glycoside hydrolase, polysaccharide lyase) identified in the environmental species *T. agarivorans* and *T. jejuense*, in line with a restricted ecological niche (i.e., fish tissues) and an exclusive protein-based predicted regimen for these pathogenic species. Moreover, the presence of insertion sequences or their scars as well as genes remnants in *T. dicentrarchi*, "*T. finnmarkense*" and strain TNO020 argues for genome reduction trends in contrast to the horizontal transfer genes in *T. jejuense* and *T. agarivorans*. These findings support the expected small genome size of bacterial pathogens compared with their nonpathogenic relatives (Weinert and Welch 2017). The core genome of the seven strains belonging to both *T. dicentrarchi* and "*T. finnmarkense*" is composed of 1,818 CDS (supplementary fig. 2C, Supplementary Material online), a value close to those computed for each species. Each strain has ~180 specific genes essentially composed of prophages remnants, restriction/modification systems, toxin/antitoxin systems and transposases encoding genes or their scars as well as genes required for the biosynthesis of exopolysaccharides that likely account for the minor intraspecies genome size differences previously mentioned. These strain-specific genes, representing the accessory genome, do not seem linked to bacterial pathogenicity as no *bona fide* toxin or virulence factor-encoding genes have been identified in this gene pool. Therefore, virulence-encoding genes likely belong to the core genome common to both *T. dicentrarchi* and "*T. finnmarkense*." Among those, peptidases containing a

Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

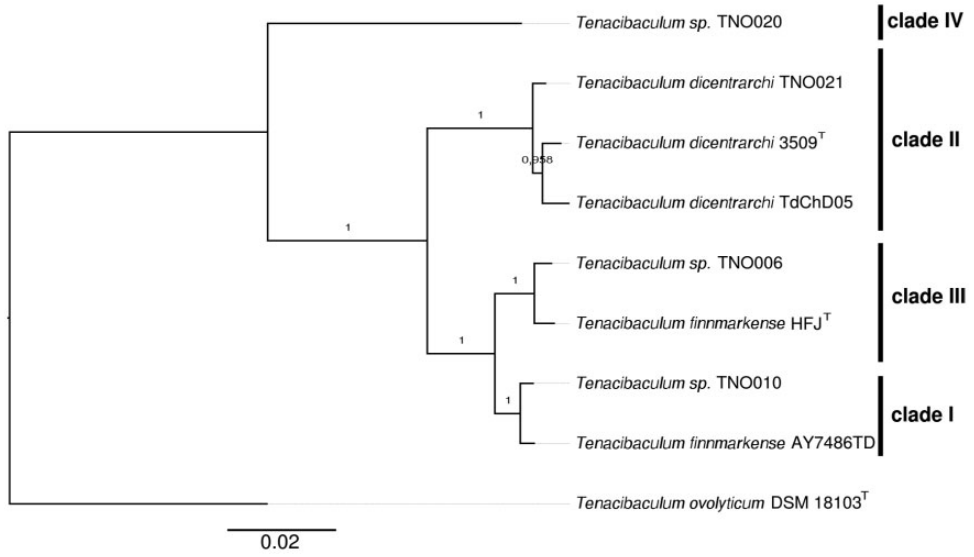


Fig. 1.—Core genome phylogeny. Phylogenetic tree inferred by PhyML with boostraping (100 replicates) using the concatenation of the 895 aligned orthologous genes. The MLST clades I–IV defined in Olsen et al. (2017) are reported.

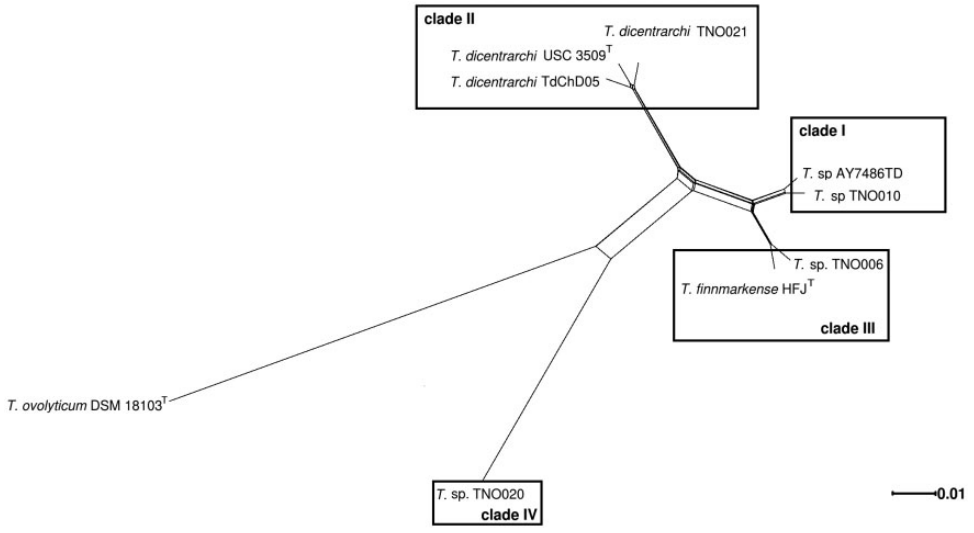


Fig. 2.—Detection of recombination. Reticulate evolutionary relationship between concatenated sequences of the 895 core genome genes visualized by the Splits Tree neighbor-net analysis. The clades defined in Olsen et al. (2017) are reported.

carboxy-terminal protein domain (TIGR04183), predicted to be required for T9SS-mediated secretion and cell surface exposure (Veith et al. 2013), were identified. These peptidases (i.e., *TFINN_2500013*, *TFINN_140038*, and *TFINN_60057* from "*T. finnmarkense*" HFJ^T and their orthologs) are likely involved in the breakdown of proteinaceous compounds and the destruction of host tissues. The presence of a M9 family protease-encoding gene (*TFINN_140038* and orthologs), similar to the 120 kDa collagenase of *Clostridium perfringens* (Matsushita et al. 1994) but different from the M43 family collagenase (encoded by *MARIT_1085*) identified in *T. maritimum* (Pérez-Pascual et al. 2017), suggests convergent evolution for some virulence-linked functions in fish-pathogenic *Tenacibaculum* species.

Conclusion

Since the pioneering work of Wakabayashi and col. on *T. maritimum* in the eighties (Wakabayashi et al. 1986), many other *Tenacibaculum* species have been described. Some of them are important fish pathogens and an unexpected diversity at different levels (e.g., genetic, fish host, geographical) has been reported (Habib et al. 2014; Olsen et al. 2017). *Tenacibaculum dicentrarchi* strains were previously identified

Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

from several farmed fish species in Spain, Norway and Chile, whereas "*T. finnmarkense*" strains were exclusively isolated in Norway so far (Olsen et al. 2017). Thanks to genomes comparison, we were able to correct the affiliation of strain AYD7486TD which actually belongs to the species "*T. finnmarkense*" rather than to *T. dicentrarchi* as previously claimed (Grothusen et al. 2016). Importantly, this result demonstrates that "*T. finnmarkense*" is also present in Chilean fish farms. Our data set suggests that *T. dicentrarchi* strains form a cohesive group whereas "*T. finnmarkense*" strains are split into two subclusters. Similar subclusters, referred to as genomovars, were reported in *Flavobacterium columnare*, another fish pathogen of the family *Flavobacteriaceae* with a broad host range. Correlations between genomovars, fish hosts and virulence have been suggested (Evenhuis and LaFrentz 2016; Olivares-Fuster et al. 2007). Hence, the same type of genomic heterogeneity observed in "*T. finnmarkense*" may account for specific traits such as host specificity or level of virulence. Strikingly, the virulence factors described in the closely related species *T. maritimum* (i.e., a sphingomyelinase, a ceramidase, a chondroitin AC lyase, a sialidase and a M43 family collagenase; Pérez-Pascual et al. 2017) have not been identified in any of the genomes described in the present study. In full agreement with this observation, a parallel evolution of pathogenicity in the species encompassed in the genus *Tenacibaculum* has been proposed (Habib et al. 2014). However, the grouping of *T. dicentrarchi* and "*T. finnmarkense*" as well as *T. soleae* and *T. ovolyticum* (Habib et al. 2014; Olsen et al. 2017; Småge et al. 2016) in a single clade exclusively encompassing fish-associated bacteria suggests evolution of these four species from a pathogenic ancestor. In addition, our data support recombination as an important force shaping genome evolution of these fish pathogens as previously observed in other members of the family *Flavobacteriaceae* (Nicolas et al. 2008; Vos and Didelot 2009). The genome sequences reported here should therefore facilitate future epidemiological studies and provide new insights into pathogenicity and niche adaptation of emerging fish pathogens.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This study was funded by the EU EMIDA ERA-NET project "Control Flavobacteriaceae Infections in European Fish farms" and by the Agence Nationale pour la Recherche (contract ANR-14-CE19-0020). This work has benefited from the facilities and expertise of the high-throughput sequencing platform of I2BC (Centre de Recherche de Gif; http://www.i2bc.paris-saclay.fr/spip.php? article399&lang=en). We also wish to thank the following structures for providing computational resources: The INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr), the LABGeM, and the National Infrastructure "France Génomique" funded as part of the Investissement d'avenir program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). A.-H. acknowledges Grant FONDECYT 1150695 and the CONICYT/FONDAP/15110027 from the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT, Chile). The authors are very grateful to Sophie Pasek and Mathilde Carpentier for fruitful discussion.

Literature Cited

Auch AF, Klenk H-P, Göker M. 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci. 2(1):142–148.

Avendaño-Herrera R, et al. 2016. Isolation, characterization and virulence potential of *Tenacibaculum dicentrarchi* in salmonid cultures in Chile. Transbound Emerg Dis. 63(2):121–126.

Avendaño-Herrera R, Toranzo AE, Magariños B. 2006. Tenacibaculosis infection in marine fish caused by *Tenacibaculum maritimum*: a review. Dis Aquat Organ. 71(3):255–266.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bayliss SC, et al. 2017. The promise of whole genome pathogen sequencing for the molecular epidemiology of emerging aquaculture pathogens. Front Microbiol. 8:121.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172(4):2665–2681.

Brynildsrud O, et al. 2014. Microevolution of *Renibacterium salmoninarum*: evidence for intercontinental dissemination associated with fish movements. ISME J. 8(4):746–756.

Cepeda C, García-Márquez S, Santos Y. 2003. Detection of *Flexibacter maritimus* in fish tissue using nested PCR amplification. J Fish Dis. 26(2):65–70.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Evenhuis JP, LaFrentz BR. 2016. Virulence of *Flavobacterium columnare* genomovars in rainbow trout *Oncorhynchus mykiss*. Dis Aquat Organ. 120(3):217–224.

FAO. 2016. The State of World Fisheries and Aquaculture 2016. Available from: http://www.fao.org/3/a-i5555e.pdf.

Ficko-Blean E, et al. 2017. Carrageenan catabolism is conferred by a complex regulon in marine heterotrophic bacteria. Nat Commun (forthcoming)

Fringuelli E, et al. 2012. Development of a quantitative real-time PCR for the detection of *Tenacibaculum maritimum* and its application to field samples. J Fish Dis. 35(8):579–590.

Grothusen H, et al. 2016. First complete genome sequence of *Tenacibaculum dicentrarchi*, an emerging bacterial pathogen of salmonids. Genome Announc. 4(1):e01756–15.

Habib C, et al. 2014. Multilocus sequence analysis of the marine bacterial genus *Tenacibaculum* suggests parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. Appl Environ Microbiol. 80(17):5503–5514.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111(1):147–164.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.

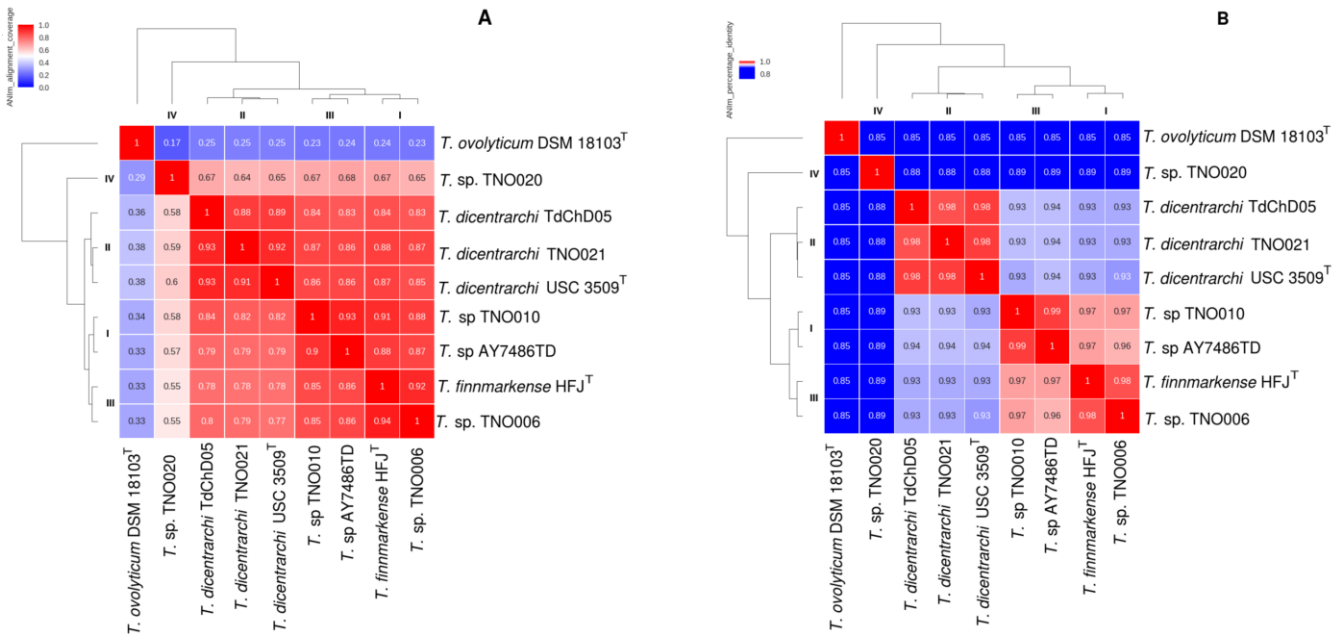
Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

Irgang R, et al. 2017. First identification and characterization of *Tenacibaculum dicentrarchi*isolated from Chilean red conger eel (*Genypterus chilensis*, Guichenot 1848). J Fish Dis. 40(12):1915–1920.
Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2):R12.
Matsushita O, Yoshihara K, Katayama S, Minami J, Okabe A. 1994. Purification and characterization of *Clostridium perfringens* 120-kilodalton collagenase and nucleotide sequence of the corresponding gene. J Bacteriol. 176(1):149–156.
Médigue C, et al. 2017. MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. Brief Bioinformatics bbx113, https://doi.org/10.1093/bib/bbx113.
Nicolas P, et al. 2008. Population structure of the fish-pathogenic bacterium Flavobacterium psychrophilum. Appl Environ Microbiol. 74(12):3702–3709.
Okonechnikov K, Golosova O, Fursov M. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28(8):1166–1167.
Olivares-Fuster O, et al. 2007. Host-specific association between *Flavobacterium columnare* genomovars and fish species. Syst Appl Microbiol. 30(8):624–633.
Olsen AB, et al. 2017. Multilocus sequence analysis reveals extensive genetic variety within *Tenacibaculum* spp. associated with ulcers in sea-farmed fish in Norway. Vet Microbiol. 205:39–45.
Olsen AB, et al. 2011. *Tenacibaculum* sp. associated with winter ulcers in sea-reared Atlantic salmon *Salmo salar*. Dis Aquat Organ. 94(3):189–199.
Pérez-Pascual D, et al. 2017. The complete genome sequence of the fish pathogen *Tenacibaculum maritimum* provides insights into virulence mechanisms. Front Microbiol. 8:1542.
Piñeiro-Vidal M, Carballas CG, Gómez-Barreiro O, Riaza A, Santos Y. 2008. *Tenacibaculum soleae* sp. nov., isolated from diseased sole (*Solea senegalensis* Kaup). Int J Syst Evol Microbiol. 58(Pt 4):881–885.
Piñeiro-Vidal M, Gijón D, Zarza C, Santos Y. 2012. *Tenacibaculum dicentrarchi* sp. nov., a marine bacterium of the family *Flavobacteriaceae* isolated from European sea bass. Int J Syst Evol Microbiol. 62(Pt 2):425–429.
Piñeiro-Vidal M, Riaza A, Santos Y. 2008. *Tenacibaculum discolor* sp. nov. and *Tenacibaculum gallaicum* sp. nov., isolated from sole (*Solea senegalensis*) and turbot (*Psetta maxima*) culture systems. Int J Syst Evol Microbiol. 58(Pt 1):21–25.
Rahmati-Holasoo H, et al. 2016. First detection of koi herpesvirus from koi, *Cyprinus carpio* L. experiencing mass mortalities in Iran: clinical, histopathological and molecular study. J Fish Dis. 39(10):1153–1163.
Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106(45):19126–19131.
Rodriguez-R LM, Konstantinidis KT. 2014. Bypassing cultivation to identify bacterial species. Microbe 9(3):111–118.
Småge SB, et al. 2016. *Tenacibaculum finnmarkense* sp. nov., a fish pathogenic bacterium of the family *Flavobacteriaceae* isolated from Atlantic salmon. Antonie Van Leeuwenhoek. 109(2):273–285.
Suzuki M. 2015. *Tenacibaculum*. In: Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, Hedlund B, Dedysh S, editors. Bergey's manual of systematics of archaea and bacteria. Chichester (United Kingdom): John Wiley & Sons, Ltd. p. 1–7. Available from: http://doi.wiley.com/10.1002/9781118960608.gbm00345 (accessed 28.08.17).
Suzuki M, Nakagawa Y, Harayama S, Yamamoto S. 2001. Phylogenetic analysis and taxonomic study of marine Cytophaga-like bacteria: proposal for *Tenacibaculum* gen. nov. with *Tenacibaculum maritimum* comb. nov. and *Tenacibaculum ovolyticum* comb. nov., and description of *Tenacibaculum mesophilum* sp. nov. and *Tenacibaculum amylolyticum* sp. nov. Int J Syst Evol Microbiol. 51(5):1639–1652.
Veith PD, et al. 2013. Protein substrates of a novel secretion system are numerous in the *Bacteroidetes* phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification, and cell-surface attachment. J Proteome Res. 12(10):4449–4461.
Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3(2):199–208.
Wakabayashi H, Hikida M, Masumura K. 1986. *Flexibacter maritimus* sp. nov., a pathogen of marine fishes. Int J Syst Evol Microbiol. 36(3):396–398.
Weinert LA, Welch JJ. 2017. Why might bacterial pathogens have small genomes? Trends Ecol Evol. 32(12):936–947.
Xu Z-X, Yu P, Mu D-S, Liu Y, Du Z-J. 2017. *Tenacibaculum agarivorans* sp. nov., an agar-degrading bacterium isolated from marine alga *Porphyra yezoensis* Ueda. Int J Syst Evol Microbiol. 67(12):5139–5143.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.

Associate editor: Howard Ochman

Downloaded from https://academic.oup.com/gbe/article-abstract/10/2/452/4817510 by INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement user on 14 February 2020

C3 – Supplementary materials



C

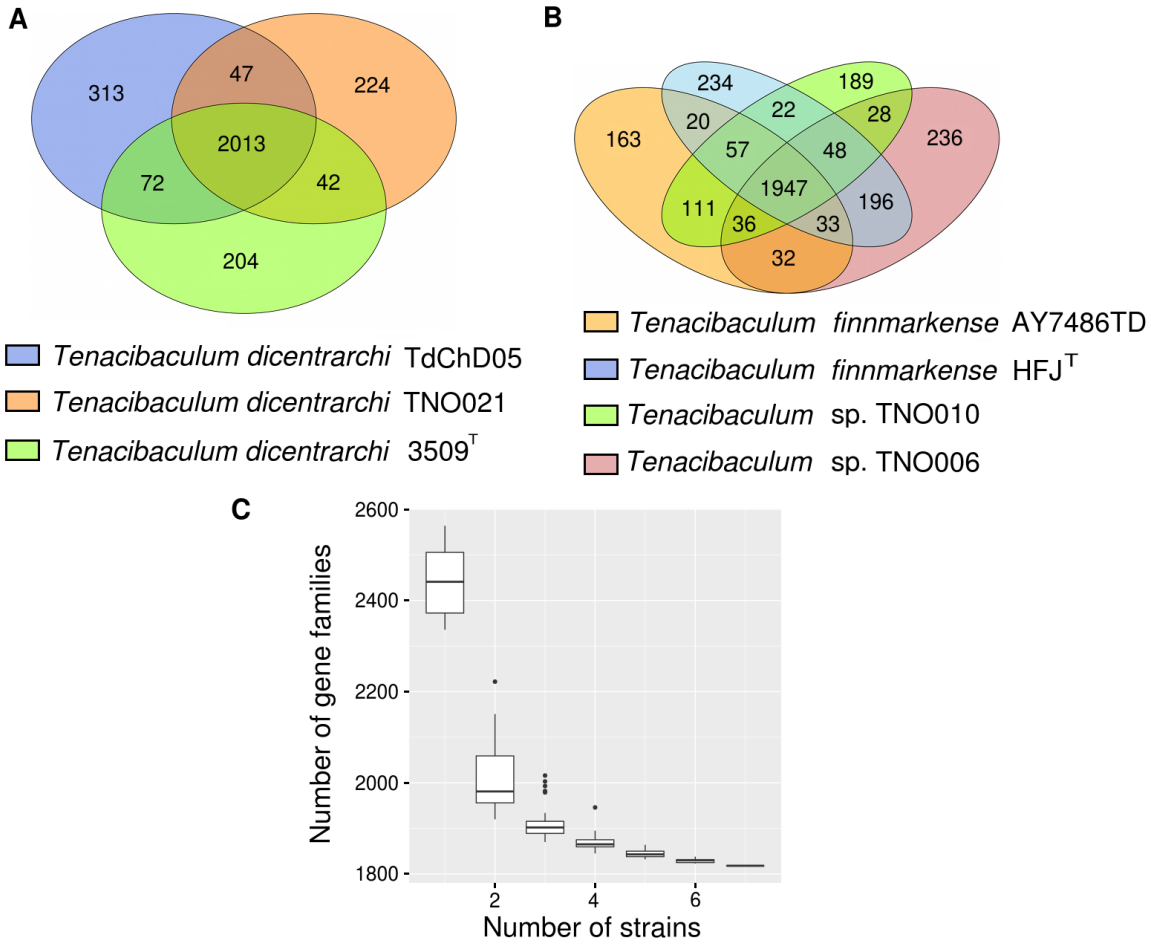
Strains	HFJ^T	TNO006	AY7486TD	TNO010	USC 3509^T	TNO021	TDCHD05	TNO020
HFJ^T	100	94.13	76.88	78.62	16.97	18.4	16.97	0.22
TNO006	-	100	77.02	78.62	15.48	18.12	16.07	0.18
AY7486TD	-	-	100	94.61	22.1	23.17	20.9	0.26
TNO010	-	-	-	100	19.53	20.21	19.33	0.23
USC 3509^T	-	-	-	-	100	94.43	94.04	0.13
TNO021	-	-	-	-	-	100	93.21	0.11
TDCHD05	-	-	-	-	-	-	100	0.14
TNO020	-	-	-	-	-	-	-	100

Tenacibaculum finnmarkense

Tenacibaculum dicentrarchi

Tenacibaculum sp.

SUPLEMENTAL FIGURE 1| ANIm results.
Heatmaps of A) Fraction of shared genomes. Blue gradient corresponds to value below 50% and red gradient to values over 50%. The dendogram and matching red squares correspond to MLST clades as described in Olsen et al. 2017, B) Average nucleotide identity. Red gradient is for values greater than 95%. Blue gradient is for values between 70 and 95% and grey color is for any value below 70% and C) Probability that two strains belong to the same species using GGDC analysis.



SUPLEMENTAL FIGURE 2| Core genome size.
Venn diagram of core/pan genomes for A) *T. dicentrarchi* strains and B) "*T. finnmarkense*" strains. C) Core-genome size evolution according to the number of *T. dicentrarchi* and "*T. finnmarkense*" genomes considered.

Table 1 : General genome features

Species	Strain	Country	Host	Date of isolation	Contigs	Total length	GC %	Predicted CDS	Reference	Accession numbers
T. finnmarkense	HFJ^T	Norway	Atlantic salmon (*Salmo salar*)	2013	96	2964507	31.01	2631	Småge et al. 2016	GCA_900239485
T. finnmarkense	AY7486TD	Chile	Atlantic salmon	2015	1	2918253	31.5	2465	Grothusen et al. 2016	GCA_001483385
T. finnmarkense	TNO006	Norway	Atlantic salmon	2011	103	2933487	30.91	2595	Habib et al. 2014	GCA_900239185
T. finnmarkense	TNO010	Norway	Atlantic salmon	1998	55	2821318	31.06	2456	Habib et al. 2014	GCA_900239495
T. dicentrarchi	3509^T	Spain	European sea bass (*Dicentrarchi labrax*)	2012	65	2683831	30.15	2345	Piñeiro-Vidal et al. 2012	GCA_900239455
T. dicentrarchi	TdChD05	Chile	Atlantic salmon	2014	47	2808633	30.09	2461	Avendaño-Herrera et al. 2016	GCA_900239345
T. dicentrarchi	TNO021	Norway	Corkwing-wrasse (*Symphodus melops*)	2010	44	2667884	30.18	2337	Olsen et al. 2017	GCA_900239305
Tenacibaculum sp.	TNO020	Norway	Atlantic salmon	1998	44	2457154	30.71	2165	Habib et al. 2014	GCA_900239505

C4 – Information complémentaire et discussion

D'après le *Genome Clustering* de MicroScope (Figure 9) *T. dicentrarchi* et *T. finnmarkense* étaient regroupés dans le même cluster. Cependant, des analyses ANI réalisées à l'aide du programme pyani [213] sont en faveur de la distinction de ces deux espèces. En effet, l'ANI réalisée sur MicroScope n'est pas paramétrable et fixe la valeur seuil de définition de l'espèce à 94%. L'avantage de pyani est qu'il ne possède pas de valeur seuil et laisse le soin à l'utilisateur de conclure en prenant en compte l'ensemble des résultats. D'après les résultats de pyani présentés dans la *Supplementary Figure 1*, la valeur moyenne des pourcentages d'ANI entre les 3 génomes de *T. dicentrarchi* et les 4 de *T. finnmarkense* est de 93%. De plus, la GGDC estime une probabilité faible d'appartenance de ces souches à la même espèce (inférieure à 20% en moyenne). Néanmoins, il reste vrai que ces deux espèces possèdent des génomes très proches (et forment un cluster distinct au sein du clade II défini à la Figure 6 : Phylogénie du genre). Bien que « *T. finnmarkense* » ne soit pas officiellement un nom d'espèce valide (car le manuscrit décrivant cette nouvelle espèce a été publiée dans l'Antonie Van Leevwenkoek mais cette dernière n'est pas incluse dans la « *List of prokaryotic names with standing in nomenclature* », (http://www.bacterio.net/tenacibaculum.html), notre article supporte la définition de *T. finnmarkense* comme nouvelle espèce sur une base génomique. Dans le cadre de cet article, les génomes des souches types de *T. dicentrarchi* et « *T. finnmarkense* » ont été séquencés et publiés, ainsi que 2 génomes supplémentaires pour chacune de ces deux espèces. Un septième génome correspondant à la souche TNO020 (souche type de l'espèce *T. piscium* en cours de description par Olsen et al.) a également été séquencé. Cette espèce serait la plus apparentée au cluster *T. dicentrarchi* – « *T. finnmarkense* ». Dans cette étude, nous avons également tenté d'identifier certains facteurs de virulence. Ceux identifiés dans l'espèce *T. maritimum* ne furent retrouvés dans aucun des génomes mentionnés ici. En revanche, nous avons pu identifier un facteur de virulence (gène TFINN_140038 chez « *T. finnmarkense* ») et ses homologues dans les autres génomes du cluster codant pour une collagénase (protéase de la famille M9), mais différent de la collagénase appartenant à la famille M43 identifiée chez *T. maritimum*. Ainsi, cet article étaye l'hypothèse émise par Habib et al. [28] selon laquelle il y aurait une évolution parallèle de la pathogénicité au sein du genre *Tenacibaculum*, avec des acquisitions indépendantes de facteurs de virulence entre les différentes espèces. Il suggère aussi la présence de mécanismes d'évolution convergente de fonctions, en lien avec la virulence, dans ce groupe d'espèces pathogènes.

D – Synthèse et arbre phylogénétique complet

Les conclusions obtenues après analyse des résultats d'ANI m'ont permis d'améliorer les affiliations taxonomiques de l'ensemble des génomes appartenant à des bactéries du genre *Tenacibaculum*. J'ai alors reconstruit un arbre phylogénétique prenant en compte ces résultats afin de proposer la phylogénie la plus complète et la plus précise possible. Celle-ci (Figures 10 et 11 ci-après) inclue 96 génomes de *Tenacibaculum* ainsi que 2 génomes de l'espèce-type du genre *Pseudotenacibaculum*, un genre très apparenté au précédent. Cette phylogénie a été réalisée à partie du *core-genome* calculée par la plateforme MicroScope avec les paramètres les plus stricts proposés (80% identité en acides aminés et 80% de couverture d'alignement). Sur les 422 familles de gènes prédites comme appartenant au *core-genome*, 409 ont été finalement retenues (gènes en monocopie dans l'ensemble des génomes considérés). Les valeurs de support des nœuds sont de 100%. La topologie présentée est proche de l'arbre issu du *Genome Clustering*. Elle est plus fiable car calculée à partir d'alignements de séquences protéiques et à l'aide du maximum de vraisemblance.

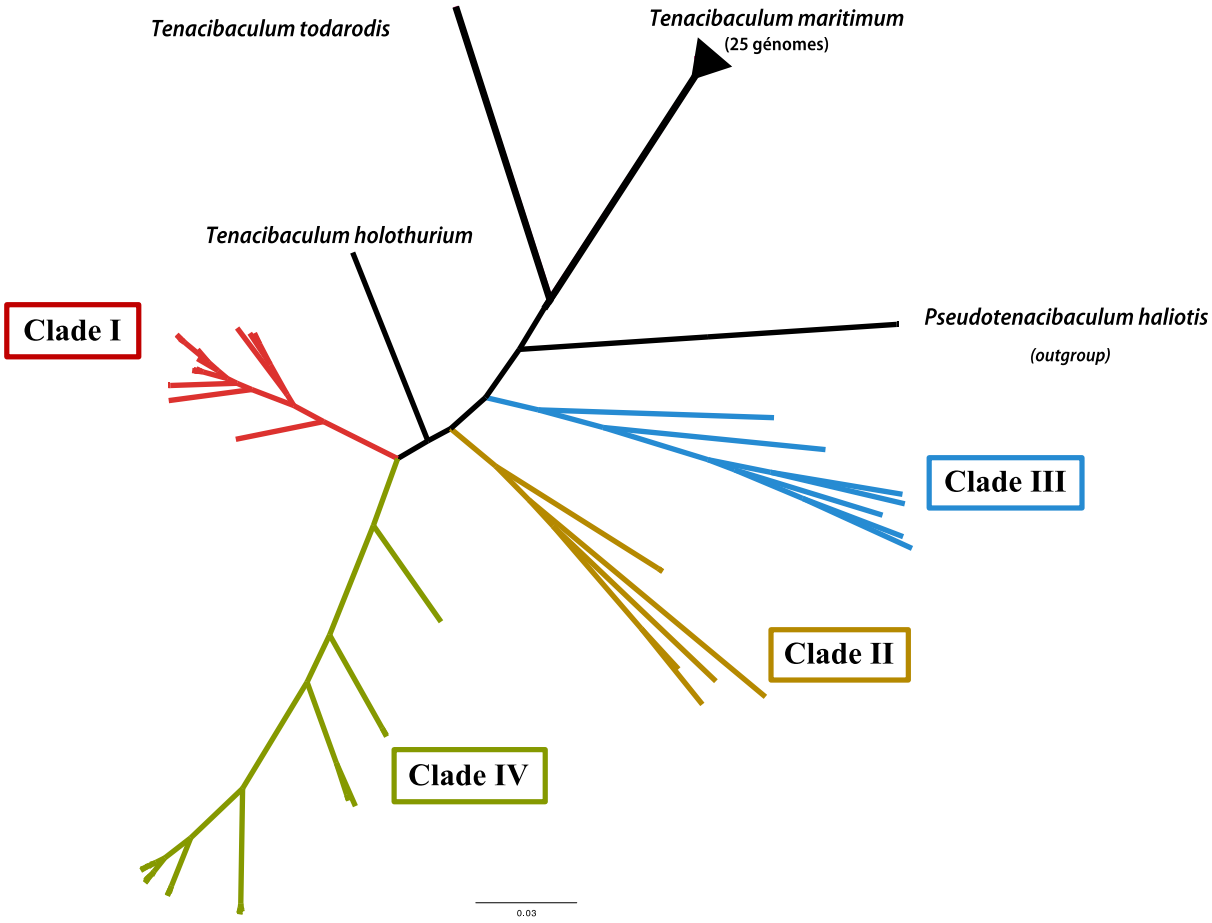


Figure 10 : Arbre phylogénétique révisé représentant une proposition comprenant 4 clades au sein du genre

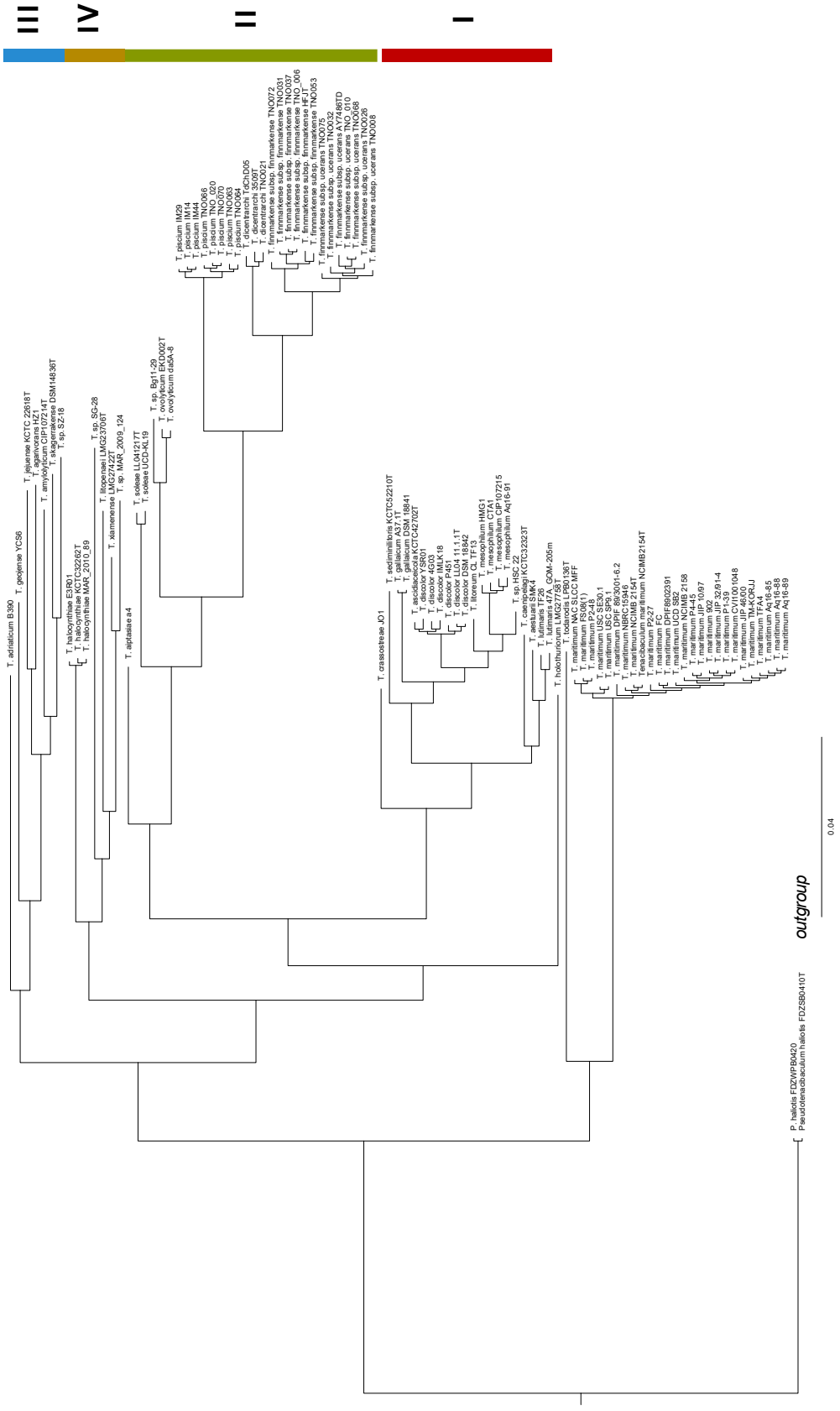


Figure 11: Arbre phylogénétique des 32 espèces du genre Tenacibaculum tenant compte des analyses d'ANI (raciné par l'outgroup Pseudotenacibaculum)

Partie 2 : Construction des spectres de référence MALDI-TOF

A – Introduction

Le MALDI-TOF est un outil efficace pour l'identification de routine, comme nous l'avons vu dans le chapitre introductif. Cependant, l'utilisation du MALDI-TOF est limitée aux espèces présentes dans les différentes bases de données disponibles. Cette limite est d'autant plus contraignante lorsque l'on travaille sur des espèces peu étudiées, comme c'est le cas pour celles appartenant au genre *Tenacibaculum*. Bien sûr, il est toujours possible pour un chercheur de réaliser des spectres de référence et de lancer la procédure d'intégration de ces données à la base de données commerciale correspondante. Cependant, c'est un processus contraignant et qui en limite l'accès à une seule base de données. Un des objectifs de ce projet doctoral était de pallier cette limite, en proposant non seulement de construire la première base de données spectrale du genre *Tenacibaculum* la plus exhaustive qui soit, mais également une méthodologie, basée sur des logiciels libres, permettant de comparer un échantillon inconnu à cette base afin d'identifier l'espèce correspondante. Pour cela, l'ensemble des spectres des souches types ont été obtenu. Pour s'assurer de la cohérence de notre collection et pour approfondir nos connaissances sur le genre, leurs génomes ont été séquencés (cf. Première partie). Puis, chaque souche type a été analysée par MALDI-TOF. Enfin, deux méthodes d'identification d'isolats ont été développées.

B – Matériel & Protocoles

B1 – Collection de souches

Afin de construire la banque de données de spectres de référence, les souches types (issues de différentes collections internationales) de 24 espèces appartenant au genre *Tenacibaculum* ont été analysées par MALDI-TOF. Jean-François Bernardet, Ingénieur de Recherche dans l'équipe IIP, s'est occupé de cet aspect.

Tableau 3: Liste des souches types pour lesquelles le spectre de référence MALDI-TOF a été réalisé

Nom de l'espèce	**Souche**
T. adriaticum	DSM 18961^T
T. aestuarii	JCM 13491^T
T. aiptasiae	LM 24004^T
T. ascidiaceicola	KCTC 42702^T
T. amylolyticum	CIP 107214^T
T. caenipelagi	KCTC 32323^T
T. crassostreae	JCM 15428^T
T. dicentrarchi	USC 3509^T
T. discolor	LL04 11-1-1^T
"*T. finnmarkense*"	HFJT^T
T. gallaicum	A37-1^T
T. geojense	KCTC 23423^T
"*T. halocynthiae*"	KCTC 32262^T
T. jejuense	KCTC 22618^T
T. litopenaei	LMG 23706^T
T. litoreum	JCM 13039^T
T. lutimaris	DSM 16505^T
T. maritimum	NCIMB 2154^T
T. mesophilum	CIP 107215^T
T. ovolyticum	EKD 002^T
T. sediminilitoris	KCTC 52210^T
T. skagerrakense	DSM 14836^T
T. soleae	LL04 12-1-7^T
T. xiamenense	LMG 27422^T

B2 – Protocole de préparation des protéines pour l'analyse MALDI-TOF

Ce protocole correspond à une extraction légère des protéines bactériennes totales. Le matériel biologique (une ou plusieurs colonies) est transféré dans un tube de 1,5 mL contenant 300µL d'eau déminéralisée, puis le mélange est homogénéisé par pipetage. 900µL d'éthanol sont ajoutés, puis le mélange est à nouveau homogénéisé par pipetage. Le mélange est centrifugé pendant 2 minutes à 13000 g et le surnageant est éliminé. L'échantillon est à nouveau centrifugé pendant quelques secondes à 13000 g afin d'éliminer toute trace résiduelle d'éthanol. Le culot est séché pendant 5 minutes. Entre 10 µL et 50 µL d'acide formique à 70% sont ensuite ajoutés et l'échantillon est homogénéisé. Un volume d'acétonitrile équivalent au volume d'acide formique utilisé dans l'étape précédente est ajouté. L'échantillon est à nouveau mélangé puis centrifugé pendant 2 minutes à 13000 g ; le culot correspond aux débris bactériens. On dépose 1 µL du surnageant sur la cible MALDI, puis on laisser sécher le dépôt. Ce dernier est alors recouvert d'1 µL de matrice HCCA et de nouveau séché. Enfin, la cible MALDI peut être chargée dans le spectromètre de masse.

Pour réaliser les spectres de référence, chaque souche type est cultivée en triplicats. Pour chaque culture, 12 spectres sont enregistrés (4 dépôts sur la cible MALDI et 3 spectres par dépôt). Il y a donc 36 spectres par souche type. Pour les isolats de terrain, 12 spectres sont enregistrés, sans réplicat biologique. L'appareil utilisé est un spectromètre de masse Bruker microflex.

B3 – Protocoles supplémentaires de préparation des échantillons

Pour évaluer nos outils d'analyse, nous voulions les tester sur des données de qualité variables. De plus, un protocole de conservation des culots bactériens en éthanol a été réalisé pour évaluer la possibilité de faciliter le transport d'isolats de terrain sous forme de matériel fixé (et donc sans danger). Cela permettrait de créer des collaborations entre des laboratoires qui ne sont pas équipés en spectromètre de masse MALDI-TOF situés dans différents endroits du monde.

Dans le cadre du projet, nous avons donc préparé et testé 3 jeux de données distincts. Chacun d'entre eux a été réalisé avec différents protocoles de préparation avant acquisition par le spectromètre de masse MALDI-TOF. Le premier correspond à l'extraction protéique totale. Il est décrit dans la partie matériel (voir partie III, paragraphe B2). Le second, correspondant au Dépôt Direct avec Acide Formique (DDFA), consiste simplement à déposer l'échantillon directement sur la plaque MALDI-TOF puis à ajouter une goutte d'acide formique pour faire l'équivalent d'une « extraction rapide ». Le dernier consiste à conserver l'échantillon plusieurs jours en éthanol (dans notre cas de 7 à 26 jours) avant de réaliser le protocole d'extraction dont l'éthanol est la première étape.

La préparation des échantillons et les acquisitions des spectres ont été réalisées par Frédéric Bourgeon dans le laboratoire BioChêneVert.

C – Méthodes

C1 – Première approche : référence spectre entier

Les spectres analysés sont traités à l'aide de scripts utilisant le langage de programmation R et les packages MALDIquant Foreign, MALDIquant, et MALDIrppa [214, 215]. L'analyse de ces données nécessite un prétraitement. Le principe reste le même quel que soit le logiciel utilisé. Premièrement, la variance de l'intensité des pics est transformée (transformation racine carrée) afin de simplifier la représentation graphique des empreintes spectrales. Les spectres sont ensuite lissés par la méthode Savitsky-Golay à 21 points [216], puis la ligne de base est corrigée à l'aide de la méthode *Statistics-sensitive Non-linear Iterative Peak-clipping algorithm* (SNIP) pour la ramener à 0 [217]. L'intensité des spectres est ensuite calibrée à l'aide de l'algorithme *Total-Ion-Current* (TIC). Enfin, comme il y a de nombreux réplicats par souche type (36 pour les spectres de référence), le spectre moyen est calculé. Pour cela, les valeurs de masses sont d'abord calibrées par un algorithme d'alignement appliqué aux 36 spectres entre eux puis le spectre moyen est établi. Toutes ces étapes correspondent à des fonctions R paramétrables. Ici, nous cherchons à obtenir une référence définie comme étant une liste restreinte mais fiable de pics. Les paramètres choisis sont ceux par défaut (méthode : *Mean Absolute Deviation*), à l'exception de la valeur seuil du rapport signal/bruit qui fut, après évaluation, fixée à 3. Les pics obtenus sont alignés et filtrés aves les paramètres par défauts. Une liste d'environ 50 pics (min. : 23, moyenne : 53.43, max. : 66) est finalement obtenue. Le schéma illustrant la méthode de création d'une référence est représenté dans la figure 12 ci-dessous.

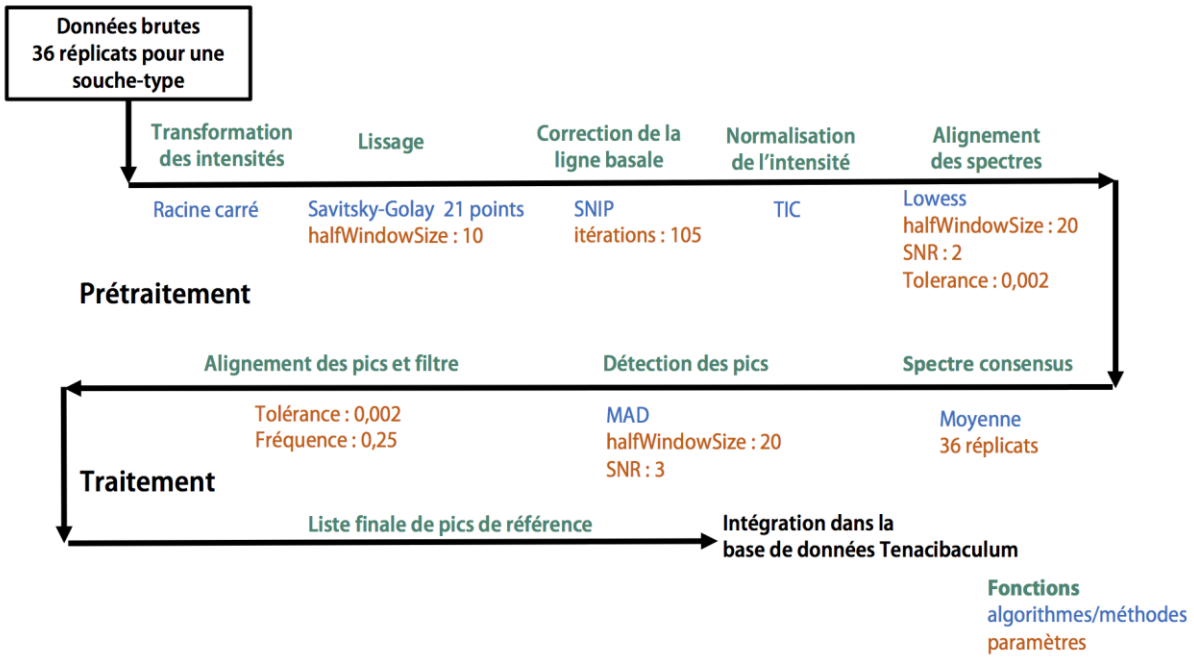


Figure 12: Schéma de création d'une référence "spectre entier"

Une fois la liste de pics spécifiques obtenue pour chaque souche type, il faut pouvoir comparer cette liste à celle obtenue à partir d'un échantillon afin de proposer une identification. La méthode d'identification consiste à traiter l'isolat de terrain selon la même procédure qui a permis d'établir les listes de pics de référence à partir des souches types. Une fois la liste de pics de l'échantillon obtenue, chaque masse est comparée successivement à chacune des références. Un schéma récapitulatif de la méthode d'identification est présenté en Figure 13.

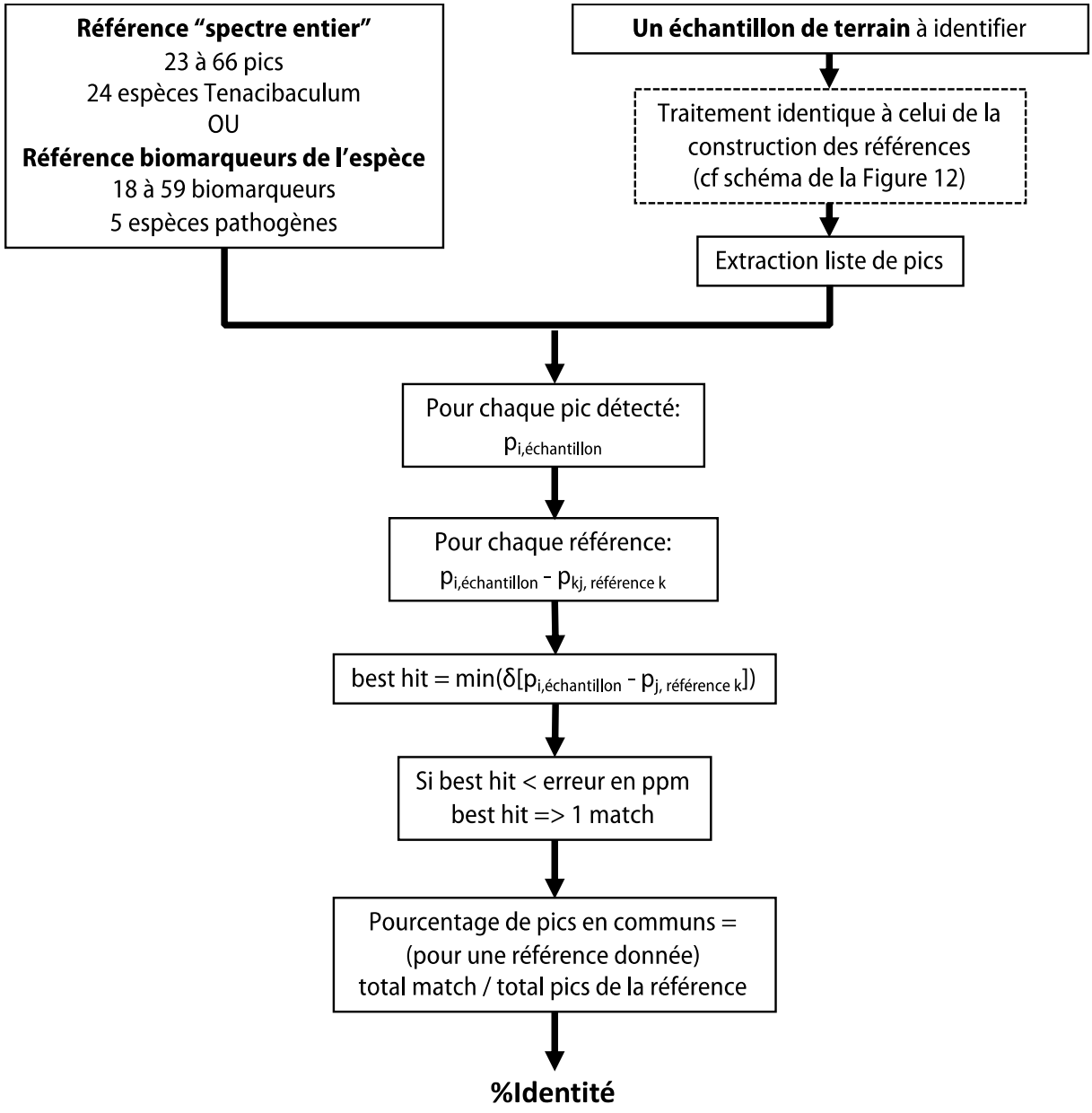


Figure 13 : Schéma de l'identification d'un isolat par spectrométrie de masse MALDI-TOF.

$p_{i,échantillon}$: i-ème pic de l'échantillon ; $p_{j,référence\ k}$: j-ème pic de la référence k ; δ : différence

Un pic observé est retrouvé dans une référence si, et seulement si, la différence absolue entre la valeur du pic observé et celle du pic de référence est inférieure à l'erreur tolérée de la mesure du MALDI-TOF. Généralement, dans la littérature, l'erreur admise est de 500 ppm (partie par millions), correspondant à une erreur acceptable de 5 Da pour une masse observée de 10 000 Da.

Nous avons donc choisi cette valeur pour notre méthode. Enfin, le nombre de pics retrouvés dans chacune des références, rapporté au nombre total de pics inclus dans cette même référence, est calculé (pourcentage d'identité). Le pourcentage d'identité permet d'identifier (ou non) l'espèce à laquelle appartient l'isolat. La valeur à partir de laquelle l'identification est considérée comme fiable est, idéalement, à définir pour chacune des espèces d'intérêt.

Dans le cadre de mon projet, le seuil « optimal » de décision a pu être obtenu pour l'espèce *T. maritimum*. Pour cela la précision de cette méthode a été évaluée en faisant varier la valeur seuil. Faute de jeux de données suffisants pour les autres espèces de *Tenacibaculum* (certaines n'étant constituées que de leur seule souche type), ce seuil n'a pas été calculé pour l'ensemble des espèces présentes dans notre banque de données. Pour déterminer la valeur optimale pour l'espèce *T. maritimum,* nous avons utilisé deux jeux de données. Nous avons construit un ensemble correspondant au contrôle positif. Il contient environ 300 acquisitions indépendantes (représentant près de 3300 spectres) correspondant à des isolats préalablement identifiés comme appartenant à l'espèce *T. maritimum* par MLST. Nous avons également construit un deuxième ensemble correspondant au contrôle négatif contenant les spectres réalisés sur les autres espèces du genre *Tenacibaculum* (environ 130 acquisitions indépendantes, correspondant approximativement à 4500 spectres). Nous avons alors considéré notre outil d'identification comme une classification binaire : un isolat appartient à l'espèce *T. maritimum* ou pas. Pour calculer la valeur de la précision, il faut d'abord calculer le nombre de vrais ou faux positifs (TP, FP) correspondant au nombre d'isolats correctement ou incorrectement identifiés comme *T. maritimum* ainsi que le nombre de vrais ou faux négatifs (TN, FN) correspondant respectivement aux nombres d'isolats correctement ou incorrectement rejetés en fonction d'une valeur seuil définie préalablement. La précision est calculée comme suit : $ACC = (TP + TN)/(TP + FP + TN + FN)$. La variation de la précision selon la valeur seuil d'identification est présentée dans la Figure 14. Nous pouvons voir que la précision est maximale (environ 85%) entre les valeurs seuil de 15% (ligne rouge) à 40% (ligne verte) (Figure 14). Nous proposons donc pour cette méthode une valeur optimale de 40%. La Figure 15 présente la distribution des pourcentages d'identité sur l'ensemble du jeu de données (dont les contrôles positif et négatif).

Pour améliorer nos outils d'identification, il faudrait faire le même travail pour toutes les espèces et faire varier le seuil de décision en fonction de l'espèce considérée. D'autre part, pour des espèces très proches il ne serait peut-être pas possible d'utiliser cette approche et il faudrait envisager une approche par biomarqueur, approche détaillée dans le paragraphe suivant.

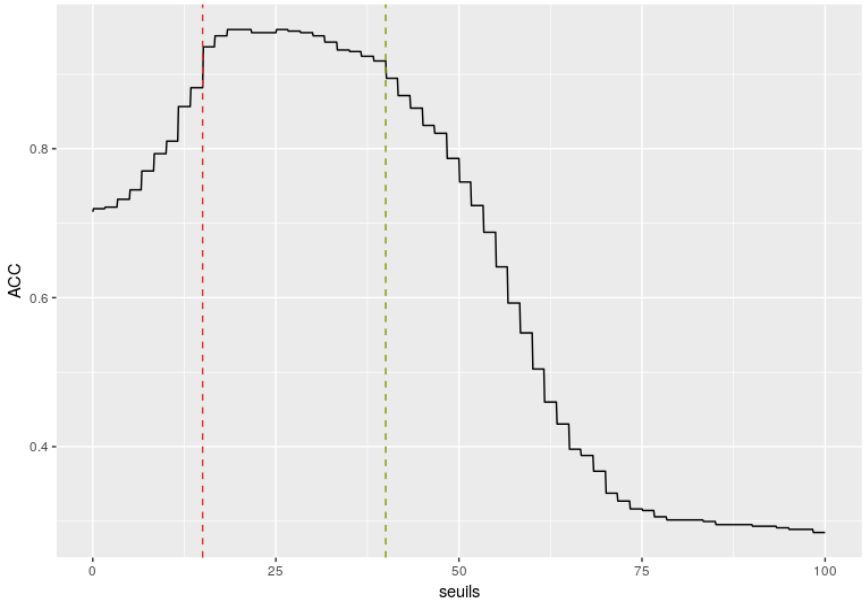


Figure 14 : Evolution de la précision en fonction de la valeur seuil de décision

Axe x : valeur seuil de pourcentage d'identité identifiant un échantillon comme appartenant à l'espèce *T. maritimum*.
Axe y : valeur de la précision obtenue sur l'intégralité du jeu de données.

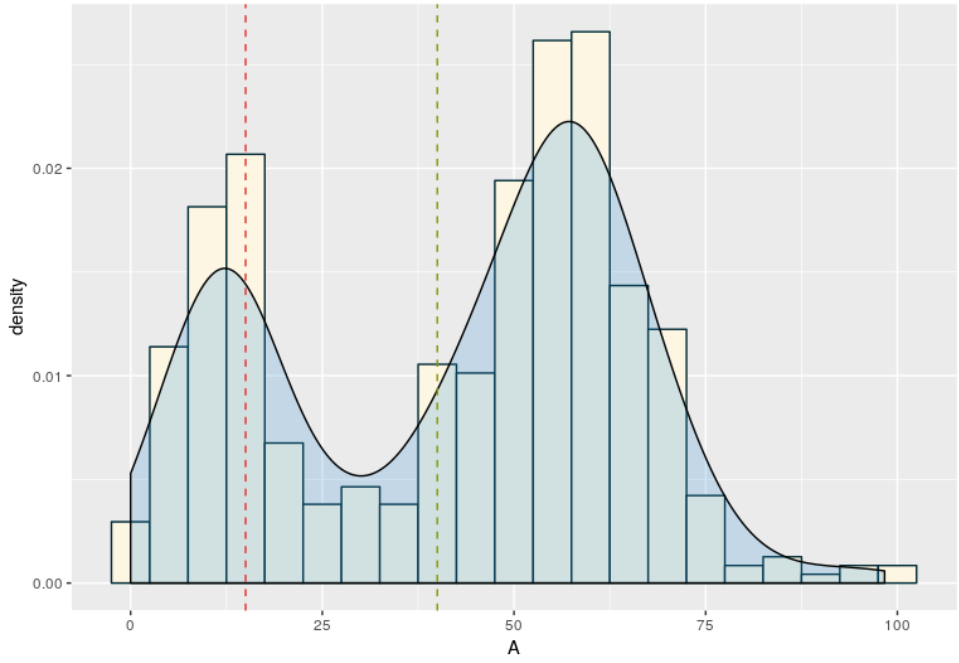


Figure 15 : Distribution des pourcentages d'identité obtenus sur l'ensemble du jeu de données

Axe x : pourcentage d'identité avec la référence *T. maritimum* mesurée sur l'intégralité du jeu de données (contrôle positif et contrôle négatif inclus)
Axe y : densité et fréquence d'observations des pourçentages d'identité

C2 – Seconde approche : biomarqueurs d'espèce

Afin de proposer une identification reposant sur des données biologiques réelles, nous avons développé une méthode basée sur des biomarqueurs propres à l'espèce. Comme précédemment évoqué, les protéines ribosomiques sont les protéines les plus représentées dans les spectres MALDI-TOF [172, 206]. Notre recherche de biomarqueurs s'est donc focalisée sur ces protéines. Le schéma de la méthode est présenté Figure 16 ci-dessous. Nous nous sommes intéressés à 4 espèces pathogènes : *T. dicentrarchi*, *T. discolor*, "*T. finnmarkense*" et *T. maritimum*. Pour définir des biomarqueurs pertinents, il est nécessaire d'avoir plusieurs génomes pour la même espèce. En effet, il est nécessaire de choisir des protéines qui ne présentent, *a priori*, aucun polymorphisme. En moyenne, environ la moitié des 54 protéines ribosomiques sont retrouvées dans les spectres MALDI-TOF. Cette approche est plus robuste dans le sens où elle se base sur l'information génomique pour identifier précisément certains pics. En moyenne, 38,4 pics (min. : 18, max. : 59) ont été identifiés comme biomarqueurs pertinents de l'espèce. Sachant que les masses observées dans les spectres sont exprimées par un rapport masse sur charge, il arrive d'observer 2 ou 3 pics correspondant à la même protéine (selon le degré d'ionisation). C'est ainsi que l'on obtient parfois plus de pics identifiés qu'il n'existe de protéines ribosomiques. La valeur seuil d'identification pour l'espèce *T. maritimum* a été calculée selon la méthode décrite dans le paragraphe précédent (partie III, §C1). La valeur seuil est de 60% (calculs détaillés partie IV, paragraphe C, Supplementary Material S7). La méthode d'identification de cette approche est identique à la méthode décrite au paragraphe précédent (§C1) et est illustrée Figure 13.

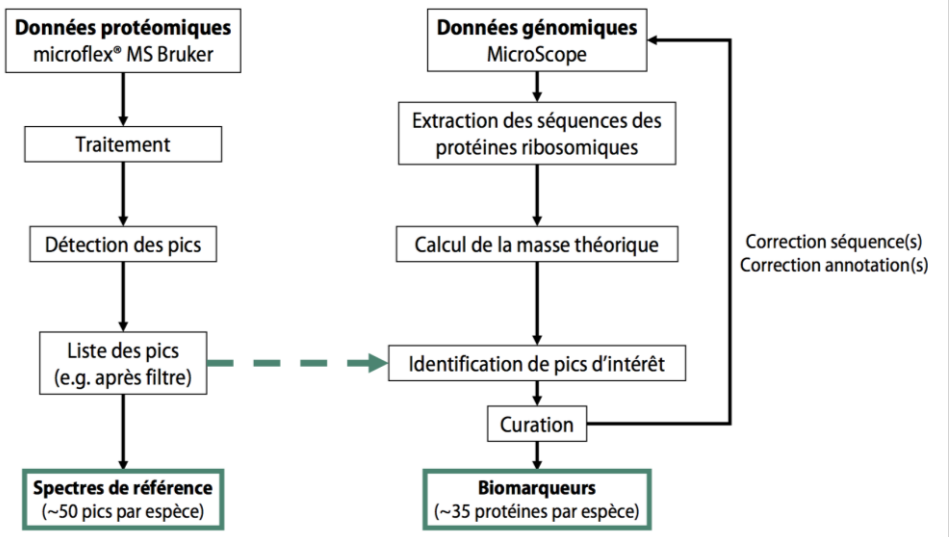


Figure 16: Schéma de création d'une référence basée sur des biomarqueurs ribosomiques

D – Comparaison des approches d'identification

Nous avons développé et utilisé deux méthodes d'identification d'isolats à partir de données de spectrométrie de masse MALDI-TOF. Ces deux méthodes sont complémentaires. La première (référence « spectre entier ») est générique. La construction de la référence est donc directe : il s'agit d'une simple détection de pics conservés (avec un seuil rapport/bruit élevé de 3).

La deuxième méthode (référence biomarqueurs de l'espèce) est moins évidente à mettre en place, car elle nécessite plus de mise au point. Il est en effet nécessaire de combiner les informations génomiques (idéalement, plusieurs génomes) aux empreintes spectrales afin d'identifier des biomarqueurs pertinents (retrouvés de manière quasi-systématique). L'utilisation de plusieurs génomes permet de privilégier les protéines ayant une séquence en acides-aminés conservée dans les génomes considérés. Ainsi, nous nous assurons que les chances d'identifier ces biomarqueurs sont élevées pour des isolats appartenant à la même espèce. Cette deuxième approche a donc été intégrée pour les espèces d'intérêt majeur (4 espèces pathogènes) et pour lesquelles nous avions plusieurs génomes à notre disposition (de 3 à 25 selon les cas). Il est évident que l'utilisation d'un grand nombre de génomes favorise la sélection des meilleurs biomarqueurs).

Afin de comparer les deux approches, nous les avons appliquées à l'ensemble des jeux de données (y compris les contrôles positif et négatif) en prenant comme référence *T. maritimum*. Le pourcentage d'identité moyen a été calculé en utilisant 3 protocoles différents de préparation d'échantillons (ceux présentés au paragraphe B3). Le premier, correspondant à l'extraction, est décrit dans la partie matériel (voir Protocole MALDI-TOF). Le second, correspondant au Dépôt Direct avec Acide Formique (DDFA), comporte deux mesures différentes. Nous avons vu que les protéines pouvaient apparaître plusieurs fois dans les spectres en fonction de leurs différents états d'ionisation. Nous avons donc une mesure qui se base sur le nombre total de pics communs (entre la référence *T. maritimum* et l'isolat considéré) et une seconde mesure qui ne prend en compte que le nombre réel de protéines identifiées dans l'échantillon considéré.

Tableau 4: Évaluation comparée des deux méthodes d'identification de spectres MALDI-TOF

Jeu de Données	Conditions	Première approche : Spectre entier (60 pics)	Seconde approche : biomarqueurs	
			Pics* (18 pics)	Protéines** (9 protéines)
Contrôle positif *T. maritimum* 130 isolats	Extraction	57,82 %	83,60 %	91,68 %
	Dépôt Direct	48,77 %	80,82 %	86,86 %
	Éthanol	54,79 %	82,99 %	93,00 %
	TOTAL	**54,03** %	**82,51** %	**90,37** %
Contrôle négatif 23 autres espèces	Extraction	12,88 %	7,02 %	10,34 %
	Dépôt Direct	11,15 %	5,98 %	9,57 %
	TOTAL	**12,04** %	**6,52** %	**9,97** %

* L'ensemble des 18 biomarqueurs (correspondants à 18 pics définis dans les spectres)

** Ces 18 biomarqueurs correspondent à 9 protéines ribosomiques monomorphes, visibles dans l'empreinte spectrale sous forme mono, di ou tri-chargée. Ce pourcentage représente le nombre de protéines retrouvées, quel que soit leur état de charge.

Nous pouvons voir d'après le tableau 4 que le protocole dépôt direct est celui où l'identification est la plus difficile. Cependant, les valeurs restent globalement comparables aux deux autres protocoles. Le gain de la seconde approche, basée sur des biomarqueurs protéiques, est notable. En effet, le pourcentage de pics retrouvés est supérieur de 30 à 40% par rapport à la première approche ce qui représente un gain significatif de sensibilité. De plus, ce pourcentage diminue de 2 à 6% pour le contrôle négatif ce qui indique un gain significatif de spécificité. Nous pouvons également voir que, pour la seconde approche, le calcul basé sur la capacité de l'outil à détecter la présence de protéines plutôt que sur la totalité des pics donne toujours un pourcentage plus élevé. Ces résultats sont encourageants, et suggèrent que l'on peut développer des outils d'identification fiables. Néanmoins, il faut noter que ce genre d'approche est intéressant seulement pour l'étude de bactéries absentes des différentes banques de données ou bien pour des études spécifiques ciblées sur un nombre limité d'espèces d'intérêt. En effet, pour que cette approche fonctionne, il faut définir des biomarqueurs à partir d'un jeu de données fiables extraites de la comparaison de génomes complets. Pour savoir si les données génomiques sont strictement nécessaires pour développer une approche biomarqueur, j'ai décidé de tester si la seconde approche pouvait être développée sans données génomiques.

Une des solutions possibles est de raisonner sur la notion de « super-spectre ». En prenant un grand nombre de spectres d'isolats variés appartenant à la même espèce (et idéalement sous plusieurs conditions : *a minima* en suivant le protocole extraction classique et le protocole DDFA), nous devrions être en mesure de définir des biomarqueurs comme étant des pics retrouvés dans l'ensemble de ces spectres (avec un seuil donné). J'ai testé cette troisième approche sur le même jeu de données. Pour identifier les biomarqueurs du « super-spectre », les spectres sont moyennés par dépôt (4 dépôts x 3 spectres par acquisition) et non plus par acquisition. De plus, seuls les pics présents dans plus de 85% des spectres ont été retenus. Ainsi, sur le jeu de données correspondant au contrôle positif (et en enlevant manuellement les spectres de mauvaise qualité), 24 pics sont obtenus. Nous sommes très proches des 18 biomarqueurs identifiés dans la seconde approche. Cette liste de 24 pics contient 10 biomarqueurs précédemment identifiés (valeurs en gras) comme protéines ribosomiques (Tableau 5). Cette approche permet d'avoir des résultats comparables à ceux de la seconde : en moyenne, le pourcentage d'identité est de 89,52% pour le contrôle positif, et de 14,25% pour le contrôle négatif. Il est donc vraisemblablement possible de développer une approche de type biomarqueur sans que les génomes soient disponibles. Cependant, cette approche doit compenser l'absence de données génomiques par un jeu de données plus important : plus le jeu de données utilisé sera grand, plus les biomarqueurs identifiés seront pertinents.

Tableau 5: Liste des biomarqueurs « super-spectre » de l'espèce T. maritimum identifiés comme protéines ribosomiques. m/M indique le clivage ou non de la première méthionine. H1, H2, H3 indiquent l'état de charge (mono-, di-, tri-chargée).

Biomarqueurs "Super-spectre"	Identification
3093	RpmH m-H2
3441	RpmG m-H2
3931	RpsU M-H2
4439	RpmB m-H2
4532	RpmJ M-H1
6884	RpmG m-H1
7863	RpsU M-H1
8877	RpmB m-H1
10028	RpsS m-H1
11357	RpsR m-H1

Partie 3 : Typage des isolats appartenant à l'espèce *T. maritimum*

A – Introduction

Comme évoqué précédemment, les spectres MALDI-TOF contiennent en théorie suffisamment d'information pour typer des isolats bactériens. Un des objectifs du projet doctoral était de développer une méthode de caractérisation d'isolats, c'est-à-dire une méthode permettant d'identifier des groupes de souches au sein de la même espèce. Cette approche nécessite une analyse poussée des empreintes spectrales d'un ensemble de spectres contenant plusieurs dizaines d'isolats. Le but de cette partie est de tester l'utilisation du spectromètre de masse MALDI-TOF comme outil épidémiologique pour l'espèce *T. maritimum*. Cela pourrait permettre de mieux comprendre la structure de la population bactérienne dans les pisciculture. Par exemple, cet outil serait capable de déterminer si la même souche (ou des souches proches) est responsable d'événements infectieux successifs dans le même bassin ou le même élevage.

Comme déjà évoqué dans l'introduction, le signal spectral est une donnée analogique, c'est-à-dire que les paramètres physico-chimiques peuvent influencer les données obtenues. A l'échelle de l'isolat, ces légères variations (disparation/apparition) de pics peuvent apporter un bruit difficile à évaluer. La caractérisation d'isolats nécessite d'utiliser un type précis d'information dans les spectres. En fait, il existe 3 types d'informations dans les données MALDI-TOF : l'intensité du signal, la perte (et le gain) de signal et le *peak shift* (décalage du signal). Une variation du signal peut être due à une modification d'expression protéique (mutations génétiques dans les séquences régulatrices) ou une adaptation phénotypique aux conditions de culture. Une perte de signal peut apparaître après une mutation non-sens ou un *frameshift* (décalage du cadre de lecture) entraînant l'apparition prématurée d'un codon stop. Elle pourrait également survenir en raison de l'absence d'un gène, au mode de préparation de l'échantillon ou aux conditions de culture. Ainsi, la variation d'intensité et l'apparition/disparition de pics donne une information ambigüe sur le génotype d'un isolat. Cependant, les *peak shift* sont la conséquence directe d'une modification de la séquence en acides aminés de la protéine après une (ou plusieurs) mutation(s) non synonyme(s) dans le gène correspondant. Ces *peak shift* se matérialisent dans les spectres par un changement de masse entrainant un décalage, en amont ou en aval, du pic théoriquement attendu pour la protéine correspondante. Ils sont donc de très bons candidats comme biomarqueurs pertinents pour le typage bactérien.

La première stratégie étudiée (sans doute un peu naïve) est basée sur le spectre entier et une classification hiérarchique non-supervisée. Elle se focalise sur un traitement binaire des empreintes, ce qui correspond à une analyse d'absence et de présence de pics. Bien que cette méthode semblât intéressante à évaluer, il s'est avéré qu'elle n'était pas adaptée à une application de routine. En revanche, elle permit de mettre en évidence les *peak shift* majeurs pour l'espèce *T. maritimum*. Nous avons ensuite imaginé une approche uniquement basée sur des *peak shifts* de biomarqueurs sélectionnés. C'est ici qu'apparaît la difficulté majeure et l'étape la plus critique de cette nouvelle approche : l'identification de ces biomarqueurs. Elle peut être réalisée par une comparaison visuelle des pics ou bien par une combinaison de méthodes bioinformatiques [181]. Nous avons développé une approche intermédiaire semi-automatisée permettant de trouver des biomarqueurs potentiels. A l'aide de 24 génomes de *T. maritimum*, nous avons pu établir une liste de candidats à partir du polymorphisme observé pour certaines protéines ribosomiques. Ceux-ci ont été ensuite recherchés manuellement dans les spectres. Finalement, 9 protéines ribosomiques ont été retenues comme biomarqueurs de

typage pour notre premier schéma de typage *Multi Peak Shift Typing* (MPST). L'outil de typage, basé sur ce schéma, a été également intégré dans une application web développée dans le cadre du projet doctoral, MALDIquantTypeR (http://genome.jouy.inra.fr/shiny/maldiquanttyper/.

B – Typage par classification non-supervisée des spectres MALDI-TOF

B1 – Matériel & Méthodes

Isolats appartenant à l'espèce *T. maritimum*

Le jeu de données utilisé pour cette étude comprend 74 isolats (74 acquisitions) en protocole extraction. Cela représente les ¾ du jeu de données final. En effet, au moment du développement de cette approche, l'ensemble du jeu de données n'était pas encore entièrement analysé par spectrométrie de masse MALDI-TOF. Cela représente tout de même approximativement 800 spectres.

Classification non supervisée des spectres

Cette approche est basée sur le même traitement initial que celui décrit dans la partie III paragraphe C1 et Figure 12. Cependant, la méthode change après le calcul des spectres moyens (Figure 16). Une matrice d'intensité de l'ensemble des pics détectés sur le jeu de données est obtenue. Dans cette matrice, chaque colonne correspond à un pic détecté, chaque ligne correspond à une acquisition (un isolat) du jeu de données et chaque case contient une valeur d'intensité (si le pic est présent dans le spectre correspondant). Cette matrice est transformée en matrice binaire selon les règles suivantes :

- Une case vide correspond à 0 (absence du pic)
- Une case non vide correspond à 1 (présence du pic)
- Si un pic est détecté avec une intensité inférieure à 20% de l'intensité maximale observée pour ce pic, il est considéré comme artéfact et la case est mise à 0

Nous avons choisi une approche binaire car, pour le typage d'isolats bactériens, une variation d'intensité n'est pas informative [181]. Cette matrice est ensuite analysée par classification hiérarchique. Comme nous voulions qu'une mesure de fiabilité de l'arbre soit calculée, nous avons choisi d'utiliser le package *pvclust*. Celui-ci calcule deux valeurs pour chacun des nœuds de l'arbre de classification obtenu : un bootstrap classique et une *p-value* similaire au SH-like de FasTree 2 (bootstrap multi-échelle). Nous espérions que cela nous permettrait de définir le nombre de groupes corrects existant dans nos données. Pour la classification hiérarchique, l'algorithme utilisé est le ward.D2. La mesure de distance utilisée est l'indice de Jaccard (mesure binaire asymétrique). Celle-ci a été choisie car nous travaillons avec une matrice binaire dans laquelle le 0 signifie une absence de pic et le 1 signifie une présence de pics (le 0 et le 1 ne sont pas interchangeables).

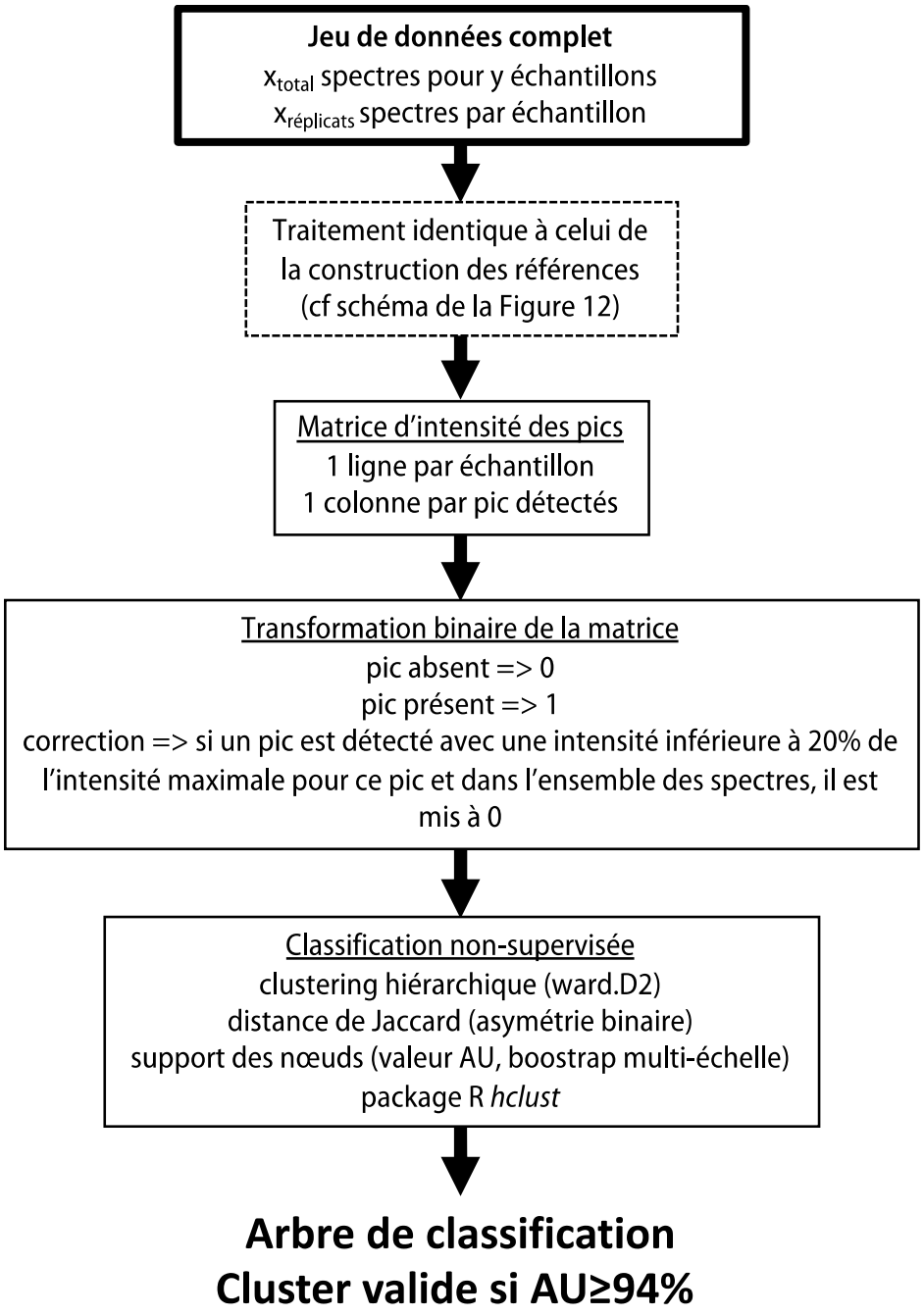


Figure 17: Schéma de classification non-supervisée des spectres MALDI-TOF

B2 – Résultats & Discussion

Dans cette partie, nous nous intéresserons aux arbres de classification hiérarchique obtenus grâce à la méthode décrite dans le paragraphe précédent. Ces arbres sont calculés avec le package R *hclust* qui permet d'obtenir des valeurs de support statistique pour chacun des nœuds. Ces valeurs nous permettrons de définir les clusters valides. Cet outil calcule deux valeurs de support : un *bootstrap* classique (BP) et un *bootstrap* multi-échelle (AU). Ces deux modes de calculs du *bootstrap* sont détaillés dans l'annexe 1. Après observation des premiers résultats, nous avons choisi d'utiliser les valeurs AU qui nous semblaient plus adaptés aux données

Le premier arbre (Figure 18), réalisé avec 1000 réplicats pour le bootsrap, présente une classification en 2 groupes (violet et orange). Les valeurs vertes à droite de chaque nœud représentent la valeur de *bootstrap* classique (BP). Les valeurs bleues à gauche de chaque nœud représentent la valeur de *bootstrap* multi-échelle (AU). Sur cet arbre, nous pouvons définir une classification binaire de nos spectres : le cluster orange (95% AU) et le cluster violet (94%)

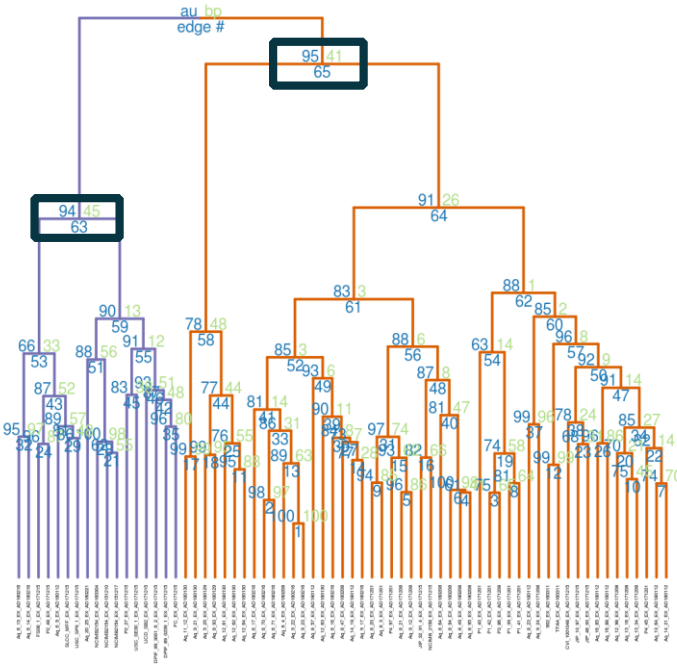


Figure 18 : Clustering hiérarchique, 2 clusters identifiés

Les deux chiffres en amont de chaque nœud indiquent à gauche (bleu) la valeur AU de bootsrap multi-échelle et à droite (vert) la valeur BP de bootstrap classique. La valeur en aval de chaque nœud (bleu) indique le numéro du nœud)

Le deuxième arbre (Figure 19), réalisé sur le même jeu de données avec 2×10^5 itérations (*bootstrap*), était intéressant. Pour la première fois, nous pouvons définir des sous-groupes au sein du cluster I (violet dans la figure précédente) et au sein du cluster II (orange). Les clusters « valides » ont une AU supérieure ou égale à 94%. Néanmoins, il montre aussi la fragilité de la classification obtenue, sensible à de nombreux paramètres comme le nombre de réplicats.

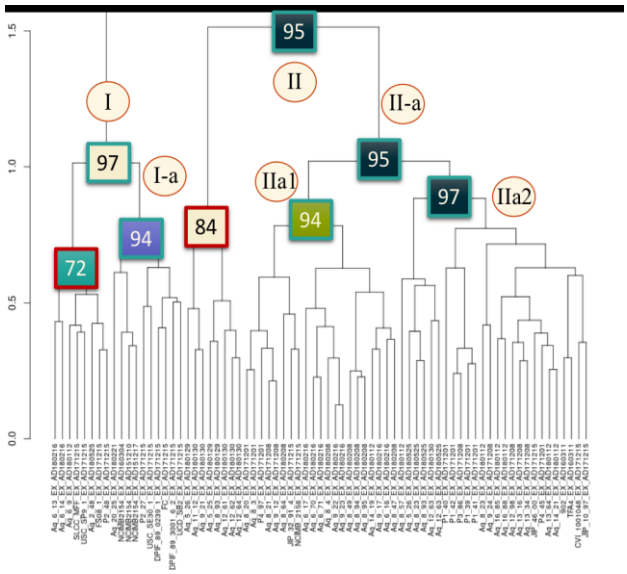


Figure 19 : Clustering hiérarchique, 6 clusters identifiés (encadrés verts)

Les valeurs encadrées indiquent la valeur AU du nœud. Un cadre vert indique que le cluster est valide tandis qu'un cadre rouge indique un cluster invalide. Les valeurs cerclées indiquent la numérotation des clusters valides.

Nous pouvons également affiner notre analyse en essayant de déterminer les pics spécifiques aux clusters identifiés. Pour cela la fonction *sda.ranking()* du package *sda* fut utilisée. Cette fonction est capable d'identifier les variables les plus discriminantes pour une classification donnée. Nous avons donc utilisé cette fonction en proposant la classification binaire (cluster orange et cluster violet) du premier arbre. Cette fonction génère une liste ordonnée des 40 variables (donc 40 pics) les plus discriminantes (Figure 20). Si l'on s'intéresse aux pics les plus discriminants numéros 1, 2, 3 et 5 (repérés par des flèches orange et violettes), nous pouvons remarquer que le pic 2 apparait à une masse deux fois plus grande que le pic 1, et de même pour les pics 3 et 5. Cela correspond en fait à une même protéine monochargée (pic 2 et 5) et dichargée (pic 1 et 3). Les barres blanches indiquent la présence du pic et les barres noires indiquent l'absence de ce pic. Ainsi, les pics 1 et 2 sont présents dans le cluster orange, mais pas dans le groupe violet. Inversement, les pics 3 et 5 sont spécifiques au cluster violet. En fait, le phénomène observé est un décalage de pics (*peak shift*). Une même protéine possède une certaine variabilité entre les isolats d'une même espèce. Ces mutations entrainent un changement de la séquence protéique et *in fine* un décalage du pic correspondant à la protéine dans le spectre. A l'aide des données génomiques, nous avons pu identifier ces 4 pics comme correspondant à une protéine ribosomique, RpmD. La description de l'identification de ces *peak shift* et de l'utilisation des génomes pour l'identification de ces derniers est décrite dans la partie suivante, car ils sont la base de la méthode finalement retenue pour la caractérisation des isolats de terrain.

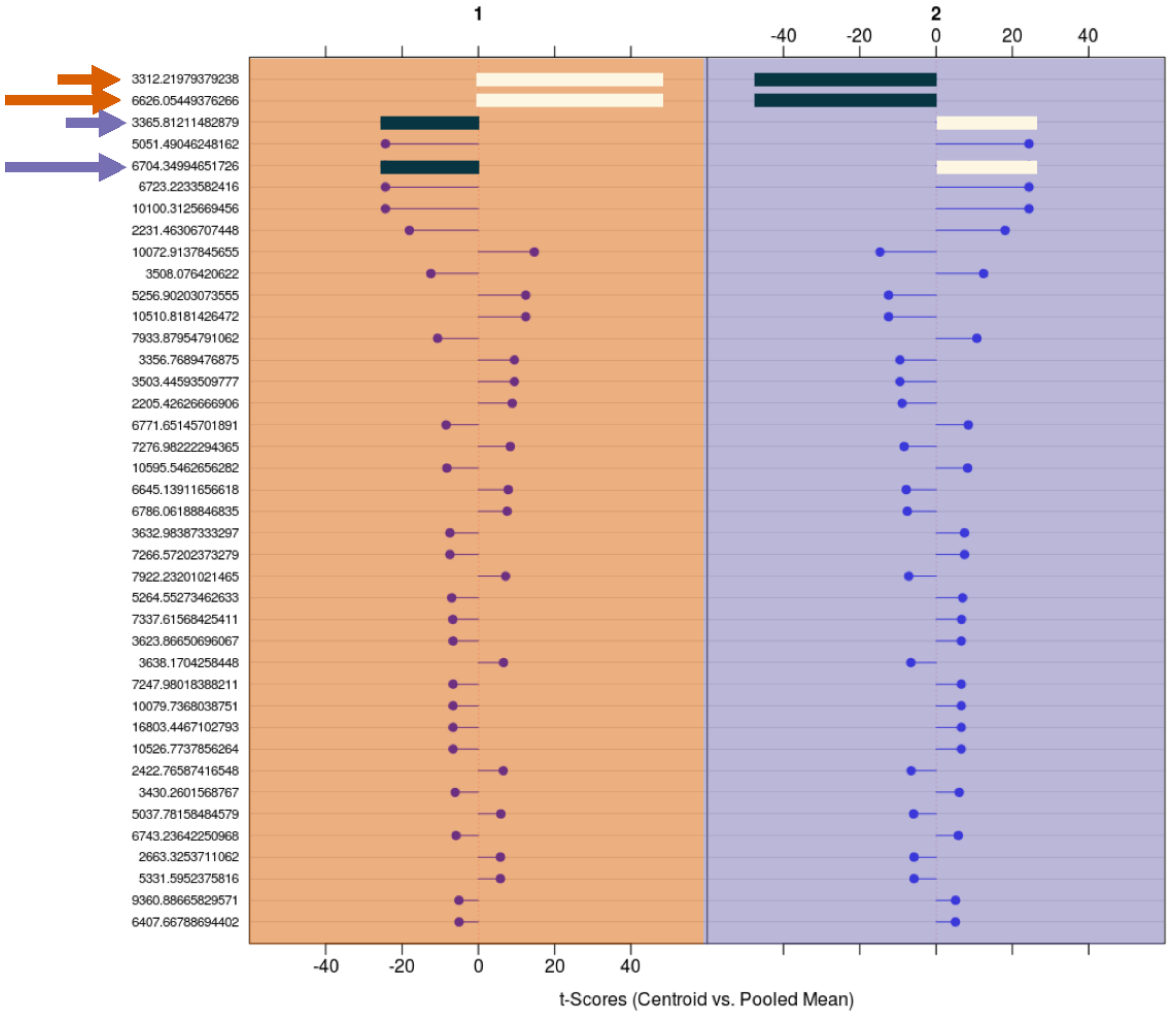


Figure 20 : Les 40 pics les plus discriminants selon une classification binaire

Les couleurs orange et violette reflètent la classification binaire proposée Figure 18. Les flèches indiquent un *peak shift* d'une protéine ribosomique. Celui-ci est observé à la masse théorique attendue (flèche longue) ainsi qu'a une valeur m/z deux fois plus petite correspondant à la protéine dichargée. Les barres blanches indiquent la présence du pic tandis que les noires indiquent l'absence.

Malgré les résultats encourageants, cette approche ne convient pas vraiment pour développer une approche diagnostique utilisable avec des isolats de terrain. Plusieurs facteurs sont à prendre en compte. Premièrement, la construction de l'arbre et l'emplacement d'un échantillon dans celui-ci nécessitent d'avoir l'intégralité du jeu de données. Deuxièmement, la classification proposée n'est valable en théorie que pour le jeu de données utilisé pour construire l'arbre. Enfin, la classification est sensible à différents paramètres qu'il est difficile de maîtriser. Outre la complexité propre au prétraitement des spectres MALDI-TOF (qui influence énormément les pics détectés), la détection des pics est aussi une étape difficile lors du traitement des spectres. De nombreux paramètres peuvent être changés comme le rapport signal sur bruit et des filtres peuvent être appliqués (comme le seuil de fréquence minimum pour qu'un pic soit conservé). Tout cela nous donne finalement une liste de pics dont il est difficile de distinguer automatiquement l'information pertinente du bruit de fond, d'autant plus à l'échelle du typage d'un isolat. Afin de résoudre tous ces problèmes, nous avons donc développé une méthode basée sur des biomarqueurs prédéfinis (des *peak-shift*) : le Multi Peak Shift Typing qui fait l'objet du second article.

C – Typage par biomarqueurs, *Multi Peak Shift Typing*

C1 – Second article : Développement d'un schéma de typage pour l'espèce T. maritimum et génomique comparée

Bridel *et al. Vet Res* (2020) 51:60
https://doi.org/10.1186/s13567-020-00782-0

VETERINARY RESEARCH

RESEARCH ARTICLE Open Access

Genetic diversity and population structure of *Tenacibaculum maritimum*, a serious bacterial pathogen of marine fish: from genome comparisons to high throughput MALDI-TOF typing

Sébastien Bridel[1,2,3], Frédéric Bourgeon[4], Arnaud Marie[2], Denis Saulnier[5], Sophie Pasek[6], Pierre Nicolas[7], Jean-François Bernardet[1] and Eric Duchaud[1*]

Abstract

Tenacibaculum maritimum is responsible for tenacibaculosis, a devastating marine fish disease. This filamentous bacterium displays a very broad host range and a worldwide geographical distribution. We analyzed and compared the genomes of 25 *T. maritimum* strains, including 22 newly draft-sequenced genomes from isolates selected based on available MLST data, geographical origin and host fish. The genome size (~3.356 Mb in average) of all strains is very similar. The core genome is composed of 2116 protein-coding genes accounting for ~75% of the genes in each genome. These conserved regions harbor a moderate level of nucleotide diversity (~0.0071 bp^{-1}) whose analysis reveals an important contribution of recombination ($r/m \geq 7$) in the evolutionary process of this cohesive species that appears subdivided into several subgroups. Association trends between these subgroups and specific geographical origin or ecological niche remains to be clarified. We also evaluated the potential of MALDI-TOF-MS to assess the variability between *T. maritimum* isolates. Using genome sequence data, several detected mass peaks were assigned to ribosomal proteins. Additionally, variations corresponding to single or multiple amino acid changes in several ribosomal proteins explaining the detected mass shifts were identified. By combining nine polymorphic biomarker ions, we identified combinations referred to as MALDI-Types (MTs). By investigating 131 bacterial isolates retrieved from a variety of isolation sources, we identified twenty MALDI-Types as well as four MALDI-Groups (MGs). We propose this MALDI-TOF-MS Multi Peak Shift Typing scheme as a cheap, fast and an accurate method for screening *T. maritimum* isolates for large-scale epidemiological surveys.

Introduction

The rapid development of intensive aquaculture has been associated with a dramatic increase in outbreaks of infectious diseases [1, 2]. Additionally, the international spread of pathogens through the trade of fish and eggs or as a response to environmental changes has been documented for some important viruses and bacteria [3, 4]. In this context, the success and sustainability of aquaculture largely depend on the understanding of the evolution and epidemiology of pathogens [1]. Among those, several species of the genus *Tenacibaculum* (family *Flavobacteriaceae*, phylum *Bacteroidetes*) are responsible for diseases collectively designated as tenacibaculosis, a very serious bacterial condition of many commercial marine

*Correspondence: eric.duchaud@inrae.fr
[1] Université Paris-Saclay, INRAE, UVSQ, VIM 78350 Jouy-En-Josas, France
Full list of author information is available at the end of the article

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

fish species leading to considerable economic losses [5, 6]. *Tenacibaculum maritimum* (formerly *Flexibacter maritimus*) was the first species to be characterized and probably the best-known pathogen in the genus. Moreover, *T. maritimum* can affect many feral, captive, and cultured fish species [5, 7] and has been repeatedly identified in many marine aquaculture systems worldwide. Diseased fish usually exhibit a diversity of external symptoms including corroded mouth, skin ulcers, fin necrosis, and rotted tail. Skin lesions are often colonized by opportunistic pathogens such as *Vibrio* spp. So far, only one vaccine is commercially available, but it is restricted to the protection of turbot. Hence, the control of *T. maritimum* outbreaks essentially relies on the use of antibiotics, sometimes combined with external disinfectants [8].

Reliable methods for studying the relationships between isolates of the same bacterial species (i.e., strain typing) are a key step for understanding the population structure, the spreading and the epidemiology of pathogens. Different typing methods have been proposed for epidemiological investigations of *T. maritimum* [9]. Three serotypes displaying varying degrees of association with host fish species have been reported [10, 11] and different molecular technics have been used to determine the intraspecific diversity of *T. maritimum* [12, 13]. Serological data were compared with several PCR-based methods [14]. In 2014, we proposed a Multi Locus Sequence Analysis (MLSA) scheme [15] that proved to be a powerful discriminating tool for isolate identification and taxonomic affiliation [16, 17]. MLSA revealed an unforeseen diversity including several, yet undescribed, pathogenic species in Norway [18]. In addition, this 11-locus sequenced-based method revealed a high number of distinct genotypes for the species *T. maritimum*, suggesting an endemic distribution of strains without significant contribution of long-distance dissemination linked to international fish movements. More recently, whole-cell matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) was used for the differentiation of several fish-pathogenic *Tenacibaculum* species [14, 19]. However, these studies did not reveal any relationships between the proteomic profiles and the source of isolation of the strains for any of the *Tenacibaculum* species analyzed, and no biomarker below the species level (e.g., serotype-specific peaks) was detected for *T. maritimum*. Meanwhile, the complete genome of the *T. maritimum* type strain [20] as well as the draft genomes of the type strains and several field isolates of *T. dicentrarchi* and "*T. finnmarkense*" have been recently published [21], paving the way to comparative genomics.

In this study, we analyzed and compared 25 genomes (including 22 newly draft-sequenced) of *T. maritimum* isolates from various geographical origins and host fish species to draw a global picture of the genomic diversity of the species. In addition, we developed a genome-based MALDI-TOF MS scheme and typed 131 field isolates. While this technique is commonly used for species identification, bacterial characterization below the species level is much more challenging, requiring the identification of subtle differences between strains [22]. We also propose a dedicated website that allows both identification and typing of new *Tenacibaculum* isolates [23].

Materials and methods

Bacterial strains

Tenacibaclum maritimum strains were grown in marine 2216E broth (Difco, Becton, Dickinson and Co., Franklin Lakes, New Jersey, USA) for 24 h at 28 °C and 170 rpm. Stock cultures were preserved in marine 2216E broth containing 20% (v/v) glycerol at −80 °C. The 25 strains used in this study are listed in Table 1 and the bacterial isolates subjected to MALDI-TOF MS analysis are listed in Additional file 1.

Genome sequencing, assembly and annotation

Following centrifugation of the liquid culture, genomic DNA was extracted from the pellet using the Wizard genomic DNA purification kit (Promega, Madison, Wisconsin, USA). The genomes were sequenced using Illumina (HiSeq, 100 paired-end or MiSeq, 300 paired-end) and genome assemblies were performed using Spades and Velvet on the PATRIC website with default settings [24]. The resulting contigs (>2000 bp) were integrated into the MicroScope platform [25].

Genome analysis and comparisons

Average Nucleotide Identity analyses were performed using the ANIm method [26] with the Python module Pyani [27] using proposed threshold of≈95–96% for species delineation. Genome annotation, including manual curation, and comparisons, including pan and core genome computation, were performed using the web interface MicroScope [28], which allows graphic visualization enhanced by a synchronized representation of synteny groups [29]. Persistent, shell and cloud genomes were computed using the PPanGGOLiN 0.1.4 software [30]. Considering the low levels of sequence divergence at typical core genome loci previously reported for *T. maritimum* [15], we chose a cutoff of 80% identity and 80% on the minimal coverage of the length between the aligned portions of two proteins to determine whether two CDSs were members of the same gene family.

The 25 genomes were aligned using Snippy [31] with the genome of strain NCIMB 2154^T (the type strain of *T. maritimum*) [20] serving as a reference. The resulting whole-genome alignment was used for phylogenetic

Table 1 General genome features.

Strain	Country	Host	Tissue	Date of isolation	Technology	Reads (post-trimming)	Contigs (>2000pb)	Total length	GC %	Coverage	Number of predicted genomic islands	SNPs vs. NCIMB 2154^T	Predicted CDS	Genbank Assembly
NCIMB 2154T	Japan	*Pagrus major*	Skin	1977	[20]	n/a	1	3453 971	32.01	1734	29	0	2774	GCA_900119795.1
TM-KORJJ	Korea	*Paralichthys olivaceus*	n/a	n/a	PacBio	n/a	1	3333 272	31.98	300	24	15 049	2735	GCA_004803875.1
NBRC 15946	Japan	n/a	n/a	n/a	HiSeq	n/a	96	3 240 791	31.80	123	20	10 084	2692	GCA_000509405.1
P1-39	France	*Dicentrarchus labrax*	Liver	2010	HiSeq (2 × 100 bp)	56 562 140	104	3 337 690	31.79	93	23	15 401	2788	GCA_902705535
P4-45	France	*Dicentrarchus labrax*	Skin	2010	HiSeq (2 × 100 bp)	54 681 392	73	3 349 546	31.81	70	27	15 707	2843	GCA_902705495
902	France	*Dicentrarchus labrax*	Skin	2013	HiSeq (2 × 100 bp)	88 844 854	68	3 372 337	31.80	85	24	15 521	2851	GCA_902705365
Aq16-85	French Polynesia	*Platax orbicularis*	Skin	2016	MiSeq (2 × 300 bp)	2 174 369	43	3 198 696	31.88	220	21	14 929	2664	GCA_902705305
Aq16-88	French Polynesia	*Platax orbicularis*	Skin	2016	MiSeq (2 × 300 bp)	3 434 486	45	3 196 671	31.88	301	21	14 966	2665	GCA_902705275
Aq16-89	French Polynesia	*Platax orbicularis*	Skin	2016	MiSeq (2 × 300 bp)	2 156 781	45	3 196 642	31.89	171	21	14 914	2666	GCA_902705375
TFA4	French Polynesia	*Platax orbicularis*	Skin	2013	MiSeq (2 × 300 bp)	2 369 550	66	3 356 632	31.82	226	33	14 961	2865	GCA_902705565
FS08(1)	Italy	*Sparus aurata*	Skin	2006	MiSeq (2 × 300 bp)	1 256 466	54	3 399 437	31.81	87	30	4424	2866	GCA_902705395
NAC SLCC MFF	Malta	*Dicentrarchus labrax*	Skin	1995	MiSeq (2 × 300 bp)	1 091 150	80	3 352 671	31.81	76	29	40 598	2795	GCA_902705345
USC SP9.1	Spain	*Salmo salar*	Skin	1993	MiSeq (2 × 300 bp)	727 120	80	3 395 385	31.84	46	30	31 252	2779	GCA_902705515
DPIF 89/3001-6.2	Tasmania	*Latris lineata*	Skin	1989	MiSeq (2 × 300 bp)	1 019 320	129	3 448 890	31.77	66	33	26 962	2788	GCA_902705315

Table 1 (continued)

Strain	Country	Host	Tissue	Date of isolation	Technology	Reads (post-trimming)	Contigs (>2000pb)	Total length	GC %	Coverage	Number of predicted genomic islands	SNPs vs. NCIMB 2154^T	Predicted CDS	Genbank Assembly
DPIF 89/0239-1	Tasmania	*Salmo salar*	Skin	1989	MiSeq (2 × 300 bp)	1 496 188	55	3 353 931	31.90	92	28	17 743	2773	GCA_902705355
USC SE30.1	Spain	*Oncorhynchus kisutch*	Mouth	1993	MiSeq (2 × 300 bp)	749 406	111	3 544 405	31.75	52	30	31 928	2928	GCA_902705525
UCD SB2	California	*Atractoscion nobilis*	n/a	1995	MiSeq (2 × 300 bp)	1 323 140	50	3 308 376	31.91	85	23	17 064	2715	GCA_902705445
JIP 32/91-4	France	*Dicentrarchus labrax*	Skin	1991	MiSeq (2 × 300 bp)	1 031 822	59	3 447 003	31.80	59	31	17 265	2872	GCA_902705385
CVI1001048	Holland	*Solea solea*	Skin	2010	MiSeq (2 × 300 bp)	1 242 180	42	3224 047	31.94	90	16	1659	2670	GCA_902705265
FC	Chile	*Scophthalmus maximus*	Eye	1998	MiSeq (2 × 300 bp)	1086 816	57	3 505 634	32.01	70	27	16 571	2921	GCA_902705415
P2-48	France	*Solea senegalensis*	Skin	2010	MiSeq (2 × 300 bp)	1 349 672	55	3 418 994	31.86	90	34	43 607	2910	GCA_902705555
P2-27	Spain	*Scophthalmus maximus*	Skin	2011	MiSeq (2 × 300 bp)	1 087 036	88	3 371 677	31.82	76	24	19 639	2821	GCA_902705465
JIP 46/00	France	*Scophthalmus maximus*	Skin	2000	MiSeq (2 × 300 bp)	1 155 334	56	3 371 335	31.89	73	25	17 215	2781	GCA_902705435
JIP 10/97	France	*Scophthalmus maximus*	Skin	1997	MiSeq (2 × 300 bp)	1 345 346	52	3 333 073	31.86	86	22	17 441	2746	GCA_902705285
NCIMB 2158	Scotland	*Solea solea*	Skin	1981	MiSeq (2 × 300 bp)	1 138 826	74	3 369 590	31.87	79	28	17 387	2797	GCA_902705425

The list of contributors is available in Additional file 1.

tree reconstruction using Gubbins [32] and the Phylip package version 3.6 released by Felsenstein in 2005, available online at [33] and originally released in 1980 [34]. Regions of high diversity (high number of single nucleotide polymorphisms [SNPs]) presumably linked to recombination events were masked at each Gubbins iteration. A Maximum Likelihood tree was built at the fifth iteration using non-masked polymorphic sites. Neighbor-joining and parsimony trees were obtained with Phylip suite v3.696 (programs dnadist, neighbor, and dnapars) after removing positions with gaps in the alignment. Custom Perl and R scripts were used for the nucleotide diversity analysis (computation of the average pairwise nucleotide diversity and of the homoplasy index) and graphical representations (including R library "ape" for the drawing of phylogenetic trees). An estimate of the ratio of recombination and mutation (r/m) based on the analysis of SNPs between pairs of closely related isolates was obtained using a two-state hidden-Markov model (HMM), as described in Duchaud et al. [35].

Preparation of bacterial samples for MALDI-TOF MS

The ethanol/formic acid extraction procedure, as described by Mellmann et al. [36] was used. Briefly, a colony of a fresh overnight culture was inoculated in 5 mL marine 2216E broth and cultivated at 28 °C for 24 h with shaking (170 rpm). The resulting bacterial culture was centrifuged at 12 000*g* in a desktop centrifuge for 2 min and the supernatant discarded. About 10 mg of the resulting bacterial pellet was transferred in a clean Eppendorf tube with 300 µL of ultra-pure water (Acros organics, New Jersey, USA) and vigorously mixed to resuspend the cells. 900 µL of 100% ethanol (VWR Chemicals, Radnor, Pennsylvanie, USA) was added into the tube and mixed again. The mix was directly processed or kept at room temperature (RT) up to 1 month for the assessment of the ethanol-fixed bacteria protocol. The tube was centrifuged at 12 000*g* in a desktop centrifuge for 2 min and the supernatant discarded. The tube was centrifuged for 2 additional minutes and the residual ethanol removed and the pellet was dried at RT. Thirty µL of 70% formic acid was added and mixed thoroughly by pipetting. An equal volume of acetonitrile was then added to the tube, mixed carefully and then centrifuged at 12 000*g* in a desktop centrifuge for 2 min. One µL of the supernatant was dropped onto a 96-spot polished steel target. The sample spot was dried at RT and 1 µL of matrix solution (α-cyano-4-hydroxycinnamic acid in 50% acetonitrile, 47.5% water and 2.5% trifluoroacetic acid) was then added. The sample spot was finally air dried again before analysis. A calibration of the MALDI-TOF mass spectrometer was performed using Bruker bacterial test standard for each series of acquisitions with a mass tolerance limit of ± 300 parts-per-million (ppm). To evaluate alternative sample preparation procedures, bacteria were cultivated on solid medium [i.e., marine 2216E broth and 15 g/L agar (Invitrogen, Illkirch, France)] at 28 °C for 24 h. About 10 bacterial colonies were collected and processed as the bacterial pellet obtained from liquid culture. For the assessment of the direct colony picking protocol, a single bacterial colony was picked with a sterile toothpick and directly deposited onto a 96-spot polished steel target and processed as the sample spot previously mentioned.

MALDI-TOF MS data acquisition

A MALDI Biotyper Microflex LT controlled by Compass flexControl software (version 3.4; Bruker Daltonics, Billerica, Massachusetts, USA) was used to generate mass spectra for all isolates. Mass spectra were acquired using automatic mode and default settings (2000 to 20 0000 Da; linear positive mode; 240 laser shots). For each isolate, twelve spectra (four spots of each isolate extracted were measured 3 times) were recorded.

Peak shift characterization using genomic data

Ribosomal protein sequences were retrieved from the MicroScope annotation platform using the 25 available *T. maritimum* genomes. The theoretical mass weight for each sequence was computed with the *mw*() function from the Peptides R-package and TermiNator [37] was used to predict the first methionine cleavage resulting in a theoretical loss of 131.2 Da. The theoretical masses obtained were compared to the peak list and spectra were screened to retrieve peak shifts using the predicted masses of presumptive polymorphic biomarkers. Several other peak shifts were identified during this step but were not kept for typing purpose because of the stringent criteria used (see section "Results").

MALDI-TOF data analysis algorithm and implementation in R

We used different R packages dedicated to mass spectrometry analysis from the BioConductor repository [38], i.e. MALDIquant Foreign, MALDIquant, MALDIrppa and MassSpecWavelet. Using these packages, any type of MALDI-TOF data can be loaded from any manufacturer. Raw data generated by Microflex® Bruker were imported into R environment by MALDIquant Foreign. The spectra pre-processing steps (i.e., intensity transformation, baseline correction, intensity correction, spectra alignment, peak detection), curve smoothing and peak detection on average spectra (mean spectra) were performed using MALDIquant, MALDIrppa and MassSpecWavelet [39], respectively. This method is based on local maxima detection at different scales, and the retained peaks are

observed at many scales as true maxima (true peaks). Thus, it theoretically gets rid of noisy peaks more efficiently than classical algorithms (e.g., about 300 peaks were detected with MALDIquant/MALDIrppa with soft parameters whereas around 100 peaks were detected with MassSpecWavelet for each *T. maritimum* average spectrum). The resulting peak list is then compared to a reference peak list that encompasses either: (i) full spectra of the type strains of *Tenacibaculum* species, (ii) species-specific biomarkers or (iii) subtyping biomarkers. We considered two peaks as matching between a sample and the reference with a tolerance of 700 ppm. MALDIquantTypeR is a home-made R application developed for MALDI type assignment [23].

Results

General genome features

The origins of the sequenced genomes, the main sequencing and assembly data and the genomic characteristics are listed in Table 1. The number of contigs (> 2 kb) for the 22 newly sequenced genomes varied from 42 to 129 depending on sequencing technology, sequencing depth and genome properties (i.e., number of repeats). Genome length estimated by the cumulated size of the contigs was 3356 Kb in average with very little variations between isolates (standard deviation (SD) = 92 Kb). Each genome contained an average of 2788 (SD = 81) predicted CDSs. The core-genome was composed of 2116 gene families (out of 5809 gene families in total), representing about 75% of the CDSs in each genome. The core-genome encompassed all the predicted toxins and virulence factors (i.e., cholesterol-dependent cytolysin, collagenase, sphingomyelinase, ceramidase, chondroitin AC lyase, streptopain family protease, sialidase, iron uptake systems and T9SS components) previously identified in strain NCIMB 2154^T [20]. In addition, the persistent-genome, equivalent to a relaxed core-genome (i.e., genes conserved in all but a few genomes) encompassed about 2500 gene families representing about 81% of the CDSs. There were about 10% of shell-genome genes (i.e., genes having intermediate frequencies corresponding to moderately conserved genes potentially associated to environmental adaptation capabilities) and about 9% of cloud-genome (i.e., genes found at a very low frequency). These two later values were likely overestimated since the newly sequenced genomes were not fully assembled and because of pseudogenization events and gaps between contigs that lead to gene fragments that artificially increase the total number of shell-genome and cloud-genome gene families. Strikingly, most of the shell- and cloud-genome genes were encompassed in genomic islands (region of genomic plasticity). Each genome contained an average of 26 (SD = 4.65) genomic islands (Table 1). Interestingly, strain FC isolated in Chile encompassed a unique (i.e., not identified in other strains), 83-kb long (MARITFC_v1_10015 to MARITFC_v1_10098) genomic island that displayed prophage characteristics (i.e., located at an Arg tRNA and bordered with an integrase/recombinase encoding gene). This island contained many genes predicted to be involved in heavy metal resistance, including genes encoding proteins of the copper oxidase family and different cobalt-zinc-cadmium efflux pumps.

We focused on the two fully PacBio-assembled genomes (i.e., strains NCIMB 2154^T and TM-KORJJ) contained in our dataset to accurately identify genes encompassed in these islands. Strain NCIMB 2154^T contained 29 genomic islands (encompassing 384 genes and corresponding to 13% of the total number of CDSs) while strain TM-KORJJ contained 24 genomic islands (encompassing 351 genes and corresponding to 12% of the total number of CDSs). Most of these islands displayed classical prophage characteristics and encompassed CDSs encoding for phage structural proteins, integrases, transposases, insertion sequences, restriction/modification systems and Vgr/Rhs elements or their scars.

Nucleotide diversity and population structure

To estimate the nucleotide diversity, we first used assembled draft genomes and computed the average nucleotide identity (ANI) between pairs of genomes. ANI values ranged between 98.18% and 100% (Additional file 2), far above the species delineation threshold of 95–96% [26], revealing a cohesive species and confirming that all the isolates included in this genomic study indeed belonged to the species *T. maritimum*.

Using a whole genome alignment built on strain NCIMB 2154^T as a reference and corresponding to 2587 083 bp of conserved concatenated sequence, we identified a total of 86 217 SNPs (mean 21 126 SNPs ± SD 9795) out of which 84 694 (98.2%) where bi-allelic. The average pairwise nucleotide divergence between isolates (Table 1) amounted to 18 496 SNPs corresponding to a diversity π of 0.0071 bp^{-1} with a maximum of 0.0152 bp^{-1}. The smallest divergence was 0.0021 bp^{-1} (5477 SNPs), reflecting an absence of very closely related isolate pairs outside of the three isolates Aq18-85, Aq18-88 and Aq18-89 (no SNPs found between them).

The whole genome alignment was subjected to several tree reconstruction methods. Trees obtained by parsimony or with the more sophisticated Gubbins method that attempts to remove recombination tracts are shown in Additional files 3 and 4, respectively. These trees tended to contain internal branches of substantial length with poor bootstrap values (see parsimony tree S3A) presumably due to recombination. To reflect more faithfully

the pairwise distances between isolates, Figure 1 presents a neighbor-joining tree based on a simple Jukes-Cantor distance. In this tree, whose topology is very similar to those of the Parsimony and Gubbins trees, branches with poor bootstrap support tended to disappear (i.e. to have length close to zero). Overall, the topologies and bootstrap values supported the division in three subgroups (corresponding to clades A, B and C) previously observed using MLST data [15]. Clades A, B and C encompassed 3, 20 and 2 strains, respectively. The core genome-based tree and the previously obtained MLST-based tree [15] showed similar clade composition with

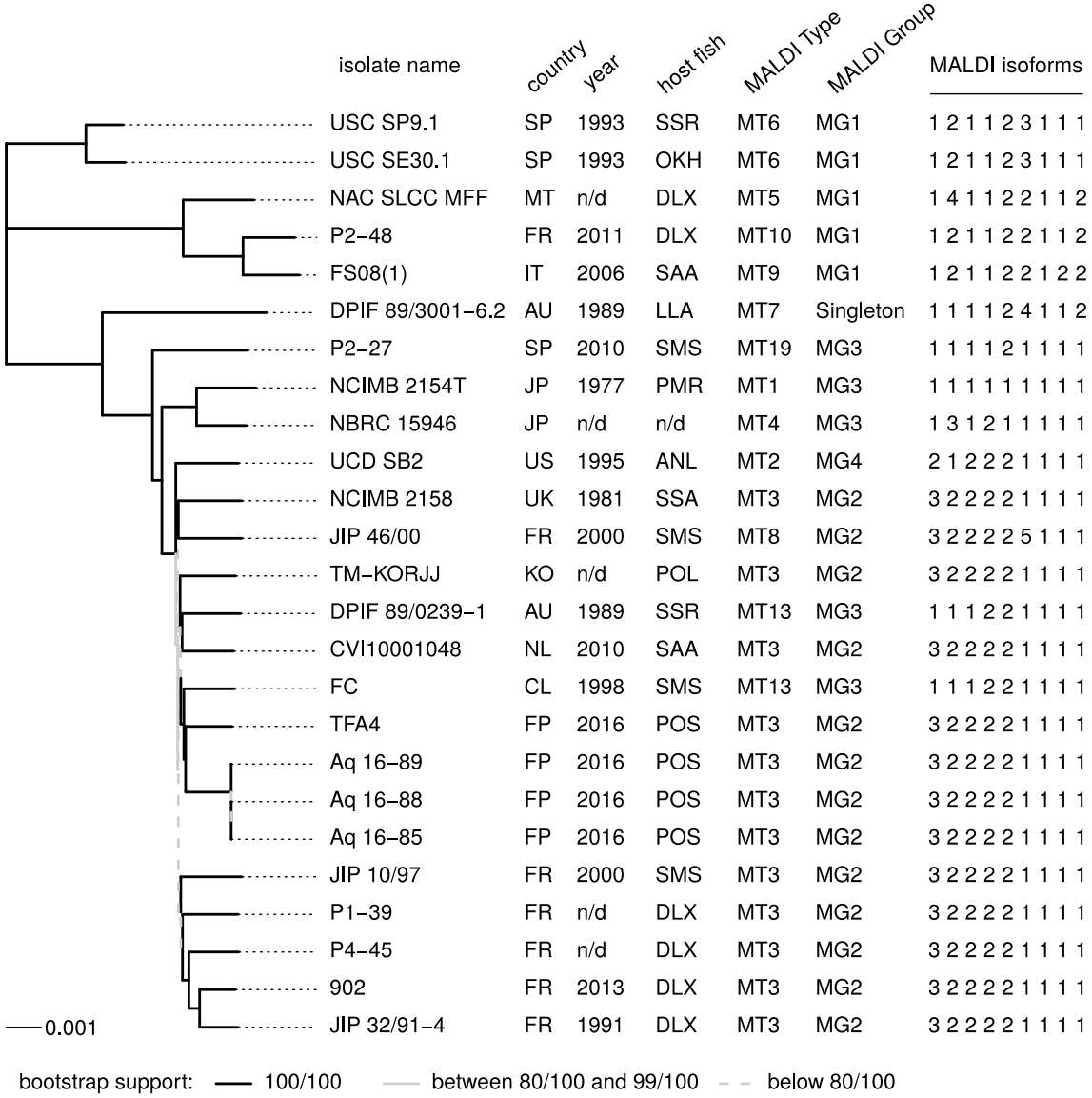


Figure 1 Neighbor-joining tree based on a Jukes-Cantor distance of the 25 *T. maritimum* genomes. The tree is based on snippy alignment without gap regions. It is reconstructed by NJ method using a Jukes-Cantor distance. 100 bootstrap replicates were made. A black line denotes a branch support of 100/100, while a grey line denotes a value between 80/100 and 99/100 and a grey dotted line a value below 80/100. The origin of strains is indicated next to the tree and the meanings of each abbreviation are detailed below. The MALDI-Type, the MALDI-Group and the corresponding MALDI isomorphic profile are plotted on the right hand side of the figure. Origin: Australia (Tasmania), AU; Chile, CL; France, FR; French Polynesia, FP; Italy, IT; Japan, JP; Malta, MT; Spain, SP; United Kingdom (Scotland), UK; USA (California), US; no data available, n/d. Fish host species: ASI, *Acanthopagrus shlegeli*; ANS, *Atrasctoscion nobilis*; CTS, *Carcharias taurus*; DLX, *Dicentrarchus labrax*; EMX, *Engraulis mordax*; LLA, *Latris lineata*; OKH, *Onchorynchus kisutch*; OMS, *Onchorynchus mykiss*; PMR, *Pagrus major*; POL, *Paralichthys olivaceus*; POS, *Platax orbicularis*; SSR, *Salmo salar*; SMS, *Scophtalmus maximus*; SQA, *Seriola quinqueradiata*; SSS, *Solea senegalensis*; SAA, *Sparus aurata*; n/d no data available.

the only exceptions of strains DPIF 89/3001-6.2 and DPIF 89/0239-1 that belong to ST24 and ST20, respectively (MLST subgroup C) whereas they belong to core genome clade B. However, the position of these two strains had very poor bootstrap support in the MLST-based tree.

Pervasive recombination was obvious from the difference between the 158 863 changes along the parsimony tree and the 87 770 changes that would have been needed in the absence of homoplasy and recombination at the 86 217 polymorphic positions (considering 1493 tri- and 30 quadri-allelic SNPs). These numbers corresponded to an apparent homoplasy index HI of 44.8% [(158 863–87 770)/158 863]. A similar HI of 44.3% [(102 755–57 185)/102 755] was obtained when examining polymorphism only within clade B (the average pairwise nucleotide divergence measured in this clade was 0.0042 bp^{-1}). To further quantify the impact of recombination, we analyzed the pattern of pairwise nucleotide divergence between closely related genomes, selecting unambiguous pairs of tips in the reconstructed trees such as to be able to distinguish private and shared polymorphism. In practice, we used for this purpose the comparison USC SP9.1 vs. USC SE30.1 (5615 SNPs) and FS08(1) vs. P2−48 (7980 SNPs). No isolates from clade B satisfied our needs of forming a clearly isolated pair. Only bi-allelic SNPs were used, and the fraction of shared polymorphism (i.e. polymorphism also found outside the pair) represented as much as 78.8% of the sites that distinguished the first pair of isolates and 69.3% for the second pair, indicating a considerable contribution of recombination to divergence since mutation is expected to produce almost exclusively private polymorphism while recombination is expected to produce tracts mixing shared and private polymorphism. A hidden Markov model served to delineate recombination tracts and lead to estimates of the ratio of recombination and mutations to nucleotide-level divergence (r/m) of 20.6 for USC SP9.1 vs. USC SE30.1 and 7.7 for FS08(1) vs. P2−48 (Additional files 5A and B, respectively).

As observed on the phylogenetic trees, strains retrieved from the same geographical origin tended to cluster together, indicating genetic relatedness. Indeed, the two strains originating from the Atlantic coast of Spain (USC SE30.1 and USC SP9.1) formed cluster C while the three strains in group-A [FS08(1), NAC SLCC MFF and P2-48] were all retrieved from countries bordering the Mediterranean Sea. The two Japanese strains NCIMB 2154^T and NBRC 15946 also appeared related. The three strains Aq16-85, Aq16-88 and Aq16-89 isolated in French Polynesia in 2016 were virtually identical to each other (0 SNPs in the 2587 083 aligned positions of the core-genome), suggesting a dissemination of a single clone between fish farms where *Platax orbicularis* are raised. In addition, these three strains were related to strain TFA4 also isolated in French Polynesia in 2013. The Chilean strain FC, also originating from the South Pacific, belonged to the same group. Five strains (i.e., JIP 10/97, P1-39, P4-45, 902 and JIP 32/91-4) retrieved from France over 22 years were also grouped together in the NJ tree (Figure 1).

Identification of potential biomarkers in genomic data

Because genomic data pointed to a cohesive species but still displaying variability to some extent, we aimed to evaluate the potential of MALDI-TOF MS for rapid screening and typing using specific sets of biomarker ions. Since half of the detected peaks in bacterial MALDI-TOF MS spectra correspond to ribosomal proteins [40], we focused on this protein set. Using genome sequence data, we first retrieved all deduced ribosomal protein sequences and predicted the first methionine removal using the Terminator software [37, 41]. The molecular weight of each deduced ribosomal protein was computed. We retrieved ribosomal protein sequences without polymorphism (i.e., invariant ribosomal protein sequences hereafter designed as monomorphic) that could serve as species biomarkers and for MALDI-TOF MS spectra internal calibration. We also retrieved ribosomal protein sequences displaying polymorphism (i.e., ribosomal protein sequences with variation hereafter designed as polymorphic) since they are likely relevant biomarkers for strain typing. Strikingly, one-third (18/54) of the ribosomal proteins were monomorphic while the others (36/54) displayed some degree of amino-acid polymorphism that mostly gave rise to a mass change (the resulting information is summarized in Additional file 6). Strikingly, the topology of the hierarchical classification tree deduced from this data set was globally congruent with that of the trees obtained using the core genome genes (Figure 1 and Additional file 3). Indeed, some ribosomal protein-encoding genes displayed clade-specific variations. For example, gene *rplU* encoded two different versions of the 50S ribosomal subunit protein L21: a 209 amino-acid long RplU protein in clades A and C and a 161 amino-acid long RplU protein in clade B. Other examples were provided by proteins RpmI and RpsP that displayed isoforms only found in clade A or by RplD that displayed an isoform only found in clade C.

Selection of invariant biomarker ions for internal calibration of spectra and species identification

Using our monomorphic biomarker candidates, we performed visual inspection of spectra obtained from 24 out of the 25 genome-sequenced strains to identify the corresponding peaks. We selected 18 peaks (Additional file 7) according to the following stringent criteria: (i) they

covered as much of the entire spectra as possible (ranging from 2958 m/z to 12 460 m/z); (ii) they corresponded to mono-, di- or tri-charged monomorphic ribosomal proteins; (iii) they occurred in all MALDI-TOF spectra of the 24 isolates; (iv) most of them were of high intensity (ranging from 98 to 1371, mean intensity = 547), even at both ends of the spectra (though intensity was globally lower at these m/z locations); and (v) they were not disrupted by the presence of other very close peaks which could degrade peak detection reliability; in other words, retained peaks had to be in a window devoid of additional peaks. Only two selected peaks (i.e., RpmJ-M-H1 and RpmE-M-H2) represented exceptions to the latter criterion. Indeed, strains USC SE30.1 and NCIMB 2158 displayed a slightly different peak shape for RpmJ-M-H1 and RpmE-M-H2, respectively, with a larger area under the curve and a flatter bump. This resulted from the presence of two close peaks with the same intensity. However, these peaks did not disrupt the signal and the peaks for RpmJ-M-H1 and RpmE-M-H2 could both be accurately detected. The 18 selected peaks corresponded to 9 ribosomal proteins with varying degrees of ionization. They could serve as internal calibration references using the *alignSpectra* function for accurate peak detection. In addition, these monomorphic biomarkers were of utmost interest for species identification and were included in the quality control process of spectra (Additional file 8).

Selection of 9 polymorphic biomarkers for strain typing and MALDI-Type attribution

Using our polymorphic biomarker candidates (Additional file 6), we performed visual inspection of the above-mentioned spectra to identify the corresponding peak shifts. We selected 8 polymorphic biomarkers corresponding to ribosomal proteins (i.e., RpmD, RpmC, RpsP, RpsN, RpsO, RpsQ, RplX and RplT) with a molecular weight ranging from 6600 Da to 13 200 Da. An additional polymorphic biomarker, RpsT, was included after screening our collection of 131 isolates (see next paragraph). Indeed, strain Aq8-57 displayed an unexpected peak shift (not correlated to any amino-acid polymorphism observed in the 25 genomes data set). Sanger sequencing of strain Aq8-57 revealed that this peak shift corresponded to a 122A > G mutation in the sequence of the *rpsT* gene leading to a R41K change in the amino-acid sequence. In addition, strain USC RPM 539.1 also displayed a peak shift not previously observed that corresponded to a 35G > A mutation in the sequence of the *rplT* gene leading to a R12K change in the amino-acid sequence. These mutations gave rise to a −27 Da shift observed in the spectra compared to the type strain used as a reference. Therefore, the 9 retained biomarkers displayed varying degrees of polymorphism (Additional file 9) ranging from two to up to five isoforms (IF) (Table 2, Figure 2). An arbitrary number was given to each IF for subsequent analysis. By convention, an IF1 numbering was given for each biomarker of the type strain NCIMB 2154^{T}.

Validation of the typing scheme using field isolates

To validate the strain typing approach, and in addition to the 24 above-mentioned spectra, we analyzed by MALDI-TOF MS 111 field isolates from worldwide origins and retrieved from a variety of fish species (Additional file 1). Among the field isolates, 56 had previously been typed using MLST [15]. Our strategy based on automated peak detection and assignation to the corresponding IF proved to be very efficient with a success rate of 97%. Only four isolates (P1-40, Aq12-62, Aq12-64 and Aq6-9) out of 111 were not fully typed corresponding to 5 (out of 1176) peak losses (0.42%). We chose an MLST-like strategy by combining IF numbering to produce the corresponding MALDI profile as proposed by Zautner et al. [42], and here referred to as MALDI-Type (MT). Using this scheme, we identified 20 MTs (Additional file 1). Using genome sequence data, one could predict the MT of a strain; for instance, strain TM-KORJJ was predicted to belong to MT3. This approach allowed unambiguous assignment for each isolate as well as the use of visualization and analysis tools developed for MLST such as the eBurst [43] and SplitsTree decompositions [44]. Indeed, using eBurst on our dataset, the 20 MTs could be grouped in 4 clusters based on connection by single-locus-variants (Figure 3A) that can be designated as MALDI-Groups (MGs). Of note, the criterion for the determination of MGs is analogous to the criterion for the determination of clonal complexes (CC) reported in the MLST strategy [45] but does not correspond to the same level of divergence between isolates. Figure 3B is complementary to the graph drawn using eBurst. It depicts a hierarchical clustering (average-link) built from MALDI-Types isomorphic profiles with the corresponding number of isolates.

The MTs and MGs are globally congruent with the core genome-based phylogeny (Figure 1). Indeed, strains USC SP9.1 and USC SE30.1 both belonged to MT6 and were encompassed in clade C. Strains NAC SLCC MFF, FS08(1) and P2-48 were encompassed in clade A and belonged to MTs 5, 9 and 10, respectively. These 3 latter MTs were linked by single locus variants. All of the above-mentioned strains were encompassed in MG1. Strain DPIF 89/0239-1 belonged to MT7, which was the only singleton among all MGs. In the core genome-based phylogeny, strain DPIF 89/0239-1 also appeared distantly related to the other strains encompassed in clade B. Strain P2-27 belonged to MT19 whereas the type strain

Table 2 The retained polymorphic biomarkers.

Biomarkers	First methionine cleavage	H+	Predicted mass	Observed mass	Delta ppm	Isoform
RpmD	S(2) 91%	1	6700.56	6703.90	498	IF1
			6638.48	6641.21	411	IF2
			6622.48	6625.88	513	IF3
RpmC	M(1) 99%	1	7245.24	7248.41	437	IF1
			7273.26	7275.95	273	IF2
			7259.31	7262.33	416	IF3
			7301.31	7303.94	360	IF4
RpsP	P(2) 98%	2	9168.45	9169.730	139	IF1
			9197.49	9199.03	167	IF2
			9183.47	9184.97	163	IF3
			9190.48	9191.34	93	IF4
			9161.44	9163.107	181	IF5
RpsT	A(2) 97%	1	9404.57	9406.69	225	IF1
			9376	9379.19	225	IF2
RpsN	A(2) 97%	1	10049.41	10051.07	165	IF1
			10061.46	10064.14	266	IF2
RpsQ	M(1) 99%	1	10097.52	10099.99	244	IF1
			10070.50	10071.79	68	IF2
RpsO	M(1) 99%	1	10521.89	10524.30	229	IF1
			10507.82	10510.00	207	IF2
RplX	M(1) 99%	1	11117.46	11119.12	146	IF1
			11135.48	11136.94	131	IF2
RplT	P(2) 88%	1	13171.38	13172.27	67	IF1
			13157.34	13157.79	34	IF2
			13144	13145.75	57	IF3

NCIMB 2154^T belonged to MT1 and strain NBRC 15946 belonged to MT4. These three strains were encompassed in MG3. Strain UCD SB2 belonged to MT8 (MG2) and also appeared more distantly related to the other strains encompassed in clade B in the core genome-based phylogeny. Finally, all the other strains belonged to MT3 (MG2) with only two exceptions, strains FC and DPIF 89/023-9 that were the only strains displaying incongruence between MT and core genome-based phylogeny.

Although the 131 strains studied were from worldwide origin, our dataset was not suitable to highlight sound association between the isolation sources and the MT. We tried several statistical analyses (AMOVA and Fisher exact test, data not shown), but each tested variable (i.e., country and year of isolation, host fish species) was drastically correlated with each other indicating that these variables are not truly independent. For instance, a high correlation between fish hosts and countries was obvious. However, and in accordance with our previous MLST study [15], we observed trends between the geographical origin of the strains and the MT. For instance, the 4 isolates belonging to MT1 and the 3 isolates belonging to MT4 all originated from Japan. In addition, these two MTs belong to the same MG3 and are linked by Double Locus Variants (DLV). The two isolates belonging to MT2 originated from California. The two isolates belonging to MT5 originated from Italy and Malta, two neighboring countries. The two isolates belonging to MT6 originated from Spain. The three isolates belonging to MT12 originated from Tasmania. In addition, 45 out of the 50 isolates from France belonged to MT3. On the other hand, strains from the same geographical origin could belong to different MTs (e.g., the 10 Tasmanian isolates belonged to 4 different MTs whereas the 5 Italian strains each belonged to a different MT). Of importance, strains retrieved from the same host fish species could belong to different, unrelated MTs (e.g., the 74 strains retrieved from *Dicentrarchus labrax* belonged to 5 different MTs).

Evaluation of reproducibility, repeatability and alternative sample preparations for MALDI-TOF MS

In order to evaluate reproducibility and repeatability, the *T. maritimum* type strain, the 22 genome-sequenced *T.*

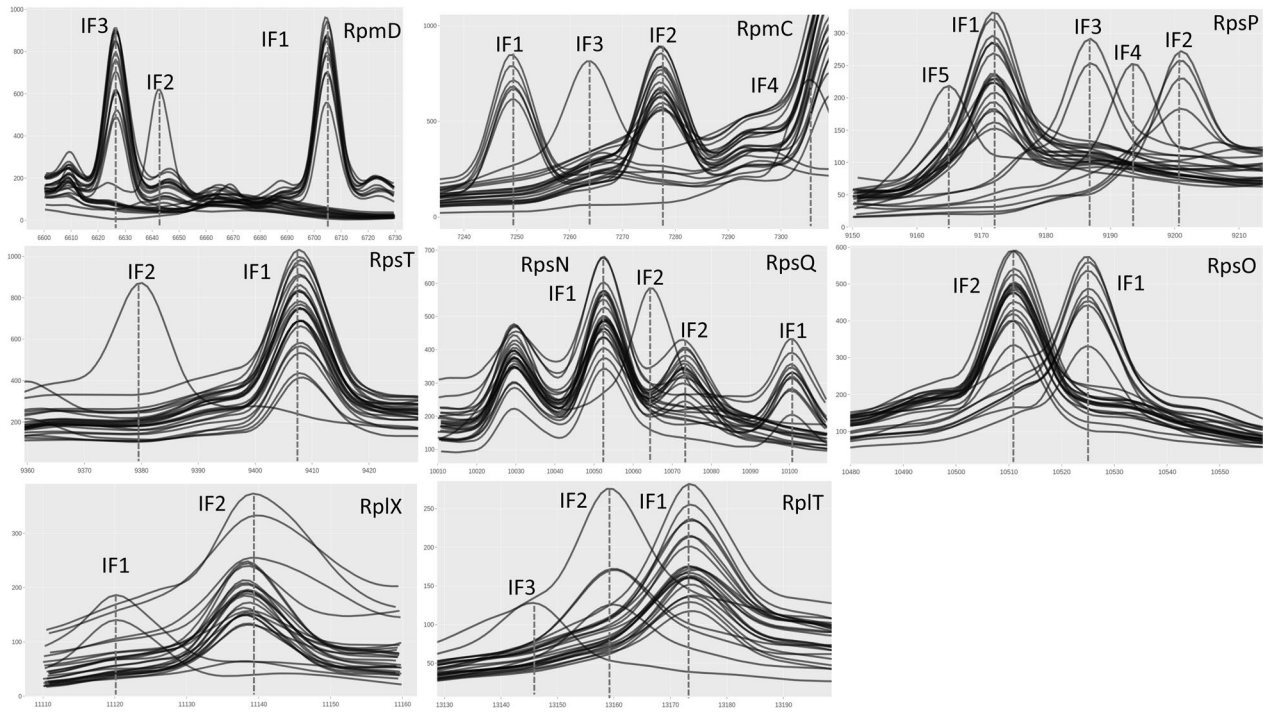


Figure 2 Characteristic peak shifts used for strain characterization and MALDI-Types definition. The 24 average spectra are plotted, corresponding to all except one sequenced isolates. Intensity varies from 200 to 1000. The RpmC IF4 (strain NAC SLCC MFF) is present but partially masked by a stronger peak shared by the other 23 isolates. The central figure displays 3 different biomarkers that correspond to 5 different peaks. The first at 10 300 m/z is shared by all isolates in our collection and corresponds to RpsS (species biomarker). The two peaks in the middle correspond to RpsN and the last two peaks correspond to RpsO, isoforms 1 and 2 respectively.

maritimum strains and the type strains of the 23 other *Tenacibaculum* species included in this study were subjected to multiple ($n \geq 3$), independent, MALDI-TOF MS data acquisitions (one acquisition corresponding to an average spectrum of four technical replicates) using the ethanol/formic acid extraction procedure from fresh bacterial culture. Using this set of reference strains, the accuracy was 100%. In addition, laboratories may use different procedures [46] for MALDI-TOF MS analysis (from sample preparation to data processing and interpretation). To address these issues, we tested several sample preparation protocols. The quickest and easiest way to acquire MALDI-TOF MS data from a bacterium is direct transfer method. We evaluated this protocol with 104 bacterial isolates arbitrarily sampled among the 135 above-mentioned isolates. Among these, only 9 strains were not fully typed (8.65%) because of about 1% peak loss. We also assessed our method with ethanol-fixed bacteria, a strategy that could facilitate the MALDI-TOF MS typing of isolates from distant locations. We evaluated the effects of long term (up to 1 month) ethanol conservation followed by the regular protein-extraction protocol. Fifty-one out of 62 isolates (82%) were fully typed after 7 to 15 days ethanol storage. However, only 15 out of 25 isolates (60%) were fully typed after 16 to 26 days ethanol storage. Hence, the conservation time significantly affected typing reliability. Of importance, IF assignments were always congruent for the same sample whatever the preparation method used.

A web-based tool for MALDI-Type assignment

We felt that a web-based tool for MT assignment of *T. maritimum* strains would facilitate data interpretation and help comparisons at a global scale in the same way as previously proposed for MLST schemes [47]. Hence, we developed a friendly application named MALDIquant-TypeR for *Tenacibaculum* MALDI-TOF data analysis that contains 3 tools. Firstly, in order to avoid misinterpretation with spectra from bacteria that do not belong to the species *T. maritimum*, we added reference spectra from 24 out of the 29 described *Tenacibaculum* species, including all the fish-pathogenic or fish-associated species (i.e., *T. dicentrarchi*, *T. discolor*, "*T. finnmarkense*", *T. gallaicum*, *T. ovolyticum* and *T. soleae*). Each reference spectrum is composed of a peak list obtained with *Tenacibaculum* type strains following Bruker's recommendations (>20 independent acquisitions). We used stringent parameters and kept peaks with a signal/noise

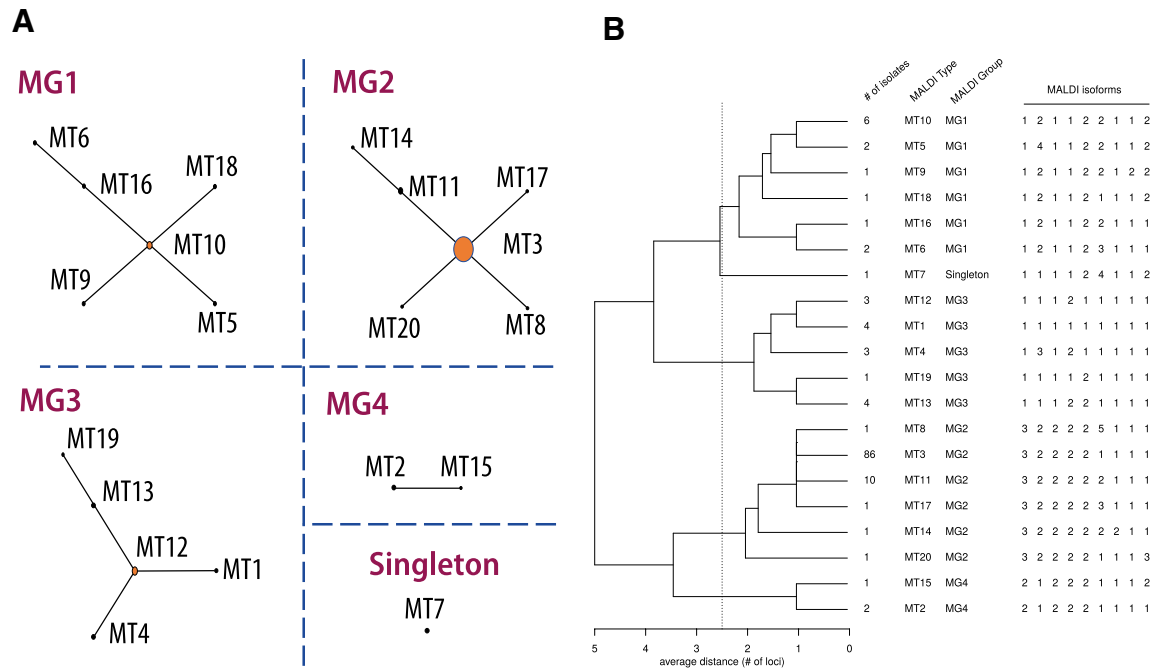


Figure 3 eBurst network, visual representation of the links between the MALDI-Types and the hierarchical tree based on MALDI-Types isomorphic profile. A eBurst defines clusters when isolates share 8 out of the 9 polymorphic biomarkers. These clusters may be considered as MALDI-Groups. Each of them contains a central MT, which represents the "founder" MT of the corresponding group. **B** The classification tree, based on average-link hierarchical clustering (average-link method), offers another view of the links existing between MALDI-Types and MALDI-Groups. The distance used for this tree corresponds to the number of differences between each MALDI profile.

ratio > 3. Thus, each *Tenacibaculum* species is defined by 20 to 60 peaks. When raw spectra from an unidentified sample are uploaded in the application, a peak list is produced using the same stringent parameters and each retained peak is compared to the reference peak file. A peak is considered matching the reference if the difference between two values is less than 700 ppm. Then, the number of matching peaks between the sample and each type strain is computed. The output is a bar-plot providing the percentage of matching peaks with the references corresponding to the 24 *Tenacibaculum* species included so far. For instance, all *T. maritimum* isolates used in this study display at least 50% of common peaks with the *T. maritimum* type strain (vs < 20% with the type strains of other *Tenacibaculum* species). This first tool suggests a taxonomic affiliation and only selects the spectra likely belonging to the species *T. maritimum.* The second tool uses the 18 *T. maritimum* specific monomorphic biomarkers previously defined. It is based on the same peak matching strategy to accurately identify a spectrum as belonging to the species *T. maritimum.* In addition, it also provides a valuable measure of spectra data quality (Additional file 8). The third tool is the typing method itself that identifies the IF of the 9 polymorphic biomarkers and provides the isomorphic profile and the resulting combination corresponding to the MT. This profile can further be processed in the same way as an MLST profile. The MALDIquantTypeR application is available online [23] hosted by the Migale platform at INRAE Jouy-en-Josas.

Discussion

The fast development of aquaculture faces an array of sanitary issues, causing important economic losses and impacting the environment and animal welfare [1, 2]. The rapid detection of pathogens and the continuous monitoring of circulating bacterial genotypes in space and time is a prerequisite for the implementation of rational control measures [48]. International and cross-sector surveillance of pathogens requires strain typing methods that must be rapid, accurate, resolutive, reproducible and affordable, and that must enable the exchange of molecular typing data via the Internet [1].

Tenacibaculosis is an ulcerative disease affecting many marine fish species of commercial interest worldwide and *T. maritimum* is considered the main causative agent of tenacibaculosis in wild and cultured fish [5]. Different typing methods for epidemiological investigations of *T. maritimum* have been proposed [9]. However, they do not fit all the above-mentioned recommendations. For

instance, the different serotyping schemes that have been proposed [10, 11, 49] appear to be unrelated and poorly discriminatory (only three serotypes have been reported to date and some strains display cross-reactivity). Moreover, conventional serology is costly, labor-intensive and requires significant technical expertise and the use of animals to raise anti-sera. On the other hand, an MLST scheme has been proposed [15] and proved to be a powerful discriminating tool for isolate identification and strain typing. However, classical MLST is also costly, labor-intensive and time consuming. Whole genome sequencing represents the "ultimate" typing methodology in terms of discriminatory power. However, it is not yet suitable for real time monitoring of large collections of bacterial isolates and is mostly used for retrospective analysis.

In this study, we first used genomic comparisons of 25 strains including 22 newly draft-sequenced genomes to draw a global picture of the genomic diversity of the species. *Tenacibaculum maritimum* appears as a cohesive bacterial species which strains are characterized by: (i) a high genomic identity (ANI > 98%); (ii) a similar genome size (~3,4 Mb); and (iii) a moderate level of nucleotide divergence (maximum 1.52% in pairwise core-genome sequence comparisons) while typical bacterial species can exhibit up to ~5% nucleotide divergence [50]. In addition, a major contribution of recombination ($r/m \geq 7$) in the evolutionary process of the species was observed, reminiscent of the situation observed in another fish-pathogenic species of the family *Flavobacteriaceae*, *Flavobacterium. psychrophilum* [35]. Our data set encompasses two fully PacBio-assembled genomes (i.e., NCIMB 2154^T and TM-KORJJ) that are perfectly collinear, pointing to a similar chromosomal organization without major genomic rearrangements.

Overall, our results based on 25 genomes are not only in good accordance with the conclusions drawn from our previous analysis using a 11 loci-based MLST data set [15], but they provide unprecedented details on genome content and organization. Tentative phylogenomic tree reconstructions using different methods are congruent and support the division in three main clades (A, B and C). In addition to core-genome genes similarities, clades A and C share some common genomic features. For example, the RplU encoding gene version is obviously different between clades A/C and clade B strains (209 amino-acid long in clade A and C strains vs a 161 amino-acid long in clade B strains). The *metE* gene, encoding methionine synthase, is full length in clade B strains but pseudogenized in clades A/C strains suggesting a non-functional methionine biosynthesis pathway in the latter. Additional clade-specific features were identified. For example, the two clade C strains display non-functional, frame shifted versions of genes of which full-length versions are present in all the other genomes examined (e.g., strain NCIMB 2154^T locus tags: *MARIT_0127*, *MARIT_0229*, *MARIT_0258*, *MARIT _0523*, *MARIT _0991*, *MARIT_1553*, *MARIT_1615*, *MARIT_1833*, *MARIT_1972*, *MARIT_2492*, *MARIT_2635* and *MARIT_2635*). These observations performed on a limited number of genomes may face some exceptions (e.g., due to gene shuffling by homologous recombination). However, the presence of such gene remnants in the genome of some strains argues for genome reduction trends as frequently observed in bacterial pathogens [51]. Strikingly, most of the variable-genome encoding genes are located in genomic islands, some restricted to a single strain while others are shared between several strains. It is therefore tempting to speculate that some islands provide a fitness advantage such as the one containing heavy metal resistance genes identified in strain FC to face an environment polluted by heavy metals. Indeed, important copper concentrations in sediments because of disposal of copper mine tailings has been documented for bays of northern Chile, the region where strain FC was isolated from. A copper-resistant *Vibrio* sp. strain was retrieved from cultured scallop [52] from the same geographical area suggesting convergent evolutionary mechanisms for heavy metal resistance in these phylogenetically unrelated marine bacteria.

In addition to providing a global picture of the genomic diversity, the use of genomic data has been a key step in the MALDI-TOF MS scheme proposed in this study by allowing a better understanding of the spectral composition. The scheme follows Sauget et al. [46] MALDI-TOF MS recommendations by using the genomic information for selecting relevant biomarkers, by choosing ribosomal proteins that should be unambiguously identified within the spectra and by focusing exclusively on pic shifts for the typing purpose. Indeed, we were able to retrieve from the spectra 18 out of 54 in silico-identified ribosomal proteins. The total number of biomarkers (18 monomorphic and 9 polymorphic) corresponds to about one-third of the detected peaks. In addition, most of these peaks display high intensity, even at both ends of the spectra. Combining genome mining and visual exploration of the spectra, we linked each noteworthy peak shift with the corresponding polymorphism in genomic sequences (Additional file 9). Strikingly, the 9 polymorphic biomarkers and their corresponding isoforms (25 in total) were defined from only 10 strains (out of 25 sequenced isolates), reflecting the cohesive nature of the *T. maritimum* species. The strategy based on highly relevant biomarkers and their selection based on stringent criteria proved to be particularly effective. Using 111 field isolates and the

ethanol/formic acid extraction procedure, automated peak detection displayed a very high level of success rate (> 99.5% of the polymorphic biomarkers identified). Because fluctuations in MALDI-TOF MS results might be caused by the use of different bacterial growth conditions, sample preparation procedures or matrix used [46], we evaluated the robustness and reliability of the proposed typing scheme by using biological and technical replicates as well as alternative sample preparations (i.e., direct transfer method and ethanol/formic extraction after ethanol conservation). As detailed above, the results are still very good when alternative sample preparation procedures are used, though the success rate may be lower, in particular using long-term ethanol conservation. The observed peak detection failures may be attributed to random ion suppression intrinsically linked to the analogical nature of the spectra [46]. This concern may, however, be simply addressed by data re-acquisition (Additional file 8).

Because peak shifts are the direct consequence of a non-synonymous mutation in a ribosomal protein, a peak shift may be considered as an allelic change as proposed by Zautner et al. [42]. We therefore treated peaks shifts as a combination of alleles. A single strain would thus display a unique combination that could be considered as a MALDI-type (MT). These MTs are therefore the equivalent of the sequence types (STs) or electrophoretic types (ETs) resulting from the MLST or MLEE schemes, respectively. In the same way as the latter two schemes, the proposed Multi Peak Shift Typing (MPST) scheme for *T. maritimum* could be enriched in the future by considering additional biomarkers and/or by the identification of additional IF in the retained biomarkers. New, yet unobserved, combinations of isoforms may also be identified in the future. The definition of new MTs will follow the same incremental process as previously proposed for MLST schemes.

Applied to our collection of *T. maritimum* isolates, this method identified 20 MTs grouped in 4 MGs. Because our goal was to propose and to evaluate a MPST scheme dedicated to the typing of *T. maritimum* strains, we used isolates from broad geographical, temporal and host-fish species origins to cover as much of the species diversity as possible. We are aware of the inherent bias of such a sampling choice. Indeed, the variables (i.e. country, year of isolation, host fish species) were highly correlated with each others; therefore, this dataset is not suitable for highlighting sound associations between isolation sources and MTs. However, we observed trends between the geographical origin of the strains and the MT as previously reported using MLST data [15]. In addition, and in line with previous MLST-based conclusions, our analysis revealed no trace of long-distance dissemination of *T. maritimum* that could be linked to the international trade of fish or eggs.

MPST schemes based on MALDI-TOF MS data may prove suitable for large-scale epidemiological studies for a number of reasons: (i) the approach is similar to MLST, and the tools used to analyze MLST data can also be used on MPST data; (ii) MALDI-TOF acquisitions are much cheaper and faster (about 20 minutes for a whole MALDI-TOF plate acquisition) than sequencing- or PCR-based MLST; (iii) 96 samples can be deposited on the same MALDI-TOF plate; and (iv) the scheme can be transposed to any bacterial species that contains polymorphic ribosomal proteins or any noteworthy polymorphic biomarker. Indeed, the peak shifts caught by MALDI-TOF MS are highly correlated to the genotype as defined in any MLST analysis. MALDI-TOF spectra databases and processing tools can be easily exported on the Internet as a webtool application in the same way as MLST web-based tools (e.g., PubMLST.org). MALDIquantTypeR is one example of such application available online [23] as well as other dedicated websites such as the Mass Spectrometry Identification platform [53]. Large-scale studies could easily benefit from worldwide data collection. We have shown that ethanol (at least short-term) conservation can be used for MALDI-TOF identification and typing. This method also suppresses biological hazard and should facilitate international transportation of bacterial samples. Other solutions may be designed for international collaboration, such as direct shipment of MALDI-TOF plates with fixed biological material, ready for acquisition. MPST analysis could also be interesting in local surveys monitoring bacterial populations in particularly sensitive geographical areas.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13567-020-00782-0.

Additional file 1. *T. maritimum* isolates used in this study and their corresponding MALDI-Types.

Additional file 2. Average Nucleotide Identity (ANI) of pairwise comparison of the *T. maritimum* isolates.

Additional file 3. Parsimony based phylogenetic tree. The tree is based on the alignment made by Snippy. It is reconstructed using the parsimony method as implemented in *dnapars* (Phylip package v3.6). The bootstrap support of each branch is computed from 100 bootstrap replicates. The three clades designated A, B, and C are labeled and delineated by vertical bars.

Additional file 4. Maximum likelihood, Gubbins-based phylogenetic tree. The tree was obtained from the whole genome alignment of 25 *T. maritimum* strains at the fifth and final iteration of Gubbins. Statistical support of nodes is indicated. The three clades designated A, B, and C are labeled and delineated by vertical bars.

Additional file 5. Overview of recombination tracts between two pairs of *T. maritimum* isolates. SNPs and recombination tracts between two closely related isolates. In 5A the strains compared are USC SP9.1 and

USC SE30.1. In 5B the strains compared are FS08(1) and P2—48. (Upper) Positions of the SNPs along the genomes. SNP index is reset every 100 SNPs for this representation. Each dot corresponds to one SNP in the comparison between the two considered isolates. Colors distinguish two types of polymorphism: in blue, polymorphism observed only between the two considered genomes; in red, polymorphism also observed among the other sequenced genomes. Areas in gray correspond to regions not covered by our alignments. SNPs in regions where probability is < 0.5 (i.e. outside predicted recombination tracts) are represented by open symbols (blue circles). (Lower) Probability of recombination tract as computed with the HMM. Estimation of the % of genome in recombination tracts is 15.8 and 19.1 for (A) and (B), respectively. Estimation of the average length of recombination tracts is 885 bp and 328 bp. for (A) and (B), respectively. Estimation of the average nucleotide diversity inside recombination tracts is 0.013/bp and 0.014/bp for (A) and (B), respectively. Estimation of the average nucleotide diversity outside recombination tracts is 9.9e-5/bp and 3.5e-4/bp for (A) and (B), respectively. Estimation of the number of SNPs inside recombination tracts is 5266 and 7013 for (A) and (B), respectively. Estimation of the number of SNPs outside recombination tracts is 216 and 737 for (A) and (B), respectively. Estimation of the number of SNPs due to mutations (extrapolated from non — recombined regions) is 256 and 910 for (A) and (B), respectively. Estimation of the ratio r/m is 20.6 and 7.7 for (A) and (B), respectively.

Additional file 6. Heatmap displaying the diversity of ribosomal protein weights. Lines correspond to strains and columns correspond to ribosomal proteins in ascending order by weight (from left to right). White lines indicate no variation of the weight of the corresponding proteins (i.e., monomorphic proteins). In purple, proteins with weight higher than the mean and in orange proteins with weight lower than the mean (i.e., polymorphic proteins).

Additional file 7. Monomorphic biomarker peaks. Screenshots of the 18 conserved peaks produced by 9 ribosomal monomorphic proteins with several degrees of ionization. The m/z values are highlighted by dotted lines and cover the entire spectrum. The red curve corresponds to the *T. maritimum* type strain average spectra. For each peak, the corresponding ribosomal protein is indicated with the degree of ionization (H1, H2 and H3 corresponding to 1, 2 and 3 H^+) and the presence (M) or absence (m) of the first methionine. Color code: red line for the *T. maritimum* type strain NCIMB 2154^T and black for the sequenced *T. maritimum* isolates.

Additional file 8. Quality control and *T. maritimum* species identification. The full dataset is composed of representatives of 24 *Tenacibaculum* species including 135 isolates belonging to the species *T. maritimum*. It encompasses 476 independent acquisitions (one acquisition corresponds to an average spectrum of several technical replicates) corresponding to 5102 spectra including technical and biological replicates. This dataset was divided into two groups: the positive control group (*T. maritimum* isolates) and the negative control group (the type strains of 23 other *Tenacibaculum* species). In order to confirm that an isolate belongs to the species *T. maritimum*, the spectra were scanned to identify the 18 *T. maritimum* monomorphic biomarkers. However, some of these biomarkers could be absent from a number of *T. maritimum* strains. Reciprocally, strains that do not belong to the *T. maritimum* species may possess some *T. maritimum* monomorphic biomarkers. In order to set up a *T. maritimum* species identity threshold, a value corresponding to the number of monomorphic biomarkers identified in a single sample was computed. The monomorphic biomarkers frequency plot obtained shows a bimodal distribution (Figure A). All samples with a score above 60% correspond to *bona fide T. maritimum* isolates while all samples with a score below 25% belong to other *Tenacibaculum* species. True and false positives correspond to isolates correctly and incorrectly identified as *T. maritimum* (TP and FP), respectively. On the other hand, true and false negatives correspond to correctly and incorrectly rejected isolates (TN and FN, respectively). Positive isolates correspond to those having an identity value above a defined threshold. Using the full dataset, the number of TP and FP and the number of TN and FN were counted. The accuracy of the tool [i.e., (TP + TN)/(TP + TN + FP + FN)] was then computed by increasing the threshold value from 0% to 100% by a 0.1% step. One could observe than the accuracy varies from 0% to 100% and is maximal (i.e., above 97%) between 25% and 60% of threshold value (Figure B). It is therefore proposed that a 60% threshold value safely identifies isolates as belonging to the *T. maritimum* species. Using this 60% threshold value, only 10 false negatives (i.e., *bona fide T. maritimum* isolates discarded) and 0 false positive (i.e., not *T. maritimum* isolates) were found out of 476 acquisitions. Among the 10 false negatives, 3 resulted from the extraction protocol, 3 from the Direct Deposit with Formic Acid (DDFA) protocol and 4 from the ethanol conservation protocol (see section "Testing alternative sample preparation for MALDI-TOF MS"). One can hypothesize that these rare false negatives correspond to technical problems. Indeed, by performing new spectra acquisitions on the 3 isolates previously analyzed by the extraction protocol, the identification score reached at least 88%, far above the safe threshold value of 60%, demonstrating that the original spectra were likely faulty. Finally, spectra from other *Tenacibaculum* species all had a score below 23%, far below the confidence threshold.

Additional file 9. Amino-acid polymorphism of the 9 retained biomarkers. Protein sequence alignments of the 9 ribosomal proteins selected as polymorphic biomarkers and their corresponding isoform (IF).

Acknowledgements
This study was funded by the AAP Ressourcement Scientifique F2E 2017 project "New tools for diagnosis of bacterial fish pathogens". The authors are very grateful to Klervi L'Her (Bio Chêne vert) for her participation in the sample preparation and MALDI-TOF acquisitions, to the many individuals listed in the legend to Additional file 1 for kindly providing bacterial strains, to Pierre-Yves Moalic (Labofarm) for fruitful discussions and to Zoé Rouy (MicroScope Platform) for her kind help with genome data management. This work has benefited from the facilities and expertise of the high-throughput sequencing facility of I2BC for its sequencing and bioinformatics expertise (Centre de Recherche de Gif [54]) and of the CEA/GENOSCOPE [55]. We also wish to thank for providing computational resources: the INRAE MIGALE bioinformatics platform [56], the LABGeM, and the National Infrastructure "France Génomique" funded as part of the Investissement d'avenir program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). SB is funded by ANRT/CIFRE grant number 2006/0707.

Authors' contributions
SB performed data analysis and writing of the manuscript. FB and AM contributed to bacterial isolation and to sample processing. DS and JFB contributed to fish sampling and bacterial isolation. SP and PN contributed to data analysis and presentation of the results. ED initiated the project, contributed to data analysis and wrote the manuscript. All authors read and approved the final manuscript.

Availability of data and materials
All genome sequencing data have been deposited in the European Nucleotide Archive (Accession numbers: GCA_902705265, GCA_902705275, GCA_902705285, GCA_902705305, GCA_902705315, GCA_902705345, GCA_902705355, GCA_902705365, GCA_902705375, GCA_902705385, GCA_902705395, GCA_902705415, GCA_902705425, GCA_902705435, GCA_902705445, GCA_902705465, GCA_902705495, GCA_902705515, GCA_902705525, GCA_902705535, GCA_902705555 and GCA_902705565).

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Université Paris-Saclay, INRAE, UVSQ, VIM 78350 Jouy-En-Josas, France. [2] Labofarm, Finalab, 22603 Loudéac, France. [3] Université de Versailles Saint-Quentin-En-Yvelines, 78180 Montigny-Le-Bretonneux, France. [4] Bio Chêne Vert, Finalab, Rue Blaise Pascal, 35220 Châteaubourg, France. [5] Ifremer, UMR EIO 241, Labex Corail, Centre du Pacifique, BP 49, Taravao, 98719 Tahiti, French Polynesia. [6] Institut de Systématique Evolution, Biodiversité, UMR 7205 Sorbonne Université MNHN CNRS EPHE, Paris, France. [7] Université Paris-Saclay, INRAE, MaIAGE 78350 Jouy-en-Josas, France.

Received: 20 January 2020 Accepted: 8 April 2020
Published online: 07 May 2020

References

1. Bayliss SC, Verner-Jeffreys DW, Bartie KL, Aanensen DM, Sheppard SK, Adams A, Feil EJ (2017) The promise of whole genome pathogen sequencing for the molecular epidemiology of emerging aquaculture pathogens. Front Microbiol 8:121
2. FAO (2016) The State of World Fisheries and Aquaculture. FAO, Rome. http://www.fao.org/3/a-i5555e.pdf
3. Adam KE, Gunn GJ (2017) Social and economic aspects of aquatic animal health. Rev Sci Tech 36:323–329
4. Rodgers CJ, Mohan CV, Peeler EJ (2011) The spread of pathogens through trade in aquatic animals and their products. Rev Sci Tech 30:241–256
5. Avendaño-Herrera R, Toranzo AE, Magariños B (2006) Tenacibaculosis infection in marine fish caused by *Tenacibaculum maritimum*: a review. Dis Aquat Organ 71:255–266
6. Gourzioti E, Kolygas M, Athanassopoulou F, Babili V (2018) Tenacibaculosis in aquaculture farmed marine fish. J Hell Vet Med Soc 67:21
7. Rahman T, Suga K, Kanai K, Sugihara Y (2014) Biological and serological characterization of a non-gliding strain of *Tenacibaculum maritimum* isolated from a diseased puffer fish *Takifugu rubripes*. Fish Pathol 49:121–129
8. Avendaño-Herrera R, Núñez S, Barja JL, Toranzo AE (2008) Evolution of drug resistance and minimum inhibitory concentration to enrofloxacin in *Tenacibaculum maritimum* strains isolated in fish farms. Aquac Int 16:1–11
9. Fernández-Álvarez C, Santos Y (2018) Identification and typing of fish pathogenic species of the genus *Tenacibaculum*. Appl Microbiol Biotechnol 102:9973–9989
10. Avendaño-Herrera R, Magariños B, López-Romalde S, Romalde JL, Toranzo AE (2004) Phenotypic characterization and description of two major O-serotypes in *Tenacibaculum maritimum* strains from marine fishes. Dis Aquat Organ 58:1–8
11. Avendaño-Herrera R, Magarinos B, Morinigo M, Romalde J, Toranzo A (2005) A novel O-serotype in *Tenacibaculum maritimum* strains isolated from cultured sole (*Solea senegalensis*). Bull Eur Assoc Fish Pathol 25:70–74
12. Avendaño-Herrera R, Rodríguez J, Magariños B, Romalde JL, Toranzo AE (2004) Intraspecific diversity of the marine fish pathogen *Tenacibaculum maritimum* as determined by randomly amplified polymorphic DNA-PCR. J Appl Microbiol 96:871–877
13. Piñeiro-Vidal M, Pazos F, Santos Y (2008) Fatty acid analysis as a chemotaxonomic tool for taxonomic and epidemiological characterization of four fish pathogenic *Tenacibaculum* species. Lett Appl Microbiol 46:548–554
14. Fernández-Álvarez C, Torres-Corral Y, Santos Y (2018) Comparison of serological and molecular typing methods for epidemiological investigation of *Tenacibaculum* species pathogenic for fish. Appl Microbiol Biotechnol 102:2779–2789
15. Habib C, Houel A, Lunazzi A, Bernardet J-F, Olsen AB, Nilsen H, Toranzo AE, Castro N, Nicolas P, Duchaud E (2014) Multilocus sequence analysis of the marine bacterial genus *Tenacibaculum* suggests parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. Appl Environ Microbiol 80:5503–5514
16. Avendaño-Herrera R, Irgang R, Sandoval C, Moreno-Lira P, Houel A, Duchaud E, Poblete-Morales M, Nicolas P, Ilardi P (2016) Isolation, characterization and virulence potential of *Tenacibaculum dicentrarchi* in salmonid cultures in Chile. Transbound Emerg Dis 63:121–126
17. Klakegg Ø, Abayneh T, Fauske AK, Fülberth M, Sørum H (2019) An outbreak of acute disease and mortality in Atlantic salmon (*Salmo salar*) post-smolts in Norway caused by *Tenacibaculum dicentrarchi*. J Fish Dis 42:789–807
18. Olsen AB, Gulla S, Steinum T, Colquhoun DJ, Nilsen HK, Duchaud E (2017) Multilocus sequence analysis reveals extensive genetic variety within *Tenacibaculum* spp. associated with ulcers in sea-farmed fish in Norway. Vet Microbiol 205:39–45
19. Fernández-Álvarez C, Torres-Corral Y, Saltos-Rosero N, Santos Y (2017) MALDI-TOF mass spectrometry for rapid differentiation of *Tenacibaculum* species pathogenic for fish. Appl Microbiol Biotechnol 101:5377–5390
20. Pérez-Pascual D, Lunazzi A, Magdelenat G, Rouy Z, Roulet A, Lopez-Roques C, Larocque R, Barbeyron T, Gobet A, Michel G, Bernardet J-F, Duchaud E (2017) The complete genome sequence of the fish pathogen *Tenacibaculum maritimum* provides insights into virulence mechanisms. Front Microbiol 8:1542
21. Bridel S, Olsen A-B, Nilsen H, Bernardet J-F, Achaz G, Avendaño-Herrera R, Duchaud E (2018) Comparative genomics of *Tenacibaculum dicentrarchi* and "*Tenacibaculum finnmarkense*" highlights intricate evolution of fish-pathogenic species. Genome Biol Evol 10:452–457
22. Hou TY, Chiang-Ni C, Teng SH (2019) Current status of MALDI-TOF mass spectrometry in clinical microbiology. J Food Drug Anal 27:404–414
23. MALDIquantTypeR by S. Bridel. http://genome.jouy.inra.fr/shiny/maldiquanttyper/. Accessed 10 Apr 2020
24. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJC, Yoo HS, Zhang C, Zhang Y, Sobral BW (2014) PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:581–591
25. Médigue C, Calteau A, Cruveiller S, Gachet M, Gautreau G, Josso A, Lajus A, Langlois J, Pereira H, Planel R, Roche D, Rollin J, Rouy Z, Vallenet D (2017) MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. Brief Bioinform 20:1071–1084
26. Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131
27. Pritchard L (2020) https://github.com/widdowquinn/pyani. Accessed 10 Apr 2020
28. MicroScope home-MaGe: microbial genome annotation & analysis platform-MicroScope-web interface system & specialized databases for (re)annotation and analysis of microbial genomes. https://mage.genoscope.cns.fr/microscope/home/index.php. Accessed 10 Apr 2020
29. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, Burlot L, Bussell X, Fouteau S, Gautreau G, Lajus A, Langlois J, Planel R, Roche D, Rollin J, Rouy Z, Sabatet V, Médigue C (2019) MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res 48:579–589
30. Gautreau G (2020) https://github.com/ggautreau/PPanGGOLiN. Accessed 10 Apr 2020
31. Seemann T (2020) https://github.com/tseemann/snippy; Accessed 10 Apr 2020
32. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15
33. PHYLIP Home Page. http://evolution.genetics.washington.edu/phylip.html. Accessed 10 Apr 2020
34. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791
35. Duchaud E, Rochat T, Habib C, Barbier P, Loux V, Guérin C, Dalsgaard I, Madsen L, Nilsen H, Sundell K, Wiklund T, Strepparava N, Wahli T, Caburlotto G, Manfrin A, Wiens GD, Fujiwara-Nagata E, Avendaño-Herrera R, Bernardet J-F, Nicolas P (2018) Genomic diversity and evolution of the fish pathogen *Flavobacterium psychrophilum*. Front Microbiol 9:138
36. Mellmann A, Cloud J, Maier T, Keckevoet U, Ramminger I, Iwen P, Dunn J, Hall G, Wilson D, Lasala P, Kostrzewa M, Harmsen D (2008) Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to 16S rRNA gene sequencing for species identification of nonfermenting bacteria. J Clin Microbiol 46:1946–1954
37. TermiNator. https://bioweb.i2bc.paris-saclay.fr/terminator3/. Accessed 10 Apr 2020
38. Bioconductor-Home. https://bioconductor.org/. Accessed 10 Apr 2020
39. Du P, Kibbe WA, Lin SM (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 22:2059–2065
40. Ryzhov V, Fenselau C (2001) Characterization of the protein subset desorbed by MALDI from whole bacterial cells. Anal Chem 73:746–750
41. Martinez A, Traverso JA, Valot B, Ferro M, Espagne C, Ephritikhine G, Zivy M, Giglione C, Meinnel T (2008) Extent of N-terminal modifications in cytosolic proteins from eukaryotes. Proteomics 8:2809–2831

42. Zautner AE, Masanta WO, Weig M, Groß U, Bader O (2015) Mass spectrometry-based phyloproteomics (MSPP): a novel microbial typing Method. Sci Rep 5:13431
43. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186:1518–1530
44. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267
45. Urwin R, Maiden MCJ (2003) Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol 11:479–487
46. Sauget M, Valot B, Bertrand X, Hocquet D (2017) Can MALDI-TOF mass spectrometry reasonably type bacteria? Trends Microbiol 25:447–455
47. Jolley KA, Bray JE, Maiden MCJ (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Welcome Open Res 3:124
48. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, Fussing V, Green J, Feil E, Gerner-Smidt P, Brisse S, Struelens M, European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM) (2007) Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect 13(Suppl 3):1–46
49. Wakabayashi H (1984) *Flexibacter* infection in cultured marine fish in Japan. Helgol Meeresunters 37:587–593
50. Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351
51. Weinert LA, Welch JJ (2017) Why might bacterial pathogens have small genomes? Trends Ecol Evol 32:936–947
52. Miranda CD, Rojas R (2006) Copper accumulation by bacteria and transfer to scallop larvae. Mar Pollut Bull 52:293–300
53. Normand AC, Becker P, Gabriel F, Cassagne C, Accoceberry I, Gari-Toussaint M, Hasseine L, De Geyter D, Pierard D, Surmont I, Djenad F, Donnadieu JL, Piarroux M, Ranque S, Hendrickx M, Piarroux R (2017) Validation of a new Web application for identification of fungi by use of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol 55:2661–2670
54. Institut de Biologie Intégrative de la Cellule https://www.i2bc.paris-saclay.fr/. Accessed 10 Apr 2020
55. CEA (2018) About Genoscope: CEA François Jacob Inst. Biol. http://www.cea.fr/drf/ifrancoisjacob/english/Pages/Departments/Genoscope/About-Genoscope.aspx. Accessed 10 Apr 2020
56. Migale platform. https://migale.inra.fr/. Accessed 10 Apr 2020

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

C2 – Supplementary materials

Supplementary material

Genetic diversity and population structure of *Tenacibaculum maritimum,* a devastating bacterial pathogen of marine fish: from genome comparisons to high throughput MALDI-TOF typing.

Sébastien Bridel[1,2,3], Frédéric Bourgeon[4], Arnaud Marie[3], Denis Saulnier[5], Sophie Pasek[6], Pierre Nicolas[7] , Jean-François Bernardet[1], and Eric Duchaud[1]*

Affiliations
1 Unité VIM, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France.
2 Labofarm, Finalab, 22603 Loudéac, France.
3 Université de Versailles Saint-Quentin-En-Yvelines, 78180 Montigny-Le-Bretonneux, France.
4 Bio Chêne Vert, Finalab, rue Blaise Pascal, 35220 Châteaubourg.
5 Ifremer, UMR EIO 241, Labex Corail, Centre du Pacifique, BP 49, 98719 Taravao, Tahiti, French Polynesia.
6 Institut de Systématique Evolution, Biodiversité, UMR 7205 Sorbonne Université MNHN CNRS EPHE, Paris, France.
7 Unité MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France.

List of contents

S1 – *Tenacibaculum maritimum* isolates and their corresponding MALDI-Types.

S2 – Average Nucleotide Identity (ANI) of pairwise comparison of the *T. maritimum* isolates.

S3A – Parsimony tree.

S3B – Maximum likelihood, Gubbins-based phylogenetic tree.

S4 – Overview of recombination tracts between two pairs of *T. maritimum* isolates.

S5 – Heatmap displaying the diversity of ribosomal protein weights.

S6 – Monomorphic biomarker peaks.

S7 – Quality control and *T. maritimum* species identification.

S8 – Amino-acid polymorphism of the 9 retained biomarkers.

S1 – *T. maritimum* isolates used in this study and their corresponding MALDI-Types

Sample	Country	Year	Host	Contributor	RpmD	RpmC	RpsQ	RpsO	RplX	RpsP	RpsT	RpsN	RplT	MALDI-Type	MALDI-Group
NCIMB 2154^{T}	JP	1977	PMR	NCIMB	1	1	1	1	1	1	1	1	1	MT1	MG3
Baxa 1y 1-1	JP	1985	ASI	RPB	1	1	1	1	1	1	1	1	1	MT1	MG3
FPC371	JP	2000	SMS	HW	1	1	1	1	1	1	1	1	1	MT1	MG3
FPC454	JP	1977	PMR	HW	1	1	1	1	1	1	1	1	1	MT1	MG3
UCD SD26	US	1995	ANS	RH	2	1	2	2	2	1	1	1	1	MT2	MG4
UCD SB2	US	1995	ANS	RH	2	1	2	2	2	1	1	1	1	MT2	MG4
DPIF 89/0329-11	AU	1989	SSR	JC	3	2	2	2	2	1	1	1	1	MT3	MG2
DPIF 89/0329-5	AU	1989	SSR	JC	3	2	2	2	2	1	1	1	1	MT3	MG2
DPIF 89/0578-4	AU	1989	SSR	JC	3	2	2	2	2	1	1	1	1	MT3	MG2
DPIF 89/0699	AU	1989	SSR	JC	3	2	2	2	2	1	1	1	1	MT3	MG2
DPIF 89/1288-8	AU	1989	OMS	JC	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 16-89	FP	2016	POS	DS	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 16-85	FP	2016	POS	DS	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 16-88	FP	2016	POS	DS	3	2	2	2	2	1	1	1	1	MT3	MG2
TFA4	FP	2016	POS	DS	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 13-47	FR	2014	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
FM1068	FR	1993	DLX	JFP	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 24/99	FR	1999	SMS	FL	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 31/99	FR	1999	SMS	FL	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/99	FR	1991	DLX	CS	3	2	2	2	2	1	1	1	1	MT3	MG2
LVDH 1577.01	FR	2001	DLX	NK	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 21/91-1	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 21/91-2	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 21/91-3	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-1	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-3	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-4	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-5	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-6	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 11-80	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 12-98	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 13-16	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 13-34	FR	2014	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 13-64	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 14-1	FR	2017	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 14-19	FR	2017	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 14-21	FR	2017	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 6-46	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 6-70	FR	2012	n/d	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 6-71	FR	2012	n/d	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 7-8	FR	2012	SAA	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-20	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-21	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-3	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-4	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-47	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-49	FR	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-94	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
902	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-12	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-22	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-23	FR	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-24	FR	2014	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 10/97	FR	2000	SMS	FL	3	2	2	2	2	1	1	1	1	MT3	MG2
JIP 32/91-4	FR	1991	DLX	JFB	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-39	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-41	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-42	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P3-86	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P4-45	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P4-97	FR	n/d	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-28	IT	2015	CTS	MLF	3	2	2	2	2	1	1	1	1	MT3	MG2
Baxa DBA-4a	JP	1986	SQA	RPB	3	2	2	2	2	1	1	1	1	MT3	MG2
Baxa GBF-8601	JP	1986	PLS	RPB	3	2	2	2	2	1	1	1	1	MT3	MG2
NCIMB 2153	JP	1976	ASI	NCIMB	3	2	2	2	2	1	1	1	1	MT3	MG2
NAC SLMG 101	MT	1995	DLX	JT	3	2	2	2	2	1	1	1	1	MT3	MG2

Sample	Country	Year	Host	Contributor	RpmD	RpmC	RpsQ	RpsO	RplX	RpsP	RpsT	RpsN	RplT	MALDI-Type	MALDI-Group
NAC SLMG 105	MT	1995	DLX	JT	3	2	2	2	2	1	1	1	1	MT3	MG2
NAC SLMG 109	MT	1995	DLX	JT	3	2	2	2	2	1	1	1	1	MT3	MG2
NAC SLMG 115	MT	1996	DLX	JT	3	2	2	2	2	1	1	1	1	MT3	MG2
NAC SLMG 120	MT	1996	DLX	JT	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-24	IT	SAA	SAA	MLF	3	2	2	2	2	1	1	1	1	MT3	MG2
P1-27	IT	DLX	DLX	MLF	3	2	2	2	2	1	1	1	1	MT3	MG2
CVI10001048	NL	2010	SAA	OH	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 6-12	SP	2012	SAA	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 5-26	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
USC RPM522.1	SP	1992	SMS	AET	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 11-12	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 12-61	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 5-25	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 5-26	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 6-17	SP	2012	SAA	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-18	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-26	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-34	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-64	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-69	SP	2012	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-88	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 8-90	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-17	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
Aq 9-21	SP	2013	DLX	ALB	3	2	2	2	2	1	1	1	1	MT3	MG2
NCIMB 2158	UK	1981	SSA	NCIMB	3	2	2	2	2	1	1	1	1	MT3	MG2
FPC386	JP	1978	PMR	ALB	1	3	1	2	1	1	1	1	1	MT4	MG3
FPC394	JP	1978	PMR	ALB	1	3	1	2	1	1	1	1	1	MT4	MG3
NBRC 15946	JP*	n/d	n/d	n/d	1	3	1	2	1	1	1	1	1	MT4	MG3
P1-26	IT	n/d	DLX	MLF	1	4	1	1	2	2	1	1	2	MT5	MG1
NAC SLMG MFF	MT	n/d	DLX	ALB	1	4	1	1	2	2	1	1	2	MT5	MG1
USC SE30.1	SP	1993	OKH	AET	1	2	1	1	2	3	1	1	1	MT6	MG1
USC SP9.1	SP	1993	SSR	AET	1	2	1	1	2	3	1	1	1	MT6	MG1
DPIF 89/3001-6.2	AU	1989	LLA	JC	1	1	1	1	2	4	1	1	2	MT7	Singleton 1
JIP 46/00	FR	2000	SMS	GG	3	2	2	2	2	5	1	1	1	MT8	MG2
FS08(1)	IT	2006	SAA	FS	1	2	1	1	2	2	1	2	2	MT9	MG1
Aq 2-43	FR	2011	SSS	ALB	1	2	1	1	2	2	1	1	2	MT10	MG1
Aq 2-48	FR	2011	DLX	ALB	1	2	1	1	2	2	1	1	2	MT10	MG1
P2-48	FR	2011	DLX	ALB	1	2	1	1	2	2	1	1	2	MT10	MG1
P1-25	IT	SAA	SAA	MLF	1	2	1	1	2	2	1	1	2	MT10	MG1
Aq 6-11	SP	2012	SAA	ALB	1	2	1	1	2	2	1	1	2	MT10	MG1
Aq 6-14	SP	2012	SAA	ALB	1	2	1	1	2	2	1	1	2	MT10	MG1
P1-29	n/d	n/d	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-46	SP	2010	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 6-16	SP	2012	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-23	SP	2012	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-24	SP	2012	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-31	SP	2012	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-80	SP	2013	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-89	SP	2013	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
Aq 8-93	SP	2013	DLX	ALB	3	2	2	2	2	2	1	1	1	MT11	MG2
UCD V2b	US	1993	ANS	RH	3	2	2	2	2	2	1	1	1	MT11	MG2
DPIF 90/1445	AU	1990	SSR	JC	1	1	1	2	1	1	1	1	1	MT12	MG3
DPIF 89/0235-3	AU	1989	OMS	JC	1	1	1	2	1	1	1	1	1	MT12	MG3
DPIF 89/0528-1	AU	1989	SSR	JC	1	1	1	2	1	1	1	1	1	MT12	MG3
DPIF 89/0239-1	AU	1989	SSR	JC	1	1	1	2	2	1	1	1	1	MT13	MG3
FC	CL	1998	SMS	JM	1	1	1	2	2	1	1	1	1	MT13	MG3
JIP 05/00(1)	FR	2000	SMS	FL	1	1	1	2	2	1	1	1	1	MT13	MG3
UCD V6f	US	1994	EMX	RH	1	1	1	2	2	1	1	1	1	MT13	MG3
Aq 8-57	SP	2012	DLX	ALB	3	2	2	2	2	2	2	1	1	MT14	MG2
UCD WSB1b	US	1994	ANS	RH	2	1	2	2	2	1	1	1	2	MT15	MG4
USC RP67.1	SP	1993	SMS	AET	1	2	1	1	2	2	1	1	1	MT16	MG1
Aq 12-63	SP	2013	DLX	ALB	3	2	2	2	2	3	1	1	1	MT17	MG2
Aq 6-13	SP	2012	SAA	ALB	1	2	1	1	2	1	1	1	2	MT18	MG1
P2-27	SP	2010	SMS	ALB	1	1	1	1	2	1	1	1	1	MT19	MG3
USC RPM539.1	SP	1993	SMS	AET	3	2	2	2	2	1	1	1	3	MT20	MG2
P1-40	FR	2010	DLX	ALB	3	2	NT	2	2	1	1	1	NT	NT	n/d
Aq 12-62	SP	2013	DLX	ALB	3	2	2	2	2	NT	1	1	1	NT	n/d
Aq 12-64	SP	2013	DLX	ALB	3	2	2	2	2	NT	1	1	1	NT	n/d
Aq 6-9	SP	2012	SAA	ALB	1	NT	1	1	2	2	1	1	1	NT	n/d

** denotes a sample of uncertain origin*

NT equivalent to non-typable, used for missing biomarkers

List of contributors

AET	A. Estévez Toranzo	Universidad Santiago de Compostla	Spain
ALB	A. Le Breton	Vet'eau	Grenade-sur-Garonne, France
CS	C. Sauvegrain	Aquanord	France Gravelines, France
DS	D. Saulnier	IFREMER	Taravao, Tahiti, French Polynesia
FL	F. Leveau	N.A.T.A, France Turbot	L'Epine, France
FS	F. Salati	State Veterinary Institute	Oristano, Italy
GG	G. Gauthier	N.A.T.A, France Turbot	L'Epine, France
HW	H. Wakabayashi	University of Tokyo	Japan
JC	J. Carson	Departement of Primary Industry and Fisheries	King Meadows, Tasmania, Australia
JM	J. Montaña	Fundaciòn Chile	Puerto Montt, Chile
JT	J. Tabone	National Aquaculture Center	Marsaxlokk, Malta
JFB	J.-F. Bernardet	Institut National de la Recherche Agronomique	Jouy-en-Josas, France
JFP	J.-F. Pépin	IFREMER	Palavas-les-Flots, France
MLF	M.L. Fioraventi	University of Bologna	Italy
NK	N. Keck	Laboratoire départemental vétérinaire de l'Hérault	Montpellier, France
NCIMB	National Collection of Industrial and Marine Bacteria	Aberdeen	Scotland
OH	O. Haenen	Central Veterinary Institute	Lelystad, the Netherlands
RH	R. Hedrick	Univeristy of California	Davis, USA
RPB	R.P. Burchard	University of Mariland	Baltimore

List of abbreviations

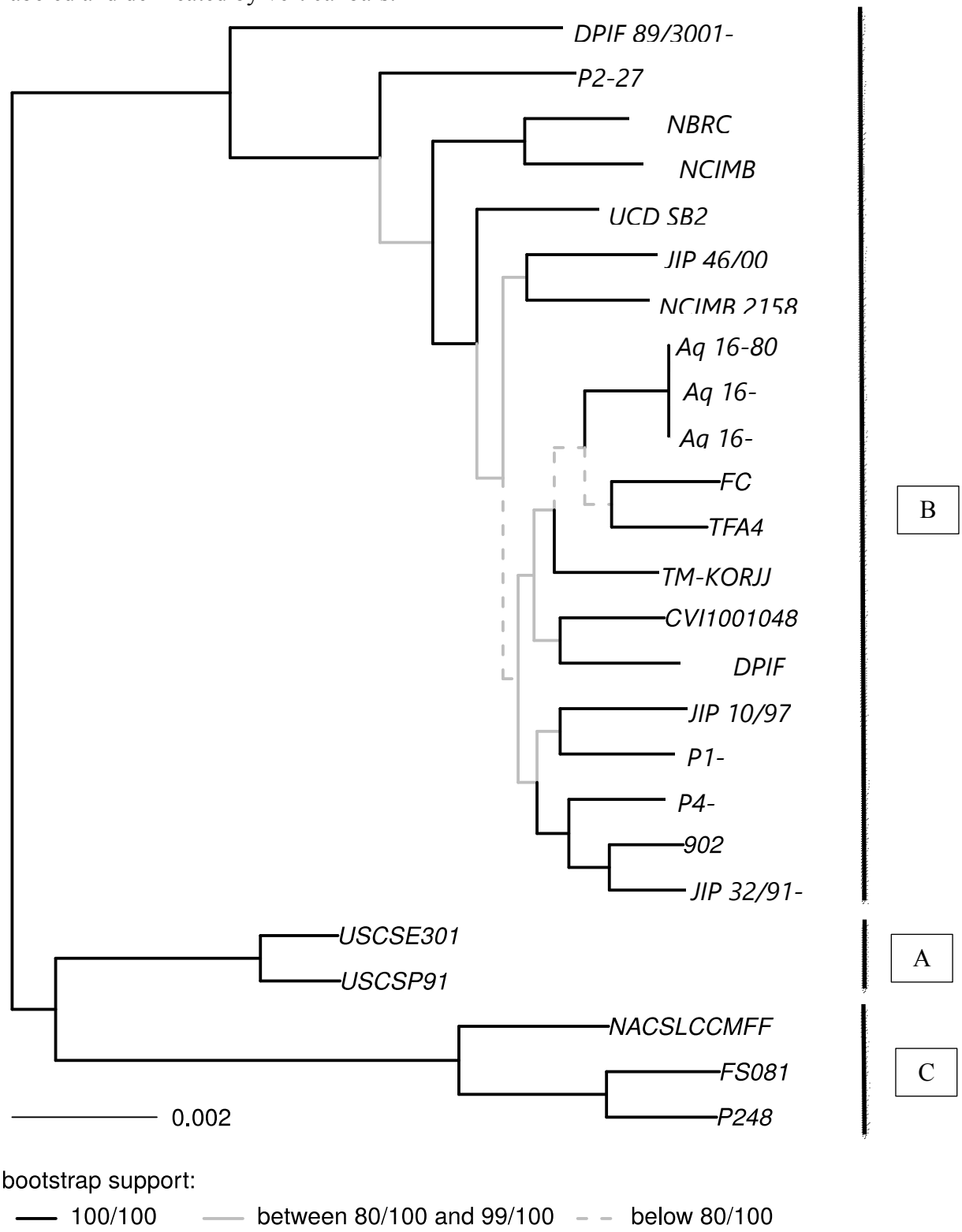
Origin			**Fish host species**	
Australia (Tasmania)	AU		*Acanthopagrus shlegeli*	ASI
Chile	CL		*Atrasctoscion nobilis*	ANS
France	FR		*Carcharias taurus*	CTS
French polynesia	FP		*Dicentrarchus labrax*	DLX
Italy	IT		*Engraulis mordax*	EMX
Japan	JP		*Latris lineata*	LLA
Malta	MT		*Onchorynchus kisutch*	OKH
Spain	SP		*Onchorynchus mykiss*	OMS
United Kingdom (Scotland)	UK		*Pagrus major*	PMR
USA (California)	US		*Paralichthys olivaceus*	POS
no data available	n/d		*Platax orbicularis*	POS
			Salmo salar	SSR
			Scophtalamus maximus	SMS
			Seriola quinqueradiata	SQA
			Solea senegalensis	SSS
			Sparus aurata	SAA
			no data available	n/d

S2 – Average Nucleotide Identity (ANI) of pairwise comparison of the *T. maritimum* isolates

	NAC SLCC MFF	P1-39	P2-27	UCD SB2	JIP 10/97	Aq 16-85	JIP 46/00	TFA4	CVI10001048	Aq 16-89	FC	P4-45	902	NCIMB 2158	Aq 16-88	NBRC 15946	NCIMB 2154T	FS08(1)	DPIF 89/3001-6.2	P2-48	JIP 32/91-4	USC SP9.1	USC SE30.1	DPIF 89/0239-1	TM-KORJJ
NAC SLCC MFF	100,00	98,41	98,45	98,38	98,38	98,37	98,36	98,37	98,37	98,37	98,39	98,40	98,38	98,38	98,37	98,49	98,41	99,26	98,31	99,28	98,38	98,67	98,65	98,38	98,38
P1-39	98,41	100,00	99,31	99,48	99,53	99,54	99,51	99,51	99,57	99,54	99,53	99,53	99,55	99,48	99,54	99,40	99,29	98,29	98,94	98,32	99,54	98,86	98,86	99,50	99,53
P2-27	98,45	99,31	100,00	99,27	99,24	99,31	99,24	99,23	99,27	99,31	99,30	99,28	99,23	99,27	99,31	99,27	99,14	98,29	98,80	98,33	99,26	98,76	98,73	99,25	99,28
UCD SB2	98,38	99,48	99,27	100,00	99,44	99,46	99,43	99,46	99,44	99,46	99,46	99,47	99,44	99,43	99,46	99,36	99,28	98,21	98,94	98,27	99,40	98,78	98,79	99,44	99,47
JIP 10/97	98,38	99,53	99,24	99,44	100,00	99,48	99,43	99,48	99,47	99,48	99,46	99,49	99,47	99,45	99,49	99,36	99,22	98,23	98,91	98,24	99,45	98,82	98,81	99,46	99,47
Aq 16-85	98,37	99,54	99,31	99,46	99,48	100,00	99,47	99,55	99,54	99,99	99,57	99,52	99,48	99,48	100,00	99,41	99,28	98,24	98,93	98,25	99,49	98,80	98,78	99,49	99,51
JIP 46/00	98,36	99,51	99,24	99,43	99,43	99,47	100,00	99,42	99,45	99,47	99,43	99,47	99,43	99,44	99,47	99,39	99,24	98,18	98,89	98,24	99,42	98,77	98,78	99,40	99,49
TFA4	98,37	99,51	99,23	99,46	99,48	99,55	99,42	100,00	99,49	99,56	99,51	99,48	99,47	99,45	99,56	99,39	99,26	98,24	98,93	98,25	99,44	98,81	98,78	99,51	99,49
CVI10001048	98,37	99,57	99,27	99,44	99,47	99,54	99,45	99,49	100,00	99,54	99,49	99,49	99,48	99,47	99,54	99,41	99,26	98,21	98,93	98,23	99,46	98,82	98,80	99,50	99,50
Aq 16-89	98,37	99,54	99,31	99,46	99,48	99,99	99,47	99,56	99,54	100,00	99,56	99,52	99,48	99,48	99,99	99,41	99,28	98,25	98,94	98,26	99,50	98,80	98,78	99,49	99,52
FC	98,39	99,53	99,30	99,46	99,46	99,57	99,43	99,51	99,49	99,56	100,00	99,47	99,49	99,43	99,55	99,41	99,28	98,22	98,92	98,29	99,43	98,82	98,81	99,46	99,50
P4-45	98,40	99,53	99,28	99,47	99,49	99,52	99,47	99,48	99,49	99,52	99,47	100,00	99,57	99,46	99,51	99,38	99,25	98,25	98,90	98,25	99,53	98,81	98,82	99,45	99,52
902	98,38	99,55	99,23	99,44	99,47	99,48	99,43	99,47	99,48	99,48	99,49	99,57	100,00	99,42	99,47	99,38	99,24	98,21	98,93	98,25	99,63	98,82	98,79	99,45	99,50
NCIMB 2158	98,38	99,48	99,27	99,43	99,45	99,48	99,44	99,45	99,47	99,48	99,43	99,46	99,42	100,00	99,49	99,39	99,27	98,18	98,90	98,23	99,45	98,77	98,76	99,42	99,47
Aq 16-88	98,37	99,54	99,31	99,46	99,49	100,00	99,47	99,56	99,54	99,99	99,55	99,51	99,47	99,49	100,00	99,40	99,28	98,24	98,92	98,26	99,49	98,80	98,78	99,49	99,50
NBRC 15946	98,49	99,40	99,27	99,36	99,36	99,41	99,39	99,39	99,41	99,41	99,41	99,38	99,38	99,39	99,40	100,00	99,55	98,35	98,96	98,37	99,36	98,77	98,78	99,38	99,39
NCIMB 2154T	98,41	99,29	99,14	99,28	99,22	99,28	99,24	99,26	99,26	99,28	99,28	99,25	99,24	99,27	99,28	99,55	100,00	98,29	98,86	98,31	99,24	98,72	98,71	99,23	99,27
FS08(1)	99,26	98,29	98,29	98,21	98,23	98,24	98,18	98,24	98,21	98,25	98,22	98,25	98,21	98,18	98,24	98,35	98,29	100,00	98,19	99,54	98,24	98,66	98,65	98,23	98,21
DPIF 89/3001-6.2	98,31	98,94	98,80	98,94	98,91	98,93	98,89	98,93	98,93	98,94	98,92	98,90	98,93	98,90	98,92	98,96	98,86	98,19	100,00	98,21	98,91	98,65	98,69	98,95	98,92
P2-48	99,28	98,32	98,33	98,27	98,24	98,25	98,24	98,25	98,23	98,26	98,29	98,25	98,25	98,23	98,26	98,37	98,31	99,54	98,21	100,00	98,25	98,65	98,64	98,26	98,24
JIP 32/91-4	98,38	99,54	99,26	99,40	99,45	99,49	99,42	99,44	99,46	99,50	99,43	99,53	99,63	99,45	99,49	99,36	99,24	98,24	98,91	98,25	100,00	98,74	98,73	99,42	99,46
USC SP9.1	98,67	98,86	98,76	98,78	98,82	98,80	98,77	98,81	98,82	98,80	98,82	98,81	98,82	98,77	98,80	98,77	98,72	98,66	98,65	98,65	98,74	100,00	99,67	98,80	98,78
USC SE30.1	98,65	98,86	98,73	98,79	98,81	98,78	98,78	98,78	98,80	98,78	98,81	98,82	98,79	98,76	98,78	98,78	98,71	98,65	98,69	98,64	98,73	99,67	100,00	98,82	98,76
DPIF 89/0239-1	98,38	99,50	99,25	99,44	99,46	99,49	99,40	99,51	99,50	99,49	99,46	99,45	99,45	99,42	99,49	99,38	99,23	98,23	98,95	98,26	99,42	98,80	98,82	100,00	99,48
TM-KORJJ	98,38	99,53	99,28	99,47	99,47	99,51	99,49	99,49	99,50	99,52	99,50	99,52	99,50	99,47	99,50	99,39	99,27	98,21	98,92	98,24	99,46	98,78	98,76	99,48	100,00

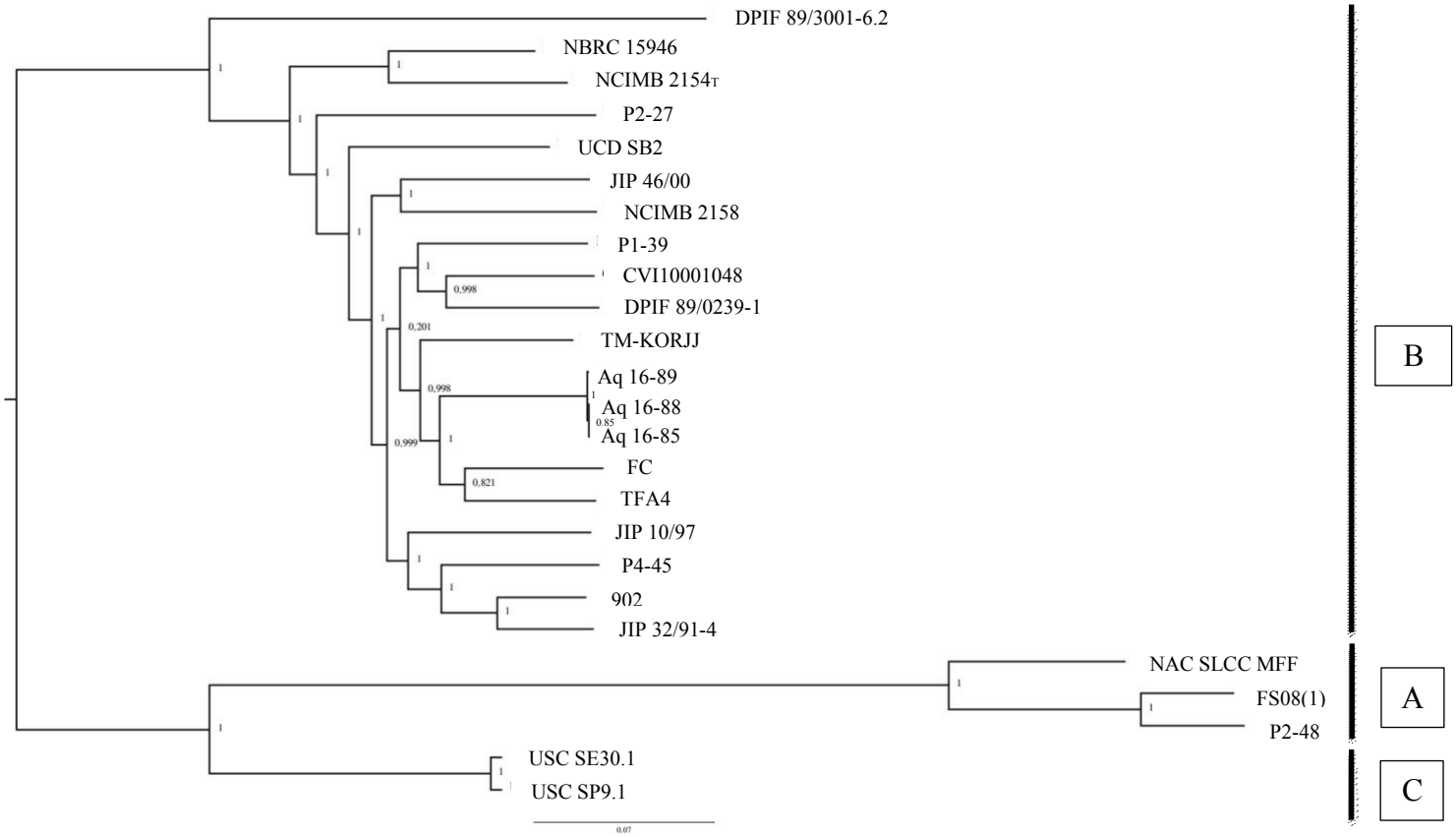
S3A – Parsimony tree

The tree is based on the alignment made by Snippy. It is reconstructed using parsimony method as implemented in *dnapars* (Phylip package v3.6). Bootstrap support of each branch is computed from 100 bootstrap replicates. The three clades designated A, B, and C are labeled and delineated by vertical bars.



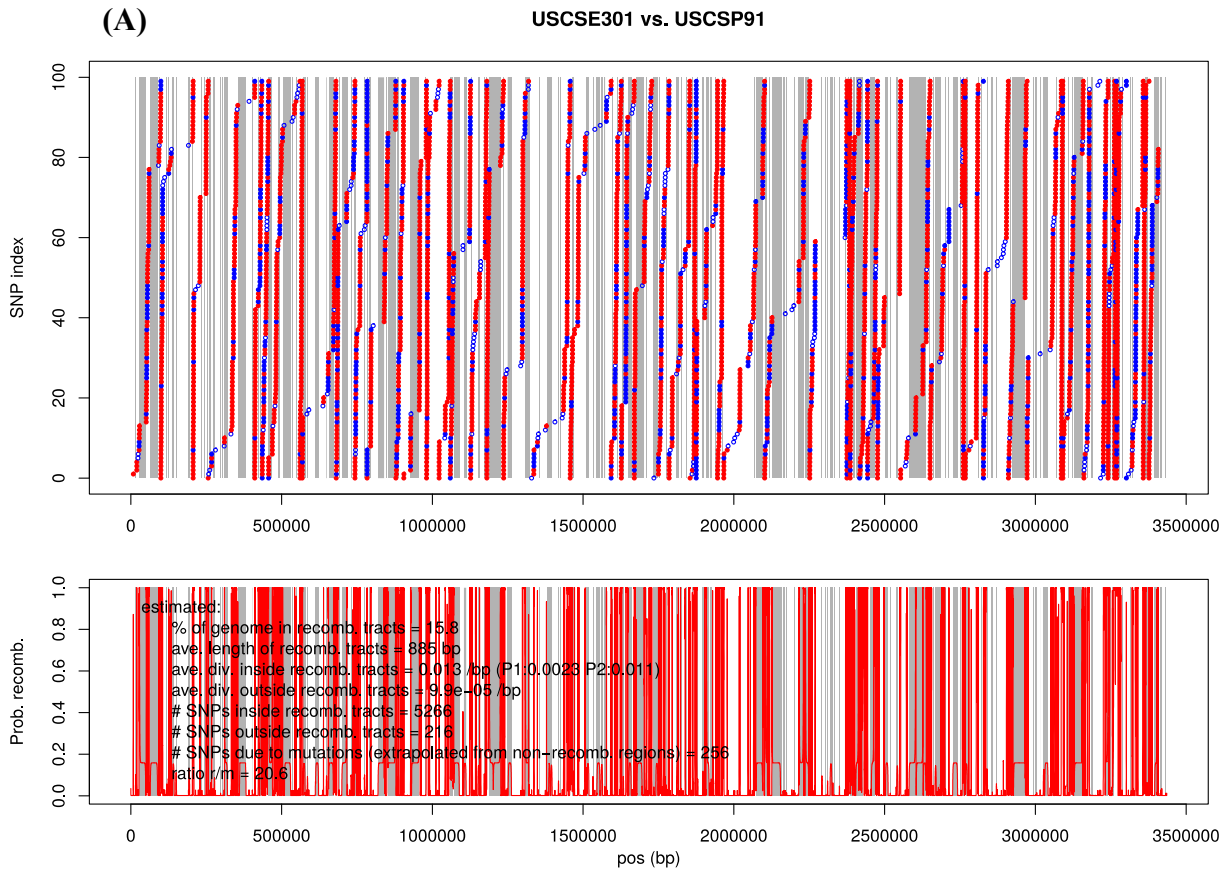
S3B – Maximum likelihood, Gubbins-based phylogenetic tree

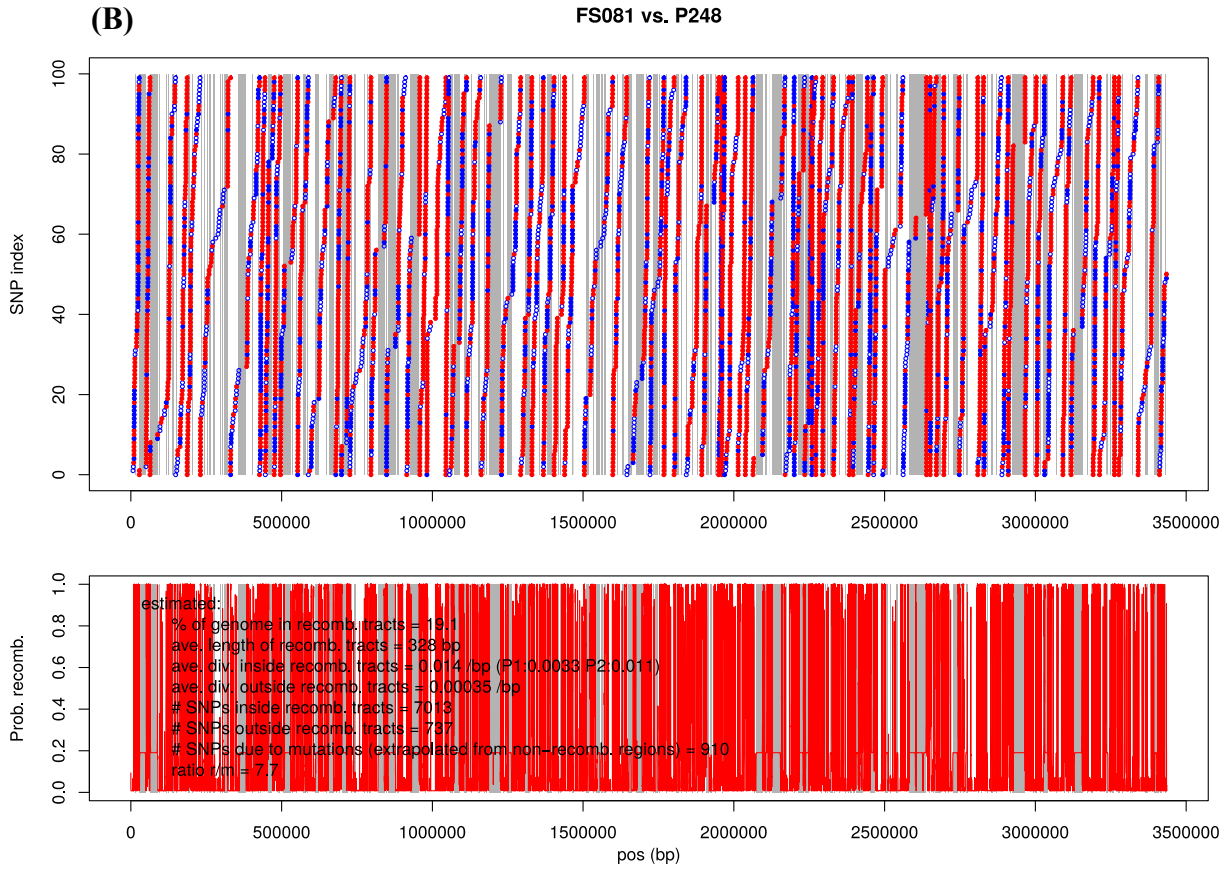
The tree was obtained from the whole genome alignment of 25 *T. maritimum* strains at the fifth and final iteration of Gubbins. Statistical support of nodes is indicated. The three clades designated A, B, and C are labeled and delineated by vertical bars.



S4 – Overview of recombination tracts between two pairs of *T. maritimum* isolates.

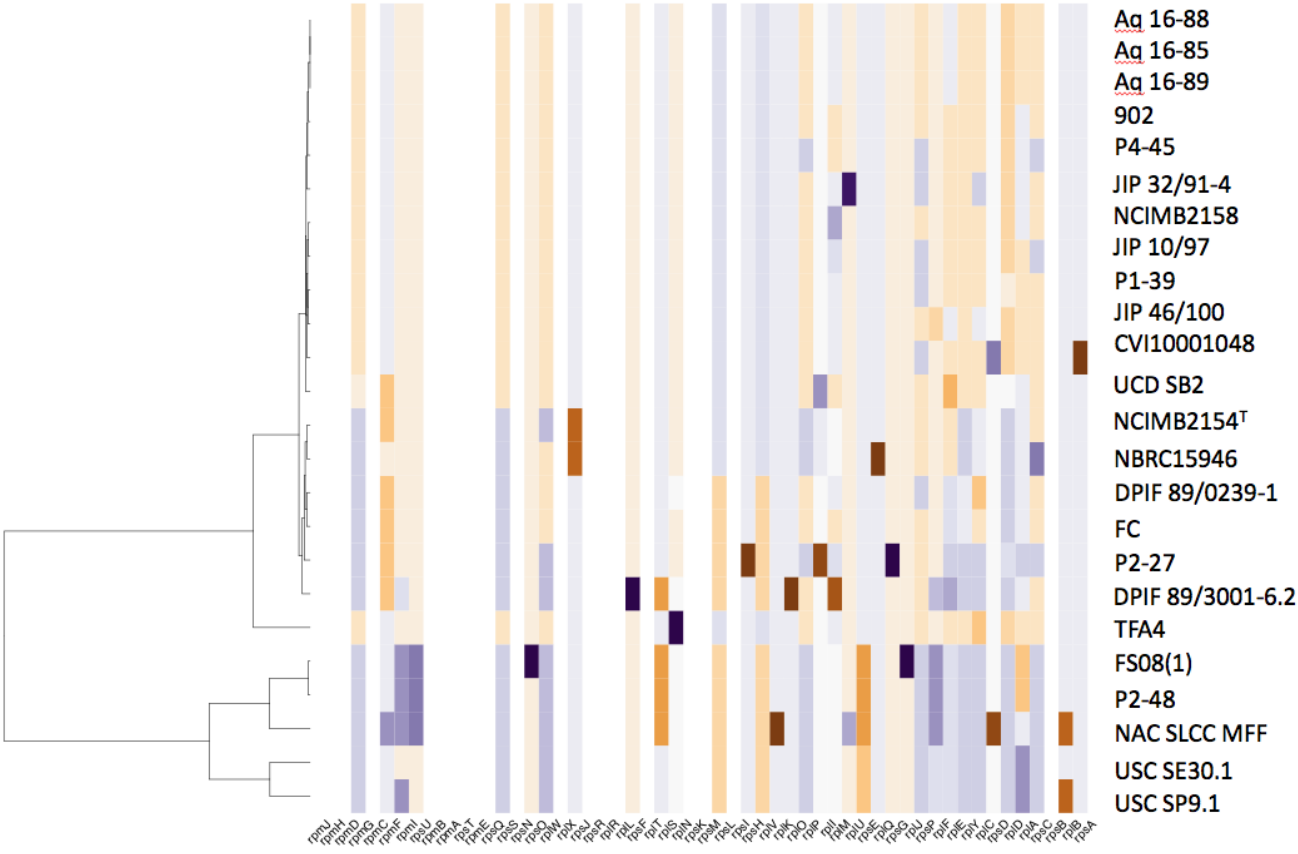
SNPs and recombination tracts between two closely related isolates. In S4A the strains compared are USC SP9.1 and USC SE30.1. In S4B the strains compared are FS08(1) and P2−48. (Upper) Positions of the SNPs along the genomes. SNP index is reset every 100 SNPs for this representation. Each dot corresponds to one SNP in the comparison between the two considered isolates. Colors distinguish two types of polymorphism: in blue, polymorphism observed only between the two considered genomes; in red, polymorphism observed also among the other sequenced genomes. Areas in gray correspond to regions not covered by our alignments. SNPs in regions where probability is < 0.5 (i.e. outside predicted recombination tracts) are represented by open symbols (blue circles). (Lower) Probability of recombination tract as computed with the HMM. Estimations of the % of genome in recombination tracts, average length of recombination tracts, average nucleotide diversity inside recombination tracts, average nucleotide diversity outside recombination tracts, number of SNPs inside recombination tracts, number of outside recombination tracts, SNPs due to mutations (extrapolated from non−recombined regions), ratio r/m are included.


(A)
USCSE301 vs. USCSP91
SNP index
Prob. recomb.
estimated:
% of genome in recomb. tracts = 15.8
ave. length of recomb. tracts = 885 bp
ave. div. inside recomb. tracts = 0.013 /bp (P1:0.0023 P2:0.011)
ave. div. outside recomb. tracts = 9.9e-05 /bp
SNPs inside recomb. tracts = 5266
SNPs outside recomb. tracts = 216
SNPs due to mutations (extrapolated from non-recomb. regions) = 256
ratio r/m = 20.6
pos (bp)


(B)
FS081 vs. P248
SNP index
Prob. recomb.
estimated:
% of genome in recomb. tracts = 19.1
ave. length of recomb. tracts = 328 bp
ave. div. inside recomb. tracts = 0.014 /bp (P1:0.0033 P2:0.011)
ave. div. outside recomb. tracts = 0.00035 /bp
SNPs inside recomb. tracts = 7013
SNPs outside recomb. tracts = 737
SNPs due to mutations (extrapolated from non-recomb. regions) = 910
ratio r/m = 7.7
pos (bp)

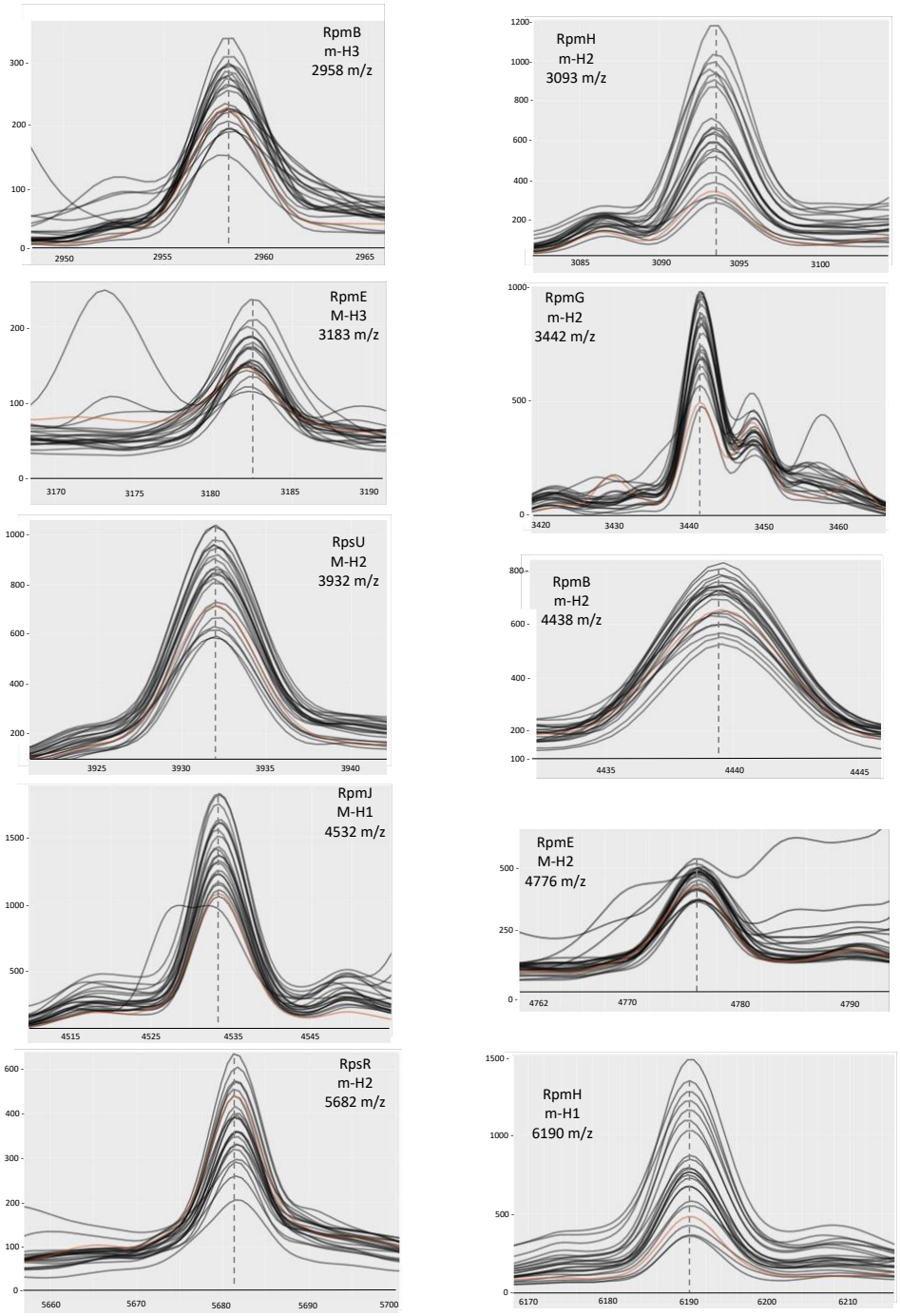
S5 – Heatmap displaying the diversity of ribosomal protein weights.

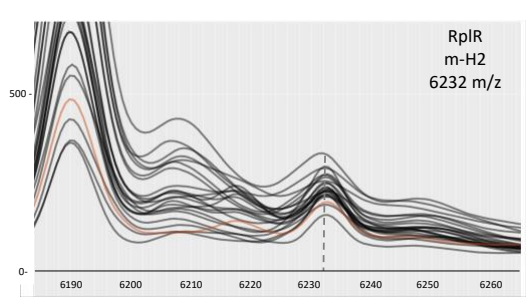
Lines correspond to strains and columns correspond to ribosomal proteins in ascending order by weight (from left to right). White lines indicate no variation of the weight of the corresponding proteins (i.e., monomorphic proteins). In purple, proteins with weight higher than the mean and in orange proteins with weight lower than the mean (i.e., polymorphic proteins).

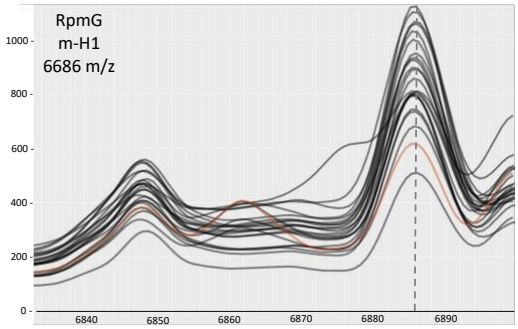


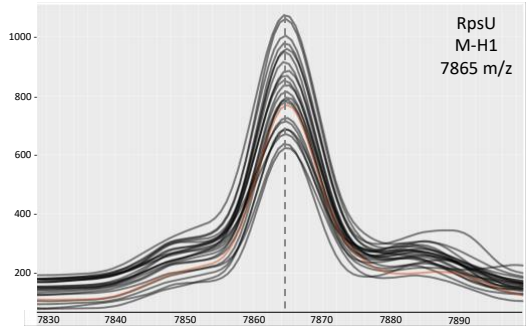
S6 – Monomorphic biomarker peaks.

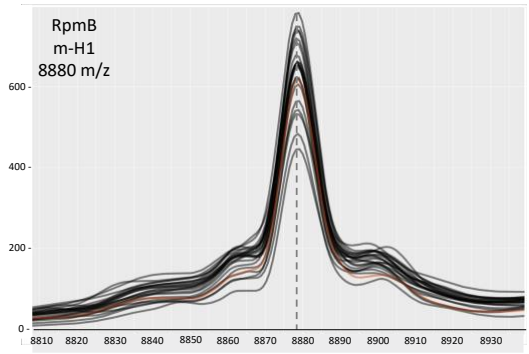
Screenshots of the 18 conserved peaks produced by 9 ribosomal monomorphic proteins with several degrees of ionization. The m/z values are highlighted by dotted lines and cover the entire spectrum. The red curve corresponds to the *T. maritimum* type strain average spectra. For each peak, the corresponding ribosomal protein is indicated with the degree of ionization (H1, H2 and H3 corresponding to 1, 2 and 3 H^+) and the presence (M) or absence (m) of the first methionine. Color code: red line for *T. maritimum* type strain NCIMB 2154^T and black for sequenced *T. maritimum* isolates.

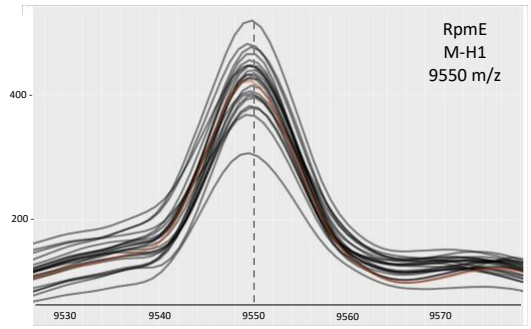


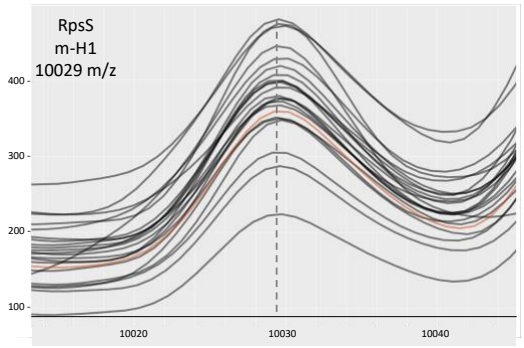

RplR
m-H2
6232 m/z
500
0
6190
6200
6210
6220
6230
6240
6250
6260

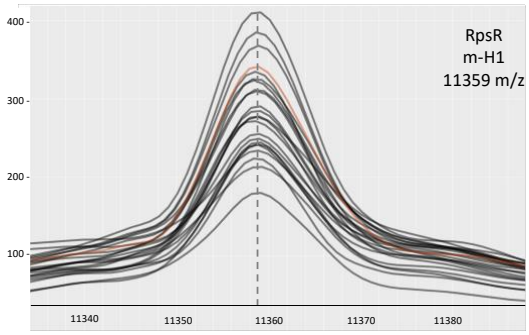

RpmG
m-H1
6686 m/z
1000
800
600
400
200
0
6840
6850
6860
6870
6880
6890

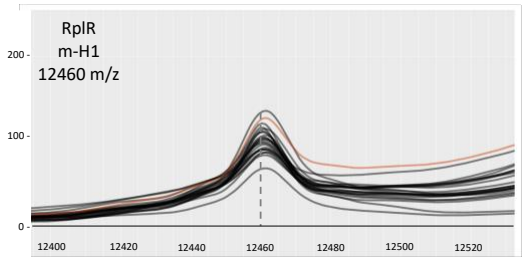

RpsU
M-H1
7865 m/z
1000
800
600
400
200
7830
7840
7850
7860
7870
7880
7890


RpmB
m-H1
8880 m/z
600
400
200
0
8810
8820
8830
8840
8850
8860
8870
8880
8890
8900
8910
8920
8930


RpmE
M-H1
9550 m/z
400
200
9530
9540
9550
9560
9570


RpsS
m-H1
10029 m/z
400
300
200
100
10020
10030
10040


RpsR
m-H1
11359 m/z
400
300
200
100
11340
11350
11360
11370
11380


RplR
m-H1
12460 m/z
200
100
0
12400
12420
12440
12460
12480
12500
12520

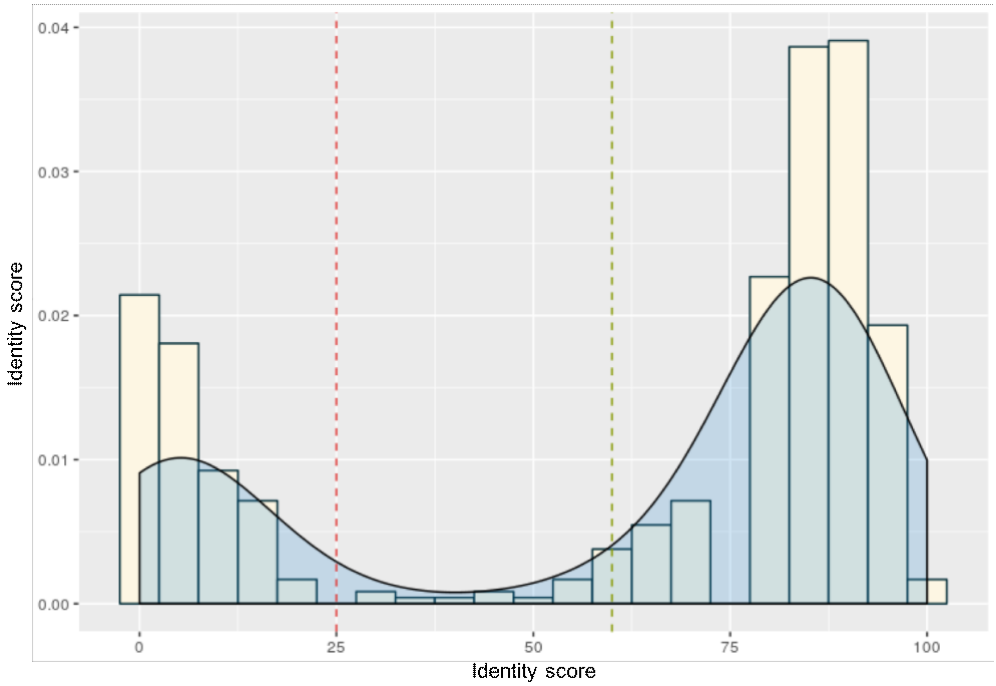
S7 – Quality control and *T. maritimum* species identification

The full dataset is composed of representatives of 24 *Tenacibaculum* species including 135 isolates belonging to the species *T. maritimum*. It encompasses 476 independent acquisitions (one acquisition corresponds to an average spectrum of several technical replicates) corresponding to 5,102 spectra that includes technical and biological replicates. This dataset was divided into two groups: the positive control group (*T. maritimum* isolates) and the negative control group (the type strains of 23 other *Tenacibaculum* species). In order to confirm that an isolate belongs to the species *T. maritimum,* the spectra were scanned to identify the 18 *T. maritimum* monomorphic biomarkers. However, some of these biomarkers could be absent from a number of *T. maritimum* strains. Reciprocally, strains that do not belong to the *T. maritimum* species may possess some *T. maritimum* monomorphic biomarkers. In order to setup a *T. maritimum* species identity threshold, a value corresponding to the number of monomorphic biomarkers identified in a single sample was computed. The monomorphic biomarkers frequency plot obtained shows a bimodal distribution (Figure A). All samples with a score above 60% correspond to *bona fide T. maritimum* isolates while all samples with a score below 25% belong to other *Tenacibaculum* species.

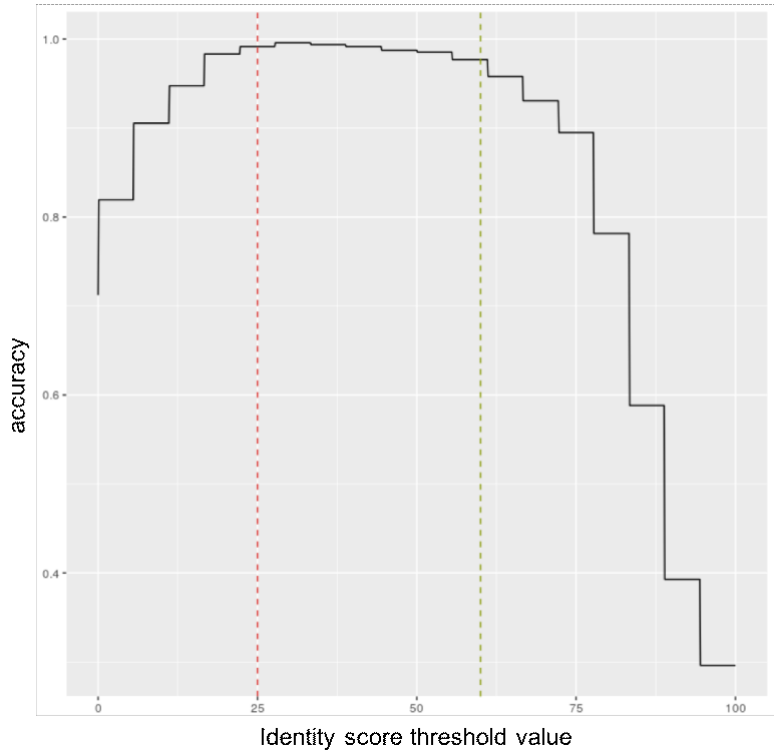
True and false positives correspond to isolates correctly and incorrectly identified as *T. maritimum* (TP and FP), respectively. On the other hand, true and false negatives correspond to correctly and incorrectly rejected isolates (TN and FN, respectively). Positive isolates correspond to those having an identity value above a defined threshold. Using the full dataset, the number of TP and FP and the number of TN and FN were counted. The accuracy of the tool [i.e., (TP+TN) / (TP+TN+FP+FN)] was then computed by increasing the threshold value from 0% to 100% by a 0,1% step. One could observe than the accuracy varies from 0% to 100% and is maximal (i.e., above 97%) between 25% and 60% of threshold value (Figure B). It is therefore proposed that a 60% threshold value safely identifies isolates as belonging to the *T. maritimum* species. Using this 60% threshold value, only 10 false negatives (i.e., *bona fide T. maritimum* isolates discarded) and 0 false positive (i.e., not *T. maritimum* isolates) were found out of 476 acquisitions. Among the 10 false negatives, 3 resulted from the extraction protocol, 3 from the Direct Deposit with Formic Acid (DDFA) protocol and 4 from the ethanol conservation protocol (see section "Testing alternative sample preparation for MALDI-TOF MS**"**). One can hypothesize that these rare false negatives correspond to technical problems. Indeed, by performing new spectra acquisitions on the 3 isolates previously analyzed by the extraction protocol, the identification score reached at least 88%, far above the safe threshold value of 60%, demonstrating that the original spectra were likely

faulty. Finally, spectra from other *Tenacibaculum* species all had a score below 23%, far below the confidence threshold.

(A)

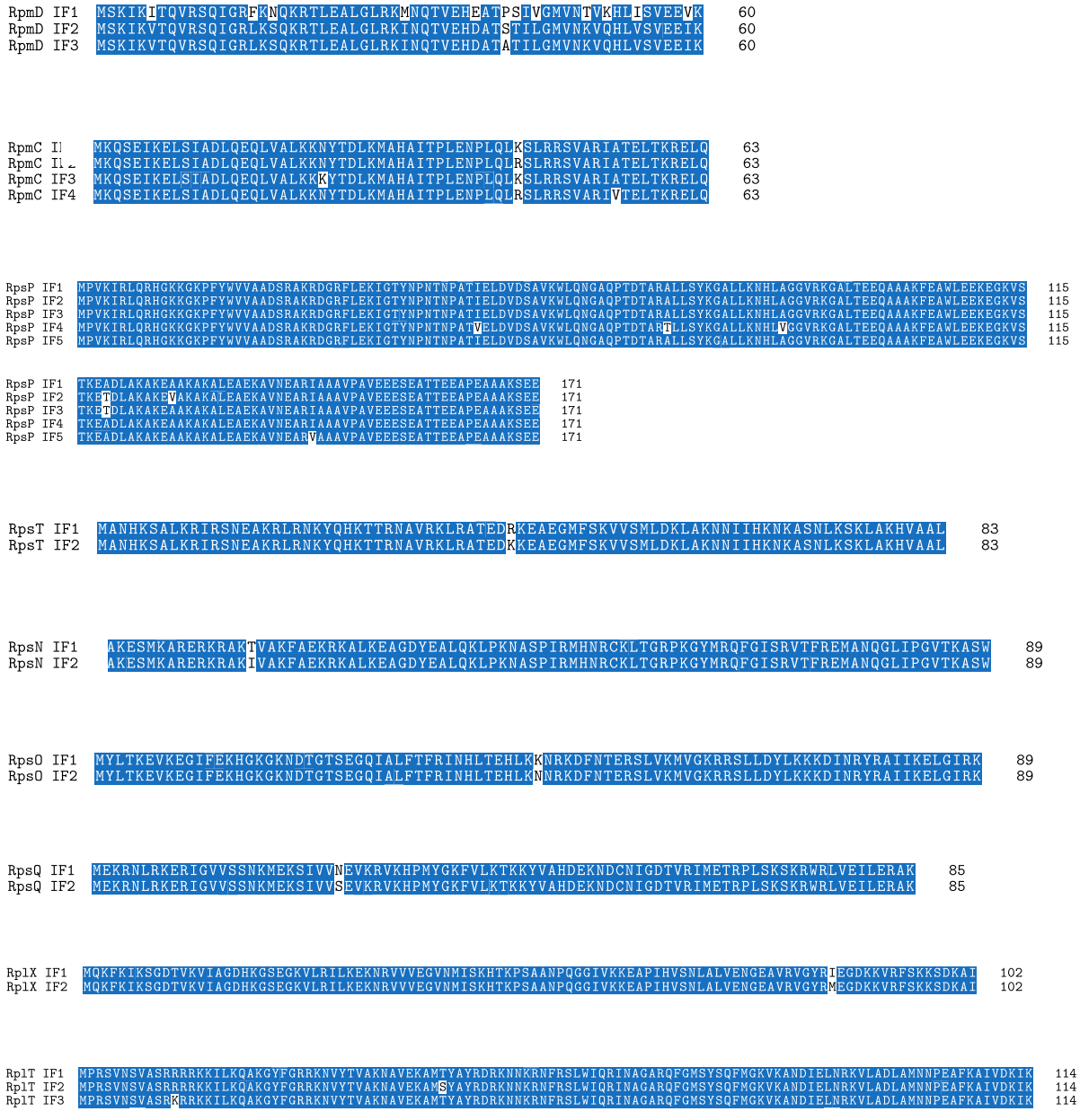


(B)



S8 – Amino-acid polymorphism of the 9 retained biomarkers

Protein sequence alignments of the 9 ribosomal proteins selected as polymorphic biomarkers and their corresponding isoform (IF).



C3 – Discussion

Cet article présente une analyse détaillée des spectres MALDI-TOF et définit une nouvelle approche de typage pour l'espèce *T. maritimum* (Figure 21) basée sur l'apport des données génomiques pour identifier des biomarqueurs caractéristiques. L'intérêt de la méthode repose sur le fait qu'une fois le schéma établi, il n'est plus nécessaire de séquencer d'autres génomes. Le typage se base sur des biomarqueurs précis dont la masse attendue est connue. S'il devait arriver qu'un échantillon possède une forme inconnue, un simple séquençage du gène correspondant par la méthode Sanger serait suffisant pour identifier le nouveau polymorphisme observé dans les spectres. L'utilisation de biomarqueurs permet d'avoir un outil de typage reproductible et fiable. Bien qu'avec le jeu de données utilisé il n'a pas été possible de définir des liens épidémiologiques clairs, cet outil devrait être en mesure de décrire les populations bactériennes présentes dans les élevages avec une précision suffisamment fine pour réaliser du suivi épidémiologique à une échelle locale.

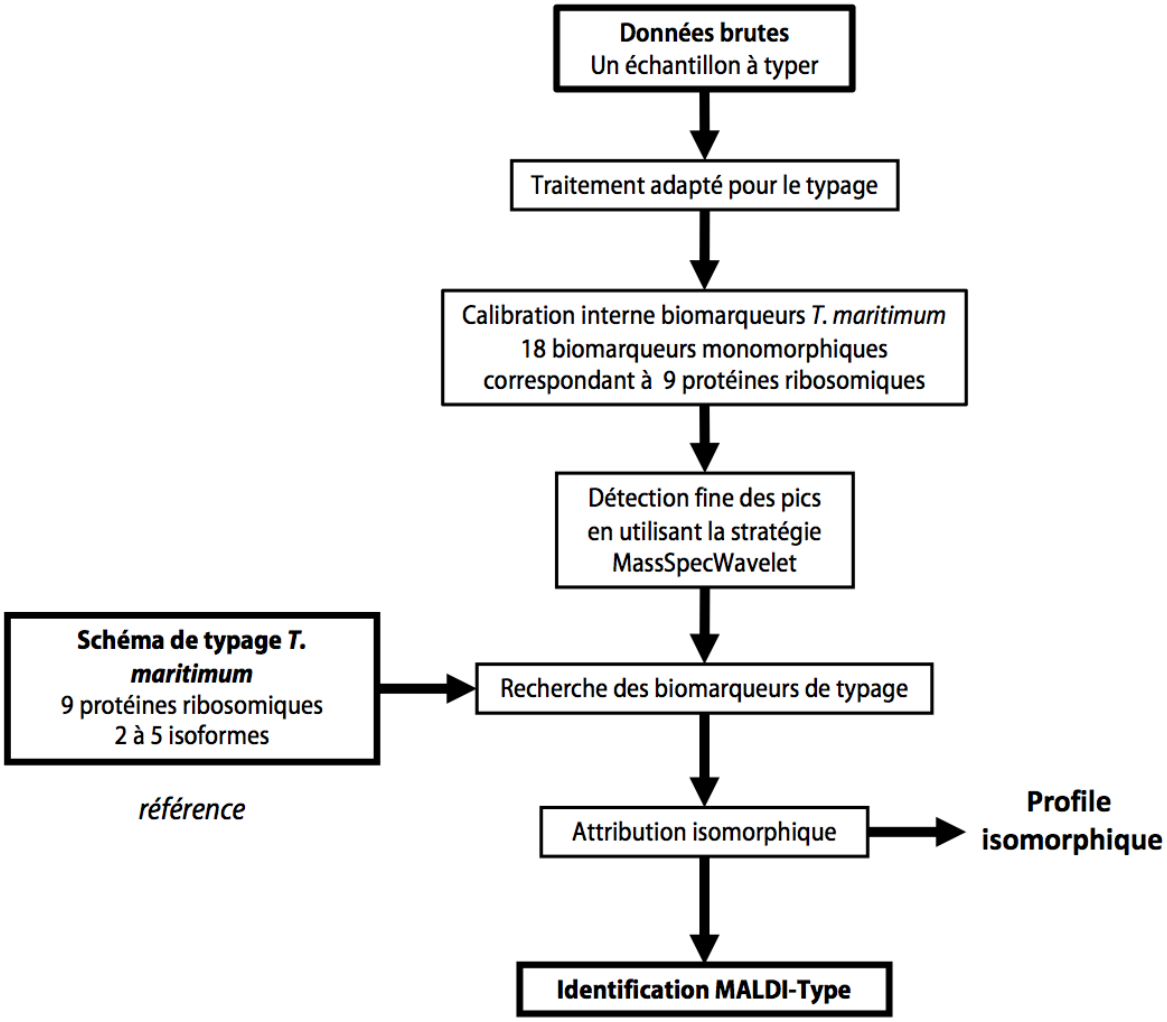


Figure 21: Schéma simplifié du typage d'un isolat par spectrométrie de masse MALDI-TOF

Un outil en ligne a également été développé qui intègre les algorithmes élaborés durant le projet doctoral. Il permet de comparer un isolat à l'intégralité des espèces

appartenant au genre *Tenacibaculum* présentes dans notre collection. Il intègre les deux méthodes d'identification présentées dans la Partie 2. Si l'échantillon est identifié comme appartenant à l'espèce *T. maritimum*, l'outil applique le schéma décrit dans cet article et propose une attribution pour chacun des 9 biomarqueurs retenus. Lorsqu'ils sont tous identifiés dans un échantillon, l'outil affiche le MALDI-type correspondant (s'il s'agit d'une combinaison appartenant à un des 20 MTs actuellement décrits). Le principe de cet outil est schématisé dans la Figure 22 ci-dessous.

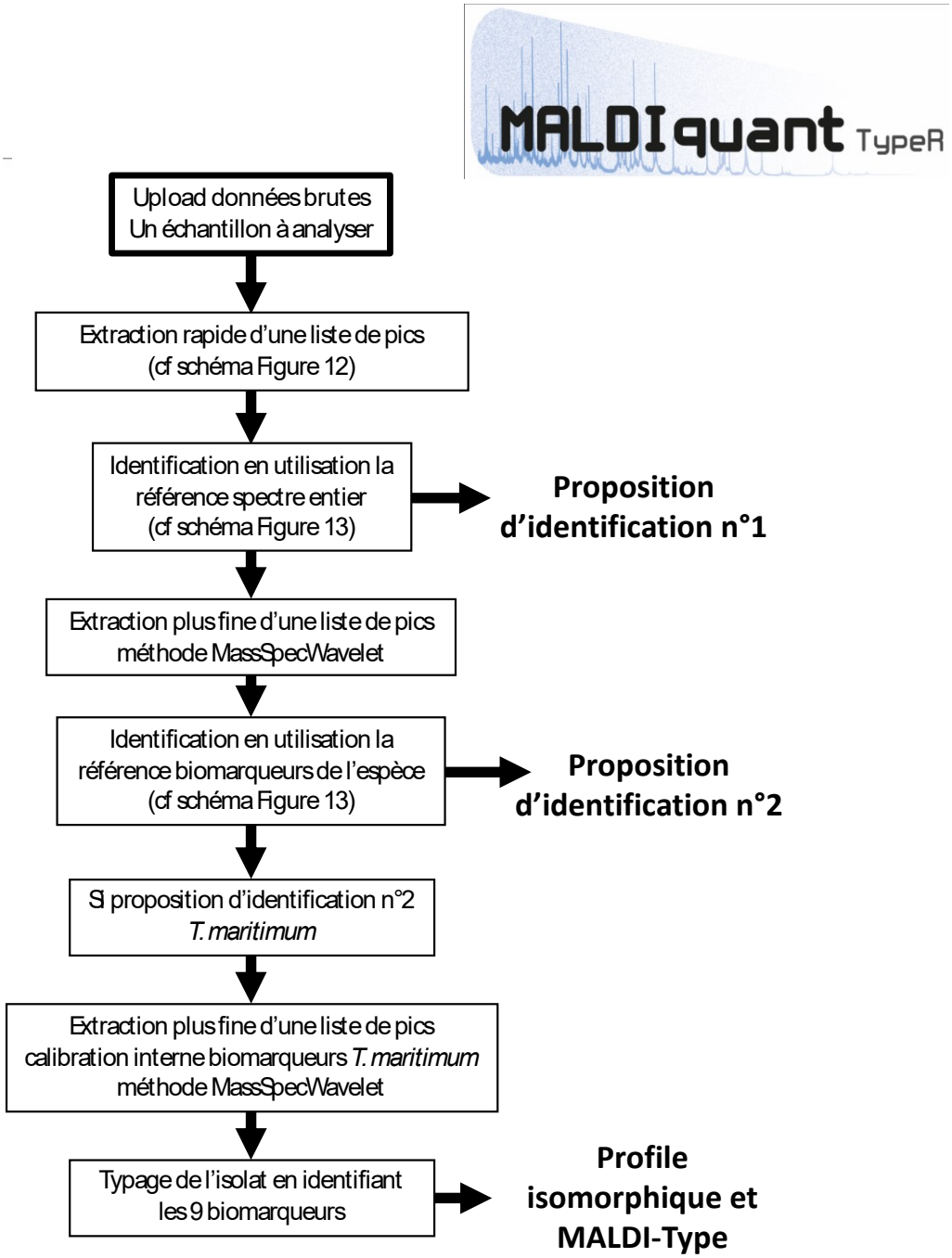


Figure 22: Schéma du flux d'analyse disponible sur l'application web MALDIquantTypeR

C4 – Recul critique de la méthode de typage par MALDI-TOF

Le typage par MALDI-TOF présente de nombreux avantages. Même si le schéma présenté ici est (en théorie) équivalent au génotype, cela n'est pas forcément toujours le cas. Dans notre étude, chaque isoforme est associé à des masses bien distinctes. Néanmoins, deux facteurs de confusion peuvent apparaître dans les spectres.

Le premier facteur consiste en une mutation dans la séquence du gène, entraînant une légère modification de la masse de la protéine correspondante : si celle-ci est inférieure à 5 Da, elle sera quasiment impossible à distinguer de la forme la plus proche. Une mutation aussi petite peut être le résultat :

- D'une seule mutation, comme un changement entre une proline et une valine entrainant une variation de 2Da (http://prospector.ucsf.edu/prospector/html/misc/mutation.htm).
- De plusieurs mutations avec des changements de masses qui se compensent, entrainant un changement mineur (ou nul) de masse.

Cela indique que la diversité génétique des biomarqueurs choisis est, peut-être, sous-estimée.

Le second facteur est plus général et est valable dans tous les spectres. Ceux-ci correspondent à une empreinte moléculaire qui peut être vue comme l'ombre projetée par une partie des protéines exprimées par la bactérie. En théorie, une même masse observée en MALDI-TOF peut être produite par différentes protéines.

La caractérisation d'isolats par MALDI-TOF est une approche efficace lorsqu'il s'agit d'étudier un grand nombre d'échantillons. Cette rapidité d'acquisition des données et d'analyse est un avantage comparé à d'autres méthodes comme la MLST. Comparée à cette dernière, l'approche par MALDI-TOF sacrifie un peu en précision pour gagner grandement en vitesse d'analyse et en coût. C'est donc l'outil idéal pour des analyses de *pre-screening* d'un jeu de données permettant de sélectionner des isolats qui pourraient être analysés plus en détail (par exemple lorsqu'ils présentent de nouvelles isoformes de biomarqueurs) par MLST ou par séquençage de génomes complets.

Étant donné la nature des échantillons étudiés dans le cadre du développement de notre méthode, nous avons vu que l'identification de liens épidémiologiques était difficile, mais c'était également le cas avec l'approche MLST (Habib et al., 2014). La prochaine étape importante serait de pouvoir appliquer un schéma de typage semblable mais sur un jeu de données bien décrit, dans lequel les relations épidémiologiques sont connues. Cela permettrait de mieux évaluer l'apport possible du MALDI-typage pour des études épidémiologiques futures.

IV – Discussion générale & Conclusion

A – Un travail collaboratif et interdisciplinaire

Durant ce projet doctoral, un travail considérable a été réalisé par de nombreuses personnes pour générer l'ensemble des données nécessaires à sa réalisation. Ce travail n'est pas forcément appréciable de prime abord, mais il est fondamental a permis de développer des outils reposant sur des données les plus solides possibles.

La première partie, commencée bien en amont du projet, fut l'établissement de la collection des souches types de l'ensemble des espèces de *Tenacibaculum* décrites dans la littérature. Celles-ci furent obtenues à partir de différentes collections internationales par Jean-François Bernardet qui s'est également chargé de les contrôler en utilisant les caractéristiques phénotypiques détaillées dans les publications décrivant les espèces correspondantes.

Ensuite, les génomes de ces souches ont été séquencés. Les séquences d'ARNr 16S obtenues ont été comparées avec celles présentes dans les bases de données (par exemple celle du NCBI) et déposées lors de la description des espèces afin de valider nos propres données génomiques. Nous avons également séquencé les génomes de 23 souches appartenant à l'espèce *T. maritimum.* L'ensemble des génomes a ensuite été intégré dans la plateforme MicroScope. Cela nous a permis de vérifier et de corriger manuellement au besoin les annotations automatiques. Ce travail long et fastidieux est nécessaire par exemple pour supprimer les sur-prédictions de gènes pouvant diminuer la qualité des études de génomique comparée. Il a surtout permis d'identifier des gènes (ou groupes de gènes) en lien avec le mode de vie de ces bactéries (gènes de virulence par exemple pour les espèces pathogènes). Ce travail a également permis d'effectuer des analyses d'ANI régulières permettant de vérifier l'affiliation taxonomique des génomes déposés dans les bases de données publiques.

Pour l'analyse par MALDI-TOF des isolats appartenant à l'espèce *T. maritimum,* nous avons utilisé plus d'une centaine d'isolats provenant de différentes collections. De plus, pour la plupart d'entre eux, 3 protocoles différents de préparation d'échantillons ont été utilisés en parallèle afin de tester la robustesse des méthodes développées. Ce travail a été réalisé par Jean-François Bernardet (INRA), Arnaud Marie (Labofarm, Finalab) et Frédéric Bourgeon (BioChêneVert, Finalab). Tout ce travail en amont a nécessité une réelle expertise en microbiologie. Finalement, l'effort investi est récompensé par la mise au point d'outils fonctionnels basés sur des données robustes.

Un autre aspect important à mentionner est le développement de l'application web MALDIquantTypeR programmé en R Shiny. La création d'un site web est un processus chronophage. C'est un challenge technique à l'interface entre la programmation R classique et la programmation web, cette dernière possédant des spécificités bien particulières. Toutefois, cet investissement est important puisqu'il permet de partager l'intégralité du travail réalisé dans le cadre de ce projet au travers d'une interface web utilisable par d'autres équipes. Cela permet également d'ajouter une plus-value intéressante au second article qui décrit la méthode de typage puisqu'elle est rendue accessible à des tiers.

B – L'épidémiologie des ténacibaculoses

Les ténacibaculoses sont des pathologies peu connues. Sur la base de données PubMed, il n'existe qu'environ 150 publications portant sur des bactéries du genre *Tenacibaculum*, correspondant essentiellement aux descriptions de nouvelles espèces, de nouveaux génomes, ou de typage par PCR ou par sérologie. Il n'existe pas encore de données épidémiologiques exploitables, la plupart des publications étant même contradictoires. Il faudrait par exemple, afin d'établir un modèle pour ces maladies, recueillir des données sur un cycle d'élevage complet. Il faudrait également mettre en place des études systématiques des pisciculture dans lesquelles des épisodes de ténacibaculose interviennent régulièrement. Il faudrait également identifier le (ou les) réservoir(s) des espèces pathogènes, ainsi que vérifier l'existence de porteurs sains (pour l'instant aucun cas n'a été décrit). La maladie se déclenche-t-elle obligatoirement lorsque l'agent pathogène est mis en contact avec les poissons ? L'agent pathogène fait-il partie de l'environnement normal des poissons et la maladie se déclare-t-elle plutôt lorsque les poissons sont « fragilisés » (stress dû à la densité de population, aux transports, aux manipulations, etc…) ? Répondre à ces questions permettrait non seulement d'améliorer nos connaissances générales sur ces pathologies mais également de définir des traitements mieux adaptés et surtout de mettre en place des mesures prophylactiques. Cela s'inscrit dans une gestion durable des pisciculture dans laquelle la prévention permettrait de réduire considérablement l'usage des antibiotiques et des traitements chimiques. Il est alors nécessaire d'avoir des outils efficaces pour conduire de telles études c'est à dire des outils d'identification et de typage fiables, rapides et peu coûteux. La spectrométrie de masse MALDI-TOF est une technologie pouvant répondre à ces critères. Elle offre des solutions pertinentes en termes de diagnostic. Les spectres de masse sont de véritables empreintes moléculaires qui, telles les empreintes digitales sur une scène de crime, sont en fait de véritables cartes d'identité des espèces bactériennes. A l'aide de données de référence solides, ces empreintes facilitent et accélèrent énormément l'identification bactérienne. De plus, il existe dans ces empreintes des signatures plus subtiles permettant de différencier des groupes de souches au sein de la même espèce, offrant ainsi un outil de typage possiblement très intéressant. Néanmoins, le développement de méthodes aussi précises demande un travail important alliant génomique, recherche de biomarqueurs, mise en place d'outils informatiques et recoupement avec des données épidémiologiques.

C – Le spectromètre de masse MALDI-TOF, outil de première ligne pour le diagnostic vétérinaire.

L'identification d'un isolat bactérien est une application de routine de la spectrométrie de masse MALDI-TOF pour les laboratoires de bactériologie. En outre, cette technologie permet également de caractériser les isolats. En l'absence de méthode générique pour ce genre d'approches, il existe malgré tous des recommandations à suivre afin de développer correctement ce type de stratégie. Celles-ci préconisent d'abord un protocole de préparation le plus homogène possible entre les différents échantillons. Cela a pour objectif de gommer au maximum les fluctuations aléatoires propres à la nature analogique des empreintes moléculaires, celles-ci étant sensibles aux caractéristiques physicochimiques du dépôt (homogénéité, forme, concentration, etc…). Dans le cadre du projet, nous avons choisi de réaliser 12 spectres répartis en 3 enregistrements pour chacun des 4 dépôts réalisés. Nous avons également choisi de réaliser et comparer 3 protocoles différents de préparation. Comme vu précédemment, le protocole extraction est généralement le plus fiable (aussi bien sur le plan de l'identification de l'espèce que du typage d'isolats). Ce résultat était attendu, car l'extrait est enrichi en protéines solubles et débarrassé des débris.

Les méthodes ont été réalisées à partir des données issues du protocole extraction. Les autres protocoles (dépôt direct et éthanol) ont permis d'une part de tester la robustesse de la méthode et d'autre part de tester des protocoles facilitant l'envoi d'isolats entre laboratoire pour l'analyse MALDI-TOF. En effet, malgré le faible coût d'une acquisition MALDI-TOF, l'achat et l'entretien de l'appareil reste un investissement important qui ne peut pas être engagé par tous les laboratoires. Au contraire, l'envoi d'échantillons fixés en éthanol est peu coûteux. Cette approche a été expérimentée à la fin de ce projet doctoral dans le cadre d'une collaboration entre l'INRA, BioChêneVert (qui réalisa les acquisitions MALDI-TOF dans le cadre de mon projet doctoral) et une équipe chilienne. Celle-ci a déjà envoyé une centaine d'isolats appartenant au cluster comprenant *T. dicentrarchi,* "*T. finnmarkense*" et *T. piscium*. Les outils développés seront également utilisés dans le cadre d'un second projet doctoral réalisé par Pierre Lopez et effectué en partenariat entre notre équipe et l'équipe de Denis Saulnier à l'IFREMER de Tahiti et Labofarm/BioChêneVert. Dans le cadre de celui-ci, plus de deux cents isolats de *Tenacibaculum* ont été collectés, dont de nombreux isolats de *T. maritimum*. Les outils détaillés ci-dessus pourront être utilisés afin de vérifier l'identification de l'ensemble des isolats de cette nouvelle collection et permettront d'attribuer un MALDI-Type à l'ensemble des isolats de *T. maritimum*. Ces outils permettront donc d'étudier la diversité locale des isolats appartenant au genre *Tenacibaculum*. De plus, de nouveaux MALDI-Types ainsi que de nouvelles isoformes des 9 biomarqueurs de typage pourront potentiellement être observés.

Le spectromètre de masse MALDI-TOF est donc un outil intéressant pour l'identification et le typage d'isolats bactériens qui peuvent s'avérer utiles à la fois dans le cadre de recherches académiques (par exemple pour des études épidémiologiques) et à la fois pour soutenir le diagnostic de routine en médecine vétérinaire (outil de screening de première ligne sur le terrain). Cela s'explique par le faible coût, la rapidité et la fiabilité des analyses MALDI-TOF pouvant également être réalisées sur un large nombre d'isolats (une plaque MALDI-TOF possédant 96 cibles). Bien que ce dernier soit fréquemment utilisé en médecine humaine pour des diagnostics de routine, son utilisation est moins répandue en médecine vétérinaire et encore moins pour l'aquaculture et la pisciculture. Il semble donc important de communiquer sur l'impact positif de l'utilisation du spectromètre de masse MALDI-TOF et plus particulièrement avec les vétérinaires aquacoles.

D – Conclusion

Le travail synthétisé dans ce manuscrit de thèse s'inscrit dans la continuité des projets de recherche menés par l'équipe Infection et Immunité des Poissons de l'unité Virologie et Immunologie Moléculaires à l'INRA de Jouy-en-Josas.

Ce travail a permis de publier des données de génomique fiables qui serviront de base robuste pour d'autres études de génomique comparée. Le gain de connaissances fondamentales apporté par les génomes des bactéries appartenant au genre *Tenacibaculum* a également permis de développer de nouveaux outils d'identification et de typage utilisant la spectrométrie de masse MALDI-TOF et d'autres en cours de développement tels que la qPCR et la PCR multiplex pour le sérotypage de souches de *T. maritimum*). Notre travail illustre également la méthode permettant de combiner une information génomique à une information spectrale.

Les élevages de poissons marins sont particulièrement sensibles aux maladies. L'épidémiologie des ténacibaculoses est inconnue. Des études de plus grande ampleur sont nécessaires pour modéliser les mécanismes favorisant l'apparition de ces maladies. Elles pourront être menées à bien si les conditions suivantes sont réunies :

- Forte collaboration entre les différents acteurs (chercheurs, vétérinaires et pisciculteurs)
- Capacité d'échantillonnage et de suivi importante et structurée au travers de commémoratifs exhaustifs (méta-données)
- Disponibilité d'outils de caractérisation d'isolats bactériens multi-échelles.
- Possibilité de mettre en place des infections expérimentales pour effectuer des corrélations entre le degré de virulence et le génotype (et/ou MALDI-type) de la souche.

Ainsi, au travers de ce projet, nous proposons que la spectrométrie de masse MALDI-TOF soit l'outil de diagnostic de première ligne pour les piscicultures. Cet outil propose en effet une identification robuste à moindre coût. En outre, comme nous l'avons montré, cette technologie permet, en parallèle, de typer les isolats bactériens ; et cela sans modifier le protocole d'acquisition de routine de spectres MALDI-TOF. C'est une plus-value indéniable face aux autres méthodes disponibles et avec pour objectif le suivi régulier des populations bactériennes locales.

Bien que le typage par spectrométrie de masse MALDI-TOF ne soit pas aussi discriminant que d'autres méthodes (par exemple, la MLST), il reste néanmoins une alternative sérieuse pour les raisons évoquées précédemment. Si la pertinence épidémiologique des « MALDI-types » – ou équivalents – a déjà été évaluée avec succès pour certaines bactéries, il sera nécessaire de faire de même pour les autres bactéries pathogènes appartenant au genre *Tenacibaculum*.

Nous pensons pouvoir répondre à cette question en perfectionnant les outils décrits ici d'une part, et en les employant pour des projets à large échelle d'autre part. Le perfectionnement de nos outils comprend 3 axes majeurs pouvant être résumés ainsi :

- Optimisation des méthodes déjà implémentées
- Diversification des données de référence pour le MALDI-TOF
- Amélioration de l'application MALDIquantTypeR (intégration de nouvelles fonctionnalités, offre d'outils supplémentaires, enrichissement des références disponibles, etc…)

Ces solutions seront utilisées dans un premier temps en tant qu'outils de criblage appliqués à de grandes campagnes d'échantillonnage, menées aussi bien par des collègues à l'étranger (Chili, Norvège) que par notre équipe (en métropole et en Polynésie Française).

V – Annexes

Annexe 1 : Bootstrap classique (BP), bootstrap multi-échelle (AU) et package R *pvclust()*.

An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?

Ryota Suzuki
ryota.suzuki@is.titech.ac.jp

Hidetoshi Shimodaira
shimo@is.titech.ac.jp

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology

How accurate are these clusters?

Cluster analysis is a method to examine the similarities between individuals. Hierarchical clustering generates a dendrogram which contains clusters which show such similarities based on the dissimilarity matrix computed by data. It offers detailed information on the relationships between individuals and hence has been often used in applied areas including microarray analysis. However, it is not clear how strong these clusters are supported by the data. The question is, "How accurate are these clusters?"

Given data and an algorithm to construct a dendrogram, hierarchical clustering generates a dendrogram which contains clusters, as shown in Figure 1. We have microarray data of sample size $n = 10$, where columns correspond to $p = 4$ individuals to be analyzed and rows correspond to $n = 10$ observations. By the resulting dendrogram, we hope to conclude, for example, that A and B are closer than B and C. If we can take samples as many as we want, we can examine this conclusion by applying the same analysis to these samples, as shown in Figure 2.

However, as is always the case, we have only one sample available. Hence we cannot deny the possibility of the situation shown in Figure 3. The figure shows the space of data divided by the resulting dendrograms, where our sample is indicated by the red star. The sample is drawn from the probability distribution centered at the blue star, which is just inside the area where B and C are closer than any other combination of two individuals. In this situation our conclusion is obtained because of random noise and does not correctly reflect the truth.

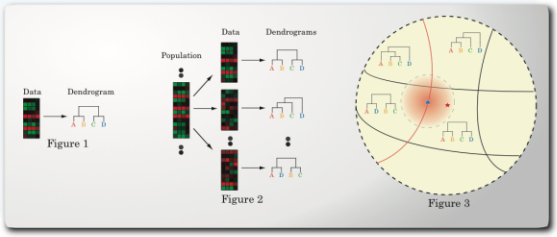


Figure 1 Figure 2 Figure 3

Bootstrap probability

How can we assess such uncertainty using only one sample? One way of achieving this is using bootstrap resampling. In bootstrap resampling we replicate data by resampling from the data itself. Figure 4 shows this situation. We randomly choose $n = 10$ observations from the original data with replication. The procedure is repeated many times, as 10,000 times in the figure. These sampled are called bootstrap samples.

We can examine the result of hierarchical clustering using bootstrap samples, through quantities called "bootstrap probabilities". Figure 5 shows the procedure in detail. First, we generate bootstrap samples. Second, we apply hierarchical clustering to each bootstrap sample. The resulting dendrograms are called bootstrap replications of dendrograms. Third, we count the number of dendrograms which contained a cluster which we hope to examine. In this example, we took the cluster (B, C), which is called a hypothesis. Finally, divide this number by the number of bootstrap samples to obtain the ratio of the dendrograms which fulfill the hypothesis.

The ratio is called bootstrap probability and it can be used to examine the certainty of the hypothesis. In this example, the bootstrap probability of the cluster (B, C) was 4%. It is quite small and it seems that we may be able to conclude that the cluster (B, C) does not exist in the true dendrogram. Indeed this conclusion is correct in an approximate sense. However the approximation is not enough good, so we will adopt a more sophisticated way.

Figure 6 shows what bootstrap probability means. In bootstrap resampling we generate a probability distribution centered at the original data, indicated as the red star. The bootstrap probability we computed is the probability that the bootstrap samples fall inside the area where B and C are most closer. It reflects the uncertainty of cluster analysis in some seanses, but this does not satisfy our objective since it is not exactly the same situation we hope to know, which is shown in Figure 3. The center of the distribution should not be the same point as our sample, but it should be inside the area of our hypothesis.

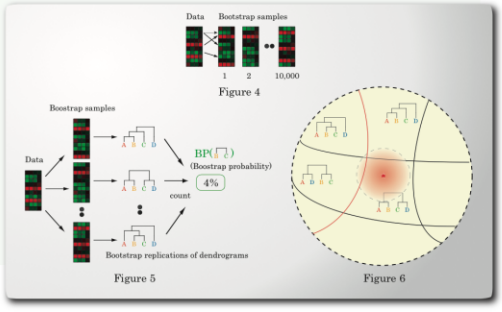


Figure 4 Figure 5 Figure 6

Multiscale bootstrap resampling

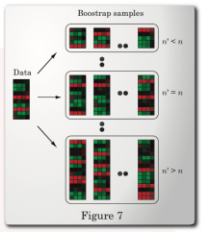


Figure 7

By comparing Figure 3 and 4, we can see that there is a gap between our objective and bootstrap probability. Multiscale bootstrap resampling [2] successfully fills this gap. In bootstrap resampling, the sample size of a bootstrap sample was n, the same as that of original data. On the other hand, we change the sample sizes to several values in multiscale bootstrap resampling. We take sample sizes n', which can be smaller than or larger than, or also can be equal to the original size n. It is shown in Figure 7. In the figure original sample size is $n = 10$, and bootstrap samples with $n' = 5$, 10, 15 are shown.

Using these bootstrap samples, multiscale bootstrap resampling computes a quantity called p-value for each hypothesis. If the p-value of a hypothesis is very small, say smaller than 5%, we can reject the hypothesis. In fact bootstrap probability is an approximation of this value, and multiscale bootstrap resampling corrects the bias of bootstrap probability.

The algorithm of multiscale bootstrap is shown in Figure 8. First, we generate bootstrap samples for each sample size. Second, we apply hierarchical clustering to each bootstrap sample to obtain the sets of bootstrap replications of dendrograms. Third, we compute bootstrap probability for each sample size. Finally, using values of bootstrap probabilities, we can estimate the p-value by fitting a theoretical equation to them. The estimated p-value is called AU (approximately unbiased) value.

Figure 9 shows what multiscale bootstrap resampling computes. AU value is equivalent to the probability that a new sample (if available) appears farther than our sample, under the given hypothesis. In the figure, the red star is our sample and the blue star is the center of the probability distribution, just inside the area of hypothesis. The red dashed curve indicate the points where the distance between data and the hypothesis is the same as our sample. AU value is the probability that a new sample is outside the area indicated by the red dashed curve. If this probability is less than, for example 5%, we can conclude that our sample is not obtained under the hypothesis. In this example AU value is 8%, so we cannot deny the possibility that the data is obtained under the hypothesis that B and C are most closer.

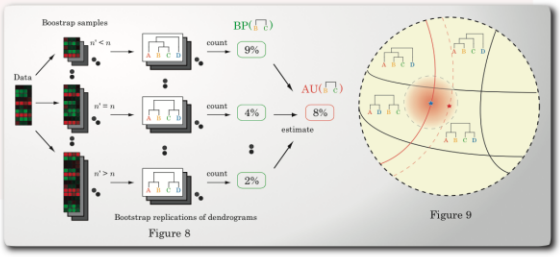


Figure 8 Figure 9

Example

We applied multiscale bootstrap resampling to the hierarchical clustering of microarray data of lung tumores, in Garber et al [1]. The data is expression pattern of $n = 916$ genes of $p = 73$ tumores, and we conducted cluster analysis of tumores. We took $n' = 467$, 542, 636, 757, 916, 1131, 1431, 1869, 2544 and 3664 for multiscale bootstrap resampling, and generated $B = 10{,}000$ replications for each sample size. Figure 10 shows the result of this analysis. The labels from A to E are classification by specialists and it can be seen that cluster analysis returns a similar result. We can also see that some clusters have quite high AU values as larger than 95%. In these cases, we can reject the hypothesis that these cluster do not exist. In other words, we can conclude that these clusters are strongly supported by the data. In this example, the cluster which contains all A's is such a case. Actually the label A means normal cell and cluster analysis seems to detect this strong feature in the data.

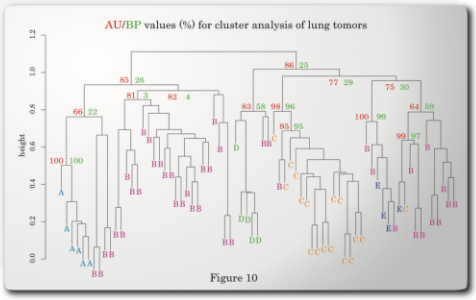


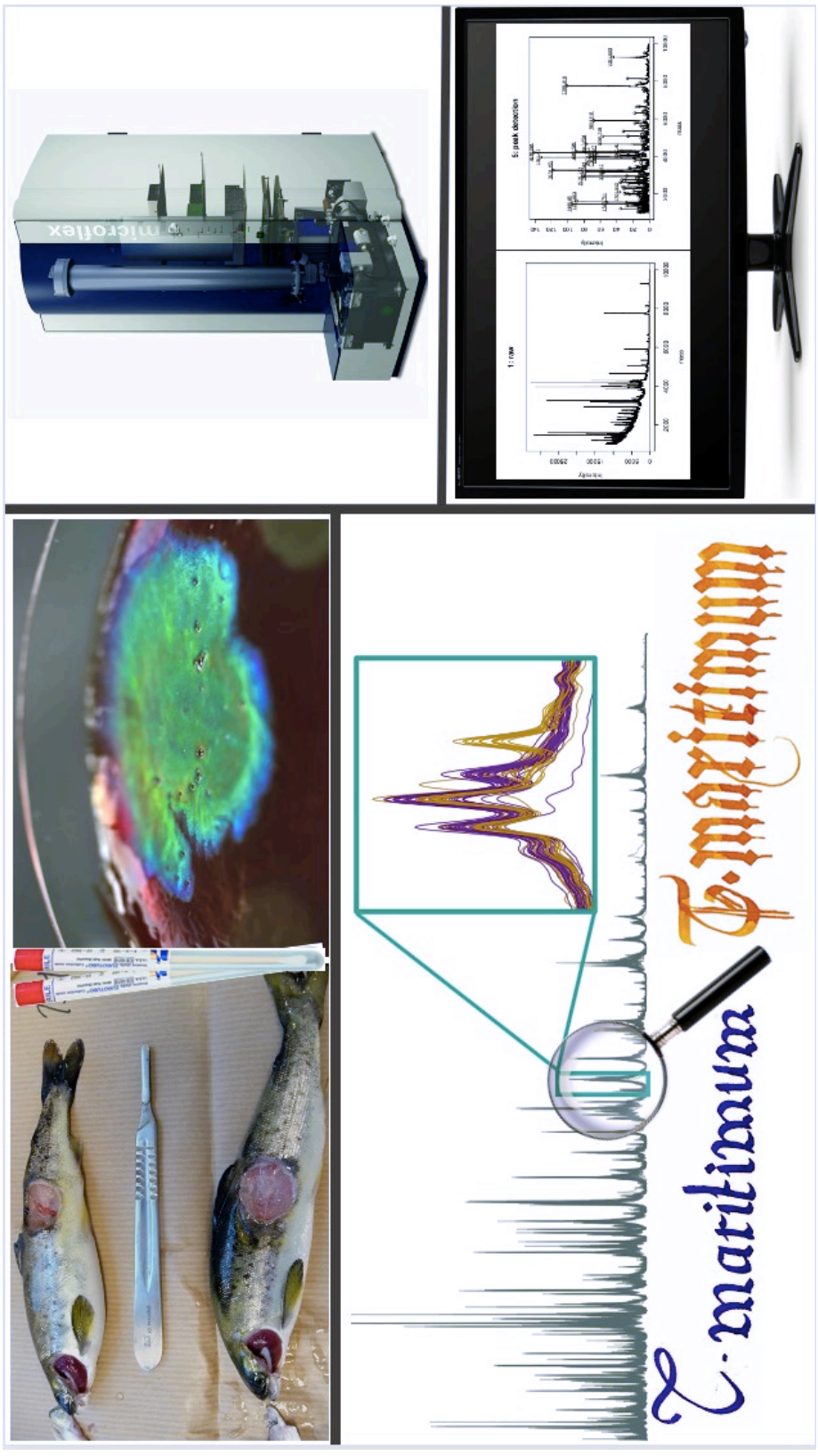
Figure 10

References

[1] Garber, M., et al., Diversity of gene expression in adenocarcinoma of the lung, Proc. Natl. Acad. Sci., USA, 98, 24:13784-13789, 2001.

[2] Shimodaira, H., An approximately unbiased test of phylogenetic tree selection, Systematic Biology, 51: 492-508, 2002.

Annexe 2 : Illustration réalisée comme support du concours Doc'J MT180. Deuxième prix présentation orale.



Annexe 3 : Affiche réalisée pour le congrès SFM 2019 ; Cité des Sciences, Paris

Caractérisation d'agents pathogènes par spectrométrie de masse MALDI-TOF : application à l'espèce *Tenacibaculum maritimum*

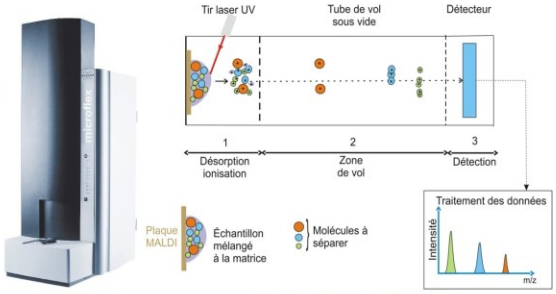
Sébastien Bridel[1,2,3], Frédéric Bourgeon[4], Arnaud Marie[2], Pierre-Yves Moalic[2], Jean-François Bernardet[1], Sophie Pasek[5] et Eric Duchaud[1]

[1] VIM, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France ; [2] Labofarm, Finalab, 22603 Loudéac, France ; [3] Université de Versailles Saint-Quentin-En-Yvelines, 78180 Montigny-Le-Bretonneux, France ; [4] Bio Chêne Vert, Finalab, Rue Blaise Pascal, 35220 Châteaubourg ; [5] Institut de Systématique Evolution, Biodiversité, ISYEB, UMR 7205 CNRS MNHN UPMC EPHE, Paris, France.

Parmi les pathologies retrouvées en pisciculture marine, celles regroupées sous le terme générique de ténacibaculoses (ou flexibactérioses marines) frappent régulièrement de nombreux élevages dans le monde entier (Fig. 1).
Les espèces bactériennes responsables, appartenant au genre *Tenacibaculum*, sont aujourd'hui mal décrites et leur identification est encore difficile. Notre étude avait pour objectif de proposer un outil de diagnostic fiable, rapide et bon marché permettant d'une part d'identifier l'espèce responsable d'un épisode infectieux et d'autre part de caractériser plus finement les isolats au sein de l'espèce la plus prévalente, *T. maritimum*.

Figure 1 : Saumon atlantique (*Salmo salar*) atteint de ténacibaculose
Photo : Carlos Sandoval

MATÉRIELS & MÉTHODES



MALDI-TOF MS Bruker microflex

Figure 2 : Principe du MALDI-TOF D'après MORIDA, académie de Montpellier

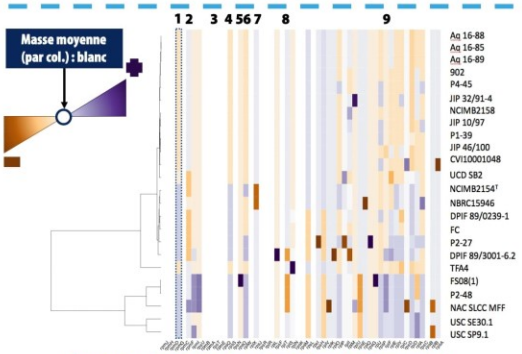


Figure 3 : Analyse *in silico* du polymorphisme des 54 sous-unités protéiques ribosomiques de l'espèce *T. maritimum* (24 génomes)

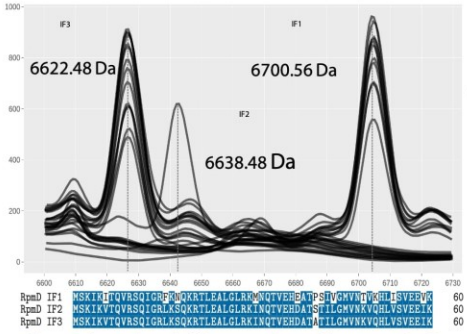


Figure 4 : Aperçu du *peak shift* correspondant au biomarqueur RpmD (50S ribosomal protein subunit L30)

Comment choisir des biomarqueurs pertinents ?

Combiner l'information **génomique** et les données **spectrales**

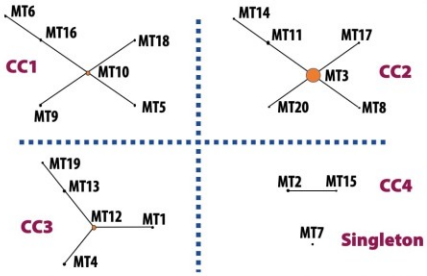


Figure 5 : Relation entre les 130 isolats, les 20 MALDI-types identifiés et les 4 complexes clonaux (réseau eBurst)

RÉSULTATS Nous avons évalué le potentiel de la spectrométrie de masse MALDI-TOF (Fig. 2) pour classer les isolats de *T. maritimum*. À partir de l'analyse croisée des données génomiques et des données spectrales, 9 biomarqueurs polymorphes ont été retenus (Fig. 3). Chaque biomarqueur possède 2 à 5 isoformes (IF) différentes. À chacune de ces IFs correspond une masse spécifique, identifiable dans les empreintes moléculaires (Fig. 4). Un isolat possède une combinaison spécifique de ces 9 IFs. Chaque combinaison spécifique est définie comme un MALDI-type (MT). Sur une centaine d'isolats de terrain, nous avons pu observer 20 MTs différents groupés en 4 complexes clonaux (Fig. 5).

CONCLUSION La spectrométrie de masse MALDI-TOF et le schéma *Multi Peak Shift Typing* (MPST) peuvent constituer une méthode pertinente pour l'analyse des isolats de *T. maritimum* dans les études épidémiologiques à grande échelle. Le MALDI-type est un équivalent des *sequence type* (ST) et des types électrophorétiques (ET) définis respectivement par MLST et MLEE. Cette approche devrait permettre de faciliter la gestion des agents pathogènes dans les élevages. La stratégie développée pourrait également être transposée à d'autres espèces.

Unité Virologie et Immunologie Moléculaires
Équipe Infection et Immunité des Poissons
78350 Jouy-en-Josas, France
sebastien.bridel@inra.fr

IX – Bibliographie

1. FAO. La situation mondiale des pêches et de l'aquaculture 2016: contribuer à la sécurité alimentaire et à la nutrition de tous. Rome (I): FAO; 2016.
2. Bayliss SC, Verner-Jeffreys DW, Bartie KL, Aanensen DM, Sheppard SK, Adams A, et al. The Promise of Whole Genome Pathogen Sequencing for the Molecular Epidemiology of Emerging Aquaculture Pathogens. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.00121.
3. Yin LK. Current Trends in the Study of Bacterial and Viral Fish and Shrimp Diseases | Molecular Aspects of Fish & Marine Biology. National University of Singapore, Singapore: Leung Ka Yin; 2004. https://www.worldscientific.com/worldscibooks/10.1142/5465. Accessed 10 Jul 2019.
4. Crane M, Hyatt A. Viruses of fish: an overview of significant pathogens. Viruses. 2011;3:2025–46.
5. Woo PTK, Leatherland JF. Fish Diseases and Disorders. CABI; 2006.
6. Alvial A, Kibenge F, Forster J, Burgos JM, Ibarra R, St-Hilaire S. The Recovery of the Chilean Salmon Industry. :83.
7. Flegel TW. Historic emergence, impact and current status of shrimp pathogens in Asia. J Invertebr Pathol. 2012;110:166–73.
8. FAO/MARD Technical Workshop on Early Mortality Syndrome (EMS) or Acute Hepatopancreatic Necrosis Syndrome (AHPNS) of Cultured Shrimp, Food and Agriculture Organization of the United Nations. Report of the FAO/MARD Technical Workshop on Early Mortality Syndrome (EMS) or Acute Hepatopancreatic Necrosis Syndrome (AHPNS) of Cultured Shrimp (under TCP/VIE/3304): Hanoi, Viet Nam, 25-27 June 2013. 2013. http://bibpurl.oclc.org/web/48266/018/i3422e/i3422e.pdf
http://www.fao.org/docrep/018/i3422e/i3422e.pdf. Accessed 10 Jul 2019.
9. Lee C-T, Chen I-T, Yang Y-T, Ko T-P, Huang Y-T, Huang J-Y, et al. The opportunistic marine pathogen *Vibrio parahaemolyticus* becomes virulent by acquiring a plasmid that expresses a deadly toxin. Proc Natl Acad Sci USA. 2015;112:10798–803.
10. Toranzo AE, Magariños B, Romalde JL. A review of the main bacterial fish diseases in mariculture systems. Aquaculture. 2005;246:37–61.
11. Avendaño-Herrera R, Toranzo AE, Magariños B. Tenacibaculosis infection in marine fish caused by *Tenacibaculum maritimum*: a review. Dis Aquat Org. 2006;71:255–66.
12. McBride MJ. The Family Flavobacteriaceae. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. The Prokaryotes: Other Major Lineages of Bacteria and The Archaea. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 643–76. doi:10.1007/978-3-642-38954-2_130.
13. Suzuki M, Nakagawa Y, Harayama S, Yamamoto S. Phylogenetic analysis and taxonomic study of marine *Cytophaga*-like bacteria: proposal for *Tenacibaculum* gen. nov. with *Tenacibaculum maritimum* comb. nov. and *Tenacibaculum ovolyticum* comb. nov., and description of *Tenacibaculum mesophilum* sp. nov. and *Tenacibaculum amylolyticum* sp. nov. Int J Syst Evol Microbiol. 2001;51 Pt 5:1639–52.
14. Tenacibaculum. List of Prokaryotic Names with standing in nomenclature. http://www.bacterio.net/tenacibaculum.html. Accessed 10 Jul 2019.
15. Suzuki M. *Tenacibaculum*. In: Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, et al., editors. Bergey's Manual of Systematics of Archaea and Bacteria. Chichester, UK: John Wiley & Sons, Ltd; 2015. p. 1–7. doi:10.1002/9781118960608.gbm00345.
16. Wakabayashi H. *Flexibacter* infection in cultured marine fish in Japan. Helgolander Meeresunters. 1984;37:587–93.
17. Apablaza P, Frisch K, Brevik ØJ, Småge SB, Vallestad C, Duesund H, et al. Primary Isolation and Characterization of *Tenacibaculum maritimum* from Chilean Atlantic Salmon Mortalities Associated with a *Pseudochattonella* spp. Algal Bloom. J Aquat Anim Health.

2017;29:143–9.
18. Ferguson HW, Delannoy CMJ, Hay S, Nicolson J, Sutherland D, Crumlish M. Jellyfish as vectors of bacterial disease for farmed salmon (Salmo salar). J Vet Diagn Invest. 2010;22:376–82.
19. Florio D, Gridelli S, Fioravanti ML, Zanoni RG. First Isolation Of *Tenacibaculum maritimum* in a captive sand tiger shark (*Carcharias*). J Zoo Wildl Med. 2016;47:351–3.
20. Hansen GH, Bergh O, Michaelsen J, Knappskog D. *Flexibacter ovolyticus* sp. nov., a pathogen of eggs and larvae of Atlantic halibut, *Hippoglossus hippoglossus* L. Int J Syst Bacteriol. 1992;42:451–8.
21. Bergh Ø, Hansen GH, Taxt RE. Experimental infection of eggs and yolk sac larvae of halibut, *Hippoglossus hippoglossus* L. Journal of Fish Diseases. 1992;15:379–91.
22. Olsen AB, Gulla S, Steinum T, Colquhoun DJ, Nilsen HK, Duchaud E. Multilocus sequence analysis reveals extensive genetic variety within *Tenacibaculum* spp. associated with ulcers in sea-farmed fish in Norway. Veterinary Microbiology. 2017;205:39–45.
23. Míguez B, Combarro MP. Bacteria associated with sardine (Sardina pilchardus) eggs in a natural environment (Ría de Vigo, Galicia, northwestern Spain). FEMS Microbiol Ecol. 2003;44:329–34.
24. Piñeiro-Vidal M, Carballas CG, Gómez-Barreiro O, Riaza A, Santos Y. *Tenacibaculum soleae* sp. nov., isolated from diseased sole (Solea senegalensis Kaup). Int J Syst Evol Microbiol. 2008;58 Pt 4:881–5.
25. López JR, Piñeiro-Vidal M, García-Lamas N, de la Herran R, Navas JI, Hachero-Cruzado I, et al. First isolation of *Tenacibaculum soleae* from diseased cultured wedge sole, Dicologoglossa cuneata (Moreau), and brill, Scophthalmus rhombus (L.). J Fish Dis. 2010;33:273–8.
26. Burioli E a. V, Varello K, Trancart S, Bozzetta E, Gorla A, Prearo M, et al. First description of a mortality event in adult Pacific oysters in Italy associated with infection by a *Tenacibaculum soleae* strain. J Fish Dis. 2018;41:215–21.
27. Piñeiro-Vidal M, Riaza A, Santos Y. *Tenacibaculum discolor* sp. nov. and *Tenacibaculum gallaicum* sp. nov., isolated from sole (*Solea senegalensis*) and turbot (*Psetta maxima*) culture systems. Int J Syst Evol Microbiol. 2008;58 Pt 1:21–5.
28. Habib C, Houel A, Lunazzi A, Bernardet J-F, Olsen AB, Nilsen H, et al. Multilocus sequence analysis of the marine bacterial genus *Tenacibaculum* suggests parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. Appl Environ Microbiol. 2014;80:5503–14.
29. Allen RC, Tu Y-K, Nevarez MJ, Bobbs AS, Friesen JW, Lorsch JR, et al. The mercury resistance (mer) operon in a marine gliding flavobacterium, *Tenacibaculum discolor* 9A5. FEMS Microbiol Ecol. 2013;83:135–48.
30. Piñeiro-Vidal M, Centeno-Sestelo G, Riaza A, Santos Y. Isolation of pathogenic *Tenacibaculum maritimum*-related organisms from diseased turbot and sole cultured in the Northwest of Spain. 2007;:7.
31. Vidal MP. Descripción de tres nuevas especies del Género *Tenacibaculum* causantes de tenacibaculosis: aspectos taxonómicos y patogenicidad. :239.
32. Piñeiro-Vidal M, Gijón D, Zarza C, Santos Y. *Tenacibaculum* dicentrarchi sp. nov., a marine bacterium of the family *Flavobacteriaceae* isolated from European sea bass. Int J Syst Evol Microbiol. 2012;62 Pt 2:425–9.
33. Avendaño-Herrera R, Irgang R, Sandoval C, Moreno-Lira P, Houel A, Duchaud E, et al. Isolation, Characterization and Virulence Potential of *Tenacibaculum dicentrarchi* in Salmonid Cultures in Chile. Transboundary and Emerging Diseases. 2016;63:121–6.
34. Irgang R, González-Luna R, Gutiérrez J, Poblete-Morales M, Rojas V, Tapia-Cammas D, et al. First identification and characterization of *Tenacibaculum dicentrarchi* isolated from

Chilean red conger eel (Genypterus chilensis, Guichenot 1848). J Fish Dis. 2017;40:1915–20.
35. Levipan HA, Irgang R, Tapia-Cammas D, Avendaño-Herrera R. A high-throughput analysis of biofilm formation by the fish pathogen *Tenacibaculum dicentrarchi*. Journal of Fish Diseases. 2019;42:617–21.
36. Småge SB, Frisch K, Brevik ØJ, Watanabe K, Nylund A. First isolation, identification and characterization of *Tenacibaculum maritimum* in Norway, isolated from diseased farmed sea lice cleaner fish Cyclopterus lumpus L. Aquaculture. 2016;464:178–84.
37. Småge SB, Frisch K, Vold V, Duesund H, Brevik ØJ, Olsen RH, et al. Induction of tenacibaculosis in Atlantic salmon smolts using *Tenacibaculum finnmarkense* and the evaluation of a whole cell inactivated vaccine. Aquaculture. 2018;495:858–64.
38. Småge SB, Brevik ØJ, Frisch K, Watanabe K, Duesund H, Nylund A. Concurrent jellyfish blooms and tenacibaculosis outbreaks in Northern Norwegian Atlantic salmon (*Salmo salar*) farms. PLoS One. 2017;12. doi:10.1371/journal.pone.0187476.
39. Bridel S, Olsen A-B, Nilsen H, Bernardet J-F, Achaz G, Avendaño-Herrera R, et al. Comparative Genomics of *Tenacibaculum dicentrarchi* and "*Tenacibaculum finnmarkense*" Highlights Intricate Evolution of Fish-Pathogenic Species. Genome Biol Evol. 2018;10:452–7.
40. Sheu S-Y, Lin K-Y, Chou J-H, Chang P-S, Arun AB, Young C-C, et al. *Tenacibaculum litopenaei* sp. nov., isolated from a shrimp mariculture pond. Int J Syst Evol Microbiol. 2007;57 Pt 5:1148–53.
41. Tajima K, Hirano T, Nakano K, Ezura Y. Taxonomical Study on the Causative Bacterium of Spotting Disease of Sea Urchin *Strongylocentrotus intermedius*. Fisheries science. 1997;63:897–900.
42. Tajima K, Hirano T, Shimizu M, Ezura Y. Isolation and Pathogenicity of the Causative Bacterium of Spotting Disease of Sea Urchin Strongylocentrotus intermedius. :4.
43. Frette L, Jørgensen NOG, Irming H, Kroer N. *Tenacibaculum skagerrakense* sp. nov., a marine bacterium isolated from the pelagic zone in Skagerrak, Denmark. Int J Syst Evol Microbiol. 2004;54 Pt 2:519–24.
44. Yoon J-H, Kang S-J, Oh T-K. *Tenacibaculum lutimaris* sp. nov., isolated from a tidal flat in the Yellow Sea, Korea. Int J Syst Evol Microbiol. 2005;55 Pt 2:793–8.
45. Jung S-Y, Oh T-K, Yoon J-H. *Tenacibaculum aestuarii* sp. nov., isolated from a tidal flat sediment in Korea. Int J Syst Evol Microbiol. 2006;56 Pt 7:1577–81.
46. Choi DH, Kim Y-G, Hwang CY, Yi H, Chun J, Cho BC. *Tenacibaculum litoreum* sp. nov., isolated from tidal flat sediment. Int J Syst Evol Microbiol. 2006;56 Pt 3:635–40.
47. Wang J-T, Chou Y-J, Chou J-H, Chen CA, Chen W-M. *Tenacibaculum aiptasiae* sp. nov., isolated from a sea anemone *Aiptasia pulchella*. Int J Syst Evol Microbiol. 2008;58 Pt 4:761–6.
48. Heindl H, Wiese J, Imhoff JF. *Tenacibaculum adriaticum* sp. nov., from a bryozoan in the Adriatic Sea. Int J Syst Evol Microbiol. 2008;58 Pt 3:542–7.
49. Lee YS, Baik KS, Park SY, Kim EM, Lee D-H, Kahng H-Y, et al. *Tenacibaculum crassostreae* sp. nov., isolated from the Pacific oyster, Crassostrea gigas. Int J Syst Evol Microbiol. 2009;59 Pt 7:1609–14.
50. Oh Y-S, Kahng H-Y, Lee D-H, Lee SB. *Tenacibaculum jejuense* sp. nov., isolated from coastal seawater. Int J Syst Evol Microbiol. 2012;62 Pt 2:414–9.
51. Kang S-J, Lee S-Y, Lee M-H, Oh T-K, Yoon J-H. *Tenacibaculum geojense* sp. nov., isolated from seawater. Int J Syst Evol Microbiol. 2012;62 Pt 1:18–22.
52. Kim Y-O, Park S, Nam B-H, Jung Y-T, Kim D-G, Jee Y-J, et al. *Tenacibaculum halocynthiae* sp. nov., a member of the family Flavobacteriaceae isolated from sea squirt Halocynthia roretzi. Antonie Van Leeuwenhoek. 2013;103:1321–7.
53. Park S, Yoon J-H. *Tenacibaculum caenipelagi* sp. nov., a member of the family

Flavobacteriaceae isolated from tidal flat sediment. Antonie Van Leeuwenhoek. 2013;104:225–31.
54. Li Y, Wei J, Yang C, Lai Q, Chen Z, Li D, et al. *Tenacibaculum xiamenense* sp. nov., an algicidal bacterium isolated from coastal seawater. Int J Syst Evol Microbiol. 2013;63 Pt 9:3481–6.
55. Wang L, Li X, Hu D, Lai Q, Shao Z. *Tenacibaculum holothuriorum* sp. nov., isolated from the sea cucumber Apostichopus japonicus intestine. Int J Syst Evol Microbiol. 2015;65:4347–52.
56. Kim Y-O, Park I-S, Park S, Nam B-H, Park J-M, Kim D-G, et al. *Tenacibaculum ascidiaceicola* sp. nov., isolated from the golden sea squirt *Halocynthia aurantium*. Int J Syst Evol Microbiol. 2016;66:1174–9.
57. Park S, Ha M-J, Jung Y-T, Kang C-H, Yoon J-H. *Tenacibaculum sediminilitoris* sp. nov., isolated from a tidal flat. Int J Syst Evol Microbiol. 2016;66:2610–6.
58. Kim Y-O, Park I-S, Park S, Nam B-H, Park J-M, Kim D-G, et al. *Tenacibaculum haliotis* sp. nov., isolated from the gut of an abalone *Haliotis discus hannai*. Int J Syst Evol Microbiol. 2017;67:3268–73.
59. Park S, Choi SJ, Won S-M, Yoon J-H. *Tenacibaculum aestuariivivum* sp. nov., isolated from a tidal flat. Int J Syst Evol Microbiol. 2017;67:4612–8.
60. Xu Z-X, Yu P, Mu D-S, Liu Y, Du Z-J. *Tenacibaculum agarivorans* sp. nov., an agar-degrading bacterium isolated from marine alga Porphyra yezoensis Ueda. Int J Syst Evol Microbiol. 2017;67:5139–43.
61. Park S, Choi J, Choi SJ, Yoon J-H. *Tenacibaculum insulae* sp. nov., isolated from a tidal flat. Int J Syst Evol Microbiol. 2018;68:228–33.
62. Shin S-K, Kim E, Yi H. *Tenacibaculum todarodis* sp. nov., isolated from a squid. Int J Syst Evol Microbiol. 2018;68:1479–83.
63. Dubnau D, Smith I, Morell P, Marmur J. Gene conservation in *Bacillus* species. I. Conserved genetic and nucleic acid base sequence homologies. Proc Natl Acad Sci U S A. 1965;54:491–8.
64. Woese CR. Bacterial evolution. Microbiol Rev. 1987;51:221–71.
65. Chen K, Neimark H, Rumore P, Steinman CR. Broad range DNA probes for detecting and amplifying eubacterial nucleic acids. FEMS Microbiol Lett. 1989;48:19–24.
66. Relman DA. The search for unrecognized pathogens. Science. 1999;284:1308–10.
67. Woese CR, Stackebrandt E, Macke TJ, Fox GE. A Phylogenetic Definition of the Major Eubacterial Taxa. Systematic and Applied Microbiology. 1985;6:143–51.
68. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;64 Pt 2:346–51.
69. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews. 2004;17:840–62.
70. Rosselló-Mora R. DNA-DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation. In: Stackebrandt E, editor. Molecular Identification, Systematics, and Population Structure of Prokaryotes. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 23–50. doi:10.1007/978-3-540-31292-5_2.
71. Sachse K, Hotzel H. Classification of isolates by DNA-DNA hybridization. Methods Mol Biol. 1998;104:189–95.
72. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA. 2009;106:19126–31.
73. Glaser P. Les puces à ADN vont-elles révolutionner l'identification des bactéries ? Med Sci (Paris). 2005;21:539–44.

74. Galeries d'identification API. bioMérieux France. https://www.biomerieux.fr/diagnostic-clinique/galeries-didentification-api. Accessed 17 Jul 2019.
75. Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. *Escherichia coli* ST131, an Intriguing Clonal Group. Clinical Microbiology Reviews. 2014;27:543–74.
76. Tacconelli E, Cataldo MA, Dancer SJ, De Angelis G, Falcone M, Frank U, et al. ESCMID guidelines for the management of the infection control measures to reduce transmission of multidrug-resistant Gram-negative bacteria in hospitalized patients. Clinical Microbiology and Infection. 2014;20:1–55.
77. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13 Suppl 3:1–46.
78. MacCannell D. Bacterial strain typing. Clin Lab Med. 2013;33:629–50.
79. Ørskov F, Ørskov I. 2 *Serotyping of Escherichia coli***. In: Bergan T, editor. Methods in Microbiology. Academic Press; 1984. p. 43–112. doi:10.1016/S0580-9517(08)70447-1.
80. Tenover FC, Arbeit RD, Goering RV, America MTWG of the S for HE of. How to Select and Interpret Molecular Strain Typing Methods for Epidemiological Studies of Bacterial Infections A Review for Healthcare Epidemiologists. Infection Control & Hospital Epidemiology. 1997;18:426–39.
81. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53:2410–26.
82. Barenfanger J, Drake C, Kacich G. Clinical and Financial Benefits of Rapid Bacterial Identification and Antimicrobial Susceptibility Testing. J Clin Microbiol. 1999;37:1415–8.
83. Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clin Infect Dis. 2009;49:1749–55.
84. Wiegand I, Hilpert K, Hancock REW. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. Nat Protoc. 2008;3:163–75.
85. Syal K, Mo M, Yu H, Iriya R, Jing W, Guodong S, et al. Current and emerging techniques for antibiotic susceptibility tests. Theranostics. 2017;7:1795–805.
86. Meyers JA, Sanchez D, Elwell LP, Falkow S. Simple agarose gel electrophoretic method for the identification and characterization of plasmid deoxyribonucleic acid. J Bacteriol. 1976;127:1529–37.
87. Goering RV. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infect Genet Evol. 2010;10:866–75.
88. Kostman JR, Edlind TD, LiPuma JJ, Stull TL. Molecular epidemiology of Pseudomonas cepacia determined by polymerase chain reaction ribotyping. J Clin Microbiol. 1992;30:2084–7.
89. O'Neill GL, Ogunsola FT, Brazier JS, Duerden BI. Modification of a PCR Ribotyping Method for Application as a Routine Typing Scheme forClostridium difficile. Anaerobe. 1996;2:205–9.
90. Cartwright CP, Stock F, Beekmann SE, Williams EC, Gill VJ. PCR amplification of rRNA intergenic spacer regions as a method for epidemiologic typing of Clostridium difficile. J Clin Microbiol. 1995;33:184–7.
91. Bidet P, Barbut F, Lalande V, Burghoffer B, Petit JC. Development of a new PCR-ribotyping method for Clostridium difficile based on ribosomal RNA gene sequencing. FEMS Microbiol Lett. 1999;175:261–6.
92. DebRoy C, Fratamico PM, Yan X, Baranzoni G, Liu Y, Needleman DS, et al. Comparison of O-Antigen Gene Clusters of All O-Serogroups of Escherichia coli and Proposal for Adopting a New Nomenclature for O-Typing. PLoS ONE. 2016;11:e0147434.

93. Stubbs SL, Brazier JS, O'Neill GL, Duerden BI. PCR targeted to the 16S-23S rRNA gene intergenic spacer region of *Clostridium difficile* and construction of a library consisting of 116 different PCR ribotypes. J Clin Microbiol. 1999;37:461–3.
94. Schumann P, Pukall R. The discriminatory power of ribotyping as automatable technique for differentiation of bacteria. Syst Appl Microbiol. 2013;36:369–75.
95. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998;95:3140–5.
96. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology (Reading, Engl). 1998;144 (Pt 11):3049–60.
97. Heym B, Le Moal M, Armand-Lefevre L, Nicolas-Chanoine M-H. Multilocus sequence typing (MLST) shows that the "Iberian" clone of methicillin-resistant *Staphylococcus aureus* has spread to France and acquired reduced susceptibility to teicoplanin. J Antimicrob Chemother. 2002;50:323–9.
98. Kriz P, Kalmusova J, Felsberg J. Multilocus sequence typing of *Neisseria meningitidis* directly from cerebrospinal fluid. Epidemiol Infect. 2002;128:157–60.
99. Nicolas P, Mondot S, Achaz G, Bouchenot C, Bernardet J-F, Duchaud E. Population Structure of the Fish-Pathogenic Bacterium Flavobacterium psychrophilum. Appl Environ Microbiol. 2008;74:3702–9.
100. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing Escherichia coli O157. Clin Infect Dis. 2015;61:305–12.
101. Bakker HC den, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z, et al. Rapid Whole-Genome Sequencing for Surveillance of *Salmonella* enterica Serovar Enteritidis - Volume 20, Number 8—August 2014 - Emerging Infectious Diseases journal - CDC. doi:10.3201/eid2008.131399.
102. Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR, Cookson BT. Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. Journal of Clinical Microbiology. 2015;53:1072–9.
103. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. J Clin Microbiol. 2016;54:333–42.
104. Kuroda M, Serizawa M, Okutani A, Sekizuka T, Banno S, Inoue S. Genome-Wide Single Nucleotide Polymorphism Typing Method for Identification of Bacillus anthracis Species and Strains among B. cereus Group Species. Journal of Clinical Microbiology. 2010;48:2821–9.
105. Chen C, Zhang W, Zheng H, Lan R, Wang H, Du P, et al. Minimum Core Genome Sequence Typing of Bacterial Pathogens: a Unified Approach for Clinical and Public Health Microbiology. Journal of Clinical Microbiology. 2013;51:2582–91.
106. Huijsmans CJJ, Schellekens JJA, Wever PC, Toman R, Savelkoul PHM, Janse I, et al. Single-Nucleotide-Polymorphism Genotyping of Coxiella burnetii during a Q Fever Outbreak in The Netherlands. Appl Environ Microbiol. 2011;77:2051–7.
107. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, et al. Population Genetics of Vibrio cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. mBio. 2011;2:e00157-11.
108. Sahl JW, Lemmer D, Travis J, Schupp JM, Gillece JD, Aziz M, et al. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microb Genom. 2016;2. doi:10.1099/mgen.0.000074.
109. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal

multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology. 2012;158:1005–15.
110. Been M de, Pinholt M, Top J, Bletz S, Mellmann A, Schaik W van, et al. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of *Enterococcus faecium*. Journal of Clinical Microbiology. 2015;53:3788–97.
111. Boers SA, Reijden WA van der, Jansen R. High-Throughput Multilocus Sequence Typing: Bringing Molecular Typing to the Next Level. PLOS ONE. 2012;7:e39630.
112. Chen Y, Frazzitta AE, Litvintseva AP, Fang C, Mitchell TG, Springer DJ, et al. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. Fungal Genetics and Biology. 2015;75:64–71.
113. Pérez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. Infect Genet Evol. 2018;63:346–59.
114. Kwong JC, McCallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. Pathology. 2015;47:199–210.
115. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17:132.
116. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep. 2016;6:27930.
117. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res. 2018;3:33.
118. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity. 2011;106:383–90.
119. Balloux F, Brønstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends Microbiol. 2018;26:1035–48.
120. Everley RA, Mott TM, Wyatt SA, Toney DM, Croley TR. Liquid chromatography/mass spectrometry characterization of *Escherichia coli* and *Shigella* species. J Am Soc Mass Spectrom. 2008;19:1621–8.
121. Ekström S, Onnerfjord P, Nilsson J, Bengtsson M, Laurell T, Marko-Varga G. Integrated microanalytical technology enabling rapid and automated protein identification. Anal Chem. 2000;72:286–93.
122. Singhal N, Kumar M, Kanaujia PK, Virdi JS. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. Front Microbiol. 2015;6. doi:10.3389/fmicb.2015.00791.
123. Wolk DM, Clark AE. Matrix-Assisted Laser Desorption Time of Flight Mass Spectrometry. Clin Lab Med. 2018;38:471–86.
124. Yates JR. Mass spectrometry and the age of the proteome. J Mass Spectrom. 1998;33:1–19.
125. Westblade LF, Garner OB, MacDonald K, Bradford C, Pincus DH, Mochon AB, et al. Assessment of Reproducibility of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Bacterial and Yeast Identification. J Clin Microbiol. 2015;53:2349–52.
126. Cherkaoui A, Hibbs J, Emonet S, Tangomo M, Girard M, Francois P, et al. Comparison of Two Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Methods with Conventional Phenotypic Identification for Routine Identification of Bacteria to the Species Level. Journal of Clinical Microbiology. 2010;48:1169–75.
127. Dupont C, Sivadon-Tardy V, Bille E, Dauphin B, Beretti JL, Alvarez AS, et al.

Identification of clinical coagulase-negative staphylococci, isolated in microbiology laboratories, by matrix-assisted laser desorption/ionization-time of flight mass spectrometry and two automated systems. Clinical Microbiology and Infection. 2010;16:998–1004.
128. Spanu T, De Carolis E, Fiori B, Sanguinetti M, D'Inzeo T, Fadda G, et al. Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to rpoB gene sequencing for species identification of bloodstream infection staphylococcal isolates. Clin Microbiol Infect. 2011;17:44–9.
129. Carbonnelle E, Grohs P, Jacquier H, Day N, Tenza S, Dewailly A, et al. Robustness of two MALDI-TOF mass spectrometry systems for bacterial identification. Journal of Microbiological Methods. 2012;89:133–6.
130. Martiny D, Dediste A, Debruyne L, Vlaes L, Haddou NB, Vandamme P, et al. Accuracy of the API Campy system, the Vitek 2 Neisseria–Haemophilus card and matrix-assisted laser desorption ionization time-of-flight mass spectrometry for the identification of Campylobacter and related organisms. Clinical Microbiology and Infection. 2011;17:1001–6.
131. Multicenter Evaluation of the Bruker MALDI Biotyper CA System for the Identification of Clinical Aerobic Gram-Negative Bacterial Isolates. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141350. Accessed 5 Aug 2019.
132. Garner O, Mochon A, Branda J, Burnham C-A, Bythrow M, Ferraro M, et al. Multi-centre evaluation of mass spectrometric identification of anaerobic bacteria using the VITEK® MS system. Clinical Microbiology and Infection. 2014;20:335–9.
133. Rychert J, Burnham C-AD, Bythrow M, Garner OB, Ginocchio CC, Jennemann R, et al. Multicenter Evaluation of the Vitek MS Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry System for Identification of Gram-Positive Aerobic Bacteria. Journal of Clinical Microbiology. 2013;51:2225–31.
134. Altun O, Botero-Kleiven S, Carlsson S, Ullberg M, Özenci V. Rapid identification of bacteria from positive blood culture bottles by MALDI-TOF MS following short-term incubation on solid media. Journal of Medical Microbiology. 2015;64:1346–52.
135. Ferreira L, Sánchez-Juanes F, González-Ávila M, Cembrero-Fuciños D, Herrero-Hernández A, González-Buitrago JM, et al. Direct Identification of Urinary Tract Pathogens from Urine Samples by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. Journal of Clinical Microbiology. 2010;48:2110–5.
136. Bar-On O, Mussaffi H, Mei-Zahav M, Prais D, Steuer G, Stafler P, et al. Increasing nontuberculous mycobacteria infection in cystic fibrosis. Journal of Cystic Fibrosis. 2015;14:53–62.
137. Donohue MJ. Increasing nontuberculous mycobacteria reporting rates and species diversity identified in clinical laboratory reports. BMC Infectious Diseases. 2018;18:163.
138. Şamlı A, İlki A. Comparison of MALDI-TOF MS, nucleic acid hybridization and the MPT64 immunochromatographic test for the identification of M. tuberculosis and non-tuberculosis Mycobacterium species. New Microbiol. 2016;39:259–63.
139. Vega-Rúa A, Pagès N, Fontaine A, Nuccio C, Hery L, Goindin D, et al. Improvement of mosquito identification by MALDI-TOF MS biotyping using protein signatures from two body parts. Parasit Vectors. 2018;11. doi:10.1186/s13071-018-3157-1.
140. Hou TY, Chiang-Ni C, Teng SH. Current status of MALDI-TOF mass spectrometry in clinical microbiology. J Food Drug Anal. 2019;27:404–14.
141. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl J m, Laurent F, et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013;18:20380.
142. Siegrist TJ, Anderson PD, Huen WH, Kleinheinz GT, McDermott CM, Sandrin TR. Discrimination and characterization of environmental strains of *Escherichia coli* by matrix-

assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS). Journal of Microbiological Methods. 2007;68:554–62.
143. Egli A, Tschudin-Sutter S, Oberle M, Goldenberger D, Frei R, Widmer AF. Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass-Spectrometry (MALDI-TOF MS) Based Typing of Extended-Spectrum β-Lactamase Producing E. coli – A Novel Tool for Real-Time Outbreak Investigation. PLOS ONE. 2015;10:e0120624.
144. Karger A, Ziller M, Bettin B, Mintel B, Schares S, Geue L. Determination of Serotypes of Shiga Toxin-Producing Escherichia coli Isolates by Intact Cell Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. Appl Environ Microbiol. 2011;77:896–905.
145. Christner M, Trusch M, Rohde H, Kwiatkowski M, Schlüter H, Wolters M, et al. Rapid MALDI-TOF Mass Spectrometry Strain Typing during a Large Outbreak of Shiga-Toxigenic Escherichia coli. PLOS ONE. 2014;9:e101924.
146. Chui H, Chan M, Hernandez D, Chong P, McCorrister S, Robinson A, et al. Rapid, Sensitive, and Specific Escherichia coli H Antigen Typing by Matrix-Assisted Laser Desorption Ionization–Time of Flight-Based Peptide Mass Fingerprinting. Journal of Clinical Microbiology. 2015;53:2480–5.
147. Mazzeo MF, Sorrentino A, Gaita M, Cacace G, Stasio MD, Facchiano A, et al. Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for the Discrimination of Food-Borne Microorganisms. Appl Environ Microbiol. 2006;72:1180–9.
148. Fagerquist CK, Garbus BR, Miller WG, Williams KE, Yee E, Bates AH, et al. Rapid Identification of Protein Biomarkers of Escherichia coli O157:H7 by Matrix-Assisted Laser Desorption Ionization-Time-of-Flight−Time-of-Flight Mass Spectrometry and Top-Down Proteomics. Anal Chem. 2010;82:2717–25.
149. Clark CG, Kruczkiewicz P, Guan C, McCorrister SJ, Chong P, Wylie J, et al. Evaluation of MALDI-TOF mass spectroscopy methods for determination of *Escherichia coli* pathotypes. Journal of Microbiological Methods. 2013;94:180–91.
150. Sauget M, Nicolas-Chanoine M-H, Cabrolier N, Bertrand X, Hocquet D. Matrix-assisted laser desorption ionization-time of flight mass spectrometry assigns *Escherichia coli* to the phylogroups A, B1, B2 and D. International Journal of Medical Microbiology. 2014;304:977–83.
151. Veenemans J, Welker M, van Belkum A, Saccomani MC, Girard V, Pettersson A, et al. Comparison of MALDI-TOF MS and AFLP for strain typing of ESBL-producing *Escherichia coli*. Eur J Clin Microbiol Infect Dis. 2016;35:829–38.
152. Novais Â, Sousa C, de Dios Caballero J, Fernandez-Olmos A, Lopes J, Ramos H, et al. MALDI-TOF mass spectrometry as a tool for the discrimination of high-risk *Escherichia coli* clones from phylogenetic groups B2 (ST131) and D (ST69, ST405, ST393). Eur J Clin Microbiol Infect Dis. 2014;33:1391–9.
153. Matsumura Y, Yamamoto M, Nagao M, Tanaka M, Machida K, Ito Y, et al. Detection of Extended-Spectrum-β-Lactamase-Producing *Escherichia coli* ST131 and ST405 Clonal Groups by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. Journal of Clinical Microbiology. 2014;52:1034–40.
154. Matsumura Y, Yamamoto M, Nagao M, Tanaka M, Takakura S, Ichiyama S. Detection of *Escherichia coli* sequence type 131 clonal group among extended-spectrum β-lactamase-producing E. coli using VITEK MS Plus matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of Microbiological Methods. 2015;119:7–9.
155. Nakamura A, Komatsu M, Kondo A, Ohno Y, Kohno H, Nakamura F, et al. Rapid detection of B2-ST131 clonal group of extended-spectrum β-lactamase–producing *Escherichia coli* by matrix-assisted laser desorption ionization–time-of-flight mass spectrometry: discovery of a peculiar amino acid substitution in B2-ST131 clonal group.

Diagnostic Microbiology and Infectious Disease. 2015;83:237–44.
156. Lafolie J, Sauget M, Cabrolier N, Hocquet D, Bertrand X. Detection of *Escherichia coli* sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies? Journal of Hospital Infection. 2015;90:208–12.
157. Schlebusch S, Price GR, Hinds S, Nourse C, Schooneveldt JM, Tilse MH, et al. First outbreak of PVL-positive nonmultiresistant MRSA in a neonatal ICU in Australia: comparison of MALDI-TOF and SNP-plus-binary gene typing. Eur J Clin Microbiol Infect Dis. 2010;29:1311–4.
158. Ueda O, Tanaka S, Nagasawa Z, Hanaki H, Shobuike T, Miyamoto H. Development of a novel matrix-assisted laser desorption/ionization time-of-flight mass spectrum (MALDI-TOF-MS)-based typing method to identify meticillin-resistant *Staphylococcus aureus* clones. Journal of Hospital Infection. 2015;90:147–55.
159. Lasch P, Fleige C, Stämmler M, Layer F, Nübel U, Witte W, et al. Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus faecium* and Staphylococcus aureus isolates. Journal of Microbiological Methods. 2014;100:58–69.
160. Wolters M, Rohde H, Maier T, Belmar-Campos C, Franke G, Scherpe S, et al. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages. International Journal of Medical Microbiology. 2011;301:64–8.
161. Boggs SR, Cazares LH, Drake R. Characterization of a *Staphylococcus aureus* USA300 protein signature using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Journal of Medical Microbiology. 2012;61:640–4.
162. Josten M, Reif M, Szekat C, Al-Sabti N, Roemer T, Sparbier K, et al. Analysis of the Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrum of *Staphylococcus* aureus Identifies Mutations That Allow Differentiation of the Main Clonal Lineages. Journal of Clinical Microbiology. 2013;51:1809–17.
163. Zhang T, Ding J, Rao X, Yu J, Chu M, Ren W, et al. Analysis of methicillin-resistant Staphylococcus aureus major clonal lineages by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI–TOF MS). Journal of Microbiological Methods. 2015;117:122–7.
164. Østergaard C, Hansen SGK, Møller JK. Rapid first-line discrimination of methicillin resistant Staphylococcus aureus strains using MALDI-TOF MS. International Journal of Medical Microbiology. 2015;305:838–47.
165. Camoez M, Sierra JM, Dominguez MA, Ferrer-Navarro M, Vila J, Roca I. Automated categorization of methicillin-resistant *Staphylococcus aureus* clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. Clinical Microbiology and Infection. 2016;22:161.e1-161.e7.
166. Sauget M, van der Mee-Marquet N, Bertrand X, Hocquet D. Matrix-assisted laser desorption ionization-time of flight Mass spectrometry can detect *Staphylococcus aureus* clonal complex 398. Journal of Microbiological Methods. 2016;127:20–3.
167. Cabrolier N, Sauget M, Bertrand X, Hocquet D. Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry Identifies *Pseudomonas aeruginosa* High-Risk Clones. Journal of Clinical Microbiology. 2015;53:1395–8.
168. Sousa C, Botelho J, Grosso F, Silva L, Lopes J, Peixe L. Unsuitability of MALDI-TOF MS to discriminate *Acinetobacter baumannii* clones under routine experimental conditions. Front Microbiol. 2015;6. doi:10.3389/fmicb.2015.00481.
169. Sachse S, Bresan S, Erhard M, Edel B, Pfister W, Saupe A, et al. Comparison of multilocus sequence typing, RAPD, and MALDI-TOF mass spectrometry for typing of β-lactam–resistant *Klebsiella pneumoniae* strains. Diagnostic Microbiology and Infectious

Disease. 2014;80:267–71.
170. Berrazeg M, Diene SM, Drissi M, Kempf M, Richet H, Landraud L, et al. Biotyping of Multidrug-Resistant *Klebsiella pneumoniae* Clinical Isolates from France and Algeria Using MALDI-TOF MS. PLOS ONE. 2013;8:e61428.
171. Rodrigues C, Passet V, Rakotondrasoa A, Brisse S. Identification of *Klebsiella pneumoniae*, *Klebsiella quasipneumoniae*, *Klebsiella variicola* and Related Phylogroups by MALDI-TOF Mass Spectrometry. Frontiers in Microbiology. 2018;9. doi:10.3389/fmicb.2018.03000.
172. Suarez S, Ferroni A, Lotz A, Jolley KA, Guérin P, Leto J, et al. Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. Journal of Microbiological Methods. 2013;94:390–6.
173. Barbuddhe SB, Maier T, Schwarz G, Kostrzewa M, Hof H, Domann E, et al. Rapid Identification and Typing of *Listeria* Species by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. Appl Environ Microbiol. 2008;74:5402–7.
174. Månsson V, Resman F, Kostrzewa M, Nilson B, Riesbeck K. Identification of *Haemophilus* influenzae Type b Isolates by Use of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. Journal of Clinical Microbiology. 2015;53:2215–24.
175. Williamson YM, Moura H, Woolfitt AR, Pirkle JL, Barr JR, Carvalho MDG, et al. Differentiation of *Streptococcus pneumoniae* Conjunctivitis Outbreak Isolates by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. Appl Environ Microbiol. 2008;74:5891–7.
176. Dunne EM, Ong EK, Moser RJ, Siba PM, Phuanukoonnon S, Greenhill AR, et al. Multilocus Sequence Typing of *Streptococcus pneumoniae* by Use of Mass Spectrometry. Journal of Clinical Microbiology. 2011;49:3756–60.
177. Nakano S, Matsumura Y, Ito Y, Fujisawa T, Chang B, Suga S, et al. Development and evaluation of MALDI-TOF MS-based serotyping for *Streptococcus pneumoniae*. Eur J Clin Microbiol Infect Dis. 2015;34:2191–8.
178. Moura H, Woolfitt AR, Carvalho MG, Pavlopoulos A, Teixeira LM, Satten GA, et al. MALDI-TOF mass spectrometry as a tool for differentiation of invasive and noninvasive *Streptococcus pyogenes* isolates. FEMS Immunol Med Microbiol. 2008;53:333–42.
179. Wang J, Zhou N, Xu B, Hao H, Kang L, Zheng Y, et al. Identification and Cluster Analysis of *Streptococcus pyogenes* by MALDI-TOF Mass Spectrometry. PLOS ONE. 2012;7:e47152.
180. Spinali S, van Belkum A, Goering RV, Girard V, Welker M, Van Nuenen M, et al. Microbial typing by matrix-assisted laser desorption ionization-time of flight mass spectrometry: do we need guidance for data interpretation? J Clin Microbiol. 2015;53:760–5.
181. Sauget M, Valot B, Bertrand X, Hocquet D. Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? Trends Microbiol. 2017;25:447–55.
182. Giacometti F, Piva S, Vranckx K, De Bruyne K, Drigo I, Lucchi A, et al. Application of MALDI-TOF MS for the subtyping of *Arcobacter butzleri* strains and comparison with their MLST and PFGE types. Int J Food Microbiol. 2018;277:50–7.
183. Rödel J, Mellmann A, Stein C, Alexi M, Kipp F, Edel B, et al. Use of MALDI-TOF mass spectrometry to detect nosocomial outbreaks of *Serratia marcescens* and *Citrobacter freundii*. Eur J Clin Microbiol Infect Dis. 2019;38:581–91.
184. Li R, Xiao D, Yang J, Sun S, Kaplan S, Li Z, et al. Identification and Characterization of *Clostridium difficile* Sequence Type 37 Genotype by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 2018;56.
185. Cheng J-W, Liu C, Kudinha T, Xiao M, Yu S-Y, Yang C-X, et al. Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify MLST clade 4

Clostridium difficile isolates. Diagn Microbiol Infect Dis. 2018;92:19–24.
186. Josten M, Dischinger J, Szekat C, Reif M, Al-Sabti N, Sahl H-G, et al. Identification of agr-positive methicillin-resistant *Staphylococcus aureus* harbouring the class A mec complex by MALDI-TOF mass spectrometry. Int J Med Microbiol. 2014;304:1018–23.
187. Schuster D, Josten M, Janssen K, Bodenstein I, Albert C, Schallenberg A, et al. Detection of methicillin-resistant coagulase-negative staphylococci harboring the class A mec complex by MALDI-TOF mass spectrometry. Int J Med Microbiol. 2018;308:522–6.
188. Lindgren Å, Karami N, Karlsson R, Åhrén C, Welker M, Moore ERB, et al. Development of a rapid MALDI-TOF MS based epidemiological screening method using MRSA as a model organism. Eur J Clin Microbiol Infect Dis. 2018;37:57–68.
189. Holzknecht BJ, Dargis R, Pedersen M, Pinholt M, Christensen JJ, Danish Enterococcal Study Group. Typing of vancomycin-resistant enterococci with MALDI-TOF mass spectrometry in a nosocomial outbreak setting. Clin Microbiol Infect. 2018;24:1104.e1-1104.e4.
190. Kang L, Li N, Li P, Zhou Y, Gao S, Gao H, et al. MALDI-TOF mass spectrometry provides high accuracy in identification of *Salmonella* at species level but is limited to type or subtype Salmonella serovars. Eur J Mass Spectrom (Chichester). 2017;23:70–82.
191. Gallegos-Candela M, Boyer AE, Woolfitt AR, Brumlow J, Lins RC, Quinn CP, et al. Validated MALDI-TOF-MS method for anthrax lethal factor provides early diagnosis and evaluation of therapeutics. Anal Biochem. 2018;543:97–107.
192. Gagnaire J, Dauwalder O, Boisset S, Khau D, Freydière A-M, Ader F, et al. Detection of Staphylococcus aureus Delta-Toxin Production by Whole-Cell MALDI-TOF Mass Spectrometry. PLOS ONE. 2012;7:e40660.
193. Jang KS, Park M, Lee JY, Kim JS. Mass spectrometric identification of phenol-soluble modulins in the ATCC® 43300 standard strain of methicillin-resistant Staphylococcus aureus harboring two distinct phenotypes. Eur J Clin Microbiol Infect Dis. 2017;36:1151–7.
194. Welker M, Van Belkum A, Girard V, Charrier J-P, Pincus D. An update on the routine application of MALDI-TOF MS in clinical microbiology. Expert Rev Proteomics. 2019;:1–16.
195. Jorgensen JH, Turnidge JD. Susceptibility Test Methods: Dilution and Disk Diffusion Methods*. Manual of Clinical Microbiology, Eleventh Edition. 2015;:1253–73.
196. Abbott AN, Fang FC. Molecular Detection of Antibacterial Drug Resistance. Manual of Clinical Microbiology, Eleventh Edition. 2015;:1379–89.
197. Kostrzewa M, Sparbier K, Maier T, Schubert S. MALDI-TOF MS: an upcoming tool for rapid detection of antibiotic resistance in microorganisms. Proteomics Clin Appl. 2013;7:767–78.
198. Sparbier K, Schubert S, Kostrzewa M. MBT-ASTRA: A suitable tool for fast antibiotic susceptibility testing? Methods. 2016;104:48–54.
199. Hrabák J, Walková R, Studentová V, Chudácková E, Bergerová T. Carbapenemase activity detection by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol. 2011;49:3222–7.
200. Burckhardt I, Zimmermann S. Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry To Detect Carbapenem Resistance within 1 to 2.5 Hours ▿ . J Clin Microbiol. 2011;49:3321–4.
201. Sparbier K, Schubert S, Weller U, Boogen C, Kostrzewa M. Matrix-assisted laser desorption ionization-time of flight mass spectrometry-based functional assay for rapid detection of resistance against β-lactam antibiotics. J Clin Microbiol. 2012;50:927–37.
202. Sparbier K, Lange C, Jung J, Wieser A, Schubert S, Kostrzewa M. MALDI Biotyper-Based Rapid Resistance Detection by Stable-Isotope Labeling. Journal of Clinical Microbiology. 2013;51:3741–8.

203. Lange C, Schubert S, Jung J, Kostrzewa M, Sparbier K. Quantitative Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Rapid Resistance Detection. Journal of Clinical Microbiology. 2014;52:4155–62.
204. Vrioni G, Tsiamis C, Oikonomidis G, Theodoridou K, Kapsimali V, Tsakris A. MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance: current achievements and future perspectives. Ann Transl Med. 2018;6. doi:10.21037/atm.2018.06.28.
205. How to choose your MALDI Matrix. Bitesize Bio. 2016. https://bitesizebio.com/27800/how-to-choose-your-maldi-soul-matrix/. Accessed 4 Nov 2019.
206. Ryzhov V, Fenselau C. Characterization of the protein subset desorbed by MALDI from whole bacterial cells. Anal Chem. 2001;73:746–50.
207. Sub-speciating Campylobacter jejuni by proteomic analysis of its protein biomarkers and their post-translational modifications. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/17022624. Accessed 4 Nov 2019.
208. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res. 2019. doi:10.1093/nar/gkz926.
209. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
210. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics. 2015;31:3997–9.
211. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490.
212. Pérez-Pascual D, Lunazzi A, Magdelenat G, Rouy Z, Roulet A, Lopez-Roques C, et al. The Complete Genome Sequence of the Fish Pathogen Tenacibaculum maritimum Provides Insights into Virulence Mechanisms. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.01542.
213. Pritchard L. Python module for average nucleotide identity analyses: widdowquinn/pyani. Python. 2019. https://github.com/widdowquinn/pyani. Accessed 3 Sep 2019.
214. Gibb S. MALDIquant: Quantitative Analysis of Mass Spectrometry Data. :16.
215. Palarea-Albaladejo J, Mclean K, Wright F, Smith DGE. MALDIrppa: quality control and robust analysis for mass spectrometry data. Bioinformatics. 2018;34:522–3.
216. Savitzky Abraham, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem. 1964;36:1627–39.
217. Ryan CG, Clayton E, Griffin WL, Sie SH, Cousens DR. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms. 1988;34:396–402.