UNIVERSITÉ EVRY VAL D'ESSONNE

Ecole Doctorale des Génomes Aux Organismes

Institut National de Recherche Agronomique

Unité de Virologie et Immunologie Moléculaires

Equipe Infection et Immunité des Poissons

THÈSE

Présentée et soutenue publiquement le 13 Novembre 2013

pour l'obtention du grade de

Docteur de l'Université d'Evry Val d'Essonne

Discipline ou Spécialité : Génomique / Biologie Cellulaire et Moléculaire

par :

Paul BARBIER

Diversité génomique des espèces bactériennes du genre *Flavobacterium*

COMPOSITION DU JURY

Président :	**Pr SGHIR Abdelghani**	Professeur, université d'Evry Val d'Essonne
Rapporteurs :	**Pr VALLEYS Tatiana**	Professeur, université Montpellier II
	Dr MICHEL Gurvan	Directeur de Recherche, CNRS
Examinateurs :	**Dr CALVEZ Ségolène**	Maître de conférence, Oniris
	Dr SAPRIEL Guillaume	Maître de conférence, UVSQ
Directeur de thèse :	**Dr DUCHAUD Eric**	Directeur de Recherche, INRA

A mon grand-père, Gilbert Fendt, qui aurait été fier de voir où le petit laboratoire de chimie qu'il m'a offert et l'inspection des tiques en Alsace m'ont mené. Merci papi…

Remerciements

Je souhaite tout d'abord remercier l'école doctorale GAO, l'université d'Evry Val d'Essonne et l'INRA pour m'avoir permis de réaliser ces travaux.

Je tiens à remercier les membres du jury, Mr le Professeur Sghir Abdelghani, le président, le Professeur Tatiana Valleys et le Docteur Gurvan Michel, les rapporteurs, le Docteur Ségolène Calvez et le Docteur Guillaume Sapriel, de m'avoir fait l'honneur d'accepter de juger mon travail.

J'adresse mes plus sincères remerciements à Eric Duchaud, mon directeur de thèse. Je n'aurais pas pu imaginer un meilleur encadrant pour ces années. L'attention qu'il a su déployer à me former, sa patience, ses précieux conseils, sa confiance et l'autonomie qu'il m'a accordée ont été autant d'éléments qui ont contribués au bon déroulement de ces années. Par sa méthode de travail, son investissement dans les projets et son esprit critique, il représente, scientifiquement, une référence et un exemple que je m'efforcerai de suivre au mieux. Au-delà de la science, j'ai beaucoup apprécié la qualité de nos échanges. Tu as réussi à me transmettre ta passion pour ce métier sans m'en cacher les difficultés et donné envie de poursuivre dans cette direction. Merci !

Je tiens à remercier sincèrement l'Unité de Virologie et Immunologie Moléculaire et toute l'équipe Infection et Immunité des Poissons pour son accueil chaleureux et son accompagnement durant ces années passés à leur côté.

J'ai une pensée particulière pour tous ceux qui m'ont aidé au quotidien dans mon travail, Brigitte Kerouault, Aurélie Lunazzi et Armel Houel. Votre aide et votre disponibilité sont pour beaucoup dans la réussite de ce projet. Merci également à Jean-François Bernardet pour sa bonne humeur, sa patience, ses conseils en microbiologie et son aide lors de la relecture des articles qui ont été d'une grande aide. Je n'oublie évidemment aucun des membres de l'équipe avec qui ces années ont été très agréables tant sur le plan humain que scientifique : Céline Chantry, Fabienne Neulat-Ripoll, Christelle Langevin, Corinne Thory, Tatiana Rochat, Elina Aleksejeva, Abdenour Benmansour, Christian Michel, Robert Laroque, Eloi Verrier et Christophe Habib. Merci à Luc Jouneau avec qui les longues journées au labo

ont été ponctuées d'agréables moments. J'espère que tu franchiras un jour la porte des laboratoires sans prendre de douche !

Un grand merci enfin à Pierre Boudinot car au-delà de l'apport scientifique, votre enthousiasme, votre aide et votre dynamisme ont largement contribué au bon déroulement de ces années. Une pensée particulière à Pierre de Kinkelin dont la richesse des connaissances et la présence manquent à tous.

Cette thèse est également le fruit de collaborations multiples. Je souhaite à ce titre remercier particulièrement Marie Touchon pour sa patience, sa réactivité, son aide, ses suggestions et ses judicieux conseils en bioinfo. Je remercie également nos collaborateurs étrangers Erina Fujiwara-Nagata, Ruben Avendano, Tom wiklund et Krister Sundel pour leur contribution. Je remercie les membres du comité de thèse, Muriel Vayssier, Frédérique Le Roux et Michel Le Hénaff d'avoir accepté de prendre le temps de regarder mon travail.

Un grand merci à l'équipe des Castors de Jouy-en-Josas pour les très bons moments que nous avons partagés sur le terrain comme en dehors. Votre accueil et votre amitié ont largement contribué à rendre très agréables toutes ces années.

Je remercie également l'association des doctorants « DOC'J » qui m'a permis d'échanger avec de nombreux jeunes scientifiques, en particulier Laure Decamps, Chris Hoze, Belén Jimenez Mena, Pauline Maisonnasse et Aude Remot qui se sont beaucoup investies pour permettre à l'association de continuer à participer et à animer la vie du centre de recherche de Jouy. Je garderai d'excellents souvenirs des colloques de jeunes scientifiques que nous avons organisés ensemble sur le centre.

Je remercie enfin chaleureusement ma famille et mes amis pour leurs encouragements et leur soutien. Je remercie particulièrement mes parents qui m'ont toujours incité et aidé à faire ce qui me plaisait dans la vie.

Résumé

Les bactéries du genre *Flavobacterium* sont retrouvées dans des types d'habitats très divers (eaux douces, océans, sols, glaciers, environnements polaires et une source chaude, entre autres). La majorité des espèces ont été isolées de l'environnement et sont considérées comme non pathogènes. Cependant, ce genre contient également trois espèces ichtyopathogènes : *F. columnare*, *F. branchiophilum* et *F. psychrophilum*. Cette dernière affecte principalement l'élevage des salmonidés et est responsable de pertes économiques importantes en France et dans le monde.

Un projet de séquençage et de comparaison des génomes de plusieurs flavobactéries pathogènes de poissons ainsi qu'isolées de différents environnements a été mis en place pour contribuer à améliorer les connaissances sur ce genre peu étudié. Les objectifs étaient l'identification des déterminants de virulence associés à la pathogénicité, la caractérisation de différents marqueurs moléculaires des traits phénotypiques associés à leur mode de vie et leur mise en relation avec les niches écologiques qu'elles colonisent.

L'analyse des génomes de plusieurs isolats de *F. psychrophilum* a permis de mettre en évidence une diversité des structures chromosomiques au sein de l'espèce et d'identifier *in silico* des cibles moléculaires prometteuses pour le développement de tests de diagnostic spécifiques de l'espèce ainsi que des cibles vaccinales potentielles. Le génome de *F. branchiophilum* a permis d'identifier des mécanismes moléculaires de virulence originaux. Les caractéristiques du génome de *F. indicum* révèlent un mode de vie « environnemental » : sa petite taille (2,9 Mpb) et ses faibles capacités de dégradation des bio-polymères suggèrent que *F. indicum* est adapté à une niche écologique restreinte. Ces nouvelles données ont permis de caractériser *in silico* de nombreux marqueurs moléculaires de caractères phénotypiques (locomotion, attachement, nutrition, virulence, etc…). En particulier, un groupe de gènes (*dnd*) rare et responsable d'une modification étonnante de la structure de la molécule d'ADN a été décrit pour la première fois chez des membres de la famille des *Flavobacteriaceae*.

Ce projet, qui a fait appel à des approches de microbiologie, de biologie moléculaire et de génomique bactérienne, a permis d'enrichir les connaissances sur les bactéries du genre *Flavobacterium* et a contribué au développement d'outils pour la santé animale.

Mots-clés : Microbiologie, Bactéries pathogènes des poissons, Génomique comparative, Génomique fonctionnelle, *Flavobacterium*

Abstract

Flavobacterium species occur in a wide range of habitats (fresh water, oceans, soils, glaciers, polar regions and a warm spring water, among others). Most species are isolated from the environment and considered non-pathogenic. However, this genus also includes three fish-pathogenic species, namely *F. columnare*, *F. branchiophilum* and *F. psychrophilum*. The latter mainly affects farmed salmonids and is responsible for serious economic losses in France and worlwide.

A comparative genomics project including several fish-pathogenic flavobacteria as well as various environmental species has been set up in order to improve the knowledge on this poorly studied genus. Our aims were the identification of virulence determinants associated with pathogenicity, the characterization of various molecular elements reflecting phenotypes associated with their life-style and the establishment of relationships with the ecological niches in which they occur.

Analysis of the genomes of several *F. psychrophilum* isolates revealed the diversity of chromosomal structures within the species and identified *in silico* promising molecular targets for the development of diagnostic tests as well as potential vaccines targets. Analysis of the *F. branchiophilum* genome enabled to identify particular molecular virulence mechanisms. The features of the *F. indicum* genome reflected its environmental lifestyle : its small size (2,9 Mbp) and its limited bio-polymers degrading abilities suggested that *F. indicum* is adapted to a quite narrow ecological niche. These new data have allowed the *in silico* identification of many molecular elements reflecting phenotypic traits (motility, adherence, nutrition, virulence, etc…). In particular, a rare gene cluster (*dnd*) responsible for an unusual DNA structure modification was described for the first time within members of the family *Flavobacteriaceae*.

This project, that combined microbiology, molecular biology and bacterial genomics approaches, enriched the knowledge on *Flavobacterium* species and contributed to the development of tools for animal health.

Keywords : Microbiology, Fish-pathogenic bacteria, Comparative genomics, Fonctional genomics, *Flavobacterium*

Table des matières

Première partie : Validation de l'assemblage des génomes complets de *F. psychrophilum* souches JIP 02/86 et THC 02/90. Diversité des structures chromosomiques au sein de l'espèce.

Deuxième partie : Etude du génome complet d'une bactérie pathogène de poisson : *Flavobacterium branchiophilum*

Troisième partie : Etude du génome complet de *Flavobacterium indicum*, isolé d'une source chaude

Quatrième partie : Etude d'un groupe de gènes particulier : les gènes *dnd* ; analyse de leur distribution au sein du phylum *Bacteroidetes*

Discussion Générale

Liste des annexes

Liste des abréviations

ADN/DNA	Acide désoxyribonucléique
ADNr/rDNA	ADN ribosomique
AOBE	Milieux de culture des flavobactéries (Anacker Ordal Broth Enriched)
ARNt/tRNA	ARN de transfert
BAC	Chromosome artificiel bactérien (Bacterial Artificial Chromosome)
BET	Bromure d'éthidium
BGD	Maladie des branchies (Bacterial Gill Disease)
CTD	Domaine Carboxy-Terminal
CWD	Maladie des eaux froides (Cold Water Disease)
dUTP	Déoxyuridine triphosphate
ICSP	Comité International de Systématique des Procaryotes
IS	Séquence d'insertion
LPSN	List of Prokaryotic names with Standing in Nomenclature
MLST	Multi Locus Sequence Typing
Mpb, kpb, pb	Méga, kilo et paire de bases, respectivement
NGS	Séquençage de nouvelle génération (Next Generation Sequencing)
PCR	Réaction de polymérisation en chaine (Polymerase Chain Reaction)
PFGE	électrophorèse en champ pulsé (pulsed field gel electrophoresis)
PUL	Polysaccharides Utilization Locus
RTFS	Maladie des alevins (Rainbow Trout Fry Syndrome)
TRIS	trihydroxyméthylaminométhane (2-amino-2-hydroxyméthyl-1,3 propanediol)

Avant-propos

Depuis mon stage de première année dans un laboratoire de thérapeutique avancée en cancérologie à Vancouver en 2006 jusqu'à mon arrivée à l'INRA en 2008, en passant par le laboratoire des maladies tropicales de l'INSERM en 2007 où j'ai travaillé sur le cycle erythrocytaire du parasite *Plasmodium falciparum*, j'ai acquis très jeune, un intérêt pour la recherche en général et pour la microbiologie en particulier.

Ma formation universitaire m'a conduit à l'obtention d'un Master de recherche en Biologie Moléculaire et d'une spécialisation en Microbiologie. Passionné par l'étude des mécanismes de l'ADN et vivement intéressé par l'étude des bactéries, la génomique bactérienne m'est apparue comme un formidable compromis entre ces deux disciplines. La cohérence entre ma formation et le sujet de ce projet doctoral a donc été déterminante dans le choix de mon projet de recherche.

La séquence d'ADN contient l'information nécessaire aux êtres vivants pour survivre et se reproduire. Déterminer cette séquence est donc utile aussi bien pour les recherches visant à comprendre comment vivent les organismes que pour des sujets appliqués. L'explosion de l'utilisation des approches génomiques, largement visible pendant mes dernières années universitaires, m'est tout de suite apparue comme un progrès majeur et une révolution dans la manière d'aborder notre compréhension du Vivant. Participer à cet élan scientifique en utilisant ces nouvelles technologies a été une source de motivation supplémentaire.

Les projets de séquençage et d'analyse de génomes sont de formidables exercices de synthèse entre plusieurs disciplines et approches de la biologie. Participer à ces projets m'est donc toujours apparu comme extrêmement formateur et enrichissant.

Introduction

Introduction Générale

La pisciculture en France et dans le monde

La consommation de poissons par l'Homme augmente (de 100 à 110 millions de tonnes de 2002 à 2006) [1]. Cependant les tonnages pêchés ont atteint aujourd'hui leurs limites et les stocks de poissons diminuent dans certaines zones du globe. Cette différence correspond donc au développement de l'aquaculture. Le secteur, en pleine mutation, a de fait connu une forte progression ces dernières années d'environ 10% par an [1]. La production mondiale piscicole a atteint 53 millions de tonnes en 2008 et aujourd'hui la majorité du poisson consommé dans le monde provient de la pisciculture [2]. Les espèces de poissons les plus élevées au monde sont les carpes, suivies du tilapia, des salmonidés et des siluriformes [2].

Si le secteur est largement dominé par l'Asie qui produit 90% du volume mondial de poissons d'élevage [2], la pisciculture est aussi présente en Europe et notamment en France. La pisciculture française d'eau douce et marine produit plus de 50000 tonnes de poisson par an et emploie aujourd'hui 2500 personnes sur plus de 600 sites de productions [3]. Avec près de 34000t produites par an, la France est le 4ème éleveur mondial de truites d'eau douce, notamment de truites arc-en-ciel (*Oncorhynchus mykiss*) qui représentent 95% de la production [3]. Cependant, la plus grande partie du poisson consommé aujourd'hui en France provient de l'importation. Les exportations sont plus souvent liées à la pisciculture marine notamment grâce à un savoir faire zootechnique français important de sélection génétique et de production d'alevins dans les écloseries.

Néanmoins les maladies pouvant sévir dans les élevages et affectant la production de poissons sont un frein majeur au développement de la filière piscicole en France et dans le monde [1], [2]. Par exemple, une grande crise sanitaire en 2008 au Chili [4] a fait perdre au niveau national environ un tiers de la production en deux ans.

De nombreuses pathologies peuvent survenir dans les élevages de poissons. Les agents pathogènes les plus fréquemment signalés par les personnes ayant réalisé des diagnostics de terrain en France en 2003 sont des bactéries (46%), des virus (20%), des parasites (18%) et d'autres agents étiologiques comme les mycoses ainsi que les pollutions et maladies nutritionnelles [5].

Les pathologies d'origine bactérienne en pisciculture

Les principales pathologies bactériennes et leurs agents étiologiques signalées en France sont la furonculose (*Aeromonas salmonicida*), la flavobactériose d'eau froide (*Flavobacterium psychrophilum*), les streptococcoses (*Streptococcus iniae*, *Lactococcus garviae* ou *Vagococcus salmoninarum*), la yersiniose (*Yersinia ruckeri*) et la flexibacteriose (*Tenacibaculum maritimum*, anciennement *Flexibacter maritimus*). Il ne s'agit pas ici de prévalence mais de déclarations qui suivant l'importance accordée par les professionnels impliqués, peuvent apparaître sur-déclarées en fonction des émergences et réglementations ou sous-déclarées par rapport à leur présence réelle en piscicultures [5]. De plus, ces déclarations peuvent être également « faussées » par un manque de connaissances à l'origine de regroupements sous une même pathologie de signes cliniques identiques dus à des agents bactériens différents.

Les analyses réalisées par les pathologistes de terrain vont souvent jusqu'à l'isolement bactérien, essentiellement pour effectuer des antibiogrammes. Ces méthodes ne permettent généralement pas de diagnostiquer la présence d'agents pathogènes chez des porteurs

asymptomatiques ou les œufs par exemple (dans le cas des transmissions verticales). Il existe peu d'outils aujourd'hui disponibles pour effectuer des diagnostics moléculaires précis et fiables, notamment pour la flavobactériose à *Flavobacterium psychrophilum*. C'est dans l'optique de répondre à ce besoin de connaissances génériques sur ces bactéries, nécessaire à long terme pour aider à une gestion efficace du diagnostic et du contrôle de ces maladies, que s'inscrit une partie des travaux de cette thèse.

La question du contrôle et de la surveillance des pathologies prend de plus en plus d'importance dans la filière aquacole. Dans cette visée, l'Institut National de la Recherche Agronomique (INRA) s'est investi, depuis une trentaine d'années, au côté des pisciculteurs et des vétérinaires aquacoles pour améliorer le diagnostic et le contrôle des nombreuses pathologies affectant les élevages piscicoles.

L'équipe Infection et Immunité des poissons (INRA, UR0892 Virologie et Immunologie Moléculaires)

L'équipe Infection et Immunité des Poissons (IIP) de l'INRA de Jouy-en-Josas a été pionnière dans l'étude de la famille *Flavobacteriaceae* et du genre *Flavobacterium*. La conduite d'une étude originale en 1996 en a fait une référence au niveau mondial dans l'étude des flavobactéries.

Le séquençage et l'annotation du génome complet de *Flavobacterium psychrophilum* (souche JIP 02/86) à été réalisé en 2007 par cette équipe [6]. Ce génome a été le premier génome disponible d'une bactérie pathogène pour les poissons. Second génome à être séquencé au sein de la famille des *Flavobacteriaceae*, il fut cependant le premier génome complet disponible pour le genre *Flavobacterium*. Cette séquence a donc été le point de départ de l'étude du genre par des approches génomiques.

Par la suite, l'équipe a utilisé cette séquence pour conduire la première étude de diversité génétique de l'espèce *F. psychrophilum* fondée sur le séquençage de plusieurs locus (ou MLST) [7]. Le développement d'outils de typage et de caractérisation moléculaires de ces bactéries pathogènes de poissons a permis d'appréhender la structure et le suivi des populations de souches isolées sur le terrain.

En parallèle, des expériences d'épreuves infectieuses sur truites arc ciel ont suggéré des différences dans la virulence et le spectre d'hôtes entre des isolats de *F. psychrophilum*. Une étude de génomique comparative intra-espèce, fondée sur le séquençage de plusieurs génomes complets d'isolats provenant d'origines géographiques et de poissons hôtes différents, a été mise en place pour essayer d'identifier les gènes de virulence et les déterminants associés à la pathogénicité. Ces données permettent également de participer, en collaboration avec des partenaires, à la mise au point d'outils de diagnostic et l'identification de candidats vaccins utilisables sur le terrain.

Enfin, la production de séquences complètes de bactéries isolées de l'environnement et retrouvées dans des niches écologiques très diverses a permis la mise en place d'une étude de génomique comparative inter-espèces au sein du genre *Flavobacterium*. Cette approche permet d'enrichir les connaissances fondamentales en écologie microbienne, de compléter les informations sur la dynamique évolutive au sein du genre *Flavobacterium* et de caractériser les éléments marqueurs de ces différentes niches écologiques.

Une famille, un genre, trois espèces ichtyopathogènes

La famille des *Flavobacteriaceae*

La famille des *Flavobacteriaceae* appartient au phylum *Cytophaga-Flavobacterium-Bacteroides* ou phylum *Bacteroidetes*, à la classe des *Flavobacteriia* et à l'ordre des *Flavobacteriales*. Evoquée pour la première fois en 1992 par Reichenbach, la véritable description de la famille des *Flavobacteriaceae* et de son genre type, le genre *Flavobacterium*, a été publiée en 1996 [8]. Cette étude réalisée sur une centaine de souches bactériennes a permis de décrire la famille des *Flavobacteriaceae* à partir d'un ensemble de caractères phénotypiques communs.

Les éléments permettant cette classification ont depuis été corrigés et enrichis à plusieurs reprises [9]. La caractérisation de nouvelles espèces cultivables au sein des *Flavobacteriaceae* est basée sur des approches dites polyphasiques. Cette approche intègre les relations phylogénétiques déduites du séquençage d'un fragment du gène codant la sous-unité 16S du ribosome, les données génomiques et phylogénétiques comme l'hybridation ADN-ADN et les caractérisations phénotypiques et chimiotaxonomiques classiques.

Le nombre de genres inclus dans la famille des *Flavobacteriaceae* ne cesse de croître. En 2013, cette famille est constituée de 109 genres bactériens différents dont les genres *Flavobacterium* et *Tenacibaculum* [10].

Niches écologiques et pouvoir pathogène

Le concept théorique de la niche écologique, introduit par Haeckel en 1866, traduit à la fois une dimension abiotique reflétant l'ensemble des conditions nécessaires à la vie de l'organisme en question et également une dimension biotique, due à la présence d'autres

organismes dans un écosystème donné. En microbiologie, la description d'une niche écologique se fait donc généralement sur la base de plusieurs paramètres comme les paramètres physico-chimiques caractérisant les milieux où évolue l'organisme et un ensemble de paramètres biologiques. Ces derniers incluent l'utilisation des ressources et nutriments disponibles et les interactions avec la communauté des espèces avoisinantes.

Les niches écologiques des représentants de la famille des *Flavobacteriaceae* sont très variées: sols, eau douce et sédiments des lacs et des rivières, eau de mer et environnements marins, boues activées, biofilms, plantes et aliments (notamment produits laitiers). Plusieurs espèces de la famille des *Flavobacteriaceae* ont été isolées de divers prélèvements effectués chez les animaux: carnivores, oiseaux, poissons d'eau douce ou d'eau de mer, amphibiens, reptiles, insectes, crustacés, échinodermes, mollusques, amibes, spongiaires et cavité buccale de l'homme.

Dans le milieu extérieur, ces bactéries joueraient un rôle important dans la dégradation de multiples substrats organiques et seraient impliquées dans les grands cycles géochimiques de recyclage de la matière organique [11]. Les espèces vivant dans le sol ou l'eau douce peuvent synthétiser par exemple des enzymes capables de dégrader la cellulose (dérivés solubles), la pectine, le xylane ou la chitine des champignons, des insectes ou des plantes [12], [13]. Les espèces vivant dans l'eau de mer peuvent dégrader l'agar, la laminarine, le xylane, le fucoidane ou les carraghénanes qui sont des polymères de glucose complexes retrouvés dans la paroi des algues [14], [15].

La majorité des espèces responsables d'infections sont parfois également isolées du milieu extérieur et se comportent comme des pathogènes opportunistes. Par exemple, les espèces pathogènes des genres *Bergeyella, Capnocytophaga, Chryseobacterium, Elizabethkingia, Empedobacter, Flavobacterium, Myroides* et *Weeksella* ne sont responsables de maladies que chez des sujets immunodéprimés, affaiblis ou placés dans de mauvaises conditions environnementales. Dans une minorité de cas, les espèces responsables d'infections ne sont pas isolées du milieu extérieur et sont des pathogènes authentiques. C'est notamment le cas pour les espèces bactériennes pathogènes pour les oiseaux du genre *Riemerella* [16].

Le genre *Flavobacterium*

La première description du genre *Flavobacterium* figure dans l'édition de 1923 du *Bergey's Manual of Determinative Bacteriology*. Elle a été enrichie plusieurs fois ces dernières années. Ce genre a notamment été décrit en tant que genre type de la famille des *Flavobacteriaceae* en 1996 [8]. L'espèce type du genre, *Flavobacterium aquatile*, a été isolée en 1889 par Mr et Mme Frankland sous le nom de *Bacillus aquatilis* à partir d'un puits d'eau en Angleterre. Au fur et à mesure de la description de nouvelles espèces bactériennes et de l'évolution de leur taxonomie, cette souche a plusieurs fois changé de nom pour finalement être appelée *F. aquatile* en 1980.

Le genre *Flavobacterium* comprenait une dizaine d'espèces en 1996, 66 espèces en 2010 et regroupe aujourd'hui, en milieu d'année 2013, plus de 120 espèces dont l'appartenance à ce genre a été approuvée par le sous-comité de taxonomie des bactéries *Flavobacterium* et *Cytophaga* du Comité International de Systématique des Procaryotes (ICSP), référence de la taxonomie bactérienne. L'évolution au cours du temps du nombre d'espèces décrites pour ce genre est présentée en Figure 1. Une liste exhaustive présentant les espèces bactériennes du genre *Flavobacterium* et leurs années de validation par le comité de taxonomie est présenté sur le site « List of Prokaryotic names with Standing in Nomenclature » (LPSN)[17].

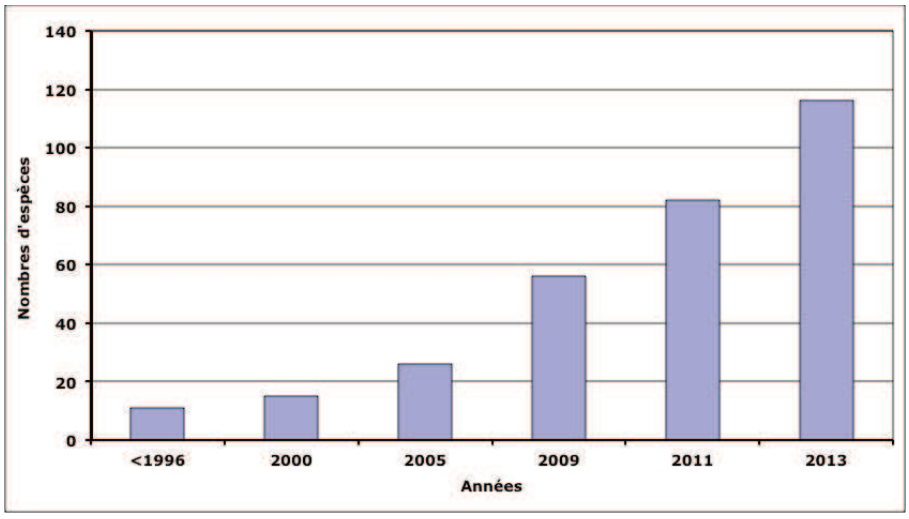


Figure 1 : Histogramme en barre représentant l'accumulation du nombre d'espèces décrites au sein du genre *Flavobacterium* au cours du temps. Le compte pour l'année 2013 a été arrêté en Janvier.

Caractérisation polyphasique des espèces du genre *Flavobacterium*

Les représentants du genre *Flavobacterium* sont des bacilles à Gram négatif, non sporulés. Le diamètre des cellules est généralement compris entre 0,2 et 0,6 μm et leur longueur est généralement comprise entre 1 et 10 μm. Certaines espèces peuvent éventuellement former des cellules filamenteuses et/ou donner des formes coccoïdes dans les vieilles cultures.

Les colonies de bacilles du genre *Flavobacterium* sont colorées en jaune pâle, vif ou orangé. Ces couleurs sont dues à la présence de pigments de type caroténoïdes ou flexirubines selon les espèces. En effet, au sein du genre *Flavobacterium*, on retrouve des organismes produisant des caroténoïdes ou des flexirubines ainsi que des espèces produisant les deux types de pigments [8].

Toutes les souches sont dépourvues de flagelle, cependant une mobilité par glissement [8], [18] est observée pour les souches de la plupart des espèces.

Toutes les espèces du genre *Flavobacterium* présentent une croissance en conditions aérobies et la plupart peuvent êtres considérés comme mésophiles avec une température optimale de croissance habituellement comprise entre 23 et 35 °C. Toutefois, plusieurs espèces sont psychrophiles ou psychro-tolérantes. C'est notamment le cas des espèces *F. psychrophilum*, *F. frigidimaris* et *F. glaciei*. Bien que l'halophilie ne soit pas une caractéristique des espèces du genre, certaines espèces isolées de milieux marins peuvent être considérées comme halophiles simples (par opposition à strictes) ou halo-tolérantes [8], [19].

Au sein du genre *Flavobacterium*, les capacités métaboliques classiquement testées pour l'identification microbiologique, telles que catalase, oxidase et nitrate réductase peuvent varier en fonction des espèces et des souches au sein même d'une espèce. Il en est de même pour la dégradation et l'utilisation des polymères de sucres plus complexes comme l'amidon par exemple [17], [9].

La proportion de bases G et C de l'ADN des représentants du genre *Flavobacterium* est comprise entre 32 et 37 mol%. En plus des caractères bactériologiques et chimiotaxonomiques (profils d'acides gras, de lipides polaires, de polyamines et de quinones

respiratoires), la description d'une espèce au sein du genre *Flavobacterium* est complétée par une analyse phylogénétique déduite du séquençage d'un fragment du gène codant la sous-unité 16S du ribosome, éventuellement complétée par celle du gène *gyrB* codant la sous-unité B de l'ADN gyrase. Ces analyses sont complétées par l'hybridation ADN-ADN avec des espèces proches du genre *Flavobacterium* et des espèces de référence pour vérifier l'appartenance d'une souche à une espèce. C'est le pourcentage d'hybridation ADN-ADN (d'une valeur seuil < 70%) qui permet de déterminer une nouvelle espèce.

Les *Flavobacterium* « environnementales »

Les différentes espèces du genre *Flavobacterium* sont retrouvées dans de très nombreux habitats et la très grande majorité d'entre elles sont isolées de l'environnement. Ces espèces peuvent vivre librement dans le milieu extérieur, ou être associées à des végétaux ou encore saprophytes. Les espèces du genre *Flavobacterium* occupent presque tous les types d'habitats en milieu tempéré. Les différentes espèces ont été isolées à partir du sol, d'eau douce et sédiments des lacs et rivières, des boues d'épurations, ou encore associées aux racines de plantes. Certainement à cause de leur abondance dans des environnements d'intérêt qui ont été les objets d'intenses campagnes scientifiques visant à récolter de nouvelles espèces, de nombreuses espèces peuvent également êtres retrouvées dans les milieux polaires et extrêmes tels que les glaciers, les sols gelés, une source d'eau chaude et des environnements pollués [19], [20], [21], [22]. C'est notamment le cas des espèces *F. indicum* isolée d'une source chaude en Inde [23], *F. frigidimaris* isolée d'eau de mer en Antarctique et *F. glaciei* isolée d'un glacier chinois [24]. Leur présence dans des milieux extrêmes laisse penser que les bactéries du genre *Flavobacterium* produisent donc de nombreuses enzymes, dont certaines adaptées au froid, qui pourraient présenter un intérêt biotechnologique [25].

La première espèce de *Flavobacterium* isolée de l'environnement qui a été séquencée est *Flavobacterium johnsoniae*. Cette bactérie commune des sols et de l'eau douce est depuis devenue un modèle d'étude de l'utilisation des polysaccharides dans des environnements oligotrophiques (souche A3) [26] et de la mobilité par glissement (souche UW101) [13].

La limite des caractérisations phénotypiques et la constante augmentation du nombre d'espèces du genre *Flavobacterium* décrites depuis quelques années ont accru le besoin de références solides permettant de définir une espèce. Les génomes complets tendent à répondre à ce besoin. En particulier, de plus en plus de génomes de bactéries isolées de l'environnement sont décrits comme appartenant au genre *Flavobacterium* sans avoir reçu une assignation taxonomique et phénotypique précise. C'est par exemple le cas des huit génomes, séquencés entre 2012 et 2013, portant le nom de *Flavobacterium* sp.. Ces organismes ont été isolés de milieux variés tels que la rhizosphère [27], l'eau douce de rivières et les sols mais peuvent aussi ne pas être cultivables et leurs séquence sont issues de données de métagénomique acquises à partir d'environnements comme les écosystèmes marins par exemple [28].

Les espèces non pathogènes isolées de l'environnement sont principalement séquencées pour répondre aux besoins de connaissances sur ce genre bactérien et identifier des traits spécifiques (par exemple *F. johnsoniae* pour le glissement et *Flavobacterium* sp. F52 pour la fixation de l'azote). Elles peuvent également servir à des fins de comparaison avec les espèces pathogènes du genre. De plus, l'analyse des caractères phénotypiques en relation avec les niches écologiques permet de faire des liens avec les informations contenues dans le génome. Il est donc très important de séquencer et travailler sur des souches bien caractérisées phénotypiquement pour permettre d'assurer ce lien lorsqu'il s'agit de souches cultivables. C'est ce lien entre génotype, phénotype et niche écologique qui permet de comprendre comment ce genre bactérien a évolué et comment ses représentants se sont adaptés à leurs niches écologiques. Dans le cas des espèces pathogènes, ce lien permettra également d'essayer de comprendre comment ces espèces se sont adaptées à leurs hôtes.

Les espèces ichtyopathogènes du genre *Flavobacterium*

Au sein du genre *Flavobacterium,* les espèces *F. columnare*, *F. branchiophilum* et *F. psychrophilum* sont pathogènes pour les poissons et notamment pour les salmonidés (truites et saumons). Isolées uniquement en eau douce, ces trois espèces pathogènes présentent des caractéristiques variables en termes de spectre d'hôtes, de signes cliniques, de températures de croissance ou encore de répartition géographique. Ces différentes caractéristiques sont résumées pour les trois espèces dans le Tableau 1.

Flavobacterium columnare a été décrit en 1922 et est une des plus anciennes bactéries connues pour infecter les poissons en eaux douces et chaudes. Elle est responsable d'une infection connue sous le nom de columnariose qui affecte de très nombreuses espèces de poissons d'eau douce comme les salmonidés, les cyprinidés, les anguillidés et est également fréquemment isolée de poissons ornementaux [29]. C'est un problème majeur pour l'élevage du poisson chat (*Ameiurus melas*) essentiellement dans le bassin du Mississipi aux USA. La maladie a une répartition mondiale et a été décrite en France en 1989. *Flavobacterium columnare* est responsable de lésions cutanées et/ou des branchies [30].

Flavobacterium branchiophilum [31] est la bactérie la plus souvent impliquée dans le syndrome "maladie des branchies" ou "bacterial gill disease" (BGD). L'infection a été décrite principalement au Japon, en Corée, aux USA et au Canada. Des facteurs prédisposant comme des stress ou une mauvaise qualité de l'eau ont été évoqués pour expliquer le développement de l'infection mais, la maladie a pu être reproduite chez des poissons élevés dans de bonnes conditions. *Flavobacterium branchiophilum* n'envahit pas les autres tissus, ne se dissémine pas dans l'organisme et n'a jamais été isolé d'organes internes. Cette bactérie présente un tropisme pour les cellules épithéliales des branchies et provoque l'asphyxie de l'hôte [32],[33].

Flavobacterium psychrophilum a été isolée pour la première fois en 1948 sous le nom de *Cytophaga psychrophila* et fut finalement classée dans le genre *Flavobacterium* en 1996 [8]. *F. psychrophilum* est l'agent responsable de la « cold water disease » (CWD) chez les poissons adultes et du « rainbow trout fry syndrome » (RTFS) chez les alevins, deux pathologies communément appelées flavobactérioses d'eau froide. Ce pathogène est une

cause majeure de mortalité de la truite arc-en-ciel en élevage et constitue une des menaces parmi les plus importantes pour l'élevage des salmonidés [34]. *F. psychrophilum* est plutôt considéré comme un pathogène strict. Toutefois, cette bactérie a été retrouvée occasionnellement dans l'eau [35] ou dans d'autres compartiments environnementaux tels que les sédiments ou les algues [36]. De plus, l'existence de poissons porteurs asymptomatiques de la bactérie laisse supposer que des populations naturelles de poissons peuvent en être probablement le réservoir [37]. La transmission « quasi-verticale » (les bactéries sont uniquement retrouvées dans le mucus et le chorion des oeufs) [38] est vraisemblablement responsable de la diffusion de la bactérie dans les régions salmonicoles à travers le commerce des œufs, en dépit de leur désinfection systématique par des dérivés iodés. Les mesures thérapeutiques actuelles reposent essentiellement sur l'utilisation d'antibiotiques [39] administrés par voie orale via l'alimentation au moment de l'apparition des signes cliniques caractéristiques.

Espèce	**Spectre d'hôtes**	**Pathologie**	**Signes cliniques**	**Mobilité par glissement**	**Température de croissance**	**Localisation géographique**
Flavobacterium psychrophilum	Etroit : salmonidés (aquaculture)	Flavobactériose d'eau froide (CWD) sur adultes Septicémie hémorragique (RFTS) sur alevins	Invasion tissulaire, septicémie hémorragique (parfois lésions externes)	Oui	De 4 à 20 °C	Toutes les régions du monde tempérées froides
Flavobacterium columnare	Large : poisson chat, carpe, tilapia, anguille, salmonidés (aquaculture, ornementaux et populations sauvages)	Columnariose	Ulcères musculaires et nécroses des tissus cutanés, des branchies et des nageoires (rares infections systémiques)	Oui	De 25 à 32 °C	Toutes les régions du monde tempérées chaudes
Flavobacterium branchiophilum	Etroit : salmonidés (aquaculture)	Flavobactériose des branchies	Colonisation des branchies (biofilm + mucus), asphyxie de l'hôte (très rarement isolée d'organes internes)	Non	De 10 à 25 °C	Japon, Corée, USA, Canada, Hongrie et Pays-Bas

Tableau 1 : Tableau résumant les différentes caractéristiques des trois espèces ichtyopathogènes du genre *Flavobacterium*.

Le séquençage de génomes complets

Séquençage de génomes

Le séquençage de l'ADN a été découvert dans la deuxième moitié des années 1970. Deux méthodes ont été développées indépendamment, l'une par l'équipe de Walter Gilbert, aux États-Unis, et l'autre par celle de Frederick Sanger, en Grande-Bretagne.

Au cours des 30 dernières années, la méthode de Sanger a été largement développée grâce à plusieurs avancées technologiques importantes telles que le développement de la synthèse chimique automatisée des oligonucléotides qui sont utilisés comme amorces dans la synthèse, l'introduction de traceurs fluorescents à la place des marqueurs radioactifs, la mise au point d'une méthode d'amplification linéaire utilisant peu de matrice et l'utilisation de séquenceurs automatiques ainsi que de l'électrophorèse capillaire.

Les premiers génomes complètement séquencés ont tout d'abord été ceux de virus puis de micro-organismes. Le séquençage complet des génomes de trois micro-organismes modèles que sont la levure *Saccharomyces cerevisiae* et les bactéries *Bacillus subtilis* et *Escherichia coli*, a été effectué grâce au travail concerté, durant plusieurs années, de laboratoires de recherche regroupés en consortiums internationaux. Les séquences de ces espèces ont été obtenues en totalité entre 1996 et 1997. Parallèlement, des entreprises privées se sont lancées de manière efficace dans le séquençage. Ainsi, le premier séquençage complet de génome bactérien a été celui de la bactérie pathogène *Haemophilus influenzae* [40] réalisé par l'entreprise TIGR (The Institute for Genomics Research) en 1995 et a précédé ceux de *Escherichia coli* K-12 [41] et de *Bacillus subtilis* [42].

Récemment, une nouvelle génération de séquenceurs à très haut débit est apparue. Ces techniques permettent de séquencer, en quelques jours (heures), plusieurs gigabases d'ADN et sont regroupées sous le nom de méthodes « NGS » pour Next (ou New) Generation Sequencing.

Séquençages de nouvelle génération (NGS)

Un ensemble de nouvelles méthodes de séquençage qui permettent de réaliser du séquençage à très haut débit est apparu à partir de 2005. Les avantages de ces technologies sont nombreux : pas d'étapes de clonage bactérien (et donc pas de biais inhérents à la construction des banques), rapidité (moins d'une semaine) et coûts beaucoup moins élevés (coût par paire de base Solexa environ 9000 fois moins cher que par le séquençage Sanger). Ces nouveaux séquenceurs sont 454 (Roche), Solexa (Illumina), SOLID (Applied Biosystem), PacBio RS (Pacific Bioscience) et Ion Torrent (Life Technologies). Les technologies SOLID, PacBio RS et Ion Torrent n'ont pas été utilisées dans les projets de séquençage réalisés dans le cadre de cette thèse et ne seront donc pas détaillées.

La technique 454 est une technique de séquençage basée sur l'amplification de l'ADN par PCR en émulsion et sur le pyroséquençage (luminescence par libération de pyrophosphate). Une présentation détaillée du principe est accessible sur le site internet de Roche [43]. Les inconvénients liés à cette technique sont le taux d'erreur assez élevé, en particulier dans les régions homopolymériques, et la fenêtre de lecture étroite (la version FLEX utilisée pour le séquençage de *F. indicum* a une taille de lecture de 225 pb – aujourd'hui les séquences obtenues sont d'environ 700 pb).

La technique Solexa est basée sur l'amplification préalable des fragments d'ADN sur une lame. Pour déterminer la séquence, des nucléotides terminateurs réversibles marqués et fluorescents sont incorporés par amplification d'un brin complémentaire. Une image de la fluorescence est prise avant que la partie fluorescente, fixée à l'extrémité 3' de la base, soit enlevée chimiquement permettant la réalisation du cycle suivant. Egalement appelée séquençage par terminaison cyclique réversible, une présentation détaillée du principe de cette technique est accessible sur le site internet d'Illumina [44]. Les avantages liés à la technique Solexa sont la rapidité du séquençage (jusqu'à un million de bases lues par seconde en 2012) et le très faible taux d'erreur. Récemment, la possibilité de réaliser des lectures « paired-end » (type de séquençage qui génère une paire de séquences (reads) séparées par une distance connue) et la possibilité de multiplexer les échantillons (plusieurs échantillons différents peuvent être séquencés en même temps) ont fait de cette technologie une méthode

de choix pour l'obtention de génomes bactériens complets *de novo* (c'est à dire s'il n'existe pas de séquence de référence) [45].

Ces nouvelles méthodes de séquençage offrent de nouvelles perspectives dans les domaines du diagnostic médical et de la métagénomique permettant une analyse des génomes de communautés de micro-organismes présents dans l'environnement sans isolement et culture préalables. Ces nouvelles méthodes rendent possibles le séquençage de génomes entiers à des coûts de plus en plus faibles, augmentant ainsi le volume de données et de connaissances disponibles. Cependant, ces avancées posent de nouveaux problèmes dans la compilation, l'étude et la fiabilité des données. En effet, les besoins de traitement et d'analyse de ces nombreuses données ont contribué au besoin de développement de solutions bioinformatiques.

Stratégies de séquençage des génomes

La méthode de séquençage de génomes entiers aujourd'hui utilisée est dite globale. L'ADN génomique est préalablement fragmenté par des méthodes enzymatiques (enzymes de restriction) ou physiques (ultrasons ou nébulisation) puis utilisé pour construire des librairies. L'approche génère des séquences de fragments du génome de manière aléatoire. Des programmes d'assemblage (assembleurs) permettent ensuite de réordonner les fragments génomiques obtenus par chevauchement des séquences communes (contigs). Chaque technologie de séquençage ayant son assembleur dédié, l'approche repose essentiellement sur les progrès récents de la bioinformatique. Cette méthode est couramment désignée sous le nom de shotgun (fusil à canon scié), ou encore Whole Genome Shotgun. Cette métaphore illustre le caractère aléatoire du séquençage de l'ADN génomique. En effet, cette stratégie est basée sur une distribution des séquences analysées suivant une loi de Poisson [46].

Les avantages de cette méthode sont la rapidité de la technique et un coût plus faible. L'inconvénient est que le traitement informatique ne permet pas d'aligner des fragments comportant des séquences répétées qui sont présentes en plus ou moins grand nombre selon les génomes. Enfin, elle nécessite une phase de finition. Cette phase consiste à déterminer *in silico* l'ordre des contigs obtenus pendant la phase aléatoire et à combler les brèches entre ces contigs. Sa durée et sa difficulté dépendent essentiellement de la qualité des librairies employées lors de la phase aléatoire, du nombre de séquences obtenues ramené à la taille totale du génome (ou couverture, nombre de fois où une position du génome est séquencée) mais essentiellement de la richesse du génome en éléments répétés et de leur localisation. Les éléments répétés en tandem sont à l'évidence de redoutables obstacles à l'obtention d'une séquence complète.

Pendant les premières années de l'ère de la génomique bactérienne, la méthode de séquençage de Sanger était la seule disponible. L'assemblage d'un génome était réalisé grâce à des logiciels d'assemblage de séquences tel que Phrap [47]. Cette stratégie a notamment été utilisée pour les premiers génomes séquencés dans l'équipe, celui de *F. psychrophilum* JIP 02/86 [6] et celui de *F. branchiophilum* FL15 [48].

Après 2005 et l'apparition des méthodes de séquençage NGS, l'explosion des solutions disponibles a compliqué la situation du choix de la stratégie à utiliser. Chaque méthode de séquençage dispose d'assembleurs qui lui sont mieux adaptés. De plus, en fonction de la complexité du génome, aucun ne permet d'assembler de manière correcte un génome complet *de novo*.

Lors d'une période de transition, des stratégies qualifiées de mixtes se sont mises en place. Elles ont été utilisées pour les génomes de *F. indicum* GPSTA100-9^T [49], de *F. psychrophilum* THC 02/90 (non publiée) qui ont été réalisé grâce à un mélange de séquences Sanger et 454, ou encore pour les génomes de *F. frigidimaris* KUC-1 et *F. glaciei* 0499^T (non publiées) réalisés avec un mélange de séquences 454 et Solexa (lectures uniques). Ces assemblages ont demandé l'utilisation de plusieurs assembleurs, situation peu idéale pour une utilisation fréquente.

La méthode de séquençage Solexa a rapidement évolué : la possibilité de faire du « paired-end », des lectures plus longues (deux fois 100 pb), un nombre de séquences produites plus important et le tout à des coûts toujours plus faibles ont permis par la suite de séquencer des génomes *de novo* en utilisant exclusivement cette technologie. Cela a permis d'augmenter la couverture et de réduire le nombre de régions non séquencées diminuant les trous dans l'assemblage de la séquence complète. Cette stratégie a par exemple été utilisée pour le séquençage des génomes complets de plusieurs souches de *F. psychrophilum* (non publiées) et celui des souches type du genre *Tenacibaculum* (non publiées), réalisés à des couvertures supérieures à 50X contre une moyenne d'environ 15-20X pour les précédents. Les assemblages de ces séquences ont donc été beaucoup plus rapides et ont permis d'augmenter la qualité générale des assemblages obtenus.

Enfin, depuis 2005, les méthodes NGS ont sans cesse progressé à une vitesse impressionnante en termes d'innovation, de qualité et de logiciels d'assemblages disponibles. Le montage d'un projet de séquençage d'un génome est un processus long qui demande du temps pour mobiliser des ressources à la fois financières, techniques, humaines et informatiques. Ce décalage entre d'une part, les innovations et progrès des NGS et, d'autre part, la réalité de la recherche, constitue également une difficulté dans le choix de la stratégie de séquençage et d'assemblage d'un génome.

Validation des assemblages de génomes

Nous essayons au laboratoire de produire des séquences de très haute qualité. L'assemblage de génomes complets destinés à être comparés nécessite une phase de vérification. Les critères de qualité d'une séquence génomique complète sont une couverture suffisante, un petit nombre de contigs et un minimum d'incertitude sur l'ordre de ces derniers. Les étapes de finition et de validation d'un projet de séquençage sont généralement longues et difficiles, dépendant à la fois de la méthode de séquençage utilisée, de la stratégie d'assemblage et de la richesse en éléments répétés des génomes. De plus, la construction des banques et librairies d'ADN génomique ou l'utilisation des programmes d'assemblage (et leur paramétrage) peuvent générer des erreurs préjudiciables à l'analyse ultérieure des séquences tels que des contigs chimériques et des décalages de phase de lecture artéfactuels. Il existe différentes techniques permettant la validation d'un assemblage au niveau « macroscopique » (scaffolding).

Une approche plus ancienne, utilisant le clonage, consiste à combiner un séquençage de plusieurs banques de fragments génomiques de tailles différentes. Il existe des banques d'inserts de taille moyenne (de 5 à 12 kpb) et des banques d'inserts de plus grande taille (50 à 100 kpb) clonés dans des vecteurs BACs (Bacterial Artificial Chromosome). Le séquençage et l'assemblage d'inserts de différentes tailles permettent d'organiser les contigs entre eux et de réduire l'incertitude de l'assemblage de certaines zones.

Une autre approche, par des cartes de restrictions obtenues par digestion de l'ADN génomique par une enzyme, peut-être utilisée pour valider un assemblage de génome. Une migration sur gel ou une lecture optique des produits obtenus permet d'obtenir un profil de restriction qui peut être comparé à un profil de restriction *in silico* du génome à valider.

Les cartes optiques sont des outils développés par la société OpGen Technologies (Madison, WI) depuis une dizaine d'années. Il s'agit de cartes de restriction ordonnées à haute résolution d'un génome dont l'obtention est grandement automatisée. Elle permet d'ordonner et d'orienter des contigs lors d'un assemblage de génome et constitue également une méthode de validation indépendante d'un projet de séquençage. Comparée à une carte de restriction *in silico* d'une séquence connue, ces cartes optiques peuvent aussi êtres utilisées en

génomique pour identifier des réarrangements génomiques tels que des insertions, des délétions, des duplications et des inversions [50]. Cette méthode peut être d'un grand intérêt pour la finition de génomes complets [51], [52], [53].

Dans certains cas, les approches et techniques de validation citées ci-dessus, les techniques classiques de biologie moléculaire (PCR) et d'analyses bioinformatiques (biais de composition en GC, assembleurs…) ne suffisent pas à valider l'organisation et la structure d'un chromosome bactérien. Les incertitudes peuvent alors être levées par la mise en place d'autres stratégies expérimentales. Ce fût notamment le cas pour les différents génomes de *F. psychrophilum* dont la stratégie de validation des structures chromosomiques mise en place est détaillé dans la première partie de ce manuscrit.

Annotation des génomes

Le séquençage d'un génome permet de prédire les fonctions de gènes contenus dans un organisme en annotant les gènes détectés dans une séquence d'ADN. L'annotation des génomes est réalisée à l'aide d'un pipeline d'annotation, AGMIAL [54], disponible sur la plateforme MIGALE de l'INRA. Cet outil d'annotation intègre la détection des gènes à partir de la séquence des bases nucléotidiques, la comparaison avec des génomes connus, un algorithme d'annotation automatique (permettant entre autres d'effectuer un choix, le plus rationnel possible, de la fonction d'un gène et une interface visuelle permettant une annotation manuelle. Cependant, cette annotation expertisée nécessite un travail manuel de vérification, long et fastidieux.

Génomique comparative et exploitation des génomes

L'intérêt pour le séquençage a été stimulé par un besoin de connaissances. En effet, séquencer le génome d'un micro-organisme permet d'avoir accès à la totalité de son information génétique. Cependant, cette connaissance a plus de valeur si elle est mise en perspective et comparée avec d'autres informations génomiques ou phénotypiques.

L'apport de la génomique comparative

Le séquençage et la comparaison de plusieurs génomes au sein d'une espèce ou d'un genre bactérien permet de définir le génome central (ou « core genome ») et le génome total (ou « pan-genome ») de l'espèce ou du genre. Le pan-génome décrit le nombre total de gènes retrouvés au moins une fois dans une espèce ou un genre [55]. En effet, les bactéries peuvent avoir de grandes variations de contenu en gènes entre souches proches. L'analyse de ces gènes permet donc de comprendre l'évolution d'un groupe bactérien, particulièrement pertinent dans l'étude des métagénomes. Cependant, elle peut aussi être utilisée dans un contexte génomique plus restreint [56]. Le pan-génome inclut le core génome contenant les gènes présents chez tous les organismes en question, le génome accessoire contenant les gènes présents dans deux ou plus des souches ou espèces, et les gènes uniques, spécifiques d'une seule souche ou espèce [55]. Le core génome inclut généralement tous les gènes responsables des aspects principaux de la physiologie et des caractères phénotypiques majeurs d'une espèce ou d'un genre. En revanche, les gènes accessoires et uniques contribuent à la diversité des souches ou des espèces et peuvent coder pour des voies métaboliques supplémentaires et des fonctions qui ne sont pas essentielles à la croissance mais qui peuvent conférer un avantage adaptatif à l'organisme, comme par exemple l'adaptation à une niche écologique, la résistance aux antibiotiques ou encore la colonisation d'un hôte particulier. De tels gènes sont généralement retrouvés dans les îlots génomiques [55].

L'analyse des gènes accessoires permet donc de comprendre les spécificités de chaque organisme séquencé. Remarquons ici que ces distinctions ne sont pas strictement biologiques car elles dépendent en partie des souches ou espèces incluses dans l'analyse ainsi que des paramètres utilisés pour les comparaisons et le regroupement des gènes.

Les études comparatives *in silico* effectuées sur les séquences des micro-organismes ont permis de progresser dans de nombreux domaines concernant notamment la structure et la diversité des génomes et en particulier l'identification de gènes de virulence et de déterminants de pathogénicité pour les espèces pathogènes [57], [58], [59]. Ces gènes peuvent êtres multiples et variés. Ce sont principalement les gènes codant pour des toxines, les systèmes d'export de macromolécules (ou systèmes de sécrétion) et les gènes codant pour des fonctions liées aux interactions avec l'hôte [tels que des gènes codant pour des fonctions d'adhésion, des gènes de synthèse des polysaccharides de surface et de molécules d'interaction avec le système immunitaire de l'hôte [6] et des systèmes de captation du fer (sidérophores)] qui sont recherchés comme déterminants de virulence.

Le contenu en gène d'un organisme reflète ses possibilités d'interactions avec son environnement. Par exemple, la présence de gènes codant les enzymes d'une voie catabolique donnée dans un génome indique que le micro-organisme est probablement capable d'utiliser telles ou telles ressources présentes dans son environnement. Egalement, comparer les contenus de plusieurs micro-organismes en gènes codant pour des systèmes d'import et d'export de macromolécules tels que les polysaccharides ou les polyamino-acides peut renseigner sur la capacité de ces micro-organismes à utiliser les ressources et nutriments présents dans leurs milieux de vie dans les conditions physico-chimiques nécessaires à leurs croissance. Le contenu en gènes d'un génome de micro-organisme reflète donc sa niche écologique.

La comparaison de données génomiques de plusieurs isolats au sein d'une espèce permet d'identifier des gènes ubiquitaires dans l'espèce. L'apport de génomes d'autres espèces proches peut permettre d'identifier alors les gènes spécifiques de cette espèce. Ces gènes spécifiques et ubiquitaires peuvent par exemple être utilisés comme cibles pour la mise au point de moyens de détection ou de vaccins.

Génomique fonctionnelle

Les approches de génomique fonctionnelle doivent permettre de valider le rôle et l'importance de gènes identifiés *in silico* dans un phénotype particulier ou dans la virulence pour les espèces pathogènes. Ces approches permettent de valoriser les projets de séquençage des génomes en assurant des liens solides entre la génomique bactérienne, la microbiologie et les relations hôtes/pathogènes. Les approches de génomique fonctionnelle nécessitent des changements d'échelle et sont souvent réalisées à l'aide de collaborations techniques avec d'autres équipes. Les stratégies employées dépendent de l'objet d'étude et restent un véritable point critique pour la mise en valeur du travail de génomique.

L'étude de l'expression des gènes par des approches transcriptomiques peut déterminer la part de gènes exprimés dans des conditions particulières. Ces analyses indiquent comment la bactérie se sert de son patrimoine génétique pour l'interaction avec son environnement. Par exemple, l'analyse du transcriptome de *F. psychrophilum*, actuellement initiée au laboratoire, permettra à terme de savoir quels gènes sont exprimés lorsque la bactérie est placée dans différentes conditions de cultures ou encore soumise à un stress (carence en nutriment, stress oxydatif, etc…).

L'analyse des protéines de surface d'un micro-organisme permet de décrire la partie externe de la bactérie qui est à l'interface de la cellule et de son environnement. Ces informations permettent de comprendre quelles protéines sont impliquées dans la relation hôte/pathogène et d'identifier les antigènes dominants et vraisemblablement reconnus par le système immunitaire de l'hôte pendant l'infection. L'analyse du « surfaceome » de *F. psychrophilum* a déjà été réalisée par des approches de purification d'extraits de membranes, de caractérisation en gels 2D et spectrométrie de masse [60], [61]. Une autre approche dite de « shaving », consistant à analyser les produits d'une digestion ménagée des protéines de surface par spectrométrie de masse, est également en cours d'utilisation.

L'obtention d'une séquence génomique permet également de mettre au point des outils génétiques pour des approches de mutagenèse. Ces approches, courantes en génétique, permettent d'appréhender la fonction d'un gène par son inactivation (insertion ou délétion) et peuvent être réalisées de manière aléatoire ou ciblée. Une collaboration avec une équipe

espagnole a permis de mettre en place une stratégie de mutagenèse de *F. psychrophilum* par transposition, à partir d'outils génétiques existants [62]. L'analyse *in silico* de la banque de mutants obtenue a permis de sélectionner les mutants dont la fonction du gène inactivé est probablement en relation avec la virulence (toxines, systèmes d'import du fer, régulateurs transcriptionels…). Les phénotypes de ces mutants peuvent être caractérisés *in vitro* par des tests de cytotoxicité cellulaire, des tests de mobilité et d'adhésion ou encore par mesure de leur capacité à produire un biofilm. Un modèle expérimental d'infection de la truite arc-en-ciel [63] permet de réaliser des épreuves infectieuses en conditions contrôlées et rend possible la caractérisation *in vivo* de la virulence des mutants. Ces travaux, actuellement en cours de réalisation, permettent d'espérer des applications directes concernant la relation hôte/pathogène avec par exemple l'obtention de mutants de virulence atténuée comme candidats vaccins. Dans cette même optique, une tentative de mutagenèse ciblée à été également initiée sur l'espèce marine *Tenacibaculum maritimum* à partir d'outils existants [64] et de la connaissance du génome. En cas de succès, les mutants obtenus pourront être testé sur un model d'infection sur poissons zèbres (*Danio renio*) habitués à l'eau saumâtre ($11 g.L^{-1}$ de sels marins).

Projet doctoral et objectifs de la thèse

Les membres de la famille des *Flavobacteriaceae* présentent une importante diversité de modes de vie et sont fortement impliqués dans les grands cycles de recyclage de la matière organique dans les écosystèmes terrestres et marins [65], [66]. Le genre *Flavobacterium* comprend actuellement trois espèces pathogènes pour les poissons que sont *F. columnare, F. branchiophilum et F. psychrophilum* ainsi que de nombreuses nouvelles espèces retrouvées dans des niches écologiques très diverses telles que les sédiments marins, les sols, les eaux douces, les glaciers, les milieux polaires et une source d'eau chaude. L'objectif des recherches est de contribuer à la connaissance scientifique en écologie microbienne et plus particulièrement sur les pathologies aquacoles. Dans cette visée, notre équipe développe des projets de génomique analytique et fonctionnelle appliquées au genre *Flavobacterium* et a mis en place, depuis quelques années, un programme de génomique comparative intra- et inter-espèces.

Lors de l'initiation du projet, les seuls génomes disponibles au sein du genre étaient ceux de *F. psychrophilum* JIP02/86 et *F. johnsoniae* UW101^{T}. L'analyse du génome de *F. psychrophilum* a permis de mettre en évidence des groupes de gènes en relation avec la colonisation, l'invasion et la destruction des tissus de l'hôte et l'analyse du génome de *F. johnsoniae* a permis de mettre en évidence des groupes de gènes en relation avec son mode vie environnemental (par opposition au mode de vie pathogène). Ainsi pour compléter les informations sur la dynamique évolutive au sein du genre *Flavobacterium*, un projet de séquençage et de comparaison des génomes complets de plusieurs bactéries pathogènes ainsi que de bactéries environnementales a été mis en place. Ce projet de génomique comparative incluant les génomes des espèces pathogènes *F. psychrophilum*, *F. columnare* (disponible depuis) et *F. branchiophilum* et des espèces environnementales *F. johnsoniae* et *F. indicum*, doit permettre d'identifier les traits génétiques associés à la pathogénicité et la virulence chez les espèces pathogènes des poissons. De plus, ces génomes d'espèces aux origines géographiques et niches écologiques différentes permettra de comprendre comment ces bactéries ont évolué et se sont adaptées à des niches écologiques aussi diverses et parfois extrêmes.

Le génome de la bactérie pathogène pour les salmonidés *F. psychrophilum* JIP 02/86, organisme modèle du laboratoire, a été séquencé en 2007 [6]. Le génome d'une autre souche phylogénétiquement éloignée [7], *F. psychrophilum* THC 02/90 a été séquencé en 2009 (non publié) afin de comparer la diversité génomique au sein de l'espèce. En 2010, le séquençage de dix nouvelles souches de *F. psychrophilum* provenant de poissons hôtes différents et d'origines géographiques variées a permis d'étendre cette étude. Comparée à celle de la souche JIP 02/86, l'organisation du chromosome de la souche THC 02/90 est différente et la validation de l'assemblage du génome de cette souche permettra d'apprécier la diversité des structures chromosomiques au sein de l'espèce *F. psychrophilum*. En termes d'applications, l'étude des génomes au sein de l'espèce *F. psychrophilum* permettra d'identifier *in silico* des cibles moléculaires prometteuses pour le développement d'un test diagnostic spécifique de l'espèce ainsi que des cibles vaccinales potentielles.

F. branchiophilum montre une différence dans la pathologie et les symptômes provoqués chez les poissons par rapport aux deux autres espèces pathogènes suggérant ainsi des mécanismes moléculaires de virulence différents. Le génome complet a donc été séquencé afin de permettre d'appréhender ces mécanismes de virulence. La comparaison avec les autres génomes permettra d'analyser l'évolution des déterminants de virulence au sein du genre, de préciser le génome central du genre *Flavobacterium* et d'essayer de comprendre les mécanismes d'adaptation de cette bactérie à sa niche écologique.

La diversité retrouvée au sein du genre *Flavobacterium* est un excellent contexte pour une étude de génomique comparative entre des organismes relativement proches aux modes de vie différents. *F. indicum* est une bactérie non pathogène considérée comme environnementale. Elle a été isolée d'une source chaude et est donc l'unique espèce thermophile connue du genre. Ces nouvelles données permettront notamment la caractérisation d'éléments moléculaires marqueurs de caractères phénotypiques et de comprendre les mécanismes d'adaptation de cette bactérie à sa niche écologique.

Pour réaliser des comparaisons inter-espèces au sein du genre *Flavobacterium*, les génomes de l'espèce pathogène *F. branchiophilum* et de l'espèce environnementale *F. indicum* ont donc été séquencés. Le séquençage et l'analyse de ces génomes sont respectivement détaillés dans la deuxième et la troisième partie de ce manuscrit.

L'objectif de cette thèse est de permettre la caractérisation d'éléments moléculaires marqueurs de caractères phénotypiques au sein d'organismes du genre *Flavobacterium*. Au cours de différents projets de séquençage, nous avons identifié un groupe extrêmement rare de gènes responsables d'une modification de la structure de la molécule d'ADN et dont le rôle fonctionnel dans la physiologie bactérienne n'est pas encore complètement élucidé aujourd'hui. Ces gènes *dnd* n'avaient jusqu'ici jamais été décrits dans la famille des *Flavobacteriaceae* et la conduite d'une étude originale, présentée dans la quatrième partie de ce manuscrit, nous a permis de décrire une nouvelle organisation de ces locus au sein des membres du phylum *Bacteroidetes*.

Ce projet repose donc sur des approches de microbiologie et de génomique bactérienne, se situant au carrefour de l'écologie microbienne, de l'évolution moléculaire et de la génétique moléculaire avec de fortes implications en santé animale pour les espèces pathogènes.

Première partie : Validation de l'assemblage des génomes complets de *F. psychrophilum* souches JIP 02/86 et THC 02/90. Diversité des structures chromosomiques au sein de l'espèce.

Introduction

Le séquençage du génome complet de *F. psychrophilum* THC 02/90 a été effectué par l'INRA peu avant mon arrivée au laboratoire. Phylogénétiquement éloignée de la souche JIP 02/86 (données MLST [7]), elle a été séquencée pour comparer la diversité génomique au sein de l'espèce et servir de support à des approches de mutagenèse. Les génomes des deux souches de *F. psychrophilum* JIP 02/86 et THC 02/90 ont été séquencés par la méthode shotgun incluant une banque de grands fragments dans le vecteur pCNS.

L'assemblage du génome de la souche THC 02/90 présente une large inversion chromosomique par rapport au génome de *F. psychrophilum* JIP 02/86. Cette inversion est située entre deux régions inversées répétées, particulièrement complexes, rigoureusement identiques, distantes de 0,8 Mpb et centrées sur l'origine de réplication ; elle ne conduit pas à une altération de leurs « GC skew » (Figures 1 et 2). Ces répétitions avec une bonne couverture de liens de clones donnent donc deux assemblages différents entre les deux souches de la même espèce. Compte tenu de cette bonne couverture de séquence dans ces zones « charnières », on peut se demander si les assemblages des génomes de *F. psychrophilum* JIP 02/86 et THC 02/90 sont exacts.

Cependant, la structure (organisation et orientation de segments chromosomiques) d'un chromosome bactérien n'est pas obligatoirement figée. Différents isoformes chromosomiques peuvent coexister dans la même préparation d'ADN génomique, suggérant des réarrangements génomiques durant la croissance de l'organisme. Notamment décrits chez *Yersinia pestis* [67], ces réarrangements chromosomiques semblent fréquents *in vivo* chez cette bactérie bien que leurs effets sur la biologie et la pathogénicité des organismes soient inconnus. De manière étonnante, pour la souche JIP 02/86, nous avons pu trouver deux clones indépendants et discordants qui suggèrent que les deux orientations du chromosome coexistent au sein de la même préparation d'ADN génomique. De plus, au sein de l'espèce *F. psychrophilum*, un taux de recombinaison très important a été observé à la suite d'une étude de diversité par MLST [7] et une autre étude, plus large, classe *F. psychrophilum* comme la bactérie ayant le plus haut taux de recombinaison homologue parmi toutes les espèces

analysées [68]. En considérant que les larges zones répétées inversées sont un substrat parfait pour la recombinaison, on peut donc se demander si on peut retrouver chez *F. psychrophilum* la présence des isoformes chromosomiques, tels ceux décrits chez *Y. pestis*.

L'étude des isoformes chromosomiques chez *Y. pestis* [67] a été réalisée par PCR combinatoires et a montré de nombreux événements de recombinaison au niveau de petites séquences d'insertion (IS) présentes en plusieurs copies. Chacune des répétitions présentes dans les génomes de *F. psychrophilum* (JIP 02/86 et THC 02/90) sont des répétitions directes et engagent également des gènes d'adhésines, séquences composées de n blocs répétés en tandem (Figure 3). Une approche par PCR nécessite donc une amplification d'environ 10 kpb. Plusieurs tentatives de PCR en conditions « long range » ont été réalisées sans permettre de conclure quant à l'organisation de ces régions.

Une approche expérimentale combinant la digestion des ADN génomiques dans des bouchons d'agarose, une migration en champ pulsé (PFGE) et l'hybridation de sondes marquées sur l'ADN immobilisé sur membrane (Southern-blot) a été réalisée pour valider les assemblages des deux génomes. Cette approche devrait également permettre de visualiser des isoformes chromosomiques minoritaires.

En 2010, les séquençages Solexa des génomes complets de dix nouvelles souches de *F. psychrophilum* ont été entrepris afin de comparer la diversité génomique au sein de l'espèce. Ces souches phylogénétiquement différentes [7], provenants de poissons hôtes et d'origines géographiques variés, ont été choisies afin de maximiser cette diversité. Les principales caractéristiques des douze souches utilisées dans cette étude sont présentées dans le Tableau 1. L'élargissement de l'investigation des structures chromosomiques au sein de l'espèce *F. psychrophilum* devait permettre également de déterminer quelle organisation du chromosome prédomine dans cette espèce.

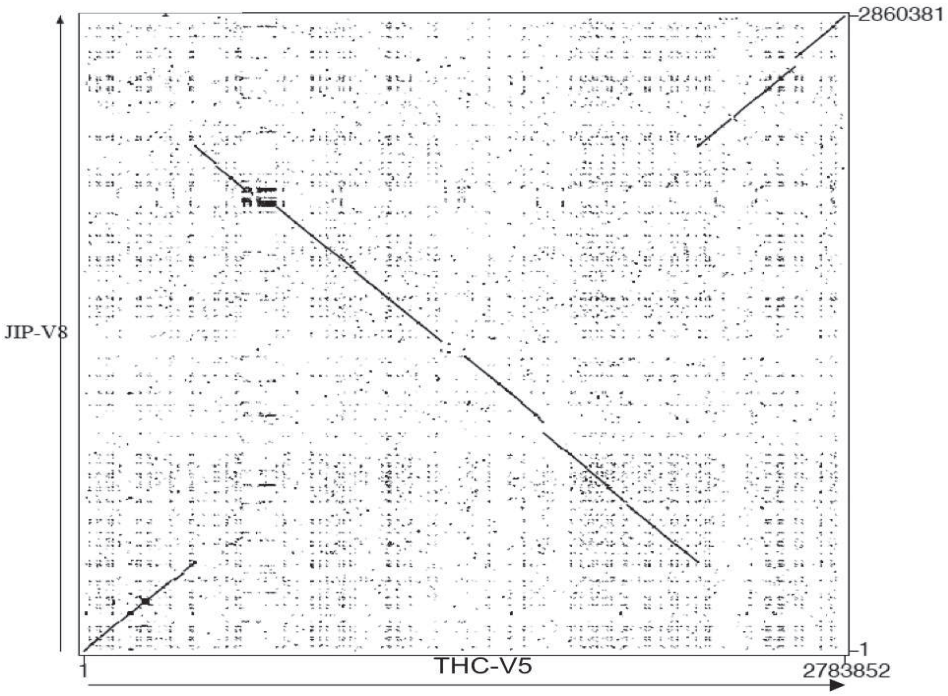


Figure 1: Graphique à points d'identité nucléotidique montrant l'inversion d'une partie du génome entre les souches de *F. psychrophilum* THC 02/90 (en abscisse) et JIP 02/86 (en ordonnée).

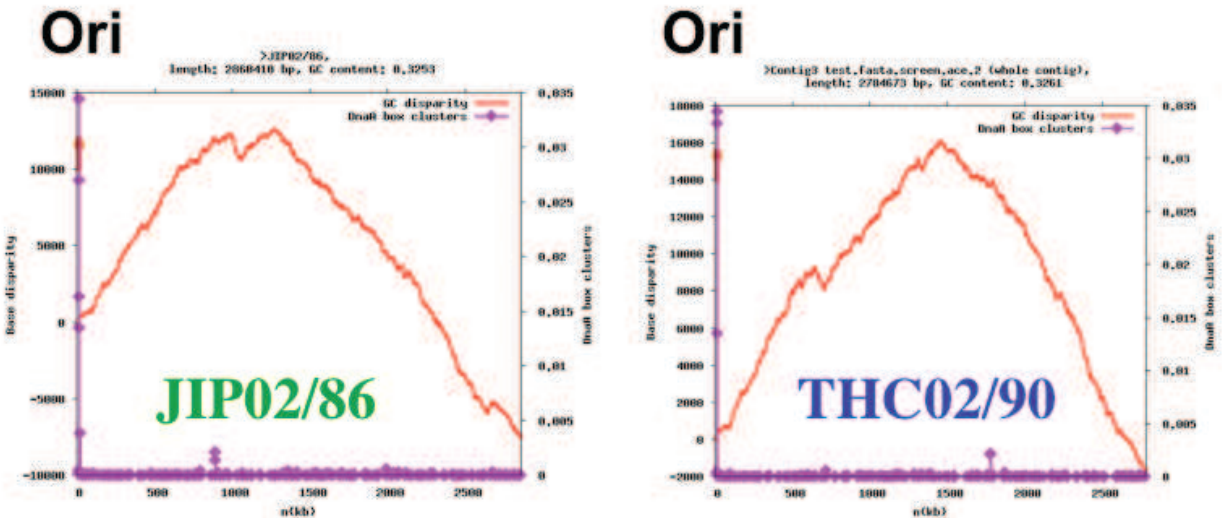
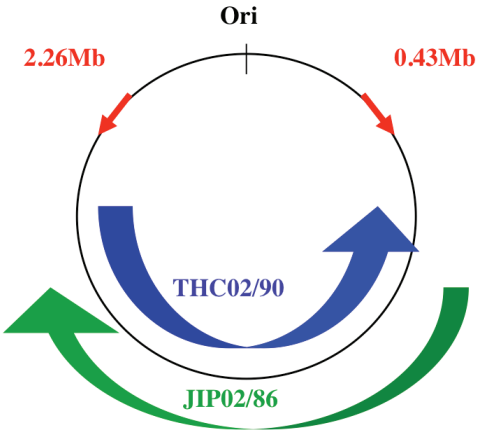


Figure 2: L'inversion d'un segment du chromosome entre les souches *F. psychrophilum* JIP 02/86 (à gauche) et THC 02/90 (à droite) est centrée sur l'origine de réplication et ne provoque pas de modification dans le biais de composition en GC (GC skew) entre les génomes. « Ori » en haut de l'axe des ordonnées indique l'origine de réplication.

A



B

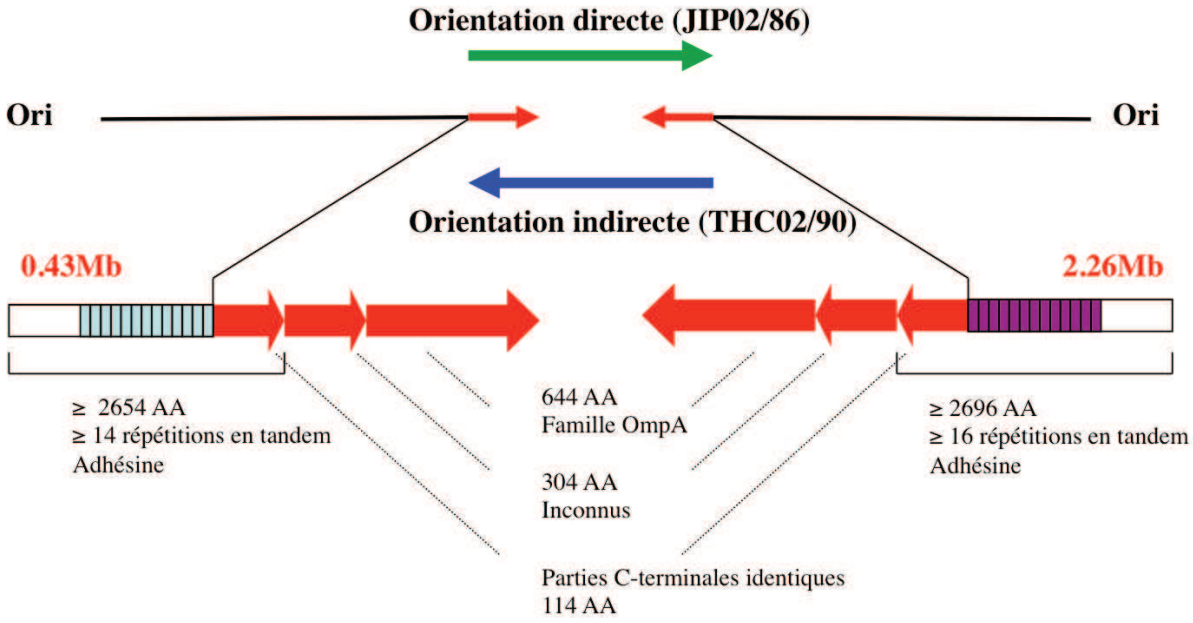


Figure 3: A. Schéma représentant l'inversion de 0,8Mpb centrée sur l'origine de réplication prédite (Ori) entre les génomes de *F. psychrophilum* JIP 02/86 (en vert) et THC 02/90 (en bleue). Les flèches rouges indiquent les deux grandes régions où l'inversion a lieu.
B. Schéma représentant la composition des régions inversées répétées. Les flèches indiquent les phases de lecture prédites. Abréviations : AA= acides aminés ; OmpA= outer membrane protein A [61].

Nom	ST/CC [7]	Origine géographique	Poisson hôte/tissu	Année d'isolement	Méthode séquençage
JIP 02/86	ST20 / CC1	EU / France	RbT/ rein	1986	Sanger Solexa
THC 02/90	ST9 / CC2	USA / Oregon	CoS / rein	1990	Sanger 454 Solexa
FPC 831	ST30	Japon	CoS / lésion pédoncule	1990	Solexa
FPC 840	ST5	Japon	Ayu / rein	1987	Solexa
JIP16-00	ST4	EU / France	RbT	2000	Solexa
JIP 08-99	ST2 / CC1	EU / France	RbT/ rein	1999	Solexa
LVDJ XP 189	ST18	EU / France	Ten / rein	1992	Solexa
NCIMB 1947^{T}	ST13 / CC2	USA / Washington	CoS / rein	?	Solexa
KU051128-10	ST66	Japon/Shiga	Eau	2005	Solexa
KU060626 -4	ST49	Japon/Shiga	Ayu/rein	2006	Solexa
KU060626-59	ST48	Japon/Shiga	Ayu/rein	2006	Solexa
KU061128-1	ST62	Japon/Shiga	Eau	2006	Solexa

Tableau 1: Tableau récapitulatif des différentes caractéristiques des souches de *F. psychrophilum* séquencées. Les relations phylogénétiques (ST/complexe clonal), les origines géographiques, les poissons hôtes et leurs tissus et l'année d'isolement des souches sont variés afin d'élargir la diversité des échantillons au sein de l'espèce. Les espèces de poissons hôtes sont RbT : truite arc-en-ciel (*Oncorhynchus mykiss*), CoS : saumon coho (*Oncorhynchus kisutch*), ayu (*Plecoglossus altivelis*) et Ten :tanche (*Tinca tinca*).

Matériels et Méthodes

1) Stratégie expérimentale : choix des enzymes de restriction, choix et obtention des sondes

Pour répondre aux questions posées, un premier travail a permis de définir la stratégie expérimentale à mettre en place. Des enzymes de restrictions ne coupant pas dans les régions répétées mais suffisamment proche d'elles pour une résolution par migration sur gel et permettant également une hybridation utilisant des sondes marquées distales ont été sélectionnées *in silico*. Nous nous sommes également assurés que les différents profils de restriction et d'hybridation attendus étaient bien différents pour les deux génomes. Les enzymes *Nco*I et *Xho*I ont été choisies pour permettre de résoudre par Southern-blot les quatre zones en question (deux zones du génome de la souche THC 02/90 et deux zones du génome de la souche JIP 02/86). Les oligonucléotides permettant la synthèse des sondes ont été obtenus grâce à une option du logiciel Consed permettant la visualisation et l'édition de l'assemblage des génomes. Ils ont été dessinés sur le génome de *F. psychrophilum* JIP 02/86 (séquences en Annexe 1). Les séquences correspondant aux sondes voulues ont été ensuite extraites et testées par blastn sur le génome de la souche THC 02/90 pour s'assurer qu'elles s'hybrideraient bien aux seuls endroits voulus et qu'elles présentaient une identité suffisante avec la séquence de cette souche (P1 : 99% ; P2 : 96% ; P3 : 98% et P4 :96% d'identité nucléotidique). Les sondes ont ensuite été obtenues par PCR à partir d'une préparation d'ADN génomique de la souche JIP 02/86. Une carte des régions, la localisation des sondes et le profil d'hybridation des sondes attendu sont présentés en Figure 4A.

2) ADN génomiques en bouchons d'agarose et électrophorèse en champ pulsé

Une culture de 15 mL de chacune des souches THC 02/90 et JIP 02/86 sont réalisées en milieu AOBE à 140 rpm pendant 48h à 18°C. La croissance des bactéries est arrêtée en phase exponentielle dans la glace fondante à 0°C puis culottées par centrifugation 30 min à 400 g et lavées trois fois dans du PBS. Un volume équivalent d'agarose « Low-melting » à 1,5% (Seaplaque, Lonza) fondu est ajouté aux suspensions bactériennes et le mélange est distribué dans des moules de 100 µL. Après solidification à 4°C, les bouchons à 0,75% d'agarose final sont démoulés et incubés à 50°C pendant la nuit sous agitation douce dans du tampon NDS préchauffé à 50°C auquel la protéinase K est ajoutée à 1mg.mL^{-1} final. Cette étape permet de lyser les bactéries. Les bouchons sont ensuite rincés 3 fois 20 min avec du tampon TE puis 2 fois 30 min avec du tampon TE supplémenté en inhibiteurs de protéases à 1X final. Les bouchons sont lavés une nouvelle fois dans le tampon TE à température ambiante (composition des tampons en Annexe 1).

L'ADN des bouchons d'agarose est digéré par les enzymes de restriction *Xho*I et *Nco*I (Fermentas, 30 U final) pendant la nuit sous agitation douce à 37°C dans le tampon adéquat + BSA 1X final. Un bouchon d'agarose de chaque souche est traité selon les mêmes conditions mais sans enzyme pour réaliser un témoin de non-digéré. Le lendemain, les bouchons d'agarose sont lavés dans 500 µL de TE pendant 30 min sous agitation douce et stockés à 4°C.

La migration en champ pulsé (PFGE) est réalisée dans un gel 1% d'agarose low melting (SeaKem, Lonza) dans la cuve CHEF-DR III System (Bio-Rad), réfrigérée à 14°C. Les marqueurs de poids moléculaire utilisés sont le Midrange II et le Yeast Chromosome PFG Marker (New England Biolabs). Les paramètres de migration choisit sont les suivants : impulsions initiales de 1 s, impulsion finales de 3s, durée totale de migration de 18h, 6 Volts.cm^{-1}, angle entre les directions des courants de 120° et courant de 105 mA. Le gel est incubé environ une heure dans un bain de bromure d'éthidium (BET) pour coloration, rincé et analysé aux UV.

3) Transfert sur membrane et hybridation

Après migration, le gel est trempé dans une solution d'HCl à 0,7 % pendant 20 min pour éliminer quelques bases puriques et fragmenter l'ADN. Après rinçage à l'eau, le gel est trempé 10 min dans une solution de NaOH à 0,4 N afin de fragiliser les liaisons phosphodiesters et de séparer les brins d'ADN. Par capillarité, l'ADN du gel est transféré sur une membrane de nylon (HYBOND N+, Amersham Science) pendant 6 heures à température ambiante. L'ADN est ensuite fixé à la membrane 2 min aux UV et cette dernière est conservée dans du tampon SSC 2X à 4°C.

Le marquage des sondes est effectué par « Random Priming » et incorporation de dUTP fixé par covalence à un haptène stéroïdien par la polymérase de Klenow, conformément aux instructions du kit DIG High Prime DNA Labeling and Detection Starter kit II (Roche Diagnostics). L'efficacité de marquage des sondes est estimée par comparaison et révélation d'une gamme de dilution des produits de la réaction de marquage.

La membrane est pré-hybridée, hybridée puis révélée avec le substrat chimio-luminescent CSPD, conformément aux instructions du kit DIG High Prime DNA Labeling and Detection Starter kit II. L'hybridation s'effectue pendant une nuit à 40°C pour la sonde P1, 44°C pour la sonde P2 et 42°C pour les sondes P3 et P4.

L'élimination de l'excès de sonde non fixée est réalisée par deux lavages de 10 min à température ambiante dans un bain de SSC 2X/SDS 0,1% et la sonde fixée de manière non spécifique est éliminée par deux lavages de la membrane de 15 minutes à 58°C dans un bain de SSC 0,1X/SDS 0,1% (composition des tampons en Annexe 1).

La membrane est exposée sur un film radiographique dans une cassette entre deux écrans pendant 20-30 min à température ambiante. Le film est révélé en chambre noire par une succession de bains (révélateur, bain d'arrêt, fixateur et rinçage à l'eau).

4) Analyse par PFGE de la diversité des structures chromosomiques par digestion avec l'enzyme *Ceu*I

Une culture de chacune des souches (Tableau 2) est traitée selon les mêmes conditions décrites au paragraphe 2 mais l'ADN immobilisé dans les bouchons d'agarose est digéré par l'enzyme de restriction *Ceu*I (New Englands Biolabs, 30 U final) pendant la nuit sous agitation douce à 37°C dans le tampon adéquat. L'enzyme *Ceu*I a été choisie pour permettre de résoudre les structures « macroscopiques » des génomes ; en effet cette enzyme coupe spécifiquement l'ADN dans les copies de rDNA [69] permettant ainsi de générer des fragments de restriction de plusieurs centaines de kpb.

La migration en champ pulsé (PFGE) est réalisée comme précédemment décrit au paragraphe 2. Le marqueur de poids moléculaire utilisés est l'ADN génomique de *Salmonella* Braenderup digéré par l'enzyme *Xba*I. Les paramètres de migration choisis sont les suivants : impulsions augmentant progressivement de 50 à 70 s pendant 17h suivies d'impulsions augmentant progressivement de 3 à 12 s pendant 6h, 6 Volts.cm^{-1}, angle entre les directions des courants de 120° et courant de 105 mA. Le gel est incubé environ une heure dans un bain de BET pour l'analyse aux UV.

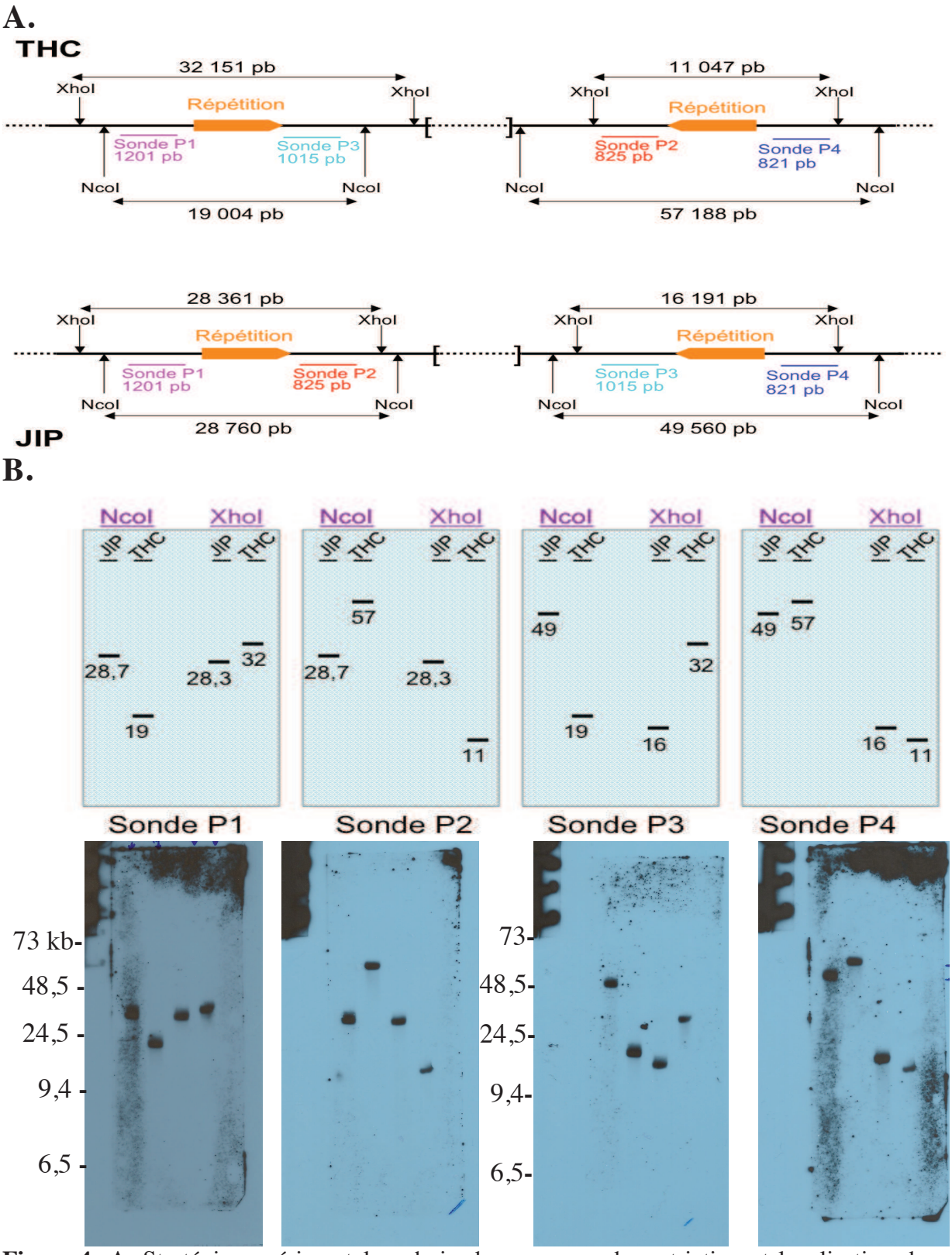


Figure 4: A. Stratégie expérimentale : choix des enzymes de restriction et localisation des sondes. **B.** Prédiction des profils d'hybridation attendus pour les deux souches (JIP 02/86 et THC 02/90) pour chaque sonde (P1 à P4) et chaque enzyme (en violet, en haut) et validation expérimentale (en bas). Les nombres sous les tirets représentent les poids moléculaires (en kpb) des signaux d'hybridation attendus.

Résultats

Validation de l'assemblage des génomes de *F. psychrophilum* souches JIP 02/86 et THC 02/90

Les profils d'hybridation obtenus par Southern-blot pour chaque sonde et chaque enzyme confirment les prédictions réalisées *in silico* sur les assemblages des génomes (Figure 4). L'orientation du chromosome dans les deux souches JIP 02/86 et THC 02/90 est donc vraisemblablement différente comme attendu par comparaison des assemblages. La méthode et l'assemblage de la souche THC 02/90 ont donc pu être validés. Une nouvelle version de l'assemblage du génome de la souche JIP 02/86 a pu être déposée dans la base de donnée du NCBI avec un commentaire stipulant que l'organisation du génome avait été vérifiée par Southern-blot (numéro d'accès : NC_009613.3).

Diversité des structures chromosomiques au sein de l'espèce *F. psychrophilum*

Les profils de restriction des douze souches de *F. psychrophilum,* obtenus par PFGE après digestion avec l'enzyme de restriction *Ceu*I, montrent une grande variété. En effet, tous sont différents des profils de restriction obtenus pour les souches JIP 02/86 et THC 02/90 (Figure 5). Cependant, les profils de restriction des pistes 3 et 7 correspondant respectivement aux souches KU061128-1 et NCIMB 1947^T sont identiques. Il en est de même pour les profils de restriction des pistes 6 et 13 correspondant respectivement aux souches FPC 840 et KU060626-4 (Figure 5).

Les profils d'hybridation obtenus par Southern-blot pour les douze souches de *F. psychrophilum* avec les sondes P3, P2 et P1 présentent également une grande variation. Selon la stratégie expérimentale définie sur les structures des génomes des souches JIP 02/86 et THC 02/90, trois sondes suffisent à révéler l'inversion d'une partie du génome. En effet, pour une organisation du génome « type THC 02/90 » les sondes P1 et P3 vont s'hybrider sur le même fragment de restriction alors que la sonde P2 s'hybride à un autre fragment de taille différente. Inversement, pour une organisation du génome « type JIP 02/86 » les sondes P1 et P2 vont s'hybrider au même fragment de restriction alors que la sonde P3 s'hybride à un autre fragment (Figure 4A).

Dans les pistes correspondant aux souches THC 02/90, JIP 16-00, JIP 08-99, LVDJ XP 189, FPC 831, FPC 840, KU061128-1 et KU060626-4 les deux sondes P1 et P3 s'hybrident, pour chaque souche, à des fragments de restriction de tailles identiques. Dans ces mêmes pistes, les signaux d'hybridation obtenus pour la sonde P2 ont un poids moléculaire différent de ceux obtenus avec les sondes P1 et P3. Ces résultats montrent que l'organisation du génome est identique au sein de ces souches. Pour la souche JIP 02/86, les sondes P1 et P2 s'hybrident à un fragment de 28,3 kpb et la sonde P3 à un fragment de 16 kpb, confirmant que l'organisation du génome est différente dans cette souche (Figure 6).

Les souches FPC 840 et THC 02/90 présentent les mêmes profils d'hybridation pour les trois sondes. Une variation dans la taille des fragments d'ADN hybridés par une même sonde est cependant observée entre toutes les autres souches. En effet, les fragments révélés par la sonde P1 présentent une variation de plusieurs kpb entres les souches. Ces résultats sont également observés pour les sondes P2 et P3 (Figure 6).

La mauvaise qualité des profils d'hybridation obtenus par Southern-blot avec la sonde P1 pour les souches NCIMB 1947^{T} et KU051128-10 ne permet pas de conclure quant à l'organisation particulière de ces dernières. Une dégradation de l'ADN génomique est observée dans les pistes correspondant à la souche KU060626-59 (Figures 5 et 6) mais nous reviendrons sur ce point dans la quatrième partie de ce manuscrit.

Présence d'isoformes chromosomiques dans les cultures de *F. psychrophilum*

Un signal d'hybridation de faible intensité est observé pour la sonde P3 dans la piste correspondant à la souche JIP 02/86. La taille du fragment de restriction ainsi révélé correspond aux prédictions réalisées *in silico* en cas de présence d'isoformes chromosomiques minoritaires (Figures 7A et 7B). Cependant, ce signal d'hybridation de faible intensité n'a pas été observé avec les autres sondes et pour les autres souches. En revanche, des bandes discrètes sont observées dans le profil de restriction par l'enzyme *Ceu*I de l'ADN de la souche JIP 02/86 (Figure 7C) ; elles correspondent également aux prédictions réalisées *in silico* en cas de présence d'isoformes chromosomiques minoritaires (non montrées).

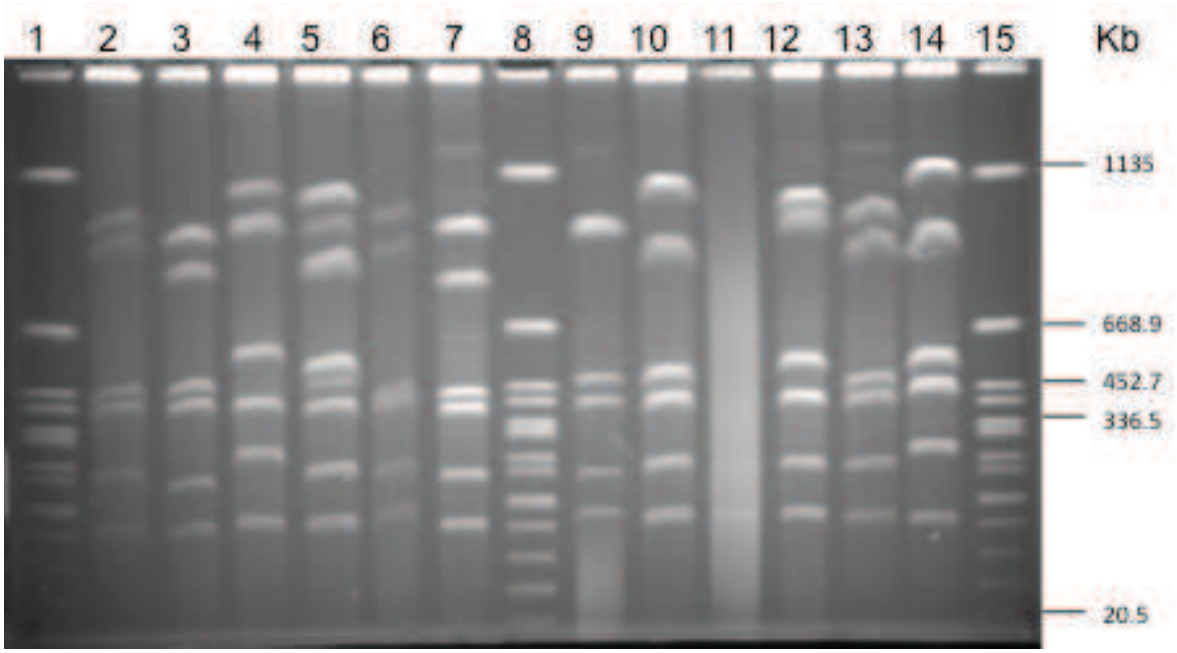
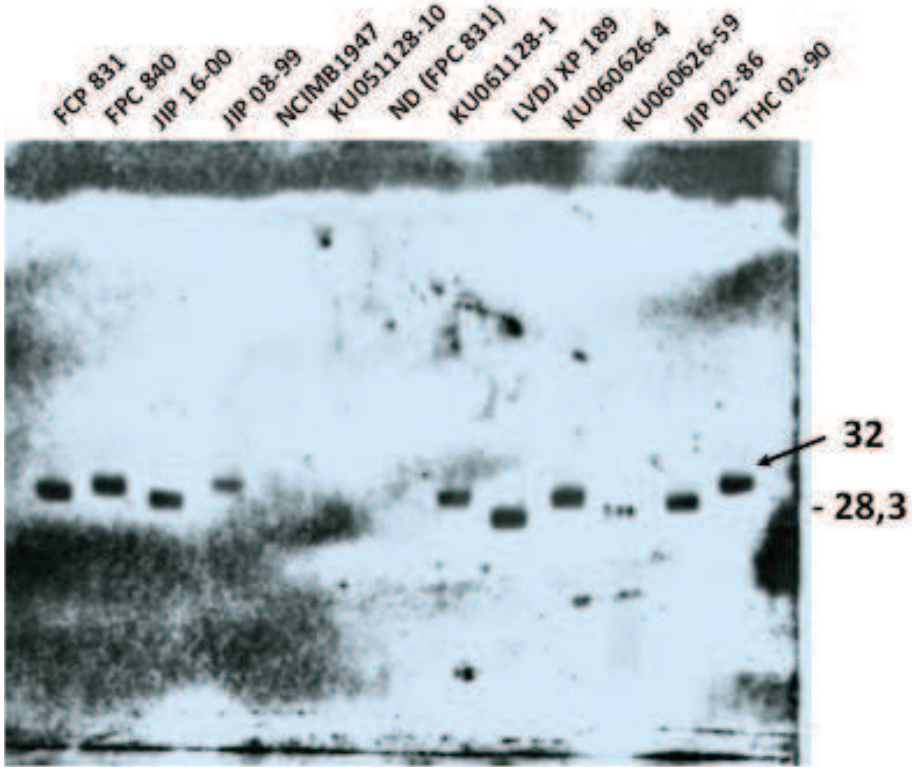
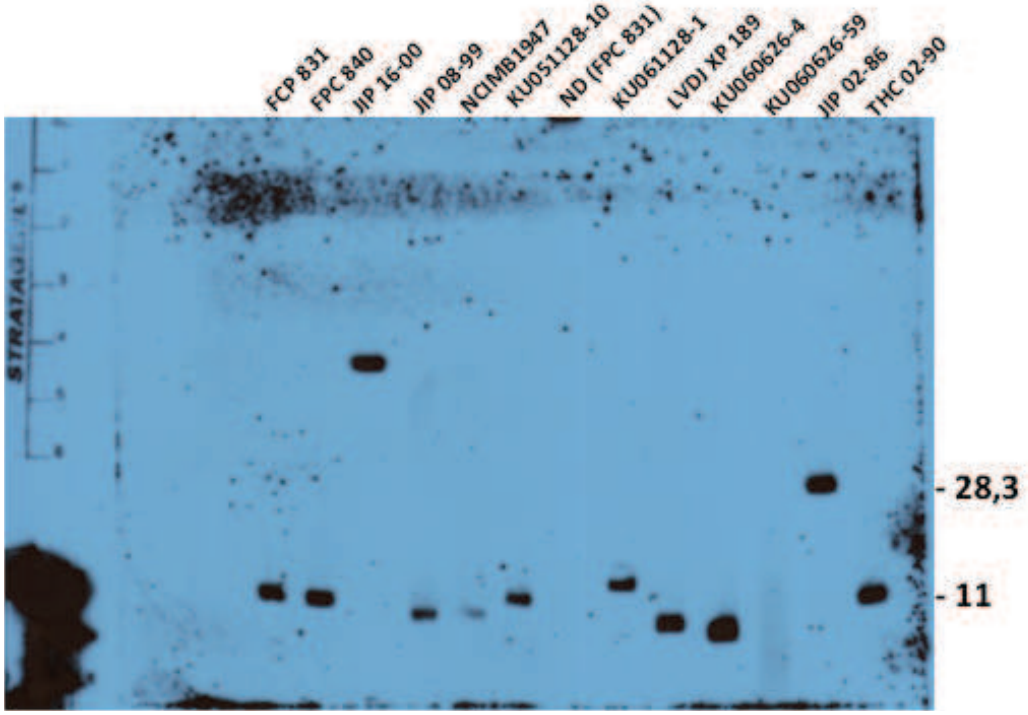


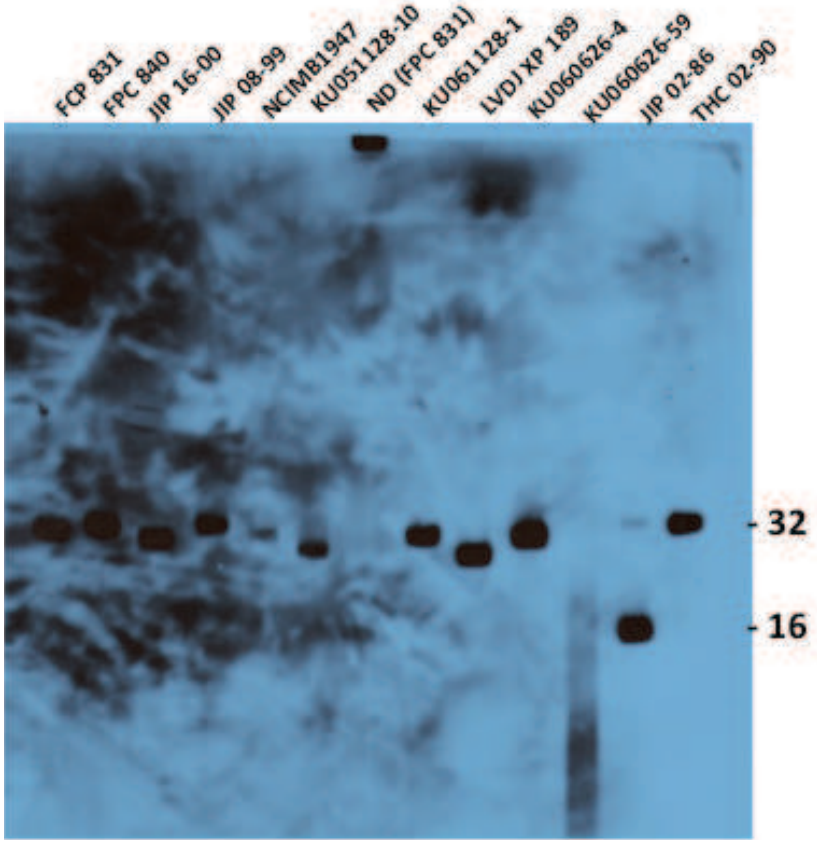
Figure 5: Analyse par PFGE des profils de restriction des douze souches de *Flavobacterium psychrophilum* après digestion de l'ADN génomique avec l'enzyme de restriction *Ceu*I. Les pistes suivantes correspondent respectivement aux souches : **2/** THC 02/90 ; **3/** KU061128-1 ; **4/** LVDJ XP189 ; **5/** JIP 02/86 ; **6/** FPC 840 ; **7/** NCIMB 1947^{T} ; **9/** KU051128-10 ; **10/** JIP 08-99 ; **11/** KU060626-59 ; **12/** FPC 831 ; **13/** KU060626-4 ; **14/** JIP 16-00.
Pistes 1, 8 et 15: Marqueurs de poids moléculaire ; les nombres sur la droite indiquent la taille des fragments en kpb.



A



B



C

Figure 6 : Résultats de Southern-blot obtenus pour les douze souches de *F. psychrophilum* après digestion de l'ADN génomique avec l'enzyme de restriction *Xho*I et hybridation avec les sondes P1 (A), P2 (B) et P3 (C). Le nom des souches est reporté en haut des pistes (ND= témoin de non digestion). Les nombres sur la droite indiquent les poids moléculaires (en kpb) des signaux d'hybridation.

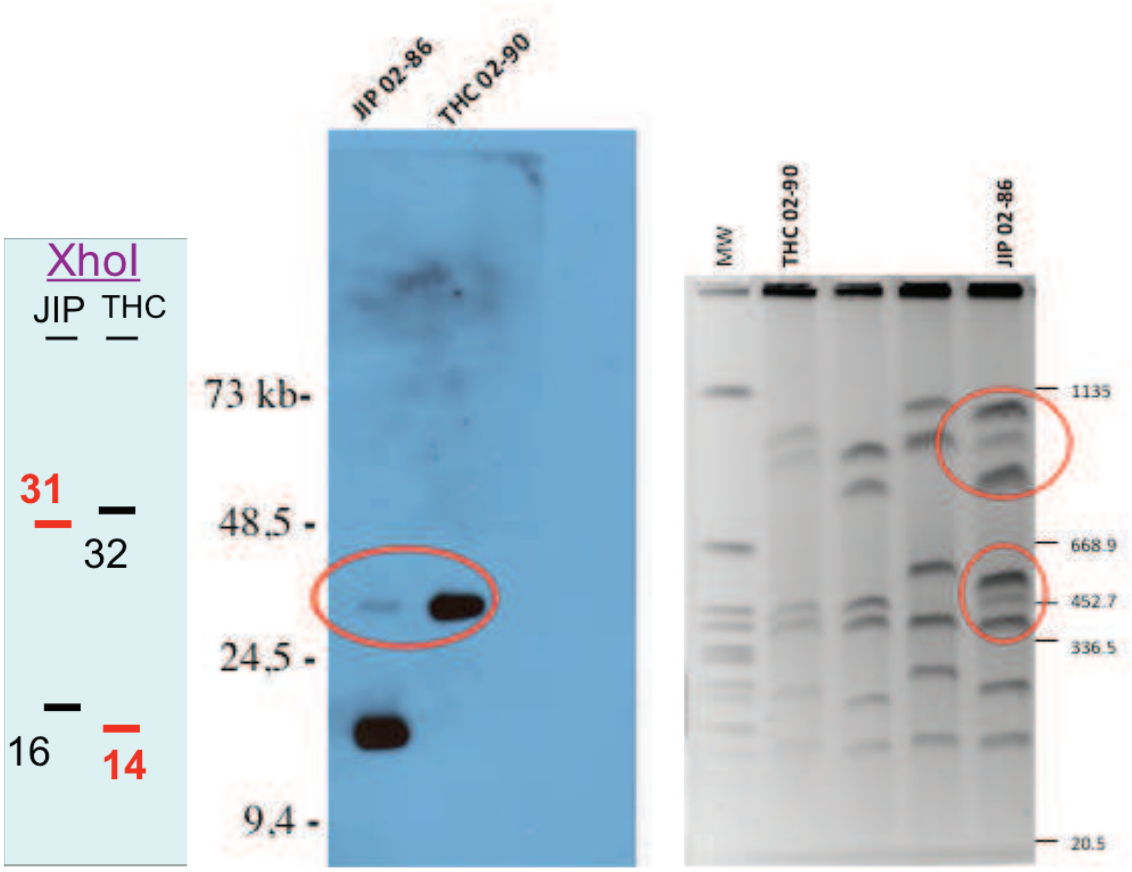


Figure 7: Présence probable d'isoformes chromosomiques dans les cultures de *F. psychrophilum* souche JIP 02/86. Les cercles rouges indiquent les indices de cette présence. **A :** Prédiction des profils d'hybridation attendus avec la sonde P3 et l'enzyme *Xho*I pour les deux souches (JIP 02/86 et THC 02/90) en cas de présence d'isoformes chromosomiques. Les prédictions ont été réalisées pour les formes majoritaires correspondant aux assemblages (en noir) et pour les formes minoritaires en cas d'inversion d'une partie du chromosome (en rouge). Les nombres aux les tirets représentent les poids moléculaires (en kpb) des signaux d'hybridation attendus. **B :** Résultats d'hybridation obtenus pour les souches JIP 02/86 et THC 02/90 avec la sonde P3 et l'enzyme *Xho*I. **C :** Analyse par PFGE des profils de restriction de souches de *Flavobacterium psychrophilum* après digestion de l'ADN génomique avec l'enzyme de restriction *Ceu*I (extrait de la Figure 5). Piste MW: Marqueur de poids moléculaire ; les nombres sur la droite indiquent la taille des fragments en kpb.

Discussion

Validation des assemblages des génomes de *Flavobacterium psychrophilum* souches JIP02/86 et THC 02/90

Les expériences de migration en champ pulsé ont permis de valider les assemblages et les structures chromosomiques (du moins majoritaires) des génomes de *F. psychrophilum* souches JIP02/86 et THC 02/90. L'orientation du chromosome dans les deux souches est donc vraisemblablement différente comme prédit à partir de la comparaison des assemblages ; ce qui permet de valider le travail et la méthode qui ont permis l'assemblage des génomes.

La comparaison de génomes complets a permis d'identifier une structure originale des génomes au sein de la même espèce. Cette inversion symétrique autour de l'origine de réplication a déjà été observée entre différentes espèces suggérant une caractéristique répandue chez les génomes bactériens et qui pourrait avoir un rôle important dans l'évolution des espèces [70].

Diversité des structures chromosomiques au sein de l'espèce *F. psychrophilum*

La variation observée entres les profils de restriction des douze souches de l'espèce *F. psychrophilum* obtenus par digestion avec l'enzyme *Ceu*I (Figure 5) montre une grande diversité de profils chromosomiques entre les souches analysées. L'enzyme choisie coupe spécifiquement l'ADN dans les copies de rDNA [69]. L'espèce *F. psychrophilum* étant pourvue de six copies distinctes [6], la résolution des différents profils de restriction « macroscopiques » des différentes souches de *F. psychrophilum* reflète un contenu génétique variable entre les souches de la même espèce. Ces différences de plusieurs dizaines de kpb correspondent probablement aux ilots génomiques dont la composition est différente selon les souches. Ces îlots génomiques contiennent les gènes accessoires qui contribuent à leur

diversité. Cependant, les profils de restriction identiques des souches KU061128-1 et NCIMB 1947^{T} et des souches FPC 840 et KU060626-4 ne révèlent aucune corrélation entre ces structures chromosomiques et un poisson hôte particulier, une origine géographique ou une année d'isolement (Tableau 1).

L'analyse des gènes accessoires est possible grâce à la connaissance des génomes de ces souches. Une analyse plus vaste du pan-génome et des génomes accessoires au sein de l'espèce *F. psychrophilum*, incluant à ce jour 22 nouveaux génomes (34 isolats au total, projet européen EMIDA), est actuellement en cours de réalisation (Thèse C. Habib). Ces gènes accessoires permettent par exemple, dans certains cas, d'expliquer l'adaptation à une niche écologique, la résistance aux antibiotiques ou encore la colonisation d'un hôte particulier [55].

Les profils d'hybridation obtenus par Southern-blot pour les douze souches de l'espèce *F. psychrophilum* avec les sondes P3, P2 et P1 ont permis de démontrer expérimentalement que l'organisation du génome de la souche THC 02/90 est retrouvée chez les génomes des souches JIP 16-00, JIP 08-99, LVDJ XP 189, FPC 831, FPC 840, KU061128-1 et KU060626-4. Retrouver cette organisation dans 8 des 9 souches résolues nous permet donc de déduire que l'organisation du génome de *F. psychrophilum* « type THC 02/90 » serait la plus commune dans l'espèce. Cependant, l'importante diversité des souches testées par Southern-blot ne permet pas de corréler l'orientation de leurs structures chromosomiques avec un poisson hôte particulier, une origine géographique ou une année d'isolement (Tableau 1).

Une variation dans la taille des fragments d'ADN hybridés par une même sonde est observée entre les différentes souches. De manière intéressante, cette observation a permis de proposer deux hypothèses.

La première hypothèse serait de penser que les sites de restriction de l'enzyme utilisée (*Xho*I) ne sont pas conservés entre les souches. La complexité de ces régions a rendu impossible leur résolution complète dans les assemblages de ces génomes, ce qui n'a malheureusement pas permis de conclure. Toutefois, avec une moyenne d'environ quatre SNP par kpb entre deux génomes de *F. psychrophilum*, la probabilité que ces

polymorphismes affectent les sites de restrictions *Xho*I (6 pb) parait faible et une autre hypothèse est privilégiée.

Cette seconde hypothèse implique une variation du nombre de blocs de répétitions dans les gènes présents dans les zones résolues par la technique de Southern-blot. Ces blocs de répétitions se situent dans des gènes codant probablement pour des adhésines, protéines exposées à la surface des bactéries (Figure 3A). On peut donc supposer que des variations de taille (expansions et contractions) de structures protéiques présentées à la surface bactérienne pourraient constituer une stratégie de contournement des réponses immunitaires de l'hôte par *F. psychrophilum*. De telles variations dans les antigènes de surface ont déjà été observées chez plusieurs bactéries pathogènes et constituent une stratégie d'échappement aux systèmes immunitaires des hôtes [71].

Présence d'isoformes chromosomiques dans les cultures de *F. psychrophilum*

Bien que des isoformes chromosomiques minoritaires aient été observées dans la souche JIP 02/86 grâce au profil de restriction avec l'enzyme *Ceu*I et lors de l'hybridation avec la sonde P3 en conditions d'exposition prolongées de la membrane (Figure 7), les résultats obtenus ne nous permettent pas de statuer sur la présence systématique d'isoformes chromosomiques dans les différentes cultures de *F. psychrophilum* réalisées. En effet, ces résultats n'ont pas été retrouvés avec les autres sondes et pour les autres souches. Les conditions de cultures des bactéries ou la limite de détection de la technique utilisée pourraient en être la cause. De telles isoformes chromosomiques ont été observées à très basses fréquences (de une à trois pour milles colonies) dans les cultures de la bactérie *Staphylococcus aureus* Mu50 [72]. Les isoformes chromosomiques observées chez *F. psychrophilum* JIP 02/86 pourraient donc probablement exister dans les autres souches de l'espèce mais avec une fréquence inférieure qui ne permettrait pas leur observation dans les conditions expérimentales utilisées.

Chez *Y. pestis*, la présence d'isoformes chromosomiques a vraisemblablement favorisé l'émergence de nouvelles souches pathogènes [67], [73]. L'étude des génomes de *Helicobacter pylori* a montré que la perte ou le gain de gènes codant pour des protéines de la membrane externe, qui modulent les interactions avec l'hôte, se produisait dans les zones charnières d'inversions chromosomiques. L'association étroite de ces inversions chromosomiques avec l'apparition et la délétion de gènes apparaît alors comme un moteur puissant dans l'évolution de cette espèce [74]. En considérant que *F. psychrophilum* est cinq fois plus « recombinogène » que *H. pylori* [68], on peut supposer que ces phénomènes d'inversion chromosomique jouent un rôle similaire dans cette espèce. La présence d'isoformes chromosomiques et les nombreux phénomènes de recombinaison [7] observés chez *F. psychrophilum* mettent en évidence une importante dynamique du génome qui permettrait peut-être d'expliquer la capacité d'adaptation de cette espèce [Stackebrandt E, Extrait de la conférence sur les bactéries du genre *Flavobacterium*. 2009, Paris].

Une étude récente, réalisée chez *S. aureus* Mu50, a mis en évidence deux variants phénotypiques (SCV et NCV) au sein d'une même culture. Le phénotype SCV avait précédemment été associé avec la persistance et la rechute des infections cliniques. Le séquençage des génomes SCV et NCV a montré l'inversion d'une partie du chromosome entre les deux variants phénotypiques, due à la recombinaison entre deux régions homologues situées de part et d'autre de l'origine de réplication. L'inversion est réversible et aboutit au changement de phénotype des deux sous-populations (« phenotypic switching »). La comparaison des profils d'expression des deux variants a montré qu'une dizaine de gènes impliqués dans le métabolisme énergétique étaient sous-exprimés chez le variant SCV sans pour autant expliquer comment l'inversion du génome pouvait aboutir a de telles modifications. Bien que les implications médicales et biologiques de l'inversion du chromosome restent à élucider, cette étude a montré que l'inversion réversible d'une partie du chromosome pourrait être une stratégie utilisée par *S. aureus* pour survivre dans un environnement défavorable [72].

Il existe également une variation de phase aboutissant à un changement de phénotype des colonies au sein des cultures sur boîte de Petri de *F. psychrophilum*, les phénotypes « rugueux » (R) et « lisse » (S). Ces phénotypes R et S sont associés à des phénomènes d'auto-agglutination des bactéries S en milieux de cultures liquides (Annexe 2). Bien que les

cultures contiennent un mélange des deux formes, il est possible de les enrichir en formes R en inoculant uniquement la phase supérieure du bouillon de culture [75]. Les bactéries S sont plus adhérentes au substrat comparées aux variants R suggérant que la variation de phase implique probablement des structures de surface autres que les protéines et lipopolysacharides exposés à la membrane externe qui ont été décrits comme similaires dans les deux types de cellules. Les bactéries R et S étant virulentes pour la truite arc-en-ciel, la distinction des phénotypes ne semble pas être liée à la virulence chez l'hôte [75].

On peut donc se demander si, comme observé chez *S. aureus*, la variation phénotypique observée dans les cultures de *F. psychrophilum* serait provoquée par l'inversion d'une partie du génome. Une stratégie expérimentale a été mise en place afin de tester cette hypothèse. Plusieurs souches de *F. psychrophilum* de phénotype S ont été converties en variants de phénotype R. Les expériences de Southern-blot, présentées ci-dessus, ont alors permis de tester l'inversion du chromosome entre les deux phénotypes R et S pour chaque souche. Malheureusement, les expériences préliminaires réalisées n'ont pas encore permis de conclure que les phénotypes R et S étaient liés à l'inversion d'une partie du chromosome bactérien.

Les changements de morphologie observés sur les colonies bactériennes sont souvent reliés à une variation de phase et sont le résultat de changements dans l'expression de gènes qui conduisent à une modification des structures de surface de la bactérie. La nature des gènes affectés par ces variations peut varier entre les différentes espèces de bactéries [71], [76]. Chez les bactéries pathogènes, une variation entre plusieurs phénotypes peut permettre à la cellule d'échapper au système immunitaire de l'hôte ou de coloniser différents tissus ou organes. Dans l'environnement, la variation de phase peut être utilisée par la bactérie comme un moyen de changer son mode vie et d'alterner par exemple entre la dissémination ou la formation de biofilms [77].

Conclusions

Les différentes approches expérimentales mises en place ont permis de valider la structure et l'assemblage des génomes des souches de *F. psychrophilum* JIP 02/86 et THC 02/90. La diversité des structures chromosomiques observée chez les souches de *F. psychrophilum* testées ne révèle cependant aucune corrélation avec un poisson hôte particulier, une origine géographique ou une année d'isolement.

Nous avons pu montrer expérimentalement que l'organisation inverse du génome de *F. psychrophilum* « type THC 02/90 » est prédominante dans l'espèce et observer une diversité dans la composition des zones résolues par l'expérience reflétant un polymorphisme important ainsi qu'une probable variation de la structure de ces régions au sein de l'espèce.

La mise en évidence d'isoformes chromosomiques chez *F. psychrophilum* JIP 02/86 suggère qu'elles pourraient probablement exister chez les autres souches de l'espèce mais à une fréquence inférieure ou dans des conditions de culture différentes de celles que nous avons utilisées. Cette inversion chromosomique pourrait être un des moteurs évolutifs de cette espèce. L'ensemble de ces résultats démontre un réel dynamisme du génome de l'espèce *F. psychrophilum* et suggère qu'il a un rôle prépondérant dans la diversification des souches au sein de cette espèce.

Deuxième partie : Etude du génome complet d'une bactérie pathogène de poisson : *Flavobacterium branchiophilum*

Introduction

Le genre *Flavobacterium* comprend une grande diversité d'espèces principalement isolées d'environnements aquatiques. Ces espèces incluent notamment trois espèces pathogènes des poissons : *F. psychrophilum, F. columnare* et *F. branchiophilum*. Si les deux premières sont parfois capables de causer des lésions branchiales chez différentes espèces de poissons, *F. branchiophilum* est la principale bactérie responsable du syndrome BGD, une pathologie affectant principalement les salmonidés. Depuis la première description de cette bactérie au Japon, *F. branchiophilum* a été détecté dans de nombreuses régions pratiquant la salmoniculture. En particulier, au Canada (Ontario) où depuis une vingtaine d'années elle est considérée comme une des maladies les plus importantes pour les élevages de salmonidés [78].

L'analyse du génome de *F. psychrophilum* a permis de mettre en évidence des groupes de gènes en relation avec la colonisation, l'invasion et la destruction des tissus de l'hôte. Des propriétés métaboliques particulières en rapport avec la réponse à différents stress et sa survie dans l'environnement avaient pu être également suggérées [6]. Contrairement à *F. psychrophilum* et à *F. columnare*, *F. branchiophilum* est considéré comme une bactérie non invasive. Les tableaux cliniques provoqués chez ses hôtes sont différents (Tableau 1). *F. branchiophilum* présente un tropisme pour les cellules épithéliales des branchies et n'envahit pas les autres tissus. La colonisation massive des branchies et les nécroses qui s'en suivent altèrent les fonctions d'osmorégulation des poissons et provoquent l'asphyxie de l'hôte [32],[33]. La pathologie est généralement caractérisée par une mortalité rapide: entre 20 à 50% de mortalité provoquée en 24h pour les épizooties les plus aigües et selon la taille des poissons [79].

L'analyse du génome de *F. johnsoniae* a permis de mettre en évidence des groupes de gènes en relation avec son mode vie environnemental (par opposition au mode de vie pathogène) avec, en particulier, des gènes codant pour l'utilisation des polysaccharides, la mobilité par glissement et la dégradation de la chitine [13].

Le génome complet de *F. branchiophilum* a donc été essentiellement séquencé pour réaliser une étude de génomique comparative afin d'essayer de comprendre les mécanismes d'adaptation de cette bactérie. La différence des symptômes provoqués chez les poissons par rapport aux deux autres espèces pathogènes suggérait des mécanismes moléculaires de virulence différents. Le génome complet devait donc permettre d'appréhender ces mécanismes de virulence et d'analyser l'évolution des déterminants de virulence au sein du genre *Flavobacterium* ainsi que de préciser le génome central du genre.

Le séquençage du génome de *Flavobacterium branchiophilum* souche FL15 (= CIP 109950) a été réalisé à partir d'un isolat de 1983 sur un alevin de poisson-chat (*Silurus glanis*) atteint de BGD provenant d'une pisciculture Hongroise [80]. Le projet a été initié en 2009 et a bénéficié d'un contrat avec le génoscope d'Evry (CNS) pour la production des séquences, l'assemblage et la finition. J'ai participé à la validation de l'assemblage final par carte optique, à l'annotation du génome et à son analyse, en particulier des îlots génomiques. Ces étapes ont été réalisées en collaboration avec les équipes Génomique Evolutive des Micro-organismes (Institut Pasteur/CNRS) et Mathématique, Informatique et Génome (INRA). La publication de cet article a été une source d'encouragement dans ma démarche car il représente ma première expérience dans le domaine de la génomique et ma première publication scientifique.

Applied and Environmental Microbiology, Nov. 2011, p. 7656–7662 Vol. 77, No. 21
0099-2240/11/$12.00 doi:10.1128/AEM.05625-11

Complete Genome Sequence of the Fish Pathogen *Flavobacterium branchiophilum*▽†

Marie Touchon,[1,2] Paul Barbier,[3] Jean-François Bernardet,[3] Valentin Loux,[4] Benoit Vacherie,[5] Valérie Barbe,[5] Eduardo P. C. Rocha,[1,2] and Eric Duchaud[3]*

Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, F-75015 Paris, France[1]; CNRS, URA2171, F-75015 Paris, France[2]; Unité de Virologie et Immunologie Moléculaires, INRA, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France[3]; Unité de Mathémathique, Informatique et Génome, INRA, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France[4]; and CEA/DSV/IG/Genoscope, Evry, France[5]

Received 26 May 2011/Accepted 3 September 2011

Members of the genus *Flavobacterium* occur in a variety of ecological niches and represent an interesting diversity of lifestyles. *Flavobacterium branchiophilum* is the main causative agent of bacterial gill disease, a severe condition affecting various cultured freshwater fish species worldwide, in particular salmonids in Canada and Japan. We report here the complete genome sequence of strain FL-15 isolated from a diseased sheatfish (*Silurus glanis*) in Hungary. The analysis of the *F. branchiophilum* genome revealed putative mechanisms of pathogenicity strikingly different from those of the other, closely related fish pathogen *Flavobacterium psychrophilum*, including the first cholera-like toxin in a non-*Proteobacteria* and a wealth of adhesins. The comparison with available genomes of other *Flavobacterium* species revealed a small genome size, large differences in chromosome organization, and fewer rRNA and tRNA genes, in line with its more fastidious growth. In addition, horizontal gene transfer shaped the evolution of *F. branchiophilum*, as evidenced by its virulence factors, genomic islands, and CRISPR (clustered regularly interspaced short palindromic repeats) systems. Further functional analysis should help in the understanding of host-pathogen interactions and in the development of rational diagnostic tools and control strategies in fish farms.

The genus *Flavobacterium*, thus far encompassing 65 validly named species, is the type genus of the family *Flavobacteriaceae*, phylum *Bacteroidetes* (3). Representatives of the genus *Flavobacterium* have colonized a wide variety of temperate and polar habitats in terrestrial, freshwater, and marine environments. They are likely of high importance for the turnover/degradation of organic matter in these ecosystems (4). Hence, members of the genus *Flavobacterium* have recently acquired important ecological interest. In addition, some strains may have biotechnological applications owing to the production of cold-adapted enzymes (40) and to their potential for bioremediation (24) or wastewater treatment (44, 47). Although most members of the genus are environmental bacteria, three *Flavobacterium* species are serious fish pathogens that severely impact freshwater aquaculture worldwide (4).

Bacterial gill disease (BGD) is characterized by the presence of numerous bacteria on the surface of the gill epithelium that severely affect the respiratory function of infected fish. Although *Flavobacterium psychrophilum* and *Flavobacterium columnare* may cause gill necrosis (4), *Flavobacterium branchiophilum* (62) is actually the main causative agent of this condition. First recognized on salmonid fish in Japan and Oregon (28, 61), BGD caused by *F. branchiophilum* has been one of the most important conditions affecting the salmonid industry in Ontario, Canada, for 2 decades (43). The disease has also been reported on salmonid as well as nonsalmonid fish in Hungary and The Netherlands (12) and in South Korea (12, 29).

In contrast with *F. psychrophilum* and *F. columnare*, *F. branchiophilum* is a fastidious, nongliding organism with a unique tropism for the gill epithelium and is usually not isolated from internal organs. The disease is characterized by explosive morbidity and mortality attributable to massive bacterial colonization of gill lamellar surfaces, causing irritation and fusion of gill filaments and lamellae. The subsequent necrosis of the gills rapidly impairs the respiratory and osmoregulatory functions (53, 63). So far, no commercial vaccine is available, and the control of BGD by bath treatments using various chemotherapeutics has met with various degrees of success (51).

Only scarce information has been available so far on the pathogenesis of BGD and on the virulence mechanisms of *F. branchiophilum*, making it difficult to adopt preventive approaches to combat this pathogen. In order to get insight into the molecular determinants, with particular emphasis on pathogenicity, we determined and analyzed the complete genome sequence of *F. branchiophilum* FL-15 (CIP 109950), isolated in 1983 from a sheatfish (*Silurus glanis*) fingerling with BGD in a Hungarian fish farm (12). Strain FL-15 was compared to isolates from Japan and Oregon and included in the original description of *F. branchiophilum* (62).

Analysis of the whole genome of *F. psychrophilum* JIP02/86 has revealed sets of genes related to colonization, invasion, and destruction of the host tissues as well as particular metabolic

* Corresponding author. Mailing address: Unité de Virologie et Immunologie Moléculaires, INRA, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France. Phone: 33 1 34 65 25 88. Fax: 33 1 34 65 25 91. E-mail: eric.duchaud@jouy.inra.fr.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 16 September 2011.

properties related to stress response and long-term survival outside the host (9). In contrast, the genome of *Flavobacterium johnsoniae* UW101^T has revealed unique sets of genes in relation to its environmental lifestyle, in particular, genes encoding polysaccharide utilization proteins, gliding motility, and novel biochemical features (35). By sequencing the whole genome of *F. branchiophilum* and performing comparative genomic studies, we aimed at understanding the adaptation of this poorly studied organism and the evolution of virulence in the genus *Flavobacterium*.

MATERIALS AND METHODS

***F. branchiophilum* genome sequencing.** To sequence the complete genome of strain FL-15, a shotgun sequencing strategy based on three different clone libraries and capillary Sanger sequencing was used to obtain a 14-fold coverage of the complete genome. The genomic DNA was fragmented by mechanical shearing. The 3-kb (library A) and 10-kb (library B) inserts were, respectively, cloned into the pcdna2.1 (Invitrogen) and pCNS (pSU18 derived) plasmid vectors while large inserts (40 kb; library C) were cloned into the fosmid vector pCC1Fos. Vector DNAs were purified and end sequenced (library A, 37,632 reads; library B, 13,056 reads; and library C, 5,568 reads) using dye terminator chemistry on ABI 3730 sequencers. The reads were assembled using the whole-genome shotgun assembler Arachne (http://www.broadinstitute.org), and the assembly was visualized by the interface Consed (CodonCode Corp., Dedham, MA). For the finishing phase, we used primer walking of clones, PCRs, and *in vitro* transposition technology (Template Generation System II Kit; Finnzyme, Espoo, Finland), corresponding to 108, 44, and 1,152 reads, respectively. All frameshift sequences were checked, and the assembly was validated by optical mapping (31).

Open reading frame (ORF) prediction and annotation. The prediction of coding sequences was generated using the self-training gene detection software SHOW (38) based on hidden Markov models ([HMMs] http://genome.jouy.inra.fr/ssb/SHOW/). The ribosome-binding sites and transcriptional terminators were detected using the SHOW and Petrin software programs (7), respectively, while the tRNA- and rRNA-encoding genes were detected using the tRNA-scan (32) and RNammer (30) software programs, respectively. Genome annotation including manual validation was performed using the AGMIAL annotation platform (6). Insertion sequences (ISs) were identified using the procedure described in Touchon and Rocha (56) and annotated using IS finder (http://www-is.biotoul.fr/) (52).

ABC transporters and proteases were classified using the ABCISSE (http://www.pasteur.fr/recherche/unites/pmtg/abc/) and MEROPS (http://merops.sanger.ac.uk/) databases, respectively. Laterally transferred regions were predicted by Alien Hunter (http://www.sanger.ac.uk/resources/software/alien_hunter/) (58). Protein localization was predicted using PSORTb, version 3.0 (64). Putative CRISPR (clustered regularly interspaced short palindromic repeats) elements were identified using CRISPRFinder (17). The hidden Markov models for the 45 CRISPR-associated (Cas) protein families described in Haft et al. (19) were obtained from the TIGRFAM database, version 6.0 (http://www.tigr.org/TIGRFAMs/). The *cas* genes were identified with these Cas HMM profiles using hmmpfam (10) with the thresholds of an E value of <0.001 and a positive score. Blastn was used for similarity searches between CRISPR spacer sequences and the 834 complete prokaryote genomes, 1,725 complete plasmid genomes, and 522 virus genomes available in GenBank. Only matches showing an E value of $<1 \times 10^{-5}$ and less than 10% difference in sequence length were retained; matches to sequences found within CRISPR loci were ignored. Prophages were identified using PhageFinder (14). Genes with homology to the transduction-like gene transfer agent (GTA) described in McDaniel et al. (36) were identified using standard Blastp search.

Assignment of orthology. Orthologs were defined by identifying unique pairwise reciprocal best hits, with at least 50% similarity in amino acid sequence and less than 20% difference in protein length. The analysis of orthology was made for every pair of *Flavobacterium* or *Flavobacteriaceae* genomes available in GenBank. The core genome consists of genes found in all genomes analyzed and was defined as the intersection of pairwise lists. Thus, the core genome of the genus *Flavobacterium* is around 1,400 genes. The core genome of the family *Flavobacteriaceae* is comprised of 595 genes detected in all 11 genomes.

Phylogenetic analyses. The reference phylogenetic tree of the family *Flavobacteriaceae* was reconstructed from the concatenated alignments of 595 proteins of the core genome obtained with MUSCLE, version 3.6 (11), and then back-translated to DNA, as is standard usage. We used Tree-Puzzle (49) to compute the distance matrix between all genomes using maximum likelihood under the HKY+G(8)+I (Hasegawa-Kishino-Yano model of nucleotide substitution with gamma distribution allowing 8 categories and a proportion of invariant sites) model. The tree of the core genome was built from the distance matrix using BioNJ (16). We made 1,000 bootstrap experiments on the concatenated sequences to assess the robustness of the topology. The topology of this tree is congruent with previous phylogenetic analyses based on 16S rRNA (33).

Homologs to FBFL15_0919 were searched using Blastp (with an E value of $<10^{-10}$) in the 249 metagenomes of the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system (34). The molecular phylogeny of these putative enterotoxin proteins has been explored by the construction of multiple sequence alignments with MUSCLE, version 3.6 (11). The phylogenetic tree was reconstructed using the maximum-likelihood method implemented in the PhyML program (version 3.0, with approximate likelihood ratio test [aLRT]) with the WAG (Wheelan and Goldman) matrix and a gamma correction for variable evolutionary rates (18). Reliability for the internal branch was assessed using the aLRT (2).

Nucleotide sequence accession numbers. The genomic sequences reported in this paper have been deposited in the EMBL database under the accession number FQ859183 for the bacterial chromosome and FQ859182 for the plasmid.

RESULTS AND DISCUSSION

General genome features. The complete genome of *F. branchiophilum* FL-15 consists of a circular chromosome of 3,559,884 bp (Fig. 1 and Table 1) and one small plasmid, pFB1, of 3,408 bp. The average G+C content is 33% for both the chromosome and the pFB1 plasmid. The chromosome is predicted to contain 2,867 protein-coding genes. We identified three rRNA operons and 44 tRNA genes. Therefore, *F. branchiophilum* contains one of the smallest subset of these genes among the sequenced genomes of the family *Flavobacteriaceae*. This is in accordance with the fastidious growth of *F. branchiophilum* strains as slow growers tend to have fewer such genes (59). The FL-15 genome encodes 36 insertion sequences, of which 8 appear incomplete (see Table S1 in the supplemental material). The pFB1 plasmid considerably differs from the pCP1 plasmid of *F. psychrophilum* JIP02/86 although they share the same size. The pFB1 plasmid is predicted to contain five genes encoding proteins including (i) a plasmid replication initiation protein, (ii) a mobilization protein similar to those previously identified on plasmids from different members of the phylum *Bacteroidetes*, and (iii) a toxin-antitoxin module.

Genome comparison. Pairwise reciprocal best hits were used to identify core protein-encoding genes, i.e., the genes shared by *F. branchiophilum* FL-15, *F. psychrophilum* JIP02/86, and *F. johnsoniae* UW101^T. Using a threshold of 50% similarity in amino acid sequence and less than 20% of difference in protein length, 1,402 core genome genes were identified (Fig. 2). The bulk of the core proteins share 70% to 90% sequence similarity (see Fig. S1 in the supplemental material). This core genome of about 1,400 genes, about half of the *F. branchiophilum* FL-15 genome, is involved in central metabolism and transcription and translation machinery. Little conservation of the gene order was observed between the genomes (see Fig. S2 in the supplemental material) although the orthologs do tend to remain at similar relative positions on the chromosome. This observed X-plot shape likely derives from the accumulation of rearrangements that tend to be symmetrical relative to the replication origin (55), which appears to be conserved in the genus *Flavobacterium*. In contrast with *F. johnsoniae* and *F. psychrophilum*, we found weak GC skew deviations (Fig. 1), suggesting extensive genome shuffling in the lineage leading to

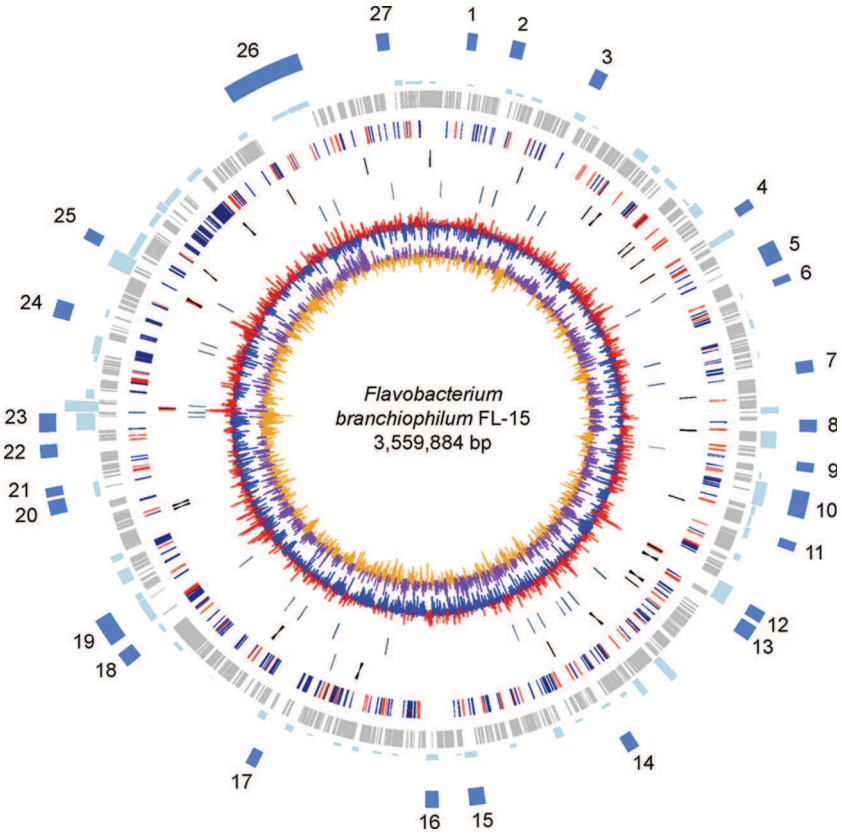


FIG. 1. Circular representation of the *F. branchiophilum* FL-15 genome. Circles represent the following (from the inside out): 1, GC skew [(G − C)/(G + C) using a 2-kbp sliding window] (purple, positive GC skew; orange, negative GC skew); 2, G+C deviation (difference between the average G+C content in a 2-kbp window and the genomic average G+C), where red areas indicate that the deviation is greater than 2 standard deviations; 3, location of ISs; 4, location of tRNA genes (black) and rRNA operons (red); 5, genes with orthologs in *F. johnsoniae* (blue) or *F. psychrophilum* (red); 6, core *Flavobacterium* genome, i.e., genes with orthologs in *F. johnsoniae* UW101^T and *F. psychrophilum* JIP02/86 (gray); 7, putative horizontal gene transfer regions detected by Alien Hunter (58), where the height corresponds to the score prediction (the higher the score, the more significant the detected region); and 8, regions specific to *F. branchiophilum* FL-15 (blue). In the outermost circle, a specific region corresponds to at least 10 consecutive noncore genes and has less than 40% of the genes present in *F. johnsoniae* UW101^T or *F. psychrophilum* JIP02/86. Characteristics of each region (defined by number) are indicated in Table S2 in the supplemental material.

F. branchiophilum. Chromosomal rearrangements may have resulted from homologous or other types of recombination between DNA repeats in the genome. Apart from the above-mentioned IS, we found a large number of Rhs (for rearrangement hot spot) elements in the genome. Rhs elements are complex genetic composites ubiquitous within the family *Enterobacteriaceae* (22). Although commonly found within accessory regions and thought to be implicated in genomic rearrangements in *Escherichia coli*, their functions are poorly understood (23). The FL-15 genome contains 65 *rhs* genes, of which 59 appear incomplete (gene remnants), and five cognate *vgr* genes, two of which are gene remnants. Most of the *rhs* and *vgr* genes are located in the predicted genomic islands (see below).

We identified 1,014 protein-encoding genes in the *F. branchiophilum* FL-15 genome that are absent from the published

TABLE 1. Chromosome features of *F. branchiophilum* FL-15 compared with those of *F. psychrophilum* JIP02/86 and *F. johnsoniae* UW101^T

Chromosome feature	Value for the parameter in:		
	F. branchiophilum FL-15	*F. psychrophilum* JIP02/86	*F. johnsoniae* UW101^T
Genome size (bp)	3,559,884	2,861,988	6,096,872
Plasmid size (bp)	3,408	3,407	No
G+C content (%)	33	32.5	34.1
No. of rRNA operons	3	6	6
No. of tRNA genes	44	49	62
Total no. of protein-coding genes	2,867	2,432	5,056
Protein coding density (%)	82.9	84.5	87.3
Avg gene length (bp)	1,030	1,003	1,061
No. of complete IS elements (pseudogene[s])	28 (8)	28 (21)	16 (1)

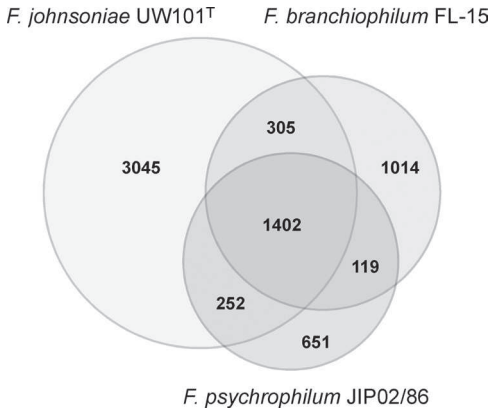


FIG. 2. Venn diagram illustrating the overlap of gene repertoires of *F. branchiophilum* FL-15, *F. johnsoniae* UW101^T, and *F. psychrophilum* JIP02/86. The core *Flavobacterium* genome contains 1,402 genes, i.e., ~50% of the *F. branchiophilum* gene repertoire and ~21% of the pan-*Flavobacterium* genome (6,788 genes).

F. psychrophilum and *F. johnsoniae* genomes. Of these, 42% are randomly distributed along the genome in small clusters of one to three genes. Taking into account the regions encompassing at least 10 noncore-genome genes, we identified 27 large genomic regions specific to *F. branchiophilum* FL-15 (Fig. 1; see also Table S2 in the supplemental material). These regions encompass 526 protein-encoding genes (52%) out of the 1,014. Regions 1, 10, 11, 18, and 20 contain genes likely involved in the detoxification of/resistance to various compounds, including heavy metals, antibiotics, and hydrogen peroxide; regions 15, 17, and 26 contain genes encoding retron-type reverse transcriptases suggesting a foreign origin; regions 8, 13, 22, and 27 contain genes encoding specific carbohydrate metabolism pathways likely involved in exopolysaccharide biosynthesis; region 16 contains phenylacetic acid degradation pathway-encoding genes; regions 5 and 24 contain sugar import- and metabolism-encoding genes; region 9 contains lipid metabolism-encoding genes; and regions 3, 19, and 26 are rich in *rhs* fragments (7, 11, and 33 fragments, respectively). Most of these genomic regions (60%) are predicted to be horizontally transferred by the Alien Hunter program, suggesting that they are indeed genomic islands. The FL-15 genome appears devoid of known prophages, gene transfer agents (GTAs), and integrative conjugative element (ICEs). Located in region 23, the FBFL15_2297 gene probably encodes an integrase, and the FBFL15_2295 and FBFL15_2299 genes probably encode excisionases; however, the neighboring genes are unrelated to prophage (i.e., phage structural phage genes) or ICE (i.e., conjugation-related genes).

CRISPR (clustered regularly interspaced short palindromic repeats) loci. The CRISPR system is thought to be an adaptative hereditary immune system of prokaryotes that allows them to cope with foreign genetic elements (26). The FL-15 genome contains three CRISPR loci with 28, 31, and 39 direct repeats (DR), respectively. While the DR length is 36 bp in all cases, the sequence is different in each CRISPR locus. The third CRISPR locus is not canonical, with three interruptions in the modular pattern. Only the first CRISPR locus appears to be located near two known CRISPR-associated (*cas*) genes (*FBFL15_1617* and *FBFL15_1622* encode Cas2 and Csn1 proteins, respectively). However, this CRISPR Nmeni subtype system (which is complete in the genome of *F. psychrophilum* JIP02/86) seems incomplete in the FL-15 genome as the *cas1* gene is absent, and the *cas2* gene is rearranged (i.e., *cas2* is downstream of the CRISPR locus and is no longer near *csn1*). Since *cas1* is probably essential for the correct functioning of CRISPR, these results suggest that these CRISPR loci are no longer functional, at least in strain FL-15.

Using nonstringent criteria (Blastn E value of $<1 \times 10^{-5}$ and less than 10% difference in sequence length), no homologous sequence (i.e., proto-spacers) has been identified in any of the plasmid/phage/bacterial complete genomes available in GenBank. This is likely the consequence of the very low number of sequenced mobile genetic elements of members of the family *Flavobacteriaceae* available thus far (one phage and seven plasmids). The lack of CRISPR in *F. johnsoniae* UW101^T and the presence of ISs in the direct vicinity of CRISPR in *F. psychrophilum* JIP02/86 suggest that CRISPR are poorly suited for the typing of *Flavobacterium* strains.

Toxins. Strikingly, the genome of strain FL-15 is devoid of most of the predicted toxins previously identified in the *F. psychrophilum* JIP02/86 genome (9). Nevertheless, it contains a peculiar set of genes likely to be involved in virulence. The *FBFL15_0919* gene probably encodes a preprotein 51% similar to the heat-labile toxin (LTA) expressed by enterotoxigenic *Escherichia coli* strains. The presence of a signal peptide and the conservation of the amino acid residues at the active site and at the NAD binding site support the functional homology between the FBFL15_0919 protein and LTA. LTA shows high sequence similarity (89%) with the cholera toxin (CTA), and both have been intensively studied as virulence factors in mammalian species (8). The LTA and CTA toxins function similarly by stimulating adenylate cyclase and provoking massive loss of water and electrolytes through the intestinal epithelium cells of the host (57). It is therefore tempting to speculate that *FBFL15_0919* has a similar mode of action that disturbs the osmoregulatory function of the epithelial cells of the gills. Indeed, these cells are of utmost importance not only for the oxygen uptake of fish but also for the excretion of urea and for the active import of salts that compensate passive salt lost in freshwater (20). To our knowledge, it is the first time a gene similar to LTA and CTA has been detected outside the phylum *Proteobacteria*. The molecular phylogeny of these proteins argues against a direct lateral gene transfer between proteobacteria and strain FL-15 (see Fig. S3 in the supplemental material).

The *FBFL15_0520* gene encodes a preprotein 47% similar to streptopain, an important streptococcal virulence factor playing a role in bacterial colonization, invasion, and inhibition of wound healing (25). In *F. psychrophilum*, this gene is absent from the genome of strain JIP02/86 (9) but present on a genomic island in the genome of strain THC02/90 (E. Duchaud, unpublished data).

F. branchiophilum has been reported to degrade gelatin and casein (62), and extracellular proteases of *F. psychrophilum* and *Porphyromonas gingivalis* (another pathogenic member of the phylum *Bacteroidetes*) have been suggested to play important roles in virulence (5, 41). Among the 11 putative secreted protease precursors identified in the FL-15 genome, two be-

long to the M1 family, one to the M36 fungalysin family, one to the M43 cytophagalysin family, one to the M12B family, and two to the S46 family. Together with five predicted peptide and amino acid importers, they might be involved in the breakdown and uptake of proteinaceous compounds during host tissue colonization.

Adhesion, motility, and secretion. Bacterial attachment is essential in the initiation of mucosal infection, and previous reports have stressed the adherence properties of *F. branchiophilum*. Formalin-killed or acetone-killed bacterial cells retained part of their adherent nature, and adherence was never totally inhibited whatever the compound tested (42). In accordance with these findings, the FL-15 genome encodes 20 predicted adhesin precursors that could be implicated in cell-cell and cell-surface interactions.

Electron micrographs suggested that *F. branchiophilum* displays pili or fimbriae on the cell surface (43, 62), and pilus-like structures were indeed purified and partially characterized (21). However, no genes encoding known pilus or fimbrial proteins were identified in the FL-15 genome. If present, the corresponding genes may be hidden within those encoding hypothetical proteins; it is also possible that production of pilus-like structures is a strain-dependent feature absent from strain FL-15.

F. branchiophilum was reported to be devoid of gliding motility, the type of motility that occurs in many *Flavobacterium* species (3). Indeed, gliding motility of strain FL-15 was never observed in our hands. Two distinct groups of genes, *gld* and *spr*, are involved in gliding motility. The *gld* genes identified so far are thought to encode the gliding motor as they are mandatory for gliding motility. In addition, several *spr* genes encoding paralogous adhesins also have roles in motility (46). Intriguingly, the FL-15 genome contains all the *gld* genes and at least some of the *spr* genes. This suggests that *F. branchiophilum* may actually be motile, but experimental conditions used so far failed to mimic natural conditions where gliding motility is expressed.

A link between the *Flavobacterium* motility apparatus (i.e., Gld proteins) and a protein translocation system (referred to as the Por secretion system, or PorSS) has been described recently (48). The fact that the *porP*, *porQ*, *porR*, *porS*, and *porT* genes are present in the FL-15 genome suggests that *F. branchiophilum* is proficient in protein secretion through the PorSS. Moreover, 31 predicted secreted proteins, including proteases and adhesins, harbor a conserved C-terminal domain (CTD) (see Table S3 in the supplemental material). This CTD, also found in proteins of other members of the phylum *Bacteroidetes* (9, 27), is involved in their attachment to the outer membrane (50). Because the proteins that possess a CTD are likely secreted by the PorSS (48), they can be promising targets for vaccine development strategies.

Metabolism. The elements of the central energy metabolism (glycolysis, tricarboxylic acid cycle, and oxidative phosphorylation) of *F. branchiophilum* are globally similar to those depicted in *F. psychrophilum* (9) and *F. johnsoniae* (35). Intriguingly, the *F. branchiophilum* genome encodes a class I fumarate hydratase (*fumA*) and an aconitate hydratase 2 (*acnB*), and the *F. psychrophilum* genome encodes a class II fumarate hydratase (*fumC*) and an aconitate hydratase 1 (*acnA*) while the *F. johnsoniae* genome encodes both type of enzymes.

In line with other *Flavobacterium* species, *F. branchiophilum* grows only by aerobic respiration and is unable to use fermentation or anaerobic respiration. Hence, the genes encoding menaquinone biosynthesis, cytochrome *c*, ATP synthase, and NADH dehydrogenases are all present. The genes *actABCDEF* and *ccsBA* encoding the recently described components of the alternative complex III menaquinol-cytochrome *c* oxidoreductase (45) and the components of system II cytochrome *c* synthetase and heme channel (15), respectively, are also present in the FL-15 genome.

The genomes of *F. psychrophilum* (9) and *F. johnsoniae* (35) both contain some amino acid catabolic pathway-encoding genes [i.e., glycine C-acetyltransferase (*kbl*), L-threonine 3-dehydrogenase (*ltd*), and phenylalanine 4-monooxygenase (*phhA*)] that are lacking in the FL-15 genome. Nevertheless, the latter contains a locus (absent in the genomes of *F. psychrophilum* and *F. johnsoniae*) of about 12 kb encompassing 10 genes (FBFL15_1502 to FBFL15_1511) similar to the *paa* gene cluster of some *E. coli* strains (13). These *paa* genes encode the only known aerobic degradation pathway for phenylacetate, which is the most common for phenylalanine (54). Moreover, this pathway occurs in various pathogens, and a connection with virulence through toxicity of reactive early intermediates was proposed (reference 54 and references therein).

Polysaccharide utilization. Consistent with the strictly aerobic lifestyle of *F. branchiophilum*, its genome lacks phosphotransferase systems for sugar import. *F. branchiophilum* is able to hydrolyze starch and to produce acid from various carbohydrates under aerobic conditions (12, 43, 62). Indeed, analysis of the FL-15 genome identified 23 predicted glycoside hydrolases likely involved in carbohydrate catabolism, of which 17 contain a signal peptide (see Table S4 in the supplemental material). Among these, *FBFL15_2407* is predicted to encode a levanase precursor, *FBFL15_2433* encodes a cellulase precursor, and *FBFL15_2455* encodes an arabinogalactan endo-1,4-β-galactosidase precursor. These genes are absent from the previously sequenced *Flavobacterium* genomes, but they were identified in members of the marine clade of the family *Flavobacteriaceae*.

Commonly found in members of the phylum *Bacteroidetes*, polysaccharide utilization involves outer membrane oligomer transport systems encoded by *susC*-like and *susD*-like genes, referred to as polysaccharide utilization loci (PULs). This capacity to degrade complex polysaccharides has been shown to be of high importance for marine members of the family *Flavobacteriaceae* (37). In the FL-15 genome, four gene clusters encoding predicted PULs were identified, one of which may not be functional due to a frameshift mutation in the *susC*-like gene. These loci encompass *susC*-like, *susD*-like, and other adjacent genes likely involved in polysaccharide utilization (i.e., encoding glycoside hydrolases, polysaccharide lyases, and carbohydrate esterases). While *F. psychrophilum* is devoid of such systems, *F. johnsoniae* appears to have an arsenal of PULs for the digestion of carbohydrates, including plant cell wall polysaccharides, which may reflect its prevalence in soil and rhizosphere habitats (35). It was therefore rather surprising to identify PULs in the FL-15 genome, especially the *FBFL15_2659-2677* locus resembling the *Fjoh_4246-4265* locus predicted to be involved in pectin utilization (35). This may

suggest that *F. branchiophilum* is able to use some plant carbohydrates or gill mucus mucopolysaccharides.

Pigments. In contrast with *F. johnsoniae* and *F. psychrophilum*, *F. branchiophilum* cells do not produce flexirubin-type pigments (62). Indeed, the *F. branchiophilum* genome is devoid of the flexirubin biosynthesis gene cluster identified in the *F. johnsoniae* and *F. psychrophilum* genomes. Nevertheless, it contains the *crtIBZY* gene cluster predicted to be involved in the biosynthesis of carotenoid-type pigments, suggesting that carotenoids are the only pigments responsible for the light yellow appearance of the colonies.

Bacterial stress genes. Most bacteria have to face many stresses, among which are UV exposure, cold/heat, starvation, and toxic chemicals (including heavy metals and antibiotics). In addition, pathogenic bacteria are confronted with a host response that includes the oxidative stress during infection and transmission.

The FL-15 genome is predicted to encode 37 proteins likely involved in detoxification. Among these, five TerZ/TerD, one TerC, and one TelA family proteins are predicted to be involved in tellurium resistance, two arsenate reductases and an arsenite efflux transporter are involved in resistance to arsenic compounds, two proteins (CopA and CopZ) are involved in resistance to copper compounds, four superoxide dismutases are involved in the elimination of superoxide radicals produced by the host oxidative burst during infection, and one DNA alkylation repair enzyme, AlkD, is involved in the repair of 7-methylguanine alkylation damage on DNA. Genes predicted to be involved in antibiotic resistance were also identified in the FL-15 genome, including one chloramphenicol acetyltransferase, two bleomycin resistance proteins, and two acetyltransferases highly similar to Vat, a staphylococcal protein inactivating the A-type compounds of virginiamycin-type antibiotics (1). Therefore, *F. branchiophilum* FL-15 seems well equipped to face various kinds of stress. This gene repertoire could be of importance for the survival of the bacterium and might provide a selective advantage during host colonization against competitors in the environment.

Summary and future directions. We reported here the complete genome sequence of *F. branchiophilum*, a serious pathogen of freshwater fish in many geographic areas. Comparison with the available genomes of other *Flavobacterium* species has revealed striking differences in chromosome organization and gene content. *F. branchiophilum* is phylogenetically more closely related to *F. psychrophilum* than to *F. johnsoniae* on the basis of 16S rRNA genes (33) and concatenated core genome proteins (see Fig. S4 in the supplemental material). However, its biochemistry resembles more that of *F. johnsoniae.* In addition, its toxins are unrelated to those of *F. psychrophilum* and were originally described in widely distant taxa. These elements point to very different paths in the evolution of virulence in *F. branchiophilum* and *F. psychrophilum*. Intense gene transfer is suggested by the large number of genomic islands and by the presence of one plasmid; together with the accumulation of genomic rearrangements, gene transfer could have an important role in the diversification of the species. Indeed, *Flavobacterium* species are highly prone to homologous recombination (39, 60) and represent the major recipients of natural gene transfer agents in the oceans (36).

This genome sequence therefore provides insights into the lifestyle of this understudied pathogen and should help in the development of rational diagnostic tools and more efficient control strategies in fish farms. In addition, it should yield molecular markers for the development of population structure analysis and epidemiological survey using, e.g., multilocus sequence typing (MLST)-based or variable-number tandem repeat (VNTR) genotyping-based strategies. The availability of this genome sequence should also facilitate the development of functional genomic studies. As new sequencing methods now provide high-throughput short reads, the genome of strain FL-15 may also serve as a reference, allowing read mapping and scaffolding of the genome sequence of other *F. branchiophilum* strains for a better understanding of intraspecies diversity.

ACKNOWLEDGMENTS

This work was supported in part by grant 07-GMGE from the Agence Nationale de la Recherche of France. P.B. is a Université Evry Val d'Essonne Ph.D. fellowship.

We are indebted to H. Wakabayashi for kindly providing strain FL-15. We also thank Stéphane Chaillou, Guillaume Achaz, and Pierre Nicolas for critical reading of the manuscript. We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources.

REFERENCES

1. **Allignet, J., and N. el Solh.** 1995. Diversity among the gram-positive acetyltransferases inactivating streptogramin A and structurally related compounds and characterization of a new staphylococcal determinant, *vatB*. Antimicrob. Agents Chemother. **39:**2027–2036.
2. **Anisimova, M., and O. Gascuel.** 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. **55:**539–552.
3. **Bernardet, J.-F., and J. P. Bowman.** 2011. Genus I. *Flavobacterium*, p. 112–154. *In* W. Whitman (ed.), Bergey's manual of systematic bacteriology, 2nd ed., vol. 4. Williams and Wilkins, Baltimore, MD.
4. **Bernardet, J.-F., and J. P. Bowman.** 2006. The genus *Flavobacterium*, p. 481–531. *In* M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer, and E. Stackebrandt (ed.), The prokaryotes, a handbook on the biology of bacteria, 3rd ed., vol. 7. Springer, New York, NY.
5. **Bertolini, J. M., H. Wakabayashi, V. G. Watral, M. J. Whipple, and J. S. Rohovec.** 1994. Electrophoretic detection of proteases from selected strains of *Flexibacter psychrophilus* and assessment of their variability. J. Aquat. Anim. Health **6:**224–233.
6. **Bryson, K., et al.** 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. **34:**3533–3545.
7. **d'Aubenton Carafa, Y., E. Brody, and C. Thermes.** 1990. Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. J. Mol. Biol. **216:**835–858.
8. **de Haan, L., and T. R. Hirst.** 2000. Cholera toxin and related enterotoxins: a cell biological and immunological perspective. J. Nat. Toxins **9:**281–297.
9. **Duchaud, E., et al.** 2007. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. Nat. Biotechnol. **25:**763–769.
10. **Eddy, S. R.** 1998. Profile hidden Markov models. Bioinformatics **14:**755–763.
11. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
12. **Farkas, J.** 1985. Filamentous *Flavobacterium* sp. isolated from fish with gill diseases in cold water. Aquaculture **44:**1–10.
13. **Ferrandez, A., et al.** 1998. Catabolism of phenylacetic acid in *Escherichia coli*. Characterization of a new aerobic hybrid pathway. J. Biol. Chem. **273:**25974–25986.
14. **Fouts, D. E.** 2006. Phage Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. **34:**5839–5851.
15. **Frawley, E. R., and R. G. Kranz.** 2009. CcsBA is a cytochrome *c* synthetase that also functions in heme transport. Proc. Natl. Acad. Sci. U. S. A. **106:**10201–10206.
16. **Gascuel, O.** 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14:**685–695.
17. **Grissa, I., G. Vergnaud, and C. Pourcel.** 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. **35:**W52–W57.
18. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.

19. **Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson.** 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. **1:**e60.
20. **Helfman, G. S., B. B. Collette, D. E. Facey, and B. W. Bowen.** 2009. The diversity of fishes: biology, evolution, and ecology, 2nd ed. Wiley-Blackwell, Oxford, United Kingdom.
21. **Heo, G. J., H. Wakabayashi, and S. Watabe.** 1990. Purification and characterisation of pili from *Flavobacterium branchiophila*. Fish Pathol. **25:**21–27.
22. **Hill, C. W., C. H. Sandt, and D. A. Vlazny.** 1994. Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. Mol. Microbiol. **12:**865–871.
23. **Jackson, A. P., G. H. Thomas, J. Parkhill, and N. R. Thomson.** 2009. Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC Genomics **10:**584.
24. **Jit, S., M. Dadhwal, O. Prakash, and R. Lal.** 2008. *Flavobacterium lindanitolerans* sp. nov., isolated from hexachlorocyclohexane-contaminated soil. Int. J. Syst. Evol. Microbiol. **58:**1665–1669.
25. **Kapur, V., et al.** 1993. A conserved *Streptococcus pyogenes* extracellular cysteine protease cleaves human fibronectin and degrades vitronectin. Microb. Pathog. **15:**327–346.
26. **Karginov, F. V., and G. J. Hannon.** 2010. The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol. Cell **37:**7–19.
27. **Karlsson, E. N., et al.** 2004. The modular xylanase Xyn10A from *Rhodothermus marinus* is cell-attached, and its C-terminal domain has several putative homologues among cell-attached proteins within the phylum *Bacteroidetes*. FEMS Microbiol. Lett. **241:**233–242.
28. **Kimura, N., H. Wakabayashi, and S. Kudo.** 1978. Studies on bacterial gill disease in salmonids I. Selection of bacterium transmitting gill disease. Fish Pathol. **12:**233–242.
29. **Ko, Y. M., and G. J. Heo.** 1997. Characteristics of *Flavobacterium branchiophilum* isolated from rainbow trout in Korea. Fish Pathol. **32:**97–102.
30. **Lagesen, K., et al.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.
31. **Latreille, P., et al.** 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics **8:**321.
32. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
33. **Madhaiyan, M., S. Poonguzhali, J. S. Lee, K. C. Lee, and S. Sundaram.** 2010. *Flavobacterium glycines* sp. nov., a facultative methylotroph isolated from the rhizosphere of soybean. Int. J. Syst. Evol. Microbiol. **60:**2187–2192.
34. **Markowitz, V. M., et al.** 2008. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. **36:**D534–D538.
35. **McBride, M. J., et al.** 2009. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. Appl. Environ. Microbiol. **75:**6864–6875.
36. **McDaniel, L. D., et al.** 2010. High frequency of horizontal gene transfer in the oceans. Science **330:**50.
37. **Michel, G., P. Nyval-Collen, T. Barbeyron, M. Czjzek, and W. Helbert.** 2006. Bioconversion of red seaweed galactans: a focus on bacterial agarases and carrageenases. Appl. Microbiol. Biotechnol. **71:**23–33.
38. **Nicolas, P., et al.** 2002. Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. Nucleic Acids Res. **30:**1418–1426.
39. **Nicolas, P., et al.** 2008. Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. Appl. Environ. Microbiol. **74:**3702–3709.
40. **Nogi, Y., K. Soda, and T. Oikawa.** 2005. *Flavobacterium frigidimaris* sp. nov., isolated from Antarctic seawater. Syst. Appl. Microbiol. **28:**310–315.
41. **O'Brien-Simpson, N. M., et al.** 2001. Role of RgpA, RgpB, and Kgp proteinases in virulence of *Porphyromonas gingivalis* W50 in a murine lesion model. Infect. Immun. **69:**7527–7534.
42. **Ostland, V. E., J. S. Lumsden, D. D. MacPhee, J. A. Derksen, and H. W. Ferguson.** 1997. Inhibition of the attachment of *Flavobacterium branchiophilum* to the gills of rainbow trout, *Oncorhynchus mykiss* (Walbaum). J. Fish Dis. **20:**109–117.
43. **Ostland, V. E., J. S. Lumsden, D. D. MacPhee, and H. W. Ferguson.** 1994. Characteristics of *Flavobacterium branchiophilum*, the cause of salmonid bacterial gill disease in Ontario. J. Aquat. Anim. Health **6:**13–26.
44. **Park, M., et al.** 2006. *Flavobacterium croceum* sp. nov., isolated from activated sludge. Int. J. Syst. Evol. Microbiol. **56:**2443–2447.
45. **Refojo, P. N., F. L. Sousa, M. Teixeira, and M. M. Pereira.** 2010. The alternative complex III: a different architecture using known building modules. Biochim. Biophys. Acta **1797:**1869–1876.
46. **Rhodes, R. G., S. S. Nelson, S. Pochiraju, and M. J. McBride.** 2011. *Flavobacterium johnsoniae sprB* is part of an operon spanning the additional gliding motility genes *sprC*, *sprD*, and *sprF*. J. Bacteriol. **193:**599–610.
47. **Ryu, S. H., et al.** 2007. *Flavobacterium filum* sp. nov., isolated from a wastewater treatment plant in Korea. Int. J. Syst. Evol. Microbiol. **57:**2026–2030.
48. **Sato, K., et al.** 2010. A protein secretion system linked to *Bacteroidetes* gliding motility and pathogenesis. Proc. Natl. Acad. Sci. U. S. A. **107:**276–281.
49. **Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler.** 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:**502–504.
50. **Seers, C. A., et al.** 2006. The RgpB C-terminal domain has a role in attachment of RgpB to the outer membrane and belongs to a novel C-terminal-domain family found in *Porphyromonas gingivalis*. J. Bacteriol. **188:**6376–6386.
51. **Shotts, E. B., Jr., and C. E. Starliper.** 1999. Flavobacterial diseases: columnaris disease, cold-water disease and bacterial gill disease, p. 559–576. *In* P. T. K. Woo and D. W. Bruno (ed.), Fish diseases and disorders, vol. 3. CABI Publishing, Oxford, United Kingdom.
52. **Siguier, P., J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler.** 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. **34:**D32–D36.
53. **Spear, D. J., H. W. Ferguson, F. W. M. Beamish, J. A. Yager, and S. Yamashiro.** 1991. Pathology of bacterial gill disease: sequential development of lesions during natural outbreaks of disease. J. Fish Dis. **14:**21–32.
54. **Teufel, R., et al.** 2010. Bacterial phenylalanine and phenylacetate catabolic pathway revealed. Proc. Natl. Acad. Sci. U. S. A. **107:**14390–14395.
55. **Tillier, E. R., and R. A. Collins.** 2000. Genome rearrangement by replication-directed translocation. Nat. Genet. **26:**195–197.
56. **Touchon, M., and E. P. Rocha.** 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. **24:**969–981.
57. **Vanden Broeck, D., C. Horvath, and M. J. De Wolf.** 2007. *Vibrio cholerae*: cholera toxin. Int. J. Biochem. Cell Biol. **39:**1771–1775.
58. **Vernikos, G. S., and J. Parkhill.** 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics **22:**2196–2203.
59. **Vieira-Silva, S., and E. P. Rocha.** 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. **6:**1000808.
60. **Vos, M., and X. Didelot.** 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. **3:**199–208.
61. **Wakabayashi, H., S. Egusa, and J. L. Fryer.** 1980. Characteristics of filamentous bacteria isolated from a gill disease of salmonids. Can. J. Fisheries Aquatic Sci. **37:**1499–1507.
62. **Wakabayashi, H., G. J. Hun, and N. Kimura.** 1989. *Flavobacterium branchiophila* sp. nov., a causative agent of bacterial gill disease of freshwater fishes. Int. J. Syst. Bacteriol. **39:**213–216.
63. **Wakabayashi, H., and T. Iwado.** 1985. Effects of a bacterial gill disease on the respiratory functions of juvenile rainbow trout, p. 153–160. *In* A. E. Ellis (ed.), Fish and shellfish pathology. Academic Press, London, United Kingdom.
64. **Yu, N. Y., et al.** 2007. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics **26:**1608–1615.

TABLE S1. Predicted insertion sequences in the FL-15 genome

IS Families	IS630	IS200/IS605	IS5	IS3	ISL3	IS1	IS1634	IS1595	IS110	IS1182	Putative
IS complete		3	1		4	1	3	5	8	2	1
Size (AA)		145-153	348		304	139	487-616	298	383	514	
Locus_tag FBFL15_		0819 1227 1814	1306		0689 1186 2867 3006	1785	0117 0141 0225	0024	0252 0642 0527 1206 1277 1716 1961 2326	1115 2429	2300
IS pseudogene	1		1	1				1		4	
Locus_tag FBFL15_	2420		(0234 + 0235)	(2827 + 2828 + 2829)				(0291 + 0292)		1867 (1871 + 1872) 2635 2637	

TABLE S2. Predicted genomic islands in the FL-15 genome

Region	Protein description	Region Size (bp)	Region Start (bp)	Region End (bp)	Number of complete genes	Number of pseudogenes	% Protein of unknown function	Putative Role	Alien Hunter detection	tRNA +/- 5 genes at each end of the region
>Reg1	**FBFL15_0050:FBFL15_0067**	**13788**	**59138**	**72926**	**18**	**0**	**61**	**Detoxification (Tellurium)**	**Yes**	**No**
11	Protein of unknown function									
7	Detoxification									
>Reg2	**FBFL15_00108:FBFL15_00123**	**18748**	**125426**	**144174**	**14**	**2**	**79**	**Unknown**	**Yes**	**No**
11	Protein of unknown function									
1	Miscellaneous Putative plasmid stabilization system									
1	Transposon and IS									
1	DNA restriction and modification (and repair)									
>Reg3	**FBFL15_00216:FBFL15_00235**	**18311**	**253369**	**271680**	**11**	**9**	**73**	**Unknown**	**Yes**	**No**
8	Protein of unknown function									
2	Miscellaneous Putative plasmid stabilization system									
1	Transposon and IS									
>Reg4	**FBFL15_0465:FBFL15_0475**	**14409**	**548676**	**563085**	**11**	**0**	**64**	**Unknown**	**No**	**tRNA-Ser : FBFL15_0479**
7	Protein of unknown function									
2	DNA restriction and modification (and repair)									
1	Transcription regulation									
1	Sensors (signal transduction)									
>Reg5	**FBFL15_0527:FBFL15_0548**	**30878**	**618903**	**649781**	**22**	**0**	**41**	**Sugar import and sugar metabolism**	**No**	**tRNA-Leu : FBFL15_0523**
9	Protein of unknown function									
6	Transport/binding proteins and lipoproteins									
2	Metabolism of carbohydrate									
2	Metabolism of amino acids and related molecules									
1	Transposon and IS									
1	Transcription regulation									
1	Metabolism of coenzymes and prosthetic groups									
>Reg6	**FBFL15_0569:FBFL15_0578**	**10080**	**672432**	**682512**	**10**	**0**	**80**	**Unknown**	**Yes**	**No**
8	Protein of unknown function									
1	Transport/binding proteins and lipoproteins									
1	Metabolism of sulfur									
>Reg7	**FBFL15_0688:FBFL15_0701**	**20427**	**799755**	**820182**	**14**	**0**	**64**	**Unknown**	**No**	**No**
9	Protein of unknown function									
1	Metabolism of carbohydrate									
1	Cell-Wall									
1	Transposon and IS									
1	Transcription regulation									
1	Sensors (signal transduction)									
>Reg8	**FBFL15_0763:FBFL15_0779**	**17894**	**890529**	**908423**	**17**	**0**	**47**	**Metabolism of carbohydrate; EPS precursors biosynthesis**	**No**	**tRNA-Thr : FBFL15_0783**
8	Protein of unknown function									
9	Metabolism of carbohydrate									
>Reg9	**FBFL15_0815:FBFL15_0826**	**13159**	**954733**	**967892**	**12**	**0**	**42**	**Metabolism of lipids**	**No**	**No**
5	Protein of unknown function									
3	Metabolism of lipids									
2	DNA restriction and modification (and repair)									
1	Transposon and IS									
1	Transcription regulation									
>Reg10	**FBFL15_0853:FBFL15_0887**	**37271**	**999418**	**1036689**	**34**	**2**	**65**	**Detoxification (Cuivre)**	**Yes**	**No**
22	Protein of unknown function									
4	Detoxification									
6	Transport/binding proteins and lipoproteins									
1	Membrane bioenergetics (electron transport chain and ATP synthase)									
1	DNA restriction and modification (and repair)									
>Reg11	**FBFL15_0913:FBFL15_0923**	**13061**	**1073708**	**1086769**	**11**	**0**	**55**	**Detoxification (O2-) + Toxin CTxA1**	**Yes**	**tRNA-Asp : FBFL15_0909**
6	Protein of unknown function									
2	Detoxification									
1	Transport/binding proteins and lipoproteins									
1	Transcription regulation									
1	Miscellaneous Probable toxin precursor									
>Reg12	**FBFL15_1008:FBFL15_1018**	**16942**	**1185032**	**1201974**	**11**	**0**	**100**	**Unknown**	**Yes**	**tRNA-Ala : FBFL15_1019**
11	Protein of unknown function									
>Reg13	**FBFL15_1029:FBFL15_1049**	**22585**	**1210007**	**1232592**	**21**	**0**	**43**	**Metabolism of carbohydrate; EPS precursors biosynthesis**	**No**	**No**
9	Protein of unknown function									
6	Metabolism of carbohydrate									
3	Metabolism of amino acids and related molecules									
2	Cell-Wall									
1	Transport/binding proteins and lipoproteins									
>Reg14	**FBFL15_1244:FBFL15_1262**	**17102**	**1455032**	**1472134**	**19**	**1**	**84**	**Unknown**	**Yes**	**No**
16	Protein of unknown function									
2	Protein modification									
1	Transport/binding proteins and lipoproteins									
>Reg15	**FBFL15_1441:FBFL15_1460**	**22453**	**1695349**	**1717802**	**16**	**3**	**88**	**Unknown/RETRON ?**	**Yes**	**No**
14	Protein of unknown function									
2	Transport/binding proteins and lipoproteins									

>Reg16	**FBFL15_1495:FBFL15_1511**	**19265**	**1764204**	**1783469**	**17**	**0**	**24**	**phenylacetic acid degradation**	**Yes**	**No**
11	Metabolism of lipids									
4	Protein of unknown function									
1	Metabolism of amino acids and related molecules									
1	Membrane bioenergetics (electron transport chain and ATP synthase)									
>Reg17	**FBFL15_1737:FBFL15_1746**	**15608**	**2040374**	**2055982**	**10**	**0**	**70**	**Unknown/RETRON ?**	**No**	**No**
7	Protein of unknown function									
1	Transposon and IS									
1	Transcription regulation									
1	RNA modification									
>Reg18	**FBFL15_1948:FBFL15_1968**	**22071**	**2279979**	**2302050**	**16**	**3**	**44**	**Detoxification (antibio??)**	**Yes**	**No**
7	Protein of unknown function									
2	Transport/binding proteins and lipoproteins									
1	Transposon and IS									
1	Detoxification									
1	Transcription regulation									
1	Sensors (signal transduction)									
1	Membrane bioenergetics (electron transport chain and ATP synthase)									
1	Metabolism of coenzymes and prosthetic groups									
1	DNA restriction and modification (and repair)									
>Reg19	**FBFL15_1981:FBFL15_2014**	**42022**	**2318328**	**2360350**	**23**	**11**	**65**	**Unknown**	**Yes**	**No**
15	Protein of unknown function									
3	Transport/binding proteins and lipoproteins									
2	Transcription regulation									
3	Miscellaneous									
>Reg20	**FBFL15_2168:FBFL15_2182**	**21143**	**2528040**	**2549183**	**15**	**0**	**47**	**Detoxification (H2O2)**	**No**	**No**
7	Protein of unknown function									
2	Transport/binding proteins and lipoproteins									
1	Transcription regulation									
1	Sensors (signal transduction)									
1	Miscellaneous									
1	Membrane bioenergetics (electron transport chain and ATP synthase)									
1	Detoxification									
1	DNA restriction and modification (and repair)									
>Reg21	**FBFL15_2186:FBFL15_2207**	**13231**	**2554956**	**2568187**	**20**	**2**	**85**	**Unknown**	**Yes**	**No**
17	Protein of unknown function									
3	Transport/binding proteins and lipoproteins									
>Reg22	**FBFL15_2249:FBFL15_2262**	**19201**	**2612939**	**2632140**	**14**	**0**	**36**	**Metabolism of carbohydrate; EPS biosynthesis and export**	**No**	**No**
5	Protein of unknown function									
6	Metabolism of carbohydrate									
1	Transport/binding proteins and lipoproteins									
1	Cell-Wall									
1	Sensors (signal transduction)									
>Reg23	**FBFL15_2280:FBFL15_2302**	**27245**	**2651194**	**2678439**	**23**	**0**	**65**	**Unknown /** Phage/plasmid maintenance system	**Yes**	**No**
15	Protein of unknown function									
4	Miscellaneous									
1	Transcription regulation									
1	Phage-related function									
1	Transposon and IS									
1	DNA restriction and modification (and repair)									
>Reg24	**FBFL15_2405:FBFL15_2425**	**25813**	**2821994**	**2847807**	**21**	**0**	**43**	**Sugar import**	**No**	**No**
9	Protein of unknown function									
4	Metabolism of carbohydrate									
2	Transcription regulation									
2	Metabolism of coenzymes and prosthetic groups									
1	Transposon and IS									
1	Transport/binding of carbohydrates									
1	Sensors (signal transduction)									
1	Metabolism of amino acids and related molecules									
>Reg25	**FBFL15_2509:FBFL15_2527**	**16741**	**2944579**	**2961320**	**17**	**1**	**65**	**Unknown**	**Yes**	**No**
11	Protein of unknown function							*(prot bacteriophage KVP40)		
2	Miscellaneous									
2	DNA restriction and modification (and repair) *									
1	Transposon and IS									
1	Transcription regulation									
>Reg26	**FBFL15_2778:FBFL15_2909**	**117186**	**3248241**	**3365427**	**93**	**35**	**82**	**Unknown/RETRON ?**	**Yes**	**No**
76	Protein of unknown function									
7	Transport/binding proteins and lipoproteins									
5	Miscellaneous									
3	Transposon and IS									
1	Phage-related function									
1	Adaptation to atypical conditions									
>Reg27	**FBFL15_3018:FBFL15_3033**	**17670**	**3484038**	**3501708**	**16**	**0**	**56**	**Metabolism of carbohydrate; EPS biosynthesis and export**	**No**	**No**
9	Protein of unknown function									
4	Metabolism of carbohydrate									
1	Transport/binding proteins and lipoproteins									
1	Transcription regulation									
1	Sensors (signal transduction)									

TABLE S3. Predicted secreted proteins harboring a conserved C-terminal domain in the FL-15 genome

Locus_tag	Description
FBFL15_2775	Probable M1 family metalloprotease precursor
FBFL15_2755	Protein of unknown function precursor; putative immunoreactive 84 kDa antigen
FBFL15_2734	Protein of unknown function precursor
FBFL15_2421	Protein of unknown function precursor
FBFL15_2422	Protein of unknown function precursor
FBFL15_2352	Probable M36 fungalysin family metalloprotease precursor
FBFL15_2375	Protein of unknown function precursor
FBFL15_2391	Probable glycoside hydrolase precursor, family 13
FBFL15_2274	Protein of unknown function precursor; putative adhesin
FBFL15_1976	Protein of unknown function precursor
FBFL15_1983	Protein of unknown function precursor
FBFL15_2011	Protein of unknown function precursor
FBFL15_1849	Protein of unknown function precursor; putative adhesin
FBFL15_1444	Protein of unknown function precursor
FBFL15_1456	RCC1 (Regulator of Chromosome Condensation) repeat domain protein precursor
FBFL15_1459	Hypothetical protein precursor
FBFL15_1118	Probable M12B family metalloprotease precursor
FBFL15_0894	Hypothetical protein precursor
FBFL15_0920	Protein of unknown function precursor; putative adhesin
FBFL15_1001	Protein of unknown function precursor
FBFL15_0699	Protein of unknown function precursor
FBFL15_0661	Putative outer membrane protein precursor
FBFL15_0601	Protein of unknown function precursor
FBFL15_0606	Protein of unknown function precursor
FBFL15_3030	Protein of unknown function precursor
FBFL15_1314	Protein of unknown function precursor
FBFL15_3033	Hypothetical protein precursor
FBFL15_0799	Protein of unknown function precursor; putative adhesin
FBFL15_2266	Probable S8 subtilisin family serine endopeptidase precursor
FBFL15_1376	Probable extracellular ribonuclease precursor
FBFL15_2171	Hypothetical protein precursor

TABLE S4. Predicted secreted glycoside hydrolases likely involved in carbohydrate catabolism in the FL-15 genome

Locus_tag	Description
FBFL15_2407	Levanase precursor. Glycoside hydrolase, family 32
FBFL15_2667	Glycoside hydrolase precursor, family 28
FBFL15_2391	Probable glycoside hydrolase precursor, family 13
FBFL15_0690	Glycoside hydrolase precursor, family 9
FBFL15_2664	Glycoside hydrolase precursor, family 88
FBFL15_2533	Putative alpha-1,2-mannosidase precursor. Glycoside hydrolase family 92
FBFL15_1325	Glycoside hydrolase precursor, family 3
FBFL15_1324	Glycoside hydrolase precursor, family 30
FBFL15_1154	Probable beta-N-acetylglucosaminidase precursor. Glycoside hydrolase family 3
FBFL15_2425	Probable glycoside hydrolase precursor
FBFL15_2455	Glycoside hydrolase precursor, family 53. Putative arabinogalactan endo-1,4-beta-galactosidase precursor
FBFL15_2612	Oligo-1,6-glucosidase precursor. Glycoside hydrolase family 13
FBFL15_2433	Glycoside hydrolase precursor, family 5
FBFL15_2661	Hypothetical protein precursor, putative glycoside hydrolase
FBFL15_2381	Probable glycoside hydrolase precursor, family 13
FBFL15_1323	Glycoside hydrolase precursor, family 30
FBFL15_1000	Beta-glucosidase precursor. Glycoside hydrolase family 3

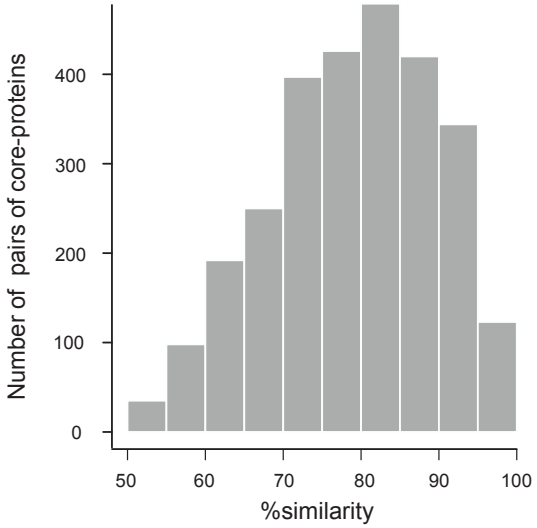


Figure S1: Distribution of the %similarity of all orthologs pairs of core-proteins. Orthologs were defined by identifying unique pairwise reciprocal best hits, with at least 50% similarity in amino acid sequence and less than 20% difference in protein length. The analysis of orthology was made for every pair of *Flavobacterium* genomes. The core genome, consisting of proteins ubiquitously found among all *Flavobacterium* genomes, was defined as the intersection of pairwise lists. The %similarity median is 80% [first quartile –third quartile: 72%-88%]

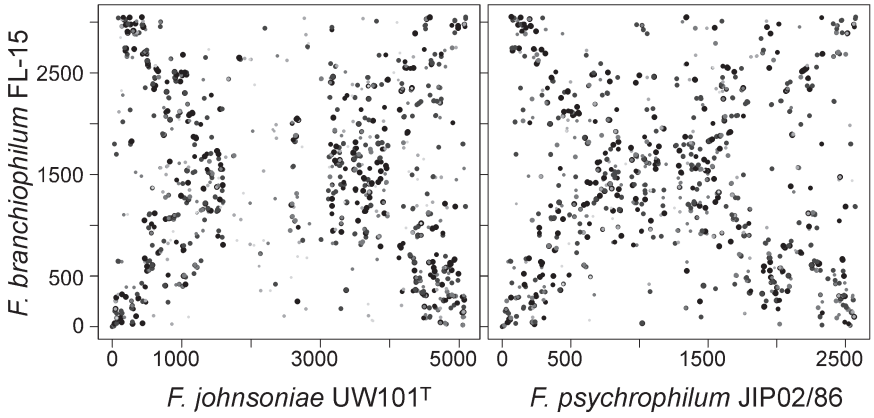


Figure S2: Plots of the position of orthologous genes between *F. branchiophilum* FL-15 and *F. johnsoniae* UW101^T (left) and between *F. branchiophilum* FL-15 and *F. psychrophilum* JIP02/86 (right). The size and gray level of each dot are proportional to the sequence similarity between orthologs.

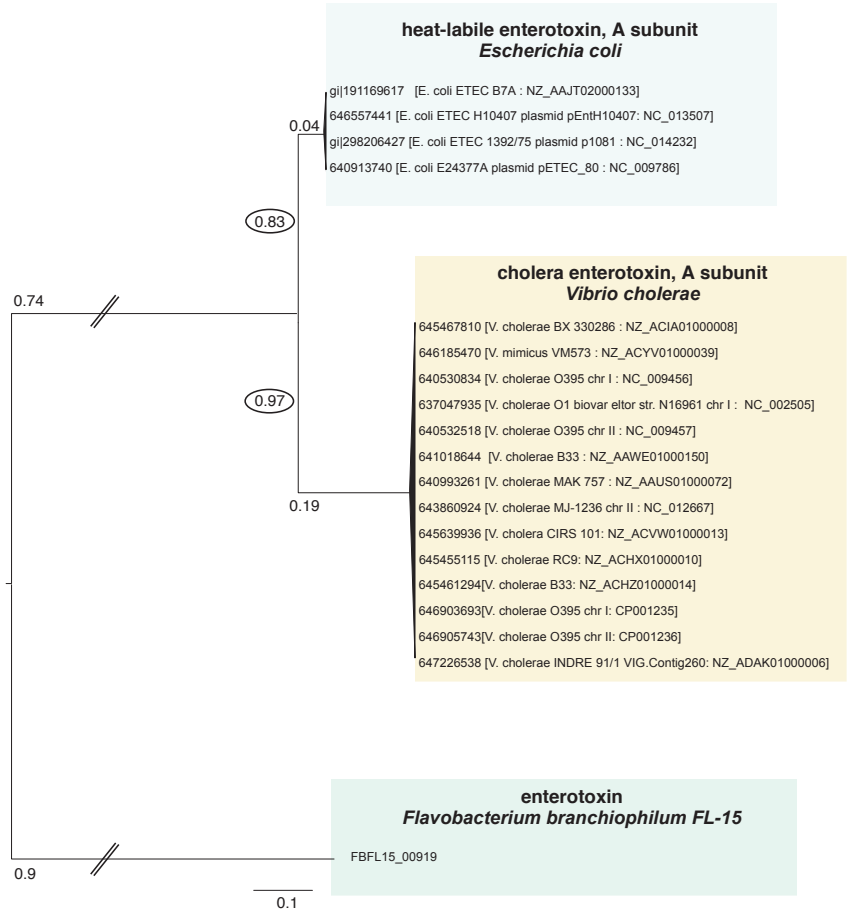


Figure S3: Molecular phylogeny of the enterotoxin homologs to FBFL15_0919. Phylogenetic tree for the enterotoxin proteins was performed using PhyML with the WAG+G model (17). Surrounded values correspond to aLRT values.

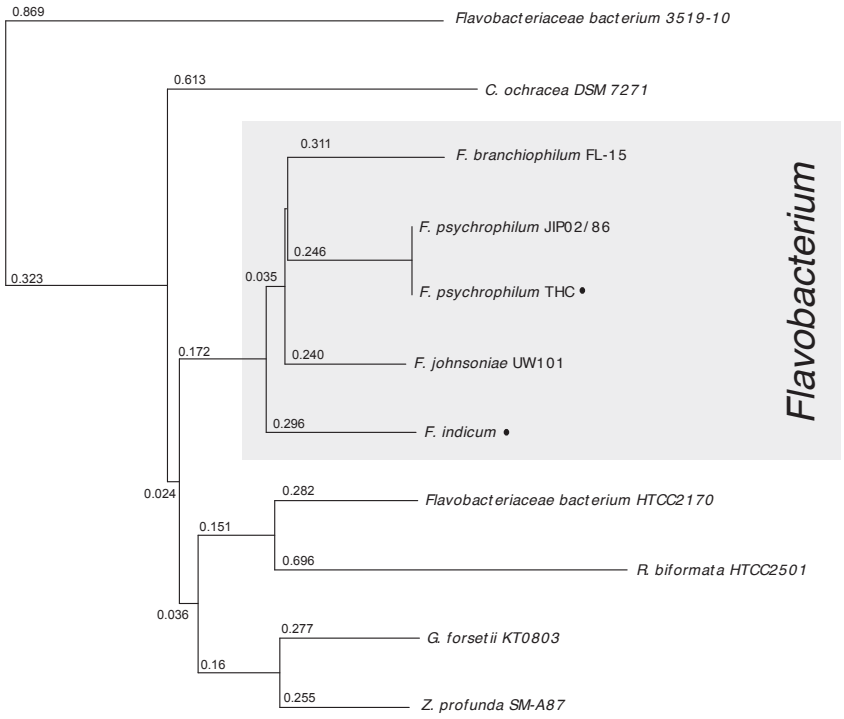


Figure S4: The tree was reconstructed from the concatenated alignments of 595 genes of the core genome of *Flavobacteriaceae* strains (see Methods). The main nodes of these branches were supported with high bootstrap values (>90%). *Flavobacterium* genus is indicated on the right part of the figure. black circles correspond to unpublished genomes.

Discussion

L'analyse du génome de *F. branchiophilum* FL15 révèle des déterminants de virulence probables remarquablement différents de ceux de l'autre espèce proche, également pathogène pour les poissons et en particulier pour les salmonidés, *F. psychrophilum*.

En effet, le génome de la souche FL15 est dépourvu de la plupart des gènes codant pour des toxines précédemment identifiées dans le génome de *F. psychrophilum* JIP 02/86 [6].

Toutefois, la souche FL15 contient un gène analogue à celui codant pour les enterotoxines exprimées par certaines souches d'*Escherichia coli* et *Vibrio cholerae*. Cette protéine contient un peptide signal indiquant qu'elle est probablement sécrétée et les acides aminés du site actif et du domaine de liaison au cofacteur NAD sont conservés entre cette protéine et les enterotoxines « heat-labile toxin » (LTA) exprimées par certaines souche d'*Escherichia coli*. Les LTA sont elles-mêmes très similaires aux toxines cholériques (CTA). Leur mode d'action moléculaire et leur implication dans la virulence d'*Escherichia coli* et *Vibrio cholerae* pour les mammifères ont déjà été étudiés [81], [82]. Il est donc possible que la protéine codée par le gène FBFL15_0919 ait un mode d'action similaire et soit responsable de la perturbation des fonctions osmorégulatrices des cellules épithéliales des branchies des poissons. C'est la première fois que ce probable analogue fonctionnel de la toxine cholérique est détecté en-dehors du phylum *Proteobacteria* mais une analyse par phylogénie moléculaire a cependant permis d'exclure la possibilité d'un transfert latéral direct entre les protéobactéries et la souche FL15 (Figure S3).

De nombreuses protéases extracellulaires, appartenant à différentes familles, ont été également identifiées dans le génome de la souche FL15. *F. branchiophilum* est présentée comme étant capable de dégrader des protéines [31] et les protéases extracellulaires jouent probablement un rôle important dans la virulence de deux autres pathogènes au sein du phylum *Bacteroidetes*, *F. psychrophilum* et *Porphyromonas gingivalis* [83], [84]. Il est donc probable que ces protéases sécrétées soient impliquées dans la dégradation des composés

Le tropisme branchial de *F. branchiophilum* est vraisemblablement lié aux capacités d'adhésion de la bactérie [33]. L'attachement des bactéries étant en effet essentiel à l'initiation d'une infection, le génome de la souche FL15 contient 20 gènes d'adhésines qui pourraient être impliqués dans les interactions avec les cellules de l'hôte.

La comparaison du génome de *F. branchiophilum* avec les autres génomes disponibles au sein du genre *Flavobacterium* a permis de mettre en évidence plusieurs éléments permettant d'expliquer comment cette bactérie s'est adaptée à son environnement.

Le mode de vie pathogène et l'adaptation sous-jacente de la bactérie à son hôte sont des conditions de vie très réduites. Comparée à la taille du génome de *F. johnsoniae*, la petite taille du génome de *F. branchiophilum* reflète probablement cette adaptation à une niche écologique restreinte. De plus, *F. branchiophilum* est une bactérie dont la culture au laboratoire est difficile et lente. Le petit nombre d'opérons d'ARN ribosomiques et d'ARN de transferts contenus dans son génome concorde avec cette observation [85].

Le chromosome de *F. branchiophilum* FL15 contient 2867 gènes. En utilisant une stratégie de comparaison des séquences protéiques prédites à partir de ces gènes et des gènes des autres génomes disponibles (*F. psychrophilum* et *F. johnsoniae*), il a été possible de préciser le génome central du genre *Flavobacterium*, composé d'environ 1400 gènes (Figure 2). La moitié environ de ces gènes sont impliqués dans le métabolisme central et les machineries cellulaires de transcription et de traduction. Ainsi, à part quelques spécificités, les éléments du métabolisme énergétique central de *F. branchiophilum* sont globalement similaires à ceux décrits chez *F. psychrophilum* et *F. johnsoniae* [6], [13]. Cependant, certaines voies de catabolisme des acides aminés, identifiées chez ces derniers, sont absentes du génome de *F. branchiophilum* et, inversement, dix gènes codant pour une voie métabolique de dégradation aérobie du phénylacétate ont été identifiés dans le génome de la souche FL15. Cette voie métabolique à été identifiée chez plusieurs pathogènes et le rôle dans leur virulence de certains intermédiaires toxiques à été proposée [86].

F. branchiophilum a été décrit comme étant capable d'hydrolyser l'amidon et d'utiliser de nombreux carbohydrates en conditions aérobies [31]. L'analyse du génome a permis d'identifier 23 glycosides hydrolases probablement impliquées dans le catabolisme

des carbohydrates (Table S4). Des systèmes de transport des oligomères à travers la membrane externe, codés par des groupes de gènes paralogues (*susCD*), sont souvent retrouvés à proximité de gènes impliqués dans l'utilisation des polysaccharides tels que les glycosides hydrolases, les polysaccharides lyases et les carbohydrates estérases. De tels systèmes d'utilisation des polysaccharides (PUL) sont communément présents chez les membres du phylum *Bacteroidetes* [87]. Le génome de *F. branchiophilum* contient 4 PULs alors que le génome de *F. psychrophilum* apparaît dénué de tels systèmes. Le génome de *F. johnsoniae*, organisme retrouvé dans les sols et la rhizosphère, en contient 33, notamment certains impliqués dans la digestion des carbohydrates issus de la paroi des cellules végétales [13]. Il est donc très surprenant d'identifier de tels locus dans le génome de la souche FL15, en particulier les gènes FBFL15_2659-2677 ressemblant à un locus de *F. johnsoniae* impliqué dans l'utilisation de la pectine [13], suggérant ainsi que *F. branchiophilum* est capable d'utiliser certains carbohydrates dérivés des plantes. Une autre hypothèse serait que *F. branchiophilum* est capable d'utiliser des polysaccharides du mucus des branchies ou les dérivés issus de leur dégradation.

La comparaison avec les génomes de *F. psychrophilum* et *F. johnsoniae* a permis d'identifier 1014 gènes spécifiques du génome de *F. branchiophilum* FL15 (Figure 2) dont plus de la moitié (52%) sont retrouvés dans des régions du chromosome constituées d'au moins dix de ces gènes. Nous avons ainsi pu identifier 27 grandes régions génomiques spécifiques de *F. branchiophilum* FL15 (Figure 1 et Table S2). En analysant les différentes fonctions codées par les gènes les constituant, il a été possible de dégager, au moins pour certaines d'entres elles, les rôles probables de ces régions dans la physiologie de la bactérie. Environ la moitié (14) des 27 régions n'ont pas de fonctions connues. Cependant, cinq régions pourraient être impliquées dans la détoxification ou la résistance à plusieurs composés tels que les métaux lourds, les antibiotiques et le peroxyde d'hydrogène ; quatre régions comporteraient des gènes impliqués dans la biosynthèse d'exopolysaccharides ; deux régions contiendraient des gènes impliqués dans l'import et le métabolisme de sucres ; une région comporte la voie métabolique de dégradation de l'acide phénylacétique ; et une région comporterait des gènes du métabolisme des lipides. Ce sont probablement ces régions et leur apport fonctionnel dans la physiologie de la bactérie qui permettent en partie d'expliquer son mode de vie. Il est également important de souligner que plus de la moitié (60%) des régions

spécifiques de *F. branchiophilum* FL15 ont vraisemblablement été acquises horizontalement et sont donc considérées comme étant des îlots génomiques.

Le génome complet de *F. branchiophilum* et la comparaison avec les deux autres génomes au sein du genre *Flavobacterium* disponibles à cette époque ont donc permis de mettre en évidence d'importantes différences en terme de contenus en gènes mais également en terme d'organisation chromosomique. En effet, on observe peu de conservation dans l'ordre des gènes du génome central entre les génomes de *F. psychrophilum*, *F. johnsoniae* et *F. branchiophilum* (Figure S2) suggérant d'importants réarrangements génomiques au sein du genre. Contrairement aux génomes de *F. psychrophilum* et *F. johnsoniae,* une faible distorsion de la composition en bases G et C (« GC skew ») a été observée dans le génome de *F. branchiophilum*, indiquant un vaste brassage du génome dans la lignée menant à *F. branchiophilum*.

Par séquençage du gène codant l'ARN ribosomal 16S, *F. branchiophilum* apparaît phylogénétiquement plus proche de *F. psychrophilum* que de *F. johnsoniae*. Cette proximité a été confirmée par une phylogénie réalisée sur la séquence concaténée des protéines du génome central (Figure S4). Cependant les caractères biochimiques de *F. branchiophilum* semblent plus proches de ceux de *F. johnsoniae* et les toxines et autres déterminants de virulence sont globalement différents de ceux identifiés chez *F. psychrophilum*. Ces éléments mettent en avant une évolution différente de la virulence chez ces deux pathogènes de salmonidés depuis l'ancêtre commun. On peut en effet penser que *F. branchiophilum* ressemblerait plus à un intermédiaire entre cet ancêtre commun et *F. psychrophilum*.

Les bactéries de la famille des *Flavobacteriaceae* sont les principales réceptrices des agents naturels de transferts de gènes dans les océans [88] et les espèces du genre *Flavobacterium* sont extrêmement sujettes aux recombinaisons homologues [7], [68]. En plus de l'accumulation des réarrangements successifs du génome, les nombreux îlots génomiques de *F. branchiophilum* FL15 suggèrent que le transfert de gènes a donc eu un rôle important dans la diversification de cette espèce.

Comprendre les mécanismes moléculaires qui confèrent à une bactérie pathogène l'aptitude à coloniser, envahir et modifier la physiologie de son hôte constitue un enjeu de première importance, tant d'un point de vue fondamental que d'un point de vue agronomique. L'analyse de ce génome a permis d'identifier des mécanismes moléculaires de virulence originaux. Il devrait également permettre de guider la conception de nouveaux outils de lutte contre les agents pathogènes piscicoles, par exemple dériver des outils diagnostiques spécifiques pour les différentes espèces pathogènes du genre.

Troisième partie : Etude du génome complet de *Flavobacterium indicum*, isolé d'une source chaude

Introduction

Les membres de la famille des *Flavobacteriaceae* présentent une importante diversité de modes de vie et sont fortement impliqués dans la dégradation des substrats organiques dans les écosystèmes terrestres et marins [65], [66].

Flavobacterium indicum est une bactérie environnementale considérée comme non pathogène. En considérant que beaucoup d'autres espèces du genre *Flavobacterium* sont retrouvées dans des milieux extrêmes comme les glaciers et les régions polaires, cette espèce est la seule thermophile connue au sein du genre. Le génome complet de *F. indicum* a été essentiellement séquencé pour contribuer à la connaissance scientifique en écologie microbienne et pour réaliser une étude de génomique comparative afin de compléter les informations sur la dynamique évolutive au sein du genre *Flavobacterium*. Son génome a été comparé aux autres génomes déjà disponibles : ceux de deux espèces pathogènes de poissons *F. psychrophilum* et *F. branchiophilum* [6], [48] et celui de l'espèce environnementale *F. johnsoniae* [13]. Ces nouvelles données ont notamment permis la caractérisation de marqueurs moléculaires de sa niche écologique.

Le séquençage du génome a été réalisé sur l'unique souche disponible de *F. indicum* (GPTSA100-9^T = CIP 109464^T), isolée en 2006 à partir d'un échantillon d'eau provenant d'une source chaude dans la province d'Assam en Inde dont la température est comprise entre 37 et 38 °C [23].

Le projet du séquençage du génome complet de *F. indicum* a été initié en 2009 et a bénéficié d'un contrat avec le génoscope d'Evry (CNS) pour la production des séquences. J'ai participé à la finition du génome (réalisation de divers assemblages intermédiaires), à la validation de l'assemblage final par carte optique (Figure C2), à l'annotation du génome et à son analyse. Ces étapes de finition, d'annotation et d'analyse ont été réalisées en collaboration avec les équipes Génomique Evolutive des Micro-organismes (Institut Pasteur/CNRS) et Mathématique, Informatique et Génome (INRA). Ce travail a été publié sous la forme d'un « Genome Announcement » dans *Journal of Bacteriology* et a fait l'objet d'une présentation orale à Turku en Finlande dont le résumé étendu est également présenté ci après. Des figures complémentaires, non-incluses dans la publication d'origine, sont également présentées.

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Complete Genome Sequence of *Flavobacterium indicum* GPSTA100-9T, Isolated from Warm Spring Water

Paul Barbier,[a] Armel Houel,[a] Valentin Loux,[b] Julie Poulain,[e] Jean-François Bernardet,[a] Marie Touchon,[c,d] and Eric Duchaud[a]

Unité de Virologie et Immunologie Moléculaires, INRA, Domaine de Vilvert, Jouy-en-Josas, France[a]; Unité de Mathématique, Informatique et Génome, INRA, Domaine de Vilvert, Jouy-en-Josas, France[b]; Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, Paris, France[c]; CNRS, Paris, France[d]; and Genoscope, Evry, France[e]

We report here the complete annotated genome sequence of *Flavobacterium indicum* CIP 109464^T (= GPTSA100-9^T), isolated from warm spring water in Assam, India. The genome sequence of *F. indicum* revealed a number of interesting features and genes in relation to its environmental lifestyle.

Members of the family *Flavobacteriaceae* occur in a variety of temperate and polar habitats in terrestrial, freshwater, and marine environments. Besides this interesting diversity of lifestyle, they have a very significant role in the degradation/turnover of the organic matter in these ecosystems (1, 2).

We determined the whole genome sequence of the type strain of *Flavobacterium indicum* (CIP 109464^T) (11) in order to perform comparative genomic studies between this environmental species and other *Flavobacterium* species whose genome sequences have been published: two fish-pathogenic species that severely impact aquaculture worldwide (1) (*F. psychrophilum* JIP02/86 [4] and *F. branchiophilum* FL15 [13]) and another environmental species, *F. johnsoniae* UW101^T (7), a model organism for characterizing gliding motility (8) and biopolymer utilization in oligotrophic environments (10).

The genome of *F. indicum* was sequenced using a combination of Sanger (ABI3730, Applied Biosystems; performed on a genomic DNA library with an average fragment length of 10 kbp cloned in pCNS) and 454 (GS-FLX, Roche) sequencing with 2.6-fold and 17.4-fold coverage, respectively. The 454 reads were assembled in 145 contigs using Newbler. These contigs and the Sanger reads were assembled in four scaffolds (39 contigs) using Phrap. Scaffolds were ordered using an optical map (OpGen Technologies) (6), and gaps were closed using primer walking on gap-spanning clones or by PCR sequencing. Genome annotation, including manual validation, was performed using the AGMIAL annotation platform (3).

The complete genome of *F. indicum* consists of a circular chromosome of 2,993,089 bp with an overall G+C content of 31.8%. The genome is predicted to carry 2,671 protein-coding genes, 55 tRNA genes, and four rRNA operons.

Genome comparison with the available genome sequences of other *Flavobacterium* species confirms a loss of synteny within the genus (7, 13), likely due to the presence of many repeats (e.g., insertion sequences and rhs elements). No CRISPR locus (5) was found in *F. indicum*, which also lacks the toxin-encoding genes previously identified in *F. psychrophilum* (4) and *F. branchiophilum* (13).

In relation to its environmental lifestyle, the *F. indicum* genome is predicted to encode 38 adhesins, likely used for binding on different surfaces, six glycoside hydrolase precursors, various endo- and exopeptidases, and one polysaccharide utilization system (PUL) (9), confirming its ability to degrade some macromolecules such as gelatin, casein, and starch (11). However, the *F. indicum* genome, which is 2-fold smaller than the 6-Mbp-long genome of *F. johnsoniae* (7), lacks the 1.93-Mbp region enriched for genes involved in polysaccharide utilization of the latter genome. This finding corroborates the weak biopolymer-degrading ability of *F. indicum* (1) likely related to a more restricted ecological niche. The gliding motility machinery (12) is probably not functional, as the *gldA* gene, encoding the gliding motor, is frameshifted and the *gldE* gene, involved in gliding, is absent, verifying the original description of *F. indicum* as a nongliding organism (11). In contrast with *F. psychrophilum* and *F. johnsoniae*, which contain flexirubin-type pigments, the yellowish-orange color of *F. indicum* (11) is attributable only to the presence of carotenoid biosynthesis genes.

Nucleotide sequence accession number. The annotated complete genome sequence of *F. indicum* CIP 109464^T reported in this paper is available in GenBank under the accession number HE774682.

ACKNOWLEDGMENTS

This work was supported in part by grant 07-GMGE from the Agence Nationale de la Recherche of France. P.B. is a Université Evry Val d'Essonne Ph.D. fellow.

We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources. We acknowledge C. Bizet (Collection de l'Institut Pasteur, Paris) for generously supplying the type strain of *F. indicum.*

REFERENCES

1. **Bernardet J-F, Bowman JP.** 2011. Genus I. Flavobacterium Bergey et al. 1923, p 112–154. *In* Whitman W (ed), Bergey's manual of systematic bacteriology, 2nd ed, vol 4. The Williams & Wilkins Co., Baltimore, MD.
2. **Brown MV, Bowman JP.** 2001. A molecular phylogenetic survey of sea-ice microbial communities (SIMCO). FEMS Microbiol. Ecol. **35**:267–275.
3. **Bryson K, et al.** 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. **34**:3533–3545.
4. **Duchaud E, et al.** 2007. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum.* Nat. Biotechnol. **25**:763–769.

Received 22 March 2012 **Accepted** 27 March 2012

Address correspondence to Eric Duchaud, eric.duchaud@jouy.inra.fr.

doi:10.1128/JB.00420-12

5. **Karginov FV, Hannon GJ.** 2010. The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol. Cell **37**(1):7–19.
6. **Latreille P, et al.** 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics **8**:321.
7. **McBride MJ, et al.** 2009. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. Appl. Environ. Microbiol. **75**:6864–6875.
8. **McBride MJ.** 2004. *Cytophaga-flavobacterium* gliding motility. J. Mol. Microbiol. Biotechnol. **7**(1–2):63–71.
9. **Reeves AR, D'Elia JN, Frias J, Salyers AA.** 1996. A *Bacteroides thetaiotamicron* outer membrane protein that is essential for utilization of maltooligosaccharides and starch. J. Bacteriol. **178**:823–830.
10. **Sack EL, van der Wielen PW, van der Kooij D.** 2011. *Flavobacterium johnsoniae* as a model organism for characterizing biopolymer utilization in oligotrophic freshwater environments. Appl. Environ. Microbiol. **77**: 6931–6938.
11. **Saha P, Chakrabarti T.** 2006. *Flavobacterium indicum* sp. nov., isolated from warm spring water in Assam, India. Int. J. Syst. Evol. Microbiol. **56**:2617–2621.
12. **Sato K, et al.** 2010. A protein secretion system linked to Bacteroidetes gliding motility and pathogenesis. Proc. Natl. Acad. Sci. U. S. A. **107**:276–281.
13. **Touchon M, et al.** 2011. Complete genome sequence of the fish pathogen *Flavobacterium branchiophilum.* Appl. Environ. Microbiol. **77**:7656–7662.

Complete Genome Sequence of *Flavobacterium indicum* GPSTA100-9^T, isolated from warm spring water

Paul Barbier[1], Jean-François Bernardet[1] & Eric Duchaud[1]

[1]*Unité de Virologie et Immunologie Moléculaires, INRA, Domaine de Vilvert, F-78350 Jouy-en-Josas, France*

Correspondence: pbarbier@jouy.inra.fr

Abstract

The number of described *Flavobacterium* species is constantly increasing and *Flavobacterium* strains occur in a wide range of ecological niches. *Flavobacterium indicum* is non-pathogenic and among the rare mesophilic bacterial species in the genus *Flavobacterium.* This environmental species, isolated from a warm spring water in Assam (India), is also of interest as a member of the *Flavobacteriaceae*, a bacterial family of high importance for the degradation/turnover of organic matter in terrestrial, freshwater and marine ecosystems [1,2]. The main objective of this study was to contribute to the scientific knowledge in the field of microbial ecology. We are using *Flavobacterium* diversity as a good context for comparative analysis of closely related organisms with different life-styles using genomic approaches. We determined the whole genome sequence of the type strain of *F. indicum* (GPSTA100-9^T = CIP 109464^T) [3] in order to perform comparative genomics studies between this environmental species and other *Flavobacterium* species whose genome sequences have been published: two fish-pathogenic species that severely impact aquaculture worldwide [1] (*F. psychrophilum* [4] and *F. branchiophilum* [5]) and another environmental species *F. johnsoniae* [6], a model organism for characterizing gliding motility [7] and biopolymer utilization in oligotrophic environments (*F. johnsoniae* strain A3 [8]).

The genome of the type strain of *F. indicum* was sequenced using a combination of Sanger (ABI3730, Applied Biosystem), performed on a genomic DNA library with an average fragment length of 10 kbp cloned in a low copies vector named pCNS; and 454 (GS-FLX, Roche) sequencing with a 2.6 fold and 17.4 fold coverage, respectively. The 454 reads were assembled in 145 contigs using Newbler. These contigs and the Sanger reads were assembled in 4 scaffolds (39 contigs) using Phrap. Scaffolds were ordered using an optical map (OpGen Technologies) [9] and gaps were closed using primer walking on gap-spanning clones or by PCR sequencing. A careful manual annotation of the genome was performed using the AGMIAL annotation platform [10].

The complete genome of *F. indicum* consists of a circular chromosome of 2,993,089 bp without plasmid with an overall G+C content of 31.8%. The genome is predicted to encode 2671 protein-coding genes, 55 tRNA genes and four rRNA operons. Genome comparisons allowed to define 23 large regions not found in previously sequenced *Flavobacterium* genomes. Twenty-two insertion sequences (IS) from four different families were identify and AlienHunter predicted 12 large regions horizontally acquired (Fig. 1), most of them carrying genes of unknown functions. Some phage scars were found in these regions. The whole genome of *F. indicum* will help in the definition of the *Flavobacterium* core-genome, which

contains most of essentials genes and important metabolic pathways shared by all sequenced genomes in the genus.

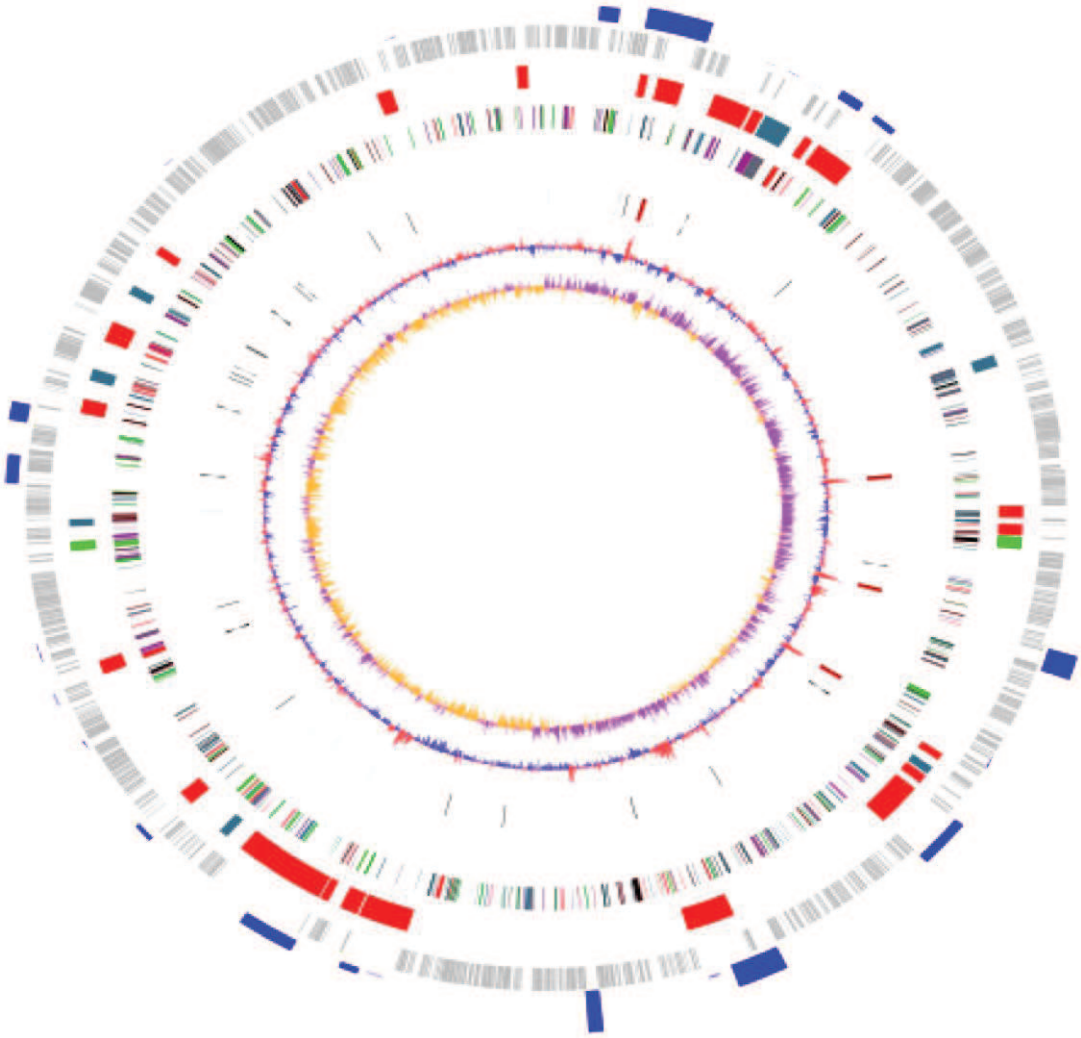


Figure 1. Circular representation of *F.indicum* whole genome. Circles represent the following (from the inside out): **1.** GC skew ; an inversion at the origin and the termini of replication is clearly visible; **2.** GC% local variations ; **3.** location of tRNA genes (black) and rRNA operons (red) ; **4.** Genes with orthologs in *F. psychrophilum* (green), *F. johnsoniae* (blue/grey) and *F. branchiophilum* (magenta) ; **5.** Regions with at least 10 genes not shared by all known *Flavobacterium* genomes. *F. indicum* specific regions (red), shared regions (at least 40% of genes) with *F. psychrophilum* (green) and *F. johnsoniae* (blue/grey) ; **6.** Core-genome genes in light grey ; **7.** Putative horizontal gene transfer regions (at least 10 genes) detected by Alien Hunter (blue).

Genome comparison with other *Flavobacterium* species confirms a loss of synteny at the genus level [5,6], likely due to the presence of many repeats (e.g. IS and rhs elements). Moreover, *F. indicum* genome, wich is two-fold smaller than the 6 Mbp-long genome of *F. johnsoniae* UW101^{T} [6], lacks the 2 Mbp region enriched in genes involved in polysaccharide utilization of the latter (Fig. 2). Among *Flavobacterium* environmental species that have been sequenced to date, *F. indicum* has the smallest genome. These findings corroborate the weak

biopolymer-degrading ability of *F. indicum* and tends to suggest that *F. indicum* is adapted to a narrow ecological niche.

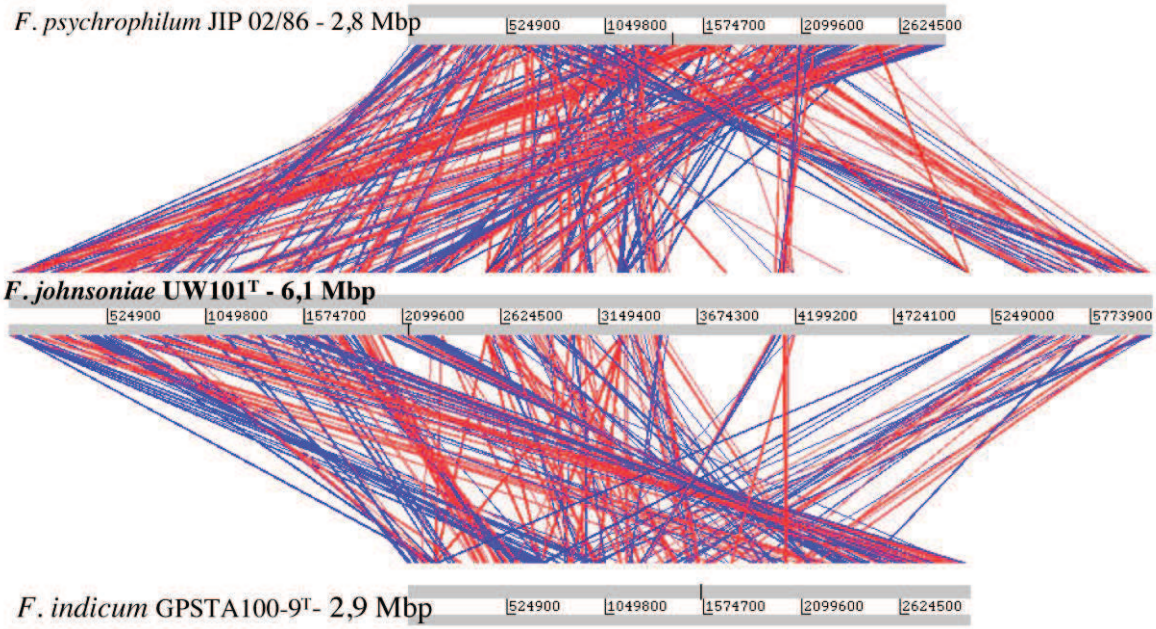


Figure 2. Nucleotidic pair-wise complete genomes comparison. The red and blue bands represent the forward and reverse matches, respectively.

The analysis of the *F. indicum* genome sequence revealed interesting features and genes in relation to its environmental life style. No CRISPR locus [11] was identified in *F. indicum,* which also lacks the toxin-encoding genes previously described in the three fish pathogenic species *F. psychrophilum* [4], *F. branchiophilum* [5] and *F. columnare* [12]. The genome is predicted to encode 38 adhesins, likely used for binding on different surfaces. Six glycoside hydrolase precursors, 50 endo- and exopeptidases and one polysaccharide utilization system (PUL) [13] were found likely in relation with its ability to degrade some macromolecules such as gelatin, casein, and starch [3]. We also identified genes encoding for secretion systems and for protein translocation across the outer membrane. We found 33 proteins with a conserved C-terminal sequence likely associated to the cell surface. In contrast with *F. psychrophilum* and *F. johnsoniae,* which contain flexirubin type pigments, the yellowish-orange color of *F. indicum* is only due to the presence of carotenoid pigments; genes for carotenoid biosynthesis were indeed found. Thirteen gliding motility machinery [14] encoding genes were also identified but the *gldA* gene, encoding the gliding motor, is frameshifted and the *gldE* gene, involved in gliding, is absent. Although not reported in the original description of *F. indicum* [3], weak gliding motility actually occurs [1].

This study provides a new genome within the genus *Flavobacterium* [15] and therefore gives additional information about genome dynamics within the genus. The small genome size and a weak biopolymer degrading ability tend to suggest an adaptation to a likely restricted ecological niche. Many features reveal the environmental life style of the bacterium: no toxins were identified and the 38 genes encoding for adhesins suggest that *F. indicum* mostly adhere to surfaces in its habitat.

References

[1] Bernardet J.-F., Bowman J.P., 2011. Genus I. *Flavobacterium* Bergey et al. 1923. In W. Whitman (ed.), Bergey's Manual of Systematic Bacteriology, pp. 112-154, 2nd ed., Vol. 4, The Williams & Wilkins Co., Baltimore

[2] Brown M.V., Bowman J.P., 2001. A molecular phylogenetic survey of sea-ice microbial communities (SIMCO). FEMS Microbiology Ecology, 35:267-275

[3] Saha P., Chakrabarti T., 2006. *Flavobacterium indicum* sp. nov., isolated from warm spring water in Assam, India. International Journal of Systematic and Evolutionary Microbiology, 56:2617-2621

[4] Duchaud E., Boussaha M., Loux V., Bernardet J.-F., Michel C., Kerouault B., Mondot S., Nicolas P., Bossy R., Caron C., Bessières P., Gibrat J.-F., Claverol S., Dumetz F., Le Hénaff M., Benmansour A., 2007. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. Nature Biotechnology, 25:763-769

[5] Touchon M., Barbier P., Bernardet J.-F., Loux V., Vacherie B., Barbe V., Rocha E.P., Duchaud E., 2011. Complete genome sequence of the fish pathogen *Flavobacterium branchiophilum*. Applied and Environmental Microbiology, 77:7656-7662

[6] McBride M.J., Xie G., Martens E.C., Lapidus A., Henrissat B., Rhodes R.G., Goltsman E., Wang W., Xu J., Hunnicutt D.W., Staroscik A.M., Hoover T.R., Cheng Y.Q., Stein J.L., 2009. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. Applied and Environmental Microbiology, 75:6864-6875

[7] McBride M.J., 2004. *Cytophaga-flavobacterium* gliding motility. Journal of Molecular Microbiology and Biotechnology, 7:63-71

[8] Sack E.L., van der Wielen P.W., van der Kooij D., 2011. *Flavobacterium johnsoniae* as a model organism for characterizing biopolymer utilization in oligotrophic freshwater environments. Applied and Environmental Microbiology, 77:6931-6938

[9] Latreille P., Norton S., Goldman B.S., Henkhaus J., Miller N., Barbazuk B., Bode H.B., Darby C., Du Z., Forst S., Gaudriault S., Goodner B., Goodrich-Blair H., Slater S., 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics, 8:321

[10] Bryson K., Loux V., Bossy R., Nicolas P., Chaillou S., van de Guchte M., Penaud S., Maguin E., Hoebeke M., Bessières P., Gibrat J.-F., 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Research, 34:3533-3545

[11] Karginov F.V., Hannon G.J., 2010. The CRISPR system: small RNA-guided defense in bacteria and archaea. Molecular Cell, 37:7-19.

[12] Tekedar H.C., Karsi A., Gillaspy A.F., Dyer D.W., Benton N.R., Zaitshik J., Vamenta S., Banes M.M., Gülsoy N., Aboko-Cole M., Waldbieser G.C., Lawrence M.L., 2012. Genome sequence of the fish pathogen *Flavobacterium columnare* ATCC 49512. Journal of Bacteriology, 194:2763-2764

[13] Reeves A.R., D'Elia J.N., Frias J., Salyers A.A., 1996. A *Bacteroides thetaiotamicron* outer membrane protein that is essential for utilization of maltooligosaccharides and starch. Journal of Bacteriology, 178:823-830

[14] Sato K., Naito M., Yukitake H., Hirakawa H., Shoji M., McBride M.J., Rhodes R.G., Nakayama K., 2010. A protein secretion system linked to *Bacteroidetes* gliding motility and pathogenesis. Proceedings of the National Academy of Sciences USA, 107:276-281

[15] Barbier P., Houel A., Loux V., Poulain J., Bernardet J.-F., Touchon M., Duchaud E., 2012. Complete genome sequence of *Flavobacterium indicum* GPSTA100-9^{T}, isolated from warm spring water. Journal of Bacteriology, 194:3024-3025

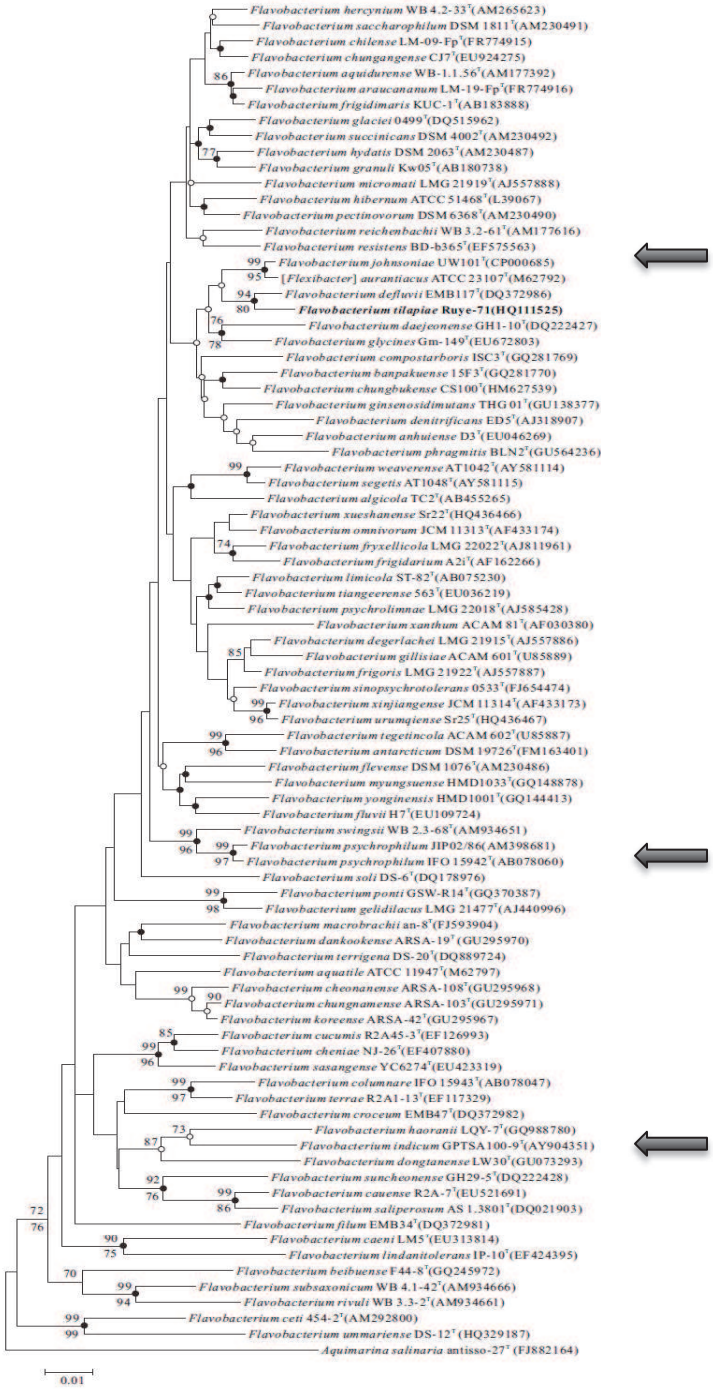
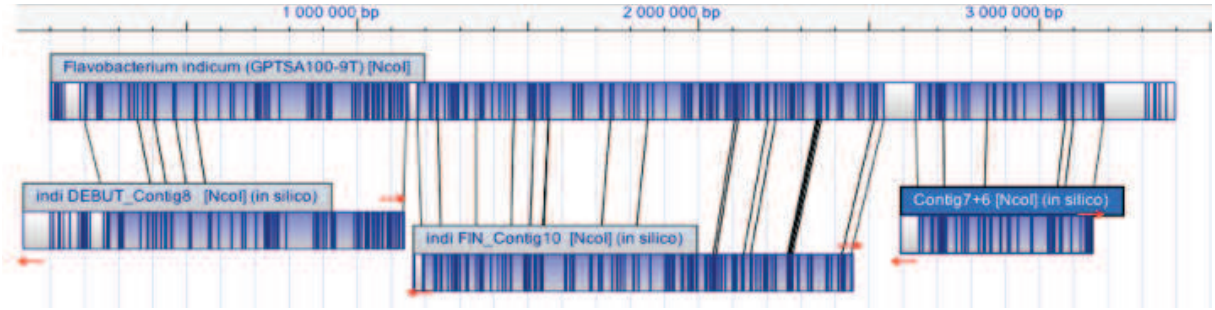


Figure Complémentaire 1 : Arbre phylogénétique du genre *Flavobacterium* construit par Neighbour-joining à partir du séquencage du gène codant la sous-unité 16S de l'ARN ribosomal. Les positions sur l'arbre de *F. indicum*, *F. psychrophilum* et *F. johnsoniae* sont indiquées par des flèches. Les nombres indiquent les valeurs de bootstrap (>70%) obtenues pour 1000 réplicats, par les méthodes d'échantillonnages neighbour-joining (au-dessus des noeuds) ou maximum-parsimony (en dessous des noeuds). *Aquimarina salinaria* est utilisé comme outgroup. L'échelle montre le nombre de substitutions par position nucléotidique. Extrait de: *Flavobacterium tilapiae* sp. nov., isolated from a freshwater pond, and emended descriptions of *Flavobacterium defluvii* and *Flavobacterium johnsoniae*. Chen et al. *Int. J. Syst. Evol. Microbiol*. 2013 63:827-834.

A



B

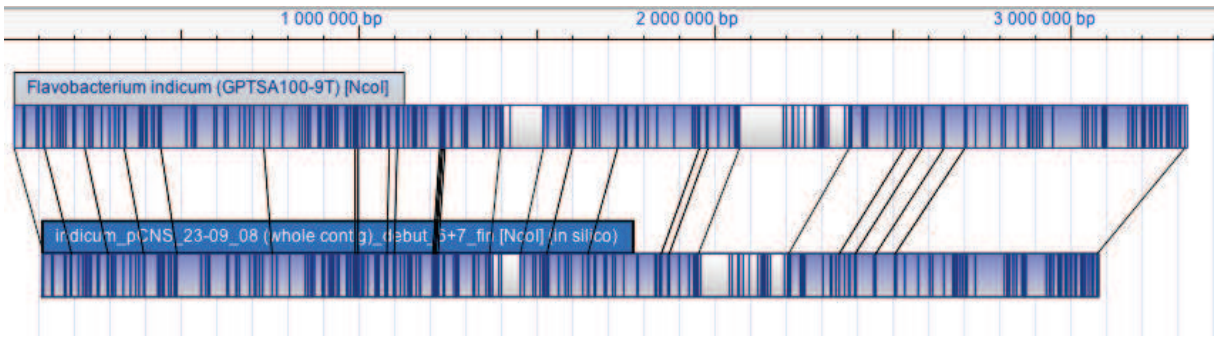


Figure Complémentaire 2 : Finition et validation de la séquence complète du génome de *Flavobacterium indicum* par carte optique. Le profil de restriction obtenu avec l'enzyme *Nco*I de la carte optique (en haut) est comparé au profil de restriction de l'assemblage généré avec la même enzyme *in silico* (en bas). Chaque trait vertical indique la position d'un site de restriction. Les fragments colorés en bleu sont reconnus de tailles identiques.

A : Détermination de l'ordre des contigs. Les flèches rouges indiquent les positions des oligonucléotides utilisés pour les PCR sur l'ADN génomique pendant la finition.

B : Validation de l'assemblage final par carte optique.

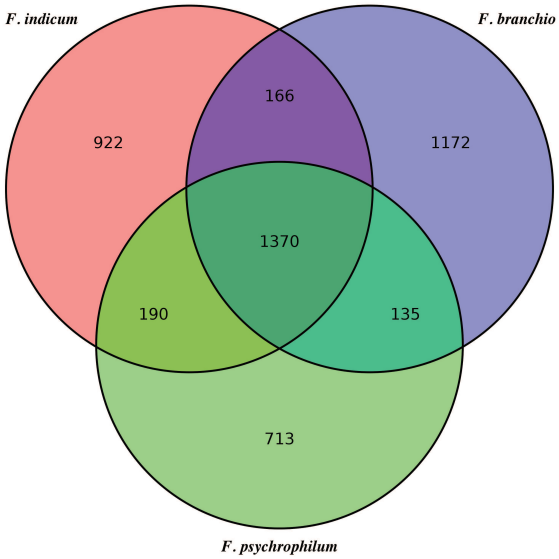


Figure Complémentaire 3 : Génome central du genre *Flavobacterium*.
Diagramme de Venn illustrant le chevauchement des répertoires de gènes entre *F. indicum* (en rouge), *F. psychrophilum* (en vert) et *F. branchiophilum* (en bleu). Les gènes orthologues ont été définit par « bidirectional best hit » (BDBH), avec les protéines comportant au moins 50% de similarité en acides aminés et moins de 20% de différence de longueur. Les gènes du génome central sont les orthologues retrouvés dans tous les génomes analysés et ont été déterminés par l'intersection des listes des comparaisons réalisées deux à deux. Le logiciel R a été utilisé pour la représentation graphique (représentation non-proportionelle).

Locus_tag	Description
KQS_00305	Probable glycoside hydrolase precursor
KQS_01055	Glycoside hydrolase precursor family 13
KQS_01095	Glycoside hydrolase group 97 family protein precursor, putative alpha-glucosidase
KQS_01100	Glycoside hydrolase precursor family 13
KQS_02290	Glycoside hydrolase precursor family 2
KQS_04245	Putative hybrid glycoside hydrolase, group 92 (20) family protein precursor

Table Complémentaire 1 : Précurseurs de glycosides hydrolases du génome de *F. indicum* probablement impliqués dans le catabolisme des carbohydrates.

Locus_tag	Description
KQS_08075	RCC1 (Regulator of Chromosome Condensation) repeat domain protein precursor
KQS_08080	RCC1 (Regulator of Chromosome Condensation) repeat domain protein precursor
KQS_08940	Protein of unknown function precursor
KQS_10725	Protein of unknown function precursor; putative adhesin
KQS_11500	Protein of unknown function precursor
KQS_11915	Hypothetical protein precursor
KQS_12070	Probable M36 fungalysin family metalloprotease precursor
KQS_12200	Psychrophilic metalloprotease Fpp2 precursor
KQS_12205	Psychrophilic metalloprotease Fpp1 precursor
KQS_12620	Protein of unknown function precursor
KQS_13110	Protein of unknown function precursor
KQS_13115	Protein of unknown function precursor
KQS_13950	Protein of unknown function precursor
KQS_00305	Probable glycoside hydrolase precursor
KQS_00365	Protein of unknown function precursor; putative immunoreactive 84 kDa antigen
KQS_05225	Protein of unknown function precursor
KQS_05690	Protein of unknown function precursor
KQS_05875	Protein of unknown function precursor; putative adhesin
KQS_07675	Probable protein of unknown function precursor
KQS_01055	Glycoside hydrolase precursor family 13
KQS_01135	Protein of unknown function precursor; putative adhesin
KQS_02660	Probable M4 thermolysin family metalloprotease precursor
KQS_02590	Protein of unknown function precursor; putative adhesin
KQS_02415	Protein of unknown function precursor
KQS_04320	Hypothetical protein precursor
KQS_04870	Protein of unknown function precursor. Putative S8A subfamily unassigned peptidases
KQS_02055	Protein of unknown function precursor; putative adhesin
KQS_09370	Protein of unknown function precursor; putative adhesin
KQS_09365	Protein of unknown function precursor; putative adhesin
KQS_05525	Protein of unknown function precursor; putative adhesin

Table Complémentaire 2 : Protéines probablement sécrétées de *F. indicum* contenants le domaine carboxy-terminal PorSS conservé.

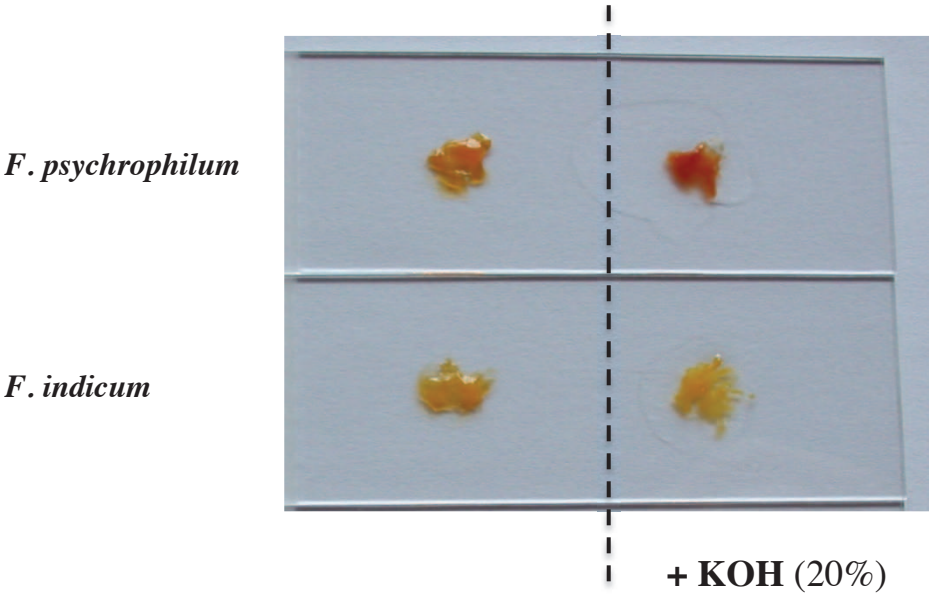


Figure Complémentaire 4 : Test de coloration à l'hydroxyde de potassium de colonies de *F. psychrophilum* et *F. indicum*. Une coloration rouge des bactéries après ajout du réactif (à droite) indique la présence de pigments de type flexirubines [9]. *F. psychrophilum*, pourvue de ces pigments, est montrée comme témoin positif. L'absence de coloration des colonies de *F. indicum* (en bas à droite) montre que cette espèce ne possède pas de pigments de type flexirubines.

Informations complémentaires et Discussion

Les bactéries du genre *Flavobacterium* colonisent des niches écologiques très diverses. Les trois génomes complets déjà disponibles à l'époque de ce travail (*F. psychrophilum*, *F. branchiophilum* et *F. johnsoniae*) sont situés sur la même branche de l'arbre phylogénétique du genre basé sur le séquençage du gène codant l'ARN ribosomal 16S, ainsi le génome de *F. indicum* a été choisi afin d'élargir cette diversité (Figure C1) et de servir à une étude de génomique comparative au sein du genre *Flavobacterium*. Son analyse a permis de mettre en avant des caractéristiques intéressantes et des gènes en relation avec son mode de vie « environnemental ».

La séquence complète a été obtenue grâce à un mélange de technologies de séquençage. La partie « shot-gun » a été réalisée en 454 (Roche) et la banque génomique, réalisée chez *E. coli* dans un plasmide à faible nombre de copies, a été séquencée par la méthode de Sanger. Une étape de finition, consistant en un séquençage des parties manquantes et un re-séquençage des parties de basse qualité, a été nécessaire. Les supers-contigs (« scaffolds ») ont ensuite été ordonnés à l'aide d'une carte optique et les parties manquantes de la séquence ont été obtenues par « marche » sur les clones chevauchants ces trous ou encore par séquençage des produits de PCR « long-range » réalisées directement sur l'ADN génomique (Figure C2). Après une étape d'annotation automatique, une étape minutieuse de validation et d'annotation manuelle a été nécessaire afin d'obtenir un génome complet de bonne qualité pouvant être utilisé pour des comparaisons.

Le génome complet de *F. indicum* consiste en un chromosome circulaire de 2, 9 Mpb sans plasmide et comporte 2671 gènes, 45 pseudo-gènes, 55 gènes de tRNA et 4 opérons d'ARN ribosomiques. Les comparaisons avec les autres génomes du genre ont permis de définir 23 grandes régions spécifiques de cet organisme. 22 IS provenant de quatre familles différentes ont été retrouvées [89] et 12 des 23 régions ont été prédites comme étant acquises

horizontalement [90] (Figure 1). La plupart contiennent des gènes de fonctions inconnues mais des traces de phages ont été identifiées dans ces régions. Cependant, l'une d'entre elles contient des gènes rarement retrouvés dans les génomes bactériens et leur étude s'est avérée extrêmement intéressante. Observés pour la première fois dans la famille des *Flavobacteriaceae* ces gènes et leur phénotype ont fait l'objet d'une publication (Article 3) [91].

La comparaison du génome de *F. indicum* avec les deux autres génomes du genre disponibles alors et séquencés au laboratoire (*F. psychrophilum* et *F. branchiophilum*) a permis de préciser le génome central du genre *Flavobacterium*, composé de 1370 gènes (Figure C3). Aujourd'hui, en ajoutant les génomes des espèces *F. johnsoniae*, *F. columnare* (publié depuis [92]), *F. frigidimaris* et *F. glaciei* (non publiés), le génome central du genre *Flavobacterium* contient environ 1000 gènes dont la majeure partie correspond aux gènes essentiellement impliqués dans les grandes fonctions cellulaires (métabolisme central, machineries de transcription, de traduction, de division, etc…)

Le génome complet de *F. indicum* a également permis de préciser les relations phylogénétiques entre les organismes déjà séquencés au sein de la famille des *Flavobacteriaceae*. L'arbre phylogénétique déduit de l'alignement des protéines concaténées du génome central de la famille (Figure S4, Article1) concorde avec ceux basés sur le séquençage du gène codant l'ARN ribosomal 16S et pourra aider à mieux permettre la définition d'espèces à l'aide d'approches génomiques.

La comparaison du génome de *F. indicum* avec d'autres génomes du genre *Flavobacterium* disponibles lors de ce travail ont permis de confirmer la perte de synténie observée au sein du genre [48]. Ces différences en terme d'organisation chromosomique suggèrent d'importants réarrangements génomiques probablement dus à la présence de nombreuses répétitions (IS et éléments rhs [93]).

Le génome de *F. indicum* est deux fois plus petit que le génome de 6,1 Mpb de *F. johnsoniae* et ne possède pas la région d'environ 2 Mpb enrichie en gènes impliqués dans l'utilisation des polysaccharides décrite chez cet organisme [13] (Figure 2). De plus, parmi les espèces du genre *Flavobacterium* isolées de l'environnement déjà séquencées, *F. indicum* a le plus petit génome (5,6 Mpb pour le génome de *F. frigidimaris* KUC-1, non publié). Ces observations corroborent la faible capacité de dégradation des bio-polymères de *F. indicum* [65] et tendent à suggérer qu'il est adapté à une niche écologique plus restreinte.

F. indicum est une bactérie non pathogène isolée de l'environnement. Comme attendu, les gènes codant pour des toxines dans les génomes des trois espèces pathogènes de poissons (*F. psychrophilum*, *F. branchiophilum* et *F. columnare* [6], [48], [92]) n'ont pas été retrouvés dans son génome. Cependant, des gènes apparaissant plus en relation avec son mode de vie « environnemental » ont pus être mis en évidence. Notamment six gènes codant des glycosides hydrolases probablement sécrétées (Table C1), ainsi qu'un système d'utilisation des polysaccharides (PUL) [87] (KQS_01055-01110 ; ressemblant au locus Fjoh_1398-1408 de *F. johnsoniae* probablement impliqué dans l'utilisation de l'amidon ou des α-glucans et dont l'expression est probablement régulée par un répresseur transcriptionel de type LacI [13]) ont été identifiés dans le génome de *F. indicum*. En plus des cinquante gènes codant pour des endo- et exopeptidases, ces différents gènes sont probablement en relation avec sa capacité à dégrader certaines macromolécules comme la gélatine, la caséine et l'amidon [23].

La sécrétion est mécanisme important dans la physiologie bactérienne permettant d'interagir avec l'environnement et plusieurs systèmes de sécrétion ont été retrouvés dans le génome de *F. indicum*. En effet, des gènes codant pour les systèmes de transport de type ABC [94], pour le système de transport Sec-dépendant ou encore pour le système Sec-indépendant (TAT) [95], [96] ont tous été identifiés dans le génome de *F. indicum*. Récemment décrit au sein du phylum *Bacteroidetes* [97], un système de translocation de protéines (PorSS) a également été identifié dans le génome de *F. indicum*. De plus, 30 protéines probablement sécrétées, incluant des peptidases, des glycosides hydrolases et des adhésines (Table C2), possèdent un domaine carboxy-terminal (CTD) conservé. Ce domaine, indiquant qu'elles sont probablement sécrétées via le système PorSS, est également retrouvé dans les protéines d'autres membres du phylum *Bacteroidetes* [98] et en particulier dans les

protéines d'autres membres du genre *Flavobacterium* [6], [48]. Ce CTD est impliqué dans la translocation des protéines du périplasme à la surface cellulaire et dans leur attachement à la membrane externe [99] en particulier de certaines adhésines, montrant également une implication de ce système de sécrétion dans la mobilité par glissement [97]. Il est intéressant de remarquer que 8 des 38 gènes codant pour des adhésines qui ont été identifiés dans le génome de *F. indicum* possèdent ce CTD. Les nombreuses adhésines identifiées sont probablement utiles à l'attachement de la bactérie à différentes surfaces et suggèrent un mode de vie fixé au substrat.

La mobilité par glissement est observée dans la plupart des espèces du genre *Flavobacterium* [65]. Deux groupes distincts de gènes, *gld* et *spr*, sont impliqués dans cette mobilité par glissement. Les gènes *gld* codent pour la machinerie moléculaire du système et les gènes *spr* codants pour des paralogues d'adhésines ont également un rôle dans cette mobilité [97], [100].

Treize gènes codant pour la machinerie de la mobilité par glissement (*gldABCDFGHIJKLMN*) et trois gènes codant pour les adhésines associés (*sprAET*) ont été identifiés dans le génome de *F. indicum*. Le décalage de phase (« frameschift ») dans le gène *gldA*, codant pour une des sous-unités du moteur moléculaire semblable aux ABC transporteur, et l'absence du gène *gldE*, impliqué dans le glissement, nous a conduit à formuler l'hypothèse que la machinerie de mobilité par glissement était probablement non-fonctionelle chez *F. indicum* et que cela confirmait la description originale de l'espèce comme étant un organisme non mobile [23].

Cependant, peu de temps après la publication du génome, l'analyse des gènes de glissement présents chez tous les membres du phylum *Bacteroidetes* capables de glisser a permis d'établir les gènes centraux du glissement (*gldBDHJKLMN* et *sprAET*). Cette étude a mis en évidence que les trois gènes encodant un homologue des ABC transporteurs (*gldA*, *gldF* et *gldG*), requis pour le glissement chez *F. johnsoniae*, étaient absents chez deux proches bactéries mobiles, suggérant que le rôle de ce transporteur n'était probablement pas aussi central dans le glissement que précédemment supposé [101]. Les auteurs remarquent également que la mobilité par glissement est souvent mal rapportée lors de la description

d'espèces au sein du phylum *Bacteroidetes* et que le glissement de *F. indicum* a été observé [102]. Ainsi il est vraisemblable que lors de la description de l'espèce comme lors de nos observations au microscope il n'est pas été possible de produire les conditions expérimentales nécessaire à l'observation du glissement. De même, le génome de *F. branchiophilum* contient les quatorze gènes *gld* (*gldA-N*) et les gènes *spr* (*sprAET*) mais cet organisme a été décrit comme étant dépourvu de mobilité par glissement. Cela suggèrerait que *F. branchiophilum* pourrait également être mobile mais que les conditions expérimentales utilisées jusqu'ici ont échoué à reproduire les conditions dans lesquelles le glissement se manifeste [48], [65].

Au sein du genre *Flavobacterium*, on retrouve des organismes produisant des pigments de type caroténoïdes ou flexirubines ainsi que des espèces produisant les deux types de pigments [8]. Ces pigments sont responsables de l'aspect jaune pâle, jaune vif ou orange des cultures.

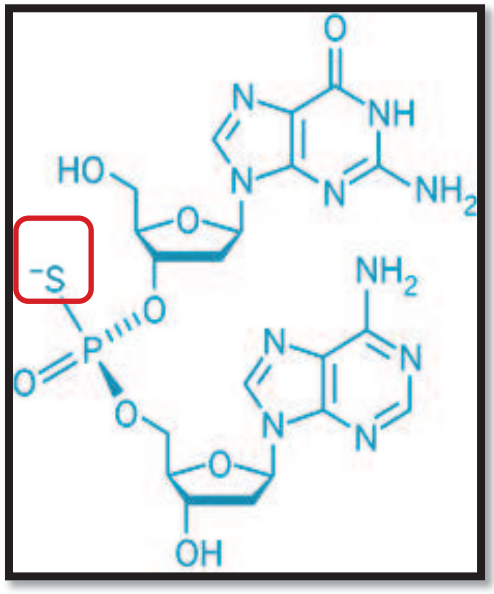
Contrairement à *F. psychrophilum* et à *F. johnsoniae*, *F. indicum* ne produit pas de pigments de type flexirubines [23] (Figure C4). En effet, son génome est dépourvu des gènes de biosynthèse des flexirubines identifiés dans les génomes de *F. psychrophilum* et *F. johnsoniae* [7], [13]. Néanmoins, les gènes *crtIBZY* codant pour les enzymes (phytoène synthase, phytoène déhydrogénase, lycopène cyclase et carotène hydroxylase, respectivement) responsables de la biosynthèse de caroténoïdes, ont été retrouvés dans le génome. Cette observation a également été réalisée dans le génome de *F. branchiophilum* [48], suggérant que la couleur jaune/orange des cultures de ces deux organismes serait uniquement due à la présence de pigments de type caroténoïdes.

Cette étude a permis de rapporter un nouveau génome complet au sein du genre *Flavobacterium*. Ces nouvelles données ont notamment permis la caractérisation de marqueurs moléculaires de sa niche écologique. Plusieurs caractéristiques du génome révèlent un mode de vie « environnemental ». La petite taille du génome et ses faibles capacités de dégradation des bio-polymères tendent à suggérer que *F. indicum* est adapté à une niche écologique restreinte.

Quatrième partie : Etude d'un groupe de gènes particulier : les gènes *dnd* ; analyse de leur distribution au sein du phylum *Bacteroidetes*

Introduction

Les cinq éléments azote, phosphore, carbone, hydrogène et oxygène ont toujours été considérés comme étant les composants canoniques de l'ADN. La récente découverte d'un sixième élément, le soufre, a bouleversé ce dogme. La phosphorothioation de l'ADN est la première modification physiologique décrite du squelette de l'ADN. L'atome de soufre remplace un atome d'oxygène non-liant du phosphate de la liaison phospho-diester (Figure ci-contre). Cette modification a lieu sur les deux brins après la réplication de l'ADN, de manière stéréospécifique (configuration chirale R des phosphates modifiés) et sélective (le contexte nucléotidique de la modification peut être propre aux souches bactériennes) [103], [104].



La phosphorothioation de l'ADN a été découverte suite à l'identification des cinq gènes *dnd* (*dndABCDE*) chez *Streptomyces lividans* [105], responsables du phénotype Dnd (pour « DNA degradation ») [106]. La dégradation chimique de l'ADN pendant l'électrophorèse est provoquée par l'attaque nucléophile des molécules du tampon de migration TRIS, activées sous l'action du champ électrique, sur les atomes de soufre, provoquant des cassures double brins dans la molécule d'ADN.

Les gènes *dnd* ont depuis été identifiés dans très peu d'espèces de bactéries et d'archéobactéries, sans relations phylogénétiques [104], [107]. Les différents contextes nucléotidiques phosphorothioatés ont été corrélés à la phylogénie des protéines Dnd mais pas à la phylogénie, basée sur le 16S, des différentes espèces concernées, suggérant que la dissémination des gènes *dnd* au sein des génomes bactériens avait lieu par transferts horizontaux [104].

Cependant, le rôle fonctionnel des modifications par phosphorothioation de l'ADN dans la physiologie bactérienne reste aujourd'hui obscur. En effet, certains auteurs ont proposé leur implication dans la régulation de l'expression des gènes [108], [109], probablement comme un signal épi-génétique [110]. D'autres auteurs ont montré l'implication des gènes *dnd* dans un système de modification couplé à un nouveau système de restriction protégeant le génome bactérien endogène de l'introduction d'ADN étranger [104], [109], [111], soutenant ainsi la comparaison avec les systèmes de restriction/modification basés sur la méthylation de l'ADN. Enfin, un article récent a démontré que la phosphorothiation peut également protéger l'ADN bactérien du stress oxydatif [112].

Au cours de différents projets, nous avons remarqué une dégradation de l'ADN génomique des souches de *F. indicum* CIP 109464^{T} et *F. psychrophilum* KU060626-59 lors de leur migration en gel d'électrophorèse. Nous avons longtemps suspecté que cette dégradation de l'ADN génomique était conséquence d'une contamination par des nucléases. Lors de l'annotation du génome de *F. indicum*, un groupe de gènes a attiré notre attention. Une étude de la littérature sur la fonction de ces gènes nous a permis d'établir un lien logique avec les phénotypes de dégradation observés.

Cette observation originale, réalisée dans le cadre d'une thèse essentiellement construite autour d'approches de génomique analytique, a permis la caractérisation d'éléments moléculaires marqueurs de caractères phénotypiques. Ces travaux ont été récemment publiés dans *FEMS Microbiology Letters*.

RESEARCH LETTER

From the *Flavobacterium* genus to the phylum *Bacteroidetes*: genomic analysis of *dnd* gene clusters

Paul Barbier[1], Aurélie Lunazzi[1], Erina Fujiwara-Nagata[2], Ruben Avendaño-Herrera[3,4], Jean-François Bernardet[1], Marie Touchon[5,6] & Eric Duchaud[1]

[1]INRA, Virologie et Immunologie Moléculaires UR892, Jouy-en-Josas, France; [2]Department of Fisheries, Kinki University, Nara, Japan; [3]Laboratorio de Patología de Organismos Acuáticos y Biotecnología Acuícola, Facultad de Ciencias Biológicas, Universidad Andrés Bello, Viña del Mar, Chile; [4]Interdisciplinary Center for Aquaculture Research (INCAR), Concepción, Chile; [5]Microbial Evolutionary Genomics, Institut Pasteur, Paris, France; and [6]CNRS, UMR3525, Paris, France

Correspondence: Eric Duchaud, INRA, Virologie et Immunologie Moléculaires UR892, Jouy-en-Josas, France.
Tel.: +33 1 34 65 25 88;
fax: +33 1 34 65 25 91;
e-mail: educhaud@jouy.inra.fr

Received 24 June 2013; revised 12 August 2013; accepted 18 August 2013.

DOI: 10.1111/1574-6968.12239

Editor: Michael Galperin

Keywords
Dnd phenotype; phosphorothioate DNA; DndEi; *Flavobacteriaceae*; *Flavobacterium indicum*; *Flavobacterium psychrophilum*.

Abstract

Phosphorothioate modification of DNA and the corresponding DNA degradation (Dnd) phenotype that occurs during gel electrophoresis are caused by *dnd* genes. Although widely distributed among Bacteria and Archaea, *dnd* genes have been found in only very few, taxonomically unrelated, bacterial species so far. Here, we report the presence of *dnd* genes and their associated Dnd phenotype in two *Flavobacterium* species. Comparison with *dnd* gene clusters previously described led us to report a noncanonical genetic organization and to identify a gene likely encoding a hybrid DndE protein. Hence, we showed that *dnd* genes are also present in members of the family *Flavobacteriaceae*, a bacterial group occurring in a variety of habitats with an interesting diversity of lifestyle. Two main types of genomic organization of *dnd* loci were uncovered probably denoting their spreading in the phylum *Bacteroidetes* via distinct genetic transfer events.

Introduction

Most members of the family *Flavobacteriaceae* (hereafter designated flavobacteria), a prominent member of the phylum *Bacteroidetes*, are free-living organisms occurring in a variety of temperate and polar habitats in terrestrial, freshwater, and marine environments (Bernardet, 2011). Because of their ecological significance (Brown & Bowman, 2001; Kirchman, 2002; Horner-Devine *et al.*, 2003), a large effort was recently undertaken to unravel their life styles including the use of genomic and metagenomic approaches (Venter *et al.*, 2004; Bauer *et al.*, 2006; Gómez-Consarnau *et al.*, 2007; González *et al.*, 2008; Oh *et al.*, 2010).

The original identification of a five-gene (*dndA-E*) cluster in *Streptomyces lividans* (Zhou *et al.*, 2005), which caused DNA degradation during electrophoresis (Zhou *et al.*, 1988), has led to the discovery of phosphorothioate (PT) DNA modification, in which sulfur replaces a nonbridging phosphate oxygen (Wang *et al.*, 2007). Yet, *dnd* genes have been identified in very few, taxonomically unrelated, bacterial and archaeal species so far (Ou *et al.*, 2009; Wang *et al.*, 2011). Phylogenies of Dnd proteins correlated with the PT sequence context strongly support the spreading of *dnd* genes by horizontal transfer (Wang *et al.*, 2011). However, the functional role of PT modifications still remains unclear. Some studies have proposed their involvement in (1) the regulation of gene expression (Eckstein, 2007; Xu *et al.*, 2010), (2) a new restriction–modification system protecting endogenous bacterial genome against foreign DNA introduction (Xu *et al.*, 2010; Wang *et al.*, 2011), and (3) the protection of bacterial DNA against oxidation (Xie *et al.*, 2012).

Our group has a long-standing history in the study of flavobacteria and is currently involved in genomic projects on fish-pathogenic and environmental *Flavobacterium* species. During distinct projects, we identified a Dnd phenotype during gel migration in two bacterial strains belonging to two different *Flavobacterium* species. As they had not been reported yet in flavobacteria, we investigated the presence of *dnd* gene clusters and performed genomic comparisons that allow to identify noncanonical *dnd* gene clusters in members of the phylum *Bacteroidetes*.

Materials and methods

Bacterial strains and growth conditions

Flavobacterium indicum CIP 109464^T was isolated from a warm spring in India (Saha & Chakrabarti, 2006), and *F. psychrophilum* KU060626-59 was isolated from a diseased ayu (*Plecoglossus altivelis*) in Japan (Fujiwara-Nagata *et al.*, 2012). *F. indicum* CIP 109464^T was grown in AOBE (Bernardet & Kerouault, 1989) for 24 h at 37 °C with shaking. *F. psychrophilum* KU060626-59, JIP 02/86, and ten other strains used as references were grown in AOBE for 48 h at 18 °C with shaking.

Pulse-field gel electrophoresis (PFGE)

Preparation of DNA in agarose blocks and restriction enzyme digestions were performed according to the protocol described by Liu & Sanderson (1995). Briefly, approximately 10^8 bacteria cells were embedded in Seaplaque agarose plugs (Lonza, Rockland, MI) and digested with proteinase K (Euromedex) for 10 h at 50 °C. Plugs were then washed in Tris–EDTA buffer with 1X protease inhibitor cocktail, and agarose-plug-embedded DNA were digested with 30 U of *I-Ceu*I or *Xho*I (New England Biolabs) for 8 h at 37 °C.

PFGE was performed using a CHEF-DR III System (Bio-Rad, Hercules, CA). Digested plugs were run on 1% Seakem agarose (Lonza, Rockland, MI) gels in 1X TBE at 14 °C with an electric field of 6V.cm^{-1} alternating in two directions at an incident angle of 120° with 1-pulse times ramped from 50 to 70 s over 17 h followed by pulse times ramped from 3 to 12 s over 6 h for the *I-Ceu*I-digested genomic DNA analysis (Fig. 1a) and 2-pulse times ramped from 1 to 3 s over 18 h for the *Xho*I-digested genomic DNA analysis (Fig. 1b). Gels were

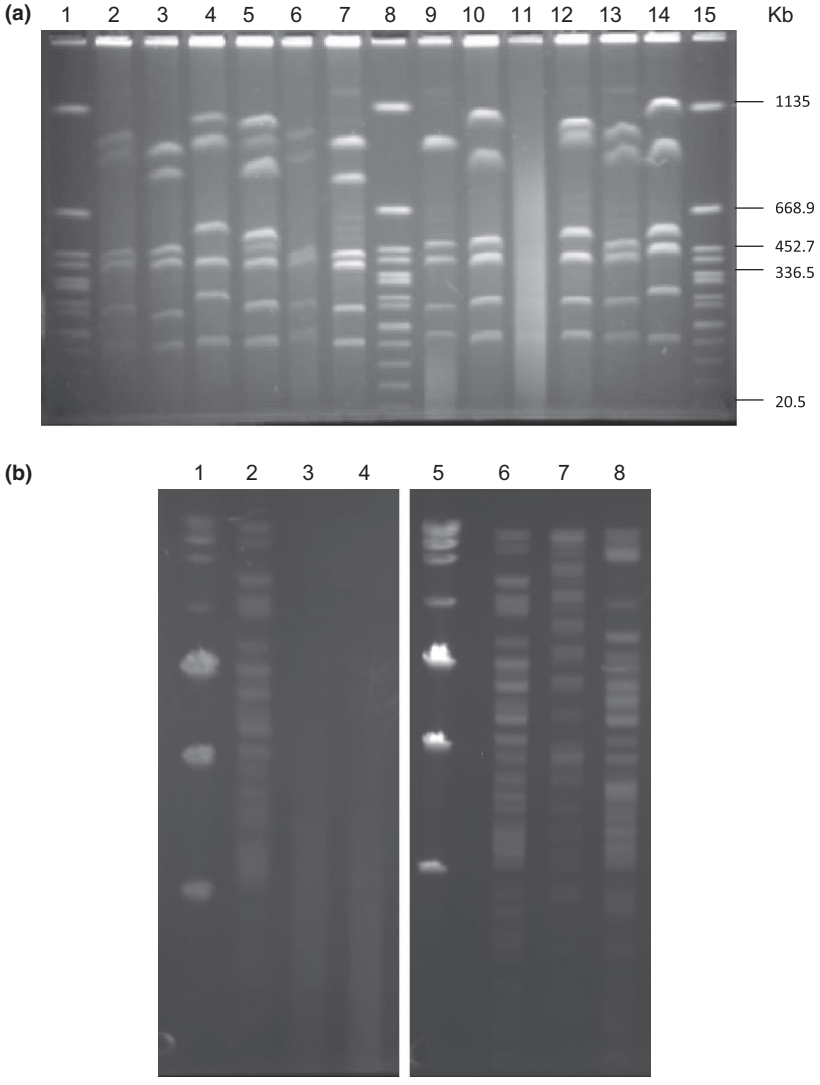


Fig. 1. (a) PFGE banding patterns of twelve *Flavobacterium psychrophilum* isolates after chromosomal DNA digestion with the *I-Ceu*I restriction enzyme. Strain KU060626-59 (lane 11) displays the typical DNA degradation phenotype (a smear pattern). Lanes 1, 8, and 15 : DNA of *Salmonella* Braenderup cleaved by *Xba*I was used as molecular size marker. Numbers on the right indicate size marker in kbp. (b) PFGE patterns of *Xho*I-digested genomic DNA displaying the typical degradation phenotype (left part) and same DNA samples run with thiourea (right part). Lanes 2 and 6, *F. psychrophilum* JIP 02/86; lanes 3 and 7, *F. indicum* GPSTA100-9$^{T;}$ lanes 4 and 8, *F. psychrophilum* KU060626-59. Lanes 1 and 5, mid-range PFG molecular marker II (New England Biolabs).

stained with 0.5 mg mL^{-1} ethidium bromide for 30 min before digital analysis (Storm; GE Healthcare).

Test of Dnd phenotype

To assess the Dnd phenotype, methods described by Ray *et al.* (1992, 1995) were used: that is, gel electrophoresis performed with Tris buffer and the use of 100 μM of thiourea to reverse the DNA degradation phenotype. The DNA of *F. psychrophilum* JIP 02/86 (Dnd^- and devoid of *dnd* genes) was used as a negative control.

RT-PCR analysis of the *dnd* genes transcripts

The total RNA of *F. indicum* CIP 109464^T was obtained from 6 mL of early-stationary-phase cultures. RNA was isolated using a guanidinium thiocyanate–phenol–chloroform extraction (TRIzol; Invitrogen) and treated with DNase I (RNase-free; Ambion). PCRs using 16S rRNA gene-specific primers were performed to determine whether RNA was free of contaminant DNA. Reverse transcription was performed using Superscript II Reverse Transcriptase (Invitrogen) and PCR with GoTaq DNA polymerase (Promega). Ten nanograms of RNA was used in each reaction. Primers used are listed in Supporting Information, Table S1.

DNA sequencing and sequence analysis

The genome sequencing of *F. psychrophilum* KU060626-59 was performed using a Solexa (Illumina, Inc., CA) single-end strategy. Reads were cleaned by adaptive trimming (home-made tool available at: http://migale.jouy.inra.fr) and assembled using VELVET (Zerbino & Birney, 2008). ORFs predictions and genome annotation were performed using the AGMIAL annotation platform (Bryson *et al.*, 2006) as previously described (Touchon *et al.*, 2011). The *de novo* assembly resulted in 55-fold coverage of a 2 588 113-bp draft genome encompassed in 254 contigs. The annotated sequence of *F. psychrophilum* KU060626-59 *dnd* gene cluster is available in GenBank under the accession number KF241852.

Identification of *dnd* gene clusters in *Bacteroidetes* and phylogenetic analysis

Flavobacterium indicum Dnd proteins were initially identified by BLASTP against the dedicated dnd database (Ou *et al.*, 2009). *Bacteroidetes* Dnd proteins were then identified by BLASTP (with e-value $< 1 \times 10^{-10}$) using Dnd proteins of *F. indicum* as a query against the 'nonredundant' database (April 2013) restricted to members of the phylum *Bacteroidetes*. Loci containing at least two of the five Dnd-protein-encoding genes confined in less than ten genes from each other were selected. Proteins homologies were confirmed by global alignments with at least 50% of similarity in amino acid sequence and < 10% of difference in protein length.

The molecular phylogeny of Dnd proteins has been explored by the construction of multiple sequence alignments with MUSCLE (version 3.6) (Edgar, 2004) and filtered with Gblocks (version 0.91b) (Castresana, 2000). The phylogenetic tree was reconstructed from the concatenated alignments of DndC and DndD proteins using the maximum-likelihood method implemented in the PHYML program (version 3.0) (Guindon & Gascuel, 2003) with the WAG matrix and a gamma correction for variable evolutionary rates. The reference tree was reconstructed from the 16S rRNA gene sequences of species of the phylum *Bacteroidetes* carrying a *dnd* gene cluster. The phylogenetic tree was built using PhyML (version 3.0) under the GTR model with gamma correction. The robustness of the tree topologies was assessed with 100 bootstraps.

Results

Identification of functional *dnd* gene clusters in *Flavobacterium* species

During genomic DNA preparations, we noticed that two *Flavobacterium* strains displayed a Dnd phenotype, which results in a smear pattern, following gel electrophoresis. It was the case of *F. psychrophilum* KU060626-59 after DNA restriction by *I-Ceu*I, the only one strain displaying a DNA smear pattern of the 12 *F. psychrophilum* stains tested (Fig. 1a). This phenotype was also observed for *F. indicum* CIP 109464^T. Indeed, *F. indicum* CIP 109464^T and *F. psychrophilum* KU060626-59 after DNA restriction by *Xho*I and following gel electrophoresis display a degradation phenotype (Fig. 1b – left part). This phenotype was maintained even with formaldehyde fixation (not shown), indicating that DNA degradation was not caused by extracellular DNase (Soto *et al.*, 2008). However, the observed degradation phenotype was completely abolished when the same DNA samples were run in the presence of thiourea (Fig. 1b – right part), as reported with PT DNA modification (Ray *et al.*, 1992, 1995). Taken together, these results indicate that *F.indicum* CIP 109464^T and *F. psychrophilum* KU060626-59 harbor a typical DNA degradation phenotype triggered by PT DNA modification.

During annotation of the *F. indicum* CIP 109464^T complete genome (Barbier *et al.*, 2012), we identified a gene cluster encoding proteins similar to Dnd proteins previously shown to be responsible of the Dnd phenotype (Zhou *et al.*, 2005). This gene cluster, encompassed in a 9450-bp region, contains five genes encoding

proteins homologous to DndABCDE (Table S2) and two additional genes of unknown function (Fig. S1). Strikingly, when comparing this locus with previously reported *dnd* clusters (Ou *et al.*, 2009), despite protein sequence similarity, gene order is not conserved. According to Alien Hunter prediction (Vernikos & Parkhill, 2006), this locus lies within a typical genomic island of 26 kbp inserted at a tRNA-Tyr gene suggesting a horizontally acquired origin.

To confirm our hypothesis that the degradation phenotype observed in *F. psychrophilum* KU060626-59 was triggered by a *dnd* locus, we performed the whole-genome shotgun of this strain. As expected, we identified on the draft genome a *dnd* gene cluster. This gene cluster contains *dndCDE* homologous genes and three additional genes of unknown function, different from those found in the *F. indicum* genome (Fig. S1). This gene cluster is encompassed in a 7732-bp contig with no sequence homology to the genome of *F. psychrophilum* JIP 02/86 (Duchaud *et al.*, 2007), suggesting its presence within a genomic island.

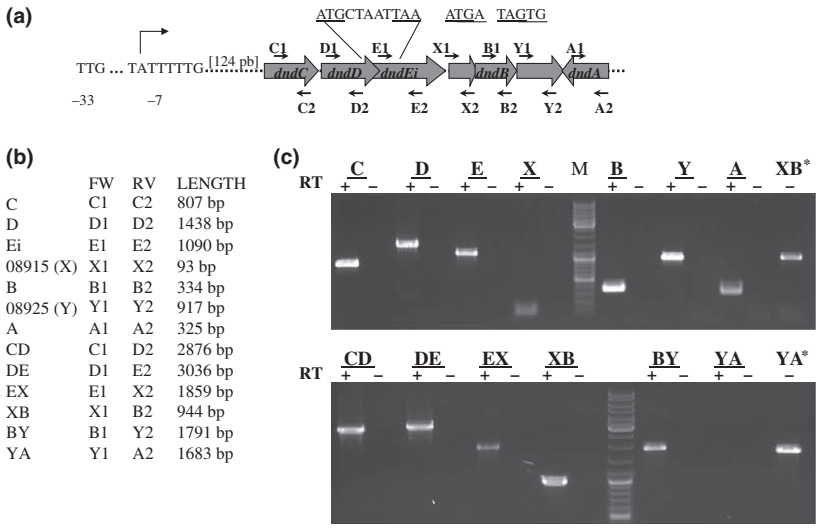
Transcriptional analysis of *F. indicum dnd* genes

Sequence analysis of the 9450-bp region of *F. indicum* CIP 109464^T containing the five *dndA-E* homologous genes and two additional genes of unknown function (KQS_08915 and KQS_008925) revealed that *dndA* and the other genes are divergently transcribed, the latter in a likely operonic structure (Figs 2a and S1).

To evidence divergent transcription of *dndA* and the hypothetical *dndCDEi,* KQS_08915, *B,* KQS_008925 operonic structure in *F. indicum* CIP 109464^T, we performed a transcriptional analysis by RT-PCR using different sets of primers (Table S1). Our RT-PCR results (Fig. 2c) suggest that *dndCDEi,* KQS_08915, *B,* KQS_008925 are cotranscribed as a single operon in *F. indicum* CIP 109464^T. The absence of DNA amplicon using primers Y1 and A2 (Fig. 2c – lane YA) confirms an independent transcription of *dndA* compared with the rest of the cluster, as previously reported in the *dnd* gene cluster of *Streptomyces lividans* (Zhou *et al.*, 2005; Xu *et al.*, 2009).

Identification of *dnd* clusters in the family *Flavobacteriaceae*

The presence of *dnd* genes has already been reported in the genome of *Microscilla marina* ATCC 23134 (Xu *et al.*, 2010), another member of the phylum *Bacteroidetes*. However, such *dnd* gene clusters have never been reported within flavobacteria so far. Using a dedicated homologs identification strategy, we identified the presence of *dnd* gene clusters in three other members of



	FW	RV	LENGTH
C	C1	C2	807 bp
D	D1	D2	1438 bp
Ei	E1	E2	1090 bp
08915 (X)	X1	X2	93 bp
B	B1	B2	334 bp
08925 (Y)	Y1	Y2	917 bp
A	A1	A2	325 bp
CD	C1	D2	2876 bp
DE	D1	E2	3036 bp
EX	E1	X2	1859 bp
XB	X1	B2	944 bp
BY	B1	Y2	1791 bp
YA	Y1	A2	1683 bp

Fig. 2. Organization of the *Flavobacterium indicum* CIP 109464^T *dnd* gene cluster and RT-PCR analysis. *dnd* gene transcripts were reverse-transcribed and amplified. The two genes of unknown function KQS_08915 and KQS_08925 are labeled X and Y, respectively. (a) Relative positions and directions of the corresponding primers are marked with blacks arrows. A putative *Flavobacterium* promoter (Chen *et al.*, 2010) was identified 124-bp upstream the start codon of *dndC*. The DNA sequences between adjacent genes are indicated at the top, with start codon and stop codon underlined. (b) Amplification products with sense primer (FW), antisense primer (RV), and their corresponding lengths. Intra-dnd gene amplification products are indicated as *dnd* gene names, while products of regions between *dnd* genes are named linking two corresponding genes such as CD. (c) Electrophoresis analysis of RT-PCR products. The amplification products are labeled as described above. Reactions omitting the reverse transcription step (-RT) and reactions without DNase treatment (*) were included in each run as negative and positive controls. DNA markers are labeled 'M' (10-kb DNA Ladder Mix, Fermentas).

flavobacteria. We also observed *dnd* clusters within the genomes of three members of the family *Prevotellaceae* and nine other members of the phylum *Bacteroidetes* (Table 1 and Fig. S1). Analysis of the GC% of the *dnd* loci and their dinucleotide distribution bias differing from the chromosomal backbone (Table 1) suggests their laterally acquired origin.

Incomplete (i.e. fewer than three Dnd-encoding genes) and probably not functional *dnd* clusters have been also found in *Cecembia lonarensis* LW9 (DndA : WP_009184828.1 and DndC : WP_009184832.1) and *Capnocytophaga canimorsus* Cc5 (DndC : YP_004740273.1 and DndE : YP_004740276.1) genomes. These *dnd* gene clusters were not further included in our analysis.

An unforeseen predicted hybrid DndE protein in members of the phylum *Bacteroidetes*

DndE was originally described as a small putative phosphoribosylaminoimidazole carboxylase of 126 residues homologous to NCAIR synthetases (Zhou *et al.*, 2005). Despite its essential role in the PT DNA modification, the biochemical function of DndE has not been demonstrated so far. Recent structural studies of DndE from *Escherichia coli* and *Salmonella enterica* indicate that it might be a nicked dsDNA-binding protein (Chen *et al.*, 2011; Hu *et al.*, 2012).

The predicted length of the DndE homologs identified to date never exceeds 141 residues (Ou *et al.*, 2009), but strikingly, the DndE protein from *F. indicum* CIP 109464^{T} (hereafter named DndEi) is predicted to be a 520-residue protein. Therefore, DndEi corresponds to an extended version of the previously described DndE (i.e. about four times longer). Indeed, the amino-terminal part of DndEi (1–128) shows 50% similarity with DndE from *Clostridium perfringens* NCTC 8239 (Table S2) corresponding to the domain PF08870 in the Pfam database (Punta *et al.*, 2012), while its carboxy-terminal (187–485) part contains an AAA+ ATPase superfamily domain found in many proteins with ATPase activity involved in a wide range of cellular processes (Iyer *et al.*, 2004). This domain, PF12846 in the Pfam database, contains a P-loop NTPase region, which exhibits the conserved nucleotide phosphate-binding motif ($^{194-}$GxxxxGKT^{-201}, where x is any residue) and a ($^{412-}$DEAH^{-416}) helicase motif (Fig. S2).

We found such *dndEi*, included in a *dnd* gene cluster, on two other genomes of the flavobacteria: *Kordia algicida* OT1 and *Riemerella anatipestifer* DSM 15868 (Table 1 and Fig. S1). DndEi homologous proteins (more than 60% of similarity in amino acid sequence and < 10% of difference in protein length with DndEi) were also detected from the genomes of seven other species of the phylum *Bacteroidetes* (Table 1). All these DndEi homologs exhibit conserved residues predicting AAA+ ATPase domain in their carboxy-terminal part (Fig. S2). Moreover, remote *dndEi*-encoding homologs (above our homology threshold criteria) or *dndEi* homologs not included in a *dnd* gene cluster could also be found in other *Bacteroidetes* (e.g. *Psychroflexus torquis*, *Capnocytophaga canimorsus* Cc5, *Leeuwenhoekiella blandensis* MED217, *Saprospira grandis* DSM 2844, and the unidentified eubacterium SCB49) as well as other bacteria outside the phylum *Bacteroidetes* (e.g. *Desulfovibrio africanus* PCS, *Methylobacterium mesophilicum* SR1.6/6, and *Xanthomonas axonopodis* pv. citrumelo F1) suggesting a widespread distribution.

Discussion

Identification of functional *dnd* genes clusters in two *Flavobacterium* species

We identified functional *dnd* gene clusters in the genome of two bacterial strains belonging to two different *Flavobacterium* species. In contrast with all previously described *dnd* loci, the typical gene order was not conserved in the *F. indicum* genome, and a hybrid DndEi-encoding gene was identified. DndEi may result from the fusion between a typical DndE and an AAA+ ATPase domain and possess a 'nicked dsDNA-binding activity' and a NTPase activity of unknown biological role. *F. psychrophilum* KU060626-59 contains only the *dndCDE* gene homologs, suggesting that the minimal functional *dnd* gene cluster is limited to *dndCDE*. It has been recently shown that IscS, another cysteine sulfurtransferase, could complement for DndA protein in *Escherichia coli* (An *et al.*, 2012) to supply this PT modification of DNA, while disruption of *dndB* does not abolish the Dnd phenotype (Liang *et al.*, 2007; Xu *et al.*, 2009). In genome of *F. psychrophilum* KU060626-59, we identified a gene encoding an IscS homologous protein. One can speculate that this gene could substitute *dndA*.

Distribution of *dnd* clusters in members of the phylum *Bacteroidetes*

All previously reported *dnd* gene clusters show a conserved genetic organization with the *dndBCDE* genes invariably oriented in the same order and direction (Ou *et al.*, 2009). The variety of genomic organization and gene composition of *dnd* loci within members of the phylum *Bacteroidetes* is obvious, and these loci seem therefore particularly prone to gene rearrangements in this phylum. Based on their gene composition and gene order, one might conclude that at least two main types of *dnd* gene clusters

Table 1. Identification of *dnd* genes homologs in members of the phylum *Bacteroidetes*

Number of *dnd* genes detected	Organism source	*dndA*	*dndB*	*dndC*	*dndD*	*dndE*	GC% *dnd* locus [GC% whole genome]	$\delta^* \times 10^{3\dagger}$	Genome fragments with lower δ^* (%)‡
3	*Flavobacterium psychrophilum* KU060626-59			AGR55439.1	AGR55440.1	AGR55441.1	28.58 [32.5]	83.234	93.956
5	*Flavobacterium indicum* CIP 109464T	YP_005357778.1	YP_005357776.1	YP_005357772.1	YP_005357773.1	YP_005357774.1§	29.30 [31.4]	94.927	95.413
4	*Flavobacteriaceae bacterium* 3519-10		YP_003095738.1¶	YP_003095736.1	YP_003095737.1	Non detected originally **	31.76 [42.7]	114.547	99.76
4	*Kordia algicida* OT1		WP_007096424.1¶	WP_007096428.1	WP_007096427.1	WP_007096426.1§	28.97 [34.2]	161.484	99.192
3	*Riemerella anatipestifer* DSM 15868			YP_004045227.1	YP_004045228.1	YP_004045229.1§	30.27 [35]	154.948	99.767
4	*Paraprevotella xylaniphila* YIT 11841		WP_008630087.1	WP_008630105.1	WP_008630103.1	WP_008630100.1	37.48 [48.5]	142.562	100
4	*Prevotella amnii* CRIS 21A-A		WP_008450453.1	WP_008450502.1	WP_008450432.1	WP_008450280.1§	34.61 [36.4]	89.942	93.416
4	*Prevotella bivia* JCIHMP010		WP_004335941.1	WP_004335930.1	WP_004335934.1	WP_004335936.1§	36.23 [39.7]	78.254	97.706
3	*Haliscomenobacter hydrossis* DSM1100			YP_004448668.1	YP_004448667.1	YP_004448662.1§	37.50 [47.1]	96.667	97.814
4	*Belliella baltica* DSM 15883	YP_006406347.1		YP_006406357.1	YP_006406356.1	YP_006406355.1	34.20 [36.8]	38.057	70.358
5	*Cyclobacteriaceae bacterium* AK24	WP_010854647.1	WP_010854648.1	WP_010854651.1	WP_010854650.1	WP_010854649.1§	38.11 [46.8]	62.260	95.982
4	*Microscilla marina* ATCC 23134	WP_002704309.1		WP_002704312.1	WP_002704313.1	WP_002704315.1	37.01 [40.6]	78.987	93.707
4	*Microscilla marina* ATCC 23134	WP_004155710.1		WP_004155718.1	WP_004155719.1	WP_004155721.1	38.58 [40.6]	76.524	98.199
3	*Parabacteroides goldsteinii* CL02T12C30			WP_007659347.1	WP_007659346.1	WP_007659344.1	35.76 [43.3]	85.266	95.249

Table 1. Continued

Number of *dnd* genes detected	Organism source	*dndA*	*dndB*	*dndC*	*dndD*	*dndE*	GC% *dnd* locus [GC% whole genome]	$\delta^* \times 10^{3\dagger}$	Genome fragments with lower δ^* (%)‡
4	*Bacteroides* sp. 2_1_33B		WP_008772058.1¶	WP_008772054.1	WP_008772055.1	WP_008772056.1§	34.18 [44.5]	97.347	97.304
3	*Bacteroides xylanisolvens* XB1A			YP_007793114.1	YP_007793115.1	YP_007793116.1	36.33 [40.7]	90.253	96.561
3	*Bacteroides finegoldii* CL09T03C10			WP_007766197.1	WP_007766203.1	WP_007766204.1	34.61 [42.3]	90.133	95.83
3	*Bacteroides faecis* MAJ27			WP_010537098.1	WP_010537099.1	WP_010537100.1§	34.63 [42.4]	77.886	93.262

GenBank accession numbers of Dnd homologs identified within species of the phylum *Bacteroidetes* by BLASTP and confirmed by global alignment.
†Dinucleotide bias analysis was adapted from the method proposed by Karlin (2001). The value δ^* denotes the dinucleotide relative abundance difference between the dnd locus and the complete genome. The δ^* value was calculated with the δp-Web program (http://deltarho.amc.nl) (van Passel *et al.*, 2005). The high δ^* values indicate a likely heterologous origin.
‡The percentage distribution of δ^* is plotted using the δp-Web tool with random host genomic fragments of equal length as input sequences (van Passel *et al.*, 2005).
§Extended version of the previously described DndE.
¶Detected above the initial cut-off, these proteins contain the DndB-conserved DGQHR domain. Schematic representations of these *dnd* gene clusters are shown in Fig. S1.
**|dentified by TBLASTN between genomic positions 1316314 and 1316682 (ACPR00000000.1) with amino acids similarity > 70% with the amino-terminal part (1–130) of DndEi (YP_005357774.1).

coexist within members of the phylum *Bacteroidetes*. One is found in the genomes of *Bacteroides xylanisolvans* XB1A, *Bacteroides finegoldii* CL09T03C10, *Paraprevotella xylaniphila* YIT 11841, *Flavobacteriaceae* bacterium 3519-10, and *F. psychrophilum* KU060626-59, while the other one occurs in the genomes of *K. algicida* OT-1, *R. anatipestifer* DSM 15868, *Bacteroides sp.* 2_1_33B, *P. amnii* CRIS 21A-A, *P. bivia* JCVIHMP010, *H. hydrossis* DSM 1100, *P. goldsteinii* CLT02T12C30, and *B. faecis* MAJ27.

The presence of *dnd* gene clusters seems to occur at low frequency in bacteria. They have never been reported so far among *Flavobacterium* species or other members of flavobacteria. In this study, *dnd* gene clusters were only identified into one *Flavobacterium* genome (i.e. *F. indicum*) of the 12 publicly available to date. Among the 28 draft genomes of *F. psychrophilum* strains from many worldwide geographic origins and different host fish (Duchaud *et al.*, 2007 and E. Duchaud, unpublished data), only strain KU060626-59 contained a *dnd* gene cluster. As such a cluster has been found in only five among the 204 sequenced genomes in the family, the presence of a complete *dnd* gene cluster in flavobacteria is a rare event. Moreover, the phylogenetic tree of concatenated DndC and DndD protein sequences reported here and the phylogenetic tree constructed on 16S rRNA gene sequences are obviously noncongruent (Fig. S3). Together with the GC% and the dinucleotide distribution bias (Table 1), this confirms that *dnd* gene clusters are the result of horizontal genetic transfer events.

Although the process of evolution and dissemination of *dnd* gene clusters across different bacterial species is unknown so far, plasmids have been proposed to play a major role in the dissemination of these clusters. In particular, large plasmids have been suggested to serve as 'natural depository' for *dnd* loci that could be probably sourced from diverse bacterial donors (He *et al.*, 2007). Indeed, *dnd* gene clusters have been never found on complete phage genomes so far. The only defined *dnd* island shown to be functionally mobile is the *S. lividans* SLG island and is known to function as a typical, self-circularizing, site-specific integrative element (He *et al.*, 2007). Using probabilistic model (HMM profiles) (Eddy, 1996) searches across the NCBI plasmid database (that contains 3867 complete plasmid genomes), we detected remote DndC-, DndD- and DndEi-encoding homologs (YP_002967204.1, YP_002967205.1, and YP_002967206.1, respectively) on the *Methylobacterium extorquens* AM1 megaplasmid (Vuilleumier *et al.*, 2009). The presence of this remote homologous gene cluster on a megaplasmid suggests that large plasmids could indeed serve as 'natural depository' and/or vectors for the spreading across bacterial phyla of *dnd* gene clusters, including the unusual cluster identified in this study.

In addition, the high degree of conservation of gene organization in all previously reported *dnd* gene clusters may suggest that these elements evolved from a common ancient ancestor (He *et al.*, 2007; Ou *et al.*, 2009). Our study reveals a contrasted situation: the variety of genomic organization and gene composition of *dnd* loci within members of the phylum *Bacteroidetes* is obvious and differs from those already described. However, the frequent absence of *dnd* islands in members of the same species and the presence of two distinct *dnd* loci within a genome, for instance in *M. marina* (Table 1 and Fig S1), confirmed that the diverse *dnd* clusters islands had been acquired independently on many occasions (Ou *et al.*, 2009). As two main types of *dnd* gene clusters coexist within members of the phylum *Bacteroidetes,* one might suggest at least two independent ways of acquisition possibly through distinct horizontal genetic transfer events where large plasmids likely play an important role.

Acknowledgements

This work was supported in part by Grant ERA-NET EMIDA PathoFish. P.B. is a Université Evry Val d'Essonne Ph.D. fellowship. We are grateful to Christophe Habib and Tatiana Rochat for their useful advices. We thank the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources. R.A.-H. acknowledges Grants FONDECYT 1110219 and the CONICYT/FONDAP/15110027 from the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT, Chile).

References

An X, Xiong W, Yang Y, Li F, Zhou X, Wang Z, Deng Z & Liang J (2012) A novel target of IscS in *Escherichia coli*: participating in DNA phosphorothioation. *PLoS One* **7**: e51265.

Barbier P, Houel A, Loux V, Poulain J, Bernardet JF, Touchon M & Duchaud E (2012) Complete genome sequence of *Flavobacterium indicum* GPSTA100-9T, isolated from warm spring water. *J Bacteriol* **194**: 3024–3025.

Bauer M, Kube M, Teeling H *et al.* (2006) Whole genome analysis of the marine *Bacteroidetes* '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol* **8**: 2201–2213.

Bernardet JF (2011) Family I. *Flavobacteriaceae* Reichenbach 1992. *Bergey's Manual of Systematic Bacteriology*, Vol. 4, 2nd edn (Whitman W, Ed), pp. 106–111, The Williams & Wilkins Co., Baltimore.

Bernardet JF & Kerouault B (1989) Phenotypic and genomic studies of *Cytophaga psychrophila* isolated from diseased rainbow trout (*Oncorhynchus mykiss*) in France. *Appl Environ Microbiol* **55**: 1796–1800.

Brown MV & Bowman JP (2001) A molecular phylogenetic survey of sea-ice microbial communities (SIMCO). *FEMS Microbiol Ecol* **35**: 267–275.

Bryson K, Loux V, Bossy R *et al.* (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res* **34**: 3533–3545.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Chen S, Kaufman MG, Bagdasarian M, Bates AK & Walker ED (2010) Development of an efficient expression system for *Flavobacterium* strains. *Gene* **458**: 1–10.

Chen F, Lin K, Zhang Z, Chen L, Shi X, Cao C, Wang Z, Liang J, Deng Z & Wu G (2011) Purification, crystallization and preliminary X-ray analysis of the DndE protein from *Salmonella enterica* serovar Cerro 87, which is involved in DNA phosphorothioation. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **67**: 1440–1442.

Duchaud E, Boussaha M, Loux V *et al.* (2007) Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. *Nat Biotechnol* **25**: 763–769.

Eckstein F (2007) Phosphorothioation of DNA in bacteria. *Nat Chem Biol* **3**: 468–475.

Eddy SR (1996) Hidden Markov models. *Curr Opin Struct Biol* **6**: 361–365.

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

Fujiwara-Nagata E, Ikeda J, Sugahara K & Eguchi M (2012) A novel genotyping technique for distinguishing between *Flavobacterium psychrophilum* isolates virulent and avirulent to ayu, *Plecoglossus altivelis altivelis* (Temminck & Schlegel). *J Fish Dis* **35**: 471–480.

Gómez-Consarnau L, González JM, Coll-Lladó M, Gourdon P, Pascher T, Neutze R, Pedrós-Alió C & Pinhassi J (2007) Light stimulates growth of proteorhodopsin-containing marine *Flavobacteria*. *Nature* **445**: 210–213.

González JM, Fernández-Gómez B, Fernàndez-Guerra A *et al.* (2008) Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (*Flavobacteria*). *P Natl Acad Sci USA* **105**: 8724–8729.

Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

He X, Ou HY, Yu Q, Zhou X, Wu J, Liang J, Zhang W, Rajakumar K & Deng Z (2007) Analysis of a genomic island housing genes for DNA S-modification system in *Streptomyces lividans* 66 and its counterparts in other distantly related bacteria. *Mol Microbiol* **65**: 1034–1048.

Horner-Devine CM, Leibold MA, Smith VH & Bohannan BJM (2003) Bacterial diversity patterns along a gradient of primary productivity. *Ecol Lett* **6**: 613–622.

Hu W, Wang C, Liang J *et al.* (2012) Structural insights into DndE from *Escherichia coli* B7A involved in DNA phosphorothioation modification. *Cell Res* **22**: 1203–1206.

Iyer LM, Leipe DD, Koonin EV & Aravind L (2004) Evolutionary history and higher order classification of AAA+ ATPases. *J Struct Biol* **146**: 11–31.
Karlin S (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**: 335–343.
Kirchman DL (2002) The ecology of *Cytophaga–Flavobacteria* in aquatic environments. *FEMS Microbiol Ecol* **39**: 91–100.
Liang J, Wang Z, He X, Li J, Zhou X & Deng Z (2007) DNA modification by sulfur: analysis of the sequence recognition specificity surrounding the modification sites. *Nucleic Acids Res* **35**: 2944–2954.
Liu SL & Sanderson KE (1995) *I-Ceu*I reveals conservation of the genome of independent strains of *Salmonella typhimurium*. *J Bacteriol* **177**: 3355–3357.
Oh HM, Kang I, Ferriera S, Giovannoni SJ & Cho JC (2010) Complete genome sequence of *Croceibacter atlanticus* HTCC2559T. *J Bacteriol* **192**: 4796–4797.
Ou HY, He X, Shao Y, Tai C, Rajakumar K & Deng Z (2009) dndDB: a database focused on phosphorothioation of the DNA backbone. *PLoS One* **4**: e5132.
Punta M, Coggill PC, Eberhardt RY *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* **40**: D290–D301.
Ray T, Weaden J & Dyson P (1992) Tris-dependent site-specific cleavage of *Streptomyces lividans* DNA. *FEMS Microbiol Lett* **75**: 247–252.
Ray T, Mills A & Dyson P (1995) Tris-dependent oxidative DNA strand scission during electrophoresis. *Electrophoresis* **16**: 888–894.
Saha P & Chakrabarti T (2006) *Flavobacterium indicum* sp. nov., isolated from warm spring water in Assam, India. *Int J Syst Evol Microbiol* **56**: 2617–2621.
Soto E, Mauel M & Lawrence M (2008) Improved pulsed-field gel electrophoresis procedure for the analysis of *Flavobacterium columnare* isolates previously affected by DNA degradation. *Vet Microbiol* **128**: 207–212.
Touchon M, Barbier P, Bernardet JF, Loux V, Vacherie B, Barbe V, Rocha EP & Duchaud E (2011) Complete genome sequence of the fish pathogen *Flavobacterium branchiophilum*. *Appl Environ Microbiol* **77**: 7656–7662.
van Passel MW, Luyf AC, van Kampen AH, Bart A & van der Ende A (2005) Deltarho-web, an online tool to assess composition similarity of individual nucleic acid sequences. *Bioinformatics* **21**: 3053–3055.
Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.
Vernikos GS & Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**: 2196–2203.
Vuilleumier S, Chistoserdova L, Lee MC *et al.* (2009) *Methylobacterium* genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. *PLoS One* **4**: e5584.
Wang L, Chen S, Xu T, Taghizadeh K, Wishnok JS, Zhou X, You D, Deng Z & Dedon PC (2007) Phosphorothioation of DNA in bacteria by dnd genes. *Nat Chem Biol* **3**: 709–710.
Wang L, Chen S, Vergin KL *et al.* (2011) DNA phosphorothioation is widespread and quantized in bacterial genomes. *P Natl Acad Sci USA* **108**: 2963–2968.
Xie X, Liang J, Pu T *et al.* (2012) Phosphorothioate DNA as an antioxidant in bacteria. *Nucleic Acids Res* **40**: 9115–9124.
Xu T, Liang J, Chen S *et al.* (2009) DNA phosphorothioation in *Streptomyces lividans*: mutational analysis of the dnd locus. *BMC Microbiol* **9**: 41.
Xu T, Yao F, Zhou X, Deng Z & You D (2010) A novel host-specific restriction system associated with DNA backbone S-modification in *Salmonella*. *Nucleic Acids Res* **38**: 7133–7141.
Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.
Zhou X, Deng Z, Firmin JL, Hopwood DA & Kieser T (1988) Site-specific degradation of *Streptomyces lividans* DNA during electrophoresis in buffers contaminated with ferrous iron. *Nucleic Acids Res* **16**: 4341–4352.
Zhou X, He X, Liang J, Li A, Xu T, Kieser T, Helmann JD & Deng Z (2005) A novel DNA modification by sulphur. *Mol Microbiol* **57**: 1428–1438.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Organization of *dnd* genes clusters in 17 members of the phylum *Bacteroidetes.*
Fig. S2. Alignment of eleven DndEi homologues detected in members of the phylum *Bacteroidetes.*
Fig. S3. (A) Inferred phylogenetic relationship of 15 members of the phylum *Bacteroidetes* carrying a *dnd* cluster (or their close relatives) on the basis of 16S rRNA gene sequences. Included in the phylogeny are 16S sequences from : *Flavobacterium psychrophilum* NCIMB 1947T (HM443879.1), *Flavobacterium indicum* CIP 109464T (NR_074422.1), *Kordia algicida* OT-1 (NR_027568.1), *Flavobacteriaceae* bacterium 3519-10 (EU694411.1), *Riemerella anatipestifer* DSM 15868 (NR_074429.1), *Belliella baltica* DSM 15883 (NR_025599.1), *Microscilla marina* NBRC 16560 (AB681071.1), *Haliscomenobacter hydrossis* DSM 1100 (NR_074420.1), *Parabacteroides goldsteinii* JCM13446T (EU136697.1), *Bacteroides faecis* MAJ27 (GQ496624.1), *Bacteroides finegoldii* DSM 17565 (NR_041313.1), *Bacteroides xylanisolvens* XB1A (NR_042499.1), *Paraprevotella xylaniphila* YIT 11841 (NR_041627.1), *Prevotella amnii* CCUG 53648T (NR_042587.1), *Prevotella bivia* ATCC 29303 (NR_044629.1). 16S sequences from *Bacteroides* sp.

2_1_33B and *Cyclobacteriaceae* bacterium AK24 were not included in this analysis. (B) Phylogenetic analysis performed on concatened DndC and DndD proteins sequences, found in all the *Bacteroidetes dnd* clusters reported. * indicate *dnd* gene clusters containing a DndEi homolog. Accessions numbers are shown in Table 1. Bootstrap values (100 replicates) are shown at the nodes and scales represent inferred evolutionary distance.

Table S1. primer sequences used in the RT-PCR analysis of the *Flavobacterium indicum dnd* locus.

Table S2. Dnd proteins homologs of *Flavobacterium* strains identified.

Supplementary Figure S1. Organization of *dnd* genes clusters in 17 members of the phylum *Bacteroidetes*. Colored arrows indicate similar ORFs, their colors referring to the matching *dnd* gene in *Streptomyces lividans 1326* (shown at the bottom as a reference). *dndA* homologs are shown in red, *dndB* homologs in green, *dndC* in yellow, *dndD* in blue and *dndE* in pink. Newly identified *dndE* homologs (*dndEi*) are shown in purple. Hatched arrows represent gene coding for proteins with a predicted function. Hatched orange arrows represent genes linked to Mobile Genetic Elements : Transposase, Integrase and Reverse transcriptase. Full grey arrows indicate genes coding for hypothetical proteins. Blacks arrows represent genes coding for proteins similar to type II Toxin/AntiToxin systems. According to the TADB database (Shao et al., 2011), these proteins show significant amino acid similarity (>65%, e-value < $4x10^{-11}$) with the BT_4529/BT_4530 Xre-Fic like type II Toxin/AntiToxin system from *Bacteroides thetaiotaomicron* VPI-5482. Dnd homologs GenBank accession numbers are shown in Table 1, GenBank accession number range of genes represented here for each locus in brackets : *F. psychrophilum* KU060626-59 [AGR55438 .. AGR55443], *F. indicum* CIP 109464^T [YP_005357772 .. 005357778], *Flavobacteriaceae* bacterium 3519-10 [YP_003095736 .. 003095739], *Kordia algicida* OT1 [WP_007096428 .. 007096424], *Riemerella anatipestifer* DSM 15868 [YP_004045227.. 004045229], *Paraprevotella xylaniphila* YIT 11841 [WP_008630105 .. 008630087], *Prevotella amnii* CRIS 21A-A [WP_008450535 .. 008450350], *Prevotella bivia* JCVIHMP010 [WP_004335924 .. 004335941], *Haliscomenobacter hydrossis* DSM1100 [YP_004448669 .. 004448662], *Belliella baltica* DSM 15883 [YP_006406357.. 006406347], *Cyclobacteriaceae* bacterium AK24 [WP_010854651 .. WP_010854647], *Microscilla marina* ATCC 23134 [WP_002704309 .. 002704315] and [WP_004155710 .. 004155721], *Parabacteroides goldsteinii* CL02T12C30 [WP_007659347 .. 007659344], *Bacteroides* sp. 2_1_33B [WP_008772053 .. 008772060], *Bacteroides xylanisolvens* XB1A [YP_007793114 .. 007793116], *Bacteroides finegoldii* CL09T03C10 [WP_007766197 .. 007766204], *Bacteroides faecis* MAJ27 [WP_010537098 .. 010537100] and *Streptomyces lividans* 1326 [ABP49156 .. ABP49152].

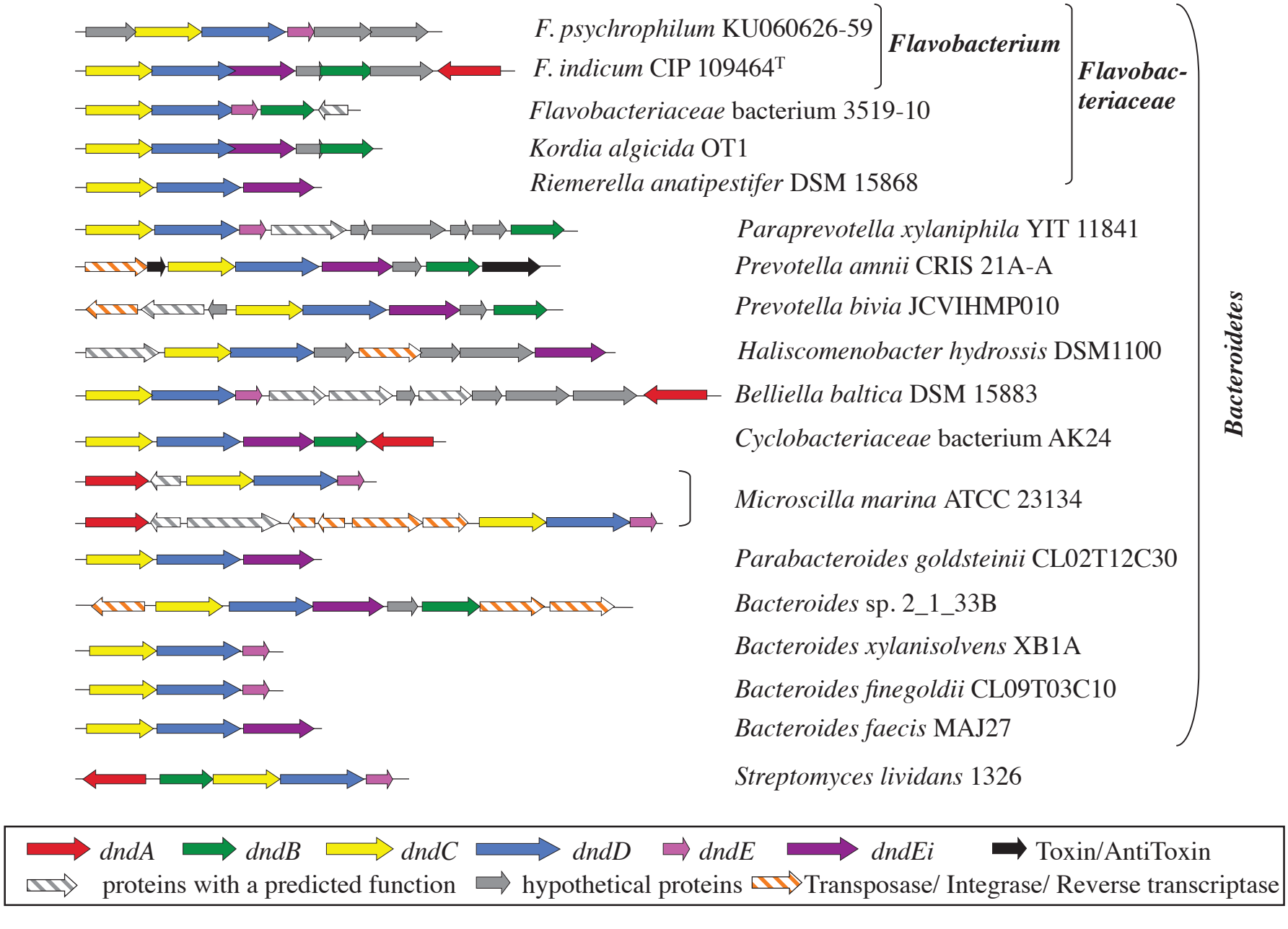
Supplementary Figure S2. Alignment of eleven DndEi homologues detected in members of the phylum *Bacteroidetes*. All proteins, including DndEi, show significant homology with the DNA sulphur modification protein DndE (blue underlined) and contain a P-loop phosphate and Mg2+ binding motif (green underlined) predicting an AAA+ ATPase domain, and a DEAH helicase box (in red) in their carboxy-terminal part. Sequences were aligned using MUSCLE (v3.6) (Edgar, 2004) and the resulting alignement was displayed with Boxshade server (www.ch.embnet.org/software/BOX_form.html). Dark background indicates amino-acid identities and shaded background indicates similarities. Included in the alignment are DndEi homologs from *Flavobacterium indicum* CIP 109464^T (YP_005357774.1), *Kordia algicida* OT1 (WP_007096426.1), *Riemerella anatipestifer* DSM 15868 (YP_004045229.1), *Prevotella amnii* CRIS 21A-A (WP_008450280.1), *Prevotella bivia* JCVIHMP010 (WP_004335936.1), *Haliscomenobacter hydrossis* DSM 1100 (YP_004448662.1), *Cyclobacteriaceae* bacterium AK24 (WP_010854649.1), *Parabacteroides goldsteinii* CL02T12C30 (WP_007659344.1), *Bacteroides* sp. 2_1_33B (WP_008772056.1), *Bacteroides faecis* MAJ27 (WP_010537100.1) and *Capnocytophaga canimorsus* Cc5 (YP_004740276.1).

Supplementary Figure S3.
A. Inferred phylogenetic relationship of 15 members of the phylum *Bacteroidetes* carrying a *dnd* cluster (or their close relatives) on the basis of 16S rRNA gene sequences. Included in the phylogeny are 16S sequences from : *Flavobacterium psychrophilum* NCIMB 1947^T (HM443879.1), *Flavobacterium indicum* CIP 109464^T (NR_074422.1), *Kordia algicida* OT-1 (NR_027568.1), *Flavobacteriaceae* bacterium 3519-10 (EU694411.1), *Riemerella anatipestifer* DSM 15868 (NR_074429.1), *Belliella baltica* DSM 15883 (NR_025599.1), *Microscilla marina* NBRC 16560 (AB681071.1), *Haliscomenobacter hydrossis* DSM 1100 (NR_074420.1), *Parabacteroides goldsteinii* $JCM13446^T$ (EU136697.1), *Bacteroides faecis* MAJ27 (GQ496624.1), *Bacteroides finegoldii* DSM 17565 (NR_041313.1), *Bacteroides xylanisolvens* XB1A (NR_042499.1), *Paraprevotella xylaniphila* YIT 11841 (NR_041627.1), *Prevotella amnii* CCUG 53648^T (NR_042587.1), *Prevotella bivia* ATCC 29303 (NR_044629.1). 16S sequences from *Bacteroides* sp. 2_1_33B and *Cyclobacteriaceae* bacterium AK24 were not included in this analysis.

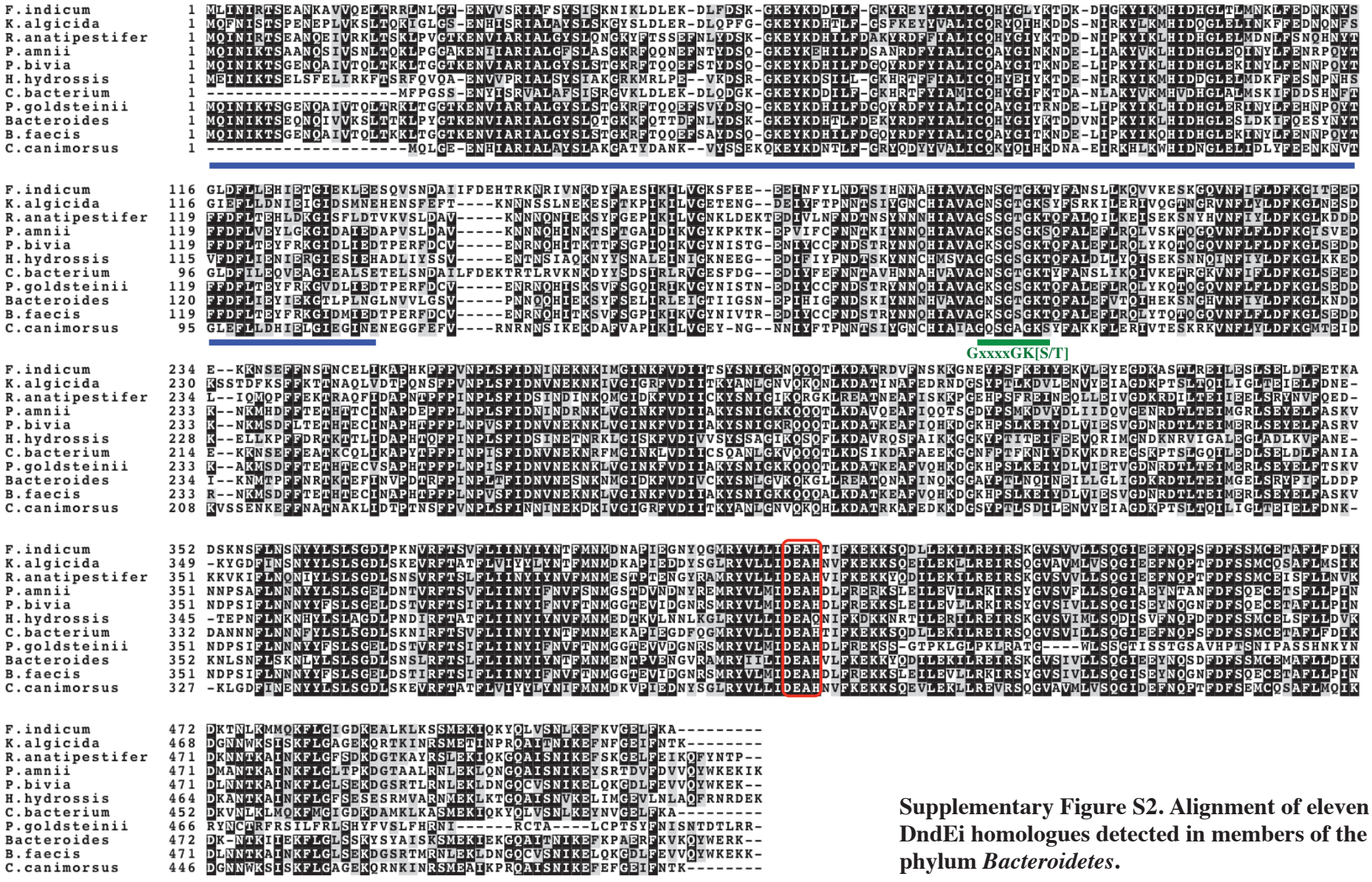
B. Phylogenetic analysis performed on concatened DndC and DndD proteins sequences, found in all the *Bacteroidetes dnd* clusters reported. * indicate *dnd* gene clusters containing a DndEi homolog. Accessions numbers are shown in Table 1. Bootstrap values (100 replicates) are shown at the nodes and scales represent inferred evolutionary distance.

Table S1 : primer sequences used in the RT-PCR analysis of the *Flavobacterium indicum dnd* locus

Table S2 –Dnd proteins homologs of *Flavobacterium* strains identified.
BLASTP results against the dedicaced dnd database (http://db-mml.sjtu.edu.cn/dndDB) (Ou et al., 2009).

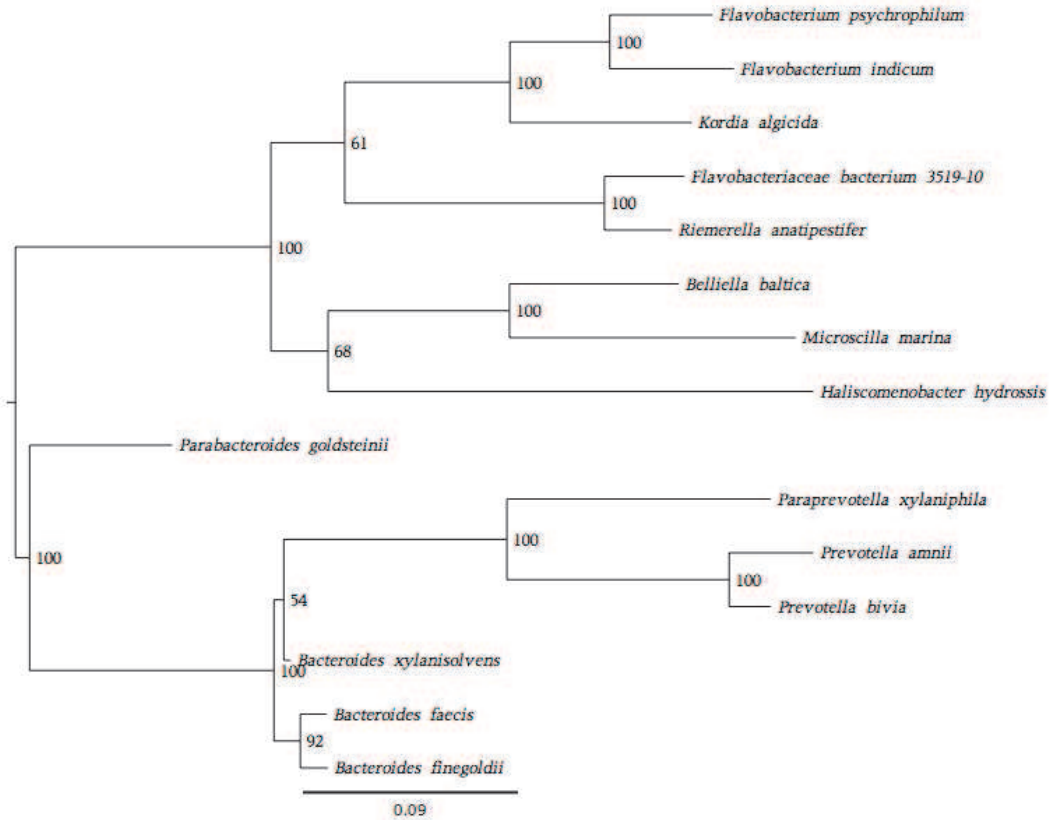


Supplementary Figure S1. Organization of *dnd* genes clusters in 17 members of the phylum *Bacteroidetes*.

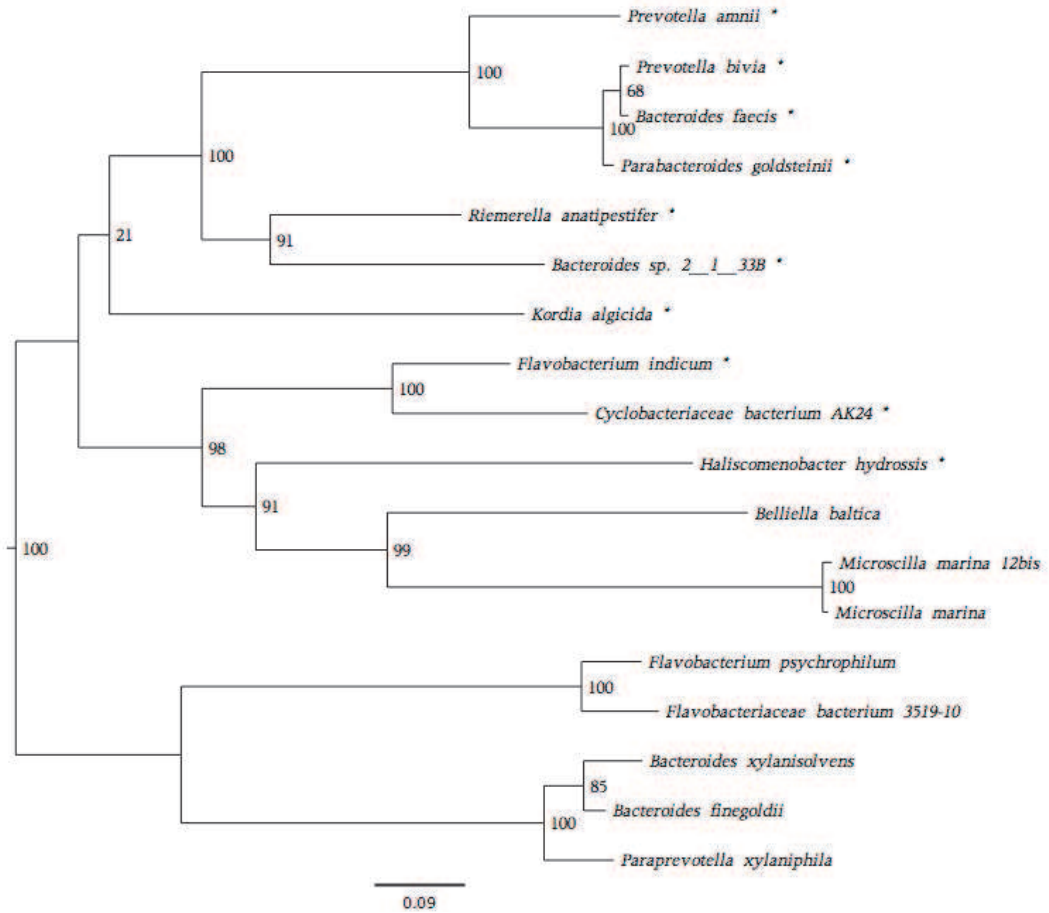


Supplementary Figure S2. Alignment of eleven DndEi homologues detected in members of the phylum *Bacteroidetes*.

A.



B.



Supplementary Figure S3.

Primer Name	Sequence
C1	AGGTTGTTTGGAGATCTCTGT
C2	GCTTTTAAAATTCTGGTCAATATTTC
D1	AAATACATTGGAAATTCTCTTAATAGAT
D2	AACGATATCAATAGTGTCACCAA
E1	TTCTAGAATCGCTTTCTCATACTC
E2	CATAACGCATCCCTTGATAG
X1	CAAGTATAGATGAAGCACTTAAGTATG
X2	TTCGATTTTGACGCTTGA
B1	GATTTACCTAATTCCTATCAGGC
B2	CTTCTTGAGCGTTTTCCC
Y1	AGCTTGTTAAACTCATCTATATTATTCA
Y2	CCATTTTCCTGAGAACTCG
A1	CGTTACTCACAGCTATCAGAGGT
A2	AATTAAAATACCTGCGTTGCT
16S rRNA SP	CTTCCGGTACGGCTACCTTG
16S rRNA AP	CCTGGCTCAGGATGAACGC

Table S1 : primer sequences used in the RT-PCR analysis of the *Flavobacterium indicum dnd* locus

Name (Accession No.)	Size (aa)	Best match (Accession No.)	Best match organism source	Best match size (aa Subject matches)	Identities/ total (percentage identity)	Similarities/ total (percentage similarity)	BlastP E-value
***Flavobacterium indicum* CIP 109464^T**							
DndC (YP_005357772.1)	440	DndC (YP_082437)	*Bacillus cereus* E33L	522 (31 - 317)	134/291 (46%)	197/291 (67%)	2.5e-90
				522 (350 - 520)	44/175 (25%)	85/175 (48%)	2.5e-90
DndD (YP_005357773.1)	698	DndD (ZP_02338392)	*Salmonella enterica* serovar Saintpaul SARA23	671 (5 - 343)	93/366 (25%)	158/366 (43%)	7.8e-34
				671 (284 - 660)	88/403 (21%)	177/403 (43%)	7.8e-34
DndEi (YP_005357774.1)	520	DndE (ZP_02642719)	*Clostridium perfringens* NCTC 8239	129 (1 - 128)	40/138 (28%)	69/138 (50%)	1.3e-09
DndB (YP_005357776.1)	409	DndB (ZP_00714765)	*Escherichia coli* B7A	395 (18 - 172)	53/175 (30%)	75/175 (42%)	6.1e-06
				395 (93 - 344)	75/280 (26%)	113/280 (40%)	7.8e-04
DndA (YP_005357778.1)	391	DndA (YP_001403519)	Candidatus *Methanoregula boonei* 6A8	390 (8 - 378)	173/377 (45%)	237/377 (62%)	3.7e-85
***Flavobacterium psychrophilum* KU060626-59**							
DndC (AGR55439.1)	454	DndC (YP_082437)	*Bacillus cereus* E33L	522 (6 - 313)	144/318 (45%)	201/318 (63%)	1.8e-85
				522 (390 - 513)	34/124 (27%)	67/124 (54%)	1.8e-85
DndD (AGR55440.1)	699	DndD (ZP_01265058)	Candidatus *Pelagibacter ubique* HTCC1002	654 (116 - 648)	159/576 (27%)	259/576 (44%)	9.7e-36
				654 (1 - 78)	28/85 (32%)	41/85 (48%)	9.7e-36
DndE (AGR55441.1)	130	DndE (ZP_02642719)	*Clostridium perfringens* NCTC 8239	129 (4 - 117)	33/115 (28%)	62/115 (54%)	1.7e-11

Table S2 : Dnd proteins homologs of *Flavobacterium* strains identified.

Discussion

Nous avons identifié des groupes de gènes *dnd* fonctionnels pour la modification par phosphorothiation du squelette des molécules d'ADN dans deux souches appartenant à deux espèces différentes du genre *Flavobacterium*. Suite à cette première observation dans la famille des *Flavobacteriaceae*, nous avons recherché la présence des gènes *dnd* au sein des membres du phylum *Bacteroidetes* et réalisé des comparaisons génomiques.

Contrairement aux locus *dnd* déjà décrits dans la littérature, l'ordre des gènes n'est pas conservé dans le génome de *F. indicum* et une protéine DndE hybride a été identifiée. Cette protéine résulte probablement de la fusion d'une DndE classique et d'un domaine AAA+ ATPase. Le génome de *F. psychrophilum* KU060626-59 ne contient que les homologues des gènes *dndCDE*, suggérant ainsi que le locus *dnd* fonctionnel minimal semble limité aux gènes *dndCDE*.

D'autre part, tous les locus *dnd* décrits jusqu'ici montraient une organisation génétique conservée avec les gènes *dndBCDE* invariablement orientés dans le même ordre et la même direction [107]. La variation en terme d'organisation et de composition en gènes des locus *dnd* identifiés au sein des membres du phylum *Bacteroidetes* est évidente et ces locus semblent particulièrement sujet aux réarrangements dans ce phylum (Figure S1). En se basant sur la composition et l'ordre des gènes on peut conclure que deux grands types de locus *dnd* coexistent au sein des génomes des membres du phylum *Bacteroidetes*.

Les gènes *dnd* sont observés à basse fréquence dans les génomes bactériens. Ils n'ont jamais été décrits jusqu'ici dans le genre *Flavobacterium* ou d'autres membres de la famille des *Flavobacteriaceae*. A travers cette étude, les gènes *dnd* ont été identifiés uniquement dans un des génomes (*F. indicum*) parmi les douze disponibles à ce jour pour le genre *Flavobacterium*. Parmi 28 « drafts » de génomes de souches de *F. psychrophilum* provenant d'origines géographiques et de poissons hôtes différents, seule la souche KU060626-59 contient un locus *dnd*. De même, le locus *dnd* a été identifié dans seulement cinq des 204 génomes des membres de la famille des *Flavobacteriaceae* disponibles à ce jour. La présence

du locus *dnd* dans les génomes des membres de cette famille est donc un événement rare. De plus, l'arbre phylogénétique réalisé à partir du des protéines DndC et DndD concaténées et l'arbre réalisé à partir des séquences du gène 16S des organismes concernés ne sont pas congruents (Figure S3). Avec la différence de GC% et le biais de distribution des dinucléotides observés entre les locus *dnd* et le reste du génome chez tous les organismes concernés (Table 1), cela confirme que la présence des gènes *dnd* est bien le résultat d'événements de transferts horizontaux.

Bien que le procédé d'évolution et de dissémination des locus *dnd* à travers différentes espèces bactériennes ne soit pas connu à ce jour, les plasmides ont été proposés pour jouer un rôle central dans la dissémination de ces locus [113]. L'identification d'homologues lointains des gènes *dndCDE* sur le megaplasmide de *Methylobacterium extorquens* suggère que les grands plasmides pourraient en effet servir de réservoirs naturels ou de vecteurs pour la dissémination des locus *dnd* à travers les phylums bactériens.

Les études précédentes suggèrent que les locus *dnd* identifiés jusqu'ici ont évolués depuis une version ancestrale commune [107], [113]. Notre recherche au sein des membres du phylum *Bacteroidetes* révèle une situation contrastée. La variation d'organisation et de composition en gènes des locus *dnd* identifiés est évidente et ces locus diffèrent de ceux précédemment rapportés. De plus, l'absence fréquente de locus *dnd* dans les membres de la même espèce ou encore la présence de deux locus distincts au sein d'un génome (*M. marina*, Table 1 et Figure S1), confirment que les divers locus *dnd* ont été acquis de manière indépendante à plusieurs occasions [107]. Comme deux grands types de locus *dnd* coexistent au sein des génomes des membres du phylum *Bacteroidetes*, cela suggère au moins deux voies d'acquisition différentes, à travers des évènements de transferts horizontaux distincts où les grands plasmides joueraient probablement un rôle important.

Discussion Générale

Qualité des assemblages: de la haute-couture au prêt-à-porter

Les premiers génomes bactériens complets, obtenus par des approches de séquençage de banques de clones au prix de lourds efforts de finition, étaient des séquences d'excellente qualité [41], [42]. Suite à l'arrivée des NGS, l'évaluation de la qualité d'un génome était dichotomique : les génomes complets d'une part et les génomes morcelés ou « drafts » d'autre part. Le nombre de « drafts » génomiques de qualité variable a par la suite augmenté de manière considérable. Les possibilités exponentielles du séquençage brut et l'importante réduction de son prix ont profondément modifié les rapports temps et coûts associés à l'obtention d'un « draft » par rapport à l'obtention d'un génome « bien fini ». Cela a permis de séquencer de plus en plus de génomes, augmentant ainsi le volume de données génomiques. Ce constat a poussé des consortiums de séquençage internationaux à redéfinir les normes des assemblages afin de mieux refléter la qualité des génomes [114]. Les six niveaux de qualité différents vont maintenant du « *Standard Draft* », séquence de qualité minimum pour un dépôt dans les bases de données, au « *Finished* », séquence complète (la plupart du temps circulaire) avec moins de une erreur pour 100000 pb. Une séquence « *Improved High-Quality Draft* » est définie comme la norme acceptable pour des comparaisons génomiques. Elle ne contient pas d'incertitude dans l'assemblage et peu de « scaffolds ». A ce jour (septembre 2013), la base de données GOLD (Genome Online Database) référençant 22465 génomes « complets » de bactéries, contient seulement 2331 génomes de qualité « *Finished* ». Cela illustre bien la difficulté à obtenir des séquences de bonne qualité.

La qualité de l'assemblage d'un génome est donc communément évaluée par un minimum de contigs de la plus grande taille possible et un minimum d'incertitude sur l'ordre de ces derniers (« scaffolding »). Cependant, en pratique, la qualité et la difficulté d'assemblage d'un génome dépendent à l'évidence de la stratégie utilisée (cf. ci-après) mais essentiellement des caractéristiques intrinsèques de l'organisme séquencé, en particulier du nombre de séquences répétées contenues dans son génome mais aussi de leur taille, de leur organisation et de leur localisation. Enfin, le temps, l'énergie et l'argent consacrés au

séquençage du génome influence également sa qualité. Les grandes régions répétées dans les génomes de *F. psychrophilum* (cf. première partie) en sont un bon exemple. La structure et la localisation de ces répétitions ont été de grands obstacles et ont demandé un certain temps de travail en finition et en validation.

La qualité d'un génome est enfin dépendante de l'utilisation prévue des données. L'obtention d'un génome *de novo* (dans un genre ou une espèce où il n'existe pas de référence) nécessite un génome d'une excellente qualité. En effet, un génome doit être le plus exact possible s'il doit servir de point de départ à des études de génomique fonctionnelles. Cela a été le cas par exemple du génome de *F. psychrophilum* JIP 02/86 qui a servi de véritable support à la mise en place des approches de génomique au sein de notre équipe. A l'inverse, lorsque l'on travaille dans un contexte donné, l'investigation de la diversité génomique entre des souches proches ne demande pas forcément d'obtenir une qualité optimale d'assemblage des génomes. Une fois la référence acquise, l'étude du polymorphisme au sein d'une espèce ou la recherche de mutations dans une même souche peut se réaliser par des méthodes de « mapping » des reads sur une séquence de référence. Les séquences d'ADN obtenues dans cette optique ne demanderont pas nécessairement d'en effectuer l'assemblage. La qualité requise d'un génome dépend donc de l'objet d'étude ou de son utilisation ultérieure [114], [115].

On pourrait considérer qu'un génome complet est un peu comme un vêtement : sa qualité dépend de l'occasion pour laquelle il est porté. On peut donc opposer une séquence génomique de référence « bien finie », considérée aujourd'hui un peu comme de la haute couture tant son obtention est minutieuse, longue et difficile ; à un classique « draft » de génome dont l'obtention est plus aisée et ressemblant alors à une pièce d'une collection de prêt-à-porter.

Stratégies de séquençage ou l'art du compromis

La génomique bactérienne a rapidement évolué ces dernières années. Les nouvelles méthodes de séquençage sont sans cesse plus performantes en terme de vitesse et de coût, généralisant ainsi leur utilisation. Le contexte technique de nos travaux de recherches est donc très dynamique. De nombreux outils d'obtention et de traitement des données brutes sont maintenant disponibles. Il est donc parfois difficile de s'y retrouver dans les différentes stratégies de séquençage (454, Solexa, IonTorrent, reads « paired-end » ou non, « mate-pair », etc…) De plus, chaque technologie de séquençage ayant son assembleur dédié, définir les bonnes stratégies à utiliser aujourd'hui pour le séquençage et l'assemblage d'un nouveau génome constitue également une difficulté. En France, de manière plus « contextuelle » et politique, les équipes ou instituts de recherche se lançant dans des approches génomiques ont eu tendance à mettre chacun en place leur propre plateforme avec leurs propres stratégies et outils « maison » d'assemblage des séquences produites. Cette situation contribue grandement à la difficulté du choix de la stratégie à utiliser pour un projet de séquençage et devrait faire naître un besoin d'uniformisation des stratégies utilisées. L'enjeu technique principal d'un projet de séquençage reste cependant de trouver un bon compromis entre le prix de revient, la qualité et la vitesse d'obtention de la séquence finale.

Les stratégies utilisées dans les différents projets de séquençage de génomes complets présentés au cours de cette thèse illustrent parfaitement à la fois le dynamisme du contexte technique des approches de séquençage mais aussi les transitions qui ont eu lieu ces dernières années dans les stratégies de séquençage.

Le génome complet de *F. branchiophilum* FL15 [48] a été obtenu uniquement par la méthode de Sanger. Les extrémités des plasmides de trois banques d'ADN génomique, aux tailles d'inserts différentes (3, 10 et 40 kpb), ont été séquencées pour une couverture totale du génome de 14X. Les avantages principaux d'utilisation de cette stratégie ont été : un faible taux d'erreur grâce à la couverture relativement haute du séquençage Sanger ; la rapidité de la phase de finition due à l'utilisation des trois banques aux tailles d'inserts différentes qui ont facilité le « scaffolding » et à l'utilisation d'une carte optique du génome. L'inconvénient a été le prix du séquençage des 56256 reads Sanger (environ 1 euro par read soit 50000 euros à

l'époque). Le compromis choisi pour la stratégie de séquençage de ce génome penche donc plus vers la qualité et la rapidité (relative) de la phase de finition.

Le génome complet de *F. indicum* GPTSA100-9^T [49] a été obtenu grâce à une stratégie de séquençage qualifiée de « mixte ». Un premier jeu de séquence a été obtenu par la méthode 454 (version FLEX) avec une couverture de 17X. Les inconvénients liés à l'utilisation de cette technique ont été des séquences relativement courtes (225 pb en moyenne) créant de nombreux contigs et n'offrant pas la possibilité d'orienter les contigs entres eux (« scaffolding »), et un taux d'erreurs très élevé en particulier dans les régions homopolymériques. Un second jeu de séquences a été obtenu par séquençage Sanger des extrémités des plasmides d'une banque d'ADN génomique (inserts de taille moyenne de 10 kpb) avec une couverture de 3X. L'avantage principal de cette technique a été la possibilité de réaliser des prédictions de liens entre les contigs et de débuter la phase de finition. L'utilisation d'une carte optique a permis de terminer cette phase de finition en ordonnant les super-contigs entres eux et de valider d'une manière indépendante l'assemblage de ce génome.

Le compromis choisi pour la stratégie d'obtention de ce génome penche donc plus vers la qualité de la séquence finale. Au vu de l'investissement financier (environ 20000 euros) et en temps de travail (début en 2009 et fin en 2011) pour l'obtention de ce génome de très bonne qualité, on pourrait dire que ce génome est « cousu main ». Il illustre toutefois les savoir-faire en génomique et la fiabilité scientifique du laboratoire.

Les génomes complets de plusieurs souches de *F. psychrophilum* (non publiées) et ceux des souches type du genre *Tenacibaculum* (non publiées) ont été obtenus par séquençage Solexa « paired-end » avec des couvertures supérieures à 50X. L'augmentation de la couverture et l'utilisation de la technologie « paired-end » ont permis de réduire le nombre de régions mal assemblées diminuant ainsi les trous dans l'assemblage des séquences complètes. Les assemblages de ces séquences ont donc été beaucoup plus rapides et ont permis d'augmenter la qualité générale des assemblages obtenus, le tout à un coût raisonnable (environ 1000 euros par génome).

Trois paramètres stratégiques influent sur l'assemblage et sur la qualité globale d'un génome : la taille des reads (en paires de bases) et le nombres de reads affectent essentiellement la couverture, tandis que la possibilité d'obtenir des lectures « paired-end » permet d'orienter les contigs entre–eux et d'obtenir des super-contigs (« scaffolds »). L'enjeu d'un projet de séquençage réside donc dans le fait de trouver le meilleur compromis entre ces trois paramètres afin d'obtenir une séquence finale de bonne qualité.

Cependant, augmenter la couverture lors des séquençages de génomes peut aboutir à l'introduction d'erreurs dans l'assemblage. Nous avons constaté que le séquençage Solexa « paired-end » avec une couverture supérieure à 100X peu d'informations supplémentaires sont ajoutées et qu'au delà de 200X des erreurs d'assemblages apparaissent (contigs chimériques). Du fait des nombreuses erreurs introduites dans les régions homopolymériques, la qualité d'un séquençage 454 « paired-end » est inférieure à la qualité d'un séquençage Solexa « paired-end » et ce même si les reads 454 sont plus longs (750 pb pour les reads 454 contre 100 pb pour les reads Solexa). Ce constat place aujourd'hui l'utilisation de la méthode de séquençage Solexa « paired-end » comme un excellent rapport qualité/prix pour l'obtention de génomes complets. Cette stratégie de premier choix pour l'obtention d'un génome est maintenant généralisée à tous nos projets de séquençage en cours. Ce compromis a été choisi afin d'optimiser les rapports de qualité, de temps et de coûts associés à l'obtention de génomes complets, en gardant bien à l'esprit qu'un jour une meilleure stratégie pourra la remplacer.

De l'importance de la finition ou la recherche de la perfection

La finition d'un génome afin d'obtenir sa séquence complète est extrêmement importante pour de multiples raisons. En effet, lorsqu'un organisme présente des caractéristiques d'intérêt, une séquence complète (donc bien finie) de son génome peut fournir une base d'étude pour le long terme. C'est notamment le cas des génomes de nombreuses espèces de bactéries modèles et pathogènes qui servent de références et de supports à de multiples études [41], [42], [116]. Ce constat a par exemple mené au re-séquençage du génome de l'organisme modèle *Bacillus subtilis* afin d'améliorer sa qualité [117].

De plus, les résultats du séquençage des génomes complets de nombreux organismes ont montrés que l'information issue de l'organisation générale d'un génome peut améliorer la compréhension de leur biologie. Par exemple, la présence d'un second chromosome chez *Vibrio cholerae* qui a été acquis comme un élément génomique distinct dans l'histoire évolutive de cette espèce [118]. Avec un « draft » du génome seul, ces observations et les expériences complémentaires qu'elles suggèrent n'auraient pas été possibles [115].

L'obtention de génomes complets destinés à être comparés demande de produire des séquences de très haute qualité. En effet, de nombreux travaux ont permis de mettre en évidence des éléments d'organisation des chromosomes bactériens directement liés à des processus cellulaires fondamentaux de la cellule bactérienne. Ces travaux indiquent un rôle prépondérant des structures bi- et tri-dimensionnelles (organisation en opérons et supra-opérons, organisation en macrodomaines, respectivement) des génomes bactériens.

Les gènes co-localisés sur le chromosome et reliés fonctionnellement peuvent être organisés en opérons qui correspondent à des unités transcriptionnelles. Des comparaisons génomiques entre génomes éloignés montrent également que les opérons remplissant des fonctions complémentaires sont souvent regroupés en super-opérons ce qui leur permet de partager la machinerie traductionnelle [119], [120]. Les chromosomes bactériens sont compactés en complexe nucléoprotéiques appelés nucléoïdes dont les structures sont intimement liées au fonctionnement du génome. Par exemple, les macrodomaines Ori et Ter regroupent respectivement l'origine et la terminaison de la réplication d'un génome. La réplication influence très fortement l'organisation chromosomique qui a elle même une incidence sur le dosage de l'expression des gènes (un gène situé à proximité de l'origine de

réplication aura un niveau d'expression plus élevé). Il a été montré que les gènes fortement sollicités lors de la phase exponentielle de croissance (gènes impliqués dans la transcription et la traduction) se regroupent effectivement près de l'origine de réplication [121]. La fourche de réplication synthétise un brin d'ADN de façon continue pour le brin direct et semi-discontinue pour le brin retardé. Ainsi, un biais de distribution des nucléotides mais aussi de densité de codage entre les deux brins est observé. Le brin direct est plus riche en guanine (G) et le brin retardé plus riche en cytosine (C). Calculer le biais de distribution en bases G et C (« GC skew », G-C/G+C) permet d'identifier l'origine et la terminaison de la réplication. La densité de gènes codants est également plus élevée sur le brin direct. La transcription et la traduction étant couplées chez les bactéries, ce biais serai dû à la plus haute probabilité de collision entre les complexes des ADN et ARN polymérases sur le brin retardé que sur le brin direct [119], [120]. Par conséquent les gènes essentiels sont majoritairement localisés sur le brin direct par rapport aux gènes dont le niveau d'expression est important [122]. L'organisation d'un chromosome bactérien est alors largement contrainte autour de l'axe de symétrie de la réplication (axe Ori-Ter) et les inversions de parties du génome qui ont lieu de part et d'autre de cet axe sont fréquemment observées [50], [70], [72], [74]. Les études de comparaison de la structure générale des chromosomes bactériens ou études de synténies ont également permis de mettre en évidence d'autres évènements aboutissant aux réarrangements des génomes tels que les délétions ou les duplications de régions génomiques.

Ces importantes découvertes de mécanismes fondamentaux du fonctionnement et de l'évolution des génomes bactériens ont été possibles grâce à la disponibilité de séquences complètes et bien finies.

Les informations de présence ou d'absence de gènes sont également nécessaires pour déduire certains événements de l'évolution d'un génome tels que la duplication ou la perte de gènes et les transferts horizontaux. L'absence d'un gène dans un « draft » génomique ne peut pas être une preuve de son absence dans le génome. En effet, des erreurs de séquençage et d'assemblage peuvent devenir des sources d'erreurs dans l'analyse ultérieure de la séquence. En particulier, la détection des gènes et leurs comparaisons (assignations d'orthologies) peuvent être perturbées.

Dans cette optique, un investissement dans les phases de finition et de validation des assemblages de génome est nécessaire. L'enjeu de l'étape de finition d'un génome bactérien consiste à améliorer la qualité globale de l'assemblage c'est-à-dire : 1) obtenir des contigs de bonne qualité, 2) orienter ces contigs les uns par rapport aux autres, 3) fusionner les contigs.

Un exemple impliquant un effort de finition d'un génome bactérien consiste à résoudre les différentes copies de l'ADN ribosomique (ADNr). Le nombre de copies de ces locus varie d'une espèce à l'autre, on retrouve par exemple six copies chez *F. psychrophilum*, trois chez *F. branchiophilum* et quatre chez *F. indicum*. Un locus d'ADNr est classiquement organisé de la manière suivante : le gène 16S codant la petite sous-unité du ribosome, le gène 23S puis le gène 5S codants pour la grosse sous-unité du ribosome. Au sein d'une souche tout les locus ne sont pas organisés de la même manière et souvent la nature des gènes codant pour les ARN de transfert (ARNt), situés entre les gènes 16S et 23S, varie d'un locus à l'autre.

Les reads couvrant ces locus d'ADNr s'assemblent tous entre eux lors de la première version d'un assemblage. Considérer alors que l'organisme ne contient qu'une copie d'ADNr est aujourd'hui une erreur répandue dans les assemblages de génomes réalisés par des non-spécialistes. De plus, l'agglomération de toutes les séquences d'ADNr en un seul contig peut aboutir à la perte dans l'assemblage de certains gènes d'ARNt essentiels. Résoudre correctement ces régions prend du temps et n'est malheureusement pas possible avec les méthodes NGS. Il est alors nécessaire d'utiliser des méthodes alternatives comme le clonage ou des PCR longue distance et le séquençage de leurs produits.

L'étape de « scaffolding » permet de résoudre l'organisation générale d'un chromosome bactérien. A l'origine, la stratégie reposait sur le séquençage des extrémités d'inserts de grandes tailles (banques de BACs par exemple). La recherche d'une extrémité d'insert dans la partie terminale d'un contig et la recherche de son autre extrémité permet d'établir des « liens clones » entres contigs et donc d'orienter ainsi les contigs entre eux. Cet échafaudage de contigs est appelé « scaffold ». L'approche de clonage va ainsi avoir un double avantage en procurant également la possibilité d'obtenir la séquence manquante entre deux contigs en dessinant des oligonucléotides qui vont permettre d'acquérir directement la séquence manquante à partir de l'insert du clone chevauchant ce trou. Cette stratégie de

« marche sur clone » est particulièrement utile dans la finition des régions comprenant des répétitions dont la structure et l'organisation sont complexes.

Les cartes optiques permettent d'ordonner et d'orienter des contigs lors d'un assemblage de génome et constituent également une méthode de validation indépendante d'un projet de séquençage. L'utilisation de cet outil a été par exemple d'un grand intérêt pendant la finition du génome de *F. indicum* GPTSA100-9^T, bien qu'elle comporte des limites. La détection optique des fragments de restrictions ne permet pas de détecter des fragments inférieurs à un kpb, aboutissant à une carte physique du génome comportant de petites inexactitudes. De plus, le logiciel utilisé pour la comparaison *in silico* de la carte physique avec l'assemblage du génome ne permet pas toujours de placer correctement des contigs inférieurs à 100 kpb. L'utilisation d'une carte optique permet donc de valider l'organisation générale d'un chromosome bactérien en organisant les « scaffolds » entre eux. Ces outils permettent de gagner un temps précieux et la baisse de leur coût de fabrication laisse espérer leur utilisation systématique pour la finition de génomes complets bactériens [51] et eucaryotes [123].

En pratique, la durée et la difficulté de la phase de finition d'un génome dépendent essentiellement des caractéristiques intrinsèques de l'organisme séquencé, en particulier, du nombre de séquences répétées contenues dans son génome, de leur organisation et de leur localisation.

Par exemple, la finition du génome de *F. indicum* a été particulièrement longue et difficile. En effet, ce génome contient de nombreuses copies d'IS, parfois organisées en tandem ou incluses dans d'autres régions répétées (Annexe 3), ainsi que de très nombreuses copies de gènes eux-mêmes composés de plusieurs blocs de répétitions (éléments rhs et gènes codant pour des adhésines). Ces régions du génome, particulièrement difficiles à résoudre, ont parfois demandé la mise en place d'une stratégie de finition bien particulière. Il a fallu séquencer intégralement les inserts des clones de la banque génomique chevauchants ces régions, réaliser des sous-assemblages de ces régions puis les ré-inclure dans l'assemblage général, un peu comme des attelles. De plus, certains clones de la banque, chevauchant la même région du génome, se sont avérés différents entre eux, une copie d'IS étant présente dans certains inserts et absente d'autres (Annexe 3). Il a fallu beaucoup de temps et d'efforts

pour comprendre l'organisation de cette région particulière mais également de beaucoup d'autres régions au sein de ce génome, rendant la phase de finition particulièrement ardue.

On peut parfois se demander quels sont les bénéfices apportés à un projet de séquençage par la phase de finition souvent coûteuse et chronophage. Outre les éléments mentionnés ci-dessus, le réel avantage d'une phase de finition pour ceux qui la réalisent est la satisfaction d'obtenir un génome le plus exact possible, à l'esthétique circulaire avantageuse. Cette recherche d'une certaine perfection est propice à l'appropriation d'une rigueur scientifique nécessaire.

Les limites de la génomique analytique et la « photographie génomique »

Les NGS ont permis de séquencer de plus en plus de génomes, augmentant ainsi le volume de données et de connaissances disponibles. Cependant, ces avancées posent de nouveaux problèmes dans l'utilisation, la fiabilité et la pertinence des données produites. En effet, le séquençage de génomes est devenu en quelques années une discipline parfois éloignée de l'expérimentation et manquant de liens solides avec la biologie des organismes séquencés.

Les approches de génomique analytique sont particulièrement efficaces pour identifier des gènes probablement impliqués dans un phénotype particulier mais restent néanmoins limitées. En effet, raisonner en terme de présence/absence de gènes n'augure pas pour autant de leur importance relative. Nous l'avons vu précédemment avec l'étude de la mobilité par glissement chez *F. branchiophilum* et *F. indicum* où la présence des gènes pourtant responsables de ce caractère n'est pas directement corrélée avec le caractère mobile des bactéries (cf. troisième partie).

Les approches de génomique fonctionnelle doivent permettre de valider le rôle et l'importance de gènes identifiés *in silico* dans un phénotype particulier ou dans la virulence pour les espèces pathogènes. Ces approches permettent d'assurer des liens solides entre la génomique bactérienne, la microbiologie et les relations hôtes/pathogènes. Les approches de génomique analytique et fonctionnelle sont par conséquent complémentaires, bien qu'elles ne répondent pas aux mêmes questions et qu'elles demandent des savoir-faire expérimentaux différents. La réalisation d'approches de mutagenèse fonctionnelle et d'analyses de l'expression des gènes par des expériences de protéomique et de transcriptomique constitue en excellent moyen de progresser sur la compréhension des fonctions des gènes. Les génomes complets que nous séquençons servent de supports à ces approches expérimentales. L'observation originale des gènes *dnd* dans nos génomes (cf. quatrième partie) a permis la caractérisation d'éléments moléculaires marqueurs de caractères phénotypiques et exprime particulièrement bien la complémentarité des deux types d'approches. Cette complémentarité des approches génomiques et fonctionnelles est aujourd'hui nécessaire dans la manière d'aborder notre compréhension du Vivant.

La découverte de possibles isoformes chromosomiques chez *F. psychrophilum* (cf. première partie) a confirmé que les génomes sont dynamiques. Ces réarrangements chromosomiques semblent fréquents chez *Yersinia pestis* et *Staphylococcus aureus* [50], [67], [72] bien que leurs effets sur la biologie et la pathogénicité de ces organismes ne soient pas complètement élucidés. Les génomes de micro-organismes et leurs structures ne sont pas figés, contrastants ainsi avec la continuité linéaire que l'on peut imaginer à partir d'une séquence d'ADN *in silico*.

De manière plus générale, séquencer le génome d'un organisme revient à décrire base après base la composition de son ADN à un instant « t ». Cependant, la composition génomique d'un organisme est susceptible de changer, du moins en partie, au cours du temps ou en fonction des conditions de culture. Par exemple, des souches de laboratoire repiquées en conditions de culture avantageuses (milieux de culture riches) ou contraignantes (milieux de culture minimum) ont tendance à accumuler des mutations [124], [125]. De même, chez *E. coli* W3110, une dégénérescence rapide du génome a pu être observée probablement en relation avec les conditions de stockage de la souche [126]. Des souches de bactéries pathogènes ont également perdu leurs caractères de virulence au fur et à mesure de leurs repiquages successifs en laboratoire. C'est souvent le cas des souches types des pathogènes humains classiques isolés il y'a parfois plus de cinquante ans [127], [128]. Des études d'évolution expérimentale, réalisées chez *E. coli* dans des temps courts (quelques jours) ou plus longs (plusieurs années), ont également montré la flexibilité des génomes permettant l'adaptation à des pressions de sélection [129], [130], [131].

Il apparaît alors évident que les séquences déposées dans les bases de données sont des objets figés alors que les génomes bactériens sont très dynamiques et continuent à évoluer au gré des sélections et des changements de l'environnement. Séquencer un génome peut être vu comme une photographie de sa séquence d'ADN. La séquence complète d'un organisme représente donc un « instantané » d'un moment unique de son histoire évolutive. Elle restera cependant relativement exacte (et paradoxalement permanente) même si l'organisme continue d'évoluer.

Le genre *Flavobacterium*

Les espèces bactériennes du genre *Flavobacterium* sont retrouvées dans des environnements étonnants et parfois considérés comme des milieux extrêmes tels que les glaciers, les sols gelés, les boues actives des stations d'épuration, des environnements pollués et une source d'eau chaude [20], [21], [22], [23], [24], [132], [133]. L'isolement de plus en plus fréquent de nouvelles espèces à partir du sol ou de la rhizosphère [27], [134] montre que le genre *Flavobacterium* est probablement encore plus répandu sur notre planète qu'initialement estimé.

Une analyse rapide de génomique comparative au sein du genre *Flavobacterium* a récemment montré que deux embranchements distincts semblent coexister au sein de ce genre, le premier serait associé au sol et à la rhizosphère tandis que le second serait associé à des environnements « aquatiques » [27]. Cependant, cette vision de la diversité génomique retrouvée au sein du genre semble un peu simpliste. En effet, l'analyse phylogénétique a été réalisée uniquement sur quelques gènes (gènes de ménage et ADNr 16S) et à partir de seulement sept génomes, biaisant probablement l'analyse. De plus, la définition de l'embranchement « aquatique » semble vague et regroupe à la fois des espèces pathogènes de poissons et des isolats pour lesquels un doute subsiste (la glace est-elle considérée comme de l'eau ? ; le sol et la rhizosphère peuvent également être des environnements très humides, etc…).

Seize génomes complets et « draft » génomiques (bases de données publiques et nos données non publiées) sont disponibles en juin 2013 pour le genre *Flavobacterium*. Ces données génomiques concernent des génomes d'espèces pathogènes ainsi que ceux d'espèces environnementales, soit isolées soit directement séquencées à partir d'échantillons. L'ensemble des caractéristiques de chaque séquence est repris dans l'Annexe 4.

Au sein du genre, les différentes séquences génomiques ont des tailles très différentes et vont de 2,5 Mpb pour *Flavobacterium* sp. SCGCAAA160-P02 jusqu'à 6 Mpb pour *F. johnsoniae* UW101 (Annexe 4). Ces différences dans la taille des génomes reflètent des capacités métaboliques variables. Les organismes ayant de petits génomes (les pathogènes

ainsi que *F. indicum* et *F. glaciei*) sont probablement associés à des niches écologiques restreintes, contrairement aux organismes comportant de plus grands génomes qui sont associés à des niches écologiques plus larges demandant une flexibilité métabolique plus importante. Cette seconde catégorie de taille de génome regroupe effectivement des organismes associés au sol ou à la rhizosphère (*F. johnsoniae* UW101, *Flavobacterium* sp. F52 et *Flavobacterium* sp. CF136) mais aussi des organismes associés à des environnements aquatiques comme un lac ou de l'eau de mer (*Flavobacterium* sp. WG21 et *F. frigidimaris* KUC1, respectivement). Il apparaît cependant, tant au niveau génomique qu'au niveau des habitats où ces représentants sont retrouvés, que l'évolution de ce genre a permis à ces nombreuses espèces de coloniser des niches écologiques très diverses.

La comparaison des génomes au sein du genre *Flavobacterium* disponibles lors de ce travail a permis de mettre en évidence des différences de contenu en gènes mais également d'organisation chromosomique. En effet, on observe peu de conservation dans l'ordre des gènes du génome central entre les génomes de *F. psychrophilum*, *F. johnsoniae* et *F. branchiophilum* (cf. deuxième partie). Cette perte de synténie, confirmée à l'échelle du genre, suggère d'importants réarrangements génomiques. La présence d'isoformes chromosomiques et les phénomènes d'évolution par recombinaison homologue observés chez *F. psychrophilum* mettent en évidence une importante dynamique du génome dans cette espèce (cf. première partie). Des éléments semblent également indiquer un vaste brassage du génome dans la lignée menant à *F. branchiophilum* (cf. deuxième partie). Ces évènements de recombinaison entre les séquences répétées des génomes ont crée une certaine variabilité dans laquelle des séquences d'ADN peuvent êtres réarrangées, dupliquées ou perdues menant certainement à la diversification des espèces. Cette dynamique des génomes observée, conduisant à la perte de synténie est vraisemblablement un moteur puissant dans l'évolution des espèces du genre *Flavobacterium*. En plus de l'accumulation des réarrangements successifs des génomes, les nombreux îlots génomiques retrouvés chez *F. psychrophilum* (cf. première partie), *F. branchiophilum* (cf. deuxième partie) et *F. indicum* (cf. troisième et quatrième parties) suggèrent que le transfert et l'acquisition horizontale de gènes ont également eu un rôle important dans la diversification des espèces au sein du genre.

La constante augmentation du nombre de nouvelles espèces cultivables décrites au sein du genre *Flavobacterium* et la limite des caractérisations phénotypiques laissent supposer qu'il englobe peut-être, dans sa définition actuelle, une entité plus vaste qu'un genre bactérien. Les séquences génomiques de représentants du genre peuvent servir de références solides et aider à la définition des espèces au sein de ce genre.

Les microbiologistes ont toujours eu besoin de classer les organismes qu'ils isolent. Dans cette optique, la taxonomie bactérienne a toujours cherché à délimiter les espèces en se basant sur une cohérence de caractères phénotypiques, phylogénétiques et génomiques. L'hybridation ADN-ADN est pour l'instant la technique qui permet de délimiter les espèces bactériennes, définies comme l'ensemble des souches partageant plus de 70% de similarité, mais cette méthode, longue et difficile à réaliser, devrait rapidement devenir obsolète à l'ère de la génomique. Le pourcentage moyen d'identité nucléotidique (ANI) entre deux génomes semble être actuellement la meilleure alternative à l'hybridation ADN-ADN pour la définition génomique d'une espèce. Cette valeur reflète les fragments de séquences nucléotidiques que l'on peut aligner entre deux génomes en ignorant les régions divergentes. Ainsi, cet outil reflète le degré de distance évolutive entre deux génomes et une valeur de 95% d'ANI représente un seuil équivalent aux 70% de réassociation des deux génomes par hybridation ADN-ADN [135], [136]. Une autre méthode alternative (MUM index [137]), semblait bien fonctionner pour délimiter les espèces mais demandait des génomes complets et ne fonctionnait pas avec les « drafts » génomiques. Récemment, une nouvelle méthode de calcul a été mise au point et permet de calculer les ressemblances entres deux génomes sans nécessairement demander des séquences complètes [138].

Le pourcentage moyen d'identité nucléotidique entre toutes les séquences génomiques disponibles pour le genre *Flavobacterium* a donc été calculé (Annexe 5). Ces résultats montrent bien que les échantillons séquencés appartiennent à des espèces différentes (sauf *F. psychrophilum* souches JIP 02/86 et THC 02/90 ; 99,35%). La valeur la plus forte est obtenue entre les génomes de deux organismes associés à la rhizophère *F. johnsoniae* UW101 et *Flavobacterium* sp. F52 (82,6%), reflétant une certaine proximité évolutive. L'arbre phylogénétique déduit de cette matrice de différences (100% - ANI) n'a cependant pas

permis de délimiter des sous-groupes d'espèces distincts associés à des niches écologiques ou des environnements particuliers (résultats non montré).

Il apparaît dans l'histoire du genre *Flavobacterium* que sa définition a plusieurs fois été modifiée et enrichie. Au fur et à mesure de l'apparition de nouvelles méthodes et outils utilisables pour la taxonomie bactérienne mais aussi de la description de nouvelles espèces, la description du genre a su évoluer. On pourrait donc proposer prochainement, avec la baisse des coûts de séquençage et l'augmentation du nombre d'espèces décrites, que les données de génomiques se substituent aux données d'hybridation ADN-ADN. Cela permettrait de définir les espèces de manière rapide et précise. Cette approche a été récemment appliquée au genre *Acinetobacter* et a montré des résultats cohérents avec les approches de taxonomie classique et de phylogénie réalisée sur les gènes du génome central [139].

Actuellement, les méthodes de métagénomique apparaissent parmi les plus pertinentes pour étudier la diversité des communautés microbiennes. Ces approches permettent également d'extraire des séquences génomiques d'organismes entiers directement à partir de l'environnement sans passer par des étapes d'isolement qui peuvent fausser notre vision de la diversité. Les représentants du genre *Flavobacterium* étant retrouvées dans de très nombreuses niches écologiques, on peut espérer que nos génomes complets (et bien annotés) serviront de références robustes pour la description d'espèces et l'annotation des gènes présents dans ces communautés complexes.

Valorisations des données

Aujourd'hui, les nombreuses méthodes d'identification moléculaire de *F. psychrophilum* existantes ne sont pas optimisées et ne répondent pas aux besoins des professionnels du secteur piscicole. De plus, aucun vaccin commercial n'est disponible pour la protection des salmonidés contre les infections à *F. psychrophilum*.

Grâce au jeu unique de données de génomes complets d'espèces pathogènes et environnementales du genre *Flavobacterium* en notre possession, les gènes spécifiques et ubiquitaires de l'espèce *F. psychrophilum* ont pu être identifiés *in silico*. Ces gènes (ou portions de gènes) sont maintenant utilisés comme cibles moléculaires prometteuses pour le développement d'un test diagnostique (qPCR TaqMan [140]) spécifique de *F. psychrophilum*. Les premiers tests réalisés sont très encourageants et permettent d'espérer leur intégration dans un « kit » facilement utilisable sur des échantillons « de terrain ». Leur utilisation pour le diagnostic et l'épidémio-surveillance dans les pisciculture représentera probablement une réelle plus-value pour les professionnels du secteur.

De plus, une stratégie de « vaccination inverse » (des Génomes Aux Organismes !!!) [141] a permis d'identifier les protéines ubiquitaires dans l'espèce *F. psychrophilum* et probablement exposées à la surface de la bactérie. Cette vingtaine de cibles vaccinales vont êtres clonées, exprimées et testées dans un model *in vivo* afin d'évaluer leur immunogénicité et leur immuno-protection. Ces candidats vaccins sont en cours de production et laissent espérer leur utilisation prochaine dans les premiers essais vaccinaux sur truites arc-en-ciel.

Le développement de ces applications a été réalisé dans le cadre d'un consortium de recherche européen (EMIDA Era-Net, projet PathoFish) incluant une dizaine de partenaires académiques et du secteur industriel.

Les données produites au cours de ce projet doctoral ont permis la réalisation de trois publications scientifiques. Si le format « *Genome Announcement* » choisi pour la publication du génome complet de *Flavobacterium indicum* est relativement court pour un article scientifique, les connaissances déduites de l'analyse de ce génome ont fait l'objet d'une communication orale à la 3ème conférence internationale sur les membres du genre *Flavobacterium* (Flavobacterium 2012) le 5 Juin 2012 à Turku en Finlande.

Conclusion

Les travaux de recherche exposés dans cette thèse s'intègrent dans un projet de compréhension globale de la diversité retrouvée au sein du genre *Flavobacterium*. Les données produites ont enrichi les connaissances fondamentales sur ce genre bactérien, aujourd'hui encore peu étudié. Ces données servent de support au développement d'approches fonctionnelles complémentaires des approches de génomique analytique.

Comprendre les mécanismes moléculaires qui confèrent aux bactéries pathogènes l'aptitude à coloniser, envahir et modifier la physiologie de leurs hôtes constitue un enjeu de première importance, tant d'un point de vue fondamental que d'un point de vue agronomique. L'utilisation abusive de grande quantités d'antibiotiques dans la lutte contre les agents pathogènes dans le secteur piscicole et leur dissémination n'est pas souhaitable aujourd'hui dans le cadre d'une gestion durable de notre environnement [142]. Améliorer le diagnostic et le contrôle des maladies piscicoles constitue donc un enjeu de première importance. Les données produites dans le cadre de ce projet doctoral devraient permettre le développement d'applications plus larges en santé animale, comme des tests de diagnostic et des candidats vaccins.

Nous avons vu qu'il n'existe aucune méthode ou stratégie idéale de séquençage et que le contexte de nos approches et travaux est dynamique. L'enjeu principal consiste donc à tenir compte des ressources matérielles, humaines, temporelles et financières qui peuvent être consacrées à la réalisation des projets de recherche. Tant sur le plan scientifique, matériel qu'économique, les obstacles inhérents à la réalisation des projets sont nombreux et nécessitent de faire des compromis permanents.

Ce projet a reposé sur des approches de microbiologie et de génomique bactérienne, se situant au carrefour de l'écologie microbienne, de l'évolution moléculaire et de la génétique moléculaire. Il a été extrêmement intéressant de réaliser que la connaissance des génomes de micro-organismes permet de mettre en place des applications avec de fortes implications en santé animale pour les espèces pathogènes.

Bibliographie

[1] FAO, *WORLD FISHERIES AND AQUACULTURE*, vol. 35. 2008, p. 176.

[2] FAO, "THE STATE OF WORLD FISHERIES AND AQUACULTURE," *Aquaculture*, vol. Electronic, p. 197, 2010.

[3] "www.franceagrimer.fr/filiere-peche-et-aquaculture/La-filiere-en-bref/La-production-de-la-filiere-peche-et-aquaculture-en-2010," 2011. .

[4] F. Mardones, A. Perez, P. Valdes-Donoso, and T. Carpenter, "Farm-level reproduction number during an epidemic of infectious salmon anemia virus in southern Chile in 2007–2009," *Preventive Veterinary Medicine*, vol. 102, no. 3, pp. 175–184, Dec. 2011.

[5] B. Guichard, "PRINCIPAUX RÉSULTATS DE L'ENQUÊTE « PATHOLOGIE DES POISSONS 2004 »," *Bulletin Epidémiologique Afssa*, no. 15, p. 5, 2004.

[6] E. Duchaud, M. Boussaha, V. Loux, J.-F. Bernardet, C. Michel, B. Kerouault, S. Mondot, P. Nicolas, R. Bossy, C. Caron, P. Bessières, J.-F. Gibrat, S. Claverol, F. Dumetz, M. Le Hénaff, and A. Benmansour, "Complete genome sequence of the fish pathogen Flavobacterium psychrophilum," *Nature Biotechnology*, vol. 25, pp. 763–769, 2007.

[7] P. Nicolas, S. Mondot, G. Achaz, C. Bouchenot, J.-F. Bernardet, and E. Duchaud, "Population structure of the fish-pathogenic bacterium Flavobacterium psychrophilum," *Applied and Environmental Microbiology*, vol. 74, no. 12, pp. 3702–9, Jun. 2008.

[8] J. Bernardet, P. Segers, M. V. Eyt, and F. Berthe, "Cutting a Gordian Knot : Emended Classification and Description of the Genus Flavobacterium , Emended Description of the Family Flavobacteriaceae , and Proposal of Flavobacterium hydatis," *International Journal of Systematic Bacteriology*, no. 85, pp. 128–148, 1996.

[9] J.-F. Bernardet, Y. Nakagawa, and B. Holmes, "Proposed minimal standards for describing new taxa of the family Flavobacteriaceae and emended description of the family," *International Journal of Systematic and Evolutionary Microbiology*, pp. 1049–1070, 2002.

[10] J. P. Euzéby, "LPSN bacterio.net." [Online]. Available: www.bacterio.net/f/flavobacteriaceae.html.

[11] G. C. J. Abell and J. P. Bowman, "Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean.," *FEMS microbiology ecology*, vol. 51, no. 2, pp. 265–77, Jan. 2005.

[12] B. Klippel, A. Lochner, D. C. Bruce, K. W. Davenport, C. Detter, L. a Goodwin, J. Han, S. Han, L. Hauser, M. L. Land, M. Nolan, G. Ovchinnikova, L. Pennacchio, S. Pitluck, R. Tapia, T. Woyke, S. Wiebusch, A. Basner, F. Abe, K. Horikoshi, M. Keller, and G. Antranikian, "Complete genome sequences of Krokinobacter sp. strain 4H-3-7-5 and Lacinutrix sp. strain 5H-3-7-4, polysaccharide-degrading members of the family Flavobacteriaceae," *Journal of Bacteriology*, vol. 193, no. 17, pp. 4545–6, Sep. 2011.

[13] M. J. McBride, G. Xie, E. C. Martens, A. Lapidus, B. Henrissat, R. G. Rhodes, E. Goltsman, W. Wang, J. Xu, D. W. Hunnicutt, A. M. Staroscik, T. R. Hoover, Y.-Q. Cheng, and J. L. Stein, "Novel features of the polysaccharide-digesting gliding bacterium Flavobacterium johnsoniae as revealed by genome sequence analysis," *Applied and Environmental Microbiology*, vol. 75, no. 21, pp. 6864–75, Nov. 2009.

[14] J. E. Johansen, P. Nielsen, and C. Sjraholm, “Description of Cellulophaga baltica gen. nov., sp. nov. and Cellulophaga fucicola gen. nov., sp. nov. and reclassification of [Cytophaga] Iytica to Cellulophaga lytica gen. nov., comb. nov.,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 49, pp. 1231–1240, 1999.

[15] J.-H. Hehemann, G. Correc, F. Thomas, T. Bernard, T. Barbeyron, M. Jam, W. Helbert, G. Michel, and M. Czjzek, “Biochemical and structural characterization of the complex agarolytic enzyme system from the marine bacterium Zobellia galactanivorans,” *Journal of Biological Chemistry*, vol. 287, no. 36, pp. 30571–30584, 2012.

[16] D. Rubbenstroth, M. Ryll, H. Hotzel, H. Christensen, J. K.-M. Knobloch, S. Rautenschlein, and M. Bisgaard, “Description of Riemerella columbipharyngis sp. nov., isolated from the pharynx of healthy domestic pigeons (Columba livia f. domestica), and emended descriptions of the genus Riemerella, Riemerella anatipestifer and Riemerella columbina.,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 63, no. Pt 1, pp. 280–7, Jan. 2013.

[17] J. P. Euzéby, “LPSN bacterio.net.” [Online]. Available: www.bacterio.net/f/flavobacterium.html.

[18] M. J. McBride, “Cytophaga-Flavobacterium Gliding Motility,” *Journal of Molecular Microbiology Biotechnology*, vol. 7, pp. 63–71, 2004.

[19] S. Van Trappen, “Flavobacterium degerlachei sp. nov., Flavobacterium frigoris sp. nov. and Flavobacterium micromati sp. nov., novel psychrophilic bacteria isolated from microbial mats in Antarctic lakes,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, no. 1, pp. 85–92, Jan. 2004.

[20] F. Zhu, “Flavobacterium xinjiangense sp. nov. and Flavobacterium omnivorum sp. nov., novel psychrophiles from the China No. 1 glacier,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 3, pp. 853–857, May 2003.

[21] S. Jit, M. Dadhwal, O. Prakash, and R. Lal, “Flavobacterium lindanitolerans sp. nov., isolated from hexachlorocyclohexane-contaminated soil,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 58, no. Pt 7, pp. 1665–9, Jul. 2008.

[22] H. S. Yoon, Z. Aslam, G. C. Song, S. W. Kim, C. O. Jeon, T. S. Chon, and Y. R. Chung, “Flavobacterium sasangense sp. nov., isolated from a wastewater stream polluted with heavy metals,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 59, no. Pt 5, pp. 1162–6, May 2009.

[23] P. Saha and T. Chakrabarti, “Flavobacterium indicum sp . nov ., isolated from warm spring water in Assam , India,” *International Journal of Systematic and Evolutionary Microbiology*, pp. 2617–2621, 2006.

[24] D.-C. Zhang, H.-X. Wang, H.-C. Liu, X.-Z. Dong, and P.-J. Zhou, “Flavobacterium glaciei sp. nov., a novel psychrophilic bacterium isolated from the China No.1 glacier,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. Pt 12, pp. 2921–5, Dec. 2006.

[25] T. Kazuoka, T. Oikawa, I. Muraoka, S. Kuroda, and K. Soda, “A cold-active and thermostable alcohol dehydrogenase of a psychrotorelant from Antarctic seawater, Flavobacterium frigidimaris KUC-1,” *Extremophiles*, vol. 11, no. 2, pp. 257–67, 2007.

[26] E. L. W. Sack, P. W. J. J. van der Wielen, and D. van der Kooij, “Flavobacterium johnsoniae as a model organism for characterizing biopolymer utilization in oligotrophic freshwater environments,” *Applied and Environmental Microbiology*, vol. 77, no. 19, pp. 6931–8, Oct. 2011.

[27] M. Kolton, S. J. Green, Y. M. Harel, N. Sela, Y. Elad, and E. Cytryn, “Draft genome sequence of Flavobacterium sp. strain F52, isolated from the rhizosphere of bell pepper (Capsicum annuum L. cv. Maccabi),” *Journal of Bacteriology*, vol. 194, no. 19, pp. 5462–3, Oct. 2012.

[28] W. Qi, G. Nong, J. F. Preston, F. Ben-Ami, and D. Ebert, “Comparative metagenomics of Daphnia symbionts,” *BMC Genomics*, vol. 10, p. 172, Jan. 2009.

[29] A. M. Declercq, F. Haesebrouck, W. Van den Broeck, P. Bossier, and A. Decostere, “Columnaris disease in fish: a review with emphasis on bacterium-host interactions,” *Veterinary Research*, vol. 44, no. 1, p. 27, Jan. 2013.

[30] A. Decostere, R. Ducatelle, and F. Haesebrouck, “Flavobacterium columnare (Flexibacter columnaris) associated with severe gill necrosis in koi carp (Cyprinus carpio L),” *Veterinary Record*, vol. 150, no. 22, pp. 694–5, 2002.

[31] H. Wakabayashi, G. J. Hun, and N. Kimura, “Flavobacterium branchiophila sp. nov., a causative agent of bacterial gill disease of freshwater fishes,” *International Journal of Systematic Bacteriology*, vol. 39, pp. 213–216, 1989.

[32] D. J. Speare, R. J. F. Markham, B. Despres, K. Whitman, and N. MacNair, “Examination of Gills from Salmonids with Bacterial Gill Disease using Monoclonal Antibody Probes for Flavobacterium Branchiophilum and Cytophaga Columnaris,” *Journal of Veterinary Diagnostic Investigation*, vol. 7, no. 4, pp. 500–505, Oct. 1995.

[33] H. Wakabayashi, T. Iwado, A. E. Ellis, and (ed.), “Effects of a bacterial gill disease on the respiratory functions of juvenile rainbow trout,” *Fish and shellfish pathology*, pp. 153–160, 1985.

[34] R. C. Cipriano and R. A. Holt, “Flavobacterium psychrophilum, cause of bacterial cold-water disease and rainbow trout fry syndrome,” *Fish disease leaflet*, vol. 86, 2005.

[35] T. Wiklund, L. Madsen, M. Bruun, and I. Dalsgaard, “Detection of Flavobacterium psychrophilum from fish tissue and water samples by PCR amplification,” *Journal of Applied Microbiology*, vol. 88, no. 2, pp. 299–307, 2000.

[36] L. Madsen, J. Møller, and I. Daalsgard, “Flavobacterium psychrophilum in rainbow trout, Oncorhynchus mykiss (Walbaum), hatcheries: studies on broodstock, eggs, fry and environment,” *Journal of Fish Diseases*, vol. 28, no. 1, pp. 39–47, 2005.

[37] Y. Chen, M. Davis, S. Lapatra, K. Cain, K. Snekvik, and D. Call, “Genetic diversity of Flavobacterium psychrophilum recovered from commercially raised rainbow trout, Oncorhynchus mykiss (Walbaum), and spawning coho salmon, O. kisutch (Walbaum),” *Journal of Fish Diseases*, vol. 31, no. 10, pp. 765–73, 2008.

[38] I. Vatsos, K. Thompson, and A. Adams, “Colonization of rainbow trout, Oncorhynchus mykiss (Walbaum), eggs by Flavobacterium psychrophilum, the causative agent of rainbow trout fry syndrome,” *Journal of Fish Diseases*, vol. 29, no. 7, pp. 441–4, 2006.

[39] C. Michel, B. Kerouault, and C. Martin, “Chloramphenicol and florfenicol susceptibility of fish-pathogenic bacteria isolated in France: comparison of minimum inhibitory concentration, using recommended provisory standards for fish bacteria,” *Journal of Applied Microbiology*, vol. 95, no. 5, pp. 1008–15, 2003.

[40] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, and J. Merrick, “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd,” *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[41] F. R. Blattner, "The Complete Genome Sequence of Escherichia coli K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, Sep. 1997.

[42] F. Kunst, N. Ogasawara, I. Moszer, a M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, and A. Danchin, "The complete genome sequence of the gram-positive bacterium Bacillus subtilis," *Nature*, vol. 390, no. 6657, pp. 249–56, Nov. 1997.

[43] Roche, "454 Sequencing Systems Technology Overview." [Online]. Available: http://454.com/resources-support/product-videos.as.

[44] "Illumina Solexa Sequencing Overview." [Online]. Available: http://www.youtube.com/watch?v=77r5p8IBwJk.

[45] S. C. Schuster, "Next-generation sequencing transforms today 's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, 2008.

[46] A. Favello, L. Hillier, and R. K. Wilson, "Genomic DNA sequencing methods," *Methods Cell Biology*, vol. 45, pp. 551–69, 1995.

[47] M. de la Bastide and W. R. McCombie, "Assembling genomic DNA sequences with PHRAP," *Curr Protoc Bioinformatics*., no. Chapter 11, p. Unit11.4, 2007.

[48] M. Touchon, P. Barbier, J.-F. Bernardet, V. Loux, B. Vacherie, V. Barbe, E. P. C. Rocha, and E. Duchaud, "Complete genome sequence of the fish pathogen Flavobacterium branchiophilum," *Applied and Environmental Microbiology*, vol. 77, no. 21, pp. 7656–62, Nov. 2011.

[49] P. Barbier, A. Houel, V. Loux, J. Poulain, J.-F. Bernardet, M. Touchon, and E. Duchaud, "Complete genome sequence of Flavobacterium indicum GPSTA100-9T, isolated from warm spring water," *Journal of Bacteriology*, vol. 194, no. 11, pp. 3024–5, Jun. 2012.

[50] S. K. Shukla, J. Kislow, A. Briska, J. Henkhaus, and C. Dykes, "Optical mapping reveals a large genetic inversion between two methicillin-resistant Staphylococcus aureus strains," *Journal of Bacteriology*, vol. 191, no. 18, pp. 5717–23, Sep. 2009.

[51] P. Latreille, S. Norton, B. S. Goldman, J. Henkhaus, N. Miller, B. Barbazuk, H. B. Bode, C. Darby, Z. Du, S. Forst, S. Gaudriault, B. Goodner, H. Goodrich-Blair, and S. Slater, "Optical mapping as a routine tool for bacterial genome sequence finishing," *BMC genomics*, vol. 8, p. 321, Jan. 2007.

[52] A. Lim, E. T. Dimalanta, K. D. Potamousis, G. Yen, J. Apodoca, C. Tao, J. Lin, R. Qi, J. Skiadas, A. Ramanathan, N. T. Perna, G. P. Iii, V. Burland, B. Mau, J. Hackett, F. R. Blattner, T. S. Anantharaman, B. Mishra, and D. C. Schwartz, "Shotgun Optical Maps of the Whole Escherichia coli O157 : H7 Genome," *Genome Research*, pp. 1584–1593, 2001.

[53] S. Reslewic, S. Zhou, M. Place, Y. Zhang, A. Briska, S. Goldstein, C. Churas, R. Runnheim, D. Forrest, A. Lim, A. Lapidus, C. S. Han, G. P. Roberts, and D. C. Schwartz, "Whole-Genome Shotgun Optical Mapping of Rhodospirillum rubrum," *Applied and Environmental Microbiology*, vol. 71, no. 9, pp. 5511–5522, 2005.

[54] K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. van de Guchte, S. Penaud, E. Maguin, M. Hoebeke, P. Bessières, and J.-F. Gibrat, "AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system," *Nucleic Acids Research*, vol. 34, no. 12, pp. 3533–45, Jan. 2006.

[55] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, “The microbial pan-genome,” *Current opinion in genetics & development*, vol. 15, no. 6, pp. 589–94, Dec. 2005.

[56] J. Reinhardt, D. Baltrus, M. Nishimura, W. Jeck, C. Jones, and J. Dangl, “De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae,” *Genome Research*, vol. 19, no. 2, pp. 294–305, Feb. 2009.

[57] A. Bhattacharyya, S. Stilwagen, N. Ivanova, M. D’Souza, A. Bernal, A. Lykidis, V. Kapatral, I. Anderson, N. Larsen, T. Los, G. Reznik, E. Selkov, T. L. Walunas, H. Feil, W. S. Feil, A. Purcell, J.-L. Lassez, T. L. Hawkins, R. Haselkorn, R. Overbeek, P. F. Predki, and N. C. Kyrpides, “Whole-genome comparative analysis of three phytopathogenic Xylella fastidiosa strains,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 19, pp. 12403–8, Sep. 2002.

[58] C. Buchrieser, C. Rusniok, F. Kunst, P. Cossart, P. Glaser, and L. Consortium., “Comparison of the genome sequences of Listeria monocytogenes and Listeria innocua: clues for evolution and pathogenicity,” *FEMS Immunol Med Microbiol.*, vol. 35, no. 3, pp. 207–13, 2003.

[59] A. Mellmann, D. Harmsen, C. a Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, and H. Karch, “Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology,” *PloS one*, vol. 6, no. 7, p. e22751, Jan. 2011.

[60] F. Dumetz, E. Duchaud, S. E. LaPatra, C. Le Marrec, S. Claverol, M.-C. Urdaci, and M. Le Hénaff, “A protective immune response is generated in rainbow trout by an OmpH-like surface antigen (P18) of Flavobacterium psychrophilum,” *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 4845–52, Jul. 2006.

[61] F. Dumetz, E. Duchaud, S. Claverol, N. Orieux, S. Papillon, D. Lapaillerie, and M. Le Hénaff, “Analysis of the Flavobacterium psychrophilum outer-membrane subproteome and identification of new antigenic targets for vaccine by immunomics,” *Microbiology*, vol. 154, no. Pt 6, pp. 1793–801, Jun. 2008.

[62] B. Alvarez, P. Secades, M. J. Mcbride, and J. A. Guijarro, “Development of Genetic Techniques for the Psychrotrophic Fish Pathogen Flavobacterium psychrophilum,” *Applied and Environmental Microbiology*, vol. 70, no. 1, pp. 581–587, 2004.

[63] C. Garcia, F. Pozet, and C. Michel, “Standardization of experimental infection with Flavobacterium psychrophilum, the agent of rainbow trout Oncorhynchus mykiss fry syndrome,” *Diseases of Aquatic Organisms*, vol. 42, no. 3, pp. 191–7, Sep. 2000.

[64] M. J. McBride and M. J. Kempf, “Development of techniques for the genetic manipulation of the gliding bacterium Cytophaga johnsoniae,” *Journal of Bacteriology*, vol. 178, no. 3, pp. 583–90, Feb. 1996.

[65] J. Bernardet and J. Bowman, *Genus I. Flavobacterium Bergey et al. 1923,*, In Bergey’. The Williams & Wilkins Co. (Baltimore, MD), 2011, pp. 112–154.

[66] M. V Brown and J. P. Bowman, “A molecular phylogenetic survey of sea-ice microbial communities,” *FEMS Microbiol. Ecol.*, vol. 35, pp. 267–275, 2001.

[67] J. Parkhill, B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, a M. Cerdeño-Tárraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, a V Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M.

Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell, “Genome sequence of Yersinia pestis, the causative agent of plague,” *Nature*, vol. 413, no. 6855, pp. 523–7, Oct. 2001.

[68] M. Vos and X. Didelot, “A comparison of homologous recombination rates in bacteria and archaea,” *The ISME journal*, vol. 3, no. 2, pp. 199–208, Feb. 2009.

[69] S. L. Liu and K. E. Sanderson, “I- Ceu I Reveals Conservation of the Genome of Independent Strains of Salmonella typhimurium,” *Journal of Bacteriology*, vol. 177, no. 11, pp. 3355–7, Jun. 1995.

[70] J. A. Eisen, J. F. Heidelberg, O. White, and S. L. Salzberg, “Evidence for symmetric chromosomal inversions around the replication origin in bacteria,” *Genome biology*, vol. 1, no. 6, Jan. 2000.

[71] M. W. van der Woude and A. J. Bäumler, “Phase and Antigenic Variation in Bacteria,” *Clinical Microbiology Reviews*, vol. 17, no. 3, pp. 581–611, 2004.

[72] L. Cui, H. Neoh, A. Iwamoto, and K. Hiramatsu, “Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 25, pp. E1647–56, Jun. 2012.

[73] A. Darling, I. Miklós, and M. Ragan, “Dynamics of Genome Rearrangement in Bacterial Populations,” *PLoS genetics*, vol. 4, no. 7, p. e1000128, 2008.

[74] Y. Furuta, M. Kawai, K. Yahara, N. Takahashi, N. Handa, T. Tsuru, K. Oshima, M. Yoshida, T. Azuma, M. Hattori, I. Uchiyama, and I. Kobayashi, “Birth and death of genes linked to chromosomal inversion,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 4, pp. 1501–6, Jan. 2011.

[75] E. Högfors-Rönnholm and T. Wiklund, “Phase variation in Flavobacterium psychrophilum: characterization of two distinct colony phenotypes,” *Diseases of Aquatic Organisms*, vol. 90, no. 1, pp. 43–53, 2010.

[76] A. Pawlik, G. Garnier, M. Orgeur, P. Tong, A. Lohan, F. Le Chevalier, G. Sapriel, A.-L. Roux, K. Conlon, N. Honoré, M.-A. Dillies, L. Ma, C. Bouchier, J.-Y. Coppée, J.-L. Gaillard, S. V Gordon, B. Loftus, R. Brosch, and J. L. Herrmann, “Identification and characterization of the genetic changes responsible for the characteristic smooth-to-rough morphotype alterations of clinically persistent Mycobacterium abscessus,” *Molecular Microbiology*, pp. 1–18, Sep. 2013.

[77] M. W. van der Woude, “Re-examining the role and random nature of phase variation,” *FEMS Microbiology Letters*, vol. 254, no. 2, pp. 190–7, Jan. 2006.

[78] V. E. Ostland, J. S. Lumsden, D. D. MacPhee, and H. W. Ferguson, “Characteristics of Flavobacterium branchiophilum, the cause of salmonid bacterial gill disease in Ontario,” *Journal of Aquatic Animal Health*, no. 6, pp. 13–26, 1994.

[79] J. W. Warren, “Diseases of hatchery fish. A disease manual,” *U. S. Fish Wildl. Serv.*, no. 3, p. 91, 1981.

[80] J. Farkas, “Filamentous sp. isolated from fish with gill diseases in cold water,” *Aquaculture*, no. 44, pp. 1–10, 1985.

[81] D. Vanden Broeck, C. Horvath, and M. J. De Wolf, “Vibrio cholerae: cholera toxin,” *Int J Biochem Cell Biol*, no. 39, pp. 1771–5, 2007.

[82] L. de Haan and T. L. Hirst, “Cholera toxin and related enterotoxins: a cell biological and immunological perspective,” *J Nat Toxins*, no. 9, pp. 281–97, 2000.

[83] J. M. Bertolini, H. Wakabayashi, V. G. Watral, M. J. Whipple, and J. S. Rohovec, "Electrophoretic detection of proteases from selected strains of Flexibacter psychrophilus and assessment of their variability," *Journal of Aquatic Animal Health*, no. 6, pp. 224–233, 1994.

[84] N. M. O. Brien-simpson, R. A. Paolini, B. Hoffmann, N. Slakeski, S. G. Dashper, and E. C. Reynolds, "Role of RgpA , RgpB , and Kgp Proteinases in Virulence of Porphyromonas gingivalis W50 in a Murine Lesion Model Role of RgpA , RgpB , and Kgp Proteinases in Virulence of Porphyromonas gingivalis W50 in a Murine Lesion Model," 2001.

[85] S. Vieira-Silva and E. P. C. Rocha, "The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics," *PLoS Genet*, vol. 6, no. 1, p. e1000808, Jan. 2010.

[86] D. J. Spear, H. W. Ferguson, F. W. M. Beamish, J. A. Yager, and S. Yamashiro, "Pathology of bacterial gill disease: sequential development of lesions during natural outbreaks of disease," *Journal of Fish Diseases*, no. 14, pp. 21–32, 1991.

[87] A. Reeves, J. D'Elia, J. Frias, and A. Salyers, "A Bacteroides thetaiotaomicron outer membrane protein that is essential for utilization of maltooligosaccharides and starch," *Journal of Bacteriology*, vol. 178, no. 3, pp. 823–30, Feb. 1996.

[88] L. D. McDaniel, E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul, "High frequency of horizontal gene transfer in the oceans," *Science*, vol. 330, no. 6000, p. 50, Oct. 2010.

[89] P. Siguier, "ISfinder: the reference centre for bacterial insertion sequences," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D32–D36, Jan. 2006.

[90] G. S. Vernikos and J. Parkhill, "Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands," *Bioinformatics*, vol. 22, no. 18, pp. 2196–203, Sep. 2006.

[91] P. Barbier, A. Lunazzi, E. Fujiwara-Nagata, R. Avendaño-Herrera, J.-F. Bernardet, M. Touchon, and E. Duchaud, "From the Flavobacterium genus to the phylum Bacteroidetes: genomic analysis of dnd gene clusters," *FEMS Microbiology Letters*, Aug. 2013.

[92] H. C. Tekedar, A. Karsi, A. F. Gillaspy, D. W. Dyer, N. R. Benton, J. Zaitshik, S. Vamenta, M. M. Banes, N. Gülsoy, M. Aboko-Cole, G. C. Waldbieser, and M. L. Lawrence, "Genome sequence of the fish pathogen Flavobacterium columnare ATCC 49512," *Journal of Bacteriology*, vol. 194, no. 10, pp. 2763–4, May 2012.

[93] C. W. Hill, C. H. Sandt, and D. A. Vlazny, "MicroReview Rhs elements of Escherichia coii : a family of genetic composites each encoding a large mosaic protein," *Molecular Microbiology*, vol. 12, no. March, pp. 865–871, 1994.

[94] A. Ponte-Sucre, Ed., *ABC Transporters in Microorganisms*. Universidad Central de Venezuela, Caracas, Venezuela: Caister Academic Press, 2009.

[95] A. P. Pugsley, "The Complete General Secretory Pathway in Gram-Negative Bacteria," *Microbiology Review*, vol. 57, no. 1, 1993.

[96] K. Wooldridge, Ed., *Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis*. Centre for Biomolecular Sciences, University of Nottingham, UK: Caister Academic Press, 2009.

[97] K. Sato, M. Naito, H. Yukitake, H. Hirakawa, M. Shoji, M. J. McBride, R. G. Rhodes, and K. Nakayama, "A protein secretion system linked to bacteroidete gliding motility and pathogenesis,"

Proceedings of the National Academy of Sciences of the United States of America, vol. 107, no. 1, pp. 276–81, Jan. 2010.

[98] E. Karlsson, M. Hachem, S. Ramchuran, H. Costa, O. Holst, S. Svenningsen, and G. Hreggvidsson, "The modular xylanase Xyn10A from Rhodothermus marinus is cell-attached, and its C-terminal domain has several putative homologues among cell-attached proteins within the phylum Bacteroidetes," *FEMS Microbiology Letters*, vol. 241, no. 2, pp. 233–42, 2004.

[99] C. a Seers, N. Slakeski, P. D. Veith, T. Nikolof, Y.-Y. Chen, S. G. Dashper, and E. C. Reynolds, "The RgpB C-terminal domain has a role in attachment of RgpB to the outer membrane and belongs to a novel C-terminal-domain family found in Porphyromonas gingivalis," *Journal of Bacteriology*, vol. 188, no. 17, pp. 6376–86, Sep. 2006.

[100] R. G. Rhodes, S. S. Nelson, S. Pochiraju, and M. J. McBride, "Flavobacterium johnsoniae sprB is part of an operon spanning the additional gliding motility genes sprC, sprD, and sprF," *Journal of Bacteriology*, vol. 193, no. 3, pp. 599–610, Feb. 2011.

[101] S. Agarwal, D. W. Hunnicutt, and M. J. McBride, "Cloning and characterization of the Flavobacterium johnsoniae (Cytophaga johnsonae) gliding motility gene, gldA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 22, pp. 12139–44, Oct. 1997.

[102] M. J. McBride and Y. Zhu, "Gliding motility and Por secretion system genes are widespread among members of the phylum bacteroidetes," *Journal of Bacteriology*, vol. 195, no. 2, pp. 270–8, Jan. 2013.

[103] L. Wang, S. Chen, T. Xu, K. Taghizadeh, J. S. Wishnok, X. Zhou, D. You, Z. Deng, and P. C. Dedon, "Phosphorothioation of DNA in bacteria by dnd genes," *Nature Chemical Biology*, vol. 3, no. 11, pp. 709–10, Nov. 2007.

[104] L. Wang, S. Chen, K. L. Vergin, S. J. Giovannoni, S. W. Chan, M. S. DeMott, K. Taghizadeh, O. X. Cordero, M. Cutler, S. Timberlake, E. J. Alm, M. F. Polz, J. Pinhassi, Z. Deng, and P. C. Dedon, "DNA phosphorothioation is widespread and quantized in bacterial genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 7, pp. 2963–8, Feb. 2011.

[105] X. Zhou, X. He, J. Liang, A. Li, T. Xu, T. Kieser, J. D. Helmann, and Z. Deng, "A novel DNA modification by sulphur," *Molecular Microbiology*, vol. 57, no. 5, pp. 1428–38, Sep. 2005.

[106] X. Zhou, Z. Deng, J. L. Firmin, D. A. Hopwood, and T. Kieser, "Site-specific degradation of Streptomyces lividans DNA during electrophoresis in buffers contaminated with ferrous iron," *Nucleic Acids Research*, vol. 16, no. 10, pp. 4341–4352, 1988.

[107] H.-Y. Ou, X. He, Y. Shao, C. Tai, K. Rajakumar, and Z. Deng, "dndDB: a database focused on phosphorothioation of the DNA backbone," *PloS one*, vol. 4, no. 4, p. e5132, Jan. 2009.

[108] F. Eckstein, "Phosphorothioation of DNA in bacteria," *Nature Chemical Biology*, vol. 3, no. 11, pp. 689–690, 2007.

[109] T. Xu, F. Yao, X. Zhou, Z. Deng, and D. You, "A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella," *Nucleic acids Research*, vol. 38, no. 20, pp. 7133–41, Nov. 2010.

[110] J. Casadesús and D. Low, "Epigenetic gene regulation in the bacterial world," *Microbiology and Molecular Biology Reviews : MMBR*, vol. 70, no. 3, pp. 830–56, Sep. 2006.

[111] S. Chen, L. Wang, and Z. Deng, "Twenty years hunting for sulfur in DNA," *Protein & cell*, vol. 1, no. 1, pp. 14–21, Jan. 2010.

[112] X. Xie, J. Liang, T. Pu, F. Xu, F. Yao, Y. Yang, Y.-L. Zhao, D. You, X. Zhou, Z. Deng, and Z. Wang, "Phosphorothioate DNA as an antioxidant in bacteria," *Nucleic Acids Research*, vol. 40, no. 18, pp. 9115–24, Oct. 2012.

[113] X. He, H.-Y. Ou, Q. Yu, X. Zhou, J. Wu, J. Liang, W. Zhang, K. Rajakumar, and Z. Deng, "Analysis of a genomic island housing genes for DNA S-modification system in Streptomyces lividans 66 and its counterparts in other distantly related bacteria," *Molecular Microbiology*, vol. 65, no. 4, pp. 1034–48, Aug. 2007.

[114] P. S. G. Chain, D. V Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter, "Genome Project Standards in a New Era of Sequencing," *Science*, vol. 326, no. October, pp. 4–5, 2009.

[115] C. M. Fraser, J. A. Eisen, K. E. Nelson, I. T. Paulsen, and S. L. Salzberg, "The Value of Complete Microbial Genome Sequencing (You Get What You Pay For)," *Journal of Bacteriology*, vol. 184, no. 23, pp. 6403–6405, 2002.

[116] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell, "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence," *Nature*, vol. 393, no. 6685, pp. 537–544, 1998.

[117] V. Barbe, S. Cruveiller, F. Kunst, P. Lenoble, G. Meurice, A. Sekowska, D. Vallenet, T. Wang, I. Moszer, C. Médigue, and A. Danchin, "From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later," *Microbiology*, vol. 155, no. Pt 6, pp. 1758–75, Jun. 2009.

[118] J. F. Heidelberg, J. a Eisen, W. C. Nelson, R. a Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellers, L. McDonald, T. Utterback, R. D. Fleishmann, W. C. Nierman, O. White, S. L. Salzberg, H. O. Smith, R. R. Colwell, J. J. Mekalanos, J. C. Venter, and C. M. Fraser, "DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae," *Nature*, vol. 406, no. 6795, pp. 477–83, Aug. 2000.

[119] E. P. C. Rocha, "Order and disorder in bacterial genomes," *Current Opinion in Microbiology*, vol. 7, no. 5, pp. 519–27, Oct. 2004.

[120] E. P. C. Rocha, "The organization of the bacterial genome," *Annual Review of Genetics*, vol. 42, pp. 211–33, Jan. 2008.

[121] E. Couturier and E. P. C. Rocha, "Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes," *Molecular Microbiology*, vol. 59, no. 5, pp. 1506–18, Mar. 2006.

[122] E. P. C. Rocha and A. Danchin, "Essentiality, not expressiveness, drives gene-strand bias in bacteria," *Nature Genetics*, vol. 34, no. 4, pp. 377–378, 2003.

[123] S. Zhou, F. Wei, J. Nguyen, M. Bechner, K. Potamousis, S. Goldstein, L. Pape, M. R. Mehan, C. Churas, S. Pasternak, D. K. Forrest, R. Wise, D. Ware, R. a Wing, M. S. Waterman, M. Livny, and D.

C. Schwartz, “A single molecule scaffold for the maize genome,” *PLoS genetics*, vol. 5, no. 11, p. e1000711, Nov. 2009.

[124] C. Honisch, A. Raghunathan, C. R. Cantor, B. Ø. Palsson, and D. Van Den Boom, “High-throughput mutation detection underlying adaptive evolution of Escherichia coli -K12,” *Genome Research*, no. 14, pp. 2495–2502, 2004.

[125] C. J. Davidson, A. P. White, and M. G. Surette, “Evolutionary loss of the rdar morphotype in Salmonella as a result of high mutation rates during laboratory passage,” *The ISME journal*, vol. 2, no. 3, pp. 293–307, Mar. 2008.

[126] C. D. Herring and B. Ø. Palsson, “An evaluation of Comparative Genome Sequencing (CGS) by comparing two previously-sequenced bacterial genomes,” *BMC genomics*, vol. 8, no. 274, Jan. 2007.

[127] C. a Fux, M. Shirtliff, P. Stoodley, and J. W. Costerton, “Can laboratory reference strains mirror ‘real-world’ pathogenesis?,” *Trends in Microbiology*, vol. 13, no. 2, pp. 58–63, Mar. 2005.

[128] J. a Lanie, W.-L. Ng, K. M. Kazmierczak, T. M. Andrzejewski, T. M. Davidsen, K. J. Wayne, H. Tettelin, J. I. Glass, and M. E. Winkler, “Genome Sequence of Avery’s Virulent Serotype 2 Strain D39 of Streptococcus pneumoniae and Comparison with That of Unencapsulated Laboratory Strain R6,” *Journal of Bacteriology*, vol. 189, no. 1, pp. 38–51, Jan. 2007.

[129] C. D. Herring, A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. Ø. Palsson, “Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale,” *Nature genetics*, vol. 38, no. 12, pp. 1406–12, Dec. 2006.

[130] T. M. Conrad, A. R. Joyce, M. K. Applebee, C. L. Barrett, B. Xie, Y. Gao, and B. Ø. Palsson, “Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations,” *Genome biology*, vol. 10, no. 10, p. R118, Jan. 2009.

[131] Z. D. Blount, J. E. Barrick, C. J. Davidson, and R. E. Lenski, “Genomic analysis of a key innovation in an experimental Escherichia coli population,” *Nature*, vol. 489, no. 7417, pp. 513–8, Sep. 2012.

[132] M. Park, S. Lu, S. H. Ryu, B. S. Chung, W. Park, C.-J. Kim, and C. O. Jeon, “Flavobacterium croceum sp. nov., isolated from activated sludge,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. Pt 10, pp. 2443–7, Oct. 2006.

[133] M. Park, S. H. Ryu, T.-H. T. Vu, H.-S. Ro, P.-Y. Yun, and C. O. Jeon, “Flavobacterium defluvii sp. nov., isolated from activated sludge,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. Pt 2, pp. 233–7, Feb. 2007.

[134] H.-Y. Weon, M.-H. Song, J.-A. Son, B.-Y. Kim, S.-W. Kwon, S.-J. Go, and E. Stackebrandt, “Flavobacterium terrae sp. nov. and Flavobacterium cucumis sp. nov., isolated from greenhouse soil,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. Pt 7, pp. 1594–8, Jul. 2007.

[135] K. T. Konstantinidis and J. M. Tiedje, “Genomic insights that advance the species definition for prokaryotes,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2567–72, Feb. 2005.

[136] J. Goris, K. T. Konstantinidis, J. a Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje, “DNA-DNA hybridization values and their relationship to whole-genome sequence similarities,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. Pt 1, pp. 81–91, Jan. 2007.

[137] M. Deloger, M. El Karoui, and M.-A. Petit, “A genomic distance based on MUM indicates discontinuity between most bacterial species and genera,” *Journal of Bacteriology*, vol. 191, no. 1, pp. 91–9, Jan. 2009.

[138] M. Richter and R. Rosselló-Móra, “Shifting the genomic gold standard for the prokaryotic species definition,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 45, pp. 19126–31, Nov. 2009.

[139] J. Z.-M. Chan, M. R. Halachev, N. J. Loman, C. Constantinidou, and M. J. Pallen, “Defining bacterial species in the genomic era : insights from the genus Acinetobacter,” *BMC Microbiology*, vol. 12, no. 1, p. 302, Jan. 2012.

[140] P. M. Holland, R. D. Abramson, R. Watson, and D. H. Gelfand, “Detection of specific polymerase chain reaction product by utilizing the 5’----3' exonuclease activity of Thermus aquaticus DNA polymerase,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 16, pp. 7276–80, Aug. 1991.

[141] A. Sette and R. Rappuoli, “Reverse Vaccinology: Developing Vaccines in the Era of Genomics,” *Immunity*, vol. 33, no. 4, pp. 530–541, 2010.

[142] F. C. Cabello, “Heavy use of prophylactic antibiotics in aquaculture: a growing problem for human and animal health and for the environment,” *Environmental Microbiology*, vol. 8, no. 7, pp. 1137–44, Jul. 2006.

Annexes

Annexe 1 : Composition des tampons et séquences des oligonucléotides pour la synthèse des sondes

Tampon NDS : EDTA 100mM pH8 et 1% N Lauroyl sarcosine

Tampon SSC 20X (Saline Sodium Citrate): 3.0 M citrate de sodium, 0.3 M NaCl, pH 7.5

Tampon SSC 2X : 100 mL SSC 20X dans QSP 1L d'eau distillée

Tampon SSC 2X/ 0,1 % SDS : 100 mL SSC 20X, 10 mL SDS 10% dans QSP 1L d'eau distillée

Tampon SSC 0,1X/ 0,1 % SDS : 5 mL SSC 20X, 10 mL SDS 10% dans QSP 1L d'eau distillée

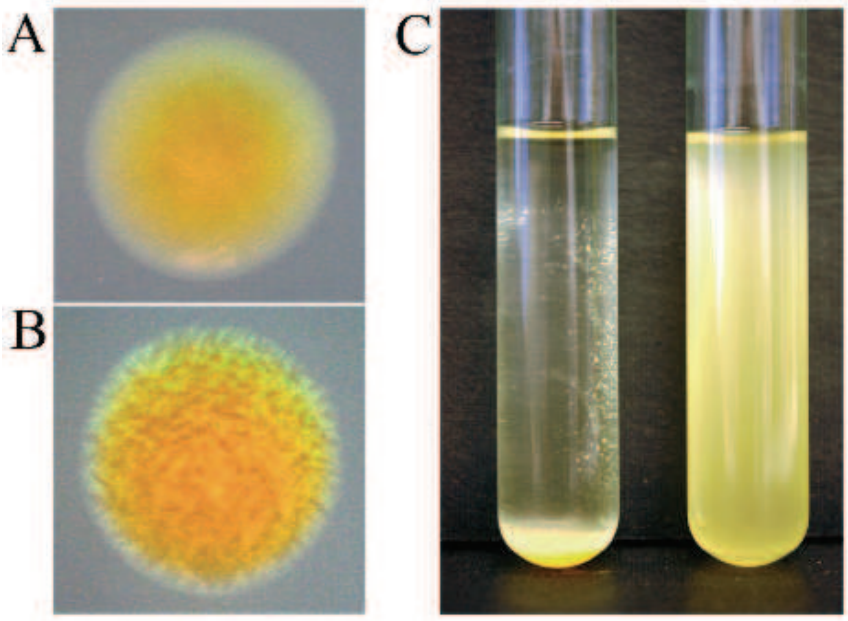
Solution de travail NaOH : 40 ml de NaOH 10N dans QSP 1L d'eau distillée

Solution de travail HCL : 20,8 ml de HCL 37% dans QSP 1L d'eau distillée

Séquences des oligonucléotides pour la synthèse des sondes :

Nom du primer	**Séquence**	**Tm en °C**
P1_fw	TGCCAACAGCATCAGTACA	61,6
P1_rev	CCTGCTGGCACAATTACA	60,9
P2_fw	CATCTCATATTTTAATCACTGTCG	59,1
P2_rev	GGAGCTTCCATTGACTAACAT	59,6
P3_fw	TACGATCCAAGAAAATGGA	58,1
P3_rev	GTTTTATCTAACTTCGTTATTCA	53,7
P4_fw	GGGAACTGGAGTTACAGAAGAT	60,1
P4_rev	GGTATCAGTTTGGGTCCG	60,4

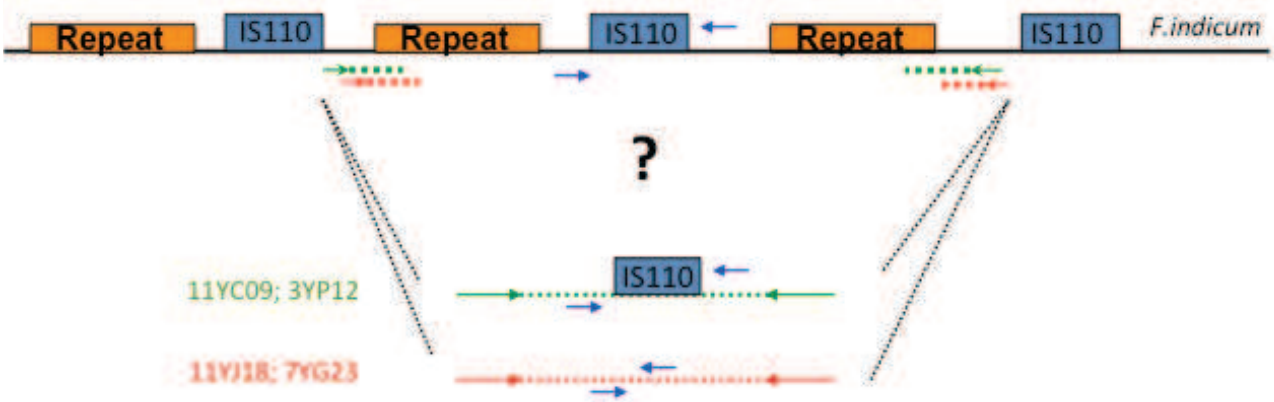
Annexe 2 : Variation de phase chez F. psychrophilum



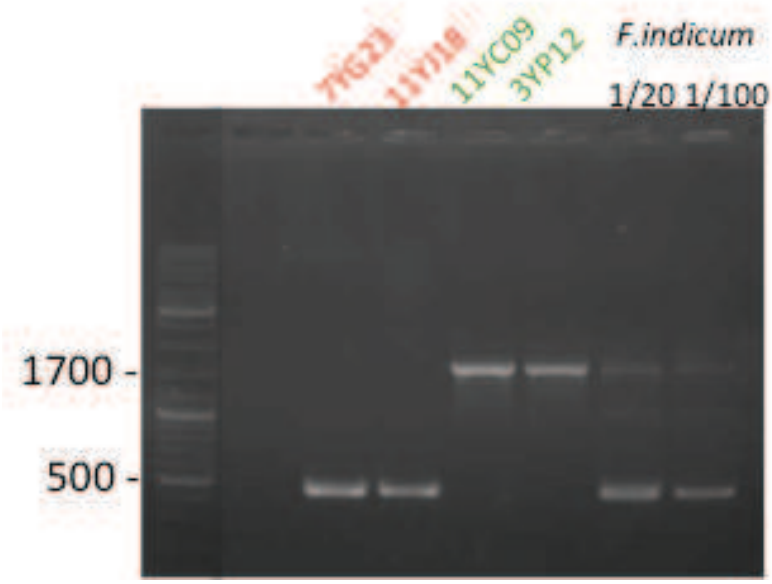
Cultures de *Flavobacterium psychrophilum* en milieux solides et liquides. Apparence des colonies « lisses » (S) en A et « rugueuses » (R) en B sur boîte de pétri. En milieu liquide (C), les colonies « lisses » s'auto agglutinent au fond du tube (gauche) tandis que les colonies « rugueuses » ne s'agglutinent pas et montrent une dispersion uniforme (droite). D'après Tom Wiklund, extrait de : Phase variation in *Flavobacterium psychrophilum*: characterization of two distinct colony phenotypes [75].

Annexe 3 : Schéma de la stratégie de finition d'une région du génome de F. indicum

A



B



A. Organisation schématique d'une région particulière du génome de *F. indicum*. Les différentes copies de l'IS100 (en bleu) sont présentes dans cette zone contenants des répétitions (en orange). En dessous, la représentation schématique de clones chevauchants cette région contenants une copies de l'IS (en vert) et ne la contenants pas (en rouge). Les flèches bleues indiquent la position des primers dessinés pour la finition de cette région.

B. Vérification des prédictions par PCR réalisés sur les inserts des clones chevauchants et l'ADN génomique de *F. indicum*. Les chiffres indiquent les poids moléculaires en pb.

Annexe 4 : Caractéristiques des séquences génomiques disponibles pour le genre Flavobacterium (Juin 2013)

Nom	Taille (pb)	Source d'isolation ou d'échantillonnage	Numéro d'accès
F. psychrophilum JIP02/86 **(p)**	2861988 **(c)**	Truite arc-en-ciel, France	NC_009613.3
F. psychrophilum THC 02/90 **(p)**	2783852 **(c)**	Saumon Coho, USA	Non publié
F. branchiophilum FL-15 **(p)**	3559884 **(c)**	Poisson chat, Hongrie	NC_016001.1
F. columnare ATCC 49512 **(p)**	3162432 **(c)**	Truite brune, France	NC_016510.2
F. johnsoniae UW101	6096872 **(c)**	Sol, Angleterre	NC_009441.1
F. indicum GPTSA100-9^{T}	2993089 **(c)**	Source d'eau chaude, Inde	NC_017025.1
F. frigidimaris KUC1	5624955	Eau de mer, Antarctique	Non publié
F. glaciei 0499^{T}	3227944	Glacier, Chine	Non publié
F. rivuli DSM 21788^{T}	4487368	Eau calcaire, Allemagne	NZ_ARKJ00000000.1
F. frigoris PS1	3934101	Lac gelé, Antarctique	NZ_AHKF00000000.1
Flavobacterium sp. F52	5336938	Rhizosphère du poivron	NZ_AKZQ00000000.1
Flavobacterium sp. CF136	5102016	Rhizosphère, USA	NZ_AKJZ00000000.1
Flavobacterium sp. SCGC AAA160-P02	2543104	Océan Atlantique, Golf du Maine	NZ_ARTD00000000.1
Flavobacterium sp. ACAM 123	3955605	Lagon marin, Antarctique	NZ_AJXL00000000.1
Flavobacterium sp. B17	4165754	Pousse de riz, Japon	NZ_BACY00000000.1
Flavobacterium sp. WG21	5198514	Lac, USA	NZ_AMYW00000000.1

(p) : espèce pathogène pour les poissons
(c) : génome circulaire

Annexe 5 : ANI pour le genre Flavobacterium

	psychrophilum JIP02/86	*psychrophilum* THC 02/90	*branchiophilum* FL-15	*columnare* ATCC 49512	*johnsoniae* UW101	*indicum* GPTSA100-9^T	*frigidimaris* KUC1	*glaciei* 0499^T	*rivuli* DSM 21788^T	*frigoris* PS1	sp. F52	sp. CF136	sp. SCGC AAA160-P02	sp. ACAM123	sp. B17	sp. WG21
psychrophilum JIP 02/86	100															
psychrophilum THC 02/90	**99,35**	100														
branchiophilum FL-15	71,8	71,4	100													
columnare ATCC 49512	72,1	71,8	69,9	100												
johnsoniae UW101	72,4	72,3	71,3	70,2	100											
indicum GPTSA100-9^T	71	70,8	70	72	70,5	100										
frigidimaris KUC1	72,5	72	71	69,8	80,4	70,1	100									
glaciei 0499^T	73,4	73	71,8	70,4	75	70,7	75,2	100								
rivuli DSM 21788^T	68,6	68,2	67,3	67,8	67,6	67,3	67,5	67,9	100							
frigoris PS1	72,8	72	71	70,2	73,4	70,1	73,7	76,5	67,6	100						
sp. F52	72,4	72	71,1	70,5	82,6	70,1	79,3	74,7	67,7	73,4	100					
sp. CF136	72,9	72,4	71,4	70,4	79,5	70,4	80,8	75,8	67,9	74,4	78,9	100				
sp. SCGC AAA160-P02	68,8	68,6	67,9	68,6	68,4	68,9	67,9	67,8	65,5	67,9	67,9	68,1	100			
sp. ACAM123	72,1	71,7	70,7	69,6	72,9	70	73,1	75,8	67,5	77,5	73	73,9	67,4	100		
sp. B17	66,4	65,3	65,1	65,1	65,9	65,7	65,2	65	64,3	65,1	65,6	65,4	65,2	64,8	100	
sp. WG21	72,2	71,5	70,7	69,5	79,3	69,9	81,4	74,9	67,8	73,6	78,7	79,6	67,7	73,2	65,5	100

Pourcentage moyen d'identité nucléotidique (ANI) entre les séquences génomiques disponibles pour le genre *Flavobacterium* en Juin 2013. Calculs (ANIb) réalisés à partir des séquences listés en Annexe 4 à l'aide du logiciel JSpecies [138].