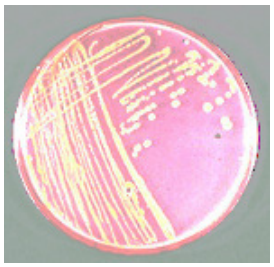
université PARIS DIDEROT PARIS 7

Master 2 Biologie Informatique — Année 2010/2011



Analyse des données de séquençage à haut débit de plusieurs souches de la bactérie *Flavobacterium psychrophilum*

INRA

Auteur : Guillaume Fernandez
Numéro d'étudiant : 2100555696
Encadrants : Dr. Pierre Nicolas, Dr. Eric Duchaud

Institut National de la Recherche Agronomique
Coencadrement des unités :
Virologie et Immunologie Moléculaires (UR892)
Mathématique, Informatique et Génome (UR1077)
Domaine de Vilvert, Centre INRA de Jouy-en-Josas

Remerciements

La thématique à l'interface entre les enjeux agronomiques, la biologie expérimentale et l'analyse *in silico* et, d'autre part, les objectifs en prise avec l'actualité scientifique étaient un défi particulièrement motivant. Je souhaite remercier particulièrement l'ensemble de l'équipe génome de MIG. Particulièrement Valentin Loux, Fabien Melchiore et Cyprien Guérin pour nos discussions sur l'assemblage des génomes et leurs limites techniques. Guillaume Launay a contribué à mon acquisition de bonnes pratiques de code dès mon arrivée et sa rigueur de travail a été de la meilleure influence. Je remercie pour leur bonne humeur quotidienne (modulo quelques lignes de codes parfois récalcitrantes) les collègues de notre bureau : Pascal, Dialecti et Frédéric. Enfin, ce stage a pu être mené grâce à mes encadrants Eric Duchaud et Pierre Nicolas qui m'ont donné l'opportunité de conduire ce stage qui m'a permis de progresser encore sur la manipulation des données génomiques.

« *It is important to emphasize that assembly is not a solved problem, in particular with very short reads, and there will continue to be considerable algorithmic improvement*. »

Zerbino & Birney, 2008 [1]

Sommaire

Introduction

A. *Flavobacterium psychrophilum* engendre des pertes économiques

Les maladies causées par les bactéries pathogènes du genre *Flavobacterium* constituent un problème important pour le développement d'une activité piscicole durable. L'espèce *Flavobacterium psychrophilum* est responsable d'une pathologie sévère chez les téléostéens de la famille des Salmonidae (truites, saumons). Celle-ci prend deux formes cliniques. La forme connue sous le nom de « maladie d'eau froide » consiste en de nombreuses lésions nécrotiques des tissus des individus adultes (voir la figure 1) particulièrement ceux de la surface du corps, des branchies, du cartilage. Les individus juvéniles souffrent d'une mortalité sévère associée à une hémorragie septicémique, forme connue sous le nom de « syndrome de la truite arc-en-ciel juvénile[1] » [2]. Les signes extérieurs de l'anémie provoquée est révélée par des branchies pâles tandis que les individus paraissent léthargiques et restent à la surface de l'eau [3].

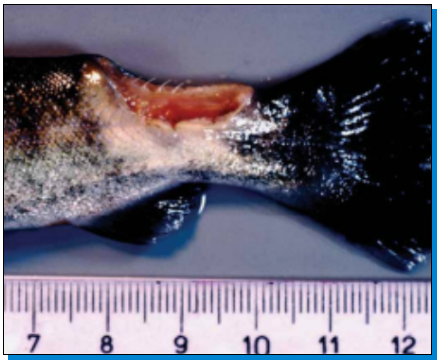


Figure 1 : profonde lésion du pédoncule caudal provoquée par *Flavobacterium psychrophilum* dans une truite arc-en-ciel (*Oncorhynchus mykiss*). Les tissus de la peau et du muscle ont été détruits, exposant la corde spinale. L'échelle est en centimètres. D'après [2].

Aucun vaccin efficace n'est pour le moment disponible et les antibiotiques à base de sulphonamides, de furanace et d'oxytetracycline (qui ont un coût économique) représentent le seul recours actuel aux infections par la bactérie [3]. Pour concevoir un vaccin il est essentiel de caractériser quelques isolats représentatifs de la diversité biologique de *F. psychrophilum*. Individualiser un cœur génomique commun et comparer les parties variables peut notamment concourir à la détermination des potentialités différentielles d'infection des différentes souches. Dans ce contexte, l'objectif de ce stage était d'assembler les génomes de nouvelles souches séquencées en 2009 et 2010 et d'y étudier les répertoires de gènes s'y trouvant.

1 En anglais : *rainbow trout fry syndrome*.

B. Caractéristiques de *Flavobacterium psychrophilum*

La bactérie *Flavobacterium psychrophilum* appartient à la famille des Flavobactéries, dans le phylum des *Bacteroidetes*. La pathologie engendrée par l'infection a été décrite dès 1846. La bactérie a été isolée en 1960 dans l'état de Washington [3]. Jusque dans les années 1980, *Flavobacterium psychrophilum* était connue uniquement en Amérique du Nord où elle infectait le saumon Coho et la truite arc-en-ciel. A la fin des années 1980 le pathogène a également été identifié en France, en Allemagne puis au Japon au début des années 1990. Le terme de maladie des eaux froides provient du fait que les épizooties sont prévalentes lorsque la température de l'eau est inférieure à 10°C[2] [3].

1) Structure de la population

Une étude de la structure de la population examinant le polymorphisme de 11 loci codants dans 50 isolats a montré une faible diversité génétique chez *Flavobacterium psychrophilum* [4], l'existence de complexes clonaux[3] témoignant d'une évolution verticale et un taux de recombinaison élevé (la plupart des isolats portent la même combinaison unique d'allèles quand 33 associations uniques différentes ont été repérées). Le type de polymorphisme mis en évidence est compatible avec une situation quasi-panmictique[4] où la recombinaison est suffisante pour disperser les allèles entre toutes les souches. Le mécanisme de recombinaison reste à démontrer, bien que des plasmides conjugatifs de 3kb [3] [5] et des transposons aient été trouvés [2]. La propagation multimodale du pathogène est suspectée (notamment par une adhérence très importante aux œufs [6]). Elle pourrait expliquer son évolution génomique dont les marqueurs témoignent à la fois d'une composante verticale et horizontale.

2 La bactérie se développe de 4 à 23°C et le temps de génération optimal de cette souche est de 2 heures à 15° C [3].

3 Les groupes d'associations uniques d'allèles dont les variants diffèrent d'un seul locus sont appelés complexes clonaux. Ces isolats qui ne se distinguent que par un seul locus sont généralement suspectées d'être le résultat d'une diversification verticale depuis un ancêtre commun.

4 Panmixie : en génétique des populations, se dit des populations où les croisements ou échanges génétiques se réalisent au hasard (sans sélection) entre tous les individus d'une population. Cela implique que les individus soient tous au contact dans le même environnement, avec pour un individu donné la même probabilité de rencontrer chacun de tous les individus identifiés de la population.

2) Caractéristiques génomiques

Les données génomiques présentées dans le tableau 1 reposent sur un premier séquençage Sanger de la souche virulente JIP02/86 (ATCC 49511) publié par le laboratoire en 2007 [2]. La stratégie était non-ordonnée[5].

Tableau 1 : caractéristiques générales du génome de la souche JIP02/86 de *F. psychrophilum*

Taille du génome	2 861 988 pdb	Longueur moyenne des gènes	1 003 pdp
Contenu GC	32,54%	CDS[6] avec une fonction prédite	1318
Séquences codantes	2432	CDS similaires à des protéines de fonction inconnue	701
Pseudogènes	20	CDS sans similarités	401
Densité codante	84,50 %	Nombre de séquences d'insertion (IS)	74

3) Génome et écologie de la bactérie

Les bases génétiques de la virulence de *F. psychrophilum* sont caractéristiques de la production de toxines, d'un système de sécrétion et d'un métabolisme spécifiques. Le génome déjà disponible en contient les traces : une dizaine de protéases sécrétées, des métalloprotéases similaires à celles connues chez les champignons pathogènes, des protéines de transport de protéases et d'adhésines (par exemple porT) et de nombreuses peptidases permettant la dégradation des protéines de l'hôte (pour y trouver une source énergétique) [2]. Des résultats expérimentaux accompagnent ces données. Par exemple, la sécrétion d'une collagénase extracellulaire de type II est avérée [7]. Le système de sécrétion Por (en abrégé PorSS) semble unique et restreint aux *Bacteroidetes* (homologues trouvés chez *Porphyromonas gingivalis* et *Flavobacterium johnsoniae*) [8]. Enfin, on trouve dans le génome des déterminants de la biosynthèse d'exopolysaccharides (quatre alginates O-acétyl-transférases), éléments du biofilm permettant l'adhérence de la bactérie à ses hôtes [2].

C. Déroulement du stage

Au cours de ce stage l'assemblage des données issues du séquençage du génome de nouvelles souches de *F. psychrophilum*, la détection des potentiels codants et une première annotation fonctionnelle ont été menés. Chacun de ces pas offre plusieurs choix techniques qui doivent être discutés en prise avec les objectifs recherchés pour obtenir le compromis de paramétrage

5 La finition a été réalisée par PCR et par séquençage direct de clones d'ADN génomique.

6 CDS : séquence codante (*coding sequence*) débutant par un signal de début de la transcription et terminant par un signal d'arrêt de la transcription.

et de procédure qui leur correspond le mieux. En l'état actuel il n'existe pas de procédure entièrement universelle de séquençage et d'assemblage des génomes.

Enfin, du fait de limitations techniques, les méthodes récentes de séquençage à haut débit durant synthèse, si elles autorisent le traitement de gros volumes de données, voient discuter la qualité de leurs produits notamment à cause des répétitions génomiques [9]. Ainsi, dans l'assemblage *de novo* du génome d'un chinois de l'ethnie Han (séquençage publié en 2010[7] à l'aide d'une méthode à haut débit[8]) 99,1% des séquences dupliquées (validées expérimentalement) étaient manquantes dans le génome par rapport au génome de référence (NCBI Build 36) [9]. D'après certains chercheurs, une comparaison critique des nouveaux génomes devrait donc être menée en comparaison avec des standards connus[9] [9]. Le laboratoire dispose d'une souche de référence obtenue avec des méthodes classiques offrant l'opportunité d'effectuer l'évaluation critique mise en avant par Eichler *et al*. Avant d'effectuer le travail de bioanalyse, il semblait donc important de déterminer la qualité des séquences de *F. psychrophilum* assemblées, afin de disposer d'un indicateur, au moins à minima, des limites du séquençage et de l'assemblage.

^-^
=°.°=

☞ Bilan du chapitre

☑ *Flavobacterium psychrophilum* induit une mortalité significative chez les Salmonidés.

☑ Aucun vaccin efficace n'est disponible pour les centres d'aquaculture.

☑ La contamination est multimodale (oralement, par contact, par les parents).

☑ L'évolution du génome de plusieurs souches montre que la diversité génétique est faible (analyse multi-locus) ; il existe des traces de transmissions horizontale (recombinaison élevée) et verticale (complexes clonaux).

☑ Le génome fini de la souche JIP02/86 a été publié en 2007, il fait 2,9 Mb et comporte 2432 séquences codantes.

☑ De nombreuses séquences similaires à des facteurs de virulence connus dans d'autres organismes pathogènes (toxicité et adhésion cellulaire) ont pu y être annotés.

☑ Des travaux montrent une perte de données issue de l'utilisation des nouvelles technologies de séquençage : une évaluation critique des nouveaux assemblages a été menée avant toute recherche fonctionnelle.

7 Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 20, 265-272

8 Séquençage Illumina produisant des séquences individuelles de l'ordre de 100 paires de bases.

9 *NGS technologies typically generate shorter sequences with higher error rates. [...] It is, therefore, expected that assembly of longer repeats and duplications will suffer from this short read length. [...] It is the responsability of the scientific community to enforce standards of quality that can be measured and assessed. […] Large molecule, high-quality sequencing should not be abandonned.* Eichler *et al*., 2011.

Matériel et méthodes

1) Le matériel d'étude

Les souches disponibles pour ce travail sont présentées dans le tableau 2. Elles ont été obtenues au Génoscope avec la méthode de séquençage à haut débit durant synthèse Solexa. L'utilisation de cette technologie permet d'obtenir une profondeur d'échantillonnage supérieure à 200X[10] pour chaque souche. Chaque jeu contient 10 millions de lectures de 108 paires de bases [10].

Tableau 2 : souches de *Flavobacterium psychrophilum et matériel de l'étude*

Souche	*N$^{um.}$ MIG/VIM*	*Témoins positifs*	*Annotation exploitée*
Fp_JIP16-00	4		
Fp_JIP08-99	8		
THC02-90	12	☑ Nouveau séquençage	version 5
Fp_LVDJ-XP189	22		
Fp_ NCIMB_1947T	23		
JIP02-86	24	☑ Nouveau séquençage	version 8
Fp_FPC_831	39		
Fp_FPC_840	40		

2) Procédure d'assemblage du laboratoire

La figure 2 décrit la procédure d'assemblage définie empiriquement au laboratoire et susceptible d'être améliorée ou modifiée.

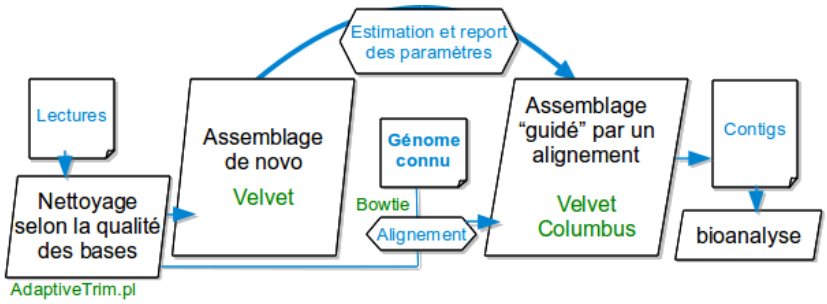


Figure 2 : procédure d'assemblage disponible au laboratoire

Pour des raisons techniques la qualité des bases décroit en fin d'analyse. Un nettoyage des lectures a été réalisé (retrait de la première et de la dernière base tant que la qualité est inférieure à un seuil de 10, élimination des lectures contenant un *N* ou plus courtes que 20

10 A 200X, le volume des lectures du séquençage représente 200 fois le volume du génome réel.

nucléotides[11]). Les lectures ont ensuite été assemblées avec Velvet [1]. Afin d'optimiser automatiquement les paramètres, une couche logicielle supplémentaire a été ajoutée à l'assembleur. Le script *VelvetOptimiser* (version 2.1.7) lance plusieurs exécutions de Velvet et, sur la base de la fragmentation du génome obtenu, élimine tous les assemblages sauf celui qui semble résulter du meilleur paramétrage. (Dans certains cas (voir les résultats) les contigs produits à cet étape ont été exploités). L'alignement des lectures avec un des génomes connus a été réalisé avec Bowtie [11]. L'utilisation du module *Columbus* [12] de Velvet a ensuite permis de prendre en compte cet alignement dans la construction de l'assemblage[12]. Les paramètres déterminés pendant les assemblages *de novo* exploratoires par *VelvetOptimiser* ont été reportés pour cet assemblage final. Un fichier de contigs a alors été obtenu. La fragmentation du génome a alors été estimée, entre autres à l'aide d'un indicateur simple, le N50.

> **Définition du N50**
> Le N50 est une mesure de la fragmentation du génome assemblé en contigs. Lorsque les contigs sont triés par taille et que l'on somme les contigs un par un du plus grand au plus petit, le N50 est la longueur du dernier contig obtenu lorsque tous les contigs déjà parcourus couvrent la moitié de la taille du génome. Plus le N50 est élevé, moins le génome est fragmenté[13].

3) Réglage de la sensibilité de l'assembleur Velvet avec le paramètre k

La paramètre k est déterminant pour la qualité des assemblages. Un paramètre k très élevé aboutit à un assemblage très fragmenté et inexploitable. Velvet utilise une approche basée sur les graphes de Bruijn pour concilier lectures courtes et succès de l'assemblage [1]. Au cours de la première partie du traitement les lectures sont découpées en morceaux de taille plus petite encore : les k-mers (*k* est taille du k-mer) [1]. Un graphe orienté est ensuite créé. Ses nœuds sont un ensemble de k-mer se recouvrant – les nœuds peuvent donc avoir une taille variable. Les nœuds peuvent être connectés par un arc direct si le dernier k-mer du nœud de départ se recoupe avec le premier k-mer du nœud de destination. De petits k-mers augmentent

11 L'équipe a précédemment déterminé que le nettoyage des lectures est nécessaire et efficace pour réduire les temps de calcul et améliorer la qualité de l'assemblage [10]

12 Dans ce rapport l'utilisation de *Columbus* sera désignée par le terme « assemblage avec référence » ou « assemblage guidé ».

13 Un mauvais assemblage peut avoir un N50 élevé. L'habitude de juger la qualité d'un alignement par le N50 est discutable. En effet, de grands contigs peuvent refléter une fusion de contigs trop agressive, introduisant des défauts d'assemblage.

la connectivité du graphe en augmentant la chance d'observer un recouvrement entre eux[14]. Quand cela arrive, il y a concaténation entre plusieurs blocs de nucléotides – figure 3.

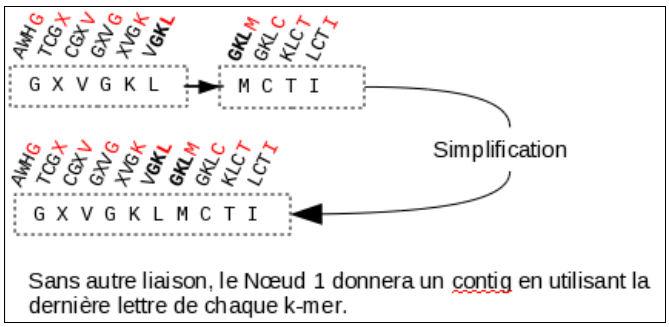


Figure 3 : exemple hypothétique de découpage des lectures en mots de taille k et d'assemblage. Les nœuds ont été reconstitués à partir d'une base de données de k-mer (AWHG, WHGC...) issue des lectures AWHGCCGKL et GKLTCI. Dans la figure, tous les caractères sont des nucléotides, même s'il sont représentés par commodité avec un alphabet de 26 lettres (Source : [1]). Pour plus de détail, voir l'annexe 2 page 24.

Cette approche bute sur les répétitions. Lorsque le k-mer *est* un motif répété des enchevêtrements se créent dans les liaisons des nœuds du graphe. Ce sont des points de cassures : obtenir une séquence en un seul morceau est impossible. (Pour comprendre en détail pourquoi, se reporter à l'annexe 3 « Répétitions et nœuds »). Plus le paramètre k est petit, plus nombreux seront les nucléotides pouvant constituer un motif fréquent, et donc des répétitions. Ainsi, de petits k-mer augmentent la connectivité du graphe mais aussi le nombre de répétitions ambigües dans le graphe. D'après Zerbino *et al.*, pour un même réglage de k, les résultats dépendent largement du génome, de la couverture, de la qualité de séquençage et de la longueur des lectures. Une approche consiste à tester plusieurs paramétrages en parallèle et à choisir de ne retenir empiriquement que celui qui produit le plus grand N50 [1]. C'est la méthode qui a été adoptée. Une autre méthode de réglage de k est d'estimer la couverture en k-mer des lectures mais elle ne sera pas abordée ici.

4) Analyse qualitative

☞ Objectifs de la méthodologie proposée

☑ Répondre à la question « comment les données issues du séquençage à haut débit influencent-elles la reconstruction des génomes de *F. psychrophilum* ? » ;

☑ Valider la procédure d'assemblage existante ou l'adapter pour obtenir des génomes convenablement assemblés.

Volga (un ensemble de scripts écrits pendant ce stage) calcule le N50 de l'assemblage, le nombre de contigs produits ou le nombre de gènes présents dans les annotations disponibles

14 D'une manière générale, l'usage de k-mer (à la place de l'approche traditionnelle d'assemblage qui recherche les recouvrements entre les lectures entières) augmente la « connectivité des lectures ». Il est également moins couteux en ressources de calcul de rechercher de petites séquences dans un dictionnaire que des longues. Ainsi, le temps de calcul et l'empreinte mémoire sont limités face au volume des projets à haut débit.

mais absents ou cassés dans les nouvelles données (première fonction : fournir des statistiques sur l'assemblage). Le programme offre également de comparer ces données avec des statistiques pour un jeu de lectures simulées. Volga produit un ensemble de contigs artificiels en découpant le génome de référence pour mimer les contigs issus d'un assemblage avec Velvet (les répétitions sont les points de cassure). La comparaison des statistiques génomiques sur les données réelles et simulées fait ressortir l'impact d'éventuels autres facteurs limitants de l'assemblage (deuxième fonction : montrer la qualité d'un assemblage par rapport à un jeu idéal). Avec Volga, la quantité d'information biologique perdue chez JIP02/86 et THC02/90 pendant la procédure d'assemblage (assemblage de novo ou assemblage guidé) a pu être testée. En fonction de tous les résultats, la procédure pour produire une version standard des souches de *Flavobacterium psychrophilum* a été adaptée.

5) Annotation du génome

Le programme SHOW (Sequence Homogeneity Watcher) développé par Pierre Nicolas a été utilisé pour détecter les potentiels codants dans les contigs assemblés [13]. La détection et classification des gènes repose sur la partition de la séquence d'ADN en régions de compositions homogènes en mots de longueurs variables. Les paramètres du modèle de Markov susceptible de représenter le potentiel codant ont été déterminés en utilisant le génome de JIP02/86 comme jeu d'entrainement. Ont été pris en compte les codons *start* et *stop*, la phase de la composition de la séquence codante, la présence d'un site de fixation du ribosome en amont de le la séquence codante et la possibilité pour les gènes de se recouvrir mutuellement. Une recherche de similarités des séquences prédites a été réalisée avec un algorithme heuristique (BLAST+). Les similarités ont été recherchées dans des séquences traduites en protéines. Les séquences à aligner étant peu divergentes la matrice de substitution utilisée était la BLOSUM80. L'attribution fonctionnelle par similarité restant une tâche sensible, seules les séquences codantes prédites avec une excellente probabilité par BLAST 2.2.25+ ont été extraites (supérieure à 0,99). Puis les alignements avec une e-value supérieure à 0,01 n'ont pas été pris en compte. Parmi celles-ci, seules les excellentes homologies ont été retenues pour l'annotation des nouvelles souches[15].

15 La politique était de minimiser le risque d'accepter une annotation fonctionnelle à tort. Le module qui analyse les sorties de BLAST+ (programmé dans le cadre du stage) voit de très bonnes annotations et les similarités de qualité juste inférieure. De plus, le module conserve uniquement la meilleure annotation fonctionnelle disponible (*i.e.* issue du meilleur alignement). De plus le module tente de confirmer une bonne similarité avec une séquence d'une souche par une très bonne similarité dans l'autre (puisque la base en comporte deux).

Résultats et discussion

A. Approche méthodologique sur les données de séquençage à haut débit

1) Souche JIP02/86

La perte d'information biologique par rapport à l'annotation faite sur la base du séquençage Sanger publié en [2] semble raisonnable. Avec le séquençage à haut débit, l'assemblage *de novo* (première étape de la procédure du laboratoire) semble satisfaisant avec une couverture du génome par les contigs proche de 90% (tableau 3). Seulement 4 des 630 contigs produits par l'assemblage ne peuvent être alignés sur la référence connue. 8 autres contigs ont pu être alignés par Volga avec une bonne probabilité que l'alignement ne soit pas dû au hasard mais une partie significative des contigs (plus de 100 paires de bases) n'a pas été inclue dans l'alignement. Ces contigs peuvent être issus d'une concaténation biologiquement abusive de lectures, ce sont potentiellement des chimères. En terme de séquences codantes, 160 gènes n'ont pas été retrouvés[16]. Sachant que le nombre de gènes référencés dans la version 8 de l'annotation est de 2508, cela représente 6,38% de séquences codantes manquantes dans les données de séquençage à haut débit. Nos résultats peuvent connaitre les biais rapportés par la communauté : l'absence d'un gène dans un assemblage peut facilement résulter d'un artefact technique empêchant sa détection plutôt que d'une réelle délétion dans le lignage [14]. Enfin, 308 gènes ne peuvent pas être trouvés en entier dans un contig. (1/ Un gène couvert[17] peut être fragmenté et réparti sur plusieurs extrémités de contigs. 2/ Il suffit qu'un gène ne soit pas couvert pour qu'il soit compté comme non-entier).

16 Les positions des 160 gènes dans le génome de référence n'étaient pas couvertes par les nouveaux contigs alignés contre celui-ci. Note : il faut qu'un gène possède au moins 100 paires de bases couvertes pour qu'il soit considéré comme présent par Volga. En effet, avec 100 paires de bases, il existe une chance que le gène soit détectable dans le génome, à défaut de contenir toute l'information biologique nécessaire à son identification fonctionnelle et structurelle.

17 Pour Volga un gène est « couvert » s'il est totalement couvert ou bien s'il y a un défaut mineur de couverture (moins de 100 paires de bases non couvertes). Cette partie des gènes non-entiers pourrait être détectée lors de la phase de recherche de gènes.

Tableau 3 : assemblage *de novo* avec k optimisé à 31. La taille du génome de JIP02/86 est de 2860382 bases.

	JIP02/86 de *novo*
Espace des *k* testés par *VelvetOptimiser*	27 à 47
k retenu pour le meilleur N50	31
Contigs produits par Velvet :	
N 50	17 228
Nombre total de contigs (>2k)	630
Taille cumulée des contigs	2 725 288
contigs non alignés	4
Chimères (contigs avec plus de 100 pb non alignées)	8
Gènes non couverts par les contigs	160
Gènes non-entiers	308
Couverture totale en nucléotides (%)	89,75

Une information connexe peut être apportée avec la simulation du lot de contigs produits par Velvet si les répétitions étaient le seul problème de l'assemblage. Les statistiques sont du même ordre de grandeur que celles des données réelles (voir le tableau 4). Le jeu de données artificiel est donc globalement bien simulé.

Tableau 4 : comparaison de l'assemblage *de novo* avec k optimisé à 31 et des statistiques pour le jeu de données simulées pour le même k

Contigs produits par Velvet	JIP « artificiel »	JIP02/86 *de novo*
N 50	20 624	17 228
Nombre total de contigs (>2k)	326	630
Taille cumulée des contigs	2 694 523	2 725 288
contigs non alignés	0	4
Chimères (contigs avec plus de 100 pb n	0	8
Gènes non couverts par les contigs	83	160
Gènes non-entiers	154	308
Couverture totale en nucléotides (%)	93,82	89,75

En revanche les données simulées sont moins fractionnées que les données réelles, laissant supposer l'intervention d'autres facteurs limitant dans l'assemblage réel, tels que les erreurs de séquençage ou des biais de couverture (absence de couverture dans une région du génome).

2) Souche THC 02/90

Un k égal à 41 a été déterminé. La perte d'information biologique semble raisonnable (tableau 5, première colonne).

Tableau 5 : assemblage *de novo* de THC02/90 avec k optimisé à 41 (et jeu de donné simulé). Le génome Sanger de THC 02/90 fait 2 783 852 bases.

	THC02/90 de *novo*	THC02/90
Espace des *k* testés par *VelvetOptimiser*	27 à 47	artificiel
k retenu pour le meilleur N50	41	
Contigs produits par Velvet :		
N 50	16 439	33 584
Nombre total de contigs (>2k)	563	241
Taille cumulée des contigs	2 591 691	2 637 350
contigs non alignés	1	0
Chimères (contigs avec plus de 100 pb non alignées)	3	0
Gènes non couverts par les contigs	83	68
Gènes non-entiers	238	118
Couverture totale en nucléotides (%)	94,50	93,97

Si les répétitions étaient le seul facteur limitant de l'assemblage, le lot de contigs produits (68 gènes non couverts) serait moins fractionné qu'il ne l'est en réalité (83 gènes non couverts) (tableau 5, deuxième colonne).

Enfin, 0,30 % des régions du génome référentiel sont couvertes plus d'une fois par les contigs (soit environ 8 300 nucléotides du référentiel)[18]. Ce nombre doit être quasiment nul (c'est le cas), autrement l'assemblage n'a aucune valeur. Les résultats de l'investigation menée à la suite de cette observation sont rapportés dans le paragraphe suivant.

☞ Bilan du paragraphe

☑ L'assemblage *de novo* d'un génome séquencé avec une technologie à haut débit produisant de courtes lectures conduit à une perte limitée d'information biologique.

☑ La première partie de la procédure du laboratoire semble validée.

☑ La valeur de k choisie en optimisant automatiquement l'assemblage selon le N50 est différente pour les souches testées : la comparaison directe des statistiques dressées par Volga pour les souches THC02/90 et JIP02/86 n'est pas l'approche adoptée. Ce sont les ordres de grandeurs de chaque souche prise indépendamment qui permettent de confirmer la bonne qualité générale de la procédure.

☑ Un module générant aléatoirement des erreurs de séquençage dans le jeu simulé pourrait être ajouté à Volga (seul le problème des répétitions est pris en compte). Dans la littérature, un exemple de simulation peut être trouvé en [15].

3) Positions du génome couvertes plus d'une fois

La couverture du génome de référence par les contigs est homogène (figure 4, THC02/90). Les 0,3 % de régions de référence couvertes plus d'une fois ne le sont que de manière très localisée. On repère aussi des défauts de couverture sous forme de pics ou de régions plus étendues.

18 Le même pourcentage a été constaté pour la souche JIP02/86.

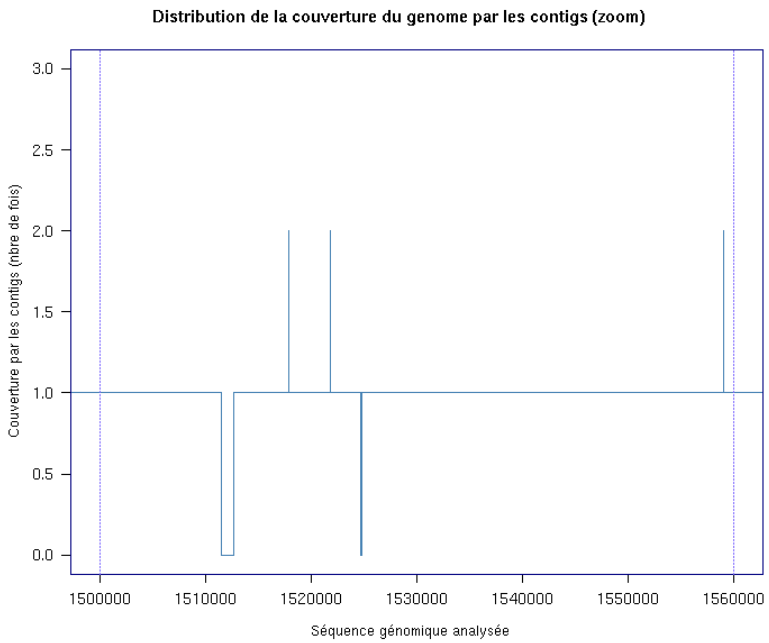
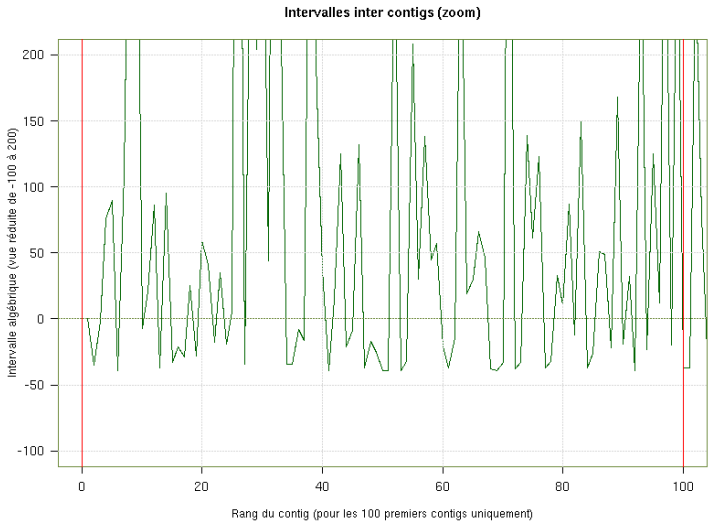


Figure 4 : représentation graphique de la couverture du génome de référence THC02/90 par les contigs. L'intervalle des nucléotides 1500000 à 1560000 a été choisi aléatoirement devant la difficulté de tracer un tel graphique sur la longueur totale des génomes. Les génomes sont couverts un nombre entier de fois compris entre 0 et 2. (Résultats semblables pour JIP 02/86 non montrés).

A contrario, nos premières simulations de données artificielles ne donnaient jamais de pics de couverture à 2X (la simulation faisait intervenir le simple découpage en segments de la référence). On a observé que ces pics de couvertures peuvent être expliqués par des recouvrements de contigs (figure 5).

Figure 5 : intervalles entre contigs pour la souche THC02/90 (100 premiers contigs). Les intervalles négatifs ont au maximum une taille de 39 nucléotides. Le paramètre k employé était de 41.



Comme expliqué en annexe 3 le chevauchement résulte des répétitions et on vérifie dans le cas présenté ici que la longueur maximale du chevauchement est k - 2.

Les contigs artificiels doivent mimer le comportement des contigs réels. Pour simplifier, une longueur de chevauchement fixe a été adoptée dans la simulation (k - 2). Les contigs ont été étendus d'une distance *k - 1* par dessus les régions répétées des deux brins (voir l'encadré

« Production des contigs artificiels »). Le chevauchement est de k-2. Ce sont les statistiques issues de cette simulation plus précise qui sont utilisées dans ce rapport.

Production des contigs artificiels

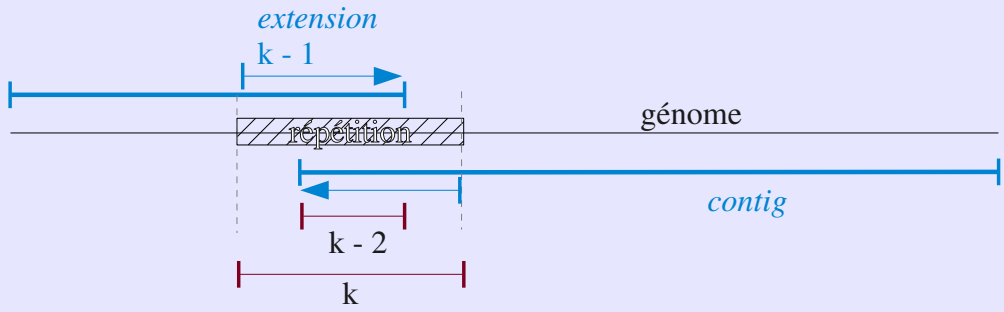


Figure 6 : le chevauchement des contigs artificiels est réglé à k-2

Exemple pour un k-mer de 4 et une répétition de taille 4 :

```
Alignement 1 :      AWH|GCC          (contig 1)
Génome :         ...AWH|GCCG|KL...
Alignement 2 :          CCG|KL       (contig 4)
```

Dans la réalité, le chevauchement peut être inférieur à k-2 (lorsque la taille de la répétition augmente, *cf.* annexe 3).

☞ Bilan du paragraphe

☑ Moins de 0,50 % des nouveaux assemblages couvrent plus d'une fois la séquence référence (valeur maximale constatée à k=31 pour JIP02/86 *de novo* : 0,41 %).

☑ Cette couverture excédentaire très localisée (pics de couverture) est dû aux répétitions et à la façon de produire les contigs à partir des nœuds du graphe.

☑ Cet événement disparaît lorsque la longueur du motif répété est suffisament grande.

4) Utilisation d'un alignement avec un génome de référence

☞ Rappel de la question initiale

☑ Dans quelle mesure utiliser un alignement comme guide d'assemblage permet-il d'améliorer la qualité de l'assemblage ? (*cf.* procédure du laboratoire).

Pour le test, nous sommes partis d'assemblages de mauvaise qualité (démarche de la couverture en k-mer, non montré), renforcés avec un alignement Bowtie et nous constatons une grande différence (cf. tableau 6).

Tableau 6 : à k = 59, assemblage *de novo* de JIP02/86 versus assemblage avec alignement préalable des lectures de JIP02/86 sur le génome de THC02/90.

Contigs produits par Velvet	JIP02/86 *de novo*	JIP sur THC
N 50	544	24 176
Nombre total de contigs (>2k)	5436	888
Taille cumulée des contigs	2503970	2 673 302
contigs non alignés	1	0
Chimères (contigs avec plus de 100 pb n	1	17
Gènes non couverts par les contigs	111	342
Gènes non-entiers	2 147	646
Couverture totale en nucléotides (%)	83,43	82,25

Il semble que les contigs soient trop agressivement reconstruits sur le modèle de la souche tierce. Le nombre de contigs produits avec une référence est relativement « bon » (tableau 6, deuxième colonne). Pourtant un k-mer de 59 paires de bases pose nativement des problèmes d'assemblage (tableau 6, première colonne) à cause des erreurs de séquençage[19]. L'utilisation du module *Columbus* aboutirait à donner trop d'importance à la référence sur laquelle sont alignées les lectures. Les spécificités de la souche assemblée pourraient être annihilées. La crédibilité de l'assemblage est douteuse avant même d'optimiser les paramètres d'assemblage.

☞ Bilan du paragraphe
☑ l'utilisation du module *Columbus* pourrait aboutir à donner trop d'importance à la référence sur laquelle sont alignées les lectures.
☑ les lectures seraient l'objet d'un nombre significatif d'erreurs de séquençage.
☑ dans les temps impartis, les souches 4, 8, 22, 23, 39 et 40 ont fait l'objet d'un assemblage *de novo* simple avec *VelvetOptimiser*.

5) Limites connues de la démarche méthodologique

Durant le stage, seul le paramètre k a été isolé pour en étudier l'influence. L'influence des autres paramètres est restée indéterminée (*i.e.* le seuil d'élimination des nœuds de faible couverture en k-mer). L'expérience semble montrer qu'après la production du graphe orienté, il existe de nombreux nœuds qui ont une très faible couverture et qui sont probablement des erreurs [1] [16]. Il est possible de nettoyer le graphe en dessous d'une couverture donnée. Le réglage par défaut de Velvet a été conservé dans ce travail mais il pourrait également faire l'objet de tests spécifiques (la moitié de la couverture en k-mer[20] [16]).

19 Pour utiliser de grands k-mers des filtres probabilistes de correction des erreurs de séquençage devraient être appliqués (P. Nicolas, discussion interne).
20 Le choix algorithmique fait ici semble résulter de l'expérience empirique des programmeurs, qui proposent comme seule alternative de régler ce seuil « selon les cas et à la convenance de l'utilisateur » [16].

6) Assemblages réalisés

Tableau 7 : caractéristiques des assemblages

	Souche	k	taille totale	N50	Contigs
24	JIP02-86	31	2 725 288	17 228	630
39	FPC_831	37	2 787 875	20 211	650
4	JIP16-00	31	2 803 614	11 806	847
12	THC02-90	41	2 691 591	16 439	563
23	NCIMB_1947T	35	2 624 740	24 870	492
40	FPC_40	35	2 695 228	16 138	561
8	JIP08-99	35	2 714 517	19 634	521
	Moyenne	35	2 728 782	18 270	614

B. Recherche de gènes dans les assemblages

Seules les meilleures prédictions ont été retenues. Le transfert à chacune de la meilleure annotation possible a été réalisé avec BLAST+ (mais aucun transfert n'a eû lieu s'il ne s'agissait pas d'une bonne similarité). Voir le tableau 8.

Tableau 8 : séquences codantes (CDS) prédites et similarités avec des annotations connues.

	CDS (Σ)	CDS 99	CDS *sim+*	CDS *sim*	CDS *ns*	CDS *trans*	CDS *gbk*
	1	2	3	4	5	6	7
JIP02-86 Sanger + SHOW + vérification manuelle							2420
JIP02-86 Sanger + SHOW	2624	2470	1146	1300	24	**2446**	2470
JIP02-86 Solexa + SHOW	3058	2811	1077	1705	29	**2782**	2811
THC02-90	2707	2505	1083	1396	26	**2479**	2505
NCIMB_1947T	2655	2445	1057	1343	45	**2400**	2445
FPC_831	2878	2642	1020	1362	259	**2382**	2641
FPC_840	2766	2525	1006	1312	207	**2318**	2525
JIP08-99	2766	2545	1099	1414	32	**2513**	2545

CDS (Σ) : nombre de CDS prédites par SHOW
CDS 99 : nombre de CDS ayant une probabilité >=0,99 (potentiels retenus)
CDS sim+ : CDS ayant une excellente similarité avec JIP0286/THC0290
CDS sim : CDS ayant une similarité avec JIP0286/THC0290
CDS ns : CDS sans similarité connue et ne faisant pas l'objet d'un transfert d'annotation
CDS trans : somme des CDS ayant fait l'objet d'un transfert de fonction (même inconnue) (sim+ et sim)
CDS gbk : somme des CDS reportées dans la fiche Genbank à l'issue de l'annotation
Seuil pour la recherche de similarité (avec une BLOSUM80) : 0,01 (probabilité que le résultat soit dû au hasard). Les alignements avec une e-value supérieure ne déclenchent pas de transfert d'annotation.

Le premier témoin est fourni par la séquence génomique complète de JIP02/86 (séquençage Sanger) publiée. Celle-ci fait ici l'objet d'une nouvelle recherche de potentiels avec SHOW. 2470 CDS sont prédites par SHOW avec une probabilité supérieure à 0,99 quand le génome publié regroupe 2420 CDS seulement. Il est possible que les CDS « en trop » soient des CDS courtes qui seraient éliminées par une vérification manuelle ou de CDS vues plusieurs fois (car réparties sur plusieurs extrémités de contigs). 2446 CDS ont fait l'objet d'un transfert

d'annotation par similarité. Ainsi, le témoin permet de vérifier que la détection des potentiels et l'attribution d'une annotation (qui peut être une fonction connue ou inconnue) se passent sans perte.

Le second témoin est issu du reséquençage de cette même souche en technologie Solexa. On s'attendrait à ce que moins de CDS soient prédites par SHOW. Toutefois, il y a légèrement plus de CDS que dans le témoin Sanger (2811 avec une bonne probabilité). Une fois encore, il peut s'agir de CDS courtes ou de CDS vues plusieurs fois. Du moins ce témoin Solexa permet-il de conclure que la prédiction de CDS est sans perte. 24 CDS ne peuvent faire l'objet d'une similarité avec des CDS déjà connues pour cette souche, ce qui représente seulement 0,97 % des CDS retenues. Là encore, une inspection au cas par cas permettrait d'en déterminer l'origine.

Les différents types d'annotations disponibles dans l'annotation d'origine de JIP02-86 et reportées pour chacune des souches sont présentées dans le tableau 9.

Tableau 9: types d'annotation de CDS. Le type n°6 est reporté tel quel dans chaque souche, et signifie qu'à la base chez JIP02-86 il n'y a pas de similarité connue pour une CDS, bien qu'il y ait une similarité entre celle-ci et la CDS prédite qui permette le transfert de cette annotation.

1 Cell envelope and cellular processes	3.2 DNA restriction/modification and repair
1.1 Cell wall	3.3 DNA recombination
1.10 Transformation/competence	3.4 DNA packaging and segregation
1.2 Transport/binding proteins and lipoproteins	3.5 RNA synthesis
1.3 Sensors (signal transduction)	3.6 RNA modification
1.4 Membrane bioenergetics (electron transport chain and ATP synthase)	3.7 Protein synthesis
1.5 Mobility and chemotaxis	3.8 Protein modification
1.6 Protein secretion	3.9 Protein folding
1.7 Cell division	4 Other functions
2.1 Metabolism of carbohydrates and related molecules	4.1 Adaptation to atypical conditions
2.2 Metabolism of amino acids and related molecules	4.2 Detoxification
2.3 Metabolism of nucleotides and nucleic acids	4.4 Phage-related functions
2.4 Metabolism of lipids	4.5 Transposon and IS
2.5 Metabolism of coenzymes and prosthetic groups	4.6 Miscellaneous
2.6 Metabolism of phosphate	5.1 Protein of unknown function similar to other proteins from the same organi
2.7 Metabolism of sulfur	5.2 Protein of unknown function similar to proteins from other organisms
3.1 DNA replication	6 No similarity
3.10 Protein degradation	

La figure 7 montre la répartition des annotations pour la souche JIP02-86 Solexa. En l'occurrence, cela représente un total de 2765 CDS[21]. On constate également que les CDS prédites pour la souche nouvelle JIP08-99 adoptent la même répartition des différentes fonctions. C'est le cas pour toutes les souches (voir le tableau 10).

21 Au lieu de 2811 : le code « fonctionnel » était absent de la description de certaines CDS de l'annotation d'origine de JIP02-86. Dans ce cas, la protéine produite a été reportée, mais les statistiques sur les types fonctionnels ne peuvent la prendre en compte.

Figure 7 : répartition des fonctions de FP JIP02-86 (reséquençage) et de FP JIP08-99

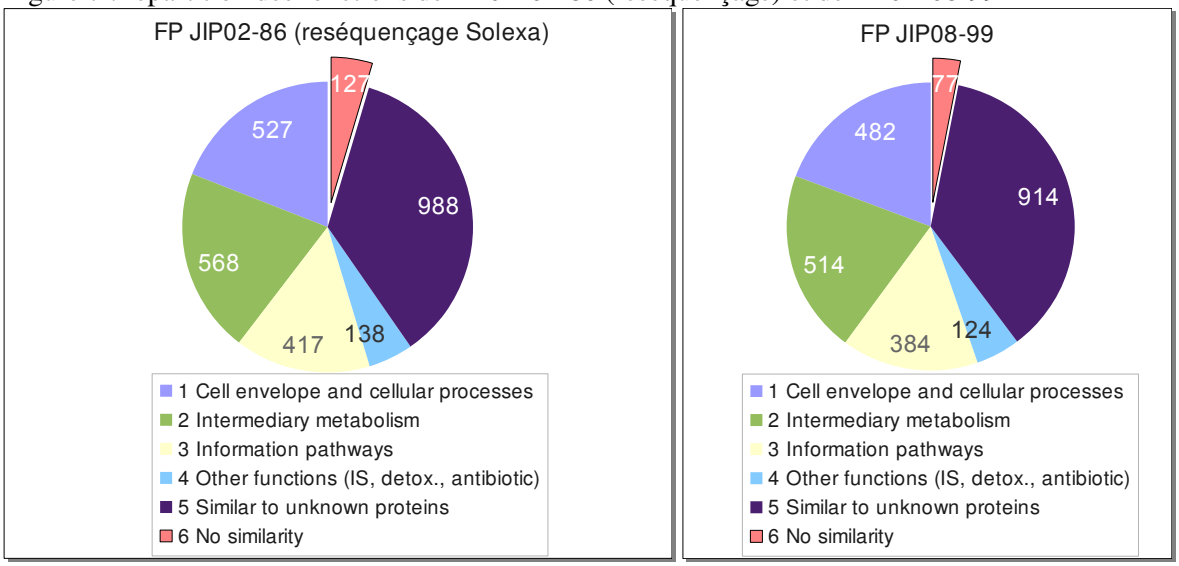


Tableau 10 : annotations des nouvelles souches par types majeurs

Type fonctionnel	*Nombre de CDS pour chaque souche*								
	JIP02-86 Solexa	JIP02-86 Sanger	THC02-90	NCIMB_ 1947T	FPC_831	FPC_840	JIP08-99	Moyenne	Ecart-type
1	527	439	486	470	473	430	482	472,4	32,1
2	568	485	509	502	502	504	514	512	26,3
3	417	381	389	388	386	382	384	389,6	12,4
4	138	124	100	111	109	107	124	116,1	13,1
5	988	884	907	864	868	860	914	897,9	45
6	127	115	44	38	23	17	77	63	44,2

Deux tiers des fonctions prédites dans les souches sont de fonction inconnue, comme dans la souche Sanger d'origine. La suite de ce travail consisterait à lever l'incertitude sur les CDS qui ont hérité de l'annotation « fonction inconnue » depuis JIP02-86. Pour ce faire, ce sous-groupe de CDS pourrait faire l'objet d'une recherche de similarités sur une base de données de génomes procaryotes plus étendue que la seule base de données *F. psychrophilum*. (Bien que cela ait déjà était fait à la base, c'est-à-dire pour JIP02-86).

On a également tenté de regarder fonction par fonction, si un enrichissement pour une souche donnée pouvait être trouvé. En effet, le fait que la répartition fonctionnelle globale montrée dans les figures précédentes soit équivalente pour toutes les bactéries ne prouve pas qu'au sein de chaque grand type fonctionnel (1, 2, etc.) la répartition des fonctions s'équilibre de la même façon pour tous les génomes. Cependant la représentation du nombre de CDS de chaque fonction (figure en annexe 1) n'a pas présenté pas de différence significative entre souches. Sur la bases des seules CDS que nous avons annotées, il n'y a pas d'enrichissement fonctionnel significatif.

Discussion et conclusion

Les lectures du séquençage de *Flavobacterium psychrophilum* posent des défis algorithmiques de part leur longueur restreinte. Le programme Velvet basé sur les graphes de Bruijn permet d'obtenir *de novo* des génomes convenablement assemblés Les génomes sont représentés par un nombre limité de contigs (de l'ordre de 700). Les points de cassures d'assemblage que sont les répétitions et les biais de séquençage semblent empêcher d'optimiser encore plus l'assemblage. Augmenter significativement la qualité de l'assemblage ne peut passer que par l'ajout de nouvelles données telles que des lectures plus longues issues de technologies plus récentes (ou plus anciennes). Cependant, plus que de parvenir à une séquence finie complète, obtenir un nombre fini de contigs de taille respectable[22] semble un objectif raisonnable. L'identification des grandes fonctions encodées par le génome est possible. Avec les séquences reconstruites pendant ce travail, il a été possible de commencer à détecter les potentiels codants et à réaliser leur annotation. Notre étude qualité préalable nous a en effet montré qu'environ seulement 6 % de séquences codantes étaient manquantes par rapport à une génome fini de *F. psychrophilum*.

L'objectif principal du travail mené est de mettre en relation des divergences génomiques avec les caractéristiques de pathogénicité et de spécificité d'hôte des souches de la bactérie. Dans l'état actuel, la première approche déployée dans ce travail n'a pas montré d'enrichissement particulier d'une des catégories fonctionnelles pour une des souches traitées dans la phase d'annotation. Plusieurs axes de travail permettront d'améliorer le résultat actuel. Le premier est de définir différemment la sensibilité du filtrage des CDS (quelle probabilité de prédiction selon le modèle de Markov est acceptée ? Quelle similarité avec une séquence connue est-elle acceptée pour transférer une annotation ?) Cependant, on a vu que les réglages adoptés ici (favorisant la spécificité plus que la sensibilité) permettent déjà de trouver et d'annoter la majorité des CDS. Le deuxième axe est de poursuivre la recherche de similarités pour attribuer une fonction aux CDS qui en manquent. De plus, une limite de ce travail a été de réaliser en première approche un BLAST unidirectionnel. Un BLAST bidirectionnel dans lequel on ne considère que les meilleurs alignements réciproques serait préférable.

22 Une des actualités scientifiques majeures en génomique est de définir ce qu'est un contig de taille respectable.

Le troisième axe consiste en une approche plus résolutive. Au lieu de commencer par rechercher les fonctions des séquences codantes prédites, il peut être tenté des les regrouper (*Single Linkage Clustering*) pour ensuite trouver les fonctions des groupes obtenus.

Le quatrième axe de travail est de rechercher les spécificités du génome au delà des séquences codantes (recherche ciblée de locus à l'aide de profils poids-position ou de modèles de Markov dédiés). Malheureusement, les assemblages tirés des lectures courtes induisent une perte des signaux génomiques transitoires[23] (aux rangs desquels, les signaux répétés des îlots de pathogénicité où les régions CRISPR – un locus d'incorporation d'éléments génétiques exogènes autorisant des réactions de défense chez les bactéries).

Les expérimentations menées à l'INRA sur la bactérie montrent qu'un certain nombre de souches de *F. psychrophilum* présentent une grande inversion génomique bordée par des segments courts répétés inversés. Les travaux n'ont pas encore montré si cette inversion est spécifique des souches ou bien des individus. Il doit être possible de caractériser le sens du segment qui fait l'objet de l'inversion pour les nouvelles souches assemblées. En effet, il existe des cibles spécifiques qui permettent déjà l'emploi de sondes radioactives de part et d'autres des locus-charnières de l'inversion pour la détermination expérimentale. L'analyse bioinformatique nécessiterait qu'un contig couvre une région charnière. Les assemblages actuels ne semblent pas offrir la lecture intègre de ces régions (Duchaud E., communication interne), ce qui pourrait constituer un cinquième axe de travail.

Pour ces deux derniers points, tant que les procédures et les techniques ne permettent pas de restituer le génome *entier* – comme avec les technologies classiques et la finition manuelle – l'utilisation de paramètres plus agressifs d'assemblage pourrait être envisagée pour rechercher des signaux transitoires qui n'arriveraient pas à émerger dans une version standard de l'assemblage[24].

23 C'est vrai avec la technologie Solexa mais aussi avec les premières générations de séquençage 454©.

24 Une version standard d'un génome pourrait être un compromis entre la longueur des contigs et l'évitement des erreurs de séquençage. Un assemblage plus agressif pourrait conduire à incorporer plus d'erreurs mais à assembler plus. Les deux versions ne répondraient pas au même objectif (le premier à la recherche des fonctions biologique, le second à la recherche de signaux plus discrets). Enfin, lorsque les nouvelles technologies de séquençage auraient évolué ce concept perdrait son sens.

Références

[1] **Daniel Zerbino, Ewan Birney**, Velvet : algorithms for de novo short read assembly using de Bruijn graphs, Genome Research, 18:821-829 (2008)

[2] **Eric Duchaud, Mekki Boussaha, Valentin Loux, Jean-François Bernardet, Christiant Michel, Brigitte Kerouault, Stanislas Mondot, Pierre Nicolas, Robert Bossy, Christophe Caron, Philippe Bessières, Jean-François Gibrat, Stéphane Claverol, Fabien Dumetz, Michel Le Hénaff, Abdenour Benmansour**, Complete genome sequence of the fish pathogen Flavobacterium psychrophilum, Nature biotechnology, 25*7*:763-769 (2007)

[3] **A. Nematollahi, A. Decostere, F. Pasmans, F. Haesebrouck**, Flavobacterium psychrophilum infections in salmonid fish, Journal of Fish Diseases, 26:563-574 (2003)

[4] **Pierre Nicolas, Stanislas Mondot, Guillaume Achaz, Catherine Bouchenot, Jean-François Bernardet, Eric Duchaud**, Population structure of the fish-pathogenic bacterium Flavobacterium psychrophilum, Applied and Environnental Microbiology, 74*12*:3702-3709 (2008)

[5] **C. Chakroun, F. Grimont, M.C. Urdaci, J.-F. Bernardet**, Fingerprinting of Flavobacterium psychrophilum isolates by ribotyping and plasmid profiling., Dis. Aquat. Organ., 29:213-218 (2008)

[6] **Motaki Kondo, Kenji Kawai, Kenrou Kurohara, Syun-ichirou Oshima**, Adherence of Flavobacterium psychrophilum on the body surface of the ayu Plecoglossus altivelis, Microbes and Infection, 4:279-283 (2002)

[7] **V.E. Ostland, P.J. Byrne, G. Hoover, H.W. Ferguson**, Necrotic myositis of rainbow trout, Oncorhynchus mykiss (Walbaum) : proteolytic characteristics of a crude extracellular preparation from Flavobacterium psychrophilum, Journal of Fish Diseases, 23:329-336 (2000)

[8] **Keiko Satoa, Mariko Naitoa, Hideharu Yukitakea, Hideki Hirakawab, Mikio Shojia, Mark J. McBridec, Ryan G. Rhodesc et Koji Nakayama**, A protein secretion system linked to bacteroidete gliding motility and pathogenesis, Proc Natl Acad Sci USA., 107*1*:276-281 (2010)

[9] **Can Alkan, Saba Sajjadian, Evan E. Eichler,**, Limitations of next-generation genome sequence assembly, Nature Methods, 8*1*:61-65 (2011)

[10] **Fabien Melchiore, Cyprien Guérin, Pierre Nicolas, Valentin Loux**, Assemblage de génomes bactériens séquencés par NGS : comparaison d'outils et choix de paramètres (Poster, 2010)

[11] **Ben Langmead, Cole Trapnell, Mihai Pop, Steven L. Salzberg**, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology, 3:R25.1-10 (2009)

[12] **Daniel Zerbino**, Using the Colombus extension to Velvet (2010)

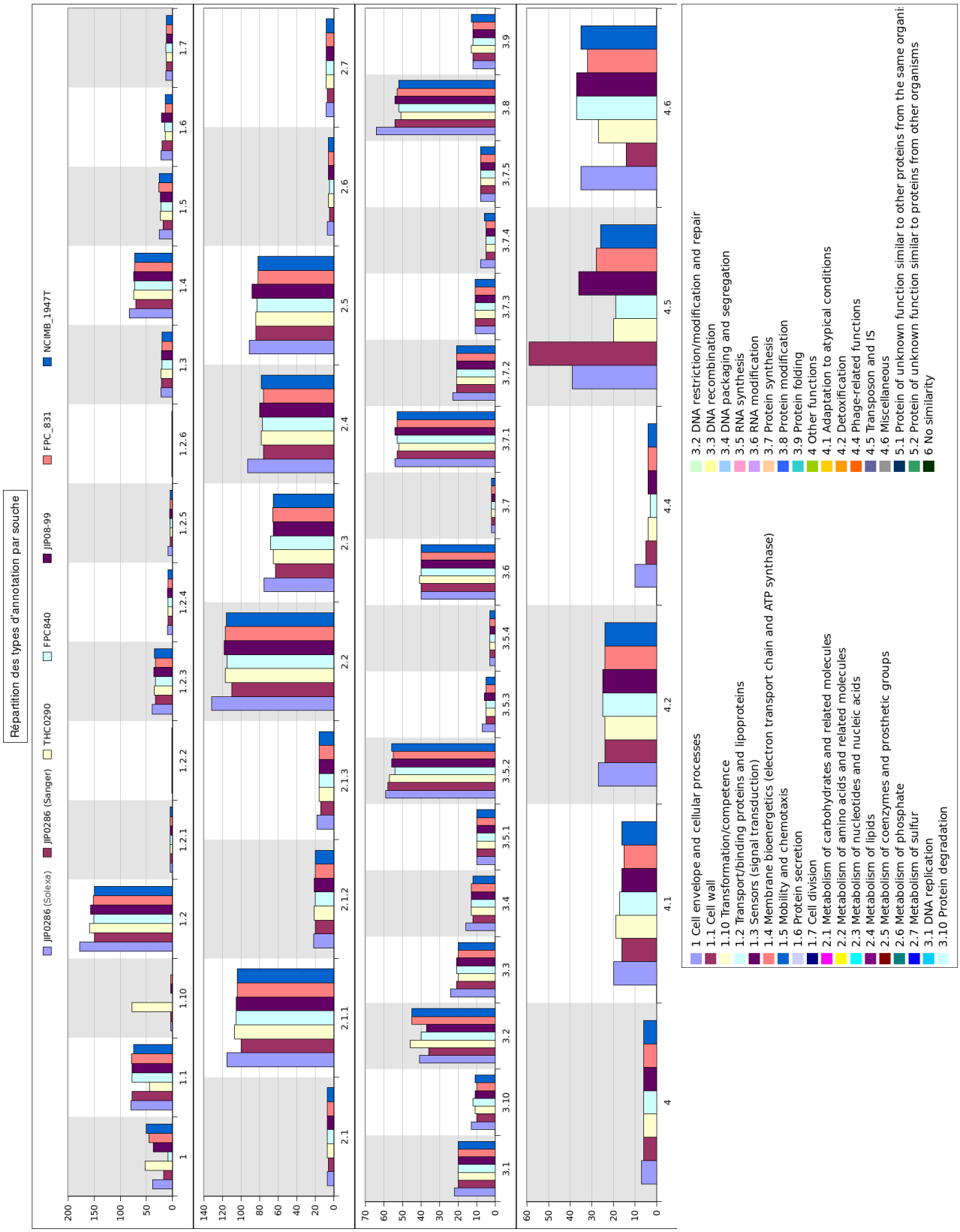
[13] **K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. ven de Guchte, S. Penaud, E. Maguin, M. Hoebeke, P. Bessières, J.-F. Gibrat**, AGMIAL : implementing an annotation strategy for prokaryote genomes as a distributed system, Nucleic Acids Research, 34*12*:3533-3545 (2006)

[14] **Ewan Birney**, Assemblies : the good, the bad, the ugly, Nature Methods, 8*1*:59-60 (2011)

[15] **Wenhyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jim Shang, Bairong Shen**, A practical comparaison of de novo genome assembly software tools for next-generation sequencing technologies, PLoS One, 6*3*:e17915 (2011)

[16] **Daniel Zerbino**, Manuel de Velvet, version 1.1 (2011)

Annexe 1 : comparaison fonctionnelle des souches



Annexe 2 : traitement des lectures par Velvet

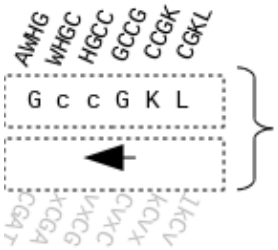
1. Découpage des lectures en mots de taille k

AWHG
WHGC
HGCC
GCCG
CCGK
CGKL

2. Constitution du « nœud »

Lecture 1

3. Constitution du « nœud » inverse-complémentaire
(pour que les recouvrements entre lectures de brins opposés soient prises en compte p.822)



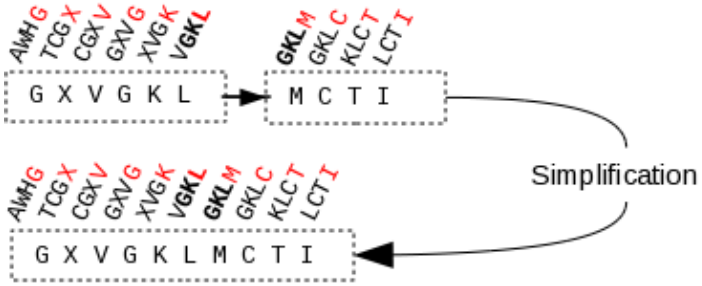
Lecture 1

Un **bloc** (ensemble des deux nœuds) est interrompu au début ou à la fin de chaque lecture – p. 822

Pour simplifier, dans notre exemple toutes les lectures sont issues du même brin.

4. Traitement de la lecture 2 et simplification

Les nœuds peuvent être connectés par un arc direct si le dernier k-mer du nœud d'origine se recoupe avec le premier k-mer du nœud de destination.
S'il n'y a pas de conflit de liaisons, les nœuds peuvent être fusionnés.



Sans autre liaison, le Nœud 1 donnera un contig en utilisant la dernière lettre de chaque k-mer.
(A l'exception du premier pris intégralement pour initialiser).

Annexe 3 : Répétitions et nœuds

Soit 2 lectures issues du séquençage :

`Lecture 1 : AWHGCCGKL` `Lecture 2 : GKLTCI`

Tous les caractères sont des nucléotides, même s'il sont représentés par commodité avec un alphabet de 26 lettres.

L'algorithme crée deux blocs : le premier court du premier au dernier k-mer issus du découpage de la lecture 1 en mots de taille *k,* le second est issu de la lecture 2. Le dernier k-mer du nœud 1 s'aligne en partie avec le premier k-mer du nœud de destination ce qui permet leur liaison.

```
N0a (AWHG) -- (WHGC) -- (HGCC)--(GCCG) -- (CCGK) -- (GGKL)
                                                    (GKLT) -- (KLTC) – (LTCI)  N0b
```

Si aucune autre liaison ne produisant un conflit logique ne vient s'ajouter au cours du reste de l'analyse des données (c'est le cas ici), il peut y avoir fusion en un unique nœud :

```
N1 (AWHG) -- (WHGC) -- (HGCC)--(GCCG) -- (CCGK) -- (GGKL) -- (GKLT) -- (KLTC) -- (LTCI)
```

La lecture 1 et la lecture 2 sont représentées dans le nœud. Il en résultera *in fine* un contig. Typiquement, si le nœud N1 n'a pas d'autres liaisons "à droite" ou "à gauche", le contig produit sera, en prenant la dernière lettre de chaque k-mer (*cf.* publication de 2008*)* : GCCGKLTCI.

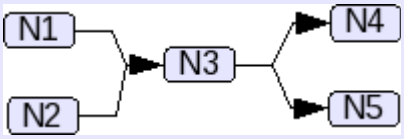
Afin de comprendre pourquoi une répétition une fragmentation du génome reconstruit, incorporons maintenant un motif répété dans plusieurs lectures.

`Lecture 1 : AWHGCCGKL`
`Lecture 2 : GKLMGCCGACG`

La construction des nœuds donne le résultat suivant :

```
      N1 (AWHG) -- (WHGC) -- (HGCC)-\       N3       /- (CCGK) -- (GGKL) N4
                                     }- (GCCG) -{
N2 (GKLM) -- (KLMG) -- (LMGC) -- (MGCC)-/            \- (CCGA) -- (CGAC) -- (GACG) N5
```

Ou, schématiquement :



Toute répétition de taille supérieure ou égale à k (R>=*k*) pose problème. Augmenter *k* masque les petites répétitions.

La lecture 1 est représentée par les nœuds (1, 3, 4) tandis que la lecture 2 est représentée par les nœuds (2, 3, 5).

L'algorithme produit de nombreux petits contigs séparés (un contig issu de chacun des nœuds 1, 2, 4 et 5). La taille du contig étant au minimum de 2*k* le nœud N3 est ensuite oublié (c'est une cassure dans le génome) – ainsi, potentiellement dans notre exemple, que les autres nœuds s'il n'atteignent pas la longueur minimale.

Lorsque la longueur de la répétition est égale à k, on observe que les k-2 dernières lettres du contig de gauche sont identiques aux k-2 premières lettres du contig de droite. Quand la longueur répétée augmente d'une unité ce chevauchement diminue d'une unité. La longueur du chevauchement est donc $k-2-(L-k)=2k-L-2$ où L est la longueur de la répétition.

Annexe 4 : liste des pièces produites

A l'issue du stage, les pièces finales disponibles sont les suivantes :

- **scripts écrits dans le cadre du stage :**

  Volga 1.03 (simulation d'un jeu de contig et extraction de statistiques d'assemblage)

  Scripts d'interfaçage des étapes de la procédure d'assemblage

  Scripts de filtrage des CDS et de transfert d'annotation (avec BioPerl)

- **données :**

  Pour reproduire les assemblages : fichiers de lectures nettoyés, fichiers d'alignement avec les génomes de référence et fichiers Genbank déjà disponibles (THC02/90, JIP02/86)

  Résultats : 56 Go d'assemblages finaux de *F. Psychrophilum* (les assemblages intermédiaires ont été supprimés par manque de place), fichiers GFF des CDS prédites, fiches d'annotation Genbank pour les nouvelles souches

- **documentation ;**

  Cahier de laboratoire (permet de refaire les assemblages) [~140 pages]

  Documentation explicite sur le fonctionnement et les points critiques de Velvet à l'intention des nouveaux utilisateurs au laboratoire [3 pages]

  Documentation technique sur la simulation Volga (modalités et algorithmes) [9 pages]

  Rapport sur les résultats méthodologiques appliqués à l'assemblage de *F. psychrophilum* (pour deux série d'expériences : procédure empirique laboratoire et procédure de la couverture en k-mer) [10 pages]

  Rapport préliminaire sur la recherche de fonctions dans les différentes souches [3 pages]

  Rapport complet s'appuyant sur l'ensemble des résultats du stage [44 pages]

Ces documents ont été repris dans le présent rapport.

High throughput sequencing of Flavobacterium psychrophilum analysis

Keywords
Early 454 sequencing - Bacterial pathogen assembly - annotation

Summary
Flavobacterium psychrophilum is currently one of the main bacterial pathogens hampering the activity of salmonid farming worldwide. Its genetic, metabolic and antigenic characterization is a manner to find consistent ways of fighting against this bacterium. The first generations of high throughput sequencing challenges the reconstruction of the genomic repetitions and replications due to the shortness of their products. This work shows that the global volume of biological information isn't lost in basic assemblies with respect with the known Sanger-based annotations, even if pinpoint information (some locus or segmental inversion) may be lost. However, bringing a reference alignment to the assembly tool needs further biological evaluation. Our work shows it succeeds in producing an assembly, but his biological fidelity hasn't be enforced. In the end, some coding regions found on the newly-assembled strains were annotated automatically by similarity.

Analyse des données de séquençage à haut débit de plusieurs souches de la bactérie Flavobacterium psychrophilum

Mots-clefs
Procaryote pathogène – technologies de séquençage – méthodologie d'assemblage - annotation

Résumé
La bactérie *Flavobacterium psychrophilum* est une bactérie pathogène des salmonidés tels que la truite et les saumons d'élevage. Sa caractérisation génétique, métabolique et antigénique fait l'objet de recherches visant à trouver des moyens de lutte efficaces. La reconstruction des génomes de souches séquencées avec les premières générations de séquenceurs à haut débit pose des défis algorithmiques liés aux répétitions et duplications du génome mais aussi à la faible longueur des séquences disponibles. Ce travail montre que la quantité d'information biologique perdue dans ces données reste limitée par rapport à un génome de référence (basé sur du séquençage Sanger). En revanche, certaines options de la procédure de travail (par exemple aider l'assembleur à l'aide d'un génome connu) paraissent plus hasardeuses et nécessiteraient une investigation expérimentale plus poussée de la fidélité biologique de leurs résultats. Enfin, sur les nouvelles souches assemblées, des gènes ont pu être identifiés et faire l'objet d'un report d'annotation par similarité.