Étude de la plasticité génomique chez Streptomyces ambofaciens : assemblage et analyse comparative du génome des souches ATCC23877 et DSM40697

Maxime Bruto

UNIVERSITE HENRI POINCARE NANCY 1

MASTER MICROBIOLOGIE

Mention Biotechnologies Microbiennes

Etude de la plasticité génomique chez *Streptomyces ambofaciens* : Assemblage et analyse comparative du génome des souches ATCC23877 et DSM40697

Par
BRUTO Maxime

Travaux dirigés par :
M^{r} Pierre LEBLOND (Professeur) et M^{elle} Annabelle Thibessard (Maitre de conférence)

Stage effectué du 04/01/2010 au 04/07/2010

Structure d'accueil : Laboratoire de Génétique et Microbiologie (UMR 1128)

SOMMAIRE

RESULTATS

DISCUSSION

BIBLIOGRAPHIE

ABREVIATIONS

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

ARNt : Acide ribonucléique de transfert

ARNr : Acide ribonucléique ribosomal

ATCC : American Type Culture Collection (Collection Américaine de Cultures)

ATP : Adénosine tri-phosphate

BLAST : Basic local alignment search tool (outil de recherche d'alignement local)

CDS : Coding DNA sequence (séquence codante d'ADN)

dNTP : Désoxyribonucléotide tri-phosphate

ddNTP : Didésoxyribonucléotide tri-phosphate

DSM : Deutsche Sammlung von Mikroorganismen (Collection Allemande de Microorganismes)

kb : kilobase

Mb : Mégabase

pb : paire de base

ORF : Open reading frame (phase ouverte de lecture)

PCR : Polymerase chain reaction (réaction de polymérisation en chaine)

INTRODUCTION

Préambule

Depuis les années 70, et l'avènement du séquençage enzymatique par la méthode de Sanger, de nombreux verrous techniques ont été levés menant au séquençage haut-débit. Aujourd'hui, le pyroséquençage et la « terminaison réversible », deux nouvelles approches de séquençage, génèrent plusieurs millions de lectures[1] par réaction, là où les techniques manuelles ne produisaient qu'une seule séquence. L'assemblage est l'étape cruciale du processus de traitement de ces données de séquençage. Cette étape, nécessitant l'utilisation d'algorithme spécifiques, permet de reconstituer tout, ou une partie, de la séquence d'un génome d'intérêt à partir des lectures.

Depuis le séquençage du génome humain par l'équipe de Craig Venter (Venter *et al.* 2001), de nombreuses séquences génomiques d'organismes bactériens ont été obtenus. La mise au point de nouvelles techniques de séquençage, toujours plus performantes en terme de débit de séquences générées, à accéléré l'accumulation des données relatives à l'organisation et au contenu génétique des organismes procaryotes. Par l'analyse comparative de ces génomes, les mécanismes permettant l'évolution des bactéries ont pu être approchés, permettant de mieux comprendre leurs succès évolutifs incomparables dans le monde vivant.

1°) Le séquençage des génomes

1.1°) Les prémices du séquençage à haut débit

Au cours des 30 dernières années, la méthode la plus utilisée pour séquencer l'ADN a été la technique de terminaison de chaine mise au point par Frederick Sanger en 1975 (Sanger et Coulson, 1975). Le principe de cette méthode est de générer toutes les molécules d'ADN simple brin complémentaires d'une matrice, commençant à la même extrémité 5'. Pour cela, un court segment d'ADN synthétique sert d'amorce pour une ADN polymérase. La terminaison de l'élongation est provoquée par l'intégration dans le fragment en cours de synthèse d'un nucléotide dépourvu du groupement hydroxyle en position 3' (ddNTP), nécessaire à la poursuite de la polymérisation de la chaine. Cette phase d'élongation se déroulant dans un milieu réactionnel où dNTP et ddNTP sont présents, la terminaison se déroule aléatoirement dans la population de fragments en synthèse (**figure 1**). C'est par cette méthode que le premier génome, celui du virus ϕX174 , fut séquencé en 1977.

Toutefois c'est l'automatisation de cette méthode à la fin des années 80, avec le développement des marquages fluorescents et de l'électrophorèse capillaire, qui a ouvert la voie du séquençage à haut débit. Pour le séquençage automatique de Sanger, les appareils utilisés permettent de séquencer en parallèle de multiples échantillons (jusqu'à 384), ce qui augmente le débit de la technique.

Malgré les progrès technique accomplis par l'automatisation du procédé de séquençage, la préparation des échantillons à séquencer passe par leur clonage dans un vecteur puis leur amplification chez *Escherichia coli*. Cette étape ne pouvant être totalement automatisée, de nouvelles pistes ont été recherchées, permettant de préparer et séquencer des millions de fragments en parallèle.

1.2°) Les nouvelles technologies de séquençage

La méthode de Sanger automatisée est considérée comme étant la stratégie de séquençage haut-débit de première génération. Depuis cinq ans, de nouvelles méthodes ont été mises au point parmi lesquelles le pyroséquençage haut-débit, les techniques de terminaison cyclique réversible

[1] : une lecture est la séquence obtenue après séquençage d'un fragment d'ADN.

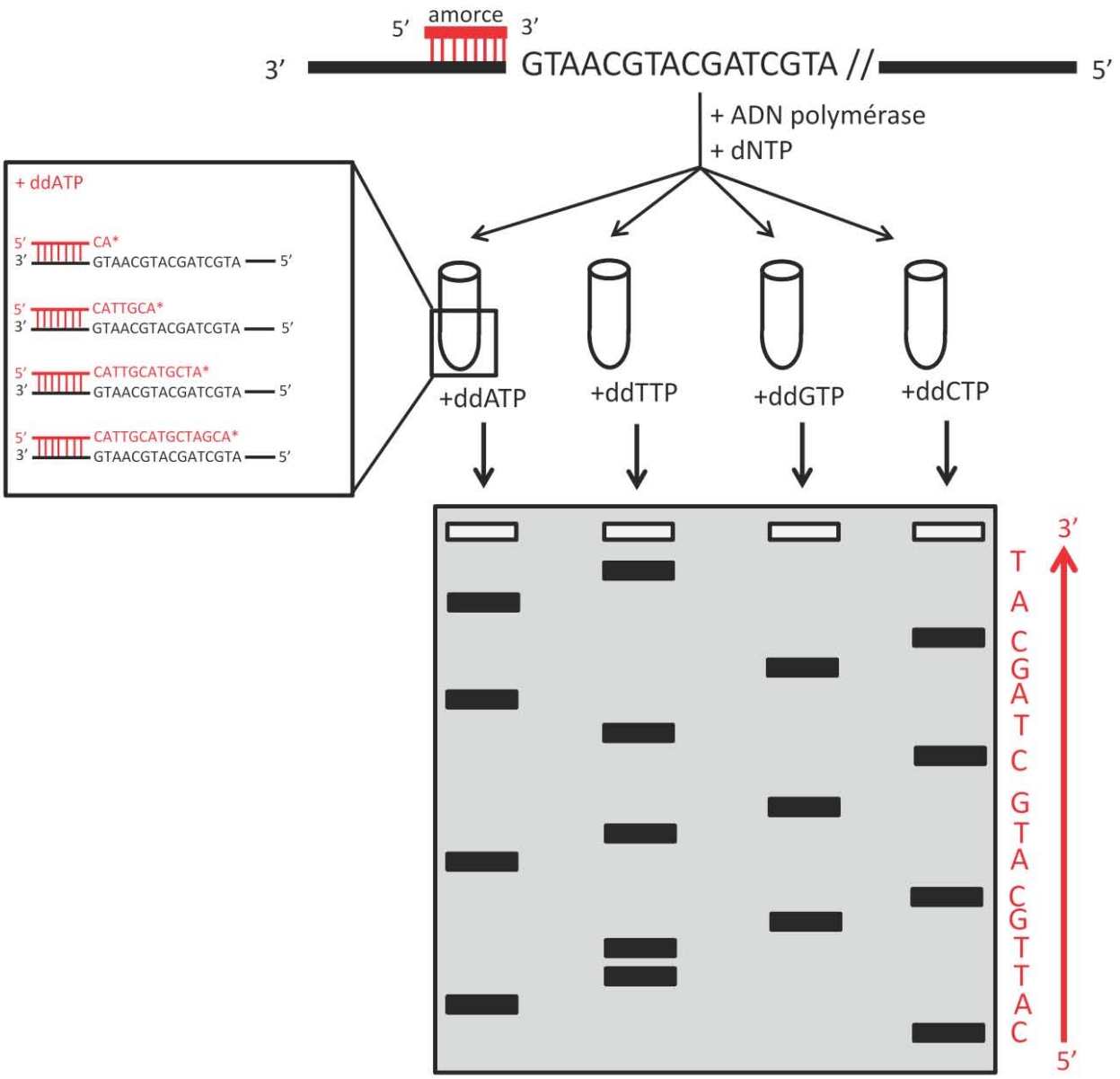


Figure 1 : *Principe du séquençage selon la méthode de Sanger.*
Chacun des quatre tubes contient la matrice ADN, des amorces, de l'ADN polymérase, des dNTP et des ddNTP. A la fin de la réaction de polymérisation, chaque tube contient des fragments de tailles variables, correspondant à l'arrêt de la synthèse pour chacun des sites d'incorporation du ddNTP. A l'origine, la séquence était déduite de la séparation sur gel de polyacrylamide des fragments d'ADN de chaque réaction de polymérisation.

(par Solexa/Illumina et Helicos) ou encore le séquençage par ligation (SOliD). Elles ont pour point commun de fournir un grand nombre de données à une vitesse telle que des projets de séquençage demandant plusieurs années avec la technique de Sanger peuvent théoriquement être achevés en quelques semaines. De plus le coût requis pour séquencer un génome est nettement plus réduit, comme en atteste le **tableau 1**, récapitulant la somme nécessaire au séquençage d'un génome d'être humain par les différentes méthodes de séquençage.

Malgré tous ces avantages, les nouvelles technologies de séquençage présentent certains biais qui rendent leur utilisation complexe. Nous allons voir avec les deux méthodes de séquençage de nouvelle génération les plus utilisées, le pyroséquençage haut-débit et la méthode de terminaison cyclique réversible Illumina/Solexa, quels peuvent être les avantages et les limites de ces méthodes.

1.2.1°) Le pyroséquençage haut-débit

Le pyroséquençage est une technique de séquençage basée sur la détection du pyrophosphate relâché lors de la réaction de polymérisation de l'ADN. Les nucléotides sont ajoutés les uns après les autres dans un ordre défini. Lorsqu'un nouveau nucléotide est ajouté, l'ATP sulfurylase va utiliser le pyrophosphate relâché lors de la polymérisation pour générer de l'ATP. Cet ATP sera utilisé par la luciférase pour oxyder la luciférine en oxyluciférine et émettre de la lumière (**figure 2**). C'est ce signal lumineux qui est détecté par une caméra puis traduit en chromatogramme (Rotberg et Leamon, 2008).

A l'aide de cette technique, il est possible de séquencer 500 millions de bases en 10 heures environ, ce qui en fait la technique de séquençage haut-débit commercialisée la plus rapide. Cette vitesse de séquençage peut être atteinte grâce à des améliorations constantes de la processivité des enzymes utilisées lors de la réaction (Leamon et Rothberg, 2007). La grande quantité de séquences générées par cette technique vient en partie de la préparation des échantillons. Chaque fragment est isolé puis amplifié par PCR sur une microbille qui sera ensuite déposée dans le puits d'une plaque de microtitration (il est possible de charger 1 à 2 millions de billes sur une même plaque)(**figure 3**). La rapidité de chaque cycle de séquençage permet d'obtenir des lectures d'une taille moyenne de 400 pb (Rothberg et Leamon, 2008).

Si un même nucléotide est incorporé dans le même cycle, un signal lumineux proportionnel au nombre de nucléotides est décelé. Au-delà de 4 à 5 nucléotides, le caractère proportionnel du signal est perdu. Ces séquences homopolymériques représentent la source majeure d'erreurs de séquençage par cette méthode, comme l'ont estimé Huse *et al.* (2007). Ils attribuent 39% des erreurs de type insertions/délétions à ce contexte nucléotidique spécifique.

1.2.2°) Le séquençage par terminaison cyclique réversible d'Illumina/Solexa

Dans le principe de séquençage d'ADN commercialisé par Illumina/Solexa, une ADN polymérase ajoute un nucléotide modifié porteur d'une molécule fluorescente et d'un terminateur labile (**figure 4**). Chaque nucléotide est porteur d'un fluorochrome spécifique. Suite à l'incorporation, le milieu réactionnel est nettoyé pour supprimer les nucléotides non-incorporés qui sont en excès. Les fluorochromes sont ensuite excités, les émissions lumineuses sont détectées par le séquenceur puis traduites en une séquence. A la fin de chacun des cycles, le fluorochrome et le terminateur en 3' sont clivés pour que la réaction de séquençage puisse continuer.

Par cette méthode, il est possible de séquencer 18 à 35 milliards de bases en 4 à 9 jours. La grande quantité de données générées par cette méthode est due à la préparation des fragments. Chaque fragment est isolé sur une plaque de verre puis amplifié par une réaction de PCR (**figure 5**). Par cette méthode, il est possible de produire 100 à 200 millions de matrices pouvant être séquencées sur une même lame (Metzker, 2010).

Bien que cette technique de séquençage soit à l'heure actuelle l'une des plus utilisées, elle présente plusieurs biais notables. Le principal problème de cette technique est la taille relativement

Provenance de la séquence génomique	Méthode utilisée	Coût estimé (US$)
J. Craig Venter	Sanger automatisée	70 000 000
James D. Watson	Pyroséquençage haut-débit	1 000 000 *
Homme chinois	Terminaison cyclique réversible (Solexa/Illumina)	500 000 *
Homme coréen	Terminaison cyclique réversible (Solexa/Illumina)	200 000 *

Tableau 1 : *Evolution du coût de séquençage d'un génome humain selon les méthodes de séquençage utilisées.*

* : Coût des réactifs uniquement. (source : Metzker, 2010)

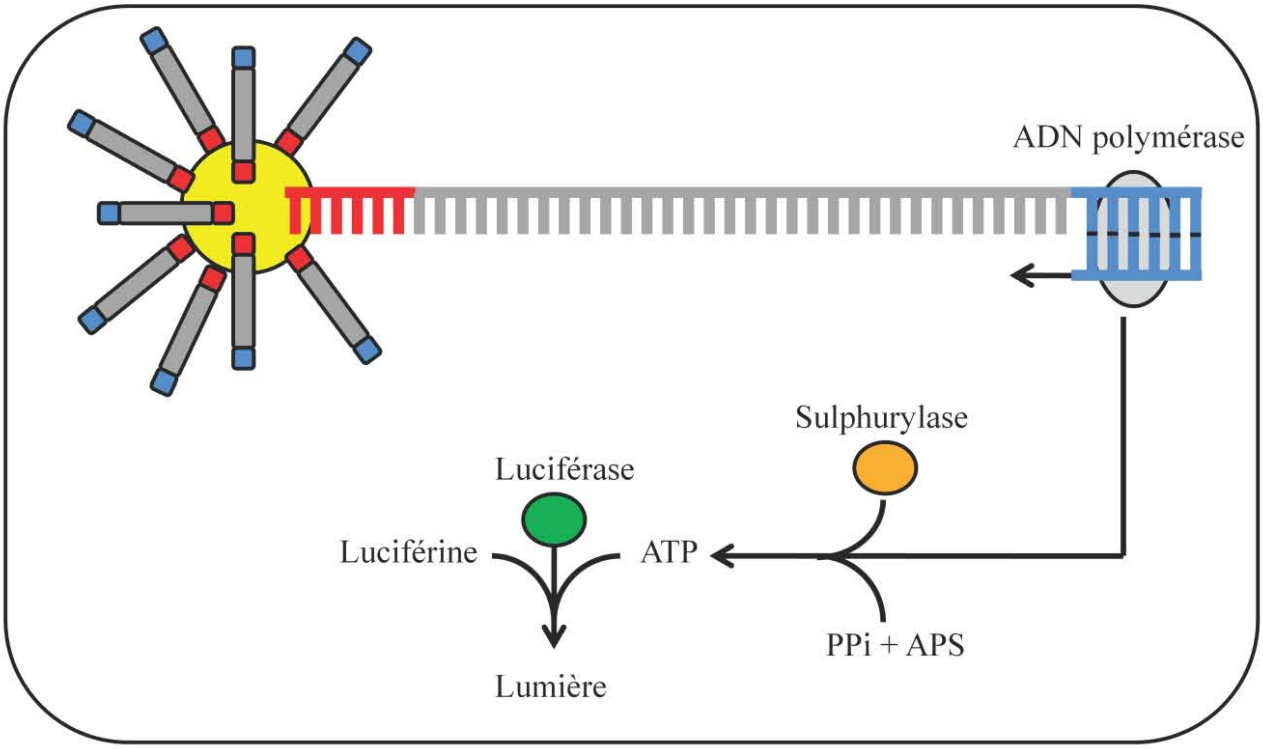


Figure 2 : *Principe du pyroséquençage haut-débit*.
L'ADN polymérase, la Bst polymérase de *Bacillus stearothermophilus*, synthétise le brin complémentaire du fragment d'ADN fixé sur la bille. A chaque cycle de réaction, un nucléotide différent est apporté. Si le nucléotide est incorporé dans le fragment en élongation, le pyrophosphate relâché est utilisé par la sulphurylase pour recréer de l'ATP en présence d'Adénosine 5' – phosphosulfate (APS). Cet ATP est ensuite utilisée par la luciférase avec de la luciférine pour produire un signal lumineux.

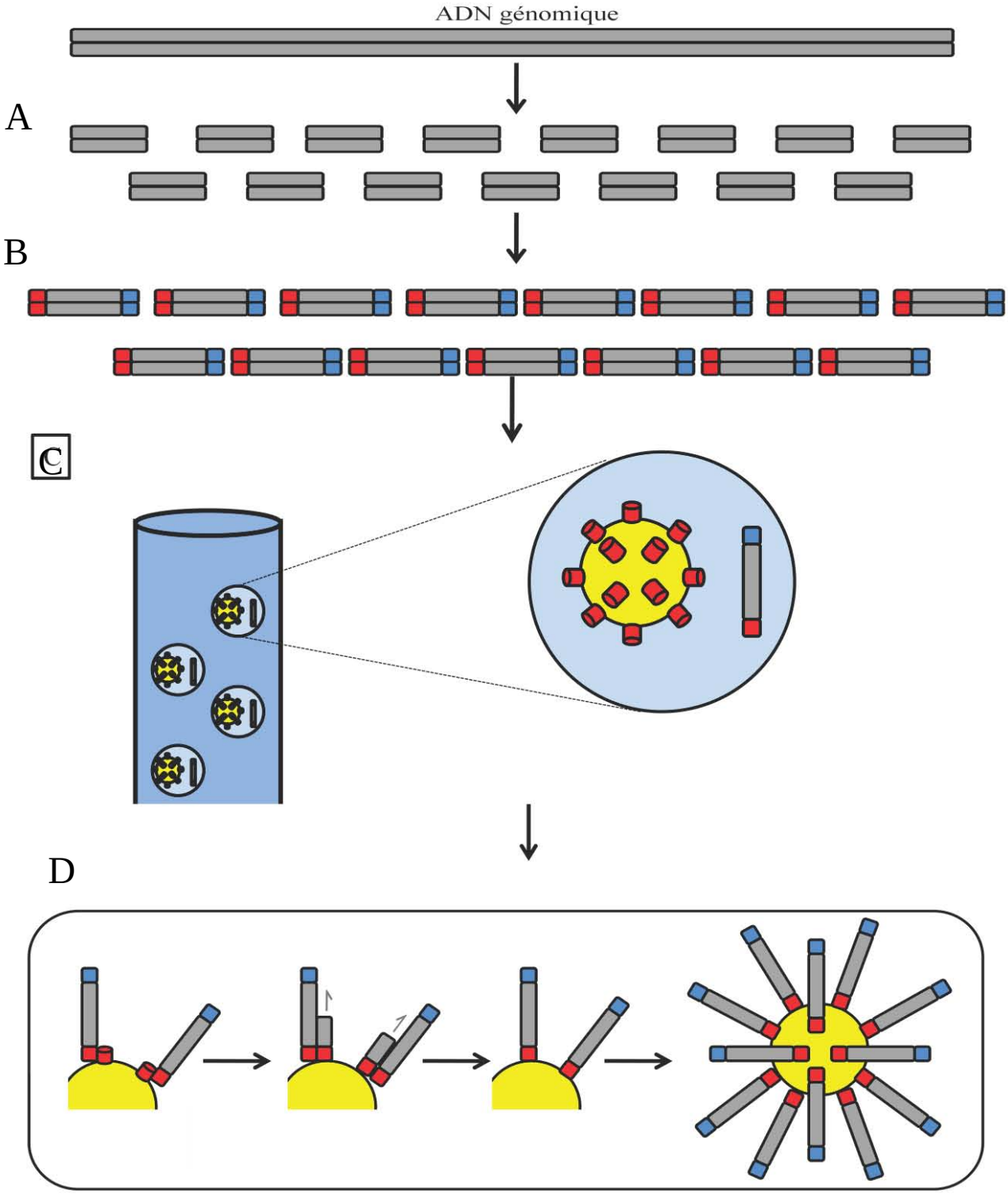


Figure 3 : *Préparation des fragments d'ADN pour le pyroséquençage haut-débit.*
A. L'ADN à séquencer est physiquement cassé par nébulisation en fragments de 400 à 1000 pb (ref manuel). B. Des séquences adaptatrices sont ajoutées aux extrémités de chaque fragment obtenu. C. L'amplification des fragments se fait par une PCR en émulsion (le milieu est composé d'eau et d'huile). Chaque bille est isolée dans une micelle (qui va agir comme un microréacteur) avec un unique fragment d'ADN. D. Les fragments possédant un adaptateur sont immobilisés à l'état simple brin sur une bille pour être amplifiés. Chaque bille porte le même fragment d'ADN en plusieurs millions de copies.

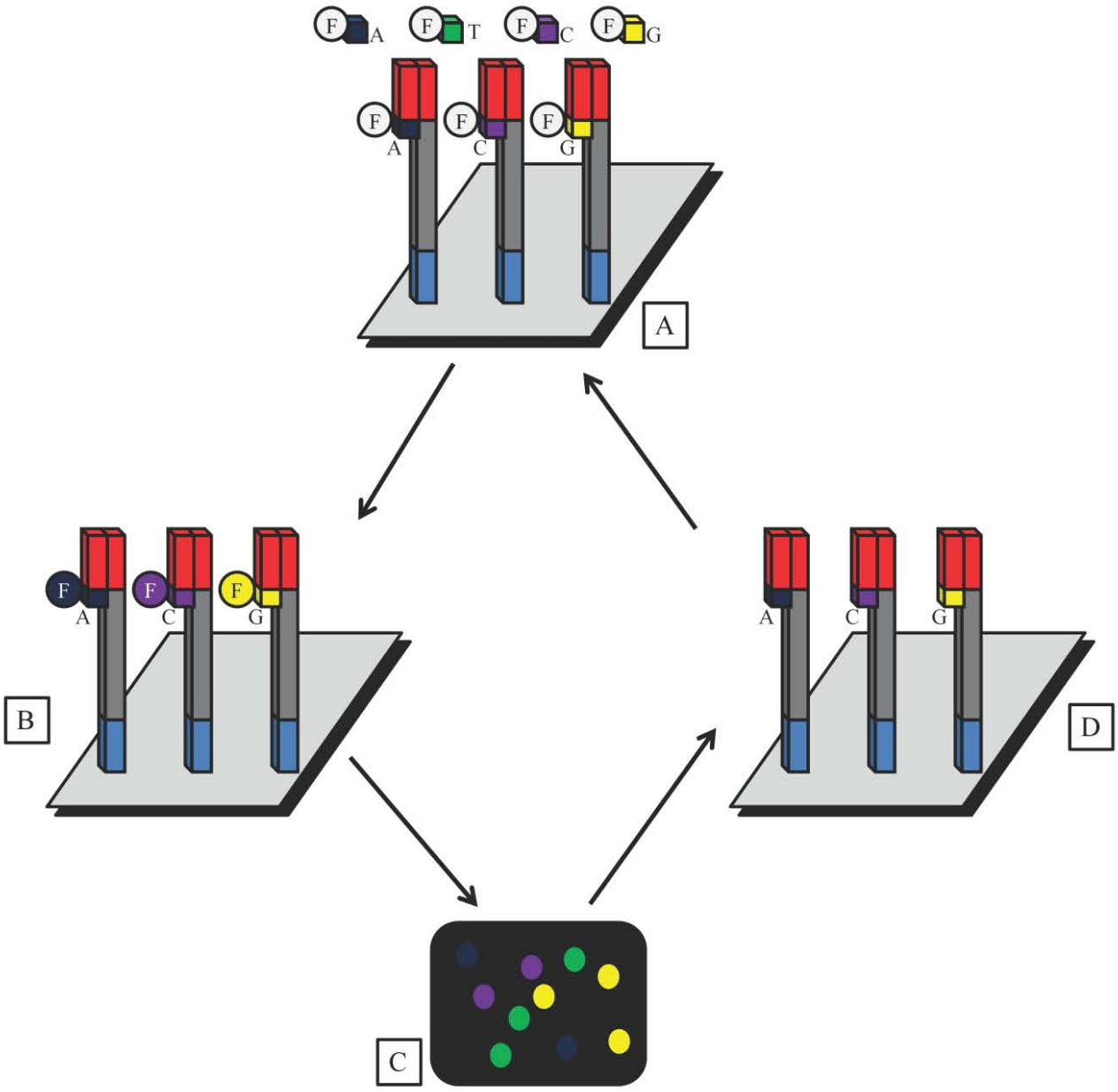


Figure 4 : *Principe du séquençage Illumina/Solexa.*
A. Des nucléotides marqués à l'aide de fluorochromes différents sont apportés au mélange réactionnel. Une ADN polymérase va les insérer dans le brin complémentaire du fragment séquencé. La présence en 3' des nucléotides ne permet l'incorporation que d'un seul nucléotide. B. Les fluorochromes sont excités grâce à des lasers permettant l'émission d'un signal lumineux, spécifique à chaque nucléotide, qui sera détecté par une caméra. C. Exemple de signaux détectés par la caméra du séquenceur. Des programmes informatiques vont traduire les signaux en séquence nucléotidique. D. Après détection, le fluorochrome et le terminateur de chaque nucléotides sont retirés pour permettre l'initiation d'un nouveau cycle.

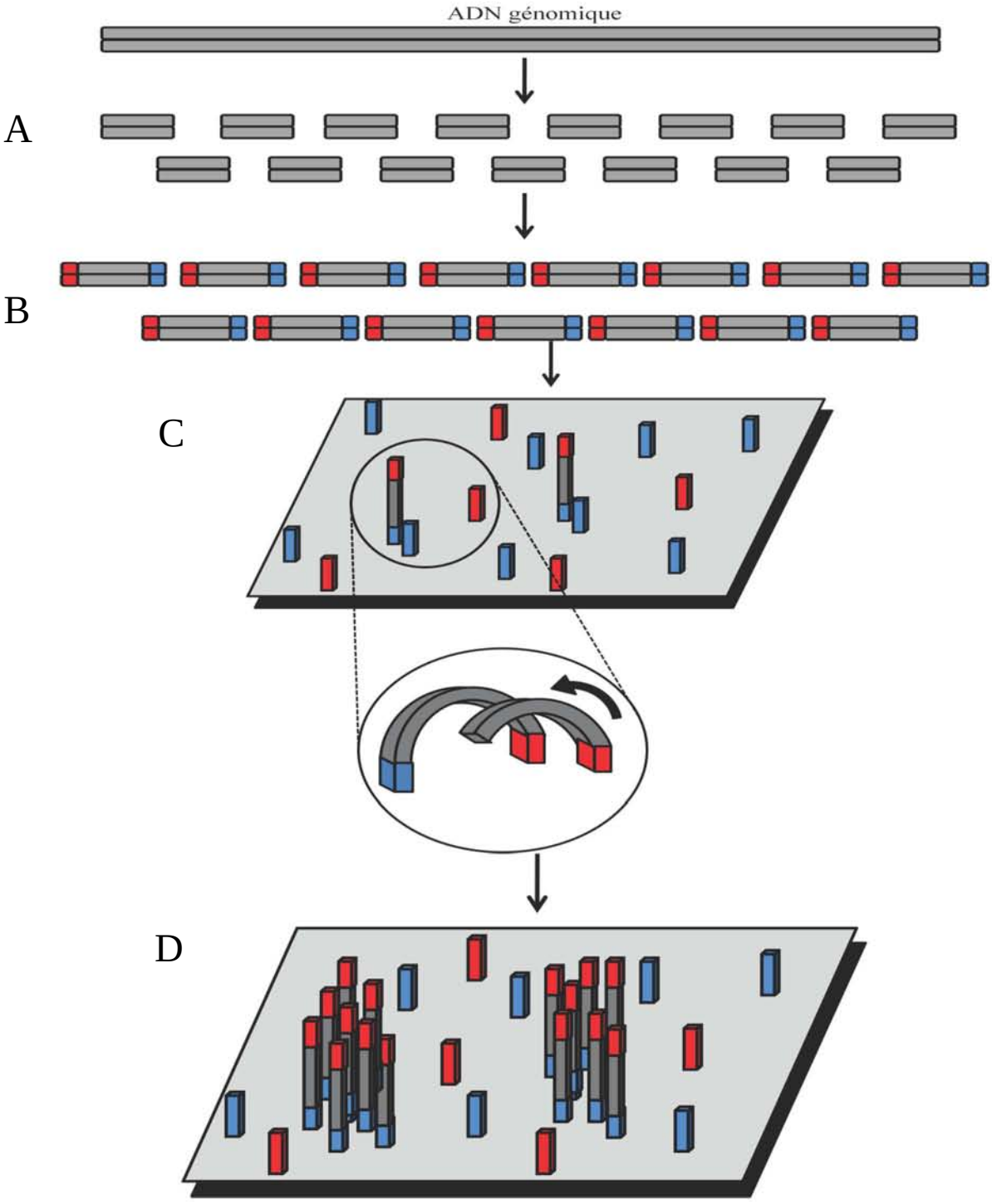


Figure 5 : *Préparation des fragments d'ADN pour le séquençage Solexa/Illumina.*
A. L'ADN à séquencer est fragmenté par nébulisation en fragments d'environ 400 à 500 pb. B. Des séquences adaptatrices sont ajoutées aux extrémités des fragments. C. Les fragments à séquencer sous forme simple brin sont immobilisés sur une plaque de verre à l'aide d'amorces s'hybridant avec un des adaptateurs. D. L'amplification des fragments se fait par pontage à l'aide de l'ADN polymérase. Cela permet l'amplification localisée d'un fragment unique en plusieurs centaines de millions de fragments identiques, appelés « polonie ».

faible des lectures générées, entre 35 et 100 pb, comparées à celles générées par le pyroséquençage, jusqu'à 400 pb, et de la méthode Sanger, jusqu'à 1 kb. L'autre biais de cette technique est la diminution de la fiabilité du séquençage aux extrémités des lectures. Dohm *et al*. (2008) ont émis l'hypothèse que l'accumulation d'erreurs à l'extrémité 3' des lectures était due au phénomène de déphasage, qui correspond à l'extension incomplète ou l'addition de multiples nucléotides au fragment séquencé. Ainsi, plus le nombre de cycles augmente, plus les décalages dans la séquence s'accumulent conduisant ainsi à une augmentation du bruit de fluorescence et une interprétation erronée des signaux lumineux (Erlich *et al.* 2008).

1.3°) Evolution et impact du séquençage haut-débit

La principale amélioration apportée par les nouvelles technologies de séquençage est la quantité de données qui peuvent être générées en une seule réaction. En effet, alors qu'il est possible de séquencer plusieurs millions d'échantillons en une seule fois, 384 séquences au maximum peuvent être générés avec les séquenceurs les plus récents utilisant la méthode de Sanger automatisée. La construction de la banque est aussi grandement facilitée : il n'est plus nécessaire de cloner les fragments dans un vecteur, puis de les amplifier chez l'hôte *Escherichia coli*, ce qui permet une économie de temps pour la préparation des matrices séquencées. A la place, l'ADN à séquencer est fragmenté puis des adaptateurs sont ligués aux extrémités des fragments obtenus avant d'être amplifiés par une méthode spécifique à chaque méthode de séquençage.

De par les progrès techniques apportés ces dernières années, le séquençage est devenu plus rapide et plus efficace. La **figure 6** montre l'évolution de ces techniques, à la fois en terme de débit mais aussi en terme de temps nécessaire à l'exécution d'un séquençage. Au fur et à mesure que de nouvelles techniques de séquençage sont mises au point, la quantité de données générées augmente sensiblement alors que le temps nécessaire à un séquençage continue de diminuer.

La diminution du coût des séquençages a notamment été motivée par le projet « 1000 $ genome », dont l'objectif est de permettre le séquençage du génome d'un Homme à des fins essentiellement médicales. Le premier génome humain, a fini d'être séquencé en 2003 pour la somme de 3 milliards de dollars. L'année suivante, l'équipe de Craig Venter a séquencé le génome de ce dernier pour la somme de 70 millions de dollars. Avec les nouvelles techniques de séquençage, un génome humain peut être obtenu pour une centaine de milliers d'euros (cf tableau 1). Pour illustrer la diminution de l'investissement nécessaire à l'obtention du génome d'un être humain, récemment le projet « 1000 génomes », visant à séquencer le génome de 1000 personnes dans le but d'apprécier la variabilité génétique chez l'homme, a été lancé pour la communauté scientifique.

2°) L'assemblage des données issues du séquençage nouvelle génération

Un assembleur est programme informatique dont la tâche est de combiner les lectures obtenues après séquençage en séquences contigües, aussi appelées contig, en se basant sur l'identité de séquences entre les lectures. De cette façon, deux lectures se chevauchant proviennent vraisemblablement de la même région. L'étape de « *scaffolding* » (échaffaudage) a pour but d'orienter et de placer les contigs les uns par rapport aux autres pour obtenir une séquence unique, bien que possédant des lacunes (soit des trous dans la séquence finale). Ces lacunes seront corrigées lors de l'étape de finition du génome qui vise à combler les trous restant dans l'assemblage par des méthodes expérimentales, comme la PCR ou l'extension d'amorces.

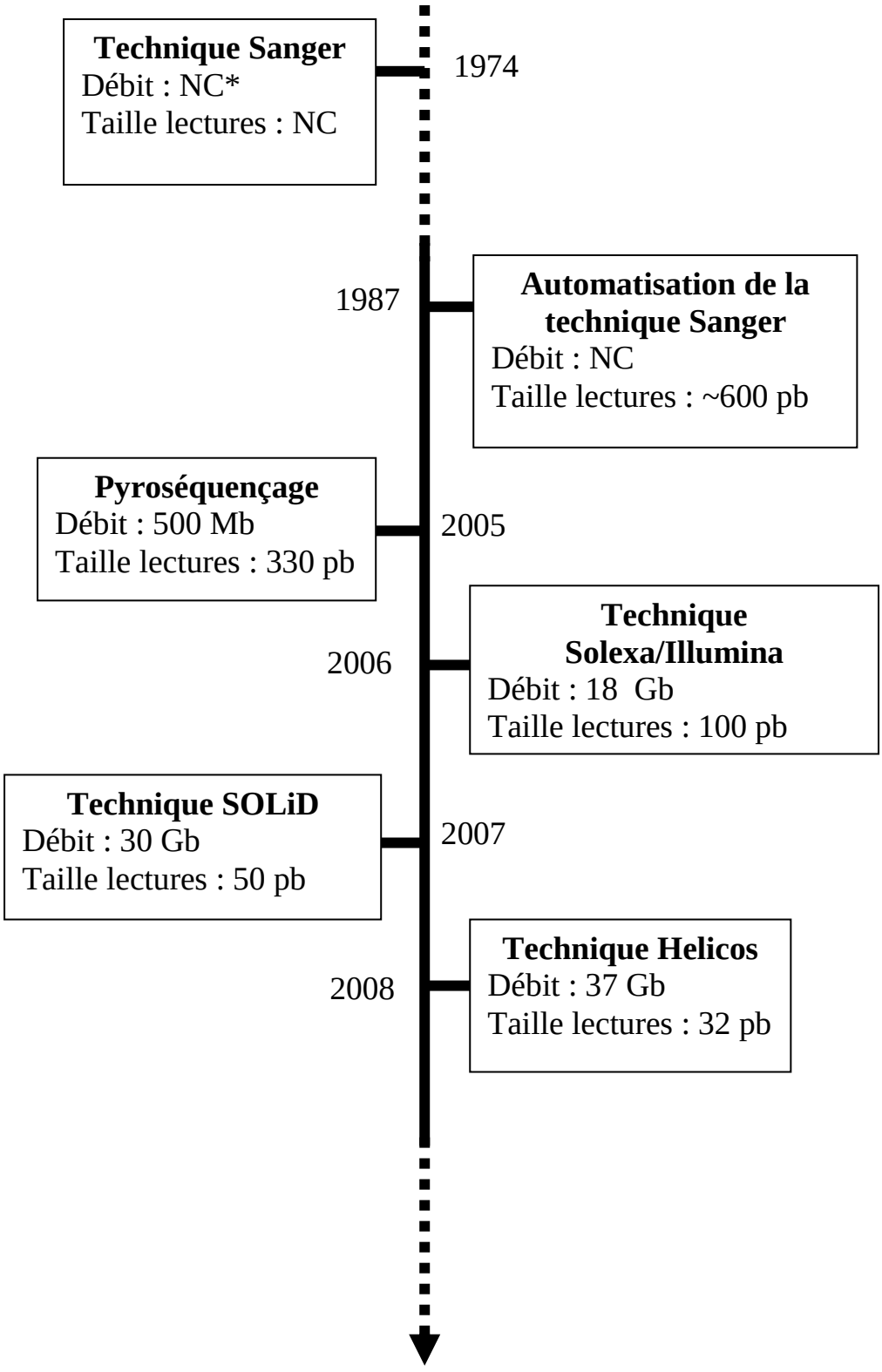


Figure 6 : *Schéma montrant l'évolution des techniques de séquençage depuis la méthode originelle de F. Sanger.*

* : Sept ans ont été nécessaires à l'équipe de Sanger pour séquencer le génome du virus ϕX174 (taille : 5386 pb) (Sanger *et al.* 1977).

NC : Non Connu.

Des indices statistiques tels que le N50[1], la taille moyenne ou la taille cumulée des contigs permettent d'estimer la qualité d'un assemblage. Le postulat est que plus les contigs formés sont grands, moins il y aura de trous dans la séquence génomique reconstituée. Toutefois, il est nécessaire de vérifier la justesse des contigs formés (la présence de contigs chimériques notamment).

2.1°) Les difficultés de l'assemblage

La principale difficulté lors d'un assemblage est la présence de répétitions dans les génomes séquencés. En effet, pour un assembleur deux lectures se chevauchant parfaitement proviennent de la même région génomique. Or, certaines séquences peuvent se retrouver de multiples fois dans un génome entrainant la formation de contigs chimériques et de trous dans l'assemblage final (**figure 7**). Pour s'affranchir de ce problème, deux solutions sont possibles. La première est d'utiliser une technique de séquençage produisant des lectures les plus longues possibles. Si une lecture est plus longue que la répétition, la présence de séquence spécifiques de la répétition permettra de résoudre le problème. L'autre solution est d'utiliser une banque de séquence dite « *paired-end* » (aussi appelée « *mate-pair* »). Les lectures ainsi produites sont regroupées par paire dont l'orientation et la distance les séparant sont renseignées, ajoutant une contrainte à l'assembleur qui doit respecter ces deux critères supplémentaires lors de la recherche de chevauchements entre séquences.

L'autre difficulté de l'assemblage de séquence se situe d'un point de vue matériel. La quantité de données générées par les nouvelles techniques de séquençage augmentant constamment, leur traitement demande l'utilisation de matériel informatique de plus en plus performant.

Pour s'affranchir de ces difficultés, les assembleurs mis au point utilisent des méthodes de calcul permettant d'alléger les ressources matérielles demandées mais aussi d'améliorer la prise en compte des répétitions.

2.2°) Les méthodes d'assemblage des séquences

2.2.1°) La méthode « *overlap-layout-consensus* »

Cette méthode est l'une des premières utilisées avec succès pour assembler une séquence génomique (Miller *et al.* 2010). Le principe général est de fusionner des séquences chevauchantes, en commençant par celles ayant les chevauchements les plus significatifs, jusqu'à formation d'une séquence unique (**figure 8**). La première étape, dite *overlap*, sert à calculer le meilleur alignement possible entre deux lectures en tenant compte des éventuelles erreurs de séquençage. Chaque alignement est pondéré par un score qui va tenir compte du pourcentage d'identité nucléotidique et de la longueur de cet alignement entre deux lectures.

L'étape de *layout* permet de déterminer l'ordre dans lequel les lectures vont être assemblées. Pour la plupart des assembleurs, des paires de lectures sont sélectionnées en fonction de leur score (les scores les plus élevés en premier) et, si l'alignement est considéré comme cohérent, le deux séquences sont fusionnées pour former un contig. Cette étape est la plus complexe de par la difficulté à déterminer la cohérence de l'alignement entre deux lectures. En effet, il est nécessaire de discriminer un chevauchement réel, où les différences entre les deux séquences sont dues aux erreurs de séquençage, d'un chevauchement dû à une répétition, où les différences sont dues aux mutations ou de vraies variations génétiques.

[1] : Le N50 est la taille du plus petit contig tel que 50 % de la longueur cumulée de l'ensemble des contigs obtenus après assemblage soit contenue dans des contigs de taille égale ou supérieure.

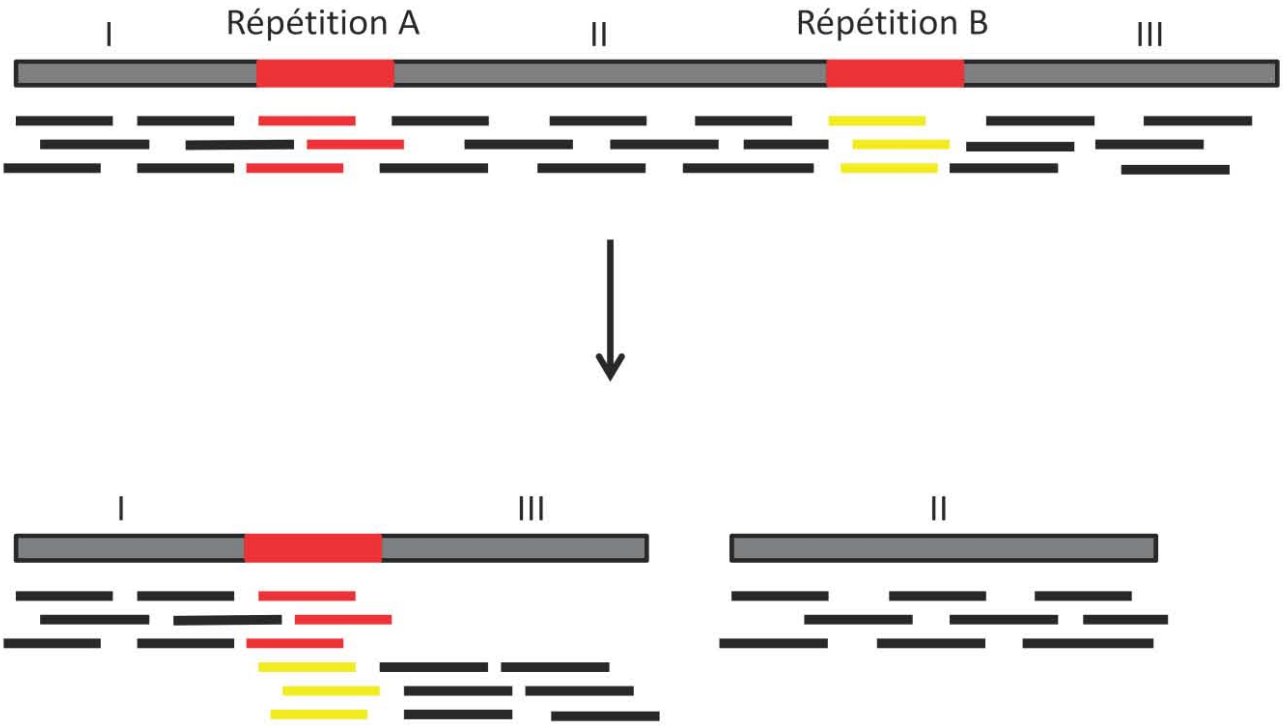
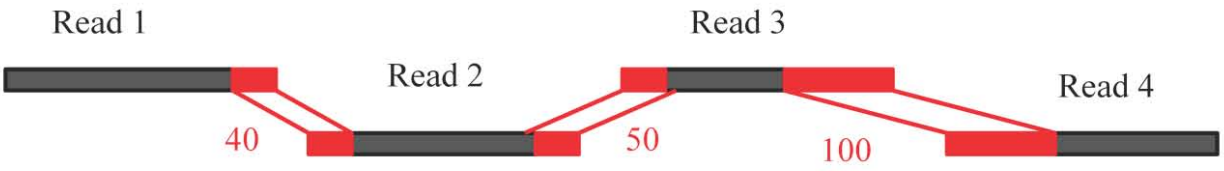


Figure 7 : *Schématisation du comportement d'un assembleur vis-à-vis d'une répétition.*

Pour l'assembleur, les lectures provenant de la répétition A (lectures rouges) et de la répétition B (lectures jaunes) sont supposées provenir de la même région génomique. Cela peut entrainer la formation de contigs chimériques et l'apparition de trous dans l'assemblage final.



Read 1

Read 2 Read3

Read 4

Figure 8 : *Schématisation du principe de fonctionnement d'un assembleur utilisant la méthode Overlap-Layout-Consensus.*

Dans un premier temps, des chevauchements entre les lectures sont déterminés en se basant sur l'identité de séquence. Chaque alignement est considéré comme significatif selon la longueur de l'alignement et le pourcentage d'identité de séquence (ici pour l'exemple il y une pondération par un score bien qu'un assembleur puisse utiliser une autre méthode pour définir si un alignement est correct). La séquence génomique est recréée en alignant l'ensemble des lectures qui se chevauchent.

La dernière étape dite de *consensus*, sert, comme son nom l'indique, à établir la séquence consensus des contigs obtenus lors de l'étape de *layout*. La méthode la plus simple est de déterminer à chaque position quelle base est la plus représentée.

2.2.2°) L'approche du graphique « De Bruijn »

Cette méthode d'assemblage a été mise au point pour s'affranchir de l'étape complexe de *layout* utilisée par la méthode « *overlap-layout-consensus* » (Pevzner *et al.* 2001). Les lectures qui doivent être assemblées sont préalablement fragmentées en sous-séquences chevauchantes de taille constante. Ces sous-séquences sont appelées K-mer (avec K correspondant à la longueur). Deux K-mer vont être assemblés si leurs séquences ne divergent que par le premier de l'un et le dernier nucléotide de l'autre (**figure 9**). Ainsi une identité parfaite doit être partagée entre les deux séquences qui se chevauchent. De cette façon il est théoriquement possible d'établir un lien entre chacun des K-mer, la présence des répétitions empêchant ce cas idéal dans la pratique, et de reconstituer une séquence génomique (**figure 9**).

Il semble paradoxal de fragmenter des lectures alors que leur longueur est importante pour la résolution des répétitions. Cependant, cette méthode demande peu de ressources computationnelles par rapport au nombre de lectures qu'il est possible d'assembler. En effet, la recherche de chevauchements par la méthode « *overlap-layout-consensus* » nécessite de calculer des alignements entre chaque lectures. Ainsi, plus la quantité de données à analyser augmente, plus le temps et les ressources informatiques nécessaires à l'assemblage augmentent. Pour l'approche de « de Bruijn », cette étape de *layout* n'a pas à être opérée pour l'assemblage, car une identité de séquence parfaite doit exister entre deux lectures pour qu'elles se chevauchent. Cette approche d'assemblage est devenue la référence pour assembler des séquences issues de techniques de séquençage produisant un grand nombre de données comme celles d'Illumina/Solexa ou de SOLiD.

Toutefois, ce principe d'assemblage est particulièrement sensible aux erreurs car une identité parfaite entre K-mer est requise pour valider un lien. Ainsi en pratique, l'assemblage de séquences par cette méthode demande des lectures possédant peu d'erreurs pour être efficace.

2.3 °) Vue d'ensemble des assembleurs et leur utilisation.

L'utilisation d'un assembleur particulier sera conditionnée par de nombreux paramètres tels que la stratégie de séquençage utilisée, la présence de « *paired-end* » ou encore la quantité de données. Les assembleurs utilisant la méthode « *overlap-layout-consensus* » sont plutôt utilisés avec des lectures allant de 100 à 800 pb (principalement avec un séquençage Sanger ou un pyroséquençage haut-débit). La méthode graphique « De Bruijn » s'emploie plutôt avec des lectures d'une taille allant de 25 à 100 pb. Cette dernière est la méthode d'assemblage la plus utilisée à l'heure actuelle grâce au succès de la plateforme Illumina/Solexa. Toutefois, avec les améliorations des techniques de séquençage, la taille des lectures ne cesse d'augmenter et la méthode « *overlap-layout-consensus* » devrait redevenir la référence dans le futur.

La plupart des assembleurs ont démontré leur efficacité sur l'assemblage de génomes bactériens car ils sont plus simples à assembler que les génomes eucaryotes, notamment par la plus faible longueur de leurs génomes et leur plus faible teneur en répétitions. Il faut toutefois noter que certains assembleurs ont été testés sur des données simulées (Butler *et al.* 2008) alors que d'autres ont montré leur efficacité sur des données réelles (MacCallum *et al.* 2009), ce dernier critère étant plus déterminant dans le choix de l'utilisation d'un assembleur.

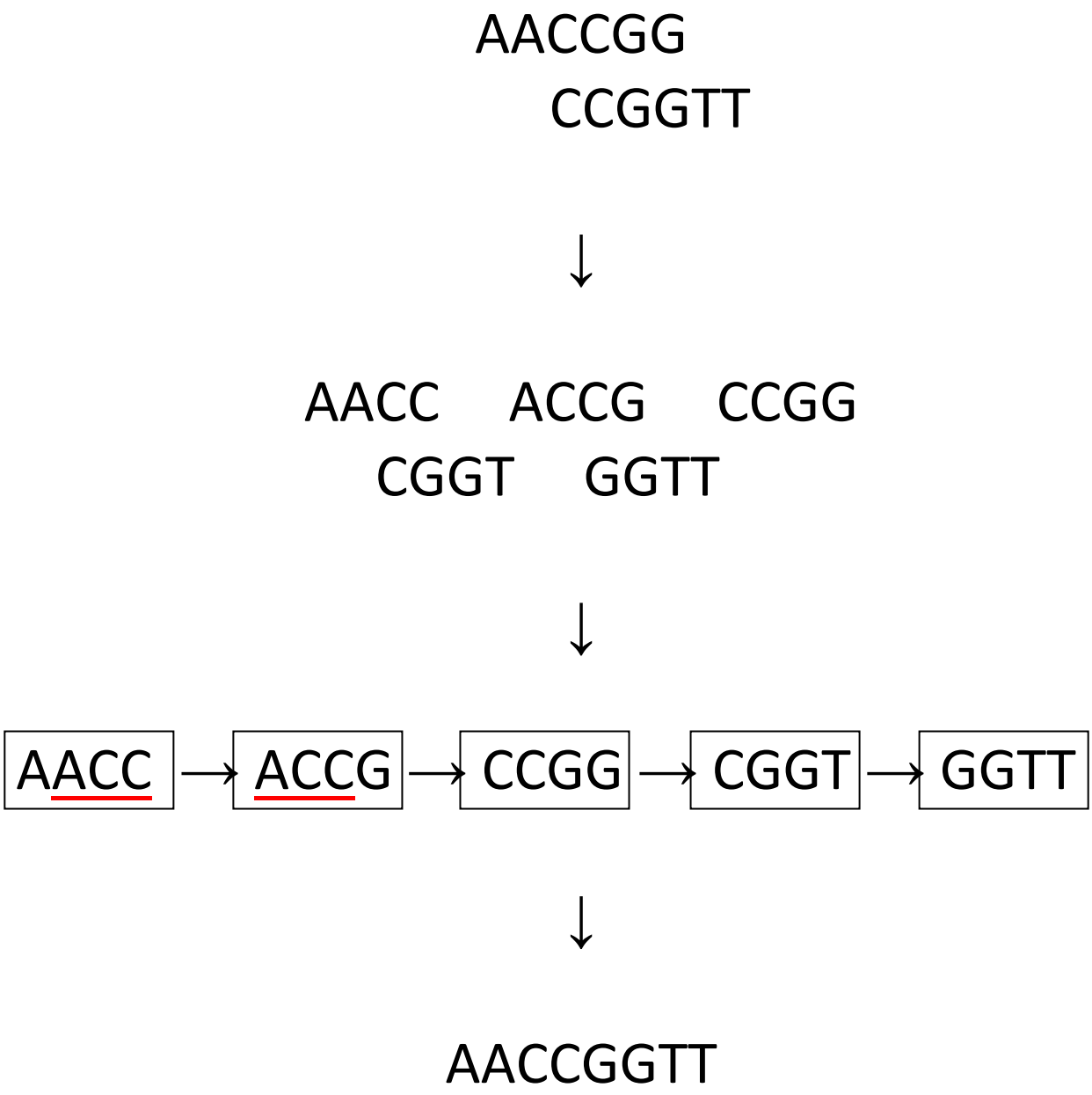


Figure 9 : *Représentation schématique de l'assemblage de deux séquences par la méthode graphique « De Bruijn ».*

Les deux séquences assemblées pour cette exemple sont découpées en 4-mer chevauchants. Il est à noter que le 4-mer CCGG qui est une répétition n'est représenté qu'une fois par la méthode graphique de « De Bruijn ». Une identité de séquence parfaite entre les 3 premiers nucléotides de l'un et les 3 derniers de l'autre est recherchée entre ces sous-séquences. Si il y a concordance exacte (comme par exemple pour les 4-mer soulignés en rouge), l'assembleur va relier les deux fragments. Le consensus obtenu fournit la séquence initiale.

3°) Le séquençage des génomes de *Streptomyces*

3.1°) Présentation des bactéries du genre *Streptomyces*

Les *Streptomyces* sont des bactéries Gram positives appartenant à l'ordre des *Actinomycetales*. Ces bactéries sont ubiquistes des sols. Leur mode croissance est original car elles se développent selon un cycle cellulaire complexe au cours duquel une différentiation morphologique et biochimique poussée (Chatter, 1993). En conditions favorables de croissance, une spore de *Streptomyces* va germer puis se développer pour former un hyphe. Le développement de ces hyphes végétatifs, correspondant à une structure coenocytique contenant de multiples génomes (20 à 50), entraine la formation d'un mycélium végétatif, permettant ainsi la recherche de substrat nécessaire à la croissance. Quand les conditions de croissance du milieu deviennent défavorables, la croissance tout d'abord rampante devient verticale et les hyphes se différencient alors en spores unigénomiques. Ainsi le cycle de croissance pourra être ré-initié (**figure 10**).

Les *Streptomyces* sont réputées pour leur importance biotechnologique. Leur métabolisme secondaire produit de nombreux composés d'intérêt, notamment des antibiotiques. En effet, les bactéries du genre *Streptomyces* produisent plus de 60% des antibiotiques utilisés en thérapie humaine ou vétérinaire (Berdy, 2005).

3.2°) Le séquençage des génomes de *Streptomyces*

3.2.1°) Caractéristiques génomiques des *Streptomyces*

Les *Streptomyces* possèdent des caractéristiques génomiques particulières avec la linéarité du génome (Lin *et al.* 1994) et le fort pourcentage en bases G/C (de l'ordre de 70%). Aux extrémités du chromosome se trouvent des séquences inverses répétées (appelées TIR pour Terminal Inverted Repeat) de taille variable selon les espèces de *Streptomyces* (de 5 kb à 1,4 Mb (Wenner *et al.* 2003)). Ces TIR sont impliquées dans le processus de réplication de l'ADN de la bactérie. De par leur repliement à l'état simple brin, les télomères vont permettre la fixation d'une protéine terminale TP (pour Terminal Protein). Le chromosome des *Streptomyces* se répliquant de façon bidirectionnelle depuis une origine de réplication centrale (Musialowski *et al.* 1994), la linéarité du réplicon pose problème aux extrémités. Le retrait du dernier fragment d'Okazaki sur le brin retardé entraine la formation d'une lacune qui sera comblée par une phase de synthèse d'ADN initiée depuis la TP (Bao et Cohen, 2003).

3.2.2°) Compartimentation génomique des *Streptomyces*

A l'heure actuelle, cinq génomes de *Streptomyces* ont été séquencés et assemblés entièrement : *Streptomyces coelicolor* (Bentley *et al.* 2002), *S. avermitilis* (Ikeda *et al.* 2003), *S. scabies*, *S. griseus* (Ohnishi *et al.* 2008) et *S. clavuligerus*.

La comparaison entre ces génomes a révélé une compartimentation originale du génome. Alors que les régions centrales des génomes sont conservées entre espèces, les régions terminales sont variables et spécifiques à chaque espèce. Il a été montré que la taille de ces régions spécifiques augmentait avec la distance phylogénétique séparant les deux organismes considérés (Choulet *et al.* 2006(a)). Des expériences précédentes chez *Streptomyces ambofaciens* et *Streptomyces lividans* ont aussi démontré que les régions terminales étaient délétables sans affecter la croissance végétative en conditions de laboratoire (Fischer *et al.* 1997). En revanche, la différenciation cellulaire et biochimique est bien souvent affectée par les réarrangements. L'étude du contenu en gènes des régions terminales a montré une densité en séquences codantes équivalente à la région centrale.

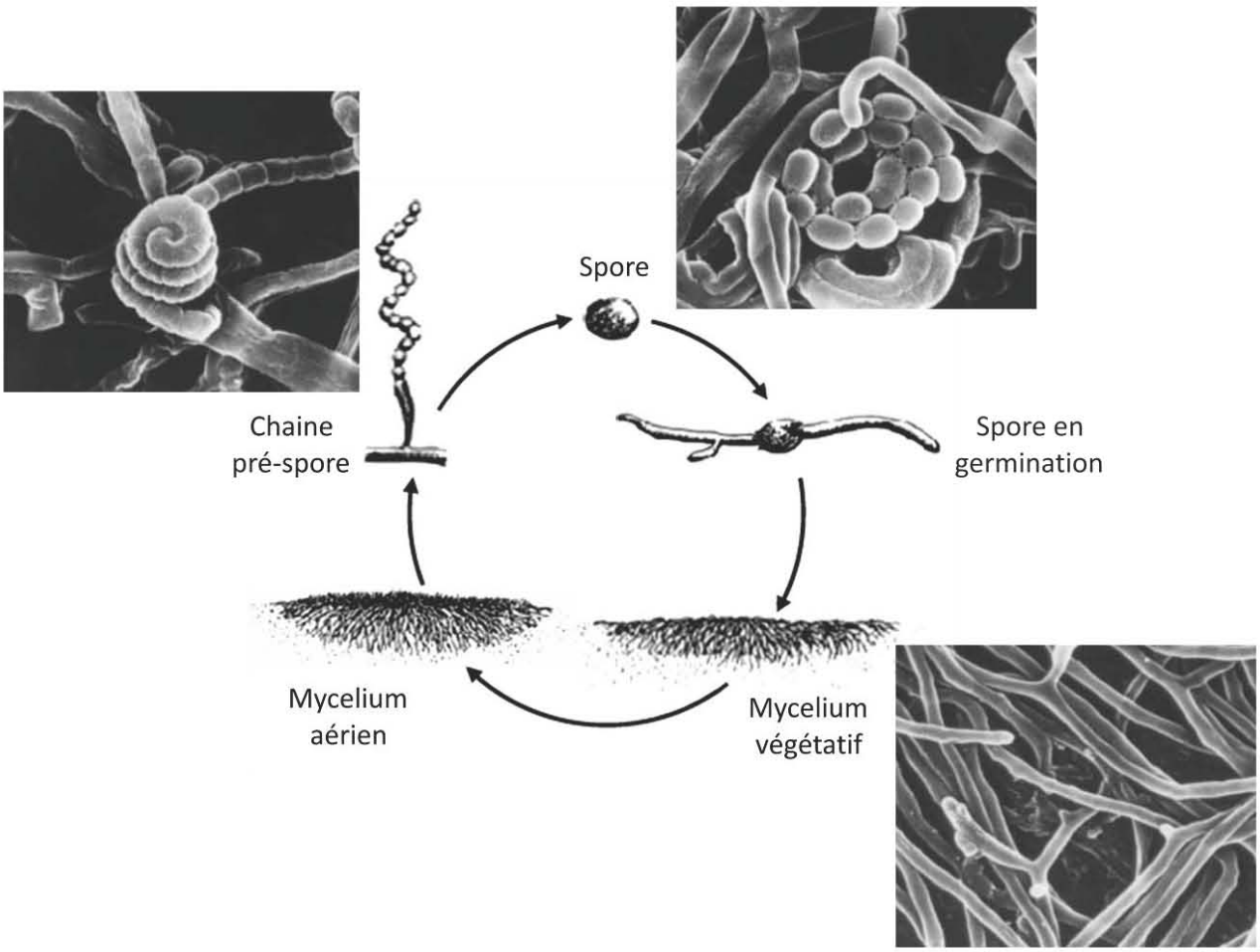


Figure 10 : *Cycle de développement des Streptomyces.*
La germination de la spore produit des filaments mycéliens qui se ramifient. Ce mycélium végétatif se différencie en mycélium aérien quand le milieu devient limitant en éléments nutritifs. Les hyphes aériens vont ensuite produire des spores qui pourront à leur tour germer.

Toutefois, alors que les gènes essentiels à l'organisme (impliqués dans la réplication, la transcription, la traduction ou le métabolisme primaire) sont situés dans le centre du génome, les gènes localisés dans les extrémités sont des gènes accessoires (Bentley *et al.* 2002). Des expériences de transcriptomique ont aussi montré qu'en conditions optimales de croissance, les gènes de la région centrale sont globalement plus exprimés que les gènes situés dans les extrémités chromosomiques, alors qu'en condition de stress l'expression des gènes est plus uniforme le long du génome (Karoonuthaisiri *et al.* 2005). Ces observations ont permis de corroborer l'hypothèse selon laquelle les gènes des régions terminales ne seraient pas nécessaire à la croissance végétative.

3.3°) Intérêts du séquençage des génomes des *Streptomyces*

Au jour du 21/06/2010, 1086 génomes bactériens ont été entièrement séquencés (source : base de données GOLD). Certains biais ont été observés quant aux choix des bactéries séquencées notamment en terme de taille du génome mais aussi de l'intérêt que représente la bactérie pour l'Homme (**figure 11**). Une grande majorité des génomes provient de trois phyla bactériens : les *Proteobacteria*, les *Firmicutes* et les *Actinobacteria* car la plupart des bactéries appartenant à ces groupes peuvent être isolées en cultures pures (**tableau 2**). Toutefois, pour le phylum des *Actinobacteria*, dont les *Streptomyces* font partie, seulement 1% des 12460 espèces ont été séquencées.

Récemment, le projet GEBA (Genomic Encyclopedia of Bacteria and Archae) a vu le jour avec pour but de séquencer une centaine de génomes bactériens sur la base de leur originalité phylogénétique (Wu *et al.* 2009). Les résultats préliminaires de cette étude ont montré que grâce à ce projet, de nouvelles familles de protéines aurait été identifiées par l'augmentation de la diversité des bactéries séquencées.

Bien que les *Streptomyces* possédent un génome de grande taille et présentent des caractéristiques compositionnelles extrêmes, le Broad Institute, l'un des plus gros centre de séquençage mondial, a engagé un programme de séquençage pour 17 souches dans le but d'identifier des nouvelles voies de biosynthèse de molécules à activité anti-microbienne. En effet, plusieurs analyses génomiques ont démontré que ces espèces recelaient une capacité de codage de métabolites secondaires bien plus importante que ne le révèlent les cribles classiques (Challis, 2008). De plus, le séquençage et l'assemblage de génomes complets de ce groupe bactérien apporte des éléments de compréhension des mécanismes évolutifs des grands génomes (Choulet *et al.* 2006(a)), linéaire de surcroit, appartenant à des organismes présentant une différenciation poussée et un métabolisme complexe.

3.4°) Le séquençage de *Streptomyces ambofaciens*

Les bactéries du genre *Streptomyces* sont aussi réputées pour leur forte instabilité génétique. Chez *Streptomyces ambofaciens*, des études réalisées au laboratoire ont montré la haute fréquence d'apparition de mutants affectés dans la pigmentation des colonies (Martin *et al.* 199). Il a été suggéré que cette forte variabilité était due à des réarrangements au niveau des extrémités du chromosome.

Un programme de séquençage, en collaboration avec le Genoscope, a permis d'obtenir la séquence partielle de deux souches de *Streptomyces ambofaciens* : la souche ATCC23877 et la souche DSM40697. Pour la souche ATCC23877, 1 367 119 pb de l'extrémité droite et 1 544 032 pb de l'extrémité gauche ont été obtenus par l'intermédiaire d'un séquençage ordonné (Choulet *et al.* 2006(a)). La taille des TIR de cette souche est de 197 936 pb.

Pour la souche DSM40697, les extrémités droite et gauche ont été séquencées pour obtenir respectivement 237 382 pb et 238 517 pb. La taille des TIR pour cette souche est de 212 655 pb. La comparaison de ces régions avec celles de la souche ATCC23877 a montré qu'une majeure partie

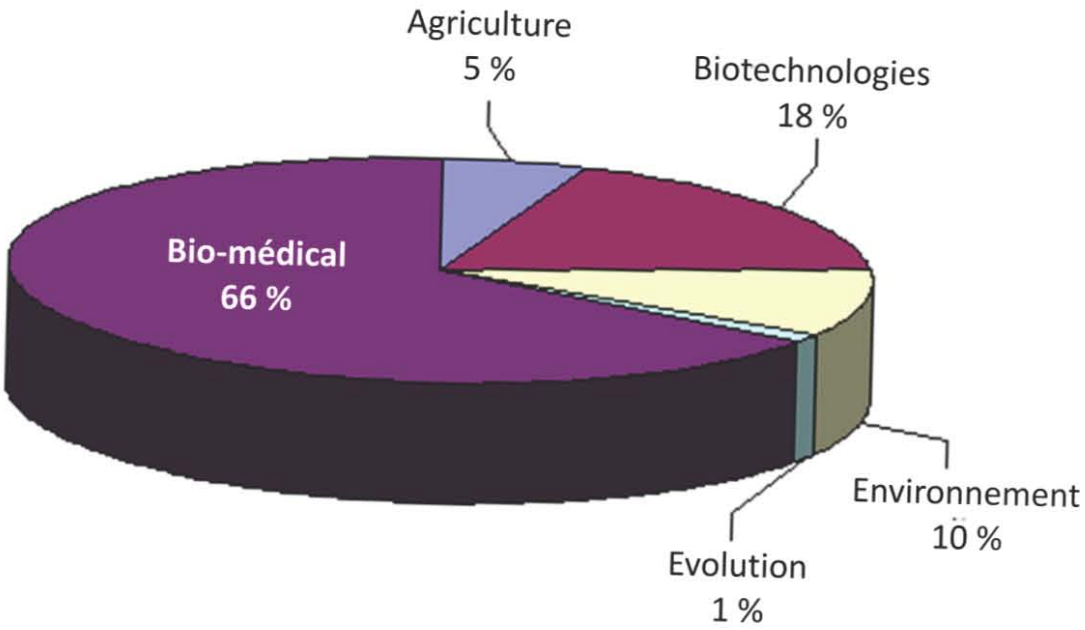


Figure 11 : *Part des financements de séquençage de génomes bactériens en fonction des domaines d'intérêt.*

Les deux tiers des financements accordés aux projets de séquençage concernent le biomédical, avec notamment le séquençage de bactéries pathogènes. Seulement 1% des financements sont accordés à des projets visant à étudier l'évolution bactérienne. (source : Base de données GOLD).

Phylum	Nombre d'isolats (pourcentage)	Nombre de génomes séquencés (pourcentage)	Pourcentage d'isolats séquencés
Proteobacteria	31 597 (46.7)	727 (47.6)	2.3
Firmicutes	15 059 (22.2)	364 (23.8)	2.4
Actinobacteria	12 460 (18.4)	130 (8.5)	1.0
Bacteroidetes	2 889 (4.3)	53 (3.5)	1.8
Cyanobacteria	2 274 (3.4)	51 (3.3)	2.2
Spirochaetes	964 (1.4)	36 (2.4)	3.7
Chlamydiae	122 (0.2)	20 (1.3)	16.4
Chlorobi	109 (0.2)	12 (0.8)	11.0
Chloroflexi	82 (0.1)	12 (0.8)	14.6
Thermotogae	97 (0.1)	10 (0.7)	10.3
Verrucomicrobia	62 (0.1)	8 (0.5)	12.9
Thermi	325 (0.5)	6 (0.4)	1.8
Fusobacteria	118 (0.2)	6 (0.4)	5.1
Planctomycetes	113 (0.2)	6 (0.4)	5.3
Aquificae	90 (0.1)	6 (0.4)	6.7
Synergistetes	42 (0.1)	3 (0.2)	7.1
TOTAL	66 403 (98,2)	1450 (95)	

Tableau 2 : *Estimation du nombre d'isolats et de génomes séquencés pour les principaux phylums bactériens*.

Le phylum des *Actinobacteria* est surligné en gris. Il présente le plus faible pourcentage d'isolats de l'ensemble des phylums bactériens (1%). (source : Wu *et al.* 2009).

des répétitions inverses terminales (les TIR) était très conservées entre les deux espèces. Cependant, les extrémités des TIR contiennent des régions spécifiques à chaque souches qui proviendraient d'échanges entre l'extrémité du réplicon et des plasmides linéaires (**figure 12**). En effet, de nombreux clusters de gènes spécifiques présentent des homologies avec des gènes plasmidiques. De plus, plusieurs évènements d'insertion/délétion, résultant probablement d'évènements de recombinaison illégitime, ont été observés à l'intérieur et à l'extérieur des TIR (Choulet *et al.* 2006(b)). Toutefois, ces observations n'ont été faites que sur les quelques 200 kb disponibles de la souche DSM40697. L'obtention de ces deux génomes de *Streptomyces ambofaciens* permettrait d'établir la première comparaison intra-spécifique chez les Streptomycètes pour découvrir quels seraient les mécanismes d'évolution rapide différenciant ces deux souches.

4°) Objectifs du travail

L'objectifs de ces travaux est d'assembler le génome de *Streptomcyes* ambofaciens ATCC23877. Pour obtenir cette souche, une stratégie de séquençage mixte faisant appel au pyroséquençage haut-débit ainsi qu'à la terminaison cyclique réversible de Solexa/Illumina a été mise en place. Si cette approche de séquençage porte ses fruits, le génome de *Streptomyces ambofaciens* ATCC23877 sera ensuite annoté à l'aide d'une plate-forme d'annotation, en partie mise au point lors de ce stage. Des données de séquençage par pyroséquençage haut-débit ont été auparavant obtenues pour la souche DSM40697 de *Streptomyces ambofaciens*. Des analyses préliminaires de génomique comparative vont être faites en se servant des différentes données de séquençage obtenues pour ces deux souches afin d'estimer le degré de spécificité les séparant.

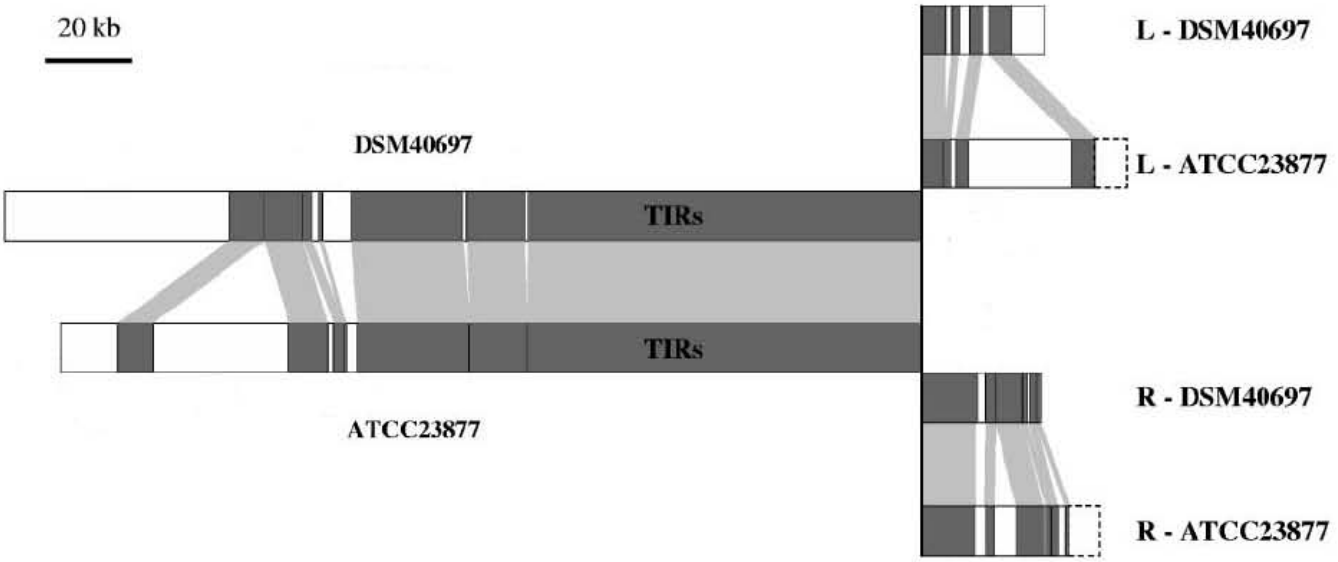


Figure 12 : *Comparaison des extrémités terminales des souches ATCC23877 et DSM40697 de Streptomyces ambofaciens.*

Les régions représentées en blanc sont spécifiques à chaque souche. A droite du schéma est représenté la comparaison entre les séquences bordant les TIRs sur les extrémités droites (R) et les extrémités gauches (L).

MATERIEL
ET METHODES

1°) Séquençage du génome de *Streptomyces ambofaciens*

1.1°) Choix des souches

Deux souches de *Streptomyces ambofaciens* ont été sélectionnées pour le séquençage. La souche ATCC23877, isolée en 1954 du sol de Peronne (région de Picardie) (Pinert-Sindico, 1954), a été choisie car elle est utilisée en industrie pharmaceutique pour la production de l'antibiotique spiramycine. La souche DSM40697, qui a été isolée d'un sol en Italie (Hütter, 1967), a aussi été retenue car des études portant sur les phénomènes d'instabilité génétique ont été faites auparavant sur ces deux souches (Leblond et Decaris, 1999).
Ces deux souches produisant les antibiotiques spiramycine, alpomycine et congocidine (Pang et al, 2004), leur appartenance à la même espèce a initialement été basée sur des caractères essentiellement phénétiques. La comparaison des séquences ITS (Internal Transcribed Spacer) des six opérons ribosomiques des deux souches a montré une identité de séquence parfaite, validant ainsi par une méthode moléculaire l'appartenance de ces deux souches à la même espèce.

1.2°) Stratégie de séquençage envisagée

Le séquençage du génome de *Streptomyces ambofaciens* a été initié avec la souche ATCC23877 au mois de Juillet 2000 en collaboration avec le Genoscope et l'Institut de Génétique et Microbiologie d'Orsay. Les extrémités terminales des deux souches ont été obtenues par un séquençage de type ordonné (Choulet *et al.* 2006(a)).

Pour obtenir la séquence génomique complète des deux souches de *Streptomyces ambofaciens*, un pyroséquençage a été initié auprès d'un prestataire, GATC. Le séquenceur utilisé est le GS FLX qui produit des lectures d'une taille moyenne de 220 pb.

Par la suite, l'obtention du génome de la souche ATCC23877 a été privilégié car l'obtention de sa séquence complète aiderait à assembler plus facilement le génome de la souche DSM40697. Un séquençage par la technique Solexa/Illumina a ainsi été commandé pour cette souche auprès de deux prestataires, GATC et Fasteris. Trois types de banques de séquences ont été utilisées pour la préparation des échantillons.

La première, dite de « *single end* » a été commandée auprès de GATC en 2008. Ce premier séquençage a été utilisé d'une part pour augmenter la couverture du génome et diminuer le nombre de contigs après assemblage, et d'autre part pour limiter les erreurs inhérentes aux séquences homopolymériques du pyroséquençage.

La seconde banque, appelée « *paired end* », permet de séquencer les deux extrémités d'un fragment matrice. Ainsi, en plus de l'information apportée par la séquence, la distance séparant les deux lectures est connue (cette distance ne peut dépasser ~800 pb). Le protocole permettant la création d'une telle banque de séquence est donné en **figure 13**. Ce type de librairie permet un assemblage plus efficace des lectures en passant au travers des répétitions. La distance séparant les lectures de la banque commandée est de 300 pb.

La troisième banque utilisée est appelée « *mate-pair* ». Comme pour le « *paired-end* », la distance séparant les deux lectures est connue, mais elle est beaucoup plus grande : de 1 à 20 kb selon les cas. Cette banque est très utile pour remettre en ordre et orienter les contigs obtenus les uns par rapport aux autres (étape de « *scaffolding* »). Toutefois, le prestataire a eu des difficultés à produire des séquences correctes en utilisant le protocole fourni par Illumina (**figure 14**). Le prestataire a donc mis au point un protocole qui leur est propre pour réussir à créer une banque correcte (**figure 15**). La distance séparant les deux lectures de la banque commandée est d'environ 3 kb.

Les détails sur le séquençage de ces fragments est donné en **figure 16**.

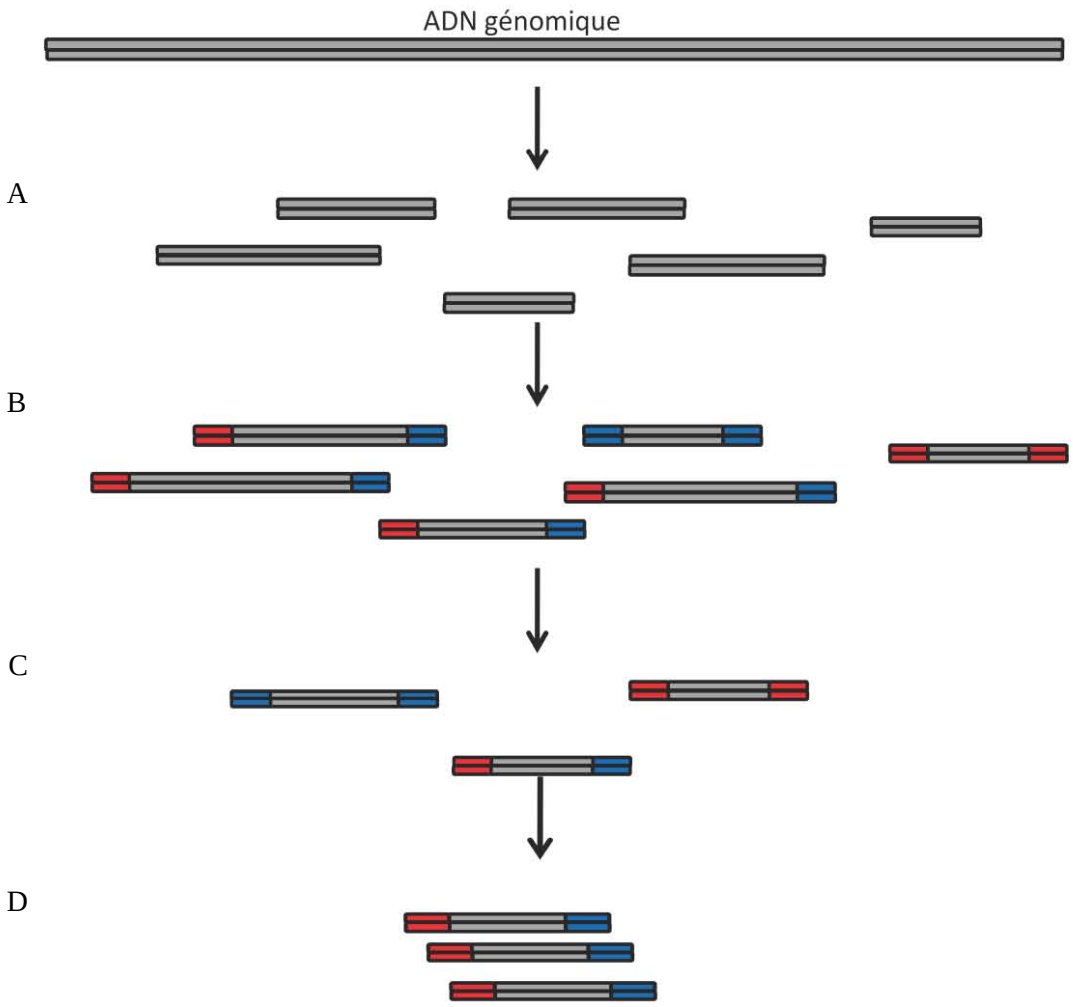


Figure 13 : *Protocole de préparation de la banque « paired-end » selon Illumina.*
A. L'ADN génomique est fragmenté par nébulisation. B. Des séquences adaptatrices sont liguées aux extrémités des fragments, puis le mélange réactionnel est mis à migrer par électrophorèse. C. Les fragments de taille désirée sont purifiés à partir du gel. La taille des fragments définira la distance séparant les deux lectures séquencées. D. Les fragments possédant les bonnes séquences adaptatrices sont sélectionnées par enrichissement spécifique de ces matrices par PCR.

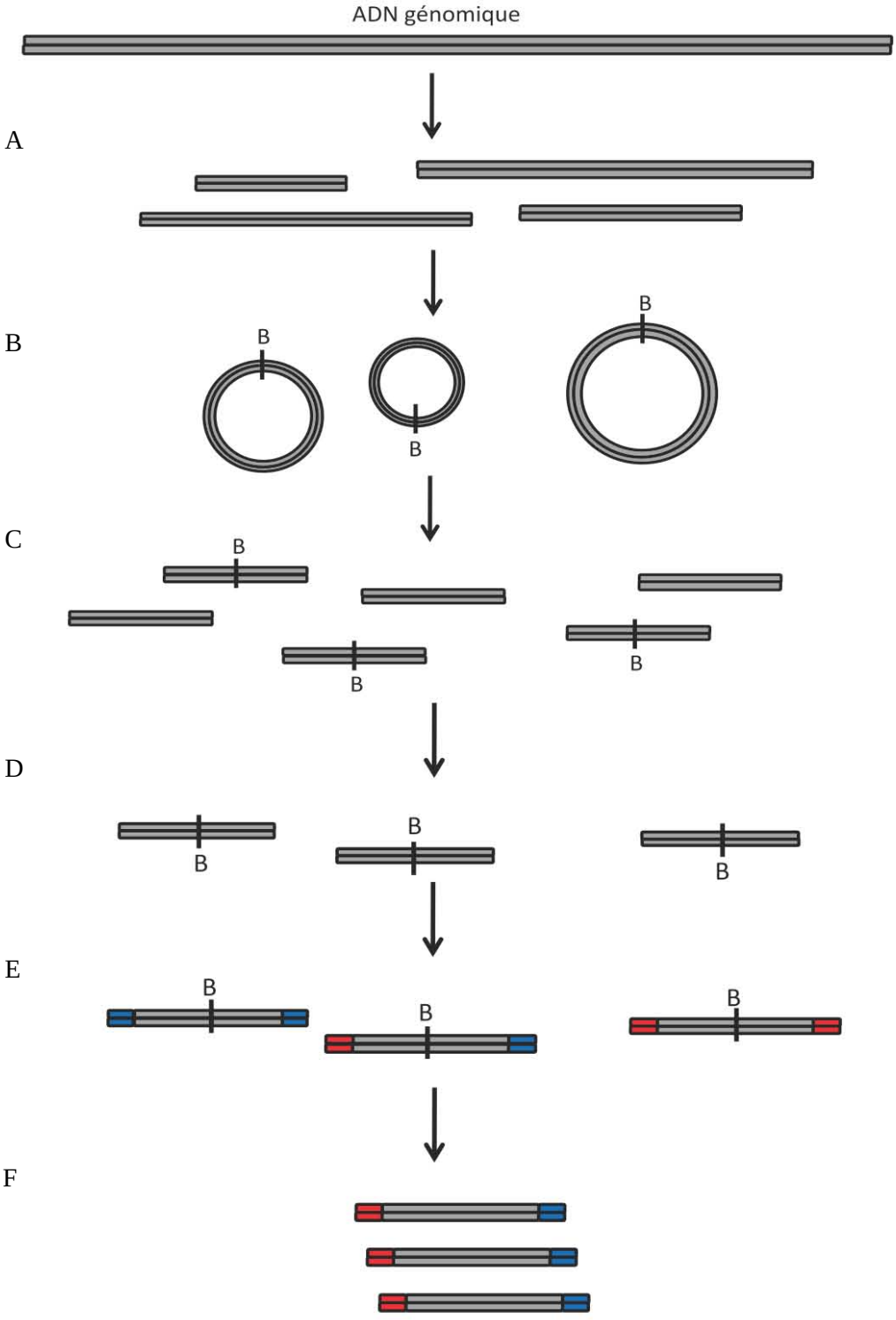


Figure 14 : *Protocole de préparation de la librairie « mate-pair » selon Illumina.*
A. L'ADN génomique est fragmenté par nébulisation. Les fragments faisant une taille comprise entre 2 et 4 kb, correspondant a la distance qui séparera les deux lectures séquencées, sont sélectionnés. B. Les extrémités sont réparées à l'aide de base biotinylées puis le fragment est circularisé. C. Une seconde étape de nébulisation permet d'obtenir des fragments d'environ 400 pb. D. Les fragments biotynilés, contenant les deux extrémités des fragments de 2-4 kb de départ, sont sélectionnés. E. Des séquences adaptatrices sont ajoutées aux extrémités des fragments isolés. F. Les fragments possédant les bons adaptateurs à leurs extrémités sont amplifiés par PCR.

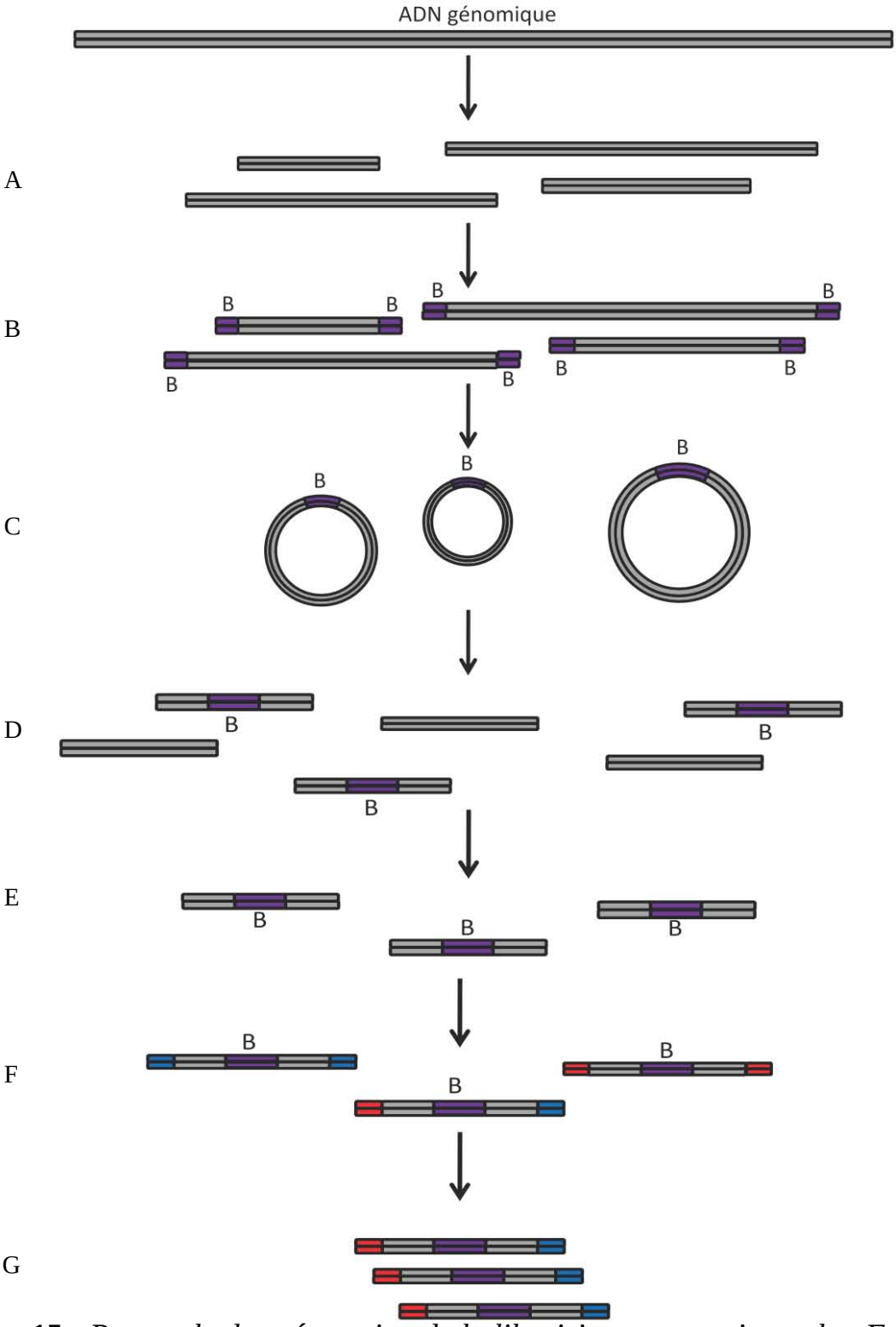


Figure 15 : *Protocole de préparation de la librairie « mate-pair » selon Fasteris.*
A. L'ADN génomique est fragmenté par nébulisation. B. Les fragments d'une taille comprise entre 2 et 4 kb sont sélectionnés, ce qui correspond à la distance séparant les deux lectures séquencées. Des séquences adaptatrices biotinylées sont ajoutées aux extrémités des fragments. C. Les adaptateurs vont permettre la circularisation du fragment. D. Une seconde nébulisation permet d'obtenir des fragments d'environ 400 pb. E. Les fragments biotinylés sont sélectionnés. F. Des séquences adaptatrices sont ajoutées aux extrémités des fragments. G. Les fragments possédant les bons adaptateurs sont ensuite sélectivement enrichis par PCR.

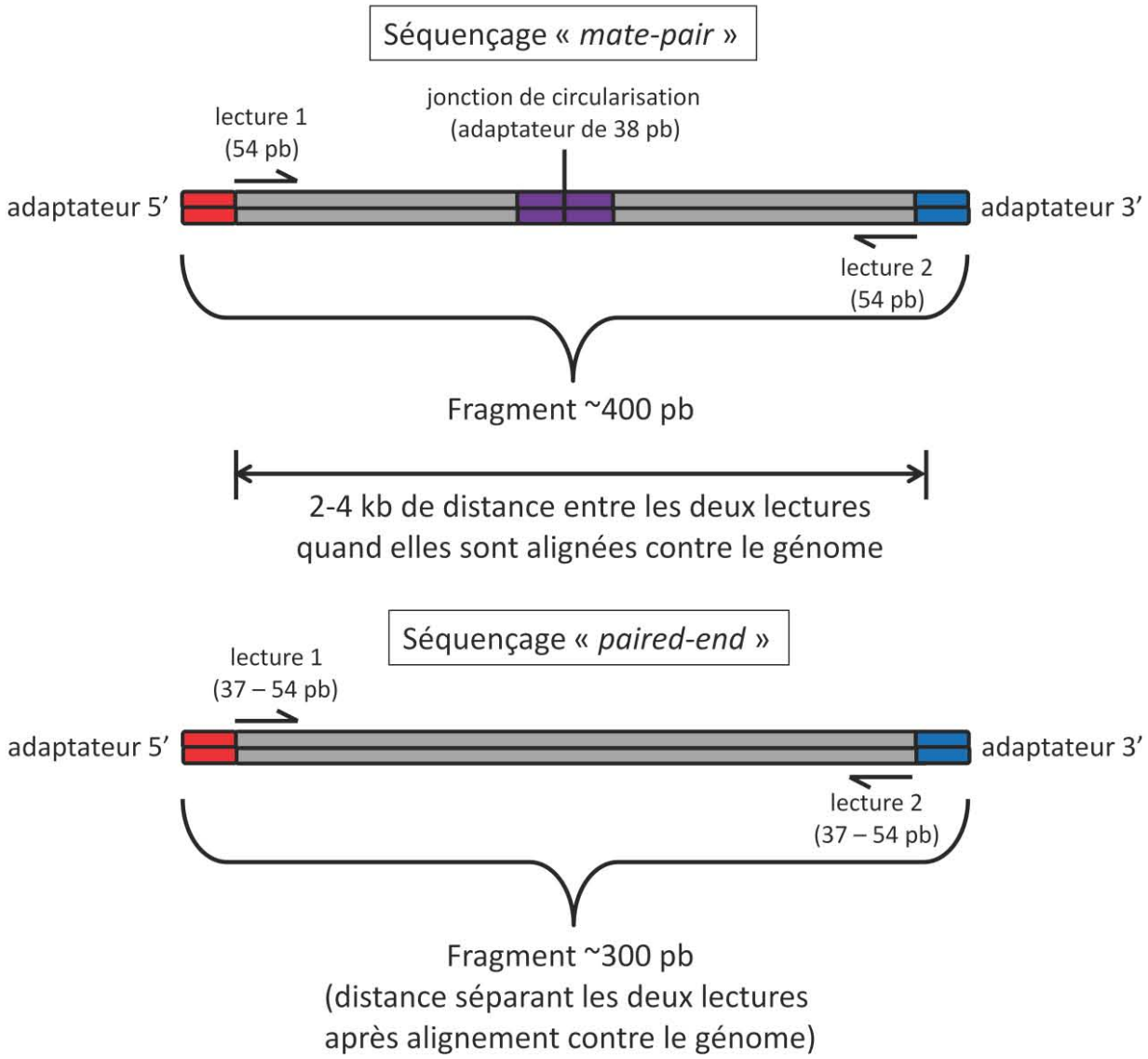


Figure 16 : *Principe du séquençage des fragments obtenus par les protocoles « paired-end » et « mate-pair ».*

Pour chaque fragments séquencés, deux lectures sont obtenues dont la distance et l'orientation sont renseignées. Cette distance est dépendante de la taille des fragments sélectionnés lors de la préparation de la banque. Pour le « *paired-end* », deux banques avec des lectures de taille différentes ont été obtenues (de 54 et 37 pb).

2°) Analyse informatique des données de séquençage

2.1°) Moyens informatiques utilisés

L'ensemble des manipulations a été faite sur un serveur hébergé par le LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications). Les langages de programmation PERL (Practical Extraction and Report Language), PHP (Hypertext PrepProcessor) et HTML (HyperText Markup Language) ont été utilisés pour le développement des programmes informatiques.

2.2°) Estimation de la qualité globale des lectures

A chaque nucléotide des lectures de pyroséquençage et Solexa/Illumina est associé un score de qualité. Pour le pyroséquençage haut-débit, ce score reflète la probabilité qu'un nucléotide soit une erreur de séquençage due à un homopolymère. En ce qui concerne la technique Solexa/Illumina, ce score est dépendant de la qualité du signal détecté en fonction du bruit de fluorescence. Dans les deux cas, un score élevé indique une probabilité plus faible que le séquenceur se soit trompé.

Dans le cas du pyroséquençage haut-débit, la composition de la séquence, la présence d'homopolymères, va directement influer sur la qualité des lectures. Ainsi, il semble peu pertinent de se baser sur ce critère pour estimer la qualité d'un pyroséquençage. Ainsi, des critères de filtration dont l'efficacité a été démontrée et publiée sont utilisés par la suite.

Pour la technique Solexa/Illumina, un programme informatique, *check_score.pl*, a été écrit afin de calculer à chaque position des lectures la qualité moyenne des nucléotides, ainsi que l'écart-type de cette moyenne.

2.3°) Filtration des lectures de pyroséquençage

Les données issues du pyroséquençage ont été envoyés dans un fichier au format SFF (Standard Flowgram Files). Le programme informatique *sff_extract.py* a été utilisé pour en tirer les séquences nucléotidiques au format FASTA ainsi qu'un fichier contenant les scores de qualité. Lors de l'extraction du fichier, les séquences des adaptateurs utilisés ont été automatiquement supprimées des lectures.

Les critères de filtration des lectures de pyroséquençage sont issus de l'étude de Huse *et al* (2007). Dans un premier temps, les lectures possédant des nucléotides non-identifiés (représentés par un N dans la séquence) n'ont pas été considérées. Par la suite, les lectures dont la taille est inférieure à 50 pb ou supérieures à 360 pb ont été supprimées. Il a été démontré que les lectures dont la longueur est trop faible ou trop élevée ont un nombre plus important d'erreurs. Les seuils de filtration ont été choisis de manière à conserver les lectures de meilleures qualités tout en limitant la perte d'information.

2.5°) Vérification de la couverture des reads sur une séquence de référence

Le programme MOSAIK a été utilisé pour aligner les lectures contre une séquence de référence. Un de ses sous-programmes, MosaikCoverage, est utilisé pour déduire de cet alignement les biais représentationnels dans la séquence de référence. Ces programmes ont été utilisés avec les paramètres optimaux suggérés. Un fichier est créé, contenant pour chaque position de la référence, le nombre de fois où ce nucléotide est représenté dans les lectures alignées. Ce fichier est ensuite utilisé par le script *plot_cov.plot* afin de permettre la représentation graphique par Gnuplot (logiciel permettant la création de graphiques) du résultat.

3°) Assemblage des séquences

3.1°) Assembleurs utilisés

Deux assembleurs ont été utilisés selon les séquences qui sont assemblées. Pour l'assemblage des lectures issues du pyroséquençage, MIRA (Chevreux *et al.* 2004) a été utilisé. Cet assembleur est particulièrement réputé pour son usage lors de projets de séquençage de génomes possédant beaucoup de répétitions. De plus, de nombreux paramètres peuvent être modifiés afin d'optimiser au maximum l'assemblage.
Pour l'assemblage de la souche ATCC23877, la commande suivante a été utilisée :
mira -projetct AllATCC_assembly –job=denovo,genome,accurate,454 –DP :ure=yes –ED :ace=yes.

Les lectures produites par la technique Solexa/Illuminaont été assemblées le programme Velvet (Zerbino et Birney, 2008). Cet assembleur, qui fonctionne selon la méthode de « de Bruijn », a démontré son efficacité dans de nombreux projets de séquençage de génomes bactériens. Le script *VelvetOptimizzer.pl* fourni avec l'assembleur a été utilisé avec la commande :
perl VelvetOptimizer.pl –s 23 –e 35 –t 1 –f « varie selon les assemblages *» -o «* varie selon les assemblages *».*

4°) Mise au point d'une plateforme d'annotation de séquence

Une plate-forme d'annotation a été développée dans le but d'automatiser au maximum la manipulation des données, depuis la détection des séquences codantes (CDS) jusqu'à l'identification de domaines fonctionnels dans les protéines prédites. L'architecture de ce pipeline est très largement inspiré du module d'annotation SAMANNOT mis au point par Frédéric Choulet au cours de sa thèse (ref). Cependant, l'utilisation de ce module demande à l'utilisateur d'être familier avec l'environnement LINUX et l'invite de commande. Ainsi, pour permettre son utilisation par un large public, une interface graphique a été développée en PHP.

La plate-forme d'annotation se divise en quatre étapes : l'annotation syntaxique primaire, l'annotation syntaxique secondaire, la classification et l'annotation fonctionnelle des protéines et enfin la suppression des annotations en cours (**figure 17**). La division de la plate-forme d'annotation permet à l'utilisateur de vérifier et de modifier le résultat de chaque étape avant de passer à la suivante.

4.1°) Annotation syntaxique primaire

Un premier script, *rawseq2embl.pl*, identifie les séquences codantes en utilisant Glimmer2, un programme de prédiction de gènes basé sur des modèles de Markov interpolés (Delcher *et al.*, 1999). L'apprentissage de Glimmer2 a été réalisé à partir de 3000 ORF de *S. coelicolor*. Un seuil minimal classique de 120 nucléotides a été fixé arbitrairement. Les résultats obtenus sont automatiquement redirigés vers un autre programme, RBSfinder, dans le but de corriger les positions des codons d'initiation de la traduction en fonction de la présence ou non d'un RBS potentiel (Suzek *et al.*, 2001). Le programme *rawseq2embl.pl* formate les résultats sous forme d'un fichier EMBL contenant les positions corrigées des CDS potentielles.

A partir de ce fichier EMBL, un second programme intégré dans le pipeline d'annotation, *BlastFromEmbl.pl*, traduit les séquences nucléiques de chaque CDS en séquence protéique et recherche pour chacune d'elles des similarités de séquences à l'aide du programme d'alignement BLASTP (Altschul *et al.*, 1997). Tout d'abord, la requête est effectuée contre la banque généraliste NR (Non Redundant) du National Centre for Biotechnology Information (NCBI) qui regroupe toutes les séquences protéiques disponibles. De par le grand nombre de séquences à aligner, l'étape

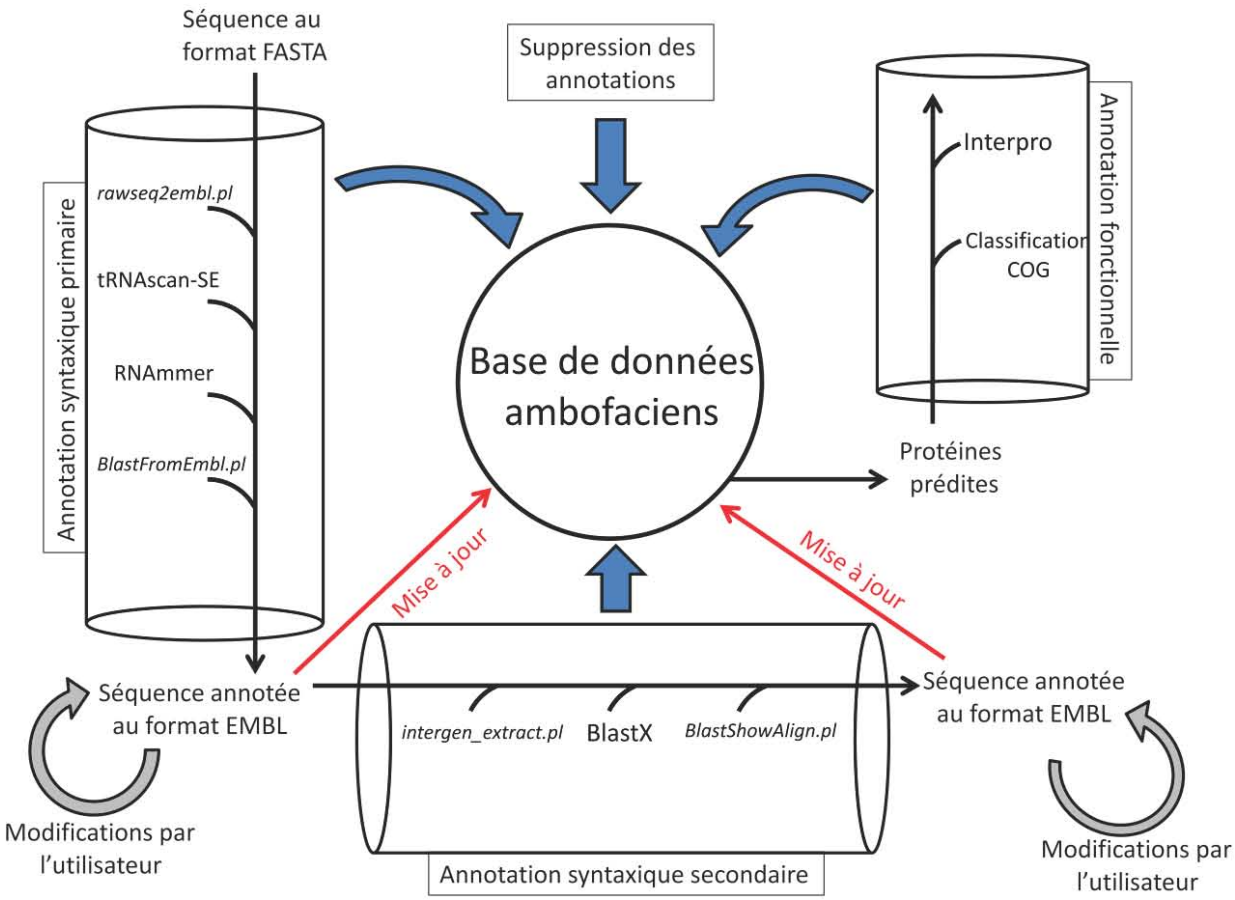


Figure 17 : *Schématisation du fonctionnement de la plate-forme d'annotation.*
L'annotation se divise en quatre modules qui interagissent avec la base de données ambofaciens pour stocker ou supprimer les résultats obtenus. A la fin des modules d'annotation syntaxique primaire et secondaire, les données trouvées sont compilées dans un fichier au format EMBL (Europen Molecular Biology Laboratory) éditable par l'utilisateur. La base de données ambofaciens peut être mise à jour à l'aide des données modifiées.

de BLAST peut prendre plusieurs jours pour être terminée sur le serveur utilisé pour les annotations. Pour réduire de temps nécessaire aux alignements, quatre BLASTP sont lancés en parallèle par « *multi-threading* ». Le script *writenote.pl* permet d'extraire les informations utiles (ex : description, score, numéro d'accession …) de chaque séquence montrant une similarité significative avec la protéine prédite et de les écrire dans le fichier EMBL. Ces informations facilitent la validation de la CDS prédite par l'annotateur.

La plate-forme inclut aussi des outils permettant de détecter des ARN non codants. Le programme tRNAscan-SE (ref) permet d'identifier dans une séquence les ARN de transfert tandis que RNAMMER (ref) permet la détection des ARN ribosomiques.

4.2°) Annotation syntaxique secondaire et recherche de RBS

Les séquences intergéniques sont extraites du génome en cours d'annotation, par le script *intergen-extract.pl*, pour rechercher des gènes fonctionnels non-prédits par Glimmer2 ou des pseudogènes. A partir des prédictions de l'annotation syntaxique primaire, les régions entre chaque gènes identifiés faisant plus de 20 pb sont extraites. Les séquences codantes potentielles sont recherchées par BlastX (traduction dans les six phases de lecture pour chaque région intergénique) contre la banque NR. Un script, *BlastShowAlign.pl*, a été écrit pour détecter automatiquement les similarités significatives (plus de 40% d'identité de séquence) et générer un fichier au format EMBL contenant les prédictions.

Une fois les annotations automatiques terminées, le programme *Find_RBS.pl* est utilisé pour détecter un RBS potentiel en amont de chaque CDS prédites. Ce programme recherche la présence d'au moins 4 des 5 nucléotides GGAGG constituant le consensus retrouvé chez les *Streptomyces* entre 4 et 13 pb en amont du codon d'initiation de la traduction (Strohl, 1992).

4.3°) Vérification des annotations

Malgré l'efficacité des programmes Glimmer2, RBSfinder et BlastX, l'annotation automatique génère des erreurs de prédiction. La nécessité d'une étape "manuelle", c'est-à-dire d'une intervention de l'annotateur après le processus automatique, est indispensable à l'obtention d'une annotation la plus proche possible de la réalité biologique. La validation manuelle des CDS a été réalisée grâce à Artemis (Rutherford *et al.*, 2000).

Deux paramètres doivent être contrôlés : la validité de chaque CDS prédite en tant que gène potentiel et la position du codon d'initiation de la traduction. L'annotation du génome des *Streptomyces* est rendue difficile du fait de la pauvreté des séquences en codons stop (riches en A/T) dans les différentes phases de lecture. Cette caractéristique résulte du pourcentage très élevé des bases G+C. Ainsi, le nombre de phases ouvertes de lecture de taille compatible avec celle d'un gène codant une protéine est si grand qu'il n'est pas possible de s'en servir comme d'un critère de détection, contrairement aux génomes à faible pourcentage en bases G+C (**figure 18**).

Deux critères ont été utilisés pour valider une CDS. Le premier est le biais très prononcé en base G+C en troisième position de codon des séquences codantes (aussi appelé GC3) (**figure 19**). Cette particularité propre aux génomes à haut pourcentage en bases G/C permet d'éliminer la plupart des erreurs de prédictions. Le deuxième critère est la présence d'homologues identifiés par BLASTP dans les banques de séquences.

4.4°) Classification et annotation fonctionnelle des protéines prédites

La base de données COG (Cluster of Orthologous Groups, (Tatusov *et al.*, 2003)) a été utilisée afin d'affecter une catégorie fonctionnelle à chaque protéine. COG définit 25 catégories

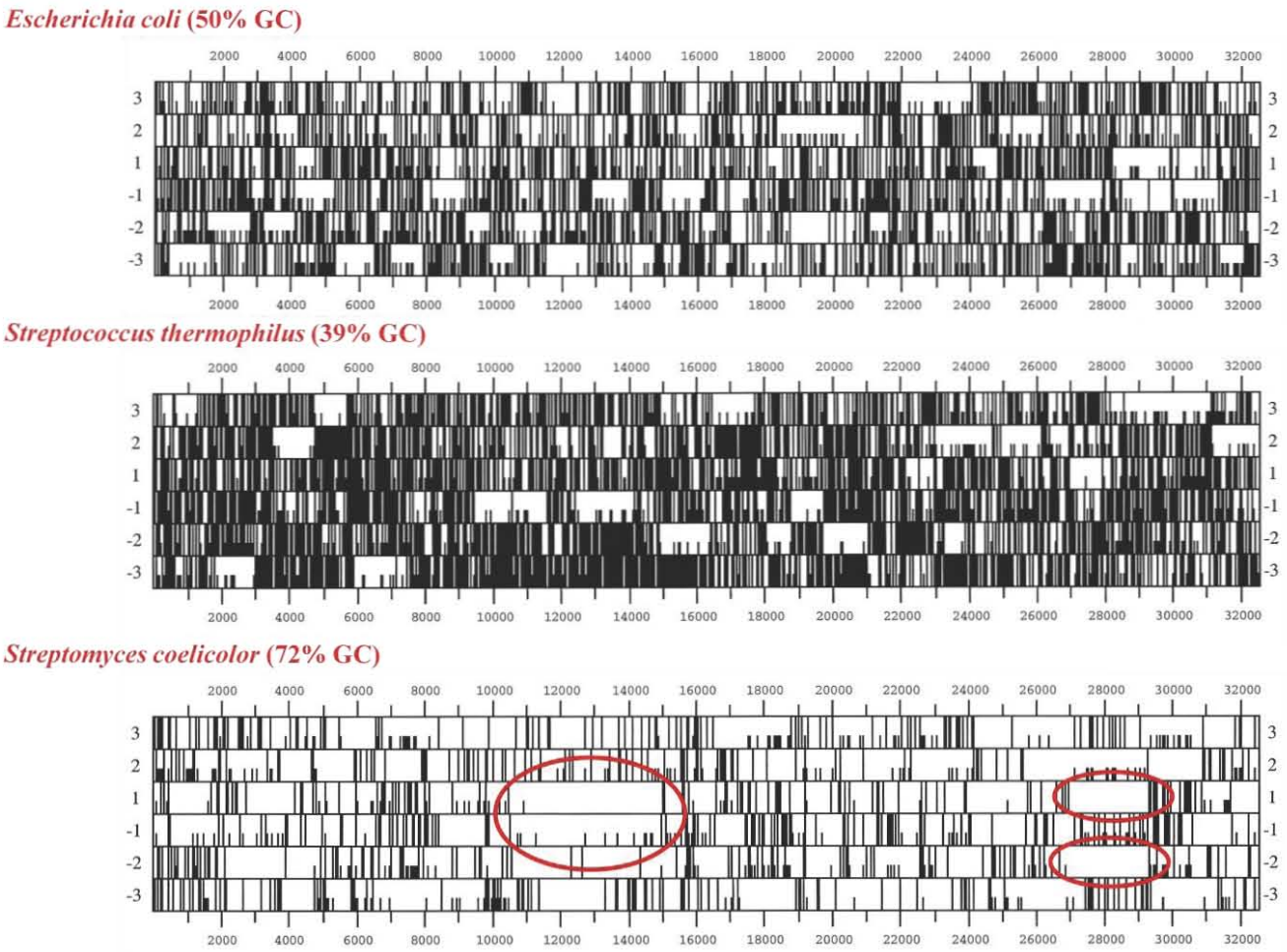


Figure 18 : *Comparaison sous DNAstrider de la densité en codons stop selon le pourcentage en G/C.*

Les traits noirs représentent les codons stop dans les 6 phases de lectures et les zones blanches des phases ouvertes de lecture. Plus le pourcentage en G/C augmente, plus il difficile de distinguer les phases ouvertes de lectures correspondant à des séquences codantes (deux exemples sont compris dans les cercles rouges pour *Streptomyces coelicolor*).

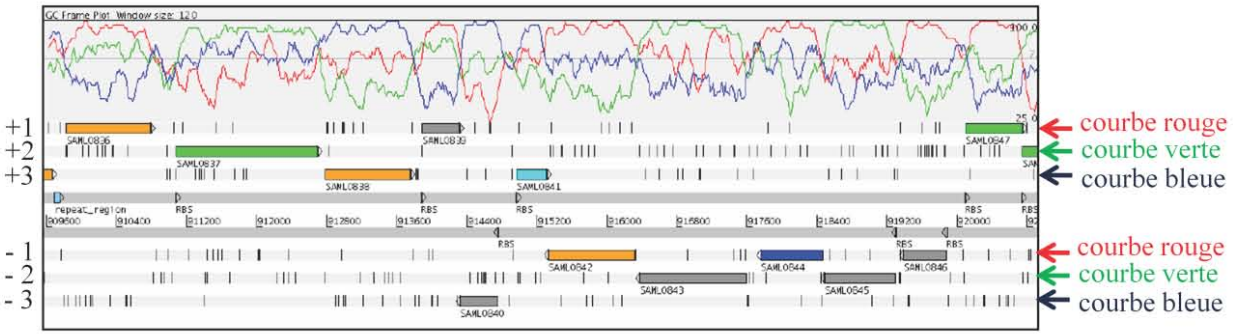


Figure 19 : *Profil de GC3 chez Streptomyces ambofaciens ATCC23877.*
Le pourcentage en bases G et C est calculé dans les 6 phases de lecture (indiquées sur la gauche) dans une fenêtre glissante, chaque phase étant représentée par une courbe de couleur distincte (indiquée à droite). Etant donné le fort biais en bases G et C (92%) à la troisième base des codons chez *S. ambofacien*s, ce signal est fréquemment utilisé pour prédire la présence et les limites des CDS dans les génomes à haut pourcentage en G+C.

regroupées en 4 grandes familles de fonctions. Elle compte 138 458 séquences protéiques regroupées en 4873 COG, eux mêmes répartis dans les 18 catégories fonctionnelles.

Ces séquences ont été formatées en base de données BLAST afin de réaliser une recherche de similarités de chaque protéine prédite chez *S. ambofaciens*. Le script *cogblast2sql.pl* permet un traitement automatique de ces résultats. Il parcourt les fichiers BLASTP et affecte à chaque protéine en cours d'annotation la catégorie fonctionnelle à laquelle appartient l'homologue le plus proche (avec plus de 30% d'identité recouvrant au moins 70% de la taille de la protéine).

Pour l'annotation fonctionnelle, de nombreuses banques de domaines protéiques sont disponibles et il n'existe en réalité aucun critère objectif pour en privilégier certaines dans la mesure où chacune d'elles possède ses propres caractéristiques. Afin de préciser au maximum la fonction des gènes prédits par le pipeline d'annotation, les banques de données protéiques regroupées dans le projet collaboratif InterPro ont été interrogées via le programme InterProScan v27.0 (Zbodnov et Apweiler, 2001). Les bases de données et méthodes de détection suivantes ont été utilisées :

- PROSITE pattern : Recherche de domaines protéiques à partir de motifs déjà connus contenus dans la base de données PROSITE.
- PROSITE profile : Recherche des domaines protéiques en se basant sur des matrices de poids.
- PRODOM : regroupe des familles de domaines protéiques classifiées automatiquement par analyse des banques SWISS-PROT et TrEMBL.
- PRINTS : Collection d'"empreintes" spécifiques de familles de protéines, chaque empreinte étant constituée de plusieurs motifs conservés.
- PFAM : collection d'alignements de séquences de familles de protéines et de domaines. Tous les alignements utilisent les séquences présentes dans SWISS-PROT et TrEMBL. Un modèle de Markov est généré pour chaque alignement et l'interrogation de nouvelles séquences se fait via HMMER2.2 (Eddy, 1998).
- SMART : Banque dédiée aux domaines "mobiles", c'est-à-dire retrouvés en association avec d'autres dans des protéines contenant de multiples domaines. Elle fonctionne également via des modèles de Markov.
- TIGRFAM : Banque de données de domaines protéiques. Elle est focalisée sur une classification des protéines par groupes présentant une fonction identique (appelés "équivalogues") et non uniquement selon la parenté phylogénétique ("orthologues").
- SUPERFAMILY : Collection de modèles de Markov créés à partir de protéines de structures connues.
- TMHMM : Programme de prédiction de domaines transmembranaires également basé sur des modèles de Markov

4.5°) Conservation et classification des données

Dans la mesure où les analyses de séquences génèrent une quantité difficilement gérable de fichiers de résultats, les contraintes de stockage et d'accès aux données se sont posées. Dans le but de structurer les données de façon ordonnée, l'utilisation de bases de données relationnelles s'est imposée comme la solution la plus efficace. PostgreSQL est un système de gestion de bases de données relationnelles libre, très performant et gratuit. La base de données « ambofaciens », créée par Birama Ndiaye (ingénieur au LORIA), a été utilisée pour le stockage et la manipulation des annotations créées.

La plate-forme d'annotation intègre de nombreux scripts permettant l'insertion des résultats des programmes de recherche décrits précédemment dans la base de données « ambofaciens ».

RESULTATS

1°) Stratégie d'assemblage des séquences

Pour obtenir la séquence génomique de la souche ATCC23877 de *Streptomyces ambofaciens*, une stratégie de séquençage mixte, par pyroséquençage et par terminaison cyclique réversible de Solexa/Illumina, a été mise en place. L'utilisation de ces deux méthodes de séquençage permet de compenser leurs limites respectives : la méthode de terminaison cyclique réversible n'est pas sensible aux homopolymères tandis que le pyroséquençage produit des lectures plus longue que celle obtenues avec la méthode Solexa/Illumina.

Pour la souche de *Streptomyces ambofaciens* ATCC23877, trois banques de séquences Solexa/Illumina de type « *paired-end* » (une avec des lectures de 36 pb et une autre avec des lectures de 54 pb) et « *mate-pair* » ont été reçues au cours de ces travaux de la part du prestataire Fasteris (cf : matériel et méthodes). Auparavant, d'autres données de séquençage Solexa/Illumina et de pyroséquençage haut-débit ont été commandées auprès d'un autre prestataire GATC (**figure 20**). Les banques « *paired-end* » et « *mate-pair* » sont spécifiquement créées pour aider à résoudre les problèmes dus aux répétitions. La distance et l'orientation de chaque paire de lectures étant connues, l'assembleur doit respecter ces deux informations lors de l'assemblage, ce qui en améliore la qualité. Pour la souche DSM40697, des lectures de pyroséquençage haut-débit ont été reçues de la part du prestataire GATC (**figure 20**).

Filtrer les lectures d'un séquençage est une étape indispensable dans l'objectif d'obtenir un assemblage cohérent. La présence d'un trop grand nombre d'erreurs dans les séquences peut conduire à la formation de contigs chimériques et l'obtention d'un mauvais assemblage. Pour les lectures obtenues par la méthode Solexa/Illumina, la filtration des lectures s'est faite en se basant sur les scores de qualité[1] attribués à chaque base par le séquenceur. Ainsi, les séquences possédant des scores élevés ont été conservées pour l'assemblage. L'analyse de la couverture[2] des lectures le long de séquences de référence a ensuite aussi été utilisée pour estimer la proportion du génome couvert (**figure 20**). Les lectures de pyroséquençage ont été filtrées selon différents critères établis par Huse *et al.* (2007). Le score de qualité attribué au cours du pyroséquençage correspond à la probabilité que la base séquencée soit erronée à cause d'un homopolymère. Ainsi, des lectures possédant de nombreux homopolymères verront leur qualité diminuer fortement même si elles ne contiennent aucune erreur. C'est pourquoi le score de qualité n'est pas pris en compte pour filtrer ou pour vérifier ces lectures.

Une fois les filtrations effectuées, différents assemblages ont été lancés dans le but d'obtenir le génome de *Streptomyces ambofaciens* ATCC23877 à l'aide de la stratégie de séquençage mixte envisagée. Un assemblage à partir des données de pyroséquençage de la souche DSM40697 a aussi été tenté pour essayer d'en obtenir la séquence génomique.

2°) Analyse des données de séquençage

2.1°) Filtration des lectures Solexa/Illumina

Pour filtrer les lectures Solexa/Illumina, quatre critères ont été utilisés :

- Si plus de deux nucléotides dans une lecture possèdent un score de qualité inférieur à 10, cette lecture n'est pas considérée. La rigueur de ce filtre permet de ne conserver que les séquences de haute qualité.
- Si il y présence d'un nucléotide non-identifié dans la séquence (un N), la lecture est supprimée. L'assembleur utilisé (VELVET) remplace arbitrairement les N dans les lectures par des A, ce qui pourrait provoquer des erreurs d'assemblage.

[1] : le score de qualité représente la probabilité que le séquenceur se soit trompé de base à la position considérée. Plus ce score est élevé et plus cette probabilité est faible.

[2] : La couverture d'une séquence est le nombre moyen de fois qu'une base de du génome est représentée dans la séquence des lectures.

Figure 20 : *Résumé du traitement des lectures effectué au cours de ces travaux.*
Les banques Solexa/Illumina utilisées au cours de ces travaux sont : « *single-end* » (appelée SE), « *paired-end* » avec des lectures de 54 pb (appelée PE 54), « *paired-end* » avec des lectures de 36 pb (appelée PE 36) et « *mate-pair* » (appellée MP).

- Si il y a plus un même nucléotide dans plus de 90% de la lecture, celle-ci est retirée. La séquence de ces lectures est peu informative et leur présence risque de perturber l'assemblage.
- Si le score moyen de la lecture est inférieur à 20, la lecture est éliminée. De façon générale, il n'a jamais été formellement prouvé que ce dernier critère de filtration pouvait améliorer de façon notable la qualité des lectures utilisées pour un assemblage. Toutefois, compte tenu de la grande quantité de séquences reçues du prestataire Fasteris des assemblages sur le matériel informatique utilisé aurait été compliquée du fait de la trop grande quantité de mémoire demandée. Ce critère de filtration a donc permis de diminuer la quantité de séquences sans pour autant diminuer la qualité globale des lectures utilisées lors de l'assemblage.

Les résultats de filtration sont résumés dans le **tableau 3**. Parmi toutes les lectures « *single-end* » considérées, aucune n'a réussi à passer les filtres, démontrant ainsi la faible qualité des lectures obtenues. Ces données sont plus anciennes que les autres (2008) et ont été obtenues sur un séquenceur d'ancienne génération (Illumina Genome Analyzer I). Cela pourrait expliquer la plus faible qualité des lectures, comparativement aux autres librairies qui ont été obtenues avec du matériel plus récent.

2.2°) Conséquences de la filtration des lectures Solexa/Illumina

Pour vérifier si ces filtres ont été efficaces, le score moyen, ainsi que l'écart-type de cette moyenne, ont été calculés pour chaque position de la séquence. De cette façon, il est possible d'apprécier et de comparer la qualité d'un grand nombre de lectures. La **figure 21** montre les résultats comparatifs entre les données filtrées et non filtrées.

Pour l'ensemble des lectures analysées, le score moyen diminue à l'extrémité 3' de la séquence. Cela traduit globalement une diminution de la fiabilité du séquençage en fin d'extension. Toutefois, après filtration des banques « *mate-pair* » et « *paired-end* » dont les lectures font 37 pb , les scores moyens sont plus élevés à l'extrémité 3' des lectures que ces mêmes extrémités avant filtration. De plus, les écart-types calculés à ces positions sont plus faibles. Ces deux observations montrent une amélioration notable de la qualité des lectures qui ont passé les filtres. Pour la banque « *single-end* », le score diminue fortement dès la 10ème position pour atteindre un score moyen d'environ 5 à la dernière position alors que, par exemple, pour la banque « *paired-end* » dont les lectures sont de longueur similaire, le score moyen minimal est d'environ 25 sans filtration. Cela confirme la conclusion faite précédemment sur la mauvaise qualité globale des lectures de la banque « *single-end* ». La dernière banque « *paired-end* » dont les lectures mesurent 54 pb a été préalablement filtrée par le fournisseur, ce qui explique la bonne qualité des séquences obtenues.

2.3°) Taux de couverture génomique des lectures Solexa/Illumina

Pour estimer la couverture des lectures obtenues, les séquences ont été alignées contre la séquence des bras chromosomiques déjà séquencés de la souche ATCC23877. Les programmes MosaikAligner et MosaikCoverage de l'ensemble MOSAIK ont été utilisés pour en déduire la couverture pour chaque base des bras droit et gauche (**tableau 4**).

Pour l'ensemble des lectures « *paired-end* », les bras gauche et droit sont respectivement couverts à 100% et 99,9% avec une couverture moyenne estimée de 127 X, c'est-à-dire que chaque base est statistiquement séquencée environ 127 fois.

Pour les lectures « *mate-pair* », seulement 9,7% du bras gauche et 9,9% du bras droit ont été séquencés dans cette banque malgré les 140 X de couverture estimée. Une représentation graphique de cette couverture est présentée en **figure 22**. Sur ces deux séquences de référence, des zones de couverture très élevée d'une cinquantaine de paire de base succèdent à de larges zones non couvertes. Cette distribution anormale des lectures explique les chiffres obtenus par MosaikCoverage : une faible proportion des bras a été séquencée mais à haute couverture.

Les lectures « *paired-end* » recouvrant la quasi-totalité des bras, il est possible d'estimer que 99,9% du génome de la souche ATCC23877 est représenté dans cette banque de séquences. Pour

Banques de lectures	SE	PE 36	MP
Nombre de lectures initiales	6 361 402	27 129 690	57 310 692
Nombre de lectures après filtration	0	23 414 266	25 735 132
Pourcentage de lectures ayant passé le filtre	0 %	86 %	45 %

Tableau 3 : *Récapitulatif des résultats de filtration des lectures Solexa/Illumina de Streptomyces ambofaciens ATCC23877.*

La banque « *paired-end* » avec lectures de 54 pb n'a pas été considérée pour ce test étant donné que les lectures ont été filtrées au préalable par le prestataire.(SE : banque « *single-end* ». PE 36 : banque « *paired-end* » avec lectures de 36 pb. MP : banque « *mate-pair* »).

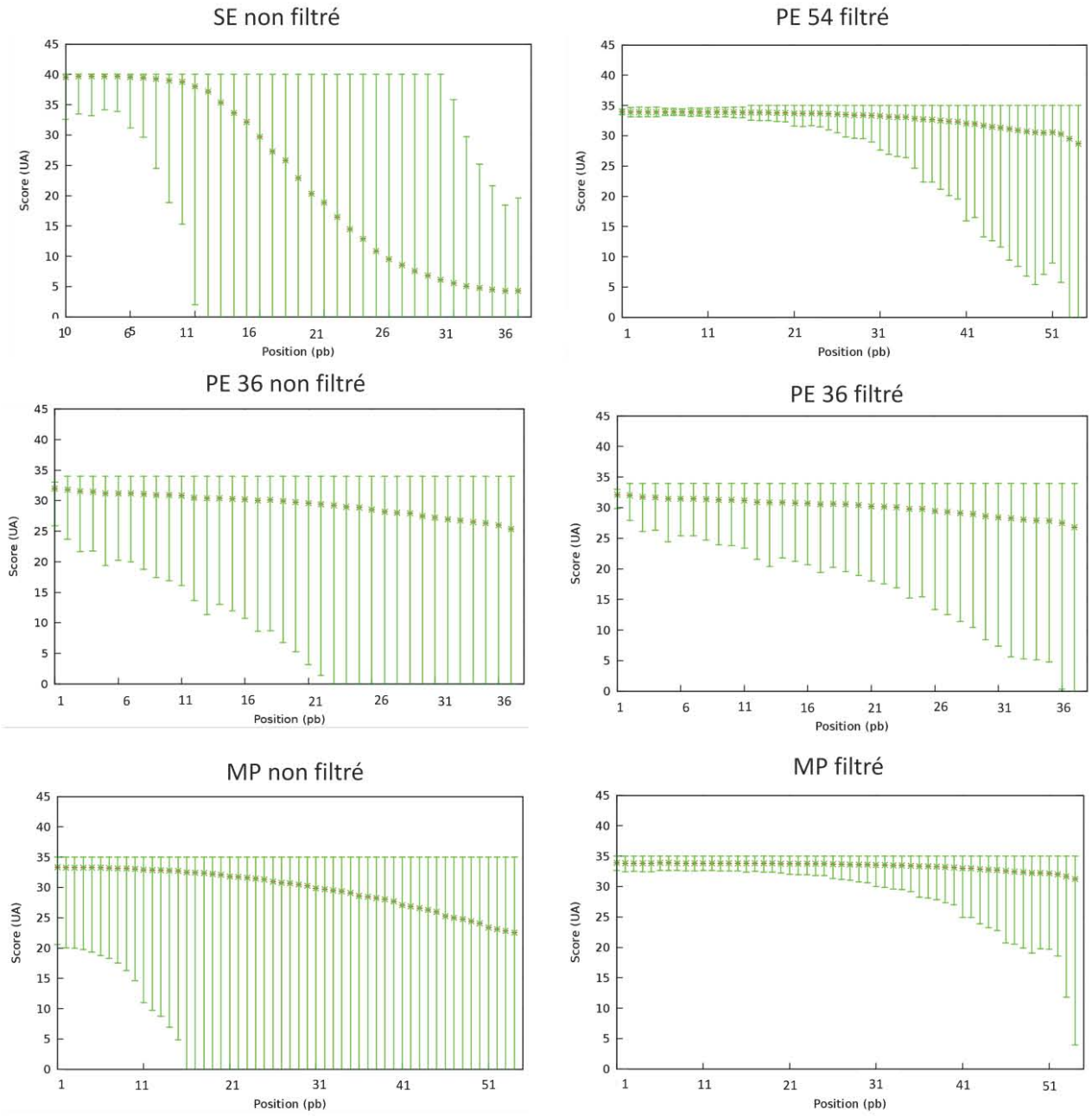


Figure 21 : *Graphiques représentant la qualité moyenne des lectures Solexa/Illumina.*

A chaque position des lectures, la qualité moyenne a été calculée, ainsi que l'écart-type de cette moyenne. (SE : banque « *single-end* », les données non filtrées sont présentées car aucune lecture n'a passé la filtration. PE 54 : banque « *paired-end* » avec lectures de 54 pb. Les données ont été fournies filtrées par le prestataire. PE 36 : banque « *paired-end* » avec lectures de 36 pb. MP : banque « *mate-pair* »).

Séquences de référence	Bras droit (1 367 119 pb)			Bras gauche (1 544 032 pb)		
Banques	PE 54	PE 37	MP	PE 54	PE 37	MP
Nombre de lectures collectées	3 161 340	23 414 266	25 735 132	3 161 340	23 414 266	25 735 132
Nombre de lectures alignées	492 344	3 838 256	3 245 938	634 356	4 573 588	3 885 478
Pourcentage de lectures alignées	15,6 %	16,4%	12,6 %	20,1%	19,5%	15,1 %
Couverture estimée	19,8 X	105,6 X	134,8 X	21,7 X	107,2 X	138,5 X
Pourcentage de la séquence référence couverte	97,9%	99,9%	9,9 %	97,8%	100%	9,6 %

Tableau 4 : *Récapitulatif des alignements des lectures « paired-end » contre les bras chromosomiques de Streptomyces ambofaciens.*

(PE 54 : banque « *paired-end* » avec des lectures de 54 pb. PE 37 : banque « *paired-end* » avec des lectures de 37 pb. MP : banque « *mate-pair* »).

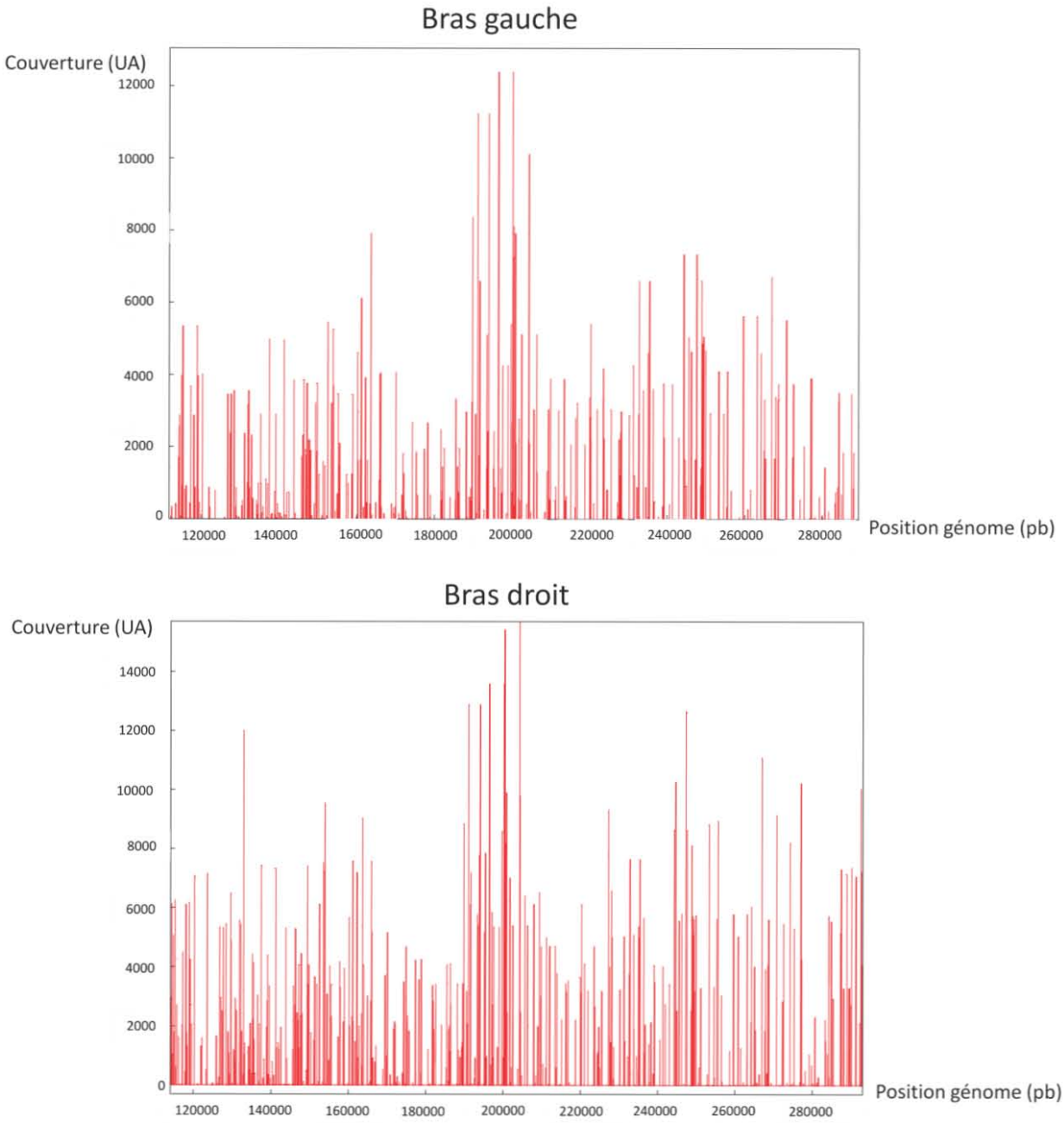


Figure 22 : *Graphiques représentant la couverture en fonction de la position génomique pour une partie des bras gauche et droit de la souche Streptomyces ambofaciens ATCC23877.*

La couverture est calculée à chaque position du génome (ou encore chaque base). Ici, une partie des bras chromosomiques est montrée pour que le résultat soit bien visible mais ce type de couverture se retrouve sur l'ensemble des bras.

les lectures « *mate-pair* », la couverture anormale montre que la banque créée ne contiendrait qu'une faible proportion du génome séquencé. Ainsi, les séquences « *mate-pair* » pourraient ne contribuer que faiblement à la progression de l'assemblage.

3°) Assemblage des séquences

3.1°) Estimation de la taille du génome de *Streptomyces ambofaciens* ATCC23877

A partir des résultats de la couverture des bras chromosomiques, il est possible d'en déduire la taille du génome attendu. Pour les lectures « *paired-end* », la couverture est estimée à 127X sachant que 1 014 625 936 pb sont dénombrées dans l'alignement. La taille du génome peut être approchée en divisant la longueur cumulée de toutes les lectures par le taux de couverture estimé (ici sur les séquences test des bras chromosomiques).

De cette façon, la longueur du génome de la souche ATCC23877 peut être estimée à environ 7 990 000 pb. Toutefois, bien que la couverture soit globalement uniforme le long des bras, certaines zones présentent un taux de couverture hautement élevé (**figure 23**). Pour le bras gauche, un pic de plus de 1900X est observé au niveau d'une séquence d'ARN ribosomique, probablement parceque ces loci sont répétés 6 fois dans le génome de *Streptomyces ambofaciens*. Pour le bras droit, des pics de haute couverture sont observés, notamment au niveau d'un groupe de gènes (cluster), codant une poly-cétone-synthétase (Pang *et al.* 2004), où de nombreuses répétitions sont trouvées. Ainsi, ces répétitions augmenteraient artificiellement le nombre de lectures alignées et ainsi, le taux de couverture sur ce cluster.

Ces régions de haute couverture pourraient faire augmenter la couverture moyenne qui a été estimée et ainsi fausser l'estimation qui en résulte. La taille du génome attendue pourrait être comprise entre 8 et 8,3 Mb.

3.2°) Assemblage des lectures de *Streptomyces ambofaciens* ATCC23877

3.2.1°) Assemblage des lectures du pyroséquençage

Les assemblages des lectures de pyroséquençage pour la souche ATCC23877 ont été faits par deux assembleurs, Newbler (utilisé par le prestataire de séquençage GATC) et MIRA (utilisé au laboratoire). Les résultats des deux assemblages sont présentés dans le **tableau 5**.

Le nombre de contigs obtenus par l'assembleur Newbler est de 2739, ce qui est supérieur au nombre de contigs créés par MIRA, 2620. Par contre, pour les contigs ayant une taille supérieure à 1 kb, qui est le seuil arbitraire de sélection d'un contig de grande taille, Newbler produit moins de contigs que MIRA avec respectivement 793 contigs contre 1345. Le N50[1] obtenu pour l'assemblage Newbler est de 13 178 pb, signifiant que la moitié de l'assemblage final de Newbler se trouve dans des contigs ayant une taille supérieure à 13 178 pb, contre 7 483 pb pour MIRA. D'après ces résultats, les contigs formés par Newbler sont globalement de plus grande taille.

Les deux assemblages obtenus par chaque assembleur, sont de qualité équivalente. Ainsi, les lectures de pyroséquençage semblent insuffisantes à elles seules pour obtenir un assemblage correct du génome de la souche ATCC23877. C'est pourquoi des lectures obtenues à partir d'un séquençage Solexa/Illumina ont été utilisées pour compléter ces données.

[1] : Le N50 est la taille du plus petit contig tel que 50 % de la longueur cumulée de l'ensemble des contigs obtenus après assemblage soit contenue dans des contigs de taille égale ou supérieure.

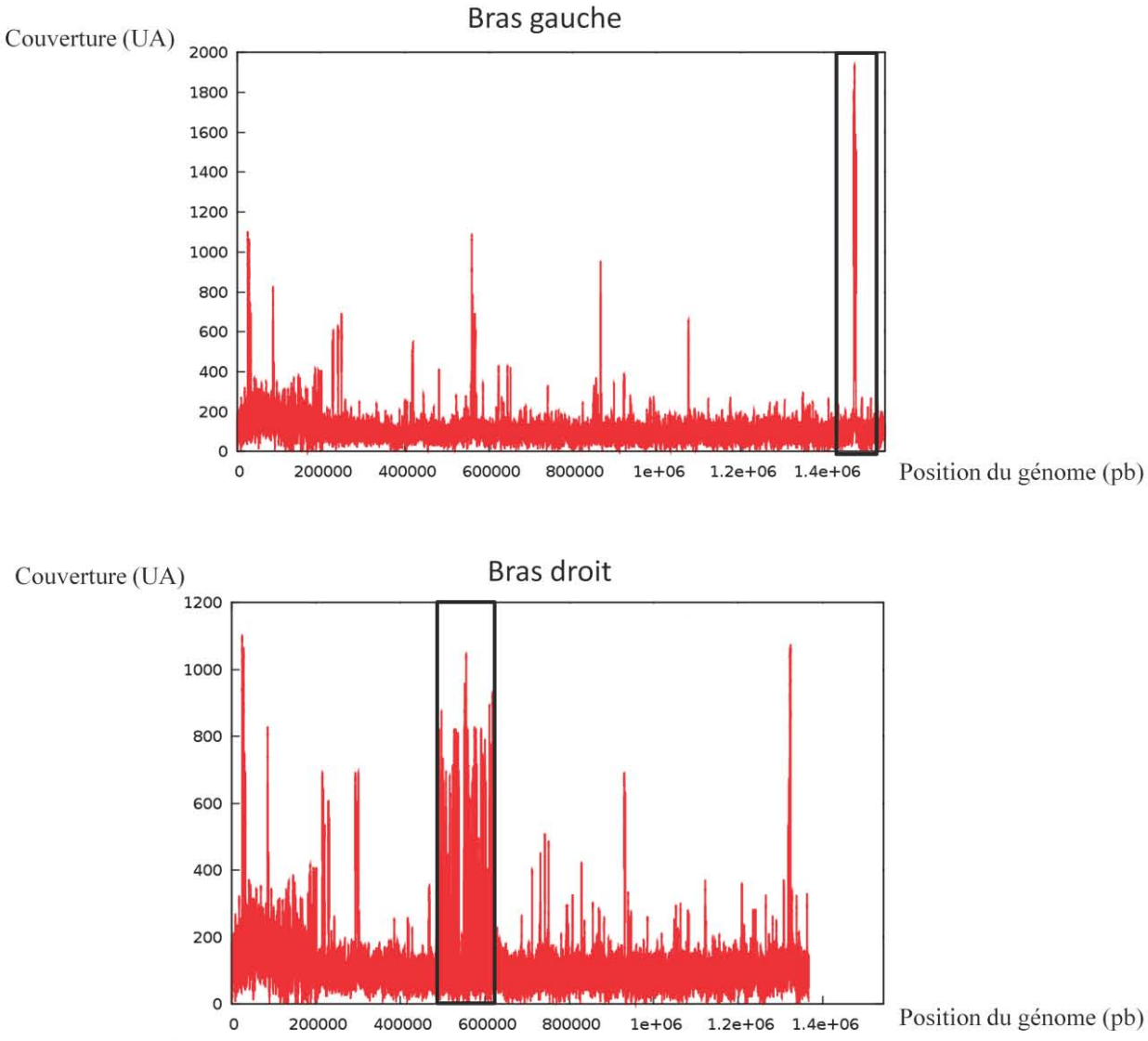


Figure 23 : *Représentation graphique de la couverture des lectures en fonction de la position génomique.*

La couverture est calculée à chaque position du génome (ou encore chaque base). Dans chaque carré noir, une zone de très haute couverture, pouvant influer sur la quantité de lectures alignées et la couverture globale, est mise en évidence.

Assembleur utilisé	Newbler	MIRA
Nombre de contigs	2739	2620
Nombre de contigs avec taille supérieure à 1 kb	793	1345
Taille du plus grand contig	89 627	89 597
N 50	13 178	7 802
Nombre de total de bases dans les contigs	8 241 017	8 217 879
Nombre total de bases dans les contigs avec taille supérieure à 1 kb	7 522 238	7 771 400
Nombre de lectures assemblées	719 730	882 440

Tableau 5 : *Récapitulatif des statistiques d'assemblage pour les lectures de pyroséquençage de Streptomyces ambofaciens ATCC23877.*

Ce tableau résume les résultats obtenus avec les deux assembleurs Newbler (assemblage fourni par le prestataire) et MIRA.

3.2.2°) Assemblage des banques « *single-end* » et « *mate-pair* »

L'analyse préliminaire des données de séquençage pour ces deux banques de séquences a montré que les lectures obtenues présentaient des lacunes, que ce soit en terme de qualité pour les « *single-end* » ou en terme de couverture pour les « *mate-pair* ». Ces séquences ont été assemblées indépendamment pour vérifier la justesse des analyses effectuées et déterminer si ces lectures pourraient être utilisées pour l'assemblage du génome de *Streptomyces ambofaciens*. Les statistiques des assemblages obtenus pour ces deux banques sont présentées dans le **tableau 6**.

Pour les lectures « *single-end* », 37 contigs sont obtenus pour une taille totale de 5675 pb. La longueur du plus grand contig est de 553 pb et le N50 est de 165 pb. Malgré le taux de couverture de 29 X, les résultats de l'assemblage sont décevants. Même si peu de contigs sont obtenus au final, leur taille cumulée ainsi que le N50 sont très faibles, montrant que ces contigs sont de taille réduite. La taille totale de ces contigs représente seulement 0,007 % de la taille estimée du génome ce qui est très faible. Les statistiques d'assemblage obtenus vont dans le même sens que l'analyse préliminaire des lectures : la banque de séquences « *single-end* » est de trop mauvaise qualité pour être utilisée dans un assemblage.

Pour les lectures « *mate-pair* », 125 contigs ont été obtenus pour une taille totale de 18 533 pb, ce qui représente 0,021% de la taille du génome estimée. Le N50 est égal à 114 pb, ce qui signifie que la moitié des contigs obtenus ont une taille inférieure à ce seuil. Pour cette banque, la couverture de 140X étant conséquente, les statistiques d'assemblage obtenus corroborent l'analyse préliminaire des lectures : la banque de séquence « *mate-pair* », bien que de bonne qualité, est inutilisable pour les assemblages à cause de la faible proportion du génome couvert par les lectures.

3.2.3°) Assemblage des lectures des banques « *paired-end* »

L'analyse des lectures obtenues pour cette banque montre une qualité globale satisfaisante et une couverture assez uniforme sur 99,9% du génome. Les assemblages obtenus devraient ainsi donner de meilleurs résultats que ceux obtenus précédemment. Plusieurs assemblages ont été testés sur les deux banques « *paired-end* » afin de déterminer si ces séquences peuvent être valorisées par un résultat de bonne facture. Les statistiques des assemblages obtenus sont résumées dans le **tableau 7**.

Pour la première banque « *paired-end* », ayant une couverture d'environ 17X, 10 940 contigs ont été obtenus pour une taille cumulée de 7 859 243 pb, ce qui représente environ 95 % de la taille du génome estimée. De plus, 2622 contigs ont une taille supérieure à 1 kb, qui est le seuil arbitraire à partir duquel on considère que les contigs sont de grande taille, pour une longueur cumulée de 4 871 641 pb, soit 57,3 % du génome de *Streptomyces ambofaciens*.

Pour la seconde banque « *paired-end* », ayant une couverture d'environ 108X, 1619 contigs ont été obtenus pour une taille totale de 8 099 218 pb, représentant environ 97,5 % de la taille estimée du génome. Si les contigs longs (dont la taille est supérieure à 1kb) sont considérés, la taille totale atteint 7 931 660. La comparaison des assemblages des deux banques « *paired-end* » révèle l'avantage apporté par une augmentation de la couverture. Bien que la portion du génome couvert varie peu, cela reste toutefois une estimation, le nombre de contigs diminue de façon notable pour la banque ayant le plus de couverture alors que dans le même temps, le N50 augmente. Ces deux critères montrent qu'augmenter la quantité de données d'un séquençage tend à améliorer la qualité finale de l'assemblage en créant des contigs de plus grande taille.

L'assemblage mixte des deux banques de lectures « *paired-end* » a été testé en ajoutant les contigs obtenus après assemblage des séquences de pyroséquençage 454. Ces contigs vont servir de référence pour l'assembleur qui va se servir de leurs séquences afin d'assembler les lectures Solexa/Illumina.

Cet assemblage possède le plus faible nombre de contigs, 1217, alors que dans le même temps, il possède le N50 et le nombre de bases totales les plus élevés avec respectivement

Banque(s) utilisée(s) pour l'assemblage	« *single-end* »	« *mate-pair* »
Nombre de lectures assemblées	6 361 402	25 735 132
Nombre contigs	37	125
Nombre contigs avec taille supérieure à 1 kb	0	1
Taille du plus grand contig	553	5378
N 50	165	114
Nombre total de bases dans les contigs	5675	18 533
Nombre total de bases dans les contigs avec taille supérieure à 1 kb	0	5378

Tableau 6 : *Récapitulatif des statistiques d'assemblage pour les banques « single-end » et « mate-pair ».*

Banque(s) utilisée(s) pour l'assemblage	PE 54	PE 36	PE 54 + PE 36 + Contigs pyroséquençage
Nombre de lectures assemblées	3 161 340	23 414 268	26 576 707
Nombre contigs	10 940	1619	1217
Nombre contigs avec taille supérieure à 1 kb	2 622	659	623
Taille du plus grand contig	8 448	123 148	125 262
N 50	1 292	20 512	22 595
Nombre total de bases dans les contigs	7 859 243	8 099 218	8 227 037
Nombre total de bases dans les contigs avec taille supérieure à 1 kb	4 871 641	7 931 660	8 099 118

Tableau 7 : *Récapitulatif des statistiques d'assemblage pour les deux banques « paired-end ».*

(PE 54 : banque « *paired-end* » avec lectures de 54 pb. PE 36 : banque « *paired-end* » avec lectures de 36 pb).

22 595 pb et 8 227 037 pb. Il semble donc, d'après les statistiques que cet assemblage soit le meilleur. L'ajout des contigs dans l'assemblage amène une faible augmentation de la couverture mais apporte, probablement par leur longueur, une meilleure résolution des répétitions. En effet, pour qu'une répétition pose problème lors d'un assemblage, il faut que sa longueur dépasse celle des lectures. Le cas échéant, la présence de séquences spécifiques de la répétition permet à l'assembleur de bien positionner les lectures entre elles. Il semble ainsi que l'augmentation du nombre de lectures assemblées apporte un bénéfice certain mais montre ses limites dans la résolution des répétitions.

3.3°) Assemblage des lectures de *Streptomyces ambofaciens* DSM40697

Pour l'assemblage du génome de la souche DSM40697, les lectures de pyroséquençage obtenues ont été utilisées avec les assembleurs Newbler et MIRA. Les résultats de l'assemblage sont présentés dans le **tableau 8**.

Les résultats de l'assemblage présentent un grand nombre de contigs, 2183, avec un N50 de 7451 pb, ce qui est assez faible, pour l'assembleur Newbler. Dans le cas de MIRA, le nombre de contigs est plus élevé, 2849, pour un N50 plus faible, 6483 pb. Toutefois, le nombre de total de bases dans les contigs est nettement plus élevé pour MIRA, avec 7 956 534, que pour Newbler, avec 7 524 790. Si on considère que les génomes des deux souches ont une taille équivalente, cela signifie que l'assembleur MIRA serait plus crédible sur ce critère. Dans les deux cas, les résultats de l'assemblage sont moins bons que pour la souche de *Streptomyces ambofaciens* ATCC23877, probablement à cause d'une plus faible quantité de données disponibles, environ 35 % de moins, pour la souche DSM40697.

3.4°) Conclusions sur les assemblages

Pour les deux souches de *Streptomyces ambofaciens* séquencées, les assemblages obtenus sont encore trop morcelés pour pouvoir supporter un processus d'annotation fonctionnelle. Toutefois, pour la souche ATCC23877, les résultats obtenus avec la stratégie de séquençage sont encourageants et la mise au point par le prestataire d'une banque « *mate-pair* » valide devrait permettre d'obtenir un assemblage suffisamment bon pour être exploité en terme de génomique descriptive et comparative. Cependant, les contigs obtenus ont été utilisés pour évaluer le degré de spécificité qui existe entre les deux souches afin de savoir si la comparaison génomique donnerait des résultats intéressants.

4°) Génomique comparative préliminaire des deux souches de *Streptomyces ambofaciens*

Pour déterminer la part de séquences spécifiques de chacune des souches de *Streptomyces ambofaciens*, les contigs obtenus après assemblage des lectures de pyroséquençage ont été utilisés, les dernières données Solexa/Illumina ayant été acquises trop tardivement pour être exploitées. Ces contigs ont été alignés les uns par rapport aux autres par le programme Nucmer du package MUMMER (Delcher *et al.* 2002) pour identifier les contigs entièrement spécifiques de chaque souche. Pour cette approche, les assemblages obtenus par MIRA ont été préférentiellement utilisés car plus morcelés. Ainsi, la recherche de contigs totalement spécifiques donnera plus de résultats. Le **tableau 9** regroupe les résultats obtenus en utilisant différents ensembles de séquences.

Quand l'ensemble des contigs sont considérés pour la souche ATCC23877, 946 contigs sont spécifiques de cette souche pour une taille cumulée de 546 841 pb. En ce qui concerne la souche DSM40697, 584 contigs sont obtenus avec une longueur totale de 273 265 pb. On observe une

Assembleur utilisé	Newbler	MIRA
Nombre de contigs	2 183	2 849
Nombre de contigs avec taille supérieure à 1 kb	1 689	1 486
Taille du plus grand contig	96 990	52 639
N 50	7 231	6 483
Nombre de total de bases dans les contigs	7 524 790	7 956 534
Nombre total de bases dans les contigs de taille supérieure à 1 kb	7 397 988	7 364 326
Nombre de lectures assemblées	583 701	572 621

Tableau 8 : *Récapitulatif des statistiques d'assemblage pour les lectures de pyroséquençage de Streptomyces ambofaciens DSM40697.*

Ce tableau résume les résultats obtenus avec les deux assembleurs Newbler (assemblage fourni par le prestataire) et MIRA.

Contigs utilisés pour l'alignement	Tous contigs ATCC23877 Tous contigs DSM40697		Contigs filtrés ATCC23877 Tous contigs DSM40697		Tous contigs ATCC23877 Contigs filtrés DSM40697	
	ATCC23877	DSM40697	ATCC23877	DSM40697	ATCC23877	DSM40697
Nombre de contigs spécifiques	930	584	817	671	1008	549
Taille totale des contigs spécifiques	546 841 pb	273 265 pb	430 348 pb	595 925 pb	881 585 pb	255 714 pb

Tableau 9 : *Tableau récapitulatif des alignements entre les contigs des souches ATCC23877 et DSM40697 de Streptomyces ambofaciens.*

Les contigs filtrés correspondent aux contigs, après assemblage des données de pyroséquençage, qui ne s'alignent pas contre les extrémités déjà étudiées au niveau intra-spécifique (273 757 pb pour l'extrémité gauche et 243 019 pb pour l'extrémité droite du génome).

spécificité notable entre les deux groupes de contigs comparés, correspondant à 6,7 % de la longueur totale de l'assemblage pour la souche ATCC23877 et 3,4 % pour la souche DSM40697.

Pour estimer la variabilité exclusivement en dehors des régions déjà séquencées, les contigs s'alignant contre les extrémités chromosomiques déjà étudiées par comparaison intra-spécifique, soit 273 757 pb pour les extrémités gauche et 243 019 pb pour les extrémités droite des deux souches, n'ont pas été considérés pour les alignements. De cette façon, la comparaison entre contigs totaux et filtrés pour ces séquences permet d'estimer quelle est la part de spécificité qui est contenue dans les régions non assemblées. Quand les contigs filtrés de la souche ATCC23877 sont alignés avec les contigs bruts de la souche DSM40697, 817 contigs lui sont spécifiques pour une taille cumulée de 430 348 pb. A l'inverse, 549 contigs sont spécifiques à la souche DSM40697, pour une taille cumulée de 258 714 pb, quand ces séquences sont comparées avec les contigs totaux de la souche ATCC23877. Quand on compare ces chiffres avec les résultats obtenus précédemment, cela signifie que sur les 546 841 pb spécifiques à la souche ATCC23877, 430 348 pb se trouveraient en dehors des extrémités chromosomiques déjà étudiées, soit 78,7% du total. De la même façon, pour la souche DSM40697, 94,7% de la spécificité détectée par cette approche se trouverait en dehors des régions déjà séquencées.

Ainsi, il semble que l'essentiel des régions spécifiques, détectées par l'approche, entre les génomes des deux souches soient localisées dans la partie non télomérique du génome, suggérant que l'obtention de leur séquence génomique complète apporterait de nombreuses informations sur les différences génétiques les séparant.

5°) Annotation fonctionnelle d'un contig spécifique de la souche ATCC23877 de *Streptomyces ambofaciens*

Après assemblage des lectures de pyroséquençage de la souche ATCC23877, le plus grand contig obtenu mesure 89 597 pb. Il se trouve par ailleurs que cette séquence n'est pas retrouvée chez la souche DSM40697. La plate-forme d'annotation mise en place a été utilisée pour identifier de façon fonctionnelle ce contig et déterminer la nature de cette région spécifique.

5.1°) Annotation syntaxique et fonctionnelle de la séquence

L'annotation syntaxique sert à identifier des régions d'intérêts biologiques, comme les gènes ou les RBS, dans une séquence. Grâce aux différents programmes de prédiction utilisés, 93 séquences codantes (ou CDS) ont été identifiées dans le contig de 89 597 pb. Aucun ARNt ou ARNr n'a été détecté (**figure 24**).

La fonction des séquences codantes prédites a d'abord été recherchée par BLASTp en utilisant la base de données non-redondante du NCBI (ou NR). Pour les 94 protéines blastées, 38 ne présentent aucune homologie de séquence avec la base de données utilisée. L'analyse des résultats montre un groupe de gènes présentant des homologies avec des protéines phagiques, et notamment des protéines structurant la capside virale (cont89_0071 ;cont89_0074 ;cont89_0083) ou des intégrases putatives (cont89_0052 ou cont89_0058) (**tableau 10**). Bien que les scores de BLAST obtenus soient faibles pour certaines de ces protéines, notamment cont89_0071 et cont89_0083, d'après le contexte génomique il est plausible que ces gènes aient une origine phagique.

Il a aussi été remarqué plusieurs protéines présentent des homologies de séquences significatives avec des protéines plasmidiques (tableau X). Chez les plasmides des *Streptomyces*, les protéines impliquées dans le transfert conjugatif présentent très souvent une certaine identité de séquence avec les protéines FtsK/SpoIIIE et Tra, qui sont des ADN translocases. La protéine codée par la CDS cont89_0089 présentant une identité de séquence avec les protéines FtsK/SpoIIIE de *Catenulispora acidiphila* et *Stackebrandtia nassauensis*, deux actinomycètes présentant différenciation morphologique analogue à celle des *Streptomyces*. Il semblerait donc que, malgré les scores de BLAST faibles qui sont obtenus, cette CDS code une protéine plasmidique. De

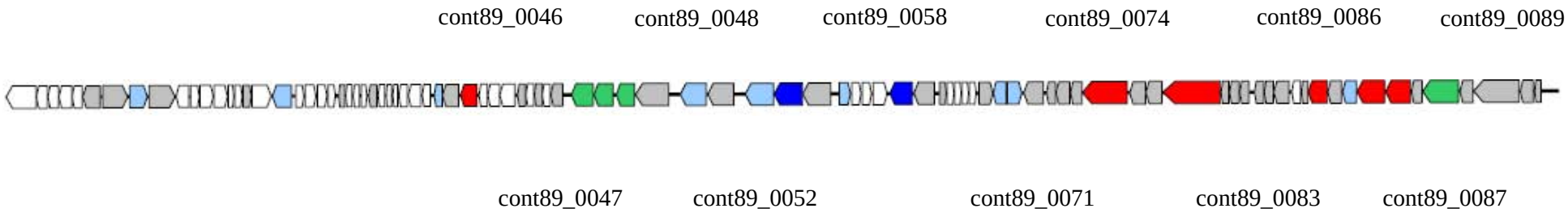


Figure 24 : *Organisation génétique des séquences codantes identifiées dans le contig annoté.*
Les CDS présentant des homologies de séquence avec des fonctions plasmidiques ou phagiques sont nommées sur le schéma.

Protéine d'entrée	Description de la protéine de la banque	Score de blast	Pourcentage d'identité	Longueur de l'alignement	e-value
cont89_0046	Putative plasmid partitioning protein *Nocardia farcinica* (189 aa)	83	35 %	176 aa	2e-14
cont89_0047	Putative plasmid partitioning protein *Mycobacterium sp. KMS* (237 aa)	181	46 %	225 aa	6e-44
cont89_0048	Putative Tra3-like protein *Streptomyces clavuligerus* (287 aa)	99	40 %	208 aa	4e-19
cont89_0052	Phage recombination protein Bet *Xylanimonas cellulosilytica* (327 aa)	191	56 %	189 aa	2e-46
cont89_0058					
cont89_0071	Phage-related protein tail component-like *Comamonas testosteroni* (2191 aa)	82	28 %	374 aa	3e-13
cont89_0074	Putative phage tail *Streptomyces sp. W9* (1552 aa)	772	45 %	1285 aa	0
cont89_0083	Putative head protein *Enterococcus faecalis* (337 aa)	95	29 %	330 aa	1e-17
cont89_0086	Phage domain protein *Streptomyces hygroscopicus* (639 aa)	674	59 %	636 aa	0
cont89_0087	Terminase large subunit, putative *Streptomyces hygroscopicus* (527 aa)	587	61 %	509 aa	1e-165
cont89_0089	Cell division FtsK/SpoIIIE *Catenulispora acidiphila* (681 aa)	65	26 %	340 aa	3e-08
	FtsK/SpoIIIE family protein *Stackebrandtia nassauensis* (726 aa)	58	25 %	342 aa	6e-06

Tableau 10 : *Récapitulatif des résultats de BLASTp apparentés à des éléments génétiques mobiles.*

Pour ce tableau, les homologies de séquences significatives (pourcentage d'identité supérieur ou égal à 30% et une longueur d'alignement représentant au moins 80 % de la taille de la protéine d'entrée) sont grisées. Toutes les bactéries présentes dans ce tableau appartiennent à l'ordre des *Actinomycetales* à l'exception de *Comamonas testosteroni* (*Burkholderiales*) et *Enterococcus faecalis (Lactobacillales).*

la même façon, la CDS cont89_0048 présente une homologie de séquence significative avec la protéine Tra de *Streptomyces clavuligerus*. Deux autres CDS, cont89_0046 et cont89_0047, présentent une bonne homologie de séquence avec des protéines impliquées dans la ségrégation des copies d'un plasmide au cours de la division cellulaire. Bien que certaines de ces protéines présentent aussi des homologies de séquence assez faibles, leur co-localisation sur le contig, à l'exception de cont89_0089, suggère la présence d'un plasmide dans la séquence étudiée.

5.2°) Recherche de domaines fonctionnels dans les protéines prédites

Pour aller plus loin dans l'annotation, des domaines fonctionnels ont été recherchés par InterproScan (ref) sur les protéines prédites dans le but de trouver une fonction aux protéines ne présentant pas d'homologie de séquence suffisantes à l'identification de leur action biologique. Les résultats sont résumés dans le **tableau 11**.

Sur les 94 protéines prédites, 20 d'entre elles se sont vu attribuées un ou plusieurs domaines fonctionnels. Les CDS cont89_0057 à cont89_0059 possèdent des domaines fonctionnels apparentés à des intégrases de phage. Pour la CDS cont89_0070, un domaine apparenté à N-acetylmuramoyl-L-alanine amidase, qui est un domaine retrouvé dans des enzymes lysant les parois de certaines bactéries, est mis évidence. Cette CDS pourrait ainsi coder un facteur lytique du bactériophage. Enfin, la CDS cont89_0088 possède un domaine terminase-like, qui est impliqué dans la translocation de l'ADN viral dans sa capside. La présence de nombreuses protéines possédant des domaines fonctionnels impliqués dans divers processus viraux (l'intégration, la lyse ou l'encapsidation) semble confirmer la présence d'une molécule phagique intégrée dans ce contig. Les nouvelles annotations semblent aussi mettre en évidence une organisation modulaire des gènes identifiés : les intégrases putatives et les protéines de capside virale putatives sont co-localisés (**figure 25**). Ce type d'organisation des gènes est souvent retrouvé dans les éléments génétiques mobiles.

5.3°) Comparaison nucléotidique du contig annoté

Au cours du processus d'annotation, il a été remarqué un grand nombre de protéines prédites présentant des homologies de séquences significatives (plus de 30 % d'identité de séquence pour un alignement d'au moins 80% de la taille de la protéine) avec des protéines identifiées dans un plasmide circulaire de *Streptomyces sp. zl12* mesurant 90 435 pb. La séquence de ce plasmide a été obtenue sur le portail du NCBI puis aligné avec le contig de la souche ATCC23877 pour vérifier si une identité de séquence nucléotidique pouvait être observée. La **figure 26** présente la visualisation sous ACT (Artemis Comparison Tool) de l'alignement obtenu.

Deux régions semblent être conservées dans le contig annoté : une de 21 467 pb et une autre de 25 561 pb. Le contig isolé dans la souche ATCC23877 présente dans ces régions une séquence qui semble conservée avec le plasmide pZL12, aussi bien en terme de synténie que d'identité nucléotidique, où les zones rouges présentent une identité de séquence supérieure à 60 %.

D'après cette comparaison, il semblerait ainsi que le contig isolé chez la souche ATCC23877 de *Streptomyces* ambofaciens soit apparenté à un plasmide.

Nom de la CDS	Domaines fonctionnels trouvés
cont89_0046	- HMMSmart : ParB-like nuclease - HMMTigr : parB part : ParB-like proteins - superfamily : ParB/Sulfiredoxin
cont89_0052	- HMMTigr : bet_lambda : phage recombination protein Bet
cont89_0057	- superfamily : lambda integrase-like, N-terminal domain
cont89_0058	- superfamily : _ DNA breaking-rejoining enzymes _ lambda integrase-like, N-terminal domain - HMMPfam : Phage integrase
cont89_0059	- superfamily : _ DNA breaking-rejoining enzymes _ lambda integrase-like, N-terminal domain
cont89_0060	- HMMPfam : Prophage CP4-57 regulatory
cont89_0070	- superfamily : N-acetylmuramoyl-L-alanine amidase-like - HMMSmart : N-acetylmuramoyl-L-alanine amidase, family 2 - HMMPfam : N-acetylmuramoyl-L-alanine amidase, family 2
cont89_0088	- HMMPfam : Terminase-like

Tableau 11 : *Tableau récapitulatif des domaines fonctionnels détectés dans les protéines prédites par l'annotation syntaxique.*

Les domaines fonctionnels probablement apparentés à des éléments génétiques mobiles sont présentés dans ce tableau.

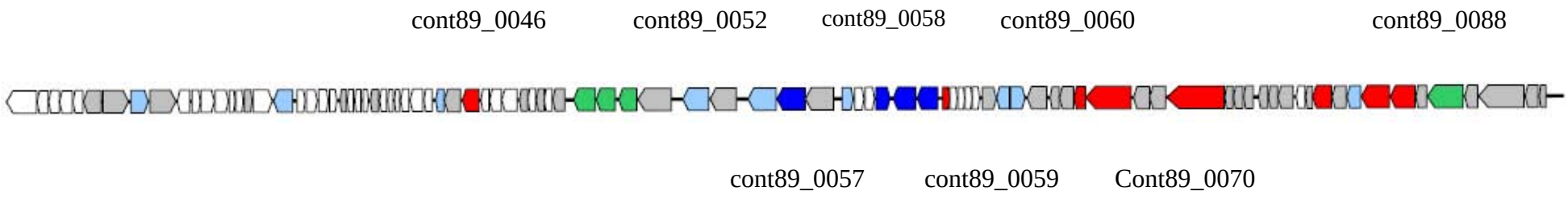


Figure 25 : *Organisation génétique des séquences codantes identifiées dans le contig annoté.*
Les CDS codant des protéines présentant des domaines apparentés à des fonctions plasmidiques ou phagiques sont nommées sur le schéma.

20 ter

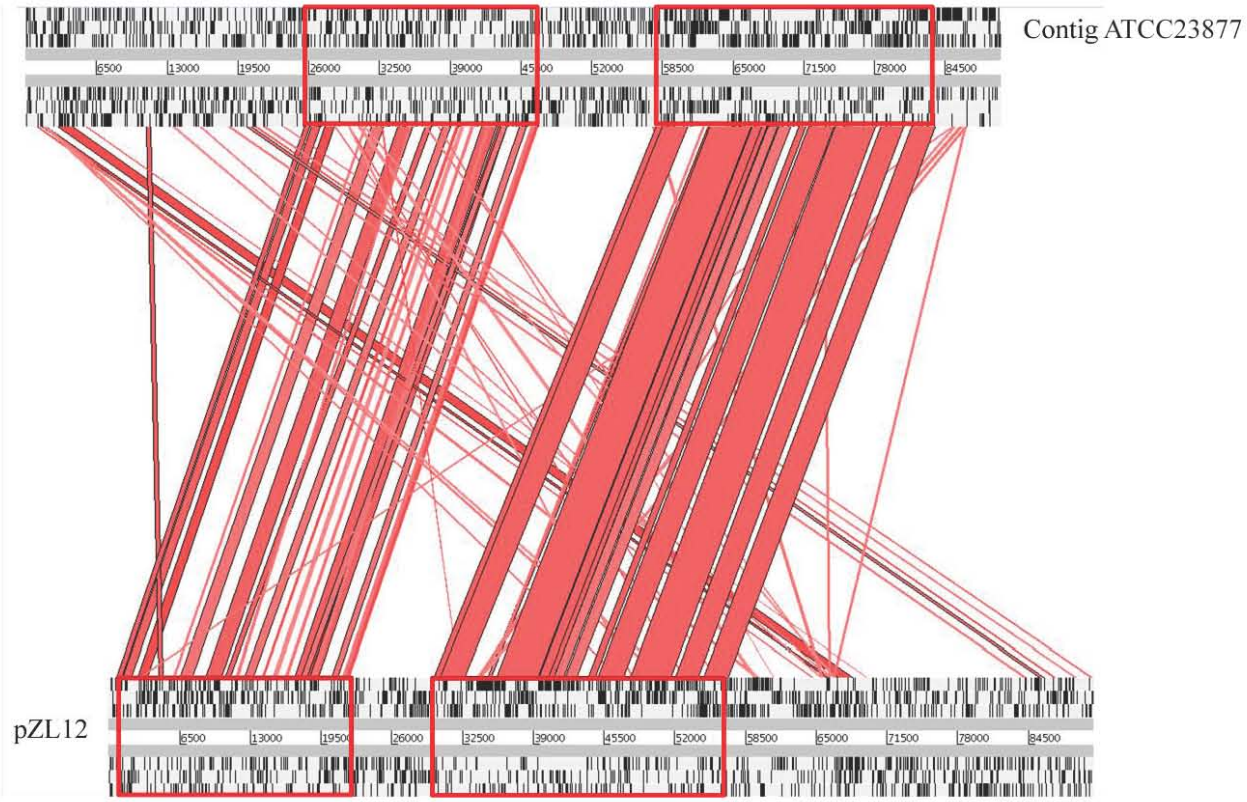


Figure 26 : *Représentation sous ACT de l'alignement entre pZL12 et le contig isolé de la souche ATCC23877.*

Pour chacune des deux séquences, ACT (Artemis Comparison Tools) représente les ORF dans les 6 cadres de lectures. En rouge, il est représenté les séquences présentant une identité de séquence nucléotidique supérieure à 60% et dont les deux séquences sont alignées dans le même sens. Le résultat final montre une conservation de séquence dans les carrés rouges.

DISCUSSION

1°) Analyse des données de séquençage et des assemblages

1.1°) Résultats obtenus pour la souche *Streptomyces ambofaciens* ATCC23877

Au cours de ces travaux, différentes collectes de données de séquences ont été traitées pour les souches ATCC23877 et DSM40697 de *Streptomyces ambofaciens*. L'essentiel des analyses ont d'abord été effectuées sur les données obtenues par terminaison cyclique réversible (Solexa/Illumina) obtenues pour la souche ATCC23877.

Les données de la banque « *single-end* » (prestataire GATC) montrent des scores de qualité (correspondant à la probabilité d'erreur de lecture à chaque position nucléotidique) globalement médiocres. Les piètres statistiques d'assemblages obtenus à partir de ces séquences ont confirmé la mauvaise qualité de ces lectures. Il est ainsi très probable que la prestation réalisée ait été de mauvaise qualité, rendant ces lectures inutilisables pour nos assemblages.

La banque « *mate-pair* » présente des scores de qualité excellents après filtration mais la couverture des lectures pose problème. Une très faible proportion du génome (moins de 10 %) semble être contenue dans la banque de séquence bien qu'à très haute couverture. Les assemblages tentés à partir de ces séquences donnent des mauvaises statistiques, montrant ainsi que la faible proportion du génome couvert pose problème pour les assemblages. Plusieurs tentatives infructueuses et la mise en place d'un protocole spécialisé ont été nécessaires au prestataire Fasteris pour créer des séquences « *mate-pair* » valides. Pour vérifier la validité de cette banque, plusieurs fragments d'ADN préparés selon le protocole expliqué en **figure 15** ont été clonés puis séquencés par le prestataire (méthode Sanger). Ces fragments sont ensuite alignés par BLAST contre les séquences de référence pour vérifier si la distance séparant les deux lectures était compatible avec celle sélectionnée par le protocole « *mate-pair* ». De cette façon, la banque « *mate-pair* » reçue a été validée. La couverture anormale du génome n'a pas pu être décelée malgré la validation de cette étape. La qualité des séquences étant excellente, l'analyse du protocole de préparation de la banque semble indiquer que l'étape défaillante serait l'amplification par PCR (cf **figure 15**), qui aurait favorisé la surreprésentation de régions génomiques très limitées. Le prestataire se penche actuellement sur le problème et va réitérer la préparation de la banque.

Les deux banques « *paired-end* » fournies par le prestataire Fasteris présentent des bons scores de qualité ainsi qu'une couverture relativement uniforme le long des séquences de référence. D'après les analyses mises en place, les lectures semblent ainsi avoir été bien préparées. Les statistiques d'assemblage obtenus sont bons mais le nombre de contigs est encore très important. Le caractère « *paired-end* » de la banque donne une information supplémentaire sur la distance séparant deux lectures, permettant une construction plus précise des contigs en empêchant la formation de chimères dues aux répétitions. Toutefois, si la distance séparant les deux lectures est plus faible que la longueur de la répétition, le bénéfice apporté par cette banque est aboli. Pour la banque « *paired-end* », la distance séparant les deux lectures est d'environ 300 pb. Ainsi, le morcellement de l'assemblage final résulte probablement de la présence de répétitions excédant cette longueur. L'obtention d'une banque « *mate-pair* » valide, dont la distance séparant les lectures est d'environ 3 kb, permettra certainement d'améliorer le résultat d'assemblage.

L'ajout des données de pyroséquençage améliore les assemblages mais leur apport reste tout de même négligeable au vu du résultat. Toutefois, la stratégie de séquençage mise en place implique l'incorporation des lectures obtenues par différentes méthodes : pyroséquençage haut-débit et terminaison cyclique réversible (« *paired-end* » et « *mate-pair* »). Des tentatives d'assemblages avec l'ensemble des séquences seront effectuées afin de confirmer le bienfondé de la stratégie de séquençage mixte.

1.2°) Résultats obtenus pour la souche *Streptomyces ambofaciens* DSM40697

Les résultats d'assemblages obtenus pour cette souche à partir des lectures de pyroséquençage, reçues du prestataire GATC, ne sont pas bons. La quantité plus faible de données assemblées par rapport à la première souche explique ces statistiques. L'obtention du génome de la souche ATCC23877 a été considérée comme prioritaire dans un premier temps pour tester si la stratégie de séquençage mixte pouvait fonctionner. Si elle fonctionne, cette séquence pourra être utilisée comme référence pour faciliter l'assemblage du génome de la souche DSM40697. Cela justifie la quantité de données moindres collectées pour le séquençage du génome de cette souche.

2°) Génomique comparative entre la souche ATCC23877 et DSM40697 de *Streptomyces ambofaciens*

La recherche de spécificité entre les deux souches a permis d'identifier des loci spécifiques. L'approche a consisté à ne considérer que les contigs totalement spécifiques de chaque souche. Ainsi, les contigs présentant à la fois des régions spécifiques et des régions conservées n'ont pas été considérés. Cela signifie que les résultats obtenus sont sous-estimés en terme de quantité de séquences spécifiques de souche.

Malgré cette approximation, environ 3% de la séquence contenue dans les contigs de la souche DSM40697 et 5% de la séquence des contigs de la souche ATCC23877 sont spécifiques. Des études de génomique comparative intra-spécifique chez *Mycoplasma agalactiae* ont montré qu'environ 10 % du génome de chaque souche était spécifique (Nouvel *et al* 2010). Le chiffre assez faible qui a été obtenu par notre approche reflète le biais qui a été proposé : il est très probable que la spécificité constatée soit sous-estimée. Toutefois, il a été montré que l'essentiel de la spécificité détectée par notre approche se trouvait en dehors des répétitions terminales (TIR) déjà étudiées au cours de comparaisons génomiques. Il est donc important d'obtenir les deux séquences complètes pour localiser les régions spécifiques.

3°) Annotation fonctionnelle de contigs spécifiques

L'analyse des séquences spécifiques de chaque souche a révélé la présence de loci présentant des homologies avec des plasmides ou des bactériophages putatifs. Certains de ces contigs possèdent aussi des homologies de séquences avec des transposons. La présence de telles séquences, révèle probablement la présence d'éléments génétiques mobiles spécifiques de chaque souche montrant l'importance du transfert horizontal dans l'évolution à court terme des lignées bactériennes.

Parmi ces séquences spécifiques, un contig de 89 597 pb a été sélectionné pour être annoté en détail afin de déterminer la nature de cette séquence. D'après l'annotation, le contig spécifique à *Streptomyces ambofaciens* ATCC23877 possède gènes apparentés à des éléments génétiques mobiles. Toutefois, le doute subsiste sur sa nature réelle. D'après les analyse *in-silico*, certains gènes seraient d'origine plasmidique tandis que d'autres auraient une origine phagique. Il semble toutefois plus probable que le contig annoté code de nombreuses fonctions d'origine plasmidique de par sa forte ressemblance avec le plasmide circulaire de *Streptomyces sp. zl12*. Ainsi, la présence de gènes impliqués dans le transfert conjugatif de ce plasmide suggère que cet élément génétique puisse être fonctionnel, et de fait, pourrait expliquer son absence de la souche DSM40697 de *Streptomyces ambofaciens*.

Toutefois, l'essentiel des annotations effectuées sont basées sur des homologies de séquences avec des protéines contenues dans les bases de données du NCBI. La fiabilité de nos résultats est, dans ce cas, dépendante de la qualité des annotations effectuées sur les séquences de ces bases de données.

Une acquisition récente de cet élément par la souche ATCC23877 ou une perte dans la souche DSM40697 pourrait constituer un évènement d'évolution rapide. Une démonstration fonctionnelle est nécessaire pour valider cette prédiction comme, par exemple, au travers de la mise en évidence expérimentale de la mobilité de l'élément au cours d'expérience de conjugaison entre les deux souches de *S. ambofaciens* (souche ATCC23877 comme donneuse et souche DSM40697 comme receveuse).

4°) Perspectives

L'obtention du génome complet de la souche ATCC23877 est le premier objectif à atteindre. La fabrication d'une banque de séquence « *mate-pair* » est indispensable pour valider les différentes étapes d'analyse et faire progresser les assemblages. Cela permettrait par ailleurs de valider la stratégie de séquençage mise en place pour l'appliquer par la suite à la souche DSM40697. Après l'obtention des deux génomes, la plateforme d'annotation mise au point pourrait être utilisée pour annoter ces deux séquences. A partir des séquences codantes prédites, les outils de génomique comparative mis au point par Frédéric Choulet (Thèse UHP, 2006) pourraient être utilisés pour étudier de façon globale les différences entre les deux génomes.

Le principal intérêt dans l'étude de la spécificité séparant les deux souches est d'appréhender les mécanismes d'évolution rapide qui façonnent le génome. Les études faites au préalable sur les extrémités terminales des deux souches ont mis en évidence de nombreux évènements d'insertion/délétion qui pourraient être dus à des mécanismes de recombinaison illégitime. Il serait intéressant de déterminer si cette forte plasticité génomique est limitée aux ragions terminales, validant ainsi l'hypothèse selon laquelle ces séquences seraient des cibles privilégiées pour l'adaptation environnementale par perte, acquisition ou création de nouvelles fonctions.

La génomique comparative des deux souches permettrait en plus d'apprécier l'importance du transfert horizontal de gènes sur l'évolution rapide des deux souches de *Streptomyces ambofaciens*.

La nature des gènes spécifiques de chaque souche pourrait aussi être recherchée dans le but d'identifier des fonctions ayant un intérêt biotechnologique. Les *Streptomyces* sont particulièrement réputées pour leurs capacités à synthétiser des métabolites secondaires et l'identification de gènes spécifiques de souches pourrait révéler des nouvelles voies de biosynthèse de métabolites d'intérêts.

BIBLIOGRAPHIE

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res. 25:3389-3402.

Bao, K., and Cohen, S.N. 2003. *Recruitment of terminal protein to the ends of Streptomyces linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication*. Genes Dev. 17: 774-785.

Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, HuangCH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. 2002. *Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)*.Nature. 417(6885):141-7

Berdy, J. 2005. *Bioactive microbial metabolites*. J Antibiot. 58: 1-26.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. *ALLPATHS: de novo assembly of whole-genome shotgun microreads*. Genome Res. 18(5):810-20.

Challis GL. 2008. *Mining microbial genomes for new natural products and biosynthetic pathways*. Microbiology. 154(Pt 6):1555-69.

Chater, K.F. (1993) *Genetics of differentiation in Streptomyces*. Annu Rev Microbiol 47: 685-713.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S. 2004. *Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs*. Genome Res. 14(6):1147-59.

Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou FX, Fourrier C, Guérineau M, Decaris B, Barbe V, Pernodet JL, Leblond P. 2006(a). *Evolution of the terminal regions of the Streptomyces linear chromosome.* Mol Biol Evol. 23(12):2361-9.

Choulet F, Gallois A, Aigle B, Mangenot S, Gerbaud C, Truong C, Francou FX, Borges F, Fourrier C, Guérineau M, Decaris B, Barbe V, Pernodet JL, Leblond P. 2006. *Intraspecific variability of the terminal inverted repeats of the linear chromosome of Streptomyces ambofaciens*. J Bacteriol. 188(18):6599-610.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. 1999. *Improved microbial gene identification with GLIMMER*. Nucleic Acids Res 27: 4636-4641.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. *Fast algorithms for large-scale genome alignment and comparison*. Nucleic Acids Res. 30(11):2478-83.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing*. Nucleic Acids Res. Sep;36(16):e105

Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ. 2008. *Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.* Nat Methods. 5(8):679-82

Fischer, G., Decaris, B., and Leblond, P. 1997. *Occurrence of deletions, associated with genetic instability in Streptomyces ambofaciens, is independent of the linearity of the chromosomal DNA*. J Bacteriol. 179: 4553- 4558.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. *Accuracy and quality of massively parallel DNA pyrosequencing*. Genome Biol. 8(7):R143.

Hütter, R. 1967. *In Systematik der Streptomycete*. Verlag, K. (ed). Basel.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. 2003. *Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis*. Nat Biotechnol. 21(5):526-31.

Karoonuthaisiri N, Weaver D, Huang J, Cohen SN, Kao CM. 2005. *Regional organization of gene expression in Streptomyces coelicolor*. Gene. 353(1):53-66.

Leamon, J.H. Rothberg, J.M. 2007.*Cramming more sequencing reactions onto microreactor chips*. Chem. Rev. 107, 3367–3376.

Leblond, P., and Decaris, B. 1999. *'Unstable' linear chromosomes: the case of Streptomyces*. In Organization of the Prokaryotic Genome. Charlebois, R.L. (ed). Washington, D.C.: American Society for Microbiology, pp. 235-261.

Lin YS, Kieser HM, Hopwood DA, Chen CW. 1994. *The chromosomal DNA of Streptomyces lividans 66 is linear*. Mol Microbiol. 1993 Dec;10(5):923-33. Erratum in: Mol Microbiol. 14(5):1103.

Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, MalekJ, McKernan K, Ranade S, Shea TP, Williams L, Young S, Nusbaum C, Jaffe DB. 2009. *ALLPATHS 2: small genomes assembled accurately and with high continuity fromshort paired reads*. Genome Biol. 10(10):R103.

Martin P, Dary A, André A, Fischer G, Leblond P, Decaris B. 1999. *Intraclonal polymorphism in the bacterium Streptomyces ambofaciens ATCC23877: evidence for a high degree of heterogeneity of the wild type clones*. Mutat Res. 430(1):75-85.

Metzker ML. 2010. *Sequencing technologies - the next generation*. Nat Rev Genet. Jan;11(1):31-46.

Miller JR, Koren S, Sutton G. 2010. Assembly *algorithms for next-generation sequencing data*. Genomics. 95(6):315-27

Musialowski, M.S., Flett, F., Scott, G.B., Hobbs, G., Smith, C.P., and Oliver, S.G. 1994. *Functional evidence that the principal DNA replication origin of the Streptomyces coelicolor chromosome is close to the dnaAgyrB region*. J Bacteriol 176: 5123-5125.

Nouvel LX, Sirand-Pugnet P, Marenda MS, Sagné E, Barbe V, Mangenot S, Schenowitz C, Jacob D, Barré A, Claverol S, Blanchard A, Citti C. 2010. C*omparative genomic and proteomic analyses of two Mycoplasma agalactiae strains: clues to the macro- and micro-events that are shaping mycoplasma diversity*. BMC Genomics. 11:86.

Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S. 2008. *Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350.* J Bacteriol. 190(11):4050-60.

Pang, X., Aigle, B., Girardet, J.M., Mangenot, S., Pernodet, J.L., Decaris, B., and Leblond, P. 2004. *Functional angucycline-like antibiotic gene cluster in the terminal inverted repeats of the Streptomyces ambofaciens linear chromosome*. Antimicrob Agents Chemother. 48: 575-588.

Pevzner PA, Tang H, Waterman MS. 2001. *An Eulerian path approach to DNA fragment assembly*. Proc Natl Acad Sci U S A. 98(17):9748-53,

Pinnert-Sindico, S. 1954. *Une nouvelle espèce de Streptomyces productrice d'antibiotiques: Streptomyces ambofaciens n. sp. caractères culturaux*. Ann Inst Pasteur (Paris). 87: 702-707.

Rothberg, J.M and Leamon, J.H. 2008. *The development and impact of 454 sequencing*. Nat Biotechnol. Oct;26(10):1117-24.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. 2000. *Artemis: sequence visualization and annotation.* Bioinformatics. 16: 944-945.

Sanger, F., Coulson, A., 1975. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase*. J. Mol. Biol. 94, 441–448.

Sanger, F., Coulson, A., Friedmann, T., Air, G., Barrell, B., Brown, N., Fiddes, J., Hutchison, C., Slocombe, P., Smith, M., 1978. *The nucleotide sequence of bacteriophage phiX174*. J. Mol. Biol. 125, 225–246.

Strohl, W.R. 1992. *Compilation and analysis of DNA sequences associated with apparent streptomycete promoters*. Nucleic Acids Res. 20: 961-974.

Suzek, B.E., Ermolaeva, M.D., Schreiber, M., and Salzberg, S.L. 2001. *A probabilistic method for identifying start codons in bacterial genomes*. Bioinformatics. 17: 1123-1130.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. *The COG database: new developments in phylogenetic classification of proteins from complete genomes*. Nucleic Acids Res. 29: 22-28.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K,

Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. *The sequence of the human genome*. Science. 291(5507):1304-51

Wenner, T., V. Roth, G. Fischer, C. Fourrier, B. Aigle, B. Decaris, and P. Leblond**.** 2003. *End-to-end fusion of linear deleted chromosomes initiates a cycle of genome instability in Streptomyces ambofaciens*. Mol. Microbiol. 50**:**411–425.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP, Eisen JA. 2009. *A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea*. Nature. 462(7276):1056-60.

Zdobnov EM, Apweiler R. 2001. *InterProScan--an integration platform for the signature-recognition methods in InterPro*. Bioinformatics. 17(9):847-8.

Zerbino DR, Birney E. 2008. *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*. Genome Res. 18(5):821-9.

RESUME :

Les *Streptomyces* appartiennent à la communauté microbienne des sols et sont confrontées à des conditions biotiques et abiotiques changeantes. La génomique comparée des espèces a révélé une organisation génétique particulière, avec la région centrale du chromosome linéaire conservée contrastant avec des régions terminales très variables.

L'objectif de ce travail est d'obtenir le génome complet de *Streptomyces ambofaciens* ATCC23877 et d'effectuer une comparaison intraspécifique avec celui de la souche DSM40697 afin d'évaluer le niveau de spécificité de lignées très fortement liées au niveau phylogénétique. Pour cela, les données de séquençage (pyroséquençage et terminaison réversible) collectées auprès de prestataires ont été évaluées en termes de qualité, et assemblées en un nombre de 'contigs' minimum. 1217 contigs ont été obtenus pour la souche ATCC 23877 regroupant 8,2 millions de nucléotides séquencés (avec un taux de couverture de 127 X pour chaque nucléotide du génome estimé). Le génome de la souche DSM40697 comprend 2849 'contigs', au minimum, pour 7,9 millions de nucléotides collectés (taux de couverture estimé à environ 15 X). La recherche de 'contigs' spécifiques de chacune des deux souches a été réalisée, et a permis d'identifier 546 kb (7%) et 273 kb (3%) respectivement spécifiques des souches ATCC23877 et DSM40607. Parmi les régions spécifiques de la souche ATCC23877, un locus de 89 kb code potentiellement des protéines homologues de bactériophages et de plasmides indiquant la présence de tout ou partie d'éléments génétiques mobiles dans cet îlot génomique.