université PARIS DIDEROT PARIS 7

MASTER 2 Biologie Informatique

Parcours professionnel

Promotion 2011 - 2012

Mise au point d'un pipeline de typage bactérien par approche pangénomique

Adrien PAIN

Lieu de stage :
Unité Mathématique Informatique et Génome (MIG)
Institut National de la Recherche Agronomique (INRA)
Domaine de Vilvert, 78 350 Jouy-en-Josas Cedex

Encadrants du stage :
Dr. Philippe Bessières et Pr. Mahendra Mariadassou

Remerciements

Je tiens tout d'abord à remercier Philippe Bessières et plus particulièrement Mahendra Mariadassou, mes encadrants de stage, pour leur accueil, leur confiance et la disponibilité dont ils font preuve depuis mon arrivée.

Un grand merci également à l'ensemble des membres de l'unité MIG pour m'avoir intégré rapidement au sein du laboratoire, pour leur bonne humeur au quotidien et pour leurs réponses à mes questions.

Bonne route à vous tous et encore une fois merci pour ces 6 mois passés ici !

Sommaire

Introduction

Les bactéries sont des organismes asexués qui se reproduisent par scissiparité. Ainsi une bactérie mère va, après division cellulaire, donner naissance à deux bactéries filles, dont les patrimoines génétiques sont identiques aux erreurs de réplication près. La principale source de variabilité des bactéries est le transfert horizontal de gènes (THG) au cours duquel une bactérie reçoit un fragment d'ADN d'une autre bactérie et l'intègre à son génome ce qui lui permet ensuite de transmettre à sa descendance ce nouveau matériel. Grâce à ce mécanisme, plusieurs formes (souches) d'une même espèce bactérienne peuvent coexister et se maintenir. Ces différentes souches partagent une grande partie de leur matériel génétique mais présentent aussi des spécificités (gènes de résistance, virulence, etc). Par exemple, l'espèce bactérienne *Escherichia coli* est une bactérie très commune dans le corps humain (la souche HS fait partie de notre flore commensale), mais d'autres souches sont très pathogènes (la souche O127:K63:H6 occasionne des diarrhées), voire même létales (la souche O104:H4 est responsable de l'épidémie du syndrome hémolytique et urémique ayant fait 27 morts en Europe en 2011).

Les bactéries sont extrêmement utiles aux hommes : elles sont par exemple utilisées dans l'industrie agroalimentaire pour la fabrication du fromage et des yaourts ou encore pour le traitement des eaux usées dans les stations d'épuration. De plus, elles participent au bon fonctionnement des autres organismes vivants. Par exemple chez l'homme les bactéries participent au processus de digestion, à la production de certaines vitamines, etc. Ces bactéries sont inoffensives, voire bénéfiques à notre santé mais dans certains cas elles peuvent êtres pathogènes : le choléra, la peste, l'anthrax sont des exemples de pathologies induites par des bactéries. Dans ce contexte, les bactéries sont énormément étudiées et des Centres de Ressources Biologiques (CRB) existent pour assurer la conservation et la description des bactéries dans les domaines précités. Afin d'aider les CRB à gérer leurs collections bactériennes, l'unité Mathématique Informatique et Génome (MIG) du centre INRA de Jouy-en-Josas souhaite mettre à leur disposition un outil automatique permettant de distinguer différentes souches d'une même espèce bactérienne (c'est à dire de faire du typage bactérien) sur la base de leurs répertoires de gènes. Un tel outil permettrait aux CRB de s'assurer que leurs collections bactériennes ne contiennent pas de duplicats indépendamment isolés de différents lieux et considérés, à tort, comme des souches bactériennes différentes.

Récemment, des expériences *in silico* ont démontré le potentiel des expériences de séquençage à haut-débit dans l'identification de souches bactériennes avec une précision

jusque-là jamais atteinte [1]. La méthode décrite par Hall et al. (2010) est basée sur une approche pangénomique. Les gènes présents dans l'ensemble des souches d'une espèce bactérienne forment le pangénome de cette espèce qui peut être divisé en 3 catégories :

- le génome de base (*core genome*) : contient les gènes présents chez toutes les souches de l'espèce ;

- le génome accessoire (*accessory genome*) : contient les gènes présents chez plusieurs souches mais pas toutes ;

- les gènes uniques : ceux qui ne sont présents que dans une seule souche.

Dans cette approche pangénomique une distance est calculée entre génomes sur la base des profils de présence/absence des gènes. Ainsi les gènes du génome de base étant présents chez tous les génomes étudiés, ils ne permettent pas de calculer une distance sur la base de leurs profils et seul le génome distribué (les gènes uniques et les gènes accessoires) sont utilisés. La première étape consiste à calculer, pour chaque paire de bactéries, la distance entre leurs génomes distribués (DGD). Pour cela, on attribue à chaque gène du génome distribué un score : 0 si le gène est présent chez un des génomes et absent chez l'autre et 1 si le gène est présent ou absent dans les deux génomes. Ces scores sont ensuite additionnés et le total est divisé par le nombre de gènes du génome distribué, ce qui permet d'obtenir la fraction de gènes partagés (FGP). La DGD comparée est obtenue en calculant 1 - FGP. Deux bactéries sont décrétées appartenir à la même souche si leur DGD est inférieure à la moyenne des DGD obtenues à laquelle on a soustrait l'écart-type. D'après l'étude réalisée par l'équipe de Hall, cette technique est plus résolutive qu'une autre technique *in silico* basée sur la comparaison des séquences du génome de base et n'aurait pas d'équivalent en termes de précision.

Ainsi le pipeline qui doit être mis en place doit permettre de faire du typage à partir des données de séquençage à haut-débit. Pour cela il doit donc en amont permettre de reconstruire le pangénome d'un ensemble de souches et en extraire le génome accessoire. De plus le laboratoire souhaite que le pipeline fonctionne sur des ordinateurs "classiques" et que les temps de calculs soient raisonnables. Le choix des différents outils à intégrer au pipeline doit donc tenir compte de ces éléments.

Matériel et méthodes

Description des données de départ

Le pipeline prend en entrée des données de séquençage à haut débit au format fasta ou fastq. Pour tester le pipeline, des lectures (ou *reads* en anglais, terme générique désignant les séquences sortant du séquenceur et qui constituent pour nous les données de départ) ont été générées à partir d'un script perl qui permet, à partir d'un génome de référence, de produire un nombre défini de lectures (au format fasta), pairées ou non, uniformément distribuées le long du génome de référence et en choisissant : leur taille, le nombre de bases entre deux lectures pairées et leur taux d'erreurs [2]. Ce script a donc été utilisé afin de générer des lectures pairées d'une longueur de 100 bases et séparées de 350 bases, ce qui pour des données réelles pourrait correspondre à des lectures produites par un séquenceur de type Illumina. De plus le nombre de lectures a été choisi afin que la couverture du génome de référence soit de 100 X. Enfin différents taux d'erreurs ont été utilisés, allant de 1 à 5 % par pas de 1 %. Les numéros d'accession des génomes de référence utilisés se trouvent en annexe (annexe 1).

Expériences réalisées

Pour tester la faisabilité de l'approche pangénomique nous avons dans un premier temps utilisé 3 génomes d'*Escherichia coli* pour chacun desquels 3 jeux de lectures ont été générés. Les génomes de référence utilisés pour cette expérience correspondent aux souches K12-MG1655, BL21 et O157:H7-Sakai. Cette expérience a été répétée 5 fois avec différents taux d'erreurs dans les lectures (de 1 à 5%) afin de tester la faisabilité de l'approche pangénomique et de déterminer le taux d'erreur maximal pour lequel le typage permet d'identifier comme tels les 3 réplicats issus de la même souche. Cette expérience a ensuite été étendue à 14 souches d'*Escherichia coli* (dont les 3 précédentes) pour lesquels 2 jeux de lectures ont été générés contre 3 auparavant.

Pour tester la sensibilité de la méthode nous avons ensuite utilisés 5 génomes de la bactérie *Bacillus anthracis*, car elle est connue pour sa faible variabilité pangénomique [3]. Pour chacun des 5 génomes, 2 jeux de lectures ont été générés. Le but de cette expérience est de déterminer si dans un cas idéal, c'est à dire avec des lectures "propres" (1% d'erreur ici), il est possible de distinguer par l'approche pangénomique des souches bactériennes très proches.

Mise au point du pipeline

Le pipeline devant fonctionner directement à partir des lectures qui sortent des appareils de séquençage à haut-débit, il est nécessaire dans un premier temps de reconstruire le répertoire de gènes des bactéries étudiées avant de procéder au typage. Pour cela il faut d'abord reconstruire pour chaque bactérie le génome à partir des lectures (étape d'assemblage) puis utiliser un programme de détection de gènes sur le génome reconstruit. Enfin la dernière étape consiste à déterminer parmi l'ensemble des gènes ceux qui font partie du génome de base, du génome accessoire et les gènes uniques.

Assemblage des génomes

Le but ici est d'assembler des génomes afin de pouvoir ensuite utiliser un programme de détection de gènes, il n'y a donc pas besoin d'un assemblage parfait. De plus on veut un outil qui soit aussi rapide que possible sachant que l'on va devoir assembler et détecter les gènes de plusieurs génomes, avec un ordinateur de configuration courante.

Pré-traitement des lectures

En général avant de procéder à l'assemblage *de novo* d'un génome on effectue un prétraitement des lectures afin d'éliminer ou de cliver celles de mauvaise qualité. Néanmoins l'utilisation de lectures générées automatiquement au format fasta ne nous permet pas d'effectuer un tel nettoyage, car nous ne disposons pas d'information sur leur qualité (mise à part le fait que l'on choisit le taux d'erreurs qu'elles contiennent au moment de leur création). Un prétraitement basé sur la redondance locale des lectures a tout de même été réalisé en utilisant le programme Khmer [4]. L'algorithme sur lequel repose ce programme est un algorithme de calcul à passage unique (*simple pass*) permettant de ramener la couverture de lectures, issues d'une expérience de séquençage aléatoire, à un niveau défini et le plus uniforme possible le long du génome. La "normalisation digitale" effectuée par Khmer permet ainsi de réduire le jeu de données de départ en éliminant les lectures les plus redondantes ce qui permet de réduire le temps et la quantité de mémoire nécessaire pour effectuer l'assemblage sans en dégrader la qualité. Pour mesurer l'abondance d'une lecture celle-ci est découpée en mots (que l'on appelle des *k-mers* et dont la taille a été fixée à 20 dans les expériences réalisées (paramètre *k*)) dont on regarde la distribution dans toutes les lectures observées jusque-là. Si la valeur médiane de la distribution des différents *k-mers* composant la lecture est inférieure à un seuil fixé par l'utilisateur (que nous avons fixé à 20 (paramètre C))

cette lecture est éliminée (car elle contient en majorité des *k-mers* qui ont souvent été observés et correspond donc à une région du génome déjà bien échantillonnée), sinon elle est conservée et les comptages de *k-mers* que comprend cette lecture sont incrémentés. De plus le programme travaille à mémoire fixe et la taille maximale utilisable est à définir par l'utilisateur. En sortie on obtient les lectures au format fasta (peu importe le format d'entrée donné) et un script python permet de filtrer les lectures pour séparer en deux fichiers distincts les lectures pairées dont la paire est complète et les lectures qui ne sont plus pairées. Au final, avec ces paramètres, la couverture des souches bactériennes utilisées passe de 100 X à un chiffre compris entre 35 et 38 X.

Assemblage des génomes avec Velvet

Velvet est un programme d'assemblage *de novo* populaire basé sur la construction d'un graphe de *de Bruijn* [5]. Le graphe de *de Bruijn* est un formalisme permettant de représenter sous forme de nœuds toutes les sous-séquences de taille *k* (fixé par l'utilisateur). Ce paramètre *k* (pour *k-mer*) est assez difficile à régler et il est recommandé d'en tester plusieurs valeurs et de garder le meilleur assemblage obtenu (d'après le paramètre d'optimisation choisi). En effet, si cette valeur est trop faible par rapport à la taille des lectures une utilisation de la mémoire trop importante risque de faire tomber en erreur le programme. À l'inverse, si la valeur de ce paramètre est trop élevée, l'assemblage en résultant risque d'être de mauvaise qualité.

Afin de lancer Velvet le script perl VelvetOptimiser a été utilisé [6]. Ce script permet d'optimiser le lancement de Velvet, notamment dans la gestion des cœurs disponibles et le lancement de plusieurs instances de Velvet simultanément pour différentes valeurs de *k-mers*. Tous les assemblages ont été lancés de façon à n'utiliser que 4 cœurs. En effet le nombre d'instance de Velvet que VelvetOptimiser pouvait lancer était limité à 2 (paramètre t) et le nombre de cœurs pour chaque instance de Velvet était limité à 2 (à l'aide d'un script bash). Afin de ne pas utiliser trop de mémoire et de s'assurer du bon fonctionnement du programme la taille minimale de *k-mers* testée correspond à 40% de la taille moyenne des lectures soit 40 dans notre cas (automatiquement ajusté à 39 par Velvet qui ne gère que les valeurs impaires de k) et la taille maximum correspond à 75% de la taille moyenne des lectures soit 75. De plus le pas entre deux valeurs consécutives de k a été réglé à 4 et le paramètre d'optimisation retenu est le N50 (paramètre par défaut du programme). Néanmoins, suivant la quantité de mémoire vive disponible sur la machine dont on dispose, ces paramètres devront être adaptés.

En effet avec ceux-ci les assemblages de génomes d'*Escherichia coli* utilisaient environ 8 GO de mémoire vive. En entrée deux fichiers sont nécessaires au programme : le premier contient les lectures pairées et le second les lectures non-pairées (suite à la normalisation digitale faite à l'aide de Khmer). En sortie le programme permet de récupérer un fichier multi-fasta contenant les différents contigs obtenus. Avant de passer à l'étape suivante un filtre est appliqué pour éliminer les contigs dont la taille est inférieure à 1kb.

Prédiction des gènes avec Prodigal

Prodigal (pour ***Pro****karyotic* ***Dy****namic Programming* ***G****enefinding* ***Al****gorithm*) est un programme de prédiction de séquences codantes procaryotes écrit en C [7]. En plus d'être très rapide (environ 2 minutes pour analyser un génome bactérien de 5 Mb) ce programme a un faible taux de faux positifs et permet de récupérer les gènes prédits directement sous forme de séquences protéiques au format multi-fasta. Le faible taux de faux positifs (on détecte une séquence codante à tort) est ici un critère important. En effet si jamais les faux positifs ne correspondent pas dans l'étape suivante à des gènes uniques ils seront utilisés dans le calcul des DGA et fausseront le résultat.

Regroupement des gènes avec OrthoMCL

OrthoMCL pour *Orthologs Markov Cluster algorithm* est un programme permettant de regrouper des séquences protéiques homologues [8]. Le programme prend en entrée les fichiers contenant les protéines des génomes à analyser au format multi-fasta (un fichier par génome) et fournit en sortie des groupes de gènes homologues. Ces groupes, contiennent les gènes orthologues (dont l'origine correspond à un événement de spéciation) ainsi que les gènes paralogues récents (dont l'origine correspond à un évènement de duplication). Après traitement du fichier de sortie qui contient pour chaque groupe la liste des gènes le composant, on obtient un tableau où chaque colonne décrit un génome et chaque ligne décrit un groupe de gènes homologues (orthologues et paralogues) (annexe 2). Chaque case de ce tableau contient le nombre de représentants d'un groupe présent dans un génome et permet donc d'identifier aisément les gènes du génome de base, les gènes accessoires et les gènes uniques.

Typage bactérien par approche pangénomique

Le but de cette étape est de discriminer les génomes bactériens donnés en entrée du pipeline en fonction de leur contenu en gènes. Dans un premier temps nous avons testé l'approche décrite par Hall et al. qui est basée sur le calcul de la DGD en fonction de la

présence ou de l'absence des gènes du génome distribué. Dans notre cas, seul les gènes du génome accessoire ont été utilisés. En effet, dans les 3 expériences réalisées, aucune souche n'était présente en un unique exemplaire, on sait donc de manière certaine que les gènes uniques détectés sont des artefacts ou des gènes accessoires considérés à tort comme des gènes uniques. Pour cette raison on parlera dans la suite de ce document de distance des gènes accessoires (DGA) et non plus de DGD. De plus, suite à l'utilisation du programme OrthoMCL, nous ne disposons pas exactement de la même information. En effet dans l'approche décrite précédemment, il suffit pour chaque orthologue putatif de déterminer s'il est présent ou absent dans le génome pour pouvoir calculer la distance entre deux génomes. Dans notre cas nous disposons, en plus, d'informations sur la présence de gènes paralogues récents (pour lesquels on suppose donc qu'ils ont la même fonction). Le calcul de la DGA entre chaque paire de génomes a donc été, dans un second temps, légèrement modifié afin de tenir compte du nombre de copies des gènes. Pour chaque gène le score qui lui est attribué vaut 1 si le gène est présent en 1 copie dans les deux génomes ou s'il est absent dans les deux génomes. Si le gène est présent en nombre a et b dans les deux génomes son score est égal au ratio suivant : $\min(a,b)/\max(a,b)$. Enfin si un gène est présent chez un génome (peu importe le nombre de copie) et absent chez l'autre génome son score vaut 0. Une fois tous les scores calculés, ils sont sommés et normalisés par la somme du nombre maximum de copies observées de chaque gène utilisé. Pour ce calcul deux sous-ensembles de groupes de gènes peuvent être utilisés : le premier contient uniquement les gènes du génome accessoire (filtrage complet) alors que le second contient en plus les gènes du génome de base à condition que le nombre d'exemplaire de ces gènes varie entre les souches (filtrage partiel). En effet dans le cas où un gène est présent chez tous les génomes étudiés mais avec un nombre de copies différents on peut utiliser cette information avec la modification du mode de calcul de la DGA. Nous avons donc voulu vérifier si cette information supplémentaire nous aide à discriminer plus précisément les différents génomes testés.

Résultats

Etude de la faisabilité de l'approche

Pour rappel, la première expérience a été réalisée à partir des 9 génomes d'*Escherichia coli* (3 génomes de la souche BL21, 3 de la souche K12-MG1655 et 3 de la souche O157:H7-Sakai). Dans un premier temps nous avons appliqué le calcul, pour chaque paire possible de génomes, de la DGA telle que décrite dans la publication de B. G. Hall et, dans un second

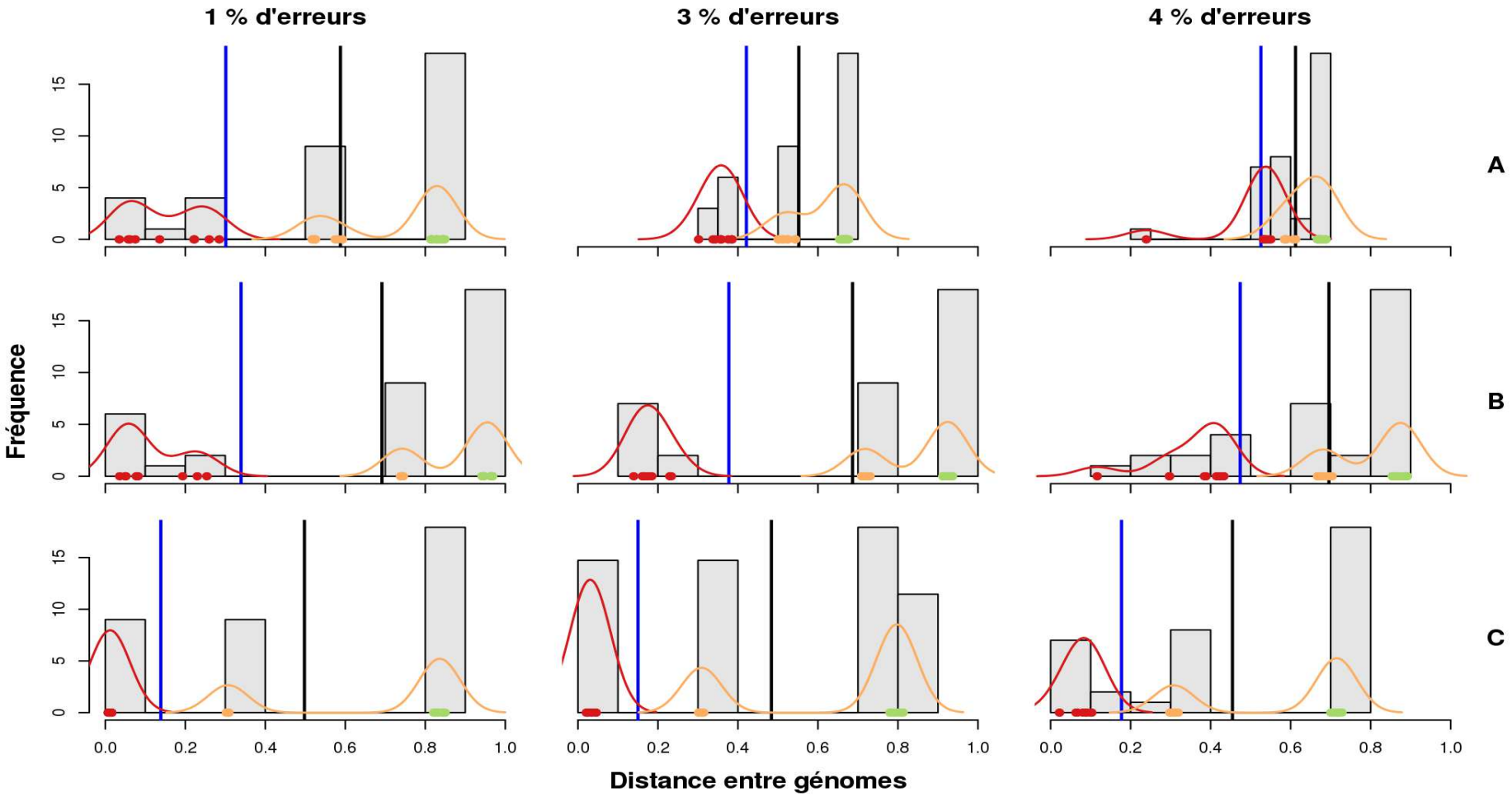


Figure 1 - Histogrammes des DGA obtenues à partir de lectures contenant de 1 à 4 % d'erreurs, avec notre méthode de calcul et un filtrage partiel des gènes (A), notre méthode de calcul et un filtrage complet des gènes (B) et la méthode de Hall (C). Pour chaque histogramme la barre verticale noire correspond à la moyenne des DGA observées, la barre verticale bleue correspond au seuil en dessous duquel on considère que deux génomes sont issus d'une même souche. Les points rouges correspondent à des distances intra-souches alors que les points orange et vert correspondent à des distances inter-souches (distances entre souches BL21 et K12-MG1655 en orange et distances entre O157 :H7-Sakai et BL21 ou K12MG1655 en vert). Enfin la courbe rouge est la courbe de densité des distances intra-souches alors que la courbe orange est la courbe de densité des distances inter-souches.

temps, le calcul prenant en compte le nombre d'exemplaires de chaque gène que nous avons mis au point (avec filtrage complet ou partiel des gènes).

Les résultats obtenus avec chacune de ces méthodes et avec des taux d'erreurs dans les données de départ allant de 1 à 4 % sont présentés dans la figure 1 ci-contre. Quand on utilise le filtrage complet des gènes, avec le calcul prenant en compte le nombre de copies des gènes, les résultats sont similaires avec ceux obtenus par la méthode décrite par Hall et al. Dans les deux cas on parvient à regrouper ensemble les génomes provenant d'une même souche, pour des taux d'erreur dans les lectures allant jusqu'à 4 %, alors qu'avec le filtrage partiel des gènes le typage ne fonctionne que jusqu'à 3 %. On observe également que les distances inter-souches sont concentrées en deux point : un groupe correspond aux distances entre les souches BL21 et les souches K12-MG1655 (en orange) tandis que l'autre groupe corresponds aux distances entre des souches O157 :H7-Sakai et BL21 ou K12-MG1655 (en vert). En effet lorsque l'on observe la matrice des distances obtenue (figure 2 ci-dessous) on observe que les distances entre les groupes BL21 et K12-MG1655 sont plus faibles que les autres distances inter-souches.

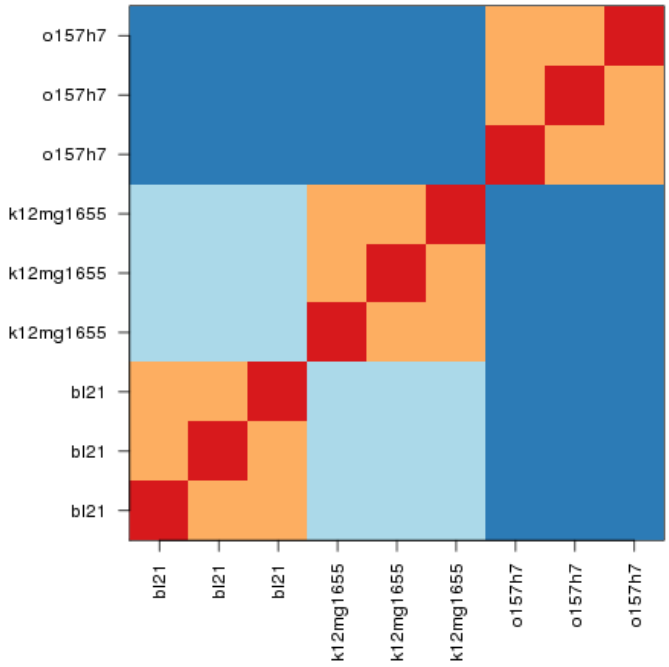


Figure 2 - Matrice des DGA. Matrice obtenue avec des lectures de départ contenant 3 % d'erreurs, notre méthode de calcul des DGA et un filtrage partiel des gènes. Les cases rouges correspondent aux distances égales à 0, en orange les cases correspondent aux distances comprises entre 0 et le seuil de décision (égal à 0,42), en bleu clair on retrouve les cases dont la distance est supérieur au seuil et inférieur à la moyenne des DGA (égale à 0,55) et en bleu foncé on retrouve les distances dont la valeur est supérieur à la moyenne des DGA.

De plus, avec le filtrage complet des gènes, l'écart entre les distances intra-souches et les distances inter-souches est plus élevé. En effet à 1 % d'erreurs dans les données de départ, cet écart est 0.2 avec la méthode de B. G. Hall alors qu'il est de 0.45 avec la notre. Jusqu'à 3 % d'erreurs ces écarts restent constants et sont donc toujours à l'avantage de notre méthode. En revanche à partir de 4 % d'erreurs dans les données de départ l'écart diminue fortement dans les deux cas : il passe à 0.16 avec la méthode de Hall et à 0.2 avec la notre. Ces résultats laissent penser que notre méthode est plus précise pour séparer des souches relativement proches. En revanche si on utilise, en plus des gènes accessoires, les gènes du génome de base dont le nombre de copies varie entre génomes, le typage ne fonctionne que jusqu'à 3 % d'erreurs dans les lectures de séquences et les écarts entre les différents groupes identifiés sont plus faibles que lorsque l'on utilise uniquement les gènes accessoires. Néanmoins, dans ce cas, on peut observer qu'une des distances intra-souche est faible par rapport aux autres. Par conséquent cette distance entraîne une augmentation de l'écart-type des DGA calculées et donc un abaissement du seuil. En calculant autrement le seuil on peut penser qu'il serait possible de séparer les distances intra-souches des distances inter-souches car elles ne se mélangent pas.

A partir de 5 % d'erreurs dans les lectures les distances intra et inter-souches (que ce soit les distances entre les souches BL21 et K12-MG1655 ou les distances entre les souches O157:H7-Sakai et BL21 ou K12-MG1655) sont complètement mélangées dans tous les cas et l'approche pangénomique pour le typage n'est plus envisageable (résultats présentés en annexe 4).

Le nombre de gènes uniques détectés dans les génomes testés semble utilisable comme indicateur de la fiabilité des résultats obtenus. En effet si deux génomes sont identifiés comme provenant de la même souche ils ont, en théorie, le même contenu en gènes et ne doivent pas contenir de gènes uniques. Ainsi à 1 % d'erreurs dans les lectures le nombre de gènes uniques dans les 9 génomes testés varie de 3 à 49 (le nombre moyen de gènes détecté par génome est de 4615). À 3 % d'erreurs ce nombre varie de 32 à 70 (pour 4755 gènes détectés par génome en moyenne) et à 4% d'erreurs dans les lectures, c'est à dire quand le typage commence à montrer des signes de faiblesse, ce nombre varie entre 30 et 259 (et est supérieur à 150 pour 7 des 9 génomes testés) pour 5704 gènes détectés par génome en moyenne. Enfin à 5% d'erreurs dans les lectures, c'est-à-dire quand le typage ne fonctionne plus du tout, le nombre de gènes uniques varie de 110 à 955 et est supérieur à 700 dans 6 cas sur 9. On peut donc en déduire que, quand le nombre de gènes uniques détectés dans chaque génome est trop

important (supérieur à 100), les résultats obtenus ne garantissent aucune certitude. De plus cette information peut être utilisée indépendamment du fait que l'on connaisse à l'avance quelles sont les souches testées. En effet à partir du moment où deux génomes ont une DGA inférieure au seuil fixé on peut douter du résultat si ces deux génomes ont un nombre trop important de gènes uniques.

Cette expérience a ensuite été étendue à 14 souches d'*Escherichia coli* pour lesquelles 2 jeux de lectures à 1% d'erreurs ont été générés. Les résultats obtenus avec notre méthode de calcul des DGA et un filtrage partiel des gènes sont présentés figure 3 (ci-dessous).

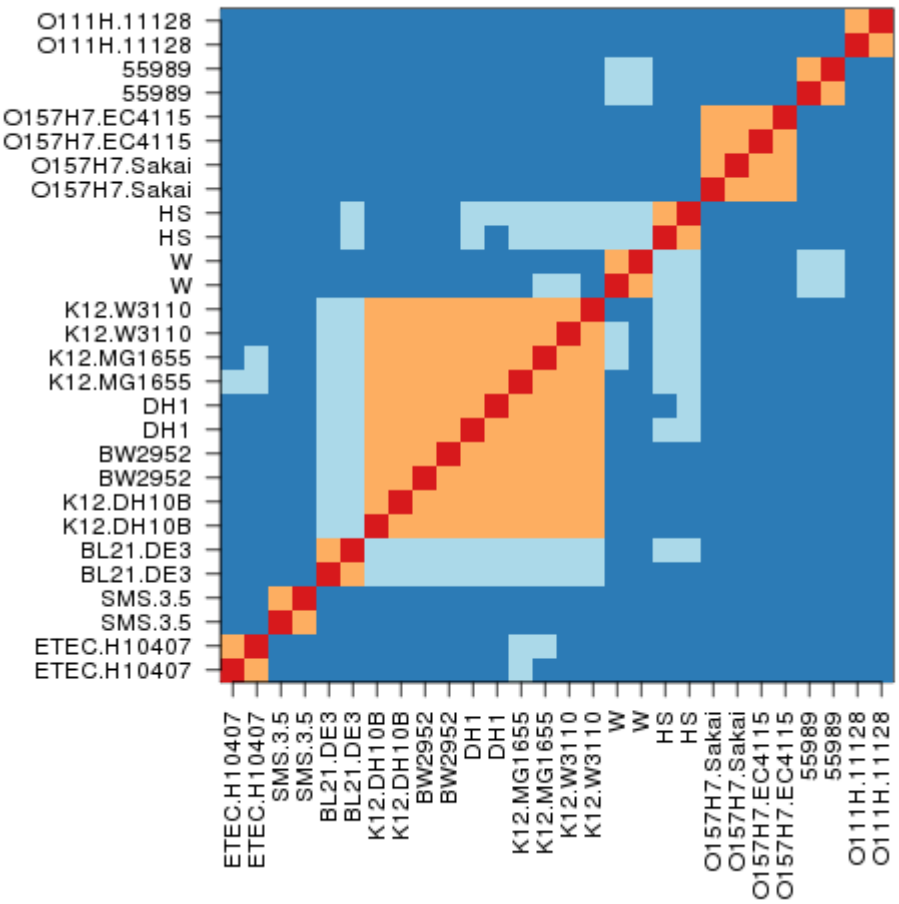


Figure 3 - Matrice des DGA. Matrice obtenue avec des lectures de départ contenant 1 % d'erreurs, notre méthode de calcul des DGA et un filtrage partiel des gènes. Les cases rouges correspondent aux distances égales à 0, en orange les cases correspondent aux distances comprises entre 0 et le seuil, en bleu clair on retrouve les cases dont la DGA est supérieure au seuil et inférieur à la moyenne des DGA et en bleu foncé on retrouve les DGA supérieures à la moyenne des DGA.

Comme on peut l'observer sur la figure 3, tous les réplicats issus d'une même souche sont regroupés ensemble. Il en va de même pour des réplicats issus de souches différentes : il n'est pas possible de distinguer les souches K12-DH10B, K12-MG1655, K12-W3110, DH1 et

BW2952. De plus le faible nombre de gènes uniques détectés dans les réplicats de ces souches (de 5 à 14) ne permet pas de se méfier du résultat. De même on ne parvient pas à distinguer la souche O157:H7-Sakai de la souche O157:H7-EC4115. Le nombre de gènes uniques détectés dans ces souches est plus élevé (de 62 à 88) mais au vu des résultats obtenus précédemment il ne permet pas non plus de se méfier du résultat. D'après le dendrogramme de l'espèce *Escherichia coli* disponible sur le site du NCBI, les souches que l'on ne parvient pas à distinguer les unes des autres sont proches entre elles (annexe 3). Ainsi les distances entre les 9 groupes que l'on parvient à créer sont relativement élevées comparées aux distances entre les souches composant les 2 groupes erronés. De plus, quand on observe la répartition des distances intra- et inter-souches (figure 5 ci-dessous), on constate qu'il n'est pas possible de distinguer complètement ces distances même si les résultats pourraient être améliorés avec un seuil plus sévère.

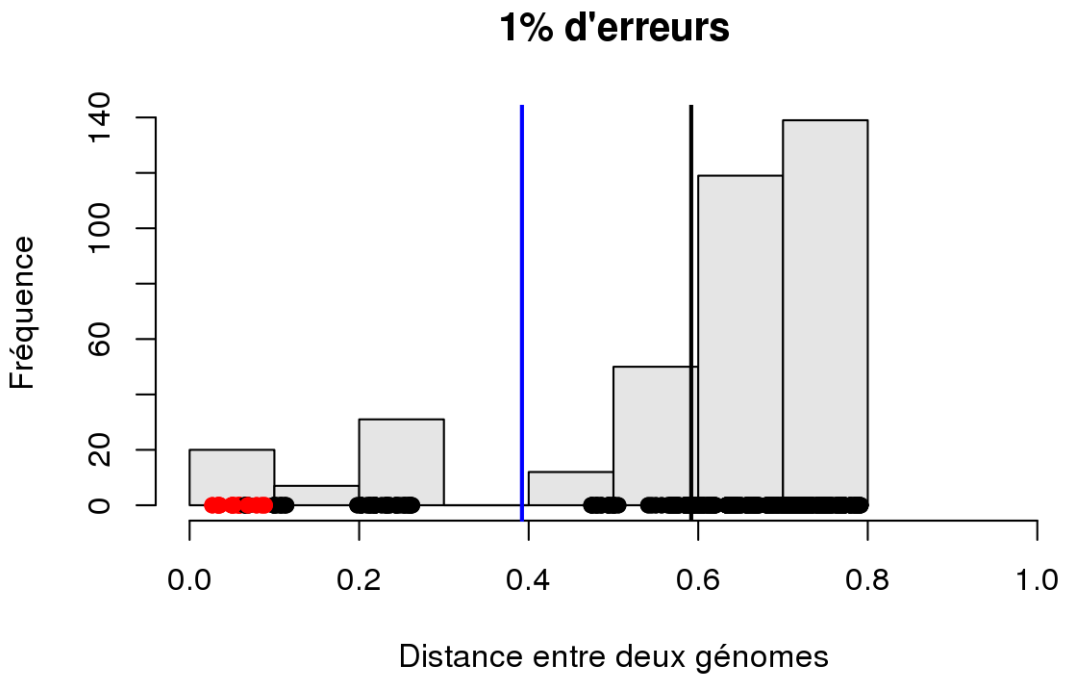


Figure 4 - Histogramme des DGA obtenues. Résultats obtenue avec des lectures contenant 1% d'erreurs, notre méthode de calcul et un filtrage partiel des gènes. La barre verticale noire correspond à la moyenne des DGA observées, la barre verticale bleue correspond au seuil en dessous duquel on considère que deux génomes sont issus d'une même souche. Les points rouges représentent des DGA intra-souches et les noirs des DGA inter-souches.

Etude de la sensibilité de la méthode

Les résultats obtenus avec 9 génomes issus de 3 souches différentes *d'Escherichia coli* démontrent la faisabilité de l'approche pangénomique pour faire du typage bactérien.

Néanmoins les résultats obtenus avec l'expérience étendue à 28 génomes de 14 souches différentes laissent à penser que cette approche n'est pas suffisamment résolutive pour distinguer des souches très proches entre elles. Pour confirmer ces résultats une expérience similaire a été réalisée à partir de souches bactériennes de la bactérie *Bacillus anthracis* qui est connue pour sa faible variabilité pangénomique.

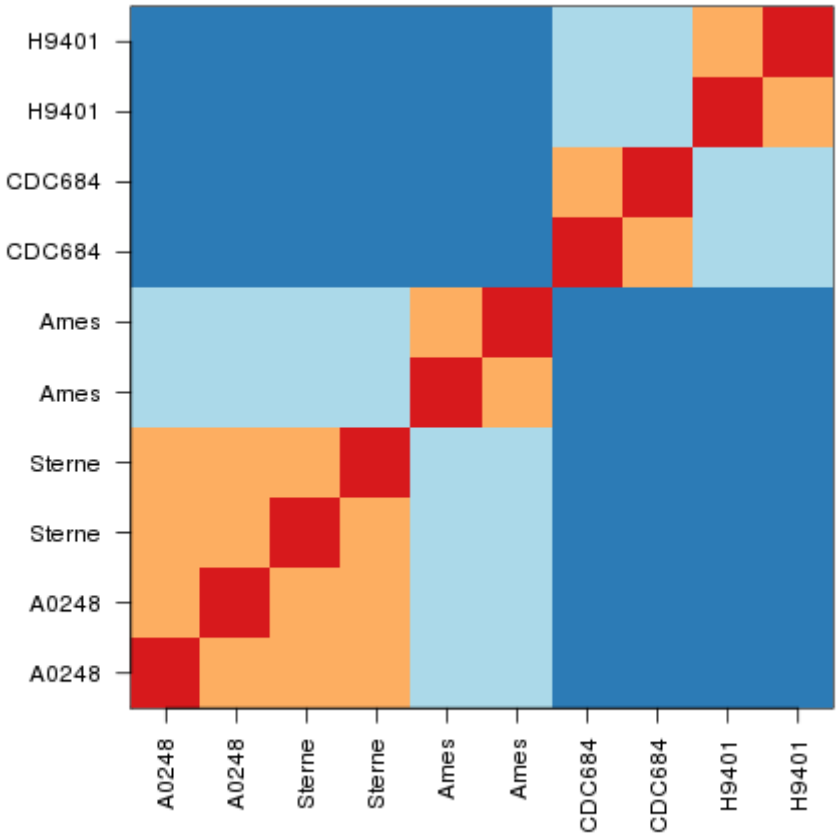


Figure 5 - Matrice des DGA. Matrice obtenue avec des lectures de départ contenant 1 % d'erreurs, notre méthode de calcul des DGA et un filtrage partiel des gènes. Les cases rouges correspondent aux distances égales à 0, en orange les cases correspondent aux distances comprises entre 0 et le seuil, en bleu clair on retrouve les cases dont la DGA est supérieure au seuil et inférieur à la moyenne des DGA et en bleu foncé on retrouve les DGA supérieures à la moyenne des DGA.

Cette expérience a été réalisée à partir de 5 souches de la bactérie pour lesquelles 2 jeux de données ont été générés avec 1 % d'erreurs. Avant même de procéder au typage on remarque que pour cette bactérie le typage n'est fait que sur la base de 31 groupes de gènes accessoires ce qui est relativement faible comparé au nombre de groupes de gènes du pangénome de ces 5 souches, détecté à l'aide d'OrthoMCL, qui est de 3848. Parmi ces 3848 groupes, 3650 correspondent à des groupes de gènes du génome de base et sont éliminés et 37

correspondent à des gènes uniques et sont donc éliminés eux aussi. Les 130 groupes de gènes restant sont des gènes du génome de base (présent dans toutes les souches observées) dont le nombre de copies varie entre les souches et peuvent donc être utilisés dans le calcul avec notre méthode (filtrage partiel). Comme on peut le voir sur la matrice obtenue avec notre méthode (figure 5) et en prenant en compte les 130 groupes de gènes supplémentaires utilisables, tous les réplicats issus d'une même souche sont regroupés ensemble. Les réplicats des souches Ames et A0248 sont également regroupés ensemble. Avec la méthode décrite par B. G. Hall les réplicats de la souche H9401 ne sont pas regroupés ensemble et la distinction entre les réplicats des souches Ames, A0248 et Sterne n'est pas faite correctement (résultats non présentés). Les résultats sont les mêmes avec notre méthode quand on ne prend en compte que les gènes du génome accessoire (filtrage complet).

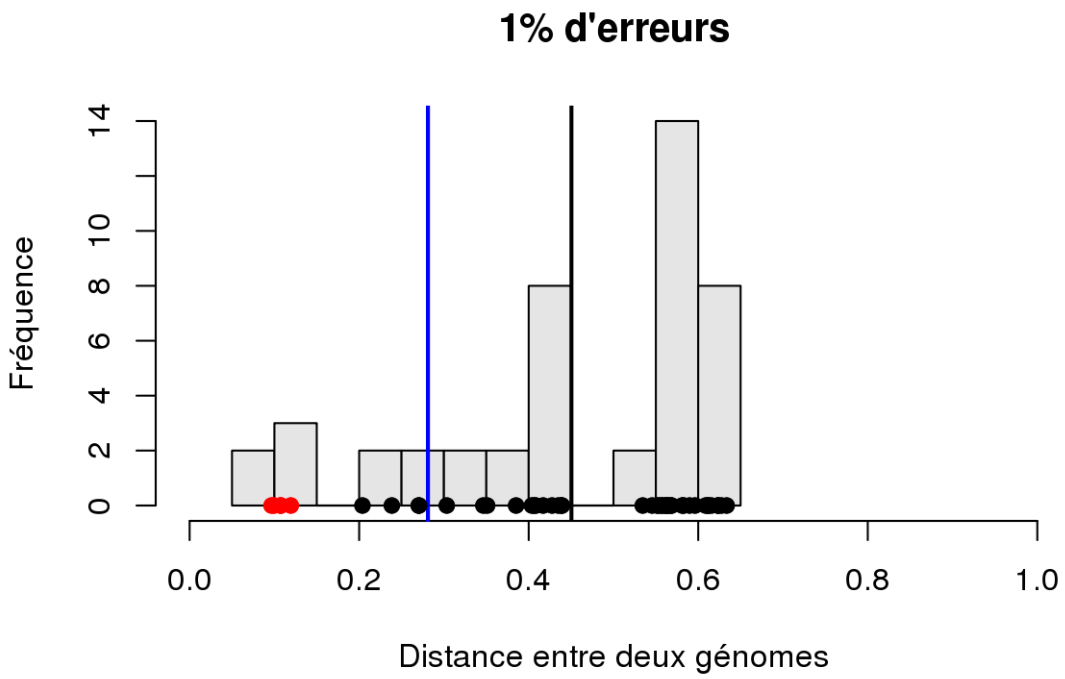


Figure 6 - Histogramme des DGA obtenues. Ici les lectures contiennent 1% d'erreurs et la méthode de calcul utilisé est celle que nous avons mise au point et appliquée sur les gènes accessoires et les gènes de base présentant une variation du nombre de copies. La barre verticale noire correspond à la moyenne des DGA observées, la barre verticale bleue correspond au seuil en dessous duquel on considère que deux génomes sont issus d'une même souche.

De plus, quand on regarde la répartition des distances obtenues avec notre méthode et en prenant en compte les gènes de base présentant une variation du nombre de copies (figure 6) on se rend compte que les distances inter-souches ne se mélangent pas aux distances intra-souches, ce qui laisse à penser qu'en calibrant autrement le seuil de décision il est possible de

les séparer. En effet dans ce cas avec un seuil égal à la moyenne à laquelle on soustrait 1,5 fois l'écart-type on arriverait à distinguer les distances intra-souches des distances inter-souches. Sans prendre en compte les gènes du génome de base dont le nombre de copie varie entre les souches, les distances intra- et inter-souches sont mélangées, ce qui annihile tout espoir de pouvoir le séparer. La même constatation prévaut quand les DGA sont calculées à l'aide de la méthode de B. G. Hall (résultats non montrés dans ce rapport). Ces résultats démontrent donc la viabilité de l'approche, en effet même avec des souches proches entre elles on peut réussir à les typer correctement à condition de faire un filtrage partiel des gènes. La calibration d'un seuil informatif est donc une étape cruciale qui nécessite encore du travail. Néanmoins, même en calibrant autrement le seuil de séparation, les DGA inter- et intra-souches obtenues au cours de l'expérience contenant les 28 génomes de 14 souches d'*Escherichia coli* ne pourraient être parfaitement séparées car elles se mélangent (sur la figure 4 on peut constater la présence d'un point noir parmi les points rouges) mais avec un seuil mieux calibré, le nombre de faux positifs serait moindre.

Conclusion

D'après les résultats montrés précédemment le typage par approche pangénomique est une alternative crédible aux méthodes existantes. L'avantage de cette méthode est qu'elle permet dans la plupart des cas de se baser sur un grand nombre de gènes là ou les méthodes traditionnelles (type MLST) n'en utilisent qu'un petit nombre. Cet avantage explique la meilleure résolution de l'approche pangénomique. De plus en prenant en compte les gènes paralogues grâce à OrhtoMCL il est possible d'augmenter le nombre de gènes utilisables : en effet en plus des gènes du génome accessoire on peut prendre en compte les gènes du génome de base à condition que ceux-ci ne soient pas présents dans le même nombre de copies dans les différents génomes à tester. Néanmoins il reste à déterminer le meilleur moyen de fixer le seuil en dessous duquel on décide que deux génomes proviennent de la même souche. En effet avec le seuil tel qu'il est fixé par B. G. Hall, on constate la présence de faux-positifs (deux génomes sont considérés, à tort, comme provenant d'une même souche) dans l'expérience sur *Bacillus anthracis*. Il est donc crucial de bien calibrer dynamiquement ce seuil. Une manière de procéder qui mériterait d'être testée serait de fixer le seuil à partir des distances intra-souches. Or pour disposer de distances intra-souches certaines (en théorie on ne possède pas cette information à l'avance puisqu'on cherche à la déterminer) on pourrait diviser, pour chaque séquençage, les données en deux lots qui fourniraient ainsi une distance

intra-souches. Pour agir ainsi, la couverture de départ des lectures doit être assez importante (environ 100 X). En effet suite à l'utilisation de Khmer la couverture effective des lectures utilisées pour les expériences décrites ici était comprise entre 35 et 38 X et s'avérait suffisante pour reconstruire les génomes. On peut donc imaginer cliver un jeu de lecture dont la couverture est de 100 X en deux sans trop dégrader la qualité de l'assemblage.

Pour confirmer ces résultats il reste à tester le pipeline sur de vraies données de séquençage. En effet l'utilisation de lectures générées artificiellement n'est pas exempte de reproches. Avec cette méthode, les lectures sont réellement distribués uniformément le long du génome (ce qui en pratique n'est jamais le cas). De plus les biais dus aux appareils de séquençage à haut-débit ne peuvent pas être reproduits à l'aide d'un simple script. Enfin les données étant générées au format fasta il n'est pas possible d'utiliser des programmes de correction de lectures car on ne possède pas d'information de qualité. Il n'est donc pas impossible qu'un tel programme de correction doive être intégré au pipeline auquel cas le programme Reptile semble être une bonne alternative [9]. Néanmoins l'utilisation d'un tel programme augmenterait considérablement les temps de calculs sans possibilité de paralléliser le traitement des différents génomes testés. Par ailleurs, de par son fonctionnement, cette méthode nécessite que tous les génomes testés ne proviennent pas de la même souche. En effet, le seuil est calculé à partir de la moyenne des distances entre les génomes et de l'écart-type. Si tous les génomes proviennent de la même souche on aura forcément parmi les résultats des faux négatifs. Cette limitation implique que l'on ne peut pas utiliser uniquement des génomes dont on ne connaît pas la souche ce qui augmente les temps de calculs.

Au final les travaux décrits dans ce rapport démontrent la viabilité du projet. La calibration du seuil de séparation des distances intra- et inter-souches peut (et doit) encore être améliorée et des tests doivent être effectués afin de déterminer s'il est nécessaire d'intégrer, ou non, un outil de correction de lectures. De plus, les temps de calculs étant un critère important du choix des programmes à intégrer au pipeline, on peut s'interroger sur la pertinence de mettre au point une application web reposant sur un cluster de calcul qui permettrait : d'effectuer l'assemblage des génomes et les blasts protéiques en parallèle ainsi qu'optimiser les paramètres d'assemblage (nombre de *k-mer* que l'on peut tester en simultanée pour chaque assemblage) ce qui permettra un gain de temps non-négligeable. Néanmoins il faudrait s'assurer, avant cela, que le temps nécessaire au téléchargement des données (*upload*) ne soit pas trop important.

Bibliographie

[1] B. G. Hall, G. D. Ehrlich, F. Z. Hu. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequences typing. *Microbiology*. **2010**. 165:1060-1068

[2] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, B. Shen. A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*. **2011**. Doi:10.1371/journal.pone.0017915

[3] P. Keim, J. M. Gruendike, A. M. Klevytska, J. M. Schupp, J. Challacombe, R. Okinaka. The genome and variation of *Bacillus anthracis. Molecular Aspects of Medecine*. **2009**, 30(6):397–405

[4] C. T. Brown, A. Howe, Q. Zhang, A. B. Pyrkosz, T. H. Brom. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. Version preprint accessible à l'adresse suivante : http://ged.msu.edu/papers/2012-diginorm

[5] D. R. Zerbino, E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. **2008**. 18:821-829

[6] S. Gladman. Velvet Optimiser. Accessible à l'adresse suivante : http://vicbioinformatics.com/software.velvetoptimiser.shtml

[7] D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **2010**. 11(1):119.

[8] L. Li, C. J. Stoeckert Jr., D. S. Roos. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res*. **2003.** 13:2178-2189.

[9] X.Yang, S. P. Chockalingam, S. Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*. Advanced Access published April 6, **2012**. Doi :10.1093/bib/bbs015

Annexes

Annexe 1 : Numéros d'accession *GenBank* des génomes utilisés

Espèce	Nom de la souche	Numéro d'accession
E. coli	55989	NC_011748.1
E. coli	BL21-DE3	NC_012947.1
E. coli	BW2952	NC_012759.1
E. coli	DH1	NC_017638.1
E. coli	ETEC-H10407	NC_017633.1
E. coli	HS	NC_009800.1
E. coli	K12-DH10B	NC_010473.1
E. coli	K12-MG1655	NC_000913.2
E. coli	K12-W3110	NC_007779.1
E. coli	O111H-11128	NC_013364.1
E. coli	O157:H7-EC4115	NC_011353.1
E. coli	O157:H7-Sakai	NC_002695.1
E. coli	SMS-3-5	NC_010498.1
E. coli	W	NC_017635.1
B. anthracis	A0248	NC_012659.1
B. anthracis	Ames	NC_003997.3
B. anthracis	CDC684	NC_012581.1
B. anthracis	H9401	NC_017729.1
B. anthracis	Sterne	NC_005945.1

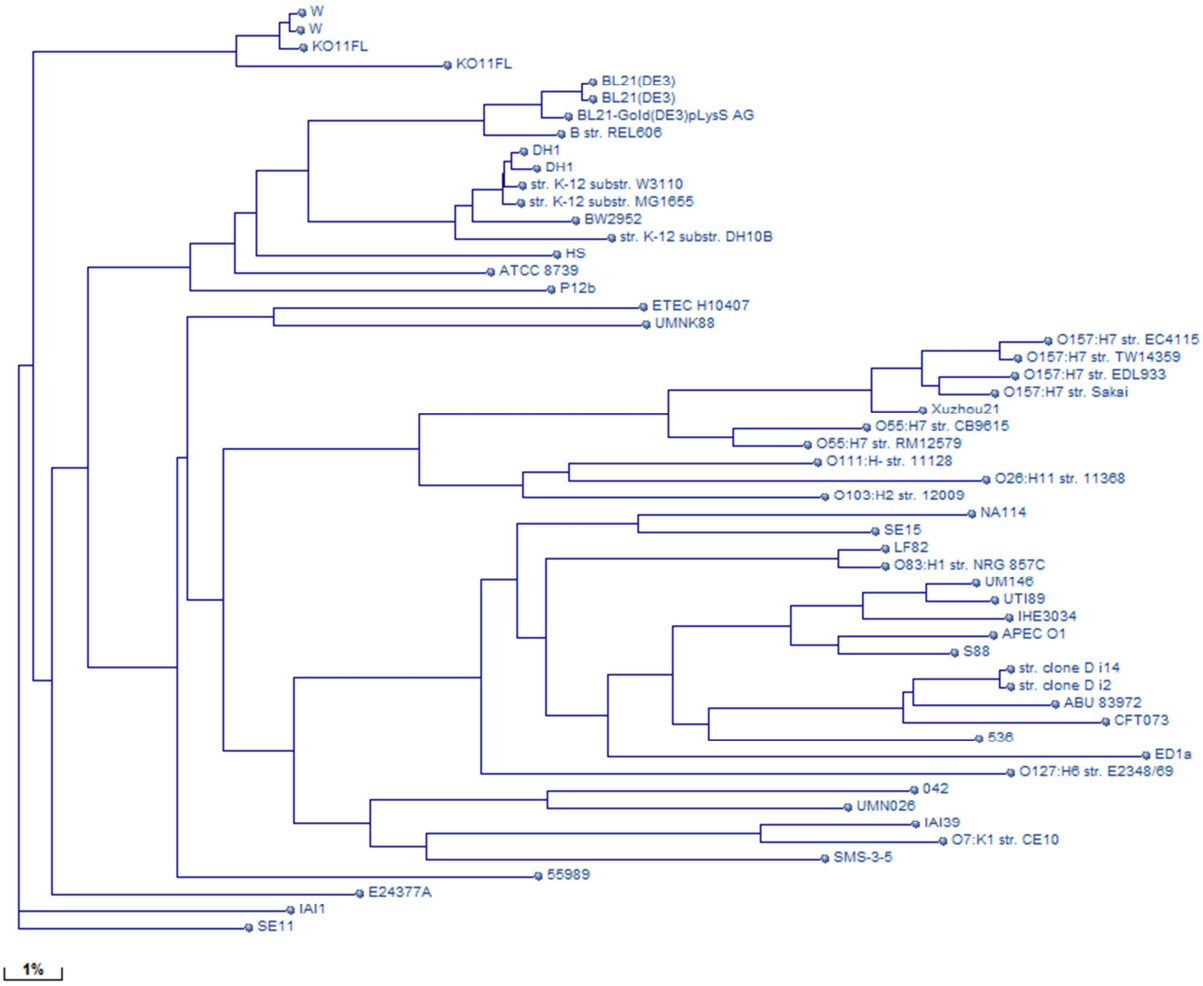
Annexe 2 – Format du fichier de sortie (après traitement) d'OrthoMCL

Groupe	Génome 1	Genome 2	Génome 3
A	1	1	1
B	2	2	2
C	1	1	1
D	2	2	1
E	0	0	1
F	2	0	0
G	1	1	0
H	1	2	0

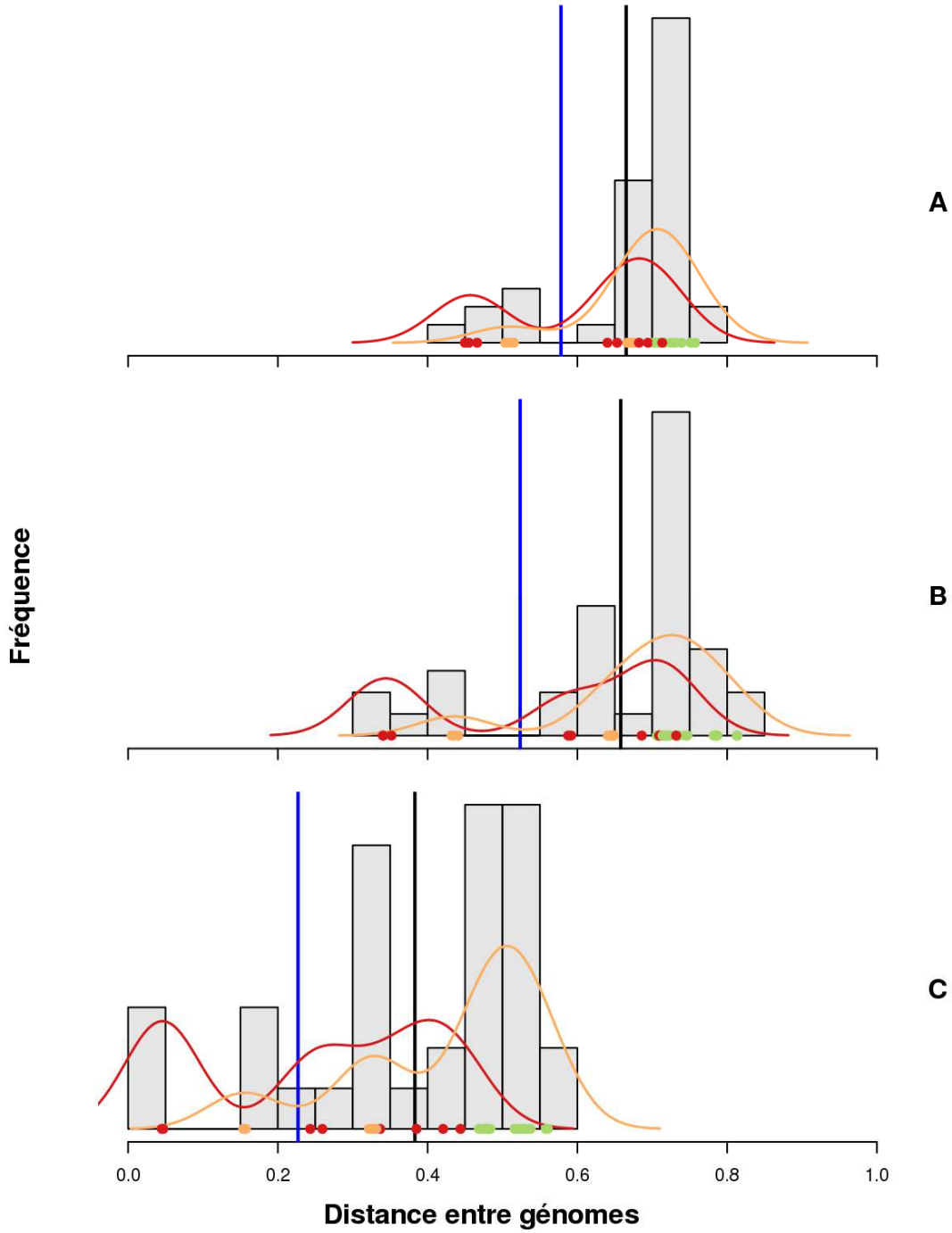
Dans cet exemple les groupes A, B et C correspondent à des gènes du génome de base. Le groupe D correspond à un gène de base dont le nombre de copie varie dans les souches étudiées. Les groupes E et F correspondent à des gènes uniques (le groupe n'est présent que dans une seule souche peu importe le nombre d'exemplaire du gène dans le groupe). Enfin les groupes G et H correspondent à des gènes du génome accessoires (c'est-à-dire qu'ils sont présent dans plusieurs, mais pas dans tous, génomes).

Annexe 3 – Dendrogramme de l'espèce *Escherichia coli*

Ce dendrogramme est disponible sur la page génome de l'espèce du site du NCBI à l'adresse suivante : http://www.ncbi.nlm.nih.gov/genome?term=escherichia%20coli



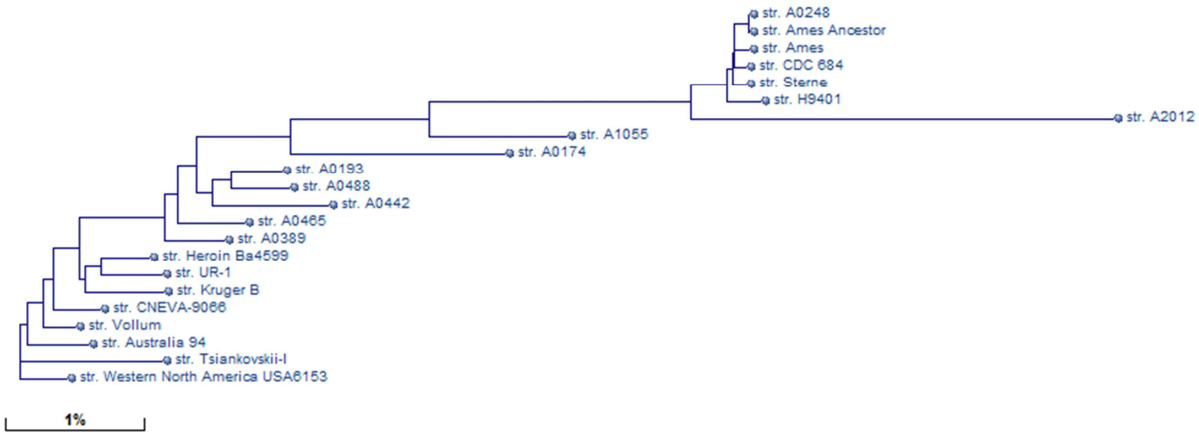
Annexe 4 – Histogrammes des distances des génomes accessoires obtenues quand les lectures de départ contiennent 5 % d'erreurs



(A) Notre méthode de calcul et filtrage partiel des gènes, (B) Notre méthode de calcul et filtrage complet des gènes. (C) Méthode de Hall (ne peut être appliquée que sur le filtrage complet des gènes). Le trait vertical noir correspond à la moyenne des DGA et le trait vertical bleu au seuil en dessous duquel on considère que deux génomes proviennent d'une même souche.

Annexe 5 – Dendrogramme de l'espèce *Bacillus anthracis*

Ce dendrogramme est disponible sur la page génome de l'espèce du site du NCBI à l'adresse suivante : http://www.ncbi.nlm.nih.gov/genome?term=bacillus%20anthracis



Les génomes de la bactérie *Bacillus anthracis* que nous avons utilisés sont tous situés dans la partie supérieur droite de ce dendrogramme et présentent entres elles moins de 0,5 % de différence (le calcul est basé sur des blasts génomiques).

Résumé

Les bactéries sont impliquées dans beaucoup d'aspects de notre vie quotidienne ce qui en fait des organismes particulièrement étudiés par la communauté scientifique. Des Centres de Ressources Biologiques (CRB) ont été créés afin d'assurer la conservation et l'étude de ces bactéries. Le but du projet présenté dans ce document est de fournir un outil aidant les CRB à gérer au mieux leurs collections bactériennes. Pour cela l'unité Mathématique Informatique et Génome du centre INRA de Jouy-en-Josas souhaite mettre un place pipeline automatique de typage bactérien par approche pangénomique à partir de données de séquençage à haut-débit. En effet les progrès réalisés ces dernières années par les technologies de séquençage à haut-débit couplé à leur démocratisation offrent la possibilité de distinguer différentes souches avec une résolution jusque-là jamais atteinte. Le pipeline qui doit être mis au point devra tout d'abord reconstruire le pangénome des souches bactérienne étudiées. Il devra ensuite identifier les bactéries provenant d'une même souche à partir des répertoires de gènes reconstruits à l'étape précédente.

Abstract

Bacteria are involved in many aspects of our daily life. For this reason these organisms are very studied by the scientific community. The aim of the project described in this report is to develop an accurate and affordable typing tool to help Biological Resource Centers to manage their bacterial collections. To do this, the MIG unit of the Jouy-en-Josas INRA's center would like to develop an automatic typing tool by pangenomic approach pipeline with high throughput sequencing data. Indeed, the recent advanced and popularization of high throughput sequencing offers the opportunity to identify evolutionary events at an unprecedented scale and resolution. The pipeline will first reconstruct the genes repertoires of studied bacteria. These repertoires will then be used to identify the genomes which come from the same strain.