Détection d'organismes génétiquement modifiés (OGM) inconnus par analyse statistique de données de séquençage haut débit

Julie Hurel

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : Génétique, génomique et bio-informatique

Par
Julie HUREL

Détection d'organismes génétiquement modifiés (OGM) inconnus par analyse statistique de données de séquençage haut débit

Thèse présentée et soutenue à Ploufragan, le 13 novembre 2020
Unité de recherche : Anses Ploufragan-Plouzané-Niort, Unité Génétique Virale et Biosécurité

Rapporteurs avant soutenance :

Stéphane ROBIN	Directeur de recherche Agro paris Tech/INRAE, Paris, France
Annabelle DEJARDIN	Chargé de recherche, INRAE, Orléans, France

Composition du Jury :

Président :

Fabien NOGUÉ	Directeur de recherche, INRAE, Versailles, France

Examinateurs :

Stéphane ROBIN	Directeur de recherche, AgroParisTech/INRAE, Paris, France
Annabelle DEJARDIN	Chargé de recherche, INRAE, Orléans, France
Francis GALIBERT	Professeur émérite, Université de Rennes 1, France
Jean-Stéphane VARRÉ	Professeur, Université de Lille 1, France

Dir. de thèse :

Stéphanie BOUGEARD	Ingénieure de recherche, ANSES, Ploufragan, France

Co-dir. de thèse :

Fabrice TOUZAIN	Chargé de projet de recherche, ANSES, Ploufragan, France

Remerciements

Tout d'abord, je tiens à remercier Fabrice Touzain, Stéphanie Bougeard et André Jestin, mes directeurs de thèse pour leur encadrement, leur disponibilité, la confiance qu'ils m'ont accordé et leur aide au bon déroulement de ce projet. Fabrice, je te remercie encore pour ta patience indéfectible, ta bienveillance et ton soutien à toute épreuve. Merci pour ton implication sans failles dans la relecture de mon manuscrit mais aussi l'écriture de l'article. Stéphanie, merci d'avoir accepté de reprendre l'encadrement de cette thèse à la suite du départ d'André. Tes nombreux conseils avisés m'ont été d'une aide précieuse notamment lors de la construction du modèle de prédiction et de la relecture de mon manuscrit. Ton domaine de compétences étant parfois éloigné de mon travail, ton regard extérieur a permis de rendre ce manuscrit accessible à un plus grand nombre de personnes (pas uniquement des bioinformaticiens !).

Je souhaite remercier particulièrement Sophie Schbath pour ces nombreux conseils en statistiques combinatoires, son aide pour l'élaboration des formules de normalisation des distances et sa disponibilité malgré ses nombreuses responsabilités.

Je remercie les membres de mon jury Stéphane Robin, Annabelle Dejardin, Francis Galibert et Fabien Nogue d'avoir accepté d'évaluer ce travail. Je tiens à remercier également les membres de mon comité de pilotage de thèse, Mathieu Rolland, Christophe Hitte, Johann Joets et Mauro Petrillo pour leurs conseils et les échanges que nous avons pu avoir. Je remercie particulièrement Mauro de m'avoir accepté pendant quatre mois au sein de son équipe au JRC en Italie et pour sa collaboration au sein de ce projet.

Je remercie chaleureusement toute l'équipe de l'unité de Génétique Virale et Biosécurité (GVB). Yannick Blanchard pour son accueil au sein de son unité et les échanges constructifs que nous avons pu avoir. Pierrick et Edouard, les as de la bioinformatique : toujours prêts à m'aider pour déboguer mes scripts en Snakemake ou en Python, un grand merci à vous deux ! Je remercie aussi Helène, Daniel, Véronique, Aurélie, Renée, Françoise, Aurélie, Cécilia, Lionel, Nolwenn et Maud pour votre gentillesse et tous les bons moments passés en salle de pause.

Merci à tous les doctorants, stagiaires et jeunes techniciens de l'ANSES que j'ai pu croiser lors de notre jambonnage hebdomadaire. Un merci particulier à Julie Eclercy, Clément Droillard et Laurent Souci avec qui j'ai débuté cette thèse. Vous avez été un grand soutien au cours de ces années passées à l'ANSES. J'ai apprécié échanger et partager avec vous, les obstacles et les victoires que j'ai traversé tout au long de cette thèse.

Merci à mes amis, fléchois, Gphy et ceux croisés lors des congrès JOBIM et

séminaires EIR-A pour ces bons moments passés ensemble. Des remerciements un peu plus particuliers à Aurélie Leobardy, une amitié marquante qui a débuté au détour d'une paillasse et d'un problème d'électricité lors de mon premier jour à l'ANSES.

Enfin, merci à ma famille, mes parents qui m'ont toujours soutenu dans mes choix et tout au long de mes études.

Table des figures

Liste des tableaux

Table des matières

Introduction

Les organismes génétiquement modifiés (OGM) sont un sujet de société inépuisable et polémique. Les OGM sont par définition des organismes dont le matériel génétique a été modifié d'une manière qui n'est pas naturelle. La sélection d'organismes parmi un grand nombre de variants n'est pas considérée comme un processus de génération d'OGM, même si le résultat pourrait présenter de très fortes similitudes avec une insertion non naturelle. Un OGM est constitué de trois parties principales : le génome hôte, une ou plusieurs séquences insérées, nommées inserts, et les séquences de jonction issues des plasmides et autres éléments utilisés pour la transgénèse (incorporation d'un ou plusieurs gènes dans le génome d'un organisme vivant) (Figure 1). Ces séquences de jonctions ont la particularité de présenter une partie de séquence appartenant au génome hôte et une partie de séquence appartenant à l'insert. L'insert est constitué dans la majorité des cas d'au moins une séquence codante (CDS) d'un gène, pour exprimer une spécificité non présente dans le génome hôte. L'insert représente une infime partie du génome OGM.

FIGURE 1 – Construction d'un OGM.

De nombreuses sociétés en biotechnologie ont développé des semences OGM dans le but de rendre les cultures résistantes à certaines maladies ou d'accélérer leur croissance pour augmenter leur rendement. La lignée de maïs OGM MON810 créée et vendue par la société américaine Monsanto donne des semences OGM qui contiennent un gène nommé *Cry1Ab* appartenant à la bactérie *Bacillus thuringensis*. Ce gène permet au maïs de produire une protéine insecticide lui conférant une résistance à certains insectes ravageurs comme la Noctuelle (*Sesamia nonagrioides*) et la Pyrale du maïs (*Ostrinia nubilalis*). De nombreuses études ont été menées sur cette lignée de maïs OGM, la teneur de la protéine codée par le gène inséré *Cry3Bb1* dans le maïs MON81017 au cours de sa vie a ainsi pu être déterminée [1] mais aussi les effets potentiels de ces cultures dans les sols exposés sur les organismes non ciblés tels que le nématode *Caenorhabditis elegans* [2, 3].

Plus récemment, en 2016, l'entreprise AquaBounty Technologies a obtenu une autorisation de commercialisation au Canada pour son saumon atlantique géné-

tiquement modifié, devenant ainsi le premier poisson OGM commercialisé à des fins alimentaires. D'après l'entreprise, ce saumon intègre deux gènes provenant d'un autre poisson, le saumon Chinook lui conférant une capacité de croissance deux fois plus importante qu'un saumon sauvage. En Amérique, la FDA (Food and Drug Administration) a qualifié le saumon AquAdvantage comme aussi sûr à manger que tout autre saumon de l'Atlantique non génétiquement modifié et aussi nutritif [4].

Alors que les États-Unis autorisent la culture des OGM, les politiques d'autorisation et de restriction des OGM diffèrent en fonction des pays bien que de nombreuses études aient été réalisées sur le sujet pour déterminer les avantages et les inconvénients à leurs cultures. En Europe, leur utilisation est encadrée et contrôlée. Les OGM sont ainsi autorisés à l'importation en Europe. La majorité des OGM importés est utilisée pour l'alimentation animale. En France, leur culture est totalement interdite. De nouvelles méthodes d'édition du génome, utilisant les mécanismes de réparation de l'ADN tel que la technique CRISPR-CAS9 ont été créées. Pour ces méthodes, la modification génétique n'est pas directement introduite par une entité extérieure mais les mécanismes de réparation de l'ADN sont détournés pour induire des modifications dans le génome. Ces organismes sont considérés comme OGM d'après la législation Européenne. Il est possible d'inactiver un gène, d'introduire une mutation ciblée ou de corriger une mutation particulière dans un génome à l'aide de ces nouvelles techniques.

Des méthodes de détection sont facilement développées sur des OGM connus, lorsque la séquence d'insert est connue. Il s'agit souvent d'un pré-requis à leur autorisation, donc leur utilisation légale. Des méthodes de détection par biologie moléculaire sont donc répertoriées dans des bases de données spécifiques comme GMO-Methods développée par le Joint Research Centre (JRC) qui est le laboratoire de référence Européen sur les OGM. Des méthodes bioinformatiques fortement basées sur des alignements directs ont été développées pour détecter des OGM partiellement connus. Une partie de leur séquence insérée étant connue, il est possible de détecter les séquences de jonctions de l'insert dans le génome OGM [5]. Cependant, un OGM correspondant à l'intégration d'ADN exogène inconnu, donc non répertoriée dans les bases de données, n'aura pas de méthode de détection rapide dédiée. Aucune indication sur la localisation ou la nature du gène inséré à l'origine de l'OGM n'est précisée. Les méthodes de biologie moléculaire sont donc impuissantes face à des OGM présentant des séquences non définies. Actuellement, aucune méthode ou outil n'est disponible pour identifier de tels OGM. Comment détecter et identifier de tels OGM ?

Cette thèse aborde cette problématique en proposant une solution bioinformatique disponible pour détecter des bactéries génétiquement modifiées à partir de données de séquençage haut débit. Le séquençage permet de lire l'information génomique, l'ADN ou l'ARN, contenue dans les cellules d'un échantillon biologique. Avec le développement de nouvelles technologies de séquençage comme Illumina, des génomes entiers peuvent être séquencés. Les données fournies par séquençage haut débit regroupent des millions, voire des milliards de courtes séquences nucléotidiques. A partir de ces courtes séquences, un génome peut être assemblé et annoté. L'annotation du génome fournit des CDS sur lesquels la méthode de

détection des OGM inconnus va s'appuyer. Ainsi, cette nouvelle méthode de détection repose sur la caractérisation de chacun des CDS en recherchant des suites de nucléotides, nommées motifs exceptionnels, propres à chaque espèce dans le but de caractériser le génome hôte et distinguer un potentiel OGM. Les CDS présentant des différences d'utilisation de ces motifs dans leur séquence nucléotidique seront identifiés comme séquences insérées, non apparentées à cette espèce. Des méthodes statistiques utilisant des motifs existent pour déchiffrer des fonctions biologiques [6], détecter un transfert horizontal de gènes [7] ou identifier des séquences d'origine virale [8]. Cependant, aucune application utilisant ces types de propriétés n'a été utilisée jusqu'à présent dans le cadre de la détection des OGM. Cette méthode de détection des OGM inconnus propose une réponse et permet de combler un manque. De plus, notre outil exploite la différence d'utilisation des nucléotides entre deux séquences, une propriété biologique présente chez les génomes possédant une modification quelconque incluant les OGM. Toutes les séquences nucléotidiques géniques présentant des motifs s'éloignant de ceux du génome hôte peuvent donc être détectées. Les gènes tronqués, fusionnés ou fortement mutés peuvent donc être identifiés.

Pour distinguer les CDS OGM des CDS sauvages, nous avons choisi de développer un modèle de prédiction. Ce modèle fait appel à l'apprentissage automatique ou machine learning. Ce choix a été motivé par les récentes avancées en terme d'algorithme et de méthodes d'apprentissage automatique dont les résultats modélisent bien la réalité. De nombreux domaines d'application utilisent des méthodes de machine learning, comme par exemple la génétique et la génomique [9] pour améliorer l'évaluation des risques microbiens et ainsi prédire le risque de maladie à l'aide de données de séquençage de nouvelle génération [10]. Le machine learning peut aussi être utilisé pour la prédiction de gènes synthétiques ou naturels [11]. L'utilisation de modèles de prédiction est aussi présent dans d'autres domaines d'application comme la finance, pour l'arbitrage financier sur un vaste marché [12] ou encore la musique avec la création d'un algorithme de recherche audio inclus dans Shazam [13]. Cependant, aucune application n'a, à ce jour, été utilisée dans le contexte de la détection d'OGM.

Dans un premier temps, la méthode proposée dans le cadre de cette thèse va nettoyer et trier les données de séquençage haut débit du génome bactérien potentiellement génétiquement modifié (GM). Les données de séquençage vont être nettoyées puis assemblées et annotées pour obtenir les CDS présents dans le génome. L'assemblage permet de reconstituer la séquence génomique qui a permis de produire les données de séquençage. L'annotation a pour vocation de traduire la séquence du génome en CDS, seules les parties qui seront traduites en protéines dans le génome sont conservées. Ces différentes étapes sont très fréquentes en analyse bioinformatique de données de séquençage. L'objectif est de trier les CDS du génome analysé obtenus après l'annotation selon deux catégories : les CDS apparentés au génome hôte et les CDS potentiellement inserts OGM. Plusieurs alignements, utilisant notamment le génome de référence de l'espèce à analyser, seront utilisés pour effectuer ce tri.

Dans un second temps, pour apprendre à différencier les CDS présents dans notre échantillon, le modèle de prédiction a besoin de jeux de données connus ap-

partenant à ces deux catégories : OGM et non OGM. Dans ce but, la littérature a été compilée pour constituer un jeu de données de CDS OGM. Les CDS non OGM utilisés comme données connues sont les CDS sauvages apparentés au génome de l'échantillon analysé. En effet, l'utilisation de ces CDS sauvages permet de bien caractériser ses propriétés géniques, contrairement aux CDS d'un génome proche comme le génome de référence de l'espèce associée qui ne parviennent pas à les approcher. La qualité des données présentes dans ce second jeu est primordiale pour ne pas induire des erreurs dans les résultats. En effet, si un CDS n'est pas correctement catégorisé, le modèle de prédiction ne pourra pas correctement s'entraîner, ce qui engendrera des erreurs dans les prédictions finales.

Ce travail est à l'interface de plusieurs domaines scientifiques. La biologie est nécessaire pour comprendre et déterminer les propriétés génétiques qui diffèrent entre un génome génétiquement modifié et un génome hôte. Le traitement des données de séquençage haut débit et l'implémentation de la méthode font appel à l'informatique. Les statistiques de motifs et l'analyse combinatoire s'entrecroisent pour identifier et rechercher des motifs exceptionnels dans les séquences nucléotidiques. De plus, elles sont aussi utilisées pour quantifier les différences de motifs présents au sein des gènes d'un génome potentiellement génétiquement modifié. Enfin, l'apprentissage automatique ou machine learning établit les critères de cette différence et identifie les séquences insérées.

Ce manuscrit est construit en 8 chapitres permettant de décrire les différentes stratégies mises en place et testées pour obtenir un outil de détection fonctionnel à partir de données de séquençage haut débit d'un génome bactérien pur potentiellement génétiquement modifié. Le chapitre 1 décrit l'état de l'art et le contexte avec les différentes méthodes de création et de détection des OGM. Le chapitre 2 présente les supports mis en place nécessaires à la détection des OGM. Les chapitres 3 et 5 décrivent les méthodes permettant de caractériser et de comparer les CDS. Les différents jeux de données OGM et non OGM nécessaires à l'application de ces méthodes sont présentés dans le chapitre 4. Un modèle de prédiction des CDS potentiellement OGM est proposé dans le chapitre 6. Le chapitre 7 énonce les résultats obtenus avec notre outil sur plusieurs jeux de données de séquençage haut débit de génomes bactériens OGM ou sauvages. Enfin, le chapitre 8 tire les conclusions et perspectives de ce travail.

*
* *

Conformément à la politique de l'Anses, cette thèse a été menée sous assurance qualité. Un plan de management de la qualité a été écrit et appliqué en s'appuyant sur les normes ISO 10006 « Système de management de la qualité – Lignes directrices pour le management de la qualité dans les projets » pour la mise en œuvre (Annexe A). Les comptes rendus des deux audits qualité réalisés sont présents à la fin de l'Annexe A. Ce plan a permis de formaliser le processus d'organisation des travaux de thèse avec pour objectifs d'assurer la qualité du résultat final et d'obtenir la garantie par un tiers de la fiabilité et de l'intégrité des résultats de recherche.

*
* *

Au cours de cette thèse, un séjour de 4 mois au sein du Joint Research Centre (JRC), laboratoire de référence Européen pour les OGM à Ispra en Italie a été réalisé. Ce séjour s'inscrit dans le parcours de l'école de recherche internationale d'Agreenium (EIR-A) qui vise à améliorer l'employabilité des doctorants par une ouverture à l'international et à les sensibiliser aux grands enjeux de société.

*
* *

Un glossaire regroupant les définitions du manuscrit est proposé à la page 215. Lors de leur première apparition les mots présents dans le glossaire sont suivis par un astérisque *.

Chapitre 1

Contexte de la détection des OGM

Sommaire

Ce chapitre présente le contexte de travail concernant la détection des OGM. Il définit les OGM, au cœur de ce sujet, détaille leur fabrication ainsi que la réglementation. Plusieurs méthodes de détection des OGM, lorsque la séquence insérée est connue ou partiellement connue, sont présentées.

1.1 Définition

Depuis 2001, la Commission Européenne définit un organisme génétiquement modifié comme un être vivant dont le matériel génétique a été modifié d'une manière qui n'est pas naturelle, donc en dehors des processus d'accouplement et/ou de recombinaison naturelle [14]. Cette modification est réalisée artificiellement par la main de l'homme pour atteindre un but défini. Le génome hôte correspond au génome sauvage, original, auquel une modification génétique est apportée pour construire un OGM. En 1972, Jackson et ses collaborateurs [15] ont développé un virus simien 40 (SV40), où un segment de l'acide désoxyribonucléique (ADN) contenant des gènes de phage lambda et l'opéron galactose de *Escherichia coli* ont été insérés : le premier OGM est ainsi créé. De nombreuses bactéries sont par la suite modifiées génétiquement, comme dans le cas de la bactérie *E. coli* dans laquelle un gène humain a été inséré dans le but de produire l'insuline humaine en 1979 [16].

1.2 Fabrication

Pour créer un OGM, différentes techniques de transfert de gènes ont été mises au point. Deux types de transferts peuvent être distingués : les transferts directs et indirects. Les transferts directs utilisent des moyens physiques ou chimiques pour forcer la pénétration de l'ADN dans les tissus ou cellules végétales tandis que les transferts indirects sont réalisés via un organisme transporteur comme avec la bactérie *A. tumefaciens* ou l'utilisation de plasmides lors d'une conjugaison bactérienne. En se focalisant sur les OGM de plantes, trois procédés différents permettant de créer un OGM vont être présentés : le transfert de gènes via la bactérie *A. tumefaciens*, la biolistique et l'édition de génome.

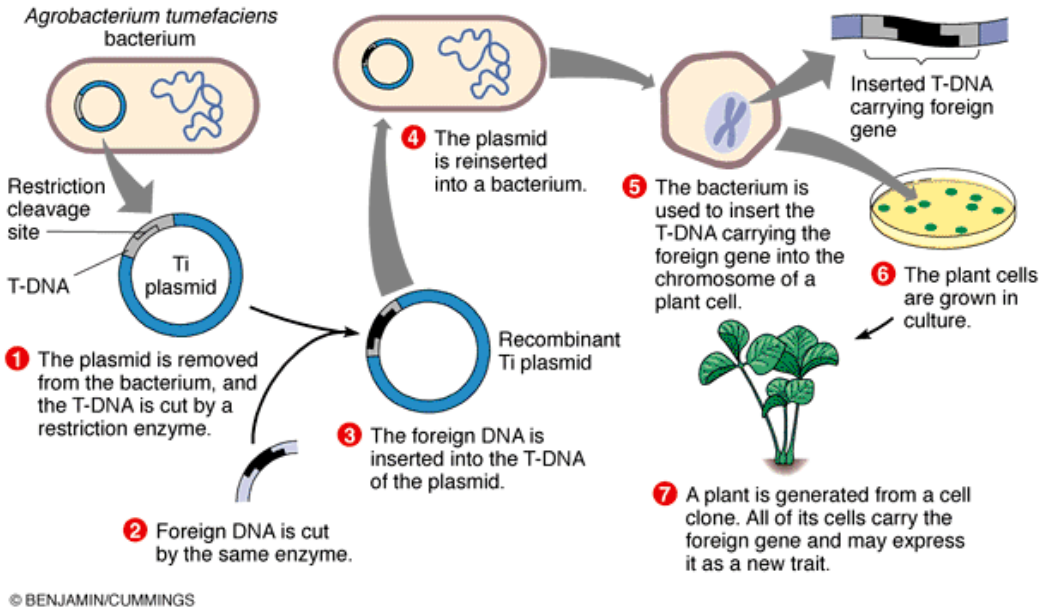


FIGURE 2 – Étapes de la fabrication d'un OGM de plante en utilisant la bactérie *A. tumefaciens* [17].

Grâce à la découverte des propriétés de bactéries telluriques phytopathogènes comme *Agrobacterium tumefaciens*, une méthode de transfert de gènes a été développée. La bactérie *A. tumefaciens* est capable de transférer une partie de son ADN dans le génome d'un végétal, pouvant ainsi servir de transporteur pour un gène d'intérêt [18, 19, 20]. Cette technique a permis de créer le premier OGM de plante en 1983, un tabac transgénique. Ce tabac peut se régénérer en plant sain après l'introduction de souches bactériennes produisant des tumeurs atténuées de la galle du collet [21]. La cellule de la bactérie *A. tumefaciens* est composée d'un chromosome et d'un plasmide inducteur de tumeur, le plasmide Ti. Cette méthode consiste à isoler ce plasmide Ti de la cellule de *A. tumefaciens*. Ensuite, la partie qui peut être transférée, nommée ADN-T, est coupée à l'aide d'une enzyme de restriction. Cette même enzyme de restriction est utilisée pour couper l'ADN étranger présentant le gène d'intérêt. Cet ADN étranger est ensuite inséré dans l'ADN-T de la bactérie *A. tumefaciens*. Le plasmide Ti modifié est replacé dans

la cellule de la bactérie *A. tumefaciens*. La bactérie transportant le gène d'intérêt insère l'ADN-T modifié dans le génome de la cellule végétale. Enfin, cette dernière est cultivée dans le but de donner naissance à de nouvelles plantes possédant l'ADN étranger (Figure 2). La transformation par *A. tumefaciens* a aussi été réalisée sur des plantes cultivées comme le maïs [22]. Récemment, une quatrième version d'un super *Agrobacterium* a été développée dans le but de palier à des taux de transferts de gène et de transformation stable nettement inférieurs chez certaines plantes [23].

Une autre technique permettant de créer un OGM se nomme la biolistique. Elle fait partie des méthodes de transfert direct. Cette technique consiste à propulser des projectiles microscopiques (billes d'or ou de tungstène) enrobés d'ADN à l'aide d'un canon à particules sur des cellules végétales. L'appareil est constitué d'une chambre reliée à une sortie pour créer un vide (Figure 3). Au sommet, un tube est temporairement isolé du reste de la chambre par un disque de rupture. Ce tube est utilisé pour contenir un gaz accélérateur. Un microsupport contenant des particules de tungstène recouvertes d'ADN, nommé microprojectiles, est placé à proximité du disque de rupture. Ensuite, un écran d'arrêt est placé entre les cellules cibles et le microsupport. Lorsque l'appareil fonctionne, de l'hélium gazeux est introduit dans le tube à grande vitesse. La pression à l'intérieur du tube étant importante, le disque de rupture se rompt propulsant ainsi les microprojectiles recouverts d'ADN contenus sur le microsupport grâce aux ondes de chocs de l'hélium. Les microprojectiles passent à travers l'écran d'arrêt pour délivrer l'ADN dans les cellules cibles. Les cellules transformées sont régénérées sur un milieu nutritif.

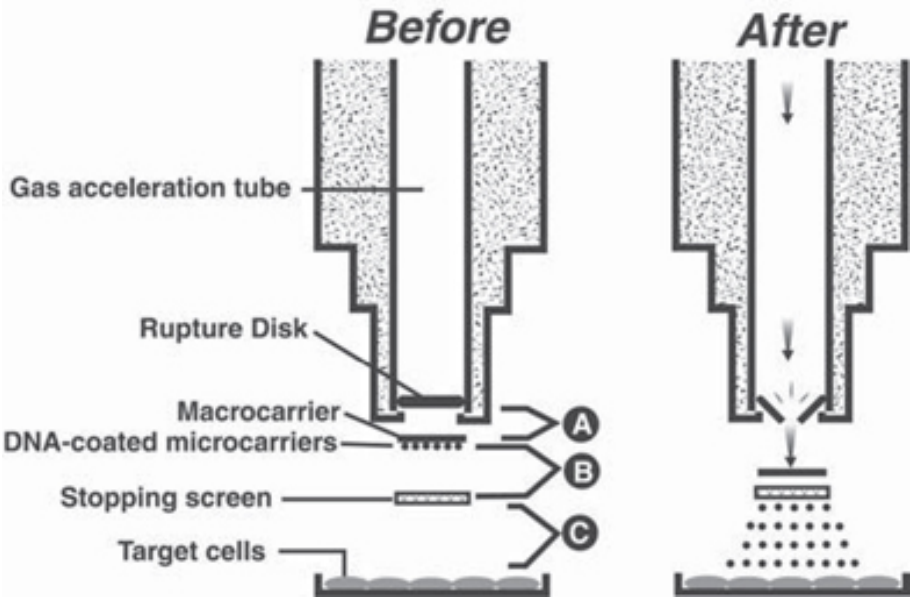


FIGURE 3 – Représentation de la création d'un OGM à l'aide de la méthode biolistique [24].

Cette technique est particulièrement adaptée aux plantes comme le riz, le blé ou le maïs qui se régénèrent difficilement et ne réagissent pas suffisamment au

transfert de gènes par *A. tumefaciens*. Cependant, elle entraîne la formation de chimères* et nécessite donc d'attendre la génération suivante pour sélectionner les descendants transformés. L'inconvénient de cette méthode, comparée à l'utilisation de *A. tumefaciens*, est qu'elle peut entraîner l'insertion de nombreuses copies du gène d'intérêt au niveau de plusieurs sites d'insertion dans le génome. Ces copies multiples peuvent inhiber l'expression du transgène dans la descendance [25].

De nouvelles techniques de sélection végétale sont récemment apparues sous l'appellation New Plant Breeding Techniques (NPBT). Parmi ces dernières, les techniques d'édition du génome peuvent produire différents types d'effets : l'inactivation, la modification ou l'insertion d'un gène. L'édition du génome permet de modifier de manière précise et ciblée la séquence génomique d'un organisme. Cette méthode est facile à mettre en place et applicable à tout type d'organisme : bactéries, plantes, animaux. Une ou plusieurs parties d'ADN d'un génome peuvent ainsi être insérées, modifiées ou supprimées. Pour cela, l'édition génomique fait appel à différents types de ciseaux moléculaires correspondant à des enzymes de restriction appelées nucléases : les méganucléases, les nucléases à doigt de zinc (ZFNs), les nucléases effectrices de type activateur de transcription (TALENs) et le système CRISPR-Cas9. Lorsque ces nucléases vont couper l'ADN, les cellules ciblées utilisent des mécanismes de réparation de l'ADN. Le principe de l'édition du génome est basé sur l'utilisation de ces mécanismes de réparation pour introduire des modifications dans le génome (Figure 4). Certaines modifications ne nécessitent pas l'insertion d'un gène exogène dans un autre organisme pour créer un OGM, notamment lorsqu'une méthodologie telle que CRISPR-Cas9 est utilisée [26].

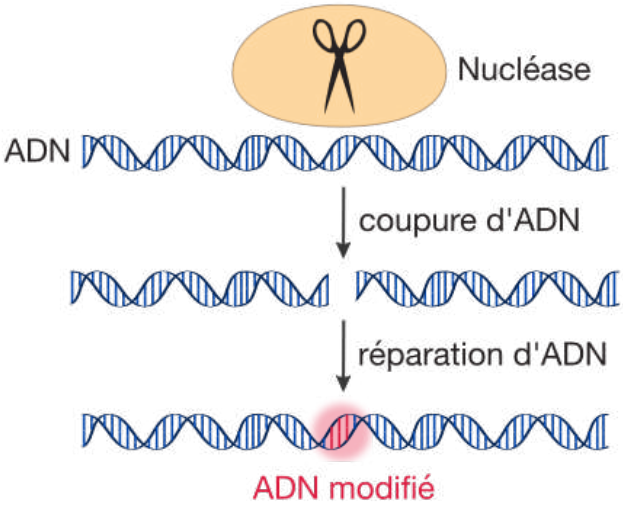


FIGURE 4 – Schéma présentant la technique d'édition du génome à l'aide d'une nucléase [27].

Récemment, Anzalone *et al.* [28] ont développé une technique d'édition du génome permettant de rechercher et de remplacer des variants génétiques sans

couper l'ADN double brin ou d'utiliser un ADN donneur. En effet, cette méthode utilise une endonucléase Cas9 catalytiquement altérée, fusionnée avec une transcriptase inverse modifiée et programmée avec un ARN guide pour écrire directement de nouvelles informations génétiques dans un site d'ADN spécifique.

1.3 Réglementation

La Communauté Européenne a adopté une politique très restrictive vis-à-vis de la diffusion et de l'utilisation des OGM. En effet, la législation Européenne exige que la présence d'OGM dans les produits alimentaires soit signalée lorsqu'ils représentent plus de 0,9% des ingrédients [29]. Ce règlement implique que tous les OGM ou produits dérivés d'OGM soient déclarés aux autorités Européennes. Les modifications de leurs séquences nucléotidiques doivent être enregistrées dans des bases de données appropriées et une méthode de détection basée sur des techniques moléculaires standards (par exemple, de type PCR) doit être fournie. Cependant, même si ces mesures existent, seuls les OGM connus sont aisément détectables. Un OGM inconnu est un organisme non documenté dont la séquence d'insert est inconnue.

Les OGM créés à l'aide de la méthode CRISPR-CAS9 par exemple, quand ils ne présentent pas d'insertion d'un gène exogène et donc ne peuvent pas être détectés, soulèvent la nécessité de la redéfinition d'un OGM comme cela est discuté par Eriksson *et al.* [30]. L'une des principales raisons de cette redéfinition est la détectabilité d'un OGM et donc l'applicabilité des lois concernant les conditions de mise sur le marché des OGM. Plusieurs projets sont présentés pour illustrer et appuyer cette demande de redéfinition, notamment le projet français GENIUS (Genome ENgineering Improvement for Useful plants of a Sustainable agriculture) mis en place par l'INRAE en collaboration avec plusieurs laboratoires privés et publics. Ce projet, d'une durée de 7 ans, a eu pour but d'utiliser l'édition génomique chez différentes espèces de végétaux en vue d'apporter des solutions à la résistance aux maladies, d'améliorer la tolérance à la salinité, de favoriser la précocité de la floraison, de modifier l'architecture de la plante ou encore sa reproduction, privilégiant des traits d'intérêt pour une agriculture plus durable [31].

1.4 Méthodes de détection

1.4.1 OGM connus

La séquence insérée dans le génome hôte représente une infime partie de la totalité du génome considéré. Pour détecter les OGM connus, plusieurs stratégies existent, classées selon deux catégories : les méthodes s'appuyant sur la détection de protéines et celles s'appuyant sur la détection d'ADN [32]. La méthode de détection des OGM la plus utilisée actuellement est la réaction en chaîne par polymérase en temps réel (qPCR). Elle appartient aux méthodes basées sur l'ADN qui sont généralement favorisées de part leur plus grande sensibilité et leur

quantification plus précise. Le principe de la méthode qPCR est de dupliquer un nombre important de séquences d'ADN connues (amplicons) à partir d'une faible quantité de matériel génétique et de mesurer la quantité d'ADN initiale. Elle peut ainsi déterminer le nombre de copies de transgène dès lors que la séquence recherchée est connue [33]. Cependant, la méthode qPCR est généralement limitée à l'amplification d'un seul amplicon. La méthode qPCR multiplexe a donc été développée pour amplifier plusieurs amplicons en même temps. Un nouveau type de PCR nommée PCR digitale (dPCR) permet une quantification absolue. La méthode dPCR partitionne le mélange réactionnel en un ensemble de réactions PCR menées en parallèle. En détection d'OGM, elle est utilisée lorsque le nombre de copies dans l'échantillon est faible ou qu'un inhibiteur de PCR est présent [34].

Des méthodes multiplexes non basées sur la qPCR ont été développées pour détecter des OGM connus : le Luminex, les microarrays ou puces à ADN et l'électrophorèse sur gel capillaire PCR (CGE) (Figure 5). Pour garantir une sensibilité suffisante lors de l'utilisation de ces différentes technologies détaillées dans les paragraphes suivants, les échantillons d'OGM nécessitent d'être amplifiés par PCR.

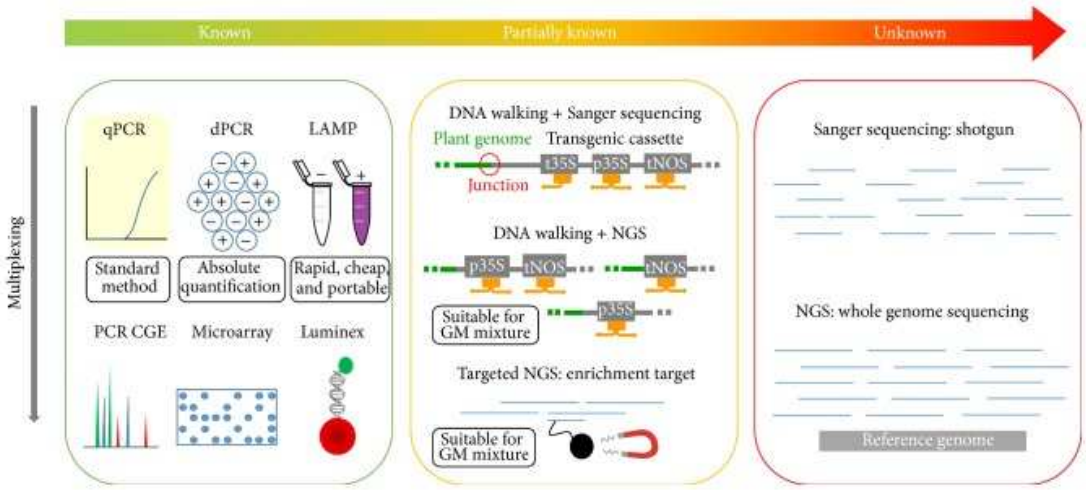


FIGURE 5 – Différentes méthodes de détection en fonction de l'information disponible sur les séquences d'inserts OGM [34].

Le Luminex [35] est un système multiplex basé sur le principe de la cytométrie de flux additionnée de deux lasers. Elle utilise des microsphères de polystyrène fluorescentes comme support de phase solide pour la détection de réactions biochimiques. L'avantage de ce système est la rapidité d'analyse des microsphères et l'utilisation de très faibles quantités de réactifs et d'échantillons. Dans le contexte de la détection OGM, cette technique permet de détecter jusqu'à 500 séquences cibles d'ADN différentes dans un échantillon. Pour cela, des ensembles de microsphères spectralement distincts sont couplés indépendamment à des sondes uniques d'acide nucléique [36].

Les puces à ADN ou microarrays ont été développées pour visualiser et analyser les niveaux d'expression de gènes dans des cellules ou organismes par rapport à un échantillon de référence. Une puce à ADN est constituée d'un support phy-

sique subdivisé en milliers de cases dans chacune desquelles une séquence d'ADN contenant un gène particulier est déposée. Pour détecter un OGM, cette technique est couplée à une PCR qui amplifie l'ADN cible grâce à des amorces spécifiques et/ou universelles qui sont ensuite hybridées sur la puce. De cette façon, plus de 250 000 cibles peuvent être détectées simultanément. Cette technique possède un débit plus important que celui de la qPCR mais une sensibilité légèrement moindre [37, 38].

L'électrophorèse capillaire sur gel (CGE) est une technique moléculaire permettant de distinguer des molécules en fonction de leur taille en leur appliquant un champ électrique. Pour détecter des OGM, cette technique est associée à une PCR multiplexe, les amorces PCR sont marquées par fluorescence pour séparer des amplicons de même taille. La sensibilité de cet électrophorèse est plus faible que celle de la qPCR [39].

La technique d'amplification isotherme de l'ADN facilitée par boucle (LAMP - loop-mediated isothermal amplification) a été développée par Notomi *et al.* [40]. Elle permet d'amplifier une grande quantité d'ADN rapidement grâce à la formation d'une structure en boucle. Peu coûteuse, elle présente une haute sensibilité et spécificité [41]. Contrairement à la qPCR, l'amplification d'ADN avec la technique LAMP ne nécessite pas de changement de température [41]. Cette technique permet également de traiter plusieurs échantillons en même temps.

1.4.2 OGM partiellement connus

De nouvelles techniques de séquençage haut débit (HTS) sont apparues à partir de 2005. Ces méthodes ont initié une nouvelle génération de séquençage (NGS) qui permet de séquencer massivement, en parallèle, des fragments d'ADN ou d'ARN appelés reads. Les avantages de ces méthodes sont la rapidité et un faible coût comparé aux méthodes de première génération (par exemple, la méthode Sanger). Ces nouvelles techniques de séquençage sont utilisées pour détecter et caractériser des OGM. Elles sont couplées à d'autres techniques (comme la qPCR) pour identifier et quantifier l'OGM [14, 42, 43]. Cependant, leur utilisation nécessite une connaissance complète ou au moins partielle de la séquence insérée et du génome hôte utilisé lors de l'alignement des reads sur une référence. En effet, des analyses bioinformatiques des ensembles de données de séquençage basées sur les similarités d'alignement ont été développées pour caractériser des éléments OGM référencés [14, 34, 44, 45].

Le DNA walking est aussi une technique qui nécessite une connaissance partielle de la séquence insérée pour caractériser et détecter des OGM dans les matrices alimentaires [47, 48, 49]. De nombreuses méthodes de DNA walking existent. L'objectif est toujours d'amplifier et de séquencer de proche en proche à partir d'une zone connue. La difficulté est de n'amplifier que les fragments d'intérêt. Le DNA walking permet donc de déterminer la séquence d'ADN de régions génomiques inconnues adjacentes à une région de séquence d'ADN connue (Figure 6). Cette technique est notamment utilisée pour identifier des régions promotrices de gènes où seule la région codante est présente. Le DNA walking couplé au NGS est utilisé pour la mise au point d'une approche d'identification des OGM non

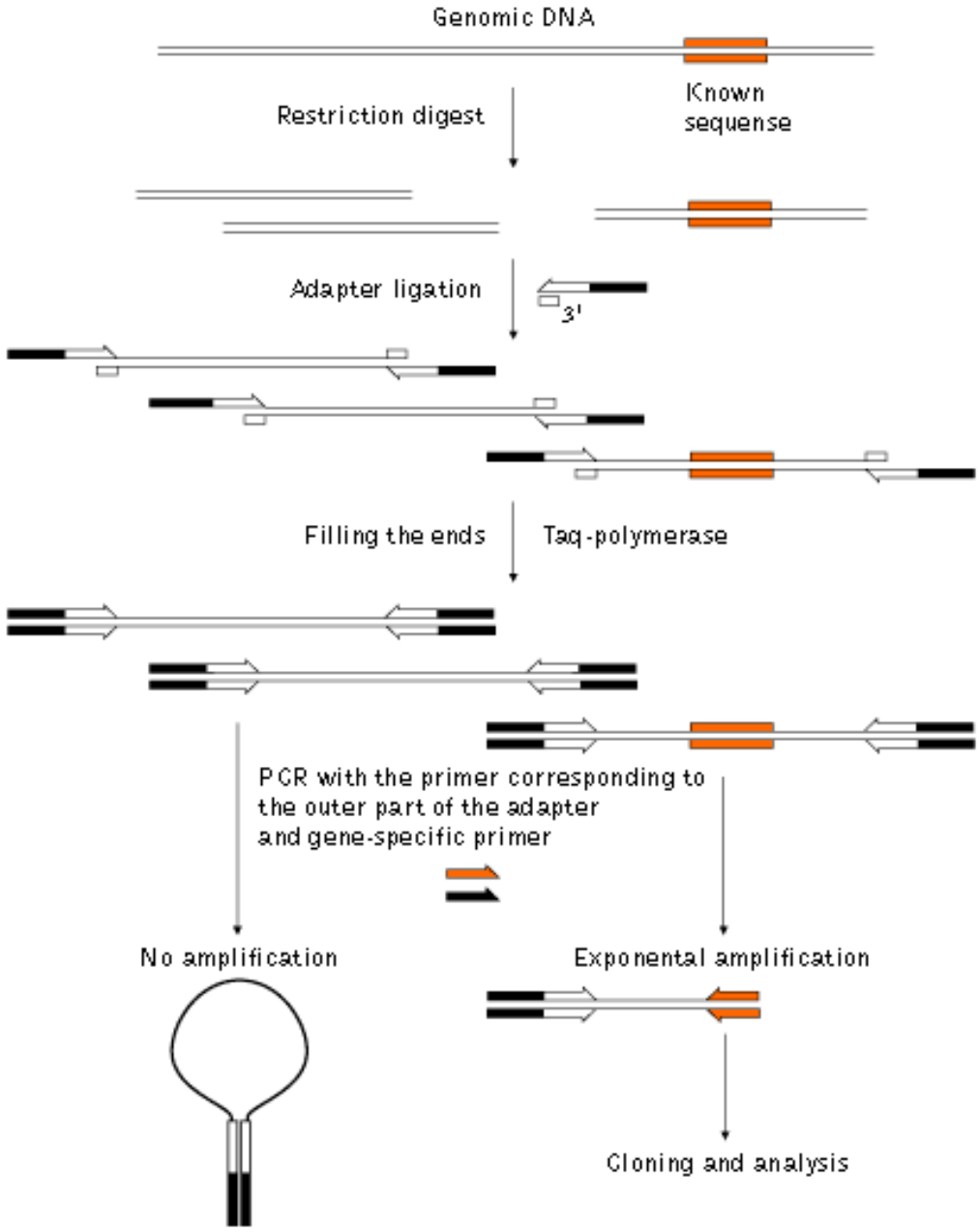


FIGURE 6 – Description de la technique de DNA walking utilisant une méthode de sélection via la formation de boucle pour éviter l'amplification à partir des extrémités aspécifiques. [46].

autorisés [50].

L'utilisation des techniques de séquençage permet de standardiser les approches de détection pour tous les types d'OGM contrairement au développement de méthodes spécifiques pour chaque OGM nécessitant l'utilisation de la PCR en temps réel. Différents outils ont donc été développés pour faciliter le travail de détection dans le cadre de routines de laboratoire pour des OGM dont la séquence est connue. Un framework* statistique permet de prédire le nombre

de reads nécessaires à la détection des séquences d'inserts à partir d'un séquençage MinION pour prouver leur intégration dans le génome hôte [43]. Avec son interface utilisateur intuitive, l'outil GMOseek aide l'utilisateur à sélectionner les analyses appropriées pour déterminer la présence d'événements de modification génétique individuels dans l'échantillon analysé. Il permet ainsi d'aider à la prise de décision et à l'interprétation des résultats dans toutes les phases d'analyses de la détection d'un OGM [51].

D'autres outils se basent sur l'utilisation de k-mers, par exemple Simka, Kraken2 ou Tallymer. Les k-mers représentent les sous-séquences existantes ayant une longueur k dans notre génome. Simka est un outil de métagénomique comparative basé sur l'analyse des k-mers [52]. Il représente chaque ensemble de données comme un spectre de k-mer et calcule plusieurs distances classiquement utilisées en écologie entre eux. Il est très rapide et performant. Cependant, Simka vise à comparer deux ensembles de séquences à partir de leur contenu en k-mers. Sa problématique n'est pas, comme dans le cas de la détection d'un insert d'OGM, de distinguer une ou quelques petites séquences dont la composition en k-mers s'éloigne de celle observée dans un génome sauvage. Kraken2 est un outil permettant de classer des séquences métagénomiques en utilisant des alignements exacts de k-mers [53]. Pour cela, il attribue une étiquette taxonomique à de courtes séquences d'ADN. Sensible et rapide, Kraken2 utilise un nouvel algorithme de classification pour supplanter les outils existants. Enfin, le comptage de k-mers permet aussi d'indexer et d'annoter de grands génomes végétaux répétitifs. Le logiciel Tallymer a ainsi permis d'indexer l'ensemble des séquences du génome du maïs [54].

1.4.3 Base de données OGM connus

Actuellement, différentes informations sont disponibles sous forme de bases de données pour l'aide à la détection d'OGM connus. Pour chaque OGM, de nombreuses spécificités sont indiquées, comme leur autorisation de mise sur le marché ou leur(s) méthode(s) de détection.

La base de données European GMO Initiative for a Unified Database System (Eugenius), créée par le Federal Office of Consumer Protection and Food Safty (BVL) en Allemagne et le Wageningen Food Safety Research (WFSR) aux Pays-Bas, a pour but de soutenir les autorités compétentes et les utilisateurs privés concernant la recherche d'informations précises sur les OGM [55]. Cette base de données indique des informations concernant la présence, la détection et l'identification des OGM dans l'Union Européenne et le monde.

Le Joint Research Centre (JRC) a aussi développé ses propres bases de données GMO-Amplicons [56, 57] et GMOMETHODS [58]. GMO-Amplicons regroupe plus de 240 000 séquences qui ont été obtenues par criblage PCR des banques de données publiques de séquences de nucléotides, incluant les brevets et les génomes entiers de plantes disponibles. Accessible via une interface web, cet outil aide les laboratoires dans la détection et l'identification d'OGM non autorisés mais aussi à retrouver les séquences liées aux OGM figurant dans la documentation relative aux brevets. GMOMETHODS est centré sur les méthodes de référence utilisées

pour l'analyse d'OGM. L'objectif est de fournir des informations fiables et harmonisées sur les méthodes d'analyses des OGM. En effet, le JRC étant laboratoire de référence de l'Union Européenne concernant les denrées alimentaires et les aliments pour animaux, ces méthodes ont été vérifiées dans le respect de la législation Européenne.

1.5 Conclusion

De nombreuses méthodes de détection des OGM connus sont disponibles. Cependant, l'utilisation des nouvelles techniques de séquençage haut débit est nécéssaire dans le cadre de la détection des OGM inconnus. Plus précisément, la méthode de séquençage Illumina est privilégiée dans ce travail. En effet, le séquençage Illumina présente une fréquence d'erreurs de 10^{-3} par rapport au séquençage IonTorrent PGM ou 454 GS Junior pour lesquels la fréquence d'erreurs est de 10^{-2} [59]. De plus, le pipeline de nettoyage (décrit dans le Chapitre 4) de la méthode mise au point utilise un assembleur qui donne d'excellents résultats d'assemblage mais ne peut traiter que des données pairées, donc provenant d'un séquençage Illumina. Enfin, le séquençage Illumina présente un plus faible coût de séquençage pour un volume de données important par rapport aux autres méthodes. On peut supposer que de nombreux génomes de plantes seront séquencés dans les prochaines années et compte tenu de la grande taille de leurs génomes en général, il est probable que le séquençage Illumina HiSeq sera privilégié.

Les technologies de séquençage Ion torrent et Minion (Oxford Nanopore) présentent un trop grand nombre d'erreurs systématiques d'homopolymères dans leur séquençage pour pouvoir être utilisées dans notre cas : cela peut entraîner des décalages de cadre de lecture par rapport au gène réel, donc une corruption des données statistiques des CDS. Si ces erreurs étaient totalement aléatoires, elles pourraient être corrigées lors de l'assemblage ou de l'alignement des reads. Plus les données de séquençage présenteront une qualité suffisante et plus les résultats seront probants et fiables en éliminant l'apparition de faux négatifs et de faux positifs. La qualité des données de séquençage est donc primordiale.

Les bases de données regroupant des informations sur les OGM connus sont une aide précieuse pour les analyses biologiques (type PCR) lorsque la séquence insérée est connue. Cependant, lorsque la séquence insérée dans l'OGM est inconnue, elles sont inutilisables. Néanmoins, dans le cadre de cette thèse, les bases de données répertoriant des séquences d'inserts OGM connus vont être utilisées en temps que référence pour réaliser une comparaison entre des séquences potentiellement inserts (provenant de l'OGM inconnu) et des séquences d'inserts OGM connus. Cette utilisation des bases de données d'inserts OGM connus pourra être réalisable uniquement si leur contenu est disponible publiquement. Dans la majeure partie d'entre elles, un moteur de recherche permet de retrouver une méthode de détection ou un amplicon en fonction de l'espèce ou de l'OGM précisé mais l'exportation de la totalité de la base de données n'est pas possible.

Chapitre 2

Mise en place de supports de détection d'OGM inconnus

Sommaire

Des pangénomes correspondant aux différents génomes hôtes, une banque de données regroupant des séquences d'inserts d'OGM connus et des données de séquençage d'OGM synthétiques ont été construits dans le but de faciliter la détection d'OGM inconnus.

2.1 Objectifs

Suivant la disponibilité des données, une sélection d'espèces OGM représentatives et variées a été réalisée. Les pangénomes correspondant à ces différentes espèces ont été créés pour représenter les différences génétiques présentes au sein d'une même espèce et ainsi pouvoir écarter les gènes appartenant à l'espèce hôte de ceux pouvant être identifiés comme des séquences d'inserts OGM. Des séquences d'inserts OGM connus ont été récoltés.

Ensuite, une banque de données d'OGM connus a été construite pour fournir un maximum d'informations connues lors de l'étape de création d'un modèle de prédiction. En effet, les données d'apprentissage nécessaires à l'élaboration de modèles de prédiction doivent comporter des données connues appartenant aux deux catégories d'échantillons (positifs et négatifs) pour pouvoir déduire des règles permettant de distinguer ces échantillons. Dans notre cas, la distinction entre les CDS inserts OGM (échantillons positifs) et les CDS apparentés au génome hôte (échantillons négatifs) est recherchée.

Enfin, n'ayant que peu de données d'OGM à notre disposition, des données de séquençage OGM ont été générées artificiellement en vue de tester notre approche. En complément, des génomes sauvages ont été sélectionnées dans le but de conforter nos résultats en ne détectant aucune séquence insérée.

2.2 Méthodes

2.2.1 Création de pangénomes

En 2005, Tettelin *et al.* [60] définissent un pangénome comme le regroupement de l'intégralité de la variété des gènes présents dans une même espèce. Un pangénome est constitué du core génome et du génome accessoire (Figure 7). Le core génome regroupe tous les gènes communs à toutes les souches de l'espèce et le génome accessoire rassemble les gènes absents d'une ou plusieurs souches et les gènes uniques, propres à chaque souche [61].

Deux types généralisés de pangénomes existent, ils sont classés en fonction du nombre de nouveaux gènes ajoutés au pangénome par génome séquencé. Un pangénome ouvert signifie que le nombre de gènes présents dans le pangénome augmente fortement avec le nombre de souches supplémentaires séquencées. En effet, à l'aide d'une extrapolation mathématique des données, Tettelin *et al.* [60] ont observé que le pangénome de *Streptococcus agalactiae* était très grand et que des gènes uniques peuvent toujours être identifiés même après que des centaines de génomes aient été séquencés. Les pangénomes ouverts sont fréquemment présents chez des espèces comme *E. coli* qui vivent dans des environnements multiples exigeant des niveaux élevés d'adaptabilité ou favorisant beaucoup les échanges de matériel génétique [62]. A l'opposé, un pangénome fermé signifie qu'après avoir séquencé quelques souches, les souches supplémentaires n'apporteront pas de nouveaux gènes au pangénome. En effet, certaines espèces comme *Bacillus anthracis*, vivent dans une niche isolée et restreinte permettant d'obtenir peu de gènes étrangers.

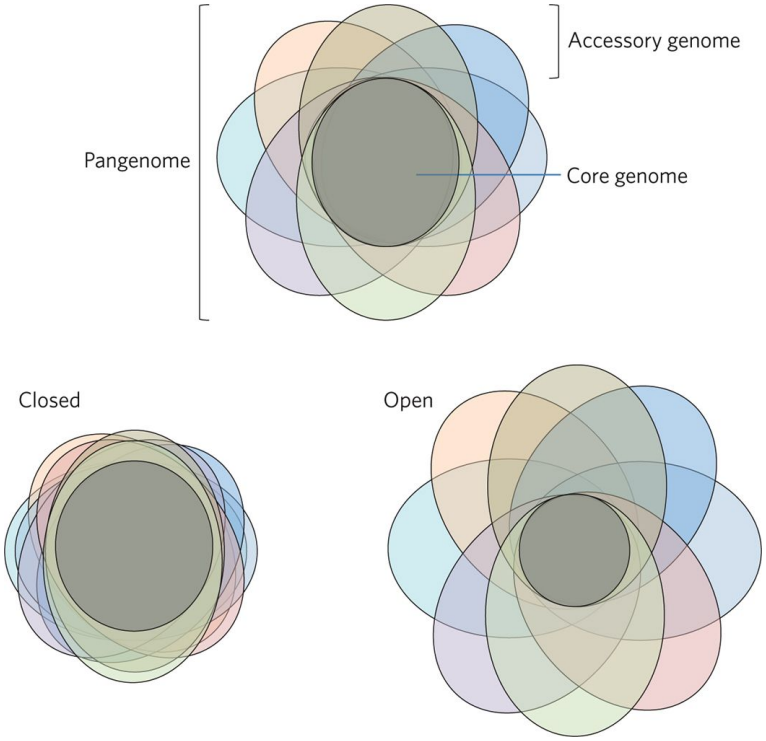


FIGURE 7 – Schéma représentant les différentes parties qui constituent un pangénome [63]. Chaque couleur de « pétale » correspondant à l'ensemble des CDS d'un génome distinct.

Pour déterminer si un pangénome est ouvert ou fermé, une approche de la loi de Heap a été suggérée par Tettelin *et al.* en 2008 [62]. Historiquement, la loi de Heap est une loi empirique qui décrit le nombre de mots distincts dans un document ou un ensemble de documents en fonction de la longueur du document. Elle est représentée par la formule suivante [64] :

$$n = K \times N^{-\alpha} \tag{2.1}$$

Dans le contexte génétique, n représente le nombre attendu de gènes pour un nombre donné de génomes, N est le nombre de génomes, et K et α sont des paramètres libres définis empiriquement [62]. Selon la loi de Heap, le pangénome est considéré ouvert lorsque $\alpha < 1$, donc le nombre de gènes augmentera pour chaque nouveau génome séquencé. A l'inverse, un pangénome est fermé lorsque $\alpha > 1$, donc le nombre de gènes n'augmentera pas après l'ajout de nouveaux génomes [62].

De nombreux outils ont été développés pour analyser et visualiser les pangénomes [65, 66, 67, 68]. Le pangénome d'une espèce permet notamment de caractériser ses différentes souches à l'aide de leurs ensembles de gènes uniques [69] mais aussi d'étudier l'impact du transfert de gènes horizontaux sur l'espèce [70]. De plus, il est utilisé pour la détection, l'identification et le suivi de nouvelles souches dans les échantillons de métagénomique à partir des sous-ensembles de gènes individuels [71]. Cependant, les pangénomes sont approximatifs. Pour le

moment, il est impossible de déterminer de manière exhaustive un pangénome complet. Aucun outil n'existe pour en garantir l'exhaustivité.

Pour détecter les séquences d'inserts présentes dans un génome d'OGM inconnu, la majorité des CDS apparentés au génome hôte seront écartés afin de réduire le nombre de CDS potentiellement inserts et ainsi faciliter la détection de l'insert réel. Dans ce but, un pangénome est donc utilisé. La construction de ce pangénome nécessite de regrouper des génomes complets appartenant à la même espèce que l'OGM inconnu. Cependant, dans notre cas, ce pangénome est approximatif, il est très difficile de regrouper l'intégralité des génomes complets pour une souche donnée et ainsi garantir que l'intégralité des CDS apparentés au génome hôte est écartée. L'important dans le cadre de notre méthode de détection est la représentativité du vocabulaire (utilisation préférentielle des codons) des CDS de l'espèce dans notre pangénome. La qualité des données de séquençage inclus dans le pangénome est déterminante pour éviter les erreurs de caractérisation des CDS du génome potentiellement OGM.

Nous nous sommes orientés vers la détermination des CDS sauvages de l'échantillon par recherche de similarités avec ceux du pangénome. Les données d'apprentissage pour l'étape de création d'un modèle de prédiction nécessitent deux ensembles : l'un pour les CDS sauvages, qui vient d'être abordé, l'autre pour les CDS OGM connus, qui est évoqué dans la partie suivante.

2.2.2 Création d'une banque de données d'inserts OGM connus

Le JRC a fourni 2641 séquences provenant de la base de données GMO-Amplicons (Chapitre 1 partie 1.4.3). Ces séquences ont ensuite été annotées avec le logiciel Prokka [72] pour déduire les CDS inserts OGM. 183 CDS inserts OGM provenant de bactéries issues de la littérature ont été ajoutés à ces séquences annotées par Prokka pour constituer la banque de données d'inserts OGM connus. Certaines des séquences d'inserts provenant de la littérature n'étant disponibles que sous forme de séquences protéiques, l'outil Backtranseq de la suite EMBOSS [73] a été utilisé pour les traduire en nucléotides à partir des usages des codons de *B. subtilis* et *B. subtilis* high disponibles dans la base de données de l'outil. Notre banque de données contient donc des inserts procaryotes et eucaryotes.

2.2.3 Construction d'OGM synthétique

Deux méthodes de construction d'OGM synthétiques ont été mises au point, la première à partir des données du maïs et la seconde à partir de données provenant de bactéries.

Une séquence d'insert provenant d'une bactérie GM a été insérée dans le génome de référence du maïs. Ensuite, les séquences consensus provenant de l'alignement entre les reads de maïs sauvage (ayant servi à créer l'OGM de maïs) et le génome de référence du maïs ont été ajoutées au génome de référence précédemment modifié pour s'assurer de garder des séquences contaminantes naturellement présentes dans des échantillons sauvages dans la création de notre génome OGM

synthétique. Enfin, ART version 2.5.8 [74], un logiciel permettant de générer des reads synthétiques de séquençage de nouvelle génération a été utilisé pour générer des données pairées provenant d'un séquençage Illumina MiSeq.

Pour les bactéries, une autre méthode de construction d'OGM synthétique a été mise au point. En premier lieu, un gène exogène est inséré entre deux CDS dans un génome de référence. Ensuite, le logiciel ART a été utilisé sur ce génome modifié en utilisant le programme art_illumina avec les paramètres : -ss HS25, -p, -l 150, -f 11, -m 200, -s 10 pour créer des données de séquençage Illumina synthétiques.

2.3 Application

Plusieurs collaborations ont été mises en place pour obtenir différents jeux de données de séquençage OGM, souvent non disponibles pour des raisons de confidentialité. Le génome complet d'un maïs (*Zea Mays*) OGM ainsi que sa souche sauvage ont été utilisés comme sources principales de données.

2.3.1 Cas du maïs

2.3.1.1 Présentation du maïs OGM

Nous avons choisi d'étudier un génome de maïs OGM pour mettre au point la méthode de détection des OGM inconnus. Plusieurs arguments ont motivé ce choix. La culture de maïs OGM est très importante, principalement aux États Unis, et il s'agit d'une des plantes les plus souvent modifiées génétiquement. Il s'agit d'un organisme modèle souvent étudié en recherche fondamentale. De plus, la grande taille de son génome en fait un modèle intéressant et complexe pour la détection d'OGM inconnus. En effet, si la détection est possible sur le maïs, alors il devrait être facile d'adapter la méthode à d'autres organismes moins complexes comme les bactéries ou les levures.

Le laboratoire de la santé des végétaux (LSV) de l'ANSES à Angers possède une unité dédiée aux OGM. Cette unité est Laboratoire National de Référence (LNR) concernant la détection et l'autorisation de mise sur le marché de plantes OGM. Le maïs OGM utilisé au cours de ce projet à été fourni par cette unité. Le séquençage pairé du maïs OGM a été réalisé en Illumina HiSeq. Les données de séquençage du maïs OGM, d'une souche sauvage de maïs ainsi que la séquence de référence de l'insert ont été mises à notre disposition.

Le fait de disposer à la fois des données de séquençage d'un OGM de maïs et de celles de son pendant sauvage, données fournies par le LSV d'Angers, permet d'appréhender les données auxquelles nous serions confrontés en l'absence d'insert (témoin négatif). Au début du projet, seul ce jeu de données complet regroupant les données nécessaires à la création d'un OGM (génome sauvage ayant servi à créer l'OGM correspondant, génome OGM et la portion insérée) était accessible.

Le maïs OGM fourni par le LSV d'Angers possède un insert nommé T25 qui confère une résistance aux herbicides à base de glufosinate d'ammonium. Cet insert T25 est composé de deux gènes *pat* et *bla*. *pat* est un gène synthétique

dérivé de la souche Tu 494 de *Streptomyces viridochromogenes* qui provoque cette résistance aux herbicides. Le gène *bla* provient de la bactérie *Escherichia coli* et confère une résistance à un antibiotique, l'ampicilline (Tableau 1).

Gène introduit	Source du gène	Produit	Fonction
pat (syn)	Forme synthétique du gène pat dérivée de la souche Tu 494 de *Streptomyces viridochromogenes*	Enzyme de phosphinothricin N-acetyltransferase (PAT)	Élimine l'activité herbicide des herbicides à base de glufosinate (phosphinothricin) par acétylation
bla	*Escherichia coli*	Enzyme de beta lactamase	Détoxifie les antibiotiques beta lactame comme l'ampicilline

Tableau 1 – Description des gènes présents dans l'insert T25 du maïs OGM. Librement adapté de [75].

La construction génétique de l'insert T25 est décrite par le graphique suivant :

FIGURE 8 – Construction génétique de l'insert T25 du maïs OGM [76].

Le mais OGM T25 a été développé par Bayer CropScience et approuvé comme culture transgénique pour les denrées alimentaires et les aliments pour animaux dans la plupart des pays du monde. Cependant, sa culture est interdite dans tous les pays exceptés les États-Unis, le Japon, le Brésil, l'Argentine et le Canada. Le développement d'une méthode de PCR multiplexe pour l'identification et la quantification du maïs OGM T25 a été développée. Elle avait pour but de simplifier et de réduire le nombre de tests nécessaires pour caractériser un OGM [77].

Le génome de référence du maïs utilisé au cours des différentes analyses du projet provient de la base de données PlantGDB : il s'agit de la version 34. Ce maïs de référence est nommé B73 [78]. Il est composé de 40000 gènes et au moins le double de transcrits compte tenu des possibilités d'épissage alternatif (plusieurs fois le même gène avec des exons différents). Ce génome de référence a été récupéré le 14 Février 2017 mais n'est actuellement plus disponible sur la base de données de PlantGDB.

2.3.1.2 Résultats relatifs aux OGM synthétique

Les résultats obtenus sur le maïs avec le procédé décrit dans la partie 2.2.3 sont très facilement identifiables. Cette technique fournit un OGM synthétique trop éloigné de la réalité en termes de propriétés statistiques. En effet, la séquence insérée ne possède pas une utilisation préférentielle des triplets de nucléotides (codons) similaire à celle du génome de référence. De plus, l'environnement du génome hôte est difficilement reproductible. L'insert est dans ce cas trop facilement identifiable. Cependant, ce procédé n'a été testé que deux fois avec le

génome de référence du maïs. La création de données synthétiques à partir du génome du maïs n'est pas conservée pour tester la méthode de détection mais elle sera utilisée par la suite avec des génomes bactériens.

2.3.2 Cas de la bactérie *B. subtilis*

Après avoir réalisé des analyses sur le génome du maïs OGM, une seconde stratégie visant la détection de bactéries génétiquement modifiées (GM) à été envisagée. Nos analyses ont été focalisées sur deux jeux de données de séquençage de la bactérie *B. subtilis*. Le premier jeu correspond à la bactérie GM fourni par le JRC et le second à une bactérie sauvage (témoin négatif).

2.3.2.1 Présentation de *B. subtilis*

B. subtilis est une bactérie ubiquitaire gram positif* appartenant aux Firmicutes, couramment retrouvée dans le sol, les sources d'eau et en association avec les plantes [79]. La séquence complète de son génome a été publiée pour la première fois en 1997 par un consortium formé principalement de laboratoires européens et japonais [79]. Depuis, les techniques de séquençage ayant évolué, le génome de *B. subtilis* a entièrement été reséquencé en utilisant des méthodes NGS et aussi nouvellement annoté [80]. Enfin, en 2018, une annotation unique du génome de *B. subtilis*, traitée manuellement a été présentée par Borriss *et al.* [81]. Cette annotation est essentiellement basée sur des données expérimentales. Pour ce projet, le génome de référence de la souche 168 de *B. subtilis* a été utilisé, il est disponible sous le numéro d'accession NC_000964.3.

Le Bavarian Hearth and Food Safety Authority (LGL) et le Hessian State Laboratory (LHL) en Allemagne ainsi que le Joint Research Centre (JRC) situé à Ispra en Italie ont effectué trois séquençage haut débit d'une bactérie *B. subtilis* GM non autorisée [82]. Le LGL, via une plateforme de séquençage, a réalisé un séquençage de la bactérie en Illumina MiSeq avec des reads pairés*. Le LHL a utilisé la technique de séquençage Illumina HiSeq 1500 avec des reads pairés. Enfin, le JRC a effectué le séquençage en 454 GS Junior. Ces trois jeux de données de séquençage haut débit d'une bactérie GM ont été mis à notre disposition par le JRC et sont disponibles publiquement [83]. Ce génome de *B. subtilis* GM contient un gène *cat* conférant une résistance à un antibiotique et inséré sur le chromosome entre deux parties du gène *recA*. De plus, quatre plasmides (Figure 9) ont aussi été insérés pour augmenter la production de riboflavine, donc de vitamine B2. Certains de ces plasmides possèdent des gènes issus des bactéries *Staphylococcus aureus*, *Bacillus amyloliquefaciens* ou *B. subtilis*. Pour obtenir ces résultats, le JRC a réalisé une analyse bioinformatique utilisant comme génome de référence pour *B. subtilis* la souche AG1839 disponible sous le numéro d'accession CP008698.1. Récemment, une révision de ce génome avec un seul plasmide pGMrib et une insertion chromosomique a été décrite par Berbers *et al.* [84]. Cet insertion chromosomique de 53 kb est intégrée dans le gène *scpA*. Dans le cadre de cette intégration, plusieurs occurrences des opérons ribDEAHT, complets ou partiels, montrent une grande similitude avec l'opéron ribDEA de *Bacillus amyloliquefaciens*. De plus, de multiples copies des gènes de résistance

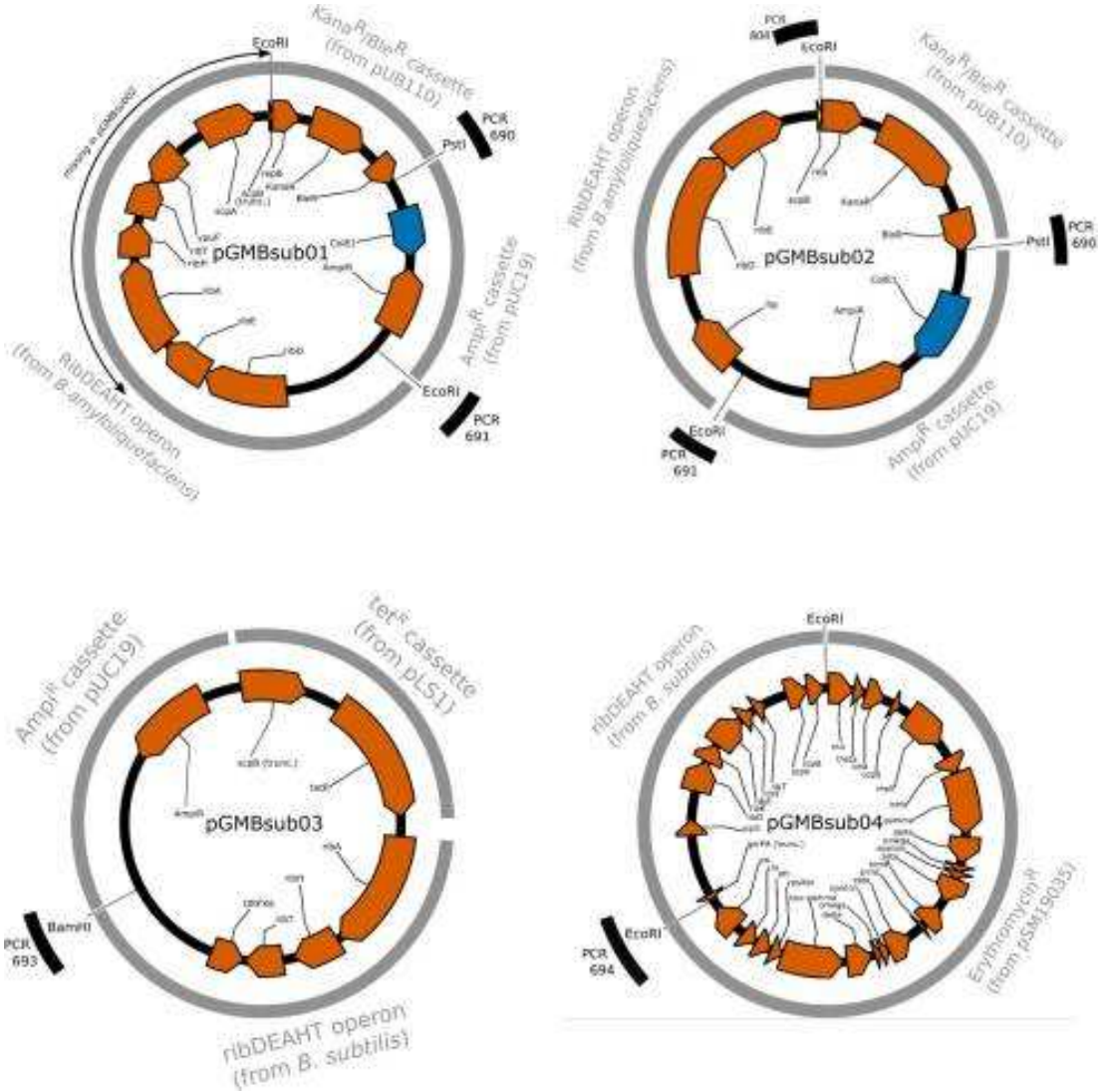


FIGURE 9 – Visualisation des 4 plasmides présents dans la bactérie *B. subtilis* GM [82].

à la bêta-lactamase et à la kanamycine sont trouvées dans l'insertion de 53 kb. Enfin, un gène de résistance à la bléomycine est présent.

Les données pairées de séquençage Illumina HiSeq 2500 d'une souche sauvage de *B. Subtilis* 9407 ont été utilisées. Ces données sont disponibles sur le Sequence Read Archive (SRA) du National Center for Biotechnology Information (NCBI) sous le numéro d'accession SRR8935610. Elles ont été soumises par l'université agricole de Chine dans le but d'en savoir plus sur ces souches en vue de la production d'antibiotiques. En effet, *B. subtilis*, régulièrement utilisée pour la production d'antibiotiques, est capable de produire plus de 24 antibiotiques ayant une importante variété de structure [85].

2.3.2.2 Pangénome de *B. subtilis*

Un pangénome de la bactérie *B. subtilis* a été créé pour répondre au besoin de notre objectif de recherche. Le pangénome de *B. subtilis* a été construit à partir de la publication de H. Brito *et al.* [86] où les références de 22 souches de *B. subtilis* correctement identifiées sont incluses. Les espèces limites de ce pangénome sont *Bacillus tequilensis* et *Bacillus vallismortis*. Notre pangénome regroupe 37 souches sauvages différentes provenant de données de séquençage haut débit de génomes complets disponibles sur le NCBI (Annexe B, Tableau 17). Les assemblages utilisés proviennent principalement de données de séquençage de type PacBio, Illumina MiSeq ou HiSeq pour éviter les erreurs de cadre de lecture liées aux autres types de séquençage.

2.3.3 Cas de la bactérie *E. coli*

Les données de séquençage de trois génomes de la bactérie *E. coli* dont le statut d'OGM est connu, ont été utilisées pour valider et ainsi tester la méthode de détection d'OGM. Le pangénome correspondant à cette bactérie a donc été créé.

2.3.3.1 Présentation de *E. coli*

E. coli est une bactérie intestinale gram négatif, dont certaines souches sont pathogènes chez l'homme ou chez l'animal. Ces souches pathogènes provoquent notamment des gastro-entérites, des diarrhées et des infections urinaires [87]. Les souches K-12 et B de la bactérie *E. coli* ont été utilisées de nombreuses fois lors d'expériences et ont permis de mieux comprendre la génétique moléculaire. Les différentes souches de quatre sous-groupes de *E. coli* ont été comparées en analysant leurs nombres de gènes dans le pangénome ou dans le core génome en fonction du nombre de génomes considérés. Ces quatre sous-groupes de *E. coli* sont six souches commensales, treize souches pathogènes intestinales, sept souches pathogènes extra-intestinales et six souches Shigella [88]. La séquence entière du génome de *E. coli* K-12 de la souche MG1655 a été présentée et annotée en 1997 [89]. Le génome de référence de la souche MG1655 de *E. coli* K-12 est disponible sous le numéro d'accession U00096.3.

Les données pairées de séquençage Illumina MiSeq provenant de trois souches GM de la bactérie *E. coli* sont disponibles sur la base de données public SRA (Sequence Read Archive) sous les numéros d'accession SRR9304542, SRR9304539, SRR9304540. Ces trois génomes sont issus de la même souche de *E. coli* mais possèdent chacun un plasmide contenant le gène *parE* provenant de *Agrobacterium tumefaciens*, *Mycobacterium tuberculosis* et *Streptococcus pyogenes* respectivement [90]. Le gène *parE* fait partie de la sous-famille des toxines-antitoxines et peut être présent sur les chromosomes et les plasmides. L'objectif de l'étude de Ames JR. *et al.* [90] était d'évaluer l'équivalence de l'activité du gène *parE* en fonction de son espèce bactérienne d'origine.

2.3.3.2 Pangénome de *E. coli*

Comme pour le pangénome de *B. subtilis*, le pangénome de la bactérie *E. coli* est constitué de génomes complets provenant de souches sauvages. Il rassemble 45 souches différentes (détaillées en Annexe B, Tableau 18) mentionnées en partie dans la publication de Touzain *et al.* [91].

2.3.4 Cas de six autres bactéries

Pour valider et tester la méthode de détection d'OGM, en complément des génomes de *E. coli*, les génomes complets de six bactéries sauvages supplémentaires ont ensuite été ajoutés pour éprouver la méthode. Leur pangénome associé a aussi été créé.

2.3.4.1 Présentation des bactéries utilisées

Les données pairées de séquençage Illumina MiSeq de six bactéries sauvages additionnelles ont été utilisées : *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Lactococcus lactis*, *Listeria monocytogenes*, *Salmonella Typhimurium* et *Staphylococcus aureus*. Ces données sont disponibles sur le SRA sous les numéros d'accession SRX6930849, ERX3965724, SRX4873416, SRX7828066, SRX7828483, SRX7817854 respectivement. Ces bactéries ont été choisies dans le but de sélectionner des espèces représentatives et variées. En effet, elles sont souvent étudiées dans la littérature et présentent des pourcentages en GC différents (*M. tuberculosis* possède un fort pourcentage en GC [92] alors que *C. jejuni* possède un pourcentage en GC plutôt faible [93]).

Les génomes de référence qui seront utilisés pour ces six bactéries, disponibles sur le site américain du NCBI, sont :

- le génome *C. jejuni* str. NCTC11168 (NC_002163.1) pour les analyses sur la bactérie *C. jejuni*,
- le génome *M. tuberculosis* str. H37Rv (NC_000962.3) pour les analyses sur la bactérie *M. tuberculosis*,
- le génome *L. lactis* str. Il1403 (AE005176.1) pour les analyses sur la bactérie *L. lactis*,
- le génome *L. monocytogenes* str. EGD-e (NC_003210.1) pour les anaylses sur la bactérie *L. monocytogenes*,
- le génome *Salmonella enterica* serovar *Typhimurium* str. LT2 (AE006468.2) pour les anlayses sur la bactérie *S. Typhimurium*,
- le génome *S. aureus* str. NCTC8325 (NC_007795.1) pour les analyses sur la bactérie *S. aureus*.

2.3.4.2 Pangénome de six bactéries additionnelles

Un pangénome pour chacune des six bactéries suivantes a été créé : *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Listeria monocytogenes*, *Lactococcus lactis* et *Salmonella Typhimurium*. Comme pour les précédents pangénomes, ils sont constitués uniquement de génomes complets provenant

de souches sauvages. Les pangénomes des bactéries *C. jejuni*, *L. lactis* et *S. Typhimurium* sont composés de 219, 107 et 226 génomes et plasmides respectivement provenant de la base de données du NCBI. Le pangénome de *L. monocytogenes* rassemble 48 génomes et plasmides issus en partie de la publication de Di *et al.* [94]. Le pangénome de *M. tuberculosis* est composé de 47 génomes et plasmides provenant en partie des publications de Yang *et al.* [95] et Kavvas *et al.* [96]. Enfin, le pangénome de *S. aureus* regroupe 49 génomes et plasmides issus de la publication de Bosi *et al.* [97]. La souche, le numéro d'accession NCBI, le nombre de gènes et la taille de chacun des génomes pour chacun des six pangénomes sont disponibles en Annexe B (Tableau 19 à 24).

2.3.4.3 Résultats OGM synthétiques

Quarante-deux génomes d'OGM synthétiques ont été créés à partir de la procédure sur les bactéries décrites dans la partie 2.2.3. Ces génomes synthétiques sont basés sur ces six bactéries sauvages (*Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Listeria monocytogenes*, *Lactococcus lactis* et *Salmonella Typhimurium*) et huit gènes artificiellement insérés provenant de bactéries exogènes, de l'homme et du riz. Ces huit gènes ont été choisis aléatoirement avec des tailles diverses. Les gènes insérés pour créer ces génomes synthétiques ainsi que leurs numéros d'accession NCBI sont décrits en Annexe D.

2.4 Conclusion

Plusieurs supports de détection ont été créés pour détecter des OGM inconnus et faciliter les tests de la méthode. Un pangénome pour chacune des espèces bactériennes considérées a été créé ainsi qu'une banque de données d'inserts OGM connus contenant des inserts procaryotes et eucaryotes. Ces deux éléments vont permettre de préparer les données d'entrées nécessaires à l'étape de création d'un modèle de prédiction en fournissant des éléments dont le statut est connu (non OGM pour les CDS du pangénome et OGM pour la banque de données d'inserts). Enfin, de nombreux jeux de données de séquençage d'OGM synthétiques ont été crées à partir de bactéries présentant des caractéristiques différentes pour évaluer et tester la méthode de détection des OGM inconnus.

Chapitre 3

Proposition d'une méthode de caractérisation des séquences codantes (CDS) exceptionnelles d'un génome

Sommaire

L'usage du codon et la recherche de motifs nucléotidiques exceptionnels dans des séquences sont les outils principaux employés pour caractériser les CDS d'un génome. Cette caractérisation des CDS présents dans le génome OGM potentiel est utilisée pour distinguer les CDS du génome hôte des CDS potentiellement GM.

3.1 Objectifs

Notre objectif est d'identifier les différences de statistiques d'utilisation entre les CDS provenant du génome hôte et les séquences codantes insérées. En effet, ces séquences introduites dans le génome hôte ont un enchaînement de nucléotides différent de celui des CDS appartenant au génome hôte. Cet enchaînement de nucléotides est spécifique à chaque génome, il est nommé le vocabulaire.

Notre but est donc de caractériser les CDS présents dans notre OGM candidat à l'aide de leur vocabulaire génomique. Cette caractéristique n'est pas spécifique

à la détection d'OGM inconnus, elle peut être appliquée dans d'autres contextes comme la mise en évidence d'un gène ayant subit un nombre de mutations important. Le vocabulaire du gène possédant des mutations s'éloignera du vocabulaire utilisé par son génome hôte et le gène possédant plusieurs mutations pourra être identifié. L'utilisation de cette propriété existante généraliste est donc appliquée à un domaine particulier, la détection d'OGM inconnus.

3.2 Méthodes

La propriété des génomes qui a été, pour partie, le support de nos investigations pour construire la méthode de détection des OGM inconnus, est l'usage préférentiel de certains triplets de nucléotides dans leurs gènes.

3.2.1 Usage du codon

L'ADN est constitué de quatre nucléotides : Adénine (A), Guanine (G), Cytosine (C) et Thymine (T). Le processus de transcription permet de transcrire un gène (ADN génomique) en ARN messager (ARNm), cet ARNm est alors traduit en protéine à l'aide d'ARN de transfert (ARNt). L'ARNm est composé de régions non codantes (qui ne sont pas traduites en protéines) et de régions codantes (qui sont traduites pour produire des protéines). L'ARNm est constitué de quatre types de bases nucléiques, A, C, G et U (Uracile). Ces quatre bases constituent le code génétique, utilisé pour traduire une séquence d'ADN (gène) en séquence protéique. Le code génétique est universel, il est identique pour les organismes vivants (sauf certaines exceptions). La reconnaissance de triplets de nucléotides, appelés codons, par les ARNt lors de la traduction des ARNm implique donc 64 combinaisons (4^3). Sur ces 64 codons existants, trois correspondent à des signaux de fin de traduction qui ne sont pas traduits, les codons stop. Il reste donc 61 codons pour coder seulement 20 acides aminés naturels, constituants des protéines. Cette spécificité permet de qualifier le code génétique de dégénéré, plusieurs codons différents pouvant correspondre au même acide aminé. Ces codons sont nommés codons synonymes. Cependant, un codon ne peut coder qu'un seul et unique acide aminé. De plus, tous les codons ne sont pas utilisés avec la même fréquence. L'utilisation non aléatoire des codons synonymes crée un biais d'usage du codon spécifique de l'espèce considérée [98] et résulte de l'amélioration du « fitness » de la cellule : la synthèse d'un nombre restreint d'ARNt diminue sa dépense énergétique. Les facteurs responsables de l'usage préférentiel de certains codons comprennent le pourcentage en GC à la troisième position du codon, l'abondance de la molécule d'ARNt, la composition générale des bases de gènes, les différences dans le niveau d'expression et dans la localisation cellulaire des gènes, la température de croissance optimale et les structures secondaires des protéines [99]. La première et la deuxième position du codon sont plus informatives quant à la structure de la protéine codée tandis que la troisième position du codon est plus informative de l'espèce considérée [100]. Les codons synonymes partagent les mêmes deux premiers nucléotides et ne diffèrent qu'à leur troisième nucléotide.

En 2016, A. A. Komar [101] résume, grâce au graphique suivant (Figure 10), l'usage du codon ainsi que l'utilisation et les fonctions des codons synonymes qui en découlent. Ce graphique permet aussi de visualiser les récents développements dans ce domaine qui ont permis d'identifier les fonctions des codons synonymes ainsi que leur utilisation.

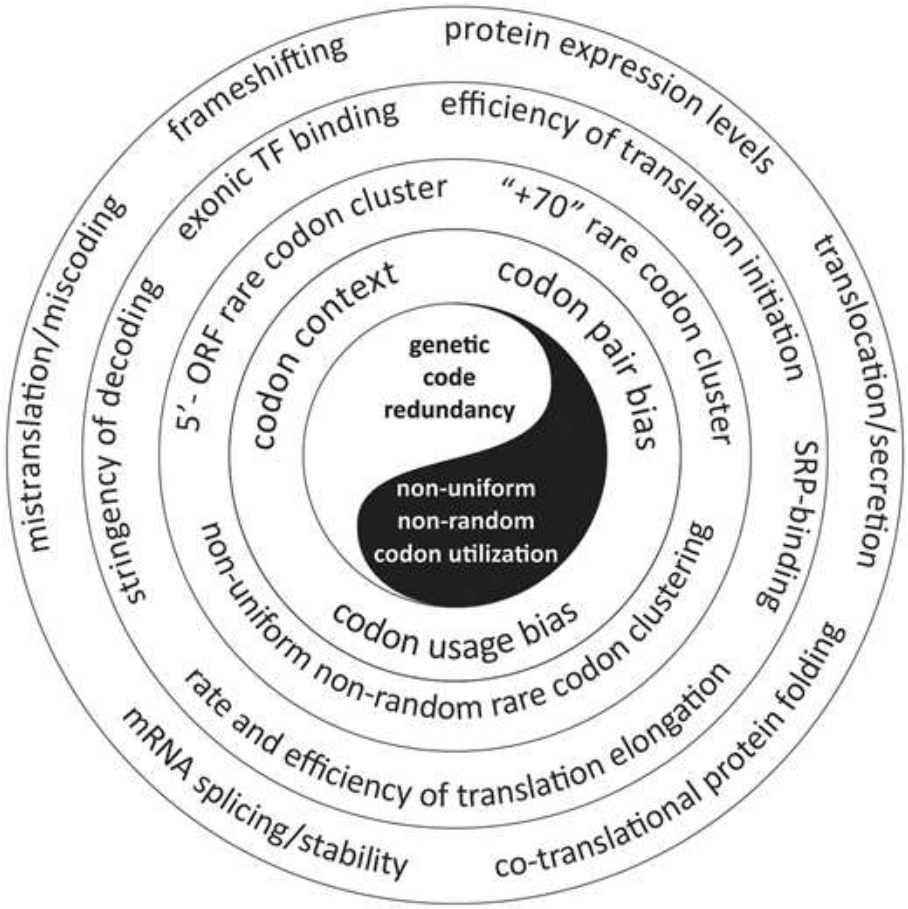


FIGURE 10 – Le ying et le yang de l'usage du codon [101].

3.2.2 Caractérisation du vocabulaire du génome

Le logiciel R'MES [102] permet de déterminer les mots exceptionnels dans une séquence, selon un modèle de Markov M donné pour une taille de mot définie. Un modèle de Markov est constitué d'états. Le passage d'un état à un autre est associé à une probabilité. Dans un modèle de Markov d'ordre k, un état donné dépend des k états précédents. Un mot de taille l ne peut être analysé que dans les modèles de $M0$ à $M(l\text{-}2)$. Des modèles de Markov plus élevés correspondraient au nombre de mots lui-même (le mot serait alors attendu). Le modèle de Markov d'ordre maximal est donc défini par $M(l\text{-}2)$. Le logiciel R'MES est généralement utilisé sur des séquences courtes pour trouver des motifs d'intérêts chez les bactéries. Le temps de calcul d'une analyse est très court et il fournit une approximation acceptable avec des résultats relativement simples à interpréter.

De nombreux paramètres sont disponibles comme la création de familles de mots ou le comptage en phase. Le logiciel R'MES est capable de réaliser une analyse de biais d'orientation, c'est-à-dire déterminer si un mot est plus souvent présent dans un sens ou dans l'autre.

Dans notre cas, le logiciel R'MES a été utilisé pour caractériser le vocabulaire d'un génome, en déterminant ses mots spécifiques et relativement fréquents. Le vocabulaire d'un génome est représenté par l'ensemble des mots d'une longueur donnée ou un sous-ensemble de ces mots particulièrement spécifiques du génome de l'hôte. Le logiciel R'MES utilise aussi une approximation de la répartition du nombre de mots. Cette approximation s'appuie sur l'une des trois lois suivantes : Gaussienne, Poisson ou Poisson composé. L'approximation Gaussienne est utilisée lorsque le génome contient des mots de plus de 100 occurrences, elle est donc adaptée aux génomes ayant une longueur importante. L'approximation de Poisson ne gère pas les chevauchements entre les mots contrairement à l'approximation de Poisson Composé ; elle n'est donc pas utilisée. L'approximation de Poisson Composé possède une précision plus importante que la méthode Gaussienne mais nécessite un temps de calcul plus long. Une indication est stipulée dans le manuel d'utilisation de R'MES [103] pour faciliter le choix de l'approximation suivant la longueur de mots choisie et la longueur de la séquence à analyser : « As a crude rule, calculate the naive, expected count of a word of length l in a sequence of length n namely n/(|A| l) ; If it is smaller than 10 then use the compound Poisson approximation, if it is greater than 100 then use the Gaussian approximation (in between is the twilight zone). »

```
Results from algorithm  : Gauss
Markov Order            : 4
Word Size               : 6
Data Set                : cds_GenomeRef#
Stat Value Interval     : ] -Inf, -0 ] U [ 0, +Inf ]
-------------------------------------------------------------------------
      word |      count |     expect |     sigma2 |      score |   rank
-------------------------------------------------------------------------
    ttggcg |      30452 |  43495.859 |  27050.080 |   -79.3089 |      1
    tagtac |      12949 |  21871.715 |  12963.670 |   -78.3670 |      2
    cgctga |      38202 |  49886.277 |  32091.906 |   -65.2235 |      3
    ctgctt |      77287 |  91679.586 |  50458.102 |   -64.0728 |      4
    tgcggc |      46935 |  56726.824 |  26413.086 |   -60.2496 |      5
    ttgccg |      29068 |  37986.270 |  23195.801 |   -58.5566 |      6
    cgctca |      33626 |  42876.398 |  25702.480 |   -57.6996 |      7
    ctcctt |      53533 |  64206.535 |  34714.469 |   -57.2866 |      8
    cgccca |      32647 |  41118.051 |  21964.293 |   -57.1582 |      9
    gagtaa |      19177 |  26257.834 |  15626.034 |   -56.6448 |     10
```

Tableau 2 – Portion de résultats issue du logiciel R'MES pour certains mots de taille 6 avec un modèle de Markov d'ordre 4 sur l'ensemble des CDS du génome de référence du maïs.

Le logiciel R'MES calcule différentes variables pour chaque mot d'une taille donnée suivant l'approximation choisie. Pour chaque mot en utilisant l'approximation Gaussienne, les variables calculées sont le nombre d'occurrences du mot

observé ou « comptage » (count), l'estimation du nombre d'occurrences attendu du mot (expect), la variance limite estimée (sigma2), le score d'exceptionnalité (score) et le classement du mot au moment de l'évaluation (rank). Les résultats sont présentés sous la forme d'un tableau trié par ordre croissant de score (Tableau 2).

En utilisant l'approximation Poisson composé, les variables calculées sont le nombre d'occurrences du mot observé, l'estimation du nombre d'occurrences attendues du mot, l'estimation du nombre prévu de trains (ensemble de motifs successifs), la probabilité de chevauchements, le score d'exceptionnalité et le classement du mot en considérant un tri par ordre croissant de scores. Le score d'exceptionnalité fourni par le logiciel R'MES évalue l'importance de la différence entre les nombres d'occurrences de mots observés et attendus.

Pour que le vocabulaire d'un génome puisse être défini à l'aide de mots en utilisant le logiciel R'MES, nous voulions que l'ensemble des mots retenus couvrent la totalité du génome avec environ 10 occurrences pour 150 nucléotides pour s'assurer de la repré'sentativité du génome étudié dans une séquence prise au hasard, même de la taille d'un read. Pour cela, l'objectif était de trouver des mots fréquents sur le génome hôte mais absents sur l'insert ou inversement, la signature du génome hôte pouvant consister en l'absence de certains mots rares, en plus de la présence de mots fréquents. La taille de mot choisie ne devait donc être ni trop petite, sinon les occurrences auraient été trop nombreuses, ni trop grande car trop peu d'occurrences auraient été trouvées dans la séquence. L'utilisation du logiciel R'MES est réalisé sur les CDS du génome hôte sans les séquences UTR* (régions non traduites), les introns* et les autres séquences intergéniques. En effet, un insert devrait être constitué d'un ou plusieurs CDS pour que celui-ci puisse s'exprimer dans le génome et ainsi conférer une ou plusieurs caractéristiques additionnelles au génome hôte. Enfin, cette propriété permet de réduire le nombre de nucléotides à analyser par rapport au génome entier, donc de réduire le temps de calcul de l'analyse.

3.2.3 Comptage des occurrences

Dans un premier temps, l'ensemble des CDS du génome hôte est soumis au logiciel R'MES [102] où le nombre d'occurrences pour chaque mot est ainsi déterminé. Dans un second temps, un comptage des mots est réalisé pour chacun des CDS appartenant aux trois jeux de données à la fin du pipeline de nettoyage (Chapitre 4 partie 4.2.1) : les CDS apparentés au génome hôte, les CDS potentiellement inserts et les CDS de la banque de données d'OGM connus avec l'algorithme Aho-Corasick (Figure 11). Ces comptages de mots de chacun des CDS des trois jeux de données sont ensuite comparés au comptage de l'ensemble des CDS du génome hôte à l'aide de calculs de distances (Chapitre 5 partie 5.2.1).

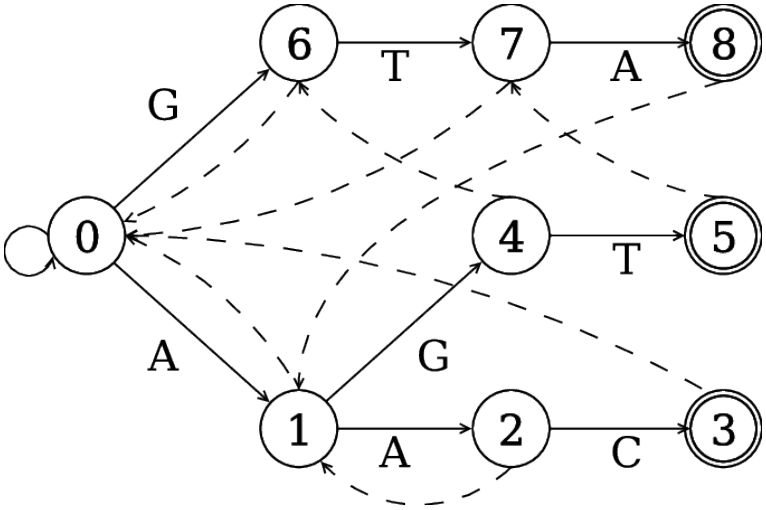


FIGURE 11 – Visualisation de l'automate de l'algorithme Aho-Corasick pour l'ensemble des motifs "AAC", "AGT" et "GTA" [104].

Publié en 1975, l'algorithme Aho-Corasick est un algorithme de recherche de chaînes de caractères (motifs) conçu par Alfred Aho et Margaret Corasick [105]. Il utilise des dictionnaires qui localisent les éléments d'un ensemble fini de chaînes dans un texte. La complexité en temps de l'algorithme est linéaire, elle correspond à la longueur des motifs considérés ajoutée à la taille du texte où ces motifs vont être recherchés. Cet algorithme est très efficace en temps de calcul et en utilisation d'espace mémoire.

3.3 Application

Les génomes du maïs et des bactéries *B. subtilis* et *E. coli* ont été utilisés pour décrire la caractérisation de leurs CDS.

3.3.1 Cas du maïs

Les particularités liées à l'usage du codon chez les plantes et plus particulièrement chez le maïs ainsi que les choix réalisés pour les paramètres du logiciel R'MES sont exposés dans les parties suivantes.

3.3.1.1 Propriétés de l'usage du codon

Dans la Figure 12, l'usage des codons sur 100 gènes a été observé chez les dicotylédones. La moyenne du pourcentage en GC de la troisième position (XXC/G) du codon est de 45,0%. Chez les monocotylédones, l'usage du codon sur 63 gènes a été observé. La moyenne du pourcentage en XXC/G est de 73,5%. Contrairement au dicotylédones, la distribution du pourcentage en GC de la troisième position des codons des monocotylédones est bimodale (Figure 12). Les gènes nucléaires des monocotylédones et des dicotylédones utilisent plus souvent des

codons finissant par C ou G que les gènes chloroplastiques ou mitochondriaux [106].

L'utilisation des codons du maïs est principalement déterminée par la composition nucléotidique et le niveau d'expression génétique. Les variations dans l'usage du codon entre les gènes pourraient être dues à des biais mutationnels au niveau de l'ADN et à la sélection naturelle agissant au niveau de la traduction de l'ARNm [107].

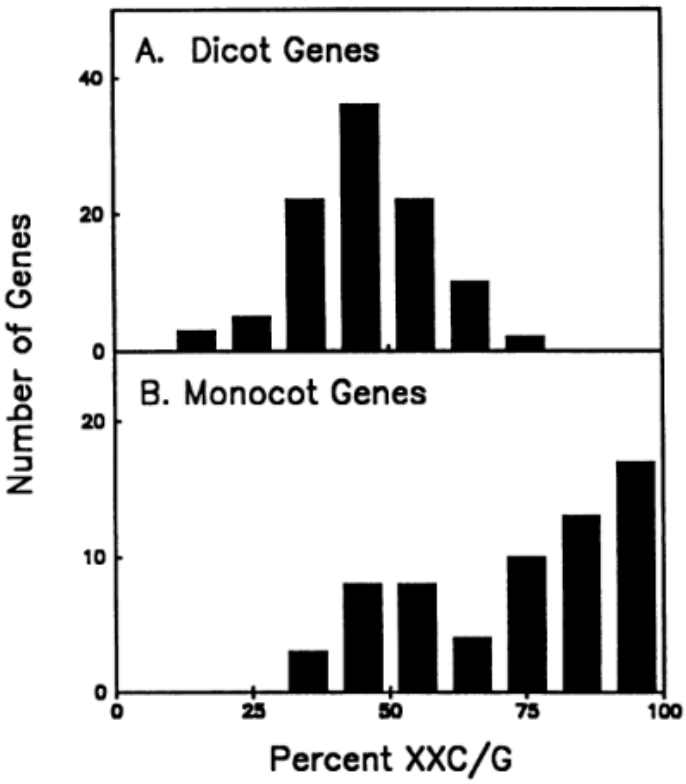


FIGURE 12 – Distribution du pourcentage des codons ayant un C ou un G en troisième position des gènes localisés dans le noyau des dicotylédones et monocotylédones [106].

Le biais d'usage du codon chez le maïs est représenté sur la Figure 13. Chez le maïs, il convient de noter que pour coder l'Alanine, dont les codons synonymes sont GCU, GCC, GCA et GCG, le codon GCC est plus fréquemment utilisé que les trois autres : 31% pour GCC contre 20% en moyenne pour les autres. De même pour la Lysine, le codon AAG sera plus souvent utilisé que le codon AAA : 40% pour AAG contre 15% pour AAA. Il est intéressant de remarquer que, pour coder la Proline, les codons synonymes CCG, CCA, CCC et CCU ont des fréquences d'utilisation très similaires (environ 13%) même si l'utilisation du codon CCG est légèrement plus importante (15%).

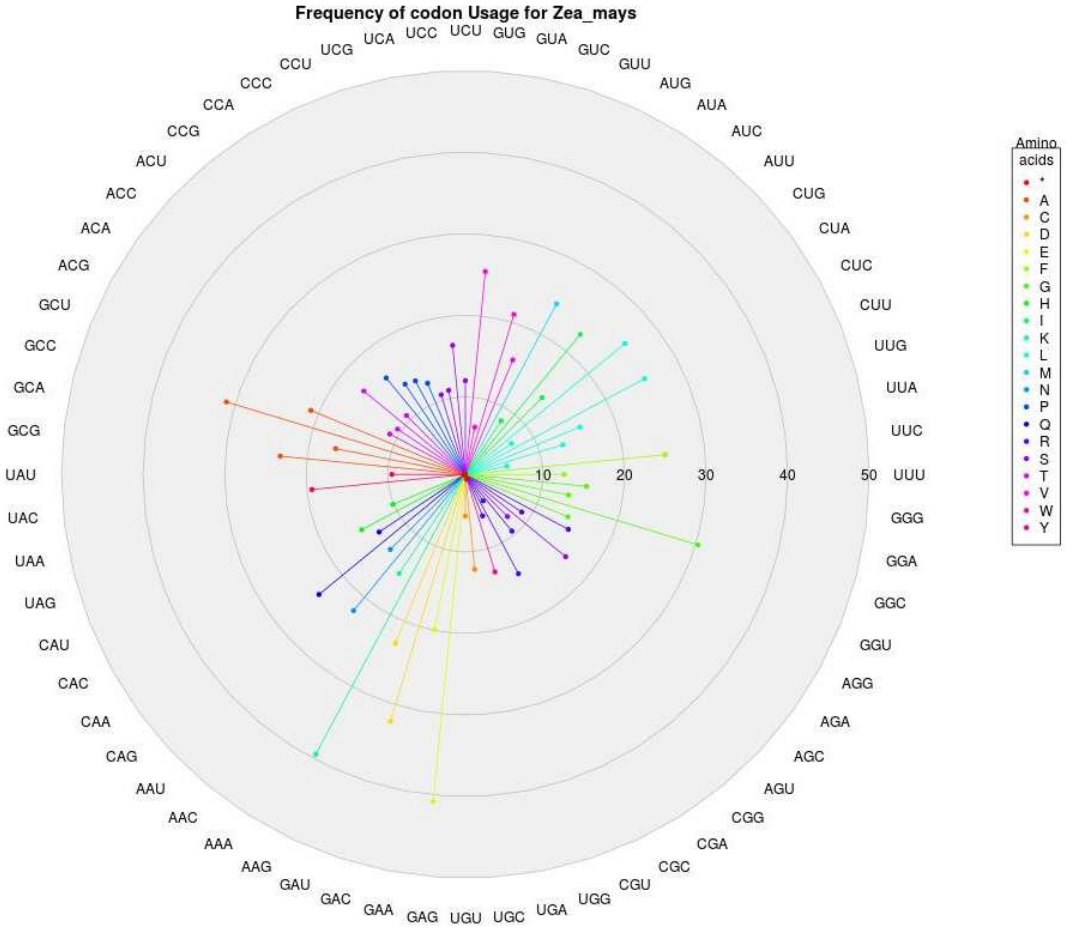


FIGURE 13 – Fréquence d'utilisation des codons pour le maïs exprimée pour 1000 codons [108]. Acides aminés : * : STOP, A : Alanine, C : Cystéine, D : Aspartate, E : Glutamate, F : Phénylalanine, G : Glycine, H : Histidine, I : Isoleucine, K : Lysine, L : Leucine, M : Méthionine, N : Asparagine, P : Proline, Q : Glutamine, R : Arginine, S : Sérine, T : Thréonine, V : Valine, W : Tryptophane, Y : Tyrosine.

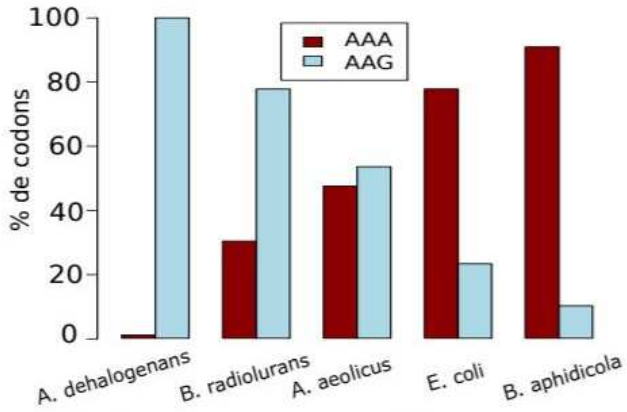
3.3.1.2 Loi de répartition du nombre de mots

Concernant le choix de la loi de répartition du nombre de mots dans le logiciel R'MES, l'approximation de Poisson a été écartée car elle ne gère pas les chevauchements. L'approximation Poisson Composé, plus précise que l'approximation Gaussienne mais nécessitant un temps de calcul plus long a aussi été écartée. Par exemple pour des mots de taille 9, le temps de calcul avec l'approximation Poisson Composé nécessite plusieurs semaines contre quelques heures avec l'approximation Gaussienne. Pour la suite des analyses, l'approximation Gaussienne est donc utilisée. Le génome de référence du maïs, d'une taille de 2182 Mb, présentant en moyenne toujours plus de 100 occurrences pour tous les mots constitués d'au plus dix lettres. Même pour des génomes bactériens, dont la taille se mesure en millions de lettres, et du fait de la prise en compte de mots de petite taille, l'approximation Gaussienne est retenue.

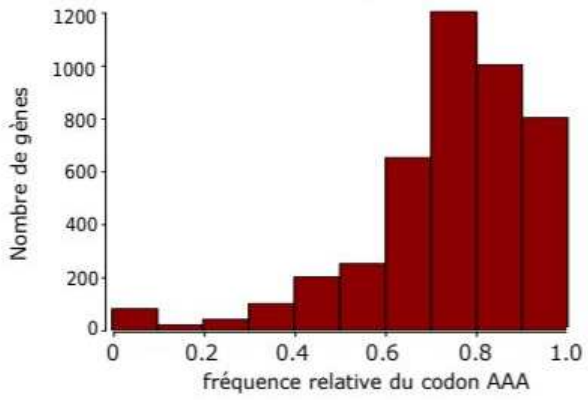
3.3.2 Cas des bactéries *B. subtilis* et *E. coli*

L'usage du codon n'est pas universel et dépend notamment de l'espèce considérée. Ainsi, le biais d'usage du codon représentant un profil d'utilisation des codons non uniformes varie selon l'organisme et représente donc une caractéristique unique de ce dernier [101]. Dans les organismes bactériens, comme *E. coli* ou *B. subtilis*, l'utilisation du codon est attribuable à l'équilibre entre la sélection naturelle et le biais de mutation [109, 110].

La différence d'utilisation des codons synonymes de la lysine chez différentes bactéries est visible sur la Figure 14. La différence d'utilisation pour le codon AAG est de 9% chez *Buchnera aphidicola* alors qu'elle est de 99% chez *Anaeromyxobacter dehalogenans* (Figure 14a). La variation du biais d'usage du codon au sein d'un organisme peut être très importante. De plus, ces différences de biais sont aussi présentes au sein d'un génome. L'utilisation du codon AAA de la lysine chez plusieurs gènes d'*E. coli* diffère (Figure 14b).



(a) Chez différentes espèces bactériennes



(b) Chez *Echerichia coli*

FIGURE 14 – Variation du biais d'usage des codons de la lysine chez différentes bactéries ou au sein de la bactérie *E. coli* [111].

Dans le Tableau 3, nous notons que les bactéries *B. subtilis* et *E. coli* présentent une utilisation différente des codons. Par exemple, pour le codon CGU codant pour l'arginine, la fréquence d'utilisation est de 7,2% par rapport à l'en-

semble des codons présents dans *B. subtilis* tandis que pour *E. coli* elle est de 15,9% soit plus du double par rapport à *B. subtilis*. Cependant ces deux bactéries présentent des similitudes pour les codons associés à l'Alanine. En effet, pour le codon GCU (un des codons associés à l'Alanine), la fréquence d'utilisation est de 18,6% par rapport à l'ensemble des codons présents dans *B. subtilis* tandis que pour *E. coli* elle est de 18,9%.

Pour les gènes de *E. coli*, il existe une forte corrélation entre les codons optimaux et le niveau d'expression des gènes. En effet, les bais de codons des protéines ribosomiques sont les plus déviants par rapport aux fréquences de codon d'un gène de *E. coli*. Pour les autres classes de gènes ayant un niveau d'expression élevé comme les gènes de modifications essentiels aux activités de traduction, les protéines chaperons* et les ARNt synthétases, les biais de codons sont moins importants. La taille des gènes influence aussi les fréquences de GC dans les troisièmes positions des codons [112].

Codons	AA	BS	EC	Codons	AA	BS	EC	Codons	AA	BS	EC	Codons	AA	BS	EC
UUU	F	30.0	24.4	UCU	S	12.7	13.1	UAU	Y	23.3	21.6	UGU	C	3.6	5.9
UUC	F	14.3	13.9	UCC	S	8.3	9.7	UAC	Y	12.6	11.7	UGC	C	4.3	5.5
UUA	L	19.8	17.4	UCA	S	14.6	13.1	UAA	*	1.9	2.0	UGA	*	0.8	1.1
UUG	L	15.8	12.9	UCG	S	6.5	8.2	UAG	*	0.5	0.3	UGG	W	10.7	13.4
CUU	L	21.8	14.5	CCU	P	10.6	9.5	CAU	H	15.7	12.4	CGU	R	7.2	15.9
CUC	L	10.7	9.5	CCC	P	3.5	6.2	CAC	H	7.5	7.3	CGC	R	8.2	14.0
CUA	L	4.9	5.6	CCA	P	7.1	9.1	CAA	Q	20.4	14.4	CGA	R	4.3	4.8
CUG	L	23.0	37.4	CCG	P	16.3	14.5	CAG	Q	18.5	26.7	CGG	R	6.9	7.9
AUU	I	36.2	29.6	ACU	T	8.7	13.1	AAU	N	22.9	29.3	AGU	S	6.8	13.2
AUC	I	27.2	19.4	ACC	T	9.0	18.9	AAC	N	17.8	20.3	AGC	S	14.4	14.3
AUA	I	9.8	13.3	ACA	T	21.6	15.1	AAA	K	48.4	37.2	AGA	R	10.5	7.1
AUG	M	26.3	23.7	ACG	T	14.9	13.6	AAG	K	20.8	15.3	AGG	R	4.1	4.0
GUU	V	18.6	21.6	GCU	A	18.6	18.9	GAU	D	33.2	33.7	GGU	G	13.0	23.7
GUC	V	17.3	13.1	GCC	A	16.5	21.6	GAC	D	19.0	17.9	GGC	G	23.3	20.6
GUA	V	13.0	13.1	GCA	A	21.1	23.0	GAA	E	48.1	35.1	GGA	G	21.8	13.6
GUG	V	17.3	19.9	GCG	A	19.8	21.1	GAG	E	22.6	19.4	GGG	G	11.2	12.3

Tableau 3 – Usage du codon exprimé en fréquence par millier de codons et leurs acides aminées (AA) correspondant pour les bactéries *Bacillus subtilis* (BS) et *Escherichia coli* (EC) [108].

3.4 Conclusion

L'usage préférentiel de certains triplets de nucléotides (codons) est au centre de la méthode détection des OGM inconnu pour distinguer la ou les séquences qui sont insérées dans un OGM. Le logiciel R'MES associé à l'approximation Gaussienne est utilisé pour définir des motifs exceptionnels dans le génome hôte. Par la suite, différents paramètres de ce logiciel comme la taille de mot ou l'ordre de Markov sont testés pour sélectionner les paramètres optimaux permettant de distinguer des CDS inserts OGM de CDS du génome hôte. Cependant, pour

valider les résultats émis avec ces paramètres, il est nécessaire d'utiliser les calculs de distance (Chapitre 5 partie 5.2.1). Pour cette raison, tous les paramètres du logiciel R'MES testés et les résultats associés sont présentés dans le Chapitre 5.

Chapitre 4

Préparation des données en vue de l'application des calculs de distance

Sommaire

La préparation des données est nécessaire pour séparer les CDS apparentés au génome hôte des CDS potentiellement OGM en vue des calculs de distance. Cette préparation permet d'obtenir des jeux de données fiables et de bonne qualité en évitant les erreurs liées à un mauvais tri des données.

4.1 Objectif

Un pipeline de nettoyage a été développé dans le but d'éliminer les CDS apparentés au génome hôte et ainsi de ne conserver que les CDS potentiellement inserts GM. L'emploi d'un pipeline de nettoyage permettant la suppression d'éléments connus est fréquemment utilisé dans le traitement des données de séquençage haut débit. Par exemple, Baron *et al.* 2018 [113]) utilisent un traitement similaire pour éliminer les reads chromosomiques afin d'assembler un plasmide de *E. coli*.

Les données de séquençage du génome potentiellement OGM inconnu sont triées pour faciliter leur traitement et ainsi mettre en évidence la ou les séquences insérées dans le génome hôte. Notre objectif est d'obtenir trois jeux de données distinct :

— les CDS apparentés au génome hôte,

— les CDS potentiellement inserts OGM,
— les CDS de la banque de données d'OGM connus.

En complément du pipeline de nettoyage, le pangénome de l'espèce considérée est utilisé pour trier les CDS de l'OGM inconnu et les données de la banque de données d'OGM connus sont filtrées pour correspondre à l'OGM inconnu analysé.

4.2 Méthodes

La méthode de détection d'OGM s'appuie sur la distinction entre CDS apparentés au génome hôte et CDS d'inserts. Un premier tri est réalisé avec le pipeline de nettoyage puis un second tri est fait à l'aide d'un pangénome de l'espèce (voire la sous-espèce) analysée dans le but de distinguer les CDS sauvages sûrs à fournir lors de la construction d'un modèle de prédiction.

4.2.1 Pipeline de nettoyage

Un pipeline de nettoyage est comme son nom l'indique une succession d'outils ou d'étapes permettant de nettoyer les données de séquençage brutes. En complément du nettoyage des données de séquençage, notre pipeline est aussi utilisé pour trier les CDS du génome potentiellement GM selon deux catégories : les CDS appartenant au génome sauvage et les CDS d'inserts OGM potentiels (Figure 15).

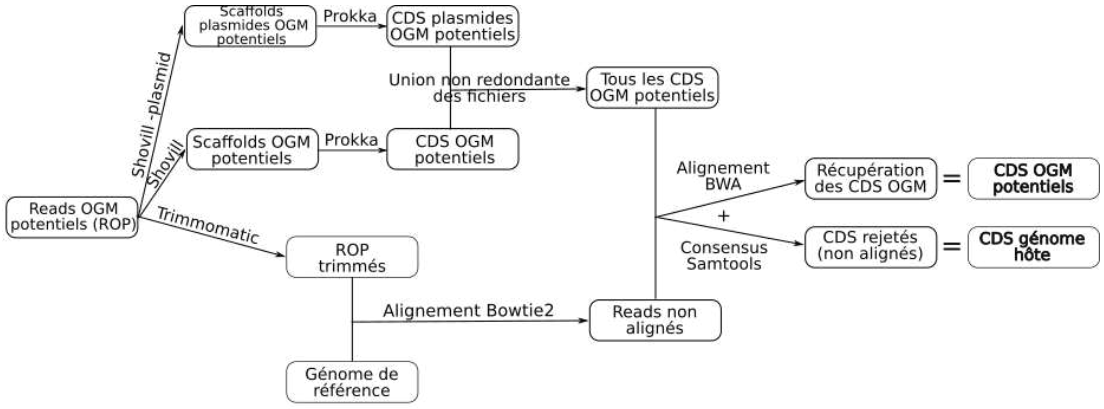


FIGURE 15 – Schéma du pipeline de nettoyage pour les données brutes de séquençage du génome suspecté d'être modifié génétiquement.

En premier lieu, les données brutes de séquençage du génome GM potentiel sont assemblées deux fois avec l'outil Shovill [114] où l'option « --trim » est utilisée pour assurer la suppression des oligomères de construction des librairies de séquençage. Un logiciel d'assemblage permet de fusionner les reads par chevauchement de régions similaires dans le but d'obtenir des contigs qui seront eux mêmes assemblés en scaffolds (Figure 16). Shovill est un pipeline utilisant l'assembleur SPAdes [115] où les étapes avant et après l'assemblage principal sont modifiées. Shovill permet ainsi d'obtenir des résultats similaires à l'utilisation de SPAdes mais en un temps réduit. L'outil Shovill est uniquement compatible avec des génomes bactériens. Le premier assemblage Shovill est réalisé en vue d'obtenir

le génome entier et le second est optimisé pour assembler les plasmides avec l'option « --plasmid » de SPAdes. Les scaffolds obtenus possédant une profondeur de couverture* inférieure à 3 sont éliminés. Ce seuil modifiable sert à éliminer toute contamination potentielle du génome bactérien, telle que la contamination par une autre librairie de séquençage ou par la matrice dans laquelle les bactéries sont situées. Les deux résultats des assemblages Shovill sont annotés à l'aide de l'outil Prokka [72] en utilisant l'option « --use-genus » qui permet de considérer prioritairement les annotations de bactéries du même genre. Les annotations sont ensuite regroupées en supprimant les doublons. En parallèle, les données brutes de séquençage de l'OGM potentiel sont nettoyées avec l'outil Trimmomatic [116]. Trimmomatic est un outil permettant de supprimer les adaptateurs ajoutés lors de la préparation des librairies de séquençage et de « trimmer » les données de séquençage brutes. C'est à dire qu'il élimine les nucléotides en bord de read présentant une qualité insuffisante (les paramètres peuvent être modifiés en fonction de la technologie du séquenceur utilisée). Il peut être utilisé sur des données pairées ou non pairées.

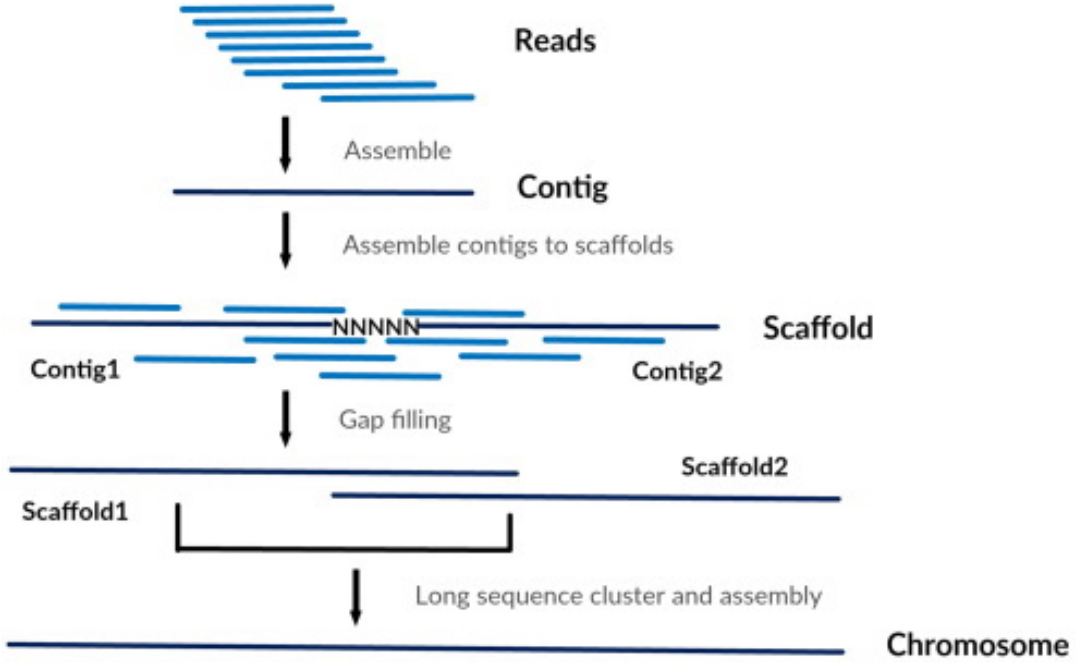


FIGURE 16 – Représentation schématique des reads, contigs et scaffolds [117].

Les données de séquençage nettoyées de l'OGM potentiel sont ensuite alignées via le logiciel d'alignement Bowtie2 [118] sur le génome de référence. Les reads qui ne se sont pas alignés (incluant tous les CDS inserts GM) sont ensuite alignés sur les CDS prédits lors des annotations précédentes issues de l'outil Prokka à l'aide de l'aligneur BWA MEM [119]. Un consensus d'alignement* des CDS couverts est ensuite déduit des reads qui se sont alignés en utilisant la suite de programmes Samtools [120].

Par défaut, l'aligneur Bowtie2 réalise des alignements pleine longueur : la totalité de la séquence d'un read est alignée sur un génome de référence préalablement

renseigné par l'utilisateur. A l'inverse, l'aligneur BWA permet quant à lui d'aligner partiellement des reads sur le génome de référence, alignant donc aussi des reads chimériques. Des reads chimériques sont des reads qui s'alignent (partiellement) sur deux portions différentes non contiguës du génome de référence (Figure 17). Il est donc plus sensible que l'aligneur Bowtie2. L'aligneur Bowtie2 permet donc d'éliminer les reads complets du génome suspect très proches de ceux du génome de référence et de conserver les reads correspondant potentiellement à des inserts GM. Le second alignement avec BWA MEM récupère les alignements partiels de CDS pouvant correspondre à des protéines tronquées ou fusionnées avec un CDS exogène.

A la fin de ces différentes étapes, les CDS présents dans le consensus d'alignement donné par Samtools, contenant les séquences d'inserts OGM potentiels, sont retenus comme CDS GM potentiels. Les CDS restants sont conservés comme CDS apparentés au génome hôte.

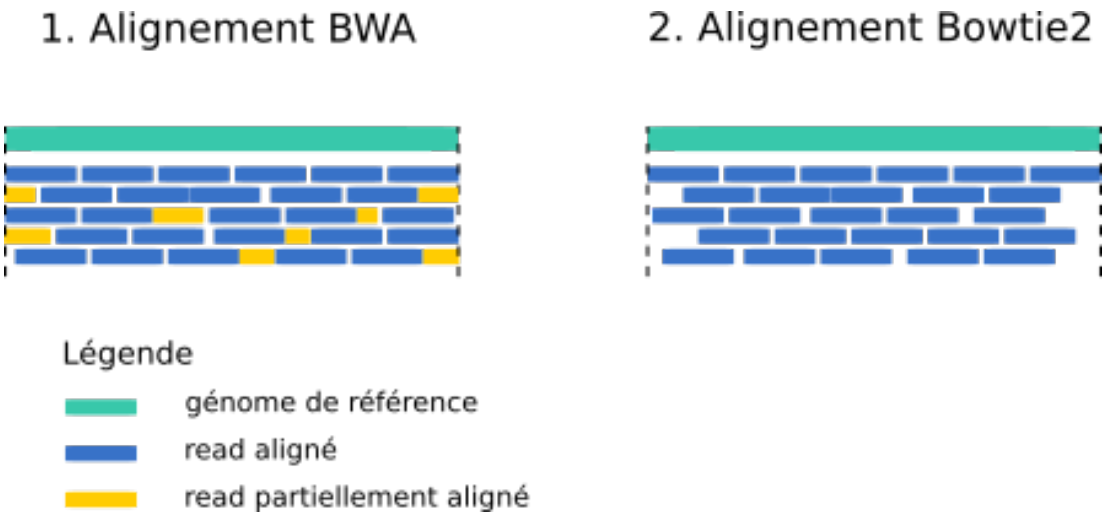


FIGURE 17 – Différences d'alignements entre les aligneurs BWA et Bowtie2.

4.2.2 Tri des CDS à l'aide du pangénome de l'hôte

Deux alignements BLASTN [121] sont réalisés entre les CDS GM potentiels fournis à la fin du pipeline de nettoyage et le pangénome de l'espèce bactérienne correspondant au génome potentiellement OGM. Le premier alignement utilise les CDS du pangénome alors que le second utilise le pangénome, comportant génomes et plasmides entiers (Figure 18). L'objectif de ces alignements BLASTN est de réaliser un second tri (le premier étant effectué par le pipeline de nettoyage) permettant de séparer les CDS apparentés au génome hôte et les CDS GM potentiels.

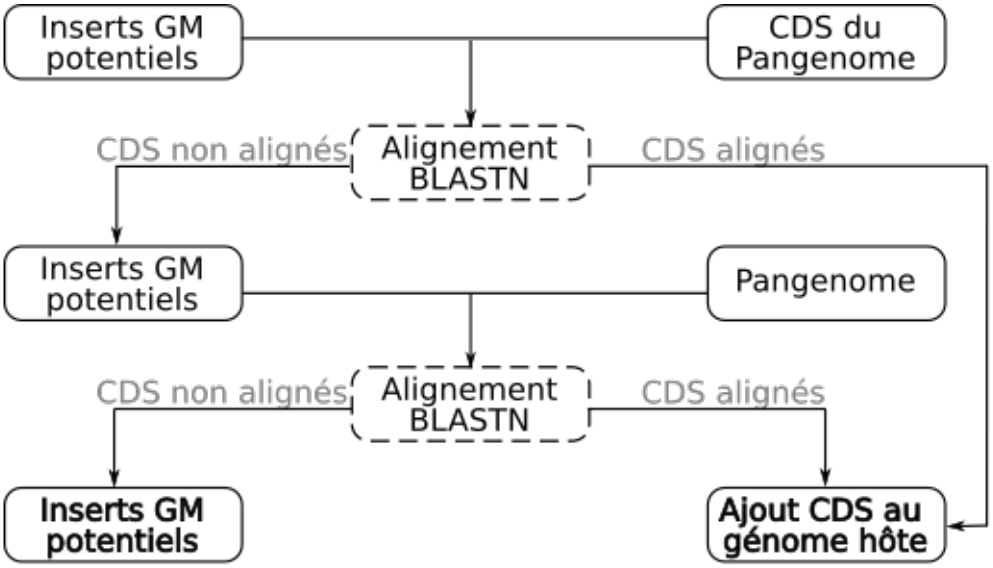


FIGURE 18 – Schéma des deux alignements BLASTN sur les CDS et les génomes/plasmides entiers du pangénome permettant d'effectuer un second tri sur les CDS OGM potentiels.

Le premier alignement BLASTN est réalisé entre les CDS GM potentiels fournis à la fin du pipeline de nettoyage et les CDS du pangénome. Cet alignement élimine les CDS appartenant au génome hôte présents dans les CDS inserts GM potentiels. Les paramètres utilisés sont les options « --gapopen » et « --gapextend » abaissées à 3 et 1 respectivement. Ces options permettent de définir les coûts d'ouverture et d'extension d'un gap* (espacement d'un ou plusieurs nucléotides) dans l'aligneur. Ces options permettent d'éviter de couper des protéines possédant un gap interne. En effet, si un CDS apparenté au génome hôte possède un gap interne de plusieurs nucléotides alors il sera caractérisé comme CDS insert OGM potentiel si les options « --gapopen » et «--gapextend » sont utilisées avec leur valeur par défaut, soit 5 et 2 respectivement. Par ailleurs, l'option « --dust no » est employée pour que tous les alignements du pangénome soient pris en compte (y compris les gènes au contenu informatif faible) : nous sommes assurés que le contenu de nos pangénomes est informatif. Cette option générale de BLASTN (« --dust yes ») n'est donc pas nécessaire. Les CDS retenus lors de ce premier BLASTN présentent un pourcentage d'identité supérieur à 98%, ou un pourcentage de couverture de la séquence recherchée (query) par HSP (High-scoring Segment Pair) supérieur à 98% et un pourcentage d'identité supérieur à 95%. Le HSP est une unité fondamentale de BLAST. Un HSP correspond à un alignement local sans gaps (insertion-délétion), obtenant l'un des meilleurs scores d'alignement dans une recherche donnée. Ensuite, les CDS dont la longueur d'alignement est incluse dans un intervalle de plus ou moins 15% de la longueur du CDS GM potentiel (longueur de la requête) et de la longueur du CDS du pangénome (longueur de l'objet) sont retenus. Ce paramètre permet d'assurer un pourcentage de couverture d'alignement élevé sur les CDS présentant un gap important par rapport aux CDS du pangénome.

Le second alignement BLASTN, réalisé sur les génomes entiers du pangénome élimine les CDS prédits par l'annotateur Prokka mais correspondant à des séquences non codantes dans notre pangénome. Tout comme le premier alignement BLASTN, les options « --gapopen » et « --gapextend » sont fixées à 3 et 1 respectivement et l'option « --dust no » est utilisée. Les CDS retenus présentent un pourcentage d'identité supérieur à 98% et un pourcentage de couverture de la requête par HSP supérieur à 98%. Seuls les CDS dont la longueur d'alignement est incluse dans un intervalle de plus ou moins 15% de la longueur du CDS GM potentiel sont retenus.

Les CDS alignés au moins une fois lors de ces deux alignements BLASTN possèdent de très fortes similarités avec le pangénome de l'espèce considérée, ils sont donc ajoutés aux CDS apparentés au génome hôte définis à la fin du pipeline de nettoyage. Les CDS non alignés sont conservés comme CDS OGM potentiels.

4.2.3 Filtrage de la banque de données d'OGM connus

La banque de données d'inserts OGM connus va être filtrée pour obtenir un ensemble de CDS OGM connus dénué des CDS sauvages de l'espèce hôte considérée. Pour cela, tous les CDS de la banque de données OGM naturellement présents dans le génome de l'espèce de la souche suspectée vont être éliminés. En effet, un insert n'est qualifié d'OGM que s'il est considéré dans un contexte différent de celui du génome dans lequel il est naturellement présent. Deux alignements BLASTN successifs sont donc réalisés pour filtrer cette banque de données. Le premier élimine les CDS OGM connus s'alignant sur les CDS du pangénome et le second, ceux s'alignant sur les CDS apparentés au génome hôte.

Pour ces deux alignements BLASTN, comme précédemment, les options « --gapopen » et « --gapextend » sont fixées respectivement à 3 et 1 pour éviter de couper des protéines ayant un gap interne. De plus, l'option « --dust no » est employée pour n'éliminer aucun alignement. Les CDS écartés lors des alignements doivent obtenir un pourcentage d'identité minimum de 95% avec un pourcentage de couverture par HSP de 98% ou un pourcentage d'identité supérieur à 98%. Ces paramètres ont été choisis pour assurer une sélectivité importante et n'éliminer que les CDS très proches du génome hôte. De plus, les parties des CDS qui s'alignent lors de ces alignements BLASTN sont supprimées de la banque de données d'OGM connus. La valeur minimale de la e-value est fixée à 0.00001 pour être assurés d'obtenir des alignements probants (dans BLASTN par défaut elle est fixée à 10).

Ensuite, la médiane pour chacune des trois distances (P L3M1, P L4M2 et F L9M7 décrites ultérieurement dans le chapitre 5 partie 5.4) est calculée entre chaque CDS du génome de l'hôte et l'ensemble de tous les CDS du génome de l'hôte. Ces médianes ont été nommées respectivement medL3M1, medL4M2 et medL9M7. Enfin, les CDS OGM connus de la banque de données dont les distances par rapport à l'ensemble de tous les CDS du génome hôte sont vérifiées (P L3M1 $<$ medL3M1 et P L4M2 $<$ medL4M2 et F L9M7 $<$ medL9M7) sont rejetés. La banque de données filtrée sera utilisée lors de l'élaboration d'un modèle de prédiction en complément des CDS apparentés au génome hôte pour constituer

les données d'apprentissage.

4.3 Application

4.3.1 Cas du maïs

Au début du projet, un pipeline de nettoyage a été mis au point à partir des données de séquençage du maïs OGM (Chapitre 2 partie 2.3.1.1). Ce pipeline est décrit en Annexe C. Ce pipeline de nettoyage utilise les CDS du génome de référence de maïs comme données connues non-OGM contrairement au pipeline de nettoyage développé avec la bactérie *B. subtilis* qui utilise les CDS apparentés au génome hôte. Il n'est pas possible d'assembler et d'annoter la totalité d'un génome de maïs dans un délai et avec des ressources informatiques raisonnables en raison de sa taille trop importante : la taille du génome du maïs est d'environ 2197.97 Mb. En comparaison, le génome de la bactérie *B. subtilis* possède une taille de 4.13 Mb. Ce pipeline est utilisé pour produire les données utilisées pour tester certains paramètres du logiciel R'MES (Chapitre 5 partie 5.3.2).

4.3.2 Cas de la bactérie *B. subtilis*

Pour les données de séquençage de la bactérie *B. subtilis*, l'assembleur SPAdes a été utilisé avec l'option « --careful ». Dans le cas de l'assemblage d'un génome bactérien, nous pouvons obtenir de mauvais assemblages si la profondeur de couverture (le nombre de reads alignés par position en moyenne) est trop importante. Il est préférable dans ce cas de sous-échantillonner les reads de façon à avoir une profondeur de couverture estimée de 80 [122]. Un script Perl développé au laboratoire a été utilisé lorsqu'il y avait besoin de sous-échantillonner les données. Le sous-échantillonnage est réalisé en s'appuyant sur un tirage aléatoire des reads retenus.

4.4 Conclusion

Un pipeline de nettoyage est crée pour nettoyer les données de séquençage et réaliser un premier tri entre les CDS apparentés au génome hôte et les CDS potentiellement OGM. Un second tri mis en place à l'aide d'alignements sur le pangénome de l'espèce analysée est réalisé. Enfin, la banque de données d'OGM connus est filtrée à l'aide des CDS du génome hôte. A la fin de ces étapes de préparation des données, trois jeux de données sont obtenus : les CDS apparentés au génome hôte, les CDS potentiellement inserts OGM et les CDS de la banque de données d'OGM connus filtrés. Ils seront utilisés en Chapitre 5.

Chapitre 5

Proposition d'une méthode de comparaison des séquences codantes (CDS) de génomes

Sommaire

Pour comparer les CDS du génome OGM inconnu potentiel, plusieurs calculs de distance sont utilisés en complément de filtres ou de paramètres spécifiques au logiciel R'MES (recherche de motifs exceptionnels dans une séquence). L'ensemble des tests effectués pour obtenir la combinaison de paramètres idéale pour distinguer au mieux les CDS inserts OGM des CDS du génome hôte est décrit dans ce chapitre.

5.1 Objectif

Pour différencier la séquence insérée (insert OGM) du génome hôte, une distance entre les CDS apparentés au génome hôte et ceux potentiellement insert OGM est calculée. L'objectif est de mettre en évidence les CDS potentiellement inserts OGM possédant une distance élevée entre eux et les CDS appartenant au génome hôte, c'est-à-dire présentant une différence de vocabulaire importante avec le génome hôte. Le résultat recherché est l'obtention d'un nombre de CDS potentiellement OGM restreint contenant l'insert bien séparé en raison de sa distance importante par rapport aux autres CDS.

Plusieurs études ont utilisé un calcul de distance pour évaluer la similarité entre deux séquences [123] et ainsi détecter des gènes issus de transferts horizontaux [7] ou identifier des séquences virales propres à une espèce [8]. Cependant, un calcul de distance n'a jamais été exploité dans le cadre de la détection d'OGM.

Des calculs de distances complémentaires sont aussi réalisés entre les CDS apparentés au génome hôte et les CDS de la banque de données d'OGM inconnus. Ces distances constitueront certaines des variables explicatives utilisées lors de l'étape de construction du modèle de prédiction (Chapitre 6).

5.2 Méthodes

5.2.1 Distances

Dans l'objectif de comparer les différences de vocabulaire entre le génome hôte et son potentiel insert, trois formules de distances ont été testées.

5.2.1.1 Principe

A la fin des étapes décrites dans le chapitre précédent (Chapitre 4 partie 4.4), trois ensembles disjoints de CDS ont été produits : les CDS de la banque de données d'OGM connus filtrés (Figure 19, a), les CDS apparentés au génome hôte (Figure 19, b) et les CDS potentiellement inserts OGM (Figure 19, c). Plusieurs calculs de distance sont réalisés sur ces différents jeux de données (Figure 19).

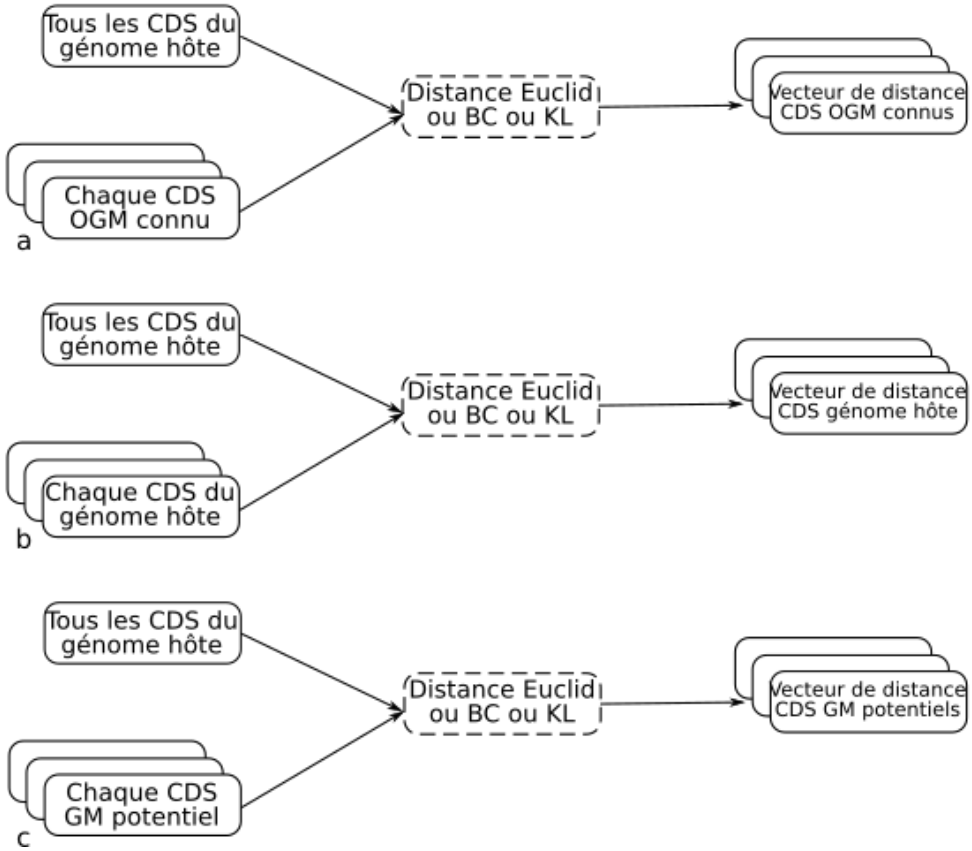


FIGURE 19 – Principe du calcul des distances Euclidienne (Euclid) ou de Bray-Curtis (BC) ou de Kullback-Leibler (KL). a - jeu de données des CDS de la banque de données d'OGM connus filtrés, b - jeu de données des CDS apparentés au génome hôte et c - jeu de données des CDS potentiellement inserts OGM.

Les distances entre l'ensemble des CDS apparentés au génome hôte et chacun des CDS appartenant aux trois jeux de données sont calculées. Par exemple, un calcul de distance est réalisé entre l'ensemble des CDS apparentés au génome hôte et un CDS potentiellement OGM ou entre l'ensemble des CDS apparentés

au génome hôte et un CDS apparenté au génome hôte. Cette opération est donc répétée avec chacun des CDS présents dans les trois jeux de données. Comparer les CDS un à un par rapport aux autres n'est pas envisageable à cause du grand nombre de combinaisons possible, donc du temps de calcul extrêmement long qui en découlerait.

Trois formules de calcul de distance sont sélectionnées pour répondre à ce problème : la distance Euclidienne, la distance de Bray-Curtis et la distance de Kullback-Leibler. La distance Euclidienne représente une distance basique, fondamentale. La distance de Bray-Curtis est associée à des calculs de distance utilisant des k-mers [124]. La distance de Kullback-Leibler est utilisée dans un objectif similaire à la caractérisation de séquences [8].

En mathématiques, une « vraie » mesure de distance vérifie les trois propriétés suivantes :

1. la symétrie : dist(A,B) = dist(B,A)
2. la séparation : dist(A,B) $\geq$ 0 et dist(A,B) = 0 si et seulement si A = B
3. l'inégalité triangulaire : dist(A,B) $\leq$ dist(A,C) + dist(C,A)

où dist(A,B) représente la distance entre les points A et B. La symétrie indique que la distance entre le point A et le point B est identique à celle entre le point B et le point A. La séparation signifie que la distance doit toujours être positive sauf lorsque A et B sont identiques, elle est alors égale à 0. Enfin, l'inégalité triangulaire indique que la distance entre deux points est inférieure à l'addition des deux autres distances d'un triangle. Les distances ne respectant pas ces propriétés sont parfois appelées dissimilarités. La distance de Bray-Curtis ne respectant pas la propriété de l'inégalité triangulaire, elle est nommée en réalité dissimilarité de Bray-Curtis. Cependant, pour faciliter la compréhension, le terme distance de Bray-Curtis est conservé dans la suite de ce manuscrit.

5.2.1.2 Distance Euclidienne

La distance Euclidienne est une distance de référence, elle est couramment utilisée pour calculer une distance entre deux points ou objets. Elle représente une distance entre un point A et un point B de la manière la plus simple et la plus courte possible. En mathématiques, cette distance est notamment utilisée dans le théorème de Pythagore. En phylogénie, elle est utilisée pour la création d'arbres phylogénétiques : une distance Euclidienne est alors calculée entre des fréquences de k-mers sans nécessiter d'alignement préalable des séquences [125].

La distance Euclidienne correspond à la racine carré des différences au carré entre les coordonnées de deux points ou objets. Son minimum est de 0 et elle ne possède pas de maximum défini.

Dans notre contexte, la distance Euclidienne, permettant de comparer une séquence S et un ensemble de séquences H en se basant sur leur composition en mots, est définie par l'équation suivante :

$$distE(S,H) = \sqrt{\sum_{m \in M} f(m,S) - f(m,H)^2} \tag{5.1}$$

où M est l'ensemble des mots sélectionnés de taille l (dans le cas où aucune sélection n'est réalisée, M est l'ensemble des 4^l mots possibles pour une taille de mot l dans l'alphabet {A, T, C, G}), m est un mot de M, f est une fonction définie par F (pour les fréquences, définie dans l'équation suivante 5.5) ou P (pour les proportions, définie dans l'équation suivante 5.6) se rapportant au nombre de mots m dans la séquence considérée.

5.2.1.3 Distance de Bray-Curtis

La distance de Bray-Curtis [126] est une distance généralement utilisée pour analyser des données biologiques et très largement employée par les scientifiques de l'environnement comme les écologistes [127]. Elle permet d'évaluer la dissimilarité entre deux échantillons donnés, en terme d'abondance d'espèces. Cette distance est égale à 1 lorsque les deux échantillons de données comparés n'ont aucun point en commun et elle est égale à 0 lorsque les deux échantillons comparés sont identiques [128, 129].

La distance de Bray-Curtis se présente sous la forme d'un rapport où le numérateur résume le degré de dissimilarité d'espèce à espèce, tandis que le dénominateur normalise l'indice [130]. Elle est l'une des distances écologiques utilisées dans l'outil de métagénomique comparative Simka [52] où chaque ensemble de données est représenté par un spectre de k-mers.

Dans notre contexte, la distance de Bray-Curtis permettant de comparer une séquence S et un ensemble de séquences H en se basant sur leur composition en mots est définie par l'équation suivante :

$$BC(S,H) = 1 - \frac{2\sum_{m\in M}[f(m,S)+f(m,H)]}{\sum_{m\in M}[f(m,S)+f(m,H)]} \tag{5.2}$$

où M est l'ensemble des mots sélectionnés de taille l (dans le cas où aucune sélection n'est réalisée, M est l'ensemble des 4^l mots possibles pour une taille de mot l dans l'alphabet {A, T, C, G}), m est un mot de M, f est une fonction définie par F (pour les fréquences, définie dans l'équation suivante 5.5) ou P (pour les proportions, définie dans l'équation suivante 5.6) se rapportant au nombre de mots m dans la séquence considérée. En pratique, S est généralement un CDS (potentiellement GM ou non) et H est l'ensemble des CDS du génome hôte déduit.

5.2.1.4 Distance de Kullback-Leibler

La distance de Kullback-Leibler utilisée est issue de la publication de Trifonov et Rabadan [8]. Elle est basée sur une formule de divergence utilisant l'estimation des fréquences pour chacune des signatures génomiques comparées. Son minimum est de 0 et elle ne possède pas de maximum défini.

Dans notre contexte, la distance de Kullback-Leibler permet de comparer une séquence S et un ensemble de séquences H en se basant sur leur composition en mots. Une divergence est d'abord calculée entre les deux vecteurs de fréquences

à l'aide de l'équation suivante :

$$diverg(S,H) = \sum_{m \in M} f(m,S) \log\left(\frac{f(m,S)}{f(m,H)}\right) \tag{5.3}$$

où M est l'ensemble des mots sélectionnés de taille l (dans le cas où aucune sélection n'est réalisée, M est l'ensemble des 4^l mots possibles pour une taille de mot l dans l'alphabet {A, T, C, G}), m est un mot de M, f est une fonction définie par F (pour les fréquences, définie dans l'équation suivante 5.5) ou P (pour les proportions, définie dans l'équation suivante 5.6) se rapportant au nombre de mots m dans la séquence considérée.

Ensuite, la distance de Kullback-Leibler est calculée selon l'équation suivante :

$$distKL(S,H) = \frac{[L_S \times diverg(S,H) + L_H \times diverg(H,S)]}{L_S + L_H} \tag{5.4}$$

Où L_S est la somme du nombre d'occurrences (ou comptage) de mots m présents dans la séquence S.

Contrairement à la distance de Bray-Curtis ou à la distance Euclidienne, la distance de Kullback-Leibler prend en compte directement des fréquences de k-mers ce qui est plus approprié dans notre cas concernant le comptage des mots. En effet, nous cherchons à mettre en évidence des mots ayant une fréquence d'apparition différente dans la séquence insérée et dans le génome hôte.

5.2.2 Types de calcul de distances

Deux types de calcul de distances ont été développés pour différencier le vocabulaire de deux ensembles de séquences. Ces types de distances permettent de normaliser les comptages de mots avant de procéder au calcul de distance avec l'une des trois distances présentées dans les parties précédentes en section 5.2.1. Le vocabulaire, noté M, qui sera utilisé dans notre méthode peut être soit l'ensemble de tous les mots d'une taille l donnée, soit un sous-ensemble de ces mots particulièrement spécifique au génome de l'hôte. Dans le second cas, la sélection des mots est préalablement effectuée à l'aide du logiciel R'MES (Chapitre 3 partie 3.2.2). Cette sélection correspond à des mots nettement sur-représentés par rapport à un modèle de Markov d'ordre donné.

5.2.2.1 En fréquences

Le calcul de distance en fréquences convertit les comptages de mots réalisés via R'MES ou l'algorithme d'Aho-Corasick (Chapitre 3 partie 3.2.3) présents dans un CDS ou un ensemble de CDS, en fréquences d'apparition par rapport au nombre total de mots dans ce CDS ou cet ensemble de CDS. Par exemple, le motif « AGTT » est trouvé deux fois dans un CDS où la somme des comptages est de 50 pour tous les motifs présents. Ce même motif est présent treize fois dans l'ensemble des CDS où la somme des comptages est de 5000 pour tous les motifs présents. La fréquence d'apparition du motif « AGTT » est donc plus rare dans l'ensemble des CDS ($13/5000 = 0.0026$) en comparaison du CDS ($2/50 = 0.04$).

La distance en fréquences est défini par l'équation suivante :

$$f(m,S) = F(m,S) = \frac{C(m,S)}{\sum_{w \in M} C(w,S)} \tag{5.5}$$

où *C(m,S)* correspond au nombre d'occurrences (ou comptage) de mots m dans la séquence S et w représente chaque mot dans M. En d'autres termes, *F(m,S)* est la fréquence normalisée de m dans S.

5.2.2.2 En proportions

Le calcul de distance en proportions a été développé pour remédier à la différence notable dans la longueur cumulée entre un CDS et l'ensemble des CDS mais aussi dans l'ordre de grandeur du nombre d'occurrences. En effet, le nombre de comptage est moins important dans un CDS comparé à un ensemble de CDS. Le calcul de distance en proportions permet donc d'équilibrer les comptages entre un CDS et l'ensemble des CDS avant d'appliquer le calcul de distance. Par exemple, le motif « AGTT » est trouvé une seule fois dans un CDS. Ce même motif est présent quinze fois dans l'ensemble des CDS. Étant donné que le nombre de nucléotides présents dans un CDS et celui dans l'ensemble des CDS est très différent, le calcul en proportion a pour but de normaliser cette différence pour les comptages de motifs.

La distance en proportions est défini par l'équation suivante :

$$f(m,S) = P_H(m,S) = C(m,S) \times \frac{\sum_{w \in M} C(w,H)}{\sum_{w \in M} C(w,S)} \tag{5.6}$$

En d'autres termes, $P_H(m,S)$ est une normalisation du comptage de m dans S équivalant à considérer m dans une séquence aussi grande que la longueur cumulée de H. Avec cette définition, lorsque $S = H$, on obtient $P_H(m,S) = C(m,H)$.

5.2.3 Visualisation des résultats

Pour visualiser des différences de distances pour différentes tailles de mots, la visualisation sous forme de boîtes à moustaches a été retenue. Pour chaque distance (Bray-Curtis ou Kullback-Leibler), une boîte à moustaches par taille de mots est générée. Cette boîte représente la distribution des distances entre l'ensemble des CDS du génome hôte et chacun de ces CDS ou chacun des CDS de la banque de données d'OGM connus. Les CDS inserts OGM sont représentés par des étoiles noires. Le corps de la boîte à moustache correspond à l'intervalle entre le premier (25% des données) et le troisième quartile (75% des données). L'objectif est que le corps de la boîte à moustache des CDS de la banque de données d'OGM connus ne chevauche pas celui des CDS appartenant au génome hôte. En effet, les distances entre l'ensemble des CDS du génome hôte et les CDS de la banque de données devraient être plus élevées par rapport aux distances sur les CDS du génome hôte. Les CDS de la banque de données d'OGM connus présentent un vocabulaire différent de celui du génome hôte.

5.3 Application

La stratégie utilisant les calculs de distances qui a été choisie résulte de tests d'hypothèses et de paramètres qui vont être détaillés dans les parties suivantes. Plusieurs combinaisons de taille de mots en fréquences et en proportions ont été testées pour prendre en compte les différentes propriétés du génome.

5.3.1 Démarches exploratoires

5.3.1.1 Test 1 : Évaluation de la cohérence des distances

Notre hypothèse est que la distance entre deux organismes totalement différents doit être élevée. En opposition, la distance entre deux organismes identiques doit être égale à 0. Pour vérifier cette hypothèse et ainsi évaluer la distance maximale entre deux espèces éloignées phylogénétiquement, le génome de référence du maïs a été comparé à la bactérie *Streptomyces avermitilis* (Numéro d'accession NCBI : NC_003155.5). Cette bactérie a été choisie car elle provient du sol, elle est donc un contaminent naturel de l'environnement. De plus, elle possède une taille importante de génome pour une bactérie (9,0 Mb) et un pourcentage moyen en GC de 70.7% alors que celui du maïs est de 46.5%. Cette différence de nucléotides caractérisée par le pourcentage en GC devrait se traduire par une distinction de vocabulaire notable entre le génome de la bactérie et celui du maïs. Ces résultats sont présentés dans le Tableau 4.

Ensembles de CDS comparés	**Distance Euclid**	**Distance KL**	**Distance BC**
GenomeRef_GenomeRef	0	0	0
GenomeRef_S.avermitilis	0.0240	1.6902	0.5250

Tableau 4 – Calcul des distances Euclidienne, de Kullback-Leibler et de Bray-Curtis en fréquences entre différents ensembles de CDS avec des mots de taille 6 et un modèle de Markov d'ordre 4 (maximal). « GenomeRef » représente l'ensemble des CDS du génome de référence du maïs et « S.avermitilis » représente l'ensemble des CDS du génome de *S. avermitilis.*

Les distances calculées entre deux ensembles de CDS identiques sont toutes égales à 0. Les distances obtenues entre l'ensemble des CDS du génome de référence du maïs et l'ensemble des CDS de la bactérie sont moins élevées qu'attendu, bien que les deux ensembles de CDS aient des vocabulaires différents du fait de leur éloignement phylogénétique. Les distances Euclidienne et de Kullback-Leibler ne possèdent pas une valeur maximale définie. A contrario, la distance de Bray-Curtis obtient une valeur maximale de 1 lorsqu'aucune similarité entre les deux échantillons n'est observée (Chapitre 5 partie 5.2.1.3). En se basant sur cette propriété de la distance de Bray-Curtis, le génome de la bactérie *S. avermitilis* et du maïs ne seraient pas aussi éloignés qu'attendu. Nous nous attendions a obtenir une valeur de la distance de Bray-Curtis plus importante, de l'ordre de 0.8, étant donné que de multiples différences existent entre le maïs et la bactérie *S. avermitilis*. Cependant, *S. avermitilis* est une bactérie provenant du sol,

donc soumise aux mêmes contraintes et environnements que le maïs. Ces points communs pourraient expliquer les relativement faibles distances obtenues.

Les distances Euclidienne, de Kullback-Leibler et de Bray-Curtis entre le génome de la bactérie *S. avermitilis* et celui du maïs de référence sont peu élevées.

5.3.1.2 Test 2 : Évaluation de l'homogénéité des distances au sein d'un organisme

L'hypothèse émise est que le vocabulaire des CDS d'un génome est homogène. Pour vérifier cette hypothèse et ainsi évaluer l'homogénéité des distances entre deux génomes, les CDS du génome de référence du maïs et les CDS du génome de *S. avermitilis* sont répartis en 10 ensembles de même taille et comparés entre eux. Si cette hypothèse est invalide, les CDS devront être séparés en fonction de leurs classes protéiques, le vocabulaire de chaque classe serait ainsi respecté et pourrait alors être comparé. En effet, les propriétés d'homogénéité attendues seraient ainsi retrouvées. Les résultats sont donnés dans le Tableau 5.

Dans le Tableau 5, l'intervalle des valeurs de distance est de 0.0028, 0.2634 et 0.0550 pour les distances Euclidienne, de Kullback-Leibler et de Bray-Curtis respectivement. Par rapport à la moyenne, les distances Euclidienne, de Kullback-Leibler et de Bray-Curtis ont un écart maximal de 8.88%, 12.75% et 7.37% respectivement, ce qui montre leur homogénéité. Ainsi, les distances obtenues en comparant les parties de génome sont homogènes pour les trois distances testées. L'homogénéité des distances est donc respectée lorsqu'un génome est partitionné.

Ensembles de CDS comparés	**Distance Euclid**	**Distance KL**	**Distance BC**
GenomeRef1_StreptoAver1	0.0218	1.3337	0.4900
GenomeRef2_StreptoAver2	0.0231	1.4441	0.5177
GenomeRef3_StreptoAver3	0.0246	1.5464	0.5381
GenomeRef4_StreptoAver4	0.0242	1.5269	0.5360
GenomeRef5_StreptoAver5	0.0243	1.5857	0.5353
GenomeRef6_StreptoAver6	0.0240	1.5972	0.5312
GenomeRef7_StreptoAver7	0.0240	1.5290	0.5289
GenomeRef8_StreptoAver8	0.0238	1.5619	0.5292
GenomeRef9_StreptoAver9	0.0248	1.5914	0.5450
GenomeRef10_StreptoAver10	0.0246	1.5693	0.5390
Moyenne	0.0239	1.5286	0.5290

Tableau 5 – Calcul de distances Euclidienne, de Kullback-Leibler et de Bray-Curtis en fréquences entre différents ensembles de CDS avec des mots de taille 6 et un modèle de Markov d'ordre 4 (maximal). « GenomeRef1-10 » représente l'ensemble des CDS de chacune des 10 parties du génome de référence du maïs et « StreptoAver1-10 » représente l'ensemble des CDS de chacune des 10 parties du génome de *S. avermitilis*.

5.3.2 Cas du maïs

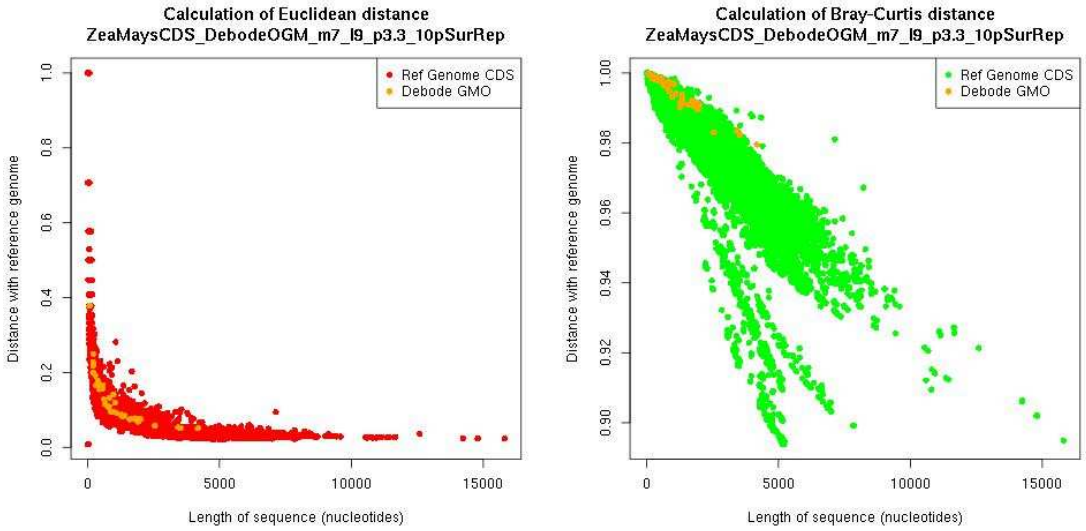
Différents paramètres du logiciel R'MES, qui permet de rechercher des motifs exceptionnels dans une séquence, ou des filtres exploitant les propriétés d'un génome sont été testés. Le but de ces différents tests est de sélectionner les variables et paramètres optimaux permettant de distinguer des CDS inserts OGM. Pour mener à bien ces tests, trois jeux de données ont été utilisés : les CDS nettoyés du génome de maïs contenant les séquences d'inserts OGM, 45 CDS d'OGM connus fournis par le Centre wallon de Recherches agronomiques (CRA-W) notés ensemble FD par la suite et les CDS du génome de référence du maïs (*Zea Mays B73*). Ces trois jeux de données ont été obtenus en utilisant la stratégie mise au point sur le génome de maïs OGM détaillée dans l'Annexe C. Le pipeline de nettoyage (décrit dans le Chapitre 4 partie 4.2.1) a été modifié en raison de la taille du génome du maïs (Chapitre 4 partie 4.3.1). La comparaison des calculs de distance est réalisée avec les CDS du génome de référence du maïs et pas les CDS apparentés au génome hôte. Cette stratégie résulte des premières investigations menées sur la caractérisation des CDS inserts OGM du maïs.

Les paramètres R'MES testés sont la longueur d'un mot, l'ordre du modèle de Markov, l'utilisation de la phase 3 expliquée ultérieurement (Chapitre 5 partie 5.3.2.5), l'utilisation de familles de mots, la concaténation des troisièmes positions des codons et le pourcentage de mots sur ou sous-représentés considéré. Ces différents paramètres sont appliqués aux trois distances sélectionnées, Euclidienne, de Bray-Curtis (BC) et de Kullback-Leibler (KL) selon deux types de normalisations des comptages de mots, en fréquences et en proportions.

Dans les parties suivantes, pour chaque distance calculée, la notation de deux paramètres R'MES cités précédemment est abrégée comme suit : L{taille de mots}M{ordre de Markov}. Par exemple pour un calcul de distance de Bray-Curtis avec une taille de mot de 9 et un modèle de Markov d'ordre 7, on nomme Distance de Bray-Curtis L9M7.

5.3.2.1 Discrimination

Avec la distance Euclidienne, plusieurs résultats provenant des tests de différents paramètres comme l'influence de la taille de mots, de la phase 3 ou de l'ordre du modèle de Markov ont été obtenus. En effet, tous ces paramètres ont été testés avec des mots de tailles 3 à 7 pour des modèles de Markov d'ordre 1 à l-2 (l étant la taille d'un mot). Cependant, pour des raisons de lisibilité, uniquement les résultats permettant d'illustrer les choix retenus sont présentés dans ce manuscrit.



(a) Distance Euclidienne

(b) Distance de Bray-Curtis

FIGURE 20 – Comparaison des distances Euclidienne et de Bray-Curtis en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 en phase 3 et un modèle de Markov d'ordre 7 (maximal pour cette taille de mot). Les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (rouge ou vert) ou chacun des CDS OGM de l'ensemble FD (orange) par rapport à la longueur des CDS.

Comme illustré sur le graphique ci-dessus (Figure 20), l'utilisation de la distance Euclidienne ne permet pas de discriminer les CDS OGM par rapport aux CDS du maïs de référence. En effet, dans la Figure 20a, les points oranges, représentant les distances Euclidienne entre chacun des CDS inserts OGM de l'ensemble FD et l'ensemble des CDS du génome de référence du maïs, sont mélangés aux points rouges, représentant les distances Euclidienne entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS. A contrario, dans la Figure 20b, la distance de Bray-Curtis regroupe les CDS OGM connus (points oranges), ceux-ci ayant des distances élevées.

Distance Euclidienne	**Mots de taille 6**	**Mots de taille 9**	**Mots de taille 12**
StreptoAver1_StreptoAver2	0.0483	0.0073	0.0021
StreptoAver3_StreptoAver4	0.0489	0.0073	0.0021
StreptoAver5_StreptoAver6	0.0552	0.0074	0.0021
StreptoAver7_StreptoAver8	0.0437	0.0074	0.0019
StreptoAver9_StreptoAver10	0.0494	0.0078	0.0021

Distance de Bray-Curtis	**Mots de taille 6**	**Mots de taille 9**	**Mots de taille 12**
StreptoAver1_StreptoAver2	0.2929	0.3205	0.3374
StreptoAver3_StreptoAver4	0.3084	0.3151	0.3702
StreptoAver5_StreptoAver6	0.3209	0.3226	0.3399
StreptoAver7_StreptoAver8	0.2858	0.3220	0.2939
StreptoAver9_StreptoAver10	0.3110	0.3284	0.3357

Distance de Kullback-Leibler	**Mots de taille 6**	**Mots de taille 9**	**Mots de taille 12**
StreptoAver1_StreptoAver2	0.4205	0.5092	Infini
StreptoAver3_StreptoAver4	0.5291	0.5143	Infini
StreptoAver5_StreptoAver6	0.5687	0.5235	Infini
StreptoAver7_StreptoAver8	0.4458	0.5290	Infini
StreptoAver9_StreptoAver10	0.4909	0.5464	Infini

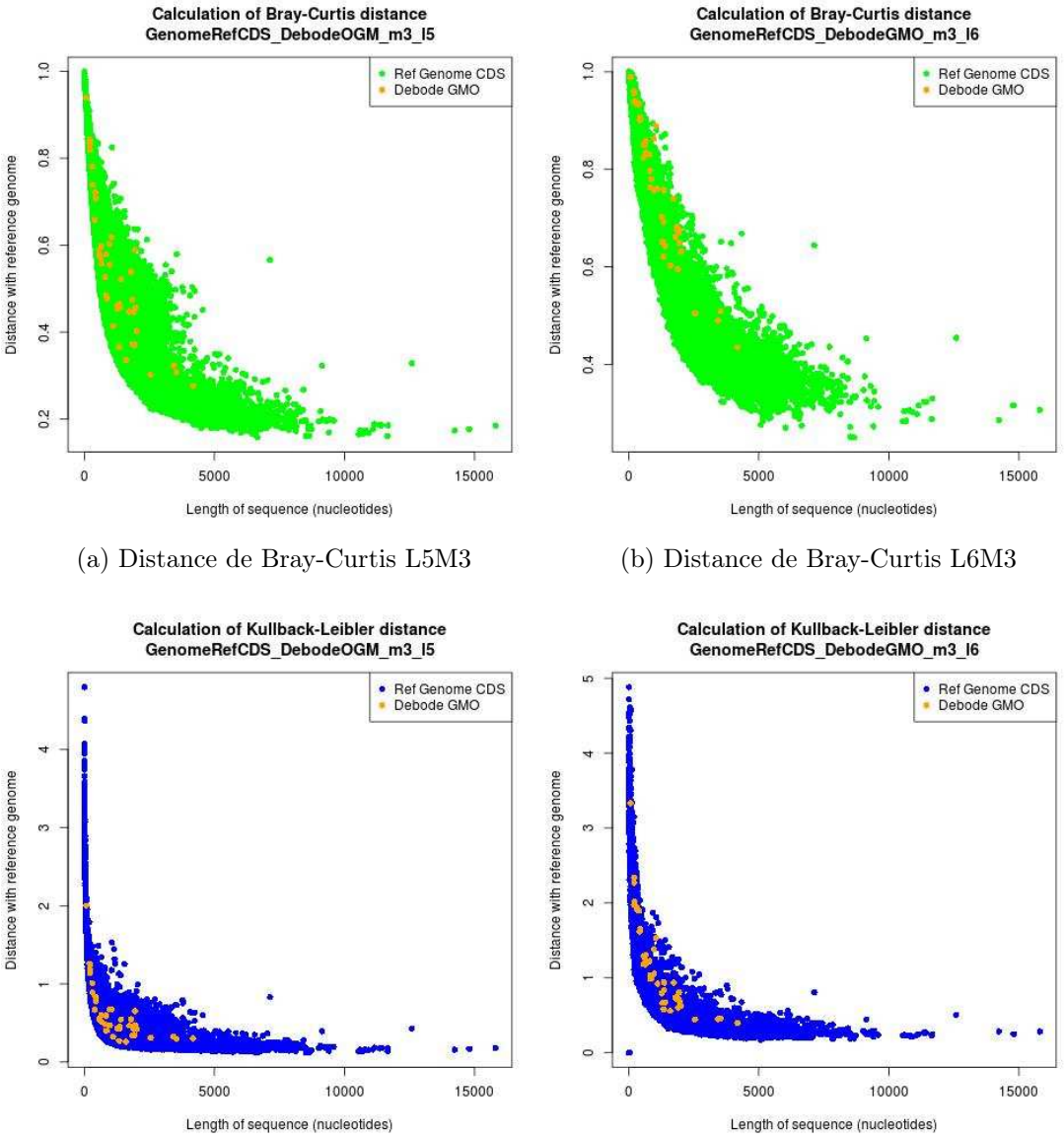
Tableau 6 – Calcul des distances Euclidienne, de Bray-Curtis et de Kullback-Leibler en fréquences pour trois tailles de mots (6, 9 et 12) et un modèle de Markov d'ordre maximal entre plusieurs ensembles de CDS distincts de la bactérie *S. avermitilis* (« StreptoAver1-10 »).

D'autres analyses, non fournies dans ce manuscrit, présentent des résultats divergents entre la distance Euclidienne et les deux autres distances. Notre hypothèse est que la distance calculée entre deux génomes ne doit pas ou peu varier en fonction de la taille d'un mot. En effet, différentes tailles de mots doivent pouvoir être comparables. Les distances Euclidiennes calculées entre deux ensembles de CDS provenant de parties différentes de la bactérie *S. avermitilis*, présentées dans le Tableau 6, montrent que la distance minimale observée diminue en fonction de la taille des mots considérés. Par exemple, avec des mots de taille 6, la distance Euclidienne entre les ensembles de CDS StreptoAver_1 et StreptoAver_2 est de 0.0483 et avec des mots de taille 12, cette distance est de 0.0021. Plus la taille d'un mot est grande, plus la distance Euclidienne diminue contrairement aux distances de Kullback-Leibler et de Bray-Curtis où la distance ne varie pas en fonction de la taille d'un mot. Pour les mots de taille 12 avec la distance de Kullback-Leibler, une valeur infinie est obtenue. La petite taille des portions du génome de *S. avermitilis* entraîne un nombre de comptages de mots de 12 lettres proche de zéro,

expliquant ainsi l'impossibilité de calculer la distance de Kullback-Leibler.

Au vu des résultats précédents et après avoir testé différents paramètres, aucune séparation nette des CDS OGM par rapport à l'ensemble des CDS du génome de référence du maïs n'a été observée, la distance Euclidienne a donc été écartée.

5.3.2.2 Influence de la taille des mots considérés



(a) Distance de Bray-Curtis L5M3

(b) Distance de Bray-Curtis L6M3

(c) Distance de Kullback-Leibler L5M3

(d) Distance de Kullback-Leibler L6M3

FIGURE 21 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en considérant des mots de taille 5 (à gauche) ou 6 (à droite) avec un modèle de Markov d'ordre 3. Les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) par rapport à la longueur des séquences.

Pour déterminer l'influence de la taille des mots considérés, plusieurs tailles de mots sont testées avec les distances de Bray-Curtis et de Kullback-Leibler entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS ou des CDS de l'ensemble FD. Avec des mots de plus de 9 lettres, le logiciel R'MES renvoie une erreur due à l'allocation mémoire insuffisante de l'ordinateur utilisé. De plus, les calculs de distance sur des mots de 10 lettres demanderaient un temps de calcul de plusieurs semaines à cause du nombre de mots considérés (4^{10} soit 1048576 mots) et du nombre important de CDS du génome de référence du maïs. Le génome de référence du maïs utilisé (*Zea Mays* B73) est décrit dans le Chapitre 2 partie 2.3.1.1 et est constitué d'environ 133 000 CDS.

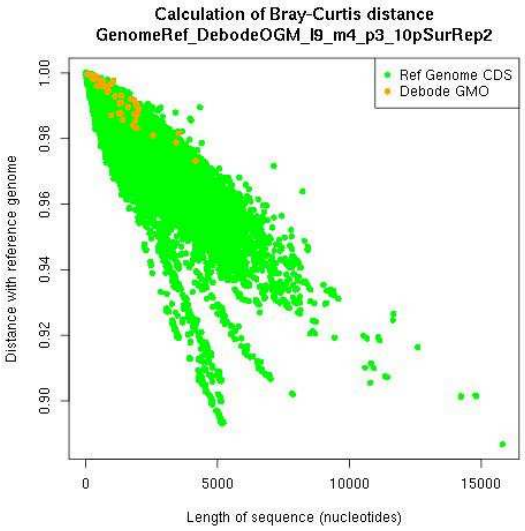
La distance minimale des CDS inserts OGM est utilisée pour comparer les différentes tailles de mots testées, de 3 à 7, avec un modèle de Markov d'ordre maximal l-2 (l étant la taille du mot). La Figure21 présente uniquement les meilleurs résultats obtenus en fonction de deux tailles de mots distincts pour illustrer le choix des paramètres retenus.

Pour les distances de Bray-Curtis et de Kullback-Leibler, la proportion des CDS ayant une distance élevée est plus importante avec une taille de mot élevée (Figure 21). La distance minimale des CDS inserts OGM est ainsi plus élevée avec des mots de taille 6 qu'avec des mots de taille 5. Cependant, les distances des CDS du génome de référence sont généralement plus élevées également, mais dans une moindre mesure. La distance minimale pour les CDS inserts OGM varie en fonction de la taille des mots.

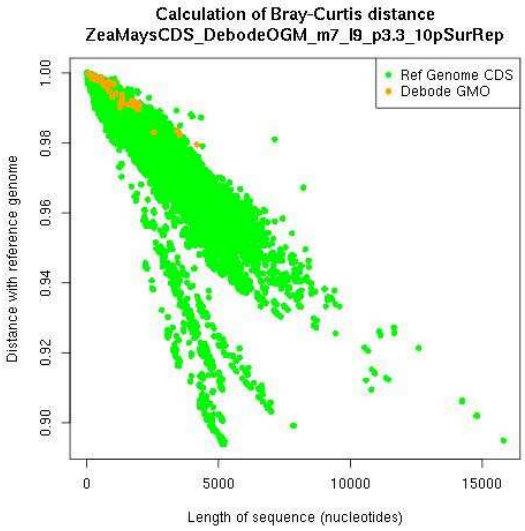
5.3.2.3 Influence de l'ordre du modèle de Markov

Les distances de Bray-Curtis et Kullback-Leibler ont été calculées entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS ou des CDS de l'ensemble FD pour déterminer l'influence de l'ordre du modèle de Markov (Figure 22). Comme décrit dans le Chapitre 3 partie 3.2.2, l'ordre de Markov est un paramètre R'MES lié au nombre d'états présents dans un modèle de Markov.

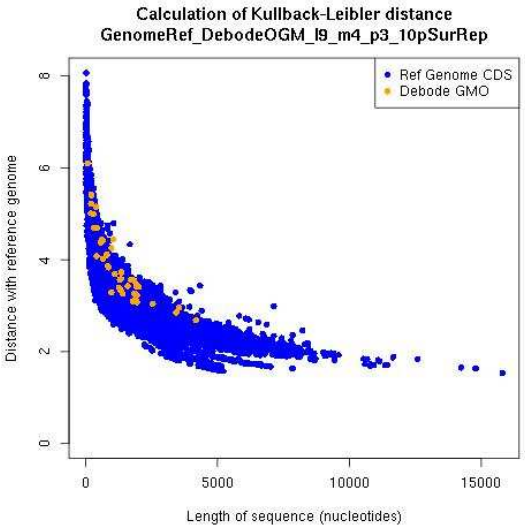
Pour les distances de Kullback-Leibler et Bray-Curtis, les valeurs de distances des CDS augmentent légèrement avec un ordre du modèle de Markov plus élevé. Les CDS inserts OGM connus sont légèrement plus groupés avec une distance plus élevée lorsque l'ordre de Markov est maximal par rapport à la taille de mots donnée. La valeur de distance minimale pour les CDS inserts OGM connus avec le modèle L9M4 est de 0,973 alors que celle pour le modèle L9M7 est de 0,980 avec la distance de Bray-Curtis. Avec la distance de Kullback-Leibler, la valeur minimale de distance pour les CDS inserts OGM connus avec le modèle L9M4 est de 2,721 alors que celle pour le modèle L9M7 est de 2,945. Un ordre du modèle de Markov maximal améliore la distance minimale des CDS inserts OGM connus. Les analyses suivantes seront réalisées avec un ordre de Markov maximal pour une taille de mots donnée.



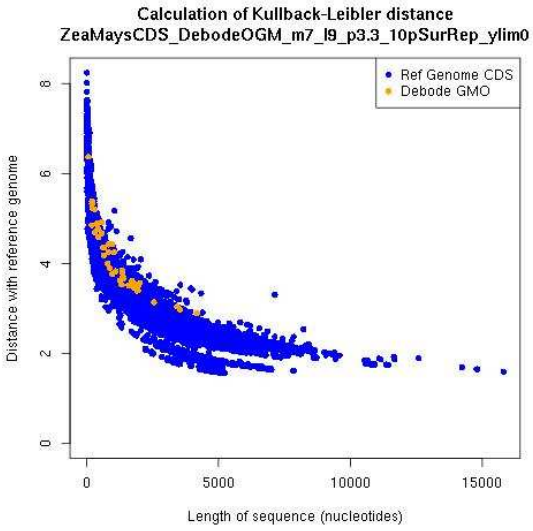
(a) Distance de Bray-Curtis L9M4



(b) Distance de Bray-Curtis L9M7



(c) Distance de Kullback-Leibler L9M4



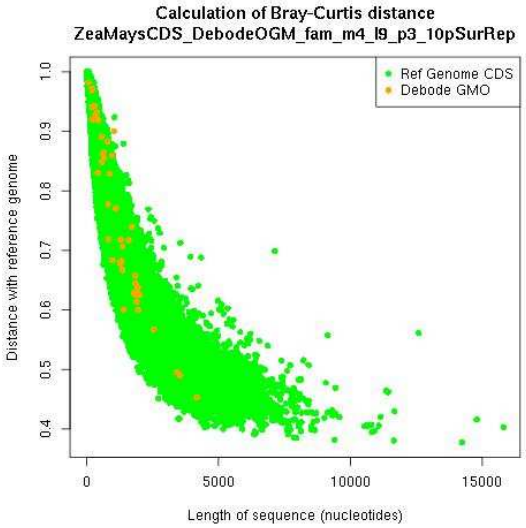
(d) Distance de Kullback-Leibler L9M7

FIGURE 22 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 en phase 3 avec un modèle de Markov d'ordre 4 (à gauche) ou d'ordre 7 (à droite) maximal pour cette taille de mot. Les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) par rapport à la longueur des séquences.

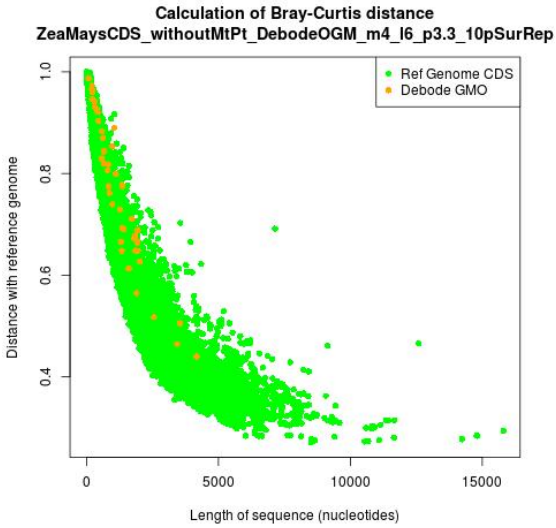
5.3.2.4 Influence des familles de mots

Le logiciel R'MES dispose d'un paramètre permettant de rechercher des familles de mots exceptionnelles ou de motifs dégénérées en spécifiant la taille des mots ainsi que le masque choisi composé de n et de #. n correspond à chacune

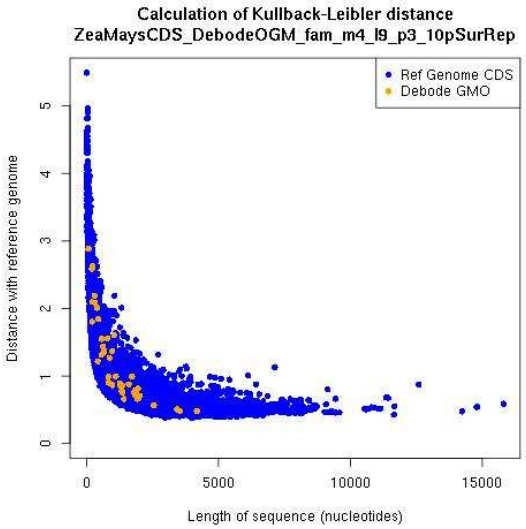
des lettres de l'alphabet A, T, C, G. Par exemple, le masque #n## produira 64 tétra-nucléotides dégénérés ayant un n en deuxième position, où # est une lettre de l'alphabet A, T, C, G.



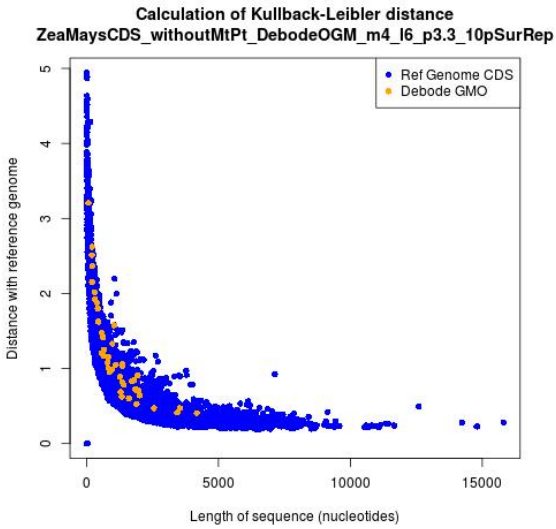
(a) Distance de Bray-Curtis L9M4



(b) Distance de Bray-Curtis L6M4



(c) Distance de Kullback-Leibler L9M4



(d) Distance de Kullback-Leibler L6M4

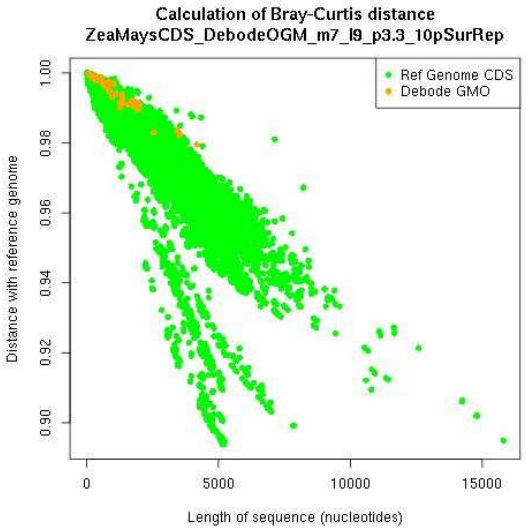
FIGURE 23 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés de taille 6 en phase 3 avec un modèle de Markov d'ordre 4 avec des familles de mots ayant pour masque n##n##n## (à gauche) ou des mots non groupés (à droite). Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) en fonction de la longueur des séquences (abscisses).

Différentes tailles de familles de mots ont été testées, avec des modèles tels que n##n## ou ##n##n##n. Les meilleurs résultats obtenus justifiant le choix de conserver ou non ce paramètre (familles de mots) sont présentés sur la Figure23. Pour les distances de Bray-Curtis et de Kullback-Leibler, les résultats obtenus par CDS entre l'utilisation de mots simples et de familles de mots sont minimes. Pour la distance de Bray-Curtis, la distance minimale des CDS OGM de l'ensemble FD est de 0,447 avec les familles de mots (L9M4) contre 0,438 sans les familles de mots (L6M4). Pour la distance de Kullback-Leibler, la distance minimale des CDS OGM de l'ensemble FD est de 0,440 avec les familles de mots (L9M4) contre 0,437 sans les familles de mots (L6M4). L'utilisation de familles de mots dans R'MES n'améliore pas les résultats obtenus. Cette option a donc été écartée des futurs calculs.

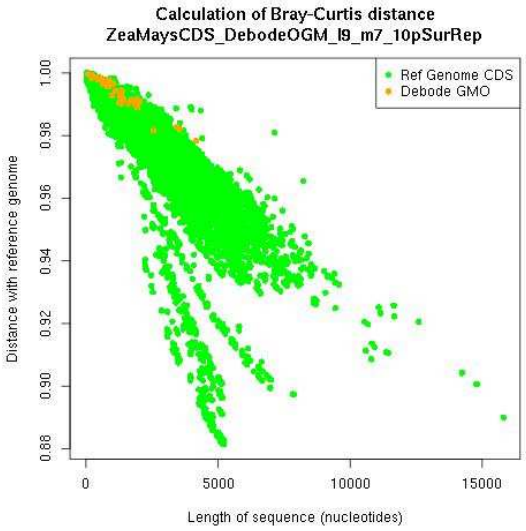
5.3.2.5 Influence du paramètre phase 3

Le paramètre phase, disponible dans le logiciel R'MES, permet de tenir compte de l'importance contextuelle périodique de la séquence. Ce paramètre est principalement utilisé lors de l'analyse de CDS pour ne considérer que les codons. En effet, fixer le paramètre phase à 3 permet de ne considérer qu'une position trinucléotidique sur 3, la troisième position du codon étant la plus spécifique de l'usage du codon (Chapitre 3 partie 3.2.1).

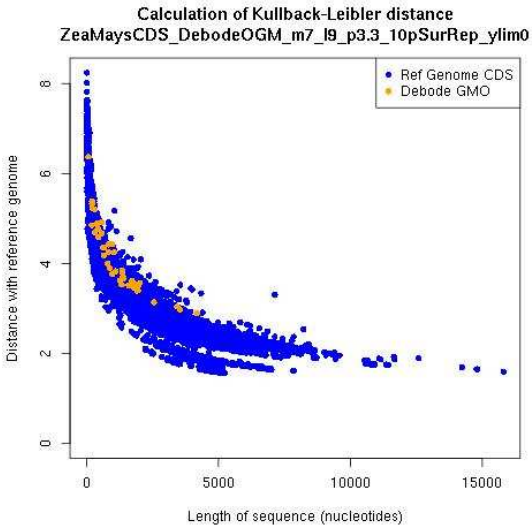
Le paramètre phase 3 a été testé avec des tailles de mots de 3 à 9 avec un ordre de Markov maximal soit *l-2* (l étant la taille d'un mot). Par exemple, avec une taille de mot de 7, l'ordre de Markov maximal est 5 (l=7 soit 7-2=5). Les meilleurs résultats permettant de conserver ou non le paramètre phase 3 sont présentés sur la Figure24.



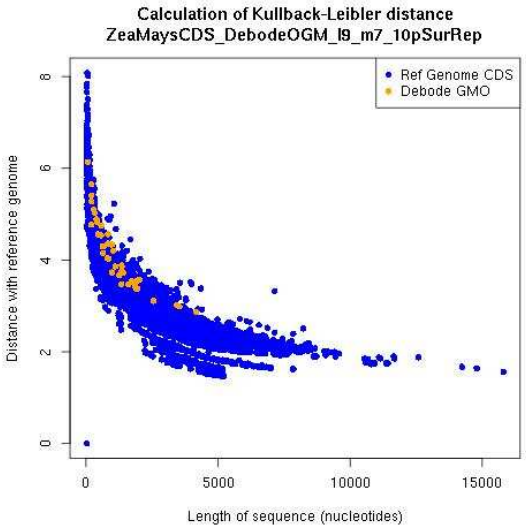
(a) Distance de Bray-Curtis L9M7 p3



(b) Distance de Bray-Curtis L9M7



(c) Distance de Kullback-Leibler L9M7 p3



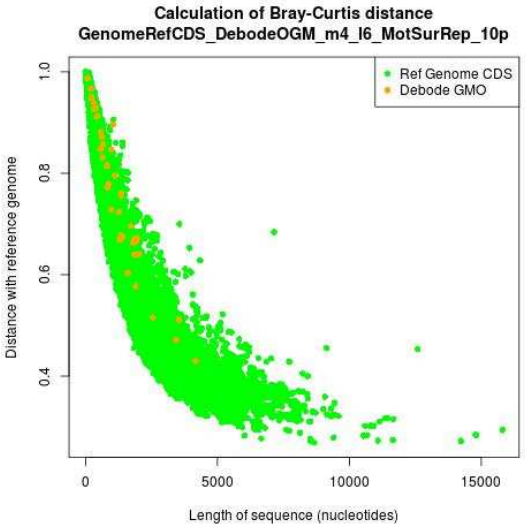
(d) Distance de Kullback-Leibler L9M7

FIGURE 24 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 avec un modèle de Markov d'ordre 7 en phase 3 (à gauche) ou sans phase (à droite). Les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) par rapport à la longueur des séquences.

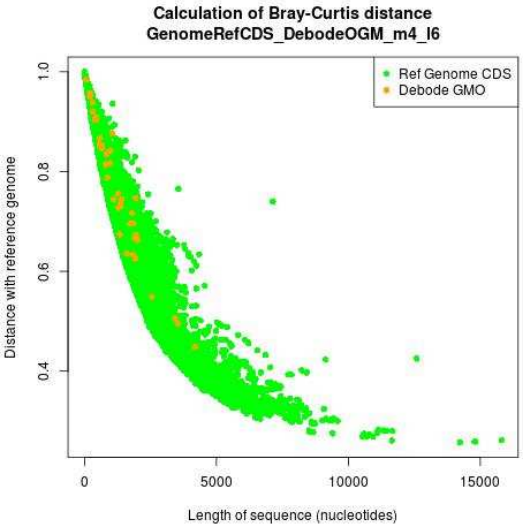
A la fois pour les distances de Bray-Curtis et de Kullback-Leibler, très peu de différences sont observables avec ou sans l'utilisation du paramètre phase 3 sur les graphiques précédents. Néanmoins, le calcul en phase 3 présente des avantages, il tient compte de l'usage du codon et ne nécessite pas un temps de calcul supplémentaire. L'usage du codon est un critère important pour caractériser les

CDS. Le paramètre phase 3 est donc conservé pour la suite des analyses.

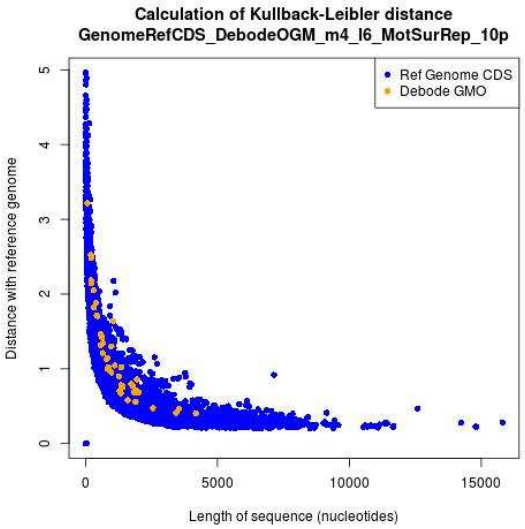
5.3.2.6 Influence du pourcentage de mots sur et/ou sous représentés



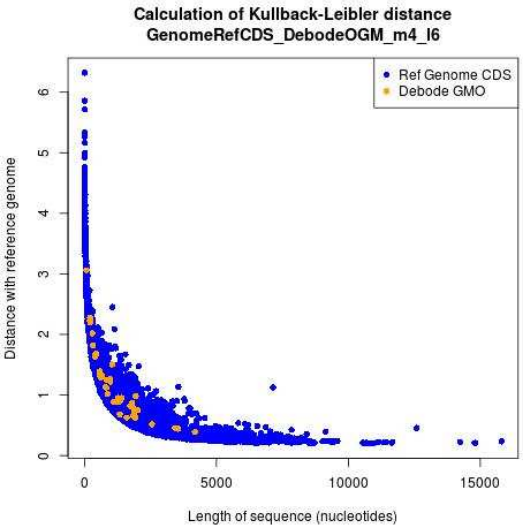
(a) Distance de Bray-Curtis L6M4 10pSurRep



(b) Distance de Bray-Curtis L6M4



(c) Distance de Kullback-Leibler L6M4 10pSurRep



(d) Distance de Kullback-Leibler L6M4

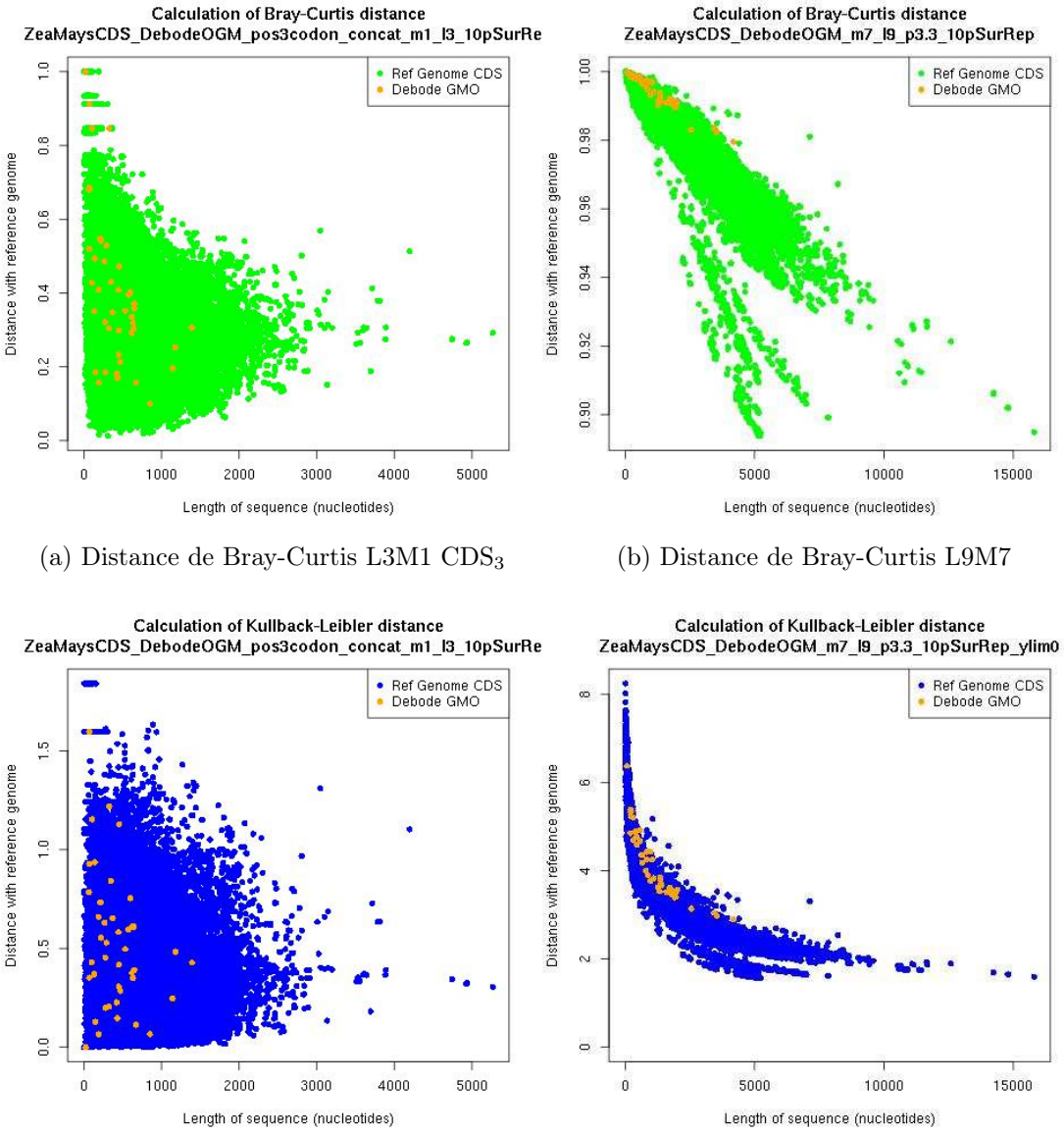
FIGURE 25 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés (à gauche) ou en considérant tous les mots (à droite) de taille 6 avec un modèle de Markov d'ordre 4. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) en fonction de la longueur des séquences (abscisses).

Ne considérer que les mots sur ou sous représentés dans un génome permet d'accentuer la spécificité du vocabulaire du génome de référence et donc mieux discriminer des CDS inserts OGM. Plusieurs combinaisons ont été testées : 1% ou 10% de mots sur-représentés et un mélange composé de 5% de mots sur-représentés et de 5% de mots sous-représentés. La Figure25 présente les meilleurs résultats obtenus qui ont permis de déduire si le pourcentage de mots sur ou sous représentés choisis a une incidence sur les calculs de distance.

Les résultats obtenus avec 1% de mots sur-représentés sont trop sélectifs, un nombre important de CDS obtient une distance à 1 avec Bray-Curtis ou très élevés avec Kullback-Leibler, du fait de l'absence d'occurrences de mots. Les distances obtenues avec un mélange de mots sur et sous représentés sont légèrement moins élevées que celles obtenues avec 10% de mots sur-représentés. Les distances obtenues avec l'utilisation de 10% de mots sur-représentés où la totalité des mots sont très similaires. Comme observé sur la Figure25, avec la distance de Bray-Curtis, la distance minimale des CDS OGM de l'ensemble FD est de 0,436 en ne considérant que 10% de mots sur-représentés contre 0,448 lorsque tous les mots sont pris en compte. Avec la distance de Kullback-Leibler, la distance minimale des CDS OGM de l'ensemble FD est de 0,443 en ne considérant que 10% de mots sur-représentés contre 0,439 avec tous les mots. Le temps de calcul est moins long en utilisant 10% de mots sur-représentés. L'utilisation de 10% de mots sur-représentés a donc été conservée pour les analyses suivantes surtout pour les calculs gourmands réalisés sur des mots de taille 9.

5.3.2.7 Influence de la concaténation des troisièmes positions des codons

Les troisièmes positions des codons sont les plus représentatives de l'utilisation des codons par un organisme (Chapitre 3 partie 3.2.1). Nous voulions réaliser des analyses R'MES avec des mots de très grande taille (plus de 21 nucléotides) en phase 3 (paramètre R'MES permettant de tenir compte de la périodicité). L'utilisation du logiciel R'MES sur des mots d'une telle taille est impossible (la taille maximale des mots acceptés est de 14). Les résultats statistiques du logiciel R'MES sont basés uniquement sur les troisièmes positions des codons. Les mêmes résultats que sur des séquences entières, peuvent ainsi être obtenus en analysant des mots de 7, 8 ou 9 lettres dans des CDS où seuls les nucléotides à la troisième position des codons, désignés par CDS_3, ont été retenus. L'ensemble des CDS_3 permet de considérer les mots de taille n contenant les informations utiles des mots de taille $3n$ dans l'ensemble des CDS. L'utilisation du logiciel R'MES sur les CDS_3 sans l'option « -phase » permet donc d'obtenir des résultats équivalents à ceux de l'analyse des mots de 21, 24 ou 27 nucléotides sur des CDS avec l'option « -phase ». La concaténation des troisièmes positions des codons a été réalisée avec des tailles de mots de 3 à 9 lettres correspondant à des séquences de 9 à 27 nucléotides avec un modèle de Markov maximal. Les meilleurs résultats obtenus pour les données du génome de maïs OGM, permettant de caractériser l'influence de la concaténation des troisièmes positions des codons par rapport à l'utilisation de la séquence entière sont présentés sur la Figure 26 et la Figure 27.



(a) Distance de Bray-Curtis L3M1 CDS_3

(b) Distance de Bray-Curtis L9M7

(c) Distance de Kullback-Leibler L3M1 CDS_3

(d) Distance de Kullback-Leibler L9M7

FIGURE 26 – Comparaison des distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 avec un modèle de Markov d'ordre 7 avec les CDS_3 (à gauche) ou la séquence entière (à droite). Une séquence entière de 9 lettres correspond à une séquence de 3 lettres avec un modèle de Markov d'ordre 1 en conservant uniquement les troisièmes positions des codons, les CDS_3. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) en fonction de la longueur des séquences (abscisses).

Pour les distances de Bray-Curtis et Kullback-Leibler, aucune amélioration n'est observable avec les distances des CDS_3 par rapport aux séquences entières (Figure 26). La totalité des distances des CDS OGM connus de l'ensemble FD

ont des valeurs moins élevées avec les CDS_3 qu'avec les séquences non concaténées (CDS). Avec la distance de Bray-Curtis, les distances entre chacun des CDS OGM de l'ensemble FD et l'ensemble des CDS du génome de référence en considérant la totalité des séquences sont incluses dans l'intervalle [0,98 ; 1,00]. Pour les mêmes distances mais avec la concaténation des troisièmes positions des codons (CDS_3), l'intervalle est de [0,10 ;1,00]. Néanmoins avec les CDS_3, il est possible de considérer un plus grand nombre de lettres donc une taille de séquence plus grande.

Les CDS possédant une longueur inférieure à 27 nucléotides ont été écartés de l'analyse à cause de la taille de mots nécessaire pour le modèle considéré. En effet, une séquence entière de 27 nucléotides, par exemple :

AG**A**CT**G**AC**T**GA**G**CC**A**TG**A**CT**C**AG**T**TG**C**

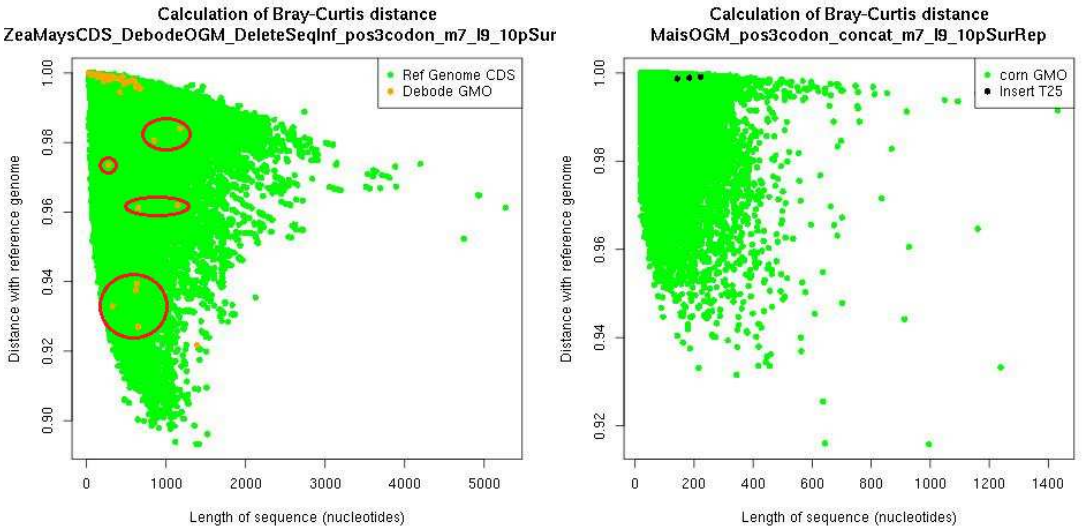
correspond à une séquence de 9 lettres en ne considérant que les troisièmes positions des codons, les CDS_3, soit avec l'exemple :

AGTGAACTC

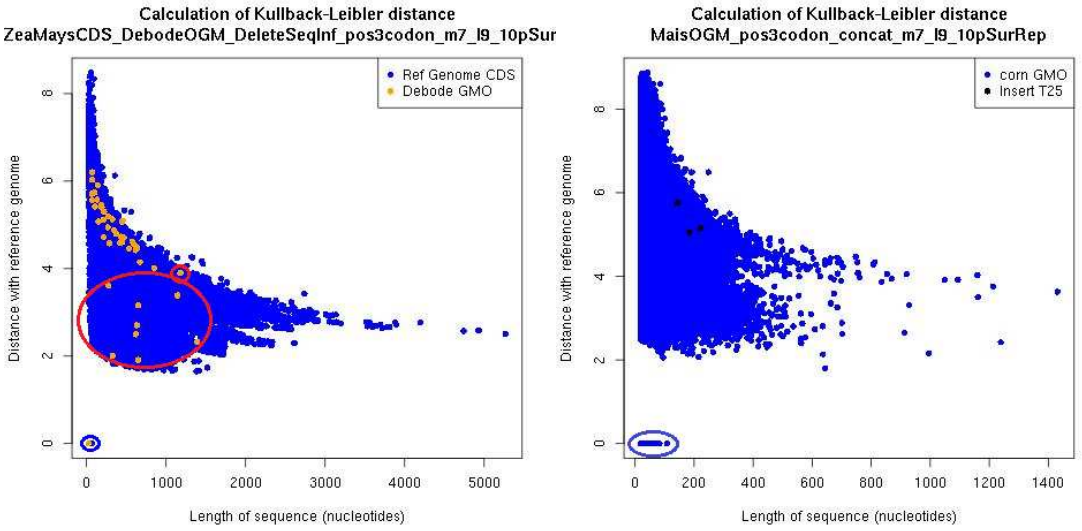
Donc un CDS possédant une longueur inférieure à 27 nucléotides ne pourra pas être utilisé, sa longueur étant inférieure à la taille de mot considérée.

Concernant les distances de Bray-Curtis et Kullback-Leibler entre les CDS OGM connus de l'ensemble FD et l'ensemble des CDS du génome de référence (Figure 27, a et c), les distances des CDS_3 OGM connus sont élevées et groupées pour la majorité d'entre elles (Figure 27, points oranges). Certaines distances des CDS_3 OGM connus de l'ensemble FD obtiennent des valeurs moins élevées avec les CDS_3 (Figure 27, cercles rouges). La plupart de ces CDS_3 correspondent à des inserts OGM de riz, espèce très proche phylogénétiquement du maïs. Cette particularité peut expliquer les faibles distances obtenues pour ces séquences par rapport à un insert de bactérie où la distance devrait être plus élevée. Les autres inserts avec une distance faible sont des inserts de maïs, ils seront par la suite éliminés car un insert provenant de la même espèce que le génome hôte ne peut pas être qualifié d'OGM.

Pour la distance de Kullback-Leibler sur les CDS_3, des distances à 0 sont observées (Figure 27, cercles bleus). Elles correspondent aux CDS_3 où aucun mot n'a été trouvé. La distance de Kullback-Leibler ne possède pas de valeur maximale contrairement à la distance de Bray-Curtis. Si aucun mot n'est trouvé dans un CDS cela signifie qu'il possède un vocabulaire éloigné du génome de référence (la sélection des mots sur-représentés est réalisée sur l'ensemble des CDS du génome de référence). Dans les analyses suivantes pour la distance de Kullback-Leibler, une distance maximale sera automatiquement allouée aux CDS ne présentant aucun mot dans leur séquence nucléotidique.



(a) Distance de Bray-Curtis L9M7 CDS_3 (b) Distance de Bray-Curtis L9M7



(c) Distance de Kullback-Leibler L9M7 CDS_3 (d) Distance de Kullback-Leibler L9M7

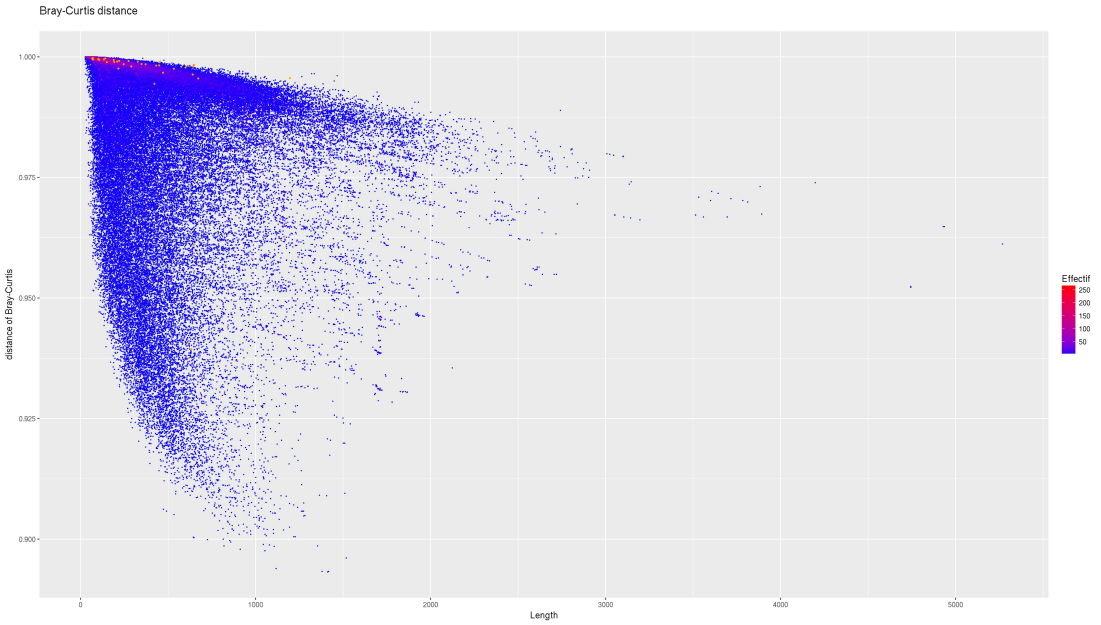
FIGURE 27 – Distances de Bray-Curtis (a et b) et de Kullback-Leibler (c et d) en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 avec un modèle de Markov d'ordre 7 avec les CDS_3 (soit une séquence entière de 27 nucléotides). A gauche, les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (vert ou bleu) ou chacun des CDS OGM de l'ensemble FD (orange) en fonction de la longueur des séquences (abscisses). A droite, les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun des CDS inserts OGM potentiels du maïs OGM (vert ou bleu) par rapport à la longueur des séquences.

Concernant les distances de Bray-Curtis et Kullback-Leibler entre les CDS inserts OGM potentiels du maïs OGM et l'ensemble des CDS du génome de

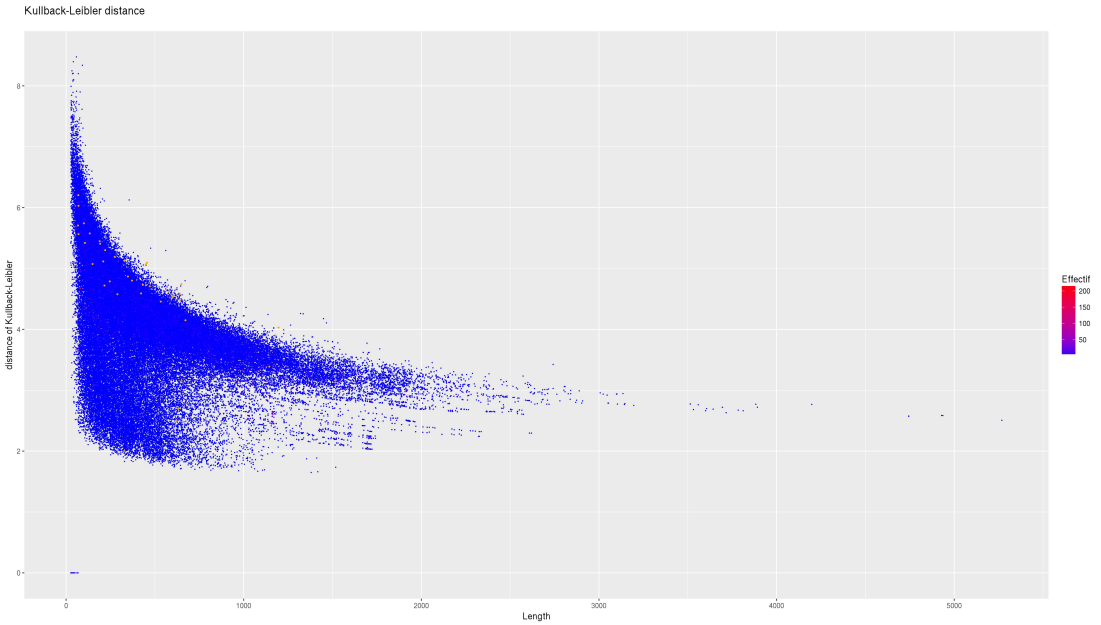
référence (Figure 27, b et c), l'utilisation des CDS_3 est bénéfique pour la distance des CDS de l'insert T25 du maïs OGM (points noirs). En effet, les distances des CDS inserts T25 sont élevées et groupées avec les CDS_3. Le calcul est plus rapide sur les séquences concaténées du fait de leurs longueurs réduites. Le paramètre CDS_3 a donc été conservé pour les prochaines analyses.

5.3.2.8 Visualisation de la densité de points sur les graphiques de représentation de distances

Sur les graphiques de distances de Bray-Curtis et de Kullback-Leibler, nous souhaitions pouvoir déterminer le nombre de points se trouvant au dessus des distances des CDS inserts OGM connus de l'ensemble FD ou pouvant être « cachés » par d'autres points (plusieurs points superposés de même taille et même distance n'étaient pas distinguables jusqu'à présent). La question des chevauchements de points non visibles sur ces graphiques de distance a donc été considérée, avec notamment l'objectif d'écarter la possibilité que des CDS sauvages avec une distance élevée puissent être nombreux mais non visibles de par notre représentation en deux dimensions.



(a) Distance de Bray-Curtis



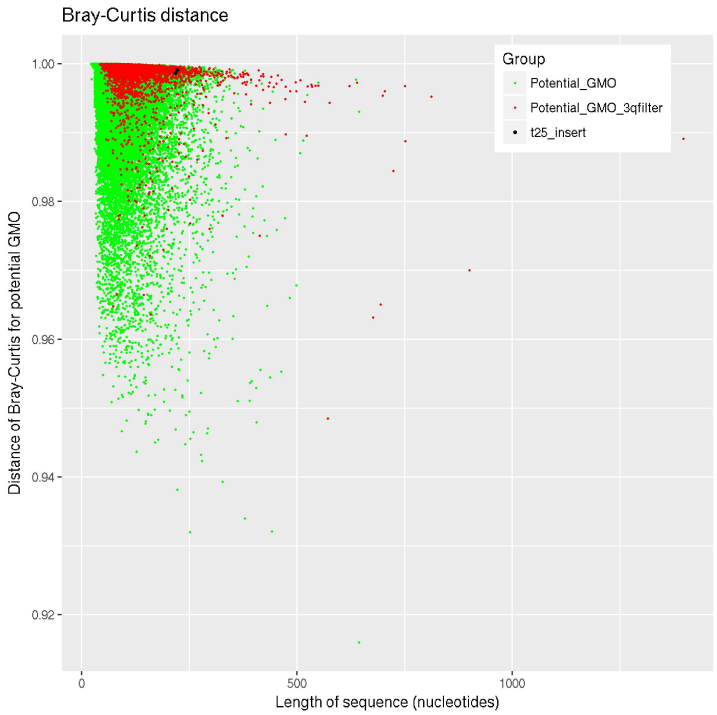
(b) Distance de Kullback-Leibler

FIGURE 28 – Densité de CDS_3 par point pour les calculs de distances de Bray-Curtis (a) et de Kullback-Leibler (b) avec un arrondi à 4 chiffres après la virgule pour les distances. Calcul des distances de Bray-Curtis en fréquences en ne considérant que 10% de mots sur-représentés de taille 9 avec un modèle de Markov d'ordre 7 avec les CDS3. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun de ses CDS (bleu) ou chacun des CDS OGM de l'ensemble FD (orange) en fonction de la longueur des séquences (abscisses).

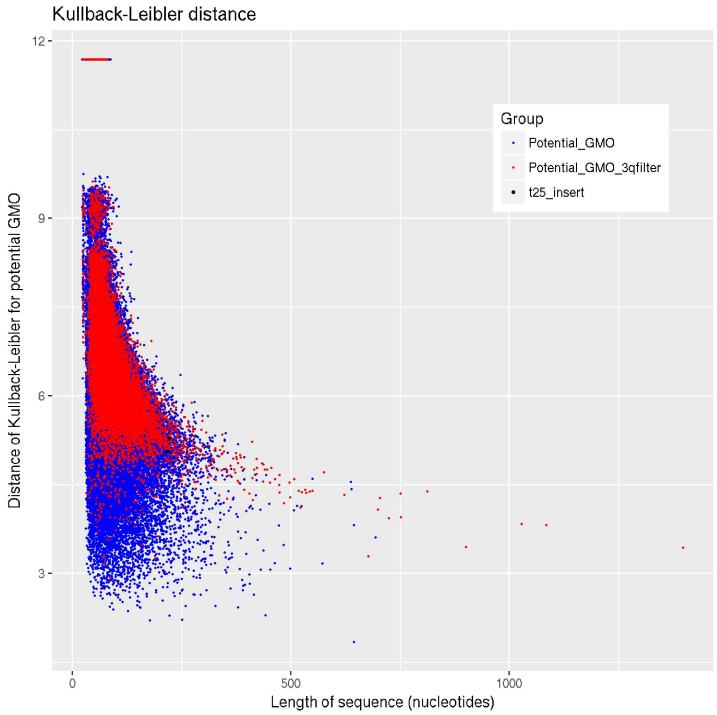
Sur les graphiques précédents (Figure 28), les chevauchements de distances sont peu présents avec un arrondi des valeurs à 4 chiffres après la virgule, surtout avec la distance de Kullback-Leibler. Pour obtenir une visualisation plus importante de la superposition des points avec la distance de Kullback-Leibler, l'arrondi après la virgule doit être de deux chiffres. Enfin, aucune zone avec une densité par point importante n'est supérieure aux distances des CDS inserts OGM connus de l'ensemble FD.

5.3.2.9 Filtre sur la distance avec des mots de taille 3 et un modèle de Markov d'ordre 1 (L3M1)

Un filtre utilisant des mots de taille 3 avec un modèle de Markov d'ordre 1 (L3M1) a été mis en place pour caractériser les distances tenant compte de l'usage du codon (Chapitre 3 partie 3.2.1). Le modèle L3M1 utilise des mots de 3 lettres dont font partie les codons. L'objectif de ce filtre est d'éliminer les CDS potentiellement OGM présentant un usage du codon proche du génome de référence, donc une distance faible. Sur la Figure29, l'insert T25 est identifié en noir, il correspond au résultat que notre méthode de détection doit fournir lorsque l'OGM de maïs est analysé. A la fin du pipeline de nettoyage sur le maïs (décrit en Annexe C), un nombre important de CDS potentiellement OGM est conservé (en vert ou bleu sur la Figure29) dont l'insert T25 fait partie. Notre but est de réduire le nombre de CDS potentiellement OGM en conservant l'insert T25 (résultat cible) lors de l'application du filtre en vue de la construction d'un modèle de prédiction.



(a) Distance de Bray-Curtis L9M7



(b) Distance de Kullback-Leibler L9M7

FIGURE 29 – Application du filtre sur la distance L3M1 (points rouges) aux distances de Bray-Curtis (a) et de Kullback-Leibler (b) en proportions des CDS_3 en ne considérant que 10% de mots sur-représentés de taille 9 en phase 3 avec un modèle de Markov d'ordre 7. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et les CDS potentiellement inserts OGM (vert ou bleu) en fonction de la longueur des séquences (abscisses).

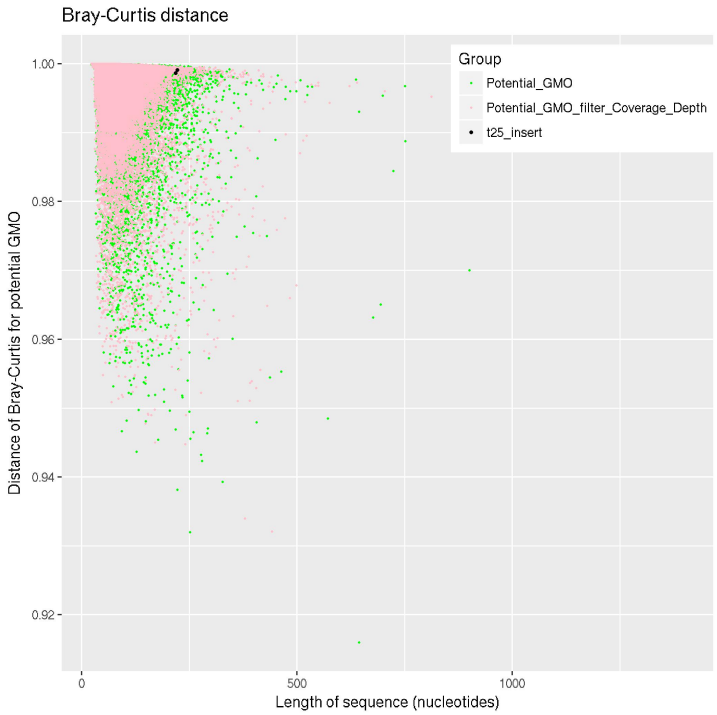
Le filtre L3M1 supprime les CDS inserts OGM potentiels qui ont une distance en L3M1 inférieure au troisième quartile des distances retrouvées entre l'ensemble des CDS du génome de référence et chacun de ses CDS car leur usage des mots de taille 3 est trop proche de celui du génome de référence. Une distance (par exemple, L9M7 avec les paramètres choisies) est ensuite calculée sur les CDS potentiellement OGM qui passent le filtre (en rouge sur la Figure29). Ce filtre a été appliqué sur plusieurs tailles de mots, de 3 à 9 avec un modèle de Markov d'ordre maximal avec des calculs de distances en fréquences et en proportions. La Figure29 présente les meilleurs résultats obtenus pour chacune des deux distances.

Sur les deux graphiques de la Figure 29, le filtre sur la taille de mots L3M1 (points rouges) permet de réduire notablement le nombre de CDS inserts OGM potentiels du maïs OGM pour les distances de Bray-Curtis et Kullback-Leibler. De plus, les CDS de l'insert T25 (points noirs) (Chapitre 2 partie 2.3.1.1) sont inclus dans le filtre et sont donc considérés comme inserts OGM potentiels. Pour la distance de Kullback-Leibler, une série de points avec des valeurs élevées est observée en haut à gauche du graphique. Ces points représentent la valeur maximale allouée pour la distance de Kullback-Leibler. Dans ces cas précis, aucun mot sur-représenté n'est trouvé dans ces CDS. Le filtre de la distance L3M1 a été conservé pour la suite des analyses sur le génome du maïs.

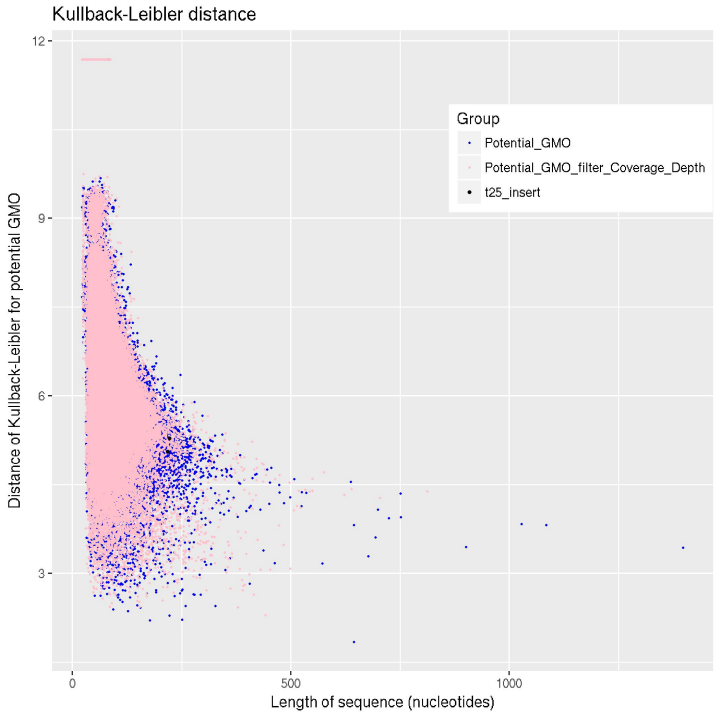
5.3.2.10 Filtre sur la profondeur de couverture

Notre hypothèse est que, chez une plante, une séquence d'insert possède une profondeur de couverture similaire à celle du génome hôte. Pour vérifier cette hypothèse, un filtre sur la profondeur de couverture* a été mis en place pour supprimer les CDS potentiellement inserts dont la profondeur de couverture du contig (séquences provenant d'un assemblage) associé est trop éloignée de celle du génome de référence. En effet, pour déterminer le seuil du filtre, la moyenne de la profondeur de couverture du chromosome le plus long du génome de référence est majorée de 120%. Cette majoration a été déterminée de façon empirique pour ne pas être trop drastique, une marge de 120% par rapport à la profondeur de couverture du plus long chromosome est donc considérée. Tous les CDS dont le contig associé présente une profondeur de couverture inférieure à ce seuil et supérieure à 2 sont conservés. Ce filtre a été appliqué sur plusieurs tailles de mots, de 3 à 9 avec un modèle de Markov d'ordre maximal avec des calculs de distances en fréquences et en proportions. La Figure30 illustre les meilleurs résultats obtenus pour chacune des deux distances.

Comme pour le filtre L3M1 précédent, l'objectif est de réduire le nombre de CDS potentiellement inserts en conservant l'insert T25 dans les résultats. Pour les distances de Bray-Curtis et Kullback-Leibler, le filtre sur la profondeur de couverture discrimine certains CDS inserts OGM potentiels. Les CDS de l'insert T25 (points noirs) sont inclus dans le filtre : ils possèdent une profondeur de couverture similaire à celle du génome de référence du maïs et supérieure à 2. Le filtre de la profondeur de couverture a donc été conservé pour la suite des analyses sur le génome du maïs.



(a) Distance de Bray-Curtis



(b) Distance de Kullback-Leibler

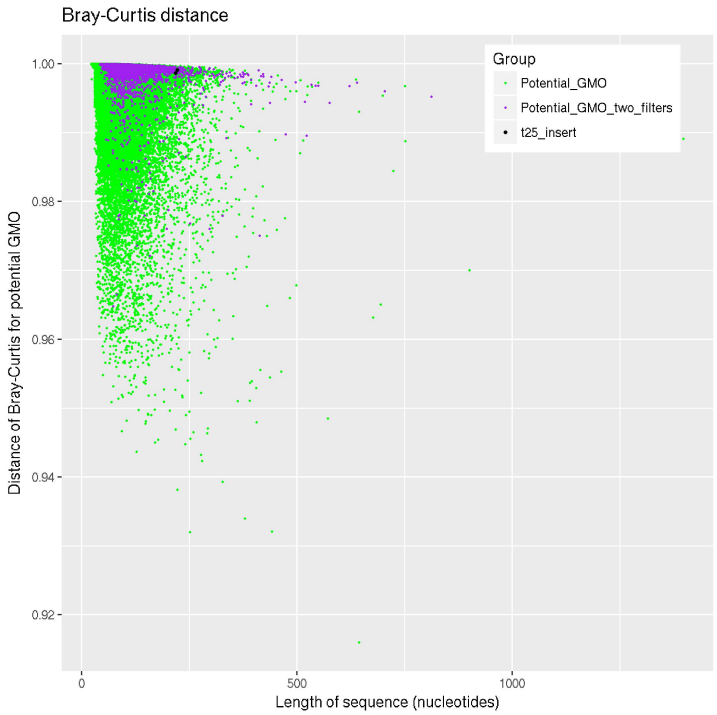
FIGURE 30 – Application du filtre sur la profondeur de couverture (points roses) aux distances de Bray-Curtis (a) et de Kullback-Leibler (b) en proportions des CDS_3 en ne considérant que 10% de mots sur-représentés de taille 9 en phase 3 avec un modèle de Markov d'ordre 7. Les données visualisées sont les distances entre l'ensemble des CDS du génome de référence du maïs et chacun des CDS potentiellement inserts OGM (vert ou bleu) par rapport à la longueur des séquences.

5.3.2.11 Bilan : paramètres les plus probants

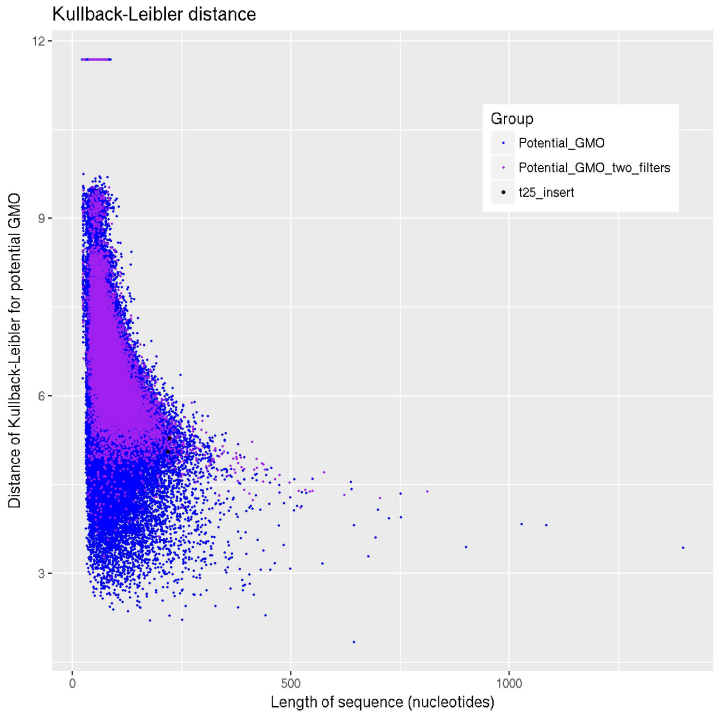
Les meilleurs paramètres obtenus pour la distance de Bray-Curtis et de Kullback-Leibler en fréquences avec le génome du maïs OGM sont la combinaison des deux filtres (L3M1 et profondeur de couverture) et la concaténation des troisièmes positions des codons en ne considérant que 10% de mots sur-représentés de taille 9 et un modèle de Markov d'ordre maximal.

Concernant les distances en proportions, les meilleurs paramètres pour la distance de Bray-Curtis sont des mots de taille 5 avec un modèle de Markov d'ordre 3 (Figure 31a) et pour la distance de Kullback-Leibler, des mots de taille 6 avec un modèle de Markov d'ordre 4 en considérant tous les mots (Figure 31b).

La distance de Bray-Curtis permet de discriminer un nombre plus important de CDS inserts OGM potentiels après l'application des deux filtres (L3M1 et profondeur de couverture) par rapport à la distance de Kullback-Leibler.



(a) Distance de Bray-Curtis L5M3



(b) Distance de Kullback-Leibler L6M4

FIGURE 31 – Application des deux filtres (points violets) aux distances de Bray-Curtis (a) et de Kullback-Leibler (b) en proportions des CDS_3 en ne considérant que 10% de mots sur-représentés de taille 9 avec un modèle de Markov d'ordre 7. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence du maïs et chacun des CDS potentiellement inserts OGM (vert ou bleu) en fonction de la longueur des séquences (abscisses).

5.3.3 Analyse d'un OGM bactérien plutôt que végétal

Pour connaître la proportion de reads d'origine bactérienne présentes dans les données de séquençage du maïs OGM fournis par le Laboratoire de la Santé des Végétaux (LSV), un alignement Bowtie2 [118] de ses reads a été réalisé sur la banque de données NCBI Nucléotides installée localement. Le résultat a ensuite été converti en sortie BLAST [121] et visualisé via le logiciel Krona [131] sur la Figure 32. Sur cette figure, le maïs OGM s'aligne à différentes variétés de maïs à environ 91% (*Zea mays*), à d'autres plantes à environ 8% et à un peu plus de 1% de bactéries sauvages. Les bactéries de l'environnement ainsi que la contamination bactérienne étant très présentes dans les données de séquençage du maïs, il est très difficile de les distinguer de celles provenant d'un insert OGM.

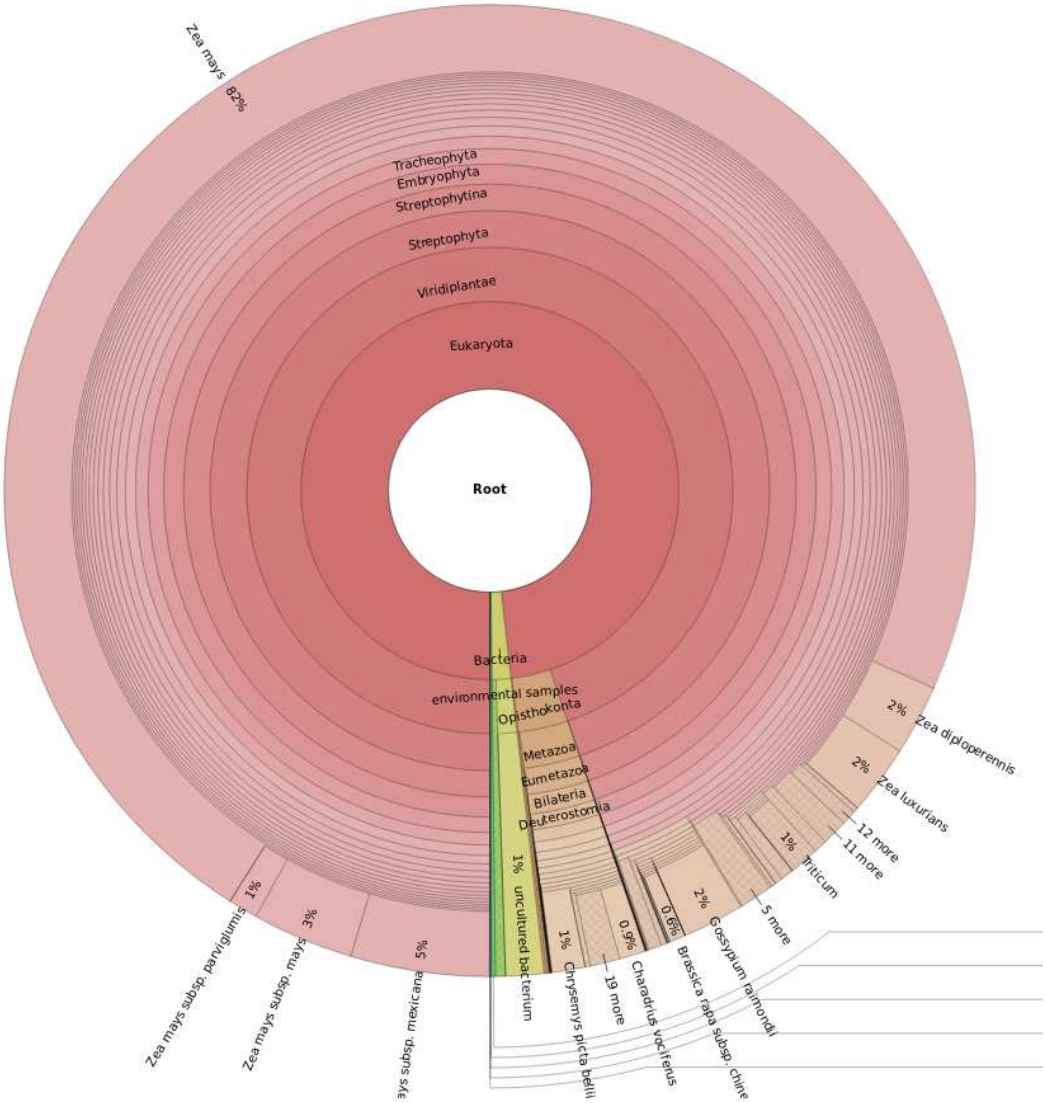


FIGURE 32 – Visualisation de la composition des reads des données de séquençage du maïs OGM (fournis par la LSV) à l'aide de l'outil Krona.

La taille du génome du maïs est élevée (2,134 Gbp) et l'ensemble de nos analyses (du pipeline de nettoyage aux calculs d'une distance comme L9M7) nécessite un temps de calcul d'environ une semaine. Pour la bactérie GM qui possède un génome de petite taille, la totalité des étapes permettant de détecter un OGM inconnu est réalisée en quelques heures. L'impossibilité de différencier les CDS appartenant à des bactéries provenant de l'environnement des CDS bactériens d'inserts GM d'une part, ainsi que le temps de calcul d'autre part, nous ont orienté vers l'analyse de bactéries GM pour affiner les paramètres à utiliser sur un cas plus simple.

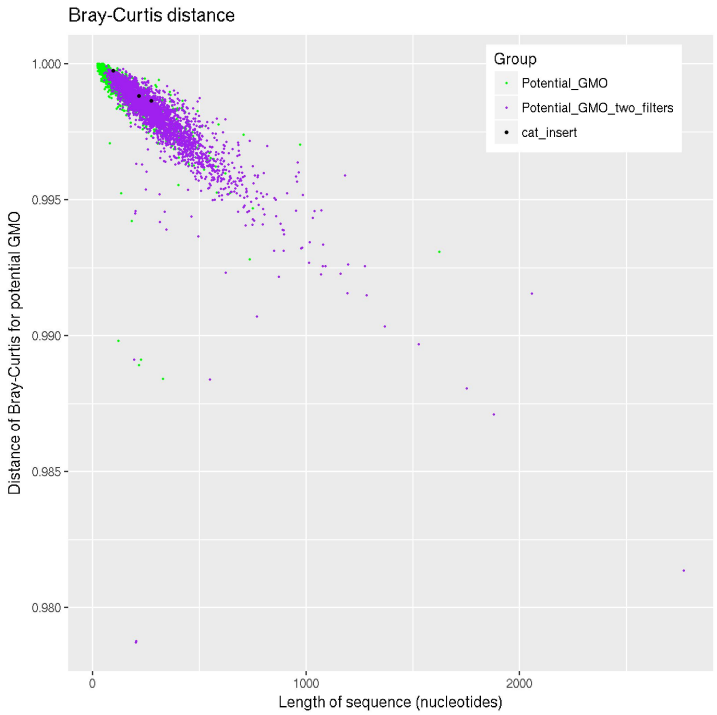
5.3.4 Cas de la bactérie *B. subtilis*

Après l'utilisation des paramètres les plus probants obtenus à partir du maïs OGM et appliqués à la bactérie *B. subtilis* GM, la stratégie de nettoyage a été modifiée pour prendre en compte les CDS du génome hôte plutôt que les CDS du génome de référence. Les distances des CDS OGM connus et potentiels sont donc comparées aux distances de l'ensemble des CDS du génome hôte. Pour les graphiques suivants, les jeux de données utilisés sont les distances entre l'ensemble des CDS du génome hôte de la bactérie *B. subtilis* GM et chacun des CDS du génome hôte, des CDS de la banque de données d'OGM connus et des CDS inserts OGM potentiels.

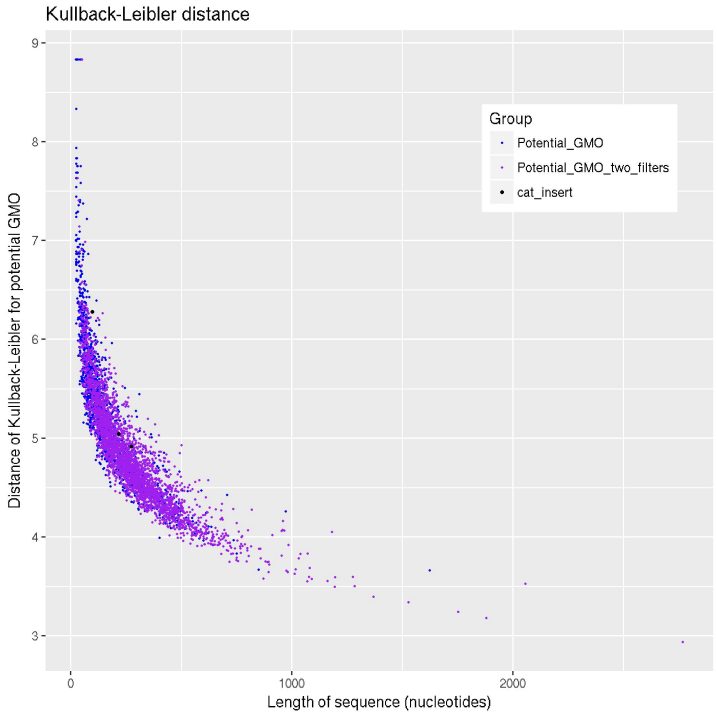
Les meilleurs paramètres obtenus avec le génome du maïs OGM sont testés avec la bactérie *B. subtilis* GM (Chapitre 5 partie 5.3.4.1). Des ajustements ont été nécessaires pour parvenir à isoler les CDS inserts OGM. Les mots de 27 lettres en ne prenant que les troisièmes positions des codons (CDS_3), tirés de l'observation chez le maïs, ont été conservés pour les analyses sur la bactérie *B. subtilis* GM. Pour chaque taille de mot, l'ordre de Markov maximal est conservé. En proportions, les distances de Bray-Curtis et Kullback-Leibler ont été testées avec tous les mots de taille 3 à 8. En fréquences, les distances de Bray-Curtis et Kullback-Leibler ont été testés avec 10% de mots sur-représentés de taille 3 à 9 pour les fréquences avec les CDS_3 (Chapitre 5 partie 5.3.4.2 et 5.3.4.3) sauf pour la distance L3M1 où tous les mots sont conservés (il y a peu de mots de 3 lettres). Le calcul des distances en proportions avec des mots de 9 lettres n'est pas réalisable sur la totalité du génome de la bactérie *B. subtilis* GM avec le logiciel R'MES.

5.3.4.1 Combinaison des filtres L3M1 et profondeur de couverture

Comme avec le maïs OGM, les filtres utilisant la distances L3M1 et la profondeur de couverture ont été appliqués aux jeux de données de la bactérie *B. subtilis* avec les mêmes paramètres.



(a) Distance de Bray-Curtis



(b) Distance de Kullback-Leibler

FIGURE 33 – Application des deux filtres (en violet) aux distances de Bray-Curtis (a) et de Kullback-Leibler (b) en fréquences des CDS_3 en ne considérant que 10% de mots de taille 9 avec un modèle de Markov d'ordre 7. Les données visualisées sont les distances (ordonnées) entre l'ensemble des CDS du génome de référence de *Bacillus subtilis* et chacun des CDS potentiellement inserts OGM (vert ou bleu) en fonction de la longueur des séquences (abscisses).

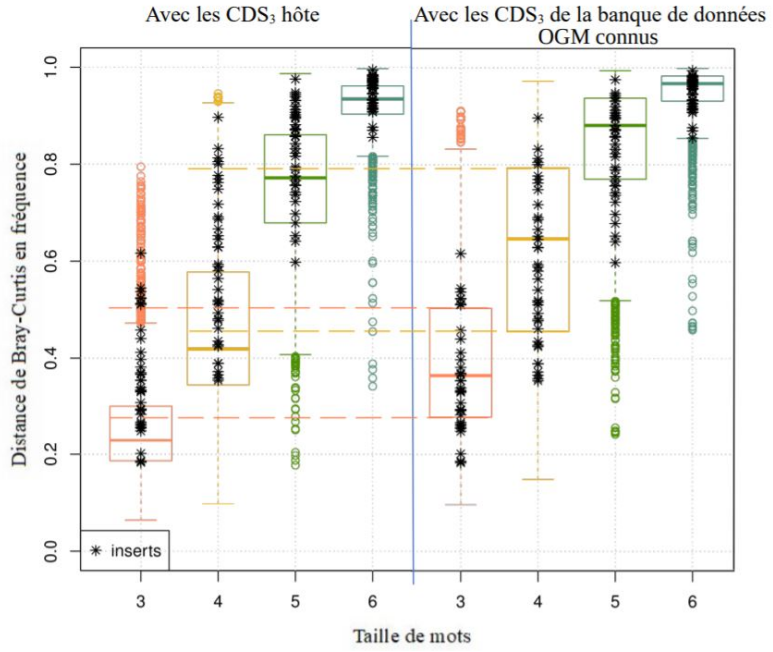
La combinaison des deux filtres, la distance L3M1 et la profondeur de couverture, ne permet pas de discriminer un nombre important de CDS sur la bactérie par rapport aux résultats obtenus pour le maïs OGM. Le filtre sur la profondeur de couverture n'est pas applicable aux bactéries. En effet, un insert peut être inséré dans un plasmide et ainsi posséder une profondeur de couverture beaucoup plus importante que le génome hôte. Les filtres usage du codon et profondeur de couverture ont été abandonnés dans les analyses suivantes sur des bactéries mais resteront d'actualité dans les versions futures de la méthode dédiées aux eucaryotes.

5.3.4.2 Influence de la taille de mots sur la distance en fréquences

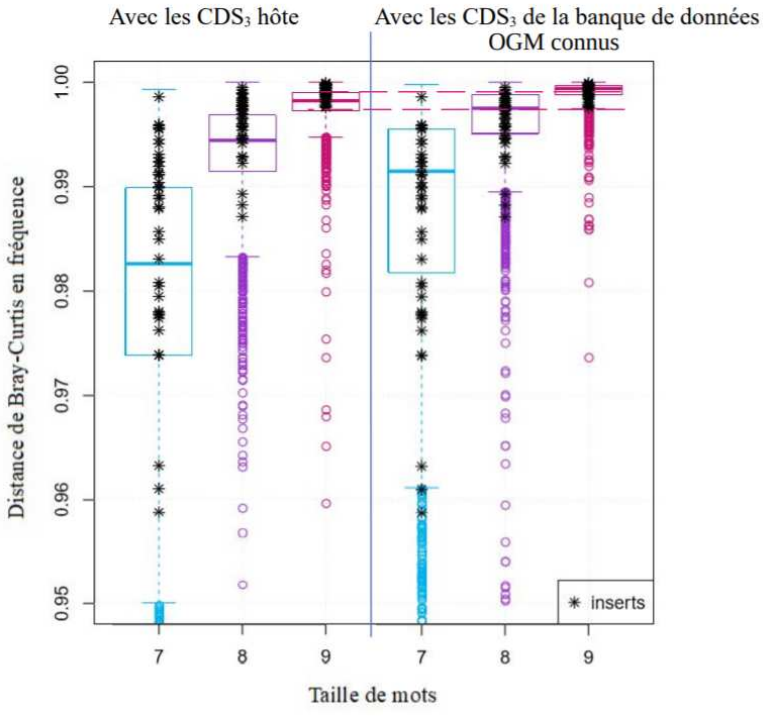
En fréquences, les tailles de mots de 3 à 9 sont traités en considérant uniquement 10% de mots sur-représentés sur des CDS_3 sauf pour L3M1 où tous les mots sont pris en compte (car il y a peu de mots de taille 3). Ces paramètres sont testés avec la distance de Bray-Curtis et la distance de Kullback-Leibler.

Sur la Figure34, les CDS_3 inserts OGM ont des distances de Bray-Curtis très diverses. Les résultats attendus sont que les CDS_3 inserts OGM ont des distances supérieures aux CDS_3 du génome hôte, donc plutôt au-dessus du corps de la boîte à moustache des CDS_3 appartenant au génome hôte. Par rapport aux CDS_3 de la banque d'inserts OGM connus, les CDS_3 inserts OGM (Figure 34, étoiles noires) devaient être à l'intérieur du corps de la boîte à moustache (50% des données) des CDS_3 appartenant à la banque de données d'inserts OGM connus.

Aucune taille de mots ne se distingue par rapport aux autres entre les deux ensembles. Cependant, avec des mots de 9 lettres et un modèle de Markov d'ordre maximal (L9M7), les CDS_3 de la banque de données OGM connus ont des distances plus élevées que les CDS_3 apparentés au génome hôte. De plus, la distance L9M7 a montré, chez le maïs, une capacité à grouper les CDS OGM de l'ensemble FD avec des valeurs de distance élevées. Le modèle L9M7 est donc privilégié pour la suite des analyses.



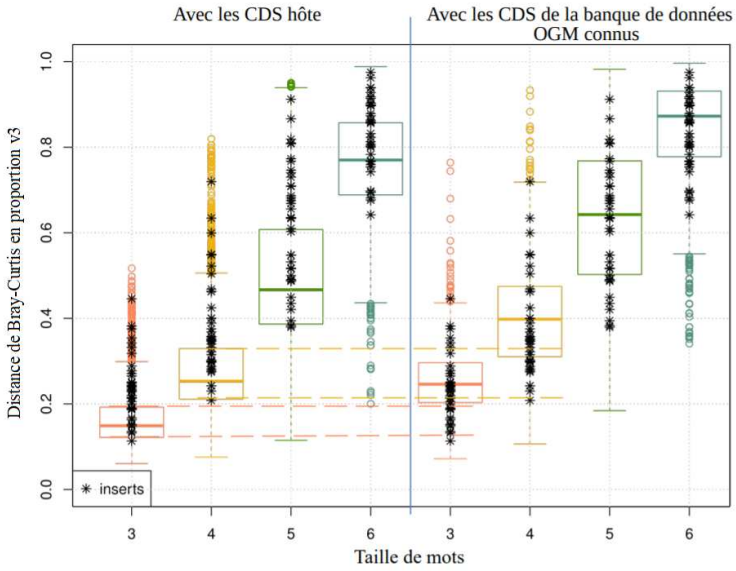
(a) Distance de Bray-Curtis en fréquence avec des mots de 3 à 6 lettres



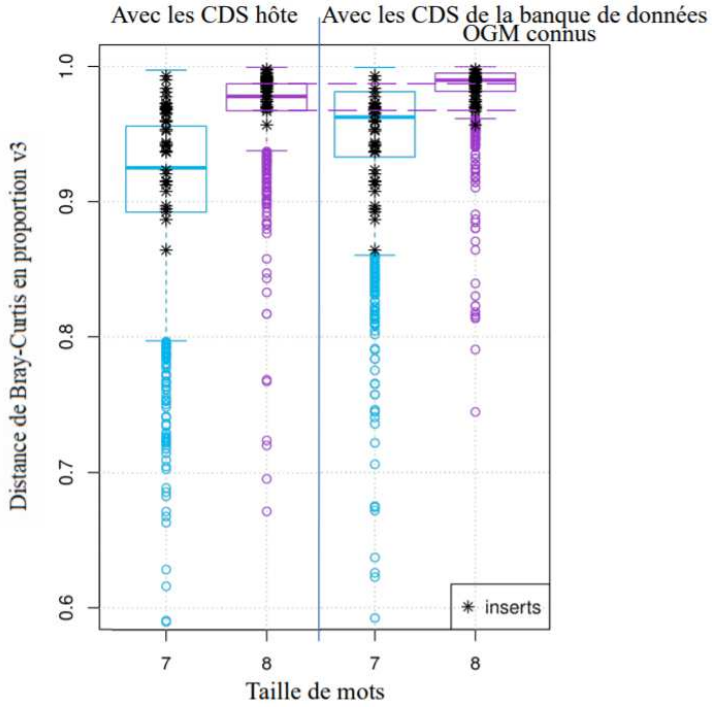
(b) Distance de Bray-Curtis en fréquence avec des mots de 7 à 9 lettres

FIGURE 34 – Comparaison des résultats de la bactérie *B. subtilis* avec des tailles de mots de 3 à 9 et un ordre de Markov maximal sur la distance de Bray-Curtis en fréquences des CDS_3 en ne considérant que 10% de mots sur-représentés sur les données des CDS hôte (partie gauche) ou la banque de données de CDS inserts OGM connus (partie droite).

5.3.4.3 Influence de la taille de mots sur la distance en proportions



(a) Distance de Bray-Curtis en proportion avec des mots de 3 à 6 lettres



(b) Distance de Bray-Curtis en proportion avec des mots de 7 et 8 lettres

FIGURE 35 – Comparaison des résultats de la bactérie *B. subtilis* avec des tailles de mots de 3 à 8 avec un ordre de Markov maximal sur la distance de Bray-Curtis en proportions en considérant tous les mots sur les données des CDS de l'hôte (partie gauche) ou la banque de données de CDS d'inserts OGM connus (partie droite).

Pour caractériser les CDS inserts d'une bactérie GM, l'influence de la taille de mots a été testée sur la bactérie *B. subtilis* GM pour toutes les tailles de mots allant de 3 à 8 pour les distances de Bray-Curtis et de Kullback-Leibler en proportions et en considérant tous les mots. Les résultats avec la distance de Bray-Curtis sont illustrés sur la Figure35.

Comme avec les distances en fréquences, les distances des CDS d'inserts OGM sont diverses. Cependant, avec une taille de mot de 3 ou 4, l'intervalle entre le premier et le troisième quartile des deux boîtes à moustaches, celle pour les CDS de l'hôte et celle pour les CDS d'OGM connus, ne se chevauchent pas ou très peu. Les distances obtenues pour ces deux tailles de mots permettent de mieux caractériser les CDS de l'hôte par rapport aux CDS d'OGM connus. Les distances en proportions L3M1 et L4M2 sont donc privilégiées pour la suite des analyses.

5.4 Conclusion

De nombreuses stratégies ont été testées à l'aide des paramètres du logiciel R'MES pour optimiser la détection d'OGM inconnus. Les résultats obtenus avec la distance de Kullback-Leibler sont proches de ceux obtenus avec la distance de Bray-Curtis. Néanmoins, cette dernière a été choisie car, avec la bactérie *B. subtilis* GM, elle permet de mieux caractériser les inserts avec des valeurs de distances élevées. De plus, elle possède une valeur maximale bien déterminée (égale à 1 lorsque les deux ensembles de données n'ont aucun point en commun et à 0 lorsque ceux-ci ont une composition identique) contrairement à la distance de Kullback-Leibler qui ne possède pas de maximum défini. Enfin, la distance de Bray-Curtis fournit des distances plus homogènes par rapport à la distance de Kullback-Leibler.

La distance de Bray-Curtis a été choisie avec trois combinaisons paramétriques différentes pour être utilisées conjointement comme variables explicatives dans l'étape de construction d'un modèle de prédiction décrite dans le Chapitre 6 :

- P L3M1 : la distance de Bray-Curtis en proportions en considérant tous les mots de 3 lettres selon le modèle de Markov d'ordre 1
- P L4M2 : la distance de Bray-Curtis en proportions en considérant tous les mots de 4 lettres selon le modèle de Markov d'ordre 2
- F L9M7 : la distance de Bray-Curtis en fréquences en considérant uniquement 10 % de mots de 9 lettres les plus sur-représentés selon le modèle de Markov d'ordre 7 sur les CDS_3.

La combinaison P L3M1 a été choisie car ce calcul caractérise les mots de trois lettres des CDS, ce qui est spécifique au vocabulaire du génome de l'hôte, y compris l'utilisation des codons. La combinaison P L4M2 caractérise les CDS en utilisant des mots de petite taille et donc plus fréquents, distincts de l'usage du codon, afin d'obtenir une distribution plus précise des mots sur le CDS. Outre son apport d'informations, la combinaison P L4M2 contribue à l'insensibilité de la méthode lorsque l'insert OGM présente un usage des codons/dicodons (deux triplets de nucléotides) optimisé par rapport au génome hôte. Enfin, la distance de Bray-Curtis en fréquences sur les CDS_3 en ne considérant que 10% de mots sur-représentés en L9M7 analyse les CDS en utilisant des mots longs et sur-

représentés (en raison de la concaténation de la troisième position du codon). Ce paramétrage rend les mots trouvés dans le CDS comparé très spécifiques (en raison de leur longueur) et adaptés à l'usage du codon du génome de l'hôte (en considérant seulement les troisièmes lettres des mots de 27 lettres du CDS).

Ces calculs de distance de Bray-Curtis avec les paramétrages P L3M1, P L4M2 et F L9M7 sont effectués sur chacun des trois ensembles de CDS obtenus à la fin du pipeline de nettoyage (Chapitre 4 partie 4.2), comme résumé dans la Figure 36. Seuls les CDS d'une longueur supérieure ou égale à 27 nucléotides sont utilisés, en raison de la longueur minimale de la taille des mots considérée dans le calcul de la distance de Bray-Curtis L9M7 en fréquences (Chapitre 5 partie 5.3.2.7).

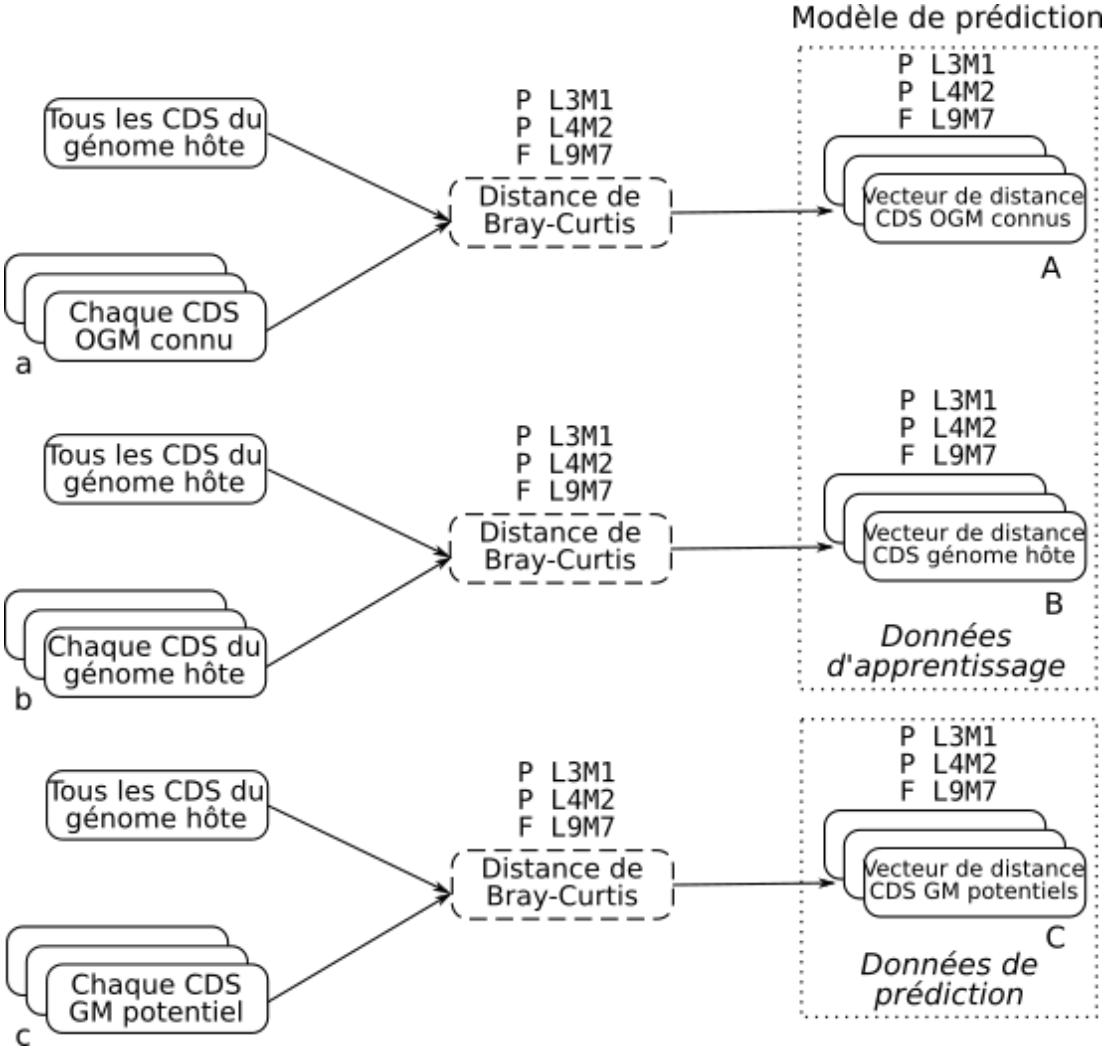


FIGURE 36 – Distances de Bray-Curtis calculées en proportions (P L3M1 et P L4M2) et en fréquences (F L9M7) pour obtenir les données d'apprentissage et de prédiction utilisées pour l'établissement d'un modèle de prédiction.

Ce chapitre permet de fournir des variables potentiellement explicatives, les calculs de distances de Bray-Curtis avec les paramétrages P L3M1, P L4M2, F L9M7, de la variable à expliquer (« OGM » ou « non OGM »).

Chapitre 6

Proposition d'un modèle de prédiction des OGM inconnus

Sommaire

Pour valider les différences établies par l'utilisation d'un calcul de distance entre le génome hôte et ces CDS potentiellement inserts, un modèle de prédiction est utilisé. Le machine learning ou apprentissage automatique est un domaine d'étude incluant l'application et le développement d'algorithmes informatiques permettant d'établir un modèle à partir de données. Les méthodes de machine learning permettent ainsi d'expliquer et de prédire des données [9, 132].

6.1 Objectifs

L'étape d'élaboration d'un modèle de prédiction a pour but de prédire les CDS inserts OGM avérés. L'objectif est de construire un modèle statistique de prédiction du statut d'un échantillon contenant potentiellement des OGM inconnus. Dans ce but, des variables explicatives sont utilisées pour caractériser les

données de l'échantillon et ainsi prédire la variable cible ou de prédiction (dans notre cas, « OGM » ou « non OGM »).

6.2 Méthodes

6.2.1 Données

Les différents tests réalisés avec les données de séquençage du maïs OGM ont constitué un travail préparatoire pour l'application de la méthode aux eucaryotes et aux bactéries. Le jeu de données de séquençage de la bactérie *B. subtilis* GM publié par le laboratoire du JRC, a constitué le mètre étalon pour élaborer notre méthode de détection des OGM inconnus sur les bactéries.

Deux jeux de données de séquençage haut débit de la bactérie *B. subtilis* sont utilisés pour cette mise au point. Le premier jeu correspond à une bactérie GM et le second à une bactérie sauvage (Chapitre 2 partie 2.3.2.1). De plus, la banque de données de CDS OGM de différentes espèces fournie par le JRC [56] et complétée par des CDS d'inserts GM bactériens issus ou déduits de la littérature (Chapitre 4 partie 4.2.3) est utilisée lors de l'apprentissage automatique.

Le nombre de variables explicatives est identique pour tous les jeux de données. Cependant, le nombre d'observations dépend du jeu de données. Ces informations sont précisées dans le Chapitre 7 partie 7.3 lors de l'application de ces données.

6.2.1.1 Variables explicatives sélectionnées

Les variables explicatives permettent de caractériser les données de l'échantillon et ainsi d'établir un modèle. Dans le chapitre précédent (Chapitre 5 partie 5.4), les distances de Bray-Curtis avec les paramétrages P L4M2, P L3M1 et F L9M7 calculées entre l'ensemble des CDS du génome hôte et chacun de ces CDS ou chacun des CDS de la banque de données d'OGM connus ou chacun des CDS inserts GM potentiels ont été sélectionnées. Ces trois distances constituent une partie des variables explicatives utilisées pour l'élaboration du modèle de prédiction.

En complément de ces distances de Bray-Curtis, d'autres variables explicatives sont calculées pour chaque CDS des trois ensembles de données (CDS du génome hôte, CDS OGM connus et CDS potentiels). Ces variables sont :

- la longueur du CDS,
- la moyenne $A_H(S)$ d'une séquence S des scores d'exceptionnalité fournis par R'MES dans le génome hôte pour L4M2 et L9M7 (Équation 6.1 ci-après),
- la densité de comptage par nucléotide dans le CDS considéré pour les mots de 4 et 9 lettres (somme des comptages $C(m,S)$ de tous les mots sélectionnés m, divisée par la longueur du CDS S)
- le pourcentage en GC du CDS considéré.

La moyenne des scores d'exceptionnalité permet d'estimer l'exceptionnalité moyenne des mots trouvés dans le CDS considéré compte tenu de la composition des CDS

du génome hôte. Elle est décrite par l'équation :

$$A_H(S) = \frac{\sum_{m \in M} K(m, H) \times C(m, S)}{\sum_{m \in M} C(w, S)} \quad (6.1)$$

où $K(m, H)$ est le score d'exceptionnalité de R'MES d'un mot m dans les CDS du génome hôte H. *C(m,S)* correspond au nombre d'occurrences de mots m dans la séquence S dans l'ensemble des mots sélectionnés M et w représente chaque mot.

Pour la création du modèle de prédiction, neuf variables explicatives sont présentes.

6.2.1.2 Données d'apprentissage et de prédiction

Les données d'apprentissage sont constituées des CDS apparentés au génome hôte d'une part et des CDS de la banque de données d'inserts OGM connus filtrés d'autre part. Ces données ont un statut OGM/non-OGM connu, elles sont utilisées pour construire un modèle visant à différencier un CDS insert GM d'un CDS appartenant au génome hôte.

Les CDS codant les ARN non messagers ont été supprimés de ces données d'apprentissage, un CDS insert OGM ne peut être uniquement constitué d'ARNr (ARN ribosomique) ou d'ARNt (ARN de transfert). Si un insert OGM était constitué uniquement d'ARNr cela conduirait probablement à une mort prématuré de la cellule. En effet, les mécanismes impliquant les ARNr (comme la traduction) représentent une machinerie cellulaire complexe, la modifier est très difficile. Si un insert OGM était constitué uniquement d'ARNt, cela reviendrait à modifier l'usage du codon car les ARNt sont impliqués dans la synthèse des protéines. L'expression d'un nombre important de gènes serait alors modifié entraînant une diminution de la valeur sélective de la bactérie. La bactérie serait ainsi moins compétitive, voire non viable.

Les données de prédiction sont constituées des CDS OGM candidats qui ne se sont pas alignés sur le pangénome (Chapitre 4 partie 4.2.2).

Dans le cas des données de *B. subtilis* GM, 4102 CDS sont identifiés comme apparentés au génome hôte et 2714 CDS comme provenant de la banque de données d'OGM connus filtrée. Le nombre de CDS potentiellement OGM est de 39 où 37 sont des CDS OGM avérés et 2 sont des CDS non-OGM.

6.2.2 Contexte statistique de la discrimination

Le machine learning utilise des algorithmes d'apprentissage qui, à partir d'un jeu de données, retournent plusieurs modèles. Les algorithmes d'apprentissage supervisés (Figure 37) utilisent, en complément des variables explicatives, une variable cible, à prédire (par exemple, « OGM » ou « non OGM » dans le cadre de ces travaux de recherche). Il existe différents types d'apprentissages supervisés en fonction de la nature de la variable cible : ceux pour résoudre des problèmes de discrimination et ceux pour résoudre des problèmes de régression. La discrimination est utilisée lorsque la variable cible est qualitative et la régression lorsque la

variable cible est numérique. Notre modèle de prédiction se situe dans un contexte supervisé où la variable cible est booléenne (« OGM » ou « non OGM »). Les méthodes standards de discrimination statistique sont donc utilisées.

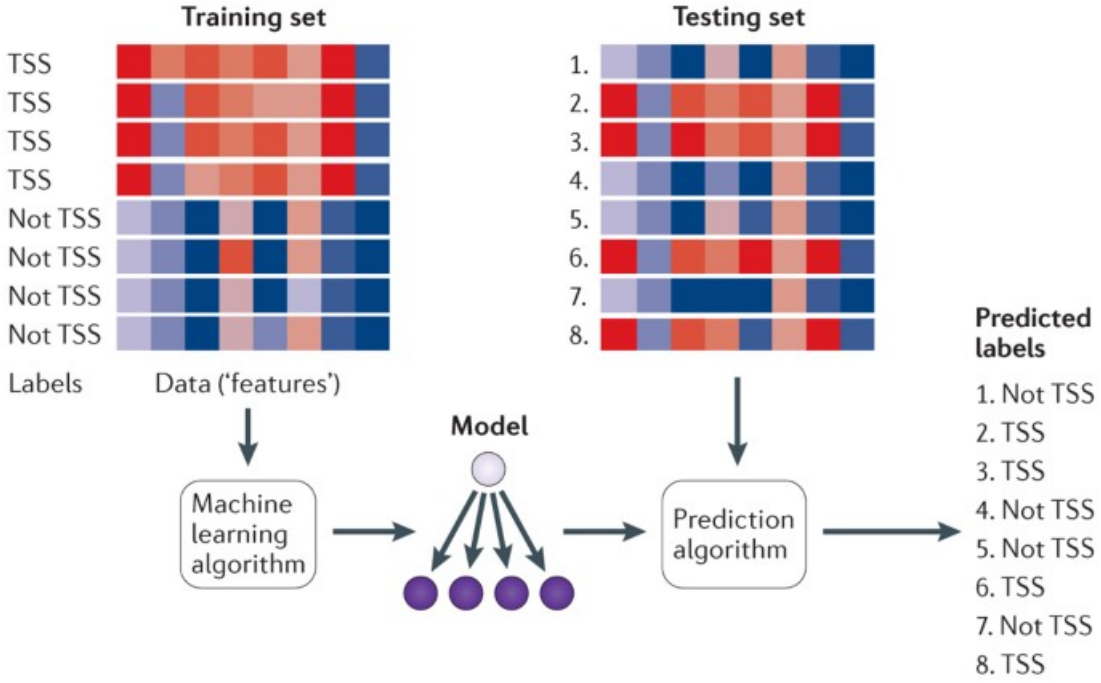


FIGURE 37 – Exemple schématique d'application de machine learning pour la prédiction des sites d'initiation de la transcription ou Transcription Start Sites (TSS). Les données d'apprentissage (Training set), constituées de séquences ADN, sont fournies en entrée d'une procédure d'apprentissage avec les informations indiquant si chaque séquence est centrée sur un TSS ou non. L'algorithme d'apprentissage produit un modèle qui est ensuite utilisé pour prédire si les séquences ADN présentent dans les données de prédiction (testing set) sont centrées sur un TSS ou non (Predicted labels) [9].

La discrimination permet d'établir une différence basée sur les variables explicatives utilisées. De nombreuses méthodes de discrimination sont disponibles (notamment dans le package caret sous le langage de programmation et logiciel R [133]).

6.2.3 Utilisation de l'apprentissage automatique

Plusieurs méthodes supervisées correspondant à des modèles de discrimination ont été appliquées à nos données. Le choix de ces méthodes a été réalisé en sélectionnant des méthodes couramment utilisées comportant à la fois des méthodes paramétriques et non paramétriques. Les méthodes paramétriques sont principalement basées sur des hypothèses de distribution des données (généralement Gaussienne) contrairement aux méthodes non paramétriques.

Douze méthodes de discrimination qualitative sont testées et comparées au

sein d'une procédure d'apprentissage automatique. Les méthodes paramétriques sélectionnées sont :

- Le modèle linéaire généralisé (Logit), qui correspond à une généralisation de la régression linéaire. Les modèles logistiques visent à construire un modèle expliquant une variable expliquée binaire à partir d'un ensemble de variable explicatives à partir d'un modèle Logit. Ce modèle est résolu par maximisation de la vraisemblance.
- L'analyse discriminante linéaire stepwise (stepLDA), qui recherche des combinaisons linéaires des variables explicatives qui séparent au mieux les observations selon les modalités de la variable de prédiction. L'option stepwise permet de sélectionner automatiquement les meilleures variables explicatives à utiliser dans le modèle.
- L'analyse discriminante quadratique stepwise (stepQDA). Cette méthode recherche des combinaisons non linéaires des variables explicatives pour séparer les observations selon les modalités de la variable à expliquer.
- La régression par les moindres carrés partiels (PLS). Elle est élaborée à partir d'une régression linéaire sur composantes. Une régression par les moindres carrées est ensuite effectuée sur ces composantes plutôt que sur les données initiales.

Les méthodes non paramétriques choisies sont :

- Les réseaux de neurones (nnet). Cette méthode inspirée de la structure et du fonctionnement des réseaux de neurones biologiques. Des réseaux représentent une perception constituée d'une ou plusieurs couches de neurones (i.e., variables explicatives) liées à la variable de prédiction par des poids. Ces poids sont mis à jour par un algorithme du gradient qui calcule le gradient de l'erreur pour chaque neurone, de la dernière couche vers la première.
- La machine à vecteur de support. Avec une fonction noyau de base radiale (svmRadial), elle permet de classer les observations selon un hyperplan construit dans l'espace des variables. Un hyperplan d'un espace vectoriel de dimension quelconque représente la généralisation des plans vectoriels d'un espace de dimension 3. A cette fin, une fonction noyau de base radiale est utilisée pour transformer les données d'entrée non linéaires en données linéaires dans un espace dimensionnel plus grand dans lequel il est probable qu'il existe une séparation.
- Les k plus proches voisins (knn). C'est une méthode basée sur des comparaisons avec les observations adjacentes. Elle utilise l'ensemble des observations pour définir un nombre k permettant d'affecter la classe d'une observation suivant ses k plus proches voisins.

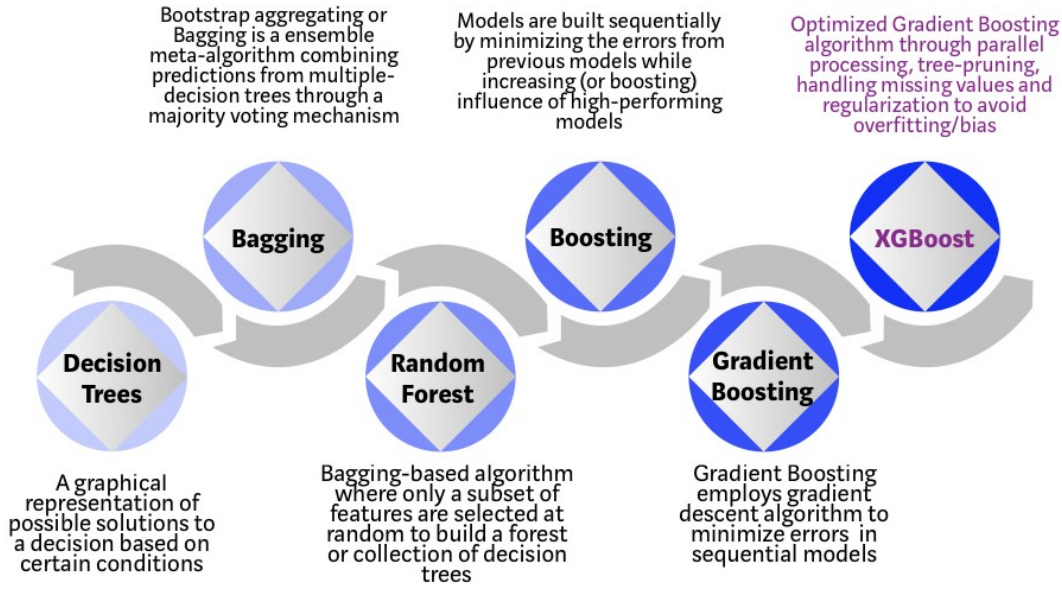


FIGURE 38 – Évolution des techniques de machine learning concernant les arbres de décision [134].

Parmi les méthodes non paramétriques, certaines sont basées sur des arbres de décision qui utilisent une représentation en forme d'arbre pour classifier les observations (Figure 38) :

- L'algorithme de discrimination C5.0 (C5.0). Cet algorithme est décrit par Quinlan [135], il ajuste des modèles d'arbres de décision pour la discrimination des données.
- Les arbres de partitionnement récursifs (rpart). Ces arbres sont inspirés de l'approche des arbres de discrimination et de régression (CART) introduits par Breiman *et al.* en 1984 [136]. Ils permettent de réaliser l'apprentissage selon un partitionnement récursif des observations grâce à des règles appliquées aux variables explicatives.
- Les arbres de discrimination avec Bagging (Treebag). Ces arbres utilisent la méthode du Bagging, contraction de Boostrap Aggregating, introduite par Breiman *et al.* en 1996 [137]. Le Bagging consiste à agréger par vote majoritaire, indépendamment les unes des autres, une collection de variables d'entrée faibles pour obtenir de meilleurs variables de prédiction.
- Les forêts aléatoires (RF). Cette méthode combine la construction de nombreux arbres de décision avec l'approche du Bagging. Un grand nombre d'arbres décisionnels est construit sur un nombre restreint de variables explicatives pour être ensuite entraîné sur des sous ensembles du jeu de données.
- L'extrême gradient boosting (xgboost). Cette méthode utilise l'algorithme de plusieurs arbres de Boosting de gradient. Le Boosting de gradient [138] combine les résultats d'un ensemble de modèles plus simples et plus faibles élaborés séquentiellement à partir des données d'apprentissage permettant ainsi une auto-amélioration des prédictions.

6.2.4 Critère de sélection de la/des méthodes les plus performantes

Pour évaluer les performances d'un modèle, plusieurs indicateurs de performances sont disponibles. Dans le cas d'un problème de classification supervisée à deux classes, seules quatre situations sont possibles lorsque le modèle attribue une classe à une observation. Ces situations sont résumées dans une matrice de confusion (Tableau 7). Les quatre situations sont les suivantes :

— Les faux positifs (FP) correspondent aux observations de la classe non-OGM mal prédits par le modèle.
— Les faux négatifs (FN) correspondent aux observations de la classe OGM mal prédits par le modèle.
— Les vrais positifs (VP) correspondent aux observations de la classe OGM bien prédits par le modèle.
— Les vrais négatifs (VN) correspondent aux observations de la classe non-OGM bien prédits par le modèle.

		Prédiction du modèle	
		Classe OGM	Classe non-OGM
Données réelles	Classe OGM	Vrais positifs	Faux négatifs
	Classe non-OGM	Faux positifs	Vrais négatifs

Tableau 7 – Matrice de confusion pour un modèle de classification à deux classes.

A partir de cette matrice de confusion, quatre critères de performances sont calculés pour évaluer la qualité du modèle :

— La sensibilité mesure la capacité d'un modèle à prédire correctement les CDS inserts OGM lorsqu'ils sont CDS inserts OGM dans la réalité. Elle est définie par l'équation suivante :

$$Se = \frac{VP}{VP + FN} \tag{6.2}$$

— La spécificité mesure la capacité d'un modèle à prédire correctement les CDS du génome hôte (non-OGM) lorsqu'ils sont CDS du génome hôte dans la réalité. Elle est définie par l'équation suivante :

$$Sp = \frac{VN}{VN + FP} \tag{6.3}$$

— Le taux de faux positifs mesure le taux d'erreurs lorsque le modèle prédit des CDS inserts OGM alors qu'ils sont des CDS du génome hôte (non-OGM) dans la réalité. Elle est définie par l'équation suivante :

$$T_{FP} = \frac{FP}{FP + VP} \tag{6.4}$$

— Le taux de faux négatifs mesure le taux d'erreurs lorsque le modèle prédit des CDS du génome hôte (non-OGM) alors qu'ils sont CDS inserts OGM dans la réalité. Elle est définie par l'équation suivante :

$$T_{FN} = \frac{FN}{FN + VN} \tag{6.5}$$

La meilleure méthode prédictive a ensuite été sélectionnée à l'aide de ces quatre critères de performance. Le critère de performance le plus important est le taux de faux positifs. En effet, la méthode ne doit pas prédire un CDS insert GM comme un CDS apparenté au génome hôte. Dans un second temps, la méthode choisie doit permettre d'obtenir des résultats très fiables, donc ayant une spécificité et une sensibilité élevées. Le taux de faux négatifs est moins contraignant car lorsqu'un CDS appartenant au génome hôte est prédit comme CDS GM, l'utilisateur pourra facilement l'écarter après des analyses complémentaires pour confirmer et vérifier l'exactitude des résultats.

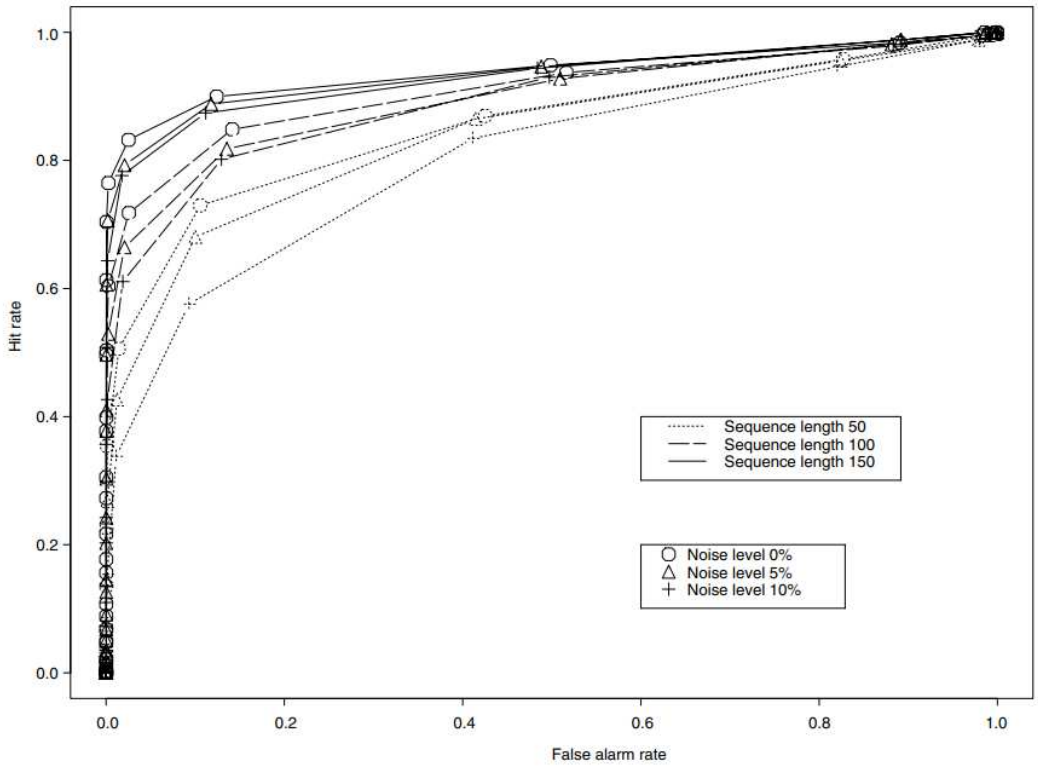


FIGURE 39 – Courbes ROC pour la détection de fragments de protéines en fonction de la longueur et du niveau de bruit des fragments [139]. L'échelle en ordonnée représente la sensibilité. L'échelle en abscisse représente 1 - spécificité.

Un autre indicateur permettant d'évaluer les performances d'un modèle est la courbe ROC (Receiver Operating Characteristic). Une courbe ROC permet de mesurer les performances d'un modèle de discrimination binaire c'est à dire une classification selon deux groupes distincts (dans notre cas, « OGM » et « non OGM »). Cette courbe trace le la sensibilité en fonction de 1 - spécificité (Exemple

sur la Figure 39). Plus l'aire sous la courbe ROC est grande, meilleurs sont les résultats.

6.2.5 Application de la validation croisée pour évaluer la qualité de modélisation et de prédiction de chaque méthode

Pour évaluer la qualité de prédiction et de modélisation de chaque modèle de discrimination, une validation croisée de type k-fold a été mise en place.

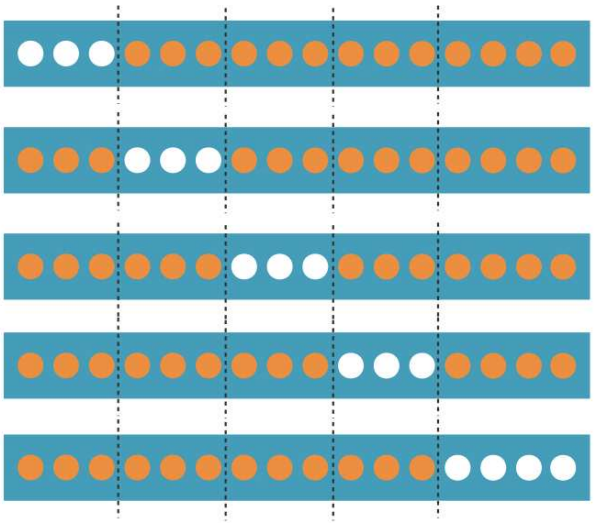


FIGURE 40 – Validation croisée 5-fold. Les données de calibration sont représentées en orange et les données de prédiction en blanc pour chacun des 5 folds [140].

Ce type de validation croisée consiste à partitionner un échantillon en k sous-échantillons de tailles similaires, à savoir un jeu de calibration composé de *k-1* sous-échantillons et un jeu de validation, le sous-échantillon k restant. Chacune des parties sera tour à tour utilisée comme données de prédiction (Figure 40). La méthode est donc appliquée k fois pour que chaque sous-échantillon soit employé une seule fois comme données de prédiction. Les données de calibration (ou d'apprentissage) évaluent la qualité de modélisation du modèle et les données de prédiction évaluent la qualité de prédiction du modèle. La validation croisée est utilisée pour éviter le sur-apprentissage [141]. Le sur-apprentissage ou overfitting signifie que le modèle est surentraîné, il modélise trop bien les données d'apprentissage entraînant généralement de mauvaises prédictions sur de nouvelles observations.

Avant d'effectuer une validation croisée, les données sont préalablement stratifiées, centrées et réduites. L'échantillonnage stratifié permet une répartition homogène des CDS inserts OGM connus dans les données échantillonnées lors de la validation croisée. La proportion d'inserts OGM est la même dans les données d'apprentissage que dans les données de prédiction. Les variables explicatives sont

centrées et réduites afin que chacune ait le même poids dans le modèle. En effet, sans ce pré-traitement, les variables explicatives de plus grandes variances (généralement celles qui sont mesurées sur des échelles plus grandes) auraient plus de poids dans la construction du modèle.

Les paramètres de chacune des douze méthodes sont optimisés suivant une première validation croisée 10-fold. Les performances des méthodes sont comparées sur la base d'une seconde validation croisée 2-fold.

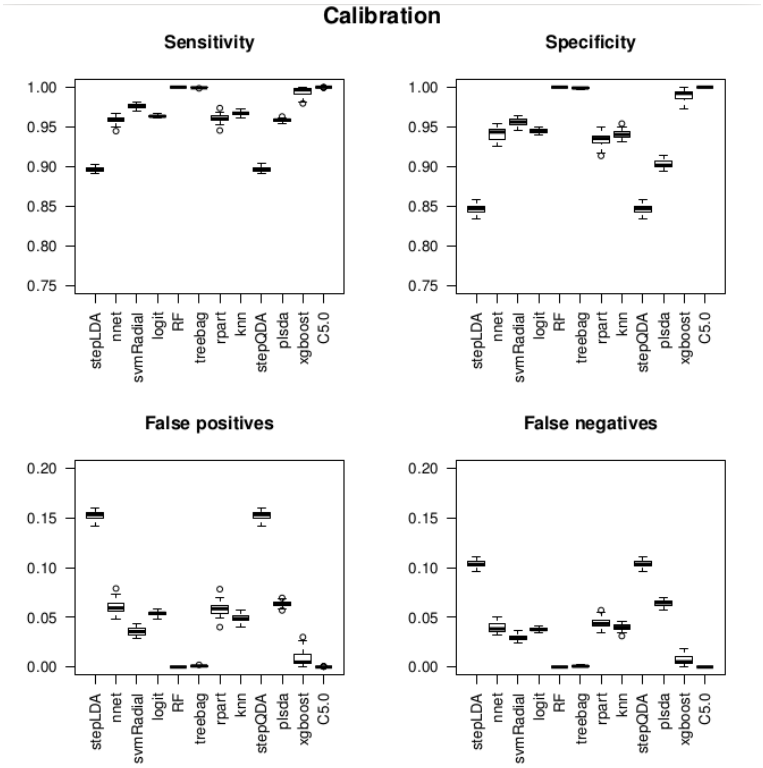
6.3 Application

6.3.1 Résultats pour chaque méthode

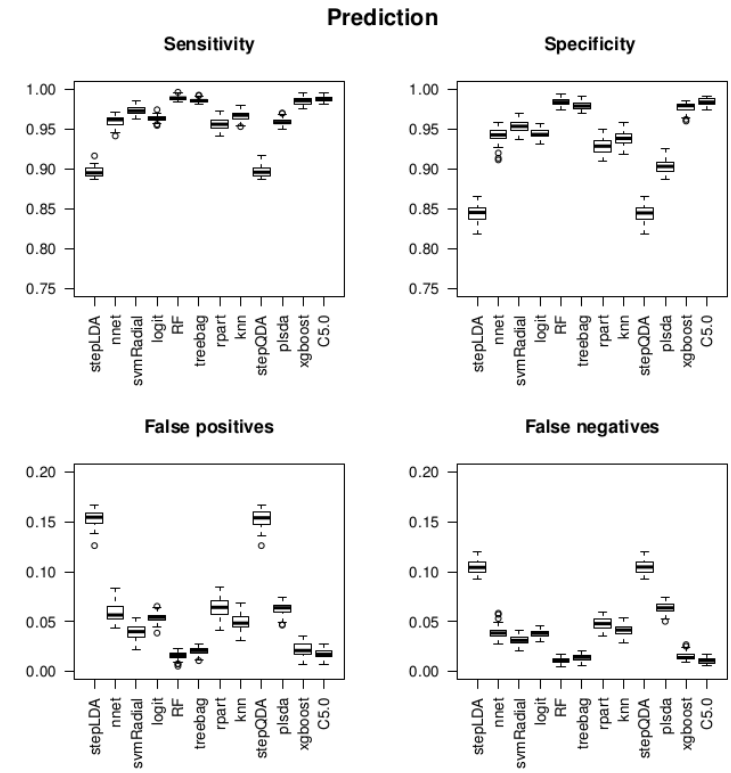
La Figure 41 illustre les résultats obtenus à partir des douze méthodes de discrimination testées 50 fois et appliquées aux données d'apprentissage de la bactérie *B. subtilis* GM. Ces redondances de processus assurent la robustesse des résultats, certaines méthodes étant non déterministes*. Les méthodes non déterministes produisent un résultat différent entre deux analyses utilisant des données d'entrées et paramètres identiques.

Les méthodes RF, C5.0, Treebag, et xgboost obtiennent les meilleurs résultats sur les données de calibration et de prédiction (Figure 41). Avec les données de calibration, les méthodes RF, Treebag et C5.0 ont une sensibilité et une spécificité de 1, un taux de faux positifs et faux négatifs à 0. La méthode xgboost obtient 0.99 en sensibilité et spécificité, un taux de faux positifs et faux négatifs à 0.01 avec les données de calibration. Avec les données de prédiction, les méthodes RF et C5.0 ont une sensibilité et une spécificité de 0.98 avec un taux de faux positifs à 0.015 et de faux négatifs à 0.01. Les méthodes Treebag et xgboost ont une sensibilité et une spécificité à 0.97 avec un taux de faux positifs et faux négatifs à 0.015. Avec les données de calibration, la méthode Logit obtient une sensibilité à 0.95, une spécificité à 0.95, un taux de faux positifs à 0.05 et un taux de faux négatifs à 0.04. Avec les données de prédiction, la méthode Logit obtient une sensibilité à 0.95, une spécificité à 0.94, un taux de faux positifs à 0.05 et un taux de faux négatifs à 0.04. D'après les critères de performances sélectionnées (partie 6.2.3), la méthode RF est privilégiée. En effet, elle présente le taux de faux positifs le plus faible des douze méthodes ainsi qu'une sensibilité et spécificité élevée avec les données de prédiction.

Pour chaque méthode, la courbe ROC (graphique représentant la sensibilité et la spécificité d'un modèle) et l'importance de l'utilisation de chaque variable parmi les neuf variables explicatives disponibles ont également été calculées.



(a) Données de calibration



(b) Données de prédiction

FIGURE 41 – Résultats des 12 méthodes de calibration (a) et de prédiction (b) sur 50 simulations avec les données d'apprentissage de la bactérie *B. subtilis* GM suivant les quatre critères de performance : la sensibilité, la spécificité, le taux de faux positifs et le taux de faux négatifs.

La Figure 42 présente l'importance des neuf variables explicatives obtenues à partir des douze méthodes de discrimination testées 50 fois et appliquées aux données d'apprentissage de la bactérie *B. subtilis* GM.

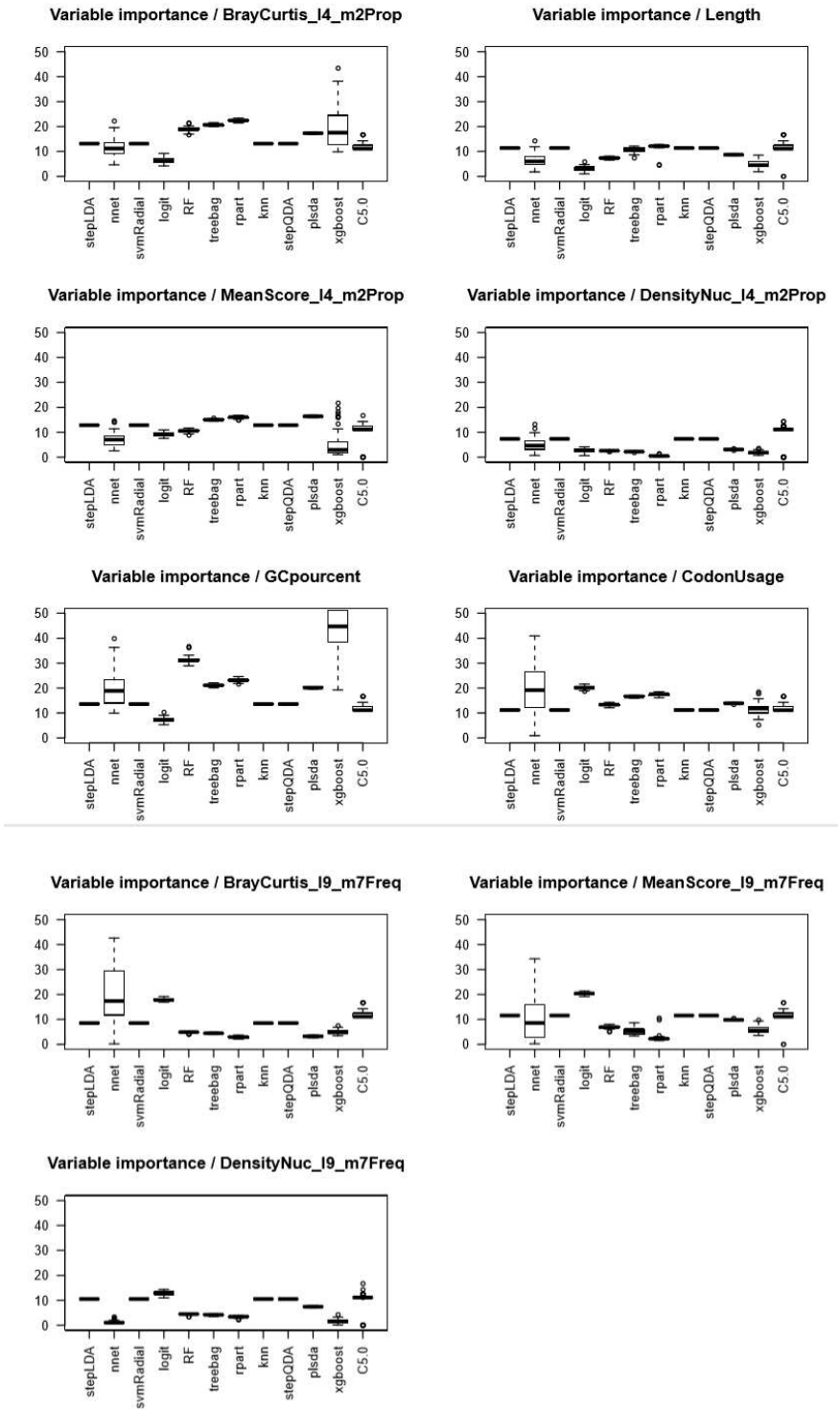


FIGURE 42 – L'importance des 9 variables utilisées pour chacune des méthodes. La variable CondonUsage correspond à la distance de Bray-Curtis avec les paramétrages P L3M1.

Les méthodes qui donnent de l'importance aux calculs de distance de Bray-Curtis avec les paramétrages P L3M1, P L4M2 et F L9M7 sont les modèles Logit et rpart. Avec la méthode RF, les variables les plus importantes pour la prédiction des CDS inserts GM sont le pourcentage en GC et la distance de Bray-Curtis en proportions L4M2. Avec la méthode Logit, les variables les plus importantes sont les distances de Bray-Curtis en fréquences L9M7 et en proportions L3M1 (Figure 42). La méthode Logit est donc privilégiée car elle utilise préférentiellement deux types de distances pour prédire les CDS OGM. De plus, la méthode Logit fait partie des méthodes paramétriques contrairement à la méthode RF qui appartient aux méthodes non paramétriques basées sur des arbres de décision.

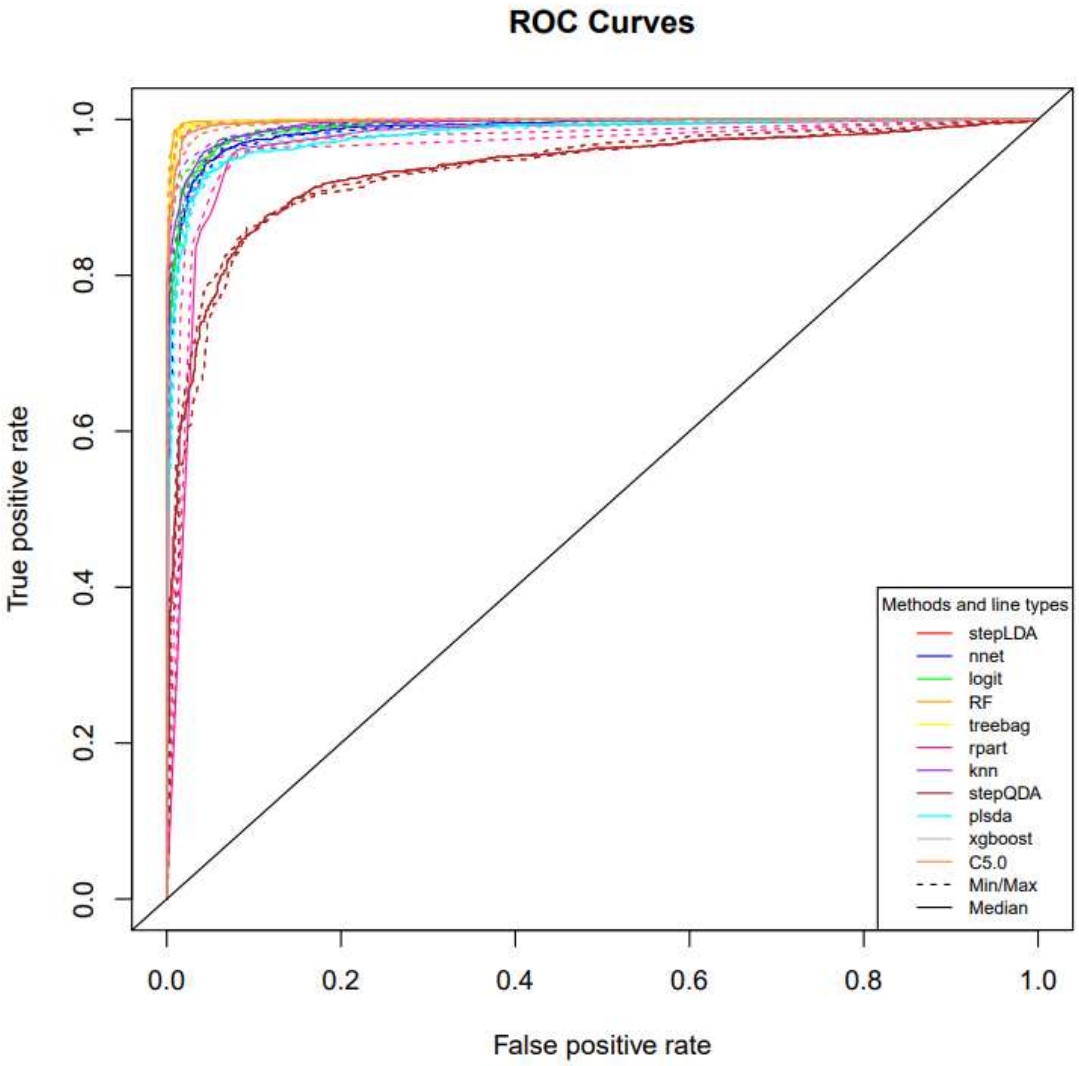


FIGURE 43 – Courbes ROC moyennes pour chacune des 12 méthodes.

Sur la Figure 43, les courbes ROC des méthodes RF et Treebag possèdent les meilleurs résultats par rapport aux autres méthodes.

6.3.2 Résultats pour deux méthodes combinées

Nos attentes sur les performances du modèle de prédiction étant très exigeantes : le taux de faux positifs et faux négatifs doit être égal à 0 et la spécificité et la sensibilité à 1, une méthode seule n'obtient pas de résultats suffisamment performants. Deux méthodes ont donc été combinées pour répondre à nos attentes.

Conformément à nos critères de performance, l'union non redondante des résultats positifs des méthodes RF et Logit a été retenue pour établir un modèle de prédiction. En effet, aucune méthode ne parvenait à approcher des critères de performances recherchés contrairement à l'utilisation des deux méthodes Logit et RF.

	Résultats pour les données de calibration		
	Logit	**RF**	**Union RF et Logit**
Faux négatifs	0.04	0	0
Spécificité	0.94	1	1
Sensibilité	0.95	1	1
Faux positifs	0.05	0	0.05

	Résultats pour les données de prédiction		
	Logit	**RF**	**Union RF et Logit**
Faux négatifs	0.04	0.01	0.01
Spécificité	0.94	0.98	0.99
Sensibilité	0.95	0.98	0.99
Faux positifs	0.05	0.01	0.09

Tableau 8 – Comparaison des résultats médians des données de calibration et de prédiction pour les méthodes RF, Logit et l'union de ces deux méthodes, réalisée sur 50 simulations en utilisant les données d'apprentissage de la bactérie *B. subtilis* GM.

La sensibilité et la spécificité globales de l'union des deux méthodes ont été déduites (Tableau 8, colonne Union RF et Logit). Les trois premiers critères de performance souhaités (le taux de faux négatifs, la spécificité et la sensibilité) sur les données de prédiction obtiennent des valeurs plus élevées ou similaires avec l'union des méthodes RF et Logit comparé aux résultats de chacune de ces deux méthodes. Cependant, le taux de faux positifs est plus élevé avec l'union des méthodes RF et Logit. Dans la réalité, l'utilisation du modèle composé de l'union des deux méthodes avec la bactérie *B. subtilis* GM permet d'obtenir un nombre de CDS inserts GM plus important qu'avec la méthode RF seule. En effet, la

méthode RF seule est plus restrictive et des CDS inserts GM sont prédits comme CDS apparentés au génome hôte (non-OGM).

Avec la bactérie *B. subtilis* GM, l'union des méthodes RF et Logit sur les données de prédiction obtient une sensibilité et une spécificité de 0.99, le taux de faux positifs est de 0.09 et le taux de faux négatifs est de 0.01. La méthode RF seule sur ces mêmes données obtient une sensibilité et une spécificité de 0.98, un taux de positifs et de faux négatifs de 0.01. La méthode Logit seule sur ces mêmes données obtient une sensibilité de 0.95, une spécificité de 0.94, un taux de faux positifs de 0.05 et un taux de faux négatifs de 0.04.

Les variables de calcul de distance (P L4M2, P L3M1 et F L9M7) font partie des variables les plus utilisées par les méthodes RF et Logit pour prédire les CDS inserts OGM. Ces deux méthodes utilisent des variables différentes (Figure 42) pour prédire correctement les CDS potentiellement OGM et n'appartiennent pas à la même famille de méthodes. La combinaison des résultats des méthodes RF et Logit apporte une valeur ajoutée pour l'identification précise d'éventuels inserts GM.

6.4 Conclusion

Avec l'union des résultats non redondants des méthodes RF et Logit, les résultats estimés avec les données de prédiction sont très proches de ceux attendus. En effet, sur les données de prédiction de la bactérie *B. subtilis* GM, le taux de faux négatif et le taux de faux positifs sont proches de 0. La spécificité et la sensibilité sur ces mêmes données sont proches de 1. Ce modèle de prédiction doit donc être testé sur d'autres jeux de données bactériens pour être validé. Cette étape de validation est décrite dans le chapitre suivant (Chapitre 7 partie 7.3).

Chapitre 7

Proposition de l'outil DUGMO

Sommaire

La méthode DUGMO pour l'acronyme Detection of Unkwown Genetically Modified Organisms est détaillé dans ce chapitre. Les résultats obtenus pour 48 génomes bactériens ainsi que les limites de détection et la rapidité d'exécution sont présentés.

7.1 Objectif

L'objectif de la méthode DUGMO, proposé dans le cadre de ce projet de recherche, est de détecter, à partir de données de séquençage Illumina, des bactéries pures génétiquement modifiées non décrites (Chapitre 1 partie 1.5). Les données multigénomiques ou métagénomiques ne sont pas prises en compte par DUGMO (comme un échantillon provenant d'un aliment OGM par exemple).

7.2 Méthode

En données d'entrée, DUGMO nécessite en plus des données de séquençage brutes, un génome de référence, un pangénome de l'espèce correspondant à la bactérie hôte et les CDS de ce même pangénome (Chapitre 2 partie 2.2.). Le pangénome doit être fiable et constitué uniquement de souches sauvages.

Dans un premier temps, DUGMO utilise un pipeline de nettoyage des données de séquençage à haut débit et une combinaison d'alignements BLASTN sur les pangénomes pour réaliser un tri et obtenir deux jeux de données distincts : les CDS apparentés au génome hôte et les CDS potentiellement GM. (Chapitre 4 parties 4.1.1, 4.2.2) Ensuite, une banque de données d'OGM connus est filtrée par rapport aux CDS du génome hôte (Chapitre 4 partie 4.1.3). Dans un second temps, différents calculs de distance de Bray-Curtis sont utilisés pour comparer les CDS apparentés au génome hôte et les CDS GM (Chapitre 5 partie 5.2.1.3). Enfin, la création d'un modèle de prédiction (Chapitre 6 partie 6.4) est réalisée à l'aide de l'union des méthodes forêts aléatoires et modèle linéaire généralisé. Cette combinaison a été sélectionnée comme étant la plus prédictive pour les données sur les bactéries GM.

7.3 Application

Les données de séquençage de trois génomes GM de la bactérie *E. coli* (Chapitre 2 partie 2.3.3.1) sont utilisées pour valider et tester la méthode. Les tests sont aussi menés sur les génomes sauvages des bactéries *C. jejuni*, *L. lactis*, *L. monocytogenes*, *M. tuberculosis*, *S. aureus*, *S. Typhimurium* et leurs données synthétiques GM associées (Chapitre 2 partie 2.3.4).

Les génomes de référence associés à chaque bactérie utilisée pour ces analyses, sont présentés dans le Chapitre 2 partie 2.3.4.1. Les différents pangénomes bactériens nécessaires à ces analyses ont été créés (Chapitre 2 partie 2.3.4.2). Enfin, une banque de données de CDS présents dans des inserts OGM de différentes espèces a été fournie par le JRC [56] et complétée par des CDS d'inserts GM bactériens issus ou déduits de la littérature.

7.3.1 Resultats pour la bactérie *B. subtilis*

Le génome de *B. subtilis* GM contient quatre plasmides de pGMBsub01 à pGMBsub04. Récemment, Berbers *et al.* [84] ont décrit une révision de ce génome avec un seul plasmide pGMrib (comprenant pGMBsub03 et pGMBsub04) et une insertion chromosomique de 53 kb (comprenant pGMBsub01 et pGMBsub02) (Chapitre 2 partie 2.3.2.1). Nos analyses ont été réalisées sur les données de Paracchini *et al.* [82]. Dans les résultats suivants, les noms de gènes fournis dans Paracchini *et al.* sont conservés et les noms de gènes ou locus apparentés trouvés dans les assemblages de Berbers *et al.* sont indiqués entre parenthèses lorsqu'ils sont différents.

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs max (3) (4)	Vrais négatifs (3)
B. subtilis GM	4102	2714	39	25	0	12	2
B. subtilis sauvage	3941	2724	4	-	0	-	4

Tableau 9 – Résultats de DUGMO pour les génomes de *B. subtilis*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (4) Estimation du nombre maximal de faux négatifs : le nombre réel ne peut être déduit en raison de l'origine inconnue des CDS, potentiellement de la famille des Bacillales. (-) Ne s'applique pas.

Les résultats présentés dans le Tableau 9 concernent les deux jeux de données de la bactérie *B. subtilis*. Après les étapes de nettoyage et de tri des données de séquençage de la bactérie *B. subtilis* GM, 4102 CDS apparentés au génome hôte et 39 CDS GM candidats sont obtenus. A la fin du filtrage de la banque de données d'OGM connus, 2714 CDS inserts OGM connus sont conservés. Pour caractériser les 39 CDS inserts potentiels à l'aide du modèle de prédiction (définir si ces CDS sont OGM ou non-OGM), les données d'apprentissage sont donc constituées de 4102 CDS hôtes (non-OGM) et 2714 CDS inserts OGM. Sur les 39 CDS inserts potentiels à prédire, l'union non redondante des résultats des méthodes RF et Logit prédit 25 CDS GM (vrais positifs), 12 CDS GM mal prédits (faux négatifs) et 2 CDS du génome hôte (non-OGM) (vrais négatifs). Les 25 CDS inserts de la bactérie *B. subtilis* GM, correctement prédits par la méthode (vrais positifs), correspondent aux principaux gènes situés sur les plasmides (riboflavine) et au gène conférant la résistance au chloramphénicol (Chapitre 2 partie 2.3.2.1). Les 12 CDS inserts mal prédits représentent le nombre maximal de faux négatifs : le nombre réel ne peut être déduit en raison de l'origine inconnue des CDS, potentiellement de la famille des Bacillales (Tableau 9, annotation (4)).

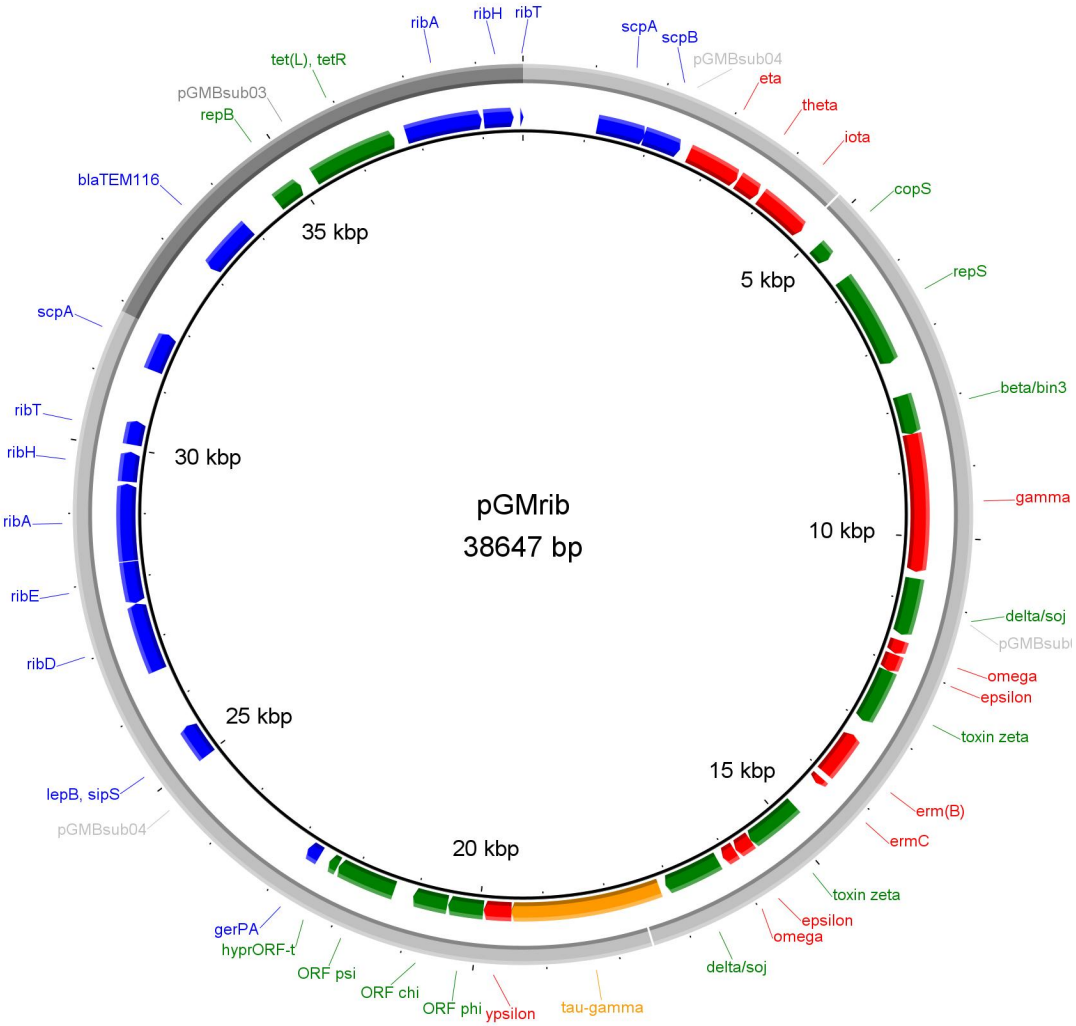


FIGURE 44 – Résultats de DUGMO sur le plasmide pGMrib de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

Pour la bactérie sauvage *B. subtilis*, les données d'apprentissage sont constituées de 2724 CDS inserts OGM retenus à la fin du filtrage de la banque de données d'OGM connus et de 3941 CDS hôtes (non-OGM) retenus après les étapes de nettoyage et de tri des données de séquençage. Après ces étapes, 4 CDS sont potentiellement GM. Sur ces CDS, aucun n'est prédit comme insert GM par le modèle de discrimination (dans le Tableau 9, ligne *B. subtilis* sauvage et la colonne Résultats DUGMO).

Concernant la bactérie *B. subtilis* GM, les portions *cat* et *recA* sur le chromosome (Chapitre 2 partie 2.3.2.1) sont toutes deux correctement détectées comme CDS inserts GM par DUGMO. Deux gènes provenant de *B. subtilis* (vrais négatifs) sont écartés par l'étape de prédiction, la sous unité delta de l'ARN polymérase chromosomique et *scpB*. DUGMO distingue la séquence chromosomique sauvage *scpB* de celle du plasmide pGMrib. Le plasmide pGMrib (Figure 44) dé-

crit par Berbers *et al.* comprend deux parties, la première a un opéron ribDEATH de *B. subtilis* reconnu et écarté des résultats de DUGMO (comprenant *lepB* (*sipS*) et *gerPA* (*GAY71_RS22375*)), la seconde contient des gènes de bacille conférant une résistance à l'érythromycine. Dans cette deuxième partie, neuf gènes sont reconnus comme GM par DUGMO (*repS* , *beta* (*bin3*), *delta* (*soj*), *zeta*, *copS*, *ORF psi*, *ORF phi*, *ORF chi*, et *hyprORF-t*, tandis que les dix autres sont écartés (*eta*, *theta*, *iota*, *gamma* (*topB*, 2 sont affichés sur Berbers *et al.* [84] Figure C, la seconde est une petite fraction de *topB*), *omega*, *epsilon*, *erm(B)*, *ermC*, *tau-gamma* et *ypsilon*). Parmi eux, les gènes *ermC* et *tau-gamma* sont écartés lors des alignements BLASTN sur le pangénome. En effet, ces gènes ont une séquence nucléotidique très proche des CDS de *B. subtilis* et sont inclus dans les données d'apprentissage des CDS de l'hôte lors de l'étape de construction d'un modèle de prédiction. La plupart des autres CDS écartés correspondent à des gènes d'entérocoques, qui sont des bacilles gram+ phylogénétiquement proches de *B. subtilis*.

Une partie du plasmide pGMrib (pGMBsub04 de Paracchini *et al.* [82]) proviendrait du plasmide breveté pMX45 pour lequel seules les origines de *B. subtilis* et les méthodes de génie génétique sont mentionnées [82]. Cependant, certains plasmides possèdent une gamme d'hôtes très étendue et il n'est pas toujours possible d'affirmer avec certitude leur origine. Dans ce cas, la distance génétique de ces onze gènes, s'ils ne proviennent pas de *B. subtilis*, ne permet pas leur discrimination. De plus, les résultats de DUGMO présentent un CDS supplémentaire, non décrit par Paracchini *et al.* [82] dans le plasmide pGMBsub04 (Figure 45). Ce CDS est trouvé deux fois par DUGMO, une fois sur le plasmide et une fois sur le chromosome. Il est présent aux positions 1 à 171 dans le plasmide pGMBsub04 (2652 à 2822 dans pGMrib) et correspond à une protéine hypothétique provenant d'un plasmide de *Enterococcus faecalis* (numéro d'accession AP018546.1).

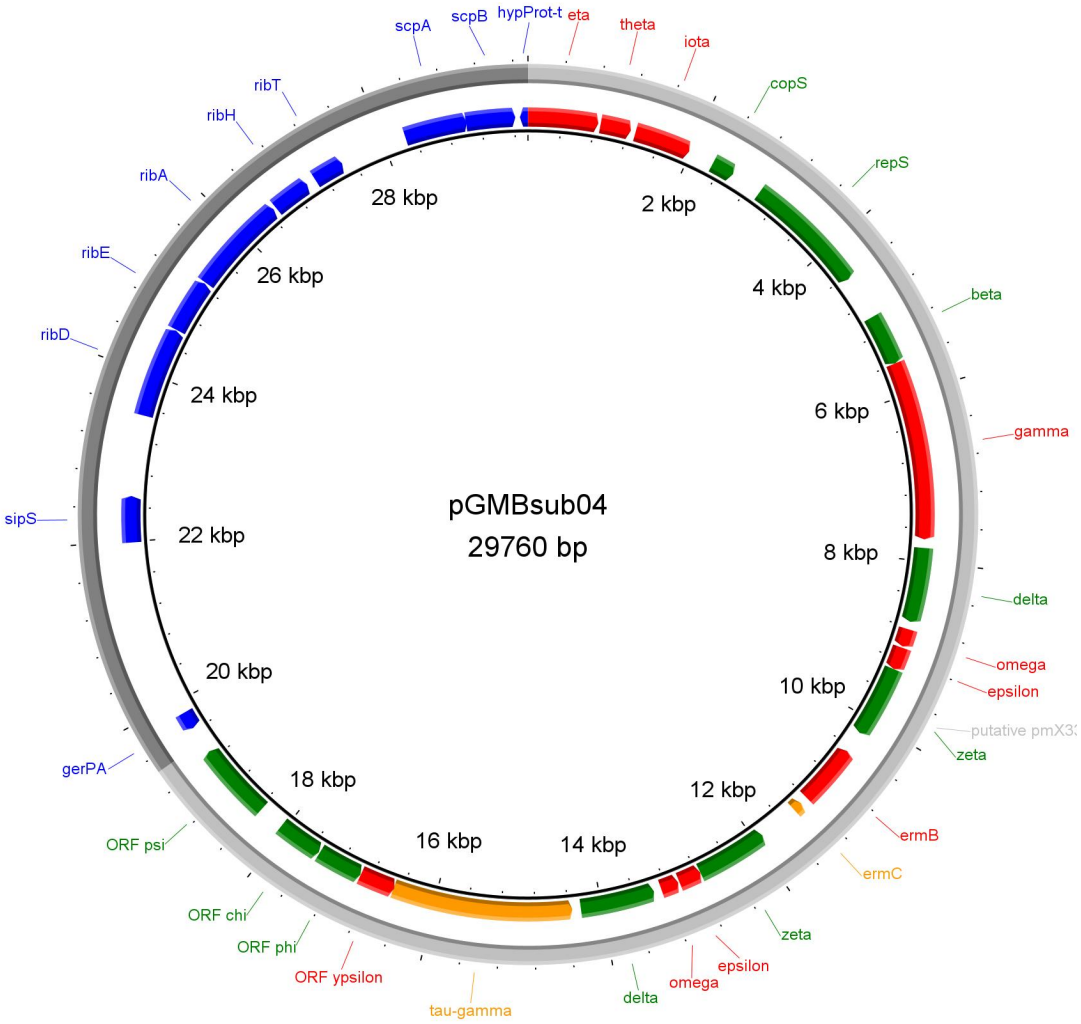


FIGURE 45 – Résultats de DUGMO sur le plasmide pGMBsub04 de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

Le plasmide pGMBsub03 de Paracchini *et al.* (Figure 46) possède uniquement deux gènes qui n'ont pas pour origine *B. subtilis*, *repB* (inclus dans *GAY71_RS22445*) et *tetR* (version abrégée de *tet(L)*) trouvés GM dans les résultats de DUGMO.

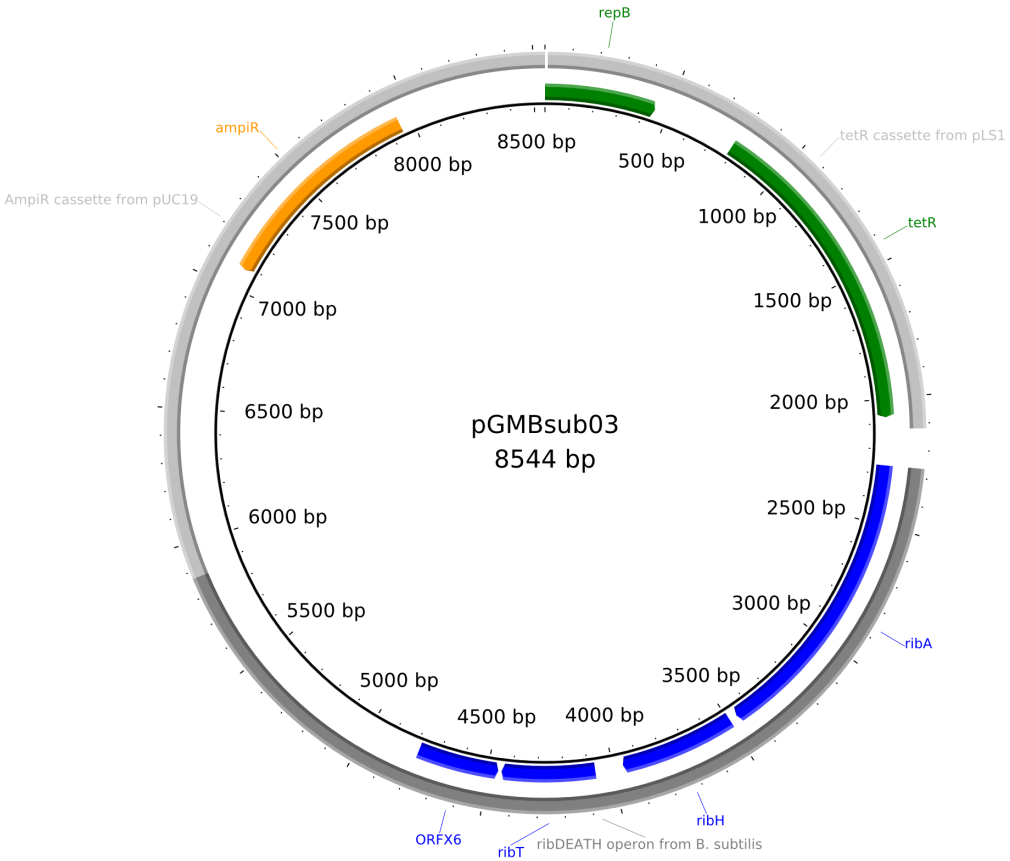


FIGURE 46 – Résultats de DUGMO sur le plasmide pGMBsub03 de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

Le plasmide pGMBsub02 (Figure 49) est une version tronquée du plasmide pGMBsub01 (Figure 47). L'insertion chromosomique de 53 kb (pGMBsub01 de Paracchini *et al.*) porte des gènes de différentes origines : *Bacillus amyloliquefaciens* et *Staphylococcus aureus*. Les gènes de *S. aureus bleR* (*GAY71_RS12530*) et *kanR* (*aadD1*) et les CDS *ribD*, *ribA*, *ribT* et *scpB* de l'opéron ribDEATH de *B. amyloliquefaciens* sont correctement détectés comme CDS GM par DUGMO. Les CDS *ampR* (*bêta lactamase TEM-116* dans le chromosome, *GAY71_RS22440* dans le plasmide pGMrib), *ribH* et *scpA* s'alignent sur les CDS du pangénome de *B. subtilis* dans le respect des paramètres BLASTN et sont donc inclus dans les données d'apprentissage des CDS de l'hôte, montrant les limites de la catégorisation des CDS pour les espèces étroitement apparentées. Cette erreur provoque une classification erronée de certains CDS, ce qui pourrait probablement expliquer que les CDS *rep* et *ribE* ne soient pas détectés par DUGMO. Enfin, l'annotation Prokka fournit deux CDS hypothétiques supplémentaires, correspondant aux postions 4759-5103 et 10112-10201 de l'opéron ribDEATH de *B. amyloliquefaciens* dans pGMBsub01 (loci *GAY71_RS12515* dans le chromosome fourni dans Berbers *et al.*). Le premier CDS est étiqueté CDS GM alors que le second non.

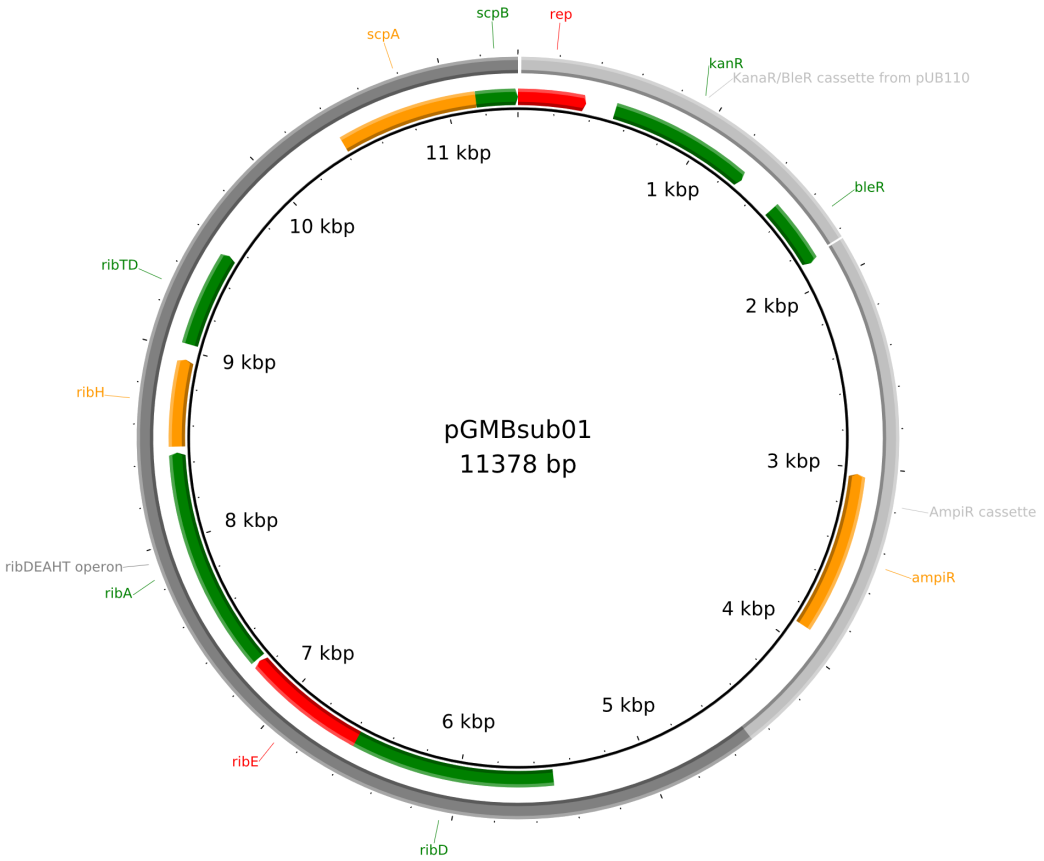


FIGURE 47 – Résultats de DUGMO sur le plasmide pGMBsub01 de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

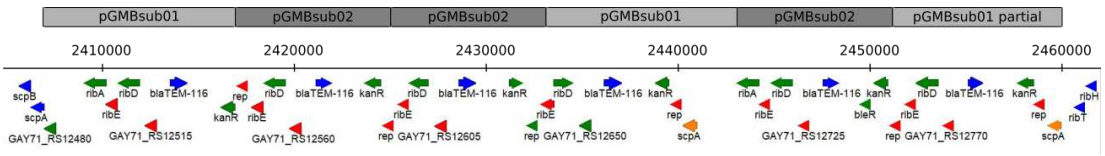


FIGURE 48 – Résultats de DUGMO sur l'insertion chromosomique de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

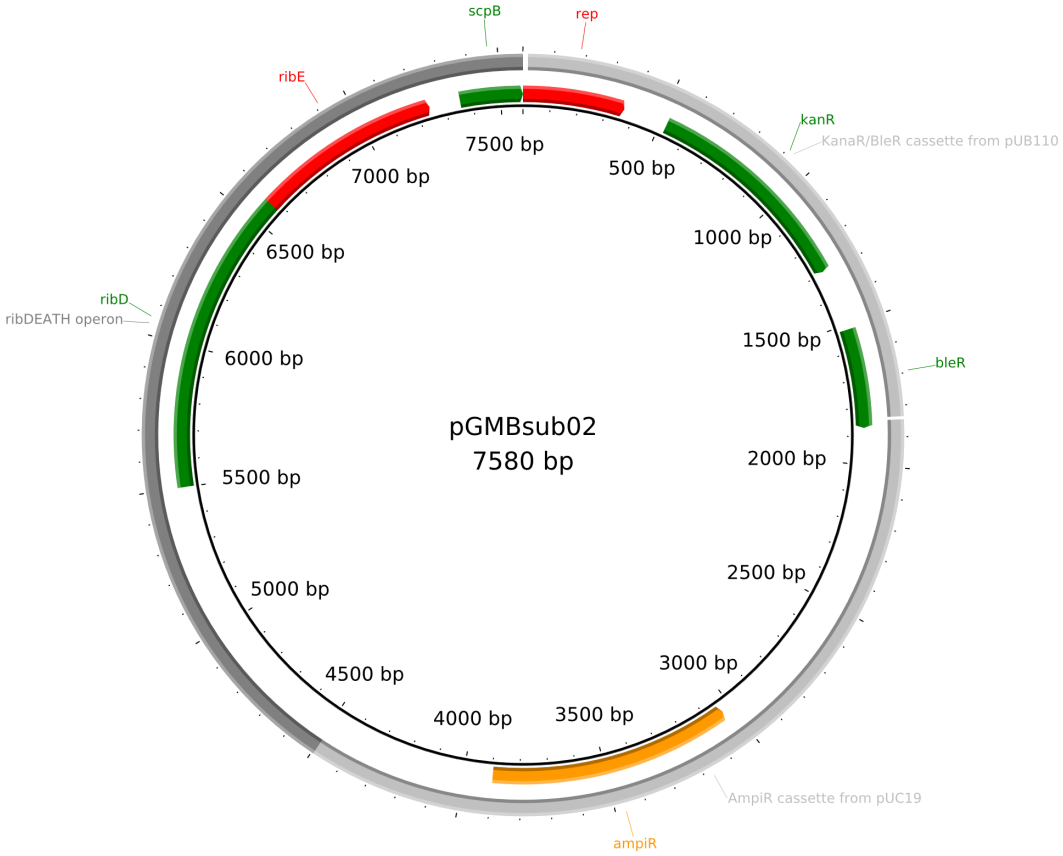


FIGURE 49 – Résultats de DUGMO sur le plasmide pGMBsub02 de la bactérie *B. subtilis* GM. Les gènes correctement prédits par DUGMO (vrais positifs) sont en vert. Les gènes mal prédits par DUGMO (faux négatifs) sont en rouge. Les gènes appartenant à *B. subtilis* sont en bleu. Les gènes s'alignant sur le pangénome de *B. subtilis* lors de l'alignement BLASTN sont en orange.

7.3.2 Résultats pour la bactérie *E. coli*

Le Tableau 10 présente les résultats de l'outil DUGMO obtenus avec les trois génomes de la bactérie *E. coli* (Chapitre 2 partie 2.3.3.1). Avec les bactéries *E. coli* GM portant les gènes de *A. tumefaciens*, *M. tuberculosis* ou *S. pyogenes*, les résultats contenaient respectivement 5, 4 et 5 CDS prédits comme inserts GM (vrais positifs). Cependant, un faux positif a été trouvé dans les résultats de *E. coli* portant des gènes de *S. pyogenes*. Le CDS trouvé correspond à un gène *aslA* annoté présent dans *E. coli* fortement tronqué dans notre souche (984 nt au lieu de 1650 nt ou plus pour ceux du pangénome). Dans le cas de l'incorporation d'une protéine fortement tronquée provenant d'une famille de protéines mal représentée dans le génome de l'hôte, cette séquence peut être considérée comme un insert GM. DUGMO permet de cibler les CDS de l'insert si le génome inconnu soumis est en fait une bactérie GM.

Sur la souche de *E. coli* portant les gènes de *S. pyogenes*, DUGMO a trouvé comme faux positif un CDS qui correspond au gène *arlS* fortement tronqué codant pour la protéine arylsulfatase. La version complète de ce gène est naturellement

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
E. coli avec un gène de *A. tumefaciens*	4033	2588	6	5	0	0	1
E. coli avec un gène de *M. tuberculosis*	4018	2589	5	4	0	0	1
E. coli avec un gène de *S. pyogenes*	4015	2587	6	5	1	0	0

Tableau 10 – Résultats de DUGMO pour les génomes de *E. coli*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ».

présente dans le génome de *E. coli*. Lors de l'analyse, l'étape de l'alignement BLASTN sur les CDS du pangénome vérifie que les longueurs des CDS correspondants ne varient pas de plus de 15%. Pour les deux autres génomes de *E. coli*, porteurs des gènes de *A. tumefaciens* ou de *M. tuberculosis*, le CDS *arlS* est de pleine longueur dans les assemblages et n'est pas détecté par DUGMO. Le CDS *arlS* tronqué dans le génome de *E. coli* portant les gènes de *S. pyogenes* est localisé à l'extrémité d'un contig (assemblage de reads) de l'assemblage de novo, et la partie manquante, 15% du gène, est retrouvée dans un autre contig. DUGMO est donc capable de détecter des CDS qui peuvent consister en une longue délétion ou insertion dans un CDS sauvage. La moyenne de la profondeur de couverture des données pour cet échantillon de *E. coli* portant des gènes de *S. pyogenes* est de 41, tandis que la moyenne de la profondeur de couverture des deux autres souches de *E. coli* portant les gènes de *A. tumefaciens* ou de *M. tuberculosis* est de 63 et 68, respectivement. Cet exemple indique que l'assemblage avec une profondeur de couverture moyenne de 41 est partiel et conduit à un faux positif, en raison de la troncature d'un CDS. Au contraire, les assemblages de données ayant une profondeur de couverture moyenne supérieure à 60 ne présentent pas de problème de troncature des gènes dans nos ensembles de données. La plupart des assembleurs de bactéries recommandent une profondeur de couverture comprise entre 80 [122] et 100. Il a été démontré que le N50* augmente avec la profondeur de couverture jusqu'à une valeur de 100 pour les génomes bactériens [142]. Une profondeur de couverture de 60 est potentiellement la limite inférieure acceptable pour obtenir un assemblage complet sans troncation de CDS. Ces observations

soulignent la nécessité d'obtenir un bon assemblage, et donc une profondeur de couverture suffisante pour obtenir des résultats satisfaisants avec DUGMO.

7.3.3 Résultats pour la bactérie *C. jejuni*

Le Tableau 11 présente les résultats de DUGMO obtenus pour la bactérie *C. jejuni* (Chapitre 2 partie 2.3.4.1).

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
C. jejuni sauvage	1561	2794	26	-	0	-	26
C. jejuni avec un gène de *L. lactis*	1441	2794	1	1	0	0	0
C. jejuni avec un gène de *L. monocytogenes*	1344	2794	1	1	0	0	0
C. jejuni avec un gène de *S. aureus*	1391	2794	1	1	0	0	0
C. jejuni avec un gène de *M. tuberculosis*	1407	2794	1	1	0	0	0
C. jejuni avec un gène de *S. Typhimurium*	1434	2794	1	1	0	0	0
C. jejuni avec un gène de *O. sativa*	1383	2794	1	1	0	0	0
C. jejuni avec un gène de *H. sapiens*	1385	2794	1	1	0	0	0

Tableau 11 – Résultats de DUGMO pour les génomes de *C. jejuni*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

Aucun CDS n'a été détecté comme insert dans les résultats de DUGMO (aucun faux positif ou faux négatif) sur le génome de *C. jejuni* sauvage et sur les

génomes synthétiques à l'exception du gène introduit (un vrai positif).

7.3.4 Résultats pour la bactérie *L. lactis*

Le Tableau 12 présente les résultats de DUGMO obtenus pour la bactérie *L. lactis* (Chapitre 2 partie 2.3.4.1).

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
L. lactis sauvage	1358	2786	0	-	0	-	0
L. lactis avec un gène de *C. jejuni*	1928	2784	1	1	0	0	0
L. lactis avec un gène de *L. monocytogenes*	1924	2779	1	1	0	0	0
L. lactis avec un gène de *S. aureus*	1904	2780	1	1	0	0	0
L. lactis avec un gène de *M. tuberculosis*	1916	2783	1	1	0	0	0
L.lactis avec un gène de *S. Typhimurium*	1951	2778	1	1	0	0	0
L. lactis avec un gène de *O. sativa*	1933	2779	1	1	0	0	0
L. lactis avec un gène de *H. sapiens*	1940	2779	1	1	0	0	0

Tableau 12 – Résultats de DUGMO pour les génomes de *L. lactis*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

Aucun CDS n'a été détecté comme insert dans les résultats de DUGMO (aucun faux positif ou faux négatif) sur le génome de *L. lactis* sauvage et sur les

génomes synthétiques à l'exception du gène introduit (un vrai positif). Cependant, pour l'analyse du génome de *L. lactis* sauvage, la profondeur minimale de couverture après assemblage a été modifiée à 50 (option « mincov » dans l'assembleur Shovill). En effet, la présence d'une contamination de poisson a été détectée dans ces données de séquençage.

7.3.5 Résultats pour la bactérie *L. monocytogenes*

Le Tableau 13 présente les résultats de DUGMO obtenus pour la bactérie *L. monocytogenes* (Chapitre 2 partie 2.3.4.1).

Aucun CDS n'a été détecté comme insert dans les résultats de DUGMO (aucun faux positif ou faux négatif) sur le génome de *L. monocytogenes* sauvage et sur les génomes synthétiques à l'exception du gène introduit (un vrai positif). Lors de la création du génome synthétique (Chapitre 2 partie 2.2.3) de *L. monocytogenes* avec le gène de *H. sapiens*, l'option de profondeur de couverture du logiciel ART (générateur de reads aléatoires) a été modifiée à -f18. La profondeur de couverture choisie dans le logiciel ART (option -f11) n'était pas suffisante pour compléter l'étape d'alignement avec l'aligneur BWA lors du pipeline de nettoyage.

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
L. monocytogenes sauvage	2753	2778	0	-	0	-	0
L. monocytogenes avec un gène de *L. lactis*	2495	2782	1	1	0	0	0
L. monocytogenes avec un gène de *C. jejuni*	2525	2778	1	1	0	0	0
L. monocytogenes avec un gène de *S. aureus*	2491	2780	1	1	0	0	0
L. monocytogenes avec un gène de *M. tuberculosis*	2547	2780	1	1	0	0	0
L. monocytogenes avec un gène de *S. Typhimurium*	2397	2782	1	1	0	0	0
L. monocytogenes avec un gène de *O. sativa*	2463	2779	1	1	0	0	0
L. monocytogenes avec un gène de *H. sapiens*	2788	2782	1	1	0	0	0

Tableau 13 – Résultats de DUGMO pour les génomes de *L. monocytogenes*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

7.3.6 Résultats pour la bactérie *M. tuberculosis*

Le Tableau 14 présente les résultats de DUGMO obtenus avec la bactérie *M. tuberculosis* (Chapitre 2 partie 2.3.4.1).

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
M. tuberculosis sauvage	4001	2809	14	-	1	-	13
M. tuberculosis avec un gène de *L. lactis*	3695	2808	1	1	0	0	0
M. tuberculosis avec un gène de *L. monocytogenes*	3646	2807	1	1	0	0	0
M. tuberculosis avec un gène de *S. aureus*	3593	2808	2	2	0	0	0
M. tuberculosis avec un gène de *C. jejuni*	3657	2808	1	1	0	0	0
M. tuberculosis avec un gène de *S. Typhimurium*	3699	2806	1	1	0	0	0
M. tuberculosis avec un gène de *O. sativa*	3676	2808	1	1	0	0	0
M. tuberculosis avec un gène de *H. sapiens*	3695	2807	1	1	0	0	0

Tableau 14 – Résultats de DUGMO pour les génomes de *M. tuberculosis*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

Dans les résultats de DUGMO sur le génome de *M. tuberculosis* sauvage, un CDS a été trouvé comme insert GM (un faux positif dans la première ligne du Tableau14). Ce CDS présente un pourcentage d'identité inférieur à 95% lors de l'alignement BLASTN sur le pangénome entier de *M. tuberculosis*. Notre hypo-

thèse est que ce CDS pourrait provenir d'un transfert horizontal de gène d'une bactérie non identifiée. Un alignement MEGABLAST [143] de ce CDS a donc été effectué sur la base de données WGS (échantillons environnementaux exhaustifs de la banque de données NCBI) sur tous les actinomycètes (ordre), mais aucun meilleur résultat n'a été fourni. Par conséquent, l'origine de ce CDS n'a pas pu être déterminée. L'aligneur MEGABLAST est utilisé pour trouver des séquences très similaires (par exemple, intra-espèces ou espèces proches). Lors de la création du génome synthétique (Chapitre 2 partie 2.2.3) de *M. tuberculosis* avec le gène de *L. lactis*, l'option de profondeur de couverture du logiciel ART a été modifiée à -f18. La profondeur de couverture choisie dans le logiciel ART (option -f11) n'était pas suffisante pour compléter l'étape d'alignement avec l'aligneur BWA lors du pipeline de nettoyage.

7.3.7 Résultats pour la bactérie *S. Typhimurium*

Le Tableau 15 présente les résultats de DUGMO obtenus avec la bactérie *S. Typhimurium* (Chapitre 2 partie 2.3.4.1).

Aucun CDS n'a été détecté comme insert dans les résultats de DUGMO (aucun faux positif ou faux négatif) sur le génome de *S. Typhimurium* sauvage et sur les génomes synthétiques à l'exception du gène introduit (un vrai positif). Lors de la création du génome synthétique (Chapitre 2 partie 2.2.3) de *S. Typhimurium* avec le gène de *H. sapiens*, l'option de profondeur de couverture du logiciel ART a été modifiée à -f18. La profondeur de couverture choisie dans le logiciel ART (option -f11) n'était pas suffisante pour compléter l'étape d'alignement avec l'aligneur BWA lors du pipeline de nettoyage.

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
S. Typhimurium sauvage	4629	2574	0	-	0	-	0
S. Typhimurium avec un gène de *L. lactis*	4080	2573	2	1	0	0	1
S. Typhimurium avec un gène de *L. monocytogenes*	4065	2571	0	1	0	0	0
S. Typhimurium avec un gène de *S. aureus*	4141	2572	1	1	0	0	0
S. Typhimurium avec un gène de *M. tuberculosis*	4106	2573	1	1	0	0	0
S. Typhimurium avec un gène de *C. jejuni*	4145	2570	1	1	0	0	0
S. Typhimurium avec un gène de *O. sativa*	4089	2573	1	1	0	0	0
S. Typhimurium avec un gène de *H. sapiens*	4501	2574	1	1	0	0	0

Tableau 15 – Résultats de DUGMO pour les génomes de *S. Typhimurium*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

7.3.8 Résultats pour la bactérie *S. aureus*

Le Tableau 16 présente les résultats de DUGMO obtenus avec la bactérie *S. aureus* (Chapitre 2 partie 2.3.4.1).

	Données			**Résultats DUGMO**			
	Apprentissage		**Prédiction**				
	Nombre de CDS du génome hôte (1)	Nombre de CDS OGM connus (2)	Nombre de CDS GM potentiels (1)	Vrais positifs (3)	Faux positifs (3)	Faux négatifs (3)	Vrais négatifs (3)
S. aureus sauvage	2485	2794	2	-	0	-	2
S. aureus avec un gène de *L. lactis*	2267	2794	1	0	0	0	0
S. aureus avec un gène de *L. monocytogenes*	2204	2792	1	1	0	0	0
S. aureus avec un gène de *C. jejuni*	2216	2793	1	0	0	1	0
S. aureus avec un gène de *M. tuberculosis*	2251	2791	1	1	0	0	0
S. aureus avec un gène de *S. Typhimurium*	2190	2783	1	1	0	0	0
S. aureus avec un gène de *O. sativa*	2544	2793	1	1	0	0	0
S. aureus avec un gène de *H. sapiens*	2156	2792	1	1	0	0	0

Tableau 16 – Résultats de DUGMO pour les génomes de *S. aureus*. (1) Après deux alignements BLASTN sur des pangénomes sans ARN non messagers. (2) Après le filtrage des CDS de la banque de données des OGM connus qui sont trop proches de l'espèce hôte. (3) Dans les CDS inserts GM potentiels. La colonne « Résultats DUGMO » détaille les résultats obtenus après l'union des résultats des méthodes RF et Logit, en utilisant les données des colonnes « Données d'apprentissage » et « Données de prédiction ». (-) Ne s'applique pas.

Pour l'analyse du génome de *S. aureus* sauvage, l'option « –dust no » a été ajoutée dans l'alignement BLASTN sur l'ensemble des alignements contre les pangénomes. Cette option masque par défaut les régions de faible complexité. Un CDS a été trouvé comme faux négatif lorsque cette option était définie par défaut. Le faux négatif présent dans l'analyse de *S. aureus* avec le gène de *C. jejuni*

est dû aux méthodes RF et Logit qui prédisent cet insert GM comme un CDS sauvage. Lors de la création du génome synthétique (Chapitre 2 partie 2.2.3) de *S. aureus* avec le gène de *O. sativa*, l'option de profondeur de couverture du logiciel ART a été modifiée à -f18. La profondeur de couverture choisie dans le logiciel ART (option -f11) n'était pas suffisante pour compléter l'étape d'alignement avec l'aligneur BWA lors du pipeline de nettoyage.

Le faux négatif correspondant au gène de *C. jejuni* inséré peut s'expliquer par l'extrême plasticité du génome de *Staphylococcus aureus* et sa propension à intégrer des gènes d'autres bactéries gram positives à faible teneur en GC, comme souligné dans [144] : "Gene transfer between Staphylococci and low-GC-content gram-positive bacteria appears to have shaped their virulence and resistance profiles". Après vérification des génomes complets présents dans le pangénome de *S. aureus*, certains d'entre eux incluent des plasmides de *Staphylococcus epidermidis* ou de *Staphylococcus argentus*. De plus, ce pangénome contient probablement des gènes de transferts horizontaux récents provenant d'autres bactéries gram + à faible teneur en GC. La caractérisation du vocabulaire génomique du génome sauvage est donc moins précise pour l'étape de construction d'un modèle de prédiction, entraînant ainsi quelques faux négatifs dans les résultats de DUGMO. *S. aureus* possède probablement un pangénome « ouvert » (Chapitre 2 partie 2.2.1). Dans le cas de telles espèces, utiliser un core-génome pour caractériser les CDS de l'hôte serait peut être plus approprié que le pangénome.

7.3.9 Limites de détection et rapidité d'exécution

Les limites de la méthode ont été évaluées à l'aide de données synthétiques spécifiques. La robustesse de la méthode à l'optimisation des dicodons (deux triplets de nucléotides) a été évaluée. Dans ce but, un gène de *B. subtilis* codant pour une riboflavine synthase a été optimisé pour l'utilisation des dicodons de *E. coli*. Des données OGM synthétiques ont été générées à partir de ce gène avec la méthode de création d'OGM bactérien décrite dans le chapitre 2 partie 2.2.3. Cette procédure a été exécutée 10 fois. Dans tous les cas, DUGMO a détecté le gène optimisé de *B. subtilis* comme un insert GM, prouvant ainsi l'insensibilité de la méthode à l'optimisation des dicodons, donc à l'optimisation des codons.

Pour évaluer la proportion de substitutions nécessaire pour rendre un gène sauvage détectable par DUGMO, des substitutions ont été introduites artificiellement dans un gène sauvage de *B. subtilis*. Deux gènes sauvages ont été testés indépendamment, une séquence courte (417 nucléotides) et une séquence longue (1317 nucléotides). Le gène long de la surfactine et le gène court CadI du génome sauvage de *B. subtilis* ont subi n substitutions aléatoires, à l'exception des substitutions dans le codon start, dans les codons stop et celles qui auraient introduit un codon stop précoce dans la séquence. Ces exclusions permettent d'éviter des événements de mutation qui seraient trop facilement détectables par DUGMO. Chaque gène modifié a ensuite été remplacé dans l'assemblage de la bactérie *B. subtilis* sauvage. Enfin, les données de séquençage Illumina associées à ces génomes modifiés ont été générées à l'aide du logiciel ART. Ce processus a été répété 10 fois. Les résultats de DUGMO indiquent qu'au-delà de 9% de mutations,

un CDS muté est détecté comme GM.

Comme DUGMO utilise des propriétés statistiques liées aux troisièmes positions des CDS, spécifiques à l'utilisation des codons (Chapitre 3 partie 3.2.1), il ne peut pas détecter des bactéries GM possédant uniquement des insertions d'ARNt ou d'ARNr, gènes qui ont un usage des codons différent de celui des gènes codant pour des ARNm et protéines. Conceptuellement, DUGMO est fait pour identifier des CDS qui n'utilisent pas le vocabulaire du génome de l'hôte. Il n'est pas capable de distinguer un gène provenant d'un transfert horizontal de gène d'un gène GM.

Le processus d'analyse des bactéries GM décrit par Paracchini *et al.* [82] a nécessité plus d'un mois de travail (communication personnelle), tandis que DUGMO n'a eu besoin que de trois heures avec 10 threads d'un processeur Core i7 pourvu de 64 Go de RAM pour rendre le même résultat. La rapidité du traitement des données est un avantage majeur de cet outil. En outre, il ne nécessite pas d'expertise préalable en matière d'OGM pour les utilisateurs finaux, sauf pour valider le nombre limité de gènes GM probables détectés dans les résultats. Cependant, l'utilisateur doit posséder des compétences en ligne de commande Linux et avoir suffisamment de connaissances biologiques pour construire un pangénome.

7.4 Conclusion

L'outil DUGMO, proposé dans le cadre de cette thèse, combine un pipeline de nettoyage des données de séquençage à haut débit, des alignements BLASTN sur les pangénomes et différents calculs de distance de Bray-Curtis associés à un modèle de prédiction des CDS inserts GM. Les résultats présentés sur les génomes GM de la bactérie *E.coli* ainsi que les génomes sauvages et les données synthétiques des six autres bactéries sont très concluants. En effet, seul deux faux positif, un avec la bactérie *M. tuberculosis* sauvage et un avec la bactérie *S. aureus* ayant un gène de *C. jejuni* sont détectés sur 48 génomes analysés. Ces résultats prouvent l'efficacité du logiciel pour la détection de CDS insérées ou la non détection lorsqu'une bactérie sauvage est analysée. L'optimisation des codons ou des dicodons d'une séquence insérée n'interfère pas dans les résultats de DUGMO. De plus, DUGMO est capable de détecter un gène sauvage comportant plus de 9% de mutations. Enfin, DUGMO est disponible pour les utilisateurs sur la plateforme github, `https://github.com/ANSES-Ploufragan/DUGMO`. Cette méthode a fait l'objet d'une publication disponible dans le journal BMC Bioinformatics [145].

Chapitre 8

Perspertives

Sommaire

Les perspectives d'évolution de la méthode DUGMO ainsi que les nouveaux défis à relever sont présentés dans ce chapitre. Ces améliorations pourront être développées dans les versions futures de DUGMO, permettant ainsi d'accroître la capacité de détection de la méthode ainsi que son champ d'application.

8.1 Ajout d'options supplémentaires

DUGMO étant basé sur l'identification de gènes dont le vocabulaire est distinct de ceux de son hôte, il devrait pouvoir détecter dans des bactéries sauvages des gènes résultant de transferts horizontaux récents. Cette hypothèse devrait être testée sur un jeu de données de séquençage où un gène résultant d'un tel transfert est présent.

Actuellement, DUGMO utilise uniquement des données de séquençage brutes provenant de la technologie de séquençage Illumina. Dans une future version, DUGMO sera aussi capable d'accepter les assemblages en tant que données d'entrée, y compris les assemblages long reads. Les assemblages long reads sont réalisés avec les technologies de séquençage de troisième génération comme Pacific Biosciences (PacBio) ou Oxford Nanopore Techologies. Cette nouvelle option permettra ainsi de supprimer les faux positifs dus à des troncatures d'assemblage de données haut débit, dans la mesure où les assemblages long reads auront été faits de façon adéquate. Ainsi, si les assemblages issus de technologie PacBio sont très propres du fait d'erreurs pratiquement aléatoires dans les reads qu'un consensus permet de corriger, la technologie Nanopore souffre d'erreurs d'homopolymères qu'il conviendra de corriger pour éviter des décalages de cadre de lecture des gènes de l'assemblage. Cela peut être fait par l'utilisation de reads illumina, ou

dans un futur sans doute proche par applications de règles obtenues à partir de données d'apprentissage.

Dans DUGMO, lors de la définition des paramètres des alignements BLASTN sur le pangénome, nous avons fait le choix de privilégier la détection de gènes tronqués au risque de détecter des CDS provenant de bactéries proches. Cependant, l'ajout d'une option dans DUGMO pourrait permettre de ne pas détecter ces gènes partiels et ainsi privilégier de manière « plus sûre » les gènes de l'espèce dans l'ensemble des CDS apparentés au génome hôte. Dans tous les cas, il est impossible d'être certains de posséder uniquement des gènes de l'espèce considérée dans l'ensemble des CDS apparentés au génome hôte (exemple illustré par le gène de l'ampicilline ubiquitaire en terme d'espèce et présent dans le plasmide pGMBsub04 du génome de *B. subtilis* GM, Chapitre 7 partie 7.3.1). Cette option pourrait être nécessaire pour l'adaptation de DUGMO à l'analyse de génomes eucaryotes.

8.2 Amélioration avec les génomes bactériens

Dans le but de consolider la méthode DUGMO utilisant des données de séquençage de génomes bactériens, l'utilisation de mots de 7 ou 8 lettres pourrait être ajoutée dans les calculs de distances. Ces tailles de mots correspondent à des mots fréquents dont font partie les motifs KOPS (FtsK Orienting Polarized Sequence). Ces motifs de 8 pb sont reconnus par la protéine FtsK orientant sa translocation [146] 138 permettant ainsi la résolution des dimères de chromosomes dans la région *diff* du chromosome bactérien. Les motifs KOPS sont distincts chez *L. lactis* et *Streptococcus agalactiae* [147], deux bactéries gram + ayant un faible pourcentage en GC. L'ajout d'une nouvelle taille de mot dans les calculs de distances pourrait donc permettre de mieux distinguer les cas de bactéries ayant des vocabulaires proches.

8.3 Adaptation aux génomes eucaryotes

L'utilisation de DUGMO pourra être étendue à la détection des OGM de végétaux et d'animaux (eucaryotes). Deux stratégies sont proposées pour répondre à cette amélioration. La première stratégie nécessite plusieurs adaptations de la méthode. En effet, le pipeline de nettoyage devrait être légèrement modifié : assembler la totalité d'un génome végétal comme le maïs par exemple est irréalisable dans un temps réduit (sa taille étant trop importante). Les données de séquençage de l'OGM potentiel devront être alignées sur le génome de référence correspondant, séparant ainsi les reads alignés appartenant au génome hôte et les reads non alignés correspondant aux inserts potentiellement. Les reads alignés permettront de fournir les consensus de CDS tirés de l'alignement et les reads non alignés seront ensuite assemblés et annotés. Enfin, les CDS potentiellement inserts seront alignés suivant des alignements BLASTN sur le pangénome. De plus, les paramètres des calculs de distance pourraient nécessiter des modifications. L'utilisation des trois calculs de distances présents actuellement dans DUGMO n'a

pas été testée sur le maïs OGM. L'ajout de nouveaux calculs de distance serait donc envisagé en plus des distances présentes actuellement dans la méthode.

La seconde stratégie pour détecter des OGM de végétaux ou d'animaux serait de partitionner les données de séquençage de l'OGM potentiel en « génome eucaryote » et plusieurs génomes procaryotes, chacun distinguable par un pourcentage en GC distinct et une profondeur de couverture éloignée de celle du génome eucaryote. Partitionner permettrait de conserver les paramètres de la méthode actuel en se focalisant sur chaque partie du génome de l'OGM potentiel. Néanmoins, le risque est de perdre de l'information ou de couper l'insert lors de l'assemblage en cas de faible profondeur de couverture. Les données de séquençage seraient coupées selon une profondeur de couverture définie et un pourcentage en GC définis. DUGMO serait ainsi appliqué sur chacun de ces sous-échantillons. Les résultats finaux seront constitués de l'intersection des résultats des différents sous-échantillons. En théorie, l'insert ne fera partie d'aucun de ces sous-échantillons en supposant un petit nombre d'espèces bactériennes contaminantes.

8.4 Adaptation à des données métagénomiques

A long terme, DUGMO pourrait traiter des données métagénomiques simples comme un mélange d'une dizaine de souches bactériennes par exemple. Cependant, le principal obstacle à cette amélioration serait la création du pangénome associé à chacune des dix souches. Une solution pourrait être de créer une base de données publique que les utilisateurs pourrait compléter dès lors qu'ils utilisent DUGMO sur un génome potentiellement OGM non présent dans cette base. Une seconde solution envisagée serait de créer de manière automatique les pangénomes correspondants en sélectionnant des génomes complets issue de séquençage Illumina ou PacBio de l'espèce suspectée via la base de données du NCBI. La création de ces pangénomes pourrait aussi être possible grâce à la base de données de séquences de référence du NCBI, refseq, en utilisant l'outil suivant : `https://github.com/kblin/ncbi-genome-download`. Néanmoins, ces deux solutions soulèvent un problème lié à la définition et à l'exactitude des données mises à disposition par la communauté scientifique et l'indispensable curration de ce qui aura été ajouté.

Le deep learning ou apprentissage profond est une technique de machine learning utilisant un nombre important de couches cachées de réseaux de neurones artificiels et nécessitant une puissance de calcul importante. De plus, cette technique peut traiter une quantité énorme de données complexes et de grande dimension par rapport au machine learning [148]. Le deep learning est principalement utilisé pour la reconnaissance faciale ou vocale mais aussi dans le cas du traitement automatique du langage. Par exemple, cette technique a aussi été récemment utilisé pour développer un modèle permettant de prédire l'évolution de la maladie chez les patients atteint de mésothéliome, un cancer agressif qui attaque souvent la muqueuse des poumons [149]. Pour diminuer le temps de calcul et faciliter l'apprentissage sur des réseaux de neurones très profonds lors de l'utilisation du deep learning, des techniques de transfer learning peuvent être employées. En effet, le transfer learning consiste à stocker les connaissances apprises par un modèle de

machine learning pour résoudre un problème spécifique et à les transférer dans un second modèle. Ce deuxième modèle sera alors capable d'appliquer et d'utiliser cette connaissance pour résoudre sa problématique plus rapidement. Le transfer learning utilise des modèles qui sont déjà entraînées.

Le deep learning et le transfer learning pourront être utilisés avec la banque de données d'OGM connus pour améliorer la détection des inserts et réduire le temps de calcul. Ces deux techniques nécessitent un volume de données important, la banque de données d'OGM connus devra donc être étoffée. Un modèle de transfer learning pourrait être créé avec l'augmentation de la quantité de données disponible et ainsi être utilisé directement dans l'outil.

8.5 Bilan

Les différentes améliorations de la méthode DUGMO consistant en l'ajout d'options qui ont été présentées peuvent être envisageables à court terme. Cependant, pour les améliorations sur l'utilisation de la méthodes sur des génomes eucaryotes et sur des données métagénomiques, des tests devront être réalisés pour valider les stratégies envisagées. Dans ces deux cas, les tests nécessaires à la validation des hypothèses émises représentent un travail important et nécessitent d'avoir accès publiquement à des données de séquençage haut débit de génomes d'OGM de plantes et d'animaux, principal facteur limitant actuellement.

Conclusion

Plusieurs méthodes de détection des organismes génétiquement modifiés (OGM) connus sont disponibles. De plus, des bases de données, regroupant des informations sur les OGM connus, permettent de fournir des informations nécessaires à leur détection. Cependant, aucune méthode bioinformatique n'est, actuellement, disponible pour détecter des séquences codantes (CDS) exogènes provenant d'un OGM où la totalité des séquences insérées sont inconnues. Pour essayer de résoudre ce problème, une méthode de détection de bactéries génétiquement modifiées inconnues à partir de données de séquençage, appelée DUGMO (Detection of Unknown Genetically Modified Organisms), a été développée. L'approche exposée dans ce manuscrit apporte les briques principales d'une méthode de détection des OGM basée sur l'utilisation de motifs statistiques pour pointer les différences de vocabulaire au sein d'un génome. L'originalité et la particularité de cette méthode réside dans l'utilisation du génome hôte et non d'un génome de référence. En effet, l'utilisation du génome hôte permet de définir le vocabulaire de l'espèce potentiellement OGM en vue de distinguer les gènes qui possèdent un vocabulaire distinct du génome hôte.

Dans le but de détecter des bactéries génétiquement modifiées inconnues, la méthode DUGMO nécessite des données de séquençage provenant de la technologie Illumina. Elle présente un faible coût de séquençage pour un volume de données important et une fréquence d'erreurs faible par rapport à d'autres techniques de séquençage haut débit.

Dans un premier temps, un pipeline de nettoyage a été développé pour nettoyer, assembler et annoter les données de séquençage d'un échantillon bactérien pur suspect pour en obtenir les séquences codantes. Ce pipeline, ainsi qu'une suite d'alignements effectuée sur le pangénome de l'espèce considérée, permettent de trier les CDS présents dans le génome suspecté selon deux catégories : les CDS apparentés au génome hôte et les CDS potentiellement inserts. Une banque de données de CDS d'OGM connus est filtrée selon les données du génome suspecté.

Dans un second temps, pour caractériser les différences de vocabulaire présentes entre le génome hôte et la séquence insérée, trois distances de Bray-Curtis avec différents paramétrages sont calculées. Ces distances sont réalisées entre l'ensemble des CDS du génome hôte et chacun de ces CDS ou chacun des CDS de la banque de données d'OGM connus ou chacun des CDS inserts potentiels.

Enfin, un modèle a été créé pour prédire les CDS inserts OGM avérés. Afin d'obtenir d'excellentes performances de prédiction, l'apprentissage automatique a été utilisée pour élaborer un modèle de prédiction. Douze méthodes de discrimination ont ainsi été testées en vue de l'élaboration de ce modèle de prédiction.

Pour cela, les distances précédemment calculées sont utilisées comme variables explicatives ainsi que six autres variables. Suivant nos critères de performance pour la prédiction des CDS (taux de faux positifs, taux de faux négatifs, spécificité et sensibilité), l'union des résultats non redondants des forêts aléatoires et du modèle linéaire généralisé a été retenue.

Pour valider et tester notre méthode de détection des OGM inconnus, la méthode DUGMO a été appliquée à 51 génomes : trois génomes de la bactérie *E. coli* génétiquement modifié, six génomes de bactéries sauvages et 42 génomes provenant de données de séquençage synthétiques. Sur les trois génomes de la bactérie *E. coli*, un seul faux positif est détecté par la méthode DUGMO. Il correspond à un gène tronqué dans l'assemblage du génome de *E. coli*. Sur les 48 autres génomes analysés, uniquement deux faux positifs sont détectés par la méthode DUGMO. L'efficacité de DUGMO est donc avérée pour la détection de CDS insérées et l'absence de détection quand une bactérie sauvage est analysée.

La méthode DUGMO présente d'excellents résultats. Elle pourrait être étendue, moyennant certains développements, à l'analyse de génomes eucaryotes ou de transferts horizontaux de gènes récents.

Annexe A

Plan Management de la qualité et comptes rendus des deux audits qualité

A.1 Plan Management de la qualité

anses
agence nationale de sécurité sanitaire alimentation, environnement, travail
Connaître, évaluer, protéger

Laboratoire de Ploufragan-Plouzané
Unité Génétique Virale et Biosécurité (U GVB)

Plan de Management de la Qualité

du projet de thèse

« Détection d'organismes génétiquement modifiés (OGM) inconnus par analyse statistique de données de séquençage haut débit »

Doctorante : **Julie HUREL**

Directeur de thèse : **André JESTIN**

Co-directeur & Chef de projet : Fabrice TOUZAIN

Rév. 00

Date d'application : 29/05/2017

Rédigé par : Julie Hurel	Approuvé par : Fabrice Touzain	Validé par : Michel Morin
Date : 19/05/2017	Date : 19/05/2017	Date : 19/05/2017
Visa : JH	Visa :	Visa :

Sommaire

1. Introduction

Selon la norme ISO 9000 (version 2015), le management qualité a pour objectif de satisfaire aux exigences des clients et de s'efforcer d'aller au-delà de leurs attentes.

La politique de management qualité de l'ANSES, défini par le directeur général, établit la finalité et les enjeux poursuivis par l'ANSES en pérennisant l'ensemble de la démarche ainsi engagée. La politique de management qualité s'étend aujourd'hui aux thèses de doctorat, projets à part entière. Ainsi, à Ploufragan-Plouzané, pour chaque projet de thèse, un plan management de la qualité doit être défini et appliqué.

Ce plan de management qualité concerne le projet de thèse intitulé « Détection d'organismes génétiquement modifiés inconnus par analyse statistique de données de séquençage haut débit ». Ce projet a été défini par le directeur de thèse et le co-directeur de thèse et sera réalisé et clôturé par le doctorant. La gestion de ce projet sous management qualité permettra de vérifier son bon déroulement par la planification, l'organisation, le suivi, la maîtrise et les rapports des différents aspects du projet. Les mots clés caractérisant la gestion sous management qualité de ce projet sont : traçabilité, organisation, gestion des écarts et révision.

Le management qualité de ce projet s'appuiera sur la norme ISO 10006 « Système de management de la qualité – Lignes directrices pour le management de la qualité dans les projets » pour l'établissement et la mise en œuvre du plan management qualité et sur la norme ISO 17025 « Prescriptions générales concernant la compétence des laboratoires d'étalonnages et d'essai » pour la gestion des ressources.

2. Le champ du plan de management qualité de la thèse

Ce plan de management qualité s'applique aux trois parties du projet de thèse :

- la conduite scientifique du projet,
- la valorisation des résultats,
- la clôture du projet.

Il couvre les travaux conduits dans l'Unité Génétique Virale et Biosécurité (U GVB), laboratoire d'accueil pour la réalisation du projet, ainsi que tous les autres collaborateurs du site même de l'ANSES inscrits dans une démarche qualité.

3. La politique de management qualité du laboratoire de l'unité

L'ANSES de Ploufragan-Plouzané développe une politique de management de la qualité en lien avec la déclaration de politique qualité du directeur du laboratoire. Suite à cet engagement, le

Laboratoire a mis en place un Service de Management Qualité (SMQ). Le SMQ s'assure de la bonne application et du suivi de la politique de management qualité dans les unités de recherche et les services du laboratoire. Il est composé d'un responsable qualité, d'un responsable qualité adjoint, d'un responsable métrologie et d'une secrétaire. Dans chaque unité et service du Laboratoire, un correspondant qualité et un correspondant métrologie sont désignés pour appliquer la politique qualité du chef d'unité ou chef de service en lien avec le SMQ.

Le Laboratoire de Ploufragan-Plouzané est composé de huit unités de recherche. Cinq d'entre-elles sont accréditées par le Comité Français d'Accréditation (COFRAC) au titre de la norme NF EN ISO 17025 :

- Unité Hygiène et Qualité des Produits Avicoles et Porcins (HQPAP),
- Unité Virologie, Immunologie, Parasitologie Aviaires et Cunicoles (VIPAC),
- Unité Virologie, Immunologie Porcines (VIP),
- Unité Mycoplasmologie, Bactériologie (MB),
- Unité Pathologie Virale des Poissons (PVP) située à Plouzané (29)).

L'unité GVB a fait le choix de développer une démarche de management qualité depuis 2004. Le chef d'unité décrit son engagement en matière de politique qualité dans une déclaration (cf. annexe 4). Un correspondant qualité a été nommé afin d'élaborer et de réviser le Plan Qualité de l'unité ainsi que les procédures et les modes opératoires spécifique au fonctionnement de l'unité.

La documentation qualité de l'ANSES de Ploufragan-Plouzané comprend le Manuel de Management Qualité, les plans qualités de chaque unité, les procédures, les modes opératoires et les supports d'enregistrement. La procédure P.ESS.P1 décrit la gestion d'un projet au sein de l'ANSES de Ploufragan-Plouzané et notamment la gestion des thèses avec pour référence la norme ISO 10006.

La rédaction du plan management qualité de la thèse fait également partie de la démarche de management qualité suivie par l'unité GVB. Elle permet ainsi de formaliser le processus d'organisation des travaux de thèse dans le but d'assurer la qualité du résultat final.

4. Les acteurs du projet

Il convient d'énumérer les acteurs du projet et de définir leur rôle respectif dans ce plan management qualité de thèse. Le projet de thèse qui va être réalisé au cours de ces trois prochaines années (2017-2019) est une thèse en co-tutelle au sein de l'unité GVB dont la structuration est mentionnée dans l'organigramme (cf. Annexe 3). Le directeur de thèse et le co-directeur de thèse

encadrent le doctorant avec des rôles complémentaires. Les relations entre les différents acteurs du projet sont définies dans l'organigramme de thèse (Annexe 1).

Le directeur de thèse est André JESTIN, son rôle est de s'assurer du bon déroulement de l'apprentissage du doctorant par la recherche et de donner une opinion générale de l'avancement de la thèse. Le co-directeur de thèse, Fabrice TOUZAIN, chef de l'équipe Bioinformatique, a également un rôle de référent. Tout comme le directeur de thèse, il dirige le projet et s'assure du bon déroulement et de l'avancement de celui-ci. Il assume la responsabilité du rôle de chef de projet. Selon la norme ISO 10006, un chef de projet est une personne qui a des responsabilités et une autorité définies pour diriger le projet et s'assurer que le système de management de la qualité du projet est établi, mis en œuvre et entretenu.

Mathieu ROLLAND et Anne-Laure BOUTIGNY de l'ANSES Laboratoire de la Santé des Végétaux à Angers et Frédéric DEBODE du Centre Wallon de recherches agronomiques participent à ce projet en fournissant des données confidentielles de séquençage haut débit d'OGM.

Le comité de suivi individuel du doctorant est composé de :

- Dr. Mathieu ROLLAND, Laboratoire de la santé des végétaux ANSES Angers,
- Dr. Christophe HITTE, tuteur pour l'école doctorale Vie Agro Santé (VAS), Ingénieur de recherche au CNRS à Rennes,
- Dr. Johann JOETS, Ingénieur de recherche à l'INRA du Moulon,
- MSc. Mauro PETRILLO, Joint Research Centre (JRC),
- Dr. André JESTIN, Directeur de thèse,
- Dr. Fabrice TOUZAIN, Co-directeur de thèse.

Ce comité assure une analyse critique de l'avancée du projet et conseillera le directeur de thèse, le co-directeur de thèse et le doctorant sur la suite du projet. Le tuteur est un scientifique de l'école doctorale VAS de Rennes. Son rôle est de s'assurer du bon déroulement de la thèse, de faire le lien avec l'école doctorale et de guider le doctorant dans son parcours professionnel.

Le doctorant a la charge de conduire son projet de thèse. En début de thèse, il va suivre les orientations et consignes de son directeur et de son co-directeur de thèse puis il va très vite adopter une position critique (au sens du débat constructif) et être indépendant dans la conduite de sa recherche. En fin de thèse, le doctorant doit être devenu un chercheur opérationnel et autonome.

Comme décrit dans la charte des thèses de l'école doctorale VAS, le doctorant s'engage à suivre certaines formations telles que des enseignements, des conférences et des séminaires de l'école doctorale au cours de son projet. L'ANSES s'engage, dans la limite de ses moyens financiers et des priorités, à permettre au doctorant d'assister à des congrès dans le cadre de son projet. Pour

cela, le doctorant remplit une fiche de vœux qui permettra l'élaboration du programme des congrès auxquels il souhaite assister pour l'année suivante.

Le management qualité se traduit aussi par une procédure d'accueil du personnel. Le personnel composant l'unité GVB est sensibilisé au management qualité et reconnu compétent dans ses activités. Une fiche personnelle est remplie par chaque personne appartenant à l'unité. Elle fixe les responsabilités et les compétences de chacun. Elle est co-signée par l'intéressé et son chef d'unité/service.

Enfin, Michel MORIN, chef du service de management qualité, aide à la formalisation du plan management qualité de la thèse et à vérifier l'application de ses dispositions.

5. La conduite du projet

5.1. Un projet formalisé

Le projet scientifique est formalisé selon la fiche de définition de projet intitulé « Détection d'organismes génétiquement modifiés inconnus par analyse statistique de données de séquençage haut débit ». L'objectif du projet est de créer un outil informatique permettant de détecter des OGM inconnus et de déterminer la séquence d'origine exogène.

Les quatre étapes établies et révisables si besoin au cours du projet sont les suivantes :

- Mise en place d'une routine de traitement des données de séquençage,
- Choix de la méthode statistique appropriée pour détecter de façon optimale des séquences d'OGM inconnus,
- Développement de l'outil à partir de la méthode statistique choisie,
- Intégration de l'outil dans la plateforme Galaxy.

Le diagramme de GANT de la thèse (Annexe 2) définit les différentes phases du projet ainsi que la rédaction du mémoire de thèse, et précise leurs durées estimées.

La documentation scientifique et technique du projet est disponible au sein de l'unité GVB et pour l'ensemble des acteurs du projet. Les différentes méthodes utilisées au cours du projet sont consignées dans les cahiers de laboratoire spécifiques au projet.

5.2. Des étapes de revue

Le planning de déroulement de la thèse, le contenu et/ou les objectifs scientifiques ainsi que le plan management qualité du projet sont révisables si nécessaire. Le doctorant est source de propositions d'évolution du projet et ce rôle doit s'affirmer par des étapes de révision au cours de l'avancement de la thèse.

Des étapes de revues de projet sont réalisées au cours du projet afin de discuter de l'avancement de ce dernier et des directions à prendre concernant chaque étape du planning (Annexe 2). De plus, elles serviront à réviser le projet si besoin. Ces revues sont réalisées à la fin de chaque étape scientifique du projet et/ou au cours des deux réunions de comité de suivi individuel du doctorant. A l'issue de chaque revue un compte rendu est réalisé par le doctorant et ensuite diffusé à chacun des acteurs et archivé.

Les propositions de modifications faites par le doctorant sont discutées avec le chef de projet et le directeur de thèse. En dernière instance, le chef de projet et le directeur de thèse décident des modifications à apporter au projet. En cas de désaccord persistant avec le doctorant, le comité de suivi individuel du doctorant arbitrera la décision.

5.3. La traçabilité assurée

La traçabilité des différents travaux et cheminements réalisées tout au long du projet est assurée par le cahier de laboratoire tenu par le doctorant et propriété de l'unité. Celui-ci contient :

- La date et l'intitulé des analyses
- La description précise des analyses, au fur et à mesure de leur avancement,
- Les lignes de commandes réalisées,
- Toute nouvelle hypothèse de travail,
- Les interprétations, critiques et commentaires sur les analyses effectuées et sur les résultats obtenus

Toutes les données (brutes et générées) sont enregistrées et sauvegardées sur un serveur de stockage externe toutes les semaines. Les comptes rendus des réunions sont numérisés et enregistrés en local sur le poste de travail ainsi que sur le serveur de stockage. Enfin, la gestion des versions du logiciel développé sera assurée par le logiciel Git utilisant la documentation correspondante sur le serveur GitLab de l'unité GVB.

5.4. Des revues du système de management de la qualité

D'après la norme ISO/EIC 17000 de 2004, un audit interne est un processus systématique, indépendant et documenté, permettant d'obtenir des enregistrements, des énoncés de faits ou d'autres informations pertinentes, et de les évaluer de manière objective pour déterminer dans quelle mesure les exigences sont respectées.

Un audit interne sur les engagements du présent plan management qualité de la thèse est réalisé par le responsable qualité. Cet audit, réalisé selon la procédure de l'ANSES de Ploufragan-Plouzané, aura lieu à la fin de la première année puis au terme de la seconde année de travail. Il permet d'évaluer le respect des dispositions décrites dans ce plan management qualité de la thèse. Si

un écart existe, il pourra être revu. Ces audits sont également destinés à examiner les processus de management en lien avec les acteurs, vérifier leur efficacité et rechercher des améliorations potentielles. Il donne lieu à un rapport d'audit, communiqué aux membres du comité de suivi individuel du doctorant.

5.5. Des contrôles pour la qualité du résultat final

Le management de projet est destiné à obtenir des résultats fiables par le souci permanent de la qualité des résultats indiquée tout au long des processus. La gestion des erreurs est réalisée grâce à un fichier log qui contient toutes les étapes lors de l'exécution d'un programme. De plus, pour valider la méthode et le logiciel finalisé, deux types de tests seront menés :

- Les tests sur des données de séquençage brutes utilisés pour mettre en place le pipeline de nettoyage et assemblage,
- Les tests en aveugle sur des séquences d'OGM inconnus.

Ces tests prendront en compte différentes natures d'inserts (insert : matériel génétique, constitué d'un ou plusieurs gènes d'origine exogène, visant à améliorer certaines caractéristiques de l'organisme sauvage modifié) pour vérifier la pertinence du logiciel.

5.6. Un processus de management des risques

D'après la norme ISO 10006, le terme risque considère, comme le terme incertitude, les aspects tant positifs que négatifs. Le management des risques du projet traite des incertitudes tout au long du projet. Ce projet sera réalisé uniquement avec des outils informatiques, les risques pouvant impacter le résultat final seront liés à la méthode statistique utilisée et aux paramètres choisis. Si le logiciel développé n'est pas conforme aux attentes fixées (distinction entre séquence OGM et séquence sauvage pour tout OGM considéré) alors le champ d'application de la méthode devra être restreint à un type d'OGM, quitte à créer plusieurs méthodes, chacune dédiée à un type d'OGM. Pour anticiper ces risques, des actions préventives seront mises en place comme la vérification de la méthode statistique par une personne spécialiste de ce domaine, Sophie SCHBATH et des tests sur des OGM de différentes nature pour évaluer, dès les premières mises en place de l'algorithme, ses limitations, son caractère universel ou non. Ces risques et actions préventives seront répertoriés sur une fiche numérisée et sauvegarder sur le serveur de stockage. Elle sera mise à jour à chaque nouveau risque déterminé par l'un des acteurs du projet.

5.7. Une revue complète des résultats lors de la clôture du projet

La revue finale consiste à formaliser les résultats, à en faire une revue complète et à les comparer le cas échéant aux résultats attendus. Elle est transmise à toute l'équipe du projet. Le

projet de thèse se clôture par la rédaction d'un manuscrit de thèse et par une soutenance devant un jury amené à valider les résultats et le mémoire de thèse. Les différents acteurs du projet interviendront pour prendre en compte les résultats pertinents. Un bilan du management qualité de la thèse sera rédigé pour les membres du jury par le responsable qualité.

6. Les équipements

Les équipements de l'unité GVB sont gérés selon les plans de qualités des unités en référence à la norme ISO 17025. L'inventaire et le suivi des équipements sont formalisés par une gestion planifiée et maîtrisée. Cependant, le matériel d'analyse présent au laboratoire ne concerne pas ce projet de thèse, les données brutes n'étant pas réalisées au Laboratoire de Ploufragan-Plouzané. Seul de l'équipement informatique sera utilisé lors de ce projet. Le doctorant dispose d'un ordinateur de travail avec les caractéristiques suivantes :

- 64 Go de RAM (mémoire vive),
- un SSD (Solid-State Drive) de 1 To,
- 2 disques durs de 3 To,
- 12 threads à 3.5 GHz.

Pour réaliser des analyses nécessitant une capacité de calcul plus importante, le doctorant pourra utiliser le serveur de calcul du laboratoire nommé Jupiter. Ses principales caractéristiques techniques sont :

- 384 Go de RAM (mémoire vive)
- 56 threads

Un service de support informatique est présent au Laboratoire de Ploufragan-Plouzané pour traiter toutes les demandes et les problèmes rencontrés concernant le matériel informatique.

7. Les consommables

Ce projet ne nécessite pas l'utilisation de consommables.

8. Les locaux

L'unité GVB est installée sur le site des Croix au sein du bâtiment F06. L'accès à ce bâtiment est contrôlé au moyen d'un digicode sur les portes d'accès extérieurs et un registre permet d'identifier les visiteurs autorisées à pénétrer dans les locaux.

L'accès au laboratoire nécessite de retirer les vêtements et chaussures dans un premier vestiaire. Les vêtements de travail sont récupérés dans un second vestiaire après le passage par un sas. Le bâtiment est en dépression par rapport à l'extérieur et possède une climatisation ainsi qu'un système de filtrage de l'air sur filtres à haute efficacité. L'unité est aménagée de manière à disposer de salles spécifiques à des activités qui sont soit en dépression soit en surpression par rapport au couloir. Actuellement, l'unité est en réaménagement pour créer des espaces bureaux supplémentaires.

9. Le coût

Le projet de thèse a une durée de trois ans à compter du 2 Janvier 2017. Il est financé par la région Bretagne et le budget est en adéquation avec le coût du projet. Dans la mesure du possible, la gestion des coûts sera suivie dans le cadre des revues de projet.

10. La gestion des écarts

10.1. Les non-conformités

Selon la norme ISO 9000, une non-conformité est une non-satisfaction d'une exigence du système qualité et/ou d'exigences convenues avec le client.

La non-conformité constatée est consignée sur une fiche de progrès qui donne lieu à des actions curatives, correctives et/ou préventives et qui est gérée selon la procédure générale P.ACT.A1. Elle spécifie la gestion des non-conformités détectées en rapport avec les consommables ou avec tout autre type de ressources. Dans notre cas, les non-conformités regroupent surtout les problèmes techniques, comme le dysfonctionnement du matériel.

10.2. Les non-confirmations

Une non-confirmation correspond à une différence importante de résultats entre deux répétitions de méthodes au cours d'une même étude. Les programmes informatiques utilisés lors de ce projet étant déterministes, ce type de problème ne devrait pas être rencontré. Cependant, si des programmes non déterministes (dont le résultat peut différer d'un lancement à un autre) devaient être utilisés, la conformité du résultat sera vérifiée. Par exemple, sur 100 lancements, le programme devra toujours fournir des résultats comparables et conformes à nos attentes.

La gestion de ces cas est consignée dans le cahier de laboratoire et discutée avec le directeur de thèse et/ou avec le co-directeur de thèse. Une non-confirmation peut provoquer une revue de projet suivant les besoins.

11. La communication

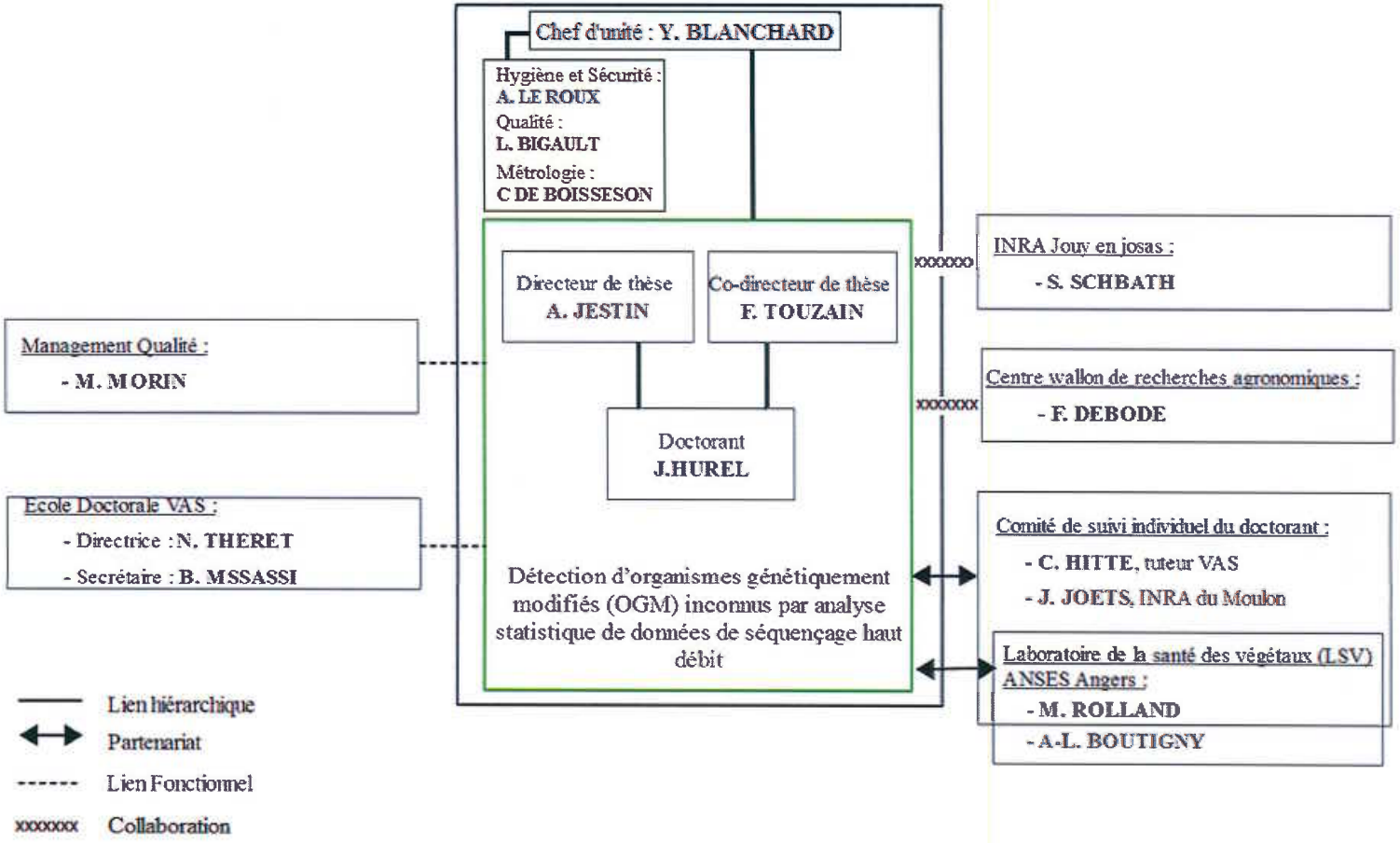
Des réunions de travail seront mise en place pour informer sur le travail qui a été réalisé et discuter de l'avancement du projet. Différents types de réunions sont déterminés en début de projet :

- Les réunions hebdomadaires avec le co-directeur de thèse permettant de suivre régulièrement la direction du projet. Elles font l'objet d'un compte rendu numérique.
- Les réunions avec le comité de suivi individuel du doctorant (CSID), une fois par an, ont pour objectif de discuter de l'avancement et des directions à prendre pour la suite du projet. Un compte rendu est réalisé par le doctorant à l'issue de ces réunions. Il est ensuite transmis à chacun des acteurs et archivé.

Les propositions de modifications faites par le doctorant sont discutées avec le chef de projet et le directeur de thèse. En dernière instance, le chef de projet et le directeur de thèse décident des modifications à apporter au projet. En cas de désaccord persistant avec le doctorant, le comité de suivi individuel du doctorant arbitrera la décision.

Le projet sera valorisé par la rédaction d'un mémoire de thèse et les résultats obtenus feront l'objet de posters, de présentations lors de congrès nationaux ou internationaux et d'au moins une publication scientifique à comité de lecture.

Annexe 1 : Organigramme de thèse

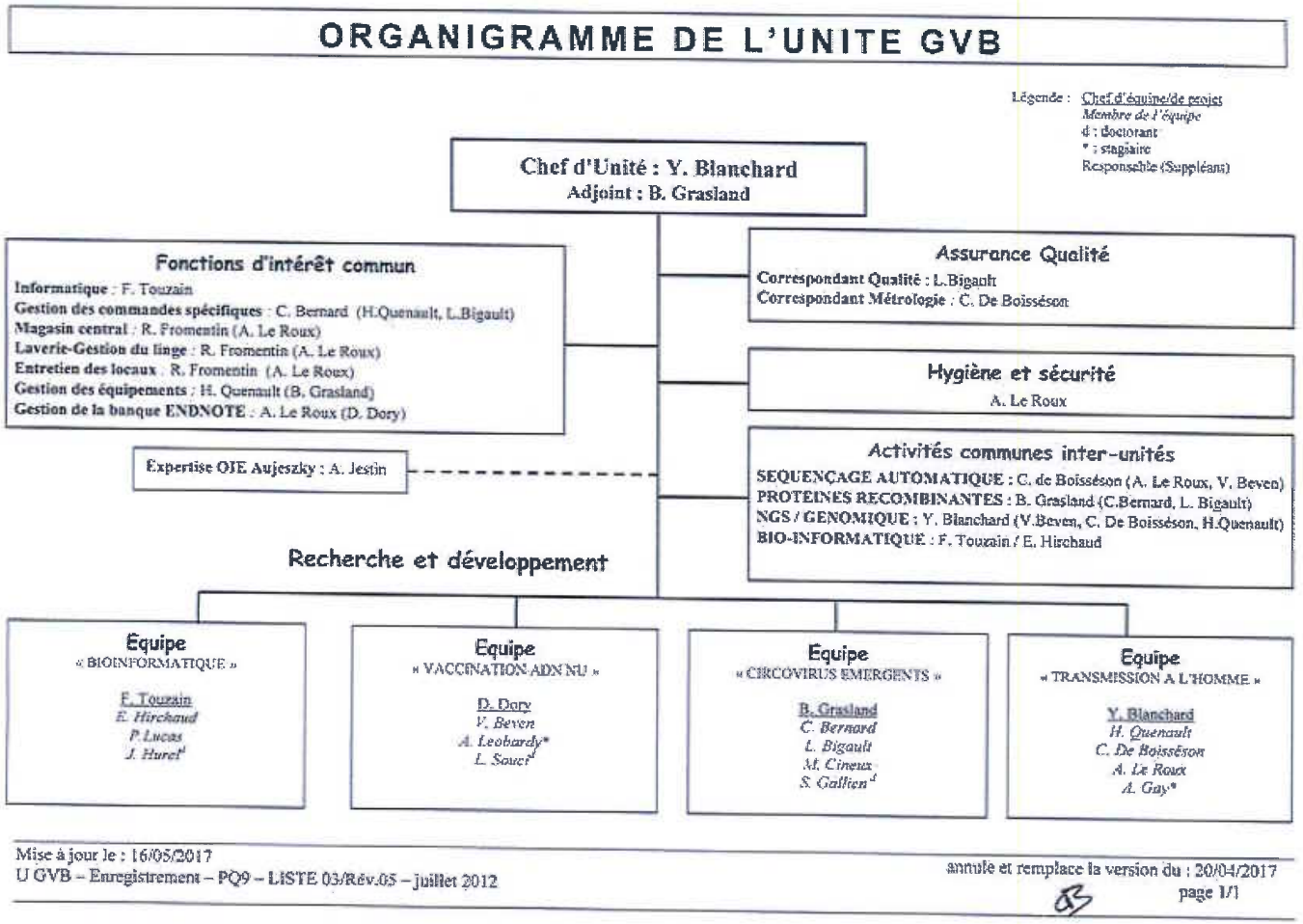


Annexe 2 : Diagramme de GANT de la thèse

Tâche	Trimestre											
	1	2	3	4	5	6	7	8	9	10	11	12
1. Bibliographie	■											
2. Mise en place d'une routine de traitement des données de séquençage		■	■									
3. Choix de la méthode statistique appropriée pour détecter de façon optimale des séquences d'OGM inconnus			■	■								
4. Développement de l'outil à partir de la méthode statistique choisie					■	■	■	■				
5. Intégration de l'outil dans la plateforme Galaxy								■	■			
6. Rédaction du mémoire de thèse									■	■	■	
7. Soutenance de thèse												■

11/10/2019 : Date limite d'envoi du manuscrit de thèse

Annexe 3 : Organigramme de l'unité GVB



Annexe 4 : La déclaration de la politique qualité de l'unité

anses
alimentation, environnement, travail

Déclaration de politique qualité de l'unité
Génétique Virale et Biosécurité

L'unité Génétique Virale et Biosécurité (UGVB) a été créée en 2000 par décision du Directeur de l'AFSSA sur proposition du Conseil Scientifique. Ses missions sont majoritairement une recherche en virologie moléculaire sur les virus émergents et les vecteurs viraux, et la description des relations avec l'hôte en faisant appel aux technologies « omiques », s'y ajoute, dans une moindre mesure , des activités d'expertise ainsi qu'un appui scientifique et technique aux autres unités de l'Anses. ». De plus, depuis 2014 l'unité héberge la plateforme de séquençage haut débit de l'Anses.

En 2004, l'UGVB s'est engagée dans la mise en œuvre d'une démarche qualité en recherche permettant d'assurer la fiabilité des résultats et la rigueur avec laquelle les projets de recherche sont conduits .Les normes NF EN ISO/CEI 17025 « prescriptions générales concernant la compétence des laboratoires d'étalonnages et d'essais » et ISO 10006 « Lignes directrices pour le management de la qualité dans les projets » constituent le référentiel sur lequel s'appuie ce système de management.

L'application de la norme NF EN ISO/CEI 17025 est la base de l'organisation interne du laboratoire, notamment pour la gestion des ressources
La norme ISO 10006 est utilisée comme guide pour réaliser les thèses sous assurance de la qualité. Toutes les thèses entreprises dans l'unité sont désormais gérées en mode projet suivant les recommandations de cette norme. Les thèses réalisées en cotutelle, avec des équipes extérieures à l'Anses, pourront être dispensées de cette gestion en mode projet.

La mise en place du système de management de la qualité dans l'unité ayant été franchie, l'UGVB doit maintenant maintenir ses efforts sur l'amélioration continue de son système qualité.

Le développement important de la bioinformatique ces dernières années et la constitution d'une équipe bioinformatique au sein de GVB, en appui entre autre de la plateforme NGS nécessite, au même titre que pour les autres activités de l'unité, la mise en œuvre d'une politique de traçabilité et de sécurisation des données. Conscient de pratiques différentes dans ce champ d'activité, par rapport aux usages en vigueur pour les autres thématiques de GVB, il conviendra aux bio-informaticiens de définir, formaliser et appliquer les moyens nécessaires pour atteindre ces objectifs de traçabilité et de sécurité. Ceci se traduira par la mise en œuvre d'actions de progrès décidées en concertation dans l'unité.

Conformément à la déclaration de politique qualité du Directeur de l'Anses laboratoire de Ploufragan-Plouzané, j'invite tous les agents de l'Unité à s'investir dans la mise en place de ce programme d'action et ceci pour un maximum d'actions de recherche.

A Ploufragan, le 02/05/17
Le Chef de l'Unité
Yannick Blanchard

***
*

Page **15** sur **15**

A.2 Attestation de management qualité de la thèse

Laboratoire de Ploufragan-Plouzané-Niort

Service Management Qualité

Thèse préparée par Julie HUREL

« Détection d'organismes génétiquement modifiés (OGM) inconnus par analyse statistique de données de séquençage haut débit »

Attestation de management qualité de la thèse

Par Eric CHORIN, chef du Service de Management Qualité

Les travaux de thèse conduits par Julie HUREL se sont déroulés sous management de la qualité, dans l'unité « Génétique Virale et Biosécurité » (U GVB) à compter du 02 Janvier 2017. Cela s'est concrétisé par la rédaction, dès le début des travaux et par la doctorante, d'un Plan de Management Qualité (PMQ) signé par elle-même, le Co-Directeur de thèse et le Responsable qualité du Laboratoire.
Véritable engagement pour le management de la thèse, ce PMQ a deux normes pour référence :

- la norme NF EN ISO/CEI 17025 (Prescriptions générales concernant la compétence des laboratoires d'étalonnages et d'essais) appliquée dans ce projet, pour la gestion des ressources au sein de l'unité GVB ;
- la norme ISO 10006 (Systèmes de management de la qualité – lignes directrices pour le management de la qualité dans les projets) mise en œuvre pour la conduite du projet de thèse.

Le PMQ fait systématiquement le lien entre le sujet de la thèse, sa mise en œuvre et la formalisation de son déroulement. Avec ce document, ce sont des éléments d'organisation et de gestion des travaux de thèse qui ont été réfléchis, décidés et formalisés. Ce plan comprend notamment l'identification des « points critiques » dans le projet de thèse, les données de sortie attendues à chaque étape identifiée et des « revues de projet » conformément à la norme ISO 10006.

La mise en œuvre du PMQ

Comme pour toute démarche de management qualité, la conduite de la thèse s'est traduite par un suivi régulier de l'encadrement avec validation des travaux de la doctorante. De nombreuses données ont été enregistrées conformément aux engagements en termes de traçabilité.
Le management de projet a donné lieu à des vérifications de l'application des dispositions d'organisation formalisées dans le PMQ. Cela s'est notamment traduit par deux audits internes de management de projet conduit par le Chef du Service Management Qualité les 10 Juillet 2018 et 29 Novembre 2019. Les documents examinés lors de ces audits sont listés en annexe de cette attestation.
Au cours de ces audits, aucune fiche de non-conformité n'a été notifiée : les dispositions du PMQ ont donc été parfaitement suivies par la doctorante.
Les exercices de traçabilité faits par l'auditeur se sont révélés conformes aux engagements de l'équipe de thèse. Plusieurs cahiers de laboratoires normalisés ont été utilisés, y compris pour les réunions de l'équipe de thèse. Les cahiers examinés sont extrêmement bien tenus et visés régulièrement par l'encadrement. La doctorante a aussi stocké ses nombreuses données dans un dossier électronique très bien structuré permettant de retrouver aisément l'ensemble des données brutes et des données

analysées.
Les réunions prévues d'avancement du projet et de suivi de la doctorante ont effectivement été réalisées, les comptes rendus sont disponibles et permettent d'identifier le but de la réunion, les données d'entrée et de sortie, la validation des différentes étapes du projet, les éventuels besoins de réorientation des recherches…
Les audits ont permis d'identifier :

- le gros travail réalisé par l'équipe et en particulier par la doctorante ;
- la qualité de la traçabilité des travaux conduits ;
- la rigueur du cheminement intellectuel de la doctorante dans la conduite de son projet.

Bilan du management qualité de la thèse

Le déroulement de la thèse sous management qualité, a d'abord permis à Julie HUREL de découvrir deux référentiels qualité qui, dans l'esprit au moins, lui seront nécessaires pour sa carrière professionnelle et directement bénéfiques pour des fonctions éventuelles au sein d'un laboratoire ou dans le management d'un projet.
En second lieu, ce management qualité a dû lui être utile sur plusieurs points :

- la formalisation la plus complète possible du projet scientifique, en particulier au démarrage ;
- l'importance de clarifier les rôles respectifs des différents acteurs du projet mais aussi de s'intégrer dans leur agenda respectif ;
- l'intérêt de réfléchir, au moins au début du projet, aux éventuels points critiques et les actions préventives nécessaires pour les maitriser. Dans le cas présent, cet aspect « point critique » a été parfaitement géré ;
- la nécessité de mettre en œuvre la traçabilité des ressources (équipements principaux suivis et vérifiés) afin de pouvoir analyser dans les meilleures conditions tout résultat inattendu ;
- l'importance de formaliser tous les résultats et les cheminements intellectuels.

Au final, je suis convaincu que Julie HUREL a su mener avec une grande rigueur son projet scientifique en ayant partagé dès le début de sa thèse avec son encadrement la démarche de management qualité.

Fait à Ploufragan le 09 Avril 2020

Eric CHORIN
Chef du Service Management Qualité

2

Annexe :

Principaux documents examinés lors de l'audit du 10 Juillet 2018 :

Examen de la mise en place de la routine d'analyse de données de séquençage du génome d'un maïs sauvage et d'un maïs potentiellement OGM, alignement des reads sur un génome de référence, assemblage des reads non alignés, puis suppression des séquences communes. Examen de l'outil R'Mes qui permet la recherche de mots exceptionnels dans une séquence. Examen du cahier de laboratoire concernant les données ayant permis le choix du critère de 10% des mots surreprésentés : la traçabilité est très satisfaisante, le cahier est régulièrement signé actant la validation et l'avancée des travaux menés. Les sauvegardes informatiques sont judicieusement planifiées. Les documents préparatoires aux comités de thèse des 22/09/2017 et 11/06/2018 sont clairs et correctement archivés. Les comptes rendus sont disponibles. En accord avec l'école doctorale, diverses formations ont été suivies par la doctorante.
Au jour de l'audit, la doctorante n'a pas eu à enregistrer de non-conformité lors de la réalisation des travaux. Par ailleurs, il n'a pas été observé de non-confirmation des hypothèses scientifiques qui aurait pu amener à modifier le PMQ relatif au projet. Les travaux avancent selon le rythme défini dans le plan de management de qualité de la thèse.

Il n'a pas été notifié de non-conformité au cours de cet audit.

Principaux documents examinés lors de l'audit du 29 Novembre 2019 :

Vu les comptes rendus des réunions hebdomadaires bien classés et archivés. Vu notamment le compte-rendu du 24/04/2019 date à laquelle il a été décidé de réorienter le travail et de ne plus se consacrer qu'aux bactéries. Les comptes rendus sont très précis. Le calendrier des réunions est très bien respecté.

Le logiciel a été mis au point sur une souche de *Bacillus subtilis* modifiée pour produire de la vitamine B12. Le logiciel est maintenant capable de détecter les séquences insérées dans les 4 plasmides qui sont différentes des séquences de subtilis. Le logiciel DUGMO a été validé par test sur 3 souches de E. coli ogm : le logiciel détecte tout ce qui est étranger à la bactérie hôte. Le logiciel a aussi été testé sur le subtilis sauvage (Ref. SRR 8935610 du NCBI) et ne montre pas de faux positif. Le logiciel est disponible sur le site github de l'Anses.

Vu le projet de publication : "DUGMO : tool for detection of unknown gmo with high throughput sequencing data" soumise à BMC Bioinformatics. La publication était l'un des produits de sortie attendus à l'issue de la these.

Les données sont sauvegardées toutes les semaines sur le serveur de l'Anses, il existe un rapport par mail pour savoir si la sauvegarde s'est effectuée correctement.

Il n'a pas été enregistré de non-conformité lors des travaux sur la période écoulée.

Une communication a été faite lors des journées doctorales anses d'octobre 2017.

Vu le cahier de labo C60091 : cahier très bien tenu, et signé très régulièrement par le co-directeur de thèse. Le cahier contient les conclusions des manipulations et aussi les chemins d'accès de l'ensemble des données stockées. La traçabilité est très bonne satisfaisante.

Il n'a pas été notifié de non-conformité au cours de cet audit.

Annexe B

Souches utilisées pour construire les pangénomes des différentes bactéries

Sur les tableaux suivant, le type de données de séquençage (colonne « niveau ») est représenté suivant quatre catégories :

- chr correspond aux données de séquençage d'un chromosome,
- wgs (whole genome sequencing) correspond aux données d'assemblage d'un génome incomplet (contigs),
- complet correspond aux données de séquençage d'un génome complet,
- plasmide correspond aux données de séquençage d'un plasmide (fini entier).

Le - indique que l'information n'est pas disponible.

B.1 Pangénome de *Bacillus subtilis*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Bacillus subtilis strain BJ3-2	CP025941.1	4443	4.2	complet
Bacillus subtilis subsp. *subtilis* strain BS155	CP029052.1	4576	4.3	complet
Bacillus subtilis subsp. *subtilis* str. OH 131.1	CP007409.1	4062	4.0	complet
Bacillus subtilis subsp. *subtilis* str. JH642 substr. AG174	CP007800.1	4361	4.2	complet
Bacillus subtilis subsp. *subtilis* str. AG1839	CP008698.1	4361	4.2	complet
Bacillus subtilis strain B-1	CP009684.1	3868	3.9	genome

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Bacillus subtilis subsp. *subtilis* strain IIG-Bs27-47-24	CP016787.1	3035	2.9	genome
Bacillus subtilis subsp. *subtilis* strain PG10	CP016788.1	2890	2.8	genome
Bacillus subtilis subsp. *subtilis* strain PS38	CP016789.1	2824	2.7	genome
Bacillus subtilis subsp. *subtilis* strain QB5412	CP017312.1	4436	4.2	genome
Bacillus subtilis subsp. *subtilis* strain QB5413	CP017313.1	5248	4.2	genome
Bacillus subtilis strain SR1	CP021985.1	4155	4.1	genome
Bacillus subtilis subsp. spizizenii strain SW83	CP030925.1	4126	4.0	chr
Bacillus subtilis subsp. *subtilis* strain IITK SM1	CP031675.1	4238	4.0	chr
Bacillus subtilis strain FB6-3	CP032089.1	4105	4.2	chr
Bacillus subtilis strain MZK05	CP032315.1	4352	4.1	complet
Bacillus subtilis strain SRCM104005	CP035164.1	4325	4.1	complet
Bacillus subtilis strain SRCM103571	CP035231.1	4329	4.1	complet
Bacillus subtilis strain SRCM103835	CP035400.1	4325	4.1	complet
Bacillus subtilis strain SRCM103622	CP035411.1	4385	4.1	complet
Bacillus subtilis strain SRCM103629	CP035413.1	4306	4.1	complet
Bacillus subtilis subsp. *subtilis* str. 168	NC_000964.3	4106	4.2	complet
Bacillus subtilis subsp. *spizizenii* str. W23	NC_014479.1	4116	4.0	complet
Bacillus subtilis BSn5	NC_014976.1	4237	4.1	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Bacillus subtilis subsp. *spizizenii* TU-B-10	NC_016047.1	4315	4.2	complet
Bacillus subtilis subsp. *subtilis* RO-NN-1	NC_017195.1	4115	4.0	complet
Bacillus subtilis subsp. *natto* BEST195 DNA	NC_017196.2	4366	4.1	complet
Bacillus sp. JS	NC_017743.1	4185	4.1	complet
Bacillus subtilis QB928	NC_018520.1	4332	4.1	complet
Bacillus subtilis subsp. *subtilis* str. BSP1	NC_019896.1	4155	4.0	complet
Bacillus subtilis XF-1	NC_020244.1	4175	4.0	complet
Bacillus subtilis subsp. *subtilis* 6051-HGW	NC_020507.1	4422	4.2	complet
Bacillus subtilis subsp. *subtilis* str. BAB-1	NC_020832.1	4124	4.0	complet
Bacillus subtilis PY79	NC_022898.1	4187	4.0	complet
Bacillus subtilis strain SG6	NZ_-CP009796.1	4223	4.0	complet
Bacillus subtilis strain B4146	NZ_-JXHR01000-001.1	4464	4.0	wgs
Bacillus subtilis strain HDZK-BYSB7	CP026608.1	5695	5.3	complet

Tableau 17 – Liste des 37 génomes utilisés pour construire le pangénome de *B. subtilis* (partiellement adapté de [86]).

B.2 Pangénome de *Echerichia coli*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Escherichia coli B strain C2566	CP014268.2	4428	4.4	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Escherichia coli 0127 :H6 strain E2348/69	NC_011601.1	5388	4.9	complet
Escherichia coli ST131 strain EC958	NZ_HG941718.1	5472	5.1	complet
Escherichia coli BL21(DE3)	NC_012892.2	4700	4.5	complet
Escherichia coli strain BA22372	NZ_CP040397.1	4833	4.7	complet
Escherichia coli strain RM14715	NZ_CP027104.1	4852	4.8	complet
Escherichia coli strain FORC_081	NZ_CP029057.1	5138	4.7	complet
Escherichia coli strain WCHEC025970	NZ_CP036177.1	5284	4.7	complet
Escherichia coli strain 2012C-4227	NZ_CP013029.1	5534	5.2	complet
Escherichia coli Nissle 1917	NZ_CP007799.1	5495	5.4	complet
Escherichia coli UM146	NC_017632.1	5049	4.9	complet
Escherichia coli CFT073	NC_004431.1	5169	5.2	complet
Escherichia coli strain ECZP248	NZ_CP034784.1	4812	4.4	complet
Escherichia coli strain BH100 substr. MG2014	NZ_CP024650.2	5102	5.0	complet
Escherichia coli strain FAM21845	NZ_CP017220.1	5139	4.9	complet
Escherichia coli strain SCEC020022	NZ_CP032892.1	5417	4.9	complet
Escherichia coli str. K-12 substr. MG1655	NC_000913.3	4395	4.6	complet
Escherichia coli strain WCHEC025943	NZ_CP027205.2	6343	4.8	complet
Escherichia coli strain FORC_069	NZ_CP023061.1	6478	5.1	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Escherichia coli strain cq9	NZ_-CP031546.1	7284	5.2	complet
Escherichia coli O157 :H7 str. Sakai DNA	NC_002695.2	5361	5.5	complet
Escherichia coli strain FORC 064	NZ_-CP022664.1	5405	4.8	complet
Escherichia coli APEC O2-211	NZ_-CP006834.2	5565	5.1	complet
Escherichia coli strain EC590	NZ_-CP016182.2	4788	4.6	complet
Escherichia coli str. Sanji	NZ_-CP011061.1	5702	4.9	complet
Escherichia coli strain SF-088	NZ_-CP012635.1	5474	5.0	complet
Escherichia coli strain 2009C-3133	NZ_-CP013025.1	5860	5.1	complet
Escherichia coli strain CFSAN029787	NZ_-CP011416.1	5633	4.9	complet
Escherichia coli O145 :H28 str. RM12581	NZ_-CP007136.1	6116	5.5	complet
Escherichia coli strain ST2747	NZ_-CP007393.1	5039	5.0	complet
Escherichia coli S88	NC_011742.1	5427	5.0	complet
Escherichia coli UMN026	NC_011751.1	5033	5.2	complet
Escherichia coli APEC O18	NZ_-CP006830.1	5149	5.0	complet
Escherichia coli strain D9	NZ_-CP010152.1	5006	4.6	complet
Escherichia coli strain D9 plasmid A	NZ_-CP010153.1	130	0.1	plasmide
Escherichia coli strain D9 plasmid B	NZ_-CP010154.1	87	0.08	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Escherichia coli strain D9 plasmid C	NZ_-CP010155.1	52	0.04	plasmide
Escherichia coli strain D9 plasmid D	NZ_-CP010156.1	7	0.005	plasmide
Escherichia coli strain S50	NZ_-CP010238.1	5239	4.8	complet
Escherichia coli strain M8	NZ_-CP010191.1	5232	4.8	complet
Escherichia coli strain M8 plasmid A	NZ_-CP010192.1	185	0.2	plasmide
Escherichia coli strain M8 plasmid B	NZ_-CP010193.1	57	0.04	plasmide
Escherichia coli strain M8 plasmid C	NZ_-CP010194.1	46	0.03	plasmide
Escherichia coli strain M8 plasmid D	NZ_-CP010195.1	11	0.005	plasmide
Escherichia coli strain C8	NZ_-CP010125.1	5233	4.9	complet
Escherichia coli strain C8 plasmid A	NZ_-CP010126.1	181	0.1	plasmide
Escherichia coli strain C8 plasmid B	NZ_-CP010127.1	52	0.04	plasmide
Escherichia coli strain C8 plasmid C	NZ_-CP010128.1	14	0.008	plasmide
Escherichia coli strain H3	NZ_-CP010167.1	4797	4.6	complet
Escherichia coli strain H3 plasmid A	NZ_-CP010168.1	59	0.05	plasmide
Escherichia coli strain H1	NZ_-CP010160.1	4994	4.8	complet
Escherichia coli strain H1 plasmid A	NZ_-CP010161.1	7	0.004	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Escherichia coli strain H1 plasmide B	NZ_-CP010162.1	4	0.004	plasmide
Escherichia coli strain S3	NZ_-CP010228.1	4773	4.6	complet
Escherichia coli O145 :H28 str. RM13516	NZ_-CP006262.1	5899	5.4	complet
Escherichia coli KO11	NC_016902.1	5217	4.9	complet
Escherichia coli KO11 plasmid pEKO1102	NC_016903.1	9	0.005	plasmide
Escherichia coli KO11 plasmid pEKO1101	NC_016904.1	136	0.1	plasmide
Escherichia coli IAI1	NC_011741.1	4786	4.7	complet
Escherichia coli APEC O1	NC_008563.1	5543	5.0	complet
Escherichia coli IAI39	NC_011750.1	4936	5.1	complet
Escherichia coli ST131 strain EC958 plasmid pEC958	NZ_-HG941719.1	173	0.1	plasmide
Escherichia coli ST131 strain EC958 plasmid pEC958B	NZ_-HG941720.1	4	0.004	plasmide
Escherichia coli UM146 plasmid pUM146	NC_017630.1	157	0.1	plasmide
Escherichia coli strain SF-088 plasmid pSF-088-1	NZ_-CP012636.1	192	0.1	plasmide
Escherichia coli strain SF-088 plasmid pSF-088-2	NZ_-CP012637.1	5	0.005	plasmide
Escherichia coli strain SF-088 plasmid pSF-088-3	NZ_-CP012638.1	3	0.004	plasmide
Escherichia coli O145 :H28 str. RM12581 plasmid pRM12581	NZ_-CP007137.1	97	0.06	plasmide
Escherichia coli O145 :H28 str. RM12581 plasmid pO145-12581	NZ_-CP007138.1	102	0.09	plasmide

Tableau 18 – Liste des 45 génomes utilisés pour construire le pangénome de *E. coli* (partiellement adapté de [91]).

B.3 Pangénome de *Campylobacter jejuni*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* NCTC 11168 = ATCC 700819	AL111168.1	1668	1.6	complet
Campylobacter jejuni subsp. doylei strain NCTC11951	LR134359.1	2052	1.9	complet
Campylobacter jejuni subsp. *jejuni* strain ATCC 33560	CP019838.1	1836	1.8	complet
Campylobacter jejuni strain NCTC11351	LN831025.1	1803	1.8	complet
Campylobacter jejuni subsp. *doylei* strain 269.97	CP000768.1	1972	1.8	complet
Campylobacter jejuni subsp. *doylei* strain FDAARGOS_295	CP027403.1	1969	1.8	complet
Campylobacter jejuni strain CFSAN054107	CP028185.1	1999	1.9	complet
Campylobacter jejuni strain CFSAN054107 plasmid pGMI16-002	CP028186.1	81	0.1	plasmide
Campylobacter jejuni strain NCTC13265	LR134498.1	1948	1.8	complet
Campylobacter jejuni strain HF5-4A-4	CP007188.1	1971	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain NADC 20827	CP045048.1	1956	1.9	complet
Campylobacter jejuni subsp. *jejuni* strain NADC 20827 plasmid p20827L	CP045046.1	54	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain NADC 20827 plasmid p20827S	CP045047.1	5	0	plasmide
Campylobacter jejuni strain NCTC13268	LR134497.1	1864	1.8	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain R14	CP005081.1	1943	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain HPC5	CP032316.1	1936	1.8	complet
Campylobacter jejuni strain CJ018CCUA	CP012221.1	1926	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain MTVDSCj16	CP017033.1	1866	1.8	complet
Campylobacter jejuni RM1221 strain RM1221	CP000025.1	1888	1.8	complet
Campylobacter jejuni strain FDAARGOS_421	CP023866.1	1887	1.8	complet
Campylobacter jejuni strain YH002	CP020776.1	1935	1.8	complet
Campylobacter jejuni strain YH002 plasmid pCJP002	CP020775.1	54	0	plasmide
Campylobacter jejuni strain FDAARGOS_262	CP022076.1	1824	1.8	complet
Campylobacter jejuni strain ZP3204	CP017856.1	1888	1.8	complet
Campylobacter jejuni strain ZP3204 plasmid pCJDM204L	CP017854.1	45	0	plasmide
Campylobacter jejuni strain ZP3204 plasmid pCJDM204S	CP017855.1	6	0	plasmide
Campylobacter jejuni strain TS1218	CP017860.1	1881	1.8	complet
Campylobacter jejuni strain TS1218 plasmid pCJDM218	CP017861.1	46	0	plasmide
Campylobacter jejuni subsp. *doylei* strain NCTC11924	LR134530.1	1845	1.8	complet
Campylobacter jejuni strain AR-0412	CP044173.1	1925	1.8	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain AR-0412 plasmid pAR-0412	CP044174.1	-	-	plasmide
Campylobacter jejuni strain CF-SAN032806	CP045789.1	1916	1.8	complet
Campylobacter jejuni strain CFSAN032806 plasmid pCF-SAN032806	CP045790.1	63	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain M129	CP007749.1	1863	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain M129 plasmid pTet-M129	CP007750.1	51	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain CLB104	CP034393.1	1855	1.8	complet
Campylobacter jejuni strain YQ2210	CP017859.1	1886	1.8	complet
Campylobacter jejuni strain YQ2210 plasmid pCJDM210L	CP017857.1	45	0	plasmide
Campylobacter jejuni strain YQ2210 plasmid pCJDM210S	CP017858.1	6	0	plasmide
Campylobacter jejuni strain CJ071CC464	CP012217.1	1851	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain 00-0949	CP010301.1	1955	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain 00-0949 plasmid pTet	CP010302.1	52	0	complet
Campylobacter jejuni subsp. *jejuni* strain 00-0949 plasmid pVir	CP010303.1	46	0	complet
Campylobacter jejuni strain IF1100	CP017863.1	1853	1.7	complet
Campylobacter jejuni strain IF1100 plasmid pCJDM100	CP017864.1	6	0	plasmide
Campylobacter jejuni strain YH003	CP041584.1	1813	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain 01-1512	CP010072.1	1954	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain 01-1512 plasmid pCj1	CP010073.1	51	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain 01-1512 plasmid pCj2	CP010074.1	46	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain 00-1597	CP010306.1	1801	1.7	complet
Campylobacter jejuni strain SCJK2	CP038862.1	1993	1.9	complet
Campylobacter jejuni strain SCJK2 plasmid p2	CP038864.1	49	0	plasmide
Campylobacter jejuni strain SCJK2 plasmid unnamed1	CP038863.1	101	0.1	plasmide
Campylobacter jejuni subsp. *doylei* strain NCTC11925	LS483295.1	1789	1.7	complet
Campylobacter jejuni strain CFSAN032806	CP023543.1	1870	1.8	complet
Campylobacter jejuni strain CFSAN032806 plasmid pCFSAN032806	CP023544.1	66	0	plasmide
Campylobacter jejuni strain CJ067CC45	CP012206.1	1823	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain 00-2544	CP006709.2	1860	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain 00-2544 plasmid unnamed	CP006710.1	44	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain 00-2538	CP006707.2	1805	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain 00-2425	CP006729.2	1801	1.7	complet
Campylobacter jejuni strain AR-0414	CP044169.1	1778	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain YH001	CP010058.1	1809	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain 14980A	CP017029.1	1799	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain 14980A plasmid pCJ14980A	CP017030.1	51	0	plasmide
Campylobacter jejuni strain 12567	CP028909.1	1774	1.7	complet
Campylobacter jejuni strain 32488	CP006006.1	1765	1.7	complet
Campylobacter jejuni strain FJ3124	CP017862.1	1777	1.7	complet
Campylobacter jejuni strain CJ074CC443	CP012216.1	1764	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain F38011	CP006851.1	1773	1.7	complet
Campylobacter jejuni strain FDAARGOS_422	CP023867.1	1771	1.7	complet
Campylobacter jejuni strain AR-0419	CP044162.1	1765	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain MTVDSCj13	CP017032.1	1746	1.7	complet
Campylobacter jejuni strain NCTC 12660	CP028910.1	1756	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain huA17	CP028372.1	1910	1.8	complet
Campylobacter jejuni subsp. *jejuni* strain huA17 plasmid unnamed1	CP028373.1	152	0.1	plasmide
Campylobacter jejuni strain WP2202	CP014742.1	1880	1.8	complet
Campylobacter jejuni strain WP2202 plasmid pCJDM202	CP014743.1	116	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain S3	CP001960.1	1818	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain S3 plasmid pTet	CP001961.1	50	0	plasmide
Campylobacter jejuni strain CJ017CCUA	CP012212.1	1772	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain 00-2426	CP006708.2	1750	1.7	complet
Campylobacter jejuni strain CJ677CC520	CP010501.1	1777	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain D42a	CP007751.1	1809	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain D42a plasmid pTet-D42a	CP007752.1	54	0	plasmide
Campylobacter jejuni strain CJ677CC032	CP010496.1	1770	1.7	complet
Campylobacter jejuni strain CJ677CC538	CP010495.1	1775	1.7	complet
Campylobacter jejuni strain NS4-5-1	CP007192.1	1723	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain RM1285	CP015209.1	1744	1.7	complet
Campylobacter jejuni strain OD267	CP014744.1	1921	1.8	complet
Campylobacter jejuni strain OD267 plasmid pCJDM67 L	CP014745.1	108	0.1	plasmide
Campylobacter jejuni strain OD267 plasmid pCJDM67 S	CP014746.1	46	0	plasmide
Campylobacter jejuni strain CJ677CC073	CP010475.1	1773	1.7	complet
Campylobacter jejuni strain CJ677CC014	CP010502.1	1777	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain 00-6200	CP010307.1	1743	1.7	complet
Campylobacter jejuni strain NS4-1-1	CP007191.1	1719	1.7	complet
Campylobacter jejuni strain CJ677CC002	CP010472.1	1773	1.7	complet
Campylobacter jejuni strain 4031	HG428754.1	1732	1.7	complet
Campylobacter jejuni strain CJ677CC100	CP010462.1	1769	1.7	complet
Campylobacter jejuni strain CJ677CC521	CP010476.1	1779	1.7	complet
Campylobacter jejuni strain CJ677CC524	CP010480.1	1773	1.7	complet
Campylobacter jejuni strain CJ677CC526	CP010477.1	1754	1.7	complet
Campylobacter jejuni strain CJ677CC523	CP010508.1	1765	1.7	complet
Campylobacter jejuni strain NCTC13255	LR134499.1	1740	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain RM3420	CP017456.1	1733	1.7	complet
Campylobacter jejuni strain NCTC13261	LR134500.1	1753	1.7	complet
Campylobacter jejuni strain CJ677CC531	CP010492.1	1767	1.7	complet
Campylobacter jejuni strain CJ677CC036	CP010479.1	1745	1.7	complet
Campylobacter jejuni strain NCTC13266	LR134496.1	1738	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain RM3196	CP012690.1	1738	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain RM3197	CP012689.1	1740	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain CJ677CC024	CP010467.1	1745	1.7	complet
Campylobacter jejuni strain RM1246-ERRC	CP022470.1	1781	1.7	complet
Campylobacter jejuni strain RM1246-ERRC plasmid pRM1246_ERRC	CP022471.1	53	0	plasmide
Campylobacter jejuni strain CJ066CC508	CP012224.1	1713	1.7	complet
Campylobacter jejuni strain CJ677CC533	CP010458.1	1760	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC12109	LR134505.1	1710	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain MTVDSCj07	CP017031.1	1708	1.7	complet
Campylobacter jejuni strain CJ515CC45	CP012210.1	1741	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain MTVDSCj20	CP008787.1	1711	1.7	complet
Campylobacter jejuni strain RM3194	CP014344.1	1782	1.7	complet
Campylobacter jejuni strain RM3194 plasmid unnamed	CP014345.1	80	0.1	plasmide
Campylobacter jejuni strain FDAARGOS_265	CP022079.1	1761	1.7	complet
Campylobacter jejuni strain FDAARGOS_265 plasmid unnamed1	CP022078.1	54	0	plasmide
Campylobacter jejuni strain CJ677CC530	CP010489.1	1743	1.6	complet
Campylobacter jejuni strain 11168H/araE	CP022559.1	1715	1.6	complet
Campylobacter jejuni strain AR-0415	CP044167.1	1757	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain AR-0415 plasmid pAR-0415	CP044168.1	-	-	plasmide
Campylobacter jejuni strain BfR-CA-14430	CP043763.1	1746	1.7	complet
Campylobacter jejuni strain BfR-CA-14430 plasmid pBfR-CA-14430	CP043764.1	45	0	plasmide
Campylobacter jejuni strain 11168H/lacY	CP022439.1	1709	1.6	complet
Campylobacter jejuni strain CJ677CC528	CP010500.1	1728	1.6	complet
Campylobacter jejuni strain CJ677CC064	CP010468.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC062	CP010493.1	1721	1.6	complet
Campylobacter jejuni strain NS4-9-1	CP007193.1	1700	1.6	complet
Campylobacter jejuni strain CJ677CC527	CP010506.1	1718	1.6	complet
Campylobacter jejuni strain CJ677CC532	CP010490.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC519	CP010471.1	1713	1.6	complet
Campylobacter jejuni strain CJ677CC013	CP010461.1	1718	1.6	complet
Campylobacter jejuni strain CJ677CC522	CP010463.1	1714	1.6	complet
Campylobacter jejuni strain CJ677CC536	CP010474.1	1715	1.6	complet
Campylobacter jejuni strain CJ677CC525	CP010469.1	1711	1.6	complet
Campylobacter jejuni subsp. *jejuni* NCTC 11168-GSv strain NCTC 11168-rNRC	CP006689.1	1715	1.6	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain NCTC 11168-BN148	HE978252.1	1708	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC 11168-K12E5	CP006685.1	1710	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC 11168-Kf1	CP006686.1	1710	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC 11168-mcK12E5	CP006687.1	1712	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC 11168-mfK12E5	CP006688.1	1711	1.6	complet
Campylobacter jejuni strain FDAARGOS_263	CP022077.1	1714	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC11168	LS483362.1	1715	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain NCTC10983	LR134511.1	1712	1.6	complet
Campylobacter jejuni strain CJ677CC094	CP010464.1	1714	1.6	complet
Campylobacter jejuni strain CJ677CC040	CP010510.1	1713	1.6	complet
Campylobacter jejuni strain AR-0413	CP044171.1	1819	1.7	complet
Campylobacter jejuni strain AR-0413 plasmid pAR-0413-1	CP044170.1	-	-	plasmide
Campylobacter jejuni strain AR-0413 plasmid pAR-0413-2	CP044172.1	-	-	plasmide
Campylobacter jejuni strain CJ677CC016	CP010481.1	1729	1.6	complet
Campylobacter jejuni strain CJ677CC034	CP010484.1	1723	1.6	complet
Campylobacter jejuni strain CJ677CC540	CP010509.1	1720	1.6	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain CJ677CC085	CP010504.1	1722	1.6	complet
Campylobacter jejuni strain CJ677CC539	CP010457.1	1729	1.6	complet
Campylobacter jejuni strain CJ677CC041	CP010482.1	1721	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain CG8421	CP005388.1	1715	1.6	complet
Campylobacter jejuni strain RM1285	CP012696.1	1700	1.6	complet
Campylobacter jejuni strain CJ677CC026	CP010470.1	1725	1.6	complet
Campylobacter jejuni strain CJ677CC039	CP010503.1	1722	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain PT14	CP003871.4	1694	1.6	complet
Campylobacter jejuni strain CJ677CC047	CP010459.1	1720	1.6	complet
Campylobacter jejuni strain CJ677CC052	CP010505.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC061	CP010511.1	1720	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain IA3902	CP001876.1	1749	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain IA3902 plasmid pVir	CP001877.1	53	0	plasmide
Campylobacter jejuni strain NCTC12851	LR134507.1	1685	1.6	complet
Campylobacter jejuni strain CJ031CC45	CP012211.1	1723	1.6	complet
Campylobacter jejuni strain NCTC 12664	CP028912.1	1694	1.6	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain CJ677CC541	CP010466.1	1713	1.6	complet
Campylobacter jejuni strain CJ677CC033	CP010497.1	1718	1.6	complet
Campylobacter jejuni strain NCTC13257	LR134502.1	1684	1.6	complet
Campylobacter jejuni strain CJ677CC012	CP010487.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC534	CP010473.1	1719	1.6	complet
Campylobacter jejuni strain CJ677CC535	CP010483.1	1723	1.6	complet
Campylobacter jejuni strain CJ677CC542	CP010499.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC537	CP010498.1	1720	1.6	complet
Campylobacter jejuni strain CJ677CC092	CP010488.1	1717	1.6	complet
Campylobacter jejuni strain CJ677CC058	CP010460.1	1720	1.6	complet
Campylobacter jejuni strain CJ677CC059	CP010494.1	1714	1.6	complet
Campylobacter jejuni strain CJ677CC078	CP010507.1	1715	1.6	complet
Campylobacter jejuni strain CJ677CC529	CP010491.1	1715	1.6	complet
Campylobacter jejuni strain CJ677CC086	CP010485.1	1711	1.6	complet
Campylobacter jejuni strain FORC_046	CP017229.1	1775	1.7	complet
Campylobacter jejuni strain FORC_046 plasmid pFORC46.1	CP017230.1	62	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni strain FORC_046 plasmid pFORC46.2	CP017231.1	41	0	plasmide
Campylobacter jejuni strain HF5-7-1	CP007190.1	1671	1.6	complet
Campylobacter jejuni strain HF5-5-1	CP007189.1	1674	1.6	complet
Campylobacter jejuni strain CJ677CC095	CP010486.1	1713	1.6	complet
Campylobacter jejuni subsp. textitjejuni strain 81116; NCTC 11828	CP000814.1	1677	1.6	complet
Campylobacter jejuni strain FORC_083	CP028933.1	1747	1.7	complet
Campylobacter jejuni strain FORC_083 plasmid pFORC_083_2	CP028934.1	70	0	plasmide
Campylobacter jejuni strain FORC_083 plasmid pFORC_083_3	CP028935.1	1	0	plasmide
Campylobacter jejuni strain CJ677CC008	CP010465.1	1707	1.6	complet
Campylobacter jejuni strain FDAARGOS_266	CP022080.1	1673	1.6	complet
Campylobacter jejuni strain 81-176_G1_B0	CP022440.1	1667	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain ATCC 35925	CP020045.1	1654	1.6	complet
Campylobacter jejuni strain CJM1cam	CP012149.1	1667	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain M1	CP001900.1	1667	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain 81-176	CP000538.1	1771	1.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Campylobacter jejuni subsp. *jejuni* strain 81-176 plasmid pTet	CP000549.1	52	0	plasmide
Campylobacter jejuni subsp. *jejuni* strain 81-176 plasmid pVir	CP000550.1	53	0	plasmide
Campylobacter jejuni strain 81-176_G1_B7	CP022551.1	1671	1.6	complet
Campylobacter jejuni strain NCTC 12661	CP028911.1	1656	1.6	complet
Campylobacter jejuni strain CJ513CC45	CP012213.1	1685	1.6	complet
Campylobacter jejuni strain NCTC12662	CP019965.1	1657	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain 35925	CP010906.1	1652	1.6	complet
Campylobacter jejuni strain BC	CP032522.1	1680	1.6	complet
Campylobacter jejuni strain CJ088CC52	CP012214.1	1674	1.6	complet
Campylobacter jejuni strain CJ090CC1332	CP012220.1	1638	1.6	complet
Campylobacter jejuni subsp. *jejuni* strain ICDCCJ07001	CP002029.1	1669	1.7	complet
Campylobacter jejuni subsp. *jejuni* strain ICDCCJ07001 plasmid pTet	CP002030.1	37	0	plasmide
Campylobacter jejuni strain CJ677CC010	CP010478.1	1751	1.7	complet

Tableau 19 – Liste des 219 génomes et plasmides utilisés pour construire le pangénome de *C. jejuni*.

B.4 Pangénome de *Lactococcus lactis*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* strain A12	LT599049.1	2700	2.7	complet
Lactococcus lactis subsp. *lactis* strain A12 plasmid pA12-1	LT599050.1	6	0	plasmide
Lactococcus lactis subsp. *lactis* strain A12 plasmid pA12-2	LT599051.1	8	0	plasmide
Lactococcus lactis subsp. *lactis* strain A12 plasmid pA12-3	LT599052.1	58	0	plasmide
Lactococcus lactis subsp. *lactis* strain A12 plasmid pA12-4	LT599053.1	85	0.1	plasmide
Lactococcus lactis subsp. *lactis* strain KF147	CP001834.1	2616	2.6	complet
Lactococcus lactis subsp. *lactis* strain KF147 plasmid pKF147A	CP001835.1	38	0	plasmide
Lactococcus lactis subsp. *lactis* strain 14B4	CP028160.1	2626	2.6	complet
Lactococcus lactis subsp. *lactis* strain 14B4 plasmid p14B4	CP028161.1	63	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06	CP015902.1	2714	2.7	complet
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06A	CP016734.1	43	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06B	CP016735.1	55	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06C	CP016736.1	29	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06D	CP034579.1	10	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06E	CP034580.1	6	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC06 plasmid pUC06F	CP034581.1	27	0	plasmide
Lactococcus lactis subsp. *lactis* strain NCDO 2118	CP009054.1	2551	2.6	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* strain NCDO 2118 plasmid pNCDO2118	CP009055.1	44	0	plasmide
Lactococcus lactis subsp. *lactis* strain 275	CP015897.1	2784	2.8	complet
Lactococcus lactis subsp. *lactis* strain 275 plasmid p275A	CP016699.1	104	0.1	plasmide
Lactococcus lactis subsp. *lactis* strain 275 plasmid p275B	CP016700.1	65	0	plasmide
Lactococcus lactis subsp. *lactis* strain 275 plasmid p275C	CP016701.1	62	0	plasmide
Lactococcus lactis subsp. *lactis* strain 275 plasmid p275D	CP016702.1	60	0	plasmide
Lactococcus lactis subsp. *lactis* strain S0	CP010050.1	2493	2.5	complet
Lactococcus lactis subsp. *lactis* strain 229	CP015896.1	2663	2.6	complet
Lactococcus lactis subsp. *lactis* strain 229 plasmid p229A	CP016694.1	59	0	plasmide
Lactococcus lactis subsp. *lactis* strain 229 plasmid p229B	CP016695.1	29	0	plasmide
Lactococcus lactis subsp. *lactis* strain 229 plasmid p229C	CP016696.1	29	0	plasmide
Lactococcus lactis subsp. *lactis* strain 229 plasmid p229D	CP016697.1	8	0	plasmide
Lactococcus lactis subsp. *lactis* strain 229 plasmid p229E	CP016698.1	51	0	plasmide
Lactococcus lactis subsp. *lactis* strain F44	CP024954.1	2350	2.4	complet
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03	CP020604.1	2560	2.5	complet
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd1	CP020605.1	10	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd2	CP020606.1	16	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd3	CP020607.1	2	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd4	CP020608.1	11	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd5	CP020609.1	6	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd6	CP020610.1	3	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain FM03 plasmid pLd7	CP020611.1	30	0	plasmide
Lactococcus lactis subsp. *lactis* strain UL8	CP015908.1	2495	2.5	complet
Lactococcus lactis subsp. *lactis* strain UL8 plasmid pUL8A	CP016704.1	6	0	plasmide
Lactococcus lactis subsp. *lactis* strain UL8 plasmid pUL8B	CP016705.1	30	0	plasmide
Lactococcus lactis subsp. *lactis* strain UL8 plasmid pUL8C	CP016706.1	3	0	plasmide
Lactococcus lactis subsp. *lactis* IO-1 strain IO-1	AP012281.1	2329	2.4	complet
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96	CP043523.1	2706	2.6	complet
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_01	CP043525.1	15	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_02	CP043526.1	91	0.1	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_03	CP043527.1	35	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_04	CP043528.1	12	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_05	CP043524.1	56	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_06	CP043518.1	26	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_07	CP043519.1	12	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_08	CP043520.1	10	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_09	CP043521.1	5	0	plasmide
Lactococcus lactis subsp. *lactis* bv. diacetylactis strain SD96 plasmid pSD96_10	CP043522.1	17	0	plasmide
Lactococcus lactis subsp. *lactis* strain G423	CP024958.1	2345	2.4	complet
Lactococcus lactis subsp. *lactis* strain CV56	CP002365.1	2516	2.5	complet
Lactococcus lactis subsp. *lactis* strain CV56 plasmid pCV56A	CP002366.1	39	0	plasmide
Lactococcus lactis subsp. *lactis* strain CV56 plasmid pCV56B	CP002367.1	28	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* strain CV56 plasmid pCV56C	CP002368.1	29	0	plasmide
Lactococcus lactis subsp. *lactis* strain CV56 plasmid pCV56D	CP002369.1	8	0	plasmide
Lactococcus lactis subsp. *lactis* strain CV56 plasmid pCV56E	CP002370.1	3	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC063	CP015905.1	2570	2.5	complet
Lactococcus lactis subsp. *lactis* strain UC063 plasmid pUC063A	CP016715.1	79	0.1	plasmide
Lactococcus lactis subsp. *lactis* strain UC063 plasmid pUC063B	CP016716.1	41	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC063 plasmid pUC063C	CP016717.1	15	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC063 plasmid pUC063D	CP016718.1	10	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC063 plasmid pUC063E	CP016719.1	11	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11	CP015904.1	2483	2.5	complet
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pCU11E	CP034572.1	8	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pUC11A	CP016720.1	65	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pUC11B	CP016721.1	52	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pUC11C	CP016722.1	18	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pUC11D	CP016723.1	17	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC11 plasmid pUC11F	CP016725.1	4	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* strain UC08	CP015903.1	2502	2.5	complet
Lactococcus lactis subsp. *lactis* strain UC08 plasmid pUC08A	CP016726.1	102	0.1	plasmide
Lactococcus lactis subsp. *lactis* strain UC08 plasmid pUC08B	CP016727.1	52	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC08 plasmid pUC08C	CP016728.1	21	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC08 plasmid pUC08D	CP034577.1	4	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC08 plasmid pUC08E	CP034578.1	7	0	plasmide
Lactococcus lactis subsp. *lactis* Il1403 strain IL1403	AE005176.1	2406	2.4	complet
Lactococcus lactis subsp. *lactis* strain G50	CP025500.1	2300	2.3	complet
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325	CP006766.1	2869	2.8	complet
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid 1	CP006767.1	5	0	plasmide
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid 2	CP007042.1	2	0	plasmide
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid 3	CP007043.1	3	0	plasmide
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid unnamed4	CP029291.1	14	0	plasmide
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid unnamed5	CP029292.1	63	0	plasmide
Lactococcus lactis subsp. *lactis* strain KLDS 4.0325 plasmid unnamed6	CP029293.1	119	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Lactococcus lactis subsp. *lactis* strain UC77	CP015906.1	2825	2.7	complet
Lactococcus lactis subsp. *lactis* strain UC77 plasmid pUC77A	CP016713.1	7	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC77 plasmid pUC77B	CP016714.1	66	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC77 plasmid pUC77C	CP034573.1	58	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC77 plasmid pUC77D	CP034574.1	47	0	plasmide
Lactococcus lactis subsp. *lactis* strain UC77 plasmid pUC77E	CP034575.1	7	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184	CP015895.1	2441	2.4	complet
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184A	CP016691.1	13	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184B	CP016692.1	6	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184C	CP016693.1	14	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184D	CP034584.1	3	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184E	CP034585.1	4	0	plasmide
Lactococcus lactis subsp. *lactis* strain 184 plasmid p184F	CP034586.1	8	0	plasmide
Lactococcus lactis subsp. *lactis* strain C10	CP015898.1	2437	2.4	complet
Lactococcus lactis subsp. *lactis* strain C10 plasmid pC10A	CP016703.1	4	0	plasmide
Lactococcus lactis subsp. *lactis* strain C10 plasmid pC10B	CP034582.1	48	0	plasmide

Organisme/nom	Numéro d'accession NCBI	Nb de gènes	Taille (Mb)	Niveau
Lactococcus lactis subsp. *lactis* strain C10 plasmid pC10C	CP034583.1	5	0	plasmide

Tableau 20 – Liste des 107 génomes et plasmides utilisés pour construire le pangénome de *L. lactis*.

B.5 Pangénome de *Listeria monocytogenes*

Organisme/nom	Numéro d'accession NCBI	Nb de gènes	Taille (Mb)	Niveau
Listeria monocytogenes stain J1816	CP006047.2	2947	2.9	complet
Listeria monocytogenes strain C1-387	CP006591.1	3052	3.0	complet
Listeria monocytogenes strain J2-064	CP006592.1	2946	2.9	complet
Listeria monocytogenes strain J2-031	CP006593.1	3020	3.0	complet
Listeria monocytogenes strain R2-502	CP006594.1	3142	3.0	complet
Listeria monocytogenes strain J1-108	CP006596.2	2981	2.9	complet
Listeria monocytogenes strain N1-011A	CP006597.1	3160	3.1	complet
Listeria monocytogenes strain J1776	CP006598.1	2996	3.0	complet
Listeria monocytogenes strain J1817	CP006599.1	2998	3.0	complet
Listeria monocytogenes strain J1926	CP006600.1	2995	3.0	complet
Listeria monocytogenes strain CFSAN023459	CP014252.1	3113	3.0	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Listeria monocytogenes strain CFSAN023459 plasmid CFSAN023459_01	CP014253.1	17	0	plasmide
Listeria monocytogenes strain CFSAN023459 plasmid CFSAN023459_02	CP014254.1	62	0	plasmide
Listeria monocytogenes strain FDA00006907	CP022020.1	3066	3.0	complet
Listeria monocytogenes strain FDA00006907 plasmid pCFSAN021445	CP022021.1	168	0.1	plasmide
Listeria monocytogenes strain AUSMDU00000224	CP045972.1	2851	2.8	complet
Listeria monocytogenes strain AUSMDU00000224 plasmid pAUSMDU00000224_01	CP045973.1	62	0	plasmide
Listeria monocytogenes serotype 1-2c str. SLCC2372 plasmid pLM1-2cUG1	FR667691.1	54	0	plasmide
Listeria monocytogenes strain SLCC2372 serotype 1/2c	FR733648.1	2936	2.9	complet
Listeria monocytogenes R479a	HG813247.1	2982	2.9	complet
Listeria monocytogenes R479a plasmid pLMR479a	HG813248.1	92	0.1	plasmide
Listeria monocytogenes serotype 4b str. F2365	NC_002973.6	2883	2.9	complet
Listeria monocytogenes EGD-e chromosome	NC_003210.1	2867	2.8	complet
Listeria monocytogenes HCC23	NC_011660.1	3031	3.0	complet
Listeria monocytogenes serotype 4b str. CLIP 80459	NC_012488.1	2882	2.9	complet
Listeria monocytogenes strain 08-5578	NC_013766.2	3144	3.1	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Listeria monocytogenes strain 08-5923	NC_013768.1	3024	3.0	complet
Listeria monocytogenes strain L99	NC_017529.1	3027	3.0	complet
Listeria monocytogenes strain M7	NC_017537.1	3031	3.0	complet
Listeria monocytogenes strain 10403S	NC_017544.1	2955	2.9	complet
Listeria monocytogenes strain J0161	NC_017545.1	3077	3.0	complet
Listeria monocytogenes strain FSL R2-561	NC_017546.1	2997	3.0	complet
Listeria monocytogenes strain Finland 1998	NC_017547.1	2902	2.9	complet
Listeria monocytogenes strain 07PF0776	NC_017728.1	2939	2.9	complet
Listeria monocytogenes strain ATCC 19117 serotype 4d	NC_018584.1	2976	2.9	complet
Listeria monocytogenes strain SLCC2378	NC_018585.1	2919	2.9	complet
Listeria monocytogenes strain SLCC2540	NC_018586.1	2951	2.9	complet
Listeria monocytogenes strain SLCC2755	NC_018587.1	3045	3.0	complet
Listeria monocytogenes strain SLCC2372	NC_018588.1	3053	3.0	complet
Listeria monocytogenes strain SLCC2479	NC_018589.1	2997	3.0	complet
Listeria monocytogenes strain SLCC2376	NC_018590.1	2810	2.8	complet
Listeria monocytogenes serotype 7 str. SLCC2482	NC_018591.1	2968	2.9	complet
Listeria monocytogenes strain SLCC5850	NC_018592.1	2911	2.9	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Listeria monocytogenes strain SLCC7179	NC_018593.1	2896	2.9	complet
Listeria monocytogenes strain L312	NC_018642.1	2880	2.9	complet
Listeria monocytogenes serotype 4b str. LL195	NC_019556.1	2936	2.9	complet
Listeria monocytogenes strain La111	NC_020557.1	2884	2.9	complet
Listeria monocytogenes strain N53-1	NC_020558.1	2894	2.9	complet

Tableau 21 – Liste des 48 génomes et plasmides utilisés pour construire le pangénome de *L. monocytogenes* (partiellement adapté de [94]).

B.6 Pangénome de *Mycobacterium tuberculosis*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Mycobacterium tuberculosis strain F1	CP010329	4400	4.43	complet
Mycobacterium tuberculosis strain F28	CP010330	4366	4.42	complet
Mycobacterium tuberculosis strain H37Ra	NC_009525	4348	4.42	complet
Mycobacterium tuberculosis strain Erdman	NC_020559	4373	4.39	complet
Mycobacterium tuberculosis strain 22103	CP010339	4345	4.4	complet
Mycobacterium tuberculosis strain 22115	CP010337	4356	4.4	complet
Mycobacterium tuberculosis strain 37004	CP010338	4375	4.42	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Mycobacterium tuberculosis strain KZN 4207	NC_016768	4324	4.4	complet
Mycobacterium tuberculosis strain KZN 605	NC_018078	4326	4.4	complet
Mycobacterium tuberculosis strain KZN 1435	NC_012943	4333	4.4	complet
Mycobacterium tuberculosis strain Haarlem	NC_022350	4322	4.41	complet
Mycobacterium tuberculosis strain F11	NC_009565	4352	4.42	complet
Mycobacterium tuberculosis strain H37Rv	NC_000962	4336	4.41	complet
Mycobacterium tuberculosis strain CDC1551	NC_002755	4351	4.4	complet
Mycobacterium tuberculosis strain 7199-99	NC_020089	4344	4.42	complet
Mycobacterium tuberculosis strain CTRI-2	NC_017524	4331	4.4	complet
Mycobacterium tuberculosis strain Kurono	NZ_-AP014573	4342	4.42	complet
Mycobacterium tuberculosis strain 26105	CP010340	4393	4.43	complet
Mycobacterium tuberculosis strain 2242	CP010335	4428	4.42	complet
Mycobacterium tuberculosis strain 2279	CP010336	4400	4.41	complet
Mycobacterium tuberculosis strain NITR203	NC_021054	4442	4.41	complet
Mycobacterium tuberculosis strain HKBS1	CP002871	4343	4.41	complet
Mycobacterium tuberculosis strain CCDC5079	NC_021251	4354	4.41	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Mycobacterium tuberculosis strain 49-02	HG813240	4350	4.42	complet
Mycobacterium tuberculosis strain 96075	CP009426	4327	4.4	complet
Mycobacterium tuberculosis strain BT1	CP002883	4337	4.4	complet
Mycobacterium tuberculosis strain BT2	CP002882	4343	4.4	complet
Mycobacterium tuberculosis strain CCDC5180	CP002885	4353	4.41	complet
Mycobacterium tuberculosis strain 323	CP010873	4403	4.41	complet
Mycobacterium tuberculosis strain ZMC13-88	CP009101	4364	4.41	complet
Mycobacterium tuberculosis strain ZMC13-264	CP009100	4365	4.41	complet
Mycobacterium tuberculosis strain KIT87190	CP007809	4341	4.41	complet
Mycobacterium tuberculosis strain K	CP007803	4334	4.39	complet
Mycobacterium tuberculosis strain 96121	CP009427	4375	4.41	complet
Mycobacterium tuberculosis strain EAI5	NC_021740	4322	4.39	complet
Mycobacterium tuberculosis strain NITR206	NC_021194	4417	4.39	complet
Mycobacterium tuberculosis strain PanR0201	CM002049.1	4352	4.41	wgs
Mycobacterium tuberculosis strain PanR0802	CM002050.1	4339	4.4	wgs
Mycobacterium tuberculosis strain PanR1005	CM002051.1	4322	4.39	wgs

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Mycobacterium tuberculosis strain PanR0704	CM002048.1	4332	4.39	wgs
Mycobacterium tuberculosis strain Beijing/NITR203	CP005082.1	4461	4.41	complet
Mycobacterium tuberculosis strain EAI5/NITR206	CP005387.1	4433	4.39	complet
Mycobacterium tuberculosis strain EAI5	CP006578.1	4341	4.41	complet
Mycobacterium tuberculosis strain UT205	HE608151.1	4212	4.29	complet
Mycobacterium tuberculosis strain 7199-99	HE663067.1	3994	4.0	complet
Mycobacterium tuberculosis strain H37Rv	AL123456.3	4367	4.41	complet
Mycobacterium tuberculosis strain TCDC11	CP046728.2	4121	4.2	complet

Tableau 22 – Liste des 47 génomes utilisés pour construire le pangénome de *M. tuberculosis* (partiellement adapté de [95] et [96]).

B.7 Pangénome de *Salmonella typhimurium*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain LT2	AE006468.2	4714	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain LT2 plasmid pSLT	AE006471.2	102	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ATCC 13311	CP009102.1	4658	4.8	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ATCC 13311 plasmid pSTY1	CP009103.1	44	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain L-3553	AP014565.1	5101	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain L-3553 plasmid pST3553	AP014566.1	159	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO4698-09	LN999997.1	4939	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00008979	CP045952.1	5188	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00008979 plasmid pAUSMDU00008979_01	CP045953.1	187	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain TW-Stm6	CP019649.1	5176	5.3	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain TW-Stm6 plasmid p275_TW-Stm6	CP019647.1	287	0.3	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain TW-Stm6 plasmid p4_TW-Stm6	CP019648.1	4	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain WW012	CP022168.1	5020	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain WW012 plasmid pWW012	CP022169.1	0	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain VNB151-sc-2315230	LT795114.1	5118	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain VNB151-sc-2315230 plasmid p1	LT795115.1	0	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain VNB151-sc-2315230 plasmid p2	LT795116.1	0	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain TJWQ005	CP040458.1	5138	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain TJWQ005 plasmid unnamed1	CP040457.1	0	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 81741	CP019442.1	5293	5.3	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 81741 plasmid unnamed1	CP019443.1	266	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 81741 plasmid unnamed2	CP019444.1	114	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_317	CP027410.1	5204	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_317 plasmid unnamed	CP027409.1	273	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-8290	CP040568.1	5037	5.1	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-8290 plasmid pCFSAN059542	CP040569.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00010530	CP045947.1	4849	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00010530 plasmid pAUSMDU00010530_01	CP045948.1	6	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain NCCP 16207	CP041976.1	5004	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain NCCP 16207 plasmid unnamed1	CP041974.1	151	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain NCCP 16207 plasmid unnamed2	CP041975.1	23	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain T000240	AP011957.1	4956	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain T000240 plasmid pSTMDT12_L	AP011958.1	139	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain T000240 plasmid pSTMDT12_S	AP011959.1	10	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014862	CP039591.1	4909	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014862 plasmid p11-0500.1	CP039592.1	121	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1808	CP014969.1	4919	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1808 plasmid pSTY1-1808	CP014970.1	112	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2009K-1640	CP014975.1	4998	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2009K-1640 plasmid pSTY1-2009K-1640	CP014976.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain DT104	HF937208.1	4918	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain DT104 plasmid II	HF937209.1	106	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 138736	CP007581.1	4997	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 138736 plasmid unnamed	CP007582.1	105	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU15	CP014358.1	4914	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU15 plasmid pYU15_94	CP014359.1	114	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014856	CP039576.1	4974	5.1	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014856 plasmid p10-3857.1	CP039577.1	191	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014856 plasmid p10-3857.2	CP039578.1	6	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain sg_wt7	CP036168.1	4943	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* str. USDA-ARS-USMARC-1810 strain USDA-ARS-USMARC-1810	CP014982.2	4855	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014879	CP040321.1	4907	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014879 plasmid p16-6397.1	CP040322.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014879 plasmid p16-6397.2	CP040323.1	9	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO2	CP014356.1	4888	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO2 plasmid pSO2_STV	CP014357.1	115	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain B3589	CP034968.1	4962	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain B3589 plasmid p3589	CP034969.1	7	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO3	CP014536.1	4882	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO3 plasmid pSO3_STV	CP014537.1	113	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2011K-1702	CP014967.1	4944	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2011K-1702 plasmid pSTY1-2011K-1702	CP014968.1	112	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7399	CP040562.1	4835	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7399 plasmid pCFSAN059545	CP040563.1	5	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39	CP011428.1	5111	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_2.7	CP011435.1	4	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_4.2	CP011434.1	4	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_4.8	CP011433.1	5	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_5.1	CP011432.1	3	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_89	CP011430.1	109	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_IncA/C	CP011429.1	178	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain YU39 plasmid pYU39_IncX	CP011431.1	48	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_321	CP022070.2	4986	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_321 plasmid unnamed1	CP022071.2	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_321 plasmid unnamed2	CP022072.2	65	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC H2662	CP014979.2	4938	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC H2662 plasmid pSTY1-H2662	CP014980.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067216	CP028318.1	4933	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067216 plasmid pSC-09-1	CP028319.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7699	CP040564.1	4913	5.0	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7699 plasmid pCFSAN059544	CP040565.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain D23580	LS997973.1	5037	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain D23580 plasmid D23580_liv_-pBT1	LS997975.1	-	-	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain D23580 plasmid D23580_liv_-pBT2	LS997976.1	-	-	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain D23580 plasmid D23580_liv_-pBT3	LS997977.1	-	-	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain D23580 plasmid D23580_liv_-pSLT-BT	LS997974.1	-	-	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* str. D23580 strain D23580	FN424405.1	4729	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UGA14	CP021462.1	5249	5.4	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UGA14 plasmid pUGA14_1	CP021463.1	389	0.4	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UGA14 plasmid pUGA14_2	CP021464.1	154	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UGA14 plasmid pUGA14_3	CP021465.1	4	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UGA14 plasmid pUGA14_4	CP021466.1	2	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST4/74	CP002487.1	4941	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST4/74 plasmid TY474p1	CP002488.1	116	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST4/74 plasmid TY474p2	CP002489.1	90	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST4/74 plasmid TY474p3	CP002490.1	11	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL1344	FQ312003.1	4941	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL1344 plasmid pCol1B9_SL1344	HE654725.1	101	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL1344 plasmid pRSF1010_SL1344	HE654726.1	12	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL1344 plasmid pSLT_SL1344	HE654724.1	101	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 798	CP003386.1	4897	5.0	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 798 plasmid p798_93	CP003387.1	104	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028S substr. GXS275	CP043399.1	4772	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM13672	CP047323.1	4825	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM13672 plasmid pRM13672	CP047324.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028S	CP001363.1	4818	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028S plasmid unnamed	CP001362.1	102	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* str. 14028S strain 14028S substr. JY996	CP043400.1	4769	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AR_0031	CP026700.1	4825	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AR_0031 plasmid unitig_1_pilon	CP026701.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PIR00538	CP025555.1	4826	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PIR00538 plasmid pPIR00538	CP025556.1	120	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028	CP034479.1	4889	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028 plasmid unnamed	CP034480.1	119	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 01ST04081	CP029840.1	5093	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 01ST04081 plasmid p01ST04081A	CP029841.1	189	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 01ST04081 plasmid p01ST04081B	CP029842.1	139	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ATCC 14028	CP034230.1	5006	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ATCC 14028 plasmid pATCC14028	CP034231.1	124	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028S substr. GXS259	CP043401.1	4771	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 14028S substr. GXS254	CP043402.1	4774	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00010529	CP045949.1	4922	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00010529 plasmid pAUSMDU00010529_01	CP045950.1	112	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain AUSMDU00010529 plasmid pAUSMDU00010529_02	CP045951.1	99	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 10ST07093	CP029839.1	5002	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 10ST07093 plasmid p10ST07093A	CP029837.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 10ST07093 plasmid p10ST07093B	CP029838.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN001921	CP006048.1	4967	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN001921 plasmid unnamed	CP006050.1	233	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN001921 plasmid unnamed2	CP006051.1	5	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN001921 plasmid unnamed3	CP006052.1	2	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7299	CP040566.1	4908	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP17-7299 plasmid pCFSAN059543	CP040567.1	133	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CFSAN018746	CP028199.1	4807	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CFSAN018746 plasmid pGMI14-001	CP028200.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM10961	CP013702.1	4772	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1896	CP014977.1	4897	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1896 plasmid pSTY1-1896	CP014978.1	176	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1899	CP007235.2	4698	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC098	CP030029.1	4747	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014851	CP038847.1	4799	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain PNCS014851 plasmid p09-0499.1	CP038848.1	119	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R	CP016385.1	5184	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R plasmid p931-3904	CP016386.1	3	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R plasmid p931IncI1	CP016387.1	102	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R plasmid p931IncI2	CP016388.1	69	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R plasmid pESBL931	CP016389.1	78	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain ST931R plasmid pSLT931	CP016390.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain U288	CP003836.1	4893	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain U288 plasmid pSTU288-1	CP004058.1	200	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain U288 plasmid pSTU288-2	CP004059.1	12	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain U288 plasmid pSTU288-3	CP004060.1	5	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain NC983	CP015157.1	4869	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain NC983 plasmid unnamed	CP015158.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL26	CP032490.1	5005	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL26 plasmid pSL26_91	CP032492.1	116	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL26 plasmid pSL26_ColRNAI	CP032493.1	6	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SL26 plasmid pSL26_IncA/C2	CP032491.1	203	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10	CP024619.1	5051	5.2	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10 plasmid p10k	CP025337.1	12	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10 plasmid p113k	CP025339.1	143	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10 plasmid p11k	CP025338.2	17	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10 plasmid p220k	CP025340.1	236	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain BL10 plasmid p3.8k	CP025336.1	6	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2009K-2059	CP014983.1	4719	4.8	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_320	CP027414.1	4982	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_320 plasmid unnamed1	CP027415.1	144	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_320 plasmid unnamed2	CP027413.1	120	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FDAARGOS_320 plasmid unnamed3	CP027416.1	10	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SARA13	CP017728.1	5019	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SARA13 plasmid pSARA13	CP017729.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UK-1	CP002614.1	4751	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain UK-1 plasmid pSTUK-100	CP002615.1	100	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 22495	CP017617.1	4949	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 22495 plasmid unnamed1	CP017618.1	112	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 22495 plasmid unnamed2	CP017619.1	1	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1880	CP014981.1	4643	4.8	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain DT2	HG326213.1	4824	4.9	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain DT2 plasmid pSLT	LN999012.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP18-6199	CP040900.1	4848	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SAP18-6199 plasmid pCFSAN074387	CP040901.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain STMU2UK	LT855376.1	4753	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain STMU2UK plasmid 2	LT855377.1	100	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 33676	CP012681.1	5063	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 33676 plasmid p33673_IncF	CP012683.1	115	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 33676 plasmid p33676_4.5	CP012684.1	5	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 33676 plasmid p33676_IncA/C	CP012682.1	198	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067217	CP028314.1	5043	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067217 plasmid pSC-31-1	CP028315.1	129	0.1	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067217 plasmid pSC-31-2	CP028316.1	217	0.2	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* var. 5- strain CFSAN067217 plasmid pSC-31-3	CP028317.1	1	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 22792	CP017621.1	4817	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain 22792 plasmid unnamed1	CP017620.1	111	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC_020	CP012144.1	4705	4.8	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2010K-1587	CP014965.1	4988	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2010K-1587 plasmid pSTY1-2010K-1587	CP016864.1	133	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2010K-1587 plasmid pSTY2-2010K-1587	CP016865.1	116	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2010K-1587 plasmid pSTY3-2010K-1587	CP016866.1	6	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2010K-1587 plasmid pSTY4-2010K-1587	CP016867.1	4	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1898	CP014971.2	4859	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1898 plasmid pSTY1-1898	CP014972.2	122	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1898 plasmid pSTY2-1898	CP014973.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain USDA-ARS-USMARC-1898 plasmid pSTY3-1898	CP014974.1	47	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain CDC 2011K-0870	CP007523.1	4634	4.8	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RSE04	CP034719.1	4875	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RSE04 plasmid pRSE04	CP034720.1	119	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21	CP032494.1	4937	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21 plasmid pSO21_118	CP032495.1	149	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21 plasmid pSO21_75	CP032497.1	101	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21 plasmid pSO21_ColRNAI_5.8	CP032499.1	12	0	plasmide

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21 plasmid pSO21_ColRNAI_6.8	CP032498.1	3	0	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain SO21 plasmid pSO21_IncA/C2	CP032496.1	126	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain VNP20009	CP007804.2	4791	4.9	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain VNP20009 plasmid pSLT_VNP20009	CP008745.1	142	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC_015	CP011365.1	4667	4.8	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC58	CP020565.1	4525	4.7	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC50	CP019383.1	4519	4.7	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain FORC88	CP029029.1	4517	4.6	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM9437	CP012985.1	4792	5.1	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM9437 plasmid pRM9437	CP014577.1	95	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain E40	CP038432.1	5027	5.0	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain E40 plasmid unnamed	CP038433.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain E40V	CP038434.1	5027	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain E40V plasmid unnamed	CP038435.1	121	0.1	plasmide
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM10607	CP013720.1	5117	5.0	complet
Salmonella enterica subsp. *enterica* serovar *Typhimurium* strain RM10607 plasmid pRM10607	CP013721.1	98	0.1	plasmide

Tableau 23 – Liste des 226 génomes et plasmides utilisés pour construire le pangénome de *S. typhimurium*.

B.8 Pangénome de *Staphylococcus aureus*

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Staphylococcus aureus subsp. *aureus* Mu3 DNA	AP009324.1	2699	2.7	complet
Staphylococcus aureus strain IT4-R	CP028470.1	2895	2.9	complet
Staphylococcus aureus subsp. *aureus* N315	NC_002745.2	2812	2.8	complet
Staphylococcus aureus subsp. *aureus* Mu50	NC_002758.2	2907	2.9	complet
Staphylococcus aureus subsp. *aureus* COL	NC_002951.2	2862	2.9	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Staphylococcus aureus subsp. *aureus* MRSA252	NC_002952.2	2873	2.9	complet
Staphylococcus aureus subsp. *aureus* MSSA476	NC_002953.3	2785	2.8	complet
Staphylococcus aureus subsp. *aureus* MW2	NC_003923.1	2783	2.8	complet
Staphylococcus aureus RF122	NC_007622.1	2730	2.7	complet
Staphylococcus aureus subsp. *aureus* USA300_FPR3757	NC_007793.1	2990	3.0	complet
Staphylococcus aureus subsp. *aureus* NCTC 8325	NC_007795.1	3148	3.1	complet
Staphylococcus aureus subsp. *aureus* JH9	NC_009487.1	3022	3.0	complet
Staphylococcus aureus subsp. *aureus* JH1	NC_009632.1	3023	3.0	complet
Staphylococcus aureus subsp. *aureus* str. Newman	NC_009641.1	2933	3.0	complet
Staphylococcus aureus subsp. *aureus* Mu3	NC_009782.1	2880	3.0	complet
Staphylococcus aureus subsp. *aureus* USA300_TCH1516	NC_010079.1	2997	3.0	complet
Staphylococcus aureus subsp. *aureus* ED98	NC_013450.1	2840	2.8	complet
Staphylococcus aureus subsp. *aureus* VC40	NC_016912.1	2658	2.6	complet
Staphylococcus aureus subsp. *aureus* M013	NC_016928.2	2766	2.7	complet
Staphylococcus aureus subsp. *aureus* TW20	NC_017331.1	3112	3.2	complet
Staphylococcus aureus subsp. *aureus* ST398	NC_017333.1	2858	2.8	complet
Staphylococcus aureus subsp. *aureus* ED133	NC_017337.1	2886	2.9	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Staphylococcus aureus subsp. *aureus* JKD6159	NC_017338.1	2798	2.8	complet
Staphylococcus aureus 04-02981	NC_017340.1	2862	2.9	complet
Staphylococcus aureus subsp. *aureus* str. JKD6008	NC_017341.1	3038	3.0	complet
Staphylococcus aureus subsp. *aureus* TCH60	NC_017342.1	2778	2.8	complet
Staphylococcus aureus subsp. *aureus* ECT-R 2	NC_017343.1	2722	2.7	complet
Staphylococcus aureus subsp. *aureus* T0131	NC_017347.1	2911	2.9	complet
Staphylococcus aureus subsp. *aureus* LGA251	NC_017349.1	2684	2.7	complet
Staphylococcus aureus subsp. *aureus* 11819-97	NC_017351.1	2864	2.8	complet
Staphylococcus aureus subsp. *aureus* 71193	NC_017673.1	2646	2.6	complet
Staphylococcus aureus subsp. *aureus* HO 5096 0412	NC_017763.1	2809	2.8	complet
Staphylococcus aureus 08BA02176	NC_018608.1	2782	2.8	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020529.1	2759	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020532.1	2762	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020533.1	2761	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020536.1	2762	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020537.1	2753	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020564.1	2759	2.7	complet

Organisme/nom	**Numéro d'accession NCBI**	**Nb de gènes**	**Taille (Mb)**	**Niveau**
Staphylococcus aureus subsp. *aureus* ST228	NC_020566.1	2765	2.7	complet
Staphylococcus aureus subsp. *aureus* ST228	NC_020568.1	2755	2.2	complet
Staphylococcus aureus M1	NC_021059.1	2907	2.9	complet
Staphylococcus aureus CA-347	NC_021554.1	2930	2.9	complet
Staphylococcus aureus	NC_021670.1	3026	3.0	complet
Staphylococcus aureus subsp. *aureus* 55/2053	NC_022113.1	2807	2.8	complet
Staphylococcus aureus subsp. *aureus* 6850	NC_022222.1	2739	2.7	complet
Staphylococcus aureus subsp. *aureus* CN1	NC_022226.1	2716	2.7	complet
Staphylococcus aureus subsp. *aureus* SA957	NC_022442.1	2794	2.8	complet
Staphylococcus aureus subsp. *aureus* SA40	NC_022443.1	2710	2.7	complet

Tableau 24 – Liste des 49 génomes utilisés pour construire le pangénome de *S. aureus* (partiellement adapté de [97]).

Annexe C

Stratégie de détection des OGM mise au point sur le génome de maïs OGM

Le génome du maïs OGM fourni par le LSV est utilisé pour développer/mettre au point la méthode de détection des OGM inconnu. Cet OGM contient l'insert T25 qui est composé de trois CDS (Chapitre 2 partie 2.3.1.1).

C.1 Nettoyage des données de séquençage

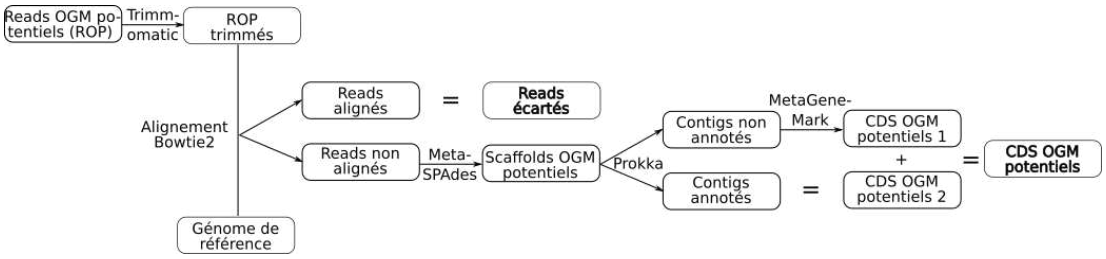


FIGURE 50 – Pipeline de nettoyage des données de séquençage du maïs OGM.

Le but du pipeline de nettoyage illustré en Figure 50 est d'éliminer les CDS appartenant au génome de référence du maïs qui sont présents dans le génome du maïs OGM et conserver uniquement les CDS potentiellement inserts. Les données de séquençage de l'OGM de maïs sont nettoyées à l'aide du logiciel Trimmomatic. Ensuite, un alignement avec l'aligneur Bowtie2 est réalisé entre le génome de référence et les données de séquençage trimmées. A la fin de cet alignement, les reads qui se sont alignés sur le génome de référence sont écartés, ils correspondent au génome de référence du maïs. Les reads non alignés (ne correspondant pas au génome de référence donc potentiellement OGM) sont ensuite assemblés à l'aide de l'assembleur MetaSPAdes et annotés par le logiciel Prokka. MetaSPAdes est un logiciel d'assemblage pour les données métagénomiques. Il est un dérivé du logiciel d'assemblage SPAdes qui a initialement été conçu et testé sur des petits génomes bactériens et fongiques. L'avantage de SPAdes est son efficacité, il est rapide et précis pour assembler les reads. Prokka est un annotateur dédié prin-

cipalement aux procaryotes, les contigs non annotés par ce dernier sont ensuite soumis au logiciel MetaGeneMark. MetaGeneMark permet de prédire des gènes provenant de données métagénomiques. La fusion des deux résultats de prédiction des logiciels Prokka et MetaGeneMark est alors conservée, les CDS nettoyés du génome contenant les séquences d'inserts OGM sont alors obtenus.

C.2 Filtrage de la banque de données d'OGM connus (noté FD) et calcul de distance

La banque de données d'OGM connus FD est filtrée à l'aide d'un alignement BLASTN sur les CDS du génome de référence. En effet, un OGM est défini par l'insertion d'une séquence exogène, provenant d'une espèce extérieure au génome sauvage. Cette étape permet donc d'éliminer tous les CDS appartenant à l'OGM présent dans la banque de données. Trois jeux de données distincts sont nécessaires pour le calcul de distance, les CDS potentiellement inserts obtenus à la fin du pipeline de nettoyage, les CDS de la banque de données d'OGM filtrés et les CDS du génome de référence. Une distance est alors calculée entre l'ensemble des CDS du génome de référence et chacun des CDS de ces trois jeux de données. Différents filtres vont être appliqué sur les calculs de distances obtenus des CDS OGM potentiels comme le filtre sur la profondeur de couverture (Chapitre 5 partie 5.3.2.10).

Annexe D

Gènes utilisés pour créer les données synthétiques OGM

Liste des numéros d'accession NCBI des huit gènes utilisés pour créer 42 jeux de données synthétiques OGM à partir de six bactéries différentes :

- CAL34210.1 : gène *ansA* de *Campylobacter jejuni*
- NP_001356419.1 : gène *RHEX* de *Homo sapiens*
- AAK04455.1 : gène *ydgD* de *Lactococcus lactis*
- NP_463608.1 : gène produisant la protéine carboxyphosphonoenolpyruvate phosphonomutase de *Listeria monocytogenes*
- NP_215994.1 : gène *ripB* de *Mycobacterium tuberculosis*
- AM411441.1 : gène *eIF4E* de *Oryza sativa*
- AAL21510.1 : gène produisant la protéine Gifsy-1 prophage de *Salmonella typhimurium*
- YP_501325.1 : gène produisant la protéine 1-pyrroline-5-carboxylate dehydrogenase de *Staphylococcus aureus*

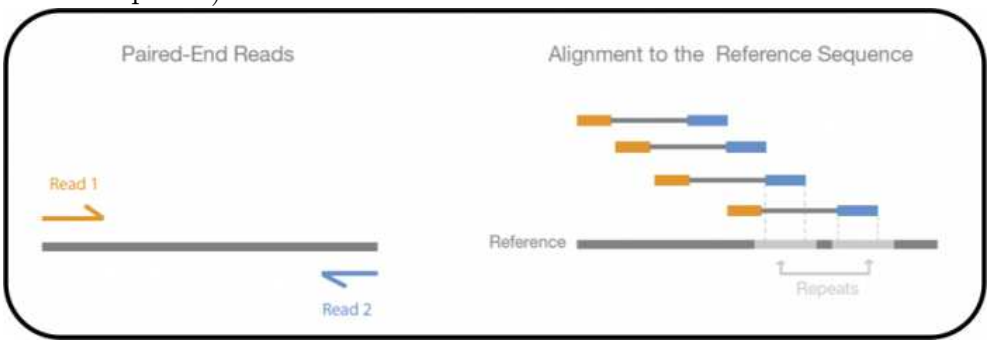
Glossaire

Chimère Une chimère est un organisme, comme une plante, formé de deux (ou plus) populations de cellules génétiquement distinctes. Une plante est appelée chimère lorsque des cellules de plus d'un génotype (code génétique) sont trouvées croissant de façon juxtaposée dans les tissus de cette plante. Par extension, des reads dits chimériques sont des reads formés de séquences non contiguës sur un génome de référence (s'alignant sur plusieurs portions du génome référence).

Coloration de Gram La coloration de Gram permet de distinguer plusieurs grandes familles de bactéries par un test simple de coloration différenciée en fonction de la composition de la paroi cellulaire.

Consensus d'alignement Un consensus d'alignement est la séquence nucléotidique la plus fréquente à chaque position d'un alignement de séquences.

Données pairées Les données pairées proviennent de séquençage de deuxième génération qui permettent de séquencer des reads pairés, ou « paired-end ». Ces reads proviennent de chacune des deux extrémités 5' et 3' du fragment d'ADN qui a été séquencé (Figure suivante [150]) et fournissent donc une indication de distance entre ces deux extrémités (pour les assembleurs qui en tirent partie).



Le séquençage de données pairées permet de séquencer les deux extrémités du fragment. Pour chaque fragment, la distance séparant les reads d'une même paire peut être estimée ainsi que les informations concernant le read apparié à celui considéré. Les séquenceurs de type Illumina fournissent ce type de données, mais ce n'est pas obligatoire.

Framework Un framework, aussi appelé infrastructure logicielle, désigne un ensemble de composants logiciels et d'outils représentant la base d'un lo-

giciel ou d'une application. Il fonctionne comme un patron et fournit un cadre pour établir les fondations d'un logiciel, facilitant ainsi le travail des développeurs.

Gap Un gap est un espacement de quelques nucléotides pouvant être généré par les programmes d'alignement dans le but de réaliser un meilleur alignement. Un gap peut être interprété par une succession d'insertions ou de délétion.

Introns Les introns sont des séquences nucléotidiques non utilisées pour la traduction de la protéine considérée. Au cours de la maturation de l'ARNm, ces séquences sont éliminées. Elles sont encadrés par deux exons. A contrario, les exons sont les séquences nucléotidiques conservées dans l'ARN après l'épissage (processus d'excision des introns).

Non déterministe Un outil ou une méthode non déterministe est un programme qui peut produire des résultats différents lorsque les données d'entrée fournies au programme sont identiques.

Profondeur de couverture La profondeur de couverture d'un séquençage représente le nombre de fois où chaque base du génome est couverte par un read à une position donnée en moyenne.

Protéine Chaperon Une protéine chaperon est une protéine dont la fonction est d'assister d'autres protéines dans leur maturation, en évitant la formation d'agrégats via les domaines hydrophobes présents sur leur surface lors de leur repliement tridimensionnel.

Région UTR (Untranslated Transcribed Region) Une région UTR est une partie non traduite en protéine de l'ARN messager. Les régions UTR ont un rôle important dans le processus cellulaire qui conduit à l'expression de l'ARN messager. Ces régions sont localisées aux extrémités 5' (en amont du codon d'initiation) et 3' (après le codon stop).

Abbréviations

ADN Acide désoxyribonucléique.

ANSES Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

ARN Acide ribonucléique.

ARNr ARN ribosomique.

ARNt ARN de transfert.

BVL Federal Office of Consumer Protection and Food Safty.

C5.0 Algorithme de classification C5.0.

CDS Séquence codante.

CRA-W Centre wallon de Recherches agronomiques.

DUGMO Detection of Unkwown Genetically Modified Organisms.

Eugenius European GMO Initiative for a Unified Database System.

FDA Food and Drugs Administration.

GM Génétiquement Modifié.

JRC Joint Research Centre.

knn K plus proches voisins.

Logit Modèle linéaire généralisé.

LSV Laboratoire de la santé des végétaux.

NCBI National Center for Biotechnology Information.

NGS Next Generation Sequencing.

nnet Réseaux de neurones.

NPBT New Plant Breeding techniques.

OGM Organisme Génétiquement Modifié.

PLS Régression par les moindres carrées partiels.

RF Forêts aléatoire.

rpart Arbres de partitionnement récursifs.

SRA Sequence Read Archive.

stepLDA Analyse discriminante linéaire stepwise.

stepQDA Analyse discriminante quadratique stepwise.

svmRadial Machine à vecteur de support avec une fonction noyau de base radiale.

Treebag Arbres de classification avec Bagging.

WFSR Wageningen Food Safety Research.

Xgboost Extrême gradient boosting.

Bibliographie

[1] Hang Thu Nguyen and Johannes A. Jehle. Expression of Cry3Bb1 in transgenic corn MON88017. *Journal of Agricultural and Food Chemistry*, 57(21) :9990–9996, November 2009.

[2] S. Höss, M. Arndt, S. Baumgarte, C. C. Tebbe, H. T. Nguyen, and J. A. Jehle. Effects of transgenic corn and Cry1Ab protein on the nematode, Caenorhabditis elegans. *Ecotoxicology and Environmental Safety*, 70(2) :334–340, June 2008.

[3] Sebastian Höss, Ralph Menzel, Frank Gessler, Hang T. Nguyen, Johannes A. Jehle, and Walter Traunspurger. Effects of insecticidal crystal proteins (Cry proteins) produced by genetically modified maize (Bt maize) on the nematode Caenorhabditis elegans. *Environmental Pollution (Barking, Essex : 1987)*, 178 :147–151, July 2013.

[4] Wayback Machine, https :/web.archive.org/web/20151121055636/http :/-www.fda.gov/downloads/forconsumers/consumerupdates/ucm473578.pdf, November 2015.

[5] Doori Park, Su-Hyun Park, Yong Wook Ban, Youn Shic Kim, Kyoung-Cheul Park, Nam-Soo Kim, Ju-Kon Kim, and Ik-Young Choi. A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. *BMC Biotechnology*, 17, August 2017.

[6] Fabrice Touzain, Marie-Agnès Petit, Sophie Schbath, and Meriem El Karoui. DNA motifs that sculpt the bacterial chromosome. *Nature Reviews. Microbiology*, 9(1) :15–26, January 2011.

[7] Morgan G. I. Langille and Fiona S. L. Brinkman. Bioinformatic detection of horizontally transferred DNA in bacterial genomes. *F1000 Biology Reports*, 1 :25, March 2009.

[8] Vladimir Trifonov and Raul Rabadan. Frequency Analysis Techniques for Identification of Viral Genetic Data. *mBio*, 1(3) :e00156–10, August 2010.

[9] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews. Genetics*, 16(6) :321–332, June 2015.

[10] Patrick Murigu Kamau Njage, Clementine Henri, Pimlapas Leekitcharoenphon, Michel-Yves Mistou, Rene S. Hendriksen, and Tine Hald. Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. *Risk Analysis*, 39(6) :1397–1413, 2019.

[11] Aditya M. Kunjapur, Philipp Pfingstag, and Neil C. Thompson. Gene synthesis allows biologists to source genes from farther away in the tree of life. *Nature Communications*, 9(1) :1–11, October 2018.

[12] Christopher Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests : Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2) :689–702, June 2017.

[13] Avery Li-Chun Wang. An Industrial-Strength Audio Search Algorithm. page 7, 2003.

[14] Arne Holst-Jensen, Yves Bertheau, Marc de Loose, Lutz Grohmann, Sandrine Hamels, Lotte Hougs, Dany Morisset, Sven Pecoraro, Maria Pla, Marc Van den Bulcke, and Doerte Wulff. Detecting un-authorized genetically modified organisms (GMOs) and derived materials. *Biotechnology Advances*, 30(6) :1318–1335, December 2012.

[15] D. A. Jackson, R. H. Symons, and P. Berg. Biochemical method for inserting new genetic information into DNA of Simian Virus 40 : circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 69(10) :2904–2909, October 1972.

[16] D. V. Goeddel, D. G. Kleid, F. Bolivar, H. L. Heyneker, D. G. Yansura, R. Crea, T. Hirose, A. Kraszewski, K. Itakura, and A. D. Riggs. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences of the United States of America*, 76(1) :106–110, January 1979.

[17] How are organisms genetically modified ?, http :/sphweb.bumc.bu.edu/otlt-/mph-modules/ph/gmos/gmos3.html.

[18] I. Zaenen, N. Van Larebeke, M. Van Montagu, and J. Schell. Supercoiled circular DNA in crown-gall inducing Agrobacterium strains. *Journal of Molecular Biology*, 86(1) :109–127, June 1974.

[19] N. Van Larebeke, G. Engler, M. Holsters, S. Van den Elsacker, I. Zaenen, R. A. Schilperoort, and J. Schell. Large plasmid in Agrobacterium tumefaciens essential for crown gall-inducing ability. *Nature*, 252(5479) :169–170, November 1974.

[20] N. Van Larebeke, C. Genetello, J. Schell, R. A. Schilperoort, A. K. Hermans, M. Van Montagu, and J. P. Hernalsteens. Acquisition of tumour-inducing ability by non-oncogenic agrobacteria as a result of plasmid transfer. *Nature*, 255(5511) :742–743, June 1975.

[21] K. A. Barton, A. N. Binns, A. J. Matzke, and M. D. Chilton. Regeneration of intact tobacco plants containing full length copies of genetically engineered T-DNA, and transmission of T-DNA to R1 progeny. *Cell*, 32(4) :1033–1043, April 1983.

[22] Yuji Ishida, Hideaki Saito, Shozo Ohta, Yukoh Hiei, Toshihiko Komari, and Takashi Kumashiro. High efficiency transformation of maize (Zea mays L.)

mediated by Agrobacterium tumefaciens. *Nature Biotechnology*, 14(6) :745–750, June 1996.

[23] Satoko Nonaka, Tatsuhiko Someya, Yasuhiro Kadota, Kouji Nakamura, and Hiroshi Ezura. Super-Agrobacterium ver. 4 : Improving the Transformation Frequencies and Genetic Engineering Possibilities for Crop Plants. *Frontiers in Plant Science*, 10, October 2019.

[24] Julie R. Kikkert, José R. Vidal, and Bruce I. Reisch. Stable transformation of plant cells by particle bombardment/biolistics. *Methods in Molecular Biology (Clifton, N.J.)*, 286 :61–78, 2005.

[25] Qin Yao, Ling Cong, Jun Li Chang, Ke Xiu Li, Guang Xiao Yang, and Guang Yuan He. Low copy number gene transfer and stable expression in a commercial wheat cultivar via particle bombardment. *Journal of Experimental Botany*, 57(14) :3737–3746, 2006.

[26] Jennifer A. Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213) :1258096, November 2014.

[27] Édition génomique, https ://fr.wikipedia.org/w/index.php ?title=édition_-génomique, February 2020. Page Version ID : 167525454.

[28] Andrew V. Anzalone, Peyton B. Randolph, Jessie R. Davis, Alexander A. Sousa, Luke W. Koblan, Jonathan M. Levy, Peter J. Chen, Christopher Wilson, Gregory A. Newby, Aditya Raguram, and David R. Liu. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 576(7785) :149–157, December 2019.

[29] EUR-Lex - 32003R1829 - EN - EUR-Lex, https ://eur-lex.europa.eu/legal-content/en/all/ ?uri=celex :32003r1829.

[30] Dennis Eriksson, Wendy Harwood, Per Hofvander, Huw Jones, Peter Rogowsky, Eva Stöger, and Richard G. F. Visser. A Welcome Proposal to Amend the GMO Legislation of the EU. *Trends in Biotechnology*, 36(11) :1100–1103, 2018.

[31] Genius - Présentation générale, https ://www6.inra.fr/genius-project/le-projet/presentation-generale.

[32] Marco Mazzara, Claudia Paoletti, Philippe Corbisier, Emanuele Grazioli, Sara Larcher, Gilbert Berben, Marc De Loose, Imma Folch, Christine Henry, Norbert Hess, Lotte Hougs, Eric Janssen, Gillian Moran, Roberta Onori, and Guy Van den Eede. Kernel Lot Distribution Assessment (KeLDA) : a Comparative Study of Protein and DNA-Based Detection Methods for GMO Testing. *Food Analytical Methods*, 6(1) :210–220, February 2013.

[33] Sylvia Broeders, Nina Papazova, Marc Van den Bulcke, and Nancy H. C. Roosens. Development of a Molecular Platform for GMO Detection in Food and Feed on the Basis of "Combinatory qPCR" Technology. 2012.

[34] Marie-Alice Fraiture, Philippe Herman, Isabel Taverniers, Marc De Loose, Dieter Deforce, and Nancy H. Roosens. Current and new approaches in GMO detection : challenges and solutions. *BioMed Research International*, 2015 :392872, 2015.

[35] Luminex Corporation | Complexity Simplified, https ://www.luminexcorp.com/.

[36] Anna Fantozzi, Monica Ermolli, Massimiliano Marini, Branko Balla, Maddalena Querci, and Guy Van den Eede. Innovative Application of Fluorescent Microsphere Based Assay for Multiple GMO Detection. *Food Analytical Methods*, 1(1) :10–17, March 2008.

[37] David Dobnik, Dany Morisset, and Kristina Gruden. NAIMA as a solution for future GMO diagnostics challenges. *Analytical and Bioanalytical Chemistry*, 396(6) :2229–2233, March 2010.

[38] M. Pla, A. Nadal, V. Baeten, C. Bahrdt, G. Berben, Y. Bertheau, A. Coll, J. P. van Dijk, D. Dobnik, J. A. Fernandez Pierna, K. Gruden, S. Hamels, A. Holck, A. Holst[U+2010]Jensen, E. Janssen, E. J. Kok, J. L. La Paz, V. Laval, S. Leimanis, A. Malcevschi, N. Marmiroli, D. Morisset, T. W. Prins, J. Remacle, G. Ujhelyi, and D. Wulff. New Multiplexing Tools for Reliable GMO Detection. In *Genetically Modified and Non-Genetically Modified Food Supply Chains : Co-Existence and Traceability*, pages 333–366. John Wiley & Sons, Ltd, 2012.

[39] Mojca Milavec, David Dobnik, Litao Yang, Dabing Zhang, Kristina Gruden, and Jana Zel. GMO quantification : valuable experience and insights for the future. *Analytical and Bioanalytical Chemistry*, 406(26) :6485–6497, October 2014.

[40] T. Notomi, H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino, and T. Hase. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*, 28(12) :E63, June 2000.

[41] Junyi Xu, Qiuyue Zheng, Ling Yu, Ran Liu, Xin Zhao, Gang Wang, Qinghua Wang, and Jijuan Cao. Loop-mediated isothermal amplification (LAMP) method for detection of genetically modified maize T25. *Food Science & Nutrition*, 1(6) :432, November 2013.

[42] Marie-Alice Fraiture, Philippe Herman, Marc De Loose, Frédéric Debode, and Nancy H. Roosens. How Can We Better Detect Unauthorized GMOs in Food and Feed Chains ? *Trends in Biotechnology*, 35(6) :508–517, 2017.

[43] Marie-Alice Fraiture, Assia Saltykova, Stefan Hoffman, Raf Winand, Dieter Deforce, Kevin Vanneste, Sigrid C. J. De Keersmaecker, and Nancy H. C. Roosens. Nanopore sequencing technology : a new route for the fast detection of unauthorized GMO. *Scientific Reports*, 8(1) :7903, May 2018.

[44] Sylvia R. M. Broeders, Sigrid C. J. De Keersmaecker, and Nancy H. C. Roosens. How to deal with the upcoming challenges in GMO detection in food and feed. *Journal of Biomedicine & Biotechnology*, 2012 :402418, 2012.

[45] Sander Willems, Marie-Alice Fraiture, Dieter Deforce, Sigrid C. J. De Keersmaecker, Marc De Loose, Tom Ruttink, Philippe Herman, Filip Van Nieuwerburgh, and Nancy Roosens. Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing. *Food Chemistry*, 192 :788–798, February 2016.

[46] Evrogen Technologies : Genome walking, http ://evrogen.com/technologies/genome-walking.shtml.

[47] Marie-Alice Fraiture, Philippe Herman, Isabel Taverniers, Marc De Loose, Dieter Deforce, and Nancy H. Roosens. An innovative and integrated approach based on DNA walking to identify unauthorised GMOs. *Food Chemistry*, 147 :60–69, March 2014.

[48] Marie-Alice Fraiture, Philippe Herman, Loic Lefèvre, Isabel Taverniers, Marc De Loose, Dieter Deforce, and Nancy H. Roosens. Integrated DNA walking system to characterize a broad spectrum of GMOs in food/feed matrices. *BMC Biotechnology*, 15(1) :76, August 2015.

[49] Marie-Alice Fraiture, Julie Vandamme, Philippe Herman, and Nancy H. C. Roosens. Development and validation of an integrated DNA walking strategy to detect GMO expressing cry genes. *BMC Biotechnology*, 18(1) :40, June 2018.

[50] Marie-Alice Fraiture, Philippe Herman, Nina Papazova, Marc De Loose, Dieter Deforce, Tom Ruttink, and Nancy H. Roosens. An integrated strategy combining DNA walking and NGS to detect GMOs. *Food Chemistry*, 232 :351–358, October 2017.

[51] Dany Morisset, Petra Kralj Novak, Darko Zupanič, Kristina Gruden, Nada Lavrač, and Jana Žel. GMOseek : a user friendly tool for optimized GMO testing. *BMC bioinformatics*, 15 :258, August 2014.

[52] Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, and Claire Lemaitre. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2 :e94, November 2016.

[53] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1) :257, November 2019.

[54] Stefan Kurtz, Apurva Narechania, Joshua C. Stein, and Doreen Ware. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, 9(1) :517, October 2008.

[55] The European GMO database, http ://www.euginius.eu/euginius/pages/-home.jsf.

[56] Mauro Petrillo, Alexandre Angers-Loustau, Peter Henriksson, Laura Bonfini, Alex Patak, and Joachim Kreysa. JRC GMO-Amplicons : a collection of nucleic acid sequences related to genetically modified organisms. *Database : The Journal of Biological Databases and Curation*, 2015, 2015.

[57] GMO-Amplicon sources, http ://data.europa.eu/89h/f7e6917f-ccc4-4c88-a622-07c8f961083e. June 2019.

[58] Laura Bonfini, Marc H. Van den Bulcke, Marco Mazzara, Enrico Ben, and Alexandre Patak. GMOMETHODS : the European Union database of reference methods for GMO analysis. *Journal of AOAC International*, 95(6) :1713–1719, December 2012.

[59] Edward J. Fox, Kate S. Reid-Bayliss, Mary J. Emond, and Lawrence A. Loeb. Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, 1, 2014.

[60] Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. Deboy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39) :13950–13955, September 2005.

[61] Computational pan-genomics : status, promises and challenges. *Briefings in Bioinformatics*, 19(1) :118–135, October 2016.

[62] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics : the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5) :472–477, October 2008.

[63] James O. McInerney, Alan McNally, and Mary J. O'Connell. Why prokaryotes have pangenomes. *Nature Microbiology*, 2(4) :1–5, March 2017.

[64] Leo Egghe. Untangling Herdan's law and Heaps' law : Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58(5) :702–709, 2007.

[65] Ye Peng, Shanmei Tang, Dan Wang, Huanzi Zhong, Huijue Jia, Xianghang Cai, Zhaoxi Zhang, Minfeng Xiao, Huanming Yang, Jian Wang, Karsten Kristiansen, Xun Xu, and Junhua Li. MetaPGN : a pipeline for construction and graphical visualization of annotated pangenome networks. *GigaScience*, 7(11), 2018.

[66] Xinyu Chen, Yadong Zhang, Zhewen Zhang, Yongbing Zhao, Chen Sun, Ming Yang, Jinyue Wang, Qian Liu, Baohua Zhang, Meili Chen, Jun Yu, Jiayan Wu, Zhong Jin, and Jingfa Xiao. PGAweb : A Web Server for Bacterial Pan-Genome Analysis. *Frontiers in Microbiology*, 9, August 2018.

[67] Jingfa Xiao, Zhewen Zhang, Jiayan Wu, and Jun Yu. A Brief Review of Software Tools for Pangenomics. *Genomics, Proteomics & Bioinformatics*, 13(1) :73–76, February 2015.

[68] George Vernikos, Duccio Medini, David R. Riley, and Hervé Tettelin. Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23 :148–154, February 2015.

[69] L. Rouli, V. Merhej, P.-E. Fournier, and D. Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7 :72–85, June 2015.

[70] Paul G. Livingstone, Russell M. Morphew, and David E. Whitworth. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Frontiers in Microbiology*, 9, December 2018.

[71] Matthias Scholz, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, 13(5) :435–438, 2016.

[72] Torsten Seemann. Prokka : rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14) :2068–2069, July 2014.

[73] P. Rice, I. Longden, and A. Bleasby. EMBOSS : the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6) :276–277, June 2000.

[74] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART : a next-generation sequencing read simulator. *Bioinformatics*, 28(4) :593–594, February 2012.

[75] T25 | GM Approval Database, http ://www.isaaa.org/gmapprovaldatabase-/event/default.asp ?eventid=102.

[76] Biosafety Unit. Record details, http ://bch.cbd.int/database/record.shtml-?documentid=14767.

[77] Maher Chaouachi, Mohamed Salem Zellama, Nesrine Nabi, Ahmed Ben Hafsa, and Khaled Saïd. Molecular Identification of Four Genetically Modified Maize (Bt11, Bt176, Mon810 and T25) by Duplex Quantitative Real-Time PCR. *Food Analytical Methods*, 7(1) :224–233, January 2014.

[78] Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth

Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The B73 maize genome : complexity, diversity, and dynamics. *Science (New York, N.Y.)*, 326(5956) :1112–1115, November 2009.

[79] F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Cordani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denziot, K. M. Devine, A. Düsterhöft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S. Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Hènaut, H. Hilbert, S. Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauël, C. Médigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Pohl, D. Portelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Roche, M. Rose, Y. Sadaie, T. Sato, E. Scanlan, S. Schleich, R. Schroeter, F. Scoffone, J. Sekiguchi, A. Sekowska, S. J. Seror, P. Serror, B. S. Shin, B. Soldo, A. Sorokin, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Togoni, V. Tosato, S. Uchiyama, M. Vandebol, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, H. Wedler, T. Weitzenegger, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H. F.

Yoshikawa, E. Zumstein, H. Yoshikawa, and A. Danchin. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature*, 390(6657) :249–256, November 1997.

[80] Valérie Barbe, Stéphane Cruveiller, Frank Kunst, Patricia Lenoble, Guillaume Meurice, Agnieszka Sekowska, David Vallenet, Tingzhang Wang, Ivan Moszer, Claudine Médigue, and Antoine Danchin. From a consortium sequence to a unified sequence : the Bacillus subtilis 168 reference genome a decade later. *Microbiology (Reading, England)*, 155(Pt 6) :1758–1775, June 2009.

[81] Rainer Borriss, Antoine Danchin, Colin R. Harwood, Claudine Médigue, Eduardo P. C. Rocha, Agnieszka Sekowska, and David Vallenet. Bacillus subtilis, the model Gram-positive bacterium : 20 years of annotation refinement. *Microbial Biotechnology*, 11(1) :3–17, 2018.

[82] Valentina Paracchini, Mauro Petrillo, Ralf Reiting, Alexandre Angers-Loustau, Daniela Wahler, Andrea Stolz, Birgit Schönig, Anastasia Matthies, Joachim Bendiek, Dominik M. Meinel, Sven Pecoraro, Ulrich Busch, Alex Patak, Joachim Kreysa, and Lutz Grohmann. Molecular characterization of an unauthorized genetically modified Bacillus subtilis production strain identified in a vitamin B2 feed additive. *Food Chemistry*, 230 :681–689, September 2017.

[83] GM Bacillus subtilis, http ://data.europa.eu/89h/2abb5c2b-3ab6-4ce4-b103-cb1c5fc7349e. 2014.

[84] Bas Berbers, Assia Saltykova, Cristina Garcia-Graells, Patrick Philipp, Fabrice Arella, Kathleen Marchal, Raf Winand, Kevin Vanneste, Nancy H. C. Roosens, and Sigrid C. J. De Keersmaecker. Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Scientific Reports*, 10(1) :4310, March 2020. Number : 1 Publisher : Nature Publishing Group.

[85] Torsten Stein. Bacillus subtilis antibiotics : structures, syntheses and specific functions. *Molecular Microbiology*, 56(4) :845–857, May 2005.

[86] Patrícia H Brito, Bastien Chevreux, Cláudia R Serra, Ghislain Schyns, Adriano O Henriques, and José B Pereira-Leal. Genetic Competence Drives Genome Diversity in Bacillus subtilis. *Genome Biology and Evolution*, 10(1) :108–124, December 2017.

[87] Abigail Clements, Joanna C. Young, Nicholas Constantinou, and Gad Frankel. Infection strategies of enteric pathogenic Escherichia coli. *Gut Microbes*, 3(2) :71–87, April 2012.

[88] Haeyoung Jeong, Valérie Barbe, Choong Hoon Lee, David Vallenet, Dong Su Yu, Sang-Haeng Choi, Arnaud Couloux, Seung-Won Lee, Sung Ho Yoon, Laurence Cattolico, Cheol-Goo Hur, Hong-Seog Park, Béatrice Ségurens, Sun Chang Kim, Tae Kwang Oh, Richard E. Lenski, F. William Studier, Patrick Daegelen, and Jihyun F. Kim. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *Journal of Molecular Biology*, 394(4) :644–652, December 2009.

[89] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)*, 277(5331) :1453–1462, September 1997.

[90] Jessica R. Ames, Meenakumari Muthuramalingam, Tamiko Murphy, Fares Z. Najar, and Christina Bourne. Expression of different ParE toxins results in conserved phenotypes with distinguishable classes of toxicity. *MicrobiologyOpen*, July 2019.

[91] Fabrice Touzain, Erick Denamur, Claudine Médigue, Valérie Barbe, Meriem El Karoui, and Marie-Agnès Petit. Small variable segments constitute a major type of diversity of bacterial genomes at the species level. *Genome Biology*, 11(4) :R45, 2010.

[92] Tohru Miyoshi-Akiyama, Kazunori Matsumura, Hiroki Iwai, Keiji Funatogawa, and Teruo Kirikae. Complete Annotated Genome Sequence of Mycobacterium tuberculosis Erdman. *Journal of Bacteriology*, 194(10) :2770–2770, May 2012. Publisher : American Society for Microbiology Journals Section : Genome Announcements.

[93] J. Parkhill, B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. M. Davies, T. Feltwell, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Moule, M. J. Pallen, C. W. Penn, M. A. Quail, M. A. Rajandream, K. M. Rutherford, A. H. van Vliet, S. Whitehead, and B. G. Barrell. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature*, 403(6770) :665–668, February 2000.

[94] Huiling Di, Lei Ye, He Yan, Hecheng Meng, Shinji Yamasak, and Lei Shi. Comparative analysis of CRISPR loci in different Listeria monocytogenes lineages. *Biochemical and Biophysical Research Communications*, 454(3) :399–403, 2014.

[95] Tingting Yang, Jun Zhong, Ju Zhang, Cuidan Li, Xia Yu, Jingfa Xiao, Xinmiao Jia, Nan Ding, Guannan Ma, Guirong Wang, Liya Yue, Qian Liang, Yongjie Sheng, Yanhong Sun, Hairong Huang, and Fei Chen. Pan-Genomic Study of Mycobacterium tuberculosis Reflecting the Primary/Secondary Genes, Generality/Individuality, and the Interconversion Through Copy Number Variations. *Frontiers in Microbiology*, 9 :1886, 2018.

[96] Erol S. Kavvas, Edward Catoiu, Nathan Mih, James T. Yurkovich, Yara Seif, Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet, Jonathan M. Monk, and Bernhard O. Palsson. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nature Communications*, 9(1) :1–9, October 2018.

[97] Emanuele Bosi, Jonathan M. Monk, Ramy K. Aziz, Marco Fondi, Victor Nizet, and Bernhard Ø Palsson. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proceedings of the National Academy of Sciences of the United States of America*, 113(26) :E3801–3809, 2016.

[98] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1) :r49, January 1980.

[99] Pamela Mukhopadhyay, Surajit Basak, and Tapash Chandra Ghosh. Synonymous codon usage in different protein secondary structural classes of human genes : implication for increased non-randomness of GC3 rich genes towards protein stability. *Journal of Biosciences*, 32(5) :947–963, August 2007.

[100] S. Majumdar, S. K. Gupta, V. S. Sundararajan, and T. C. Ghosh. Compositional correlation studies among the three different codon positions in 12 bacterial genomes. *Biochemical and Biophysical Research Communications*, 266(1) :66–71, December 1999.

[101] Anton A. Komar. The Yin and Yang of codon usage. *Human Molecular Genetics*, 25(R2) :R77–R85, October 2016.

[102] Sophie Schbath and Mark Hoebeke. R'MES : A Tool to Find Motifs with a Significantly Unexpected Frequency in Biological Sequences. In *Advances in Genomic Sequence Analysis and Pattern Discovery*, volume Volume 7 of *Science, Engineering, and Biology Informatics*, pages 25–64. WORLD SCIENTIFIC, January 2011.

[103] doc/rmes3.1.0cmake.userguide.pdf · master · Sophie Schbath / RMES.

[104] Charalampos S. Kouzinopoulos, John-Alexander M. Assael, Themistoklis K. Pyrgiotis, and Konstantinos G. Margaritis. A Hybrid Parallel Implementation of the Aho–Corasick and Wu–Manber Algorithms Using NVIDIA CUDA and MPI Evaluated on a Biological Sequence Database. *International Journal on Artificial Intelligence Tools*, 24(01) :1540001, February 2015.

[105] Alfred V. Aho and Margaret J. Corasick. Efficient string matching : an aid to bibliographic search. *Communications of the ACM*, 18(6) :333–340, June 1975.

[106] W. H. Campbell and G. Gowri. Codon usage in higher plants, green algae, and cyanobacteria. *Plant Physiology*, 92(1) :1–11, January 1990.

[107] Hanmei Liu, Rui He, Huaiyu Zhang, Yubi Huang, Mengliang Tian, and Junjie Zhang. Analysis of synonymous codon usage in Zea mays. *Molecular Biology Reports*, 37(2) :677–684, February 2010.

[108] Codon Usage Database, https ://www.kazusa.or.jp/codon/.

[109] P. M. Sharp, M. Stenico, J. F. Peden, and A. T. Lloyd. Codon usage : mutational bias, translational selection, or both ? *Biochemical Society Transactions*, 21(4) :835–841, November 1993.

[110] D. C. Shields and P. M. Sharp. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Research*, 15(19) :8023, October 1987.

[111] Marc Bailly-Bechet. Biais de codons et régulation de la traduction chez les bactéries et leurs phages. page 263.

[112] S. Karlin, J. Mrázek, and A. M. Campbell. Codon usages in different gene classes of the Escherichia coli genome. *Molecular Microbiology*, 29(6) :1341–1355, September 1998.

[113] Sandrine Baron, Laetitia Le Devendec, Fabrice Touzain, Eric Jouy, Pierrick Lucas, Claire de Boisséson, Emeline Larvor, and Isabelle Kempf. Longitudinal study of Escherichia coli plasmid resistance to extended-spectrum cephalosporins in free-range broilers. *Veterinary Microbiology*, 216 :20–24, March 2018.

[114] Torsten Seemann. Faster SPAdes assembly of Illumina reads. Contribute to tseemann/shovill development by creating an account on GitHub, July 2019. original-date : 2016-09-05T23 :50 :20Z.

[115] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes : a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 19(5) :455–477, May 2012.

[116] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15) :2114–2120, August 2014.

[117] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2) :83–90, March 2017.

[118] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4) :357–359, March 2012.

[119] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14) :1754–1760, July 2009.

[120] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16) :2078–2079, August 2009.

[121] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410, October 1990.

[122] Bastien Chevreux, Thomas Wetter, and Sándor Suhai. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. In *German Conference on Bioinformatics*, 1999.

[123] T. J. Wu, Y. C. Hsieh, and L. A. Li. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, 57(2) :441–448, June 2001.

[124] Derrick E. Wood and Steven L. Salzberg. Kraken : ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3) :R46, March 2014.

[125] Elizabeth S. Allman, John A. Rhodes, and Seth Sullivant. Statistically Consistent k-mer Methods for Phylogenetic Tree Reconstruction. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 24(2) :153–171, February 2017.

[126] J. Roger Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, February 1957.

[127] C. Ricotta and J. Podani. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 31 :201–205, September 2017.

[128] K. Robert Clarke, Paul J. Somerfield, and M. Gee Chapman. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, 330(1) :55–80, March 2006.

[129] Pierre Legendre and Miquel De Cáceres. Beta diversity as the variance of community data : dissimilarity coefficients and partitioning. *Ecology Letters*, 16(8) :951–963, 2013.

[130] Pierre Legendre. Interpreting the replacement and richness difference components of beta diversity. *Global Ecology and Biogeography*, 23(11) :1324–1334, 2014.

[131] Brian D. Ondov, Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, 12 :385, September 2011.

[132] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2) :325–340, 2018.

[133] Max Kuhn. *The caret Package*.

[134] Vishal Morde. XGBoost Algorithm : Long May She Reign !, https ://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d, April 2019.

[135] Steven L. Salzberg. C4.5 : Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3) :235–240, September 1994.

[136] L. Breiman, J. Friedman, and C. Olshen, R. and Stone. *Classification and Regression Tree*. New York, chapman & hall edition, 1984.

[137] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2) :123–140, August 1996.

[138] Jerome H. Friedman. Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5) :1189–1232, October 2001.

[139] Pierre Baldi and Søren Brunak. *Bioinformatics : The machine learning approach*. MIT Press, 1998.

[140] Mettez en place un cadre de validation croisée, https ://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308241-mettez-en-place-un-cadre-de-validation-croisee.

[141] Zhenyu Jia. Controlling the Overfitting of Heritability in Genomic Selection through Cross Validation. *Scientific Reports*, 7(1) :13678, October 2017. Number : 1 Publisher : Nature Publishing Group.

[142] Aarti Desai, Veer Singh Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, and Abhay Jere. Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. *PLOS ONE*, 8(4) :e60204, April 2013.

[143] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 7(1-2) :203–214, April 2000.

[144] Steven R. Gill, Derrick E. Fouts, Gordon L. Archer, Emmanuel F. Mongodin, Robert T. Deboy, Jacques Ravel, Ian T. Paulsen, James F. Kolonay, Lauren Brinkac, Mauren Beanan, Robert J. Dodson, Sean C. Daugherty, Ramana Madupu, Samuel V. Angiuoli, A. Scott Durkin, Daniel H. Haft, Jessica Vamathevan, Hoda Khouri, Terry Utterback, Chris Lee, George Dimitrov, Lingxia Jiang, Haiying Qin, Jan Weidman, Kevin Tran, Kathy Kang, Ioana R. Hance, Karen E. Nelson, and Claire M. Fraser. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. *Journal of Bacteriology*, 187(7) :2426–2438, April 2005.

[145] Julie Hurel, Sophie Schbath, Stéphanie Bougeard, Mathieu Rolland, Mauro Petrillo, and Fabrice Touzain. DUGMO : tool for the detection of unknown genetically modified organisms with high-throughput sequencing data for pure bacterial samples. *BMC Bioinformatics*, 21(1) :284, July 2020.

[146] Sarah Bigot, Omar A. Saleh, Christian Lesterlin, Carine Pages, Meriem El Karoui, Cynthia Dennis, Mikhail Grigoriev, Jean-François Allemand, François-Xavier Barre, and François Cornet. KOPS : DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. *The EMBO journal*, 24(21) :3770–3780, November 2005.

[147] Sophie Nolivos, Fabrice Touzain, Carine Pages, Michele Coddeville, Philippe Rousseau, Meriem El Karoui, Pascal Le Bourgeois, and François Cornet. Co-evolution of segregation guide DNA motifs and the FtsK translocase in bacteria : identification of the atypical Lactococcus lactis KOPS motif. *Nucleic Acids Research*, 40(12) :5535–5545, July 2012.

[148] Zhiqiang Zhang, Yi Zhao, Xiangke Liao, Wenqiang Shi, Kenli Li, Quan Zou, and Shaoliang Peng. Deep learning in omics : a survey and guideline. *Briefings in Functional Genomics*, 18(1) :41–57, 2019.

[149] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G. Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10) :1519–1525, October 2019.

[150] Paired-End vs. Single-Read Sequencing Technology, https ://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html. Library Catalog : www.illumina.com.

UNIVERSITÉ DE RENNES 1

Titre : Détection d'organismes génétiquement modifiés inconnus par analyse statistiques de données de séquençage haut débit

Mots clés : OGM inconnus, détection, statistiques

Résumé : L'Union Européenne a adopté une politique très restrictive vis-à-vis de la diffusion et de l'utilisation des organismes génétiquement modifiés (OGM), dont l'utilisation dans l'alimentation est mal acceptée par les consommateurs. Bien qu'un seuil maximal existe pour qu'un aliment soit étiqueté « sans OGM », ne sont aisément détectables que les OGM connus. Un OGM est constitué principalement d'un génome hôte et d'une séquence insérée par un procédé non naturel conférent une propriété particulière à l'organisme comme la résistance à certaines maladies. Depuis quelques années, des OGM dont la séquence insérée n'est pas connue ont été produits, non détectables par des approches utilisées jusqu'à présent (de type PCR). D'où la nécessité de créer un outil de détection d'OGM inconnus, objet de cette thèse, s'appuyant sur les avancées récentes en terme de séquençage haut débit.

Statistiquement, chaque organisme a une fréquence d'utilisation des nucléotides dans son génome qui lui est propre. Toute introduction de matériel génétique étranger va modifier localement les fréquences d'utilisation des nucléotides dans cette région, entraînant ainsi des fréquences d'utilisation des nucléotides différentes de celles de l'organisme hôte. En se basant sur cette affirmation, un outil de détection d'OGM inconnu a été mis au point à partir de données de séquençages bactériens dès lors que cet OGM résulte de l'insertion d'un gène étranger, de la troncation ou de la fusion d'un gène pouvant appartenir au génome hôte. L'outil a été testé sur 4 génomes bactériens OGM, 7 génomes bactériens sauvages et sur 42 génomes synthétiques. Les résultats démontrent l'efficacité de la méthode développée ne présentant qu'un gène faux positif et en identifiant plus de 99% des gènes d'inserts OGM.

Title : Detection of unknown genetically modified organisms by statistical analysis of high-throughput sequencing data

Keywords : unknown GMO, detection, statistics

Abstract : The European Union has adopted a very restrictive policy towards the dissemination and use of genetically modified organisms (GMOs), whose use in food is not well accepted by consumers. Although a maximum threshold exists for a food to be labelled "GM-free", only known GMOs are easily detectable. A GMO consists mainly of a host genome and a sequence inserted by a non-natural process that confers a particular property on the organism, such as resistance to certain diseases. In recent years, GMOs with an inserted sequence that is not known have been produced that are not detectable by approaches used until now (PCR-type). Hence the need to propose a tool for the detection of unknown GMOs, the subject of this thesis, based on recent advances in terms of high-throughput sequencing.

Statistically, each organism has a specific frequency of nucleotide use in its genome. Any introduction of foreign genetic material will locally alter the nucleotide use frequencies in that region, resulting in different nucleotide use frequencies compared to those of the host organism. Based on this assertion, an unknown GMO detection tool has been developed from bacterial sequencing data when the GMO results from the insertion of a foreign gene, the truncation or fusion of a gene that may belong to the host genome. The tool has been tested on 4 GMO bacterial genomes, 7 wild bacterial genomes and 42 synthetic bacterial genomes. The results demonstrate the effectiveness of the method developed by presenting only one false positive gene and identifying more than 99% of the genes of GMO inserts.