MASTER SCIENCES
ET NUMERIQUE POUR LA SANTÉ

Master 2 Sciences et Numérique pour la Santé
Parcours Bioinformatique, Connaissances, Données

HMSN307 : Projet BCD M2
Rapport bibliographique

Critères de qualité des assemblages des génomes

Auteur :
Anna Tran

Encadrantes :
Sèverine Bérard
Anne-Muriel Chifolleau
Tuteur :
Alban Mancheron

Version du
19 novembre 2019

Table des matières

Introduction

Depuis le développement des nouvelles technologies de séquençage, le coût du séquençage a considérablement diminué. Ces dernières permettent de générer de nombreuses données appelées *reads* (lectures). Un des enjeux liés à cette avancée est la reconstruction des génomes. Cette dernière est possible grâce à différentes étapes (Figure 1) qui permet l'assemblage de génomes.

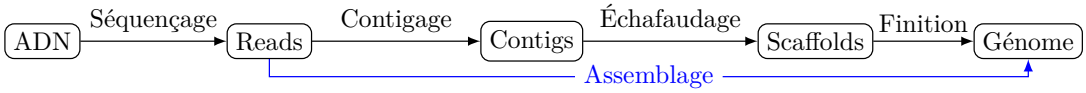

FIGURE 1 – Pipeline de reconstruction d'un génome

Il est possible de distinguer quatre étapes principales :

- Le séquençage qui permet d'obtenir les *reads* qui sont une succession de nucléotides de taille variable (quelques centaines de paires de bases pour les *reads* courts et quelques kilobases pour les *reads* longs).
- Le contigage qui permet la production de plus grand morceaux de génomes appelés contigs en utilisant les chevauchements entre les *reads*.
- L'échafaudage qui permet d'ordonner et d'orienter les contigs les uns par rapport aux autres afin d'obtenir les scaffolds.
- La finition qui permet de terminer proprement le génome en séquençant de nouveau autour des zones d'incertitudes.

Il est possible de retrouver les assemblages dans différentes bases de données dont Assembly créée par NCBI et ENA par EMBL. Celles-ci permettent d'obtenir des informations relatives aux assemblages notamment le niveau d'assemblage atteint ("génome complet", "chromosome", "scaffolds" ou "contigs"). La plupart des assemblages s'arrêtent au niveau "scaffolds" puisqu'il n'est pas toujours possible d'atteindre l'étape de finition. Cela s'explique par le fait que les scaffolds sont particulièrement sujets aux erreurs car générés à partir de *reads* courts[1]. Ainsi, il est important de pouvoir évaluer de manière correcte la qualité des assemblages afin de pouvoir les améliorer. Les données actuelles contiennent des erreurs dans les *reads* (substitution, insertion, ...), des zones difficiles à séquencer et des "données brutes" différentes. De plus, le contigage ainsi que l'échafaudage sont des heuristiques et donc le résultat obtenu en sortie n'est pas exact. Par ailleurs, plus d'une équipe peut travailler sur un même génome. Il est donc possible d'avoir plusieurs jeux de données différents provenant de multiples sources. Ainsi, pour chaque organisme, plusieurs assemblages d'un génome sont susceptibles d'être disponibles.

Il a été suggéré que faire du "méta-assemblage" en combinant plusieurs assemblages d'un même génome pourrait permettre d'améliorer la qualité des assemblages[2]. Ainsi dans le cadre de ce projet bibliographique, l'intérêt s'est essentiellement porté sur la recherche des métriques permettant l'évaluation de la qualité. Les métriques ont été divisées en deux catégories : il y a celles dites "quantitatives" qui concernent essentiellement l'analyse de l'assemblage et celles que nous avons appelés "qualitatives" qui

dérivent de l'annotation du génome. Pour évaluer la qualité, il existe différents outils : QUAST[3] et BUSCO[4] qui ont permis de mettre en avant un certain nombre de métriques classiquement prises en compte. Néanmoins, d'autres métriques ont été également étudiées afin d'élargir les critères par rapport aux méthodes existantes de fusion des assemblages.

Les outils permettant d'évaluer la qualité des assemblages (QUAST[3] et BUSCO[4]) feront l'objet d'une première partie. Les métriques quantitatives étudiées seront abordées dans une seconde partie. Enfin, les perspectives seront détaillées en dernière partie.

1 Exploration des méthodes d'évaluation existantes

Différents outils ont été développés afin de pouvoir évaluer la qualité de l'assemblage. Cette première partie concernera les deux principaux outils utilisés qui sont QUAST[3] et BUSCO[4]. Les métriques seront détaillées plus amplement dans la section 2 ("Critères d'évaluation quantitatifs de l'assemblage").

1.1 QUAST

QUAST est un outil d'évaluation de métriques quantitatives permettant ainsi aux utilisateurs d'avoir une vue d'ensemble sur la qualité du génome assemblé.

QUAST v4.*[3] calcule des métriques (N50, taille du contig/scaffold le plus long, ...) tout comme dans certaines méthodes existantes, notamment Plantagora[5], GAGE[6], GeneMark.hmm[7] et GlimmerHMM[8]. Par ailleurs, il introduit une nouvelle métrique : le NAx. Il utilise le module NUCmer de MUMmer[9] afin d'aligner l'assemblage sur une séquence de référence et ainsi évaluer des métriques. Il calcule également des métriques pour les assemblages possédant une annotation ou ne présentant pas de séquence de référence.

Une nouvelle version de QUAST v5.*[10] a été développée. Cette dernière permet d'analyser de plus grands génomes ainsi que la prise en compte de nouvelles métriques (plus détaillées dans la section 2.4) basées sur les k-mers et des spécificités liées aux génomes eucaryotes (abondance des transposons...). Une des autres améliorations est le concept d'assemblage supérieur (*upper bound assembly*) basé sur le fait que la référence ne peut pas être totalement reconstruite grâce à des *reads* étant donné les répétitions et les régions peu couvertes du génome. L'assemblage supérieur permet d'estimer la limite supérieure de complétude et contiguïté qui peut être atteinte théoriquement par un logiciel d'assemblage en partant d'un ensemble de *reads* donné.

K-mer : séquence de longueur k

Pour la construction de l'assemblage supérieur (Figure 2), les *reads* sont dans un premier temps alignés sur la référence afin de détecter les régions qui ne présentent pas de couverture. Par la suite, les répétitions dites longues sont marquées. Ces dernières sont des répétitions dont la taille excède la taille médiane de l'insert de la librairie *paired-end*. Parmi les séquences répétées, seules celles qui sont répétées au moins deux fois dans le génome sont conservées. Ces longues répétitions peuvent causer des ambuiguités et ne peuvent être résolues que grâce à des *reads* longs ou *mate-paires*. Les fragments de la séquence génomique ne présentant pas de répétitions sont appelés *upper bound* contigs. Les chevauchements entre les contigs et les *reads* longs ou *mate-pair* permettent d'obtenir les *upper bound* scaffolds. Une fois que les *upper bound* scaffolds sont obtenus, les trous adjacents aux contigs sont remplis par des séquences correspondantes provenant du génome de référence ou par des N le cas échéant. Il faut noter que

Paired-end : séquençage par paires avec des *reads* séparés par une distance connue appelée insert

les estimations des limites supérieures ne concernent que les métriques basées sur l'alignement et ne s'appliquent pas aux analogues de ces métriques calculés sans référence.

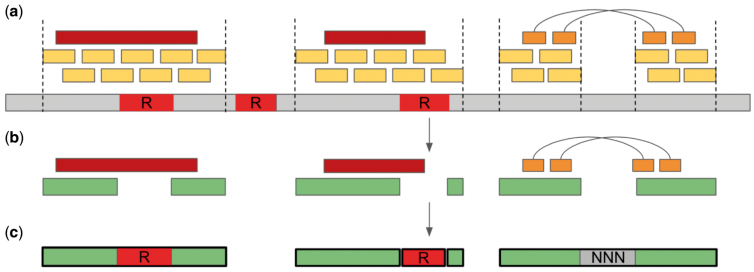

FIGURE 2 – Construction de l'assemblage supérieur
Légende : séquence répétée (rouge), *read* long (marron), *paired-ends reads* (jaune), *mate-pairs reads* (orange), contigs (vert), séquence de référence (gris) Source : Versatile genome assembly evaluation with QUAST-LG, Bioinformatics. 2018 ;34(13) :i142-i150. doi :10.1093/bioinformatics/bty266

Par ailleurs, dans cette version de QUAST, l'aligneur NUCmer a été remplacé par Minimap2[11] pour les *reads* longs (issus des technologies PacBio ou Nanopore) afin de maintenir une vitesse d'exécution similaire pour différentes tailles de génome, notamment les plus grands pour lesquels la version précédente ne marchait pas. Pour les *reads* courts issus de la technologie Illumina, BWA-MEM[12] est utilisé.

1.2 BUSCO

BUSCO[4] est une autre stratégie pouvant être utilisée afin d'évaluer la qualité des assemblages et des annotations. Pour cela, elle utilise les copies uniques d'orthologues universelles afin d'estimer la complétude de l'annotation d'un génome ainsi que des métriques similaires à celles présentées dans QUAST via l'utilisation de la base OrthoDB[13].

ORTHOLOGUES : gènes homologues (hérités d'un ancêtre commun) issus de la spéciation

OrthoDB v9.1 inclut un total de 5 756 espèces, fournissant ainsi des groupes d'orthologues pour les clades de : 3 663 bactéries, 330 métazoaires, 227 champignons, 31 plantes, 345 archées et 1 157 virus. Parmi les métazoaires, il y a maintenant 172 verterbrés et 133 arthropodes.

BUSCO utilise :
— BLAST+[14] pour la recherche de séquences
— Augustus[15] pour la prédiction des gènes basés sur les profils de blocs
— HMMER[16] pour la recherche des profils issue des modèles de Markov cachés

Pour évaluer la qualité des assemblages, les régions candidates sont recherchées via tBLASTn à l'aide de séquences consensus de BUSCO. Ces séquences consensus sont dérivées de profils de modèle de Markov cachés (HMM) construits à partir des alignements multiples de séquences d'orthologues et capturent les acides aminés pouvant être alignés et conservés dans l'ensemble des espèces. Les séquences orthologues sont choisies parmi les groupes orthologues OrthoDB des principales espèces, nécessitant la présence d'orthologues sous forme de gènes à copie unique dans la grande majorité (>90%) des espèces.

Par la suite, la structure des gènes est prédite grâce à Augustus avec les profils de bloc BUSCO. Enfin, ces gènes prédits sont évalués en utilisant les profils HMMER ainsi que les profils BUSCO lignées-spécifiques et sont classifiés (Figure 3).

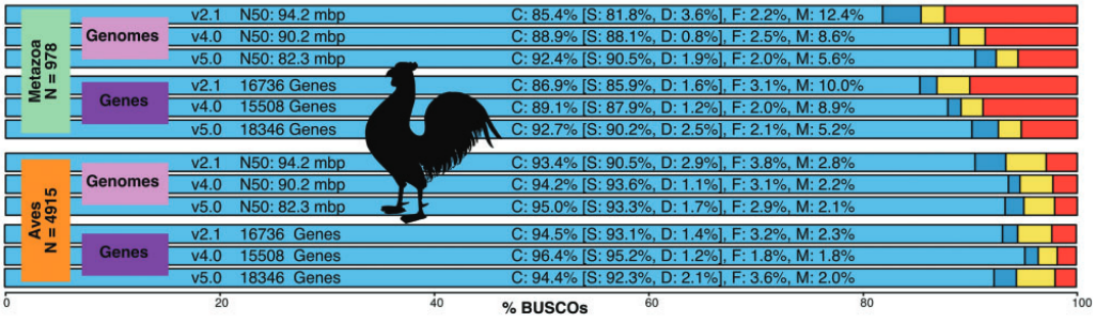

FIGURE 3 – Exemple de classification BUSCO : Comparaison de différentes versions du génome de poulet et de leurs ensembles de gènes annotés avec la lignée des oiseaux (*Aves*) et des métazoaires (*Metazoa*)

Légende : Les graphiques montrent les proportions de gènes classés comme complets (bleu clair et foncé), complets en copie unique (bleu clair), complets dupliqués (bleu foncé), fragmentés (jaune) et manquants (rouge). Source : BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution, 35(3), 543-548.

2 Critères d'évaluation quantitatifs de l'assemblage

Différentes publications proposent des métriques dont certaines sont universelles et d'autres plus spécifiques selon le procédé d'analyse. Pour la plupart des méthodes existantes, il est nécessaire d'avoir accès à une séquence de référence ou aux fichiers contenant les *reads*. Nous allons voir en détail dans cette partie les métriques calculées par QUAST[3] et BUSCO[4].

2.1 Métriques basées sur l'analyse des contigs

Les métriques basées sur l'analyse des contigs ou des scaffolds (Table 1) sont calculées à partir des données d'assemblage. Elles sont classiquement calculées de manière systématique. Par exemple le N50 est une métrique de référence notamment citée dans l'Assemblathon 2[2] et QUAST[3].

Ces métriques peuvent être calculées sans séquence de référence sauf pour le NGx. QUAST[3] propose également pour ses métriques de filtrer selon la taille des contigs afin d'éliminer ceux de taille courte qui ne sont que peu utilisés.

Parmi ces métriques, il est possible de retrouver le nombre de contigs, la taille du contig le plus long, la longueur totale de l'assemblage, le Nx, le NGx, le NAx, le nombre de contigs ayant une taille supérieure à x et la somme des longeurs de contigs ayant une taille supérieur à x.

La nouvelle métrique introduite par QUAST[3] est basée sur les blocs alignés dérivés du Nx (x allant de 0 à 100) est appelée NAx. Le Nx est défini comme la taille du contig de longueur L pour laquelle les contigs de taille supérieure ou égale à L couvrent au moins x% de l'assemblage (exemple du N50 en figure 4).

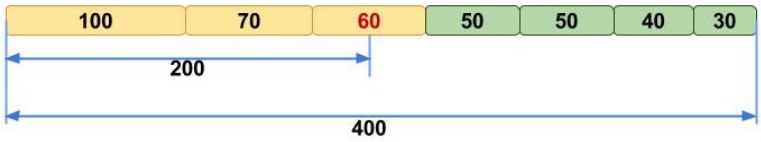

FIGURE 4 – Représentation du N50 : La taille du génome vaut ici 400 paires de bases (pb). Les différents contigs (jaunes et verts) sont ordonnés dans l'ordre décroissant selon leur taille. Pour couvrir 50% du génome, il faut que les tailles cumulées des contigs couvrent 200 pb. Le N50 vaut ici 60 pb.
Source : What's N50 ? `https://www.molecularecologist.com/2017/03/whats-n50/`

Pour calculer le NAx, les contigs sont séparés en blocs d'alignement au niveau des points de cassure. Un point de cassure est défini par :

- Deux ou plus alignements distincts couvrant le point de cassure
- Un écart entre la séquence flanquante de droite et celle de gauche inférieur à 1000 bases
- Les régions flanquantes de gauche et de droite sur le même brin et chromosome que la référence

Les NAx sont calculés à partir de ces blocs et non pas à partir des contigs originaux.

Métrique	**Séquence de référence**	**Explication**
Nombre de contigs		
Taille du contig le plus long		
Longueur totale de l'assemblage		
Nx (x compris entre 0 et 100)		Taille de contig de longueur minimum L pour laquelle les contigs de taille supérieure ou égale à L couvrent au moins x % de l'assemblage
NGx (x compris entre 0 et 100)	x	Taille de contig de longueur minimum L pour laquelle les contigs de taille supérieure ou égale à L couvrent au moins 50% de la référence
NAx (x compris entre 0 et 100)		Taille de contig de longueur minimum L pour laquelle les contigs de taille supérieure ou égale à L couvrent au moins x % de l'assemblage calculé à partir des blocs alignés
Nombre de contigs >x nucléotides		Nombre de séquences dont la taille est supérieure à x
Somme des longueurs de contigs >x nucléotides		Somme des longueurs de contigs avec une taille supérieure à x

TABLE 1 – Tableau récapitulatif des métriques basées sur l'analyse des contigs

2.2 Métriques basées sur les assemblages incohérents et les variations structurales

Les métriques abordées dans cette partie (Table 2) décrivent les variations structurales dans les contigs. Ces dernières nécessitent la disponibilité d'une séquence de référence.

Ces différentes métriques ont pour but de détecter les erreurs (mauvais assemblage par un outil ou erreurs de séquençage) et de potentiellement mettre en avant des variations structurales au sein des différents assemblages (dans ce cas-ci, entre l'assemblage étudié et la référence). En effet, les génomes peuvent être sujets à des arrangements structuraux, à des larges insertions et délétions et à des répétitions.

Métrique	Séquence de référence	Explication
Nombre d'assemblages incohérents	x	Nombre de positions dans l'assemblage où la séquence flanquante gauche s'aligne à plus de 1 kpb de la séquence flanquante droite sur la référence (relocalisation) ou se chevauchent sur plus de 1 kpb (relocalisation) ou les séquences flanquantes s'alignent sur différents brins (inversion) ou différents chromosomes (translocation)
Nombre de contigs contenant des erreurs d'assemblage	x	Nombre de contigs contenant des assemblages incohérents
Nombre de contigs non alignés	x	Nombre de contigs ne s'alignant par sur la référence
Nombre de contigs mappés de manière ambigüe	x	Nombre de contigs pouvant s'aligner à différentes localisations sur la référence

TABLE 2 – Tableau récapitulatif des métriques basées sur les assemblages incohérents et les variations structurales

2.3 Métriques basées sur la représentation du génome

Les métriques basées sur la représentation du génome et des éléments fonctionnels (Table 3) requièrent pour la plupart une séquence de référence.

Métrique	Séquence de référence	Explication
Fraction du génome	x	Nombre de bases total divisé par la taille de la référence
Ratio de duplication	x	Nombre de bases alignées (nombre total moins les bases non alignées) divisé par la taille de la référence
% GC		Pourcentage de bases G et C
Nombre de mismatch pour 100 kb	x	Nombre moyen de mismatch pour 100 000 bases alignées
Nombre d'indels pour 100 kb	x	Nombre moyen d'insertions ou délétions d'un nucléotide pour 100 000 bases alignées
Nombre de gènes	x	Nombre de gènes dans l'assemblage (complet ou partiel) basé sur l'annotation du génome
Nombre d'opérons	x	Nombre d'opérons similaire à des gènes basé sur l'annotation du génome
Nombre de gènes prédits		Nombre de gènes prédits par le module prédiction de QUAST
Nombre de gènes constitutifs manquants		
Nombre moyen d'orthologues par gènes constitutifs		
% de gènes constitutifs détectés ayant plus d'un orthologue		
Score BUSCO		Évaluation quantitative en terme de contenu génique attendu

TABLE 3 – Tableau récapitulatif des métriques basées sur la représentation du génome

Des précisions peuvent être apportées à ces différentes métriques. Dans le cas de la fraction du génome, une base dans la référence est considérée comme alignée si au moins un contig présente au moins un alignement pour la base. Les contigs issus de régions répétées peuvent potentiellement s'aligner sur la référence à plusieurs endroits. Ainsi, la base est comptée de multiples fois dans la quantité.

Concernant le ratio de duplication, un ratio supérieur à 1 pourrait indiquer que l'assemblage contient des contigs couvrant la même région du génome de référence.

Sans distinction, les *mismatchs* tiennent compte à la fois du polymoprhisme nulcéotidique (*Single Nucleotide Polymorphisms* SNP) qui ont une réalité biologique et des erreurs de séquençage.

Pour le nombre de gènes, le calcul est basé sur une annotation du génome de référence fournie par l'utilisateur. Un gène est considéré comme partiellement couvert si l'assemblage ne contient pas le gène complet mais au moins 100 paires de bases (bp) du gène. Si aucune liste n'est fournie, le calcul du nombre de gènes prédits est fait à partir des résultats de GeneMark.hmm[7] pour les génome procaryotes et GlimmerHMM[8] pour les génomes eucaryotes.

Concernant le score BUSCO, les correspondances sont classées comme "complètes" (C) si la longueur correspond à la longueur attendue du profil issu d'OrthoDB. Si elles ne sont retrouvées qu'une fois, elles sont classées dans la catégorie "copie unique" (S) sinon elles sont classées comme "dupliquées" (D) quand elles se trouvent en plusieurs copies. Les correspondances partiellement récupérées sont classées comme "fragmentées" (F). Enfin les séquences qui ne présentent pas d'orthologie sont classés comme "manquantes" (M). Ainsi, le format obtenu est le suivant : C : x%[S : x%,D : x%],F : x%,M : x%.

2.4 Métriques basées sur les k-mers

Les métriques rencontrées dans la table 4 sont basées sur l'analyse des k-mers uniques présents dans la séquence de référence et dans l'assemblage. Si la valeur de k est suffisamment grande, les k-mers uniques semblent être largement répandus sur le génome ce qui permet d'apprécier sa complétude ainsi que son exactitude. Ces métriques ont pour avantage d'éviter les incohérences liées aux éléments transposables notamment le mapping à de multiples endroits des *reads*. Les éléments transposables ne contenant que peut de k-mers uniques.

ÉLÉMENT TRANSPOSABLE : séquence répétée dans le génome

QUAST[3] utilise KMC3[17] afin de recenser les k-mers uniques dans la séquence de référence. Le pourcentage de ces k-mers détectés dans l'assemblage représente la complétude.

L'exactitude est calculée à partir d'un petit sous-ensemble de k-mers distribués uniformément. Ce sous-ensemble est sélectionné de façon à ce que tous les k-mers soient distants d'au moins 1 kilobase (kb) sur la référence. Le sous-ensemble sélectionné est soumis à KMC3 qui recherche les contigs possédant au moins deux k-mers provenant de la sélection. La position de chaque k-mer détecté est examinée et comparée à une liste de k-mers consécutifs. Un marqueur est défini comme la position d'un k-mer sur un contig C dans une liste de k-mers consécutifs k_1, k_2, ..., k_n. La distance entre k_i et k_{i+1} est semblable entre le contig et la référence à plus ou moins 5%. Afin de vérifier la corrélation, la position des marqueurs est également analysée : m_1 et m_{i+1} doivent avoir une position similaire à celle sur la référence.

Par ailleurs, les mauvais raccordements sont défini comme étant des translocations et relocations basées sur les k-mers. QUAST traite un mauvais raccordement comme une translocation lorsque les marqueurs sont situés sur deux chromosomes différents. Si les deux marqueurs sont présent sur un même chromosome et qu'il y a une incohérence entre la référence et le contig alors cela sera considéré comme une relocation.

Métrique	Séquence de référence	Explication
Complétude basée sur les k-mers	x	Pourcentage de k-mers uniques provenant de la référence retrouvés dans l'assemblage
Nombre de k-mers mal raccordés	x	Nombre total de mauvais raccordements basés sur les k-mers issus de l'assemblage
Longueur correcte basée sur les k-mers (%)	x	Pourcentage de la longueur totale de tous les contigs considérés comme corrects étant donné l'analyse des k-mer uniques
Longueur basée sur les k-mers mal raccordés (%)	x	Pourcentage de la longueur totale de tous les contigs contenant au moins un k-mer mal raccordé
Longueur totale des contigs sans marqueurs basés sur les k-mers(%)	x	Pourcentage de la longueur totale de tous les contigs sans marqueurs de k-mer

TABLE 4 – Tableau récapitulatif des métriques basées sur les k-mers

2.5 Autres métriques

Au delà des métriques vues précédemment, d'autres métriques ont été proposées notamment dans le cadre de l'Assemblathon 2[2] par exemple qui est un exercice d'assemblage du génome qui utilise des *reads* de séquençage à partir d'un mélange de technologies NGS. D'autres méthodes permettent également d'étudier de nouvelles métriques, sans génome de référence.

2.5.1 Métriques issues de l'Assemblathon 2

Dans le cadre de l'Assemblathon 2, trois génomes de vertébrés ont été étudiés : *Meiopsittacus unduiatus* (oiseau), *Mayiandia zebra* (poisson) et *Boa constrictor constrictor* (serpent). Afin de comparer et classer les assemblages fournis par les différentes équipes, dix métriques ont été analysées (Figure 5).

FIGURE 5 – Métriques analysées au cours de l'assemblathon 2

Parmi les métriques, le NG50 des contigs et scaffolds est retrouvé ainsi que le nombre de gènes constitutifs. Ces dernières sont des métriques retrouvées de manière systématique. D'autres métriques se basent également sur l'analyse des scaffolds, des *reads* via l'outil REAPR et de données expérimentales, notamment les cartes optiques et les fosmides.

FOSMIDE : plasmide hybride issu d'*E. coli*

Une métrique est dérivée de l'analyse des scaffolds. En utilisant une longueur approximative de 25 kbp pour un gène moyen de vertébré, il a été calculé le pourcentage de la taille estimée de l'assemblage étant composé de scaffolds assez grands pour contenir un gène.

Le principe de la carte optique consiste à établir un profil enzymatique de molécules d'ADN génomique basé sur la taille et l'ordre des fragments obtenus après digestion et/ou marquage. La compilation de plusieurs profils issus de plusieurs molécules permet d'établir une carte complète de restriction d'un génome, appelée carte optique d'un génome complet (« Whole Genome Map »). Les applications d'une carte optique sont multiples. Une des premières applications a été la validation et l'amélioration d'assemblage de novo.

Deux métriques sont issues de l'analyse des cartes optiques :

— Couverture de niveau 1 : indicateur de la précision de l'assemblage global. Plus la couverture est élevée, mieux cela est.

— Niveaux de couverture 1 + 2 + 3 : indiquent la quantité d'assemblage correctement alignée sur une carte optique, même si elle est due à des échafaudages chimériques. Plus le niveau est haut, plus l'assemblage est bon.

L'analyse des fosmides est également un moyen de valider expérimentalement l'assemblage. Les fosmides permettent de cloner efficacement et de maintenir de manière stable l'ADN inséré. La constitution de banque d'ADN de génomes complexes est ainsi facilitée avec une représentation plus complète et une analyse plus rapide de la structure

génomique.

Trois métriques sont issus de l'analyse des fosmides :

— Couverture du fosmide : calculée à l'aide de l'outil COMPASS[18], elle reflète la quantité de région validée du fosmide (VFR) capturés par les assemblages
— Validité du fosmide : calculée à l'aide de l'outil COMPASS, mesure la quantité de l'ensemble pouvant être validée par les VFR (plus elle est élevée, mieux c'est.)
— Score de synthèse des étiquettes VFR de scaffold (paires de courtes séquences aux extrémités de chaque séquence VFR) : nombre de paires d'étiquettes VFR correspondant à un même scaffold, multiplié par le pourcentage de paires d'étiquettes mappant de manière unique à la bonne distance. Plus le pourcentage est élevé mieux cela est.

L'outil REAPR[19] a également été utilisé pour analyser la qualité des assemblages en utilisant les *paired-end reads* mappés sur l'assemblage afin d'évaluer un ensemble de métriques pour chaque base de l'assemblage. Les *reads* courts permettent de mesurer les erreurs locales tels que les SNP ainsi que les petites insertions ou délétions. Quant aux *reads* longs, ils sont utilisés pour localiser les erreurs structurales tels que les translocations. Une métrique a été associée à l'outil REAPR par l'attribution d'un score récapitulatif.

Certaines de ces métriques n'ont pas pu être appliquées à toutes les séquences en particulier celles concernant les fosmides. Ainsi, seuls sept métriques ont été évaluées parmi les dix citées. De plus, il faut noter que les métriques vues ici ont été appliquées à des organismes spécifiques. Elles ne sont donc pas forcément applicables à tous les organismes de manière générale.

2.5.2 Métriques calculées en absence de référence

La majorité des métriques présentées dans les sections précédentes nécessitent la disponibilité d'un génome de référence. Or, cela pose un problème pour l'évaluation de la qualité des séquences obtenues *de novo*. Ainsi, il a été proposé une méthode sans utilisation d'une référence. Cette dernière est un module issu de la méthode SGA[20].

Les métriques calculées sont les estimations :

— de la taille attendue du génome.
— de la distribution de la taille des inserts.
— du taux d'erreur par base.

L'étude de ces métriques permet à la fois d'évaluer la qualité des données prises en entrée des logiciels mais cela pourrait également permettre d'améliorer l'assemblage en sélectionnant ainsi un jeu de données qui permettrait de maximiser les métriques étudiées.

Dans un premier temps, la méthode utilise le FM-Index afin de compter le nombre d'occurrences d'un motif particulier P de la librairie de *reads*. Par la suite, ces derniers sont échantillonnés de manière aléatoire afin que les calculs des métriques soient efficaces sans avoir à tester le jeu de données complet. Pour chaque métrique, les calculs sont menés sur plusieurs échantillons pour que les résultats soient pertinents.

FM-Index : structure de compression basée sur la Transformée de Burrows-Wheeler permettant de compter le nombre d'occurrences d'un motif et de les localiser

Afin d'estimer la taille du génome, les k-mers sont comptés puis le calcul suivant

est appliqué :

$$G = (1 - \widehat{w_0}) \frac{n(l-k+1)}{\widehat{\lambda}}$$

avec G l'estimation de la taille de génome, $n(l-k+1)$ le nombre de k-mers total, $\widehat{w_0}$ l'estimation de la proportion de k-mers contenant des erreurs et $\widehat{\lambda}$ l'estimation de la moyenne d'apparition d'un k-mer.

Afin d'aider à résoudre les longues répétitions au sein du génome, les paires de lectures sont obtenues en séquençant les deux extrémités d'un fragment d'ADN. La gamme des tailles de *reads* est déterminé lors de la préparation de l'ADN pour le séquençage.

Pour s'assurer que la distribution des tailles correspond à la distribution attendue, celle-ci peut être estimée. Dans un premier temps, une paire de *reads* X et Y est échantillonnée à partir du FM-Index. Une recherche gloutonne est effectuée sur le graphe de De Bruijn simulé à partir des k-mers afin de trouver le premier k-mer de taille 51 [1] du *read* X en sélectionnant la branche ayant la couverture la plus élevée comme prochain sommet de recherche. La recherche s'arrête lorsque le premier k-mer de taille 51 du *read* Y est rencontré. Si une marche complète de X à Y est trouvée, la longueur de la marche correspond à la taille du fragment en nucléotides.

Les erreurs de séquençage peuvent altérer l'assemblage. La plupart des méthodes permettent la prise en compte de *mismatches* ou de trous menant ainsi à la présence de faux positifs (*reads* ayant été faussement reliés) parmi les contigs/scaffolds dans l'assemblage.

Pour estimer le taux d'erreurs au sein d'un *read*, les chevauchements *read-read* qui possèdent en partie des alignements exacts sont analysés. Un *read* R est d'abord échantillonné à partir du FM-Index et les *reads* ayant des k-mers de longueur 31 communs avec ce dernier sont recherchés. Le chevauchement doit avoir une longueur d'au moins 50 paires de bases (bp) et le pourcentage d'identité doit être d'au moins 95%. Un alignement multiple est construit à partir de R et du chevauchement par paire pour les *reads* atteignant les critères cités précédemment. Enfin, pour chaque colonne de l'alignement, une séquence consensus est déterminée. Une base $R[j] = b$ est considérée comme incorrecte si :

— b ne correspond pas à la base d'indice j de la séquence consensus.
— au moins trois *reads* supportent la base de la séquence consensus.
— moins de quatre *reads* supportent la base b.

Le taux d'erreur à la position j est calculé de la manière suivante :

$$\epsilon_j = \frac{\sum_{i=1}^{M} I[\text{base j incorrecte dans le read i}]}{M}$$

2.6 Discussion

Les deux principaux outils utilisés pour l'évaluation de la qualité des assemblages sont QUAST et BUSCO. Les métriques calculées par QUAST nécessitent pour la plupart la présence d'une séquence de référence. Dans le cas de BUSCO, il faut des références d'organismes proches de celui étudié mais aussi générer une base de données contenant les orthologues. Ces métriques sont pour la plupart des métriques dites quantitatives qui se basent essentiellement sur l'analyse de la séquence.

1. Les tailles sont celles définis par défaut dans la méthode.

L'ensemble des métriques sont présentées dans l'annexe A. Certaines métriques sont communes aux deux méthodes et d'autres sont spécifiques. Les plus communes telles que la taille des contigs, le N50, le nombre de gènes prédits ou encore le pourcentage de GC sont présentées par les deux outils. D'autres métriques sont plus spécifiques. Pour QUAST, l'analyse de la séquence est plus poussée notamment via l'analyse des k-mers. Pour BUSCO, l'analyse se tourne essentiellement sur la composition du génome.

Différentes études ont montré que d'autres critères pouvaient être pris en compte comme vu dans l'Assemblathon 2. Il peut être intéressant au delà des métriques calculés à l'aide de l'outil informatique d'avoir accès à des données expérimentales permettant de confirmer ou d'infirmer une hypothèse notamment en utilisant par exemple des cartes optiques ou les fosmides. Par ailleurs, une majorité des métriques est calculée à partir d'une séquence de référence ou en faisant l'homologie avec un organisme proche. Il est difficile de calculer des métriques *de novo*.

3 Conclusion et perspectives

Les outils analysés ainsi que les métriques vues au cours de ce rapport bibliographique montrent une certaine diversité parmi ce qui peut être étudié. En effet, l'analyse peut se faire au niveau des contigs/scaffolds, de l'assemblage complet, de la représentation du génome mais aussi au niveau des k-mers.

Ces outils permettent aux biologistes d'analyser un certain nombre de métriques. Néanmoins, peu d'entre elles sont réellement étudiées. En effet, l'expérience a montré que l'évaluation de la qualité ne se faisait que sur un nombre limité de critères quantitatifs notamment :

- le nombre de contigs/scaffolds (moins il y en a, mieux cela est)
- la taille des contigs : maximum, moyenne, médiane et N50
- le nombre de "N" pour remplir les trous dans l'assemblage (moins il y en a, mieux cela est)

D'autres métriques liées à l'annotation sont également étudiées par exemple le nombre de gènes identifiés ainsi que la proportion de gènes conservés.

Parmi les critères qualitatifs (liés à l'annotation), il est possible de s'intéresser à l'annotation à proprement parler des gènes ainsi qu'à l'annotation dérivant de la *Gene Ontology* (GO).

Gene Ontology : structuration de la description des gènes et des produits géniques

Concernant l'annotation des gènes, deux stratégies ont pu être observées. Dans un premier cas, il est possible de comparer les gènes prédits grâce à différents logiciels (par exemple : Twinscan et Genscan) afin de déterminer leurs paramètres optimaux et ainsi obtenir de meilleures statistiques concernant les caractéristiques des gènes tels que la taille du gène, la taille de l'ARNm, la longueur des introns et exons, etc... via la méthode EVAL[21]. Cette dernière fait parti d'une suite d'outils nommé AEGeAn qui permet l'analyse du génome pour différents loci.

Locus : Emplacement précis d'un gène sur le chromosome qui le porte

ParsEval[22] est une autre méthode permettant la comparaison de deux sets d'annotation en calculant des statistiques sur les exons, les séquences codantes (CDS) et les régions non-traduites (UTR).

Enfin, au niveau de la GO, un score peut être défini pour évaluer les termes associés à l'assemblage[23]. Ce dernier peut être utilisé afin de suivre les changements dans les annotations GO au fil du temps mais aussi évaluer la qualité des annotations GO disponibles pour des processus biologiques spécifiques.

L'interprétation de la combinaison des différentes métriques reste difficile pour les biologistes et les outils de méta-assemblage restent difficile à utiliser. Ainsi au cours du stage du second semestre, l'étude plus fine des critères qualitatifs liés à l'annotation, la recherche de nouvelles métriques ainsi que la comparaison d'assemblages avec la prise en compte de critères supplémentaires par rapport aux méthodes existantes de fusion pourraient permettre à terme d'améliorer les méta-assemblages en sélectionnant un certain nombre de critères.

Références

[1] Martin Hunt et al. « A comprehensive evaluation of assembly scaffolding tools ». In : *Genome Biology* 15.3 (mar. 2014), R42. doi : 10.1186/gb-2014-15-3-r42.

[2] Keith R Bradnam et al. « Assemblathon 2 : evaluating de novo methods of genome assembly in three vertebrate species ». In : *GigaScience* 2.1 (2013), p. 10.

[3] Alexey Gurevich et al. « QUAST : quality assessment tool for genome assemblies ». In : *Bioinformatics* 29.8 (2013), p. 1072–1075. doi : 10.1093/bioinformatics/btt086. eprint : /oup/backfile/content_public/journal/bioinformatics/29/8/10.1093_bioinformatics_btt086/2/btt086.pdf.

[4] Felipe A. Simão et al. « BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs ». In : *Bioinformatics* 31.19 (2015), p. 3210–3212. doi : 10.1093/bioinformatics/btv351. eprint : /oup/backfile/content_public/journal/bioinformatics/31/19/10.1093_bioinformatics_btv351/2/btv351.pdf.

[5] Roger Barthelson et al. « Plantagora : Modeling Whole Genome Sequencing and Assembly of Plant Genomes ». In : *PLOS ONE* 6.12 (déc. 2011), p. 1–8. doi : 10.1371/journal.pone.0028436.

[6] Weijun Luo et al. « GAGE : generally applicable gene set enrichment for pathway analysis ». In : *BMC Bioinformatics* 10.1 (mai 2009), p. 161. doi : 10.1186/1471-2105-10-161.

[7] Alexander V Lukashin et Mark Borodovsky. « GeneMark. hmm : new solutions for gene finding ». In : *Nucleic acids research* 26.4 (1998), p. 1107–1115.

[8] William H Majoros, Mihaela Pertea et Steven L Salzberg. « TigrScan and GlimmerHMM : two open source ab initio eukaryotic gene-finders ». In : *Bioinformatics* 20.16 (2004), p. 2878–2879.

[9] Arthur L Delcher, Steven L Salzberg et Adam M Phillippy. « Using MUMmer to identify similar regions in large sequence sets ». In : *Current protocols in bioinformatics* 1 (2003), p. 10–3.

[10] Alla Mikheenko et al. « Versatile genome assembly evaluation with QUAST-LG ». In : *Bioinformatics* 34.13 (2018), p. i142–i150. doi : 10.1093/bioinformatics/bty266. eprint : /oup/backfile/content_public/journal/bioinformatics/34/13/10.1093_bioinformatics_bty266/1/bty266.pdf.

[11] Heng Li. « Minimap2 : pairwise alignment for nucleotide sequences ». In : *Bioinformatics* 1 (2018), p. 7.

[12] Heng Li. « Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM ». In : *arXiv preprint arXiv :1303.3997* (2013).

[13] Robert M Waterhouse et al. « OrthoDB : a hierarchical catalog of animal, fungal and bacterial orthologs ». In : *Nucleic acids research* 41.D1 (2012), p. D358–D365.

[14] Christiam Camacho et al. « BLAST+ : architecture and applications ». In : *BMC bioinformatics* 10.1 (2009), p. 421.

[15] Mario Stanke et al. « AUGUSTUS : ab initio prediction of alternative transcripts ». In : *Nucleic acids research* 34.suppl_2 (2006), W435–W439.

[16] Robert D Finn, Jody Clements et Sean R Eddy. « HMMER web server : interactive sequence similarity searching ». In : *Nucleic acids research* 39.suppl_2 (2011), W29–W37.

[17] Marek Kokot, Maciej Długosz et Sebastian Deorowicz. « KMC 3 : counting and manipulating k-mer statistics ». In : *Bioinformatics* 33.17 (2017), p. 2759–2761.

[18] Andrew Low, Nicolas Rodrigue et Alex Wong. « COMPASS : the COMPletely Arbitrary Sequence Simulator ». In : *Bioinformatics* 33.19 (2017), p. 3101–3103. doi : 10.1093/bioinformatics/btx347. eprint : /oup/backfile/content_public/journal/bioinformatics/33/19/10.1093_bioinformatics_btx347/4/btx347.pdf.

[19] Martin Hunt et al. « REAPR : a universal tool for genome assembly evaluation ». In : *Genome biology* 14.5 (2013), R47.

[20] Jared T Simpson. « Exploring genome characteristics and sequence quality without a reference ». In : *Bioinformatics* 30.9 (2014), p. 1228–1235.

[21] Evan Keibler et Michael R Brent. « Eval : a software package for analysis of genome annotations ». In : *BMC bioinformatics* 4.1 (2003), p. 50.

[22] Daniel S Standage et Volker P Brendel. « ParsEval : parallel comparison and analysis of gene structure annotations ». In : *BMC bioinformatics* 13.1 (2012), p. 187.

[23] Teresia J Buza et al. « Gene Ontology annotation quality analysis in model eukaryotes ». In : *Nucleic acids research* 36.2 (2008), e12–e12.

A Tableau récapitulatif des métriques présentes dans BUSCO et/ou QUAST

	Métrique	Séquence de référence nécessaire	Explication	BUSCO	QUAST
Analyse des contigs	Nombre de contigs			x	x
	Contig le plus long			x	x
	Longueur totale			x	x
	N50 (ou 25, 75)		Taille du contig de longueur L pour laquelle les contigs de taille supérieur ou égale à L couvrant au moins 50% de l'assemblage	x	x
	Nombre de contigs >X (nt)		Nombre de séquence supérieur à une taille x	x	x
	NG50 (ou 25,75)	x	Taille du contig de longueur L pour laquelle les contigs de taille supérieur ou égale à L couvrant au moins 50% de la référence		x
	NA50 (ou 25,75)	x	Taille du contig de longueur L pour laquelle les contigs de tailles supérieure ou égale à L couvre au moins 50% de l'assemblage à partir des blocs alignés		x
	Somme des longueurs de contigs >X (nt)		Somme des longueurs de contigs avec une taille supérieur à X	x	
Assemblage incohérents et variations structurales	Nombre d'assemblage incohérent	x	Nombre de positions dans l'assemblage où la séquence flanquante gauche s'aligne à plus de 1 kpb de la séquence flanquante droite sur la référence (relocalisation) ou se chevauchent sur plus de 1 kpb (relocalisation) ou les séquences flanquantes s'alignent sur différents brins (inversion) ou différents chromosomes (translocation)		x
	Nombre de contigs contenant des erreurs d'assemblage	x	Nombre de contigs contenant des assemblages incohérents		x
	Nombre de contigs non alignés	x	Nombre de contigs ne s'alignant pas sur la référence		x
	Nombre de contigs mappés de manière ambigüs	x	Nombre de contigs pouvant s'aligner à différentes localisation sur l'assemblage		x
Représentation du génome et des éléments fonctionnels	Fraction du génome	x	Nombre de bases total divisé par la taille de la référence		x
	Ratio de duplication	x	Nombre de bases alignés (nombre total moins les bases non alignés) divisé par la taille de la référence		x
	Nombre de mismatch pour 100kb	x	Nombre moyen de mismatch pour 100 000 bases alignées		x
	Nombre d'indels pour 100kb	x	Nombre moyen d'insertions ou délétions d'un nucléotide pour 100 000 bases alignées		x
	Nombre d'opérons	x	Nombre d'opérons similaire à des gènes basé sur l'annotation du génome		x
	% GC		Pourcentage de bases G et C	x	x
	Nombre de gènes	x	Nombre de gènes dans l'assemblage (complet ou partiel) basé sur l'annotation du génome	x	x
	Nombre de gènes prédits			x	x
	Nombre de gènes constitutifs manquants			x	
	Nombre moyen d'orthologues par gènes constitutifs			x	
	% de gènes constitutifs détecté ayant plus d'un orthologue			x	
	Score BUSCO		Évaluation quantitative en terme de contenu génique attendu	x	
Analyse des k-mers	Complétude basée sur les k-mers	x	Pourcentage de k-mer unique provenant de la référence retrouvé dans l'assemblage		x
	Nombre de k-mers mal raccordés	x	Nombre total de mauvais raccordements basés sur k-mer dans l'assemblage		x
	Longueur correcte basée sur les k-mer (%)	x	Pourcentage de la longueur totale de tous les contigs considéré comme correct étant donné l'analyse des k-mer unique		x
	Longueur basée sur les k-mers mal raccordés (%)	x	Pourcentage de la longueur totale de tous les contigs contenant au moins un k-mer mal raccordé		x
	Longueur totale des contigs sans marqueurs basé sur les k-mers(%)	x	Pourcentage de la longueur totale de tous les contigs sans marqueurs de k-mer		x

TABLE 5 – Tableau récapitulatif des métriques présentes dans BUSCO et/ou QUAST