Assemblage métagénomique d'écosystèmes complexes avec différentes technologies de séquençage de 3ème génération

Nicolas Maurice

▶ **To cite this version:**

Nicolas Maurice. Assemblage métagénomique d'écosystèmes complexes avec différentes technologies de séquençage de 3ème génération. Bio-informatique [q-bio.QM]. 2023. hal-04142837

HAL Id: hal-04142837
https://inria.hal.science/hal-04142837

Submitted on 27 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Rapport de Stage de Master

Assemblage métagénomique d'écosystèmes complexes avec différentes technologies de séquençage de 3ème génération

Auteur :

Nicolas MAURICE

Master Bioinformatique, Biologie Computationnelle

Université de Bordeaux

Encadrantes :

Claire LEMAITRE, Chargée de recherche à Genscale, INRIA

Clémence FRIOUX, Chargée de recherche à Pleiade, INRIA

03/01/2023 - 30/06/2023

Table des matières

1 Remerciements **4**

2 Introduction **5**

3 Etat de l'art **6**
3.1 Communautés microbiennes 6
3.1.1 Diversité des écosystèmes microbiens 6
3.1.2 Comment capturer cette diversité ? 6
3.2 Assemblage métagénomique 7
3.2.1 Assembleurs métagénomiques 7
3.2.2 Évaluation d'assemblages métagénomique 8

4 Matériel et Méthodes **9**
4.1 Jeux de données 9
4.1.1 Bmock12 9
4.1.2 Zymo D6331 9
4.1.3 Bio-Collective 139445 10
4.1.4 Champ de concombre 10
4.2 Pipeline 10
4.3 Assemblage 11
4.4 Evaluation 11
4.4.1 Avec références 11
4.4.2 Sans références 12
4.4.3 Avec binning 13

5 Résultats **14**
5.1 Pipeline 14
5.1.1 Présentation 14
5.1.2 Guide d'utilisation 15
5.1.3 Guide de modification 15
5.2 Comparaison des assembleurs 15
5.2.1 Lectures longues à fort taux d'erreurs 15
5.2.2 Lectures longues à haute fidélité 19
5.3 Effet du sur-séquençage 21
5.4 Comparaison des métriques 23
5.5 Assemblage d'écosystèmes complexes 24

6 Discussion **27**

7 Annexes **32**

Table des figures

5.1 Graphe de lien des règles de la pipeline 14
5.2 Comparaison des résultats d'assemblage selon la technologie de séquençage et l'assembleur, sur la communauté Bmock12 18
5.3 Comparaison des résultats d'assemblage selon l'assembleur, sur la communauté ZymoD6331. 20
5.4 Comparaison de l'effet de stratégies de surséquencage sur la qualité de l'assemblage de la communauté Bmock12. 22
5.5 Recrutement des lectures dans les contigs selon l'assembleur et l'écosystème étudié. 25
5.6 Qualité du binning selon l'assembleur et l'écosystème étudié 26

7.1 Annexe : Comparaison des résultats d'assemblage selon l'assembleur, sur la communauté ZymoD6331. 33

Liste des tableaux

4.1 Tableau récapitulatif des différents jeux de données. 10
4.2 Critères de complétude, contamination et contiguïté des bins par qualité 13

5.1 Tableau de comparaison des métriques par références, par binning, ou sans références 23

Chapitre 1

Remerciements

Un grand merci à Claire Lemaitre et Clémence Frioux, mes encadrantes pour ce stage, pour leurs conseils avisés et les multiples relectures de ce rapport. Merci en particulier pour leur offre de thèse, en espérant que ces trois prochaines années se passent encore mieux que ces six mois de stage.

Merci au reste de l'équipe Genscale, et plus largement aux membres des équipes Dyliss et Genouest avec qui nous partageons nos bureaux et nos pauses. Merci en particulier à Riccardo Vicedomini et Gaëtan Benoit, les experts locaux en assemblage métagénomique, ainsi qu'à Victor Epain pour avoir organisé les soutenances blanches des stagiaires, et Marie Le Roïc, pour son accompagnement administratif.

Merci aux équipes gérant les serveurs de calcul Genouest et Plafrim, sans lesquels ma pipeline tournerait encore.

Un grand merci à tous ceux qui m'ont préparé pour ce stage et/ou soutenu durant sa durée, ma mère, mes soeurs, mes frères, mes amis et professeurs.

Pour finir, je tiens à remercier Rennes et à sa population pour son accueil chaleureux, en particulier Nicolas Jia Kenneth.

Chapitre 2

Introduction

Au cours de ces six dernier mois, j'ai effectué un stage au centre INRIA de l'Université de Rennes dans l'équipe Genscale, sous la supervision de Claire Lemaitre et Clémence Frioux, portant sur l'assemblage métagénomique d'écosystèmes complexes avec différentes technologies de séquençage de 3ème génération.

L'étude des communautés complexes est un problème ouvert, et une partie des approches utilisées pour les étudier passe par l'analyse du génome des espèces composant ces communautés. L'une des approches les plus prometteuses pour l'obtention de ces génomes est l'assemblage métagénomique, technique visant à séquencer l'ADN contenu dans un échantillon environnemental, puis à essayer de reconstruire et séparer les génomes de chacune des unités taxonomiques présentes dans cet échantillon. Il existe plusieurs études évaluant ce genre d'assemblages, mais elles se concentrent généralement sur des écosystèmes simples et sur des assembleurs dédiés aux lectures courtes ou au lectures longues basse fidélité.

Ainsi, j'ai cherché à évaluer la performance d'assembleurs dédiés aux lectures longues haute fidélité, en cherchant pour cela à comparer des métriques d'évaluation pour l'évaluation métagénomique sans référence et à explorer l'impact de la profondeur de séquençage et de la présence de souches sur la qualité de l'assemblage. L'objectif était de fournir des recommandations méthodologiques pour séquencer et assembler des écosystèmes complexes.

J'ai pour cela développé une pipeline d'assemblage et d'évaluation métagénomique, modifiable, réutilisable et disponible à cette adresse : `https://gitlab.inria.fr/stage_nmaurice/metagenomic_benchmark`. Je vais, dans ce rapport, introduire l'état de l'art sur l'assemblage métagénomique et son évaluation, puis vais décrire les méthodes que j'ai utilisé pour mon propre benchmark, puis vais présenter et discuter les résultats obtenus.

Chapitre 3

Etat de l'art

3.1 Communautés microbiennes

3.1.1 Diversité des écosystèmes microbiens

Les végétaux sont intimement liés à leurs microbiotes, communautés microbiennes vivant à l'intérieur ou à la surface de la plante. Ces micro-organismes sont impliqués dans de nombreux processus, comme l'acquisition de nutriments, la résistance à des stress biotiques ou abiotiques, ou encore le parasitisme de la plante hôte[1]. Ainsi, l'étude de ces microbiotes végétaux permettrait de mieux comprendre ces processus et leur régulation, pour pouvoir mieux surveiller l'impact de perturbations environnementales sur un écosystème, mais aussi pour potentiellement optimiser la croissance de plantes cultivées et limiter leur susceptibilité aux pathogènes et leur dépendances aux engrais et pesticides[2].

Ces microbiotes sont riches en diversité, et ce à chaque échelle taxonomique, allant du règne à la souche. Ils contiennent des bactéries, champignons et virus, mais aussi des protistes, des nématodes, des algues et quelques archées[3], chacun de ces clades présentant à son tour une forte diversité d'espèces et de souches. Les facteurs promouvant cette diversité sont eux même multiples. Il y a d'une part la proximité avec le sol, un des écosystèmes les plus riches en diversité microbienne[4], mais aussi la diversité des hôtes ayant co-évolué avec leur communauté microbienne[5], et la régulant selon les conditions environnementales, plus les différentes parties de la plante pouvant être colonisées[3].

Étant donnés ces enjeux, il devient important de capturer cette diversité génétique.

3.1.2 Comment capturer cette diversité ?

La méthode standard d'étude microbienne passe par l'isolation et la culture de l'organisme. En revanche, tous les micro-organismes ne sont pas cultivables en conditions de laboratoire[6], et la diversité de certaines communautés rend irréaliste l'isolation individuelle de chacun de ses membres cultivables.

Une autre approche courante est le séquençage par amplicon, généralement le gène de l'ARN ribosomique 16s ou 18s, mais cette approche ne permet que de capturer la diversité d'un règne à la fois, selon les amorces sélectionnées. De plus, elle ne permet pas de capturer la diversité d'une espèce à l'échelle de la souche, leur génome n'étant pas assez divergent pour pouvoir être détecté avec seulement un gène, et ce malgré des différences phénotypiques. Une autre faiblesse de cette approche est qu'elle n'indique que la diversité taxonomique des espèces, et ne permet en aucun cas de reconstituer des génomes ou bien même des gènes[7].

Une troisième approche, que j'explorerai dans la suite de ce rapport, permet en revanche à la fois d'éviter l'isolation de chacun des membres de la communauté et de reconstituer leurs gènes et leur génomes : la métagénomique plein-génome, ou *whole genome shotgun metagenomics*. Cette technique consiste à séquencer aléatoirement des fragments d'ADN, puis à les assembler en contigs, fragments du génome de taille plus importante.

3.2 Assemblage métagénomique

En assemblage classique, on séquence un individu, ou une population clonale, pour obtenir des lectures, ou *reads*, fragments de séquence nucléique. On assemble ensuite ces lectures en contigs, plus long fragments de séquence nucléique.

En assemblage métagénomique, on séquence un échantillon issu d'un écosystème plutôt qu'un individu, puis l'on cherche à rassembler les contigs en bins, clusters de lectures appartenant à une même espèce ou une même souche, avec idéalement un bin correspondant à un génome.

L'assemblage métagénomique présente de nombreuses difficultés que l'on ne retrouve pas en assemblage classique. En effet, lorsque l'on retrouve des séquences faiblement représentées, il n'est pas aisé de déterminer s'il s'agit d'une erreur de séquençage ou bien de la réalité biologique d'une espèce faiblement abondante. Les régions conservées entre plusieurs espèces posent aussi un problème, non seulement car il n'est pas aisé de déterminer avec quel organisme les placer, mais car il est possible de réaliser des assemblages chimériques, mélangeant des séquences issues de plusieurs organismes dans un même contig ou bin.

Enfin, les souches amplifient ces problèmes, de par leur proximité taxonomique, et donc génétique. En effet, plus des unités taxonomiques sont proches, plus il est aisé de confondre les erreurs de séquençage de l'une avec la réalité terrain de l'autre. De même, une grande partie du génome est conservée entre souches, ce qui rend les assemblages chimériques bien plus fréquents.

3.2.1 Assembleurs métagénomiques

Il existe une grande variété de technologies de séquençage et d'assembleurs métagénomiques dédiés à l'assemblage de chacun de ces types de données.

Pour les lectures courtes de type Illumina, la plupart des assembleurs se basent sur le principe du graphe de Bruijn. Ils se basent sur la découpe des lectures en kmers, ou mots de taille k, puis de la création d'un graphe de ces kmers. Les assembleurs métagénomiques se basant sur des lectures courtes, comme MEGAHIT, MetaSPADES ou IDBA-UD, produisent des assemblages généralement très fragmentés[8], et ont étés fortement étudiés[9].

Ainsi, nous nous concentrerons sur les assembleurs métagénomiques utilisant des technologies de séquençage de 3ème génération, ou lectures longues. La plupart de ces lectures longues, comme les lectures PacBio CLR ou ONT, ont un fort taux d'erreurs. Les lectures PacBio CLR, ou *Continuous long reads*, peuvent mesurent près d'une dizaine de kilo-bases de long, et présentent un taux d'erreur proche des 10%. Les lectures ONT sont quant à elle, plus longues, près d'une vingtaine de kilo-bases, et bien que restant riches en erreurs, s'améliorent continuellement, passant de 10% à, pour les versions les plus récentes, un taux d'erreurs de 3%.

Une première approche est l'assemblage hybride, cherchant à combiner les lectures longues et les lectures courtes dans un même assemblage. Il n'existe que peu d'assembleurs métagénomiques dédiés à l'assemblage hybride, OPERA-MS[10] en étant un. Il est aussi basé sur les graphes de Bruijn, mais est capable de prendre avantage des lectures longues pour résoudre des ambiguïtés structurales que les lectures courtes n'arrivent pas à résoudre, l'objectif étant de combiner la précision des lectures courtes à la contiguïté des lectures longues.

Il est aussi possible de réaliser un assemblage métagénomique en utilisant uniquement des lectures longues, comme le font Miniasm[11], Metaflye[12], Canu[13] ou encore Lathe[14]. Ce type d'assembleurs se base généralement sur des graphes de recouvrement, qui ont l'avantage de ne pas découper les lectures en sous-mot, et prennent ainsi avantage de la longueur des lectures longues.

Enfin, il existe une nouvelle technologie de séquençage de 3ème génération, les lectures longues PacBio Hi-Fi. Elles ont l'avantage d'être à la fois longues (près d'une dizaine de kilo-bases de long, comme les lectures PacBio CLR) et précises (près d'1% d'erreur, comme pour les lectures Illumina). Metaflye[12] et Hifiasm-meta[15] sont deux assembleurs de lectures longues haute fidélité se basant sur des graphes de recouvrement, et prenant avantage du faible taux d'erreur des lectures haute fidélité. MetaMDBG[16] est aussi un assembleur de lectures longues haute fidélité, mais se base sur l'assemblage de kmers de *minimizers* dans un graphe de Bruijn.

3.2.2 Évaluation d'assemblages métagénomique

Il existe plusieurs études évaluant des assembleurs métagénomiques sur une grande variété de données[9][17][8]. En revanche, aucune n'évalue l'assemblage de lectures haute-fidélité sur des environnements de complexité variable.

L'étude de Wang et Al. de 2020[9] évaluant uniquement des assemblages lectures courtes et mélangeant données réelles et simulées, arrive à la conclusion que, dans un écosystème complexe, plus un génome est abondant et est différent des autres génomes du métagénome, mieux il est reconstitué.

L'étude de Meyer et Al. de 2022[17] évalue des assemblages lectures courtes, longues et hybrides, sur des données simulant divers écosystèmes plus ou moins complexes. Ils observent aussi que les génomes abondants et différent des autres génomes du métagénome sont mieux assemblés, et que les écosystèmes complexes produisent généralement des assemblages de moins bonne qualité. Seul Flye et metaFlye sont évalués pour les lectures longues, et, pour les assemblages hybrides, le seul assembleur généraliste évalué est OPERA-MS. Ils remarquent également que les assemblages hybrides sont supérieurs aux assemblage lectures courtes, à la fois en terme de complétude et de contiguïté.

L'étude de Z. Zhang et Al de 2023[8] évalue des assemblages lectures courtes, liées, longues, et hybrides, sur des données simulées, réelles et issues de simili-communautés. Ils concluent aussi que les espèces les moins abondantes d'un métagénome sont très difficiles à reconstituer, en particulier pour les lectures courtes. Pour les lectures longues, plusieurs assembleurs sont évalués, mais Canu, Lathe et metaFlye sortent du lot, tandis que pour les assemblages hybrides, ils observent que les assembleurs basés sur les graphes de recouvrement (metaFlye-subassemblies, DBG2OLC) ont une meilleure contiguité que ceux basés sur les graphes de de Bruijn (OPERA-LG, OPERA-MS). Ils n'évaluent aucun assembleur dédié aux lectures longues à haute fidélité, et n'évaluent leurs trois meilleurs assembleurs lectures-longues que sur un seul jeu de données lectures longues haute fidélité, et concluent que l'utilisation de lectures longues haute fidélité augmente la qualité de l'assemblage. Cette troisième étude ayant été publiée pendant le stage et après que le choix des assembleurs ait été effectué, n'a donc pas influencé mon choix d'assembleurs à évaluer.

Ainsi, il existe à la fois un manque d'études évaluant des assembleurs dédiés à l'assemblage Hi-Fi, et un manque de d'évaluation de l'impact de la complexité d'une communauté sur la qualité des assemblages métagénomiques.

Au cours de ce stage, j'ai donc dévellopé une pipeline d'assemblage et d'évaluation métagénomique dédiée aux lectures longues, et l'ai utilisé pour nottament assembler diverses communautés présentant divers niveaux de complexité avec trois assembleurs Hi-Fi : MetaMDBG[16], MetaflyeMetaflye[12] et Hifiasm-meta[15]. Pour pouvoir comparer ces assembleurs sur des communautés complexes, j'ai aussi cherché à évaluer diverses métriques sans-référence, et ai exploré l'impact de la profondeur de séquençage et de la présence de souche sur la qualité de l'assemblage.

Chapitre 4

Matériel et Méthodes

4.1 Jeux de données

J'ai utilisé cinq communautés microbiennes lors du développement et de l'évaluation de ma pipeline, sélectionnées pour me permettre d'étudier une large gamme de technologies de séquençage et donc d'assembleurs, mais aussi d'étudier divers environnements plus ou moins complexes. Un tableau récapitulatif regroupant le nom des jeux de données, la communauté microbienne desquels ils sont issus, des technologies de séquençage utilisées, de la taille du jeux de donnée en nombre de bases, de la longueur médiane des lectures, et de leur qualité médiane, est disponible à la fin de cette section (table 4.1).

4.1.1 Bmock12

Cette communauté[18] est une simili-communauté, ou *mock community*, formée en mélangeant des ADNs connus, en proportion contrôlée, après extraction mais avant séquençage. Cela permet non seulement de simplifier la complexité d'une communauté réelle, mais aussi de disposer d'une référence avec laquelle comparer l'assemblage. Par rapport aux simulations de lectures in-silico, les simili-communautés permettent aussi de reproduire plus fidèlement les erreurs et biais de séquençage.

Elle comprend 12 espèces appartenant à 9 genres bactériens, présentant une forte variabilité dans leurs abondances, allant jusqu'à un facteur 17 entre l'espèce ayant la plus grande molarité d'ADN et celle ayant la plus petite. Il est important de noter que l'une des espèces, *Micromonospora coxensis* DSM 45161 n'est pas retrouvée dans les lectures, vraisemblablement suite à une erreur de pipetage lors de la création de cette communauté [18].

Des séquençages Illumina (TruSeq SBS v.4), PacBio CLR (RSII v. C4) et ONT (R9.4.1), sont disponibles dans la banque de données SRA, et leur numéros d'accesions sont respectivement SRX4901583, SRX4901585 et SRX4901586[18]. Afin de pouvoir mieux comparer les technologies PacBio et ONT, j'ai aléatoirement sous-échantillonné ce dernier à l'aide de rasusa[19] pour que les deux séquençages aient la même profondeur.

4.1.2 Zymo D6331

Cette communauté[20] est aussi une simili-communauté, cette fois formée en mélangeant les micro-organismes de séquence connue en proportion contrôlée avant d'extraire l'ADN. En plus des avantages précédemment cités, cette méthode permet de reproduire les biais d'extraction d'ADN que l'on retrouve dans une communauté réelle.

Elle comprend deux espèces de levures, une espèce d'archée, et quatorze espèces de bactéries, dont *E. coli* avec cinq souches. Leurs abondances sont encore plus variables que pour Bmock12, allant jusqu'à un facteur 2×10^6 entre l'espèce ayant la plus grande molarité d'ADN et celle ayant la plus petite.

Un séquençage PacBio Hi-Fi (Sequel II, SMRT Cell 8M) est disponible au numéro d'accession SRX9569057[20] dans la banque de données SRA.

4.1.3 Bio-Collective 139445

Cette communauté[21] est issue du mélange de quatre échantillons fécaux d'humains adultes suivant un régime végétalien. Les communautés réelles correspondent exactement à la réalité terrain et sont donc indispensables afin de valider la pipeline, mais sont généralement plus complexes que des simili-communautés et les espèces présentes, et donc leur génomes et leurs abondances relatives, sont inconnues, rendant l'évaluation plus complexe.

Un séquençage PacBio Hi-Fi (Sequel II, SMRT Cell 8M) est disponible au numéro d'accession SRX11580195[21] dans la banque de données SRA.

4.1.4 Champ de concombre

Ces deux communautés sont issues d'un échantillon de sol chacun, prélevées dans un champ de concombres. L'un des échantillons a été prélevé au contact de la racine, dans la rhizosphère, tandis que le second est prélevé plus loin dans le sol du champs, dans l'espace inter-rang. Les deux jeux de données on été séquençés à l'aide de la technologie PacBio Hi-Fi (Sequel II, SMRT Cell 8M), mais ne sont pas encore publics.

Jeu de données	Communauté	Séquençeur	Bases (Gb)	Taille Médiane	Qualité Médiane
Bmock CLR	Bmock12	PacBio RSII v. C4	1.5	5.9 Kb	8.5
Bmock ONT full		ONT R9.4.1	3.7	17.9 Kb	9.5
Bmock ONT sub			1.5	18.0 Kb	9.5
Bmock Illumina		Illumina TruSeq SBS v.4	64.4	151 b	30.5
Zymo HiFi	Zymo D6331	PacBio Hi-Fi Sequel II, SMRT Cell 8M	18.0	8.1 Kb	39.6
Vegan gut	Bio-Collective 139445		15.2	8.4 Kb	39.6
rhizosphère	Concombre rhizosphère		19.5	5.8 Kb	42.5
inter-rang	Concombre inter-rang		13.4	6.0 Kb	43.8

TABLE 4.1 – **Tableau récapitulatif des différents jeux de données.** Le nombre de bases total du jeu de données (en gigabases), la taille médiane (en kilobases (Kb) ou en bases (b) des lectures et la qualité médiane (en score Phred) ont étés calculées par l'outil NanoPlot [22].

4.2 Pipeline

J'ai utilisé Snakemake [23] pour réaliser une pipeline permettant de traiter mes données de séquençage. Snakemake est basé sur des règles liant des fichiers d'entrée, des fichiers de sortie et des commandes bash. Ces liens forment un graphe permettant à Snakemake de savoir quels fichiers d'entrée sont requis et quelle séquence de commandes sont à exécuter pour obtenir les fichiers désirés. Chaque règle est indépendante, et isole une partie du traitement, facilitant le parallélisme et la modularisation de l'ensemble. J'utilise généralement ces commandes bash pour appeler un programme avec des paramètres générés selon les noms des fichiers d'entrée. Cela permet de facilement modifier une étape de la pipeline sans avoir à toucher au Snakefile, fichier pricipal de la pipeline.

J'utilise aussi un fichier de configuration, nommé config.yaml, afin de pouvoir sélectionner les traitements et jeux de données sans avoir à modifier le code.

Slurm[24] est un gestionnaire de charge de travail, qui permet de gérer les ressources d'un serveur de calcul par un système de tâches, exécutées par ordre de priorité. J'intègre les directives

slurm dans la pipeline, ce qui permet de paralléliser les traitements de la pipeline et d'accéder à des ressources de calcul (en terme de CPUs et de RAMs, par exemple) auquelles je n'ai pas accès sur un ordinateur individuel. J'ai développé et utilisé cette pipeline sur deux serveurs de calcul, Genouest[25] et plafrim[26].

Conda[27] est un gestionnaire de packages et d'environnements, gérant les dépendances d'une grande variété de programmes. J'intègre des directives pour générer et utiliser un environnement conda pour chaque règle nécessitant l'utilisation de programmes externes, ce qui permet de déployer la pipeline dans à peu près n'importe quel environnement, sans que les personnes utilisatrices ait à elle-même installer la moindre dépendance.

Cette pipeline est téléchargeable sur gitlab au lien suivant : `https://gitlab.inria.fr/stage_nmaurice/metagenomic_benchmark/-/tree/main/workflow/scripts`

4.3 Assemblage

Des cinq assembleurs métagénomiques que j'ai évalués, deux sont dédiés aux assemblages métagénomiques à partir de lectures longues à fort taux d'erreurs de type Nanopore ou PacBio CLR, correspondant au jeux de données Bmock12 : Miniasm et Opera-MS (ce dernier requérant à la fois des lectures longues et des lectures courtes de type Illumina). Deux autres assembleurs, Hifiasm-Meta et Meta-MDBG, assemblent plutôt des lectures longues haute-fidélité PacBio, correspondant aux jeux de données Zymo, Vegan gut, rhizosphère et inter-rang. Le dernier assembleur, Metaflye, peut être capable, selon les paramètres utilisés, d'assembler des lectures longues à fort ou faible taux d'erreurs.

Afin d'accélérer l'assemblage j'utilise l'option -t (ou –num-processors pour Opera-MS) afin de répartir au moins une partie des calculs pour chaque assemblage sur plusieurs CPUs. Pour faciliter la modularisation de ma pipeline, chaque assembleur prend en entrée un fichier .fastq contenant des lectures (à part Opera MS, prenant trois fichiers .fastq en entrée), et génère un fichier nommé assembly.fasta, contentant des contigs, en sortie. Si ce n'est pas le cas, je rajoutes une étape afin de renommer le fichier, de le décompresser ou éventuellement de convertir son format. Chaque assembleur a été utilisé sur chacun des jeux de données qu'il pouvait traiter.

J'utilise Flye 2.9, en utilisant l'option –meta pour exécuter MetaFlye[12], une version plus adaptée à l'assemblage métagénomique que Flye classique. Afin d'adapter l'assemblage au lectures que l'on cherche à assembler, j'utilise l'option –pacbio-raw pour les données de type PacBio CLR, –nano-raw pour ONT, ou –pacbio-hifi pour PacBio Hi-Fi.

Miniasm 0.3[11] n'aligne pas lui-même les lectures et requiert en entrée un alignement au format .paf. J'utilise donc minimap2 2.24[28], avec l'option ava-pb pour les données de type PacBio CLR ou ava-ont pour ONT pour adapter l'alignement à la technologie de séquençage. J'utilise ensuite Miniasm, qui génère un fichier .gfa, que je convertit en .fasta à l'aide de la commande `awk '/^S/{print ">"$2"\n"$3}' <fichier.gfa>`. Pour polir cet assemblage, j'utilise Flye 2.9 avec l'option –polish-target (et la même adaptation au jeu de données que pour MetaFlye), qui prend en entrée l'assemblage à polir et les reads, pour produire un nouvel assemblage.

J'utilise Opera-MS 0.9.0 [10], qui prend trois fichier .fastq en entrée, un contenant des lectures longues, et deux contenant des lectures courtes paired-ends. L'objectif étant d'évaluer des assembleurs de-novo, j'utilise l'option –no-ref-clustering pour éviter d'utiliser des références lors de l'assemblage.

J'utilise les paramètres par défault pour les deux assembleurs dédié à l'assemblage hi-fi, Hifiasm meta[15] version hamtv0.3.1 et la version du commit 8b5419f de Meta-mdbg[16].

4.4 Evaluation

4.4.1 Avec références

j'ai choisi d'utiliser 5 métriques avec références, qui se basent sur l'alignement et la comparaison entre le génome de référence et les contigs. Ces métriques sont : la fraction du génome reconstituée, le nombre de mismatches pour 100k paires de bases, le nombre de misassemblies, le ratio de

duplication et le NGA50. La fraction du génome reconstituée se calcule en divisant le nombre de bases alignées du génome de référence par la taille de ce génome, et indique la complétude de l'assemblage, ou à quel point l'assemblage a pu reconstruire une grande partie du génome. Plus elle est élevée, plus l'assemblage est complet, et est donc de bonne qualité.

La NGA50 correspond à la taille du plus petit contig aligné au génome de référence nécessaire pour que la somme des contigs alignés d'au moins cette taille soit égale à au moins la moitié (50%) de la taille du génome. Elle indique le niveau de contiguïté de l'assemblage, et plus la NGA50 est élevée, plus on a de longues sections de génome assemblée en un seul morceau, ce qui indique un assemblage de bonne qualité. Je ne la calcule pas automatiquement, mais certains résultats sont communiqués en terme de NGA50 normalisée, c'est à dire qu'elle est divisée par la taille du génome. Cela permet de comparer l'assemblage d'espèces de différentes tailles, et permet de rapidement visualiser qu'une espèce pour laquelle la NGA50 normalisée est égale à 1 a été reconstruit en un seul contig.

Le nombre de mismatches pour 100kb correspond au nombre de différences d'un nucléotide entre l'assemblage et l'alignement, divisé par la taille totale de l'alignement et multiplié par 100000. Le nombre de misassemblies correspond au nombre de différences structurales entre l'assemblage et le génome de référence. Ces deux métriques mesurent la précision de l'assemblage, est plus elles sont faibles, plus l'assemblage colle à la réalité.

Le ratio de duplications correspond au nombre de bases alignées de l'assemblage divisé par le nombre de bases alignées du génome de référence. Une valeur supérieure à 1 indique qu'une même partie du génome correspond à plusieurs contigs, et que le génome a donc été au moins partiellement dupliqué lors de l'assemblage. Plus cette valeur est proche de 1, meilleur est la qualité de l'assemblage

J'utilise metaQuast[29] 5.2 pour évaluer la qualité d'un assemblage si des génomes de référence sont disponibles. Cet outil aligne les contigs avec les génomes de référence, puis évalue séparément l'assemblage de chaque génome, en produisant notament les métriques précédemment citées, dans une série de fichiers tabulés, un par métrique évaluée.

Un des objectifs de la pipeline étant de faciliter la prise de décision et donc de fournir un résultat facile à interpréter, j'ai choisi de communiquer les résultats sous formes de moyennes pour chaque métrique. En revanche, un autre objectif est d'évaluer la capacité des assembleurs à reconstituer des espèces plus ou moins abondantes, et j'ai donc décidé, pour chaque métrique, de fournir trois moyennes plutôt qu'une : une pour les espèces faiblement abondantes (profondeur de séquençage inférieure à 10X), une pour les espèces moyennement abondantes (profondeur de séquençage comprise entre 10X et 50X), et une pour les espèces fortement abondantes (profondeur de séquençage supérieure à 50X). Ces seuils ont été choisis empiriquement, les espèces fortement abondantes étant généralement bien assemblées, et les espèces faiblement abondantes étant généralement mal assemblées.

Pour calculer cette profondeur de séquençage, j'utilise minimap2 [28] 2.24, en utilisant l'option map-pb pour PacBio CLR, map-ont pour ONT, et map-hifi pour PacBio Hi-Fi, pour aligner les lectures aux génomes de référence. J'utilise ensuite samtools [30] 1.17, une première fois avec l'option sort afin de trier les alignements par coordonnées, une étape essentielle afin de pouvoir appliquer dans un second temps samtools coverage, qui calcule la profondeur de séquençage de chaque génome de référence, et écrit ses résultats dans un fichier .tsv.

J'ai crée un programme python regroupant ensuite ces informations en utilisant la bibliothèque pandas[31][32] 1.5. J'importe les deux fichiers .csv, puis calcule les moyennes pour les trois groupes d'abondance, et écris dans un fichier à la fois les moyennes et, pour si la personne l'utilisant désire plus de détails, les informations individuelles pour chaque génome de référence.

4.4.2 Sans références

Si aucun génome de référence n'est disponible, il est quand même possible d'évaluer un assemblage, avec des métriques telles que la proportion de lectures ou de bases alignées, ou bien la longueur d'assemblage, le nombre de contigs, le N50 et le L50.

La proportion de lectures alignées se calcule en divisant le nombre de lectures alignées à l'assemblage par le nombre de lectures totales. Cette métrique indique à quel point l'assembleur a su utiliser toutes les lectures, et plus elle est proche de 100%, plus l'assemblage est complet.

De même, la proportion des longueurs d'alignement se calcule en divisant le nombre de base de

chaque alignement lectures-assemblages par la longueur totale des lectures. Une valeur élevée peut indiqué que toutes les lectures ont été utilisée en leur entièreté, mais peut aussi indiquer, surtout au dessus de 100%, que plusieurs contigs sont alignés à la même lecture, et donc une duplication de l'assemblage.

La longueur d'assemblage est la somme des longueur des contigs. Une valeur élevée peut indiquer un bon assemblage, mais, de même, peut aussi indiquer une duplication de l'assemblage.

Un nombre de contigs élevé peut indiquer que de nombreux génomes ont été reconstruit, mais peut aussi indiquer que les génomes qui ont été reconstruits sont fortement fragmentés.

Le N50 correspond à la taille du plus petit contig nécessaire à ce que la somme de tous les contigs d'au moins cette taille soit supérieure à la moitié de la taille de l'assemblage, et le L50 correspond au nombre de contigs nécessaire à cette somme. Tout comme le nombre de contigs, un L50 élevé et un N50 bas peuvent à la fois indiquer une abondance de génomes reconstruits ou un fort fractionnement de ceux-ci. Néanmoins, un N50 de l'ordre de grandeur de la taille d'un génome peut indiquer qu'au moins la moitié de l'assemblage est composé de génomes reconstruits en un seul contig chacun.

Les proportions de lectures et de longueurs alignées se basent sur un alignement, et j'utilise donc minimap2 [28] 2.24 pour aligner les lectures sur l'assemblage, avec l'option map-pb, map-ont ou map-hifi, selon le type de lectures utilisées (PacBio CLR, ONT ou PacBio Hi-Fi, respectivement). La sortie est un fichier .paf, un format plus simple à traiter que le .bam ou le .sam.

J'ai développé un programme en c++, prenant en entrée le fichier .paf, ainsi que les lectures et contigs elle-mêmes, afin de calculer les métriques précédemment citées. Il est important de noter que, pour la longueur d'assemblage, le nombre de contigs, le N50 et le L50, je passe par une étape de filtrage afin d'éliminer les contigs de moins de 1000 paires de bases et donc de mettre sur un pied d'égalité les assembleurs ayant ce filtre de ceux ne l'ayant pas.

4.4.3 Avec binning

Un autre moyen d'évaluer un métagénome sans références, est de passer par une étape de binning, puis d'évaluer le résultat de ce binning. Dans mon cas, j'utilise MetaBat2[33] 2.12, qui clusterise les contigs en fonction de leur couverture par les lectures et de leur fréquence tétranucléotidique, afin de former des bins, sensées chacune contenir tout ou au moins une partie du génome d'une espèce ou d'une souche. Le paramètre –percentIdentity, par défaut 97, permet de contrôler à partir de quel pourcentage d'identité une lecture est considérée comme alignée sur un contig, je l'ai baissé à 90, afin de permettre aux assemblages issues de lectures longues, surtout celles à fort taux d'erreur, de quand même être regroupés en bins.

J'évalue ensuite ce binning avec checkM[34] 1.2, un outil qui estime la taxonomie, la complétude et la contamination à l'aide de gènes marqueurs. J'utilise ensuite les résultats de checkM dans un programme python , pour compter le nombre de bins correspondant à divers critères de qualité (table 4.2). Une bin ne peut appartenir qu'à une seule catégorie, et appartient à la catégorie de plus haute qualité à laquelle elle peut prétendre. Un assemblage présentant plus de bins de meilleure qualité est considéré comme un meilleur assemblage.

	Complétude	Contamination	Contigs
Presque complète (NC)	>=99	>=1	1
Haute qualité (HQ)	>=90	>=5	//
Moyenne qualité (MQ)	>=50	>=10	//
Basse qualité (LQ)	>0	>=10	//
Sans qualité (NQ)	//	//	//

Table 4.2 – Critères de complétude, contamination et contiguïté des bins par qualité

Chapitre 5

Résultats

5.1 Pipeline

5.1.1 Présentation

La pipeline que j'ai réalisé est téléchargeable sur gitlab à l'adresse suivante : `https://gitlab.inria.fr/stage_nmaurice/metagenomic_benchmark`.

Le schéma de cette pipeline (fig. 5.1) présente les diverses règles. Une majorité (Metaflye_assembly, miniasm_assembly (suivie de flye_polishing), meta_MDBG_assembly et operams_assembly) sont liées à l'assemblage lui même. assembly_quality_check encapsule metaQUAST, produisant les données brutes traitées par assembly_refrences_based_stats, en utilisant les données de couvertures caluées par references_merger et coverage _calculator, pour produire un rapport de statistiques basées sur des références. assembly_references_free_stats, utilisant un programme que j'ai écrit en c++, complié par la règle complie_cpp, calcule et produit un rapport de statistiques basées sur l'alignement des contigs avec les lectures, ainsi que sur les longueurs des contigs. metabat2 produit un binning, évalué par checkm, dont les informations sont synthétisées dans un rapport statistique par binning_stats. reads_quality_check calcule des statistiques sur les lectures elles-mêmes, indépendamment de l'assemblage. Enfin, all définit les fichiers auquel la pipeline s'attend en sortie, calculés automatiquement en fonction du fichier de configuration.

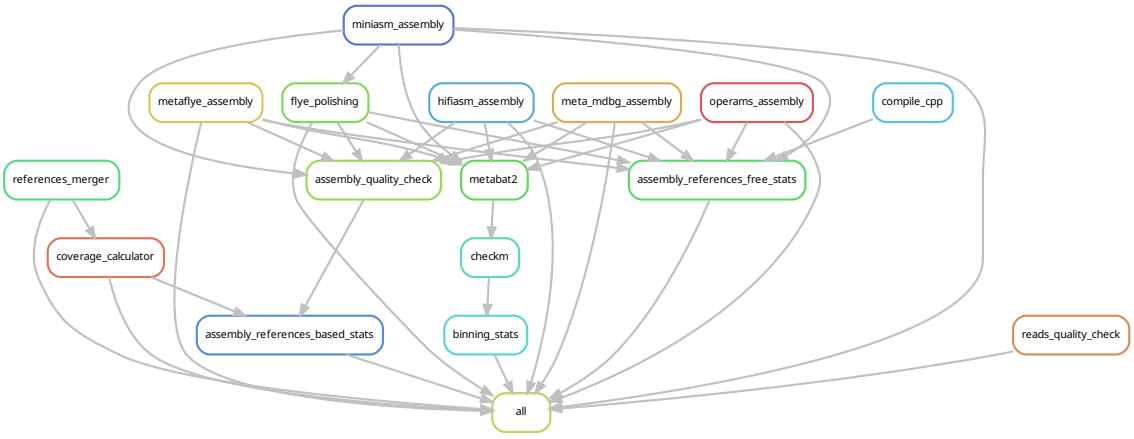

FIGURE 5.1 – **Graphe de lien des règles de la pipeline**, obtenu avec la commande snakemake –rulegraph all. Chaque noeud représente une règle, et les arêtes indique que la sortie d'une règle correspond à l'entrée d'une autre règle.

5.1.2 Guide d'utilisation

Après avoir télécharger cette pipeline sur un cluster de calcul disposant de slurm, vérifiez que mamba est installé. Ensuite, lancez le programme initialise_environnement.sh, dans le dossier config, qui s'occupera d'installer snakemake, opera-MS et metaMDBG, ainsi que de créer des dossiers nécéssaires au bon fonctionnement du programme. La pipleine est maintenant installée.

Avant de l'utiliser, il faut placer vos données d'entrée, des lectures au format fastq, dans le dossier data/input_reads, et leur métadonnées dans le ../data/runs_metadata. Ces métadonnées sont, pour chaque jeu de données, un fichier au nom identique à celui du jeu de données, si ce n'est que l'extention .fastq est remplacée par un .txt. Chaqun de ces fichiers contiens une seule ligne de texte indiquant la technologie de séquençage utilisée : "PacBio RS II" pour les PacBio CLR, "MinION" pour les ONT ou "pacbio-hifi" pour les PacBio Hi-Fi. Il est aussi important d'indiquer le nom du jeu de donnée dans le fichier config.yaml, dans le dossier config. Il suffit ensuite de commenter et décomenter les noms des jeux de données, assembleurs et métriques d'évaluation pour configurer un run.

Pour utiliser des métriques par référence, il faudra aussi mettre les génomes de référence dans le dossier data/input_reference_genomes, au format fasta.

Avant chaque utilisation, activez l'environement conda en appelant le programme activate_environement.sh. Vous pouvez prévisualiser le déroulement de la pipeline avant de la lancer avec snakemake -np all, qui présentera les règles qui vont être lancées, ainsi que les fichiers d'entrée requis et les fichiers de sortie prévus. Si cette prévisualisation vous convient, lancez en job slurm pipeline.sh. Vous pouvez suivre le déroulement de la pipeline avec les commandes slurm habituelles.

En sortie, vous obtiendrez les assemblages eux même dans le dossier data/assemblies, et une série de rapports d'évaluation dans le dossier data/stats_reports, dans des fichiers texte afin de faciliter la consultation sur cluster à distance.

5.1.3 Guide de modification

De par la modularité de ma pipeline, il est assez aisé de la modifier. Pour rajouter un assembleur lectures longues basse ou haute fidélité, ou un assembleur hybride, il suffit de rajouter une nouvelle règle dans la Snakefile, de créer le programme d'encapsulation correspondant et de rajouter le nom de cet assembleur dans le fichier de configuration.

Pour des modifications plus drastiques, telles que de nouvelles métriques ou de nouvelles catégories d'assembleur, le procédé est le même, si ce n'est qu'il faut aussi modifier la règle all, afin d'ajouter un fichier auquel la pipeline s'attend en sortie, et potentiellement de modifier la structure du fichier de configuration (par exemple, rajouter une option pour choisir quel(s) assembleurs lectures courtes utiliser.

5.2 Comparaison des assembleurs

5.2.1 Lectures longues à fort taux d'erreurs

En utilisant la pipeline décrite plus tôt, j'ai assemblé les jeux de données à lectures longues, ONT et PacBio, issus de la simili-communauté Bmock12 à l'aide de deux assembleurs : Metaflye et Miniasm, les assemblages de ce dernier ayant été poli par Flye, puis j'ai évalué ces assemblages avec références. Afin de faciliter la comparaison des technologies de séquençage, j'utilise le jeu de données sous-échantilloné d'ONT, afin que les deux jeux de données aient la même profondeur de séquençage. Il est important de noter que, pour chaque génome de référence, ONT apparaît comme ayant une plus grande couverture, ce qui est potentiellement dû à un meilleur alignement des lectures sur les génomes de références, ou bien à une plus forte contamination dans l'échantillon PacBio.

Lors des analyses, je me référerai à quatre groupes : les espèces fortement abondantes, ayant une profondeur de séquançage superieure à 50X pour ONT (*Muricauda sp. ES050*, *Halomonas sp. HL 4*, *Marinobacter sp. LV10R510 8*, et *Halomonas sp.HL 93*), les espèces moyennement abondantes, dont la couverture est comprise entre 10X et 50X (*Marinobacter sp. LV10MA510 1*,

Thioclava sp. ES032, *Psychrobacter sp. LV10R520 6*, et *Cohaesibacter sp. ES047*), et les espèces faiblement abondantes, ayant une couverture inférieure à 10X (*Micromonospora echinofusca DSM 43913*, *Micromonospora echinaurantiaca DSM 43904*, et *Propionibacteriaceae bacterium ES041*). *Micromonospora coxensis* DSM 45161 ne sera pas considéré dans l'analyse, puisque ayant une abondance de 0 suite à une erreur de manipulation lors de la création de la simili-communauté[18].

Pour les assemblages des lectures ONT par Metaflye, on remarque que les espèces les plus abondantes sont généralement mieux assemblées, que ce soit en terme de complétude (fraction du génome reconstitué), contiguité (NGA50 normalisée) ou précision (nombre de mismatches pour 100kb et nombre de misassemblies), a deux exceptions près, les *Halomonas* (fig. 5.2). On ne remarque en revanche pas de grosse différence au niveau des ratio de duplication, très proches de 1, à l'exception des *Halomonas*.

Pour se concentrer sur cette exception, on observe que les deux autres espèces fortement abondantes (*Muricauda sp. ES050* et *Marinobacter sp. LV10R510 8*) sont extrêmement bien assemblées (en moyenne, 100% du génome est reconstitué, dont 99% en un seul contig, avec seulement 51 mismatches pour 100kb et un ratio de duplication de 1), alors que les deux *Halomonas* sont non seulement moins complètes (96% du génome reconsitué), mais surtout bien plus fragmentées (NGA50 normalisé de 6%), bien moins précises, avec plus de 10 fois plus de mismatches (626 pour 100kb), et partiellement dupliquées (ratio de duplication de 1.07). Ces observations sont cohérentes avec d'autres études indiquant la difficulté à assembler des génomes si des unités taxonomiques génétiquement proches sont présentes dans le métagénome[9][17].

Les espèces moyennement abondantes sont aussi complètes (près de 100% du génome reconstruit), mais plus fragmentées que les espèces fortement abondantes (NGA50 normalisée de 69%) et avec près de 3 fois plus de mismatches (148 pour 100kb). Malgré leur plus faible abondance, elles restent en moyenne mieux reconstituées que les *Halomonas*, indiquant que l'abondance n'est pas le seul facteur influençant la qualité d'un assemblage.

Les espèces faiblement abondantes sont quant à elles bien moins complètes, avec seulement 52% du génome reconstruit en moyenne, et bien moins précises, avec plus de 10 fois le nombre de mismatches que les espèces moyennement abondantes (2773 pour 100kb). Il est compliqué d'évaluer la fragmentation, car la métrique usuelle, le NGA50, requiert qu'au moins 50% du génome ait été reconstitué, et que seul *Propionibacteriaceae bacterium ES041* remplit ce critère (89% du génome reconstitué, NGA50 normalisée de 4%). Néanmoins, on peut raisonnablement supposer que des génomes aussi incomplets sont fortement fragmentés.

Pour l'assemblage de ces lectures par miniasm, on observe les mêmes tendances (fig. 5.2), et l'assemblage des espèces fortement abondantes (*Halomonas* exclues) et moyennement abondantes sont similaires entre les deux assembleurs. Néanmoins, miniasm peine à assembler les *Halomonas*, que ce soit en terme de complétude (91%, contre 96% pour Metaflye) ou duplication (1.32, contre 1.07), mais les deux assembleurs restent équivalents en terme de contiguïté ou de précision. Pour les espèces faiblement abondantes, miniasm échoue complètement l'assemblage, ne réussissant à assembler que 0.56% du génome.

On observe des tendances similaires pour l'assemblage des lectures PacBio (fig. 5.2) (assembleurs équivalents à haute abondance, mais Metaflye est meilleur que miniasm à faible abondance et pour reconstituer des génomes proches), si ce n'est que Metaflye assemble mieux les espèces moyennement abondantes que miniasm. En effet, l'assemblage est bien plus complet (96% pour Metaflye, contre 62% pour miniasm), bien moins fragmenté (NGA50 normalisée de 18%, contre 9% (calculé sur 2 génomes, les deux autres étant trop peu complets pour pouvoir calculer la NGA50)), et plus précis (66, contre 117 mismatches pour 100kb)(1 misassembly en moyenne, contre 16). On peut expliquer ces observations par le fait que les lectures PacBio ont une moins bonne couverture pour chaque génome (probablement du au plus faible taux d'erreur et à la plus grande longueur des lectures ONT), et que miniasm a des difficultés à assembler les génomes peu couverts. Je rappelle que les lectures ONT ont été sous échantillonnées aléatoirement afin que les deux échantillons aient la même couverture, et ainsi permettre une comparaison entre les deux technologies de séquençage.

Si l'on compare les assemblages des lectures ONT et PacBio par Metaflye, on peut observer d'assez grandes différences (fig. 5.2). À haute abondance (*Halomonas* exclues), les assemblages des lectures ONT sont équivalentes à ceux des lectures PacBio en terme de complétude, légèrement

moins précises (51 contre 27 mismatches pour 100kb), mais surtout bien plus contiguës (NGA50 normalisé de 99%, contre 39%). Pour les espèces moyennement abondantes, les assemblages des lectures ONT sont plus complètes (fraction du génome reconstitué de près de 100%, contre 96%), légèrement moins précises (148 contre 116 mismatches pour 100kb) mais là encore bien plus contiguës (NGA50 de 69%, contre 18%) ;

Pour les espèces faiblement abondantes, les assemblages des lectures ONT sont aussi plus complètes (52%, contre 26%), mais bien moins précises (2773 contre 753 mismatches pour 100kb), la contiguïté est difficilement évaluable due à la faible complétude, mais très mauvaise pour les deux assembleurs.

En revanche, les deux types de lectures produisent des assemblages de qualité similaire pour les *Halomonas*, en terme de complétude, contiguïté ou précision.

Ainsi, plus l'abondance augmente, mieux un génome est reconstitué. Ce n'est pas le seul facteur, et des espèces ou souches proches peuvent aussi compliquer l'assemblage. À forte abondance, Metaflye et miniasm produisent généralement des assemblages de qualité similaire, mais seul Metaflye est capable de reconstituer les génomes des espèces les moins abondantes. Les lectures ONT produisent généralement un assemblage plus complet et moins fragmenté, mais les lectures PacBio produisent un assemblage plus précis

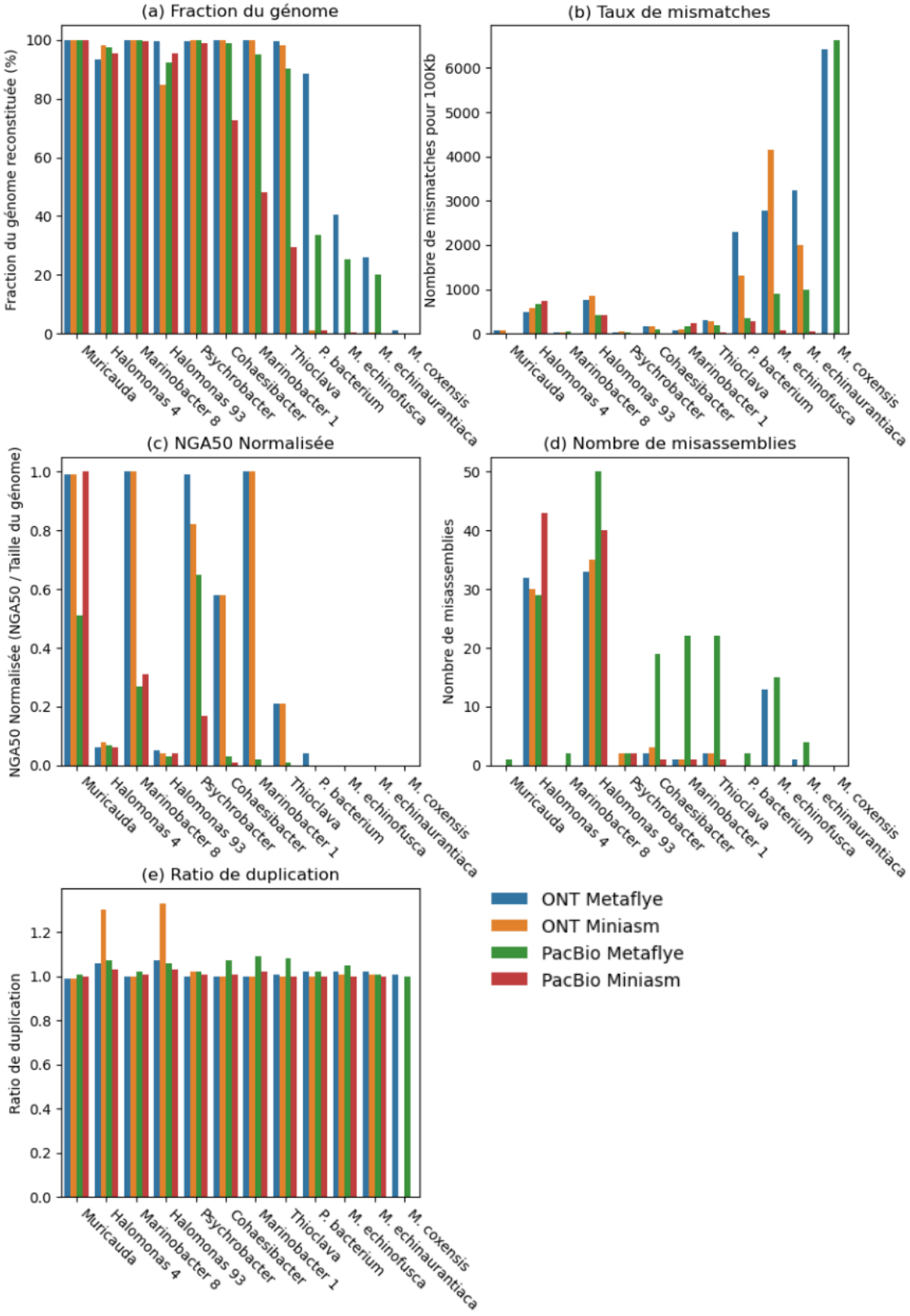

FIGURE 5.2 – **Comparaison des résultats d'assemblage selon la technologie de séquençage et l'assembleur, sur la communauté Bmock12.** (a) Fraction du génome,(b) Taux de mismatches, (c) NGA50 normalisée, (d) Nombre de misassemblies, (e) Ratio de duplication. Les lectures ONT sont sous-échantillonnées pour être à la même profondeur de séquençage que les lectures PacBio. Les assemblages Miniasm ont été polis par Flye-polish. Les espèces sont rangées par ordre décroissant d'abondance relative, de gauche à droite.

5.2.2 Lectures longues à haute fidélité

Toujours avec la même pipeline, j'ai assemblé les lectures longues à haute fidélité PacBio, issues de la communauté Zymo D6331, à l'aide de trois assembleurs : Metaflye, Hifiasm-meta, et MetaMDBG, puis j'ai évalué ces assemblages avec références. Lors des analyses, je me référerai à quatre groupes : les 5 *Escherichia coli*, les 10 autres espèces fortement abondantes (couverture > 50X, *Escherichia coli* exclus), 3 espèces faiblement abondantes (couverture < 10X, mais >1X), et 3 espèces très faiblement abondantes (couverture < 1X). La figure 5.3 présente les résultats sous forme de boxplot plutôt que de barplots afin de condense l'information des 21 unités taxonomiques. Le barplot complet est disponible en annexes (fig. 7.1).

Pour les espèces fortement abondantes, les trois assembleurs sont excellents et équivalents en terme de complétude (près de 100% du génome reconstitué en moyenne pour les trois assembleurs), précision (moins de 10 mismatches pour 100km en moyenne), mais seuls Metaflye et MetaMDBG parviennent à avoir une excellente contiguïté (NGA50 normalisée de 97% pour metaMDBG et 98% pour Metaflye), tandis qu'Hifiasm-meta produit un assemblage très fragmenté (NGA50 normalisée de 5%). De plus, Hifiasm-meta produit un assemblage très fortement dupliqué (ratio de duplication de 21, contre 1 pour les deux autres assembleurs) et riche en misassemblies (33 contre 1 en moyenne pour les autres assembleurs) (fig. 5.3).

Pour les *Escherichia coli*, on observe une baisse de la qualité de l'assemblage, malgré une complétude restant excellente et équivalente entre les assembleurs (environ 99% du génome reconstitué en moyenne). Ils sont en revanche très fragmentés (NGA50 normalisée de 6% pour metaMDBG, 4% pour Metaflye et 2% pour Hifiasm-meta), et bien moins précis, surtout metaMDBG et Metaflye (695 mismatches pour 100kb pour metaMDBG et 625 pour Metaflye, contre 224 pour hifiam-meta). Les assemblages sont fortement dupliqués, surtout pour Hifiasm-meta (ratio de duplication de 9.0, contre 2.0 pour Metaflye et metaMDBG) et riches en misassemblies (90 pour Hifiasm-meta, 150 pour metaMDBG, et 137 pour Metaflye) (fig. 5.3).

Pour les espèces faiblement abondantes, on observe aussi une baisse de la qualité de l'assemblage par rapport aux espèces les plus abondantes, que ce soit en terme de complétude (85% du génome reconstitué pour metaMDBG, 79% pour Hifiasm-meta, et 66% pour Metaflye), de contiguïté (NGA50 normalisée de 18% pour Hifiasm-meta, 8% pour metaMDBG, et incalculable pour Metaflye) et de précision (146 pour Hifiasm-meta, 138 pour metaMDBG, et 128 pour Metaflye), ou du nombre de misassemblies, surtout pour metaMDBG et Hifiasm-meta (1318 pour metaMDBG et 1390 pour Hifiasm-meta, contre 757 pour Metaflye). Pour les espèces très faiblement abondantes en revanche, aucun des assembleur n'est efficace, ne reconstituant que 2% du génome pour Hifiasm-meta et Metaflye, et 5% pour MetaMDBG (fig. 5.3).

Ainsi, on confirme que les génomes les plus abondants et n'ayant pas d'autres génomes trop similaires dans le métagénome sont mieux reconstitués. Meta-MDBG produit un excellent assemblage pour les espèces les plus abondantes et reconstruit la plus grande fraction du génome des espèces faiblement et très faiblement abondantes. Metaflye rivalise avec MetaMDBG pour les espèces fortement abondantes, et produit des assemblages d'espèces faiblement abondantes les plus contigus. Hifiasm-meta produit des assemblages généralement fortement dupliqués et fragmentés.

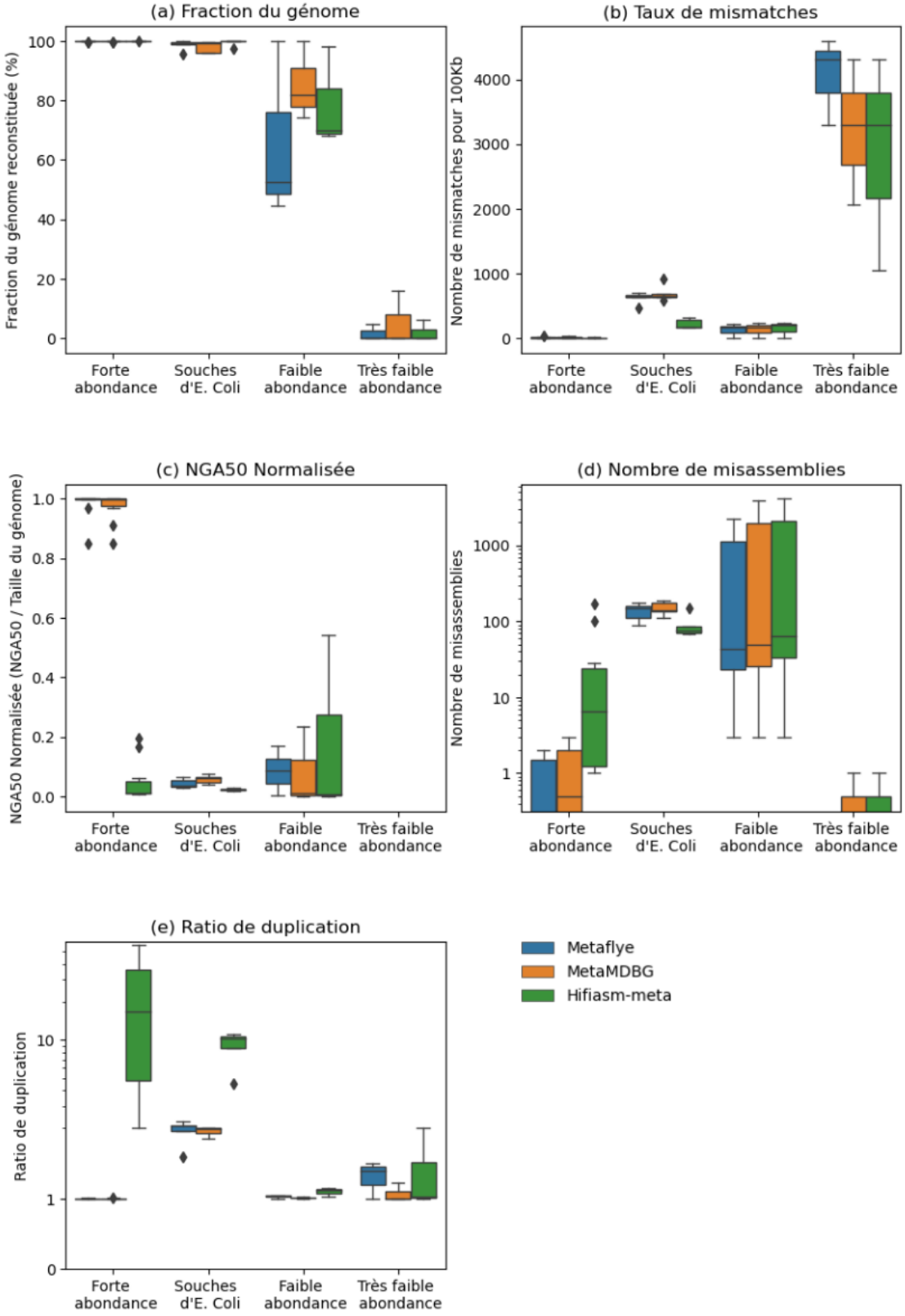

FIGURE 5.3 – **Comparaison des résultats d'assemblage selon l'assembleur, sur la communauté ZymoD6331.** (a) Fraction du génome,(b) Taux de mismatches, (c) NGA50 normalisée, (d) Nombre de misassemblies, (e) Ratio de duplication. La catégorie "forte abondance" regroupe 10 espèces ayant une profondeur de séquençage > 50X, "souches d'E. coli" regroupe 5 souches ayant une profondeur de séquençage > 50X, "faible abondance" regroupe 3 espèces ayant une profondeur de séquençage comprise entre 1 et 10X, et "très faible abondance regroupe 3 espèces ayant une profondeur de séquençage inférieure à 1X.

5.3 Effet du sur-séquençage

Les expériences précédentes ont montré qu'une espèce plus abondante est généralement mieux reconstituée. En revanche, il nous reste à déterminer si augmenter cette abondance est toujours bénéfique et augmente la qualité des assemblages à chaque niveau d'abondance relative. On cherche ainsi à déterminer si augmenter la profondeur de séquençage augmente effectivement la qualité de l'assemblage. Pour cela, j'ai comparé l'assemblage par Metaflye des lectures ONT issues de la communauté Bmock12, avec un jeu de données à 3.7Gb (Taille cumulée de toutes les lectures de 3.7 giga bases), et une version sous échantillonnée à 1.5Gb (le sous-échantillonage ayant été originalement prévu pour pouvoir comparer les technologies PacBio et ONT) avec Metaflye. Cette communauté disposant aussi de lectures courtes illumina, j'ai aussi effectué un assemblage hybride avec OPERA-MS.

Afin de facilité les comparaisons, les espèces sont groupées en fonctions de leur abondance dans le jeu de données ONT sous-séquencé, en espèces fortement, moyennement ou faiblement abondantes, et ce même si le surséquençage les aurait fait changer de catégorie.

En terme de complétude, augmenter la profondeur de séquençage augmente fortement la fraction du génome reconstituée pour les espèces faiblement abondantes (52% => 97%), et légèrement pour les *Halomonas* (96% => 99%), et ne la dégrade pas pour les espèces déjà reconstituées à 100% dans le jeu de données sous-échantillonné (fortement ou moyennement abondantes).

La contiguïté est améliorée seulement pour les espèces les moins abondantes, précédemment incalculable due à une trop faible portion du génome reconstruit, maintenant avec une NGA50 normalisée de 28%. L'augmentation de la profondeur de séquençage n'affecte pas la contiguïté des autres espèces.

La précision est améliorée pour les espèces faiblement abondantes (mismatches 2773 => 1532), moyennement abondantes (148 => 74), mais ne semble pas avoir d'effet sur les *Halomonas*, (626 => 577), et diminue même pour les espèces fortement abondantes (51=> 225).

Augmenter la profondeur de séquençage n'a pas l'air d'avoir d'effet sur le nombre de misassemblies.

Augmenter la couverture avec des lectures courtes pour un assemblage hybride est une approche efficace pour améliorer la qualité de l'assemblage des espèces peu abondantes, à la fois en terme de complétude (fraction du génome reconstitué : 52% => 99%), de fractionnement (NGA50 : incalculable => 2%) et de précision (mismatches pour 100kb : 2773 =>57).

En revanche, Opera-MS échoue à assembler les génomes fortement abondants, *Halomonas* inclus (NGA50 : 98% => 2%), et même les génomes moyennement abondants sont moins bien assemblés, que ce soit en terme de complétude (100% => 84%) ou de fractionnement (NGA50 normalisé : 69% => 20%).

Si l'on combine les deux approches, en utilisant Opera-MS avec ONT surséquencé, on obtient un résultat assez similaire à simplement utiliser Opera-MS sans surséquençage. Pour les génomes faiblement abondants, la fraction du génome reconstituée est légèrement meilleure (99% => 100%), mais il y a légerement plus de mismatches (57% => 84%).

Ainsi, augmenter la profondeur de séquençage en lectures longues (1.5 => 3.7 Gb) améliore grandement la qualité de l'assemblage des espèces faiblement abondantes en terme de complétude, de fractionnement et de précision. Cela augment aussi la qualité de l'assemblage des espèces moyennement abondantes, mais seulement en terme de précision. L'effet est plus ambigu pour les espèces fortement abondantes, pour lesquelles la complétude n'augmente que pour les *Halomonas*, et la précision baisse pour les autres.

De même, Opera-MS, en ajoutant des lectures courtes à l'assemblage, améliore grandement la qualité de l'assemblage des espèces faiblement abondante et, comparé au sur-séquencage en lectures longue, offre une complétude légèrement meilleure, une bien meilleure précision mais une moins bonne contiguïté. En revanche, Opera-MS n'arrive pas à reconstituer les espèces fortement ou moyennement abondantes.

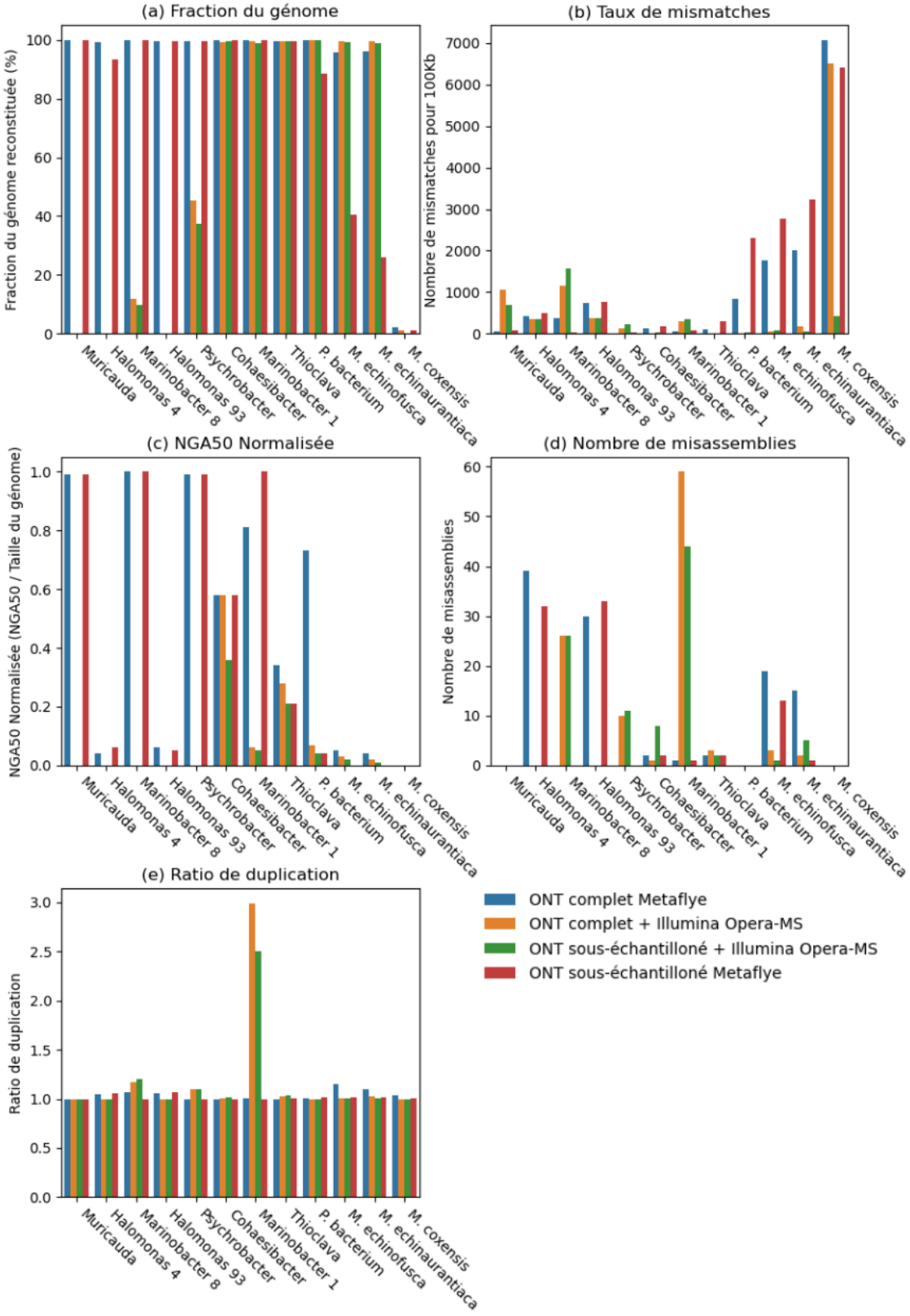

FIGURE 5.4 – **Comparaison de l'effet de stratégies de surséquencage sur la qualité de l'assemblage de la communauté Bmock12.** (a) Fraction du génome,(b) Taux de mismatches, (c) NGA50 normalisée, (d) Nombre de misassemblies, (e) Ratio de duplication. Le jeu de données ONT complet contient 3.7 Gb (en nombre de bases total dans les lectures), et la version sous-séquencée contient 1.5 Gb. Le jeu de données Illumina contient 64.4 Gb. Les espèces sont rangées par ordre décroissant d'abondance relative, de gauche à droite.

5.4 Comparaison des métriques

		Par référence					Par binning				Sans références			
Dataset	Assembler	Fraction du génome	NGA50 normalisée	mismatches pour 100kb	Misassemblies	Ratio de duplication	Presque complètes	Haute qualité	Moyenne qualité	Basse qualité	Longueur d'assemblage	N50	Ratio de lectures alignées	Ratio de longueurs alignées
Zymo HiFi	Hifiasm-meta	100%	2%	31	47	33.8	0	0	2	57	771480745	20996	100%	546%
	Metaflye	100%	87%	85	18	1.1	8	3	0	8	64180720	2028387	100%	123%
	MetaMDBG	100%	87%	95	21	1.1	9	3	0	6	69309381	2140404	100%	114%
Bmock ONT complet	MetaFlye	100%	60%	309	12	1.0	0	0	7	1	50223719	3211433	100%	91%
	Miniasm	98%	54%	298	11	1.2	0	0	6	2	42105477	3567321	99%	101%
	Opera-MS	26%	5%	540	9	1.2	0	1	0	9	44456846	218962	100%	79%
Bmock ONT sous échantillonné	MetaFlye	98%	60%	315	11	1.0	0	0	5	2	39160320	3568443	100%	90%
	Miniasm	95%	58%	338	11	1.1	0	0	4	3	32752974	3569160	100%	94%
	Opera-MS	25%	3%	553	9	1.1	0	1	1	4	41831907	152481	96%	76%
Bmock PacBio	MetaFlye	95%	25%	232	17	1.0	0	4	1	4	37348379	133166	93%	70%
	Miniasm	88%	28%	206	14	1.0	0	3	1	5	22936213	683904	85%	63%
	Opera-MS	21%	0%	547	4	1.0	0	1	1	2	34240056	68227	91%	49%

TABLE 5.1 – **Tableau de comparaison des métriques par références, par binning, ou sans références**

Il est compliqué de comparer les métriques, puisque celles avec références se font pour chaque génome de chaque assemblage, alors que celles sans référence se font seulement pour chaque assemblage. Puisque les métriques basées sur le binning et sans-références n'offrent qu'un aperçu global, j'ai décidé de les comparer aux moyennes des métriques avec références, pondérées par la profondeur de séquençage de chaque génome(table 5.1).

4 Assemblages Bmock ont à la fois un N50 de l'ordre de grandeur du méga et un ratio des longueurs d'alignement superieur ou égal à 90%. Ces assemblages sont extrèmement bien reconstruits, ayant une fraction du génome reconstitué à plus de 95% et une NGA50 normalisée supérieure à 50%. En effet, leur N50 est de l'ordre de grandeur de la taille d'un génome bactérien, indiquant qu'au moins la moitié de l'assemblage est constitué de génomes reconstitués en un seul contig. De plus, le ratio des longueurs d'alignement indique qu'il y a eu un bon alignement entre lectures et contigs, et que l'assemblage capture donc bien une bonne partie de la diversité de l'environnement.

Pour Zymo, on observe encore une fois la corrélation entre bon assemblage et N50 de la taille d'un génome bactérien, mais on remarque aussi que, malgré son fort ratio de longueurs d'alignement, Hifiasm-meta produit un assemblage fortement dupliqué. On remarque ainsi que, toute communauté confondues, des 7 assembleurs ayant un ratio de duplication superieur à 1, les seuls ayant un ratio de duplication inferieur à 100% sont les trois ayant la plus petite fraction du génome reconstituée du groupe

En revanche, d'autres métriques, comme la longueur totale de l'assemblage, ou le ratio de lectures utilisées, ne semblent pas être liées à la vrai qualité de l'assemblage.

Pour les métriques basées sur les bins, en se concentrant sur les trois jeux de données de la communauté Bmock, on remarque qu'aucun assemblage ne peut former de bins prèsque complètes. Les deux assemblages produisant le plus grand nombre de bins de haute qualité (PacBio assemblé par Metaflye et Miniasm) sont ceux ayant le plus faible taux de mismatches. Les trois autres assemblages ayant réussi à former une bin de haute qualité sont formés par Opera-MS (sur les trois jeux de données), qui présente en apparence un fort taux de mismatches (la moyenne étant pondérée par l'abondance) mais produit les assemblages d'espèces peu abondantes les plus précis, l'une d'entre elle ayant pu produire cette bin de haute qualité.

Pour la communauté Zymo, on remarque que deux assembleurs (Metaflye et MetMDBG) parviennent à produire un assemblage produisant des bins prèsque complètes, ce que l'on peut potentiellement expliquer par le plus faible taux de mismatches des assemblages Hi-Fi par rapport aux lectures longues basses fidélité. On remarque en revanche que l'assemblage de Hifiasm-meta ne produit que des assemblages de basse ou moyenne qualité, ce qui pourrait être expliqué par son fort ratio de duplication (ou moins probablement car les assemblages Opera-MS sont tout autant fragmentés, par son faible NGA50).

Si des références ne sont pas disponibles, il est complexe d'évaluer la qualité d'un assemblage métagénomique. En utilisant des métriques basées sur l'alignement des contigs avec les lectures, on se confronte au problème de ne pouvoir évaluer la complétude qu'en supposant que l'assemblage

ne soit pas dupliqué (hypothèse souvent invalide dans les communautés complexes, dans lesquelles les assemblages de souches d'une même espèce sont souvent dupliqués).

Les métriques basées sur l'évaluation du binning sont qualitatives, mais des bins presque complètes indiquent une bonne contiguïté et un très faible taux de mismatches, et des bins de bonne qualité indiquent un faible taux de mismatches.

5.5 Assemblage d'écosystèmes complexes

Bien que l'assemblage semble fonctionner pour un écosystème simple comme une une simili-communauté, il nous reste à évaluer l'efficacité pour des écosystèmes plus complexes. En effet, dans les données réels, la multiplicité des souches pour une même espèce, un cas qui gène l'assemblage , est la norme.

On peu observer dans la figure 5.5 (a) que bien que près de 100% des lectures sont alignées à au moins une contig de l'assemblage pour tous les assembleurs dans les communautés relativement simples que sont ZymoD6331 et le microbiote humain, cette proportion décroît grandement pour les écosystèmes plus complexes que sont l'espace inter-rang et la rhizosphère du champ de concombre. Pour ces écosystèmes plus complexes, les contigs produites par metaMDBG s'alignent avec le plus de lectures des échantillons (en prenant l'espace inter-rang comme exemple, 70%, contre 43% pour Hifiasm-meta et seulement 33% pour Metaflye).

Si on ce concentre maintenant sur la figure 5.5 (b), on remarque que Metaflye et MetaMDBG ont un plus grand ratio des longueurs d'alignements pour le microbiote humain que pour la communauté Zymo (147% contre 123% pour Metaflye, 124% contre 114% pour MetaMDBG). Cela pourrait soit être du à un alignement de meilleur qualité, pour lequel les contigs sont mieux alignées avec les lectures, mais cela parrait improbable étant donnés les résultats de la figure 5.6 a et b, expliqués plus tard. Une piste plus probable est que la présence de souches augmente la duplication de l'assemblage, car pour ces deux assembleurs, seules les souches ont étés fortement dupliquées dans l'assemblage de la communauté ZymoD6331 (fig. 5.3 e)

Au contraire, le ratio des longueurs d'alignement d'Hifiasm-meta décroît entre zymo et le microbiote humain. On peut encore une fois expliquer ça soit par une dégradation de la qualité de l'assemblage ou une baisse réelle du taux de duplication de l'assemblage. Les bins étant de meilleur qualité (fig 5.6 a et b), on peut supposer que l'assemblage est moins dupliqué, ce qui reste cohérent avec la figure 5.3 e, car on peut supposer qu'une communauté plus complexe présente moins d'espèces fortement abondantes (très fortement duppliquées pour Hifiasm-meta), et plus de souches ou d'espèces faibleement abondantes.

Pour les communautés issues du champ de concombre, le ratio des longueurs d'alignement est bien plus faible, ce qui est lié en grande partie au plus faible ratio de lectures alignées, mais pas que. En effet, pour les trois assemblages de la rhizosphère, le ration des lectures alignées est superieur au ratio des longueurs d'alignements (39% contre 37% pour Hifiasm-meta, 26% contre 21% pour Metaflye, et 68% contre 50% pour MetaMDBG), ce qui indique qu'une plus faible partie de chaque lecture a pu être alignée au contigs, ou que seule les lectures les plus courtes ont été alignées au contigs.

Une observation surprenante au niveau de la comparaison de l'espace inter-rang et de la rhizosphère du champ de concombre, est que l'espace inter-rang, bien que théoriquement complexe, a un plus grand ratio de lectures alignées et de longueurs d'alignement (fig. 5.5). Cela pourrait potentiellement être expliqué par la profondeur de séquençage plus importante pour la rhizosphère que pour l'espace inter-rang(19.5Gb contre 13.4Gb), introduisant plus d'espèces faiblement abondantes et de variations, et complexifiant donc l'assemblage.

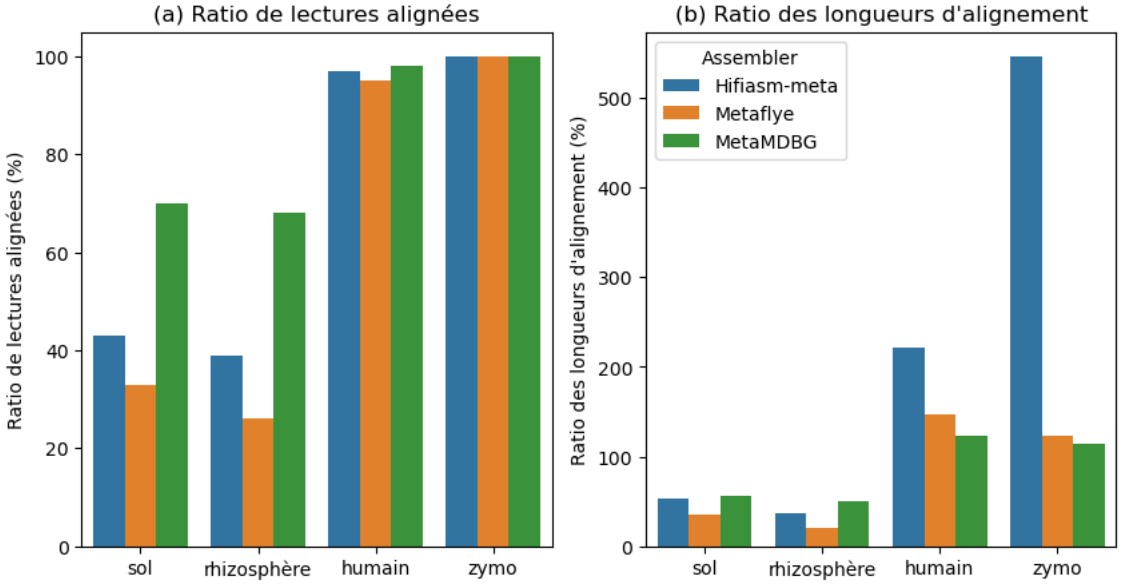

FIGURE 5.5 – **Recrutement des lectures dans les contigs selon l'assembleur et l'écosystème étudié.** (a) le ratio de lectures alignées, (b) le ratio des longueurs d'alignements. Le ratio de lectures alignées est calculé en divisant le nombre de lectures alignées à au moins une contig de l'assemblage par le nombre total de lectures du jeu de données. Le ratio des longueurs d'assemblage est calculé en divisant la somme des longueurs des alignements lectures-contigs par la somme des longueurs des lectures.

Après binning de chacun des assemblages, on observe qu'Hifiasm-meta produit pour chaque assemblage, plus de bins que les deux autres assembleurs (fig. 5.6). En revanche, ce n'est pas forcément un gage de qualité, puisque pour la communauté Zymo, Hifiasm-meta produit plus de bins qu'il n'existe d'unité taxonomiques dans la communauté. On peut aussi observer que MetaMDBG produit, pour chaque communauté, plus de bins presque complètes, de haute qualité, et de moyenne qualité que les deux autres assembleurs (à une exeption près, pour la rhizophère, ou metaMDBG produit 41 bins de moyenne qualité, contre 42 pour Hifiasm-meta), ce qui pourrait être expliqué par le fait qu'il produit des contigs s'alignant avec plus de lectures (fig. 5.5, et capture donc une plus grande part de la biodiversité des écosystèmes complexes.

Les bins ont en général une meilleur qualité pour les communautés simples. En effet, seuls les assemblages de la communauté Zymo et du microbiote humain produisent des bins presque complètes, et la proportion des bins de basse qualité augmente avec la complexité de la communauté (Pour metaMDBG, en ignorant les bins sans qualité, elle passe de 33% pour Zymo à 64% pour le microbiote humain et enfin à 83% pour les deux communautés du champ de concombre). La quantité de bins décroît aussi avec la complexité. En ignorant Zymo, pour qui le nombre d'unité taxonomique est très faible, le nombre de bins (sans qualité exclus) décroît (430 pour le microbiote humain, 287 pour la rhizosphère, et 218 pour l'espace inter-rang).

Ainsi, tous les assembleurs peine à reconstituer des écosystèmes complexes, capturant une plus faible portion de la diversité et produisant des contigs de moins bonne qualité.

En revanche, metaMDBG arrive mieux à reconstituer ces écosystèmes complexes que Metaflye et MetaMDBG, capturant une plus grande partie de la diversité de l'écosystème et produisant des bins de meilleur qualité.

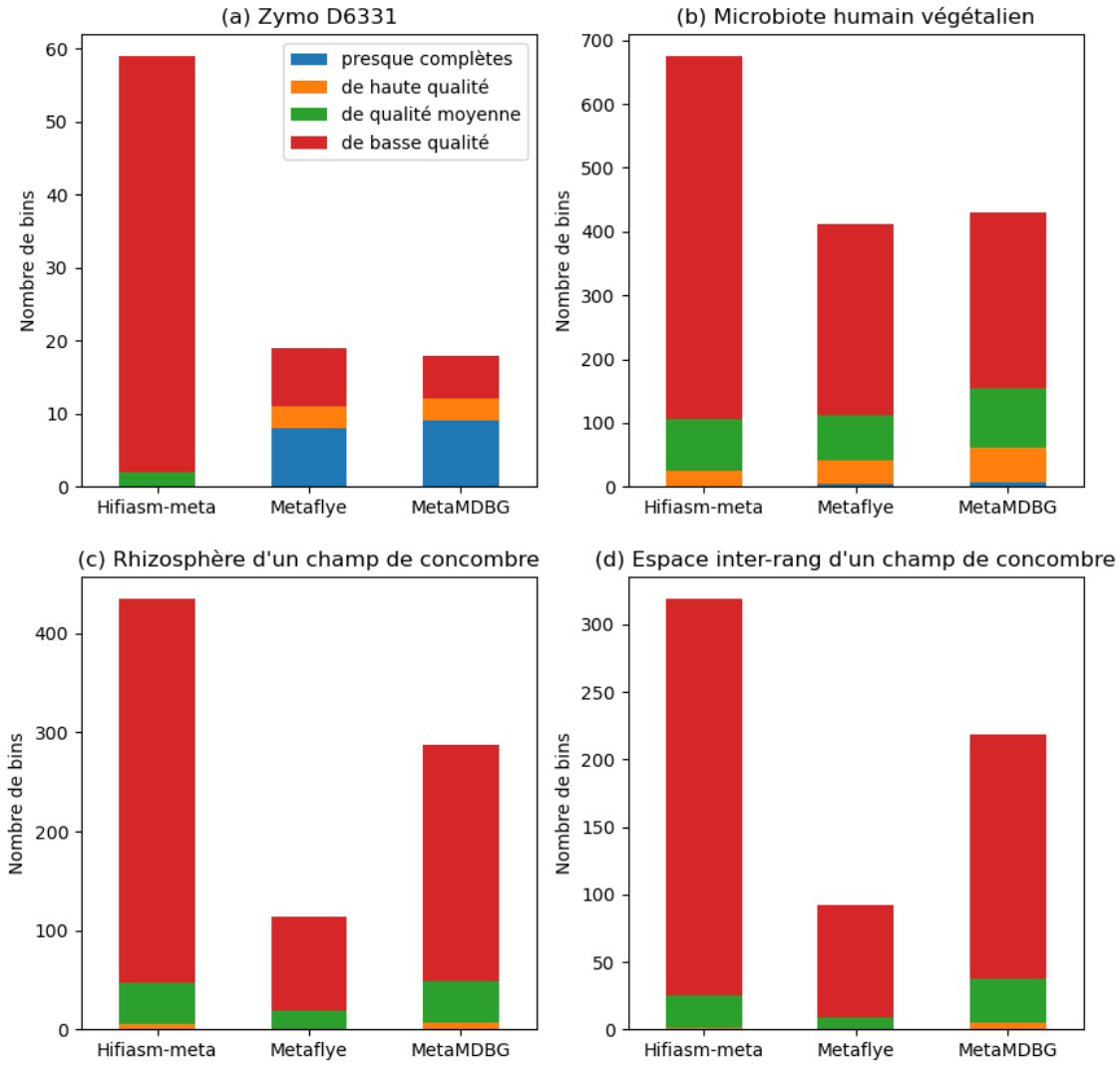

FIGURE 5.6 – **Qualité du binning selon l'assembleur et l'écosystème étudié**. Binning réalisé par MetaBat2 et évalué par CheckM. Les 4 niveaux de qualité des bins sont décrits dans la table 4.2, mais les bins sans qualité n'ont pas étés inclues

Chapitre 6

Discussion

Lors de ce stage, j'ai développé une pipeline d'assemblage et d'évaluation métagénomique, dédiée aux lectures issues des technologies de séquençage de 3ème génération, disponible sur `https://gitlab.inria.fr/stage_nmaurice/metagenomic_benchmark`.

Je l'ai ensuite utilisée pour explorer l'impact de la profondeur de séquençage et de la présence d'unités taxonomiques proches sur la qualité de l'assemblage de chaque espèce d'un métagénome, et l'impact de l'augmentation de la profondeur de séquençage ou de la complexité de la communauté sur la qualité de l'assemblage d'un métagénome. J'ai pour cela évalué diverses métriques d'évaluation afin de faciliter la comparaison d'assemblages d'écosystèmes complexes.

Ainsi, la principale difficulté rencontrée lors de l'assemblage d'un écosystème complexe est la présence de souches et d'espèces faiblement abondantes. Augmenter la profondeur de séquençage aide généralement à assembler ces espèces faiblement abondantes, mais peut parfois compliquer l'assemblage des espèces les plus fortement abondantes. Les lectures longues haute fidélité permettent de produire des assemblages bien plus précis que les lectures longues basse fidélité. Lorsque des références ne sont pas disponibles, la qualité des bins est généralement un bon indicateur de la qualité de l'assemblage. Metaflye, mais surtout MetaMDBG, arrivent à relativement bien assembler les souches et les espèces faiblement abondantes dans des simili-communauté, et MetaMDBG semble produire des assemblages de meilleure qualité pour les communautés plus complexes.

Néanmoins, il est encore possible d'améliorer cette pipeline et ces comparaisons, et voilà quelques pistes :

Il est toujours intéressant d'implémenter plus d'assembleurs, notamment hybrides (metaFlye-subassemblies, DBG2OLC, Opera-LG) ou lectures longues (Canu, Lathe), ce qui pourrait permettre non seulement de potentiellement obtenir un assemblage de meilleur qualité, mais aussi d'obtenir plus de points de données pour comparer différentes métriques et communautés.

Il pourrait aussi être intéressant, pour mieux évaluer les assembleurs dépendant d'une étape de polissage (notamment miniasm), d'implémenter une plus grande variété de polisseurs, comme Racon ou Pilon. Dans l'état actuel de ma pipeline, en comparant Miniasm et Metaflye, il est compliqué de déterminer si Metaflye est vraiment un meilleur assembleur, ou si Flye polish n'est pas adapté au polissage des assemblages Miniasm.

De même, pour l'évaluation de la qualité des bins, il n'est pas aisé de déterminer si un assemblage produisant des bins de meilleure qualité est lui même de meilleur qualité, ou s'il est simplement mieux adapté au binneur utilisé. Cela est d'autant plus vrai que Metabat2 a originellement été développé pour faire du binning sur des assemblages de lectures courtes Ainsi, implémenter plus de binneurs (notamment ceux ayant été évalué par Yue et Al.[35], ou un assembleur dédié aux lectures longues[36]), et laisser à la personne utilisatrice le choix duquel ou desquels utiliser, pourrait permettre de résoudre en partie cette ambiguïté.

Une autre piste d'amélioration est au niveau de la facilité et flexibilité d'utilisation. Plutôt qu'avoir un emplacement prédéfini pour les lectures, les génomes de référence et les métadonnées d'entrée, il pourrait être intéressant de permettre à la personne utilisatrice d'inclure un chemin dans le fichier de configuration, et ainsi éviter à la personne utilisant la pipeline d'avoir à déplacer ses fichiers. Il serait aussi possible d'encoder directement les métadonnées dans la pipeline, et ainsi éviter la création de fichier, peu pratique, et source d'erreur si la personne utilisatrice oublie de le

créer.

Enfin, pour les métriques avec références, il pourrait être avantageux de pouvoir entrer les abondances relatives de chaque génome manuellement, plutôt que de les calculer automatiquement. En effet, il est parfois compliqué de déterminer, en présence de souche, l'origine d'une lecture, et pour la communauté Zymo, les cinq souches d'*E. Coli* ont été déterminées comme ayant une abondance similaire, ce qui n'est pas cohérent avec la spécification de cette simili-communauté[20].

Pour les comparaisons entre assembleurs sur des communautés simples, il y a un grand manque de quantité de données, rendant impossible une conclusion définitive. Tester les différents assembleurs sur différentes simili-communautés permettrait de vérifier si les tendances que l'on observe sont spécifiques à une communauté en particulier, et si elle le sont, d'explorer quels facteurs pourraient jouer sur le succès d'un assembleur plutôt que d'un autre. Cela permettrait aussi d'être plus confiants dans ces résultats. Par exemple, il n'y a que 2 espèces fortement abondantes, *Halomonas* exclues, dans la communauté Bmock12, ce qui rend impossible de séparer les facteurs d'abondance des facteurs spécifiques à l'espèce.

Il serait aussi intéressant de tester ces assembleurs sur une même communauté ayant été séquencée à la fois en lecture longue basse-fidélité et haute-fidélité, afin de pouvoir comparer les technologies d'assemblage.

Pour de ce qui est de l'influence de la profondeur de séquençage, il pourrait être intéressant d'intégrer le sous-échantillonage à la pipeline, soit pour automatiquement mettre sur un pied d'égalité plusieurs séquençages et ainsi pouvoir mieux les comparer, soit pour explorer plus rigoureusement l'influence de la profondeur de séquençage sur la qualité de l'assemblage.

Il pourrait aussi être intéressant de pousser plus loin l'analyse des bins lorsque des références sont disponibles, en analysant avec référence le contenu de chaque bin. J'ai commencé à faire ce genre d'analyses, et des résultats préliminaires montrent, pour Bmock, que les espèces les moins bien reconstituées (*Halomonas* et espèces faiblement abondantes) ont tendance à être séparées sur plusieurs bins, ou à être mélangées à un autre génome dans une même bin.

Sans référence, il pourrait être intéressant de comparer les bins entre elles, afin de savoir si les mêmes espèces sont retrouvées par chaque assembleur.

Pour de ce qui est de la comparaison des métriques, il y a un manque de communautés suffisamment complexes avec références, qui permettraient notamment de comprendre la relation entre les métriques basées sur l'alignement des contigs avec les lectures et la qualité de l'assemblage.

Bibliographie

[1] R. Mendes, P. Garbeva, and J. M. Raaijmakers, "The rhizosphere microbiome : significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms," *FEMS Microbiology Reviews*, vol. 37, pp. 634–663, Sept. 2013.

[2] M. S. Afridi, M. A. Javed, S. Ali, F. H. V. De Medeiros, B. Ali, A. Salam, Sumaira, R. A. Marc, D. H. M. Alkhalifah, S. Selim, and G. Santoyo, "New opportunities in plant microbiome engineering for increasing agricultural sustainability under stressful conditions," *Frontiers in Plant Science*, vol. 13, 2022.

[3] P. Trivedi, J. E. Leach, S. G. Tringe, T. Sa, and B. K. Singh, "Plant–microbiome interactions : from community assembly to plant health," *Nature Reviews Microbiology*, vol. 18, pp. 607–621, Nov. 2020.

[4] T. P. Curtis, W. T. Sloan, and J. W. Scannell, "Estimating prokaryotic diversity and its limits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 10494–10499, Aug. 2002.

[5] C. R. Fitzpatrick, J. Copeland, P. W. Wang, D. S. Guttman, P. M. Kotanen, and M. T. J. Johnson, "Assembly and ecological function of the root microbiome across angiosperm plant species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, pp. E1157–E1165, Feb. 2018.

[6] C. E. Davies, K. E. Hill, M. J. Wilson, P. Stephens, C. M. Hill, K. G. Harding, and D. W. Thomas, "Use of 16S Ribosomal DNA PCR and Denaturing Gradient Gel Electrophoresis for Analysis of the Microfloras of Healing and Nonhealing Chronic Venous Leg Ulcers," *Journal of Clinical Microbiology*, vol. 42, pp. 3549–3557, Aug. 2004.

[7] S. Gupta, M. S. Mortensen, S. Schjørring, U. Trivedi, G. Vestergaard, J. Stokholm, H. Bisgaard, K. A. Krogfelt, and S. J. Sørensen, "Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing," *Communications Biology*, vol. 2, pp. 1–7, Aug. 2019.

[8] Z. Zhang, C. Yang, W. P. Veldsman, X. Fang, and L. Zhang, "Benchmarking genome assembly methods on metagenomic sequencing data," *Briefings in Bioinformatics*, vol. 24, p. bbad087, Mar. 2023.

[9] Z. Wang, Y. Wang, J. A. Fuhrman, F. Sun, and S. Zhu, "Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences," *Briefings in Bioinformatics*, vol. 21, pp. 777–790, May 2020.

[10] D. Bertrand, J. Shaw, M. Kalathiyappan, A. H. Q. Ng, M. S. Kumar, C. Li, M. Dvornicic, J. P. Soldo, J. Y. Koh, C. Tong, O. T. Ng, T. Barkham, B. Young, K. Marimuthu, K. R. Chng, M. Sikic, and N. Nagarajan, "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," *Nature Biotechnology*, vol. 37, pp. 937–944, Aug. 2019.

[11] H. Li, "Minimap and miniasm : fast mapping and de novo assembly for noisy long sequences," *Bioinformatics*, vol. 32, pp. 2103–2110, 03 2016.

[12] M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. L. Smith, and P. A. Pevzner, "metaFlye : scalable long-read metagenome assembly using repeat graphs," *Nature Methods*, vol. 17, pp. 1103–1110, Nov. 2020.

[13] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu : scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation," *Genome Research*, vol. 27, pp. 722–736, May 2017.

[14] E. L. Moss, D. G. Maghini, and A. S. Bhatt, "Complete, closed bacterial genomes from microbiomes using nanopore sequencing," *Nature Biotechnology*, vol. 38, pp. 701–707, June 2020.

[15] X. Feng, H. Cheng, D. Portik, and H. Li, "Metagenome assembly of high-fidelity long reads with hifiasm-meta," *Nature Methods*, vol. 19, pp. 671–674, June 2022.

[16] "Metamdbg, a lightweight assembler for long and accurate metagenomics reads."

[17] F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, D. Bertrand, J. J. Brito, C. T. Brown, J. Buchmann, A. Buluç, B. Chen, R. Chikhi, P. T. L. C. Clausen, A. Cristian, P. W. Dabrowski, A. E. Darling, R. Egan, E. Eskin, E. Georganas, E. Goltsman, M. A. Gray, L. H. Hansen, S. Hofmeyr, P. Huang, L. Irber, H. Jia, T. S. Jørgensen, S. D. Kieser, T. Klemetsen, A. Kola, M. Kolmogorov, A. Korobeynikov, J. Kwan, N. LaPierre, C. Lemaitre, C. Li, A. Limasset, F. Malcher-Miranda, S. Mangul, V. R. Marcelino, C. Marchet, P. Marijon, D. Meleshko, D. R. Mende, A. Milanese, N. Nagarajan, J. Nissen, S. Nurk, L. Oliker, L. Paoli, P. Peterlongo, V. C. Piro, J. S. Porter, S. Rasmussen, E. R. Rees, K. Reinert, B. Renard, E. M. Robertsen, G. L. Rosen, H.-J. Ruscheweyh, V. Sarwal, N. Segata, E. Seiler, L. Shi, F. Sun, S. Sunagawa, S. J. Sørensen, A. Thomas, C. Tong, M. Trajkovski, J. Tremblay, G. Uritskiy, R. Vicedomini, Z. Wang, Z. Wang, Z. Wang, A. Warren, N. P. Willassen, K. Yelick, R. You, G. Zeller, Z. Zhao, S. Zhu, J. Zhu, R. Garrido-Oter, P. Gastmeier, S. Hacquard, S. Häußler, A. Khaledi, F. Maechler, F. Mesny, S. Radutoiu, P. Schulze-Lefert, N. Smit, T. Strowig, A. Bremges, A. Sczyrba, and A. C. McHardy, "Critical Assessment of Metagenome Interpretation : the second round of challenges," *Nature Methods*, vol. 19, pp. 429–440, Apr. 2022.

[18] V. Sevim, J. Lee, R. Egan, A. Clum, H. Hundley, J. Lee, R. C. Everroad, A. M. Detweiler, B. M. Bebout, J. Pett-Ridge, M. Göker, A. E. Murray, S. R. Lindemann, H.-P. Klenk, R. O'Malley, M. Zane, J.-F. Cheng, A. Copeland, C. Daum, E. Singer, and T. Woyke, "Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies," *Scientific Data*, vol. 6, p. 285, Nov. 2019.

[19] M. B. Hall, "Rasusa : Randomly subsample sequencing reads to a specified coverage," *Journal of Open Source Software*, vol. 7, p. 3941, Jan. 2022.

[20] "ZymoBIOMICS Gut Microbiome Standard." https ://zymoresearch.eu/products/zymobiomics-gut-microbiome-standard.

[21] M. Blood, "Data Release : Human Microbiome Samples Demonstrate Advances in HiFi-Enabled Metagenomic Sequencing." https ://www.pacb.com/blog/data-release-human-microbiome-samples-demonstrate-advances-in-hifi-enabled-metagenomic-sequencing/, Aug. 2021.

[22] "Nanoplot — plotting tool for long read sequencing data and alignments."

[23] F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster, "Sustainable data analysis with Snakemake," Tech. Rep. 10 :33, F1000Research, Jan. 2021. Type : article.

[24] A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM : Simple Linux Utility for Resource Management," in *Job Scheduling Strategies for Parallel Processing* (D. Feitelson, L. Rudolph, and U. Schwiegelshohn, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 44–60, Springer, 2003.

[25] "GenOuest bioinformatics – Development, expertise and resources for bioinformatics."

[26] "PlaFRIM – Plateforme Fédérative pour la Recherche en Informatique et Mathématiques."

[27] "Conda — conda documentation."

[28] H. Li, "Minimap2 : pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, pp. 3094–3100, Sept. 2018.

[29] A. Mikheenko, V. Saveliev, and A. Gurevich, "MetaQUAST : evaluation of metagenome assemblies," *Bioinformatics*, vol. 32, pp. 1088–1090, Apr. 2016.

[30] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, 02 2021. giab008.

[31] T. pandas development team, "pandas-dev/pandas : Pandas," Feb. 2020.

[32] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[33] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "MetaBAT 2 : an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, p. e7359, July 2019.

[34] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, pp. 1043–1055, July 2015.

[35] Y. Yue, H. Huang, Z. Qi, H.-M. Dou, X.-Y. Liu, T.-F. Han, Y. Chen, X.-J. Song, Y.-H. Zhang, and J. Tu, "Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets," *BMC Bioinformatics*, vol. 21, p. 334, Dec. 2020.

[36] A. Wickramarachchi and Y. Lin, "Binning long reads in metagenomics datasets using composition and coverage information," *Algorithms for Molecular Biology*, vol. 17, p. 14, Dec. 2022.

Chapitre 7

Annexes

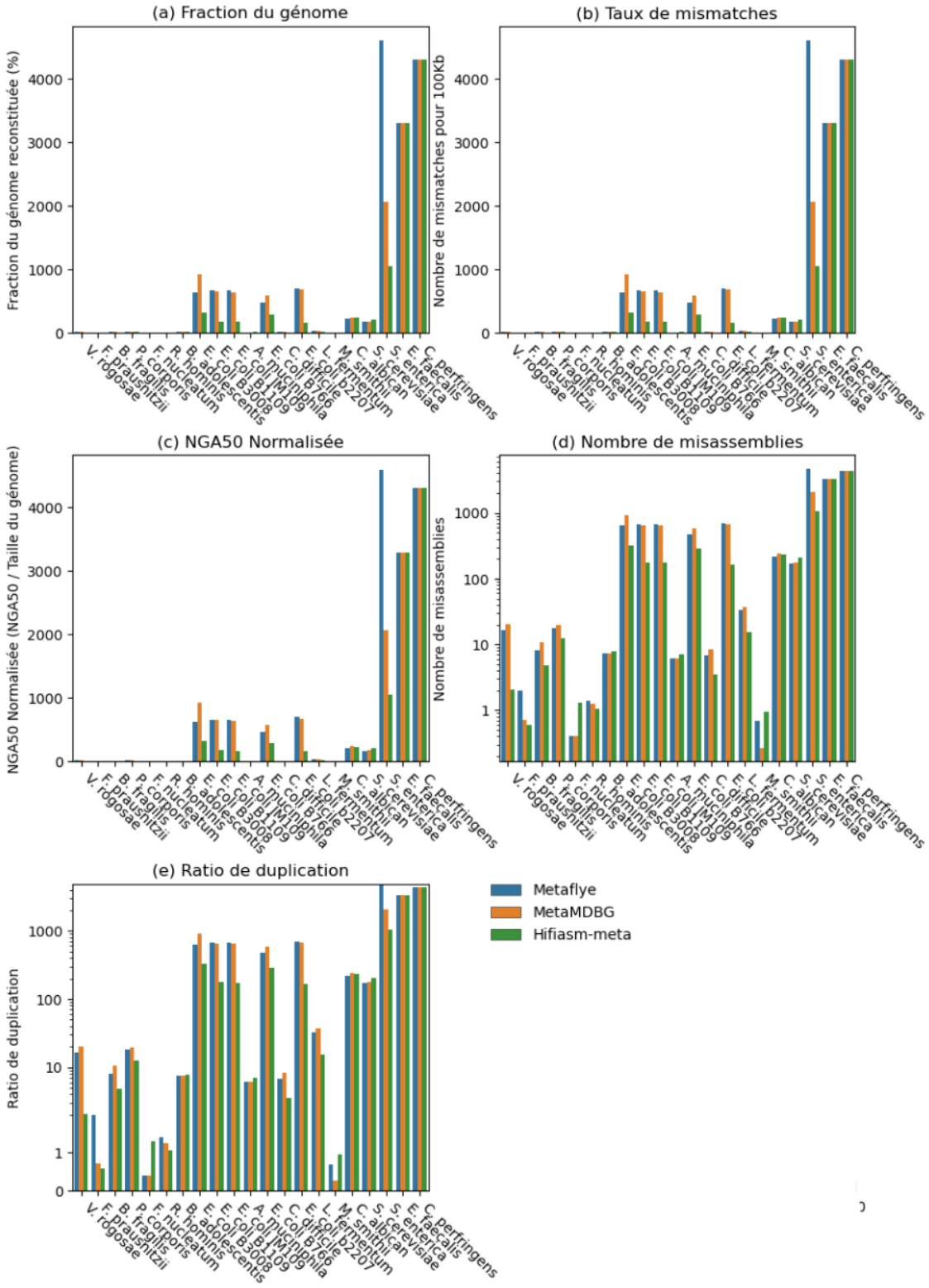

FIGURE 7.1 – **Comparaison des résultats d'assemblage selon l'assembleur, sur la communauté ZymoD6331.** (a) Fraction du génome,(b) Taux de mismatches, (c) NGA50 normalisée, (d) Nombre de misassemblies, (e) Ratio de duplication. Les espèces sont rangées par ordre décroissant d'abondance relative, de gauche à droite.