Métagénomique et métatranscriptomique

Cervin Guyomar, Claire Lemaitre

Chapitre **1**

Métagénomique et métatranscriptomique

Cervin Guyomar [1,1], Claire Lemaitre [3]

1 : iDiv – German Centre for Integrative Biodiversity Research, Deutscher Platz 5e, D-04103 Leipzig, Germany
2 : INRAE, UMR 1349 IGEPP, Le Rheu, France
3 : Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000 Rennes, France

1.1. Qu'est ce que la métagénomique ?

1.1.1. *Motivations et contexte historique*

Les micro-organismes (bactéries, virus et eucaryotes unicellulaires) représentent une part invisible mais importante de la biomasse, que ce soit par leur abondance ou leur diversité. Des estimations indiquent que la seule biomasse bactérienne représenterait près de 15% de la biomasse totale (Bar-On et al. 2018). Ce compartiment du vivant abrite également une immense diversité. Des micro-organismes sont présents dans tous les écosystèmes, y compris les plus extrêmes, et on estime que le nombre d'espèces microbiennes pourrait atteindre un billion (10^{12}) (Locey & Lennon 2016). Ils sont intégrés dans leurs écosystèmes, où ils assurent une multitude de fonctions. Ainsi, la totalité des macro-organismes sont par exemple associés à des micro-organismes impliqués entre autres dans leur métabolisme, ou leur santé, et l'étude des communautés microbiennes dans leurs environnements est une tâche essentielle.

Bien que l'étude de ces communautés soit ancienne, elle a longtemps été restreinte à l'utilisation de techniques d'imagerie permettant simplement d'observer

des caractères morphologiques. Dans ce cadre, seuls des organismes susceptibles d'être mis en culture pouvaient être étudiés. Ainsi, avant l'essor des technologies de biologie moléculaire, seule une étude à faible résolution d'une mince fraction des microbes existants était possible.

Les progrès des techniques de biologie moléculaire ont permis de contourner ces obstacles et ont contribué à révolutionner la microbiologie. L'avènement du séquençage Sanger a permis d'accéder à la structure et à la fonction des génomes bactériens. En particulier, l'ARN ribosomique est décrit comme un marqueur permettant de classer taxonomiquement les espèces. Ces approches sont appliquées à des données environnementales par Pace et collaborateurs, qui proposent en 1986 de séquencer l'ARN ribosomique directement dans l'environnement (Pace et al. 1986), sans passer par une étape de culture bactérienne. Cette idée permet de s'affranchir du biais de cultivabilité, qui rendait jusqu'alors invisible une large fraction de la diversité microbienne (Rappé & Giovannoni 2003). Proposé en 1998 par Handelsman et collaborateurs (Handelsman et al. 1998), le terme de *métagénomique* désigne le séquençage direct de l'ADN dans un milieu, qui permet potentiellement d'accéder aux génomes de tous les membres d'une communauté.

1.1.2. *Les données métagénomiques*

1.1.2.1. *Métagénomique ciblée et plein-génome*

On distingue deux principaux types de données métagénomiques. La métagénomique ciblée ou *metabarcoding* consiste à l'amplification puis au séquençage d'une région particulière du génome, appelée marqueur génomique. Une région fréquemment utilisée est l'ADN ribosomique 16S des bactéries, qui est un excellent marqueur phylogénétique. À l'inverse, la métagénomique plein-génome ou *shotgun* consiste au séquençage de tout l'ADN contenu dans l'échantillon. Plutôt que d'amplifier une région spécifique du génome, tout l'ADN du génome ou du métagénome est découpé aléatoirement en fragments séquencés avec les techniques haut-débit classiques. L'ensemble des génomes des membres de la communauté peut donc être séquencé.

Un certain flou existe sur la dénomination donnée à la métagénomique ciblée. Selon de nombreux auteurs, le terme de métagénomique est peu adapté, car il s'agit d'une technique ciblée sur une petite portion du génome (Esposito & Kirschberg 2014). Aussi, dans le reste du chapitre, qui porte essentiellement sur les techniques *shotgun*, nous emploierons le terme de métagénomique pour désigner la métagénomique plein-génome.

La métagénomique ciblée est utilisée pour caractériser taxonomiquement un échantillon. Parce que seule une petite fraction du génome est séquencée, elle est moins coûteuse que la métagénomique plein-génome, et permet d'identifier des

organismes plus rares pour un effort de séquençage équivalent. Pour ces raisons, elle a été prioritairement utilisée dans les première étapes de la métagénomique, ainsi que pour de grands projets de catalogage de la diversité bactérienne.

Par rapport à la métagénomique ciblée, l'information obtenue en métagénomique plein-génome est à la fois plus volumineuse, plus difficile à interpréter et plus riche. L'information plein génome permet d'atteindre une meilleure résolution taxonomique que les marqueurs utilisés en métagénomique ciblée. La métagénomique *shotgun* souffre moins des biais liés à l'amplification des séquences ciblées, ce qui la rend plus à même de représenter quantitativement des communautés. Enfin et surtout, alors que la métagénomique ciblée informe uniquement sur la composition taxonomique des communautés, la métagénomique plein génome renseigne sur le potentiel fonctionnel des communautés, à travers les répertoires des gènes de ses membres. Ainsi, la méthode *shotgun* est capable de répondre aux questions classiques posées en métagénomique sur les membres des communautés microbiennes qui sont "Qui-sont-ils ?" et "Que sont-ils capables de faire ?", tandis que la métagénomique ciblée est difficilement capable de répondre à la seconde.

En contrepartie, le séquençage d'un échantillon par la métagénomique plein génome est nettement plus coûteux, car une importante quantité de lectures doit être séquencée pour atteindre une couverture suffisante pour les organismes moins abondants. Toutefois, grâce aux progrès continuels des technologies de séquençage, cette technique se démocratise largement, et les grands projets de métagénomique se sont multipliés. On peut citer le projet MetaSoil (Delmont et al. 2011) pour l'étude du microbiome du sol, ou le projet HMP pour *Human Microbiome Project* (Turnbaugh, Ruth, Ley, Fraser-Liggett, Knight & Gordon 2007) qui vise à étudier les différents microbiotes humains. Métagénomiques *shotgun* et ciblée peuvent également être utilisées conjointement. Elles sont complémentaires, notamment dans le cas des communautés les plus complexes, où les organismes rares ne peuvent être identifiés que par des méthodes ciblées. On peut citer le projet TARA Océans (Bork et al. 2015*a*), qui vise à explorer la diversité microbienne des océans, et combine ces deux approches.

1.1.2.2. *Lectures de troisième génération*

L'émergence des technologies de séquençage à haut débit dans les années 2000 a marqué le début de l'essor de la métagénomique. Par rapport aux technologies précédentes, cette nouvelle génération permet de séquencer d'importants volumes de données pour un coût raisonnable. Les profondeurs de séquençage ainsi atteintes permettent ainsi d'explorer la quasi-intégralité des communautés bactériennes, avec des taux d'erreur faibles.

En revanche, ces technologies restent limitées pour certaines applications. Dans le cas du metabarcoding, la faible longueur des lectures amoindrit la résolution taxonomique atteinte, tandis que des biais lors de l'étape d'amplification empêchent

la bonne quantification des espèces présentes. En métagénomique plein-génome, l'assemblage des génomes des communautés est fortement limité par la longueur des lectures.

La principale restriction du séquençage de seconde génération est la longueur limitée des lectures, qui rend par exemple difficile voire impossible certains problèmes d'assemblage. Les technologies de séquençage les plus récentes peuvent être qualifiées de "longue portée" (Sedlazeck et al. 2018). Elles permettent par exemple de séquencer des lectures plus longues, pouvant atteindre le million de bases (technologies Pacific Biosciences ou NanoPore), ou bien de relier de courtes lectures provenant de la même région génomique (technologies 10X genomics et Hi-C). Ces techniques ont un potentiel élevé pour répondre à certains des problèmes posés par la métagénomique et exposés dans le paragraphe précédent. Ainsi, les plus longues lectures permettent d'améliorer les assemblages métagénomiques. Les techniques telles que le Hi-C peuvent être appliquées afin de distinguer les lectures provenant d'organismes différents (Burton et al. 2014). Cependant, le débit proposé par les technologies de séquençage de longues lectures ne rivalise pas à l'heure actuelle avec le séquençage de seconde génération. Ce paramètre étant critique dans les applications métagénomiques qui nécessitent un important effort de séquençage pour appréhender la diversité des communautés, l'usage des nouvelles méthodes reste conditionné à leurs progrès technologiques futurs.

1.1.3. *Défis bioinformatiques pour la métagénomique*

Si les données métagénomiques permettent en principe de porter un nouveau regard sur des communautés jusqu'alors mal connues, ces nouvelles données possèdent certaines particularités qui justifient et nécessitent le développement de méthodes dédiées.

1.1.3.1. *Volume de données*

Les séquençages de données métagénomiques peuvent générer des volumes de données importants. Dans des écosystèmes complexes tels que le sol ou l'eau de mer, un important effort de séquençage est nécessaire pour caractériser les organismes rares (Welch & Huse 2011, Roesch et al. 2007). À titre d'exemple, un échantillon du projet TARA Océans peut contenir près de 300 millions de lectures. Par ailleurs, de nombreuses études nécessitent le séquençage et l'analyse conjointe de plusieurs dizaines ou centaines d'échantillons, afin de comparer des écosystèmes différents.

Ce volume de données important pose des problèmes à toutes les étapes de l'analyse. Des outils et bases de données dédiés existent pour le stockage, l'indexation et le catalogage de telles données (IMG/MER (Chen et al. 2019), CAMERA (Seshadri et al. 2007), MG-RAST (Keegan et al. 2016*a*), et MGnify (Mitchell et al. 2019)). D'un point de vue algorithmique, la recherche de solutions

performantes permettant de gérer une multitude de jeux de données de taille importante est une priorité.

1.1.3.2. *Diversité génomique*

Les communautés bactériennes présentent généralement un continuum de diversité, découpé en plusieurs niveaux taxonomiques. Classiquement, on distingue au sein de ces communautés plusieurs espèces microbiennes différentes, dont les génomes peuvent présenter des régions homologues, par exemple suite au transfert horizontal d'un gène. Les abondances de ces espèces sont mesurées par leur couverture par les lectures métagénomiques, et elles peuvent être très déséquilibrées.

Par ailleurs, chaque espèce est représentée par un nombre variable d'individus, qui peuvent présenter des génotypes différents, incluant courts variants et variations structurales. Contrairement aux données génomiques dont la ploïdie est connue, les données métagénomiques abritent donc un nombre très important de variations. Ces différents variants peuvent également être quantifiés en mesurant leur couverture. En fonction de l'effort de séquençage fourni, certains variants rares pourront facilement être confondus avec les erreurs de séquençage, ou avec des régions provenant d'une autre espèce de la communauté.

Appréhender cette diversité est donc complexe, en particulier lorsqu'il s'agit de comparer différentes communautés pouvant abriter des espèces distinctes. Les tâches classiques de la génomique, telles que l'assemblage ou la recherche de variants, sont rendues difficiles par ce polymorphisme (Sczyrba et al. 2017), et nécessitent le développement d'algorithmes dédiés. Bien souvent, une manière d'appréhender ce problème est de restreindre l'analyse à des unités taxonomiques opérationnelles distinguables par des critères arbitraires (par exemple un seuil de similarité), en ne considérant pas la variabilité au sein de ces unités qui peut pourtant avoir des impacts fonctionnels. Cette simplification de la diversité métagénomique permet d'employer des métriques de diversité issues de l'écologie. Parmi ces métriques, on peut citer l'index de Shannon qui permet de mesurer la diversité dans un échantillon (diversité alpha) ou des mesures comme la distance de Jaccard ou la dissimilarité de Bray-Curtis qui permettent de traduire la dissimilarité entre échantillons (diversité beta). Cette dernière catégorie de métriques est décrite dans la section 1.4.1.

1.2. "Qui sont-ils ?" : Caractérisation taxonomique des communautés microbiennes

La caractérisation taxonomique, ou *metagenomic profiling*, vise à répondre à une question qui ne se pose généralement pas lors de l'étude d'un seul organisme, en décrivant au niveau taxonomique les organismes présents dans l'échantillon. Il s'agit, à partir de lectures métagénomiques, d'identifier et éventuellement de quantifier les organismes qui sont présents au sein d'une communauté. Plus formellement, on peut

distinguer deux problèmes, celui de l'identification des unités taxonomiques présentes et celui de quantification de chacune de ces unités dans l'échantillon. Le problème d'identification prend en entrée un ensemble de séquences nucléiques et renvoie une liste d'identifiants taxonomiques distincts. Le problème de quantification prend en entrée l'ensemble de séquences et la liste de taxons renvoyée par le premier problème et renvoie une valeur numérique associée à chaque identifiant taxonomique, représentant la quantité absolue (nombre de lectures par exemple) ou relative (pourcentage de lectures) de chaque taxon dans l'échantillon.

Cette section vise à présenter les différentes familles de méthodes qui permettent de répondre à cette question. Ces méthodes diffèrent notamment par le recours ou non à des bases de données existantes, construites à partir des génomes déjà séquencés et identifiés. On distingue ainsi des méthodes *avec référence* ou *sans référence*. Généralement, les méthodes qui reposent lourdement sur ces bases de données se révèlent difficilement capables d'identifier des organismes non caractérisés précédemment, ce qui est un problème critique en métagénomique. Par ailleurs, la performance de l'assignation taxonomique varie selon les différentes techniques. Certaines fournissent au mieux un inventaire des espèces présentes, tandis que d'autres permettent d'étudier des variations plus fines entre des souches microbiennes. C'est sous l'angle de ces deux aspects que cette section présente les différentes méthodes permettant la caractérisation taxonomique d'échantillons métagénomiques. Nous nous pencherons brièvement sur les méthodes dédiées aux données de *barcoding*, avant d'étudier plus en détail celles permettant l'assignation taxonomique de lectures issus de métagénomique plein-génome, recourant ou non à des bases de données de référence.

1.2.1. *Méthodes pour la métagénomique ciblée*

Le séquençage environnemental d'amplicons, ou métagénomique ciblée, est une méthode répandue pour l'analyse de la composition taxonomique d'une communauté microbienne. Parce qu'il ne s'agit pas d'une méthode plein-génome, un effort de séquençage modéré permet de séquencer une grande majorité des organismes présents. Une première approche pour analyser de telles données consiste à aligner les séquences contre des bases de données de séquences de marqueurs phylogénétiques (principalement l'ADN ribosomique 16S), tels que Silva (Quast et al. 2013) ou GreenGenes (DeSantis et al. 2006). Les lectures sont alignées sur un alignement de référence de marqueurs phylogénétiques, généralement à l'aide d'un programme d'alignement multiple tel que SINA (Pruesse et al. 2012) pour Silva. Étant donnée la taille restreinte des séquences des marqueurs utilisés, cette tâche reste relativement rapide. En revanche, cette méthode ne permet pas de traiter les organismes absents de ces bases de référence, et elle présente un certain nombre de biais liés notamment aux étapes d'amplification. Une autre approche, suivie par exemple par les outils Qiime (Caporaso et al. 2010) et Mothur (Schloss et al. 2009),

consiste à regrouper les lectures partageant une similarité de séquence très forte, de manière à former des OTUs (Operational Taxonomic Units). Les bases de données de référence peuvent ensuite être utilisées pour annoter les OTUs ainsi obtenus par un taxon.

La principale faiblesse du metabarcoding pour la caractérisation taxonomique de communautés provient du fait que bien souvent un unique marqueur est utilisé pour décrire l'échantillon. Ces marqueurs offrent rarement une résolution taxonomique suffisante pour déterminer quelles sont les espèces présentes dans l'échantillon, et *a fortiori* ne permettent pas d'identifier avec fiabilité différentes souches. Par ailleurs, d'éventuels transferts horizontaux de gènes d'ARN ribosomique peuvent également tromper l'assignation taxonomique (Schouls et al. 2003). Enfin, la quantification de l'abondance des taxons peut être perturbée par des biais d'amplification ou des variations de nombre de copies du marqueur considéré (Schouls et al. 2003).

1.2.2. *Méthodes plein-génome avec référence*

Ces méthodes consistent à comparer les lectures métagénomiques à une information de référence qui prend la forme de collections de gènes ou de génomes qui ont déjà été assignés à un taxon. L'approche la plus souvent employée consiste à aligner les lectures soit sur des génomes complets, soit sur des séquences qui ont été désignées comme des marqueurs taxonomiques. Alternativement, et pour répondre aux problèmes de passage à l'échelle sur les jeux de données de grande taille, d'autres méthodes permettent cette assignation sans recourir à de l'alignement.

1.2.2.1. *Alignement de séquences*

Dans le cas du séquençage d'une communauté composée d'organismes pour lesquels un grand nombre de génomes de référence est disponible, une première approche est d'aligner les lectures métagénomiques contre cette collection de génomes. Le programme Blast (Altschul et al. 1990*b*) a été majoritairement utilisé pour l'analyse des premiers jeux de données métagénomiques, mais des aligneurs plus rapides tels que DIAMOND (Buchfink et al. 2014) ont également été conçus pour mieux passer à l'échelle sur des jeux de données importants. Le programme MEGAN (Huson et al. 2016) est une référence dans l'analyse d'alignements de jeux de données métagénomiques. À partir d'un résultat de type Blast, il permet d'assigner une lecture à un taxon, qui est le plus petit ancêtre commun des taxons avec lesquels cette lecture s'aligne. Il propose en complément de multiples options de visualisation. En corrigeant le nombre d'alignements par la longueur des génomes de référence, des outils tels que GAAS (Angly et al. 2009) ou GRAMMY (Xia et al. 2011) permettent de quantifier l'abondance des différents taxons. La résolution atteinte par ces outils est hautement dépendante de la densité de la base de données de référence utilisée. Par ailleurs, l'assignation de lectures à des génomes de référence partageant une forte similarité de séquence (des souches proches par

exemple) nécessite l'emploi de méthodes dédiées, qualifiées de *strain tracking*. Les outils Pathoscope (Francis et al. 2013) et Sigma (Ahn et al. 2015) permettent d'effectuer cette tâche. Ces logiciels reposent en particulier sur des outils statistiques. Pathoscope repose sur un modèle bayésien, tandis que Sigma repose sur un modèle probabiliste pour estimer la probabilité qu'une lecture provienne d'un génome donné avec une optimisation par maximum de vraisemblance.

L'alignement des lectures métagénomiques sur des bases de données de référence présente deux inconvénients majeurs. Le premier est lié au temps nécessaire à l'alignement des séquences. En effet, le passage à l'échelle sur des jeux métagénomiques pouvant comporter des centaines de millions de lectures est difficile. Puisqu'il est nécessaire d'aligner chaque lecture, le temps dévolu à l'alignement augmente rapidement avec la taille des projets de séquençage. Ensuite, cette approche est très dépendante de la quantité et de la qualité des génomes de référence disponibles. La plupart des microorganismes constituant les échantillons métagénomiques ne sont pas cultivables, et le nombre de génomes de référence est limité. Ainsi, à l'exception des bactéries d'intérêt médical pour lesquelles de nombreuses souches sont connues, ces méthodes ne permettent pas d'atteindre une résolution supérieure à l'espèce. Finalement, en dehors de communautés modèles, de telles approches ne permettent pas d'étudier finement la diversité et le potentiel fonctionnel des microbiomes.

1.2.2.2. *Utilisation de marqueurs phylogénétiques*

Les méthodes basées sur l'alignement de lectures contre des génomes complets rencontrent des problèmes de passage à l'échelle dus à la taille des séquences de référence et au nombre de lectures à traiter. En conséquence, certaines des techniques d'assignation taxonomique utilisant des références se contentent d'un ensemble de séquences de gènes au lieu de génomes complets. Amphora (Wang & Wu 2013) utilise 31 gènes présents en copie unique dans les génomes bactériens. La faible longueur de ces séquences de référence favorise le passage à l'échelle par rapport aux méthodes d'alignement contre des génomes entiers. Par rapport à la métagénomique ciblée, l'emploi de plusieurs gènes distincts diminue les erreurs dues à d'éventuels transferts horizontaux. Le fait que ces gènes soient présents en copie unique permet également une meilleure quantification de l'abondance des espèces. Enfin, la résolution obtenue est supérieure à celle permise en métagénomique ciblée car ces gènes évoluent généralement plus rapidement que l'ARN ribosomique 16S. Dans le meilleur des cas, des souches bactériennes distinctes peuvent éventuellement être distinguées. Metaphlan (Truong et al. 2015) utilise quant à lui une base de données contenant près d'un million de gènes spécifiques de certains taxons. Des outils complémentaires permettent de détecter l'existence de différentes souches bactériennes à partir de profils de SNPs au sein des gènes marqueurs, ou par la présence/absence de gènes d'un pangénome identifié à partir de génomes de référence. L'une des limites de cette approche est qu'il est difficile de mettre à jour la

base de données de référence par de nouveaux génomes en identifiant de nouveaux marqueurs spécifiques.

Les outils reposant sur des bases de données de séquences de marqueurs phylogénétiques offrent de bons résultats lorsque la communauté étudiée est bien représentée dans la base de données (Sankar et al. 2015). Ils présentent en revanche des limites pour caractériser des souches bactériennes, et nécessitent pour ce faire des bases de données très fournies, ce qui restreint ces analyses à des communautés modèles.

1.2.2.3. *Méthodes sans alignement*

L'alignement de grands ensembles de lectures sur une large collection de génomes complets est une tâche coûteuse en temps de calcul. Pour pallier ce problème, sans restreindre l'analyse à des marqueurs phylogénétiques, des méthodes d'assignation taxonomique sans alignement ont été développées. La solution retenue pour permettre le passage à l'échelle de l'assignation taxonomique métagénomique est d'utiliser de courtes séquences, nommées *k-mers*, soit des séquences de taille k. Du fait de leur courte longueur, il est possible d'énumérer et d'indexer tous les k-mers présents dans une collection de génomes. Le premier outil utilisant cette technique est Kraken (Wood & Salzberg 2014). Kraken embarque une large base de données, contenant les k-mers présents dans près de 25000 génomes de bactéries, archées et virus. Pour chaque k-mer d'une lecture, l'index est interrogé pour obtenir le plus petit ancêtre commun des génomes le possédant. Cette information pour chaque k-mer est ensuite utilisée pour associer un taxon à la lecture. Une place est laissée à l'incertitude : si aucun génome n'est suffisamment proche de la lecture, cette dernière est associée à un niveau taxonomique plus élevé. Afin de permettre des performances compatibles avec le volume des données métagénomiques, ces outils nécessitent l'usage de structures de données particulières. Kraken repose par exemple sur une table de hash, construite à l'aide de minimiseurs de k-mers, ce qui permet de requêter rapidement des k-mers voisins dans la lecture, qui ont de grandes chances de partager un même minimiseur (voir le chapitre **??**). Dans son mode le plus rapide, Kraken peut assigner près de 4 millions de lectures par minute, ce qui en fait un outil très rapide. En revanche, il est nécessaire de charger en mémoire l'intégralité de l'index, ce qui nécessite des quantités élevées de mémoire (près de 70 GB pour l'index complet) et limite le nombre de génomes pouvant être inclus dans la base de données. Cette approche est raffinée dans l'outil Clark (Ounit et al. 2015), qui construit un index plus léger à partir de *k-mers* spécifiques de chaque génome. Plus récemment, les outils Kaiju (Menzel et al. 2016) et Centrifuge (Kim et al. 2016) ont été développés en utilisant une autre structure d'indexation ne reposant pas sur une table de hash des k-mers, mais sur un FM-index (voir le chapitre **??**), et qui permet de réduire l'usage mémoire et d'améliorer l'assignation. Ainsi, Centrifuge est par exemple capable de stocker 4300 génomes procaryotes dans un index occupant 4 Go de mémoire.

1.2.2.4. *Limites des méthodes basées sur des références*

Toutes ces méthodes reposent fortement sur des bases de données constituées à partir de génomes déjà séquencés. Cette approche se heurte à la faible représentativité de ces bases de données. Si un grand nombre d'espèces ont été séquencées, le nombre de souches disponibles pour une même espèce varie grandement. Ainsi, la base de données Refseq comporte à ce jour près de 144,000 génomes provenant d'environ 11,000 espèces, soit près de 13 génomes par espèce en moyenne. Cet effort de séquençage important ne représente pourtant qu'une infime partie des espèces bactériennes existantes (grossièrement estimé à plus d'un milliard dans (Dykhuizen 2005)). De plus, 60 % de ces espèces ne sont représentées que par une seule souche, et 14% des espèces les mieux représentées représentent 90% des génomes de RefSeq. Parmi ces espèces abondamment séquencées et assemblées, la plupart sont d'intérêt biomédical. Ainsi, les 3 espèces les plus abondantes dans RefSeq sont *Escherichia coli*, *Salmonella enterica* et *Staphylococcus aureus*, avec près de 10 000 souches chacune. L'assignation taxonomique de lectures issues de communautés peu étudiées souffre ainsi d'une faible résolution taxonomique : la caractérisation de la communauté se limite le plus souvent à l'inventaire des espèces présentes. Pourtant, des différences fonctionnelles importantes peuvent s'expliquer par des variations génomiques à des échelles inférieures. Pour répondre à cette limite, quelques approches ont été développées pour rechercher et exploiter les variations par rapport aux séquences de référence. ConStrains (Luo et al. 2015) et StrainPhlan (Truong et al. 2017) utilisent des profils de SNPs, détectés sur les gènes marqueurs de la base de données Metaphlan. S'il est ainsi possible de reconstruire une phylogénie des différentes souches identifiées, cette analyse est restreinte à quelques gènes, et ne permet donc pas d'accéder à la séquence génomique complète des micro-organismes, ce qui empêche d'évaluer l'impact fonctionnel des différentes souches.

1.2.3. *Méthodes sans référence*

La diversité microbienne étant en grande partie inconnue, les méthodes requérant la comparaison à des bases de données de référence montrent rapidement leurs limites. Des méthodes qualifiées de *de novo* ont été développées pour identifier de nouveaux génomes à partir de données métagénomiques en recourant pas ou peu à des génomes de référence, et sont le principal moyen d'identifier les membres des communautés microbiennes. Dans ces méthodes, le principal objectif est de regrouper les lectures provenant du même organisme.

1.2.3.1. *Assemblage metagénomique*

1.2.3.1.1. Principe de l'assemblage

Les limites actuelles des méthodes de séquençage rendent impossible de séquencer en une seule lecture des génomes complets. L'assemblage est la tâche permettant de transformer une multitude de courtes lectures en des portions plus longues du génome

(voir chapitre **??**). Les programmes d'assemblage génèrent généralement un graphe représentant les chevauchements entre les lectures. On distingue deux principales familles d'assembleurs, en fonction du type de graphe utilisé par l'assemblage. Les assembleurs à *overlap graph* tels que Celera (Denisov et al. 2008) utilisent les lectures complets comme nœuds du graphe. Les assembleurs reposant sur des graphes de *De Bruijn*, tels que Abyss (Simpson et al. 2009) recourent quant à eux à un graphe de k-*mers* (mots de taille k) plus petits que les lectures, qui rend plus facile la détection de chevauchements.

Dans ce graphe, l'assembleur recherche des chemins, dont la séquence va former des contigs, représentatifs du génome de l'organisme séquencé. Bien que ce soit possible, il est rare qu'un génome soit assemblé sous la forme d'un unique contig, car les graphes d'assemblage sont souvent complexes, ce qui contraint l'assembleur à interrompre les contigs. En particulier, c'est la présence de répétitions dans le génome qui fait qu'un assemblage est généralement découpé en contigs plus petits que le génome. Par exemple, lorsqu'un graphe de *De Bruijn* est utilisé pour l'assemblage, toute répétition d'un k-mer au sein du génome se traduit par une structure en "X" dans le graphe. Cette structure ne peut être résolue par l'assembleur sans information extrinsèque, ce qui force le programme à interrompre les contigs au niveau de telles répétitions.

Une autre source de complexité dans la structure de ces graphes est l'existence de chemins alternatifs, qui peuvent être dus soient à des erreurs de séquençage, soit à de véritables variations de la séquence génomique (par exemple les deux allèles d'un individu diploïde). Ces variants génèrent dans le graphe des structures en forme de bulle. La stratégie employée par la plupart des assembleurs est de retirer les k-mers les moins abondants dans le graphe, qui correspondent généralement à des erreurs de séquençage, et à fusionner les bulles restantes dues au polymorphisme.

Il est ensuite nécessaire d'évaluer la qualité d'un assemblage. Pour cela, un premier critère important est la longueur et le nombre des contigs. Un bon assemblage est constitué d'un faible nombre de contigs de grande taille, dont la somme des longueurs approche la longueur du génome ciblé. Un indicateur fréquemment utilisé et qui synthétise ces critères est le N50, qui est la longueur minimale permettant de couvrir au moins la moitié du génome avec des contigs plus grands. Toutefois, bien qu'étant des méthodes *de novo*, la bonne évaluation des assemblages nécessite également de vérifier la véracité des contigs, par exemple en les alignant au génome de référence attendu quand cela est possible. C'est rendu possible par des outils tels que Quast (Gurevich et al. 2013), qui reportent le nombre d'erreurs commises lors de l'assemblage.

1.2.3.1.2. Méthodes d'assemblage métagénomique

Les méthodes d'assemblage citées précédemment ont été développées dans l'objectif d'assembler un unique génome provenant d'une unique espèce.

Lorsqu'elles sont appliquées à des données métagénomiques, comme dans (Venter et al. 2004), les principales difficultés rencontrées en génomique classique sont exacerbées à cause de la diversité présente dans les communautés bactériennes.

Premièrement, une multitude d'espèces peuvent être représentées en quantités déséquilibrées dans les données de séquençage. Les assembleurs traditionnels utilisent l'information de la couverture du génome pour identifier des répétitions et les erreurs de séquençage. En contexte métagénomique, où l'abondance des différentes espèces varie, et où des régions génomiques peuvent être partagées par plusieurs espèces, cette stratégie n'est plus valable. Ainsi, dans (Venter et al. 2004), les génomes les plus couverts ont été considérés comme des répétitions par l'assembleur Celera, et une étape préalable a été nécessaire pour mieux les assembler. Comme indiqué précédemment, dans le cas d'un assemblage par graphe de *De Bruijn*, toute séquence répétée de longueur supérieure à k interrompt les contigs. Cette situation se produit fréquemment en métagénomique, où des espèces apparentées et pouvant partager certains gènes sont présentes. Les régions répétées entre les génomes à assembler s'ajoutent aux répétitions à l'intérieur d'un génome, et complexifient l'assemblage. Deuxièmement, le polymorphisme existant au sein des communautés bactériennes complique l'assemblage de plusieurs manières. De nombreux génotypes ou souches d'une même espèce microbienne peuvent être séquencés au sein d'un échantillon métagénomique. Ce polymorphisme n'est pas équitablement réparti le long des génomes, et il est difficile d'assembler conjointement les régions conservées et caractéristiques de souches différentes. D'éventuelles variations structurales au sein de la population compliquent également la tâche.

Afin de résoudre ces problèmes, des algorithmes dédiés ont été développés pour l'assemblage *de novo* de données métagénomiques, tels que IDBA-UD (Peng et al. 2012), MetaVelvet(Namiki et al. 2012), metaSPAdes (Nurk et al. 2017) ou MegaHit (Li et al. 2015). Tous ces programmes présentent différentes particularités algorithmiques qui, en principe, permettent à la fois le passage à l'échelle sur de larges jeux de données métagénomiques, et l'assemblage de mélanges d'espèces (voir la revue Ayling et al. (2020)). Par exemple, MetaVelvet utilise les différences de couverture et la connectivité d'un graphe de *De Bruijn* pour le séparer en sous graphes qui sont ensuite assemblés séparément grâce à l'algorithme de Velvet. À ce jour, les assembleurs considérés comme les plus performants sont ceux qui reposent sur une approche dite *multi-k* (Vollmers et al. 2017). Ils emploient successivement des graphes de *De Bruijn* construits avec des tailles de k-mers croissantes, les petites valeurs de k étant plus adaptées aux génomes peu couverts, et les plus grandes aux génomes plus abondants. Parmi ces assembleurs, IDBA-UD tient compte des fortes variations de couvertures présentes en métagénomique pour appliquer des seuils de simplification du graphe différents. MetaSPAdes ré-utilise l'algorithme de SPAdes qui était déjà conçu pour tenir compte des variations de couverture présentes dans les données de type single-cell. En plus de l'approche multi-k, il incorpore dans le

graphe l'information associée aux reads pairés, et parcourt le graphe de contigs pour identifier des contigs chimériques, des répétitions inter-espèces et d'éventuelles variations intra-spécifiques. Enfin, MegaHit se distingue par l'usage de *graphes de De Bruijn succincts*, une structure de données efficace qui permet d'atteindre un bon compromis entre qualité de l'assemblage et faible usage des ressources de calcul.

1.2.3.1.3. Limites de l'assemblage métagénomique

Malgré le développement de ces outils dédiés, le problème de l'assemblage métagénomique n'est pas résolu. Le challenge CAMI (Sczyrba et al. 2017) a permis de confronter différents outils sur des thématiques propres à la métagénomique, incluant l'assemblage. Les résultats illustrent la difficulté de cette tâche avec les méthodes actuelles. Dans ce concours, pour le jeu de données de "haute complexité", qui contient 596 génomes dont 399 montrant plus de 95% d'ANI, l'assemblage le plus long ne couvre que 70% de la communauté, au prix de près de 8000 erreurs dans l'alignement avec les génomes de référence ciblés. La qualité des assemblages obtenus dépend tout d'abord de la couverture des génomes. Naturellement, les génomes les moins couverts sont difficilement assemblés. De manière moins intuitive, certaines assembleurs peinent à assembler des génomes très abondants. Une parade employée par certains assembleurs (tels que MegaHit (Li et al. 2015)) est d'utiliser différentes tailles de k-mers, adaptées à des abondances différentes. En complément, la présence de génomes fortement apparentés dans la communauté rend difficile l'assemblage de ces espèces pour tous les outils considérés. Ce point semble particulièrement problématique, étant donné l'existence au sein de la plupart des communautés d'un continuum de diversité entre les individus.

Par ailleurs, la question de la pertinence de représenter l'assemblage d'un métagénome sous forme de séquences linéaires peut se poser. Des bulles créées par le polymorphisme ponctuel ou des branchements dus aux variations structurales sont présentes dans le graphe d'assemblage, mais absentes des contigs. Les assembleurs suppriment ces variations, en écrasant les bulles et en arrêtant les contigs lorsque des branchements se produisent. Le résultat est un ensemble de contigs, dont la séquence est un consensus de celles des organismes séquencés, et dans lequel les variations structurales ne sont pas représentées. Une tendance actuelle est d'aller au-delà de cette représentation linéaire d'un génome par des séquences vers une représentation sous forme de graphe. Les assemblages sont construits à partir de graphes, dont les régions linéaires sont extraites pour donner des contigs. Par rapport au graphe d'assemblage, les contigs généralement donnés en sortie du programme d'assemblage contiennent donc moins d'information. Formellement, ces structures sont des graphes bi-dirigés, où les noeuds sont des séquences nucléotidiques et les arrêtes représentent des chevauchements entre séquences. Puisque les séquences d'ADN peuvent être lues dans les deux sens, 4 types de chevauchements sont possibles (*forward-forward*, *forward-reverse*, *reverse-forward* et *reverse-reverse*). Ainsi, il est proposé de remplacer les génomes de référence linéaires par des graphes

rendant compte des variations génomiques (Paten et al. 2017), et certains assembleurs tels que metaSPAdes (Nurk et al. 2017) incluent dans leurs résultats un graphe au format GFA où la diversité structurale des génomes peut être observée.

Finalement, il est à ce jour impossible d'assembler complétement et fidèlement les organismes d'un métagénome, et de rendre compte de la diversité génomique dans ces communautés. Les assembleurs retournent des contigs les plus longs possible, tout en évitant de construire des chimères en assemblant des lectures issues de différents organismes. La diversité intra-spécifique est le plus souvent ignorée, de manière à retourner un consensus des génomes présents. L'assemblage n'est donc pas suffisant pour caractériser la diversité métagénomique d'un échantillon, et des outils complémentaires sont nécessaires pour retrouver les contigs issus d'une même espèce.

1.2.3.2. *Binning de séquences métagénomiques*

Les méthodes de binning ont pour but de regrouper des séquences de même origine taxonomique. Elles prennent en entrée des contigs préalablement assemblés, et les placent dans des clusters en fonction de leur origine taxonomique supposée. Étant donné qu'aucune information de référence n'est utilisée, ces bins ne sont pas associés à un taxon, mais ils peuvent ensuite être assemblés relativement facilement par des méthodes classiques, ce qui permet en théorie de reconstituer les génomes complets des membres de la communauté.

Une première famille de méthodes de binning utilise le contenu nucléotidique des séquences, en faisant l'hypothèse que le contenu nucléotidiques ou en mots d'une certaine taille est homogène le long du génome d'une espèce et est différent entre deux espèces suffisamment éloignées phylogénétiquement. La première méthode de ce type est TETRA (Teeling et al. 2004), qui calcule pour chaque séquence des profils de tétranucléotides et la corrélation entre ces profils, ce qui permet de regrouper les lectures provenant d'organismes proches. La limite majeure des méthodes basées sur le contenu nucléotidique est qu'elles ne s'appliquent qu'à des fragments génomiques de grande taille (près de 10 kilobases), dans lesquels le contenu nucléotidique est représentatif de celui du génome entier. Elles ne peuvent donc s'appliquer qu'à des contigs préalablement assemblés et de grande taille, ce qui limite fortement leur intérêt. En plus de la composition nucléotidique, il est possible de classifier des contigs en fonction de leur couverture, selon l'hypothèse que des contigs avec des couvertures similaires proviennent des mêmes génomes. Certaines méthodes utilisent la couverture *différentielle* entre des jeux de données provenant par exemple d'endroits différents. L'idée sous-jacente est que des séquences provenant du même organisme, en plus d'avoir des niveaux de couverture similaires dans un échantillon, auront une couverture qui covarie de la même manière au sein de plusieurs échantillons. Les outils utilisant à la fois l'information de couverture et de composition obtiennent généralement les meilleures performances dans le binning

de contigs. Ce sont ces outils qui se sont imposés comme les plus performants pour le binning de séquences. On peut citer parmi les principaux logiciels Concoct (Alneberg et al. 2014), GroopM (Imelfort et al. 2014) ou MaxBin (Wu et al. 2014). Ces multiples alternatives retournent parfois des résultats différents, ce qui a encouragé le développement d'outils pour la comparaison visuelle de leurs résultats (VizBin (Laczny et al. 2015)), ou de validation de la qualité des bins (CheckM (Parks et al. 2015)).

La principale limite rencontrée par ces techniques est qu'elles regroupent des contigs, issus d'un assemblage préalable. Dans ces contigs, la majorité de la diversité intra-spécifique a été éliminée par l'assembleur. Il est donc difficile, voire impossible pour ces méthodes de caractériser des variations génomiques fines, comme la présence de différentes souches bactériennes. Par ailleurs, le binning est sensible aux éventuelles erreurs lors de l'assemblage, comme la création de contigs chimériques. Enfin, certaines particularités génomiques rendent le binning difficile : une région génomique particulièrement riche en variants pour une souche particulière peut par exemple être affectée à un nouveau cluster différent du reste du génome.

Peu d'outils proposent un binning au niveau des lectures, car elles sont trop courtes pour apporter une signature génomique fiable, et trop nombreuses pour proposer des outils passant à l'échelle sur de grands jeux de données. LSA (Latent Strain Analysis) (Cleary et al. 2015) est un outil original permettant le binning de lectures métagénomiques, qui permet un assemblage indépendant de chaque bin. La méthode repose sur la comparaison de profils d'abondance de k-mers au sein de plusieurs échantillons. À partir d'une matrice donnant l'abondance de chaque k-mer dans chaque jeu de données, LSA procède à une décomposition en valeurs singulières (SVD) qui permet de sélectionner des k-mers montrant le même profil de covariance, et provenant vraisemblablement du même génome. Une étape de clustering permet ensuite de construire des ensembles de lectures associées. Les bins ainsi reconstruits peuvent permettre d'isoler des souches distinctes, ce qui correspond à une résolution rarement atteinte par les méthodes de binning de contigs.

1.3. "Que font-ils" : Métagénomique fonctionnelle

L'une des questions centrales de l'étude des communautés microbiennes est la caractérisation de l'effet fonctionnel de leurs différentes composantes. La métagénomique permet de séquencer l'intégralité des génomes d'une communauté, et par conséquent d'accéder à leur contenu en gènes. Il est donc possible de décrire des fonctions inédites qui peuvent avoir un grand intérêt, que ce soit pour la santé (par exemple par la recherche d'antibiotiques (Garmendia et al. 2012)), l'agro-alimentaire (De Filippis et al. 2017) ou l'énergie (Tiwari et al. 2018). Comme pour la caractérisation taxonomique, on peut distinguer deux problèmes de caractérisation fonctionnelle, celui de l'identification des fonctions présentes et celui

de quantification de chacune de ces fonctions dans l'échantillon. Le problème d'identification prend en entrée un ensemble de séquences nucléiques, le plus souvent issues de l'assemblage d'un échantillon métagénomique, et renvoie une liste d'identifiants fonctionnels distincts, comme par exemple des identifiants d'enzymes (numéro EC) ou des fonctions issues d'ontologies de gènes (Gene ontology). Le problème de quantification prend en entrée l'ensemble de séquences et la liste des fonctions renvoyée par le premier problème et renvoie une valeur numérique associée à chaque fonction, représentant la quantité absolue (nombre de lectures par exemple) ou relative (pourcentage de lectures) de chaque fonction dans l'échantillon. Néanmoins ces deux tâches sont rendues difficile par la complexité des données métagénomiques et le caractère non modèle de la plupart des organismes étudiés.

1.3.1. *Prédiction et annotation de gènes*

Une fois les génomes de la communauté bactérienne assemblés, la prédiction de séquences codantes sur ces génomes peut être effectuée de manière assez similaire à ce qui est fait en génomique classique, soit par comparaison à des bases de données protéiques, ou bien *ab initio*, ce qui permet de détecter de nouveaux gènes. Dans le premier cas, les séquences nucléotidiques sont converties en séquences protéiques suivant les 6 phases de lectures possibles, puis comparées à des bases de données de protéines connues. Ce type d'analyse est par exemple permis par l'outil BlastX (Altschul et al. 1990*b*). Étant donné l'incomplétude des bases de données de référence, les méthodes *ab initio* sont privilégiées en métagénomique. Différents outils ont été développés expressément pour cet usage, tels que MetaGeneMark (Zhu et al. 2010) ou Orphelia (Hoff et al. 2009). Les difficultés liées à la métagénomique sont essentiellement dues à la longueur généralement plus courte des contigs par rapport aux assemblages d'une seule espèce isolée. D'autre part, la plupart des techniques reposent sur des modèles (HMM, machine learning) entrainés à partir de génomes connus.

L'étape suivante consiste à assigner une fonction aux gènes détectés, qu'elle soit métabolique, structurale ou régulatrice. Le processus est analogue à celui réalisé en génomique classique, avec un nombre de gènes généralement très supérieur. Il s'agit de comparer les séquences protéiques obtenues à l'étape précédente avec des bases de données protéiques. Celles-ci sont nombreuses, on peut citer KEGG (Kanehisa 2004), PFAM (Finn et al. 2014) ou Uniprot (Bateman et al. 2017). Ces différentes bases de données ne couvrent pas toutes les fonctions connues, et des outils tels qu'InterPro (McDowall & Hunter 2011) ou MG-RAST (Keegan et al. 2016*b*) permettent d'en exploiter plusieurs.

1.3.2. *Métatranscriptomique*

Afin de répondre aux interrogations sur les fonctions assurées par une commuanuté microbienne, la méta-transcriptomique complète efficacement la métagénomique. En effet, la présence d'un gène ne garantit pas son expression. La métatranscriptomique permet d'obtenir un aperçu de l'expression des gènes dans un échantillon à un moment et dans des conditions données.

Les données métatranscriptomiques sont généralement dominées par les ARNs ribosomiques. S'il peuvent être utilisés pour comparer l'activité de différents taxons, ils sont néanmoins peu informatifs sur les fonctions assurées par ces organismes. Toutefois, différents protocoles permettent de filtrer ces séquences avant leur séquençage (Sultan et al. 2014).

Tout comme en transcriptomique conventionnelle, deux approches sont possibles pour analyser de telles données : l'alignement des lectures sur des génomes de référence connus, où l'assemblage *de novo* de métatranscriptomes. Néanmoins, en métatranscriptomique, ces approches sont limitées soit par le caractère incomplet des bases de données existantes, soit par la difficulté à assembler correctement les lectures en raison d'importantes variations de couverture et du risque de chimérisme.

Dans les deux cas, relativement peu d'outils dédiés à la métatranscriptomique ont été développés (Aguiar-Pulido et al. 2016), mais des pipelines dédiés existent, tels MG-Rast (Keegan et al. 2016*a*) ou HUMAnM (Abubucker et al. 2012)). Les méthodes *de novo* reposent sur des assembleurs transcriptomiques tels que *Trinity* (Celaj et al. 2014), et utilisent par exemple RSEM (Li & Dewey 2011) pour la quantification de l'expression des différents transcrits. On peut également citer des assembleurs conçus pour l'assemblage métatranscriptomique, tels que IDBA-MT (Leung et al. 2013), IDBA-MTP (Leung et al. 2014) ou TAG (Ye & Tang 2016), qui prennent en compte les spécificités de la métatranscriptomique et permettent d'assembler moins de contigs chimériques. Alternativement, les approches basées sur des bases de données de référence utilisent également le plus souvent des outils classiques tels que Bowtie (Langmead & Salzberg 2012) ou Blast (Altschul et al. 1990*a*) pour aligner les lectures.

1.3.3. *Reconstruction de réseaux métaboliques*

L'une des finalités de l'annotation des gènes est la reconstruction de réseaux métaboliques. Si les réseaux peuvent être reconstruits indépendamment pour chaque espèce de la communauté assemblée et annotée, une perspective intéressante offerte par la métagénomique est de reconstruire le réseau de la communauté microbienne dans son ensemble. L'outil MetaPath (Liu & Pop 2010) aligne sur un méta-réseau générique des lectures métagénomiques afin d'identifier quelles en sont les composantes présentes dans un jeu de données. Il est à noter que cette méthode se

base sur l'alignement de lectures sur un réseau métabolique, et ne nécessite donc pas d'employer les techniques d'assemblage et de *binning* présentées précédemment. L'analyse fonctionnelle de séquençages métagénomiques peut par exemple permettre l'identification de nouvelles voies métaboliques, qui peuvent être utiles à l'hôte de communautés symbiotiques (Cecchini et al. 2013). Dans le cas d'échantillons environnementaux, il est possible d'identifier les composés pour lesquels un organisme dépend de son environnement (Borenstein et al. 2008). En comparant les topologies des réseaux métaboliques de plusieurs espèces vivant dans le même environnement, il est possible d'identifier des coopérations (Levy et al. 2015) ou des compétitions (Kreimer et al. 2012) au sein d'une communauté.

Les données métagénomiques offrent ainsi la possibilité d'accéder au fonctionnement de communautés complexes. Cette étape intervient cependant après de nombreuses autres analyses, et peut souffrir d'erreurs survenues au cours de l'assemblage, de l'assignation taxonomique ou de l'annotation, *a fortiori* dans le cas d'organismes peu connus.

1.4. Métagénomique comparative

La métagénomique comparative consiste à comparer des échantillons métagénomiques d'un point de vue génomique. On cherche à estimer des mesures de similarité (ou dissimilarité) entre échantillons ou communautés, pris dans leur ensemble. Le résultat est une matrice de taille $N \times N$ où N est le nombre d'échantillons à comparer. Chaque cellule de la matrice indique la similarité (ou inversement la dissimilarité) entre une paire d'échantillons.

Les ensembles d'échantillons comparés peuvent être des séries temporelles ou des échantillons collectés à des localisations géographiques différentes ou dans des conditions différentes. Puisqu'un échantillon métagénomique correspond à l'image d'une communauté à un instant, une localisation et dans des conditions donnés, les projets métagénomiques sont généralement constitués de plusieurs échantillons. C'est la comparaison des échantillons, associés à leur méta-données, qui permet ensuite d'extraire des connaissances. Parmi les grands projets de métagénomique qui possèdent cette dimension comparative, on peut citer le projet Human Microbiome Project (Lloyd-Price et al. 2017) (HMP) dont l'ambition est d'explorer le microbiome humain dans divers tissus et parmis différents individus sains (plus de 690 échantillons), ou bien le projet TARA oceans (Bork et al. 2015*b*) qui a récolté environ 2000 échantillons d'eau de mer à différentes profondeurs et pour différentes tailles d'organismes dans une centaines de localisations géographiques autour du globe.

Il existe deux types de méthodes de métagénomique comparative : celles qui se basent sur la diversité préalablement identifiée dans chaque échantillon (comme par exemple la composition taxonomique obtenue grâce aux approches présentées

précédemment), et celles qui ne se basent sur aucun pré-traitement des échantillons, ni n'utilisent de connaissances *a priori* et comparent directement toutes les séquences des échantillons. Ces dernières méthodes sont regroupées sous le terme de métagénomique comparative *de novo*.

1.4.1. *Métagénomique comparative avec estimation de la diversité*

Un premier ensemble des méthodes de métagénomique comparative s'inscrit dans la continuité des approches décrites précédemment de caractérisation des échantillons, soit en termes de taxonomie, soit en termes de fonctions. Les échantillons sont ainsi chacun représentés par un ensemble d'éléments (taxons, OTUs, bins, gènes, ou fonctions). Pour une paire d'échantillons donnée, la similarité dépend des nombres d'éléments qu'ils ont en commun et qui leurs sont spécifiques. Il existe de nombreux indices de similarité utilisés en écologie qui combinent ces nombres différemment, et qu'on peut regrouper principalement dans deux familles : les indices qualitatifs et les indices quantitatifs (voir (Legendre & Cáceres 2013) pour une classification plus fine de ces indices). La première famille traite les éléments de manière égale qu'ils soient rares ou très abondants dans les échantillons et utilise seulement l'information de présence ou d'absence des éléments. Dans cette famille, l'indice le plus classique est la distance de Jaccard qui est le rapport entre le cardinal de l'intersection et le cardinal de l'union des ensembles comparés ($J(A,B) = \frac{|A \cap B|}{|A \cup B|}$). A l'inverse, les indices quantitatifs utilisent l'information d'abondance des éléments dans chaque échantillon. Si deux échantillons possèdent les mêmes éléments, ils pourront tout de même être différenciés si les éléments ne sont pas présents dans les mêmes proportions. Une conséquence est également que les éléments les plus abondants auront plus de poids dans le calcul de l'indice de similarité que les éléments rares. La dissimilarité de Bray-Curtis (Bray & Curtis 1957) est un des indices les plus populaires de cette catégorie. Elle est définie par la formule suivante, où N_{iA} est l'abondance de l'élément i dans l'échantillon A :

$$BC(A,B) = 1 - 2\frac{\sum_{i=1}^{p} min(N_{iA}, N_{iB})}{\sum_{i=1}^{p} N_{iA} + N_{iB}}$$

Ces indices sont implémentés dans certains logiciels d'analyse de données métagénomiques tels que MEGAN (Huson et al. 2016), qui proposent également des visualisations de données multidimensionnelles (par exemple par PCoA).

1.4.2. *Métagénomique comparative de novo*

Étant données les difficultés à établir un bon inventaire taxonomique ou fonctionnel de métagénomes, une alternative est de comparer les échantillons par

leurs séquences directement, sans tenter de les assigner dans des OTUs ou des fonctions. On parle alors de métagénomique comparative *de novo*. L'objectif de ces méthodes est d'estimer le contenu génomique partagé entre deux ensembles de lectures.

Une approche naïve consiste à comparer (par exemple par alignement de séquences) toutes les lectures du premier échantillon à toutes les lectures du second échantillon. Les paires de lectures ayant une forte similarité de séquence sont dites similaires et sont supposées provenir du même taxon. La similarité entre les deux échantillons peut alors être définie par leur pourcentage de lectures similaires. Plus formellement, deux ensembles de séquences sans aucune annotation sont donnés en entrée, et le problème consiste à renvoyer en sortie le nombre de séquences du premier ensemble qui sont similaires à au moins une séquence du second ensemble et vice-versa.

Cette approche naïve se heurte à deux problèmes de passage à l'échelle. Le premier vient de la comparaison de deux échantillons. Un échantillon plein-génome contient des centaines de millions de lectures qu'il faut comparer aux séquences de l'autre échantillon. Le deuxième vient du fait qu'il faut calculer $O(N^2)$ mesures de similarité entre toutes les paires des N échantillons considérés.

1.4.2.1. *Approches sans alignement (alignment-free)*

Pour traiter ce problème de passage à l'echelle, les approches de métagénomique comparative *de novo* ont remplacé les étapes d'alignement de séquences qui sont très coûteuses en temps de calcul par des comparaisons exactes de k-mers (mots de taille k). A l'inverse des lectures, les k-mers sont très rapides à traiter car leur comparaison se fait de manière exacte : deux k-mers sont identiques ou non. Les valeurs de k utilisées sont généralement grandes ($k > 21$), afin que la plupart des k-mers soient spécifiques à un génome et que leurs abondances dans l'échantillon soit proportionnelles à celle des génomes d'où les k-mers proviennent. Ces grandes tailles de k-mer posent alors des problèmes d'indexation et de stockage en mémoire, puisque l'espace des k-mers (4^k) croît de manière exponentielle avec k.

Les méthodes Compareads (Maillet et al. 2012) et Commet (Maillet et al. 2014) estiment pour chaque lecture d'un échantillon si elle est similaire à au moins une lecture de l'autre échantillon en calculant le nombre de k-mers partagés entre la lecture et l'échantillon pris dans son ensemble. Pour cela, l'ensemble des k-mers d'un échantillon est indexé dans une structure de données à faible empreinte mémoire, le filtre de Bloom (Bloom 1970). Cette structure de données est un simple tableau de bits, associé à une ou plusieurs fonctions de hachage, et qui permet de requêter très rapidement l'existence d'un élément dans un ensemble. Une particularité de cette structure est qu'elle est probabiliste : les collisions de la ou des fonctions de hachage n'étant pas gérées, des faux positifs peuvent exister dans l'étape de requêtage à un taux que l'on peut contrôler (voir le Chapitre **??**). Cette approche est de l'ordre de 30

fois plus rapide que Blast mais reste trop longue lorsque le nombre N d'échantillons est grand puisque les étapes d'indexation et de requêtes doivent être répétées entre N et N^2 fois.

1.4.2.2. *Comparaison de spectres de k-mers*

Les méthodes les plus efficaces actuellement vont encore plus loin et oublient la structure en lectures des jeux de données. Dans la méthode Simka (Benoit et al. 2016), chaque échantillon est vu comme un ensemble de k-mers. On peut alors calculer les mêmes indices de similarité vus dans la section précédente en remplaçant les espèces/OTUs/gènes par les différents k-mers avec leurs abondances dans les échantillons. La difficulté principale est alors le comptage et la représentation en mémoire de ces ensembles d'éléments beaucoup plus grands que dans le cas de compositions taxonomiques ou fonctionnelles. Par exemple, dans un seul échantillon classique de métagénomique plein génome, on peut observer plusieurs milliards de 21-mers distincts. A titre d'exemple, une matrice d'abondance, avec en lignes les k-mers distincts et en colonnes les 690 échantillons du projet HMP, nécessiterait plusieurs centaines de TeraOctets pour être stockée en mémoire. Pour répondre à ce défi de mémoire, la méthode Simka compte les k-mers en utilisant l'écriture sur disque (méthode inspirée de DSK (Rizk et al. 2013) et KMC2 (Deorowicz et al. 2015) qui utilisent la notion de *minimiseur* pour partitionner l'ensemble des k-mers, voir Chapitre **??**) et ne stocke jamais la matrice d'abondances en mémoire mais calcule les indices de similarité de manière itérative et parallèle, k-mer après k-mer. Ainsi, en quelques heures, les indices de similarité entre les 690 échantillons du projet HMP (32 milliards de lectures) peuvent être calculés et produisent des résultats de classification d'échantillons similaires aux méthodes basées sur l'assignation taxonomique (Benoit et al. 2016).

1.4.2.3. *Approches de sous-échantillonnage des k-mers*

Si Simka utilise l'ensemble des k-mers présents pour estimer les indices de similarité, d'autres approches se basent sur un sous-échantillonnage de l'espace des k-mers pour résoudre le problème de dimension des données. C'est le cas de MetaFast (Ulyantsev et al. 2016) qui effectue en amont de la comparaison une étape d'assemblage de novo assez grossière des lectures, afin de sélectionner les k-mers présents dans des composantes du graphe de de Bruijn respectant certains critères topologiques et d'abondance. Les composantes sélectionnées jouent alors le rôle de références et l'abondance de chaque composante est quantifiée dans chaque échantillon par l'intermédiaire des abondances des k-mers qui la composent. L'assemblage est un moyen de sélectionner et de compacter des k-mers afin de réduire leur dimension. Cette stratégie est intéressante car la sélection se base sur des critères qui ont un sens biologique : sélection de k-mers partagés par plusieurs échantillons, élimination des k-mers issus d'erreurs de séquençage ou provenant de régions très complexes à assembler. Cependant, l'inconvénient majeur de cette approche est son coût en ressources de calcul (temps et mémoire) qui provient

naturellement de l'étape d'assemblage puisque le graphe de De Bruijn doit être construit et représenté en mémoire et sa taille est de l'ordre du nombre de k-mers distincts. Par exemple, cette méthode ne passe pas à l'échelle sur les données du projet HMP.

Les approches les plus efficaces en métagénomique comparative sont celles basées sur un sous-échantillonnage "aléatoire" de l'espace des k-mers. C'est la méthode Mash (Ondov et al. 2016), développée principalement pour la comparaison de génomes, qui a introduit ce type de méthodes. Mash s'appuie sur la technique Minhash (Broder 1997) qui est une approche statistique permettant d'estimer l'indice de Jaccard entre deux ensembles en n'utilisant que quelques milliers de leurs éléments. Chaque échantillon à comparer est représenté par une liste triée de n k-mers, appelée *signature* ou *sketch* en anglais (avec n très petit par rapport à l'ensemble des k-mers présents). La sélection aléatoire et le tri des k-mers sont effectués grâce à une fonction de hachage uniforme qui pour un k-mer donné renvoie un entier sur 64 bits (avec une probabilité de collision proche de 0). Les n k-mers distincts de l'échantillon ayant les plus petites valeurs de hachage constituent sa signature. L'indice de Jaccard entre 2 échantillons est alors calculé en appliquant la formule de Jaccard sur les n premiers k-mers de l'union des deux signatures. En termes de performances, il semble peu probable de pouvoir être plus efficace que Mash. En effet, pour calculer la signature d'un ensemble de m k-mers, seule une liste triée de n éléments doit être stockée en mémoire, la complexité en temps est presque linéaire avec m (si $n \ll m$) et les signatures des différents échantillons peuvent être calculées en parallèle. Enfin, le calcul des distances est extrêmement rapide compte tenu de la taille des signatures ($n = 1000$ en général). Son point faible se situe au niveau des résultats fournis qui se limitent à une estimation de l'indice de Jaccard. Cette mesure ne prend en compte que la présence-absence des k-mers et non leur abondance. Si elle est bien adaptée pour comparer des génomes, elle l'est beaucoup moins pour comparer des échantillons métagénomiques ayant des profils de diversité en espèces très variables ou simplement des ensembles de lectures avec des erreurs de séquençage. La méthode SimkaMin a résolu ce problème en étendant l'approche de Mash à l'indice de Bray-Curtis et en permettant de filtrer les k-mers rares (Benoit et al. 2019). Les k-mers sont sélectionnés de la même manière avec une fonction de hachage, indépendemment de leur abondance, mais les abondances des k-mers sélectionnés sont stockées dans la signature et utilisées dans le calcul de l'indice de Bray-Curtis. SimkaMin est environ 10 fois plus rapide que Simka avec de très faibles empreintes mémoire et disque, sans impacter qualitativement les résultats des analyses en aval.

Ces approches à base de k-mers ont été développées pour les données de séquençage de seconde génération. Le faible taux d'erreurs de séquençage et la profondeur de séquençage souvent importante dans ces données permettent notamment d'utiliser des valeurs de k assez grandes, généralement entre 20 et 30. Ces approches sont ainsi peu adaptées aux données de séquençage de troisième

génération, car leur taux d'erreur plus important implique que la plupart des k-mers possèdent au moins une erreur de séquençage ce qui ne permet plus de les comparer directement sans édition entre plusieurs échantillons. Pour ce type de données plus bruitées, d'autres types de graines pourraient être envisagées comme les graines espacées.

1.5. Conclusion

La métagénomique est un outil prometteur pour l'étude des communautés microbiennes, d'une part car elle permet de s'affranchir du biais de cultivabilité et ainsi de décrire l'ensemble des organismes en interaction, et d'autre part car elle apporte une résolution inédite, qui permet de décrire à la fois la structure, la fonction et l'évolution de ces communautés. Pour les mêmes raisons, la métagénomique soulève également de nombreux défis méthodologiques du fait de la complexité de ces communautés et du faible degré de connaissance de leurs membres. Dans ce chapitre, nous avons présenté un ensemble de méthodes permettant de répondre aux principales questions adressées à la métagénomique.

Bien souvent, le choix de ces outils résulte d'un compromis entre complexité de la communauté, disponibilité de données de référence, et précision (souvent taxonomique) des résultats obtenus. L'état de l'art permet ainsi de décrire finement la structure et la fonction de communautés de référence, de caractériser les membres de communautés non modèles, ou d'appréhender plus grossièrement les systèmes les plus complexes. Progresser sur ces questions nécessite des avancées à la fois du côté de la bio-informatique et de la biologie.

Du point de vue de la bio-informatique, un sujet majeur de la métagénomique demeure de caractériser le plus finement possible le contenu des métagénomes. Se pose en particulier le problème de la résolution taxonomique atteinte par les outils actuels d'assemblage et de *binning*. Les génomes des organismes vivant dans le même environnement interagissent entre eux, et un métagénome résulte d'un *continuum* de diversité, qui rend l'assemblage difficile. En particulier, un des défis récents de la métagénomique est la description plus fine des génomes, à travers l'assemblage à des niveaux taxonomiques plus fins (Segata 2018). Parallèlement, l'émergence du séquençage des longues lectures promet de faciliter l'assemblage métagénomique (Kolmogorov et al. 2019), mais aussi l'assignation taxonomique (Dilthey et al. 2019).

Les questions biologiques posées par la métagénomique varient grandement en fonction des communautés étudiées. Cependant, étant donné l'augmentation de la qualité et quantité des données et le développement des méthodes pour les traiter, les études métagénomiques permettent d'apporter des réponses de plus en plus fines. L'étude du microbiome humain est le fer de lance de la recherche en métagénomique, et illustre parfaitement l'évolution de la discipline. Tandis que sa première phase

visait à cataloguer les espèces présentes dans les différents microbiotes humains (Turnbaugh, Ley, Hamady, Fraser-Liggett, Knight & Gordon 2007), la seconde, achevée en 2019 (iHMP Research Network Consortium et al. 2019), proposait une approche intégrative et longitudinale, permettant d'observer l'évolutions de profils -omiques sous différentes conditions. Plus qu'un simple assemblage d'espèces, il s'agit donc de considérer un microbiote comme une entité complexe, en constante évolution, tout en considérant les liens fonctionnels entre ses constituants. Ainsi, l'étude de ces systèmes pourrait bénéficier d'apports de la biologie évolutive, mettant en valeur les déterminants de l'évolution de ces communautés (Proctor 2019).

Bibliographie

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B. et al. (2012), 'Metabolic reconstruction for metagenomic data and its application to the human microbiome', *PLoS computational biology* **8**(6), e1002358.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K. & Narasimhan, G. (2016), 'Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis : supplementary issue : bioinformatics methods and applications for big metagenomics data', *Evolutionary Bioinformatics* **12**, EBO–S36436.

Ahn, T. H., Chai, J. & Pan, C. (2015), 'Sigma : Strain-level inference of genomes from metagenomic analysis for biosurveillance', *Bioinformatics* **31**(2), 170–177.
URL: *http ://bioinformatics.oxfordjournals.org/content/early/2014/10/22/bioinformatics.btu641.full*

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F. & Quince, C. (2014), 'Binning metagenomic contigs by coverage and composition', *Nature Methods* **11**(11), 1144–1146.
URL: *http ://www.nature.com/nmeth/journal/v11/n11/abs/nmeth.3103.html*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990*a*), 'Basic local alignment search tool', *Journal of molecular biology* **215**(3), 403–410.

Altschul, S. F., Gish, W., Miller, W., Myers, W. E. & Lipman, D. J. (1990*b*), 'Basic local alignment search tool', *Journal of Molecular Biology* **215**, 402–410.
URL: *https ://www.sciencedirect.com/science/article/pii/S0022283605803602*

Angly, F. E., Willner, D., Prieto-Davó, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D. A., Barott, K., Cottrell, M. T., Desnues, C., Dinsdale, E. A., Furlan, M., Haynes, M., Henn, M. R., Hu, Y., Kirchman, D. L., McDole, T., McPherson, J. D., Meyer, F., Miller, R. M., Mundt, E., Naviaux, R. K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B. & Rohwer, F. (2009), 'The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes', *PLoS Computational*

Biology **5**(12), e1000593.
URL: *http ://journals.plos.org/ploscompbiol/article ?id=10.1371/journal.pcbi.1000593*

Ayling, M., Clark, M. D. & Leggett, R. M. (2020), 'New approaches for metagenome assembly with short reads', *Briefings in Bioinformatics* **21**(2), 584–594.

Bar-On, Y. M., Phillips, R. & Milo, R. (2018), 'The biomass distribution on earth', *Proceedings of the National Academy of Sciences* **115**(25), 6506–6511.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S. & Zhang, J. (2017), 'UniProt : The universal protein knowledgebase', *Nucleic Acids Research* **45**(D1), D158–D169.
URL: *https ://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1099*

Benoit, G., Mariadassou, M., Robin, S., Schbath, S., Peterlongo, P. & Lemaitre, C. (2019), 'SimkaMin : fast and resource frugal de novo comparative metagenomics', *Bioinformatics* .

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D. & Lemaitre, C. (2016), 'Multiple Comparative Metagenomics using Multiset k-mer Counting', *PeerJ Computer Science* **2**, e94.
URL: *http ://arxiv.org/abs/1604.02412*

Bloom, B. H. (1970), 'Space/time trade-offs in hash coding with allowable errors', *Communications of the ACM* **13**(7), 422–426.

Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. (2008), 'Large-scale reconstruction and phylogenetic analysis of metabolic environments', *Proceedings of the National Academy of Sciences* **105**(38), 14482–14487.
URL: *http ://www.pnas.org/cgi/doi/10.1073/pnas.0806162105*

Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E. & Wincker, P. (2015*a*), 'Tara Oceans studies plankton at Planetary scale', *Science* **348**(6237), 873.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/25999501*

Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E. & Wincker, P. (2015*b*), 'Tara oceans studies plankton at planetary scale'.

Bray, J. R. & Curtis, J. T. (1957), 'An ordination of the upland forest communities of southern wisconsin', *Ecological Monographs* **27**(4), 325–349.

Broder, A. Z. (1997), On the resemblance and containment of documents, *in* 'Compression and Complexity of Sequences 1997. Proceedings', IEEE, pp. 21–29.

Buchfink, B., Xie, C. & Huson, D. H. (2014), 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods* **12**(1), 59–60.
URL: *http ://www.nature.com/articles/nmeth.3176*

Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. (2014), 'Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps', *G3: Genes|Genomes|Genetics* **4**(7), 1339–1346.
URL: *http ://g3journal.org/lookup/doi/10.1534/g3.114.011825*

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010), 'QIIME allows analysis of high-throughput community sequencing data.', *Nature methods* **7**(5), 335–6.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/20383131 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC3156573*

Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J., Henrissat, B., Remaud-Siméon, M., Monsan, P. & Potocki-Véronèse, G. (2013), 'Functional Metagenomics Reveals Novel Pathways of Prebiotic Breakdown by Human Gut Bacteria', *PLoS ONE* **8**(9), e72766.
URL: *http ://dx.plos.org/10.1371/journal.pone.0072766*

Celaj, A., Markle, J., Danska, J. & Parkinson, J. (2014), 'Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation', *Microbiome* **2**(1), 39.

Chen, I., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J. & Huntemann, M. (2019), 'Varghese 609 n, white jr, seshadri r et al : Img/m v. 5.0 : an integrated data management and 610 comparative analysis system for microbial genomes and microbiomes', *Nucleic 611 Acids Res* **47**(D1), D666–D677.

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S. & Alm, E. J. (2015), 'Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning', *Nature Biotechnology* **33**(10), 1053–1060.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/26368049\nhttp ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=I*

De Filippis, F., Parente, E. & Ercolini, D. (2017), 'Metagenomics insights into food fermentations', *Microbial biotechnology* **10**(1), 91–102.

Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R. & Vogel, T. M. (2011), 'Accessing the soil metagenome for studies of microbial diversity', *Applied and Environmental Microbiology* **77**(4), 1315–1324.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/21183646 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC3067229*

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S. & Sutton, G. (2008), 'Consensus generation and variant detection by Celera Assembler', *Bioinformatics* **24**(8), 1035–1040.

Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. (2015), 'Kmc 2 : fast and resource-frugal k-mer counting.', *Bioinformatics* **31**(10), 1569–1576.
URL: *http ://dx.doi.org/10.1093/bioinformatics/btv022*

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. (2006), 'Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB', *Applied and Environmental Microbiology* **72**(7), 5069–5072.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/16820507 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC1489311*

Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. (2019), 'Strain-level metagenomic assignment and compositional estimation for long reads with metamaps', *Nature communications* **10**(1), 1–12.

Dykhuizen, D. (2005), 'Species Numbers in Bacteria.', *Proceedings. California Academy of Sciences* **56**(6 Suppl 1), 62–71.
URL: *http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=3160642&tool=pmcentrez&rendertype=abstract*

Esposito, A. & Kirschberg, M. (2014), 'How many 16S-based studies should be included in a metagenomic conference ? It may be a matter of etymology', *FEMS Microbiology Letters* **351**(2), 145–146.
URL: *https ://academic.oup.com/femsle/article-lookup/doi/10.1111/1574-6968.12375*

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J. & Punta, M. (2014), 'Pfam : The protein families database'.

Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. A. & Johnson, W. E. (2013), 'Pathoscope : Species identification and strain attribution with unassembled sequencing data', *Genome Research* **23**(10), 1721–1729.
URL: *http ://genome.cshlp.org/content/23/10/1721.abstract*

Garmendia, L., Hernandez, A., Sanchez, M. & Martinez, J. (2012), 'Metagenomics and antibiotics', *Clinical Microbiology and Infection* **18**, 27–31.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013), 'QUAST : Quality assessment tool for genome assemblies', *Bioinformatics* **29**(8), 1072–1075.
URL: *https ://academic.oup.com/bioinformatics/article/29/8/1072/228832*

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998), 'Molecular biological access to the chemistry of unknown soil microbes : A new frontier for natural products', *Chemistry and Biology* **5**(10), R245–R249.
URL: *https ://www.sciencedirect.com/science/article/pii/S1074552198901089*

Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. (2009), 'Orphelia : Predicting genes in metagenomic sequencing reads', *Nucleic Acids Research* **37**(SUPPL. 2).

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J. & Tappu, R. (2016), 'MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data', *PLoS Computational Biology* **12**(6), e1004957.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/27327495 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC4915700*

iHMP Research Network Consortium, I. H. et al. (2019), 'The integrative human microbiome project', *Nature* **569**, 641–648.

Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P. & Tyson, G. W. (2014), 'GroopM : an automated tool for the recovery of population genomes from related metagenomes', *PeerJ* **2**, e603.
URL: *https ://peerj.com/articles/603*

Kanehisa, M. (2004), 'The KEGG resource for deciphering the genome', *Nucleic Acids Research* **32**(90001), 277D–280.
URL: *https ://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh063*

Keegan, K. P., Glass, E. M. & Meyer, F. (2016*a*), Mg-rast, a metagenomics service for analysis of microbial community structure and function, *in* 'Microbial Environmental Genomics (MEG)', Springer, pp. 207–233.

Keegan, K. P., Glass, E. M. & Meyer, F. (2016*b*), MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function, *in* 'Methods in molecular biology (Clifton, N.J.)', Vol. 1399, pp. 207–233.
URL: *http ://link.springer.com/10.1007/978-1-4939-3369-3_13*

Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. (2016), 'Centrifuge : Rapid and sensitive classification of metagenomic sequences', *Genome Research* **26**(12), 1721–1729.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/27852649 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC5131823*

Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E. & Pevzner, P. (2019), 'metaflye : scalable long-read metagenome assembly using repeat graphs', *bioRxiv* p. 637637.

Kreimer, A., Doron-Faigenboim, A., Borenstein, E. & Freilich, S. (2012), 'NetCmpt : A network-based tool for calculating the metabolic competition between bacterial species', *Bioinformatics* **28**(16), 2195–2197.

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., Coronado, S., der Maaten, L. V., Vlassis, N. & Wilmes, P. (2015), 'VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data', *Microbiome* **3**(1).

Langmead, B. & Salzberg, S. L. (2012), 'Fast gapped-read alignment with bowtie 2', *Nature methods* **9**(4), 357.

Legendre, P. & Cáceres, M. D. (2013), 'Beta diversity as the variance of community data : dissimilarity coefficients and partitioning', *Ecology Letters* **16**(8), 951–963.

Leung, H. C., Yiu, S.-M. & Chin, F. Y. (2014), Idba-mtp : a hybrid metatranscriptomic assembler based on protein information, *in* 'International Conference on Research in Computational Molecular Biology', Springer, pp. 160–172.

Leung, H. C., Yiu, S.-M., Parkinson, J. & Chin, F. Y. (2013), 'Idba-mt : de novo assembler for metatranscriptomic data generated from next-generation sequencing technology', *Journal of Computational Biology* **20**(7), 540–550.

Levy, R., Carr, R., Kreimer, A., Freilich, S. & Borenstein, E. (2015), 'NetCooperate : A network-based tool for inferring host-microbe and microbe-microbe cooperation', *BMC Bioinformatics* **16**(1).

Li, B. & Dewey, C. N. (2011), 'Rsem : accurate transcript quantification from rna-seq data with or without a reference genome', *BMC bioinformatics* **12**(1), 323.

Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. (2015), 'MEGAHIT : An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics* **31**(10), 1674–1676.

URL: *https ://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv033*

Liu, B. & Pop, M. (2010), 'Identifying differentially abundant metabolic pathways in metagenomic datasets', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6053 LNBI**(Suppl 2), 101–112.
URL: *http ://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-5-S2-S9*

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G. et al. (2017), 'Strains, functions and dynamics in the expanded human microbiome project', *Nature* **550**(7674), 61.

Locey, K. J. & Lennon, J. T. (2016), 'Scaling laws predict global microbial diversity', *Proceedings of the National Academy of Sciences* **113**(21), 5970–5975.

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J. & Gevers, D. (2015), 'ConStrains identifies microbial strains in metagenomic datasets', *Nature Biotechnology* **33**(10), 1045–1052.

Maillet, N., Collet, G., Vannier, T., Lavenier, D. & Peterlongo, P. (2014), Commet : Comparing and combining multiple metagenomic datasets, *in* 'Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014', IEEE, pp. 94–98.
URL: *http ://ieeexplore.ieee.org/document/6999135/*

Maillet, N., Lemaitre, C., Chikhi, R., Lavenier, D. & Peterlongo, P. (2012), 'Compareads : comparing huge metagenomic experiments', *BMC Bioinformatics* **13**(Suppl 19), S10.
URL: *http ://www.biomedcentral.com/1471-2105/13/S19/S10/abstract*

McDowall, J. & Hunter, S. (2011), 'InterPro protein classification.', *Methods in molecular biology (Clifton, N.J.)* **694**, 37–47.
URL: *http ://link.springer.com/10.1007/978-1-60761-977-2\nhttp ://www.ncbi.nlm.nih.gov/pubmed/21082426*

Menzel, P., Ng, K. L. & Krogh, A. (2016), 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', *Nature Communications* **7**.

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J. et al. (2019), 'Mgnify : the microbiome analysis resource in 2020', *Nucleic Acids Research* .

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. (2012), 'MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly from short sequence reads', *Nucleic Acids Research* **40**(20), e155.
URL: *http ://nar.oxfordjournals.org/content/40/20/e155*

Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. (2017), 'metaspades : a new versatile metagenomic assembler', *Genome research* **27**(5), 824–834.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. (2016), 'Mash : Fast genome and metagenome distance estimation using MinHash', *Genome Biology* **17**(1), 132.
URL: *http ://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x*

Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. (2015), 'CLARK : fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers', *BMC Genomics* **16**(1), 236.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/25879410 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC4428112*

Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. (1986), The analysis of natural microbial populations by ribosomal rna sequences, *in* 'Advances in microbial ecology', Springer, pp. 1–55.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. (2015), 'CheckM : Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research* **25**(7), 1043–1055.

Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. (2017), 'Genome graphs and the evolution of genome inference', *Genome Research* **27**(5), 665–676.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/28360232 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC5411762*

Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. (2012), 'IDBA-UD : A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth', *Bioinformatics* **28**(11), 1420–1428.
URL: *http ://bioinformatics.oxfordjournals.org/content/28/11/1420.full*

Proctor, L. (2019), 'Priorities for the next 10 years of human microbiome research'.

Pruesse, E., Peplies, J. & Glöckner, F. O. (2012), 'SINA : Accurate high-throughput multiple sequence alignment of ribosomal RNA genes', *Bioinformatics* **28**(14), 1823–1829.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/22556368 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC3389763 https ://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts252*

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. (2013), 'The SILVA ribosomal RNA gene database project : Improved data processing and web-based tools', *Nucleic Acids Research*

41(D1), D590–D596.
URL: *http ://academic.oup.com/nar/article/41/D1/D590/1069277/The-SILVA-ribosomal-RNA-gene-database-project*

Rappé, M. S. & Giovannoni, S. J. (2003), 'The Uncultured Microbial Majority', *Annual Review of Microbiology* **57**(1), 369–394.
URL: *http ://www.annualreviews.org/doi/10.1146/annurev.micro.57.030502.090759*

Rizk, G., Lavenier, D. & Chikhi, R. (2013), 'Dsk : k-mer counting with very low memory usage.', *Bioinformatics* **29**(5), 652–653.
URL: *http ://dx.doi.org/10.1093/bioinformatics/btt020*

Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., Daroub, S. H., Camargo, F. A., Farmerie, W. G. & Triplett, E. W. (2007), 'Pyrosequencing enumerates and contrasts soil microbial diversity', *ISME Journal* **1**(4), 283–290.
URL: *http ://www.nature.com/articles/ismej200753*

Sankar, S. A., Lagier, J. C., Pontarotti, P., Raoult, D. & Fournier, P. E. (2015), 'The human gut microbiome, a taxonomic conundrum', *Systematic and Applied Microbiology* **38**(4), 276–286.
URL: *https ://www.sciencedirect.com/science/article/pii/S0723202015000454 ?via%3Dihub*

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. & Weber, C. F. (2009), 'Introducing mothur : Open-source, platform-independent, community-supported software for describing and comparing microbial communities', *Applied and Environmental Microbiology* **75**(23), 7537–7541.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/19801464 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC2786419*

Schouls, L. M., Schot, C. S. & Jacobs, J. A. (2003), 'Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group', *Journal of Bacteriology* **185**(24), 7241–7246.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/14645285*

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., Demaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiutė, M., Hansen, L. H., Sørensen, S. J., Chia, B. K., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y. W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H. H., Liao, Y. C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard,

B. Y., Pop, M., Klenk, H. P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T. & McHardy, A. C. (2017), 'Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software', *Nature Methods* **14**(11), 1063–1071.
URL: *http ://www.nature.com/doifinder/10.1038/nmeth.4458*

Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. (2018), 'Piercing the dark matter : Bioinformatics of long-range sequencing and mapping', *Nature Reviews Genetics* **19**(6), 329–346.
URL: *http ://www.nature.com/articles/s41576-018-0003-4*

Segata, N. (2018), 'On the Road to Strain-Resolved Comparative Metagenomics', *mSystems* **3**(2), e00190–17.
URL: *http ://msystems.asm.org/lookup/doi/10.1128/mSystems.00190-17*

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. (2007), 'Camera : a community resource for metagenomics', *PLoS biology* **5**(3), e75.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, I. (2009), 'ABySS : A parallel assembler for short read sequence data', *Genome Research* **19**(6), 1117–1123.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H. & Yaspo, M.-L. (2014), 'Influence of rna extraction methods and library selection schemes on rna-seq data', *BMC genomics* **15**(1), 675.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004), 'Application of tetranucleotide frequencies for the assignment of genomic fragments', *Environmental Microbiology* **6**(9), 938–947.
URL: *http ://doi.wiley.com/10.1111/j.1462-2920.2004.00624.x*

Tiwari, R., Nain, L., Labrou, N. E. & Shukla, P. (2018), 'Bioprospecting of functional cellulases from metagenome for second generation biofuel production : a review', *Critical reviews in microbiology* **44**(2), 244–257.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. & Segata, N. (2015), 'MetaPhlAn2 for enhanced metagenomic taxonomic profiling'.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. (2017), 'Microbial strain-level population structure & genetic diversity from metagenomes', *Genome Research* **27**(4), 626–638.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/28167665*

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. (2007), 'The human microbiome project', *Nature* **449**(7164), 804.

Turnbaugh, P. J., Ruth, E., Ley, M. H., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. (2007), 'The human microbiome project', *Nature* **449**.
URL: *http ://dx.doi.org/10.1038/nature06244*

Ulyantsev, V. I., Kazakov, S. V., Dubinkina, V. B., Tyakht, A. V. & Alexeev, D. G. (2016), 'Metafast : fast reference-free graph-based comparison of shotgun metagenomic data.', *Bioinformatics* .
URL: *http ://dx.doi.org/10.1093/bioinformatics/btw312*

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. & Smith, H. O. (2004), 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science* **304**(5667), 66–74.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/15001713*

Vollmers, J., Wiegand, S. & Kaster, A.-K. (2017), 'Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-not only size matters !', *PloS one* **12**(1), e0169662.

Wang, Z. & Wu, M. (2013), 'A phylum-level bacterial phylogenetic marker database', *Molecular Biology and Evolution* **30**(6), 1258–1262.
URL: *https ://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst059*

Welch, D. B. & Huse, S. M. (2011), 'Microbial Diversity in the Deep Sea and the Underexplored "Rare Biosphere"', *Handbook of Molecular Microbial Ecology II : Metagenomics in Different Habitats* **103**(32), 243–252.
URL: *http ://www.ncbi.nlm.nih.gov/pubmed/16880384 http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=PMC1524930*

Wood, D. E. & Salzberg, S. L. (2014), 'Kraken : ultrafast metagenomic sequence classification using exact alignments.', *Genome biology* **15**(3), R46.
URL: *http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?artid=4053813&tool=pmcentrez&rendertype=abstract*

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. (2014), 'MaxBin : an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.', *Microbiome* **2**(1), 26.
URL: *http ://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-2-26*

Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A. & Sun, F. (2011), 'Accurate genome relative abundance estimation based on shotgun metagenomic reads.', *PloS one* **6**(12), e27992.
URL: *http ://journals.plos.org/plosone/article ?id=10.1371/journal.pone.0027992*

Ye, Y. & Tang, H. (2016), ‘Utilizing de bruijn graph of metagenome assembly for metatranscriptome analysis’, *Bioinformatics* **32**(7), 1001–1008.

Zhu, W., Lomsadze, A. & Borodovsky, M. (2010), ‘Ab initio gene identification in metagenomic sequences’, *Nucleic Acids Research* **38**(12).