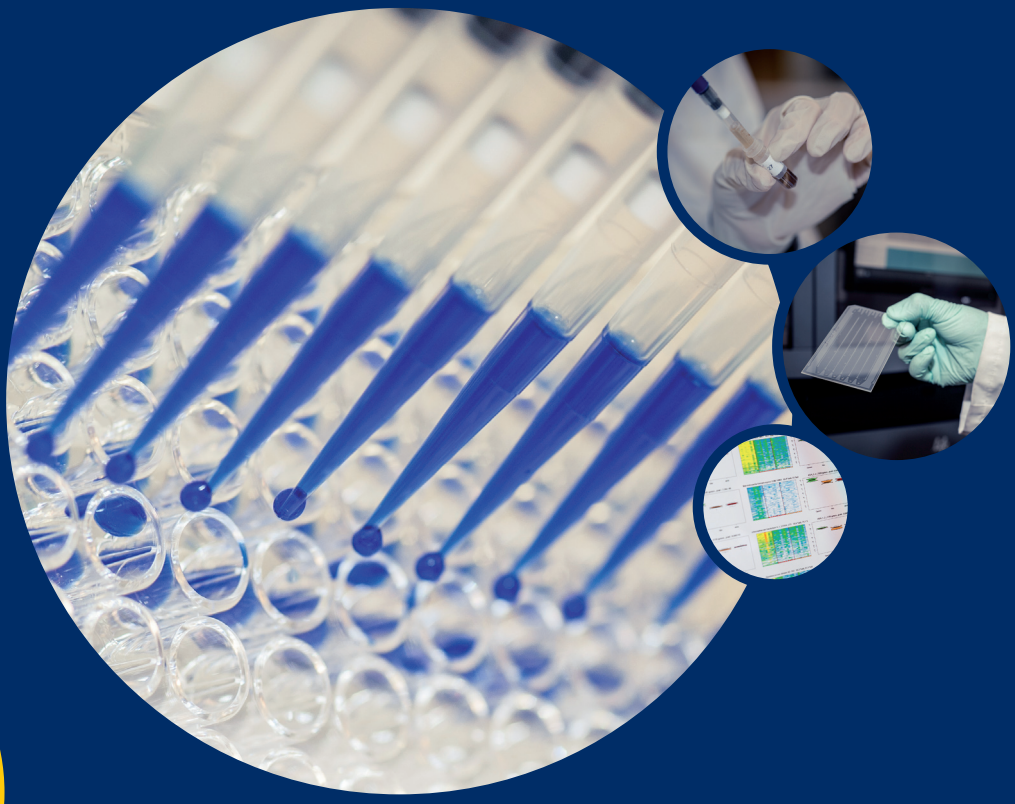
Savoir faire

La métagénomique

Développements et futures applications

M.-C. Champomier-Vergès, M. Zagorec, coord.



éditions Quæ

La métagénomique

Développements et futures applications

Marie-Christine Champomier-Vergès,
Monique Zagorec, coord.

Éditions Quæ

Collection *Savoir-faire*

Faut-il travailler le sol ?
Acquis et innovations pour une agriculture durable
F. Laurent, J. Roger-Estrade, J. Labreuche
2014, 192 p.

Les clémentiniers et autres petits agrumes
C. Jacquemond, F. Curk, M. Heuzet
2014, 368 p.

Torrents et rivières de montagne
Dynamique et aménagement
A. Recking, D. Richard, G. Degoutte, coord.
2013, 352 p.

Qualité du cacao
L'impact du traitement post-récolte
M. Barel
2013, 104 p.

Analyse de sensibilité et exploration de modèles
Application aux sciences de la nature et de l'environnement
R. Faivre, B. Looss, S. Mahévas, D. Makowski, H. Monod, éd.
2013, 352 p.

De la domestication à la transgénèse
Évolution des outils pour l'amélioration des plantes
A. Gallais
2013, 176 p.

Éditions Quæ
RD 10, 78026 Versailles Cedex, France

 ISBN 978-2-7592-2294-0 ISSN 1952-1251

Sommaire

Avant-propos

Le terme de métagénomique a été introduit pour la première fois en 1998 par Handelsman et ses collaborateurs[1], même si des études que l'on qualifierait aujourd'hui d'études métagénomiques avaient été publiées dès 1996 par Stein et ses collaborateurs[2]. Ce terme associe le mot « méta » (du grec, signifiant transcendant) au mot génomique, la discipline qui analyse un organisme (un ensemble, une entité vivante) au niveau de son génome. Le génome quant à lui est l'ensemble du matériel génétique, donc des gènes, d'un organisme. La métagénomique est bien effectivement une analyse qui transcende — se situe au-dessus de — l'analyse d'un ensemble de gènes. C'est surtout dans le domaine de la microbiologie que se distinguent de nombreux travaux faisant appel à cette nouvelle discipline. À ce jour, presque tous les éléments présents sur Terre sont étudiés par métagénomique. L'environnement, les sols, les nuages, l'air, l'eau des lacs et des océans, ainsi que les communautés microbiennes associées au monde végétal (la rhizosphère, la phylosphère) et animal (les microbiotes hébergés par les animaux) font l'objet d'études métagénomiques. La communauté scientifique des microbiologistes s'est emparée de cette discipline. Ainsi le sol et l'eau (les océans) ont été parmi les premiers éléments à être investigués par métagénomique. En effet, les limites des méthodes classiques de microbiologie pour décrire les communautés bactériennes extrêmement riches de ces domaines étaient connues. Pour exemple, on estimait que moins de 1 % des bactéries du sol étaient cultivables en milieu de laboratoire. Les microbiotes associés au règne animal et en particulier ceux portés par l'homme ont également bénéficié de la métagénomique car plus de 70 % des bactéries hébergées dans le tractus digestif humain n'étaient pas non plus cultivables avec les milieux et conditions de laboratoire disponibles. L'analyse par métagénomique d'autres écosystèmes, tels que ceux des aliments dont les communautés microbiennes étaient mieux connues, a été plus tardive. Elle a cependant révélé des éléments nouveaux, notamment la présence d'espèces déjà connues mais que l'on considérait comme

1. 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. *Chem. Biol.*, 5, R245-R249.
2. Stein J.L., Marsh T.L., Wu K.Y., Shizuya H., DeLong E.F., 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-Kilobase-pair genome fragment from a planktonic marine Archaeon. *J. Bacteriol.*, 178 (3), 591-599.

des espèces environnementales et qui n'avaient pas été recherchées dans les aliments. De fait, les aliments véhiculent des communautés microbiennes provenant de différents environnements (sols, eaux, végétaux, microbiotes animaux, etc.). Ces communautés microbiennes des aliments, une fois ingérées, peuvent transiter voire coloniser durablement le tractus digestif humain après consommation. Ainsi, les frontières entre les communautés microbiennes des mondes vivants (animal, végétal) et minéral (sol, air, eau) commencent à tomber, nous faisant prendre conscience de l'omniprésence du monde bactérien dans notre quotidien.

La métagénomique permet une profondeur d'analyse qui n'était pas accessible auparavant. En effet, les techniques de séquençage ont atteint un tel débit que des espèces bactériennes minoritaires peuvent être détectées maintenant parmi des centaines de milliers ou des millions de lectures alors que les approches de microbiologie classique ou même de criblage ne le permettaient pas. Une quantification relative des différentes communautés bactériennes est également possible. Enfin, grâce au déluge de séquences génomiques bactériennes désormais disponibles, les données de métagénomique peuvent aboutir à l'identification des communautés microbiennes au niveau du genre et même de l'espèce.

Avec la même logique et des techniques similaires, la métatranscriptomique a vu le jour. Elle permet de connaître les gènes exprimés ainsi qu'une quantification relative de leur niveau d'expression. Ainsi, non seulement les communautés microbiennes mais également les fonctions qu'elles expriment potentiellement peuvent être connues.

La métagénomique est encore une discipline jeune. On lui reproche souvent de n'être que descriptive. Cependant, elle permet une description bien plus exhaustive qu'auparavant et des analyses approfondies allant au-delà de la description commencent à voir le jour.

1

Techniques et matériels pour la production des données brutes de séquençage métagénomique

Benoît Remenant, Stéphane Chaillou

Introduction

Au cours des dix dernières années, la métagénomique a permis des avancées considérables en écologie microbienne, notamment pour la mesure de la diversité et de l'évolution du monde microbien. De nombreux laboratoires de recherche sont désormais engagés dans ce champ thématique, entraînant le développement d'un vaste pan de connaissances méthologiques et d'expertises pour lesquelles il est préférable de posséder un guide pratique. En effet, il existe plusieurs façons d'analyser le métagénome d'un environnement donné. L'avancée rapide qu'ont connue les technologies de séquençages permet aujourd'hui de séquencer l'ensemble de l'ADN d'un échantillon (séquençage direct), les analyses computationnelles se chargeant ensuite de réassembler les séquences obtenues entre elles et de les réattribuer à des organismes ou à des catégories biologiques fonctionnelles (de type *Clusters of Orthologous Proteins*). Ce type d'analyses donne accès à l'ensemble des fonctions d'un écosystème mais génère une telle quantité de données qu'il peut être parfois difficile d'extraire les informations pertinentes. Une autre façon de procéder est de transférer l'ADN métagénomique dans une banque de clones, et de cribler cette banque pour l'expression de fonctions en particulier. On parle alors de métagénomique fonctionnelle, et seuls certains clones, dont la réponse lors d'un criblage aura été intéressante, seront séquencés. Cette approche, beaucoup plus ciblée, permet notamment la découverte de nouvelles molécules ou d'enzymes, à partir de l'environnement étudié. La métatranscriptomique, plus communément étiquetée sous le terme d'analyse RNA-Seq est, quant à elle, une démarche qui vise à séquencer en masse l'ADNc issu de la rétrotranscription des ARN messagers d'un

environnement complexe. Cette stratégie représente une étape souvent menée en aval de la métagénomique, car elle nécessite l'acquisition préalable de références pan-, supra- ou méta-génomiques (échelle de l'espèce, du genre ou de l'écosystème, respectivement). Elle permet entre autres d'étudier le rôle de facteurs environnementaux sur l'expression des fonctions biologiques clés d'un écosystème. Toutefois, si l'objectif est seulement d'étudier la diversité microbienne en termes de richesse taxonomique (aussi appelée diversité α), ou pour une analyse différentielle de la structure des communautés microbiennes (diversité β), une approche de type séquençage d'un amplicon peut être réalisée, typiquement l'ADN ribosomique codant pour la petite sous-unité 16S (SSU) pour les procaryotes ou 18S pour les eucaryotes. Dans ce cas, seules les informations des taxons présents dans l'échantillon seront analysées. Ainsi, ces approches variées répondent chacune à un niveau différent du questionnement scientifique sur les écosystèmes microbiens, lequel est lié à sa complexité et au degré des connaissances déjà acquises par la communauté scientifique. Pour les environnements encore peu connus et très stratifiés, on préférera la démarche d'analyse taxonomique par amplicon, pour ensuite initier des stratégies adaptées mais plus globales de métagénomiques et métatranscriptomiques.

Un projet de métagénomique se divise donc en deux parties distinctes : la première partie implique des techniques de génomique et de séquençage, la seconde implique les analyses computationnelles des données de séquençage. Nous traiterons dans ce chapitre de la partie génomique et séquençage, et résumerons les pratiques courantes utilisées pour mener avec succès un projet de métagénomique.

Quelle que soit l'approche, l'ADN ou l'ARN séquencé se doit d'être le plus représentatif possible de l'échantillon de départ. De plus, l'échantillon lui-même se doit d'être le plus représentatif possible de l'environnement dans lequel il est prélevé. L'échantillonnage et la préparation des échantillons sont donc autant d'étapes importantes, car susceptibles d'introduire des biais significatifs, qui fausseront les analyses malgré l'utilisation de protocoles avancés de normalisation statistique.

Il n'existe pas à ce jour de méthode ou de procédure infaillible. Néanmoins, avant que les données de séquençage ne soient produites, il est nécessaire de porter une réflexion préalable sur trois étapes clés :

1 - le design expérimental de l'échantillonnage,

2 - la préparation de l'échantillon et l'extraction des acides nucléiques,

3 - les procédures éventuelles d'amplification par réaction en chaîne par polymérase (*polymerase chain réaction*, PCR) associées ou non à la technique de séquençage.

Design expérimental et échantillonnage

L'échantillonnage est séquentiellement la première et la plus cruciale des étapes d'un projet de métagénomique. Un échantillonnage se définit comme la réduction massive d'un ensemble et/ou d'une population donnés par la sélection de certaines sous-unités, dans le but d'obtenir un échantillon vrai. Dans un échantillonnage correct, chaque constituant doit avoir la même probabilité d'être sélectionné, et l'intégrité de chaque élément sélectionné doit être respectée, de sorte que les résultats de l'analyse de l'échantillon puissent être attribués à l'ensemble du matériel

(Gy, 2004). Pour que cette condition soit respectée, l'environnement échantillonné devrait être parfaitement homogène, c'est-à-dire que tous ses éléments constitutifs devraient être identiques. Mais si l'homogénéité peut facilement être définie mathématiquement, elle n'est jamais observée dans les systèmes réels. Au mieux, il peut s'agir d'un système hétérogène bien homogénéisé, où chacun des constituants est distribué aléatoirement dans la matrice, mais souvent la distribution de chacun des constituants dans le matériau est modulaire, voire ségrégée. L'échantillonnage représente donc une source importante d'erreur ou d'imprécision (Gerlach *et al.*, 2003), qu'il convient au mieux de connaître et de maîtriser.

Dans un environnement aussi complexe que le sol, il a été montré qu'une analyse d'un amplicon par pyroséquençage pouvait faire l'objet de nombreux biais, rendant l'analyse de la richesse spécifique peu reproductible (Zhou *et al.*, 2011). Augmenter l'effort d'échantillonnage, en adaptant la stratégie d'échantillonnage à l'environnement étudié, et augmenter le nombre de réplicats (biologiques et techniques) seraient donc le meilleur moyen d'améliorer la reproductibilité. Quoi qu'il en soit, le design expérimental devrait être dicté par la question posée et le type de matrice étudiée, plutôt que par des restrictions techniques ou opérationnelles (Thomas *et al.*, 2012 ; Zhang *et al.*, 2013).

Les aspects spatiaux et temporels sont importants à prendre en compte. Il est ainsi évident que la taille de l'échantillon et la répartition spatiale des éléments que l'on souhaite doser doivent être corrélées pour obtenir des résultats représentatifs de la situation étudiée. L'imprécision fondamentale due à l'échantillonnage peut surtout être réduite en diminuant le diamètre des plus grosses particules. Le broyage de l'échantillon a pour résultat de diminuer la taille des particules et de les homogénéiser, permettant ainsi de travailler sur des échantillons plus petits (Gerlach *et al.*, 2003), ce qui s'avère bien plus efficace que de travailler sur une fraction aliquote. Selon une démarche un peu semblable, des stratégies d'assemblage dites de *pooling* sont parfois réalisées entre échantillons indépendants afin d'homogénéiser la trop forte variabilité des communautés observées à une faible profondeur d'analyse. Toutefois, ces stratégies présentent parfois le risque d'introduire des biais d'observation entre espèces ou entre phylum (Manter *et al.*, 2010). Par ailleurs, il est préférable de réaliser le *pooling* après les étapes éventuelles d'amplifications par PCR. Dans le cas contraire, le *pooling* d'ADN réduit la capacité à capturer la diversité des échantillons assemblés.

Dans un environnement hétérogène, il existe plusieurs niveaux de structuration spatiale des communautés microbiennes. Dans le sol, un sous-échantillonnage centimétrique peut montrer une hétérogénéité de la structure des peuplements microbiens qui n'était pas apparente dans un échantillonnage plus distant (Ritz *et al.*, 2004). Ainsi, en fonction de la question posée, la structure des communautés microbiennes peut aussi bien être recherchée à micro-échelle (en cohérence avec les unités fonctionnelles du sol) qu'à une échelle plus grande, pour des aspects plus globaux de la biodiversité. En somme, la probabilité d'observer une espèce au sein d'une communauté très large dans un environnement très hétérogène dépendra de son abondance dans la communauté, de la taille de l'échantillon et de la distribution spatio-temporelle des espèces de manière individuelle (Plotkin et Muller-Landau, 2002).

Toutefois, en raison de l'extrême complexité des écosystèmes naturels, il reste relativement difficile de mesurer l'abondance et la nature de toutes les espèces ou de toutes les unités biologiques fonctionnelles à toutes les échelles écologiques d'un environnement. Afin de tester les meilleures stratégies à mettre en œuvre (nombre contre taille des échantillons) lors d'une analyse métagénomique sur un environnement *a priori* peu connu, il est conseillé par exemple d'établir des simulations mathématiques appliquées à des pré-tests aléatoires d'échantillonnage et volontairement sous-dimensionnés (Zhou *et al.*, 2013). D'une autre manière, ces prédictions mathématiques peuvent être conduites sur des jeux de données simulés construits à partir des connaissances déjà acquises sur ces écosystèmes. Il conviendra alors d'ajuster la profondeur de séquençage (nombre de séquences par échantillon) à la stratégie choisie. Certaines de ces études, menées par exemple sur des communautés microbiennes de sol, ont démontré l'avantage des grandes quantités d'échantillons à faible effort de séquençage par rapport aux petits nombres d'échantillons analysés avec une grande profondeur de séquençage (Kuczynski *et al.*, 2010). Toutefois, le séquençage à forte profondeur restera de prédilection pour la compréhension des mécanismes écologiques subtils ou pour appréhender le rôle d'espèces sous-dominantes ou rares.

Préparation des échantillons et extraction des acides nucléiques

Isoler l'ADN à partir d'échantillons environnementaux est une étape critique, d'autant plus que la préparation des librairies pour les technologies de séquençage à haut débit requiert de grandes quantités d'ADN, de l'ordre d'une dizaine, voire plusieurs dizaines, de microgrammes. L'objectif est de lyser les cellules, par des moyens chimiques, physiques et/ou biologiques, pour extraire uniquement leur ADN. En effet, la co-extraction de composés naturels est susceptible de rendre l'utilisation ultérieure de l'ADN extrait difficile, en empêchant par exemple la lyse des cellules, en le dégradant ou en le capturant, ou encore en agissant comme inhibiteur des réactions enzymatiques (Wilson, 1997). Ainsi, dans les sols, les composés phénoliques, l'acide humique ou les métaux lourds sont connus pour être des inhibiteurs de PCR. Dans l'environnement alimentaire, les inhibiteurs sont plutôt le glycogène ou le gras, alors que dans les échantillons cliniques, l'hémoglobine, l'urée ou l'héparine sont concernés.

Les inhibiteurs enzymatiques peuvent parfois être contournés en jouant sur les paramètres de la réaction (par exemple en ajoutant du diméthylsulfoxyde lors d'une réaction de PCR), mais il est plus avantageux encore d'essayer d'obtenir un ADN avec le moins possible de contaminants environnementaux. Différents types de traitements peuvent être combinés pour maximiser la lyse des cellules. Pour isoler spécifiquement les acides nucléiques, des purifications à l'aide de phénol/chloroforme permettent de se débarrasser des protéines et de certains contaminants. Cependant, ces méthodes sont souvent fastidieuses à mettre en place et des kits commerciaux pour l'extraction d'ADN existent. En général, ces kits sont développés pour toutes sortes de matrices (sols, plantes, sang, matrices alimentaires, etc.) et permettent d'isoler l'ADN (en général sur des colonnes d'affinités ou sur des billes

aimantées) en se débarrassant des principaux contaminants. Par ailleurs, certaines pratiques assez courantes dans les analyses de détection d'espèces bactériennes par PCR quantitative consistent à éliminer les effets des inhibiteurs par dilution de l'ADN extrait ou de l'échantillon (Fode-Vaughan *et al.*, 2001). L'efficacité des kits reste néanmoins très dépendante de la nature de l'échantillon et les rendements d'extraction peuvent être très mauvais (sols contaminés par des métaux lourds par exemple). Ainsi, les méthodes plutôt traditionnelles de purification utilisant le phénol/chloroforme bénéficient toujours d'un meilleur rendement global quand on les compare aux kits d'extraction sur colonnes d'affinité. Si le choix quantité plutôt que qualité ne se pose pas pour les écosystèmes microbiens enrichis (de type microbiote intestinal), le rendement d'extraction reste néanmoins un point crucial pour les environnements paucimicrobiens (comme les eaux souterraines, l'air, les biofilms de surfaces, les biopsies et les aliments non-fermentés).

Pour pallier à la co-extraction de composés naturels inhibiteurs lors de l'extraction directe d'ADN métagénomique, une méthode indirecte peut être également utilisée, en séparant d'abord les bactéries de leur environnement, puis en réalisant une extraction d'ADN *ex-situ*. Ce type de stratégie s'applique aussi lorsque la communauté microbienne ciblée est associée à un hôte (invertébrés, plantes) dont le génome, particulièrement large, risque de noyer les séquences d'origine microbienne dans les efforts de séquençage. Pour cela, des méthodes par centrifugations successives à faible vitesse sont utilisées pour séparer les cellules de la matière environnementale (Lindahl et Bakken, 1995), ou le recours à des centrifugations sur gradients différentiels, tels ceux utilisés sur couches de Nicodenz (Lindahl et Bakken, 1995). D'autres techniques, plus sophistiquées, impliquent des étapes de filtrations et de cytométrie de flux afin d'enrichir la fraction ciblée comme par exemple la fraction virale d'un écosystème (Venter *et al.*, 2004 ; Angly *et al.*, 2006 ; Palenik *et al.*, 2009). Les méthodes indirectes permettent de diminuer sensiblement la présence d'inhibiteurs et d'extraire des fragments d'ADN de plus grande taille (> 20 kpb). Cependant, avec ce type d'extraction, les bactéries solidement adsorbées à leur matrice sont en général éliminées, apportant un biais potentiel en termes de diversité lors de l'analyse.

Toutes les bactéries ne réagissent pas de manière égale face à une même méthode d'extraction d'ADN. Des bactéries peuvent être protégées des étapes de lyse à l'intérieur des micro-environnements de l'échantillon. De plus, les bactéries à Gram-positif et à Gram-négatif possèdent des différences fondamentales dans la structure de leurs enveloppes respectives qui font qu'elles ne sont pas sensibles aux mêmes stress physiques ou chimiques. Le moyen le plus sûr pour évaluer un protocole d'extraction reste la stratégie d'ensemencement artificiel de l'échantillon environnemental par des quantités variables d'une espèce ou d'un jeu d'espèces (le terme *spiking* est communément utilisé en anglais). Cette démarche permet de s'assurer que l'extraction de l'ADN fonctionne de manière homogène entre les différents genres, voire entre les différents phylums bactériens. Ainsi, le protocole peut éventuellement être modifié et amélioré en amont du processus de collecte des échantillons. Le *spiking* est donc fortement à conseiller si l'échantillonnage concerne des environnements difficiles à obtenir (par exemple les fonds marins hydrothermaux)

ou si l'échantillonnage de microbiotes s'effectue sur des cohortes cliniques d'individus ou d'animaux. De manière alternative, il est possible de réaliser du *spiking* directement avec de l'ADN en quantité connue et provenant d'une espèce bien identifiée. Cette fois, on vise à déterminer la présence potentielle d'inhibiteur de PCR dans l'ADN extrait. Le *spiking* est un atout pour l'expérimentateur mais il ne peut malheureusement pas résoudre la problématique liée aux espèces non connues et/ou non cultivables.

D'une manière générale, ces dernières années, les méthodes d'extraction d'ADN ont peu évolué. La variabilité entre les protocoles métagénomiques réside donc principalement dans le choix des kits d'extraction, dans la procédure de broyage des échantillons, dans le choix de séparer ou non le microbiote de sa matrice. Dans certains cas, le délai entre l'étape d'extraction de l'ADN et sa congélation pour conservation peut être crucial. Dans le cas du microbiote fécal humain par exemple, une analyse comparative de différents projets indépendants a démontré que la variabilité de composition du microbiote observée dans la méta-analyse était principalement due aux biais techniques des protocoles d'extraction d'ADN et de séquençage plutôt qu'aux différentes procédures en aval, d'assemblage, de tri, d'annotation et d'analyses biostatistiques (Lozupone *et al.*, 2013). Progressivement, des initiatives se développent pour changer les paradigmes techniques obsolètes et unifier les méthodes d'extraction d'ADN afin d'assurer des compatibilités comparatives et envisager une analyse plus globale du microbiome terrestre sur des dizaines de milliers d'écosystèmes (Knight *et al.*, 2012).

L'amplification par PCR est au cœur de nombreuses études en métagénomique que ce soit en tant que processus d'enrichissement d'un locus génétique ou d'un ensemble métagénomique (amplicon ADNr 16S, par exemple), ou en tant que processus de séquençage. Le processus de PCR est compétitif, si bien que les espèces à faible abondance seront amplifiées proportionnellement avec moins d'efficacité que celles présentes à forte abondance dans l'échantillon (Reysenbach *et al.*, 1992 ; Polz et Cavanaugh, 1998). De nombreux autres paramètres peuvent être à l'origine de biais et provoquer l'émergence de séquences chimériques. La formation de ces chimères provient d'une extension avortée au cours d'un cycle lors de la PCR. Le produit d'extension avorté peut néanmoins servir d'amorce à une molécule matrice non spécifique lors du cycle suivant, générant ainsi une séquence artificielle qui n'a aucune réalité biologique. Le degré de formation des chimères peut être très variable car il dépend du degré de conservation du locus amplifié et des conditions d'amplification lors de la PCR (Haas *et al.*, 2011). Dans le cas de l'ADNr 16S, qui est l'exemple type d'une molécule très conservée entre espèces, le taux de formation de chimères peut atteindre 30 % des amplicons produits (Wang et Wang, 1997). Ce taux dépendra également d'autres paramètres tels que les amorces choisies (Wu *et al.*, 1991), le pourcentage de guanine et cytosine (GC %) des régions cibles à amplifier (Reysenbach *et al.*, 1992) et des cycles d'amplifications (Ishii et Fukui, 2001). Par exemple, il a été démontré que la région V6-V9 de l'ADNr 16S est plus sensible (~ 3 %) à la création de chimères que la région V1-V5 (Haas *et al.*, 2011). De nombreux sites web dédiés à l'analyse des amplicons de l'ADNr 16S ou à la détection des molécules chimériques proposent des guides de bonnes

pratiques pour les amplifications par PCR dédiées à la métagénomique. Toutefois, il n'est pas possible techniquement de réduire totalement ce biais et les séquences chimériques doivent faire l'objet d'une recherche spécifique par des méthodes computationnelles.

En plus de la genèse de séquences chimériques, les étapes de PCR, et particulièrement celles impliquées dans les étapes ultérieures de séquençage, peuvent générer des distorsions dans la réelle abondance des espèces. La technologie de pyroséquençage 454 est connue pour légèrement sous-représenter les séquences dont le GC % est supérieur à 60 % (Pinto et Raskin, 2012). L'utilisation des « tags » lors du séquençage multiplex peut aussi être à l'origine de biais d'amplification liés à ces séquences code-barres (Cai *et al.*, 2013).

Afin de minimiser l'ensemble de ces biais, il est fortement recommandé que le produit issu de la PCR qui sera analysé par séquençage haut débit résulte d'un mélange équimolaire de produits de PCR réalisés en triplicata. De façon identique, le séquençage devra faire l'objet d'une analyse en réplicats (2 minimum). Cette stratégie de bonnes pratiques doit être prise en compte dès l'amont du projet et sur le calcul du dimensionnement du séquençage car de nombreuses revues scientifiques, spécialisées dans les études métagénomiques, exigent désormais leur mise en œuvre comme critère de publication.

Le séquençage à haut débit

Depuis le premier brin d'ADN séquencé en 1977, par les équipes de Frederick Sanger d'une part (Sanger *et al.*, 1977), et de Allan Maxam et Walter Gilbert d'autre part (Maxam et Gilbert, 1977), les technologies de séquençage ont très rapidement évolué. Les technologies de séquençage à haut débit (ou NGS, pour *Next Generation Sequencing*) permettent aujourd'hui de séquencer en parallèle des centaines de milliers de brins d'ADN, et donc de diminuer drastiquement les coûts, rendant le séquençage beaucoup plus accessible à la majorité des équipes de recherche.

Toutes les méthodes NGS actuellement disponibles ne se valent pas en termes de coûts. Même si ceux-ci ont beaucoup diminué, le séquençage reste un poste de dépense conséquent dans un projet de recherche. Les différentes technologies NGS diffèrent également au niveau de leur précision (taux d'erreur de séquençage) et du nombre et de la longueur des séquences générées. Le choix d'une technologie de séquençage a également un impact sur le traitement des fichiers en aval. Selon la technique choisie, les fichiers de sortie ne seront pas les mêmes (c'est particulièrement vrai pour les technologies utilisant les espaces colorimétriques), et pourront ou devront être traités de manières différentes lors des analyses computationnelles.

Les technologies NGS ont vu le jour vers 2005 et n'ont pas cessé d'évoluer depuis cette date avec une fréquence d'apparition de nouvelles générations tous les 4 ans environ. Les améliorations successives ont d'abord visé l'augmentation de la longueur des lectures et de la profondeur de séquençage. Les échelles de séquençage sont désormais tellement importantes que le stockage des données et les capacités computationnelles des clusters de calculs sont désormais les

facteurs limitants à l'analyse. Les nouvelles générations encore en développement se focalisent donc davantage sur la réduction des coûts, sur une meilleure rentabilité de l'investissement des machines, sur leur miniaturisation et sur des technologies qui visent à minimiser les biais en essayant d'obtenir une information de séquençage la plus proche de l'état natif de l'ADN ou de l'ARN au sein des organismes. Du fait de ces évolutions permanentes, il est peu judicieux d'en faire ici une synthèse comparative qui risque d'être rapidement obsolète. On pourra se référer à la version anglophone de l'encyclopédie libre Wikipedia (http://en.wikipedia.org/wiki/DNA_sequencing), dont la mise à jour est entretenue par une communauté scientifique très active, et qui regorge d'informations très détaillées sur les technologies NGS.

Avantages et désavantages pour la métagénomique des technologies NGS les plus populaires

Pyroséquençage 454/Roche (apparition en 2005)

Avec la méthode du pyroséquençage (technologie 454), la séquence d'ADN est déterminée en direct lors de la synthèse d'ADN qui s'effectue lors d'une PCR en émulsion (Ronaghi *et al.*, 1996). Pour avoir un signal détectable lors de la lecture de la séquence, les brins d'ADN sont individuellement fixés à des billes microscopiques, chacune de ces billes étant par la suite placée séparément dans un des puits d'une plaque picotitre (comportant au total 1,7 million de puits). L'ADN est alors amplifié pour donner un grand nombre de molécules identiques par puits. Pour lire la séquence d'ADN, les désoxyribonucléotides triphosphates sont ajoutés séquentiellement, et à chaque incorporation de nucléotide complémentaire sur le brin d'ADN, la libération d'un pyrophosphate inorganique conduit à une réaction enzymatique couplée à une luciférase qui produit alors un photon. Ces photons sont détectés par un capteur colorimétrique, et la succession des signaux détectés permet d'interpréter quelles bases ont été ajoutées et dans quel ordre.

Le pyroséquençage permet d'obtenir des lectures très longues par rapport aux autres méthodes de séquençage de nouvelle génération. Cette longueur de lecture obtenue facilite grandement l'assemblage ultérieur des séquences, et permet également de séquencer de grands fragments d'un gène en particulier dans une communauté microbienne (gène de l'ADNr 16S par exemple), faisant de cette technologie la référence pour l'analyse taxonomique de la diversité par séquençage. Cependant, cette méthode est aussi la plus onéreuse, et le risque d'obtenir des erreurs de séquençage est en général plus élevé notamment dans les régions homopolymériques où le signal colorimétrique sature. L'autre désavantage du pyroséquençage est sa proportion à introduire des insertions/délétions (souvent d'une paire de base), un phénomène qui génère lors des analyses taxonomiques sur amplicon (type ADNr 16S) un phénomène d'inflation de la diversité (Schloss *et al.*, 2011). Toutefois, le pyroséquençage reste la technologie la plus fiable pour de longues lectures, notamment avec la nouvelle génération de GS-FLX++ titanium capables de lire jusqu'à 900 pb. Il s'agit donc d'une technique de choix pour l'assemblage de génomes riches en régions répétées.