Initiation à la bioinformatique pour la microbiologie
Dr Anne JAMET
(Médecin microbiologiste Hôpital Necker)

Sources:
A Primer on Infectious Disease Bacterial Genomics (Lynch 2016)
Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing (Ronholm 2016)

Le séquençage à haut débit (HTS pour high-throughput sequencing aussi appelé NGS pour next-generation sequencing) a transformé la recherche biomédicale. La baisse des coûts et le développement d'outils informatiques ont entraîné l'adoption généralisée de cette technologie.
Le séquençage Sanger (souvent appelés méthodes de "séquençage traditionnel") est toujours très utilisé pour séquencer des fragments d'ADN de quelques kilobases. Le séquençage Sanger exige plus de temps pour générer les données que pour les analyser; en revanche, les plates-formes NGS peuvent produire des quantités massives de données relativement rapidement par rapport au temps souvent long nécessaire à leur analyse et à leur interprétation. La capacité à stocker, gérer, analyser et interpréter les données issues du NGS est donc aujourd'hui un enjeu important en biologie.

Design expérimental

Chaque échantillon doit être associé à une description précise. Ces données décrivant chaque échantillon sont connues sous le nom de "métadonnées" et sont cruciales pour l'interprétation biologique des résultats. Des informations minimales telles que la source, l'emplacement et la date de collecte doivent être fournies pour chaque échantillon. Les données cliniques associées aux échantillons (sexe, âge, pathologies sous-jacentes, terrain, traitements...) sont particulièrement précieuses. En effet, si on a une collection de plusieurs centaines de bactéries

séquencées sans métadonnée on se contentera de comparer des génomes bactériens. Par contre, si chaque génome est assorti d'informations cliniques on pourra chercher dans ces génomes des facteurs de virulence associés à la présence de certains symptômes chez les patients (pathogénomique).

Le NGS permet de séquencer rapidement des dizaines voire des centaines de souches mais il ne faut pas oublier que l'une des premières et des plus importantes étapes de toute investigation scientifique est la génération d'une hypothèse. Il faut bien poser la question scientifique à laquelle on veut répondre avant de s'engager dans un projet de séquençage car même si les coûts du séquençage diminuent, le temps d'analyse peut engendrer une perte de temps et donc d'argent importante si le projet est mal défini au départ.
Gardez donc à l'esprit ceci : « Réfléchir d'abord ; séquencer après ».

Par exemple, un seul passage (« run ») d'un petit séquenceur MiSeq d'Illumina peut produire jusqu'à 15 gigabases. La plupart des projets nécessitent plusieurs « runs » de MiSeq. Par conséquent, l'analyse de tels ensembles de données peut prendre beaucoup de temps et de ressources.
Néanmoins, la plupart des analyses de base sur un seul ou une dizaine de génomes bactériens peuvent être réalisées sur de simples ordinateurs personnels (PC avec au moins 8 Go de RAM, 500 Go de stockage et 4 CPU) avec les logiciels et la configuration appropriés.
Les grandes étapes d'un séquençage NGS sont résumées dans la figure 1.

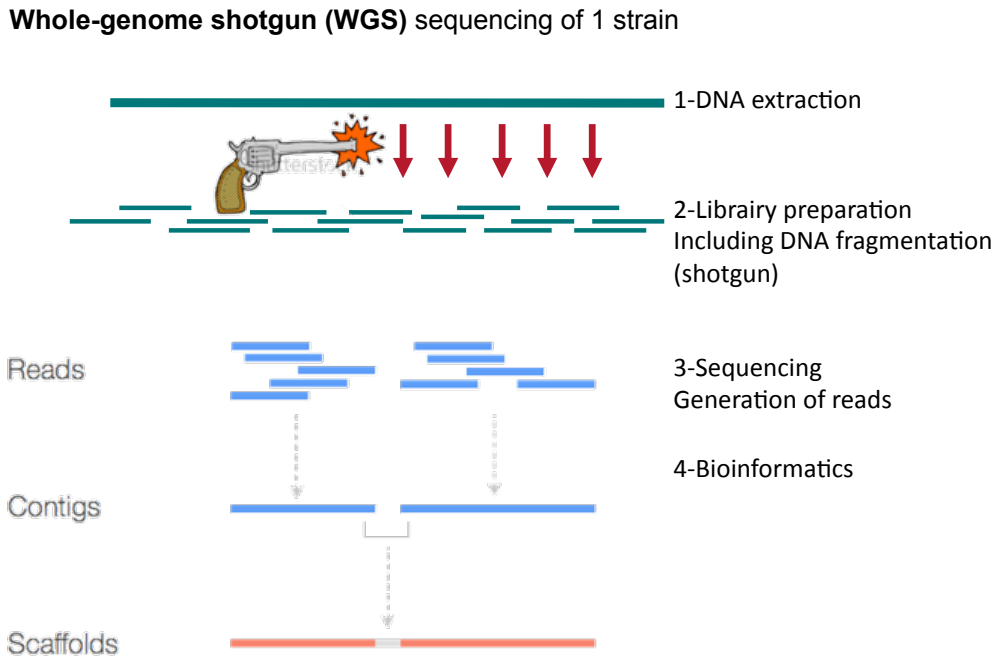

Figure 1. Les grandes étapes du séquençage NGS.

Extraction de l'ADN

La composante expérimentale des projets NGS, comprend le recueil, le traitement des échantillons et la génération de données.
Le traitement des échantillons débute par l'isolement des bactéries en culture et l'extraction d'ADN, qui sont des étapes réalisables par la majorité des laboratoires de biologie moléculaire ou de microbiologie.
Néanmoins, la qualité de l'ADN extrait est cruciale pour obtenir des séquences de qualité et la technique d'extraction doit souvent être mise au point en fonction de l'espèce bactérienne. Certaines espèces de bactéries à Gram-positif nécessitent l'utilisation d'enzymes spécifiques pour lyser leur paroi (lysostaphine pour les Staphylocoques, Mutanolysine pour les Streptocoques...). Dans le cas des Mycobactéries la lyse de la paroi nécessite souvent un traitement mécanique (type Fastprep).
Une des étapes clés du séquençage est la détermination très précise de la qualité et de la quantité de l'ADN extrait. En effet, l'ADN extrait va devoir être « préparé » pour le séquençage. Cette « préparation » de l'ADN va se faire à l'aide de kits commerciaux appelés « kits de préparation de librairies ». La plupart des protocoles de préparation des librairies sont très sensibles à la concentration de l'ADN d'où l'importance d'obtenir une quantification très précise. La concentration en ADN va être déterminée par deux méthodes de quantification : l'absorbance UV (p. ex. spectrophotomètre type NanoDrop) et par système de fluorescence (p. ex. Qubit). Les approches par fluorescence sont plus précises pour la quantification que les méthodes fondées sur l'absorbance UV. Par contre, c'est le spectrophotomètre qui va permettre la détection des impuretés résiduelles présentes dans la suspension d'ADN après l'étape d'extraction.

Le NGS est beaucoup plus sensible que le séquençage Sanger aux impuretés associées à l'ADN extrait. Les impuretés sont problématiques car elles ont un impact négatif sur de nombreuses étapes enzymatiques pendant la préparation des librairies.
Tous les échantillons d'ADN doivent être évalués pour la présence d'impuretés de type protéines, de composés organiques et d'autres inhibiteurs d'enzymes tels que les sels biliaires ou les glucides (par exemple, les capsules bactériennes).
La pureté de l'ADN est évaluée en calculant les rapports d'absorbance, à savoir A 260/A 280 (le rapport de l'absorbance à 260 nm divisé par la lecture à 280 nm) et A 260/A 230 (Figure 2); plus les rapports sont faibles plus il y a de contaminants présents. Un faible rapport A 260/A 280 (inférieur à 1,8) suggère la présence de micelles de protéines ou de phénol; les acides nucléiques qui ne sont pas complètement remis en suspension peuvent disperser la lumière, ce qui entraîne également un faible rapport A 260/A 280. Une absorption élevée à 230 nm est causée par une contamination par des particules (p. ex. particules de silice), des précipités de cristaux de sel (c.-à-d. thiocyanate de guanidine, LiCl ou NaI), des ions phénolate, des solvants et d'autres composés organiques. Un échantillon d'ADN est considéré comme suffisamment pur lorsque les rapports A 230/A 260/A et 260/A 280 sont d'au moins 1,8. Un rapport A 260/A 280 élevé (supérieur à 2,1) indique généralement la présence d'ARN ; ceci peut être corrigé en traitant avec une enzyme RNase qui va dégrader l'ARN de l'échantillon sans impacter l'ADN. La contamination protéique ou phénolique est indiquée par des rapports A 230/A 260 supérieurs à 0,5. Le nettoyage post-extraction des ADN en utilisant des colonnes de purification peut permettre d'éliminer les impuretés.

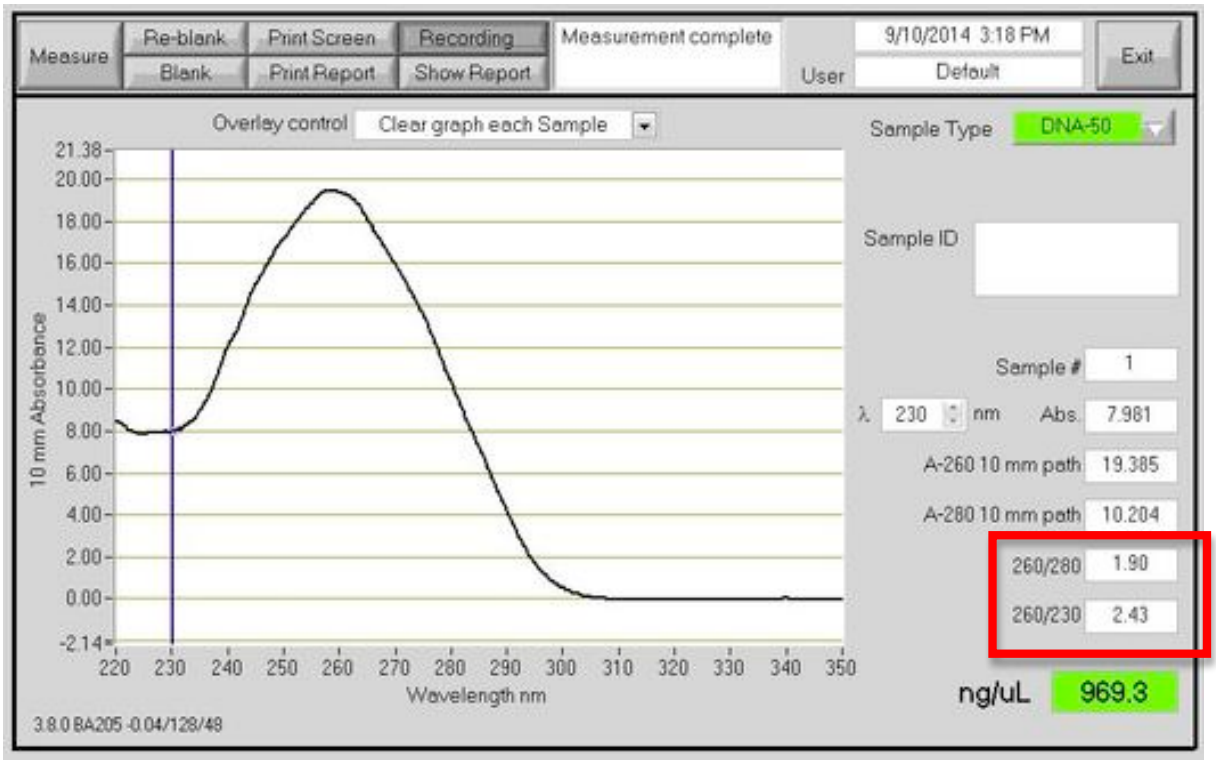

Figure 2. Résultat du dosage d'un échantillon d'ADN de bonne qualité par Nanodrop (absorbance UV).

Préparation des librairies et séquençage

Certains laboratoires de recherche ou de microbiologie hospitaliers sont déjà équipés de leur propre séquenceur. Une autre possibilité très fréquente est l'envoi des échantillons soit à une plateforme locale de séquençage soit à un centre extérieur de séquençage.

Selon le séquenceur qui va être utilisé, les séquences générées vont avoir des caractéristiques différentes, notamment leur longueur et le nombre d'erreurs qu'elles contiennent.

- La longueur de lecture : les séquenceurs à lectures courtes (« short reads ») produisent des lectures entre 75 et 1 000 pb et les séquenceurs à lectures longues (« long reads ») produisent des lectures de 1 000 à > 30 000 pb.
Les plates-formes NGS à lectures courtes les plus courantes comprennent les plates-formes HiSeq, NextSeq et MiSeq d'Illumina et les plates-formes Ion PGM et S5 de Thermo Fisher. Les technologies de séquençage à plus longues lectures sont les systèmes PacBio Sequel et PacBio RSII de Pacific Biosciences et les systèmes MinION et PromethION d'Oxford Nanopore Technologies.
Plus les reads sont longs plus ils vont permettre de résoudre les problèmes posés par les séquences répétées de l'ADN, et notamment l'assemblage des régions flanquées par des séquences répétées (voir ci-dessous le chapitre sur l'assemblage des génomes).

La technologie la plus utilisée est celle d'Illumina dont les séquenceurs produisent des short reads (en général 100 ou 150 nucléotides de long).

Lors de la préparation de la librairie pour le séquençage on va commencer par fragmenter l'ADN à séquencer en petits fragments d'environ 1 000 pb. Cette fragmentation initiale de l'ADN est appelée « shotgun ». Ces fragments pourront être ensuite séquencés soit seulement à partir d'une de leurs extrémités (« single-end ») soit à partir de leurs deux extrémités (c'est le cas le plus fréquent appelé « paired-end ») (Figure 3). Les techniques de séquençage basées sur une étape de « shotgun » sont appelées « whole-genome shotgun » (WGS) et peuvent porter sur un seul génome ou sur un échantillon contenant un mélange de bactéries (métagénomique shotgun).

Les kits de séquençage ont un numéro de "cycle", qui est le nombre de fois que l'instrument ajoutera un nucléotide à la copie du fragment d'ADN. Par exemple, un "kit de 300 cycles" pourrait théoriquement synthétiser jusqu'à 300 nucléotides. En mode « single-end », l'instrument séquencera un fragment de 300 pb dans une seule direction. En mode « paired-end », l'instrument séquencera le même fragment dans les deux directions (150 pb à partir de chaque extrémité). Par conséquent, si les fragments initiaux d'ADN ont une longueur de 1 000 pb (taille de l'insert), le mode « paired-end » pourrait générer des séquences de 150 pb aux deux extrémités du fragment de 1 000 pb, laissant un espace intermédiaire (distance intérieure) de 700 pb qui n'est pas séquencé. La distance intérieure connue entre les lectures « paired-end » peut être appliquée de manière algorithmique pour aider à résoudre les problèmes des régions répétées.

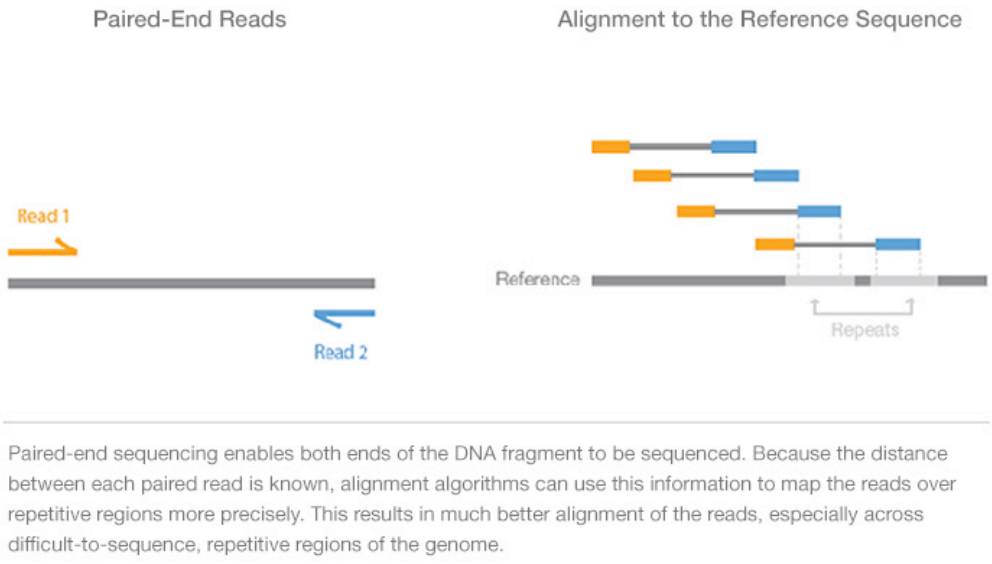

Figure 3. Paired-end reads

Connaître la biologie du génome des microbes que l'on veut séquencer peut grandement aider au choix de la technique de séquençage. Par exemple, les organismes monomorphes tels que *Bacillus anthracis* ou *Mycobacterium tuberculosis* dont le génome varie très peu d'une souche à l'autre peuvent être séquencés de façon appropriée en utilisant de courtes lectures. Les organismes dont les génomes contiennent des séquences répétées multiples (p. ex., opérons ribosomiques et séquences d'insertion) et de l'ADN étranger acquis (p. ex., prophage et îles génomiques) peuvent nécessiter de combiner des lectures courtes et des lectures longues afin d'assembler le génome de façon appropriée.

- Types et taux d'erreur. Les types et les taux d'erreur varient selon les technologies. Les technologies à lectures courtes comme Illumina ont des taux d'erreur plus faibles que les technologies à lectures longues, plus comparables à ceux du séquençage traditionnel de Sanger à ~0.1%. Les erreurs peuvent généralement être compensées avec une profondeur de séquençage suffisante. Une plus grande profondeur de séquençage signifie qu'une même base est lue plusieurs fois (Figure 4).

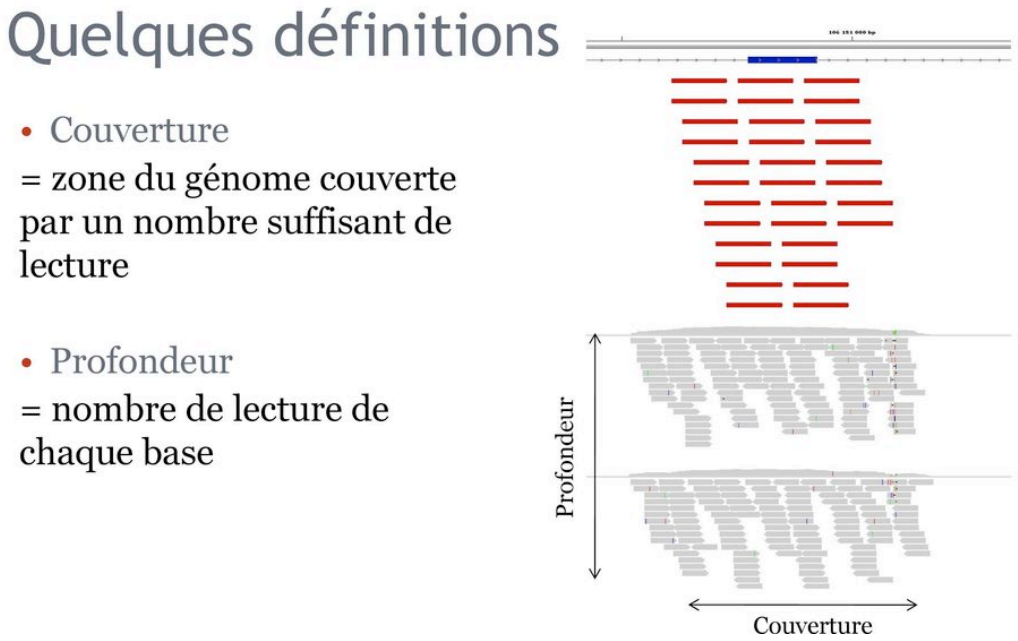

Figure 4. Couverture et profondeur.

- Couverture : Le terme "couverture" est souvent utilisé de manière interchangeable avec "profondeur" ou "redondance de séquence" et fait référence au nombre de fois qu'une base est représentée dans les données brutes de séquençage. Les séquences produites par l'instrument ne sont pas réparties également sur le génome et, par conséquent, le terme couverture est souvent rapporté comme étant la couverture moyenne (par exemple, couverture de 10×) et est utilisé pour planifier à l'avance le nombre d'échantillons à placer simultanément dans un run de séquençage. En effet, pour réduire les coûts on séquence plusieurs échantillons en même temps lors d'un run (c.-à-d. les échantillons sont multiplexés). Si on met trop d'échantillons en même temps il y a un risque d'obtenir des couvertures trop faibles et de ne pas pouvoir faire une analyse bioinformatique correcte par la suite.
Sachant que les lectures ne seront pas uniformément réparties sur l'ensemble du génome, il peut être judicieux de surestimer la couverture requise pour chaque échantillon afin que les régions à faible couverture soient suffisantes pour les analyses en aval, comme la recherche des variants nucléotidiques (SNP). Par exemple, si on souhaite une couverture minimale de 50×, on va chercher à obtenir une couverture pour chaque échantillon de 75 à 90× afin de s'assurer que toutes les régions répondent à l'exigence de couverture minimale de 50x. Tous les fournisseurs de plates-formes NGS disposent des ressources nécessaires pour fournir l'information théorique nécessaire au calcul du nombre de génomes qui peuvent être combinés sur un cycle une fois que la couverture désirée a été stipulée.
La couverture moyenne théorique (C) peut être calculée avec l'équation de Lander-Waterman comme C = LN/G, où L est la longueur de la lecture, N est le nombre de lectures et G est la longueur du génome en paires de bases. Ainsi, il faut 1 million de reads de 150 pb pour obtenir une couverture de 30x d'un génome d'*E. coli* de 5 Mégabases.

Analyse de la qualité des séquences et nettoyage (“trimming”)

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) est un logiciel open-source populaire qui est le plus utilisé pour une vue d'ensemble de la qualité des lectures. FastQC produit un rapport composé d'une série de graphiques pour des aspects tels que la qualité de base et le contenu en G+C des lectures séquencées. Le rapport est évalué par FastQC et reçoit une note de "pass", "warning" ou "failure" sur la base de critères intégrés prédeterminés (Figure 5).

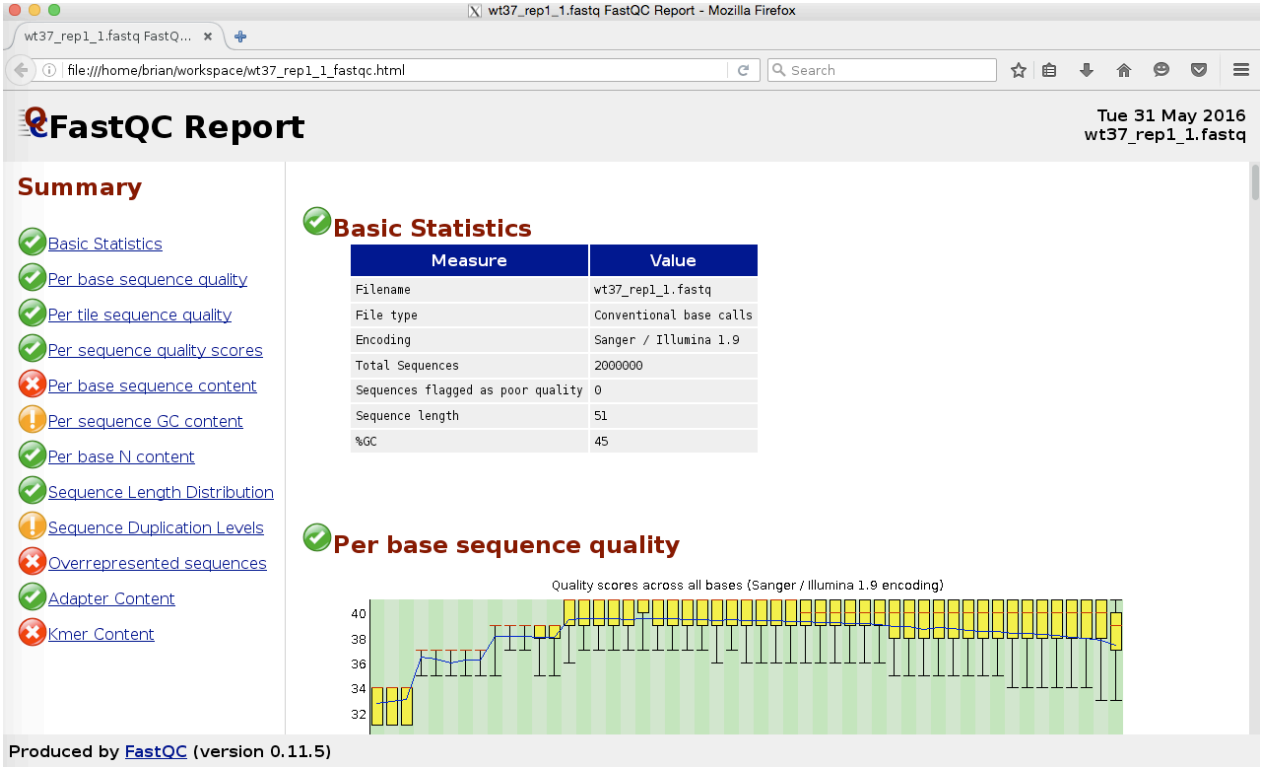

Figure 5. FastQC report.

Afin de générer des lectures de qualité supérieure pour des analyses plus rigoureuses en aval, on peut effectuer le nettoyage des lectures appelé « trimming ». Le nettoyage des lectures se fait en enlevant les lectures de mauvaise qualité, en masquant (en remplaçant les bases de mauvaise qualité par un "N" pour représenter une base "indéterminée"), en coupant les extrémités de mauvaise qualité des lectures, et en enlevant les adaptateurs et autres artefacts de séquençage. Les logiciels de nettoyage des lectures les plus utilisés sont FASTX-Toolkit et Trimmomatic.

Une autre étape de contrôle de la qualité du séquençage est la détection et l'élimination possible des séquences d'ADN contaminantes. Les séquences d'adaptateurs qui ont été ajoutés aux extrémités des fragments d'ADN pendant la préparation de la librairie peuvent être détectées et supprimées par des logiciels tels que Trimmomatic ou Cutadapt. D'autres sources de contamination comprennent le phage de contrôle (phiX utilisé par Illumina) ou un problème lors de la préparation de l’échantillon accidentellement contaminé/mélangé. Les programmes conçus pour identifier les contaminants comprennent Kraken, Deconseq et Genome Peek.
S'assurer que les données de séquençage sont de haute qualité et nettoyées de toute contamination potentielle augmentera la qualité des résultats d'analyse en aval.

L'assemblage *de novo*

L'assemblage *de novo* est défini comme la reconstruction d'un génome à partir des reads sans l'aide d'un génome de référence. Plus techniquement, l'assemblage *de novo* est le processus informatique de reconstruction de séquences consensuelles contiguës plus longues (appelées contigs) en déterminant le chevauchement le plus long et le placement optimal de lectures plus courtes. Le résultat de cette première approche automatisée est considéré comme un brouillon de génome (« draft »). Si des informations supplémentaires telles que des séquences à lectures longues sont disponibles, ces contigs peuvent être ordonnés pour générer des contigs encore plus longs ; l'assemblage *de novo* résultant peut alors être classé comme un "génome de haute qualité". La désignation de "génome fini" nécessite la résolution de toutes les erreurs d'assemblage et des incertitudes de séquençage.
Le niveau de finition recherché pour les génomes d'un projet dépendra des exigences du projet de séquençage telles que définies dans la phase de planification du projet. En effet, obtenir un génome fini coûte beaucoup plus cher et n'est pas forcement utile.

Le logiciel actuellement le plus utilisé pour l'assemblage *de novo* des génomes bactériens est SPAdes. Nous tenons à souligner qu'il n'est pas nécessaire de comprendre la théorie mathématique derrière tous les nouveaux assembleurs lorsque l'on a seulement besoin de faire des choses simples à partir des données de séquençage. Il est toutefois important de comprendre que tous les assembleurs ont leurs forces et leurs limites et il est nécessaire de savoir évaluer la qualité d'un assemblage. Le rendement d'un assembleur est influencé par la biologie du génome (p. ex. éléments répétitifs, taille globale, plasmides extrachromosomiques multiples, etc.), la nature des données (p. ex. longueur des séquences, profondeur de couverture et uniformité) et les ressources informatiques disponibles.
Les statistiques pour évaluer les assemblages génomiques comprennent le nombre total et la longueur des contigs. Un assemblage de bonne qualité a, en général, un nombre de contigs inférieur à 100. La longueur totale cumulée de tous les contigs doit être proche de la taille attendu du génome (par exemple : 5 mégabases pour un *E. coli* et 3 mégabases pour un *S. aureus*).
Une autre mesure de la qualité est la longueur N50. Une valeur N50 de longueur N signifie que la moitié des bases de l'assemblage appartient à des contigs ayant une longueur L<N. Ainsi, la longueur N50 correspond à la longueur du contig qui avec les contigs de taille plus grande contiennent la moitié des bases du génome (Figure 6).

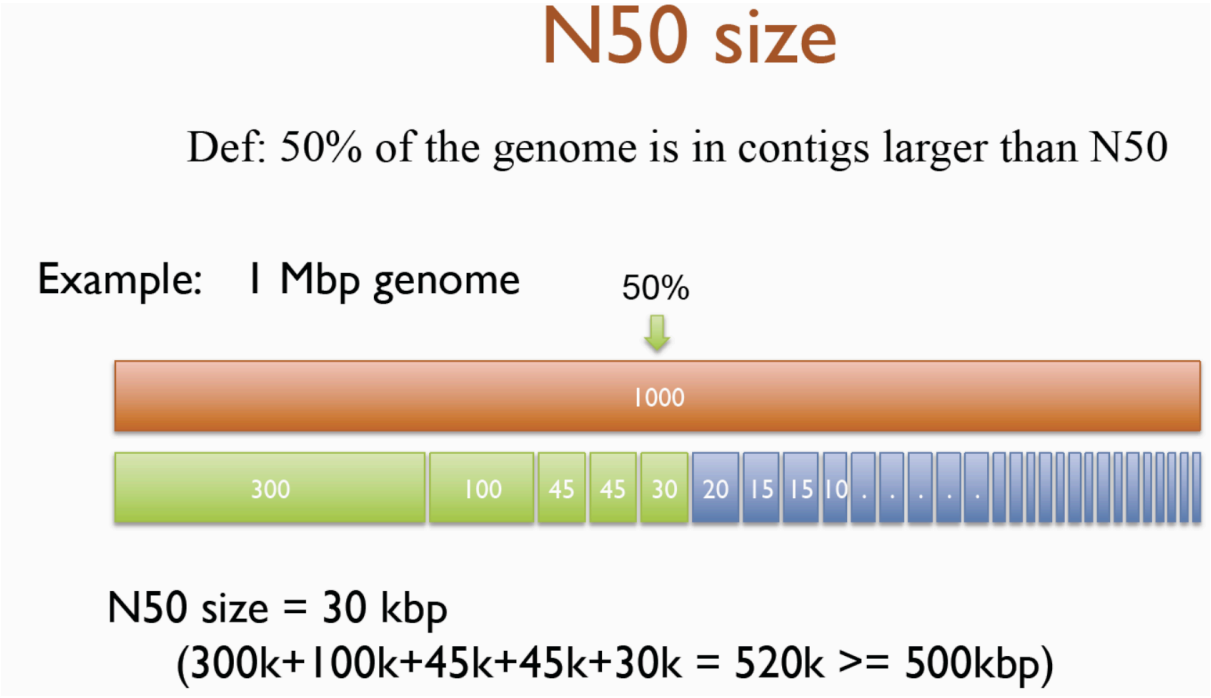

Figure 6. Définition de la N50.

Ces statistiques n'évaluent toutefois que la contiguïté des séquences assemblées, et non leur exactitude !
L'exactitude d'un assemblage peut être évaluée en mappant les lectures originales (« reads ») sur l'assemblage *de novo* pour identifier les régions avec une couverture anormalement élevée (pouvant correspondre à une séquence répétée) ou anormalement faible. Il existe un nombre croissant de programmes pour faciliter les évaluations d'assemblage tels que Quast.
Il peut être utile de tester plusieurs assembleurs *de novo* et de comparer la qualité des assemblages pour choisir le meilleur (cela va dépendre de l'espèce bactérienne).

Annotation du génome bactérien

L'annotation du génome est le processus qui consiste à identifier les caractéristiques biologiquement importantes contenues dans un génome et à joindre des informations descriptives à ces caractéristiques. L'annotation du génome est typiquement l'une des premières étapes appliquées après l'assemblage des séquences. Elle peut être effectuée sur des génomes incomplets (« draft ») ou complets.
Si on veut faire une analyse comparative détaillée d'un groupe de génomes, il y a un risque à utiliser des génomes incomplets puisque les caractéristiques qui existent dans le génome réel peuvent ne pas être présentes dans l'assemblage « draft » (si certaines régions n'ont pas été couvertes par le séquençage ou s'il existe des erreurs d'assemblage).

Les caractéristiques typiquement annotées dans les génomes bactériens sont les gènes codant les protéines, appelés séquences codantes (CDS), et les gènes non codants, tels que l'ARNr et l'ARNt. On peut aussi annoter les pseudogènes, les répétitions palindromiques courtes régulièrement espacées (CRISPR), les transposons et les intégrons. Les génomes bactériens ont une structure et une organisation génétique prévisible qui se prêtent bien à des approches automatisées.

L'annotation du génome peut être divisée en deux tâches principales : l'annotation structurale et l'annotation fonctionnelle.

L'annotation structurale, communément appelée prédiction de gènes, implique l'identification de l'emplacement des gènes codant pour les protéines (CDS) et des gènes non codants (ARNt et ARNr). Les fonctions des gènes codant des protéines sont diverses et difficiles à déterminer. Ces gènes doivent faire l'objet d'une annotation fonctionnelle pour déduire leur fonction biologique probable.

Les programmes de prédiction des gènes bactériens sont de 2 types : intrinsèques (ou *ab initio*), qui tentent d'identifier les séquences codantes en se basant uniquement sur les informations contenues dans le génome ou le contig, et extrinsèques, qui utilisent une base de données de séquences codantes de protéines bactériennes préalablement identifiées et vérifiées pour aider à l'identification des gènes dans un génome nouvellement séquencé.
Aujourd'hui, les outils de recherches de gènes *ab initio* sont les plus utilisés puisqu'ils apprennent à partir des informations génétiques contenues dans le génome cible et parviennent à identifier plus de gènes pour les espèces peu étudiées. En revanche, les outils de recherches de gènes extrinsèques s'appuient sur le contenu génétique généralement issu des organismes modèles comme *E. coli*, qui peuvent être éloignés du génome nouvellement séquencé, ce qui réduit la précision de la prédiction.

La tâche principale des outils de recherches de gènes *ab initio* est de distinguer les séquences codantes des séquences non codantes au sein d'un ensemble de tous les cadres de lecture ouverts (ORF ; séquences génomiques contiguës flanquées de codons stop dans la même phase de lecture), qui peuvent ou non contenir des séquences codantes. Ces programmes ont des précisions prédictives de l'ordre de 95 à 98 %, bien qu'ils puissent avoir de la difficulté à identifier les gènes acquis à partir d'ADN étranger (car leurs caractéristiques divergent du reste du génome) et les gènes des génomes à forte teneur en G+C.

Il est important de savoir que les bases de données de séquences publiques contiennent un nombre substantiel d'ORF codants ou non codants mal annotés. L'annotation automatique à partir de ces bases de données sans vérification supplémentaire peut favoriser la misannotation des gènes et la propagation des erreurs. Il faut donc privilégier la base de données RefSeq du NCBI, qui contient des séquences de référence de haute qualité.

Peut-on prédire *in silico* la résistance aux antibiotiques ?

Bien que l'analyse traditionnelle de la résistance par culture (antibiogramme) soit en général, relativement rapide (2-3j) et peu couteuse, les isolats bactériens peuvent maintenant être examinés *in silico* pour détecter la présence de gènes de résistance immédiatement après le séquençage. Plusieurs logiciels en ligne ont été créés pour identifier les gènes de résistance dans les génomes entiers ou partiellement assemblés (drafts) comme le site ResFinder. Par ailleurs, il est aussi simple de détecter un gène de résistance présent dans des données provenant d'un isolat d'une seule espèce ou que dans un mélange de souches issues d'un échantillon complexe (ex : métagénomes de fèces). Néanmoins la présence d'un gène de résistance ne permet pas de dire s'il est pas, peu ou très exprimé.
Différents travaux sur *E. coli* ont démontré une corrélation de 97,8 % entre les résultats des tests phénotypiques et génotypiques (Tyson 2015). Cependant, il y a encore des limites au WGS pour la prédiction de la résistance aux antibiotiques. En effet, les algorithmes de détection ne sont pas encore capables de prédire le résultat global d'une combinaison de mécanismes de résistance. Par exemple, chez les Entérobactéries, la résistance aux carbapénèmes peut résulter d'une combinaison de perte de porine et d'enzymes bêta-lactamases à spectre étendu (BLSE). Les gènes impliqués dans de nouveaux mécanismes de résistance doivent également être identifiés en laboratoire avant d'être ajoutés aux bases de

données, ce qui représente un travail considérable de mise à jour permanente des bases de données. De plus, il existe aussi des cas où le mécanisme précis ainsi que les gènes impliqués dans la résistance ne sont pas encore connus.

L'épidémiologie et la comparaison de souches

Les séquences des génomes microbiens contiennent une variabilité entre les isolats en raison de l'accumulation de mutations et de la recombinaison entre bactéries. La variabilité de séquence qui existe peut être exploitée pour comparer les souches.
L'analyse comparative des séquences ou des génomes comprend souvent des méthodes phylogénétiques. La phylogénétique est l'étude des relations évolutives entre les organismes. Les arbres phylogénétiques peuvent être déduits en analysant les variations physiques (phénotype) ou les variations génétiques présentent dans les organismes que l'on étudie.

La phylogénétique peut être appliquée pour la comparaison de segments d'ADN qui peuvent être ou non des gènes ou même des génomes complets ; dans ce dernier cas on parle de phylogénomique.

Quelques définitions

- Le polymorphisme nucléotidique ou polymorphisme d'un seul nucléotide (SNP ou SNV, single-nucleotide polymorphism ou single-nucleotide variant) est la variation (polymorphisme) d'une seule base du génome, entre individus d'une même espèce lors de la comparaison de plusieurs génomes.
- Lorsqu'un SNP dans un gène est trouvé dans au moins deux isolats, la version du gène avec le SNP est décrite comme un nouvel allèle du gène.

- Les SNPs peuvent se trouver dans les régions codantes (gènes) ou dans les régions non codantes d'un génome. Les SNP présents dans les gènes peuvent être synonymes s'ils ne modifient pas la séquence codante ou non synonymes si ils modifient la séquence des acides aminés. L'existance de mutations synonymes est due à la redondance du code génétique. Le terme SNP informatif est utilisé pour décrire un SNP qui est partagé par deux souches ou plus dans un alignement et supporte donc le regroupement phylogénétique de deux isolats ou plus. Pour détecter les SNP, une comparaison doit être effectuée par l'alignement de deux ou plusieurs séquences.

- le core génome (génome cœur) : ensemble des gènes communs à toutes les souches d'une même espèce (gènes orthologues). Transmis de manière verticale d'une bactérie à ses descendantes.

- le génome accessoire ou variable : ensemble des gènes présents uniquement dans une souche ou dans seulement quelques souches. Peut être transmis par recombinaison (de manière horizontale) entre 2 bactéries qui peuvent ne pas appartenir à la même espèce. La recombinaison peut perturber les méthodes phylogénétiques. Lorsque des séquences d'ADN similaires sont échangées entre bactéries, si la source de recombinaison est externe à la population étudiée, les régions échangées peuvent contenir un nombre de SNP plus élevé que celui observé ailleurs dans la population. Cela a pour conséquence d'augmenter faussement la distance génétique et pourrait conduire à exclure à tord une transmission directe dans une étude épidémiologique. La gestion de la recombinaison homologue implique souvent l'identification et l'exclusion des signaux phylogénétiques des régions ayant subie une recombinaison. Ainsi les analyses phylogénétiques se focalisent en général sur le core-génome en ne laissant que les régions issues de la descendance verticale.

- le pan-génome : ensemble des gènes du génome cœur et des génomes accessoires de toutes les souches de l'espèce.

La phylogénomique

La phylogénomique est l'application des données du séquençage complet des génomes à l'étude de l'évolution. Les méthodes phylogénétiques traditionnelles utilisent souvent un ou plusieurs gènes pour retracer l'évolution d'un organisme. Le NGS permet d'étendre ces méthodes avec des informations provenant de l'ensemble du génome, ce qui permet la reconstruction de phylogénies à très haute résolution. Cela a des applications dans la surveillance des maladies infectieuses et des épidémies, où les phylogénies reconstituées à partir des génomes peuvent compléter ou remplacer les méthodes de typage traditionnelles comme le MLST, l'électrophorèse sur gel en champ pulsé (PFGE) ou le sérotype.

Les méthodes d'analyse phylogénétique, peuvent être basées soit sur des alignements, soit sur des profils alléliques (méthodes « gène par gène »).

Phylogénies basées sur les alignements et les SNPs

Les phylogénies du génome entier basées sur l'alignement reposent généralement sur la génération puis l'analyse d'un alignement à séquences multiples (MSA), c'est-à-dire un alignement de caractères nucléotidiques ou d'acides aminés où chaque ligne représente un isolat et chaque colonne dans l'alignement représente une homologie hypothétique au niveau d'une base ou d'un acide aminé (Figure 7).

L'identification des SNPs se fait soit par une approche basée sur un génome de référence commun, soit par une approche sans référence. Dans une approche fondée sur une référence, un génome assemblé sert de référence commune à l'identification des SNPs pour tous les génomes que l'on veut comparer. Ceci est souvent réalisé par mapping et recherche de variants.

L'un des moyens les moins intensifs en calcul pour détecter les SNPs est l'utilisation d'une référence. Une fois que les lectures ont été mappées sur un génome de référence, un logiciel peut être utilisé pour identifier les SNPs entre les génomes. Les approches basées sur une référence présentent l'avantage de pouvoir identifier l'emplacement exact et le gène correspondant pour chaque variant par rapport à la référence. Cependant, le choix du génome de référence peut être problématique car s'il est trop éloigné des génomes à comparer cela peut biaiser le résultat (on va rater des SNPs s'ils sont dans des gènes absents du génome de référence).

Il est également possible d'effectuer une analyse de SNPs des génomes après les avoir assemblés *de novo*. Cette approche exige que les contigs assemblés soient annotés, puis que les ORFs soient comparés.

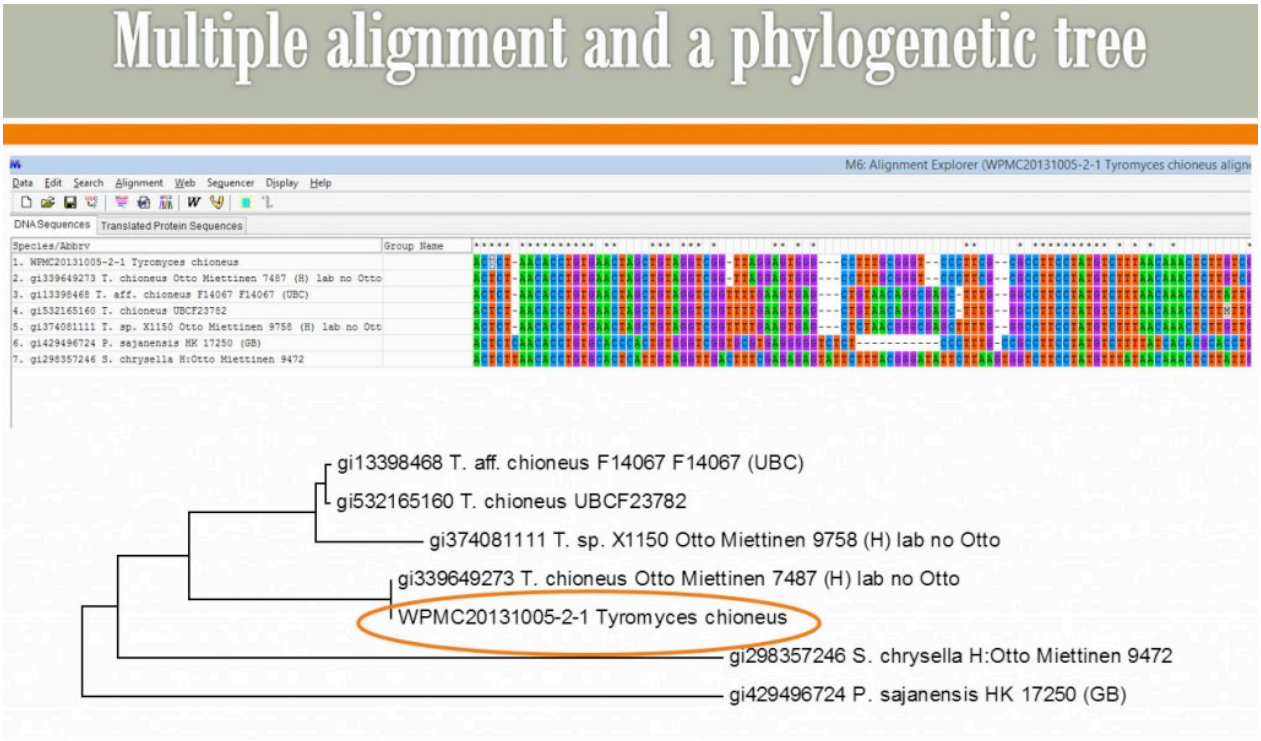

Figure 7. MSA et arbre phylogénétique.

Phylogénies basées sur les profils alléliques (« gènes par gènes »)

Une approche alternative aux méthodes basées sur l'alignement est l'approche gène par gène.

La parenté entre les souches est basée sur une distance entre chaque génome définie par le nombre d'allèles partagés pour chaque gène du schéma de MLST.

Le typage MLST (Multilocus sequence typing) est basé sur la variabilité des allèles de 7 gènes de ménage, c'est à dire des gènes appartenant au core-genome de l'espèce et qui assurent les fonctions indispensables. Le séquençage des 7 gènes est généralement réalisé par la technique traditionnelle à faible débit de Sanger. Les séquences des gènes sont comparées à une base de données établie de manière internationale, telle que pubMLST/BIGSdb, pour déterminer les allèles individuels et ensuite le « sequence type » (ST) de l'isolat. Il est maintenant moins coûteux de séquencer un génome entier par NGS que de séquencer les gènes MLST individuellement ; par conséquent, il existe plusieurs sites web qui sont capables de calculer les ST à partir d'une séquence génomique entière, comme le Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/MLST/). Les isolats ayant le même ST sont définis comme clonaux ayant un ancêtre commun. Le problème de cette technique de typage moléculaire basée sur seulement 7 gènes est qu'elle n'est pas assez résolutive pour comparer des souches génétiquement proches. Les souches pathogènes au sein d'une espèce donnée appartiennent souvent à quelques ST dominants retrouvés dans le monde entier. Il faut donc se baser sur d'avantage de gènes pour comparer des souches lorsqu'elles font partie du même ST. Des schémas de typage basés sur plusieurs centaines de gènes (et non plus seulement 7) ont donc été développés. C'est sur ce principe que repose le core-genome MLST (cgMLST) qui permet la classification rapide d'isolats étroitement apparentés et donne des résultats comparables à la classification générée par des phylogénies basées sur l'alignement.

Comparaison entre les méthodes basées sur l'alignement et les SNPs et les méthodes gènes par gènes

Les méthodes basées sur l'alignement et les SNPs soulèvent une question importante : combien de SNP sont nécessaires pour conclure que deux souches sont liées ? Bien que plusieurs études aient été réalisées à ce sujet, les réponses sont complexes et aucun consensus n'a été établi. Par exemple, les isolats de *L. monocytogenes* épidémiologiquement liés examinés dans une étude rétrospective différaient entre eux par moins de 10 SNPs. Mais dans la même étude, deux isolats récupérés à un jour d'intervalle chez le même patient différaient de 21 SNPs !

Une étude rétrospective de 183 isolats séquencés d'*E. coli* O157:H7 ayant des liens épidémiologiques connus a défini les isolats liés comme ayant moins de cinq différences de SNP, et en moyenne, une seule différence de SNP a été trouvée entre les isolats de patients appartenant à la même famille.

Le nombre de SNPs pour déterminer quels isolats sont suffisamment proches pour pouvoir évoquer une transmission directe entre 2 personnes est donc différent selon les espèces. Par ailleurs, il faut savoir que le nombre de SNPs identifiés lorsque l'on compare 2 génomes va être différent selon, entre autres, la qualité du séquençage des génomes, le génome de référence que l'on aura choisi, les paramètres choisis pour les logiciels d'identification des SNPs. Les méthodes basées sur l'alignement et les SNPs présentent donc des difficultés de reproductibilité.

Les méthodes gènes par gènes telles que le cgMLST, ont une résolution comparable à celle des méthodes basées sur l'alignement et bénéficient d'un schéma standard (et d'une nomenclature de classification associée qui permet la compatibilité des données). Ainsi, les approches gènes par gènes sont particulièrement utiles pour l'intégration des génomes nouvellement séquencés dans un contexte global. Jusqu'à présent, l'utilisation des méthodes de cgMLST a été limitée par le fait que chaque laboratoire a eu tendance à développer son propre schéma de typage plutôt que d'essayer d'utiliser ceux des autres laboratoires. Cependant, il est probable que les méthodes « gène par gène » soient de plus en plus pertinentes si les laboratoires parviennent à s'uniformiser en utilisant des bases de données internationales.

La métagénomique "shotgun" en clinique

Il faut faire la distinction entre l'échantillon (le matériel qui a été prélevé sur le patient) et l'isolat (un organisme qui a été cultivé et isolé de cet échantillon). La technique de shotgun (WGS) se réfère généralement au séquençage du génome d'un seul organisme. Cependant, cet organisme ne représente souvent qu'une petite fraction de la diversité microbienne totale présente dans l'échantillon clinique dont il est issu.

En revanche, la métagénomique shotgun vise le séquençage des échantillons d'une manière globale. Cette approche est particulièrement pertinente pour les scénarios cliniques où l'agent pathogène d'intérêt ne peut être prédit et/ou est fastidieux (c.-à-d. qu'il a des exigences de culture complexes).
L'absence d'étape de culture peut réduire considérablement le temps de traitement de l'échantillon et permettre l'identification d'agents pathogènes nouveaux et/ou non cultivables par les techniques de culture classiques.

Différents échantillons présentent cependant des défis différents. Les sites d'échantillonnage faciles à prélever (p. ex., les fèces et les expectorations) ont généralement un microbiote local, de sorte qu'il peut être difficile de distinguer l'agent pathogène à l'origine de la maladie des microbes colonisateurs. Inversement, les sites qui sont habituellement stériles (p. ex., le liquide céphalorachidien, le liquide pleural) présentent une bien meilleure occasion pour la métagénomique de contribuer aux soins cliniques.
Les données métagénomiques sont plus complexes à analyser que les données WGS d'une seule bactérie et s'appuient sur des outils de calcul sophistiqués. De telles approches sont difficiles à mettre en œuvre, car très exigeantes sur le plan informatique et ne pourront probablement pas être déployées en microbiologie clinique dans un avenir proche. Les plates-formes dans le « cloud » pourraient être une solution dans les laboratoires de diagnostic clinique.

Conclusion

Malgré ses immenses promesses et ses premiers succès, il est difficile de prédire si et quand le WGS modifiera complètement les normes actuelles en microbiologie clinique. Il existe plusieurs difficultés importantes dans sa mise en œuvre en routine pour diagnostiquer et caractériser les infections microbiennes. Il s'agit notamment des coûts actuels du WGS, qui restent loin d'être négligeables (malgré la croyance répandue que les coûts de séquençage ont chuté) ; d'un manque de formation en bioinformatique chez les microbiologistes cliniques et d'une possible résistance culturelle à la bioinformatique ; du manque d'infrastructure informatique nécessaire dans la plupart des hôpitaux ; de l'insuffisance des bases de données internationales de référence en génomique microbienne, nécessaires pour établir des protocoles bioinformatiques efficaces, normalisés et agréés.