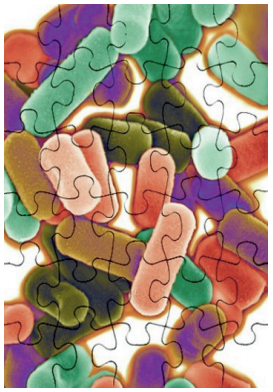
médecine/sciences 2014 ; 30 : 1144-51

Le séquençage haut-débit

Vers un diagnostic basé sur la séquence complète du génome de l'agent infectieux

Christophe Audebert[1,4], David Hot[2,4], Yves Lemoine[3,4], Ségolène Caboche[3,4]

> Lors d'une infection, l'identification précise de l'agent infectieux est de première urgence. Après les méthodes phénotypiques qui restent largement employées, la biologie moléculaire a fait son entrée dans les laboratoires de microbiologie clinique, permettant l'identification du pathogène par la comparaison de son empreinte moléculaire à une banque de données. Le séquençage haut-débit permettrait de dépasser cette seule identification en exploitant la totalité de la connaissance du génome du pathogène. Ce passage de l'empreinte au portrait-robot, soutenu par un nombre croissant d'études préliminaires, permettrait une meilleure adaptation du traitement aux spécificités du pathogène. Cependant, plusieurs obstacles doivent être franchis afin que l'exploitation des données de séquençage haut-débit devienne une réalité. <



[1] Gènes Diffusion, Douai, France ;
[2] U1019, UMR8204, Université de Lille, France ;
[3] FRE3642, Université de Lille, France ;
[4] Pegase-Biosciences, Institut Pasteur de Lille,
1, rue du professeur Calmette,
59019 Lille, France.
segolene.caboche@pasteur-lille.fr

Séquençage haut-débit

Principes généraux du séquençage multi-parallélisé

La notion de séquençage haut-débit est apparue en 2000 avec la 1re génération, aujourd'hui oubliée, de machines proposées par la société Lynx Therapeutics. Cinq ans après, avec l'arrivée des séquenceurs de 2e génération, le séquençage multi-parallélisé n'aura de cesse d'être la source de plus en plus d'applications dans les laboratoires de recherche. Depuis, le nombre de génomes complètement séquencés et accessibles dans les banques de données a cru de façon exponentielle. Cette diffusion technologique s'est accompagnée d'une baisse brutale du coût associé à l'opération de séquençage : l'analyse du 1er génome humain, accomplie en 2001, aura coûté environ trois milliards de dollars ; la même analyse coûte un millier de dollars 13 ans plus tard.

En 2014 cohabitent deux générations de séquenceurs : les séquenceurs de 2e génération, dont une partie est déclinée en séquenceurs de paillasse aux débits généralement suffisants pour les applications en microbiologie ; et les séquenceurs de 3e génération. Ces derniers, plus sensibles, permettent de s'affranchir de l'étape d'amplification clonale, caractéristique qui les rend plus rapides que leurs aînés.

Le séquençage haut-débit a pour principe de base la parallélisation de réactions permettant le séquençage de courtes lectures d'une librairie (*reads*, voir *Glossaire*). Il permet d'obtenir la séquence d'un génome complet (séquençage pangénomique, voir glossaire) ou de parties de génome bornées par PCR (*polymerase chain reaction*, séquençage ciblé).

Les applications liées au séquençage haut-débit font intervenir une phase de biologie humide et une phase de biologie sèche *(Figure 1)*. En laboratoire de biologie moléculaire, la phase de biologie humide commence par une extraction et une purification de l'ADN du pathogène à séquencer. Le pathogène pourra être isolé préalablement ou pas ; la modalité envisagée aura un impact principalement sur la puissance de séquençage nécessaire et les différents types d'analyses bioinformatiques à réaliser. Par exemple, séquencer un échantillon complexe constitué d'ADN de plusieurs microorganismes mélangé à une majorité d'ADN humain nécessitera plus de puissance de séquençage que le séquençage du génome d'un microorganisme isolé.

Le séquençage haut-débit (2e génération) repose ensuite sur trois étapes successives *(Figure 1)* :

DOI : 10.1051/medsci/20143012018

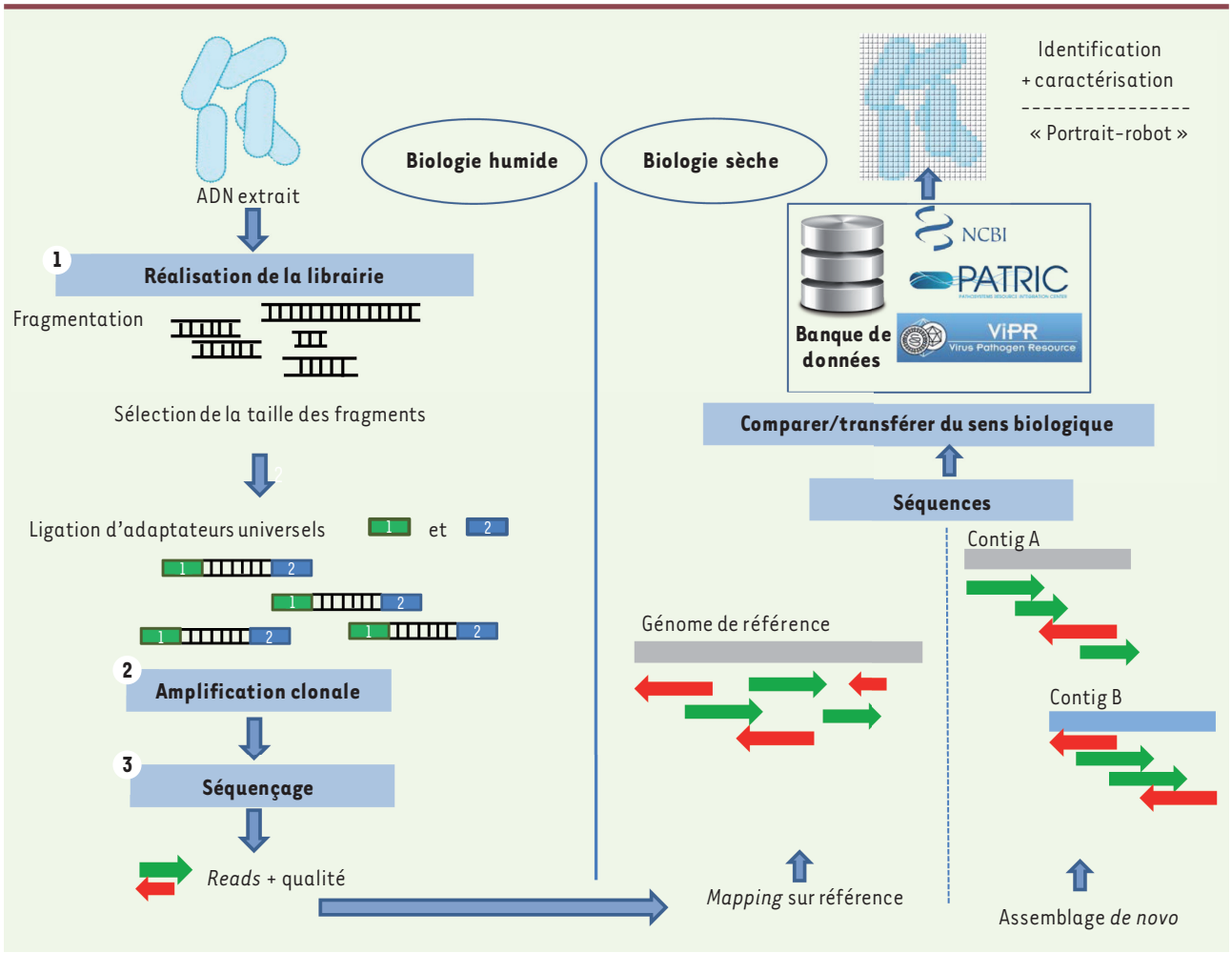


Figure 1. *Étapes successives de biologie humide et sèche aboutissant à l'établissement du portrait-robot de l'agent infectieux suspecté.*

• La réalisation d'une librairie : l'ADN extrait est le plus souvent fragmenté, suite à un traitement enzymatique, avant que ces fragments soient ligaturés avec des adaptateurs universels.

• Une amplification clonale moléculaire, permettant de multiplier la quantité de molécules matricielles ; elle est généralement réalisée par PCR, où des amorces ciblent les adaptateurs universels bornant les fragments de la librairie.

• Le séquençage multi-parallélisé : les fragments matriciels issus de l'étape précédente sont individualisés pour être lus afin de produire autant de séquences nucléotidiques courtes.

Les modes d'amplification clonale peuvent varier selon les fournisseurs de séquenceurs, et font aujourd'hui l'objet d'une simplification des procédures à l'aide d'automates dédiés. Les technologies de séquençage se distinguent principalement par leurs modes de détection produisant un signal qui est fonction des bases polymérisées lors de l'opération de séquençage multi-parallélisé. Ainsi, la société Roche s'est appuyée sur la méthode du pyroséquençage (*voir glossaire*), la société Illumina sur le séquençage par synthèse de nucléotides fluorescents « terminateurs » réversibles, et la société Ion Torrent sur la lecture du pH suite à la libération d'un ion H^+ après la polymérisation d'un nucléotide natif.

La phase de biologie sèche a pour but de générer de l'information biologique pertinente à partir des *reads* (par exemple, l'identification de gènes présents, de variations par rapport à une référence, etc.) à l'aide d'outils bioinformatiques *(Figure 1)*. Dans le but d'obtenir des séquences plus longues à partir des *reads*, deux approches peuvent être utilisées : le *mapping* sur référence et l'assemblage *de novo*. Le *mapping* sur référence consiste à repositionner les *reads* sur un génome de référence. Cette référence est le plus souvent constituée par le génome de l'organisme séquencé ou par un génome proche, complètement annoté et disponible dans les banques de données publiques. Dans le cas d'un séquençage à partir d'un prélèvement, la référence n'est pas connue *a priori*. Cependant, il est possible d'identifier, à l'issue du séquençage, un ou plusieurs génomes qui serviront de références en

exploitant l'homologie entre les *reads* et les génomes disponibles dans les banques publiques. Le *mapping* sur référence permet, par comparaison, de reconstruire la séquence génomique de l'organisme séquencé et d'identifier les différences entre la séquence obtenue et le génome de référence. Dans le cas où il n'existe pas de génome proche de celui de l'organisme étudié dans les banques de données, l'assemblage *de novo* est alors utilisé. Il permet de construire, sans *a priori*, des séquences plus longues (*contigs*) à partir des *reads*. Une fois les séquences obtenues par *mapping* ou assemblage *de novo*, l'étape suivante est d'en extraire les informations biologiques pertinentes en les comparant aux séquences annotées contenues dans les banques de données.

Le *mapping* sur référence nécessite bien souvent moins de puissance de séquençage et permet, à la fois, de reconstruire le génome séquencé en bénéficiant d'un modèle proche constitué par le génome de référence, d'annoter aisément le génome séquencé, et de le comparer à sa référence (mutations, gènes/plasmides présents, tronqués ou absents).

Laboratoire de microbiologie : gagner du temps pour un traitement optimal

Les laboratoires de microbiologie clinique ont pour mission de fournir rapidement des éléments de caractérisation du (ou des) microorganisme(s) à l'origine d'une maladie infectieuse. Les données du tableau clinique orientent le choix des prélèvements ; les données issues des analyses biologiques complètent ces informations pour permettre de choisir la thérapeutique optimale.

L'identification du (ou des) pathogène(s), cause principale ou aggravante de la maladie infectieuse, est généralement réalisée après isolement, par des méthodes phénotypiques (galerie API[1]). Cependant, les méthodes de biologie moléculaire, telles que les méthodes de diagnostic par PCR en temps réel ciblant un locus spécifique du microorganisme testé, permettent l'identification sans avoir recours à un isolement préalable. Ces méthodes de biologie moléculaire sont très sensibles, standardisables et d'une mise en place assez aisée. Une fois l'identification réalisée, d'autres méthodes peuvent être utilisées dans le but de réaliser une discrimination infra-spécifique. Ces méthodes moléculaires exploitent les locus polymorphes de l'ADN microbien, principe à la base d'outils de génotypage mené par séquençage Sanger ou par la mesure de la taille des produits PCR générés. C'est, par exemple, le cas des MLST (*multilocus sequence typing*) ou de l'analyse du nombre de séquences répétées sur plusieurs locus (MLVA, pour *multiple loci variable number tandem repeat analysis*). Ces analyses visent à attribuer un code barre au microorganisme analysé ; ce même code barre permettra d'interroger une banque d'empreintes pour aboutir à une identification clonale.

Depuis peu, la spectrométrie de masse permet, par l'analyse des empreintes spectrales des biomarqueurs, une identification rapide par comparaison avec une banque d'empreintes [1]. La spectrométrie de masse a pour seul objectif ici de dresser un profil protéique caractéristique d'un microorganisme, et non de déterminer la nature protéique de ces biomarqueurs. Ces méthodes d'identification sont basées sur la comparaison des empreintes (détection de molécules spécifiques dans le cas de la spectrométrie de masse, identification de caractéristiques génétiques pour les méthodes de biologie moléculaire) du microorganisme suspecté pathogène avec celles d'une banque privée ou publique d'empreintes. La comparaison permet l'identification ; l'analyse est assez simple à réaliser ; la force du système tient à la quantité d'empreintes associées à autant de taxons et à la richesse des banques d'empreintes.

À l'instar de la spectrométrie de masse qui, initialement, nécessitait un équipement lourd dédié à la recherche de pointe, le séquençage haut-débit est devenu un outil de plus en plus accessible pour des applications de routine. Les séquenceurs de paillasse ciblent le marché des laboratoires d'analyses biologiques. Ainsi, Ion Torrent (Life Technologies) a lancé son séquenceur haut-débit judicieusement nommé PGM (*personal genome machine*), devançant de peu un autre acteur du domaine, Illumina. Ce dernier, avec le MiSeqDx, devient le 1e fournisseur d'un séquenceur agréé par la FDA (*Food and drug administration*) pour des applications de diagnostics *in vitro*. Si la recherche de mutations liées à des cellules cancéreuses est actuellement l'application très largement prépondérante, la technologie du séquençage haut-débit est envisagée comme outil d'investigation en cas d'infections ne pouvant pas être élucidées par les méthodes usuelles [2].

Une identification basée sur le génome complet : vers un changement de paradigme ?

Réductionnisme contre complétude : empreinte contre portrait-robot

L'approche réductrice faisant intervenir des empreintes moléculaires est une méthode imparfaite, mais néanmoins efficace dans une grande majorité des cas. Par principe, ces méthodes d'identification ne tiennent pas compte de la plasticité des génomes microbiens [3]. Par exemple, les intégrons, qui peuvent être considérés comme un système de capture de gènes, sont répandus chez les bactéries à Gram négatif ; ces éléments jouent un rôle majeur dans les phénomènes de résistance à de multiples antibiotiques[2]. Ainsi, derrière un même

[1] Il s'agit d'un système standardisé de microtubes contenant des tests biochimiques miniaturisés ainsi qu'une base de données pour l'identification des bactéries.

[2] Voir le numéro thématique de *m/s* sur « Résistance aux antibiotiques : un enjeu international » paru en novembre 2010.

« code barre » issu d'analyses MLST/MLVA peuvent se cacher plusieurs souches aux profils de résistances antimicrobiens divers [4].

Palliant les défauts des conclusions déduites d'une identification incomplète, le séquençage haut-débit a fait l'objet de plusieurs études de preuves de concept permettant d'envisager son emploi au sein des laboratoires de microbiologie clinique *(Tableau I)*. Si certaines études ont montré l'implémentation efficace sur plate-forme de séquençage haut-débit des techniques d'empreintes moléculaires [5], dans le cadre de la microbiologie clinique, il semble que le séquençage pangénomique soit une stratégie qui permette de caractériser finement l'entité séquencée, et ce pour plusieurs raisons :

- La possibilité, selon certains modes d'utilisation, d'effectuer le séquençage haut-débit sans isolement préalable.
- L'élucidation des cas de co-infections ; le séquençage haut-débit permet de séquencer le microorganisme prédominant ainsi que les microorganismes minoritaires, rendant possible l'identification et la caractérisation de variants rares.
- L'amélioration de la sensibilité.
- La détection des mutations référencées et associées à des systèmes de résistance/virulence.
- La détection des nouvelles mutations, signatures de souches émergentes dont l'association ou non à de nouveaux mécanismes de résistance/virulence pourra être analysée.

L'un des principaux avantages de cette méthodologie découle du fait que le praticien, au-delà de la simple connaissance taxonomique de l'agent infectieux, peut faire un diagnostic sur la base de la connaissance quasi exhaustive du génome de ce même agent pathogène. Cette connaissance lui permet de réaliser des antibiogrammes *in silico*, d'estimer le pouvoir pathogène d'une souche et, *in fine*, de proposer un traitement ciblé. Cette stratégie peut contribuer à éviter l'emploi, sur une trop longue période, d'antibiotiques à large spectre favorisant l'émergence de souches résistantes.

Le séquençage haut-débit permet de caractériser des génomes complets associés à une infection, ce qui permet d'envisager une évolution des pratiques hospitalières. En effet, actuellement, les informations issues de l'antibiogramme permettent d'adapter le traitement. La connaissance des communautés de gènes impliquées dans une infection particulière permettrait d'envisager de nouvelles approches thérapeutiques certainement plus ciblées, et donc vraisemblablement plus efficaces. Même si les informations génomiques sont, dans la plupart des cas, inutiles, et que la seule identification phénotypique reste suffisante pour la bonne prise en charge du patient, il existe néanmoins certains cas d'infections où la connaissance du génome permettra l'accès à un traitement efficace : par exemple, lorsque l'infection est complexe (co-infection), lorsque des microorganismes multirésistants sont impliqués, ou encore lorsque le microorganisme impliqué est difficilement cultivable.

La révolution génomique appliquée au diagnostic d'agent infectieux : une utopie ou une réalité en marche ?

Le séquençage haut-débit est un moyen d'accéder à une vision complète (généralement supérieure à 99 %) d'un génome. Pour une utilisation de routine, la tendance est à la simplification des procédures techniques dont le temps de réalisation se réduit à seulement quelques heures. Le séquenceur délivre des données brutes, ainsi que des morceaux plus ou moins fidèles d'un puzzle comme autant d'éléments constitutifs du génome séquencé, qu'il faudra reconstituer à l'aide d'outils bioinformatiques. L'exploitation rapide, rationnelle, et la plus automatisée possible, est l'enjeu de l'étape de bioanalyse. Contrairement aux méthodes analytiques d'ores et déjà utilisées en routine, le séquençage haut-débit a le défaut d'être une technologie jeune et dont les procédés d'exploitation des données ne font pas, à ce jour, consensus au sein la communauté scientifique.

Les limites et défis actuels que pose l'utilisation du séquençage haut-débit dans un contexte clinique se situent donc majoritairement au niveau de la partie analytique. En effet, les outils disponibles aujourd'hui sont difficilement accessibles aux utilisateurs non experts : l'installation des *pipelines* (*voir glossaire*) est difficile, nécessitant des compétences informatiques, et les outils manquent d'interfaces graphiques intuitives permettant une utilisation facile [6]. Une réflexion sur la centralisation des données au sein de banques dédiées et sur la standardisation des formats doit également être menée, afin de regrouper efficacement les données pertinentes issues du séquençage, et améliorer nos connaissances [7]. Enfin, une fois les premières limites dépassées, les *pipelines* ergonomiques utilisant les banques de données nettoyées devront être intensivement testés sur de larges jeux de données cliniques avant d'être validés et adoptés dans les laboratoires cliniques [6]. Des difficultés demeurent quant à la généralisation de l'approche du séquençage haut-débit dans l'élucidation de maladies infectieuses, excluant dans l'immédiat une utilisation en routine. Malgré tout, pour des applications particulières, telles que la prise en charge d'infections à bactéries multirésistantes, le recours au séquençage haut-débit est envisagé à large échelle [8, 9].

Du concept à l'utilisation de routine

À l'occasion de la prise en charge d'un patient souffrant d'une maladie infectieuse, un tableau clinique est dressé qui ouvre le champ à des hypothèses étiologiques. S'appuyant sur ces premières constatations, des échantillons sont prélevés. Un nombre croissant d'études *(Tableau I)* utilisent le séquençage haut-débit avec pour objectif d'exploiter ces données pour les intégrer au sein d'un arbre décisionnel clinique. Ainsi, l'étude de Sherry *et al*. en 2013 [19], a démontré que

Utilisation du séquençage haut-débit		Plate-forme de séquençage	Diagnostic par connaissance du génome pathogène	Pathogènes	Réf
Avec isolement	Études épidémiologiques	MiSeq (Illumina)	Étude de faisabilité, en contexte hospitalier, où un séquenceur haut-débit de paillasse a permis d'augmenter la résolution par rapport à des stratégies moléculaires usuelles	*Staphylococcus aureus* et *Clostridium difficile*	[18]
		PGM (Ion Torrent)	Projet pilote dans lequel un séquenceur haut-débit de paillasse a mis en évidence des souches multirésistantes	*Escherichia coli* multirésistante	[19]
	Évolution des génomes	GA IIx (Illumina)	Travaux précurseurs (2010) qui ont montré une meilleure résolution du séquençage haut-débit pour envisager le suivi de la transmission d'un agent pathogène de personne à personne	*Staphylococcus aureus* résistant à la méticilline, isolats cliniques divers	[20]
		GA IIx (Illumina)	La technologie est employée pour élucider le mécanisme d'apparition de résistances virales	Virus de la grippe A H7N7	[21]
Sans isolement	Métagénomique clinique	PGM (Ion Torrent)	Preuve de principe : intégration au niveau du diagnostic de données métagénomiques	Microbiote pulmonaire	[22]
		HiSeq2500 et MiSeq (Illumina)	Afin d'éviter un isolement préalable, cette étude suggère l'emploi de méthodes métagénomiques	*Shiga-toxigenic Escherichia coli*	[13]
	Microbiologie unicellulaire	GA IIx (Illumina)	Dans un contexte clinique, première démonstration de l'efficacité d'une méthode de séquençage complet d'un génome issu de la capture d'une seule cellule	*Porphyromonas gingivalis*	[23]
		GAIIx et HiSeq (Illumina)	Isolement d'une cellule par immunocapture suivi d'une amplification pangénomique avant séquençage complet	*Chlamydia trachomatis*	[11]

Tableau I. ***Études pilotes utilisant le séquençage haut-débit dans le cadre de l'identification et/ou la caractérisation fine d'un agent infectieux.***

cinq jours sont nécessaires pour l'obtention et l'analyse des données, et qu'un budget en réactifs de 300 US$ par souche permettait de réaliser le séquençage d'*E. coli* multirésistantes. L'exploitation de ces données de séquençage a permis de conforter les résultats obtenus par antibiogramme mais, au-delà, ces données ont pu discriminer deux souches dont les antibiogrammes étaient identiques, mais qui possédaient des gènes de résistance différents.

Si, actuellement, il semble que l'isolement du pathogène soit le plus souvent le préalable à une extraction d'ADN, puis un séquençage pangénomique, plusieurs études laissent entrevoir la perspective d'un séquençage sans isolement préalable. Cette dernière approche permettrait de réduire encore le temps d'analyse tout en s'affranchissant des diverses étapes de culture

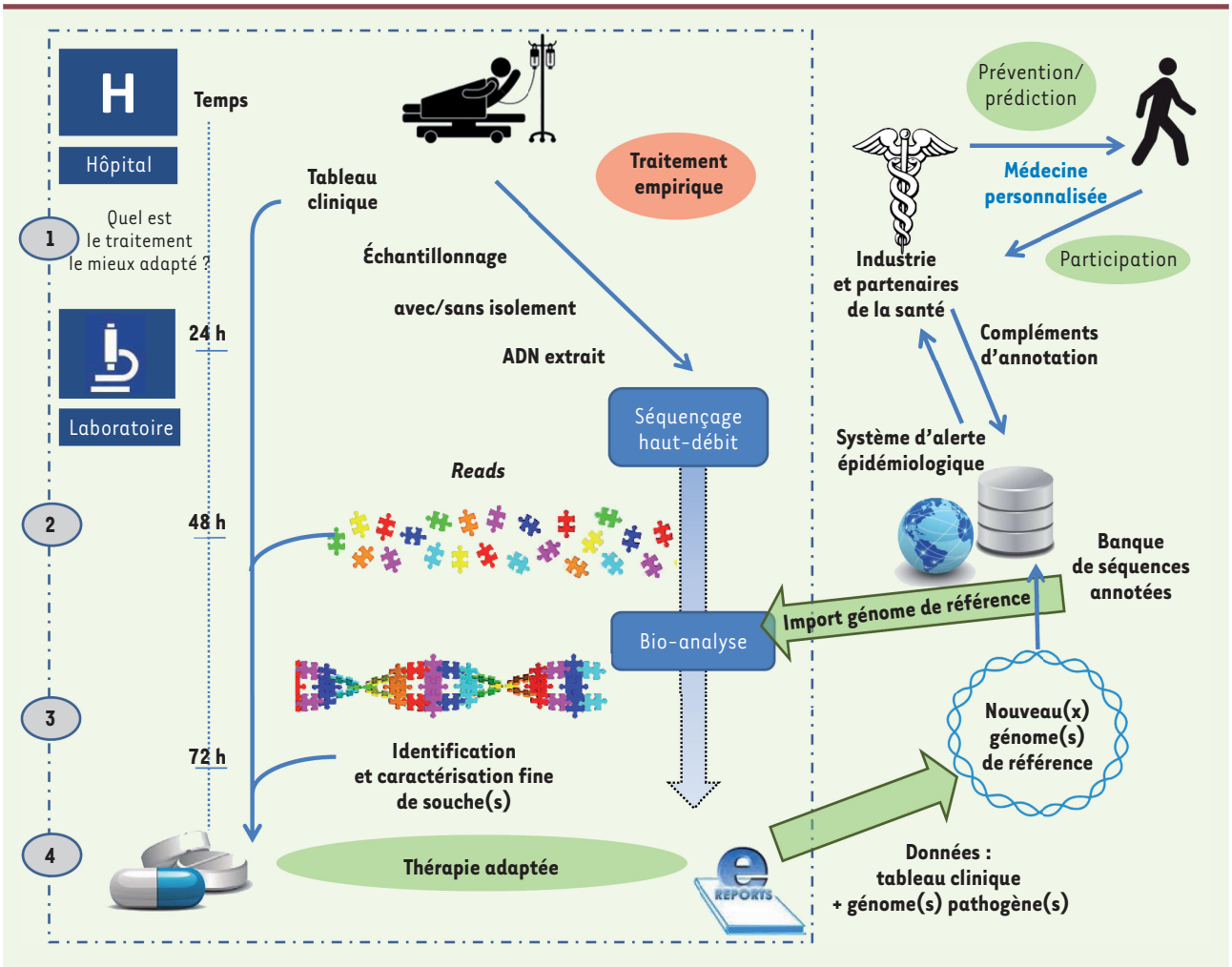


Figure 2. *Une organisation possible de la gestion et de l'exploitation des données issues du séquençage haut-débit : vers un diagnostic génomique.*

(isolement, antibiogramme) ; se passer de ces étapes est un défi entrain d'être relevé par les méthodes de séquençage haut-débit [10]. Des stratégies, telles que le séquençage haut-débit du génome de cellules isolées par des méthodes d'immunocapture [11], ou encore les approches métagénomiques (*voir glossaire*) [12, 24], sont envisagées. Cette dernière approche vise à séquencer tous les ADN microbiens présents afin de les identifier, voire de les caractériser. Si la technique reste encore onéreuse et assez ardue à implémenter au niveau d'un laboratoire de microbiologie clinique, elle permet : (1) de connaître la communauté des gènes impliqués dans une pathologie infectieuse donnée, (2) d'échapper à la contrainte de la mise en culture, et (3) d'éviter l'emploi de stratégies de biologie moléculaire avec *a priori* (emploi d'amorces ou sondes spécifiques). Ainsi, Loman *et al.* [13] ont présenté une approche visant à identifier et caractériser par séquençage direct, à partir d'une matrice complexe (échantillons de fèces), les souches impliquées dans un épisode infectieux à *Escherichia coli* entérohémorragiques survenu en Allemagne en 2011. En premier lieu, cette approche a réussi à identifier/caractériser la souche principale à l'origine supposée de l'infection, une *E. coli* O104:H4, mais aussi, pour un échantillon, une souche de *Clostridium concisus* non détectée par les approches conventionnelles. Avec isolement (pour ce qui concerne une implémentation rapide), ou sans isolement préalable (qui demandera encore de simplifier les méthodologies à employer), le séquençage haut-débit apparaît comme une approche possible, applicable en routine au sein des laboratoires de microbiologie. Son exploitation, pour être efficace, peut être envisagée en deux étapes *(Figure 2)* :

- Une phase où l'urgence prime ; l'information pertinente est celle nécessaire et suffisante à la prise en charge optimale du patient.
- Une phase de décentralisation des données : les séquences sont hébergées dans une banque de données et analysées plus finement pour être exploitées ultérieurement.

GLOSSAIRE

Métagénomique : un procédé méthodologique qui consiste à exploiter l'information du contenu génétique d'un échantillon issu d'un environnement complexe pour extrapoler la diversité microbienne et l'abondance relative des entités taxonomiques constitutives de ce même échantillon.

***Pipeline* :** procédé informatique qui consiste à lier et automatiser les diverses étapes visant à l'analyse de données.

Pyroséquençage : procédé de détection alternatif à la méthode Sanger : en opposition à cette dernière dans laquelle les nucléotides correspondant à la matrice sont séquentiellement incorporés. Les nucléotides polymérisés seront détectés suite à l'intervention d'enzymes aboutissant à l émission d'un signal lumineux.

***Reads* :** cette terminologie anglophone se réfère aux données sorties des séquenceurs haut-débit qui représentent des lectures nucléotidiques de fragments, dont la longueur et la fidélité sont fonction de la technologie de séquençage.

Séquençage haut-débit : de nombreuses appellations sont liées à cette technologie. Parmi celles-ci, le séquençage multi-parallélisé est la plus appropriée. Cette dénomination est un équivalent francophone des terminologies anglophones telles que NGS (*next-generation sequencing*), HTS (*high-throughput sequencing*), et *deep sequencing*.

Séquençage pangénomique : séquençage intégral d'un génome ayant pour équivalent anglophone WGS (*whole genome sequencing*) ; il s'oppose au séquençage ciblé.

Un temps pour l'urgence

Les séquenceurs de paillasse proportionnés pour séquencer des génomes microbiens en quelques heures ont une capacité suffisante et représentent un investissement acceptable pour un laboratoire de microbiologie. Actuellement, entre 48 et 72 heures sont nécessaires pour obtenir des *reads*. Cette information de base doit être prise en charge à l'aide d'un logiciel intégré qui permettra d'aboutir à l'identification d'un génome et, surtout, de mettre en évidence l'annotation pertinente pour étayer un diagnostic et choisir un protocole thérapeutique le plus efficace possible.

Un temps pour l'expertise poussée et l'exploitation des données

Par principe, le séquençage haut-débit peut permettre le décodage de l'intégralité d'un génome pathogène impliqué dans un épisode infectieux, des épidémies ou des attaques bioterroristes [14]. L'homogénéité du type de données (séquences ADN) permet une exploitation par un nombre considérable d'experts. L'épidémie allemande à *E. coli* O104:H4 de l'été 2011 est le premier exemple d'une mise en commun de données de séquençage exploitées par une large communauté scientifique, notamment grâce à un site internet dédié, l'*E. coli* O104:H4 *genome analysis crowdsourcing website* [15]. Une analyse bioinformatique fine du génome complet permet d'acquérir de nouvelles connaissances sur l'organisme étudié et de compléter les banques de données utilisées lors de la première phase d'urgence. Des développements d'outils bioinformatiques spécifiques et standardisés en fonction des applications peuvent être menés également durant cette phase et utilisés ultérieurement durant la phase d'urgence. Les connaissances acquises (par exemple sur les facteurs de virulence et les gènes de résistances) peuvent être exploitées par les industries et partenaires de la santé [16]. Cette exploration peut être utilisée pour la prédiction et la prévention des maladies infectieuses (par exemple en adaptant une stratégie vaccinale). Enfin, cette organisation permet d'optimiser l'organisation épidémiologique et en santé publique car chaque souche peut être caractérisée individuellement ; la propagation du pathogène peut être retracée géographiquement et en temps quasi réel.

Conclusions

La question n'est pas tant de savoir si le séquençage haut-débit va être utilisé au niveau des laboratoires de microbiologie clinique mais quand. Depuis 2011 et l'arrivée des séquenceurs haut-débit de paillasse qui démocratisent cette technologie, de nombreux auteurs ont fait évoluer cette question. L'intérêt indéniable de cette technologie réside dans la mise à disposition d'un portrait-robot de l'agent pathogène incriminé avec une définition incomparable. Cette connaissance quasi exhaustive du génome microbien à la source de l'infection permettra au clinicien de bâtir une stratégie thérapeutique différenciée, adaptée et économe.

Le passage de la preuve de concept à l'application de routine comporte encore quelques verrous. En effet, au niveau de l'implémentation technique, la tendance est à la simplification grâce à l'optimisation des protocoles opératoires et à la mise à disposition d'automates dédiés. Il n'en va pas de même en ce qui concerne l'exploitation des données brutes. Le défi est à relever par les bioinformaticiens qui devront concevoir et développer des logiciels permettant l'interprétation automatisée des données brutes produites par les séquenceurs. Si, dans la plupart des cas, les méthodes actuelles d'identification phénotypique sont amplement satisfaisantes, il en est d'autres (mise en culture délicate, importance de la connaissance de la variabilité génomique infra-spécifique) pour lesquelles le séquençage haut-débit, avec la connaissance exhaustive du génome pathogène qu'il permet, peut représenter une approche nouvelle.

Actuellement, l'utilisation dans le cadre du diagnostic précoce de cancers est l'une des applications phare du séquençage haut-débit pour le développement de la médecine dite personnalisée. Néanmoins, l'accessibilité de données génomiques du pathogène et de son hôte laisse entrevoir qu'une médecine personnalisée

(mais aussi participative, prédictive et préventive[3]) est applicable aux maladies infectieuses, une nouvelle médecine au sein de laquelle le séquençage haut-débit tient un rôle central [17]. ◊

SUMMARY

High-throughput sequencing: towards a genome-based diagnosis in infectious diseases

During a pathogen outbreak, the emergency resides in the identification and characterization of the infectious agent. In addition to the traditional phenotypic methods which are still widely used, the molecular biology is nowadays a common approach of clinical microbiology labs and the pathogen can be identified by comparing its molecular fingerprint to a data-bank. High-throughput sequencing should allow overcoming this single identification to exploit the whole information encoded in the pathogen genome. This evolution, supported by an increasing number of proof-of-concept studies, should result in moving from detection through fingerprints to the use of the pathogen whole genome; this forensic profile should allow the adaptation of the treatment to the pathogen specificities. From concept to routine use, many parameters need to be considered to promote high-throughput sequencing as a powerful tool to help physicians and clinicians in microbiological investigations. ◊

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Carbonnelle E, Nassif X. Utilisation en routine du MALDI-TOF-MS pour l'identification des pathogènes en microbiologie médicale. *Med Sci (Paris)* 2011 ; 27 : 882-8.
2. Caboche S, Audebert C, Hot D. High-throughput sequencing, a versatile weapon to supporte genome-based diagnosis in infectious diseases: applications to clinical bacteriology. *Pathogens* 2014 ; 3 : 258-79.
3. Ehrlich GD, Post JC. The time is now for gene- and genome-based bacterial diagnostics : You say you want a revolution. *JAMA Intern Med* 2013 ; 173 : 1405-6.
4. Roetzer A, Diel R, Kohl TA, *et al.* Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 2013 ; 10 : e1001387.
5. Boers SA, van der Reijden WA, Jansen R. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One* 2012 ; 7:e39630.
6. Nocq J, Celton M, Gendron P, *et al.* Harnessing virtual machines to simplify next-generation DNA sequencing analysis. *Bioinformatics* 2013 ; 29 : 2075-83.
7. Carrico JA, Sabat AJ, Friedrich AW, Ramirez M. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. *Euro Surveill* 2013 ; 18 : 20382.
8. Clark TG, Mallard K, Coll F, *et al.* Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLoS One* 2013 ; 8 : e83012.
9. Daum LT, Fischer GW, Sromek J, *et al.* Characterization of multi-drug resistant *Mycobacterium tuberculosis* from immigrants residing in the USA using Ion Torrent full-gene sequencing. *Epidemiol Infect* 2014 ; 142 : 1328-33.
10. Oyola SO, Gu Y, Manske M, *et al.* Efficient depletion of host DNA contamination in malaria clinical sequencing. *J Clin Microbiol* 2013 ; 51 : 745-51.
11. Seth-Smith HM, Harris SR, Skilton RJ, *et al.* Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. *Genome Res* 2013 ; 23 : 855-66.
12. Pallen MJ. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology* 2014 ; 27 : 1-7.
13. Loman NJ, Constantinidou C, Christner M, *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 2013 ; 309 : 1502-10.
14. Boissinot M, Bergeron MG. Génomique et bioterrorisme. *Med Sci (Paris)* 2003 19 : 967-71.
15. Pritchard L, Holden NJ, Bielaszewska M, *et al.* Alignment-free design of highly discriminatory diagnostic primer sets for *Escherichia coli* O104:H4 outbreak strains. *PLoS One* 2012 ; 7 : e34498.
16. Chen C, Zhang W, Zheng H, *et al.* Minimum core genome sequence typing of bacterial pathogens : a unified approach for clinical and public health microbiology. *J Clin Microbiol* 2013 ; 51 : 2582-91.
17. Bengoechea JA. Infection systems biology: from reactive to proactive (P4) medicine. *Int Microbiol* 2012 ; 15 : 55-60.
18. Eyre DW, Golubchik T, Gordon NC, *et al.* A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open* 2012 ; 2 : e001124.
19. Sherry NL, Porter JL, Seemann T, *et al.* Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. *J Clin Microbiol* 2013 ; 51 : 1396-1401.
20. Harris SR, Feil EJ, Holden MT, *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010 ; 327 : 469-74.
21. Jonges M, Welkers MR, Jeeninga RE, *et al.* Emergence of the virulence-associated PB2 E627K substitution in a fatal human case of highly pathogenic avian influenza virus A(H7N7) infection as determined by Illumina ultra-deep sequencing. *J Virol* 2014 ; 88 : 1694-1702.
22. Salipante SJ, Sengupta DJ, Rosenthal C, *et al.* Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS One* 2013 ; 8 : e65226.
23. McLean JS, Lombardo MJ, Ziegler MG, *et al.* Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Res* 2013 ; 23 : 867-77.
24. Bernardo P, Albina E, Eloit M, Roumagnac P. Métagénomique virale et pathologie. *Med Sci (Paris)* 2013 ; 29 : 501-8.

[3] Le concept de médecine P4 a été introduit par Leroy Hood (*P4 Medicine: Personalized, Predictive, Preventive, Participatory. A change of view that changes everything*).

TIRÉS À PART

S. Caboche