Immuno-analyse et biologie spécialisée (2008) **23**, 260—279

Disponible en ligne sur www.sciencedirect.com

ScienceDirect

journal homepage: http://france.elsevier.com/direct/IMMBIO/

REVUES GÉNÉRALES ET ANALYSES PROSPECTIVES

Les techniques de séquençage de l'ADN : une révolution en marche. Première partie

DNA sequencing technologies: A revolution in motion. Part one

J. Lamoril[a,*], N. Ameziane[b], J.-C. Deybach[a], P. Bouizegarène[a], M. Bogard[c]

[a] *Laboratoire de biochimie et génétique moléculaire, hôpital Louis-Mourier, 178, rue des Renouillers, 92700 Colombes, France*
[b] *Laboratoire de biologie polyvalente, centre hospitalier de Sens, 89100 Sens, France*
[c] *Laboratoire de biochimie et biologie moléculaire, centre hospitalier de Meaux, 77100 Meaux, France*

Reçu le 17 juillet 2008 ; accepté le 28 juillet 2008
Disponible sur Internet le 23 septembre 2008

KEYWORDS
Sequencing;
Sanger;
Pyrosequencing;
Whole genome amplification

MOTS CLÉS
Séquençage ;
Sanger ;
Pyroséquençage ;
Amplification génome entier

Summary DNA sequencing is an essential tool in molecular biology and applied biosciences. Described in the late 1970s, this method had enormously increased the possibilities of genetic research. DNA sequencing is now routinely used in molecular biology laboratories. This technology has allowed sequencing of various and important genomes as the human one. Numerous innovations to improve DNA sequencing have been realized and new technologies have been described. The main objective of this article made up of two parts is to present the actual sequencing methodologies and the main evolutions in progress in this field. Individual human sequencing is not so far. In addition to ethical questions that will rise, other questions need to be considered. For example, how will we interpret the many genetic variants in regard to predisposition to disease and to phenotype? Many studies in progress will answer these questions. In any case, a revolution is in motion.

Résumé Le séquençage d'ADN est devenu un outil essentiel en biologie moléculaire tant en médecine que dans de nombreuses autres disciplines des sciences de la vie. Le séquençage a été décrit il y a environ 30 ans et n'a cessé d'évoluer depuis cette période. Cette méthode est devenue une technique courante dans les laboratoires de biologie moléculaire. Les connaissances acquises grâce à cette méthode et la possibilité de séquencer des génomes de grande taille, tel que le génome humain, ont amené les chercheurs à développer des techniques de séquençage de plus en plus sophistiquées. Cet article, composé de deux parties, présente les techniques actuellement utilisées pour séquencer l'ADN, qu'il soit humain ou d'autre origine, et les méthodes de séquençage en développement. Ces dernières constituent un réel bouleversement. Le séquençage à l'échelle individuelle n'est plus loin. En dehors des problèmes

* Auteur correspondant.
Adresse e-mail : jerome.lamoril@lmr.ap-hop-paris.fr (J. Lamoril).

0923-2532/$ – see front matter
doi:10.1016/j.immbio.2008.07.016

éthiques qu'elle soulève, cette révolution pose de nouvelles questions, par exemple : comment interpréterons-nous les nombreuses variations génétiques observées chez un individu, quelles en seront les conséquences sur ses prédispositions génétiques aux maladies et autres risques, quels en seront les retentissements sur le phénotype ? De nombreuses études en cours cherchent les réponses. Dans tous les cas, la révolution est en marche.

Introduction

Depuis la description de la structure de l'ADN en 1955 jusqu'à nos jours, la biologie a connu une suite de remarquables progrès technologiques dont le séquençage constitue l'un des évènements clés. En ce début de troisième millénaire déjà riche en nouvelles technologies, nous assistons à une nouvelle révolution dans le domaine du séquençage. À travers deux articles, nous souhaitons vous présenter ces changements. Dans une première partie, nous exposerons les différentes méthodes de séquençage dans leur aspect actuel et dans une seconde partie, les bouleversements technologiques en marche dans ce domaine.

Le séquençage

Le séquençage de l'ADN constitue une méthode dont le but est de déterminer la succession linéaire des bases A, C, G et T prenant part à la structure de l'ADN. La lecture de cette séquence permet d'étudier l'information biologique contenue par celle-ci. Étant donné l'unicité et la spécificité de la structure de l'ADN chez chaque individu, la séquence de l'ADN permet de nombreuses applications dans le domaine de la médecine, comme, par exemple, le diagnostic, les études génétiques, l'étude de paternité, la criminologie, la compréhension de mécanismes physiopathologiques, la synthèse de médicaments, les enquêtes épidémiologiques. Dans de nombreuses publications, le terme séquençage peut se retrouver sous deux dénominations différentes qu'il est important de connaître. Dans les études de génomes, le terme de *reséquençage* (expression pouvant prêter à confusion) est utilisé à la place de séquençage. Cette dénomination, essentiellement utilisée en génétique, désigne le séquençage d'un segment d'ADN suivi de la comparaison du résultat obtenu avec celui d'une séquence de référence connue. Un autre terme est également employé : le séquençage *de novo*. Dans ce cas, il s'agit du séquençage d'un génome pour lequel il n'existe pas de séquence référence. Il s'agit donc de la détermination d'une séquence inconnue. Dans notre article, sauf dans quelques cas, nous parlerons de séquençage au sens large du terme sans distinguer le *reséquençage* du séquençage *de novo*, les techniques utilisées étant généralement les mêmes.

Intérêt général du séquençage

Le génome humain contient six milliards de bases, soit la quasi-totalité de notre patrimoine génétique. De plus, nous possédons un petit génome indépendant, celui de la mitochondrie (d'environ 16 500 bases). La séquence complète du génome humain contenu dans le noyau de la cellule sous forme d'ADN a été finalisée en 2006. Par ailleurs, les génomes de nombreux agents infectieux, de mammifères et de plantes ont également été sequencés dans leur totalité (nombre d'entre eux sont accessibles sur le site www.ncbi.nlm.nih.gov/Genomes). Leur connaissance a modifié considérablement les recherches biomédicales et biologiques en ouvrant de vastes panoramas dans le domaine de la médecine (diagnostic, thérapeutique, prédiction, pronostic, prévention…) et dans de nombreuses autres disciplines biologiques (anthropologie, agronomie, environnement…). La progression des connaissances croît à une vitesse spectaculaire. Le séquençage est de fait un remarquable instrument nécessaire à la compréhension des cycles de la vie dans leur globalité. Il devrait permettre d'améliorer la santé humaine et l'équilibre écologique de la planète. Les techniques de séquençage évoluent et leurs applications s'élargissent (Tableau 1). Par ailleurs, le séquençage a pu « se démocratiser » dans de nombreux laboratoires, en partie depuis la description de la *polymerase chain reaction* (PCR) en 1985, suivie de sa diffusion très large dans les laboratoires de biologie moléculaire. Depuis 2000, outre la PCR que nous ne décrirons pas, de nouvelles techniques de séquençage se sont développées. Elles constituent un progrès technologique révolutionnaire et seront présentées dans la seconde partie de cet article.

Tableau 1 Quelques applications du séquençage.

Diagnostic et traitement de nombreuses maladies humaines (exemples : cancers, maladies infectieuses, maladies héréditaires…)
Informations sur le génome (structure, fonction, évolution) et étude des variations du génome (polymorphismes bialléliques, insertions, délétions, insertions/délétions (appelées aussi indéls), réarrangement de gènes, variation du nombre de copies de gènes, duplication (ou plus)
Variants génétiques associés à une pathologie (par exemple, le diabète)
Analyse de méthylation du génome (études épigénétiques et méthylome)
Analyse microbiologique (identification d'espèces, taxonomie, études épidémiologiques, génotypage à but pronostique et/ou thérapeutique)
Tests de paternité et médecine légale
Police scientifique
Pharmacogénétique
Études anthropologiques

Figure 1 Technique de Maxam-Gilbert. Cette technique est une méthode chimique de traitement de l'ADN. Un fragment amplifié par PCR et marqué radioactivement par le phosphore radioactif (P^{32}) est modifié par un agent chimique, par exemple l'hydralazine. Celle-ci modifie les bases C et T et en milieu alcalin, uniquement les bases C (comme dans ce schéma). Dans un second temps, l'addition de pipéridine casse de manière aléatoire et au moins une fois au niveau de chaque base C modifié. On obtient donc des fragments de taille différente.

Les techniques de séquençage de l'ADN

Les deux premières techniques de séquençage de l'ADN, celle de Maxam-Gilbert [25] et celle de Sanger [66] ont été décrites en 1977. À noter que les deux premières publications rapportant un séquençage datent de 1973 [25,49]. Il s'agissait du séquençage de l'opérateur Lac et de l'ARNm de celui-ci. La technique de Sanger a révolutionné le monde de la biologie moléculaire en permettant de décrypter différents génomes, tel que celui du génome humain complètement déchiffré en 2006 ou d'autres génomes, le génome bactérien, par exemple, le premier d'entre eux étant celui d'*Haemophilus influenzae*, complètement décrit en 1995 [18]. Bien que les techniques de séquençage évoluent, comme nous allons le voir dans cet article, la méthode de Sanger continue d'être la méthode la plus employée dans le monde à l'heure actuelle.

La technique de Maxam-Gilbert

Cette technique est pratiquement abandonnée de nos jours. Nous la décrirons brièvement pour des raisons historiques. Cette méthode, publiée parallèlement à celle de Sanger en 1977, par son caractère révolutionnaire a grandement contribué à l'histoire de la biologie moléculaire. Il s'agit d'une méthode chimique de séquençage. Les réactifs clivent spécifiquement après chacune des bases A, C, G, [A + G], [C + T]. Cette technique est basée sur la propriété de certains agents chimiques, l'hydrazine, le diméthyl sulfate (DMS) et l'acide formique, de modifier les bases de l'ADN. Dans un second temps, la pipéridine est ajoutée et « casse » les brins d'ADN au niveau des bases modifiées. Les agents chimiques sont utilisés dans des conditions telles qu'ils n'agissent qu'avec un faible pourcentage des bases de l'ADN étudié. Le DMS agit au niveau des bases « G ». L'acide formique agit au niveau des bases « A + G ». L'hydrazine agit au niveau des bases « C + T » (en milieu alcalin, l'hydrazine agit uniquement au niveau des « C »). L'ADN à séquencer est marqué à une extrémité. Le plus souvent, il s'agit d'un marqueur radioactif. Le produit de séquence est déposé sur un gel d'acrylamide, puis la séquence lue après autoradiographie (Fig. 1 et 2). L'ADN étudié peut être simple ou double brin. Cette technique permettait d'analyser des fragments allant jusqu'à 500 pb.

La technique de Sanger

La diffusion de la méthode de Sanger, la commercialisation d'automates utilisant des fluorophores quatre couleurs ainsi que le déploiement de la PCR dans les laboratoires ont considérablement amélioré les procédures de séquençage. La méthode de Sanger a en effet rapidement dépassé la méthode de Maxam-Gilbert pour la remplacer et reste à ce jour la principale méthode de séquençage utilisée dans les laboratoires. Son principe est le suivant. Dans un premier temps, il est nécessaire d'amplifier l'ADN cible par PCR, puis de le dénaturer afin d'obtenir un ADN simple brin. À l'aide d'une amorce spécifique et complémentaire du brin étudié (sens ou antisens), identique ou différente de celle utilisée pour la PCR, une ADN polymérase effectue alors la synthèse de l'ADN complémentaire à partir de cette amorce. De l'extrémité 5′ vers l'extrémité 3′, cette enzyme ajoute les désoxyribonucléotides-triphosphates (dNTP) complémentaires et de manière aléatoire et inconstante des didéoxyribonucléotides triphosphates (ddNTP), par exemple un ddGTP sera parfois ajouté à la place d'un dGTP. La réaction se faisant dans un seul tube, les ddNTP (ddATP, ddGTP, ddCTP et ddTTP) sont marqués à l'aide de fluorophores différents pour chaque ddNTP (fluorophores « quatre couleurs »). Lorsqu'un ddNTP est incorporé à la place d'un dNTP, l'ADN polymérase ne peut plus continuer sa polymérisation. La réaction d'extension s'arrête (en effet, le didéoxynucléotide ne possède pas de groupe 3′-hydroxyle indispensable

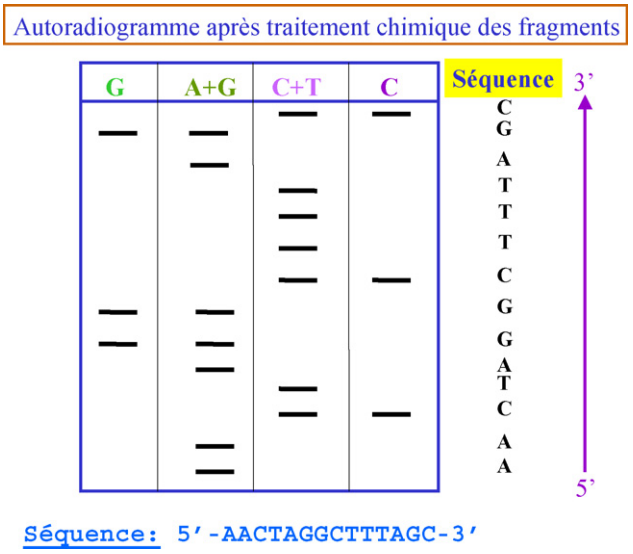

Figure 2 Technique de Maxam-Gilbert. Dans quatre tubes différents, l'ADN cible est traité par chacun des produits de modification spécifique de base (hydralazine C + T ; hydralazine C en milieu alcalin ; diméthyl sulfate G ; acide formique A + G), suivi d'un traitement par la pipéridine. Les fragments coupés aléatoirement et au moins une fois après chaque base spécifique sur l'ADN cible sont de taille différente. La migration de ces derniers dans un gel d'acrylamide spécifique suivie d'une autoradiographie permet de déduire la séquence de l'ADN au cours de la lecture du gel dans le sens 5′ → 3′ de bas en haut du gel.

à la réaction de polymérisation de l'enzyme). Statistiquement, au cours de la réaction, pour chaque « base » de l'ADN cible, au moins une fois, un ddNTP complémentaire sera incorporé à la place d'un dNTP. Par conséquent, à la fin de la réaction, nous obtiendrons des fragments de taille différente. L'analyse de la réaction est ensuite effectuée. Différentes méthodes d'analyse sont possibles. Aujourd'hui, l'électrophorèse capillaire réalisée sur un automate de séquençage est la méthode de choix. Lors de la migration, chaque fragment (contenant un ddNTP marqué par un fluorophore) sera excité par un laser et le signal obtenu analysé par un logiciel spécifique. L'analyse informatique des signaux permet d'obtenir la séquence étudiée, par exemple, sous forme d'un électrophorégramme, de lecture manuelle aisée mais souvent fastidieuse (Fig. 3). Des logiciels d'analyse des séquences peuvent être utilisés. Dans tous les cas, l'analyse d'un fragment d'ADN après PCR se fait toujours à l'aide d'une amorce sens et antisens afin de confirmer la séquence (et une éventuelle anomalie de séquence). En général, cette technique permet d'obtenir des séquences de longueur comprise entre 400 et 850 pb [30]. Comme déjà indiqué, cette technique, décrite pour la première fois en 1977, reste la plus utilisée dans les laboratoires, notamment en milieu hospitalier. À titre d'exemple, de nombreux laboratoires hospitaliers utilisent un séquenceur commercialisé par la société Applied Biosystems permettant l'analyse de séquences en plaques de 96 ou 384 puits par électrophorèse capillaire (analyse multicapillaire en parallèle). Ce séquenceur contient un, quatre, huit, 16, 48 ou 96 capillaires selon le modèle. Ainsi, le modèle ABI3130Xl (96 puits, 16 capillaires) permet le séquençage d'environ 400 pb/puits en trois heures (soit 28,8 kb pour la plaque entière). En sachant que cette machine permet de lire environ 18 bases par seconde (pour des 96 capillaires), un an serait nécessaire pour séquencer un génome humain à l'aide de 100 machines utilisées en parallèle, en recouvrant cinq fois le génome (équivalent à cinq séquençages du génome), minimum nécessaire pour s'assurer de l'absence d'erreurs et en supposant que le temps de préparation de ces machines et des échantillons soient négligeable. D'autre développement de la méthode de Sanger sont néanmoins en cours et notamment la miniaturisation de la technique. À titre d'exemple, récemment des auteurs ont réussi à séquencer 600 pb en 6,5 minutes à l'aide d'une puce constituée d'une microfluidique permettant une électrophorèse avec un capillaire de 7,5 cm de long constitué d'une polymère spécifique [21]. D'autres technologies ont donc été développées pour améliorer le rendement, la rapidité et le coût du séquençage.

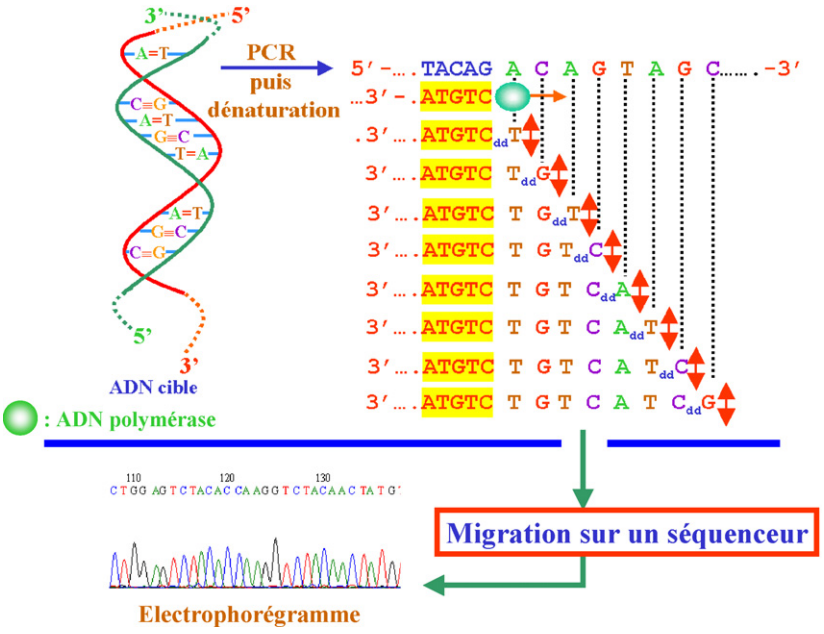

Figure 3 Principes du séquençage selon la méthode de Sanger. Après dénaturation du produit amplifié par séquençage, l'un des deux brins (ici, le brin sens) s'hybride à une amorce spécifique. Pour la simplicité du schéma, nous avons pris une amorce de 5 pb, la taille habituelle des amorces étant de 20 pb environ. Le mélange réactionnel contient, outre les tampons et l'ADN polymérase, des déoxynucléotides triphosphates (dNTP, dA-, dC-, dG-, dT-TP) mais aussi des didéoxynucléotides triphosphates (ddNTP, ddA-, ddC-, ddG-, ddT-TP). L'incorporation aléatoire d'un ddNTP à la place d'un dNTP ne permet plus la polymérisation par l'ADN polymérase. L'extension s'arrête. À la fin de la réaction de séquence effectuée selon des cycles thermiques identiques à ceux de la PCR (on parle de PCR asymétrique, une seule amorce étant utilisée au lieu de deux), nous avons des fragments de taille différente. Ces fragments sont soumis à migration dans un champs électrique. Il s'agit le plus souvent d'une électrophorèse capillaire. Chaque ddNTP étant marqué par un fluorophore différent, un signal lumineux sera généré, spécifique de la base didéoxy incorporée. Les fragments étant de taille différente et la résolution allant jusqu'à une base de différence, il sera simple de recueillir ce signal et en déduire la séquence. Les signaux lumineux sont analysés par un logiciel spécifique, et le résultat de l'analyse peut être lu, par exemple, sous forme d'un électrophorégramme de lecture facile. Des logiciels d'interprétation des séquences sont également disponibles. Pour confirmer un résultat, toute réaction de séquence d'un fragment d'ADN est systématiquement faite sur le brin sens et le brin antisens.

La technique shotgun (séquençage aléatoire globale). Elle a été utilisée massivement pour le séquençage d'un grand nombre de génomes notamment bactériens (séquençage *de novo*). Schématiquement, cette méthode consiste à fragmenter le génome entier à étudier en petits fragments d'ADN à l'aide de moyens mécaniques. Sur chaque fragment, une réaction de ligation permet de fixer de courtes séquences d'ADN appelées adaptateurs, ces derniers servant d'amorce pour la PCR. Ces fragments sont ensuite intégrés dans des plasmides et constituent une bibliothèque (*library*) de fragments aléatoires d'ADN simple brin. Ils sont ensuite amplifiés par PCR, par exemple, puis séquencés à l'aide de la méthode de Sanger. À l'aide de logiciels informatiques, les séquences sont ensuite alignées et recoupées par chevauchement. Ces séquences représentent plusieurs blocs de séquences continus : on parle alors de « contig ». Certaines séquences manquent : ce sont des trous de séquences (*gap*). Ceux-ci sont comblés par séquençage à partir des séquences déjà déterminées. À partir de ces dernières, sont dessinées des amorces de PCR servant pour l'amplification en direction de ces trous. Cette technique nécessite de séquencer le génome cible plusieurs fois (minimum cinq fois), tant pour éviter les erreurs de séquençage que pour s'assurer du maximum de chevauchement entre les séquences et faciliter l'ordonnancement des séquences. Cette technique est essentiellement utilisée pour des génomes de petite taille comme ceux des bactéries.

Le principe de séquençage de grands génomes. L'automatisation à l'aide de robots a permis de séquencer de nombreux génomes dont le génome humain (pour plus de détails sur une des méthodes utilisées, voir le site jgi.doe.gov/education/how). Le travail a été considérable et n'a pu être effectuée que grâce à une robotisation et une informatisation poussées dans des instituts dédiés à ces séquençages à grande échelle. À ce jour, plus de 280 génomes prokaryotes ont ainsi été totalement séquencés. Pour réaliser le séquençage d'un grand génome (eucaryote ou procaryote, par exemple), une carte physique du génome est d'abord constituée. Il s'agit d'établir des repères sur le génome à l'aide de marqueurs spécifiques de chaque chromosome. Schématiquement, deux types de marqueurs sont utilisés :

- les polymorphismes (avec étude de la liaison génétique entre ceux-ci). Ce type de marqueurs a été étudié soit à l'aide d'enzymes de restriction (on parle alors de *restricted fragment length polymorphisms* [RFLP]), soit à l'aide de microsatellites (séquences répétées dont le nombre de répétition est le plus souvent variable [polymorphe] au même locus) ;

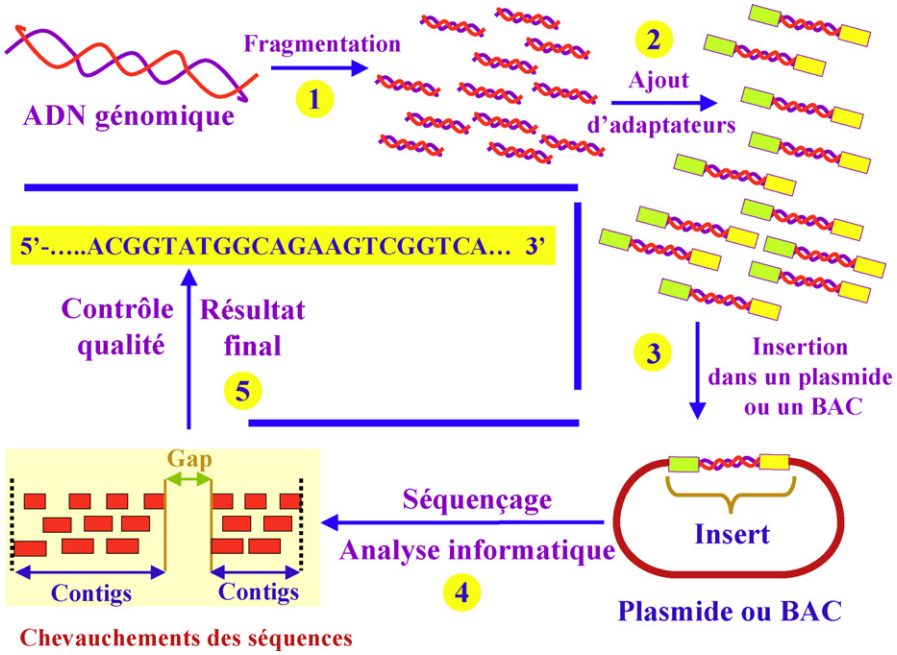

Figure 4 Principe du séquençage après shotgun. **1** : la première étape consiste à fragmenter l'ADN génomique à l'aide de moyens mécaniques par exemple. **2** et **3** : après ajout d'adaptateurs par exemple (rectangles), les fragments d'ADN sont clonés (insérés) dans un vecteur (plasmide ou BAC, par exemple). L'ensemble est mis dans des bactéries (le plus souvent après transformation) pour amplification. **4** : après extraction du vecteur et de son insert à partir de la bactérie, le séquençage selon la méthode de Sanger peut être réalisée, par exemple, à l'aide soit d'amorces spécifiques, soit d'une amorce universelle. L'analyse informatique des résultats permet d'établir une succession de séquences se chevauchant (et formant par regroupement des *contigs*). Parfois, des trous (*gaps*) existent. Pour séquencer ces trous, on utilise le plus souvent des amorces à partir des *contigs* encadrant le trou. **5** : un génome est en fait séquencé cinq à dix fois, ce qui permet de limiter les erreurs et d'augmenter les chances de tout séquencer. À chaque étape de ces préparations, des contrôles de qualité s'assurent de la validité des résultats et la séquence finale peut être mise à disposition.

- les séquences uniques polymorphes ou non appelées aussi *sequence tagged sites* (STS).

La carte physique permet de limiter la zone à étudier et ainsi de faciliter la reconstitution du génome à séquencer. Plusieurs méthodes existent. Nous en décrirons une de manière simplifiée. Pour permettre le séquençage du génome, il est nécessaire de le fragmenter (Fig. 4). Pour cela, l'ADN est cassé mécaniquement. Les nombreux fragments obtenus sont ensuite insérés au hasard dans des plasmides (ces fragments sont alors appelés inserts). Ils peuvent être aussi insérés dans des cosmides ou des fosmides (vecteurs particuliers dérivés de plasmides, utilisés pour des inserts de plus grande taille) [3]. L'ensemble est introduit dans des bactéries (*Escherichia coli*) par transformation. Les fragments insérés ont une taille d'environ 2 à 4 kb. On obtient alors des clones dont on peut étudier les marqueurs et les STS. Pour cela, il est nécessaire de sélectionner les bactéries ayant incorporé le plasmide/insert (la transformation ne présente pas un rendement de 100 %). Des marqueurs de sélection sont utilisés (par exemple, résistance à des antibiotiques et système de couleurs des colonies s'assurant du succès de la transformation). Étant donné le nombre important de bactéries à sélectionner, plusieurs robots sont utilisés. Ils reconnaissent et sélectionnent les bactéries ayant incorporé le plasmide avec l'insert. La connaissance de la carte génétique et des STS facilitent ensuite la sélection des clones à séquencer de manière ordonnée. D'autres robots permettent de récupérer l'ensemble plasmide/insert après culture des bactéries et d'extraire l'ADN de l'ensemble plasmide/insert. Cet ADN est alors amplifié non par PCR mais par *rolling circle amplification* (RCA). L'ADN est ensuite séquencé à l'aide d'hexamères aléatoires par la technique de Sanger. Cette réaction est également réalisée à l'aide de robots. Les séquences sont ensuite lues par électrophorèse capillaire à l'aide d'automates de séquençage de 384 capillaires. De nombreux fragments sont obtenus et l'analyse informatique regroupe les séquences chevauchantes pour obtenir une séquence continue. Cette séquence, une fois validée, sera la séquence de référence pour d'autres études sur le génome cible. Dans le cas du génome humain, plutôt que des plasmides, ce sont des chromosomes bactériens artificiels appelés *bacterial artificial chromosomes* (BAC, pouvant contenir des inserts d'une taille allant jusqu'à 300 kb) et parfois des *phage artificial chromosomes* (PAC, de même capacité) qui ont été utilisés. Ces clones ont ensuite été ordonnés (positionnés les uns par rapport aux autres et le long des chromosomes humains). Pour cela, l'utilisation des profils de restriction, la connaissance de la carte génétique et des STS ainsi que l'étude par hybridation des clones entre eux ont été nécessaires. Le séquençage a alors été réalisé sur les clones chevauchant pour obtenir la séquence humaine de référence. Depuis le séquençage initial du génome à l'aide de la technique shotgun précédemment décrite, d'autres approches similaires ont été développées pour permettre un séquençage plus rapide. Par exemple et schématiquement, après fragmentation de l'ADN et addition d'adaptateurs aux extrémités par ligation, les différents fragments ainsi préparés sont fusionnés ensemble au hasard et forment ainsi une bibliothèque (*library*) de fragments aléatoires simple brin. Chaque fragment est ensuite immobilisé sur une bille, chaque bille ne contenant qu'un seul fragment aléatoire. Une PCR émulsion est réalisée (cette technique est décrite dans la partie II de l'article) [15,32,33]. Chaque amplification est clonale et permet d'obtenir des millions de copies d'un même ADN par puits. Un automate de séquence peut alors démarrer le séquençage dans chaque puits (par pyroséquençage, par exemple) et la séquence sera déduite après analyse informatique. Cette dernière méthode est plus rapide et moins chère que la méthode utilisant le séquençage par la méthode de Sanger [80].

Tableau 2 Recommandations pour l'identification bactérienne par séquençage de l'ARNr 16S [35].

Catégorie	Recommandations
Souche à séquencer	En cas de difficulté d'identification phénotypique ou de difficulté identifiée et connue pour l'analyse par le gène ARNr 16S, il est nécessaire de séquencer en parallèle un autre gène commun (par exemple, le gène *rpoB*).
Séquençage de l'ARNr 16S	Séquencer au minimum 500—525 pb Pour obtenir moins de 1 % d'ambiguïté, séquencer 1300—1500 pb
Critères d'identification d'espèce	Au moins > 99 % d'identité de séquence (idéalement > 99,5 %) par rapport à la souche référence dont les caractères ont été définis par les études d'hybridation ADN/ADN En cas de différence < 0,5 % par rapport à l'espèce la plus proche, il est nécessaire d'évaluer les autres propriétés de la souche, notamment au niveau phénotypique pour permettre l'identification définitive.

L'analyse de méthylation du génome et profil de méthylation du génome (méthylome). Le séquençage permet de déterminer la succession des bases sur l'ensemble du génome. L'étude de ce dernier a montré que certes la séquence du génome était fondamentale pour la compréhension de nombreux mécanismes cellulaires mais il demeurait certaines anomalies et particularités génétiques inexpliquées résultant de la structure de ce génome. Schématiquement, elles ont pour origine un ensemble de mécanismes aboutissant à des modifications phénotypiques sans atteinte génotypique. On parle alors d'épigénétique. Il n'y a donc pas d'altération de la séquence de l'ADN (qui est donc normale) mais altération de la structure de ce génome. En ce qui concerne la chromatine, on observe des modifications au niveau des histones et un remodelage des nucléosomes. L'épigénétique joue donc un rôle important. Ces modifications épigénétiques sont étudiés par diverses techniques dont le séquençage. En effet, parmi les modifications observées qu'il nous est impossible de décrire en totalité dans cet article, on retrouve des

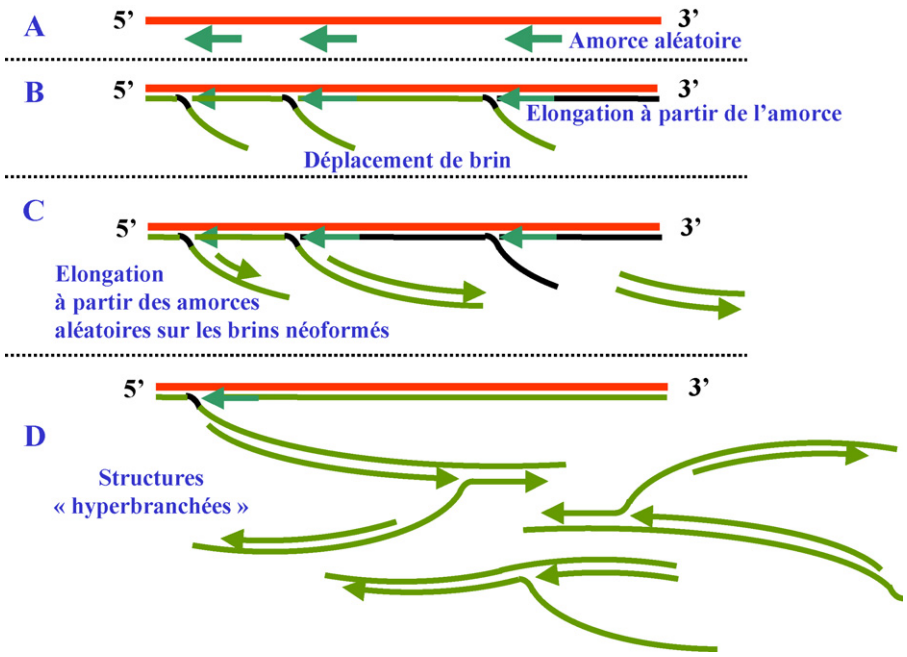

Figure 5 Principe de l'amplification par MDA. **A** : des amorces aléatoires hybrident à l'ADN cible simple brin. **B** : l'élongation du brin complémentaire est effectuée par l'ADN polymérase à partir des amorces aléatoires. Les brins se trouvant en amont sont déplacés. **C** : les brins d'ADN déplacés servent de matrice pour les amorces aléatoires et permettent l'élongation de nouveaux brins. **D** : les multiples activités de l'ADN polymérase et les nombreux déplacements consécutifs aboutissent à des structures de type « hyperbranchées ». L'ensemble de ces réactions aboutit à la génération de nombreuses copies de l'ADN cible initial.

méthylations de l'ADN sur les cytosines de zones riches en bases GC, zones appelées aussi îlots CpG (plus précisément, une méthylation sur le carbone 5 des cytosines). Ces méthylations jouent un rôle fonctionnel majeur, notamment dans la régulation de l'expression de gènes [6]. Le séquençage est un moyen d'étudier ces méthylations normales ou non. La méthode de Sanger est utilisée après traitement de l'ADN génomique au bisulphite de sodium, technique décrite pour la première fois en 1992 [22]. Ce traitement a été adapté dans un second temps au séquençage par pyroséquençage (voir infra) [75]. Schématiquement, avant amplification du produit à analyser par PCR, l'ADN génomique est traité avec du bisulphite de sodium. Celui-ci possède la propriété de transformer les cytosines non méthylées en uracile alors que les cytosines méthylées ne sont pas modifiées. Après PCR, l'uracile sera transformé en thymine et le séquençage permettra de distinguer bien évidemment les thymines (correspondant à une cytosine non méthylées sur l'ADN génomique) des cytosines (cytosine méthylée sur l'ADN génomique). Les nouvelles générations de séquenceur (traitées dans la seconde partie de cet article) permettent aussi l'analyse de méthylation du génome. Des projets de recherche sur l'épigénome sont en cours (par exemple, le projet épigénome humain démarré en 2000 ; www.epigenome.org).

Un exemple de séquençage : l'identification bactérienne par séquençage du gène ARN 16S. Le séquençage du gène codant pour l'ARN ribosomal 16S (ARNr 16S) permet l'étude de la phylogénie et de la taxonomie bactériennes [35]. Ce gène de 1500 pb environ est présent dans la plupart des bactéries (sous forme d'opéron ou de famille multigènes) et sa fonction n'a pas changé au cours de l'évolution. Par ailleurs, sa taille est suffisante pour une analyse informatique puissante. Au 9 juillet 2008, 9263 espèces de bactéries ont pu être classées grâce au séquençage de l'ARNr 16S (www.bacterio.cict.fr/number.html#total). Le séquençage de l'ARNr 16S par sa simplicité a rendu la classification des bactéries plus aisée même si la référence absolue (*gold standard*) pour l'identification de nouvelles espèces et leur classification taxonomique reste l'hybridation ADN/ADN de réalisation plus lourde et plus difficile. C'est en 1994 que des chercheurs ont proposé et établi l'utilisation potentielle du séquençage de l'ARNr 16S pour la définition d'espèces en microbiologie [72]. En pratique, le séquençage de l'ARNr 16S permet l'identification du genre et de l'espèce d'un isolat. Cette méthode est intéressante notamment lorsque les classiques méthodes biochimiques ne permettent pas d'identifier clairement une bactérie (par méthode commerciale ou non). Le séquençage de l'ARNr 16S permet l'identification du genre bactérien dans plus de 90 % des cas, de l'espèce bactérienne dans 65—85 %. Selon les études, 1 à 14 % des isolats étudiés demeurent non identifiées après utilisation de cette méthode [35]. Dans tous les cas, l'identification par séquençage de l'ARNr 16S reste supérieure aux méthodes bactériologiques conventionnelles. Ce séquençage présente néanmoins des limites. À titre d'exemple, dans le genre *Bacillus*, deux souches — *Bacillus globisporus* et *Bacillus psychrophilus* — partagent plus de 95 % de séquences identiques sur leurs gènes ARNr 16S alors que la technique de référence d'hybridation ADN/ADN montre 23—50 % d'identité [19]. Certaines études ont établi une liste de genres bactériens posant des difficultés d'identification à l'aide de cette technique (à titre d'exemple, le genre *Campylobacter* pour les espèces « non » *Jejuni coli*, les mycobactéries à croissance rapide, le genre *Actinomyces*). Malgré ces inconvénients, le séquençage de

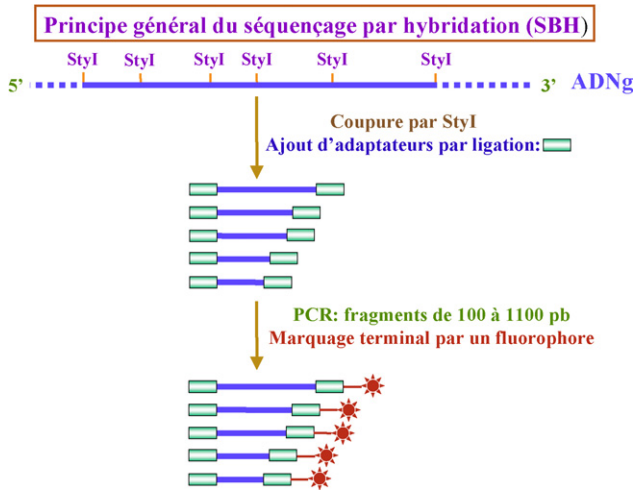

Figure 6 Principe du SBH. Pour permettre le SBH, l'ADN génomique est d'abord coupé en plusieurs fragments à l'aide d'une enzyme de restriction (par exemple, *StyI*). Des oligonucléotides (appelés adaptateurs) sont ajoutés aux extrémités 3′ et 5′ des fragments d'ADN double brins pour permettre l'amplification génique par PCR à l'aide d'amorces complémentaires de ces adaptateurs. Des fragments dont la taille varie entre 100 et 1100 pb sont obtenus. Ils sont ensuite marqués à l'aide d'un fluorophore puis hybridés sur la puce d'ADN.

l'ARNr 16S joue un rôle important dans l'identification bactérienne, que ce soit pour l'identification de bactéries inconnues ou pour celles au profil biochimique ambigu. Pour aider à mieux interpréter le séquençage de l'ARNr 16S, des recommandations ont été publiées (Tableau 2). Bien évidemment, les autres techniques de séquençage, par exemple, le pyroséquençage (voir infra) sont également utilisées pour réaliser de telles études [31].

L'amplification génome complet ou *whole genome amplification* (WGA)

L'étude du génome humain par séquençage nécessite une quantité suffisante d'ADN. Grâce aux techniques d'amplification génique dont la PCR constitue la principale méthode, il est possible d'analyser des échantillons contenant peu d'ADN. Néanmoins, dans certains cas, la quantité d'ADN est insuffisante ou très faible alors que le nombre d'études à effectuer sur un ADN peut être important (par exemple, séquençage, étude de polymorphismes, quantification...). Jusqu'au début des années 2000, la principale technique permettant d'avoir de l'ADN quasi-infinie reposait sur l'immortalisation de lymphocytes après infection par le virus d'Epstein-Barr, technique lourde et coûteuse. Afin de contourner cet obstacle, des techniques d'amplification de l'ensemble du génome appelée aussi WGA ont été décrites afin d'obtenir une quantité suffisante d'ADN génomique à analyser [45]. La difficulté de ce principe technique d'amplification distinct de la PCR est de deux ordres :

- éviter l'amplification préférentielle d'allèle et/ou l'absence d'amplification de certaines zones chromosomiques (appelé *allele drop-out* [ADO]) ;
- empêcher un déséquilibre d'amplification entre les locus chromosomiques.

L'amplification du génome doit donc être équilibrée, c'est-à-dire qu'elle doit être quantitativement identique pour chaque région du génome. Parmi les techniques publiées, la plus utilisée est la méthode d'amplification par déplacement multiple également appelée MDA (*multiple displacement amplification*) décrite pour la première fois en 2002 [13]. À noter pour mémoire que d'autres techniques d'amplification tout génome ont été décrites avant la description de la MDA telles que la préamplificaton par extension d'amorce, *primer extension preamplification* (PEP) et la PCR à l'aide d'amorces dégénérées, *degenerate oligonucleotide primed* PCR (DOP-PCR) [74,79]. Nous ne les décrirons pas. La MDA est basée sur les propriétés particulières d'une ADN polymérase issue d'un bactériophage, le bactériophage Φ29. Cette enzyme possède trois caractéristiques importantes :

- une activité de polymérisation très rapide ;
- une grande fidélité de recopiage ;
- une activité de déplacement de brin.

La première technique utilisant cette enzyme fut décrite en 1998. Il s'agissait d'une technique d'amplification isotherme à partir d'un ADN circulaire, technique appelée aussi

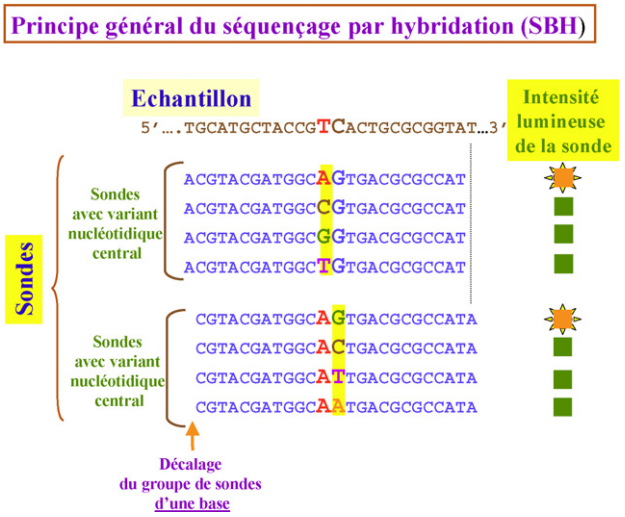

Figure 7 Principe du SBH. Plusieurs constructions sont possibles. Nous en donnerons un exemple. Les puces d'ADN sont constituées de courtes sondes oligonucléotidiques dans lesquelles chaque variant possible (A, C, G et T) est représenté au milieu de la sonde (en couleur différente pour chaque base sur le schéma). Quatre sondes au minimum de 25 pb sont donc présentes pour interroger chaque position nucléotidique (dans notre exemple, deux positions adjacentes). Le plus souvent, huit sondes sont présentes au minimum afin d'étudier les brins sens et antisens ainsi que de petites insertions et délétions. En cas de complémentarité exacte entre la sonde et la séquence, le signal d'hybridation mesuré sera maximum (en rouge). Après mesure des signaux d'hybridation, l'analyse informatique permet de reconstruire la séquence exacte à l'aide de logiciels spécifiques.

amplification par enroulement de cercle ou plus simplement RCA [43]. Cette méthode utilisée par la suite dans d'autres applications permettait l'amplification de courts fragments d'ADN. Une variante de cette technique a ensuite été décrite. En utilisant des amorces hexamères aléatoires (*random hexamers*) modifiées en 3′ par l'addition d'un groupement thiophosphate les protégeant d'une dégradation possible liée à l'activité exonucléasique 3′—5′ de la Φ29 ADN polymérase, il a été démontré qu'on pouvait réaliser une amplification dite hyperbranchée appelée RCA multiamorcée (*multiple-primed* RCA [MP-RCA]) au court de laquelle de plus grands fragments pouvaient être amplifiés (5 Mb, voire plus) [14]. En 2002, il a été démontré que cette même enzyme pouvait également répliquer des fragments linéaires [13]. Le principe de cette technique WGA est le suivant : des amorces hexamères aléatoires hybrident au hasard et en de nombreuses localisations sur l'ADN cible. La Φ29 ADN polymérase réplique l'ADN à partir de ces amorces. Au cours de la polymérisation, l'enzyme atteint d'autres sites d'initiation de la réplication, conséquences de l'hybridation des amorces aléatoires et de leur réplication parallèle. Du fait de ses propriétés de déplacement d'ADN double brin, l'enzyme déplace les fragments amplifiés. La réplication continue. L'ADN déplacé sert de nouveau pour l'initiation de la réplication à partir d'autres hexamères aléatoires. L'ensemble donne un aspect d'ADN « hyperbranché » (Fig. 5). Les fragments d'ADN ainsi obtenu ont une taille moyenne d'environ 10 kb (voire plus). Même des fragments riches en GC (> 80 % GC, zones difficiles à amplifier par PCR) ont été amplifiés avec succès. Cette méthode permet l'analyse d'ADN par de nombreuses techniques de biologie moléculaire dont le séquençage [45]. Comparée à la stratégie actuellement la plus utilisée (extraction d'ADN génomique suivie de PCR, puis séquençage), la stratégie WGA suivie de séquençage permet d'obtenir des résultats identiques [47]. Une publication a démontré qu'en partant d'une quantité d'ADN de 0,3 ng (équivalent à 45 cellules), on obtenait des résultats similaires à ceux obtenus avec des quantités d'ADN nettement supérieures [44]. Une autre étude a comparé la technique d'amplification WGA suivie de PCR, puis de séquençage versus la technique de PCR sur ADN total suivie de séquençage. Une discordance de 9 % a été observée (sujets considérés homozygotes alors qu'ils étaient hétérozygotes) [55]. La discordance observée était probablement liée à la faible quantité d'ADN de départ. Pour limiter ce risque, il a été démontré qu'une dilution de l'ADN amenant à une concentration d'une seule molécule d'ADN (une cellule contient 6 pg d'ADN) suivie de MDA améliorait grandement le résultat après PCR : dans ce dernier exemple, deux méthodes d'amplification étaient synergiquement utilisées [40]. La MDA a également montré son intérêt dans l'étude des micro-organismes. Il a été possible d'amplifier de l'ADN génomique d'une seule spore ou d'une seule bactérie [23,61]. En pratique, la MDA a démontré sa supériorité sur les autres techniques d'amplification tout génome telle que les techniques PEP et DOP-PCR et représente la méthode de choix pour amplifier de l'ADN génomique dans de nombreuses indications et, plus particulièrement, quand la quantité d'ADN de départ est faible [56,59]. Il est cependant nécessaire que l'ADN extrait soit de bonne qualité. Ainsi, il faut rester prudent dans l'interprétation après amplification par MDA lorsque des ADN d'archives congelés/décongelés plusieurs fois sont étudiés [12].

Le séquençage par hybridation (*sequencing by hybridization* [SBH])

En 1988, une nouvelle technologie basée non pas sur la migration de fragments en électrophorèse mais sur l'hybridation est décrite. Cette technique reprend le principe du Southern blot mais à grande échelle et sur un support miniaturisé. L'idée proposée était de déterminer la fréquence d'hybridation de courtes séquences nucléotidiques à un ADN génomique, d'assembler l'ensemble de ces courtes séquences parfaitement hybridées en une séquence unique et de comparer celles-ci à un ADN de référence [16,46]. Ce principe a permis le développement des puces d'ADN et de nombreuses applications dont le SBH. Il est impossible de décrire simplement et brièvement cette technique. Par conséquent, nous décrirons schématiquement le principe de la technique, une revue récente sur les puces d'ADN et ses applications ayant été récemment publiée dans cette revue [9]. La fabrication de ces puces et le principe de séquençage qui en découle ont été rendus possibles grâce aux nombreuses avancées réalisées dans des domaines variés tels que la fabrication des puces informatiques (dont découle l'idée de puce d'ADN), les synthèses d'oligonucléotides, la photolithographie, les imprimantes à jet d'encre mais aussi grâce aux progrès de la biologie moléculaire, de l'optique physique, de la robotisation, de l'informatique, des progrès sur la connaissance des génomes et, bien évidemment, grâce à l'application des règles de Watson-Crick pour l'appariement des bases d'ADN. Schématiquement, les puces d'ADN constituent une collection de sondes courtes (oligonucléotides ou oligomères) fixées de manière ordonnée sur un support solide. Le principe général du SBH est basé sur l'idée que de longues séquences d'ADN peuvent être obtenues par le chevauchement de nombreux oligomères spécifiques (courte séquence d'ADN ou sondes) après hybridation entre l'ADN étudié et ces oligomères [71,73]. Par exemple, avec les trois octamères suivants :

- ATCAGGTC ;
- TCAGGTCT ;
- CAGGTCTG.

On peut définir l'unique décamère, ATCAGGTCTG. En connaissant la position des oligomères, leur séquence et les résultats de l'hybridation, à l'aide de calculs mathématiques, il est possible de reconstituer la séquence totale du fragment étudié. La longueur optimale des sondes nécessaires pour la fixation de l'ADN cible dépend de la complexité de ce dernier (Fig. 6 et 7). À titre d'exemple, il a été démontré que l'utilisation de sondes de 11—15 nucléotides fixées sur une surface étaient suffisantes pour l'analyse d'un ADN de 10^9 pb et pouvaient constituer une méthode de choix. Dans ce dernier cas, il a été calculé que $4{,}2 \times 10^6$ oligonucléotides de 11 pb étaient nécessaires. Toutefois, un défi consiste en la nécessité de discriminer les duplexes sans misappariement de ceux avec misappariement au cours de l'hybridation. Les premiers résultats furent concluants et encourageants [73]. D'autres études ont suivi. Ainsi, une première détermination du séquençage de l'ADN mitochondrial décrite en

1996 montrait les difficultés rencontrées en appliquant une telle technique [11]. En effet, dans ce travail, un certain nombre de limitations a été observé : nécessité de générer un ARN après transcription *in vitro* de l'ADN cible, séquençage d'un seul des deux brins et insuffisances du logiciel d'interprétation des données. En 1998, une étape supplémentaire a été franchie avec le séquençage des exons 5 à 8 du gène codant pour le gène p53. Pour séquencer les 2000 kb de l'ADN cible, les auteurs préparèrent 16 384 variants de 7 pb de longueur. Après amplification par PCR de l'ADN cible, celui-ci était hybridé aux sondes fixées sur un support [17]. Les résultats montrèrent cependant que toutes les mutations n'étaient pas mises en évidence (notamment, les délétions et insertions importantes) et que même si les résultats étaient confirmés à 100 % par la méthode de référence pour le séquençage, la méthode de Sanger (le séquençage « classique »), 10 % des sondes n'avaient pas hybridé. Enfin, certaines régions étaient difficilement analysables (par exemple, les régions riches en motifs CA). Les difficultés rencontrées pour ces motifs étaient probablement une conséquence de leur structure secondaire dont les caractéristiques d'hybridation aux sondes étaient modifiées. Une autre étude réalisée en 2000 a comparé les performances des puces de séquençage et la méthode de Sanger [77]. Pour cette étude, les auteurs ont analysé des mutations sur le gène codant pour la p53, l'un des gènes les plus fréquemment mutés dans les cancers (gène sur lequel plus de 600 mutations sont connues). Pour cette étude, la société Affymétrix, spécialisé dans la fabrication de puces d'ADN (*DNA chips*) a réalisé des puces constituées de 65 000 sondes de 18 pb chacune. À l'aide de ces sondes, l'analyse des exons 2 à 11 du gène a pu être réalisée en étudiant les brins sens et antisens. L'ensemble des sondes permettait l'analyse de la séquence normale et des mutations ponctuelles ainsi que les délétions d'une paire de base, la douzième base en partant de l'extrémité 3′ étant la base modifiée (A, C, G, T ou délétion d'1 pb). Par ailleurs, des sondes spécifiques de 300 mutations connues pour ce gène étaient également ajoutées. Dans ce cas, 12 sondes (six en sens et six en antisens) étaient synthétisées pour chaque mutation, la substitution sur chaque sonde étant localisée en différents points de celle-ci. Pour chaque position, il existait donc cinq sondes. Un logiciel et un scanner de puces spécifiques ont permis la lecture, puis l'interprétation des puces. La concordance (résultats identiques) entre la méthode de Sanger et la méthode SBH était de 81 %, avec un avantage pour le séquençage par SBH dont la détection atteint 94 % des mutations versus 87 % pour la méthode de Sanger. Néanmoins, la méthode par SBH n'a pas détecté certaines mutations (six mutations sur les 108 échantillons analysés dans cette étude). Celles-ci correspondaient à des délétions ou des insertions de 3 à 15 pb. Le taux de détections de minidélétions ou -insertions (> 1 pb) est donc moindre dans la technique par SBH. D'autres études comparant la technique SBH à la méthode de Sanger ont retrouvé des résultats similaires [2,28,29]. Quelques années plus tard, en 2004, une équipe a décrit le séquençage des 16,5 kb de l'ADN mitochondrial par cette technique. Faisant suite à la première étude réalisée en 1996, cette seconde génération de puce pour le séquençage de l'ADN mitochondrial par SBH a constitué un progrès important. Cette puce appelée MitoChip par les auteurs (fabriquée par la société Affymétrix) a permis le séquençage des 29 366 pb incluant l'ADN mitochondrial et les séquences plasmidiques de contrôle servant de contrôle positif d'hybridation. Cette puce contenait les sondes complémentaires des séquences sens et antisens de l'ADN mitochondrial (chaque sonde mesurait 25 pb). Pour détecter une mutation sur l'ADN, seule la treizième base de chaque sonde était modifiée et pouvait être une des quatre bases A, T, G ou C. La quantité d'ADN nécessaire était faible, puisque 300 ng d'ADN ont suffi pour ce séquençage (il faut 100 ng pour une PCR) et le nombre de PCR nécessaires au séquence réduit à trois (pour un séquençage selon la méthode de Sanger, le nombre de PCR est compris entre 12 et 32 selon les analyses). Avant l'hybridation, le produit issu de la longue PCR était fragmenté par une DNase et marqué pour permettre une lecture du fluorophore. Le résultat était aussi bon que ceux obtenus avec la méthode de Sanger, puisque la méthode permette d'obtenir 96 % de la séquence analysable et une reproductibilité de 100 %. Par ailleurs, la méthode a démontre sa grande sensibilité puisque l'étude a permis de montrer que les mutations pouvaient être détectées jusqu'à 2 % d'hétéroplasmie (mélange de mitochondries normales et anormales ; 2 % d'hétéroplasmie signifie ainsi 2 % de mitochondries mutées parmi l'ensemble des mitochondries de la cellule) nettement supérieure à la technique de Sanger (taux de détection de l'hétéroplasmie : environ 10 %) [48]. Le séquençage de l'ADN mitochondriale à l'aide de cette méthode a été repris par la suite dans d'autres études [41]. Elle reste cependant imparfaite (par exemple, dans les zones riches en GC responsables d'artéfacts). Bien entendu, la technique de séquençage par SBH a été décrite dans des domaines autres que la génétique humaine. Dans les années 2000, outre la société Affymétrix, d'autres sociétés se sont crées pour le développement de puces d'ADN et pour certaines, pour le séquençage par SBH. En 2008, une société domine cependant les autres, il s'agit de la société Affymétrix, précurseur dans ce domaine et leader incontesté des puces d'ADN (www.affymetrix.com). À titre d'exemple, la société Affymétrix a développé une puce de séquençage pour l'étude du génome de l'hépatite B [58] ou pour séquencer d'autres virus, par exemple le coronavirus responsable du *severe acute respiratory syndrom* (SARS) [78]. En pratique, les puces d'ADN pour séquençage ont démontré leur capacité à séquencer un génome, notamment dans le cadre du reséquençage. Cependant, dans cette indication, elles restent encore imparfaites, certaines mutations n'étant pas décelées par les puces, les points sensibles étant :

- les zones riches en GC ;
- les délections ou insertions importantes ;
- les séquences répétées.

De plus, leur coût important ne permet pas l'utilisation de cette technique en pratique quotidienne. Les puces sont cependant utilisées dans le domaine de la recherche, soit pour le reséquençage soit pour l'étude de variants de bases, les *single nucleotide polymorphism* (SNP). La société Affymétrix commercialise actuellement des puces permettant l'étude des polymorphismes humains répartis sur l'ensemble du génome (*genome wide human SNP array*). Cette puce permet l'étude de 1,8 millions de SNP mais aussi permet d'étudier les variations du nombre de copies de segments

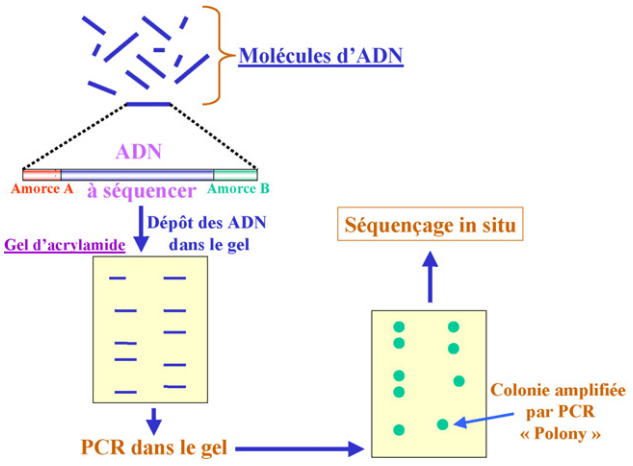

Figure 8 PCR sur colonies, principe de la *polony* et du séquençage sur *polony*. Un ensemble de molécules d'ADN issues d'une bibliothèque d'un ADN génomique (collection d'ADN) sur lesquelles ont été ajoutées des amorces universelles A et B en 3′ et 5′ par réaction de ligation sont incorporées dans un gel de polyacrylamide contenant tous les réactifs nécessaires à une PCR. Une seule molécule d'ADN est présente par site de dépôt dans le gel. L'amplification par PCR est donc réalisée dans le gel à partir d'une seule molécule. Chaque colonie correspond à une seule molécule d'ADN amplifiée par PCR. Après PCR, les colonies (*polonies*) sont dénaturées in situ et séquencées après addition d'une amorce de séquence. Le séquençage est réalisé par addition séquentielle et en série d'un seul nucléotide fluorescent (chaque nucléotide dA-, dG-, dC-, dTTP ayant une fluorescence spécifique). Après addition d'un nucléotide (par exemple, dATP), un lecteur type scanner lira la fluorescence si le nucléotide a été incorporé par l'ADN polymérase présente dans le gel. Après élimination par lavage des nucléotides non incorporés, le nucléotide suivant (par exemple, dCTP) est ajouté. La réaction continue.

d'ADN (*copy number variants* [CNV] ; 900 000 variants CNV détectés). Bien qu'il ne s'agisse pas à proprement parler de séquençage mais de la recherche de variants alléliques identifiés, il s'agit d'une technique similaire à la SBH et qui présente probablement un avenir plus prometteur pour les puces d'ADN que la technique par SBH. L'avenir des puces semble donc essentiellement tourné vers l'identification de marqueurs génétiques (génotypage) tels que des SNP ou des CNV ainsi que l'analyse d'expression des gènes plutôt que vers le séquençage proprement dit [5,9,26]. Certains auteurs continuent cependant à travailler sur des développements de cette technique, telle que la *shotgun*-SBH exposée dans la seconde partie [60].

La méthode de PCR sur colonies (dite amplification clonale) ou *PCR colonies* (*polony*)

La méthode *polony* a été décrite pour la première fois en 1999 pour l'étude de génomes bactériens [53]. Le principe de cette technique consiste à séparer physiquement des fragments d'ADN génomique, puis de les amplifier de telle sorte qu'ils restent séparés (une molécule d'ADN amplifiée par PCR et par colonie) (Fig. 8). Les produits amplifiés sont appelés *polony* (abréviation de *PCR colony*). Le séquençage peut ensuite se faire sur chaque molécule d'ADN individuellement et en parallèle par incorporation d'un dNTP fluorescent. Après addition du dNTP, le fluorophore est clivé (chimiquement ou photochimiquement) et la fluorescence émise recueillie pour analyse informatique. Un second dNTP fluorescent est alors ajouté et le processus se répète ainsi. Ce système de séquençage est aussi appelé séquençage par fluorescence in situ (*fluorescence in situ sequencing* [FISSEQ]). Mille à 10 000 molécules peuvent ainsi être séquencées parallèlement et indépendamment. Initialement décrite par immobilisation de l'ADN sur un gel d'acrylamide [52,54], cette technique de séparation des molécules d'ADN a inspiré d'autres développements tels que la PCR émulsion et la capture sur billes (voir seconde partie). Il a ainsi été possible de séquencer 30,1 millions pb d'une souche d'*E. coli* en 2,5 jours avec une précision de 99,7 % [69].

Le pyroséquençage

Cette nouvelle technique a été publiée pour la première fois en 1998 alors que le principe a été décrit en 1985 [1]. Avec les nouveaux automates de séquençage, elle est en train de supplanter progressivement la méthode de Sanger. Il s'agit d'une méthode permettant d'analyser la synthèse d'ADN cible en temps réel. On parle de séquençage par synthèse d'ADN [63]. Le principe de base de la méthode consiste à hybrider une amorce à l'ADN cible (amplifié par PCR), puis à ajouter séquentiellement et dans l'ordre une base à partir de l'extrémité 3′ de l'amorce. Chaque base est marquée par un fluorophore différent dont le signal est mesuré par bioluminescence à condition que la base complémentaire de la cible soit incorporée. La séquence est déduite en fonction de l'ordre d'incorporation des nucléotides sur l'ADN complémentaire de la cible néosynthétisée. Quatre enzymes sont nécessaires pour la réaction : une ADN polymérase, une ATP sulfurylase, une luciférase et une apyrase. Le mélange réactionnel contient, par ailleurs, les substrats de ces différentes enzymes : adénosine phosphosulfate (APS), D-luciférine, l'amorce de séquence (complémentaire de l'ADN cible). Les nucléotides alphathio-dATP (dATP-αS), dCTP, dGTP, dTTP sont ajoutés de manière cyclique un par un, toujours dans le même ordre et successivement. Une caméra CCD mesure le signal de bioluminescence produit (Fig. 9 et 10). Cette méthode performante totalement automatisée permet de séquencer de courts fragments d'ADN (en moyenne, 60 bp pour l'automate commercialisé par Biotage et en moyenne 106 pb sur l'automate 454 commercialisé par Roche), voire, dans certains cas, jusqu'à 200 pb [24]. La limitation du nombre de bases séquencées est liée à l'inhibition progressive de l'apyrase par l'accumulation de déoxymononucléotide phosphate (dNMP) et de son produit intermédiaire le déoxydinucléotide phosphate (dNDP, Fig. 9) ou l'élimination incomplète des nucléotides résiduels après lavage [51]. La société suédoise Biotage commercialise des automates de pyroséquençage (www.biotage.com). Cette technique est également utilisée sur d'autres automates de séquence tel que le 454 (société Roche, voir seconde partie). Le pyroséquençage, du fait de la faible longueur des bases pouvant être lues, a eu longtemps une indication limitée dans le séquençage. L'arrivée d'automates, tel que le 454 permettant le séquençage massif parallèle, a permis le développement de cette technique à une échelle bien plus

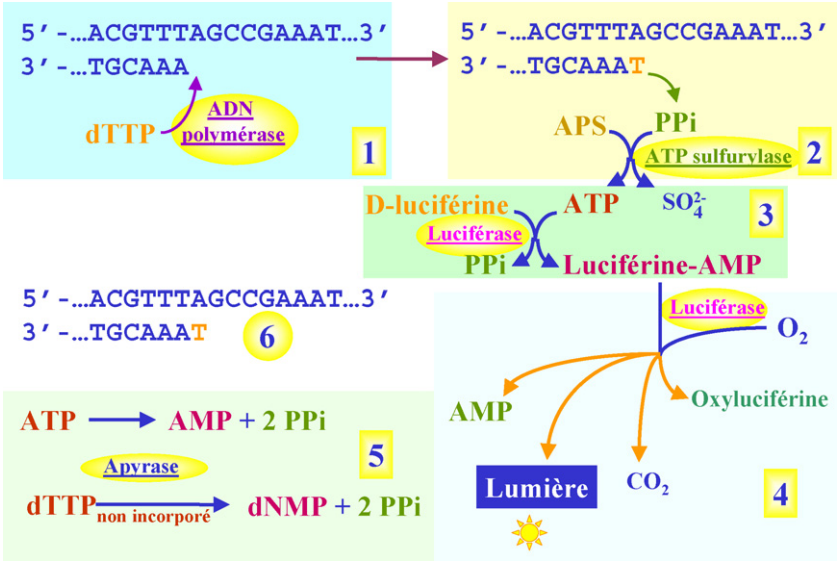

Figure 9 Principe du pyroséquençage. Il s'agit d'une technique d'addition séquentielle de nucléotides en temps réel. Prenons le premier cas de figure où l'automate ajoute la base complémentaire T à l'ADN cible. **1** : l'ADN polymérase ajoute le déoxynucléotide dTTP à l'ADN cible à partir de l'amorce de séquence déjà hybridée. Par cette réaction de polymérisation, un pyrophosphate inorganique (PPi) est libéré. **2** : le pyrophosphate inorganique réagit avec l'APS (adénosine phosphosulfate), substrat de l'enzyme ATP sulfurylase produisant ainsi de l'ATP (adénosine triphosphate). **3** : L'ATP réagit avec un substrat de la luciférase, la D-luciférine aboutissant à la production d'un produit intermédiaire, la luciférine-AMP. **4** : la luciférine-AMP en présence d'oxygène est transformée immédiatement par la luciférase en oxyluciférine, en CO_2, en AMP et en un signal lumineux mesuré par la caméra CDD. À noter que dans la réaction de pyroséquençage, la luciférase permet la production de lumière à partir d'ATP. Elle peut aussi en produire à partir de dATP. Ce nucléotide est donc remplacé dans la réaction de pyroséquençage par un nucléotide modifié, le dATP-αS [60]. **5** : le dTTP en excès non incorporé ainsi que l'ATP sont ensuite dégradés par l'apyrase (sur l'automate de la société Biotage). Dans un autre système de pyroséquençage (automate 454 de la société Roche), la séquence cible est capturée sur des billes et les dNTP et l'ATP non incorporés au cours de la réaction sont éliminés par lavage et non par action d'une apyrase. L'addition d'un autre nucléotide peut avoir lieu. Cette réaction est cruciale. En effet, elle permet de s'assurer que le signal lumineux mesuré correspond au nucléotide spécifiquement ajouté. **6** : l'automate a donc mesuré un signal lumineux pour la base T. La réaction peut se continuer par l'addition d'un autre nucléotide. Lorsqu'un nucléotide non complémentaire est ajouté par l'automate (par exemple, un dATP en regard de la base G de l'ADN cible), celle-ci ne sera pas incorporée par l'ADN polymérase et sera directement dégradée par l'apyrase. Il n'y aura pas de signal lumineux.

grande (voir seconde partie). Le pyroséquençage est beaucoup utilisée pour l'étude de variants alléliques, notamment les polymorphismes biallélliques (SNP). Du fait de l'addition séquentielle de bases, lorsque les SNP ne sont pas trop éloignés les un des autres, cette technique permet l'étude directe d'haplotypes (association de plusieurs SNP sur un même chromosome). Il s'agit de la seule technique actuelle capable de déduire un haplotype de l'analyse directe du génome. Cette technique est également appliquée dans le génotypage bactérien et viral ainsi que dans l'analyse de méthylation en épigénétique.

Le séquençage par spectrométrie de masse

La spectrométrie de masse a évolué considérablement ces 20 dernières années. Bien que réservée à certains laboratoires de recherche, ses applications en biologie moléculaire sont nombreuses et toujours en évolution. Plusieurs variantes de spectrométrie de masse existent (par exemple, *matrix-assisted laser desorption ionization time-of-flight mass spectrometry* [MALDI-TOF MS], ionisation par électrospray ou par analyse de Fournier). Bien qu'actuellement abandonné pour le séquençage, nous évoquerons schématiquement le principe du séquençage de la MALDI-TOF MS. En effet, les recherches sur le séquençage à l'aide de cette technique ont permis de l'améliorer considérablement et de faire évoluer cette technique vers d'autres utilisations en biologie moléculaire [50]. Cette technique, décrite pour la première fois en 1988, a d'abord été rapportée pour l'analyse des protéines. L'analyse des acides nucléiques s'est fait dans un second temps (début des années 1990). Le principe de la méthode de spectrométrie MALDI-TOF est simple. L'ADN cible est séché à température ambiante sur une matrice constituée d'acide hydroxypicolique. Cette substance possède la propriété d'absorber les UV sans interagir avec l'ADN. L'ADN est exposé à de courtes impulsions laser UV (dont l'énergie est absorbée par la matrice) désorbant ainsi l'ADN dans la phase gazeuse (l'ADN sous forme d'ions est expulsé de la matrice). Les ions ADN sont monovalents et l'ADN intact. Juste après la désorption, une impulsion électrique de grande intensité permet l'extraction, puis l'accélération de l'ADN ionisé dans un champs électrique. L'ADN parcourt alors une distance connue dans le vide (dans un tube d'environ 1 m de long) pendant un temps donné (*time of flight*) et acquiert une certaine énergie cinétique. Le temps relatif pour parcourir cette distance est proportionnel à la masse de la molécule. En fin de parcours, l'ADN rentre en collision avec un

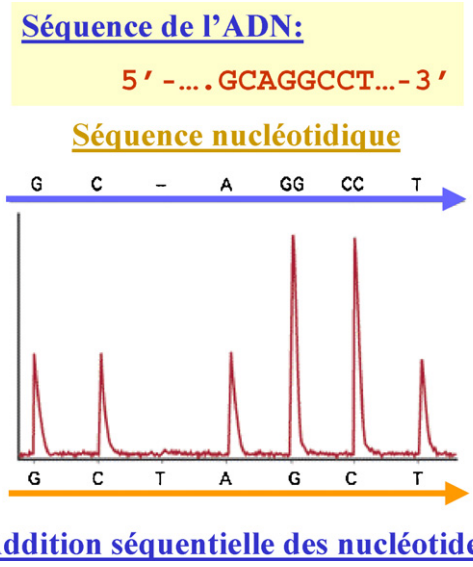

Figure 10 Exemple de pyrogramme. En abscisse, en dessous du diagramme, l'addition séquentielle des nucléotides par l'automate. En ordonnée, l'intensité du signal lumineux mesuré. Lorsque plusieurs bases successives sont présentes sur la séquence, le signal mesuré est proportionnel au nombre de bases identiques présentes. Ainsi, lorsque deux G se suivent (séquence « GG »), l'intensité du pic lumineux est double. Lorsque aucune base n'est ajoutée, le nucléotide est dégradé (absence de complémentarité avec la séquence ADN cible) : il n'y a pas de signal lumineux.

détecteur mesurant le temps de parcours de la molécule depuis l'impulsion laser de départ. La première description de cette application pour le séquençage consistait à réaliser un séquençage classique par la méthode de Sanger, puis de détecter les molécules par MALDI-TOF, la masse des monobrins séquencés étant détectée (au lieu de les séparer dans un gel comme dans la technique électrophorétique) [27]. Le MALDI-TOF présente certains avantages : il est automatisable et rapide (un spectre est analysé en une milliseconde et en quelques secondes, une centaines de spectres peuvent être obtenus). Le rapport charge/masse mesuré par l'appareil est une propriété de l'ADN monobrin indépendante de sa structure secondaire. En 1996, il était possible de séquencer en moyenne 89 pb [70]. Depuis cette date, des progrès considérables ont été réalisés dans la rapidité et la masse de données analysables alors que le nombre de bases séquencées ne changeait pas, limitant son intérêt. La société Sequenom (www.sequenom.com) s'est spécialisé dans le MALDI-TOF appliqué à la biologie moléculaire. Actuellement, cette technique s'est orientée vers des applications telles que le génotypage ou l'analyse de méthylation, le séquençage ayant été abandonné [80].

Le séquençage d'une seule molécule d'ADN

En routine, le séquençage en milieu hospitalier est réalisé sur de l'ADN génomique le plus souvent amplifié par PCR. L'ADN est extrait à partir d'un groupe plus ou moins important de cellules. Le séquençage d'une seule molécule d'ADN représente une technique alternative capable de séquencer une seule molécule d'ADN à la fois (amplifiée ou non au préalable) [4]. Ce principe possède deux avantages : d'abord, il n'est pas nécessaire d'amplifier l'ADN ou, si c'est le cas, peu de cycles d'amplification sont nécessaires, ensuite, le séquençage peut être réalisé en temps réel (absence de cycles répétés ou de succession de réactions enzymatiques). Nous citerons brièvement ces méthodes développées dans la seconde partie :

Technique de détection par sonde atomique (scanning probe). Elle utilise un instrument développé pour les nanotechnologies, le microscope par force atomique. La sonde de ce dernier scanne chaque base d'ADN pour en déduire la sequence.

Le séquençage par l'exonucléase. Bien que décrite dans les années 1980, cette technique est redevenu d'actualité. Après transcription d'un ADN par incorporation de nucléotides fluorescents (fluorophores spécifique pour chaque nucléotide) à l'aide d'une ADN polymérase, chaque molécule d'ADN est fixée sur un support (bille, par exemple), puis circule dans un microcapillaire où elle est digérée par une exonucléase. Cette enzyme libère séquentiellement chaque nucléotide fluorescent dont la lecture se fait en temps réel dans un canal microfluidique.

Le séquençage par synthèse ou cyclic reversible termination (CRT) [64]. Dans cette approche, en parallèle et de manière cyclique, chaque molécule d'ADN est séquencée par addition du nucléotide complémentaire (fluorescent) catalysée par une enzyme. La réaction est suivie en temps réel par une caméra. Chaque nucléotide intégré possède un groupe protecteur qui arrête la synthèse de l'ADN. Le fluorophore est ensuite éliminé ainsi que le groupe protecteur par une autre enzyme. Le cycle peut redémarrer. Quelques sociétés développe ce procédé (par exemple, Pacific Biosciences, www.pacificbiosciences.com ; VisiGen, www.visigen.com).

Le séquençage après traversée de nanopores. Une molécule d'ADN traverse un nanopore [62]. Le passage de chaque nucléotide à travers un nanopore (par exemple, une alpha-hémolysine) soumis à un courant électrique provoque une variation de ce courant nucléotide-dépendant au cours de la traversée du nanopore. Ce principe décrit en 1996 semble prometteur [37]. Cette technique est rapide, ne nécessite pas de réactifs ni d'amplification de l'ADN [62].

Le séquençage du génome humain, le projet Hugo

La première séquence connue d'un être vivant date de 1977 : il s'agit génome du bactériophage ΦX174 séquencé par la méthode de Sanger [65]. En 1998, le premier génome d'un animal (en fait, un ver de terre), *Caernohabditis elegans*, fut publiée [67]. En 1990, un projet considéré comme fou fut initié par les américains : ce projet surnommé *Human Genome project* (Hugo) avait pour objectif de déterminer la séquence complète de l'ADN génomique humain (contenu dans le noyau de la cellule). Ce projet colossal a duré 13 ans et a réuni plusieurs équipes internationales (la majorité d'entre elles américaines). En parallèle, des instituts publics (d'abord américain, puis rejoint par des institutes européennes et asiatiques) et un institut privé Celera Genomics se lançaient aussi dans cette course. Bien que la stratégie choisie de séquençage

Tableau 3 Exemple du séquençage de J. Craig Venter, 62 ans, américain d'origine britannique en bonne santé : quelques données [42].

Échantillon	Sang total Arbre généalogique présenté sur trois générations
Technique de séquençage	Méthode de Sanger (séquençage *de novo*) — lecture de 800 pb en moyenne Méthode *shotgun* tout-génome Lecture de 32 millions de séquences (7,5 fois le génome humain) Résultat par alignement de séquences multiples
Caryotype	Normal
Méthode de confirmation des SNP	Puces d'ADN Affymétrix Billes de génotypage Illumina
Méthode pour évaluer les variants du nombre de copies de gène, *copy number variants* (CNV)	Puces d'hybridation génomique comparative appelées aussi *comparative genomic hybridization* (CGH)
Nombre de gènes potentiels	20 000—25 000 11 250 gènes de fonction connue (58 % des gènes)
Proportion de l'ADN codant pour un gène dans le génome	2 %
Nombre de variants (polymorphismes SNP, microdélétions, microinsertions, insertions/délétions [indéls], inversions, duplications...) — étude d'haplotypes	4 195 960 (dont 78 % SNP) dont 1 288 319 variants nouveaux, soit 12,3 Mb (référence utilisée : base de données du *National Center for Biotechnology Information* (NCBI) dont 62 % répertoriés dans la base de données de SNP, dbSNP)
Relation génotype/phénotype — Interprétation difficile (exemples) 44 % des gènes comprennent au moins un variant à l'état hétérozygote	*Antécédents familiaux de maladie cardiovasculaire* Hétérozygote dans le gène *Klotho* (*KL*, Omim 604824) pour deux variants (F352V et C370S) associé à une diminution du risque cardiovasculaire Homozygote 5A/5A dans le promoteur du gène *matrix metalloproteinase 3* (*MMP3*, Omim 185250) associé à une augmentation du risque d'infarctus du myocarde. Nombreux variants aux conséquences inconnues ou paradoxales (rôle des gènes modificateurs, de leurs produits et de l'environnement). Par exemple Délétion hétérozytoge de 4 pb du gène ACOX2 (*Acyl-CoA oxidase 2*, Omim 609751) créant une protéine anormale tronquée. Cette enzyme est impliquée dans le métabolisme lipidique. Génotype pour le gène *lactase* (*LCT*, Omim 603202) compatible avec une intolérance au lactose, non rapportée par le sujet.
Difficultés principales	Séquençage des zones répétées Alignement des séquences Confirmation des séquences variantes (notamment délétions, insertions, indéls) Incertitudes sur certains variants : nécessité de couvrir plus de sept fois le génome humain pour atteindre un taux d'exactitude du séquençage proche de 100 % (au moins, 20 fois idéalement)

ait été différente entre les deux groupes (public et privé), dans tous les cas, la technique de séquençage utilisée a été la méthode de Sanger. Le premier chromosome complètement séquencé fut le chromosome 22 en 1999. En 2001, 95 % de l'ADN humain était séquencé [39,76]. Le séquençage final fut achevé en mai 2006 avec le séquençage complet du chromosome 1. Il est impossible de décrire ici l'histoire incroyable de cette aventure humaine : pour plus d'informations sur ce projet scientifique mondial, nous renvoyons le lecteur au site internet www.ornl.gov/sci/techresources/Human_Genome/home.shtm (en anglais).

Tableau 4 Quelques caractéristiques des principales techniques de séquençage.

	Méthode de Sanger	Pyroséquençage	SBH	Séquençage en temps réel d'un seul ADN	Séquençage par synthèse	Séquençage d'une seule molécule (nanotechnologies)
Avantages	Technique bien rodée Lecture de longs fragments (500—800 pb voire 1000 pb) Faible taux d'erreur Miniaturisation possible	Lecture de courts fragments (100 pb)	Intéressant pour Le reséquençage L'étude de SNP	Séquençage d'une molécule au sein d'un mélange complexe Utilisable en séquençage de novo et en reséquençage	Séquençage d'une molécule au sein d'un mélange complexe Utilisable en séquençage de novo et en reséquençage	Quantité minime d'ADN
Inconvénients	Faible rendement de lecture (longue durée d'analyse d'un génome)		Lecture de courtes séquences (25 pb) Mal adapté au séquençage de novo Séquences répétées non analysables	Non commercialisé	Lecture de courtes séquences (15 pb)	Perte de signal possible si ADN altéré ou échec du signal

Tableau 5 Défis du séquençage et exemples de paramètres modulant les performances des séquenceurs.

Réduire les coûts
- Parallélisation massive (10^5 à 10^8 réactions en parallèle)
- Miniaturisation

Optimiser le débit de séquençage
- Degré de parallélisation
- Vitesse de détection du signal
- Intérêt des nanopores (débit de 100—1000 bases/seconde versus 0,17 base/seconde pour l'électrophorèse capillaire)

Temps d'utilisation de la machine (absence de travail continu)
- Degré d'automatisation
- Temps entre les migrations appelés aussi « *runs* » (réaction et/ou migration et/ou analyse des échantillons)
- Temps de travail de l'opérateur
- Réanalyse des échecs de séquençage

Réduire le taux d'erreurs
- Contrôle de qualité intra-analyse
- Reséquencer une/plusieurs fois un même échantillon (lectures multiples d'un même échantillon)

Séquençage complet (sans zones non analysées)
- Difficulté d'analyse de certaines séquences, par exemple, les séquences répétées mononucléotidiques (cf. poly [A]), les séquences en épingle à cheveu (*hairpin*)
- Longueur de la séquence analysée (Sanger : jusqu'à 800 pb versus pyrosequencing et autres systèmes, 25—100 pb).

La totalité du génome est plus difficile à obtenir avec de courtes séquences.

Les courtes séquences (25—100 pb) sont un moyen puissant de reséquencer un génome alors que le séquençage de novo nécessite le plus souvent d'obtenir des longueurs supérieures à 100 pb

Les deux premiers séquençages individuels : les génomes de James Watson et de Craig Venter

En juin 2007, était publié le premier génome humain complet d'une seule personne à savoir James Watson, âgé de 79 ans, codécouvreur de la structure de l'ADN (accessible sur le site web : jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence). Alors que le projet de séquençage du génome humain a coûté trois milliards de dollars et a duré 13 ans (achevé en 2006), celui de James Watson a coûté un million de dollars et a été réalisé en deux mois. Il a été effectué sur un séquenceur Genome Sequencer FLX (société 454 Life Sciences, Baylor College of Medicine, Houston, Texas, États-Unis, voir seconde partie de l'article). Quatre mois après, l'institut Craig Venter publiait le génome complet de Craig Venter [42]. Contrairement à celui de James Watson, celui-ci a été séquencé selon la technique classique de Sanger. Ces deux séquençages individuels constituent une étape majeure vers la médecine personnelle. Par ailleurs, il existe actuellement une compétition entre plusieurs sociétés (par exemple, les sociétés Illumina, Applied Biosystems et 454 Life Sciences) pour réaliser le séquençage de 100 génomes

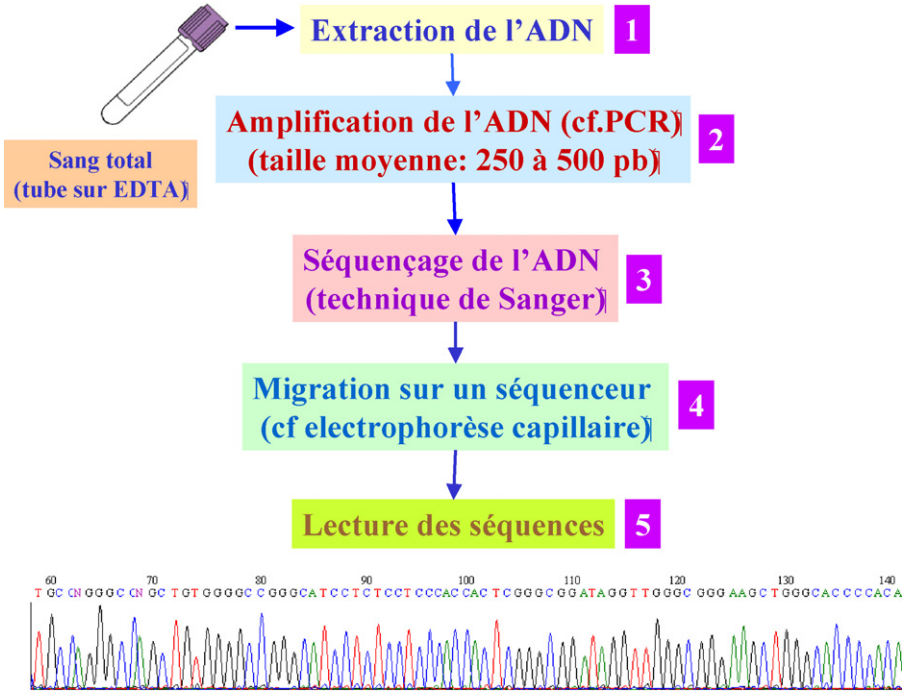

Figure 11 Le séquençage en milieu hospitalier : exemple pratique. **1** : bien que tout type d'échantillon biologique puisse être utilisé, le séquençage d'un sujet est essentiellement réalisé à partir de sang total. Du sang est prélevé dans un tube contenant l'anticoagulant EDTA (anticoagulant qui n'inhibe pas l'ADN polymérase utilisée pour la PCR). De nombreux kits d'extraction d'ADN sont disponibles et permettent une extraction rapide de l'ADN. Lorsque le laboratoire extrait de nombreux échantillons, il peut aussi utiliser un automate d'extraction. L'ARN peut aussi être séquencé après extraction. Dans ce cas, il est nécessaire de le transformer en ADN complémentaire (ADNc) par une réaction de transcription inverse (ou reverse transcription). **2** : pour séquencer l'ADN, ce dernier doit d'abord être amplifié. La *polymerase chain reaction* (PCR) est la principale technique d'amplification utilisée. La taille des échantillons amplifiés est variable. Selon la zone à amplifier, elle peut aller de 250 pb à 500 pb en moyenne. À titre d'exemple, dans le cas de gènes, il est habituel de séquencer les exons avec leurs jonctions exon/intron, la partie 3′ non codante (NC) du gène ainsi que le promoteur. Le nombre de PCR pour un gène varie donc en fonction de la taille de ce dernier. Ainsi, un gène possédant cinq exons pourra avoir au minimum sept PCR (une pour le promoteur, une pour chaque exon [et jonctions exon/intron] et une pour la partie 3′NC). **3** : après PCR, il est le plus souvent nécessaire de purifier le produit d'amplification. Des kits permettent de réaliser cette étape. La réaction de séquençage est ensuite réalisée à l'aide d'autres kits spécifiques. Les recommandations internationales préconisent de séquencer chaque produit amplifié à l'aide d'amorces de séquence sens et antisens. Par conséquent, deux réactions de séquence au minimum sont réalisées pour chaque fragment amplifié. Ces amorces peuvent être les mêmes ou différentes de celles utilisées pour la PCR. Ces réactions ont lieu en général en microplaques (de 96 puits, par exemple). Les réactions de séquence sont ensuite purifiées pour éliminer notamment les désoxynucléotides, les didéoxynucléotides et l'amorce de séquence non incorporés ainsi que l'enzyme. **4** : la microplaque contenant les réactions de séquences purifiées est alors déposé dans un automate de séquence pour permettre la migration des échantillons (par exemple, par électrophorèse capillaire), la lecture se faisant après excitation par un laser. Les signaux sont transmis à un ordinateur qui permet leur interprétation à l'aide d'un logiciel spécifique. Ainsi, avec un séquenceur 16 capillaires, pour des fragments de 350 pb, le temps nécessaire pour la migration d'une plaque de 96 échantillons et son analyse est d'environ 3 heures. **5** : La lecture des séquences est effectuée soit manuellement et visuellement, séquence par séquence, soit à l'aide de logiciel(s) permettant automatiquement la détection de variant(s) (mutation ou polymorphisme) au sein de la séquence. Ces logiciels ne sont pas fiables à 100 %. Il est donc souvent nécessaire de contrôler visuellement les séquences.

Tableau 6 Estimation des performances de quelques techniques de séquençage.

Technologie	Coût de la base (en dollars 2007)	Débit (bases/seconde)	Exactitude (%)	Longueur de la séquence lisible (paires de base)
Sanger	10^{-3}	24	99,7	~800
Pyroséquençage	$25{,}10^{-5}$	1234	99	~110
Polony	$11{,}10^{-5}$	400	99	~12
Génome à 1000 $	$3{,}10^{-8}$	100 000	99,7	>60

Base du calcul : étude d'un génome diploïde, couverture de séquençage multipliée par six [64].

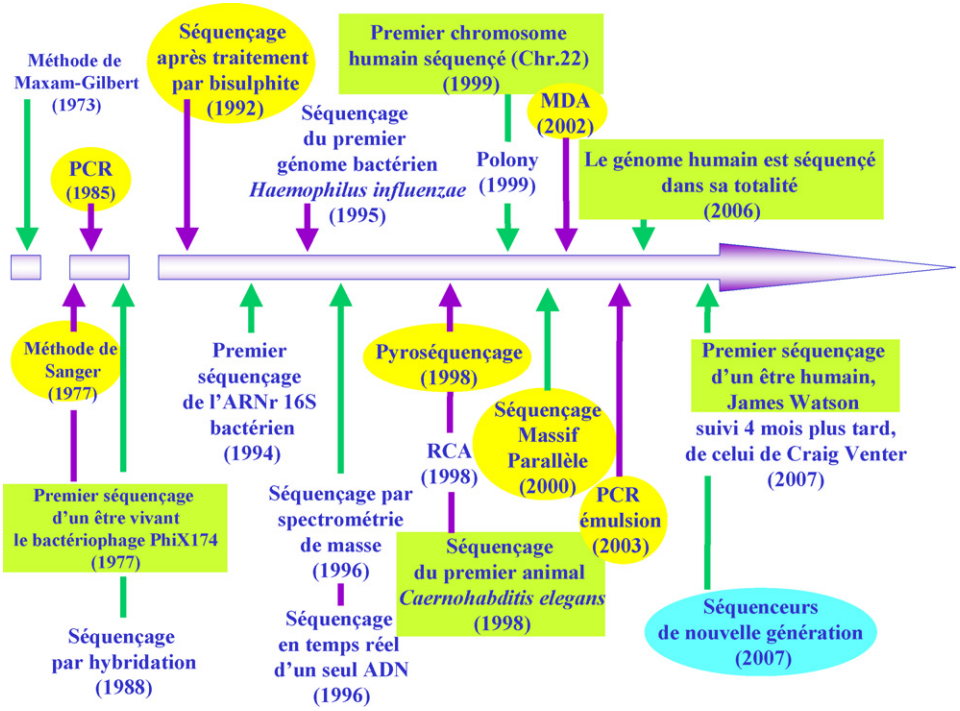

Figure 12 Quelques étapes importantes démontrant l'évolution des progrès du séquençage.

humains en dix jours : le gagnant de cette compétition obtiendra dix millions de dollars (le prix Archon Genomics de la Fondation X Prize). En parallèle de ces différents projets importants, en 2003, un autre projet ambitieux a été lancé, le projet de génome personnel (*Personal Genome project*) proposé par George Church de la célèbre Harvard Medical School de Boston, États-Unis. Ce projet ne semblait pas réalisable à cette période. L'évolution des techniques de séquençage rendent plus probable et même certain sa réalisation dans un très proche avenir. Dans ce projet (accessible sur le site www.personalgenomes.org), George Church souhaite séquencer 100 000 volontaires et intégrer ces données à celles de l'environnement et leurs données phénotypiques [8]. Mais, quel peut-être l'intérêt de ces séquençages et que nous apprennent-ils ? Par exemple, que nous a apporté le séquençage de James Watson ? Sept mille quatre-vingt-treize polymorphismes bialléliques (SNP) ont été trouvés. Deux mille huit cent quatre-vingt-neuf d'entre eux sont considérés comme bénins, c'est-à-dire sans conséquence fonctionnelle, 522 sont possiblement inducteurs de maladie, 286 probablement pathologiques et 201 dont les conséquences possibles ne sont pas connues. Bien qu'actuellement, ce résultat soit décevant dans le sens où il ne permet pas de prédire précisément les risques pour lesquels une prévention serait possible pour James Watson ou Craig Venter (Tableau 3), il introduit le concept de médecine génomique individuelle dans une pratique (pas encore quotidienne) et non plus en théorie : nous ne sommes plus dans le domaine de la science-fiction. Il s'agit donc d'un pas important vers le séquençage génomique individuel. Même si, à ce jour, l'interprétation reste décevante, on peut s'attendre à une prochaine carte génomique individuelle précise avec une interprétation claire des risques encourus dès la naissance pour chacun d'entre nous. Bien évidemment, des problèmes éthiques apparaissent dès que ce sujet est abordé, problèmes qui sortent du cadre du présent article.

Encore un défi : le génome humain individuel à 1000 dollars

Le projet de séquençage du génome humain qui a duré 13 ans a coûté environ trois milliards de dollars. En 2004, le NIH américain (l'équivalent de l'Inserm aux États-Unis) a lancé un nouveau défi : le génome d'un seul être humain pour un coût total de 1000 dollars [68]. En 2008, l'objectif n'est pas encore atteint. Les séquenceurs de nouvelle génération basés sur l'analyse massive en parallèle de l'ADN semblent prometteurs pour atteindre cet objectif. La seconde partie de cet article exposera les séquenceurs de nouvelle génération. D'autres projets de séquençage de nombreux individus et à prix bas sont en cours de réalisation [36]. Bien qu'encore réservé aux instituts de recherche ayant des moyens financiers considérables, le passage au séquençage total individuel est en passe de devenir une réalité. Une révolution biomédicale est en cours. La seconde partie vous présentera les faits.

À quoi sert le séquençage ?

Le séquençage du génome humain et d'autres organismes vivants constitue une étape majeure pour comprendre l'organisation des êtres vivants. L'étude du génome humain permet l'élucidation toujours en progrès de notre fonctionnement, de nos différences et de nos similitudes (études des polymorphismes et des variations du nombre de copies dans le génome). Du séquençage sont sorties de nouvelles disciplines biologiques, telles que la génomique,

la transcriptomique, la protéomique... L'étude des gènes, de leur variabilité, de leur expression, de leur régulation, de leur fonctionnement, de leur organisation participent à la compréhension du vivant et permettront des applications dans le domaine de la santé, de la prévention et des traitements. Le séquençage comparatif de plusieurs génomes humains n'en est qu'à ses débuts [38]. Certains gènes impliqués dans des maladies héréditaires à transmission mendélienne, notamment, étaient déjà connus avant le séquençage du génome humain (le gène *CFTR* et la mucoviscidose, par exemple). Depuis sa complétion, une foison de gènes impliqués dans de nombreuses pathologies ont été découverts (voir le site Omim pour les maladies à transmission mendélienne, *Online mendelian inheritance in men*, www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM). Par ailleurs, le projet de séquençage humain et le reséquençage en général ont été possibles grâce à l'essor de la bio-informatique, outil clé de la génomique. L'étude des variabilités humaines au niveau des SNP a permis la réalisation d'une carte des variations génétiques humaines à l'échelle mondiale (projet HapMap, voir le site www.hapmap.org). Le séquençage est donc une méthode fondamentale en médecine humaine et dans de nombreuses autres disciplines biologiques.

Les nouveaux défis du séquençage

En médecine, et plus particulièrement en génétique humaine, le séquençage a permis l'analyse des maladies monogéniques à transmission mendélienne, puis a évolué vers la description de variations dans les maladies polygéniques et multifactorielles. De nombreuses avancées ont déjà été réalisées et les progrès continuent (Tableau 4). Même si certains problèmes ne sont pas encore résolus (Tableau 5), de véritables usines à séquençage ont été mises en place et participent à la connaissance des génomes (humain et autres). L'arrivée du séquençage massif en parallèle permet désormais le séquençage de grands fragments de génome (> 1 Mb) à des prix de plus en plus bas [57] (Tableau 5).

Conclusion

Le séquençage de l'ADN par la méthode de Sanger est actuellement la méthode de choix dans les laboratoires hospitaliers et de recherche (Fig. 11). Sans cesse améliorée depuis plus d'une dizaine d'années, elle semble atteindre aujourd'hui ses limites même si des améliorations, notamment la miniaturisation de cette technique sont en développement [20,21]. Outre les limites technologiques, le coût du séquençage selon cette méthode est encore élevé (Tableau 5). Nous avons évoqué de manière schématique certains aspects des nouvelles technologies en cours de développement. La Fig. 12 résume quelques étapes des progrès du séquençage. La seconde partie de cet article traitera des nouvelles générations de séquenceurs et des révolutions que ces nouvelles machines apportent et apporteront dans un futur proche [7]. Par ailleurs, il est important de noter que l'arrivée dans un futur proche du séquençage tout génome (humain) posera de nombreux problèmes éthiques [10,34]. Il est aussi important de remarquer que même si le séquençage des génomes quelles que soit leur origine (animale ou non) apporte des informations fondamentales, elles ne suffisent pas à comprendre l'ensemble des phénomènes observés au niveau des interactions moléculaires de la cellule et de son environnement. Parmi les mécanismes agissant sur les génomes, outre l'environnement et d'autres facteurs, les modifications épigénétiques (qui sortent du cadre de cet article) jouent également un rôle fondamental dans l'expression des gènes et le fonctionnement du génome [6] (Tableau 6).

Sites internet

La première séquence humaine d'un être humain clairement identifié, celle de James Watson codécouvreur de l'hélice de l'ADN : http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/.

De nombreuses bases de données de séquences (eucaryotes, procaryotes, archéobactéries, viroïdes, virus, parasites...) sont disponibles sur internet. Il est impossible de tous les citer. À titre d'exemple, nous donnons quelques références.

Base de données de génomes (procaryotes et eucaryotes).

www.ncbi.nlm.nih.gov/sites/entrez?db=genome.

De nombreuses bases de données sont disponibles en dehors de ce site.

Base de données de bactéries.

Microbial Genome Database for Comparative Analysis (MBGD) : mbgd.genome.ad.jp.

Projets de génomes en cours d'étude.

www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj.

Références

[1] Ahmadian A, Ehn M, Hober S. Pyrosequencing: history, biochemistry and future. Clin Chim Acta 2006;363:83—94.
[2] Ahrendt SA, Halachmi S, Chow JT, Wu L, Halachmi N, Yang SC, et al. Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array. Proc Natl Acad Sci U S A 1999;96:7382—7.
[3] Ameziane N, Bogard M, Lamoril J. Principes de biologie moléculaire en biologie clinique. Collection Campus Référence Elsevier; 2005.
[4] Bayley H. Sequencing single molecules of DNA. Cur Opin Chem Biol 2006;10:628—37.
[5] Beaudet AL, Belmont JW. Array-based DNA diagnostics: let the revolution begin. Ann Rev Med 2008;59:113—29.
[6] Beck S, Rakyan VK. The methylome: approaches for global DNA méthylation profiling. Trends Genet 2008;24:231—7.
[7] Blow. DNA sequencing: generation next-next. Nat Methods 2008;5:267—74.
[8] Blow N. The personal side of genomics. Nature 2007;449:627630.
[9] Bogard M, Ameziane N, Lamoril J. Microarray d'ADN et profils d'expression des gènes. Première partie : concept, fabrication et mise en œuvre. Immunoanal Biol Spec 2008;23:71—88.
[10] Caulfield T, McGuire AL, Cho M, Buchanan JA, Burgess MM, Danilczyk U, et al. Research ethics recommendations for whole-genome research: consensus statement. PLoS Biol 2008;6:e73.

[11] Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610—4.
[12] Croft Jr DT, Jordan RM, Patney HL, Shriver CD, Vernalis MN, Orchard TJ, et al. Performance of whole-genome amplified DNA isolated from serum and plasma on high-density single nucleotide polymorphism arrays. J Mol Diagn 2008;10: 249—57.
[13] Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 2002;99:5261—6.
[14] Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res 2001;11:1095—9.
[15] Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A 2003;100:8817—22.
[16] Drmanac R, Labat I, Brukner I, Crkvenjakov R. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 1989;4:114—28.
[17] Drmanac S, Kita D, Labat I, Hauser B, Schmidt C, Burczak JD, et al. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. Nat Biotechnol 1998;16:54—8.
[18] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496—512.
[19] Fox GE, Wisotzkey JD, Jurtshuk Jr P. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 1992;42:166—70.
[20] Fredlake CP, Hert DG, Mardis ER, Barron AE. What is the future of electrophoresis in large-scale genomic sequencing? Electrophoresis 2006;27:3689—702.
[21] Fredlake CP, Hert DG, Kan CW, Chiesl TN, Root BE, Forster RE, et al. Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 min. Proc Natl Acad Sci U S A 2008;105:476—81.
[22] Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 1992;89:1827—31.
[23] Gadkar V, Rillig MC. Suitability of genomic DNA synthesized by strand displacement amplification (SDA) for AFLP analysis: genotyping single spores of arbuscular mycorrhizal (AM) fungi. J Microbiol Methods 2005;63:157—64.
[24] Gharizadeh B, Nordström T, Ahmadian A, Ronaghi M, Nyrén P. Long-read pyrosequencing using pure 2′-deoxyadenosine-5′-O′-(1-thiotriphosphate) Sp-isomer. Anal Biochem 2002;301: 82—90.
[25] Gilbert W, Maxam A. The nucleotide sequence of the lac operator. Proc Natl Acad Sci U S A 1973;70:3581—4.
[26] Gresham D, Dunham MJ, Botstein D. Comparing whole genomes using DNA microarrays. Nat Rev 2008;9:291—302.
[27] Guo BC. Mass spectrometry in DNA analysis. Anal Chem 1999;71:333—7.
[28] Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441—7.
[29] Hacia JG, Makalowski W, Edgemon K, Erdos MR, Robbins CM, Fodor SP, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18: 155—8.
[30] Hall N. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 2007;210:1518—25.
[31] Heller LC, Jones M, Widen RH. Comparison of DNA pyrosequencing with alternative methods for identification of mycobacteria. J Clin Microbiol 2008;46:2092—4.
[32] Holmberg K. Organic reactions in microemulsions. Curr Opin Colloid Interface Sci 2003;8:187—96.
[33] Hori M, Fukano H, Suzuki Y. Uniform amplification of multiple DNAs by emulsion PCR. Biochem Biophys Res Commun 2007;352:323—8.
[34] Hudson KL, Holohan MK, Collins FS. Keeping pace with the times — The genetic information non-discrimination Act of 2008. New Engl J Med 2008;358:2661—3.
[35] Janda JM, Abott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils and pitfalls. J Clin Microbiol 2007;45:2761—4.
[36] Kaiser J. A plan to capture human diversity in 1000 genomes. Science 2008;319:395.
[37] Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A 1996;93: 13770—3.
[38] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56—64.
[39] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860—921.
[40] Lespinet V, Terraz F, Recher C, Campo E, Hall J, Delsol G, et al. Single-cell analysis of loss of heterozygosity at the ATM gene locus in Hodgkin and Reed-Sternberg cells of Hodgkin's lymphoma: ATM loss of heterozygosity is a rare event. Int J Cancer 2005;114:909—16.
[41] Lévêque M, Marlin S, Jonard L, Procaccio V, Reynier P, Amati-Bonneau P, et al. Whole mitochondrial genome screening in maternally inherited non-syndromic hearing impairment using a microarray resequencing mitochondrial DNA chip. Eur J Hum Genet 2007;15:1145—55.
[42] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol 2007;5:e254.
[43] Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet 1998;19: 225—32.
[44] Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvänen AC. Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. Nucleic Acids Res 2003;31: e129.
[45] Lovmar L, Syvänen AC. Multiple displacement amplification to create long lasting source of DNA for genetic studies. Hum Mutat 2006;27:603—14.
[46] Lysov IuP, Florent'ev VL, Khorlin AA, Khrapko KR, Shik VV. Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method. Dokl Akad Nauk SSSR 1988;303:1508—11.
[47] Mai M, Hoyer JD, McClure RF. Use of multiple displacement amplification to amplify genomic DNA before sequencing of the alpha and beta haemoglobin genes. J Clin Pathol 2004;57:637—40.
[48] Maitra A, Cohen Y, Gillespie SE, Mambo E, Fukushima N, Hoque MO, et al. The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. Genome Res 2004;14:812—9.
[49] Maizels NM. The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of *Escherichia coli*. Proc Natl Acad Sci U S A 1973;70:3585—9.
[50] Marziali A, Akeson M. New DNA sequencing methods. Annu Rev Biomed Eng 2001;3:195—223.

[51] Mashayekhi F, Ronaghi M. Analysis of read length limiting factors in Pyrosequencing chemistry. Anal Biochem 2007;363:275—87.
[52] Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A 1977;74:560—4.
[53] Mitra RD, Church GM. In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res 1999;15(27):e34.
[54] Mitra RD, Shendure J, Olejnik J, Edyta-Krzymanska-Olejnik, Church GM. Fluorescent in situ sequencing on polymerase colonies. Anal Biochem 2003;320:55—65.
[55] Murthy KK, Mahboubi VS, Santiago A, Barragan MT, Knöll R, Schultheiss HP, O'Connor DT, et al. Assessment of multiple displacement amplification for polymorphism discovery and haplotype determination at a highly polymorphic locus, MC1R. Hum Mutat 2005;26:145—52.
[56] Nakano M, Nakai N, Kurita H, Komatsu J, Takashima K, Katsura S, et al. Single-molecule reverse transcription polymerase chain reaction using water-in-oil emulsion. J Biosci Bioeng 2005;99:293—5.
[57] Oetting WS. Large scale DNA sequencing: new challenges emerge — The 2007 human genome variation society scientific meeting. Hum Mutat 2008;29:765—8.
[58] Pas SD, Tran N, de Man RA, Burghoorn-Maas C, Vernet G, Niesters HG. Comparison of reverse hybridization, microarray, and sequence analysis for genotyping hepatitis B virus. J Clin Microbiol 2008;46:1268—73.
[59] Petti CA. Detection and idendification of microorganisms by gene amplification and sequencing. CID 2007;44: 1108—14.
[60] Pihlak A, Baurén G, Hersoug E, Lönnerberg P, Metsis A, Linnarsson S. Rapid genome sequencing with short universal tiling probes. Nat Biotechnol 2008;26:676—84.
[61] Raghunathan A, Ferguson Jr HR, Bornarth CJ, Song W, Driscoll M, Lasken RS. Genomic DNA amplification from a single bacterium. Appl Environ Microbiol 2005;71:3342—7.
[62] Rhee M, Burns MA. Nanopore sequencing technology: research trends and applications. Trends Biotechnol 2006;24: 580—6.
[63] Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science 1998;281:363—5.
[64] Ryan D, Rahimi M, Lund J, Mehta R, Parviz BA. Toward nanoscale genome sequencing. Trends Biotechnol 2007;25:385—9.
[65] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977;265:687—95.
[66] Sanger F, Nicklen S. Coulson AR DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74:5463—7.
[67] Sequencing Consortium Genome sequence of the nematode C. elegans *Caernahbaditis elegans*: a platform for investigating biology. Science 1998; 282:2012—18.
[68] Service RF. The race for the 1000$ genome. Science 2006;311:1544—6.
[69] Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005;309:1728—32.
[70] Smith LM. Sequence from spectrometry: a realistic prospect? Nat Biotechnol 1996;14:1084—7.
[71] Southern EM. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. Trends Genet 1996;12:110—5.
[72] Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 1994;44:846—9.
[73] Strezoska Z, Paunesku T, Radosavljević D, Labat I, Drmanac R, Crkvenjakov R. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. Proc Natl Acad Sci USA 1991;15(88):10089—93.
[74] Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 1992;13:718—25.
[75] Tost J, Gut IG. DNA methylation analysis by pyrosequencing. Nat Protoc 2007;2:2265—75.
[76] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304—51.
[77] Wen WH, Bernstein L, Lescallett J, Beazer-Barclay Y, Sullivan-Halley J, White M, et al. Comparison of TP53 mutations identified by oligonucleotide microarray and conventional DNA sequence analysis. Cancer Res 2000;60:2716—22.
[78] Wong CW, Albert TJ, Vega VB, Norton JE, Cutler DJ, Richmond TA, et al. Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. Genome Res 2004;14:398—405.
[79] Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A 1992;89:5847—51.
[80] Ziebolz B, Droege M. Toward a new era in sequencing. Biotechnol Annu Rev 2007;13:1—26.

Pour en savoir plus

NCBI : www.ncbi.nlm.nih.gov/
Base de données de SNP : www.ncbi.nlm.nih.gov/SNP/
Omim : www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM