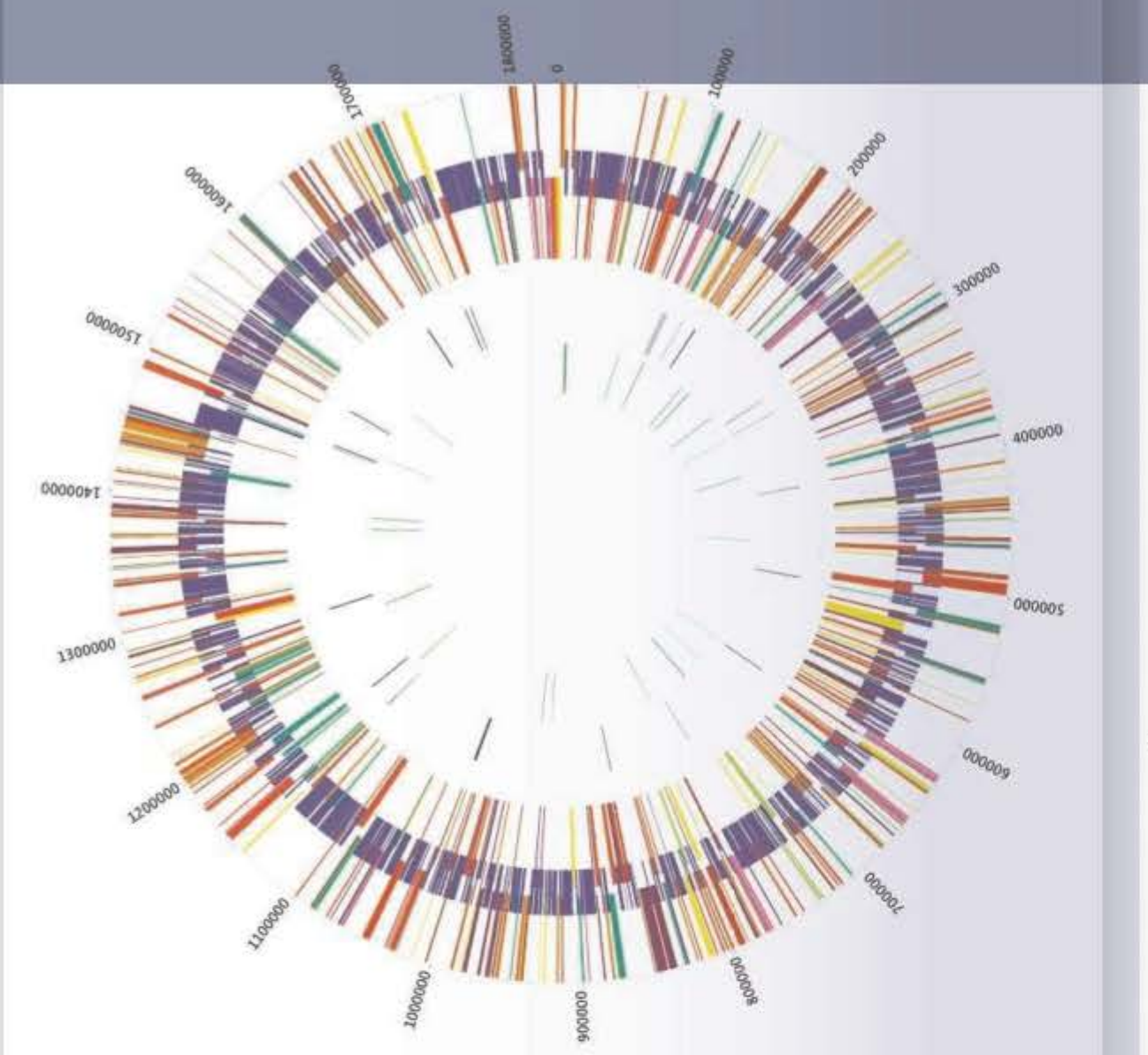


CHAPITRE **18**

La génomique

Aperçu du chapitre

18.1 Les cartes du génome

18.2 Le séquençage des génomes

18.3 Programmes génomiques

18.4 Annotation des génomes et banques de données

18.5 La génomique comparative et fonctionnelle

18.6 Applications de la génomique

Introduction

Au cours des 30 dernières années, notre connaissance de la biologie a progressé de façon extraordinaire. Après l'isolement des premiers gènes au milieu des années 1970, les chercheurs ont obtenu la première séquence complète d'un génome, celle de la bactérie *Haemophilus influenzae* (représentée ci-dessus, avec la même couleur pour la même fonction). On a terminé une séquence grossière complète du génome humain au détour du 21e siècle. Autrement dit, la science est passée du clonage d'un gène à la détermination se la séquence d'un million des paires de bases après 20 ans et d'une séquence de 3 milliards de paires de bases après cinq années de plus, pour pouvoir actuellement séquencer 20 milliards de paires de bases en quelques heures pour moins de 1000 $. Au chapitre 17, nous avons envisagé la biotechnologie et, dans ce chapitre, nous allons voir comment certaines de ces technologies ont été appliquées à l'analyse des génomes complets. La génomique implique la participation de la génétique classique et moléculaire, ainsi que de la biotechnologie, pour l'étude de la structure, du fonctionnement et des relations des génomes.

18.1 Les cartes des génomes

Objectifs

1. *Distinguer les cartes génétiques et physiques.*
2. *Décrire les avantages et les inconvénients des cartes de restriction, chromosomiques et obtenues par STS.*
3. *Montrer les relations entre une carte génétique et une carte physique.*

Nous utilisons les cartes pour retrouver certains lieux et, selon la nature de ces lieux, nous pouvons nous servir de différents types de cartes. En génomique, nous pouvons localiser un gène sur un chromosome, dans une région du chromosome et, enfin, trouver sa localisation la plus précise dans la séquence d'ADN du chromosome. Pour obtenir une carte au niveau de la séquence d'ADN, il faut connaître toute la séquence du génome, chose qui fut autrefois hors de portée de la technologie. Connaître la séquence complète est toutefois inutile en l'absence d'autres informations, par exemple si l'on ne sait pas quelle partie du génome intervient dans tels phénotypes. On trouve ces informations dans les cartes génétiques. En comparaison des cartes géographiques,

retrouver un gène dans la séquence du génome humain, c'est un peu comme essayer de retrouver une maison sur une carte du monde.

Plusieurs méthodes permettent d'analyser des génomes entiers

Nous construisons des sortes différentes de cartes de génomes à partir de types différents de données et pour des types différents d'analyses. Il existe fondamentalement deux sortes de cartes : des cartes génétiques et les cartes physiques. Toutes deux sont basées sur des **marqueurs génétiques**, qui peuvent être des différences décelables entre les individus. Les **cartes génétiques** sont sommaires, elles montrent la position relative des marqueurs génétiques, basée sur la fréquence de recombinaison (voir chapitre 13). Les **cartes physiques** localisent avec précision les marqueurs génétiques dans le génome, et la carte physique définitive est la séquence complète du génome. Enfin, la génomique fait la synthèse des différents types de cartes du génome et permet une analyse à grande échelle.

Les cartes génétiques donnent les distances relatives entre les marqueurs génétiques

Les cartes génétiques, ou cartes de liaison, sont basées sur le mécanisme de recombinaison à la méiose, qui modifie la façon dont les allèles sont liés sur les chromosomes (voir chapitre 13). Ces « marqueurs génétiques » peuvent être des gènes, détectés par des différences phénotypiques, ou des différences dans les séquences d'ADN décelées par PCR ou digestion par les endonucléases de restriction, comme on l'a vu au chapitre 17. Pour les cartes génétiques classiques, les croisements contrôlés permettent de connaître le nombre de recombinants pour des paires particulières de marqueurs. Ce n'est évidemment pas faisable chez les humains ; la fréquence des recombinaisons est donc estimée à partir des données préexistantes concernant les relations familiales et les pedigrees, ainsi que des analyses statistiques.

Sur une carte génétique, les distances sont exprimées en centimorgans (cM), un cM correspondant à une fréquence de recombinaison de 1 % entre deux locus. Deux locus distants de 1 cM sont relativement proches, puisqu'ils n'ont qu'une chance sur cent d'être séparés à la méiose. Les premières cartes génétiques humaines reprenaient les gènes pour la susceptibilité aux maladies, localisés sur les différents chromosomes, mais ce n'étaient pas des cartes générales. Cette situation n'a évolué qu'avec l'arrivée des marqueurs moléculaires, qui ne correspondent par à des phénotypes observables (voir chapitre 13). Non seulement c'était une source de nombreux marqueurs polymorphes, mais on pouvait en outre les intégrer facilement aux cartes moléculaires. La carte génétique humaine est aujourd'hui très dense, avec un marqueur par cM environ.

Les cartes génétiques sont très importantes parce qu'elles situent les caractères génétiques sur le chromosome, mais elles souffrent aussi de restrictions. En premier lieu, la répartition des crossing-over, et des recombinaisons, n'est pas aléatoire. De nombreuses observations montrent en fait actuellement l'existence de points chauds pour la recombinaison chez les humains, les souris et beaucoup d'autres organismes étudiés. De plus, la fréquence des recombinaisons peut varier parmi les individus et dépendre de la structure de la chromatine. Tout cela signifie que la relation entre la distance génétique (en centimorgans) et la distance effective (en paires de bases, ou pb) varie dans le génome. Mais il reste possible de mettre en parallèle une séquence génomique complète et une carte génétique dense pour constater la complémentarité des représentations du génome.

Les cartes physiques donnent les distances exactes entre marqueurs génétiques

Contrairement à une carte génétique, qui donne la position relative d'un marqueur dans le génome, une carte physique donne sa position exacte. Les plus anciennes cartes physiques dépendaient des premiers outils de la biologie moléculaire, les enzymes de restriction. La sophistication des cartes physiques a suivi celle des technologies.

La forme ultime des cartes physiques est la localisation d'un grand nombre de marqueurs génétiques sur la séquence d'ADN complète du génome. Sur une carte physique, les distances entre marqueurs est donnée en paires de bases (1000 paires de bases, ou pb correspondent 1 kilobase, ou kb). On peut obtenir une carte physique d'uns segment d'ADN sans connaître la séquence et sans savoir si l'ADN code des gènes. En fait, beaucoup de programmes de séquençage débutent par la création de cartes physiques. Il existe trois grands types de cartes physiques : (1) les cartes de restriction, établies à l'aide des endonucléases de restriction, (2) les cartes chromosomiques (basées sur les techniques décrites au chapitre 17) et les cartes des sites servant de marqueurs (STS)

Les cartes de restriction

Les cartes de restriction ne conviennent généralement pas pour des molécules d'ADN dépassant 50 kb ; elles ne sont donc utiles que pour les génomes de certains organites et virus. Les premières cartes physiques ont été obtenues en coupant l'ADN par différentes enzymes de restriction, seules ou en combinaison (figure 18.1). On obtient une carte en analysant la répartition des fragments obtenus.

On peut appliquer cette technique à des génomes plus importants si l'ADN est d'abord découpé en fragments plus petits. Pour les plus grands fragments d'ADN, ce processus est répété, puis utilisé pour réunir ensuite les fragments en un segment continu, ou **contig**, en se basant sur leur taille et leurs chevauchements.

Les cartes chromosomiques

Les cytologistes qui étudiaient les chromosomes au microscope optique ont constaté qu'en utilisant des colorants différents, ils pouvaient obtenir des répartitions reproductibles de bandes sur les chromosomes. Ils pouvaient ainsi identifier tous les chromosomes et les diviser en régions en fonction de la répartition de ces bandes. L'utilisation de colorants différents permet de construire des cartes cytologiques de l'ensemble du génome. Sur ces cartes, chaque chromosome est composé de deux bras divisés eux-mêmes en plusieurs régions et sous-régions. Toutes les cartes ultérieures ont été basées sur ces cartes peu précises. Ces cartes physiques à grande échelle rappellent une carte d'un pays, du fait qu'elles renferment tout le génome, mais à faible résolution.

Les cartes cytologiques permettent d'identifier les anomalies chromosomiques liées à des maladies humaines, comme la leucémie myéloïde chronique. Dans cette maladie, une translocation réciproque s'est produite entre un chromosome 9 et un chromosome 22 (figure 18.2*a*) ; elle est responsable d'une forme modifiée de tyrosine kinase constamment active, entraînant la prolifération des leucocytes et, par conséquent, la leucémie.

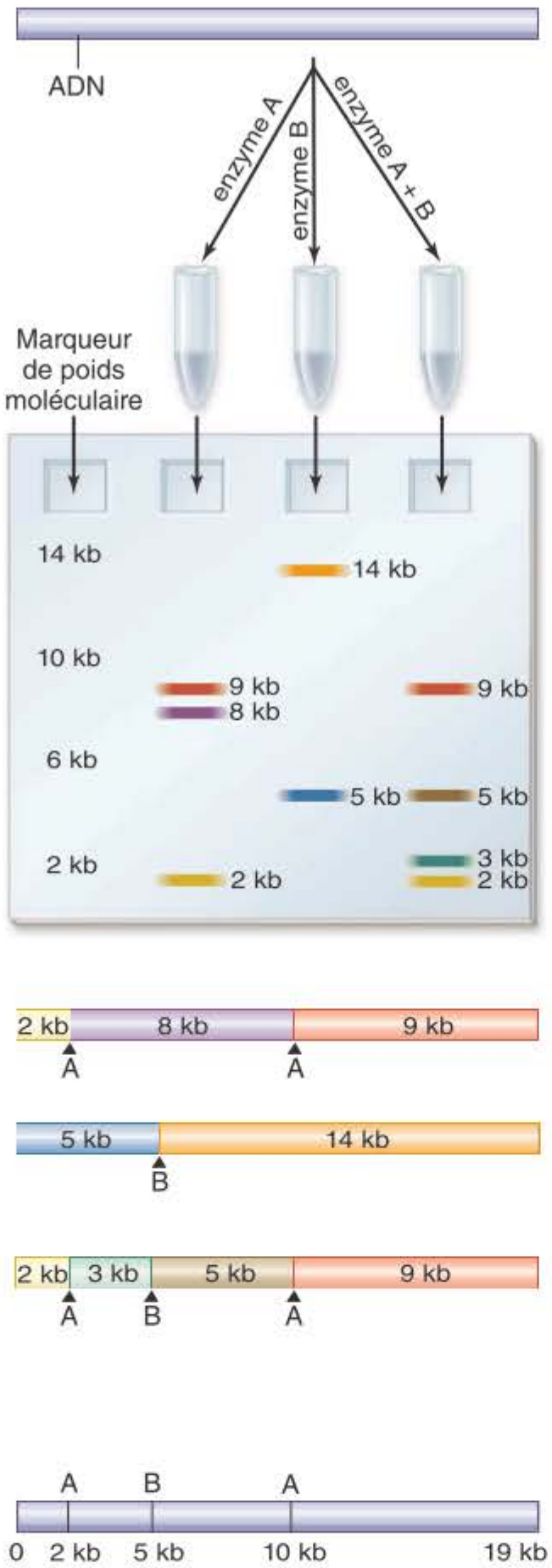
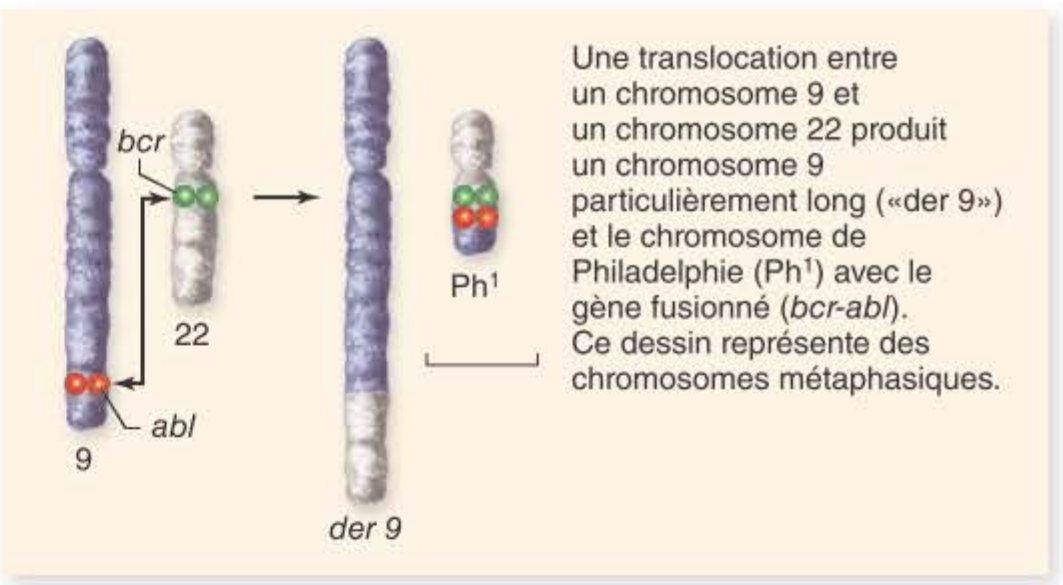
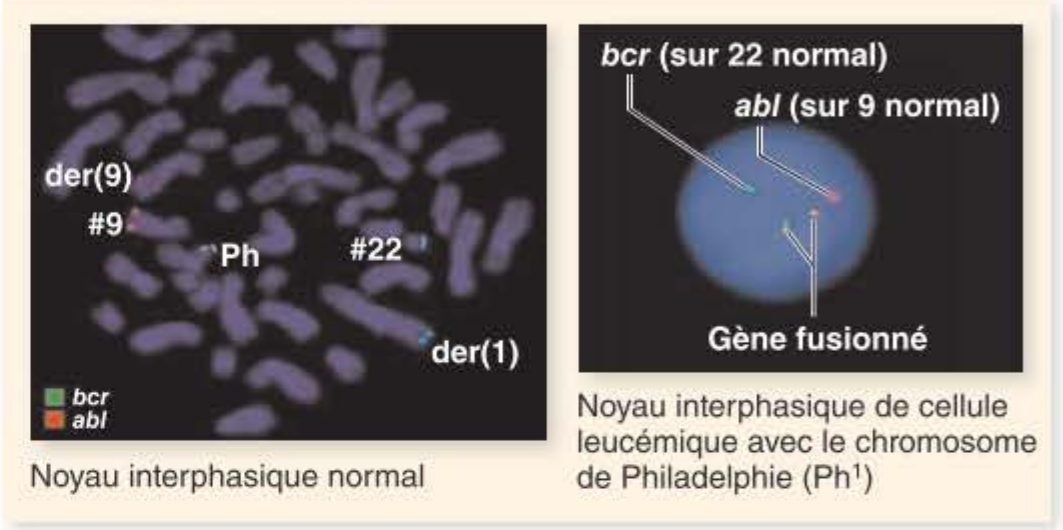


Figure 18.1 On peut utiliser les enzymes de restriction pour construire une carte physique. L'ADN est digéré par deux enzymes de restriction différentes séparément et simultanément. Les fragments sont séparés par électrophorèse en fonction de leur taille. On peut localiser les sites en comparant la taille des fragments provenant des réactions individuelles et de la réaction combinée.



a.



b.

Figure 18.2 Application de l'hybridation avec fluorescence in situ (FISH) pour faire le lien entre l'ADN cloné et les cartes cytologiques. ***a.*** Chromosomes d'un caryotype humain montrant la translocation entre les chromosomes 9 et 22. ***b.*** FISH avec une sonde *bcr* (*en vert*) et *abl* (*en rouge*). La coloration jaune signale les gènes fusionnés (combinaison des fluorescences verte et rouge). Le gène *abl* et le gène fusionné *bcr-abl* codent tous deux une tyrosine kinase, mais le gène fusionné s'exprime toujours.

Question Comment pouvez-vous expliquer la répartition des taches rouges, vertes et jaunes sur la partie gauche de la figure 18.2*b*.

Pour l'obtention des cartes chromosomiques, on peut aussi utiliser l'hybridation in situ avec fluorescence, dont il a été question au chapitre 17 (figure 18.2*b*). À l'origine, le niveau de résolution de la FISH était d'environ 1 Mb (mégabase) ; on ne pouvait donc distinguer deux points d'un chromosome séparés par moins de 1 Mb. À cause de cette restriction, FISH n'était réellement utile que pour voir si une séquence d'ADN particulière se trouvait sur un chromosome particulier. En 1990, les progrès techniques ont permis l'utilisation des chromosomes en interphase aussi bien qu'en métaphase, et la résolution de FISH a été ainsi portée à 25 kb. Ce progrès permettait donc d'appliquer la technique pour des cartes plus fines.

Cartes des sites servant de marqueurs

La technique de restriction est un moyen relativement rapide et aisé de construire des cartes précises à partir de quantités assez faibles d'ADN. D'autre part, on obtient des cartes de grandes molécules d'ADN à faible résolution par la méthode FISH, techniquement délicate. Les sites servant de marqueurs (STS) constituent une alternative intéressante, combinant les avantages de la restriction et de FISH. Cette méthode permet de construire rapidement des cartes physiques détaillées de grandes molécules d'ADN avec moins de problèmes techniques.

Un STS (*sequence-tagged site*) est un court segment monocaténaire d'ADN génomique qui peut être amplifié par PCR (voir chapitre 17). L'objectif est de voir si deux STS sont sur la même molécule d'ADN. Cela implique la fragmentation aléatoire de l'ADN en morceaux qui se chevauchent. Si deux marqueurs STS se trouvent régulièrement ensemble sur des fragments d'ADN, c'est qu'ils sont relativement proches. S'ils ne sont pas régulièrement ensemble sur les fragments d'ADN, c'est vraisemblablement qu'ils sont plus éloignés (figure 18.3).

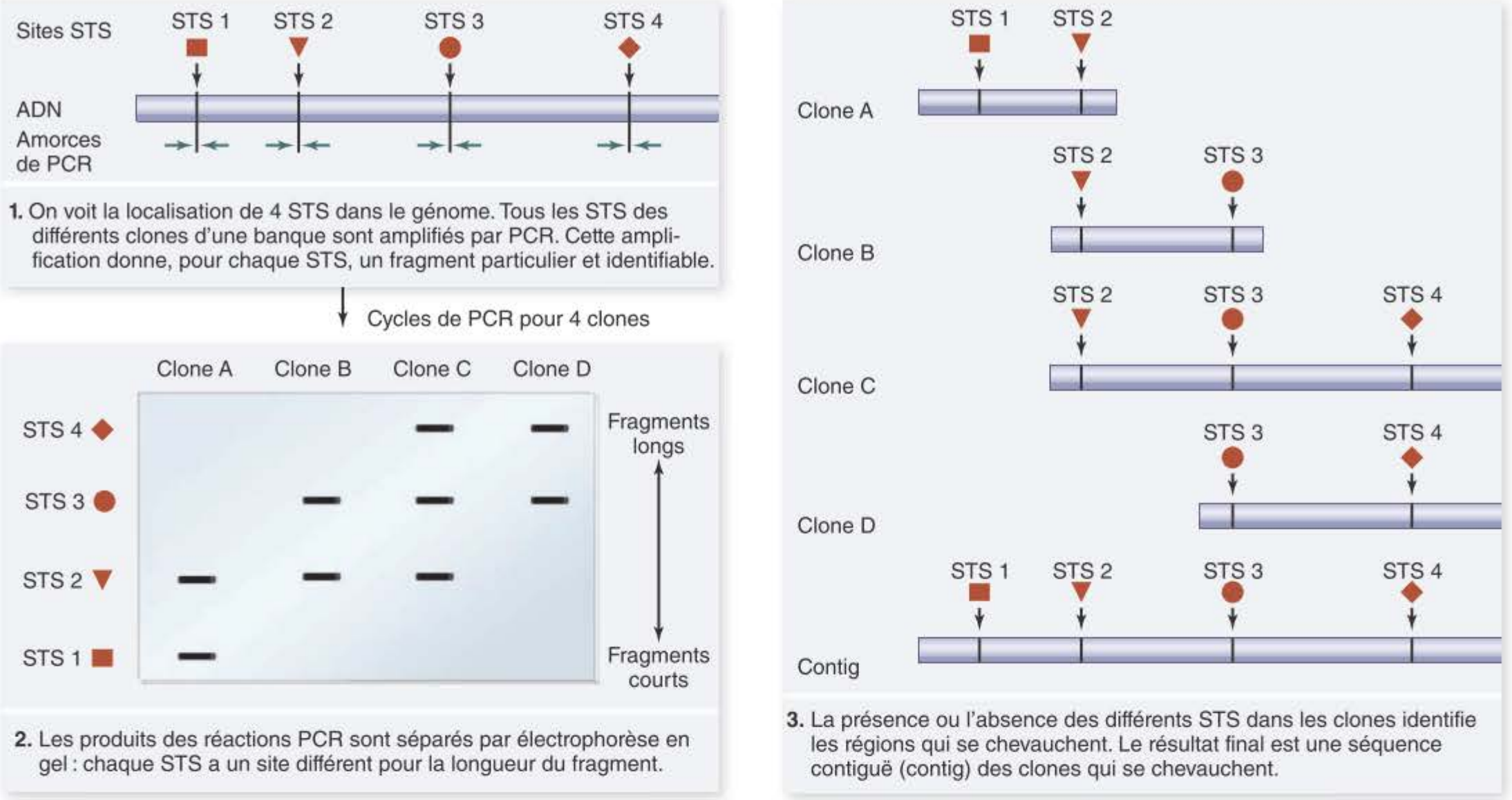


Figure 18.3 Construction d'une carte physique de sites à séquences marquées. La présence de repères, les STS, dans le génome humain permet d'entamer la construction d'une carte physique à une échelle suffisante pour servir de base au séquençage de l'ensemble du génome. *(1)* Des amorces (*flèches vertes*) qui reconnaissent des STS individuels sont ajoutées à un segment d'ADN cloné, puis l'ADN est amplifié par la réaction en chaîne de la polymérase (PCR). *(2)* Les produits de la PCR sont séparés en fonction de leur taille et les STS présents dans les différents clones sont identifiés. *(3)* Les segments d'ADN clonés sont alignés en fonction des STS pour construire un contig.

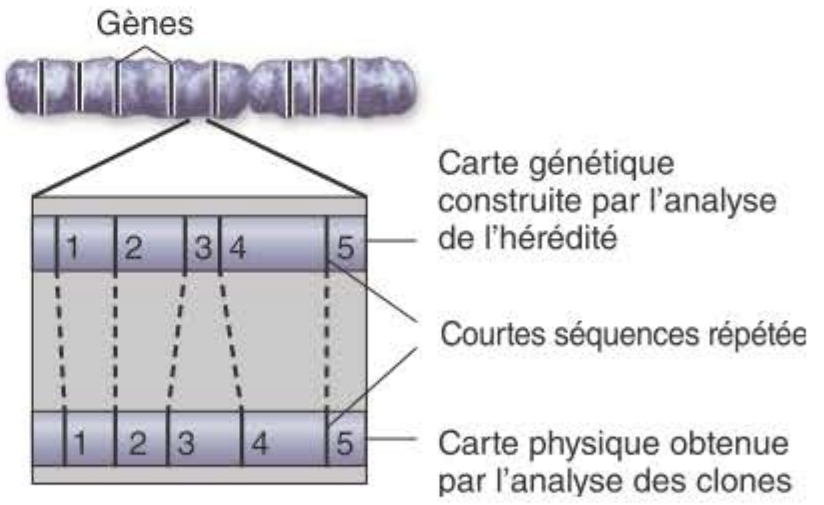
Analyse des données Vous pourriez construire la carte physique en vous servant d'une partie des clones du tableau 2. Citez les combinaisons possibles vous permettant de construire cette carte avec le plus petit nombre de clones possible.

De même que les cartes génétiques sont basées sur la fréquence à laquelle les marqueurs sont séparés par recombinaison, l'utilisation des STS repose sur la fréquence de séparation des marqueurs par fragmentation de l'ADN. Les marqueurs les plus proches seront séparés moins souvent que ceux qui sont éloignés. Une cassure est d'autant plus probable entre deux marqueurs qu'ils sont plus éloignés.

On peut réunir les fragments d'ADN grâce aux STS en recherchant les régions qui se chevauchent. En raison de la densité importante des STS dans le génome humain et de leur identification aisée dans les clones d'ADN, les chercheurs ont pu construire des cartes physiques à grande échelle de l'énorme génome de 3,2 milliards de paires de bases au milieu des années 1990 (figure 18.3). Les STS représentent un échafaudage permettant d'assembler les séquences génomiques.

Il est possible de mettre en relation les cartes physiques et les cartes génétiques

Les cartes génétiques permettent d'identifier les régions du génome qui interviennent dans des phénotypes particuliers, mais elles ne nous disent rien sur la nature de ces « gènes ». La corrélation entre les cartes physiques et génétiques nous permet de trouver la séquence des gènes localisés génétiquement.



Le problème, pour retrouver des gènes est la résolution des cartes génétiques, qui n'est actuellement guère aussi fine que pour la séquence du génome. Dans le génome humain, des marqueurs distants d'un cM peuvent se trouver à un million de paires de bases l'un de l'autre.

Comme les marqueurs utilisés pour les cartées génétiques sont actuellement surtout des marqueurs moléculaires, on peut les localiser facilement dans une séquence génomique. Tout gène cloné peut aussi être inséré dans la séquence génomique et sur la carte génétique. Une corrélation automatique est ainsi établie entre les deux cartes. En théorie, on peut facilement localiser, dans la séquence d'ADN du génome, les gènes localisés sur une carte génétique. En pratique, c'est parfois plus

problématique parce que la relation entre distance génétique et distance effective diffère au sein du génome. Un cM sur la carte génétique peut en fait correspondre à des nombres différents de paires de bases dans des régions différentes du génome.

Questions d'apprentissage 18.1

Il existe des cartes physiques et génétiques des génomes. Parmi les cartes physiques, on a des cartes cytogénétiques des bandes chromosomiques et des cartes de restriction. Les cartes génétiques sont corrélées aux cartes physiques grâce à des marqueurs d'ADN comme les sites servant de marqueurs (STS) caractéristiques de chaque génome. Alors que les cartes génétiques situent des marqueurs, comme les gènes, les uns par rapport aux autres, les cartes physiques les localisent à des endroits spécifiques sur une séquence d'ADN. Les cartes génétiques et physiques sont utilisées lors du séquençage dans les programmes génomiques.

- *Comment deux marqueurs peuvent-ils se trouver à 10 cM l'un de l'autre sur une carte génétique, alors que deux marqueurs situés ailleurs sur la même carte sont distants de 12 cM, bien que ces deux paires de gènes soient séparés par la même quantité d'ADN ?*

18.2 Le séquençage des génomes

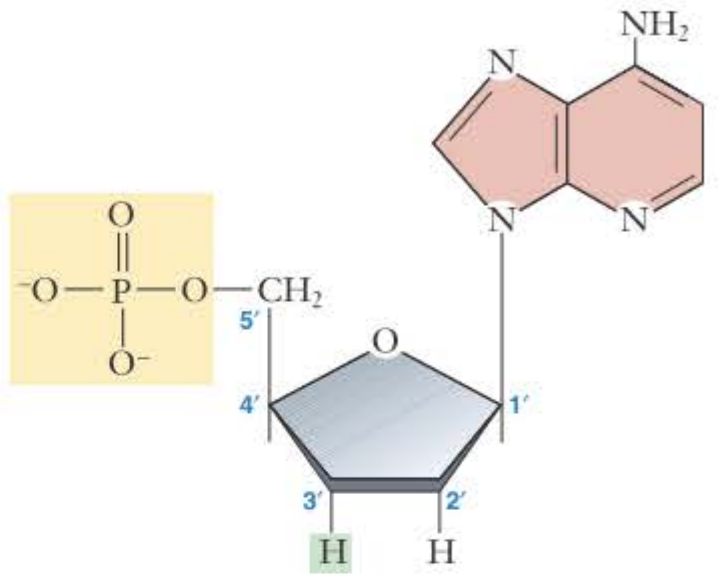
Objectifs

1. *Montrer les différences entre le séquençage avec terminateur didésoxy et les technologies de séquençage de dernière génération.*
2. *Comparer les méthodes clone-contig et shotgun de séquençage et assemblage du génome.*

La carte physique ultime est la séquence des paires de bases de tout un génome. Grâce au séquençage de l'ADN rapide et à grande échelle, le séquençage de génomes entiers est à la portée des laboratoires de petite taille et des sociétés biotechnologiques. À la base, toutes les méthodes de séquençage adaptent la réplication de l'ADN à des conditions in vitro. Les premières méthodes utilisaient des nucléotides modifiés pour arrêter la réplication et limiter la réplication à un seul brin de l'ADN. Les progrès rapides de la génomique ont été à l'origine de nouvelles techniques de séquençage automatisées beaucoup plus rapides.

Le séquençage du génome avec terminateur didésoxy reste la méthode la plus importante

Des progrès spectaculaires ont été apportés aux techniques de séquençage de l'ADN au cours des dix dernières années. Dans certains cas cependant, les anciennes méthodes restent utiles. Une de ces méthodes, mise au point par Fred Sanger dans les années 1970, rappelle la réplication de l'ADN. Cette technique repose sur l'utilisation de **didésoxynucléotides** dans les réactions de séquençage de l'ADN. Les didésoxynucléotides fonctionnent comme terminateurs de chaînes et arrêtent la réplication quand ils sont incorporés. Aucun nucléotide de l'ADN ne possède de groupement hydroxyle sur le carbone 2' du sucre, tandis que les didésoxynucléotides sont dépourvus d'un groupement OH 3'. Au chapitre 14, on a vu que les bases s'ajoutent à l'extrémité en croissance 3' d'un brin d'ADN par une réaction entre le groupement OH du nucléotide précédent et le triphosphate du suivant (voir figure 14.12). Un didésoxynucléotide peut être incorporé, mais ce sera l'extrémité de la chaîne en croissance, parce qu'il ne possède pas OH.



À l'origine, le séquençage avec terminateur didésoxy était manuel et il fallait effectuer quatre réactions distinctes, chacune avec un même didésoxynucléotide et les quatre désoxynucléotides. Une fois incorporé, le didésoxynucléotide bloque la synthèse au niveau d'une base spécifique. Chaque réaction donne donc un lot de fragments se terminant chacun par une base particulière, et les fragments provenant des quatre réactions se terminent par l'une des quatre bases possibles. Une électrophorèse sur gel à haute définition sépare les fragments dont la longueur diffère d'une seule base. On obtient ainsi une échelle de fragments permettant de lire la séquence du bas (le fragment le plus court) au sommet (les plus longs fragments).

Pour l'automatisation, chaque didésoxynucléotide est marqué par un colorant fluorescent et les quatre sont utilisés dans la même réaction de synthèse d'ADN. Comme précédemment, l'incorporation aléatoire des didésoxynucléotides donne un lot de fragments se terminant par des bases spécifiques mais, dans ce cas, chaque base est représentée par une couleur. Les fragments marqués sont séparés par électrophorèse dans un tube capillaire et chaque molécule passe par un laser qui provoque la fluorescence. Le nucléotide terminant le brin d'ADN est identifié par la couleur de la fluorescence et automatiquement signalé à un ordinateur (figure 18.4). Cette méthode reste encore laborieuse, mais l'automatisation permet d'effectuer en parallèle des milliers de réactions de séquençage. Il est ainsi possible de séquencer près de 900 kb d'ADN par jour. D'une certaine façon, ce fut la première étape dans l'élaboration des plateformes de séquençage parallèle massif caractéristiques des techniques de séquençage de dernière génération.

Le séquençage de dernière génération fait massivement appel à des technologies parallèles pour accroître la vitesse

Les techniques de séquençage de dernière génération ont fait leur apparition il y a une dizaine d'années. Depuis lors, des améliorations très importantes ont été apportées quant à la vitesse du séquençage, le coût par base séquencée et la longueur de la séquence lue à chaque étape de séquençage. Grâce à cela, le nombre de génomes séquencés et

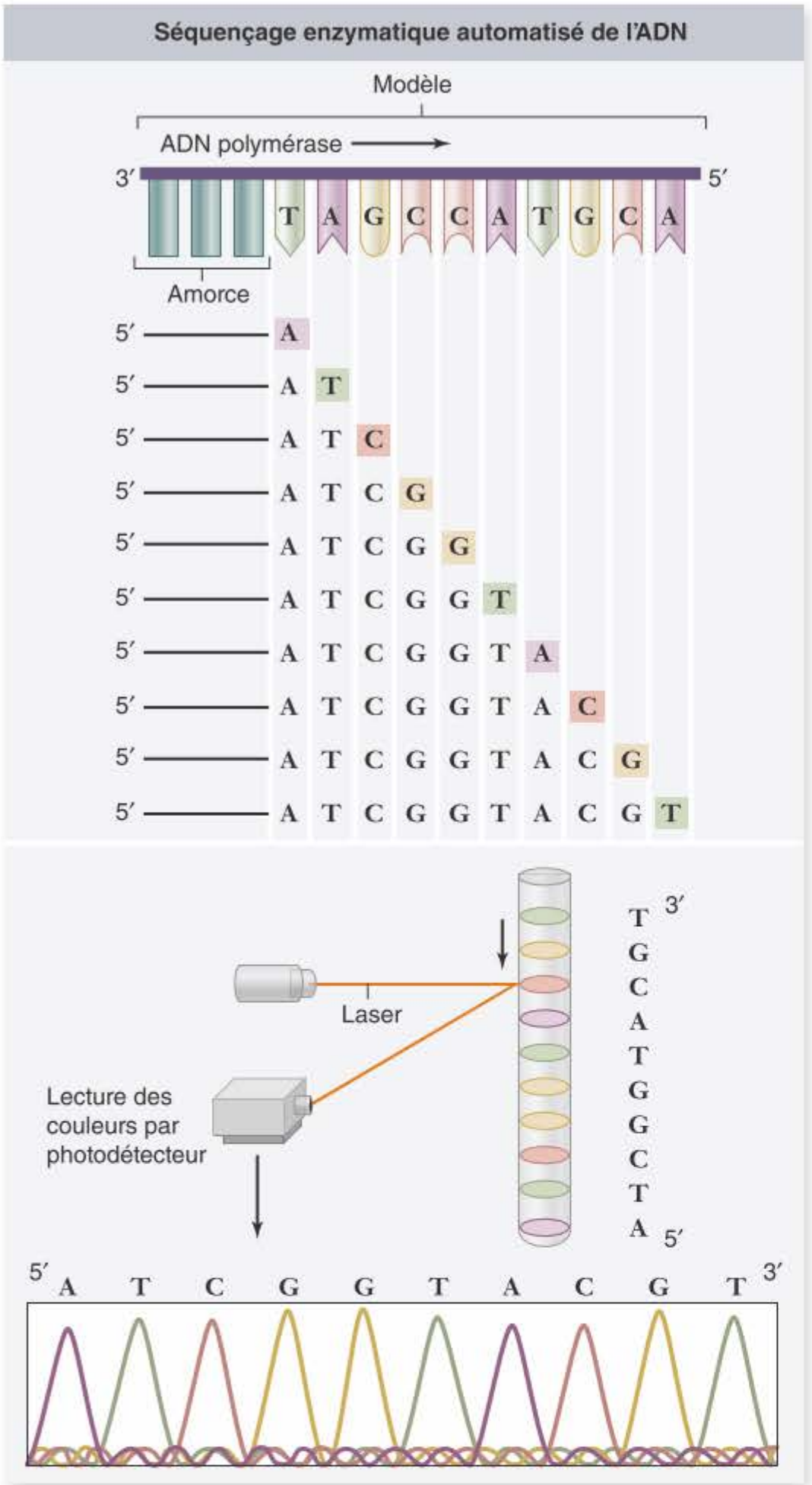


Figure 18.4 Séquençage didésoxy automatisé de l'ADN. La séquence à étudier est représentée au-dessus sous la forme d'un brin matrice et d'une amorce destinée à l'ADN polymérase. Dans le séquençage automatisé, chaque ddNTP est marqué par un colorant fluorescent différent, ce qui permet d'effectuer la réaction dans un même tube. Les fragments provenant des réactions sont représentés. Si l'électrophorèse est réalisée dans un tube capillaire, un laser localisé dans le bas du tube excite les colorants et chacun émet une couleur différente détectée par un photorécepteur.

reséquencés a fortement augmenté au cours des 10 dernières années. À l'avenir, cette masse de données pourrait se traduire par des découvertes concernant de nouvelles relations évolutives et améliorer la personnalisation des soins de santé et le diagnostic.

Beaucoup de nouvelles technologies sont disponibles sur le marché, mais elles ont toutes des traits communs. Premièrement, contrairement au séquençage didésoxy, elles permettent de séquencer l'ADN sans devoir d'abord produire une banque génomique par clonage conventionnel (voir chapitre 17). Deuxièmement, au lieu de milliers de réactions préparées et analysées simultanément, on peut effectuer simultanément des millions de réactions de séquençage ; c'est pourquoi on parle d'une technologie « massivement parallèle ». Troisièmement, on peut se passer de l'électrophorèse qui était nécessaire et laborieuse. Les réactions de séquençage se déroulent désormais simultanément en solution et sont enregistrées directement par l'appareil de séquençage (figure 18.5). Grâce à l'accélération du séquençage et à la réduction des

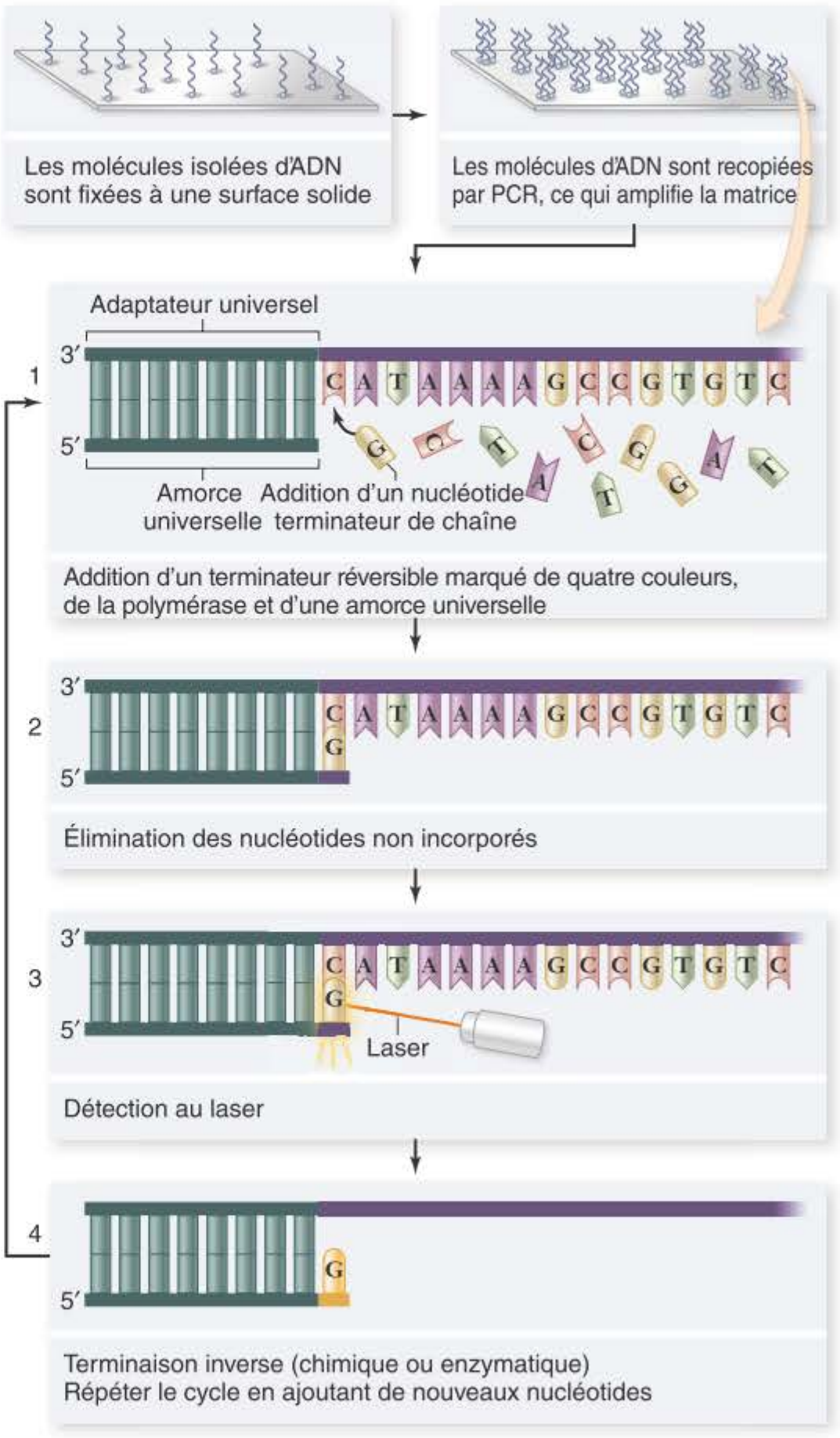


Figure 18.5 Séquençage de dernière génération d'Illumina. Dans le séquençage de dernière génération d'Illumina, les molécules d'ADN matrice sont attachées à une surface solide et amplifiées par PCR pour obtenir des îlots d'ADN. Quatre nucléotides terminateurs marqués par fluorescence sont ensuite ajoutés et, après leur incorporation au nouvel ADN, la synthèse s'arrête. La couleur du nucléotide incorporé est lue, le nucléotide est modifié par voie chimique de façon à pouvoir ajouter un autre nucléotide. Cette addition d'un unique nucléotide est répétée de façon à lire la séquence d'ADN dans chaque îlot.

coûts, il est possible de séquencer certains génomes bactériens en quelques heures seulement.

En dépit de l'accélération du séquençage et de la chute des coûts, un inconvénient de certaines technologies est la longueur de la séquence d'ADN – la *longueur lue* – obtenue à chaque réaction de séquençage. La longueur lue diffère notablement en fonction des technologies et va de 35 bases à un exemple rare de 20 000 bases. En raison de ces différences, les techniques diffèrent en fonction des applications : certaines conviendront pour le séquençage des génomes, tandis que d'autres seront plus appropriées au diagnostic médical. La majorité des technologies habituelles donnent des longueurs de lecture comprises entre 50 et 1000 pb. Si les fragments sont plus courts, l'assemblage des pièces individuelles en génome complet est plus compliqué. En raison de ce problème, on a créé de nouveaux algorithmes pour l'assemblage génomique.

En 2014, on a signalé le séquençage d'un génome humain pour 1000 $. C'est près de 10 000 fois moins que le coût du séquençage d'un génome de la taille de celui de l'homme il y a une dizaine d'années.

Les fragments séquencés sont assemblés en séquences complètes

La tâche la plus complexe, lors d'un séquençage à grande échelle, est l'assemblage des pièces individuelles en un génome complet. Le génome est d'abord coupé en morceaux chevauchants qui sont séquencés, mais il faut ensuite les réassembler en réunissant les courtes régions chevauchantes. Plus courts sont les fragments séquencés et plus il y a d'ADN répété dans le génome ; plus il est difficile aussi de reconstituer le puzzle. Deux stratégies sont possibles ; (1) assembler d'abord des morceaux d'un chromosome, puis voir comment s'adaptent les morceaux plus grands (**assemblage clone-contig**), ou (2) essayer de réunir tous les morceaux à la fois, plutôt que par étapes (**assemblage shotgun**).

L'assemblage shotgun

L'assemblage shotgun est intéressant parce qu'il ne repose sur aucune carte génétique ou physique. Aucune information préalable sur le génome séquencé n'est donc nécessaire. Le premier génome séquencé, celui de la bactérie *Haemophilus influenzae*, a été séquencé, puis assemblé par la méthode shotgun. Ce génome a été coupé en fragments longs d'environ 1,5 à 2 kb qui ont été clonés dans des vecteurs appropriés (voir chapitre 17). Les extrémités des fragments ont été séquencées puis, en se basant sur les chevauchements, on a pu recoller les fragments d'ADN pour obtenir le génome complet de 1,8 Mb (figure 18.6*b*) Cette stratégie aboutit rarement à une séquence génomique parfaitement assemblée. Plusieurs autres moyens sont utilisés pour combler les lacunes qui n'apparaissent qu'après l'assemblage des séquences.

L'assemblage clone-contig

Les méthodes clone-contig (figure 18.6*a*) sont utilisées pour assembler les fragments d'ADN séquencés provenant de génomes plus complexes, comme ceux des eucaryotes. Contrairement à l'assemblage shotgun, elles ont besoin d'une carte physique pour aboutir au réassemblage d'une séquence génomique complète. Le génome est découpé en fragments plus longs – environ 1 à 1,5 Mb. Ces fragments sont ensuite localisés dans un segment plus long d'ADN, appelé **contig**, en utilisant des marqueurs génétiques tels que les STS ou les sites d'endonucléases de restriction. On peut séquencer les fragments d'ADN de 1 à 1,5 Mb (mégabases) et les assembler par la technique shotgun (figure 18.6*b*).

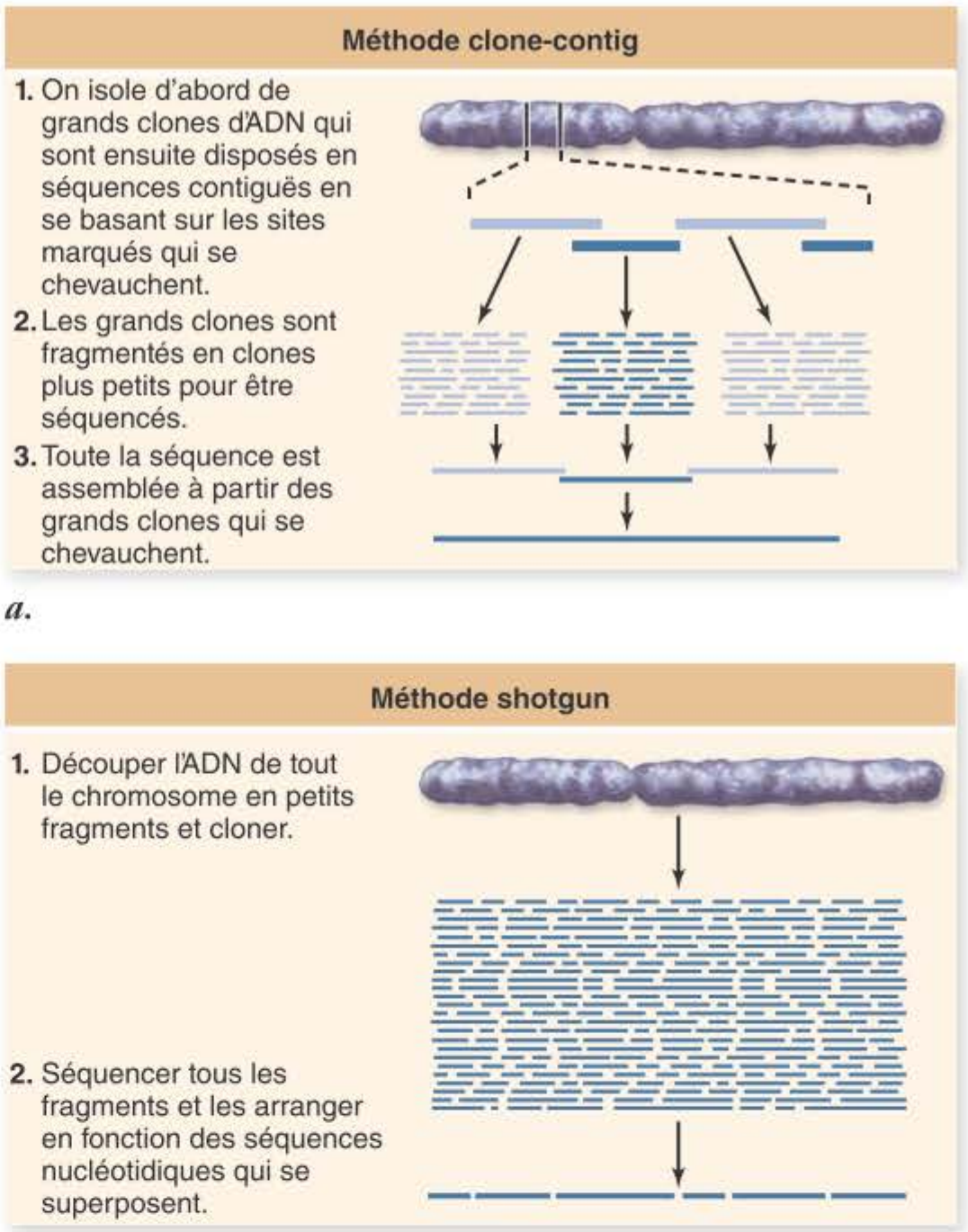


Figure 18.6 Comparaison des méthodes de séquençage. ***a.*** La méthode clone-contig utilise un grand nombre de clones assemblés en régions chevauchantes par les STS. Après leur assemblage, on peut les découper en fragments plus courts pour le séquençage. ***b.*** Dans la méthode shotgun, l'ensemble du génome est découpé en petits clones et séquencé. Des algorithmes informatiques assemblent la séquence finale d'ADN en se basant sur les séquences de nucléotides chevauchantes.

Questions d'apprentissage 18.2

À cause de la taille énorme des génomes, le séquençage a besoin de séquenceurs automatiques traitant en parallèle de nombreux échantillons. Deux méthodes ont été mises au point pour séquencer des génomes entiers : l'un utilise des clones déjà alignés dans des cartes physiques (séquençage et assemblage clone-contig) ; l'autre implique le séquençage aléatoire des clones et l'utilisation d'un ordinateur pour assembler la séquence finale (séquençage et assemblage shotgun). Dans les deux cas, un ordinateur puissant est nécessaire pour l'assemblage de la séquence finale.

- ***Quels sont les avantages et les inconvénients des différentes méthodes de séquençage et d'assemblage de génomes complets ?***

18.3 Les programmes génomiques

Objectifs

1. *Décrire les résultats du Projet Génome Humain.*
2. *Montrer pourquoi le séquençage du génome humain était une entreprise relativement simple en comparaison de celui du génome de blé.*
3. *Montrer les avantages potentiels du projet génome du cancer.*

La technologie de séquençage automatisé a donné une masse de données sur les séquences. Grâce à cela, les chercheurs ont pu étudier des problèmes complexes et dépasser l'analyse des gènes individuels. Les programmes de séquençage ne sont cependant, par eux-mêmes, que des analyses descriptives qui ne nous apprennent rien sur l'organisation des génomes, ils ne s'intéressent pas à la fonction des produits des gènes ni aux interactions pouvant exister entre eux. Les travaux basés sur les résultats des programmes génomiques nous ont apporté des réponses et, en même temps, de nouvelles énigmes.

Le Projet Génome Humain a séquencé et cartographié la plus grande partie du génome humain

Le Projet Génome Humain (PGH) a débuté officiellement en 1990, mais on peut faire remonter ses origines au milieu des années 1980. Le principal objectif était l'obtention de cartes génétiques et physiques du génome humain et il a finalement abouti au séquençage et à l'assemblage effectifs du génome. En 1995, on a obtenu une carte physique couvrant 94 % du génome humain, avec des marqueurs distants en moyenne de 199 kb. Puis, en 1996, on a publié une carte génétique situant les marqueurs à environ 1,6 cM les uns des autres. Depuis lors, le nombre de marqueurs et la précision des cartes se sont notablement accrus. En même temps, les progrès du séquençage automatique réduisaient le coût du séquençage jusqu'à environ 40 cents par base. Finalement, le terrain était prêt pour les programmes clone-contig et shotgun utilisés pour séquencer et assembler le génome humain.

En 1998, Craig Venter, fondateur de l'Institut pour la Recherche Génomique, déclara que cette société privée, à but lucratif, allait entrer en compétition directe avec le PGH pour le séquençage du génome humain. Venter prétendait que sa stratégie de séquençage et assemblage shotgun serait plus rapide et moins chère que la méthode clone-contig du PGH C'était peut-être vrai, mais la société de Venter, qui s'appelle actuellement Celera Genomics, se basait aussi sur les cartes physiques du PGH accessibles au public. Il n'est pas certain que la méthode de Venter aurait abouti si le génome avait été vaste et complexe. Il est clair que ce fut un succès avec le génome beaucoup plus petit et plus simple de *Haemophilus influenzae*. En juin 2000, une séquence grossière du génome humain fut annoncée conjointement par le PGH, financé par l'État, et par la société privée de Venter, mais plusieurs publications reconnaissaient l'existence de nombreuses lacunes et erreurs. Comme beaucoup de séquences manquantes avaient peu de chance de concerner des gènes, on a généralement admis qu'il s'agissait d'une version du génome humain parfaitement fonctionnelle, mais incomplète.

En avril 2003, le consortium Projet Génome Humain annonçait la finalisation du séquençage du génome humain. La séquence définitive diffère de la séquence brute par moins d'erreurs et de lacunes et une meilleure couverture. Il a fallu 13 années pour arriver à la séquence définitive, et un coût de quelque 2,7 milliards de dollars pour le contribuable américain, en se basant sur la valeur du dollar américain en 1991 ; aujourd'hui, certains séquenceurs très chers et très précis peuvent séquencer un génome humain pour environ 1000 $ en 2 à 3 heures seulement.

Pendant longtemps, les généticiens ont estimé que le génome humain comportait quelque 100 000 gènes. Une des conclusions du Projet Génome Humain fut que le nombre de gènes était d'environ 20 000. Cela ne représente qu'environ 1,5 fois le nombre de gènes de la drosophile et près de la moitié du riz (figure 18.7). Les humains, les souris et le poisson-globe ont à peu près le même nombre de gènes ; il est évident que la complexité de l'organisme ne dépend pas seulement du nombre de gènes dans son génome, ni de la taille de ce génome.

Analyse des données Si le génome humain comporte environ 3 milliards de paires de bases, 20 000 gènes et que 1 % seulement code des gènes, quelle est la longueur approximative d'un gène ?

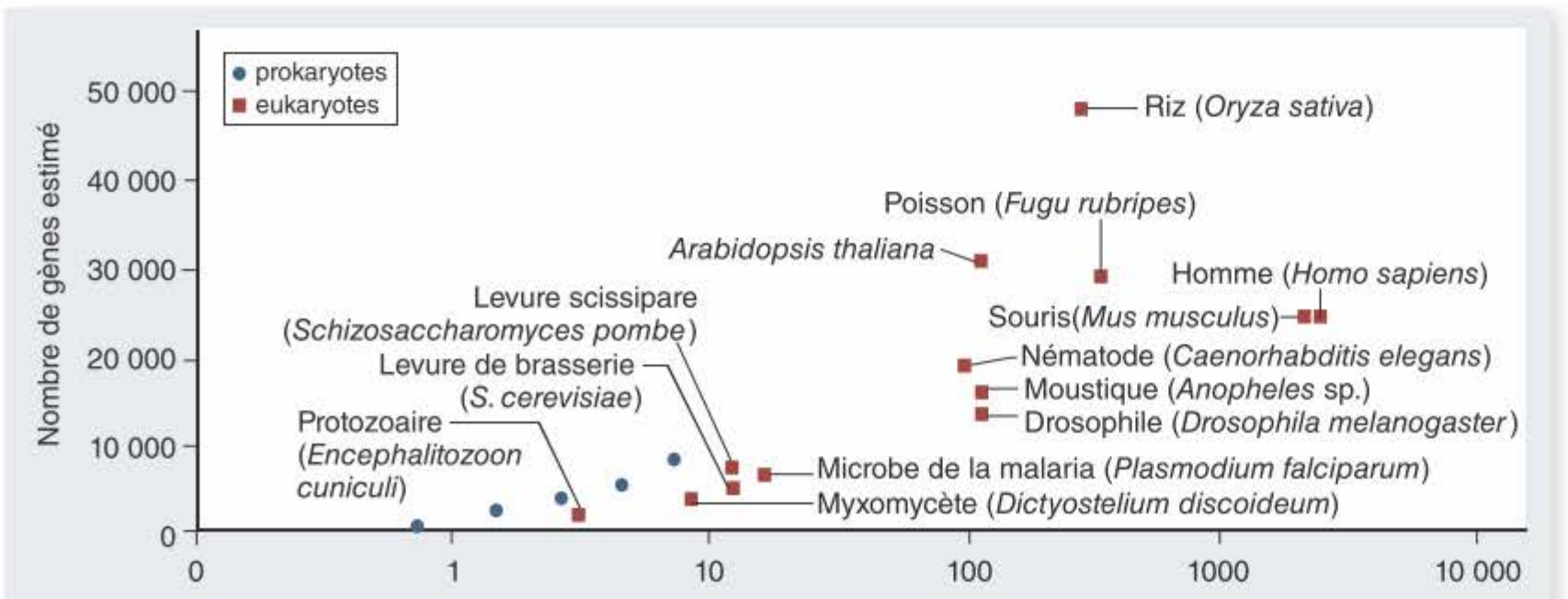


Figure 18.7 Taille et complexité des génomes. Les génomes eucaryotes sont généralement plus grands et possèdent plus de gènes que les génomes procaryotes, bien que la taille de l'organisme ne soit pas le facteur déterminant. Le génome de la souris est plus grand que celui de l'homme, et celui du riz renferme plus de gènes que les deux autres.

Le Projet Génome du Blé illustre les difficultés d'assemblage des génomes complexes

Le blé, *Triticum aestivum*, intervient dans l'alimentation d'environ 30 % de la population mondiale. Une meilleure connaissance de la génomique du blé peut être utile pour l'amélioration des rendements et de la résistance à la sécheresse et aux maladies par sélection. Elle peut aussi éclairer l'évolution des espèces de blé. En raison de l'explosion des populations, de la réduction des ressources en eau et du changement climatique global, l'application des connaissances scientifiques à l'amélioration des plantes deviendra critique pour garantir des ressources alimentaires globales suffisantes.

Une version du génome du blé a été publiée en 2012 ; elle ne couvrait cependant que 70 % des régions renfermant des gènes et ne situait pas les gènes sur les différents chromosomes. En 2014, les chercheurs ont publié une séquence génomique préliminaire avec 124 201 locus géniques, dont 75 000 étaient localisés sur les chromosomes. Dans ce contexte, le plus grand chromosome de blé a été complètement séquencé, annoté et assemblé en une structure sept fois plus longue que tout le génome de la plante modèle qu'est *Arabidopsis thaliana*.

Le blé est un hexaploïde (voir chapitre 24) provenant de deux hybridations. Cela signifie que les cellules somatiques du blé comprennent trois génomes différents, représentés par A, B et D, chacun composé de quelque 5,5 milliards de pb d'ADN répartis entre les 7 chromosomes de chaque génome ancestral. Le séquençage et l'assemblage du génome de blé rencontrent trois obstacles, (1) sa longueur de 16,5 gB (gigabases), la présence de 80 % d'ADN répétitif et (3) sa polyploïdie.

La polyploïdie du génome de blé entraîne une duplication des gènes, la redondance des séquences et l'abondance de l'ADN répétitif. La méthode de séquençage et d'assemblage la plus rapide du génome, la technique shotgun, est inapplicable pour ce type de génome complexe. Avec autant d'ADN répétitif et de duplications de gènes, il est difficile d'assembler correctement les morceaux d'ADN en chromosomes. On doit donc appliquer la méthode clone-contig aux chromosomes individuels, et cette technique exige nettement plus de temps et d'argent. On espère que les 20 autres chromosomes pourront être séquencés et assemblés en 2017.

Les projets sur le génome du cancer sont à la recherche des causes génétiques du cancer

Le cancer tue environ 1500 personnes par jour aux États-Unis. Cette fréquence risque de croître avec l'allongement de la durée de vie moyenne. Des années de recherche ont souligné la base génétique du cancer. Il faut plusieurs mutations pour transformer une cellule normale en cellule cancéreuse – il s'agit d'une série progressive de mutations. Nous avons également identifié deux catégories de gènes impliqués : les oncogènes, capables de provoquer le cancer par mutations positives ou activation inadéquate, et les gènes suppresseurs de tumeurs responsables du cancer par mutations entraînant la perte de leur fonction. On peut aussi considérer ces dernières comme un type positif, « accélérateur » de mutations, et négatif « frein » (voir chapitre 10).

Grâce à la génomique, on a pu compléter cette simple ébauche en comparant les génomes tumoraux aux génomes des tissus normaux correspondants. À ce jour, on a séquencé plus de 5000 échantillons de tumeurs et tissus normaux correspondants. On a pu ainsi identifier les gènes associés à un ou plusieurs types de tumeurs. Les mutations trouvées dans un génome tumoral se répartissent en « mutations pilotes », qui interviennent dans la progression du cancer, et « mutations passagères », qui s'accumulent sans aboutir au cancer. On a identifié plus de 200 gènes différents comme des pilotes potentiels, oncogènes ou suppresseurs de tumeurs, et cette liste n'est pas exhaustive. Ils interviennent dans les voies de transmission des signaux, la réparation de l'ADN, le contrôle de l'expression génique, la structure de la chromatine et même dans le métabolisme. Certains gènes sont mutés dans de nombreux types de tumeurs et des tumeurs différentes peuvent découler de mutations plus ou moins nombreuses.

Cette avalanche de données a clarifié le sujet et, en même temps, fait apparaître de nouveaux problèmes à analyser. Ce n'est pas un schéma simple, avec un ensemble de gènes conduisant toujours à un type spécifique de cancer ; on observe certaines règles. Il est clair aussi que même les types spécifiques de tumeurs sont hétérogènes. Sur le long terme, la caractérisation de chaque type de tumeur sera utile pour le diagnostic et pour la définition de traitements basés sur le génotype du patient et des mutations présentes dans chaque tumeur.

Questions d'apprentissage 18.3

Le Projet Génome Humain a coûté 2,7 milliards de dollars et 13 ans de travail. En comparaison, on pourrait séquencer et assembler le génome du blé, beaucoup plus grand et complexe, pour une fraction de ce montant et deux fois moins de temps. Le séquençage et l'assemblage des génomes de grande taille, complexes et très répétitifs sont un défi technologique. La génomique du cancer propose d'identifier les lots de mutations susceptibles d'indiquer une prédisposition au cancer, de prévoir la progression de la maladie ou d'anticiper la réactivité à des traitements particuliers.

- ***Expliquez pourquoi les génomes complexes, comme celui du blé, sont techniquement plus difficiles à séquencer que les génomes relativement simples.***

18.4 Annotation des génomes et banques de données

Objectifs

1. ***Expliquer pourquoi nous avons besoin de génomes annotés.***
2. ***Comparer les types d'ADN trouvés dans les génomes.***
3. ***Évaluer les conclusions du projet ENCODE.***

Les étapes de séquençage et assemblage d'un programme génomique aboutissent à une masse de données. Ajoutez à ces données, toutes celles qui découlent des analyses ultérieures des séquences et structures génomiques et vous êtes aussitôt confronté au problème de l'organisation et du stockage des données. Les données concernant la séquence et les résultats de leur analyse sont stockés dans différentes banques de données qui peuvent être consultées, en général gratuitement.

Bien que son obtention soit laborieuse, la séquence du génome elle-même a intrinsèquement peu d'intérêt. Le nombre et le type de

gènes présents dans le génome sont plus importants, ainsi que le rôle de ces gènes dans le phénotype. L'affectation de ce type d'information à une séquence génomique est l'*annotation du génome*.

L'annotation du génome assigne une fonction aux séquences d'ADN

Une annotation est une étiquette attachée à une séquence d'ADN conservée dans une banque de données qui l'identifie comme un gène. Cette étiquette peut donner des informations sur le gène, sa structure et sa fonction et sur le produit qui lui correspond finalement (ARN ou protéine). L'annotation peut aussi rappeler comment on a découvert la fonction du gène. Le but ultime est d'essayer de donner, sur les séquences d'ADN du génome, les informations utiles en biologie.

L'annotation des génomes est un processus en grande partie automatisé et contrôlé par des experts si nécessaire. L'annotation débute souvent par l'application d'algorithmes informatiques au repérage, dans l'ADN, des régions susceptibles de renfermer un gène. Ces algorithmes peuvent rechercher des séquences contenant des régions régulatrices telles que les promoteurs, ou un codon de départ suivi de codons correspondant à des acides aminés, et finalement un codon stop. Ces séquences sont désignées comme **cadres ouverts de lecture**, ou **ORF** (*open reading frames*). On peut aussi rechercher des signaux d'épissage pour prévoir une structure intron/exon et, en conséquence, la séquence et la structure d'un ARN qui peut être synthétisé, et ainsi la séquence d'acides aminés codée par le gène supposé.

? Question En vous basant sur la séquence d'un génome, comment pourriez-vous prédire le nombre de gènes qu'il contient. Pourquoi votre estimation pourrait-elle être inexacte ?

Recherche de fonctions dans d'autres espèces : l'algorithme BLAST

Dès qu'un gène ou un produit de gène potentiel a été identifié, il est possible d'en déduire la fonction si l'on trouve une séquence similaire dans une autre espèce. Cette recherche est possible grâce à l'algorithme de recherche BLAST (pour *Basic Local Alignment Search Tool*). Cette recherche peut aussi être automatisée pour accélérer l'annotation. Avec un ordinateur en réseau, les séquences de gènes ou de produits géniques potentiels sont soumis au serveur BLAST et les séquences similaires déjà présentes dans la banque de données sont recherchées. Une ressemblance suffisante avec d'autres gènes ou produits de gènes chez une autre espèce pourrait suggérer que la fonction a été conservée.

L'application des programmes informatiques à la recherche des gènes, à la comparaison des génomes et à leur assemblage ne constitue qu'une des nouvelles méthodes de la génomique réunies sous le terme de **bioinformatique**.

Les marqueurs de séquence exprimée identifient les gènes transcrits

Un autre moyen d'identifier des séquences géniques potentielles dans les génomes est l'isolement de tout les ARNm produits dans différents tissus et leur transformation en ADNc pour le séquençage, puis la localisation de l'ADNc dans la séquence génomique (figure 18.8). Ce processus peut être simplifié en ne séquençant qu'une ou les deux extrémités du plus grand nombre possible d'ADNc. Ce processus peut être automatisé et ces courts fragments d'ADNc ont été appelés **marqueurs de**

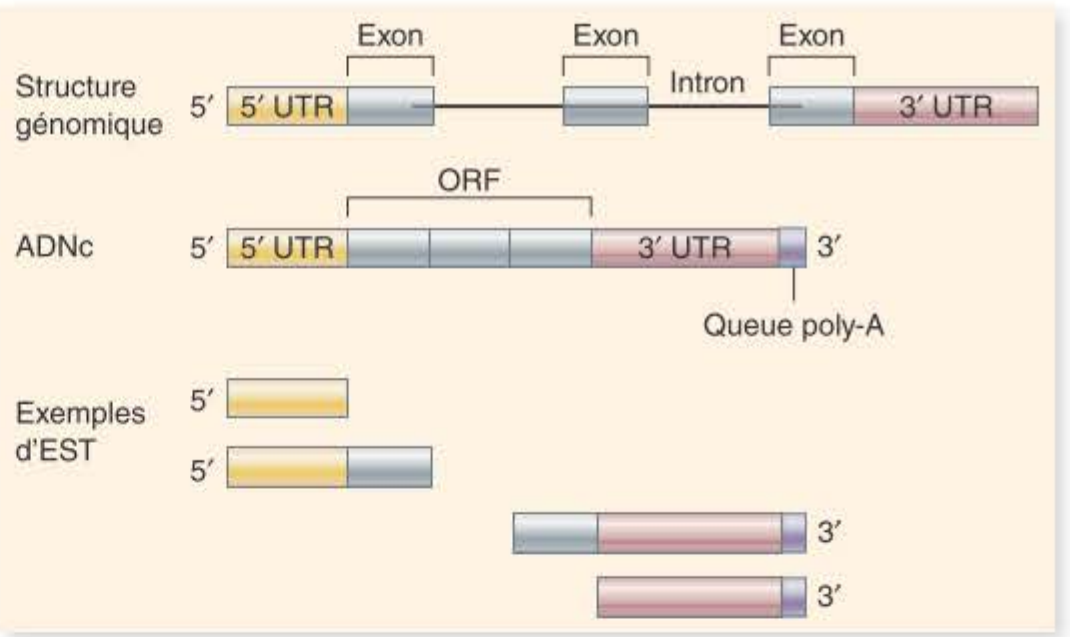


Figure 18.8 On peut utiliser les EST pour localiser, sur la carte du génome, les gènes exprimés. Le séquençage des extrémités des EST permet d'aligner les séquences, de prédire la structure des ADNc et de les localiser sur les structures génomiques correspondantes. Le séquençage de nombreuses EST permet de déterminer les structures génomiques complètes.

séquence exprimée, ou **EST** (pour *expressed sequence tags*). Une EST est une autre forme de STS pouvant être incluse dans les cartes physiques. On peut également utiliser les EST dans les recherches BLAST ; une séquence ressemblant à des EST d'autres espèces peut donner des idées sur la fonction de la protéine codée.

On a utilisé les EST pour identifier 87 000 ADNc dans différents tissus humains. Environ 80 % de ces ADNc étaient inconnus auparavant. Les quelque 20 000 gènes que l'on estime présents dans le génome humain peuvent produire ces 87 000 ADNc différents à cause de l'*épissage alternatif* (voir chapitre 15).

Les génomes contiennent de l'ADN codant et non codant

Les séquences codantes sont les séquences d'ADN qui servent à la synthèse de protéines. Restent toutes les séquences qui n'interviennent pas dans la synthèse de polypeptides, qui sont les séquences non codantes. L'annotation du génome identifie la répartition des séquences codantes et les expériences ultérieures peuvent révéler la fonction de ces gènes et de leurs produits. Dans certains cas, on connaît la fonction de l'ADN non codant ; par exemple, on connaît bien les gènes produisant des molécules d'ARN fonctionnelles comme les ARNt, ARNr et miARN. Par contre, beaucoup d'ADN non codants sont moins bien caractérisés.

L'ADN non codant des eucaryotes

Une caractéristique importante des génomes eucaryotes est la quantité d'ADN non codant. Le Projet Génome Humain en a donné une image particulièrement étonnante. Chacune de nos cellules contient environ 2 m d'ADN mais, sur ces deux mètres, 2,5 cm seulement représentent des séquences codantes ! Près de 99 % de l'ADN de nos cellules sont de l'ADN non codant.

Les gènes codant des protéines sont disséminés dans le génome humain sous forme de paquets au sein de l'ADN non codant. On a décrit sept grands types d'ADN humain non codant (Le tableau 18.1 montre la composition du génome humain, y compris l'ADN non codant) :

L'ADN non codant au sein des gènes. Comme on l'a vu au chapitre 15, un gène humain est formé de nombreux morceaux

TABLEAU 18.1 Catégories de séquences d'ADN trouvées dans le génome humain

Catégorie	Description
Gènes codant des protéines	Portions traduites de quelque 20 000 gènes dispersés parmi les chromosomes
Introns	ADN non codant, représentant la majeure partie de tous les gènes humains
Segments dupliqués	Régions dupliquées du génome
Pseudogènes (gènes inactifs)	Séquences possédant les caractéristiques d'un gène, mais non fonctionnelles
ADN de structure	Hétérochromatine constitutive, située près des centromères et des télomères
Séquences répétées simples	Répétitions de quelques nucléotides, comme CGG, présentes des milliers de fois.
Éléments transposables	21 % : longs éléments intercalés (LINE), qui sont des transposons actifs ; 13 % : courts éléments intercalés (SINE), qui sont des transposons actifs ; 8 % : rétrotransposons contenant de longues répétitions terminales (LTR) aux deux bouts ; 3 % : ADN de transposons fossiles
ARN non codants	ARN qui ne codent pas des protéines, mais ont des fonctions régulatrices, dont beaucoup sont encore inconnues

d'ADN codant (exons) séparés par de l'ADN non codant (introns). Les introns représentent environ 24 % du génome humain, et les exons, moins de 1,5 %

L'ADN de structure. Certaines régions des chromosomes restent très condensées, étroitement spiralées, et ne sont pas transcrites. Ces portions, appelées *hétérochromatine constitutive*, sont surtout localisées autour du centromère ou aux extrémités des chromosomes, au niveau des télomères.

Séquences répétées simples. Des **séquences répétées simples** (**SSR**, pour *simple sequence repeats*) sont dispersées le long des chromosomes. Une SSR est une séquence d'un à six nucléotides, comme CA ou CGG, répétée des milliers de fois. Les SSR peuvent provenir d'erreurs de réplication et leur nombre peut changer à la suite d'erreurs lors de la recombinaison homologue. Elles constituent environ 3 % du génome humain.

Segments dupliqués. Ce sont des blocs de séquences génomiques de 10 000 à 300 000 pb qui ont été dupliqués et se sont déplacés soit au sein du chromosome, soit vers un chromosome non homologue.

Les pseudogènes. Ce sont des gènes inactifs qui peuvent avoir perdu leur fonction à la suite d'une mutation.

Les éléments transposables. Quarante-cinq pour cent du génome humain sont des séquences d'ADN capables de se déplacer dans le génome. Certaines de ces séquences codent des protéines qui permettent leurs déplacements, mais beaucoup ne le font pas. Ces éléments sont décrits dans cette section en raison de leur importance.

Les gènes de microARN. Caché dans l'ADN non codant se trouve un mécanisme de contrôle de l'expression génique. Considéré d'abord comme un "déchet", cet ADN code les microARN, (miARN), qui sont transformés après transcription en séquences de 21 à 23 bases, mais ne sont jamais traduits. On a identifié environ 10 000 miARN différents complémentaires d'un ou plusieurs ARNm matures. Ces miARN contrôlent certains processus complexes du développement chez les eucaryotes par régulation négative de la traduction.

Les longs ARN non codants. Outre les nombreux petits ARN, comme les microARN, qui ne sont pas traduits en protéines mais ont un rôle régulateur, des dizaines de milliers d'ARN non codants plus longs contrôlent vraisemblablement l'expression des gènes. Ce monde caché récemment découvert de réseaux de régulation révèle un nouveau niveau de complexité dans le contrôle précis de l'expression génique. Le rôle des longs ARN non codants est important pour la physiologie et le développement. On a seulement élucidé la fonction d'ARN non codants longs d'environ 200 pb et la fonction biologique éventuelle des autres n'est pas claire.

Les éléments transposables : de l'ADN mobile

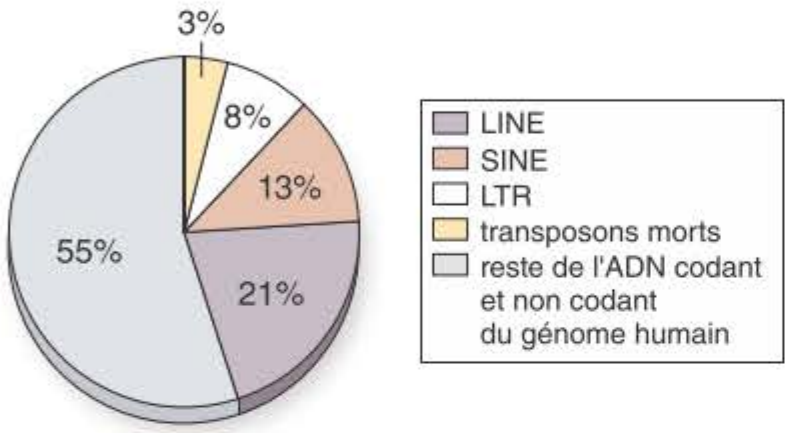
Les **éléments transposables**, aussi appelés *transposons* et *éléments génétiques mobiles*, sont des séquences d'ADN capables de se déplacer d'un endroit à un autre dans le génome. Les éléments transposables disposent de plusieurs moyens pour se déplacer. Dans certains cas, le transposon est dupliqué et l'ADN dédoublé va à un autre endroit du génome. Dans ce cas, le nombre de copies des transposons augmente. D'autres types de transposons sont excisés et s'insèrent ailleurs dans le génome. On parlera du rôle des transposons dans l'évolution du génome au chapitre 24.

Les chromosomes humains renferment quatre types d'éléments transposables. Environ 21 % du génome sont représentés par des **longs éléments nucléaires intercalés** (**LINE**, pour *long interspersed nuclear elements*). Ces éléments anciens et très efficaces sont longs d'environ 6000 paires de bases et renferment tout ce qui est nécessaire à la transposition. Les LINE codent une transcriptase inverse capable de synthétiser une copie d'ADNc de l'ARN transcrit du LINE. Il en résulte un segment bicaténaire qui peut se réinsérer dans le génome au lieu d'être traduit en protéine. On parle de rétrotransposons parce que ces éléments utilisent un ARN comme intermédiaire.

Les **courts éléments nucléaires intercalés** (**SINE**, pour *short interspersed nuclear elements*) ressemblent aux LINE, mais ils ne codent pas les enzymes de transposition et ne peuvent être transposés sans l'équipement de transposition de ces derniers. Nichées parmi les LINE du génome, il existe plus d'un demi million de copies d'un élément

SINE appelé Alu (du nom de l'enzyme de restriction qui coupe dans cette séquence). Le SINE Alu comporte 300 paires de bases et représente 10 % du génome humain. Alu peut utiliser les enzymes du LINE dont il fait partie pour migrer vers un autre endroit du chromosome. Les séquences Alu peuvent aussi sauter à l'intérieur de séquences codantes et provoquer une mutation.

On trouve encore deux autres sortes d'éléments transposables dans le génome humain : 8 % de ce génome sont consacrés à des rétrotransposons appelés **longues répétitions terminales** (**LTR**, pour *long terminal repeats*). Bien que le mécanisme de transposition soit un peu différent de celui des LINE, les LTR utilisent aussi la transcriptase inverse pour produire des copies bicaténaires et réintégrer le génome. D'autre part, des transposons morts occupent environ 3 % du génome ; ce sont des transposons qui ont perdu, par mutation, les signaux nécessaires à la réplication et ne peuvent plus se déplacer.



Question Comment pensez-vous que ces éléments répétés pourraient affecter l'ordre des gènes ?

L'objectif du programme ENCODE est l'identification de tous les éléments fonctionnels du génome humain

On connaît depuis longtemps la présence d'ADN non codant dans les génomes, mais son importance biologique restait obscure. Comme on ne lui connaissait aucune fonction, cet ADN non codant était souvent désigné comme « ADN poubelle ». À la suite d'analyses de grande ampleur de l'ADN non codant, certains ont suggéré de revoir le terme « ADN poubelle ». Il vaudrait mieux considérer l'ADN dont on ne connaît pas la fonction comme « ADN non fonctionnel ».

Le programme ENCODE (pour *Encyclopedia of DNA Elements*) est une tentative collaborative d'identifier tous les éléments fonctionnels du génome humain. La principale conclusion de ce projet fut que 80 % des séquences du génome humain sont fonctionnels. De ces 80 %, environ 62 % sont considérés comme actif dans la transcription, bien que l'on ne connaisse pas la fonction de la plus grande partie de l'ARN transcrit ou susceptible d'être transcrit. Les scientifiques ont beaucoup discuté la signification de ces chiffres en terme de fonction biologique, et l'association ENCODE a revu son estimation pour situer le taux d'ADN fonctionnel entre 20 et 80 %.

Définition de la fonction selon ENCODE

ENCODE considère comme élément fonctionnel toute séquence d'ADN qui entraîne la synthèse d'une protéine, qui est transcrite ou qui possède une signature biochimique distincte et reproductible (figure 18.9). On a une signature biochimique si une protéine s'unit à

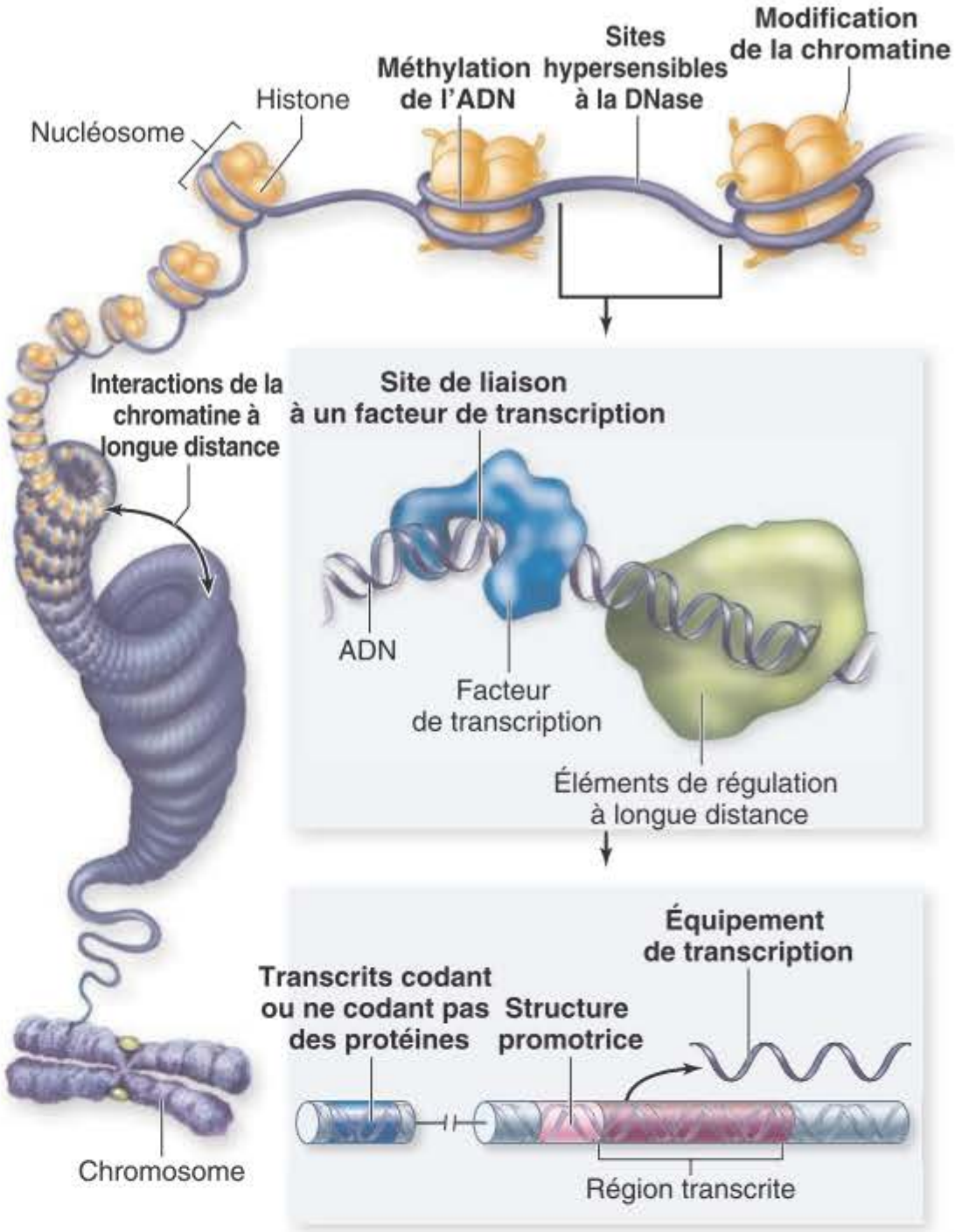


Figure 18.9 Éléments d'ADN fonctionnels définis par ENCODE. ENCODE est basé sur l'identification d'une signature biochimique de la fonction. Les signatures biochimiques sont, par exemple, le type de méthylation des séquences d'ADN, des modifications dans les histones de la chromatine, la susceptibilité à la DNase I, suggérant la présence d'une activité de transcription à cet endroit, des séquences identifiées comme des promoteurs, la production de transcrits codants ou non codants, des éléments régulateurs à longue distance, comme les amplificateurs, et des interactions à longue distance dans la chromatine.

une séquence d'ADN, si la chromatine se recourbe pour des interactions à longue distance, ou si une séquence d'ADN montre une modification particulière de la disposition des histones. Étant donné l'existence probable de différences notables entre ces fonctionnalités dans les différents types de cellules, ENCODE a obtenu cette valeur de 80 % en testant les éléments fonctionnels des génomes de 144 souches différentes de cellules humaines en culture.

ENCODE a été critiqué pour avoir utilisé des souches de cellules en culture et de cellules souches. Il est possible que ces souches ne représentent pas la situation du génome des cellules de l'organisme et que leur activité au niveau de la transcription soit plus grande que dans les cellules des tissus. Il existe aussi probablement des différences importantes entre les éléments fonctionnels au cours des différents stades du développement de l'organisme et dans des conditions environnementales différentes.

La fonction biologique

La définition de la fonction employée par ENCODE n'est pas nécessairement synonyme de fonction biologique. Une séquence d'ADN

qui est transcrite, et donc biochimiquement fonctionnelle pour ENCODE, ne donne pas nécessairement un ARN dont la fonction est indispensable à l'organisme. De même, le fait que l'on ait prouvé le rôle biologique d'un ou plusieurs ARN transcrits ne signifie pas que tous les ARN transcrits ont un rôle biologique. La définition de l'élément fonctionnel pourrait être étendu pour inclure toute séquence d'ADN répliquée et, pour cette définition biochimique, 100 % du génome serait fonctionnel, puisque 100 % du génome est répliqué au cours de la division cellulaire.

Fonction par effet de sélection

Les conclusions d'ENCODE ont aussi été critiquées par certains évolutionnistes. La plupart de ces derniers, et beaucoup de biologistes en général, adhèrent à une définition de la fonction par effet de sélection. Selon cette définition, la fonction d'une caractéristique est soumise à iune sélection purificatrice. Par exemple, la fonction sélectionnée du poumon des mammifères est l'échange gazeux. Les pressions sélectives ont façonné cette fonction. Les changements de volume de la poitrine pendant la respiration sont dus au gonflement et au dégonflement des poumons, mais ces changements de volume n'ont pas été sélectionnés comme une fonction des poumons.

Si une séquence génétique a une fonction particulière, elle finira par disparaître au cours du temps à cause de l'accumulation de mutations aléatoires, à moins d'une neutralisation par sélection naturelle. Une séquence fonctionnelle ne peut échapper à la destruction que si une sélection purificatrice la protège de l'accumulation de mutations qui dégradent la fonction. Beaucoup de biologistes soutiendraient que seules les séquences d'ADN soumises à une sélection purificatrice devraient être considérées comme fonctionnelles. Ces séquences ont plus de chances d'être conservées dans une lignée spécifique et aussi d'être conservées dans des espèces apparentées. Sur la base de ces critères, 5 à 15 % seulement du génome serait fonctionnel.

Il est clair qu'il existe des points de vue contradictoires quant à la fonctionnalité des séquences du génome humain, et la fonctionnalité de 80 % proposée par ENCODE reste discutable. Cela étant, les techniques et les modes de collecte des données proposés par ENCODE font progresser notre connaissance des fonctions des séquences de l'ADN génomique. Les résultats d'ENCODE concernant les fonctions biochimiques de l'ADN non codant sont également à l'origine d'une carte impressionnante utilisable par d'autres pour explorer la pertinence biologique des éléments fonctionnels identifiés dans le génome humain.

Questions d'apprentissage 18.4

Quand un génome a été séquencé, assemblé et déposé dans une banque de données adéquate, ses éléments fonctionnels sont annotés avec les informations utiles. On peut identifier les éléments fonctionnels de différentes manières, comme par déduction, en appliquant le programme BLAST. Seule une fraction du génome possède l'information pour la synthèse des protéines ; le reste de l'ADN non codant peut avoir une fonction. À cause de l'ambiguïté du terme « fonction », les conclusions concernant le pourcentage d'ADN fonctionnel dans un génome doivent être prises avec prudence.

- *Sur la base des principes de la sélection naturelle, comment pourriez-vous expliquer le nombre élevé d'éléments transposables dans le génome humain ?*

18.5 La génomique comparative et fonctionnelle

Objectifs

1. *Expliquer l'intérêt de la génomique comparative pour étudier les propriétés des génomes.*
2. *Montrer comment la génomique fonctionnelle permet d'étudier les fonctions des génomes.*
3. *Décrire les relations entre le génome, le transcriptome et le protéome.*

La génomique comparative se base sur ce que l'on connaît d'un génome pour en étudier un autre. On peut, par exemple, utiliser la fonction connue d'un gène d'un organisme pour imaginer la fonction d'un gène semblable présent dans le génome d'un organisme apparenté. Il s'avère que 60 % des gènes intervenant dans le déclenchement des cancers humains se retrouvent dans le génome de la drosophile. On peut donc étudier les fonctions de ces gènes chez la drosophile, ce qui est nettement plus facile que leur analyse chez les humains. La génomique comparative nous renseigne aussi sur le degré de parenté entre les organismes, sur le déroulement de fonctions semblables chez des organismes différents, sur ce qui fait que des espèces sont différentes et même sur le nombre minimum de gènes nécessaires pour construire une cellule fonctionnelle – question intéressante pour les biologistes de synthèse (section 18.6).

La génomique fonctionnelle est un développement de la génomique qui étudie la relation entre le génotype et le phénotype. Le phénotype de l'organisme est déterminé par le mode d'expression des gènes et par une interaction de l'organisme avec son environnement. Par exemple, en comparant l'expression des gènes dans une cellule saine et dans une cellule malade, il est possible de savoir quels gènes peuvent intervenir dans la maladie.

La génomique comparative révèle des régions conservées dans les génomes

Une des leçons remarquables découlant de la séquence du génome humain est la grande ressemblance entre les humains et les autres organismes. Plus de la moitié des gènes de la drosophile ont leur correspondant chez l'homme. Parmi les mammifères, les similitudes sont encore plus grandes. Les humains ne possèdent que 300 gènes qui ne sont pas représentés dans le génome de la souris. La génomique comparative est un autre domaine très prometteur pour répondre aux questions évolutives. La comparaison des nombreux génomes procaryotes déjà séquencés indique un transfert latéral de gènes plus important qu'on ne l'avait soupçonné auparavant.

Génomique comparative basée sur la synténie

La génomique comparative permet de comparer les génomes sur une grande échelle en tenant compte de la synténie. La *synténie* désigne le maintien de la répartition des segments d'ADN dans des génomes apparentés. On peut utiliser les cartes physiques pour rechercher la synthénie dans des génomes non séquencés. Les comparaisons avec le segment

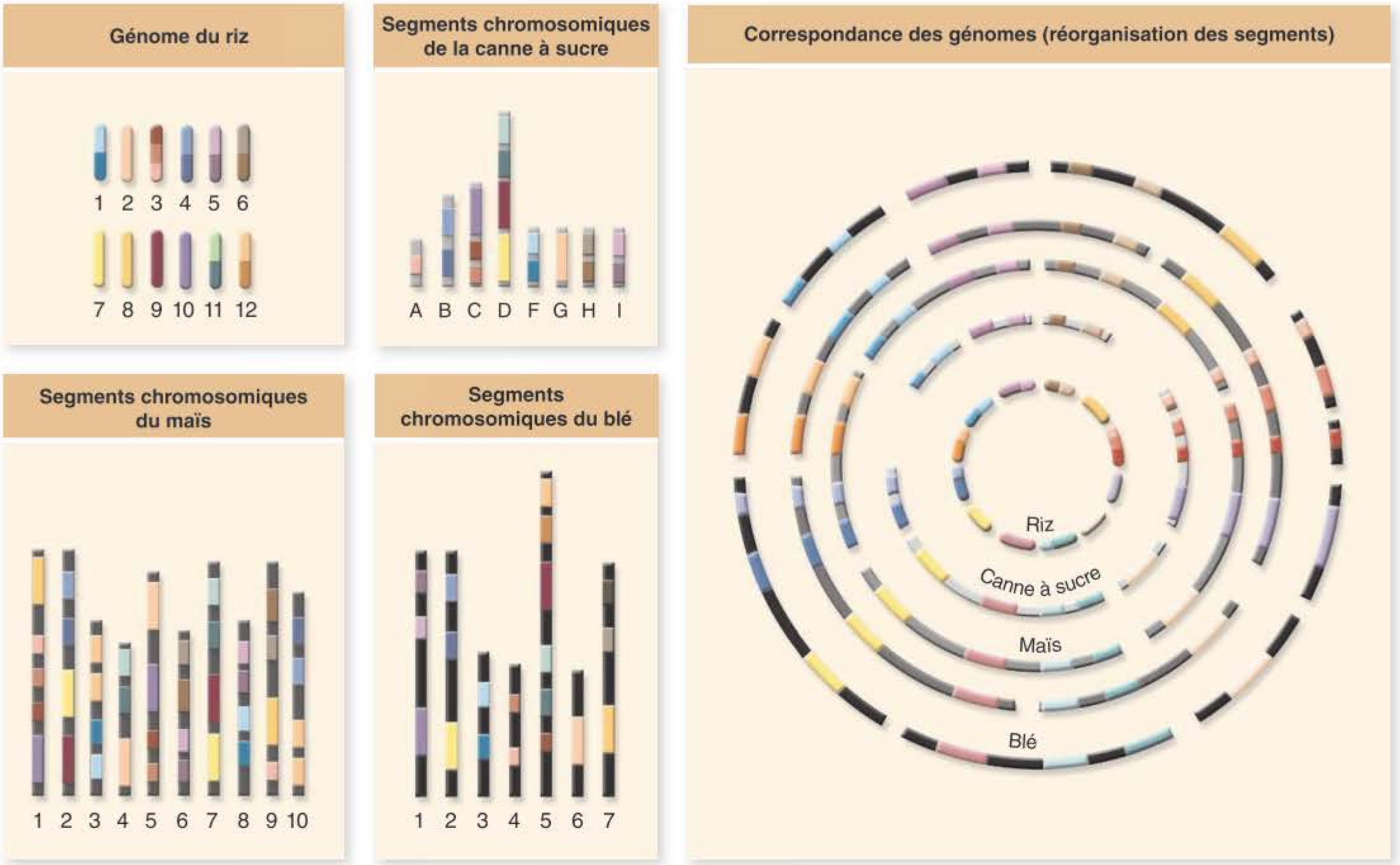


Figure 18.10 Les génomes des céréales sont composés de segments chromosomiques semblables répartis différemment. Les bandes de même couleur représentent des morceaux d'ADN qui ont été conservés dans les différentes espèces mais ont été réarrangés. En découpant les chromosomes des principales espèces de céréales en morceaux et en réarrangeant les morceaux, les chercheurs ont constaté que les éléments qui composent le génome du riz, de la canne à sucre, du maïs et du blé sont très conservés. Cela signifie que la répartition des segments du génome ancestral des graminées a été modifiée au cours de l'évolution.

synténique séquencé d'une autre espèce peut donner des informations sur le génome non séquencé.

Prenons par exemple le riz et les céréales apparentées, maïs, orge et blé. Seuls les génomes du riz et du maïs ont été complètement séquencés. Bien que ces plantes soient séparées depuis plus de 50 millions d'années, les chromosomes du riz, du maïs, du blé et d'autres céréales montrent une synthénie importante (figure 18.10). Quand on considère le génome, «le riz, c'est le blé», et les informations glanées dans les cartes physiques et génétiques des génomes du maïs et du riz permettent d'accélérer le séquençage en cours du génome de blé.

Connaissant la séquence d'ADN des génomes du riz et du maïs, il devrait être beaucoup plus facile d'identifier et d'isoler des gènes de céréales à génome plus long. L'analyse des séquences d'ADN des céréales pourrait être importante pour l'identification des gènes associés à la résistance aux maladies, à la productivité, à la qualité nutritive et à la croissance potentielle.

La génomique fonctionnelle permet de connaître la fonction des gènes au niveau du génome

Pour bien comprendre la contribution du génome à la structure et au fonctionnement de l'organisme, nous avons besoin de caractériser ses produits – les molécules d'ARN et les protéines. Cette information est essentielle pour la compréhension de la biologie, de la physiologie, du développement et de l'évolution. La génomique fonctionnelle permet de faire la liaison entre le génotype de l'organisme et son phénotype.

La génomique fonctionnelle fait appel à une série de techniques à haut rendement pour étudier les produits du génome, pour voir comment ils sont synthétisés et comment ils se modifient au cours du développement ou en réponse à l'environnement. Dans la génomique fonctionnelle, on peut reconnaître trois orientations distinctes, mais apparentées : (1) l'étude de toutes les molécules d'ARN produites par le génome (le transcriptome), (2) l'étude des protéines produites par le génome (le protéome) et (3) l'étude des interactions et des produits des interactions entre protéines. Nous envisagerons la protéomique plus tard dans cette section, mais nous allons d'abord donner un aperçu de quelques techniques utilisées pour l'étude du transcriptome.

Les micro-alignements d'ADN

Des **micro-alignements d'ADN**, ou **puces à ADN**, ont été créés pour l'analyse de l'expression génique au niveau du génome entier. Ces puces permettent d'analyser les modifications de l'expression des gènes au cours du développement ou en réponse à l'environnement, et même pendant l'évolution d'une maladie. Ces analyses ont besoin d'une annotation adéquate du génome et ces données sont utilisées pour construire un alignement comprenant toutes les séquences codantes (figure 18.11).

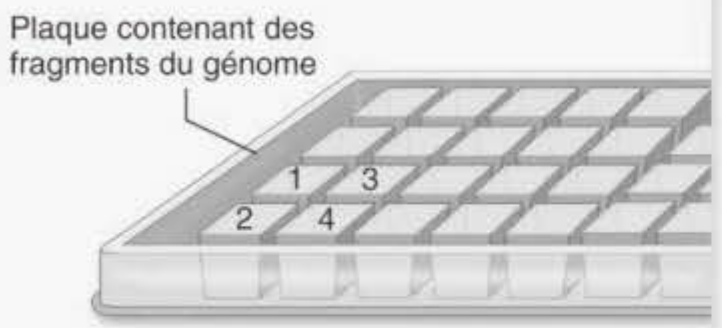
DÉMARCHE SCIENTIFIQUE

Hypothèse: *les fleurs et les feuilles exprimeront certains gènes communs.*

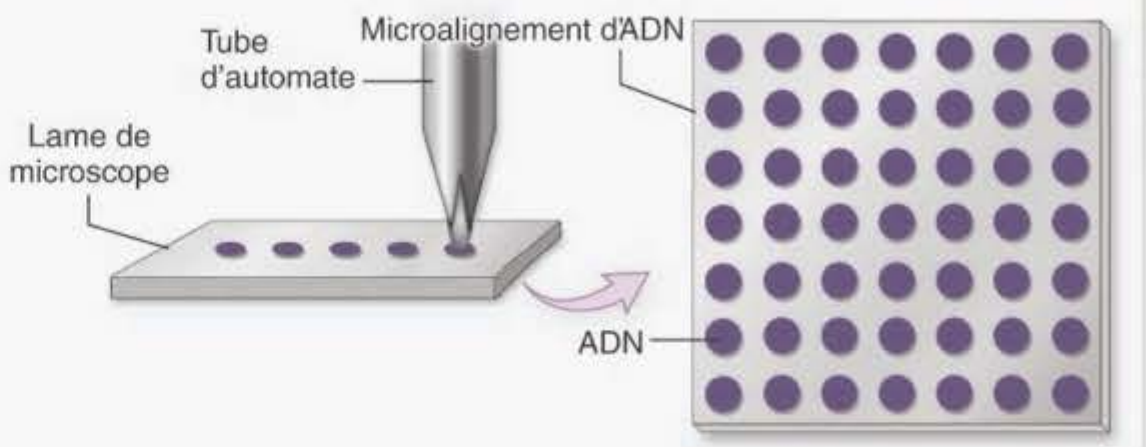
Prédiction: *si l'on utilise des ARNm isolés de fleurs et de feuilles d'*Arabidopsis thaliana *comme sondes sur un microalignement du génome d'*Arabidopsis, *les deux lots d'amorces s'hybrideront avec les séquences communes et uniques.*

Test:

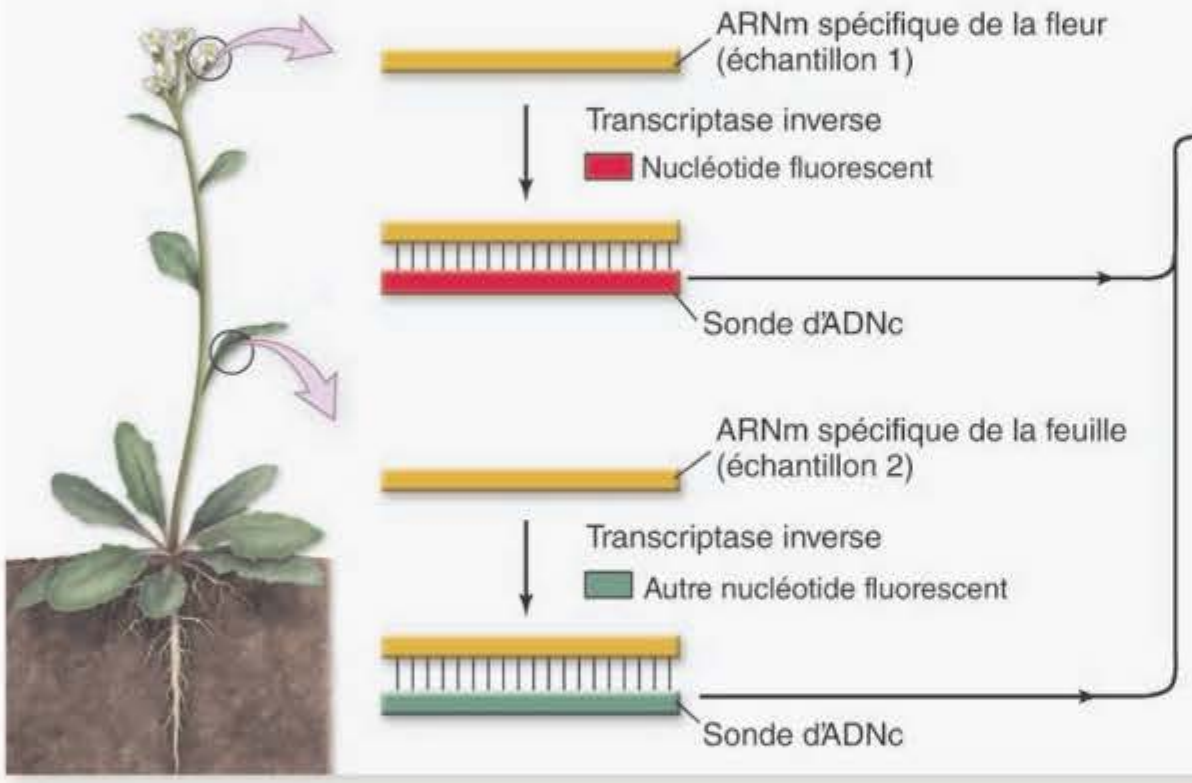
1. Partir d'un microalignement du génome d'*Arabidopsis*. Les fragments génomiques d'*Arabidopsis* uniques, amplifiés par PCR (1, 2, 3, 4…) sont dans tous les puits de la plaque



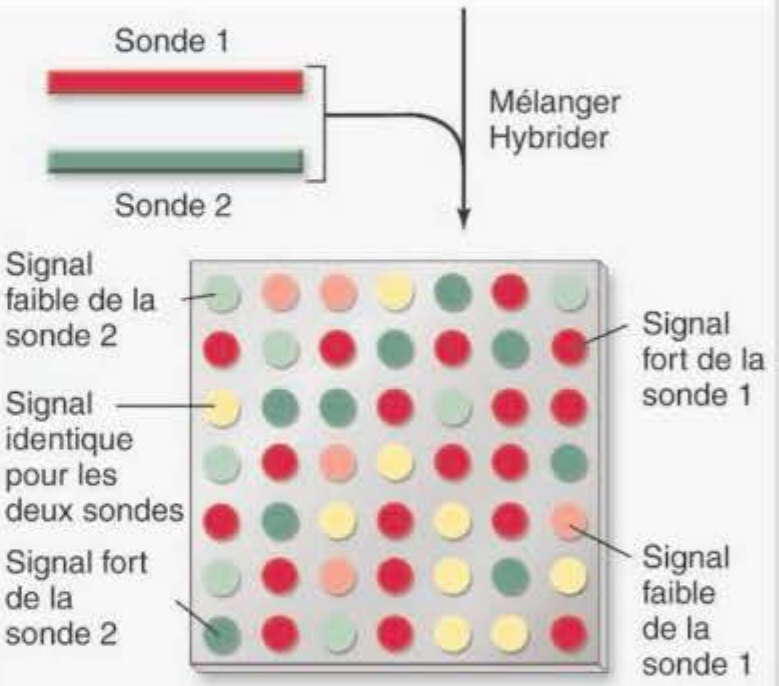
2. L'ADN est déposé sur la lame de microscope.



3. Isoler l'ARNm de fleurs et de feuilles, le transformer en ADNc et le marquer par fluorescence. On obtient des échantillons d'ARNm des deux tissus. On prépare des sondes de chaque échantillon avec un nucléotide fluorescent différent pour chacun.



4. Microalignement avec ADNc marqué. Les deux sondes sont mélangées et hybridées avec le microalignement. On analyse les signaux fluorescents du microalignement.



Résultat: *les taches jaunes représentent les séquences hybridées aux ADNc des fleurs et des feuilles. Les taches rouges représentent les gènes exprimés seulement dans les fleurs. Les taches vertes correspondent aux gènes qui ne s'expriment que dans les feuilles.*

Conclusion: *certains gènes d'*Arabidopsis *s'expriment dans les fleurs et dans les feuilles, mais d'autres ne s'expriment que dans les fleurs ou dans les feuilles.*

Autres expériences: *comment pourriez-vous utiliser les microalignements pour voir si les gènes exprimés dans les fleurs et dans les feuilles sont des gènes « de ménage » ou spécifiques des fleurs et des feuilles ?*

Figure 18.11 Les micro-alignements

Pour préparer une puce à ADN, un robot dépose des fragments d'ADN sur une plaque en verre ou en silicone pour former un alignement de points. Chaque fragment d'ADN correspond à une séquence génique spécifique et l'ensemble de la puce représente tous les transcrits codant potentiels, chacun présent une seule fois. L'utilisation d'une puce est illustrée à la figure 18.11. Il est possible d'utiliser les puces dans des expériences d'hybridation avec des ARNm marqués d'origines diverses. On obtient ainsi un aperçu général des gènes actifs et inactifs dans des conditions ou états divers. On peut aussi préparer des puces avec différentes formes de la séquence d'ADN de façon à ne détecter que les gènes avec polymorphisme d'un seul nucléotide.

Les micro-alignements souffrent cependant de certaines limites. Il faut connaître la séquence du génome pour préparer une puce avec les séquences appropriées. Si la puce est préparée pour rechercher des transcrits polymorphes rares dans un échantillon, il faut aussi être au courant de ce polymorphisme. À cause de certaines de ces restrictions, et en raison des progrès technologiques, d'autres méthodes sont devenues populaires pour l'étude du transcriptome.

Analyse en série de l'expression des gènes

SAGE (*serial-analysis of gene expression*) est une technique à haut débit permettant l'examen simultané de tous les ARNm présents dans un échantillon – comme un groupe de cellules cancéreuses. Pour ce faire, l'ARNm est isolé à partir d'un échantillon et transformé en ADNc par la transcriptase inverse, comme on l'a vu au chapitre 17. L'ADNc est découpé en petits fragments qui sont réunis en longues chaînes. Ces chaînes sont clonées dans des vecteurs et séquencées par les techniques de dernière génération. Les fragments d'ADNc sont suffisamment courts pour permettre l'identification des ARNm correspondants. SAGE présente trois avantages par rapport aux puces d'ADN. Premièrement, il n'est pas nécessaire de connaître le génome. Deuxièmement, grâce à la quantité d'information obtenue par le séquençage de dernière génération, il est possible d'identifier les molécules d'ARNm les plus rares. Troisièmement, l'estimation de la quantité d'ADNc correspondant à chaque ARNm est beaucoup plus précise qu'avec les puces.

Séquençage de l'ARN (RNA-seq)

Le séquençage de l'ARN ressemble beaucoup à SAGE. Comme SAGE, l'RNA-seq donne une vue d'ensemble de tous les ARNm d'un échantillon, mais il le fait par séquençage direct de l'ADNc. Comme pour SAGE, le séquençage passe par les techniques de dernière génération. On peut modifier l'RNA-seq en fonction des problèmes expérimentaux qui se présentent. Par exemple, on peut s'en servir pour identifier les frontières entre introns et exons, localiser le polymorphisme d'un nucléotide unique dans des génomes non séquencés et déterminer le rôle des différents allèles d'un gène dans le phénotype. Comme SAGE, l'RNA-seq pose un problème : c'est souvent la protéine, plutôt que l'ARN, qui définit le phénotype. La présence d'un ARN détecté par RNA-seq ou SAGE ne tient pas compte d'un contrôle possible de l'ARN après la transcription et de ses conséquences pour la quantité de protéine. L'analyse du transcriptome d'une cellule peut être une source importante d'information, mais il est parfois essentiel d'envisager aussi l'ensemble des protéines d'une cellule à un moment particulier. Ces travaux concernent un volet de la génomique appelé *protéomique*

La protéomique établit un catalogue des protéines codées par le génome

Il est plus difficile d'interpréter les données du protéome que celles du génome ou du transcriptome. Beaucoup de protéines possèdent en effet des groupements fonctionnels ou d'autres structures chimiques ajoutées après la traduction. Ces additions peuvent modifier les activités et la localisation des protéines dans la cellule. En outre, beaucoup de protéines ne fonctionnent que dans des structures quaternaires, et elles interagissent les unes avec les autres pour réguler la structure et le fonctionnement des cellules. La protéomique doit encore faire face au défi – lié à l'épissage alternatif de l'ARN – un même gène peut coder beaucoup de protéines (figure 18.12). Comme le transcriptome, le protéome est très dynamique. La séquence des nucléotides est généralement constante, mais le transcriptome diffère selon le type de tissu, le stade de développement et les conditions environnementales. Le protéome est dynamique parce que le transcriptome est dynamique.

Beaucoup de techniques à haut débit ont été mises au point pour étudier le protéome. Le choix de la technique dépend du problème posé. Une technique de plus en plus utilisée pour l'analyse du protéome est la *spectroscopie de masse*. Cette technique détermine le rapport entre la masse et la charge des molécules et permet d'identifier les petites molécules sans erreur possible. Pendant de nombreuses années, elle n'a pas été appliquée aux protéines à cause de leur taille. Les nouvelles techniques combinent cependant une purification et une ionisation permettant l'application de la spectroscopie de masse à l'analyse de mélanges complexes de protéines (figure 18.13). Un avantage de cette technique est la possibilité d'identifier des modifications survenant après la traduction, comme le phosphorylation.

La protéomique fait également appel à la technologie des puces déjà décrite dans cette section. L'ADN réparti sur la puce est remplacé par des anticorps qui reconnaissent des protéines spécifiques. Si un mélange de protéines isolé d'un échantillon est appliqué sur une puce d'anticorps, ces derniers se fixent aux protéines spécifiques comme dans les immuno-essais décrits au chapitre 17. On peut ainsi identifier rapidement des protéines spécifiques, ou biomarqueurs, présentes dans un mélange. Utile pour les programmes de recherche, cette technique pourrait aussi être appliquée au diagnostic et au criblage des maladies.

Bioinformatique et protéomique

Tout comme l'analyse du génome et du transcriptome, les analyses à haut débit du protéome sont à l'origine d'une énorme quantité de données. Ces données auraient peu d'intérêt s'il n'était pas possible de les étudier et de les interpréter. La bioinformatique se situe à la frontière entre l'informatique et la biologie. Elle applique la programmation, les mathématiques et la modélisation à l'analyse de la masse de données biologiques. Ses applications à la protéomique permettent l'identification rapide des protéines découvertes lors des expériences de collecte de

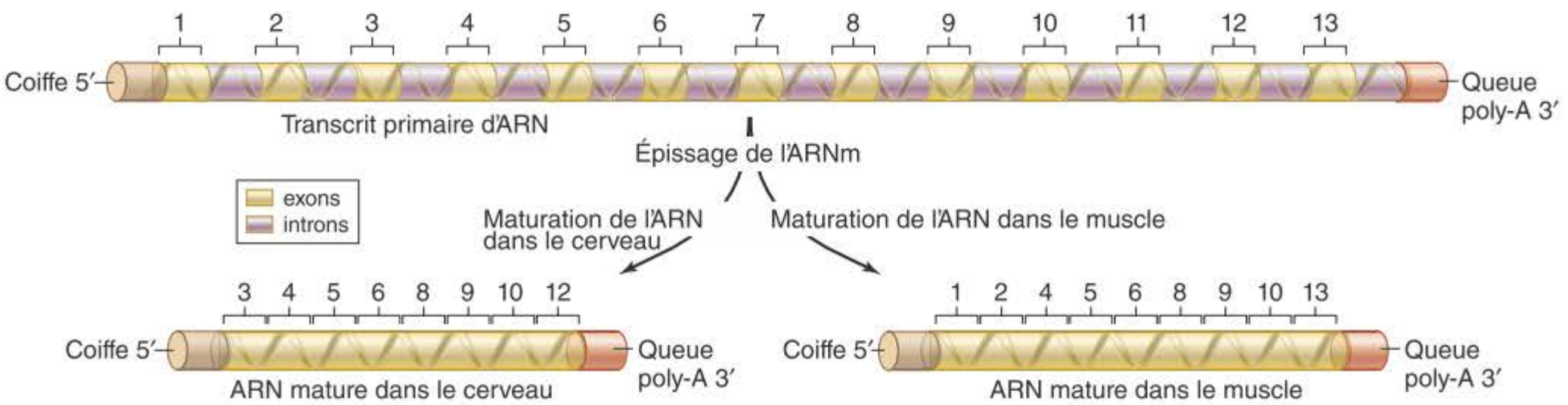


Figure 18.12 L'épissage alternatif peut entraîner la transcription d'ARNm différents à partir de la même séquence codante. Dans certaines cellules, des exons peuvent être excisés en même temps que les introns voisins, donnant ainsi des protéines différentes. L'épissage alternatif explique pourquoi 20 000 gènes humains peuvent coder trois ou quatre fois plus de protéines.

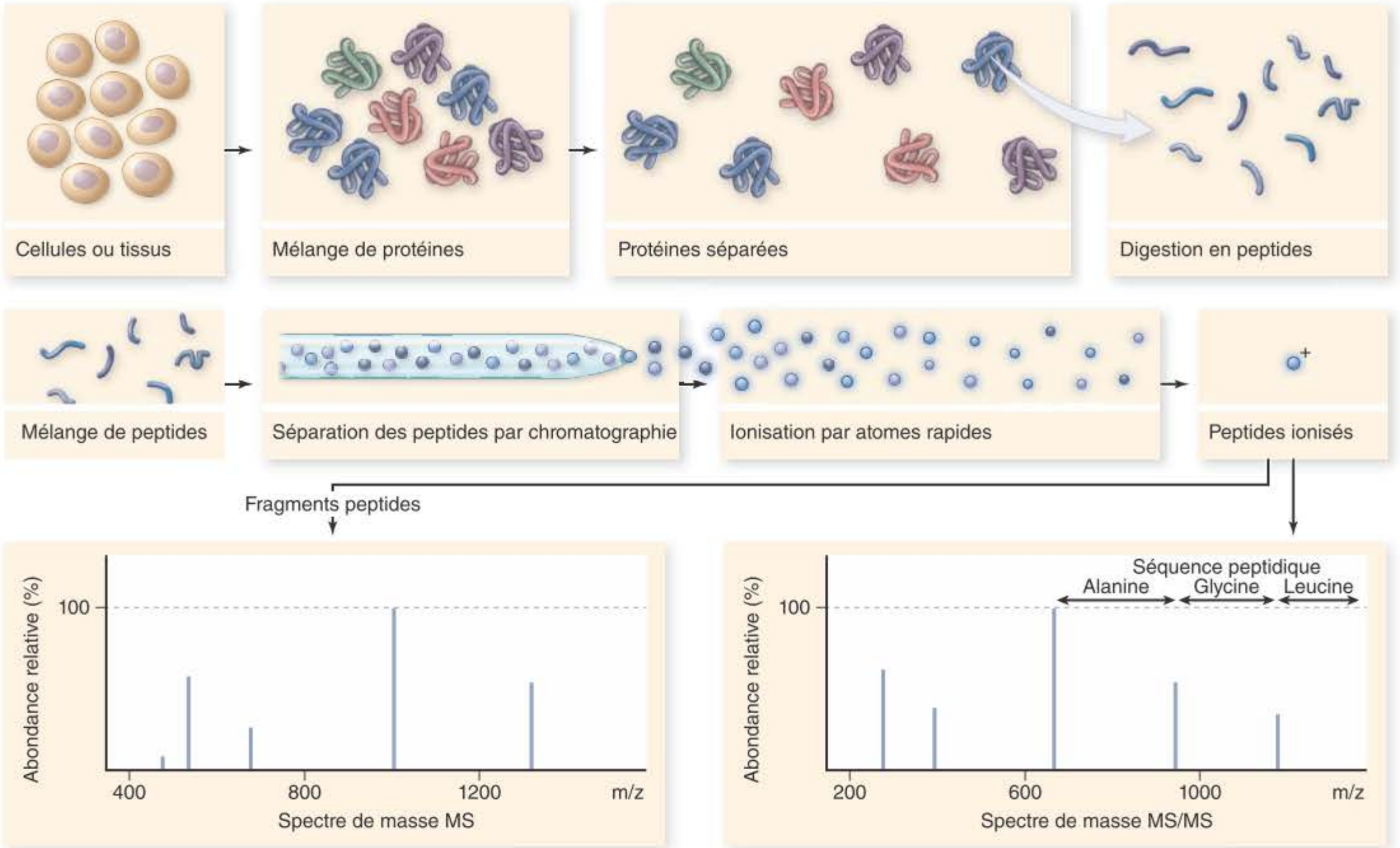


Figure 18.13 On peut utiliser la spectroscopie de masse pour l'analyse des protéines. On peut isoler les protéines des cellules ou des tissus et les séparer partiellement par des gels ou des techniques biochimiques. Les protéines partiellement purifiées ou individuelles sont digérées en petits peptides par des protéases et ionisées. Les peptides chargés peuvent être séparés en fonction de leur masse dans un spectromètre. Des algorithmes informatiques peuvent faire correspondre les fragments peptidiques avec un certain type de digestion pour essayer d'identifier les protéines. D'un autre côté, on peut encore fragmenter les peptides et déterminer la composition en acides aminés par spectrographie de masse. (spectre SM, abondance des ions classés en fonction du rapport entre leur masse et leur charge ; spectre SM/SM, abondance des ions classés en fonction du rapport entre la masse et leur charge quand ils ont été obtenus par ionisation en tandem ; m/z, rapport entre la masse d'un ion et sa charge.)

données à haut débit, l'annotation des génomes grâce aux informations fournies par la spectrométrie de masse et les puces de protéines, et la prédiction de la structure et de la fonction des protéines.

? Question Comment pourriez-vous annoter la fonction d'un gène dans une banque de données génomique si vous avez accès à la séquence d'acides aminés d'une protéine obtenue par une expérience de protéonomique ?

Prédiction de la structure et de la fonction des protéines

La protéomique se distingue de la chimie traditionnelle des protéines par l'application des nouvelles méthodes informatiques à l'identification rapide et à la caractérisation d'un grand nombre de protéines. Comme pour la génomique, le défi est lié à l'échelle envisagée.

Idéalement, nous voudrions découvrir la structure et la fonction de la protéine à partir de la séquence du gène qui la code. Pour l'instant, nous en sommes réduits à comparer la séquence de la nouvelle protéine aux séquences de protéines déjà identifiées. On peut parfois imaginer la structure et la fonction en supposant que des séquences similaires d'acides aminés donneront une structure semblable. Cependant, il vrai aussi que des séquences très différentes d'acides aminés peuvent aboutir à des structures semblables. Dans ce cas, on ne peut imaginer des fonctions semblables que si l'on dispose d'informations sur la structure.

Proposer une fonction en se basant sur la structure est encore plus délicat. Si leurs séquences et structures sont semblables, les protéines ont souvent des fonctions semblables, mais on connaît de nombreux exemples où ce n'est pas le cas. Beaucoup de fonctions semblables sont le fait de protéines avec des structures très différentes. Cette analyse est utile si nous connaissons l'histoire évolutive de la protéine. Nos déductions sur les structures et fonctions communes sont plus fiables si nous comparons des protéines dérivées d'un ancêtre commun dont les fonctions n'ont pas divergé.

Jusqu'il y a peu, il était très difficile de prédire les différents modes de pliage des chaînes d'acides aminés en structures secondaires et tertiaires (un exemple de structure protéique prédite est illustré à la figure 18.14). Un algorithme informatique essaye de prédire la structure des protéines en recherchant les séquences d'acides aminés caractéristiques d'un domaine protéique. L'algorithme prend ensuite de courts fragments du domaine protéique et les compare à des plis connus de protéines dont on connaît la structure. Un programme informatique

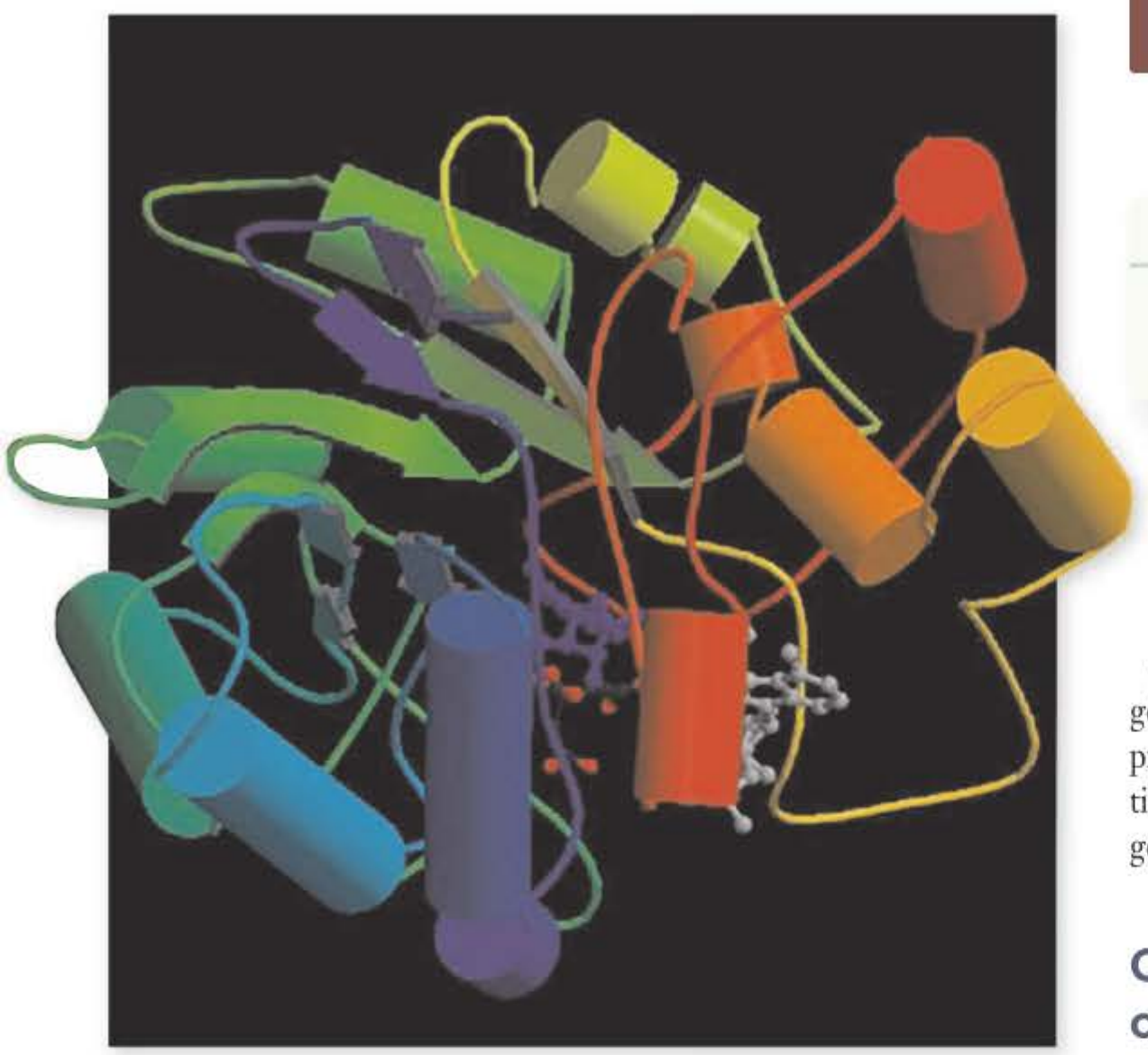


Figure 18.14 Schéma d'une enzyme obtenu par informatique.
On peut retrouver la structure des protéines connues dans des banques de données, comme celle de l'aldose réductase humaine, représentée ici. Les motifs secondaires ont des couleurs différentes.

utilise ensuite ces données pour construire un modèle montrant comment tous les fragments pourraient s'organiser pour donner des structures spécifiques. Quand la même structure apparaît dans des simulations indépendantes, il est probable qu'elle est correcte.

? Question Quelle est la relation entre génome, transcriptome et protéome ?

Questions d'apprentissage 18.5

En comparant différents génomes, les généticiens peuvent déduire les relations structurelles, fonctionnelles et évolutives entre les gènes et les protéines, ainsi que les relations entre les espèces. La génomique fonctionnelle offre des outils pour entamer l'étude des fonctions des gènes du génome. On peut appliquer les micro-alignements d'ADN, SAGE er RNA-seq pour l'étude du transcriptome, mais ces différentes méthodes ont des avantages et des inconvénients. La protéomique implique l'analyse des protéines produites par le génome dans des conditions spécifiques. Il est possible de faire des prédictions à propos de la structure et de la fonction des protéines, mais ces analyses sont complexes et nécessitent des ordinateurs puissants.

- *Pourquoi la création du transcriptome d'une espèce est-elle une étape importante dans l'étude de son protéome ?*

18.6 Applications de la génomique

Objectif

1. *Évaluer les conséquences éthiques et sociales de l'utilisation des données de la génomique.*

Nous pouvons nous servir de la génomique pour répondre à des questions fondamentales concernant la biologie. Nous pouvons, par exemple, mieux comprendre pourquoi une espèce est fondamentalement différente d'une autre et comment le génome intervient dans la détermination du phénotype si nous connaissons les fonctions du génome et les relations évolutives entre génomes spécifiques. Plusieurs problèmes sociétaux et éthiques découlent de l'application des informations dont nous disposons aujourd'hui sur la structure et la fonction du génome.

On a utilisé les données génomiques pour créer la première cellule « synthétique ».

La vie est caractérisée par un ensemble de propriétés découlant de l'assemblage d'atomes en molécules, de molécules en macromolécules et de macromolécules en structures cellulaires. Les mécanismes contrôlant la synthèse de ces protéines et les processus aboutissant à l'édification des structures cellulaires sont sous le contrôle du génome. Mais quelle est la taille minimale d'un génome capable de construire une cellule capable de se répliquer de façon autonome ?

Nous pouvons avoir une idée du plus petit génome requis pour entretenir la vie en recherchant le plus petit génome d'un organisme vivant. Le plus petit génome viral compte 5386 pb, mais beaucoup de biologistes ne considèrent pas les virus comme vivants parce qu'ils ne sont pas capables de se répliquer indépendamment. Le génome de la mitochondrie semi-autonome est long d'environ 16 600 pb ; la mitochondrie est cependant incapable de se répliquer indépendamment de sa cellule « hôte ». Le plus petit génome d'une bactérie autonome actuellement connu est celui de *Mycoplasma genitalium*, avec quelque 500 000 pb. Cela implique que la taille minimale d'un génome d'une cellule vivante se situe quelque part entre 16 600 et 500 000 pb.

En 2010, les scientifiques du J. Craig Venter Institute ont annoncé avoir conçu un génome de *Mycoplasma mycoides* à l'aide d'un ordinateur, synthétisé l'ADN, assemblé l'ADN en chromosome à l'intérieur de cellules de levure, et transféré le chromosome dans la bactérie *Mycoplasma capricolum*. Le génome de la bactérie réceptrice a été détruit au cours du processus et le nouveau génome synthétique était capable de contrôler le développement de la nouvelle cellule bactérienne, baptisée *Mycoplasma mycoides* JCVI-syn1,0 (figure 18.15). De cette façon, il est possible de tester expérimentalement un génome minimal.

La génomique peut aider à identifier et traiter les maladies

La révolution génomique a produit des millions de gènes qu'il faut étudier. La génomique peut avoir un impact énorme sur la santé humaine. Des mutations survenant dans un seul gène peuvent expliquer un

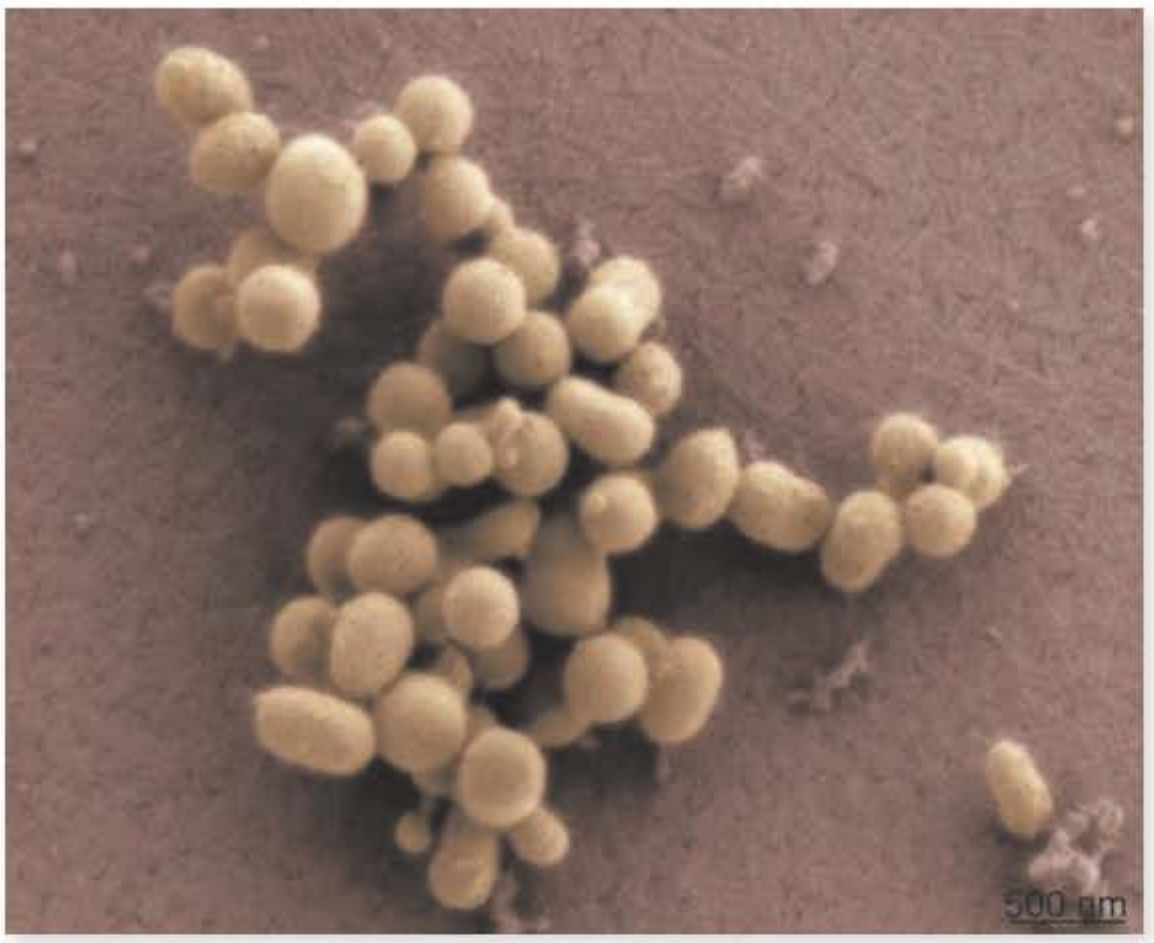


Figure 18.15 Photo, au microscope électronique à balayage, de bactéries synthétiques *Mycoplasma mycoides* JCVI-syn1.0.

nombre limité de maladies héréditaires. Mais, avec la génomique, on pourrait étudier tous les caractères influencés par la génétique.

La génomique est susceptible de conduire à de nouveaux médicaments, mais on voit son impact immédiat sur le diagnostic. Les progrès technologiques et la découverte des gènes ont optimisé la diagnose des anomalies génétiques. Le diagnostic est également utilisé pour l'identification des individus. Par exemple, les courtes répétitions en tandem (STR) découvertes à la suite des recherches en génomique, ont été des outils de diagnostic médico-légal pour identifier les restes des victimes de l'attaque terroriste du World Trade Center de New York le 11 septembre 2001.

On s'est particulièrement intéressé aux armes biologiques après l'attentat du 11 septembre aux États-Unis. Cinq personnes sont mortes et 17 autres ont été infectées suite à l'envoi par la poste d'enveloppes contenant des spores du charbon (anthrax). Grâce au séquençage du génome, on a identifié la source des bactéries. Une différence de 10 pb seulement entre souches a permis au FBI de remonter à la source dans un unique flacon de bactéries utilisé dans un programme de recherche sur les vaccins à l'institut de recherches médicales de l'armée pour les maladies infectieuses. Finalement, personne ne fut accusé de ces attaques par le FBI, mais un chercheur soupçonné s'est suicidé et un autre a été blanchi. Après cet incident, les centres de contrôle et de prévention des maladies ont classé les bactéries et virus susceptibles d'être des cibles pour le bioterrorisme (tableau 18.2).

La connaissance du génome humain est aussi une source d'informations pour l'étude génétique des maladies complexes. Ces recherches au niveau du génome examinent les relations existant entre marqueurs génétiques, comme les SNP, et certaines caractéristiques ou maladies. Les recherches en cours concernent des caractères tels que la taille et pratiquement toutes les maladies à connotation génétique. De cette façon, on a identifié un grand nombre de gènes, mais pas encore toute la diversité génétique existante.

La génomique soulève des questions sociales et éthiques

La génomique permet d'élucider la relation entre notre génotype et notre phénotype et d'améliorer le diagnostic et le traitement de certaines maladies. L'utilisation de ces données implique des responsabilités sociales et éthiques. Il faut bien tenir compte des conséquences sociales et éthiques concernant la propriété des données génétiques et leur utilisation, de l'éventuel mauvais usage de ces données et même de l'opportunité d'effectuer toute expérience génomique.

Reconstruction du virus de la grippe de 1918

En mai 2005 furent publiés une série d'articles décrivant le séquençage et le réassemblage du virus de la grippe espagnole de 1918. Ce virus avait tué entre 20 et 50 millions de personnes. Au cours des années 1990, plusieurs groupes de chercheurs, y compris dans les centres de contrôle des maladies à Atlanta et à l'institut de pathologie des forces armées à Washington, ont découvert des fragments du génome viral, les ont séquencés et ont reconstruit le virus complet. On a constaté que ce virus avait le même niveau de mortalité que le virus d'origine dans des tests sur d'autres mammifères, comme les souris. On ne l'a évidemment pas testé sur des humains.

Ces expériences ont soulevé des questions sérieuses chez des scientifiques, des politiciens, des associations citoyennes et auprès du public en général. D'un côté, pouvoir étudier le génome du virus et les propriétés du virus lui-même, permettrait de produire des vaccins plus efficaces et de prévenir une nouvelle pandémie. Les adversaires de ces expériences considèrent que les vagues avantages potentiels de la reconstruction du virus ne compensent pas le risque de voir un scientifique malheureux ou négligent libérer le virus accidentellement ou délibérément. Beaucoup se posent aussi des questions sur l'utilisation des données par des terroristes qui pourraient créer des formes de virus utilisables comme armes biologiques.

Après une vaste enquête effectuée par la communauté scientifique, une recherche minutieuse des centres de contrôle des maladies et sous la supervision du National Science Advisory Board for Biosecurity, il fut décidé que les avantages l'emportaient sur les risques et que ce travail pouvait être accessible au public. De telles décisions sont évidemment subjectives, mais avec une bonne information sur la génomique, il

TABLEAU 18.2 Pathogènes prioritaires pour la recherche génomique

Pathogène	Maladie	Génome*
Virus variolique	Variole	Complet
Bacillus anthracis	Charbon	Complet
Yersinia pestis	Peste	Complet
Clostridium botulinum	Botulisme	En cours
Francisella tularensis	Tularémie	Complet
Filovirus	Ébola et fièvre hémorragique de Marburg	Complets tous les deux
Arénavirus	Fièvre de Lassa et fièvre hémorragique d'Argentine	Complets tous les deux

* Il existe de nombreuses souches de ces virus et bactéries. « Complet » signifie que l'une au moins a été séquencée. Par exemple, la souche Florida de l'anthrax fut la première à être séquencée.

est possible de prendre des décisions plus raisonnées dans un domaine aussi controversé.

Propriété et brevetage des gènes

Vous pensez peut-être que vous êtes le propriétaire des données concernant vos gènes mais, jusqu'en 2013, il semblait que quelqu'un d'autre pouvait posséder des droits sur les séquences de vos gènes. Dans les années 1990, une société, Myriad Genetics, breveta deux gènes que l'on savait impliqués dans le développement de certaines formes de cancer du sein. En brevetant les gènes *BRCA1* et *BRCA2*, la société s'assurait le droit d'utiliser les séquences pour tester la présence de formes particulières de ces gènes. Ce fut à l'origine d'un test permettant de déceler une prédisposition à certains types de cancer du sein. Étant donné les risques plus élevés de cancer du sein chez les femmes porteuses de formes particulières de *BRCA1* et *BRCA2* et l'importance du diagnostic, beaucoup ont considéré comme une atteinte aux libertés civiles le fait de ne pas pouvoir garantir un second diagnostic.

Après une longue bataille devant les tribunaux fédéraux, menée principalement par l'American Civil Liberty Union (ACLU), une plainte fut déposée contre Myriad Genetics devant la cour suprême des Etats-Unis. À l'unanimité, la cour statua que Myriad Genetics n'avait pas inventé les gènes et ne pouvait donc pas breveter leurs séquences. Dans son jugement, la cour notait que certaines séquences dérivées – par exemple les séquences d'ADNc – et leurs applications, pouvaient être brevetées parce qu'elles avaient été créées.

? Question Quels arguments pourrait-on avancer pour appuyer une demande de brevet pour un gène synthétique qui augmente la production de lipides dans des algues utilisées pour l'obtention de biocarburant ?

Questions d'apprentissage 18.6

Le séquençage du génome, son annotation et la génomique comparative et fonctionnelle ont abouti à la découverte de gènes impliqués dans les maladies. On peut utiliser certains de ces gènes pour améliorer le diagnostic ou préciser les pronostics. On peut aussi utiliser la génomique pour étudier le déclenchement d'épidémies. La génomique implique des responsabilités sociales et éthiques. En consultant les spécialistes en bioéthique, la police et diverses associations intéressées, les instances de régulations peuvent mieux s'informer sur les expérimentations et l'utilisation adéquate des données.

- *Une cellule obtenue en remplaçant son génome par une séquence construite par l'homme représente-t-elle une nouvelle forme de vie ?*

Résumé

18.1 Les cartes des génomes

Plusieurs méthodes permettent d'analyser des génomes entiers.
La génomique utilise des cartes génétiques et physiques pour localiser des repères génétiques, ou marqueurs, dans le génome.

Les cartes génétiques donnent les distances relatives entre les marqueurs génétiques.

Les cartes de liaisons sont utilisées pour déterminer la position relative des marqueurs génétiques dans le génome. Elles reposent sur la séparation des marqueurs lors de la recombinaison ; plus les marqueurs sont éloignés, plus probable est leur séparation.

Les cartes physiques donnent les distances exactes entre marqueurs génétiques.

On peut obtenir des cartes de restriction, des cartes chromosomiques et des cartes de STS pour localiser effectivement les marqueurs dans la séquence du génome (figures 18.1-18.3). La création des cartes physiques implique une certaine connaissance de la séquence d'ADN du génome.

Il est possible de mettre en relation les cartes physiques et les cartes génétiques.

On peut mettre en parallèle les cartes physiques et génétiques. Tout gène susceptible d'être cloné peut être placé dans la séquence du génome et localisé. On ne peut cependant pas faire correspondre exactement les distances.

18.2 Le séquençage des génomes

Le séquençage du génome avec terminateur didésoxy reste la méthode la plus importante.
Malgré les énormes progrès engrangés par les technologies de séquençage au cours des dernières années, on utilise encore le séquençage avec terminateur didésoxy dans certains cas. Cette technique repose sur une modification chimique des nucléotides afin de clôturer la synthèse de l'ADN par un nucléotide fluorescent (figure 18.4).

Le séquençage de dernière génération fait massivement appel à des technologies parallèles pour accroître la vitesse.

Le coût et la durée de séquençage des génomes entiers ont été réduits d'environ 10 000 fois depuis la mise en route du Projet Génome Humain. Les séquenceurs de dernière génération n'ont plus besoin de cloner les morceaux d'ADN et l'ADN isolé est analysé dans des milliers de réactions simultanées pour aboutir à une masse de données (figure 18.5).

Les fragments séquencés sont assemblés en séquences complètes.

Les méthodes shotgun de séquençage et assemblage recréent la séquence d'ADN du génome à partir de très petits morceaux d'ADN chevauchants. Les méthodes clone-contig se servent de morceaux d'ADN plus longs et d'une approche consistant à construire des fragments chevauchants de plus en plus longs jusqu'au réassemblage de tout le génome (figure 18.6).

18.3 Les programmes génomiques

Le Projet Génome Humain a séquencé et cartographié la plus grande partie du génome humain.

Une compétition pour le séquençage et l'assemblage du génome humain a permis des progrès technologiques rapides et la clôture du projet en avance sur les délais. Le génome humain contient nettement moins de gènes que prévu (figure 18.7).

Le Projet Génome du Blé illustre les difficultés d'assemblage des génomes complexes.

Le séquençage et l'assemblage des génomes longs et très répétitifs sont plus coûteux et prennent plus de temps. Les structures génomiques complexes impliquent la technique clone-contig.

Les projets sur le génome du cancer sont à la recherche des causes génétiques du cancer.

En identifiant les ressemblances et les différences dans les génomes et épigénomes des mêmes cancers on peut améliorer le diagnostic, prédire les rechutes et administrer des médicaments mieux adaptés.

18.4 Annotation des génomes et banques de données

L'annotation du génome assigne une fonction aux séquences d'ADN.

Les gènes sont identifiés par la recherche de cadres ouverts de lecture dans les séquences d'ADN. On peut parfois utiliser les outils servant à la recherche de séquences correspondant à un gène potentiel pour en déduire la fonction du gène. L'identification d'un EST peut faciliter l'annotation en apportant une information supplémentaire (figure 18.8).

Les génomes contiennent de l'ADN codant et non codant.

Dans l'ADN codant des protéines, on trouve des gènes à copie unique, des duplications, des familles multigéniques et des groupes en tandem. L'ADN non codant des eucaryotes représente jusqu'à 99 % du total. Approximativement 45 % du génome humain se compose d'éléments transposables, comme les LINE, SINE et LTR (tableau 18.1).

L'objectif du programme ENCODE est l'identification de tous les éléments fonctionnels du génome humain.

Le consortium ENCODE estime que 20 à 80 % du génome humain est fonctionnel. On peut se poser la question de l'utilité de définir la fonction sans intégrer les contraintes évolutives intervenant dans la fonction du génome (figure 18.9).

18.5 La génomique comparative et fonctionnelle

La génomique comparative révèle des régions conservées dans les génomes.

Plus de la moitié des gènes de la drosophile ont leur équivalent chez les humains. Les éléments transposables constituent la plus grande différence entre notre génome et celui des chimpanzés.

La génomique fonctionnelle permet de connaître la fonction des gènes au niveau du génome.

La génomique fonctionnelle a besoin de techniques expérimentales à haut débit et de la bioinformatique pour analyser la fonction des gènes et de leurs produits. Les micro-alignements d'ADN permettent de contrôler en même temps l'expression de tous les gènes d'une cellule (figure 18.11). SAGE et RNA-seq sont des techniques plus globales pour l'étude du transcriptome

La protéomique établit un catalogue des protéines codées par le génome.

La protéomique caractérise toues les protéines produites par une cellule. Elle est compliquée à cause de la nature dynamique du protéome, des modifications survenant après la traduction et du rôle de l'épissage alternatif responsable de la production de protéines différentes à partir d'un même gène (figures 18.12-18.14).

18.6 Applications de la génomique

On a utilisé les données génomiques pour créer la première cellule « synthétique ».

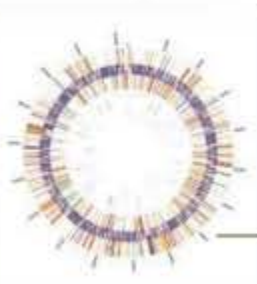
On a créé une cellule se répliquant de façon autonome en planifiant un génome sur ordinateur, en synthétisant chimiquement l'ADN du génome, en assemblant ce génome dans la levure et en l'introduisant dans une bactérie (figure 18.15).

La génomique peut aider à identifier et traiter les maladies.

L'identification de certains gènes dans des programmes génomiques peut faciliter et améliorer le diagnostic de certaines maladies. À l'avenir, une médecine entièrement personnalisée peut signifier des traitements taillés sur mesure en fonction du génotype de l'individu.

La génomique soulève des questions sociales et éthiques.

Des groupes sociaux et bioéthiques contrôlent soigneusement les expériences, telles que celles qui impliquent la création d'une cellule de synthèse et la reconstitution de virus dangereux. On ne peut pas breveter des gènes individuels, mais on peut le faire pour des produits de synthèse, comme un ADNc correspondant au gène.



Questions

COMPRÉHENSION

1. Une carte génétique donne
   a. la séquence de l'ADN du génome
   b. La position relative des gènes sur les chromosomes.
   c. la localisation des sites de coupure par les enzymes de restriction dans une séquence d'ADN connue.
   d. la répartition des bandes sur un chromosome.
2. Qu'est-ce qu'un STS,
   a. Une séquence unique dans l'ADN qui peut servir à construire une carte.
   b. Une séquence répétée de l'ADN qui peut servir à construire une carte.
   c. Un élément amont permettant de placer sur la carte la région 3' d'un gène.
   d. Une séquence d'ADN qui chevauche d'autres fragments séquencés lors de l'assemblage du génome par clone-contig.

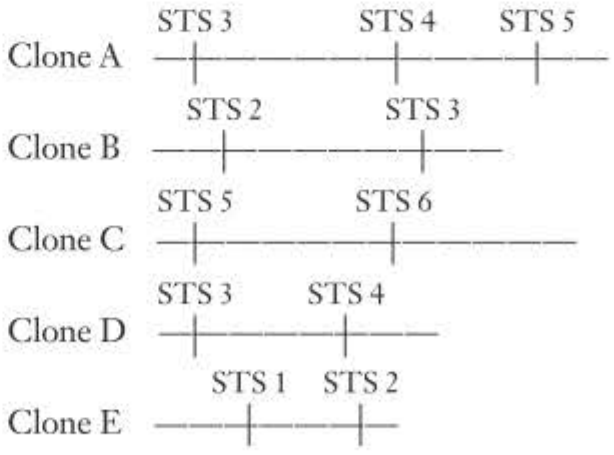
3. Quel est le nombre approximatif de gènes dans le génome humain ?
   a. 2500
   b. 10 000
   c. 20 000
   d. 100 000
4. Un cadre ouvert de lecture (ORF) se distingue par la présence
   a. d'un codon stop.
   b. d'un codon de départ.
   c. d'une séquence d'ADN assez longue pour coder une protéine.
   d. Tout ces choix sont corrects.
5. Qu'est-ce qu'une recherche BLAST ?
   a. Un mécanisme d'alignement des régions consensus au cours du séquençage d'un génome entier.
   b. La recherche de séquences géniques semblables dans d'autres espèces.
   c. Une méthode de criblage d'une banque d'ADN.
   d. Un moyen d'identifier les ORF.
6. Pourquoi l'ADN répétitif est-il un problème pour le séquençage shotgun et l'assemblage du génome ?
   a. L'ADN répétitif n'est pas facile à séquencer.
   b. Les fragments séquencés ne sont pas uniques.
   c. L'ADN répétitif contient trop de GC.
   d. Il est trop coûteux et trop laborieux.
7. Qu'est-ce qu'un protéome ?
   a. L'ensemble de tous les gènes codant des protéines.
   b. L'ensemble de toutes les protéines codées par le génome.
   c. L'ensemble de toutes les protéines présentes dans une cellule.
   d. La séquence des acides aminés d'une protéine.
8. Pourquoi ne peut-on pas utiliser le transcriptome pour prédire le protéome avec une exactitude totale ?
   a. Il ne peut être séquencé comme le génome.
   b. Le transcriptome est trop dynamique pour permettre des prévisions.
   c. Tous les gènes ne sont pas transcrits.
   d. Beaucoup de transcrits sont soumis à l'épissage alternatif et donnent des protéines différentes.

APPLICATION

1. Si, par la génomique, on montre la présence de la même mutation chez tous les malades du cancer qui n'ont pas répondu à un traitement particulier, que peut-on en conclure ?
   a. Que la mutation était responsable de l'inefficacité du traitement.
   b. Que la mutation serait un bon outil de diagnostic.
   c. Que les enfants de ces patients courent un risque accru d'être atteints par ce cancer ?
   d. Que les patients ont tous le même type de cancer.
2. ENCODE définit la fonction sur une base chimique. Laquelle des approches génomiques suivantes pourrait améliorer l'exactitude de l'estimation de la fonctionnalité du génome chez les humains ?
   a. La génomique fonctionnelle.
   b. La génomique comparative.
   c. Le séquençage et l'assemblage du génome.
   d. Les choix a, b et c pourraient améliorer l'exactitude de l'estimation.
3. Si j'avais synthétisé une protéine en unissant chimiquement les acides aminés et ajouté quelques acides aminés supplémentaires pour mieux stabiliser la protéine et donc l'améliorer pour le traitement d'une maladie, quel argument pourrait faciliter l'obtention d'un brevet pour cette protéine ?
   a. La protéine est utile et je devrais donc être payé pour ma découverte.
   b. La protéine est synthétique et différente de la forme naturelle.
   c. La protéine a été synthétisée par voie chimique.
   d. La production de la protéine est coûteuse et le brevet m'aiderait à couvrir les frais.
4. Le génome de *Xenopus tropicalis* est tétraploïde. Quelles difficultés cela pourrait-il entraîner pour le séquençage et l'assemblage du génome ?
   a. L'assemblage du génome peut être difficile à cause de la présence de portions dupliquées.
   b. Le séquençage difficile à cause des LINE.
   c. On ne peut pas établir des cartes génétiques et physiques.
   d. b et c sont corrects.
5. Quelle information peut-on attendre d'un micro-alignement d'ADN ?
   a. La séquence d'un gène particulier.
   b. La présence de gènes dans un tissu particulier.
   c. Le mode d'expression des gènes.
   d. Les différences entre génomes
6. Des propositions suivantes concernant la technologie des micro-alignements et le cancer, laquelle est correcte ?
   a. Un micro-alignement d'ADN peut déterminer le type de cancer.
   b. Un micro-alignement d'ADN peut mesurer la réponse d'un cancer au traitement.
   c. Un micro-alignement d'ADN peut servir à prévoir si un cancer produira des métastases.
   d. Tous ces choix sont corrects.
7. La protéomique comparative implique la comparaison des protéomes de deux cellules ou tissus différents ou dans des conditions différentes. Dans certaines cellules cancéreuses, la protéine Retinoblastoma paraît souvent avoir une masse moléculaire plus grande que dans les cellules non cancéreuses, bien que les séquences d'acides aminés soient identiques dans les deux cas. Comment pourrait-on expliquer cette différence ?
   a. Une modification après la traduction.
   b. L'épissage alternatif.
   c. L'association à d'autres protéines.
   d. La mutation du gène codant la protéine.

RÉVISION

1. Vous entamez le séquençage d'un génome. Vous avez isolé des clones d'une banque de chromosomes artificiels de bactéries (BAC) et cartographié les inserts de ces clones par les STS. Utilisez les STS pour aligner les clones dans une séquence contiguë du génome (un contig).



2. On peut utiliser la génomique pour savoir si une épidémie est naturelle ou "intentionnelle". Expliquez ce qu'un chercheur en génomique devrait étudier si l'on suspectait une épidémie intentionnelle d'une maladie comme le charbon ?