Les génomes et leur évolution

21

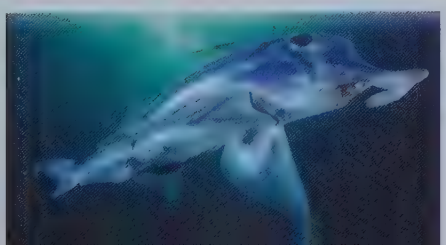
VOS OUTILS INTERACTIFS

Consultez votre MANUEL NUMÉRIQUE, qui vous donne accès aux **animations**, aux **exercices** et à la plateforme d'**anatomie interactive**.

▲ **Figure 21.1** **Quelles différences génomiques distinguent l'humain du chimpanzé ?**

CONCEPTS CLÉS

- **21.1** Le projet Génome humain a favorisé la mise au point de techniques de séquençage plus rapides et moins onéreuses
- **21.2** Les scientifiques utilisent la bio-informatique pour analyser les génomes et leurs fonctions
- **21.3** Les génomes varient en taille, en nombre de gènes et en densité génique
- **21.4** Les eucaryotes multicellulaires possèdent beaucoup d'ADN non codant et de nombreuses familles multigéniques
- **21.5** Les duplications, les réarrangements et les mutations de l'ADN contribuent à l'évolution du génome
- **21.6** La comparaison des séquences génomiques fournit des indices sur l'évolution et le développement



◀ **Chimère éléphant (*Callorhinchus milii*).**

Lire dans les feuilles de l'arbre de la vie

Dans l'arbre de la vie, le chimpanzé (*Pan troglodytes*) est notre plus proche parent vivant. Le garçon de la **figure 21.1** et son compagnon le chimpanzé observent attentivement la même feuille, mais un seul des deux est en mesure d'en parler. Comment expliquer cette différence entre deux primates qui ont en commun une si grande part de leur histoire évolutive ? Grâce aux avancées réalisées dans la technologie du séquençage, nous pouvons maintenant répondre à des questions aussi fascinantes que celle-ci sous l'angle de leurs fondements génétiques. Plus loin dans ce chapitre, il sera question du gène *FOXP2* qui intervient dans la vocalisation et qui diffère entre les deux espèces.

Le génome du chimpanzé a été séquencé deux ans après la publication de la séquence complétée du génome humain. Maintenant qu'il est possible de comparer base par base notre génome avec celui du chimpanzé, on peut s'attaquer à cette question : quelles différences dans l'information génétique rendent compte des caractéristiques distinctes de ces deux espèces de primates ?

En plus d'avoir déterminé les séquences des génomes de l'humain et du chimpanzé, les chercheurs ont décodé les séquences génomiques complètes d'*Escherichia coli* (*E. coli*) et celles d'un grand nombre d'autres procaryotes, ainsi que de nombreux eucaryotes, dont *Zea mays* (maïs), *Drosophila melanogaster* (mouche du vinaigre, ou drosophile), *Octopus bimaculoides* (pieuvre à deux points de Californie) et *Callorhinchus milii* (chimère éléphant ; voir la photo à gauche). En 2014, on a publié une séquence de haute qualité du génome d'*Homo neanderthalensis* (néandertalien), une espèce disparue étroitement liée aux humains modernes. Déjà intéressants en eux-mêmes, ces génomes nous fournissent de plus des renseignements précieux sur l'évolution

et sur d'autres processus biologiques. En étendant la comparaison humain-chimpanzé aux génomes d'autres primates et d'animaux plus éloignés, on arrivera sûrement à connaître les séries de gènes responsables des caractéristiques qui définissent un groupe donné. Au-delà de cet exercice, les comparaisons avec les génomes des bactéries, des archées, des eumycètes, des protistes et des végétaux devraient nous éclairer sur la longue histoire évolutive de gènes anciens qui nous sont tous communs.

Maintenant que les séquences de génomes entiers sont connues, les scientifiques peuvent étudier des ensembles complets de gènes et leurs interactions grâce à la **génomique**. Les travaux de séquençage qui alimentent cette approche ont généré d'énormes volumes de données, et ils continuent aujourd'hui sur cette lancée. Le besoin de traiter ce déluge d'information toujours croissant a donné le jour à la **bio-informatique**, un domaine de l'informatique qui met ses méthodes de calcul au service de l'organisation et de l'analyse de données biologiques.

Nous commencerons le présent chapitre en examinant deux approches de séquençage du génome et certains progrès accomplis en bio-informatique et ses applications. Nous résumerons ensuite ce que nous ont appris les génomes séquencés jusqu'à maintenant, puis nous décrirons la composition du génome humain comme un génome représentatif d'un eucaryote multicellulaire complexe. Enfin, nous étudierons les hypothèses actuelles qui nous aident à saisir comment sont apparus les génomes et comment l'évolution des mécanismes de développement a pu engendrer la grande diversité de la vie sur Terre aujourd'hui.

CONCEPT **21.1**

Le projet Génome humain a favorisé la mise au point de techniques de séquençage plus rapides et moins onéreuses

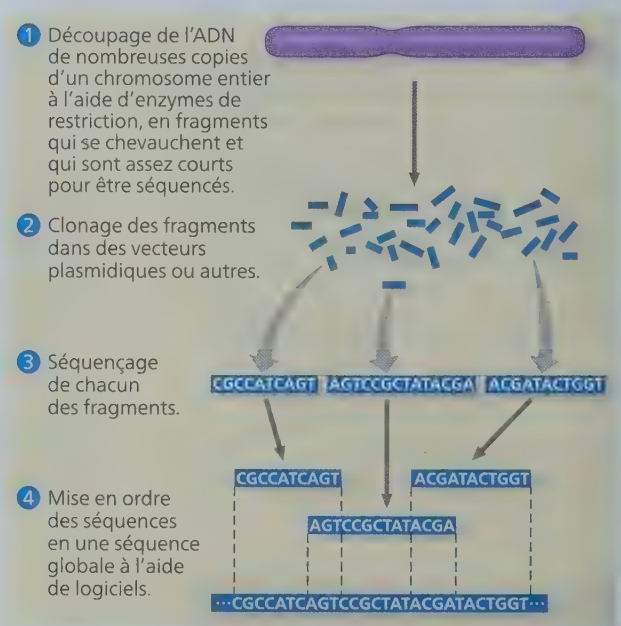
Le séquençage du génome humain, connu sous le nom de **projet Génome humain**, est un ambitieux projet de recherche. Il a été lancé officiellement en 1990 sous l'égide d'un consortium international financé par le secteur public et réunissant des scientifiques œuvrant dans des universités et des instituts de recherche. Ce projet a entraîné la création de 20 grands centres de séquençage répartis dans 6 pays, en plus d'une quantité d'autres laboratoires travaillant sur de petits projets.

Alors que le séquençage du génome humain était presque terminé, en 2003, la séquence de chaque chromosome a été analysée et décrite dans une série de publications, dont la dernière, parue en 2006, portait sur le chromosome 1. À ce stade, le séquençage était considéré comme « pratiquement terminé ».

Le but ultime de la cartographie d'un génome est de déterminer la séquence nucléotidique complète de chaque chromosome. Pour le génome humain, cette étape a été réalisée grâce à des séquenceurs, à l'aide de la méthode de terminaison de chaîne par un didésoxyribonucléotide (méthode didésoxy, en abrégé) décrite au concept 20.1. Malgré l'avènement de l'automatisation, le séquençage des 3 milliards de paires de bases d'un jeu haploïde de chromosomes humains représentait toujours une tâche monumentale. Parmi les événements qui ont eu un effet décisif sur le projet Génome humain, la mise au point d'une technique de séquençage plus rapide constitue un fait marquant (voir le concept 20.1). Au cours des années, les améliorations apportées ont réduit la longueur de chaque étape, permettant d'accélérer la vitesse de séquençage de façon impressionnante. Alors qu'au cours des années 1980 un laboratoire productif arrivait à séquencer quotidiennement 1 000 paires de bases, en l'an 2000 chaque centre de recherche qui travaillait au projet Génome humain séquençait 1 000 paires de bases *à la seconde*. Depuis 2016, les appareils automatisés les plus couramment utilisés peuvent séquencer près de 25 millions de paires de bases par seconde. Toutefois, les concepteurs de certaines des techniques encore plus récentes prétendent qu'il est possible de séquencer 66 milliards de paires de bases par seconde. Les méthodes qui peuvent analyser du matériel biologique très rapidement et produire d'énormes volumes de données sont dites « à haut débit ». Les séquenceurs sont un exemple d'appareils à haut débit.

On a recours à deux approches complémentaires pour produire une séquence complète. L'approche initiale est une technique méthodique qui repose sur l'utilisation d'une source antérieure de données sur la génétique humaine. En 1998, cependant, le biologiste moléculaire J. Craig Venter fonde une société (Celera Genomics) et déclare qu'il a pour objectif d'établir le séquençage complet du génome humain à l'aide d'une autre technique appelée **séquençage en aveugle sur l'ensemble du génome**. Cette technique débute par le clonage et le séquençage de fragments d'ADN pris au hasard. Ensuite, de puissants programmes informatiques analysent les très nombreuses courtes séquences déchiffrées qui, en se recouvrant partiellement, reconstituent la séquence complète (**figure 21.2**).

▼ **Figure 21.2 Le séquençage en aveugle sur l'ensemble du génome.** La technique conçue par J. Craig Venter et ses collègues de Celera Genomics consiste à cloner (voir la figure 20.4) et à séquencer des fragments d'ADN aléatoires, puis à les classer les uns par rapport aux autres.



HABILETÉS VISUELLES ▶ Plutôt que d'être placés de façon ordonnée, les fragments à l'étape 2 semblent éparpillés. Comment cette représentation correspond-elle à l'approche utilisée ?

Aujourd'hui, l'approche de séquençage en aveugle sur l'ensemble du génome est toujours utilisée, même si des techniques de séquençage plus récentes (voir la figure 20.3) donnent des résultats plus rapidement et à moindre coût. Dans ces nouvelles techniques, un grand nombre de fragments d'ADN plus petits (d'une longueur d'environ 300 paires de bases) sont séquencés simultanément, et des logiciels assemblent rapidement la séquence complète. En raison de la précision de ces méthodes, il est possible de séquencer les fragments directement, c'est-à-dire sans passer par l'étape du clonage (étape 2 de la figure 21.2). Alors qu'il a fallu 13 ans pour séquencer le premier génome humain, au coût de 100 millions de dollars, il n'a fallu que 4 mois en 2007 pour séquencer celui de James Watson (codécouvreur de la structure de l'ADN) à l'aide de nouvelles techniques, et l'opération n'a coûté qu'environ 1 million de dollars. Depuis 2016, il est possible de séquencer le génome d'une personne en un seul jour au coût d'environ 1 000 $ US.

Ces progrès techniques ont également facilité la mise en œuvre de la **métagénomique** (du grec *meta*, « ce qui dépasse »), une approche dans laquelle l'ADN d'une communauté entière d'espèces (un *métagénome*) est extrait d'un échantillon prélevé dans l'environnement et séquencé. Là aussi, un logiciel se charge de trier les séquences fragmentaires et de les assembler en génomes individuels spécifiques. Cette technique a l'avantage de permettre le séquençage de l'ADN de populations microbiennes disparates, ce qui élimine la nécessité de cultiver chaque espèce séparément en laboratoire, une difficulté qui a limité l'étude de nombreuses espèces microbiennes. Jusqu'à maintenant, les scientifiques ont appliqué cette démarche pour analyser le génome des communautés microbiennes présentes dans des environnements aussi divers que l'intestin humain et les sols anciens de l'Arctique. En effet, dans une étude menée 2014, on a pu caractériser des douzaines d'espèces de la région de l'Arctique qui formaient depuis 50 000 ans une communauté dans laquelle vivaient des animaux, des plantes et des microorganismes.

À première vue, les séquences génomiques des humains et d'autres organismes ne sont que des listes monotones de bases nucléotidiques (une succession interminable de millions de A, de T, de C et de G) auxquelles il est difficile de donner un sens. Pour décrypter cette quantité phénoménale de données, il a donc fallu mettre au point de nouvelles méthodes d'analyse, que nous décrivons dans la prochaine section.

RETOUR SUR LE CONCEPT 21.1

1. Décrivez l'approche de séquençage en aveugle sur l'ensemble du génome.

Voir les réponses proposées à l'appendice A.

CONCEPT 21.2

Les scientifiques utilisent la bio-informatique pour analyser les génomes et leurs fonctions

Jour après jour, chacun des quelque 20 centres de séquençage travaillant au projet Génome humain a produit quotidiennement un nombre considérable de séquences d'ADN. Devant cette accumulation de données, il est devenu rapidement nécessaire de coordonner les travaux afin d'être en mesure de suivre toutes les séquences. Aussi, les chercheurs et les responsables gouvernementaux engagés dans le projet Génome humain se sont-ils donné pour objectif d'établir des bases de données centralisées, de perfectionner les logiciels d'analyse et de rendre toutes ces ressources facilement accessibles sur internet.

La centralisation des ressources pour l'analyse des séquences génomiques

L'accès aux ressources bio-informatiques et le partage plus rapide des données ont permis aux chercheurs du monde entier d'accomplir de grands progrès dans l'analyse des séquences d'ADN. Par exemple, aux États-Unis, le National Center for Biotechnology Information (NCBI) a vu le jour en 1988 grâce aux activités concertées de la National Library of Medicine (NLM) et des National Institutes of Health (NIH) pour mettre sur pied le projet Génome humain. Le NCBI héberge aujourd'hui un site internet (www.ncbi.nlm.nih.gov) qui comporte d'importantes ressources bio-informatiques et qui propose des liens vers des bases de données, des logiciels et quantité d'informations sur la génomique et des sujets connexes. Par ailleurs, trois centres génomiques en relation avec le NCBI ont créé des sites semblables : le Laboratoire européen de biologie moléculaire, la Banque de données génétiques du Japon et le BGI (autrefois nommé Beijing Genome Institute) à Shenzhen, en Chine. D'autres sites internet hébergés par des individus ou des petits groupes de laboratoires s'ajoutent à ces sites d'envergure et complets. Des sites plus petits fournissent souvent des bases de données et des logiciels conçus à des fins plus circonscrites, comme l'étude des modifications génétiques et génomiques correspondant à un type particulier de cancer.

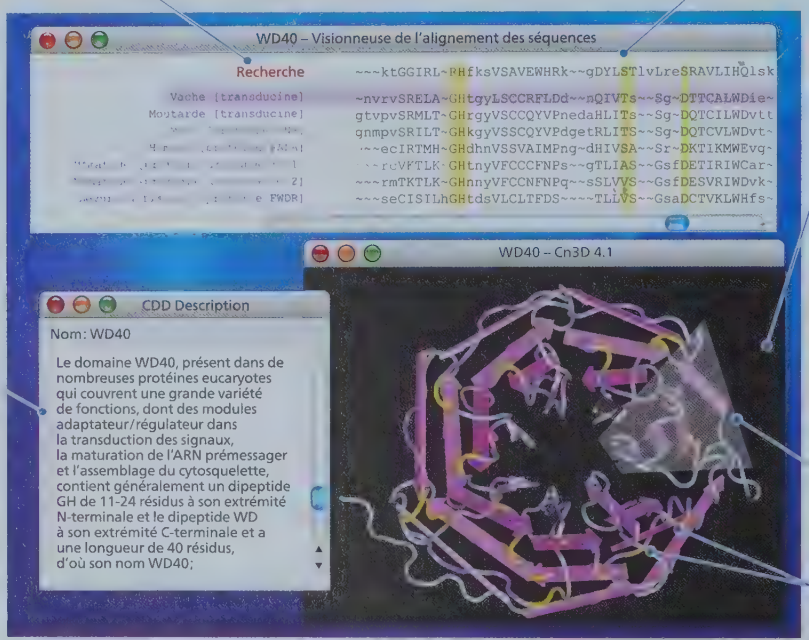
La base de données des séquences du NCBI est nommée GenBank. En juin 2016, elle comprenait les séquences de 194 millions de fragments d'ADN génomique, pour un total de 213 milliards de paires de bases ! GenBank est constamment mise à jour, et la quantité de données qu'elle contient augmente rapidement. Toute séquence dans la banque peut être extraite et analysée à l'aide de logiciels disponibles sur le site internet du NCBI ou ailleurs.

BLAST, l'un des programmes informatiques les plus largement utilisés qui est accessible sur le site du NCBI, permet à l'utilisateur de comparer une séquence d'ADN avec chaque séquence dans la GenBank, base par base. Un chercheur peut donc repérer des régions similaires dans d'autres gènes d'une même espèce ou parmi les gènes d'autres espèces. Un autre logiciel permet de comparer des séquences prédites de protéines. Et un troisième peut chercher n'importe quelle séquence de polypeptides afin de trouver des segments *conservés* (communs) d'acides aminés (domaines) dont la fonction est connue ou présumée. De plus, ce logiciel peut afficher un modèle tridimensionnel du domaine en question ainsi que d'autres renseignements pertinents (**figure 21.3**). Il existe même un programme informatique capable d'aligner et de comparer une collection de séquences, soit d'acides nucléiques, soit de polypeptides, et de les schématiser sous la forme d'un arbre évolutif basé sur les relations entre les séquences. (La figure 21.17 présente un tel schéma.)

▼ Figure 21.3 Les outils bio-informatiques accessibles dans internet. Un site internet administré par le National Center for Biotechnology Information (NCBI) permet aux scientifiques et au public d'accéder aux séquences d'ADN et de protéines et à d'autres données stockées. Le site comprend un lien vers une base de données sur les structures de protéines (Conserved Domain Database, CDD) qui peut trouver et décrire des domaines similaires dans des protéines apparentées; il comprend également un logiciel (Cn3D; en anglais, *See in 3D*) qui présente des modèles de domaines. Cette figure montre certains résultats obtenus lors de la recherche de régions de protéines comparables à une séquence d'acides aminés présente dans une protéine de melon brodé, *Cucumis melo* var. *reticulatus* (plus connu sous le nom de cantaloup). Le domaine WD40 est fréquent dans les protéines codées par des génomes eucaryotes. Il joue souvent un rôle crucial dans les interactions moléculaires pendant la transduction des signaux.

1 Dans cette fenêtre, une séquence partielle d'acides aminés provenant d'une protéine inconnue de melon brodé («Recherche», «*Query*» en anglais) est mise en correspondance avec les séquences similaires d'autres protéines trouvées par le programme informatique. Chaque séquence représente un domaine nommé WD40.

2 Quatre empreintes caractéristiques du domaine WD40 sont mises en évidence en jaune. (La similarité des séquences est basée sur les aspects chimiques des acides aminés, de sorte que les acides aminés dans la région des empreintes caractéristiques ne sont pas toujours identiques.)



3 Le programme Cn3D présente un modèle en ruban tridimensionnel de la transducine de vache (la protéine mise en évidence en violet dans la fenêtre Visionneuse de l'alignement des séquences, *Sequence Alignment Viewer* en anglais). Cette protéine est la seule parmi celles qui sont illustrées dont la structure a été déterminée. La similarité des séquences des autres protéines et de celle de la transducine de vache semble indiquer que leurs structures sont probablement semblables.

4 La transducine de vache contient sept domaines WD40, dont l'un est mis en évidence en gris.

5 Les segments en jaune correspondent aux motifs caractéristiques de WD40 mis en évidence en jaune dans la fenêtre au-dessus.

6 Cette fenêtre présente des informations au sujet du domaine WD40 provenant de la *Conserved Domain Database*.

Deux instituts de recherche, la Rutgers University et la University of California à San Diego, hébergent pour leur part une base de données renfermant toutes les structures tridimensionnelles des protéines qui ont été déterminées, soit la Worldwide Protein Data Bank, que l'on peut consulter à l'adresse www.wwpdb.org. Il est possible de faire pivoter les structures afin d'observer une protéine sous tous les angles. Tout au long de ce manuel, vous trouverez des images de structures protéiques tirées de cette base de données.

Il existe une foule de ressources accessibles gratuitement aux chercheurs dans le monde entier. Penchons-nous maintenant sur les types de questions que les scientifiques peuvent essayer de résoudre à l'aide de ces ressources.

L'identification des gènes codant pour des protéines et la compréhension de leurs fonctions

À l'aide des séquences d'ADN disponibles, les généticiens peuvent étudier les gènes directement sans recourir à l'approche génétique classique qui exige d'établir la fonction d'un gène inconnu à partir du phénotype. Toutefois, cette approche plus récente nous confronte à un nouveau défi : quelle est la véritable fonction du gène ? À partir d'une longue séquence d'ADN fournie par une base de données comme GenBank, l'objectif des scientifiques consiste à identifier tous les gènes qui codent pour des protéines dans la séquence et, finalement, à déterminer leurs fonctions. Ce processus, nommé **annotation d'un gène**, identifie un gène en se fondant sur trois éléments de preuve.

D'abord, il faut rechercher certaines séquences indiquant la présence des gènes à l'aide d'un ordinateur. La démarche habituelle consiste à se servir de logiciels pour parcourir les séquences stockées et rechercher celles qui sont associées aux signaux de départ et d'arrêt de la transcription et de la traduction, ainsi qu'aux sites d'épissage de l'ARN ; il est également possible de trouver d'autres indices de la présence de gènes codant pour des protéines. Ces logiciels permettent aussi de repérer certaines courtes séquences qui codent pour des ARNm connus. Des milliers de séquences de ce type, nommées *étiquettes de séquences exprimées* (ou EST, pour *expressed sequences tags*), ont été obtenues à partir de séquences d'ADNc et sont actuellement répertoriées dans des bases de données informatisées. Ce type d'analyse

détermine les séquences qui pourraient correspondre à des gènes auparavant inconnus codant pour des protéines.

Même si on connaissait déjà près de la moitié des gènes humains avant même le début du projet Génome humain, ceux qui étaient auparavant inconnus ont pu être identifiés par une analyse des séquences d'ADN. Une fois ces gènes identifiés, la deuxième étape consiste à recueillir des indices sur leur identité et leur fonction en utilisant un logiciel pour comparer leur séquence nucléotidique à celle de gènes connus d'autres organismes. En raison des redondances du code génétique, la séquence de l'ADN elle-même peut varier davantage d'une espèce à l'autre que la séquence de la protéine. Par conséquent, les scientifiques qui analysent les protéines comparent souvent la séquence prédite des acides aminés d'une protéine avec celle d'autres protéines. Enfin, l'identité des gènes doit être confirmée par ARN-seq (voir la figure 20.13) ou par une autre méthode afin de démontrer que l'ARN ciblé est réellement exprimé par les gènes présumés.

Parfois, une séquence nouvellement décodée correspond, du moins en partie, à la séquence d'un gène ou d'une protéine d'une autre espèce dont la fonction est déjà bien connue. Par exemple, une chercheuse en phytologie travaillant sur les voies de signalisation dans le cantaloup (melon brodé) serait enthousiaste si elle constatait qu'une partie de la séquence d'acides aminés d'un gène qu'elle aurait identifié correspondait à celui du domaine WD40 présent chez d'autres espèces. Ce domaine est une partie fonctionnelle d'une protéine (voir la figure 21.3). De nombreuses protéines eucaryotes comportent des domaines WD40, et on sait qu'ils agissent dans les voies de transduction des signaux. Il est également possible que la séquence du nouveau gène soit semblable à une séquence déjà rencontrée, mais dont la fonction est encore inconnue. Il arrive aussi que la séquence soit entièrement différente de tout ce qui a été vu auparavant. C'est ce qui s'est avéré pour près du tiers des gènes d'*E. coli* lorsque son génome a été séquencé. Dans le dernier cas, on déduit habituellement la fonction de la protéine par le biais d'une série d'études biochimiques et fonctionnelles. Les analyses biochimiques visent à déterminer la structure tridimensionnelle de la protéine, de même que d'autres propriétés, comme les sites de liaison potentiels avec d'autres molécules. Les études fonctionnelles portent habituellement sur l'*inactivation* (le blocage ou l'inhibition) du gène d'un organisme, afin de déterminer comment le phénotype est affecté. Le système CRISPR-Cas9 décrit à la figure 20.14 est un exemple de technique expérimentale utilisée pour bloquer la fonction d'un gène.

La compréhension des gènes et de l'expression génétique au niveau des systèmes

La capacité de traitement impressionnante des outils de la bioinformatique permet l'étude de jeux complets de gènes et de leurs interactions, de même que la comparaison des génomes provenant d'espèces différentes. La génomique est une extraordinaire source de nouveaux éclairages sur des questions fondamentales concernant l'organisation du génome, la régulation de l'expression génétique, la croissance et le développement, et l'évolution.

Dans un projet de recherche nommé ENCODE (Encyclopedia of DNA Elements), réalisé entre 2003 et 2012, on a fait appel à une stratégie informationnelle. En effet, le projet avait pour but de recueillir un maximum de renseignements sur les éléments fonctionnels importants du génome humain en utilisant différentes techniques expérimentales sur divers types de cellules mises en culture. Les chercheurs cherchaient les gènes codant pour des protéines, les gènes pour les ARN non traduits ainsi que les séquences participant à la régulation de l'expression génétique (comme les amplificateurs et les promoteurs). Les chercheurs ont également caractérisé de façon exhaustive les modifications de l'ADN et des histones ainsi que la structure de la chromatine. Il s'agit en fait de caractéristiques épigénétiques puisqu'elles influent sur l'expression génétique sans toutefois modifier la séquence des bases nucléotidiques (voir le concept 18.3). La deuxième phase du projet, à laquelle ont participé plus de 440 scientifiques de 32 groupes de recherche, a atteint son apogée en 2012 avec la publication simultanée de 30 articles portant sur plus de 1 600 grands ensembles de données. Ce projet est d'une très grande portée puisqu'il permet de comparer les résultats de différents travaux spécifiques et, par conséquent, de dresser un portrait plus complet du génome entier.

La découverte la plus importante est assurément que plus de 75 % du génome est transcrit à un certain moment dans au moins un des types de cellules étudiés, même si moins de 2 % du génome code pour des protéines. De plus, on a pu associer des fonctions biochimiques à des éléments de l'ADN totalisant au moins 80 % du génome humain. À l'heure actuelle, on réalise en parallèle des projets visant à analyser de façon similaire les génomes de deux organismes modèles, soit *Caenorhabditis elegans* (nématode) et *Drosophila melanogaster* (drosophile) afin d'en apprendre davantage sur les différents types d'éléments fonctionnels. Puisqu'il est possible de réaliser des expériences de génétique et de biologie moléculaire sur ces espèces, les tests effectués sur les activités d'éléments d'ADN potentiellement fonctionnels dans leur génome révéleront beaucoup d'informations sur le mode de fonctionnement du génome humain.

Comme le projet ENCODE visait à analyser des cellules mises en culture, les possibilités d'application cliniques étaient limitées. Un projet connexe, soit le Roadmap Epigenomics Project, a été mis sur pied pour caractériser l'*épigénome* (caractéristiques épigénétiques du génome) de centaines de types cellulaires et de tissus humains. Le but était de se concentrer sur l'épigénome des cellules souches, des tissus normaux d'adultes matures et des tissus spécifiques prélevés chez des personnes atteintes de certaines maladies comme un cancer ou une maladie neurodégénérative ou auto-immune. En 2015, les résultats obtenus pour 111 tissus ont été publiés dans une série d'articles. La possibilité d'établir le foyer d'origine (tissu) du cancer à partir des cellules d'une tumeur secondaire en se fondant sur la caractérisation de l'épigénome des tissus figurait parmi les découvertes les plus utiles.

La biologie des systèmes

Les progrès scientifiques enregistrés dans le domaine du séquençage des génomes et de l'étude d'ensembles de gènes ont incité les scientifiques à se lancer dans l'étude systématique de jeux de protéines et de leurs propriétés (comme leur abondance, leurs modifications après la traduction et leurs interactions). Ce nouveau domaine de recherche porte le nom de **protéomique**. (Un protéome est l'ensemble des protéines exprimées par une cellule ou par un groupe de cellules.) Les protéines, et non les gènes qui les codent, sont les molécules qui assurent la plupart des diverses fonctions cellulaires. Il faut donc découvrir à quel moment et dans quels lieux elles sont produites dans un organisme et étudier

la façon dont elles interagissent en réseau, si l'on veut comprendre le fonctionnement des cellules et des organismes.

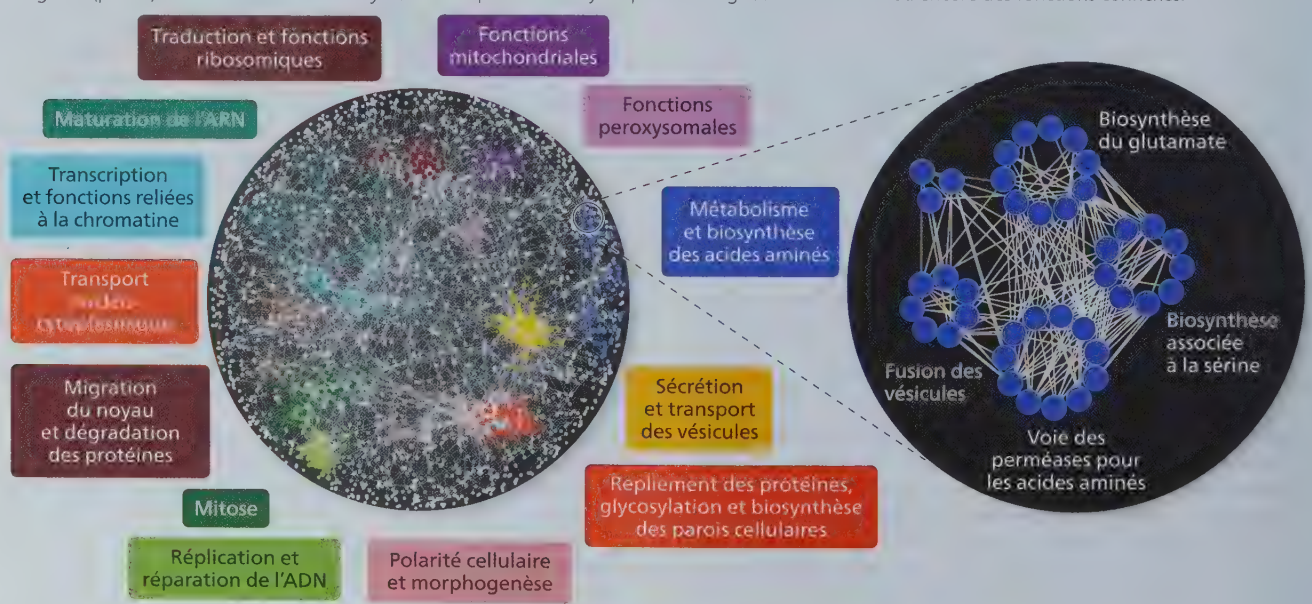
Grâce à la génomique et à la protéomique, les biologistes moléculaires sont en train d'acquérir une vision de plus en plus globale du monde vivant. À l'aide des outils que nous avons décrits, ils ont commencé à dresser des catalogues de gènes et de protéines, des listes complètes de tous les « morceaux » qui contribuent au fonctionnement des cellules, des tissus et des organismes. Grâce à ces catalogues, les chercheurs ont pu détourner leur attention des composants individuels (gènes et protéines) pour se concentrer sur l'intégration fonctionnelle au sein des systèmes biologiques. Comme vous vous en souvenez, le concept 1.1 examine cette approche de la **biologie des systèmes**, dont l'objectif est de représenter par modèles le comportement dynamique de systèmes biologiques entiers en se fondant sur l'étude des interactions entre les composants de l'organisme. Étant donné le vaste éventail de données générées par ces types d'études, les avancées informatiques et bio-informatiques sont essentielles à l'étude de la biologie des systèmes.

Une application importante de l'approche de la biologie des systèmes consiste à définir les circuits de gènes et les réseaux d'interactions entre les protéines. Afin de cartographier le réseau d'interactions protéiques chez la levure *Saccharomyces cerevisiae*, par exemple, les chercheurs ont utilisé des techniques perfectionnées pour neutraliser des paires de gènes, une paire à la fois, créant des cellules doublement mutantes. Ils ont ensuite comparé la compatibilité de chaque double mutant (basée en partie sur la taille de la colonie de cellules formée) à celle prédite à partir des compatibilités des deux mutants uniques. Les chercheurs ont conclu que si la compatibilité observée correspondait à la prédiction, alors il n'y avait pas d'interaction entre les produits des deux gènes. En revanche, si la compatibilité observée était supérieure ou inférieure à celle prédite, c'est qu'il y avait eu interaction dans la cellule entre les produits de ces deux gènes. À l'aide d'un logiciel, ils ont alors élaboré un modèle graphique en cartographiant les produits géniques par rapport à certains emplacements dans le modèle. Pour y arriver, ils se sont fondés sur la similitude des interactions entre les protéines. Ils ont ainsi représenté le tout sous forme d'une « carte fonctionnelle », comme celle que montre la **figure 21.4**. Il a fallu se servir d'ordinateurs puissants et faire appel à des outils mathématiques et à des logiciels nouvellement mis au point pour traiter le grand nombre d'interactions protéine-protéine générées par cette expérience et les intégrer dans la carte complétée. L'approche de la biologie des systèmes a donc réellement été rendue possible grâce aux avancées de la technologie informatique et de la bio-informatique.

L'application de la biologie des systèmes à la médecine

Le projet Cancer Genome Atlas (Atlas génomique du cancer) est un autre exemple de la biologie des systèmes dans lequel on analyse simultanément plusieurs gènes et produits de gènes en interaction en tant que groupe. Réalisé sous la direction conjointe du National Cancer Institute et des NIH, ce projet vise à déterminer comment les modifications dans les systèmes biologiques peuvent causer le cancer. Un projet d'une durée de trois ans, achevé en 2010, avait pour but de découvrir toutes les mutations communes dans trois types de cancer (cancer du poumon, cancer de l'ovaire et glioblastome du cerveau) en comparant les séquences des gènes et les modes d'expression génétique des cellules cancéreuses avec ceux des cellules normales. Les travaux sur le glioblastome ont confirmé le rôle de plusieurs gènes

▼ Figure 21.4 L'approche de la biologie des systèmes appliquée aux interactions protéiques. Cette carte des interactions protéiques globales révèle les interactions probables (lignes) parmi environ 4 500 produits de gènes (points) chez la levure *Saccharomyces cerevisiae*. Les points de même couleur représentent les produits géniques intervenant dans l'une des 13 fonctions cellulaires de couleur similaire mentionnées autour de la carte. Les points blancs représentent les protéines n'ayant pas été assignées à une fonction d'une couleur donnée. La portion élargie montre des détails supplémentaires d'une région de la carte où les produits géniques (points bleus) effectuent la biosynthèse, l'incorporation des acides aminés ou encore des fonctions connexes.



soupçonnés et ont permis d'en isoler quelques-uns jusqu'alors inconnus, révélant de nouvelles cibles possibles pour des thérapies. La stratégie s'est avérée si fructueuse pour ces trois types de cancer qu'elle a été étendue à dix autres types, choisis parce qu'ils sont répandus et souvent mortels chez les humains.

Comme en témoigne le Roadmap Epigenomics Project décrit précédemment, les techniques à haut débit sont de plus en plus utilisées dans l'étude du cancer. Cela s'explique par leur rapidité croissante et par la diminution de leur coût. D'ailleurs, plutôt que de séquencer uniquement les gènes codant pour une protéine, on séquence les génomes entiers de plusieurs tumeurs d'un type particulier, ce qui permet aux scientifiques de mettre en évidence des anomalies chromosomiques courantes ainsi que tout autre changement systématique dans ces génomes aberrants.

En plus du séquençage du génome entier, des puces de silicium et de verre contenant un microréseau de la plupart des gènes humains connus sont maintenant utilisées pour analyser les modes d'expression génétique chez les patients qui souffrent de divers cancers et d'autres maladies (**figure 21.5**). De plus en plus, on abandonne les tests sur microréseaux au profit de l'ARN-seq (voir la figure 20.13). L'analyse des gènes surexprimés ou sous-exprimés en présence d'un cancer donné permet aux médecins d'adapter le traitement au profil génétique unique de leurs patients et aux caractéristiques particulières de leur cancer. Cette approche a été utilisée pour caractériser des sous-ensembles de plusieurs cancers particuliers, ce qui a permis de mettre au point des traitements mieux ciblés. Le cancer du sein en est un bon exemple (voir la figure 18.27).

Dans quelques années, les dossiers médicaux contiendront peut-être un catalogue de la séquence d'ADN, une sorte de code-barres génétique, dans lequel seraient mises en évidence des régions associées à un risque accru de maladies particulières. L'utilisation de ces séquences pour une médecine personnalisée (prévention et traitement des maladies) possède un énorme potentiel.

La biologie des systèmes est une façon très efficace d'étudier les propriétés émergentes à l'échelle moléculaire. Nous avons vu au chapitre premier que, selon le thème des propriétés émergentes, des propriétés nouvelles apparaissent à chaque niveau successif de la complexité biologique à la suite du réarrangement des éléments constitutifs du niveau inférieur (voir le concept 1.1). Plus nous pourrons en apprendre sur l'arrangement et les interactions des composantes des systèmes génétiques, plus nous approfondirons notre compréhension des organismes entiers. Dans le reste du présent chapitre, nous passerons en revue ce que les études génomiques nous ont appris jusqu'à maintenant.

◀ **Figure 21.5 Une puce à microréseau de gènes humains.** De minuscules points d'ADN fixés en rangées ordonnées sur cette plaquette de silicium représentent presque tous les gènes du génome humain. À l'aide de cette puce, les chercheurs peuvent analyser les modes d'expression pour tous ces gènes simultanément (voir la figure 20.12).

RETOUR SUR LE CONCEPT 21.2

1. Quel rôle joue internet dans la recherche actuelle en génomique et en protéomique ?
2. Expliquez les avantages de l'approche de la biologie des systèmes pour étudier le cancer par rapport à l'approche de l'étude d'un seul gène à la fois.
3. **FAITES DES LIENS ▶** Le projet pilote ENCODE a montré que plus de 75 % du génome est transcrit en ARN, beaucoup plus que ce qui peut être attribuable aux gènes qui codent pour des protéines. Revoyez les concepts 17.3 et 18.3 et proposez quelques rôles que peuvent jouer ces ARN.
4. **FAITES DES LIENS ▶** Au concept 20.2, vous avez pris connaissance des études d'association sur l'ensemble du génome. Expliquez comment ces études utilisent l'approche de la biologie des systèmes.

Voir les réponses proposées à l'appendice A.

CONCEPT 21.3

Les génomes varient en taille, en nombre de gènes et en densité génique

À ce jour, des milliers de génomes ont été complètement séquencés, alors que des dizaines de milliers d'autres génomes sont en cours de séquençage ou sont considérés comme des ébauches permanentes (parce que les efforts qu'il faudrait déployer pour les terminer n'en vaudraient pas la peine). Parmi les génomes en cours de séquençage, on dénombre environ 3 400 métagénomes. Dans le groupe des génomes complètement séquencés, on compte ceux d'environ 5 000 bactéries et de plus de 240 archées. Chez les eucaryotes, on a séquencé la totalité du génome de près de 300 espèces et on dispose d'ébauches permanentes pour plus de 2 600 espèces, dont des vertébrés, des invertébrés, des protistes, des eumycètes et des végétaux. Dans la prochaine section, nous discuterons des connaissances acquises sur la taille des génomes, le nombre de gènes et la densité génique en prêtant une attention particulière aux tendances générales.

La taille des génomes

La comparaison des trois domaines (bactéries, archées et eucaryotes) révèle une différence générale dans la taille du génome des procaryotes et celle des eucaryotes (**tableau 21.1**). Malgré quelques exceptions, la plupart des génomes bactériens comportent entre 1 et 6 millions de paires de bases (Mb). Le génome d'*E. coli*, par exemple, possède 4,6 Mb. Les dimensions des génomes d'archées se situent pour la plupart dans le même intervalle que celles des génomes bactériens. (Souvenez-vous cependant que beaucoup moins de génomes d'archées ont été complètement séquencés, de sorte que ce portrait pourrait changer.) Les génomes eucaryotes tendent à être plus imposants : le génome de la levure unicellulaire *Saccharomyces cerevisiae* (un eumycète) comporte environ 12 Mb, alors que celui de la majorité des animaux et des végétaux, qui sont des multicellulaires, compte au moins 100 Mb. Il y a 165 Mb dans le

Tableau 21.1 La taille du génome et le nombre estimé de gènes*

Organisme	Taille du génome haploïde (Mb)	Nombre approximatif de gènes	Gènes par Mb
Bactéries			
Haemophilus influenzae	1,8	1 700	1 080
Escherichia coli	4,6	4 300	930
Archées			
Archaeglobus fulgidus	2,2	2 500	1 130
Methanosarcina barkeri	4,9	3 700	760
Eucaryotes			
Saccharomyces cerevisiae (levure, un eumycète)	12	6 300	520
Utricularia gibba (utriculaire flottante)	82	28 500	350
Caenorhabditis elegans (nématode)	100	45 100	450
Arabidopsis thaliana (plante de la famille moutarde)	157	32 700	210
Drosophila melanogaster (drosophile)	123	17 300	140
Daphnia pulex (puce d'eau)	200	31 000	155
Zea mays (maïs)	2 130	44 500	20
Ailuropoda melanoleuca (panda géant)	2 400	21 600	9
Homo sapiens (humain)	3 080	<20 000	7
Paris japonica (plante japonaise)	149 000	AD	AD

* Le nombre de gènes inclut les gènes d'ARN (transcrits, non traduits). Certaines valeurs présentées dans le tableau pourraient être modifiées à mesure que l'analyse des génomes se poursuit. Mb: million de paires de bases ou mégabase; AD: aucune donnée.

génome de la drosophile, et 3 000 dans le génome humain, soit de 500 à 3 000 fois plus que dans une bactérie typique.

À l'exclusion de cette différence générale entre les procaryotes et les eucaryotes, une comparaison de la taille des génomes chez les eucaryotes ne réussit pas à révéler une relation systématique entre la taille du génome et le phénotype de l'organisme. Par exemple, le génome de *Paris japonica*, une plante japonaise, contient 149 milliards de paires de bases (149 000 Mb), alors que celle d'*Utricularia gibba*, ou l'urticaire flottante, ne contient que 82 Mb. La taille du génome d'une amibe unicellulaire, *Polychaos dubium*, en cours de séquençage, est un exemple encore plus frappant puisque son génome compterait 670 milliards de paires de bases (670 000 Mb). En comparant plus précisément deux espèces d'insectes, il s'avère que le génome du grillon (*Anabrus simplex*) renferme 11 fois plus de paires de bases que celui de *Drosophila melanogaster*. Il y a un large éventail de tailles de génomes au sein des groupes des eucaryotes unicellulaires, des insectes, des amphibiens et des végétaux, et une gamme moins étendue chez les mammifères et les reptiles.

Le nombre de gènes

Le nombre de gènes varie également entre les procaryotes et les eucaryotes: les bactéries et les archées possèdent en général moins de gènes que les eucaryotes. Les bactéries libres (non parasites) et les archées ont de 1 500 à 7 500 gènes, tandis que chez les eucaryotes ce nombre varie entre 5 000 environ pour les eumycètes unicellulaires (levures) et plus de 40 000 pour certains eucaryotes multicellulaires.

Chez les eucaryotes, le nombre de gènes que possède une espèce est souvent plus faible que ce que laisse supposer la taille du génome. En examinant le tableau 21.1, vous pouvez constater que la taille du génome du nématode *C. elegans* est de 100 Mb et que celui-ci contient environ 45 100 gènes. Par comparaison, le génome de *Drosophila melanogaster* est un peu plus gros (123 Mb), mais il ne renferme que 17 300 gènes, ce qui constitue moins de la moitié du nombre de gènes de *C. elegans*.

Si on examine un exemple qui nous concerne plus directement, on note que le génome humain contient 3 080 Mb, un nombre bien supérieur à la taille du génome de *Drosophila melanogaster* ou de *C. elegans*. Au début du projet Génome humain, les biologistes s'attendaient à isoler entre 50 000 et 100 000 gènes une fois terminé le séquençage. Cette estimation était fondée sur le nombre de protéines humaines connues. À mesure que le projet progressait, il a fallu revoir plusieurs fois les prévisions à la baisse. Aussi, dans le projet ENCODE dont il a été question précédemment, on a fixé ce nombre à environ 20 000. Le nombre relativement faible, comparable au nombre de gènes du nématode *C. elegans*, a surpris les biologistes, qui s'attendaient à ce que les gènes humains soient beaucoup plus nombreux.

Quels attributs génétiques permettent donc à l'humain (et aux autres vertébrés) de fonctionner sans posséder plus de gènes qu'un nématode? Un important facteur est que les séquences codantes des génomes des vertébrés sont plus «productives», parce que la fréquence des épissages extensifs différentiels est plus élevée dans les transcrits d'ARN. Souvenons-nous que, par le biais de ce processus, un seul gène est en mesure d'engendrer plus d'un polypeptide (voir la figure 18.13). Un gène humain typique contient environ 10 exons, et on estime qu'au moins 90% de ces gènes multiexons peuvent subir un épissage différentiel: certains gènes sont exprimés sous des centaines de formes différentes, d'autres en seulement deux formes. Les scientifiques n'ont pas encore répertorié toutes les différentes formes résultant d'épissages différentiels, mais il est clair que le nombre de protéines codées dans le génome humain excède de beaucoup le nombre proposé de gènes.

À cela s'ajoute la diversité des polypeptides résultant des modifications post-traductionnelles comme le clivage ou l'ajout de glucides dans différents types de cellules ou à divers stades de développement. Enfin, la découverte des miARN et d'autres petits ARN qui jouent des rôles de régulation (voir le concept 18.3) introduit une nouvelle variable. Certains scientifiques croient que ce niveau de régulation supplémentaire, lorsqu'il est présent, peut contribuer à une plus grande complexité des organismes à partir d'un nombre limité de gènes.

La densité génique et l'ADN non codant

Il est possible d'évaluer la densité génique chez différentes espèces en comparant la taille du génome et le nombre de gènes qu'il possède. En d'autres mots, on peut déterminer le nombre de gènes présents sur une longueur donnée d'ADN. Quand on compare les génomes de bactéries, d'archées et d'eucaryotes, on constate que les eucaryotes ont généralement des génomes plus gros, mais qu'ils renferment moins de gènes pour un nombre donné de paires de bases. Les humains possèdent des centaines ou des milliers de fois plus de paires de bases dans leurs génomes que la plupart des bactéries, comme nous l'avons déjà noté, mais, en moyenne, seulement de 5 à 15 fois plus de gènes ; par conséquent, la densité des gènes est plus faible chez les humains (voir le tableau 21.1). Même les eucaryotes unicellulaires, comme les levures, possèdent moins de gènes par million de paires de bases que les bactéries et les archées. Parmi les génomes entièrement séquencés, ce sont les humains et les autres mammifères qui présentent la densité génique la plus faible.

Dans tous les génomes bactériens étudiés jusqu'à maintenant, la majeure partie de l'ADN est constituée de gènes qui codent pour des protéines, de l'ARNt ou de l'ARNr ; les petites quantités d'ADN qui restent sont principalement constituées de séquences régulatrices non transcrites, comme les promoteurs. De plus, le segment nucléotidique le long d'un gène bactérien qui code pour des protéines n'est pas interrompu par des séquences non transcrites (il n'a pas d'introns). Par contre, dans le génome des eucaryotes, la plus grande partie de l'ADN n'est pas transcrite en protéines ni ne code pour des molécules d'ARN de fonction connue, et l'ADN comporte des séquences régulatrices plus complexes. En fait, les humains possèdent 10 000 fois plus d'ADN non codant que les bactéries. Chez les eucaryotes multicellulaires, une partie de cet ADN est présente sous forme d'introns dans le gène. En fait, ce sont les introns qui comptent pour la majeure partie de la différence dans la longueur moyenne entre les gènes des humains (27 000 paires de bases) et ceux des bactéries (1 000 paires de bases).

En plus des introns, les eucaryotes multicellulaires ont une grande quantité d'ADN non codant pour des protéines, situé entre les gènes. À la section suivante, nous décrirons la composition et l'arrangement de ces grands segments d'ADN dans le génome humain.

RETOUR SUR LE CONCEPT 21.3

1. Selon la meilleure estimation actuelle, le génome humain contient environ 20 000 gènes. Cependant, il est évident que le nombre de polypeptides différents dans les cellules humaines est bien supérieur à 20 000. Quels processus peuvent expliquer cette divergence ?
2. Le site de la base de données GOLD (Genomes Online Database) du Joint Genome Institute offre des renseignements sur les projets de séquençage génomique. Visitez la page https://gold.jgi.doe.gov/statistics et décrivez les renseignements qui y figurent. Quelle est la proportion de projets portant sur des génomes bactériens qui présentent un intérêt médical ?
3. Quels processus évolutifs pourraient expliquer que les procaryotes ont des génomes plus petits que les eucaryotes ?

Voir les réponses proposées à l'appendice A.

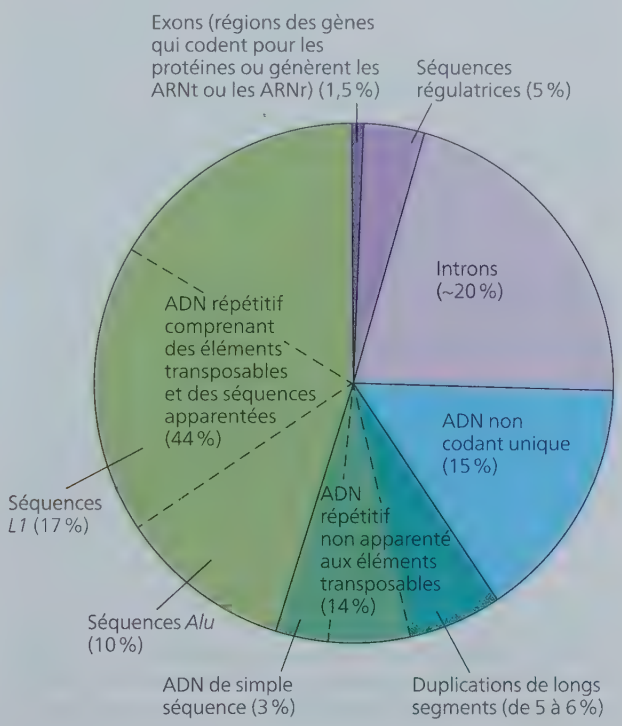
CONCEPT 21.4

Les eucaryotes multicellulaires possèdent beaucoup d'ADN non codant et de nombreuses familles multigéniques

Dans la majeure partie du chapitre et, bien sûr, dans la présente partie, nous avons mis l'accent sur les gènes qui codent pour des protéines. Pourtant, les régions codantes de ces gènes et les gènes pour les ARN non codants comme l'ARNr, l'ARNt et le miARN ne constituent qu'une petite partie du génome de la plupart des eucaryotes multicellulaires. Par exemple, une fois que la séquence complète du génome humain a été connue, il est devenu clair que seule une petite partie (environ 1,5 %) est transcrite en protéines ou code pour des ARNr ou des ARNt. La **figure 21.6** montre ce qui est connu des autres 98,5 %.

Les séquences régulatrices et les introns apparentés aux gènes représentent respectivement 5 % et environ 20 % du génome humain. Les autres séquences, situées entre les gènes fonctionnels, comportent un ADN non codant unique (une seule copie), tels des fragments de gènes et des **pseudogènes**, c'est-à-dire des anciens gènes qui ont accumulé des mutations et ne produisent

▼ Figure 21.6 Les types de séquences d'ADN dans le génome humain. Les séquences des gènes qui sont transcrites en protéines ou qui codent pour des molécules d'ARNr ou d'ARNt ne forment que 1,5 % du génome humain (en violet foncé dans le diagramme à secteurs), alors que les introns et les séquences régulatrices associés aux gènes (violet pâle) en forment le quart. La majeure partie du génome humain ne code pas pour des protéines (bien qu'une grande proportion génère de l'ARN), et une bonne part de ce génome est constituée d'ADN répétitif (en vert foncé et en vert pâle).



plus de protéines fonctionnelles. (Les gènes qui produisent les petits ARN non codants ne constituent qu'un faible pourcentage du génome, distribué entre 20 % d'introns et 15 % d'ADN non codant unique.) Toutefois, presque tout l'ADN situé entre les gènes fonctionnels est formé d'**ADN répétitif**, c'est-à-dire d'un grand nombre de copies de séquences nucléotidiques présentes dans le génome.

En effet, un grand nombre de ces génomes comportent beaucoup de séquences d'ADN qui ne codent pas pour des protéines, ni ne sont transcrites pour produire des ARN de fonctions connues. Dans le passé, cet ADN non codant a souvent été décrit par le terme « ADN poubelle ». Cependant, au cours des 10 dernières années, une comparaison des génomes a démontré la persistance de cet ADN dans divers génomes, sur des centaines de générations. Par exemple, les génomes de l'humain, du rat et de la souris contiennent près de 500 régions d'ADN non codant qui portent des séquences *identiques*. Il s'agit d'un niveau de conservation de la séquence plus élevé que ce qu'on observe dans les régions qui codent pour des protéines chez ces espèces. Il se pourrait donc que ces régions non codantes remplissent des fonctions essentielles. De plus, les résultats du projet ENCODE dont il a été question précédemment mettent en évidence les rôles essentiels joués par la plupart des ADN non codants. Dans les pages qui suivent, nous examinerons la répartition des gènes et des séquences non codantes d'ADN dans les génomes des eucaryotes multicellulaires ; le génome humain nous servira de principal exemple. La structure du génome nous renseigne beaucoup sur la façon dont les génomes sont apparus et continuent à évoluer, comme nous le verrons au concept 21.5.

Les éléments transposables et les séquences apparentées

Les procaryotes et les eucaryotes possèdent des portions d'ADN capables de se déplacer d'un endroit à un autre dans le génome ; chez certaines plantes, ces portions mobiles d'ADN peuvent constituer jusqu'à 75 % du génome. Ces segments d'ADN, dits *éléments génétiques transposables* ou simplement **éléments transposables**, se déplacent d'un site dans l'ADN d'une cellule vers un site cible différent grâce à la *transposition*, un processus de recombinaison d'un certain type. On nomme parfois ces éléments transposables « gènes sauteurs », mais en réalité ils ne se séparent jamais complètement de l'ADN de la cellule. Au contraire, les sites d'origine et les nouveaux sites de l'ADN sont fortement rapprochés par des enzymes et d'autres protéines qui replient l'ADN. Étonnamment, environ 75 % de l'ADN répétitif humain (44 % du génome humain entier) est constitué d'éléments transposables et de séquences qui leur sont associées.

La première démonstration de l'existence de tels segments d'ADN mobiles a été fournie par la généticienne américaine Barbara McClintock pendant qu'elle effectuait des expériences de croisement sur le maïs (*Zea mays*) au cours des années 1940 et de la décennie suivante (**figure 21.7**). Alors qu'elle étudiait des plants de maïs sur plusieurs générations, la scientifique a relevé des changements dans la couleur des grains qui ne pouvaient s'expliquer que par l'existence d'éléments génétiques mobiles. Ces éléments génétiques influeraient sur les gènes de la couleur des grains à partir d'autres emplacements dans le génome, interrompant les gènes de sorte que la couleur du grain était changée. La découverte de Barbara McClintock a été reçue avec beaucoup de scepticisme et a été pratiquement oubliée à l'époque. Ses travaux minutieux et ses idées visionnaires ont finalement été confirmés de nombreuses années plus tard lorsqu'on a trouvé des éléments transposables chez les bactéries. En 1983, alors âgée de 81 ans, Barbara McClintock a reçu un prix Nobel pour ses recherches novatrices.

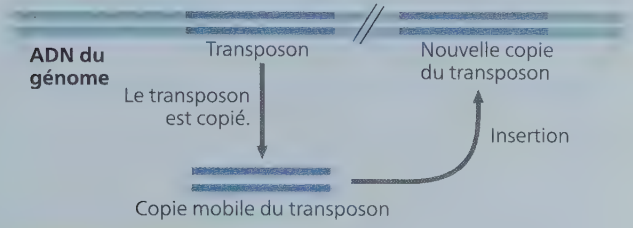
▼ **Figure 21.7 L'effet des éléments transposables sur la couleur de grains de maïs.** Barbara McClintock a été la première à proposer le concept d'éléments génétiques mobiles après avoir observé des bigarrures dans la couleur des grains d'un épi de maïs (à droite).

Le déplacement des transposons et des rétrotransposons

Il existe deux types d'éléments transposables chez les eucaryotes. Le premier comprend les **transposons**, qui se déplacent à l'intérieur d'un génome par l'intermédiaire d'un ADN. Les transposons peuvent se déplacer grâce à un mécanisme de type « couper-coller », qui enlève l'élément du site original, ou de type « copier-coller », qui laisse une copie au site original (**figure 21.8**). Les deux mécanismes nécessitent une enzyme, la transposase, qui est généralement codée par la séquence du transposon.

La plupart des éléments transposables dans les génomes eucaryotes sont du second type, les **rétrotransposons**, qui sont transportés à l'intérieur du génome par l'intermédiaire d'un ARN. Celui-ci est une transcription de l'ADN du rétrotransposon. De plus, les rétrotransposons laissent toujours une copie au site original au cours de la transposition (**figure 21.9**). Pour pouvoir s'introduire dans un autre site, l'intermédiaire d'ARN doit être reconverti en ADN sous l'action d'une transcriptase inverse, une enzyme encodée dans le rétrotransposon. La transcriptase inverse est également encodée par des rétrovirus, comme vous l'avez appris au concept 19.2. En fait, il est possible que les

▼ **Figure 21.8 Le déplacement des transposons.** Le déplacement des transposons soit par un mécanisme «copier-coller» (illustré ici), soit par un mécanisme «couper-coller», fait intervenir un intermédiaire d'ADN bicaténaire qui est inséré dans le génome.



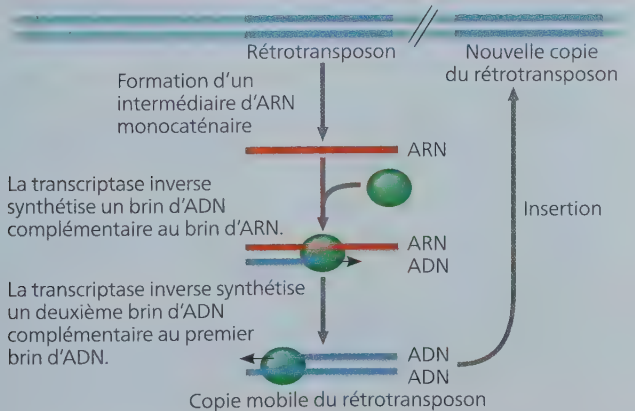
HABILETÉS VISUELLES ► En quoi cette figure serait-elle différente si elle illustrait le mécanisme «couper-coller»?

rétrovirus descendent de rétrotransposons. Une autre enzyme cellulaire catalyse l'insertion de l'ADN (reconverti par la transcriptase inverse) à un nouveau site.

Les séquences apparentées aux éléments transposables

Une multitude d'exemplaires d'éléments transposables et de séquences qui leur sont apparentées sont dispersés dans l'ensemble du génome eucaryote. Une seule de ces unités est habituellement longue de plusieurs centaines, voire de milliers de paires de bases. Les «copies» dispersées sont semblables, mais généralement non identiques. Certaines d'entre elles sont des éléments transposables capables de se déplacer. Les enzymes requises pour ce déplacement peuvent être encodées par un élément transposable, incluant celui qui se déplace. Certaines autres unités sont des séquences apparentées qui ont complètement perdu la capacité de se déplacer. Les éléments transposables et les séquences apparentées constituent de 25 à 50% du génome de la plupart des mammifères (voir la figure 21.6), et ces pourcentages sont encore plus élevés chez les amphibiens et les plantes supérieures. En fait, la très grande taille des génomes de certains végétaux est attribuable non pas à des gènes supplémentaires, mais à des éléments transposables additionnels. Par exemple, le génome du maïs est composé à 85% de telles séquences!

▼ **Figure 21.9 Le déplacement des rétrotransposons.** Le déplacement commence par la formation d'un intermédiaire d'ARN monocaténaire. Le reste des étapes est essentiellement identique à une partie du cycle de réplication des rétrovirus (voir la figure 19.9).



Chez les humains et les autres primates, une grande partie de l'ADN apparenté aux éléments transposables est formée d'une famille de séquences semblables, les *séquences Alu*. À elles seules, ces séquences représentent environ 10% du génome humain. Les séquences *Alu*, qui peuvent être répétées près de 1 million de fois chez l'humain, ont une longueur d'environ 300 nucléotides. Elles constituent des séquences beaucoup plus courtes que la plupart des éléments transposables fonctionnels et elles ne codent pour aucune protéine. Cependant, de nombreuses séquences *Alu* sont transcrites en ARN, et l'on croit que certains de ces ARN favorisent la régulation de l'expression génétique.

Un pourcentage encore plus élevé (17%) du génome humain est constitué d'un type de rétrotransposon, les séquences *LINE-1*, ou *L1*. Elles sont beaucoup plus longues que les séquences *Alu* (environ 6 500 paires de bases) et leur vitesse de transposition est généralement très faible. Cependant, des chercheurs travaillant avec des rats ont découvert que les rétrotransposons *L1* sont plus actifs dans les cellules du cerveau en développement que dans les autres types cellulaires. D'après eux, ces rétrotransposons exerceraient des effets différentiels sur l'expression génétique dans le développement des neurones, contribuant ainsi à la grande diversité des types de cellules neuronales (voir le concept 48.1).

Bien que de nombreux éléments transposables codent pour des protéines, ces dernières n'interviennent pas dans des fonctions cellulaires normales. Elles peuvent même inactiver certains gènes par leur insertion. Une forme d'hémophilie héréditaire, par exemple, est causée par une protéine devenue anormale (le facteur VIII), en raison de l'insertion dans le chromosome X d'un transposon provenant du chromosome 22. Par conséquent, on inclut habituellement les éléments transposables ainsi que d'autres séquences répétitives dans la catégorie des ADN «non codants».

Les autres ADN répétitifs, dont l'ADN de simple séquence

L'ADN répétitif qui n'est pas apparenté aux éléments transposables est probablement apparu à la suite d'erreurs survenues au cours de la réplication ou de la recombinaison de l'ADN. Cet ADN représente autour de 14% du génome humain (voir la figure 21.6). Environ un tiers de ce pourcentage (de 5 à 6% du génome humain) consiste en des duplications de longs segments d'ADN de 10 000 à 30 0000 paires de nucléotides chacun. Ces longs segments semblent avoir été copiés d'un locus chromosomique à un autre site, sur le même chromosome ou sur un chromosome différent, et incluent probablement quelques gènes fonctionnels.

Contrairement à des copies dispersées de longues séquences, des segments d'ADN connus sous le nom d'**ADN de simple séquence** contiennent de nombreux exemplaires de courtes séquences répétitives en tandem, comme dans l'exemple suivant (montrant un seul brin d'ADN):

...GTTACGTTACGTTACGTTACGTTACGTTAC...

Dans ce cas, l'unité répétée (GTTAC) consiste en 5 nucléotides. Les unités répétées peuvent contenir jusqu'à 500 nucléotides, mais elles en renferment souvent moins de 15, comme dans cet exemple. Les **répétitions courtes en tandem** ou **STR** (pour *short tandem repeat*) comportent des unités de 2 à 5 nucléotides;

nous avons décrit leur utilisation pour comparer des profils génétiques au concept 20.4 (voir la figure 20.24). Pour un génome donné, le nombre d'exemplaires de l'unité répétée peut varier d'un site à l'autre. Il pourrait y avoir plusieurs centaines de milliers de répétitions de l'unité GTTAC à un site, mais seulement la moitié de ce nombre à un autre. L'analyse des STR est effectuée sur des sites choisis pour leur nombre relativement faible de répétitions. Le nombre de répétitions peut varier d'une personne à l'autre et pour un même site chez une même personne, il peut varier d'un allèle à l'autre (puisque les humains sont diploïdes). L'analyse des STR permet de représenter cette diversité au moyen de profils génétiques spécifiques à chaque individu. Dans l'ensemble, l'ADN de simple séquence forme 3 % du génome humain.

Dans un génome donné, une grande partie de l'ADN de simple séquence est située dans les télomères et dans les centromères. Cela permet de penser que cet ADN remplit un rôle structural dans les chromosomes. L'ADN des centromères joue un rôle essentiel au cours de la séparation des chromatides, pendant la division cellulaire (voir le concept 12.2). De plus, il contribue peut-être à structurer la chromatine contenue dans le noyau pendant l'interphase – et ce, conjointement avec l'ADN de simple séquence situé à un autre emplacement. L'ADN de simple séquence des télomères (extrémités des chromosomes) permet d'éviter la perte de gènes lorsque l'ADN est raccourci à chaque réplication (voir le concept 16.2). L'ADN des télomères protège également les chromosomes en se liant à des protéines ayant pour fonction d'empêcher les extrémités de ceux-ci de se dégrader ou de se joindre à d'autres chromosomes.

Les répétitions courtes comme celles décrites ici nuisent au séquençage en aveugle sur l'ensemble du génome, car elles empêchent les ordinateurs de rassembler, de façon précise, les séquences fragmentaires. Les régions constituées d'ADN de simple séquence sont en grande partie responsables de l'incertitude qui entoure toute estimation de la taille d'un génome entier, et elles expliquent pourquoi on considère certaines séquences comme des ébauches permanentes.

Les gènes et les familles multigéniques

Nous terminerons notre examen des divers types de séquences d'ADN dans les génomes d'eucaryotes par une étude plus détaillée des gènes. Souvenez-vous que les séquences d'ADN qui codent pour les protéines ou donnent naissance aux ARNt ou aux ARNr ne constituent pas plus de 1,5 % du génome humain (voir la figure 21.6). Si l'on inclut les introns et les séquences régulatrices associés aux gènes, la quantité totale d'ADN apparenté aux gènes (codant et non codant) représente environ 25 % du génome humain. Autrement dit, seulement environ 6 % (1,5 % de 25 %) de la longueur du gène moyen est représentée dans le produit final du gène.

De nombreux gènes eucaryotes, tout comme ceux des bactéries, ne renferment qu'un exemplaire de la plupart des gènes, c'est-à-dire qu'il n'y a qu'un seul exemplaire par jeu haploïde de chromosomes. Mais, dans le génome humain et dans celui de nombreux autres animaux et végétaux, ces gènes uniques ne forment qu'environ la moitié de l'ADN apparenté à des gènes. Le reste se trouve sous forme de **familles multigéniques**, des ensembles de gènes identiques ou très semblables.

Dans les familles multigéniques composées de séquences de gènes *identiques*, ces séquences sont habituellement groupées en tandem et, à l'exception importante des gènes des histones, elles codent pour de l'ARN. On peut citer l'exemple de la famille de séquences d'ADN identiques qui contiennent toutes les gènes codant pour les trois plus grandes molécules d'ARNr (**figure 21.10a**). Celles-ci sont transcrites à partir d'une même unité de transcription, qui est répétée en tandem des centaines ou des milliers de fois en un ou plusieurs regroupements dans le génome des eucaryotes multicellulaires ; chez l'humain, on a trouvé près de 300 de ces séquences identiques réparties sur 5 chromosomes. Ces nombreux exemplaires d'unités de transcription d'ARN aident les cellules à fabriquer les millions de ribosomes nécessaires à la synthèse protéique. Le transcrit primaire est découpé de façon à donner trois molécules d'ARNr qui forment ensuite des sous-unités ribosomiques en se combinant avec des protéines et un autre type d'ARNr (ARNr 5S).

Des exemples classiques de familles multigéniques constituées de gènes *non identiques* sont les deux familles apparentées qui codent pour les globines, groupe de protéines comprenant les sous-unités polypeptidiques α et β de l'hémoglobine. L'une de ces familles (située sur le chromosome 16 chez les humains) code pour diverses formes de la α-globine ; l'autre (située sur le chromosome 11), pour plusieurs formes de la β-globine (**figure 21.10b**). Les diverses formes de chaque sous-unité s'expriment à des stades distincts du développement, ce qui permet à l'hémoglobine de remplir ses fonctions de façon efficace, malgré les changements dans le milieu où l'individu se développe. Chez l'humain, par exemple, les formes d'hémoglobine de l'embryon et du fœtus ont une plus grande affinité pour la molécule d'oxygène (O_2) que celles qui existent chez l'adulte. Cela permet d'assurer un transfert efficace de l'O_2 de la mère au fœtus. Les familles multigéniques des globines comprennent également plusieurs pseudogènes.

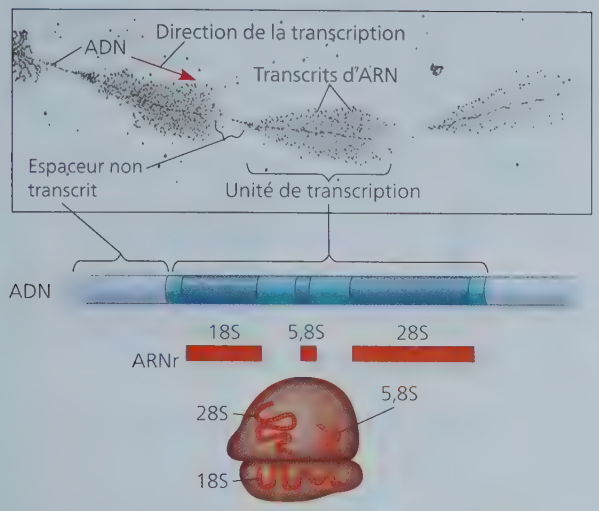
Au concept 21.5, nous examinerons l'évolution des deux familles multigéniques de la globine alors que nous verrons comment la classification des gènes en familles a permis de comprendre l'évolution des génomes. Nous nous pencherons également sur quelques-uns des processus qui ont formé les génomes de diverses espèces au cours de l'évolution.

RETOUR SUR LE CONCEPT 21.4

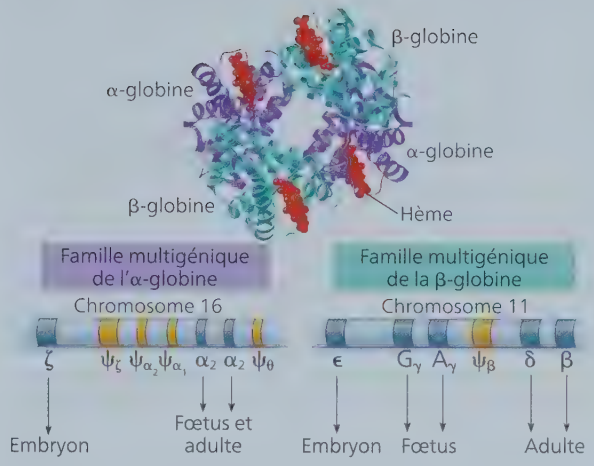
1. Discutez des caractéristiques qui font que les génomes des mammifères sont plus gros que ceux des procaryotes.
2. **HABILETÉS VISUELLES ▶** Parmi les trois mécanismes décrits aux figures 21.8 et 21.9, lequel ou lesquels aboutissent à une copie qui reste sur le site d'origine et apparaît aussi sur un nouveau site ?
3. Comparez l'organisation de la famille de gènes des ARNr et celle des familles de gènes des globines. Expliquez pour chacune comment l'existence d'une famille de gènes apporte des avantages aux organismes.
4. **FAITES DES LIENS ▶** Assignez chaque segment d'ADN dans le haut de la figure 18.8 à un secteur dans le diagramme de la figure 21.6.

Voir les réponses proposées à l'appendice A.

▼ **Figure 21.10** Les familles de gènes.



(a) Partie d'une famille de gènes de l'ARN ribosomique. La micrographie ci-dessus montre trois exemplaires (il en existe des centaines) d'unités de transcription d'ARNr présentes dans la famille de gènes de l'ARNr d'une salamandre (MET). Chacune des «plumes» correspond à une unité en cours de transcription par une centaine de molécules d'ARN polymérase (les points foncés situés le long de l'ADN). Ces molécules se déplacent de gauche à droite (flèche rouge). Les transcrits d'ARN en cours de synthèse s'allongent se détachent peu à peu de l'ADN (de gauche à droite), ce qui explique l'aspect penniforme. Le diagramme sous la micrographie illustre une unité de transcription. Celle-ci comprend les gènes de trois types d'ARNr (en bleu foncé), adjacents à des régions transcrites mais enlevées par la suite (en bleu moyen). Un seul transcrit, après maturation (épissage), donne une molécule de chacun des trois ARNr (en rouge), les composants clés d'un ribosome.



(b) Familles multigéniques de l'α-globine et de la β-globine. Chez l'adulte, l'hémoglobine est formée de quatre sous-unités polypeptidiques, soit deux α-globines et deux β-globines, comme l'illustre le modèle moléculaire. Les gènes (en bleu foncé) qui codent pour les globines α et β appartiennent à deux familles, disposées sur les chromosomes 16 et 11, comme dans l'illustration. L'ADN non codant (en bleu pâle) qui sépare les gènes fonctionnels d'une même famille comporte des pseudogènes (dorés et désignés par la lettre grecque ψ), des versions inactives de gènes fonctionnels (qui ne produisent plus de protéines fonctionnelles). Les gènes et les pseudogènes sont désignés par des lettres grecques, comme vous l'avez constaté précédemment pour les α-globines et les β-globines. Certains gènes ne sont exprimés que chez l'embryon ou le fœtus.

HABILETÉS VISUELLES ▶ Dans la micrographie de la partie (a), comment pourriez-vous déterminer le sens de la transcription s'il n'était pas indiqué par la flèche rouge ?

CONCEPT **21.5**

Les duplications, les réarrangements et les mutations de l'ADN contribuent à l'évolution du génome

ÉVOLUTION Maintenant que nous savons de quoi est constitué le génome humain, voyons ce que sa composition nous révèle sur son évolution. Les mutations constituent le fondement des modifications à l'échelle du génome ; elles sont à l'origine de son évolution. Il semble que les premières formes de vie ne possédaient qu'un génome minimal, qui se limitait aux gènes nécessaires à la survie et à la reproduction. Si tel était le cas, l'évolution a dû être caractérisée par une augmentation de la taille du génome, c'est-à-dire que le matériel génétique supplémentaire fournissait la matière première pour la diversification des gènes. Dans la présente section, nous allons d'abord décrire comment les exemplaires supplémentaires du génome en tout ou en partie peuvent apparaître, puis nous examinerons les processus subséquents qui peuvent mener à l'évolution des protéines (ou des molécules d'ARN) possédant des fonctions légèrement différentes ou entièrement nouvelles.

La duplication des jeux complets de chromosomes

Un accident au cours de la méiose, comme l'incapacité de séparer les chromosomes homologues pendant la méiose I, peut donner naissance à un ou à plusieurs jeux supplémentaires de chromosomes, ce qui entraîne un état de *polyploïdie* (voir le concept 15.4). Bien que de tels accidents soient généralement létaux, ils facilitent parfois l'évolution des gènes. Chez un organisme polyploïde, un jeu de gènes peut fournir les fonctions essentielles à l'organisme. Les gènes dans le ou les jeux supplémentaires peuvent diverger en accumulant les mutations, et ces variations persistent si l'organisme qui les porte survit et se reproduit. C'est de cette façon que les gènes dotés de nouvelles fonctions évoluent. Pourvu qu'une copie d'un gène essentiel soit exprimée, la divergence d'une autre copie peut mener à sa protéine codée qui joue un nouveau rôle, changeant ainsi le phénotype de l'organisme.

Ultimement, l'accumulation de mutations peut avoir pour résultat l'apparition d'une nouvelle espèce. La polyploïdie est rare chez les animaux, mais elle est relativement fréquente chez les végétaux, en particulier chez les plantes à fleurs. Selon certains botanistes, jusqu'à 80 % des espèces végétales existant

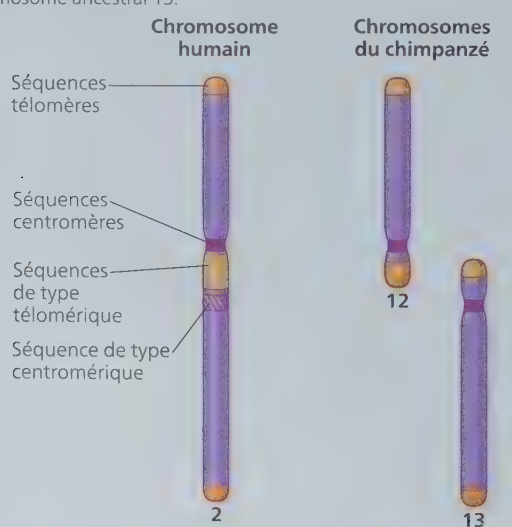
aujourd'hui montreraient des signes de polyploïdie ancestrale. Au concept 24.2, vous apprendrez comment une polyploïdie mène à la spéciation chez les végétaux.

Les modifications de la structure chromosomique

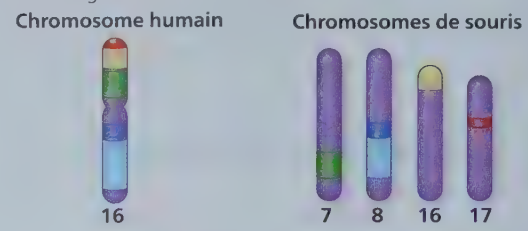
Grâce au très grand nombre de données recueillies récemment sur les séquences génomiques, nous pouvons maintenant comparer les structures chromosomiques fines de nombreuses espèces différentes. Ces données nous permettent en effet de faire des inférences au sujet des processus de l'évolution qui façonnent les chromosomes et qui peuvent entraîner la spéciation. Par exemple, les scientifiques savent depuis longtemps qu'au cours des derniers 6 millions d'années, lorsque les ancêtres des humains et des chimpanzés ont divergé en tant qu'espèces, la fusion de deux chromosomes ancestraux dans la lignée humaine a mené à des nombres haploïdes différents pour les humains ($n = 23$) et les chimpanzés ($n = 24$). Les bandes dans les chromosomes colorés donnent à penser que les versions ancestrales des chromosomes actuels 12 et 13 du chimpanzé ont effectué une fusion bout à bout, pour former le chromosome 2 chez un ancêtre de la lignée humaine. Dans le cadre du projet Génome humain, le séquençage et l'analyse du chromosome 2 de l'humain ont apporté des preuves convaincantes à l'appui de ce modèle (**figure 21.11**).

Dans une autre étude d'une plus grande portée, les chercheurs ont comparé la séquence de l'ADN de chaque chromosome humain avec la séquence du génome entier de la souris (**figure 21.12**). Une partie de cette étude a démontré que de grands segments de gènes du chromosome 16 humain trouvent leur équivalent sur quatre chromosomes de souris (7, 8, 16 et 17),

▼ Figure 21.11 Chromosomes de l'humain et du chimpanzé. Les positions des séquences de type télomérique et de type centromérique sur le chromosome humain 2 (à gauche) correspondent à celles des télomères sur les chromosomes 12 et 13 du chimpanzé et à celle du centromère sur le chromosome 13 du chimpanzé (à droite). Cela donne à penser que les chromosomes 12 et 13 chez un ancêtre humain ont effectué une fusion bout à bout pour former le chromosome humain 2. Le centromère du chromosome ancestral 12 est resté fonctionnel sur le chromosome humain 2, contrairement à celui du chromosome ancestral 13.



▼ Figure 21.12 Chromosomes de l'humain et de la souris. Des séquences d'ADN très semblables à de grands segments du chromosome 16 de l'humain (les zones colorées dans le schéma) sont présentes sur les chromosomes 7, 8, 16 et 17 de la souris. On peut en déduire que les séquences d'ADN de chaque segment sont restées ensemble dans les lignées de la souris et de l'humain depuis le temps où ils ont divergé d'un ancêtre commun.



ce qui indique que les gènes de chaque segment sont restés ensemble, à la fois dans les lignées de la souris et celles de l'humain, depuis le temps où ils ont divergé d'un ancêtre commun.

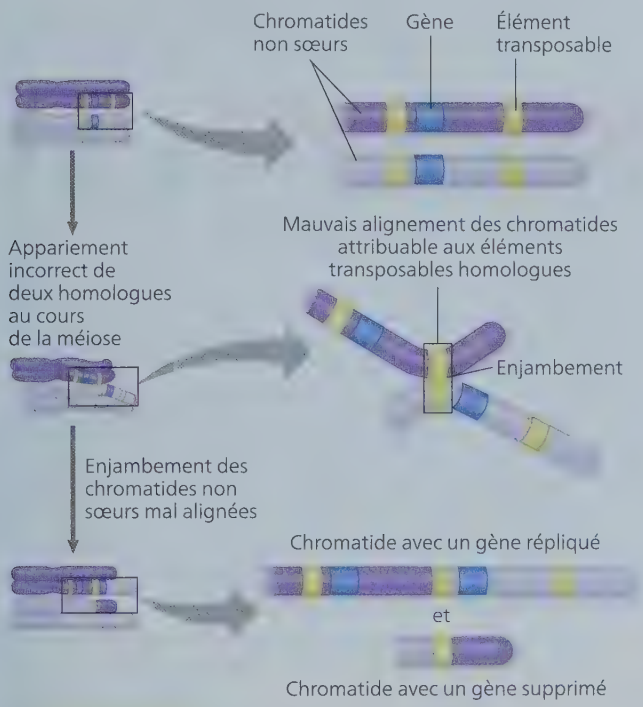
La même comparaison entre les chromosomes des humains et ceux de six autres espèces de mammifères a également permis aux chercheurs de reconstruire l'histoire évolutive des réarrangements chromosomiques chez ces huit espèces. Ils ont découvert de nombreuses duplications et inversions qui se sont produites au cours de la recombinaison méiotique et dans lesquelles l'ADN rompu a été raccordé incorrectement. Le taux de ces événements semble s'être accéléré il y a près de 100 millions d'années, soit environ 35 millions d'années avant l'extinction des grands dinosaures et l'augmentation rapide du nombre d'espèces de mammifères. La coïncidence apparente est intéressante parce qu'il se pourrait bien que les réarrangements chromosomiques aient contribué à l'apparition de nouvelles espèces. Bien que deux individus ayant des arrangements différents puissent encore s'accoupler et produire des descendants, ceux-ci posséderaient deux jeux non équivalents de chromosomes, rendant la méiose inefficace, voire impossible. Par conséquent, les réarrangements chromosomiques conduiraient à deux populations incapables de s'accoupler l'une avec l'autre, une étape conduisant à la formation de deux espèces séparées. (Vous en apprendrez plus à ce sujet au concept 24.2.)

La même étude a mis en lumière un sujet susceptible d'avoir des répercussions sur le plan médical. En effet, l'analyse des points de rupture associés aux réarrangements a établi l'existence de sites spécifiques sur lesquels les remaniements chromosomiques se sont produits à maintes reprises. Certains de ces « points chauds » de recombinaison correspondent à des emplacements de réarrangements chromosomiques dans le génome humain associés à des maladies congénitales (voir le concept 15.4).

La duplication et la divergence de régions d'ADN de la taille d'un gène

Les erreurs au cours de la méiose peuvent également aboutir à la duplication de régions chromosomiques plus petites que celles que nous venons d'examiner, incluant des segments de la longueur de gènes individuels. Un enjambement inégal à la prophase I de la méiose, par exemple, peut donner un chromosome portant une délétion et un autre avec une duplication d'un gène particulier. Des éléments transposables constituent des sites homologues où des chromatides non sœurs peuvent effectuer un enjambement, même lorsque les autres régions de ces chromatides ne sont pas correctement alignées (**figure 21.13**).

▼ Figure 21.13 La duplication de gènes attribuable à un enjambement inégal. La recombinaison au cours de la méiose entre des exemplaires d'un élément transposable (en jaune) flanquant le gène (en bleu) est un des mécanismes par lequel un gène ou un autre segment d'ADN peut être dupliqué. Cette recombinaison entre des chromatides non sœurs mal alignées de chromosomes homologues produit une chromatide ayant deux exemplaires du gène et une chromatide sans aucun exemplaire. (Les gènes et les éléments transposables sont représentés uniquement pour la région d'intérêt.)

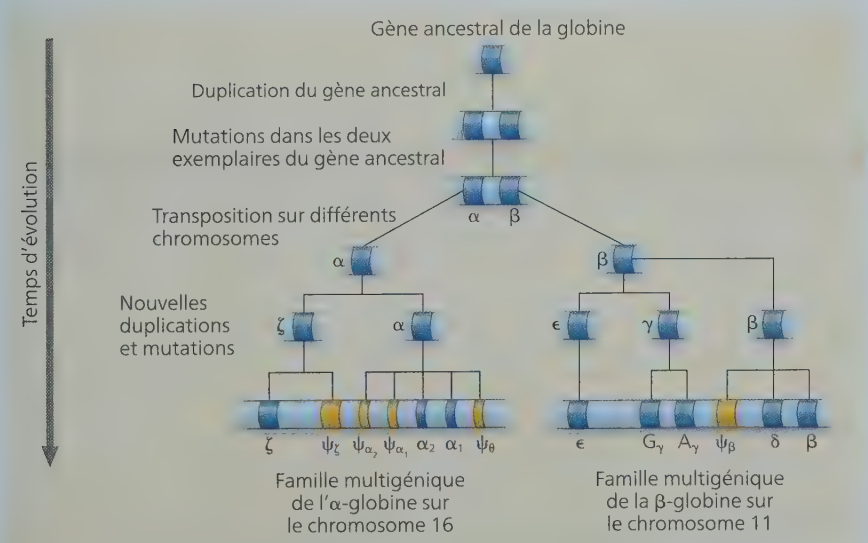


FAITES UN DESSIN ► Examinez comment se produit un enjambement à la figure 13.9. Dans la section centrale de la figure ci-dessus, tracez une ligne traversant les portions qui produisent la chromatide du haut dans la partie du bas de la figure. Utilisez une couleur différente pour faire la même chose pour l'autre chromatide.

De plus, un glissement peut survenir pendant la réplication et entraîner un déplacement de la matrice par rapport au nouveau brin complémentaire en formation. Il s'ensuit qu'une partie du brin matrice n'est pas copiée par le mécanisme de réplication, ou encore qu'elle l'est deux fois. Il y a donc délétion ou duplication d'un segment de l'ADN. Il est facile d'imaginer comment de telles erreurs peuvent se produire dans des régions de séquences répétées. Le nombre variable d'unités répétées d'ADN de simple séquence à un site donné, utilisées pour l'analyse des courtes répétitions en tandem, est probablement attribuable à de telles erreurs. L'existence de familles multigéniques, comme la famille des globines, fournit une preuve que des événements moléculaires tels que l'enjambement inégal et le glissement de matrice pendant la réplication de l'ADN peuvent être à l'origine de la duplication de gènes.

L'évolution des gènes à fonctions apparentées : les gènes de la globine humaine

La figure 21.10b présente l'organisation des familles multigéniques de l'α-globine et de la β-globine telles qu'elles existent dans le génome humain actuel. Maintenant, voyons comment des événements comme la duplication peuvent mener à l'évolution des gènes dont les fonctions sont connexes, comme ceux des globines. En comparant des séquences de gènes à l'intérieur d'une famille multigénique, il est possible d'entrevoir l'ordre dans lequel les gènes sont apparus. La reconstitution de l'histoire de l'évolution des gènes de la globine à l'aide de cette approche indique qu'ils ont tous évolué à partir d'un gène ancestral commun. De plus, ce gène ancestral a subi une duplication et une divergence au sein des gènes ancestraux de l'α-globine et de la β-globine il y a 450 à 500 millions d'années. Par la suite, chacun de ces gènes a été dupliqué à plusieurs reprises, et les copies ont alors divergé les unes à la suite des autres pour donner les membres des familles actuelles (**figure 21.14**). En fait, le gène ancestral commun de la globine a également donné naissance à la myoglobine, la protéine musculaire qui stocke l'O_2, et à la leghémoglobine, une protéine végétale. Ces deux dernières protéines fonctionnent comme des monomères et leurs gènes font partie d'une « superfamille des gènes de la globine ».



◄ Figure 21.14 Le modèle proposé de la séquence des événements menant à l'apparition des familles multigéniques de l'α-globine et de la β-globine à partir d'un seul gène ancestral de la globine.

? Les éléments dorés sont des pseudogènes. Expliquez comment ils ont pu apparaître après une duplication génique.

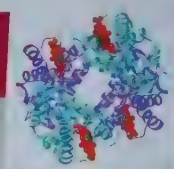
Après les événements de duplication, les différences entre les gènes des familles de la globine sont sans doute apparues à la suite de mutations qui se sont accumulées dans les exemplaires des gènes pendant de nombreuses générations. Selon le modèle actuel, la fonction nécessaire assurée par une α-globine, par exemple, était remplie par un gène, alors que d'autres copies du gène de l'α-globine accumulaient des mutations aléatoires. De nombreuses mutations peuvent avoir eu un effet négatif sur l'organisme et d'autres peuvent n'avoir produit aucun effet ; toutefois, il se peut que quelques mutations modifient avantageusement la fonction de la protéine pour l'organisme à un stade particulier de sa vie sans entraîner pour autant de changements substantiels pour ce qui est de sa capacité de transport de l'O_2. La sélection naturelle a probablement agi sur ces gènes modifiés pour les maintenir dans la population.

Dans l'exercice de la rubrique **Habiletés scientifiques**, vous pourrez comparer les séquences d'acides aminés des protéines de la famille des globines pour constater la façon dont ces comparaisons ont servi à produire le modèle de l'évolution des gènes de la globine présenté à la figure 21.14. L'existence de

DÉMARCHE SCIENTIFIQUE

HABILETÉS SCIENTIFIQUES

Lire un tableau des séquences d'acides aminés



▶ Hémoglobine.

■ COMMENT LES SÉQUENCES D'ACIDES AMINÉS DES GÈNES DE LA GLOBINE HUMAINE ONT-ELLES DIVERGÉ PENDANT LEUR ÉVOLUTION ? ■ Les séquences d'ADN évoluent de façon divergente, en accumulant des mutations qui peuvent générer des différences dans la séquence des acides aminés de la protéine codée. Il est possible d'étudier l'histoire évolutive au niveau moléculaire en comparant les séquences des différentes protéines d'une même famille de gènes. Pour créer un modèle de l'histoire évolutive des gènes de la globine (voir la figure 21.14), les chercheurs ont comparé les séquences d'acides aminés des polypeptides codés par ces gènes. L'hypothèse sous-entendue dans cette démarche est que plus les séquences de protéines se ressemblent, plus les gènes qui les codent sont de proches parents évolutifs. Dans cet exercice, vous analyserez différentes comparaisons des séquences d'acides aminés des polypeptides de la globine pour mettre en évidence les relations évolutives.

■ MÉTHODE ■ Les scientifiques ont mis en évidence les séquences d'ADN de chacun des huit gènes de la globine, et ils les ont traduits en séquences d'acides aminés. Par la suite, ils ont utilisé un logiciel pour aligner les séquences (les tirets indiquant les trous dans une séquence) et calculer le pourcentage d'identité pour chaque paire de globines. Ce pourcentage représente le nombre de positions auxquelles se trouve un acide aminé identique par rapport au nombre total d'acides aminés dans une chaîne polypeptidique. Les données sont présentées dans un tableau pour illustrer les comparaisons par paires.

■ RÉSULTATS ■ Le tableau ci-dessous montre un exemple d'alignement par paires – celles des séquences d'acides aminés de l'α_1-globine (alpha-1 globine) et de la ζ-globine (zêta-globine). Les acides aminés sont désignés selon le code standard à une seule lettre. À gauche de la séquence, on trouve le numéro du premier acide aminé de la ligne. On a calculé le pourcentage d'identité des séquences d'acides aminés de l'α_1-globine et de la ζ-globine en divisant le nombre d'acides aminés équivalents (86, soulignés en jaune) par le nombre total de positions (143), multiplié par 100. On a obtenu un pourcentage d'identité de 60 % pour la paire α_1-ζ, comme le montre le tableau des acides aminés ci-dessous. La même formule a été utilisée pour calculer le pourcentage d'identité pour les autres paires de globines.

Globine	Alignement des séquences d'acides aminés des globines
α_1	1 MVLSPADKTNVKAAWGKVGAHAGEYGAEAL
ζ	1 MSLTKTERTIIVSMWAKISTQADTIGTETL
α_1	31 ERMFLSFPTTKTYFPHFDLSH-GSAQVKGH
ζ	31 ERLFLSHPQTKTYFPHFDL-HPGSAQLRAH
α_1	61 GKKVADALTNAVAHVDDMPNALSALSDLHA
ζ	61 GSKVVAAVGDAVKSIDDIGGALSKLSELHA
α_1	91 HKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
ζ	91 YILRVDPVNFKLLSHCLLVTLAARFPADFT
α_1	121 PAVHASLDKFLASVSTVLTSKYR
ζ	121 AEAHAAWDKFLSVVSSVLTEKYR

Tableau du pourcentage d'identité des acides aminés pour chaque paire de globines

		Famille α			Famille β				
		α_1 (alpha 1)	α_2 (alpha 2)	ζ (zêta)	β (bêta)	δ (delta)	ε (epsilon)	Aγ (gamma A)	Gγ (gamma G)
Famille α	α_1	-----	100	60	45	44	39	42	42
	α_2		-----	60	45	44	39	42	42
	ζ			-----	38	40	41	41	41
Famille β	β				-----	93	76	73	73
	δ					-----	73	71	72
	ε						-----	80	80
	Aγ							-----	99
	Gγ								-----

Créé à l'aide des données du National Center for Biotechnology Information (NCBI).

DÉMARCHE SCIENTIFIQUE

HABILETÉS SCIENTIFIQUES

Lire un tableau des séquences d'acides aminés (*suite*)

INTERPRÉTEZ LES DONNÉES ▼

1. Il est à noter que dans le tableau des acides aminés, les données sont disposées de façon à ce qu'on puisse comparer chaque paire de globines. (a) Certaines cases du tableau présentent des lignes pointillées plutôt qu'un chiffre. En tenant compte des paires comparées pour ces cases, quel pourcentage d'identité correspond aux lignes pointillées ? (b) Les cases situées dans la moitié inférieure du tableau sont vides. En vous fondant sur les renseignements déjà fournis dans le tableau, inscrivez les valeurs manquantes dans les cases vides. Pourquoi est-il logique que ces cases aient été laissées vides ?
2. Plus longue est la période de temps écoulé à partir de la duplication d'un gène, plus grande est la possibilité que les séquences de nucléotides aient divergé, ce qui pourrait entraîner des différences dans la séquence d'acides aminés des deux protéines produites. (a) En vous fondant sur cette prémisse, nommez les deux gènes qui divergent le plus l'un de l'autre. Quel est le pourcentage d'identité des acides aminés (communs) entre leurs polypeptides ? (b) Utilisez la même approche pour identifier les deux gènes de la globine dont la duplication est la plus récente. Quel est le pourcentage d'identité entre eux ?
3. D'après le modèle de l'histoire évolutive du gène de la globine présenté à la figure 21.14, les gènes de l'α-globine et de la β-globine sont apparus après la duplication et les mutations d'un gène ancestral. Ces gènes ont ensuite fait l'objet d'autres duplications et mutations. Quelles caractéristiques de l'ensemble de données appuient ce modèle ?
4. Dressez une liste ordonnée de tous les pourcentages d'identité du tableau en plaçant la valeur de 100 % au sommet de la liste. Près de chaque chiffre, indiquez à quelle paire de globines correspond la valeur. Utilisez une couleur pour les globines de la famille α et une autre couleur pour celles de la famille β. (a) Comparez l'ordre des paires de votre liste à leur position dans le modèle de la figure 21.14. L'ordre des paires met-il en évidence la même «proximité» relative, entre les différents membres de la famille des globines, que celle observée dans le modèle ? (b) Comparez les pourcentages d'identité des paires dans un même groupe (α ou β) à ceux des paires dans les deux groupes.

Pour en savoir plus : R. C. Hardison, Globin genes on the move, *Journal of Biology* 7 : 35.1-35.5 (2008).

plusieurs pseudogènes parmi les gènes fonctionnels des globines fournit une preuve additionnelle en faveur de ce modèle : les mutations aléatoires dans ces «gènes» au fil de l'évolution ont détruit leur fonction.

L'évolution des gènes assurant de nouvelles fonctions

Au cours de l'évolution des familles de gènes de la globine, la duplication des gènes et les divergences subséquentes ont donné naissance à des membres de cette famille de gènes codant pour des protéines qui assument des fonctions similaires (transport de l'O_2). Cependant, un exemplaire d'un gène dupliqué peut également subir des modifications qui amènent la protéine à remplir une fonction complètement nouvelle. Les gènes pour le lysozyme et l'α-lactalbumine en sont de bons exemples.

Le lysozyme est une enzyme qui aide à protéger les animaux contre l'infection bactérienne en hydrolysant les parois cellulaires des bactéries (voir la figure 5.16) ; l'α-lactalbumine est une protéine non enzymatique qui intervient dans la production du lait chez les mammifères. Les séquences des acides aminés et les structures tridimensionnelles (**figure 21.15**) de ces deux protéines sont très semblables. On trouve les deux gènes chez les mammifères, mais seul le lysozyme est présent chez les oiseaux. Ces constatations portent à croire que, quelque temps après la séparation des lignées aboutissant aux mammifères et aux oiseaux, le gène du lysozyme a subi une duplication au sein de la lignée des mammifères, mais pas dans celle des oiseaux. Par la suite, une copie du gène du lysozyme dupliqué a évolué vers un gène codant pour l'α-lactalbumine, une protéine de fonction entièrement nouvelle associée à une caractéristique fondamentale des mammifères, soit la production de lait. Dans une étude récente, des biologistes évolutionnistes ont analysé les génomes de vertébrés afin d'y déceler des gènes présentant des séquences similaires. La famille des lysozymes semble compter au moins huit membres qui sont largement répartis parmi les espèces de mammifères. À l'heure actuelle, on ignore quelles sont les fonctions accomplies par l'ensemble des produits géniques de cette famille, mais il sera intéressant de découvrir si elles varient autant que celles des lysozymes et des α-lactalbumines.

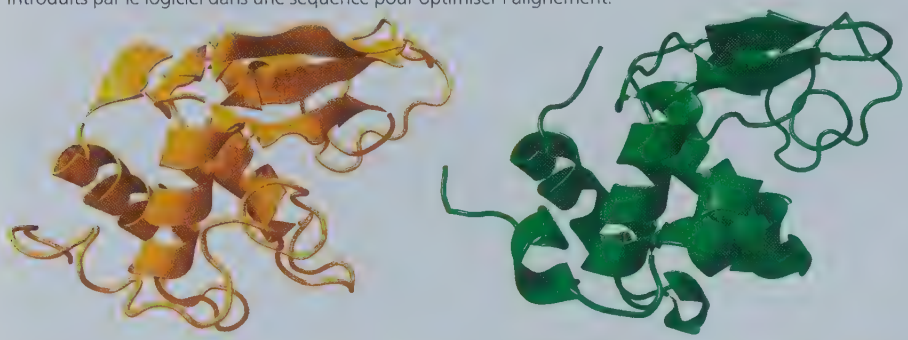
Mis à part la duplication et la divergence des gènes entiers, le réarrangement de séquences d'ADN existantes dans les gènes a également contribué à l'évolution du génome. La présence d'introns peut avoir favorisé l'évolution de protéines nouvelles en facilitant la duplication et le brassage des exons dans le génome, comme nous le verrons maintenant.

Les réarrangements de parties de gènes : la duplication d'exons et le brassage d'exons

Au concept 17.3, vous avez vu qu'un exon code souvent pour un domaine, région structurale ou fonctionnelle distincte d'une protéine. Nous avons déjà vu qu'un enjambement inégal au cours de la méiose peut conduire à la duplication d'un gène sur un chromosome et à sa perte par le chromosome homologue (voir la figure 21.13). Selon un processus semblable, un exon donné dans un gène peut subir une duplication dans un chromosome et une délétion dans l'autre. Le gène dont l'exon a été dupliqué pourrait coder pour une protéine contenant une deuxième copie du domaine encodé. Cette modification de

▼ Figure 21.15 Comparaison de deux protéines: le lysozyme et l'α-lactalbumine. Cette figure présente les structures similaires **(a)** du lysozyme et **(b)** de l'α-lactalbumine, sous forme de modèles en ruban informatisés, et en **(c)** une comparaison des séquences d'acides aminés des deux protéines. Les acides aminés sont disposés en groupes de 10 pour faciliter la lecture, et désignés au moyen de codes à une seule lettre (voir la figure 5.14). Les acides aminés identiques sont surlignés en jaune, et les traits montrent les trous introduits par le logiciel dans une séquence pour optimiser l'alignement.

FAITES DES LIENS ► Même si les séquences d'acides aminés ne sont pas totalement identiques d'une protéine à une autre, la structure et les propriétés de ces acides aminés peuvent être similaires et, par conséquent, ils peuvent jouer des rôles comparables dans ces protéines. En vous reportant à la figure 5.14, examinez les acides aminés non identiques aux positions 1 à 30 et notez les cas où les acides aminés des deux séquences présentent une acidité ou une basicité similaire.



(a) Lysozyme **(b) α-lactalbumine**

Lysozyme	**1**	KVFERCELAR	TLKRLGMDGY	RGISLANWMC	LAKWESGYNT	RATNYNAGDR
α-lactalbumine	**1**	KQFTKCELSQ	LLK--DIDGY	GGIALPELIC	TMFHTSGYDT	QAIVENN--E
Lysozyme	**51**	STDYGIFQIN	SRYWCNDGKT	PGAVNACHLS	CSALLQDNIA	DAVACAKRVV
α-lactalbumine	**51**	STEYGLFQIS	NKLWCKSSQV	PQSRNICDIS	CDKFLDDDIT	DDIMCAKKIL
Lysozyme	**101**	RDPQGIRAWV	AWRNRCQ-NR	DVRQYVQGCG	V	
α-lactalbumine	**101**	D-IKGIDYWL	AHKALCT--E	KLEQWLCEKL	-	

(c) Alignement des séquences d'acides aminés du lysozyme et de l'α-lactalbumine

la structure de la protéine pourrait renforcer sa fonction en augmentant sa stabilité, en amplifiant sa capacité à se lier à un ligand particulier ou en altérant une autre propriété quelconque. Un bon nombre de gènes codant pour des protéines possèdent de multiples copies d'exons apparentés, qui sont probablement apparues par duplication suivie d'une divergence. Le gène qui code pour le collagène, protéine de la matrice extracellulaire, en est un bon exemple. Le collagène est une protéine de structure (voir la figure 5.18) dont la séquence d'acides aminés est très répétitive, ce qui se reflète dans le schéma répétitif des exons dans le gène du collagène.

On peut également imaginer le mélange occasionnel et l'appariement de différents exons soit dans un gène, soit entre deux gènes différents (non alléliques), à la suite d'erreurs survenues au cours de la recombinaison méiotique. Ce *brassage d'exons* pourrait conduire à la production de nouvelles protéines dotées d'une nouvelle combinaison de fonctions. Examinons, par exemple, le gène codant pour l'activateur tissulaire du plasminogène (tPA). Le tPA est une protéine extracellulaire qui aide à limiter la coagulation sanguine. Il possède quatre domaines de trois types, tous codés par un exon; un des exons est présent en deux exemplaires. Étant donné que chaque type d'exon se trouve aussi dans d'autres protéines, il se pourrait que la version actuelle du gène codant pour le tPA soit apparue à l'issue d'une série de brassages d'exons à la suite d'erreurs survenues pendant la recombinaison méiotique et la duplication subséquente d'un de ces exons (**figure 21.16**).

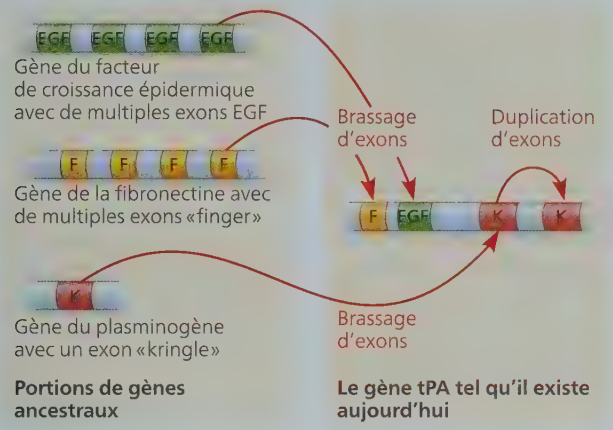
La contribution des éléments transposables à l'évolution du génome

La persistance des éléments transposables en tant que fraction substantielle de certains génomes eucaryotes est tout à fait compatible avec l'idée qu'ils jouent un rôle important dans la formation d'un génome au fil de l'évolution. Ces éléments peuvent contribuer de plusieurs façons à l'évolution du génome. Ils peuvent faciliter la recombinaison, dérégler les gènes cellulaires ou les éléments de contrôle et transporter des gènes entiers ou des exons individuels à de nouveaux emplacements.

Les éléments transposables de séquences similaires dispersées dans l'ensemble du génome permettent la recombinaison entre différents chromosomes (non homologues) en fournissant des régions homologues pour un enjambement (voir la figure 21.13). La plupart de ces recombinaisons sont probablement nuisibles, car elles provoquent des translocations chromosomiques et d'autres modifications dans le génome potentiellement létales. Mais, au fil de l'évolution, une telle recombinaison occasionnelle peut avoir des effets favorables sur l'organisme. (Évidemment, pour que la modification soit transmissible, elle doit se produire dans une cellule qui produit un gamète.)

Le déplacement d'un élément transposable peut avoir diverses conséquences directes. Par exemple, si un tel élément « saute » et s'insère au milieu d'une séquence d'un gène qui code pour une protéine, il empêchera la production d'un transcrit normal du

▼ Figure 21.16 L'évolution d'un nouveau gène par brassage d'exons. Des erreurs durant la méiose peuvent avoir déplacé les exons codant pour un domaine particulier de protéines différentes. À partir des formes ancestrales des gènes codant pour le facteur de croissance épidermique (EGF, pour *epidermal growth factor* en anglais), pour la fibronectine et pour le plasminogène (à gauche), des copies d'exons se sont déplacées vers le gène codant pour l'activateur tissulaire du plasminogène, tPA (à droite), qui s'est développé après le brassage et la duplication de ces exons. La duplication subséquente de l'exon «kringle» (K) du gène du plasminogène après son déplacement pourrait expliquer les deux copies de cet exon dans le gène tPA actuel.



HABILETÉS VISUELLES ▶ En vous reportant à la figure 21.13, décrivez les étapes par lesquelles les éléments transposables dans les introns ont facilité le brassage d'exons illustré ci-dessus.

gène. Les introns offrent une sorte de «zone de sécurité» qui n'exerce aucun effet sur le transcrit puisque l'élément transposable sera épissé. Par contre, si un élément transposable s'insère dans une séquence régulatrice, la transposition peut accroître ou réduire la production d'une ou de plusieurs protéines. Dans les grains de maïs de Barbara McClintock, la transposition était responsable des deux types d'effets sur les gènes codant pour les enzymes de synthèse des pigments, ce qui pouvait donner des grains plus pâles ou plus foncés. Nous avons évoqué antérieurement un autre exemple : celui d'un des éléments *Alu* qui produit des ARN régulant l'expression des gènes humains.

Au cours d'une transposition, un élément transposable peut déplacer un gène ou un groupe de gènes vers une nouvelle position dans le génome. Ce mécanisme est probablement responsable de la localisation des familles multigéniques de l'α-globine et de la β-globine sur différents chromosomes humains, de même que de la dispersion des gènes de certaines autres familles. En suivant un mouvement similaire de dispersion, un exon d'un gène peut s'insérer dans un autre gène grâce à un mécanisme s'apparentant à celui du brassage d'exons au cours de la recombinaison. Par exemple, un exon peut être inséré par transposition dans l'intron d'un gène codant pour une protéine. Si l'exon inséré est retenu dans le transcrit d'ARN au cours de l'épissage d'ARN, la protéine en voie de synthèse comportera un domaine additionnel qui peut lui conférer une nouvelle fonction.

Même si tous ces processus entraînent la plupart du temps des effets létaux, nuisibles ou même nuls, il peut dans quelques cas survenir de petites modifications avantageuses qui se transmettront de génération en génération. Il se crée alors une diversité génétique qui fournit plus de matière première pour la sélection naturelle. La diversification des gènes et de leurs produits est un facteur important dans l'évolution de nouvelles espèces. Par conséquent, l'accumulation des modifications du génome de chaque espèce fournit des archives de son histoire évolutive. Pour lire ces archives, il faut être en mesure de reconnaître les changements génomiques. La comparaison de génomes de différentes espèces nous permet de le faire et bonifie notre compréhension de l'évolution des génomes. Vous en apprendrez plus sur ces sujets dans la prochaine section.

RETOUR SUR LE CONCEPT 21.5

1. Décrivez trois exemples de processus cellulaires erronés qui aboutissent à des duplications d'ADN.
2. Expliquez comment des exons multiples peuvent être apparus dans les gènes ancestraux de l'EGF et de la fibronectine illustrés à gauche dans la figure 21.16.
3. Il semble que les éléments transposables contribuent de trois façons à l'évolution du génome. Quelles sont-elles ?
4. **ET SI ? ▶** En 2005, des scientifiques islandais ont rapporté la découverte d'une grande inversion chromosomique présente chez 20 % des Européens du Nord, et ils ont noté que les femmes islandaises porteuses de cette inversion avaient beaucoup plus d'enfants que les autres femmes. Selon vous, qu'arrivera-t-il à la fréquence de cette inversion dans la population islandaise chez les générations futures ?

Voir les réponses proposées à l'appendice A.

CONCEPT 21.6

La comparaison des séquences génomiques fournit des indices sur l'évolution et le développement

ÉVOLUTION Un chercheur a comparé l'état actuel de la biologie à l'Âge des découvertes (au 15e siècle), qui a été marqué par les améliorations majeures apportées aux instruments de navigation et à la conception des navires. Au cours des 30 dernières années, le séquençage des génomes et la collecte de données ont progressé à pas de géant. Dans la foulée, on a mis au point de nouvelles techniques pour évaluer l'activité génique dans le génome entier et conçu des stratégies raffinées pour comprendre comment les gènes et leurs produits agissent de concert dans des systèmes complexes. Il ne fait aucun doute que le domaine de la biologie est à l'aube d'un monde nouveau.

Les comparaisons entre les séquences génomiques issues d'espèces différentes nous en apprennent beaucoup sur l'histoire évolutive de la vie, de la très ancienne à la plus récente. Dans le même ordre d'idées, les études comparatives des programmes génétiques qui commandent le développement embryonnaire chez des espèces différentes commencent à éclairer les mécanismes à l'origine de la grande diversité des formes de vie présentes aujourd'hui. Dans la dernière section du chapitre, nous examinerons ce que ces deux approches nous ont révélé.

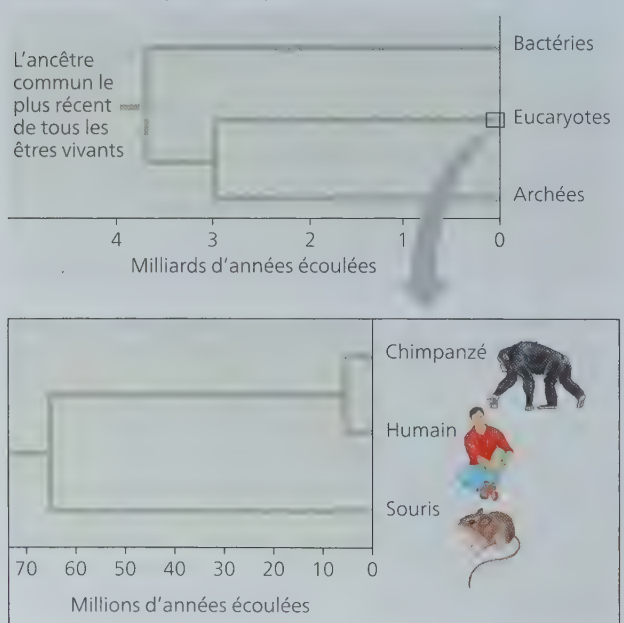
La comparaison des génomes

Deux espèces différentes sont d'autant plus étroitement apparentées dans leur histoire évolutive que les séquences de leurs gènes et de leurs génomes sont semblables. En effet, le peu de temps écoulé n'a pas encore permis aux mutations et aux autres changements de s'accumuler. La comparaison de génomes d'espèces étroitement apparentées a apporté un éclairage sur les événements plus récents de l'évolution, alors que la comparaison entre les génomes des espèces très distantes nous aide à comprendre l'histoire évolutive ancienne. Dans les deux cas, le fait de connaître les caractéristiques communes ou distinctives des groupes améliore notre compréhension de l'évolution des organismes et des processus biologiques. Comme vous l'avez appris au concept 1.2, on peut représenter les relations entre les espèces au cours de l'évolution par un diagramme arborescent où chaque point d'embranchement marque la divergence de deux lignées. La **figure 21.17** montre les relations, au cours de l'évolution, de quelques groupes et espèces que nous allons examiner.

La comparaison entre espèces distantes

Il est possible de faire la lumière sur les relations évolutives parmi les espèces qui ont divergé les unes des autres il y a longtemps en déterminant quels gènes sont restés semblables (autrement dit, *hautement conservés*) parmi des espèces distantes. En effet, les comparaisons effectuées entre les séquences de gènes spécifiques de bactéries, d'archées et d'eucaryotes indiquent que ces trois groupes ont divergé il y a entre 2 et 4 milliards d'années et confirment de façon convaincante qu'ils constituent bel et bien les trois domaines fondamentaux du monde vivant (voir la figure 21.17).

▼ Figure 21.17 Les relations au cours de l'évolution entre les trois domaines de la vie. Le diagramme en arborescence du haut illustre la divergence ancienne des bactéries, des archées et des eucaryotes. Dans le médaillon, la portion de la lignée des eucaryotes montre la divergence la plus récente de trois espèces de mammifères examinées dans le présent chapitre.



En plus de leur intérêt en biologie de l'évolution, les études comparatives génomiques montrent également que les recherches menées sur des organismes modèles contribuent à mieux comprendre la biologie en général et la biologie humaine en particulier. Il est surprenant de constater combien des gènes très anciens peuvent être similaires chez des espèces disparates. En 2015, on a réalisé une étude pour évaluer si la version humaine de 414 gènes jouant un rôle important chez les levures fonctionnait de façon comparable dans les cellules de levure. Étonnamment, les chercheurs en sont venus à la conclusion que 47 % des gènes de levure pouvaient être remplacés par leur équivalent humain. Ce résultat frappant souligne l'origine commune des levures et des humains, deux espèces distantes.

La comparaison entre espèces étroitement apparentées

Les structures des génomes de deux espèces étroitement apparentées sont probablement similaires en raison de leur divergence relativement récente. Une longue histoire commune explique également le petit nombre de différences entre les gènes que révèle la comparaison de leurs génomes. Il est alors plus facile d'établir des corrélations entre les différences génétiques particulières et les variantes phénotypiques des deux espèces. Ce type d'analyse s'avère fort utile pour les chercheurs qui veulent comparer le génome humain avec les génomes du chimpanzé, de la souris, du rat et d'autres mammifères. En identifiant les gènes communs chez ces espèces, qui sont toutes des mammifères, on devrait obtenir des indices sur ce qui caractérise un mammifère, alors qu'en distinguant les gènes communs uniquement chez les chimpanzés et les humains, donc en excluant les rongeurs, on devrait en connaître plus sur les primates. Et, bien sûr, la comparaison du génome humain avec celui du chimpanzé nous aide à répondre à la question absolument fascinante que nous avons posée au début du présent chapitre : quelles différences génomiques distinguent un humain d'un chimpanzé ?

Une analyse de la composition globale des génomes de l'humain et du chimpanzé, qui semblent avoir divergé il y a seulement environ 6 millions d'années (voir la figure 21.17), révèle quelques différences d'ordre général. Quand on examine les substitutions d'un seul nucléotide, on constate que les génomes ne diffèrent que par 1,2 %. Cependant, lorsque les chercheurs ont étudié des segments d'ADN plus longs, ils ont été surpris de trouver une différence supplémentaire de 2,7 %, en raison des insertions ou des délétions de plus grandes régions du génome de l'une ou l'autre espèce ; un grand nombre des insertions étaient des duplications ou d'autre ADN répétitif. En fait, un tiers des duplications observées chez les humains sont absentes du génome des chimpanzés, et certaines de ces duplications contiennent des régions associées à des maladies humaines. Il y a plus d'éléments *Alu* dans le génome humain que dans celui du chimpanzé, et ce dernier contient de nombreuses copies d'un provirus rétroviral absent chez les humains. Toutes ces observations fournissent des indices concernant les forces qui ont dû entraîner les deux génomes dans des directions différentes, mais notre compréhension du mécanisme de cette divergence est encore incomplète.

À l'instar des chimpanzés, les bonobos sont une espèce de singes africains étroitement apparentés aux humains. Le séquençage du génome des bonobos, achevé en 2012, a démontré que

dans certaines régions, les séquences humaines et celles du chimpanzé ou du bonobo présentaient une plus grande similarité que celle observée entre les séquences du chimpanzé et celles du bonobo. Une comparaison aussi exhaustive de trois espèces étroitement apparentées permet de reconstituer leur histoire évolutive de façon beaucoup plus précise.

Nous ignorons dans quelle mesure les caractéristiques distinctes de chaque espèce dépendent des différences génétiques mises en évidence par le séquençage du génome. Afin de découvrir le fondement des différences phénotypiques entre les humains et les chimpanzés, les biologistes étudient des gènes spécifiques et des types de gènes qui distinguent les deux espèces ; ils les comparent ensuite avec leurs contreparties chez d'autres mammifères. Les résultats obtenus montrent qu'un certain nombre de gènes changent (évoluent) apparemment plus rapidement chez l'humain que chez le chimpanzé ou la souris. C'est le cas notamment des gènes intervenant dans la défense contre le paludisme et la tuberculose et au moins d'un gène qui commande la grosseur du cerveau. Quand on classe les gènes par fonction, on s'aperçoit que les gènes codant pour des facteurs de transcription semblent évoluer plus rapidement que tous les autres. Cette observation est logique parce que les facteurs de transcription assurent la régulation de l'expression génétique et jouent donc un peu le rôle de chef d'orchestre du programme génétique dans son ensemble.

Le gène d'un facteur de transcription, *FOXP2* (**figure 21.18**), témoigne d'une modification rapide dans la lignée humaine. Selon plusieurs sources de données, le produit du gène *FOXP2* régulerait les gènes intervenant dans la vocalisation chez les vertébrés. D'une part, des mutations de ce gène peuvent provoquer de graves troubles de la parole et du langage chez les humains. D'autre part, le gène *FOXP2* est exprimé dans les cerveaux des diamants mandarins et des canaris au moment où ces oiseaux chanteurs apprennent leurs chants. Mais la preuve la plus convaincante vient peut-être d'une expérience d'invalidation génétique (*knockout*) au cours de laquelle les chercheurs ont neutralisé le gène *FOXP2* chez la souris et ont analysé le phénotype produit (voir la figure 21.18). La souris homozygote mutante présentait des malformations du cerveau et était incapable d'émettre des vocalisations ultrasoniques normales ; par ailleurs, des souris possédant une copie défectueuse du gène ont également présenté des problèmes importants de vocalisation. Ces résultats corroborent l'idée selon laquelle le produit du gène *FOXP2* active les gènes qui jouent un rôle dans la vocalisation.

Plus récemment, un autre groupe de recherche a apporté de nouveaux éclaircissements sur la question ; les chercheurs ont remplacé chez des souris le gène *FOXP2* par une copie « humanisée » codant pour la version humaine de la protéine (deux acides aminés diffèrent entre l'humain et le chimpanzé), qui serait responsable de la capacité de parler des humains. Bien que les souris aient été généralement en bonne santé, leurs vocalisations étaient légèrement différentes ; elles présentaient également des modifications des cellules du cerveau dans une région reconnue comme jouant un rôle dans le langage chez les humains.

En 2010, on a séquencé le génome du néandertalien à partir d'une très petite quantité d'ADN génomique préservé et, en 2014, on disposait d'une séquence de bonne qualité. Le néandertalien (*Homo neanderthalensis*) et l'être humain (*Homo sapiens*) appartiennent tous deux au même genre et sont donc relativement proches parents évolutifs (voir le concept 34.7). La reconstitution de leur histoire évolutive, d'après la comparaison du génome des deux espèces, laisse croire que certains groupes d'humains et de néandertaliens ont coexisté et se sont croisés pendant un certain temps avant que le néandertalien ne s'éteigne, il y a environ 30 000 ans. Bien que le néandertalien ait parfois été dépeint comme un être primitif qui ne pouvait qu'émettre des grognements, la séquence du gène *FOXP2* découverte chez lui code pour une protéine identique à celle observée chez l'humain. Il est donc possible qu'il ait été doté d'une certaine forme de langage. Cette capacité, ainsi que d'autres similitudes génétiques, exige de réévaluer notre perception des espèces éteintes qui nous sont apparentées.

L'histoire du gène *FOXP2* est un excellent exemple illustrant comment des approches différentes peuvent se compléter en révélant des phénomènes biologiques d'une importance générale. Les expériences sur le gène *FOXP2* ont porté sur des souris comme modèles pour les humains parce qu'il aurait été contraire à l'éthique d'effectuer de telles expériences chez ceux-ci, sans compter que cela n'aurait pas été pratique. Les souris et les humains ont divergé il y a environ 65,5 millions d'années (voir la figure 21.17) et ont en commun environ 85 % de leurs gènes. Il est possible d'exploiter cette similitude génétique dans l'étude des troubles génétiques. Si les chercheurs connaissent l'organe ou le tissu qui est atteint par un trouble génétique, ils peuvent chercher les gènes qui sont exprimés à ces emplacements chez les souris.

Même si elle est plus éloignée de l'espèce humaine, la drosophile est une espèce utile qui peut servir de modèle pour étudier certains troubles chez les humains, comme la maladie de Parkinson et l'alcoolisme. Quant aux nématodes (vers), ils ont permis de recueillir un vaste éventail de données sur le vieillissement. D'autres travaux de recherche sont en cours pour étendre les études génomiques à beaucoup plus d'espèces, notamment des espèces délaissées dans diverses branches de l'arbre de la vie. Ces études feront progresser notre connaissance de l'évolution, bien entendu, mais également de tous les aspects de la biologie, y compris la santé humaine et l'écologie.

La comparaison des génomes au sein d'une même espèce

Notre capacité à analyser des génomes entraîne une autre conséquence intéressante : elle accroît notre compréhension du spectre des variations génétiques chez les humains. Étant donné la brève histoire de l'espèce humaine (probablement autour de 200 000 ans), le nombre de variations de l'ADN chez les humains est faible en comparaison de celles qu'on observe chez de nombreuses autres espèces. Une bonne part de notre diversité semble résulter de polymorphismes mononucléotidiques (les SNP). Les SNP correspondent à un site d'une seule paire de bases présentant une variation chez au moins 1 % de la population (voir le concept 20.2) ; ils sont habituellement détectés par le séquençage de l'ADN. Dans le génome humain, les SNP se produisent en moyenne environ une fois par 100 à 300 paires de bases. Les scientifiques ont déjà repéré l'emplacement de plusieurs millions de sites SNP dans le génome humain et continuent d'en découvrir d'autres. Ils sont répertoriés dans différentes bases de données à l'échelle mondiale. L'une de ces bases de données est gérée par le National Center for Biotechnology Information (NCBI), et il est possible de la consulter à l'adresse www.ncbi.nlm.nih.gov/SNP/.

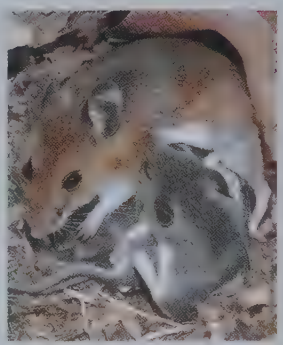
DÉMARCHE SCIENTIFIQUE

INVESTIGATION

Quelle est la fonction d'un gène (*FOXP2*) qui évolue rapidement dans la lignée humaine ?*

HYPOTHÈSE Si le gène *FOXP2* a pour fonction la phonation chez l'humain, il devrait avoir une fonction similaire chez d'autres espèces de mammifères, telles les souris.

EXPÉRIENCE Plusieurs sources de données confirment le rôle du gène *FOXP2* dans le développement de la parole et du langage chez les humains et de la vocalisation chez d'autres vertébrés. En 2005, Joseph Buxbaum et ses collaborateurs de la Mount Sinai School of Medicine et de plusieurs autres institutions ont testé la fonction du gène *FOXP2*. Ils ont utilisé la souris, un organisme modèle dont les gènes peuvent facilement être neutralisés et qui est représentatif des vertébrés qui vocalisent : les souris émettent des cris ultrasoniques (sifflements) pour manifester leur stress. Les chercheurs ont utilisé le génie génétique pour produire des souris chez lesquelles un ou les deux exemplaires de *FOXP2* ont été bloqués.



Type sauvage : deux exemplaires normaux de *FOXP2*

Hétérozygote : un exemplaire de *FOXP2* interrompu

Homozygote : les deux exemplaires de *FOXP2* interrompus

Puis, ils ont comparé les phénotypes de ces souris. Voici les résultats concernant deux des caractères qu'ils ont examinés : l'anatomie du cerveau et la vocalisation.

EXPÉRIENCE 1 Les chercheurs ont fait des coupes fines des sections du cerveau et les ont colorées avec des réactifs qui permettent de visualiser l'anatomie du cerveau au moyen d'un microscope à fluorescence (UV).

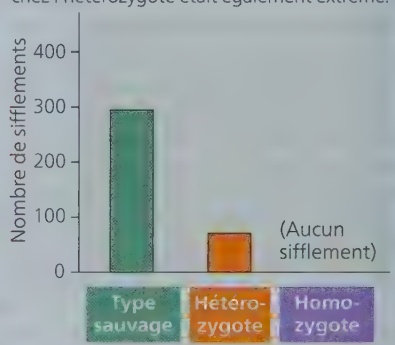
RÉSULTATS

Expérience 1 : Le blocage des deux exemplaires de *FOXP2* a provoqué des anomalies du cerveau dans lesquelles les cellules ont été désorganisées. Les effets phénotypiques sur le cerveau des hétérozygotes, avec un exemplaire bloqué, ont été moins graves. (Dans les micrographies ci-dessous, chaque couleur révèle une cellule ou un type de tissu différents.)

Type sauvage | Hétérozygote | Homozygote

EXPÉRIENCE 2 Pour provoquer un stress, les chercheurs ont séparé chaque souriceau de sa mère dès sa naissance et ont relevé le nombre de sifflements ultrasoniques qu'il produisait.

Expérience 2 : Le blocage des deux exemplaires de *FOXP2* a provoqué une absence de vocalisation ultrasonique en réaction au stress. L'effet sur la vocalisation chez l'hétérozygote était également extrême.



CONCLUSION *FOXP2* joue un rôle déterminant dans le développement des systèmes fonctionnels de communication chez la souris. Les résultats s'ajoutent à la preuve fournie par les études effectuées sur les oiseaux et les humains, corroborant l'hypothèse que *FOXP2* agirait de façon similaire chez divers organismes.

Source des données : W. Shu et coll., Altered ultrasonic vocalization in mice with a disruption in the *FOXP2* gene, *Proceedings of the National Academy of Sciences* 102 : 9643-9648 (2005).

ET SI ? ▶ Étant donné que les résultats confirment le rôle du gène *FOXP2* de la souris dans la vocalisation, on peut se demander si la protéine FOXP2 humaine est un régulateur clé de la parole. En supposant que vous connaissez les séquences d'acides aminés des protéines FOXP2 humaines du type sauvage et mutant et de la protéine FOXP2 du chimpanzé de type sauvage, comment pourriez-vous étudier cette question ? Quels autres indices pourriez-vous obtenir en comparant ces séquences à celles de la protéine FOXP2 de la souris ?

* Cette recherche a été effectuée il y a plus d'une décennie, et la façon dont on traite les animaux en recherche a beaucoup évolué depuis. Les protocoles doivent maintenant être approuvés par des comités d'éthique qui veillent à ce que la souffrance animale soit évitée ou à tout le moins atténuée le plus possible. Cependant, des expériences passées menées selon des protocoles qui ne seraient probablement pas approuvés de nos jours fournissent parfois de précieuses informations qu'il ne faut pas négliger.

Au cours de cette recherche, les scientifiques ont également trouvé d'autres variations, notamment des inversions, des délétions et des duplications dans certaines régions chromosomiques. Mais la découverte la plus surprenante a été l'importante occurrence du polymorphisme associé au nombre de copies d'un gène particulier ou d'une région génétique. Certains individus ont un plus grand nombre de copies d'une séquence donnée d'ADN, au lieu des deux copies normales (une sur chaque chromosome homologue). La variabilité du nombre de copies est due à des duplications ou à des délétions qui se sont produites de façon désordonnée au sein de la population. Une étude réalisée sur 40 personnes a permis de découvrir plus de 8 000 différences quant au nombre de copies mettant en cause 13 % des gènes du génome. Il se peut que ces différences ne représentent qu'un petit sous-ensemble du total. Étant donné que ces différences incluent des segments d'ADN beaucoup plus longs que les nucléotides uniques des SNP, la variabilité du nombre de copies est plus susceptible d'avoir des conséquences phénotypiques et de jouer un rôle dans les maladies et les affections. À tout le moins, l'incidence élevée de la variabilité du nombre de copies sème le doute quant à la signification de l'expression « génome humain normal ».

Les variances en nombre de copies, les SNP et les variations dans l'ADN répétitif comme les courtes répétitions en tandem (STR) sont des marqueurs génétiques utiles pour étudier l'évolution humaine. Dans une étude, les génomes de deux Africains provenant de communautés différentes ont été séquencés. Le premier était l'archevêque Desmond Tutu, le défenseur des droits civils sud-africain et membre de la tribu bantoue, la population majoritaire en Afrique du Sud ; le second était un chasseur-cueilleur du nom de !Gubi, de la communauté khoisan de Namibie, une population minoritaire en Afrique qui est probablement la plus ancienne branche de la lignée humaine. La comparaison a révélé de nombreuses différences, comme on pouvait s'y attendre. L'analyse a alors été élargie pour comparer les régions du génome de !Gubi qui codent pour des protéines avec celles de trois autres Khoisans (qui se sont déclarés Bochimans) vivant à proximité. On a trouvé plus de différences entre les génomes de ces quatre Africains qu'entre ceux d'un Européen et d'un Asiatique. Ces observations illustrent la très grande diversité génétique parmi les génomes africains. À mesure que s'étendront ces travaux de comparaison, nous serons de plus en plus aptes à répondre à des questions importantes concernant les différences entre les populations humaines et leurs voies de migration au fil du temps.

La conservation généralisée des gènes du développement chez les animaux

Les biologistes spécialistes du champ disciplinaire de la biologie de l'évolution du développement (surnommé *évo-dévo*) comparent les processus de développement de divers organismes multicellulaires. Ils cherchent à comprendre comment ces mécanismes sont apparus et comment des changements qui les touchent peuvent modifier les caractéristiques existantes d'un organisme ou en créer de nouvelles. L'avènement des techniques de la biologie moléculaire et le récent déluge de données en génomique nous révèlent que les génomes d'espèces apparentées, mais qui se distinguent de façon étonnante par leurs formes, peuvent ne présenter que des différences mineures dans la séquence ou, plus important encore, dans la régulation des gènes. Par ailleurs, la découverte du fondement moléculaire de ces différences nous aidera à comprendre les origines de la myriade de formes diverses qui cohabitent sur Terre, ce qui constitue un apport d'information à notre étude de l'évolution de la vie.

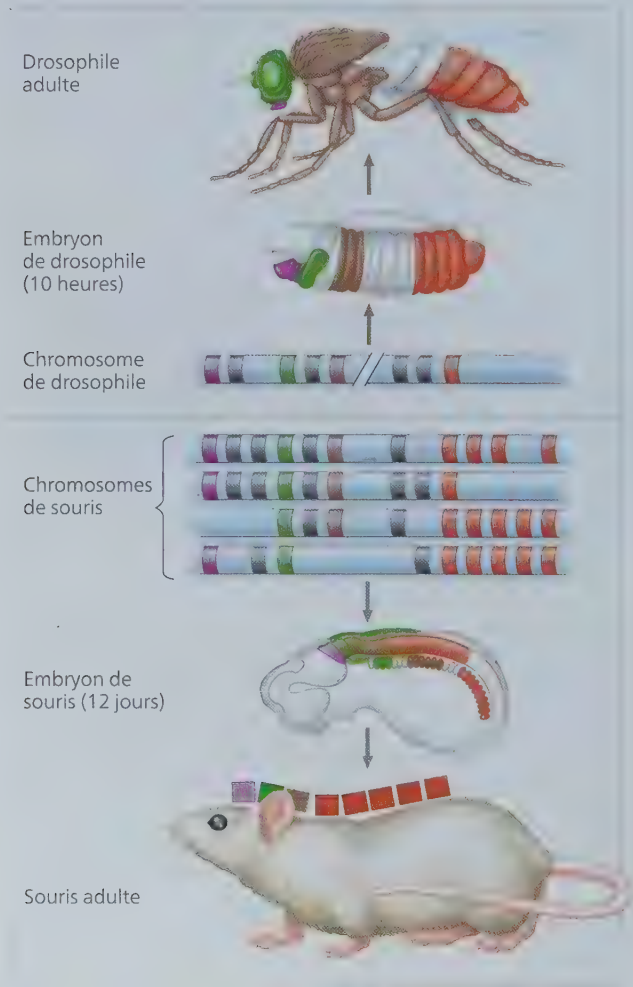
Le concept 18.4 traitait des gènes homéotiques de *Drosophila melanogaster* (voir la figure 18.20), qui codent pour des facteurs de transcription assurant la régulation de l'expression des gènes et pour l'identité des segments corporels dans la drosophile. L'analyse moléculaire des gènes homéotiques a montré qu'ils comprennent tous une séquence de 180 nucléotides nommée **boîte homéotique**. Celle-ci code pour un *domaine homéotique* de 60 acides aminés dans les protéines encodées. On a découvert une séquence nucléotidique identique ou très semblable dans les gènes homéotiques de nombreux vertébrés et invertébrés. En fait, la ressemblance entre les séquences nucléotidiques des humains et des drosophiles est tellement surprenante qu'un chercheur a déclaré qu'il considérait les drosophiles comme des « petites personnes dotées d'ailes ». La ressemblance s'étend même à l'organisation de ces gènes : chez les vertébrés, les gènes homologues aux gènes homéotiques ont conservé le même ordre qu'ils occupaient sur les chromosomes de drosophiles (**figure 21.19**). On a également trouvé des séquences contenant une boîte homéotique dans les gènes régulateurs d'eucaryotes beaucoup plus éloignés, dont des végétaux et des levures. Ces ressemblances nous permettent de conclure que la séquence d'ADN de la boîte homéotique est apparue très tôt au cours de l'histoire de la vie ; de plus, elle doit être assez précieuse pour avoir été conservée à peu près intacte chez les animaux et les végétaux durant des centaines de millions d'années.

Chez les animaux, les gènes homéotiques sont nommés gènes *Hox* (une abréviation qui réfère à *Homeobox*, en anglais), parce que les gènes homéotiques ont été les premiers gènes trouvés qui présentaient cette séquence. Plus tard, on a repéré d'autres gènes à boîte homéotique, mais qui n'agissent pas comme des gènes homéotiques, c'est-à-dire qu'ils ne déterminent pas directement l'identité des parties de l'organisme. Cependant, la plupart sont liés au développement, du moins chez les animaux, ce qui laisse penser qu'ils jouent un rôle fondamental dans ce processus depuis des temps reculés. Par exemple, chez la drosophile, les boîtes homéotiques sont présentes non seulement dans les gènes homéotiques, mais aussi dans les gènes *bicoïd*, qui régissent la polarité de l'œuf (voir les figures 18.21 et 18.22), dans plusieurs gènes de segmentation et dans le gène régulateur principal du développement de l'œil.

Les chercheurs ont découvert que le domaine homéotique se lie à l'ADN lorsque la protéine agit en tant que facteur de transcription. Ailleurs dans la protéine, les domaines plus variables interagissent avec d'autres facteurs de transcription, ce qui fait que la protéine contenant un domaine homéotique reconnaît certains amplificateurs et régule les gènes associés. Les protéines contenant un domaine homéotique assurent probablement la régulation du développement en coordonnant la transcription d'un ensemble de gènes du développement qu'elles activent ou désactivent. Chez la drosophile et d'autres espèces animales, diverses combinaisons de gènes à boîte homéotique sont actives dans les différentes parties de l'embryon. L'expression sélective des gènes régulateurs et les fluctuations de cette expression dans le temps et dans l'espace sont essentielles à la réalisation des plans d'organisation corporelle.

Les biologistes du développement ont découvert qu'en plus des gènes homéotiques, de nombreux autres gènes participant au développement sont très bien conservés d'une espèce à l'autre. La plupart de ceux-ci codent pour les composants des voies de signalisation. L'extraordinaire ressemblance entre les gènes de développement particuliers chez diverses espèces animales suscite la question suivante : comment les mêmes gènes peuvent-ils jouer un rôle dans le développement des animaux dont les formes diffèrent tellement d'une espèce à l'autre ?

▼ Figure 21.19 La conservation de gènes homéotiques chez la drosophile et la souris. Les gènes homéotiques commandant la forme des structures antérieures et postérieures de l'organisme sont placés dans le même ordre sur les chromosomes de la drosophile et de la souris. Chacune des bandes colorées qui figurent ici sur les chromosomes désigne un gène homéotique. Chez la drosophile, tous ces gènes se situent sur le même chromosome. Chez la souris et les autres mammifères, on trouve le même ensemble de gènes ou des ensembles similaires sur quatre chromosomes. Les couleurs renvoient aux parties de l'embryon où ces gènes s'expriment et aux régions correspondantes de l'organisme adulte. On constate que la disposition des gènes sur le chromosome reflète fidèlement la disposition des structures de l'animal sur lesquelles ils agissent. Tous ces gènes sont presque identiques chez les drosophiles et les souris, excepté ceux qui sont représentés par des bandes noires ; ces derniers se ressemblent moins chez les deux espèces.



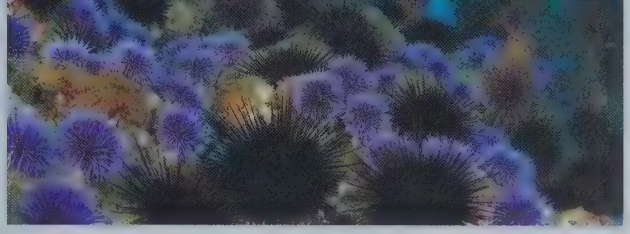
Les études en cours semblent indiquer les éléments de réponse suivants. Dans certains cas, des changements minimes dans les séquences de régulation de gènes particuliers causent des transformations des modes d'expression génétique qui peuvent mener à des modifications majeures de la forme d'un organisme. Par exemple, les divers modes d'expression des gènes *Hox* le long de l'axe corporel des insectes et des crustacés peuvent expliquer la variation du nombre de segments porteurs de pattes chez ces animaux étroitement apparentés (**figure 21.20**). Dans d'autres cas, des gènes similaires commandent des processus différents de développement chez les divers organismes, ce qui entraîne la diversité des formes du corps. Ainsi, plusieurs gènes *Hox* sont exprimés aux stades embryonnaire ou larvaire des oursins, des animaux non segmentés qui ont un plan d'organisation corporelle très différent de celui des insectes et des souris. Les oursins parvenus à l'âge adulte fabriquent leur coquille ayant la forme d'une pelote à épingles ; la photo à la page suivante présente deux espèces d'oursins vivants. Les oursins font partie des organismes utilisés depuis longtemps dans les études d'embryologie classique (voir le concept 47.2).

Dans ce dernier chapitre de la partie portant sur la génétique, nous avons appris comment les études de la composition génomique et la comparaison des génomes de différentes espèces mettent en évidence les mécanismes de l'évolution des génomes. De plus, en comparant les programmes de développement, il est possible de constater que l'unité de la vie se révèle dans la similitude des mécanismes moléculaires et cellulaires qui servent à

▼ Figure 21.20 L'effet des différences dans l'expression du gène *Hox* au cours du développement des crustacés et des insectes. Des changements dans les modes d'expression des gènes *Hox* se sont produits au cours de l'évolution, depuis la divergence des insectes d'un ancêtre crustacé. Ces changements expliquent en partie les plans d'organisation corporelle différents **(a)** de la crevette des salines, *Artemia salina*, un crustacé, et **(b)** de la sauterelle, un insecte. L'illustration montre en couleurs distinctes les régions du corps de l'animal adulte correspondant à l'expression de quatre gènes *Hox* qui déterminent la formation de parties particulières du corps au cours du développement de l'embryon. Chaque couleur représente un gène *Hox* spécifique.

(a) Expression de quatre gènes *Hox* dans la crevette des salines *Artemia salina*. Trois des gènes *Hox* sont exprimés ensemble dans une seule région (indiquée par les rayures) et déterminent quels segments sont dotés de pléopodes. Le quatrième gène (en bleu-vert) détermine quels sont les segments génitaux.

(b) Expression, chez la sauterelle, des quatre mêmes gènes *Hox*. Chez la sauterelle, chaque gène *Hox* est exprimé dans une zone distincte et détermine la nature de cette zone.



établir le plan d'organisation corporelle, bien que les gènes commandant le développement puissent différer parmi les organismes. Ces ressemblances entre les génomes reflètent l'existence d'ancêtres communs de la vie sur Terre. Mais les différences jouent aussi un rôle essentiel, parce que ce sont elles qui ont fait apparaître l'extraordinaire diversité des organismes actuels. Le reste du présent ouvrage va au-delà des molécules, des cellules et des gènes. Il vous amènera à explorer le vivant au niveau des organismes et de leur environnement.

RETOUR SUR LE CONCEPT 21.6

1. Doit-on s'attendre à ce que le génome du macaque (un singe) ressemble plus au génome de la souris ou à celui de l'humain ? Expliquez votre réponse.
2. Les boîtes homéotiques sont des séquences d'ADN qui assistent le développement embryonnaire gouverné par les gènes homéotiques. Comme elles sont communes aux drosophiles et aux souris, expliquez pourquoi ces animaux ne se ressemblent pas davantage.
3. **ET SI ? ▶** Le génome humain comporte trois fois plus d'éléments *Alu* que celui du chimpanzé. Selon vous, comment ces éléments *Alu* supplémentaires sont-ils apparus dans le génome humain ? Proposez un rôle qu'ils pourraient avoir joué dans la divergence de ces deux espèces.

Voir les réponses proposées à l'appendice A.

RÉVISION DU CHAPITRE 21

Consultez votre MANUEL NUMÉRIQUE, qui vous donne accès aux **animations**, aux **exercices** et à la plateforme d'**anatomie interactive**.

Résumé des concepts clés

CONCEPT 21.1

Le projet Génome humain a favorisé la mise au point de techniques de séquençage plus rapides et moins onéreuses (p. 484 à 485)

- Favorisé par d'importantes avancées réalisées dans les technologies de séquençage, le **projet Génome humain** était en grande partie terminé en 2003.
- Dans l'approche de **séquençage en aveugle sur l'ensemble du génome**, le génome entier est découpé en un grand nombre de petits fragments dont les extrémités se chevauchent et qui sont séquencés. Après quoi, un programme informatique reconstitue la séquence complète.

? Pourquoi le projet Génome humain a-t-il permis de mettre au point des techniques de séquençage de l'ADN plus rapides et moins onéreuses ?

CONCEPT 21.2

Les scientifiques utilisent la bio-informatique pour analyser les génomes et leurs fonctions (p. 485 à 489)

- L'analyse informatique des séquences génomiques aide à l'**annotation d'un gène**, une opération qui consiste à identifier des séquences codant pour des protéines. Les méthodes pour déterminer la fonction d'un gène comprennent la comparaison des séquences des gènes nouvellement découverts avec celles de gènes connus dans d'autres espèces et l'observation des effets de l'inactivation expérimentale de gènes.
- Dans la **biologie des systèmes**, les scientifiques utilisent les outils informatiques de la **bio-informatique** pour comparer les génomes et étudier les jeux de gènes et de protéines comme des systèmes entiers (**génomique** et **protéomique**). Les études comprennent les analyses à grande échelle des interactions des protéines, les éléments de l'ADN fonctionnel et les gènes qui sont à l'origine de certains troubles médicaux.

? Quelle a été la découverte la plus importante du projet pilote ENCODE ? Pourquoi le projet a-t-il été étendu à des espèces autres que l'espèce humaine ?

CONCEPT 21.3

Les génomes varient en taille, en nombre de gènes et en densité génique (p. 489 à 491)

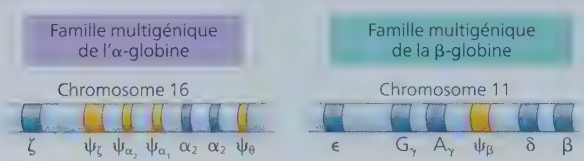
	Bactéries	Archées	Eucaryotes
Taille du génome	1 à 6 Mb (pour la plupart)		Entre 10 et 4 000 Mb pour la plupart, mais quelques-uns sont beaucoup plus gros
Nombre de gènes	1 500 à 7 500		Pour la plupart, 5 00 à 45 000
Densité génique	Plus élevée que chez les eucaryotes		Plus faible que chez les procaryotes (chez les eucaryotes, une densité plus faible est associée à des génomes plus gros)
Introns	Aucun dans les gènes codant pour des protéines	Présents dans certains gènes	Présents dans la plupart des gènes des eucaryotes multicellulaires, mais prévalant seulement dans quelques gènes des eucaryotes unicellulaires
Autres ADN non codants	Très peu		Parfois présents en grandes quantités ; généralement plus d'ADN non codant répétitif chez les eucaryotes multicellulaires

? Comparez la taille des génomes, le nombre de gènes et la densité génique (a) dans les trois domaines et (b) parmi les eucaryotes.

CONCEPT 21.4

Les eucaryotes multicellulaires possèdent beaucoup d'ADN non codant et de nombreuses familles multigéniques (p. 491 à 495)

- Seulement 1,5 % du génome humain code pour des protéines ou donne naissance à des ARNr ou à des ARNt ; le reste est de l'ADN non codant, incluant les **pseudogènes** et l'**ADN répétitif** de fonction inconnue.
- Le type le plus abondant d'ADN répétitif chez les eucaryotes multicellulaires se compose d'**éléments transposables** et de séquences apparentées. Chez les eucaryotes, il y a deux types d'éléments transposables : les **transposons**, qui se déplacent ou se copient par l'intermédiaire d'un ADN, et les **rétrotransposons**, qui sont plus nombreux et qui se déplacent ou se copient par l'intermédiaire d'un ARN.
- L'autre ADN répétitif inclut des séquences courtes non codantes qui sont répétées en tandem des milliers de fois (**ADN de simple séquence**, qui comprend les **STR**) ; ces séquences dominent dans les centromères et les télomères, où elles jouent probablement des rôles structuraux au sein des chromosomes.
- Bien que de nombreux gènes eucaryotes soient présents dans un exemplaire par jeu de chromosomes haploïdes, il arrive que d'autres (la plupart, chez certaines espèces) constituent des membres d'une famille de gènes ; c'est le cas des familles de globines humaines :



? Selon vous, comment la fonction des éléments transposables pourrait-elle expliquer leur prévalence dans l'ADN non codant humain ?

CONCEPT 21.5

Les duplications, les réarrangements et les mutations de l'ADN contribuent à l'évolution du génome (p. 495 à 501)

- Des erreurs survenant durant la division cellulaire peuvent donner naissance à des copies supplémentaires d'une partie ou de l'ensemble des jeux de chromosomes ; les gènes dans le ou les jeux supplémentaires peuvent alors diverger si un jeu accumule des modifications de séquences. La polyploïdie, qui est plus fréquente chez les végétaux que chez les animaux, intervient dans la spéciation.
- La comparaison de la structure chromosomique des génomes de différentes espèces fournit de l'information sur les relations au cours de l'évolution. Il se peut que les réarrangements de chromosomes au sein d'une espèce donnée aient contribué à l'émergence de nouvelles espèces.
- Les gènes codant pour les diverses globines apparentées ont évolué à partir d'un gène ancestral commun de la globine qui a subi une duplication et une divergence en gènes ancestraux des α-globines et des β-globines. Des duplications subséquentes de ces gènes et des mutations au hasard ont donné naissance aux gènes actuels des globines, qui codent tous pour des protéines qui se lient à l'oxygène. Les copies de quelques gènes dupliqués ont divergé au cours de l'évolution à un point tel que les fonctions de leurs protéines encodées (comme le lysozyme et l'α-lactalbumine) sont maintenant passablement différentes.
- Les remaniements d'exons dans et entre les gènes au cours de l'évolution ont produit des gènes contenant de multiples copies d'exons semblables ou plusieurs exons différents dérivés d'autres gènes.
- Le déplacement d'éléments transposables ou la recombinaison entre des copies du même élément peut engendrer des combinaisons de nouvelles séquences qui sont favorables à l'organisme. Ces mécanismes peuvent altérer les fonctions des gènes ou leurs modes d'expression et de régulation.

? Comment le réarrangement chromosomique peut-il mener à l'émergence de nouvelles espèces ?

CONCEPT 21.6

La comparaison des séquences génomiques fournit des indices sur l'évolution et le développement (p. 501 à 507)

- La comparaison des génomes provenant d'espèces très divergentes et étroitement apparentées fournit de l'information précieuse sur l'histoire de l'évolution au cours des temps plus anciens et des périodes plus récentes, respectivement. On peut aussi obtenir de l'information sur l'évolution d'une espèce au moyen de l'analyse des polymorphismes mononucléotidiques (SNP) et des différences en nombre de copies parmi les individus de cette espèce.
- Les biologistes spécialistes du champ disciplinaire de la biologie de l'évolution du développement (*évo-dévo*) ont montré que les gènes homéotiques et quelques autres gènes associés au développement des animaux contiennent une **boîte homéotique** dont la séquence est hautement conservée chez diverses espèces animales. Des séquences apparentées sont présentes dans les gènes de végétaux et de levures.

? Quel type d'information peut-on obtenir en comparant les génomes d'espèces étroitement apparentées ? D'espèces très distantes ?

Évaluation

NIVEAU 1 : CONNAISSANCES ET COMPRÉHENSION

1. La bio-informatique inclut tous les éléments suivants, sauf :
   a) l'utilisation de programmes informatiques pour aligner les séquences d'ADN.
   b) l'utilisation de la biotechnologie pour combiner l'ADN provenant de deux sources différentes dans une éprouvette.
   c) la mise au point d'outils informatiques pour l'analyse des génomes.
   d) l'utilisation d'outils mathématiques pour donner un sens aux systèmes biologiques.

2. Les gènes homéotiques :
   a) codent pour des facteurs de transcription qui assurent la régulation de l'expression des gènes commandant des structures anatomiques spécifiques.
   b) n'existent que chez *Drosophila melanogaster* et les autres arthropodes.
   c) sont les seuls gènes qui contiennent un domaine homéotique encodé par la boîte homéotique.
   d) codent pour des protéines qui forment des structures anatomiques chez la drosophile.

NIVEAU 2 : APPLICATION ET ANALYSE

3. Deux protéines eucaryotes ont un domaine en commun, mais elles sont pour le reste très différentes. Parmi les processus suivants, lequel est le plus susceptible d'avoir contribué à ce phénomène ?
   a) La duplication génique.
   b) L'épissage différentiel.
   c) Le brassage d'exons.
   d) La modification d'histones.

4. FAITES UN DESSIN ▶ Voici les séquences d'acides aminés (identifiés par leur symbole en une lettre ; voir la figure 5.14) de quatre courts segments de la protéine FOXP2 provenant de six espèces (pas nécessairement en ordre dans les séquences ci-dessous) : chimpanzé (C), orang-outan (O), gorille (G), macaque rhésus (R), souris (S) et humain (H). Ces segments contiennent toutes les différences d'acides aminés entre les protéines FOXP2 de ces espèces.

1. ATETI...PKSSD...TSSTT...NARRD
2. ATETI...PKSSE...TSSTT...NARRD
3. ATETI...PKSSD...TSSTT...NARRD
4. ATETI...PKSSD...TSSNT...SARRD
5. ATETI...PKSSD...TSSTT...NARRD
6. VTETI...PKSSD...TSSTT...NARRD

À l'aide d'un surligneur, marquez d'une couleur tout acide aminé qui varie parmi les espèces. (Colorez cet acide aminé dans toutes les séquences.) Puis, répondez aux questions qui suivent.

a) Les séquences C, G, R sont identiques. Quelles lignes correspondent à ces séquences ?
b) La séquence de l'humain diffère de celle des espèces C, G et R par deux acides aminés. Soulignez les deux différences dans la séquence H.
c) La séquence O diffère de celle des espèces C, G et R par un acide aminé (V plutôt que A) et de la séquence H par trois acides aminés. Quelle ligne correspond à la séquence O ?
d) Dans la séquence S, entourez les acides aminés qu'on ne retrouve pas dans les séquences C, G et R, et tracez un carré autour de ceux qu'on ne retrouve pas dans la séquence H.
e) Les primates et les rongeurs ont divergé il y a entre 60 et 100 millions d'années, et les chimpanzés et les humains ont divergé il y a environ 6 millions d'années. Comparez les différences d'acides aminés entre la souris et les espèces C, G et R avec les différences entre l'humain et les espèces C, G et R ? Qu'en concluez-vous ?

Voir les réponses proposées à l'appendice A.