Stratégies de séquençage

6

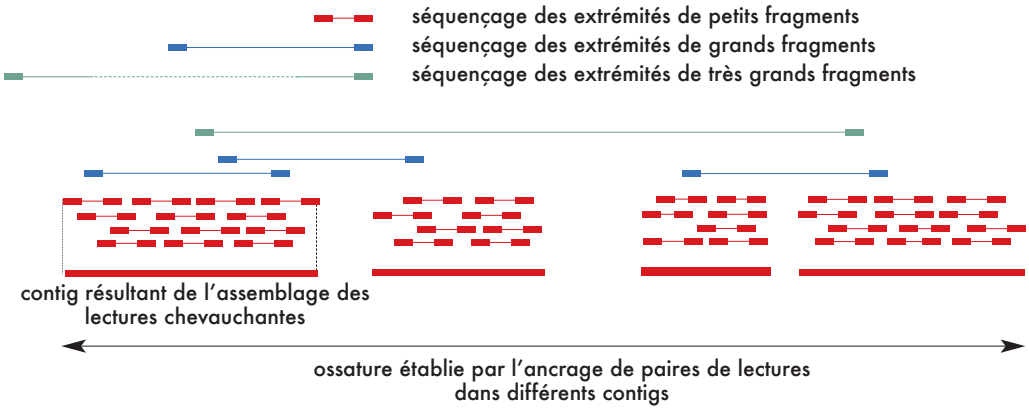
Le séquençage des grands génomes reste une entreprise ardue, pour laquelle différentes stratégies existent. Elles répondent à une même difficulté : sachant que la lecture des réactions de séquençage ne livre que des séquences de 1000 nucléotides au plus, comment reconstituer l'ensemble de la séquence du génome, plusieurs milliers (bactéries) à plusieurs millions (mammifères) de fois plus longue ?

La comparaison des "lectures" individuelles permet certes d'identifier des séquences chevauchantes et d'assembler le puzzle de proche en proche. Toutefois, l'existence de séquences répétées, voire de régions entières dupliquées dans le génome, complique l'assemblage. En outre, l'échantillonnage des séquences lues peut laisser des trous, des régions du génome pour lesquelles on ne dispose pas de lecture. Il faut donc accumuler un nombre de lectures équivalant à plusieurs fois la taille du génome (au moins 4 à 5 fois) pour couvrir de façon satisfaisante le génome et réussir un premier assemblage.

L'assemblage des lectures chevauchantes ne produit donc pas une séquence continue entière, mais des blocs de séquence, nommés contigs. Or les lectures sont obtenues à partir des deux extrémités de fragments d'ADN de taille variable, et vont donc par paires. Cette information d'appariement peut permettre de jeter des ponts entre deux contigs. De la sorte, les contigs sont ordonnés et orientés à mesure que la couverture du génome augmente.

L'assemblage "brut" de séquences issues de l'ensemble du génome, avec l'aide apportée par l'appariement des lectures, constitue une stratégie nommée séquençage aléatoire global ("whole genome shotgun"). Cette méthode fonctionne bien avec les génomes bactériens, qui sont petits (de l'ordre de quelques millions de nucléotides) et pratiquement dénués de séquences répétées. Son application a été plus problématique dans le cas de la drosophile (où d'autres types de ressources ont été utilisées) et de l'anophèle. Enfin, sa contribution réelle au séquençage du génome humain, long de 3 milliards de nucléotides et constitué pour moitié de séquences répétées, reste vivement controversée *(voir la fiche* Génome humain*)*.

Dans le cas des grands génomes, la stratégie du séquençage aléatoire global consiste à obtenir des séquences appariées à différentes échelles génomiques. On procède donc au séquençage des extrémités de grands fragments génomiques (plusieurs centaines de milliers de nucléotides), qui permettent d'assembler les contigs en vastes ossatures et de détecter certaines erreurs à l'intérieur des contigs. *(suite au dos)*



La stratégie du séquençage aléatoire global

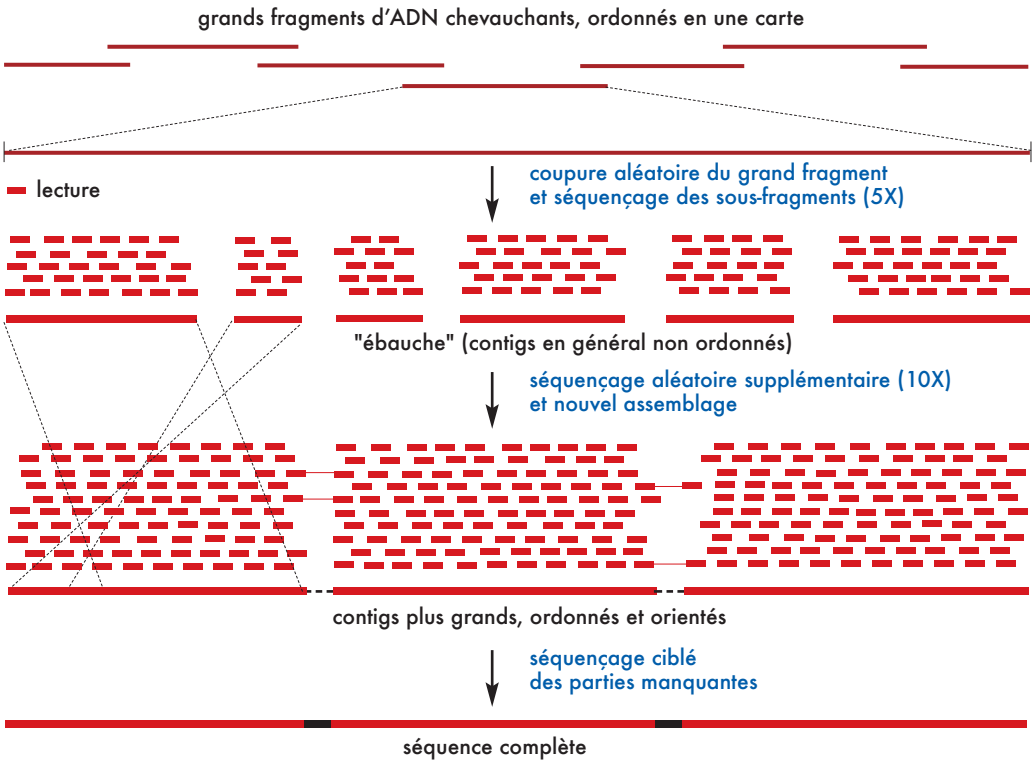
Stratégies de séquençage (suite)

6

Une stratégie différente – dite "clone par clone" ou hiérarchique – a été adoptée par divers consortiums internationaux pour le séquençage du génome humain et d'autres grands génomes (riz, arabette, ver, ...). L'assemblage des lectures chevauchantes n'est plus réalisé à l'échelle de l'ensemble du génome, mais à l'échelle de grands fragments génomiques (appelés abusivement "clones"), préalablement ordonnés en une carte. Cette compartimentation permet de diminuer la difficulté posée par les séquences répétées et, surtout, de diriger le travail de "finition" sur des régions précises. Cette étape de finition est indispensable dans la perspective d'une annotation exhaustive et précise de la séquence.

La contrepartie est un effort de cartographie pour ordonner les grands fragments génomiques *(voir la fiche* Cartographie*)*. Ce travail fait appel à diverses ressources : marqueurs des cartes physiques et génétiques, assemblage des grands fragments chevauchants sur la base de leur profil de coupure par des enzymes ("fingerprint"), et enfin séquences d'extrémités des grands fragments. Cette dernière ressource peut être utilisée pour construire la carte des grands fragments de proche en proche, à mesure que le séquençage progresse, et pour choisir les fragments chevauchants les moins redondants en vue du séquençage *(voir la fiche* Génome humain*)*.

Avec un volume de séquences lues équivalant à 5 ou 6 fois la taille d'un génome comme celui de l'être humain (profondeur de 5X), il est possible d'assembler clone par clone une "ébauche" où l'on obtient pour chaque clone quelques dizaines de contigs. Toutefois, ces contigs initiaux ne sont en général ni ordonnés, ni orientés. En augmentant la profondeur jusqu'à 10X, on obtient des contigs plus longs, moins nombreux, et surtout ordonnés et orientés grâce aux paires de lectures ou à d'autres informations. Reste alors à entreprendre une étape fastidieuse de lissage et de finition, pour garantir à tout endroit moins d'une erreur tous les 10 000 nucléotides, et pour boucher les trous résiduels par un travail local.



La stratégie de séquençage "clone par clone"