Etat actuel du séquençage et de la connaissance du génome des espèces animales

Alain Vignal

INRA Prod. Anim.,
2011, 24 (4), 387-404

Etat actuel du séquençage et de la connaissance du génome des espèces animales

A. VIGNAL
INRA, UMR444 Laboratoire de Génétique Cellulaire, F-31326 Castanet-Tolosan, France
Courriel : Alain.Vignal@toulouse.inra.fr

Avec l'avènement des nouvelles technologies, le séquençage de génomes eucaryotes entiers est devenu monnaie courante et les génomes de toutes les espèces animales d'intérêt agronomique seront très bientôt décryptés et présentés sous forme de «*draft sequences*» (séquences brouillon) plus ou moins complètes. Le re-séquençage d'individus et/ou de populations est déjà une réalité, permettant la mise en évidence de régions du génome sous influence de la sélection.

A de rares exceptions près, la question n'est plus tellement de savoir si le génome d'un animal de rente va être séquencé, mais plutôt dans quelles conditions, avec quelles technologies et si cette séquence va pouvoir s'appuyer sur des cartes. En effet, l'utilisation des nouvelles technologies de séquençage permet maintenant la production massive de données à un coût relativement bas et les étapes limitantes sont au niveau de l'assemblage de ces données en séquences de chromosomes complets et bien ordonnés. Le degré de finition et le type d'information que l'on pourra retirer d'un génome séquencé dépendra donc beaucoup de l'effort consenti à la production de données annexes telles que des cartes ou des séquences de transcrits et à leur intégration avec les données de séquence. Le point le plus délicat à traiter n'est donc plus la production des séquences, mais la capacité à pouvoir les assembler et les analyser correctement. C'est la qualité de cet assemblage, qui va conditionner par la suite celle des annotations structurales* (position, structure des gènes et autres éléments fonctionnels du génome), mais aussi de manière importante la qualité de l'annotation fonctionnelle* (fonction des éléments annotés). Cependant, quel que soit son degré d'avancement, un génome séquencé sera toujours une mine d'informations à la fois pour les recherches en génétique et structure des populations qu'en fonctionnement des systèmes biologiques.

1 / Génome, séquence, chromosomes, réplication de l'ADN

1.1 / Génome et séquence

Le génome représente la totalité de l'information génétique codée par l'ADN pour un organisme donné. La séquence d'un génome est l'ordre dans lesquelles se trouvent les 4 bases azotées ou nucléotides : Adenine (A), Cytosine (C), Guanine (G) et Thymine (T), utilisées comme un alphabet de 4 lettres, le long de filaments d'ADN. Cette séquence n'est pas aléatoire, car elle spécifie les informations permettant d'exprimer les caractères visibles de l'organisme (phénotypes), en interaction avec le milieu. Une fonction majeure de l'ADN est de coder pour les gènes, dont la plupart spécifient la fabrication des protéines, qui assurent une part importante des fonctions connues de la cellule et de l'organisme. La célèbre structure en double-hélice de l'ADN, élucidée en 1953 par James Watson et Francis Crick, permet en outre à l'ADN d'avoir des propriétés auto-réplicatives, grâce à la complémentarité des bases : A ⇔ T et C ⇔ G (Encadré 1). C'est cette structure en double-hélice, composée de deux brins complémentaires (brin «+» ou «sens», et brin «-», «complémentaire» ou «réverse»), à laquelle on se réfère quand on parle de longueurs de séquences en paires de bases (pb*). Par convention, la séquence d'un seul brin est reportée, la séquence du brin complémentaire pouvant être déduite directement grâce à la complémentarité des bases. Les informations lues par la machinerie cellulaire en revanche, peuvent concerner aussi bien l'un ou l'autre des brins, voire les deux à la fois. Par exemple, les gènes et donc les protéines peuvent être codées aussi bien par le brin sens que le brin complémentaire de la double hélice (figure 1).

Le nombre de gènes estimés pour le génome humain est de l'ordre de 25 000 à 30 000 (Roest Crollius *et al* 2000, Southan 2004). Un gène peut coder pour une ou plusieurs protéines et certains gènes ne codent que pour un ARN qui ne sera pas traduit en protéine. S'il apparaît donc que seulement 1,5% de la séquence des génomes de mammifères codent pour des protéines, au moins 5% du génome semblent sous sélection, suggérant qu'entre 5 et 20% du génome auraient des fonctions importantes (Pheasant et Mattick 2007).

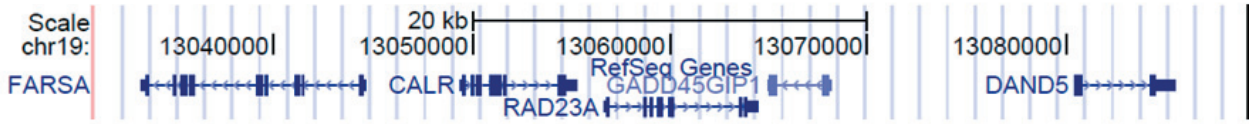
1.2 / Chromosomes, cellules haploïdes et diploïdes

Les chromosomes, composés d'ADN et de protéines, ont été identifiés comme porteurs de l'information génétique dès le début du 20ème siècle. A titre d'exemple, chez l'Homme, cette information est répartie le long de 23 chromosomes et le génome complet est constitué d'une séquence d'environ 3 milliards de nucléotides (ou paires de bases).

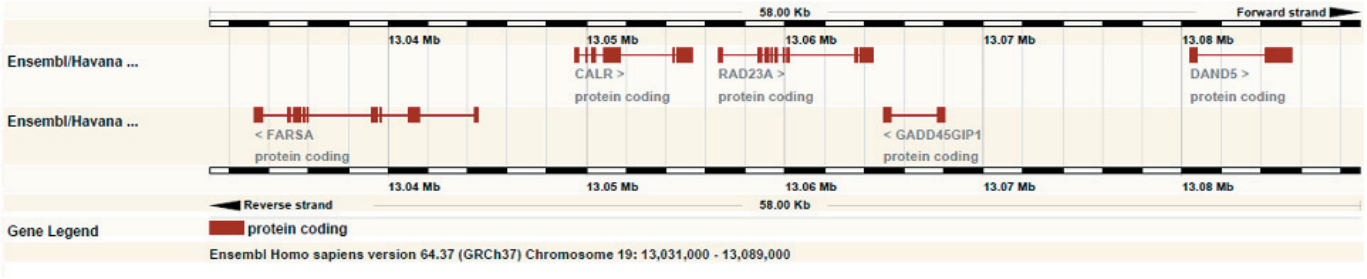
* Voir glossaire.

Figure 1. *Région génomique présentée sur les sites UCSC et Ensembl.*
La même région a été sélectionnée dans les sites de l'UCSC et Ensembl et la fenêtre présentant les gènes a été extraite. Les lignes (boîtes) épaisses représentent les exons de gènes, séparés par les introns représentés par des lignes fines de même couleur. Sur les cinq gènes présentés, trois (CALR, RAD23A, DAND5) sont codés par le brin sens et deux (FARSA, GADD45GIP1) par le brin réverse, comme indiqué par les flèches dans les introns de gènes (UCSC) ou à côté de leur nom (Ensembl).

UCSC http://genome.ucsc.edu/



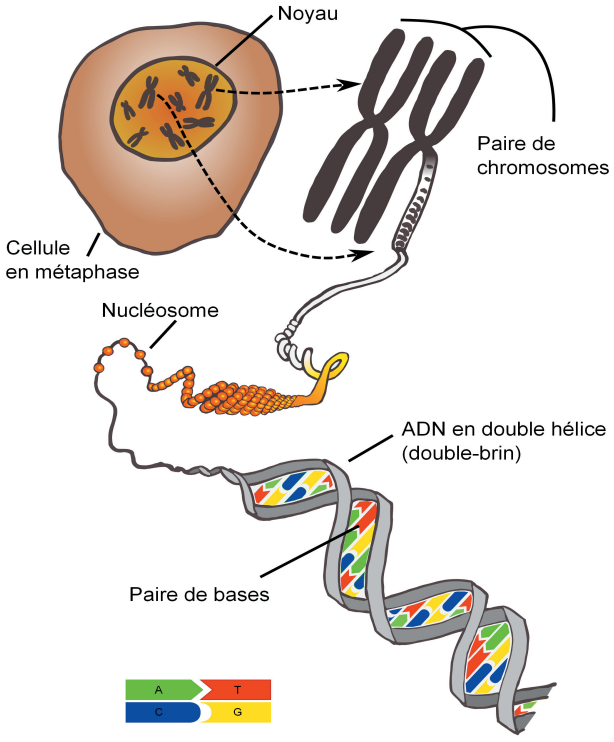
Ensembl http://www.ensembl.org/index.html



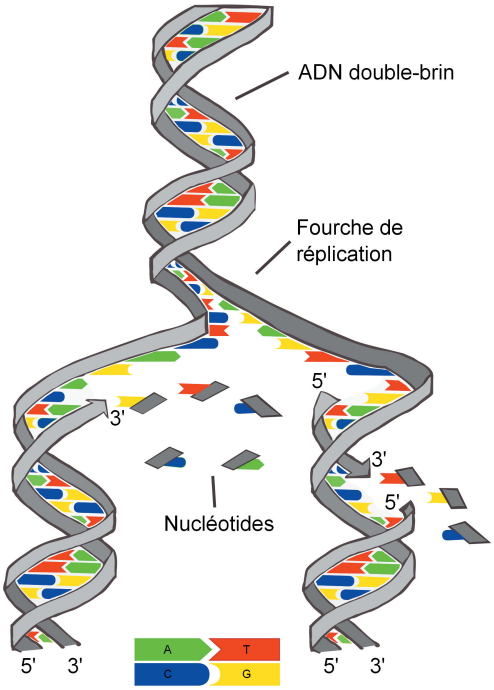
Encadré 1. *Génome, chromosomes, réplication de l'ADN*

A : Cellule, chromosomes, ADN. La séquence d'un génome représente la totalité de l'information génétique contenue dans l'ADN pour un organisme donné. L'ADN est lui-même une composante essentielle des chromosomes. Déchiffrer cette séquence revient à déterminer l'ordre dans lequel les 4 bases azotées ou nucléotides : Adenine (A), Cytosine (C), Guanine (G) et Thymine (T), s'enchaînent le long des filaments d'ADN. La séquence n'est pas aléatoire, car elle spécifie des messages permettant d'exprimer les caractères visibles de l'organisme (phénotypes), en interaction avec le milieu. Chez l'Homme, cette information est répartie dans n = 23 chromosomes et une séquence de près de 3 milliards de nucléotides constituent le génome complet. Toute cellule de l'organisme, à de rares exceptions près, telles que les globules rouges des mammifères qui n'ont pas de noyau et donc de chromosomes et les gamètes qui n'ont qu'une copie du génome, contient la totalité du génome en double copie, chacune provenant d'un des deux parents. Chez l'Homme, une cellule somatique contient donc $2n$ = 46 chromosomes.

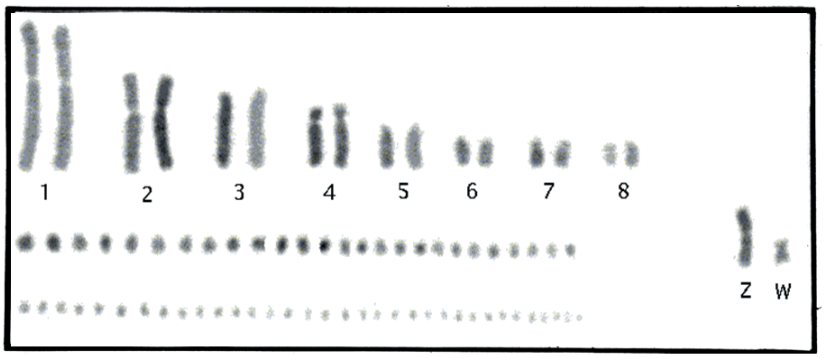
La fonction la plus connue de l'ADN est de coder pour les gènes, dont la plupart spécifient la fabrication des protéines, qui assurent la majorité des fonctions de la cellule et de l'organisme. Le nombre estimé de tels gènes pour le génome humain est de l'ordre de 25 000 à 35 000. D'autres gènes servent à produire des molécules d'ARN, qui ne seront pas traduites en protéines, mais qui auront des fonctions importantes en participant à des complexes riboprotéiques, à des mécanismes de régulation de l'expression des gènes et à des phénomènes épigénétiques. Pour une espèce donnée, l'ordre des gènes le long de la séquence d'ADN est défini et sera déduit de la séquence.



B : Réplication de l'ADN. La célèbre structure en double hélice de l'ADN, élucidée en 1953 par James Watson et Francis Crick, permet à l'ADN d'avoir des propriétés auto-réplicatives, grâce à la complémentarité des bases : A ↔ T et C ↔ G. Les deux brins d'une double hélice seront chacun recopiés, permettant à une cellule de réaliser une distribution équitable de l'information génétique à ses deux cellules filles lors de la mitose.



C : Caryotype de poule. La forme en X souvent utilisée pour représenter un chromosome correspond en fait à un stade très particulier, la métaphase de la mitose, où chaque chromosome est entièrement répliqué juste avant la répartition des deux copies dans les deux cellules filles. En réalité, cette forme en X n'est pas toujours visible, même au stade de la métaphase. Chaque chromosome est visible sous forme de bâtonnet simple. Les numéros indiquent les paires de chromosomes. Les chromosomes Z et W sont les gonosomes, équivalents des chromosomes X et Y chez les mammifères.



La représentation la plus courante des chromosomes est celle que l'on peut observer au cours de la division cellulaire, lors de la métaphase de la mitose ou de la méiose, pendant lesquelles ils sont condensés. En dehors de ces phases spécifiques, pendant l'interphase, les chromosomes sont décondensés et impossibles à distinguer dans le noyau sans utiliser des marqueurs spécifiques. En fonction du stade plus ou moins avancé de la condensation qui a lieu lors des divisions cellulaires, les chromosomes seront visibles sous forme de bâtonnets simples plus ou moins allongés (les deux copies qui seront réparties entre les cellules filles sont encore accolées) ou doubles (les deux copies sont détachées, mais sont encore retenues ensemble au niveau du centromère). C'est cette dernière représentation qui donne la structure en X des chromosomes souvent utilisée dans les représentations schématiques.

Lors de la fécondation d'un ovule par un spermatozoïde, chacune de ces deux cellules gamètes haploïdes apporte la copie unique du génome qu'elle contient, soient 23 chromosomes et 3 milliards de pb* chez l'Homme, pour former une cellule diploïde contenant $2n$ = 46 chromosomes. Le génome est donc présent dans cette nouvelle cellule œuf, ou zygote, en deux exemplaires quasiment identiques, aux variations individuelles transmises par les deux parents près. Cette cellule originale se divise alors par mitoses successives, transmettant la totalité des deux copies du génome à ses descendantes. Elle est donc à l'origine des quelques 10^{14} de cellules composant le corps et qui contiennent, à de rares exceptions près, les deux copies du génome héritées des parents. Le génome des deux parents est brassé lors de la meiose, grâce aux phénomènes de recombinaison entre chromosomes homologues, la division cellulaire produisant les gamètes : spermatozoïdes et ovules, qui sont des cellules haploïdes. Un exemple de caryotype de poule est donné (Encadré 1C).

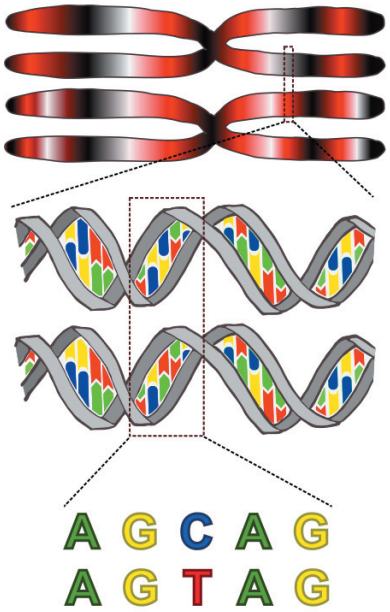
1.3 / Réplication de l'ADN

Les cellules doivent réaliser une copie à l'identique de tout l'ADN génomique qu'elles contiennent, afin de transmettre leurs copies du génome aux deux cellules filles lors de la mitose. Ceci est rendu possible par la structure en double-hélice de l'ADN et la complémentarité des bases nucléotidiques et est réalisé par des complexes enzymatiques contenant notamment de l'ADN polymérase (Encadré 1B). Cette propriété est utilisée pour de nombreuses applications, les plus courantes étant la PCR (*Polymerase Chain Reaction)* et le séquençage. Pour ces deux techniques très répandues, une ADN polymérase est utilisée pour réaliser des copies *in vitro* d'une séquence à analyser.

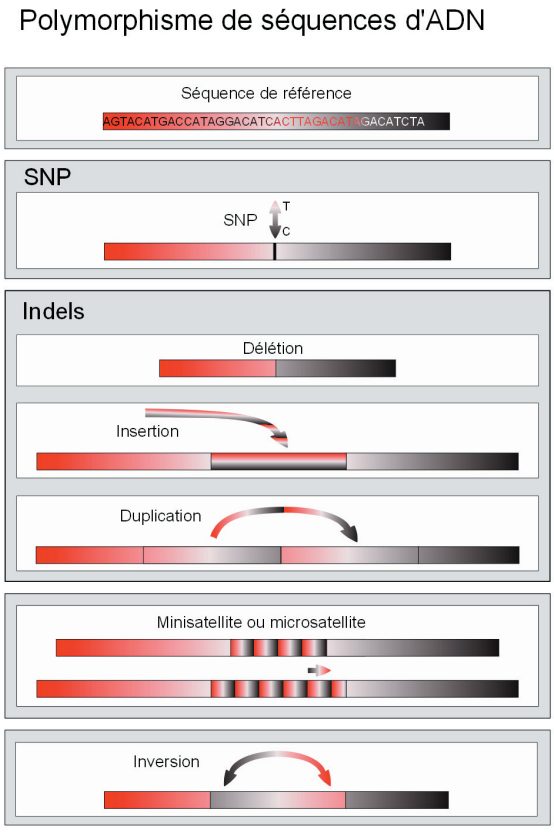
1.4 / Polymorphisme

La séquence du génome est légèrement variable entre plusieurs individus d'une même espèce, cette variabilité étant principalement due à des changements ponctuels de nucléotides dans la séquence (SNP* : *Single Nucleotide Polymorphism*) (Encadré 1D), à des insertions ou délétions plus ou moins longues (InDel* : *Insertion-Deletion*) ou à des inversions de séquence (Encadré 1E). Le polymorphisme existant dans un gène peut altérer son fonctionnement, les effets les plus couramment observés étant une variation du niveau d'expression ou de la structure de la protéine codée. De telles variations pourront avoir des effets plus ou moins importants sur le phénotype et seront à l'origine des variations interindividuelles. Il est notable cependant, que la plupart du polymorphisme mis en évidence par les programmes de séquençage n'a pas d'influence connue sur la variabilité des phénotypes ou l'apparition d'une maladie génétique. Chez l'Homme, en prenant deux copies

D : Polymorphisme SNP. Les deux copies répliquées d'un chromosome en vue de la transmission lors de la mitose sont identiques, tandis que les deux copies héritées de chacun des deux parents et formant une paire de chromosomes allèles, auront des séquences légèrement différentes, représentant une partie de la diversité génétique de la population. Ici, un SNP* (*Single Nucleotide Polymorphism*).



E : Différents types de variations nucléotidiques. Les différences de séquence entre chromosomes allèles peuvent concerner des changements ponctuels d'une base (SNP*), mais peuvent également être de toute autre forme : délétions, insertions ou inversions de fragments de séquence de longueur parfois importante (de quelques pb* à plusieurs centaines de kb*) ; duplications en tandem ou non (duplications pouvant impliquer deux ou plusieurs chromosomes) ; minisatellites ou microsatellites : séquences de 2 à 10 pb* (microsatellites) ou plus (minisatellites) répétées en tandem et dont le nombre varie d'un chromosome allèle à un autre.



du génome sélectionnées au hasard dans la population, on observe en moyenne un SNP tous les 1000 pb*, soient environ 3 millions de SNP au total. En multipliant le nombre de génomes séquencés pour une espèce, le nombre de SNP connus devient très élevé. Par exemple, plus de 10 millions de SNP communs ont été identifiés dans le génome humain (Altshuler *et al* 2010).

2 / Stratégies de séquençage et d'assemblage

2.1 / Cartes, assemblage en contigs* et scaffolds* (supercontigs*)

Un génome (haploïde) typique de mammifère est composé de 3 milliard de paires de bases (pb*), réparties sur plusieurs chromosomes, tandis qu'une lecture de séquence classique produite par une réaction basée sur la méthode de Sanger* et lue sur un séquenceur capillaire ne fait qu'un millier de pb*. Il faut donc un minimum de 3 millions de lectures (ou séquences) pour une couverture minimale. Dans la pratique, il en faut beaucoup plus car il est nécessaire que les lectures se chevauchent (figure 2) à la fois pour assurer la continuité de la séquence et pour corriger les erreurs éventuelles, afin de générer des séquences consensus, assemblées en contigs*.

Afin de minimiser le problème de l'ordonnancement des plusieurs dizaines de millions de lectures nécessaires, la première séquence du génome humain produite par le consortium public a été entreprise après avoir réalisé des cartes détaillées : génétiques, d'hybrides irradiés* et de contigs* de BAC*. Ces cartes permettaient de guider le choix de clones BAC* (*Bacterial Artificial Chromome*) à séquencer. Chacun des clones ne faisant que 100 à 200 kb* et étant localisé sur la carte, l'assemblage du génome humain était simplifié (Lander *et al* 2001). Cependant, une stratégie alternative de séquençage aléatoire global (WGS* : *Whole Genome Shotgun*), sans l'appui de cartes, a été proposée et utilisée pour le génome humain (Venter *et al* 2001). Chacune des deux stratégies avait ses avantages et inconvénients en termes de rapidité, de coût et de couverture du génome (Istrail *et al* 2004) et très rapidement, une stratégie intermédiaire a été adoptée, consistant à produire un grand nombre de lectures par du séquençage aléatoire global (WGS), tout en s'appuyant sur des cartes génétiques, RH (*Radiation Hybrid*), FPC* (*FingerPrint Contig**) et FISH* (*Fluorescent In Situ Hybridisation*) (figure 3). Sans cette dernière méthode

Figure 2. *Assemblage des lectures en contigs.*
Chaque lecture (1 kb* en moyenne pour du séquençage selon la méthode de Sanger ; 35 à 500 pb* pour du séquençage parallèle) est alignée par similarité de séquence avec l'ensemble des autres lectures produites, afin de construire des contigs de lectures chevauchantes, dont la taille sera limitée par les problèmes de couverture du génome dus à des raisons statistiques liées à la distribution aléatoire des lectures, mais aussi à des problèmes techniques ou liés à la structure du génome (présence de séquences répétées, biais de composition en bases…).

de cartographie, les séquences ne peuvent pas être assignées aux chromosomes. Cette stratégie a été utilisée pour le séquençage de la poule, premier animal de rente séquencé (Hillier *et al* 2004) et est encore la plus utilisée actuellement, dans le cas où l'assemblage en chromosomes complets est prévu.

Avec l'avènement des nouvelles technologies, permettant un accroissement significatif du volume de données produit tout en en diminuant le prix, un nombre croissant d'espèces est séquencé sans l'appui de cartes. Ces dernières deviennent donc maintenant le facteur limitant et des séquences de génomes sont produites sans que l'on sache attribuer les données à des chromosomes, ce qui était inconcevable il y a peu de temps (Li *et al* 2010). Si de tels assemblages peuvent sembler d'un apport limité pour aborder l'étude de la structure globale des génomes, des réarrangements chromosomiques inter-spécifiques, ou pour le clonage positionnel, ils sont tout de même très utiles pour la détection de gènes et l'étude de leur structure et peuvent servir de point de départ pour de nombreuses études telles que l'expression génique* ou la phylogénomique*.

Les lectures de séquence produites par WGS* sont assemblées par voie informatique en lectures chevauchantes (contigs*), permettant la production d'une séquence continue (figure 2). Cependant, pour diverses raisons tenant principalement à la distribution statistique des lectures, à la présence de séquences répétées dans le génome (figure 4) et à des difficultés de production de lectures pour certaines régions ayant des caractéristiques de séquence particulières (par exemple un taux de nucléotides (G + C) particulièrement élevé ou faible), la couverture du génome est loin d'être optimale. De très nombreux contigs*, souvent plusieurs dizaines de milliers, dont la taille est de l'ordre de la dizaine ou de la centaine de kb seulement, sont à mettre en regard de chromosomes dont la taille fait plusieurs dizaines de Mb*. Les portions de séquence continue des contigs* sont ensuite assemblés en supercontigs* (ou *scaffolds**) en utilisant des lectures appariées (Encadré 2C ; figure 5). Jusqu'à récemment ces données provenaient uniquement de séquences d'extrémités de clones BAC*, cosmides* ou fosmides*, et/ou des lectures plus ou moins complètes de clones BAC*. Des lectures appariées peuvent maintenant être produites directement par les nouvelles générations de séquenceurs par séquençage des deux extrémités des fragments d'ADN (*paired-ends**) ou en circularisant les fragments à séquencer (*mate-pairs**).

Les assemblages sont réalisés à l'aide de programmes informatiques tels que Arachne (Batzoglou *et al* 2002), Phusion (Mullikin et Ning 2003), Atlas (Havlak *et al* 2004) ou PCAP (Huang *et al* 2003) pour les séquençages classiques de type Sanger* ou tels que Euler, Velvet, AllPaths, ABySS ou SOAPdenovo (Miller *et al* 2010) pour les nouvelles technologies de séquençage parallèle générant un grand nombre de lectures courtes. Un même jeu de données peut générer des assemblages différents en fonction des assembleurs utilisés, montrant l'importance des algorithmes et de la bioinformatique.

Afin d'optimiser la phase d'assemblage des génomes, le choix de l'individu à séquencer est primordial. En effet, lors de la construction des contigs*, il est nécessaire de tenir compte de l'existence éventuelle de polymorphisme : l'individu séquencé peut avoir deux allèles différents pour un même locus, ce dont il faut tenir compte lors de l'assemblage. Cependant, des séquences très similaires peuvent exister en multiples copies dans le génome, dont l'origine peut être par exemple une duplication récente (séquences paralogues). Il est de ce fait parfois difficile de décider, lors de l'assemblage, entre deux allèles d'une même séquence et deux séquences paralogues (figure 6).

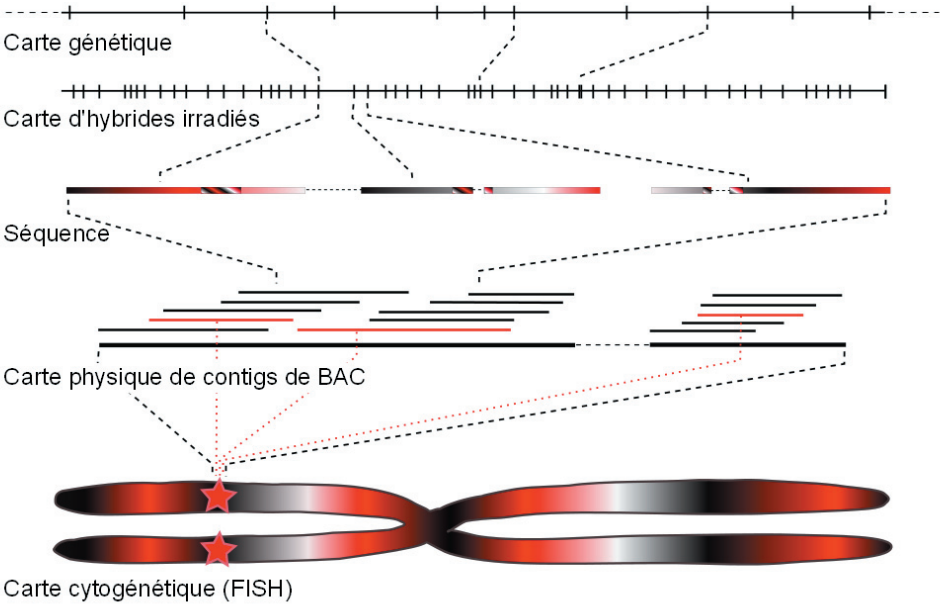
2.2 / Critères de qualité des assemblages

Un score de qualité phred (Ewing et Green 1998, Ewing *et al* 1998) est assigné à chaque base lue en fonction de paramètres liés aux signaux détectés par les séquenceurs. Ces scores sont utilisés pour calculer un score pour la séquence consensus. La qualité de la couverture du génome est estimée par la mesure de continuité N50 sur les contigs* et les supercontigs : au moins 50% du génome est couvert par des contigs* ou supercontigs de longueur supérieure à N50.

Les trois grandes catégories de qualité d'assemblage sont la «séquence finie», la «*working draft*» (brouillon de travail) et «*low-coverage*» (basse couverture). Les conventions habituelles stipulent que pour une séquence finie, 95% de la séquence des chromosomes doit être contenue dans une séquence continue de haute qualité, sans contaminations ou inversions. Au sens strict, elle n'est donc pas tout à fait finie, certaines régions étant pratiquement impossible à séquencer. Pour arriver à ce stade, les finitions requièrent obligatoirement du séquençage ciblé et seuls parmi les vertébrés, les génomes de l'Homme et de la souris (Church *et al* 2009, Ihgsc 2004) ont atteint cet objectif. La séquence «*working draft*» est incomplète et des erreurs d'assemblage

Figure 3. *Intégration des différents types de carte.*

L'étape finale d'assemblage consiste à rassembler les contigs* et scaffolds* en utilisant des informations d'autres types telles que les cartes de contigs FPC* de BAC*, des cartes cytogénétiques, des cartes génétiques* et d'hybrides irradiés*. Chaque méthode de cartographie présente des défauts spécifiques et permet d'obtenir une couverture du génome plus ou moins complète. La carte génétique* permet de localiser à la fois des marqueurs moléculaires et des caractères phénotypiques. Une de ses limites provient du fait que pour placer un marqueur, il faut qu'il soit polymorphe, avec des allèles différents sur les chromosomes des parents des familles utilisées pour la carte. La carte d'hybrides irradiés* permet la localisation de tout marqueur pour lequel on peut développer un test PCR. La carte de contigs de BAC* a une résolution élevée et est composée de clones de grande taille chevauchants, contenant plusieurs gènes. Un chromosome n'est généralement pas couvert par un contig d'un seul tenant. La carte cytogénétique* a une faible résolution, mais est la seule permettant de relier les données à des chromosomes. Cependant, la localisation par FISH* n'est possible que pour des clones de grande taille, comme les BAC*. Tout comme la carte de contigs de BAC*, la séquence est composée de fragments (contigs*, scaffolds*), interrompus par intervalles. Afin de localiser les nombreux scaffolds* de la séquence, il est nécessaire de les relier aux autres cartes.



peuvent exister mais cependant, des critères de qualité tels qu'une proportion minimum du génome couverte par des bases d'une valeur de qualité donnée et un minimum de contamination par des séquences d'autres organismes (E. coli, vecteurs…), doivent être requis. Pour atteindre cet objectif par la technique de Sanger, le séquençage avec une profondeur minimale à 6 à 7 équivalents-génomes (6-7 X) est nécessaire. La séquence «*low coverage*» en revanche, est très incomplète et correspond le plus souvent à l'assemblage en contigs* et supercontigs de couvertures basses de génomes (2X).

Les séquenceurs parallèles permettent de générer un volume important de données à un coût moindre que la technique de Sanger, mais la profondeur de séquence nécessaire sera supérieure. La production typique de séquence nécessaire à l'assemblage *de novo* d'un génome de mammifère, tenant compte des possibilités techniques actuelles des séquenceurs parallèles, qui permettent d'obtenir des lectures de 100 pb* appariées, est la suivante : 45 équivalents de couverture du génome (45 X) à partir d'une banque avec des inserts de 180 pb*, 45 X avec des inserts de 3 kb*, 5X avec des inserts de 6 kb* et si possible,

Figure 4. *Assemblage en contigs et séquences répétées.*

La présence de séquences répétées va induire des erreurs d'assemblage. En début d'assemblage, elles seront détectées soit à l'aide de logiciels de détection de motifs répétés, soit par comparaison à des bases de données, ou par détection de zones présentant une profondeur anormalement élevée du nombre de lectures. Les lectures ainsi détectées seront éliminées, provoquant un morcellement de la séquence. Les multiples contigs créés par le morcellement seront ensuite rassemblés en utilisant des informations sur les séquences appariées et les séquences répétées seront réintroduites plus tard dans la mesure du possible.

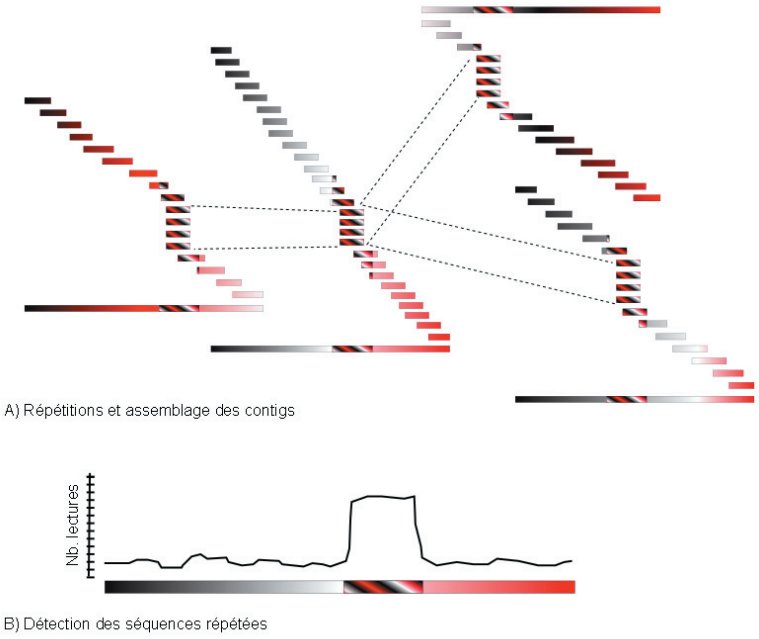


Figure 5. *Principe de l'assemblage de la séquence en contigs*, scaffolds* et chromosomes.*

Après élimination des séquences répétées, les multiples contigs de séquence continue de petite taille sont assemblés en utilisant des informations sur les séquences appariées (paired-ends et mate-pairs). En construisant des banques de séquençage avec des distances variables entre lectures, la construction des scaffolds* est améliorée.

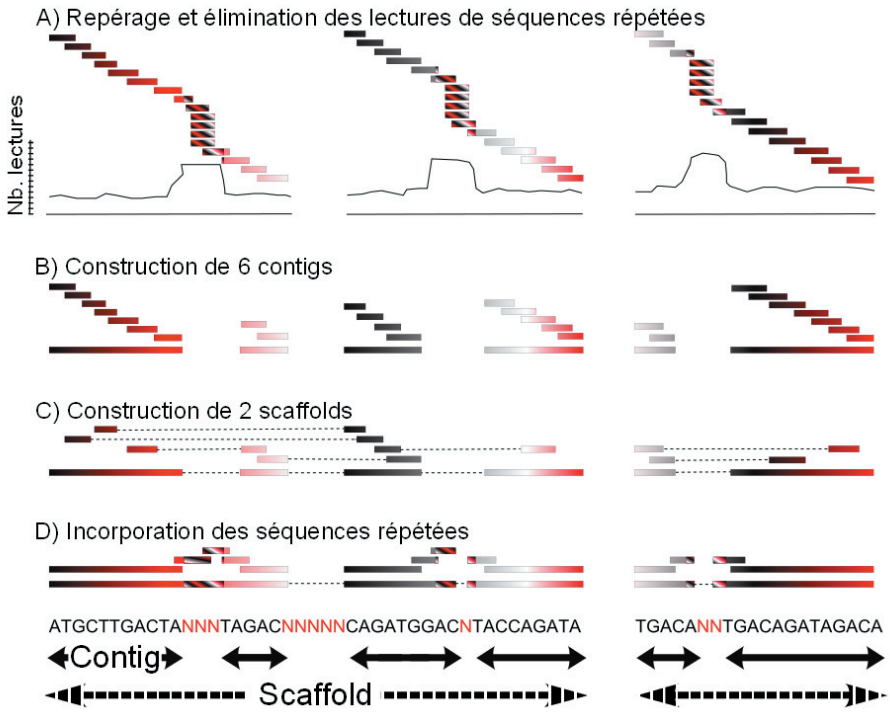
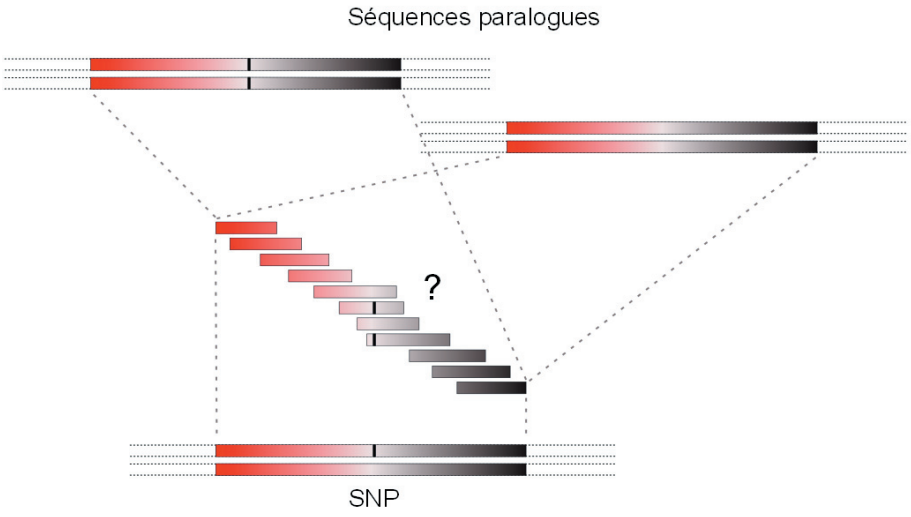


Figure 6. *Assemblage, SNP* et séquences paralogues*.*

Lors de l'assemblage des lectures, il faut tenir compte du polymorphisme présent chez l'individu séquencé. Quand des lectures divergent seulement pour quelques bases, il peut être difficile de choisir entre les assembler en un seul contig avec un SNP* ou en deux contigs séparés, en faisant l'hypothèse de l'existence de deux séquences paralogues*, récemment dupliquées dans le génome. Idéalement, en sélectionnant un individu totalement consanguin, les différences nucléotidiques entre lectures ne devraient pas être dues à des SNP*, ce qui facilite l'assemblage.



inclure 1X de lectures d'extrémités de clones de type fosmides* ou BAC* (Gnerre *et al* 2011).

2.3 / Recherche de SNP*

Les SNP* font maintenant partie des outils indispensables en génétique et sont largement utilisés pour des études d'association, permettant de tester des génomes complets dans des populations afin de détecter des zones du génome influant des caractères phénotypiques. Une autre application importante des SNP* est leur utilisation pour la sélection génomique, pour laquelle les génotypes déterminés pour des marqueurs en densité élevée sur le génome d'individus de phénotype connu, servent à prédire les phénotypes d'individus des générations suivantes. Ces nouvelles applications ont été rendues possibles grâce à la grande densité en marqueur SNP* existant dans les génomes et les nouvelles technologies de génotypage permettant l'étude simultanée de plusieurs centaines de milliers de ces marqueurs simultanément. Lors d'un programme de séquençage de génome, l'utilisation d'un individu consanguin facilitera l'assemblage du génome au détriment de la détection de SNP* (figure 6). La recherche de SNP* se fera par conséquent par séquençage à faible profondeur d'un nombre élevé d'individus, choisis dans des populations d'intérêt. L'alignement sur la séquence de référence permet ensuite la détection et la localisation des différences nucléotidiques. Parmi plusieurs millions de SNP* détectés par séquençage, quelques centaines de milliers seront retenus pour la fabrication de puces de génotypage en vue d'études exhaustives sur un grand nombre d'individus.

3 / Evolution de la capacité et des techniques de séquençage

3.1 / Les premiers vertébrés séquencés

La connaissance sur la séquence complète de génomes de vertébrés a progressé de manière très rapide au cours des vingt dernières années, avec les premières réalisations pour l'Homme il y a tout juste dix ans (Lander *et al* 2001, Venter *et al* 2001). La taille du génome avait été considérée comme un obstacle majeur par certains détracteurs du projet de séquençage du génome humain (*Human Genome Project* : HGP), suggérant qu'une évolution majeure des techniques de séquençage serait nécessaire avant de se lancer dans le projet, mais finalement celui-ci a été réalisé dans un réseau international de centres

de séquençage disposant chacun de l'ordre d'une centaine de séquenceurs classiques utilisant la technique de Sanger*. Il est amusant de constater à ce propos, qu'une des stratégies proposée pour obtenir un premier répertoire complet des gènes de vertébrés, avait été de séquencer des génomes de poissons, dix fois plus compacts que celui de l'Homme, mais que finalement, ces séquences ont été terminées et publiées juste après (Aparicio *et al* 2002, Jaillon *et al* 2004). La souris et le rat, vertébrés modèles majeurs pour la biologie humaine ont également été séquencés rapidement (Waterston *et al* 2002, Gibbs *et al* 2004).

3.2 / Le séquençage de génome «Sanger*» se répand

a) Une multiplication des génomes complets

Une fois réalisée la séquence de l'Homme, la «force de frappe» constituée dans les centres de séquençage pouvait se tourner vers d'autres génomes de vertébrés. L'expérience accumulée avait permis d'affiner les stratégies de production et d'analyse des données et d'envisager le séquençage d'un génome de plusieurs Gb à un coût moindre et dans un laps de temps de l'ordre d'une ou deux années seulement. Les génomes de la poule, du chimpanzé, de l'opossum et du chien ont ainsi été rapidement séquencés, suivis par un nombre de plus en plus important d'espèces, dont la vache, le cheval et le porc.

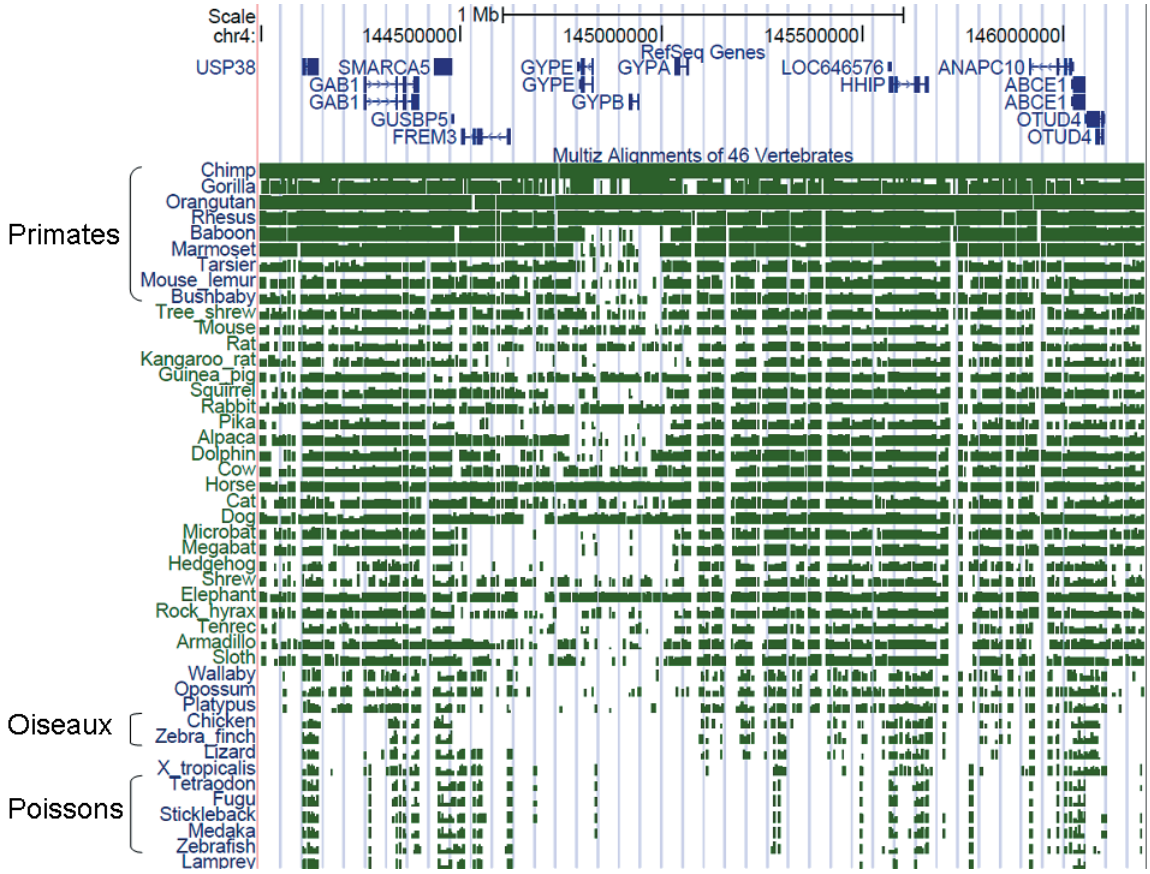
b) Les séquençages partiels

Le séquençage de nombreux mammifères permet la recherche de régions fonctionnelles par empreinte phylogénétique. Dans cette approche, l'alignement multiple de séquences de plusieurs espèces permet la détection des régions pour lesquelles la séquence est particulièrement conservée. Ces régions sont considérées comme étant fonctionnellement contraintes et peuvent aider à la détection d'exons* de gènes ou à celle de régions non-codantes fonctionnelles (figure 7). Pour cela, le *Broad Institute* (Cambridge, MA, USA) a initié un programme de séquençage partiel (entre 2 et 6-7 X de profondeur, selon les espèces) de 24 mammifères : *the Mammalian Genome Project* (Coordination : Broad Institute, USA). Dans le cadre de ce projet, une première séquence partielle du lapin a été obtenue.

3.3 / L'introduction du séquençage parallèle

Les anciens séquenceurs capillaires sont basés sur la détermination automatique de la taille de fragments d'ADN par électrophorèse dans des capillaires et détection fluorescente (Encadré 2). L'utilisation de quatre fluorochromes, permet de lire les quatre bases simultanément. Il est cependant nécessaire pour chaque lecture, de réaliser une réaction de Sanger* séparée sur des clones bactériens et/ou des fragments d'ADN produits spécifiquement par PCR.

Figure 7. *Alignement multiple de séquences de vertébrés sur le génome humain* (site UCSC : http://genome.ucsc.edu/).
Bleu : gènes ; vert : séquences conservées. Ces alignements multiples permettent de mettre en évidence des empreintes phylogénétiques. Certains gènes tels que *USP38* présentent une conservation de séquence très forte chez les vertébrés. d'autres tels que *GYPB* et *GYPE* sont moins bien conservés. Une région non codante présente dans la séquence entre *GYPA* et LOC646576 est bien conservée chez les vertébrés, suggérant une importance fonctionnelle. D'une manière générale, une forte conservation de séquence chez les primates, espèces proches de l'Homme, ne permet pas de distinguer de signal spécifique et à l'inverse, les poissons ne permettent pas de mettre en évidence toutes les régions conservées. On peut toutefois noter une conservation limitée des séquences pour les gènes *GYPB* et *GYPE* pour certains primates.



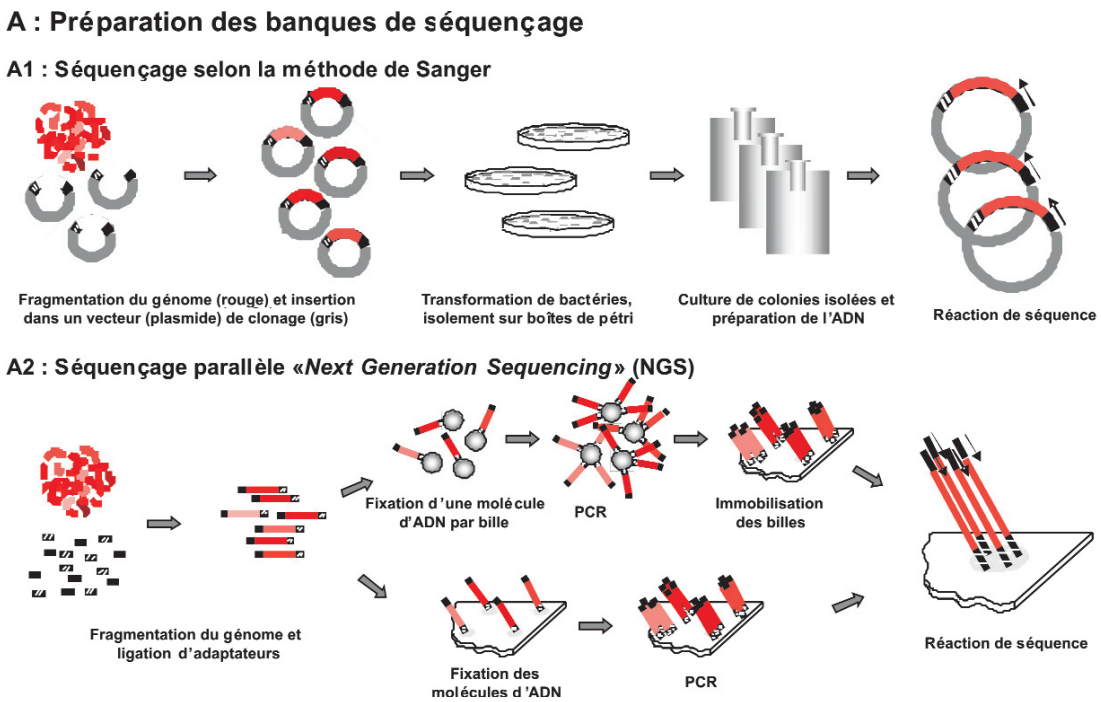
Encadré 2. *Les techniques de séquençage Sanger et NGS (*Next Generation Sequencing*), ou séquençage parallèle.*

Les techniques de séquençage actuellement utilisées tirent parti de la structure en double hélice de l'ADN, qui permet la copie fidèle de la séquence du génome lors de la réplication cellulaire. Les lectures obtenues par la technique de Sanger sont de l'ordre de 1000 paires de bases (pb*), tandis que par les NGS, elles sont de l'ordre de 75 à 100 pb* (Illumina ou ABI Solid) ou 500-800 pb* (Roche 454). La technologie produisant les lectures les plus courtes est aussi celle qui permet de produire le plus de lectures (plusieurs millions simultanément) et une quantité de séquence plus importante. Le taux et le type d'erreur va varier selon la technologie NGS, mais n'est à ce jour pas aussi bas qu'en séquençage Sanger.

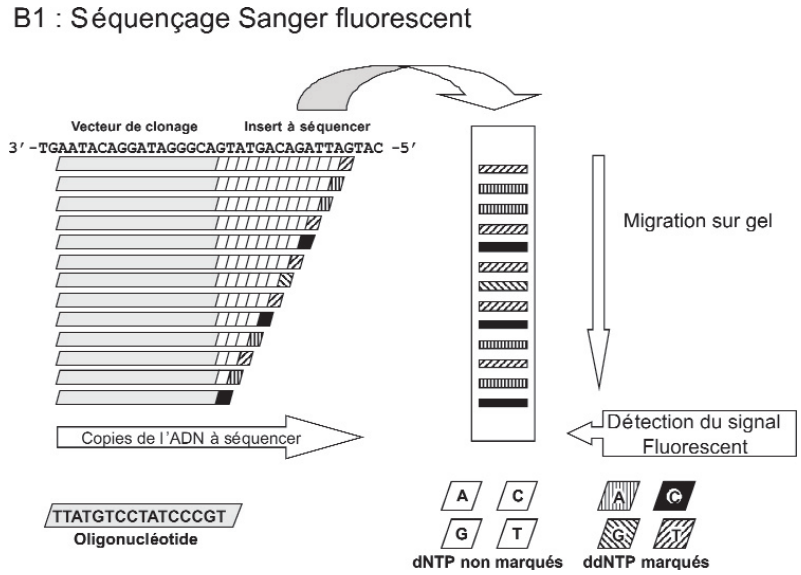
A : Afin de produire un signal d'intensité suffisante pour être détecté, il est nécessaire de réaliser de nombreuses copies du fragment à séquencer. Un ensemble de fragments prêts à être séquencés est une banque d'ADN.

A1 : Le séquençage par la méthode de Sanger nécessite que les fragments d'ADN soient amplifiés dans des clones bactériens, impliquant des étapes de culture individuelle pour chaque fragment. Une alternative est l'amplification individuelle des fragments par PCR.

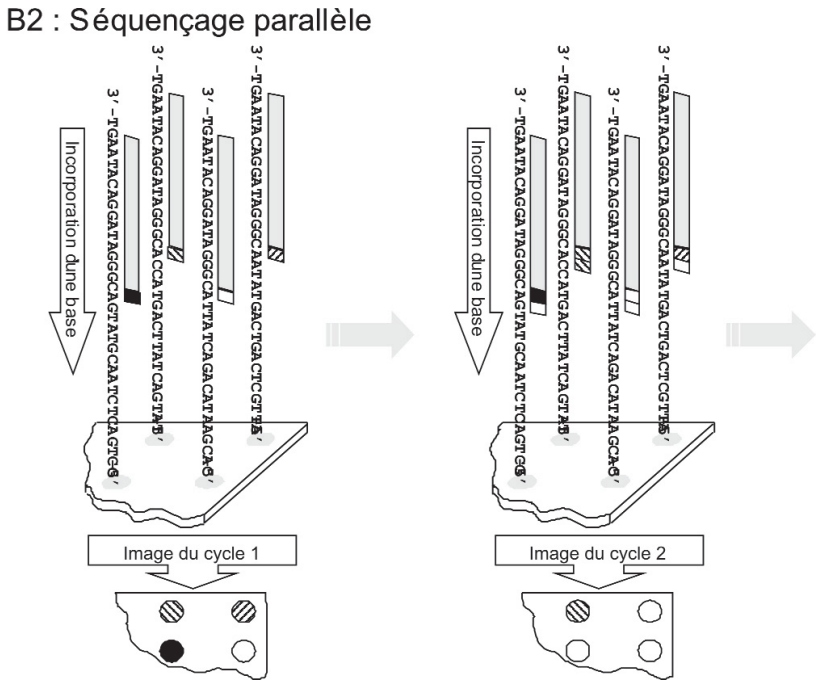
A2 : Pour le séquençage parallèle, des molécules uniques d'ADN reçoivent des adaptateurs ADN de séquençage par ligation, puis sont isolées soit par fixation sur des microbilles, soit directement sur un support solide plan. Une PCR directement sur le support solide ou sur l'ensemble des billes, permet alors d'amplifier chaque molécule spécifiquement. Dans le cas de l'amplification sur billes, les billes sont alors ensuite immobilisées sur un support plan.



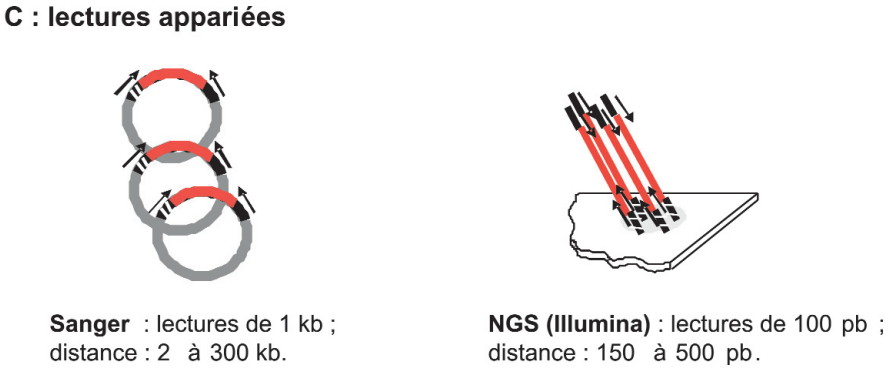
B1 : Un oligonucléotide de synthèse complémentaire à la séquence du vecteur de clonage sert d'amorce pour la copie de l'insert à séquencer, réalisée à l'aide d'une polymérase. L'incorporation aléatoire de bases modifiées (ddNTP) provoque l'arrêt de l'élongation. Chacun des 4 ddNTP étant marqué avec un fluorophore différent, la détection est réalisée automatiquement après séparation des fragments synthétisés selon leur taille par électrophorèse capillaire. L'ordre de passage des fragments détermine la séquence.



B2 : Les fragments d'ADN amplifiés et immobilisés sur support solide sont lus simultanément en plusieurs cycles. A chaque cycle, une seule base est incorporée et une image de l'ensemble des fragments est réalisée.



C : Il est possible de produire des lectures de séquence appariées. Dans le cas du séquençage Sanger, la distance entre les lectures va dépendre du vecteur de clonage utilisé : 2 à 10 kb* pour des plasmides* ; 100 à 300 kb* pour des BAC* et 40 kb* exactement pour des fosmides*. Les lectures appariées peuvent être produite de manière directe, avec une distance entre lectures de 200 à 500 pb* (Illumina) et de manière indirecte (non montré), avec une distance allant de 3 à 10 kb* (Illumina) ou jusqu'à 40 kb* (Roche 454).



Le principe de la nouvelle génération de séquenceurs, dite NGS (*Next Generation Sequencing*) est le séquençage parallèle d'un très grand nombre de fragments (de l'ordre du million) d'ADN immobilisés, par lecture des bases au fur et à mesure de leur incorporation. Ces séquenceurs NGS sont utilisés soit en appui au séquençage classique, soit comme seule source de données. Dans ce dernier cas, bien que les profondeurs de séquençage pratiquées soient de l'ordre de 60 X voire plus, les scaffolds* obtenus sont plus courts qu'avec le séquençage classique et ne sont pas assignés a priori à des chromosomes. Ils constituent néanmoins une source importante d'informations sur les gènes présents dans une espèce ou sur le polymorphisme, d'autant que ces données sont le plus souvent complétées par le séquençage d'EST* (*Expressed Sequence Tags*) pour l'annotation structurale* et le séquençage d'individus multiples ou de populations pour la recherche de SNP*, d'insertions-délétions (InDels*) ou d'autres variations nucléotidiques. Les scaffolds peuvent cependant être intégrés après leur constitution à des cartes génétiques, d'hybrides irradiés et/ou physiques.

3.4 / Stockage, présentation des données

Les données de séquences de génomes complets sont stockées dans les bases de données internationales : DDBJ/EMBL/GenBank, qui sont trois sites miroir respectivement au Japon, aux USA et en Europe. Pour la visualisation des données de séquence et d'annotation* (position des gènes, fonctions, similarités entre espèces, polymorphisme…), il existe trois principaux browsers de génome : Ensembl (http://www.ensembl.org/index.html) (figure 1), UCSC (http://genome. ucsc.edu/) (figure 1) et NCBI (http://www.ncbi.nlm.nih.gov/projects/mapview/). Selon le type d'analyse que l'on voudra effectuer et les informations que l'on voudra retirer, l'un ou l'autre de ces sites sera le plus approprié.

La plupart des mises à jour de versions d'assemblage de génomes correspondent à des évolutions faisant suite à l'intégration de données nouvelles : augmentation de la profondeur, changement de technologie de séquençage, séquençage de finition, etc. D'autres améliorations dans les assemblages sont dues aux évolutions des algorithmes d'assemblage, tenant mieux compte des problèmes dus à l'ADN répété et aux duplications de séquences. Par exemple deux versions de l'assemblage du génome de la vache ont été publiées simultanément à partir des mêmes données de séquence, ce qui soulève un point important concernant leur disponibilité, avec le problème du choix d'un génome de référence qui sera le seul disponible actuellement dans les browsers Ensembl et UCSC. Ce point avait déjà été soulevé lors de la publication du génome de la souris, pour lequel deux logiciels d'assemblage : Arachne et Phusion avaient été utilisés. L'assemblage de référence produit par Arachne avait alors été choisi en raison d'une continuité (N50) plus grande des séquences à précision comparable (Waterston *et al* 2002). Bien que certains browsers de génomes présentent actuellement toujours les anciennes versions des assemblages avec les nouvelles, il n'est pas possible de visualiser de manière simple deux versions pour une même espèce. Ce problème deviendra d'autant plus important que les données de re-séquençage de génomes commencent à affluer et qu'il faudra aussi tenir compte du polymorphisme d'insertions-délétions et d'inversions. Quoi qu'il en soit, l'échange de données sera difficile en l'absence d'un système de coordonnées stable (Church et Hillier 2009).

4 / Les espèces d'élevage

Les espèces d'intérêt agronomique ont été séquencées à des époques différentes, en fonction de leur intérêt biologique ou de leur importance économique. Il s'ensuit une variété de situations tant du point de vue des stratégies utilisées que de la qualité des résultats obtenus. Ne sera décrite ici, pour les espèces qui nous intéressent, que la production d'une séquence de référence. Cependant, les programmes de séquençage d'un génome s'accompagnent le plus souvent de la recherche de SNP* et de séquençage de transcrits pour l'annotation*.

Bovins (*Bos taurus*) - Les publications traitant de l'assemblage du génome du bovin démontrent bien toute la difficulté qu'il y a, de reconstituer un génome à partir des données brutes de séquence. En effet, un des assemblages : Btau_4.0 (Liu *et al* 2009), qui est celui reconnu comme étant la référence et est trouvé dans les browsers de génomes Ensembl et UCSC, est basé uniquement sur les lectures WGS* et de clones BAC*, assemblées à l'aide de la suite de programmes Atlas (Havlak *et al* 2004). Cet assemblage est placé sur les chromosomes en utilisant les données de cartographie et est notamment contrôlé par la cartographie génétique* de plus de 17 000 SNP*, dont 99,2% sont positionnés correctement. L'autre assemblage : UMD3 (Zimin *et al* 2009), utilise les mêmes données brutes et des algorithmes d'assemblage différents. De plus, l'orientation de certains contigs*, voire leur assignation chromosomique, a été réalisée par alignement sur la séquence humaine, afin d'étendre la couverture de la séquence. Au final, 2,65 Gb de séquence sont assignés aux chromosomes pour UMD3, contre 2,47 pour Btau_4.0 et seul UMD3 contient de la séquence pour le chromosome X. Les valeurs contig* N50 sont de 93,1 et 103,7 kb* respectivement. Il existe de plus des régions en désaccord entre les deux assemblages. Un projet de séquençage du buffle est en cours.

Porc (*Sus scrofa*) - Ne pouvant pas bénéficier d'un financement global pour le séquençage, le consortium pour le séquençage du génome du porc s'est tourné vers une approche de séquençage de clones BAC* sélectionnés après cartographie FPC* (Humphray *et al* 2007). Cette approche a permis d'obtenir une séquence avec une profondeur de 4 X tout en optimisant la couverture. Depuis, des lectures en WGS* Illumina (30 X) ont été générées (BGI : *Beijin Genomics Institute*, Chine et Sanger Institute, UK) (Archibald *et al* 2010a). La séquence s'appuie sur des cartes, notamment des cartes RH (INRA).

Mouton (*Ovis aries*) - En comparaison avec les bovins, les ressources financières disponibles pour le séquençage du génome du mouton sont limitées. Cependant, les données de cartographie générées par les divers groupes travaillant sur la génétique moléculaire de cette espèce ne sont pas négligeables (Archibald *et al* 2010b), comprenant des cartes génétiques et d'hybrides irradiés*, plusieurs centaines de milliers de SNP*, ainsi qu'un génome virtuel assemblé grâce à des séquences d'extrémités de BAC* alignées sur les génomes humain, bovin et du chien (Dalrymple *et al* 2007). Le séquençage est principalement réalisé au BGI (75 X paired-end* d'une brebis Texel ; banques de 170 pb* à 40 kb*) et à ARK genomics, UK (45 X d'un bélier Texel qui avait servi à la construction d'une banque BAC* ; inserts de 200 à 500 pb* ; mate-pairs* 3 à 8 kb* prévus). Un assemblage primaire est prévu à l'aide des données de séquençage de la brebis et les données du bélier serviront à combler les trous, à identifier des SNP* et commencer l'étude du chromosome Y.

Chèvre (*Capra aegagrus*) - Malgré une importance moindre dans le monde des ruminants du moins dans les pays développés, de nombreux travaux ont permis l'établissement de cartes génétiques et la constitution et la caractérisation de banques de clones BAC*. Un séquençage du génome est prévu au BGI et son assemblage pourra s'appuyer sur les cartes existantes.

Cheval (*Equus caballus*) - La séquence a été réalisée par la technique de Sanger* par le Broad Institute, avec production de lectures d'extrémités de plasmides* de 4 kb* (5X), de 10 kb* (1,4X), de fosmides* de 40 kb* (0,4X) et d'extrémités de BAC* (Wade *et al* 2009). Le chromosome 11 a un centromère récent sans ADN satellite. Le déséquilibre de liaison est intermédiaire entre celui observé chez l'Homme et le chien et des haplotypes de grande longueur sont conservés entre races.

Lapin (*Oryctolagus cuniculus*) - Bien que pouvant être considéré comme une espèce de moindre importance du point de vue agronomique, le lapin est utilisé dans le domaine biomédical. Son génome n'a pas été sélectionné au début pour un séquençage complet, mais sa position phylogénétique lui a permis d'être sélectionné avec les 24 autres espèces du programme «*Mammalian Genome project*» pour un séquençage «*low-coverage*» de 2X de profondeur (données consultables dans les archives Ensembl). Cette production initiale a été suivie depuis par un séquençage avec une profondeur de 7X (Ensembl). L'amélioration des statistiques de continuité est notable : contig* N50 est passé de 3,2 à 64,6 kb* et scaffold N50 de 54 kb* à 35,3 Mb*. L'INRA participe par l'intégration de données de cartographie, principalement de clones BAC* localisés par FISH* (Chantry-Darmon *et al* 2006) et à travers la cartographie du CMH. Un projet de détection de SNP* auquel participe l'INRA est coordonné par N. Ferrand (Porto, Portugal)

Volailles - Outre les intérêts biologiques et agronomiques pour lesquels les oiseaux sont séquencés, le fait qu'ils aient des caryotypes stables et que trois stratégies différentes ont été employées pour l'assemblage de leurs génomes nous éclaire sur leur influence sur la qualité des résultats.

La poule (*Gallus gallus*) est le premier animal de rente à avoir été séquencé en raison de sa position phylogéné-

tique, comme premier représentant des oiseaux et avec comme intérêt pratique l'annotation* du génome humain (Hillier *et al* 2004). Sa séquence a été mise à jour en 2006. Cette séquence est réalisée par la méthode de Sanger* et s'appuie sur des cartes génétiques, RH (INRA) et FPC*. A ce jour, les statistiques de qualité sont : plus grand contig* = 442 kb* ; plus grand scaffold = 33 Mb* ; contig* N50 = 36 kb* ; scaffold N50 = 7,1 Mb*. La séquence de la dinde (*Meleagris gallopavo*) vient d'être publiée (Dalloul *et al* 2010) en utilisant une combinaison de deux techniques NGS : Roche 454 et Illumina GAII, s'appuyant sur une carte FPC* et une carte génétique. L'analyse des données montre que 4,6% de l'assemblage ne sont couverts que par une seule des technologies (2,3% chacune), suggérant l'intérêt de l'utilisation systématique des deux plates-formes pour l'assemblage de futurs génomes. Cependant, la taille des contigs* et scaffolds* est inférieure à celles obtenues pour la poule en Sanger* (plus grand contig* = 90 kb* ; plus grand scaffold = 9 Mb* ; contig* N50 = 12,6 kb* ; Scaffold N50 = 1,5 Mb*). Les estimations de coût sont de plus de $10 millions pour la poule et de moins de $250 000 pour la dinde.

Les microchromosomes, spécificité des génomes aviaires, sont mal couverts voire absents pour les trois génomes d'oiseaux publiés : poule, diamant mandarin et dinde.

Un séquençage du canard (*Anas platyrhynchos)* entièrement réalisé en Illumina GAII (BGI, Chine) et assemblé sans l'appui de cartes est en cours d'analyse. Les scaffolds* sont de petite taille (plus grand contig* = 263,7 kb* ; plus grand scaffold = 5,9 Mb*, contig* N50 = 26 Kb* et scaffold N50 = 1,2 Mb*) (Y. Huang communication personnelle). Finalement, la production de données pour la caille vient d'être initiée avec 15 X produits en Illumina GAII (D. Burt communication personnelle).

Poissons - Les génomes de cinq espèces de poissons présentant des intérêts pour la biologie fondamentale ont rapidement été séquencés. Le fugu (*Takifugu rubripes*) et le tétraodon (*Tetraodon nigroviridis*) (Aparicio *et al* 2002, Jaillon *et al* 2004) en raison de la compaction de leur génome et avec pour seul but l'annotation* du génome humain (Roest Crollius *et al* 2000) ; le médaka (*Oryzias latipes*) (Kasahara *et al* 2007) et le poisson zèbre ou zebrafish (*Danio rerio*) (http://www.sanger.ac.uk/Projects/D_rerio/) pour leur intérêt en biologie du développement et l'épinoche (*Gasterosteus aculeatus*) (http://www.broadinstitute.org/models/stickleback) pour l'étude de nombreux caractères en relation avec la spéciation.

Plus récemment, des projets de séquençage d'espèces d'intérêt agronomique ont été entrepris :

- Tilapia du Nil (*Oreochromis niloticus*) : séquençage 7 X (*Broad Institute*, USA) ; cartes génétiques (Cnaani *et al* 2004) ; cartes physiques (Katagiri *et al* 2005) ;

- Bar (loup) (*Dicentrarchus labrax*) : séquence d'extrémités de BAC* alignées sur la séquence de l'épinoche, recherche de SNP*, séquençage 3 X en Sanger*, 3 X en Roche 454 et 20 X en Illumina (Kuhl *et al* 2010a, 2010b) ; cartes RH (Guyon *et al* 2010) ;

- Morue de l'atlantique ou cabillaud (*Gadus morhua*) : carte génétique* (Hubert *et al* 2010) et séquençage 15 X Roche 454 en cours, suivi de paired-ends* en Sanger* (Johansen *et al* 2009) ;

- Poisson chat (*Ictalurus punctatus*) : cartes génétiques, physiques (BAC*), séquençage NGS commencé (USDA, USA) (Liu 2010) ;

- Saumon atlantique (*Salmo salar*) : cartes génétiques, SNP*, clones BAC* (Lorenz *et al* 2010), séquençage prévu 4 X d'extrémités de plasmides*, fosmides* et clones BAC* (janvier 2011, Beckman Coulter), suivi de séquençage NGS et intégration avec les cartes génétiques et de clones BAC* (Davidson *et al* 2010) ;

- Truite arc-en-ciel (*Oncorhynchus mykiss*) : cartes génétiques (Guyomard *et al* 2006, Rexroad *et al* 2008), cartes physiques (Palti *et al* 2009, 2011), séquences d'extrémités de BAC* (Genet *et al* 2011). Le séquençage est en cours (INRA et Genoscope, France).

Les distances évolutives existant entre les diverses espèces de poisson sont très importantes et limitent les possibilités d'alignement d'un génome sur l'autre. Le séquençage d'une espèce doit donc être complet et ancré sur des cartes. Par exemple, la divergence entre le saumon et la truite arc-en-ciel, deux espèces qu'on considère comme proches, est estimée à 75 millions d'années, et est de plus de 200 millions d'années entre les salmonidés et le plus proche des cinq premiers poissons séquencés (Palti *et al* 2009). En comparaison, la divergence entre mammifères est souvent inférieure à 100 millions d'années.

5 / Utilisation des séquences génomiques

L'utilisation d'une séquence de génome est devenue indispensable pour les recherches en génomique et des retombées pratiques sont déjà présentes pour certaines espèces.

5.1 / Structure du génome, gènes, annotation*, expression

Le but premier de la séquence complète d'un génome est de réaliser un inventaire complet des gènes d'une espèce. Cependant, la séquence seule n'est qu'une suite de lettres et la détection des gènes passe par des méthodes faisant appel pour la plupart à des données supplémentaires telles que la séquence de transcrits et d'EST* ou la comparaison de séquences entre espèces. L'alignement de séquences de transcrits sur le génome permet de repérer les exons* et donc les gènes, tandis que la comparaison des génomes de plusieurs espèces permet la mise en évidence des séquences conservées et donc la détection de gènes, mais aussi de séquences non-codantes biologiquement importantes telles que les gènes non-codants ou les régions régulatrices. La connaissance de la séquence du génome et son annotation permet donc de proposer rapidement des listes de gènes candidats lors des approches de clonage positionnel basées sur leur proximité physique avec des marqueurs génétiques liés à des caractères d'intérêt.

Un apport important de l'assemblage de génomes, est de permettre d'accéder aux séquences régulatrices des gènes, qui peuvent être souvent impliquées dans la variation de caractères phénotypiques en modulant les niveaux d'expression.

5.2 / Génétique, sélection, gestion des populations

La détection de variations interindividuelles de séquences, ou polymorphisme, est réalisée en alignant des séquences d'individus ou de mélanges d'individus de populations différentes sur la séquence de référence. Les différences ainsi détectées sont pour la plupart de type SNP*, mais peuvent également être de courtes insertions ou délétions nucléotidiques (Indels*), des variations du nombre de copies de séquences pouvant ou non contenir des gènes (CNV*), voire des inversions de fragments de séquence de courte taille

ou pouvant aller jusqu'à plus d'un Mb (Bansal *et al* 2007) (Encadré 1E). La plupart des polymorphismes n'ont pas ou peu de conséquences phénotypiques, mais leur étude détaillée permet d'une part de détecter puis de préciser les régions du génome impliquées dans l'expression de caractères phénotypiques et d'autre part de comprendre l'historique des populations naturelles ou sélectionnées (variations d'effectifs, sélection, migrations, croisements…). Ainsi, les études de re-séquençage de populations pour l'étude du polymorphisme permettent maintenant de détecter des régions du génome ayant subi des pressions de sélection importantes, au point d'entraîner une diminution locale de la variabilité. Par exemple, une étude récente de re-séquençage de mélanges d'ADN génomique d'individus de lignées de poules de chair et de ponte, ainsi que de poule de jungle, a permis d'obtenir une résolution allant jusqu'au niveau du gène pour des gènes sélectionnés depuis la domestication de la poule, tels que le gène *BCDO2*, responsable de la coloration jaune de la peau ou le gène *TSHR*, connu pour son implication dans le contrôle photopériodique de la reproduction. D'autres gènes potentiellement impliqués dans les différences entre les lignées de type chair et ponte, dont certains co-localisés avec des QTL ont été mis en évidence dans cette étude (Rubin *et al* 2010).

La détection de plusieurs millions de SNP* puis leur intégration dans des outils de génotypage à grande échelle et à coût raisonnable permettent de mettre en œuvre des méthodes de sélection prenant en compte l'information moléculaire, telle que la sélection génomique ou GWMAS (*Genome Wide Marker Assisted Selection*).

5.3 / Conservation évolutive, phylogénomique

La comparaison simultanée des génomes de plusieurs espèces permet la détection de séquences plus ou moins bien conservées au cours de l'évolution. En effet, deux facteurs majeurs agissant sur le degré de conservation de séquences d'ADN sont d'une part la distance phylogénétique entre les espèces étudiées : plus deux espèces sont éloignées, plus leurs séquences seront différentes du fait des mutations, et d'autre part la pression de sélection agissant sur la vitesse d'évolution des séquences du fait de leur importance fonctionnelle. Le plus souvent, cette pression de sélection sera négative, empêchant la diversification des séquences par mutation, afin de conserver les fonctions vitales. Il existe cependant des pressions de sélection positives, tendant à favoriser la diversification des séquences.

Typiquement, les signaux de conservation détectés entre espèces éloignées, telles que les poissons et les mammifères, correspondront le plus souvent aux exons* de gènes (Roest Crollius *et al* 2000), permettant la détection des nombreux gènes pour lesquels la conservation structurale et fonctionnelle des protéines codées est nécessaire. Cependant, à ce niveau de distance phylogénétique, seule une faible part de la séquence des génomes pourra effectivement être alignée, entraînant une sensibilité relativement faible de détection (tous les gènes ne seront pas détectés). A l'inverse, entre espèces très proches telles que des primates, les séquences d'ADN n'auront pas encore eu le temps de diverger par mutations et bien que la sensibilité sera élevée, les séquences s'alignant avec un fort taux de similarité sur une part importante du génome, la réalité biologique de ces alignements et donc leur spécificité sera faible : une conservation élevée de séquence entre deux espèces de primates ne sera pas forcément un bon indicateur de l'importance de sa fonction. Des espèces intermédiaires peuvent être utilisées et en utilisant un plus grand nombre d'espèces proches et des alignements multiples, le cumul des séquences permet la détection d'un signal phylogénétique (figure 7). Certain gènes sont mieux conservés que d'autres au cours de l'évolution et la vitesse d'évolution des gènes peut appuyer une hypothèse que l'on aura sur leur type de fonction. Par exemple, les gènes impliqués dans des fonctions métaboliques de base ont tendance à varier moins rapidement au cours de l'évolution que d'autres impliqués dans les mécanismes de résistance aux pathogènes, qui auront une évolution plus rapide. Parfois, des régions en dehors de gènes connus sont très conservées, suggérant une importance fonctionnelle autre : gène non-codant, région régulatrice, etc. (figure 7).

Les comparaisons de génomes inter-espèces permettent également de mettre en évidence des différences de composition en gènes. Des gènes pourront être absents chez certaines espèces, suggérant une perte de la fonction correspondante, des familles de gènes subiront des expansions ou des contractions du nombre de leurs membres. Bien que ces différences puissent suggérer des pistes sur les différences physiologiques entre espèces ou sur les fonctionnalités de gènes, il faut avoir à l'esprit que les séquences de génomes peuvent être incomplètes et il faut le plus souvent appuyer une analyse préliminaire par des travaux plus poussés tels que l'analyse des gènes voisins sur la séquence et/ou le re-séquençage des régions concernées.

5.4 / Epigénétique

Pour une espèce, un individu donné, un même génome permet d'exprimer une multitude de phénotypes cellulaires. En effet, lors de la différenciation, le même génome exprimera un répertoire différent de gènes en fonction des tissus, de la réponse à des signaux internes à l'organisme, ou venant de l'environnement. Ces variations de répertoire exprimé sont le plus souvent dues à des modifications autour du génome (modifications épigénétiques) et non à des modifications de la séquence (sauf dans de très rares cas, tels que les recombinaisons somatiques V(D)J des gènes des immunoglobulines). Un mécanisme épigénétique fréquent et facile à aborder, grâce à des techniques de séquençage spécifiques permettant de repérer les bases méthylées, est le niveau de méthylation de l'ADN, qui joue un rôle dans la conformation de la chromatine et l'accessibilité de l'ADN aux facteurs de transcription. En particulier, des dinucléotides CpG dans des promoteurs de gènes, dont les cytosines ne sont pas méthylées, suggèrent un état potentiellement actif (Jammes et Renard 2010). D'autres marques épigénétiques importantes sont les états d'acétylation et de méthylation des histones, autours desquels l'ADN est enroulé. Des techniques d'immuno-précipitation d'ADN à l'aide d'anticorps spécifiques des différents états des histones, suivies de séquençage, permettent d'étudier précisément les différents états possibles de la chromatine et nécessitent l'alignement sur un génome de référence.

Conclusion

Les techniques de séquençage sont en évolution constante et les progrès réalisés pour les espèces d'intérêt agronomique, sont très rapides. Cependant, le degré de finition pour chacune des espèces sera variable en fonction de la technologie utilisée, mais aussi de la possibilité de générer et d'intégrer des données de cartographie. Chaque technique de cartographie des génomes a ses avantages et inconvénients et participe à l'assignation chromosomique, l'ordonnancement des séquences le long des chromosomes, la détection et la correction d'erreurs, etc.

En plus du séquençage de nouvelles espèces, les technologies de séquençage parallèle permettent, grâce à une diminution des coûts, aux volumes de données produits et au développement de logiciels appropriés, d'envisager le re-séquençage de génomes en multipliant les individus, groupes d'individus ou populations analysés, qui seront comparés aux génomes de référence. Ceci

ouvre des perspectives nouvelles, permettant par exemple de voir les effets de la sélection sur la structure des génomes, pointant directement dans certains cas vers les gènes sélectionnés. En revanche, la masse de données produites pose la question de leur stockage et traitement informatique.

L'évolution des techniques de séquençage est rapide et la capacité des séquenceurs parallèles croît rapidement, nécessitant des adaptations fréquentes tant au niveau de la production des données, qui nécessite une technicité certaine pour la production des banques de séquençage et l'utilisation des séquenceurs, qu'au niveau de leur analyse par utilisation de logiciels développés très récemment et en constante évolution. Au niveau de la stratégie à adopter pour un projet de séquençage, de nombreuses options sont possibles, en fonction de la question posée et des moyens disponibles, tant financiers que biologiques (disponibilité des échantillons). Le séquençage *de novo* posera des problèmes différents du re-séquençage avec alignement sur un génome de référence. L'alignement sur un génome de référence nécessitera des analyses plus ou moins sophistiquées, selon qu'il s'agisse de la même espèce ou d'une espèce plus ou moins proche. Le séquençage parallèle sera utilisé de plus en plus pour les études d'expression de gènes (RNAseq) par séquençage de transcrits, ce qui pose de nouveau des problèmes spécifiques de définition des plans d'expérience tels que le nombre d'échantillons, de répétitions, ainsi que des problèmes spécifiques pour l'analyse des données, tels que les normalisations, la prise en compte des exons, de la couverture partielle des transcrits et de l'épissage alternatif. Ces études de RNAseq permettront d'améliorer notablement l'annotation structurale et fonctionnelle des gènes, ainsi que d'aborder au niveau du génome des questions nouvelles telles que la détermination de l'origine parentale des transcrits exprimés (sont-ils plus exprimés à partir du chromosome d'un des deux parents).

Finalement, une troisième génération de séquenceurs est annoncée, produisant les données de séquençage en temps réel, c'est-à-dire au fur et à mesure que l'ADN polymérase incorpore des nucléotides, ou lors du passage d'une molécule d'ADN à travers une nano pore. Cette nouvelle génération de séquenceurs, si elle tient ses promesses, devrait entraîner une nouvelle diminution des coûts, l'obtention beaucoup plus rapide des données de séquence et des lectures plus longues. Ceci pourrait ouvrir les possibilités d'utilisation du séquençage comme outil de diagnostic et permettre la recherche de génotypes de variants rares dans des populations.

Glossaire

Annotation structurale : repérage des coordonnées des diverses structures dans le génome, telles que les gènes.

Annotation fonctionnelle : renseignements sur les fonctions des séquences, le plus souvent pour les gènes.

BAC : *Bacterial Artificial Chromosome*. Vecteur de clonage permettant l'obtention de clones bactériens contenant un grand fragment d'ADN génomique (taille > 100 kb*). Les BAC assemblés en contigs* sont à la base des cartes physiques du génome.

Carte cytogénétique : carte des chromosomes. Réalisée par localisation visuelle (FISH*) au microscope de fragments d'ADN sur les chromosomes au stade métaphase de la mitose.

Carte d'hybrides irradiés : réalisée en testant par PCR la présence ou l'absence de fragments d'ADN dans une collection de clones d'hybrides irradiés (RH*). Deux fragments d'ADN sont proches sur le génome s'ils sont trouvés fréquemment dans les mêmes clones.

Carte génétique : obtenue par l'étude de la ségrégation dans des familles ou des populations, de marqueurs polymorphes, soit moléculaires, soit phénotypiques, deux séquences étant d'autant plus proches qu'elles sont souvent transmises ensemble lors de la méiose.

Clonage positionnel : stratégie visant à identifier un gène responsable de l'expression d'un phénotype en utilisant des informations de position sur le génome.

Contig : ensemble de clones (le plus souvent des BAC*) ou de lectures de séquence ordonnés grâce à des informations sur leur parties chevauchantes.

Cosmide : vecteur de clonage permettant l'obtention de clones bactériens contenant des fragment d'ADN génomique de taille avoisinant les 50 kb*.

CNV : *Copy Number Variation.* Polymorphisme du génome correspondant à la variation du nombre de copies d'une séquence, pouvant dans certains cas contenir un ou plusieurs gènes.

EST : *Expressed Sequence Tag*. Séquences étiquettes (partielles) de transcrit, obtenues par séquençage aléatoire d'ARN.

Expression génique : études visant à estimer le niveau de production (expression) des gènes en fonction d'états physiologiques ou de tissus différents.

Exon : fraction de la partie codante d'un gène eucaryote. Les gènes des organismes eucaryotes sont le plus souvent fractionnés en plusieurs séquences d'ADN dans le génome, les exons, séparés entre eux par d'autres séquences (introns*).

FISH : *Fluorescent In Situ Hybridisation*. Hybridation de sondes d'ADN marquées à l'aide d'un fluorochrome, sur des chromosomes au stade métaphase de la mitose. Permet la réalisation de la carte cytogénétique.

Fingerprinting : technique permettant d'estimer très grossièrement la similarité entre des séquences d'ADN sans les séquencer, par la comparaison des longueurs de bandes produites par des enzymes de restriction coupant l'ADN à des sites précis.

Fosmide : vecteur de clonage permettant l'obtention de clones bactériens contenant des fragment d'ADN génomique de taille déterminée et égale à 40 kb*.

FPC : *FingerPrint Contig**. Contig* de clones (généralement des BAC*) ordonnés par la technique du fingerprinting, afin d'obtenir une carte physique du génome.

Homologues : séquences similaires en raison d'une origine évolutive commune.

Hybride irradié : cellule hybride obtenue par fusion entre cellules hôte d'une espèce et donneuse d'une autre espèce, contenant une fraction aléatoire du génome de l'espèce donneuse, après cassures par irradiation, reconstitution aléatoire de chromosomes ou insertion dans des chromosomes de la cellule hôte et rétention partielle. Deux séquences proches sur le génome sont en probabilité dans les mêmes clones RH, tandis que deux séquences distantes ont une probabilité faible d'être conservées ensemble.

Indel : *Insertion-deletion*. Polymorphisme de présence ou absence d'un ou plusieurs nucléotides.

Intron : séquence non-codante dans les gènes, séparant les exons, qui codent pour une protéine.

Kb : kilobase ; séquence de mille paires de bases (pb*).

Mate-pair : séquences appariées (1 à 10 kb* de distance), produites en circularisant les fragments d'ADN, puis par séquençage à travers le point de jointure.

Mb : mégabase ; séquence d'un million de paires de bases (pb*) de longueur.

Paired end : séquences appariées produites par la lecture des deux extrémités de courts fragments d'ADN (moins de 500 pb dans le cas des nouvelles technologies de séquençage).

Paralogues : séquences homologues* résultat de la duplication d'une séquence ancestrale dans le génome. Il s'agit de deux (ou plus) séquences similaires par homologie dans un même génome.

Pb : paire de base ; unité de séquence d'ADN, représentée par une base et sa complémentaire-inverse sur l'autre brin.

Phylogénomique : utilise les méthodes de la génomique et de la phylogénie. Par la comparaison de génomes entiers, permet de mettre en évidence des pertes et gains de gènes dans les génomes, ainsi que leur variabilité moléculaire, afin (entre autres buts) d'aider à prédire leur fonctions.

Plasmide : vecteur de clonage permettant l'obtention de clones bactériens contenant des fragment d'ADN génomique de taille allant de 500 pb* à 10 kb* environs.

Orthologues : séquences homologues* entre deux espèces.

RH : *Radiation Hybrid* (hybride irradié*)

Sanger (méthode de) : Méthode de séquençage publiée en 1977 (Sanger *et al* 1977) et encore utilisée de nos jours avec les séquenceurs à électrophorèse capillaire.

Scaffold : ensemble de contigs* de séquence reliés entre eux par des informations apportées par des lectures appariées (*mate-pairs** ou *paired-ends**).

Supercontig : nom alternatif pour les scaffolds*.

SNP : *Single Nucleotide Polymorphism*. Polymorphisme ponctuel de nucléotide : substitution d'une base par une autre dans une séquence.

WGS : *Whole Genome Shotgun*. Production de lectures de séquence d'un génome entier de manière aléatoire.

Références

Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F. *et al*, 2010. Integrating common and rare genetic variation in diverse human populations. Nature, 467, 52-58.

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M. *et al*, 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science, 297, 1301-1310.

Archibald A.L., Bolund L., Churcher C., Fredholm M., Groenen M.A., Harlizius B., Lee K.T., Milan D., Rogers J., Rothschild M.F., Uenishi H., Wang J., Schook L.B., 2010a. Pig genome sequence--analysis and publication strategy. BMC Genomics, 11, 438.

Archibald A.L., Cockett N.E., Dalrymple B.P., Faraut T., Kijas J.W., Maddox J.F., Mcewan J.C., Hutton Oddy V., Raadsma H.W., Wade C., Wang J., Wang W., Xun X., 2010b. The sheep genome reference sequence: a work in progress. Anim. Genet., 41, 449-453.

Bansal V., Bashir A., Bafna V., 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res., 17, 219-230.

Batzoglou S., Jaffe D.B., Stanley K., Butler J., Gnerre S., Mauceli E., Berger B., Mesirov J.P., Lander E.S., 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res., 12, 177-189.

Chantry-Darmon C., Urien C., De Rochambeau H., Allain D., Pena B., Hayes H., Grohs C., Cribiu E.P., Deretz-Picoulet S., Larzul C., Save J.C., Neau A., Chardon P., Rogel-Gaillard C., 2006. A first-generation microsatellite-based integrated genetic and cytogenetic map for the European rabbit (*Oryctolagus cuniculus*) and localization of angora and albino. Anim. Genet., 37, 335-341.

Church D.M., Goodstadt L., Hillier L.W., Zody M.C., Goldstein S. *et al*, 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol., 7, e1000112.

Church D.M., Hillier L.W., 2009. Back to Bermuda: how is science best served? Genome Biol., 10, 105.

Cnaani A., Zilberman N., Tinman S., Hulata G., Ron M., 2004. Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. Mol. Genet. Genomics, 272, 162-172.

Dalloul R.A., Long J.A., Zimin A.V., Aslam L., Beal K. *et al*, 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biol., 8, pii: e1000475.

Dalrymple B.P., Kirkness E.F., Nefedov M., Mcwilliam S., Ratnakumar A., Barris W., Zhao S., Shetty J., Maddox J.F., O'grady M., Nicholas F., Crawford A.M., Smith T., De Jong P.J., Mcewan J., Oddy V.H., Cockett N.E., 2007. Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. Genome Biol., 8, R152.

Davidson W.S., Koop B.F., Jones S.J., Iturra P., Vidal R., Maass A., Jonassen I., Lien. S., Omholt S.W., 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol, 11, 403.

Ewing B., Green P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res., 8, 186-194.

Ewing B., Hillier L., Wendl M.C., Green P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res., 8, 175-185.

Genet C., Dehais P., Palti Y., Gao G., Gavory F., Wincker P., Quillet E., Boussaha M., 2011. Analysis of BAC-end sequences in rainbow trout: content characterization and assessment of synteny between trout and other fish genomes. BMC Genomics, 12, 314.

Gibbs R.A., Weinstock G.M., Metzker M.L., Muzny D.M., Sodergren E.J. *et al*, 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 428, 493-521.

Gnerre S., Maccallum I., Przybylski D., Ribeiro F.J., Burton J.N. *et al*, 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA, 108, 1513-1518.

Guyomard R., Mauger S., Tabet-Canale K., Martineau S., Genet C., Krieg F., Quillet E., 2006. A type I and type II microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) with presumptive coverage of all chromosome arms. BMC Genomics, 7, 302.

Guyon R., Senger F., Rakotomanga M., Sadequi N., Volckaert F.A., Hitte C., Galibert F., 2010. A radiation hybrid map of the European sea bass (*Dicentrarchus labrax*) based on 1581 markers: Synteny analysis with model fish genomes. Genomics, 96, 228-238.

Havlak P., Chen R., Durbin K.J., Egan A., Ren Y., Song X.Z., Weinstock G.M., Gibbs R.A., 2004. The Atlas genome assembly system. Genome Res., 14, 721-732.

Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C. *et al*, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 432, 695-716.

Huang X., Wang J., Aluru S., Yang S.P., Hillier L., 2003. PCAP: a whole-genome assembly program. Genome Res., 13, 2164-2170.

Hubert S., Higgins B., Borza T., Bowman S., 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). BMC Genomics, 11, 191.

Humphray S.J., Scott C.E., Clark R., Marron B., Bender C. *et al*, 2007. A high utility integrated map of the pig genome. Genome Biol., 8, R139.

Ihgsc, 2004. Finishing the euchromatic sequence of the human genome. Nature, 431, 931-945.

Istrail S., Sutton G.G., Florea L., Halpern A.L., Mobarry C.M. *et al*, 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc. Natl. Acad. Sci. USA, 101, 1916-1921.

Jaillon O., Aury J.M., Brunet F., Petit J.L., Stange-Thomann N. *et al*, 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature, 431, 946-957.

Jammes H., Renard J.P., 2010. Epigénétique et construction du phénotype, un enjeu pour les productions animales ? In : Robustesse, rusticité, flexibilité, plasticité, résilience… les nouveaux critères de qualité des animaux et des systèmes d'élevage. Sauvant D., Perez J.M. (Eds). Dossier, INRA Prod. Anim., 23, 23-42.

Johansen S.D., Coucheron D.H., Andreassen M., Karlsen B.O., Furmanek T., Jorgensen T.E., Emblem A., Breines R., Nordeide J.T., Moum T., Nederbragt A.J., Stenseth N.C., Jakobsen K.S., 2009. Large-scale sequence analyses of Atlantic cod. N Biotechnol., 25, 263-271.

Kasahara M., Naruse K., Sasaki S., Nakatani Y., Qu W. *et al*, 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature, 447, 714-719.

Katagiri T., Kidd C., Tomasino E., Davis J.T., Wishon C., Stern J.E., Carleton K.L., Howe A.E., Kocher T.D., 2005. A BAC-based physical map of the Nile tilapia genome. BMC Genomics, 6, 89.

Kuhl H., Beck A., Wozniak G., Canario A.V., Volckaert F.A., Reinhardt R., 2010a. The European sea bass Dicentrarchus labrax genome puzzle: comparative BAC-mapping and low coverage shotgun sequencing. BMC Genomics, 11, 68.

Kuhl H., Tine M., Hecht J., Knaust F., Reinhardt R., 2011. Analysis of single nucleotide polymorphisms in three chromosomes of European sea bass *Dicentrarchus labrax*. Comp.. Biochem. Physiol. Part D Genomics Proteomics, ***6, 70-75.***

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C. *et al*, 2001. Initial sequencing and analysis of the human genome. Nature, 409, 860-921.

Li R., Fan W., Tian G., Zhu H., He L. *et al*, 2010. The sequence and de novo assembly of the giant panda genome. Nature, 463, 311-317.

Liu Y., Qin X., Song X.Z., Jiang H., Shen Y., Durbin K.J., Lien S., Kent M.P., Sodeland M., Ren Y., Zhang L., Sodergren E., Havlak P., Worley K.C., Weinstock G.M., Gibbs R.A., 2009. *Bos taurus* genome assembly. BMC Genomics, 10, 180.

Liu Z., 2010. Development of genomic resources in support of sequencing, assembly, and annotation of the catfish genome. Comp. Biochem. Physiol. Part D Genomics Proteomics.

Lorenz S., Brenna-Hansen S., Moen T., Roseth A., Davidson W.S., Omholt S.W., Lien S., 2010. BAC-based upgrading and physical integration of a genetic SNP map in Atlantic salmon. Anim. Genet., 41, 48-54.

Miller J.R., Koren S., Sutton G., 2010. Assembly algorithms for next-generation sequencing data. Genomics, 95, 315-327.

Mullikin J.C., Ning Z., 2003. The phusion assembler. Genome Res., 13, 81-90.

Palti Y., Luo M.C., Hu Y., Genet C., You F.M., Vallejo R.L., Thorgaard G.H., Wheeler P.A., Rexroad C.E., 3rd, 2009. A first generation BAC-based physical map of the rainbow trout genome. BMC Genomics, 10, 462.

Palti Y., Genet C., Luo M.C., Charlet A., Gao G., Hu Y., Castano-Sanchez C., Tabet-Canale K., Krieg F., Yao J., Vallejo R.L., Rexroad C.E., 2011. A first generation integrated map of the rainbow trout genome. BMC Genomics, 12, 180.

Pheasant M., Mattick J.S., 2007. Raising the estimate of functional human sequences. Genome Res., 17, 1245-1253.

Rexroad C.E., 3rd, Palti Y., Gahr S.A., Vallejo R.L., 2008. A second generation genetic map for rainbow trout (*Oncorhynchus mykiss*). BMC Genet., 9, 74.

Roest Crollius H., Jaillon O., Bernot A., Dasilva C., Bouneau L., Fischer C., Fizames C., Wincker P., Brottier P., Quetier F., Saurin W., Weissenbach J., 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nat. Genet., 25, 235-238.

Rubin C.J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T., Ka S., Hallbook F., Besnier F., Carlborg O., Bed'hom B., Tixier-Boichard M., Jensen P., Siegel P., Lindblad-Toh K., Andersson L., 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature, 464, 587-591.

Sanger F., Nicklen S., Coulson A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA, 74, 5463-5467.

Southan C., 2004. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics, 4, 1712-1726.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J. *et al*, 2001. The sequence of the human genome. Science, 291, 1304-1351.

Wade C.M., Giulotto E., Sigurdsson S., Zoli M., Gnerre S. *et al*, 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science, 326, 865-867.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F. *et al*, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420, 520-562.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R., Schatz M.C., Puiu D., Hanrahan F., Pertea G., Van Tassell C.P., Sonstegard T.S., Marcais G., Roberts M., Subramanian P., Yorke J.A., Salzberg S.L., 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol., 10, R42.

Résumé

Depuis la première publication du génome humain en 2001, les génomes de nombreuses espèces animales d'intérêt agronomique ont été séquencés ou sont en cours de séquençage. Après la réalisation de l'assemblage des génomes pour les espèces majeures par la méthode de Sanger* et des séquenceurs capillaires, la nouvelle génération de séquenceurs parallèles (NGS : Next Generation Sequencing) permet maintenant le séquençage rapide et à bas coût d'un nombre bien plus grand d'espèces. Un rappel sur les divers principes du séquençage et de l'assemblage des génomes, permettant de mieux comprendre leurs avantages et leurs limites, est suivi d'une description de l'état d'avancement de la connaissance des génomes des animaux d'intérêt agronomique. Finalement, quelques exemples de domaines d'utilisation des génomes assemblés sont rapidement décrits.

Abstract

Current state of genome sequencing in animal species

Since the first publication of the human genome in 2001, many animal species of agronomic interest have been sequenced or are being sequenced. After the production of genome assemblies by the Sanger* method with capillary sequencers, the new generation of parallel sequencing machines (NGS : Next Generation Sequencing), now allows the rapid production of assemblies for a much larger number of species at low cost. An overview of the diverse principles of sequence production and assembly, highlighting their advantages and limits, is followed by a description of the state of progress of the knowledge in animal species of agronomic importance. Finally, a few examples of application areas of genome assemblies are described.

VIGNAL A., 2011. Etat actuel du séquençage et de la connaissance du génome des espèces animales. In : Numéro spécial, Amélioration génétique. Mulsant P., Bodin L., Coudurier B., Deretz S., Le Roy P., Quillet E., Perez J.M. (Eds). INRA Prod. Anim., 24, 387-404.