2014. Journées Recherche Porcine, 46, 7-18.

Le séquençage du génome du porc : apport des nouvelles technologies de séquençage à la génomique et à la génétique porcines

Bertrand SERVIN (1), Claire ROGEL-GAILLARD (2), Denis MILAN (1), Jean-Pierre BIDANEL (2) Juliette RIQUET (1)

(1) INRA, UMR444, LGC, 31320 Toulouse, France
(2) INRA, UMR1313, GABI, 78352 Jouy-en-Josas, France

bertrand.servin@toulouse.inra.fr, juliette.riquet@toulouse.inra.fr

Le séquençage du génome du porc : apport des nouvelles technologies de séquençage à la génomique et la génétique porcines

Depuis une vingtaine d'années, l'évolution des technologies de biologie moléculaire a permis de révolutionner nos connaissances sur les génomes et, chez les animaux domestiques, d'améliorer significativement la sélection via des approches de sélection génomique (notamment chez les bovins laitiers). Parmi les développements technologiques majeurs, la mise au point d'une nouvelle génération de séquenceurs (NGS : Next Generation Sequencing) permet, depuis 2006, de déterminer la séquence individuelle de n'importe quel individu. En 2008, 100 fois plus rapidement que pour l'obtention de la première séquence du génome humain et pour un coût également divisé par 100, une seconde séquence humaine a été publiée (Nature 452, no.7189 (2008) : 872-876). Depuis, les données issues de cette technologie haut débit (HTS : High Throughput Sequencing) se sont accumulées pour de très nombreuses espèces et trouvent des applications dans des domaines de recherche fondamentale et appliquée extrêmement variés. En 2012, une première séquence du génome porcin a été produite dans le cadre d'un consortium international (Nature 491, no. 7424 (2012) : 393-398). Le but de cette communication est de faire un état des lieux de l'apport de cette technologie à l'étude des génomes, plus particulièrement chez le porc. En s'appuyant sur les travaux réalisés chez d'autres espèces, une réflexion prospective sur l'impact de ces nouvelles données génomiques en génétique et en sélection porcine sera proposée.

Genome sequencing: contribution of new sequencing technologies in swine genetics and genomics

For twenty years, the development of new technologies in molecular biology have revolutionized our knowledge of genomes and, in livestock species, have significantly improved selection via new selection genomic approaches (essentially in dairy cattle). Among the major technologies, the development of a new generation of sequencers (NGS: Next Generation Sequencing) since 2006, allows the genome sequence of any individual to be acquired. In 2008, 100 times faster than for the first human genome sequence, and also for a cost divided by 100, a second human sequence was published (Nature 452, no.7189 (2008) : 872-876). Since this period, the data from this high throughput technology (HTS: High Throughput Sequencing) have been accumulated for many species and new extremely varied applications in fundamental and applied researches have been developed. In 2012, a first sequence of the pig genome was produced as part of an international consortium (Nature 491, no. 7424 (2012): 393-398). The purpose of this communication is to make an inventory of the contribution of this technology to the study of genomes, particularly in pigs. Based on data obtained in other species, a prospective view on the impact of these new genomic data in genetics and pig breeding is also proposed.

INTRODUCTION

En 1977, deux techniques permettant de déterminer base à base la composition nucléotidique d'une séquence d'ADN *(Acide Désoxyribonucléique, support de l'information héréditaire)* ont été développées indépendamment. Pour cette découverte majeure en biologie moléculaire, Walter Gilbert et Frederick Sanger ont été récompensés en 1980 par le prix Nobel de chimie. Dans les années qui suivirent, alors que la technique dite de Maxam et Gilbert (basée sur une succession de modifications chimiques et de clivages de l'ADN) était progressivement abandonnée, l'utilisation de la technique de séquençage "Sanger" s'est généralisée et a bénéficié de développements améliorant son efficacité et son innocuité pour le manipulateur. Dans les années 90, des séquenceurs automatiques permettant d'obtenir simultanément par la technologie Sanger, jusqu'à 96 lectures de 600 bases ont été commercialisés. Les performances de ces séquenceurs de première génération ont alors permis d'entreprendre le séquençage complet du génome humain dans le cadre d'un consortium international:

(http://web.ornl.gov/sci/techresources/Human_Genome/project/timeline.shtml : *lien permettant d'accéder à l'historique des évènements majeurs du projet Génome Humain*). En 2003, à l'issue de 13 années de travaux et pour un coût de plusieurs centaines de millions de dollars, une première version de la séquence du génome humain a été publiée. Parallèlement, ce vaste projet a encouragé la recherche technologique de méthodes alternatives fiables destinées à gagner du temps tout en diminuant les coûts et l'intervention humaine. Depuis 2005, de nouvelles technologies répondant à ces critères, essentiellement proposées par trois grands groupes (Roche, Illumina et Life Technologies), ont débouché sur une seconde génération de séquenceurs regroupés sous l'appellation NGS (« Next generation sequencing »). Cette seconde révolution dans le domaine du séquençage a dès lors permis l'acquisition de séquences de génome de très nombreuses espèces et débouché sur l'analyse fine et exhaustive de mécanismes nouveaux de régulation des génomes.

1. LE SEQUENÇAGE DU GENOME DU PORC

Dans la continuité des travaux réalisés chez l'homme, et bénéficiant des développements technologiques et bio-informatiques, plusieurs génomes ont été séquencés pour un coût et dans un laps de temps moindres. Rapidement après la séquence du génome humain, les séquences des génomes de la souris, du rat, de la poule, du chimpanzé, de l'opossum et du chien ont été obtenues, puis celles des génomes du bovin, du cheval et du porc. L'ensemble de ces données a été acquis entre 2002 et 2012, années au cours desquelles les technologies de séquençage nouvelle génération ont très rapidement évolué. La stratégie adoptée par les différents consortiums a donc différé en fonction des années de réalisation du séquençage. Entre les années 2000 et 2005, le séquençage d'un génome était réalisé à partir d'une carte physique de l'espèce. Le principe était, dans un premier temps, de construire une (des) collection(s) de colonies ou clones bactériens appelés BAC (« bacterial artificial chromosomes ») dans lesquelles étaient insérés de grands fragments d'ADN (160 à 180 kb) issus du génome de l'espèce étudiée.

La production de milliers de clones permettait de garantir que l'ensemble du génome étudié soit représenté, fragmenté dans ces collections. Dans un second temps l'ensemble de ces clones BAC était ordonné les uns par rapport aux autres tout au long du génome et une sous-sélection de clones garantissant une couverture homogène du génome (« Minimum Tiling Path ») était sélectionnée afin de procéder à leur séquençage. Depuis 2005, une part de plus en plus importante de données a été obtenue à partir du séquençage en parallèle de fragments d'ADN répartis aléatoirement sur le génome (stratégie WGS – « whole genome shotgun » ou séquençage aléatoire global). Cette seconde stratégie a largement bénéficié des performances associées aux séquenceurs de seconde génération et des progrès de la bio-informatique. En effet, une des difficultés du séquençage par WGS est d'ordonner les séquences d'ADN les unes par rapport aux autres et de les positionner sur le génome. Ce travail dit d'assemblage est réalisé en recherchant les séquences dites chevauchantes (qui se recouvrent partiellement). Un alignement de séquences chevauchantes permet d'obtenir des enchainements plus longs appelés contigs. Lorsqu'aucune séquence supplémentaire ne peut être ajoutée par chevauchement à un contig, un nouveau contig est construit. Bien que des trous subsistent entre contigs, il est toutefois possible d'orienter et d'ordonner certains d'entre eux sous la forme de « scaffolds » («échafaudages» : en Français, blocs *de séquence incomplète comprenant plusieurs contigs séparés par des séquences inconnues matérialisés par des N successifs*). Le nombre et la longueur de ces scaffolds sont des indicateurs de la qualité de la séquence à un instant t.

1.1. Obtention d'une séquence porcine de référence

Le Consortium de séquençage du génome du porc, regroupant universitaires, représentants de gouvernements et de l'industrie, a été créé en Septembre 2003 à Jouy-en-Josas afin d'assurer une coordination internationale pour le séquençage du génome porcin. L'objectif du consortium est de contribuer aux recherches agronomiques et biomédicales par l'acquisition de données et le développement d'outils résultant du séquençage de ce génome (Schook *et al.*, 2005).

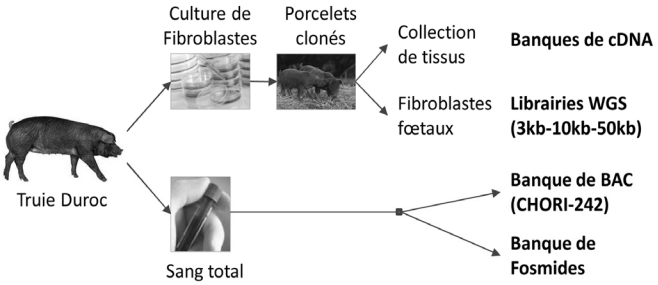

Figure 1 – Origine des différentes sources d'ADN utilisées pour obtenir la séquence de référence du génome du porc. La grande majorité des banques utilisées ont été construites à partir d'un seul animal de race Duroc, afin de faciliter l'assemblage de la séquence en limitant le nombre de variants ponctuels ou de structure.

Afin de faciliter l'obtention d'une séquence de qualité, la stratégie a consisté à n'utiliser que les séquences issues de l'ADN génomique d'un seul animal pour éviter que des remaniements de structure existant entre individus n'entraînent des erreurs d'assemblage.

Différentes banques d'ADN ont donc été réalisées à partir d'une femelle de race Duroc (Tabasco) afin de disposer de fragments d'ADN génomique de tailles différentes ou des fragments d'ADN correspondant à la partie codante du génome (Figure 1).

La stratégie de séquençage du génome du porc et l'utilisation des différentes banques sont résumés dans la Figure 2. Une première version de la séquence du génome du porc (Draft V9) a été publiée en 2009 (Humphray *et al.*, 2007). Dans un second temps l'ajout de séquences réalisées à partir de librairies WGS, et l'utilisation de séquences d'ADNc (*ADN obtenu à partir d'ARN messager, qui correspond donc à une partie codante du génome*) permettant l'annotation du génome *(détermination des éléments fonctionnels)* ont permis de proposer en 2012 une nouvelle version de la séquence de référence porcine (Draft V10.2) (Groenen *et al.*, 2012). L'apport respectif des différentes librairies à la production de la séquence de référence actuelle est résumé sur la Figure 2.

Actuellement la version V10.2 comprend 2,80 Gb de séquences agencées sous la forme de 5 343 scaffolds. La taille moyenne des scaffolds est de 436 176 bases et la valeur N50 de 637 332 *(le paramètre N50 indique que 50% des scaffolds ont une taille d'au moins la valeur indiquée)*. A une résolution plus fine, la valeur N50 des contigs est de 80 720 bases ; à titre comparatif cette valeur est supérieure à celle obtenue pour la séquence bovine (Btau3.1 : N50 = 76 449 bases) mais de qualité moindre que celle obtenue pour la séquence du cheval (EquCab2 : N50 = 112 381 pb). La qualité de l'agencement des séquences constituant cette séquence de référence a également été estimée par comparaison avec les cartes publiées du génome du porc ce qui a permis de confirmer que la qualité de cette version était bonne.

Dans les années à venir des versions améliorées de cette séquence seront certainement publiées mais d'ores et déjà, la disponibilité de la version V10.2 permet d'entreprendre des analyses inenvisageables jusqu'à présent.

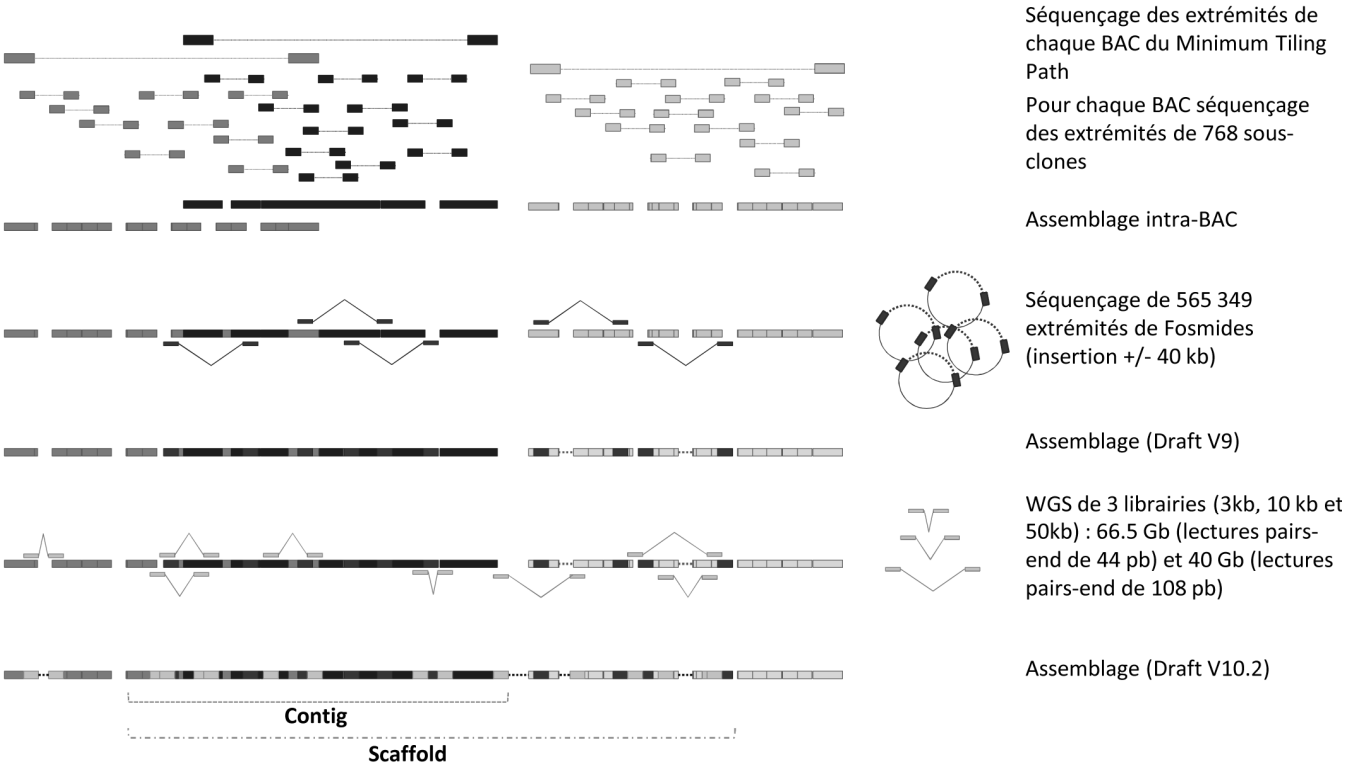

Figure 2 - Stratégie de séquençage basée sur une cartographie physique préalable du génome du porc.
Utilisation respective des différentes banques d'ADN pour compléter et affiner la qualité de l'assemblage des versions successives de la séquence de référence.

***BAC** : chromosome artificiel de bactérie ; **minimum tiling path** : nombre minimum de fragments permettant de couvrir l'ensemble du génome ; **fosmides** : plasmides pouvant contenir des inserts d'ADN de 40000 pb ; **WGS** : séquençage aléatoire global ; **contig** : ensemble de fragments de séquences chevauchantes alignées ; **scaffold** : ensemble de contigs orientés et ordonnés.*

1.2. Premières retombées du séquençage du génome du porc

1.2.1. *Construction d'une puce générique de 60 000 marqueurs génétiques*

Parallèlement au projet destiné à établir une première séquence de référence de la femelle Duroc Tabasco, des travaux complémentaires de séquençage partiel d'un plus grand nombre d'individus ont été menés afin de rechercher des variations de séquence ponctuelles (polymorphismes). Des mélanges d'échantillons d'ADN comprenant de 23 à 36 individus originaires de différents pays ont été constitués, chaque mélange représentant une race différente (Duroc, Piétrain, Landrace, Large White et sanglier). A l'issue du séquençage partiel d'une partie de l'ADN de chaque mélange, les lectures obtenues ont été alignées sur le génome de référence de Tabasco. Au total, 372 886 variations d'une seule paire de bases (SNP pour « Single Nucleotide Polymorphism »)

ont été identifiées. Des filtres successifs destinés (1) à éliminer les SNP dont un des allèles avait une très faible fréquence, (2) à choisir les marqueurs informatifs dans plusieurs races, (3) à obtenir une couverture homogène de l'ensemble du génome, ont permis de sélectionner un lot de marqueurs afin de constituer une puce générique comprenant 64 232 marqueurs SNP (Ramos *et al.*, 2009).

1.2.2. *Comparaison des génomes du porc et d'autres mammifères*

Le porc domestique est un mammifère euthérien appartenant à l'ordre des cétartiodactyles, un clade distinct de celles des rongeurs et primates, mais regroupant les sous-ordres des suinés (familles des suidés et des pécaris), les ruminants, les hippopotamidés et des cétacés. Le dernier ancêtre commun de l'homme et du porc daterait de 79 à 97 millions d'années. De nombreuses données génétiques indiquent que cette espèce serait apparue dans le sud-est asiatique au début du Pliocène (5,3 à 3,5 millions d'années), période de forte fluctuation climatique. Les études de phylogénie moléculaire sont grandement facilitées par l'accès aux séquences d'ADN et de protéines. L'alignement de la séquence du génome du porc aux séquences de sept autres génomes de mammifères a ainsi permis de mettre en évidence les blocs de synténies (segment de chromosome conservé au cours de l'évolution ; l'ordre des gènes est maintenu et identique entre plusieurs espèces) conservés entre espèces et, a contrario, les points de cassure impliqués dans la spéciation des suidés. Au total 192 régions de remaniement évolutifs (EBR : evolution breakpoint regions) ont été identifiées. L'analyse du contenu en ADN répété et en gènes de ces régions a révélé la présence de motifs répétés particuliers (LTR-ERV1 et ADN satellite), et un enrichissement en gènes connus pour être impliqués dans la perception sensitive du goût et des odeurs. L'analyse comparative des six génomes de mammifères les mieux annotés a également permis de révéler qu'en moyenne (à l'échelle du génome) les taux de mutation du génome du porc étaient comparables à ceux des autres espèces de mammifères, mais qu'une pression de sélection plus importante semblait avoir particulièrement affecté les gènes impliqués dans la réponse immunitaire. Cette évolution particulière des gènes impliqués dans l'immunité est également reflétée par leur nombre. Par rapport aux autres mammifères, des mécanismes de duplication ou d'expansion au sein d'une même famille de gènes ont induit une multiplication du nombre de gènes impliqués dans l'immunité dans un premier temps dans le génome des cétartiodactyles puis de façon spécifique et indépendante dans les génomes porcin et bovin.

1.2.3. *Le porc comme modèle biomédical*

L'analyse comparative de la séquence porcine avec la séquence de référence humaine a également permis de documenter l'utilisation possible du porc comme espèce modèle pour l'homme. Le porc est d'ores et déjà un modèle biomédical important ; la possibilité de générer des animaux transgéniques et des knockouts (*animaux dont un gène a été artificiellement inactivé*) en combinaison avec des techniques de transfert nucléaire a permis de constituer artificiellement de nombreux modèles porcins de pathologies humaines. La comparaison des séquences de ces deux espèces avait comme objectif de répertorier l'ensemble des mutations partagées naturellement par l'homme et le porc. Parmi l'ensemble des gènes annotés, 112 positions correspondant à des acides aminés connus pour être impliqués chez l'homme dans des maladies multifactorielles comme le diabète, la dyslexie, les maladies de Parkinson ou d'Alzheimer ont été identifiées et permettraient, en utilisant le porc comme modèle expérimental, d'affiner la compréhension du rôle de ces mutations dans les maladies humaines.

L'autre domaine d'utilisation du porc en médecine humaine est la xénotransplantation d'organes porcins chez l'homme. Néanmoins la présence de rétrovirus endogènes porcins (REPV) dans le génome du porc représente un risque d'infection zoonotique pour l'homme. Pour l'heure, l'analyse de la séquence de référence obtenue a permis de dresser la phylogénie des différents REPV intégrés au génome du porc. La classe des γ-PERV (et γ-PERV like) est la classe la plus fréquente suivie par la classe β. Il semble également qu'une forte homologie du groupe γ1 avec la classe γ-PERV murine soit le reflet d'une transmission récente de l'ERV-γ1 de la souris au porc.

Ces premières analyses ont été réalisées dans la cadre des travaux du consortium de séquençage. Dans la continuité de ce travail, d'autres projets de recherches ont pu être menés grâce aux technologies de séquençage de nouvelle génération. Actuellement, plusieurs technologies sont disponibles sur le marché et permettent de répondre à une gamme extrêmement large de questions de recherche en biologie et en génétique.

2. LES NOUVELLES TECHNOLOGIES DE SEQUENÇAGE

2.1. D'une technologie à l'autre

Trois plateformes de séquençage basées sur des technologies différentes ont été développées en parallèle ces dernières années : (1) la plateforme 454 (*454 Life Sciences*) repose sur la technique du pyroséquençage dont le principe est de détecter la libération de pyrophosphate lors de l'incorporation de nucléotides au cours de la synthèse de novo d'un brin d'ADN à partir d'une matrice. En mars 2007, *454 Life Sciences* fut rachetée par Roche Diagnostics. (2) La plateforme développée par la société Solexa repose sur une stratégie plus proche de la méthode de Sanger. La molécule d'ADN à séquencer est mise en présence des quatre types de nucléotides, chacun associé à un fluorophore différent *(émettant un signal de couleur différente permettant de reconnaitre chaque nucléotide).* Le nucléotide nécessaire à l'élongation du brin complémentaire est alors incorporé mais une modification de sa structure empêche l'élongation de se poursuivre. A chaque cycle d'incorporation un signal lumineux spécifique du nucléotide présent est émis et enregistré. Fin 2006, la compagnie Illumina a racheté l'entreprise Solexa. (3) La troisième est la plateforme SOLiD (Sequencing by Oligo Ligation Detection) developpée par Agencourt Bioscience Corporation. Son principe est basé sur une réaction d'hybridation et de ligation chimique d'oligonucléotides complémentaires à la matrice d'ADN au cours de cycles successifs. En 2006 Agencourt Bioscience Corporation a été racheté par Applied Biosystems.

Les premiers séquenceurs commercialisés par ces trois compagnies étaient destinés à acquérir parallèlement la séquence de millions de fragments d'ADN et différentes versions successives de matériel ont été proposées (Metzker, 2010). L'objectif de cette synthèse n'est pas de faire une


	Applied Biosystem	ROCHE			ILLUMINA			Life Technologies		
	3730xl	GS Junior	454 FLX	454 FLX XL+	MiSeq	HiSeq 2000	GA IIx	Ion Torrent PGM	Ion Proton	SOLiD 5500Xl
Méthode de séquençage	Sanger	Synthèse (Pyroséquençage)			Synthèse			Synthèse (Détection de H+)		Ligation
Amplification de la matrice	PCR ou bactérien	PCR en émulsion			Bridge-PCR			NON	NON	PCR en émulsion
Taille moyenne des lectures (bases)	650	400	400	1 000	2 x 150	2 x 100	2 x 150	100	100	2 x 60
Exactitude de séquençage (%)	99,99%	99%	99%	99%	99,9%	99,9%	99,9%	99%	98%	99,999%
Capacité de séquençage / Run (Mb)	0,06	50	500	900	1 500	600 000	95 000	1 000	100	150 000
"Equivalent Génome de mammifère" / Run (X)	0,00000000002 X	0,016 X	0,16 X	0,3 X	0,5 X	200 X	32 X	0,33 X	0,03 X	50 X
Coût ($) / Run	96	1 100	6 200	6 200	750	20 000	11 500	1 000	1 000	10 500
Durée / Run	2 h	10 hr	10 hr	20 hr	1 jour	10 jours	14 jours	90 min	2 hr	8 jours
Coût ($) / 1 X	$ +100 000 000	$68 750	$38 750	$20 670	$1 500	$100	$360	$3 030	$33 333	$210

Figure 3 - Bilan synthétique des caractéristiques des différents séquenceurs de seconde génération proposés par les sociétés ROCHE, Illumina et Life technologies. Les trois séquenceurs encadrés les plus à droite correspondent aux machines de grande capacité. A titre comparatif, le séquenceur encadré à gauche est le dernier modèle de séquenceur capillaire de première génération.

description et une liste exhaustive des différentes technologies disponibles sur le marché ; la Figure 3 est destinée à résumer et à comparer les critères clefs des trois technologies de séquençage les plus souvent utilisées.

Cinq critères majeurs sont généralement rapportés et comparés pour évaluer les performances de ces technologies:

(1) la taille des lectures (nombre de bases successives lues) ; les solutions proposées permettent de lire de 60 bases à chaque extrémité d'un fragment (2x60) à 400 bases environ. Dans l'idéal, ce paramètre doit être le plus grand possible car disposer de lectures de grande taille facilite l'assemblage des séquences obtenues. Plus la taille est petite plus le taux de lectures écartées est important et plus les coûts associés aux travaux de bio-informatique sont élevés. L'objectif à court terme serait de disposer d'une technologie permettant de séquencer en continu des fragments de plusieurs milliers de base. Actuellement les lectures les plus longues sont obtenues à l'aide du dernier modèle 454 (FLX XL+).

(2) L'exactitude de séquençage ; l'optimum à atteindre serait une qualité de 100%. Les chiffres rapportés pour les différentes technologies varient de 99% à 99,999% et semblent donc comparables et très élevés. Il est cependant important de souligner que le nombre de bases lues lors d'une réaction de séquence étant très important, le nombre de bases erronées l'est aussi (1 % d'erreur sur 900 Mb lues induit 9 000 000 de bases erronées). De nouveau un lourd travail d'analyse et de "nettoyage" des données est donc nécessaire avant de pouvoir valider la séquence obtenue.

(3) La capacité d'une réaction (run) de séquence exprimée en bases ou en *équivalent génome* (chez les mammifères 1X = 3 000 Mb). L'idéal est de pouvoir obtenir en une réaction de nombreux équivalents-génomes. Disposer de 1X de séquence ne signifie pas que le génome a été séquencé une fois de manière exhaustive, mais que le nombre de bases lues au hasard est de 3.10^9. Le nombre de lectures obtenu par position du génome sera distribué selon une loi de Poisson ; certaines régions du génome seront donc représentées plusieurs fois, d'autres pas du tout. Actuellement la capacité la plus importante est proposée par le séquenceur HiSeq 2000 d'Illumina (200X par réaction).

(4) Le coût, que l'on souhaite le plus bas possible et ...

(5) La durée de réaction que l'on souhaite également la plus courte possible ; actuellement les réactions des séquenceurs de grande capacité durent de 1 à 2 semaines. Jusqu'à présent, bien que performantes, aucune des différentes technologies proposées ne permet de répondre parfaitement à l'ensemble de ces différents critères, et le développement de nouvelles solutions est en cours.

D'ores et déjà chacune des trois compagnies phares propose un modèle de "petit" séquenceur permettant de séquencer quelques dizaines à centaines de megabases en quelques heures. Néanmoins ces nouvelles solutions n'offrent pas d'améliorations notables sur les autres critères. Les solutions les plus prometteuses en cours de développement sont issues de technologies novatrices combinant les dernières avancées dans la nanofabrication, la chimie de surface et l'optique. L'ère du séquençage de troisième génération est ouverte. Le principe de cette troisième génération peut être résumé comme le séquençage d'une molécule unique (ADN ou ARN sans l'intermédiaire d'un ADNc) de grande taille, sans étape de pré-amplification, rapidement et à faible coût. Le séquençage direct de molécules (SMS : Single Molecule Sequencing) est en cours de développement selon trois grandes technologies : (1) le séquençage en temps réel impliquant la synthèse du brin

d'ADN complémentaire via une ADN polymérase (SMRT : Single Molecule Real Time) (Pacific Biosciences...) ; (2) le séquençage par détection des bases successives d'une molécule d'ADN au travers de nanopores (Oxford Nanopores, NABSys, NobleGen...) et (3) le séquençage basé sur des techniques de microscopie (Life technology, crackerBio...). Pour l'heure aucune de ces technologies ne remplit parfaitement les exigences de la technologie idéale de séquençage ; il est néanmoins certain qu'à une échéance plus ou moins lointaine une ou plusieurs solutions simples permettront d'obtenir pour quelques centaines de dollars la séquence à façon d'un individu (Figure 4).

La présentation des séquenceurs GridION et MinION (mini système de séquençage sur clé USB) faite par Oxford Nanopores en février 2012 en est l'illustration.

A **B**

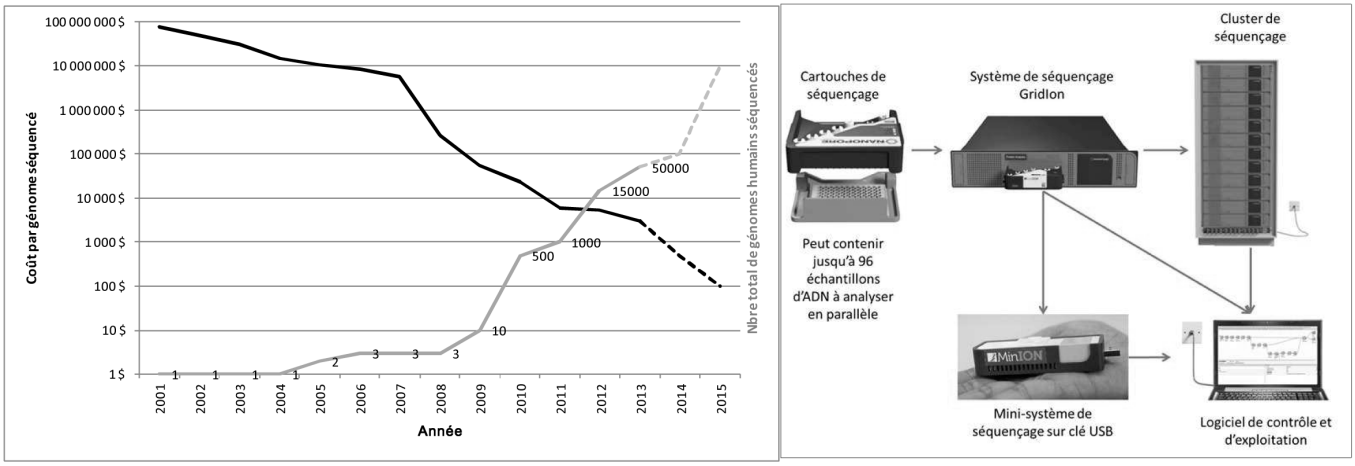

Figure 4 - A : Evolution des coûts pour obtenir la séquence d'un génome entre 2001 et 2013 (les valeurs des deux années suivantes en pointillés sont prévisionnelles) (trait noir) ; évolution dans le même temps du nombre de génomes d'humains différents séquencés (trait gris). **B** : Illustration des solutions MinION et GridION proposés par la société Oxford Nanopores.

2.2. Les domaines d'application

La possibilité d'acquérir de manière exhaustive à l'échelle du génome des données de séquences moléculaires permet de d'apporter des réponses à une large gamme de questions biologiques (Figure 5).

L'application majeure des technologies NGS est le séquençage de génomes entiers. La diminution des coûts et du temps nécessaire à l'acquisition des données a permis d'envisager l'utilisation de cette approche pour le reséquençage de novo de génomes. Chez l'homme, la comparaison de la séquence obtenue d'un individu unique (J. Craig Venter) à la séquence de référence publiée par le consortium international de séquençage de génome humain avait permis de révéler que ces séquences différaient pour 3,2 millions de SNP et 900 000 variants de structure. Ce taux d'environ 1 ‰ de SNP a par la suite été confirmé grâce aux séquences obtenues pour d'autres individus (Venter *et al.*, 2001). Bien que le nombre de différences augmente lorsque deux animaux de races distantes sont analysés, des taux moyens équivalents ont été obtenus chez les espèces domestiques. L'existence de nombreux variants et remaniements de structure a impulsé les approches de caractérisation systématiques des génomes par reséquençage, soit d'individus (programmes 1000 génomes : http://www.1000genomes.org/), soit de populations cellulaires (the cancer genome atlas : http://cancergenome.nih.gov/) afin de rechercher de façon exhaustive l'ensemble des mutations potentielles responsables de maladies. A terme une génomique personnalisée devrait voir le jour basée sur des analyses de corrélation entre la séquence d'un individu et ses données médicales (Ng *et al.*, 2009). Cette approche de médecine humaine n'est pas sans rappeler les approches de prédiction génomique développées chez les animaux domestiques. Le rapport qualité-prix des séquences obtenues actuellement reste néanmoins un facteur limitant et une des voies d'amélioration des technologies NGS est destinée à obtenir une couverture de 30 à 40X pour quelques centaines de dollars.

D'ores et déjà chez le porc, le reséquençage complet du génome de quelques individus a été réalisé dans le cadre du projet ANR SwAn (reséquençage de deux cas et de deux contrôles dans le cadre de la recherche de mutations prédisposant aux anomalies congénitales).

Pour l'heure, des stratégies alternatives permettent de réduire les coûts en limitant la région reséquencée à une fraction du génome. Ces régions peuvent être soit ciblées aléatoirement à l'issue d'une digestion du génome et du séquençage d'une fraction de quelques pourcents de l'ADN digéré, soit ciblées à l'aide de stratégies de "capture" préalable de régions particulières. Alors que l'enrichissement pour une région d'intérêt nécessite le développement à façon d'un outil de capture, des puces génériques (et de plus faible coût) ont également été développées pour permettre le reséquençage spécifiquement des régions codantes du génome (séquençage d'exome) (Bamshad *et al.*, 2011).

Chez le porc, une capture et le reséquençage de la région du MHC (Major Histocompatibility Complex) du porc à l'aide d'une puce dédiée ont été réalisés afin de répertorier l'ensemble de la variabilité existante chez des individus de races différentes (Projet ANR CapSeqAn).

En dehors du séquençage de l'ADN du génome d'un individu, les technologies NGS ont grandement facilité la caractérisation de modifications épigénétiques *(modifications chimiques du génome autres que des modifications nucléotidiques)* comme les méthylations, les variations de structure de la chromatine et les modifications post-traductionnelles des histones *(principaux constituants protéiques des chromosomes, étroitement liés à l'ADN)* qui interviennent dans les processus cellulaires de régulation du génome.

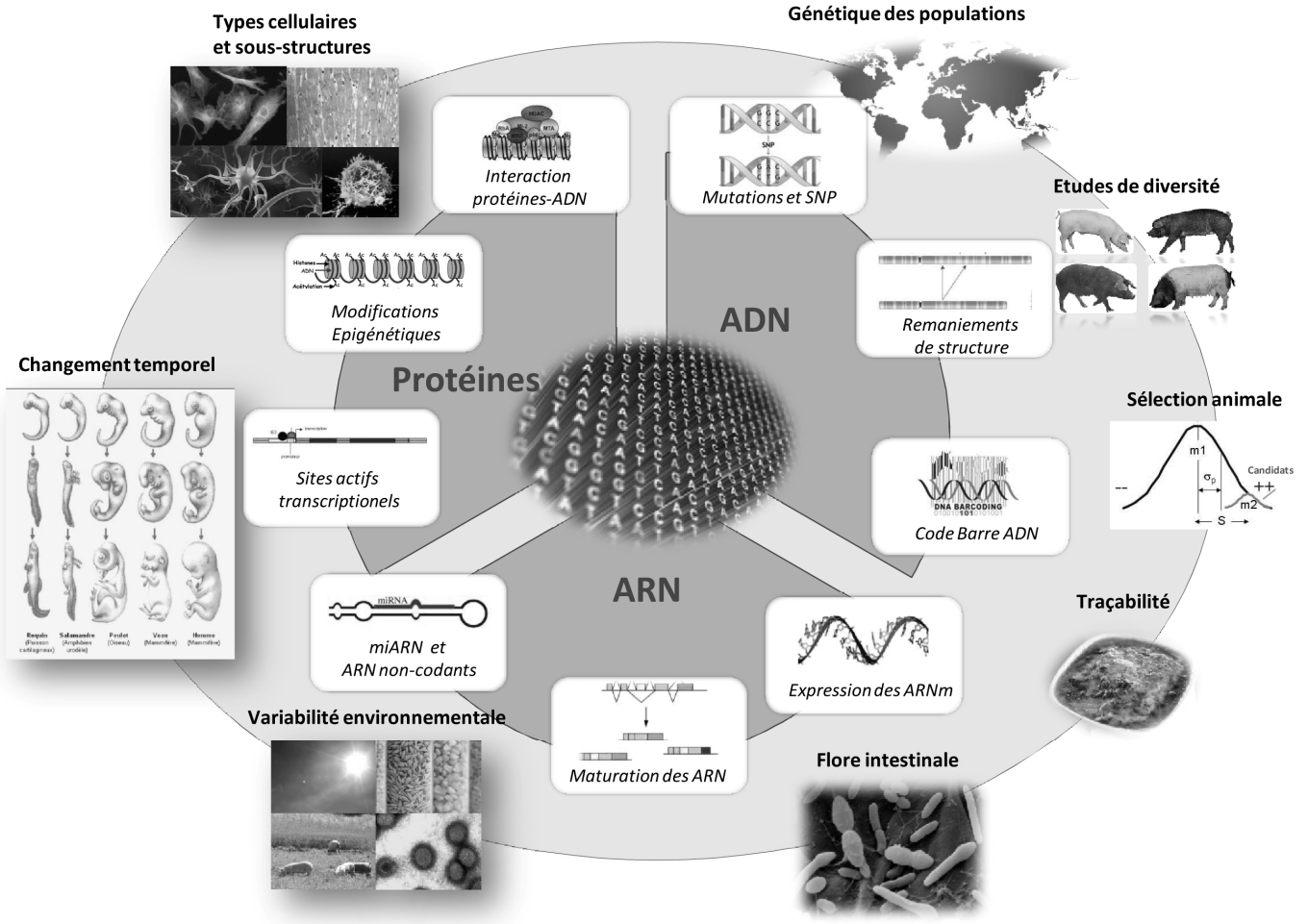

Figure 5 - Illustration des différentes études réalisables grâce aux nouvelles technologies de séquences et domaines dans lesquels ces études peuvent être réalisées.

Dans le domaine de l'analyse fonctionnelle de la régulation des gènes, certaines stratégies permettent de caractériser les interactions ADN-protéines et les sites de fixation de facteurs de transcription, de manière exhaustive à l'échelle du génome. La majorité de ces approches est basée sur un même principe : un anticorps spécifique de l'élément analysé (protéine du type facteur de transcription, ChIP-Seq, ou 5-methylcytosine MeDIP-Seq) est utilisé afin de capturer les portions de l'ADN génomique portant ces éléments. La fraction capturée est séquencée et les lectures obtenues sont alignées sur le génome de référence de l'espèce étudiée. Les régions du génome surreprésentées dans les lectures correspondront respectivement aux régions méthylées du génome ou aux régions comprenant un site de fixation de la protéine d'intérêt. Sur la base de cette technologie une carte de méthylome et de de l'hydroxyméthylome spermatique du verrat est en cours de réalisation à l'INRA ; des travaux destinés à comparer le méthylome du foie d'animaux de races différentes ont également été publiés (Bang *et al.*, 2013).

La très grande capacité des technologies NGS a également rendu possible l'accès à des données jusqu'alors inexplorées et le développement d'un nouveau domaine de recherche : la métagénomique. Celle-ci correspond à la caractérisation simultanée des génomes d'une population complexe de bactéries (microbiote) d'un milieu donné (intestin, océan, sols, air...). Les nouvelles techniques de séquençage permettent de séquencer l'ADN des microbiotes, donnant ainsi accès à la fraction bactérienne non cultivable (plus de 80% des bactéries intestinales anaérobie sont non cultivables). Les différences nucléotidiques permettent de caractériser les différentes souches, de différencier les espèces apparentées, et la représentation de chaque lecture permet une estimation quantitative de l'abondance de chaque espèce bactérienne. Chez l'homme et les espèces domestiques le domaine majeur d'application de la métagénomique est la caractérisation de la flore intestinale. Les premiers résultats obtenus dans ce domaine tendent à prouver que la composition de la flore d'un individu peut influencer ses caractéristiques phénotypiques et contribuer au déterminisme de certaines maladies (Pedersen *et al.*, 2013).

Le dernier et vaste domaine d'application des technologies NGS est l'analyse du transcriptome destinée à caractériser et

quantifier les populations d'ARN d'un échantillon (Farajzadeh *et al.*, 2013). L'avantage majeur de cette approche par rapport aux technologies d'hybridation sur puce disponibles jusqu'à présent est l'absence d'a priori. Les analyses réalisées à l'aide de puce permettent uniquement de révéler (présence/absence d'hybridation) et de quantifier l'expression des gènes représentés sur la puce. Par séquençage, tous les transcrits (même les plus faiblement exprimés) sont accessibles. La résolution de l'information est également améliorée : de faibles différences de séquences entre gènes d'une même famille ou entre allèles d'un même gène peuvent être prises en compte dans l'interprétation des résultats. Il est donc désormais possible de caractériser l'expression de différentes isoformes, l'expression différentielle entre allèles d'un même gène, les différentes formes de messagers issus d'un même gène (splicing alternatif) ainsi que les différentes populations d'ARN codant et non-codant (tRNA, rRNA, miRNA, siRNA) produits par un génome. L'analyse du transcriptome de différents tissus (ovaire, tissu adipeux sous–cutané...) a été réalisée dans le cadre de différents projets de recherche chez le porc, afin de parvenir à une caractérisation la plus exhaustive possible de la population de messagers présents à un stade donné, une condition et /ou un organe.

3. APPORTS DES NGS A LA GENETIQUE ET A LA SELECTION

Comme nous l'avons vu précédemment, les nouvelles technologies de séquençage permettent d'obtenir des informations génomiques quasi-exhaustives sur les échantillons analysés. Nous allons décrire maintenant l'importance de cette information pour la recherche en génétique, en étudiant trois domaines d'application distincts : l'étude de l'évolution des espèces et des populations qui les constituent, la caractérisation de mutations causales intervenant dans le déterminisme des caractères et enfin la sélection.

3.1. Génomique des populations

La connaissance exhaustive des variants génétiques existants dans une espèce est un grand pas en avant pour la compréhension de son histoire évolutive.

Avant la disponibilité d'information de séquence complète, des marqueurs génétiques peu nombreux et à taux de mutation relativement élevé (marqueurs microsatellites) étaient classiquement utilisés. Bien qu'ils soient assez informatifs, les caractéristiques de ces marqueurs limitaient les capacités d'inférence sur l'histoire des populations pour deux principales raisons. Premièrement, du fait de leur taux de mutation élevé, ils ne permettent pas de remonter très loin dans le passé. Deuxièmement, le nombre de marqueurs analysés reste faible, en particulier pour des raisons de coût de développement et de génotypage, ce qui ne permet d'étudier que des phénomènes affectant le génome dans son ensemble, sans pouvoir distinguer des évolutions particulières des gènes (e.g. des gènes d'adaptation).

Une importante évolution a été la disponibilité de puces de génotypage à haute densité (de l'ordre de 60 000 SNP chez le porc). Cet outil permet de pallier les deux problèmes précédents dans une certaine mesure, mais en ajoute un nouveau. En effet, les SNP présents sur les puces ne sont pas des SNP choisis aléatoirement sur le génome, il s'agit plutôt d'un sous-ensemble de SNP polymorphes dans de nombreuses populations. Ce choix est fait pour s'assurer que la puce soit informative et utilisable dans le plus de populations possibles au niveau mondial, en particulier dans des programmes de sélection. Il s'agit en revanche d'un sous-ensemble de SNP assez mal adapté aux études en génomique des populations, car le biais de recrutement ainsi induit est à même de biaiser les conclusions des analyses. Ainsi, une puce SNP spécifiquement dédiée aux études de génomique des populations humaines a été mise au point pour corriger les problèmes induits par les puces génériques, plus adaptées aux études médicales de populations d'origine européenne.

La possibilité d'obtenir des informations exhaustives de séquence individuelle ou de mélanges d'individus d'une même population permet d'éviter ce problème de biais de recrutement et constitue en quelque sorte le niveau ultime d'information utilisable pour les études de génomique des populations. La disponibilité croissante de séquences de génomes complets, du fait de la baisse du coût de séquençage, a par ailleurs stimulé les développements méthodologiques pour mettre au point des modèles d'analyse permettant de les exploiter. On peut citer en particulier le modèle PSMC (Li et Durbin, 2011) qui permet d'estimer les tailles d'une population dans le passé à partir de la séquence complète d'un seul individu de cette population. Cette approche est aujourd'hui classiquement utilisée et l'a été en particulier chez le porc (Groenen *et al.* 2012). A partir de séquences complètes de quelques sangliers asiatiques et européens cette analyse met en évidence une augmentation de la population européenne de sangliers après la colonisation et un impact plus grand de la dernière glaciation sur les populations européennes que sur les populations asiatiques (Figure 6).

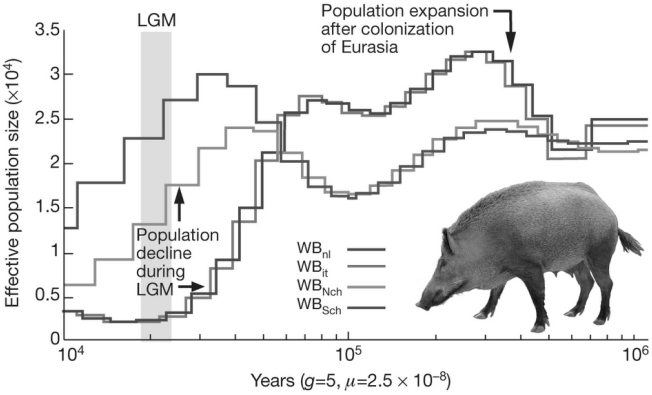

Figure 6 - Histoire démographique des sangliers, estimée à partir de séquences de génomes complets. *nl :* Pays-Bas, *it* : Italie, *Nch(Sch) :* Chine du Nord (du Sud). LGM : dernière glaciation. Tirée de Groenen *et al.* (2012)

Au-delà de la reconstitution d'une histoire démographique globale, affectant tout le génome des individus, la disponibilité de données de séquence permet également d'étudier l'histoire de *régions* génomiques.

Plus particulièrement, il s'agit de rechercher des régions du génome se comportant de façon significativement différente de la moyenne. Ce comportement peut s'expliquer par l'effet de la sélection et les régions concernées sont dites porteuses de *signatures de sélection*.

Les données de séquences complètes apportent la possibilité d'identifier les mutations sélectionnées elles-mêmes.

Par exemple, la Figure 7 présente une signature de sélection dans des populations bovines autour du gène MC1R, identifiée à partir de données de reséquençage complet (projet 1000 génomes bovins). Les deux mutations causales pour des phénotypes de couleur de robe sont retrouvées à la fois par des analyses haplotypiques (graphique du haut) et simple marqueur (graphique du milieu). Une des conséquences de la sélection sur ces mutations est une réduction de la diversité génétique (hétérozygotie) locale (graphique du bas).

3.2. Caractérisation de mutations impliquées dans le déterminisme génétique des caractères

Un autre domaine d'application qui profite de la disponibilité de séquences de génomes complets est la caractérisation du déterminisme génétique des caractères, par la facilitation de l'identification de mutations causales. Schématiquement, les mutations causales pour des caractères d'intérêt peuvent se décomposer en quatre types :

1. Des mutations fréquentes d'effet fort. Il s'agit de mutations associées à des gènes majeurs, comme par exemple chez le porc les mutations des gènes RN (Milan *et al.,* 2000) ou Halotane (Otsu *et al.*, 1991). Bien qu'elles soient plus faciles à caractériser, le nombre de mutations de ce type ségrégeant dans une population est assez faible.

2. Des mutations fréquentes ($P > 1$ %) d'effet modéré (QTL). Le nombre de mutations de ce type est typiquement assez élevé, et elles peuvent participer de manière substantielle au déterminisme génétique des caractères.

3. Des mutations rares ($P < 1$ %) d'effet fort. Le nombre de mutations de ce type est assez dépendant de l'histoire évolutive des populations. Dans les populations humaines d'origine européenne, on estime qu'elles sont assez nombreuses. Leur participation à la variance génétique totale est cependant probablement assez faible. Elles peuvent en revanche permettre d'identifier des gènes et des voies métaboliques importants pour le déterminisme des caractères.

4. Des mutations d'effet faible. Elles sont très difficile voire impossible à caractériser ; cependant les analyses de génétique quantitative démontrent que pour certains caractères, leur contribution à la variance génétique est très importante.

L'apport des nouvelles technologies de séquençage concerne l'identification de mutations des types 2 (QTLs) et 3 (mutations rares d'effet fort), en utilisant des stratégies différentes.

3.2.1. *Mutations causales de QTLs*

Pour la recherche de mutations sous-jacentes aux QTLs (*Quantitative Trait Nucleotide, QTN*), la stratégie la plus commune aujourd'hui consiste à utiliser des études d'association tout génome (GenomeWide Association Studies, GWAS), dont il a été montré qu'elles étaient potentiellement plus puissantes que des analyses familiales (Risch et Merinkangas, 1996). Brièvement, le principe consiste à génotyper de nombreux (typiquement plus de 1000) individus peu apparentés, mesurés pour le caractère d'intérêt pour un nombre de marqueurs très élevés (de 100 000 à un million de marqueurs), ces chiffres pouvant varier en fonction de l'espèce et du caractère étudié. Un test statistique de l'association génotype / phénotype est ensuite effectué marqueur par marqueur.

En génétique animale, une stratégie hybride combinant analyse familiale et GWAS est souvent employée.

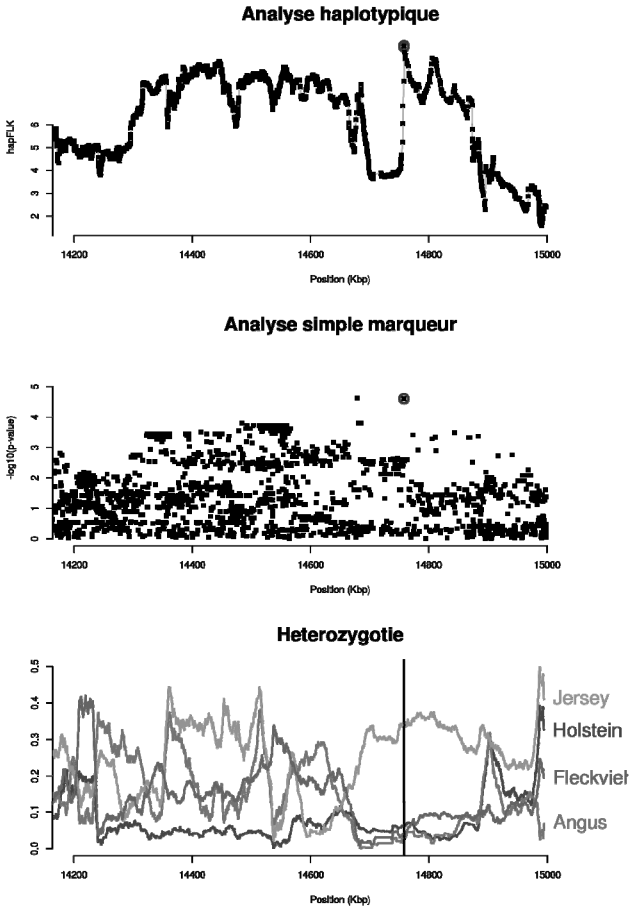

Figure 7 - Recherche de signatures de sélection le long du chromosome 18 bovin à l'aide des polymorphismes identifiés à partir des séquences obtenues dans le cadre du projet 1000 génomes bovins. Les deux mutations causales du gène MC1R connues affectant la couleur de la robe sont indiquées par un cercle. La position du gène MC1R est indiquée par la barre verticale du graphique du bas. Trois des quatre populations présentent des hétérozygoties réduites dans la zone, de manière cohérente avec leur phénotype.

La disponibilité de séquences génomiques complètes peut permettre de disposer des génotypes de tous les polymorphismes fréquents existant dans une population, et donc des QTNs dans une étude d'association. Cependant, reséquencer l'ensemble des individus d'une GWAS reste très couteux. Pour pallier ce problème de coût, le principe est d'utiliser des méthodes d'imputation de génotypes (Marchini *et al.,* 2007 ; Servin et Stephens, 2007). Le principe d'imputation a pour prérequis de disposer d'un petit échantillon d'individus (~100) reséquencés de manière assez peu profonde (~4X) et représentatifs de la population d'intérêt (*échantillon panel*), et d'un échantillon GWAS génotypé pour une puce SNP. Il est alors possible d'entraîner sur l'échantillon panel des modèles statistiques de la diversité haplotypique dans la population pour prédire les séquences des individus de l'échantillon GWAS et ensuite effectuer les tests d'association pour tous les polymorphismes connus. C'est une des motivations principales pour la mise en place de nombreux programmes de reséquençage de populations (projets 1000 génomes humain, mais également bovin, arabidopsis ...).

La Figure 8 (1000 Genomes Project Consortium, 2010) présente un exemple d'utilisation du projet 1000 génomes humains pour l'analyse d'association tout génome.

L'utilisation des données de séquence complète permet d'une part de trouver de nouvelles associations et d'autre part d'augmenter la précision de localisation d'associations connues.

3.2.2. *Mutations rares d'effet fort*

L'autre type de mutation potentiellement accessible à l'analyse génétique consiste en des mutations d'effet fort mais de fréquence faible ($P < 1$ %) dans les populations. Dans ce cas, la stratégie d'imputation précédente n'est pas adaptée car la découverte de mutations rares nécessite : (i) de séquencer de nombreux individus (du fait de leur faible fréquence) et (ii) de faire un séquençage profond pour avoir des génotypes fiables.

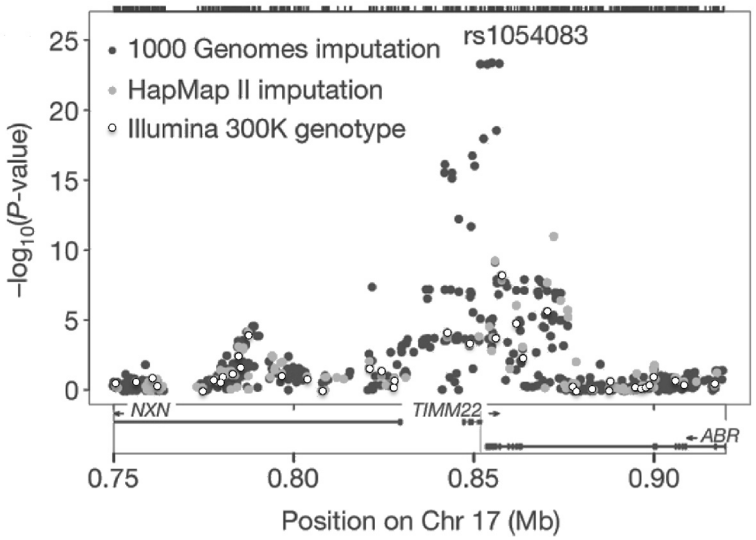

Figure 8 - Exemple d'analyse d'association par imputation de génotypes utilisant le projet 1000 génomes humains. Les SNP en blanc sont des SNP génotypés, les SNP en gris sont des SNP imputés à partir du panel HapMap2, les SNP en noir sont les SNP imputés à partir du projet 1000 génomes. La plus forte association a été obtenue avec le marqueur rs1054083 dans la région du gène TIMM22. Tirée de "1000 Genomes Project Consortium" *et al.* (2010).

Comme il n'est pas encore envisageable d'effectuer un séquençage profond de milliers d'individus sur l'ensemble du génome, une stratégie alternative consiste à reséquencer uniquement la partie codante du génome, beaucoup plus petite (environ 1 % du génome, soit 30Mb chez l'homme ou le porc) : le séquençage d'exome. Cette stratégie fait l'hypothèse que les mutations rares d'effet fort affectent la séquence codante ou interne des gènes, ce qui semble biologiquement réaliste. Elle a permis par exemple d'identifier le gène et les mutations responsables du syndrome de Miller à partir du séquençage de l'exome de quatre individus atteints non apparentés (Ng *et al.*, 2010). Là encore, un projet collaboratif de grande ampleur a été mis en place chez l'homme pour valoriser au mieux cette stratégie, conduisant au séquençage de plus de 6500 exomes (Fu *et al.*, 2013).

3.3. Séquence et sélection génomique

Les travaux sur l'intérêt des données de séquences dans des programmes de sélection génomique (SG) n'ont été développés que très récemment. Ils n'envisagent pas aujourd'hui l'utilisation directe des données de séquences dans l'évaluation génomique, mais tirent partie du fait que la connaissance des séquences permet d'accéder à la majeure partie des 20 à 30 millions de SNP présents sur le génome. Il devient ainsi possible de choisir les SNP les plus pertinents, i.e. très proches sur le génome et en très fort déséquilibre de liaison avec les mutations causales, permettant si possible d'identifier des haplotypes conservés entre populations. La connaissance des séquences permettra également, parallèlement à l'accumulation de données de génotypage, d'identifier plus efficacement les mutations causales (QTN). Ces informations nouvelles doivent permettre d'obtenir des équations de prédiction génomique plus robustes, moins sensibles aux risques de perte d'association entre marqueurs et QTN au cours des générations de sélection, et, potentiellement, de disposer de prédictions plus performantes dans le cas d'évaluations multi-populations. L'amélioration de la qualité de la prédiction doit ensuite permettre d'appréhender et de prendre en compte de façon plus efficace la complexité de l'architecture génétique des caractères (interactions entre gènes, interactions génotype x milieu,…).

CONCLUSION

La disponibilité d'une séquence de référence chez le porc offre une base fondamentale pour l'ensemble des recherches en génétique chez cette espèce. Les nouvelles technologies de séquençage peuvent permettre d'accéder pour un coût réduit à la diversité des génomes, mais également des transcriptomes, méthylomes… au sein et entre les populations porcines.

L'intérêt de ces technologies pour notre compréhension du déterminisme génétique des caractères, de leur physiologie mais également de l'histoire évolutive de l'espèce et de ses relations avec les espèces apparentées a d'ores et déjà été démontré.

Chez d'autres espèces, y compris d'animaux d'élevage, des programmes à grande échelle de caractérisation par séquençage de populations sont mis en place, en particulier dans le cadre de larges collaborations de consortiums internationaux. Ce type de programme n'existe pas encore chez le porc, mais pourrait voir le jour si une communauté suffisamment importante se constituait.

REFERENCES BIBLIOGRAPHIQUES

- 1000 Genomes Project Consortium, Abecasis G.R., Altshuler D., Auton A., Brooks L.D., Durbin R.M., Gibbs R.A., Hurles M.E., McVean G.A, 2010. A map of human genome variation from population-scale sequencing. Nature, ,467, 1061-1073
- Bamshad M.J., Ng S.B., Bigham A.W., Tabor H.K., Emond M.J., Nickerson D.A., Shendure J., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet., 12, 745-55.
- Bang W.Y., Kim S.W., Kwon S.G., Hwang J.H., Kim T.W., Ko M.S., Cho I.C., Joo Y.K., Cho K.K., Jeong J.Y., Kim C.W., 2013. Swine liver methylomes of Berkshire, Duroc and Landrace breeds by MeDIPS. Anim. Genet., 44, 463-466.
- Farajzadeh L., Hornshøj H., Momeni J., Thomsen B., Larsen K., Hedegaard J., Bendixen C., Madsen L.B., 2013. Pairwise comparisons of ten porcine tissues identify differential transcriptional regulation at the gene, isoform, promoter and transcription start site level. Biochem. Biophys. Res. Commun., 438. 346-352.
- Fu W., O'Connor T.D., Jun G., Kang H.M., Abecasis G., Leal S.M., Gabriel S., Rieder M., Altshuler D., Shendure J., Nickerson D., Bamshad M., NHLBI Exome Sequencing Project, Akey J., 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature, 493, 216-220.
- Groenen M.A.M. (Swine Genome Sequencing Consortium), 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature, 491, 393–398
- Humphray S.J., Scott C.E., Clark R., Marron B., Bender C., Camm N., Davis J., Jenks A., Noon A., Patel M., Sehra H., Yang F., Rogatcheva M.B., Milan D., Chardon P., Rohrer G., Nonneman D., de Jong P., Meyers S.N., Archibald A., Beever J.E., Schook L.B., Rogers J., 2007. A high utility integrated map of the pig genome. Genome Biol., 8, R139.
- Li H., Durbin R., 2011, Inference of human population history from individual whole-genome sequences. Nature, 475, 493-496.
- Marchini J., Howie B., Myers S., McVean G., Donnelly P., 2007, A new multipoint method for genome-wide association studies via imputation of genotypes. Nat. Genet., 39, 906-913.
- Metzker M.L., 2010. Sequencing technologies - the next generation. Nat. Rev. Genet., 11, 31-46.
- Milan D., Jeon J.T., Looft C., Amarger V., Robic A., Thelander M., Rogel-Gaillard C., Paul S., Iannuccelli N., Rask L., Ronne H., Lundström K., Reinsch N., Gellin J., Kalm E., Roy P.L., Chardon P., Andersson L., 2000. A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. Science, , 288, 1248-1251.
- Ng P.C., Murray S.S., Levy S., Venter J.C., 2009. An agenda for personalized medicine. Nature, 461, 724-726.
- Ng S.B., Buckingham K.J., Lee C., Bigham A., Tabor H., Dent K., Huff C., Shannon P., Jabs E., Nickerson D., Shendure J., Bamshad M., 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet., 42, 30-35.
- Otsu K., Khanna V.K., Archibald A.L., MacLennan D.H., 1991. Cosegregation of porcine malignant hyperthermia and a probable causal mutation in the skeletal muscle ryanodine receptor gene in backcross families. Genomics,11, 744-750.
- Pedersen R., Ingerslev H.C., Sturek M., Alloosh M., Cirera S., Christoffersen B.Ø., Moesgaard S.G., Larsen N., Boye M., 2013. Characterisation of gut microbiota in Ossabaw and Göttingen minipigs as models of obesity and metabolic syndrome. PLoS One, 8, e56612.
- Ramos A.M., Crooijmans R.P., Affara N.A., Amaral A.J., Archibald A.L., Beever J.E., Bendixen C., Churcher C., Clark R., Dehais P., Hansen M.S., Hedegaard J., Hu Z.L., Kerstens H.H., Law A.S., Megens H.J., Milan D., Nonneman D.J., Rohrer G.A., Rothschild M.F., Smith T.P., Schnabel R.D., Van Tassell C.P., Taylor J.F., Wiedmann R.T., Schook L.B., Groenen M.A., 2009. Design of a high density SNP genotyping assay in the pig using SNP identified and characterized by next generation sequencing technology. PLoS One, 4, e6524.
- Risch N., Merikangas L., 1996. The future of genetic studies of complex human diseases. Science, 273, 1516-1517.
- Servin B., Stephens M., 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PloS Genetics 3, e114.
- Schook L.B., Beever J.E., Rogers J., Humphray S., Archibald A., Chardon P., Milan D., Rohrer G., Eversole K., 2005. Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. Comp. Funct. Genomics, 6, 251-255.
- Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., et al (274 auteurs), 2001. The sequence of the human genome. Science, 291, 1304-1351.

2014. Journées Recherche Porcine, 46.