

Compte rendu de la journée de veille consacrée aux "Nouvelles techniques de séquençage haut débit et impact sur la génomique animale"

Directoire opérationnel d'AGENAE
26 Mai 2008

Au compte rendu de la journée du 26 Mai organisée pour le directoire opérationnel d'AGENAE, est joint en annexe une note de Denis Milan (Génétique Animale Toulouse) rédigée suite à la veille technologique réalisée dans le cadre de la Génopole de Toulouse, aux discussions ayant eu lieu lors de cette journée ou d'autres réflexions dans le cadre d'IBISA.

Objectif de la journée

Cette journée a été organisée dans le cadre des journées consacrées à la veille technologique au sein du directoire opérationnel d'AGENAE. Devant les enjeux des techniques de séquençage Haut débit qui vont modifier en profondeur la manière de travailler en génomique animale, la journée avait pour but de faire le point sur la question avec des représentants des plateformes nationales du CNS et du CNG (P. Wincker & I. Gut), ainsi qu'avec un certain nombre d'utilisateurs ayant déjà une expérience dans le domaine de l'étude de la diversité (P. Taberlet), du métagénome (D. Ehrlich) ou de l'étude du transcriptome et de la régulation de l'expression (A. Jacquier, M. Crespi & M. Werner). Outre des scientifiques INRA (membres du DO Agenae, représentants des départements APA), la journée était ouverte aux partenaires professionnels d'AGENAE, membres du DO Agenae.

Programme des interventions

- 10h Denis Milan (INRA Toulouse)
Introduction sur le séquençage Haut Débit
- 10h30 Patrick Wincker (CNS Evry)
Apport des Nouvelles Technologies au séquençage de génomes au Genoscope
- 11h15 Ivo Gut (CNG Evry)
Le Séquençage Haut débit au CNG
- 12H Dusko Ehrlich (INRA Jouy)
Etude de métagénome : le Projet MetaHit

- 13h45 Pierre Taberlet (UJF, Grenoble)
Les nouvelles avancées techniques du séquençage de l'ADN, et les implications en sciences de l'environnement
- 14h30 Alain Jacquier & Helen Neil (Institut Pasteur)
Adaptation de la technique du "3'-longSAGE" au séquençage à haut débit pour l'analyse génomique des transcrits cryptiques (CUT) chez la levure
- 15h15 Martin Crespi (CNRS Gif sur Yvette)
La diversité des petits ARNs non-codants des plantes
- 16h Michel Werner (CEA Saclay)
Apports des approches de ChIP-chip et ChIP-Seq dans l'étude de la transcription chez la levure et la souris

Participants

Outre les orateurs, ont participé à cette réunion : P. Monget (PHASE Tours), X. Vignon (PHASE Jouy), Y. Guiguen (PHASE Rennes), A. Eggen (GA Jouy), C. Rogel-Gaillard (GA Jouy), P. Martin (GA Jouy), V. Ducrocq (GA), C. Donnadiou (GA Génopole Toulouse), M. Douaire (GA/APA Rennes), B. Coudurier (APA Tours), P. Humblot (UNCEIA), L. Journaux (UNCEIA).

Résumé des interventions à partir des notes prises par Xavier Vignon (Phase, Jouy)

1) D. Milan (INRA Toulouse, Génopole) Introduction sur le séquençage haut débit

Les nouvelles technologies de séquençage permettent une baisse drastique des coûts des investigations dans ce domaine. Par exemple le génome humain dont le séquençage a coûté environ 3 milliards d'euros (pour une taille d'environ 3 Gb) serait actuellement séquençé pour 30 Millions d'euros avec des techniques classiques; une opération de reséquençage de ce même génome peut s'envisager pour 50 000 euros. L'objectif des développements technologiques est d'arriver au séquençage du génome humain pour 1000 \$.

Trois technologies sont actuellement disponibles :

- Technologie Roche 454 (5 machines en France), séquençage de fragments de 400b fixés sur billes, permet séquençage de 150Mb à 0,5Gb en un run.
- Technologie Solexa/Illumina (7 machines en France), séquençage de fragments de 35b, fixés sur plaque, permet le séquençage de 1Gb à 3Gb.
- Technologie SOLiD, en gros équivalente à Solexa.
-

Ces équipements sont une véritable « révolution » puisqu'ils permettent d'envisager un séquençage de 0,5 à 3 Gb en un run. Le prix des machines actuelles est de 500 000 à 600 000 € ; mais c'est surtout le coût des réactifs, de main d'œuvre et d'environnement informatiques qui est primordial.

2) Patrick Wincker (CNS-Genoscope Evry). Apport des Nouvelles Technologies au séquençage de génomes au Genoscope

Optimisation des performances sur les machines du Genoscope :

- Sur Roche 454, à partir de septembre 2008 1 250 000 lectures de 400 bases par run (500Mb) au lieu de 400 000 lectures de 250 bases actuellement.
- Taux d'erreurs (0.8% sur Roche, 0,4% sur Solexa) seront diminués en combinant les lectures Roche et Solexa.

Des combinaisons de techniques différentes par exemple Séquençage Roche 15 X + 50 X avec Illumina permettent d'obtenir à cout limité une séquence très convenable (sans toutefois accéder aux séquences répétées). Dans le cas d'Acinobacter (génome de $3.5 \cdot 10^6$) cette stratégie a laissé seulement 400 erreurs dont beaucoup correspondent à des délétions/insertions d'homopolymères.

Pour du séquençage de BAC, la stratégie pour l'instant mise en œuvre correspond au multiplexage de 6 BAC. Il y a toutefois des biais dans la représentation des BAC. Par exemple, sur une manip de référence le BAC le plus présent regroupe 23 % des séquences, le moins représenté est à 6.7 %, au lieu des 16.6 % attendus sur chaque BAC.

Sur Solexa comme sur Roche, les stratégies Paired ends, permettant de disposer de paires de séquences dont on connaît l'orientation respective et la distance, permet d'améliorer grandement la qualité de l'assemblage.

En 2009-2010, apparition de technologies de 3^{ème} génération : technologie nanopore (passage de fragments ADN dans des nanopores et séquençage en fonction de charges électriques des bases).

Réflexions pour l'avenir : penser plutôt en plateforme Nationale que Régionale ou locale pour le séquençage (nécessité de compétences pointues sur place). Les équipements du Genoscope ne sont utilisés qu'à 50% de leurs possibilités. Patrick recommande plutôt une spécialisation des équipes sur une technique ou une autre. Il est difficile de savoir tout faire (surtout pour l'analyse bioinfo).

3) Ivo Gut (CEA-CNG – Evry) **Le Séquençage Haut débit au CNG**

Présentation de programmes de génotypage SNP en complément du séquençage. Présentation de collaborations avec A. Eggen sur les SNP chez le bovin (Genome wide association). Evocation de projets similaires sur le porc avec D. Milan et la Brebis avec C. Moreno et coll.

Quatre Illumina couplés à 4 serveurs (8 processeurs, 72Gb de RAM et 90 TO de stockage) sont utilisés au CNG pour des études de reséquençage (Long Range- PCR), expression de gènes (digital-whole RNA), séquençage de fragments immuno-précipités (ChIP-Seq) et méthylation d'ADN (MeDIP-Seq).

Dans le cadre du séquençage d'une région d'intérêt, la stratégie pour l'instant privilégiée par le CEA est plus le séquençage de fragments Long Range PCR que le séquençage du résultat d'un enrichissement par hybridation sur puces génomiques Nimblegen. L'avantage de la stratégie Long Range étant que l'on est sûr de séquencer les fragments amplifiés, alors que les captures sur puce ne garantissent pas que l'on va retrouver l'ensemble des séquences de la région dans la fraction enrichie.

Promesses pour la technologie de 3^{ème} génération : séquençage d'un génome humain en 12h pour 1000\$!

4) D. Erlich (INRA- Jouy) **Etude de métagénome : le Projet MetaHit**

Présentation du projet de Metagénomique. MetaHIT qui a pour objectifs :

- séquençage des gènes et génomes de la flore microbienne de l'intestin humain,
- établissement de profils de gènes du microbiote intestinal,
- trouver des associations gènes microbiens-maladies (liens populations bactériennes – obésité),
- développer les approches bioinformatiques.

Il y a plus de gènes bactériens exprimés en nous que de gènes humains. 70 à 80 % de ces bactéries ne sont pas cultivables. Les nouvelles techniques de séquençage appliquées au métagénome de l'intestin humain vont nous permettre d'avoir accès à ce domaine encore inconnu.

Projet de 4 ans (depuis Jan 2008), 20 millions d'euros, implication de Danone et UCB Pharma.

Acquisition d'un analyseur SOLiD à Jouy pour fin Juin. Grande capacité de séquençage (6Gb en 1 run), mais analyses longues (1 semaine pour la prep d'échantillon, le dépôt sur lame et l'analyse (2 lames à 3Gb par lame). Projet à forte connotation biomédicale/santé.

A titre d'exemple pour montrer l'intérêt de ces approches : 1) On peut suivre l'évolution du métagénome lors de la perte de poids d'un obèse. Cette évolution est elle uniquement une conséquence, ou participe t'elle à la perte de poids. Quels sont les gènes impliqués ? Etude des débouchés thérapeutiques potentiels. 2) Le Microbiome d'une souris obèse implantée dans une souris normale induit une prise de poids plus importante que si l'on plante le microbiome d'une autre souris normale.

5) Pierre Taberlet (Univ J Fourier – Grenoble) **Les nouvelles avancées techniques du séquençage de l'ADN, et les implications en sciences de l'environnement**

Présentations d'applications nouvelles accessibles grâce aux progrès des technologies de séquençage. La sensibilité et la fiabilité des méthodes permettent d'accéder à de nouvelles données sur la Biodiversité et l'Ecologie:

- Des extraits de sols permettent d'identifier la présence de genres ou d'espèces sur certains territoires (essais sur le permafrost, on peut remonter jusqu'à 25 000 ans et potentiellement jusqu'à 500 000 ans),
- Identification de la présence d'espèces de grenouilles à partir de la détection de leur ADN mitochondrial dans l'eau d'une mare,
- Identification de régimes alimentaires à partir de fèces pris sur le terrain. Des exemples de succès chez des limaces, des oiseaux, des ours, des marmottes... Mais attention aux résultats artéfactuels : des ADN de tomates et de cannabis ont été trouvés dans les crottes d'un ongulé d'altitude dans le Jura, à cause de la proximité d'un refuge !!

Enfin, en génomique des populations, les techniques d'AFLP pourraient maintenant être remplacées par le séquençage (notamment de fractions du génome amplifiées par AFLP ou d'autres techniques de représentation réduite du génome) grâce à la possibilité d'avoir beaucoup de marqueurs sur beaucoup d'individus. Mais les techniques de microsatellite resteront sans doute utilisées dans les études de parenté.

6) Hélène Neil & Alain Jacquier (Inst. Pasteur – Paris) Adaptation de la technique du "3'-longSAGE" au séquençage à haut débit pour l'analyse génomique des transcrits cryptiques (CUT) chez la levure

L'exemple est donné par l'étude des « Cryptic Unstable Transcripts » (CUT) chez la levure. Les CUT sont des transcrits nucléaires qui sont régulés de manière particulière dans le noyau. A partir de mutants ne dégradant pas ces transcrits, la technique SAGE a été adaptée pour créer une banque de fragments « tagués » qui sont ensuite séquencés.

Ces CUT pourraient être aussi nombreux que les transcrits de gènes. Ils font en moyenne 300 bases et sont assez souvent positionnés juste en amont des ORF.

L'objectif est de cartographier ces unités de transcription qui pourraient présenter un nouveau mécanisme de régulation de l'expression du génome.

7) Martin Crespi (CNRS – Gif) Analyse de la diversité des petits ARN chez les plantes.

Des préparations de petits ARN ont été séquencés à l'aide de la technologie 454 afin d'identifier de nouveaux miRNA, spécifiques des plantes: transacting-siRNA, natural antisense-siRNA, repeat associated-siRNA. L'étude porte sur l'expression des petits ARN et leur participation dans la régulation du développement des racines (différenciation des cellules souches du méristème) et de l'étude de la réponse au stress abiotique sur des plantes modèles (*Medicago truncatula*). Le séquençage est réalisé sur 454 et l'analyse bioinformatique permet identifier des miRNAs (avec un structure tige-boucle) parmi ces populations de petits ARNs. La diversité des petits ARNs ainsi que leur fréquence dans la population est directement déduit de cette analyse. La méthodologie utilisée désormais est Solexa parce qu'elle permet des lectures très courtes mais à une très grande échelle.

8) Michel Werner (CEA – Saclay) Apports des approches de ChIP-chip et ChIP-Seq dans l'étude de la transcription chez la levure et la souris

Etude d'interactions protéines-ADN par ChIP-seq sur modèles levure et souris.

Sur ces modèles, le principe de ChIP on Chip est appliqué à l'étude de la formation de complexes de transcriptions spécifiques de RNA pol I, pol II ou pol III. Le développement de la technique ChIP-seq consiste ensuite à séquencer en très haut débit les séquences d'ADN immuno-précipitées. Chez la souris, on a recours à l'insertion de « tags » par recombinaison homologue (recombineering technique) dans des cellules ES (il existe maintenant des lignées contenant diverses protéines

taggées : RNA polymérase, variants d'histones, cf M. Werner). L'emplacement de machineries de transcription peut ainsi être cartographié dans les cellules dans diverses situations physiologiques, par exemple :

- transcription spécifique d'ARNs régulateurs par RNA polIII,
- dérégulation de RNA polIII dans certains cancers,
- recrutement de TFIIIC dans la propagation de l'hétérochromatine...

La grosse difficulté de ces techniques réside dans l'analyse bioinformatique. S'il existe des logiciels d'analyse de ChIP-seq, il n'y a rien en stockage, gestion et utilisation de bases de données.

9) Conclusions, discussions, recommandations à la fin de la journée:

La courte discussion qui a suivi a mis en avant les points suivants :

- L'importance de la bioinformatique est revenue à plusieurs reprises dans la journée. Même si de gros centres de traitement bioinformatique pourraient se mettre en place à l'avenir, il faut penser à développer la bioinfo en complément dans les labos du fait que chaque étude requiert des analyses spécifiques.
- Pour les animaux domestiques pour lesquels on commence à avoir beaucoup de QTL, la manière de préparer l'avenir est de préparer la bioinformatique plutôt que de prévoir l'achat de machines chères, à renouvellement rapide et gourmandes en compétences.
- P. Wincker et I. Gut nous conseillent si du séquençage doit se faire à l'INRA, de viser la constitution de plateformes plus spécialisées que le CNS ou le CNG. CNS et CNG sont en mesure d'accueillir des projets INRA : le CNS sélectionne sur appel d'offre, et les machines ne sont pas saturées ; le CNG est très sollicité, mais sélectionne continuellement des projets d'intérêts qui peuvent être courts si le protocole est bien établi.
- Le point de vue des professionnels de l'élevage est que le besoin est plutôt vers le génotypage (SAM 2ème génération), mais le re-séquençage sera certainement très utilisé sur des régions identifiées.

Note sur les enjeux du séquençage Haut débit à l'INRA

Denis Milan, Aout 2008

Des constats ...

Tout le monde perçoit l'impact très fort de technologies de séquençage haut débit sur la recherche. **La possibilité de produire à bas cout un très grand nombre de séquences donnera un avantage concurrentiel très fort à ceux qui maîtriseront ces techniques et les mettront en œuvre au bon moment.** Cela touchera tous les domaines, on peut penser notamment : 1) Dans le cadre de la recherche de QTL, à la possibilité de séquencer plusieurs haplotypes QTL caractérisés dans une région, pour identifier les mutations les différenciant, 2) Dans le cadre de projet de transcriptomique : à l'étude différentielle de petits ARN ou de transcrits peu exprimés, à l'étude de l'épissage alternatif, 3) dans le cadre de la régulation de l'expression, à l'identification de l'ensemble des séquences génomiques interagissant avec un facteur de transcription donné, à l'étude des régulations épigénétiques (méthylation différentielle ...) 4) dans le cadre de projets diversité ou de métagénomique à un accès massif à de très nombreuses séquences.

Au niveau technique, le domaine est en constante évolution, et les investissements réalisés peuvent très vite être dépassés. A l'heure actuelle, il existe deux types de familles de technologie : 1) la technologie Roche permettant d'accéder à des séquences de *grands fragments de l'ordre de 400 bases*, 2) la technologie de type Illumina / Applied biosystems donnant accès à un *nombre encore plus grand de petites séquences de l'ordre de 35 b*. D'autres technologies sont en émergence, notamment pour la réalisation du séquençage direct de molécules uniques. Si elles se stabilisent ces techniques pourraient avoir des débits très importants, et ne plus dépendre d'une étape d'amplification PCR qui introduit inévitablement des biais (certaines séquences étant plus difficile que d'autres à amplifier, notamment les séquences riches en %GC). A titre personnel, je pense que la concurrence s'exerce essentiellement sur le segment des petits fragments. Il ne semble pas y avoir vraiment de concurrence (en tout cas pour l'instant) autour de la production de fragments longs de l'ordre de 400 bases. *Les machines Roche pourraient donc rester pertinentes plus longtemps que les autres¹.*

Tous les orateurs de la journée sans exception de réflexion menée par le DO d'AGENAE, ont mis en exergue **l'importance cruciale de l'environnement informatique et de l'analyse bioinformatique** pour pouvoir tirer partie des séquences produites en très grand nombre. Dans ce domaine, il faut savoir gérer de très grandes quantités de données (de l'ordre de 1 To par run sur Illumina), puis savoir traiter (avant tout savoir assembler) les séquences.

Faut-il investir dans ce domaine à l'INRA ?

Les gros projets prioritaires ou des projets pilotes d'intérêt particulier pourront toujours passer au CNS ou au CNG. A coté de ces gros projets, **devant la multiplicité des projets qui feront appel à ces techniques, les recherches menées à l'INRA nécessitent à mon avis des investissements propres de l'INRA²**. L'enjeu majeur à l'INRA est à mon avis davantage sur l'accompagnement des programmes de taille intermédiaire, qui ponctuellement auront besoin de ces techniques haut débit, plutôt que sur les gros projets prioritaires de type Micalis qui auront toujours un accès naturel aux capacités du CNS.

On pourrait dire qu'à coté des services rendus par les centres nationaux CNS / CNG, ces services aux programmes de taille intermédiaire pourraient être externalisés sur des entreprises commerciales. A mon avis, les équipes de recherche ont besoin d'une mutualisation de l'expérience, de formation, d'un

¹ Par ailleurs, à coté des techniques à haut-débit, il est important de garder en tête que l'on aura toujours des besoins de séquençage classique de fragments individuels sur des machines de type 3730 d'Applied Biosystems.

² A titre d'exemple, même si les gros programmes de séquençage classique ou de génotypage SNP massifs initiaux sont réalisés au CNS et au CNG, il a été vital pour les équipes INRA de disposer également de capacités propres pour pouvoir mener à bien leurs programmes sans dépendre totalement d'arbitrage de priorités réalisés en dehors de l'INRA. Il n'y a pas de raison de penser différemment pour le séquençage HD.

accompagnement et d'un conseil pour l'élaboration des projets de recherche, ce qui dépasse largement le service qui peut être apporté par des entreprises commerciales. Tout cela implique donc la nécessité d'investissements en interne à l'INRA.

Coté séquenceurs HD : **Il ne faut pas se précipiter et investir trop massivement car la technologie actuellement proposée sera vite obsolète**, mais au vu des enjeux il faut mettre en place une politique d'accès réel des équipes de recherche à ces technologies. Les équipes qui ne le feront pas seront vite très pénalisées par rapport à la concurrence. Pour permettre cet accès sans surinvestir, **il est donc nécessaire d'investir sur une ou quelques plateformes collectives INRA ayant réellement une vocation et une culture d'ouverture**, apportant à un grand nombre d'équipes INRA un accès mutualisé à ces technologies. Ces plateformes devront s'organiser en réseau.

Coté bioinformatique : Si les séquenceurs HD actuels seront très certainement dépassés assez rapidement, **les besoins informatiques resteront**. Il est urgent de développer encore plus les moyens humains et l'expertise dans ce domaine. En effet, quelle que soit la machine qui les a produits, il y aura toujours à traiter et assembler des centaines de millions de séquences, et l'expérience acquise dès aujourd'hui restera valable. **L'enjeu majeur et le facteur limitant pour les équipes actuellement investies dans le domaine du séquençage HD n'est pas de produire les séquences, mais de savoir les traiter**. *L'INRA doit investir des moyens humains dans cette expertise de façon coordonnée entre les domaines animaux, végétaux et microbiens*. Dans le domaine animal, Sigenae a une légitimité évidente pour développer ces compétences. Il faudra également être attentif au renouvellement de l'infrastructure informatique nécessaire.

Ces évolutions nécessitent de s'appuyer sur le dispositif collectif des plateformes INRA. Elles induisent une convergence des objectifs des plateformes de séquençage/ génotypage et des plateformes transcriptome. Le succès de ces approches HD nécessitera également très clairement une implication très forte des plateformes bioinformatiques. Il faut en tenir compte dans la structuration du dispositif collectif de l'INRA.

Des points concrets d'organisation ...

Coté structuration du dispositif pour les investissements internes à l'INRA : Au niveau biologique comme au niveau informatique, il me semble qu'il faut identifier les étapes communes à toutes les applications (en gros le séquençage lui-même, le stockage, le nettoyage et le contrôle qualité des séquences). Ces étapes devraient mobiliser des personnes identifiées et positionnées sur des plateformes à même de garantir la qualité des résultats. **Pour les étapes biologiques en amont, comme pour les étapes bioinformatiques en aval, il me semble qu'il faut développer des compétences spécifiques en lien avec les équipes de recherche**, car les plateformes n'auront jamais assez de personnel pour prendre en charge l'ensemble du projet de l'échantillon biologique à l'annotation des séquences produites (CNS et CNG nous alertent en disant que les plateformes ne peuvent pas tout faire par elle-même, et doivent disant ils se spécialiser). Les plateformes serviront également de nœud central pour la mutualisation, ou a minima la mise en réseau, des compétences développées par des experts référant externes aux plateformes.

Pour la *partie biologie* : les plateformes devront pouvoir accueillir des biologistes des équipes de recherche à même de développer les compétences spécifiques en amont du séquençage (production des bibliothèques à partir de cDNA, d'ADN méthylé, de banques de représentation réduites par AFLP ou digestion, de produits de ChIP ou CHIP ...). Une fois les compétences maîtrisées pour un projet pilote, l'équipe plateforme pourra servir de relai (avec l'appui des experts du projet pilote) pour mettre à disposition ces compétences aux autres équipes de recherche.

Pour la *partie bioinformatique* : lors des projets pilotes d'analyse, les chefs de projets investis sur un type de projet (reséquençage de génome, analyse de transcrits, découverte de petits ARN, assemblage d'une espèce en se servant du génome modèle d'une espèce proche ...) auront à cœur d'organiser le traitement des données en développant des pipelines permettant de faciliter l'analyse des manips du même type qui seront réalisés ultérieurement. Si le chef de projet n'est pas

directement situé au sein d'une plateforme, il faudra organiser la récupération des chaînes de traitement développées pour leur mise en œuvre ultérieure (là aussi avec l'appui des experts ayant développés ces chaînes).

Comme indiqué plus haut, il est enfin fondamental de mobiliser des forces *pour aider les équipes à monter leur projet*. Même si les prix ont beaucoup baissés, le coût (tout compris : réactifs, machine, temps de personnel pour le traitement des données ...) d'une manip sur un Roche reste de l'ordre de 8-10 k€. Le pire serait de mener des projets sous dimensionnés pour lesquels les données produites seraient ininterprétables (comme par exemple l'assemblage de novo d'un jeu de séquences produit avec une profondeur insuffisante). Les plateformes devront être à même de pouvoir conseiller les équipes sur le montage et le dimensionnement de leurs projets.

Enfin, si les investissements doivent se raisonner au sein de l'INRA, ils doivent également se raisonner en tenant compte du paysage régional. **En complément des capacités nationales du CNS/CNG, il sera particulièrement important en région d'éviter les doublons localement entre partenaires du GIS IBISA**, en donnant à des plateformes partagées des missions dépassant les limites des organismes. Une telle politique permettra d'optimiser les investissements, et de ménager des capacités d'évolution dans les années futures.