

PYROSEQUENÇAGE POUR LE DEVELOPPEMENT D'EST ET DE SNP AVIAIRES

Pitel Frédérique¹, Vignal Alain¹, Leroux Sophie¹, Fève Katia¹, Vignoles Florence¹, Tircazes Aurélie¹, Morisson Mireille¹, Marty Amandine¹⁰, Donnadieu Cécile^{1,10}, Milan Denis^{1,10}, Gourichon David², Minvielle Francis³, Leterrier Christine⁴, Arnould Cécile⁴, Bernadet Marie-Dominique⁵, Marie-Etancelin Christel⁶, Basso Benjamin⁶, Hérault Frédéric⁷, Lecerf Frédéric⁷, Besnard Joël², Calenge Fanny⁸, Beaumont Catherine⁸, Klopp Christophe⁹, Diot Christian⁷

¹UMR444 LGC, INRA-ENVT, 31326 CASTANET-TOLOSAN ; ²UE1295 PEAT, INRA, 37380 NOUZILLY ; ³UMR1236 GDA, INRA-AgroParisTech, 78352 JOUY-EN-JOSAS ; ⁴UMR85 PRC, INRA-CNRS-Haras Nationaux-Univ. Tours, 37380 NOUZILLY ; ⁵UE89 PFG, INRA, 40280 BENQUET ; ⁶UR631 SAGA, INRA, 31326 CASTANET-TOLOSAN ; ⁷UMR598 GA, INRA-AgroCampus Rennes, 35042 RENNES ; ⁸UR83 URA, INRA, 37380 NOUZILLY ; ⁹UR875 BIA, INRA, 31326 CASTANET-TOLOSAN ; ¹⁰Plateforme Génomique INRA Toulouse-Auzeville 31326 CASTANET-TOLOSAN

RÉSUMÉ

Le but du programme est de combler les déficits en marqueurs observés pour trois espèces aviaires : la caille, le canard et la poule. La stratégie choisie est l'obtention, à partir de plusieurs individus de lignées d'intérêt, de SNP (Single Nucleotide Polymorphism, polymorphisme d'un nucléotide) par une nouvelle technologie de séquençage à haut débit (séquenceur 454 GS-FLX, Roche). Nous séquençons des représentations réduites du génome, en sélectionnant d'une part des fragments de restriction d'ADN génomique - les mêmes chez tous les individus - et d'autre part les transcrits qui représentent globalement la partie du génome correspondant aux gènes exprimés. Ces expérimentations sont réalisées à partir d'échantillons d'ADN ou d'ARN issus d'individus de lignées à l'origine de croisements existants, pour chacune des trois espèces.

Les données générées par plusieurs "runs" de séquence seront traitées *in silico* : contigage à haut débit, recherche de SNP, comparaison avec les banques de séquences connues...

En plus de l'intérêt que représente la production d'un très grand nombre de SNP nouveaux, cette technologie devrait permettre de mieux séquencer les régions riches en (G+C) correspondant aux plus petits des microchromosomes pour lesquels il n'y a pas de séquence chez la poule.

La comparaison des séquences des transcrits obtenues chez la caille et le canard avec la séquence du génome de la poule permettra d'établir une "cartographie virtuelle" des SNP obtenus, grâce à la grande conservation de synténie existant entre ces trois espèces.

ABSTRACT

The aim of the project is to fill the lack in markers observed for three avian species (quail, duck and poultry). The chosen strategy is to obtain, by using individuals from several lines of interest, SNP (Single Nucleotide Polymorphism) by a high-throughput sequencing technology (sequencer 454 GS-FLX, Roche). Two reduced representations of the genome are sequenced: size-selected digested genomic DNA and transcriptome, which globally represents the part of the genome corresponding to the expressed genes. These experiments are realized from samples of DNA or RNA from individuals of lines from existing crosses, for each species. The data generated by several sequencing runs will be *in silico* analyzed. Besides the interest of the production of a very large number of new SNP, this technology should allow to sequence GC rich regions corresponding to the smallest microchromosomes for which there is no sequence in chicken. The comparison of the transcriptome sequences in quail and duck with the chicken genome assembly will allow to establish a "virtual cartography" of the obtained SNP, thanks to the synteny conservation existing between these three avian species.

INTRODUCTION

La diversité de la production des volailles en France nous a amenés à poursuivre des recherches en génétique moléculaire pour trois espèces : la poule, le canard et la caille. Celles-ci visent à identifier les gènes majeurs et QTL (Quantitative Trait Loci) responsables de la variabilité des phénotypes observés dans des croisements ou des populations. Pour cela, il est nécessaire de disposer d'un nombre suffisant de marqueurs génétiques faciles et économiques à génotyper, tels que les SNP (Single Nucleotide Polymorphism), et d'outils pour analyser l'expression des gènes, telles que des collections d'EST. Pour les espèces non encore séquencées, l'une des voies d'obtention de ces SNP est le séquençage à haut-débit d'une fraction du génome, comparée entre plusieurs individus ou à une séquence de référence.

Le but du présent programme est de combler les déficits en marqueurs observés pour ces trois espèces, en développant un grand nombre de SNP, par pyroséquençage à haut débit (séquenceur 454 GS-FLX, Roche) de deux représentations réduites des génomes : les transcrits d'une part, sous forme de cDNA, et une sélection de fragments de restriction d'ADN génomique d'autre part, permettant dans les deux cas de s'assurer de la bonne représentation des microchromosomes.

Chez la poule, bien que déjà riche en marqueurs, la couverture actuelle du génome par la séquence et les cartes génétiques ne permet pas la détection de tous les QTL existants. En effet, avec une absence quasi-totale des dix plus petits microchromosomes, ces régions denses en gènes sont presque totalement exclues des criblages du génome, faute de marqueurs disponibles.

Le canard est une espèce pour laquelle une recherche de QTL impliqués dans le comportement, la résistance au portage de *Salmonelles*, et surtout la qualité des produits et le métabolisme hépatique sous-jacent, a été entreprise dans le cadre du projet ANR GENECAN. La création de familles informatives a été réalisée et les phénotypes sont en cours. Cependant, 112 marqueurs microsatellites seulement sont informatifs dans ce croisement (Feve et al., 2009). Il s'ensuit une très mauvaise couverture du génome, rendant impossible l'identification d'une bonne part des QTL impliqués. Sachant que la totalité des marqueurs disponibles dans les bases de données a été testée, que le développement de marqueurs microsatellites est particulièrement difficile chez les oiseaux et que le génotypage de SNP est plus aisé et moins coûteux, le développement de marqueurs supplémentaires, de type SNP, est pleinement justifié. Le développement d'un panel d'hybrides irradiés est prévu chez le canard. Là encore, notre projet devrait permettre d'identifier de nombreux fragments permettant la réalisation d'une carte RH, et donc de s'assurer d'une couverture du génome homogène pour améliorer l'efficacité de détection des QTL.

Espèce avicole secondaire, la caille n'en connaît pas moins une production de niche significative en France (premier producteur européen avec l'Espagne) et elle est très répandue en Asie. La coexistence en France et en Europe de deux espèces de *Coturnix*, la caille japonaise d'élevage - parfois soumise à des relâchés accidentels ou illégaux dans la nature - et la caille des blés sauvage, souligne l'intérêt de développer des marqueurs moléculaires à des fins à la fois de gestion de la faune et d'études phylogénétiques. Animal modèle dans plusieurs domaines (embryologie, toxicologie...), la caille permet également d'étudier des caractères non observés chez la poule, comme les couleurs du plumage déterminées par le gène *ASIP* (agouti) et ayant des effets associés sur d'autres caractères (Hiragaki et al., 2008, Nadeau et al., 2008). La possibilité de produire des croisements à moindre coût (financier et temporel) qu'avec la poule, nous a également amenés à mettre en place des lignées divergentes de cailles, sélectionnées pour des caractères d'intérêt (comportement, ponte...) et à l'origine de la mise en évidence de QTL, notamment de croissance et de comportement (Beaumont et al., 2005, Minvielle et al., 2005). Concernant un caractère de comportement, l'immobilité tonique, nous disposons actuellement de deux dispositifs de croisements avancés AIL (Advanced Intercross Lines) de génération F6 et F7. Ce matériel animal doit permettre - si la densité de marqueurs informatifs disponibles est suffisante - de définir très finement la région du génome impliquée dans la variabilité du caractère, avec l'objectif d'identifier le gène responsable. Des cartes génétiques sont disponibles mais ont une densité insuffisante pour l'analyse d'AIL.

La possibilité de disposer de marqueurs SNP et d'EST en très grand nombre pour ces trois espèces permettra de réaliser des cartes comparées de haute densité. L'intégration d'EST dans les cartes génétiques permettra de faciliter la comparaison des cartes entre espèces, tandis que le développement de marqueurs dans les régions non codantes facilitera l'intégration avec les cartes d'hybrides irradiés. Le séquençage des transcrits est un préalable nécessaire au développement d'outils d'analyse du transcriptome tels que des puces à oligonucléotides, qui permettront d'ajouter des expressions de gènes aux caractères étudiés dans le cadre des programmes en cours, autorisant finalement la conduite de projets « QTL d'expression ». Il n'est d'ailleurs pas exclu, et est même probable avec l'évolution favorable des coûts de séquençage, que le pyroséquençage de cDNA à haut-débit puisse s'y substituer totalement (Torres et al., 2008, Weber et al., 2007).

La densité plus grande de gènes sur les microchromosomes, ainsi que le choix d'une enzyme de restriction ciblant les régions riches en (G+C), permettra également d'orienter favorablement le développement de marqueurs dans les trois espèces étudiées.

1. MATERIELS ET METHODES

1.1 Détection de SNP par séquençage réduit du génome fragmenté

Le principe de la technique est de séquencer un mélange de fragments de restriction d'ADN génomique d'une fenêtre de taille donnée (Altshuler et al., 2000). Ce mélange est obtenu par digestion enzymatique d'ADN génomique issu d'une extraction au NaCl (Roussot et al., 2003), migration sur gel d'agarose et découpe dans le gel de la fraction de taille à séquencer, puis purification de celle-ci sur membrane de silice. Ceci permet de reséquencer la même fraction de génome d'un individu à un autre, au polymorphisme de restriction près. Le choix des enzymes de restriction chez la poule a été fait afin de privilégier les microchromosomes, qui sont riches en séquences (G+C). Une analyse de l'assemblage actuel montre que l'enzyme de restriction *MspI* (site de coupure CCGG) produit 15000 fragments de taille comprise entre 600 et 700 paires de bases, dont par exemple 1780 sur le chromosome 1 (8,9/Mb) et 193 sur le chromosome 28 (42,8/Mb), soit un facteur d'enrichissement de 4,8, ce qui est tout à fait satisfaisant, compte tenu que d'autres techniques plus complexes telles que le tri de chromosomes ne sont pas exemptes de contaminations par les macrochromosomes. On peut même supposer, compte tenu de la corrélation négative entre le pourcentage en (G+C) et la taille du microchromosome, que l'enrichissement sera encore meilleur pour les plus petits microchromosomes manquants (Hillier et al., 2004). Une autre stratégie est également testée, plus économique en ADN et permettant d'identifier l'origine individuelle des séquences obtenues, mais induisant un biais de représentation de certains fragments : l'utilisation d'une sous-représentation du génome issue d'une digestion enzymatique suivie d'une amplification PCR d'une partie du mélange obtenu, grâce à la technique d'AFLP (Amplified Fragment Length Polymorphism). Chez la caille, une carte AFLP étant disponible, cette stratégie devrait également nous permettre de connaître a priori la localisation génétique d'une partie des séquences obtenues. Après extraction comme ci-dessus et digestion par les enzymes de restriction *TaqI* et *EcoRI*, une ligation des adaptateurs est suivie par une pré-amplification classique (Roussot, et al., 2003), à l'exception de l'utilisation d'amorces phosphorylées et identifiées par de courtes séquences, appelées "tags" (Tableau 1). Les amplifiats ainsi obtenus sont séquencés sur le séquenceur GS-FLX 454.

Si le niveau de polymorphisme est aussi élevé chez le canard et la caille que chez la poule (entre 4 et 5 SNP par kb, lors de la comparaison de 2 chromosomes seulement, selon les lignées comparées), on devrait obtenir des marqueurs en nombre suffisant pour effectuer des criblages de génome à moyenne densité avec une bonne couverture pour les 3 espèces en utilisant les technologies les plus récentes. Un run de

séquençage par espèce est donc réalisé : en digestion enzymatique chez la poule (6 F0 par lignée), en AFLP chez la caille (2 F0 par lignée), et avec l'une de ces 2 techniques chez le canard, choisie en fonction des résultats obtenus pour les deux autres espèces. Chaque run de séquençage (environ 1 million de lectures) doit permettre d'obtenir une profondeur d'au moins 30x sur l'ensemble des quelques 15 000 fragments (soit 30 000 extrémités) obtenus par digestion ou AFLP, qui représentent un peu moins de 1% du génome.

1.2 Détection de SNP par séquençage du transcriptome

La méthode de détection de SNP à partir de cDNA séquencés ("transcriptome sequencing") a prouvé son efficacité (Barbazuk et al., 2007) et devrait s'avérer fructueuse également dans nos espèces aviaires ; chez la poule, le taux de polymorphisme, plus élevé que chez l'homme (Zhao et al., 2003), est de 2,1 SNP/kb dans les exons et de 3,4 SNP/kb dans les régions 3' non-codantes entre lignées de chair et pondeuses, les taux ne faiblissant pas de manière drastique dans les analyses intra-populations (Wong et al., 2004). Il est donc fort probable que pour la caille et le canard un nombre similaire de SNP puisse également être obtenu.

Les cDNA sont obtenus à partir d'échantillons d'ARN issus d'individus des lignées à l'origine des croisements existants, pour obtenir un maximum de SNP informatifs. Les ARNm, sélectionnés par passage sur colonnes oligo dT25, sont convertis en cDNA double-brins hémiméthylés, par l'utilisation d'une amorce comprenant un polyT et un site de reconnaissance de l'enzyme de restriction *GsuI* (Tableau 1), l'action de la reverse transcriptase, puis d'enzymes de synthèse du second brin (*E. Coli* RNase H, DNA polymérase I et DNA ligase). La queue polyA, peu compatible avec le pyroséquençage, est éliminée par digestion avec l'enzyme *GsuI*.

L'objectif est d'obtenir une profondeur suffisante pour détecter plusieurs milliers de SNP de manière fiable, objectif atteint avec un minimum de 30 Mb de séquence par lignée (Barbazuk, et al., 2007), limite qui sera dépassée grâce à l'utilisation du séquenceur GS-FLX, et ses quelques 400-500 Mb par run.

1.3 Analyse des données

Les données de séquences générées sont traitées *in silico* par l'équipe Sigenae, à l'aide d'outils déjà développés : contigage à haut débit, recherche de SNP et sélection des plus fiables, comparaison avec les banques de séquences connues...

2. RESULTATS ET DISCUSSION

Les tissus prélevés - choisis afin de maximiser la complexité du transcriptome - dépendent des analyses ultérieures effectuées à partir de ces données, et donc de l'espèce étudiée : chez la caille nous avons prélevé le cerveau d'animaux appartenant aux trois lignées à

l'origine des AIL, chez la poule des tissus d'embryons de lignées à l'origine d'un croisement QTL, chez le canard plusieurs tissus provenant d'animaux des 2 lignées de canard commun à l'origine du croisement QTL, et d'animaux de l'espèce Barbarie. La comparaison des résultats obtenus pour les canards commun et de Barbarie devrait en outre apporter un jeu de données préalable à l'analyse différentielle des transcriptomes de ces deux espèces. Les ARN ont été extraits de ces tissus, les cDNA double-brins sont en cours de synthèse. L'ADN de poule, caille et canard utilisé pour les runs "génomiques" du programme a été extrait. Des fragments AFLP ont été obtenus chez la caille. A ce jour, seul un demi run, préliminaire, a été réalisé à la génopole toulousaine sur ces fragments (AFLP de caille). Un total de 372 270 séquences ont

été obtenues, pour 4928 contigs (figure 1). La recherche de SNP est en cours. Un exemple d'assemblage est donné sur la figure 2.

CONCLUSION

En plus de l'intérêt que représente la production d'un très grand nombre de SNP nouveaux, cette technologie devrait permettre de mieux séquencer les régions riches en (G+C) correspondant aux plus petits des microchromosomes pour lesquels il n'y a pas de séquence chez la poule.

La comparaison des séquences d'EST obtenues chez la caille et le canard avec la séquence du génome de la poule permettra d'établir une "cartographie virtuelle" des SNP obtenus, grâce à la grande conservation de synténie existant entre ces trois espèces.

REMERCIEMENTS

Ce projet est financé par l'AIP BioRessources 2008, INRA. Les auteurs remercient Patrice Dehais (SIGENAE), Céline Noirot et Gérald Salin (génopole Toulouse Midi-Pyrénées) pour l'analyse préliminaire des séquences.

REFERENCES BIBLIOGRAPHIQUES

- Altshuler D, Pollara V, Cowles CR, Van Etten W, Baldwin J, Linton L et Lander ES, 2000. *Nature*, (407), 513-6.
- Barbazuk WB, Emrich SJ, Chen HD, Li L et Schnable PS, 2007. *Plant J*, (51), 910-8.
- Beaumont C, Roussot O, Feve K, Vignoles F, Leroux S, Pitel F, Faure JM, Mills AD, Guemene D, Sellier N, Mignon-Grasteau S, Le Roy P et Vignal A, 2005. *Anim Genet*, (36), 401-7.
- Feve K, Bounet M, Vignoles F, Leroux S, Bardes S, Vignal A et Marie-Etancelin C., 2009, 8ième JRA , St Malo, 25 et 26 mars 2009.
- Hillier LW, et al., 2004. *Nature*, (432), 695-716.
- Hiragaki T, Inoue-Murayama M, Miwa M, Fujiwara A, Mizutani M, Minvielle F et Ito S, 2008. *Genetics*, (178), 771-5.
- Minvielle F, Kayang BB, Inoue-Murayama M, Miwa M, Vignal A, Gourichon D, Neau A, Monvoisin JL et Ito S, 2005. *BMC Genomics*, (6), 87.
- Nadeau NJ, Minvielle F, Ito S, Inoue-Murayama M, Gourichon D, Follett SA, Burke T et Mundy NI, 2008. *Genetics*, (178), 777-86.
- Roussot O, Feve K, Plisson-Petit F, Pitel F, Faure JM, Beaumont C et Vignal A, 2003. *Genet Sel Evol*, (35), 559-72.
- Torres TT, Metta M, Ottenwalder B et Schlotterer C, 2008. *Genome Res*, (18), 172-7.
- Weber AP, Weber KL, Carr K, Wilkerson C et Ohlrogge JB, 2007. *Plant Physiol*, (144), 32-42.
- Wong GK, et al., 2004. *Nature*, (432), 717-22.
- Zhao Z, Fu YX, Hewett-Emmett D et Boerwinkle E, 2003. *Gene*, (312), 207-13.

Tableau 1. Oligonucléotides utilisés

Les Tags utilisés (NNNN) sont : TACG, TAGA, TATC, TCAG, TCGT, TCTA, TGAT, TGCA

Nom de l'oligonucléotide	Séquence
Adaptateurs de ligation AFLP	
Adaptateur TaqI U	GAC GAT GAG TCC TGA C
Adaptateur TaqI L	CGG TCA GGA CTC AT
Adaptateur EcoRI U	CTC GTA GAC TGC GTT ACC
Adaptateur EcoRI L	CTG ACG CAA TGG TTA A
Amorces d'amplification AFLP à une base d'ancrage (A)	
EcoApNNNN	P-NNNNCTGCGTTACCAATTCA
TaqApNNNN	P-NNNNGATGAGTCCTGACCGAA
Amorce de reverse transcription	
Gsu(T)16VN	GAGAGAGAGACTGGAGTTTTTTTTTTTTTTTTTVN

Figure 1. Répartition des tailles des contigs obtenus

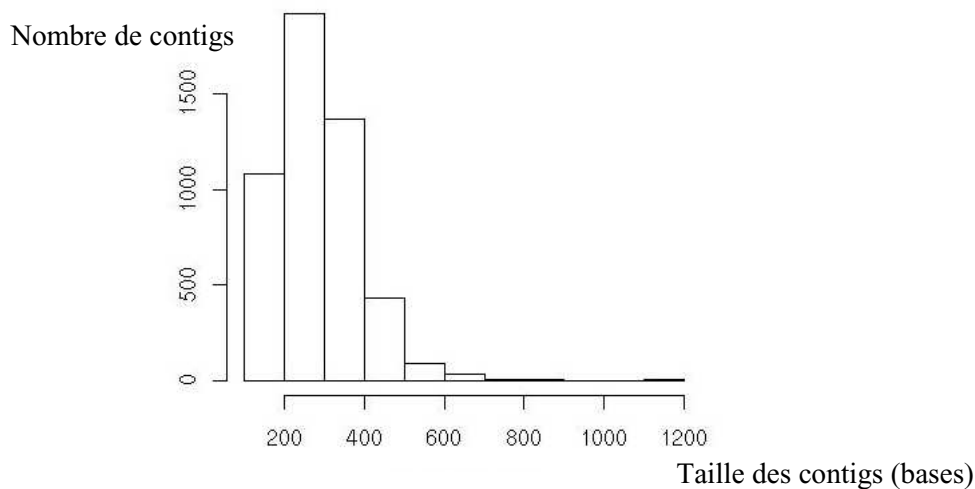


Figure 2. Exemple de SNP détectés sur un contig (AFLP cailles)

L'alignement des séquences visualisé par clview (<http://compbio.dfci.harvard.edu/tgi/software/>) donne un consensus (jaune), les bases communes (bleu) et les polymorphismes (bleu clair/rouge). On voit ici clairement deux haplotypes (A-TAT-A-A / C-GTA-T-C), et un SNP correspondant probablement à une erreur de séquence (flèche rouge).

