

DEVELOPPEMENT D'UNE PUCE DE GENOTYPAGE HAUTE DENSITE 600K POUR LE CANARD COMMUN ET LE CANARD DE BARBARIE

Thébault Noémie¹, Riquet Juliette¹, Diot Christian², Brard-Fudulea Sophie³, Guémené Daniel³, Alletru Bernard⁴, Cornil Maude⁴, Bouleau Pascal⁴, Blanchet Magali⁵, Le Mignon Guillaume⁵, Demeure Olivier⁵, Vignal Alain¹

¹GenPhySE UMR 1388, Université de Toulouse, INRA, ENVT, 31320 CASTANET-TOLOSAN

²PEGASE UMR 1348, Agrocampus Ouest, INRA, 35590 SAINT-GILLES

³SYSAF, Centre INRA Val de Loire, 37380 NOUZILLY

⁴Gourmaud Sélection, Groupe Orvia, 85260 SAINT ANDRE 13 VOIES

⁵Grimaud Frères Sélection, Groupe Grimaud, La Corbière, 49450 ROUSSAY

sophie.brard-fudulea@inra.fr

RESUME

La sélection génomique est aujourd'hui largement utilisée pour l'amélioration génétique des animaux et des plantes. Chez le canard, cette méthode pourrait permettre d'améliorer le progrès génétique pour des caractères non-mesurables sur les candidats à la sélection (létaux, exprimés chez un seul sexe, ou mesurés sur l'hybride pour la sélection des lignées pures). Malgré l'intérêt potentiel de l'utilisation de la sélection génomique chez le canard, aucun outil moderne de génotypage n'est disponible pour les canards commun (*Anas platyrhynchos*) et de Barbarie (*Cairina moschata*). Cette étude décrit la finalisation de l'assemblage du génome du canard de Barbarie, et la mise au point d'un outil de génotypage haut-débit et haute-densité pour les 2 espèces : une puce 600K Thermo Fisher SNP. L'assemblage du génome du Barbarie a permis d'obtenir 3702 scaffolds avec un N50 de 2,4 Mb. Les scaffolds ont ensuite été ordonnés en chromosomes, par alignement sur le génome du canard commun. Pour l'identification des SNP, des pools (mélanges) d'ADN de canards colvert, de Rouen, ainsi que de lignées commerciales de canards Pékin et de Barbarie ont été séquencés. Les séquences ont été alignées sur les génomes de référence correspondant. A l'issue des contrôles sur la qualité, 8,4 et 12,2 millions de SNP avec une Minor Allele Frequency (MAF) $\geq 0,05$ ont été identifiés pour le canard de Barbarie et pour le canard commun respectivement. Les SNP ont été filtrés afin de maximiser le nombre de marqueurs informatifs par population et d'assurer une bonne couverture du génome. A l'issue de ces tris, 343 950 SNP d'une part et 331 241 SNP d'autre part ont été choisis pour les canards commun et de Barbarie.

ABSTRACT

Development of a high-density 600K SNP array for Muscovy and Common Ducks

Genomic selection is widely used for genetic improvement of many plant and animal species. This evaluation method could potentially increase genetic gain in ducks for traits that cannot be measured on selection candidates (lethal, expressed in one sex only, or measured on hybrid duck for purebred selection). Implementation of genomic selection depends on the availability of genotyping tools, such as SNP chips, but so far none were developed neither for common duck (*Anas platyrhynchos*) nor Muscovy duck (*Cairina moschata*). This paper describes the assembly of the Muscovy duck genome (the common duck genome is already available), and the design of a 600K Thermo Fisher SNP chip. The common duck genome was used as reference for the assembly of the Muscovy genome: 3702 scaffolds were produced, with a N50 of 2.4 Mb. SNP were identified from sequence data: several Muscovy and common ducks populations (Rouen duck, Mallard duck, Pekin duck), were collected, and 50 samples were pooled independently for each population. After raw quality controls, 8.4 million SNPs and 12.2 million SNP were identified for Muscovy and common ducks respectively. Filters based on Minor Allele Frequency and genome coverage were used, and finally 343 950 SNP were kept for the common duck, and 331 241 SNP for the Muscovy duck.

INTRODUCTION

La mise en œuvre de la sélection génomique, qui révolutionne depuis plusieurs années l'amélioration génétique de plusieurs espèces, repose sur l'obtention des génotypes grâce à l'utilisation de puces SNP. Aucun outil de génotypage n'existe pour les deux espèces de canard Pékin et Barbarie, ainsi que pour le mulard produit de leur croisement, tous trois d'intérêt zootechnique en France. Il existe cependant un avantage potentiel d'évaluations génomiques pour des caractères difficiles à mesurer, voire non-mesurables sur les candidats à la sélection : robustesse, résilience, adaptabilité, résistances aux maladies, caractères de production et/ou de découpe mesurés sur le mulard pour la sélection des lignées des deux espèces parentales, caractères de reproduction mesurés sur les femelles pour la sélection des mâles... Le projet CanArray vise à développer une puce 600K SNP Thermo Fisher utilisable pour génotyper des canards communs et de Barbarie. Pour le développement de la puce, le génome du canard commun était disponible (Huang *et al.* 2013), mais pas celui du canard de Barbarie. Cet article décrit dans un premier temps la finalisation de l'assemblage du génome du canard de Barbarie, et dans un second temps l'identification et le choix de marqueurs SNP pour la première puce de génotypage mise au point pour les canards communs et de Barbarie.

1. MATERIEL & METHODES

1.1 Assemblage du génome du canard de Barbarie

Les séquences d'ADN disponibles pour assembler le génome du canard de Barbarie proviennent d'un individu mâle consanguin. Elles ont été obtenues dans le cadre du programme AGENAVI SeqVol, en collaboration avec l'Université Washington (Saint-Louis, USA) et se composent de 3702 scaffolds, de 0,1 à 15 Mb, pour une longueur totale de 1,109 Gb. L'assemblage des scaffolds en chromosomes a été réalisé par alignement sur le génome du canard commun.

1.1.1 Alignement et chaînage

L'alignement a été réalisé avec l'outil LAST (Frith, 2011). Les paramètres optimaux pour la réalisation de l'alignement ont été obtenus avec l'outil last-train (Hamada *et al.* 2017) : les paramètres par défaut sont utilisés par ce logiciel, puis ajustés itérativement en fonction des résultats obtenus jusqu'à stabilisation de la qualité de l'alignement. En raison de régions de similarité plus ou moins importantes entre les deux espèces, des blocs d'alignements sont obtenus, séparés par des régions ne s'alignant pas.

Le chaînage de ces blocs a été réalisé afin de les relier et de rajouter les séquences voisines sur le génome mais non alignées, tout en supprimant les blocs d'alignements chevauchants. Un script a été développé afin de permettre à l'utilisateur d'indiquer la distance maximale autorisée entre deux alignements pour le chaînage, avec pour objectif de produire les chaînes les plus longues possibles et sans recouvrement des séquences.

1.1.2 Filtre des alignements et assignation chromosomique

Des problèmes de scaffolds du canard de Barbarie s'alignant sur plusieurs régions du génome du canard commun compliquent grandement l'assignation chromosomique. La plupart sont dus à des micro-alignements de très petite taille causés par les séquences répétées, mais d'autres peuvent être dus à des problèmes d'alignements réels sur deux chromosomes, soit dus à des problèmes d'assemblage des scaffolds, soit à des séquences récemment dupliquées dans le génome. Ceci empêche une assignation chromosomique automatique des scaffolds. Afin de résoudre ce problème, le ratio entre le plus long alignement et le second en taille a été calculé pour chaque scaffold, sur lequel l'application d'un seuil a permis d'éliminer les micro-alignements. Les scaffolds filtrés ont été assignés à leur chromosome et retournés en cas d'alignement anti-sens avec la fonction `reverse_complement` de BioPython (Cock *et al.* 2009). Les scaffolds ne s'alignant pas ou n'ayant pas passé les filtres précédemment appliqués ont été intégrés au chromosome Unknown.

1.2 Construction de la puce

1.2.1 Séquençage

Un séquençage en mélange (pools) de 50 animaux par population a été réalisé sur des animaux échantillonnés parmi des populations françaises de canard colvert, de canard de Rouen, de canard Pékin et de canard de Barbarie. Pour les canards Pékin et de Barbarie, les échantillons proviennent des lignées expérimentales de l'INRA et des lignées des sélectionneurs français Grimaud Frères Sélection et Orvia Gourmaud Sélection. Compte-tenu de l'absence du chromosome W dans les deux génomes de référence des canards Pékin et de Barbarie, seuls des mâles ont été prélevés, à l'exception des Colverts pour lesquels des échantillons mâles et femelles ont été fournis.

Les bibliothèques de séquençage ont été réalisées avec le kit TruSeq DNA PCR, et les pools ont été séquencés avec la technologie Illumina HiSeq3000 à une profondeur de 50X.

1.2.2 Détection et caractérisation des SNP

Les séquences obtenues dans les différents pools ont été alignées sur les génomes de référence correspondants avec l'algorithme BWA (Li & Durbin, 2009), les alignements triés, puis les duplicats de PCR ont été marqués avec les outils Picard SortSam et MarkDuplicates (<http://broadinstitute.github.io/picard/>). La recalibration de la qualité des données de séquençage conservées a été réalisée suivant le protocole proposé par Genome Analysis Tool Kit (McKenna *et al.* 2010).

Les fréquences alléliques ont été estimées avec le script Python pool-hmm.py (Boitard *et al.* 2013). Un autre script développé en Python regroupe les populations à analyser et conserve les positions pour lesquelles au moins une population présente un allèle différent de celui connu dans le génome de référence.

1.2.3 Qualité technique et spécificité génomique

Les SNP bi-alléliques polymorphes ont été sélectionnés dans les régions de couverture > 25X (plus de 25 lectures validées pour le locus), filtrés afin de ne garder que ceux à +/- 50 pb d'une insertion-délétion, et distants d'au moins 35 paires de bases d'un autre SNP (contrainte due à la longueur de l'amorce mise au point par Thermo Fisher). Finalement, les SNP retenus ont été alignés sur les génomes de référence des espèces correspondantes sur une longueur de 71 pb (le polymorphisme et 35 bases de part et d'autre), et seuls les alignements uniques ont été conservés. Un jeu de SNP communs aux deux espèces a également été identifié en alignant les séquences flanquantes des marqueurs identifiés chez les canards communs et de Barbarie.

1.2.4 Répartition et informativité

Le choix des SNP a tenu compte de leur répartition sur le génome pour différents niveaux de MAF donnés. La distance idéale entre 2 SNP a été estimée à 1,6 kb (600K marqueurs pour couvrir 1 Gb). Le choix des SNP a été réalisé en plusieurs étapes : il y a eu dans un 1^{er} temps un choix de SNP avec une MAF $\geq 0,15$ dans 75% des populations et répartis sur l'ensemble du génome, puis l'ajout de SNP avec le même critère de MAF dans 60% des populations, et enfin de SNP avec une MAF d'au moins 0,05 toutes populations confondues.

Un ordre de priorité des SNP pour le design de la puce a été défini : priorité 1 si le SNP était distant du SNP voisin de plus de 1,6 kb, priorité 2 si la distance était inférieure, priorité 3 si les allèles du polymorphisme étaient complémentaires (A/T ou C/G).

2. RESULTATS & DISCUSSION

2.1 Assemblage du génome du canard de Barbarie

2.1.1 Alignement et chaînage

Après alignement avec Last, 2397 des 3702 scaffolds de départ s'alignent sur le génome du canard commun avant chaînage et filtration. Les scaffolds ne s'alignant pas sont en fait de très petite taille et ne représentent qu'une couverture totale de 16,5 Mb. La séquence de ces scaffolds est soit trop divergente entre les deux espèces, spécifique du Barbarie ou pas séquencée pour des raisons techniques chez le Pékin. Parmi les scaffolds alignés, seuls 1177, de petite taille, s'alignent sur une cible unique et sont donc assignés à un chromosome, pour un total de 3,2 Mb seulement. Les 1220 scaffolds restants couvrent 1 108 980 Mb et s'alignent chacun sur plusieurs régions, notamment à cause de très courtes régions de similarité entre les deux espèces, dont la probabilité de présence augmente avec la longueur des scaffolds (Figure 1A). En effet, ces grands scaffolds ont dans la plupart des cas la très grande majorité de leur séquence alignée sur un unique chromosome Pékin, et les régions alignées sur d'autres chromosomes sont beaucoup plus petites (moins de 100kb).

L'étape de chaînage, qui permet de ne retenir que l'alignement principal, a conduit à supprimer un total de 8,74 Mb. La figure 1A et 1B présente un exemple d'alignement du scaffold 0 du canard de Barbarie sur le chromosome 4 du canard commun, avant et après chaînage.

2.1.2 Filtre des alignements

Après chaînage, certains scaffolds ont toujours des alignements multiples. Pour les résoudre, le seuil pour le filtre sur le ratio entre le plus long alignement et le second en taille a été fixé à 1% pour les scaffolds de taille inférieure au N50, et à 3% pour les scaffolds de taille supérieure au N50, afin de tenir compte du fait qu'un scaffold plus long a plus de chances de s'aligner à plusieurs endroits du génome. Un filtre sur le pourcentage d'alignement a été fixé à 60% afin d'éliminer les faux-positifs.

Ces 2 filtres ont conduit à l'élimination de 618 scaffolds. *In fine*, le génome du canard de Barbarie aligné sur le génome du canard commun compte 1780 scaffolds, pour une couverture du génome de 0,966 Gb, et un N50 de 2,194 Mb.

2.1.3 Assignment chromosomique des scaffolds

Les résultats de l'assignment chromosomique des scaffolds sont présentés dans la figure 2. Les longueurs des chromosomes sont du même ordre de grandeur chez le canard de Barbarie, le canard commun et la Poule. La taille du chromosome

unknown est calculée comme la somme des séquences non assignées. Le chromosome unknown est plus petit chez le canard de Barbarie que chez le canard commun car du fait de l'évolution des technologies, la longueur des séquences générées est plus longue chez le Barbarie que chez le canard commun.

2.2 Construction de la puce

2.2.1 Détection des SNP

Un total de 9,2 millions et 23,6 millions de loci polymorphes ont été identifiés chez le canard de Barbarie et le canard commun respectivement. Le nombre de loci polymorphes plus élevé chez le canard commun est dû au séquençage de populations très différenciées : le Colvert et le Rouen d'une part (environ 14 millions de SNP identifiés), et les lignées de canard Pékin d'autre part (environ 9 millions de SNP identifiés).

2.2.2 Choix des SNP

Le tableau 1 présente le nombre de SNP conservés à l'issue de chaque étape de sélection. A noter que pour l'élimination des SNP trop proches les uns des autres, les loci polymorphes des Colvert n'ont pas été pris en compte car ils auraient conduit à l'élimination d'un trop grand nombre de marqueurs du canard commun. En revanche, ces lignées ont été utilisées pour calculer les MAF et la répartition sur le génome des loci conservés.

Les listes finales de SNP soumises à Thermo Fisher comptaient 472K SNP pour le canard commun, et 446K SNP pour le canard de Barbarie. Seuls 1K SNP commun aux 2 espèces ont été identifiés dans ces 2 listes. A l'issue d'échanges avec Thermo Fisher basés sur des tests réalisés sur les amorces *in silico*, un total de 673 886 SNP seront présents sur la puce :

343 950 SNP pour le canard commun, et 331 241 SNP pour le canard de Barbarie.

CONCLUSION

L'assemblage du génome du canard de Barbarie réalisé dans ce projet avec pour référence le génome du canard commun a permis d'obtenir le 1^{er} génome de référence pour cette espèce, d'une taille totale proche d'1 Gb. La production de ce génome assemblé était un prérequis essentiel pour le choix de marqueurs pour la puce 600K. Le séquençage des populations françaises a permis la production de ressources génomiques en quantité et en qualité suffisantes pour le choix des marqueurs de la puce. Les marqueurs ont été sélectionnés afin d'assurer une bonne couverture du génome, avec des MAF variées. Intra-espèce, la proportion de SNP variables dans les populations de canard Pékin et de Barbarie ayant servi à l'identification des marqueurs devrait être d'au moins 80%. Le test de la puce prévu en 2019 permettra de confronter les performances attendues de cet outil aux résultats des 1^{ers} génotypages, et de faire avancer la réflexion sur l'intégration de la génomique dans les schémas de sélection de canards.

Remerciements

Les auteurs remercient les Régions Pays de la Loire-Bretagne et Nouvelle-Aquitaine pour le soutien financier apporté au projet, préalablement labellisé par les pôles de compétitivité Valorial et Agri-Sud-Ouest, respectivement. Les auteurs remercient également Mr. Eric Poullain (SARL du Domaine de la Feuge) pour la fourniture d'échantillons de canard Colvert. Les auteurs remercient également Eric Poullain (SARL du Domaine de la Feuge) pour la fourniture d'échantillons de canard Colvert.

REFERENCES BIBLIOGRAPHIQUES

- Boitard, S., Kofler, R., Françoise, P., Robelin, D., Schlötterer, C., Futschik, A., 2013. *Mol Ecol Resour* 13, 337–340
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. *Bioinformatics* 25, 1422–1423
- Fillon, V., Vignoles, M., Crooijmans, R.P.M.A., Groenen, M. a. M., Zoorob, R., Vignal, A., 2007. *Anim. Genet.* 38, 303–307
- Frith, M.C., 2011. *Nucleic Acids Res* 39, e23–e23
- Hamada, M., Ono, Y., Asai, K., Frith, M.C., 2017. *Bioinformatics* 33, 926–928
- Huang, Y., Li, Y., Burt, D.W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Yiqiang, Li, J., Yi, K., Feng, H., Zhu, P., Li, B., Liu, Q., Fairley, S., Magor, K.E., Du, Z., Hu, X., Goodman, L., Tafer, H., Vignal, A., Lee, T., Kim, K.-W., Sheng, Z., An, Y., Searle, S., Herrero, J., Groenen, M.A.M., Crooijmans, R.P.M.A., Faraut, T., Cai, Q., Webster, R.G., Aldridge, J.R., Warren, W.C., Bartschat, S., Kehr, S., Marz, M., Stadler, P.F., Smith, J., Kraus, R.H.S., Zhao, Yaofeng, Ren, L., Fei, J., Morisson, M., Kaiser, P., Griffin, D.K., Rao, M., Pitel, F., Wang, J., Li, N., 2013. *Nature Genetics* 45, 776–783
- Li, H., Durbin, R., 2009. *Bioinformatics* 25, 1754–1760

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. *Genome Res.* 20, 1297–1303

Figure 1. Alignement des 15 Mb du scaffold 0 du canard de Barbarie sur le génome du canard commun avant chaînage (A) et après (B). Chaque couleur correspond à un chromosome différent (chromosome 4 en bleu foncé). En abscisse : position sur le scaffold 0 en paires de bases. En ordonnée : position sur le chromosome du canard commun en paires de bases.

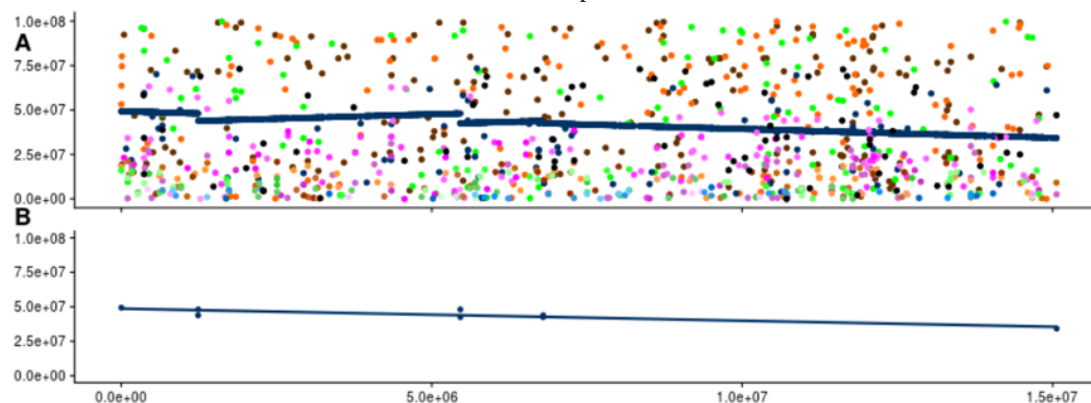


Figure 2. Histogramme de la taille des chromosomes pour la Poule (en bleu), le canard de Barbarie (en orange) et le canard commun (en vert). Les numéros des chromosomes sont indiqués, Un : chromosome unknown (séquences non assignées).

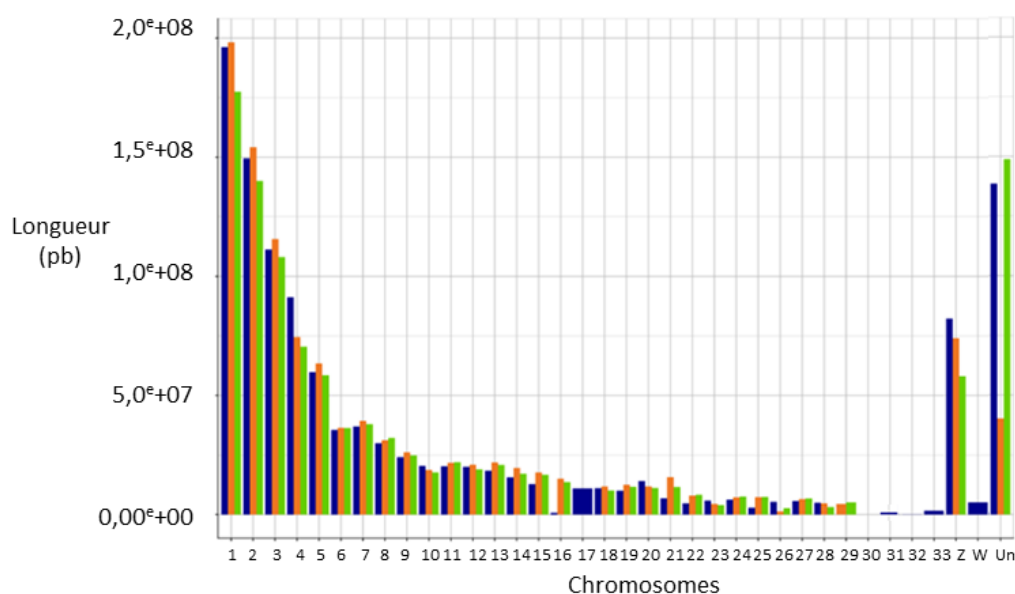


Tableau 1. Nombre de loci conservés pour la construction de la puce à chaque étape de tri.

Tri réalisé	Nombre de SNP chez le canard commun	Nombre de SNP chez le canard de Barbarie
Loci polymorphes dans au moins une population	23,6M	9,2M
Absence d'indel à au moins 50 pb	12,6M	6,2M
Suppression des SNP séparés par moins de 35 pb	8,1M	4,5M
SNP bi alléliques & profondeur de lecture comprise entre 25 et 100	8,0M	4,4M
Ségrégation dans au moins 75% des populations & MAF > 0,05	2,2M	2,0M
Alignement unique sur le génome de référence	2.1M	1.9M
Répartition sur le génome en fonction de la MAF	472K	446K