Development and validation of high-density SNP array in ducks

Marc Teissier, Noémie Thébault, Juliette Riquet, Christian Diot, Sophie Brard-Fudulea, Daniel Guemene, Bernard Alletru, Maude Cornil, Pascal Bouleau, Magali Blanchet, et al.

Development and validation of high-density SNP array in ducks

Marc Teissier[1], Noémie Thébault[2], Juliette Riquet[2], Christian Diot[3], Sophie Brard-Fudulea[1], Daniel Guémené[1], Bernard Alletru[4], Maude Cornil[4], Pascal Bouleau[4], Magali Blanchet[5], Guillaume Le Mignon[5], Olivier Demeure[5], Alain Vignal[1]

[1]SYSAAF, Centre INRA Val de Loire, 37380 Nouzilly, France; [2]GenPhySE UMR 1388, Université de Toulouse, INRA, ENVT, 31320 Castanet-Tolosan, France; [3]PEGASE UMR 1348, Agrocampus Ouest, INRA, 35590 Saint-Gilles, France; [4]Gourmaud Sélection, Groupe Orvia, 85260 Saint Andre 13 Voies, France; [5]Grimaud Frères Sélection, Groupe Grimaud, La Corbière, 49450 Roussay, France;

Corresponding author: sophie.brard-fudulea@inra.fr

Keywords: SNP, genotype, Common duck, Muscovy duck, Mule duck

Introduction

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 design of a 600K ThermoFisher SNP chip, and the validation of the array on common duck, Muscovy duck, and on their hybrid the mule duck.

Material and Methods

Design of the HD SNP chip

Sequence data for SNP discovery was generated from French duck populations including common and Muscovy ducks: several commercial lines from Grimaud Frères Sélection and Orvia Gourmaud Sélection, experimental lines from INRA, a mallard breed used as game bird, and Rouen duck. Blood samples were collected on 50 males in each population, as chromosome W is absent from reference genomes of both species. DNA was pooled by population, and each pool was sequenced with the Illumina HiSeq3000, at a 50X depth.

Sequence data generated for each pool were aligned on the corresponding reference genome (*Anas platyrhynchos* genome from Huang *et al.* 2009, and *Cairina moschata* genome from Thébault *et al*. 2019**)** using BWA algorithm (Li & Durbin 2009), alignments were sorted, and PCR duplicates were detected using Picard SortSam and MarkDuplicates. (http://broadinstitute.github.io/picard/). Quality calibration of the remaining sequence data was performed using the Genome Analysis Tool Kit guideline (McKenna *et al*. 2010). Allelic frequencies were estimated with the Python script pool-hmm.py (Boitard *et al.* 2013).

Within each species, bi-allelic polymorphic SNPs were selected only if they met the following requirements: sequencing depth of > 25X, no insertion nor deletion in the 50 bp before and after the SNP, and at least 35 bp between two consecutive SNPs. Only unique alignments on reference genomes were used. Primers of Muscovy duck SNPs and common duck SNPs were aligned to identify SNPs that could work in both species. After filtering on technical requirements, SNP were chosen based on their distribution along the genome and their minor allele frequency (MAF). First, only SNPs with a MAF ≥ 0.15 in 75% of populations and regularly distributed along the whole genome were kept; then SNPs with a MAF ≥ 0.15 in 60% of populations were added to the list, and finally the SNP list was completed with SNP which MAF was ≥ 0.05 considering all populations.

Validation of the array

Blood samples were collected in commercial lines of Muscovy duck, common duck and mule duck provided by Orvia Gourmaud Sélection and Grimaud Frères Sélection, including the commercial lines that were sequenced previously for SNP discovery. Trios of sire-dam-offspring were chosen to test for SNPs mendelian inheritance. INRA Gentyane lab performed DNA extraction and genotyping. For calibration purpose, two plates of 96 common duck and 96 Muscovy duck were genotyped first, with all lines represented on these plates. After genotyping, ThermoFisher engineers analyzed the first two plates to design library files adapted to each species. A third library containing SNPs of both species was produced ("both" library). Then, four more plates were genotyped: one plate of 96 mule ducks, and three plates with 182 common ducks and 106 Muscovy ducks.

For each species, all genotyped individuals were analyzed using Axiom Analysis Suite software (https://downloads.thermofisher.com/Axiom_Analysis_Suite_v_4.0.1_User_Guide.pdf) with the corresponding library. Samples were filtered on Dish Quality Control > 0.82 (contrast of fluorescence signal, computed on monomorphic loci) and Quality Control Call Rate > 0.90 (call rate computed on high quality SNPs identified by ThermoFisher in calibration plates), and SNPs categories were determined. Only SNPs from categories recommended by ThermoFisher were kept (PolyHighResolution, NoMinorHomozygous and MonoHighResolution). Genotypes were then exported to Plink software for analysis (Chang *et al*. 2015). SNPs were filtered on MAF > 0.05 (across

lines), HWE *p-value* > 0.001 (across lines) and call rate > 0.90 (across lines). Individuals with 10% or more missing SNPs were discarded. Mendelian inconsistencies were counted in trios: trios with more than 5% of inconsistencies were not considered for filtering of the SNPs, and SNPs with more than 10% of mendelian inconsistencies were discarded.

The same filters as those described for common duck and Muscovy duck were applied to mule duck genotypes, analyzed with the "both" library gathering all the SNPs. In addition, common duck were analyzed with the Muscovy library, and Muscovy ducks were analyzed with the common library, to check if some SNP supposed to be specific of one species could work in the other.

Result and Discussion

Design of the HD SNP chip

9.2 million and 23.6 million loci were detected for the Muscovy duck and the common duck respectively. The higher number of SNPs in the common duck is due to the sequencing of the Rouen duck and of the mallard breed used as game bird, because these two breeds are very differentiated from common duck commercial lines.

Table 1 shows the numbers of SNP kept after each filtering step. SNPs identified in the mallard game breed were not considered for filtering on the minimum distance between SNPs, because too many SNP of commercial common lines would have been discarded. The final SNP lists submitted to ThermoFisher counted 472K SNPs for the common duck and 446K SNPs for the Muscovy duck, with only 1,3 K SNP in common between the two species. This low number of shared SNPs could be due to the phylogenetic distance between the two species that was estimated between 13.5 and 20.2 million years.

After rectifications asked by ThermoFisher to ensure that a maximum of SNP primers would work, the SNP chip was composed of 673 886 SNP: 343 950 SNP for common duck, 331 241 SNPs for Muscovy duck, including 1305 SNP in common between the two species.

Table 1. Number of loci kept after each filter for the chip design

Filter	Number of SNPs in common duck	Number of SNPs in Muscovy duck
Loci polymorphic in at least one population	23,6M	9,2M
No indel in the 50 bp on each side of the SNP	12,6M	6,2M
At least 35 bp between each SNP	8,1M	4,5M
Bi-allelic SNPs & sequencing depth comprised between 25 and 100	8,0M	4,4M
Unique alignment on the adequate reference genome	2,2M	2,0M
Segregation of the SNP in at least 75% populations & MAF > 0,05	2.2M	1.9M
Spreading of the SNP along the genome according to different MAF levels.	472K	446K

Validation of the array

Among the 278 common duck samples analyzed in Axiom Analysis Suite with the common duck library, one sample had no genotyping result, 13 were discarded because of Dish QC, and four because of QC call rate. For Muscovy duck, 202 samples were genotyped. Ten samples were discarded because of Dish QC, and three because of QC call rate. The poor quality of the results of 12 common ducks and 11 Muscovy ducks was due to a pipetting error on one line on the two calibration plates.

Table 2 presents the number of SNPs in the different categories defined by Axiom Analysis Suite. When each species were analyzed with their own library, 281K and 282K SNPs were in PolyHighResolution category for the common duck and the Muscovy duck respectively. After filtering for MAF, HWE, call rate and mendelian inheritance (Table 3), 240K SNPs were kept in the Common duck, and 245K SNPs were kept in the Muscovy duck. Mule ducks were analyzed with the "both" library: 245K SNP were in PolyHighResolution category, and for more than 152K SNPs one of the homozygous genotype was not observed. After filtering for MAF, HWE, call rate and mendelian errors, 138K SNP remained for the mule duck.

Then genotypes of each species were analyzed with the library designed for the other species. The same filters as those described above were applied. Table 3 shows that 4 531 Muscovy SNPs work in the common duck, and so does 3 795 common SNPs in the Muscovy duck. These results are slightly better than what was expected, because during the array design only 1305 SNP where found to be shared by both species. These additional polyvalent SNPs may have been missed during the design phase because of an early discard in the filtering process.

Table 2. Number of SNPs in each categories for ducks analyzed with their library, and with the library of the other species. Library including the whole chip was used for mule duck

	Common duck		Muscovy duck		Mule duck
Axiom Analysis Suite category	Common library	Muscovy library	Muscovy library	Common library	„Both" library
Total	343,950	331,241	331,241	343,950	673,886
PolyHighResolution	281,385	5,954	282,861	4,645	245,513
NoMinorHomozygous	2,639	14,368	1,181	34,863	152,121
MonoHighResolution	245	144,082	228	143,613	11,901
Not recommended	59,681	166,837	46,971	160,829	264,351

Table 3. Number of SNPs kept after filters on MAF, HWE, call rate and mendelian inheritance. Only SNPs from Axiom Analysis Suite recommended categories were considered

	Common duck		Muscovy duck		Mule duck
Filter	Common library	Muscovy library	Muscovy library	Common library	„Both" library
Recommended SNPs	284,269	164,404	284,270	183,121	409,535
MAF	282,428	5,571	283,878	4,333	358,638
HWE	240,674	4,536	245,365	3,795	138,498
Missingness per SNP	240,674	4,536	245,365	3,795	138,498
Mendelian inconsistencies	240,656	4,531	245,192	3,795	138,498

Conclusion

The design and validation of a 670K HD SNP chip for ducks is the main outcome of the CanArray project. 84% of SNPs have high quality genotyping results according to Thermo Fisher standards, and a few thousands SNPs are polymorphic in both species and in their hybrid. The development of this tool is a major step for French breeding companies towards the use of genomics in their breeding programs.

Acknowledgments

Authors thank Région Pays de la Loire-Bretagne, Région Nouvelle Aquitaine, and AgenAvi for financial support of the CanArray project; competitiveness clusters Valorial and Agri-Sud Ouest for labelling of the project. Authors also thank Mr. Eric Poullain (SARL du Domaine de la Feuge) for providing Colvert samples.

BOITARD, S., KOFLER, R., FRANÇOISE, P., ROBELIN, D., SCHLÖTTERER, C. and FUTSCHIK, A. (2013) Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Molecular Ecology Resources* **13**: 337–340.

CHANG, C.C., CHOW, C.C., TELLIER, L.C., VATTIKUTI, S., PURCELL, S.M. and LEE, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4.

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. and LI, N. (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genetics* **45**: 776–783.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. and DEPRISTO, M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. **20**: 1297–1303.

THEBAULT, N., RIQUET, J., DIOT, C., BRARD-FUDULEA, S., GUEMENE, D., ALLETRU, B., CORNIL, M., BOULEAU, P., BLANCHET, M., LA MIGNON, G., DEMEURE, O. and VIGNAL, A. (2019) Développement d'une puce de génotypage haute-densité 600K pour le canard commun et le canard de Barbarie. *13èmes Journées de la Recherche Avicole et Palmipèdes à Foie Gras*, Tours, pp. 80-84.