Derivation of parentage and breed-of-origin of alleles in a crossbred broiler dataset

M.P.L. Calus¹, J. Vandenplas¹, B. Hulsegge¹, R. Borg², J.M. Henshall², R. Hawken²

¹ Wageningen University & Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, The Netherlands
mario.calus@wur.nl (Corresponding Author)
² Cobb-vantress Inc., Siloam Springs, AR 72761-1030, USA

Summary

Pig and poultry breeding programs rely on crossbreeding. With genomic selection, widespread use of crossbred performance in breeding programs comes within reach. Commercial crossbreds, however, may have unknown pedigrees and their genomes include DNA from two to four different breeds, depending on the crossbreeding scheme. SNP information allows: 1) derivation of parentage, provided that genotypes of parents are available, and 2) derivation of breed-of-origin of alleles in crossbreds, provided that sufficient genotypes of purebred animals are available to determine frequencies of segregating haplotypes for each of the parental breeds. We derived both parentage and breed-of-origin of alleles in a broiler dataset that comprised 5882 purebred and 10,943 three-way crossbred offspring that were generated by natural mating of 164 purebred sires to 660 purebred and 1031 F1 crossbred hens. Numbers of offspring per sire had a very skewed distribution, ranging from 1 to 275 crossbreds and 1 to 155 purebreds. Breed-of-origin could be derived for 99.74% of the alleles of the 1031 F1 crossbred hens and for 98.10% of the alleles of the 10,943 three-way crossbred offspring. Visual inspection of the assigned breed-of-origin, however, suggested that there are some errors in assignment of the maternal alleles. Further tuning of the algorithm, or adding more purebred animals of the dam lines to the analysis, may help to resolve those errors. The achieved percentage of assignment to the sire line appears sufficient to proceed with subsequent analyses requiring only the breed-of-origin of the paternal alleles to be known.

Keywords: parentage assignment, breed-origin of alleles, crossbreds, broilers

Introduction

Most poultry and pig breeding programs aim at improving crossbred (CB) performance, by performing selection in the purebred (PB) parental lines, based on PB performance. Basing selection on CB performance may ultimately yield higher genetic gain, but requires using phenotypic information of CB animals in the breeding program, which is tedious or even impossible using pedigree recording. The recent availability of cost-efficient high-throughput genotyping enables genotyping CB animals, which improves the accuracy of linking CB information back to PB selection candidates by using genomic prediction.

Validating the expected accuracy of genomic prediction directly using CB performance, requires an accurate estimate of the purebred-crossbred correlation, and a sufficiently large
dataset to empirically compute prediction accuracies. As a first step to characterize a dataset generated for these purposes, the objectives of this study were to investigate the distribution of the number of offspring per sire and the efficiency of deriving breed-of-origin of alleles in a CB broiler dataset.

Material and methods

The data used in this study was generated in an experimental set-up of three generations including line A males, that had both line A purebred offspring and A(BC) three-way CB offspring (Figure 1). In addition to this key set of individuals, the data included line A dams and grandparents, as well as BC dams, B granddams and C grandsires. All animals in the parent and offspring generation were generated using natural mating. No pedigree information was available for any of the animals. All animals were genotyped, and the sex of the animals was derived from the SNP chip information.

Genotype and pedigree data

The data included 56,557 SNPs and 22,346 animals, after usual edits for call rate per animal (>95%) and SNPs (>90%), and after removing mitochondrial SNPs or SNPs located on the sex chromosomes. During the parentage derivation, inconsistent genotypes between derived parent-offspring pairs were set to missing. A total of 416 SNPs were removed because they had more than 1% inconsistencies between offspring and derived parents. After the parentage derivation, any missing genotypes were imputed using FImpute (Sargolzaei et al., 2014).

Derivation of parentage

Parentage was derived using a homemade algorithm that was implemented in Fortran 90, using parallel computing. The algorithm is based on the principle that the number of opposing homozygotes (#OH) between an individual and its parents is expected to be zero, or slightly higher to accommodate for genotyping errors. Required input data includes a genotype file, and optionally: the sex of all the animals (in this case derived from the SNP chip information), a list of target animals for which the parentage needs to be derived, and a list of candidate parents. The implemented algorithm involves a number of steps. Genotypes were initially recoded as 0, 1 and 2. All 1’s were replaced by 5, which coded for missing genotypes. The #OH between any pair of individuals, avoiding comparing an individual with itself, was then simply computed by counting the number of SNPs where the sum of their genotypes was 2. Candidate parents were sorted within individual based on increasing #OH. Per animal, those candidate parents with #OH less than 1%, a threshold empirically derived from the data, were retained. From this shortlist, considering their sex, all possible pairs of male and female parents were evaluated using a trio approach. In this approach, when one of the three animals had a missing genotype (coded as 5), the genotypes of all three animals were replaced by a 1, such that no mismatch was induced in the next step. Then, for each locus the genotype of the offspring was subtracted from the sum of the genotypes of the parent pair. When the three genotypes match together, the result should be 0, 1 or 2, because the gametes “unused” by the offspring could give rise to another possible offspring genotype. If there was only one parent pair for which less than 1% of the SNPs failed this test, then both parents were assigned. If multiple parent pairs matched with the genotypes of the offspring, then no parents were assigned for this animal.
Derivation of breed-of-origin of alleles

For the A(BC) offspring, we aimed to derive the breed-of-origin of alleles for each of their alleles, using the BOA (assigning Breed-of-origin of alleles) approach (Sevillano et al., 2016; Vandenplas et al., 2016). The BOA approach involves: i) simultaneously phasing genotypes of PB and CB animals using AlphaPhase1.1 and making use of the (derived) pedigree information (Hickey et al., 2011); ii) building haplotype libraries for each breed using phased haplotypes of PB animals; and iii) assigning breed-of-origin of alleles of CB animals based on their phased haplotypes, the frequencies of those haplotypes in each of the breeds, and the breed composition of CB animals (i.e. A(BC)). Percentages of allele assignments per breed were compared against the (expected) breed composition of the CB animals. The data used for the BOA analysis included the A(BC) offspring and their ancestors, and any additional PB animals that were available in the data but not assigned to be parents or grandparents of the CB animals. The total numbers of PB animals included in the analyses, were 8167 for line A (including the PB offspring), 353 for line B and 719 for line C.

Results

A principal component analysis on the genomic relationship matrix (computed as the first method of VanRaden (2008)) showed that the three breeds are clearly separated (Figure 1). The BC dams were located in between the B and C breed, while the A(BC) animals were located in between their A and BC parents.

Derivation of parentage involved one analysis that used 8 threads and took less than 5m wall clock time. For the analysis two groups of animals were defined: 1) animals for which parentage was required, and 2) all possible parents and grandparents. The first group included all A purebred and A(BC) offspring, and all A and BC animals from the parental generation. The second group included all A and BC animals from the parental generation, and all A, B and C animals from the grandparental generation. Full parentage could be derived for 16,825 animals from the final offspring generation. In total, 164 purebred sires sired the entire offspring generation, and 660 purebred and 1031 F1 hens were identified as dams. Numbers of offspring per sire ranged from 1 to 275 CB and 1 to 155 PB (distributions are shown in Figure 3), and per dam from 1 to 37 CB and 1 to 32 PB (distributions not shown). Of the 16,825 offspring with both parents identified, 77.5, 96.5, 78.0 and 93.3% of their, respectively, paternal grandsires, paternal granddams, maternal grandsires, and maternal granddams were assigned by deriving parentage of the parental generation. The remaining grandparents were not present in the genotype data, but were derived from the relationships between offspring with (some) unknown grandparents (results not shown).

The BOA analysis was performed for GGA7 for the 1031 F1 BC hens and the 10,943 A(BC) animals for which both parents were derived from the data (Table 1). For the BC hens, on average 49.87% of their alleles were assigned to both lines B and C, yielding a total assignment of 99.74%. For the A(BC) animals, on average 49.99% of their alleles was assigned to line A, 24.15% to line B and 23.96% to line C.

Table 1. Average breed composition and percentage of alleles of the BC and A(BC) animals assigned to any of the three breeds.

<table>
<thead>
<tr>
<th>(Average) breed composition</th>
<th>Breed assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


### Discussion

The derivation of parentage showed that the distribution of the number of offspring of the sires was very skewed. There were sires that were allowed to breed but did not have any offspring, while the sire with the largest number of offspring had 275 CB and 155 PB offspring. This result suggests that in practical breeding programs with natural mating, the contribution of males to the generated offspring is highly variable. Our results also imply that natural mating in the current set-up can reduce the power to compute the pure-bred crossbred correlation, because the data is not balanced in terms of the numbers of offspring per sire.

The BOA approach enabled to assign 99.74% of the alleles in the BC hens and 98.10% in the A(BC) animals to the lines of origin. These results are better than earlier results obtained in three-way CB animals in simulated data (Vandenplas et al., 2016), and three-way CB pigs in empirical analyses (Sevillano et al., 2016). Visual inspection of the assigned breed-of-origin suggests, however, that for the dam lines there may be some errors (Figure 4). This is observed from a number of relatively short haplotypes that are assigned to one of the dam lines in the middle of a haplotype of the other dam line, resulting in more recombination events on a chromosome than expected. Focussing more on longer haplotypes removed some of those probably incorrect recombinations, but also especially reduced the total percentage of assignments for the A(BC) animals, as their percentage assignments to the B and C lines dropped below 20%. These results indicate that further tuning of the BOA procedure is necessary, or that more PB animals from the dam line should be added to the data, to increase the accuracy of assigning breed-of-origin of maternal alleles of the A(BC) animals in this particular dataset. The achieved percentage of assignment to the sire line appears sufficient to proceed with subsequent analyses requiring only the breed-of-origin of the paternal alleles to be known.

### Acknowledgements

Financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership “Breed4Food” code BO-22.04-011-001-ASG-LR-3) is acknowledged. Cobb Europe is gratefully acknowledged for making population, phenotypic and genotypic data available.

### List of References


Figure 1. Design of the experimental setup. A, B and C are the parental lines, including grandparents (GP) and parents (P).

Figure 2. PCA (principal component analysis) plot of the genotype data.
Figure 3. Number of PB (A) and CB (B) offspring per sire.

Figure 4. Breed-origin of alleles for 20 random A(BC) animals. Line A, B and C alleles are indicated in green, blue and red, respectively.