Assessment of population genetic diversity accounting for genomic relationships of founders that have unknown pedigree relationships

A. Arakawa, M. Taniguchi & S. Mikawa

Institute of Livestock and Grassland Science, National Agriculture and Food Research Organization, 2 Ikenodai, Tsukuba, 305-0901, Ibaraki, Japan
aisaku@affrc.go.jp (Corresponding Author)

Summary

Genetic relatedness is a fundamental concept in animal breeding and conservation genetics. Typically, pedigree information is available for calculating an additive relationship () matrix, while ancestors of founder individuals are assumed unrelated. Recently, genome-wide dense SNP markers have been available for measuring genetic similarity between all pairs of individuals, to construct a so-called matrix, and moreover, a hybrid () matrix of the and matrices as proposed for single-step GBLUP. A concept of metafounders has been advocated for constructing founder genetic relationships. Our objective was to investigate relationships between a matrix using all individuals in pedigree with , matrices and an matrix with metafounder (). We generated data having different two scenarios regarding relatedness among founders; linkage equilibrium (LE) and linkage disequilibrium (LD). Correlations of the matrix of whole individuals with the or matrices were better in the LE scenario than those in the LD scenario. On the other hand, in the LD scenario, regression coefficients of elements of on elements in or were less than the desired value of 1, and intercepts were higher than 0. In both scenarios, constructed using a small number of genotyped individuals departed from the matrix, whereas, that using a moderate number of individuals with the genotypes resulted in better adjusted matrices relative to .

Keywords: conservation, genetic diversity, endangered local breeds, & metafounder

Introduction

Genetic diversity is an important parameter for conservation of endangered local livestock populations. Local livestock breeds are typically kept in small farms, and therefore, genetic diversities for these breeds are expected to be very low, leading to reduction in reproductive performance. However, pedigree information is not available in many cases, because these small farms rarely record family information of all offspring. Therefore, expected genetic diversity estimated by pedigree information is often much higher than realized genetic diversity estimated by genomic information.

Several methods for assessing the genetic diversity have been developed based on using either pedigree or molecular information (Lacy, 1995; Oliehoek *et al*., 2006). The methods based on pedigree information use relatedness of individuals calculated using an additive relationship () matrix, whereas the methods of molecular information are based on genetic similarity calculated using DNA markers (Lynch & Ritland, 1999; Oliehoek *et al*., 2006). Recently, genome-wide dense single nucleotide polymorphism (SNP) markers have been used for evaluating genetic diversity (VanRaden *et al*., 2011; Toro *et al*., 2011; Gómez-Romano *et al*., 2013; de Cara *et al*., 2013). Furthermore, Legarra *et al*., (2009) proposed a hybrid () matrix for a single-step GBLUP (ssGBLUP): the matrix incorporates a genomic relationship () matrix into the matrix. However, genomic and pedigree relationships may not be compatible with each other, because pedigree information assumes the founders in base population to be unrelated, whereas genomic information shows varying relatedness among founder animals. Therefore, Legarra *et al*., (2015) presented the concept of metafounders to account for relatedness and inbreeding among founders in the base population.

In our study, we investigated relationships between the matrix using all individuals in pedigree with the , matrices and matrix with metafounder ().

Material and methods

Methods

We followed an expression of Legarra *et al*., (2009). The matrix partitioned in genotyped and ungenotyped individuals and the matrices are

, and ,

respectively, where and comprise elements for genotyped and ungenotyped individuals, respectively, **()** comprise elements relating genotyped (ungenotyped) and ungenotyped (genotyped) individuals, and is a genomic relationship matrix which is expressed as , where is a matrix of 1s and is half number of SNP.

Legarra *et al*., (2015) proposed a metafounder concept, in which pseudo-individuals describe relationships within the base population of pedigree. The and the matrices under the metafounder concept are

, and **,**

respectively, where is an ancestral relationship. According to Garcia-Baccino *et al*., (2017), was estimated using a generalized least squares method.

Simulation Scheme

In order to compare among these genetic relationship matrices, a dataset was generated by QMSim software (Sargolzaei & Schenkel, 2009). The base population consisted of 5 sires and 20 dams, with each sire mated randomly to 4 dams, and each dam produced 5 progenies. 10 discrete generations after the base population were carried out. Each individual had ten chromosomes of 100 cM length, and each chromosome had 1,000 biallelic markers that were randomly distributed, so a total of 10,000 markers was generated. We assumed two different scenarios, where markers in the base generation were assumed to be in either linkage equilibrium (LE) or linkage disequilibrium (LD). For the LE situation, alleles in the base population were sampled from a uniform distribution with equal frequencies. For the LD scenario, a historical population was started with an effective population size of 400, and the effective population size gradually decreased to the size of 40 during 100 discrete generations.

For construction of the and matrices, the SNP data in the last generation (G10) was assumed to be completely obtained from all individuals and randomly sampled from 0, 10, 50 or 100 individuals from generations 1 to 9 except for the base population. Correlations between inbreeding coefficients based on the matrix constructed using genotypes of all individuals in pedigree () and based on the other relationship matrices and regressions of on these were calculated. Moreover, we calculated correlations and regressions of elements of the matrix with elements of the other. Ten replicates were analysed.

Results and Discussion

Table 1 shows the comparisons of the inbreeding coefficients in the LE and LD situations. Correlations between and based on pedigree in the LE and LD were 0.83 ± 0.02 and 0.72 ± 0.05, respectively, suggesting that existence of ancestral relationship among the founders caused the reduction in the similarity between and pedigree . In the LE scenario, the correlations of the and matrices increased with an increase in the number of individuals with genotyping, and the regression coefficients were more than 0.9. However, in the LD scenario, the correlations of the matrices decreased with increasing the number of the genotyping individuals. Moreover, intercepts and coefficients for regression of the on based on the matrices departed from desired values of 0 and 1, respectively, whereas, for based on matrices with more than 150 genotyped individuals, the results of the regressions were better, suggesting that matrix could not completely explain inbreeding degrees within the founder population.

Correlations of the with the other relationship matrices, and coefficients and intercepts for regression of the on these relationships are shown in Table 2. In the LE scenario, the correlations between the and the other matrices were more than 0.9, except for using genotypes of only generation 10, whereas, the correlations in the LD scenario were lower than these in the LE scenario and ranged from 0.80 to 0.89. When the LE scenario, coefficients and intercepts for regression of the matrix on the or any matrices were the desired value of 1.0 and 0, respectively, whereas, when the LD scenario, the coefficients of regression were lower than these in the LE, ranging from 0.76 to 0.47. In the matrices, when using the samples of only G10 or G10 plus random 10 individuals, the coefficients of regression were higher than 1.0 and the intercepts were lower than 0 in the both scenarios. When the genotype information of G10 plus more than 50 individuals were used, these biases were eliminated.

Previous studies showed that genomic inbreeding based on a matrix using a mean allele frequency of 0.5 correlated with pedigree inbreeding (VanRaden *et al*., 2011; Marras *et al*., 2014) and inbreeding calculated using runs of homozygosity (Bjelland *et al*., 2013; Marras *et al*., 2014). Inbreeding based on a matrix includes information of both ancient and recent relatedness, whereas, pedigree inbreeding captures only recent relatedness. In endangered breeds, founders within the base population in pedigree are generally related with each other, although their pedigree is unknown. Therefore, genomic inbreeding is likely to be better for assessing their genetic diversity than pedigree inbreeding. However, it is not feasible to obtain all DNA samples of founder animals. Therefore, the matrix is an attractive approach for evaluating degree of genetic diversity for endangered livestock breeds. However, in the case of the existence of relatedness among founders in the base population, the matrix will depart from the matrix using all individuals (Tables 1 and 2), because the matrix cannot explain relationships of ancestors of founders. The concept of metafounders can explain a certain amount of relatedness among founders in the base population. In our study, if DNA samples are obtained from approximately 10% of the population, the matrix were more similar with the matrix than either the or matrices. Further research using simulation and real data are needed to confirm this approach gives better results for constructing conservation programs of endangered breeds.

Acknowledgements

We thank Andres Legarra for his valuable comments. This work was partially supported by the program for the promotion of Agu pork from Okinawa prefecture.

List of References

Bjelland D.W., K.A. Weigel, N. Vukasinovic & J.D. Nkrumah, 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *J. Dairy Sci*. **96**: 4697-4706.

de Cara, M.Á.R., B. Villanueva, M.A. Toro & J. Fernández, 2013. Using genomic tools to maintain diversity and fitness in conservation programmes. *Mol. Ecol*. **22**: 6091-6099.

Garcia-Baccino, C.A., A. Legarra, O.F. Christensen, I. Misztal, I. Pocrnic, Z.G. Vitezica & R.J.C. Cantet, 2017. Metafounders are related to F_{st} fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol*. **49**: 34.

Gómez-Romano F., B. Villanueva, M.Á.R. de Cara & J. Fernández, 2013. Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. *Genet. Sel. Evol*. **45**: 38.

Lacy, R.C., 1995. Clarification of genetic terms and their use in the management of captive populations. *Zoo Biol.* **8**: 565-578.

Legarra, A., I. Aguilar & I. Misztal, 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci*. **92**: 4656-4663.

Legarra, A., O.F. Christensen, Z.G. Vitezica, I. Aguilar & I. Misztal, 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* **200**: 455-468.

Lynch, M. & K. Ritland, 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1735-1766.

Marras G., G. Gaspa, S. Sorbolini, C. Dimauro, P. Ajmone-Marsan, A. Valentini, J.L. Williams & N.P.P Macciotta, 2014. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Anim. Genet*. **46**: 110-121.

Oliehoek, P.A., J.J. Windig, J.A.M. van Arendonk & P. Bijima, 2006. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**: 483-496.

Sargolzaei, M. & F.S. Schenkel, 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**: 680-681.

Toro, M.A., L.A. García-Cortés & A. Legarra, 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet. Sel. Evol*. **43**: 27.

VanRaden, P.M., K.M. Olsen, G.R. Wiggans, J.B. Cole & M.E. Tooker, 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci*. **94**: 5673-5682.

Table 1. Results of correlations between inbreeding coefficient () of genomic relationship matrix () and of additive (), hybrid () or the matrices with metafounder (), and regressions of on the other inbreedings.

Matrix[1]	Linkage equilibrium			Linkage disequilibrium		
	Correlation	Regression[2]	Intercept[2]	Correlation	Regression[2]	Intercept[2]
	0.83 ± 0.02	1.02 ± 0.06	0.00 ± 0.00	0.73 ± 0.05	0.76 ± 0.08	0.23 ± 0.01
	0.86 ± 0.02	0.99 ± 0.04	0.00 ± 0.00	0.71 ± 0.05	0.42 ± 0.04	0.22 ± 0.01
	0.86 ± 0.02	0.99 ± 0.03	0.00 ± 0.00	0.68 ± 0.06	0.40 ± 0.04	0.20 ± 0.01
	0.87 ± 0.01	0.97 ± 0.03	0.00 ± 0.00	0.64 ± 0.06	0.41 ± 0.05	0.17 ± 0.01
	0.88 ± 0.01	0.97 ± 0.02	0.00 ± 0.00	0.64 ± 0.05	0.45 ± 0.06	0.15 ± 0.02
	0.56 ± 0.07	0.92 ± 0.11	0.01 ± 0.02	0.63 ± 0.03	1.00 ± 0.09	0.04 ± 0.04
	0.79 ± 0.03	1.04 ± 0.04	0.02 ± 0.01	0.74 ± 0.04	0.94 ± 0.05	0.06 ± 0.01
	0.85 ± 0.02	0.98 ± 0.03	0.03 ± 0.00	0.78 ± 0.03	0.91 ± 0.04	0.05 ± 0.01
	0.87 ± 0.02	0.96 ± 0.02	0.02 ± 0.00	0.80 ± 0.03	0.91 ± 0.04	0.04 ± 0.01

[1] Subscripts indicate SNP sampling schemes; G10 expresses the sampling only individuals in generation 10, and G10 plus 10, 50 or 100 express the samples of G10 plus additional 10, 50 or 100 individuals, respectively, which were randomly sampled from generations 1 to 9 except for the base population.

[2] , where is a regression coefficient, and is an intercept.

Table 2. Results of correlations between genomic relationship matrix () and additive (), hybrid () or the matrices with the metafounder (), and the regression of the matrix on the other matrices.

Matrix[1]	Linkage equilibrium			Linkage disequilibrium		
	Correlation	Regression[2]	Intercept[2]	Correlation	Regression[2]	Intercept[2]
	0.92 ± 0.01	1.02 ± 0.06	0.00 ± 0.00	0.85 ± 0.03	0.76 ± 0.08	0.47 ± 0.02
	0.92 ± 0.01	1.00 ± 0.02	0.00 ± 0.00	0.83 ± 0.04	0.49 ± 0.03	0.47 ± 0.02
	0.93 ± 0.01	1.00 ± 0.01	0.00 ± 0.00	0.81 ± 0.04	0.47 ± 0.04	0.43 ± 0.02
	0.93 ± 0.01	1.00 ± 0.01	0.00 ± 0.00	0.80 ± 0.04	0.51 ± 0.05	0.36 ± 0.02
	0.94 ± 0.01	1.00 ± 0.00	0.00 ± 0.00	0.81 ± 0.04	0.57 ± 0.05	0.31 ± 0.02
	0.81 ± 0.02	1.66 ± 0.10	-0.48 ± 0.09	0.81 ± 0.02	1.55 ± 0.18	-0.54 ± 0.20
	0.91 ± 0.01	1.26 ± 0.03	-0.14 ± 0.02	0.86 ± 0.03	1.14 ± 0.06	-0.12 ± 0.05
	0.93 ± 0.01	1.08 ± 0.01	-0.03 ± 0.00	0.88 ± 0.02	1.03 ± 0.03	-0.02 ± 0.02
	0.94 ± 0.01	1.04 ± 0.01	-0.02 ± 0.00	0.89 ± 0.02	1.01 ± 0.02	-0.01 ± 0.01

[1] Subscripts indicate SNP sampling schemes; G10 expresses the sampling only individuals in generation 10, and G10 plus 10, 50 or 100 express the samples of G10 plus additional 10, 50 or 100 individuals, respectively, which were randomly sampled from generations 1 to 9 except for the base population.

[2] , where is a regression coefficient, is an intercept, is the relationship matrix (, or), and is a vectorization operation.