Limited dimensionality of genomic information and effective population size

I. Pocrnic[1], D.A.L. Lourenco[1], Y. Masuda[1], A. Legarra[2] & I. Misztal[1]

[1] University of Georgia, Department of Animal and Dairy Science, 30602 Athens, GA, USA
ipocrnic@uga.edu (Corresponding Author)
[2] Institut National de la Recherche Agronomique, GenPhySE, 31326 Castanet-Tolosan, France

Summary

Past studies suggested limited dimensionality of the genomic SNP information was related to effective population size (Ne). The objective of this study was to estimate that dimensionality with simulated and with livestock (Holstein, Jersey, Angus, pigs, and chicken) data sets. That dimensionality can be defined as the number of non-negligible singular values of gene content, or the number of non-negligible eigenvalues of genomic relationship matrix (GRM). In this study, eigenvalue analysis determined the numbers of largest eigenvalues corresponding to 90, 95, 98, and 99% of the variation of the GRM for each population. With many genotyped animals and SNP markers, the numbers corresponding to 90, 95, and 98% approached NeL, 2NeL and 4NeL, respectively, where L is genome length in Morgans. Realized accuracies were calculated for single-step GBLUP (ssGBLUP) with an algorithm for inversion of GRM that takes into account the limited dimensionality of the genomic information. Realized accuracies peaked with the dimensionality corresponding to 98 to 99% of variation depending on population, indicating that 1 to 2% of variation in the GRM was due to noise. However, the accuracies were only slightly reduced at half the optimum dimensionality. Subsequently, the dimensionality of the genomic information was estimated at about 14,000 for Holstein and Angus cattle, 12,000 for Jersey cattle, and 6000 for pigs and chickens, which corresponds approximately to 3NeL. Based on interpolation of simulated and real data with L of 30 Morgans, approximate Ne was 149 for Holsteins, 101 for Jerseys, 113 for Angus, and 44 for chickens; for pigs and L of 20 Morgans, approximate Ne was 48. Limited dimensionality of the genomic information has serious implications for genomic prediction and possibly GWAS.

Keywords: effective population size, genomic recursion, genomic relationship matrix, single-step GBLUP

Introduction

Genomic selection is a widely applied methodology, based on usage of SNP chips with the preferred density about 50k, and constant intention to use greater densities for achieving greater accuracies. Because inheritance of genomic material is in blocks of DNA, one of the key variables in predicting the theoretical accuracies are effective number of chromosome segments (Me), that are dependent on effective population size (Ne) and genome length in Morgan (L). The Me distribution is given by Stam (1980), and revisited several times, e.g. in Goddard (2009). Brard and Ricard (2015) reviewed the Me formulae and accuracy, and concluded that formulae are inconsistent and not appropriate.

Increasing the SNP density did not significantly increase accuracy, and pointed to conclusion that genomic information in a population is contained in the limited number of Me

or effective SNP markers (ESM) (Misztal, 2016). Based on the theory behind Me and ESM, genomic information should be a function of Ne, and Ne can be derived through the eigenvalue analysis of GRM.

The objective of the study was to test the limited dimensionality of genomic information for livestock (Holstein, Jersey, and Angus cattle, pigs, and chicken), and for idealized, simulated populations. Secondly, eigenvalue profiles of GRM and their distributions were analysed to establish a connection between accuracy and Ne.

Material and methods

Data

The simulated and livestock datasets were described in detail in previous studies (Pocrnic et al., 2016a & 2016b), with relevant information summarised in Table 1. Six populations with Ne ranging from 20 to 200 were simulated using QMSim (Sargolazaei & Schenkel, 2009). Each of the populations consisted of 10 recent, non-overlapping generations undergoing random mating without selection, and phenotypic records for all animals from generations 1 through 9. Generations 8 to 10 were genotyped, and the genome consisted of 30, 1-M long chromosomes with about 5,000 randomly distributed biallelic QTLs. Livestock datasets were provided by Holstein Association USA, AGIL ARS (USDA), American Angus Association, Genus PIC, and Cobb-Vantress.

Table 1. Summary of data for each specie/breed and simulated data.

Species/Breed	Phenotypes	Animals in pedigree	Genotypes
Holstein	11.6 M	10.7 M	77 k
Jersey	4.2 M	2.4 M	75 k
Angus	6 M	8.2 M	81 k
Pig	400 k	2.4 M	23 k
Chicken	197 k	199 k	16 k
Simulation	225 k	250 k	75 k

Models and computations

For each dataset, the GRM was created as in VanRaden (2008). Then, the number of the largest eigenvalues of explaining 90, 95, 98, or 99% of the variation was determined for each GRM. Genomic evaluations used ssGBLUP (Aguilar et al., 2010) with the Algorithm for Proven and Young inverse of the GRM (Misztal et al., 2014; Masuda et al., 2016); the number of core animals was equal to the number of eigenvalues as computed previously. Accuracies of prediction were calculated as correlations between true and estimated breeding values (simulated data), forward prediction or predictability. The dimensionality for each data set was defined based on the number of core animals that maximized accuracy.

Results and discussion

Number of eigenvalues and effective population size

The number of eigenvalues that explain 90, 95, 98, and 99% of variance in the simulated

populations is given in Figure 1A. For smaller Ne, the numbers for the 90, 95, and 98% approximately equal NeL, 2NeL, and 4NeL, respectively. For larger Ne, the curves are depressed. In general, the dimensionality of GRM cannot be larger than the minimum number of genotyped individuals and the number of SNPs. Distribution of eigenvalues for the livestock datasets is shown in Figure 1B and was compared with the simulated populations. Numbers of eigenvalues for both real and simulated datasets are shown in Table 2. The rank of the GRM is limited by the number of SNPs and genotyped animals in the study, and should be at least 12 times larger than the number of segments (MacLeod et al., 2005). If we assume this was true for number of eigenvalues at 90%, Ne can be estimated by interpolation of real and simulated data, with estimates shown in Table 2.

Table 2. Estimated and simulated effective population size (Ne), number of eigenvalues explaining 98%, and optimal and actual chip density.

Ne	Dimensionality at 98%	Optimal chip density[1]	Actual chip used
20[2]	3.7 k	44 k	50 k
48[3] (Pig)	4.1 k	49 k	37 k
40[2]	6.2 k	74 k	50 k
44[4] (Chicken)	4.2 k	50 k	39 k
80[2]	9.6 k	115 k	50 k
101[4] (Jersey)	11.5 k	138 k	61 k
113[4] (Angus)	10.6 k	127 k	38 k
120[2]	12.2 k	146 k	50 k
149[4] (Holstein)	14 k	168 k	61 k
160[2]	14.1 k	169 k	50 k
200[2]	15.5 k	186 k	50 k

[1] Defined as 12 times number of segments (MacLeod et al., 2005)
[2] Simulated effective population size
[3] Estimated value for effective population size based on genome length of 20 Morgan
[4] Estimated value for effective population size based on genome length of 30 Morgan

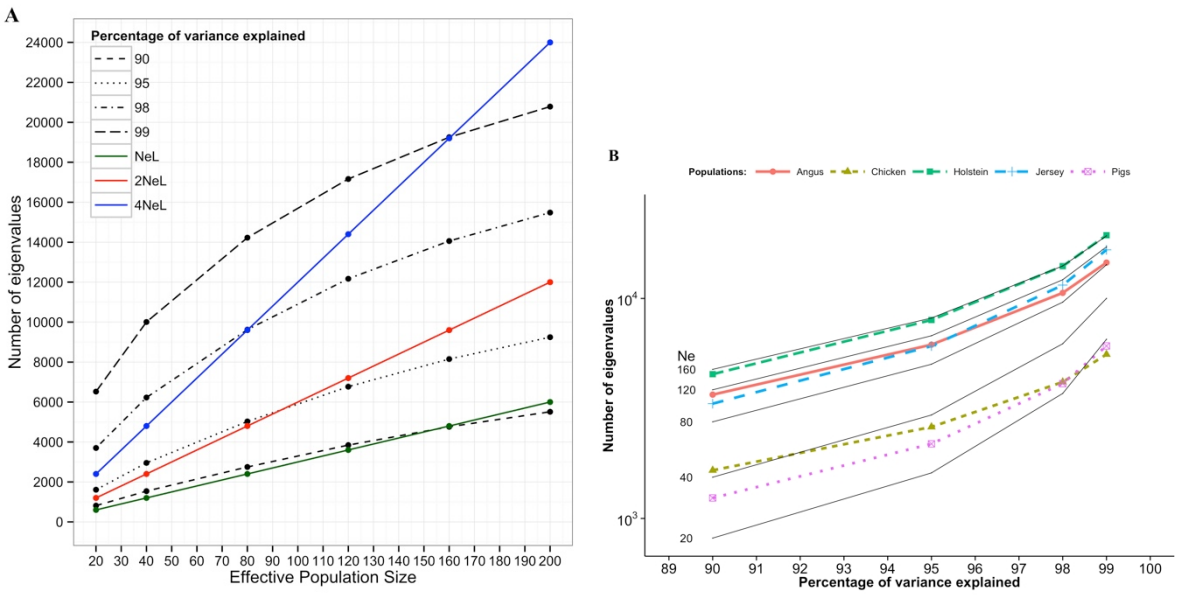

Figure 1. Number of largest eigenvalues that explain 90, 95, 98, and 99% of the variation in the genomic relationship matrix. A: simulated populations with different effective population sizes (Ne). Additional lines show NeL, 2NeL and 4NeL, where L=30 Morgans; B: for chickens, pigs, Angus, Jersey, Holstein and simulated populations (solid lines).

Dimensionality that maximizes accuracy of predictions

Figure 2A shows accuracies for simulated populations as a function of the number of eigenvalues needed to explain a given variance. Populations with smaller Ne, have fewer Me or ESM to estimate, smaller prediction error variance, and thus greater accuracy. For all Ne, the accuracy was maximized by assuming the dimensionality corresponding to 98% of the variation. Differences from 90 to 98% were small. Similar results were confirmed with the livestock datasets, as shown on Figure 2B for Jersey cattle. Results for other species/breeds were shown by Pocrnic et al. (2016b).

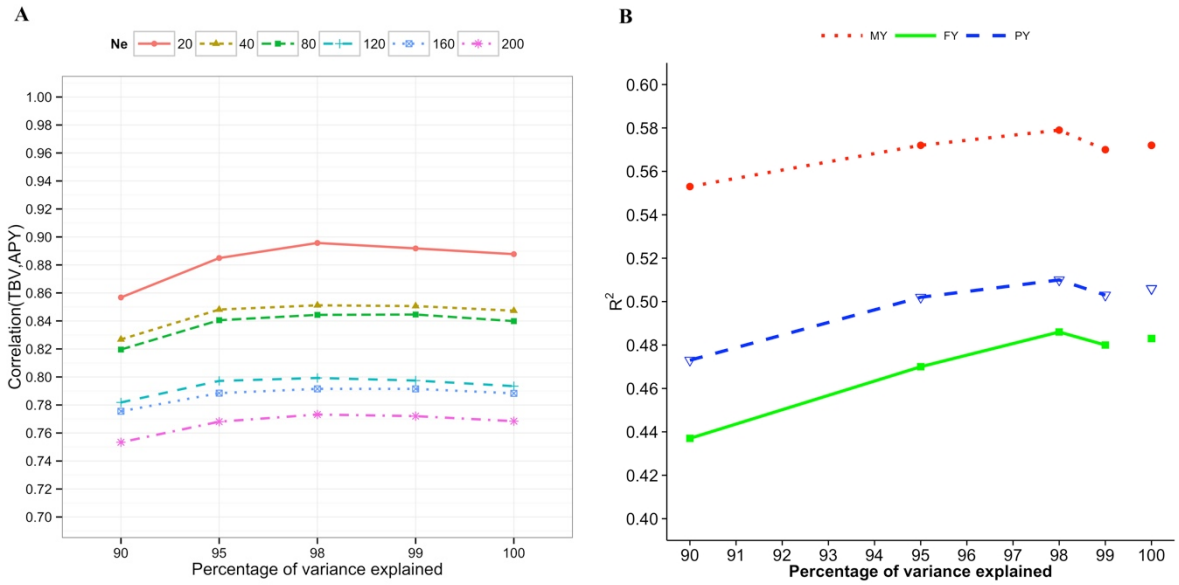

Figure 2. Accuracies of genomic estimated breeding values. A: for simulated populations where number of core animals was defined as the number of eigenvalues that explain 90, 95, 98, and 99% of the variation in the genomic relationship matrix (GRM); B: for 305-day milk yield (MY), fat yield (FY) and protein yield (PY) of Jersey cattle. Values for 100% correspond to the regular inverse of the GRM.

Conclusion

With many animals and SNPs, the dimensionality of the genomic information is approximately the number of eigenvalues that explain 98% of the variation in the GRM. It corresponds to about 4NeL chromosome segments. In livestock species, the dimensionality varies from about 4k in commercial pigs and broilers to 14k in Holsteins. The segments appear to be of a variable size, with NeL segments accounting for more than 90% of the variance. The limited dimensionality enables the computation of the optimal size of the SNP chip for each species, allows low-cost genomic predictions for millions of genotyped animals, and are likely to affect the resolution of GWAS.

List of References

Aguilar, I., I. Misztal, A. Legarra & S. Tsuruta, 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128: 422-428.

Brard, S. & A. Ricard, 2015. Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132: 207-217.

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 136: 245-257.

MacLeod, A.K., C.S. Haley, J.A. Woolliams & P. Stam, 2005. Marker densities and the mapping of ancestral junctions. Genet. Res. 85: 69-79.

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D.A.L. Lourenco, B.O. Fragomeni & T.J. Lawlor, 2016. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J. Dairy. Sci. 99: 1968-1974.

Misztal, I., 2016. Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics 202: 401-409.

Misztal, I., A. Legarra & I. Aguilar, 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97: 3943-3952.

Pocrnic, I., D.A.L. Lourenco, Y. Masuda, A. Legarra & I. Misztal, 2016a. The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics 203: 573-581.

Pocrnic, I., D.A.L. Lourenco, Y. Masuda & I. Misztal, 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genet. Sel. Evol. 48: 82.

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

Stam, P., 1980. The distribution of fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35: 131-155.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423.