PREGSF90 – POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs

Ignacio Aguilar, Ignacy Misztal, Shogo Tsuruta, Andres Legarra, Huiyu Wang

PREGSF90 – POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs

I. Aguilar,[1] I. Misztal,[2] S. Tsuruta,[2] A. Legarra,[3] H. Wang[4]
[1]Instituto Nacional de Investigación Agropecuaria, Uruguay,
[2]University of Georgia, Athens, USA, [3]INRA, France, [4]Genus Plc, Hendersonville, TN, USA

ABSTRACT: Single step genomic methodology provides a unified framework to integrate phenotypic, pedigree and genomic information in prediction of breeding values for all individuals and in estimation of marker effects. Computational tools for the implementation of the single-step methodology are presented. Methodology, quality control of samples and different options to create genomic relationship matrices are discussed. Computing time for construction and inversion of genomic relationship matrices and large-scale genome-wide association study for heat tolerance in milk yield are presented.

Key words: genomic selection; ssGBLUP; genome-wide association; genomic relationship matrix.

INTRODUCTION

The single step genomic methodology provides a unified framework to integrate phenotypic, pedigree and genomic information in prediction of breeding values for all individuals (Aguilar et al. (2010); Christensen & Lund (2010)) and marker effect estimation (Wang et al. (2012)).

This unified approach modifies the pedigree-based relationship matrix to include a genomic relationship matrix (e.g. VanRaden (2008))), and the resulting mixed model equations involve the regular inverse of the numerator relationship matrix, the inverse of the genomic relationship matrix and the inverse of the pedigree-based relationship matrix for genotyped individuals (Aguilar et al. (2010); Christensen & Lund (2010)).

Minimal modifications of current software are necessary in order to incorporate extra relationship matrices. Adding such extra relationship matrices to current software for genetic evaluation and variance component estimation results in the application of genomic information in a broad kind of models and species (Misztal et al. (2010)).

Different genomic relationship matrices based on different assumptions were proposed (Amin et al. (2007); VanRaden (2008); Yang et al. (2010)), and in general several quality control of genotypes samples are involved in genomic analyses (e.g. (Wiggans et al. (2009); Wiggans et al. (2010))) which can affect the successful implementation of the single-step genomic methodology.

Using an equivalent model, estimation of SNP effects can be done using genomic estimated breeding values (GEBV) and the genomic relationship matrix (Stranden & Garrick (2009)), and methodology to include un-genotyped individuals as in a single-step framework was proposed (Wang et al. (2012)).

Thus, the main objective is to introduce computing tools for the implementation of single-step genomic and for the marker effect estimation using the BLUPF90 family of programs (Misztal et al. (2002))

MATERIALS AND METHODS

PREGSF90. This program is an interface to process the genomic information for the BLUPF90 family of programs. Although it was developed to help the implementation of the genomic selection following the single-step methodology ((Legarra et al. (2009); Misztal et al. (2009); Aguilar et al. (2010)), it can be use to apply different quality controls on genotypes, construction and inversion of genomic relationship matrices and pedigree relationship matrix for a subset of individuals using efficient computing methods (Aguilar et al. (2011)) and provide several outputs to detect possible errors with genotypes (e.g. (Simeone et al. (2011))).

Inputs for using PREGSF90 are a genotype file with marker information coded as (0/1/2/5) denoting homozygous, heterozygous and homozygous genotypes and missing SNP, respectively; a renumber pedigree file and a file with cross reference relating samples to renumber id's in the pedigree file. Creation of such files is simplified by using a renumbering tool (RENUMF90) distributed with the package.

In a normal run, this program will read all samples and apply a default quality control on SNP and samples, and create and store on disk. A file with extra relationship matrices need to implement singe-step genomic selection. This extra information is then used by analysis programs for variance component estimation (AIREMLF90, GIBBSxF90, THRGIBBSxF90) or solution of mixed model equations (BLUPF90, BLUP90IOD). However, simplification of the implementation is achieved if a SNP file is provided to the analysis programs, and then the PREGSF90 will be called automatically. Running PREGSF90 as a stand-alone program could be useful to perform different quality controls and data cleaning.

Quality Control. Quality control (QC) is performed by default for monomorphic, allele frequency, call rate (samples and SNP) and parent-progeny conflicts (samples, SNP). All have default parameters but can be changed by using appropriate options in parameter files. Potential duplicated samples are informed but not removed from analyses. Other QC include to check SNP from departure from Hardy-Weinberg equilibrium (HWE) and high correlated SNP as described in Wiggans et al. (2009)), with SNP at the same position and linkage disequilibrium calculation and filtering. In such cases when chromosomes and positions in genome is required, an extra file with map information needs to be provided. Specific chromosomes can be removed from the analysis, and sex chromosomes

Table 1. Computing time[§] for creation and inversion of genomic relationship matrix with different number of genotyped individuals.

Number of genotypes (thousand)	Genomic Relationship Matrix 50 k	
	Creation (min)	Inversion (min)
10	0.6	0.1
30	5.4	3
50	15	14
70	30	36
100	60	122
120	140	208
150	215	406

[§]Intel(R) Xeon(R) CPU E7- 4870 @ 2.40GHz, 24 CPU, 750 GB RAM.

can be specified in order to exclude them from parent-progeny and HWE analyses.

Mendelian Conflicts. If a parent-progeny check for Mendelian conflicts is found, the sample of the progeny will be removed. Having such a pair of conflicts in a single-step genomic implementation results in a non-positive definite relationship matrix with genotyped and non-genotyped individuals. Outputs with statistics are generated in order to establish a mislabeled sample or wrong pedigree assignment.

Genomic Relationship. Genomic relationship matrices are constructed with the following general formula: G = ZDZ'/k, where Z is center matrix Z=M-2p, with M the maker allele count matrix, and p, the allele frequency: D a diagonal matrix with weights, and k a scale parameter.

From this general formula different genomic relationship matrices can be constructed, using available options, by changing the values of the allele frequency (i.e. allele frequency from a file, calculated from samples or 0.5) , or by modifying the scale parameter (i.e. sum (2 p q), trace of ZDZ', etc.). Default options for the genomic relationship matrix include, allele frequency calculated from the samples, scale parameter as sum[2 (pq)] and D as an identity matrix.

Inspecting means and correlations from diagonal and off-diagonals between both matrices can provide a simple detection for incorrect or mislabeled samples that do not match with the pedigree file. In the extreme case with very low correlations for off-diagonals, the analysis will be stopped.

For blending the pedigree relationship matrix with the genomic relationship matrix, the scale of both matrices need to match. Results from several studies (Chen et al. (2011); Forni et al. (2011); Vitezica et al. (2011)) show that an appropriate blending and matching both matrices results in unbiased variance components or in prediction of unbiased genomic breeding values.

Output files. Several statistics are printed out, but also genomic and pedigree relationship matrices can be stored in files for diagnostic or other research studies.

`POSTGSF90.` This program calculate SNP effects as described in Wang et al. (2012)), and running in sequential runs with `PREGSF90` and `BLUPF90` and allow to use differential weights in the genomic relationship matrix and allow different SNP variances.

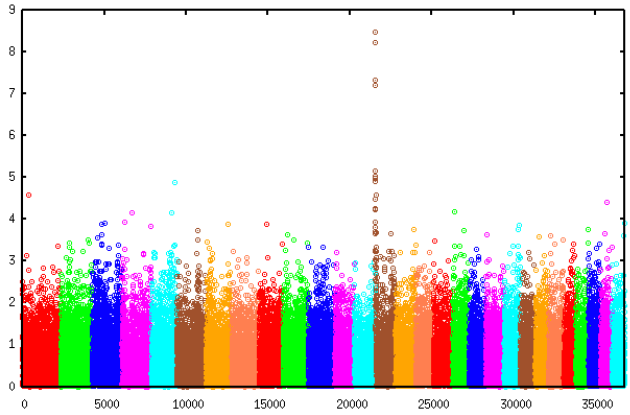

Figure 1. SNP effects for milk yield on first parity.

For each trait and random correlated effects, GEBV will be decomposed into SNP effects and stored in files. The SNP effects could be outputs as the single effect or as moving average of contiguous markers. If required, the variance explained by segments will be calculated, where segments could be defined with equal size of SNP or base on the position, (e.g. 1M BP)

Plots of SNP effects or variances explained by segments will be created and if required, high quality graphs can be created with R packages.

Availability. Compiled version in Linux / MacOSX / Windows of both programs are distributed with the BLUPF90 package [http://nce.ads.uga.edu/] and the documentation with descriptions of options: [http://nce.ads.uga.edu/wiki/doku.php].

RESULTS AND DISCUSSION

Although the computation of the inverses of such matrices has a cubic cost regarding the number of genotyped individuals, the efficient methods implemented in PREGSF90 can support a large number of genotypes. Table 1 presents the computing time for construction and inversion of the genomic relationship for different number of genotyped samples.

Large-scale genome-wide scan was performed for a multiple-trait test-day model that accounts for heat tolerance. Approximately 90 millions of test day records from US Holstein cows (AIPL, ARS, USDA) were combined with public weather data, involving 9 million

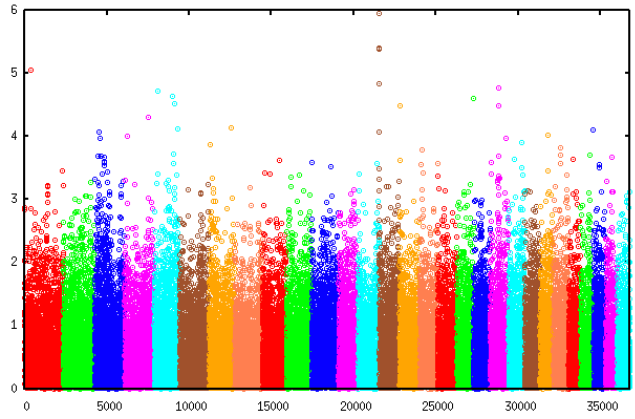

Figure 2. SNP effects for heat stress effect on milk yield on first parity.

animals in the pedigree. Details of the data used are in Aguilar et al. (2010)). Genotypes for 3,800 bulls using the Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA) were provided by the Animal Improvement Programs Laboratory, Agricultural Research Service, USDA (Beltsville, MD).

A complete evaluation includes creation and inversion of the genomic matrix, solving the mixed model equations, and estimation of SNP effects, taking about 16 hours. In the Figure 1 and 2, Manhattan plots for SNP effects for milk yield with and without heat stress are presented.

CONCLUSION

Computational tools for the implementation of single-step genomic evaluation and genome-wide association studies were presented. Different options allow checking for data quality and creation of a genomic relationship matrix in an efficient way for medium-large scale problems. Performing genome-wide scan with phenotypic information with no genotypes can be easily incorporated.

LITERATURE CITED

Aguilar, I., Misztal, I., Johnson, D. L., et al. (2010). J. Dairy Sci. 93: 743–752

Aguilar, I., Misztal, I., Legarra, A., et al. (2011). J Anim Breed Genet 128: 422-8

Aguilar, I., Tsuruta, S., Misztal, I. (2010). J. Anim. Breed. Genet. 127: 235–241

Amin, N., van Duijn, C. M., Aulchenko, Y. S. (2007). PLoS ONE 2: e1274

Chen, C. Y., Misztal, I., Aguilar, I., et al. (2011). J Anim Sci. 89: 2673-2679

Christensen, O., Lund, M. (2010). Genet. Sel. Evol. 42: 2

Forni, S., Aguilar, I., Misztal, I. (2011). Genet. Sel. Evol. 43: 1

Legarra, A., Aguilar, I., Misztal, I. (2009). J. Dairy Sci. 92: 4656–4663

Misztal, I., Aguilar, I., Legarra, A., et al. (2010). Proc 9th WCGALP. Leipzig, Germany.

Misztal, I., Legarra, A., Aguilar, I. (2009). J. Dairy Sci. 92: 4648–4655

Misztal, I., Tsuruta, S., Strabel, T., et al. (2002). Proc 7th WCGALP. Montpellier, France.

Simeone, R., Misztal, I., Aguilar, I., et al. (2011). J. Anim. Breed. Genet. 128: 386-393

Stranden, I., Garrick, D. J. (2009). J. Dairy Sci. 92: 2971–2975

VanRaden, P. M. (2008). J. Dairy Sci. 91: 4414–4423

Vitezica, Z. G., Aguilar, I., Misztal, I., et al. (2011). Genetics Research 93: 357-366

Wang, H., Misztal, I., Aguilar, I., et al. (2012). Genetics Research 94: 73-83

Wiggans, G. R., Sonstegard, T. S., VanRaden, P. M., et al. (2009). J. Dairy Sci. 92: 3431–3436

Wiggans, G. R., VanRaden, P. M., Bacheller, L. R., et al. (2010). J. Dairy Sci. 93: 2287–2292.

Yang, J., Benyamin, B., McEvoy, B. P., et al. (2010). Nat Genet 42: 565–569.