Bioinformatics, 2019, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genetics and Population Analysis

bWGR: Bayesian Whole-Genome Regression

Alencar Xavier[1,2], William M Muir[2] and Katy M Rainey[2, *]

[1]Corteva Agrisciences, 8305 NW 62nd Ave, Johnston IA 50131; [2]Purdue University, 915 W State St, West Lafayette IN 47907

* krainey@purdue.edu

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Whole-genome regressions methods represent a key framework for genome-wide prediction, cross-validation studies, and association analysis. The bWGR offers a compendium of Bayesian methods with various priors available, allowing users to predict complex traits with different genetic architectures.

Results: Here we introduce bWGR, an R package that enables users to efficient fit and cross-validate Bayesian and likelihood whole-genome regression methods. It implements a series of methods referred to as the Bayesian alphabet under the traditional Gibbs sampling and optimized Expectation-Maximization. The package also enables fitting efficient multivariate models and complex hierarchical models. The package is user-friendly and computational efficient.

Availability and implementation: bWGR is an R package available in the CRAN repository. It can be installed in R by typing: `install.packages("bWGR")`

Contact: alencar.xavier@corteva.com, krainey@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide markers have been increasingly deployed for the prediction of complex traits since the concept of genomic prediction was introduced (Meuwissen et al. 2001). Whole-genome regression (WGR) methods predict traits as a linear combination of marker effects that capture quantitative trait loci (QTL) and the relationship among individuals (Habier et al. 2007). A large variety of models exist, each with different prior assumptions that are optimized for a specific genetic architecture (de los Campos et al. 2013). Evaluating the parameterizations of genomic information in prediction models to suit different genetic architectures can enhance prediction accuracy.

Few statistical packages enable genome-wide prediction, including rrBLUP, BGLR and VIGoR (Endelman 2011, Pérez and de los Campos 2014, Onogi and Iwata 2016). Genome-wide models are sensitive to the algorithm implementation, such that two implementations of the same model often lead to reasonably different results (Gianola et al. 2009, Lehermeier et al. 2013). In a user-friendly framework, the bWGR package implements a compendium of likelihood and Bayesian methods, via expectation-maximization (EM) and Markov Chain Monte Carlo (MCMC), at univariate and multivariate level. It also implements a mixed model solver that enables modeling replicated observations, computing marker effects using link functions and accounting for nuisance parameters.

2 Markov Chain Monte Carlo methods

MCMC methods constitute the most popular set of WGR (Gianola 2013). These include Bayesian Ridge Regression, BayesA, BayesB (Meuwissen et al. 2001), BayesC, BayesCpi, BayesDpi (Habier et al. 2011), Bayesian LASSO (Park and Casella 2008), and Reproducing Kernel Hilbert Spaces (RKHS) regression (de los Campos et al. 2010). The variable selection of BayesB and BayesC was implemented through Gibbs Sampling unconditional prior (Kuo and Mallick 1998) and Metropolis-Hasting for BayesCpi and BayesDpi. In our models, the prior specifications are similar but not identical to the BGLR package (Pérez and de los Campos 2014). We kept the models less hierarchical like those originally proposed by Meuwissen et al. (2001), with restricted Bayesian learning (Lehermeier et al. 2013) to avoid under- and over-regularization. These methods can be performed either from bWGR's generalized function "wgr" or by their standalone implementation written entirely in C++ (Eddelbuettel et al. 2011). The generalized function "wgr" enable users to combine a whole-genome regression with a kernel

method, such as combining BayesB and RKHS. It also has an exclusive feature as it enables the subsampling of Markov chains to save time and computational power (Xavier et al. 2017).

3 Expectation-Maximization methods

EM methods provide an elegant and efficient way to reduce the computation time due to MCMC iterations (Shepherd et al. 2010). Iterative procedures may replace Gibbs sampling by updating parameters with the expectation as opposed to sampling and averaging the posteriors. This algorithmic variation of the traditional MCMC solver of the Bayesian methods was proposed by Meuwissen et al. (2009). These EM Bayesian methods can calibrate WGR without loss in accuracy (Lopez et al. 2019). From the Bayesian alphabet implemented via EM, the package provides implementations of BayesA ("emBA"), BayesB ("emBB"), BayesC ("emBC") and Bayesian Lasso ("emBL"). The package also includes a Gaussian maximum likelihood ("emML") and an elastic-net ("emEN"), and the fast Laplace model ("emDE") (Xavier et al. 2019).

To facilitate cross-validation studies, the bWGR implements "emCV" and "mcmcCV" which allow k-fold cross-validations and leave-a-level-out cross-validation (LLO), where the cross-validations are performed on the phenotypes or true breeding values (TBV), if provided.

4 Multivariate methods

The package provides a ridge-type ("mrr") and kernel-type ("mkr") function for multivariate regressions that enable simultaneous modeling of two or more response variables. Both implementations were based on an efficient Gauss-Seidel (Legarra and Misztal 2008) paired with an efficient first-derivative estimation of EM-REML like variance components (Schaeffer 1986) written in C++. These implementations are fast and memory efficient by avoiding explicit matrix inversion or Kronecker products, also robust to a relatively large number of traits and accept missing-values. The multivariate regression functions do not offer the modeling flexibility of other packages but the computation time to fit the model is approximately 8% of those fit with REML implementations (Gilmour et al. 1995, Covarrubias-Parazan 2016) and without the burden of MCMC methods (Hadfield 2010, Montesinos-López et al. 2019).

5 Hierarchical mixed models

The functions implemented via MCMC and EM enable simple models with one or two random effects. For more complex models, the function "mixed" enables fitting models with multiple fixed and random effects, with or without marker information through link functions.

Design to be a multi-purpose function, users can use the "mixed" function to run a wide range of models, from phenotypic analysis to single-step models. The function can estimate best linear unbiased predictors (BLUPs), marker effects, and variance components, while accounting for environmental factors and other nuisance parameters.

6 Additional tools

The bWGR package is a self-coined toolbox for genetic analysis. In this section we will briefly describe some of the key additional functions.

6.1 Relationship matrices: "GRM" creates the genomic relationship described by VanRaden (2008) and "GAU" generates a Gaussian kernel often deployed for RKHS regression (de los Campos et al. 2010);

6.2 Genotyping imputation: "markov" implements a forward Markov model that accounts for the linkage disequilibrium among neighbor markers, and "IMP" imputes missing values with the expected value of the marker (marker average).

6.3 Spatial analysis: Providing field coordinates and phenotype, the spatial covariate function "SPC" creates covariates on neighbor plots (Lado et al. 2013). The "SPM" function generates a design matrix for spatial adjustment (Muir 2005, Gilmour et al. 1995).

6.4 Miscellaneous: "CNT" centralizes markers for a better blending a posteriori and unbiasedness. "SibZ" creates a WGR-compatible matrix from pedigree. "emGWA" runs a ridge regression coupled with genome-wide association studies that outputs values for prediction and inference.

6.5 Two random effects: Hybrid breeding models often fit two random terms, such as Additive-Dominance and Parent1-Parent2. Besides the function "mixed", simpler stand-alone functions include: "BayesA2", "BayesB2", "BayesRR2", "emML2", "mrr2X" and "mkr2X".

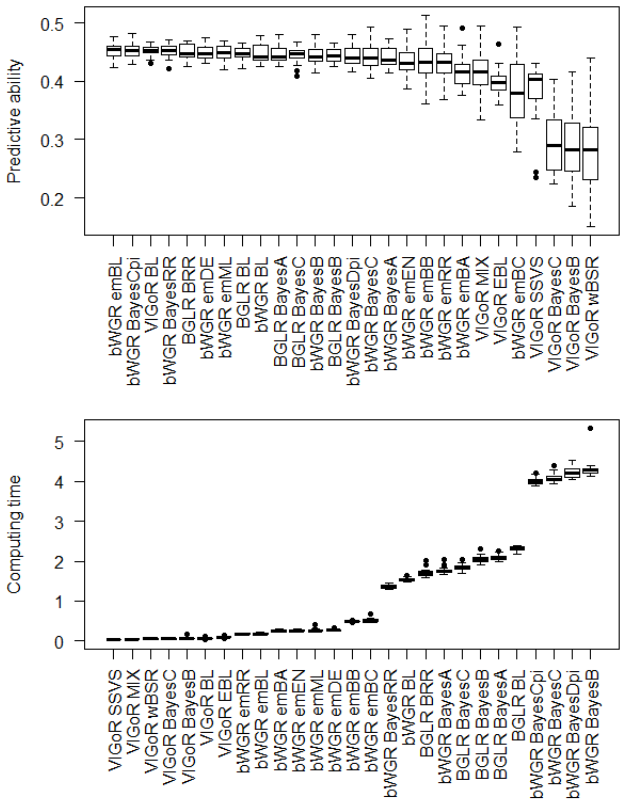

Fig. 1. Comparison between bWGR, BGLR and VIGoR packages. Predictive ability (top) as the correlation between predicted and observed values and computing time (bottom) in seconds to fit a whole-genome regression, 5-fold cross-validation repeated 20x on the wheat dataset available on the BGLR package.

Conclusions

The bWGR package has implemented a series of whole-genome regression methods in Bayesian framework that covers a variety of priors to enable accurate genome-wide prediction of complex traits across various genetic architectures. Implementations are available in the traditional MCMC framework as well as efficient EM methods. The package also enables efficient multivariate and hierarchical modeling. The package focuses on statistically sound methodologies implemented for high computational performance and prediction accuracy.

Funding

Conflict of Interest: none declared.

References

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PloS one*, **11**(6), e0156744.

de los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Calus M. P., 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**(2): 327–345.

de los Campos G., Gianola D., Rosa G. J., Weigel K. A., Crossa J., 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **92**(04): 295–308.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., ... and Bates, D. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**(8), 1-18.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, **4**(3), 250-255.

Gianola D. 2013 Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* **194**(3): 573–596.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**(1), 347-363.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440-1450.

Habier D., Fernando R. L., Kizilkaya K., Garrick D. J., 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**(1): 186.

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, **177**(4), 2389-2397.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, **33**(2), 1-22.

Kuo, L., Mallick, B. (1998). Variable selection for regression models. Sankhyā: *The Indian Journal of Statistics, Series B*, 65-81.

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., ... von Zitzewitz, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes, Genomes, Genetics*, **3**(12), 2105-2114.

Legarra, A., Misztal, I. (2008). Computing strategies in genome-wide selection. *Journal of Dairy Science*, **91**(1), 360-366.

Lehermeier, C., Wimmer, V., Albrecht, T., Auinger, H. J., Gianola, D., Schmid, V. J., Schön, C. C. (2013). Sensitivity to prior specification in Bayesian genome-based prediction models. *Statistical applications in genetics and molecular biology*, **12**(3), 375-391.

Lopez, M. A., Xavier, A., Rainey, K. M. (2019). Phenotypic Variation and Genetic Architecture for Photosynthesis and Water Use Efficiency in Soybean (Glycine max L. Merr). *Front. Plant Sci.* **10**(680).

Meuwissen, T. H., Solberg, T. R., Shepherd, R., Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution*, **41**(1), 2.

Meuwissen T. H. E., Hayes B. J., Goddard M. E., 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819–1829.

Montesinos-López, O. A., Montesinos-López, A., Luna-Vázquez, F. J., Toledo, F. H., Pérez-Rodríguez, P., Lillemo, M., Crossa, J. (2019). An R Package for Bayesian Analysis of Multi-environment and Multi-trait Multi-environment Data for Genome-Gased Prediction. *G3: Genes, Genomes, Genetics*, g3-400126.

Muir, W. M. (2005). Incorporation of competitive effects in forest tree or animal breeding programs. *Genetics*, **170**(3), 1247-1259.

Onogi, A., and Iwata, H. (2016). VIGoR: variational Bayesian inference for genome-wide regression. *Journal of Open Research Software*, 4(1).

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681-686.

Pérez P., de los Campos G. 2014 Genome-wide regression & prediction with the BGLR statistical package. *Genetics* **198**(2): 483–495.

Schaeffer, L. R. (1986). Pseudo expectation approach to variance component estimation. *Journal of Dairy Science*, **69**(11), 2884-2889.

Shepherd, R. K., Meuwissen, T. H., and Woolliams, J. A. (2010). Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC bioinformatics*, **11**(1), 529.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, **91**(11), 4414-4423.

Xavier, A. (2019). Efficient Estimation of Marker Effects in Plant Breeding. *G3: Genes, Genomes, Genetics*, **9**(11): 1-12.

Xavier, A., Xu, S., Muir, W., Rainey, K. M. (2017). Genomic prediction using subsampling. *BMC bioinformatics*, **18**(1), 191.