SOFTWARE METAPAPER

VIGoR: Variational Bayesian Inference for Genome-Wide Regression

Akio Onog[1] and Hiroyoshi Iwata[1]

[1] Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences,
The University of Tokyo, Tokyo 113-8657, Japan
onogiakio@gmail.com
Corresponding author: Akio Onogi

Genome-wide regression using a number of genome-wide markers as predictors is now widely used for genome-wide association mapping and genomic prediction. We developed novel software for genome-wide regression which we named VIGoR (variational Bayesian inference for genome-wide regression). Variational Bayesian inference is computationally much faster than widely used Markov chain Monte Carlo algorithms. VIGoR implements seven regression methods, and is provided as a command line program package for Linux/Mac, and as a cross-platform R package. In addition to model fitting, cross-validation and hyper-parameter tuning using cross-validation can be automatically performed by modifying a single argument. VIGoR is available at https://github.com/Onogi/VIGoR. The R package is also available at https://cran.r-project.org/web/packages/VIGoR/index.html.

(1) Overview

Introduction

Linear regression methods where a number of genome-wide markers are used as predictors are currently used in genomic prediction [1–3] and genome-wide association mapping [4–6]. Bayesian regression methods have attracted particular attention and a number of variations have been proposed [2], which typically use Markov chain Monte Carlo (MCMC) algorithms for parameter inference (e.g., [1, 7, 8]). Consequently, the software currently available for Bayesian regression (e.g., GenSel [9], AlphaBayes [10], GS3 [11], BayeZ [12], and BGLR [13]) is mainly based on MCMC. However, because of the computational burden associated with MCMC, analyzing huge datasets, such as those consisting of hundreds of thousands of markers, within realistic time scales is often unfeasible. Moreover, intensive cross-validation (CV) for evaluating predictive ability or for tuning hyperparameters is also difficult, even in moderate-sized datasets. These shortcomings of MCMC-based methods have hampered the widespread application of Bayesian methods to genome-wide association mapping and genomic prediction. To tackle the shortcomings of the MCMC-based software, we developed novel software for whole-genome regression in a Bayesian framework, which we named VIGoR (variational Bayesian inference for genome-wide regression). VIGoR is based on variational Bayesian inference (VB), which is computationally much faster than MCMC, and can implement seven regression methods: Bayesian lasso (BL) [14], extended Bayesian lasso (EBL) [15], weighted Bayesian shrinkage regression (wBSR) [16], BayesB [1], BayesC [8], stochastic search variable selection (SSVS) [17], and Bayesian mixture regression (MIX) [18]. BL, EBL, and wBSR implemented by VIGoR were used for genomic prediction of rice agronomic traits [19], and genomic prediction and association mapping in tomato [20]. EBL was also used for genomic prediction of rice heading date [21]. The command line program (CLP) package for the Linux/Mac platform is available at https://github.com/Onogi/VIGoR. A pdf manual is also available at the URL [24]. The R package is cross-platform and is available at https://github.com/Onogi/VIGoR and from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/VIGoR/index.html. The pdf manual contains the explanations of both the CLPs and the R functions.

Regression methods and algorithms

VIGoR assumes the following linear regression model:

$$y_i = \sum_{j=1}^{F} z_{ij}\alpha_j + \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p + \varepsilon_i$$

where y_i is the phenotypic value (response variable) of individual i, F is the number of covariates other than markers, z_{ij} is the covariate corresponding to the effect α_j, P is

the number of markers, γ_p is the indicator variable that takes 0 or 1, x_{ip} is the genotype of marker p, β_p is the effect of marker p, and ε_i is the residual. The indicator variables are fixed to 1 except in wBSR. The residual, ε_i, is assumed to follow a normal distribution with mean = 0 and variance = $1/\tau_0^2$. α_j is assumed to be proportional to a constant value, i.e., assumed to follow a non-informative prior distribution. The prior distribution of β_p differs among the regression methods (**Table 1**). These prior distributions were proposed to select important markers (i.e., markers strongly associated with phenotypes) efficiently from given markers. Variational Bayesian algorithms for these regression methods which are implemented by VIGoR are illustrated in the pdf manual of VIGoR [24]. All the regression methods require the user specifying hyperparameters to determine the shapes of prior distributions. In **Table 2**, hyperparameters to be specified by the user are presented for each regression method. We briefly describe how to specify these hyperparameter values in the next section. In **Table 3**, we present the references of the regression methods and variational Bayesian inference.

Implementation and architecture

Both the CLP and R packages consist of two programs/functions, *vigor* and *hyperpara* (**Fig. 1**). *Vigor* conducts genome-wide regression analyses, and has three main functions: fitting regression models to data (***Model fitting***), fitting models after tuning hyperparameter values using CV (***Model fitting after hyperparameter tuning***), and evaluating predictive ability of regression methods by CV (***Cross-validation***). Using ***Model fitting***, users can peform variable selection (association mapping) by fitting genome-wide regression models to data and estimating marker effects. This is the default function. Using ***Model fitting after hyperparameter tuning***, users can estimate marker effects with the hyperparameters tuned automatically using CV. Using ***Cross-validation***, users can evaluate the predictive ability of regression models using CV.

Vigor requires the phenotypic values (response variables), and the marker genotypes (predictor variables) as mandatory input information (**Fig. 1**). In addition, all the regression methods implemented by *vigor* require hyperparameter values that users should specify (**Fig.1** and **Table 2**).

	Hierarchical level		
	1st	**2nd**	**3rd**
BL	$\beta_p \sim N\left(0, \dfrac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \dfrac{\lambda^2}{2}\right)$	$\lambda^2 \sim G\left(\varphi, \varpi\right)$
EBL	$\beta_p \sim N\left(0, \dfrac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \dfrac{\delta^2 \eta_p^2}{2}\right)$	$\delta^2 \sim G\left(\varphi, \varpi\right)$ $\eta_p^2 \sim G\left(\psi, \theta\right)$
wBSR	$\beta_p \sim N\left(0, \sigma_p^2\right)$ $\gamma_p \sim Bernoulli\left(\kappa\right)$	$\sigma_p^2 \sim \chi^{-2}\left(\nu, S^2\right)$	
BayesB	$\beta_p \sim N\left(0, \sigma_p^2\right)$ if $\rho_p = 1$ $\beta_p = 0$ if $\rho_p = 0$	$\sigma_p^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
BayesC	$\beta_p \sim N\left(0, \sigma^2\right)$ if $\rho_p = 1$ $\beta_p = 0$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
SSVS	$\beta_p \sim N\left(0, \sigma^2\right)$ if $\rho_p = 1$ $\beta_p \sim N\left(0, c\sigma^2\right)$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
MIX	$\beta_p \sim N\left(0, \sigma_A^2\right)$ if $\rho_p = 1$ $\beta_p \sim N\left(0, \sigma_B^2\right)$ if $\rho_p = 0$	$\sigma_A^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\sigma_B^2 \sim \chi^{-2}\left(\nu, cS^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	

Table 1: Prior distributions of the marker effects and indicator variables.[a]

[a]Prior distributions of the marker effects have three-level hierarchical structures for BL and EBL, and two-level for the other methods.

BL, Bayesian lasso; EBL, extended Bayesian lasso; wBSR, weighted Bayesian shrinkage regression; SSVS, stochastic search variable selection; MIX, Bayesian mixture regression; *N*, normal distribution; *Inv-G*, inverse-gamma distribution; *G*, gamma distribution; *Bernoulli*, Bernoulli distribution; χ^{-2}, scaled inverse-chi-square distribution.

Regression	Hyperparameters[a]	Less influential hyperparameters (default values)	Influential hyperparameters determined by *hyperpara*
BL	φ, ω	φ (1.0)	ω
EBL	φ, ω, ψ, θ	φ (0.1), ω (0.1), ψ (1.0)	θ
wBSR	v, S^2, κ	v (5.0)	S^2
BayesB	v, S^2, κ	v (5.0)	S^2
BayesC	v, S^2, κ	v (5.0)	S^2
SSVS	c, v, S^2, κ	v (5.0)	c, S^2
MIX	c, v, S^2, κ	v (5.0)	c, S^2

Table 2: Hyperparameters required by *vigor*.
[a]For each regression model, the hyperparameters in this table correspond to those listed in Table 1. Among these hyperparameters, κ of wBSR, BayesB, BayesC, SSVS, and MIX is determined by the user. The other hyperparameters are set as default or can be determined by the function *hyperpara*.
BL, Bayesian lasso; EBL, extended Bayesian lasso; wBSR, weighted Bayesian shrinkage regression; SSVS, stochastic search variable selection; MIX, Bayesian mixture regression

|	Regression methods	Variational Bayesian algorithm
BL	[14]	[22]
EBL	[15]	[22]
wBSR	[16]	[23]
BayesB	[1]	[24]
BayesC	[8]	[25]
SSVS	[17]	[24]
MIX	[18]	[24]

Table 3: References of the regression methods and variational Bayesian algorithms.

To make specification feasible, we provide the other program/function, *hyperpara*, which calculates the values of hyperparameters that influence on inference, based on the values of hyperparameters that influence less and several assumptions of the genetic architecture (**Table 2**). Because the values of less influential hyperparameters are determined by default, users only input the assumptions of the genetic architecture to *hyperpara*. The required assumptions are (1) proportion of phenotypic variance (variance of response variable) that can be explained by the markers (predictor variables) (referred to as *Mvar*), and (2) proportion of markers with non-zero effects (referred to as κ). For example, when κ and *Mvar* are 0.01 and 0.5, respectively, this setting corresponds to an assumption that a half of phenotypic variance is explained by 1 % of markers. Based on this assumption, *hyperpara* calculates the values of influential hyperparameters. Explanations of

the calculation of hyperparameter values are provided in the pdf manual of VIGoR [24].

The CLPs, *vigor* and *hyperpara*, were written with C, and are distributed as standalone pre-compiled programs. Source codes are also available at https://github.com/Onogi/VIGoR. The programs can be built, for example, by typing gcc vigor.c -o vigor (Mac) or gcc vigor.c -o vigor -lm (Linux). The default function of vigor is **Model fitting**. ***Model fitting after hyperparameter tuning*** and ***Cross-validation*** can be conducted by adding options *-t* and *-c*, respectively (**Fig. 2**). The phenotypic values (response variables), and the marker genotypes (predictor variables) are provided as text files.

The R function *vigor* calls a C function from C library which is included in the package. Thus, calculation speed of the R function is almost equivalent to that of the CLP *vigor. Hyperpara* was developed with R. Both *vigor*
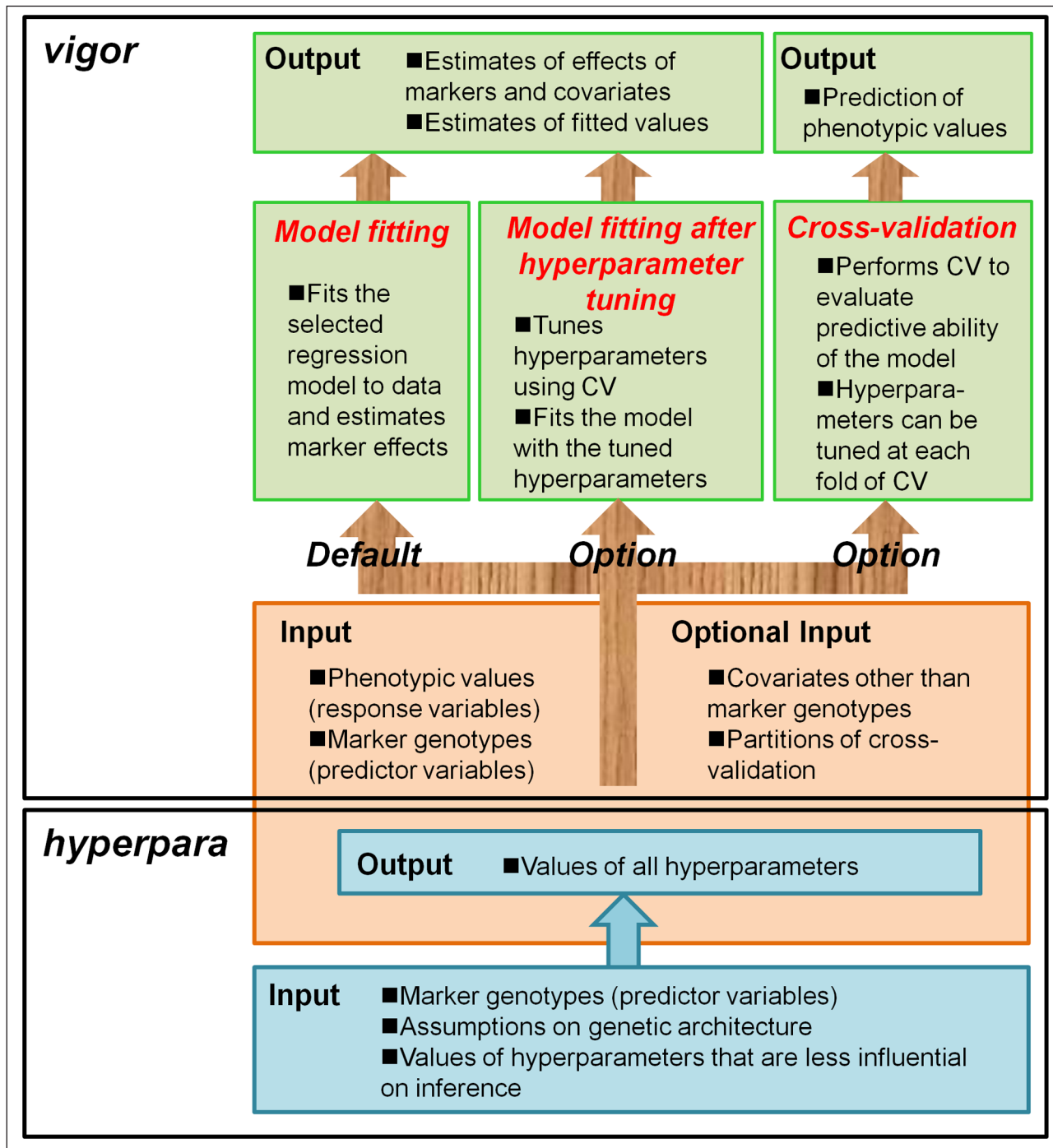
Figure 1: Overview of analysis by VIGoR. VIGoR consists of two programs/functions, *vigor* and *hyperpara*. *Vigor* conducts genome-wide regression analysis and has three main functions; ***Model fitting***, ***Model fitting after hyperparameter tuning***, and ***Cross-validation***. The former two functions output the estimates of the marker and covariate effects, and the fitted values. The last function outputs the predicted values obtained by cross-validation. *Vigor* requires three kinds of input information, phenotypic values (response variable), marker genotypes (predictor variable), and hyperparameter values. Hyperparameter values can be determined by the user or by using *hyperpara*. *Hyperpara* calculates the values of hyperparameters that influence on inference, based on the assumptions of the genetic architecture and values of hyperparameters that influence less.

and *hyperpara* have no dependency to other R packages except for those included in system library. The usages of the R functions are similar to those of the CLPs (**Fig. 2**). The phenotypic values and the marker genotypes are input to *vigor* as a vector and a matrix objects, respectively. The default function of *vigor* is ***Model fitting***. ***Model fitting***

after hyperparameter tuning and *Cross-validation* can be executed by adding arguments "tuning" and "cv", respectively (**Fig. 2**).

Both the CLP and R packages have advantages. The advantage of the CLP package is that the CLP *vigor* can accepts PED files of PLINK [26] and allele dosage

vigor	Command line program	R function
Model fitting	$./vigor Pheno.txt Geno.txt BL 1 0.1	>vigor (Pheno, Geno, "BL", c(1, 0.1))
Model fitting after hyperparameter tuning	$./vigor Pheno.txt Geno.txt BayesB 5 1 0.01 -v 5 1 0.1 -t	>vigor (Pheno, Geno, "BayesB", + matrix (c(5, 1, 0.01, 5, 1, 0.1), nc=3, byrow=T), + "tuning")
Cross-validation	$./vigor Pheno.txt Geno.txt BayesC 5 1 0.01 -c 5	>vigor (Pheno, Geno, "BayesC", c(5, 1, 0.01), "cv", 5)
hyperpara	$./hyperpara Geno.txt 0.5 wBSR 0.01	>hyperpara (Geno, 0.5, "wBSR", 0.01)

Figure 2: Examples of the usage of *vigor* and *hyperpara*. "sample.pheno.txt" and "Pheno$Height" are the example file and object that contain the phenotypic values (response variables), and "sample.geno.txt" and "Geno" are the example file and object that contain the marker genotypes (predictor variables). These files and objects are included in the command line program (CLP) and R packages, respectively. The regression methods are specified by their abbreviations (e.g., BL, BayesB, and wBSR). The argument(s) immediately after the regression methods are the hyperparameter values. Hyperparameters should be ordered as in the second column of **Table 2**. In the example of ***Model fitting***, 1 and 0.1 are the values of φ and ω of Bayesian lasso, respectively. In the example of ***Model fitting after hyperparameter tuning***, two hyperparameter value sets, [$v = 5$, $S^2 = 1$, $\kappa = 0.01$] and [$v = 5$, $S^2 = 1$, $\kappa = 0.1$], are provided using the -v option (CLP) and as a matrix (R function). The better set is chosen using cross-validation (CV), and model fitting is performed automatically with the chosen set. The -t option (CLP) and the argument "tuning" (R function) indicate this procedure. In the example of ***Cross-validation***, a five-fold CV is performed. The -c option (CLP) and "cv" (R function) indicate CV, and the argument immediately after this option/argument (here 5) is the fold number. In the example of *hyperpara*, the second (0.5) and fourth (0.01) arguments are the values of *Mvar* and κ, respectively (see the main text for the explanations of *Mvar* and κ).

files of Beagle [27] as the marker genotype file. Both PLINK and Beagle are popular association mapping and genotype imputation software, respectively. Thus, the users of PLINK or Beagle will easily perform analyses of VIGoR. Meanwhile, the advantage of the R package is that it can visualize the analysis results easily, that is, Manhattan plots can be drawn with a one-row R code (see the pdf manual or R documentation). It is also easy to evaluate prediction accuracy when ***Cross-validation*** is executed, by calculating Pearson correlation coefficient between the predicted and true values (see the pdf manual or R documentation). Users can select the packages according to their analysis purposes or environments.

Quality control
The CLPs for Linux were compiled under the Linux kernel release 3.13.0-24-generic with a X86-64 machine. We have not tried the programs in other releases of Linux kernel. The CLPs for Mac were compiled under OS X ver. 10.6.8. We verified that the programs run under a recent version, ver. 10.9.5.

The R functions were made under R version 3.0.2, and the package was build using Mac (OS X ver. 10.6.8). We verified that the package can be loaded and the functions run in Windows 7/8, Mac (OS X ver. 10.6.8 and 10.9.5), and Linux (3.13.0-24-generic).

(2) Availability
Operating system
CLP package: Linux (kernel 3.13.0-24-generic), and Mac (OS X ver. 10.6.8 and 10.9.5).

R package: Windows 7/8, Linux (kernel 3.13.0-24-generic), and Mac (OS X ver. 10.6.8 and 10.9.5).

Programming language
C (CLP programs) and C and R (R functions).

Additional system requirements
None.

Dependencies
The R package requires installation of R (http://cran.r-project.org/).

List of contributors
Akio Onogi and Hiroyoshi Iwata
Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo.

Software location
Archive
  ##### *Name*
  CLP package: GitHub
  R package: GitHub and CRAN
  ##### *Persistent identifier*
  CLP package: https://github.com/Onogi/VIGoR
  R package: https://cran.r-project.org/web/packages/VIGoR/index.html
  ##### *Licence*
  MIT license
  ##### *Publisher*
  Akio Onogi

Date published
20/05/2015

Language
English.

(3) Reuse potential

Because both the CLPs and R functions run by specifying only a few arguments, these programs will be approachable for geneticists who are interested in association mapping or genomic prediction. In addition, both the CLP and R functions *vigor* and *hyperpara* can accept predictor variables other than the marker genotypes. Therefore, although we focus on genome-wide regression here, VIGoR can be applied into various problems where variable selection is required for huge data. Thus, VIGoR will have a wide reuse potential.

Competing Interests
The authors declare that they have no competing interests.

References

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

2. **Karkkainen, H P** and **Sillanpaa, M J** 2012 Back to basics for Bayesian model building in genomic selection. *Genetics*, 191: 969–987. DOI: http://dx.doi.org/10.1534/genetics.112.139014

3. **de los Campos, G, Hickey, J M, Pong-Wong, R, Daetwyler, HD** and **Calus, M P** 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193: 327–345. DOI: http://dx.doi.org/10.1534/genetics.112.143313

4. **Xu, S** 2003 Estimating polygenic effects using markers of the entire genome. *Genetics*, 163: 789–801.

5. **Karkkainen, H P** and **Sillanpaa, M J** 2012 Robustness of Bayesian multilocus association models to cryptic relatedness. *Ann Hum Genet*, 76: 510–523. DOI: http://dx.doi.org/10.1111/j.1469-1809.2012.00729.x

6. **Hoffman, G E, Logsdon, B A** and **Mezey, J G** 2013 PUMA: a unified framework for penalized multiple regression analysis of GWAS data. *PLoS Comput Biol*, 9: e1003101. DOI: http://dx.doi.org/10.1371/journal.pcbi.1003101

7. **de los Campos, G, Naya, H, Gianola, D, Crossa, J, Legarra, A, Manfredi, E, Weigel, K** and **Cotes, J M** 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182: 375–385. DOI: http://dx.doi.org/10.1534/genetics.109.101501

8. **Habier, D, Fernando, R L, Kizilkaya, K** and **Garrick, D J** 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12: 186. DOI: http://dx.doi.org/10.1186/1471-2105-12-186

9. **Fernando, R** and **Garrick, D J** 2008 GenSel: User Manual for a Portfolio of Genomic Selection Related Analyses Animal Breeding and Genetics, Iowa State University, Ames, IA. Available at http://bigs.ansci.iastate.edu/bigsgui.

10. **Hickey, J M** and **Tier, B** 2009 AlphaBayes (Beta): Software for polygenic and whole genome analysis. User Manual. University of New England, Armidale, Australia. Available at https://sites.google.com/site/hickeyjohn/alphabayes.

11. **Legarra, A, Ricardi, A** and **Filangi, O** 2010 GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCπ and Bayesian Lasso). Available at http://snp.toulouse.inra.fr/~alegarra/.

12. **Janss, L L G** 2010 Bayz manual version version 2.03 Janss Biostatistics, Leiden, The Netherlands. Available at http://www.bayz.biz/.

13. **Perez, P** and **de Los Campos, G** 2014 Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* 198:483-495. DOI: http://dx.doi.org/10.1534/genetics.114.164442

14. **Park, T** and **Casella, G** 2008 The Bayesian lasso. *J Am Stat Assoc*, 103: 681–686. DOI: http://dx.doi.org/10.1198/016214508000000337

15. **Mutshinda, C M** and **Sillanpaa, M J** 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* 186:1067-1075. DOI: http://dx.doi.org/10.1534/genetics.110.119586

16. **Hayashi, T** and **Iwata, H** 2010 EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genet,* 11: 3. DOI: http://dx.doi.org/10.1186/1471-2156-11-3

17. **George, E I** and **McCulloch, R E** 1993 Variable selection via Gibbs sampling. *J Am Stat Assoc*, 88: 881–889. DOI: http://dx.doi.org/10.1080/01621459.1993.10476353

18. **Luan, T, Woolliams, J A, Lien, S, Kent, M, Svendsen, M** and **Meuwissen, T H** 2009 The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics*, 183: 1119–1126. DOI: http://dx.doi.org/10.1534/genetics.109.107391

19. **Onogi, A, Ideta, O, Inoshita, Y, Ebana, K, Yoshioka, T, Yamasaki, M** and **Iwata, H** 2015 Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). *Theor Appl Genet*, 128: 41–53. DOI: http://dx.doi.org/10.1007/s00122-014-2411-y

20. **Yamamoto, E, Matsunaga, H, Onogi, A, Kajiya-Kanegae, H, Minamikawa, M, Suzuki, A, Shirasawa, K, Hirakawa, H, Nunome, T, Yamaguchi, H, Miyatake, K, Ohyama, A, Iwata, H** and **Fukuoka, H** 2016 A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci. Rep.* DOI: http://dx.doi.org/10.1038/srep19454

21. **Onogi, A, Watanabe, M, Mochizuki, T, Hayashi, T, Nakagawa, H, Hasegawa, T** and **Iwata, H** 2016 Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. *Theor Appl Genet*. (accepted). DOI: http://dx.doi.org/10.1007/s00122-016-2667-5

22. **Li, Z** and **Sillanpaa, M J** 2012 Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics*, 190: 231–249. DOI: http://dx.doi.org/10.1534/genetics.111.134866

23. **Hayashi, T** and **Iwata, H** 2013 A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics*, 14: 34. DOI: http://dx.doi.org/10.1186/1471-2105-14-34

24. **Onogi, A** and **Iwata, H** 2015 Documents for VIGoR (May 2015). Available at https://github.com/Onogi/VIGoR.

25. **Carbonetto, P** and **Stephens, M** 2012 Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal*, 7: 73–108. DOI: http://dx.doi.org/10.1214/12-BA703

26. **Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, M A R, Bender, D, Maller, J, Sklar, P, de Bakker, P I W, Daly, M J** and **Sham, P C** 2007 PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, 81: 559–575. DOI: http://dx.doi.org/10.1086/519795

27. **Browning, S R** and **Browning, B L** 2007 Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81: 1084–97. DOI: http://dx.doi.org/10.1086/521987