

1 **ShinyGPAS: Interactive genomic prediction**
2 **accuracy simulator based on deterministic**
3 **formulas**

4 Gota Morota*

5 *Department of Animal Science, University of Nebraska-Lincoln, Lincoln,
6 Nebraska 68583

7 Keywords: deterministic equation, genomic prediction, interactive graph, shiny, simulation,
8 web

9

10 Running title: Shiny-based genomic prediction accuracy simulator

11

12 Corresponding author:

13 Gota Morota

14 Department of Animal Science

15 University of Nebraska-Lincoln

16 PO Box 830908

17 Lincoln, NE 68583-0908, USA.

18 E-mail: morota@unl.edu

19

20 Abstract

21 **Background:** Deterministic formulas highlight the relationships among prediction accuracy
22 and potential factors influencing prediction accuracy prior to performing computationally
23 intensive cross-validation. Visualizing such deterministic formulas in an interactive man-
24 ner may lead to a better understanding of how genetic factors control prediction accuracy.
25 **Results:** The software to simulate deterministic formulas for genomic prediction accuracy
26 was implemented in R and encapsulated as a web-based Shiny application. ShinyGPAS
27 (Shiny Genomic Prediction Accuracy Simulator) simulates various deterministic formulas
28 and delivers dynamic scatter plots of prediction accuracy vs. genetic factors impacting pre-
29 diction accuracy, while requiring only mouse navigation in a web browser. ShinyGPAS is
30 available at: <https://chikudaisei.shinyapps.io/shinygpas/>. **Conclusion:** ShinyGPAS is a
31 shiny-based interactive genomic prediction accuracy simulator using deterministic formulas.
32 It can be used for interactively exploring potential factors influencing prediction accuracy
33 in genome-enabled prediction, simulating achievable prediction accuracy prior to genotyp-
34 ing individuals, or supporting in-class teaching. ShinyGPAS is open source software and
35 it is hosted online as a freely available web-based resource with an intuitive graphical user
36 interface.

37 Background

38 Prediction of genomic values from high-dimensional single nucleotide polymorphisms is a
39 primal interest in animal breeding and quantitative genetics (Meuwissen et al., 2001; God-
40 dard, 2017). A deterministic formula such as the one proposed by Daetwyler et al. (2008)
41 highlights the relationship between prediction accuracy and potential factors influencing pre-
42 diction accuracy. In general, deterministic formulas compute expected predictive correlation
43 (or prediction R^2 of phenotypes) on the basis of a number of factors that are potentially
44 useful to assess prediction accuracy before performing computationally demanding cross-
45 validation (CV). It also allows us to decide the optimal design for training populations. Not
46 only theoretical derivations of deterministic formulas but also their applications are active
47 research areas. For instance, Brard and Ricard (2015) recently performed comparison and
48 meta-analysis of deterministic formulas. Erbe et al. (2013) inferred parameters that influence
49 prediction accuracy in deterministic formulas via a maximum likelihood. Collectively, these
50 studies have shed new light on alternative aspects of factors influencing predictive perfor-
51 mance that may not be obvious from empirical genome-enabled prediction analysis based on
52 CV.

53 In particular, visualizing such deterministic formulas may lead to a better understanding
54 of how genetic factors control prediction accuracy. Typically, visualization involves gener-
55 ating a static two-dimensional graph, where the y-axis is the genomic prediction accuracy
56 and the x-axis is one of the factors influencing prediction accuracy, while keeping the other
57 factors constant. Given that this type of static graph is a snapshot of complex dynamic
58 system, if users want to change parameters, they need to re-type and re-execute the code.
59 To overview the whole landscape of genomic prediction simulation, we need an efficient vi-
60 sualization tool that is capable for generating interactive as well as dynamic graphs. The
61 objective of this article is to describe a Shiny-based web application called ShinyGPAS (Shiny
62 Genomic Prediction Accuracy Simulator), which produces interactive graphs and offers an
63 intuitive graphical user interface (GUI) for simulating genomic prediction accuracy based on

64 deterministic formulas.

65 **Software description**

66 **Overview of software architecture**

67 ShinyGPAS is implemented entirely in R, which is an open source programming language
68 and environment for performing statistical computing and data visualization (R Core Team,
69 2017). The GUI is provided by the shiny R package (Chang et al., 2017), a web application
70 framework for R. ShinyGPAS is a Shiny application that leverages R and the shiny package
71 to construct an intuitive framework for deterministic formulas using dynamic interaction and
72 visualization. The ShinyGPAS user interface is shown in Figure 1. Although ShinyGPAS is
73 R-based software, it does not require users to either be familiar with the programming lan-
74 guage nor download the software on a local computer. The underlying R code is encapsulated
75 by Shiny and offered as cohesive web-based software to be usable solely by mouse navigation
76 in a web browser. This increases accessibility to the software, especially for users with less
77 R programming experience. ShinyGPAS is deployed through the cloud-based shinyapps.io
78 platform for hosting Shiny web applications (<https://www.shinyapps.io/>).

79 **Deterministic formulas**

80 ShinyGPAS currently delivers six simulators based on deterministic formulas described in
81 a) Daetwyler et al. (2008, 2010), b) Goddard (2009), c) Goddard et al. (2011), d) de los
82 Campos et al. (2013), e) Karaman et al. (2016), and f) Wientjes et al. (2016). The first five
83 formulas predict accuracy or squared prediction accuracy within populations whereas the
84 last one is designed for multipopulation including multienvironment and multitrait scenar-
85 ios. Deterministic formulas are functions derived from the combinations of the number of
86 individuals in a reference set, the number of independent chromosome segments underlying
87 the trait, the effective population size, the proportion of genetic variance explained by the
88 molecular markers, and heritability. Shiny-based interactive application offers the imple-
89 mentation of dynamic deterministic formulas, allowing to evaluate the simultaneous impact

90 of all the parameters described above on the degree of prediction accuracy. A user can click
91 a link located within each deterministic formula simulator to access original journal articles.
92 Below are deterministic formulas currently implemented in ShinyGPAS.

- Daetwyler et al. (2008, 2010)

$$r = \sqrt{\frac{Nh^2}{Nh^2 + M_e}}$$

93 where N is the number of individuals, h^2 is the heritability, and M_e is the number of
94 independent chromosome segments.

- Goddard (2009)

$$r = \sqrt{1 - \frac{\lambda}{2N\sqrt{\alpha}} \log\left(\frac{1 + \alpha + 2\sqrt{\alpha}}{1 + \alpha - 2\sqrt{\alpha}}\right)}$$

95 where λ is $M_e/(h^2 \log(2N_e))$, α is $1 + 2(M_e/Nh^2 \log(2N_e))$, and N_e is the effective
96 population size.

- Goddard et al. (2011)

$$r = \sqrt{b \frac{Nbh^2/M_e}{1 + Nbh^2/M_e}}$$

97 where b is the proportion of genetic variance explained by the markers.

- de los Campos et al. (2013)

$$R^2 = [1 - (1 - b)^2]h^2$$

98 where b is the average regression coefficient of the marker-based genomic relationships
99 on causal loci derived genomic relationships.

- Karaman et al. (2016)

$$R^2 = h_M^2 \left[\frac{Nh_M^2}{Nh_M^2 + M_e} \right]$$

100 where h_M^2 is the genomic heritability, which is the proportion of phenotypic variance
 101 that is explained by regression on markers.

- Wientjes et al. (2016)

$$r = \sqrt{\begin{bmatrix} b_{AC}r_{G_{AC}}\sqrt{\frac{h_A^2}{M_e}} & b_{BC}r_{G_{BC}}\sqrt{\frac{h_B^2}{M_e}} \end{bmatrix} \begin{bmatrix} \frac{h_A^2}{M_e} + \frac{1}{N_A} & r_{G_{AB}}\sqrt{\frac{h_A^2 h_B^2}{M_e}} \\ r_{G_{AB}}\sqrt{\frac{h_A^2 h_B^2}{M_e}} & \frac{h_B^2}{M_e} + \frac{1}{N_B} \end{bmatrix}^{-1} \begin{bmatrix} b_{AC}r_{G_{AC}}\sqrt{\frac{h_A^2}{M_e}} \\ b_{BC}r_{G_{BC}}\sqrt{\frac{h_B^2}{M_e}} \end{bmatrix}}$$

102 where b_{AC} is the square root of the proportion of the genetic variance in predicted
 103 population C explained by the markers in training population A, $r_{G_{AC}}$ is the genetic
 104 correlation between populations A and C, h_A^2 is heritability in population A, b_{BC} is
 105 the square root of the proportion of the genetic variance in predicted population C
 106 explained by the markers in training population B, $r_{G_{BC}}$ is the genetic correlation
 107 between populations B and C, h_B^2 is heritability in population B, N_A is the number of
 108 individuals in population A, N_B is the number of individuals in population B, and $r_{G_{AB}}$
 109 is the genetic correlation between populations A and B. This deterministic formula
 110 combines two populations A and B to predict prediction accuracy in population C
 111 (Wientjes et al., 2016).

112 Program input

113 A typical workflow starts from selecting one of the tab panels on the top (Figure 1) and
 114 then moving to a preferred deterministic formula simulator. Each of deterministic formula
 115 captures a different aspect of the genotype-phenotype map in the context of genomic predic-
 116 tion accuracy. Thus, navigating interactively visualized deterministic formulas may highlight

117 the common patterns as well as differences among them. A suite of available parameters
118 such as h^2 , h_M^2 , N , M_e , N_e , and b are located in the sidebar panel. Shiny slider provides
119 possible input values one can choose from pre-defined ranges. Users can pick a preferred
120 value by a simple mouse navigation. A radio button located on the bottom offers possible
121 options for factors influencing prediction accuracy to be used to determine the x-axis. The
122 Shiny reactive expressions are utilized in ShinyGPAS to efficiently cache results and ease
123 computational burden to ensure high speed of processing during an interactive session.

124 **Program output**

125 Rendering interactive graphs from deterministic formulas are achieved by the plotly R pack-
126 age (Sievert et al., 2017). The main engine plotly.js, which is built on top of JavaScript and
127 the visualization library D3.js, was used to create a scatter plot. The y-axis is pre-fixed with
128 prediction accuracy (r) or squared prediction accuracy (R^2). Users can choose the x-axis
129 from one of the parameters including h^2 , h_M^2 , N , M_e , N_e , or b . A scatter plot is dynamically
130 updated when users vary slider input values of factors influencing prediction accuracy. The
131 plotly.js generates a scatter plot with a toolbar coupled with useful zooming in and zooming
132 out capabilities. Also, hovering the mouse pointer over a specific point of plot shows the
133 exact values of x and y axes. A multipopulation genomic prediction simulation is enabled
134 by the plotly 3D scatter plot functionality, where x and y axes take parameters from two
135 training populations and z-axis shows prediction accuracy. Rotating the 3D scatter plot is
136 possible around all x, y and z axes to inspect prediction accuracy from different surfaces.
137 In addition, the toolbar provides features such as download button, box select, lasso select,
138 autoscale, reset, and toggle spike lines features for interactivity. ShinyGPAS is available at:
139 <https://chikudaisei.shinyapps.io/shinygpas/>.

140 Conclusions

141 A Shiny application has great potential to deliver interactive data analysis and visualiza-
142 tion in a web browser. Yet there is limited application of this type of tool in animal
143 breeding and quantitative genetics. The use of Shiny framework allows users to convert
144 deterministic formulas of genomic prediction accuracy into interactive graphics in an en-
145 gaging and straightforward manner. ShinyGPAS can be used for interactive exploration
146 of potential factors influencing prediction accuracy in genome-enabled prediction, simula-
147 tion of achievable prediction accuracy prior to genotyping individuals, or supporting in-
148 class teaching. The ShinyGPAS source code has been made publicly available on GitHub:
149 <https://github.com/morota/ShinyGPAS>.

150 **Declarations**

151 **Authors' contributions**

152 GM designed and developed the software, and wrote the manuscript.

153 **Funding**

154 This work was supported in part by the University of Nebraska startup funds to GM.

155 **Competing interests**

156 The authors declare that they have no competing interests.

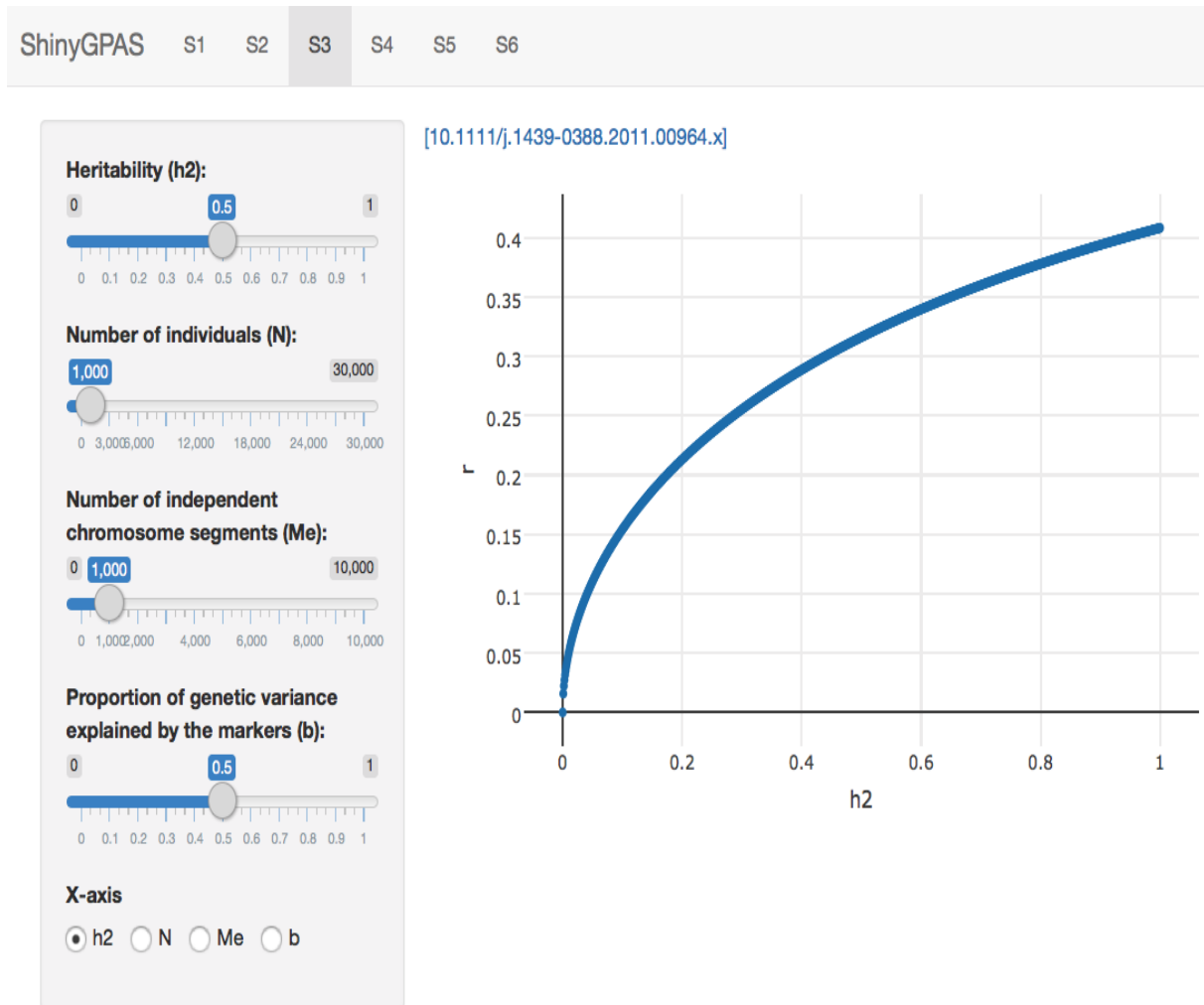


Figure 1: Each deterministic formula is implemented in a tab on the top. The y-axis is the prediction accuracy or squared prediction accuracy and the x-axis is one of the parameters. Parameters including heritability, the number of individuals, the number of independent chromosome segments, effective population size, and the proportion of genetic variance explained by the markers can be set by the user.

157 References

- 158 Brard, S. and Ricard, A. (2015). Is the use of formulae a reliable way to predict the accuracy
159 of genomic selection? *Journal of Animal Breeding and Genetics*, 132(3):207–217.
- 160 Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application*
161 *Framework for R*. R package version 1.0.3.
- 162 Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact
163 of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031.
- 164 Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the
165 genetic risk of disease using a genome-wide approach. *PloS one*, 3(10):e3395.
- 166 de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013).
167 Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS*
168 *Genet*, 9(7):e1003608.
- 169 Erbe, M., Gredler, B., Seefried, F. R., Bapst, B., and Simianer, H. (2013). A function
170 accounting for training set size and marker density to model the average accuracy of
171 genomic prediction. *PLoS One*, 8(12):e81046.
- 172 Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long
173 term response. *Genetica*, 136(2):245–257.
- 174 Goddard, M. (2017). Can we make genomic selection 100% accurate? *Journal of Animal*
175 *Breeding and Genetics*, 134(4):287–288.
- 176 Goddard, M., Hayes, B., and Meuwissen, T. (2011). Using the genomic relationship matrix
177 to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*,
178 128(6):409–421.
- 179 Karaman, E., Cheng, H., Firat, M. Z., Garrick, D. J., and Fernando, R. L. (2016). An upper
180 bound for accuracy of prediction using GBLUP. *PloS one*, 11(8):e0161054.

- 181 Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value
182 using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.
- 183 R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foun-
184 dation for Statistical Computing, Vienna, Austria.
- 185 Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy,
186 P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.6.0.
- 187 Wientjes, Y. C., Bijma, P., Veerkamp, R. F., and Calus, M. P. (2016). An equation to predict
188 the accuracy of genomic values by combining data from multiple traits, populations, or
189 environments. *Genetics*, 202(2):799–823.