Check for updates

Editorial: Statistical methods for analyzing multiple environmental quantitative genomic data

Zitong Li[1]* and Lucia Gutierrez[2]

[1]CSIRO Agriculture and Food, Canberra, ACT, Australia, [2]Department of Agronomy, University of Wisconsin-Madison, Madison, WI, United States

Editorial on the Research Topic
Statistical methods for analyzing multiple environmental quantitative genomic data

Phenotypic variation arises from the combined effects of genetic and environmental factors, including interactions between them (Lynch and Walsh, 1998). Studying the relationship between phenotypes, genotypes, and environments using sophisticated statistical models becomes a crucial Research Topic in quantitative genetics (Crossa et al., 2021). Recent advancements in high-throughput genotyping (Hu et al., 2021) and phenotyping (Gill et al., 2022) measurement techniques have enabled the acquisition of large-scale genomic, phenomic, and environmental data by quantitative geneticists. This Research Topic highlights several novel statistical analytical tools that can effectively leverage high-dimensional data to gain a deeper understanding of genotype-environment interactions (GEI) (Elias et al., 2016; van Eeuwijk et al., 2016) and use them to predict phenotype outcomes.

One important research direction in plant and animal breeding is genomic selection or genomic prediction (GP) (Meuwissen et al., 2001), a molecular breeding technique that uses genome-wide datasets to predict the genomic estimated breeding values (GEBV) or genotypic values of individuals for economically important traits. Many quantitative traits, such as yield, have a very complex genetic architecture (Doerge, 2002; Bernardo, 2016). Therefore, incorporating environmental data into the genomic prediction model and properly describing gene-environment interactions (GEI) is crucial to provide promising predictions of individual's performance. Different strategies have been used to accounting for GEI in GP models and it is still an area of active research (Jarquín et al., 2021). A natural modelling approach includes extending GP mixed models to incorporate a covariance relationship matrix between environments based on phenotypic information (Piepho, 1998; Burgueño et al., 2012; Lado et al., 2016; Malosetti et al., 2016) or environmental covariates (Jarquín et al., 2014). These can be included as either linear or non-linear kernels (Costa-Neto et al., 2020). Models based on observed covariance among environments cannot however predict the performance of individuals in untested environments (Heslot et al., 2014). An alternative to predict the performance of individuals in untested environments is to use environmental covariates in either partial-least squares regressions (Crossa et al., 1999; Monteverde et al., 2019) or for genotype-specific reaction norms in random regression models (Schaeffer, 2004; Buntaran et al., 2021) or P-splines (Bustos-Korts et al., 2021). Due

to high availability of environmental covariates and their highly correlated nature, not all environmental covariates are equally informative (Bustos-Korts et al., 2015) and variable selection has been proven useful in improving the performance (Neyhart et al., 2022).

Conventional prediction models, such as reaction norm model (Jarquín et al., 2014) or genomic best linear unbiased prediction model (G-BLUP) for GEI, face particular challenges when used to predict untested lines in new environments. As an improvement, Montesinos-López et al. demonstrated the use of partial least squares (PLS) regression approaches (Boulesteix and Strimmer, 2007) for conducting multiple environmental genomic prediction in 14 real data sets. The PLS method can simultaneously account for G, E and GEI effects for genomic prediction, and the multiple case studies have demonstrated that PLS can provide more accurate prediction compared to conventional GP methods for lines in new environments. Montesinos-López et al. further extended the PLS approach to a multivariate PLS regression that can simultaneously analyse multiple traits and it showed supervisor prediction performance to single trait PLS as well as G-BLUP because the MPLS approach can account for the correlation among traits and can therefore borrow strength from each other during the analysis.

When incorporating more data into a genomic prediction model, the numerical computation for parameter estimation may become infeasible. To overcome this challenge, Manthena et al. evaluated a series of dimensional reduction methods such as random projection, random and deterministic sampling, and shrinkage methods which were applied to reduce the dimension of the SNP data ahead of the multiple environment GP analyses. The study demonstrated that some of these methods were effective not only on reducing the computational cost, but also can maintain and sometimes even improve the predictability. However, the paper also concludes that there is no dimensional reduction approach which can constantly outperform other methods across data sets. Future efforts are needed to develop more robust dimensional reduction methods compile with the genomic prediction. Additionally, dimension reduction can also be conducted guided by linkage disequilibrium (LD) (Slatkin, 2008) or the correlation structure among loci. Jin et al. developed a LD network approach to model the correlation among genome-wide markers and cluster them into LD blocks using an efficient sparse graphical learning approach, and the dimension reduction within each LD block using classical principal component analysis. Interestingly, this approach is initially proposed for studying local adaptation using population genomics (Jones et al., 2012) data, but can also be applicable as a tool for dimensional reduction for GP data.

GP models such as reaction norm model and PLS are able to predict outcomes on the basis of GEI, but they cannot be used to identify genes that are associated with GEI. For the gene discovery purpose, Onogi et al. developed a data driven approach named Environmental Covariate Search Affecting Genetic Correlations (ECGC). The ECGC firstly calculated the genetic covariance between the pairwise environments, and then considered the correlation coefficients as the "trait" in the genome-wide association study to identify significant SNPs associated with the environmental stimuli. The ECGC approach was applied on a large-scale soybean data set, which yielded biological meaningful results.

As a conclusion, this Research Topic collects a series of modern quantitative genomic methods that can effectively analyse large-scale genomic, phenomic and environmental data sets, with the aim to either predict individuals' outcome of quantitative traits or to identify important genes that are linked to genotype by environment interactions. We are hopeful that these new analytical tools can provide useful additions to the existing quantitative genetic methods for analysing high dimensional biological data sets and can also inspire new research development in this existing research area, especially to meet challenges of big data arising in this post-genomic era.

Author contributions

ZL and LG conceived the study questions and designed the research. ZL and LG drafted or critically revised significant parts of the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bernardo, R. (2016). Bandwagons I, too, have known. *Theor. Appl. Genet.* 129 (12), 2323–2332. doi:10.1007/s00122-016-2772-5

Boulesteix, A. L., and Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brieflings Bioinforma.* 8, 32–44. doi:10.1093/bib/bbl016

Buntaran, H., Forkman, J., and Piepho, H.-P. (2021). Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: Accuracy and precision. *Theor. Appl. Genet.* 134 (5), 1513–1530. doi:10.1007/s00122-021-03786-2

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52 (2), 707–719. doi:10.2135/cropsci2011.06.0299

Bustos-Korts, D., Boer, M. P., Chenu, K., Zheng, B., Chapman, S., and van Eeuwijk, F. A. (2021). Genotype-specific P-spline response surfaces assist interpretation of regional wheat adaptation to climate change. *silico Plants* 3 (2), diab018. doi:10.1093/insilicoplants/diab018

Bustos-Korts, D., Malosetti, M., Chapman, S., and van Eeuwijk, F. (2015). "Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics," in *Crop*

Systems Biology: Narrowing the gaps between crop modelling and genetics. Editors X. Yin, and P. C. Struik (Cham: Springer), 55–82.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106. doi:10.1038/s41437-020-00353-1

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: Optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12, 651480. doi:10.3389/fpls.2021.651480

Crossa, J., Vargas, M., van Eeuwijk, F. A., Jiang, C., Edmeades, G. O., and Hoisington, D. (1999). Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theor. Appl. Genet.* 99, 611–625. doi:10.1007/s001220051276

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* 3, 43–52. doi:10.1038/nrg703

Elias, A. A., Robbins, K. R., Doerge, R. W., and Tuinstra, M. R. (2016). Half a century of studying genotype × environment interactions in plant breeding experiments. *Crop Sci.* 56, 2090–2105. doi:10.2135/cropsci2015.01.0061

Gill, T., Gill, S. K., Saini, D. K., Chopra, Y., de Koff, J. P., and Sandhu, K. S. (2022). A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics* 2, 156–183. doi:10.1007/s43657-022-00048-z

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127 (2), 463–480. doi:10.1007/s00122-013-2231-5

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum. Immunol.* 82, 801–811. doi:10.1016/j.humimm.2021.02.012

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucour, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1

Jarquín, D., de Leon, N., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I., et al. (2021). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front. Genet.* 11, 592769–592811. doi:10.3389/fgene.2020.592769

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61. doi:10.1038/nature10944

Lado, B., Barrios, P. G., Quincke, M., Silva, P., and Gutiérrez, L. (2016). Modeling genotype × Environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* 56 (5), 2165–2179. doi:10.2135/cropsci2015.04.0207

Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits.* Sunderland, MA: Sinauer Associates.

Malosetti, M., Bustos-Korts, D., Boer, M. P., and van Eeuwijk, F. A. (2016). Predicting responses in multiple environments: Issues in relation to genotype × environment interactions. *Crop Sci.* 56 (5), 2210–2222. doi:10.2135/cropsci2015.05.0311

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Monteverde, E., Gutierrez, L., Blanco, P., Pérez de Vida, F., Rosas, J. E., Bonnecarrère, V., et al. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (oryza sativa L) grown in subtropical areas. *G3 Genes.|Genomes|Genetics* 9 (5), 1519–1531. doi:10.1534/g3.119.400064

Neyhart, J. L., Silverstein, K. A. T., and Smith, K. P. (2022). Accurate predictions of barley phenotypes using genomewide markers and environmental covariates. *Crop Sci.* 62, 1821–1833. doi:10.1002/csc2.20782

Piepho, H-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theor. Appl. Genet.* 97, 195–201. doi:10.1007/s001220050885

Schaeffer, L. R. (2004). Application of random regression models in animal breeding. *Livest. Prod. Sci.* 86 (1–3), 35–45. doi:10.1016/S0301-6226(03)00151-9

Slatkin, M. (2008). Linkage disequilibrium — Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. doi:10.1038/nrg2361

van Eeuwijk, F. A., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × Environment interactions? *Crop Sci.* 56 (5), 2119–2140. doi:10.2135/cropsci2015.06.0375