PERSPECTIVE

GLOBAL GENOMIC PREDICTION IN HORTICULTURAL CROPS: PROMISES, PROGRESS, CHALLENGES AND OUTLOOK

Craig HARDNER (✉)[1], Satish KUMAR[2], Dorrie MAIN[3], Cameron PEACE[3]

1 Queensland Alliance for Agriculture and Food Innovation University of Queensland, St Lucia QLD, 4072, Australia.
2 New Zealand Institute Plant & Food Res Ltd, Hawkes Bay Research Centre, Havelock North, 4130, New Zealand.
3 Department of Horticulture, Washington State University, Pullman, WA 99164, USA.

Received January 25, 2021

Correspondence: c.hardner@uq.edu.au

Only when all contribute their firewood can they build up a big fire (众人拾柴火焰高).

Horticultural crops are a major source of high value nutritious food, and new improved cultivars developed through breeding are required for sustainable production in the face of abiotic and biotic stresses, and to deliver novel, premium products to consumers. However, grower confidence in the performance of new germplasm, particularly across environmental variability, is important for commercial adoption and germplasm-environment matching to optimize production.

Knowledge on the stability of germplasm across environments for cultivar selection and breeding in horticultural crops is limited. Normally, evaluation of replicated (usually clonally) germplasm over several locations is done to generate information on germplasm-environment matching, however, these experiments are expensive in horticultural tree crops due to the large size and longevity of the experimental unit, the need for replication within and across sites, and high costs of phenotyping. In the absence of information for genotype-environment matching, breeding tends to focus on developing domestic cultivars for local environments using small experiments with limited or no overlap of germplasm between programs. In addition, confidence in the performance of imported germplasm in exotic regions is generally developed by assuming germplasm will be stable across environments, or from ad hoc knowledge accumulated by local risk-taking early adopters.

Genomic prediction might provide an efficient approach for combining historical data across breeding programs and selection trials to increase confidence in the performance of domestic germplasm in local environments, and imported germplasm in exotic environments. Genomic prediction exploits linkage between markers and QTLs in a training population (genotyped using genome-wide marker and phenotyped for traits of interest) to develop a prediction model that can then be applied to genotypic data for a selection population to predict genetic performance in the absence of direct phenotypic data[1]. Usually, with the number of markers being greater than the number of germplasm entries (i.e., number of markers >> number of observations), a distribution of marker effects is assumed to reduce the number of parameters requiring estimation. A common distributional assumption is Gaussian (i.e., all loci have small effect). This model is equivalent to the standard definition of a quantitative trait, and the genomic realized relationship matrix (GRRM) estimated from the marker similarities is equivalent to direct modeling of individual marker effects on phenotype using this distribution[2]. One advantage of the GRRM is that it can be directly incorporated into established linear mixed model approaches that are more flexible than the Bayesian approaches required to model other marker effect distributions. Given that the GRRM models track replication of individual alleles across germplasm, rather than replication of the whole genomes which occurs when individuals are vegetative propagated, the models can be used to connect unreplicated germplasm within and across locations to improve prediction accuracy in local and exotic environments. Therefore, here we define the term *global genomic prediction* to describe the hypothesis that "historical phenotypic data from multiple breeding programs are a sample of the experienced environment and their data sets can be connected through genome-wide multivariate prediction models"[3].

Several studies have evaluated global genomic prediction in horticultural tree crops. The first study was undertaken using data on crispness of apples evaluated across 662 (unreplicated) entries across three sites in the USA (Wenatchee, WA; St Paul, MN; and Geneva, NY) that had been genotyped using an 8K SNP array[4]. A GRRM was estimated from the SNP data, with each site-by-trait-by-age considered as a unique attribute (*sensu* Falconer[5]), and a factor analytic model was used to estimate the entry-by-attribute covariance matrix. This multivariate approach allowed the combining of data collected by different methods. The analysis suggested genomic effects for texture were highly correlated across the Geneva and Wenatchee sites, but were less well correlated with effects at St Paul. A subsequent study in sweet cherry [6] with 597 entries phenotyped for timing of fruit maturity across four sites (Prosser, WA, USA; Balandran and Bourran, France; and Forli, Italy) (and two years at each site) using a similar statistical model estimated high additive genomic correlation (>0.9) among environments and high prediction accuracy (0.9 to 1.0). In recent unpublished research on peach with 506 entries phenotyped for sweetness across ten environments (four sites and multiple years within sites), we observed an increase of about 20% in prediction accuracy by combining data into a single analysis compared to only using data from a single location to train the prediction model.

Several challenges to implementing global genomic prediction have been encountered. Considerable effort is required to collate, standardize, and curate datasets contributed from different sources. Names of genetically identical entries need to be standardized, particularly if language differences exist. The identity of SNP loci and alleles across multiple datasets needs to be well defined, particularly if SNP genotypes are to be imputed across different genotyping platforms to increase marker density. Correct description of complex sampling designs in different environments is required (e.g., Hardner et al.[7]) to reduce non-genetic phenotypic noise. Often collinearity is observed in the GRRM, and therefore bending (i.e., replacing negative Eigen values with slightly positive values) is required to obtain model solutions. Confounding of population structure with testing location might also be an issue for some datasets, with further research is required to develop models that account for heterogeneity in allele frequencies across subpopulations.

Several challenges also exist for translation of global genomic prediction to genetic improvement in practice. While genome-assisted parental selection (GAPS) using additive effects is a credible alternative to phenotypic selections[8], performance of GAPS across different sites/environments remains to be investigated. For prediction of clonal values (i.e., predicted phenotype based on total genetic variation captured by vegetative propagation of superior individuals – which is the important for identifying elite cultivars), relationship matrices of non-additive genetic effects are generally more sparse. Thus, the genetic architecture of non-additive effects is likely to be estimated with less precision and might be confounded with additive genetic effects, meaning the accuracy of predicted clonal values is expected to be lower than for breeding values. While accuracy of breeding values was not greatly improved by inclusion of non-additive effects, bias was reduced[9]. Global genomic prediction supports prediction of performance of entries into environments in which they have not been tested through the use of correlated genetic effects, however, interpretation of these predictions needs qualification for traits that require expression of other traits (e.g., expression of fruit texture requires trees to flower and set fruit). Lastly, the intellectual property of contributors of data and germplasm to combined datasets needs to be respected and protected.

Global genomic prediction fundamentally relies on collaboration to leverage, and add to, the latent value of existing data of individual breeding and selection programs. A preliminary online portal has been developed with the Genome Database for Rosaceae to predict peach sweetness at four sites in the USA (Fresno, CA; College Station, TX; Clarksville, AR; and Seneca, SC) using genome-wide genotypic data uploaded by any user into the unpublished model used for the analysis of peach described above. Although crops from the family Rosaceae have been the focus of initial development, this approach can be readily applied to other crops. The approach is being extended into other traits and larger datasets, improving understanding of the drivers of genotype-by-environment interaction and enabling practical tools in genetic improvement for responding to climate change.

Acknowledgements

The support of the USDA NIFA Specialty Crop Research Initiative project “RosBREED 2: Combining disease resistance with horticultural quality in new rosaceous cultivars” (2014-51181-22378) and National Tree Genomics Program—Phenotype Prediction (AS17000) funded by the Hort Frontiers Advanced Production Systems Fund, part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from The University of Queensland, Queensland State Government, and contributions from the Australian Government is acknowledged.

REFERENCES

1. Meuwissen T H E, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 2001, **157**(4): 1819–1829
2. Hayes B J, Visscher P M, Goddard M E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetical Research,* 2009, **91**(1): 47–60
3. Iezzoni A F, McFerson J, Luby J, Gasic K, Whitaker V, Bassil N, Yue C, Gallardo K, McCracken V, Coe M, Hardner C, Zurn J D, Hokanson S, van de Weg E, Jung S, Main D, da Silva Linge C, Vanderzande S, Davis T M, Mahoney L L, Finn C, Peace C. RosBREED: bridging the chasm between discovery and application to enable DNA-informed breeding in rosaceous crops. *Horticulture Research,* 2020, **7**(1): 177
4. Hardner C M, Kumar S, Peace C M, Luby J, Evans K M. Reconstructing relationship matrices from dense SNP arrays for the prediction of genetic potential in unreplicated multilocation plantings of apple progeny. In: Onus N, Currie A, eds. *Xxix International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes,* 2016, 275–281
5. Falconer D S. The problem of environment and selection. *American Naturalist,* 1952, **86**(830): 293–298
6. Hardner C M, Hayes B J, Kumar S, Vanderzande S, Cai L, Piaskowski J, Quero-Garcia J, Campoy J A, Barreneche T, Giovannini D, Liverani A, Charlot G, Villamil-Castro M, Oraguzie N, Peace C P. Prediction of genetic value for sweet cherry fruit maturity among environments using a 6K SNP array. *Horticulture Research,* 2019, **6**(1): 6
7. Hardner C M, Evans K, Brien C, Bliss F, Peace C. Genetic architecture of apple fruit quality traits following storage and implications for genetic improvement. *Tree Genetics & Genomes*, 2016, **12**(2): 20
8. Kumar S, Molloy C, Munoz P, Daetwyler H, Chagne D, Volz R. Genome-enabled estimates of additive and nonadditive genetic variances and prediction of apple phenotypes across environments. *G3-Genes Genomes Genetics,* 2015, **5**(12): 2711–2718
9. Kumar S, Hilario E, Deng C H, Molloy C. Turbocharging introgression breeding of perennial fruit crops: a case study on apple. *Horticulture Research,* 2020, **7**(1): 47