Field Crops Research 154 (2013) 12–22

Contents lists available at ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr

The use of unbalanced historical data for genomic selection in an international wheat breeding program

Julie C. Dawson[a,*], Jeffrey B. Endelman[a], Nicolas Heslot[a,b], Jose Crossa[d], Jesse Poland[c], Susanne Dreisigacker[d], Yann Manès[d,1], Mark E. Sorrells[a], Jean-Luc Jannink[a,e]

[a] Department of Plant Breeding and Genetics, Cornell University, 240 Emerson Hall, Ithaca, NY 14853, United States
[b] Limagrain Europe, CS3911, Chappes 63720, France
[c] USDA-ARS and Department of Agronomy, Kansas State University (KSU), 4011 Throckmorton Hall, Manhattan, KS 66506, United States
[d] International Maize and Wheat Improvement Center (CIMMYT), Int. Apdo. Postal 6-641, 06600 Mexico, DF, Mexico
[e] USDA-ARS R.W. Holley Center, Cornell University, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:
Received 26 April 2013
Received in revised form 25 July 2013
Accepted 25 July 2013

Keywords:
Genomic selection
Genotype-by-environment interaction
Wheat breeding
Unbalanced data
Mega-environments

ABSTRACT

Genomic selection (GS) offers breeders the possibility of using historic data and unbalanced breeding trials to form training populations for predicting the performance of new lines. However, when using datasets that are unbalanced over time and space, there is increasing exposure to different genotype – environment combinations and interactions that may make predictions less accurate. Global cross-validated genomic prediction accuracies may be high when using large historic datasets but accuracies for individual years using a forward-prediction approach, or accuracies for individual locations, are often much lower. The objective of this study was to evaluate the overall accuracy of genomic predictions for untested genotypes using an unbalanced dataset to train a genomic selection model, and to explore ways of combining genomic selection and genotype-by-environment (G×E) interaction models to better target untested lines to different locations. Using the International Center for Maize and Wheat Improvement's (CIMMYT) Semi-Arid Wheat Yield Trials (SAWYT) we assessed the accuracy of genomic predictions and the potential to subset these nurseries using the concept of mega-environments (ME) adapted to a genomic selection context. We found that there was no difference in accuracy between models accounting for G×E interactions and global models. Data-driven methods of clustering locations based on similarities in genomic predictions also failed to improve accuracies within clusters. Using a simulation based on the empirical SAWYT data, we found that if there were different true genotypic values between clusters, there was an advantage to modeling G×E in prediction models. In the SAWYT dataset it appears that there is not a consistent pattern of genotype-by-environment interaction among the ME, and this dataset is not balanced enough to partition into new clusters that have predictive power.

1. Introduction

1.1. Potential of genomic data to improve the utility of unbalanced historical datasets

The ubiquity of unbalanced historic datasets in plant breeding programs is a longstanding challenge. Breeding lines are selected and promising lines are advanced to a point where enough seed is available for multi-locational trials of candidates for release. Other than possibly one or two long term checks, the entries in these trials change yearly. The use of phenotypic data from relatives and ancestors has been limited by the challenge of maintaining adequate pedigree records and the expense of obtaining genotypic data on hundreds of lines at early stages in a breeding program.

This situation is rapidly changing, however, as new genotyping platforms have made it possible to obtain high-density markers at very low cost (Elshire et al., 2011; Poland and Rife, 2012). Software

Abbreviations: CIMMYT, International Center for Maize and Wheat Improvement; FA, factor analytic model; G-BLUP, genotypic covariance matrix-best linear unbiased predictor model; GID, genotypic identification number; G×E, genotype by environment; GL, global model; GS, genomic selection; IN, interaction model; ME, mega-environment; SAWYT, semi-arid wheat yield trial; SP, cluster-specific model.

* Corresponding author.
E-mail address: jcd11@cornell.edu (J.C. Dawson).
[1] Present address: Syngenta Seeds Cereal Research and Development, Paris, France.

0378-4290
http://dx.doi.org/10.1016/j.fcr.2013.07.020

and database tools have become available that make it possible to keep track of phenotypic, genotypic and pedigree records for thousand of individuals over many years. Because of improvements in genotyping and statistical methods that can handle this type of data, genomic selection has become possible for many breeding programs. Genomic selection makes predictions of performance for new lines or improves estimates of performance for these lines by using phenotypic and genomic data from related genotypes (Heffner et al., 2009; Lorenz et al., 2011). While there is much interest in using genomic selection and historical data to improve current selection programs, this has not yet been put into practice because of logistical difficulties in assembling data and questions about the best data and sets of genotypes ("training populations") to use in model training for individual breeding programs or target environments. The challenges are greater for breeding programs that span very diverse environmental conditions, as these programs are faced with high levels of G×E interactions, both across years within a target population of environments, and among target populations of environments.

The question of how to use historical data most effectively in the presence of large G×E interactions is particularly relevant for the international breeding programs of the Consultative Group on International Agricultural Research (CGIAR). These breeding programs typically conduct selection in a limited number of locations and then distribute new breeding lines and varieties for use by national agricultural research and breeding programs in diverse countries and regions. International selection programs could benefit from being better able to use data returned by international collaborators, especially if data could be used to define regions with different patterns of genotypic performance to better target particular environmental conditions. National breeding programs could greatly benefit from using data from other programs with similar environmental conditions, combined with their own historical data and the international center data, in order to increase their power to detect superior lines for their target environments.

The objective of this study was to assess the accuracy of genomic predictions in a large unbalanced dataset. We used CIMMYT's international semi-arid wheat yield trial (SAWYT), with data reported on grain yield for genotypes sent out by CIMMYT over a period of 17 years. See Appendix A for more information and references on the CIMMYT international yield trials. We first looked at the global accuracy and variation in accuracy over time. Then we tested different methods of accounting for G×E interaction when making genomic predictions. This included the most common methods for including a G×E component of variation in classic phenotypic analysis, with the inclusion of genomic data to address the issue of unbalanced genotypes in the trials over time (background information on G×E interaction analysis is presented in Appendix A). We also used simulated data to examine how different methods of accounting for G×E responded to changing levels of genotypic balance and G×E in multi-year, multi-locational trials. Our goal was to evaluate prediction models that could enable international breeding programs to target lines from their selection candidate nurseries to particular types of environments using information from related genotypes in international trials.

2. Materials and methods

2.1. Genotypic data

The wheat genotypes included in SAWYT 1–17, indexed by their genotypic identification number (GID), were characterized using genotyping-by-sequencing following the same procedure as described in Poland et al. (2012). A total of 45,818 SNP markers were obtained, and 34,843 were retained with a maximum of 70% missing data for each individual marker. The marker-based, additive relationship matrix ($\mathbf{A}_m$) for the 622 genotypes was calculated with the function A.mat in R package rrBLUP, version 4.1 (R Development Core Team, 2012; Endelman, 2011), which centers (but does not standardize) each marker by the population mean (VanRaden, 2008). Missing data were imputed with the "EM" option in A.mat, which implements a multivariate normal expectation-maximization (EM) algorithm (details in Poland and Rife, 2012).

2.2. Data curation

Phenotypic data for the SAWYT was obtained from the CIMMYT bioinformatics unit after initial data cleaning to remove outliers. Yield was the most complete trait; out of a total of 723 trials in the dataset, yield was measured in 611 individual trials and 237 unique locations over 17 years (planted in years 1992–2009, excluding 1993 because no SAWYT was sent out that year). Each year a separate set of lines was sent to international collaborators who requested seed for the trial. Most trials in the SAWYT had three replications in the first year (1992), two replications in years 2–5 (1994–1997) and two replications with incomplete blocks within reps starting with the sixth year. Data were curated to keep only genotypes with GBS marker data available (622 total). Repeated checks (Dharwar Dry and Cham 6) were eliminated in all years subsequent to their first occurrence because a single check was not considered adequate to characterize environmental conditions or G×E over the 17 years of data available.

Two criteria were used to identify and eliminate trials with errors in matching genotypes and phenotypes. The ratio $V_a/(V_a+V_e)$, the proportion of variance due to additive genetic effects, was used to eliminate trials where genotypes did not match phenotypic data. Trials with this ratio less than 0.01 were considered to have errors. Variance components V_a & V_e were calculated using the relationship matrix $\mathbf{A}_m$ with the kin.blup function in the package rrBLUP (Endelman, 2011). Replicated trials were also curated based on having $V_g/(V_g+V_e)$ greater than 0.01. Variance components V_g & V_e were calculated using the lmer function in the R package lme4 (Bates et al., 2012). Low values were likely due to mismatched plots of the same genotype from the different reps of the trial. This ratio is referred to as the repeatability or broad-sense heritability, the proportion of variance due to all genetic variance effects, calculated with variance components assuming independent genotypes. This step was used only to eliminate trials that must have had errors in data reporting, and thus near-zero values of these ratios. The final dataset had a total of 168 unique locations and 424 individual trials.

2.3. Clustering methods for grouping similar locations

Because global genomic predictions may not be the most relevant to individual breeding programs in particular environments, several methods were used to attempt to group locations into similar environmental clusters to improve the accuracy of genomic predictions within each cluster. CIMMYT has defined global mega-environments (ME) using climatic patterns, farming systems, water regimes, and the incidence of biotic and abiotic stress in the major wheat growing regions of the world. Trials were assigned to ME by breeders and the bioinformatics group at CIMMYT. A complete description of these classifications is available in Rajaram et al. (1993). The ME present in the SAWYT database, as well as their average yields and yield variability, are shown in Table 1. ME 3, with acidic soil, was only present in two years at one location in the SAWYT dataset and so was eliminated. The remaining ME were represented in at least 16 of the 17 years. ME 2/4, alternating high and low rainfall, was considered to be part of ME 2 to balance the ME representation across years, and the subsets 4A, 4B and 4C of