BioSystems 146 (2016) 102–109

Contents lists available at ScienceDirect

BioSystems

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

# Phenomic prediction of maize hybrids

Christian Edlich-Muth[a,b,*], Moses M. Muraya[c,1], Thomas Altmann[c], Joachim Selbig[a,b]

[a] Bioinformatics Group, Institute for Biochemistry and Biology, University of Potsdam, 14476, Germany
[b] Max Planck Institute of Molecular Plant Physiology, Potsdam 14476, Germany
[c] Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Stadt Seeland, Germany

ARTICLE INFO

*Article history:*
Received 22 February 2016
Received in revised form 13 May 2016
Accepted 16 May 2016
Available online 19 May 2016

*Keywords:*
Hybrid prediction
LASSO
Regression
Maize
Phenomics

ABSTRACT

Phenomic experiments are carried out in large-scale plant phenotyping facilities that acquire a large number of pictures of hundreds of plants simultaneously. With the aid of automated image processing, the data are converted into genotype-feature matrices that cover many consecutive days of development. Here, we explore the possibility of predicting the biomass of the fully grown plant from early developmental stage image-derived features. We performed phenomic experiments on 195 inbred and 382 hybrid maizes varieties and followed their progress from 16 days after sowing (DAS) to 48 DAS with 129 image-derived features. By applying sparse regression methods, we show that 73% of the variance in hybrid fresh weight of fully-grown plants is explained by about 20 features at the three-leaf-stage or earlier. Dry weight prediction explained over 90% of the variance. When phenomic features of parental inbred lines were used as predictors of hybrid biomass, the proportion of variance explained was 42 and 45%, for fresh weight and dry weight models consisting of 35 and 36 features, respectively. These models were very robust, showing only a small amount of variation in performance over the time scale of the experiment. We also examined mid-parent heterosis in phenomic features. Feature heterosis displayed a large degree of variance which resulted in prediction performance that was less robust than models of either parental or hybrid predictors. Our results show that phenomic prediction is a viable alternative to genomic and metabolic prediction of hybrid performance. In particular, the utility of early-stage parental lines is very encouraging.



## 1. Introduction

Today's highest-yielding maize varieties are hybrids that are obtained in a single cross of two elite inbred parental lines (Fig. 1a). The parental lines are usually chosen from two complementary genetic pools (here 'Dent' and 'Flint') and are selfed for several generations so that they contain two identical copies of each chromosome. Thus crossing two parental lines results in an (almost) genetically identical F1 generation. Maize hybrids strongly exhibit the phenomenon of heterosis (Shull, 1908), defined as the improved performance in a given trait by the hybrid compared to its parents (see Eq. (1)). As Fig. 1b illustrates, hybrids produce on average two times more biomass than inbred lines. Heterosis and single-cross hybrids have been exploited commercially for a century, and have contributed about half of the increase in yield in that period, the other half resulting from improvements in farming technology (Duvick, 2005). Nowadays, commercial breeders develop a large number of new parental lines each year, resulting in an ever-growing panel of parental lines that can be crossed with one another. Since the number of potential hybrids is the product of the number of parental lines, exploring 'hybrid space' (Fig. 1c) experimentally is infeasible. Thus computational methods have to step into the breach and extend the knowledge that is available for a small number of hybrids (blue squares in Fig. 1c) to all of hybrid space. This approach is called hybrid prediction. Its task is to find the best combinations of parental lines given a series of measurements on all parental lines and a test set of hybrids that have been evaluated for a trait of (economic) interest. In the setting that we are going to describe, the trait of interest is biomass, dry weight (DW) or fresh weight (FW), while the measurements are multi-dimensional times series (MTS) where the variables are phenomic image-derived features. In the framework of regression, the MTS are the predictors and the biomass the response.

The MTS are obtained in an automated fashion in a greenhouse where the plants travel on carriers into an imaging chamber where

* Corresponding author at: Max Planck Institute of Molecular Plant Physiology, Potsdam 14476, Germany.
*E-mail address:* edlich@mpimp-golm.mpg.de (C. Edlich-Muth).
[1] Present address: Department of Plant Sciences, Chuka University, PO Box 60400 Chuka, Kenya.

http://dx.doi.org/10.1016/j.biosystems.2016.05.008
0303-2647/

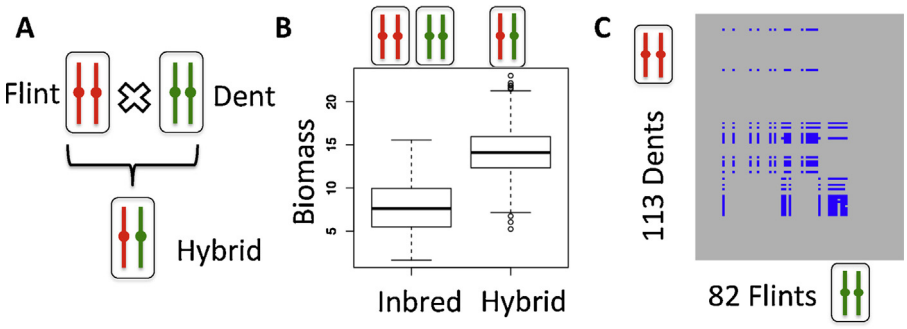


**Fig. 1.** The hybrid principle. (a) A Dent and a Flint inbred line (red and green respectively) combine in a cross to give rise to a hybrid. Both Dent and Flint are taken to be homozygous for each locus, therefore each pair of chromosomes are identical. (b) Hybrid vigour (heterosis). Hybrids are vastly superior in performance measures such as biomass. (c) Hybrid space. The panel of nearly 200 inbred lines could give rise to about 10,000 hybrids. The blue dots on the grey background represent the 400 hybrids that were experimentally tested. Thus only 4% of hybrid space is probed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

cameras that operate at different spectral ranges (ultra-violet, visible and infra-red) take pictures from several angles (top, side) every day, for the duration of the experiment (here 33 days). This experimental setup is often termed phenomics since, like other -omics data, the result is a large number of measurements that are recorded in parallel without human interference. The actual features of the MTS are obtained by image processing routines, and describe for example the size of convex hull of the plant or the number of leaves.

Phenomic MTS have not previously been used as predictors of hybrid biomass, in contrast to genomics and metabolomics data (Schrag et al., 2010; Riedelsheimer et al., 2012; Technow et al., 2014). Among the methods that have been used for genomic prediction (Ogutu and Piepho, 2014) the LASSO (Tibshirani, 1994) has the advantage that it carries out feature selection and model fitting at the same time. Moreover, extensions of the LASSO have been developed that allow the joint selection of a group of variables (Yuan and Lin, 2006). This is of particular interest here because the MTS can be represented as grouped data, where the group membership is determined by the correlation structure of the features.

In the following, we will explore several avenues for carrying out sparse regression with the LASSO, using phenomic MTS as predictors and with the aim of predicting hybrid biomass as response. Importantly, methods are presented for both hybrid-from-hybrid and hybrid-from-inbred prediction. We also investigate whether heterosis of early phenotypic traits is predictive of final biomass.

## 2. Materials and methods

### 2.1. Plant material and datasets

The data were collected in an automated greenhouse facility where every sample was phenotyped almost every day. There were four greenhouse experiments (labelled '1244', '1403', '1304' and '1343'), two with inbred (1244, 1403) and two with hybrid genotypes (1304, 1343). Inbred experiments consisted of 195 genotypes with two samples each. Hybrid experiments consisted of 382 genotypes with one sample each. Phenotyping started 16 days after sowing (DAS) and continued until DAS 48, by which time the plants were about 2.5 m high and had almost reached their maximal size. We therefore refer to plants at DAS 48 as "fully grown". Every measurement of a carrier on a particular day resulted in a set of 286 features. The experiments were carried out at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, Germany. Air temperature, light hours and watering scheme were identical in all experiments. In contrast to conventional greenhouses, the carriers with the plants also change position every day, so that there is no positional (block or otherwise) effect.

Initially, every carrier contained four plants of the same genotype. Two plants were harvested at the three-leaf stage ('harvest-1') and their fresh weight (FW) and dry weight (DW) were manually determined. Harvest-1 was carried out between 27 and 29 DAS. All experiments were terminated on DAS 48, when again DW and FW were determined ('harvest-2'). The raw data are available at https://edal.ipk-gatersleben.de (Arend et al., 2014).

The acquired images were automatically processed as described previously (Klukas et al., 2012, 2014) resulting in a multivariate time-series (MTS) over 33 days (16–48 DAS) and 286 features. A detailed description of all features can be found at http://iap.ipk-gatersleben.de/documentation.pdf. Since harvest-1 of two plants took place in the middle of each experiment, the time series are effectively split into an early and late phase (before and after harvest-1, Fig. 3).

Missing time points in the two time half-series were imputed by locally weighted polynomial regression using the R-function `loess` (Cleveland and Shyu, 1992). Any remaining points were imputed with a version of k-Nearest-Neighbour (`knn`) optimised for large matrices (Troyanskaya et al., 2001) with $k = 3$. Next, all samples of the same genotype within a dataset were aggregated by the median. Finally, the two experiments were combined by averaging. The resulting dataset is best thought of as a $n \times p \times q$ cube spanned by genotype, feature and time (see Fig. 2a).

To reduce the size of the datasets and simultaneously reduce noise in the time series, three consecutive and non-overlapping time points were averaged by the mean. Each cube at that stage contained 9 time points (19, 22, 25, 31, 34, 37, 40, 43 and 48 DAS) where e.g. 19 DAS is the average of the time points 18, 19 and 20 DAS. Note that the intervals spanning harvest-1 (DAS 27 to 29) were removed.

Inbred and hybrid MTS were processed in the same manner resulting in two data cubes, $\mathcal{I}$ and $\mathcal{H}$. In addition, we computed a third cube named $\mathcal{J}$ by applying the formula for mid-parent heterosis (Hallauer et al., 2010):

$$\mathcal{J}_{ijk} = \frac{\text{hybrid value}}{\text{average parental value}} = \frac{2\mathcal{H}_{ijk}}{\mathcal{I}_{djk} + \mathcal{I}_{fjk}}, \quad (1)$$

where $\mathcal{J}_{ijk}$ denotes mid-parent heterosis of hybrid $i$ in feature $j$ at time $k$. $\mathcal{H}_{ijk}$ is the value of hybrid $i$ in feature $j$ at time $k$. The two expressions in the denominator, $\mathcal{I}_{djk}$ and $\mathcal{I}_{fjk}$, are the values of the two parental lines of hybrid $i$, denoted by the subscripts $d$ and $f$ (for Dent and Flint); $\mathcal{I}$ is the inbred cube with the same feature and