Leveraging breeding values obtained from random regression models for genetic inference of longitudinal traits

Malachy Campbell[1], Harkamal Walia[1], and Gota Morota[2]

[1]Department of Agronomy and Horticulture, University of Nebraska Lincoln, Lincoln, NE, USA 68583

[2]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA 24061

Abbreviations: BLUP, best-linear unbiased prediction; GEBVs, Genomic estimated breeding values; GWAS, genome-wide association study; PSA, projected shoot area; QTL, quantitative trait loci; RDP1, rice diversity panel 1; DAT, days after transplant; RR, random regression; SMR, single marker regression; SNP, single nucleotide polymorphism; TP, single time point;

Corresponding author:

Malachy Campbell

Department of Agronomy and Horticulture

University of Nebraska Lincoln

Lincoln, Nebraska 68583

Email: campbell.malachy@gmail.com

Abstract

Understanding the genetic basis of dynamic plant phenotypes has largely been limited due to lack of space and labor resources needed to record dynamic traits, often destructively, for a large number of genotypes. However, the recent advent of image-based phenotyping platforms has provided the plant science researchers with an effective means to non-destructively evaluate morphological, developmental, and physiological processes at regular, frequent intervals for a large number of plants throughout development. The statistical frameworks typically used for genetic analyses (e.g. genome-wide association mapping, linkage mapping, and genomic prediction) in plant breeding and genetics are not particularly amenable for repeated measurements. Random regression (RR) models are routinely used in animal breeding for the genetic analysis of longitudinal traits, and provide a robust framework for modeling traits trajectories and performing genetic analysis simultaneously. We recently used a RR approach for genomic prediction of shoot growth trajectories in rice. Here, we have extended this approach for genetic inference by leveraging genomic breeding values derived from RR models for rice shoot growth during early vegetative development. This approach provides improvements over a conventional single time point analyses for discovering loci associated with shoot growth trajectories. This RR approach uncovers persistent, as well as time-specific, transient quantitative trait loci. This methodology can be widely applied to understand the genetic architecture of other complex polygenic traits with repeated measurements.

1 Introduction

A plant's phenotype at any given time is the manifestation of numerous biological processes that have occurred prior to the capture of the phenotype. In most genetic mapping studies, plants are phenotyped at one or few discrete time points. While this may be sufficient for end point traits, such as yield or grain quality, other agronomically important traits such as plant height or vigor are not static and vary continuously throughout development. Given the dynamic nature of these traits, it is likely that some genes will have a time-dependent contribution to the phenotype. Approaches that consider such infinite-dimensional traits as static, fail to fully capture the dynamic processes that have led to the phenotype and may not uncover contribution of time-specific loci.

Recording phenotypic measurements across development in genetic mapping populations is typically limited due to high space and labor demands to record a trait, often destructively, for a large number of genotypes. However, with the advent of image-based phenotyping platforms, researchers can now capture morphological, developmental, and physiological processes non-destructively with higher temporal resolution for a large number of plants (Fraas and Lüthen, 2015; Simko et al., 2016; Shakoor et al., 2017; Tardieu et al., 2017; Araus et al., 2018). Moreover, the growth of the unmanned aerial vehicle industry in recent years has provided many low-cost hardware options that can be outfitted with cameras, facilitating the collection of temporal phenotypes in field settings (Yang et al., 2017). While the use of these platforms is becoming more routine in plant genetics, the statistical frameworks typically used for genetic analyses (e.g. genome-wide association mapping, linkage mapping, and genomic prediction) in plant breeding and genetics are not amenable for longitudinal traits.

Several studies in recent years have sought to elucidate the genetic basis of longitudinal traits through genome-wide association studies (GWAS) or linkage mapping. For instance, Moore et al. (2013) and Würschum et al. (2014) utilized linkage mapping at discrete time points to identify time-specific quantitative trait loci (QTL) associated with root gravitropism and plant height, respectively. While these approaches may be effective, by considering the phenotype at only a single time point they do not leverage the covariance among time points and may have reduced statistical power compared to approaches that consider the entire trait trajectory in regression modeling. Several studies have leveraged a "two-step" approach for functional association mapping (Bac-Molenaar et al., 2015; Campbell et al., 2017). Here, a function is fit to phenotypic records for each genotype, and summarizes the trait trajectories using few parameters. These parameters are then used as derived phenotypes in subsequent GWAS analyses. However, with these "two-step" approaches information is lost between the curve fitting and genetic analysis steps. The residuals from the first curve fitting step likely contains important information regarding persistent environmental effects which are not considered in subsequent genetic analysis. We hypothesize that an approach that unifies the curve fitting and genetic analysis into a single framework is likely to be better than the single time point or a "two-step" longitudinal approach.

Random regression (RR) models provide a robust framework for modeling trait trajectories and performing genetic analysis simultaneously (Schaeffer, 1994; Huisman et al., 2002; Schaeffer, 2004). Covariance functions, such as spline or polynomial functions, are used to model trait trajectories for each line and sufficiently capture the covariance across time points while estimating fewer parameters (Kirkpatrick et al., 1990; Meyer, 1998; White et al., 1999; Strabel and Misztal, 1999; Pool et al., 2000; Huisman et al., 2002; Schaeffer, 2004; Misztal, 2006). Regression coefficients are treated as random effects, and therefore allow values to vary between individuals. Genomic estimated breeding values (GEBVs) for regression coefficients are obtained using a mixed model, and using simple algebra, GEBVs can be obtained for any time throughout the continuous trait trajectory (Mrode, 2014)

GEBVs represent the summation of all additive genetic effects across the genome for a given individual. Goddard (2009) showed that GEBVs predicted using genomic relationships (e.g. genomic best linear unbiased prediction (gBLUP)) are equivalent to those predicted from regression on markers. Given this equivalence, marker effects can be easily calculated from GEBVs, thus genetic inference (e.g. GWAS) can be performed. While this approach is different compared to conventional single marker regression GWAS (SMR-GWAS) approaches, it offers several advantages. First, 100,000s of statistical tests are typically run for SMR-GWAS, and as a result, a stringent p-value threshold must be used to limit false discoveries (Hayes, 2013). Thus, loci recovered using SMR-GWAS approaches typically account for only a fraction of the total genetic variance for a trait (Yang et al., 2010). Whole-genome BLUP approaches (i.e. SNP-BLUP or GBLUP) assume an infinitesimal model in which all loci have some, albeit small, contribution to the phenotype (Hayes, 2013). Thus, by considering all markers simultaneously small-effect QTL are recovered and more genetic variation can be captured compared to SMR-GWAS (Yang et al., 2010). BLUP approaches shrink marker effects towards zero, and thus may not be appropriate for simple traits that are regulated by few loci with large effects. However, for complex polygenic traits these assumptions are reasonable, and should yield biologically meaningful results. In the case of RR, GEBVs can be calculated at each time point and can be leveraged to examine the contribution of loci across a trait trajectory or the time axis.

In a recent study, we used a RR approach for genomic prediction of shoot growth trajectories in rice (Campbell et al., 2018). The utilization of longitudinal phenotypes with RR captured greater genetic variation compared to single time point approach, and significantly improved prediction accuracies. Here, we have leveraged GEBV derived from RR models to examine the genetic architecture of shoot growth through a 20-day period during early vegetative development. Here, we show that this approach can be used for genetic inference of shoot growth trajectories, and uncovers persistent, as well as time-specific QTL. Furthermore, we show that the RR approach uncovers considerably more associations compared to a conventional single time point analysis.

2 Materials and Methods

2.1 High-throughput phenotyping

Phenotypic data was collected for 357 diverse rice accessions from the Rice Diversity Panel 1 (RDP1) (Zhao et al., 2011). The plant materials, experimental design, and image processing are described in detail in Campbell et al. (2018). Briefly, 378 lines were phenotyped at the Plant Accelerator, Australian Plant Phenomics Facility, at the University of Adelaide, SA, Australia from February to April 2016. In this period three experiments were conducted where experiment consisted of a partially replicated design with 54 randomly selected lines having two replicates in each experiment. The plants were grown on greenhouses benches for 10 days after transplanting (DAT) and were loaded on the imaging system and watered to 90% field capacity at 11 DAT.

The plants were imaged daily from 13 to 33 DAT using a visible (red–green–blue camera; Basler Pilot piA2400–12 gc, Ahrensburg, Germany) from two side-view angles separated by 90° and a single top view. The LemnaGrid software was used to extract "plant pixels" from the RGB images using a color classification strategy, and noise (i.e. small areas of non-plant pixels) in the image were removed using a series of erosion and dilation steps. Projected shoot area (PSA) was calculated as the sum of the "plant pixels" from the three RGB images, and was used as a measure of shoot biomass. Outlier plants at each time point were detected at each time point using the 1.5(IQR) rule. Outliers were plotted and those that exhibited abnormal growth patterns were removed. A total of 2,604 plants remained for downstream analyses.

2.2 Predicting genomic breeding values

2.2.1 Random regression

Trajectories for PSA across the 20 time points was modeled using a RR model with Legendre polynomials. The model is the same that was used for genomic prediction in Campbell et al. (2018). The model is described below using the notation of Mrode (2014)

$$\text{PSA}_{tjk} = \mu + \sum_{k=0}^{2} \phi(t)_{jk}\beta_k + \sum_{k=0}^{2} \phi(t)_{jk}u_{jk} + \sum_{k=0}^{1} \phi(t)_{jk}s_{jk} + e_{tjk} \tag{1}$$

Here β_k is the fixed second-order Legendre polynomial to model the mean PSA trajectory for all lines, u_{jk} and s_{jk} are the k^{th} random regression coefficients for additive genetic effect and random experiment of line j, and e_{tjk} is the random residual. The order of β was selected based on visual inspection of the PSA over the 20 days. The random additive genetic effects (u) are modeled using a second-order Legendre polynomial, and the experiment effects (s) are modeled using a first-order Legendre polynomial.

In matrix notation, the model is

$$\mathbf{y} = \mathbf{Zu} + \mathbf{Qs} + \mathbf{e}, \tag{2}$$

Here, $\mathbf{y}$ is PSA over the 20 days; $\mathbf{Z}$ and $\mathbf{Q}$ are incidence matrices corresponding to the random additive genetic effect ($\mathbf{u}$), and random experimental effect ($\mathbf{s}$), respectively; and $\mathbf{e}$ is the random residual error. Note that $\mathbf{u}$ and $\mathbf{s}$ are vectors of random regression coefficients for the additive genetic and experimental effects, respectively.

For the random terms we assume $\mathbf{u} \sim N(0, \mathbf{G} \otimes \boldsymbol{\Omega})$, $\mathbf{s} \sim N(0, \mathbf{I} \otimes \mathbf{P})$, and $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{D})$. Here, $\boldsymbol{\Omega}$ is a 3×3 covariance matrix of random regression coefficients for additive genetic effects; $\mathbf{P}$ is a 2×2 covariance matrix of random regression coefficients for experiment effect;

and $\mathbf{D}$ is a diagonal matrix that allows for heterogeneous variances over the 20 time points. $\mathbf{Z}$ and $\mathbf{Q}$ are covariable matrices where the ith row contains the orthogonal polynomials for the ith day of imaging. Thus, $\mathbf{Z}$ is the covariable matrix for the additive genetic effects with a dimension of $t \times nk$ where nk is the order of Legendre polynomial for the additive genetic effect multiplied by the number of individuals with phenotypic records and t refers to 20 days of records. Similarly, $\mathbf{Q}$ is a $t \times ns$ covariable matrix for the experiment effect, where ns is the the order of the Legendre polynomial for the experiment effect (e.g. 1) times the number of experiments (e.g. 3).

A genomic relationship matrix ($\mathbf{G}$) was calculated using VanRaden (2008).

$$\mathbf{G} = \frac{\mathbf{W_{sc}}\mathbf{W}'_{\mathbf{sc}}}{m} \tag{3}$$

Here, $\mathbf{W_{sc}}$ is a centered and scaled $n \times m$ matrix, where m is 33,674 single nucleotide polymorphism (SNPs) and n is the 357 genotyped rice lines. Variance components and gBLUPs were obtained using ASREML (Release 4.0) (Gilmour et al., 2015).

GEBVs at each time point can be obtained following to Mrode (2014). For line j at time t, the GEBVs can be obtained by $\text{gBLUP}_{jt} = \phi_t \hat{u}_j$; where ϕ_t is the row vector of the matrix of Legendre polynomials of order 2.

2.2.2 Single time point

The following mixed model approach was used to fit gBLUPs at each time point

$$\mathbf{y} = \mathbf{Zu} + \mathbf{Qs} + \mathbf{e}, \tag{4}$$

with all vectors and matrices defined as above. However note that $\mathbf{u}$ and $\mathbf{s}$ are vectors of GEBV and experiment effects. Moreover, here we assume the random terms are as follows

$\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$, $\mathbf{s} \sim N(0, \mathbf{I}\sigma_s^2)$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. A genomic relationship matrix ($\mathbf{G}$) was calculated as above and used for prediction.

2.3 Genome-wide association analyses

2.3.1 Estimating marker effects from GEBVs

GEBVs ($\hat{\mathbf{g}}$) can be parameterized as $\hat{\mathbf{g}} = \hat{\boldsymbol{\beta}}\boldsymbol{W_{sc}}$, where $\mathbf{W_{sc}}$ is a matrix of marker genotypes, as defined above, and $\hat{\boldsymbol{\beta}}$ is a vector of allele substitution effects. $\hat{\boldsymbol{\beta}}$ can be obtained using BLUP

$$BLUP(\boldsymbol{\beta}) = \mathbf{W'_{sc}}(\mathbf{W_{sc}}\mathbf{W'_{sc}})^{-1}\left[\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1}\mathbf{y}. \tag{5}$$

where σ_g^2 and σ_e^2 are genetic and residual variances, respectively.

Given BLUP of GEBVs is

$$BLUP(\mathbf{g}) = \left[\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1}\mathbf{y}, \tag{6}$$

BLUP of marker effects can be obtained using the following linear transformation

$$BLUP(\boldsymbol{\beta}) = \mathbf{W'_{sc}}(\mathbf{W_{sc}}\mathbf{W'_{sc}})^{-1}BLUP(\mathbf{g}). \tag{7}$$

This relationship was leveraged to solve for marker effects from breeding values for both RR and single time point (TP) analyses.

2.3.2 Variance of SNP effects

The variance of marker effects at each time point obtained through TP or RR approaches was calculated following the methods outlined by Duarte et al. (2014). Since we solve for RR-derived GEBVs at each time point, the models for the TP and RR approaches can both be parameterized as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Qs} + \mathbf{e}, \tag{8}$$

where all vectors and matrices are defined as above except $\mathbf{b}$, which is the average PSA at each time point and $\mathbf{X}$ is an incidence matrix that relates the mean PSA to the observations.

The variance of SNP effects is obtained using

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}\hat{\mathbf{g}}) = \mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}\mathrm{Var}(\hat{\mathbf{g}})\mathbf{G}^{-\mathbf{1}}\mathbf{W}_{\mathbf{sc}}, \tag{9}$$

and $\mathrm{Var}(\hat{\mathbf{g}})$ can be obtained using

$$\mathrm{Var}(\hat{\mathbf{g}}) = \mathrm{Var}(\mathbf{g}) - \mathbf{C}^{\mathbf{aa}} = \mathbf{G}\sigma^{\mathbf{2}}_{\mathbf{g}} - \mathbf{C}^{\mathbf{aa}}. \tag{10}$$

Substituting the expression above into the expression for $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ we obtain

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}(\mathbf{G}\sigma^{\mathbf{2}}_{\mathbf{g}} - \mathbf{C}^{\mathbf{aa}})\mathbf{G}^{-\mathbf{1}}\mathbf{W}_{\mathbf{sc}} &(11)\\ &= \mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}\mathbf{W}_{\mathbf{sc}}\sigma^2_g - \mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}\mathbf{C}^{\mathbf{aa}}\mathbf{G}^{-\mathbf{1}}\mathbf{W}_{\mathbf{sc}}. &(12)\end{aligned}$$

Here, $\mathbf{C^{aa}}$ is obtained by inverting the coefficient matrix of the mixed model equation outlined by Henderson (1984), and is provided below.

$$\mathbf{C^{aa}} = \sigma_e^2(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'} + \mathbf{G}^{-\mathbf{1}}\lambda)^{-1}, \lambda = \frac{\sigma_e^2}{\sigma_g^2}. \tag{13}$$

2.3.3 Obtaining p-values for marker effects

SNP effects for SNP_j at time t were divided by their corresponding $\text{Var}(\hat{\boldsymbol{\beta}})$ using

$$\text{SNP}_{jt} = \frac{\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta})}} \tag{14}$$

The p-values for marker effects were calculated as 1 minus the cumulative probability density of the absolute value of SNP_{jt}, and this number was subsequently multiplied by two. This is summarized as follows.

$$p\text{-value}_{SNP_{jt}} = 2(1 - \phi(|\text{SNP}_{jt}|)). \tag{15}$$

Following Zhao et al. (2011) a threshold of 1×10^{-4} was used to declare significant loci.

3 Results and Discussion

To identify loci associated with shoot growth trajectories in rice, we utilized a novel RR approach that allows for trait trajectories to be modeled across time points. Shoot growth trajectories were recorded for 357 diverse rice accessions over a period of twenty days during early vegetative growth (13 - 33 DAT). A RR model was fitted to the shoot growth trajectories, which included a fixed second-order Legendre polynomial, a random second-order Legendre polynomial for the additive genetic effect, a first-order Legendre polynomial for the environmental effect, and heterogeneous residual variances. GEBVs were predicted for each accession at each of the 20 time points as described in Campbell et al. (2018), and was used to estimate marker effects at each time point. Results from the RR were compared with a conventional single time point approach in which GEBVs were predicted at each time point using a conventional mixed model and were used to estimate marker effects.

3.1 RR-GWAS recovers more significant associations and increases predicted marker effect sizes

With RR models, the incorporation of the covariance structure of multiple measurements should lead to a more accurate partitioning of phenotypic variation into genetic and environmental components, and improve genetic inference. To demonstrate the advantages of a longitudinal genetic inference approach over a conventional TP approach, significant marker effects were compared between the RR and TP approaches. A 89% increase in the number of significant associations ($p < 10^{-4}$) were observed with the RR approach compared to the conventional TP model. A total of 717 non-redundant SNPs were found to be significantly associated with shoot growth trajectories at one or more time points using the RR approach, while 379 were found using the TP approach. Correlations in SNP effects estimated using the two approaches showed a very high agreement ($r = 0.847$), however predicted marker effects ($\hat{\boldsymbol{\beta}}$) obtained using the RR were considerably larger than the single time point analysis (Fig 1). For instance, $\hat{\boldsymbol{\beta}}$ for the RR approach ranged from -309.6 to 366.3 across all days, while for the TP approach $\hat{\boldsymbol{\beta}}$ ranged from -106.0 to 114.3. These differences are evident in the Manhattan plots on days 1 and 20 pictured in Fig 2. Manhattan plots for each of the 20 time points is provided as supplemental Figures S1, S2, S3, S4. These results indicate that the utilization of information across all time points with the RR improves the ability to detect significant associations as well as increases the predicted marker effect sizes compared to a model that utilizes information at only a single time point.

Figure 1: Correlation and distribution of SNP effects from random regression (RR) and single time point (TP) analysis. (A) Correlation between SNP effects for the random regression (β_{RR}) and single time point analyses (β_{TP}). SNPs highlighted in red are those that were statistically significant in the RR approach ($p < 1 \times 10^{-4}$). The grey broken lines depicts a one-to-one relationship between β_{RR} and β_{TP}. Distribution of SNP effects across all twenty time points from the TP analyses (B) and RR analysis (C).

Figure 2: Manhattan plots for RR and TP approaches on days 1 and 20. (A,B) Manhattan plots for RR approach on days 1 and 20, respectively. (C,D) Manhattan plots for TP approach on days 1 and 20, respectively. $|\beta|$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

These results suggest that the inclusion of time axis for genetic inference improve the ability to recover significant associations. Several other studies have showed similar improvements in the estimation of variance components and genetic inference using different approaches for longitudinal traits. For instance, De Andrade et al. (2002) showed a longitudinal approach that leveraged pedigree data and systolic blood pressure measurements collected at three time points improved heritability estimates compared with a single time point approach. While in the context of GWAS, Das et al. (2011) used a novel functional GWAS (fGWAS) approach and identified several new variants associated with body mass index collected at four time points in humans. Moreover, using simulated data the authors show that the statistical power exceeds 0.8 with a false positive rate of less than 0.1 for sample sizes greater than 1,000. Similar gains for GWAS have been demonstrated in both plants, animals, and humans (Xu et al., 2014; Campbell et al., 2015; Yi et al., 2015; Lund et al., 2008).

3.2 RR-GWAS reveals the dynamic genetic architecture of shoot growth responses in rice

For many traits, such as growth, genetic effects are expected to vary across time. These temporal genetic effects can be effectively captured using a RR approach. To examine the dynamic genetic architecture of shoot growth trajectories, significant SNPs from the RR approach were selected and those within a 200 kb window were merged to a single QTL. The 200 kb window that we used, corresponds to the average linkage disequilibrium in rice (Zhao et al., 2011). For the RR approach, a total of 342 significant QTL were detected at one or more time points, while for the TP approach only 142 significant QTL were detected.

To dissect the dynamic genetic architecture of shoot growth in rice, significant QTLs were classified into four categories: persistent QTL (QTL detected at all 20 time points), long-duration QTL (those with significant associations at more than 12, but less than 20 time points), mid-duration (QTL with associations at 6 - 12 time points), and short-duration QTL (those with associations at fewer than 6 time points). Of these categories, far more persistent QTLs were detected, with a total of 128 observed at all 20 time points. Short duration QTL also showed a large number of significant QTL (127), while relatively few were detected for long and mid-duration QTL (32 and 55 QTL, respectively). The frequency of significant QTL for each category were calculated at each time point and plotted as a function of time (Fig S5). For all classes, a large number of QTL were detected on the first and last days (day 1 and day 20, respectively). For instance, all long-duration QTL were present on days 1, 7 and 17-20. While for mid-duration QTL, 94% were detected on the first day of imaging, 93% on the fifth day, and 87% on the last day of imaging. Interestingly, for both short and mid-duration QTLs, less than 10% were detected from day 9-16. The p-values across all 20 time points for a subset of highly significant QTL are provided in Figure 3. Collectively, these results indicate that the shoot growth is regulated by numerous loci that have both transient and persistent effects throughout early vegetative growth.

Figure 3: Heatmap showing time-specific QTL. A subset of significant QTL identified with RR approach are pictured. The x-axis indicates the days of imaging and the y-axis shows the chromosome and intervals for the QTL. For each QTL, the most significant SNP within the interval at each time point were selected. The grey color scale indicates a non-significant association, while the red color scale indicates a statistically significant association ($p < 1 \times 10^{-4}$).

The importance of time-specific QTL has been demonstrated in both plants and animals (Moore et al., 2013; Bac-Molenaar et al., 2015; Campbell et al., 2017, 2015). For instance, using a single time point linkage mapping approach, Moore et al. (2013) showed several time-specific QTL associated with root gravitropic responses in Arabidopsis. Moreover, many of these QTL harbored candidate genes known to influence root growth, root gravitropism, or hormone transport and signaling. Bac-Molenaar et al. (2015) collected rosette growth trajectories over a period of 20 days for a diverse panel of 324 Arabidopsis accessions. A growth function was fit for each accession, and model parameters were used for GWAS. The authors showed that many associations detected for model parameters were also detected at a few time points using a single time point GWAS approach. While few longitudinal studies have been performed in rice, our previous studies have identified time-specific QTL for shoot growth and salt stress responses (Campbell et al., 2015, 2017).

4 Conclusion

New phenotyping platforms has provided the plant science community with a suite of tools to collect high-dimensional temporal phenotypic data. With these temporal dataset, quantitative genetic approaches that can leverage the covariance across time points must be fully utilized to realize the potential of these data for genomic prediction and genetic inference. Here, we show that the RR framework that has been extensively developed in animal breeding can be extended to genetic inference in plants. This approach can effectively be used to identify QTL with time-specific effects. To date, this is the first application of random regression models for genetic inference of a longitudinal trait in a major crop.

Acknowledgements

Funding for this research was provided by the National Science Foundation (United States) through Award No. 1238125 to Harkamal Walia, and Award No. 1736192 to Harkamal Walia and Gota Morota.

Supplemental Materials

SupplementalData.pdf: Figures S1-S5.

Author Contributions

Study was conceived by H.W., G.M., and M.C.; phenotyping was performed by M.C. and H.W.; M.C. and G.M. performed all analyses; M.C. wrote the manuscript, and editorial comments were provided by H.W. and G.M.

Data Accessibility

The full datasets and all code used in this study is available via GitHub (https://github.com/malachycampb
RR-GEBVs-for-genomic-inference-of-longitudinal-traits)
and the WRCHR website (WRCHR.org).

References

Araus, J. L., S. C. Kefauver, M. Zaman-Allah, M. S. Olsen, and J. E. Cairns, 2018: Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science*.

Bac-Molenaar, J. A., D. Vreugdenhil, C. Granier, and J. J. Keurentjes, 2015: Genome-wide association mapping of growth dynamics detects time-specific and general quantitative trait loci. *Journal of Experimental Botany*, **66 (18)**, 5567–5580.

Campbell, M. T., Q. Du, K. Liu, C. J. Brien, B. Berger, C. Zhang, and H. Walia, 2017: A comprehensive image-based phenomic analysis reveals the complex genetic architecture of shoot growth dynamics in rice. *The Plant Genome*, **10 (2)**.

Campbell, M. T., A. C. Knecht, B. Berger, C. J. Brien, D. Wang, and H. Walia, 2015: Integrating image-based phenomics and association analysis to dissect the genetic architecture of temporal salinity responses in rice. *Plant Physiology*, **168 (4)**, 1476–1489.

Campbell, M. T., H. Walia, and G. Morota, 2018: Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant Direct*, **2 (9)**.

Das, K., and Coauthors, 2011: A dynamic model for genome-wide association studies. *Human genetics*, **129 (6)**, 629–639.

De Andrade, M., R. Guéguen, S. Visvikis, C. Sass, G. Siest, and C. I. Amos, 2002: Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, **22 (3)**, 221–232.

Duarte, J. L. G., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel, 2014: Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics*, **15 (1)**, 246.

Fraas, S., and H. Lüthen, 2015: Novel imaging-based phenotyping strategies for dissecting crosstalk in plant development. *Journal of Experimental Botany*, **66 (16)**, 4947–4955.

Gilmour, A., B. Gogel, B. Cullis, S. Welham, and R. Thompson, 2015: Asreml user guide release 4.1 structural specification. *Hemel Hempstead: VSN International Ltd.*

Goddard, M., 2009: Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, **136 (2)**, 245–257.

Hayes, B., 2013: Overview of statistical methods for genome-wide association studies (gwas). *Genome-wide association studies and genomic prediction*, Springer, 149–169.

Henderson, C., 1984: *Applications of linear models in animal breeding.*

Huisman, A., R. Veerkamp, and J. Van Arendonk, 2002: Genetic parameters for various random regression models to describe the weight data of pigs. *Journal of Animal Science*, **80 (3)**, 575–582.

Kirkpatrick, M., D. Lofsvold, and M. Bulmer, 1990: Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, **124 (4)**, 979–993.

Lund, M. S., P. Sorensen, P. Madsen, and F. Jaffrézic, 2008: Detection and modelling of time-dependent qtl in animal populations. *Genetics Selection Evolution*, **40 (2)**, 177.

Meyer, K., 1998: Estimating covariance functions for longitudinal data using a random regression model. *Genetics Selection Evolution*, **30 (3)**, 221.

Misztal, I., 2006: Properties of random regression models using linear splines. *Journal of Animal Breeding and Genetics*, **123 (2)**, 74–80.

Moore, C. R., L. S. Johnson, I.-Y. Kwak, M. Livny, K. W. Broman, and E. P. Spalding, 2013: High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. *Genetics*, genetics–113.

Mrode, R. A., 2014: *Linear models for the prediction of animal breeding values*. CABI.

Pool, M., and Coauthors, 2000: Reduction of the number of parameters needed for a polynomial random regression test day model. *Livestock Production Science*, **64 (2-3)**, 133–145.

Schaeffer, L., 1994: Random regressions in animal models for test-day production in dairy cattle. *World Congress of Genetics Applied Livestock Production, 1994*, Vol. 18, 443–446.

Schaeffer, L., 2004: Application of random regression models in animal breeding. *Livestock Production Science*, **86 (1-3)**, 35–45.

Shakoor, N., S. Lee, and T. C. Mockler, 2017: High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current Opinion in Plant Biology*, **38**, 184–192.

Simko, I., J. A. Jimenez-Berni, and X. R. Sirault, 2016: Phenomic approaches and tools for phytopathologists. *Phytopathology*, **107 (1)**, 6–17.

Strabel, T., and I. Misztal, 1999: Genetic parameters for first and second lactation milk yields of polish black and white cattle with random regression test-day models. *Journal of Dairy Science*, **82 (12)**, 2805–2810.

Tardieu, F., L. Cabrera-Bosquet, T. Pridmore, and M. Bennett, 2017: Plant phenomics, from sensors to knowledge. *Current Biology*, **27 (15)**, R770–R783.

VanRaden, P. M., 2008: Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91 (11)**, 4414–4423.

White, I., R. Thompson, and S. Brotherstone, 1999: Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science*, **82 (3)**, 632–638.

Würschum, T., and Coauthors, 2014: Mapping dynamic qtl for plant height in triticale. *BMC Genetics*, **15 (1)**, 59.

Xu, Z., X. Shen, W. Pan, A. D. N. Initiative, and Coauthors, 2014: Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS One*, **9 (8)**, e102 312.

Yang, G., and Coauthors, 2017: Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. *Frontiers in Plant Science*, **8**, 1111.

Yang, J., and Coauthors, 2010: Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, **42 (7)**, 565.

Yi, G., and Coauthors, 2015: Genome-wide association study dissects genetic architecture underlying longitudinal egg weights in chickens. *BMC Genomics*, **16 (1)**, 746.

Zhao, K., and Coauthors, 2011: Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications*, **2**, 467.

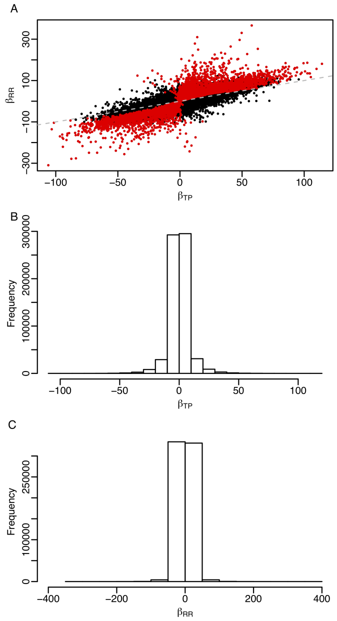

Figure 1: Correlation and distribution of SNP effects from random regression (RR) and single time point (TP) analysis. (A) Correlation between SNP effects for the random regression (β_{RR}) and single time point analyses (β_{TP}). SNPs highlighted in red are those that were statistically significant in the RR approach ($p < 1 \times 10^{-4}$). The grey broken lines depicts a one-to-one relationship between β_{RR} and β_{TP}. Distribution of SNP effects across all twenty time points from the TP analyses (B) and RR analysis (C).

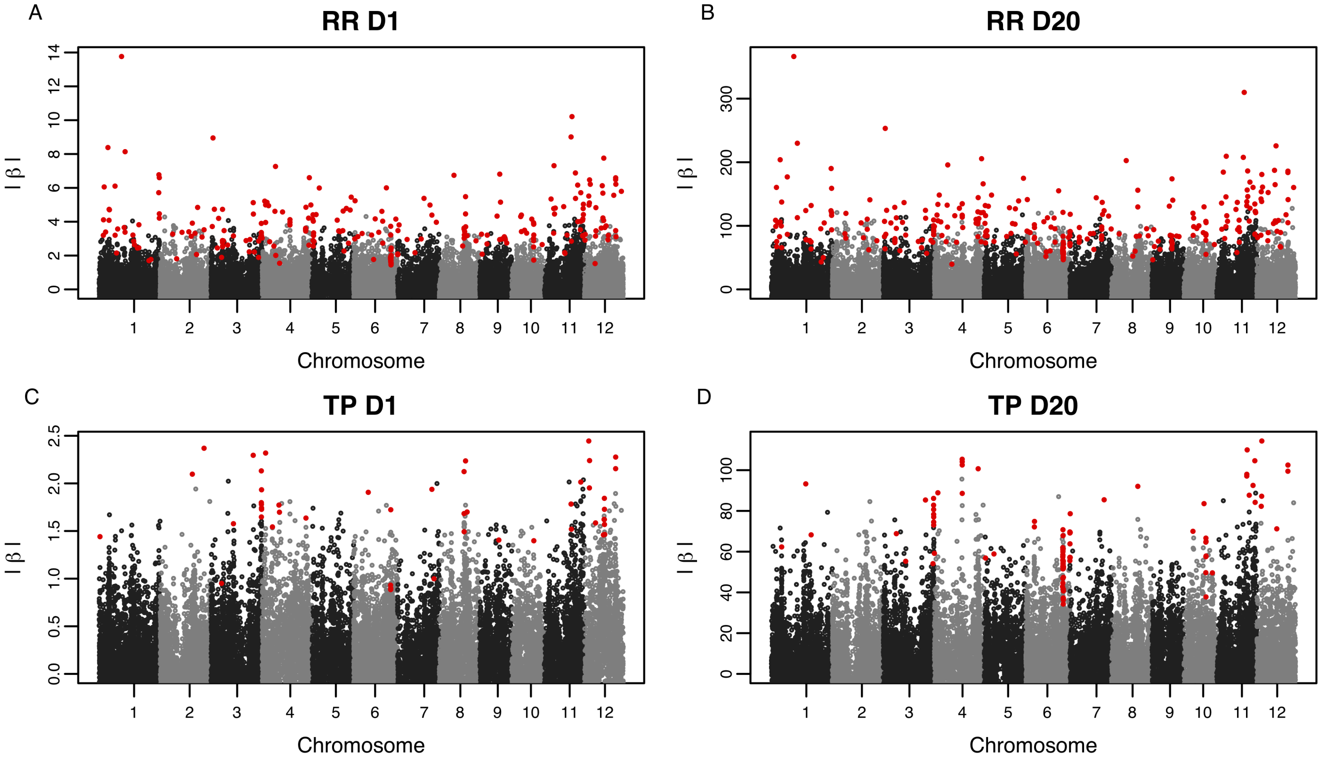

Figure 2: Manhattan plots for RR and TP approaches on days 1 and 20. (A,B) Manhattan plots for RR approach on days 1 and 20, respectively. (C,D) Manhattan plots for TP approach on days 1 and 20, respectively. $|\beta|$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

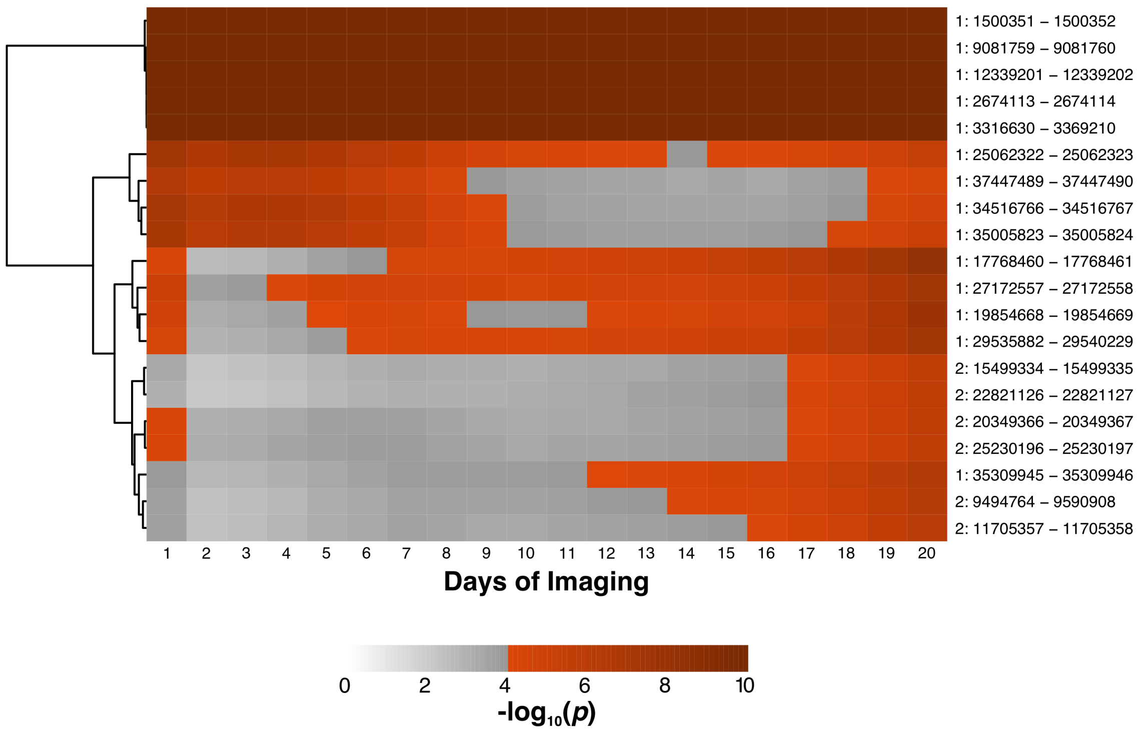

Figure 3: Heatmap showing time-specific QTL. A subset of significant QTL identified with RR approach are pictured. The x-axis indicates the days of imaging and the y-axis shows the chromosome and intervals for the QTL. For each QTL, the most significant SNP within the interval at each time point were selected. The grey color scale indicates a non-significant association, while the red color scale indicates a statistically significant association ($p < 1 \times 10^{-4}$).