Running Title: Genetic inference of longitudinal traits

Core Ideas:

- Random regression models are an appealing framework for GWAS of longitudinal traits
- This approach provides improvements over a conventional single time point analyses for GWAS
- We identify QTL with transient and persistent effects on shoot growth in rice

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

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

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

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

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 community 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 using 33,674 SNPs. In this study, 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. The 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 the contributions 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). In the two-step approach, a function is fit to phenotypic records for each genotype that summarizes the trait trajectories using a 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 contain important information regarding persistent environmental effects that 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; Sun et al., 2017). 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; Sun et al., 2017). In a recent study Sun et al. (2017) utilized a RR approach with cubic splines in wheat to obtain best linear unbiased predictions of secondary traits derived from high-throughput hyperspectral and thermal imaging. 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 accuracy. In the current study, we have leveraged GEBVs derived from

RR models to examine the genetic architecture of shoot growth through a 20-day period during early vegetative development. 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 greenhouse benches for 10 days after transplanting (DAT) and were loaded on the imaging system and watered to 90% field capacity at 11 DAT.

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 greenhouse 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 areas projected in two dimensional space from each of the three RGB images, and was used as a measure of shoot biomass. Previous studies have shown a high correlation between PSA and conventional destructive

measures of shoot biomas (Golzarian et al., 2011; Campbell et al., 2015; Neilson et al., 2015; Knecht et al., 2016). A depiction of PSA collected from RGB images is provided as Figure S1. Outlier plants at each time point were detected at each time point using the 1.5(IQR) rule. Briefly, the distribution of PSA at each day was split into quartiles and the interquartile range (IQR) was calculated as the difference between the third and first quartiles. Points that were either less than Q1 − 1.5 (IQR) or Q3 + 1.5(IQR) were considered as outliers. 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)

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

PSA_{tij} is PSA on day t for line j within experiment i. $_k$ is the fixed second-order Legendre polynomial to model the mean PSA trajectory for all lines, u_{jk} and s_{ik} are the kth random regression coefficients for additive genetic effect and random experiment effects, and e_{tij} 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

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

$\mathbf{y}$ is a vector of PSA over the 20 days and is of order n, where n is the number of observations. $\mathbf{X}$ is an $n \times k_f$ covariable matrix where the number of columns is equal to the order of Legendre polynomial used to model fixed effects (k_f). The matrices $\mathbf{Z}$ and $\mathbf{Q}$ are covariable matrices for the random additive genetic and random experimental effects, respectively. The number of rows for $\mathbf{Z}$ is n and the number of columns corresponds to the order of Legendre polynomial times the number of lines used to fit the additive genetic effect ($q * k_g = 357 * 3 = 1,071$). The dimension of $\mathbf{Q}$ is $n \times e * k_s$ where k_s is the order of Legendre polynomial used to fit the permanent environmental effects and e is the number of experiments. We assume $\mathbf{u} \sim N(0, \mathbf{G} \otimes \mathbf{\Omega})$, $\mathbf{s} \sim N(0, \mathbf{I} \otimes \mathbf{P})$, and $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{D})$. Here, $\mathbf{\Omega}$ and $\mathbf{P}$ are the covariance matrices for the RR coefficients for the additive genetic and permanent environmental effects, and $\mathbf{D}$ is a diagonal matrix that allows for heterogeneous variances over the 20-time points.

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

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

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

Solving the mixed model equation will give three RR coefficients for each line. Using these RR coefficients, GEBVs at each time point can be obtained. For line j the predicted genetic values (GEBV) at each time point is given by $GEBV_j = \mathbf{\Phi_g}\hat{u}_j$ (Mrode, 2014). $\mathbf{\Phi_g}$ is the matrix of Legendre polynomials used for fitting the additive genetic effects. A detailed

explanation of the RR model is provided in the appendix.

2.2.2 Single time point

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

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

The matrices $\mathbf{X}, \mathbf{Z}$ and $\mathbf{Q}$ correspond to incidence matrices for the fixed, random additive genetic and random experimental effect, respectively. Moreover, the dimensions for $\mathbf{X}, \mathbf{Z}$ and $\mathbf{Q}$ are $n \times 1$, $n \times q$ and $n \times e$. We assume the random terms are distributed 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 of breeding values at each time point.

2.3 Genome-wide association analyses

2.3.1 Estimating marker effects from GEBVs

GEBVs ($\hat{\mathbf{u}}$) can be parameterized as $\hat{\mathbf{u}} = \hat{\boldsymbol{\beta}}\boldsymbol{W_{sc}}$, 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

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

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

Given BLUP of GEBVs is

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

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

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

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

2.3.2 Variance of SNP effects

The variance of marker effects was calculated following the methods outlined by Duarte et al. (2014). Briefly, the variance of marker effects can be obtained via linear transformation of the variance of GEBVs ($\hat{\mathbf{u}}$).

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\mathbf{W}'_{\mathbf{sc}}\mathbf{G}^{-\mathbf{1}}\hat{\mathbf{u}}) \tag{8}$$

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

The prediction error variance (PEV) of $\hat{\mathbf{u}}$ is

$$\mathrm{PEV}(\hat{\mathbf{u}}) = \mathbf{C}^{\mathbf{22}} = \mathrm{Var}(\mathbf{u}) - \mathrm{Var}(\hat{\mathbf{u}}) \tag{10}$$

$$= \mathbf{G}\sigma_g^2 - \mathrm{Var}(\hat{\mathbf{u}}) \tag{11}$$

$\mathbf{C}^{\mathbf{22}}$ is obtained by inverting the coefficient matrix of the mixed model equation provided

in the appendix, and extracting the elements corresponding to additive genetic effects Henderson (1984). Thus, by rearranging equation 10, the variance of predicted breeding values is

$$\mathrm{Var}(\hat{\mathbf{u}}) = \mathbf{G}\sigma_g^2 - \mathbf{C^{22}}\sigma_e^2 \tag{12}$$

For the TP approach $\mathbf{C^{22}}$ is a $q \times q$ matrix, and diagonal elements correspond to the PEV of breeding values. Since the MME is solved for each time point independently, the above procedure can be used to obtain the variance of SNP effects on each day. However for the RR approach, $\mathbf{C^{22}}$ is $q * k_g \times q * k_g$ and represents the PEV for the additive genetic RR coefficients. Thus, to obtain $\mathrm{Var}(\hat{\mathbf{u}})$ at each time point, we define a new matrix $\mathbf{C^{22*}}$ that is $q * d \times q * d$ where d is the number of time points (e.g. 20). This is given by

$$\boldsymbol{C^{22*}} = \boldsymbol{\Phi_g^*} \boldsymbol{C^{22}} \boldsymbol{\Phi_g^{*\prime}} \tag{13}$$

$\boldsymbol{\Phi_g^*}$ is a $q * d \times k_g * q$ block matrix where the diagonal sub-matrices consist of Legendre polynomials at each standardized time interval. This approach is analogous to that described by Mrode (2014), and is described in greater detail in the appendix.

2.3.3 Obtaining p-values for marker effects

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

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

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 20 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 131% increase in the number of significant associations ($p < 10^{-4}$) was observed with the RR approach compared to the conventional TP model. A total of 442 SNPs were found to be significantly associated with shoot growth trajectories at one or more time points using the RR approach, while 191 were found using the TP approach. Correlations in SNP effects estimated using the two approaches showed a very high agreement ($r = 0.85$), 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 -299.1 to 295.0 across all days, while for the TP approach $\hat{\boldsymbol{\beta}}$ ranged from -104.6 to 112.3. These differences are evident in the distribution of marker effects pictured in Fig 1. Manhattan plots for each of the 20-time points is provided as supplemental Figures S2, S3, S6, S7, and the corresponding Q-Q plots are provided as supplemental Figures S10, S8, S9, S5. 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 20-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. -$\log_{10}(p)$ 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 lon-

gitudinal 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 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; Huang et al., 2010). For the RR approach, a total of 26 significant QTL were detected at one or more time points, while for the TP approach only 15 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 13 observed at all 20-time points. Short duration QTL also showed the fewest number of significant QTL (2). While five and six QTL were detected for long and mid-duration QTL, respectively. The frequencies of significant QTL for each category were calculated at each time point and plotted as a function of time (Fig S7). The majority of long-duration QTL were detected towards the end of the experiment (day 8 onward), while short-duration QTL were detected only from days 1 to 4. Mid-duration QTL were detected at all time points. The p-values across all 20-time points for all 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., 2015, 2017). 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 have 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. In this study, 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: Appendix; Figures S1-S6.

Author Contributions

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

Golzarian, M. R., R. A. Frick, K. Rajendran, B. Berger, S. Roy, M. Tester, and D. S. Lun, 2011: Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods*, **7 (1)**, 2.

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.*

Huang, X., and Coauthors, 2010: Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, **42 (11)**, 961.

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., W. G. Hill, and R. Thompson, 1994: Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genetics Research*, **64 (1)**, 57–69.

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

Knecht, A. C., M. T. Campbell, A. Caprez, D. R. Swanson, and H. Walia, 2016: Image Harvest: an open-source platform for high-throughput plant image processing and analysis. *Journal of Experimental Botany*, **67 (11)**, 3587–3599.

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.

Neilson, E. H., A. Edwards, C. Blomstedt, B. Berger, B. L. Møller, and R. Gleadow, 2015: Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a c4 cereal crop plant to nitrogen and water deficiency over time. *Journal of Experimental Botany*, **66 (7)**, 1817–1832.

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.

Sun, J., J. E. Rutkoski, J. A. Poland, J. Crossa, J.-L. Jannink, and M. E. Sorrells, 2017: Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*, **10 (2)**.

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.

Figures

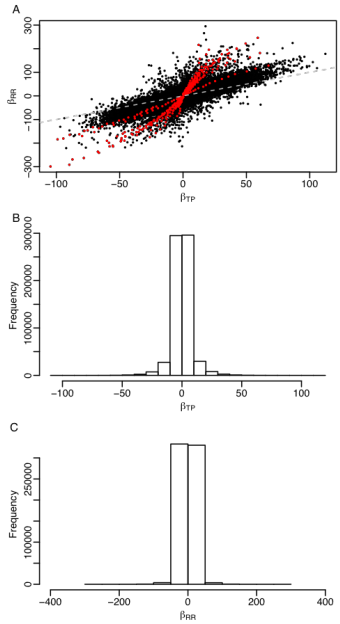

Figure 1: Correlation and distribution of SNP effects from random regression (RR) and single time point (TP) analysis. (A) Correlation between SNP effects for the random regression (β_{RR}) and single time point analyses (β_{TP}). SNPs highlighted in red are those that were statistically significant 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 20-time points from the TP analyses (B) and RR analysis (C).

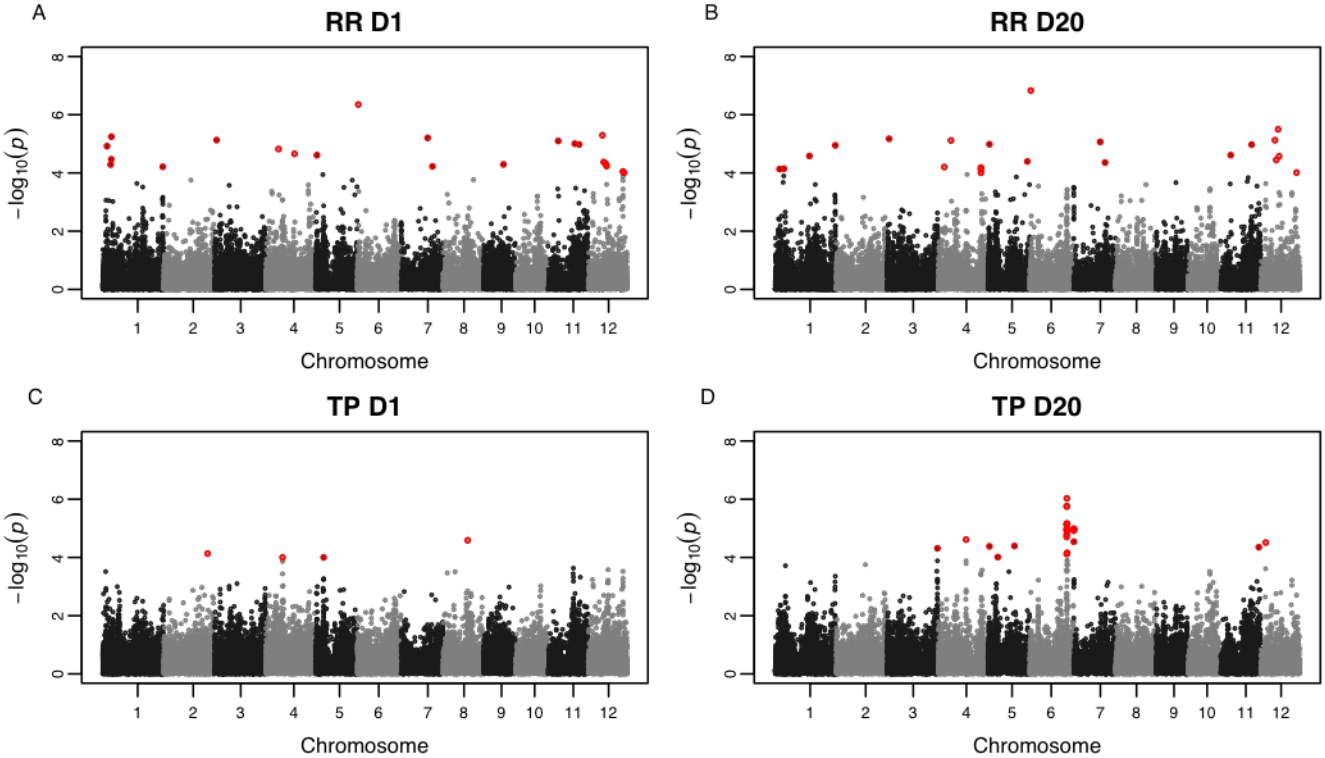

Figure 2: Manhattan plots for RR and TP approaches on days 1 and 20. (A,B) Manhattan plots for RR approach on days 1 and 20, respectively. (C,D) Manhattan plots for TP approach on days 1 and 20, respectively. -$\log_{10}(p)$ is shown on the y-axis. Statistically significant SNPs are highlighted in red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

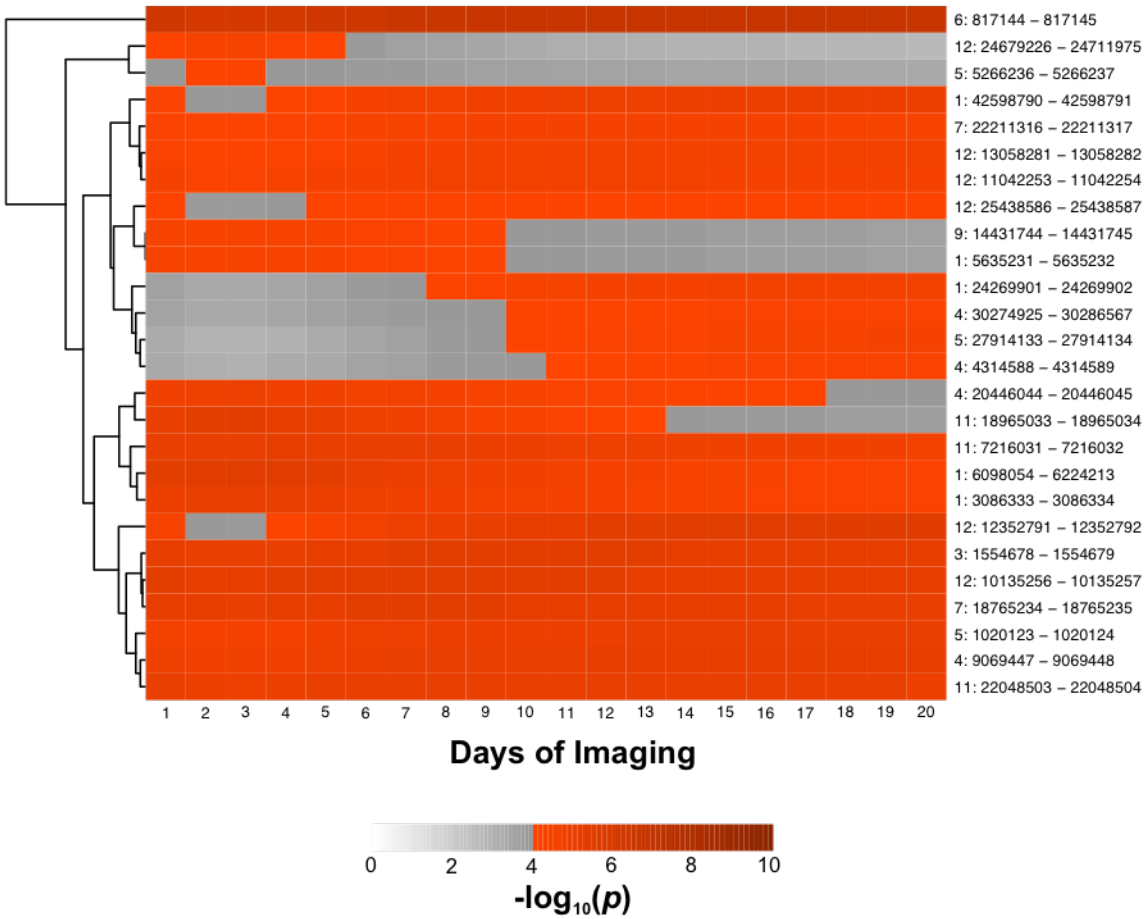

Figure 3: Heatmap showing time-specific QTL. A subset of significant QTL identified with RR approach are pictured. The x-axis indicates the days of imaging and the y-axis shows the chromosome and intervals for the QTL. For each QTL, the most significant SNP within the interval at each time point were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

Supplemental Data

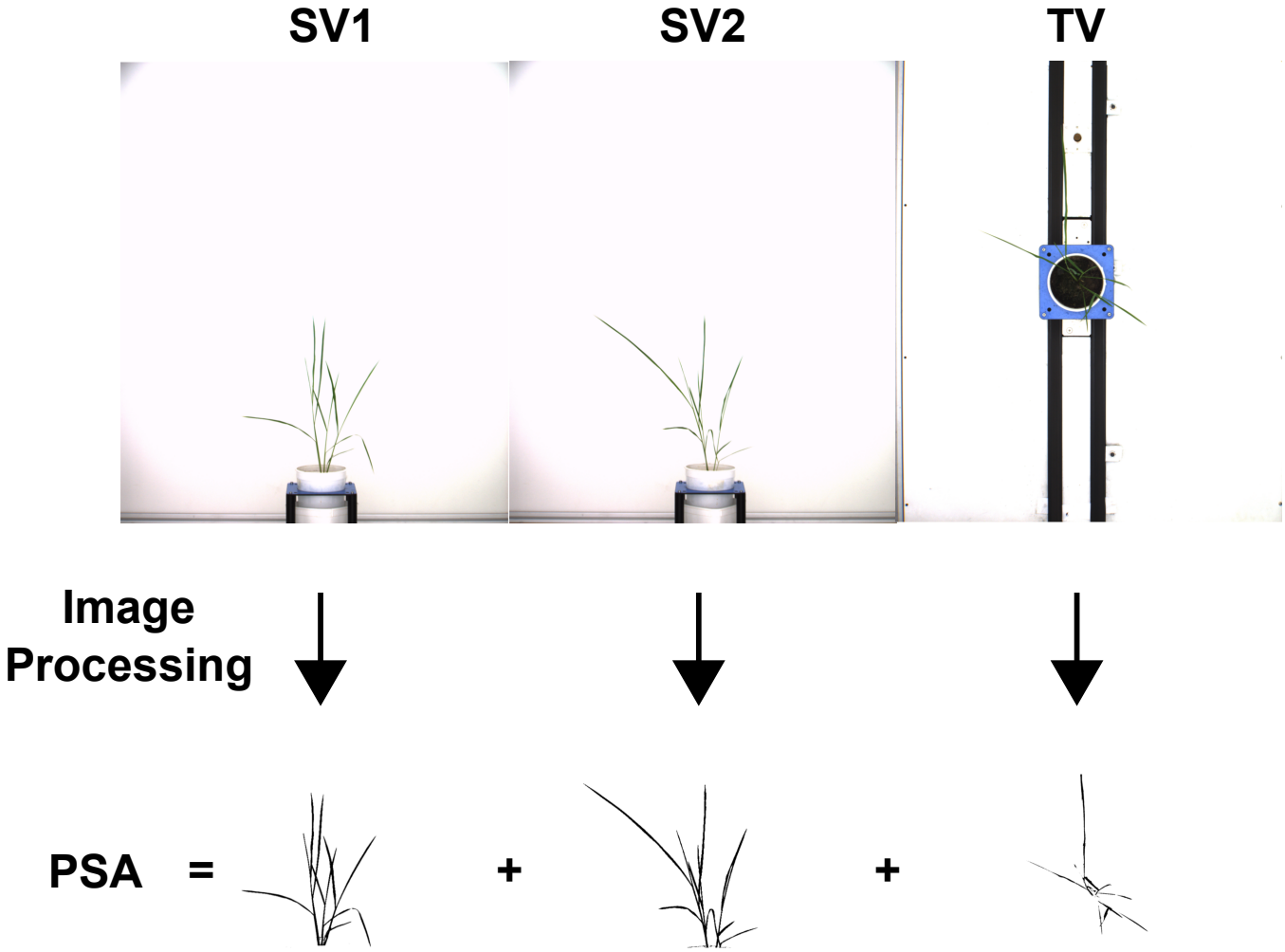

Figure S1: Visual depiction of assessing shoot biomass using projected shoot area. The plant is imaged from three perspectives. The two side view (SV) images are separated by 90 degrees. TV: top view; PSA: projected shoot area

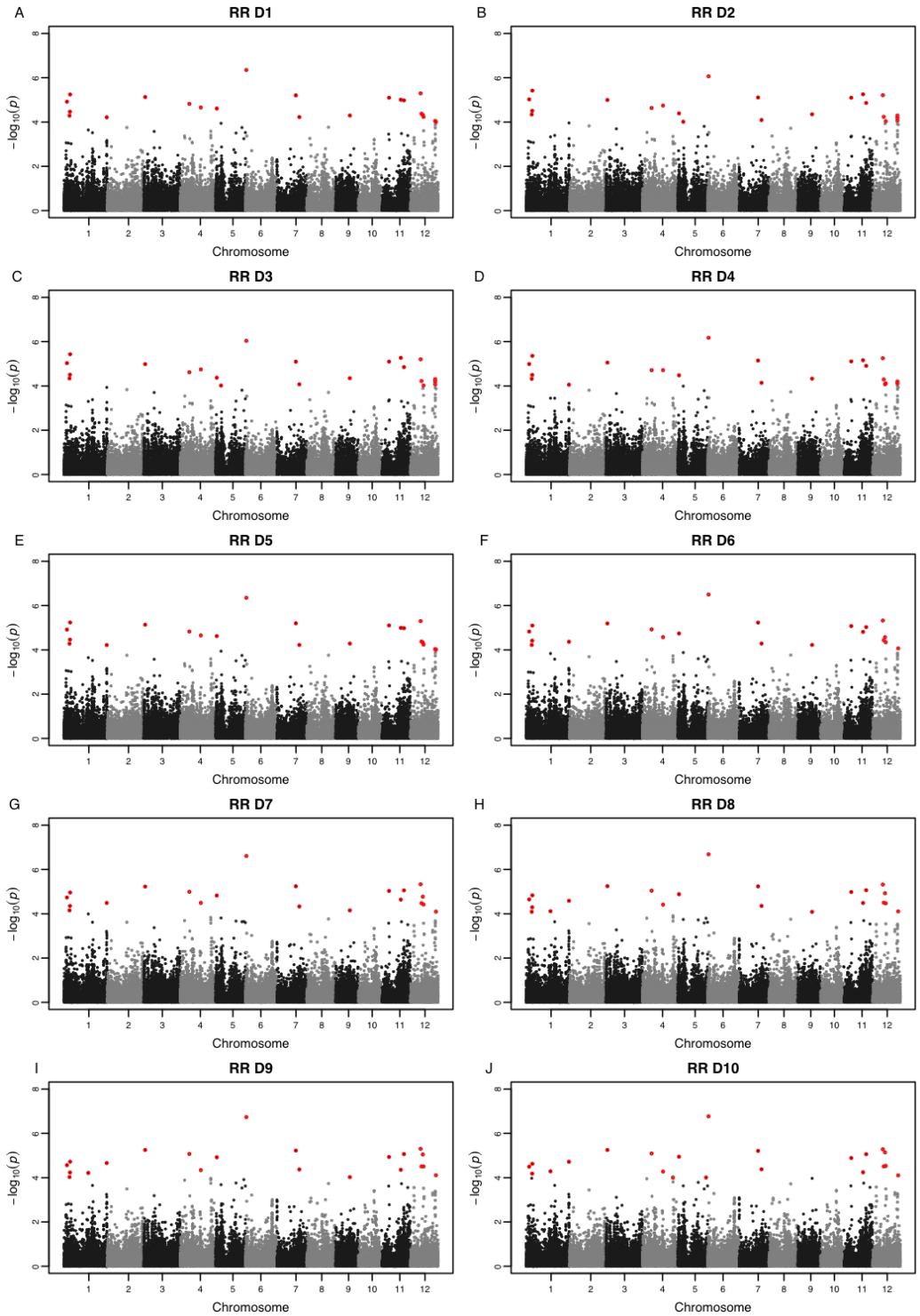

Figure S2: Manhattan plots for RR approach at days 1 to 10. Each panel represents a single time point. -$\log_{10}(p)$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

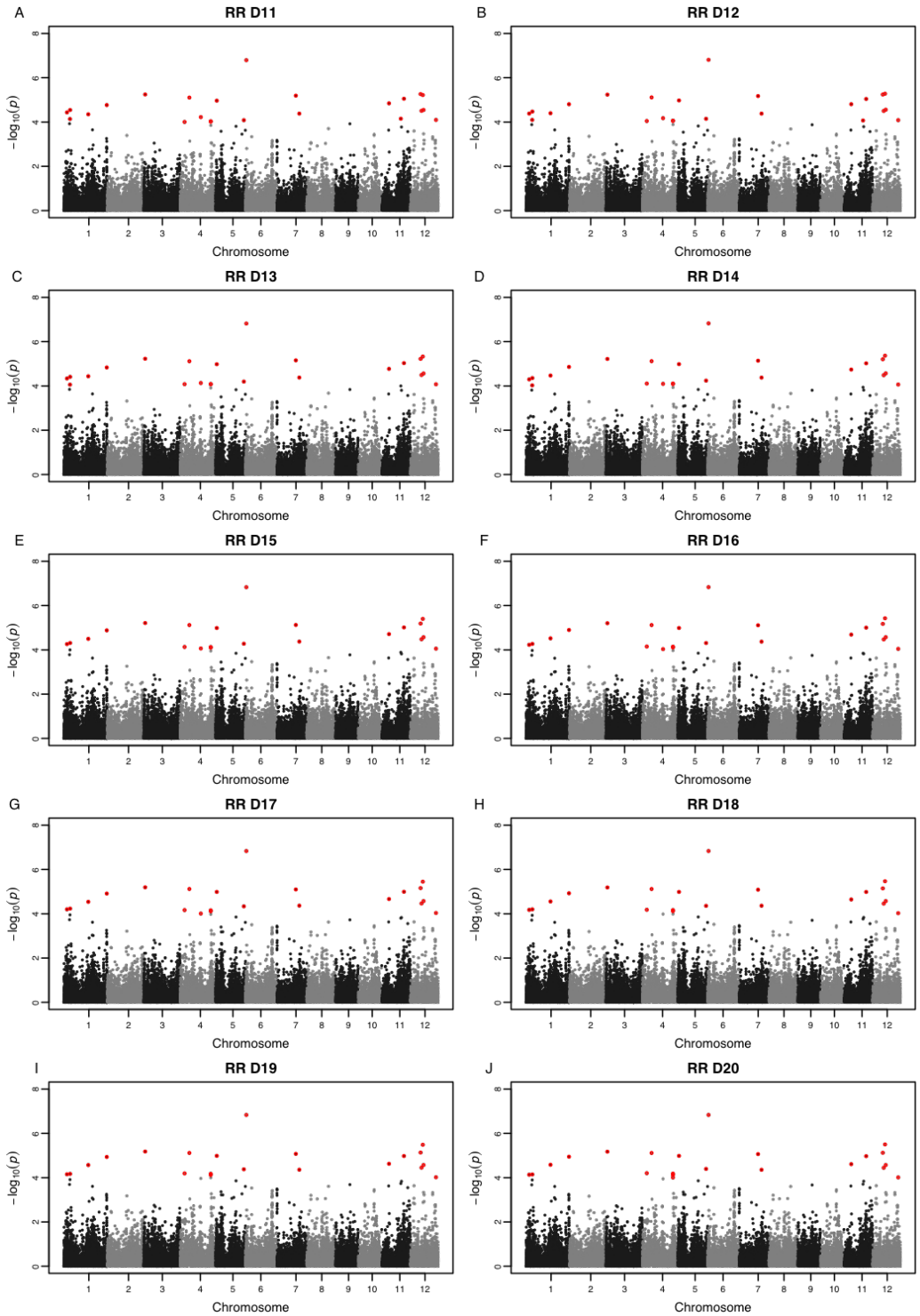

Figure S3: Manhattan plots for RR approach at days 10 to 20. Each panel represents a single time point. -$\log_{10}(p)$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

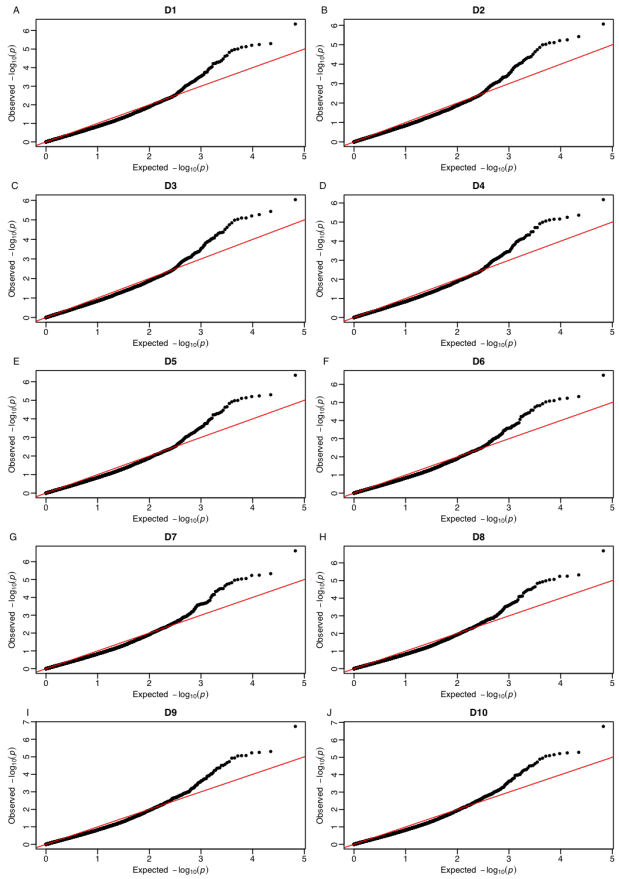

Figure S4: QQ plots for RR approach at days 1 to 10. Each panel represents a single time point. The observed -$\log_{10}(p)$ is shown on the y-axis, while the expected -$\log_{10}(p)$ is shown on the x-axis.

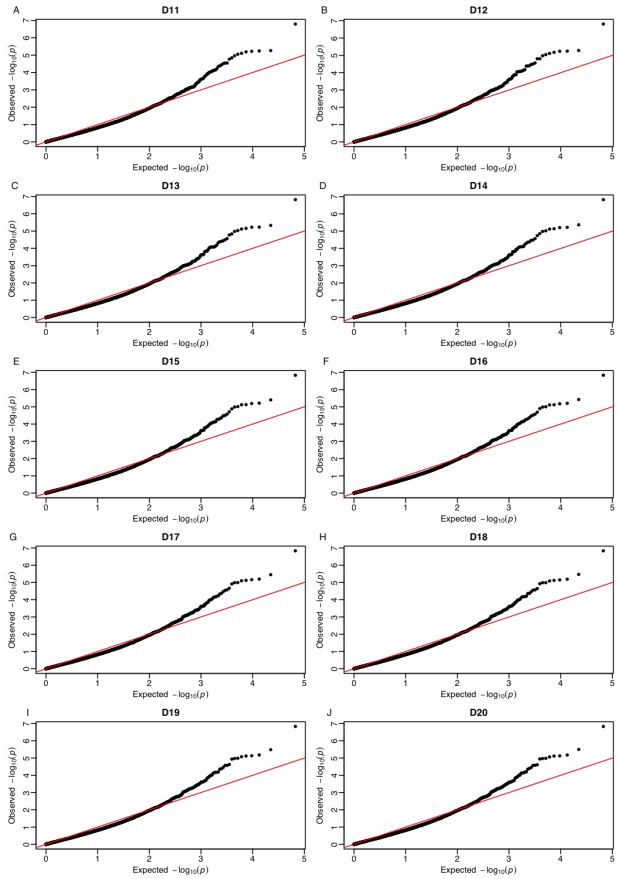

Figure S5: QQ plots for RR approach at days 11 to 20. Each panel represents a single time point. The observed -$\log_{10}(p)$ is shown on the y-axis, while the expected -$\log_{10}(p)$ is shown on the x-axis.

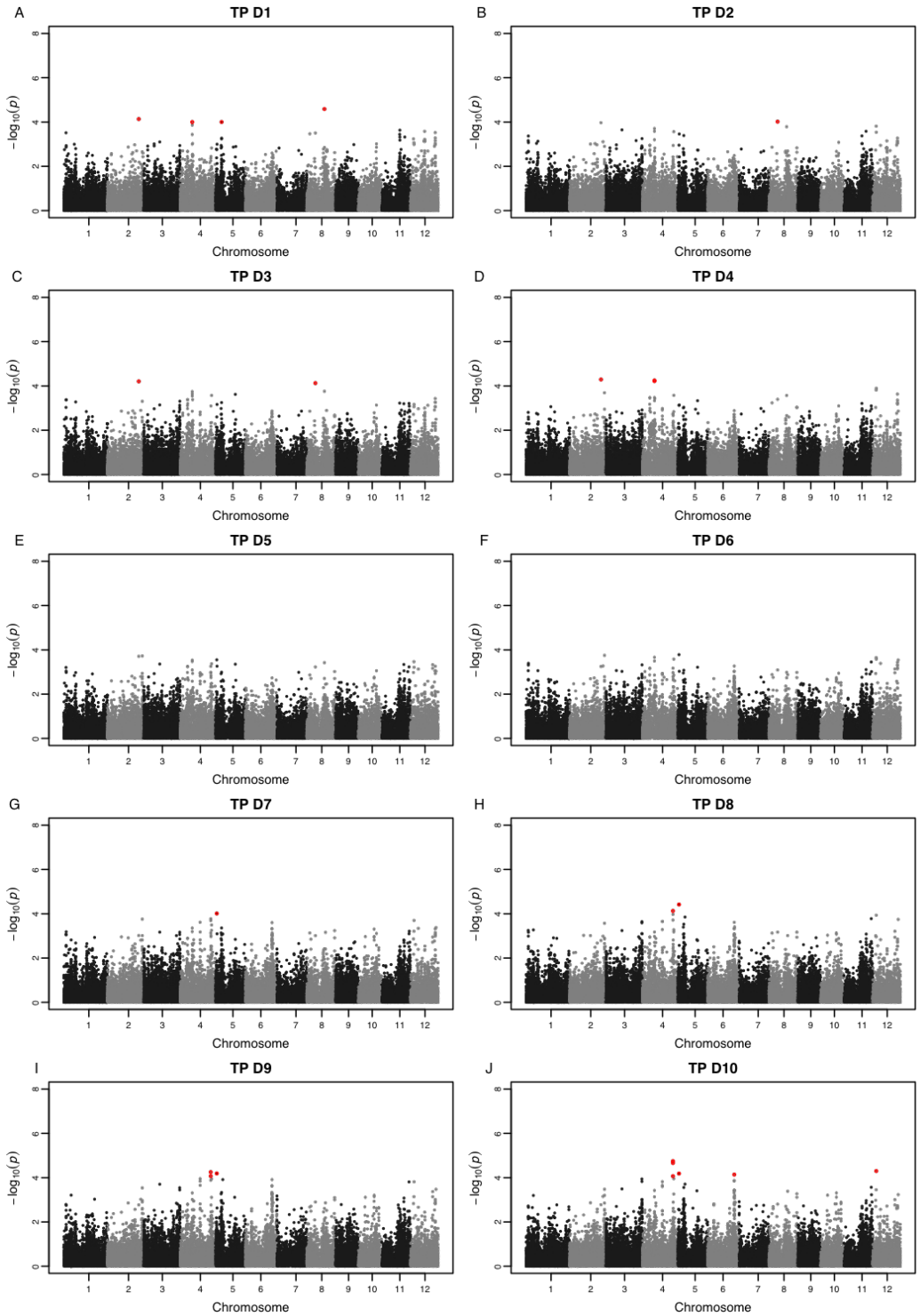

Figure S6: Manhattan plots for TP approach at days 1 to 10. Each panel represents a single time point. -$\log_{10}(p)$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

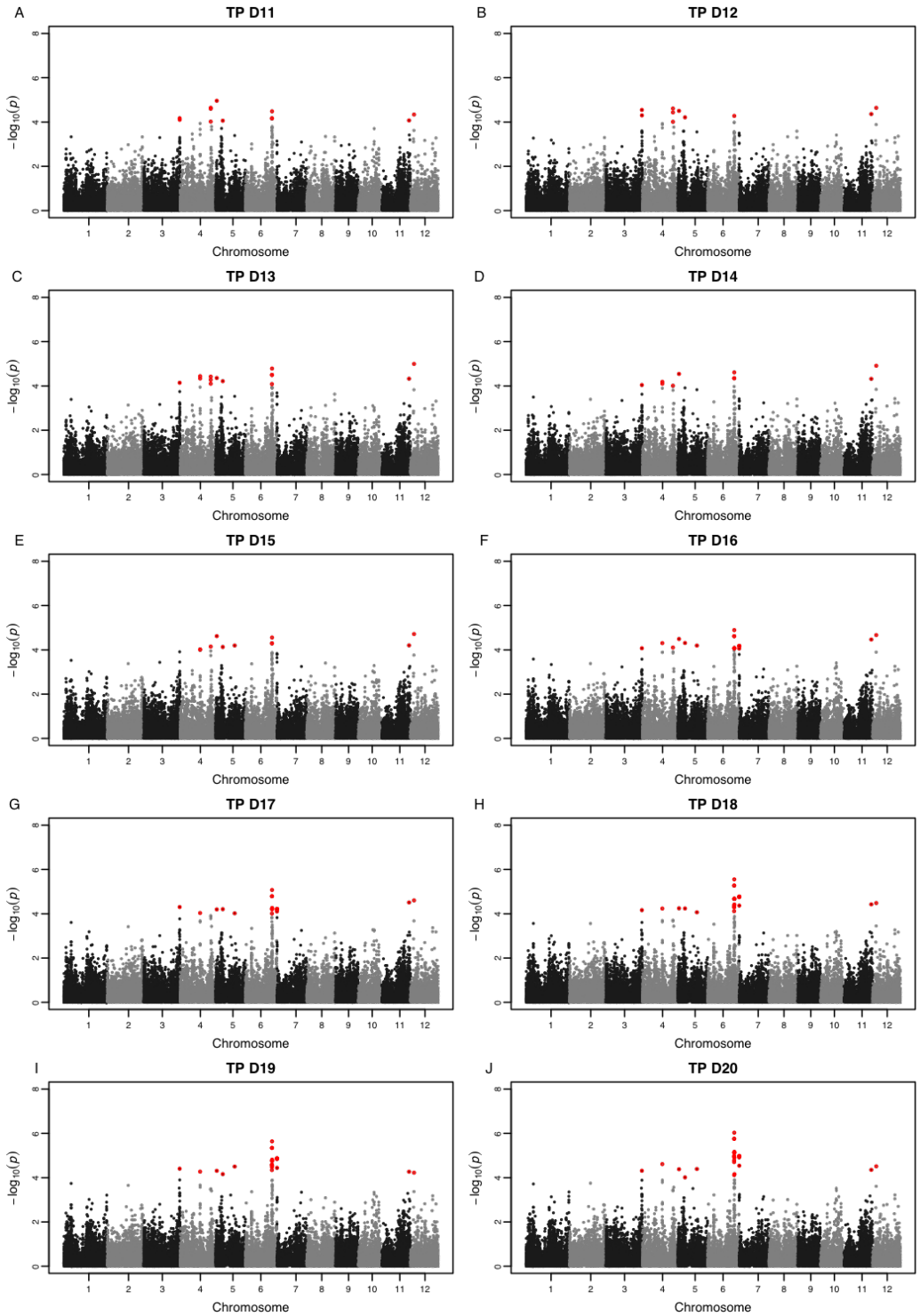

Figure S7: Manhattan plots for TP approach at days 11 to 20. Each panel represents a single time point. -$\log_{10}(p)$ is shown on the y-axis. Statistically significant SNPs are highlighted in red ($p < 1 \times 10^{-4}$).

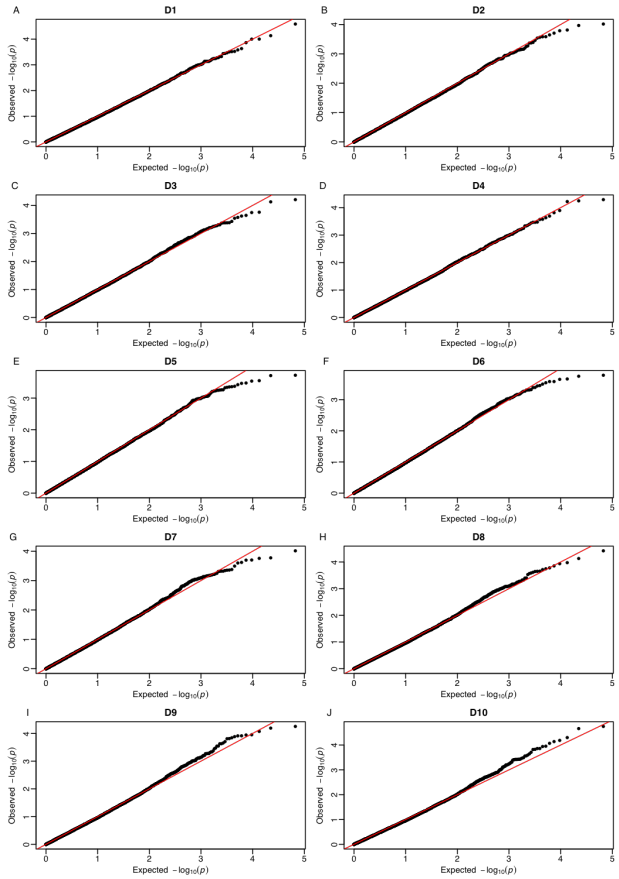

Figure S8: QQ plots for TP approach at days 1 to 10. Each panel represents a single time point. The observed -$\log_{10}(p)$ is shown on the y-axis, while the expected -$\log_{10}(p)$ is shown on the x-axis.

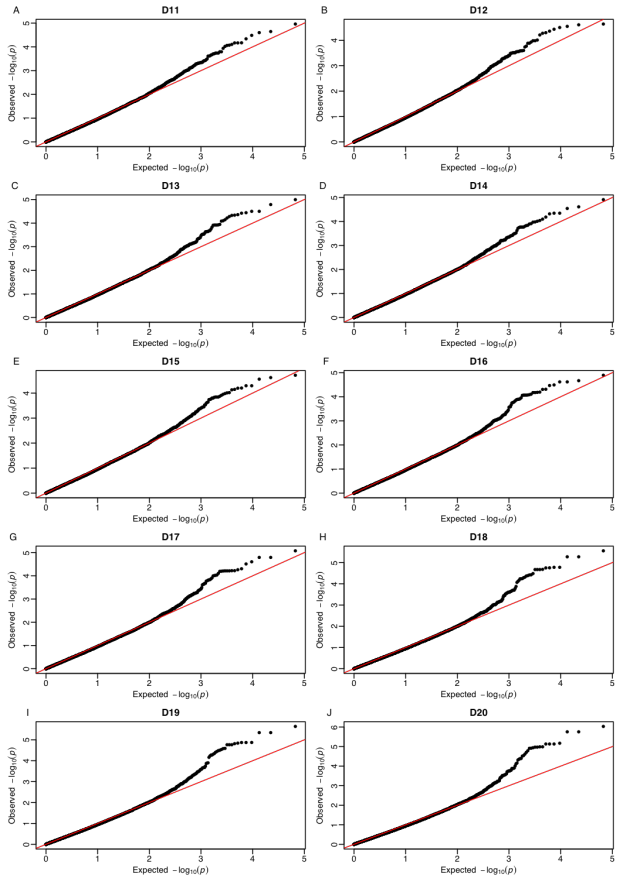

Figure S9: QQ plots for TP approach at days 11 to 20. Each panel represents a single time point. The observed -$\log_{10}(p)$ is shown on the y-axis, while the expected -$\log_{10}(p)$ is shown on the x-axis.

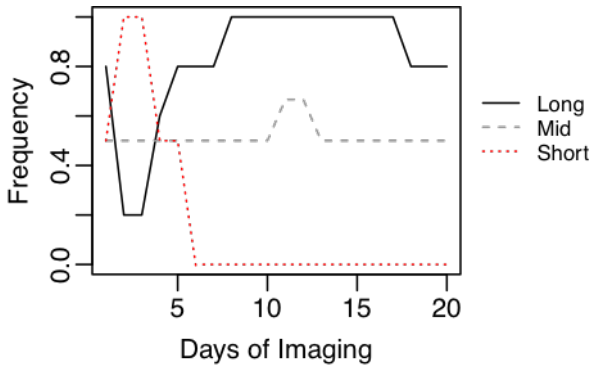

Figure S10: Frequency of time-specific QTL. Long refers to long-duration QTL that were detected on more than 12, but less than 20 days. Mid refers to mid-duration QTL that were detected between 6 to 12-time points. Short indicates short-duration QTL, which were detected at fewer than 6-time points. Frequency was determined by dividing the number of QTL detected at time t by the total number of QTL for a given class.

Appendix

Random regression gBLUP using Legendre polynomials

Polynomial functions are an attractive approach to model longitudinal data, as they require no prior knowledge of the shape of trait trajectories and can be estimated using linear modeling approaches. However, there is often a high correlation between components of the polynomial function. Orthogonal polynomials on the other hand, such as Legendre polynomials, have the same attractive characteristics of polynomial functions and also reduce the correlation between polynomial components. Legendre polynomials are defined on a standardized time interval $[-1, 1]$ using $m = 2 \cdot \frac{t-t_0}{t_n-t_0} - 1$, where t_0 is the first time point with data, and t_n is the last time point in the data set (Kirkpatrick et al., 1990, 1994).

Consider a simple case where we wish to partition a process measured over three time points (y) into fixed (μ) and random (α) time dependant effects. The RR model can be defined as $y(t) = \mu(t) + \alpha(t) + \epsilon$. We can obtain a "full" fit using a second-order Legendre polynomial. The first two Legendre polynomials are $P_o(x) = 1$ and $P_1(x) = x$. Subsequent polynomials can be calculated using $P_t + 1(x) = \frac{1}{n+1}((2n+1)xP_n(x) - nP_{n-1}(x))$. Thus, for $P_2(x)$ the Legendre polynomial is $\frac{1}{2}(3x)P_1(x) - 1P_0(x))$. These Legendre polynomials are then normalized using $\phi_k(t) = \sqrt{\frac{2n+1}{2}}P_k(t)$, giving $\phi_0(t) = 0.7071$, $\phi_1(t) = 1.2247(t)$ and $\phi_2(t) = 2.317(t^2) - 0.7906$. Two matrices can be defined, $\mathbf{\Lambda}$ and $\mathbf{M}$, that store the coefficients for the Legendre polynomials and the standardized time values, respectively.

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{16}$$

$$\mathbf{\Lambda} = \begin{bmatrix} 0.7071 & 0 & 0 \\ 0 & 1.2247 & 0 \\ -0.7906 & 0 & 2.3717 \end{bmatrix} \tag{17}$$

Multiplying the two gives $\mathbf{\Phi}$ where each row vector corresponds to the series of Legendre polynomials at each standardized time interval.

$$\mathbf{\Phi} = \begin{bmatrix} 0.7071 & -1.2247 & 1.5811 \\ 0.7071 & 0 & -0.7906 \\ 0.7071 & 1.2247 & 1.5811 \end{bmatrix} \tag{18}$$

The covariance matrix for the RR coefficients is given by $\mathbf{K}$. The full covariance matrix ($\mathbf{V}$) among all three time points can be obtained via $\mathbf{V} = \mathbf{\Phi K \Phi}'$.

In the following study we aimed to assess the genetic and environmental covariances for shoot growth measured across a period of 20-time points. To this end, we utilized a RR model that modeled the fixed population mean (β) growth trajectories using a second-order Legendre polynomial, and the random genetic (u) and experimental effects (s) using a second-order and first-order Legendre polynomial respectively. Following the example above, these time-dependant processes can be described using a linear combination of $\mathbf{\Phi}$. The covariances at each time point for the random genetic and experimental effects are given by $\mathbf{V}_g = \boldsymbol{\Phi_g \Omega \Phi_g}'$ and $\mathbf{V}_s = \boldsymbol{\Phi_s \mathbf{P} \Phi_s}'$, respectively. The matrices $\mathbf{\Omega}$ and $\mathbf{P}$ represent the covariance matrices for the RR coefficients for the genetic and experimental effects, respectively. Thus, the dimensions of $\mathbf{\Omega}$ 3×3 and $\mathbf{P}$ is 2×2.

Defining the mixed model equation

The following random regression model was used to model trajectories for PSA across the 20-time points and obtain estimates for $\mathbf{\Omega}$ and $\mathbf{P}$

$$\text{PSA}_{tijk} = \mu + \sum_{k=0}^{2} \phi_{jtk}\beta_k + \sum_{k=0}^{2} \phi_{jtk}u_{jk} + \sum_{k=0}^{1} \phi_{itk}s_{ik} + e_{tijk} \tag{19}$$

In matrix notation, the model can be written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Qs} + \mathbf{e} \tag{20}$$

$\mathbf{y}$ is a vector with an order equal to the number of observations and contains the PSA over the 20 days. $\mathbf{X}$ is an covariable matrix for the fixed effects where the number of rows is equal to the number of observations (n) and the number of columns is equal to the order of Legendre polynomial used to model fixed effects (k_f). The matrices $\mathbf{Z}$ and $\mathbf{Q}$ are covariable matrices for the random additive genetic and random experimental effects, respectively. The number of rows for $\mathbf{Z}$ is equal to the number of observations and the number of columns corresponds to the order of Legendre polynomial times the number of lines used to fit the additive genetic effect ($q * k_g = 357 * 3 = 1,071$). For $\mathbf{Q}$ the number of columns would be 6 ($e * k_s = 3 * 2$) and the number of rows would be equal to the number of observations. We assume $\mathbf{u} \sim N(0, \mathbf{G} \otimes \mathbf{\Omega})$, $\mathbf{s} \sim N(0, \mathbf{I} \otimes \mathbf{P})$, and $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{D})$. Here, $\mathbf{\Omega}$ and $\mathbf{P}$ are the covariance matrices for the RR coefficients for the additive genetic and permanent environmental effects. $\mathbf{D}$ is a diagonal matrix that allows for heterogeneous variances over the 20-time points.

The mixed model equation (MME) is

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} & \mathbf{X'R^{-1}Q} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}} \otimes \mathbf{\Omega} & \mathbf{Z'R^{-1}Q} \\ \mathbf{Q'R^{-1}X} & \mathbf{Q'R^{-1}Z} & \mathbf{Q'R^{-1}Q} + \mathbf{I} \otimes \mathbf{P} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \\ \mathbf{Q'R^{-1}y} \end{bmatrix} \tag{21}$$

Solving the above MME will give three RR coefficients for each line for the random genetic effects. Using these RR coefficients, the genetic values at each time point can be obtained as described above. For line j, the predicted genetic values (gBLUP) at each time point is given by $gBLUP_j = \mathbf{\Phi}\hat{u}_j$.

Constructing the covariable matrices

For each term, we define a matrix of Legendre polynomials evaluated at each time point. Recall that both the fixed and random additive genetic effect are modeled using a second-order Legendre polynomial. Thus the matrix of Legendre polynomials for the fixed and random additive genetic effect for the first and last three time points is, ($\boldsymbol{\Phi_f}, \boldsymbol{\Phi_g}$, respectively) are

$$\boldsymbol{\Phi_f} = \boldsymbol{\Phi_g} \qquad = \begin{bmatrix} 0.707 & -1.225 & 1.581 \\ 0.707 & -1.096 & 1.108 \\ 0.707 & -0.967 & 0.688 \\ \vdots & \vdots & \vdots \\ 0.707 & 0.967 & 0.688 \\ 0.707 & 1.096 & 1.108 \\ 0.707 & 1.225 & 1.581 \end{bmatrix} \tag{22}$$

For the environmental effect, the matrix of Legendre polynomials ($\boldsymbol{\Phi_s}$) is of order $t \times 2$

and for the first and last three time points is

$$\mathbf{\Phi_s} = \begin{bmatrix} 0.707 & -1.225 \\ 0.707 & -1.096 \\ 0.707 & -0.967 \\ \vdots & \vdots \\ 0.707 & 0.967 \\ 0.707 & 1.096 \\ 0.707 & 1.225 \end{bmatrix} \tag{23}$$

The covariable matrix $\mathbf{X}$ is defined as $\mathbf{X} = \mathbf{X^o \Phi_f}$ where $\mathbf{X^o}$ is a vector of 1 with length $q * e$. Similarly, we define matrices $\mathbf{Z}$ and $\mathbf{Q}$ as

$$\mathbf{Z} = \mathbf{Z^o} \otimes \mathbf{\Phi_g} \tag{24}$$

$$\mathbf{Q} = \mathbf{Q^o} \otimes \mathbf{\Phi_s} \tag{25}$$

$\mathbf{Z^o}$ and $\mathbf{Q^o}$ are incidence matrices that allocate temporal records to individuals and experiments respectively. The order of $\mathbf{Z^o}$ is $q * e \times q$ and $\mathbf{Q^o}$ is $q * e \times e$ (q is the number of individuals, e is the number of experiments).

Calculating Var($\hat{\beta}$) at each time point

The objective is to calculate SNP effects at each time point. Recall for a univariate gBLUP approach (e.g. the single time point approach), SNP effects can be obtained from breeding values through a simple linear transformation given by

$$\text{BLUP}(\hat{\beta}) = \mathbf{W'_{sc} G^{-1} \hat{u}} \tag{26}$$

Thus,

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\mathbf{W'_{sc}G^{-1}\hat{u}}) \tag{27}$$

$$= \mathbf{W'_{sc}G^{-1}}\mathrm{Var}(\mathbf{\hat{u}})\mathbf{G^{-1}W_{sc}} \tag{28}$$

The prediction error variance (PEV) of $\boldsymbol{\hat{u}}$ is

$$\mathrm{PEV}(\mathbf{\hat{u}}) = \mathbf{C^{22}}\sigma_e^2 = \mathrm{Var}(\mathbf{u} - \mathbf{\hat{u}}) \tag{29}$$

$$= \mathrm{Var}(\mathbf{u}) - \mathrm{Var}(\mathbf{\hat{u}}) \tag{30}$$

$$= \mathbf{G}\sigma_g^2 - \mathrm{Var}(\mathbf{\hat{u}}) \tag{31}$$

By rearranging the equation above we obtain

$$\mathrm{Var}(\mathbf{\hat{u}}) = \mathbf{G}\sigma_g^2 - \mathbf{C^{22}}\sigma_e^2 \tag{32}$$

To calculate the variance of the SNP effects, we can introduce equation 34 into equation 27 giving

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{W'_{sc}G^{-1}}(\mathbf{G}\sigma_g^2 - \mathbf{C^{22}}\sigma_e^2)\mathbf{G^{-1}W_{sc}} \tag{33}$$

$$= \mathbf{W'_{sc}G^{-1}W_{sc}}\sigma_g^2 - \mathbf{W'_{sc}G^{-1}C^{22}G^{-1}W_{sc}}\sigma_{\mathbf{e}}^{\mathbf{2}} \tag{34}$$

At each time point, σ_g^2 is extracted from the corresponding the diagonal element of the matrix of genetic covariances for each time point, given by $\boldsymbol{\Phi_g\Omega\Phi'_g}$.

$\mathbf{C^{22}}$ is obtained by inverting the coefficient matrix of the MME (equation 21), and is of order $q * k_g \times q * k_g$. The diagonal elements of $\mathbf{C^{22}}$ contain the PEV for the RR coefficients

for the additive genetic effect. To obtain the variance of SNP effects at each time point, $\mathbf{C^{22}}$ must be transformed so that the diagonal elements correspond to the PEV for GEBVs at each time point. We will refer to this $q*d \times q*d$ matrix as $\mathbf{C^{22}}^*$. Following (Mrode, 2014), PEV for individual i at each time point can be obtained by taking the diagonal elements of $\text{PEV}_i = \boldsymbol{\Phi_g}\mathbf{C_{ii}}\boldsymbol{\Phi'_g}$. $\mathbf{C_{ii}}$ is a 3×3 submatrix of RR coefficients from $\mathbf{C_{22}}$ for individual i. To extend this approach to the full $\mathbf{C_{22}}$ matrix, we construct a block matrix of $\boldsymbol{\Phi_g}$ ($\boldsymbol{\Phi^*_g}$) via $\boldsymbol{\Phi^*_g} = \mathbf{I} \otimes \boldsymbol{\Phi_g}$ and obtain $\boldsymbol{C^{22*}}$ by

$$\boldsymbol{C^{22*}} = \boldsymbol{\Phi^*_g}\boldsymbol{C^{22}}\boldsymbol{\Phi^{*\prime}_g} \tag{35}$$

Thus, $\mathbf{C^*_{22}}$ is $q*t \times q*t$ and the diagonal elements are the PEV for GEBVs at each time point. Finally, to calculate the variance of SNP effects at each time point at each, we extract the corresponding elements of $\boldsymbol{C^{22*}}$ and introduce them into **??**. Calculation of p-values from this point is straight forward. $\boldsymbol{C^{22}}$ for each time point can be extracted from $\boldsymbol{C^{22*}}$, and calculation of p-values follows the procedures outlined in Materials and Methods.