2018

Functional Modeling of Plant Growth Dynamics

Y. Xu
University of Nebraska - Lincoln

Y. Qiu
University of Nebraska - Lincoln

James C. Schnable
University of Nebraska-Lincoln, schnable@unl.edu

Published May 24, 2018

The Plant Phenome Journal

Science Notes

Functional Modeling of Plant Growth Dynamics

Yuhang Xu, Yumou Qiu, and James C. Schnable*

Core Ideas

- Functional ANOVA methods are beneficial in analyzing time series phenotypic datasets.
- Scoring all plants or plots on the same days is challenging in large-scale experiments.
- Plants phenotyped on non-overlapping days can be compared using functional ANOVA.

Recent advances in automated plant phenotyping have enabled the collection of time series measurements from the same plants of a wide range of traits at different developmental time scales. The availability of time series phenotypic datasets has increased interest in statistical approaches for comparing patterns of change among different plant genotypes and different treatment conditions. Two widely used methods of modeling growth with time are pointwise analysis of variance (ANOVA) and parametric sigmoidal curve fitting. Pointwise ANOVA yields discontinuous growth curves, which do not reflect the true dynamics of growth patterns in plants. In contrast, fitting a parametric model to a time series of observations does capture the trend of growth; however, these models require assumptions regarding the true pattern of plant growth. Depending on the species, treatment regime, and subset of the plant life cycle sampled, these assumptions will not always hold true. We have developed a different approach—functional ANOVA—which yields continuous growth curves without requiring assumptions regarding patterns of plant growth. We compared and validated this approach using data from an experiment measuring the growth of two maize (*Zea mays* L. ssp. *mays*) genotypes under two water availability treatments during a 21-d period. Functional ANOVA enables a nonparametric estimation of the dynamics of changes in plant traits with time without assumptions regarding curve shape. In addition to estimating smooth curves of trait values with time, functional ANOVA also estimates the derivatives of these curves, e.g., growth rates, simultaneously. Using two different subsampling strategies, we demonstrate that this functional ANOVA method enables the comparison of growth curves among plants phenotyped on non-overlapping days with little reduction in estimation accuracy. This means that functional ANOVA based approaches can allow larger numbers of samples and biological replicates to be scored in a single experiment given fixed amounts of phenotyping infrastructure and personnel.

One of the primary goals of both classical and quantitative genetic research is to link genotypic variation to phenotypic variation by identifying specific genetic variants that produce defined changes in phenotype. In the last several decades, advances in DNA sequencing have drastically increased the throughput and decreased the cost of quantifying genotypic variation across individuals. Today, the vast majority of the time and cost of plant genetic research is devoted to capturing and quantifying phenotypic data, a process that remains slow and both cost and labor intensive. The bottleneck of phenotypic data collection has driven interest in automated and high-throughput approaches to collecting plant phenotypes. High-throughput plant phenotyping platforms use cameras or other sensors to capture nondestructive measurements of plant traits from dozens to thousands of plants per day (Fahlgren et al., 2015b; Miller et al., 2007). Because these measurements are both automated and nondestructive, the same traits can be measured from the same plants repeatedly throughout the life cycle of a plant. Unlike single time point measurements, time series trait data enable the quantification of the dynamics of plant growth and development. Biomass data collected from maize recombinant inbred lines and association populations have demonstrated that different genetic loci are identified using data from different time points in development (Muraya et al., 2017; Zhang et al., 2017). However, statistical approaches for both

Y. Xu and Y. Qiu, Dep. of Statistics, Univ. of Nebraska, Lincoln, NE 68503; J.C. Schnable, Center for Plant Science Innovation, Dep. of Agronomy and Horticulture, Univ. of Nebraska, Lincoln, NE 68503.

Plant Phenome J. 1:170007 (2018)
doi:10.2135/tppj2017.09.0007

Received 19 Sept. 2017.
Accepted 14 Feb. 2018.
*Corresponding author (schnable@unl.edu).

Abbreviations: FDA, functional data analysis; QTL, quantitative trait loci.

dealing with the particular complexities of time series phenotypic measurements and extracting as much information as possible from repeated phenotypic measurements remains an ongoing area of development within plant biology and quantitative genetics.

One approach to dealing with high-density time series data is to conduct independent quantitative trait loci (QTL) or association analyses at each individual time point measured (Kwak et al., 2014; Moore et al., 2013). Under the ANOVA setup, we call this method pointwise ANOVA, as ANOVA is performed at each time point individually. However, this approach generally requires that all plants be scored at all time points analyzed. In addition, it does not leverage the potential of repeated measurements to increase the accuracy with which true values at a given time point can be estimated. Another approach is to fit particular functions such as logistic curves to the data (Deng et al., 2012; Xavier et al., 2017). However, this parametric inference approach will produce accurate results only if the assumptions of the growth model function are satisfied by the observed data. Commonly used growth curve models (sigmoidal curves) generally require data from across the entire life cycle of the plant, which can limit the types of phenotyping data to which these models can be applied. For example, many greenhouses or ground-based phenotyping systems can only be used to gather data from plants below a fixed height limit (Fahlgren et al., 2015a; White and Conley, 2013). For taller crops such as maize or bioenergy sorghum [*Sorghum bicolor* (L.) Moench], only a portion of the life cycle can be phenotyped without exceeding these height limits.

Functional data analysis (FDA) (Ramsay and Silverman, 2005; Yao et al., 2005) is another approach that can be applied to the analysis of time series phenotypic datasets. This alternative approach combines many of the strengths of both pointwise ANOVA and parametric modeling approaches to the analysis of time series phenotypic datasets. In FDA, data-driven nonparametric approaches (Cleveland and Devlin, 1988; Fan and Gijbels, 1996; Jacoby, 2000; Ramsay and Silverman, 2005) are used to fit the trend of a data series with time. Unlike pointwise ANOVA, FDA makes very flexible assumptions about the distribution of time points (Yao et al., 2005). Multiple observations taken from the same plant with time will show a degree of correlation, and if correctly harnessed, these correlations can be used to increase the accuracy with which different effects can be estimated. However, this correlation structure is often missed or captured incorrectly by time series analysis. In FDA, a mixed random effect term (Yao et al., 2005) is used to explain the correlation structure among the data. Statistical inference can also be used to obtain confidence bands for the estimated curves, again taking into account the temporal dependence of the data. Functional data analysis has been applied to the analysis of plant phenotypic data in several recent cases. For example, FDA has been used to analyze different levels of variation in root gravitropism data (Xu et al., 2017) and dominant variation in phenotype data has been extracted by FDA and applied to further analysis, such as multivariate QTL mapping (Kwak et al., 2016). Compared with a previously proposed approach of fitting cubic B-spline to individual plants when estimating root growth rates (Beemster and Baskin, 1998), the method proposed here pools information across different plants to provide smoothing estimates for the mean growth curves and genotype, treatment, and their interaction effects with time, together with the derivatives of those functions. Furthermore, by utilizing information on variation across biological replicates, the proposed method can also generate confidence bands around the estimated growth curves.

In studies aimed at comparing genotypes or treatments, optimal experimental design emphasizes collecting measurements as close to simultaneously as possible for all plants within the study to avoid increased variance across measurements resulting from both developmental and diurnal changes in the measured phenotype. However, in larger quantitative genetic studies using high-throughput phenotyping technologies, this requirement for simultaneous data collection can become a major bottleneck limiting the number of plants and number of accessions that can be included within a single experiment. For example, the University of Nebraska–Lincoln's Greenhouse Innovation Center has the capacity to image approximate 400 plants per day, while significantly more total plants can be grown in parallel (Ge et al., 2016). Similar systems such as the Bellwether phenotyping system also have the capacity to grow more plants simultaneously than can be imaged during the course of a single day (Fahlgren et al., 2015a). Phenotypes collected from unmanned aerial vehicles suffer from a similar constraint on how many plots can be imaged per day, with the additional constraint that unsuitable weather conditions—high wind, thunderstorms etc.—can result in missing data from particular sites on particular dates, producing unbalanced final phenotypic datasets. In many cases, FDA can provide a way to address this issue by permitting the reconstruction of growth curves using relatively small numbers of measurements spaced across a large period of development, thus generating predicted values for any time points not scored. Our proposed method can produce a subsample estimator based on half of the observed data for individual plants with only minimal decreases in accuracy relative to estimates constructed from the entire dataset. In addition, we obtain accurate estimator values when different batches of plants are phenotyped on alternating days relative to each other.

Methods

Experimental Design, Growth Conditions, and Imaging

Our B73 plants were grown from a seed source validated using RNA-sequencing single nucleotide polymorphism calling to match the B73 genotype used to generate the maize reference genome (Liang and Schnable, 2016). Fast Flowering Mini-Maize-A seeds were provided by Morgan E. McCaw and have also been subjected to 24× whole genome resequencing (McCaw et al., 2016). All plants were grown at the University of Nebraska–Lincoln's Greenhouse Innovation Center. Plants were sown into 5.7-L pots with Fafard germination mix and watered to a target weight of 5.4 kg. From 6 d

after planting (DAP) to 26 DAP, plants were imaged using an RGB camera from angles offset from each other by 90°. Until 10 DAP, each plant was rewatered to a target weight of 5.4 kg. From 11 DAP (the 6th day since the beginning of imaging) to the end of the experiment, drought-treated plants received no additional water, while well-watered plants continued to be rewatered to a target weight of 5.4 kg each day. Further details on experimental design and growth conditions were provided by Ge et al. (2016).

Extraction of Pixel Counts from RGB Images

An RGB image processing procedure (Ge et al., 2016) was applied to extract plant sizes from the acquired images. A threshold was applied to the contrast of green intensity and the average intensity of red and blue to separate the plant pixels from the background. The majority of the background in our imaging chamber was white. Therefore, the plant areas could be obtained efficiently by such a comparison. The total pixel counts of the extracted plant were considered as a measurement of the plant size.

Pointwise ANOVA Model

Let $y_i(t_j)$ be the area of the ith maize plant measured at time t_j, where $i = 1, ..., n$, $n = 60$ is the sample size, and $j = 1, ..., m$, $m = 20$ is the number of measured days. Define genotype indicator G_i as follows: $G_i = 1$ if the ith maize is of Genotype B73 and $G_i = 0$ if the ith maize is of Genotype FFMM-A. Similarly, define the environment indicator W_i as follows: $W_i = 1$ if the ith maize plant is well watered and $W_i = 0$ if the ith maize plant is water stressed. A natural way to model the growth with time is to use the following pointwise ANOVA model:

$$y_i(t_j) = \mu_j + G_i g_j + W_i w_j + G_i W_i \gamma_j + \varepsilon_i(t_j) \quad [1]$$

where μ_j is the plant area of water-stressed FFMM-A maize at time t_j, g_j is the genotype effect function at time t_j, w_j is the treatment effect function at time t_j, γ_j is the genotype × environment interaction at time t_j, and $\varepsilon_i(t_j)$ is a zero-mean random variable.

It is interesting to know whether genotype × environment interactions exist. To explore this, we tested the genotype × environment interaction in a pointwise manner. The results can be summarized as

Day 1: $P = 0.346$	Day 8: $P = 0.799$	Day 15: $P = 0.834$
Day 2: $P = 0.579$	Day 9: $P = 0.495$	Day 16: omitted
Day 3: $P = 0.696$	Day 10: $P = 0.592$	Day 17: $P = 0.869$
Day 4: $P = 0.622$	Day 11: $P = 0.705$	Day 18: $P = 0.997$
Day 5: $P = 0.662$	Day 12: $P = 0.886$	Day 19: $P = 0.915$
Day 6: $P = 0.761$	Day 13: $P = 0.687$	Day 20: $P = 0.737$
Day 7: $P = 0.851$	Day 14: $P = 0.675$	Day 21: $P = 0.793$

Because all genotype × environment interactions were insignificant, we revised Eq. [1] and used the following pointwise ANOVA model:

$$y_i(t_j) = \mu_j + G_i g_j + W_i w_j + \varepsilon_i(t_j) \quad [2]$$

The resulting estimates are denoted as $\hat{\mu}_j$, $\hat{g}_j$, and $\hat{w}_j$, where $j = 1, ..., m$. Interpolating the corresponding estimates resulted in Fig. 1. However, the estimated functions are not smooth. We advocate the following functional ANOVA method.

Functional ANOVA Model

We assume the following functional ANOVA model for plant growth:

$$y_i(t) = \mu(t) + G_i g(t) + W_i w(t) + \varepsilon_i(t) \quad [3]$$

where $\mu(t)$ is the growth function of the water-stressed FFMM-A maize, $g(t)$ is the genotype effect, $w(t)$ is the treatment effect, and $\varepsilon_i(t)$ is a zero-mean random process. We assume $\mu(t)$, $g(t)$, and $w(t)$ are smooth functions with continuous second derivatives, which is a key difference between pointwise ANOVA and functional ANOVA. To recover the underlying functions and their dynamics, namely velocity and acceleration, we use penalized smoothing splines (Ramsay and Silverman, 2005).

We first represent $\mu(t)$ using a rank K spline basis expansion:

$$\mu(t) = \sum_{j=1}^{K} \beta_{\mu,j} B_{r_1,j}(t)$$

where $\beta_{\mu,j}$ is a coefficient and $B_{r_1,j}(t)$ is an order r_1 B-spline basis function. We chose $K = 12$ for a reduced rank representation and let B-spline basis functions have equally spaced interior knots on [0,20]. Because we were interested in estimating velocity and acceleration functions smoothly, we chose order $r_1 = 6$. Define

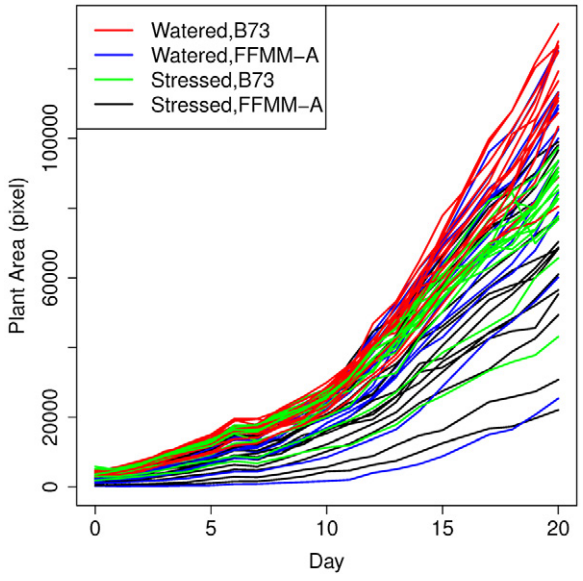

Fig. 1. The comparison of (a) estimated growth curves and (b) the estimated main effect functions using pointwise ANOVA and functional ANOVA for the dataset including two genotypes and two treatments. The estimated curves with the open circles are pointwise ANOVA estimates.

$\boldsymbol{\beta}_\mu = (\beta_{\mu,1}, ..., \beta_{\mu,K})^T$ and $\mathbf{B}(t) = (B_{6,1}, ..., B_{6,K})^T(t)$. Denote the r_2th derivative of $\mathbf{B}(t)$ as $\mathbf{B}^{(r_2)}(t)$. Then $\mu(t)$ can be rewritten as $\mu(t) = \mathbf{B}(t)^T\boldsymbol{\beta}_\mu$. Similarly, we approximate other functions as $g(t) = \mathbf{B}(t)^T\boldsymbol{\beta}_g$ and $w(t) = \mathbf{B}(t)^T\boldsymbol{\beta}_w$. To estimate the vectors of parameters $\boldsymbol{\beta}_\mu$, $\boldsymbol{\beta}_g$, and $\boldsymbol{\beta}_w$, penalized smoothing splines minimize the following penalized sum of squares:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left[y_i(t_j) - \mathbf{B}(t_j)^T\boldsymbol{\beta}_\mu - G_i\mathbf{B}(t_j)^T\boldsymbol{\beta}_g - W_i\mathbf{B}(t_j)^T\boldsymbol{\beta}_w\right]^2 + \lambda_1\boldsymbol{\beta}_\mu^T\boldsymbol{\Omega}\boldsymbol{\beta}_\mu + \lambda_2\boldsymbol{\beta}_g^T\boldsymbol{\Omega}\boldsymbol{\beta}_g + \lambda_3\boldsymbol{\beta}_w^T\boldsymbol{\Omega}\boldsymbol{\beta}_w \quad [4]$$

where λ_l for l = 1, 2, 3 are smoothing parameters and $\boldsymbol{\Omega} = \int \mathbf{B}^{(r_2)}(t)[\mathbf{B}^{(r_2)}(t)]^T dt$ is a penalty matrix. Let $\lambda = \lambda_1 = \lambda_2 = \lambda_3$ for simplicity and set $r_2 = 4$ because we penalize the second derivatives. For a given smoothing parameter λ, an explicit form of solutions can be obtained when minimizing Eq. [4] (Ramsay and Silverman, 2005). Generalized cross-validation (GCV) is a popular method to choose smoothing parameters (Ramsay and Silverman, 2005). The GCV function is a smooth function of λ. To locate the optimal smoothing parameter that minimizes the GCV function, we used a simple grid search approach. After the optimal smoothing parameter was found, the penalized sum of squares in Eq. [4] was minimized to obtain the estimates $\hat{\boldsymbol{\beta}}_\mu$, $\hat{\boldsymbol{\beta}}_g$, and $\hat{\boldsymbol{\beta}}_w$. Accordingly, the obtained estimates for the smooth functions are $\hat{\mu}(t) = \mathbf{B}(t)^T\hat{\boldsymbol{\beta}}_\mu$, $\hat{g}(t) = \mathbf{B}(t)^T\hat{\boldsymbol{\beta}}_g$, and $\hat{w}(t) = \mathbf{B}(t)^T\hat{\boldsymbol{\beta}}_w$. The confidence bands for the estimated curves $\hat{\mu}(t)$, $\hat{g}(t)$, and $\hat{w}(t)$ can be obtained by a linear transformation of the joint confidence intervals of the regression coefficients $\boldsymbol{\beta}_\mu$, $\boldsymbol{\beta}_g$, and $\boldsymbol{\beta}_w$ of the B-spline basis functions in Eq. [4]. The 95% confidence bands for the estimated genotype and treatment effects in our study were calculated by using the "fda" R package.

One advantage of using the penalized smoothing splines technique is that it readily yields different derivatives of the target smooth curves. For example, the estimates of the first and second derivative of $\mu(t)$ are $\hat{\mu}^{(1)}(t) = [\mathbf{B}^{(1)}(t)]^T\hat{\boldsymbol{\beta}}_\mu$ and $\hat{\mu}^{(2)}(t) = [\mathbf{B}^{(2)}(t)]^T\hat{\boldsymbol{\beta}}_\mu$, respectively. In general, Eq. [4] can be adapted to allow the number of observations to be different for each plant and the time points to be unequally spaced, which is a significant advantage of the functional ANOVA approach relative to conventional pointwise ANOVA. This advantage can be quite useful in determining imaging strategies. For example, image data could be collected at higher density early in development, when error and/or plant-to-plant variation is high, and at lower densities close to maturity when growth rates are low and the ratio of measurement error to mean values also declines.

Overall, when using the complete dataset, the difference between the estimates provided by the two methods was relatively small, as shown in Fig. 1. However, note that for the pointwise ANOVA, the estimates are fitted at each time point t_j, so the obtained growth curves and main effects curves are discontinuous, which does not reflect the natural growth of plants. In contrast, the functional ANOVA assumes that the main effects and interactions are smooth functions with time with continuous second derivatives. The dynamics of plant development (namely velocity and acceleration), as well as confidence bands for those curves, can be obtained by the functional ANOVA, which cannot be provided by the pointwise ANOVA.

Results

The plant high-throughput phenotyping datasets used in this study were taken from a factorial experiment with 60 plants divided equally into two genotypes (B73 and FFMM-A) and equally into two treatments (well watered and drought stressed) (Fig. 2) (Ge et al., 2016). The two genotypes were selected because B73 is a widely used reference genotype that is a typical representative of

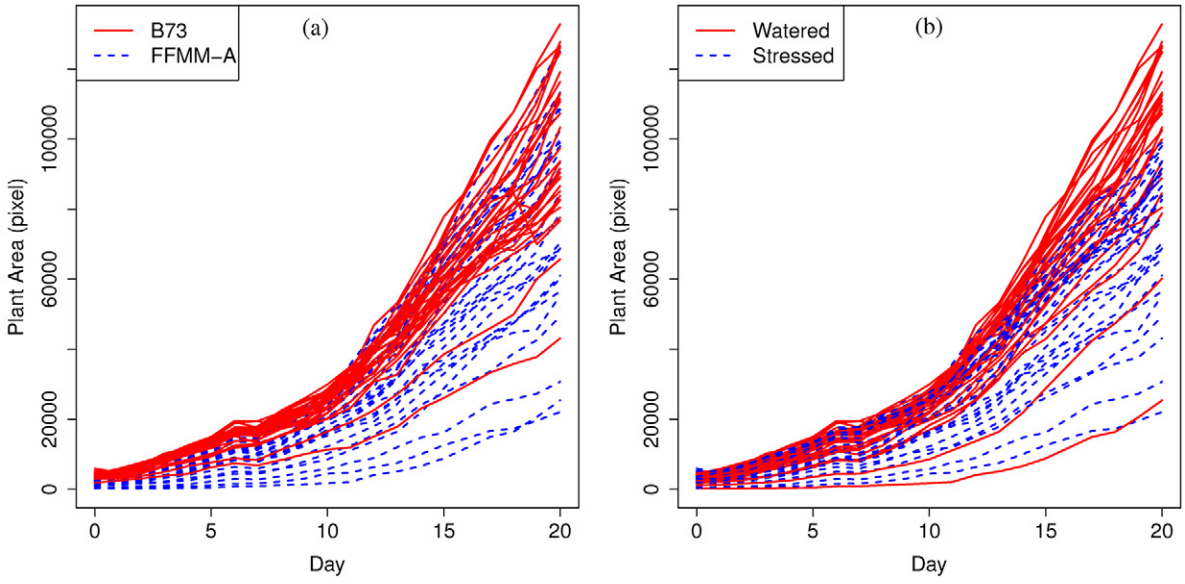

Fig. 2. Estimated plant size for each individual phenotyped within this dataset, which includes two genotypes and two water treatments. Day 0 corresponds to 6 d after planting; Day 20 corresponds to 26 d after planting.

moderate temperate maize, while FFMM-A represents one extreme end of the distribution of plant life cycle speed and plant architecture present within domesticated maize. Plants were imaged daily for a total of 21 d, excluding Day 16 as a result of a technical failure. As previously reported, B73 grew faster and larger than FFMM-A, and well-watered maize grew faster and larger than drought-stressed plants (Fig. 3) (Ge et al., 2016).

Estimating Genotype and Treatment Effects Using Spline Fitting

Vegetative biomass accumulation in maize and many other crops is generally assumed to follow a sigmoidal growth curve (Erickson, 1976). The cumulative increase of total C fixed as the plant produces additional leaves enables the growth of either more or larger leaves, creating the acceleration portion of the growth curve, while later in development much C is devoted to reproductive development, slowing the accumulation of additional vegetative biomass, which ultimately plateaus, producing a final S-shaped curve. The dataset used in this study did not extend into reproductive development and thus captured only the first phase of the sigmoidal biomass accumulation pattern, producing J-shaped curves as shown in Fig. 3.

Applying penalized spline smoothing to the data, we obtained the estimated growth under different conditions shown in Fig. 4a. As expected, for each genotype, well-watered plants were consistently larger than drought-stressed plants. In addition, plants from the accession B73 were consistently larger than those of FFMM-A. At early stages of plant development, genotype played a larger role in determining plant biomass than did water treatment. From Day 1 to Day 16, both well-watered and drought-stressed B73 plants

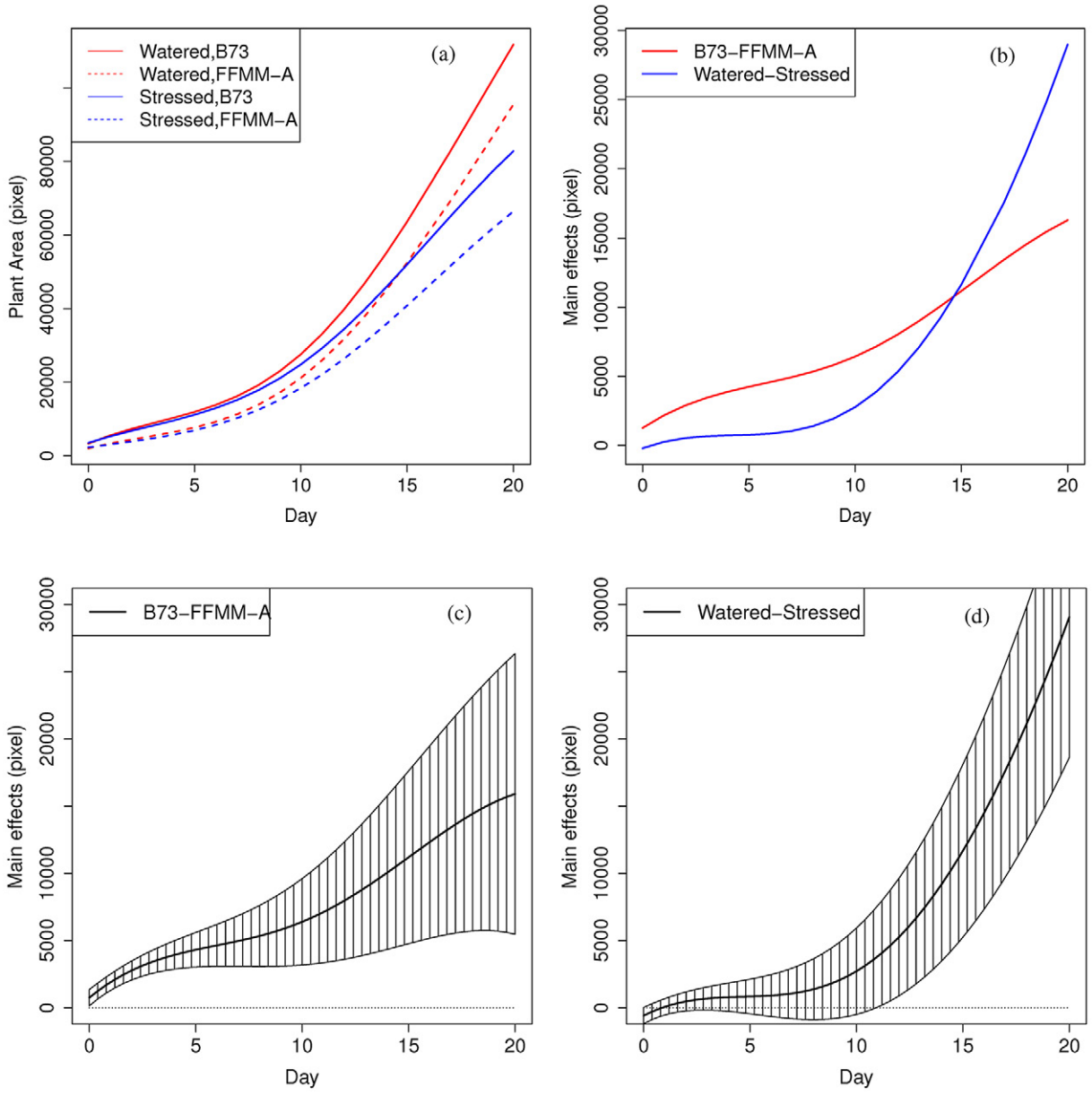

Fig. 3. Plants classified based on (a) genotype and (b) water treatment.

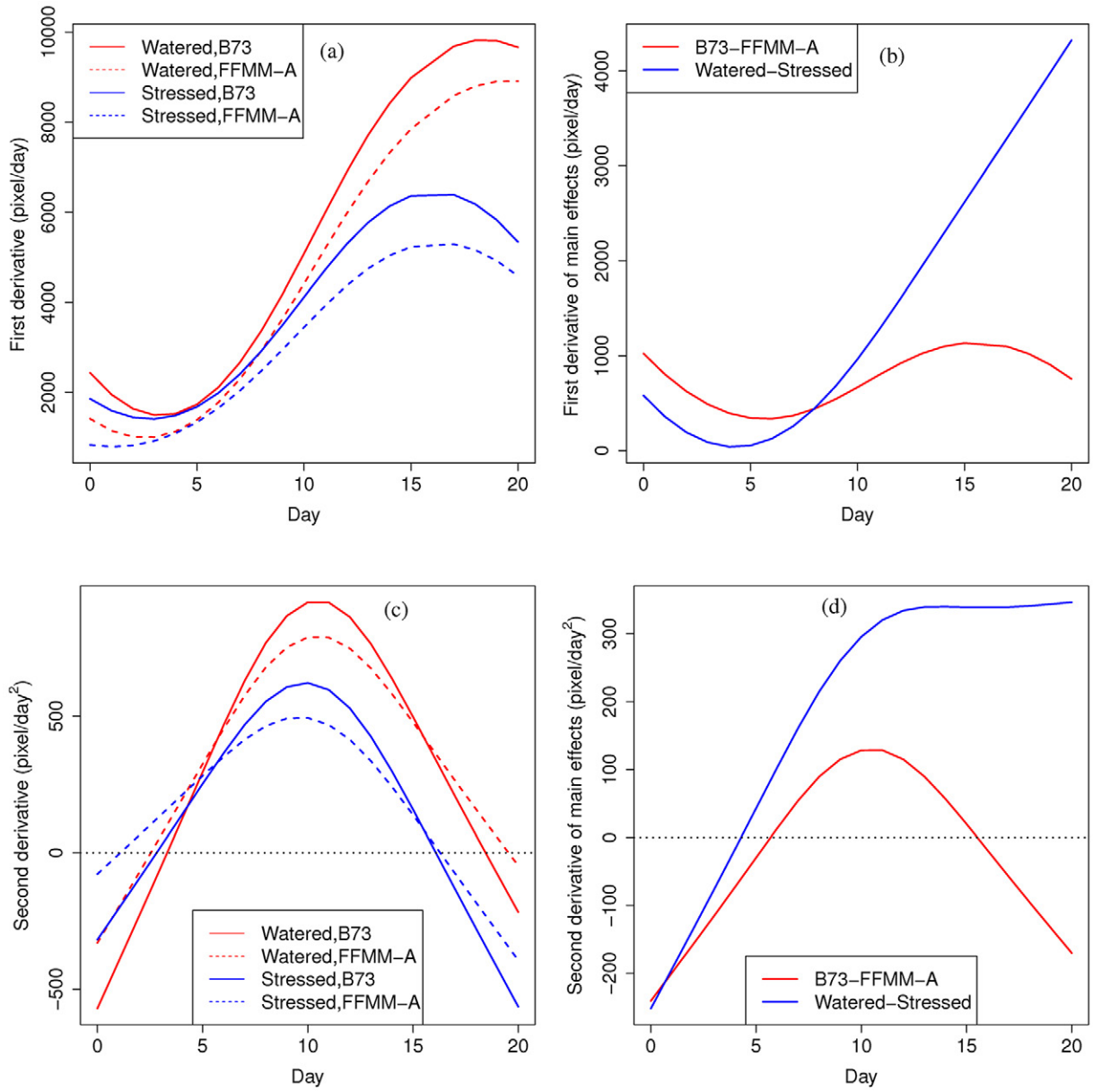

Fig. 4. (a) Growth curves estimated for each genotype–treatment combination, (b) estimated effect sizes for genotype and treatment, (c) estimated genotype effect with 95% confidence bands, and (d) estimated treatment effect with 95% confidence bands.

were consistently larger than FFMM-A plants in either water treatment. After Day 16, the biomass of well-watered FFMM-A exceeded that of drought-stressed B73.

Figure 4b shows the estimated main effect functions for both genotype and water treatment. Both effect functions are monotonically increasing, but they exhibit very different shapes. The effect function for genotype is close to linear and increases steadily. The 95% confidence band in Fig. 4c shows that the effect function is significant throughout the whole experimental period. However, the effect function for water treatment shows an obvious J shape, starting at a low value and increasing very slowly for the first third of the experiment and then growing rapidly. The estimated treatment effect function is close to zero during the first few days because the drought stress started from Day 6. The 95% confidence band in Fig. 4d indicates that the treatment effect function is not significant until the second half of the experiment, which is reasonable because drought stress may take a few days to act significantly. The intersection of the two main effects functions coincides with the finding in Fig. 4a.

Dynamic Changes in Growth Rate

The results for the dynamics of the growth curves are summarized in Fig. 5. Figure 5a shows the estimated growth velocity functions under different conditions. Similarly for each genotype, the growth velocity for non-water-stressed maize was consistently higher than for the drought-stressed maize; within each water treatment, B73 consistently grew faster than FFMM-A. All growth velocity curves show an S shape: during the early period,

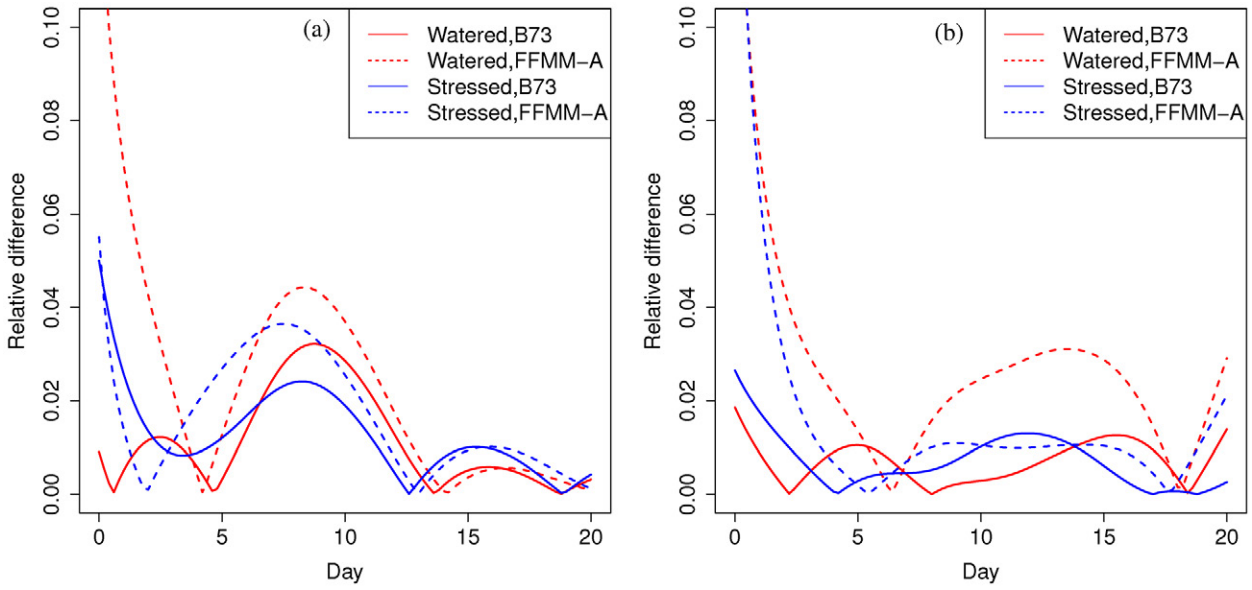

Fig. 5. The estimated first derivative of (a) growth curves and (b) the main effect function of genotype and water treatment, and the estimated second derivative of (c) growth curves and (d) the main effect function of genotype and water treatment.

the growth rates decreased slightly; during the middle period, the growth rates increased sharply; for the later period, the growth rates declined again. Interestingly, the early period for B73 was about 2 or 3 d longer than for FFMM-A, and the late period of non-water-stressed maize was about 2 or 3 d longer than that of drought-stressed maize. From the velocity perspective, this coincides again with the finding that the genotype effect plays an important role during the early period but the treatment effect plays an important role during the late period. Figure 5b shows the estimated main effect velocity functions. Similarly, the two main effect functions show different shapes: after the early period of decrease, the genotype effect on the rate of growth increases slightly followed by a decrease, but the watering effect on the rate of growth increases sharply and keeps increasing.

Figure 5c shows the estimated growth acceleration functions under different conditions. Each curve in Fig. 5c exhibits a parabola-like shape, with the maximum acceleration located around the 10th day of the experiment. Figure 5d shows the estimated main effect acceleration functions. The treatment effect on the acceleration of growth seems consistently higher than the genotype effect except for the first few days. Both acceleration functions increased during the first half of the experiment. However, for the latter half, the watering effect on acceleration became close to a constant, about 340, whereas the genotype effect on acceleration decreased dramatically.

Functional ANOVA for Comparing Growth with Non-overlapping Time Points

To investigate the estimation efficiency when plants were not phenotyped every day and to test the prediction accuracy of functional ANOVAs for plant areas when phenotype measurements were not recorded, comparisons via cross-validation were made between the full dataset and subsampled datasets.

The data were subsampled in two ways. In the first scenario, only measurements from odd-numbered days were retained (10 d in total) for all the plants. This subsampling tested the effect of reducing the number of days of imaging for a single experiment, allowing more independent experiments to be conducted in parallel using the same infrastructural capacity for phenotypic data acquisition. We named the first scenario "subsampling by dates". In the second scenario, all the plants were equally divided into two groups among the two genotypes and two treatments. For the first group, only measurements from odd-numbered days were retained, while for the remaining plants in the second group, only measurements from even-numbered days were retained. This subsampling tested the effect of measuring different subsets of plants in an experiment at different time points, which would allow experiments with a large number of genotypes or large sample sizes within each genotype to be conducted given a fixed facility capacity for phenotypic data acquisition. We named the second scenario "subsampling by plant replicates".

Growth rates together with genotype and treatment effects were estimated using the two subsampling approaches described above using the same functional ANOVA procedure as for the full dataset. With the exception of the first several days when all the plants were quite small, as shown in Fig. 6, functional ANOVA with half of the data produced reliable estimates that were within 5% deviation from the estimation using the entire data set. One explanation for the large relative difference at early time points—when the plants were small—is that for small plants, the ratio of the variation of the phenotypic trait over its mean value is high. In this study, the variation included the biological variation among different plants and measurement error from extracting plant features from images. When plants are small, the measurement error may be large compared with the mean value of the traits, which would

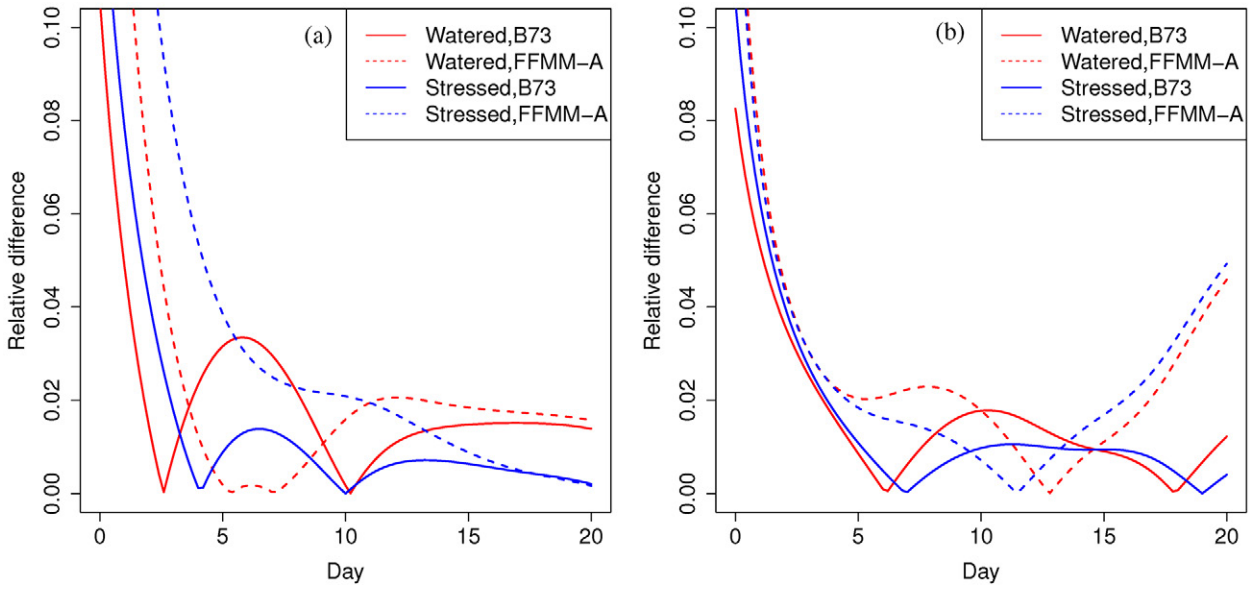

Fig. 6. The relative difference in estimated growth curves between the whole dataset, which includes two genotypes and two treatments, and (a) a dataset subsampled to include data from only every other day (subsampling half by dates, all plants measured on the same days) and (b) a dataset subsampled by plant replicates (plants split into two groups measured on alternating days).

result in higher SD/mean ratios at early developmental stages of plants. Compared with the results when all the plants were phenotyped every day, the average relative estimation difference caused by reducing the imaging frequency was 1.64% (subsampling by dates), and that caused by decreasing the number of daily phenotyped plants was 1.45% (subsampling by plant replicates). Given those small estimation differences, the functional ANOVA approach is able to recover the entire genotype and treatment effects with time even if the plant images are recorded on only half the time of the whole experiment or only half of those plants are imaged every day. Data were further subsampled to one-fourth of the all data points collected. In subsampling by dates, data from all plants from only 5 d (1st, 6th, 11th, 16th, and 21st days) were used; in subsampling by plant replicates, the 60 plants were divided into four groups with a roughly equal size within each combination of treatment and genotype. Only one group of data was used to construct the dataset used for functional data analysis. As shown in Fig. 7, the relative estimation differences remain quite small, with an average of 2.82% in the case of subsampling by dates and 2.13% in the case of subsampling by plant replicates. When the amounts of data become even smaller, the estimates become much less accurate for subsampling by dates because there are not enough time points

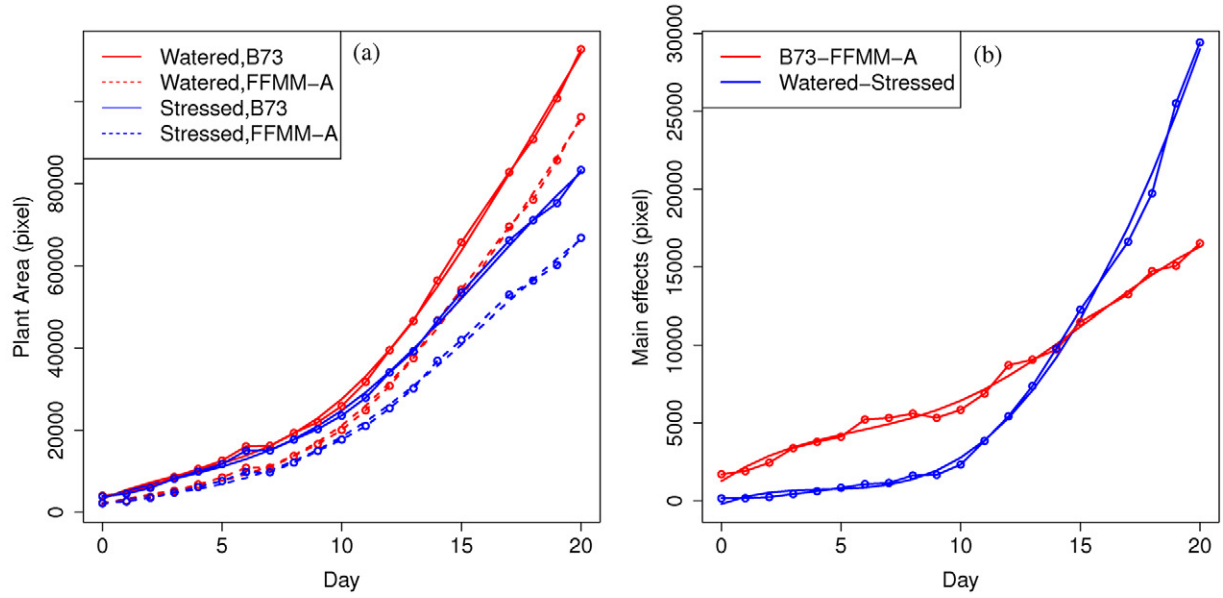

Fig. 7. The relative difference in estimated growth curves between the whole dataset, which includes two genotypes and two treatments, and (a) a dataset subsampled to include data from only every 4 d (subsampling 1/4 by dates, all plants measured on the same days) and (b) a dataset subsampled by plant replicates (plants split into four groups measured alternately on different days).

to estimate the whole growth curve. Accuracy also decreases in subsampling by plant replicates, but the rate of decrease is slower because data from different combinations of genotype and treatment are available on all the days.

One natural question is how to go about identifying optimal sampling frequencies for imaging when different levels of priority are given to minimize experimental cost or maximize measurement accuracy. One general rule of thumb is that if the curvature of the growth function or the degree of measurement error is high, the number of data points needed to estimate the underlying curve with a given level of accuracy will increase (Ramsay and Silverman, 2005). In the dataset used here, measurement error was moderate and the curvature of the estimated underlying functions was relatively low, and thus once-per-day measurements represents a denser sampling scheme than necessary as demonstrated by the results when subsampling one-half or one-quarter of the total data points. The topic of identifying optimal sampling rates in functional datasets is an area of ongoing investigation (Ji and Müller, 2017). The proposed technical methods may hold promise in working with larger and more complex phenomic datasets than the one used for initial investigation of functional data analysis for plant phenomics.

Discussion

In this study, we used functional data analysis as a nonparametric method to model plant growth with time. This nonparametric approach effectively incorporates neighborhood information when fitting the underlying growth curve and produces more accurate estimates of genotype and treatment effects with time (Ramsay and Silverman, 2005). Intuitively, plant biomass at time t_0 is highly related to that at the previous time point, $t_0 - 1$, and the following one, $t_0 + 1$. This sharing of data between nearby time points also provides increased accuracy for predictions of plant traits at time points not sampled in the experiment. Unlike parametric approaches, the functional data analysis method outlined above is data driven rather than model driven and thus is applicable to a wider range of treatments, genotypes, and developmental stages and adaptive to temporally dependent observations. Compared with parametric modeling approaches, nonparametric methods such as the one used in this study are flexible with regard to patterns of growth that do not match prior assumptions regarding the growth pattern of plants. In addition, they adjust for the temporal dependence effect in statistical inference, which is generally not considered in parametric approaches.

The nonparametric regression approach used in this study requires fewer assumptions about how traits change with time than fitting parametric curves to data. However, the translation of a curve defined by a single function into a small number of quantitative phenotypic variables that can be used for mapping genetic variants through QTL or genome-wide association study analysis is a more straightforward process than performing the same translation for a nonparametrically defined curve. For example, after fitting a sigmoidal growth curve, a researcher might perform separate quantitative genetic analyses to identify the genetics controlling the timing of the inflection point of the curve, the slope of the curve at the inflection point, and the total change in value between the bottom and total horizontal asymptotes. Further work is needed to identify the most informative summary statistics for describing the behavior of nonparametrically defined curves such as the timing of the point with the highest first derivative value, the maximum value of the first derivative, etc.

The mean function $\mu(t)$ and the effect functions $g(t)$ and $w(t)$ are obviously significant for the data we have analyzed. However, sometimes there might be many effect functions in the functional ANOVA model and some effect functions are close to zero, so it would be essential to test whether these effect functions are zero or not. For this purpose, generalized likelihood ratio tests may be conducted (Fan et al., 2001), but this was out of the scope of this study. The pointwise ANOVA requires observations from all combinations of genotypes and treatments at the same time point for analyzing the genotype × environment interaction effects. This may not be feasible in some experimental designs because the imaging process of all the plants cannot be finished within a single day. In contrast, functional ANOVA enables genotype × environment interaction analysis on such non-overlapping datasets by borrowing information from the adjacent dates. The number and the location of knots used in this study may not be optimal. However, because both the number of knots and the smoothing parameter control the smoothness of the functions in penalized splines, choosing the number of knots in penalized splines is not as important as it is in regression splines. We used equally spaced knots because of their simplicity and also due to the fact that the mean effect functions are relatively smooth.

Finally, we demonstrated that our proposed method is robust to missing data and non-overlapping sampling dates between subsets of samples within a single experiment. The necessity to collect measurements from all or nearly all individuals at each time point within an experiment is a major constraint on high-throughput phenotyping studies in both the greenhouse, where plant measurements are limited by the throughput of imaging systems, and the field, where plant measurements are limited by the availability of human labor and weather suitable for phenotyping. The wider adoption of functional data analysis in the analysis of plant phenotyping data and awareness of the increased flexibility it provides for sampling data within experimental designs should lead to larger and more statistically robust experiments in the future.

Additional information

The raw image data used in this study are hosted at CyVerse under doi:10.7946/P22K7V.x.

Acknowledgments

We wish to acknowledge Yufeng Ge and Piyush Pandey for their willingness to share prepublication data, Yang Zhang for critical evaluation of an early draft of this manuscript, and Zhikai Liang for helpful feedback during the initial conception of this experiment.

Author contributions

Yuhang Xu, Yumou Qiu, and J.C. Schnable conceived the experiments, Yuhang Xu and Yumou Qiu conducted the analysis, and Yuhang Xu, Yumou Qiu, and J.C. Schnable wrote the manuscript. The authors have no competing financial interests.

References

Beemster, G.T.S., and T.I. Baskin. 1998. Analysis of cell division and elongation underlying the developmental acceleration of root growth in *Arabidopsis thaliana*. Plant Physiol. 116:1515–1526.

Cleveland, W.S., and S.J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83:596–610.

Deng, J., J. Ran, Z. Wang, Z. Fan, G. Wang, M. Ji, et al. 2012. Models and tests of optimal density and maximal yield for crop plants. Proc. Natl. Acad. Sci. 109:15823–15828. doi:10.1073/pnas.1210955109

Erickson, R.O. 1976. Modeling of plant growth. Annu. Rev. Plant Physiol. 27:407–434. doi:10.1146/annurev.pp.27.060176.002203

Fahlgren, N., M. Feldman, M.A. Gehan, M.S. Wilson, C. Shyu, D.W. Bryant, et al. 2015a. A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in *Setaria*. Mol. Plant 8:1520–1535. doi:10.1016/j.molp.2015.06.005

Fahlgren, N., M.A. Gehan, and I. Baxter. 2015b. Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. Curr. Opin. Plant Biol. 24:93–99. doi:10.1016/j.pbi.2015.02.006

Fan, J., and I. Gijbels. 1996. Local polynomial modelling and its applications. Monogr. Stat. Appl. Probab. 66. CRC Press, Boca Raton, FL.

Fan, J., C. Zhang, and J. Zhang. 2001. Generalized likelihood ratio statistics and Wilks phenomenon. Ann. Stat. 29:153–193. doi:10.1214/aos/996986505

Ge, Y., G. Bai, V. Stoerger, and J.C. Schnable. 2016. Temporal dynamics of maize plant growth, water use, and leaf water content using automated high throughput RGB and hyperspectral imaging. Comput. Electron. Agric. 127:625–632. doi:10.1016/j.compag.2016.07.028

Jacoby, W.G. 2000. Loess: A nonparametric, graphical tool for depicting relationships between variables. Elect. Stud. 19:577–613. doi:10.1016/S0261-3794(99)00028-1

Ji, H., and H.-G. Müller. 2017. Optimal designs for longitudinal and functional data. J.R. Stat. Soc., Ser. B 79:859–876. doi:10.1111/rssb.12192

Kwak, I.-Y., C.R. Moore, E.P. Spalding, and K.W. Broman. 2014. A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. Genetics 197:1409–1416. doi:10.1534/genetics.114.166306

Kwak, I.-Y., C.R. Moore, E.P. Spalding, and K.W. Broman. 2016. Mapping quantitative trait loci underlying function-valued traits using functional principal component analysis and multi-trait mapping. G3: Genes, Genomes, Genet. 6:79–86.

Liang, Z. and J.C. Schnable. 2016. RNA-seq based analysis of population structure within the maize inbred B73. PLoS One 11:e0157942. doi:10.1371/journal.pone.0157942

McCaw, M.E., J.G. Wallace, P.S. Albert, E.S. Buckler, and J.A. Birchler. 2016. Fast-flowering mini-maize: Seed to seed in 60 days. Genetics 204:35–42. doi:10.1534/genetics.116.191726

Miller, N.D., B.M. Parks, and E.P. Spalding. 2007. Computer-vision analysis of seedling responses to light and gravity. Plant J. 52:374–381.

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 195:1077–1086. doi:10.1534/genetics.113.153346

Muraya, M.M., J. Chu, Y. Zhao, A. Junker, C. Klukas, J.C. Reif, and T. Altmann. 2017. Genetic variation of growth dynamics in maize (*Zea mays* L.) revealed through automated non-invasive phenotyping. Plant J. 89:366–380. doi:10.1111/tpj.13390

Ramsay, J.O., and B.W. Silverman. 2005. Functional data analysis. 2nd ed. Springer, New York. doi:10.1002/0470013192.bsa239

White, J.W., and M.M. Conley. 2013. A flexible, low-cost cart for proximal sensing. Crop Sci. 53:1646–1649. doi:10.2135/cropsci2013.01.0054

Xavier, A., B. Hall, A.A. Hearst, K.A. Cherkauer, and K.M. Rainey. 2017. Genetic architecture of phenomic-enabled canopy coverage in *Glycine max*. Genetics 206:1081–1089. doi:10.1534/genetics.116.198713

Xu, Y., Y. Li, and D. Nettleton. 2017. Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. J. Am. Stat. Assoc. doi:10.1080/01621459.2017.1366907

Yao, F., H.-G. Müller, and J.-L. Wang. 2005. Functional data analysis for sparse longitudinal data. J. Am. Stat. Assoc. 100:577–590. doi:10.1198/016214504000001745

Zhang, X., C. Huang, D. Wu, F. Qiao, W. Li, L. Duan, et al. 2017. High-throughput phenotyping and QTL mapping reveals the genetic architecture of maize plant growth. Plant Physiol. 173:1554–1564. doi:10.1104/pp.16.01516