Do we need to mention non-parametric regression and sorghum height in the title?

Time-series GWAS using high-throughput phenotyping and functional principal components analysis in sorghum

Chenyong Miao[1], Alejandro D. Pages[2], Yuhang Xu[3], and James C. Schnable[1*]

[1]Center for Plant Science Innovation, Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, US, 68503
[2]Computer Science Department, University of Nebraska-Lincoln, Lincoln, NE, US, 68503
[3]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, US, 68503
[*]schnable@unl.edu

ABSTRACT

background
results
conclusions

Introduction

Genome sequencing and genotyping technologies have facilitate the success of genome-wide association studies (GWAS) in the past decade. A lot of significant genes have been identified by GWAS in both plant and human studies[?,?,?,?]. Compared to the highly developed sequencing and genotyping technologies, the phenotyping technology as another important GWAS component has also been developed rapidly in recent years[?,?,?,?]. Previous studies have shown that leveraging image-based phenotypes can facilitate genetic mapping and gene discoveries due to its noninvasive, less laborious and high-throughput advantages. For example, Yang et al.[?] conducted GWAS to dissect the genetic architecture of 15 rice agronomic traits with image-based measurements. Some new associated loci were identified in addition to some well-known genes, which suggested the potential power of integrating image-based phenotyping in GWAS. The key challenge of using image-based phenotypes is how to accurately extract biological meaningful trait information from images. Although some available software packages can help researchers perform plant trait extraction from normal RGB images without coding too much[?,?], customized scripts and advanced image processing techniques such as machine learning methods are required to get accurate measurements from complex image types such as hyperspectral images.

Traditional GWAS use phenotypes measured at a single time point which is usually at the mature stage for the whole population. However, the life cycle of a plant is a dynamic process affected by different genes at various stages. Previous genetic mapping studies have shown that some associated genes are only detectable at particular developmental stages[?,?,?,?,?]. Therefore, results from a single time point GWAS is hard to provide deep insights into the genetic resources across the different developmental stages[?]. To conduct a time-series GWAS project, plants in the mapping population need to be measured individually at various growth stages, which is hard to complete using the traditional phenotyping tools. Recent development of the high-throughput phenotyping facilities equipped with various imaging sensors make it possible to automatically screen each plant in a population throughout the whole growth period[?].

Once the time-series phenotypes are available, the common way to map the dynamic trait is to conduct GWAS on each measurement point and summarize the final results. However, mapping the dynamic traits directly is more efficient and biological meaningful. For example, mapping the growth curve can provide straight forward information of plant growth. There are several nonlinear models can be used to fit plant growth curves such as monomolecular and logistic functions[?]. Then mapping dynamic traits directly can be done by using values of the fitted parameters for the function to conduct GWAS. The method of fitting nonlinear models is only suitable for the simple dynamic traits with the growth curves similar to available functions. Recently, functional principal components analysis (FPCA) as a more general method had been applied to the analysis the more complex plant phenotypic data[?,?]. Compared to nonlinear function fitting method, FPCA is more flexible and can be widely applied in either simple and complex growth curves.

In the present study, we performed a time-series GWAS on the dynamic trait of plant height in the sorghum association panel (SAP)[?]. Using the Lemna Tec high throughput phenotyping facility enable us to screen each plant in the whole SAP

this paragraph
summary of ou
sults. Shall we
it in the abstra
Is it routine pu
the summary p
graph at the en
discussion part

including 320 individuals for two months covering the flowering stages for most of plants. In total, 12,800 photos including both RGB and hyperspectral images were generated from the camera sensors equipped in the facility. The accurate height for each plant in SAP was estimated at each time point after applying a neural network to distinguish different plant plants in hyperspectral images. Then the smooth plant height growth curve for each plant was obtained using a non-parametric regression method. In order to remove the plant growth disequilibrium among the population at the same date point, plant height curves were centered from the days after planting to the new coordinate based on days after panicle emergence. Both common sequential GWAS on plant height at each date point and the FPCA-based GWAS on the whole growth curve were conducted to dissect the genetic architecture of sorghum plant height.

Before conducting this time-series GWAS project in 2017 in the greenhouse, the plant height scores for the same population including 350 accessions were also measured in the field before harvesting the sorghum seeds in the Fall 2016. Using the height values measured in the field to perform GWAS, strong signals associated with dwarf genes were observed, which is consistent to the previously study using the same population[?]. However, when using a subset of SAP including only 295 accessions adopted in the time-series GWAS project, no dwarf genes were identified due to the less GWAS power of a smaller population size. The purged 55 accessions in the time-series GWAS project is due to the height limitations of the imaging chambers in the high throughput phenotyping facility. In contrast, by leveraging the high throughput phenotyping technology and accurate measurements from hyperspectral images, all cloned dwarf genes were identified using either the sequential or FPCA-based GWAS.

feel wired to n
tion this before
showing time-s
GWAS results

Results

Plant segmentations from RGB and hyperspectral images

First, the plant pixels in RGB images was segmented from background based on a green index method widely used in plant segmentations[?]. However, this method results in non-monotonically increasing growth curves in some accessions [Figure 1A] The reason is that the estimated height is measured at the leaf apex instead of the panicle + stalk height. In order to get the more informative plant height, i.e. the panicle + stalk height, another segmentation method was applied on the hyperspectral images. Compared to the RGB images with only 3 channels (Red, Green, and Blue), the hyperspectral images generated from our imaging facility contain 254 channels from 546nm to 1700nm partially covering both visible and near infrared spectrums for each plant on each run [Figure 2A]. Previous studies have shown that the reflectance patterns at different plant parts have different signatures in maize hyperspectral images[?]. Similarly, the different reflectance patterns of leaf, stalk, and panicle pixels in sorghum photos were also observed in this study [Figure 2B]. The leaf and stalk parts have very similar reflectance patterns compare to panicle part at the visible spectrum, which makes them hard to be separated in normal RGB images. However, their patterns are clearly separated in the near infrared spectrum region. Although the stalk and panicle parts have almost the same trend from 750nm to 1700nm wavelengths, the relative reflectance values are clearly different. Based on the different signatures within the four pixel classes across wavelengths, a neural network was trained to distinguish them with the prediction accuracy reaches to 96% in the cross validation process. Using the trained neural network, leaf, stalk and panicle pixels are clearly separated as shown in Figure 2C.

how to refer th
blue line in the
figure referenc

The plant heights were then recalculated based on the new segmented images with only considering plant stalk and panicle parts [Figure 1B]. Finally, a smooth growth curve was obtained for each plant after applying the nonparametric regression method [Figure 1C]. As the new segmentation method removes the influences of leaves in terms of the calculation of plant height, a overall monotonically increasing growth curve and a faster growth rate were observed for each plant. In order to prove that the new measurements of the plant height are more informative, the broad sense heritability was estimated using the plants with at least two replicates. Overall, the broad sense heritability of the heights estimated from the hyperspectral images performs better than the broad sense heritability of leaf height estiamted from normal RGB images [Figure S1].

the nonparame
or a nonparam
ric? Can we re
the method par
here to guide r
ers to check de
for the regressi
method?

Mapping sorghum plant height associated genes using sequential GWAS

In the traditional plant GWAS projects, the traits are usually measured at the mature stage in the whole population, such as the plant height[?] and grain characters[?], which guarantees that most of the plants are at the same stage when conducting the measurements. However, in the time-series GWAS, the dynamic trait will be measured at different time points, which may cause plants are at different developmental stages when conducting the measurements. Such measurements contain the confounding factors of plant growth disequilibrium due to the diversity within the mapping population such as the flowering time. The confounding factors need to be controlled in the GWAS to reduce the false positive SNPs associated with the confounding factors rather than the real measured trait[?,?,?].

sequential or ti
series, which c
better to use?

The trait interested in this study is the plant height that changes over time. However, the inconsistent developmental stages exist among different lines in the SAP population as they were collected in various regions around the world[?]. For example, some plants grow to flowering stages 10 days earlier than other plants largely due to the genetic variance [Figure 3A]. Moreover, some replicates in the same line were also inconsistent on the flowering time largely due to the environmental variance. To

is the confound
factor flowerin
time? I didn't
any flowering
lated genes in
sequential GW
based on DAP
we say the con
founding facto
the inconsisten
velopmental st

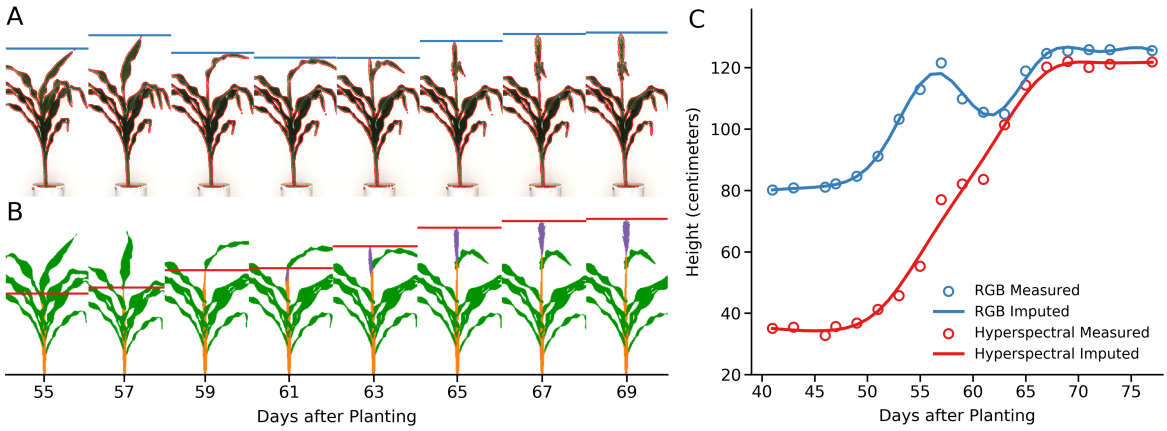

Figure 1. Plant growth curves with heights extracted from RGB and hyperspectral images. (A) RGB images with segmentation showing plant leaf height changes over time. (B) Hyperspectral images with segmentation of different plant parts showing plant stem + panicle height changes over time. (C)Plant leaf heights estimated from RGB images with missing values imputed by non-parametric regression (Blue). Plant stalk + panicle heights estimated from hyperspectral images with missing values imputed by non-parametric regression (Red).

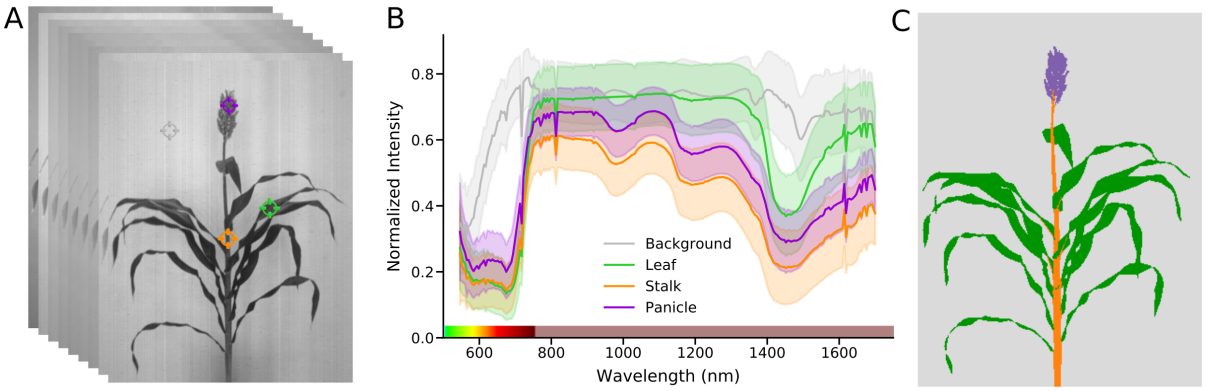

Figure 2. Sorghum leaf, stalk, and panicle segmentation from hyperspectral images. (A) A representation of a hyperspectral data cube with 254 individual wavelengths from 546nm to 1700nm. (B) Generalized reflectance patterns of leaf, stem, panicle and background pixels. Visible and near infrared spectrums are indicated in the color at the bottom. (C) Individual sorghum plant segmented into different organs after applying a trained neural network followed by further post-processing using OpenCV.

remove the confounding factor of the inconsistent developmental stages when focusing on the plant height, all the growth curves were realigned based on the days after the panicle emergence (DAPE) relative to the initial days after planting (DAP) [Figure 3B]. After the realignment, the overall pattern of growth curves in the population looks more consistent and comparable at each date point compared to the old way which is based on days after planting [Figure 3].

Then the sequential GWAS were conducted on each date point based on both DAP and DAPE to compare the effects of the realignment. The sorghum plant height is well known largely controlled by four dwarf genes *Dw*1, *Dw*2, *Dw*3, and *Dw*4 from classical genetic studies[?,?]. Except the *Dw*4, the other 3 dwarf genes had been cloned and the positions of these three genes had been confirmed in the reference genome, which provides the ground truth to validate our own sequential GWAS results[?,?,?,?]. After the realignment processing, the sequential GWAS detected all the three cloned dwarf genes at least in 6 date points [Figure 4]. Interestingly, the *Dw*3 only was identified before the panicle emergence, which is consistent to the previous study that *Dw*3 only controls the height below the flag leaf, which means the effect of *Dw*3 is hard to be detected due to the

I removed the CNN part iden fying flowering time is. We ca still talk about chine learning make predictio on flowering ti in the discussi part. In that ca shall we remov 'machine learn in our title?

date point or ti point? we run GWAS on each single day

influences of the flag leaf-to-apex interval increases after the emergence of panicle[?]. In contrast, time-series GWAS based on DAP only detected a consecutively strong signal on *Dw*3, which suggests there are more false positives without the realignment processing.

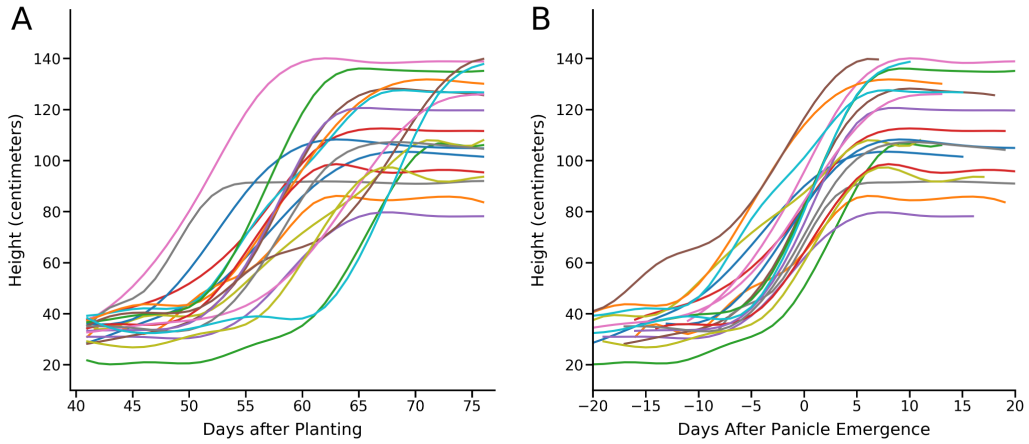

Figure 3. Comparison of plant growth curves based on DAP and DAPE (A) Growth curves of 20 samples in SAP based on Days after Planting (DAP). (B) Growth curves of the same 20 samples based on Days after Panicle Emergence (DAPE).

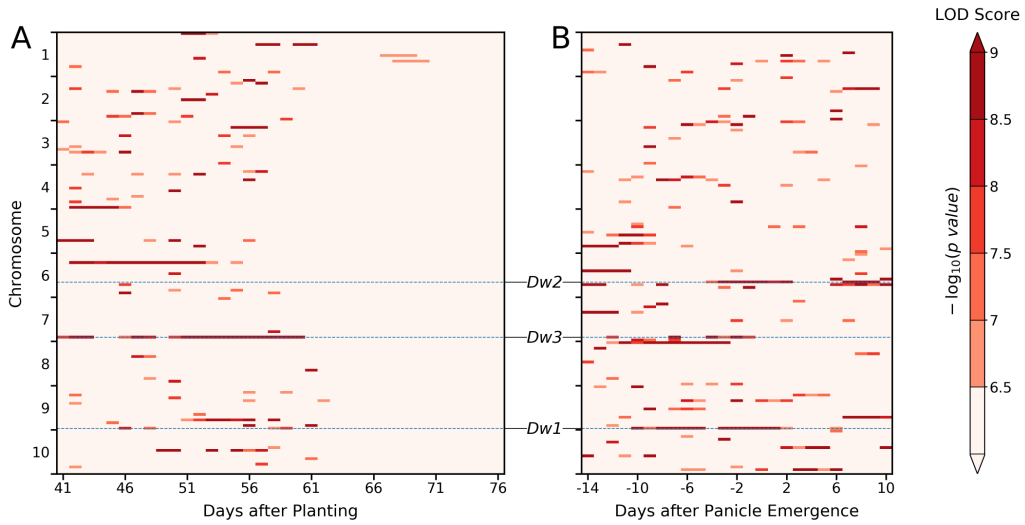

Figure 4. The heatmap of sequential GWAS based on two different time indexes. (A) Sequential GWAS from 42 to 75 days after planting (DAP). (B) Sequential GWAS from -20 to 10 days after panicle emergence (DAPE). The orange and read colors indicate the genomic regions containing SNPs significantly associated with the trait based on a significant threshold of *p*-value $10^{-6.53}$. Light pink indicates genomic regions without significant SNPs. The locations of the three cloned dwarf genes were indicated between the two heatmaps.

Mapping developmental trait of plant height using FPCA

Although the sequential GWAS based on DAPE from last part detected all the cloned dwarf genes, there are a lot of noisy signals in the sequential GWAS that cannot map to any plant height associated genes. In addition, sequential GWAS is high computational cost for current GWAS algorithms as researchers have to run GWAS on each available date point, especially

when facing large population with over 1 million SNPs. Therefore, we applied the FPCA to compress the whole growth curves to several principal components which can be further used as the phenotypes to run GWAS. The FPCA results show that the first two principal components explain over 96% variance of growth curves in the whole population, which suggest the growth patterns can be well characterized to the two main components [Figure 5]. The mean function derived from FPCA is very close to the empirical mean curve which is the average heights over each time point [Figure 5A]. The first principal component explaining 68.07% of total variance reflects the general low and high plants over time [Figure 5B]. The second principal component explaining 27.81% of total variance captures the height changes in two different stages, before and after the flowering time [Figure 5C]. The third and fourth principal components totally only explain about 4% of total variance, which indicates these two components can be treated as residuals.

Then MLM and FarmCPU GWAS were run on the four principal components as the phenotype to test if the dwarf genes are still detectable. Surprisingly, all three cloned dwarf genes were identified using FarmCPU GWAS model [Figure 6]. *Dw*1 and *Dw*2 were identified by FarmCPU model using the first principal component as the phenotype [Figure 6A]. The distance of the nearest significant SNP to *Dw*1 (Sobic.009G229800) is 19Kb. Although the nearest significant SNP is 2.5Mb away from *Dw*2 gene (Sobic.006G067700), it is the only significant SNP identified on chromosome 6 by FarmCPU model and it is also within the *Dw*2 hot region identified by GWAS MLM model. Furthermore, the hot region identified here is consistent to the previous GWAS study on sorghum plant height using the same GWAS method and the same population but with larger population size[?] [Figure 6A, Supplementary S2]. In addition to the identified dwarf genes using the first principal component, several height related orthologs also detected near the other significant SNPs. For example, an ortholog which is 48Kb away from the second significant SNPs in chromosome 2 encodes the adenylate kinase isoenzyme which is an essential stem growth factor in Arabidopsis[?] [Table S1]. Using the second principal component as the phenotype, FarmCPU model identified the *Dw*3 (Sobic.007G163800), 49Kb away from the only significant SNP in chromosome 7 [Figure 6B]. Some height related orthologs were also detected near the other significant SNPs. For example, an ortholog near the first significant SNP in chromosome 3 is one of the IAA transcription factors which are involved in important plant growth and development processes such as apical dominance, root formation, and shoot elongation[?,?,?] [Table S1]. In contrast, no dwarf gene identified using the rest two small principal components, which indicates the growth curve of sorghum plant height can be well represented by two main components [Figure S3].

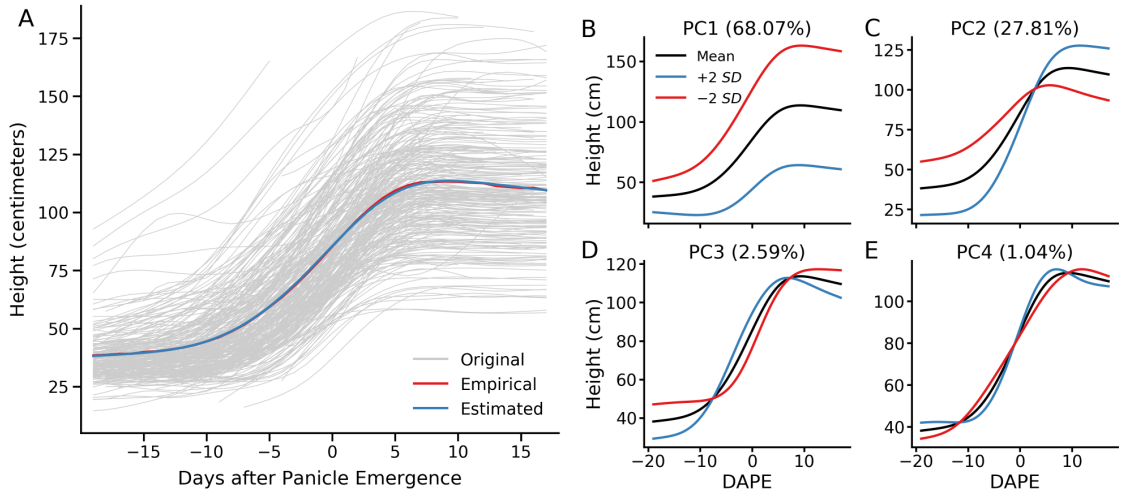

Figure 5. The estimated mean function and first four principal components derived from FPCA (A) The mean function of the growth curve estimated using FPCA (Blue) and the empirical mean curve which is the average heights over each time point (Red). (B) The predicted growth curves based on the first principal component plus or minus two SD from mean. (C) The predicted growth curves based on the second principal component plus or minus two SD from mean. (D) The predicted growth curves based on the third principal component plus or minus two SD from mean. (E) The predicted growth curves based on the fourth principal component plus or minus two SD from mean. Variance explained by each principal component is indicated in the parentheses of the title.

CSP,[?] ADP,[?]
PWWP,[?] GA At,[?,?] GA Maize,[?]

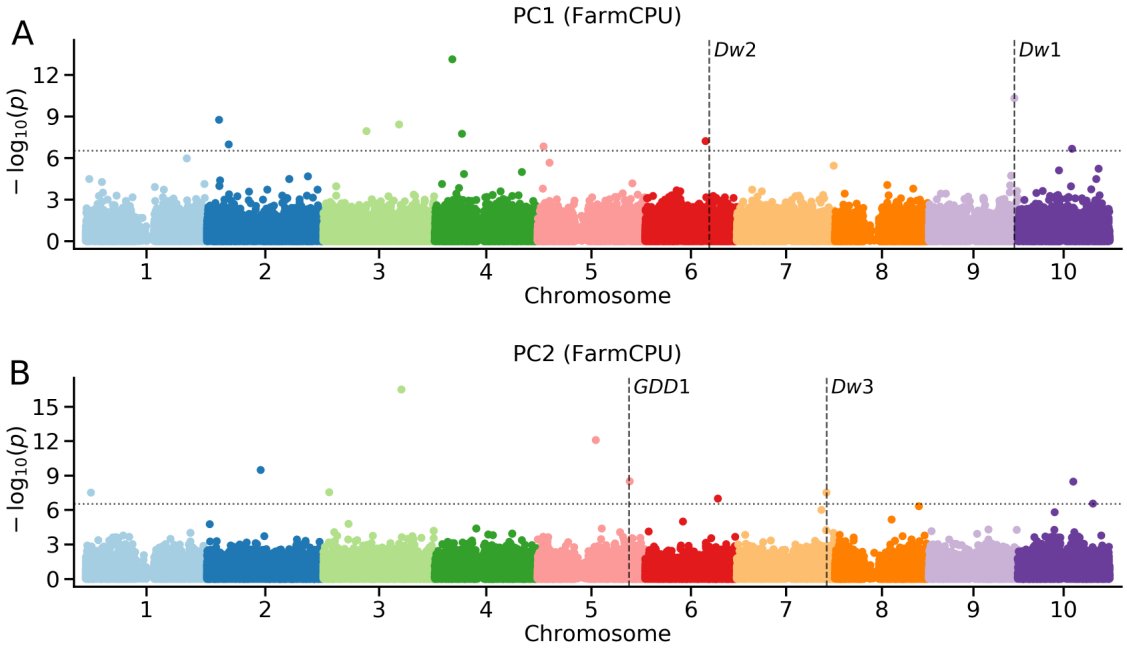

Figure 6. The Manhattan plots of the first two principal components estimated from FPCA using FarmCPU model. (A) GWAS results on the first principal component using FarmCPU model. (B) GWAS results on the second principal component using FarmCPU model. The locations of the cloned dwarf genes as well as the *GDD1* in sorghum are indicated by the vertical dash lines.

PGX,[?]
bHLH,[?]
ERAD,[?] NAC,[?]
dw3,br2[?]
dw1,[?],[?] arf at,[?] arf maize,[?] sumo,[?]

Discussion

As the *Dw*3 only controls the heights below the flag leaf, we assume it is corresponding to the first stage of the second principal component.

We are also curious if there is any significant SNP can be mapped to genes related to the upper part of the plant height, which is corresponding to the second stage of the principal component. Interestingly, there is gene (Sobic.005G144600) near the significant SNP at the end of chromosome 5, which is a homolog to the *GDD1* [Figure 6B]. *GDD1* is the Gibberellin dependent dwarf1 in rice[?].

ortholog or homolog here?

The mutant can greatly reduce the rice spikes by down-regulating the GA (gibberellic acid) biosynthesis. The same gene is also detected in a GWAS study on the sorghum panicle length[?].

Any plant with height higher than 3 meters was removed in SAP population due to the limit of our imaging chamber, which will potentially reduce the power to perform association study. In order to validate our assumption, we run GWAS on the whole population with the heights measured at the mature stage in the field in 2017. As a result, we detected three strong signals associated with Dw1, Dw2, and Dw3, which is consistent to the Morris study in 2008 using the same population [Morris paper]. Then we rerun the GWAS only use the samples adopted in this project and there is no significant SNPs identified. However, using time-series GWAS, as shown in the results, we still detected the important dwarf genes, which indicate time-series GWAS is a robust method in association study.

height, important, plant fitness and agricultural performance. increase harvest uniformity yield gains and height reductions of wheat (Tritticum spp.) and rice (Oriyza sativa) in the Green Revolution (khush 2001) its high heritability and the ease of its measurement.

which is also consistent to the previous study that *Dw3* only controls the lower part of plant height[?].

talk about the hot region on chr 6 including both ma1 and dw2

FarmCPU: high accuracy but low precision from our comparison.

The estimated heritability of sorghum height in this study is over 90% at different developmental stages, which is consistent to the previous reports that height in sorghum is a high heritability traits[?]. The heritability of maize height was estiamted to be > 90%[?]

The only significant SNP identified by FarmCPU on chr6 from the first component is 2.5Mb away from the Dw2 gene (gene name). The Dw2 is dwarf gene with the largest effect controlling the height in sorghum [ref 2008?] and it had been detected in several previous GWAS projects with a large hot region (4Mb) around Dw2 using mixed linear model, which is consistent to our mixed linear model results [supp x]. Compared to MLM, FarmCPU can control the effect of the previous identified SNPs to detect more significant SNPs with smaller effect and only report one significant SNP in the hot region, which can explain why only one significant SNPs identified in the hot region and not too close to the candidate gene.

time-series GWAS based on high throughput phenotyping can be widely performed in the future studies to dissect plant genetic architectures and facilitate the plant breeding programs.

In addition to the main dwarf genes, sequential GWAS based on DAPE also detected the *SP1* associated with the branch elongation[?] and a cell elongation related gene *TB1*[?].

Methods

Measurement of Plant Height in RGB Images

First, plant and other pixels in RGB images were separated by the green index 1.12, which is calculated from 2*Green /(Red + Blue). Then the segmentation is further cleaned by ... in opencv. Finally, the pixel counts were

Plant height calling from hyperspectral images

Fit plant growth curves using non-parametric smoothing

Let Y_{ij} be the jth observed phenotype of the ith plant/genotype, made at day t_{ij}, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,m_i$, where m_i is the total number of days observed for the ith plant. To model the plant growth, we propose to use the following non-parametric model

$$Y_{ij} = \mu(t_{ij}) + e(t_{ij}) \tag{1}$$

where $\mu(\cdot)$ is a mean function of the phenotype development and $e(t_{ij})$ is a zero-mean process associated with ith plant/genotype observed at t_{ij}. Let $B(t) = (B_1,\ldots,B_K)^T(t)$ be a vector of B-spline basis functions, where K is the number of basis functions. The estimated mean function can be expressed as $\hat{\mu}(t) = B^T(t)c = \sum_{k=1}^{K} B_k(t)c_k$, where c is a vector of coefficients of length K obtained using penalized least squares approach[?,?].

CNN to estimate the date of panicle emergence

The flowering date for each plant was detected by a convolutional neural network. The model testing accuracy and training info were shown in figure S1 and table S1.

Run GWAS

FarmCPU was used in this study to run GWAS [farmcpu ref]. The first five principal components as the fixed effect and relatedness matrix as the random effect were fitted in the model to control the false positives due to the population structure in the population. The parameter "maxLoop" was set to 10 for the maximum number of iterations and "FaST-LMM" was chosen for the "method.bin" parameter. The significant p-value cutoff was set to Bonferroni 0.05 calculated by 0.05/(independent SNPs). The Independent SNPs in the entire population was estimated by Genetic Type I error calculator (GEC) [gec 2012 ref]. Finally, 2.98E-7 was used as the p-value threshold.

0.1 Find height associated orthologs in maize

slow LD decay in the SAP. It's hard to map single gene or near-single gene resolutions. mapping resolution for GWAS. we quantified the average extent of LD decay and localized patterns of LD for each chromosome. On average, LD decays to 50% of its initial value by 1 kb and to background levels ($r2 < 0.1$) within 150 kb.

The much faster LD decay in maize populations.

Functional principal components analysis

We reduce the dimension of the phenotype development to a few principal components scores using FPCA. In FPCA, the process $e(t_{ij})$ in (1) is decomposed into two parts:

$$e(t_{ij}) = \sum_{l=1}^{L} \xi_{i,l}\phi_l(t_{ij}) + \varepsilon_{ij}, \tag{2}$$

where $\xi_{i,l}$ are zero-mean principal components scores with variance λ_l, $\phi_l(t_{ij})$ are eigenfunctions corresponding to principal components scores, and ε_{ij} are zero-mean measurement errors with constant variance. In FPCA, eigenfunctions are orthonormal, namely $\int \phi_{l_1}(t)\phi_{l_2}(t)dt = 0$, for all $l_1 \neq l_2$ and $\int \phi_l^2(t)dt = 1$, so the characteristics of phenotype development for the ith genotype can be represented by its principal components scores $\xi_{i,l}, l = 1,2,\ldots,L$. The variance of the principal components scores, $\lambda_l, l = 1,2,\ldots,L$, are sorted in decreasing order, so the first few principal components scores usually capture the majority of variation in the phenotype data. We also use B-spline bases for the approximation of eigenfunctions. The variance λ_l and eigenfunctions $\phi_l(\cdot)$ are estimated by eigenvalue decomposition and the principal components scores $\xi_{i,l}, l = 1,2,\ldots,L$ are estimated using best linear unbiased prediction[?,?]. A data-driven method[?] is used to decide the number of eigenfunctions, L.

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Figures and tables can be referenced in LaTeX using the ref command, e.g. Figure **??** and Table **??**.

Supplementary materials

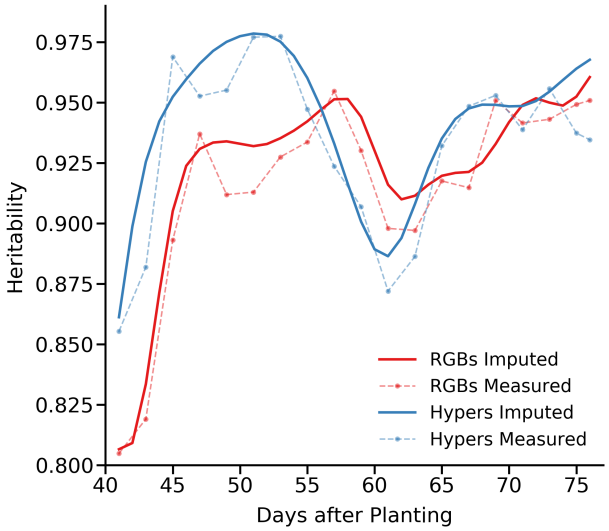

Figure S1. Heritability comparison between plant leaf height and stalk+panicle height. The estimated heritability of plant leaf heights changes over time (Red). The estimated heritability of plant stalk+panicle heights changes over time (Blue). Dot: heritability calculation based on raw heights. Solid line: heritability calculation based on imputed heights from non-parametric regression.

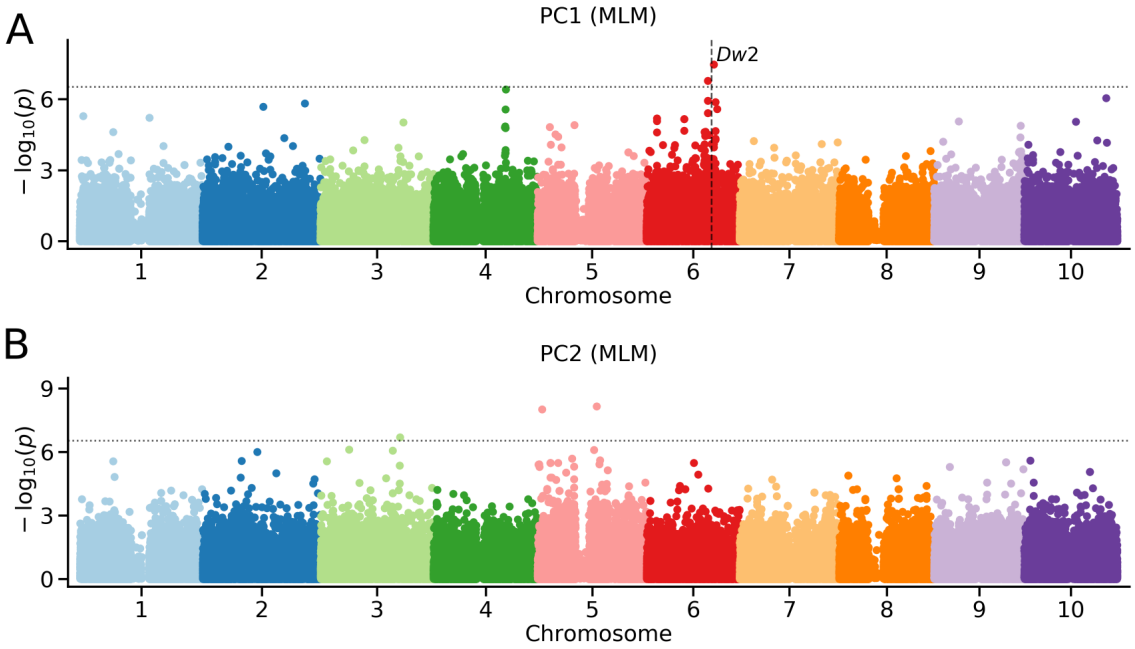

Figure S2. The Manhattan plots of the first two principal components estimated from FPCA using GWAS MLM model. (A) GWAS results on the first principal component using MLM model. (B) GWAS results on the second principal component using MLM model. The locations of the cloned dwarf genes in sorghum are indicated by the vertical dash lines.

Chromosome	Position	P value	MAF	Effect size	Principal Component	Candidate gene	Distance to SNP	Description	Reference
2	8,148,924	1.77E-09	0.33	-23.4	1st	Sobic.002G078400	1 Kb	ADP-ribosylation factor (ARF)	\cite{gebbie2005genes}
						Sobic.002G077000	112 Kb	CSP41A	\cite{beligni2008arabidopsis}
						Sobic.002G076700	155 Kb	isopentenyl transferase involved in the CTK biosynthesis	\cite{miyawaki2006roles, brugiere2008member, liu2011auxin}
2	14,415,404	1.08E-07	0.38	18.4	1st	Sobic.002G116200	48Kb	Adenylate kinase isoenzyme (an essential stem growth factor)	\cite{feng2012identification}
5	3,869,612	1.47E-07	0.4	-18.4	1st	Sobic.005G041000	117 Kb	NAC-transcription factor containing no apical meristem (NAM) protein domain	\cite{xie2000arabidopsis}
6	40,251,061	6.11E-08	0.03	-50.6	1st	Sobic.006G057000	31Kb	maize barren stalk2 and rice LAX2 orthologs	\cite{yao2019barren, tabuchi2011lax}
9	57,019,645	4.86E-11	0.24	-27.8	1st	Sobic.009G229800	19 Kb	Sorghum dwarf 1 involved in the brassinosteroid (BR) signaling	\cite{hilley2016identification, yamaguchi2016sorghum}
						Sobic.009G229200	28Kb	IAA transcription factor	\cite{xing2011genome, reed2001roles, song2009characterization}
						Sobic.009G230800	74Kb	gibberellin 2-oxidase	\cite{curtis2005modification, wang2018unveiling}
						Sobic.009G231800	169 Kb	auxin response factor	\cite{pekker2005auxin, shen2010functional, petsch2015novel}
1	3,344,877	3.09E-08	0.13	-13.8	2nd	Sobic.001G046700	106Kb	sterol methyltransferase	\cite{carland2010sterol}
						Sobic.001G042901	127Kb	isopentenyl transferase involved in the CTK biosynthesis	\cite{miyawaki2006roles, brugiere2008member, liu2011auxin}
3	3,979,698	2.81E-08	0.14	-13.7	2nd	Sobic.003G043000	1 Kb	Enhancer of polycomb-like transcription factor protein interacts with PWWP domain-containing proteins	\cite{tan2018peat}
						Sobic.003G044900	132Kb	IAA transcription factor	\cite{xing2011genome, reed2001roles, song2009characterization}
						Sobic.003G045900	233 Kb	gibberellin 3-oxidase	\cite{sakamoto2003genetic, chen2014maize, hu2008potential}
7	59,771,916	3.06E-08	0.17	-12	2nd	Sobic.007G163800	50 Kb	Auxin transporter	\cite{multani2003loss, blakeslee2007interactions}

Table S1. SNPs with significant association to height related orthologs

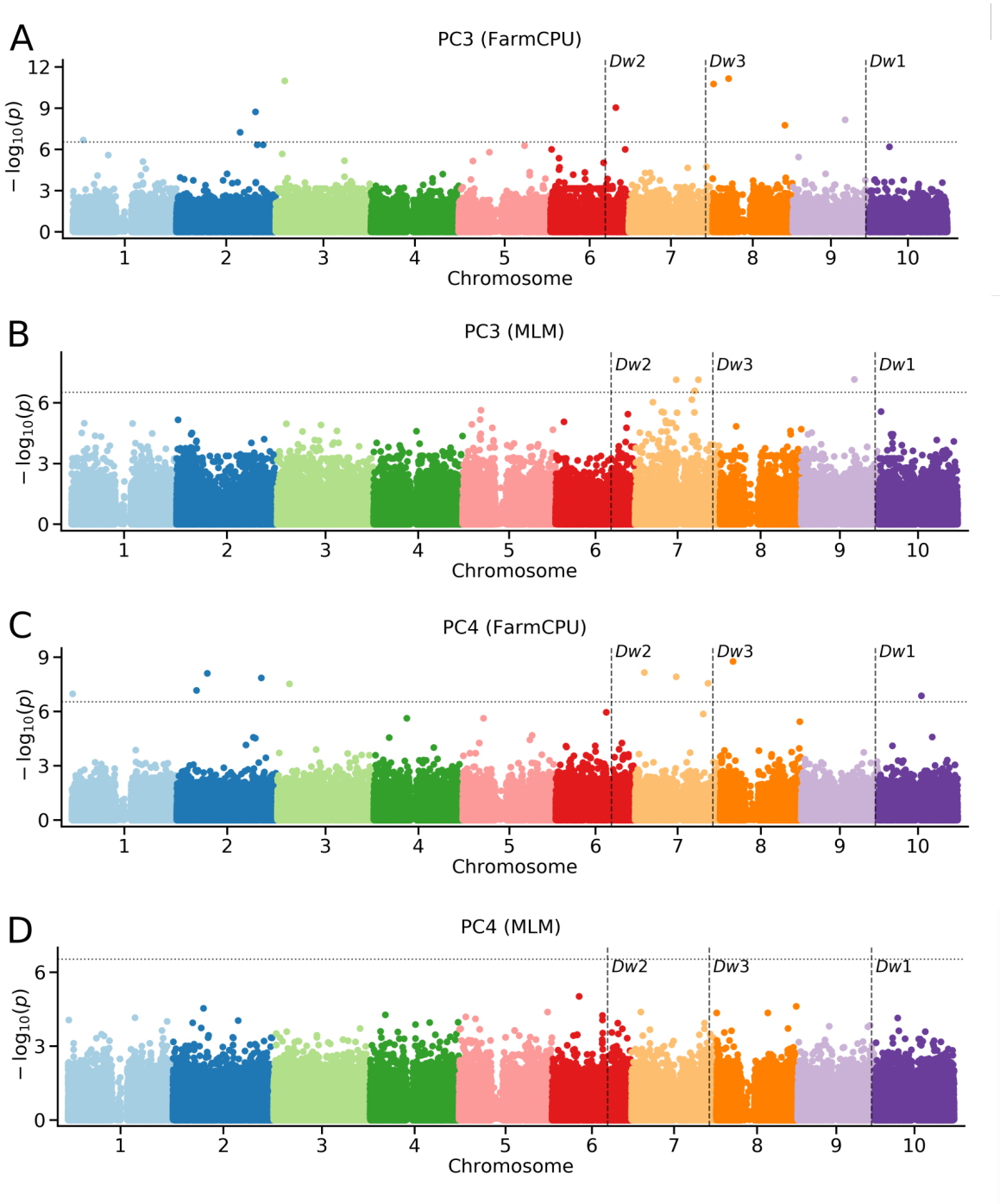

Figure S3. The Manhattan plots of the third and fourth principal component estimated from FPCA using GWAS FarmCPU and MLM models. (A) GWAS results on the third principal component using FarmCPU model. (B) GWAS results on the third principal component using MLM model. (C) GWAS results on the fourth principal component using FarmCPU model. (D) GWAS results on the fourth principal component using MLM model. The locations of the cloned dwarf genes in sorghum are indicated by the vertical dash lines.