HIGH-THROUGHPUT GENOTYPING ANALYSES AND IMAGE-BASED PHENOTYPING IN *Sorghum bicolor*

A Dissertation

by

RYAN FRANKLIN MCCORMICK

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	John Mullet
Committee Members,	Nancy Amato
	Wolfgang Bangerth
	Bruce Riley
Chair of Intercollegiate Faculty,	Dorothy Shippen

May 2017

Major Subject: Genetics

ABSTRACT

Sorghum bicolor is a valuable plant grown commercially for grain, forage, sugar, and lignocellulosic biomass production. Increasing yields for these applications without increasing inputs is necessary to sustainably meet future food and fuel demand. The generation of superior plant cultivars that produce more without increased input is facilitated by methods that can rapidly and accurately acquire plant genotypic and phenotypic data, and this dissertation describes the development and application of genomic and phenomic methods to improve crop productivity. The sensitivity and specificity with which genetic variants are called from sorghum genomic sequence data was improved by developing a variant calling workflow; this workflow interrelates different sources of genomic sequence data to inform the modern machine learning techniques implemented within the Broad Institute's Genome Analysis Toolkit (GATK). Genetic variants called in this manner have been used to dissect the genetic basis of agriculturally important traits and improve the sorghum reference genome assembly. Additionally, to increase the rate at which the morphology of plants can be evaluated, an image-based phenotyping platform was developed to acquire measurements of sorghum shoot architecture traits using a depth camera. Depth images of plants are used to generate 3D reconstructions, and these reconstructions are used to measure phenotypes, to identify the genetic bases of shoot architecture, and as input to plant and crop modeling applications. This research facilitates the rapid and accurate acquisition of the data necessary to increase the rate of crop improvement.

ACKNOWLEDGEMENTS

The work described in this dissertation was made possible by the contributions of multiple individuals, organizations, software development teams, and funding agencies.

I thank my Ph.D. committee chair, Professor John Mullet, and the members of the committee, Professors Nancy Amato, Wolfgang Bangerth, and Bruce Riley, for their guidance and support during this work. I thank the current and former members of the Mullet lab for their assistance and advice, particularly Sandra Truong, Sergio Hernandez, and Dr. Daryl Morishige. I thank the Texas A&M Institute for Genome Sciences and Society for maintaining the TIGSS (formerly the WSGI) HPC cluster used for many of the analyses reported herein. I thank the following funding sources: Department of Energy Great Lakes Bioenergy Research Center (BER DE-FC02-07ER64494), Department of Energy ARPA-E TERRA project (DE-AR0000596), and the Perry Adkisson Chair in Agricultural Biology. I also thank external collaborators at the Department of Energy Joint Genome Institute and the Hudson Alpha Institute for Biotechnology.

I thank the developers and communities surrounding the numerous open-source software projects used for their useful software, helpful advice, forum discussions, and for making their source code available. These projects and organizations include: Linux, the GNU Operating System, Canonical, the Standard C++ Foundation, Python software foundation, the R core team, PCL, OpenCV, FFTW, Meshlab, VTK, GATK, IGV, Pandas, Matplotlib, SciPy, R/qtl, ggplot2, OpenAlea, libfreenect2, the Blender Foundation, Marcin Kalicinski for RapidXML, Martin Hansen for Biopieces, Ole Tange for GNU Parallel, REPET, Sun Grid Engine, HTSeq,

GMOD and Chado, JBrowse, Foswiki, Code::Blocks, TexWorks, SequenceServer, the Document Foundation for LibreOffice, and others that may have been used but that were failed to be mentioned here. The author also thanks the Stack Exchange community for their constructive technical insights.

And finally, I thank the journal editors and anonymous reviewers that provided constructive feedback on the manuscripts that are reproduced herein, as well as the journals themselves for permitting reproduction within this dissertation, and bioRxiv for hosting preprints.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor John Mullet of the Department of Biochemistry and Biophysics, Professor Nancy Amato of the Department of Computer Science and Computer Engineering, Professor Wolfgang Bangerth of the Department of Mathematics, and Professor Bruce Riley of the Department of Biology.

Work for the dissertation was completed by Ryan McCormick in collaboration with Sandra Truong and John Mullet, co-authors on the manuscripts reproduced herein.

Funding Sources

This work was supported by the following funding sources: Department of Energy Great Lakes Bioenergy Research Center (BER DE-FC02-07ER64494), Department of Energy ARPA-E TERRA project (DE-AR0000596), and the Perry Adkisson Chair in Agricultural Biology.

NOMENCLATURE

BQSR	Base Quality Score Recalibration
CV(RMSD)	Coefficient of Variation of the RMSD
Dw	Dwarf
DAP	Days After Planting
DG	Digital Genotyping
GATK	Genome Analysis Toolkit
GBS	Genotyping By Sequencing
IF	Independent Family
indel	insertion/deletion
LOD	Logarithm of Odds
MD	Mean Difference
MLOD	Maximum Logarithm of Odds
PCL	Point Cloud Library
RAD-seq	Restriction-enzyme Associated DNA sequencing
RIG	Recalibration and Interrelation of sequence data with the GATK
RIL	Recombinant Inbred Line
RMSD	Root-Mean-Square Difference
SNP	Single Nucleotide Polymorphism
QTL	Quantitative Trait Locus
VTK	Visualization Toolkit
VQSLOD	Variant Quality Score Log of Odds
VQSR	Variant Quality Score Recalibration
WGS	Whole Genome Sequence

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Projected increases in global population size and economic affluency require global agricultural productivity to roughly double between 2016 and 2050 (Alexandratos *et al.*, 2012). This demand for increased production comes at a time when yield gains are slowing and the world seeks to mitigate the environmental damage caused by agricultural intensification (Foley *et al.*, 2005, 2011; Alexandratos *et al.*, 2012). Moreover, plant based bioenergy solutions (e.g., lignocellulosic biofuels) represent potential energy alternatives to fossil fuels. Many promising bioenergy crops are capable of high productivity with low inputs on marginal lands not optimal for food crops, and improved biomass production will continue to increase the viability of lignocellulosic biofuels (Somerville *et al.*, 2010). As such, simultaneously increasing crop productivity while minimizing crop inputs stands as a critical challenge for food security, energy security, and environmental health in the 21st century.

One strategy for improving crop productivity is the development of genetically superior cultivars with exceptional performance in target production environments. Favorable alleles can be enriched via selection or directly identified, and favorable combinations of alleles can be deployed into elite cultivars via traditional breeding or genome engineering (Cobb *et al.*, 2013; Park *et al.*, 2015). Modern implementations of this process (e.g., genomic selection) require knowledge regarding the state of the genome of many individuals combined with extensive phenotypic screening. This dissertation focuses on the topic of rapid acquisition and conversion of large amounts (i.e. terabytes) of genomic and phenotypic data to actionable information from which breeding decisions or genetic associations can be determined.

As the world's fifth most produced cereal crop (`http://www.fao.org/faostat`) and a promising bioenergy crop, sorghum is a multi-purpose plant useful for grain, forage, sugar, and lignocellulosic biomass production (Mullet *et al.*, 2014). Sorghum performs C4 photosynthesis, making it more efficient under hot and dry environments relative to C3 plants. Evolutionary relationships between sorghum and other important food and bioenergy grasses, including rice, maize, sugarcane, and *Miscanthus*, ensure that advances in sorghum are often translatable to other crops, and also that sorghum can benefit from the extensive research effort already invested in maize and rice (Paterson *et al.*, 2009). Sorghum is also readily amenable to genetic analyses due its diploid nature, ability to make controlled crosses, and a sequenced reference genome (800 Mbp) (Paterson *et al.*, 2009). Since sorghum is already used for commercial production, gains in sorghum productivity are readily translatable to economic impacts in production settings. All of these factors make sorghum a practical system for development and testing of tools for genetic improvement.

This dissertation describes approaches to increase the rate and accuracy with which genetic variation can be identified from genomic sequence data and image-based phenotyping approaches; these genetic and phenotypic data can be used to increase the rate of productivity gains in sorghum (Figure 1.1). Specifically, this dissertation reproduces manuscripts that introduce and document the (1) development of high-throughput genotyping analyses that assist the rapid identification of genetic loci of interest from genomic sequence data, (2) development of image-based phenotyping analyses to enable rapid measurement of sorghum characteristics, and (3) application of the developed methods to examine the genetic architecture of agriculturally important traits in sorghum. Continued progress in these research areas will contribute new methods to increase the rate of sorghum improvement and will facilitate the identification and deployment of traits useful for crop performance.

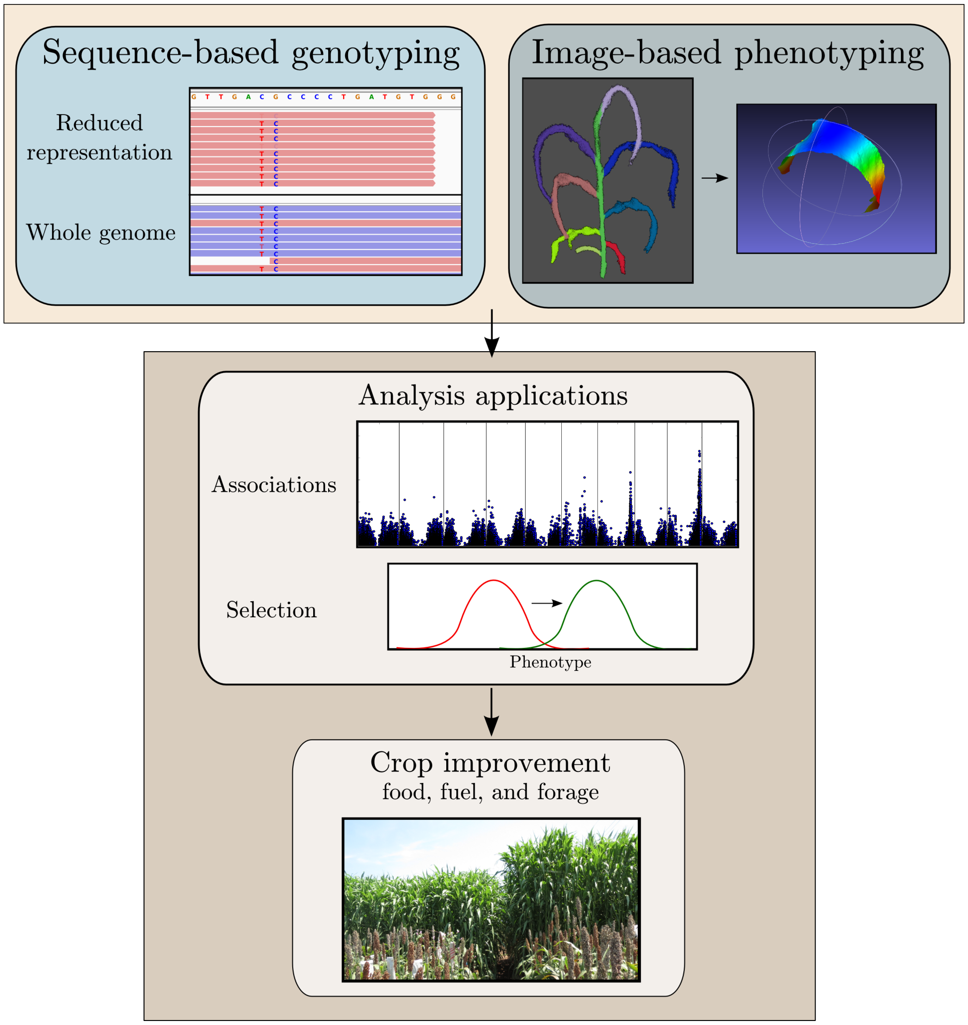

Figure 1.1: The development and application of novel genomics and phenomics approaches are necessary to increase the rate of crop improvement. Approaches that integrate multiple sources of genomic sequence information from different scales, including reduced representation, exome, and whole genome resequencing can leverage prior knowledge to rapidly and accurately identify genetic variation. These can be combined with image-based phenotyping platforms capable of acquiring longitudinal data on plant growth to dissect the genetic factors underlying plant performance as well as to drive accurate selection decisions. Ultimately, these will increase the rate at which improved plant cultivars can be developed.

2. RIG: RECALIBRATION AND INTERRELATION OF GENOMIC SEQUENCE DATA WITH THE GATK [1]

2.1 Overview

Recent advances in variant calling made available in the Genome Analysis Toolkit (GATK) enable the use of validated single-nucleotide polymorphisms and indels to improve variant calling. However, large collections of variants for this purpose often are unavailable to research communities. We introduce a workflow to generate reliable collections of single-nucleotide polymorphisms and indels by leveraging available genomic resources to inform variant calling using the GATK. The workflow is demonstrated for the crop plant *Sorghum bicolor* by (i) generating an initial set of variants using reduced representation sequence data from an experimental cross and association panels, (ii) using the initial variants to inform variant calling from whole-genome sequence data of resequenced individuals, and (iii) using variants identified from whole-genome sequence data for recalibration of the reduced representation sequence data. The reliability of variants called with the work flow is verified by comparison with genetically mappable variants from an independent sorghum experimental cross. Comparison with a recent sorghum resequencing study shows that the workflow identifies an additional 1.62 million high-confidence variants from the same sequence data. Finally, the workflows performance is validated using *Arabidopsis* sequence data, yielding variant call sets with 95% sensitivity and 99% positive predictive value. The Recalibration and Interrelation of genomic sequence data with the

[1] Reprinted from McCormick, R. F., S. K. Truong, and J. E. Mullet, 2015 RIG: recalibration and interrelation of genomic sequence data with the GATK. G3: Genes — Genomes — Genetics 5: 655-665 under the Creative Commons Attribution Unported License (`http://creativecommons.org/licenses/by/3.0`), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Copyright ©2015 McCormick *et al.*

GATK (RIG) workflow enables the GATK to accurately identify genetic variation in organisms lacking validated variant resources.

2.2 Introduction

The decreasing cost of high-throughput sequencing has led to a proliferation of template preparation methods and sequence data (Sims *et al.*, 2014). The abundance of sequence data has motivated an interest in leveraging available data to identify genetic variation, and software development has kept pace with this demand as exemplified by the Broad Institute's open-source Genome Analysis Toolkit (GATK). The GATK can integrate evidence for variants from multiple samples with joint genotyping, and it enables the use of validated single-nucleotide polymorphisms (SNPs) and indels to improve the accuracy of variant calling. Additionally, the GATK's methods are implemented in a manner amenable to reads originating from a variety of template preparation methods and sequencing platforms (DePristo *et al.*, 2011). However, many research communities lack the large, validated collections of SNPs and indels necessary for the GATK's Best Practices procedures because of the investment necessary to produce and curate such collections (Van der Auwera *et al.*, 2013). As an alternative to large-scale variant validation studies, we developed the Recalibration and Interrelation of genomic sequence data with the GATK (RIG) workflow to integrate information from multiple genomic sources and identify reliable sets of variants.

The GATK has gained extensive adoption in the human genomics research community due in part to the methods it uses to account for known error sources during variant calling; accounting for these error sources enables the GATK to consistently outperform other modern variant callers in benchmarking studies (DePristo *et al.*, 2011; Nekrutenko and Taylor, 2012; Liu *et al.*, 2013; Pirooznia *et al.*, 2014). Multiple sources of error exist, including incomplete or incorrect reference assemblies, erroneous realignment of reads to the reference genome (particularly in low complexity regions and around indels), inaccurate base quality scores, and suboptimal variant filtration parameters (DePristo *et al.*, 2011; Li, 2014). Features of the GATK address a number of these error sources, and we briefly describe three of the features most relevant to the design of the RIG workflow. The first feature is Base Quality Score Recalibration (BQSR), where the base quality scores assigned by the sequencer are corrected with scores empirically determined from the read group data using validated variants; these recalibrated scores more accurately reflect the true reliability of the base calls, thus correcting biases introduced by sequencing platforms (Li *et al.*, 2004). The second feature is the GATK's joint genotyping methodology that can integrate the evidence for a variant from many samples on reasonable time scales; this allows data from thousands of samples to be considered when evaluating the existence of a variant. The third feature is Variant Quality Score Recalibration (VQSR), where raw variant calls are assigned probabilities of being true variants based on the behavior of training variants in the raw variant calls using machine learning techniques (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). These probabilities allow users to decide which variants to use in downstream analyses based on desired levels of specificity and sensitivity, where high specificity indicates a low false-positive rate, and high sensitivity indicates a low false-negative rate. Two of these three features, BQSR and VQSR, require a collection of reliable variants to function effectively, and their benefits are inaccessible without such a resource.

Although many research communities lack large, validated resources of known SNPs and indels, some communities, namely agricultural research communities, often have access to a variety of genomic data sources that can be used to identify reliable genetic variants for use with the RIG workflow. Two characteristics influence the optimal use of these data sources with the RIG workflow: (i) the method used to produce the source's raw data from which variants are called, and (ii) the experimental design behind the source. First, many methods are available to produce the raw data from which variants are called, including reduced representation sequencing, whole-genome sequencing (WGS), SNP chips, Sanger sequencing, and RNA sequencing. Variants identified from two different methods can be considered more reliable than those identified in only one, as they are less likely to be artifacts introduced by a specific method. The RIG workflow can take advantage of multiple data sources by using variants found from one data source to inform the analysis of a second, read-based data source; by providing variants obtained from orthogonal methods, the reliability of variant resources used in BQSR and VQSR can be improved. Second, the experimental design behind the source also influences the reliability of the variants obtained from the source. Two experimental design elements influencing the reliability of a genomic variant are (i) if the variant segregates according to Mendelian expectations, and (ii) how often the variant is observed in independent samples. Genotyping large experimental crosses provides variants where Mendelian violations can be identified and the variants are observed at high frequencies in independent samples; as such, experimental crosses represent one of the most reliable sources of genetic variants. Association panels or population samples can also provide a reliable source of variants given a sufficiently large sample size and minor allele frequency. When Mendelian violations cannot be identified or when sample sizes are small (as is common with resequencing designs), variants are considered less reliable. For our use case with sorghum, we (i) generated an initial set of variants using reduced representation sequence data from an experimental cross and association panels, (ii) used the initial variants to inform variant calling from

WGS data of resequenced individuals, and (iii) used variants identified from WGS data for recalibration of the reduced representation sequence data. By considering the method used to produce the raw data from which variants are called and the experimental design behind the data source, available genomic sequence data can be optimally leveraged to improve variant calling and subsequent analyses.

Here we present the RIG workflow to formalize the process of incorporating available genomic sequence resources when calling SNPs and indels with the GATK. The RIG workflow is designed to leverage available genomic data in a manner that maximizes the information available to joint genotyping and to produce collections of reliable variants sufficiently large to perform BQSR and VQSR; this provides the benefits of the GATK's methods even in the absence of a large collection of validated variants, and it is readily applicable to organisms with a reference genome sequence and moderate sequence data resources. As an example, we describe the RIG workflow using *Sorghum bicolor* sequence data and show that it readily interrelates reduced representation and WGS data to generate variant calls. We evaluate the performance of the RIG workflow for sorghum sequence data using a collection of genetically validated variants, and we compare the output of the RIG workflow with variant calls from a recent sorghum study. Finally, we validate the workflow with *Arabidopsis* sequence data and show that high sensitivity and specificity is readily achieved.

2.3 Materials and methods

2.3.1 Sorghum analyses

The RIG workflow described in the Section 2.4 was designed as a generalization of our use cases in leveraging existing *Sorghum bicolor* genomic resources to take advantage of the GATK's strengths. Here we describe the process of transitioning from exclusive use of the naive pipeline to use of the initial informed and informed pipelines as an example of executing the RIG workflow and constructing variant resources (Figure 2.1, Figure 2.2, Figure 2.3, Figure 2.4).

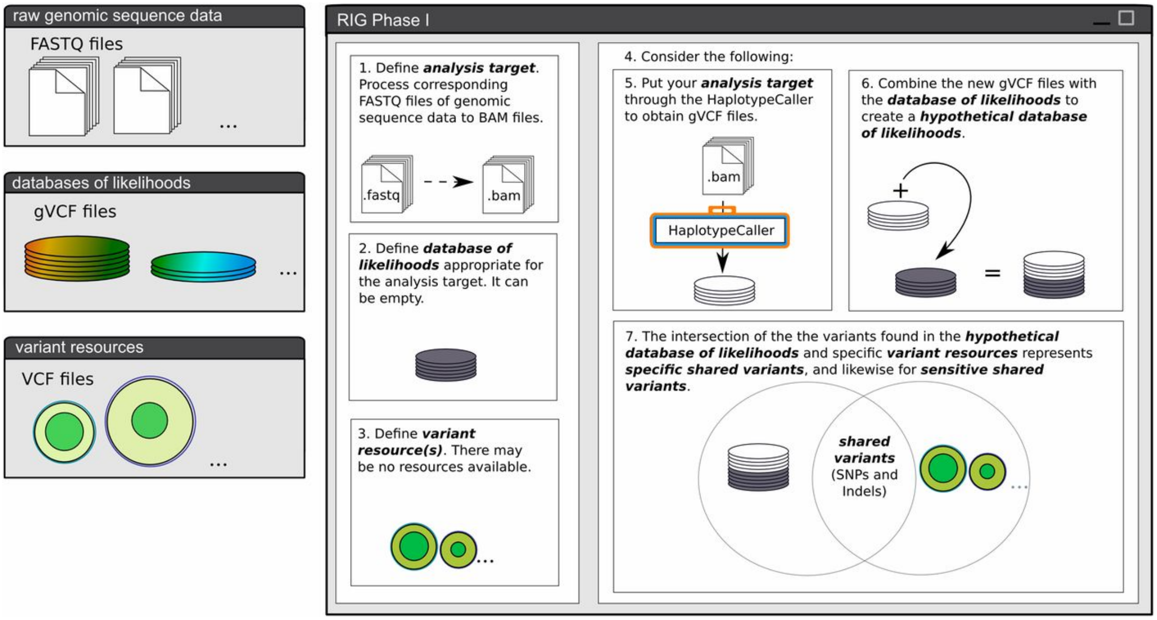
raw genomic sequence data
FASTQ files
...
databases of likelihoods
gVCF files
...
variant resources
VCF files
...
RIG Phase I
1. Define ***analysis target***. Process corresponding FASTQ files of genomic sequence data to BAM files.
.fastq
.bam
2. Define ***database of likelihoods*** appropriate for the analysis target. It can be empty.
3. Define ***variant resource(s)***. There may be no resources available.
...
4. Consider the following:
5. Put your ***analysis target*** through the HaplotypeCaller to obtain gVCF files.
.bam
HaplotypeCaller
6. Combine the new gVCF files with the ***database of likelihoods*** to create a ***hypothetical database of likelihoods***.
+
=
7. The intersection of the the variants found in the ***hypothetical database of likelihoods*** and specific ***variant resources*** represents ***specific shared variants***, and likewise for ***sensitive shared variants***.
shared variants
(SNPs and Indels)
...

Figure 2.1: Phase I of the RIG workflow. Phase I of the RIG workflow defines the five entities necessary for the execution of Phase II. Once the first three entities, the analysis target, database of likelihoods, and variant resource(s) are defined, the user considers a hypothetical case based on those first three entities to estimate the contents of the remaining two: the hypothetical database of likelihoods and the shared variants. If a user is unable to make a prediction regarding the latter two entities, the entities can either be treated as empty sets, or the user can use the GATK to carry out the necessary procedures to generate an estimate. Once all five entities are defined, the user can proceed to Phase II. RIG, Recalibration and Interrelation of genomic sequence data with the GATK; GATK, Genome Analysis Toolkit.

At the time of transitioning from the naive pipeline to the initial informed and informed pipelines with *Sorghum bicolor* sequence data, we had access to reduced representation sequence data generated internally by Digital Genotyping using the

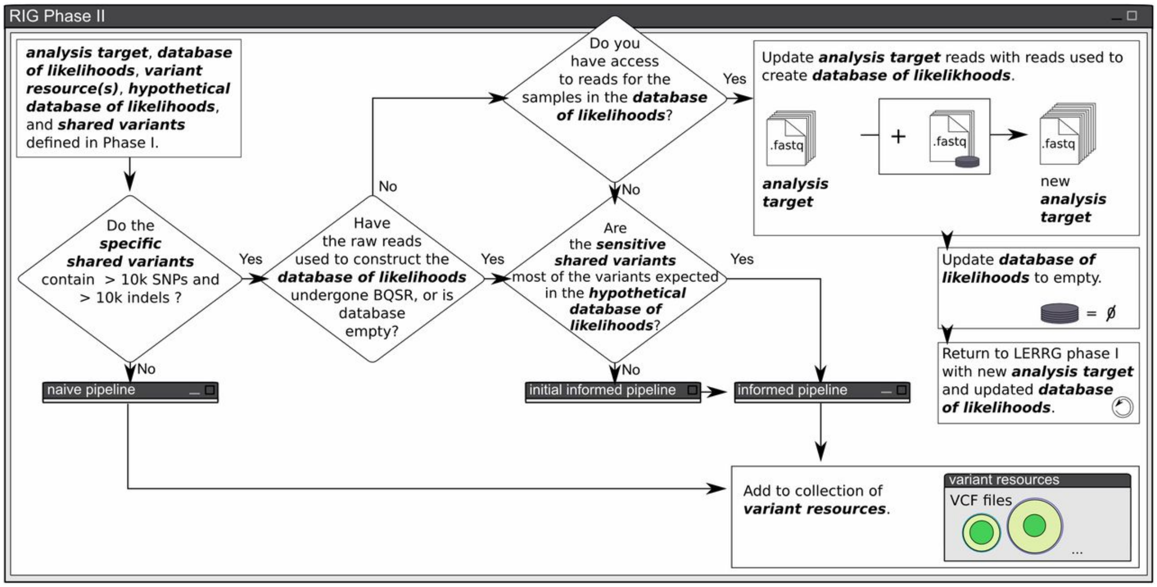
RIG Phase II
analysis target, database of likelihoods, variant resource(s), hypothetical database of likelihoods, and shared variants defined in Phase I.
Do the specific shared variants contain > 10k SNPs and > 10k indels ?
Yes
No
naive pipeline
Have the raw reads used to construct the database of likelihoods undergone BQSR, or is database empty?
Yes
No
Do you have access to reads for the samples in the database of likelihoods?
Yes
No
Are the sensitive shared variants most of the variants expected in the hypothetical database of likelihoods?
Yes
No
initial informed pipeline
informed pipeline
Update analysis target reads with reads used to create database of likelikhoods.
.fastq
analysis target
+
.fastq
.fastq
new analysis target
Update database of likelihoods to empty.
= ∅
Return to LERRG phase I with new analysis target and updated database of likelihoods.
Add to collection of variant resources.
variant resources
VCF files
...

Figure 2.2: Phase II of the RIG workflow. Phase II of the RIG workflow determines whether VQSR, BQSR, or both are appropriate, given the entities defined in Phase I. The workflow always proceeds through an analysis pipeline, characterized as the naive, the initial informed, and the informed pipelines shown in Figure 2.3. The end result of the workflow is the production of a variant resource that can be used in future analyses. RIG, Recalibration and Interrelation of genomic sequence data with the GATK; VQSR, Variant Quality Score Recalibration; BQSR, Base Quality Score Recalibration.

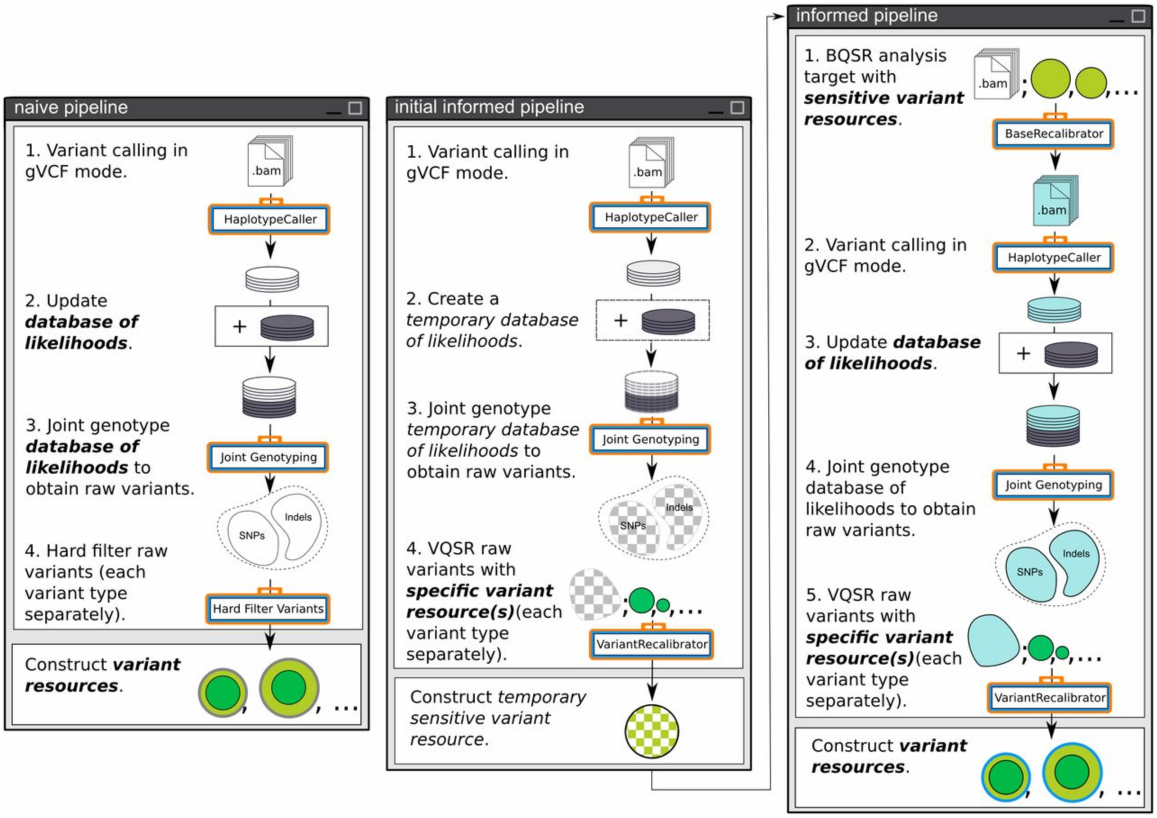
naive pipeline
1. Variant calling in gVCF mode.
.bam
HaplotypeCaller
2. Update ***database of likelihoods***.
+
3. Joint genotype ***database of likelihoods*** to obtain raw variants.
Joint Genotyping
SNPs
Indels
4. Hard filter raw variants (each variant type separately).
Hard Filter Variants
Construct ***variant resources***.
, , ...
initial informed pipeline
1. Variant calling in gVCF mode.
.bam
HaplotypeCaller
2. Create a *temporary database of likelihoods*.
+
3. Joint genotype *temporary database of likelihoods* to obtain raw variants.
Joint Genotyping
SNPs
Indels
4. VQSR raw variants with ***specific variant resource(s)***(each variant type separately).
; , , ...
VariantRecalibrator
Construct *temporary sensitive variant resource*.
informed pipeline
1. BQSR analysis target with ***sensitive variant resources***.
.bam
; , , ...
BaseRecalibrator
.bam
2. Variant calling in gVCF mode.
HaplotypeCaller
3. Update ***database of likelihoods***.
+
4. Joint genotype database of likelihoods to obtain raw variants.
Joint Genotyping
SNPs
Indels
5. VQSR raw variants with ***specific variant resource(s)***(each variant type separately).
; , , ...
VariantRecalibrator
Construct ***variant resources***.
, , ...

Figure 2.3: RIG pipelines. These are analysis pipelines that are traversed as part of Phase II of the RIG workflow. They correspond to cases where neither BQSR nor VQSR are appropriate (naive pipeline), where only VQSR is appropriate (initial informed pipeline), or where both BQSR and VQSR are appropriate (informed pipeline). When traversed, the informed pipeline emulates the GATK's Best Practices (Van der Auwera *et al.*, 2013). RIG, Recalibration and Interrelation of genomic sequence data with the GATK; BQSR, Base Quality Score Recalibration; VQSR, Variant Quality Score Recalibration; GATK, Genome Analysis Toolkit.

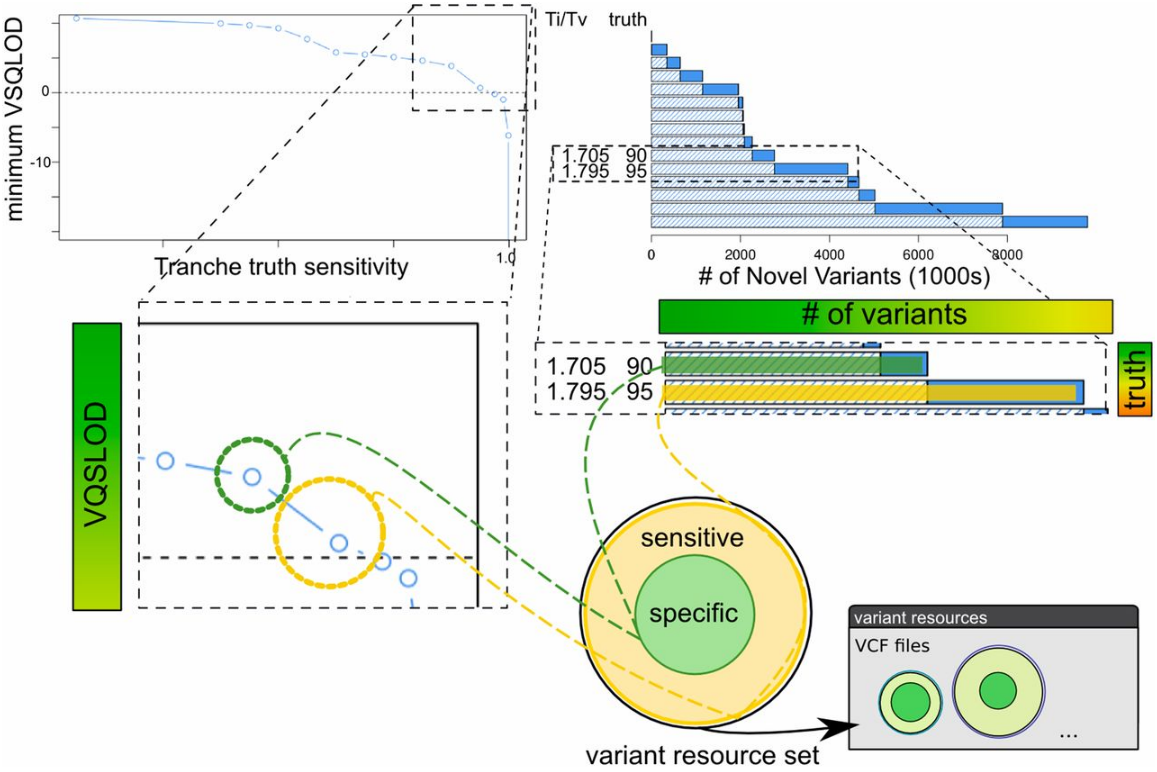

Figure 2.4: Construction of variant resources. After VQSR, multiple tranches are evaluated to choose specific and sensitive sets of variants for use in downstream analyses and to designate as variant resources. Tranches correspond to VQSLOD cutoffs above which a specified percentage of the variants designated as truth during VQSR are retained in the tranche. For example, a 95% tranche indicates the VQSLOD cutoff at which 95% of the variants designated as truth during VQSR would be retained. Accordingly, lower tranche percentages have greater specificity, lesser sensitivity, and contain fewer variants, and lower percentage tranches are subsets of greater percentage tranches. Here we show a 90% tranche being chosen as the specific variant resource and the 95% tranche being chose as the sensitive variant resource; both are subsequently added to the collection of variant resources. Note that the specific variant resource generated here is a subset of the sensitive variant resource. VQSR, Variant Quality Score Recalibration; VQSLOD, logarithm of odds ratio that a variant is real vs. not under the trained Gaussian mixture model.

restriction enzyme NgoMIV and a collection of WGS data generated from multiple groups (Zheng *et al.*, 2011; Evans *et al.*, 2013; Mace *et al.*, 2013; Morishige *et al.*, 2013). Using reduced representation sequence data for a 423 member recombinant inbred line population, we used the naive pipeline to produce variant calls (Burow *et al.*, 2011; Truong *et al.*, 2014). Preprocessing of the reads prior to variant calling, including read-mapping to version 1 of the *Sorghum bicolor* reference assembly, was performed using Picard (`http://broadinstitute.github.io/picard/`) and BWA (Paterson *et al.*, 2009; Li and Durbin, 2010). Variants were genetically mapped with R/qtl, and variants segregating as expected in these calls were used to create a Family Reference Variant Resource that contained 6849 SNPs and 2164 indels (Broman *et al.*, 2003). The Family Reference Variant Resource was considered a highly specific variant resource. Of note, the genetic positions of markers found in this manner are also being used to anchor unplaced super contigs in the *Sorghum bicolor* reference genome assembly (J. Schmutz, personal communication). Similarly, reduced representation sequence data for 733 sorghum germplasm samples processed with the naive pipeline were used to produce a Population Reference Variant Resource containing 62,022 SNPs and 20,801 indels. This variant resource was considered specific because we enforced a genotyping rate of 60% and a minor allele frequency of 0.05. The hard filtering parameters that we use in the naive pipeline for reduced representation sequence data can be found within the implementation on GitHub at `https://github.com/MulletLab/RIG`.

We sought to use these variants found in reduced representation sequence data to improve the analysis of WGS data of the 49 publicly available WGS data sources (Zheng *et al.*, 2011; Evans *et al.*, 2013; Mace *et al.*, 2013). To do this, we chose 10 individuals from the 49 that represented diverse sorghum germplasm accessions. We attempted to maximize diversity so that the sensitive variant resource constructed after the initial informed and informed pipelines had been executed would include many of the variants present in the next group of individuals for BQSR; this enabled use of the informed pipeline in the following iterations as the remaining 39 samples were processed (individuals were processed 10 at a time due to hard disk space limitations).

With the Family and Population Reference Variant Resources and the 10 WGS samples as analysis targets, we met the requirements for VQSR but not BQSR (Figure 2.1, Figure 2.2, Figure 2.3). As such, we followed the initial informed pipeline. For VQSR, the Family and Population Reference Variant Resources were both designated as truth, training, and known variants, and priors set to 15.0 and 7.0, respectively. Although these settings worked for our use case, they may not always be applicable; however, we typically follow these general rules: only highly specific variant resources should be designated as truth; variant resources designated as training do not need to be highly specific, but their priors should be set accordingly; all resources designated as truth should also be designated as training; and resources designated as truth and training can also be designated as known. Other details along with the annotations used for training the SNP and indel Gaussian mixture models can be found with the implementation on GitHub at `https://github.com/MulletLab/RIG`.

Generating variant resources following VQSR is a highly user-driven process that depends largely on the users confidence in the variant resources designated as truth for VQSR, and it requires examining multiple tranches resulting from VQSR (Figure 2.4). Tranches represent cutoffs based on variant resources designated as truth during VQSR, and they are generated by considering the VQSLOD scores (logarithm of odds ratio that a variant is real vs. not under the trained Gaussian mixture model) of truth variants that are present in the recalibrated raw variants. For example, if 90% of the truth variants found in the raw variants had a VQSLOD score over 1.5, then the 90% tranche would contain all variants in the raw variants that had a VQSLOD score over 1.5. We typically pick two tranches after VQSR, a specific tranche and a sensitive tranche, by examining the behavior of VQSLOD scores of multiple tranches (Figure 2.4). Specific tranches typically come from tranches where the VQSLOD score changes by small amounts even as the tranche percentage is decreased, and sensitive tranches are typically a non-negative VQSLOD score tranche that is more inclusive than the specific tranche.

Having generated a temporary sensitive variant resource from the initial 10 WGS samples using the initial informed pipeline, we proceeded down the informed pipeline with those 10 samples to generate a sensitive Whole-Genome Sequence Variant Resource. We then iteratively processed the remaining 39 samples in groups of 10 (9 on the final iteration) using the informed pipeline and updating the Whole-Genome Sequence Variant Resource each iteration; we continued to use only the Family and Population Reference Variant Resources for VQSR (to enforce that variants designated as truth for VQSR had been identified using a different sequencing template preparation method), and we used the newest sensitive Whole-Genome Sequence Variant Resource for BQSR. Upon completion of all 49 genomes, we used the newest Whole-Genome Sequence Variant Resources for BQSR and VQSR of the association panel data (sensitive for BQSR and specific for VQSR) to generate the Sensitive Population Reference Variant Resource (97.5% tranche) that was used for the genome-wide association study.

The Independent Family (IF) set used to examine the recalibration of WGS variants was constructed from a second biparental recombinant inbred line population (Xu *et al.*, 2000). Variants from this population were generated in the same fashion as the Family Reference Variant Resource (i.e., using NgoMIV Digital Genotyping, the naive pipeline, and checking for Mendelian segregation in R/qtl). The Raw, Sensitive, and Specific sets used in the comparison with the IF set were derived from the 100% tranche, the 95% tranche, and 75% tranche of the recalibrated WGS variants (Table A.1 and Table A.2, and Figure A.1). The Raw, Sensitive, and Specific sets used for comparison with the Gramene42-Mace2013 set originate in the same manner, but excluded indels, SNPs on super contigs, and variants not found in 1 of the 47 samples to be comparable with the Gramene42-Mace2013 set. Variants and genotypes for 171 individuals from the Sensitive Population Reference Variant Resource were used with downstream analysis tools to perform the association mapping described and depicted in Figure A.2 and Table A.3.

2.3.2 Arabidopsis analyses

Publicly available WGS for five accessions (ICE50 ICE134, ICE150, ICE213, and Leo-1) from Cao *et al.* (2011) were processed using the naive pipeline and stringently hard filtered (parameters available on GitHub). Publicly available Sanger sequence for 20 accessions (Ag-0, Bor-1, Br-0, Ei-2, Got-7, Gu-0, Hr-5, Kin-0, Kondara, Ms-0, Mz-0, NFA-8, Nok-3, PNA-17, Rmx-A02, Sorbo, Sq-8, Uod-1, Wa-1, Yo-0) were obtained from the Supporting Information of Nordborg *et al.* (2005). Publicly available WGS for the same 20 accessions from Schmitz *et al.* (2013) were processed through the initial informed pipeline, and VQSR was performed using the stringently filtered variants from Cao *et al.* (2011) as a training set (prior of 7.0) and as a truth set. The resulting 95% tranche was used for BQSR as the WGS data for the 20 accessions were then processed through the informed pipeline. The Cao *et al.* (2011) variants were again used for VQSR. All alignments and variant calling were done against the version 10 *Arabidopsis* reference genome (Arabidopsis Genome Initiative *et al.*, 2000).

To estimate error rates of the RIG workflow for WGS data, the resulting variant calls for the 20 accessions were compared to Sanger data from Nordborg *et al.* (2005) and variants from the Gramene database build 43, accessed January 2015 (Monaco *et al.*, 2014). This requires the assumption that the Sanger data were 100% specific (i.e., no false positives), and that the combination of the Sanger data and the Gramene build 43 variants were 100% sensitive (i.e., no false negatives). Although the WGS data strongly suggest that these assumptions are false, this still provides a useful baseline for comparison; however, we expect that the true sensitivity and specificity achieved in this comparison are greater than the values obtained since false positives in the Sanger data translate to decreased sensitivity and false negatives in the Sanger data and Gramene build 43 translate to decreased positive predictive value. Genomic intervals used to evaluate performance were defined as a subset of the 861 intervals from Nordborg *et al.* (2005). Because many of the Sanger reads had an abundance of Ns at the beginning and end of the read, 50 bp from the ends of each interval were removed. Excluding intervals that did not have >90% of the bases covered at greater than 15 depth in all 20 WGS samples and >90% coverage in all 20 Sanger samples yielded 419 intervals that covered 200,887 bp of the genome.

Two of the accessions (Got-0 and Ms-0) were dropped from the comparison due to extensive disagreement between the Sanger variants and the WGS variants, potentially due to not truly being the same accession. We also found a number of sites that were heterozygous in the WGS accessions that that had been manually curated by Nordborg *et al.* (2005) to Ns in the Sanger data. Because this generates what appears to be a false positive in the WGS data, we used the Sanger data to identify false negatives, and variants from both the Sanger data and Arabidopsis variants contained in Gramene build 43 to identify false positives. Variants from the Nordborg *et al.* (2005) Sanger data contained in the designated genomic intervals but not contained in a tranche of the WGS data were considered false negatives for the purpose of calculating sensitivity. Variants contained in the WGS data but not in either the Nordborg *et al.* (2005) or the variants present in Gramene build 43 were considered false positives for the purpose of calculating positive predictive value. Variant site counts used for calculating sensitivity and positive predictive value are available in Table A.4.

For the comparison, we report positive predictive value instead of specificity as a metric for false positives since the number of true negatives is far larger than the number of false positives, always leading to specificity values greater than 99.9%. As such, the performance of a tranche with a sensitivity of 95% and a positive predictive value of 99% is interpreted as a tranche where 95% of the true variants that existed were called and that 99% of the variants called are true variants.

2.3.3 Code and hardware

Our implementation of the workflow and pipelines are available on GitHub at `https://github.com/MulletLab/RIG` as a series of Bash scripts to serve as an example, to provide the annotations we used for hard filtering and VQSR, and to list all of the additional software version numbers used. GATKs Scala-based job submission controller, Queue, is suggested for implementing pipelines for the GATK for distributed computing resources; our implementation is in Bash because we experienced slowdowns in job submissions over time when using Queue (v3.1-1) on the Whole System Genomics Initiative cluster present at Texas A&M University.

2.4 Results

2.4.1 RIG: recalibration and interrelation of genomic sequence data with the GATK

The RIG workflow is a generalization of procedures to leverage existing genomic data when using the GATK v3.0+. Specifically, the workflow determines whether

VQSR and/or BQSR are appropriate to perform, and the workflow iteratively constructs reliable variant resources for future use with the GATK. The procedures of the RIG workflow are divided into two phases: Phase I, where the entities necessary for workflow execution are defined (Figure 2.1), and Phase II, where those entities are used to execute the workflow (Figure 2.2).

2.4.2 RIG Phase I: define RIG entities

Phase I of the RIG workflow defines the five entities necessary for execution of Phase II (Figure 2.1). The first entity is an analysis target. The analysis target contains the sequence reads from which the user intends to call variants. Reads of the analysis target should be preprocessed by read-mapping, duplicate marking (if applicable), and indel realignment; this entity is depicted as a stack of BAM-format files in Figure 2.1. The second entity is a database of likelihoods. The database of likelihoods contains the likelihood that a variant exists at a genomic position for all evaluated positions; this database consists of one or more gVCF-format flat files obtained from past GATK analyses of analysis targets produced by similar template preparation methods (i.e., a database of likelihoods for WGS samples should not be used with an analysis target of reduced representation samples). This entity is depicted as a stack of circles in Figure 2.1, and it can be defined as empty. The third entity is a set of variant resources. These are one or more files of VCF-format variant calls, and these calls should be conceptually (and physically, if necessary) partitioned into one or both of two categories: specific variant resources with low false positive rates, and sensitive variant resources with low false negative rates; a specific resource is necessary for VQSR and a sensitive resource is necessary for BQSR. As with the database of likelihoods, the variant resources can be empty and likely will be when first executing the workflow. The fourth and fifth entities can either be (i) constructed hypothetically based on a users expectations of the first three entities, or (ii) they can be empirically determined by performing the necessary analyses with the first three entities using the GATK. The fourth entity is a hypothetical database of likelihoods that is generated after adding the genotype likelihoods called from the analysis target to the existing database of likelihoods. The fifth entity is a set of shared variants. Shared variants are variants contained in both the hypothetical database of likelihoods and in the chosen variant resources; shared variants can be specific, sensitive, or both (or empty) depending on the classification of the variant resource they were found in. Once all five entities are defined, the analysis target, the database of likelihoods, the variant resources, the hypothetical database of likelihoods, and the shared variants, a user can proceed to Phase II.

2.4.3 RIG Phase II: execute analysis

The initial question of Phase II of the RIG workflow determines whether VQSR is appropriate based on the number of variants contained within the specific shared variants (RIG recommends at least 10,000 SNPs and 10,000 indels; Figure 2.2). A specific variant resource is required since false positives negatively impact the training of the Gaussian mixture models during VQSR, whereas false negatives have lesser effect. If the specific shared variants do not satisfy these criteria, then the RIG workflow enters the naive pipeline in which called variants are hard filtered using user-designated filtration criteria such as depth (Figure 2.3). Variants passing user-designated filtration criteria can then be added to the collection of variant resources. Once the naive pipeline has been used to analyze enough analysis targets, the collection of variant resources may be sufficiently large to answer yes to the initial question.

If the specific shared variants contain at least 10,000 SNPs and 10,000 indels, the next question addresses if the samples and variants in the database of likelihoods (if it is not empty) had previously undergone BQSR. If not, and if the reads corresponding to the samples used to generate the database of likelihoods are available, then the analysis target is updated with those reads, the database of likelihoods is set to empty, and the user returns to RIG Phase I (Figure 2.1) with the new analysis target and the empty database of likelihoods.

If the reads used in the construction of the database of likelihoods had previously undergone BQSR, or if the database of likelihoods is empty, then the final assessment determines whether the analysis target and the sensitive shared variants are appropriate for BQSR. A sensitive variant resource is necessary since false negatives cause BQSR to treat true variants in the analysis target as errors and will skew quality scores down, whereas false positives have a lesser chance to skew quality scores up. If BQSR is appropriate, the user follows the informed pipeline which emulates the GATKs Best Practices (Van der Auwera *et al.*, 2013). If BQSR is not appropriate, the user first uses the initial informed pipeline in which VQSR is performed on the raw variants to generate a temporary sensitive variant resource which is used during the execution of the informed pipeline that immediately follows the initial informed pipeline (Figure 2.3).

Construction of variant resources and adding them to the collection of variant resources is the end step of any path through the RIG workflow (Figure 2.4). Deciding the criteria for generating the variant resources at the end is a highly user-driven process that should consider the specific properties of the analysis target. For example, we generate highly specific variant resources from experimental crosses based on markers that segregate as expected (see the Section 2.3). Additionally, the variant annotations used for hard filtering and VQSR should differ based on how reads should behave in the analysis target; that is, reduced representation data should not use the same annotations as WGS data for hard filtering and VQSR because reads are not distributed around variants in a similar manner. To provide an example, we discuss the methods we use to select VQSR tranches and construct variant resources in Figure 2.4 and in the section Materials and Methods (Section 2.3. We also have made our code available on GitHub at `https://github.com/MulletLab/RIG` as an example and to provide the parameters and variant annotations we use.

2.4.4 Interrelation of genomic data enables a specificity and sensitivity framework for variant calls

In accordance with the RIG workflow, we used reduced representation data of an experimental cross and association panels to enable both BQSR and VQSR of WGS data of 49 resequenced individuals for the crop plant Sorghum bicolor. By interrelating data sources produced by different template generation methods with the RIG workflow, we enforced that the variants used to train the VQSR Gaussian mixture models that determine a variants VQSLOD score (logarithm of odds ratio that a variant is real vs. not real under the trained Gaussian mixture model) were found orthogonally, providing additional confidence that the variants used for training were real variants. Additionally, the differences in reliability of the training variants due to the different experimental designs were also considered for training of the VQSR models; variants from the experimental cross were assigned a higher prior likelihood of being correct than those from the association panels. By following the RIG workflow, each SNP and indel in the raw WGS variants was assigned a VQSLOD score that reflects its reliability. Figure 2.5 depicts the process of interrelating data for VQSR and the resulting VQSLOD scores of variant calls. While interrelating data from different template generation methods may be optimal, we also obtained good performance by following similar processing logic using only Arabidopsis WGS data.

In this way, the RIG workflow enables one of the greatest strengths of the GATK: the ability to put variant calls in a probabilistic framework that allows users to define where on the sensitivity and specificity spectrum the variants should sit for their target downstream application.

2.4.5 Evaluation of recalibrated variants from the RIG workflow

Although a formal evaluation of the accuracy of variant calling pipelines remains unfeasible for nonsimulated sequence data (Li, 2014), we estimated the performance of the workflow using both sorghum and Arabidopsis sequence data. For the sorghum data, we compared the variants called from sorghum WGS data via the RIG workflow to (i) a collection of reliable variants that were not used to train the VQSR models and (ii) a previously published sorghum variant calling analysis. We then used the sorghum WGS variants to recalibrate reduced representation data, and used the recalibrated variants for a genome-wide association study. Lastly, we further validated the performance of the RIG workflow using publicly available Sanger sequence and WGS data from Arabidopsis. Evaluation of recalibrated sorghum variants:

First, we examined the overlap between the recalibrated sorghum WGS variants and a collection of reliable variants that were not used to train the VQSR models. This collection of reliable variants, hereafter referred to as the Independent-Family (IF) set, originated from a biparental cross genotyped using a reduced representation method; the IF set was obtained in a similar manner to the Family Reference Variant Resource that was used for training during VQSR, and the IF set represented a set of highly specific, genetically mappable variants (see the section Materials and Methods). Of the 10,737 SNPs and 3740 indels in the IF set, 10,557 SNPs and 3632 indels had also been called from the 49 WGS samples (of which 2 samples represented the parents of the biparental cross). The IF variants present in the recalibrated

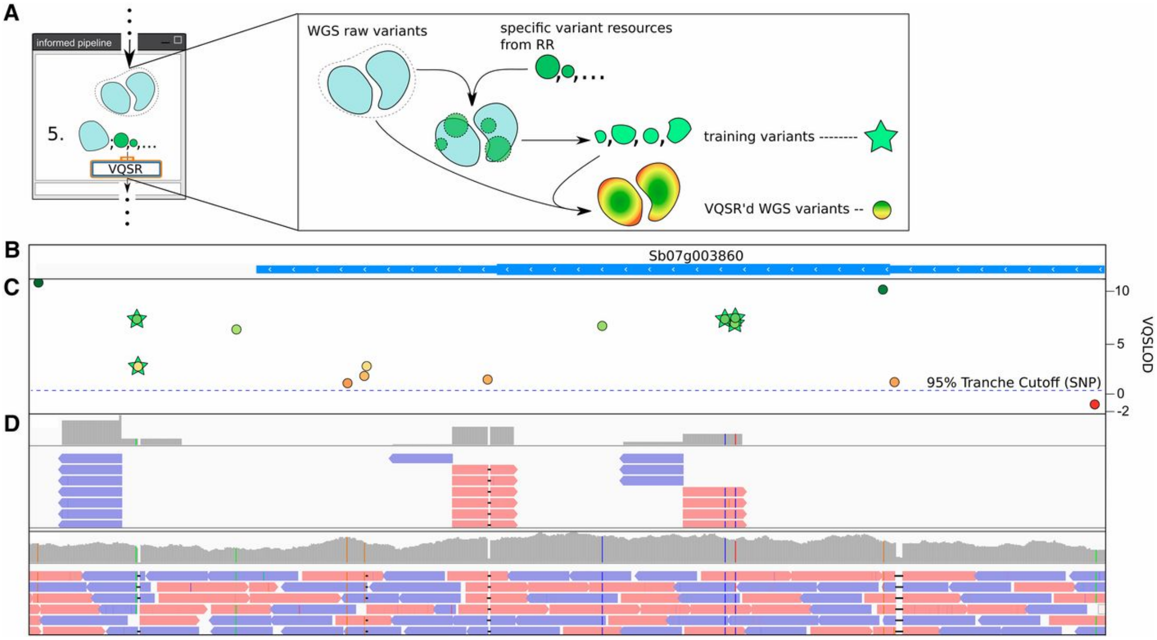

Figure 2.5: Interrelation of different genomic sequence data sources using the RIG workflow. (A) Schematic of how variants from reduced representation sequence (RR) data present in whole-genome sequence (WGS) data can be used to VQSR the WGS raw variants and assign VQSLOD scores to those variants. (BD) Visualization of the genomic region of Sb07g003860, a gene involved in sorghum midrib coloration (Bout and Vermerris, 2003). (B) the Sbi1.4 gene annotation; (C) shows the assigned VQSLOD scores for variants called in the region from WGS data; (D) shows the depth of coverage and mapped sequence reads for reduced representation and WGS data, respectively, for one sorghum line (BTx642). The RIG workflow enables variants called in the reduced representation sequence data to be used to inform and recalibrate the WGS analyses, and vice versa. This puts all of the variant calls into the GATKs probabilistic framework whereby variants can be filtered based on their reliability. Users interested in more sensitive or specific call sets can choose more inclusive or exclusive tranches, respectively, by changing the cutoff indicated by the blue dotted line in Panel C. The common and standardized file formats emitted by the GATK enable downstream interoperability between analysis and visualization tools, such as the Integrative Genomics Viewer that produced (B) and (D) (Thorvaldsdóttir *et al.*, 2013). RIG, Recalibration and Interrelation of genomic sequence data with the GATK; VQSR, Variant Quality Score Recalibration; VQSLOD, logarithm of odds ratio that a variant is real vs. not under the trained Gaussian mixture model; GATK, Genome Analysis Toolkit.

WGS variants had median VQSLOD scores of 8.22 and 5.29 for SNPs and indels, respectively, suggesting that the trained Gaussian mixture models correctly assigned true variants with highly positive VQSLOD scores (Figure A.1, Table A.1, and Table A.2). Furthermore, the proportion of IF set variants that were also contained in the 95% and 75% tranches correspond to their respective tranche cutoffs, indicating that the tranche cutoffs were functioning as expected. Since tranche cutoffs represent the VQSLOD score over which a certain proportion of variants from the designated VQSR truth set will be retained, we expected the proportion of IF variants present in each tranche to approximate the tranche cutoff. As expected, proportions of the IF set retained in each tranche were similar to the tranche cutoff. For example, the 95% SNP tranche retained 97% of the SNPs in the IF set, and the 95% indel tranche retained 94% of the indels in the IF set (Table A.2). These results indicate that the Gaussian mixture models for the WGS data were adequately trained and that the tranche cutoffs were functioning as expected.

Second, we compared the recalibrated sorghum WGS variants to a previously published sorghum variant calling analysis. The previous study from Mace *et al.* (2013) called SNPs and indels from 47 sorghum WGS samples; the SNP calls were recently made available as part of Gramene build 42 (accessed September 2014), hereafter referred to as the Gramene42-Mace2013 set (Monaco *et al.*, 2014). After excluding noncomparable variants from the calls produced by the RIG workflow (i.e., indels, SNPs on super contigs, and variants not found in the 47 samples), we obtained a Raw set comprised of 18,160,612 SNPs. We constructed an additional two sets from this Raw set for comparison: the Sensitive set, derived from the 95% tranche and comprised of 8,071,250 SNPs, and the Specific set, derived from the 75% tranche and comprised of 3,353,064 SNPs. Of the 6,450,628 SNPs in Gramene42-Mace2013 set, 5,002,099 were present in the Raw set. It is difficult to conclusively attribute the 1,448,529 SNP difference to any specific factors, and high discordance between different variant callers is not uncommon (O'Rawe *et al.*, 2013); we note that Mace *et al.* (2013) did not perform BQSR nor realignment around indels prior to calling SNPs, and they also used a different SNP calling algorithm. The overlapping 5,002,099 SNPs were used to compare the distribution of VQSLOD scores between the four sets (Figure 2.6). Because the VQSLOD score of all of the SNPs in the comparison were assigned under the same Gaussian mixture model and because the model was adequately trained as shown by the IF validation, comparisons of the relative sensitivity and specificity between the sets can be made. Given two sets of variants with similar VQSLOD distributions, the larger of the two sets contains more variants that are as likely to be true positives than the smaller set and is thus more sensitive. Furthermore, given two sets of variants where the VQSLOD distribution of one set contains a greater proportion of high VQSLOD score variants, the set with the greater proportion of high VQSLOD score variants contains variants that are more likely to be true positives and is thus more specific. As such, we find that the Raw set is the most sensitive but least specific; correspondingly, the Specific set is the most specific but least sensitive (Figure 2.6). The Sensitive set produced by the RIG workflow shows a dramatic improvement over the Gramene42-Mace2013 set in that it contains 1,620,622 more SNPs than the Gramene42-Mace2013 set while the median VQSLOD score remains similar with fewer negative VQSLOD scores, suggesting that the RIG workflow enabled greatly increased sensitivity without a corresponding loss in specificity.

As a final validation of the workflow with sorghum variants, we used a set of variants from reduced representation sequence data that had been recalibrated with WGS data to reproduce genome wide association results from the sorghum literature. There were 171 individuals contained within our reduced representation samples that

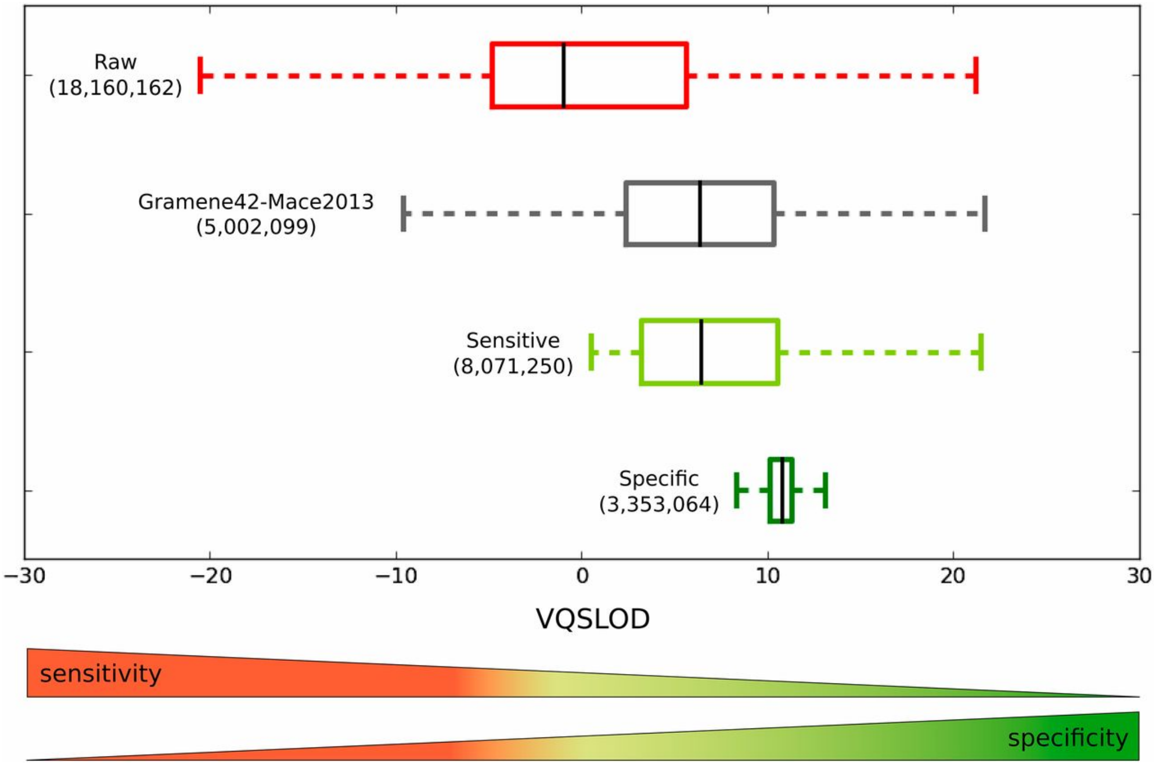

Figure 2.6: Comparison of VQSLOD score distributions for RIG-produced variant sets and a variant set from a previous study. VQSLOD (log of odds that a variant is real vs. not under the trained Gaussian mixture model) scores were calculated during VQSR of SNPs found in whole-genome sequence data using a Gaussian mixture model trained using SNPs originally found in reduced representation sequence data. For the 5,002,099 SNPs from Gramene42-Mace2013 that had been assigned VQSLOD scores in the Raw set produced by the RIG workflow, the median VQSLOD score is similar to the median of the 8,071,250 SNPs in the Sensitive set. The Sensitive set contains 1,620,622 more SNPs than the 6,450,628 SNPs in Gramene42-Mace2013, suggesting that the RIG-enabled VQSR allowed for a considerably more sensitive call set without a corresponding loss in specificity. VQSLOD, logarithm of odds ratio that a variant is real vs. not under the trained Gaussian mixture model; RIG, Recalibration and Interrelation of genomic sequence data with the GATK; VQSR, Variant Quality Score Recalibration; SNP, single-nucleotide polymorphism.

had also previously been phenotyped as part of a sorghum association panel (Brown *et al.*, 2008). After recalibrating the reduced representation data with the WGS data, we used the genotypes for these 171 individuals and phenotypes from Brown *et al.* (2008) to calculate genome wide associations (Figure A.2 and A.3) and reproduced known sorghum height QTL (Morris *et al.*, 2013; Higgins *et al.*, 2014). As such, the recalibrated reduced representation variants produced by the RIG workflow are useful for common downstream analyses, and these analyses are readily executable due to the GATKs use of standard file formats.

2.4.6 Evaluation of recalibrated Arabidopsis variants

Some organisms may not have sufficient data available from different template preparation methods to execute the RIG workflow as we did for sorghum. As such, we validated the performance of the RIG workflow using only WGS data as both the source of reliable variants and the analysis target. Efficacy of RIG was determined by comparing the variant calls produced by RIG from publicly available *Arabidopsis* WGS data against a collection of known variants from Sanger sequence data and variants present in the Gramene database (build 43; accessed January 2015) (Nordborg *et al.*, 2005; Cao *et al.*, 2011; Schmitz *et al.*, 2013; Monaco *et al.*, 2014).

The comparison used variant calls from 419 genomic intervals spanning 200,887 bp (containing at least 2,850 SNP and 375 indels) for 18 *Arabidopsis* accessions. Variants from the Sanger sequence data not present in the RIG variants were considered false negatives, and RIG variants not present in either the Sanger data or the Gramene build 43 set were considered false positives; these values were used to estimate sensitivity and positive predictive value of multiple tranches produced with RIG from *Arabidopsis* WGS data (Table 2.1 and Table A.4). This yielded a conservative estimate of RIG variant calls whereby 95% sensitivity and 99% positive predictive

Tranche (%)	Sensitivity (%)	Positive Predictive Value (%)
100.0	99.9	93.7
99.9	99.3	95.4
99.0	94.9	99.2
97.5	92.0	99.3
95.0	89.3	99.4
75.0	54.3	99.6

Table 2.1: Performance of tranches from *Arabidopsis* WGS sequence data. Sensitivity and positive predictive value of multiple tranches of recalibrated variants from Arabidopsis WGS data were calculated using variants found in Sanger sequence data from Nordborg et al. (2005) for sensitivity; variants found in both the Sanger sequence data and in Gramene (build 43) were used to estimate positive predictive value (Table A.4). For simplicity, the tranche percentage corresponds to both the SNP and the indel tranche. We note that these values are not generally applicable to other RIG analyses and these should not be taken as representative of how tranches in other analyses will behave; tranches should be chosen based on the reliability of the variants designated as truth for VQSR. WGS, whole-genome sequencing; SNP, single-nucleotide polymorphism; RIG, Recalibration and Interrelation of genomic sequence data with the GATK; VQSLOD, logarithm of odds ratio that a variant is real vs. not under the trained Gaussian mixture model; VQSR, Variant Quality Score Recalibration.

value are achieved in one tranche with the RIG workflow (the 99.0% tranche in this case). As shown in the sorghum data, larger percentage tranches are more sensitive but less specific; smaller percentage tranches are less sensitive but more specific. The optimal choice of tranche will, again, depend on the downstream application for which the variant set will be used. We note that the sensitivity does not correspond with the tranche cutoffs as well as they did in the sorghum validation; this may be a result of the low sensitivity of the Sanger variants due to manual removal of variant calls by Nordborg *et al.* (2005) during data curation. Ultimately, this *Arabidopsis* validation in combination with the sorghum validation demonstrates that the RIG workflow can produce accurate call sets from a variety of genomic data sources.

2.5 Discussion

The GATK has been shown to outperform other variant calling methods in benchmarking studies, and the RIG workflow enables the analysis benefits afforded by the GATK to research communities lacking validated variant resources (Liu *et al.*, 2013; Pirooznia *et al.*, 2014). RIG also provides access to features absent in current reduced representation sequence data analysis platforms. Two popular reduced representation sequence data analysis solutions, TASSEL and Stacks, are highly specialized for their respective data sources (GBS and RAD-seq, respectively), and they perform well in their target domains; however, they lack features that readily allow the interrelation of WGS with reduced representation sequence data, as well as the ability to accurately call indels (Catchen *et al.*, 2011; Glaubitz *et al.*, 2014). RIG provides a means to access both of these features, as well as benefit from accuracy gains from BQSR, joint genotyping, and VQSR. For organisms with a reference genome, the RIG workflow stands as a useful analysis alternative applicable to both reduced representation and WGS data, and RIG is also readily applicable to exome and RNA-seq data due to the GATKs flexibility. However, because the GATK, and by extension, RIG, cannot operate without a reference genome, software like TASSEL and Stacks will continue to fill important analysis roles, although this may change if software like dDocent, which allows users to take advantage of some of the GATKs benefits even in the absence of a complete reference genome, gain adoption (Puritz *et al.*, 2014). Ultimately, RIG was developed in the context of a genetics lab seeking accurate variant calls from multiple sequence data sources for agriculturally important organisms with a reference genome, and we expect it will be beneficial to those with similar use cases.

The RIG workflow requires that the shared variants are comprised of 10,000 SNPs and 10,000 indels for VQSR; however, the GATK developers have successfully used considerably fewer to good effect (DePristo *et al.*, 2011). We chose 10,000 for both SNPs and indels as the requirement because we have obtained useful results using these values; the values are not a hard rule. As such, the user can construct their own values by evaluating the VQSR and BQSR reports produced by the GATK to determine whether (i) the Gaussian mixture models were adequately trained to distinguish between variants of differing reliability, and (ii) whether the empirically determined base quality score recalibrations appear reasonable for the sequencing platform.

In cases in which sequence data from different template preparation methods are not available, it will not be possible to identify shared variants from orthogonal approaches as we did with sorghum sequence data. We ensured the variants designated as truth for VQSR originated from an analysis target produced by a different method (e.g., variants found in reduced representation data were used for VQSR of a WGS analysis target). This enforced that variants used in VQSR were found in two independent template preparation methods to approximate variants found using orthogonal methods. Since such genomic resources may not always be available, we also evaluated performance of a use case where only WGS data were available, and we showed that high levels of sensitivity and positive predictive value can be achieved using only WGS data. In cases in which the analysis target is the only source of variants, we and other GATK users have had some success by taking the analysis target through the naive pipeline, hard filtering to generate a temporary sensitive variant resource, and using that temporary sensitive variant resource to BQSR the analysis target. This procedure is iteratively repeated until the BQSR results from the current and preceding iteration converge, and then a specific variant resource is generated by stringently hard filtering to use as a bootstrapped variant resource in

VQSR. This ultimately skews VQSR based on the annotations used to hard filter the variants during bootstrapping, but communities lacking sufficient data sources may find this procedure to be an acceptable alternative.

The RIG workflow enables research communities to use the GATK (i) to interrelate different sequencing template preparation methods such as reduced representation and WGS into common, standardized file formats; (ii) accurately call genetic variants from genomic sequence data; and (iii) to iteratively refine variant resources. The RIG workflow will contribute to progress in construction of more complete catalogs of genetic variation, and the ability to readily interrelate variants from different sequence data sources using the GATK will increase the rate at which variants associated with a phenotype lead to the identification of the genetic variation that causes the phenotype.

3. 3D SORGHUM RECONSTRUCTIONS FROM DEPTH IMAGES IDENTIFY QTL REGULATING SHOOT ARCHITECTURE [1]

3.1 Overview

Dissecting the genetic basis of complex traits is aided by frequent and nondestructive measurements. Advances in range imaging technologies enable the rapid acquisition of three-dimensional (3D) data from an imaged scene. A depth camera was used to acquire images of sorghum (*Sorghum bicolor*), an important grain, forage, and bioenergy crop, at multiple developmental time points from a greenhouse-grown recombinant inbred line population. A semiautomated software pipeline was developed and used to generate segmented, 3D plant reconstructions from the images. Automated measurements made from 3D plant reconstructions identified quantitative trait loci for standard measures of shoot architecture, such as shoot height, leaf angle, and leaf length, and for novel composite traits, such as shoot compactness. The phenotypic variability associated with some of the quantitative trait loci displayed differences in temporal prevalence; for example, alleles closely linked with the sorghum *Dwarf3* gene, an auxin transporter and pleiotropic regulator of both leaf inclination angle and shoot height, influence leaf angle prior to an effect on shoot height. Furthermore, variability in composite phenotypes that measure overall shoot architecture, such as shoot compactness, is regulated by loci underlying component phenotypes like leaf angle. As such, depth imaging is an economical and rapid method to acquire shoot architecture phenotypes in agriculturally important plants like sorghum to study the genetic basis of complex traits.

[1] Reprinted with permission from McCormick, R. F., S. K. Truong, and J. E. Mullet, 2016, 3D sorghum reconstructions from depth images identify QTL regulating shoot architecture. Plant Physiology 172: 823-834 (www.plantphysiol.org). Copyright ©2016 American Society of Plant Biologists.

3.2 Introduction

The rate-limiting step for crop improvement and for dissecting the genetic bases of agriculturally important traits has shifted from genotyping to phenotyping, creating what is referred to as the phenotyping bottleneck (Houle *et al.*, 2010; Furbank and Tester, 2011). Alleviating the phenotyping bottleneck for agriculturally important plants will help the world meet the increasing food and energy demands of the growing global population (Somerville *et al.*, 2010; Alexandratos *et al.*, 2012; Cobb *et al.*, 2013). Approaches to alleviate the plant phenotyping bottleneck fall into two broad categories: approaches that increase the number of individuals that can be grown and evaluated (Fahlgren *et al.*, 2015b) and approaches that predict performance *in silico* to prioritize individuals to grow and evaluate (Hammer *et al.*, 2010; Technow *et al.*, 2015). Both of these approaches will be instrumental for increasing the rate of crop improvement, and both approaches are facilitated by advances in image-based phenotyping; multiple plant measurements can be acquired rapidly from images, and data from image-based phenotyping approaches also can inform performance prediction (Spalding and Miller, 2013; Pound *et al.*, 2014). As such, the development of image-based phenotyping platforms for agriculturally important plant species is a high priority for plant biology and crop improvement (Minervini *et al.*, 2015).

The diversity of crop species and the variety of traits of interest have resulted in the development of a number of different platforms for plant phenotyping (Cobb *et al.*, 2013; Li *et al.*, 2014). Commercial platforms, including the Scanalyzer series from Lemnatec (`http://www.lemnatec.com/products/`; accessed February 2016) and the Traitmill platform from CropDesign (`http://www.cropdesign.com/general.php`; accessed February 2016), have gained adoption in the research community and have promoted the development of additional software (beyond that which the respective companies provide) to analyze the images produced by the platform (**?**Hartmann *et al.*, 2011; Fahlgren *et al.*, 2015a). A variety of noncommercial platforms and methods developed by the research community also exist and have been demonstrated to perform well (White *et al.*, 2012; Fiorani and Schurr, 2013; Sirault *et al.*, 2013; Pound *et al.*, 2014). Several platforms have been deployed at sufficiently large scale to examine genomic loci underlying complex traits in crop plants such as barley (*Hordeum vulgare*) (Honsdorf *et al.*, 2014), pepper (*Capsicum annuum*) (van der Heijden *et al.*, 2012), maize (*Zea mays*) (Liu *et al.*, 2011), rice (*Oryza sativa*) (Campbell *et al.*, 2015), and wheat (*Triticum aestivum*) (Rasheed *et al.*, 2014). These successful applications of image-based phenotyping to understand the genetic bases of complex crop traits represent only a small fraction of the imaging modalities and crop species available for study. Sorghum (*Sorghum bicolor*) is the fifth most produced cereal crop in the world and is a promising bioenergy feedstock (Mullet *et al.*, 2014). Recent work has demonstrated that optimization of plant canopy architecture has the potential to improve sorghum productivity (Ort *et al.*, 2015; Truong *et al.*, 2015). As such, we sought to develop an image-based platform to examine the genetic bases of shoot architecture traits in sorghum. While commercial products like the Scanalyzer and Traitmill systems are capable of exerting fine control and extensive automation for aboveground architecture measurements, these and other current systems did not meet our specifications for phenotyping in terms of cost of entry, portability, output, throughput, or potential applicability in field phenotyping scenarios (Biskup *et al.*, 2007; Sirault *et al.*, 2013; Pound *et al.*, 2014). Thus, we sought to develop an economical (i.e. less than $1,000 U.S.) image acquisition and processing pipeline capable of nondestructively assaying sorghum canopy architecture in a portable and semiautomated fashion.

Previous work has demonstrated the potential of commercial-grade depth sensors in measuring plant architecture (Chéné *et al.*, 2012; Azzari *et al.*, 2013; Paulus *et al.*, 2014). Therefore, we used the time-of-flight depth sensor onboard Microsoft Kinect for Windows version 2 to capture depth images from multiple perspectives of individual sorghum plants, and these images were processed to construct three-dimensional (3D) representations of the imaged plants. In this manner, three replicates of 99 lines from a sorghum biparental recombinant inbred line (RIL) population were imaged at multiple time points during 1 month of development, and the resulting point clouds were registered, meshed, and segmented to generate 3D reconstructions of the imaged plants. Measurements from the segmented meshes and genotypes for the RIL population were used to identify quantitative trait loci (QTLs) underlying shoot architecture traits. We report QTLs for shoot architecture traits such as shoot height, leaf angle, and leaf length, and we demonstrate that the relative contributions to phenotypic variability of the QTLs change with respect to time. We also discuss our image analysis procedures and make our code available as part of the growing body of low-cost, open-source, image-based plant phenotyping solutions.

3.3 Results

3.3.1 3D sorghum reconstructions from depth images

To efficiently make plant architecture measurements, a portable, economical, semiautomated image acquisition and processing pipeline was developed. Image acquisition was performed using a laptop, a tripod supporting a time-of-flight depth camera, and a turntable (Figure B.1). Plants were manually transported between a greenhouse and the nearby imaging station, and, for each plant, a series of 12 depth and 12 red-green-blue (RGB) images were acquired as the plant made a 360° rotation on the turntable. Following acquisition, images were transferred to a work station and processed (Figure 3.1).

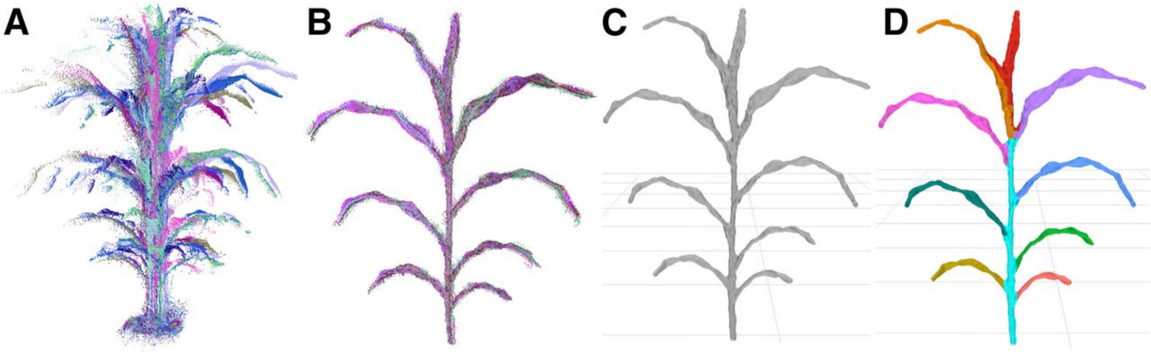

Figure 3.1: Processing of image data to segmented meshes. A, Point clouds are sampled from multiple perspectives around the plant. B, The point clouds are registered to the same frame and combined. C, The combined cloud is meshed to generate a set of polygons approximating the surface of the plant. D, The mesh is segmented into a shoot cylinder, leaves, and an inflorescence (if one exists; Figure B.2), and phenotypes are measured automatically.

Most of the processing steps use generally applicable procedures available in open-source libraries and software, including registration, cleaning, and meshing of the point clouds (Cignoni *et al.*, 2008; Rusu and Cousins, 2011; Buch *et al.*, 2013; Kazhdan and Hoppe, 2013). General solutions for the segmentation of features like leaves and stems from plants, however, remain less developed, especially for 3D plant representations (Paproki *et al.*, 2012; Paulus *et al.*, 2014; Xia *et al.*, 2015). Because of this, we developed a segmentation procedure for our particular application to partition the plant mesh into component parts. The final result of the semiautomated processing pipeline was a plant mesh segmented into a shoot cylinder, an inflorescence (when present; Figure B.2), and individual leaves, with each individual leaf assigned a relative order of emergence (Figure 3.1).

A total of 297 plants representing triplicate plantings of 99 plants (97 RILs and

Trait Type	Measurement	Description of measured trait
Composite	Shoot height	Vertical distance from the lowest shoot point to the highest shoot point, including leaves and inflorescence
	Shoot surface area	Surface area of the entire shoot
	Shoot center of mass	Vertical distance from the lowest shoot point to the shoots center of mass
	Shoot compactness	Surface area of the smallest convex polyhedron that contains the entire shoot (i.e. convex hull surface area)
Organ level	Shoot cylinder height	Vertical distance from the lowest shoot cylinder point to the highest shoot cylinder point
	Leaf length	Length of a leaf
	Leaf surface area	Surface area of a leaf
	Leaf width	Width of a leaf
	Leaf angle	Angle at which a leaf emerges from the shoot cylinder

Table 3.1: Summary of the subset of traits automatically measured from the plant mesh used for the reported QTL analyses. Additional descriptions of the methods used to obtain the measurements are found in Section B.1

the two parental lines) from the BTx623 × IS3620C sorghum mapping population were grown in a greenhouse environment (Burow *et al.*, 2011). Because image-based phenotyping is nondestructive, the same plant can be sampled at multiple time points to enable change over time to be monitored. All 297 plants were imaged at four time points over a 17-d interval starting 27 d after planting (DAP). The four time points, consisting of more than 14,000 depth images and representing nearly 1,200 samples, were processed to segmented meshes. As such, an individual plant was represented by a time course of four segmented meshes, and a RIL was represented by three biological replicates (Figure 3.2). A series of measurements from each mesh was then automatically acquired (Table 3.1).

To compare the measurements obtained from the image acquisition and process-

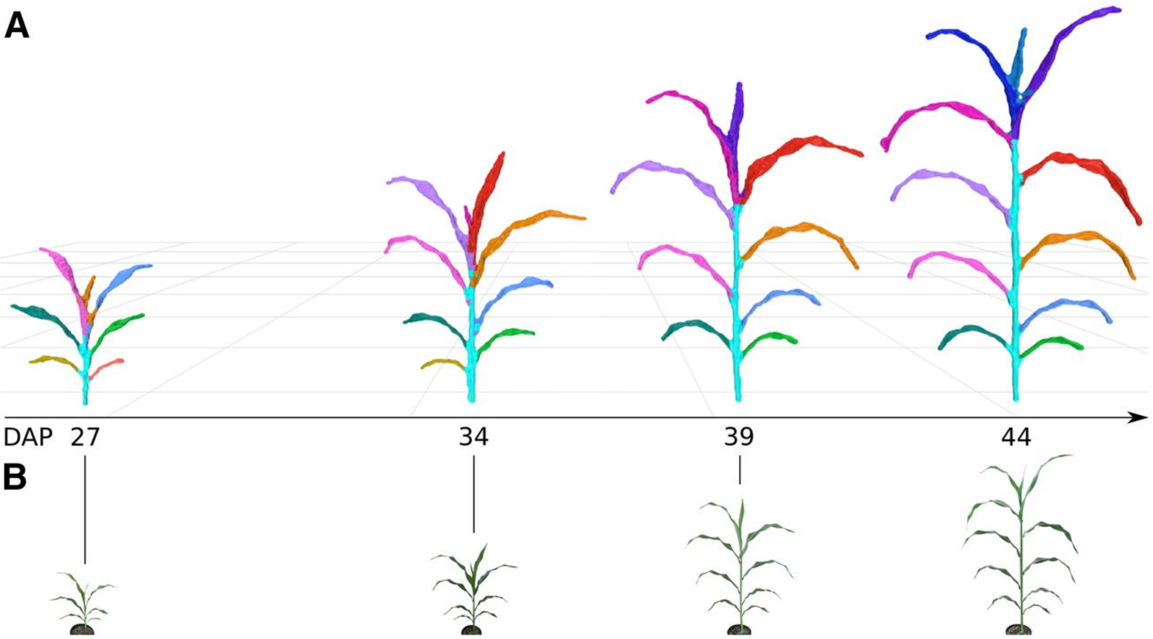

Figure 3.2: Plant growth over time. A, Segmented meshes for replicate 3 of RIL 175 are depicted at four different DAP time points. Leaf colors represent individual segmented leaves and have been assigned manually to enable tracking of the same leaf between meshes; Figure B.3 depicts how color is assigned automatically by the platform. The shoot cylinder is colored cyan. Meshes are depicted at the same relative scale. B, RGB data (not to scale) that correspond to the imaged plants and were coacquired with depth images; Figure B.3 depicts original RGB images.

ing platform with standard physical measurements of plant morphometric traits, 15 plants (with 140 leaves) from the experiment were imaged, and then leaf and stem measurements were obtained from harvested plants 62 d after planting. Shoot height, shoot cylinder height, leaf angle, leaf width, leaf length, and leaf area were compared. Leaf width and leaf length were measured using both a measuring tape and an LI-3100C Area Meter (LI-COR), and leaf area was measured using only the LI-COR instrument. Comparisons between the measurements indicated that the image-based measurements performed at least as well as the LI-COR leaf-scanning instrument for leaf width and leaf length relative to hand measurements with a measuring tape (Fig. 3.3). The RMSD between manual measurements and image-based measurements for leaf length and leaf width were 7.94 and 1.84 cm, respectively; this indicated marginally better performance than the RMSDs between manual measurements and the LI-COR instrument for leaf length and leaf width, which were 9.41 and 1.94 cm, respectively. Leaf area measurements made with the depth imaging platform and with the LI-COR instrument were well correlated (ρ of 0.92), although the image-based platform reported, on average, larger values of leaf area than the LI-COR instrument, with a mean difference of 52.45 cm^2. Leaf angle was measured with an RMSD of 9° and a ρ of 0.95 relative to hand measurements, and shoot cylinder height was measured with an RMSD of 7 cm and a ρ of 0.99. Measurements of shoot height showed the lowest correlation ($\rho = 0.63$ and RMSD = 11 cm) due to three outlier points; these outlier points likely represent errors in manual measurement due to the inherent difficulty in identifying the true maximum height point of the shoot in an unbiased way during manual measurement. We also note two leaf measurement outliers in both leaf length and leaf area that occurred because the image-based platform failed to fully reconstruct two of the leaves that were in the same vertical plane as the sensor. Ultimately, image-based measurements were well correlated with manual measurements, and the coefficient of variation of the RMSD for the measurements ranged from 0.07 to 0.3 (within the same range as measurements made using standard instrumentation). As such, measurements made with the phenotyping platform have utility for applications such as QTL mapping.

3.3.2 Genetic bases of imaged traits

To determine if the platform could be used to identify genetic loci regulating shoot architecture, measurements obtained from the plant meshes were associated with genetic data from the RIL population. Genotypes for members of the BTx623 × IS3620C RIL population were obtained previously and available to construct a genetic map for mapping QTLs for the image-based phenotypes across multiple developmental time points (Morishige *et al.*, 2013; Truong *et al.*, 2014; McCormick *et al.*, 2015). Measurements obtained from plant meshes were grouped into two categories: organ-level measurements and composite measurements. Organ-level measurements are segmentation dependent and measure organ-level plant architecture, such as leaf length and shoot cylinder height; composite measurements are segmentation independent and measure overall shoot architecture, such as shoot height and shoot compactness (Table 3.1; Figures B.4 and B.5).

QTL mapping of organ-level traits identified seven unique genomic intervals with significant contributions to phenotypic variability (Figures 3.4 and B.6; Table B.2). A genome-wide scan under a single-QTL model was used to examine the following phenotypes across the four time points: the average value of leaves 3, 4, and 5 for leaf length, width, surface area, and inclination angle as well as shoot cylinder height. Significant QTLs identified from a genome-wide scan under a single-QTL model were used as an initial model for stepwise model traversal to identify the most likely penalized multiple-QTL model (Manichaikul *et al.*, 2009); the overlapping LOD-2

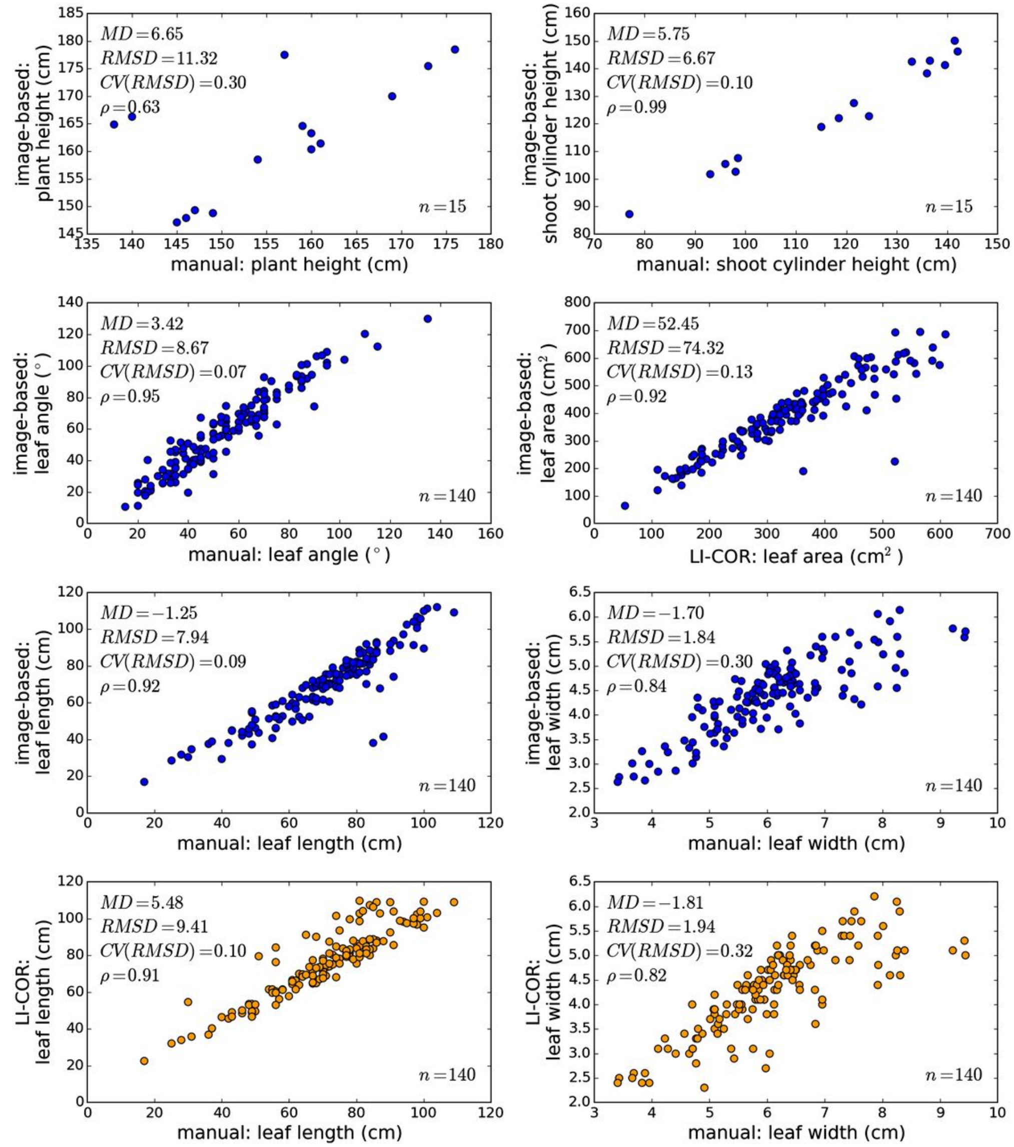

Figure 3.3: Comparison of image-based measurements with measurements made using standard methods. Axes represent measurements made via one of three methods: image-based measurements made from plant meshes, manual measurements made with a measuring tape or protractor, and measurements with a LI-COR LI-3100C Area Meter. Plots with an axis representing image-based measurements are colored blue, and plots without an axis representing image-based measurements are colored orange. Leaf area measurements made with the platform include abaxial and adaxial leaf surfaces, so the image-based area measurements were divided by two for comparison with LI-COR measurements of area. MD, Mean difference between measurements; RMSD, root-mean-square difference; CV(RMSD), coefficient of variation of the RMSD given the range of data on the bottom axis; , Pearson product-moment correlation coefficient; n, number of samples from which the differences and coefficients were calculated.

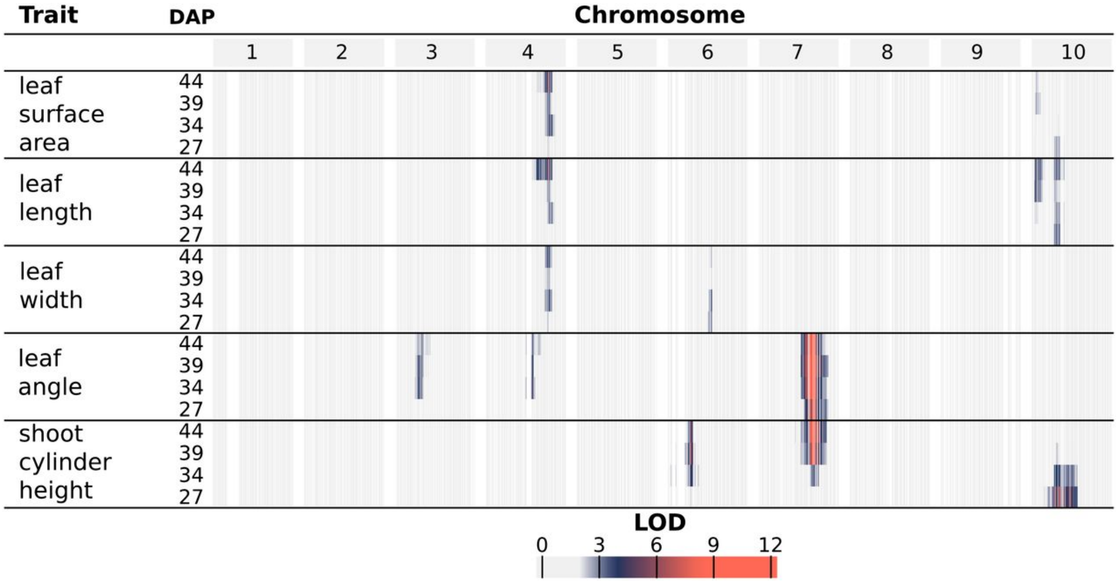

Figure 3.4: Log of the odds (LOD) profiles for organ-level traits. For each phenotype, LOD profiles are based on chromosome-wide scans of chromosomes with QTLs based on the most likely multiple-QTL models found by model selection (Figure B.6). Each row represents a different trait, and within each trait are four nested rows that each represents a different time point (DAP). Each group of columns represents a chromosome, and each column represents a marker at its genetic position. Cells are colored by marker LOD for the phenotype at the particular time point.

intervals of these multiple-QTL models define unique intervals on chromosomes 3, 4, 6, 7, and 10 (Table B.1).

A major source of variation in shoot architecture in the BTx623 × IS3620C RIL population is *Dwarf3* (*Dw3*), a sorghum dwarfing gene on chromosome 7 at 59.8 Mb. The parents of the imaged RIL population, BTx623 and IS3620C, are fixed for nonfunctional and functional forms, respectively, of the *Dw3* gene, which encodes an auxin efflux protein that has pleiotropic effects on stem elongation and additional architecture traits like leaf angle (Multani *et al.*, 2003; Truong *et al.*, 2015). A significant association between *Dw3* and shoot cylinder height is not observed until the second time point (34 DAP), while different alleles of *Dw3* introduce significant variability in leaf angle by the earliest time point (27 DAP). This is likely because *Dw3* impacts height by impacting stem elongation and the stem has not yet begun to elongate substantially by the earliest time point; as such, the nonfunctional *dw3* allele caused smaller leaf angles prior to any significant effect on stem elongation (Multani *et al.*, 2003; Truong *et al.*, 2015). Similar to *Dw3*, the effects of *Dw2*, a sorghum dwarfing gene on chromosome 6 near 42 Mb (but not yet cloned), are significantly associated with shoot cylinder height after the first time point (34, 39, and 44 DAP); unlike *Dw3*, *Dw2* is not significantly associated with any other pleiotropic effects on leaf morphology. However, an interval distinct from *Dw2* is observed on chromosome 6 near 51 Mb for leaf width.

A large interval on chromosome 10 was significantly associated with variability in leaf length and surface area as well as shoot cylinder height. While the LOD-2 intervals for these traits overlapped when comparing all phenotype-by-time point combinations, the LOD-2 interval for leaf surface area at 39 DAP was distinct from any shoot cylinder height intervals. Additionally, the significant association of the interval with shoot cylinder height is lost after 34 DAP, while the association is maintained with leaf traits throughout the time course, suggesting that multiple QTLs that regulate shoot architecture are present on chromosome 10 (Supplemental Table S1).

An interval on chromosome 4 was associated with multiple leaf traits, including length, width, and surface area, measured as the average value of leaves 3, 4, and 5 when counting green leaves starting from the bottom of the plant at the time of acquisition. Further analysis showed that plants with BTx623 alleles of an insertion/deletion (indel) marker at the leaf length maximum log of the odds (MLOD) coordinate (chromosome 4; 62.45 Mb) had a leaf length of 50.1 cm when averaged across the four time points. This was 5.6 cm longer than plants with IS3620C alleles, which had a leaf length of 44.5 cm when averaged across the four time points. Additionally, the platform captured changes in leaf length over time; plants with BTx623 alleles increased from an average length of 44.2 cm to an average length of 54.8 cm over the 17 d, whereas plants with IS3620C alleles had leaves that increased from an average length of 40.1 cm to an average length of 47.5 cm (Fig. 3.5).

Because segmentation-dependent traits represent organ-level traits that are often manually measured, QTLs identified via the image-based platform for organ-level traits were compared with QTLs identified previously for similar traits in the BTx623 × IS3620C population and previous reports on the sorghum dwarfing loci *Dw2* and *Dw3* (Hart *et al.*, 2001; Feltus *et al.*, 2006; Brown *et al.*, 2008; Mace and Jordan, 2011; Morris *et al.*, 2013; Higgins *et al.*, 2014). Most organ-level QTL intervals found in this study overlap with comparable or related traits from previous field studies (Table 3.2). Of note, some of the intervals, like chromosome 6 near 51 Mb and chromosome 4 near 62 Mb, may have multiple genes that each affect different traits or genes with pleiotropic effects, since these intervals were associated with diverse leaf morphology traits across the studies. Additionally, the genes involved could be environmentally responsive, since related but different traits were associated for the intervals when comparing the greenhouse-based and field-based studies (e.g. leaf length in this study versus leaf pitch, but not leaf length, in the previous study, where leaf pitch measures the length of the leaf from the leaf base to the apex of the naturally curved leaf). Overall, there was extensive overlap between the QTL intervals identified in previous work and those identified using the imaging platform, suggesting that these genomic loci exert phenotypic effects across multiple studies and environments.

In addition to capturing components of plant architecture like leaf morphology, the image-based measurements also capture overall architecture traits that integrate

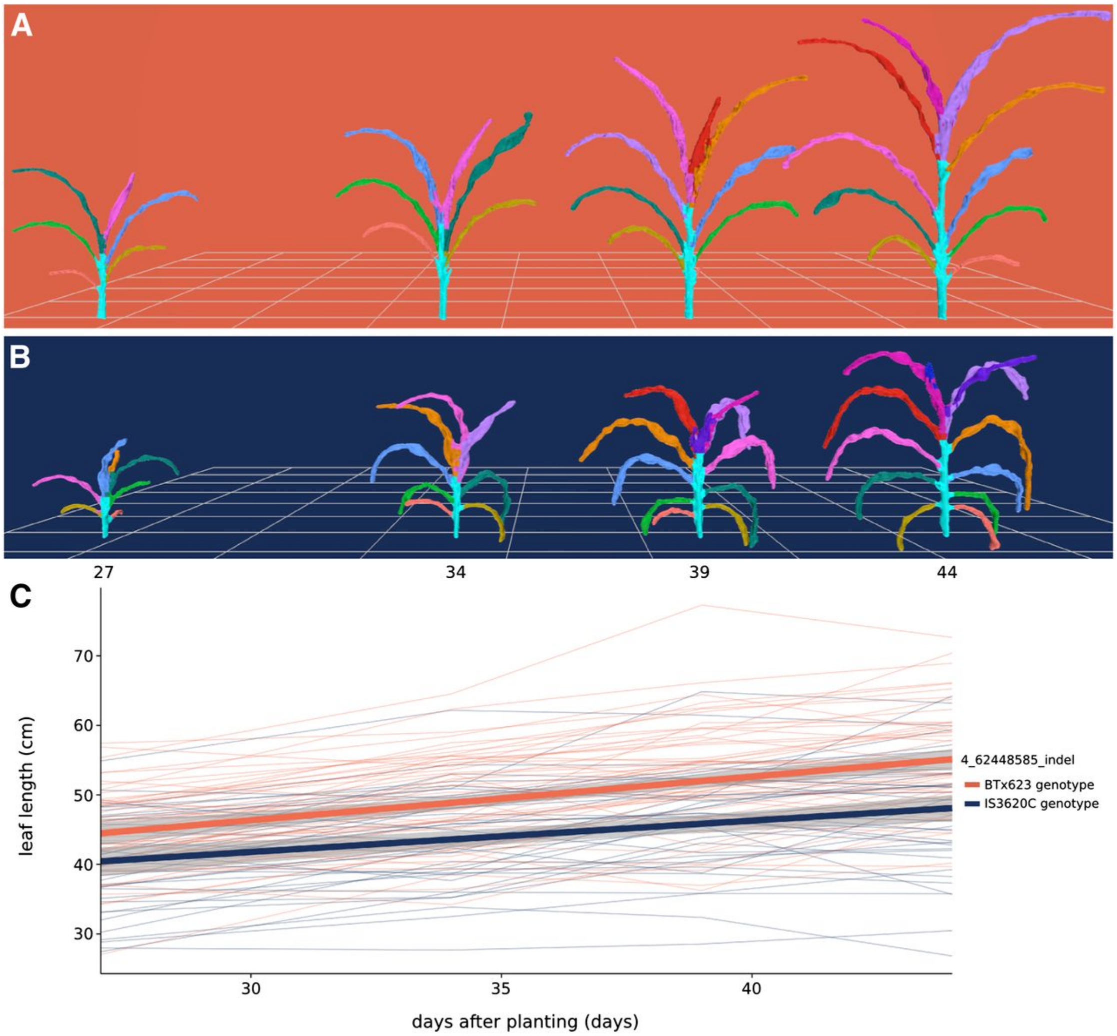

Figure 3.5: Organ-level measurement of average leaf length over time. A and B, Meshes displaying development over time for a plant bearing BTx623 alleles (A; RIL 257) and a plant bearing IS3620C alleles (B; RIL 306) of an indel marker on chromosome 4 that had the MLOD for leaf length. C, Change in average leaf length over time. Each thin line in the plot represents the average leaf length of a RIL (n = 3) colored by its genotype. Leaf length was calculated as the average of the third, fourth, and fifth leaves counting from the bottom, corresponding to the light green, dark green, and blue leaves in A and B. The two thick lines represent a linear fit for each genotype and 95% confidence intervals.

Chromosome	Origin	Interval Begin	Max MLOD Coordinate	Interval End	Related Traits with Overlapping Intervals	Prior Locus Names
4	Image based	57.48	62.45	63.40	Leaf length*, leaf surface area, leaf width	*QLcv.txs-D2*, *QLpt.txs-D*
	Literature	61.86	-	65.07	Leaf curve, leaf pitch	
6	Image based	40.10	42.77	44.83	Shoot cylinder height*	*Dw2*
	Literature	39.72	-	42.64	Preflag leaf height	
	Image based	48.45	50.97	55.08	Leaf width*	*QLcv.txs-I*
	Literature	53.73	-	56.52	Leaf curve	
7	Image based	59.51	59.65	59.99	Leaf angle*, shoot cylinder height	*Dw3*
	Literature	59.821905	-	59.829910	Leaf angle, culm height, culm uniformity	
10	Image based	1.23	2.00	8.21	Leaf length*, leaf surface area	*QLcv.txs-G*, *QLpt.txs-G*, *QLln.txs-G*, *QHtu.txs.G*, *QHGT_meta1.10*
		5.27	7.46	52.24	Shoot cylinder height*	
	Literature	1.11	-	5.76	Leaf length*, leaf surface area, leaf width	
		6.40	-	13.05	Leaf length, culm height, culm uniformity	

Table 3.2: Comparison of QTL intervals identified using image-based phenotyping with previously reported QTL intervals in the literature. Most QTL intervals identified with the platform overlapped with QTLs or causal genes reported previously for related phenotypes (Hart *et al.*, 2001; Feltus *et al.*, 2006; Brown *et al.*, 2008; Mace and Jordan, 2011; Morris *et al.*, 2013; Higgins *et al.*, 2014). *Dw3* was cloned previously (Multani *et al.*, 2003). For image-based QTL intervals, the LOD-2 interval and peak coordinate for the phenotype with the maximum MLOD are reported, and the phenotype name is indicated by an asterisk; maximum MLOD coordinates are not applicable to intervals from the literature, indicated by a dash. Table B.1 contains all identified organ-level QTLs. Leaf pitch and leaf curve are both measures of Euclidean distance from the leaf base to the apex of the curved leaf blade and from the leaf base to the leaf tip, respectively (Feltus *et al.*, 2006).

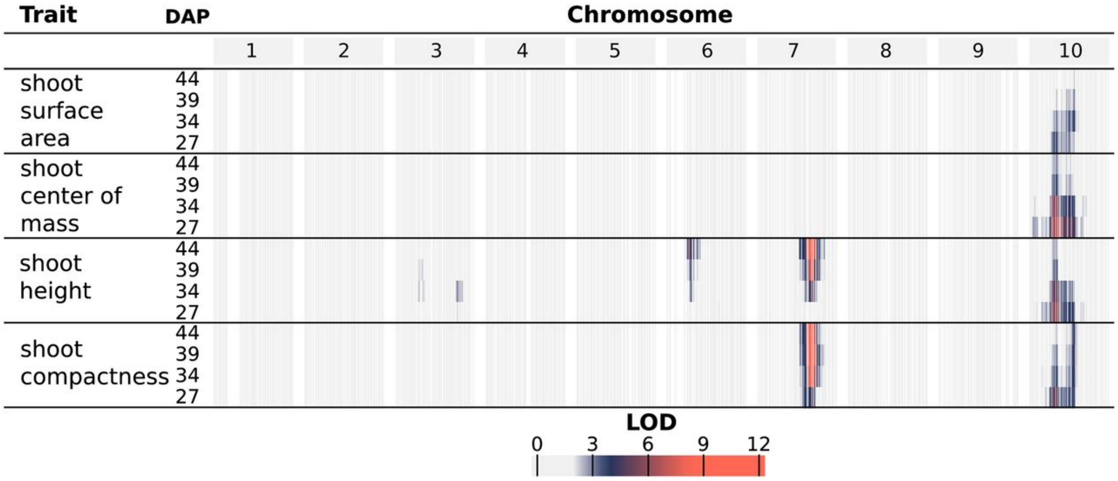

Figure 3.6: LOD profiles for composite traits. For each phenotype, LOD profiles are based on chromosome-wide scans of chromosomes with QTLs based on the most likely multiple-QTL models found by model selection (Figure B.7). Each row represents a different trait, and within each trait are four nested rows that each represents a different time point (DAP). Each group of columns represents a chromosome, and each column represents a marker at its genetic position. Cells are colored by marker LOD for the phenotype at the particular time point.

component traits. These composite measurements are difficult or impossible to measure by hand and integrate how component traits interact to influence overall plant architecture and, ultimately, how a plant canopy intercepts solar radiation. One specific example of such a measurement is shoot compactness, measured as the surface area of the convex hull of a plant mesh. Shoot compactness is influenced by factors like leaf angle and the height and planarity of a plant (Supplemental Fig. B.5). Accordingly, a strong association between Dw3 and shoot compactness is present at all time points due to the consistent effects of *Dw3* on leaf angle and later effects of *Dw3* on stem growth (Fig. 3.6). As such, composite traits represent measures of overall plant architecture and integrate the interrelationships between component phenotypes. Additional composite traits examined were shoot surface area, shoot center of mass, and shoot height, as described in Table 3.1.

QTL mapping of the selected composite traits identified four genomic intervals with significant contributions to phenotypic variability (Fig. 3.6; Supplemental Fig. B.9; Supplemental Table B.2). Since composite traits are expected to be driven by phenotypic variation in their component traits (and thus correlated), the composite trait QTLs are discussed in the context of organ-level QTLs with shared intervals. All of the composite traits were significantly associated with a large interval on chromosome 10 at early stages of development (27 and 35 DAP). Consistent with the observation of nonoverlapping QTL intervals for organ-level traits of leaf morphology and shoot cylinder height on chromosome 10, at least two QTLs are likely present in the interval; canopy compactness is a trait influenced by both leaf morphology and shoot height, and distinct LOD peaks were observed, one at 6 Mb and one at 52 Mb (Supplemental Table B.2).

Interestingly, one interval unique to the composite trait measurements was identified on chromosome 3 near 66 Mb for shoot height, indicating that there are additional component traits driving variability in overall architecture that remain to be resolved and explained by organ-level traits. Alternatively, the impact of the QTLs on individual organ-level traits is relatively small, and only the combined effects across multiple individual traits provide sufficient power for detection. As such, these composite traits represent a useful approach for detecting novel genetic loci.

Due to the effect of *Dw3* on shoot cylinder height and leaf angle, a strong association is present for shoot height and shoot compactness at the *Dw3* locus; likewise, *Dw2* is associated with shoot height. To further quantify the influence of *Dw3*, the shoot heights of individuals bearing different alleles of an indel marker near *Dw3* were compared. Plants that have the dominant, functional *Dw3* allele increase in height from, on average, 60.2 to 112.6 cm over the 17-d imaging interval, and plants with nonfunctional *dw3* alleles increase in average height from 56.8 to 93.2 cm (Fig.

3.7). Fitting a linear model to the data, *Dw3* plants grew vertically at a rate of 3.1 cm d^{1}, whereas *dw3* plants grew at a rate of 2.2 cm d^{1} between 27 and 44 DAP. Nondestructive, image-based phenotyping combined with high-throughput genotyping has great potential for parameterizing plant functional-structural modeling and performance prediction with genotype-specific rates of growth.

3.4 Discussion

A time-of-flight depth camera was used to image sorghum plants from a RIL population, and we developed an image processing pipeline to reconstruct 3D sorghum plants and make automated measurements from the reconstructions. Measurements made in this manner are sufficiently rapid and accurate to enable the identification of multiple genetic loci regulating shoot architecture. As such, we demonstrate that depth imaging represents a useful approach for high-throughput phenotyping of crop plant architecture for the genetic dissection of complex traits.

While the platform successfully identified QTLs regulating sorghum architecture (Figures 3.4 and 3.6), a number of improvements will be necessary prior to its applicability in even larger scale studies. First, the acquisition process will need to be improved. Registration artifacts were a recurring problem, introduced by nonrigid transformations of plant leaves caused by leaf shaking on the turntable, the registration methods used, and sensor noise in acquisition. Multiple potential solutions for these are available, including the use of a registration algorithm capable of handling nonrigid transformations (Zheng *et al.*, 2010; Bucksch and Khoshelham, 2013; Brophy *et al.*, 2015), the use of multiple sensors, the use of real-time mesh construction procedures like Kinect Fusion to average sensor data and rapidly reconstruct the plant (Izadi *et al.*, 2011), or the use of a model-based approach to fit a geometric plant model to the acquired points (Quan *et al.*, 2006; Ma *et al.*, 2008). Second, the

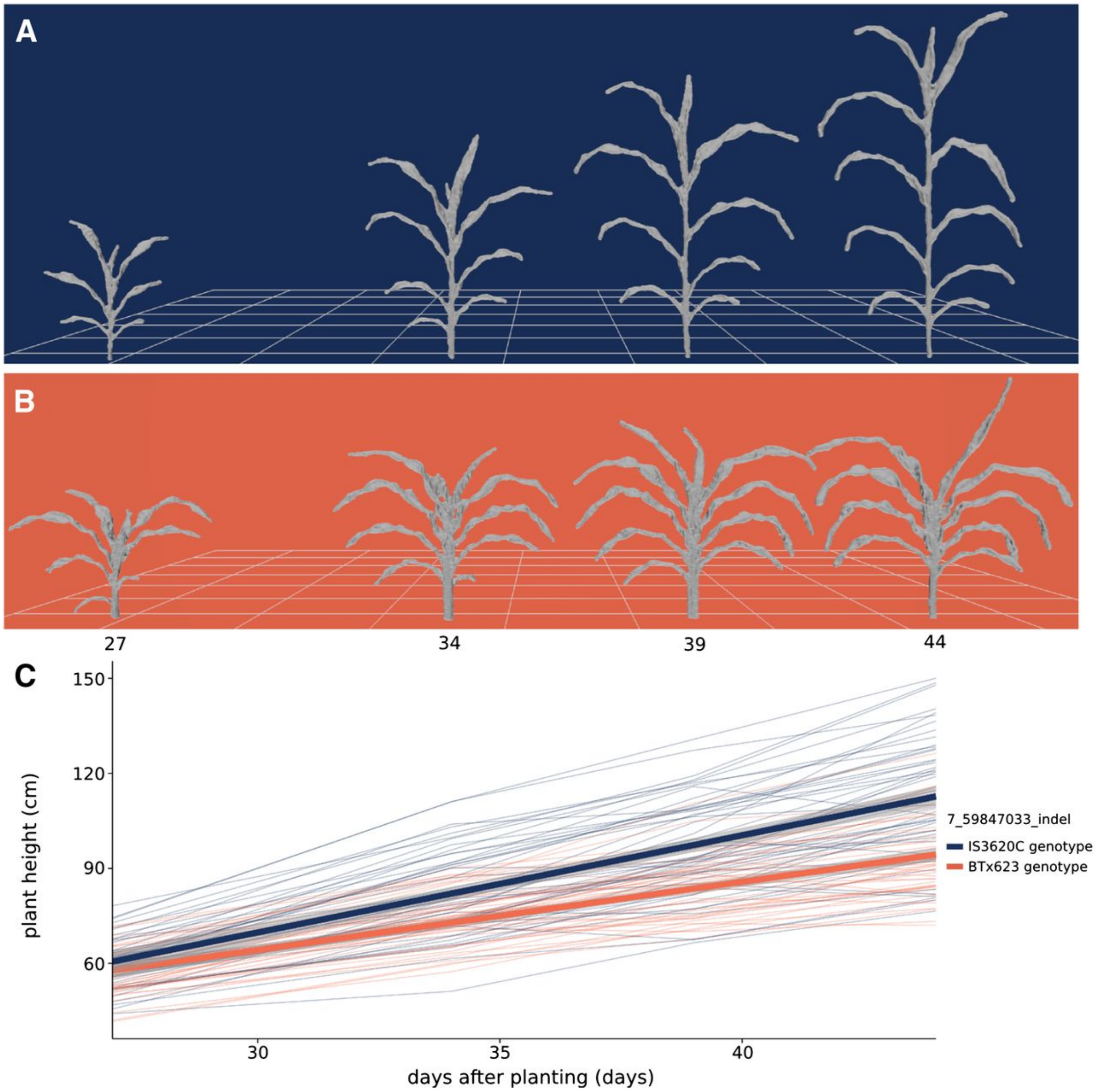

Figure 3.7: Composite measurement of shoot height over time. A and B, Meshes displaying development over time for a plant bearing IS3620C alleles (A; RIL 175) and BTx623 alleles (B; RIL 19) of an indel marker closely linked with the *Dw3* gene, an auxin transporter that regulates plant height. C, Change in plant height over time. Each thin line in the plot represents the average height of a RIL (n = 3) colored by its genotype at the *Dw3* locus. Shoot height was measured as the vertical distance from the lowest shoot point to the highest shoot point, including leaves and inflorescence (Table 3.1). The two thick lines represent a linear fit for each genotype at *Dw3* and 95% confidence intervals.

segmentation procedure will need to be improved to better distinguish leaves that are in contact with one another, to better automatically identify the shoot cylinder of the plant, and to potentially make it applicable to other grass or plant species. Progress in data-driven approaches that automatically cluster points into stem and leaf organs using feature histograms indicate that segmenting point clouds directly represents a viable option, at least for high-resolution laser scans (Paulus *et al.*, 2013; Wahabzada *et al.*, 2015). Segmenting the point cloud directly may provide the most general solution for both controlled-environment and field applications, where reconstruction prior to segmentation is difficult due to occlusion. Approaches that can accurately segment the point cloud directly also could enable automated fitting of generalized plant or organ models to the segmented cloud, potentially yielding methods that can automatically reconstruct and measure complex plant scenes.

A major benefit of image-based phenotyping is its nondestructive nature because insight into the temporal onset of genetic regulation is valuable in dissecting its mechanistic basis. Markers tightly linked with *Dw3*, a gene encoding an auxin transporter, are associated with leaf inclination angle and shoot compactness prior to their association with shoot height and shoot cylinder height, suggesting that changes in auxin transport caused by different *Dw3* alleles introduce variability in leaf development and overall shoot compactness prior to large effects on stem elongation (Figures 3.4, 3.6, and 3.7). Additionally, variation in the shoot cylinder height at the earliest time point is most associated with an interval on chromosome 10 (Figure 3.4). This QTL is the primary driver of variability in shoot height and shoot cylinder height until the variability in stem growth introduced by alleles of *Dw2* and *Dw3* increases, and it may be related to the timing of a developmental transition (Figures 3.4 and 3.6). It is likely that multiple QTLs are present on chromosome 10, given that distinct LOD peaks at 2, 7, and 52 Mb were observed; additional experimentation will be necessary to resolve the contributions and temporal prevalence of specific QTLs in the interval.

Many of the QTLs identified via image-based phenotyping overlapped with QTLs for comparable traits discovered in prior field experiments (Table 3.2). These shared QTLs represent good candidates for continued investigation, as they display robust phenotypic effects across multiple experiments and conditions. Notably, despite sharing overlapping intervals, the associated traits sometimes differed. For example, this study identified significant associations between leaf length, width, and surface area with an interval on chromosome 4; a similar interval was identified in previous work for leaf curve and leaf pitch, but it was not significantly associated with leaf length in the previous study (Table 3.2). While all of these traits are aspects of leaf morphology and share relationships, additional experimentation will be necessary to determine whether these represent one QTL with pleiotropic effects (as observed with *Dw3*), one QTL with different environmental responses, different QTLs with overlapping intervals, or some combination of these possibilities.

3.5 Conclusion

Depth imaging and subsequent processing enabled the rapid acquisition of multiple shoot architecture phenotypes from a sorghum RIL population, and genetic loci contributing to variation in shoot architecture were identified. Depth cameras represent a practical tool to rapidly measure plant morphology, and their application to plant phenotyping alongside other imaging modalities will be useful for both controlled-environment and field phenotyping scenarios. Integrated platforms that merge image-based phenotyping approaches, genetics, and performance modeling will enable rapid improvements in understanding plant biology and will promote the selection and engineering of plants for superior performance in target applications.

3.6 Materials and methods

3.6.1 Plants, greenhouse conditions, manual measurements, and image acquisition

A total of 98 RILs from the sorghum (*Sorghum bicolor*) BTx623 × IS3620C recombinant inbred mapping population and the two parents (Burow *et al.*, 2011) were planted in triplicate with five seeds per pot in C600 pots of Sunshine MVP soil (BWI) in a College Station, Texas, greenhouse on July 7, 2015. Plants were thinned to one plant per pot after germination. Plants were fertilized with Osmocote Classic (13-13-13; Everris International) and watered on demand. Tillers and senesced leaves were removed regularly. Each of the three replicates of the 100 lines was grown on a separate greenhouse table, and differences in shoot morphology were visibly apparent in the population throughout development (Figures B.8 and B.9). Seeds for one of the RILs failed to germinate (RIL 3), leaving three replicates of 99 lines for which images were acquired. Plants were imaged at 27, 34, 39, and 44 DAP. Fifteen of the plants were imaged at 62 DAP, harvested, and manually measured to compare the performance of the platform relative to standard measurement techniques. Manual measurements of leaf angle were made with a protractor, and shoot height, shoot cylinder height, leaf length, and leaf width were measured using a measuring tape. Additionally, leaf length, leaf width, and leaf area were measured using an LI-3100C Area Meter (LI-COR).

Image acquisition was performed using a Microsoft Kinect for Windows version 2 sensor and the Kinect for Windows SDK (version 2.0). Twelve RGB and 12 depth image frames were acquired at approximately 3-s intervals, and the images were saved to disk on a laptop while the Kinect for Windows version 2 sensor was positioned on a tripod in front of an Arqspin 12-inch motorized turntable that rotated the imaged plant (Figure B.1). Plants were transported manually to and from the greenhouse to the nearby imaging station. Images were transferred from the laptop to a work station for subsequent processing.

3.6.2 Processing images to acquire plant measurements

Procedures for processing images to acquire plant measurements and alternative methods that were explored are explained in Appendix Section B.1. Here, brief descriptions of procedures used for the reported analysis are outlined. For each plant, the point cloud contained in each depth image was automatically cleaned and registered to generate a single 3D point cloud using available open-source libraries and algorithms, including OpenCV (`http://opencv.org`; accessed February 2016) and PCL (Fischler and Bolles, 1981; Besl and McKay, 1992; Rabbani *et al.*, 2006; Rusu *et al.*, 2008; Rusu and Cousins, 2011; Buch *et al.*, 2013). This point cloud was inspected manually, acquisition and/or registration errors were corrected manually using MeshLab (Cignoni *et al.*, 2008), and the cleaned point cloud was meshed to generate a set of polygons representing the surface of the plant using available open-source software (Bernardini *et al.*, 1999; Corsini *et al.*, 2012; Kazhdan and Hoppe, 2013). The plant mesh was then segmented into a shoot cylinder (composed of the stem and leaf sheaths), individual leaves, and an inflorescence (when present; Figure B.2). The shoot cylinder and inflorescence were labeled manually. Following this, individual leaves were segmented using an automated procedure we developed that uses supervoxel adjacency and geodesic paths across the adjacency graph to identify leaf tips and grow leaf regions (Dijkstra, 1959; Surazhsky *et al.*, 2005; Papon *et al.*, 2013).

Multiple measurements were automatically obtained from each mesh, both at the level of the whole plant (i.e. segmentation-independent, composite traits) and at the organ level (i.e. segmentation-dependent, organ-level traits). The traits measured are described in Table 3.2. Descriptions of how these traits were measured from the plant mesh are provided in Appendix Section B.1, and graphical depictions of selected measurements are shown in Supplemental Figures B.4 and B.5. Additional implementation details can be found with the code base (see Section 3.6.4 below).

3.6.3 QTL mapping and comparison with prior QTL studies from the literature

Genotypes for the BTx623 × IS3620C RIL population were generated previously using Digital Genotyping, a restriction enzyme-based, reduced-representation sequencing assay (Morishige *et al.*, 2013). Genotypes were called using the naive pipeline of the RIG workflow with the GATK, and the genetic map was constructed as described previously with marker orderings relative to the version 3 assembly of the sorghum reference genome, Sbi3 (Department of Energy-Joint Genome Institute `http://phytozome.jgi.doe.gov`; accessed February 2016); this resulted in a genetic map with 10,787 markers (McKenna *et al.*, 2010; Goodstein *et al.*, 2012; Truong *et al.*, 2014; McCormick *et al.*, 2015). Both single- and multiple-QTL mapping were performed with R/qtl (Broman *et al.*, 2003). For single-QTL mapping (i.e. testing a single-QTL model), the complete marker set of 10,787 markers was used. Measurements of a trait for each of the three replicates of a RIL were averaged; average trait values were normalized using empirical normal quantile transformation prior to QTL mapping, so that the same permutation threshold would apply to all phenotype-by-time point combinations (Peng *et al.*, 2007). A genome-wide scan under a single-QTL model for each phenotype-by-time point combination was performed (Figures B.6 and B.7). If any of the reported phenotype-by-time point combinations had a marker with a LOD greater than 3.28 (the 95% threshold obtained from 25,000 permutations), its LOD-2 interval (the coordinates of the flanking markers where the LOD had dropped by 2 units below the peak value) was retained.

The positions (centimorgans) with the largest LOD within each LOD-2 interval for each phenotype-by-time point combination were retained to initialize multiple-QTL mapping.

For multiple-QTL mapping, a subset of 1,209 markers was obtained by enforcing a minimum marker distance of 0.8 centimorgans; significant peak-LOD markers from single-QTL mapping intervals were added back to the set if they were dropped, resulting in 1,224 markers used for multiple-QTL mapping. The genetic coordinates of the markers with the largest LOD for each LOD-2 interval from single-QTL mapping of each phenotype-by-time point combination were used to seed model selection for multiple-QTL mapping as implemented in R/qtl (Manichaikul *et al.*, 2009). Main effect, heavy chain, and light chain penalties (3.2, 4.38, and 1.94, respectively) for model selection were obtained as 95% thresholds from 25,000 permutations of the appropriate statistics. The multiple-QTL models with the largest penalized LOD for each phenotype-by-time point combination are reported (Tables 3.2, B.1, and B.2; Figures B.6 and B.7). For a given phenotype, the maximum LOD across all time points characterized the MLOD of the phenotype (Kwak *et al.*, 2014). A longitudinal QTL model for each phenotype that contained QTLs at the MLOD coordinates was used to generate the chromosome-wide LOD profile scans (Figures 3.4 and 3.6).

To compare QTLs found in this study with existing QTLs in the literature, the physical coordinates relative to the sorghum version 1 reference assembly, Sbi1, for QTLs in the BTx623 × IS3620C population were obtained; Mace and Jordan (2011) determined these physical coordinates using a consensus map and QTLs identified by Hart *et al.* (2001) and Feltus *et al.* (2006). The coordinates of *Dw2* and *Dw3* were obtained from Morris *et al.* (2013) and Multani *et al.* (2003). The corresponding locations of the markers in Sbi3 were obtained using Biopieces (`www.biopieces.org`) for sequence extraction and BLAST via a local instance of Sequenceserver (Altschul *et al.*, 1997; Paterson *et al.*, 2009; Priyam *et al.*, 2015). Physical locations relative to Sbi3 were used as the QTL intervals for comparison with this study.

3.6.4 Code and data availability

The C++, Bash, and Python code written for image acquisition and processing, the R code written for QTL mapping, the genotype and phenotype data, and the full multiple-QTL models for each phenotype-by-time point combination can be found on GitHub at

`https://github.com/MulletLab/SorghumReconstructionAndPhenotyping`

For each imaged plant, its depth images, a single RGB image, and the segmented mesh can be found at the Dryad Digital Repository (`http://dx.doi.org/10.5061/dryad.9vs26`).

4. CONCLUSION

Sustainably meeting the projected food and fuel demands of the future in a carbon neutral manner will necessitate increased crop outputs without increasing input. Towards the goal of sustainably increasing crop productivity, this dissertation described two contributions to improve high-throughput genotyping and phenotyping in agricultural systems. While additional work remains to improve the genotyping of individuals from sequence data, crop improvement is currently limited predominately by phenotyping rate, otherwise referred to as the phenotyping bottleneck. The field of high-throughput plant phenotyping is still relatively new, and much work remains to develop cost-effective and robust automated platforms for high-throughput phenotyping of crop plants both in the field and in greenhouse settings (Bao and Tang, 2016; Dengyu *et al.*, 2016; Gélard *et al.*, 2016; Jiang *et al.*, 2016; Zhang *et al.*, 2016a,b). Interpreting the variety of measurements obtained with these methods promises to be a challenge in of itself, as traits resulting from various image transformations might not have an intuitive biological interpretations; this has even spawned a new term, cryptotype, to represent phenotypes that measure abstract traits (e.g., a principal component) capable of separating individuals into pre-defined classes (Chitwood and Topp, 2015). As throughput continues to improve for both genotyping and phenotyping technologies, novel analysis approaches will need to be developed that can convert terabytes of genotypic and phenotypic information into actionable conclusions.

An alternative means to solving the phenotyping bottleneck is to reduce the number of individuals that need to be phenotyped using predictive models. One promising research area is that of performance prediction, whereby information about an individual's genome and its target production environment can be used to predict how the individual will perform (Cooper *et al.*, 2016; Hammer *et al.*, 2016). This type of performance prediction is readily enabled by the work presented in this dissertation, as the genomic information and plant reconstructions are amenable to functional-structural plant modeling and light interception simulations to determine how efficiently a given crop canopy intercepts incident solar radiation (Figure 4.1). Integrating genotypic information, high-throughput phenotyping, and performance predictions promises to rapidly improve the rate of crop improvement by enabling informed decisions on which germplasm should be prioritized for field trials, as well as which traits should be targeted for engineering traits that improve productivity.

Ultimately, improving the rate of crop improvement will require useful genomics and phenomics approaches that can be integrated within existing breeding strategies along with novel modeling and engineering methods to generate robust and high-performing cultivars. Doing so will require interdisciplinary teams of plant physiologists, crop breeders, robotics engineers, computer scientists, synthetic biologists, and mathematical biologists, and the lessons learned will further inform metabolic and genome engineering as plants are designed for specific end applications beyond producing grain or lignocellulosic biomass (Medford and Prasad, 2014; Nemhauser and Torii, 2016).

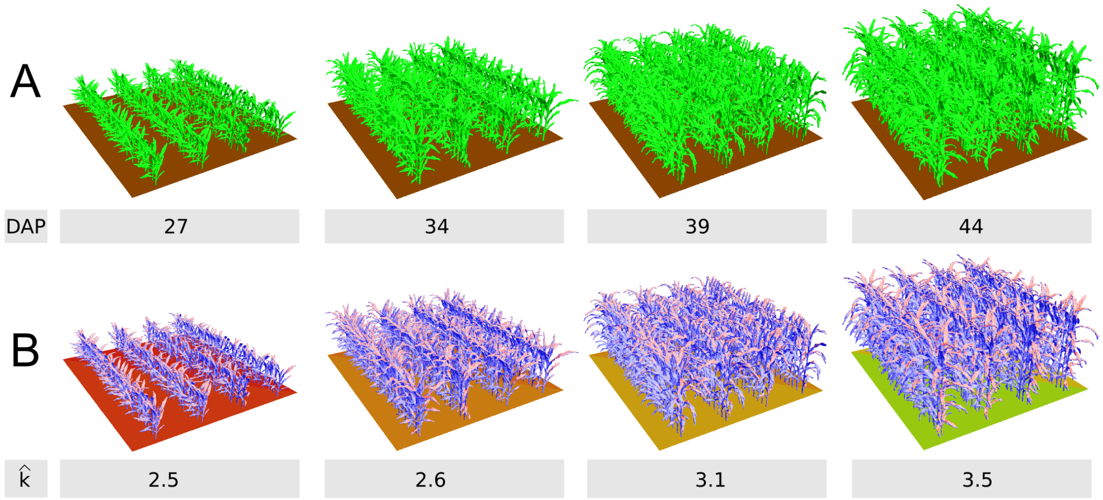

Figure 4.1: Simulated field planting and light interception of a 3D plant reconstructions. Plant reconstructions of a RIL line at each of four timepoints were used to simulate how the plant canopy would intercept solar radiation using OpenAlea and ScanAlea at each developmental time point (Chelle and Andrieu, 1998; Chelle *et al.*, 2004; Pradal *et al.*, 2008). (A) Simulated field plantings with plant reconstructions colored green. (B) Lighting simulation using a nested radiosity model where surfaces are colored based on the amount of intercepted light; the ground intercepts more light at early developmental states prior to canopy closure. The coefficient of light extinction, $\hat{k}$, is obtained by fitting the light distribution at layers down the canopy (scaled to account for plant height) to Beer-Lambert's Law, and the coefficient quantifies the observation that the top of the canopy intercepts more light as it closes during development.

REFERENCES

Alexandratos, N., J. Bruinsma, *et al.*, 2012 World agriculture towards 2030/2050: the 2012 revision. ESA Working Paper **3**.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**: 3389–3402.

Arabidopsis Genome Initiative, *et al.*, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**: 796–815.

Azzari, G., M. L. Goulden, and R. B. Rusu, 2013 Rapid characterization of vegetation structure with a Microsoft Kinect sensor. Sensors **13**: 2384–2398.

Bao, Y., and L. Tang, 2016 Field-based robotic phenotyping for sorghum biomass yield component traits characterization using stereo vision. In *5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture*, volume 49. Elsevier, 265–270.

Barber, C. B., D. P. Dobkin, and H. Huhdanpaa, 1996 The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software **22**: 469–483.

Benbouzid, D., R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. Kégl, 2012 MultiBoost: a multi-purpose boosting package. The Journal of Machine Learning Research **13**: 549–553.

Bernardini, F., J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, 1999 The ball-pivoting algorithm for surface reconstruction. IEEE Transactions on Visualization and Computer Graphics **5**: 349–359.

Besl, P. J., and N. D. McKay, 1992 Method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**: 239–256.

Biskup, B., H. Scharr, U. Schurr, and U. Rascher, 2007 A stereo imaging system for measuring structural parameters of plant canopies. Plant, Cell & Environment **30**: 1299–1308.

Boudon, F., C. Pradal, T. Cokelaer, P. Prusinkiewicz, and C. Godin, 2012 L-Py: an L-system simulation framework for modeling plant architecture development based on a dynamic language. Frontiers in Plant Science **3**: 76.

Bout, S., and W. Vermerris, 2003 A candidate-gene approach to clone the sorghum brown midrib gene encoding caffeic acid O-methyltransferase. Molecular Genetics and Genomics **269**: 205–214.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889–890.

Brophy, M., A. Chaudhury, S. S. Beauchemin, and J. L. Barron, 2015 A method for global non-rigid registration of multiple thin structures. In *2015 12th Conference on Computer and Robot Vision*. IEEE, 214–221.

Brown, P. J., W. L. Rooney, C. Franks, and S. Kresovich, 2008 Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics **180**: 629–637.

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics **81**: 1084–1097.

Buch, A. G., D. Kraft, J.-K. Kamarainen, H. G. Petersen, and N. Krüger, 2013 Pose estimation using local structure-specific shape and appearance context. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2080–2087.

Bucksch, A., and K. Khoshelham, 2013 Localized registration of point clouds of botanic trees. IEEE Geoscience and Remote Sensing Letters **10**: 631–635.

Burow, G., R. Klein, C. Franks, P. Klein, K. Schertz, *et al.*, 2011 Registration of the

BTx623/IS3620C recombinant inbred mapping population of sorghum. Journal of Plant Registrations **5**: 141–145.

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

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender, *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nature Genetics **43**: 956–963.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait, 2011 Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes—Genomes—Genetics **1**: 171–182.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4**: 7.

Chelle, M., and B. Andrieu, 1998 The nested radiosity model for the distribution of light within plant canopies. Ecological Modelling **111**: 75–91.

Chelle, M., J. Hanan, and H. Autret, 2004 Lighting virtual crops: the CARIBU solution for open L-systems. In *4th International Workshop on Functional-Structural Plant Models*. FSPM, 194.

Chéné, Y., D. Rousseau, P. Lucidarme, J. Bertheloot, V. Caffier, *et al.*, 2012 On the use of depth camera for 3D phenotyping of entire plants. Computers and Electronics in Agriculture **82**: 122–127.

Chitwood, D. H., and C. N. Topp, 2015 Revealing plant cryptotypes: defining meaningful phenotypes among infinite traits. Current Opinion in Plant Biology **24**: 54–60.

Cignoni, P., M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, *et al.*, 2008 MeshLab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conference*. The Eurographics Association, 129–136.

Cobb, J. N., G. DeClerck, A. Greenberg, R. Clark, and S. McCouch, 2013 Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theoretical and Applied Genetics **126**: 867–887.

Cooper, M., F. Technow, C. Messina, C. Gho, and L. R. Totir, 2016 Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. Crop Science **56**: 2141–2156.

Cornea, N. D., D. Silver, and P. Min, 2007 Curve-skeleton properties, applications, and algorithms. IEEE Transactions on Visualization and Computer Graphics **13**: 530–548.

Corsini, M., P. Cignoni, and R. Scopigno, 2012 Efficient and flexible sampling with blue noise properties of triangular meshes. IEEE Transactions on Visualization and Computer Graphics **18**: 914–924.

Dengyu, X., G. Liang, L. Chengliang, and H. Yixiang, 2016 Phenotype-based robotic screening platform for leafy plant breeding. 5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture **49**: 237–241.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics **43**: 491–498.

Dijkstra, E. W., 1959 A note on two problems in connexion with graphs. Numerische Mathematik **1**: 269–271.

Evans, J., R. F. McCormick, D. Morishige, S. N. Olson, B. Weers, *et al.*, 2013 Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. PLoS One **8**: e79192.

Fahlgren, N., M. Feldman, M. A. Gehan, M. S. Wilson, C. Shyu, *et al.*, 2015a A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in setaria. Molecular Plant **8**: 1520–1535.

Fahlgren, N., M. A. Gehan, and I. Baxter, 2015b Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Current Opinion in Plant Biology **24**: 93–99.

Feltus, F., G. Hart, K. Schertz, A. Casa, S. Kresovich, *et al.*, 2006 Alignment of genetic maps and QTLs between inter-and intra-specific sorghum populations. Theoretical and Applied Genetics **112**: 1295–1305.

Fiorani, F., and U. Schurr, 2013 Future scenarios for plant phenotyping. Annual Review of Plant Biology **64**: 267–291.

Fischler, M. A., and R. C. Bolles, 1981 Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**: 381–395.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, *et al.*, 2005 Global consequences of land use. Science **309**: 570–574.

Foley, J. A., N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, *et al.*, 2011 Solutions for a cultivated planet. Nature **478**: 337–342.

Furbank, R. T., and M. Tester, 2011 Phenomics–technologies to relieve the phenotyping bottleneck. Trends in Plant Science **16**: 635–644.

Gélard, W., P. Burger, P. Casadebaig, N. Langlade, P. Debaeke, *et al.*, 2016 3D plant phenotyping in sunflower using architecture-based organ segmentation from 3D point clouds. In *5th International Workshop on Image Analysis Methods for the Plant Sciences*.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, *et al.*, 2014

TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One **9**: e90346.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research **40**: D1178–D1186.

Hammer, G., G. McLean, A. Doherty, E. van Oosterom, and S. Chapman, 2016 *Sorghum Crop Modeling and Its Utility in Agronomy and Breeding*. Number Agron. Monogr. 58. American Society of Agronomy and Crop Science Society of America, Inc., Madison, WI, USA.

Hammer, G. L., E. van Oosterom, G. McLean, S. C. Chapman, I. Broad, *et al.*, 2010 Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. Journal of Experimental Botany **61**: 2185–2202.

Hart, G. E., K. F. Schertz, Y. Peng, and N. H. Syed, 2001 Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. Theoretical and Applied Genetics **103**: 1232–1242.

Hartmann, A., T. Czauderna, R. Hoffmann, N. Stein, and F. Schreiber, 2011 HT-Pheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinformatics **12**: 148.

Higgins, R. H., C. S. Thurber, I. Assaranurak, and P. J. Brown, 2014 Multiparental mapping of plant height and flowering time QTL in partially isogenic sorghum families. G3: Genes—Genomes—Genetics **4**: 1593–1602.

Honsdorf, N., T. J. March, B. Berger, M. Tester, and K. Pillen, 2014 High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. PLoS One **9**: e97047.

Houle, D., D. R. Govindaraju, and S. Omholt, 2010 Phenomics: the next challenge. Nature Reviews Genetics **11**: 855–866.

Izadi, S., D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, *et al.*, 2011 KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, 559–568.

Jiang, Y., C. Li, and A. H. Paterson, 2016 High throughput phenotyping of cotton plant height using depth images under field conditions. Computers and Electronics in Agriculture **130**: 57–68.

Kazhdan, M., and H. Hoppe, 2013 Screened poisson surface reconstruction. ACM Transactions on Graphics **32**: 29.

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.

Li, H., 2014 Towards better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics **30**: 2843–2851.

Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**: 589–595.

Li, L., Q. Zhang, and D. Huang, 2014 A review of imaging techniques for plant phenotyping. Sensors **14**: 20078–20111.

Li, M., M. Nordborg, and L. M. Li, 2004 Adjust quality scores from alignment and improve sequencing accuracy. Nucleic Acids Research **32**: 5183–5191.

Liu, X., S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang, 2013 Variant callers for next-generation sequencing data: a comparison study. PLoS One **8**: e75619.

Liu, Y., C. Subhash, J. Yan, C. Song, J. Zhao, *et al.*, 2011 Maize leaf temperature responses to drought: Thermal imaging and quantitative trait loci (QTL) mapping. Environmental and Experimental Botany **71**: 158–165.

Ma, W., H. Zha, J. Liu, X. Zhang, and B. Xiang, 2008 Image-based plant modeling by knowing leaves from their apexes. In *19th International Conference on Pattern Recognition*. IEEE, 1–4.

Mace, E., and D. Jordan, 2011 Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. Theoretical and Applied Genetics **123**: 169–191.

Mace, E. S., S. Tai, E. K. Gilding, Y. Li, P. J. Prentis, *et al.*, 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nature Communications **4**.

Manichaikul, A., J. Y. Moon, Ś. Sen, B. S. Yandell, and K. W. Broman, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics **181**: 1077–1086.

McCormick, R. F., S. K. Truong, and J. E. Mullet, 2015 RIG: recalibration and interrelation of genomic sequence data with the GATK. G3: Genes— Genomes— Genetics **5**: 655–665.

McCormick, R. F., S. K. Truong, and J. E. Mullet, 2016 3D sorghum reconstructions from depth images identify QTL regulating shoot architecture. Plant Physiology **172**: 823–834.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

Medford, J. I., and A. Prasad, 2014 Plant synthetic biology takes root. Science **346**: 162–163.

Minervini, M., H. Scharr, and S. A. Tsaftaris, 2015 Image analysis: The new bottleneck in plant phenotyping. IEEE signal processing magazine **32**: 126–131.

Monaco, M. K., J. Stein, S. Naithani, S. Wei, P. Dharmawardhana, *et al.*, 2014

Gramene 2013: comparative plant genomics resources. Nucleic Acids Research **42**: D1193–D1199.

Morishige, D. T., P. E. Klein, J. L. Hilley, S. M. E. Sahraeian, A. Sharma, *et al.*, 2013 Digital genotyping of sorghum–a diverse plant species with a large repeat-rich genome. BMC Genomics **14**: 448.

Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah, *et al.*, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences **110**: 453–458.

Mullet, J., D. Morishige, R. McCormick, S. Truong, J. Hilley, *et al.*, 2014 Energy sorghum a genetic model for the design of C4 grass bioenergy crops. Journal of Experimental Botany **65**: 3479–3489.

Multani, D. S., S. P. Briggs, M. A. Chamberlin, J. J. Blakeslee, A. S. Murphy, *et al.*, 2003 Loss of an MDR transporter in compact stalks of maize *br2* and sorghum *dw3* mutants. Science **302**: 81–84.

Nekrutenko, A., and J. Taylor, 2012 Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature Reviews Genetics **13**: 667–672.

Nemhauser, J. L., and K. U. Torii, 2016 Plant synthetic biology for molecular engineering of signalling and development. Nature Plants **2**: 16010.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biology **3**: e196.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, *et al.*, 2013 Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Medicine **5**: 28.

Ort, D. R., S. S. Merchant, J. Alric, A. Barkan, R. E. Blankenship, *et al.*, 2015 Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proceedings of the National Academy of Sciences **112**: 8529–8536.

Papon, J., A. Abramov, M. Schoeler, and F. Worgotter, 2013 Voxel cloud connectivity segmentation-supervoxels for point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2027–2034.

Paproki, A., X. Sirault, S. Berry, R. Furbank, and J. Fripp, 2012 A novel mesh processing based technique for 3D plant analysis. BMC Plant Biology **12**: 63.

Park, S.-Y., F. C. Peterson, A. Mosquna, J. Yao, B. F. Volkman, *et al.*, 2015 Agrochemical control of plant water use using engineered abscisic acid receptors. Nature **520**: 545–548.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, *et al.*, 2009 The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**: 551–556.

Paulus, S., J. Behmann, A.-K. Mahlein, L. Plümer, and H. Kuhlmann, 2014 Low-cost 3D systems: suitable tools for plant phenotyping. Sensors **14**: 3001–3018.

Paulus, S., J. Dupuis, A.-K. Mahlein, and H. Kuhlmann, 2013 Surface feature based classification of plant organs from 3D laserscanned point clouds for plant phenotyping. BMC Bioinformatics **14**: 1.

Peng, B., K. Y. Robert, K. L. DeHoff, and C. I. Amos, 2007 Normalizing a large number of quantitative traits using empirical normal quantile transformation. In *BMC Proceedings*, number Suppl 1. BioMed Central Ltd, S156.

Pirooznia, M., M. Kramer, J. Parla, F. Goes, J. Potash, *et al.*, 2014 Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics **8**: 14.

Pound, M. P., A. P. French, E. H. Murchie, and T. P. Pridmore, 2014 Automated recovery of three-dimensional models of plant shoots from multiple color images. Plant Physiology **166**: 1688–1698.

Pradal, C., S. Dufour-Kowalski, F. Boudon, C. Fournier, and C. Godin, 2008 OpenAlea: a visual programming and component-based software platform for plant modelling. Functional Plant Biology **35**: 751–760.

Priyam, A., B. J. Woodcroft, V. Rai, A. Munagala, I. Moghul, *et al.*, 2015 Sequenceserver: a modern graphical user interface for custom BLAST databases. Biorxiv **1**: 033142.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics **81**: 559–575.

Puritz, J., C. M. Hollenbeck, and J. R. Gold, 2014 dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ **2**: e431.

Quan, L., P. Tan, G. Zeng, L. Yuan, J. Wang, *et al.*, 2006 Image-based plant modeling **25**: 599–604.

Rabbani, T., F. Van Den Heuvel, and G. Vosselmann, 2006 Segmentation of point clouds using smoothness constraint. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **36**: 248–253.

Rasheed, A., X. Xia, F. Ogbonnaya, T. Mahmood, Z. Zhang, *et al.*, 2014 Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. BMC Plant Biology **14**: 128.

Rusu, R. B., N. Blodow, and M. Beetz, 2009a Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation*. IEEE, 3212–3217.

Rusu, R. B., and S. Cousins, 2011 3D is here: Point cloud library (PCL). In *IEEE International Conference on Robotics and Automation*. IEEE, 1–4.

Rusu, R. B., A. Holzbach, N. Blodow, and M. Beetz, 2009b Fast geometric point labeling using conditional random fields. In *IEEE/RSJ International Conference* *on Intelligent Robots and Systems*. IEEE, 7–12.

Rusu, R. B., Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, 2008 Towards 3D point cloud based object maps for household environments. Robotics and Autonomous Systems **56**: 927–941.

Schapire, R. E., and Y. Singer, 1999 Improved boosting algorithms using confidence-rated predictions. Machine Learning **37**: 297–336.

Schmitz, R. J., M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola, *et al.*, 2013 Patterns of population epigenomic diversity. Nature **495**: 193–198.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics **15**: 121–132.

Sirault, X., J. Fripp, A. Paproki, P. Kuffner, C. Nguyen, *et al.*, 2013 PlantScan: a three-dimensional phenotyping platform for capturing the structural dynamic of plant development and growth. In *7th International Conference on Functional-Structural Plant Models*. FSPM, 45–48.

Somerville, C., H. Youngs, C. Taylor, S. C. Davis, and S. P. Long, 2010 Feedstocks for lignocellulosic biofuels. Science **329**: 790–792.

Spalding, E. P., and N. D. Miller, 2013 Image analysis is driving a renaissance in growth measurement. Current Opinion in Plant Biology **16**: 100–104.

Surazhsky, V., T. Surazhsky, D. Kirsanov, S. J. Gortler, and H. Hoppe, 2005 Fast exact and approximate geodesics on meshes. ACM Trans. Graph. **24**: 553–560.

Technow, F., C. D. Messina, L. R. Totir, and M. Cooper, 2015 Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS One **10**: e0130855.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.

Briefings in Bioinformatics **14**: 178–192.

Truong, S. K., R. F. McCormick, D. T. Morishige, and J. E. Mullet, 2014 Resolution of genetic map expansion caused by excess heterozygosity in plant recombinant inbred populations. G3: Genes—Genomes—Genetics **4**: 1963–1969.

Truong, S. K., R. F. McCormick, W. L. Rooney, and J. E. Mullet, 2015 Harnessing genetic variation in leaf angle to increase productivity of *Sorghum bicolor*. Genetics **201**: 1229–1238.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, *et al.*, 2013 *From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

van der Heijden, G., Y. Song, G. Horgan, G. Polder, A. Dieleman, *et al.*, 2012 SPICY: towards automated phenotyping of large pepper plants in the greenhouse. Functional Plant Biology **39**: 870–877.

Wahabzada, M., S. Paulus, K. Kersting, and A.-K. Mahlein, 2015 Automated interpretation of 3D laserscanned point clouds for plant organ segmentation. BMC bioinformatics **16**: 248.

White, J. W., P. Andrade-Sanchez, M. A. Gore, K. F. Bronson, T. A. Coffelt, *et al.*, 2012 Field-based phenomics for plant genetics research. Field Crops Research **133**: 101–112.

Xia, C., L. Wang, B.-K. Chung, and J.-M. Lee, 2015 In situ 3D segmentation of individual plant leaves using a RGB-D camera for agricultural automation. Sensors **15**: 20463–20479.

Xu, W., P. K. Subudhi, O. R. Crasta, D. T. Rosenow, J. E. Mullet, *et al.*, 2000 Molecular mapping of QTLs conferring stay-green in grain sorghum (*Sorghum bicolor* L. Moench). Genome **43**: 461–469.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Zhang, C., H. Gao, J. Zhou, A. Cousins, M. O. Pumphrey, *et al.*, 2016a 3D robotic system development for high-throughput crop phenotyping. In *5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture*, volume 49. Elsevier, 242–247.

Zhang, J., L. Gong, C. Liu, Y. Huang, D. Zhang, *et al.*, 2016b Field phenotyping robot design and validation for the crop breeding. In *5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture*, volume 49. Elsevier, 281–286.

Zheng, L.-Y., X.-S. Guo, B. He, L.-J. Sun, Y. Peng, *et al.*, 2011 Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biology **12**: R114.

Zheng, Q., A. Sharf, A. Tagliasacchi, B. Chen, H. Zhang, *et al.*, 2010 Consensus skeleton for non-rigid space-time registration. Computer Graphics Forum **29**: 635–644.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR RIG: RECALIBRATION AND INTERRELATION OF GENOMIC SEQUENCE DATA WITH THE GATK [1]

Table A.1: Recovery of variants in the Independent-Family set within the WGS sets. The table shows the intersection of variants between the Independent-Family (IF) set and the Raw, Sensitive, and Specific sets from 49 WGS samples. The IF set is comprised of genetically mappable variants from a biparental cross, and the Raw, Sensitive, and Specific sets correspond to the variant calls generated from 49 WGS samples at the 100%, 95%, and 75% tranches, respectively, for both the SNP and indel models. The Independent-Family set was not used to train the VQSR Gaussian mixture models that assigned VQSLOD scores to the WGS variants. Variants not recovered in the WGS Raw set can either be false positives in the IF set or false negatives in the Raw set. False negatives in the Raw set can occur if the variant did not have sufficient coverage in the WGS data. False positives in the IF set can occur if, in the reduced representation data, a true variant (e.g., an indel) caused errors in read mapping that produced an artifactual variant (e.g., a SNP); such an artifactual variant will segregate with the true variant and appear to be genetically mappable. While procedures like indel realignment should resolve these cases, the way reads stack and the high depth of some loci acheived with reduced representation methods can prevent accurate local reassembly. These data show that most of the variants from the reduced representation IF data are identified in the WGS data and that sensitivity decreases with descending tranches.

	# SNPs	% SNP	# indels	% indel
Independent-Family (IF)	10,737	100%	3,740	100%
IF ∩ Raw	10,557	98%	3,632	97%
IF ∩ Sensitive	10,211	95%	3,402	91%
IF ∩ Specific	7,966	74%	2,330	62%

[1] Reprinted from "RIG: recalibration and interrelation of genomic sequence data with the GATK" by McCormick, RF, Truong, SK, and Mullet, JE, (2015) Genes — Genomes — Genetics 5(4) 655-665 under the Creative Commons Attribution Unported License (`http://creativecommons.org/licenses/by/3.0`), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Copyright ©2015 McCormick *et al.*

Table A.2: Comparison of the Independent-Family set with WGS tranches. The intersections of variants from the Independent-Family (IF) set with each of the WGS variant sets were compared (see Table A.2). Assuming that the IF variants represent "true" variants, the tranche cutoffs are in good agreement with how many of the IF variants were present in the tranche (even though the IF variants were not used to train the VQSR Gaussian mixture models). For example, the 95% SNP tranche represents the minimum VQSLOD cutoff whereby 95% of the "true" variants provided to VQSR would be retained. Accordingly in our data, the 95% WGS SNP tranche contains 97% of the available IF set SNPs, suggesting that the models were appropriately trained and that the tranche cutoffs functioned as expected.

	# SNPs	**% SNPS**	**SNP tranche**	**min(VQSLOD of tranche)**
IF ∩ Raw	10,557	100%	100%	-39,962.6819
IF ∩ Sensitive	10,211	97%	95%	0.4462
IF ∩ Specific	7,966	75%	75%	7.1643
	# indels	**% indels**	**indel tranche**	**min(VQSLOD of tranche)**
IF ∩ Raw	3,632	100%	100%	-39,645.5822
IF ∩ Sensitive	3,402	94%	95%	1.1027
IF ∩ Specific	2,330	64%	75%	4.6878

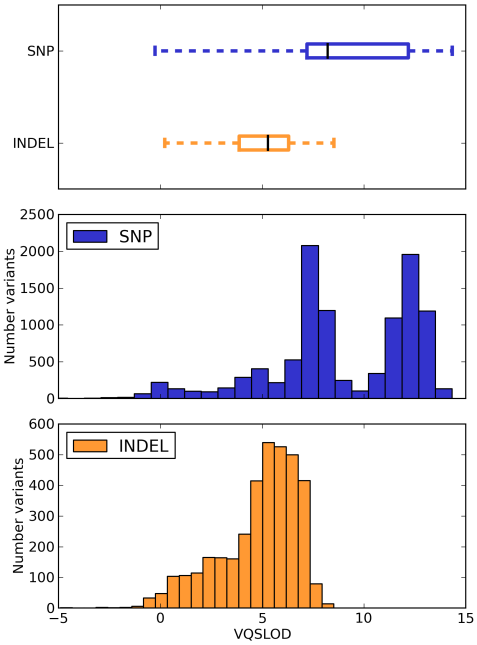

Figure A.1: Distributions of VQSLOD scores for variants from the WGS Raw set that were also contained in the Independent-Family (IF) set. The VQSLOD distributions of the 10,557 SNPs and 3,632 indels from the WGS raw set that were also in the IF set are plotted here as box plots and as histograms (see Tables A.1 and A.2). The median VQSLOD score of the SNPs and indels were 8.22 and 5.29, respectively, suggesting that the trained Gaussian mixture models correctly assigned true variants with positive VQSLOD scores. Variants from the IF set with low VQSLOD scores (e.g. < 0) potentially represent the false positives described in the caption of Table A.1 that were also called in the WGS data. Alternatively, they are true variants that did not receive sufficient coverage in the WGS data to provide strong evidence for their existence. The two peaks of the bimodal distribution of SNP VQSLOD scores correspond to whether or not certain variant annotations had been calculated by the GATK's HaplotypeCaller. Certain variant annotations, such as MQRankSum and ReadPosRankSum, are only calculated when a sample contains a mixture of reads displaying both the reference allele and the alternate allele for the variant; these annotations were typically not assigned to variants for which every sample was genotyped as homozygous. Both MQRankSum and ReadPosRank sum were used as annotations for training during VQSR; the lower VQSLOD peak consists mostly of variants assigned these annotations, and the larger VQSLOD peak consists mostly of variants that were not assigned these annotations. This suggests that these two annotations were often associated with less reliable variants in the resequenced sorghum lines which is expected given the inbred nature of most of the lines. A similar effect was seen with indels, though not as extreme.

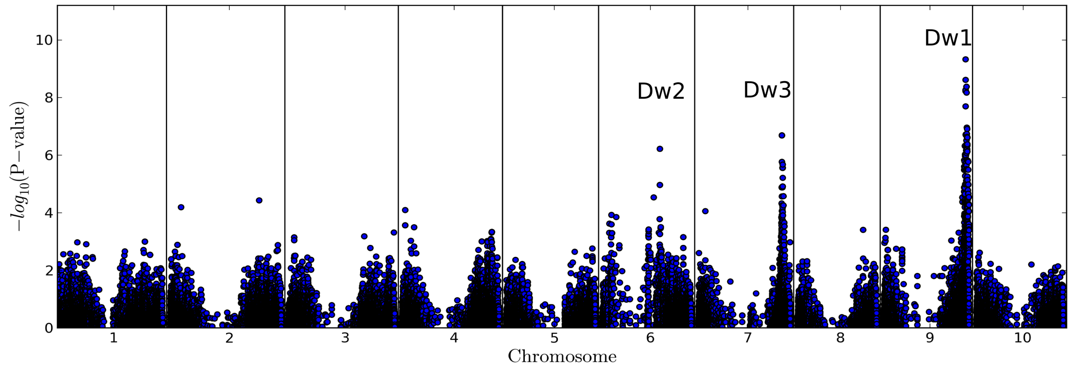

Figure A.2: Genome-wide associations for preflag leaf height using RIG-generated variants called from reduced representation data. Of the 733 sorghum germplasm samples used to generate the Population Reference Variant Resource as part of the RIG workflow, 171 of the lines had been previously phenotyped by (Brown *et al.*, 2008). After producing a recalibrated, sensitive variant resource with the RIG workflow, missing genotypes were filled in using Beagle v4 release 1274 (Browning and Browning, 2007). Variants were pre-processed (minor allele frequency > 5%) and converted to PLINK binary format using PLINK v1.90-1 (Purcell *et al.*, 2007; Chang *et al.*, 2015). The 171 phenotypes from (Brown *et al.*, 2008) were normalized using an Empirical Normal Quantile Transformation (ENQT) (Peng *et al.*, 2007). Using GCTA v1.24.3, a genomic relationship matrix was generated and associations were calculated using GCTA's mixed linear model implementation (Yang *et al.*, 2011). As shown in Table A.3, this analysis reproduced known QTL at the sorghum dwarfing loci *Dw1*, *Dw2*, and *Dw3* on chromosomes 9, 6, and 7, respectively (Morris *et al.*, 2013; Higgins *et al.*, 2014).

Table A.3: Comparison GWAS results from RIG-generated variants to previously reported results. The RIG column lists the position of the most significant marker identified by the GWAS described in Figure A.2. The Literature column lists the position of significant peaks reported by (Morris *et al.*, 2013) for Dw1 and Dw2 and the position of the cloned gene for *Dw3* (Multani *et al.*, 2003). Recalibrated variants identified from reduced representation sequence data using the RIG workflow are capable of reproducing known sorghum genome wide associations.

Locus	**Chromosome**	**RIG (Mbp)**	**Literature (Mbp)**
Dw2	6	40.2	39.7 - 42.6
Dw3	7	58.4	58.6
Dw1	9	57.2	57.2

Table A.4: Variant site counts used to calculate sensitivity and positive predictive value for each tranche. Subsets of each of the six tranches (75.0%, 95.0%, 97.5%, 99.0%, 99.9%, and 100.0%) were used for determining sensitivity and positive predictive value. Sensitivity was calculated using $\frac{(Tranche \cap Nordborg)}{Nordborg}$. Positive predictive value was calculated using $\frac{(Tranche \cap Nordborg)+((Tranche \setminus Nordborg) \cap Gramene43)}{Tranche}$. For example, the sensitivity of the 75.0% tranche is $\frac{1762}{3243} = 0.543$ and the positive predictive value is $\frac{1762+20}{1789} = 0.996$

Variant Source	**Number Variant Sites**
Nordborg 2005	3243
75.0%	1789
75.0% ∩ Nordborg 2005	1762
(75.0% \ Nordborg 2005) ∩ Gramene43	20
75.0% \ (Nordborg 2005 ∪ Gramene43)	7
95.0%	3014
95.0% ∩ Nordborg 2005	2897
(95.0% \ Nordborg 2005) ∩ Gramene43	98
95.0% \ (Nordborg 2005 ∪ Gramene43)	19
97.5%	3107
97.5% ∩ Nordborg 2005	2982
(97.5% \ Nordborg 2005) ∩ Gramene43	103
97.5% \ (Nordborg 2005 ∪ Gramene43)	22
99.0%	3212
99.0% ∩ Nordborg 2005	3078
(99.0% \ Nordborg 2005) ∩ Gramene43	109
99.0% \ (Nordborg 2005 ∪ Gramene43)	25
99.9%	3589
99.9% ∩ Nordborg 2005	3220
(99.9% \ Nordborg 2005) ∩ Gramene43	205
99.9% \ (Nordborg 2005 ∪ Gramene43)	164
100.0%	3716
100.0% ∩ Nordborg 2005	3241
(100.0% \ Nordborg 2005) ∩ Gramene43	240
100.0% \ (Nordborg 2005 ∪ Gramene43)	235

APPENDIX B

SUPPLEMENTARY MATERIAL FOR 3D SORGHUM RECONSTRUCTIONS FROM DEPTH IMAGES IDENTIFY QTL REGULATING SHOOT ARCHITECTURE [1]

[1] 1Reprinted with permission from McCormick, R. F., S. K. Truong, and J. E. Mullet, 2016, 3D sorghum reconstructions from depth images identify QTL regulating shoot architecture. Plant Physiology 172: 823-834 (www.plantphysiol.org). Copyright ©2016 American Society of Plant Biologists.

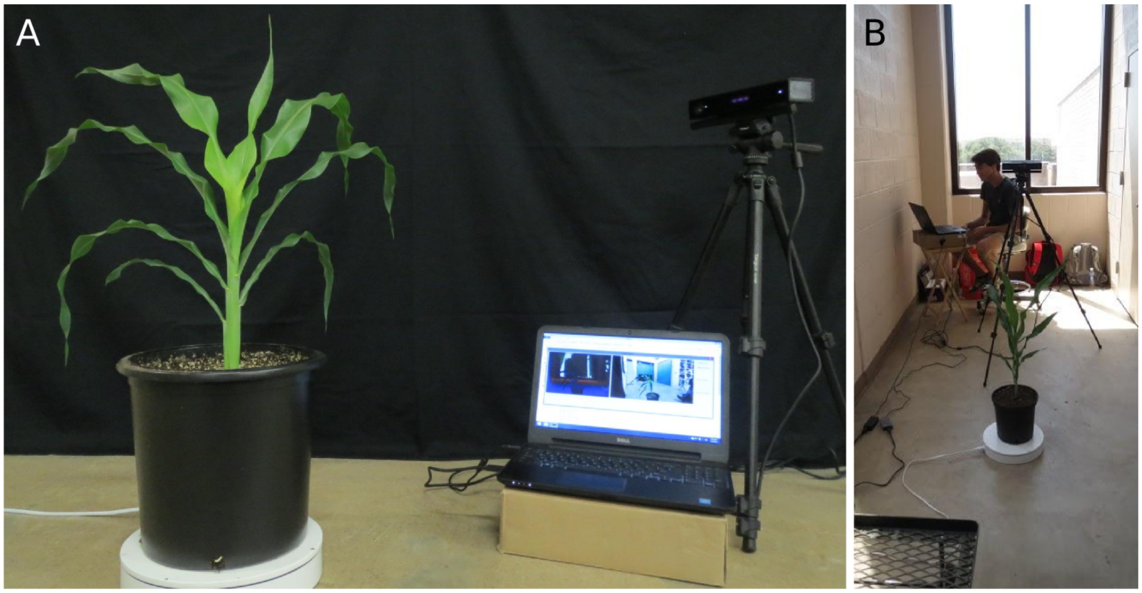

Figure B.1: Imaging platform. (A) Components of the imaging platform. The major components include the turntable, a plant, the depth sensor, and the laptop running the acquisition software. (B) The imaging platform in use. The imaged plant spins on the turntable during acquisition, and a user operates the semi-automated image acquisition software.

B.1 Image processing methods, potential alternatives, and future development

After acquiring depth images from multiple perspectives around the plant, the point clouds stored in the images were cleaned, registered, meshed, and segmented, and the resulting segmented meshes were automatically measured for a number of traits. The procedures used are detailed here, along with discussion of why particular methods were used, and discussion of some alternative approaches. The point cloud data from the Microsoft Kinect for Windows V2 were stored as depth images and processed using the open source libraries OpenCV (`http://opencv.org`; accessed February 2016) and PCL (Rusu and Cousins, 2011). For each plant, the 12 individual point clouds were registered to the same reference frame using iterative closest point

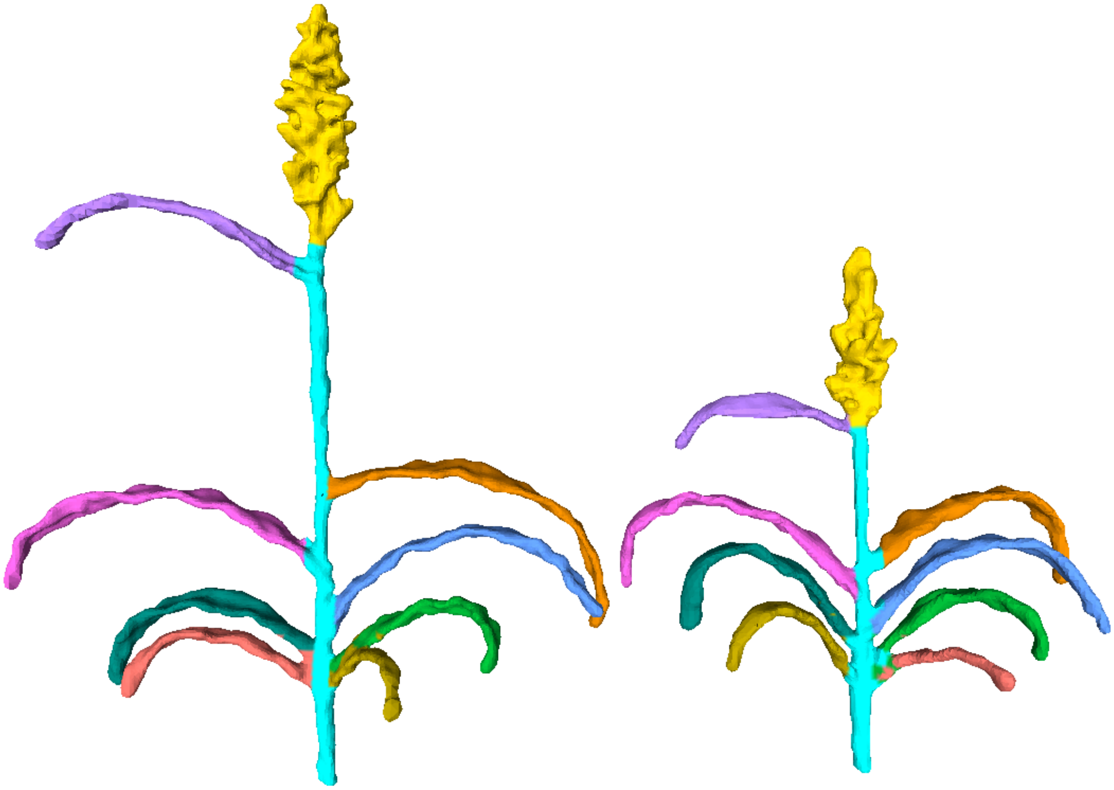

Figure B.2: Plants with inflorescences. RIL 182 (left) and RIL 374 (right) at 44 DAP. Inflorescences are colored gold. Meshes are depicted at the same relative scale. All 1200 segmented meshes are available (see Section 3.6.4).

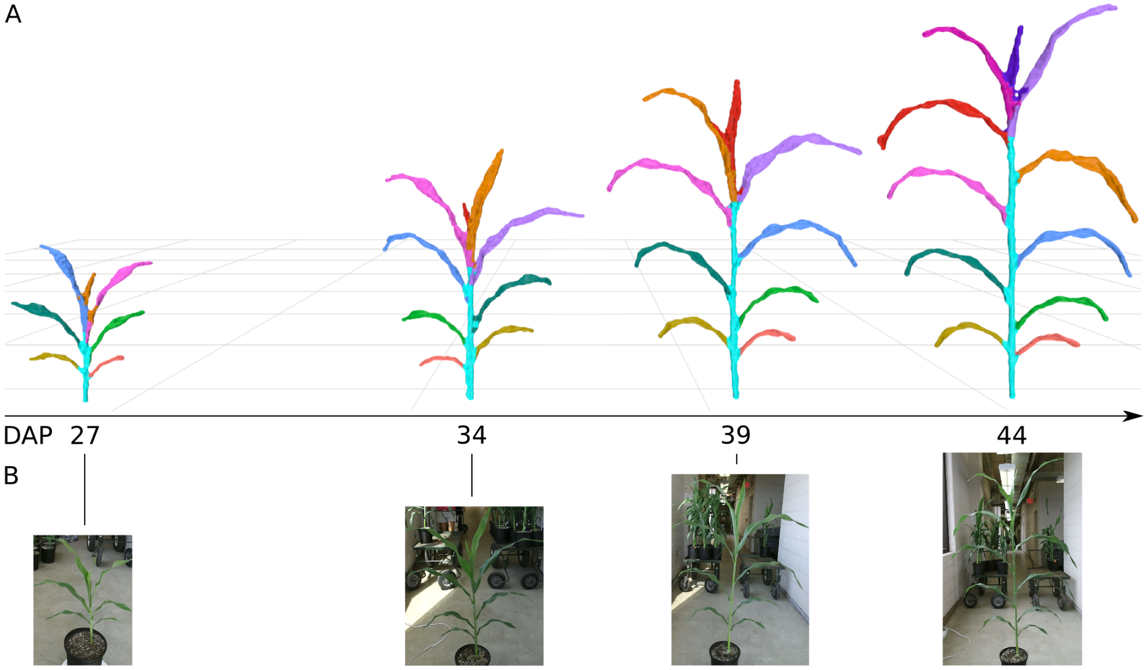

Figure B.3: Plant growth over time. (A) Segmented meshes for replicate 3 of RIL 175 are depicted at 4 different days after planting (DAP) timepoints. Leaf colors represent individual segmented leaves and are developmentally ordered for each individual mesh; the same leaf color between two meshes do not necessarily correspond to the same leaf. The shoot cylinder is colored cyan. Meshes are depicted at the same relative scale. (B) Corresponding RGB images that were co-acquired with the depth images. RGB images are not to scale.

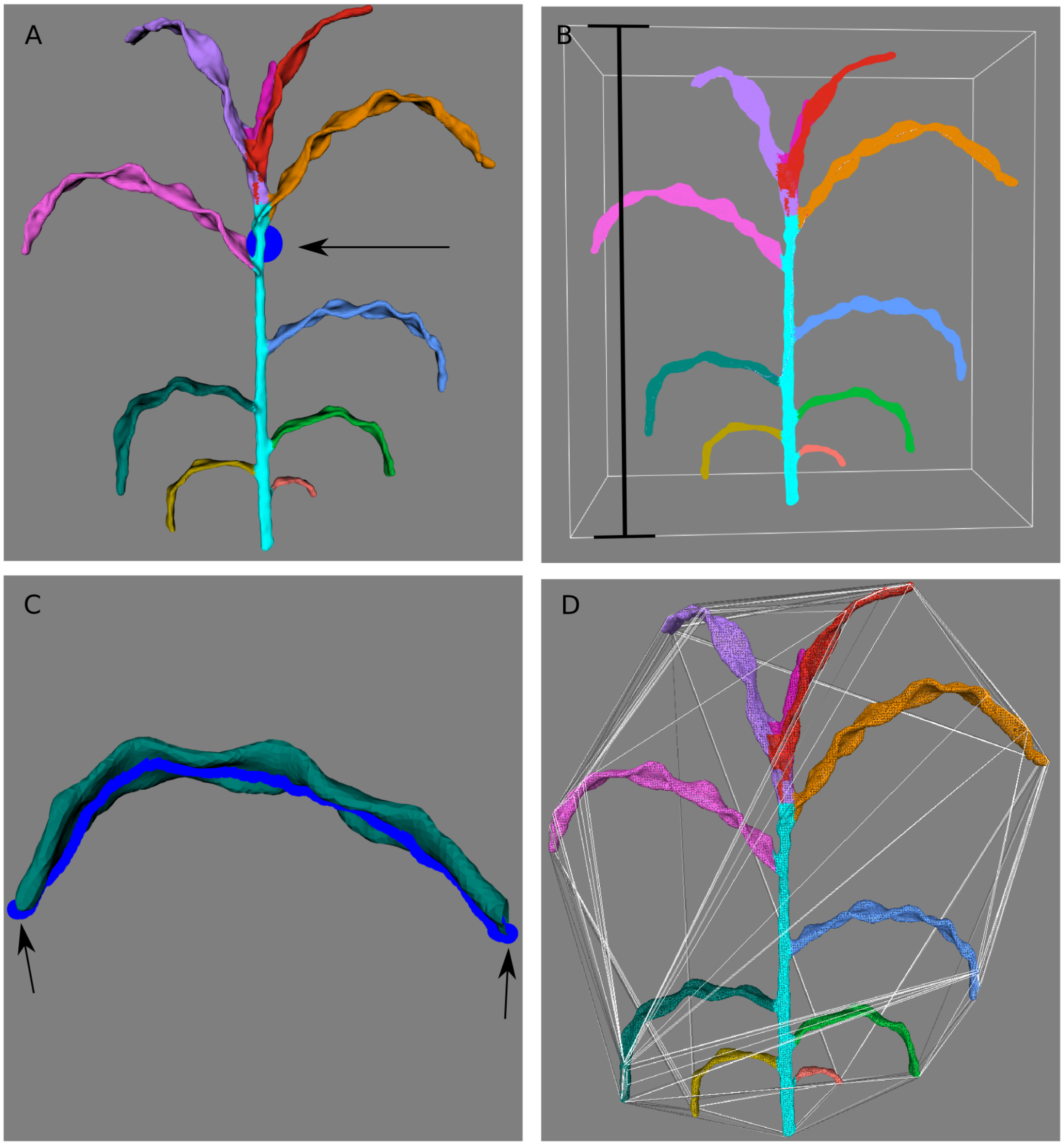

Figure B.4: Visual depiction of selected measurements. (A) The shoot center of mass (i.e. the mesh centroid) is depicted in blue. (B) Shoot height corresponds to the vertical length of the axis-aligned bounding box of the mesh. Shoot cylinder height corresponds to the same measurement, but for the axis-aligned bounding box of the shoot cylinder (cyan). (C) The path corresponding to leaf length is depicted in blue with the two end points indicated by arrows; leaf length is measured as the length of the longest graph geodesic of the leaf mesh. (D) Shoot compactness was measured as the surface area of the convex hull; the convex hull is shown around the plant mesh. The mesh shown corresponds to RIL 268 at 34 DAP. Many of the trait measurements are correlated, particularly composite traits with organ-level traits (Figure B.5).

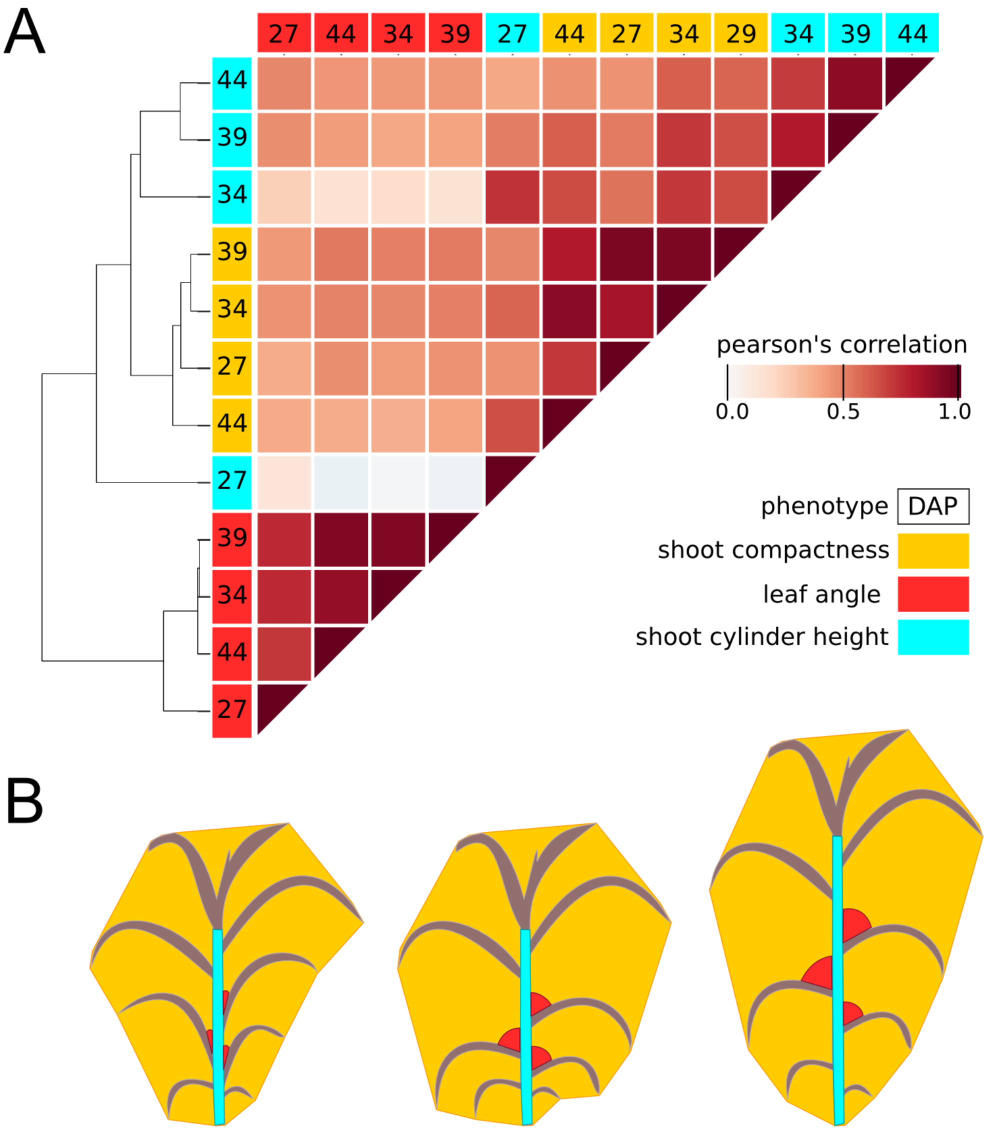

Figure B.5: Composite traits integrate multiple architecture traits. Organ-level traits are components of composite traits. The relationship between shoot compactness, leaf angle, and shoot cylinder height are depicted here. (A) The Pearson product-moment correlation coefficient matrix for three traits across four timepoints: the organ-level traits of leaf angle and shoot cylinder height, and the composite trait of shoot compactness. Shoot compactness is measured as the surface area of the smallest polyhedron that contains the 3D plant mesh (i.e. the convex hull surface area), and it is highly correlated with both of these traits at all timepoints. Leaf angle is only highly correlated with shoot cylinder height once *Dw3* begins to influence shoot cylinder height (i.e., after the first timepoint). (B) Graphical, two-dimensional representation of how leaf angle and shoot cylinder height influence convex hull area (depicted as a polygon).

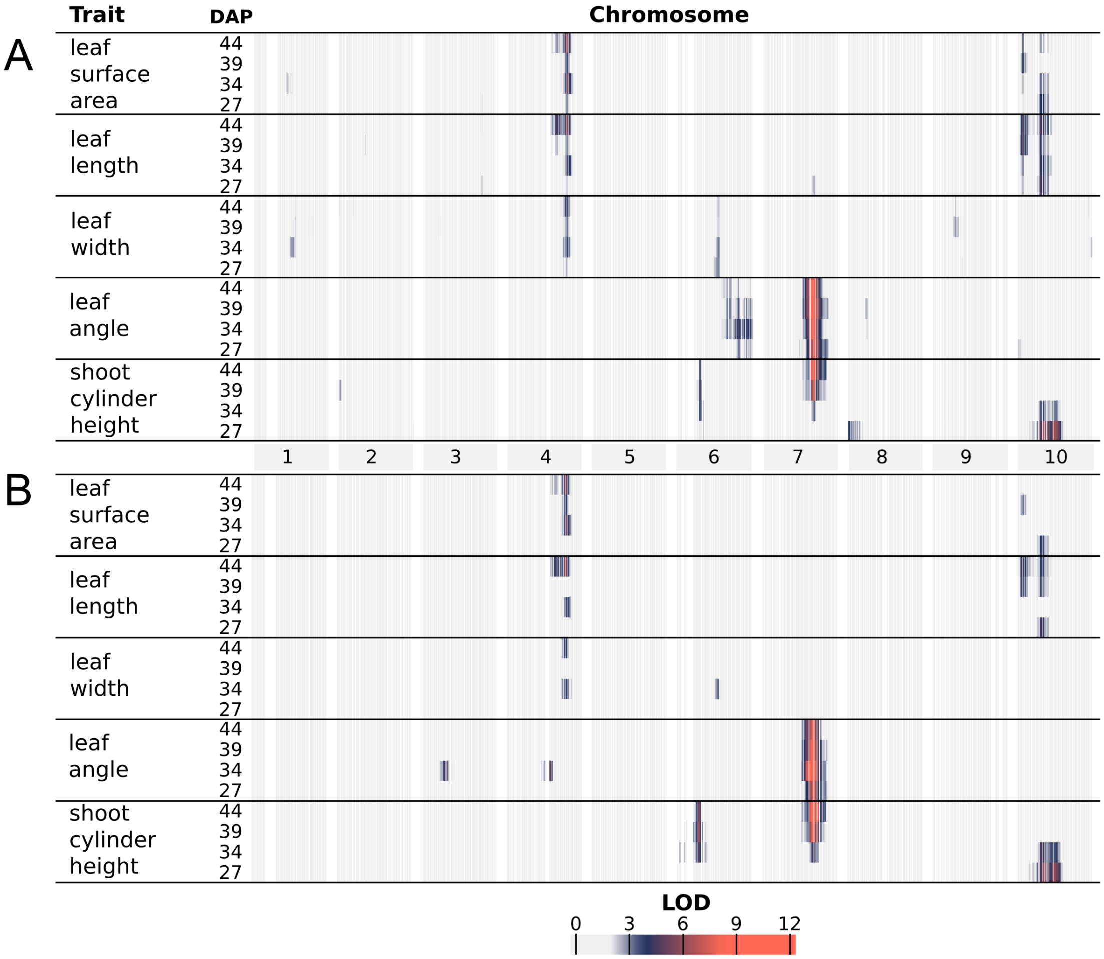

Figure B.6: QTL mapping steps for organ-level traits leading to the final LOD profiles shown in Figure 3.4. (A) LOD profiles for a genome-wide scan under a single-QTL model. (B) LOD profiles for chromosome-wide scans of chromosomes with QTL based on the most likely multiple-QTL model found by model selection for each phenotype by timepoint combination. Each row represents a different trait, and within each trait are four nested rows that each represents a different timepoint (days after planting; DAP). Each group of columns represents a chromosome, and each column represents a marker at its genetic position. Cells are colored by marker LOD for the phenotype at the particular timepoint. Panel B differs from Figure 3.4 in that a comprehensive QTL model that includes all of the QTL found via model selection for a given phenotype (across all timepoints) was used for the chromosome-wide scans for each timepoint in Figure 3.4.

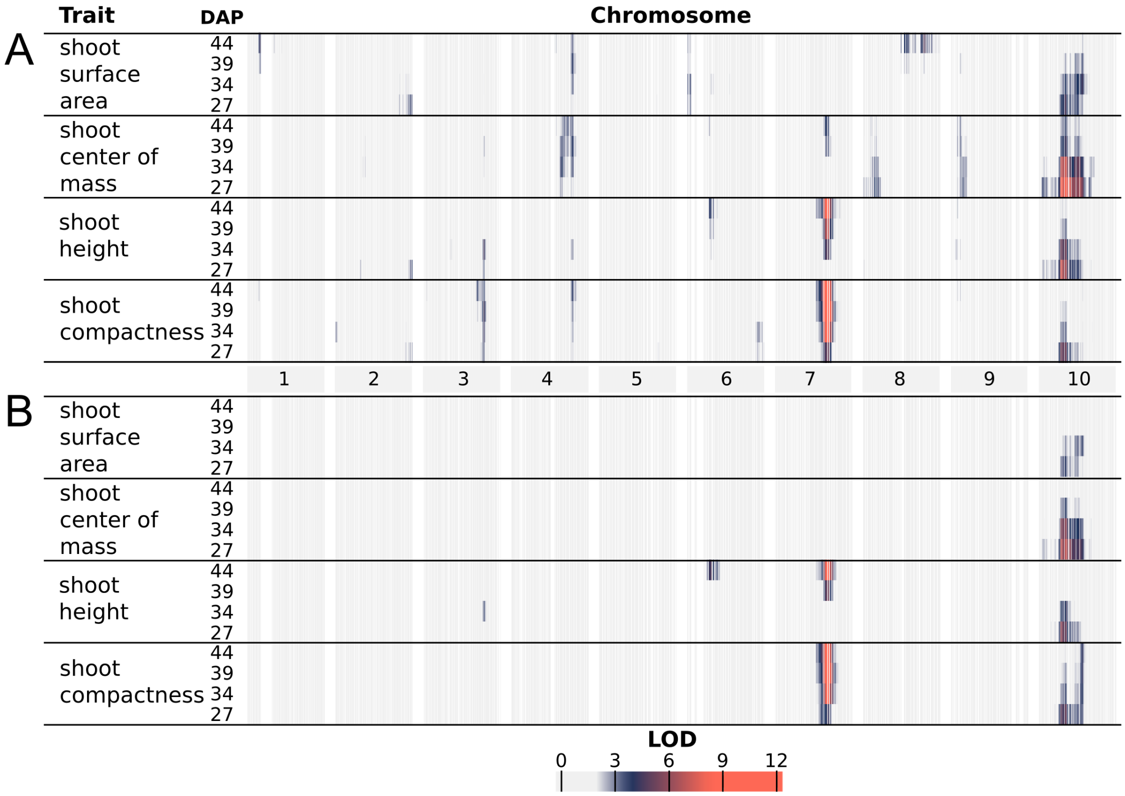

Figure B.7: QTL mapping steps of composite traits leading to the final LOD profiles shown in Figure 3.6. (A) LOD profiles for a genome-wide scan under a single-QTL model. (B) LOD profiles for chromosome-wide scans of chromosomes with QTL based on the most likely multiple-QTL model found by model selection for each phenotype by timepoint combination. Each row represents a different trait, and within each trait are four nested rows that each represents a different timepoint (days after planting; DAP). Each group of columns represents a chromosome, and each column represents a marker at its genetic position. Cells are colored by marker LOD for the phenotype at the particular timepoint. Panel B differs from Figure 3.6 in that a comprehensive QTL model that includes all of the QTL found via model selection for a given phenotype (across all timepoints) was used for the chromosome-wide scans for each timepoint in Figure 3.6.

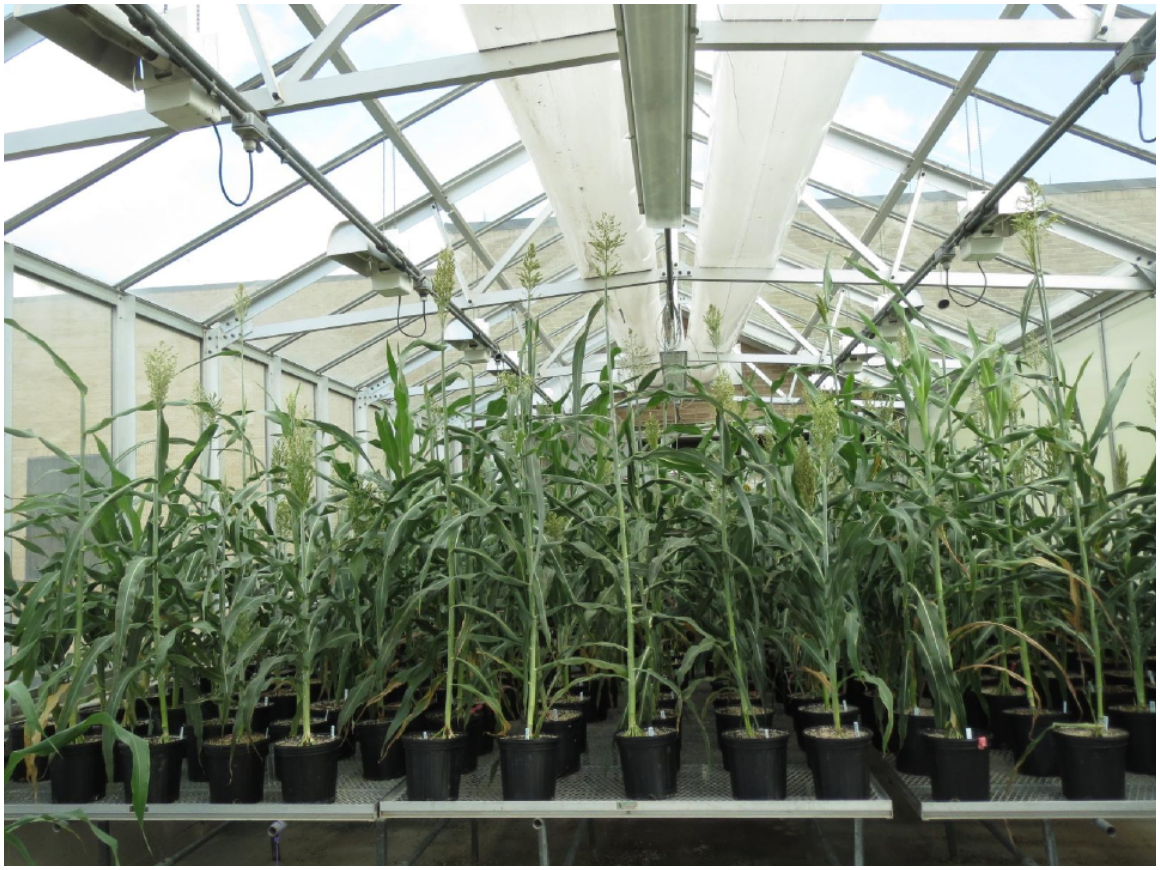

Figure B.8: BTx623 x IS3620C RIL population in greenhouse. 99 members of the BTx623 x IS3620C sorghum mapping population were planted in triplicate (97 RILs and 2 parental lines; 99 plants per table, three tables, replicate 1 on left, 2 in middle, and 3 on right). Phenotype data for each individual in each replicate are available (see Code and Data Availability in the main text). All QTL analyses reported in the main text used the average value of the three replicates for a RIL.

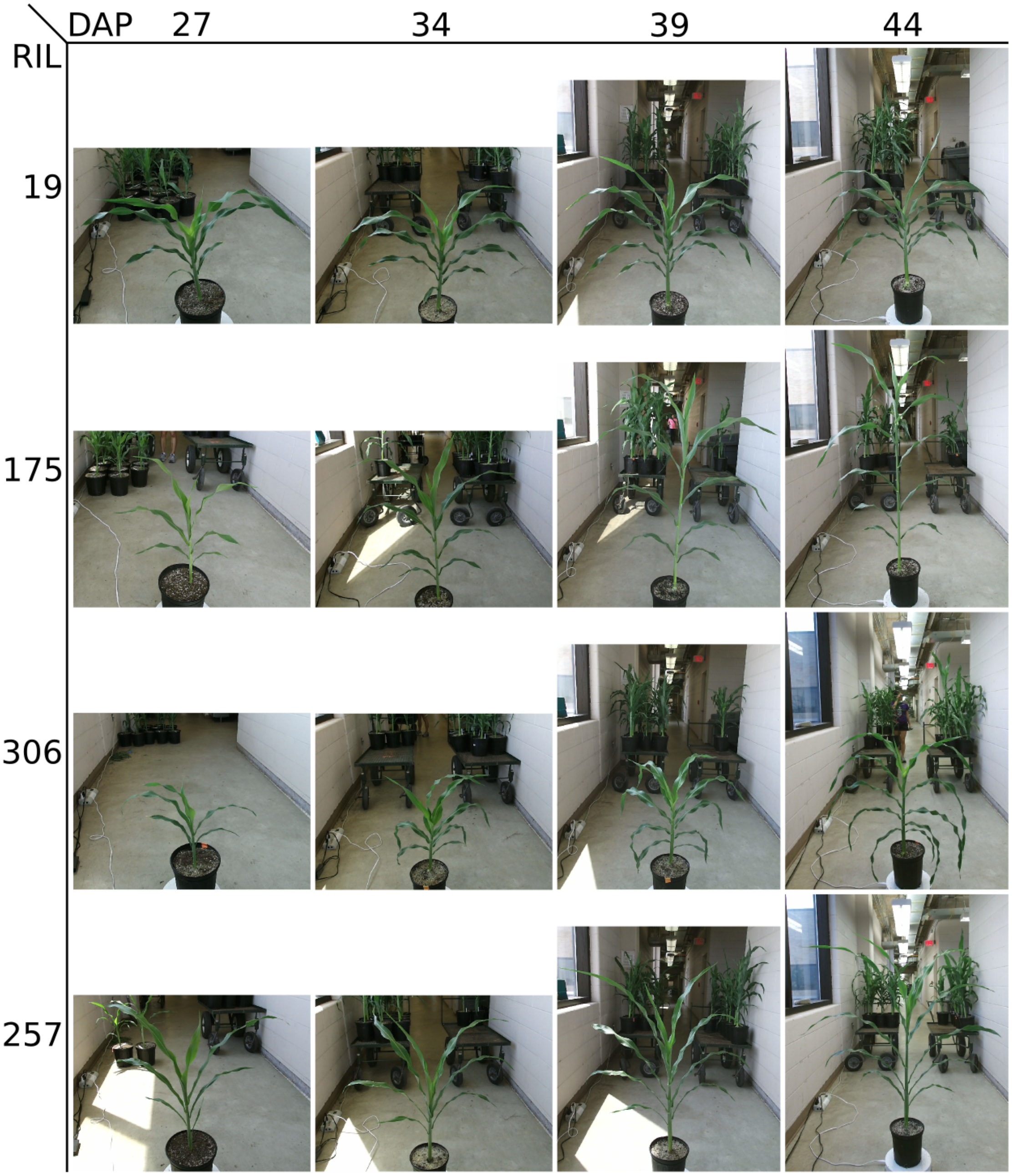

Figure B.9: Plants from the BTx623 x IS3620C RIL population display variation in shoot morphology. The images depict the plants represented by the meshes in Figures 3.5 and 3.7 in the main text.

Table B.1: QTL intervals by phenotype for organ-level traits. The LOD-2 intervals of QTL in the multiple-QTL model obtained for each phenotype by timepoint combination. The maximum LOD (MLOD) of an interval is indicated by a * and was used for the multiple-QTL model of the phenotype for chromosome-wide scans (Figure 3.4).

	chr	**interval begin (Mbp)**	**peak coordinate (Mbp)**	**interval end (Mbp)**	**peak LOD**
leaf surface area					
DAP 27	10	1.77	5.72	9.86	3.41*
DAP 34	4	62.22	62.91	63.91	4.89
DAP 39	4	60.89	62.45	63.85	3.46
	10	1.15	1.87	2.67	3.28
DAP 44	4	61.40	62.45	63.52	5.04*
leaf length					
DAP 27	10	5.13	5.72	8.60	4.05
DAP 34	4	62.07	62.91	64.09	3.90
DAP 39	10	1.23	2.00	8.21	4.15
DAP 44	4	57.48	62.45	63.40	5.56*
	10	1.23	2.00	8.21	4.55*
leaf width					
DAP 27	-	-	-	-	-
DAP 34	4	60.89	62.60	64.43	3.84*
	6	48.45	50.97	55.08	3.42*
DAP 39	-	-	-	-	-
DAP 44	4	60.89	62.83	63.75	3.76
leaf angle					
DAP 27	7	58.48	59.87	60.28	6.89
DAP 34	3	7.11	9.61	11.46	4.13*
	4	51.93	55.31	56.73	4.56*
	7	59.51	59.65	59.99	10.74*
DAP 39	7	59.48	59.65	59.99	9.76
DAP 44	7	59.20	59.65	59.99	8.14
shoot cylinder height					
DAP 27	10	5.27	7.46	52.24	6.03*
DAP 34	6	0.25	42.67	46.02	4.21
	7	58.48	59.63	61.10	3.98
	10	5.27	7.46	52.52	4.16
DAP 39	6	40.10	42.77	44.83	5.60*
	7	59.05	59.85	59.99	7.90
DAP 44	6	40.10	43.30	44.83	4.88
	7	59.20	59.65	59.99	9.81*

Table B.2: QTL intervals by phenotype for composite traits. The LOD-2 intervals of QTL in the multiple-QTL model obtained for each phenotype by timepoint combination. The maximum LOD (MLOD) of an interval is indicated by a * and was used for the multiple-QTL model of the phenotype for chromosome-wide scans (Figure 3.6).

	chr	**interval begin (Mbp)**	**peak coordinate (Mbp)**	**interval end (Mbp)**	**peak LOD**
shoot surface area					
DAP 27	10	5.06	6.97	52.52	3.55
DAP 34	10	5.55	51.93	52.82	3.82*
DAP 39	-	-	-	-	-
DAP 44	-	-	-	-	-
shoot center of mass					
DAP 27	10	5.40	7.48	49.62	6.60
DAP 34	10	5.27	7.46	8.21	6.72*
DAP 39	10	5.27	7.46	48.94	5.68
DAP 44	-	-	-	-	-
shoot height					
DAP 27	10	5.27	5.63	7.59	5.87*
DAP 34	3	65.26	66.41	69.08	3.63*
	10	5.27	7.46	8.60	4.11
DAP 39	7	58.85	59.53	60.95	4.86
DAP 44	6	40.10	44.37	47.42	4.20*
	7	59.05	59.65	59.99	8.73*
shoot compactness					
DAP 27	7	57.29	59.65	60.47	4.02
	10	5.06	5.63	52.52	4.86*
DAP 34	7	59.20	59.63	59.99	8.72
	10	5.27	51.93	52.82	3.89
DAP 39	7	59.20	59.63	59.99	10.87*
	10	5.55	52.24	52.82	3.95
DAP 44	7	59.05	59.53	59.99	9.67
	10	6.54	52.24	52.82	3.95

(ICP) and prerejective random sample consensus (RANSAC) if ICP failed to provide a good fit (Besl and McKay, 1992; Buch *et al.*, 2013). Points corresponding to the plant pot were removed using RANSAC to fit circles, planes, and cylinders to identify points belonging to the pot (Fischler and Bolles, 1981). After pot removal, the registration was refined, and the combined point clouds were cleaned by removing outlier points and using region growing to retain points corresponding to the plant (Rabbani *et al.*, 2006; Rusu *et al.*, 2008). These procedures were automated using classes from the OpenCV and PCL libraries.

The combined clouds were then visually examined, and artifacts arising from sensor noise in acquisition, non-rigid plant transforms caused by airflow or leaf shaking during acquisition, and errors in registration were manually corrected using MeshLab (Cignoni *et al.*, 2008). Point clouds from progressively later timepoints required progressively more manual cleaning, likely due either to increased leaf shaking caused by the turntable as the plants grew larger, or increased sensor noise as the plants increased in distance from the sensor; as such, the final two timepoints (DAP 48 and 55) were not processed and not included in the analysis. Once a point cloud was prepared, ball-pivoting and Poisson-disk sampling were performed (automated with MeshLab server) to create a point cloud with oriented abaxial and adaxial leaf point normals, and the point cloud was then meshed using Screened Poisson Surface Reconstruction (Bernardini *et al.*, 1999; Corsini *et al.*, 2012; Kazhdan and Hoppe, 2013).

Points in the mesh corresponding to the shoot cylinder of the plant (composed of leaf sheaths and stem) and the inflorescence (when present) were then segmented using a machine learning approach. A subset of the meshes were manually labeled and fast point feature histograms were calculated for all points in all meshes (Rusu *et al.*, 2009a,b). Features from labeled meshes were used to train a multi-class classifier using AdaBoost.MH with real valued decision stumps as implemented in MultiBoost, and the classifier was then used to label the remaining meshes (Schapire and Singer, 1999; Benbouzid *et al.*, 2012). While this worked for many meshes, some meshes had unsatisfactory labeling, and we opted to manually label all shoot cylinder and inflorescence points for the final reported analysis. Point labels were assigned as specific colors in the mesh, with cyan and gold corresponding to the shoot cylinder and inflorescence, respectively.

Following this, individual leaves were segmented using an automated procedure. The vertices in the mesh were first clustered into supervoxels, and supervoxel adjacency was determined (Papon *et al.*, 2013). Geodesic lengths (calculated via Dijkstras algorithm) across the supervoxel adjacency graph were used to iteratively label unlabeled supervoxels (Dijkstra, 1959; Surazhsky *et al.*, 2005). The lowest shoot cylinder-labeled supervoxel was used as the starting point of a geodesic path, and the most distant unlabeled supervoxel was identified. If the supervoxel had a sufficiently large geodesic length and euclidean distance from any other leaf-labeled supervoxel, the supervoxel was considered a new leaf tip and labeled as such; otherwise, the supervoxel was labeled based on the label of adjacent supervoxels (all points comprising the supervoxel were similarly labeled). Once all leaf tips had been identified, leaf segmentations were refined by finding paths between the shoot cylinder and the leaf tips in the developmental order that the leaves had emerged from the shoot cylinder such that developmentally older leaves were prioritized to improve segmentation of leaves found in the whorl.

Multiple measurements were automatically obtained from each mesh, and the entirety of the measurements made and the methods for making them can be found within the code base. The manuscript reports measurements of shoot height, shoot cylinder height, shoot center of mass, shoot compactness, shoot surface area, leaf length, leaf width, leaf surface area, and leaf angle. Brief descriptions of how these traits were measured are provided below, and graphical depictions of selected measurements are shown in Figure B.4 and Figure B.5.

Shoot height was measured as the height of the axis-aligned bounding box of the entire plant mesh after the first principal component of the shoot cylinder was aligned to the axis representing height. Shoot cylinder height was measured in the same manner as shoot height, but only for mesh vertices labeled as shoot cylinder (i.e. colored cyan). The shoot center of mass is the average height of the vertices in the mesh (i.e. the height of the centroid). Shoot compactness was measured as the surface area of the convex hull of the plant mesh using PCL's interface with libqhull (Barber *et al.*, 1996). Shoot surface area was measured as the summation of the area of individual polygons comprising the mesh.

Leaf measurements were made by first finding the largest connected mesh of vertices with the same leaf label; that is, if during segmentation not all vertices labeled as the particular leaf were connected, the largest connected mesh was retained as the leaf mesh. Leaf length was measured as the length of the longest of all shortest paths between two vertices in the mesh (i.e. the longest graph geodesic in the graph formed by the mesh edges and vertices). Leaf width was approximated by modeling the leaf as a box of known length (the maximum geodesic length), and surface area (the summation of polygon area), and height (leaf thickness, fixed as 1 cm for all leaves). Leaf angle was measured starting at the leaf vertex with the minimum geodesic length to the bottom of the shoot cylinder (this leaf vertex is referred to as the leaf base), and the path with the maximum geodesic length was found for the leaf mesh starting from the leaf base. A right triangle was formed by three points: (a) point α, the leaf base, (b) point β, reached by traveling along the path of α's maximum geodesic from α to 76 mm along the geodesics path, and (c) point γ, reached by moving up the vertical axis by the vertical distance between α and β. Leaf angle was calculated as the angle between (α, β) and (α, γ).

A variety of additional methods were considered for registration, meshing, segmentation, and measurement, but they either did not work consistently with the dataset or were outside the scope of work to implement. A few of these alternatives are discussed here.

Registration of the point clouds could potentially be improved by using a non-rigid registration approach (Zheng *et al.*, 2010; Bucksch and Khoshelham, 2013; Brophy *et al.*, 2015). Some non-rigid registration approaches depend on skeletonization, and we attempted mesh skeletonization using thinning and potential field approaches described by Cornea *et al.* (2007). While this worked for some meshes, performance was not sufficiently consistent to merit adoption for a non-rigid registration approach. Future development of the platform will consider additional testing of skeletonization and non-rigid registration approaches.

The meshing procedure employed combines two common meshing algorithms, ball-pivoting and Screened Poisson Surface Reconstruction (Bernardini *et al.*, 1999; Kazhdan and Hoppe, 2013); the use of Screened Poisson Surface Reconstruction is potentially suboptimal since it assumes a water tight surface, and thin, flat leaves are reconstructed as thicker than they otherwise should be. For our application, we were predominately interested in approaches that would reconstruct a fully connected and consistent leaf surface so that we could compare leaf measurements across different genotypes. In our hands, meshing approaches like ball-pivoting (that do not impose watertightness) inconsistently meshed leaf surfaces due to the nature of the point cloud thickness at leaves; in the resulting mesh, some portions of the leaf would have faces with normals corresponding to both an abaxial and adaxial leaf surface, and some portions with only one or the other (which precluded consistent measurements across plants). This was mitigated to some extent by applying smoothing operations on the point cloud, but it still failed to account for the fact that leaves often have ripples, tears, curls, and folding that caused treatment as a 2D surface to not consistently work well given the Kinect's resolution.

After testing multiple approaches, the final approach uses ball-pivoting (which mostly captures the abaxial and adaxial leaf surfaces and correctly orients point normals for those surfaces), followed by Poisson-disk sampling to sample the ball-pivoting mesh to points, followed by Screened Poissson Surface Reconstruction of those points; this provided the most consistent results across the thousand meshes (whereas using one meshing approach alone did not). Future development will consider alternative meshing approaches that are potentially better suited to the generally thin, flat nature of leaves.

Automated stem segmentation was attempted in a number of ways, though none had sufficiently satisfactory performance on the entire dataset. The first attempt used region growing as described in Rabbani *et al.* (2006); this worked well on some meshes, but the transition between stems and leaves, and individual leaves, were often too smooth to obtain consistent segmentation. Additionally, RANSAC was tested for use in identifying the shoot cylinder, though this also often performed poorly in distinguishing the whorl from the shoot cylinder. Lastly, we tested a machine learning approach described above using a multi-class classifier and point features to distinguish shoot cylinder from leaf points. Of the three methods, this had the best performance, though still insufficient for use in downstream measurements. As such, we opted to manually label stems and segment leaves automatically given information on the stem. Efficient segmentation of the mesh into individual organs remains an outstanding issue for rapidly making accurate organ-level measurements.

Future development will also consider improvements to measurement techniques, particularly for leaf morphology. Implementing an approach that models the leaf as a B-spline will better describe traits including leaf angle, leaf pitch, and leaf curvature, and potentially improve the ability to formally describe plant architecture obtained from images as an L-system, such as with OpenAlea and L-py, for functional-structural plant modeling (Pradal *et al.*, 2008; Boudon *et al.*, 2012).