To my parents
ACKNOWLEDGMENTS

Thank you, to my parents for the unconditional support. Thank you, to Dr. Melanie J. Correll for taking a chance on me and giving me this opportunity. Thank you, to Dr. Vallejos, Dr. Kiker, and Dr. Gezan for their guidance as committee members. Thank you, to Dr. Jones, Dr. Boote, Dr. Clavijo Michelangeli, Dr. Zhang, Dr. Bhakta, Mrs. Porter, Dr. Rao, and Mr. Ricaurte who have advised and inspired me through this journey.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>7</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>11</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 GENERAL INTRODUCTION</td>
<td>13</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>13</td>
</tr>
<tr>
<td>Crop Models</td>
<td>14</td>
</tr>
<tr>
<td>Genotyping</td>
<td>15</td>
</tr>
<tr>
<td>Phenotyping</td>
<td>17</td>
</tr>
<tr>
<td>Computational tools for identifying the G, G x E interactions</td>
<td>18</td>
</tr>
<tr>
<td>Next-Gen Crop Models; Gene-based Models</td>
<td>19</td>
</tr>
<tr>
<td>The Common Bean</td>
<td>21</td>
</tr>
<tr>
<td>Objectives and Thesis Organization</td>
<td>21</td>
</tr>
<tr>
<td>2 A DYNAMIC GENE-BASED MODEL FOR MAIN STEM NODE NUMBER</td>
<td>23</td>
</tr>
<tr>
<td>Introduction</td>
<td>23</td>
</tr>
<tr>
<td>Materials And Methods</td>
<td>26</td>
</tr>
<tr>
<td>Plant Materials And Field Sites</td>
<td>26</td>
</tr>
<tr>
<td>Phenotyping</td>
<td>27</td>
</tr>
<tr>
<td>Molecular Marker And Linkage Mapping</td>
<td>27</td>
</tr>
<tr>
<td>Dynamic QTL Effect Model</td>
<td>27</td>
</tr>
<tr>
<td>Gene-Based Common Bean Model Framework</td>
<td>29</td>
</tr>
<tr>
<td>Model Calibration And Evaluation</td>
<td>32</td>
</tr>
<tr>
<td>Results</td>
<td>33</td>
</tr>
<tr>
<td>Dynamic QTL Effect Modules</td>
<td>33</td>
</tr>
<tr>
<td>Evaluation of GB-CBM</td>
<td>37</td>
</tr>
<tr>
<td>Discussion</td>
<td>39</td>
</tr>
<tr>
<td>Conclusion</td>
<td>46</td>
</tr>
<tr>
<td>3 INTEGRATION OF A MAIN STEM LEAF AREA MODULE</td>
<td>58</td>
</tr>
<tr>
<td>Introduction</td>
<td>58</td>
</tr>
<tr>
<td>Materials And Methods</td>
<td>61</td>
</tr>
<tr>
<td>Plant Material and Field Sites</td>
<td>61</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Site Management Summary</td>
<td>48</td>
</tr>
<tr>
<td>2-2</td>
<td>The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for 1/duration from emergence to flowering (RF)</td>
<td>49</td>
</tr>
<tr>
<td>2-3</td>
<td>The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for maximum number of nodes along the main stem (MSNODmax) module</td>
<td>50</td>
</tr>
<tr>
<td>2-4</td>
<td>The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for main stem node addition rate (NAR)</td>
<td>51</td>
</tr>
<tr>
<td>3-1</td>
<td>The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for potential leaf area for nodes on the main stem (MSLAmax) module</td>
<td>77</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>The framework of the gene-based Common Bean Model (GB-CBM)</td>
</tr>
<tr>
<td>2-2</td>
<td>The QTL that were identified from multi-environment composite interval QTL mapping for time from planting to flowering (TF), maximum number of nodes along the main stem (MSNOD\text{max}), and node addition rate (NAR) in the common bean RI population.</td>
</tr>
<tr>
<td>2-3</td>
<td>The simulated with daily time steps from the RF (anthesis days after emergence) (A and B), MSNOD\text{max} (maximum main stem node number) (C and D), and node number during the linear phase of node addition predicted by NAR (node addition rate) (E and F) module versus observed data.</td>
</tr>
<tr>
<td>2-4</td>
<td>Simulation results using the GB-CBM to predict main stem node number over time for 187 genotypes (RILs), across the five field sites.</td>
</tr>
<tr>
<td>2-5</td>
<td>Simulated versus observed plots of main stem node number using the GB-CBM across time for 187 genotypes (RILs) by field sites.</td>
</tr>
<tr>
<td>2-6</td>
<td>Simulated versus observed plots of Anthesis (days after planting; ADAP) using the GB-CBM for 187 genotypes (RILs) by field sites.</td>
</tr>
<tr>
<td>3-1</td>
<td>Observed leaf area [cm\text{2}] for node positions 1~5 (LA1-LA5) for the two parents (Jamapa (right) and Calima (left)) across the four field sites.</td>
</tr>
<tr>
<td>3-2</td>
<td>Box plot of observed potential leaf area [cm\text{2}] for node positions 1~5 (MSLAMax\text{1-5}) for the training data set of 171 RILs across the four field sites.</td>
</tr>
<tr>
<td>3-3</td>
<td>The framework of the gene-based Common Bean Model (GB-CBM).</td>
</tr>
<tr>
<td>3-4</td>
<td>The QTL that were identified from multi-environment composite interval QTL mapping for the potential leaf area for nodes on the main stem (MSLAMax) in the common bean RI population.</td>
</tr>
<tr>
<td>3-5</td>
<td>Simulated results of the 171 training set RILs for the seasonal linear mixed effect model.</td>
</tr>
<tr>
<td>3-6</td>
<td>Simulation results using the GB-CBM to predict main stem leaf area [cm\text{2}] over time for 187 genotypes (RILs), across the five field sites.</td>
</tr>
<tr>
<td>3-7</td>
<td>Simulation results using the GB-CBM to predict main stem leaf area [cm\text{2}] over time for the first 5 nodes. The leaf area of each node position (1~5) for the parents across the four field sites with observed data.</td>
</tr>
</tbody>
</table>
3-8 Simulated versus observed plots of main stem leaf area using the GB-CBM across time for 187 genotypes (RILs) by field sites ..................................................................................................................85

A-1 Environment sensitivity analyses for rate of progress toward flowering (RF) ...............91

A-2 Environment sensitivity analyses for node addition rate (NAR) module .........................92

A-3 Simulation results using the GB-CBM to predict main stem node number over days after planning for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with observed data. ..................................................................................................................93

A-4 Simulated versus observed plot of main stem node number using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA)..........................................................94

A-5 Simulated versus observed plot of anthesis days after planting using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA). ..........................................................95

A-6 Simulation results using the GB-CBM to predict total main stem leaf area (cm²) over days after planning for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with observed data ........................................................................................................96

A-7 Simulated versus observed plot of total main stem leaf area up to senescence using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) ...............97
LIST OF ABBREVIATIONS

E  Environment effect.; the effect of environmental conditions on plant growth and development.
G  Genetic effect; the effect of genetics on plant growth and development.
G x E  Gene-based common bean model; the presented model.
GSP  Genotype Specific Parameter; empirical parameters in existing crop models that have genotype specific values after parameter estimation.
HTP  High Throughput Phenotyping; the method of measuring plant phenotypes through automated processes for large scale data acquisition.
MSNODmax  Main Stem Node Number max; maximal or final main stem node number.
MSLA  Main Stem Leaf Area; leaf area [cm$^2$] on the main stem.
MSLAmax  Main Stem Leaf Area max; the maximal leaf area [cm$^2$] for a node on the main stem.
NAR  Node Addition Rate; the rate of main stem node appearance.
PAR  Photosynthetically Active Radiation; frequencies of light that are absorbed by plant leaves for photosynthesis.
QTL  Quantitative Trait Locus; a section of the chromosome that has been grouped together to contribute to an observed trait. QTLs can contain multiple genes.
RF  Rate of progress towards flowering; the inverse of duration towards flowering.
RIL  Recombinant inbred lines; offspring derived from successive inbreeding. RILs represent some of the possible genetic diversity when given the parent genetic make-up.
A GENE-BASED CROP MODEL OF THE COMMON BEAN FOR EARLY VEGETATIVE GROWTH OF THE MAIN STEM AND REPRODUCTIVE DEVELOPMENT

By

Christopher Hwang

August 2016

Chair: Melanie J. Correll
Major: Agricultural and Biological Engineering

The next generation of crop models will be able to predict crop performance under varying environments by incorporating the crop's genetics with environment and management information. Since these gene-based crop models can predict crop performance in targeted environments without having to conduct multi-environment experiments, they will accelerate the time for plant breeders to develop new cultivars. To build a gene-based common bean (Phaseolus vulgaris L.) model, main stem node addition rate, final main stem node number, potential leaf area for nodes on the main stem, and rate of progress towards flowering modules were developed. Each module was developed with linear mixed-effect models to identify the crop’s genetics (i.e., quantitative trait loci (QTL)), the daily environment factors (E), and QTL x E interactions to estimate traits across the entire season. The modules were then converted to dynamic, QTL-effect modules by estimating parameters based on a daily time step. A novel model framework was then developed to simulate populations across environments. Field data was collected in Citra, FL; Fargo, ND; Palmira and Popayan, Colombia; and Puerto Rico between 2010 and 2011 and used for model development and parameter estimation. Model evaluation with data collected from Palmira in 2016 showed reliable prediction with $R^2$, 

11
%RMSE, %Bias, and d-stat for main stem node number (0.9, 22.97, 2.21, and 0.97), flowering days after planting (0.73, 9.37, -5.92, and 0.78), and main stem leaf area (0.81, 52.99, -3.86, and 0.93). This thesis demonstrates one approach to integrate process-based modules to develop a gene-based crop model.
CHAPTER 1
GENERAL INTRODUCTION

Statement of the Problem

The world is transitioning from one of food abundance to one of food scarcity (Full Planet, Empty Plates, Brown). In July 2015, the United Nations released their latest global population projections of 9.7 billion by 2050 with the continent of Africa accounting for over half of the world’s population growth (UN report, 2015). With such growth, an increase in global food production is needed yet global arable land percentage has stagnated at around 10.9% since 2006 (World Bank, 2016). In addition, crop systems place significant stress on natural resources (e.g., water, soil, and minerals) further limiting production of food. Globally, agriculture accounted for 59.6% of fresh water withdrawals in 2014 (World Bank, 2016) and weather variability and frequency of extreme events (e.g., droughts and heat waves) are projected to increase in the future due to climate change. Recent extreme heat and drought events across the globe have resulted in national cereal production losses of 9~10% (Lesk et al. 2016). To increase food production in response to these threats, higher yielding cultivars that require fewer natural resources, better field management, and cultivar selection for climate change are also required.

Plant breeders can design cultivars suited for targeted environments and for resource conservation as one step towards improving global food security. However, costs associated with performing field trials limits the breeders’ ability to create crops that are adapted for climate change. Breeding cycles also take course over 5 to 10 years to produce few marketable cultivars. A solution to this problem is to build tools that simulate crop growth and development in varying environmental conditions as affected by genetics, environment, and management; thus allowing the breeder to perform virtual experiments (Hatfield and Walthall, 2015). Currently, large breeding organizations have developed statistical computational tools to predict plant
performance in contrasting environments, but these models have limitations since they look at whole season outputs rather than the dynamic events that plants undergo within the season. Mechanistic crop models are computational tools that simulate dynamic crop growth and developmental processes that are used by farmers, policy makers, and economists to optimize field management, and for decision support; and have the potential to be a tool for plant breeders. However, mechanistic ecophysiology crop models, in their current state, are used scarcely by plant breeders due to the costs associated with collecting data to build and calibrate these models (Langridge et al. 2011). In addition, crop models are limited in their abilities to incorporate genetic information that would be useful to the breeder. Now, technological improvements to sequence the genetic information of plants and measure their responses to the environment on large scales have generated large quantities of genomic and phenomic datasets that are now readily available. A crop model that links these datasets to predict plant growth and development would be a valuable tool for plant breeders by providing insight on target selection (Langridge et al. 2011). This thesis presents a novel, mechanistic, crop model that included genetics, environmental effects, and gene-by-environment interactions to model early vegetative growth and reproductive developmental processes; a tool for plant breeders.

**Crop Models**

Mechanistic crop models simulate different plant growth and developmental stages on a biophysical process level as affected by the cultivar, soil, climate, and management with multiple computational algorithms to estimate crop growth and yield. These algorithms can range from modules that simulate canopy architecture and light interception for assimilate production; life-cycle simulation; soil nitrogen and water balance; to carbon balance and growth of plant parts as affected by assimilate production, phenological stage, and sink strength of the individual organs (Boote et al. 2013). Therefore, crop models can be used to optimize sowing date, fertilizer
application, and irrigation events to maximize yield and profit. Furthermore, these computational tools can be used to simulate multiple growing seasons with crop rotations or simulate growing conditions in hypothetical climatic conditions. Thus, they are used by farmers for field management; policy makers for decision support; and economists for local and global agricultural trade. Unfortunately, current crop models are limited in their connection to plant genetics (Boote et al. 1996). Although they include empirically-derived parameters that allow them to simulate performance of different varieties, they still do not take into account the gene-by-environment (G x E) or epistatic genetic interactions at the level of the individual plant processes simulated (e.g., node addition, leaf area expansion). Many crop models contain a specific type of empirical parameter termed genotype specific parameter (GSP) that allow them to simulate cultivar-specific environmental sensitivities. The empirical nature of these GSPs result in the need to conduct multi-environmental experiments with detailed field measurements when new cultivars are released. These experiments are time consuming and costly. However, since these models are built on frameworks that include environmental sensitivities that allow them to simulate dynamic growth and development processes, there is potential to incorporate genetics (G) at a process level, including interactive effects among G, environment (E), and management (M). Recent developments in the fields of genetics, robotics, machine vision, as well as computational power now allows for the crop modeling community to update existing model frameworks to expand its understanding on the genotype-to-phenotype link. Advancements in genomics, phenomics, and computational technologies is discussed next.

Genotyping

New DNA sequencing technologies have reduced costs, and paved the way for researchers to conduct genome-wide association studies (GWAS) using a variety of genetic data with different resolutions. For example, quantitative trait loci (QTL) linkage maps have been
used for identifying the connection between the plant’s genetics and a trait (Yin et al. 2000; Ma et al. 2002; Reymond 2004; Yin et al. 2004; Nakagawa et al. 2005; Boer et al. 2007; Uptmoor et al. 2008; Xu et al. 2011; Yang et al. 2011; Gu et al. 2014; Jiang et al. 2015). QTLs are sections of the chromosome that contribute to an observed trait, and can contain multiple genes. Creating QTL-based linkage maps is an ideal first step to identify regions of the chromosome attributing to the inspected trait because of their low production cost, and relatively smaller data size compared to other higher resolution maps (i.e., single nucleotide polymorphism (SNP)-based map). QTL regions attributing to a trait can be further sequenced for higher resolution maps to identify specific genes or SNPs.

Single nucleotide polymorphism (SNP)-based maps are the latest development in sequencing technologies that detect single DNA base pair differences in the genome of different cultivars (Gaitán-Solís et al. 2008). SNP-based high-throughput genotyping has great potential to develop high density genetic maps for GWAS of large populations (Wang et al. 1998; Gaitán-Solís et al. 2008). These SNP technologies have already been used in genotyping studies for a wide range of species: modern maize (Wright et al. 2005), sugarbeet (Schneider et al. 2001), barley (Kanazin et al. 2002), sorghum (Hamblin et al. 2004), rice (Feltus et al. 2004; Huang et al. 2010; W. Yang et al. 2014), and bean / soybean (Zhu et al. 2003; Hyten et al. 2006; Hyten et al. 2010; Blair et al. 2013; Schmutz et al. 2014; Gaitán-Solís et al. 2008; Hart and Griffiths 2015).

In this study, we conducted our analyses with a QTL-based linkage map since the SNP-based sequencing of our genotypes were not available at the time. Despite the lower cost of genetic data, the major limitation to large scale genotype-to-phenotype analyses are the time and cost of the phenotyping (Huang and Han 2014; Han and Huang 2013).
Phenotyping

Large-scale phenotyping methods such as the use of unmanned aerial vehicles (UAVs), robotics, and sensor technologies, commonly referred to as high throughput phenotyping (HTP), are reducing costs and time for collecting field measurements. These HTP technologies can identify biotic (e.g., disease and plant infections) and abiotic (e.g., nitrogen deficiency or drought) stresses within a field with less human labor than traditional methods (Rousseau et al. 2013). Not only can HTP detect diseases and stress more readily, when compared to traditional visual inspection, they are more reliable as well (Sherwood et al. 1983; Bock et al. 2008; Bock et al. 2010; Poland and Nelson 2011).

A variety of in-field, aerial, and platform based HTP systems have been developed to non-destructively measure plant phenotypes over their growth cycle. These methods can be categorized as: spectral analyses for plant shoot and canopies health, thermography-based soil-plant-atmosphere continuum assessments, optical analyses of above ground plant organ and canopy growth, and root phenotyping in lab and field conditions (Walter et al. 2015). HTP image analyses using leaf fluorescence has been demonstrated to be a robust method in identifying stages of Common Bacterial Blight as it infects bean leaves and locating the regions of chlorosis and necrosis on individual leaves (Rousseau et al. 2013). A platform based rice phenotyping facility has also been constructed in China that measures a variety of variables on morphology, biomass, and yield over the course of plant growth and post-harvest (Yang et al. 2014). The system was also able to predict yield-related traits with the measured morphological and biomass traits. As our ability to phenotype large population trials with low cost HTP technologies grows, the scientific community will be able to accelerate linking the genotype of a plant to its phenotype to improve our understanding of plant genetic networks across varying environments and allow us to investigate traits of higher complexity. As part of these advancements, a variety
of computational methods have recently been developed to identify the genes (G) and gene by environment (G X E) interactions affecting a plant trait.

**Computational tools for identifying the G, G x E interactions**

The majority of analyses used to identify the genes associated with a trait under different environments, most commonly utilize one of three statistical methods: interval mapping, functional mapping, and genomic selection. The details of the procedures for interval mapping have been described in Boer et al. 2007 and thus are only summarized here. In the context of developing gene-based models for specific plant growth and developmental traits, interval mapping is used to identify a mathematical model that reflects the G, E, and G x E interactive effects as well as their magnitudes over a time period relevant to the trait. The underlying assumption of interval mapping is that G, E, and G x E interactions have static values through a plant’s life cycle. This assumption of static values in a trait has limitations since a plant can reach a trait value by many different growth trajectories. To address this limitation, functional mapping methods were developed (Wu and Lin 2006). Functional mapping is a method that not only identifies the genetic markers associated with the trait growth trajectories, but also the fluctuations in the G and G x E interactive effects over time period (Ma et al. 2002; Wu and Lin 2006; Jiang et al. 2015). Functional mapping has shown great promise for predicting developmental traits that follow functional growth patterns (e.g., logistic growth) with linkage maps of varying densities. Recent trends in analyses to find the G, G x E interactions affecting a trait have gravitated towards a novel strategy that takes advantage of the larger data size of SNP-based marker maps termed genomic selection (Heffner et al. 2009; Shengqiang et al. 2009; Lopez-Cruz et al. 2015; Technow et al. 2015). Genomic selection develops a model that utilizes all the available genetic marker information in the linkage map to estimate trait values. Thus, genomic selection methods take full advantage of the SNP-based linkage maps being produced.
Recently, genomic selection methods have been able to take into account G x E interactive effects (Lopez-Cruz et al. 2015). While all these approaches provide promise for developing models for integrating the G, G x E interactions, for the purpose of this study, we implemented interval mapping to develop a gene-based crop model.

**Next-Gen Crop Models; Gene-based Models**

To develop gene-based crop models, multiple approaches with varying degrees of complexity have been proposed. The most common and simplest approach is to take model parameters (genotype specific parameters, GSP) and make them functions of the genes that are associated with the parameter (White and Hoogenboom 1996; White and Hoogenboom 2003; Yin and Struik 2004). The dry bean (*Phaseolus vulgaris* L.) model in DSSAT (CROPGRO-Bean) was modified to include seven genes that were assumed to affect phenology, growth habit, and seed size. Linear functions of these genes were used to predict the genotype specific parameters (GSPs) based on the original functions for computing process rates (White and Hoogenboom 1996). This gene-based model accurately predicted phenology but was unable to accurately predict yield variations (Hoogenboom et al. 1997). Another example of this process is when the GSPs in the CROPGRO-Soybean model were converted into mathematical functions of the *E* loci, a day length-sensitive loci, and used to simulate the reproductive development behavior of soybean cultivars. The modified model accounted for 75% of the variance in maturity date in independent cultivars from Illinois, USA based on weather data and 4 of the 6 *E* loci, which were found using SSR-linked markers (Messina et al. 2006). Chenu et al. (2009) modified the APSIM maize model (B. a. Keating et al. 2003) to allow parameters for leaf and silk elongation to be affected by QTLs and these were used to simulate maize growth under drought conditions. The study was able to construct a QTL network affecting the examined traits, and identify the best combination of traits for yield under the management practices in the
experiment. Another example of this is when the CERES-Maize model was modified to simulate leaf expansion, longevity, and senescence by each leaf position that also took into account some genetic effect through the addition of three empirical GSPs for leaf area simulations (Lizaso et al. 2003). The original and modified CERES-Maize models were used to simulate field trials for 5 cultivars in Morris, MN, Gainesville, FL, and Honolulu, HI between 1982 and 1986. In general, the update to the CERES-Maize was successful since the RMSE for leaf area index was lower for the modified model simulations which ranged from 0.29 to 1.25. This modified model simulated leaf area expansion and senescence by node positions as well as cultivar performance variations through GSPs. Others used parameters for maize leaf width, length, and elongation rate based on relevant QTL marker information as well as QTL x E effects to predict vegetative growth and development under non-stressed growth conditions (Reymond et al. 2004). More recently, a study using some of the same data from this thesis looked at leaf area and dry weight growth trajectories as affected by QTL effect variations over time for the common bean (Jiang et al. 2015). The study was able to simulate a range of logistic growth rate and inflection point variations for leaf area and dry weight of a family of RILs as informed by temporal QTL effect fluctuations. However, that study only included Palmira (PA) and Popayan (PO), Colombia sites whereas here all field locations (i.e., Citra, FL (CT); PA, PO, Fargo, ND (ND), and Puerto Rico (PR)) were used in either model development or evaluation for the gene-based models presented.

Recently, whole-genome prediction methods (statistical approaches) when linked with crop models, have shown increased accuracy in prediction of yield in new environments in comparison with using statistical approaches alone (Technow et al. 2015).

More complex approaches for integrating genetics with crop models have been suggested and include simulating gene expression over the course of plant development or including
polypeptide translation information (White and Hoogenboom 2003). The higher level of biological networks/levels may increase errors within the models, and it has been argued that further increasing the level of complexity in regards to the genetics may not be necessary for further improving crop models as breeding tool if they capture the physiological basis of the traits (Yin et al. 2004; Hammer et al. 2010). Taken together, there are a variety of approaches for predicting plant processes based on the genetic and environment information and likely no one approach will be best in all cases. The type of modeling approach chosen will largely be based on the availability of data and the goal of the project (e.g., targeting a specific process or overall crop yield).

The Common Bean

The common bean (*Phaseolus vulgaris* L.) is the legume with the highest level of direct consumption around the world and is an important protein and nutrient source for the malnourished poor in Latin America and Africa (Broughton et al. 2003). The genome of the bean is relatively small with 11 chromosomes, and its sequence was recently published (Schmutz et al. 2014). The common bean also has high genetic diversity due to the two separate domestication gene pools of Andean and Mesoamerican origins (Gepts et al. 2008). Genetic marker maps developed for the common bean can be used as road maps for other genomes of higher complexity such as soybean (Hyten et al. 2010; Hyten et al. 2008) and cowpea (Muchero et al. 2009). The combination of these attributes in the common bean and its rapid growth cycle of 60–120 days make it an ideal crop candidate for developing gene-based models.

Objectives and Thesis Organization

A gene-based common bean model (GB-CBM) that simulated early vegetative growth and reproductive development processes for main stem node number and leaf area by node positions, flowering days after planting, and final main stem node number for the common bean
is presented. These process simulations took into account the genotype of the crop (represented by QTLs), plant response to the environment, and genotype-by-environment interaction (i.e., QTL x E) effects. To construct such a model, a procedure was developed for each process module (chapters 2 and 3) which were then all integrated together in a novel model framework (chapters 2 and 3). In total, 4 modules were developed and then nested in the GB-CBM framework. These modules included a rate of progress towards flowering (RF, modified from Bhakta 2015), main stem node addition rate (NAR, modified form Zhang 2015), final main stem node number (MSNODmax, chapter 2), and potential leaf area for nodes on the main stem (MSLAmx, chapter 3). Chapters 2 and 3 present the approach for building the modules of the model and integrating them together for module evaluation. Chapter 4, presents the future directions of this work and next steps for building a gene-based crop model.
CHAPTER 2
A DYNAMIC GENE-BASED MODEL FOR MAIN STEM NODE NUMBER

Introduction

Tools that integrate genetic, environment and management information to predict crop performance in contrasting environments are needed to meet global food demands and assist plant breeders in designing new cultivars for increased yield (Hatfield and Walthall, 2015). Crop models are biophysical process-based simulation tools that predict crop growth and yield for a range of soil, climate, and management conditions. However, although they include empirically-derived parameters that allow them to simulate performance of different varieties, they still lack the integration of actual genetic information and thus are limited in their connection to plant genetics (Boote et al. 1996). Of note is that parameters termed “genetic coefficients” or “Genotype-Specific Parameters (GSPs)” that describe phenology, plant architecture (leaf area, number and plant dimensions), and biomass allocation in existing crop models are empirically estimated and not yet linked to any gene(s). They do not take into account gene-by-environment (G x E) or G x G interactions at the level of individual processes that are considered in the models. This lack of genetic information as explanatory variables results in the necessity to conduct multi-environment experiments when new cultivars are released. This process is time consuming, costly, and limits the utility of crop models in plant breeding programs and other practical applications. This omission of G and G x E information in crop models is not surprising since many of these models were developed before this type of information was known. However, these models do include environmental sensitivities, relationships, and traits that allow them to simulate dynamic growth and development processes. These processes and the parameters that quantify them can now potentially be used to guide the incorporation of genes at a process level, including interactive effects among G, E, and management (M).
The advances in genomics, phenomics (phenotyping), and computational technologies within the last decade have given scientists the unprecedented opportunity to understand the shaping of a given crop phenotype by the complex interactions among genotype, environment, and management. For example, new DNA sequencing technologies have increased the number of genetic markers for identifying genes associated with phenotypic traits. Large-scale phenotyping methods such as the use of unmanned aerial vehicles (UAVs), robotics, and sensor technologies are reducing costs and time for collecting field phenotype measurements. Also, new computational and statistical tools are rapidly advancing our ability to identify genes and environmental factors that affect crop traits. In spite of the technical advances and statistical sophistication of gene mapping approaches, few researchers have tackled the prediction of phenotypic vegetative and reproductive development of a genotype as affected by G, E, and G x E using daily (or shorter time step) environmental inputs. Nor do most studies reveal biological insights into the mechanisms of crop performance in specific environments (Technow et al. 2015). Since crop models have the capacity to model daily vegetative and reproductive development from a mechanistic standpoint, and we can now quantify G x E interactions at a process level, integrating this information into crop models provides an opportunity to start building next generation of gene-based crop models.

The earliest and most common approach for integrating genetic information into crop models is linking specific genes to model parameters (i.e., the model’s GSPs that had been estimated from field data, see White and Hoogenboom, 1996; 2003, and a more recent review by Yin and Struik, 2010). For example, GSPs in the CROPGRO-Soybean model were converted into mathematical functions of day length-sensitive genes (E loci), which were used to simulate the flowering and maturity behavior of soybean based on genetic information of cultivars.
(Messina et al. 2006). Chenu et al. (2009) modified the APSIM maize model (B. a. Keating et al. 2003) with parameters for leaf and silk elongation being computed by quantitative trait loci (QTL) and were able to simulate maize growth under drought conditions. Others have shown that using whole-genome prediction methods (statistical approaches) when linked with crop models, have increased accuracy in prediction of yield in new environments in comparison with using statistical approaches alone (Technow et al. 2015). While these approaches appear to be promising methods for integrating genetics into crop models, current crop models lack specific gene-by-environment interactions at a process level and many assume uniform environmental responses across genotypes.

QTL analyses can dissect the genetic architecture of complex traits, and in combination with statistical methods, such as mixed effect models, it is possible to estimate the genetic, environmental, and G x E effects on the phenotype (Boer et al. 2007; Chenu et al. 2009; Peiffer et al. 2014). We propose that these mixed effect approaches can be used to identify QTL, E, and QTL x E interactions underlying specific crop processes, using knowledge and experience from previous crop models, and later integrate them together to build a gene-based crop model that predicts crop performance based on genetic, environment, and management data. One advantage to this approach is that crop models already have subroutines (modules) that simulate different processes, such as phenological development, leaf area expansion, dry matter accumulation, and seed growth that are integrated together to simulate overall crop growth (Jones et al. 2001; 2003). Specific subroutines modeling selected biological processes could be modified to incorporate G, E, and G x E effects on those particular processes; the modified subroutines can then provide more comprehensive information to the central model, which integrates information from multiple subroutines (Boote et al. 2013). Also, many studies have already demonstrated the
usefulness of using this type of QTL-based approach to modify modules within crop models (Yin et al. 2000; Reymond et al. 2003; Nakagawa et al. 2005; Messina et al. 2006; Uptmoor et al. 2008; Chenu et al. 2009).

Here, we describe a prototype gene-based model that simulates the main stem node number over time and flowering date for common bean as affected by the genotype of the crop (represented by QTLs), and its response to the environment, and genotype-by-environment interactions (QTL x E) by integrating dynamic QTL effect models for daily development rate of progress toward appearance of the first flower (RF\(t\), modified from Bhakta 2015a), node addition rate (NAR\(t\), Zhang 2015), and daily maximum main stem node number (MSNOD\(_{\text{max}}\)). The modules described here incorporate relationships previously identified through a mixed-effect statistical model approach (QTL, E, and QTL x E), knowledge of physiological processes, and a daily time step with corresponding E inputs that vary daily (e.g., day length, temperature, solar radiation). The described approach traces a path toward building a next generation of gene-based crop models using QTL, E, and QTL x E interaction effects on separate development and growth processes, which are hypothesized to be more capable of predicting the phenotype of specific genotypes over a range of environments.

**Materials And Methods**

**Plant Materials And Field Sites**

The details on the RI family of the common bean that were used in these studies can be found in Bhakta et al. (2015b). Briefly, the RI family of 187 genotypes was generated from a cross between the determinate Andean cultivar, *Calima*, with an indeterminate Mesoamerican cultivar, *Jamapa*, for these studies (Bhakta et al. 2015b). The population was developed through single seed descent to the 11\(^{\text{th}}\) generation, and bulked to the 14\(^{\text{th}}\) generation (F\(_{11:14}\)). The population was then planted across five field sites: Citra, FL (CT); Palmira, Colombia (PA);
Popayan, Colombia (PO); Isabela, Puerto Rico (PR); and Prosper, North Dakota (ND). Details on the field sites are provided in Table 2-1 with the weather files in supplemental table (Table A 2-1 weather information). The experimental design followed a latinized, row-column design with three blocks (3 plots of each genotype, and 6 to 9 plots for each parent line). Details of the experiment design are presented in Clavijo Michelangeli (2014).

**Phenotyping**

Two types of phenotypic data were collected. The first data were non-destructive measurements in which 6 plants per plot (marked after emergence) were observed every 2 to 3 days for developmental time-to events (such as time to first anthesis). For each replication in each block, the number of days it took 50% of the plants to reach anthesis was determined. The second type of data included weekly destructive samplings of 3 replicates (one plant per plot) performed after emergence of the first true leaf. Samples were collected at each site depending on the availability of plants and measurements of node numbers were recorded for the main stem on each day of sampling. These data (number of nodes on the main stem on day \( t \), \( N_{\text{obs}}(t) \), duration between emergence and first flower, TF, and maximum node number on the main stem, \( \text{MSNOD}_{\text{max}} \)) were used to develop dynamic QTL effect models for node addition rate, rate of progress toward flowering, and maximum node number on the main stem, respectively.

**Molecular Marker And Linkage Mapping**

A SNP-based linkage map was constructed with the RI family using the genotyping-by-sequencing (GBS) method described in Bhakta et al. (2015b). The final linkage map comprised 513 molecular markers with an average interlocus distance of 1.9 cM.

**Dynamic QTL Effect Model**

Boer et al. (2007) have described in detail procedures to develop statistical linear mixed effect models to evaluate, for a given phenotype, the magnitude of the G (QTL), E, and G (QTL)
x E interactive effects on traits, which were implemented in this study using all 187 RILs and the five environments. The procedures were based on multi-environment, single trait analysis with an unstructured error variance-covariance matrix. Accordingly, this methodology was applied to fit statistical models for the traits to identify significant QTL, E, and QTL x E factors for the selected processes in this study using GenStat 15th edition (Payne et al. 2012). The linear mixed effect models developed by Bhakta (2015a) for time to flowering (TF) and by Zhang (2015) for node addition rate (NAR) used the mean value of environmental variables over the time period of each process for each RI line to predict a static trait (i.e., the time to flower or node addition rate). Bhakta (2015a) developed a linear mixed model for time to first flower using genotypic and environmental variables averaged over the time between planting and first flower (minimum temperature, maximum temperature, solar radiation, and day length). Here, a dynamic interpretation of the model was obtained by using daily values of the environmental variables to predict daily rate of progress to first flower (1 / duration from emergence to anthesis, d−1); this is the approach used in the CROPGRO-Bean model (Jones et al. 2003) and most dynamic crop models. Similarly, we transformed the node appearance rate linear mixed model developed by Zhang (2015), which used environmental variables averaged over the time during the linear phase of node addition for each genotype at a site into a dynamic model that responds to daily environmental variables. By integrating the daily rates, the dynamic rate models account for variations in QTL, E and QTL x E interactions over time, which is important for predicting crop development in the field. For example, we used daily mean temperature (TMEAN(t)) values, even though Bhakta (2015a) found that the individual temperature covariates (average max and min temperatures) over an extended time period affected the rate of development toward
flowering trait. However, all of these covariates are highly correlated and present similar predictive power.

The general scheme of the dynamic QTL and environmental effects model for each trait (i.e., two daily development rate traits, RF(t) and NAR(t), and one static trait (final main stem node number, N_final) estimated with the MSNOD_max module), is presented in Eq. 2-1 below.

\[ y(t) = \mu + y_{ai} \cdot \sum_{i=1}^{E} (E_i(t) - \bar{E}_{Ei}) + y_{bj} \cdot \sum_{j=1}^{QTL} QTL_j + y_{ck} \cdot \sum_{j=1}^{GxE} QTL_j \cdot (E_i(t) - \bar{E}_{Ei}) \]  

(2-1)

where a phenotypic trait (\(y(t)\)) for a day (\(t\)) is predicted with the trait general mean (\(\mu\)); trait environment effect parameter (\(y_{ai}\)) for the \(i^{th}\) environment covariate (ECV); daily ECV values (\(E_i(t)\)); ECV means across sites for a trait (\(\bar{E}_{Ei}\)); trait QTL effect parameters (\(y_{bj}\)) for the \(j^{th}\) QTL marker; QTL marker values (\(QTL_j\)); trait QTL x E parameters (\(y_{ck}\)); and trait QTL x E effects \([QTL_j \cdot (E_i(t) - \bar{E}_{Ei})]\). Each dynamic QTL effect model (RF(\(t\)) and NAR(\(t\))) thus uses daily environmental inputs with the equation structure shown in Eq. 2-1 whereas the MSNOD_max module uses averages environmental inputs over time from planting to the current simulation day until anthesis to predict final number of nodes on the main stem.

**Gene-Based Common Bean Model Framework**

A prototype, gene-based Common Bean Model (GB-CBM) was developed by combining the three modules that simulate rate of progress toward first flower (RF(\(t\))), maximum main stem node number (MSNOD_{max}), and node addition rate (NAR(\(t\))) and integrating the daily-predicted rates to simulate day of first flower, node numbers on the main stem over time for common bean, and final number of main stem nodes (N_final). The structural layout of the GB-CBM is presented in Fig. 2-1. Input files for weather, management, field observations, genetic marker information (QTL), and gene-based module parameter values are read at program initiation. A genotype (G) in the population is selected and simulated for each of the five sites (S) Citra, FL (CT); Fargo,
ND (ND), Palmira, Colombia (PA); Popayan, Colombia (PO); and Isabela, Puerto Rico (PR). For each genotype at a site \(G_s\) variables are computed based on daily weather information from the start (planting) to the last day of the experiment (DAY). The GB-CBM then runs the modules for flowering \(((\text{duration to flowering})^{-1}, \text{RF})\), main stem node addition rate (NAR), and maximum main stem node number \(\text{MSNOD}_{\text{max}}\) to simulate the dynamic changes in main stem node number and time to first flower for each \(G_s\). This process is repeated for the entire population to produce a simulated distribution of main stem node numbers over time, days to first flower across multiple sites, and final node number.

The \(\text{RF}(t)\) and \(\text{NAR}(t)\) rate modules were run each simulation day based on the allelic makeup at relevant QTLs and the daily environmental conditions for each RI line and across all five environments. The day of first flower was simulated for each line growing in each environment by the computed development progress, \(\text{TF}(t)\). \(\text{TF}(t)\) was computed by numerically integrating daily computations of \(\text{RF}(t)\), starting at emergence (Eq. 2-2).

\[
\text{TF}(t) = \text{TF}(t-1) + \text{RF}(t) \cdot dt,
\]
for \((\text{emergence date} \leq t \text{ and } \text{TF} < 1)\) \hspace{1cm} (2-2)

Simulated nodes on each day \(\text{N}(t)\) were obtained by numerical integration of the NAR values as follows:

\[
\text{N}(t) = \text{N}(t-1) + \text{NAR}(t) \cdot dt
\]
for \((\text{TF} < 1)\) for determinate genotypes

or (any TF for indeterminate genotypes)

and \((\text{N}(t-1) < \text{MSNOD}_{\text{max}}(t-1), \text{for all genotype})\) \hspace{1cm} (2-3)

starting at time \(t\) equal to or greater than emergence day. For determinate genotypes, numerical integration with \(\text{NAR}(t)\) is performed as long as first flower has not occurred and node number is less than the maximum main stem node number \(\text{N}_{\text{final}}\). For indeterminate genotypes, numerical integration with \(\text{NAR}(t)\) is performed beyond first flower and while node number is less than the
final main stem node number ($N_{\text{final}}$). $dt$ is the numerical integration time step ($dt = 1$ day in this study). For determinate genotypes, final main stem node number ($N_{\text{final}}$) is set at first flower or when node number reaches the maximum main stem node number estimate of the previous day. For indeterminate genotypes, final main stem node number ($N_{\text{final}}$) is set when node number reaches the estimate of maximum main stem node number estimate of the previous day ($\text{MSNOD}_{\text{max}}(t-1)$).

In the GB-CBM, the QTL (i.e., JC28) region that contains the $\text{FIN/TFL1Y}$ gene defines the determinacy of a genotype (Repinski et al. 2012). That QTL region determines whether a terminal inflorescence will develop on the main stem. Due to the bi-parental nature of our RI population, Calima (determinate) markers are assigned the value of +1 while Jamapa (indeterminate) markers are assigned -1. Therefore, a genotype is indeterminate and continues to add nodes up to the maximum main stem node number after flowering if it has a QTL marker value of -1 for that region, otherwise node addition on the main stem ends on the day of first flower appearance (i.e., determinate genotypes stop main stem node addition at anthesis). Molecular marker data used in the models are provided in supplemental table (Table A.2 Genotype). In contrast to the daily development rate traits ($\text{NAR}(t)$ and $\text{RF}(t)$), the $\text{MSNOD}_{\text{max}}$ module was initiated at emergence for each genotype at a given site, but used average environment values from planting to current simulation day ($t$) to estimate daily maximum main stem node numbers. The module was terminated upon simulated anthesis and maximum main stem node numbers were set. Further studies are needed to determine the exact window for the environment as they affect apical node development and differentiation. Parameters for each dynamic QTL effect model were estimated for daily time steps described in the following section.
Model Calibration And Evaluation

All modules were built using the R programing language (version 3.2.3; R Core Team) and the dynamic QTL effect model for each module were calibrated using the nonlinear least squares algorithm in the minpack.lm package (Elzhov et al. 2015) as implemented in the statistical package R to estimate parameters for the explanatory terms in Eq. 2-1 for each of the three traits modeled in this study. Initial values for each parameter were those identified from the linear mixed effect model for each process. The data were divided into two sets in this study: one was for dynamic QTL effect model development and parameter estimation (a training set of RILs across five sites) and the other was for evaluating predictions using combinations not used in the estimation process (evaluation set). Because we are using dynamic models instead of simple static equations to model node number and time of first flower, we used the approach described by Wallach et al. (Chapter 1, 2014) to numerically simulate values using G, E, and parameter inputs for comparison with each observed value to compute errors. To estimate the parameters associated with the QTL and E components of the model for RF(t), errors between observed and simulated duration between emergence and first flower, across all environments (S) and genotypes (G), was used to estimate parameters of the dynamic QTL effect model for first flowering (Eq. 2-4).

\[
SSQError(FL) = \sum_{S=1}^{5} \sum_{G=1}^{171} (FL - FL_{obs})^2
\]  

(2-4)

where FL is simulated day of first flower and FL_{obs} is observed first flower day. Similar to the criterion for fitting the RF(t) dynamic model, the sum of square error between observed number of nodes measured on each date during the linear phase of node addition (N_{obs}(t)) and simulated number of nodes on those same dates (N(t)) were used to estimate the parameters of the dynamic model for node addition rate that minimizes the sum of squared errors (SSQError(N)) between
observed and simulated node numbers across all environments (S), genotypes (G), and observation dates (t) (Eq. 2-5).

$$SSQError (N) = \sum_{S=1}^{5} \sum_{G=1}^{171} \sum_{1}^{aobs} (N(t) - N_{obs}(t))^2$$

(2-5)

The MSNOD_{max} module parameters were estimated based on minimizing the sum of squared errors (SSQError(N_{final})) between observed (N_{final,obs}) and simulated maximum node numbers (N_{final}) across all environments (S), and genotypes (G) (Eq. 2-6).

$$SSQError (N_{final}) = \sum_{S=1}^{5} \sum_{G=1}^{171} (N_{final} - N_{final,obs})^2$$

(2-6)

where N_{final} is simulated final node number determined by using average environment values from planting to anthesis for each genotype at a site.

Model evaluations using the evaluation set of genotypes were performed with the two common bean parents as well as an additional 14 RI genotypes for RF(t) and MSNOD_{max} and 7 RI genotypes for NAR(t) at the five sites using $R^2$, %RMSE, %Bias, and d-statistics (Willmott et al. 1985).

**Results**

**Dynamic QTL Effect Modules**

The three dynamic QTL effect models that were used for the RF, MSNOD_{max}, and NAR modules included a total of 22 QTLs with 10 of these having QTL x E interactions. The QTLs for each module were designated as TF{i} for QTLs found to affect time from planting to first flower (within the RF module), MSNi for QTLs found to affect maximum main stem node number (within the MSNOD_{max} module), and NAR{i} for QTLs found to influence node addition rate (within the NAR module), with i denoting the order which the QTLs were named (Fig. 2-2). Subsets of QTLs associated with each of the three traits were found on the same chromosome segment. For example, QTLs TF2, MSN2, and NAR2, were found in the same region of
chromosome 1. Given the fact that recombination is significantly suppressed in this region, it is highly unlikely that these QTLs will be easily resolved by recombinational analysis. However, we must point out that these QTLs are in the same region occupied by FIN/TFL1Y, a gene that has been found to control growth habit and therefore affect main stem node number (Repinski et al. 2012). Accordingly, genotypes with TF2, MSN2, and NAR2 with value +1 were determinate and stopped main stem node addition at first flower, while indeterminate genotypes (TF2, MSN2, and NAR2 with value -1) continued node addition up to the maximum node number after anthesis. Additional experimentation and molecular analyses will need to be performed to confirm the role of these QTLs/gene and environmental covariates in node development and time to first flower, and to extend the model for QTL x QTL interactions. In addition, we simplified the approach for temperature responses by using daily mean temperature (TMEAN(t)) instead of hourly or minimum and maximum daily temperatures.

The dynamic QTL effect model for RF(t) is shown below (Eq. 2-7) with estimated parameter values for the dynamic, daily model. Temperature and day length interactions have been shown to affect flowering time for common bean by White & Kornegay (1994) and were included in this analysis but not in the mixed effect model presented by Bhakta (2015a). The parameter IDs, estimated parameter values and standard errors (SE) for the RF module are reported in Table 2-2.

\[
RF(t) = 0.029 + 7.5 \times 10^{-4}(TMEAN(t) - 21.35) - 7.3 \times 10^{-6}(SRAD(t) - 18.31) - 2.2 \times 10^{-5}(DL(t) - 12.7) \\
- 3.3 \times 10^{-4}(TMEAN(t) - 21.35)(DL(t) - 12.7) \\
+ 9.8 \times 10^{-4} \cdot TF1 + 1.7 \times 10^{-3} \cdot TF2 - 3.9 \times 10^{-4} \cdot TF3 + 2.0 \times 10^{-4} \cdot TF4 \\
- 1.5 \times 10^{-4} \cdot TF5 + 8.9 \times 10^{-4} \cdot TF6 - 5.3 \times 10^{-4} \cdot TF7 - 3.1 \times 10^{-4} \cdot TF8 \\
- 3.4 \times 10^{-4} \cdot TF9 - 9.7 \times 10^{-5} \cdot TF10 + 2.6 \times 10^{-4} \cdot TF11 - 6.6 \times 10^{-5} \cdot TF12 \\
+ TF2\left(-3.6 \times 10^{-5}(TMEAN(t) - 21.35)\right) \\
+ TF3\left(6.7 \times 10^{-5}(TMEAN(t) - 21.35) - 1.1 \times 10^{-5}(DL(t) - 12.7)\right) \\
+ TF5\left(5.5 \times 10^{-5}(TMEAN(t) - 21.35)\right)
\]
The first term on the right hand side of Eq. 2-7 is the overall average rate of progress toward flowering across sites. The value of 0.029 d⁻¹ indicates that on average, the time between emergence and first flower across all genotypes and sites was 34.5 days. The 4th term indicates that an hour increase above 12.7 hrs in day length would result in a 2.2E-3 lower rate of development from the general mean of 0.029 rate of daily progress towards first flower. Increasing the day length by one hour will increase the time to first flower from 34.5 to 37.3 days provided all other variables were kept at their average values. This timing will also vary as a function of QTL alleles and their interactions with specific environmental variables as indicated in Eq. 2-7. This effect is analogous to the photoperiod sensitivity (PPSEN) parameter currently used in the DSSAT CROPGRO-Bean model to simulate development rate towards anthesis as affected by photoperiod. The Calima QTL allele, TF2^{Cal}, will have a (+1) coefficient and therefore would increase the daily rate by a factor of 1.7E-3 from the general mean rate of 0.029 towards first flower as a result of that QTL effect. Similarly, the same QTL allele will decrease the rate by a factor of 3.6E-5 for a one degree increase in temperature above 21.35 °C. In contrast, the Jamapa allele, TF2^{Jam}, will have the opposite effect. The sensitivity of the RF module to environmental factors can be seen in supplemental Figure (Fig. A.1). However, not all Calima alleles affect the time to first anthesis in the same direction. For instance, although TF3^{Cal} will have a (+1) coefficient, the parameter value (-6.0E-4) of this QTL is negative indicating that the Calima allele of TF3 actually decreases the rate in contrast to the TF2^{Cal} effect.

Next, we present the parameters and equation developed for the MSNOD_{max} module (Eq. 2-8). The parameter IDs, estimated parameter values, and standard errors (SE) for the MSNOD_{max} module are reported in Table 2-3.
\[ MSNOD_{\text{max}}(t) = 12.37 + 0.43(T\text{MEAN}(0:t) - 21.85) + 0.10(SR\text{AD}(0:t) - 18.74) + 1.2(DL(0:t) - 12.81) \]

\[-0.43 \cdot MSN1 - 3.56 \cdot MSN2 - 0.63 \cdot MSN3 - 0.20 \cdot MSN4 - 0.60 \cdot MSN5 + 0.32 \cdot MSN6 \]

\[ + MSN2 \cdot [-0.08(T\text{MEAN}(0:t) - 21.85) - 0.05(SR\text{AD}(0:t) - 18.74) - 0.62(DL(0:t) - 12.81)] \]

\[ + MSN6 \cdot [-0.02(T\text{MEAN}(0:t) - 21.85) + 0.01(SR\text{AD}(0:t) - 18.74)] \]

\[ (2-8) \]

where T\text{MEAN}(0:t) is daily average temperatures averaged between planting (\( t = 0 \)) and the current simulation day (\( t \)) up to simulated anthesis. Eq. 2-8 indicates that a one degree increase above 21.85 °C in the mean temperature would result in additional 0.43 nodes from the general mean of 12.37 maximum nodes as a result of the temperature effect. The Calima QTL allele, MSN2\text{Cal}, will have a (+1) coefficient and therefore would decrease the maximum node number by 3.56 from the general mean of 12.37 maximum nodes as a result of the QTL effect. Similarly, the same QTL allele will decrease the maximum nodes by 0.08 nodes for a one degree increase in temperature above 21.85 °C. In contrast, the Jamapa allele, MSN2\text{Jam}, will have the opposite effect.

The model for NAR(\( t \)) is shown below (Eq. 2-9) with estimated parameter values for the dynamic model. The parameter IDs, calibrated parameter values, and standard errors (SE) for the NAR module are reported in Table 2-4.

\[ NAR(t) = 0.252 + 2.0 \cdot 10^{-2}(T\text{MEAN}(t) - 21.51) - 7.9 \cdot 10^{-4}(SR\text{AD}(t) - 17.38) + 4.4 \cdot 10^{-3}(DL(t) - 12.74) \]

\[-6.0 \cdot 10^{-3} \cdot NAR1 + 7.0 \cdot 10^{-3} \cdot NAR2 + 8.2 \cdot 10^{-3} \cdot NAR3 - 4.5 \cdot 10^{-3} \cdot NAR4 \]

\[ + NAR1 \cdot [-1.9 \cdot 10^{-3}(DL(t) - 12.74)] \]

\[ + NAR2 \cdot [2.1 \cdot 10^{-3}(T\text{MEAN}(t) - 21.51)] \]

\[ (2-9) \]

where T\text{MEAN}(t) is the daily average temperature for simulation day (\( t \)). The average rate of node appearance in the study of 0.252 indicates that there was 3.97 days between the appearances of successive leaf tips. Eq. 2-9 indicates that a one degree increase above 21.51 °C in the temperature for a day would result in a 0.02 faster daily rate from the general mean of 0.252 rate of node addition per day as a result of the temperature effect. This linear temperature
response is analogous to temperature response functions in existing DSSAT CROPGRO-Bean model, where cultivar development rate is calculated with non-linear, piecewise temperature response function $f(T_{\text{base}}, T_{\text{opt}})$. The Calima QTL allele, NAR2\textsuperscript{Cal}, will have a +1 QTL value and therefore would increase the rate of node appearance by 7.0E-3 from the general mean of 0.252 nodes per day as a result of the QTL effect. Similarly, the same QTL allele will increase the node addition rate by 2.1E-3 nodes per day for a one degree increase in temperature above 21.51 °C. In contrast, the Jamapa allele, NAR2\textsuperscript{Jam}, will have the opposite effect. The sensitivity of the NAR module to environmental factors can be seen in supplemental figure (Fig. A.2).

**Evaluation of GB-CBM**

The RF module operating on daily time steps was able to capture the delay in flowering that was observed in ND (Fig. 2-3 A, B) since the dynamic QTL effect model had day length (DL), temperature x DL interaction (TMEAN x DL), and QTL x DL interaction terms. The RF module had an evaluation of $R^2$, %RMSE, and %Bias values of 0.75, 10.4, and -1.1, respectively across locations (Fig. 2-3 B). The MSNOD\textsubscript{max} module did not perform as well as the other two modules for the evaluation set ($R^2$, %RMSE, and %Bias values of 0.27, 33.36, and 0.15, respectively across locations; Fig. 2-3 D). It should be noted that a 0.15% bias would only result in over simulation of about 1.0 nodes or less. Our assumption of using ECV values over time periods of planting to current simulation day up to simulated anthesis to predict maximum main stem node number require additional studies. There are likely additional environmental covariates that are affecting the variation in maximum main stem node number that were not considered. A source and sink relationship between photosynthesis and assimilate allocation could also be a driving force behind the final main stem node number. The few replications used ($n = 3$), frequency of measurements for each genotype (a sample per week), and sampling with different plants limit all of these analyses as well. The NAR module predicted node number over
the linear phase of node addition well for the evaluation set but with high bias ($R^2$, %RMSE, and %Bias values of 0.93, 24.64, and 20.51, respectively across locations; Fig. 2-3 F). The somewhat high bias in predicted node number ($N$) is partly due to a propagation of error since predicted rates (each with some bias) were integrated for each day over the course of the simulation. This bias may also be partly due to initialization of the node number to a value of 0.0 at emergence (e.g., we assumed that the first node appears at emergence). Based on intercept values from a node addition rate linear regression analyses, time of first node appearance is likely genotype-by-environment specific. Therefore, an additional module should be added for the duration from emergence to appearance of first node for improved node development simulations.

Integrating the three modules together for the GB-CBM (Fig. 2-1) provides the time series simulation for main stem node numbers for all 187 genotypes across the five sites (Fig. 2-4). The emergence of plants was delayed in ND and can be seen relative to the other sites (Fig. 2-4 ND vs. CT, PA, PO, PR). For determinate genotypes such as Calima, either simulated first flower or maximum main stem node number stopped the addition of main stem nodes. For indeterminate genotypes such as Jamapa, the MSNOD$_{\text{max}}$ module set the maximum main stem node number. The separation of the two groups can clearly be seen and is due to the strong effect of the QTLs called TF2, MSN2, and NAR2. These QTLs either represent the action of the TFL1Y/FIN gene, or the action of separate genes tightly linked to FIN (Fig. 2-4, gray lines). The time series plots for node number showed the observed data from both Calima and Jamapa in CT, in which there were fewer node numbers at later time points, likely due to the extreme temperatures that may have caused failures in node formation for single observed plants sampled (Fig. 2-4, CT). This lower node number was associated with a large range of phenotypic responses found in CT (e.g., increased number of branches).
A comparison of the GB-CBM simulation results for node number across sites with all of the observed data from the 187 RILs shows that the GB-CBM model had fairly good predictions of node number with an average $R^2$, %RMSE, %Bias, and Willmot agreement index of 0.72, 35.28, 15.65, and 0.89, respectively (Fig. 2-5). The overall GB-CBM node number simulation performance was brought down by CT and PR results with $R^2$, %RMSE, %Bias, and Willmot agreement index of 0.62, 54.3, 31.62, and 0.83, respectively in CT and 0.61, 36.58, 18.55, and 0.85, respectively in PR. Crops in CT experienced several days of hot temperatures and thus had a greater variability in their node number compared to other sites PO and PA (Fig. 2-5 A vs. C, D). The poor GB-CBM performances in warm conditions (CT and PR) suggest that additional heat stress modules are needed. The poor GB-CBM performance in ND suggests that additional day length or day length and temperature interaction terms are needed. The reason for the high bias can be explained by the NAR module and was discussed previously but is more prominent in the warmer sites.

The GB-CBM simulated anthesis days after planting (ADAP) fairly well, except in the case of ND, with an average $R^2$, %RMSE, %Bias, and Willmot agreement index values of 0.68, 6.4, -1.99, and 0.88, respectively (Fig. 2-6). The model did a relatively poorer job of capturing ADAP in ND ($R^2 = 0.45$) compared to the other sites (Fig. 2-6), and is likely due to the fact that ND was the only site with long days and limited the accuracy in identifying long-day effects for the RF module (Fig. 2-6 B). The RF module appears to require additional adjustment in the low temperature responses for ADAP since there was some bias (-4.02%) in the prediction of ADAP for the coolest site, Popayan (PO; Fig. 2-6 D).

Discussion

Several approaches to develop gene-based, crop models with varying degrees of complexity have been suggested (White and Hoogenboom 2003). The least complicated
approach is to incorporate additive and epistatic effects as linear models that predict traditional GSPs (genetic coefficients) into crop models, if data are available that include wide ranges in environmental factors and considerable variability in genotypes evaluated. A similar approach was proposed for the dry bean (Phaseolus vulgaris L.) simulation model in DSSAT (CROPGRO-Bean) in which seven genes that were assumed to affect phenology, growth habit, and seed size were used to predict GSPs based on linear functions of the genes, and the GSPs were used with the original functions for computing process rates (White and Hoogenboom 1996). This gene-based model accurately predicted phenology but was unable to accurately predict yield variations (Hoogenboom et al. 1997). Integration of genetic information into other crops such as soybean (Glycine max L.; (Messina et al. 2006)), maize (Zea mays L.; (Reymond et al. 2003; 2004; Chenu et al. 2009)), rice (Oryza sativa; (Gu et al. 2014)), and barley ( Hordeum vulgare L.; (Yin et al. 2000a; 200b; 2004)) resulted in varying degrees of success. For example, GSPs in the CROPGRO-Soybean model were converted into mathematical functions of the (day length-sensitive) $E$ loci and used to simulate the reproductive development behavior of soybean cultivars. The modified model accounted for 75% of the variance in maturity date in independent cultivars from Illinois, USA based on weather data and 4 of the 6 $E$ loci, which were found using SSR-linked markers (Messina et al. 2006). Chenu et al. (2009) modified the APSIM maize model (B. a. Keating et al. 2003) to allow parameters for leaf and silk elongation to be affected by QTL interactions to simulate maize growth under drought conditions. The study was able to construct a QTL network affecting the examined traits, and identify the best combination of traits for yield under the management practices in the experiment. Technow et al. (2015) recently demonstrated the utility of integrating a maize model with approximate Bayesian computation (ABC) algorithm for G and G x E effects to improve genomic prediction. Although the study
used synthetic data, the ABC algorithm improved the maize model prediction accuracies relative to using statistical relationships based on markers alone.

More complex approaches for integrating genetics with crop models have been suggested and include simulating gene expression over the course of plant development or including polypeptide translation information (White and Hoogenboom 2003). The higher level of biological networks/levels may increase errors within the models, and it has been argued that further increasing the level of complexity in regards to the genetics may not be necessary for further improving crop models as breeding tool if they capture the physiological basis of the traits (Yin et al. 2004; Hammer et al. 2010). The approach suggested here would allow flexibility in designing modules at the desired level of biological complexity (also discussed in Yin et al. 2004). New gene-based modules can be built to replace sets of calculations for specific crop growth or developmental processes already in the existing CROPGRO-Bean model. These granular modules would be designed to incorporate G, E, and G x E factors to improve model capabilities to simulate performance of multiple genotypes across a range of environments. These modified models would have the capacity to quantify crop performance when new cultivars are developed or to test existing ones in target environments without having to conduct costly multi-location experiments.

The early stages of plant development are just one part of the growth and development processes of plants but they are important processes that also affect yield. In common bean genotypes with a determinate growth habit, the terminal meristem makes a transition from the vegetative to the reproductive phase thus ending the addition of nodes on the main stem (Ojehomon and Morgan, 1969). In indeterminate common bean genotypes, nodes continue to be added on the main stem after the reproductive phase has begun and this continues (Kwak et al.}
2012) until achieving some maximal node number. The rate of node/leaf addition depends on temperature, the genotype, and CO$_2$ levels (Reddy et al., 1995; Vallejos and Pearcy 1987) and is associated with levels of miR156, squamosal-like proteins and cytochrome P450 genes in Arabidopsis (Schwarz et al. 2008; Wang et al. 2008). Node addition rate and the rate of progress toward flowering were both found to be under genetic and environmental control for the bean RILs used in these studies (Zhang 2015; Bhakta 2015a).

We have constructed a prototype gene-based Common Bean Model (GB-CBM) to simulate early vegetative and reproductive development by integrating dynamic QTL effect models for node addition rate, rate of progression to anthesis, and the maximal main stem node number. The focus of this work was to illustrate an approach for integrating dynamic QTL effect models of daily process rate traits into a modular, dynamic model for predicting early growth and development based on QTL, E, and QTL x E factors previously identified by linear mixed effect statistical models. The component models in this study were based on assumptions that the QTLs, E, and QTL x E are the same factors affecting trait development. For this reason, dynamic simulations of these traits can be accomplished through daily time steps using the daily values of the relevant environmental factors. As a result, the highly influential effects were captured with this approach while maintaining simplicity.

The prototype GB-CBM presented here could be expanded at different levels. First, additional modules for the growth of the main stem could include internode and leaf expansion rates, and rate of addition of branches. Also, the expansion of the GB-CBM for more complex traits that additionally affect yield will require modules for other processes (e.g., photosynthesis, leaf area expansion, seed and pod growth, senescence among others). Second, additional studies are needed to expand the identity of QTLs that had significant effects on the modeled traits.
Further work with diversity panels could expand the QTLs repertoire by identifying new QTLs of those that have already been identified. Finally, the assumption of linearity of environment effects on a trait is another limitation in the current GB-CBM version since most biological processes have nonlinear responses to the environment. For example, many developmental and growth responses including node addition and flowering show a temperature response that is a piecewise function of base and optimal temperatures, such as used in the CROPGRO-Bean model (Jones et al. 2003). Here, the model did a poorer job of simulating node development for the extreme temperatures (cold in PO and hot in CT), so incorporating nonlinear temperature functions for describing the dynamic processes is likely to improve predictions. This would be similar to functional mapping that targets the genes that control growth and development by treating these biological processes as nonlinear dynamic traits rather than static phenotypes (Ma et al. 2002; Wu et al. 2003; Wu and Lin 2006; Yang et al. 2009). Functional mapping identifies and estimates the effect of dynamic QTLs by testing whether these parameters display significant differences among genotypes (Wu and Lin 2006). Incorporating functional mapping approaches may lead to models that better represent the underlying biological responses of dynamic traits.

Management (M) factors such as plant density, water, and nutrients were not included in the presented model. Conceivably, one could extend the approach to include M effects and G x E x M interactions, if phenotype in M conditions were included.

The dynamic processes modeled in this study (NAR(t) and RF(t)) are also included in the existing CROPGRO-Bean crop model. This is one reason that these two processes were selected. However, the functional relationships used in the existing bean model are very different from the equations developed by the dynamic QTL effect approach used in this paper (Eq. 2-7 and Eq. 2-9 for RF(t) and NAR(t), respectively). These differences deserve attention here. In the CROPGRO-
Bean model, neither of these processes was dependent on genes/QTLs as they are in the approach described in this paper. In fact, NAR(t) (called TRIFOL in the CROPGRO-Bean model) was assumed to have the same cardinal temperature dependency for all genotypes. Temperature was the only environmental variable used to compute daily node addition rates; this function is piece-wise linear (using hourly temperatures) with a base temperature of 5 °C, an optimum temperature of 27 °C above which further increases in temperature do not increase node appearance rates, and two other temperature thresholds that describe a slowing rate (above 37 °C) and no development above 45 °C. These two upper thresholds are highly uncertain. Eq. 2-9 includes daily mean temperature, daily solar radiation, day length and four QTLs. Figure A.2 (Supplemental material) shows the temperature effect developed from data in this study to be linear, with a base temperature somewhat lower than 10 °C that varies among genotypes. Additional work is needed to evaluate whether incorporating nonlinear functions for environmental effects in the dynamic QTL effect analyses, based on our knowledge of these effects, would improve results.

Comparison of the dynamic QLT effect model (Eq. 2-7) with the process model used to predict daily rate of progress toward flowering, RF(t), in the CROPGRO-Bean model is somewhat more complex. In the existing bean model, a multiplicative model formulation is used as follows:

\[ RF(t) = MR \cdot f(T_{\text{hour}}(t)) \cdot g(P(t)) \]  

where MR is a maximum daily rate that is a GSP (normally estimated from field data), \( f(T_{\text{hour}}(t)) \) is a nonlinear function of hourly temperature \( T_{\text{hour}}(t) \) and assumed to be the same function for all genotypes, and \( g(P(t)) \) is a nonlinear function of day length (Boote et al. 2003) with GSPs for critical day length (CSDL) and for sensitivity to increases in day length above the
critical threshold (PPSEN). Comparing this equation with Eq. 2-7 shows that 12 QTLs affect RF(t), and of course the equation uses only linear responses to environmental factors and includes interactive terms for temperature and day length. Figure A.1 shows that the dynamic QTL effect model produced similar responses to day length as had been assumed in the existing bean model, with the negative slope of RF(t) being analogous to PPSEN. However, the effect of day length depends on temperature in Eq. 2-7 due to the TMEAN(t)·DL(t) term in the equation, and these effects vary among genotypes. This figure also shows that the relationship with temperature varies with genotype, which had not been considered in the existing bean model.

Based on these results, one may suggest that process models be developed using dynamic QTL effect methods, using prior knowledge of the functional relationships that crop modelers have demonstrated in specific physiological studies (as demonstrated here). For example, one could implement Eq. 2-7 in place of existing computations in the CROPGRO-Bean model for flowering time simulations. One could even introduce hourly temperatures, similar to those used in current crop models. This would result in equations that would be very different from Eq. 2-10, and look more like Eq. 2-7 but use nonlinear instead of linear terms. An alternative to this approach would be to use the linear mixed effect models that can be used by researchers who are exploring G, E, M, and G x E x M interactions. Then, crop modelers could use that information to revise their original modules for different processes as more information is developed, making the terms in the functions that they use, similar to Eq. 2-10, depend on genes and G x E interactions. For example, T_b and T_opt values for the temperature function in Eq. 2-10 can be predicted with identified G, E, and G x E factors. We believe that both of these approaches have merit and should be pursued. One advantage of expanding mixed effect models to model dynamic processes is that software could be created to develop nonlinear mixed effect models of
dynamic processes and this could encourage more involvement of geneticists and bioinformaticians in gene-based crop modeling.

One of the major implications of this study is the critical need for phenotyping data that have wide variations in genetic characteristics combined with wide variations in environmental variables (including temperature and day length). In fact, for this study there were only five environments, which limited the reliability and inference of some of the parameters estimated relative to sensitivities to E and G x E. True model evaluation will require that new field sites that were not used in the calibration process be validated. One promising approach might be to make further use of the many yield trials that are conducted by plant breeders in different states and regions. Additionally, automated phenotyping techniques will improve the feasibility of these efforts. By combining genotype information with phenotypic information, it may be possible to make rapid advances toward more holistic, gene-based crop development, growth, and yield models.

**Conclusion**

Although traditional crop models are able to reproduce some G x E interactive effects on yield through GSPs, they have not adequately represented G x E interactive effects at the level of dynamic growth and development processes. The approach demonstrated here incorporates these important interactions at a process level, which are likely to enrich these G x E interaction effects on yield. Empirical GSPs have to be estimated for every genotype, which is costly and time consuming. We showed that there is potential for quantifying rates of vegetative and reproductive development of crops with dynamic QTL effect models based on G and E information identified from mixed effect statistical approaches, using data from multiple locations, and that this could eliminate the need for one to conduct experiments in multiple locations when a new cultivar is released, if genes of the cultivars are known. Nonlinear,
dynamic, QTL effect models are needed so that functional relationships known to represent effects of E on dynamic processes can be incorporated to better reflect the biology than the current linear format. Although the approach used to model vegetative and reproductive development processes in this study was successful, it is not yet clear whether this approach can improve component modules of other growth processes, such as dry matter growth and partitioning to grain yield. Next generation crop models may not quickly incorporate genetic information to improve all aspects of their performance but it is clear that the approach proposed and demonstrated in this paper will improve simulation of some processes and that it lends itself to incremental evolution toward improved models as new research is conducted. In the future, it should be possible for one to genotype a new cultivar and be able to predict crop performance in a range of environments with good accuracy.
<table>
<thead>
<tr>
<th>Site(^a)</th>
<th>CT</th>
<th>PA</th>
<th>PR</th>
<th>PO</th>
<th>ND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>29 39’ N</td>
<td>03 29’ N</td>
<td>18 28’ N</td>
<td>02 25’ N</td>
<td>47 00’ N</td>
</tr>
<tr>
<td>Longitude</td>
<td>82 06’ W</td>
<td>76 81’ W</td>
<td>61 02’ W</td>
<td>76 62’ W</td>
<td>96 47’ W</td>
</tr>
<tr>
<td>(b)Elevation (m)</td>
<td>60</td>
<td>1000</td>
<td>128</td>
<td>1800</td>
<td>280</td>
</tr>
<tr>
<td>Previous Culture</td>
<td>Fallow</td>
<td>Beans</td>
<td>Beans</td>
<td>Fallow</td>
<td>Wheat</td>
</tr>
<tr>
<td>Soil texture</td>
<td>Sand</td>
<td>Clay</td>
<td>Clayey Kaolinite</td>
<td>Medium Loam</td>
<td>Silt/Clay Loam</td>
</tr>
<tr>
<td>Fertilization ([\text{kg} , \text{ha}^{-1}])</td>
<td>N-P-K:136-60-112</td>
<td>40 (Urea)</td>
<td>55(N-P-K:10-10-10)</td>
<td>N-P-K:129-96-80.3</td>
<td>No fertilizer</td>
</tr>
<tr>
<td>Irrigation</td>
<td>Central Pivot</td>
<td>Rain Fed</td>
<td>Drip</td>
<td>Rain Fed</td>
<td>Rain Fed</td>
</tr>
<tr>
<td>Plant Density ([\text{plants} , \text{m}^{-2}])</td>
<td>4.3</td>
<td>3</td>
<td>3.9</td>
<td>4.3</td>
<td>3.3</td>
</tr>
<tr>
<td>Row spacing ([\text{cm}])</td>
<td>90</td>
<td>120</td>
<td>100</td>
<td>90</td>
<td>150</td>
</tr>
<tr>
<td># of blocks</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td># of genotypes</td>
<td>168</td>
<td>174</td>
<td>128</td>
<td>178</td>
<td>176</td>
</tr>
<tr>
<td>Measurement frequency</td>
<td>weekly</td>
<td>Weekly</td>
<td>weekly</td>
<td>weekly</td>
<td>weekly</td>
</tr>
</tbody>
</table>

\(^a\)Abbreviations of CT, PA, PR, PO, and ND represent corresponding sites of Citra, FL (CT); Palmira, Colombia (PA); Popayan, Colombia (PO); Isabela, Puerto Rico (PR); and Prosper, North Dakota (ND), \(^b\)Meters above sea level
Table 2-2. The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for 1/duration from emergence to flowering (RF)

<table>
<thead>
<tr>
<th>Significant Term</th>
<th>Parameter ID</th>
<th>Estimated value (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RF</td>
<td>RF(_b)</td>
<td>0.029 (1.5E-4)</td>
</tr>
<tr>
<td>TMEAN(^a)</td>
<td>RF(_a1)</td>
<td>7.5E-4 (3.6E-5)</td>
</tr>
<tr>
<td>SRAD(^a)</td>
<td>RF(_a2)</td>
<td>-7.3E-6 (1.4E-5)</td>
</tr>
<tr>
<td>DL(^a)</td>
<td>RF(_a3)</td>
<td>-2.2E-3 (8.9E-5)</td>
</tr>
<tr>
<td>TMEAN x DL</td>
<td>RF(_a4)</td>
<td>-3.3E-4 (2.3E-5)</td>
</tr>
<tr>
<td>TF1</td>
<td>RF(_b1)</td>
<td>9.8E-4 (1.1E-4)</td>
</tr>
<tr>
<td>TF2</td>
<td>RF(_b2)</td>
<td>1.7E-3 (1.3E-4)</td>
</tr>
<tr>
<td>TF3</td>
<td>RF(_b3)</td>
<td>-3.9E-4 (1.5E-4)</td>
</tr>
<tr>
<td>TF4</td>
<td>RF(_b4)</td>
<td>2.0E-4 (1.3E-4)</td>
</tr>
<tr>
<td>TF5</td>
<td>RF(_b5)</td>
<td>-1.5E-4 (1.2E-4)</td>
</tr>
<tr>
<td>TF6</td>
<td>RF(_b6)</td>
<td>8.9E-4 (1.2E-4)</td>
</tr>
<tr>
<td>TF7</td>
<td>RF(_b7)</td>
<td>-5.3E-4 (9.9E-5)</td>
</tr>
<tr>
<td>TF8</td>
<td>RF(_b8)</td>
<td>-3.1E-4 (8.9E-5)</td>
</tr>
<tr>
<td>TF9</td>
<td>RF(_b9)</td>
<td>-3.4E-4 (9.0E-5)</td>
</tr>
<tr>
<td>TF10</td>
<td>RF(_b10)</td>
<td>-9.7E-5(9.0E-5)</td>
</tr>
<tr>
<td>TF11</td>
<td>RF(_b11)</td>
<td>2.6E-4 (1.5E-4)</td>
</tr>
<tr>
<td>TF12</td>
<td>RF(_b12)</td>
<td>-6.6E-5 (1.5E-4)</td>
</tr>
<tr>
<td>TF2 x TMEAN(^a)</td>
<td>RF(_c1)</td>
<td>-3.6E-5 (3.4E-5)</td>
</tr>
<tr>
<td>TF3 x TMEAN(^a)</td>
<td>RF(_c2)</td>
<td>6.7E-5 (3.7E-5)</td>
</tr>
<tr>
<td>TF3 x DL(^a)</td>
<td>RF(_c3)</td>
<td>-1.1E-3 (7.1E-5)</td>
</tr>
<tr>
<td>TF5 x TMEAN(^a)</td>
<td>RF(_c4)</td>
<td>5.5E-5 (2.6E-5)</td>
</tr>
<tr>
<td>TF7 x DL(^a)</td>
<td>RF(_c5)</td>
<td>-2.6E-4 (5.9E-5)</td>
</tr>
<tr>
<td>TF12 x SRAD(^a)</td>
<td>RF(_c6)</td>
<td>-6.4E-6 (1.3E-5)</td>
</tr>
<tr>
<td>TF12 x DL(^a)</td>
<td>RF(_c7)</td>
<td>-3.9E-4 (5.8E-5)</td>
</tr>
</tbody>
</table>

\(^a\): Mean values across sites for TMEAN[^\(^\circ\) C] : SRAD[MJ·d\(^{-1}\)] : DL[hr] are 21.35 : 18.31 : 12.7, respectively

\(^b\): Estimated values are attained from non-linear least squares algorithm
Table 2-3. The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for maximum number of nodes along the main stem (MSNOD$_{\text{max}}$) module.

<table>
<thead>
<tr>
<th>Significant Term</th>
<th>Parameter ID</th>
<th>Estimated value (SE)$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean MSNOD$_{\text{max}}$</td>
<td>MSNOD$_{\text{maxb}}$</td>
<td>12.37 (0.13)</td>
</tr>
<tr>
<td>TMEAN$^a$</td>
<td>MSNOD$_{\text{maxa1}}$</td>
<td>0.43 (0.05)</td>
</tr>
<tr>
<td>SRAD$^a$</td>
<td>MSNOD$_{\text{maxa2}}$</td>
<td>0.10 (0.03)</td>
</tr>
<tr>
<td>DL$^a$</td>
<td>MSNOD$_{\text{maxa3}}$</td>
<td>1.2 (0.08)</td>
</tr>
<tr>
<td>MSN1</td>
<td>MSNOD$_{\text{maxb1}}$</td>
<td>-0.43 (0.12)</td>
</tr>
<tr>
<td>MSN2</td>
<td>MSNOD$_{\text{maxb2}}$</td>
<td>-3.56 (0.15)</td>
</tr>
<tr>
<td>MSN3</td>
<td>MSNOD$_{\text{maxb3}}$</td>
<td>-0.63 (0.10)</td>
</tr>
<tr>
<td>MSN4</td>
<td>MSNOD$_{\text{maxb4}}$</td>
<td>-0.20 (0.10)</td>
</tr>
<tr>
<td>MSN5</td>
<td>MSNOD$_{\text{maxb5}}$</td>
<td>-0.60 (0.10)</td>
</tr>
<tr>
<td>MSN6</td>
<td>MSNOD$_{\text{maxb6}}$</td>
<td>0.32 (0.12)</td>
</tr>
<tr>
<td>MSN2 x TMEAN$^a$</td>
<td>MSNOD$_{\text{maxc1}}$</td>
<td>-0.08 (0.05)</td>
</tr>
<tr>
<td>MSN2 x SRAD$^a$</td>
<td>MSNOD$_{\text{maxc2}}$</td>
<td>-0.05 (0.04)</td>
</tr>
<tr>
<td>MSN2 x DL$^a$</td>
<td>MSNOD$_{\text{maxc3}}$</td>
<td>-0.62 (0.09)</td>
</tr>
<tr>
<td>MSN6 x TMEAN$^a$</td>
<td>MSNOD$_{\text{maxc4}}$</td>
<td>-0.02 (0.05)</td>
</tr>
<tr>
<td>MSN6 x SRAD$^a$</td>
<td>MSNOD$_{\text{maxc5}}$</td>
<td>0.01 (0.03)</td>
</tr>
</tbody>
</table>

$a$: Mean values across sites for TMEAN[$^\circ$ C] : SRAD[MJ·d$^{-1}$] : DL[hr] are 21.85 : 18.74 : 12.81, respectively

$b$: Estimated values are attained from non-linear least squares algorithm running the algorithm presented in Eq. 7
Table 2-4. The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for main stem node addition rate (NAR)

<table>
<thead>
<tr>
<th>Significant Term</th>
<th>Parameter ID</th>
<th>Estimated value (SE)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean NAR</td>
<td>--</td>
<td>0.252 (4E-3)</td>
</tr>
<tr>
<td>TMEAN&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NAR&lt;sub&gt;a1&lt;/sub&gt;</td>
<td>2.0E-2 (5.7E-4)</td>
</tr>
<tr>
<td>SRAD&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NAR&lt;sub&gt;a2&lt;/sub&gt;</td>
<td>-7.9E-4 (3.2E-4)</td>
</tr>
<tr>
<td>DL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NAR&lt;sub&gt;a3&lt;/sub&gt;</td>
<td>4.4E-3 (8.0E-4)</td>
</tr>
<tr>
<td>NAR1</td>
<td>NAR&lt;sub&gt;b1&lt;/sub&gt;</td>
<td>-6.0E-3 (1.0E-3)</td>
</tr>
<tr>
<td>NAR2</td>
<td>NAR&lt;sub&gt;b2&lt;/sub&gt;</td>
<td>7.0E-3 (1.0E-3)</td>
</tr>
<tr>
<td>NAR3</td>
<td>NAR&lt;sub&gt;b3&lt;/sub&gt;</td>
<td>8.2E-3 (9.2E-4)</td>
</tr>
<tr>
<td>NAR4</td>
<td>NAR&lt;sub&gt;b4&lt;/sub&gt;</td>
<td>-4.5E-3 (9.0E-4)</td>
</tr>
<tr>
<td>NAR1 x DL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NAR&lt;sub&gt;c1&lt;/sub&gt;</td>
<td>-1.9E-4 (6.6E-4)</td>
</tr>
<tr>
<td>NAR2 x TMEAN&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NAR&lt;sub&gt;c2&lt;/sub&gt;</td>
<td>2.1E-3 (5.2E-4)</td>
</tr>
</tbody>
</table>

<sup>a</sup>: Mean values across sites for TMEAN[° C] : SRAD[MJ·d<sup>-1</sup>] : DL[hr] are 21.51 : 17.38 : 12.74, respectively  
<sup>b</sup>: Estimated values are attained from non-linear least squares algorithm
Figure 2-1. The framework of the gene-based Common Bean Model (GB-CBM) with the modules of rate of progress from emergence to flowering (RF), maximum node number on the main stem (MSNODmax), and main stem node addition rate (NAR). The input files include the daily weather and genotype information. Emergence days after planting and last day of experiment for a location (END) set simulation run time for a genotype at a site. Time to flowering (TF) is the state variable that accounts for phenology (anthesis). JC28 is the QTL region in the model that defines determinacy of a genotype for this model.
Figure 2-2. The QTL that were identified from multi-environment composite interval QTL mapping for time from planting to flowering (TF), maximum number of nodes along the main stem (MSNOD\textsubscript{max}), and node addition rate (NAR) in the common bean RI population. $T_{Fi}$ are markers (Bhakta 2015a) that were used in the RF module, $MSN_i$ are markers that were used in the MSNOD\textsubscript{max} module, and $NAR_i$ are markers that were used in the NAR module (Zhang 2015). Markers with QTL x E are denoted with the * symbol. Bars denote the 1 LOD intervals while whiskers denote the 2 LOD intervals from the peak LOD value for each identified QLT marker.
Figure 2-3. The simulated with daily time steps from the RF (anthesis days after emergence) (A and B), MSNOD$_{\text{max}}$ (maximum main stem node number) (C and D), and node number during the linear phase of node addition predicted by NAR (node addition rate) (E and F) module versus observed data with 1:1 lines are shown for the calibration set (A, C, E) and evaluation set (B, D, F) across the five sites (CT, ND, PA, PO, PR). The evaluation set included the parents, Calima (CAL) in blue and Jamapa (Jam) in orange and RILs in grey with 14 additional lines for RF and MSNOD$_{\text{max}}$, and 7 additional lines for NAR. The analyses of these plots included $R^2$, %RMSE, and %Bias.
Figure 2-4. Simulation results using the GB-CBM to predict main stem node number over time for 187 genotypes (RILs), across the five field sites (CT, Citra, ND, North Dakota, PA, Palmira, PO, Popayan, PR, Puerto Rico) with observed data for the parents Jamapa (JAM) in triangle and Calima (CAL) in circles. The gray lines represent the RILs with them segregating based on JC28 QTL.
Figure 2-5. Simulated versus observed plots of main stem node number using the GB-CBM across time for 187 genotypes (RILs) by field sites (Citra (A), North Dakota (B), Palmira (C), Popayan (D), Puerto Rico (E)) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are black, while indeterminate genotypes are grey. The parameters were estimated across all sites but plotted to compare performance at each site. The lines represent 1:1 relationships.
Figure 2-6. Simulated versus observed plots of Anthesis (days after planting; ADAP) using the GB-CBM for 187 genotypes (RILs) by field sites (Citra (A), North Dakota (B), Palmira (C), Popayan (D), Puerto Rico (E)) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are black, while indeterminate genotypes are grey. The parameters were estimated across all sites but plotted to compare performance at each site. The lines represent the 1:1 relationships.
CHAPTER 3
INTEGRATION OF A MAIN STEM LEAF AREA MODULE

Introduction

Mechanistic crop models simulate plant growth and development as a way to estimate yield for given conditions (e.g., cultivar, soil, climate, and management). These computational tools play vital roles for farmers, policy makers, and economists for field management, resource management, as well as in decision support to meet the rising global food demand. Many components within these models interact at different biophysical levels and complexities to affect yield predictions. An important component for yield simulation is photosynthesis and assimilate production which are dependent on leaf area growth and development. As an example, an ensemble of 23 maize models were used to simulate yield predictions in Lusignan, France; Ames, USA; Rio Verde, Brazil; and Morogoro, Tanzania as part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Bassu et al. 2014). These models were initially used to predict yield after calibration with only soil, management, and crop phenology information. Then additional soil condition, and in-season measurements were provided for further model calibration. The higher level of data allowed for finer adjustments of final leaf length, number, specific leaf area, and increased the yield simulation accuracy across the models. Therefore, accurate simulations of leaf area (i.e., addition and expansion rate of leaves) during the growing season is a fundamental component for any crop model (Lizaso et al. 2003).

Plants adjust their rate of leaf appearance and expansion based on a variety of environmental and developmental signals through different mechanisms (Granier et al. 2009). These mechanisms include cell cycle regulation (Beemster et al. 2003; Gonzalez et al. 2012), tissues extensibility (Pien et al. 2001), as well as turgor and sugar signaling (Gibson 2005). Individual leaf sizes have been shown to be adjusted through cell number (Chiera et al. 2002;
Rymen et al. 2007) or cell size (Radin and Eidenbock 1984; Lecoeur et al. 1995). Both cell number and size in leaves have also been shown to be affected by environmental cues such as temperature and water availability (Durand et al. 1995; Taylor et al. 2003; Kavanová et al. 2006; Kavanová et al. 2008). Leaf development in dicotyledonous leaves such as the common bean have been separated into three phases: proliferation (i.e., exponential cell division and tissue expansion), expansion, and maturation (Beemster et al. 2005). Environmental stresses have the greatest effect on final individual leaf size when they occur during the initial phase (i.e., proliferation). Water deficit during the proliferation phase has been shown to affect leaf expansion rate for 2 weeks even after an irrigation event (Lecoeur et al. 1995; Granier et al. 1999a). Additionally, the reduction in incident light during the same timeframe of leaf development has also been shown to affect the rate of expansion until the end of leaf growth (Granier et al. 1999b). These studies demonstrate that environmental conditions during specific time points of a leaf’s life cycle have great effects on leaf size over time, and consequently the plant canopy leaf area. For models to more accurately simulate canopy leaf area as affected by environmental variations or stresses during the growing season, leaf expansion and final leaf size should be estimated for each individual node.

There are a variety of methods of depicting canopy leaf area in crop models on daily or thermal time scales. Crop models such as the DSSAT CROPGRO-Bean model (Hoogenboom et al. 1994) simulate leaf area of crops based on the leaf area index (LAI), where LAI is defined to be total vegetative leaf area over ground area covered by the plant. LAI values are then used with solar radiation and incident angles to estimate photosynthetic active radiation (PAR) capture, and further simulate assimilate production. The simulation of LAI is itself affected by assimilate production, the phenological stage of the plant, as well as source-sink relationships. Not
surprisingly, many have found the accurate simulation of leaf area as one of the important components in crop models to accurately predict crop yield (Carberry, 1991; Carberry et al. 1989; Keating et al. 1992; Lizaso and Ritchie, 1997). It is possible that accurate leaf area models can be built with more detail in their algorithms to include canopy spatial variability (i.e., leaf area distribution along the canopy), environmental (E), physiological, gene-by-environment (G x E) interactions, and stressors as they affect cell division and growth as leaves are formed. Lizaso et al. 2003 demonstrated an approach to modify the CERES-Maize model to simulate cultivar-specific leaf area growth and development variations of each individual leaf for maize. While the modified model showed improvements for LAI estimations, the model framework still relied on empirical genotype specific parameters (GSP)s rather than genetic marker information. These GSPs depict cultivar-specific biophysical process or trait variations (e.g., leaf number, size, and rate of expansion) of different cultivars grown on a range of soil, climate, and management conditions. There are inherent limitations of model that operate on GSPs to depict the genetic (G) and G x E interaction effects at the level of individual processes that are considered in the models. Concretely, to estimate GSPs for a new genotype(s), multi-environment experiments with detailed crop growth and development measurements are required. This process is time consuming, costly, and limits the utility of crop models in plant breeding programs since breeders grow thousands of lines for a given breeding trial and have limited resources to collect the detailed ecophysiological measurements necessary to estimate these GSPs. Furthermore, plant breeders would benefit most from crop models that are capable of providing decision support on target selection prior to growth trials (Langridge and Fleury 2011). The processes and the parameters in existing model frameworks can, however, be used to guide the incorporation of
genes at a process level, including interactive effects among G, environment (E), and management (M).

Here, we describe a main stem leaf area module (MSLA) that was integrated into the gene-based common bean model (GB-CBM) presented in chapter 2. This module allows the GB-CBM to simulate main stem leaf area growth on each node position for the common bean as affected by the genotype of the crop (represented by QTLs (quantitative trait loci)), its response to the environment, and genotype-by-environment interactions (QTL x E). The MSLA module is comprised of a dynamic, QTL-effect model that estimates the potential leaf area for nodes on the main stem, a potential leaf size by relative node position partitioning model, and a logistic leaf area expansion model. The MSLA module presented here is just one approach to modeling leaf development, and it is expected that this can be combined with additional modules (e.g., photosynthesis) to simulate assimilate production over contrasting or hypothetical environments. These modules along with dynamic QTL-effect modules for node addition rate (NAR), final main stem node number (MSNODmax), and rate of progress toward first flower (RF) described in chapter 2 can simulate early vegetative growth of the common bean and the timing of reproductive development.

Materials And Methods

Plant Material and Field Sites

Details on plant material (including bean genotypes), field sites, phenotyping, and molecular marker were reported previously (Chapter 2). In this study, a leaf area module was developed by using the trifoliate leaf area (cm$^2$) on the main stem over time to model leaf development. For Citra (CT), North Dakota (ND), Palmira (PA), and Popayan (PO), the leaves of the first five nodes ($N$: 1~5) were recorded separately ($LA_N$) and the leaves beyond the fifth node were consolidated and recorded for the first five harvests. After the fifth harvest, total leaf
area on the main stem was recorded as main stem leaf area (MSLA). Studies in Puerto Rico (PR) did not separate leaves on the main stem and only recorded MSLA starting from the first harvest and continuing through the final harvest. Therefore, PR data on MSLA was not used in model development. A total of 171 of the 187 recombinant inbred lines (RIL) grown in CT, ND, PA, and PO were used for model development and calibration. The remaining 16 RILs from CT, ND, PA, and PO along with the PR data (MSLA) were used for model evaluation.

**Linear Mixed Effect Model**

The procedures to build the linear mixed effect model for leaf area follow the methodology reported by (Boer et al. 2007) and was also described in detail in Chapter 2. Here, a linear mixed-effect model was developed to evaluate the maximum potential leaf area of a node on the main stem, and used to determine the magnitude of the G (genetic via QTLs), E (environment), and G x E interactive effects on the trait. Briefly, the procedure consists of three steps: simple interval mapping (SIM), composite interval mapping (CIM), and QTL environment expression analysis. SIM identifies a variance-covariance matrix with the possibility of heterogenetic genetic variance across individual sites and heterogenetic genetic correlation amongst pairs of sites by moving a putative QTL along the genome at each marker position. CIM detects site-specific, QTL effects by including molecular markers detected from SIM as well as flanking marker values as predictors. QTL environment expression analyses detects specific environmental factors or combinations of factors being examined that best reproduce the variations in site-specific QTL effects.

The linear mixed effect model developed here predicts the potential (maximum) leaf area for a node on the main stem termed main stem leaf area max (MSLAmax). The maximum main stem leaf area for each node position (MSLA\textsubscript{max,\textit{N}}) for a given genotype at each site was calculated. MSLA\textsubscript{max} was set to the maximum among MSLA\textsubscript{max,1-5}, and the node position at
which MSLAmax was observed (Nlmax) was recorded. Nlmax was used to compute environment factor values for the examined trait. The observed MSLAmax values can be attributed to two stages of leaf development: cell division during primordia development, and cell filling resulting in leaf expansion during the vegetative growth stage (Granier et al. 2009; Beemster et al. 2005). A study on early vegetative growth in soybean found that cells for the third trifoliate (LA3) were formed in the shoot apical meristem (SAM) 4 days into the experiments while LA3 appeared (~10cm²) on the 16th day of experiment (Chiera et al. 2002); indicating cells for LA3 were formed in the SAM two nodes prior to the appearance of LA3. Thus, two nodes prior to the appearance of Nlmax was assumed to be the time frame for the primordial stage of cell formation in the SAM for the leaves with the largest observed leaf area. The environmental conditions from the time when two nodes prior to the appearance of Nlmax to observation of MSLAmax was assumed to affect the cell division and filling as well as leaf expansion, and attribute to the observed potential leaf area. Therefore, environment factors used in the QTL environment expression analysis were the average values from appearance of Nlmax-2 (i.e., starting at the time two nodes prior) to the observation of MSLAmax for a given genotype at a site. The environment factors used in this module were mean temperature (TMEAN, °C), maximum temperature (TMAX, °C), minimum temperature (TMN, °C), day-night temperature difference (Tdif, °C), solar radiation (SRAD, MJ m⁻² d⁻¹), and day length (DL, h).

This study used 171 RILs at the four sites (CT, ND, PA, and PO) to developed a linear mixed effect model. The procedure was based on multi-environment, single trait analysis with a first-order factor analytic error variance-covariance matrix. The methodology was applied to fit statistical models for MSLAmax to identify significant QTL, E, and QTL x E factors using GenStat 15th edition (Payne et al. 2012).
Dynamic QTL Effect Model

The dynamic QTL effect model for MSLAmax is presented in Eq. 3-1:

\[
\text{MSLAmax}(N) = \text{MSLAmax}_0 + y_{ai} \sum_{i=1}^{E} (E_i((N-2):N) - \overline{E_i}) + y_{bj} \sum_{j=1}^{QTL} \text{QTL}_j + y_{ck} \sum_{i=1}^{QTLxE} \text{QTL}_i \cdot \left(E_i((N-2):N) - \overline{E_i}\right)
\]

(3-1)

where the maximum potential leaf area of a node on the main stem at the time the \(N^{th}\) node is simulated (MSLAmax\((N)\)) is predicted with mean observed MSLAmax across sites (MSLAmax\(_0\)); trait environment effect parameter \((y_{ai})\) for the \(i^{th}\) environment covariate (ECV); mean ECV values between the simulation time of two nodes prior and current node number \((E_i((N-2):N))\); ECV means across sites for the trait \((\overline{E_i})\); trait QTL effect parameters \((y_{bj})\) for the \(j^{th}\) QTL marker; QTL marker values \((\text{QTL}_j)\); trait QTL x E parameters \((y_{ck})\); and trait QTL x E effects \((\text{QTL}_i \cdot \left(E_i((N-2):N) - \overline{E_i}\right))\).

Main Stem Leaf Area by Node Position

The maximum leaf area on the first five nodes of the main stem for the parents occurred between the 3\(^{rd}\) and 5\(^{th}\) node, which depended on the site (Fig. 3-1). For the model, we set the 4\(^{th}\) node to be the location of the largest leaf (Nlmax) based on the fact that, on average, the largest leaf was observed at this position across the four field sites. The leaf size partitioning of potential leaf area at individual nodes has been developed on similar studies on soybean models (Hofstra et al.1977). MSLAmax\(_N\) is computed with Eq. 3-2:

\[
\text{MSLAmax}_N = \begin{cases} 
\text{MSLAmax} - \text{MSLAmax} \cdot (p_1 \cdot (\text{Nlmax} - N)) & \text{if } N \leq \text{Nlmax} \\
\text{MSLAmax} - \text{MSLAmax} \cdot (p_2 \cdot (\text{Nlmax} - N)) & \text{if } N > \text{Nlmax} \\
\text{MSLAmax}_N = \text{MSLAmax}_1 & \text{if } \text{MSLAmax}_N < \text{MSLAmax}_1 
\end{cases}
\]

(3-2)

where potential leaf area for a node position \(N\) is linearly regressed from MSLAmx with \(p_1\) for nodes less than or equal to Nlmax and \(p_2\) for nodes after Nlmax. To prevent negative MSLAmx\(_N\)
values, we set the smallest MSLAmax_N to be that of the first node position. The parameter value for p₁ was the slope of the linear regression of observed MSLAmax₁₋₄ leaf area [cm²] over node positions (i.e., nodes 1~4) for all 171 RILs across the 4 sites and had an average value of 0.15 [cm² · node number⁻¹] (Fig. 3-2, dotted line) while p₂ was the slope of the linear regression of observed MSLAmax₄₋₅ leaf area [cm²] over node positions (i.e., nodes 4 and 5) and had the average value of 0.35 (Fig. 3-2, solid line) for all genotypes in CT, ND, PA, and PO.

Logistic growth models have been used to model leaf area (Niklas et al. 1994; West et al. 2001; Jiang et al. 2015). Therefore, we utilized this model in our leaf area module. Equation 3-3 is a modified logistic model for leaf area growth rate on a given main stem node:

\[
\frac{\Delta L_A_N}{\Delta t} = a \cdot L_A_N(t) \cdot \left(1 - \frac{L_A_N(t)}{MSLAMAX_N}\right) \cdot NAR(t) \quad \text{for} \ (N \leq N(t) \text{ and } L_A_N(t) \leq MSLAMAX_N) \quad (3-3)
\]

where the change in leaf area (\(\Delta L_A_N\), cm² day⁻¹) for a node position N is calculated with the expansion shape parameter (\(a\), plastochron⁻¹) estimated by Eq. 3-5; leaf area by node position on simulation day (\(L_A_N(t)\), cm²); potential leaf area for the node position (MSLAMAX_N, cm²) predicted by Eq. 3-1 and partitioned by Eq. 3-2; and the rate of node appearance on simulation day (NAR(t), nodes day⁻¹ i.e., plastochrons). Leaf area expansion rate by node position is calculated up to the number of nodes on the main stem on simulation day (N(t)), and as long as the leaf area is less than the predicted potential area for a node position. Simulated leaf area for a node position on each day (\(L_A_N(t)\)) was obtained by numeric integration with the computed rate on each daily time step, and is computed up to the estimated MSLAMAX_N for each node position (Eq. 3-4).

\[
L_A_N(t) = L_A_N(t-1) + \Delta L_A_N(t) \cdot \Delta t \quad \text{for} \ (N \leq N(t) \text{ and } L_A_N(t-1) \leq MSLAMAX_N) \quad (3-4)
\]

The leaf expansion shape parameter (\(a\)) is presented in Eq. 3-5

\[
a = ln \left \{ \frac{L_A(N)(MSLAMAX-LA_0)}{MSLAMAX-LA_0-LA(N)-LA_0} \right \} \cdot \frac{1}{N} \quad (3-5)
\]
with mean MSLAmax ($\overline{MLA_{max}}$) for all genotypes across the 4 sites; leaf area on a node for a
given node number on the main stem (LA(N), cm$^2$ plastochron$^{-1}$); initial leaf area (LA$_0$, cm$^2$)
here assumed to be 10 cm$^2$. On average, leaves fully expanded during the time for 4 additional
nodes to appear for the 171 training RIL set across the four sites. The leaf expansion shape
parameter ($a$) was solved for all genotypes at the four sites at the point when there was 95% leaf
expansion with 3.8 additional nodes ((LA(3.8) = 0.95·MLAmax)). Thus, “$a$” was calculated to
be an average of ~1.5 [plastochron$^{-1}$] for the model to simulate full leaf expansion in 4
plastochrons.

**Gene-Based Common Bean Model Framework Update**

The structural layout of the modified GB-CBM is presented in Fig. 3-3, where a main
stem leaf area (MSLA) module is integrated to the GB-CBM presented in chapter 2 to simulate
main stem leaf area by successive node positions (LA$_N$) over time in addition to day of first
flower, node numbers on the main stem over time, and final number of main stem nodes. The
MSLA module contains the dynamic QTL effect model for the potential leaf area for nodes on
the main stem (MSLAmax, Eq. 3-1); the potential leaf size by relative node position partitioning
model (MSLAmax$_N$, Eq. 3-2); and the logistic leaf area growth model (Eq. 3-3). The GB-CBM
follows the identical initialization and imbedded genotype and site looping procedure described
in chapter 2 to simulate traits for the genotype population across multiple sites.

The MSLA module is initiated after the other modules (i.e., after the rate of progress
towards flowering (RF), the final main stem node number (MSNODmax), and the main stem
node addition rate (NAR) modules). These modules are implemented on each simulation day
after emergence. After simulated first flower, the NAR module runs for all genotypes but
numeric integration for main stem node number ($N(t)$) only occurs for indeterminate genotypes
up to the predicted final node number (MSNODmax). In the GB-CBM, the QTL (i.e., JC28) region that contains the FIN/TFL1Y gene defines the determinacy of a genotype as described in chapter 2. If a new node on the main stem is simulated (floor(N(t)) > floor(N(t-1))), the MSLA module sets initial leaf area here assumed to be 10cm² and predicts MSLAmax with the dynamic QTL effect model (Eq. 3-1) based on the allelic makeup at relevant QTLs and mean environmental values relevant to N(t). For the first two nodes, average environmental values are calculated from planting to current simulation day. For node positions at or after 3, average environmental values are calculated from the simulation day when two nodes prior to the current node number was formed to current simulation day. MSLAmaxN is further predicted with the potential leaf area by relative node position partitioning model (Eq. 3-2) given the value of MSLAmax and the difference between node N(t) and the node set to have the largest leaf (Nlmax); MSLAmaxN is then locked in for node position N. After MSLAmaxN is set, the MSLA module numerically integrates for leaf area expansion on each node position (Eq. 3-4) with the rates of leaf area expansion (Eq. 3-3) on each daily time step up to the predicted MSLAmaxN value. Thus, the modified GB-CBM is able to simulate main stem leaf area by node position for a variety of genotypes across multiple environments.

**Model Calibration And Evaluation**

The dynamic QTL effect model for MSLAmax (Eq. 3-1) was calibrated using the nonlinear least squares algorithm in the minpack.lm package (Elzhov et al. 2015) using the R programing language (version 3.2.3; R Core Team). Initial values for each parameter were those identified from the linear mixed effect model. The data was divided into two sets in this study: one was for dynamic QTL effect model development and parameter estimation (a training set of 171 RILs across four sites (i.e., CT, ND, PA, and PO)) and the other was for evaluating
predictions using combinations not used in the estimation process (16 RILs) and the data set from PR (152 RILs).

The MSLAmax module parameters were estimated based on minimizing the sum of squared errors (SSQError(MSLAmax)) between observed (MSLAmaxonbs) and simulated maximum leaf area for a node (MSLAmaxsim) across all environments (S), and genotypes (G) (Eq. 3-6).

$$SSQError(MSLAmax) = \sum_{S=1}^{4} \sum_{G=1}^{171} (MSLAmax_{sim} - MSLAmax_{obs})^2$$

(3-6)

where MSLAmaxsim is the simulated maximum leaf area for a node determined by using average environment values between simulated appearance of Nlmax-2 and Nlmax (here, Nlmax was set at 4), and relevant QTL markers for each genotype at a site.

Two sets of model evaluations were performed. One used the observed and simulated MSLAmax values for the two common bean parents as well as an additional 16 RI genotypes at CT, ND, PA, and PO. The other used observed and simulated MSLA values for the entire population grown in PR. Both evaluations were performed with $R^2$, %RMSE, %Bias, and d-statistics (Willmott et al. 1985).

**Results**

**Dynamic QTL Effect Module**

The dynamic QTL effect model for MSLAmax that was based off the included 8 QTLs with 6 of these having QTL x E interactions. The QTLs for this module were designated as MSLAi for QTLs found to affect main stem leaf area, with i denoting the order which the QTLs were named (Fig. 3-4).

The dynamic QTL effect model for the potential leaf area for nodes on the main stem when the Nth node is formed (MSLAmax(N)) is shown below (Eq. 3-7) with estimated parameter values for the dynamic, daily time step model. Here t is the simulation time between the
formation of the N\textsuperscript{th} node and two nodes prior. For N \leq 2, t is between planting and current node number. The parameter IDs, estimated parameter values and standard errors (SE) for the MSLA\textsubscript{max} module are reported in Table 3-1.

\[ MSLA_{max}(N) = 149.9 + 15.25 \cdot (T_{MEAN}(t) - 21.33) - 3.60 \cdot (SRAD(t) - 18.34) + 3.69 \cdot (DL(t) - 13.09) \\
+ 11.02 \cdot MSLA1 - 5.46 \cdot MSLA2 + 7.53 \cdot MSLA3 + 5.35 \cdot MSLA4 \\
+ 19.57 \cdot MSLA5 - 9.98 \cdot MSLA6 - 7.84 \cdot MSLA7 - 2.63 \cdot MSLA8 \\
+ MSLA1 \left(0.82 \cdot (SRAD(t) - 18.34)\right) \\
+ MSLA4 \left(0.66 \cdot (T_{MEAN}(t) - 21.33) + 4.43 \cdot (DL(t) - 13.09)\right) \\
+ MSLA5 \left(2.55 \cdot (T_{MEAN}(t) - 21.33) - 0.50 \cdot (SRAD(t) - 18.34)\right) \\
+ MSLA6 \left(-1.43 \cdot (T_{MEAN}(t) - 21.33)\right) \\
+ MSLA7 \left(-1.57 \cdot (T_{MEAN}(t) - 21.33)\right) \\
+ MSLA8 \left(0.11 \cdot (T_{MEAN}(t) - 21.33)\right) \tag{3-7} \]

The first term on the right hand side of Eq. 3-7 is the average MSLA\textsubscript{max} across sites. The value of 149.9 cm\textsuperscript{2} indicates the average potential leaf area for nodes on the main stem for all genotypes and sites. The 2\textsuperscript{nd} term indicates that a one-degree increase above 12.33 °C in temperature would result in a 15.25 cm\textsuperscript{2} increase from the general mean of 149.9 cm\textsuperscript{2}; provided all other variables were kept at their average values. The potential leaf area will also vary as a function of QTL alleles and their interactions with specific environmental variables as indicated in Eq. 3-7. The *Calima* QTL allele, MSLA1\textsuperscript{Cal}, will have a (+1) coefficient and therefore would increase the potential leaf area by 11.02 cm\textsuperscript{2} from the general mean of 149.9 cm\textsuperscript{2} as a result of that QTL effect. Similarly, the same QTL allele will increase the potential leaf area by 0.82 cm\textsuperscript{2} for a one MJ\textperiodcentered d\textsuperscript{-1} increase in solar radiation above 18.34 MJ\textperiodcentered d\textsuperscript{-1}. In contrast, the *Jamapa* allele, MSLA1\textsuperscript{Jam}, will have the opposite effect.

**Evaluation of GB-CBM**

The MSLA\textsubscript{max} module operating on daily time steps was able to capture the smaller leaf sizes that were observed in PO, and larger leaf sizes in PA (Fig. 3-5 C) since the dynamic QTL
effect model had 5 temperature x QTL interactions. The dynamic QTL effect module had $R^2$, %RMSE, and %Bias values of 0.38, 29.26, and -2.62, respectively across locations (Fig. 3-5 C). The poor prediction in CT can be attributed to high phenotype variability due to heat stress, while the poor prediction in ND could be caused by the model not capturing additional DL x temperature x QTL interactions in ND since ND was the only site with long DL. Across locations the %RMSE of 29.26 indicates that the MSLAmax module has a maximal prediction error of $\pm 90 \text{ cm}^2$ when estimating potential leaf area for nodes on the main stem.

Integrating the four modules together for the GB-CBM (Fig. 3-3) provided the time series simulation of main stem leaf area for all 187 genotypes across the five sites (Fig. 3-6). For determinate genotypes such as Calima, either simulated first flower or maximum main stem node number stopped the addition of main stem nodes. For indeterminate genotypes such as Jamapa, the MSNOD$_\text{max}$ module set the maximum main stem node number. The MSLA and MSLAmax module operating in tandem, then simulated leaf area expansion on each main stem node position. The time series plots for main stem leaf area showed the observed data decreased over time after a maximal value, and at different time points and rates (Fig. 3-6). This drop in leaf area is senescence, and is not part of the MSLA simulation algorithm at this time. The partitioning of potential leaf size by relative node position (Eq. 3-2) accurately simulated potential leaf areas for the first five node positions (Fig. 3-7) across different locations. The dynamic shape of the expansion curve for each node position can be attributed to the expansion shape parameter (a), logistic growth function, as well as the NAR module operating on daily time steps (Eq. 3-3).

A comparison of the GB-CBM simulation results for total main stem leaf area across sites with all of the observed data from the 187 RILs showed that the GB-CBM model had fairly good
predictions of leaf area with an average $R^2$, %RMSE, %Bias, and Willmot agreement index of 0.63, 71.82, 33.68, and 0.85, respectively (Fig. 3-8). Model performance was much better in PA and PO (Fig. 3-8 C and D) with $R^2$, %RMSE, %Bias, and Willmot agreement index of 0.82, 33.19, 8.1, and 0.95, respectively in PA and 0.7, 44.94, 4.16, and 0.9, respectively in PO. While accuracy decreased in CT and ND (Fig. 3-8 A and B) with $R^2$, %RMSE, %Bias, and Willmot agreement index of 0.63, 71.82, 33.68, and 0.85, respectively in CT and 0.64, 84.97, 49.15, and 0.84, respectively in ND. This is a result from the poor MSLAmax module performance in CT and ND (Fig. 3-5 C). When the GB-CBM model simulated a new node position, the MSLAmax module was not able to accurately predict the maximum potential leaf area for a node across the main stem in CT and ND. That inaccurate value was then further partitioned to the potential leaf size for the newly formed node. As a result, the errors for each node position accumulated, and the GB-CBM main stem leaf area simulation accuracy was reduced. Of note, PR was not included in the training data set and can be treated as a model evaluation data set. However, there are limitations to the collected destructive samples. CT trials were subjected to heat as well as biotic stresses and PR had many outlier measurements. Therefore, GB-CBM model evaluation with independent data should be addressed with further field trials. An independent field trial consisting of Calima, Jamapa, with additional seven RILs (RIJC007, RIJC015, RIJC131, RIJC250, RIJC337, RIJC059, and RIJC220) was conducted in Palmira, Colombia at the International Center for Tropical Agriculture in 2016. The data was used to evaluate the GB-CBM’s ability to simulate main stem node number (Appendix Fig. A.3 and Fig. A.4), flowering days after planting (Appendix Fig. A.5), and main stem leaf area expansion of successive nodes (Appendix Fig. A.6 and Fig. A.7).


**Discussion**

There are many types of models that predict leaf area and these range from those that model individual leaves based on their formation at the cellular and expansion levels others that include genetic parameters (GSPs) for leaf expansion that could incorporate QTL information, models that have temporal G and G x E effects on leaf area and dry weight growth curves and models that include 3D canopy light interception. For example, the CERES-Maize model was modified to create the IXIM-Maize model that simulates leaf expansion, longevity, and senescence by each leaf position that takes into account some genetic effect through the addition of three empirical GSPs for leaf area simulations (Lizaso et al. 2003). The original and modified CERES-Maize models were used to simulate field trials for 5 cultivars in Morris, MN, Gainesville, FL, and Honolulu, HI between 1982 and 1986. In general, this modification to the CERES-Maize was successful since the RMSE for LAI was lower for the modified model simulations and ranged from 0.29 to 1.25 when compared to unmodified model. This new model allowed for simulation of leaf area expansion and senescence by node positions as well as cultivar performance variations through GSPs. However, the model framework still relied on empirically estimated GSPs rather than GSPs that were based on genetic marker information as described next.

A common approach to integrate genetics into existing crop models is to predict parameter (GSPs) with regression models of genetic marker values. For example, a rice ecophysiology model was modified to use single sequence repeats (SSR) to estimate an array of model parameters in order to better account for genetic effects on the duration of vegetative growth, maximal plant height, and specific leaf area of new leaves for a family of RILs (Gu et al. 2014). This version of the model was able to simulate yield [g·m⁻²] under well-water conditions accurately with $R^2$ and RMSE of 0.72 and 0.1, respectively. Yield simulations under drought
stressed conditions, however, were less reliable with $R^2$ and RMSE of 0.57 and 0.23, respectively. The researchers assumed that the complex G x E interactive effects on the plant processes would be taken into account by the ecophysiology model framework. However, others have shown the benefits of including G x E interaction effects as informed by marker information. For example, model parameters for maize leaf width, length, and elongation rate were estimated by relevant QTL marker information to account for QTL as well as QTL x E effects on vegetative growth and development under non-stressed growth conditions (Reymond et al. 2004). In addition to QTL and QTL x E, the study was also able to identify QTL x QTL (i.e., epistatic effects) on leaf length and width. The study found the QTLs associated with leaf width did not vary between treatments, while the detected QTL, QTL x E, QTL x QTL effects were different for leaf length. The 3D canopy architecture on light interception for rice as affected by QTLs have also been performed (Xu et al. 2011). The 3D model was able to simulate canopy architecture variations as affected by planting densities to accurately simulate grain yield with an $R^2$ of 0.96. Further work on such studies are still needed to include G x E interactive effects and 3D canopy architecture.

Recently, a model was developed for leaf area and leaf dry weight by modeling the growth trajectory as affected by QTLs over time for the common bean using a subset of the data presented here (Jiang et al. 2015). That study used a method termed functional mapping (Ma et al. 2002; Aubert et al. 2006; Wu and Lin 2006; Yang et al. 2009) to simulate logistic growth rates with inflection point variations for leaf area and dry weights of the population by making the parameters in the models a function of the associated heterochronic QTLs. However, the model was only based on data from PA and PO, Colombia. The accuracy of the simulated
trajectories remains to be analyzed with observation values. In addition, the QTLs and their interactions with logistic equation parameters did not include explicit QTL x E terms.

We have built upon the GB-CBM presented in chapter 2 to simulate main stem leaf area growth as affected by early vegetative growth (i.e., node appearance), reproductive development (i.e., flowering), and G, E, and G x E interaction effects (i.e., a dynamic, QTL-effect model for potential leaf area for a node across the main stem). This approach reflects the physiology and morphology found in plant canopy development. The model can differentiate between canopies with many small leaves and ones with few large leaves; an important distinction when modeling canopy architecture. The GB-CBM also simulates the temporal and environmental variability interactions with genetics as plants form and expand leaves on each node. A possible route for using this model would be to nest the GB-CBM model in the existing DSSAT framework to better estimate leaf area based on genetic makeup of a plant (e.g., the DSSAT CROPGRO-Bean).

While the GB-CBM with the main stem leaf area module could predict MSLA in general, there were several limitations. For example, the assumption of linearity of environment effects on a trait is a limitation of the GB-CBM (see discussion from chapter 2). Here, model simulations of leaf area development in hot and long DL environments (hot in CT and long DL in ND) were less reliable, so incorporating nonlinear temperature and photoperiod response functions for describing the dynamic processes is likely to improve predictions. Additional trials with heat tolerant genotypes may identify QTLs attributing to heat resistance to improve high temperature simulations. Experiments in sites with long DL and cool temperatures may also improve model simulations by identifying additional QTL x temperature x DL interaction effects. Further evaluations are needed to tease out the effects of the reproductive phase (i.e.,
flowering, pod filling, and assimilate allocation and mobilization) on leaf area expansion and senescence. In addition, other limitations on the model have been described in chapter 2.

Despite the limitations to this gene-based crop model, the modular design allows it to readily incorporate a variety of additional information, including the timing of the stressors (e.g., water deficit, incident light reduction) which have been found to affect final leaf sizes greater when the stresses occurred in the early phase of leaf cellular development (Lecoeur et al. 1995; C. Granier et al. 1999a; Granier et al. 1999b). Others have also found P stress to affect the rate of leaf appearance and expansion (Chiera et al. 2002) and incorporating this information can improve the model. To go beyond QTLs, future studies with diversity panels that include SNP (single nucleotide polymorphisms) based genomic maps, and genomic selection algorithms (Technow et al. 2015; Lopez-Cruz et al. 2015) could allow the GB-CBM to cross over to other populations; leveraging the diverse genetic pools of the common bean, and the massive quantities of genomic and phenomic data collected by bean breeders worldwide.

**Conclusion**

Although traditional crop models are able to reproduce some G x E interactive effects on leaf area through GSPs, they have not adequately represented G x E interactive effects nor the temporal environmental variability at the level of dynamic growth and development processes for each leaf of the canopy. Thus, assimilate production under biotic and abiotic stress simulations are less reliable. The approach demonstrated here incorporates some of these important interactions at a process level of each leaf on the main stem, which are likely to enrich these G x E interaction effects on yield and better incorporate stresses in crop models. We demonstrated an approach to simulate leaf area growth by node positions that incorporates dynamic QTL, E, and QTL x E effects via a mixed effect model for potential leaf area, leaf size
partitioning by node position, and a logistic expansion model for leaf growth. The inclusion of these models for main stem leaf area into the other modules (chapter 2) is a step towards building a crop model to predict yield based on G, E, and G X E interactions.
Table 3-1. The significant terms in the dynamic QTL effect model showing the parameter IDs and estimated parameter values with standard errors (SE) for potential leaf area for nodes on the main stem (MSLAmmax) module.

<table>
<thead>
<tr>
<th>Significant Term</th>
<th>Parameter ID</th>
<th>Estimated value (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean MSLAmmax</td>
<td>MSLAmmax₀</td>
<td>149.9 (3.40)</td>
</tr>
<tr>
<td>TMEANᵃ</td>
<td>MSLAmmax₁</td>
<td>15.25 (0.66)</td>
</tr>
<tr>
<td>SRADᵃ</td>
<td>MSLAmmax₂</td>
<td>-3.60 (0.40)</td>
</tr>
<tr>
<td>DLᵃ</td>
<td>MSLAmmax₃</td>
<td>3.69 (1.05)</td>
</tr>
<tr>
<td>MSLA1</td>
<td>MSLAmmax₄₅₆₇₈</td>
<td>11.02 (1.77)</td>
</tr>
<tr>
<td>MSLA2</td>
<td>MSLAmmax₃₂</td>
<td>-5.46 (1.50)</td>
</tr>
<tr>
<td>MSLA3</td>
<td>MSLAmmax₃₃</td>
<td>7.53 (1.47)</td>
</tr>
<tr>
<td>MSLA4</td>
<td>MSLAmmax₄₅₆₇₈</td>
<td>5.35 (1.58)</td>
</tr>
<tr>
<td>MSLA5</td>
<td>MSLAmmax₅₆₇</td>
<td>19.57 (1.79)</td>
</tr>
<tr>
<td>MSLA6</td>
<td>MSLAmmax₅₆₇</td>
<td>-9.98 (1.59)</td>
</tr>
<tr>
<td>MSLA7</td>
<td>MSLAmmax₅₆₇</td>
<td>-7.84 (1.80)</td>
</tr>
<tr>
<td>MSLA8</td>
<td>MSLAmmax₅₆₇</td>
<td>-2.63 (1.84)</td>
</tr>
<tr>
<td>MSLA1 x SRADᵃ</td>
<td>MSLAmmax₇₈</td>
<td>0.82 (0.43)</td>
</tr>
<tr>
<td>MSLA4 x TMEANᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>0.66 (0.65)</td>
</tr>
<tr>
<td>MSLA4 x DLᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>4.43 (1.06)</td>
</tr>
<tr>
<td>MSLA5 x TMEANᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>2.55 (0.62)</td>
</tr>
<tr>
<td>MSLA5 x SRADᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>-0.50 (0.45)</td>
</tr>
<tr>
<td>MSLA6 x TMEANᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>-1.43 (0.61)</td>
</tr>
<tr>
<td>MSLA7 x TMEANᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>-1.57 (0.69)</td>
</tr>
<tr>
<td>MSLA8 x TMEANᵃ</td>
<td>MSLAmmax₉₀₁</td>
<td>0.11 (0.69)</td>
</tr>
</tbody>
</table>

a: Mean values across sites for TMEAN[^[° C] : SRAD[MJ·d⁻¹] : DL[hr] are 21.33 : 18.34 : 13.09, respectively
b: Estimated values are attained from non-linear least squares algorithm runing the algorithm presented in Eq. 3-1
Figure 3-1. Observed leaf area [cm²] for node positions 1-5 (LA1-LA5) for the two parents (Jamapa (right) and Calima (left)) across the four field sites (Citra [CT], North Dakota [ND], Palmira [PA], and Popayan [PO]).
Figure 3-2. Box plot of observed potential leaf area [cm$^2$] for node positions 1~5 (MSLAmx$_{1-5}$) for the training data set of 171 RILs across the four field sites (Citra [CT], North Dakota [ND], Palmira [PA], and Popayan [PO]). Dotted line represent the linear regression of observed MSLAmx$_{1-4}$ values over node positions 1~4, and solid line represent the linear regression of observed MSLAmx$_{4-5}$ values over node positions 4 and 5.
Figure 3-3. The framework of the gene-based Common Bean Model (GB-CBM) with the modules of rate of progress from emergence to flowering (RF), maximum node number on the main stem (MSNOD\textsubscript{max}), main stem node addition rate (NAR), and main stem leaf area (MSLA). The input files include the daily weather and genotype information. Emergence days after planting and last day of experiment for a location (END) set simulation run time for a genotype at a site. Time to flowering (TF) is the state variable that accounts for phenology (anthesis). JC28 is the QTL region in the model that defines determinacy of a genotype for this model.
Figure 3-4. The QTL that were identified from multi-environment composite interval QTL mapping for the potential leaf area for nodes on the main stem (MSLAmx) in the common bean RI population. LAi are markers that were used in the MSLAmx module. Markers with QTL x E are denoted with the * symbol. Bars denote the 1 LOD intervals while whiskers denote the 2 LOD intervals from the peak LOD value for each identified QLT marker.
Figure 3-5. Simulated results of the 171 training set RILs for the seasonal linear mixed effect model (A), dynamic QTL effect model performance with calibration set of 171 RILs (B), and evaluation set of 16 RILs (C) for the MSLAmax module versus observed data with 1:1 lines across the four sites (CT, ND, PA, PO). The evaluation set included the parents, Calima (CAL) in blue and Jamapa (JAM) in orange and RILs in grey with 16 additional lines. The analyses of these plots included $R^2$, %RMSE, and %Bias.
Figure 3-6. Simulation results using the GB-CBM to predict main stem leaf area [cm$^2$] over time for 187 genotypes (RILs), across the five field sites (CT, Citra, ND, North Dakota, PA, Palmira, PO, Popayan, PR, Puerto Rico) with observed data for the parents Jamapa (JAM) in triangle and Calima (CAL) in circles. The grey lines represent the RILs with them segregating based on JC28 QTL.
Figure 3-7. Simulation results using the GB-CBM to predict main stem leaf area [cm²] over time for the first 5 nodes. The leaf area of each node position (1~5) for the parents Jamapa (JAM) on the right and Calima (CAL) on the left are plotted across the four field sites (CT, Citra, ND, North Dakota, PA, Palmira, PO, Popayan) with observed data.
GB-CBM MSLA Simulation Performance

Figure 3-8. Simulated versus observed plots of main stem leaf area using the GB-CBM across time for 187 genotypes (RILs) by field sites (Citra (A), North Dakota (B), Palmira (C), Popayan (D), Puerto Rico (E)) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are black, while indeterminate genotypes are grey. The parameters were estimated across all sites but plotted to compare performance at each site. The lines represent 1:1 relationships.
CHAPTER 4
CLOSING STATEMENT

Given the projections of global population and uncertain climate conditions of the future, efficient global agriculture production and resource management will be critical to safeguarding our food security. Plant breeding organizations currently have statistical tools based on genetics to provide decision support to design cultivars based on target selection in varying environments. However, these statistical methods assess whole season outputs (e.g., yield) rather than the dynamic ecophysiological mechanisms of plant growth and development that underlie the targeted traits (e.g., yield predictions based on canopy leaf area, assimilate production, developmental stage, and source-sink relations). A crop model that takes into account the genetics to predict plant growth and development with a mechanistic approach would be a valuable tool for plant breeders by providing insight on target selection (Langridge et al. 2011).

This thesis developed a novel, mechanistic, crop model (i.e., the gene-based common bean model (GB-CBM)) that included genetics, environmental effects, and gene-by-environment interactions to simulate early main stem vegetative growth and reproductive developmental process.

The GB-CBM simulates main stem node number and leaf area by successive node positions over time, flowering days after planting, and final main stem node number for a recombinant inbred line (RIL) population of the common bean across 5 environments (Citra, FL (CT); Fargo North Dakota (ND); Palmira (PA) and Popayan (PO), Colombia; and Isabel Puerto Rico (PR)). The model takes into account the genetic (G), environmental (E), and gene-by-environment (G x E) interactive effects, and the developmental processes are integrated in such a way to depict the observed dynamic biological growth of the plants. The work demonstrated an approach of developing dynamic, QTL-effect trait modules and further integrating them in a
novel model framework to simulate trait variations of a genotype population in contrasting environmental conditions. Although the presented methodology was successful in simulating early main stem vegetative growth and flowering time, the GB-CBM has certain limitations that are described next.

Example limitations to the GB-CBM range from limitations in the data collected (only 5 sites, with only three destructive samples in some cases, only one long day length etc.) to limits on the assumptions to build and implement the modules within the GB-CBM. As an example, the main stem node addition rate (NAR) and rate of progress towards flowering (RF) modules described in chapter 2 are limited in their assumption that growth and developmental rates have linear responses to temperature gradients. While crop models generally simulate rates of growth with non-linear temperature functions (Parent and Tardieu 2014). As described in detail in chapter 2, the positive bias from the NAR module propagates with each numeric integration for main stem node number (MSNOD); and as a result MSNOD is over-predicted during the growing cycle across sites (Fig. 2-5) and in the model evaluation (Fig. A-4). By including a piece-wise temperature function operating on hourly time-steps with base and optimal temperatures, the values predicted by the NAR module can be reduced in colder temperature (PO) and tapered-off in hotter climates (CT and PR). The RF module on its own also has a slight negative bias in its daily predicted rates of progress toward flowering, and as a result flowering days after planting is under-predicted across sites (Fig. 2-6) and in Colombia in 2016 (Fig. A-5). Furthermore, the under-prediction is more evident in colder temperatures (PO) than warmer climates (CT and PR). Similarly, by incorporating a piece-wise temperature function, it could be possible to increase the daily rates of progress predicted by the RF module with greater magnitudes in colder than warmer temperature conditions.
The final main stem node number (MSNODmax) module, developed as part of this thesis, has several limitations as well. It is unclear the time point at which MSNODmax is set in relation to phonological stages (e.g., flowering) from the training data set (2011), nor has temporal shoot apical meristem cell differentiations been studied in depth for bean. It is also possible that the value of final main stem node number is a result of phenological events (i.e., flowering time) and source-sink relations rather than a dynamic, QTL-effect module. Additional studies on these aspects can help to determine if this is the case. Also, the partitioning of leaf area by relative node position presented in chapter 3 was limited due to several factors. For example, the node position (between 1 and 5) where the largest leaf was observed (Nlmax) varied amongst genotype and across locations in the training dataset. However, there was not enough variations in the node position at which maximum leaf area was found to develop a mixed-effect model. In addition, a study in which leaf area for nodes beyond the 5th node should be conducted to develop a leaf size partitioning by relative node position model that has higher accuracy for node positions beyond the 5th node, and provides further information of the value of Nlmax. Additional dynamic, QTL-effect modules for other traits can be added to the GB-CBM to improve model simulations and expand model capacity. For example, a senescence module can be added to the main stem leaf area module to improve leaf area simulations of the main stem, while branching modules (e.g., branching node addition rate, branching leaf area) can be added to the GB-CBM to expand leaf area simulations to branches.

As we begin to develop these gene-based crop models one aspect that is needed is to expand data collection with novel plant populations, use field trials in contrasting environments, and incorporate single nucleotide polymorphisms (SNP)-based maps. New plant populations with tolerance to different environmental stresses such as heat and drought can be used to not
only identify the genetics attributing to these characteristics, but also explore approaches to allow the GB-CBM to simulate multiple populations in contrasting environments in efficient model frameworks. Growing the RIL population from this study in additional field sites with contrasting environments (e.g., long day-length with cold temperature conditions) may identify additional E, E x E, QTL x E, and QTL x E x E interaction effects on the trait modules included in the GB-CBM. The QTL regions for the plant growth and developmental traits identified with the interval mapping approach in this study can span over many QTL markers within the common bean genome. Each of the QTLs can also contain multiple genes. By incorporating SNP-based maps, the base pair differences attributing to the trait variation can be identified to improve our understanding of the G and G x E interactive effects at a much greater resolution. Genomic selection algorithms rather than interval mapping approaches will be able to take advantage of the high marker resolution of these SNP-based maps.

Although the approach used to develop a model for vegetative and reproductive development processes by building a new model framework was successful, it is not yet clear whether this approach can be implemented on component modules of other growth processes, such as dry matter growth and partitioning. Furthermore, it is unclear the time required and financial costs associated to developing gene-based models such as the GB-CBM for grain yield predictions. A potential route to arrive at yield predictions with mechanistic, gene-based crop models is to leverage the frameworks of existing crop models that already operate on algorithms developed over decades of research. As discussed in chapter 2, the approaches used to simulate MSNOD (i.e., NAR module) and flowering time (i.e., RF module) follow the original numeric integration of rates design of DSSAT and other mechanistic crop models. For these traits, replacing algorithms with the novel, dynamic, QTL-effect modules are much more feasible when
compared with replacing the algorithms for leaf area simulation. There is little overlap in the
methods of simulating leaf area between the existing framework in DSSAT and the presented
GB-CBM. Thus, it would be difficult to readily integrate this module into the crop model
directly. However, there are other methods of parallel computing and automatic model correction
algorithms termed model multifidelity that may be able to retain the crop models well established
framework while adding more genetic information. This method allows for simultaneous
computations of multiple models of varying complexity, and automatically corrects prediction
outputs. Rather than re-writing code from existing crop models with complex frameworks, it
could be possible to externally introduce simpler trait models that take into account G, E, and G
x E interactive effects with model multifidelity methods to introduce the G and G x E effects on
traits underrepresented in current crop models.

To maintain food security in the future, there is a demand in the scientific and plant
breeding community for mechanistic, gene-based crop models. This study demonstrated an
approach in developing such a computational tool. The methodology and model presented
successfully achieved its objectives to simulate dynamic phenotypic variations of a genetic
population across multiple environments, but room for improvement is present to achieve a gene-
based crop model for yield predictions. In the future, it should be possible to directly engineer
cultivars or more efficiently select candidate genotypes between breeding cycles for specific
breeding objectives and target environments as informed by mechanistic, gene-based crop
models; to reduce time and cost for breeding programs and meet global food demands.
Figure A-1. Environment sensitivity analyses for rate of progress toward flowering (RF) with parent lines highlighted. Grey lines represent the RI population. For each day length (DL) sensitivity analysis, mean temperature and solar radiation values were held at 21.35 °C and 18.31 MJ m$^{-2}$ d$^{-1}$, respectively. Simulated rate was then predicted with DL values from 10 to 14 h in increments of 0.5 h. For each solar radiation (SRAD) sensitivity analysis, mean temperature and day length values were held at 21.35 °C and 12.7 h, respectively. Simulated rate was then predicted with SRAD values from 10 to 30 MJ m$^{-2}$ d$^{-1}$ in increments of 0.5 MJ m$^{-2}$ d$^{-1}$. For each mean temperature (TMEAN) sensitivity analysis, solar radiation and day length values were held at 18.31 MJ m$^{-2}$ d$^{-1}$ and 12.7 h, respectively. Simulated rate was then predicted with TMEAN values from 10 to 30 °C in increments of 0.5 °C. For each temperature by day length interaction (TMEAN x DL) sensitivity analysis, solar radiation and day length values were held at 18.31 MJ m$^{-2}$ d$^{-1}$ and 14 h, respectively. Simulated rate was then predicted with TMEAN values from 10 to 30 °C in increments of 0.15 °C.
Figure A-2. Environment sensitivity analyses for node addition rate (NAR) module with parent lines highlighted. Grey lines represent the RI population. For each day length (DL) sensitivity analysis, mean temperature and solar radiation values were held at 21.51°C and 17.38 MJ m$^{-2}$ d$^{-1}$, respectively. Simulated rate was then predicted with DL values from 10 to 14 h in increments of 0.5 h. For each solar radiation (SRAD) sensitivity analysis, mean temperature and day length values were held at 21.51°C and 12.74 h, respectively. Simulated rate was then predicted with SRAD values from 10 to 30 MJ m$^{-2}$ d$^{-1}$ in increments of 0.5 MJ m$^{-2}$ d$^{-1}$. For each mean temperature (TMEAN) sensitivity analysis, solar radiation and day length values were held at 17.38 MJ m$^{-2}$ d$^{-1}$ and 12.74 h, respectively. Simulated rate was then predicted with TMEAN values from 10 to 30 in increments of 0.5. For each temperature by day length interaction (TMEAN x DL) sensitivity analysis, solar radiation and day length values were held at 17.38 and 14 h, respectively. Simulated rate was then predicted with TMEAN values from 10 to 30 °C in increments of 0.15 degrees.
Figure A-3. Simulation results using the GB-CBM to predict main stem node number over days after planning for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with observed data. Determinate observed values are in circles, Indeterminate observed values are in triangles, and the simulated values are represented by the lines.
Figure A-4. Simulated versus observed plot of main stem node number using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are circles, while indeterminate genotypes are triangles. The lines represent the 1:1 relationship.
Figure A-5. Simulated versus observed plot of anthesis days after planting using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are circles, while indeterminate genotypes are triangles. The lines represent the 1:1 relationship.
Figure A-6. Simulation results using the GB-CBM to predict total main stem leaf area (cm\(^2\)) over days after planting for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with observed data. Determinate observed values are in circles, Indeterminate observed values are in triangles, and the simulated values are represented by the lines.
Figure A-7. Simulated versus observed plot of total main stem leaf area up to senescence using the GB-CBM for Calima (CAL), Jamapa (JAM), and 7 RILs in Palmira (PA) with $R^2$, %RMSE, %Bias, and Willmot agreement index. Determinate genotypes are circles, while indeterminate genotypes are triangles. The lines represent the 1:1 relationship.
LIST OF REFERENCES


Jiang, Libo, Jose A Clavijo, Lidan Sun, Xuli Zhu, Mehul S Bhakta, Salvador A Gezan, Melissa Carvalho, C Eduardo Vallejos, and Rongling Wu. 2015. “Plastic Expression of Heterochrony Quantitative Trait Loci (H QTLs) for Leaf Growth in the Common Bean (Phaseolus Vulgaris”).”


BIOGRAPHICAL SKETCH

Christopher Hwang received his B.S. in agricultural and biological engineering from the University of Illinois at Urbana Champaign in 2014. After his undergraduate education, he continued his graduate education in the Department of Agricultural and Biological Engineering at the University of Florida.