Machine learning analytics for predictive breeding

Zhanyou Xu

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Recommended Citation
https://lib.dr.iastate.edu/etd/18248

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Machine learning analytics for predictive breeding

by

Zhanyou Xu

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
William D. Beavis, Co-major Professor
Steven Cannon, Co-major Professor
Silvia Cianzio
Jode Edwards
Julie A. Dickerson

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

Copyright © Zhanyou Xu, 2020. All rights reserved.
# TABLE OF CONTENTS

**ABSTRACT**.....................................................................................................................................v  

**CHAPTER 1. GENERAL INTRODUCTION** ................................................................................1  
Dissertation organization........................................................................................................... 4  

**CHAPTER 2. LITERATURE REVIEW** .........................................................................................6  
Part 1: An abiotic trait: Soybean iron deficiency chlorosis (IDC) ............................................ 6  
Part 2: A biotic trait: Soybean sudden death syndrome (SDS) ................................................ 19  
Interaction network of genes associated with traits of interest .............................................. 23  
Part 3: Spatial data analyses for adjusting phenotypic data..................................................... 24  
What is spatial data analysis? ............................................................................................. 25  
Spatial pattern recognition and phenotypic adjustment to improve prediction accuracy .............................................................................................................................. 25  
Spatial adjustment for phenotypic data .............................................................................. 27  
Spatial autoregressive regression (SAR) models ............................................................... 28  
Model comparison and Lagrange Multiplier Test (LMT) .................................................. 30  
Application of SAR models ............................................................................................... 31  
Part 4: Genomic prediction methods and machine learning algorithms .................................. 39  

**CHAPTER 3. GEOSPATIAL STATISTICS FOR SOYBEAN IRON DEFICIENCY CHLOROSIS** .................................................................................................................................60  
Abstract.................................................................................................................................... 60  
Introduction ............................................................................................................................. 61  
Data and methods .................................................................................................................... 68  
Data .................................................................................................................................... 68  
Analytic methods................................................................................................................ 72  
Performance metrics to compare the models ..................................................................... 79  
Heatmap and Lagrange Multiplier Test.............................................................................. 81  
Relative efficiency (RE) ..................................................................................................... 82  
Results and discussion ............................................................................................................. 82  
Results from the P-spline model SpATS............................................................................ 82  
Spatial effective dimension (EDs) and the importance of surface trend by f(row):f(range) .............................................................................................................................. 85  
Variance components analysis and importance of surface trend by f(row):f(range) ...... 86  
Comparison metrics among the eight models ................................................................. 88  
Relative efficiency (RE) of the spatial autoregressive (SAR) analyses ......................... 96  
Lagrange multiplier test (LMT) ........................................................................................ 97  
Statistical experiment design for spatial analysis versus breeding practice .................... 98  
IDC hill plot size and spatial variation............................................................................... 99  
The tensor product penalized splines may work better for continuous data type than for ordinal data type ............................................................................................................... 101  
Conclusion ............................................................................................................................. 101  
Conflict of interest................................................................................................................. 102
Supplemental data available .................................................................................................. 102
Acknowledgments ................................................................................................................. 102
References ............................................................................................................................. 102

CHAPTER 4. ALGORITHMIC MODELING VS. DATA MODELING: WILL THE
ALGORITHMIC OUTPERFORMS DATA MODELING FOR IRON DEFICIENCY
CHLOROSIS RESISTANCE? .................................................................................................109
Abstract ................................................................................................................................ 109
Introduction ........................................................................................................................... 110
Data and Methods .................................................................................................................. 117
Data .................................................................................................................................. 117
Analytic methods .............................................................................................................. 117
Measurement metrics for prediction accuracy ................................................................. 118
Results and discussion ........................................................................................................... 120
IDC overall prediction accuracy from data modeling, rrBLUP, and logistic regression . 120
IDC prediction results from algorithmic modeling, support vector machine (SVM) ..... 121
IDC prediction results from algorithmic modeling, Naïve Bayes (NB) ....................... 124
IDC prediction results from algorithmic modeling, K-nearest neighbor (KNN) .......... 125
IDC prediction results from algorithmic modeling, artificial neural network (ANN) ..... 126
IDC prediction results from algorithmic modeling, Gradient Boosting Machine
(GBM) .............................................................................................................................. 127
IDC prediction results from algorithmic modeling, random Forest (RF) ..................... 129
Summary from algorithmic modeling: SVM, NB, KNN, AN, GBM, and RF .............. 130
Summary of the model performance measured by AUC from both algorithmic and
data models ....................................................................................................................... 132
Models separation by principal component analysis (PCA) ............................................ 134
Algorithmic vs. data modeling: Multiple comparisons and groups .............................. 135
Conclusion ............................................................................................................................. 136
Conflict of interest ................................................................................................................. 136
Supplemental data available .................................................................................................. 137
Acknowledgments ................................................................................................................. 137
References ............................................................................................................................. 137

CHAPTER 5. MACHINE LEARNING FOR SOYBEAN SDS ANALYSIS ....................................157
Abstract ................................................................................................................................ 157
Introduction ........................................................................................................................... 158
Materials and analytic method .......................................................................................... 160
Materials ............................................................................................................................... 160
Analytic Method .................................................................................................................... 161
Model and parameter selection for Random Forest analysis ........................................... 161
Model and parameter selection for stepwise regression analysis ..................................... 162
Model and parameter selection for GWAS analysis by Tassel ........................................ 162
Results and discussions ......................................................................................................... 163
Predictive variable selection results from random forest analysis ................................... 163
Predictive variable selection results from stepwise regression by JMP 12 Pro ............... 164
Predictive variable selection results from GWAS by Tassel 5 ....................................... 165
Comparison of marker selection result among the three methods: random forest,
stepwise regression, and GWAS ................................................................. 166
Results of the selection of the overlapped SNPs from the three analyses: random
forest, stepwise regression, and GWAS .................................................. 168
SNP-SNP interaction network .................................................................. 169
Conclusions ............................................................................................ 170
Conflict of interest ................................................................................... 170
Supplemental data available ...................................................................... 170
Acknowledgments .................................................................................... 171
References ............................................................................................... 171

CHAPTER 6. DISTRIBUTION AND EVOLUTION OF COTTON FIBER DEVELOPMENT
GENES IN THE FIBRELESS G. RAIMONDII GENOME ........................................ 190
Abstract .................................................................................................... 190
Introduction ............................................................................................ 191
Materials and Methods .......................................................................... 194
  Sources of the cotton Unigenes used in this study .................................. 194
  Mapping of cotton EST Unigenes to chromosomes of the tetraploid (G. hirsutum) At
  and Dt subgenomes, and the diploid (G. raimondii) D genome .................. 195
Results ...................................................................................................... 196
  Distribution of fiber initiation Unigenes among chromosomes of the tetraploid At
  subgenome and the diploid D genome .................................................... 196
  Distribution of fiber initiation Unigenes among chromosomes of the tetraploid Dt
  subgenome and the diploid D genome .................................................... 198
  Distribution of fiber elongation Unigenes among chromosomes of the tetraploid At
  subgenome and the diploid D genome .................................................... 199
  Distribution of fiber elongation Unigenes among chromosomes of the tetraploid Dt
  subgenome and the diploid D genome .................................................... 201
  Distribution of fiber second cell wall deposition (SCWD) Unigenes among
  chromosomes of the diploid D genome and the tetraploid At and Dt subgenomes .... 201
  Overall summary of the distribution of fiber initiation, fiber elongation, and SCWD
  Unigenes in recombination hotspot regions between the tetraploid At and Dt
  subgenomes vs. the diploid D genome .................................................... 203
  Chromosomal distribution of the At- or Dt-specific fiber initiation Unigenes .......... 204
  Chromosomal distribution of the At- or Dt-specific fiber elongation Unigenes .......... 205
  Visualization of the fiber Unigenes among At subgenome, Dt subgenome, and
  diploid D genomes ................................................................................ 207
Discussion .............................................................................................. 208
  Distribution of the genetically mapped sequences in the tetraploid cotton At and Dt
  subgenomes .......................................................................................... 208
  Distribution of fiber development Unigenes in recombination hotspot regions of the
tetraploid At and Dt subgenomes, and the diploid D genome ....................... 209
  Implications of this study in future applications ....................................... 210
Conclusions ............................................................................................ 211
Acknowledgment .................................................................................... 211
References ............................................................................................... 212

CHAPTER 7. GENERAL CONCLUSION ................................................................. 216
ABSTRACT

Prediction accuracies of genomic selection methods are affected by the quality of the phenotypic and genotypic data and by the use of appropriate analytic models in the training sets. This research focuses on the impact of data quality for ordinal traits. Ordinal scores of traits are typical for various types of stress tolerance and resistance. Established spatial models developed for continuous quantitative traits were unknown whether they can effectively adjust the spatial autocorrelation for ordinal traits with sharp transitions patterns among groups of plots in experimental field trials. The effectiveness of the spatial adjustments was systematically compared with eight different spatial models using soybean iron deficiency chlorosis (IDC) as an example. After incorporation of the spatial pattern recognition to provide adjusted ordinal data, a comparison of prediction accuracies between algorithmic modeling and data modeling approaches were systematically conducted. The results revealed that genomic prediction accuracies could be dramatically improved by both machine learning models and geospatial spatial analyses. Overall, algorithmic modeling outperforms data modeling methods for the soybean IDC ordinal data type. Further, machine learning algorithms provide higher prediction accuracy than traditional statistical data models in terms of sensitivity, specificity, and overall accuracy.
CHAPTER 1. GENERAL INTRODUCTION

Genomic prediction is an advanced form of marker-assisted selection in which genetic markers covering the whole genome are used so that all genes and QTLs associated with traits of interest are in linkage disequilibrium with the markers. Genomic prediction includes two steps: Step I is to train the prediction model using lines with both genotypic and phenotypic data as training sets. Step II is to predict breeding values using genotypic information of the un-phenotyped lines, the testing sets, with the trained model and estimated parameters from step I.

Genomic prediction is becoming a widely adopted tool for breeding. It makes use of historical genotypic and phenotypic information to predict genotyped but not phenotyped lines for selection decisions. The advantage of genomic prediction is that the prediction accuracy will be improved with high quality phenotypic and genotypic data accumulated across years in the training sets. Accuracies of genomic prediction are affected by the quality of genotypic and phenotypic data and by an appropriate statistical model. These three components play different roles in their effect on prediction accuracy. Among them, phenotypic data quality is the most important component but most difficult to control because of both unpredictable weather patterns from year to year and genotype by environment interactions from testing locations to locations. Of them, spatial variability in the field caused by soil property, local weather conditions, and experimental activity (such as the movements of machines during plowing, tilling, and other procedures.) is the major factor that can decrease prediction accuracies. Hence, there is a need for spatial analysis of the phenotypic data before combining phenotypic data from multi-environment trials within a year and across years. A proper phenotypic spatial analysis is a crucial prerequisite for accurate calibration of genomic prediction procedures (BERNAL-VASQUEZ et al. 2014).
Spatial variation analysis has experienced remarkable growth in recent years in terms of theory, methods, and applications in econometrics and geo-statistics. Many models to account for the spatial variation have been developed and successfully applied to increase phenotypic data quality (Anselin 1988a; Anselin and Florax 1995; Anselin et al. 1995; Anselin et al. 1996; Anselin 2003; Anselin et al. 2004; Anselin et al. 2010; Technow 2015b). In wheat breeding, phenotypic data quality was improved by a mixed spatial model using moving means as a covariate; subsequently, genomic prediction accuracy was increased with the spatial adjustment. In rye, a model with column and row effects generated the highest prediction accuracy of all the tested spatial models (Bernal-Vasquez et al. 2014). To my knowledge, spatial variation analysis in breeding has been applied only to adjust continuous variables, such as grain yield, plant height, flowering time, thousand kernels weight, etc. I am unaware of any published reports on the application of spatial analyses to ordinal variables, such as scores for biotic and abiotic stress traits. For example, Sudden Death Syndrome and Iron Deficiency Chlorosis when scores are recorded by experts on a scale of 1 to 9, where a score of 1 represents the most resistant and 9 represents the most susceptible.

An analytical model is the second most important factor that affects genomic prediction accuracy. There are two major categories of analytical models used with training data sets (Breiman 2001b). One assumes that the data are generated by a pre-defined distribution-based stochastic data model, such as normal/Gaussian, Poisson, binomial, logistic, exponential, etc. (Baldi and Moore 2012; Moore et al. 2012). These models are referred to as data models. The other uses algorithmic approaches and treats the relation between phenotypes and genotypes as a black box with unknown distributions. This second group of analytic approaches are referred to as algorithm models and include machine learning techniques such as support vector machine
(SVM), random forest (RF), artificial neural network (ANN), Naïve Bayes Classifier (NBC), k-nearest neighbor (KNN) and so on (GUTIERREZ AND SAFARI BOOKS ONLINE (FIRM)). Logistic regression and ridge-regression BLUP (rrBLUP) belong to data models because the prerequisite assumptions for these models are that the data follow s-shape and bell-shaped distributions, respectively (MOORE et al. 2012). In contrast, majority of the machine learning and deep learning algorithms (SVM, ANN, KNN, RF, GBM) belong to algorithm models because no predefined distribution assumptions are needed (CAELLI AND BISCHOF 1996; LAVINE et al. 2004; MCDERMOTT et al. 2013; KRUPPA et al. 2014; BEHMANN et al. 2015). In the past decades, comparisons between data models and algorithm models have been conducted revealing that sometimes data models produce more accurate predictions than algorithm models, sometimes they are less accurate (BREIMAN 2001b; COX et al. 2001; IOANNIDIS 2005). These results suggest that estimates of accuracy depend on whether the data meet the prerequisite statistical distribution and model assumptions (HOWARD et al. 2014). Complex quantitative traits, including biotic traits, such as soybean sudden syndrome (SDS), flowering and relative maturity (RM), abiotic trait, such as iron deficiency chlorosis (IDC), will be used in this study as examples to compare the prediction accuracy between the data models and the algorithm models.

The objectives of this dissertation included:

1) To assess the effectiveness of spatial models to correct spatial patterns in ordinal data from field plots. Because none of the existing data model methods were capable of removing spatial patterns exhibiting sharp phase changes in ordinal data, a unique algorithmic model was developed for these types of patterns.

2) To compare the impacts of spatial adjustments to ordinal phenotypes on genomic prediction accuracies.
3) To compare algorithmic and data models based on prediction accuracies of spatially adjusted ordinal data.

4) To identify the network of genes and QTLs that may control SDS using machine learning algorithms, especially by multivariate adaptive regression splines (MARS), which can detect a high degree of interactions.

5) To investigate the distribution, evolution, and function of recombination hotspot regions in sub-genome Dt, which its ancestor diploid D genome does not produce any spinnable fibers in the tetraploid cotton genome.

**Dissertation organization**

This dissertation is organized into six chapters and one appendix of word file generated from R markdown scripts, which contains all the R codes and results.

Chapter 1 provides a general introduction and organization of this dissertation

Chapter 2 provides a literature review, including four parts: 1): An abiotic trait: Soybean iron deficiency chlorosis (IDC); 2): A biotic trait: Soybean sudden death syndrome (SDS); 3): Historical and current research on spatial data analysis; 4): Genomic selection and machine learning algorithms

Chapter 3 covers the comparison results of the phenotypic data adjustment from different spatial models. This chapter represents a manuscript for publication.

Chapter 4 covers comparison accuracies of predictions from data and algorithmic models, including rrBLUP, logistic regression, GBM, SVM, KNN, Naïve Bayes classification, and deep learning ANN. This chapter is being prepared as a manuscript for publication. This chapter represents a manuscript for publication.
Chapter 5 covers the results for SDS analysis to identify the interaction network of genes which may control SDS resistance pathways using machine learning algorithms. This chapter represents a manuscript for publication.

Chapter 6 is a published manuscript using bioinformatics and computational methods to study the distribution and evolution of genes in recombination hotspot regions in the cotton genome. This manuscript was developed and published using computational tools learned early in the BCB curricula at Iowa State University and is only loosely related to the theme of this dissertation.

The appendix is a word document generated from R markdown files that contain all the R codes used in this study and derived tables and figures from the R scripts.
CHAPTER 2. LITERATURE REVIEW

Part 1: An abiotic trait: Soybean iron deficiency chlorosis (IDC)

Iron is an essential micronutrient for the plant. Besides being needed in chlorophyll, iron is also involved in the energy transfer of the plant, is part of metabolism enzymes, and is needed in the root nodule formation associated with the N-fixation. Iron deficiency chlorosis (IDC) in soybeans is caused by the inability of the plant to utilize iron in the soil. Without enough soluble iron for plants, chlorophyll production is hampered, and the plant will suffer and possibly die. IDC may occur when IDC susceptible varieties are grown on calcareous soils (Niebur and Fehr 1981). IDC mainly found in the Midwest of the US, especially Iowa, Minnesota, Nebraska, and South and North Dakota (Froehlich and Fehr 1981; Franzen and Richardson 2000).

IDC is expressed in new leaf tissue, and symptoms typically appear on younger leaves between the first and third trifoliate growth stages from vegetative V1 to V3 (Lin et al. 1998). The typical symptom of IDC is the yellowing leaves with interveinal chlorosis, while the veins remain green (Goos and Johnson 2000). Chlorosis is the result of low chlorophyll formation due to iron deficiency. If the adverse soil condition for IDC development is not severe and short, soybean plants may be able to absorb sufficient soluble iron and recover from the chlorosis, and yellowing symptoms cannot be recognizable by breeders. If the adverse soil condition for IDC development continues deteriorating, with severe iron deficiency, leaf edges and growing points become necrotic. With the progress of necrosis, the leaves will die and fall off the plant, and eventually, the entire plant will die. The bird view of IDC symptoms in a field in the Midwest of the US is patches of an area with yellowing plants scattered across the field.

Soybeans are the second-most-planted field crop in the United States after corn, ~ 80 million acres planted from 2010, and with a record-high 85.1 million acres planted in 2015 based
on the National Agriculture Statistics Service (NASS). More than 80% of US soybean acreage is concentrated in the upper Midwest (http://www.usda.gov/nass/pubs/todayrpt/acrg0615.pdf). A big portion of the soybean production areas in the upper Midwest is prone to IDC. Yield reduction from IDC was observed as early as in the 1960s (BURAU 1965). The first study of a quantitative measure of yield loss regression on chlorosis score showed that the relationship between yield loss and IDC score is linear and has a 20% yield reduction for each unit increase in chlorosis based on chlorosis scores of 1, for no yellowing symptom, to 5, for severe yellowing, leaf and plant die, (FROEHLICH et al. 1980). Based on a survey of 79 soybean producers in Upper Midwest regions, 99% of the soybean producers indicated that IDC was a major production issue, and 24% of their soybean crops were affected by IDC, generating an estimated 12 bu/acre yield loss per annum (HANSEN et al. 2003). Soybean iron deficiency planting acreage is increasing from 1979 to 2012. A 160% increase in soybean production area into regions with soil pH of 7.2 or greater in the past 30 years. This increase of soybean production area into iron deficiency prone regions has led to yield losses of 340 million tons and worth an estimates $120 million dollars per year (HANSEN et al. 2004). Current IDC trends in soybean production areas are expected to continue, thus, minimizing or eliminating yield lost due to IDC is critical.

A large number of research studies have been conducted to better understand the soil and environmental properties associated with iron deficiency chlorosis in soybean (BURAU 1965; FROEHLICH et al. 1980; TRIMBLE AND FEHR 1983; KAUR et al. 1984; HINTZ et al. 1987; CLARK et al. 1988; LOEPPERT et al. 1988; LONGNECKER AND WELCH 1990; SINGH AND DAYAL 1992; GONZALES et al. 1998; LUCENA 2000; HELMS et al. 2010; ALVAREZ-FERNANDEZ et al. 2011; BLOOM et al. 2011). Collectively, these studies outline a complex relationship between soybean
IDC symptoms and directly and indirectly causal factors into three aspects: 1) soil characteristics, 2) weather conditions, and 3) soybean genotype and physiology.

Several soil factors including, but not limited to, the following: soil moisture, soluble salt, calcium carbonate (CaCO3), bicarbonate (HCO3-), and nitrate (NO3-) content, and soil pH (BURAU 1965; LOEPPERT et al. 1988; MORRIS et al. 1990; FRANZEN AND RICHARDSON 2000; BLOOM et al. 2011). While all of these soil factors have been implicated with iron deficiency chlorosis, not all have been consistently associated with chlorosis in both field and nutrient solution experiments. The conclusion from these studies is that these soil factors determine the form of iron, soluble, or insoluble for soybean. Soils usually have a sufficient quantity of iron, but it might not be in the required soluble and ready to be absorbed by soybean plants. If the final results of iron form the combination of the soil factors is ferric hydroxide Fe (OH) 3, which is the soluble form of Fe in Fe (III), ferric. But, this soluble Fe(III) becomes insoluble with pH value from 7.2 to 8.4 and a high level of calcium carbonate (MORRIS et al. 1990).

Soil nitrate is a causative factor in iron deficiency chlorosis in soybeans (BLOOM et al. 2011). While soybean plants have the ability to fix N through root nodules, they will take up nitrate directly from the soil when it is available. When root takes up nitrate, they release bicarbonate. Overtime, bicarbonate levels can increase in soil, and the accumulated bicarbonate reduces the level of soluble iron, which eventually lead to the development of IDC symptoms.

Weather also plays a role in IDC symptom development. Optional weather condition for IDC symptom development includes cool temperature and wet soils. When soils are wet, there is limited air exchange with the atmosphere, which causes a buildup of carbon dioxide in the soil. The carbon dioxide is produced by roots and soil microbes through respiration. The amount of bicarbonate in the soil is proportional to the amount of carbon dioxide, and as carbon dioxide
increases, so does bicarbonate. This increase will rapidly neutralize the acidity around the soybean root. The amount of bicarbonate in the soil has been positively correlated with IDC in soybean in the field (Kaiser et al. 2014).

Based on the physiological response to Iron deficiency in the soil, plants were divided into two types: Strategy I and Strategy II plants (Marschner et al. 1986). Under iron stress, Strategy I plants, through their roots, will respond to the iron deficiency through three approaches: (1) acidifying the rhizosphere through excretion of protons from H+-ATPases, (2) reducing chelated Fe3+ to Fe2+ mediated by plasma membrane ferric (chelate) reductases, and (3) transferring soluble Fe2+ across the plasma membrane and into the cytoplasm via divalent iron transporters. In contrast, Strategy II plants involve excretion of phytosiderophores that bind to the insoluble form of ferric iron (Fe3+), creating a soluble complex that Strategy II plants can uptake (Marschner et al. 1986; Romheld and Marschner 1986). Soybean and other dicot plants belong to Strategy I plants.

Genetic analysis and mapping of IDC QTL for marker-assisted selection (MAS) breeding have accumulated a wealth of knowledge and can be divided into three stages: bi-parental population-based mapping, connected population mapping, or network population mapping (NPM), and genome-wide association study (GWAS).

IDC genetic inheritance of the single dominance/recessive gene model was first reported in 1943 (Weiss 1943). From the analysis of 24 bi-parental populations between and among four IDC efficient and six inefficient lines, Weiss concluded that dominant allele control iron efficiency and the recessive allele is conditioning iron inefficiency in iron utilization. This single gene control of IDC was confirmed and modified by Dr. Cianzio that IDC was controlled by a single major gene model but additionally observed quantitative inheritance patterns, which
indicated that IDC is also controlled by modifying genes (Cianzio and Fehr 1980). In contrast to the single gene theory of IDC, analysis from different genetic populations developed from different IDC tolerant donors show that IDC is quantitatively inherited (Fehr 1982). Subsequent studies conducted with two different mapping populations with different donors by Lin confirmed that IDC displayed both single codominant gene and polygenic inheritance patterns in Anoka x A7 and Pride B2/1412 x A1/55 mapping populations, respectively (Lin et al. 1997; Lin et al. 2000b). Except above intraspecific bi-parental mapping population via Glycine Max x Glycine Max, an interspecific bi-parental mapping population between Glycine Max (A86-356022) and Glycine Soja (PI468916) was used to screen for IDC QTL, and three markers were associated with IDC QTL from the mapping set, but could not be confirmed in the tester set which both mapping and tester set were from the same population (Diers et al. 1992). In order to determine the efficiency of SSR markers in selecting for IDC resistance in breeding populations, a superior IDC tolerance, and moderate yield potential line, A97-770012, crossed with a pioneer line, P9254, with moderate IDC resistance and superior yield potential (Charlson et al. 2003). From this breeding population, three SSR markers were associated with chlorosis scores at each location, but the identities of the associated markers were different at each location. From the same team, with one more year phenotypic data and 108 SSR markers which were previously mapped and associated with 8 IDC QTLs on eight molecular linkage groups, 3 of 24 polymorphic SSR markers associated with IDC tolerance in three different locations, only one SSR marker, satt481, associated with IDC tolerance across all three environments and account for 12% of the total phenotypic variation (Charlson et al. 2005). Most recently, QTL mapping for iron and zinc concentration in the seeds identified one major QTL for iron accumulation on chromosome 20, which explained 21.5% of the phenotypic variation. This QTL was in the
marker interval pa_515-1-satt239, with pa_515-1 previously associated with iron efficiency QTL (Lin et al. 1997; King et al. 2013).

In summary, IDC is a complicated quantitative trait and controlled by both major and multiple minor genes/QTL from bi-parental mapping studies. The complex inheritance of IDC is population-dependent and influenced by a significant genotype x environment interaction (Charlson et al. 2003; De Cianzio and Fehr 1982; Rodriguez De Cianzio and Fehr 1980; Weiss 1943).

Quantitative traits like IDC are difficult traits for breeders to improve through traditional phenotypic selection methods due to the inability of a breeder to effectively select and stack numerous favorable alleles that confer IDC tolerance. Environment-independent marker-assisted selection (MAS) has been successfully used for simple qualitative traits, but not large-scale used for quantitative traits. The first challenge of using markers to select for quantitative trait loci is the ability to identify the markers that are associated with IDC tolerance, and the marker selection works across different environments/locations and years. Due to heterogeneous soil factors and the presence of genotype x environment (G x E) interactions, it is difficult to distinguish genotypic sources of variation from the environmental components responsible for chlorosis (Froechlich and Fehr, 1981; Naeve and Rehm, 2006).

Bi-parental population-based mapping has been successfully used for mapping qualitative traits but has the following limitations: 1) usually needs a large population size, but it is difficult for soybean population development due to the difficulty of manually pollination. To effectively detect marker-trait associations in complex traits such as IDC, large population sizes (>400) are required in order to provide adequate detection power for small effects QTLs (Bernardo 2004). The mapping population sizes used for QTL studies so far posted in SoyBase
(https://soybase.org/) all have less than 150 recombinants inbreed lines (RILs) or F2/F2 lines. These smaller population sizes have limited power to detect minor effect QTL typical of a quantitative trait; 2) QTL identified from the bi-parental mapping population is population-dependent. Frequently, QTLs detected in one population are not detected in other populations, or QTLs detected in one population fail to exhibit the same or similar gene effect when validated in other genetic backgrounds; 3) small effect quantitative QTL have gene by environment effects.

To overcome these drawbacks of bi-parental mapping of IDC QTL, connected population mapping using many breeding populations that shared crossing parents among these populations was proposed to map universal IDC QTLs.

Markers explaining phenotypic variation were identified in one population but were not significantly associated with IDC in another population, and QTL were mapped via a bi-parental population, but the same QTLs cannot be mapped when the QTL donor parent cross with another susceptible parent. Both cases indicated the mapped QTL are population/genetic background-dependent. In order to identify QTLs that perform stably across different genetic backgrounds, connected population, or network population mapping (NPM) was employed to discover QTLs. The two main advantages of the NPM are: 1) population size was dramatically increased compared with single bi-parental mapping populations. For example, parent “A” crosses with Parents B, C, D, E, F, and parent B cross with Parent C, D, M, N, and Q. If each population has 100 lines, the total population size will be 1,000 lines; 2) NPM can overcome QTL population-dependent issue. For the above example, if the QTL from donor parent A works in the background A x B, A x C, … A x F, then the identified QTL works across different populations. Otherwise, if the QTL from donor parent A works only in one of the many populations, then that population-specific QTL will not be discovered by NPM. The first research used the NPM
approach to map IDC was conducted by Ilene Jones (master’s degree thesis). From 8 IDC tolerant and susceptible parents, a total of 13 RIL populations were collected from Syngenta soybean IDC breeding programs (Figure 1). A total of 50 IDC QTLs were identified from the 13 disconnected bi-parental populations. In contrast, only 22 QTLs were identified in all the 8 IDC tolerant parents.

Figure 1. The network of connected mapping populations. Parent names were labeled inside rectangle boxes, and the lines show the crosses and connected network of populations. Twelve populations evaluated for iron deficiency chlorosis are depicted in blue; one population evaluated for yield is depicted in red; five populations evaluated for iron deficiency chlorosis and yield are depicted in purple.

Another project using connected network association mapping is soybean nested association mapping (NAM), which targets are to map genes/QTLs controlling soybean yield and other key traits (https://www.soybase.org/SoyNAM/). A total of 40 import soybean varieties and cultivars selected from both domestic and international exotic germplasm crossed to a common hub parent, IA3023. With 140 recombinant inbred lines (RILs) were developed from
each population, the total population size of the NAM is 5,600, which should be powerful enough to detect small effects QTLs. NAM has been successfully applied to map QTLs controlling thousand-grain weight in barley (Maurer et al. 2016), stem rust resistance in wheat (Bagain et al. 2016), and gray leaf spot resistance in maize (Benson et al. 2015).

Although the NPM mapping method overcomes the QTL population-dependent issue, its disadvantage is that you have to develop a large/moderate number of connected populations, and the QTLs identified limited in the small number of parents. A powerful and complementary to bi-parental and NPM mapping strategy, genome-wide association study (GWAS) that can be used to exploit historical recombination events accumulated from artificial selection (breeding) and natural selection pressure, was employed to map IDC QTLs (Wang et al. 2008; Mamidi et al. 2011; Mamidi et al. 2014). From the point of analysis method for the QTL mapping view, GWAS has applied a more advanced statistical model than that of bi-parental and NPM mapping. Both bi-parental and NPM mapping usually use simple single marker ANOVA analysis and/or composite interval mapping (CIM), which is a single-factor analysis of variance (SFA). In contrast, GWAS can use mixed linear model (MLM) except the general linear model (GLM), which the MLM treats the predictive variables as a random effect, plus including population structure, kinship among the testing lines to reduce the number of false-positive QTLs (Mamidi et al. 2011; Vuong et al. 2015; Zhang et al. 2016). With the MLM analysis model, two SSR markers (Satt14 and Satt239) were found to be associated with IDC in two independent association mapping panels, which consist of major public and private company lines. Verification of the two markers shows that the lines with the tolerant alleles at both of these two loci have significantly lower IDC scores than lines with only one or no tolerant alleles (Wang et al. 2008). Comparing with the GWAS by low resolution (84 Satt markers) of SSR marker,
GWAS with much higher density (1536 SNPs) SNP chip technology of Golden Gate and two relatively bigger association mapping panels (141 and 141 lines, respectively) was conducted, and seven genomic regions were significantly associated with IDC across two years phenotypic data. The seven loci accounted for 43.7% of the phenotypic data variation in 2005 and 47.6% in 2006 (MAMIDI et al. 2011). Most recently, GWAS was conducted to map the IDC genomic regions with genotype-by-sequencing technology and identified seven major effect QTLs in seven chromosomes (MAMIDI et al. 2014). From this report, a total of 12 candidate genes were associated with iron metabolism mapped near the 7 IDC QTLs, supporting the polygenic nature of soybean IDC.

The ultimate goal of QTL mapping, no matter which approach was employed, either bi-parental population-based mapping, or connected network population mapping, or GWAS, is to improve breeder’s selection accuracy via marker-assisted selection. There are several challenges that breeders face to implement the markers to stack the IDC QTLs to increase IDC breeding efficiency and accuracy:

Challenge I: no major universal IDC QTL has been discovered yet. For example, a single QTL for soybean cyst nematode (SCN) resistance from PI88788, it is a universal SCN resistant QTL and provides resistance to 99% commercial varieties (DELHEIMER et al. 2010). This QTL provides stable resistance across global soybean germplasm, and it is a population- and environment-independent QTL. By now, for IDC tolerance QTL discovery, the only major QTL was mapped in a population between Anoka x A7, the QTL can explain more than 70% of the phenotype variation (CIANZIO AND FEHR 1982; FEHR 1982). No further reports show that this major IDC QTL can provide the same or similar resistance in other genetic backgrounds, or at least, the author was not successful in querying out reports about successfully transferring this
major IDC QTL to the other populations. Instead, from the same team/lab, two QTLs on MLG I/chromosome 20 and MLG N/chromosome 3 were identified in both the Pride B216 x A15 and the Anoka x A7 mapping population, respectively (LIN et al. 2000a), but, no common DNA marker was present in both populations, and the authors concluded that the markers identified in the two mapping populations couldn’t be used in marker-assisted selection because of a general lack of common markers between the two populations (LIN et al. 2000a). The reason/mechanism why this QTL can explain 72% phenotype variation and perform like qualitative trait but has not been widely transferred to other germplasm was being further investigated by identifying the candidate genes underlying this major QTL through molecular breeding, fine mapping, and transcriptome sequencing (PEIFFER et al. 2012). Two gene encoding transcription factors within the major QTL were identified by transcriptome sequencing and confirmed by real-time PCR (PEIFFER et al. 2012).

One report to validate whether SSR markers previously reported associate with IDC QTLs can be used for marker assistant selection for a breeding population was conducted by Charlson (CHARLSON et al. 2005). In the report, a total of 108 SSR markers genetically linked to 8 QTLs on eight chromosomes were used to make MAS selections in a breeding population between superior and moderate IDC tolerance to test MAS selection accuracy and efficiency. Among the 108 SSR IDC markers, 24 of them are polymorphic, which the 22.22% (24/108) polymorphic rate is consistent across all soybean germplasm. The remaining 84 markers are monomorphic or homozygous in the two parental lines of the breeding population and thus not useful in screening for iron efficiency. The report did not show whether these 84 SSR markers are positively or negatively fixed in the two parental parents, even though the author wants to know and hope that they were positively fixed because the positively fixed genotype indirectly
proves the accuracy of the 84 SSR markers. Of the 24 polymorphic markers, three of them were associated with IDC resistance. However, only one marker, satt481 accounted for 12% of the total phenotypic variation, was associated with IDC tolerance across environments. Interestingly, marker Satt481 is in MLG L/chromosome 19 and is located at neither MLG N/chromosome 3 nor MLG I/chromosome 20 which are the two major QTLs reported before (Rodriguez de Cianzio and Fehr 1980; Lin et al. 1997; Lin et al. 2000a). The author wants to know whether these two QTLs were among the 84 monomorphic markers and whether they are positively or negatively fixed in the progeny from this breeding population. No matter which was the case, this example indicated that MAS might not be an ideal approach for complex quantitative trait IDC selection.

Challenge II: There is no efficient way to use MAS to stack many minor QTLs for IDC tolerance breeding by comparing the limited population size versus the number of IDC QTL to fix. By now, many IDC QTLs have been reported. From the Soybase, 39 IDC QTLs were deposited from 4 bi-parental mapping populations. From the connected network population mapping, 50 QTLs were detected via composite interval mapping (CIM) using the disconnected bi-parental model, and 22 QTLs were detected via CIM using the connected network population mapping (Thesis, Ilene Jones). From GWAS, with the strictly controlling of population structure, 42 and 88 QTLs were identified from the phenotypic data collected in 2005 and 2006, respectively (Mamidi et al. 2011), by the same group, 33 QTLs distributed on 10 of the 20 chromosomes were identified from the genotype-by-sequencing data (Mamidi et al. 2014). More recently, Transcriptome sequence profiling of soybean (Glycine max, L. Merr) near-isogenic lines Clark (PI548553, iron efficient) and IsoClark (PI547430, iron inefficient) grown under Fe-sufficient and Fe-limited conditions discovered 835 candidate genes in the Clark (PI548553)
genotype and 200 candidate genes in the IsoClark (PI547430). Of these candidate genes, and only a small portion (7~10%) of the differently expressed genes were mapped to previously identified QTL regions (O'ROURKE et al. 2009). Comparing with these large numbers of IDC QTLs/candidate genes identified so far, Soybean is a strictly self-pollinated crop, and it is very challenging to develop large size populations to breed IDC tolerant varieties via MAS selection. For example, if an IDC population has five complimentary IDC QTLs, one mandatory herbicide marker (such as Roundup Ready RR2) and one mandatory SCN marker, then we need to select seven loci, the least F2 population size need to be \((2)^7=128\) lines in order to get one plant with all the seven loci positively fixed. This is only for the required agronomic traits, and usually, breeders need to have 30 to 50 lines for yield variation. This leads to the population size to \(128 \times 30 = 3,540\) lines per population, which is impossible for large-scale soybean breeding programs.

QTL-based MAS strategy has been quite useful for the manipulation of large-effect alleles with known association to a marker but not for quantitative traits that have still required extensive field testing (MOREAU et al. 2004). The methods of MAS or marker-assisted recurrent selection assume that the user knows which alleles are favorable, and what their average effects on the phenotype are (BERNARDO AND CHARCOSSET 2006). This assumption is viable for major-gene traits but not for quantitative traits that are influenced by many loci of small effect and the environment. In locus identification and effect estimation for such traits, much uncertainty will remain (MELCHINGER et al. 1998). To deal with quantitative traits, new statistical approaches that could account for this uncertainty were needed to generate the best predictions possible.

In conclusion, traditional marker-assisted selection has been ineffective for traits that are complex and affected by many genes and environmental factors, each with a small effect
What is the solution for a complex trait with many minor effect alleles? Genomic selection and machine learning for predictive breeding have arisen from the conjunction of new high-throughput high-density chip genotyping technologies, new statistical methods, and machine learning algorithms to analyze the big data sets and provided a promising alternative to QTL-based MAS breeding.

**Part 2: A biotic trait: Soybean sudden death syndrome (SDS)**

Sudden death syndrome (SDS) of soybean (*Glycine max* (L.) Merr.), caused by the soil-borne fungal pathogen *Fusarium virguliforme*. It was first observed in Arkansas in 1971 and has become widespread throughout the north-central United States since then (Hartman *et al.* 2015). The fungus infects soybean root systems and produces toxins that are translocated to the leaves, resulting in premature defoliation and pod abortion (Jin *et al.* 1996), and eventually leading to grain yield reduction. In recent years, SDS ranked among the top three most damaging diseases of soybean in the United States after Soybean Cyst Nematode (SCN) and Phytophthora. In the Midwestern soybean-producing area of the U.S., it is estimated that SDS has resulted in average losses valued at $190 million a year and the disease is spreading and intensifying (Wrather and Koennig 2006).

SDS early symptoms of SDS are diffuse chlorotic mottling and crinkling of the leaves. Later, leaf tissue between the major veins turns yellow, then dies and turns brown. Soon after, the leaflets die and shrivel. In severe cases, the leaflets will drop off, leaving the petioles attached. For further diagnosis from the other disease, the cortical tissue of a plant with SDS will exhibit tan to light brown streaks, whereas the cortex of a healthy plant will be white from the split lower stem.

Disease resistance breeding is believed to be the most effective approach to control SDS. Since no soybean genotypes confer complete resistance or immunity to this disease, soybean
breeders still rely on quantitative tolerance to SDS. For this purpose, many QTLs have been mapped and published via either bi-parental populations or genome-wide association studies. The first two QTLs mapped by random amplified polymorphic DNA (RPAD) marker are in chromosome 6 (linkage group LG C2) and chromosome 3 (LG N), which jointly accounted for 34% of total phenotypic variability in mean SDS disease incidence (Hnetkovsky et al. 1996). Two more QTLs form the same bi-parental mapping population was mapped in chromosomes 3 and 18 (LG G). Together, the four QTL accounted for about 65% of total phenotypic variability in mean disease incidence and 50% in mean disease severity (Chang et al. 1996). Four near-isogeneic lines (NILs) derived from a recombinant inbred line (RIL), ExF34, with heterozygous within regions of linkage group C2 and G having different resistance to SDS indicated that the QTL on linkage group G and C2 confer separate components of resistance to SDS (njiti et al. 1998). With the application of microstate simple sequence repeat (SSR) DNA markers in QTL mapping, three more QTLs were mapped by the same group using the same mapping population (Forrest x Essex), and they were mapped in chromosome 3, 6, and 20 (LG I), interaction test of the mapped 7 QTLs indicated QTL action was additive among the loci underlying resistance to SDS (IQBal et al. 2001). The 8th SDS QTL was mapped in C2 from variety “Douglas” with R² 0.14 (Njiti et al. 2002). From US northern varieties Minsoy and Noir1, 9th QTL was mapped in chromosome 19 (LG L) with R² 0.14 (NJITI AND LIGHTFOOT 2006). By dissection resistance into root infection and leaf scorch during soybean SDS development using the cross segregating for both SDS and Soybean cyst nematode (SCN), the 10th SDS QTL was discovered on chromosome 18 underlying resistance to SDS leaf scorch with R² 0.13 and SCN resistance (Kazi et al. 2008). With a much higher resolution map constructed by high-density SNP markers, seven new SDS QTLs were discovered except confirming the previous mapped 7 SDS QTLs from the bi-parental
mapping population between PI438489B’ by ‘Hamilton’ (ABDEMALJID et al. 2012). These 7 “new” mapped QTLs made the total number of SDS QTLs reached to a total of 17 QTLs, 11th to 17th. So far, all the 17 SDS QTLs were identified from bi-parental mapping populations.

In order to exploit historical recombination events accumulated from artificial selection (breeding) and natural selection pressure, a powerful and complementary to bi-parental mapping strategy, genome-wide association study (GWAS), was used to map SDS QTLs. A total of 20 QTLs were discovered from two association mapping panel, and 7 of the 20 loci was overlapped to previously mapped QTL intervals, and 13 of them have not been mapped before (WEN et al. 2014). Similarly, from more recently GWAS mapping, two novel SDS QTLs were identified from chromosomes 3 and 18 (BAO et al. 2015). By integrating epistasis analysis with GWAS, genome-wide epistasis study (GWES) used employed to discover both SDS additive and interaction effects for SDS, four new SDS QTLs and 12 SNP-SNP interaction associated with SDS resistance were identified and the epistatic effects increase SDS resistance by 5 to 20% and provide a substantial complement to additive effects. (ZHANG et al. 2015)

From above literature search and summary, 17 SDS QTLs were mapping by bi-parental mapping populations and stored in soyBase.org [http://www.soybase.org](http://www.soybase.org), and 19 novel SDS QTLs, 13 from (WEN et al. 2014), two from (BAO et al. 2015), four from (ZHANG et al. 2015) were mapped from GWAS, with a total of 36 unique SDS QTLs were mapped. From epistasis based GWES, 12 SNP-SNP interaction were identified.

Even though many SDS tolerant QTLs have been mapped, SDS QTL-based selection was not widely used for SDS resistant variety breeding. There may be many reasons behind it, such as 1) no SDS QTL has been proved to work across different populations or genetic backgrounds. QTLs were mapped in bi-parental populations, and these QTLs were population-specific and did
not work in other genetic backgrounds. From the testing the usefulness of 10 SDS QTLs across six populations, no single QTL works across these six populations (LUCKEW et al. 2013); 2) no stable major SDS QTL has been identified. A total of 17 QTLs were mapped from bi-parental mapping populations, and their $R^2$ values range from 2% to 63% with small population sizes ranged from 40 to 80 per population (MEKSEM et al. 1999), but there is no major SDS QTLs as stable and major as soybean Cyst nematode SCN resistant QTL from PI 88788 which has been provided 99% of the SCN resistance (DIERS et al. 1997); this observation was further confirmed by GWAS study which 20 QTLs were mapped ranged from 5.3% to 11.6% with average 7.4% (WEN et al. 2014); 12 QTLs ranged from 6% to 9% with average 7.5% (ZHANG et al. 2015); 3) impractical to stack multiple QTLs with minor QTL effects. Since SDS QTLs are quantitative, and the effects are less than 10%, breeding SDS commercial varieties need to stack multiple QTLs. Soybean is self-pollinated cleistogamous (pollinated before the flower opens) crop, with current pollination technology, it would be impractical for soybean breeders to pyramid more than 6 QTLs from different populations into a single genetic background because an unusually large population is needed (LUCKEW et al. 2013); 4) both bi-parental mapping and GWA study in soybean mapped only the additive gene/QTL effects via single locus significant test or interval mapping, and epistasis and other gene interaction effects in soybean breeding have not to be reported by now. QTL- breeding either based on individual SNP selection or combination of SNPs as haplotype assistant selection. However, additive-effect only based selection may not be sufficient to explain the complexity of disease causality. It has been established that gene-gene/SNP-SNP interactions may have a higher impact on discovering causality of complex human disease (REAMS et al. 2011; LIN et al. 2013). There may exists two levels of interaction: first level interaction is pairwise interaction between QTLs/genes; second level of interaction is
three-dimensional interaction among all these QTLs which may coordinate and cross-regulate each other in a network and function as one integrated unit. When only partial QTLs of the network were tracked and traced in MAS selection, the critical or bottleneck part of the network may be monomorphic and block the pathway; this may lead to the failure of QTL-based breeding for the complexity of the trait of SDS.

**Interaction network of genes associated with traits of interest**

New studies and potential solutions have been publishing to understand human diseases and may be applied to soybean SDS breeding. SNP-by-SNP interaction network study has been reported to dissect human disease. A synergetic SNP-by-SNP network of alcoholic addiction was constructed by two steps: step 1 to form the SNP-by-SNP interaction network based on prior biological knowledge and their correlation between the functional relationships of their genes; step 2 to prioritize disease-risk SNPs via their differentially inherited properties in identity by descent (IBD) (Li et al. 2011a). A weighted SNP interaction hub network was constructed for human complex diseases and traits using whole-genome genotype data (Kogelman and Kadarmideen 2014). By combing two powerful machine learning methods: random forest (RF) and multivariate adaptive regression spline (MARS), an SNP-by-SNP interaction network associated with prostate cancer was constructed (Lin et al. 2012; Lin et al. 2013). All these reports from the study of human diseases show that SNP-by-SNP network-based prediction provides a promising opportunity to increase prediction accuracy. The goal of this study was to investigate the interaction among soybean SDS QTLs and to build both DNA and protein level interaction network of soybean SDS resistance by integrating the strengths of the four machine learning methods through 2 steps: step 1 to find the overlap SNPs from parametric-based stepwise regression with its strength in identifying a subset of most predictive variables from a large number of variables, such as high-density SNP chip data (Mitchell et al. 2001), genome-
wide association with its strength in exploiting historical recombination events at the population level to identify more QTLs across different germplasm (Teo 2008), and non-parametric based random forest with its strength in selecting important variables based on decision tree (Chen et al. 2011); step 2 to identify the interaction by MARS (Friedman and Roosen 1995) with its strength in detecting SNP-by-SNP interaction patterns (Lin et al. 2012). Here we present the first DNA level QTL interaction network, protein by protein interaction (PPI) network for soybean SDS. The identification of these QTLs in the interaction network will increase our understanding of mechanisms underlying SDS resistance, and network-based prediction by MARS via R package “earth” (http://www.milbo.users.sonic.net/earth/) provides novel marker selection system for breeding soybean lines with SDS resistance.

**Part 3: Spatial data analyses for adjusting phenotypic data**

Spatial data refers to types of data objects or elements that are present in a geographical space or horizon. Spatial data is also known as geospatial data, spatial information, or geographic information. It is the data or information that identifies the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and breeding testing sites, etc. Spatial data is usually stored as coordinates and topology and is data that can be mapped. Spatial data is often accessed, manipulated, or analyzed through Geographic Information Systems (GIS).

In this dissertation, spatial data refers to any phenotypic data taken from field coordinated by either by a combination of row and column numbers, or combination of Global Positioning System (GPS) longitude and latitude coordinates, such as testing plots planted by AccuRow’s GPS-guided planter (Raven Industries Sioux Falls, South Dakota, USA).
What is spatial data analysis?

Spatial analysis is a set of techniques for analyzing spatial data. The results of spatial analysis are dependent on the locations of the objects being analyzed. Software that implements spatial analysis techniques requires access to both the locations of objects and their attributes. The objectives of spatial analysis are to determine the spatial distribution of a variable, the relationship between the spatial distribution of variables, and the association of the variables of an area. It refers to the analysis of phenomena distributed in space and having physical dimensions (MAGUIRE et al. 2005). In GIS, there are four traditional types of spatial analysis: spatial overlay and contiguity analysis, surface analysis, linear analysis, and raster analysis.

In this dissertation, spatial analysis refers to spatial surface analysis with different models to remove spatial autocorrelation and spatial dependence, which is "everything is related to everything else, but near things are more related than distant things." This is the well-known “The First Law of Geography” by Waldo Tobler and is the foundation of the fundamental concepts of spatial dependence and spatial autocorrelation adjustment (CHAKRABORTY 2011; KLIPPEL et al. 2011; WESTLUND 2013). Specifically, the row and columns are used as spatial coordinates to locate the experiment plots to analyze the distribution of the plots, autocorrelation of the neighbor plots, and patterns of the plots.

Spatial pattern recognition and phenotypic adjustment to improve prediction accuracy

Figure 2. Diagram of genomic prediction. The training materials are genotyped and phenotyped to 'train' the genomic selection (GS) prediction model. Genotyped but not phenotyped breeding lines are then fed into the model to calculate genomic estimated breeding values (GEBV) for these lines.
Consider a genomic prediction analytic flowchart, such as depicted in Figure 2. There are three essential elements that will affect the accuracy of estimated predictions used for the selection: 1) quantity and quality of genotypic data, 2) quantity and quality of phenotypic data, 3) choice of appropriate analytic models for predictions. Among the three elements, the quantity and quality of genotypic assays are the most advanced. It is possible to obtain high-quality assays, as judged by reliability and repeatability, at sufficient density throughout the genome to assure those causative alleles for all traits of interest are in linkage disequilibrium (LD) in any arbitrary sample from a population of interest.

In contrast, the quality of phenotypic data, as judged by reliability and repeatability, is not at the same level as genotypic data. This is due in part to non-genetic sources of variability that affect the phenotype. These non-genetic effects include field soil characteristics that are not constant across a field as well as weather characteristics that are not consistent in space or time. It is hoped that phenotyping costs will decrease with the emergence of low-cost image acquisition and processing through computer-aided robots and drones, but data from these technologies are still being adapted to meet the data requirements of plant breeding objectives.

Many phenotypic traits, such as soybean Iron Deficiency Chlorosis (IDC), are strongly influenced by spatial variation in soil pH, soluble iron content, moisture, as well as other unknown environmental factors. Field plot experimental designs attempt to account for non-genetic sources of variability through the use of check plots (Muller et al. 2010), replication and blocking. Examples of field plot designs that use these techniques include a randomized complete block (RCB) and incomplete block designs such as $\alpha$-lattice and augmented designs. Although incomplete block design, such as $\alpha$-lattice design, usually account for a large amount
of field plot heterogeneity, often patterns of spatial variability remain; especially if the orientation of irregular spatial patterns are not consistent with block arrangements (LEISER et al. 2012).

**Spatial adjustment for phenotypic data**

Analyses of Variance (ANOVA) using ordinary least square (OLS) are based on three assumptions: Independence 1) model parameters are independent, 2) residual variability is identically distributed, and 3) the residual variability is normally distributed. Residual variability from field-based phenotypic data (grain yield, biomass, biotic, abiotic scores) often shows irregular spatial patterns. These data without adjustment by spatial analysis presents a challenge to the parametric and distribution assumptions. These problems are typically seen as various representations of spatial structure or non-independence. The spatial structure of the data can introduce spatial dependence into both the outcome, the predictors, and the model residuals. These data are correlated among neighbors, named autocorrelation, and the closer the distance between neighbors, the tighter that correlations are. Both positive and negative autocorrelations for either the dependent variable, the model predictors, or the model residuals.

Various spatial adjustment techniques have been developed and have been shown to significantly improve heritability and repeatability estimates resulting in a more effective selection of targeted traits. Based on the spatial adjustment mechanism and theory, three groups of approaches have been published:

1) **Group 1**: Spatial Autoregressive Regression (SAR) models including spatial lag models, spatial error models, and mixed spatial models (ANSELIN 2001; ANSELIN 2003; DORMANN et al. 2007a)

2) **Group 2**: Moving grid adjustment: a grid or pattern consisting of neighboring plots are predefined for each plot, the mean of the plots included in the grid is calculated and used
as a covariate to account for the spatial variation (TECHNOW 2015b). This approach has been reported to increase genomic selection accuracy in wheat breeding (LADO et al. 2013a).

3) Group 3: Tensor product penalized spline models. These models include both bilinear polynomial and smooth splines components. The bilinear polynomial component includes three sub-variables: row spatial trend, column spatial trend, and interaction between row and columns. The smooth spline part includes five smooth additive spatial components (Rodríguez-Álvarez et al., 2017)

Spatial autoregressive regression (SAR) models

Spatial effects can be divided into two categories: spatial dependence and spatial heterogeneity. Spatial dependence deals with autocorrelation, and spatial heterogeneity refers to structural instability, either in the form of non-constant error variances in a regression model as heteroscedasticity or the form of variable regression coefficients (ANSELIN 2001; ANSELIN 2003). In the spatial regression model, autocorrelation among neighboring observations is handled by autoregression. This definition of autoregression is that a particular observation is a linear combination of its neighboring values. This autoregression introduces dependence into the data. Instead of specifying the autoregression structure directly, spatial autocorrelation is introduced through a global autocorrelation coefficient and a spatial proximity measure. There are three basic forms of the spatial autoregressive model (SAR): the spatial lag, the spatial error, and mixed spatial models (SOKAL et al. 1998a; SOKAL et al. 1998b; DORMANN 2007).

The spatial lag models

The spatial lag model was derived from ordinary least square (OLS) by adding a spatial “lag” term (ANSELIN 1988b; ANSELIN AND FLORAX 1995). The spatial lag model is represented as:
$Y = X\beta + \epsilon$

Where $Y$ is the dependent variable, vector $n \times 1$, $X$ is the matrix of $n \times k$ of independent variables, $\beta$ is the vector of regression parameters to be estimated from the data $k \times 1$, and $\epsilon$ is the model residuals, vector $n \times 1$, which are assumed to be distributed as a Gaussian random variable with mean 0 and constant variance-covariance matrix $\Sigma$.

The spatial lag model introduces autocorrelation into the regression model by lagging the dependent variables themselves (Anselin and Florax 1995; Pace and LeSage 2010). The model is specified as:

$$Y = \rho W_y + X\beta + \epsilon$$

Where $\rho$ is the autoregressive coefficient, which tells us how strong the resemblance is, on average, between $Y_i$ and its neighbors. The matrix $W_y$ is the spatial weight matrix, describing the spatial network structure of the observations. The spatial weight matrix can be estimated either based on the distances between the $i^{th}$ and $j^{th}$ sites or plots or based on the correlation between any two points. The spatial lag model is appropriate when the focus of interest is the assessment of the existence and strength of spatial interaction. This is interpreted as substantive spatial dependence in the sense of being directly related to a spatial model (Anselin 2003).

**The spatial error model**

In comparison with the spatial lag model, the spatial error model does not add the autocorrelation covariates in the outcome. The spatial error model assumes that the autoregressive process occurs only in the error term and not in the response nor in predictor variables. In this case, the traditional regression model $Y = X\beta + \epsilon$ is complemented by a term $\lambda W_u$, which represents the spatial structure $\lambda W$ in the spatially dependent error. In the equation, the autocorrelation is added as part of the spatial covariates in the residual data as a covariate.
\[ Y = X\beta + \lambda W u + \epsilon \]

Where \( \lambda \) is the spatial autoregression coefficient, and \( W u \) is the spatial structure in the spatially dependent error term \( \epsilon \) (Dormann et al. 2007a).

The spatial error autocorrelation model, as its name stages, controls for the nuisance of correlated errors in the data that are attributable to an inherently spatial process, or to spatial autocorrelation in the measurement errors of measured and possibly unmeasured variables in the model.

**The mixed spatial model**

The mixed spatial model, including both response and predictor variables, was written as:

\[ Y = \rho W y + X\beta + WX_y + \epsilon \]

Where term \( WX_y \) describes the regression coefficients \( (y) \) of the spatially lagged predictors \( (WX) \) (Dormann et al. 2007a; Sparks 2011). This mixed spatial model is different from the general spatial mixed model, which refers to a model that has both random and fixed effect spatial variables (Negash et al. 2014). In order to distinguish the two terms, we use the SAR-mixed model.

**Model comparison and Lagrange Multiplier Test (LMT)**

Three spatial autoregressive models: \( SAR_{error}, SAR_{lag}, SAR_{mixed} \) all allow us to model spatial dependence in the data. Which model best fits data with spatial patterns? A common metric for comparison is the Lagrange multiplier score test (LMT). LMT compares the models using methods based on the relative change in the first derivative of the likelihood function around the maximum likelihood (Anselin 1988a; Baltagi and Bresson 2011; Robinson and Rossi 2014).
Application of SAR models

Before applying any statistical model to correct for spatial variation, spatial autocorrelation needs to be confirmed to impact the phenotypic data. Spatial autocorrelation can be visualized by heatmap of either the original phenotypic data or the residual plots, and can be measured by one of three standard parameters: Moran’s I (PERRY et al. 2002), Geary’s c, and semi-variograms (ISAAKS AND SRIVASTAVA 1989). These three metrics of either similarity (Moran’s I and Geary’s C) or variances of any two data points/plots. Among the three spatial evaluation parameters, Moran’s “I” is the most used statistics to judge whether there is any spatial autocorrelation (ANSELIN 2001). The Moran’s “I” ranges from -1 to +1. A higher positive Moran’s “I” indicates high spatial autocorrelation, which implies positive neighbors tend to cluster together. A lower negative Moran’s “I” is an indication that high and low values are interspersed. When Moran’s “I” is near or equal to zero, there is no spatial autocorrelation, which can be interpreted as the data are randomly distributed.

Applications of spatial analyses began with nearest neighbor methods (Papadakis, 1937) based on averages of neighboring plot responses as a covariate or by transforming the dependent variable by subtracting the moving averages of neighboring plots to account for spatial effects (STROUP 2002). A more sophisticated neighbor adjustment technique known as geostatistical kriging added spatial variance-covariance structure in the mixed model framework to account for spatial variability (KRIGE 1994). (KISSLING AND CARL 2008) compared three spatial autoregressive regression modes (SAR_{error}, SAR_{lag}, SAR_{mixed}) to species distribution in New Zealand and the results showed that SAR_{error} model was the best model and performed well in all cases in terms of minimum residual spatial autocorrelation (minRSA), maximum model fit (R2), or Akaike information criterion (AIC). In contrast, SAR_{lag} and SAR_{mixed} models did not
control type I error well and generated unpredictable biases in model parameter estimates. By integrating experiment design with different spatial model analysis, a comparison between the nearest neighbor method and mixed spatial models was conducted, and similar results were obtained (LITTELL et al. 1998). By comparing different experimental design in the presence of spatial variability, Stroup et al. (2002) concluded that incomplete block design designed specifically to mitigate spatial effects within block heterogeneity are more powerful than complete block design used measured by detection power. Explicitly, mixed spatial models used with incomplete block designs resulted in the best results (STROUP 2002).

A more recent and extensive comparison of models covering a range of spatial patterns was assessed by (LEISER et al. 2012) using the effectiveness of selection as a criterion for evaluating seed yields from legume and cereal crops grown in field trials. No single model could describe the spatial pattern in all situations. For each trial, the most appropriate model was identified. These ‘best’ models varied in their efficiencies for genotype comparisons being highest for Kabuli chickpea and considerably lower for barley and lentil. The identification of the best model, and its use, led to different adjusted mean seed yields for the genotypes, which, in turn, changed their ranking. Compared with classical approaches, spatial analysis increased the efficiency of genotype selection for further higher-level field experiment evaluation.

In parallel to time series analysis, spatial stochastic processes are categorized as spatial autoregressive (SAR) and spatial moving average (SMA) process (FINGLETON 2008). The moving grid adjustment method is a spatial method based on SMA theory and used in plant breeding trials to adjust phenotypic value variation caused by environmental effects. The adjustment is made by using phenotypic information from the nearest neighbors as a covariate. Unlike in other nearest neighbor methods, the moving grid adjustment is not determined by a
measure of distance but by inclusion in a grid of certain size and shape predefined by the researcher (TECHNOW 2015b). The trial is arranged in a spatial row-column like the design. Each entry is then associated with a cell, respectively, with a row and column number. A grid is constructed around each cell (= entry), and the mean of the cells included in the grid is calculated. The moving mean of the \( i \)th entry, denoted as \( x_i \), is calculated as:

\[
x_i = \frac{\sum_j p_{j,obs} \cdot I(p_{j,obs} \in G_i)}{\sum_j I(p_{j,obs} \in G_i)}
\]  

(TECHNOW 2015b)

The grid of entry \( i \) is denoted by \( G_i \) and \( I(\cdot) \) is an indicator function that takes the value one if the condition is satisfied and 0 if not. The observed phenotypic values of all entries which could potentially be included in \( G_i \) are denoted by \( P_{j,obs} \). The value \( x_i \) is taken as a measure of the growing conditions for the entry \( i \) and is used as a covariate to calculate an adjusted phenotypic value \( P_{i,adj} \) according to the following formula:

\[
P_{i,adj} = P_{i,obs} - b(x_i - \bar{x})
\]

Where \( P_{i,obs} \) are the observed phenotypic values of the \( i \)th entry, \( \bar{x} \) is the mean of all the \( x_i \).

Two prerequisite conditions must be met in order to make the moving grid work for spatial variation adjustment successful: 1) the entries must not influence the values of their covariates \( (x_i) \), and 2) the entries must have been randomly assigned to plots in the field. If one or both are not met, the spatial adjustment can lead to biased outcomes.

The procedure to use moving grid adjustment includes two steps: 1) pre-define a grid consisting of one, two, three, etc. neighbors. The shape of the grid can be irregular (Figure 3). A single grid can be applied to the entire location, or different grid can be applied to sub-locations; 2) calculate the mean of the grid for each entry and use it as a covariate.
Applying moving means as a covariate to account for spatial variability in soybean IDC was reported as early as 1991 (Diers et al. 1991). One of the objectives of the study was to determine if the moving-mean analysis procedure would reduce the error variance in Fe-efficiency tests and improve selection among genotypes. These results show that, although the moving-mean analysis reduced the error variance, it did not adjust means in a way that would have resulted in improved selection compared with unadjusted means. In wheat breeding, comparison between unadjusted yields with adjusted yield by moving mean showed an increase of correlation coefficient between two trials within an experiment from 0.41 to 0.48. This increase indicated an average increase of within-trial precision of 20% by using moving mean analysis (Clarke et al. 1994). In the past two decades, moving average or mean-adjustment for spatial variability was widely used for post-data collection (Mur 1999; Hoeff et al. 2004; Bakrim and Aboutajidine 2006; Wu et al. 2006; Basrak and Tafro 2014).
(Mur and Angulo 2007) compared moving grid adjustments with spatial autoregressive models using Monte Carlo analyses. They concluded that the power of the models increased, but the choice of model depended on the situation and the criteria associated with the objectives of the research. The most recent application of moving grid mean adjustment was in genomic prediction in hybrid wheat breeding (Lado et al. 2013a). Among the tested four models, a mixed model using moving means as a covariate was found to best fit the data, and the estimated accuracies from genomic prediction accuracy was increased for grain yield, thousand kernel weight, and days to heading. The conclusion from the study is that the correction of spatial variation is a crucial ingredient to increase prediction accuracy in genomic selection models.

The tensor product is an outer product of two coordinate vectors, a special case of the Kronecker product of matrices. The outer product of two coordinate vectors $u$ and $v$ and the matrix $w = u \odot v$ such that the coordinates satisfy $w_{i,j} = u_i v_j$. For example, an experiment field has 20 rows 30 columns, we can think them as two vectors: row vector $u_i = (u_1, u_2, u_3, \ldots u_{20})$ with 20 x 1 dimension, a column vector $v_j = (v_1, v_2, v, \ldots v_{30})^T$ with 1 x 30 dimension, then the tensor product of row and column vector is a matrix that has 20 x 30 matrix plots.

The term P-spline stands for "penalized B-spline." It refers to using the B-spline representation where the coefficients are determined partly by the data to be fitted, and partly by an additional penalty function that aims to impose smoothness to avoid overfitting (Chiu et al. 1996; Eilers and Marx 1996). P-splines can be either one or more than one dimension. The tensor product penalizes splines (P-splines) is a 2-dimension P-splines used to account for the interaction between row and column variation trends in the field.
Spatial analysis models can be divided into two categories based on how to model the spatial trends: the traditional autoregressive model with spatial variance-covariance structures and smoothing spline methods that model spatial trends with splines (Rodríguez Alvarez et al., 2016). The first report using smoothing techniques in agricultural experimental field trials was in 1985 (Green et al. 1985). The report showed that least-squares smoothing trend plus independent error model for the environmental effects in the yield of field plot experiment worked well for estimating the main treatment effects and unknown spatial trend. The least-square smoothing approach can be extended from one-dimension to two-dimension smoothing splines. In 1999, the two-dimension splines were extended to cubic smoothing splines in a multi-environment field experiment for a variety of testing of lupine (Kempton et al. 1999; Verbyla et al. 1999). In the proposed model, the cubic smoothing splines were used to conjunction with fixed and random effects, random coefficients, and variance modeling to provide simultaneous modeling of trends and covariance structure. The results showed that the linear mixed model representation of the cubic smoothing spline provides flexible spatial adjustments. In 2003, a semiparametric spatial model was implemented with an optimal search for the degree of smoothing in barley field trials (Durban et al. 2003). Results from the model comparison between the semiparametric with smoothing spline and classical analyses of variance and with alternative spatial models showed that semiparametric models generated an interpretable insight into spatial variation in the field and can improve the precision of parameter estimates. A two-dimension smoothing surface model, named tensor product penalized splines were first proposed to model spatial dependence in 2003 (Eilers and Marx 2003). Most recently, this two-dimension smooth surface model, named tensor product penalized splines were implemented in an R package named “SpATS” (Rodríguez Alvarez et al., 2016). The model implemented in
SpATS is a mixed model representation of P-splines with a generalized framework for the analysis of field trials. It can include both fixed and random parameters, such as genotypic effects and the corrections for rows and columns. The model equation is:

\[ y = Z_r c_r + Z_c c_c + \int (u, v) + \epsilon \]

\[ \int (u, v) = 1_n \beta_0 + u \beta_1 + v \beta_2 + u \odot v \beta_3 + f(u) + g(v) + u \odot h(v) + v \odot h(u) + f(u,v)(u, v) \]

- **Y**: \( y_1, y_2, \ldots, y_{720} \), a 720 x 1 column vector (for easy demo, we use the number from the paper, use 15 rows by 48 columns, with a total of 720 plots)

- **Z_r**: 720 x 15 matrix

- **Cr**: 15 x 1 column vector \( C_r = (c_{r1}, c_{r2}, \ldots, c_{r15}) \) with \( Cr \sim N(0, \sigma^2_r 15) \), random effect coefficients for rows

- **Z_c**: 720 x 48 matrix

- **Cc**: 48 x 1 column vector \( C_c = (c_{c1}, c_{c2}, \ldots, c_{c15}) \) with \( Cc \sim N(0, \sigma^2_r 148) \), random effect coefficients for columns

- **u**: \( = (u_1, u_2, \ldots, u_{720}) \) column vector

- **v**: \( = (v_1, v_2, \ldots, v_{720}) \) column vector

- **f (u,v)**: \( f (u_1,v_1), \ldots, f (u_{720},v_{720}) \) with \( f (...) \) representing a smooth bivariate function

- \( \beta_0 \): the intercept

- \( \beta_1 \): the linear trends along the row direction

- \( \beta_2 \): the linear trends along the column direction
• $\odot$ denotes the element-wise vector product

• $\beta_3$: the linear interaction trend

• $f_u(u)$: is a smooth trend along the rows, identical for all columns (i.e., a main smooth effect)

• $f_v(v)$: is a smooth trend among the columns, identical for all rows

• $u \odot h_v(v) \text{ and } v \odot h_u(u)$: are linear by smooth interaction trends. For instance, $u \odot h_v(v)$ is a varying coefficient surface trend, consisting of functions, linear in the rows, for each column, but with slopes that change smoothly along with the columns, $h_v(v)$

• $f_{(u,v)}(u, v)$: is a smooth-by-smooth interaction trend jointly defined over the row and column directions

In this complex model, the term $f_{(u,v)}$ is the tensor product with two-dimension P-splines. The model can be understood as consisting of two parts: bilinear polynomial and a smoothing adjustment, highlighted in yellow and green, respectively. The functions $f_{(u),f(v),h_{(u)},h(v)}$ are constructed with variations of one-dimension P-splines. Since this model contains both fixed and random effects, it is a mixed spatial model, and each of the smoothing components and their amount of smoothing portion can be computed using restricted maximum likelihood (REML).

The proposed SPATS model has the following properties: 1) it can explicitly estimate the spatial trend in the field; 2) it provide repeated and fast estimation; 3) it can still generate desired results even with a large portion of missing data, and 4) it can be used for the data with non-normal response data.

Although spatial autocorrelation was defined decades ago, its application has been limited by computation capacity and software availability (Anselin 1988b; Anselin and
FLORAX 1995; PACE AND LE SAGE 2010). Three types of models were developed and have been used for spatial adjustments of some quantitative traits. Spatial autoregressive based models adjust spatial variation based on either neighbor distance or correlation as a covariate; in contrast, moving grid average mean adjustment uses the mean of the neighbor plots as a covariate. Smoothing trend surface splines and tensor product penalized P-splines represent the most recent theoretical model to account for patterns of spatial variation. All three types of models have been reported to increase phenotypic data quality and hence to increase the heritability of the traits of interests. There is no report on the application of these statistics to understand spatial patterns for ordinal agronomic traits such as IDC and SDS.

**Part 4: Genomic prediction methods and machine learning algorithms**

Genomic selection or genome-wide selection (GS) is an approach proposed by Meuwissen (MEUWISSEN et al. 2001) to improve the quantitative trait breeding that uses genome-wide marker information to estimate the genetic breeding value (EBV) (MEUWISSEN et al. 2001; BERNARDO 2008). GS uses a training pool of individuals that have been both genotyped and phenotyped to develop the model that takes genotypic data from a breeding pool of untested individuals and produces genomic estimated breeding values (GEBVs). In the original GS proposed by Meuwissen, in contrast with QTL-based MAS breeding, there is neither QTL mapping needed nor marker selection to minimize the biased estimations of marker effect. This approach, with all the markers in the model, usually leads to more predictors, p, need to be estimated than the number of observations, n. This is the so-called “large p and small n problem” (JANNINK et al. 2010). With more predictor variables than the number of the observations, least-square estimation, or standard multiple linear regression can’t be used to estimate the marker effects. And using all the markers in the model can lead to overfitting the model, which usually have poor predictive ability. To overcome these two problems, several methods, best linear
unbiased prediction (BLUP) (Kolbehdari et al. 2006; Kolbehdari et al. 2007; Ober et al. 2011), ridge regression (Endelman 2011; Atashi and Gorgani-Firouzjah 2015), Bayesian regression (Meuwissen et al. 2001), kernel regression (Gianola and van Kaam 2008; Nosedal-Sanchez et al. 2012; Garcia-Cortes et al. 2014), has been proposed to develop prediction models for GS.

Methods to estimate the marker effects for GS have been changing since the invention of GS. From the original report, Meuwissen proposed to model them as random effects and calculate their best linear unbiased prediction (BLUP) (Meuwissen et al. 2001). The random effects were drawn from a normal distribution with mean 0, variance $\sigma_g^2$ ($X_i \sim N(0, \sigma_g^2)$), where $\sigma_g^2$ was estimated by dividing the total genetic variance by the number of effects (Habier et al. 2007; Iwata 2011). The assumption of evenly distributed and equal variance of the genetic causation marker was not realistic and improved to a more realistic model: BayesA and BayesB. In the BayesA model, each marker has an effect and each effect was drawn from a normal distribution with its own variance which was sampled from a scaled inverted chi-square distribution; in comparison with BayesA, BayesB has added a probability factor, $\pi$, for marker effects, and not all the markers have effects, some of the markers may have 0 effects which are more realistic. In contrast to directly estimate each individual marker effects by rrBLUP, or BayesA, or BayesB, and in order to reduce the overfitting the model by adding all the markers, latent variable, which is extracted as linear combinations of the predictors and is used for response prediction, was introduced from dimension-reduce regression methods such as Principal component analysis (PCA) (Zhang et al. 2012) and partial least square regression PLS (Vargas et al. 1998; Fu et al. 2012). In PCA, the latent variables are selected to explain as much of the variation in the original predictors as possible. In PLS, the latent variables are
chosen so that the relationship between the latent variables and responses is as strong as possible (JANNINK et al. 2010).

Genomic prediction accuracy was usually estimated by Pearson correlation, $r$, which measures the degree of correlation or consistency between genomic estimated breeding values (GEBV) and the observed phenotypic values (HEFFNER et al. 2010; HEFFNER et al. 2011a; HEFFNER et al. 2011b). Different breeding experiment designs for genomic selections have been proposed and studied in maize (LIAN et al. 2015; GORJANC et al. 2016), barley (IWATA AND JANNINK 2011; SCHMID AND THORWARTH 2014; SALLAM et al. 2015b), apple (MURANTY et al. 2015), and animals (HAYES et al. 2009a; MONTALDO et al. 2012). Regardless of the breeding designs, the prediction accuracy needs to be high enough for genomic selection to be time and cost-effective. The expected prediction accuracy $E(r_{mg})$ has been previously defined as a function of population size, $N$, heritability of the trait of interested, $h^2$, and number of effective markers that affect the trait of interested, $M_e$. The formula is:

$$E(r_{mg}) = \left[ \frac{N h^2}{(N h^2 + M_e)} \right]^{1/2}$$ (DAETWYLER et al. 2008; HAYES et al. 2009b)

From the above equation, the heritability of the trait of interested, $h^2$, is fixed for the specific trait, population size $N$ is under control of breeders, the unclear variable and hard to figure out is the number of effective chromosome fragment, $M_e$ that related the specific trait. The $M_e$ pertains to the number of the idealized concept of independent chromosome fragments, with each fragment containing a QTL-marker pair (DAETWYLER et al. 2008; LIAN et al. 2014). When the marker density is high enough to cover every functional gene in the genome, $M_e$ can be estimated on the basis of the effective population size and the genome size (LORENZ 2013; LORENZ AND SMITH 2015).
In order to estimate the prediction accuracy correctly using the equation from Daetwyler (DAETWYLER et al. 2008), four statistical assumptions are required to meet: 1) the marker effects were derived from simple linear regression; 2) each marker–QTL pair were assumed independent of other markers–QTL pairs, neither linkage nor interaction (epistasis); 3) different marker–QTL pairs were assumed to have equal variances; 4) each marker–QTL pair was assumed in complete linkage disequilibrium (LD). Since not all markers in the high-density chip in completely linkage disequilibrium block and to accounting for incomplete linkage disequilibrium between an SNP and a QTL, the equation was updated by Lian et al. (LIAN et al. 2014) to retain the first three assumptions but to relax the 4th assumption by adding an extra variable $r^2_{MM/2}$ in the equation as below:

$$E(r_{mg}) = r^2_{MM/2} \left[ \frac{N h^2}{(r^2_{MM/2} N h^2 + M_e)} \right]^{1/2} \text{(LIAN et al. 2014)}$$

Where $r^2_{MM/2}$ equals to the mean squared correlation between a marker and QTL when the QTL is assumed to be at the midpoint of the two markers.

Prediction accuracy of the GS has been reported from different species by different statistical models with different populations or panel sizes. Here are the two representative reports with either large number lines per population from barley or a large number of populations in maize. In barley, by analysis of 1536 SNPs and 647 lines with four traits differing in genetic architecture from 5 years phenotypic data in barley, Sallam et al. concluded that higher trait heritability, $h^2$, in the training population and simpler trait architecture were associated with higher prediction accuracy, and fixation of markers/loci associated with a trait over time was most clearly associated with reduced prediction accuracy, and the prediction accuracy ranged from 0.03 to 0.99 across the four traits and five progeny sets (SALLAM et al. 2015a). In maize, the progeny of 969 bi-parental populations were genotyped with 31 to 119 (with a mean of 70)
polymorphic SNPs, and GEBVs were calculated for seven traits and the prediction accuracy range from -0.59 to 1.03 with a mean of 0.45 for grain yield, and range from -0.34 to 0.96 with a mean of 0.59 for moisture, and authors concluded that it is difficult to predict the accuracy, $r_{MG}$ in advance, but the rule of thumb based on $r_{MM/2}^2(Nh^2)^{1/2} > 8$ can help to increase prediction accuracy (LIAN et al. 2014).

Package rrBLUP has been widely used for GS because of its open-source and easy use. But rrBLUP can only be applied to estimate additive marker effects. Dominance effect was added in the linear mixed model:

$$y = Xb + Za + Zd + e$$

(NISHIO AND SATOH 2014)

where $y$ is the vector of phenotypes; “$b$” is the vector of fixed effects; “$a$” and “$d$” are the vector of additive and dominance genetic effects; “$X$” and “$Z$” are incident matrices for the fixed effects, additive, and dominance genetic effect, respectively; and $e$ is the vector of residuals.

Results from simulated data show that GBLUP with both dominance and additive effects is a feasible approach to improve genetic performance in crossbred populations with large dominance genetic variation and identify mating systems or best crosses with a good combining ability (PEREZ-RODRIGUEZ et al. 2013; NISHIO AND SATOH 2014). A more complicated model that can include additive, dominance, and epistasis was implemented in r package “sommer,” and results from both simulated and field phenotypic data show prediction accuracy from “sommer” can be increased in species displaying heterotic effects, which require the estimate of general combing ability (GCA)/dominance and special combining ability (SCA)/epistasis (COVARRUBIAS-PAZARAN 2016).

Since GS was first proposed in 2001 (MEUWISSEN et al. 2001), it has been widely studied, and the results clearly indicated that GS was more predictive than classical marker-
assisted selection in both empirical and simulated data (LORENZANA AND BERNARDO 2009; HEFFNER et al. 2010; HEFFNER et al. 2011b; LORENZ 2013; LORENZ AND SMITH 2015). But results from some of the studies show that there is a very wide range of prediction accuracy, for example, from -0.59 to 1.03 for maize grain yield (LIAN et al. 2014); from -0.41 to 0.94 for palm oil yield (CROS et al. 2015). There may be many reasons for these observations. One of the reasons probably is because of the choice of method/models that are chosen for the data analysis. The traditional methods for data analysis, such as linear regression and linear discriminant analysis, are based on predefined distributions with required model assumptions. These methods can have a high prediction accuracy only for the data that meet the model requirements. For example, three assumptions for linear regression include 1) Linear relationship between the phenotype and marker; 2) Multivariate normality; 3) No multicollinearity among the predictors variable; for example, SNP markers. Clearly, most of the data set for GS does not strictly comply with 2 of the three assumptions, linear relationship, and multicollinearity because most SNPs from the high-density chip are linked or co-segregated with the trait of interested. Another example of Fisher’s linear discriminant regression requires that the classes can be separated linearly and each of two classes can be represented by a unimodal Gaussian distribution (Bishop 2006).

From data analysis to draw a conclusion, two different mechanism approaches have been applied to analyze the data: pre-defined distribution-based stochastic data model (normal, passion, chi-square, etc.) and non-distribution algorithmic models (random forest classification, and support vector machine algorithm, etc.). The statistical community has been committed to the almost exclusive use of data models. Some of the researches have led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting,
current problems if the data sets do not meet the prerequisite model assumptions (Breiman 2001b; Cox et al. 2001). In contrast, non-distribution based algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling. The majority of the machine learning algorithms do not need model assumptions and prior knowledge of the data distribution; they belong to algorithm modeling and provide an alternative solution to distribution based data models to improve GS prediction accuracy.

Machine learning (ML) is a branch of artificial intelligence dealing with the design and development of algorithms that allow computers to evolve behaviors based on historical empirical data (Lavine et al. 2004; Veronese et al. 2013). It is a mathematical optimization process through computational statistical analysis of the data. ML is an algorithm that can learn from data without relying on rules-based programming. The primary goal of ML algorithms is to automatically learn to recognize the complex pattern and make intelligent decisions based on data. Comparing with traditional statistics that need predefined distributions with prerequisite statistical assumptions, ML allows the detection of patterns and extraction of important information from raw data even if the underlying data model and probability distribution are unknown (Breiman 2001a).

Typically, the data for ML is arranged in an “m” by “n” matrix $X_{m \times n}$ where each row ($x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,n}]$, I from 1 to n phenotype observations, and j from to m, number of markers). In some applications (supervised ML), target variables or labels $y_i$ is corresponding to $x_i$. Based on whether a $x_i$, has a label $y_i$. ML was divided into two main types: supervised and unsupervised ML (Crisci et al. 2012; Behmann et al. 2015). Unsupervised learning is the machine learning task of inferring a function to describe hidden structure from unlabeled data.
(no \( y_i \) in the dataset), while Supervised learning is the task of inferring a function from labeled training data. The most known example of unsupervised ML is clustering, for example, k-mean clustering. But, clustering can be supervised ML if true clustering of the data is known, such as clustering people by sex, or nationality, or education level, in contrast, clustering people by their height or weight, this clustering is unsupervised learning.

In contrast to the one-step clustering unsupervised ML algorithms, a supervised ML process is divided into two steps: in step I, a model (F) is learned by a specific ML algorithm from the labeled training data. The ML algorithm itself derives from these training data suitable model by optimizing the model quality via a method-inherent principle. For example, SVM classified, is maximum space margin principle is applied (CORTES AND VAPNIK 1995), whereas, with the neural network, a non-convex optimization with regard to prediction accuracy is applied (DAUTILIA AND GUERRA 1996; SWAIN 1997; ESHGHIZADEH et al. 2012). The goal of the learning process is the generation of a model that reveals the task-relevant patterns without relying on unjustifiable statistical assumptions; in step II, the derived model from step I is applied to new, unknown data and predict the best labels \( y \) from the test sample \( X \) by \( y_i = F(x_i) \). In the field of IDC breeding, the label \( y_i \) could be the IDC tolerance score 1 to 9, and the data sample \( x_i \) could be the marker data of line \( i \).

There are many ML algorithms available in both R and python packages. Here we go over the most popular ML algorithms that have been applied for different fields and will be employed for IDC score prediction 1 to 9 and IDC to three tolerant level classification, tolerant/resistant as “R,” susceptible as “S” and moderate tolerant/resistant as “M.”

Artificial Neural Network (ANN) models the relationship between a set of input signals or markers and an output signal or score using a model derived from our understanding of how a
biological brain responds to stimuli from sensory inputs. Just as a brain uses a network of interconnected cells called neurons to create a massive parallel processor, ANN uses a network of artificial neurons or nodes to solve learning problems (Lantz 2015). ANN can be applied to solve problems that involve complex but unknown relationships between variables and non-linear relationship between predictive and response variables (Crick 1989). ANN has many advantages over statistical-based techniques. They can learn from existing data and therefore do not require a prior model or statistical distribution and the outputs of a type of ANN, the multi-layer perceptron (MLP) can be interpreted as probabilities of an event occurring (Watts and Worner 2008; Watts and Worner 2012). ANN can be used for both supervised and unsupervised ML contexts. For modeling the relationship between the input and the output in ANN, additional neurons, so-called “hidden layers” are inserted between the predictive and response variables. These hidden layers are the training processes of ANN which transform the inputs signals/markers to a local output (Bishop 1995). The ANN algorithms optimize the weights of the neurons in the hidden layers with regards to a task-specific prediction function (Behmann et al. 2015).

ANN has been compared to traditional statistical models, and results show that the ANN has consistently outperformed the statistical models with respect to prediction accuracy. Comparison between ordination and ANN on the influence of environmental variables on the abundance of aquatic insects showed that the application of ANN with various methods of variable pre-selection increased the precision of the prediction (Wagner et al. 2000). By comparing logistic regression and ANN-based classifiers for bacterial growth, ANN-based classifiers outperformed the logistic regression-based counterparts (Hajmeer and Basheer 2003). Results from the comparison between the statistical regression model and ANN on the
prediction of the dynamic density of river phytoplankton showed that the ANN model presented a high performance in prediction accuracy (Jeong et al. 2006). A literature review of ANN in solar radiation prediction and designing solar systems showed that ANN gives good accuracy in terms of the prediction error of less than 20%. And ANN, as compared to other empirical models, is capable of dealing with many input meteorological parameters, which make it more accurate and reliable (Qazi et al. 2015). Prediction accuracy of asymptomatic cirrhosis in chronic hepatitis C patients among three models of ANNs and logistic regression models (LogIt), and aspartate aminotransferase-to-platelet ratio index (APRI) showed that ANNs generated the best positive predictive value (0.86 vs. 0.67 and 0.56) and positive likelihood ratio (40.2 vs. 13.4 and 8.4), respectively. In the validation set, the discrimination power of ANNs (0.76) was significantly higher than those of LOGIT (0.67) and APRI (0.67) (Cazzaniga et al. 2009). In conclusion, ANN provide a promising option for prediction and will be employed for IDC prediction.

A Support Vector Machine (SVM) can be imagined as a surface that creates a boundary between points of data plotted in multidimensional that represent examples and their feature values. The goal of an SVM is to create a flat boundary called a hyperplane, which divides the space to create fairly homogeneous partitions on either side (Lantz 2015). The model of an SVM consists of a subset of data instances $x_i$—so-called support vecrors. They define a hyperplane separating classes into the feature space while its positions is unambiguously defined by the principle of maximum margin, thus facilitating the learning of models with optimal generalization performance (Behmann et al. 2015). SVMs are used for both supervised classification and regression, providing linear and non-linear discrimination function, and unsupervised clustering.
SVM has been compared to traditional statistical models and other ML algorithms, and results show that the SVM has outperformed the statistical models and other ML algorithms with respect to prediction accuracy. Results from the comparison of the decision tree (DT), SVM, and Bayesian network (BN) approaches for classification of falls in Parkinson's disease show that BN and SVM produce consistently higher accuracy with average sensitivity and specificity > 92% than DT with average sensitivity 88%, average specificity 72%. And both SVM and BN are not prone to imbalanced data comparing with DT, which needs to adjust for the misclassification cost. However, DT provides a straightforward, interpretable result and thus is appealing for helping to identify important items related to falls and to generate fallers' profiles (Sarini et al. 2015). Similar results from the direct comparison of SVM and ANN classification methods in land use showed that SVM and ANN with a total accuracy of 90.67% and 91.67% are superior. SVM had a better performance in separating classes with similar spectral profiles. In addition, SVM showed a better performance in delineating class borders in comparison with the neural network method. The authors concluded that both SVM and ANN showed satisfactory results, but the method of SVM proved better with a difference of 1% and 2% in overall accuracy and kappa coefficient, respectively. This was an expected outcome because SVMs are designed to locate an optimal separating hyperplane, while ANNs may not be able to locate this separating hyperplane (Mokhtari and Najafi 2015). The super classification power of SVM was further confirmed by a comparison of SVM, ANN, and spectral angle mapper (SAM) algorithms for crop classification using linear imaging and scanning data (Kumar et al. 2015). The results from the comparison indicate better performance of SVM and ANN algorithms in comparison to SAM for the classification. The overall accuracies obtained by SVM and ANN were 93.45% and 92.32%, respectively, whereas the lower accuracy of 74.99% was achieved using the SAM
algorithm. Results derived from SVM, ANN and SAM classification algorithms were validated with the ground truth information acquired by the field visit on the same day of satellite data acquisition. From clinic prostatic cancer data, the comparison results between SVM and ANN show that the area under curve (AUC, another accuracy measure of prediction) of SVM and ANN is 0.805 and 0.719, respectively (p = 0.020), in the pre-operative prediction of advanced prostate cancer. And authors concluded that the performance of SVM is superior to ANN in the pre-operative prediction of advanced prostate cancer (Kim et al. 2011).

Results from some experimental comparisons show that SVM is as good as other ML algorithms. An experimental comparison between SVM and random forest (RF) for hyperspectral image land cover classification shows that the performance of SVM is comparable to that of RF (Abe et al. 2014). The high prediction accuracy on land coverage classification was achieved by both SVM and ANN. The overall accuracy values of ANN classifiers and SVM classifiers were over 97% for land cover classification. SVM classifiers had slightly higher accuracy than ANN classifiers. The authors concluded that ANN and SVM methods should be especially useful for land cover classification (Guo et al. 2012).

Decision tree learning uses a decision tree as a predictive model that maps observations about an item to conclusions about the item's target value. There are two main types of decision trees: classification tree and regression tree. Classification tree analysis is when the predicted outcome is the class to which the data belongs, where regression tree analysis is when the predicted outcome is a continuous numeric value. There are three ensemble methods which predict the target value by constructing more than one decision trees:

**Bagging** decision trees, One of the first ensemble methods to gain widespread acceptance, used a technique called bootstrap aggregating or bagging for short. Bagging
generates a number of training datasets by bootstrap sampling the original training data. These datasets are then used to generate a set of models using a single learning algorithm. The models' predictions are combined using voting (for classification) or averaging (for numeric prediction) (Lantz 2015).

**Boosting Trees.** Another common ensemble-based method, is called boosting because it boosts the performance of weak learners to attain the performance of stronger learners. Similar to bagging, boosting uses ensembles of models trained on resampled data and a vote to determine the final prediction. There are two key distinctions. First, the resampled datasets in boosting are constructed specifically to generate complementary learners. Second, rather than giving each learner an equal vote, boosting gives each learner's vote a weight based on its past performance. Models that perform better have greater influence over the ensemble's final prediction.

**Random Forest decision trees (RF):** is the most widely used decision tree-based ML algorithm. Random forest is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and feature selection, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean-prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests combine versatility and power into a single machine learning approach. As the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large datasets, where the so-called "curse of dimensionality" might cause other models to fail. RF tends to be easier to use and less prone to overfitting (Chiu 2015; Lantz 2015).

RF has been applied to the different research field, and results show very promising output that RF usually achieve higher prediction accuracy and less computational time. Results
from the comparison between a traditional decision tree with an RF classifier on operational radar-only crop type classification show that accuracies were improved by approximately 4%-5% over traditional boosted decision trees with gains of up to 7% in the accuracies of specific classes. In addition, the RF classifier provided large performance gains in terms of processing time relative to the decision tree classifier (Deschamps et al. 2012).

RF algorithm does not always outperform other ML algorithms. Results from a comparison of Bayesian approaches to classification and regression trees (BCART) and RF to identify classifiers for childhood leukemia show that BCART had higher accuracy and, in particular better prediction of relapse (higher sensitivity) than RF. BCART also had better performance than RF in identifying important genes that predict whether a patient will relapse (O'Leary et al. 2009). By comparing of Random Forest and Adaboost Tree in Ecosystem Classification in East Mojave Desert, the authors concluded that the Adaboost tree could provide higher classification accuracy than RF, while the RF could be more efficient in computation (Miao and Heaton 2010). Results from a comparison of random forests, boosting, and SVM for the genomic selection show that the correlations between the predicted and true breeding values were 0.547 for boosting, 0.497 for SVMs, and 0.483 for RF, indicating better performance for boosting than for SVMs and RF. And authors concluded that accuracy was highest for boosting, intermediate for SVMs, and lowest for RF but differed little among the three methods and relative to ridge regression BLUP (RR-BLUP) (Ogutu et al. 2011). From a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, ANN, SVM, classification trees and RF, more complicated results were achieved which SVM showed the larger overall classification accuracy (Median (Me) = 0.76) an area under the ROC (Me = 0.90). However this method showed high specificity (Me = 1.0) but low
sensitivity \( (M_e = 0.3) \). In contrast to SVM, RF ranked second in overall accuracy \( (M_e = 0.73) \) with high area under the ROC \( (M_e = 0.73) \) specificity \( (M_e = 0.73) \) and high sensitivity \( (M_e = 0.64) \) (MAROCO et al. 2011).

Naive Bayes classifiers (NBC) are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features, which is why it was called “naïve” Bayes. Bayesian probability theory is rooted in the idea that the estimated likelihood of an event, or a potential outcome, should be based on the evidence at hand across multiple trials, or opportunities for the event to occur. Classifiers based on naïve Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new features. It's a simple idea, but it results in a method that often has results on par with more sophisticated algorithms (LANTZ 2015).

NBC is best applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the overall probability of an outcome. While many machine learning algorithms ignore features that have weak effects, Bayesian methods utilize all the available evidence to subtly change the predictions. If a large number of features have relatively minor effects, taken together, their combined impact could be quite large (LANTZ 2015).

NBC algorithm has been used for many research fields, and results show that NBC provides another option for ML prediction. By comparing the performance of NBC, an open-source Tariff Method (OTM), and InterVA-4 on three datasets covering about 21,000 child and adult deaths, Miasnikof observed that NBC scored the highest cause-specific mortality fraction
accuracy across the datasets (median 0.88, range 0.87-0.93), followed by InterVA-4 (median 0.66, range 0.62-0.73) and OTM (median 0.57, range 0.42-0.58). The authors concluded that NBC outperforms current similar causes of death classifiers (MIASNIKOF et al. 2015). Results from comparison between NBC and logistic regression for diagnosing breast masses using ultrasound imaging show that NBC showed significant accuracy variation (0.733 +/- 0.035 to 0.840 +/- 0.029, P < 0.002) with the choice of the total 34 features, but the performance of logistic regression was relatively unchanged under feature selection (0.839 +/- 0.029 to 0.859 +/- 0.028, P = 0.605). The probabilities of malignancy determined by NBC and logistic regression after feature selection showed a significant correlation (R-2 = 0.87, P < 0.0001) of the two methods. The authors concluded that the diagnostic performance of NBC and logistic regression could be comparable, but logistic regression is more robust with higher prediction accuracy. Since the probability of malignancy cannot be measured directly, a high correlation between the probabilities derived from two basic but dissimilar models increases confidence in the predictive power of machine learning models for characterizing solid breast masses on ultrasound (CARY et al. 2012). By comparing the C4.5 Decision Tree and NBC using the Ljubljana Breast Cancer dataset, Sivakumar concluded that both the classifiers achieve good accuracy on the dataset (SIVAKUMARI et al. 2009)

Regression-based ML can be grouped into two categories based on whether it is linear: linear regression will be used in this study include Ridge-regression BLUP (rrBLUP), least absolute shrinkage and selection operator (LASSO), Stepwise regression (SWR). Nonlinear regressions include Multivariate adaptive regression splines (MARS) and Multinomial logistic regression (MLR). Since rrBLUP, LASSO, SWR will be used as a reference, and other ML
algorithms will be compared with them with regards to prediction accuracy and computational efficiency, here we will summarize literature only on MARS and MLR:

MARS is a form of regression analysis introduced by Jerome H. Friedman in 1991. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically model nonlinearities and interactions between variables. As an example, consider the relationship between a quantitative trait, $Y$, and a continuous predictor, $X$. Assume $X$ is related to $Y$ via a non-linear function. MARS approximates this function by dividing the $X$ variable up into smaller regions. Within each region, MARS then fits a simple linear regression model. MARS was proposed as a flexible procedure to organize relationships between a set of input variables and the target-dependent that includes both additive effects from individual variables and epistasis effects from interactions among variables. It is a nonparametric statistical method based on a divide and conquers strategy in which the training data sets are partitioned into separate piecewise linear segments (splines) of differing gradients (slope). MARS makes no assumptions about the underlying functional relationships between dependent and independent variables (ZHANG AND GOH 2016). MARS has two important functions: calculate the epistasis among predictive variables and divided a non-linear relationship between predictive and response variables into multiple but connected linear regression. The term "MARS" is trademarked and licensed to Salford Systems. In order to avoid trademark infringements, an R package named "Earth" (http://www.milbo.users.sonic.net/earth/) was used for this research.

MARS has been recognized as an alternative to linear and additive-only model. Results from MARS show it provides a very promising solution to a complex problem (COK et al. 2004). York et al. conducted a comparison among linear and non-linear polynomial and non-linear MARS from simulations and concluded that the power obtained from MARS was an
average of 1.4 times higher than that of polynomials. Their results also showed that the power was higher even if the regression curve was linear, that gains increased with the complexity of the curve, and that for high non-linear curves, model-free methods such as MARS might be the only alternative (York et al. 2006). By comparing MARS with backpropagation neural network (BPNN) for prediction of pile drivability, Zhang and Goh concluded that BPNN and MARS models for the analyses of pile drivability provide similar predictions and can thus be used for predicting pile drivability. MARS can be considered to be more computationally efficient than BPNN, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding, and pruning (Zhang and Goh 2016). Comparison analysis between linear-regressions based multiple linear regression (MLR) and MARS for predicting liquefaction-induced lateral spread show that MARS outperforms MLR in terms of predictive accuracy. MARS automatically models non-linearity and interactions between variables without making any specific assumptions. Furthermore, it is able to provide the relative importance of the input variables and give insights into where significant changes in the data may occur (Goh and Zhang 2014). MARS’s high prediction accuracy was observed by shuffling MARS, and adaptive neuro-fuzzy inference system as tools for quantitative structure-activity relationship (QSAR) models study of severe acute respiratory syndrome (SARS) inhibitors (Jalali-Heravi et al. 2009). The accuracy and robustness of the shuffling MARS model in predicting inhibition behavior were illustrated using leave-one-out and leave-multiple-out cross-validation techniques. Jalali-Heravi concluded that the shuffling MARS model is superior and can be considered as a tool for predicting the inhibitory behavior of SARS drug-like molecules. Results from a most recent comparison among least square support vector machine (LSSVM), MARS, and M5 model tree (M5Tree) in modeling river water pollution indicated that the LSSVM and MARS models
had almost same accuracy and they performed better than the M5Tree model in modeling monthly chemical oxygen demand. The average root-mean-square error (RMSE) of the LSSVM and M5Tree models was decreased by 1.47% and 19.1% using the MARS model, respectively. The overall results indicated that the MARS and LSSVM models could be successfully used in estimating monthly river water pollution (Kisi and Parmar 2016).

MARS was also specifically used to identify the SNP-SNP interaction network in angiogenesis genes associated with prostate cancer aggressiveness (Lin et al. 2013). The model, including epistasis from SNP-SNP interaction, increase prediction accuracy.

Logistic regression is a form of the probabilistic statistical classification model, which can be used to predict class labels based on one or more features. The classification is done by using the logit function to estimate the outcome probability. One can use logistic regression by specifying the family as a binomial while using the general linear model (glm) function (Chiu 2015). Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression and can be divided into two categories based on the number of outcomes: simple logistic regression if the output has only two options, such as yes or no, 0 or 1, resistant or susceptible; in contrast, Multinomial logistic regression has more than two possible discrete outcomes, such as IDC score 1, 2, 3, 4, 5 or resistance as 1, moderate resistance as 2, moderate susceptible as 3, and susceptible as 4 and 5.

The logistic regression algorithm has been used to ML prediction, and results show that Logistic regression had superior performance compared with other ML algorithms. From a comparison of regression trees, logistic regression, generalized additive models (GAMs), and MARS for predicting acute myocardial infarction mortality, Austin et al. reported that the accuracy measured by mean ROC curve area for the regression tree models was 0.762, while the
mean ROC curve area of a simple logistic regression model was 0.845. The mean ROC curve areas for the other methods ranged from a low of 0.831 to a high of 0.851. The study shows that logistic regression performed better than regression trees do for predicting mortality. But, the logistic regression model had performance comparable to that of GAMs and MARS (Austin 2007). From the same research team, by comparing the predictive accuracy of regression trees with that of logistic regression models for predicting in-hospital mortality in patients hospitalized with heart failure, Austin et al. observed that the accuracy measured by the area under the receiver operating characteristic curve (ROC) for logistic regression models ranged from 0.747 to 0.775, whereas the corresponding values for regression trees ranged from 0.620 to 0.651, and authors concluded that Logistic regression predicted more accurately than did the regression trees in general.

Since both simple and multinomial logistic regression are used more years than the new ML algorithms, such as SVM, ANN, these new ML algorithms usually outperform logistic regression, and several reports have shown this trend (Hajmeer and Basheer 2003; Eftekhar et al. 2005; Shadabi et al. 2009; Chen et al. 2012).

From the above literature search and review, many ML algorithms have been proposed and applied in different research areas. Each ML algorithm has its advantages and disadvantages. Some of them apply a totally different mechanism for classification and regression. For example, decision tree-based prediction algorithms (RF and boosting) employed totally different mathematics and statistics from SVM. For IDC prediction, IDC score will be treated as a numeric number for regression analysis, and IDC score will be converted into three categories resistance “R” as 1, moderate resistance “MR” as 2, moderate susceptible “MS” for 3, and
susceptible “S” as 4 and 5 for classification prediction. All the ML algorithms will be employed via ether R or Python packages, or both for comparison.
CHAPTER 3. GEOSPATIAL STATISTICS FOR SOYBEAN IRON DEFICIENCY CHLOROSIS

Zhanyou Xu¹, Steven B. Cannon², William D Beavis¹,³*

¹ Bioinformatics and Computational Biology, Iowa State University, Ames Iowa 50014
² USDA Agricultural Research Service, Corn Insects, and Crop Genetics Research Unit, Ames, Iowa 50010
³ Department of Agronomy, Iowa State University Ames, Iowa 50010

Modified from a manuscript to be submitted to Crop Science

Abstract

Accounting for field variation patterns plays a crucial role in interpreting phenotype data and, thus, in plant breeding. Several spatial models have been developed to account for spatial variation, showing that spatial models can successfully increase the quality of phenotype measurement and subsequent selection accuracy for continuous data types such as grain yield and plant height. For stress traits in which the phenotypic data is recorded in ordinal data scores, such as in iron deficiency chlorosis (IDC), the spatial adjustment has not been well studied. The objective of the research described here is to evaluate methods for spatial adjustment of ordinal data, using soybean IDC as an example. Comparisons of adjustment effectiveness for spatial autocorrelation were conducted among three different groups of models. The models can be divided into three groups: group I, moving average grid adjustment; group II, geospatial autoregressive regression (SAR) models; and group III, tensor product penalized P-splines. Results from the model comparison show that the effectiveness of the models depends on the severity of field variation, the irregularity of the variation pattern, and the model used. The geospatial SAR models outperform the other models for ordinal IDC data. Prediction accuracy for the lines planted in the IDC high-pressure area is 11.9% higher than that of the lines planted...
in low IDC pressure regions. The relative efficiency of the mixed SAR model is 175%, relative to the baseline ordinary least square model.

Abbreviations: iron deficiency chlorosis (IDC), geospatial autoregressive regression (SAR), relative efficiency (RE), ordinary least square with range and row (OLS w/ RR), first-order autoregressive (AR1)

**Introduction**

Iron deficiency chlorosis (IDC) in soybeans is caused by the inability of the plant to utilize iron in the soil. Without enough iron, chlorophyll production is hampered, and the plant will suffer and possibly die. IDC is expressed in new leaf tissue, and symptoms typically appear on younger leaves, between the first and third trifoliate growth stages, vegetative stages V1 to V3 (Lin *et al.* 1998). The typical symptoms of IDC are the yellowing of leaves with interveinal chlorosis, while the veins remain green (Goos and Johnson 2000).

Soybean IDC affects yield. Soybeans are the second-most-planted field crop in the United States after corn, with a record-high of 90.16 million acres planted in 2016 (National Agriculture Statistics Service; NASS). IDC has a 20% yield reduction for each unit increase in chlorosis, based on chlorosis scoring on a 1 (no yellowing symptoms) to 5 (severe yellowing of leaves and the plant dies) scale (Froehlich *et al.* 1980). Soybean planting acreage in IDC-prone regions has increased from 1979 to 2017, with a 160% increase in soybean production area into IDC-prone regions with soil pH of 7.2 or greater in the past 30 years (Hansen *et al.* 2003). This increase of soybean production area into IDC-prone regions has led to yield losses of 340 million tons, worth an estimated 120 million dollars per year (Hansen *et al.* 2004). Current IDC trends in soybean production areas are expected to continue. Thus, minimizing or eliminating yield lost due to IDC is critical.
Because soil micro-environmental variation causes testing location heterogeneity, it is difficult to find large and uniform fields of calcareous soil that can be used for evaluation of IDC, which results in more experimental error than is desirable for selection among genotypes (Niebur and Fehr 1981). Environmental conditions for soybean to develop IDC symptoms are ephemeral, usually existing for a couple of weeks during the V1 to V3 stages of soybean development. Fields chosen for IDC testing are selected based on both historical IDC pressure records and potential for IDC conditions detected at the time of planting. In fields known to exhibit IDC, the exact locations within the fields may change from year to year, depending on rainfall prior to planting, and the rate at which soil moisture evaporates in the early growing season. Within a testing site, IDC pressure usually varies either by ranges and rows from year to year, leading to different levels and patterns of IDC expression with spatial autocorrelations. To find the exact locations of IDC pressure within a field each year, breeders planted the IDC-susceptible varieties in plots throughout the field, augmenting the varieties planted in each incomplete block. Seed companies usually test thousands of lines each year at each location, with at most two replicates; thus, some lines may be planted in high and some in low IDC-pressure areas. The ephemeral nature of IDC spatial and temporal variation for high and low IDC pressure cannot be planned for, given the small plot evaluations of early-stage field trials, and thus require spatial models to correct for variation in IDC pressure within fields.

IDC phenotype scores typically range from 1 (the most resistant) to 5 (the most susceptible) in reports by academic field breeders (Cianzio et al. 1979), or from 1 to 9 by commercial plant breeding organizations. In both cases, the scores represent ordinal data (Gaspar 2010). While plant breeders attempt to create ordinal data that captures the gradual nature between little to extreme stress, IDC scores nevertheless often change sharply from 1 to 9
in actual fields, without continuous transitions. IDC symptoms in Iowa and Southern Minnesota often appear to consist of oval-shaped patches (Figure 1) due to the location of soil moisture. In summary, ordinal data such as IDC creates phenotyping and selection challenges for plant breeders due to the genetic complexity of the trait, the scoring of IDC as ordinal data, and the use of small hill plots in early-stage evaluations, in fields with ephemeral patches of IDC.

Figure 1. Spatial circle variation patterns in the soybean IDC field.

Accurate phenotypic data is the most important factor for both visual-based phenotype selection and marker-assisted selection. High-quality phenotypic data relies on both experimental design and an accurate assessment of the phenotypes. Since IDC phenotypes are, to some degree, opportunistic, standard IDC resistance and susceptible checks or controls are included in field trials to estimate both the overall IDC pressure across the testing site and used as a reference to measure and adjust IDC scores of new lines. To balance the dilemma of the arrangement of check plots and to maximize the number of test-line entries with spatial field variation, many experimental designs, such as augmented design (SPEHAR 1994), modified augmented designed (SCHAALJE et al. 1987; MAY et al. 1989), partial replicated design (CLARKE AND STEFANOVA 2011) and augmented partially replicated (P-rep) design (WILLIAMS et al. 2011; MOEHRING et al. 2014; WILLIAMS et al. 2014), incomplete block alpha lattice design (YAU 1997; PIEPHO et al. 2006), have been developed to account for field trend and variation. At the same time, different
statistical models have been developed to account for field variation and phenotype data quality control. Phenotype data quality control is the process of removing non-genetic variation caused by environmental noise from the estimated genotype values. Phenotypic data variation and subsequent patterns in the field, or spatial variation, have been studied for many decades, especially in the geostatistics and econometrics fields (Anselin 2003; Anselin et al. 2004; Anselin 2006; Mobley et al. 2006; Anselin et al. 2010). Spatial models are required because of the autocorrelation among neighbor plots, which violates the identical independent distribution (iid) assumption for ordinary least square (OLS) regression, which has been used for the general linear regressions (Ugrinowitsch et al. 2004). Various spatial adjustment techniques have been developed to account for the spatial autocorrelation and have been shown to significantly improve the precision and repeatability of quantitative phenotypes. Collectively, these models can be clustered into three groups of spatial models based on the time of model development and the optimization mechanism used to adjust the spatial variation.

The first group includes the moving grid mean adjustment models (Townley-Smith and Hurd 1973; Mak et al. 1978; Rosielle 1980; Diers et al. 1991; Clarke et al. 1994). The moving mean spatial analysis has become more popular recently because the R package “mvngGrAd” implements the analysis. The package gives the user flexibility to pre-define any grid or pattern consisting of neighboring plots. The mean of the plots included in the grid is calculated and used as a covariate to account for the spatial variation (Technow 2015b). In contrast with the spatial autoregressive (SAR) model, which treats spatial variation anisotropic along with different directions (Dormann and Wokrina 1997), the moving mean average models treat the spatial variation as isotropic, in which the covariates among neighbor plots are simple means of the neighbor plots within a user-defined grid. This approach has been reported
to adjust the spatial variation successfully and thus has increased genomic selection accuracy from 0.231 to 0.37 for grain yield and from 0.436 to 0.614 for days to heading in wheat breeding, respectively (LADO et al. 2013a).

The second group (and also the most extensively studied one) is the spatial autoregressive regression (SAR) models, which include autocorrelation among neighbor plots in the model as a covariate parameter to model the correlation variation among the neighbors. These models have the property that the closer the plots, the higher correlation between the neighbors (TOBLER 1979). These models include spatial lag models, spatial error models, spatial lag plus error mixed models (ANSELIN 2001; DORMANN 2007), first-order autoregressive regression AR(1) or one-dimensional spatial analysis with row, column, or row + column (GLEESON AND CULLIS 1987), and the extended two-dimensional spatial analysis including the interaction between row and columns (CULLIS AND GLEESON 1991).

This group of models focuses on optimizing the variance-covariance structure of the residuals among neighbor plots. Evidence from a systematic comparison of covariance structures among experimental, spherical, Gaussian, linear, linear log, anisotropic power, and anisotropic exponential, show that AR(1) was generally not an optimal option for spatial analysis, and different covariance structures are needed to account for spatial variation in different trial sites (HU AND SPIlke 2009) because each trial site has different variation patterns. Within this group of models, field variations were also divided into a local trend or small-scale variation within a block or experiment, and global trend or large scale variation across all the experiments or the entire trial sites (WILKINSON et al. 1983). The nearest neighbor analysis model was developed to correct both local and global neighbors for the field variation (AINSLEY et al. 1995; PIEPHO et al. 2008). More complex models with polynomials were also developed to account for additive
effects for either row or column, and non-additive epistasis interaction between the rows and columns (Federer 1998). To remove the local and global field variation effectively, Gilmour et al. proposed a sequential spatial model schema which the first step is to remove local trend variation by fitting a two-dimensional range by row AR(1); the second step is to remove the global trend variation by fitting one-dimensional polynomials or splines in the direction of rows or columns (Gilmour et al. 1997; Kempton et al. 1999). To select the best model for each trial site, the sequential spatial model will run both model selection and model variable selection manually by applying graphical diagnostic tools to the spatial models. This manual model selection process was further extended with more optional models for the comparison and enhanced for model selection efficiency (Stefanova et al. 2009). Overall, SAR models have been improved as to the effectiveness of the spatial variation correction, and SAR analysis is becoming a routine for field data analysis for geostatistics and econometrics (Dhrymes 2017).

The third group of models to account for spatial variation is “Tensor product penalized spline models,” short for P-splines. P-splines have been used to account for both local and global spatial variation in tree genetics via Bayesian mixed model, with the accuracy of breeding values increasing by 66% (Cappa and Cantet 2007). This Bayesian mixed model was extended to the 2-dimensional smoothed surface in an individual-tree based model using a tensor product of linear, quadratic, and cubic splines for rows and columns, and the accuracy of breeding values for the offspring was increased by 46.03% (Cappa et al. 2011). Most recently, an advanced p-spline-based model was proposed and developed as the R package Spatial Analysis of field Trials with Splines, “SpATS” (Rodríguez-Álvarez et al. 2016). This model includes both bilinear polynomial and smooth splines components: 1) The bilinear polynomial component consists of three sub-variables: row spatial trend, spatial column trend, and interaction between
row and columns; and 2) The smooth spline part contains five smooth additive spatial components. This approach uses two-dimensional P-spline ANOVA representation of the anisotropic smooth surface formulated in a mixed model via SpATS. SpATS is a better choice than SAR in two aspects, according to the author: 1) SpATS is a one-step modeling approach to spatial analysis by fitting a general SpATS model to analyze all the trials from different trials sets. SpATS approach overcomes the sequential spatial model selection for different testing sites and can be used to large-scale, high-throughput and routine multiple environmental trial (MET) analysis; 2) SpATS can fit both local and global variation, isotropic and anisotropic variation, one-dimensional and two-dimensional variation into one model, and optimize the best parameters to remove the noise from phenotypic data from the true genotypic values without overfitting. With one model fitting all the trials, it minimizes the chance of using different models selected for different trials – which might lead to biases against different genotypes from different locations because of the different model variables selected (VELAZCO et al. 2017a). The SpATS was tested by modeling spatial trends in sorghum breeding fields, and the results show that the improvement in precision and predicted genotypic values from SpATS analysis were equivalent to those obtained using the best SAR sequential model selection for each trial (VELAZCO et al. 2017a).

The above three groups of spatial models to adjust field variation have been mainly applied in econometrics and geostatistics and a few for plant breeding for the continuous data type, such as crop grain yield and plant height (DIERS et al. 1991; LADO et al. 2013b; VELAZCO et al. 2017a).

In contrast to continuous yield data, the soybean IDC score is a discrete ordinal variable, whereas the moving grid, SAR, and P-spline models were developed and applied to continuous
variables. In the applications, adjusted continuous trait phenotypes have improved precision and the repeatability of the field trials. To our knowledge, the application of these methods to adjust for non-genetic spatial patterns in ordinal traits such as IDC is limited to a single publication in which the moving grid was applied to IDC in soybean (Diers et al. 1991). Also, while most of the published reports about the spatial analysis models have applied a few models, most reports have not systematically compared the effectiveness of models for data obtained from fields with different levels of severity and irregular, discontinuous spatial patterns. The effectiveness of adjustments made by spatial models depends not only on the model parameters but also on the severity of the spatial variation and the irregularity of the variation pattern in the experiment field. The objective of the research reported herein is 1) to apply six different geospatial as well as two OLS models as the baseline to three datasets with different levels of severity of spatial variation, variation patterns, different experimental design; 2) and to systematically compare models for spatial adjustment using R², AIC, residual standard error, Moran’s I, and prediction accuracy.

**Data and methods**

**Data**

Soybean IDC data were collected from experiments planted in historically “IDC-hot” sites, which were selected to minimize the IDC spatial variation by past years’ IDC records and the current year’s IDC status. Soybean IDC pressures vary from year to year for the same site, and testing sites are not always good enough for testing IDC across years. Before planting the testing lines in the historically “IDC-hot” testing sites in the current year, several IDC susceptible varieties were planted about ten days earlier than the typical planting date, in order to assess IDC pressure. These standard IDC controls were then rogued, and new testing lines were
planted in these testing sites if the susceptible varieties showed IDC symptoms. Otherwise, these sites were not used as IDC testing sites.

A total of five data sets, four experimental, and one simulated data set, with a total of 11,602 testing lines, were used for this study (Table 1). Dataset 1 was simulated to mimic the common circular IDC spatial variation pattern. The parameters used for the simulation and their variance proportions are summarized in Table 2. Dataset 1 contains 1,050 simulated genotypes with a total of 2,100 IDC scores. The field layout is 42 ranges by 50 rows, a randomized complete block design with two replicates.

In contrast to simulated dataset 1, datasets from #2 to #5 are experimental field data from 2016. They were selected to represent the four different IDC-prone regions. For each set, the experimental design was a six-by-six alpha-lattice, with 32 testing lines and four checks. These testing materials are in early development stages, planted with two replicates in one location, and the average number of replicates is 1.84 reps/line across all the experiments (Highlighted in Table 1). The reason for less than two replicates per testing line is because of the low emergence rate.

Table 1. Summary of the five data sets used in the study.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Exp. design</th>
<th>No. rows</th>
<th>No. range</th>
<th>No. entries</th>
<th>Ave No replicates</th>
<th>No. plots</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 1</td>
<td>RCBD</td>
<td>50</td>
<td>42</td>
<td>1,050</td>
<td>2.00</td>
<td>2,100</td>
<td>simulated</td>
</tr>
<tr>
<td>Data set 2</td>
<td>α-lattice</td>
<td>220</td>
<td>26</td>
<td>2,774</td>
<td>1.83</td>
<td>5,074</td>
<td>Iowa</td>
</tr>
<tr>
<td>Data set 3</td>
<td>α-lattice</td>
<td>24</td>
<td>220</td>
<td>2,652</td>
<td>1.79</td>
<td>4,740</td>
<td>Minnesota</td>
</tr>
<tr>
<td>Data set 4</td>
<td>α-lattice</td>
<td>110</td>
<td>56</td>
<td>2,719</td>
<td>1.88</td>
<td>5,124</td>
<td>Red River Valley</td>
</tr>
<tr>
<td>Data set 5</td>
<td>α-lattice</td>
<td>100</td>
<td>60</td>
<td>2,407</td>
<td>1.81</td>
<td>4,363</td>
<td>Nebraska</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>504</td>
<td>404</td>
<td>11,602</td>
<td><strong>1.84</strong></td>
<td>21,401</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Summary of parameters used for data simulation. The first column is the list of variance components. Columns 2, 3, and 4 are the variance values, the distribution used for the simulation, and the percentage of the variance components, respectively.

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>Variance</th>
<th>Distribution</th>
<th>Percentage of SD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location SD</td>
<td>1.5</td>
<td>normal</td>
<td>14.42</td>
</tr>
<tr>
<td>Experiment SD</td>
<td>0.5</td>
<td>Poisson</td>
<td>4.81</td>
</tr>
<tr>
<td>Line SD</td>
<td>2.1</td>
<td>Normal</td>
<td>20.19</td>
</tr>
<tr>
<td>Range SD</td>
<td>0.2</td>
<td>Normal</td>
<td>1.92</td>
</tr>
<tr>
<td>Row SD</td>
<td>1.0</td>
<td>Normal</td>
<td>9.62</td>
</tr>
<tr>
<td>Rep SD</td>
<td>0.1</td>
<td>Normal</td>
<td>0.96</td>
</tr>
<tr>
<td>Pattern SD</td>
<td>3.8</td>
<td>Normal</td>
<td>36.54</td>
</tr>
<tr>
<td>Residual</td>
<td>1.2</td>
<td>Normal</td>
<td>11.54</td>
</tr>
</tbody>
</table>

IDC score scales from 1 for most tolerant to 9 for the most susceptible are used for this research. The IDC score criteria are 1 = green leaves (no chlorosis) to 9 = dead plants, following the rating scale presented in the figure below (Figure 2). The IDC scores are treated as an ordinal data type.

![IDC score reference from 1 to 9 as an ordinal data type. 1 = green leaves (no chlorosis) to 9 = dead plants. Notice that only six images were displayed here. Each image was taken from one hill plot with 3 to 8 plants at growth stage V2 to V4.](image-url)
Among the five data sets, three classes of distinctive spatial variation patterns, the Red River Valley (RRV) pattern, the IA and Minnesota (IA/MN) pattern, and the Kansas and Nebraska (KS/NE) pattern, were observed from the heatmap of the IDC raw data (Figure 3). The three IDC field spatial patterns are consistent with the three soybean IDC-prone soil types, which were clustered by principal component analysis (PCA) based 15 soil character measurements (bottom images of Figure 3). Five testing sites from the North Dakota and Manitoba regions were clustered as the RRV group (five green dots in the bottom right-hand figure), four testing sites from Iowa and Minnesota regions were clustered together as the IA/MN group (four blue dots in the bottom right-hand figure), and one testing site from Nebraska was clustered as the NE group (Figure 3). KS/NE regions have relatively uniform IDC scores without a noticeable spatial pattern, and no spatial model is needed for the analysis.

The Red River Valley (RRV) IDC data has a column by column spatial variation with block effects, which can be corrected by block design and corresponding statistical analysis. In contrast, the IA/MN IDC testing sites show totally different spatial patterns from that of RRV and KS/NE regions. Thus, IDC data from the IA/MN region has spatial autocorrelation and needs spatial autoregressive analysis. Among the five testing sites from the IA/MN region, two of them showed different spatial variation patterns (Figure 3 from data set 2 and 3) from the other three locations and were selected for this research. All the results hereafter are based on the three data sets: two data sets from the IA/MN region and one simulated data set (Table 1 and Figure 3).
Figure 3. Three IDC field variation patterns from the heatmap (top three images) corresponding to the three IDC zones, which were characterized by soil types. The bottom left-hand image is an IDC testing field with IDC spatial variability. The bottom right-hand figure shows the three types of testing locations by soil properties.

**Analytic methods**

Including ordinary linear square (OLS), a total of eight models were compared (Table 3).

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Model No</th>
<th>Spatial Term</th>
<th>R package</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>M1</td>
<td>None</td>
<td>RMS</td>
<td>(F RANK E. HARRELL 2015)</td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>M2</td>
<td>None</td>
<td>RMS</td>
<td>(F RANK E. HARRELL 2015)</td>
</tr>
<tr>
<td>MovingGrid</td>
<td>M3</td>
<td>Mean of grid</td>
<td>mvngGrAd</td>
<td>(TECHNOW 2015a)</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>M4</td>
<td>Lag</td>
<td>spdep</td>
<td>(BIVAND 2013)</td>
</tr>
<tr>
<td>SAR + error</td>
<td>M5</td>
<td>Error</td>
<td>spdep</td>
<td>(BIVAND 2013)</td>
</tr>
<tr>
<td>SAR Durbin</td>
<td>M6</td>
<td>Lag+ Error</td>
<td>spatialreg</td>
<td>(BIVAND AND PIRAS 2015)</td>
</tr>
<tr>
<td>ASReml AR1</td>
<td>M7</td>
<td>AR1(range): AR1(row)</td>
<td>ASReml-R</td>
<td>(BUTLER et al. 2017)</td>
</tr>
<tr>
<td>B/P-Spline</td>
<td>M8</td>
<td>psanova(range, row)</td>
<td>SpATS</td>
<td>(RODRIGUEZ-ÁLVAREZ 2018)</td>
</tr>
</tbody>
</table>
Model 1: Ordinary least square (OLS) without the range and row covariates

OLS is a linear least-square method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function representing a set of explanatory variables by the principle of least squares. Traditionally the regression model does not include parameters for the spatial dependence of the experimental units. The general equation for the OLS with p variables can be written as:

\[ Y = \alpha_{L_n} + X\beta + \varepsilon \quad (DHRYMES \ 2017) \]

Where Y and \( \varepsilon \) are (n x 1) vector of the values of the response variable and the errors for the various observations, respectively. \( L_n \) is a (n x 1) vector of ones associated with the constant term parameter \( \alpha \) to be estimated. \( X \) is an (n x p) matrix of regressors or the design matrix. \( \beta \) is (1 x p) vector of the parameters to be estimated.

OLS analysis for model 1 was conducted with the ols() function implemented the R package “regression modeling strategies (RMS)” (FRANK E. HARRELL 2015).

Model 2: Ordinary least square (OLS) with range and row

Model 2 analysis was similar to model 1 except two variables, range, and the row was added in the model

\[ Y = \alpha_{L_n} + X\beta + \text{ROW} + \text{RANGE} + \varepsilon \quad (DHRYMES \ 2017) \]

Model 3: Moving grid adjustment

The “moving grid average adjustment” is a spatial method to adjust for environmental variation in field trials. It is most common in field trials with few replicates, for early-stage breeding materials. All the raw data were aligned into a row by range rectangle layout. A grid is predefined based on the field variation pattern around each cell (= entry), and each observed value was adjusted by the values from the neighbor plots within the predefined grid. The mean of the cells included in the grid is calculated as the equation below (TECHNOW 2015c):
\[ x_i = \frac{\sum_j p_{j,obs} \cdot I(p_{j,obs} \in G_i)}{\sum_j I(p_{j,obs} \in G_i)}, \quad P_{t,obs} = P_{t,obs} - b(x_i - \bar{x}) \]

Where

- \( x_i \) is the moving mean of the ith entry
- \( G_i \) is the grid of entry i and \( I(\cdot) \) is an indicator function that takes the value “1” if the condition is satisfied and “0” if not?
- \( P_{j,obs} \) are the observed phenotypic values of all entries which are included in \( G_i \)
- \( \bar{x} \) is the mean of all \( x_i \)
- \( b \) is the regression coefficient in the linear model.
- \( P_{t,obs} \) are the adjusted phenotypic values of all entries.

The layout of the grid (Supplemental Figure 1) used to adjust the phenotype value.

The model executed for each data set as:

\[ \text{movingGrid (rows = no.of.rows, columns = no.of.range, obsPhe=raw.observed.values, shapeCross=list(1:2,1:2,1:2,1:2), layers=c(1:1), excludeCenter = TRUE)} \]

where “shapeCross” is to set up the shape of the moving grid, “excludeCenter” is to define whether the center from each grid is included/excluded to calculate the mean.

**Model 4: Spatial autoregressive lag model**

When a value in one plot depends on the values of its neighbors, the errors are no longer uncorrelated and may not have the independent and identically distributed (iid). Depending on the nature of the spatial dependence, OLS will be either inefficient, with incorrect standard
errors, or will be biased and inconsistent (Bivand 2006). When IDC scores depend not only on
genetics but also on its scores from neighboring plants, the spatial lag model was introduced to
correct for autocorrelation (Baltagi and Liu 2008). In the spatial lag model, the spatial
components were specified on the dependent variable, IDC scores. This setting leads to spatial
filtering of the variable, which is averaged over the surrounding neighborhood defined in W,
called the spatially lagged variable. The spatial lag model can be specified as:

\[
Y = \rho W Y + X \beta + \epsilon
\]

Where \( \rho \) is the autoregressive lag coefficient, which tells us how strong the resemblance
is, on average, between \( Y_i \) and its neighbors; if \( \rho \) is not significantly different from 0, then the
spatial lag model becomes traditional OLS regression model.

\( y_i \) is the \( i^{th} \) IDC score, and \( y_j \) is all the neighbor’s IDC scores around \( i^{th} \) IDC score.
\( y_i \) stands for one of \( n \) observed IDC scores, \( y_j \) stands for more than one IDC scores.

\( W_Y \) is a spatial weight matrix with \( n \times p \) rows and ranges, describing the spatial
correlation structure of the observations.

\( X \) is an \( n \times p \) matrix of regressors or the design matrix; \( \beta \) is the \( p \times 1 \) vector of estimated
coefficients.

Analysis of IDC scores with the spatial lag model was conducted via R package
“spdep” (Bivand 2013) and “spatilreg” (Bivand and Piras 2015).

**Model 5: Spatial autoregressive error model**

In contrast to the spatial lag model treating autocorrelation as a lag component in the
response variable, the spatial error model regarding the autocorrelation as part of the error term.
The spatial error model incorporates a local and a spillover element in the variance-covariance
matrix of the error term in a linear regression model (Anselin and Moreno 2003). Formally, the model can be written as 

\[ Y = X \beta + e \quad \text{and} \quad e = \lambda W_e + V \]

Where \( \lambda \) is the spatial error coefficient, if \( \lambda \) is not significant from 0, the spatial error model becomes an OLS regression model. \( W_e \) are the weight matrix to adjust the error correlation in the residuals.

Analysis of IDC scores with the spatial error model was conducted via R package “spdep” (Bivand 2013) and “spatilreg” (Bivand and Piras 2015).

**Model 6: Spatial Durbin mixed model**

A limitation of the spatial lag or spatial error models is that they can include either an autoregressive lag or a spatial error covariate in the model. In reality, some fields have complexed spatial variations with both autocorrelation lag and spatial error. Also, the dependencies in the spatial autoregressive relationships don’t only occur in the dependent variable but may also be present in the independent variables. The spatial Durbin mixed model was developed to account for a dependent, the autocorrelation lag, spatial error, and independent variables (Dormann et al. 2007b; Bektı 2012). The model can be written as

\[ Y = \rho W Y + X \beta + WX \theta + \epsilon \]

Where 

\( Y, X, \rho, \) and \( \beta \) are defined as above 

\( WY \): is the spatially lagged offering IDC scores accounting for various spatial dependencies with \( W \) defined as (n x n) spatial weight matrix 

\( \rho W Y \): Endogenous interaction effect 

\( \theta \): (k x 1) vector of unknown parameters
\( \theta WX \): Exogenous interaction effect

Implementation of the spatial Durbin model was carried out via the R package “spatilreg” (Bivand and Piras 2015).

**Model 7: AR1 by AR1 via ASReml-R**

The ASReml mixed model is widely used in plant and animal breeding and quantitative genetics. It also provides functions for spatial autoregressive analysis. ASReml-R is the only commercial R package in this study and used to compare whether it will outperform the free spatial analysis packages. One update in the new ASReml-R version 4 for spatial analysis, ASReml changed the random formula and error (rcov) component “rcov = ar1(range):ar1(row)” to residual = ar1(range):ar1(row) (Butler et al. 2009; Butler et al. 2017).

The model used for the analysis is: asreml(fixed = IDC_scores ~1, random = ~ LINCD, residual = ~ar1(range):ar1(row), data = IDC.data)

**Model 8: P-spline mixed model via SpATS**

The P-spline approach models field trends using a smooth bivariate function of the range and row \( f(\text{range}, \text{row}) \), represented by a 2D P-splines (Rodríguez-Álvarez et al. 2016; Velazco et al. 2017b). The P-spline technique optimizes the fitted surface by penalizing the spatial effects. The degree of penalization over the fitted spatial variation trend is determined by smoothing parameters. The 2D range by row surface is decomposed into a sum of linear components and univariate and bivariate smooth function as:

\[
\begin{align*}
    f(\text{range}, \text{row}) &= X_s \beta_s + Z_s s \\
    \text{where:} \\
    X_s \beta_s &= \beta_{s1} \text{row} + \beta_{s2} \text{range} + \beta_{s3} \text{row. range} \\
    Z_s s &= f_1(\text{row}) + f_2(\text{range}) + h_3(\text{row}).\text{range} + h_4(\text{range}).\text{row} + f_5(\text{row},\text{range}) \\
    \beta_{s1} \text{row}: \text{linear trend by row}
\end{align*}
\]
\( \beta_{s2\text{range}} \): linear trend by range or column

\( \beta_{s3\text{row.range}} \): linear interaction trend by row \( \times \) range

\( f_1(\text{row}) \): main smooth trend across rows

\( f_2(\text{range}) \): main smooth trends across ranges

\( h_3(\text{row.range}) \): interaction trends between linear range by smooth surface row

\( h_4(\text{range.row}) \): interaction trends between linear row by a smooth surface range

\( f_5(\text{row.range}) \): smooth-by smooth trends between ranges and rows

The SpATS mixed model can be written as:

\[
Y = X\beta + X_s\beta_s + Z_s s + Z_u u + Z_g g + e \quad (\text{VELAZCO et al. 2017b})
\]

Where:

\( X, \beta, X_s\beta_s, Z_s s \) are the same as above

\( Z_u u \): u is the sub-vector of a random row and range effects accounting for discontinuous field variation; \( Z_u = [Z_{\text{row}} | Z_{\text{range}}] \) is the design matrix

\( Z_g g \): g is the vector of random effects of genotypic effects of the testing lines or hybrids; \( Z_g \) are the design matrix for the genotype effects.

The final SpATS model R scripts used in this study for P-spline is as below:

```
P-Spline.model =

SpATS(response="IDC_Scores", genotype="LINCD", genotype.as.random = TRUE,
         spatial = ~PSANOVA(row, range, nseg = c(10,10),
                          degree = 3, nest.div = 2), fixed = NULL,
         control = list(tolerance = 1e-03, monitoring =1),
         data = IDC.data)
```
The most different term in this model is the term “spatial” which is an auxiliary function used for modeling the spatial or environmental effect as a two-dimensional penalized tensor-product of marginal B-spline basis functions with anisotropic penalties based on the PSANOVA approach (Ebeling et al. 2006; Lee et al. 2013). Inside spatial, “nseg” stands for the number of segments in the P-splines, 10 and 10 segments were used for both range and row, respectively. Parameter “degree” stand for numerical the order of the polynomial of the B-spline basis for each marginal. Degree of 3, cubic B-splines were used for the IDC analysis. Parameter “nest.div” is a divisor of the number of segments (nseg) to be used for the construction of the nested B-spline basis for the smooth-by-smooth interaction component. In this case, the nested B-spline basis will be constructed, assuming a total of nseg/nest.div segments. The value was set to 2 for the IDC analysis.

**Performance metrics to compare the models**

The following statistical parameters were employed to compare model effectiveness. They are 1) $R^2$; 2) Akaike’s information criterion ($AIC$); 3) residual standard error ($RSE$); 4) Moran’s I index; 5) p-value of Moran’s I; 6) prediction accuracy I (whole data set); 7) and prediction accuracy II (cross-prediction accuracy). Not all of the models generate all of these parameters. For fairness of comparisons, all the parameters were calculated via the following equations manually:

1) $R^2$: $R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$ (James 2013)

where $SS_{residual}$ is the sum square of the residual from the model, and $SS_{total}$ is the total sum square from the data.

2) $AIC$: $AIC = 2k - 2ln(L) = -2(log-likelihood) + 2K$ (James 2013)

Where $K$ is the number of model parameters (the number of variables in the model plus the
intercept). Log-likelihood is a measure of model fit. The higher the number, the better the fit. For the AIC value, the smaller, the better of the model.

3) Residual standard error (RSE) = square.root (MSE)

where MSE is mean square error

4) Moran’s I index: The spatial autocorrelation index (Moran’s I) measures spatial autocorrelation based on both feature locations and feature values simultaneously (FORTIN AND DALE 2009). With a set of features and an associated attribute, Moran’s I index evaluates whether the pattern expressed is clustered, dispersed, or random. Moran’s I test provides a way to check whether there is spatial autocorrelation in the field data and whether residuals from a spatial model are not correlated or randomly distributed, with the iid property. Moran’s I test value is between -1 to +1. A value of “-1” indicates the large and small values intersperse across the field, and the data are negatively auto-correlated, while “+1” indicates high IDC scores surrounded by high IDC scores or low IDC scores surrounded by low IDC score; these scores are positively auto-correlated. If all the residuals from a model are iid, and there is no autocorrelation, then the Moran’s I should be close to zero or equal to zero. The formula to calculate the Moran’s I index as below:

\[
\text{Moran's I index} = \frac{N}{\sum_{i} \sum_{j} w_{ij}} \frac{\sum_{i} \sum_{j} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2} \quad \text{(BIVAND et al. 2009)}
\]

Moran’s I index was calculated via `moran.test` from the R package “spdep,” as below (order of the data is important using the function “moran.test.” The data need to be the same order as the one in the weight matrix of the rectangle):

```r
moran.test(IDC.data/residuals, listw = ds2_weightMatrix, alternative = "two.sided", na.action = na.omit )
```
Where $N$ is the number of spatial units indexed by $i$ and $j$; $X$ is the variable of interest; $\bar{X}$ is the mean of $X$; $w_{i,j}$ is the spatial weight matrix. If the “I” statistically closely to 0 or equal to 0 from the residuals, then the spatial adjustment is successful. If the I value is close to either -1 or +1, with p-values <0.05, then the spatial adjustment with the testing model does not work well to remove the spatial autocorrelation.

5) P-value of Moran’s I: The null hypothesis for the Moran’s I test in that there is no autocorrelation among the data in the area, and the data collected randomly distributed. If the p-value from Moran’s I test is not significant, or p-value > 0.05, the spatial distribution of feature values may be the result of random spatial processes.

6) Prediction accuracy I (whole data set): is calculated as

$$r_{(observed,fitted)} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (CLEWER \ 2001)$$

Where

- $n$ is the number of data points or sample size
- $x_i$ from 1 to $n$ is the observed values
- $y_i$ from 1 to $n$ is the fitted values from the model

7) Prediction accuracy II (cross prediction): is the correlation coefficient between the overserved and predicted for the testing lines planted in the low IDC pressure area. The formula for prediction accuracy II is the same as prediction accuracy I except using the testing lines planted in the low IDC pressure regions.

**Heatmap and Lagrange Multiplier Test**

Heatmaps of IDC field layout were made using the python package “seaborn” and the R package “fields” (NYCHKA 2016). The Lagrange Multiplier Test was carried out to select the
best model among all the spatial autoregressive (SAR) models via function “lm.LMtests “ from “spdep” package.

**Relative efficiency (RE)**

The mean squares error (MSE) from each analysis were used to estimate the relative efficiency (RE) reference to the MSE of the OLS:

\[
\text{Relative efficiency} = \frac{\text{MSE of the OLS}}{\text{MSE of each of the eight model}} \times 100
\]

(ABD EL-MOHSEN AND ABO-HEGAZY 2013)

**Results and discussion**

**Results from the P-spline model SpATS**

Results from data set 1 show a spatial 2D surface trend in the heatmap of “fitted spatial trend,” which indicates that the model correctly identified where the boundary of the pattern located (Figure 4, bottom left-hand heatmap). From the phenotypic best linear unbiased prediction BLUPs (Figure 4, bottom middle), a circle pattern existed in the heatmap, whereas the genotypes were randomly simulated based on a normal distribution. At the same time, a noticeable circle-pattern was left in the residual plot (Figure 4, top right-hand heatmap), showing that the model did not remove the spatial autocorrelation pattern completely. The Moran’s I index of the residual plots is 0.0748 with p-value 5.21e-07, which was much smaller than the threshold p-value 0.05. Comparing the Moran’s I index, 0.5152 from the raw IDC data, a p-value less than 2.2e-16, the spatial autocorrelation coefficient is dramatically reduced, from 0.5152 to 0.0748, for the SpATS model. But, in terms of goodness-of-fit, SpATS did fit the data well since there is still autocorrelation left in the residual plots. The results from the data set 1 analysis are that the P-spline model works to some degree, but it cannot completely remove the spatial autocorrelation in the plots.
Results from data set 2 indicate that more severe and different spatial patterns exit in field #2 than that from the simulated data set 1. Figure 5 displays the fitted spatial trend by the surface function of range and row f(row, range) from the P-spline fittings. Note that the heatmap of the fitted trend uses a finer continuous-like grid than that of the field plots, which was smoothed by 2D P-splines. The spatial surfaces displayed irregular patchy patterns across field 2, and discrepancy of the spatial trends between raw and fitted patterns existed. The heat map of the residual plots has a very similar spatial pattern as both fitted values, and raw IDC score data, which show that there is still spatial autocorrelation left among the residual plots.

Comparing the spatial trends from data set 1 and 2, the SpATS model works the best for data set 3 (Figure 6). The fitted spatial trend matched the raw field pattern very well, and heatmaps of both genotypic BLUPs and residuals look randomly distributed. No clear noticeable spatial pattern exists like that from data sets 1 and 2, even though Moran’s I index is still significant from 0 (p-value < 0.05).

Figure 4. Heatmaps and histogram from data set 1 by P-spline via SpATS. The rectangle layout is 42 rows (X-axis) by 50 ranges (Y-axis) with a total of 2,100 plots for each of the five heat maps. Six plots in this figure and 1) top left is the heatmap for the raw IDC data; 2) top middle is the heatmap of the fitted or predicted IDC data; 3) top right figure is the heatmap of the residual plots; 4) bottom left figure is the fitted spatial trend; 5) bottom middle is the phenotypic best linear unbiased prediction BLUP values or the genetic effects for each testing lines; 6) bottom right histogram is the distribution of the genotypic BLUPs. The x-axis for the last figure (bottom right) is the BLUP values from -2 to +2, and the y-axis is the number of lines.
Figure 5. Heatmaps and histogram from data set 2. The rectangle layout is 26 rows (X-axis) by 220 ranges (Y-axis) with a total of 5,720 plots. The white area on the right side of each heatmap is the 646 plots without data, from either with emergences but no IDC score data or without emergence at all. The arrangements of the six images are the same as that of Figure 6.

Figure 6. Heatmaps and histogram from data set 3. The rectangle layout is 24 rows (X-axis) by 220 ranges (Y-axis) with a total of 5,280 plots. The white area on the right side of each heatmap is the 540 plots without data, from either with emergences but no IDC score data or without emergence at all. The arrangements of the six images are the same as that of Figure 6.
Overall results from the three residual heatmaps obtained from the P-spline model are that there are spatial patterns or autocorrelations left in the residuals. These results are different from the two reports in that spatial patterns have effectively been removed by 2D P-spline surfaces, and residuals are true random noise in sorghum (Velazco et al. 2017a) and barley (Rodríguez-Álvarez et al. 2016). We believe that the reason may be because that yield is a continuous numerical variable, and the variation is more continuous in the field, which is different from the ordinal IDC score from 1 to 9 from small hill plots. IDC score could be 1 (completely resistant to IDC in one plot), and neighboring plots could be 9 (most susceptible), without gradual transitions; this may result in spatial pattern left in the residual plots.

**Spatial effective dimension (EDs) and the importance of surface trend by f(row):f(range)**

Spatial effective dimension is a measure of the complexity of the “smooth” model, in which the larger the EDs, the more complex the model (Rodríguez-Álvarez et al. 2016). The shapes of evident patchy spatial patterns of the IDC scores were best modeled by the integration of one-dimensional range or row trends (functional trend row or range and surface range or row) and two-dimensional trends (interaction function trend row by range “Row: Range,” surface trend range by linear function trend row “F(Range): Row,” linear functional trend range by surface trend row “Range: F(Row),” surface trend interaction range by row “F(Range): F(Row))” (Table 4). Overall one-dimension surface range trends are more complicated than that of the row, in which all the f(range) > f(row) in the three data sets as 3.0 > 2.3 in data set 1, 7.0 > 6.6 in data set 2, and 5.8 > 3.3 in data set 3, respectively (Table 4). From the two-dimension level, 2D surface trend rows by surface trend ranges “f(row):f(range)” are the most significant spatial terms in the model. The percentage of the effective spatial dimensions are 58.42, 44.12, and 53.92% for the data set 1, 2, and 3, respectively. The results from table 4 indicate that 2D surfaces generated by P-splines in the spatial model play a major role in correcting
autocorrelation in all the three data sets. Contributions of the 2D surface “F(Range): F(Row)” are 58.42, 44.12, 53.92% for the data sets 1, 2 and 3, respectively. Another observation is that the linear trend row plays a big role (13.9 in data set 2), and the linear trend range did not contribute at all (0 in the data set 2). Opposite results were obtained from data set 3: 24.6 for the linear trend range and 17.6 for the linear trend row, respectively. The differences in linear trend ranges and rows between data set 2 and 3 indicate the spatial structures of the data set 2 and 3 are different.

Overall results from the three data sets by the spatial effective dimension analysis are that both one-dimension 1D and 2D trend surfaces from P-splines are the major contributions for the fitted spatial trend.

Table 4. Summary of effective spatial dimension (EDs) of the smooth surface components fitted by the SpATS model and effect dimension from the OLD model (EDm) from the three data sets. The numbers in bold are the largest variable in each column.

<table>
<thead>
<tr>
<th>Variables Name</th>
<th>Data Set 1</th>
<th></th>
<th>Data Set 2</th>
<th></th>
<th>Data Set 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EDs</td>
<td>EDm</td>
<td>EDs</td>
<td>EDm</td>
<td>EDs</td>
<td>EDm</td>
</tr>
<tr>
<td>Range</td>
<td>10.1</td>
<td>42</td>
<td>0.0</td>
<td>220</td>
<td>24.6</td>
<td>221</td>
</tr>
<tr>
<td>Row</td>
<td>41.1</td>
<td>50</td>
<td>13.9</td>
<td>28</td>
<td>17.6</td>
<td>24</td>
</tr>
<tr>
<td>Row:Range</td>
<td>1.0</td>
<td>1</td>
<td>1.0</td>
<td>1</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>F(Range)</td>
<td>3.0</td>
<td>11</td>
<td>7.0</td>
<td>11</td>
<td>5.8</td>
<td>11</td>
</tr>
<tr>
<td>F(Row)</td>
<td>2.3</td>
<td>11</td>
<td>6.6</td>
<td>11</td>
<td>3.3</td>
<td>11</td>
</tr>
<tr>
<td>F(RANGE):ROW</td>
<td>0.0</td>
<td>11</td>
<td>6.5</td>
<td>11</td>
<td>5.8</td>
<td>11</td>
</tr>
<tr>
<td>RANGE:F(ROW)</td>
<td>1.8</td>
<td>11</td>
<td>1.6</td>
<td>11</td>
<td>10.0</td>
<td>11</td>
</tr>
<tr>
<td>F(RANGE):F(ROW)</td>
<td><strong>83.3</strong></td>
<td>121</td>
<td><strong>28.9</strong></td>
<td>36</td>
<td><strong>79.7</strong></td>
<td>121</td>
</tr>
<tr>
<td>% F(RANGE):F(ROW)</td>
<td><strong>58.42</strong></td>
<td>46.90</td>
<td><strong>44.12</strong></td>
<td>10.94</td>
<td><strong>53.92</strong></td>
<td>29.44</td>
</tr>
<tr>
<td>Total</td>
<td>142.6</td>
<td>258</td>
<td>65.5</td>
<td>329</td>
<td>147.8</td>
<td>411</td>
</tr>
</tbody>
</table>

Variance components analysis and importance of surface trend by f(row):f(range)

Results of variance components analysis of the three random (LINCD, Range, and Row), five surface variables (f(RANGE), f(ROW), f(RANGE): ROW, RANGE:f(ROW)), and one 2D surface range trend by surface row trend, f(RANGE):f(ROW)) were summarized in Table 5. These show that 1) three datasets have large differences in the level of spatial variances, with
dataset 1 having the smallest total variance 431.5 (Table 5) and dataset 2 having the largest variance 6597.2 (15-fold larger than that of dataset 1); 2) the tensor product of P-Splines term, surface range trend by surface row trend \( f(\text{RANGE}):f(\text{ROW}) \), accounts for the majority of the variance in dataset 1 and 2 – up to 85.6% and 89.6%, respectively; and 3) dataset 3 has different spatial variation patterns from datasets 1 and 2 – which can be inferred from the much smaller variance component from surface range trend by surface row trend 15.14%. In contrast, linear range trend by surface row trend, range:\( f(\text{row}) \), takes up to 78.1% in dataset 3. One observation is that variance components for LINCDs are very small: 0.1%, 0.01%, and 0.05% for datasets 1, 2 and 3, respectively. The overall mean-variance components for LINCDs is only 0.05% across the three data sets. Similar results were observed for the residual variance, in which the mean percentage of variance component is only 0.1% across three datasets (last column in Table 5).

The overall results from the variance components analysis show that spatial variation along the rows is much larger than that from ranges. The mean percentages related to surface row term “\( f(\text{ROW}) \)” are 5.95% for “\( f(\text{ROW}) \)”, 26.42% for “\( \text{RANGE}:f(\text{ROW}) \)”, and 63.44% for “\( f(\text{RANGE}):f(\text{ROW}) \)” with a total of 95.81% (last column in Table 5).

Table 5. Summary of variance component analysis from the tensor product panelized P-splines. Var stands for variance, type for data type: “R” for random and “S” for surface variable.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
<th>Data set 1</th>
<th>Data set 2</th>
<th>Data set 3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Var</td>
<td>% var</td>
<td>Var</td>
<td>% var</td>
<td>Var</td>
</tr>
<tr>
<td>LINCD</td>
<td>R</td>
<td>0.43</td>
<td>0.10</td>
<td>0.71</td>
<td>0.01</td>
</tr>
<tr>
<td>Range</td>
<td>R</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Row</td>
<td>R</td>
<td>0.07</td>
<td>0.02</td>
<td>0.20</td>
<td>0.00</td>
</tr>
<tr>
<td>f(RANGE)</td>
<td>S</td>
<td>4.97</td>
<td>1.15</td>
<td>39.05</td>
<td>0.59</td>
</tr>
<tr>
<td>f(ROW)</td>
<td>S</td>
<td>55.63</td>
<td>12.89</td>
<td>223.32</td>
<td>3.39</td>
</tr>
<tr>
<td>f(RANGE): ROW</td>
<td>S</td>
<td>0.31</td>
<td>0.07</td>
<td>352.16</td>
<td>5.34</td>
</tr>
<tr>
<td>RANGE:f(ROW)</td>
<td>S</td>
<td>0.55</td>
<td>0.13</td>
<td>66.26</td>
<td>1.00</td>
</tr>
<tr>
<td>f(RANGE):f(ROW)</td>
<td>S</td>
<td>369.15</td>
<td>85.56</td>
<td>5912.83</td>
<td>89.63</td>
</tr>
<tr>
<td>Residual</td>
<td>R</td>
<td>0.34</td>
<td>0.08</td>
<td>2.61</td>
<td>0.04</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>431.47</td>
<td>100.00</td>
<td>6597.15</td>
<td>100.00</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Comparison metrics among the eight models

Visualization via heatmaps of “fitted spatial trend” and residuals plots show us model qualitative effects – specifically, whether the spatial model can remove the overall autocorrelation in the data. Results of effective spatial dimension and variance component analysis show us which spatial trend variables are more important than others from only one model via SpATS. Neither gives us the answer which model is the best among all the models tested via the three data sets. Summary of the side-by-side comparison among the eight models analyzed via data set 1 in Table 6 shows us that the model 8, the P-spline model via SpATS is the best because it has the largest $R^2$ 0.8931, prediction accuracy 0.9473, and residual standard error 0.5582.

In contrast, moving grid adjustment is the least desired model because it has the smallest $R^2$ (0.3409), lowest prediction accuracy (0.6246), and the largest Moran’s I index (0.9492). From data set 1, model OLS with range and row ranks the second-best based on prediction accuracy (0.9200), which is much higher than that from OLS without the range and row in the model. The big difference between model “OLS + range + row” and OLS without the range and row indicated row and range terms are important for correcting the spatial variation caused by row and range. Another noticeable observation is for ASReml, it has the second-largest $R^2$ (0.7289) and second smallest residual standard error (0.6364), but prediction accuracy (0.8538) is low and ranked the sixth of eight models in the order from the best to the worst. If Moran’s I index is used to rank the models, SAR + mixed model is the best since it has the smallest Moran’s I index (0.0349) and the largest p-value of the Moran’s I index (0.02179). Prediction accuracy II from all the eight models is lower than that of prediction accuracy I indicated that the testing materials planted in high IDC pressure area have higher prediction accuracy than that planted in low or no IDC pressure areas. As to the AIC values, which balance between the number of parameters in
the model and model log-likelihood, is not comparable among the eight models. For example, for
the moving grid average, there are only two parameters used in the model to define the grid to
calculate the average values. For P-spline, there are more parameters (142.6 parameters for data
set 1 and 147.8 for data set 3 in Table 4) to fine-tune the boundary of the spatial trends and thus
has much bigger AIC than that from the other models. These model superiority rank differences
based on different evaluation parameters indicated each evaluation parameter has its advantages
and disadvantages as to correcting the spatial patterns in the IDC data.

The results of the model superiority comparison from Table 6 are generally consistent
with the eight heatmaps of the residual plots from data set 1 (Figure 7). The residual heatmap
from model 1, 2, 4, 5, and 6 show a “baseball” pattern, which indicates that there is spatial
autocorrelation left in the residuals and these models did not remove the spatial pattern, whereas
3, 7, and 8 show a circle pattern, which is similar as the raw IDC scores. The heatmap of SpATS
has the most random residuals, and its legend bar has the smallest scales (from -2 to 1), whereas
the rest of them have scales ranging from -2 to 2 or 3. The residual heatmap from “MovingGrid”
is the typical case for positive autocorrelation close to 1, where all the residuals inside the circle
are -1 and +3 outside the circle. Moran’s I index of the residual heatmap from the “MovingGrid”
model also is the largest and is close to 1.

Combining the results from Table 6 and Figure 7, the best models are the “SAR + mixed”
and “P-Spline via SpATS,” and the least desirable model is the “MovingGrid” from the data set
1 analysis. The differences in model effectiveness among the eight models are obviously
reflected by the residual standard errors and residual heatmaps.
Table 6. Summary of the comparison among different models via $R^2$, residual variance, prediction accuracy, Moran’s I index from data set 1. Highlighted are the most desirable values in each column.

<table>
<thead>
<tr>
<th>Model compared</th>
<th>$R^2$ values</th>
<th>AIC values</th>
<th>Residual SE*</th>
<th>Moran’s I index</th>
<th>P-value of Moran’s I</th>
<th>Prediction Accuracy I</th>
<th>Prediction accuracy II</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>0.5200</td>
<td>8978</td>
<td>1.7590</td>
<td>0.5151</td>
<td>2.2e-16</td>
<td>0.7200</td>
<td><strong>0.6882</strong></td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>0.6862</td>
<td>6630</td>
<td>1.0060</td>
<td>0.5159</td>
<td>2.2e-16</td>
<td>0.9200</td>
<td>0.6192</td>
</tr>
<tr>
<td>MovingGrid</td>
<td>0.3409</td>
<td>261</td>
<td>1.0631</td>
<td>0.9492</td>
<td>2.2e-16</td>
<td>0.6246</td>
<td>0.6246</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>0.6580</td>
<td>7794</td>
<td>0.8476</td>
<td>0.4575</td>
<td>2.2e-16</td>
<td>0.8846</td>
<td>0.4146</td>
</tr>
<tr>
<td>SAR + error</td>
<td>0.6811</td>
<td>7658</td>
<td>0.8017</td>
<td>0.4599</td>
<td>2.2e-16</td>
<td>0.8977</td>
<td>0.4279</td>
</tr>
<tr>
<td>SAR + mixed</td>
<td>0.6780</td>
<td>7646</td>
<td>0.7771</td>
<td><strong>0.0349</strong></td>
<td><strong>0.02179</strong></td>
<td>0.9052</td>
<td>0.4422</td>
</tr>
<tr>
<td>ASReml AR1</td>
<td>0.7289</td>
<td>2325**</td>
<td>0.6364</td>
<td>0.2681</td>
<td>2.2e-16</td>
<td>0.8538</td>
<td>0.3581</td>
</tr>
<tr>
<td>P-Spline</td>
<td><strong>0.8931</strong></td>
<td>10065</td>
<td><strong>0.5582</strong></td>
<td>0.0748</td>
<td>5.21e-07</td>
<td><strong>0.9473</strong></td>
<td>0.5124</td>
</tr>
</tbody>
</table>

*Residual SE: residual standard error

**from Asreml: AIC= -2*asreml.Obj$loglik + 2 * length( asreml.Obj$gammas )

Prediction accuracy I: correlation coefficient between predicted and the true IDC scores of all the lines
Prediction accuracy II: accuracy for the lines not planted in the high IDC pressure regions

A comparison of results from data set 2 is summarized in Table 7 and Figure 8. From the highlighted values in Table 7, the group of spatial autoregression with either lag, or error, or Durbin mixed model are the best of the eight models. Based on $R^2$, residual standard error (RSE), and Moran’s I index, SAR + error term as a covariate in the model ranks the best. If based on prediction accuracy, Model SAR + mixed is the best since it has the highest prediction values: 0.9486 and 0.8435 for prediction accuracy I and II, respectively. In contrast to the results from data set 1, the model AR1® AR1 via ASReml ranks the second worst as $R^2$, which is only 0.555, and it is only better than that of the “MovingGrid” model. The biggest result difference between data set 1 and 2 is the P-spline model via SpATS ranked the second-worst model based on prediction accuracy and the third-worst based on $R^2$ in data set 2, whereas P-spline model ranked the best in data set 1 (Tables 6 and 7). From AIC values of the eight model, SpATS has the largest AIC value of 50,678. The low prediction accuracy may come from overfitting.

Another result is that models “OLS w/o RR,” “OLS w/ RR,” and “SAR + mixed” have very similar prediction accuracies, which are 0.9428, 0.9429, and 0.9486, respectively. The very
similar results between models “OLS w/o RR” and “OLS w/ RR” indicated that range and row effects are not as big as that in the data set 1. This result is consistent with the total spatial effective dimensions (EDs) in Table 4, which data set 2 has a total of only 65.5 EDs with a large field of 220 range by 26 rows, whereas data set 1 has 142.6 EDs with a small field of 50 ranges by 42 rows.

All the eight residual heatmaps from the eight models show a clear spatial variation pattern (Figure 10), which is very similar to the raw IDC score heatmap (Figure 5). Heatmaps also show why models ASReml and SpATS rank the second worst because there is a clear separation between high and low IDC scores, and spatial variation patterns were left in the
residual plots, which are the accurate indicators to judge whether a spatial model has removed
the pattern. The models “SAR + mixed” and “SAR + error” have the most randomly distributed
residuals among the eight heatmaps.

Overall, the winners are the “SAR + mixed” and “SAR + error,” and the loser is the
“MovingGrid” from data set 2 analysis in Table 7 and Figure 8.

Table 7. Summary of the comparison among different models via R², residual variance, prediction
accuracy, Moran’s I index from data set 2 (field ID 2040). OLS w/o RR stands for ordinary least
square without the range and row; OLS w/ RR for ordinary least square with range and row.
Highlighted are the most desirable values in each column.

<table>
<thead>
<tr>
<th>Model compared</th>
<th>R² value</th>
<th>AIC value</th>
<th>RSE Value</th>
<th>Moran’s I index</th>
<th>P-value</th>
<th>Moran’s I</th>
<th>Prediction Accuracy 1</th>
<th>Prediction Accuracy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>0.666</td>
<td>22657</td>
<td>1.937</td>
<td>0.2148</td>
<td>2.2e-16</td>
<td>0.9428</td>
<td>0.8295</td>
<td>0.8283</td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>0.648</td>
<td>22793</td>
<td>1.923</td>
<td>0.2131</td>
<td>2.2e-16</td>
<td>0.9429</td>
<td>0.8283</td>
<td></td>
</tr>
<tr>
<td>MovingGrid</td>
<td>0.385</td>
<td>6788</td>
<td>1.8094</td>
<td>0.8664</td>
<td>2.2e-16</td>
<td>0.5839</td>
<td></td>
<td>0.7426</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>0.685</td>
<td>22521</td>
<td>1.3458</td>
<td>0.1655</td>
<td>2.2e-16</td>
<td>0.8356</td>
<td>0.5540</td>
<td></td>
</tr>
<tr>
<td>SAR + error</td>
<td>0.715</td>
<td>22479</td>
<td>1.3080</td>
<td>0.1518</td>
<td>2.2e-16</td>
<td>0.9486</td>
<td>0.8435</td>
<td>0.8435</td>
</tr>
<tr>
<td>SAR + mixed</td>
<td>0.701</td>
<td>22581</td>
<td>1.3408</td>
<td>0.1778</td>
<td>2.2e-16</td>
<td>0.9486</td>
<td>0.8435</td>
<td>0.8435</td>
</tr>
<tr>
<td>ASReML AR1</td>
<td>0.555</td>
<td>14689</td>
<td>1.6765</td>
<td>0.3962</td>
<td>2.2e-16</td>
<td>0.8088</td>
<td>0.5033</td>
<td>0.5033</td>
</tr>
<tr>
<td>B/P-Spline</td>
<td>0.565</td>
<td>50678</td>
<td>1.6156</td>
<td>0.2639</td>
<td>2.2e-16</td>
<td>0.7636</td>
<td></td>
<td>0.4636</td>
</tr>
</tbody>
</table>

Figure 8. Heatmaps of residuals from data set 2 derived from the eight models. Each of the images
has 220 ranges by 26 rows. The X-axis is the row, and Y-axis is the range.
The results from the eight models from data set 3 are very similar (Table 8) to those from data set 2. The model “SAR + mixed” is best because it has the highest \( R^2 \) (0.9491), prediction accuracy (0.9746), and smallest RSE (0.5827) (highlighted numbers in Table 8). However, different results from data set 3 are that model “SAR + lag,” instead of model “SAR + error” like from data set 2, rank the best since it has the smallest Moran’s I index 0.022 and this value is not significant from 0 (p-value = 0.233 > 0.05). The results of the three SAR models, “SAR + lag,” “SAR + error,” and “SAR + mixed,” are very similar to all the parameters for model superiority. Heatmaps of the eight models from data set 3 in Figure 9 have the largest contrast, which models 3 and 7, having almost the same pattern as the heatmap of the raw IDC scores. Models 4, 5, 6, and 8 barely show any spatial patterns in the residual plots. Comparing the legend scales, the models “SAR + error” and “SAR + mixed,” have the smallest residual ranges from -3 to 5, and all the others have the residual ranges from -4 to +4 except that from “MovingGrid” with residual ranges from -2 to +4. The most noticeable heatmap is the one from model “SAR + lag,” in which the residuals are all random noise, with no spatial pattern observed, and Moran’s I close to 0 and with p-value bigger than 0.05. The model “SAR + lag” is the clear winner based on its residual heatmap. Prediction accuracy of I is higher than prediction accuracy II. Prediction accuracy I from data sets 2 and 3 are 0.9746 and 0.9486, which are 11.34% and 12.46% higher respectively than prediction accuracy II with an average 11.90% from the best model SAR + Mixed.

Overall, the winners are the “SAR + mixed” and “SAR + lag,” and the loser is the “MovingGrid” from data set 2 analysis summarized in Table 7 and Figure 8.
Table 8. Summary of the comparison results among the eight models via $R^2$, AIC, residual standard errors, prediction accuracy, Moran’s I index from data set 3. Highlighted are the most desirable values in each column. The values in bold are the largest in each column.

<table>
<thead>
<tr>
<th>Model compared</th>
<th>$R^2$ value</th>
<th>AIC value</th>
<th>RSE value</th>
<th>Moran’s I index</th>
<th>P-value of Moran’s I</th>
<th>Prediction Accuracy 1</th>
<th>Prediction Accuracy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>0.8390</td>
<td>20298</td>
<td>1.8440</td>
<td>0.1456</td>
<td>2.2e-16</td>
<td>0.9534</td>
<td>0.8020</td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>0.8442</td>
<td>20145</td>
<td>1.8150</td>
<td>0.1358</td>
<td>2.2e-16</td>
<td>0.9541</td>
<td>0.8033</td>
</tr>
<tr>
<td>MovingGrid</td>
<td>0.5685</td>
<td><strong>5204</strong></td>
<td>1.6790</td>
<td>0.9306</td>
<td>2.2e-16</td>
<td>0.6569</td>
<td>0.4049</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>0.9076</td>
<td>18251</td>
<td>0.7852</td>
<td><strong>0.0220</strong></td>
<td><strong>0.2330</strong></td>
<td>0.9672</td>
<td>0.8471</td>
</tr>
<tr>
<td>SAR + error</td>
<td>0.9450</td>
<td>17283</td>
<td>0.6059</td>
<td>0.0568</td>
<td>4.5e-08</td>
<td>0.9723</td>
<td>0.8654</td>
</tr>
<tr>
<td>SAR + mixed</td>
<td><strong>0.9491</strong></td>
<td>17180</td>
<td><strong>0.5827</strong></td>
<td>0.0826</td>
<td>1.8e-15</td>
<td><strong>0.9746</strong></td>
<td><strong>0.8753</strong></td>
</tr>
<tr>
<td>ASReml AR1</td>
<td>0.6399</td>
<td>12603</td>
<td>1.5503</td>
<td>0.4233</td>
<td>2.2e-16</td>
<td>0.8448</td>
<td>0.5004</td>
</tr>
<tr>
<td>B/P-Spline</td>
<td>0.7494</td>
<td>41241</td>
<td>1.2935</td>
<td>0.1169</td>
<td>2.2e-16</td>
<td>0.8684</td>
<td>0.5732</td>
</tr>
</tbody>
</table>

Figure 9. Heatmaps of the residuals from data set 3 derived from the eight models.
Figure 10. Boxplot of the residual standard error (RSE). Y axis is the residual standard errors. First three boxplots from left belong to Group I, middle three belongs to Group II, and the two from the right belong to group III.

Table 9. Results of the Kruskal-Wallis test.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Model No</th>
<th>RSE Rank</th>
<th>Significant tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>M1</td>
<td>21.67</td>
<td>a</td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>M2</td>
<td>17.67</td>
<td>ab</td>
</tr>
<tr>
<td>MovingGrid</td>
<td>M3</td>
<td>16.00</td>
<td>abc</td>
</tr>
<tr>
<td>ASReml AR1</td>
<td>M7</td>
<td>12.00</td>
<td>bcd</td>
</tr>
<tr>
<td>B/P-Spline</td>
<td>M8</td>
<td>9.33</td>
<td>cd</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>M4</td>
<td>9.33</td>
<td>cd</td>
</tr>
<tr>
<td>SAR + error</td>
<td>M5</td>
<td>7.33</td>
<td>d</td>
</tr>
<tr>
<td>SAR + mixed</td>
<td>M6</td>
<td>6.67</td>
<td>d</td>
</tr>
</tbody>
</table>

From the comparisons among the eight models through three data sets, it looks like model 6, “SAR + mixed,” works the best for data sets 2 and 3, and model 8, “P-Spline via SpATS,” works the best for data set 1, based on the results in Tables 6, 7, and 8 and Figures 7, 8, and 9. With the limited knowledge of the distribution of the 3 RSE data points per model from the three data sets, we chose to conduct a distribution-free nonparametric Kruskal-Wallis test to assess whether there is a significant difference among the eight models. The significance test in
the last column of Table 9 shows that four significant levels as “a,” “b,” “c,” and “d” were reached. Model 1, “OLS w/o RR “ is the worst, and Model 1, 2, and 3 are not significantly different. Similarly, model 6, “SAR + mixed” is the best, models 4, 5, 6, 7, and 8 are not significantly different based on the Kruskal-Wallis test by ranks. Based on the boxplot (Figure 10), three groups were observed, where models 1, 2, and 3 are the group I, models 4, 5, and 6 are group II, and Models 7 and 8 are groups III. Interestingly, these three groups clustered by RSE are consistent with the model’s mathematical and spatial covariate structures. With the criterion of the smaller RSE, the better the models, models 4, 5, and 6 in group II rank the best. Pairwise t-test among the three groups shows that the group I and II are extremely significantly different, reaching to the level p-value < 0.001, Groups II and III show significant differences with P-value < 0.01. Group I and III are statistically different at a p-value of 0.1, but not 0.05 (Table 10).

Table 10. Results of pairwise t-test of the groups mean.

<table>
<thead>
<tr>
<th>Model group pairs</th>
<th>Mean of 1st group</th>
<th>Mean of 2nd group</th>
<th>Mean difference</th>
<th>t-test stats</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups I vs II</td>
<td>1.78</td>
<td>0.66</td>
<td>1.12</td>
<td>-6.3367</td>
<td>0.00039**</td>
</tr>
<tr>
<td>Groups I vs III</td>
<td>1.78</td>
<td>1.42</td>
<td>0.36</td>
<td>-2.0133</td>
<td>0.08396</td>
</tr>
<tr>
<td>Groups II vs III</td>
<td>0.66</td>
<td>1.42</td>
<td>-0.76</td>
<td>3.5051</td>
<td>0.00993*</td>
</tr>
</tbody>
</table>

Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Overall, the Group III models, “SAR + mixed,” “SAR + error,” “ and “SAR + lag” are the best options for spatial autoregressive analysis for ordinal data type as IDC.

Relative efficiency (RE) of the spatial autoregressive (SAR) analyses

The results in Table 11 summarize the relative efficiencies of the eight models, compared with model 1, ordinary least square (OLS) without range and row (OLS w/o RR), in the second to the last column of Table 11. Model 6, “SAR + mixed,” “has the highest relative efficiency 420.82%, which indicates that four more replicates of model 1 experimental design are needed in
order to reach the same residual error as that from model 6. All the relative efficiencies from the other seven models are bigger than 100%, showing that range and row are very important for field spatial variation correction. The last column of Table 11 shows the relative efficiency of the seven models compared with model 2, ordinary least square (OLS) with “range” and “row” in the model (OLS w/ RR). All the six models with “range” and “row,” from Model 3 to Model 8, have bigger than 100% relative efficiencies and varied from 104.23% for Model 3 to the largest 175.66%. The variation of relative efficiency among the six models indicates that the way of modeling “range” and “row” for the spatial autocorrelation resulted in big differences. The overall analysis of the relative efficiency of the different models shows the significant importance of applying appropriate spatial autoregressive models to correct spatial field variation.

Table 11. Summary of relative efficiency (RE) of spatial analysis vs. Model 1 and 2

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Model Number</th>
<th>Mean RSE</th>
<th>RE to M1 (%)</th>
<th>RE to M2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS w/o RR</td>
<td>M1</td>
<td>1.8467</td>
<td>100.00</td>
<td>85.63</td>
</tr>
<tr>
<td>OLS w/ RR</td>
<td>M2</td>
<td>1.5813</td>
<td>136.37</td>
<td>100.00</td>
</tr>
<tr>
<td>MovingGrid</td>
<td>M3</td>
<td>1.5172</td>
<td>148.15</td>
<td>104.23</td>
</tr>
<tr>
<td>ASReml AR1</td>
<td>M7</td>
<td>1.2877</td>
<td>205.65</td>
<td>122.80</td>
</tr>
<tr>
<td>SAR + lag</td>
<td>M4</td>
<td>0.9929</td>
<td>345.94</td>
<td>159.27</td>
</tr>
<tr>
<td>B/P-Spline</td>
<td>M8</td>
<td>1.1558</td>
<td>255.29</td>
<td>136.82</td>
</tr>
<tr>
<td>SAR + error</td>
<td>M5</td>
<td>0.9052</td>
<td>416.19</td>
<td>174.69</td>
</tr>
<tr>
<td>SAR + mixed</td>
<td>M6</td>
<td>0.9002</td>
<td>420.82</td>
<td>175.66</td>
</tr>
</tbody>
</table>

Lagrange multiplier test (LMT)

Group II models are the best choice for the data tested in this research, but the computation time is much slower than that from the other two groups of models. For data set 2, it took >10 hours to run the “SAR + mixed” model with 5,720 data points, a total of ~25 hours to tune the three data sets via a desktop Windows 10 with 16Gb Ram and core i7 CPU, whereas SpATS took less than 15 minutes to run the three datasets. The Lagrange Multiplier Test (LMT)
run on each data set took less than 1 minute and provided an overview of the models (Table 12).

LMT reports the estimates of tests among the models for spatial dependence.

Table 12. Compare among five autoregressive based models from the Lagrange Multiplier test statistics

<table>
<thead>
<tr>
<th>Spatial models</th>
<th>Variables Tested for</th>
<th>Model categories</th>
<th>dependence estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR+error</td>
<td>for error dependence</td>
<td>Spatial error model</td>
<td>2730.3</td>
</tr>
<tr>
<td>RLMerr</td>
<td>lag variance except LMerr</td>
<td>error + possible lag</td>
<td>108.98</td>
</tr>
<tr>
<td>SAR+lag</td>
<td>for lagged variable</td>
<td>Spatial lag model</td>
<td>2728.1</td>
</tr>
<tr>
<td>RLMlag</td>
<td>for error variance except LMlag</td>
<td>lag + error model</td>
<td>106.75</td>
</tr>
<tr>
<td>SAR+mixed</td>
<td>for both error and lag model</td>
<td>Spatial Durbin mixed</td>
<td>2837.1</td>
</tr>
</tbody>
</table>

From Table 12, the model with the largest dependence estimates should be chosen for SAR analysis. From the dependence estimates, spatial lag and spatial error models have very similar values, and their dependence estimates are 2728.1 and 2730.3, respectively. Tests for the possible presence of lagged variance except for LMerr, RLMerr, is 108.98. Similarly, tests for the possible presence of error variance (except LMlag, RLMlag), is 106.75. These two values are very similar. The last test is for both error and lag model, “SAR+mixed,” its dependence estimate is 2837.1, which is bigger than any of the two models as expected. Based on this LMT test results, “SAR+mixed” should be selected for the data set1 to correct the autocorrelations.

**Statistical experiment design for spatial analysis versus breeding practice**

The basic principles of experimental designs are randomization, replication, and local control (Mead et al. 2012). For IDC testing, replication, and local control with incomplete alpha-lattice design can be easily implemented. But the randomization of previously untested lines may be violated. Breeders usually group the testing lines by families for easy visual evaluations. For example, a typical breeding program might evaluate 32 recently developed lines per family from 400 families for a total of 12,800 lines. Soybean breeders usually organize the IDC evaluations of the lines into 400 trials, where the lines are randomized within each trial. But
the 400 trials representing variability among families are usually not randomized across the field site. Rather, the families are arranged by pedigree and relative maturity for purposes of visual comparisons within families and to accommodate operational considerations such as avoiding inter-plot competition between early and late maturing lines and avoiding damage to plots from mechanical harvest. The combine and other equipment will begin with plots that are planted with early maturing families and proceed through the field as the plots mature.

If a family is created by crossing between a resistant line and a resistant line or moderately resistant by moderately susceptible lines, there will be patches of resistant or susceptible lines, and the families will create patterns unrelated to the non-genetic field IDC patterns. This practice may confound the genotype pattern caused by experiment design with spatial variation patterns caused by environmental factors. Under this circumstance, the spatial analysis may lead to biased selections. Randomization of experiments and trials is a prerequisite for the three groups of spatial analyses. In the context of spatial variation, such as is observed with IDC, complete randomization of the testing lines in the testing site is needed in order to minimize biased selections from the confounding factors, and to add pedigree information in the spatial model to adjust the effects caused by the breeding practices.

**IDC hill plot size and spatial variation**

Plot size may affect spatial patterns and, subsequently, the effectiveness of spatial autoregressive models. Soybean IDC testing lines were planted in hill plots, not row-plots, and the plot size is very small (Figure 11). Each hill plot contained eight seeds. Hill plots were spaced 15 inches from center to center following up the ranges and 10 inches from center to center between rows (left-hand image in Figure 11), whereas yield plot size is usually much larger than that of IDC hill plot (right-hand image in figure 11). The geospatial statistical experiment plot/unit is usually larger than a whole soybean yield trial – similar, for example, to
the number of flu patients in a county. There are reports that differences in plot size impact the effectiveness of spatial models. Results from uniformity trials testing show that large plot sizes are needed to control field heterogeneity and spatial variation (KNORZER et al. 2013). However, contrary to this theory that large plot sizes should be better, evidence from a 28-year case study for optimizing experimental designs show that relative efficiency of the experiment design was 240% versus RCBD when the plot size decreased from 5.6 m to 1.4 m row length (CASLER 2013). Similar results from a comparison of different spatial models among correlated error, nearest neighbor analysis, and autoregressive regression AR(1) indicates that smaller plot size is more efficient to capture the spatial variation and thus increase the relative efficiency of the experimental design (SRIPATHI et al. 2017). Both small plot size and ordinal data type of IDC scores may cause P-spline via SPATS high residual standard error in this research.

Figure 11. Images of IDC hill plots vs. yield testing plots. Images were taken of the 2016 IDC evaluation trial in Fargo, ND, at plant growth stages V2 and V4. The image in the left-hand shows the expected chlorosis phenotype of two susceptible testing lines (top two hill plots) and two resistant testing lines (bottom two hill plots).
The tensor product penalized splines may work better for continuous data type than for ordinal data type

Results from the publication from the author of SpATS show that tensor product panelized splines worked very well for hybrid wheat data for both Chilean and Australian wheat field data sets (RODRÍGUEZ-ÁLVAREZ et al. 2016), sorghum grain yield and plant height (VELAZCO et al. 2017a). Repeating the analysis on these datasets, we got the same results. However, when this method was applied to soybean IDC data sets, two unexpected results were obtained: 1) from the effectiveness dimension analysis of the decomposed of the model variables, the genotype or line effectiveness accounts for about 90% of the total effectiveness, while in terms of genotype effectiveness, the tensor product term “f(ROW):f(RANGE)” accounts for less than 10% of the effectiveness. In contrast to the effectiveness dimension component analysis, the variance component analysis shows that tensor product term “f(ROW):f(RANGE)” account for over 90% of the total variance, whereas genotype accounts for only less than 1%;

These unexpected results were communicated with the author of SPATS, and the author responded that the IDC data type is ordinal, not continuous data. Most likely, the ordinal data type is the reason for the biased results, and the author is working on the model and will release a higher version of SPATS with the capacity of analysis of ordinal data type.

Conclusion

The effectiveness of spatial models depends on many factors, such as the combination of the soil characters, weather conditions, and trialing activities, the severity of the spatial variation, and other types of irregular patterns. From the comparison results of residuals standard error (RSE), $R^2$, prediction accuracy, AIC, and their heatmaps generated for the eight models, none of them can completely remove the spatial autocorrelation for the ordinal data type in the three data sets and get completely randomly distributed residuals. However, the spatial autoregressive
(SAR) approach (with either lag, or error, or Durbin mixed as a covariate) generated more random residual plots; and most of the time, it was able to smooth the spatial surface and identify and correct spatial trends in the data sets better than the other two groups of models. The tensor product panelized P-splines method works the best for the simulated data set 1, which has only one spatial pattern (a circle). As to the computation time and user-friendliness, P-spline via SpATS is the fastest and the easiest to run. The higher residual standard error (RSE) and lower $R^2$ from data set 2 and 3 via SpATS than that from SAR may come from some degree of overfitting the model with large data sets (Both data sets 2 and 3 have more than 5,000 plots).

Conflict of interest

The authors declare that there is no conflict of interest.

Supplemental data available

Supplemental material is available online for this article. R codes and the data sets used for this research are available online.

Acknowledgments

We thank Syngenta, Iowa State University, and USDA-ARS for the financial support of this research. This research was supported in part by the US. Department of Agriculture, Agricultural Research Service, project 5030-21000-062-00D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References


Cappa, E.P. and R.J.C. Cantet. 2007. Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. Canadian Journal of


Supplemental Figure 1. Layout of the grid used to calculate the mean

\[ y_1 \quad y_2 \quad y_3 \]
\[ y_4 \quad y_5 \quad y_6 \quad y_7 \quad y_8 \]
\[ y_9 \quad y_{11} \quad y_{12} \quad y_{13} \]
\[ y_{14} \quad y_{15} \quad y_{16} \quad y_{17} \quad y_{18} \]
\[ y_{19} \quad y_{20} \quad y_{21} \]
CHAPTER 4. ALGORITHMIC MODELING VS. DATA MODELING: WILL THE ALGORITHMIC OUTPERFORMS DATA MODELING FOR IRON DEFICIENCY CHLOROSIS RESISTANCE?

Zhanyou Xu¹, Andreomar Kurek¹, Steven B. Cannon², Williams D. Beavis³*

¹Bioinformatics and Computational Biology, Iowa State University, Ames Iowa 50014
²USDA-Agricultural Research Service, Corn Insects, and Crop Genetics Research Unit, Ames, IA 50010
³Department of Agronomy, Iowa State University Ames, Iowa 50010.

Modified from a manuscript to be submitted to Crop Science

Abstract

Soybean iron deficiency chlorosis (IDC) is one of the major yield-reducing factors in the U.S. upper Midwest. Marker-assisted selection (MAS) has not been successful for breeding for IDC tolerance, due to the complexity of trait. The genomic prediction has been extensively applied to increase selection accuracy for continuous numeric traits such as yield and plant height. For ordinal data types such as IDC, which are typically scored on a scale of 1-9 (resistant to susceptible), genomic prediction methods have not been systematically compared. Data modeling methods include logistic regression and ridge-regression BLUP (rrBLUP), and algorithm modeling methods include neural networks, decision trees, Bayesian inference, and support vector machines. The objectives of this research 1) to select the best model measured by discard specificity, selection sensitivity, selection precision, overall prediction accuracy, and receiver operating curve (ROC); 2) to compare prediction accuracies between data modeling approaches (rrBLUP and logistic regression) and algorithmic modeling methods (random forest (RF), gradient boosting algorithm (GBM), support vector machine (SVM), K-nearest neighbors (KNN), Naïve Bayes (NB), and artificial neural network (ANN)). We find that 1) Among the
eight tested models, SVMs generated the highest discard specificity for de-selecting IDC susceptible lines, while RF has the highest selection sensitivity for selecting IDC resistant lines, and RF has the highest overall accuracy; 2) for the soybean IDC ordinal data type, algorithmic modeling outperforms data modeling methods.

Introduction

Soybean iron deficiency chlorosis (IDC) is one of the major stress traits that lead to yield losses in the U.S. upper Midwest. The average annual yield losses are estimated at 340 million tons, worth an estimated 120 million dollars per year (Hansen et al. 2004). DNA marker-assisted selection (MAS) for IDC tolerant breeding has generally not been successful, even though more than 100 IDC QTLs have been mapped by parental mapping, network association, nested association mapping (NAM), and genome-wide association studies (GWAS) (Charlson, Cianzio, et al., 2003, Cianzio and Fehr, 1980, King, Peiffer, et al., 2013, Lin, Cianzio, et al., 1997, Lin, Grant, et al., 2000, Mamidi, Chikara, et al., 2011, O'Rourke, Charlson, et al., 2007, Wang, McClean, et al., 2008). This is likely due both to the genetic complexity of the trait and the difficulty in phenotyping the trait, which still relies on labor-extensive field IDC screening.

Several hurdles block soybean breeders from using markers to stack the IDC QTLs to breed for IDC tolerance. The first hurdle is that there are no major, universal QTLs for IDC, in contrast (for example) to the soybean cyst nematode QTL derived from PI88788, which provides resistance to 95% of the commercial soybean varieties (Mitchum, 2016). The only major IDC QTL was mapped in a population between Anoka x A7, which identified a QTL that explains more than 70% of the phenotype variation (Cianzio and Fehr, 1982, Fehr, 1982). However, no further reports have reported that this major IDC QTL provides similar resistance in other genetic backgrounds. The second hurdle is that there is no efficient way to use MAS to stack many QTLs for IDC tolerance breeding due to the limited population size versus the number of
minor-effect IDC QTLs – many of which have been reported, including 41 IDC QTLs from bi-
parental mapping populations, 50 QTLs from connected network population mapping, 88 QTLs
from GWAS QTLs deposited at SoyBase (https://soybase.org/) and described in (Mamidi,
Chikara, et al., 2011, Wang, McClean, et al., 2008). A total of 835 candidate genes in the IDC
tolerant line Clark (PI548553) were identified by transcriptome sequencing (O'Rourke, Nelson,
et al., 2009). Considering a large number of IDC QTLs/candidate genes identified so far, it is
very challenging to develop sufficiently large populations for breeding IDC tolerant varieties via
MAS selection. For example, a minimum of 16,384 progenies from an F2 population is needed
in order to select one plant with seven fixed IDC QTLs. The third hurdle is that identified IDC
QTL effects are dependent on the particular population and genetic background (Lin, Grant, et
al., 2000), due to gene-by-gene interactions (epistasis). All in all, traditional marker-assisted
selection for IDC tolerant breeding has been ineffective because IDC trait is complex and
affected by many genes and environmental factors, each with small effect (Jannink, Lorenz, et
al., 2010, Rodriguez de Cianzio and Fehr, 1982).

Genomic selection (GS) and machine learning for predictive breeding have arisen from
the conjunction of new high-density genotyping and whole-genome sequencing technologies,
new statistical methods, and machine learning algorithms. GS uses a training pool of individuals
that have been both genotyped and phenotyped, to train the model that takes genotypic data from
un-phenotyped individuals and produces genomic estimated breeding values (GEBV) (Bernardo,
2008, Meuwissen, Hayes, et al., 2001). In contrast to QTL-based MAS breeding, neither QTL
mapping nor marker selection is required for GS.

GS accuracy is generally estimated with the Pearson correlation, r, which measures the
degree of correlation or consistency between GEBV and the observed phenotypic values
Breeding experiment designs for GS have been proposed and studied in maize (Gorjanc, Jenko, et al., 2016, Lian, Jacobson, et al., 2015), barley (Iwata and Jannink, 2011, Sallam, Endelman, et al., 2015, Schmid and Thorwarth, 2014), apple (Muranty, Troggio, et al., 2015), and animals (Hayes, Bowman, et al., 2009, Montaldo, Casas, et al., 2012). Regardless of the breeding designs, the prediction accuracy needs to be high enough for genomic selection to be time and cost-effective. The expected prediction accuracy $E(r_{mg})$ has been previous defined as a function of population size (N), heritability of the trait of interest ($h^2$), and number of effective markers that affect the trait of interest ($M_e$). The formula is:

$$E(r_{mg}) = \left[ \frac{N h^2}{N h^2 + M_e} \right]^{1/2} \text{ (DAETWYLER et al. 2008; HAYES et al. 2009b)}$$

From the above equation, the heritability of the trait of interest is fixed for the specific trait, and the population size is under the breeder’s control. The variable that is more difficult to determine is the number of effective markers, $M_e$, that relate to the specific trait. The $M_e$ pertains to the number of the idealized concept of independent chromosome fragments, with each fragment containing a QTL-marker pair (Daetwyler, Villanueva, et al., 2008, Lian, Jacobson, et al., 2014). When the marker density is high enough to cover every functional gene in the genome, $M_e$ can be estimated on the basis of the effective population size and the genome size (Lorenz, 2013, Lorenz and Smith, 2015).

Four statistical assumptions are required to estimate the prediction accuracy correctly using the equation (Daetwyler, Villanueva, et al., 2008): 1) the marker effects can be derived from simple linear regression; 2) each marker–QTL pair is independent of another marker–QTL pairs; 3) different marker–QTL pairs have equal variances; 4) each marker–QTL pair is in complete linkage disequilibrium (LD). Not all markers in the high-density chip are in complete
linkage disequilibrium with QTL. The equation was updated by Lian et al. (Lian, Jacobson, et al., 2014) to retain the first three assumptions but to relax the 4th assumption by adding an extra variable \( r^2_{MM/2} \) in the equation as below:

\[
E(r_{mg}) = r^2_{MM/2} \left[ \frac{N h^2}{(r^2_{MM/2} N h^2 + M_e)} \right]^{1/2} \quad (\text{LIAN et al. 2014})
\]

Where \( r^2_{MM/2} \) equals to the mean squared correlation between a marker and QTL when the QTL is assumed to be at the midpoint of the two markers.

Prediction accuracy of GS has been reported from different species by different statistical models with different populations or panel sizes. Here are the two representative reports with either a large number of lines from barley or a large number of populations in maize. In barley, results from an analysis of 1536 SNPs and 647 lines with four traits differing in genetic architecture from 5 years phenotypic data show that higher trait heritability, h2, in the training population and simpler trait architecture were associated with higher prediction accuracy, and fixation of markers/loci associated with a trait over time was most clearly associated with reduced prediction accuracy, and the prediction accuracy ranged from 0.03 to 0.99 (Sallam, Endelman, et al., 2015). In maize, the progeny of 969 bi-parental populations were genotyped with 31 to 119 (with a mean of 70) polymorphic SNPs, and GEBVs were calculated, and the prediction accuracy range from -0.59 to 1.03 with a mean of 0.45 for grain yield and authors concluded that it is difficult to predict the accuracy, \( r_{MG} \) in advance, but the rule of thumb based on \( r^2_{MM/2}(N h^2)^{1/2} > 8 \) can help to increase prediction accuracy (Lian, Jacobson, et al., 2014).

Ridge-regression BLUP (rrBLUP) has been widely used for GS because of its open-source and ease of use, including for the data sets with a larger number of markers than the number of phenotype data points (Endelman, 2011). The early version of rrBLUP can only be...
applied to estimate additive marker effects. The ability to accommodate the dominance effect was added in 2014 (Nishio and Satoh, 2014). Results from simulated data show that rrBLUP with both dominance and additive effects is a feasible approach to improve genetic performance in crossbred populations with large dominance genetic variation and identify mating systems or best crosses with a good combining ability (Nishio and Satoh, 2014, Perez-Rodriguez, Gianola, et al., 2013). A more complicated model that can include additive, dominance, and epistasis was implemented in r package “sommer,” and results from both simulated and field phenotypic data show prediction accuracy from “sommer” can be increased in species displaying heterotic effects, which require the estimate of general combing ability (GCA)/dominance and special combining ability (SCA)/epistasis (Covarrubias-Pazaran, 2016).

Since GS was first proposed in 2001 (Meuwissen, Hayes, et al., 2001), it has been widely studied, with results indicating that GS has more predictive power than classical marker-assisted selection in both empirical and simulated data (Heffner, Jannink, et al., 2011, Heffner, Lorenz, et al., 2010, Lorenz, 2013, Lorenz and Smith, 2015, Lorenzana and Bernardo, 2009). However, results from some of the studies show that there is a very wide range of prediction accuracy, for example from -0.59 to 1.03 for maize grain yield (Lian, Jacobson, et al., 2014), and from -0.41 to 0.94 for palm oil yield (Cros, Denis, et al., 2015). One reason for the variability is probably the choice of statistical models used for the data analysis. Traditional methods for data analysis, such as linear regression and linear discriminant analysis, are based on predefined distributions with required model assumptions. These methods can have a relatively high prediction accuracy only when the data meet the model requirements, and the model emulates the nature of the data (Breiman, 2001). For example, three assumptions for linear regression include 1) Linear relationship between the phenotype and marker; 2) Multivariate normality; 3) No
multicollinearity among the predictor variable (e.g. SNP markers). In reality, most of the data sets for GS do not strictly comply with the assumptions of linear relationship and multicollinearity, because most SNPs from high-density chip are linked with the trait of interested.

From analyzing the data to drawing final conclusions, two different modeling approaches have been applied to analyze the data. These approaches are distinct enough that they have been termed statistical “cultures” by Breiman(Breiman, 2001). The “data modeling” approach uses pre-defined, distribution-based stochastic data models; examples include linear regression, logistic regression, Poisson, and chi-square modeling. The “algorithmic modeling” approach uses a distribution-free algorithm; examples include random forest classification, support vector machines, and artificial neural networks. The statistical community has generally been committed to the use of data modeling, with about 98% of the statistical analysis using the data modeling approach, according to Breiman’s estimation in 2001. In contrast, distribution-free based algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used on large complex data sets and as a more accurate and informative alternative to data modeling. The majority of the machine learning algorithms do not need model assumptions and prior knowledge of the data distribution; these belong to the algorithmic modeling approach and provide an alternative to distribution-based data modeling for improving GS prediction accuracy.

There are several extensive studies of comparing the prediction accuracy among different models for continuous traits, such as grain yield, plant height, thought kernel weight, etc. (Hayes, Bowman, et al., 2009, Lorenzana and Bernardo, 2009, Ogutu, Piepho, et al., 2011, Ogutu, Schulz-Streeck, et al., 2012, Ratcliffe, El-Dien, et al., 2015, Shu, Wu, et al., 2011, Zargar, Raatz,
et al., 2015). In contrast, only a few GS studies have been published that use ordinal categorical phenotypes. These studies include genomic-enabled prediction of ordinal data with bayesian logistic ordinal regression for maize gray leaf spot (Montesinos-Lopez, Montesinos-Lopez, et al., 2015) and GS prediction for resistance to pear black spot resistance (Iwata, Hayashi, et al., 2013). In addition, most GS model complications use overall Pearson correlation coefficients to compare between the predicted GEBV and observed values as prediction accuracy, but this measure is not specifically focused on “selection sensitivity” (the accuracy of selecting the resistant without false susceptible lines), “selection precision” (the accuracy of selecting resistant without false resistant lines), or “discard specificity” (the accuracy of discarding the susceptible without false resistant lines). In a large breeding program, a breeder wants to know what stocks should be selected or discarded. Hence, we evaluate model superiority using the selection sensitivity, discard specificity, the selection precision, the balance between false-positive and false-negative rate by the area under the curve (AUC), and the overall accuracy among the models. The eight models include logistic regression and ridge regression (from the data modeling approach), and the Naïve Bayes (NB), Random Forest (RF), K-nearest Neighbour (KNN), Support Vector Machine (SVM), Gradient Boost Machine (GBM), an Artificial Neural Network (ANN) (from the algorithmic modeling). The objectives of this research are to systematically compare the model superiority for ordinal data, represented by soybean IDC tolerance, between the data- and algorithmic modeling approaches. Specifically, we ask: 1) What model has the highest selection sensitivity? 2) What model has the highest discard specificity? 3) What model has the highest overall prediction accuracy? 4) How do models from the two statistical modelings “cultures” compare in terms of prediction accuracy?
Data and Methods

Data

A total of 1,000 F6 derived F10 lines from four years of a breeding program were used for this study. These lines were tested from 2013 to 2016, with two replicates per location, and two to four locations per year, depending on the line development stages. Phenotypic scores ranged from 1 to 9, with “1” as most resistant and “9” as most susceptible. The 1,000 lines were genotyped using 1200 SNPs, assayed by a 1.5K Illumina chip (1200 of the SNPs were of sufficient quality for use in the study).

Analytic methods

The IDC phenotype best linear unbiased prediction (BLUP) was estimated from multiple locations and years’ data via the R package “lme4” and then scaled into 1 for tolerant and 0 for susceptible. The mixed model for calculating the phenotype IDC BLUP is: \[
\text{IDC score} = \mu + \text{Year} + \text{Location} + \text{Line} + \text{experiment} + \text{Location x Line} + \text{Year x Line} + e.
\] All terms are treated as random effects except overall mean \(\mu\).

Genotype data were imputed with the “impute” R package. An artificial neural network (ANN) uses two hidden layers with the “backdrop” algorithm for the prediction, using the “neuralnet” package. The GBoost is a recently developed, deep learning algorithm. It is an optimized distributed gradient boosting machine (GBM), and it is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost has won the reputation of “winning every machine learning competition”.

Random forest (RF), rrBLUP, logistic regression, KNN, SVM, and Naïve Bayes (NB) analyses were conducted via “randomForest,” “rrBLUP,” “stats,” “class,” “e1071”, “naive Bayes” packages, respectively. All the R codes were saved as an R markdown file as supplemental material for this report.
Measurement metrics for prediction accuracy

Pearson correlation coefficients

Pearson correlation coefficients between the predicted and observed values

\[ \rho_{x,y} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]  

(OTT AND LONGNECKER 2010)

Where \( \text{cov}(X,Y) \) is the covariance of X and Y, \( \sigma_x \) is the standard deviation of X, \( \sigma_y \) is the standard deviation for Y. \( \bar{x} \) and \( \bar{y} \) are the means of X and Y, respectively. \( x_i \) and \( y_i \) are the observation values for X and Y vectors from 1 to n, respectively.

Selection sensitivity (SS)

Selection sensitivity is the true positive rate, and it measures the proportion of positives that are correctly identified (GADDIS AND GADDIS 1990). The selection sensitivity is also known as the probability of detection. In this research, selection sensitivity is used to measure the accuracy to predict the resistant IDC lines with the risk of discarding resistance as susceptible lines. The higher the selection sensitivity (close to 1), the less risk of discarding resistant lines as “false susceptible lines.” The equation is as below:

\[
\text{Selection sensitivity} = \frac{\sum \text{true positives (TP)}}{\sum \text{true positives} + \sum \text{false negatives}} = \frac{\Sigma \text{true IDC R lines}}{\Sigma \text{true R lines} + \Sigma \text{false Susceptible lines}}
\]

Discard specificity (DS)

Discard specificity measures the proportion of negatives that are correctly identified as negatives. In this research, specificity is used to measure the accuracy to discard the susceptible IDC lines with the risk of keeping the susceptible lines in the future extensive yield testing, which will result in wasting resources. The closer to 1 the discard specificity, the lower risk keep susceptible lines in the future breeding program. The equation is as below:
Discard specificity = \[ \frac{\Sigma \text{true negatives}}{\Sigma \text{true negatives} + \Sigma \text{false positives}} = \frac{\Sigma \text{true IDC susceptible lines}}{\Sigma \text{true susceptible lines} + \Sigma \text{false resistant lines}} \]

**Selection precision (SP)**

Selection precision is the ratio of true positives to combined true and false positives. In this research, selection precision is used to measure the accuracy to predict the resistant IDC lines with the risk of selecting the susceptible lines as resistant lines in the breeding program, which will result in wasting resources to test susceptible lines in the subsequent extensive yield testing. The higher the selection precision (close to 1), the less risk of selecting susceptible lines in the future breeding program.

Selection precision = \[ \frac{\Sigma \text{true positives (TP)}}{\Sigma \text{true positives} + \Sigma \text{false positive}} = \frac{\Sigma \text{true R lines}}{\Sigma \text{true R lines} + \Sigma \text{false R lines}} \]

**Overall prediction accuracy (OPA)**

Overall accuracy is the percentage of true positives and true negative out of the testing lines.

Accuracy = \[ \frac{\Sigma \text{true positive} + \Sigma \text{true negatives}}{\Sigma \text{total populations}} = \frac{\Sigma \text{true resistant lines} + \Sigma \text{true susceptible lines}}{\Sigma \text{total lines}} \]

The area under the curve (AUC) and the receiver operating characteristic curve (ROC) curve is generated by plotting the true positive rate (TPR) as the y-axis against the false positive rate (FPR) as x-axis at various threshold settings. In most of the classification, 0.5 is used as the default threshold to differentiate the positives from the negatives. ROC curve is plotted with all the thresholds between 0 to 1. The area under the receiver operating characteristic curve (ROC) is AUC, and it used to estimate the model performance and model stability where the closer to 1 the AUC, the better the model (Pepe et al. 2006; Peterson and Coleman 2008). AUC was calculated via the “ROCR” package (Sing et al. 2005).
Results and discussion

IDC overall prediction accuracy from data modeling, rrBLUP, and logistic regression

GS prediction accuracy from rrBLUP is larger than that of logistic regression (Table 1). The Pearson correlation coefficients are 0.8737, 0.8742, and 0.8738 for three repeats of 10-fold cross-validations, respectively, whereas the prediction accuracy for logistic regression is 0.562. Ridge regression is better than logistic regression because of not only the higher prediction accuracy but also the feature selection of logistic regression. The data set used in this research is 1,000 phenotypes by 1200 SNP markers, a typical case in which the number of markers is larger than the number of samples. Logistic regression analysis can’t be conducted with this data due to more unknown variables than the number of observations, which results in inestimable for the 1200 SNP marker. A total of 200 or more markers have to be removed from the data set before fitting the logistic regression. Ridge regression overcome this limitation by adding a shrinkage lambda factor and can estimate all the 1200 markers in the model. The smaller standard deviation from rrBLUP, 0.0275, 0.0209, and 0.0240 for the three cross-validations, than that of logistic regression, 0.0471, indicate that rrBLUP has higher model stability than logistic regression. The mean discard specificity, selection sensitivity, and selection precision of the 10-fold cross-validation are 0.5651, 0.5599, and 0.5418 from logistic regression, respectively. Since ridge regression can only provide numeric predictions as IDC scores, it can’t provide a classification of resistance or susceptible so that there is no specificity, sensitivity, and precision estimates from rrBLUP. Results from both prediction accuracy and standard deviation from the 10-fold cross-validation show that rrBLUP is better than logistic regression for ordinal IDC scores. Since rrBLUP is the widely used genomic prediction algorithm and provide relatively high prediction accuracy (Jeffrey 2011; Ogutu et al. 2012; Piepho et al. 2012), the accuracy from rrBLUP is used as the benchmark to compare with the rest of the models in this research.
Table 1. Prediction accuracy results from 10-fold cross-validation by rrBLUP and logistic regression with 1000 lines by 1200 SNP markers. The 10-fold cross-validations were repeated three times for the rrBLUP which is the most frequently used GS model for plant breeding.

<table>
<thead>
<tr>
<th>Cross Validation</th>
<th>Ridge Regression BLUP</th>
<th>Logistic Regression Accuracy*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Fold #</td>
<td>Repeat1</td>
<td>Repeat2</td>
</tr>
<tr>
<td>1</td>
<td>0.8862</td>
<td>0.8853</td>
</tr>
<tr>
<td>2</td>
<td>0.8883</td>
<td>0.8667</td>
</tr>
<tr>
<td>3</td>
<td>0.8978</td>
<td>0.8631</td>
</tr>
<tr>
<td>4</td>
<td>0.9007</td>
<td>0.8855</td>
</tr>
<tr>
<td>5</td>
<td>0.8614</td>
<td>0.8619</td>
</tr>
<tr>
<td>6</td>
<td>0.8762</td>
<td>0.8357</td>
</tr>
<tr>
<td>7</td>
<td>0.8554</td>
<td>0.9063</td>
</tr>
<tr>
<td>8</td>
<td>0.8791</td>
<td>0.8570</td>
</tr>
<tr>
<td>9</td>
<td>0.8069</td>
<td>0.8875</td>
</tr>
<tr>
<td>10</td>
<td>0.8852</td>
<td>0.8929</td>
</tr>
<tr>
<td>Sub mean</td>
<td>0.8737</td>
<td>0.8742</td>
</tr>
<tr>
<td>SD</td>
<td>0.0275</td>
<td>0.0209</td>
</tr>
</tbody>
</table>

*: Logistic regression accuracy was calculated based on 900 SNPs instead of the 1200 SNPs, whereas the accuracy of rrBLUP was calculated based on all the 1200 SNPs.

**IDC prediction results from algorithmic modeling, support vector machine (SVM)**

The overall mean prediction accuracy of SVM from the first repeated 10-folder cross-validation is 0.6710 (table 2), which is 23.22% lower than the baseline 0.8739 from rrBLUP.

When the selection sensitivity, discard specificity, selection precision, and overall accuracy were compared, big differences were observed between selection sensitivity and discard specificity. Selection sensitivity is 0.6075, which is the accuracy that true IDC tolerance lines can be identified without false susceptible lines. The 60.75% selection sensitivity also indicates that there are about 40% of the resistant lines would be incorrectly discarded. In contrast, discard specificity is 0.9665 and much bigger than the selection sensitivity 0.6075 (Figure 1), which indicates that we can safely discard the predicted susceptible IDC lines. The difference between the selection sensitivity and discard specificity t-test is highly significant, with a p-value of 1.077e-11. High specificity means that the ability to identify the true susceptible IDC lines is
much higher than that of identifying true resistant lines. High selection precision 0.9882 indicates that the false-positive chances are very low, and the true positive IDC resistance rate is high. The overall accuracy results from both specificity and precision show that SVM can correctly identify susceptible lines with a much higher rate than that of the identification of tolerant lines. These results were confirmed from the second repeat of the 10-fold cross-validation (Table 3). Discard specificity and selection precision from the second 10-fold cross-validation is 0.9713 and 0.9886, respectively, which are much higher than sensitivity, 0.6133 (p-value 1.01e-16). From the first and second repeats of 10-fold cross-validations, the results are: 1) SVM algorithm is not as good as rrBLUP in term of overall prediction accuracy; 2) but SVM has high discard specificity and can successfully predict susceptible IDC lines which are very useful to de-select IDC susceptible lines and exclude these lines in breeding early stage from subsequent resource-extensive yield testing.

Table 2. Selection sensitivity, discard specificity, selection precision, and overall accuracy from 10-fold cross-validation by SVM with 1000 lines by 1200 markers. Results from the first repeated 10-fold cross-validation

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold _01</td>
<td>0.6000</td>
<td>0.9500</td>
<td>0.9796</td>
<td>0.6700</td>
</tr>
<tr>
<td>Fold _02</td>
<td>0.5682</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6200</td>
</tr>
<tr>
<td>Fold _03</td>
<td>0.7432</td>
<td>0.9615</td>
<td>0.9821</td>
<td>0.8000</td>
</tr>
<tr>
<td>Fold _04</td>
<td>0.5765</td>
<td>0.8667</td>
<td>0.9608</td>
<td>0.6200</td>
</tr>
<tr>
<td>Fold _05</td>
<td>0.6220</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6900</td>
</tr>
<tr>
<td>Fold _06</td>
<td>0.5949</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6800</td>
</tr>
<tr>
<td>Fold _07</td>
<td>0.6296</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.7000</td>
</tr>
<tr>
<td>Fold _08</td>
<td>0.5233</td>
<td>0.9286</td>
<td>0.9783</td>
<td>0.5800</td>
</tr>
<tr>
<td>Fold _09</td>
<td>0.5465</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6100</td>
</tr>
<tr>
<td>Fold _10</td>
<td>0.6711</td>
<td>0.9583</td>
<td>0.9808</td>
<td>0.7400</td>
</tr>
<tr>
<td>Mean</td>
<td>0.6075</td>
<td>0.9665</td>
<td>0.9882</td>
<td>0.6710</td>
</tr>
<tr>
<td>SD</td>
<td>0.0639</td>
<td>0.0439</td>
<td>0.0138</td>
<td>0.0666</td>
</tr>
</tbody>
</table>
Table 3. Prediction sensitivity, specificity, precision, and overall accuracy from 10-fold cross-validation by SVM with 1000 lines by 1200 markers and IDC scores adjusted by spatial analysis, repeated second 10-fold cross-validation

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold _01</td>
<td>0.5402</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6000</td>
</tr>
<tr>
<td>Fold _02</td>
<td>0.6154</td>
<td>0.9545</td>
<td>0.9796</td>
<td>0.6900</td>
</tr>
<tr>
<td>Fold _03</td>
<td>0.6282</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.7100</td>
</tr>
<tr>
<td>Fold _04</td>
<td>0.5349</td>
<td>0.9286</td>
<td>0.9787</td>
<td>0.5900</td>
</tr>
<tr>
<td>Fold _05</td>
<td>0.6506</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.7100</td>
</tr>
<tr>
<td>Fold _06</td>
<td>0.5714</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6400</td>
</tr>
<tr>
<td>Fold _07</td>
<td>0.5823</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6700</td>
</tr>
<tr>
<td>Fold _08</td>
<td>0.6625</td>
<td>0.9500</td>
<td>0.9815</td>
<td>0.7200</td>
</tr>
<tr>
<td>Fold _09</td>
<td>0.7067</td>
<td>0.8800</td>
<td>0.9464</td>
<td>0.7500</td>
</tr>
<tr>
<td>Fold _10</td>
<td>0.6410</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.7200</td>
</tr>
<tr>
<td>Mean</td>
<td>0.6133</td>
<td>0.9713</td>
<td>0.9886</td>
<td>0.6800</td>
</tr>
<tr>
<td>SD</td>
<td>0.0555</td>
<td>0.0419</td>
<td>0.0176</td>
<td>0.0540</td>
</tr>
</tbody>
</table>

Figure 1. Boxplot of selection sensitivity, discard specificity, selection precision, and overall accuracy by SVM (Blue) and rrBLUP (green) from two repeats of 10-fold cross-validations. The four box plots from left to right are all from SVM, and the 5th box is from rrBLUP as the baseline for comparison.
IDC prediction results from algorithmic modeling, Naïve Bayes (NB)

Over prediction accuracy of NB is 0.8370 (Table 4), which is similar to that from rrBLUP, 0.8739 (Table 1). The prediction accuracy difference between NB and rrBLUP is not significant (p-value 0.06) But, in contrast to the results from SVM which has very high discard specificity but low selection sensitivity rates, Naïve Bayes has the lowest discard specificity 0.7405, but highest selection sensitivity 0.9357 with overall accuracy 0.8370 (Table 4 and Figure 2). The difference level between the discard specificities and selection sensitivities from the 10-fold cross-validation by t-test is extremely significant with a p-value of 8.519e-07.

Table 4 Selection sensitivity, discard specificity, selection precision, and overall accuracy from 10-fold cross-validation by Naïve Bayes with 1,000 lines by 1,200 markers.

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold_01</td>
<td>0.9388</td>
<td>0.7059</td>
<td>0.7541</td>
<td>0.8200</td>
</tr>
<tr>
<td>Fold_02</td>
<td>0.9200</td>
<td>0.8000</td>
<td>0.8214</td>
<td>0.8600</td>
</tr>
<tr>
<td>Fold_03</td>
<td>0.9643</td>
<td>0.7727</td>
<td>0.8438</td>
<td>0.8800</td>
</tr>
<tr>
<td>Fold_04</td>
<td>0.9020</td>
<td>0.7551</td>
<td>0.7931</td>
<td>0.8300</td>
</tr>
<tr>
<td>Fold_05</td>
<td>0.9412</td>
<td>0.7755</td>
<td>0.8136</td>
<td>0.8600</td>
</tr>
<tr>
<td>Fold_06</td>
<td>0.9574</td>
<td>0.6792</td>
<td>0.7258</td>
<td>0.8100</td>
</tr>
<tr>
<td>Fold_07</td>
<td>0.9804</td>
<td>0.7347</td>
<td>0.7937</td>
<td>0.8600</td>
</tr>
<tr>
<td>Fold_08</td>
<td>0.8913</td>
<td>0.6481</td>
<td>0.6833</td>
<td>0.7600</td>
</tr>
<tr>
<td>Fold_09</td>
<td>0.9574</td>
<td>0.6792</td>
<td>0.7258</td>
<td>0.8100</td>
</tr>
<tr>
<td>Fold_10</td>
<td>0.9038</td>
<td>0.8542</td>
<td>0.8704</td>
<td>0.8800</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>0.9357</strong></td>
<td><strong>0.7405</strong></td>
<td><strong>0.7825</strong></td>
<td><strong>0.8370</strong></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.0301</td>
<td>0.0634</td>
<td>0.0589</td>
<td>0.0380</td>
</tr>
</tbody>
</table>

These results indicate that Naïve Bayes can correctly predict the true tolerance IDC lines than predict the susceptible lines. If this trend is true for all data sets, by combining the results from SVM and Naïve Bayes, both true tolerant and susceptible IDC lines can be identified correctly, and use the selection sensitivity from NB to select the IDC resistant lines, and used the discard specificity from SVN to discard IDC susceptible lines.
Figure 2. Boxplot of selection sensitivity, discard specificity, selection precision, and overall accuracy by Naïve Bayes (Blue) and by rrBLUP (green) from two repeats of 10-fold cross-validations. The four box plots from left to right are all from Naïve Bayes, and the very right-hand boxplot is from rrBLUP as the baseline for comparison.

**IDC prediction results from algorithmic modeling, K-nearest neighbor (KNN)**

The overall mean prediction accuracy of 10-fold cross-validation from KNN is 0.8110 (table 5), which is about 7.2% lower than that of rrBLUP, 08739 (Table 1). The results of the t-test between the accuracy by rrBLUP indicated by green boxplot (Figure 3) and by KNN, the blue boxplot is significant with a p-value of 5.137e-05. The difference between selection sensitivity and discard specificity is not significant with p-value 0.08058. Neither reaches the significant different level alpha 0.05 between selection sensitivity and precision.

Table 5. Selection sensitivity, discard specificity, prediction precision, and overall accuracy from 10-fold cross-validation by KNN with 1000 lines by 1200 markers.

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold_01</td>
<td>0.8367</td>
<td>0.8039</td>
<td>0.8039</td>
<td>0.8200</td>
</tr>
<tr>
<td>Fold_02</td>
<td>0.8000</td>
<td>0.8800</td>
<td>0.8696</td>
<td>0.8400</td>
</tr>
<tr>
<td>Fold_03</td>
<td>0.7679</td>
<td>0.8864</td>
<td>0.8958</td>
<td>0.8200</td>
</tr>
<tr>
<td>Fold_04</td>
<td>0.7451</td>
<td>0.9388</td>
<td>0.9268</td>
<td>0.8400</td>
</tr>
<tr>
<td>Fold_05</td>
<td>0.8824</td>
<td>0.7347</td>
<td>0.7759</td>
<td>0.8100</td>
</tr>
<tr>
<td>Fold_06</td>
<td>0.8298</td>
<td>0.8302</td>
<td>0.8125</td>
<td>0.8300</td>
</tr>
<tr>
<td>Fold_07</td>
<td>0.8235</td>
<td>0.7959</td>
<td>0.8077</td>
<td>0.8100</td>
</tr>
<tr>
<td>Fold_08</td>
<td>0.7174</td>
<td>0.7963</td>
<td>0.7500</td>
<td>0.7600</td>
</tr>
<tr>
<td>Fold_09</td>
<td>0.7021</td>
<td>0.8491</td>
<td>0.8049</td>
<td>0.7800</td>
</tr>
<tr>
<td>Fold_10</td>
<td>0.7692</td>
<td>0.8333</td>
<td>0.8333</td>
<td>0.8000</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>0.7874</strong></td>
<td><strong>0.8349</strong></td>
<td><strong>0.8280</strong></td>
<td><strong>0.8110</strong></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.0571</td>
<td>0.0575</td>
<td>0.0545</td>
<td>0.0256</td>
</tr>
</tbody>
</table>
IDC prediction results from algorithmic modeling, artificial neural network (ANN)

The average selection sensitivity, discard specificity, selection precision, and overall accuracy by ANN from the 10-fold cross-validation are 0.9094, 0.8981, 0.8952, and 0.9020, respectively (Table 6). The significant difference tests among these four metrics are not significant with p-values > 0.05. But the difference of t-test is significant with a p-value of 0.0088 (<0.01) between the accuracy 0.902 by ANN and 0.8739 by rrBLUP from 10-fold cross-validation (Table 6 and Figure 4). Even though these two means from 10-fold cross-validation, 0.902 and 0.8739, are very close, their different level is significant. This is because the accuracy from the 10-fold cross-validation repeats very well and consistent across the 10-folds. These results indicate that model ANN performs very well across ten different random folds. The standard deviations are 0.0311, 0.0315, 0.0394, and 0.0199 for selection sensitivity, discard specificity, selection precision, and overall accuracy, respectively. These relatively small standard deviations indicate that the model is stable, and overfitting is unlikely for future prediction.
Table 6. Selection sensitivity, discard specificity, selection precision, and overall accuracy from 10-fold cross-validation by ANN with 1000 lines by 1200 markers.

<table>
<thead>
<tr>
<th>Cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold _01</td>
<td>0.9512</td>
<td>0.8814</td>
<td>0.8478</td>
<td>0.9100</td>
</tr>
<tr>
<td>Fold _02</td>
<td>0.8727</td>
<td>0.9111</td>
<td>0.9231</td>
<td>0.8900</td>
</tr>
<tr>
<td>Fold _03</td>
<td>0.9487</td>
<td>0.8852</td>
<td>0.8409</td>
<td>0.9100</td>
</tr>
<tr>
<td>Fold _04</td>
<td>0.8704</td>
<td>0.8696</td>
<td>0.8868</td>
<td>0.8700</td>
</tr>
<tr>
<td>Fold _05</td>
<td>0.8909</td>
<td>0.9333</td>
<td>0.9423</td>
<td>0.9100</td>
</tr>
<tr>
<td>Fold _06</td>
<td>0.9400</td>
<td>0.8800</td>
<td>0.8868</td>
<td>0.9100</td>
</tr>
<tr>
<td>Fold _07</td>
<td>0.9200</td>
<td>0.9600</td>
<td>0.9583</td>
<td>0.9400</td>
</tr>
<tr>
<td>Fold _08</td>
<td>0.9200</td>
<td>0.8600</td>
<td>0.8679</td>
<td>0.8900</td>
</tr>
<tr>
<td>Fold _09</td>
<td>0.9020</td>
<td>0.9184</td>
<td>0.9200</td>
<td>0.9100</td>
</tr>
<tr>
<td>Fold _10</td>
<td>0.8776</td>
<td>0.8824</td>
<td>0.8776</td>
<td>0.8800</td>
</tr>
<tr>
<td>Mean</td>
<td><strong>0.9094</strong></td>
<td><strong>0.8981</strong></td>
<td><strong>0.8952</strong></td>
<td><strong>0.9020</strong></td>
</tr>
<tr>
<td>SD</td>
<td>0.0311</td>
<td>0.0315</td>
<td>0.0394</td>
<td>0.0199</td>
</tr>
</tbody>
</table>

Figure 4. Boxplot of selection sensitivity, discard specificity, selection precision, and overall accuracy by ANN (Blue) and by rrBLUP (green) from two repeats of 10-fold cross-validations. The four box plots from left to right are from ANN, and the very right-hand boxplot is from rrBLUP. X-axes are the model names, and Y-axis are coefficient values.

**IDC prediction results from algorithmic modeling, Gradient Boosting Machine (GBM)**

Results from GBM are similar to those from ANN. The average selection sensitivity, discard specificity, selection precision, and overall accuracy by XGBoost-GBM from the 10-fold
cross-validation are 0.9290, 0.8904, 0.8944, and 0.9090, respectively (Table 7). The differences among these four metrics are not significant with p-values > 0.05. The difference of t-test is significant with a p-value of 0.0118 (<0.05) between accuracy by ANN and by rrBLUP from 10-fold cross-validation (Figure 5). Even though these two means from 10-fold cross-validation, 0.909 and 0.8739, are very close, their different level is moderately significant.

Table 7. Selection sensitivity, discard specificity, selection precision, and overall accuracy from 10-fold cross-validation by XGBoost_GBM with 1000 lines by 1200 markers.

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold_01</td>
<td>0.9796</td>
<td>0.9412</td>
<td>0.9412</td>
<td>0.9600</td>
</tr>
<tr>
<td>Fold_02</td>
<td>0.9400</td>
<td>0.9200</td>
<td>0.9216</td>
<td>0.9300</td>
</tr>
<tr>
<td>Fold_03</td>
<td>0.9107</td>
<td>0.8636</td>
<td>0.8947</td>
<td>0.8900</td>
</tr>
<tr>
<td>Fold_04</td>
<td>0.8431</td>
<td>0.9184</td>
<td>0.9149</td>
<td>0.8800</td>
</tr>
<tr>
<td>Fold_05</td>
<td>0.8627</td>
<td>0.8776</td>
<td>0.8800</td>
<td>0.8700</td>
</tr>
<tr>
<td>Fold_06</td>
<td>0.9574</td>
<td>0.8868</td>
<td>0.8824</td>
<td>0.9200</td>
</tr>
<tr>
<td>Fold_07</td>
<td>0.9020</td>
<td>0.8367</td>
<td>0.8519</td>
<td>0.8700</td>
</tr>
<tr>
<td>Fold_08</td>
<td>0.9565</td>
<td>0.8519</td>
<td>0.8462</td>
<td>0.9000</td>
</tr>
<tr>
<td>Fold_09</td>
<td>0.9574</td>
<td>0.8491</td>
<td>0.8491</td>
<td>0.9000</td>
</tr>
<tr>
<td>Fold_10</td>
<td>0.9808</td>
<td>0.9583</td>
<td>0.9623</td>
<td>0.9700</td>
</tr>
<tr>
<td>Mean</td>
<td><strong>0.9290</strong></td>
<td><strong>0.8904</strong></td>
<td><strong>0.8944</strong></td>
<td><strong>0.9090</strong></td>
</tr>
<tr>
<td>SD</td>
<td>0.0478</td>
<td>0.0419</td>
<td>0.0401</td>
<td>0.0354</td>
</tr>
</tbody>
</table>

Figure 5. Boxplot of selection sensitivity, discard specificity, selection precision, and overall prediction accuracy by GBM (Blue) and by rrBLUP (green). The four box plots from left to right are GBM, and the very right-hand boxplot is from rrBLUP. X-axes are the machine and deep learning algorithms; Y-axis is the accuracies from 0 to 1.
IDC prediction results from algorithmic modeling, random Forest (RF)

Overall results from RF is better than that from rrBLUP (Table 8). The mean selection sensitivity, discard specificity, selection precision, and overall prediction from 10-fold cross-validations are 0.9340, 0.9820, 0.9369, and 0.9580, respectively. The mean discard specificity is significantly larger than the selection sensitivity (p-value 9.88e-05). The p-value of the significant t-test between the overall accuracy of RF and rrBLUP is 1.344e-07. The small p-value indicated that the accuracy of RF is extremely significant and higher than that of rrBLUP. The standard deviation of overall accuracy from RF is 0.0132, which is much smaller than that of rrBLUP, 0.0241 (Table 1). Results from both overall accuracy and standard deviation of the accuracy from the 10-fold cross-validations indicate that RF outperforms rrBLUP.

![Boxplot of selection sensitivity, discard specificity, selection precision, and overall prediction accuracy by GBM (Blue) and by rrBLUP (green).](image)

**Figure 6.** Boxplot of selection sensitivity, discard specificity, selection precision, and overall prediction accuracy by GBM (Blue) and by rrBLUP (green). The four box plots from left to right are GBM, and the very right-hand boxplot is from rrBLUP. X-axes are the measurement parameters. Y-axes are the coefficients.
Table 8. Selection sensitivity, discard specificity, selection precision, and overall accuracy from 10-fold cross-validation by Random Forrest (RF) with 1000 lines by 1200 markers.

<table>
<thead>
<tr>
<th>cross-validation parameter</th>
<th>Selection sensitivity</th>
<th>Discard specificity</th>
<th>Selection precision</th>
<th>Overall prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold_01</td>
<td>0.9804</td>
<td>0.9184</td>
<td>0.9259</td>
<td>0.9500</td>
</tr>
<tr>
<td>Fold_02</td>
<td>0.9773</td>
<td>0.9464</td>
<td>0.9348</td>
<td>0.9600</td>
</tr>
<tr>
<td>Fold_03</td>
<td>1.0000</td>
<td>0.9231</td>
<td>0.9231</td>
<td>0.9600</td>
</tr>
<tr>
<td>Fold_04</td>
<td>1.0000</td>
<td>0.9583</td>
<td>0.9630</td>
<td>0.9800</td>
</tr>
<tr>
<td>Fold_05</td>
<td>0.9643</td>
<td>0.9773</td>
<td>0.9818</td>
<td>0.9700</td>
</tr>
<tr>
<td>Fold_06</td>
<td>0.9623</td>
<td>0.9149</td>
<td>0.9273</td>
<td>0.9400</td>
</tr>
<tr>
<td>Fold_07</td>
<td>1.0000</td>
<td>0.8936</td>
<td>0.9138</td>
<td>0.9500</td>
</tr>
<tr>
<td>Fold_08</td>
<td>1.0000</td>
<td>0.9388</td>
<td>0.9444</td>
<td>0.9700</td>
</tr>
<tr>
<td>Fold_09</td>
<td>0.9773</td>
<td>0.9464</td>
<td>0.9348</td>
<td>0.9600</td>
</tr>
<tr>
<td>Fold_10</td>
<td>0.9583</td>
<td>0.9231</td>
<td>0.9200</td>
<td>0.9400</td>
</tr>
<tr>
<td>Mean</td>
<td>0.9820</td>
<td>0.9340</td>
<td>0.9369</td>
<td>0.9580</td>
</tr>
<tr>
<td>SD</td>
<td>0.0170</td>
<td>0.0242</td>
<td>0.0210</td>
<td>0.0132</td>
</tr>
</tbody>
</table>

Summary from algorithmic modeling: SVM, NB, KNN, ANN, GBM, and RF

The combined results of the six algorithmic models can be summarized as follows: 1) The highest discard specificity is from SVM (0.9713; Table 3). SVM should be the first choice is to de-select susceptible lines to exclude them from further extensive yield testing to minimize wasting breeding resources. 2) The highest selection sensitivity is from the random forest (0.982; Table 6). If you have more predicted resistant than your breeding program needs, RF forest should be selected to predict the resistant lines for next round extensive testing; 3) the highest prediction accuracy of selecting IDC resistant lines with the risk of keeping the susceptible lines in the future breeding program, measured by selection precision, is 0.9886 from SVM. If you have limited number of predicted resistant lines to fit your IDC breeding program, Algorithm SVM should be selected to select the tolerant lines for further resource-extensive testing; 3) Overall accuracy from ANN, GBM, and RF are all higher than that from rrBLUP; 4) Results
from the width of the boxplots of each algorithm, ANN and RF have the smallest box width which indicates the model performance of ANN and RF are very stable across the two repeats of 10 folds cross-validations (Figure 5). In contrast, the width of the boxplot from SVM is the largest which shows that the accuracy from SVM across the 10 -fold cross-validations varies the most and model SVM is not stable; 5) Logistic regression has the lowest accuracy from selection sensitivity, discard specificity, selection precision, and overall accuracy. Overall accuracies of ANN, GBM, and RF, 0.902, 0.909, 0.958, are significantly higher than that of the baseline rrBLUP, 0.8739, by 3.12%, 3.89%, and 9.32%, respectively. The selection sensitivity from SVM, discard specificity from RF, and selection precision from SVM, 0.982, 0.9713, 0.9886, are extremely higher than that of the overall accuracy of rrBLUP, 0.8739, by 12.37%, 11.15%, and 13.13% respectively (Table 1, 2, 8 and figure 7). The accuracy of logistic regression is extremely significantly lower than the six algorithmic models with the p-values all less than 0.01 (figure 7). Overall, the algorithmic models outperform the data models in regard to selection sensitivity, discard specificity, selection precision, and overall accuracy.

Figure 7. Summary of the comparisons among eight models. From left to right in order: overall accuracy of the baseline rrBLUP, logistic regression, SVM, KNN, NB, ANN, GBM, RF, selection sensitivity of RF, discard specificity by SVM and selection precision from SVM. X-axes are the models; Y-axis is the coefficients of the efficient. Black dots are the outliers, and the numbers above each line are the p-values of the significant pairwise tests. The value at the top of this figure is the overall p-value from the Kruskal-Wallis test.
Summary of the model performance measured by AUC from both algorithmic and data models

Table 9. The area under the curve (AUC) of the receiver operating characteristics (ROC) curves

<table>
<thead>
<tr>
<th>Folds</th>
<th>Logistic</th>
<th>SVM</th>
<th>NB</th>
<th>KNN</th>
<th>ANN</th>
<th>GBM</th>
<th>RF</th>
<th>rrBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0.55</td>
<td>0.91</td>
<td>0.89</td>
<td>0.52</td>
<td>0.94</td>
<td>0.98</td>
<td>0.98</td>
<td>0.81</td>
</tr>
<tr>
<td>02</td>
<td>0.5</td>
<td>0.89</td>
<td>0.93</td>
<td>0.48</td>
<td>0.95</td>
<td>0.93</td>
<td>0.99</td>
<td>0.69</td>
</tr>
<tr>
<td>03</td>
<td>0.53</td>
<td>0.84</td>
<td>0.92</td>
<td>0.52</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
<td>0.72</td>
</tr>
<tr>
<td>04</td>
<td>0.51</td>
<td>0.86</td>
<td>0.91</td>
<td>0.36</td>
<td>0.95</td>
<td>0.95</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td>05</td>
<td>0.55</td>
<td>0.88</td>
<td>0.91</td>
<td>0.44</td>
<td>0.92</td>
<td>0.96</td>
<td>0.99</td>
<td>0.82</td>
</tr>
<tr>
<td>06</td>
<td>0.47</td>
<td>0.83</td>
<td>0.82</td>
<td>0.40</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
<td>0.71</td>
</tr>
<tr>
<td>07</td>
<td>0.55</td>
<td>0.91</td>
<td>0.92</td>
<td>0.52</td>
<td>0.91</td>
<td>0.97</td>
<td>0.99</td>
<td>0.79</td>
</tr>
<tr>
<td>08</td>
<td>0.44</td>
<td>0.81</td>
<td>0.91</td>
<td>0.57</td>
<td>0.94</td>
<td>0.94</td>
<td>0.99</td>
<td>0.74</td>
</tr>
<tr>
<td>09</td>
<td>0.52</td>
<td>0.91</td>
<td>0.91</td>
<td>0.44</td>
<td>0.93</td>
<td>0.95</td>
<td>0.98</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>0.52</td>
<td>0.87</td>
<td>0.85</td>
<td>0.45</td>
<td>0.91</td>
<td>0.94</td>
<td>0.95</td>
<td>0.82</td>
</tr>
<tr>
<td>mean</td>
<td>0.51</td>
<td>0.87</td>
<td>0.9</td>
<td>0.47</td>
<td>0.93</td>
<td>0.95</td>
<td>0.98</td>
<td>0.76</td>
</tr>
<tr>
<td>SD</td>
<td>0.036</td>
<td>0.036</td>
<td>0.035</td>
<td>0.064</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Results from the AUC values of the 10-fold cross-validation (Table 9) and ROC curves from all the data of each model show that random forest is best among the eight models. Random forest is the most stable model from its standard deviation of 0.014 (Table 9). ANN, GBM, and random forest are very similar in regard to both ROC and standard deviation. In contrast, KNN and logistic regression have very small ROC values indicate both models are sensitive to the threshold of probability to separate the IDC resistant vs. susceptible lines (Figure 8). AUC of SVM is lower than that of random forest. This observation is consistent with an overall low accuracy but higher specificity from SVM. KNN is the worst model from both the lowest AUC value and the highest standard deviation, 0.064.
Figure 8. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve from eight models. The X-axis is a false positive rate (FPR); the y-axis is the true positive rate (TPR). On the top of each image, the first part before the colon is the model name, and the number after the colon is the AUC values, which are from 0 to 1. The numbers on the right of each image are the cutoffs used to calculate the false and true positive rates.

Results from the significant pairwise t-test of the AUC values show that logistic regression and KNN are statistically the same accuracies (p-value 0.088, figure 9), and they are the worst among the eight models. All the algorithmic models, SVM, NB, ANN, GBM, and RF, have higher AUC values than that of rrBLUP and reach extremely significant different levels (p-value < 0.001) from rrBLUP and logistic regression except KNN. Overall, the algorithmic models outperform the data models in regard to the AUCs.
Figure 9. Summary of pairwise significant and overall Kruskal-Wallis test results. The values above horizontal lines are the p-values of the significant pairwise test. The value on the top of the image is the p-values of the overall Kruskal-Wallis test. The highlighted horizontal lines inside each box represent the means. The black dots are the outliers from different models. X-axes are the models. Y-axis is the AUC values.

Models separation by principal component analysis (PCA)

Results from PCA analysis using the IDC selection sensitivity, discard specificity, selection precision, prediction overall accuracy, and AUC show that SVM is KNN separated in two regions as the first and second groups (Figure 10). Logistic regression and rrBLUP are spatially located near each other and clustered as the third group. When the figure is zoomed in, NB is separated from ANN, GBM, and RF and clustered as the fourth group. ANN, GBM, and RF are the three best models and clustered as the fifth group. Overall, the models are differentiated based on their prediction capability for IDC resistance scores.
Algorithmic vs. data modeling: Multiple comparisons and groups

Results from the non-parametric Kruskal-Wallis test show that objective-specific selection sensitivity, discard specificity, selection precision for IDC prediction are the best “A” group, and they are significantly better than that from the overall accuracy from both algorithmic and data models (Table 10). With these observations, for IDC resistant breeding, different models should be applied for selecting resistant and de-selecting susceptible lines. Random forest ranks the best model with combined accuracy and AUC values. ANN and GBM rank the third-best models as group “c.” In contrast, the baseline data model, rrBLUP, ranks the 5th, which is the same as SVM clustered as the group as “e.” Logistic regression ranks the worst in the multiple comparisons and group analysis. Comparing with rrBLUP, selection precision from SVM and discard specificity from RF are 20.76% and 19.9% better than that of rrBLUP,
respectively. Overall, the prediction accuracy of algorithmic modeling is better than that of data modeling in regard to IDC resistant prediction.

Table 10. Kruskal Wallis test and group results from both AUC and accuracy

<table>
<thead>
<tr>
<th>Models Names</th>
<th>Overall mean *</th>
<th>% increase to rrBLUP</th>
<th>Kruskal ranks</th>
<th>Groups (α=0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision_SVM</td>
<td>0.989</td>
<td>20.76</td>
<td>172.3</td>
<td>a</td>
</tr>
<tr>
<td>Sensitivity_RF</td>
<td>0.982</td>
<td>19.90</td>
<td>166.6</td>
<td>ab</td>
</tr>
<tr>
<td>Specificity_SVM</td>
<td>0.971</td>
<td>18.56</td>
<td>158.3</td>
<td>ab</td>
</tr>
<tr>
<td>RF</td>
<td>0.969</td>
<td>18.32</td>
<td>154.9</td>
<td>b</td>
</tr>
<tr>
<td>GBM</td>
<td>0.931</td>
<td>13.68</td>
<td>125.1</td>
<td>c</td>
</tr>
<tr>
<td>ANN</td>
<td>0.918</td>
<td>12.09</td>
<td>114.2</td>
<td>c</td>
</tr>
<tr>
<td>NB</td>
<td>0.866</td>
<td>5.74</td>
<td>84.5</td>
<td>d</td>
</tr>
<tr>
<td>rrBLUP</td>
<td>0.819</td>
<td>0.00</td>
<td>66.4</td>
<td>e</td>
</tr>
<tr>
<td>SVM</td>
<td>0.771</td>
<td>-5.86</td>
<td>61.2</td>
<td>e</td>
</tr>
<tr>
<td>KNN</td>
<td>0.641</td>
<td>-21.73</td>
<td>33.8</td>
<td>f</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.537</td>
<td>-34.43</td>
<td>18.7</td>
<td>g</td>
</tr>
</tbody>
</table>

Overall mean*: Overall mean of both area under the curve (AUC) & Accuracy*: refers to the average of the 10 AUC and ten accuracies from each of the models.

**Conclusion**

Among the eight tested models, SVM generated the highest discard specificity, and was the best model for de-selecting IDC susceptible lines, while RF had the highest selection sensitivity, and was the best model for selecting IDC resistant lines. RF has the overall highest IDC prediction accuracy and the highest AUC.

In an overall comparison for the soybean IDC ordinal data, the algorithmic modeling approaches outperformed the data modeling approaches, consistent with Breiman’s conclusions (2001) about the two statistical modeling cultures.

**Conflict of interest**

The authors declare that there is no conflict of interest.
Supplemental data available

Supplemental material is available online for this article. R codes and the data sets used for this research are available online.

Acknowledgments

We thank Syngenta and Iowa State University for the financial support of this research. This research was supported in part by the US. Department of Agriculture, Agricultural Research Service, project 5030-21000-062-00D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References


Anselin, L., R. J. G. M. Florax and SpringerLink (Online service), 1995 New Directions in Spatial Econometrics, pp. 1 online resource. in Advances in Spatial Science,. Springer Berlin Heidelberg,, Berlin, Heidelberg.
Anselin, L., S. J. Rey and SpringerLink (Online service), 2010 Perspectives on Spatial Data Analysis, pp. 1 online resource. in Advances in Spatial Science, The Regional Science Series,. Springer-Verlag Berlin Heidelberg,, Berlin, Heidelberg.


Bernardo, R., 2008 Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Science 48: 1649-1664.


Dormann, C. F., J. M. McPherson, M. B. Araujo, R. Bivand, J. Bolliger et al., 2007a Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609-628.


Eftekhar, B., K. Mohammad, H. E. Ardebili, M. Ghodsi and E. Ketabchi, 2005 Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC medical informatics and decision making 5: 3-3.


James, G., 2013 An Introduction to Statistical Learning with Applications in R / by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. New York, NY : Springer New York : Imprint: Springer.


Lado, B., I. Matus, A. Rodriguez, L. Inostroza, J. Poland et al., 2013b Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 (Bethesda) 3: 2105-2114.


Li, C., Y. Li, J. Xu, J. Lv, Y. Ma et al., 2011a Disease-driven detection of differential inherited SNP modules from SNP network. Gene 489: 119-129.


Miao, X., and J. S. Heaton, 2010 A Comparison of Random Forest and Adaboost Tree in Ecosystem Classification in East Mojave Desert. 2010 18th International Conference on Geoinformatics.


Muller, B. U., A. Schutzenmeister and H. P. Piepho, 2010 Arrangement of check plots in augmented block designs when spatial analysis is used. Plant Breeding 129: 581-589.


Pepe, M. S., T. Cai and G. Longton, 2006 Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve. Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve 62: 221-229.


Technow, F., 2015a R package mvngGrAd: moving grid adjustment in plant breeding field trials.

Technow, F., 2015b R package mvngGrAd: moving grid adjustment in plant breeding field trials. R package version 0.1.5.: 1-12.


Teo, Y. Y., 2008 Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 19: 133-143.


CHAPTER 5. MACHINE LEARNING FOR SOYBEAN SDS ANALYSIS

Zhanyou Xu¹, Keting Chen¹, Andreomar Kurek², Steven B. Cannon³, Williams D. Beavis²*

¹Bioinformatics and Computational Biology, Iowa State University, Ames Iowa 50014
²Department of Agronomy, Iowa State University Ames, Iowa 50010.
³USDA-Agricultural Research Service, Corn Insects, and Crop Genetics Research Unit, Ames, IA 50010

Modified from a manuscript to be submitted to Scientific Reports

Abstract

Sudden death syndrome (SDS) is a serious threat to soybean production that can be managed with host plant resistance. A total of 32 QTLs have been mapped from different mapping populations at different labs for marker assistant selection (MAS) breeding. However, QTL-based SDS breeding was not successful because SDS was a complicated trait and controlled by many QTLs in the genome. These QTLs do not function independently, and they may coordinate and function as a collaborative network. The objective of this research is to build a soybean QTL network that coordinates SDS as an integrated functional unit. The research was conducted in the following three steps: first, to identify SDS QTLs by a combination of three methods, including genome-wide association study (GWAS), stepwise regression, and random forest (classification or regression?). A list of 88 significant markers overlapped from the three methods was identified. Second, an SNP-by-SNP interaction network was built from these 88SNPs, and 73 of the 88SNPs interacted with each and were added as the nodes of the network; third, a protein-by-protein interaction (PPI) network was converted from the DNA level network, and 48 of the 73 SNP sequences have corresponding protein hit under e-value at E-10; lastly, the PPI network was aligned against Arabidopsis PPI network. Our results show that the identified
32 QTLs co-control SDS as a network and this network has similarity as the network from Arabidopsis. This network-based MAS may provide a novel way for breeding complex traits, such as quantitative controlled disease and grain yield.

**Introduction**

Soybean Sudden death syndrome (SDS) of soybean [Glycine max (L.) Merr.] is caused by Fusarium virguliforme, a soil-borne fungal pathogen that infects and colonizes the soybean root and causes root rot (Roy et al. 1997). SDS spread rapidly and widely across the major soybean-producing regions in the United States since the first observation of SDS on soybean in Arkansas in 1971 (Hartman et al. 2015). Annually yield losses were estimated at $120 million and ~330 million in 2009 (Koenning and Wrather 2010). Developing resistant cultivars with marker-assisted selection is believed to be the most effective method for controlling SDS. In the past decades, many SDS QTL has been mapped through bi-parental, GWAS association study (Achenbach et al. 1996; Chang et al. 1997; Luckew et al. 2013; Wen et al. 2014a; Zhang et al. 2015).

Even though many SDS tolerant QTLs have been mapped, SDS QTL-based selection was not successfully used for SDS resistant variety breeding. Several factors affect the usage of the mapped QTLs: 1) no SDS QTL has been proved to work across different populations or genetic backgrounds. QTLs were mapped in bi-parental populations, and these QTLs were population-specific and did not work in other genetic backgrounds. From the testing the usefulness of 10 SDS QTLs across six populations, no single QTL works across these six populations (Luckew et al. 2013); 2) no stable major SDS QTL across different genetic backgrounds has been identified. A total of 17 QTLs were mapped from bi-parental mapping populations, and their R2 values range from 2% to 63% with small population sizes ranged from 40 to 80 individual lines per population (Meksem et al. 1999), but there is no major SDS QTL has stable and consistent QTL
effects across different mapping populations (Wen et al. 2014b) 3) impractical to stack multiple QTLs. Since SDS QTLs are quantitative and their effects less than 10%, breeding SDS commercial varieties need to stack multiple QTLs. Soybean is self-pollinated cleistogamous (pollinated before the flower opens) crop, with current pollination technology, it would be impractical for soybean breeders to pyramid more than 6 QTLs from different populations into a single genetic background because the unusually large population is needed (Luckew et al. 2013); 4) both bi-parental mapping and GWA study in soybean mapped only the additive gene/QTL effects via single locus significant test or interval mapping, and epistasis and other gene interaction effects in soybean breeding have not to be reported by now. QTL- breeding either based on individual SNP selection or combination of SNPs as haplotype assistant selection. However, additive-effect only based selection may not be sufficient to explain the complexity of disease causality. It has been established that gene-gene/SNP-SNP interactions may have a higher impact on discovering the causality of complex human disease (Lin et al. 2013; Reams et al. 2011). There may exist two levels of interaction: first level interaction is pairwise interaction between QTLs/genes; the second level of interaction is three or higher dimensional interaction among all these QTLs, which may coordinate and cross-regulate each other in a network and function as one integrated unit. When only partial QTLs of the network were tracked and traced in MAS selection, the critical or bottleneck part of the network may be monomorphic and block the pathway; this may lead to the failure of QTL-based breeding for the complexity of the trait of SDS.

**Interaction network of genes associated with a trait of interest**

New studies of the network of genes have been publishing to understand human diseases and may be applied to soybean SDS breeding. SNP-by-SNP interaction network study has been
reported to dissect human disease. A synergetic SNP-by-SNP network of alcoholic addiction was constructed by two steps: step 1 to form the SNP-by-SNP interaction network based on prior biological knowledge and their correlation between the functional relationships of their genes; step 2 to prioritize disease-risk SNPs via their differentially inherited properties in identity by descent (IBD) (Li et al. 2011a). A weighted SNP interaction hub network was constructed for human complex diseases and traits using whole genome genotypic data (Kogelman and Kadarmideen 2014). By combing two powerful machine learning methods: random forest (RF) and multivariate adaptive regression spline (MARS), an SNP-by-SNP interaction network associated with prostate cancer was constructed (Lin et al. 2013; Lin et al. 2012). All these reports from the study of human diseases show that SNP-by-SNP network-based prediction provides a promising opportunity to increase prediction accuracy. The objectives of this study were:

1) to identify the overlapped SDS QTL and genes by stepwise regression, GWAS, and RF
2) to identify the interaction QTLs and genes by Multivariate adaptive regression splines
3) to build a DNA interaction network of SDS QTLs and genes
4) to align the soybean SDS network with Arabidopsis network for function comparison

**Materials and analytic method**

**Materials**

Three SDS genotypic and phenotypic data sets were used for this study: two publicly available data sets and one simulation data set. The first public data set is from USDA Germplasm Resources Information Network (GRIN), and a total of 6,851 production introduction (PI) lines have phenotypic data, and 6,774 of them were genotypic data from 50k chip downloaded from soyBase (http://www.soybase.org/dlpages/). The second public data set is supplemental data of a report (ZHANG et al. 2015) and downloaded from the “Plant Journal
The simulated data set has a total of 1,400 lines with SDS scores and 3,000 SNPs with both epistasis and additive effects in the simulation model.

**Analytic Method**

Overlapped significant SNPs were selected by conducting a combination of three statistical analysis, including random forest (Chen and Liu 2005; Lin et al. 2012), stepwise regression (JMP version 11, SAS Inc., NC 275134 USA) (Hosmer et al. 1978), GWAS (Han and Zhang 2011; Li et al. 2011b; Hwang et al. 2014) by TASSEL 5 (Bradbury et al. 2007), and

**Model and parameter selection for Random Forest analysis**

For the random forest analysis, r package “randomForest” was used. Build-in function “tuneRF” was conducted to tune randomForest for the optimal “mtry” parameter, which is how many numbers of random forest analyses/tries for best prediction accuracy. The parameters to search the best “mtry” are listed as the following:

```r
bestmtry <- tuneRF(SNP-matrix, IDC-score, ntreeTry=100, stepFactor=1.5, improve=0.01, trace=TRUE, plot=TRUE, dobest=FALSE)
```

The random forest model, “randomForest.model”, was fitted as follows:

```r
randomForest.model <- randomForest(SNP-matrix, IDC-score, mtry=2, ntree=5000, keepForest=TRUE, importance=TRUE)
```

The most significant variables were identified by two parameters: %IncMSE and IncNodePurity. %IncMSE is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model; IncNodePurity is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all decision trees.
Model and parameter selection for stepwise regression analysis

Both forward and backward stepwise regressions were applied with the following stopping rules: 1) Minimum of Akaike information criterion (AIC); 2) minimum Bayesian information criterion (BIC); 3) and maximum R2. Since each SNP has three levels (AA, Aa, and aa), all the markers were treated as categorical variables, and “whole effects” was used in the JMP Pro V12 under marker selections rule.

http://www.jmp.com/support/help/The_Stepwise_Report.shtml. All markers were treated as fixed effects.

Model and parameter selection for GWAS analysis by Tassel

![Flowchart of the project](image)

Figure 1. Flowchart of the research. TASSEL: Trait Analysis by Association, Evolution, and Linkage; GWAS: Genome-Wide Association study; SNP: Single nucleotide polymorphism; JMP is a computer program for statistics developed by the JMP business unit of SAS Institute.
Networks were built by three methods: SNP-SNP interaction network was conducted using algorithm Multivariate Adaptive Regression Splines (MARS) (Lin et al. 2012) by R package “earth” (Put et al. 2004); SNP-SNP correlation network was build up based on the correlation coefficients calculated by “spearman” in R; Protein-protein interaction network was built using Matching-based Integrative GRAPh Aligner (MI-GRAAL) (Kuchaiev and Przulj 2011). Networks were visualized via Cytoscape (Kilcoyne et al. 2009). The procedures of this research were summarized in figure 1.

Results and discussions

Predictive variable selection results from random forest analysis

The most important markers based on both %IncMSE and IncNodePurity were identified, and the top 10 markers were summarized as table 1 (the entire list of the significantly associated markers were in supplemental table1) and Figure 2. When a different number of decision trees were used for the predictive variable selection, the absolute values of the %IncMSE and IncNodePurity for each of the markers are different, but the ranking of the importance of these SNPs are relatively stable. For example, marker, MK9001870, have 19.2826 and 28.2777 from ntree=5,000 and ntree=10,000, respectively, but they are the number 1 marker from both analyses. Some of the markers, such as number 10 marker, SNP0770A, it has both different values of %IncMSE (13.0985 and 19.6492) and different rankings (10 and 7) from the two analyses. Overall, the top 10 markers from both analyses are the same.

Table 1. A subset of SNPs significantly associated with SDS tolerance from random forest variable importance analysis

<table>
<thead>
<tr>
<th>SNP</th>
<th>%IncMSE ntree5000</th>
<th>IncNodePurity ntree5000</th>
<th>%IncMSE ntree10000</th>
<th>IncNodePurity ntree10000</th>
<th>Rank by ntree5000</th>
<th>Rank by ntree10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001870</td>
<td>19.2826</td>
<td>5.4339</td>
<td>28.2777</td>
<td>5.5924</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MK9001554</td>
<td>17.8890</td>
<td>8.1157</td>
<td>23.4339</td>
<td>7.3949</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>MK9001649</td>
<td>17.8074</td>
<td>5.8266</td>
<td>25.8305</td>
<td>6.0226</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>MK9001383</td>
<td>17.0462</td>
<td>8.0559</td>
<td>23.8517</td>
<td>7.6102</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 1 Continued

<table>
<thead>
<tr>
<th>SNP</th>
<th>%IncMSE ntree5000</th>
<th>IncNodePurity ntree5000</th>
<th>%IncMSE ntree10000</th>
<th>IncNodePurity ntree10000</th>
<th>Rank by ntree5000</th>
<th>Rank by ntree10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001009</td>
<td>16.4826</td>
<td>4.0653</td>
<td>21.7691</td>
<td>3.9425</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>MK9001683</td>
<td>15.7043</td>
<td>5.2855</td>
<td>21.4679</td>
<td>4.7095</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>MK9000892</td>
<td>13.6430</td>
<td>2.5712</td>
<td>17.9750</td>
<td>2.5339</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>MK9002129</td>
<td>13.5826</td>
<td>7.8892</td>
<td>17.7629</td>
<td>7.9651</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>MK9001520</td>
<td>13.4302</td>
<td>4.0816</td>
<td>19.5529</td>
<td>4.1909</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>MK9000757</td>
<td>13.0985</td>
<td>4.8219</td>
<td>19.6492</td>
<td>4.9430</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 2. The Variable importance plot from “varImpPlot(model.rf)” show the graphical representation of the selected predictive variables

Predictive variable selection results from stepwise regression by JMP 12 Pro

A total of 498 SNPs were selected by JMP when minimum AIC was used as a stopping rule for variable selection. In contrast, only 191 SNPs were selected when minimum BIC was used as a stopping rule for model optimization. The top 10 SNPs selected from AIC and BIC screening were summarized in table 5.2. From the side-by-side comparison of the SNPs selected between AIC and BIC, the majority of the SNPs have similar p-value, the sum of the square, and R2. Such as the top 3 SNPs, MK9001383, MK9001520, and MK9001979, have very similar
values and ranked the top 3 in both AIC and BIC screening processes. From the table below, one of the 10 SNPs, MK9001253, it was more important from AIC than that from BIC.

Table 2 A subset of SNPs significantly associated with SDS tolerance from stepwise regression analysis (table sorted by R2 from AIC analysis)

<table>
<thead>
<tr>
<th>SNP Name</th>
<th>P-value AIC</th>
<th>SS AIC</th>
<th>R2 AIC</th>
<th>AIC Value</th>
<th>P-value BIC</th>
<th>SS BIC</th>
<th>R2B IC</th>
<th>BIC value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001383</td>
<td>0</td>
<td>52.78</td>
<td>0.07</td>
<td>2932.85</td>
<td>0</td>
<td>45.89</td>
<td>0.06</td>
<td>2957.55</td>
</tr>
<tr>
<td>MK9001520</td>
<td>0</td>
<td>29.65</td>
<td>0.11</td>
<td>2880.87</td>
<td>0</td>
<td>29.56</td>
<td>0.10</td>
<td>2907.46</td>
</tr>
<tr>
<td>MK9001979</td>
<td>0</td>
<td>21.55</td>
<td>0.14</td>
<td>2843.13</td>
<td>0</td>
<td>21.61</td>
<td>0.13</td>
<td>2871.18</td>
</tr>
<tr>
<td>MK9001253</td>
<td>0</td>
<td>16.60</td>
<td>0.16</td>
<td>2814.47</td>
<td>0.0106</td>
<td>1.36</td>
<td>0.65</td>
<td>2224.36</td>
</tr>
<tr>
<td>MK9000831</td>
<td>0</td>
<td>16.82</td>
<td>0.18</td>
<td>2784.43</td>
<td>0</td>
<td>14.19</td>
<td>0.17</td>
<td>2822.63</td>
</tr>
<tr>
<td>MK9000322</td>
<td>0</td>
<td>17.42</td>
<td>0.21</td>
<td>2752.08</td>
<td>0</td>
<td>17.65</td>
<td>0.19</td>
<td>2791.44</td>
</tr>
<tr>
<td>MK9001266</td>
<td>0</td>
<td>14.62</td>
<td>0.23</td>
<td>2725.07</td>
<td>0</td>
<td>9.26</td>
<td>0.40</td>
<td>2506.25</td>
</tr>
<tr>
<td>MK9000005</td>
<td>0</td>
<td>15.80</td>
<td>0.27</td>
<td>2663.29</td>
<td>0</td>
<td>13.35</td>
<td>0.21</td>
<td>2768.85</td>
</tr>
<tr>
<td>MK9001379</td>
<td>0</td>
<td>13.84</td>
<td>0.29</td>
<td>2635.48</td>
<td>0</td>
<td>6.48</td>
<td>0.39</td>
<td>2526.35</td>
</tr>
<tr>
<td>MK9001878</td>
<td>0</td>
<td>15.04</td>
<td>0.31</td>
<td>2603.74</td>
<td>0.0071</td>
<td>1.25</td>
<td>0.72</td>
<td>2211.34</td>
</tr>
</tbody>
</table>

**Predictive variable selection results from GWAS by Tassel 5**

A total of 372 and 158 markers were screened out from the two models, general linear model (GLM) and mixed linear model (MLM), respectively, when a loose threshold $p<0.05$ was used. The number of overlapped markers that were identified by both GLM and MLM was 147, which 93% (147/158*100) of the markers identified from MLM were also screened out from GLM. Comparing the Manhattan and quantile-quantile plots (Figure 3) from the two models, the $-\log P$ values from MLM is smaller than that from GLM as we expected because of the additional kinship and population structure were added in the mixed model. The significantly associated markers from the two models were summarized in supplemental table S5.2. A subset of markers (top 10 based on the R2 from GLM) significant associated with SDS was list in table 5.3.

Another observation is that some of the overlapped markers identified by GLM and MLM have different ranks, such as MK9000740, it was ranked as #2 important SNP, in contrast, it was ranked as #12 important SNP (second row in table 3).
Table 3 A subset of SNPs significantly associated with SDS tolerance from GWAS (table sorted by chromosome first and physical map position second)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Pos</th>
<th>p-value GLM</th>
<th>Marker R2 GLM</th>
<th>Rank R2 GLM</th>
<th>p-value MLM</th>
<th>Marker R2 MLM</th>
<th>Rank R2 MLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001266</td>
<td>99</td>
<td>0</td>
<td>1.47E-08</td>
<td>0.04746</td>
<td>1</td>
<td>5.26E-06</td>
<td>0.03250</td>
<td>1</td>
</tr>
<tr>
<td>MK9000740</td>
<td>99</td>
<td>0</td>
<td>5.99E-09</td>
<td>0.03531</td>
<td>2</td>
<td>2.86E-04</td>
<td>0.01505</td>
<td>12</td>
</tr>
<tr>
<td>MK9000024</td>
<td>99</td>
<td>0</td>
<td>1.23E-08</td>
<td>0.03475</td>
<td>3</td>
<td>4.99E-04</td>
<td>0.01492</td>
<td>14</td>
</tr>
<tr>
<td>MK9001520</td>
<td>06</td>
<td>9</td>
<td>6.10E-09</td>
<td>0.03331</td>
<td>4</td>
<td>1.40E-06</td>
<td>0.02405</td>
<td>3</td>
</tr>
<tr>
<td>MK9000252</td>
<td>14</td>
<td>34</td>
<td>1.47E-09</td>
<td>0.03308</td>
<td>5</td>
<td>1.95E-04</td>
<td>0.01374</td>
<td>18</td>
</tr>
<tr>
<td>MK9001009</td>
<td>14</td>
<td>34</td>
<td>2.54E-09</td>
<td>0.03107</td>
<td>6</td>
<td>2.49E-04</td>
<td>0.01271</td>
<td>20</td>
</tr>
<tr>
<td>MK9001858</td>
<td>14</td>
<td>43</td>
<td>1.61E-08</td>
<td>0.03093</td>
<td>7</td>
<td>1.05E-04</td>
<td>0.01575</td>
<td>11</td>
</tr>
<tr>
<td>MK9001178</td>
<td>14</td>
<td>43</td>
<td>1.64E-08</td>
<td>0.02962</td>
<td>8</td>
<td>7.42E-05</td>
<td>0.01619</td>
<td>9</td>
</tr>
<tr>
<td>MK9000915</td>
<td>99</td>
<td>0</td>
<td>5.85E-08</td>
<td>0.02935</td>
<td>9</td>
<td>1.16E-03</td>
<td>0.01169</td>
<td>22</td>
</tr>
<tr>
<td>MK9001554</td>
<td>06</td>
<td>48</td>
<td>7.91E-08</td>
<td>0.02814</td>
<td>10</td>
<td>3.50E-06</td>
<td>0.02150</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 3. Manhattan and Q-Q plot from GWAS. The left-hand plot is the Manhattan plot of the soybean 20 Chromosomes; the right-hand plot is the Q-Q plot.

Comparison of marker selection result among the three methods: random forest, stepwise regression, and GWAS

The importance of markers was ranked based on marker effect or R2 from each of the three methods, and comparison results were summarized in supplemental Table 5.4. Top 10 of the SNPs based on stepwise regression, and the other two methods were list in table 5.4. Such as marker, MK9001383, was ranked as the most important SNP from both stepwise regression and random forest, but not identified by GWAS. The reason that MK9001383 was not screened out by GWAS is that this marker has minimum allele frequency (MAF) <0.05 and was excluded.
from GWAS analysis. When this marker was included in GWAS analysis regardless of restriction of MAF, it was screened as the most important SNP, too. From the top 10 SNPs, 6 of them (MK9001520, MK9000448, MK9000631, MK9001266, MK9001878, MK9001379) were screened out by all three methods indicating these SNPs are important. Rankings among the three methods of the 6 SNPs are different, such as MK9001520, it was ranked as #2, #4, #2 by stepwise regression, GWAS, and random forest, respectively. The three rankings for this marker from the three methods are very similar. In contrast, marker, MK9000448, was ranked as #3, #155, and #54 by the three methods, respectively, and their rankings are very different. The number of overlapped SNPs among the three methods was counted and summarized in the last three columns of table 5.4. Three overlapped numbers of SNPs of the three pairwise methods are 241 between random forest and GWAS, 132 between random forest and stepwise regression (SWR), and 117 between SWR and GWAS. GWAS and RF identified many more SNPs (241) than that of SWR and GWAS (117), from the three comparisons.

Table 4. Comparison of the SNPs selected from stepwise regression (SWQ), Random forest (RF), and GWAS.

<table>
<thead>
<tr>
<th>SNP Name</th>
<th>Rank by SWR</th>
<th>Rank by GWAS</th>
<th>Rank by RF</th>
<th>SWR &amp; RF (y/n, 132)</th>
<th>SWR &amp; GWAS (y/n, 117)</th>
<th>GWAS &amp; RF (y/n, 241)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001383</td>
<td>1</td>
<td>NA</td>
<td>1</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>MK9001520</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>MK9000448</td>
<td>3</td>
<td>155</td>
<td>54</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>MK9000631</td>
<td>4</td>
<td>15</td>
<td>11</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>MK9002169</td>
<td>5</td>
<td>NA</td>
<td>119</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>MK9001266</td>
<td>6</td>
<td>1</td>
<td>10</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>MK9001878</td>
<td>7</td>
<td>35</td>
<td>25</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>MK9001272</td>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>MK9000831</td>
<td>9</td>
<td>NA</td>
<td>8</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>MK9001379</td>
<td>10</td>
<td>34</td>
<td>5</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>
Results of the selection of the overlapped SNPs from the three analyses: random forest, stepwise regression, and GWAS

A total of 74 significant SNPs was overlapped among the three methods, GWAS, stepwise regression, and random forest. Based on the physical interval of ±5Mb and genetic distance ±5cM, the 74 SNPs were mapped in 38 loci in the 20 chromosomes. Ten of the 38 regions were mapped before, and the rest 28 loci are new. The top 10 SNPs based on R2 and P-values from GWAS were listed in table 5.5, and the summary of the distribution of the 74 SNPs was listed in table 5.6.

Table 5. Top 10 overlapped SNPs from three analysis: random forest, GWAS, and stepwise regression

<table>
<thead>
<tr>
<th>NK</th>
<th>LG</th>
<th>Chr</th>
<th>cM</th>
<th>Physical</th>
<th>P-Value</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK9001266</td>
<td>99</td>
<td>99</td>
<td>999.0</td>
<td>999</td>
<td>1.47E-08</td>
<td>0.04746</td>
</tr>
<tr>
<td>MK9000740</td>
<td>B2</td>
<td>14</td>
<td>102.4</td>
<td>34323117</td>
<td>5.98E-09</td>
<td>0.03531</td>
</tr>
<tr>
<td>MK9000024</td>
<td>B2</td>
<td>14</td>
<td>102.0</td>
<td>33571228</td>
<td>1.23E-08</td>
<td>0.03475</td>
</tr>
<tr>
<td>MK9001520</td>
<td>C2</td>
<td>6</td>
<td>95.7</td>
<td>9119063</td>
<td>6.09E-09</td>
<td>0.03331</td>
</tr>
<tr>
<td>MK9000252</td>
<td>B2</td>
<td>14</td>
<td>102.0</td>
<td>33571228</td>
<td>1.47E-09</td>
<td>0.03308</td>
</tr>
<tr>
<td>MK9001009</td>
<td>B2</td>
<td>14</td>
<td>102.4</td>
<td>34323117</td>
<td>2.53E-09</td>
<td>0.03107</td>
</tr>
<tr>
<td>MK9001858</td>
<td>B2</td>
<td>14</td>
<td>107.7</td>
<td>43188623</td>
<td>1.61E-08</td>
<td>0.03093</td>
</tr>
<tr>
<td>MK9001178</td>
<td>B2</td>
<td>14</td>
<td>105.9</td>
<td>42946869</td>
<td>1.64E-08</td>
<td>0.02962</td>
</tr>
<tr>
<td>MK9000915</td>
<td>B2</td>
<td>14</td>
<td>101.8</td>
<td>32380601</td>
<td>5.85E-08</td>
<td>0.02935</td>
</tr>
</tbody>
</table>

Table 6. Summary of the overlapped SNPs from three statistical methods across the 20 linkage groups

<table>
<thead>
<tr>
<th>Linkage group</th>
<th>#. Of SNPs</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1A</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1B</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SNP-SNP interaction network

Two networks from the 74 SNPs were obtained, one has only two nodes (showed with red dashed arrow in figure 4, bottom left, and Figure 5, and Figure 6); the other one has 71 nodes. These results indicated that there is a big network that controls soybean SDS. Further investigation of the two nodes network shows that these two nodes are in two different chromosomes, and their QTL effect size is 0.2 in the disease score range from 1 to 9.

Figure 4. the SNP by SNP interaction network

PPI network in Arabidopsis

Figure 5. PPI network in Arabidopsis obtained by blast soybean DNA against Arabidopsis PPI database

Network alignment by MI-GRAAL
Conclusions

1. overlapped QTLs and genes are mainly additive

2. The degree of interaction five was identified, and these five markers located in 5 different chromosomes, indicating that there is a coordinated network of QTLs functioning as an integrated unit.

3. The DNA network of QTL for SDS was constructed and aligned with the Arabidopsis network.

Conflict of interest

The authors declare that there is no conflict of interest.

Supplemental data available

Supplemental material is available online for this article. R codes and the data sets used for this research are available online.
Acknowledgments

We thank Syngenta, Iowa State University, and USDA-ARS for the financial support of this research. This research was supported in part by the US. Department of Agriculture, Agricultural Research Service, project 5030-21000-062-00D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

References


Anselin L (1988a) LAGRANGE MULTIPLIER TEST DIAGNOSTICS FOR SPATIAL DEPENDENCE AND SPATIAL HETEROGENEITY. Geographical Analysis 20:1-17


Bernal-Vasquez AM, Mohring J, Schmidt M, Schonleben M, Schon CC, Piepho HP (2014) The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. Bmc Genomics 15


Burau RG (1965) An investigation of soil factors in iron deficiency chlorosis of soybeans. Diss Abstr 26:603-603


imaging. In: Bosch JG, Doyly MM (eds) Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy


Cianzio SRD, Fehr WR, Anderson IC (1979) GENOTYPIC EVALUATION FOR IRON-DEFICIENCY CHLOROSIS IN SOYBEANS BY VISUAL SCORES AND CHLOROPHYLL CONCENTRATION. Crop Science 19:644-646


Diers BW, Voss BK, Fehr WR (1991) MOVING-MEAN ANALYSIS OF FIELD-TESTS FOR IRON EFFICIENCY OF SOYBEAN. Crop Science 31:54-56


Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketaebchi E (2005) Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC medical informatics and decision making 5:3-3


Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4:250-255


Froehlich DM, Fehr WR (1981) AGRONOMIC PERFORMANCE OF SOYBEANS WITH DIFFERING LEVELS OF IRON-DEFICIENCY CHLOROSIS ON CALCAREOUS SOIL. Crop Science 21:438-441


Gorjanc G, Jenko J, Hearne SJ, Hickey JM (2016) Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. Bmc Genomics 17


Gutierrez DD, Safari Books Online (Firm) Machine learning and data science : an introduction to statistical learning methods with R. First edition. edn, p 1 online resource (1 volume)


Hintz RW, Fehr WR, Cianzio SR (1987) POPULATION DEVELOPMENT FOR THE SELECTION OF HIGH-YIELDING SOYBEAN CULTIVARS WITH RESISTANCE TO IRON-DEFICIENCY CHLOROSIS. Crop Science 27:707-710


Ioannidis JPA (2005) Why most published research findings are false. Plos Medicine 2:696-701


Li FG, Wang ZP, Hu G, Li H (2011b) [Current status of SNPs interaction in genome-wide association study]. Yi Chuan 33:901-910


Melchinger AE, Utz HF, Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149:383-403


Miao X, Heaton JS (2010) A Comparison of Random Forest and Adaboost Tree in Ecosystem Classification in East Mojave Desert. 2010 18th International Conference on Geoinformatics


Nychka D (2016) Tools for Spatial Data


Robinson PM, Rossi F (2014) Improved Lagrange multiplier tests in spatial autoregressions. Econometrics Journal 17:139-164


Teo YY (2008) Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 19:133-143

Townleysmith TF, Hurd EA (1973) USE OF MOVING MEANS IN WHEAT YIELD TRIALS. Canadian Journal of Plant Science 53:447-450

Trimble MW, Fehr WR (1983) MIXTURES OF SOYBEAN CULTIVARS TO MINIMIZE YIELD LOSS CAUSED BY IRON-DEFICIENCY CHLOROSIS. Crop Science 23:691-694


CHAPTER 6. DISTRIBUTION AND EVOLUTION OF COTTON FIBER DEVELOPMENT GENES IN THE FIBRELESS G. RAİMONDII GENOME

Zhanyou Xua, Jing Yuc, Russell J. Kohela, Richard G. Percya, William D. Beavisb, Dorrie Mainc and John Z. Yua*

a USDA-ARS, Southern Plains Agricultural Research Center, College Station, Texas, USA
b Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA
c Department of Horticulture, Washington State University, Pullman, Washington, USA

Modified from a manuscript published in Genomics

Abstract

Cotton fibers represent the largest single cell in plants, and they serve as models to study cell development. This study investigated the distribution and evolution of fiber Unigenes anchored to recombination hotspots between tetraploid cotton (Gossypium hirsutum) At and Dt subgenomes, and within a parental diploid cotton (G. raimondii) D genome. Comparative analysis of At vs D and Dt vs D showed that 1) the D genome provides many fiber genes after its merger with another parental diploid cotton (G. arboreum) A genome although the D genome itself does not produce any spinnable fiber; 2) similarity of fiber genes is higher between At vs D than between Dt vs D genomic hotspots. This is the first report that fiber genes have higher similarity between At and D than between Dt and D. The finding provides new insights into cotton genomic regions that would facilitate genetic improvement of natural fiber properties.

Keywords

Cotton (Gossypium spp.), fiber development, EST Unigenes, recombination hotspots, gene distribution, genome evolution
Introduction

Cotton (Gossypium spp) fiber cells grow from an initial 11 to 22 μm to a mature length ranging from 1.0 to 6.0 cm, i.e., cells expand 1,000 to 3,000 times longer than their original size (Kim and Triplett 2001). There are only a few cells in the plant kingdom that exhibit such a large range of growth and development as cotton fibers. Because of its size and the relative ease of observation, cotton fiber is an ideal model for genetic and genomic studies of plant cell initiation, growth, and development. Since cotton fibers are unicellular, cell elongation can be evaluated independently of cell division. Cotton fiber cells grow and mature over a period of 40 to 50 days, allowing for observation at its various stage of development. Growth and development can be divided into four major stages: fiber initiation (from -3 to 3 days post anthesis or DPA), fiber elongation (0 to ~25 DPA), secondary cell wall deposition or SCWD (~19 to 45 DPA), and maturation or dehydration (45 to 50 DPA) (Schubert et al. 1973; Wilkins and Jernstedt 1999). The spatial and temporal regulation of growth and metabolic activity linked to the development of cotton fibers suggests that a large number of genes are involved in fiber development (Wilkins and Arpat 2005). The number of genes in the cotton genome is estimated to be in the range 40,000-50,000 (Rabinowicz et al. 2005), and 75%-94% of these genes are thought to be expressed to some extent in cotton fibers (Hovav et al. 2008b).

Expressed sequence tags (ESTs) for cotton fiber development have been sequenced and annotated using several technology platforms (Udall et al. 2006; Rapp et al. 2010; Lacape et al. 2012; Zhu et al. 2012; Yoo et al. 2013; Zou et al. 2013). Cotton ESTs have been systemically collected, assembled, clustered into Unigenes, formatted and deposited into the cotton genome database, CottonGen (http://www.cottongen.org/). Of 442,954 ESTs, most have been assembled into 21,698 contigs leaving 128,218 singletons (Yu et al. 2014).
Cultivated allotetraploid cottons (G. hirsutum and G. barbadense) contain two subgenomes, At and Dt, which most cotton geneticists believe are derived from a spinnable fiber capable “A” genome species (G. arboreum) (Wendel 1989) and a non-spinnable fiber capable “D” genome (G. raimondii) (Wendel 1989; Wendel et al. 1995). Cotton polyploidization has altered the function of the A and D ancestral diploid genomes, conferring emergent properties, such as higher fiber yield and quality (Xu et al. 2010; Paterson et al. 2012). An RFLP-QTL mapping study showed that more QTLs associated with fiber quality and yield are located on the Dt subgenome than the At subgenome of the cultivated tetraploid, even though the ancestral D-genome diploid progenitor does not produce spinnable fibers (Jiang et al. 1998).

In contrast, differential expression of RNA transcripts from the At subgenome is consistent with the evolution of spinnable fibers in the A-genome (Yang et al. 2006). Mapped fiber EST-derived SSRs, however, showed that both At and Dt subgenomes are equally important to fiber development (Han et al. 2006). Systematic mapping of fiber development Unigenes and transcription factors in At and Dt subgenomes showed that the At subgenome provides more genes for fiber development, suggesting that it continues to function similar to its spinnable fiber capable diploid A genome ancestor. On the other hand, the Dt subgenome, with its non-spinnable fiber capable D genome ancestor, provides more transcription factors that regulate the expression of the fiber genes in the At subgenome (Xu et al. 2010).

Evidence from gene expression patterns in interspecific hybrid F1, and synthetic and natural allopolyploid cottons using RNA-Seq reads from leaf transcriptome showed that genome-wide expression levels are more likely to come from the A genome in the diploid hybrid and natural allopolyploids, whereas the direction was reversed in the synthetic allopolyploid (Yoo et al. 2013). Results from deep sequencing of two cultivated cotton species showed that twice as
many contigs were derived from the Dt-genome as the At subgenome (Lacape et al. 2012).
Furthermore, comparison of 1,500 pairs of homoeologous genes from At- and Dt-subgenomes revealed a greater Dt-genome transcription throughout fiber development (Hovav et al. 2008a; Hovav et al. 2008b; Yoo and Wendel 2014).

The cotton fiber transcriptome can be clustered into two regions: conserved regions and recombination hotspots. Recombination hotspots are genomic regions within genetic linkage maps that exhibit high densities of markers (Newman et al. 1992; Wan et al. 2008; Smagulova et al. 2011), and the linkage maps are constructed on the basis of recombination of adjacent DNA markers or fragments in recombination hotspots. The association of markers and expressed genes is the basis for marker-assisted selection (MAS) in breeding. Based on the observations in various plant species it is reasonable to assume that the markers genetically located in the cotton linkage maps represent recombination-active regions in the cotton genome. Some of these regions may be important for cotton fiber yield and quality. ESTs derived cotton Unigenes were sequenced and assembled from functional regions in the genome. The overlapped regions between recombination hotspots and functional genomic areas were the primary targets for investigation.

Recently, the diploid D genome of wild cotton (G. raimondii) was sequenced independently by the Beijing Genomics Institute (BGI) (Wang et al. 2012) and the Joint Genome Institute (JGI) (Paterson et al. 2012). Herein, we utilized bioinformatics tools to investigate the genomic distribution and evolution of fiber development Unigenes among the tetraploid At and Dt subgenomes, and the diploid D genome (both BGI and JGI) in genomic regions that are likely associated with high frequencies of recombination. Fiber development Unigenes were anchored to recombination hotspot regions among the tetraploid At and Dt subgenomes, and the sequenced
diploid D genome. We accomplished this study by using cotton EST Unigenes, the sequenced D genome, BAC-end sequences from USDA-ARS, College Station, Texas, and all marker sequences that were publicly available for the development of simple sequence repeat (SSR), restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), single nucleotide polymorphism (SNP) markers in cotton http://www.cottongen.org (Yu et al. 2014).

**Materials and Methods**

**Sources of the cotton Unigenes used in this study**

A total of 437,185 cotton ESTs collected from 132 fiber cDNA libraries (Supplemental Table 1) were contributed by 18 research groups (Ji et al. 2003; Udall et al. 2006; Al-Ghazi et al. 2009; Yu et al. 2012). A total of 44,250 Unigenes assembled from 167,508 ESTs of 30 cotton fiber cDNA libraries were used in this study (Supplemental Tables 1 and 2). Unigenes were clustered into four groups based on fiber development stages: fiber initiation genes which are specifically expressed genes during fiber initiation stage from -3 to 3 days post anthesis or DPA; fiber elongation genes which are specifically expressed genes during fiber elongation stage from 0 to ~25 DPA; SCWD genes are specifically expressed during second cell wall deposition stage from ~19 to 45 DPA. In addition, a total of 23,460 sequences of genetically mapped markers (RFLP, SSR and SNP) (Yu et al. 2014) were also used in this study (Supplemental Table 3). Among the 23,460 markers, a total of 2,159 marker sequences were previously developed from diploid D genome cotton *G. raimondii*, although these markers were not directly inferred from the sequenced reference D genome http://www.cottongen.org (Yu et al. 2014).
Mapping of cotton EST Unigenes to chromosomes of the tetraploid (G. hirsutum) At and Dt subgenomes, and the diploid (G. raimondii) D genome

Two independently published diploid D genome sequences (Paterson et al. 2012; Wang et al. 2012) were downloaded and formatted into a local database to represent the non-spinnable fiber capable cotton genome sequence. Sequences from genetically mapped DNA markers were collected from published genetic maps (Blenda et al. 2012; Yu et al. 2012), downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), and subsequently formatted into a database to represent the fiber capable tetraploid Upland cotton (G. hirsutum).

**Fig. 1.** Flowchart to study the distribution of fiber development Unigenes in functional recombination hotspot regions between 13 pairs of At vs. D, and Dt vs. D chromosomes. JGI: Joint Genome Institute; BGI: Beijing Genomics Institute; EST: Expressed sequence tags; SCWD: Secondary cell wall deposition. *Housekeeping genes (PELEMAN et al. 1989) referred to both fiber and non-fiber development related Unigenes. These genes, not included in this study, are very conservative to species with very low frequency of changes.
A total of 44,250 fiber Unigenes (Supplemental Table 2; Udall et al. 2006; Yu et al. 2014) were downloaded for a BLAST comparison with tetraploid Upland cotton to locate the Unigenes in subgenomes At and Dt as well as in the diploid D genome. The Blast program “blastall” was downloaded from NCBI to perform the analysis with the criteria for a sequence match based on an E value less than 1e-20 and greater than 76 base pairs match. This working process from collecting marker and Unigene sequences, D genome sequences from BGI and JGI, to Blast and compare the distribution of the 13 chromosome pairs were summarized in Fig. 1.

**Results**

**Distribution of fiber initiation Unigenes among chromosomes of the tetraploid At subgenome and the diploid D genome**

Although BGI and JGI independently sequenced the D genome (*G. raimondii*) using seeds from different germplasm accessions, and with different genome coverages, we observed similar results from comparisons of fiber initiation Unigenes between At subgenome and D genome on all 13 chromosomes. A total of 793 fiber initiation Unigenes were anchored to the 13 chromosomes of tetraploid At subgenome. Of this total, 475 were found in the JGI-sequenced D genome. Among the 13 chromosome pairs of the At and D genomes, chromosome pair 10 had the highest similarity rate of 89.2% (25/28) between At and D (Table 1). Of a total of 25 Unigenes anchored in At subgenome chromosome 10, all also were found in D genome chromosome 10, and three of these Unigenes were found to have two copies in the D genome chromosome 10. In contrast, the chromosome 5 At and D pairing had the lowest shared percentage of fiber initiation Unigenes. A total of 109 Unigenes were anchored in chromosome 5 of the At subgenome, but only 35 of them were found in chromosome 5 of the D genome. The average frequency of shared Unigenes of all the 13 pairs of chromosomes is 58.4% (0.584).
A similar comparison between the 13 At chromosomes and 13 BGI-sequenced D genome chromosomes revealed the chromosome 10 pairing to have the highest frequency of shared Unigenes 89.3% (0.893). Therefore, comparisons of the shared fiber initiation Unigenes of the At subgenome chromosome 10 and the D genome chromosome 10 were the same for the JGI and BGI sequences. Likewise, comparisons of the shared fiber initiation Unigene frequencies between the At subgenome chromosomes 6 and 7 and the D genome chromosomes 6 and 7 were similar for the JGI and BGI sequences although there were slightly fewer shared fiber initiation Unigenes between the JGI and BGI sequences.

Table 1. Summary of fiber initiation Unigenes in recombination hotspot regions of 26 chromosomes in the tetraploid At and Dt subgenomes vs. the two diploid D genomes sequenced by BGI and JGI, respectively.

<table>
<thead>
<tr>
<th>Subgenome At or Dt</th>
<th># Initiation Unigenes</th>
<th>Diploid D (JGI)</th>
<th># Initiation Unigenes</th>
<th>% D:At/Dt (JGI)</th>
<th>Diploid D (BGI)</th>
<th># Initiation Unigenes</th>
<th>% D:At/Dt (BGI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr.01(At01)</td>
<td>43</td>
<td>Chr.02(D01)</td>
<td>24</td>
<td>55.81</td>
<td>Chr.02(D01)</td>
<td>14</td>
<td>32.56</td>
</tr>
<tr>
<td>Chr.02(At02)</td>
<td>76</td>
<td>Chr.05(D02)</td>
<td>36</td>
<td>47.37</td>
<td>Chr.05(D02)</td>
<td>27</td>
<td>35.53</td>
</tr>
<tr>
<td>Chr.03(At03)</td>
<td>76</td>
<td>Chr.03(D03)</td>
<td>27</td>
<td>35.53</td>
<td>Chr.03(D03)</td>
<td>14</td>
<td>18.42</td>
</tr>
<tr>
<td>Chr.04(At04)</td>
<td>56</td>
<td>Chr.12(D04)</td>
<td>42</td>
<td>75.00</td>
<td>Chr.12(D04)</td>
<td>13</td>
<td>23.21</td>
</tr>
<tr>
<td>Chr.05(At05)</td>
<td>109</td>
<td>Chr.09(D05)</td>
<td>35</td>
<td>32.11</td>
<td>Chr.09(D05)</td>
<td>23</td>
<td>21.10</td>
</tr>
<tr>
<td>Chr.06(At06)</td>
<td>34</td>
<td>Chr.10(D06)</td>
<td>60</td>
<td>56.67</td>
<td>Chr.10(D06)</td>
<td>59</td>
<td>57.63</td>
</tr>
<tr>
<td>Chr.07(At07)</td>
<td>34</td>
<td>Chr.01(D07)</td>
<td>44</td>
<td>77.27</td>
<td>Chr.01(D07)</td>
<td>48</td>
<td>70.83</td>
</tr>
<tr>
<td>Chr.08(At08)</td>
<td>86</td>
<td>Chr.04(D08)</td>
<td>31</td>
<td>36.05</td>
<td>Chr.04(D08)</td>
<td>33</td>
<td>38.37</td>
</tr>
<tr>
<td>Chr.09(At09)</td>
<td>89</td>
<td>Chr.06(D09)</td>
<td>59</td>
<td>66.29</td>
<td>Chr.06(D09)</td>
<td>45</td>
<td>50.56</td>
</tr>
<tr>
<td>Chr.10(At10)</td>
<td>25</td>
<td>Chr.11(D10)</td>
<td>28</td>
<td>89.29</td>
<td>Chr.11(D10)</td>
<td>28</td>
<td>89.29</td>
</tr>
<tr>
<td>Chr.11(At11)</td>
<td>86</td>
<td>Chr.07(D11)</td>
<td>28</td>
<td>32.56</td>
<td>Chr.07(D11)</td>
<td>24</td>
<td>27.91</td>
</tr>
<tr>
<td>Chr.12(At12)</td>
<td>36</td>
<td>Chr.08(D12)</td>
<td>29</td>
<td>80.56</td>
<td>Chr.08(D12)</td>
<td>32</td>
<td>88.89</td>
</tr>
<tr>
<td>Chr.13(At13)</td>
<td>43</td>
<td>Chr.13(D13)</td>
<td>32</td>
<td>74.42</td>
<td>Chr.13(D13)</td>
<td>28</td>
<td>65.12</td>
</tr>
<tr>
<td>Sub D:At</td>
<td>793</td>
<td>475</td>
<td>388</td>
<td>58.38</td>
<td>388</td>
<td>47.65</td>
<td></td>
</tr>
<tr>
<td>Chr.15(Dt01)</td>
<td>101</td>
<td>Chr.02(D01)</td>
<td>18</td>
<td>17.82</td>
<td>Chr.02(D01)</td>
<td>21</td>
<td>20.79</td>
</tr>
<tr>
<td>Chr.17(Dt02)</td>
<td>12</td>
<td>Chr.05(D02)</td>
<td>25</td>
<td>48.00</td>
<td>Chr.05(D02)</td>
<td>18</td>
<td>66.67</td>
</tr>
<tr>
<td>Chr.14(Dt03)</td>
<td>58</td>
<td>Chr.03(D03)</td>
<td>21</td>
<td>36.21</td>
<td>Chr.03(D03)</td>
<td>3</td>
<td>5.17</td>
</tr>
<tr>
<td>Chr.22(Dt04)</td>
<td>41</td>
<td>Chr.12(D04)</td>
<td>35</td>
<td>85.37</td>
<td>Chr.12(D04)</td>
<td>11</td>
<td>26.83</td>
</tr>
<tr>
<td>Chr.19(Dt05)</td>
<td>94</td>
<td>Chr.09(D05)</td>
<td>26</td>
<td>27.66</td>
<td>Chr.09(D05)</td>
<td>6</td>
<td>6.38</td>
</tr>
<tr>
<td>Chr.25(Dt06)</td>
<td>29</td>
<td>Chr.10(D06)</td>
<td>23</td>
<td>79.31</td>
<td>Chr.10(D06)</td>
<td>21</td>
<td>72.41</td>
</tr>
<tr>
<td>Chr.16(Dt07)</td>
<td>46</td>
<td>Chr.01(D07)</td>
<td>26</td>
<td>56.52</td>
<td>Chr.01(D07)</td>
<td>32</td>
<td>69.57</td>
</tr>
<tr>
<td>Chr.24(Dt08)</td>
<td>53</td>
<td>Chr.04(D08)</td>
<td>15</td>
<td>28.30</td>
<td>Chr.04(D08)</td>
<td>15</td>
<td>28.30</td>
</tr>
</tbody>
</table>
Table 1 Continued

<table>
<thead>
<tr>
<th>Subgenome</th>
<th># Initiation Unigenes</th>
<th>Diploid D (JGI)</th>
<th>% Initiation Unigenes</th>
<th>Diploid D (BGI)</th>
<th># Initiation Unigenes</th>
<th>% Initiation Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr.23(Dt09)</td>
<td>34</td>
<td>Chr.06(D09)</td>
<td>22</td>
<td>64.71</td>
<td>Chr.06(D09)</td>
<td>23</td>
</tr>
<tr>
<td>Chr.20(Dt10)</td>
<td>37</td>
<td>Chr.11(D10)</td>
<td>18</td>
<td>48.65</td>
<td>Chr.11(D10)</td>
<td>17</td>
</tr>
<tr>
<td>Chr.21(Dt11)</td>
<td>8</td>
<td>Chr.07(D11)</td>
<td>31</td>
<td>25.81</td>
<td>Chr.07(D11)</td>
<td>33</td>
</tr>
<tr>
<td>Chr.26(Dt12)</td>
<td>24</td>
<td>Chr.08(D12)</td>
<td>24</td>
<td>100.00</td>
<td>Chr.08(D12)</td>
<td>19</td>
</tr>
<tr>
<td>Chr.18(Dt13)</td>
<td>32</td>
<td>Chr.13(D13)</td>
<td>26</td>
<td>81.25</td>
<td>Chr.13(D13)</td>
<td>27</td>
</tr>
<tr>
<td><strong>Sub D: Dt</strong></td>
<td><strong>569</strong></td>
<td><strong>310</strong></td>
<td><strong>53.82</strong></td>
<td><strong>246</strong></td>
<td><strong>45.96</strong></td>
<td></td>
</tr>
</tbody>
</table>

There are 76 fiber initiation Unigenes in the At genome located on chromosome 3, but only 14 of them were found on chromosome 3 of the D genome. The number of fiber initiation Unigenes observed on chromosomes 6, 7 and 10 of the D genome, when compared to the number observed on the At genome, did not conform to our expectation. Since the D genome itself does not produce any spinnable fibers, it was our expectation that it should provide fewer copies of fiber initiation genes in its chromosomes than that in At chromosomes.

**Distribution of fiber initiation Unigenes among chromosomes of the tetraploid Dt subgenome and the diploid D genome**

Assignment of fiber initiation genes to the 13 chromosomes of the Dt genome (Dt01-Dt13) or chromosome 14-26 are summarized in the first column of Table 1, based on previous reports (Rong et al. 2004; Wang et al. 2008; Lacape et al. 2009; Blenda et al. 2012; Zhao et al. 2012; Wang et al. 2013). First, consider the comparison of fiber initiation Unigenes in the chromosomes of the Dt subgenome and the JGI-sequenced D genome. Chromosome 15 (Dt01) had the lowest frequency of shared fiber initiation Unigenes with D01. A total of 101 fiber initiation Unigenes were assigned to Dt01, but only 18 of the 101 were found in D01 (frequency = 0.178). At the other extreme, chromosome 26 (Dt12) shared all 24 fiber initiation Unigenes with D12. This pair of highly conserved chromosomes was previously observed and reported by Xu et al. (2008). On average across the 13 pairs of Dt and D chromosomes, 53.8% (0.538) of the
fiber initiation Unigenes were shared. This was significantly (p<0.05) less than the average frequency of shared fiber initiation Unigenes among 13 pairs of At vs. D chromosomes (Table 1). This was the first observation that diploid D chromosomes shared more fiber initiation Unigenes with At than that with Dt after the union of the diploid A and D genomes, based on genome-wide Unigenes anchored in recombination hot regions. This result suggests sequence changes of the D genome after the merger and polyploidization of the A and D diploid genomes. Similar results were obtained from the comparison between 13 pairs of Dt and D chromosomes sequenced by BGI. The highest fiber initiation Unigene sharing rate among the 13 pairs of chromosomes was 84.38% between chromosomes 18 (Dt13) and D13. The second highest sharing rate was 79.17% between chromosomes 16 (Dt12) and D12. The lowest sharing rate of only 5.17% occurred between chromosomes 14 (Dt03) and D03. There were 58 fiber initiation Unigenes anchored to chromosome 14, but only three of them were found in chromosome D03. The differences in results between the BGI and JGI sequences, while generally showing common trends, suggests that there might be differences in sequenced regions and depths between BGI and JGI.

**Distribution of fiber elongation Unigenes among chromosomes of the tetraploid At subgenome and the diploid D genome**

On average, 69.2% (0.692) of the fiber elongation Unigenes were shared across the 13 pairings of At genome and JGI D genome chromosomes. Comparisons of five chromosome pairs (At01:D01, At10:D10, At12:D12, At04:D04, At09:D09) of the At and JGI-D pairing produced a shared Unigene frequency greater than 95% (0.95) for the fiber elongation (Table 2). The smallest frequency 35.2% (0.352) of shared fiber elongation genes occurred between Dt05 and D05. In the comparison between the At genome and the D genome sequenced by BGI, four chromosome pairs (At01:D01, At10:D10, At12:D12, At09:D09) had high frequencies of shared
fiber elongation Unigenes (Table 2). The least frequent shared fiber elongation Unigenes was 28.2% (0.282) between At05 and D05. Comparing the results from fiber elongation Unigenes with those from fiber initiation Unigenes, the average sharing rates were 69.18% for fiber elongation versus 53.82% for fiber initiation (At vs. JGI-D), and 55.47% from fiber elongation versus 45.96% from fiber initiation (At vs. BGI-D). There were more fiber elongation Unigenes shared between At and D chromosomes than fiber initiation Unigenes.

<table>
<thead>
<tr>
<th>Subgenome At or Dt</th>
<th># Elongation Unigenes</th>
<th>Diploid D (JGI)</th>
<th># Elongation Unigenes</th>
<th>% D:At/Dt (JGI)</th>
<th>Diploid D (BGI)</th>
<th># Elongation Unigenes</th>
<th>% D:At/Dt (BGI)</th>
</tr>
</thead>
</table>
Table 2 Continued

| Subgenome At or Dt | # Elongation Unigenes | Diploid D (JGI) | # Elongation Unigenes | % D:At|Dt (JGI) | Diploid D (BGI) | # Elongation Unigenes | % D:At|Dt (BGI) |
|--------------------|-----------------------|----------------|-----------------------|--------|----------|----------------|-----------------------|--------|
| Chr.18(Dt13)       | 27                    | Chr.13(D13)    | 18                    | 66.67  | Chr.13(D13) | 21            | 77.78                 |
| chr.13(D13)        | 260                   | 58.08          | 204                   | 51.14  |
| Sub D: Dt          | 344                   |                |                       |        |

Distribution of fiber elongation Unigenes among chromosomes of the tetraploid Dt subgenome and the diploid D genome

A total of 344 fiber elongation Unigenes (Table 2) were anchored to 13 Dt chromosomes, which were much fewer than 495 in 13 At chromosomes. Based on the comparative studies of 13 chromosome pairs between Dt subgenome and JGI-sequenced D genome, three chromosome pairs (Chr.16:D07, Chr.25:D06, Chr.23:D09) had the highest sharing rate, with the highest rate of 96.3% occurring between Chr.23 and D09. In contrast, the lowest sharing rate was 20.45% between Chr.15 and D01. In the Dt subgenome and BGI-sequenced D genome pairings, the average sharing rate among the 13 chromosome pairs was 58.08%. Both the highest (Chr.16:D07, Chr.25:D06, Chr.23:D09) and the lowest (Chr.15: D01) sharing rates among chromosomes were the same for the JGI- and BGI- sequenced D genomes, indicating that both sequenced genomes covered the major elongation Unigenes (Table 2).

The lower rate of shared Unigenes for fiber elongation in 13 Dt:D pairings than in 13 At:D pairings (58.08% vs. 69.18% in JGI sequence and 51.14% vs. 55.47% in BGI sequence) mirrored the results obtained for fiber initiation Unigenes.

Distribution of fiber second cell wall deposition (SCWD) Unigenes among chromosomes of the diploid D genome and the tetraploid At and Dt subgenomes

As with fiber initiation and fiber elongation, SCWD Unigenes were anchored to the 13 chromosome pairs for comparative studies. Only 17 SCWD Unigenes were anchored to the 13 At chromosomes, and 13 SCWD Unigenes (Table 3) anchored to the 13 Dt chromosomes. Due
to the small number of the SCWD Unigenes in each individual chromosome, only the total number of SCWD Unigenes in the At subgenome and the D genome, as well as the Dt subgenome and the D genome were compared. In the comparison of the At subgenome and JGI-sequenced D genome, all the 17 SCWD Unigenes found anchored to the At chromosomes were also found in the D genome, but not always on the same chromosomes, resulting in a sharing rate of 31.41%. In the comparison of the At and the BGI-sequenced D genome, 13 of the 17 SCWD Unigenes in At were found in D, with an average sharing rate of 19.23%, nine of the 13 SCWD Unigenes anchored to the Dt chromosomes were found in the D chromosomes.

In the comparison of the SCWD Unigenes present in the Dt subgenome and the D genome sequence, not only were all the 13 SCWD Unigenes found in the Dt also found in the D chromosomes, but five SCWD Unigenes had two copies in the D chromosomes for a total of 18 (Table 3). The average sharing rate between 13 pairs of Dt and D chromosomes was 8.21%.

<table>
<thead>
<tr>
<th>Subgenomes</th>
<th>SCWD Unigenes</th>
<th>Diploid D (JGI)</th>
<th>% D:At</th>
<th>Diploid D (BGI)</th>
<th>% D:At</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr.01(At01)</td>
<td>0</td>
<td>Chr.02(D01)</td>
<td>3</td>
<td>0.00</td>
<td>Chr.02(D01)</td>
</tr>
<tr>
<td>Chr.02(At02)</td>
<td>0</td>
<td>Chr.05(D02)</td>
<td>4</td>
<td>0.00</td>
<td>Chr.05(D02)</td>
</tr>
<tr>
<td>Chr.03(At03)</td>
<td>1</td>
<td>Chr.03(D03)</td>
<td>0</td>
<td>0.00</td>
<td>Chr.03(D03)</td>
</tr>
<tr>
<td>Chr.04(At04)</td>
<td>0</td>
<td>Chr.12(D04)</td>
<td>1</td>
<td>0.00</td>
<td>Chr.12(D04)</td>
</tr>
<tr>
<td>Chr.05(At05)</td>
<td>2</td>
<td>Chr.09(D05)</td>
<td>1</td>
<td>50.00</td>
<td>Chr.09(D05)</td>
</tr>
<tr>
<td>Chr.06(At06)</td>
<td>0</td>
<td>Chr.10(D06)</td>
<td>3</td>
<td>0.00</td>
<td>Chr.10(D06)</td>
</tr>
<tr>
<td>Chr.07(At07)</td>
<td>3</td>
<td>Chr.01(D07)</td>
<td>1</td>
<td>33.33</td>
<td>Chr.01(D07)</td>
</tr>
<tr>
<td>Chr.08(At08)</td>
<td>4</td>
<td>Chr.04(D08)</td>
<td>0</td>
<td>0.00</td>
<td>Chr.04(D08)</td>
</tr>
<tr>
<td>Chr.09(At09)</td>
<td>4</td>
<td>Chr.06(D09)</td>
<td>1</td>
<td>25.00</td>
<td>Chr.06(D09)</td>
</tr>
<tr>
<td>Chr.10(At10)</td>
<td>0</td>
<td>Chr.11(D10)</td>
<td>1</td>
<td>0.00</td>
<td>Chr.11(D10)</td>
</tr>
<tr>
<td>Chr.11(At11)</td>
<td>1</td>
<td>Chr.07(D11)</td>
<td>1</td>
<td>100.00</td>
<td>Chr.07(D11)</td>
</tr>
<tr>
<td>Chr.12(At12)</td>
<td>1</td>
<td>Chr.08(D12)</td>
<td>1</td>
<td>100.00</td>
<td>Chr.08(D12)</td>
</tr>
<tr>
<td>Chr.13(At13)</td>
<td>1</td>
<td>Chr.13(D13)</td>
<td>1</td>
<td>100.00</td>
<td>Chr.13(D13)</td>
</tr>
</tbody>
</table>

Table 3 Summary of SCWD Unigenes in recombination hotspot regions of 26 chromosome pairs between the tetraploid At and Dt subgenome vs. the two diploid D genomes.
Overall summary of the distribution of fiber initiation, fiber elongation, and SCWD Unigenes in recombination hotspot regions between the tetraploid At and Dt subgenomes vs. the diploid D genome

To provide an overall view of the distributions from all the fiber development stages: fiber initiation, fiber elongation, and SCWD, the mean values from Tables 1, 2 and 3 are summarized in Table 4. For the D-genome sequence assembled by JGI, the proportion of Unigenes shared between At and D was larger than that shared between Dt and D: 58.38% vs. 53.82% for fiber initiation, 69.18% vs 58.08% for fiber elongation, and 31.41% vs 8.21% for SCWD. For the D-genome sequence assembled by BGI, similar results were observed that the proportion of Unigenes shared between At and D was larger than that shared between Dt and D: 47.65% vs 45.96% for fiber initiation, 55.47% vs 51.14% for fiber elongation, and 19.23% vs. 6.41% for SCWD. Statistical mean difference t-test by SAS 9.4 (SAS Institute Inc, North Carolina, USA) showed that the difference between the mean values of At vs D and Dt vs D was statistically significant with the P-value of < 0.05.
Table 4 Overall summary of the distribution of shared fiber initiation, fiber elongation, and SCWD Unigenes in recombination hotspot regions between the two tetraploid subgenomes and the two diploid D genomes

<table>
<thead>
<tr>
<th>Fiber development stage and Institute name</th>
<th>Number of fiber Unigenes between At and D (%)</th>
<th>Number of fiber Unigenes between Dt and D (%)</th>
<th>Difference between At vs D and Dt vs D (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JGI</td>
<td>BGI</td>
<td>JGI</td>
</tr>
<tr>
<td>Initiation</td>
<td>58.38</td>
<td>47.65</td>
<td>53.82</td>
</tr>
<tr>
<td>Elongation</td>
<td>69.18</td>
<td>55.47</td>
<td>58.08</td>
</tr>
<tr>
<td>SCWD</td>
<td>31.41</td>
<td>19.23</td>
<td>8.21</td>
</tr>
</tbody>
</table>

T-test  

<table>
<thead>
<tr>
<th></th>
<th>JGI</th>
<th>BGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value=0.0155</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chromosomal distribution of the At- or Dt-specific fiber initiation Unigenes

Data presented in Tables 1 to 4 show the distribution of fiber development Unigenes of the At and Dt subgenomes that are shared with the diploid D chromosomes. In contrast, Table 5 summarizes fiber initiation Unigenes specific to the At and Dt subgenomes but lacking in the D chromosomes. Fiber initiation Unigenes anchored to the At chromosomes without any copy in the D genome were treated as At-specific Unigenes, with a similar postulation being made for Dt-specific Unigenes. Chromosome 21 or Dt11 had eight fiber-initiation Unigenes, and all of these Unigenes had at least one copy in the diploid D genome. There was no At specific fiber-initiation Unigene in chromosome Dt11. Among the chromosomes, the highest percentage of At or Dt specific Unigenes (i.e., not found in the D genome) was 23.53% between chromosomes 01 and D01. Eight of the 34 fiber-initiation Unigenes anchored to C01 were not found the D genome. The average percentage of At or Dt specific fiber-initiation Unigenes (not found in the D genome) were 13.33% and 14.22%, respectively. These results indicate that there is no significant difference (p>0.05) as to the At-specific and Dt-specific fiber-initiation Unigenes by comparing between At vs. D and Dt vs. D genomes.
Table 5 Summary of fiber initiation Unigenes not found in the diploid D genomes

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Total # Unigenes</th>
<th># At or Dt specific Unigenes</th>
<th># At or Dt Unigenes shared with the D genome</th>
<th>% At or Dt specific Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>c01</td>
<td>34</td>
<td>8</td>
<td>26</td>
<td>23.53</td>
</tr>
<tr>
<td>c02</td>
<td>43</td>
<td>3</td>
<td>40</td>
<td>6.98</td>
</tr>
<tr>
<td>c03</td>
<td>64</td>
<td>6</td>
<td>58</td>
<td>9.38</td>
</tr>
<tr>
<td>c04</td>
<td>40</td>
<td>2</td>
<td>38</td>
<td>5.00</td>
</tr>
<tr>
<td>c05</td>
<td>79</td>
<td>13</td>
<td>66</td>
<td>16.46</td>
</tr>
<tr>
<td>c06</td>
<td>29</td>
<td>6</td>
<td>23</td>
<td>20.69</td>
</tr>
<tr>
<td>c07</td>
<td>26</td>
<td>4</td>
<td>22</td>
<td>15.38</td>
</tr>
<tr>
<td>c08</td>
<td>80</td>
<td>5</td>
<td>75</td>
<td>6.25</td>
</tr>
<tr>
<td>c09</td>
<td>80</td>
<td>6</td>
<td>74</td>
<td>7.50</td>
</tr>
<tr>
<td>c10</td>
<td>23</td>
<td>2</td>
<td>21</td>
<td>8.70</td>
</tr>
<tr>
<td>c11</td>
<td>60</td>
<td>11</td>
<td>49</td>
<td>18.33</td>
</tr>
<tr>
<td>c12</td>
<td>35</td>
<td>6</td>
<td>29</td>
<td>17.14</td>
</tr>
<tr>
<td>c13</td>
<td>39</td>
<td>7</td>
<td>32</td>
<td>17.95</td>
</tr>
<tr>
<td>c14</td>
<td>39</td>
<td>5</td>
<td>34</td>
<td>12.82</td>
</tr>
<tr>
<td>c15</td>
<td>51</td>
<td>5</td>
<td>46</td>
<td>9.80</td>
</tr>
<tr>
<td>c16</td>
<td>38</td>
<td>5</td>
<td>33</td>
<td>13.16</td>
</tr>
<tr>
<td>c17</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>22.22</td>
</tr>
<tr>
<td>c18</td>
<td>31</td>
<td>4</td>
<td>27</td>
<td>12.90</td>
</tr>
<tr>
<td>c19</td>
<td>60</td>
<td>8</td>
<td>52</td>
<td>13.33</td>
</tr>
<tr>
<td>c20</td>
<td>33</td>
<td>4</td>
<td>29</td>
<td>12.12</td>
</tr>
<tr>
<td>c21</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0.00</td>
</tr>
<tr>
<td>c22</td>
<td>41</td>
<td>8</td>
<td>33</td>
<td>19.51</td>
</tr>
<tr>
<td>c23</td>
<td>26</td>
<td>4</td>
<td>22</td>
<td>15.38</td>
</tr>
<tr>
<td>c24</td>
<td>45</td>
<td>8</td>
<td>37</td>
<td>17.78</td>
</tr>
<tr>
<td>c25</td>
<td>24</td>
<td>5</td>
<td>19</td>
<td>20.83</td>
</tr>
<tr>
<td>c26</td>
<td>20</td>
<td>3</td>
<td>17</td>
<td>15.00</td>
</tr>
<tr>
<td>Total</td>
<td>1057</td>
<td>140</td>
<td>917</td>
<td>13.25</td>
</tr>
</tbody>
</table>

Chromosomal distribution of the At- or Dt-specific fiber elongation Unigenes

Fiber elongation Unigenes on chromosomes of the At or Dt subgenomes were compared with elongation Unigenes on the D genome chromosomes to identify At-specific or Dt-specific Unigenes. Tetraploid chromosome 21 or Dt11 had seven fiber elongation Unigenes (Table 6), all of which were also found in the D genome, and therefore there were no At-specific fiber
elongation Unigenes located on chromosome 21. These results corresponded to results for fiber initiation Unigenes, and indicated that chromosome Dt11 was the most conservative in the distribution of fiber development Unigenes. In contrast, chromosome 22 had 22 fiber elongation Unigenes, six of which had no copy in the D genome and therefore the largest percentage (27.27%) of Dt-specific Unigenes. The average percentage of At-specific or Dt-specific fiber elongation Unigenes among the 13 chromosomes was 12.75% and 14.11%, respectively. There was no significant difference between these two estimates (P>0.05). As the results shown in both Tables 5 and 6, there was no significant difference between At-specific and Dt-specific Unigenes in the percentage of subgenome specific fiber development Unigenes.

**Table 6** Summary of fiber elongation Unigenes not found in the diploid D genome

<table>
<thead>
<tr>
<th>chromosome</th>
<th># total Unigenes</th>
<th># At or Dt specific Unigenes</th>
<th># At or Dt Unigenes shared with the D genome</th>
<th>% At or Dt specific Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>c01</td>
<td>22</td>
<td>3</td>
<td>19</td>
<td>13.64</td>
</tr>
<tr>
<td>c02</td>
<td>34</td>
<td>4</td>
<td>30</td>
<td>11.76</td>
</tr>
<tr>
<td>c03</td>
<td>49</td>
<td>3</td>
<td>46</td>
<td>6.12</td>
</tr>
<tr>
<td>c04</td>
<td>33</td>
<td>2</td>
<td>31</td>
<td>6.06</td>
</tr>
<tr>
<td>c05</td>
<td>71</td>
<td>9</td>
<td>62</td>
<td>12.68</td>
</tr>
<tr>
<td>c06</td>
<td>16</td>
<td>2</td>
<td>14</td>
<td>12.50</td>
</tr>
<tr>
<td>c07</td>
<td>21</td>
<td>4</td>
<td>17</td>
<td>19.05</td>
</tr>
<tr>
<td>c08</td>
<td>62</td>
<td>6</td>
<td>56</td>
<td>9.68</td>
</tr>
<tr>
<td>c09</td>
<td>53</td>
<td>5</td>
<td>48</td>
<td>9.43</td>
</tr>
<tr>
<td>c10</td>
<td>22</td>
<td>4</td>
<td>18</td>
<td>18.18</td>
</tr>
<tr>
<td>c11</td>
<td>56</td>
<td>12</td>
<td>44</td>
<td>21.43</td>
</tr>
<tr>
<td>c12</td>
<td>23</td>
<td>3</td>
<td>20</td>
<td>13.04</td>
</tr>
<tr>
<td>c13</td>
<td>33</td>
<td>4</td>
<td>29</td>
<td>12.12</td>
</tr>
<tr>
<td>c14</td>
<td>30</td>
<td>6</td>
<td>24</td>
<td>20.00</td>
</tr>
<tr>
<td>c15</td>
<td>44</td>
<td>3</td>
<td>41</td>
<td>6.82</td>
</tr>
<tr>
<td>c16</td>
<td>21</td>
<td>2</td>
<td>19</td>
<td>9.52</td>
</tr>
<tr>
<td>c17</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>25.00</td>
</tr>
<tr>
<td>c18</td>
<td>27</td>
<td>2</td>
<td>25</td>
<td>7.41</td>
</tr>
<tr>
<td>c19</td>
<td>55</td>
<td>7</td>
<td>48</td>
<td>12.73</td>
</tr>
<tr>
<td>c20</td>
<td>25</td>
<td>4</td>
<td>21</td>
<td>16.00</td>
</tr>
</tbody>
</table>
Table 6 Continued

<table>
<thead>
<tr>
<th>chromosome</th>
<th># total Unigenes</th>
<th># At or Dt specific Unigenes</th>
<th># At or Dt Unigenes shared with the D genome</th>
<th>% At or Dt specific Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>c21</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0.00</td>
</tr>
<tr>
<td>c22</td>
<td>22</td>
<td>6</td>
<td>16</td>
<td>27.27</td>
</tr>
<tr>
<td>c23</td>
<td>27</td>
<td>5</td>
<td>22</td>
<td>18.52</td>
</tr>
<tr>
<td>c24</td>
<td>48</td>
<td>3</td>
<td>45</td>
<td>6.25</td>
</tr>
<tr>
<td>c25</td>
<td>19</td>
<td>3</td>
<td>16</td>
<td>15.79</td>
</tr>
<tr>
<td>c26</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>18.18</td>
</tr>
<tr>
<td>Total</td>
<td>839</td>
<td>106</td>
<td>733</td>
<td>12.63</td>
</tr>
</tbody>
</table>

Visualization of the fiber Unigenes among At subgenome, Dt subgenome, and diploid D genomes

An example of fiber development Unigenes in chromosome D01 segment that were aligned in genome browser was shown in Fig. 2 below (Engels et al. 2006; Yu et al. 2014).

Several cotton fiber Unigenes conferring fiber initiation (Cotton_08929) and fiber SCWD (Cotton_19118) are located in this genomic region. The results of all cotton Unigenes anchored in the tetraploid At and Dt subgenomes as well as in the diploid D genome in this study were deposited into the cotton genome database CottonGen (http://www.cottongen.org/), and they will be available once this manuscript is accepted and all can be visualized in genome browser.

Fig. 2. Distribution of cotton fiber development Unigenes in chromosome D01. The upper half of the figure provides an overview of the cotton Unigenes in chromosome D01 from 0 bp to 58,000,000 bp. There are four groups of Unigenes listed in the figure as fiber initiation, fiber elongation, secondary cell wall deposition (SCWD), and non-fiber Unigenes. Within each of four groups, there are three subgroups of Unigenes that were derived from diploid genome D, tetraploid subgenome At and subgenome Dt.
Discussion

Distribution of the genetically mapped sequences in the tetraploid cotton At and Dt subgenomes

This study is the first report that the similarity or proportion of fiber development Unigenes in recombination hotspot regions between At and D chromosomes is higher than that between Dt and D chromosomes. Prior to this study, we expected that the diploid D genome, considered as the most probable donor for the Dt subgenome of the tetraploid cotton, would have a greater Unigene similarity with the Dt subgenome than with the At subgenome. The actual results obtained from this study indicated otherwise and they were beyond our original expectations. There may be several reasons resulting to this observation: the first, the mapped sequences used in this study are from genetically mapped markers and BAC-end sequences. These sequences represent the recombination-active regions, since genetic maps are constructed based on recombination between markers or between markers and traits of interest. The distribution of the mapped marker sequences between At and Dt subgenomes may differ and thus may affect the results of this research. When a total of 3,347 sequence tagged sites (STS), spanning 4,447 cM, were mapped in tetraploid cotton, the genetic map of subgenome At was 9.5% longer than that of subgenome Dt (Rong et al. 2004). However, when a total of 1,900 EST-based SSR loci, spanning 3,426 cM, were mapped, the genetic map of Dt subgenome was 4.5% longer in genetic length than that of At subgenome (Guo et al. 2007). Although the ratios of subgenome sizes between At and Dt vary, the differences are not significant (P>0.05). These variations may also result from differences in mapping population sizes, as well as in the numbers and sources of DNA markers used in the studies (Rong et al. 2004; Park et al. 2005; Guo et al. 2007; Yu et al. 2011). Latest mapping results further confirmed that two subgenomes At and Dt of the tetraploid cotton are equivalent in recombination frequencies between At and Dt.
despite the extra repetitive DNA in the At subgenome (Yu et al. 2012). Based on the summary of previous mapping reports we can draw the conclusion that the distribution of the genetically mapped sequences between the At and Dt subgenomes in modern tetraploid cottons are equal, since recombination rates in the At and Dt subgenomes are the same.

**Distribution of fiber development Unigenes in recombination hotspot regions of the tetraploid At and Dt subgenomes, and the diploid D genome**

Cotton genomes contain two components of nuclear DNA: recombination active and recombination inactive (conserved). Conserved regions are identical or similar sequences across species or genomes, and cannot be genetically mapped by using markers due to the absence of recombination and the lack of polymorphism (Reinisch et al. 1994; Saintenac et al. 2011). In contrast, recombination-active regions are readily mapped due to the occurrence of crossover events between two chromosomes. Genome regions that are well mapped with markers or sequences represent recombination-active regions of the genome, and therefore are the best regions for investigations of the evolution of the fiber development genes, and they are the target regions for cotton MAS breeding. The results from this study are consistent with previous reports that the Dt subgenome contributes more fiber QTLs than the At subgenome (Jiang et al. 1998; Paterson et al. 2012). In an investigation of QTLs for fiber-related traits, Jiang et al. (1998) reported that 10 of the 14 QTLs (71.42%) were located on Dt subgenome linkage groups. Only four QTLs (28.58%) mapped to overlapping intervals of LGs A02 and D03 (Jiang et al. 1998). Ten years later, 43 QTLs for fiber quality were mapped in an Asiatic A-genome cotton (*G. arboreum*), only two of the 43 (4.65%) were found in common with the At subgenome of a tetraploid cotton, but eight of the 43 QTLs (18.60%) were mapped to the Dt subgenome (Ma et al. 2008). The reminders were not mapped in either subgenome. Another reason that more fiber development Unigenes were genetically mapped in Dt or D genome than in At may result from
housekeeping genes which are typically constitutive genes are required for the maintenance of basic cellular function, and they are expressed in all cells of an organism under normal and pathophysiological conditions (Peleman et al. 1989). These housekeeping genes are so important that they are very conservative and monomorphic between the two mapping parents, and thus they cannot be genetically mapped. This hypothesis agrees with another hypothesis that the At subgenome contained “favorable” alleles for fiber development prior to cultivation (Ma et al. 2008), and also consistent with report that subgenome At provides more fiber development housekeeping genes/QTLs than Dt, but subgenome Dt provides more transcript regulatory factors to improve fiber properties with both natural and human selection pressures (Xu et al. 2010).

**Implications of this study in future applications**

Successful breeding is dependent upon hybridization of complementary parental lines, followed by identification of the best performing progeny lines with the highest recombination for desired traits. Historically this process has been performed through observation of plant phenotypes, but now the use of DNA markers allows for MAS of parents and identification of progeny with the maximum accumulation of desired QTLs. The best combinations of the genes/QTLs from each chromosome would result from DNA recombination, thus no best combination if no recombination. This study focused on the recombination hotspot regions to study the evolution of cotton fiber development Unigenes. The observation that At and D genomes shared more fiber development Unigenes than did Dt and D genomes in recombination hotspot regions suggests that we could mine and incorporate more beneficial DNA from the D genome into the tetraploid Upland cotton for the improvement of other important agronomic traits such as fiber yield, disease resistance, and stress tolerance. This suggestion supports cotton
geneticists who believed that the benefits of sequencing the diploid D genome were greater than its potential as a tool and stepping stone to sequencing the more complex tetraploid genomes of commercial cottons. The benefit of sequencing the D genome first among the *Gossypium* genus may outweigh its selection merely because of its small genome size. Subsequently with the recently finished sequence of A genome cotton (Li *et al.* 2014), what we have observed in this study could be further verified by comparative genomic analysis between Upland cotton’s tetraploid At and Dt subgenomes with both diploid A and D genomes.

**Conclusions**

Knowledge about the genomic distribution of fiber development in cultivated tetraploid cottons is essential for fiber improvement, but research is complicated by two different subgenomes (At and Dt) that control fiber initiation, elongation, and maturation processes. This study reveals that the two subgenomes do not contribute equally to the various phases of fiber development. The Dt subgenome has provided many fiber development genes after its merger with the At subgenome but the latter has greater similarity of fiber genes with the ancestor D genome. The findings are unexpected, due to the fact that the D genome does not produce any spinnable fiber and it is the most probable donor of the Dt subgenome. These insights into the evolution of the cotton genomes and the distribution of fiber development genes will guide efforts to exploit and manipulate important genes enhancing fiber properties. Identification of chromosome regions associated with fiber genes will allow for accelerated and efficient breeding through MAS for the maximum accumulation of desired genes.

**Acknowledgment**

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors. This research was supported by USDA-ARS CRIS Project 3091-
21000-038-00D. Mention of trade names or commercial products in this manuscript is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The U. S. Department of Agriculture is equal opportunity provider and employer.

References


CHAPTER 7. GENERAL CONCLUSION

Evidence from biotic and abiotic traits study from this research concluded that machine learning has higher prediction accuracy than traditional statistics. Machine learning analytics is a promising tool and will play a key role in plant breeding in the post-genomic era.