Application of Machine Learning and Hyperspectral Imaging in Plant Phenomics Research

Kshitiz Dhakal

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Crop and Soil Environmental Sciences

Song Li, Committee Chair
Joseph C. Oakes
Gota Morota
Bo Zhang
Bingyu Zhao

January 25, 2023
Blacksburg, VA

Keywords: Phenotyping, Drones, Machine Learning, Edamame, Hyperspectral Imaging

Application of Machine Learning and Hyperspectral Imaging in Plant Phenomics Research

Kshitiz Dhakal

ABSTRACT (ACADEMIC)

The digital imaging technology, spatial analyses tool, and computer vision methods can be used to extract traits related branching pattern, canopy cover, and pod location in edamame in high throughput fashion. Using genome-wide association study, we identified several single nucleotide polymorphisms (SNPs) that were associated with those traits, which can be used in marker-assisted selection to develop the edamame varieties that are more adaptable to mechanical harvesting and give more yield, along with understanding the physiological mechanisms for better shoot architecture traits and better yield. Spectroscopy-based machine learning method can be used to identify the optimal harvest time of edamame. Hyperspectral imaging combined with computer vision and machine learning methods can be used to quantify the levels of Deoxynivalenol (DON) in wheat kernels in high through put fashion

Application of Machine Learning and Hyperspectral Imaging in Plant Phenomics Research

Kshitiz Dhakal

ABSTRACT (GENERAL AUDIENCE)

The digital imaging technology, geographical analyses tool, and computer vision (a technique that enables computers and systems to get meaningful information from images) methods can be used to extract traits-related branching pattern, canopy cover, and pod location in edamame for many plant populations in short time using less labor and resources. Using genome-wide association study, we identified several genetic markers that were associated with those traits. These markers can be used in marker-assisted selection to develop the edamame varieties that are more adaptable to mechanical harvesting and give more yield, along with understanding the physiological mechanisms for better shoot architecture traits and better yield. We used spectral signatures of different edamame at several harvesting time along with machine learning methods to identify the optimal harvest time of edamame. Hyperspectral imaging (a technique that analyzes a wide spectrum of light instead of just assigning primary colors (red, green, blue) to each pixel) when combined with computer vision and machine learning methods can be used to quantify the levels of vomitoxin (chemical that causes vomiting and feed refusal in animal and humans) for larger wheat kernel samples in a cheaper and faster way.

DEDICATIONS

To my parents, and my lovely wife. I love you all.

ACKNOWLEDGMENTS

Firstly, I extend my deepest gratitude to my major advisor Dr. Song Li for his advice, support, valuable comments, expertise, faith, and encouragement. I feel very blessed and lucky to be a part of his research team. Without his guidance and suggestions, this journey would have been much tougher. I place on record, my sincere thanks to all my committee members Drs. Gota Morota, Joseph C. Oakes, Bingyu Zhao, and Bo Zhang.

To my first American advisor, Dr. Barbara E Liedl. at WVSU for always believing in me. For showing me how fulfilling a career in research can be and for showing me what plant breeding is. I would like to thank all my lab mates Jiyoung Lee, Qian Zhu, James R. Friel, Qi Song, Qi Li, Missi Zhang, and Kassaye Belay for assisting me with field and lab work; Patricia Donovan, and Carrie Edwards for introducing me with the remote sensing resources; lab members from Zhang Lab for helping me during the data collection at Kentland; Aashish Poudel for helping me with the codes during and after the class projects.

I extend my sincere appreciation to my parents, Mr. Meghraj Dhakal & Mrs. Meena Dhakal. Whatever I have achieved to date is all because of their blessings, enduring love, hard work, sacrifice, and trust in me that I can achieve something in life. Thanks for always being there for me during my thick and thins. I have no words to express how thankful I am to my dear wife, Mrs. Paru Gautam. Even when I was low, her faith in me has always helped me rise. I would not have made it this far without her continuous support, effort, time, guidance, love, and encouragement. Also, thank you so much for backing me up in my field work whenever I was short on help. Lastly, I am immensely thankful to God, who has always showered me with his blessings throughout my journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

High-throughput Phenotyping (HTP) for plant breeding and research in the post-genomic era

The global crop yields are expected to be doubled by 2050 to feed not only the rapidly increasing global population but also the current population who are shifting their diets to meat and dairy consumption. There is also an increasing demand for biofuels consumption to reduce the dependency on fossil fuel (Ray et al. 2013). It has been difficult to increase rate of the food production to catch up with the geometrically increasing world population and changing dietary shift (Cassman 1999; Hafner 2003; Peltonen-Sainio, Jauhiainen, and Laurila 2009). One way to increase the global productivity of major food and feed crops can be crop improvement via plant breeding which rely on both genotyping and phenotyping (Hickey et al. 2019; Schmidhuber and Tubiello 2007). The genotype-phenotype concept introduced by Johannson (1909), says that the overall constitution of an organism including all characteristics is a phenotype and it can be assessed by a multiple analytical methods such as morphology, physiology, anatomy and the chemistry of an organism.

Although there has been rapid advancement in genotyping but a lack of suitable phenotyping data (phenomics data) to understand the genetic contribution to the phenotypic variation has limited the progress in crop improvement programs (Junker et al. 2015). Precise and accurate phenomics data obtained by pairing with decreased labor input that can achieved with automation, remote control, and image analysis pipelines can abridge this ‘phenotyping gap’ by linking gene sequence to phenotypic variation in plant structure, development, composition, and performance (Cobb et al. 2013; Fiorani, Schurr, and others 2013; Furbank and Tester 2011; Houle, Govindaraju, and Omholt 2010). These modern phenomics tools can acquire the phenotypic traits

like plant development, architecture, growth, biomass, and photosynthesis, for hundreds to thousands of plants in a single day, hence referred to as high-throughput phenotyping (HTP) (Rahaman et al. 2015; Walter, Studer, and Kölliker 2012). Here we are discussing the two cases where we used HTP in this dissertation.

Case I: Using drones for canopy cover study and how that can benefit Edamame production such as increase yield and help with weed control

Unmanned Aerial Vehicles (UAVs) have been used in recent years for phenotypic data acquisition due to its improved spatial, spectral, and temporal resolution as well as being versatile, no-destructive method, and cost-beneficial for high-throughput plant phenotyping (Colomina and Molina 2014; Maimaitijiang et al. 2017; V Sagan et al. 2019; Vasit Sagan et al. 2019). Canopy cover is one of the important traits in plants which can naturally suppress weeds and has high genetic correlation with soybean grain yield (Casagrande et al. 2022; Cicek et al. 2006; Mansur et al. 1996; Sabbagh et al. 2020; Wilcox and Sediyama 1981). Canopy cover (CC) data have been successfully collected via UASs in multiple legumes (Cazenave et al. 2019; Sarkar et al. 2020; Xavier et al. 2017). Xavier (2017) performed time-series CC phenotyping among several recombinant inbred lines (RILs) of a soybean nested association mapping (SoyNAM) population and identified QTLs. Edamame, also known as green soybean or vegetable soybean, is a nutritious food source of protein, isoflavones, and vitamins (Lee et al. 2018; Mahoussi et al. 2020; Mentreddy et al. 2002). Edamame has become a popular snack food in the United States and many countries. Early weed control is important in edamame, and once the canopy covers the space between the rows, the weed naturally gets suppressed. Also, the edamame yield can be improved by improving canopy cover as the rate of photosynthesis is extremely sensitive to light intensity, which depends upon the spatio-temporal variability of canopy (Casagrande et al. 2022; Khan et al. 2015; Khush

2005; Reynolds, Van Ginkel, and Ribaut 2000; Rötter et al. 2015). In this dissertation, we collected CC data and performed GWASs to find the candidate genes that might be controlling the CC in edamame. This is discussed in the first and second section of Chapter II.

Case II: Tackling harvesting loss using images along with manual vs automatic traits measurements

Although United States is the major producer of grain soybeans, but most of the frozen edamame products consumed here are imported from Asia. The major hindrances for commercial production of edamame in the United States are the efficiency of mechanical harvest and the cost of hand harvesting (Lord et al., 2019; Mebrahtu and Mullins 2007). Research has found that mechanical harvest gives harvest efficiency between 54 and 85% for plant with 55–66 cm in height (Zandonadi, Coolong, and Pfeiffer 2010). In a more recent study performed by Dhaliwal and Williams (2020), it was found that with higher plant density, the number of branches and pod mass/vegetative mass ratio decrease whereas height and leaf area index increase for all varieties tested. Also, in the same study it was found that 86–95% of marketable pods can be harvested using the machine harvester. To improve edamame harvest efficiency, it is necessary to quantify the shoot architecture traits. The major challenge in plant phenotyping community is that it is difficult to quantify these traits. Using a mini-core collection of edamame, a topological approach called persistent homology (Li et al., 2017, 2019) was used to quantify the shoot architecture in topological space but it was time consuming it needed manual labelling of lots of plant parts. In this dissertation, we tested machine learning methods for obtaining branching pattern and pod location traits and performed GWASs on those phenotypic traits. This is discussed in the third section of Chapter II.

Sorting edamame pods using RGB images and spectral signatures of the pods

The ideal time of harvesting the edamame pods is sometime between R6 and R7 growth stages, just before pods beginning to turn yellow and when moisture and seed weight approach their maximum levels (Moseley et al. 2021; Yu et al. 2021). Harvesting edamame outside of its optimal harvest time can potentially threaten the marketability of the pods. For example, harvesting too early can lead to reduced yield, sweetness, and size of seeds, while harvesting too late leads to fibrous and yellow seeds (Carson 2010). RGB images have been used to determine harvesting maturity in pea pods (El-Raie et al. 2012). Also using spectral signatures, the fruit/pod maturity have been assessed in peanut (Zou et al. 2019), strawberry (Su et al. 2021), camelina (Jiang et al. 2022), and canola (Singh et al. 2021). In this dissertation, we tested machine learning methods for sorting edamame pods based on harvesting time using RGB images and spectral signatures as discussed in the first and second sections of Chapter III.

Using hyperspectral images for DON content quantification in wheat

Hyperspectral imaging has been used for the early identification of several diseases in several plants. Research on Sugarbeet for early detection and differentiation of Cercospora leaf spot, leaf rust and powdery mildew diseases based on spectral vegetation indices showed that the early differentiation between healthy and inoculated plants can be achieved (Rumpf et al., 2010). A study in soybean was carried out using hyperspectral imaging to diagnose charcoal rot and the model achieved a classification accuracy of 95.73% (Nagasubramanian et al., 2019). The VIS-NIR (400-900nm) HSI camera used to perform a large-scale screen for kernel in small grain demonstrated that HSI can classify kernels regarding whether DON toxin is above the consumption limit or not (Barbedo et al., 2018). This is discussed in Chapter IV.

RESEARCH OBJECTIVES

The specific objectives of the research in this dissertation were to:

1. To determine the genetic control of canopy cover in edamame with drone phenotyping.

2. To determine the genetic control of pod location and branching pattern with image analysis and machine learning.

3. To sort edamame pods based on disease, size, and maturity time using RGB images.

4. To sort edamame pods based on maturity time using spectral signatures.

5. To quantify DON content in wheat kernels using hyperspectral imaging.

REFERENCES

Carson, Luther C. 2010. "Cultivation and Nutritional Constituents of Virginia Grown Edamame." Virginia Tech.

Casagrande, Cleiton Renato, Gustavo César Sant'ana, Anderson Rotter Meda, Alexandre Garcia, Pedro Crescêncio Souza Carneiro, Maicon Nardino, and Aluizio Borem. 2022. "Association between UAV High-Throughput Canopy Phenotyping and Soybean Yield." *Agronomy Journal*.

Cassman, Kenneth G. 1999. "Ecological Intensification of Cereal Production Systems: Yield Potential, Soil Quality, and Precision Agriculture." *Proceedings of the National Academy of Sciences* 96(11):5952–59.

Cazenave, Alexandre-Brice, Kushendra Shah, Tresa Trammell, Michael Komp, Justin Hoffman, Christy M. Motes, and Maria J. Monteros. 2019. "High-Throughput Approaches for Phenotyping Alfalfa Germplasm under Abiotic Stress in the Field." *The Plant Phenome Journal* 2(1):1–13.

Cicek, Mine S., Pengyin Chen, M. A. Saghai Maroof, and Glenn R. Buss. 2006. "Interrelationships among Agronomic and Seed Quality Traits in an Interspecific Soybean Recombinant Inbred Population." *Crop Science* 46(3):1253–59.

Cobb, Joshua N., Genevieve DeClerck, Anthony Greenberg, Randy Clark, and Susan McCouch. 2013. "Next-Generation Phenotyping: Requirements and Strategies for Enhancing Our Understanding of Genotype--Phenotype Relationships and Its Relevance to Crop Improvement." *Theoretical and Applied Genetics* 126(4):867–87.

Colomina, Ismael, and Pere Molina. 2014. "Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 92:79–97.

Dhaliwal, Daljeet S., and Martin M. Williams. 2020. “Economically Optimal Plant Density for Machine-Harvested Edamame.” *HortScience* 55(3):368–73. doi: 10.21273/HORTSCI14642-19.

El-Raie, A. E., Y. A. Bader, H. E. Hassan, and R. Khamis. 2012. “COLOR IMAGE ANALYSIS FOR DETERMINING HARVEST TIME OF BEAN AND PEA PODS.” *Misr Journal of Agricultural Engineering* 29(1):307–20.

Fiorani, Fabio, Ulrich Schurr, and others. 2013. “Future Scenarios for Plant Phenotyping.” *Annu. Rev. Plant Biol* 64(1):267–91.

Furbank, Robert T., and Mark Tester. 2011. “Phenomics--Technologies to Relieve the Phenotyping Bottleneck.” *Trends in Plant Science* 16(12):635–44.

Hafner, Sasha. 2003. “Trends in Maize, Rice, and Wheat Yields for 188 Nations over the Past 40 Years: A Prevalence of Linear Growth.” *Agriculture, Ecosystems \& Environment* 97(1–3):275–83.

Hickey, Lee T., Amber N Hafeez, Hannah Robinson, Scott A. Jackson, Soraya Leal-Bertioli, Mark Tester, Caixia Gao, Ian D. Godwin, Ben J. Hayes, and Brande B. H. Wulff. 2019. “Breeding Crops to Feed 10 Billion.” *Nature Biotechnology* 37(7):744–54.

Houle, David, Diddahally R. Govindaraju, and Stig Omholt. 2010. “Phenomics: The next Challenge.” *Nature Reviews Genetics* 11(12):855–66.

Jiang, Hongzhe, Yilei Hu, Xuesong Jiang, and Hongping Zhou. 2022. “Maturity Stage Discrimination of Camellia Oleifera Fruit Using Visible and Near-Infrared Hyperspectral Imaging.” *Molecules* 27(19):6318.

Johannsen, Wilhelm. 1909. *Elemente Der Exakten Erblichkeitslehre*. Gustav Fischer.

Junker, Astrid, Moses M. Muraya, Kathleen Weigelt-Fischer, Fernando Arana-Ceballos, Christian Klukas, Albrecht E. Melchinger, Rhonda C. Meyer, David Riewe, and Thomas Altmann. 2015. “Optimizing Experimental Procedures for Quantitative Evaluation of Crop Plant Performance in High Throughput Phenotyping Systems.” *Frontiers in Plant Science* 5:770.

Khan, Mudasir Hafiz, Zahoor Ahmad Dar, Sher Ahmad Dar, and others. 2015. “Breeding Strategies for Improving Rice Yield—a Review.” *Agricultural Sciences* 6(05):467.

Khush, Gurdev S. 2005. “What It Will Take to Feed 5.0 Billion Rice Consumers in 2030.” *Plant Molecular Biology* 59(1):1–6.

Lord, Nick. n.d. *Administrator, 1890 Extension Program*.

Maimaitijiang, Maitiniyazi, Abduwasit Ghulam, Paheding Sidike, Sean Hartling, Matthew Maimaitiyiming, Kyle Peterson, Ethan Shavers, Jack Fishman, Jim Peterson, Suhas Kadam, and others. 2017. “Unmanned Aerial System (UAS)-Based Phenotyping of Soybean Using Multi-Sensor Data Fusion and Extreme Learning Machine.” *ISPRS Journal of Photogrammetry and Remote Sensing* 134:43–58.

Mansur, L. M., J. H. Orf, K. Chase, T. Jarvik, P. B. Cregan, and K. G. Lark. 1996. “Genetic Mapping of Agronomic Traits Using Recombinant Inbred Lines of Soybean.” *Crop Science* 36(5):1327–36.

Mebrahtu, Tadesse, and Chris Mullins. 2007. *Efficiency of Mechanical Harvest for Immature Vegetable Soybean Pods 1*. Vol. 58.

Moseley, David, Marcos Paulo Da Silva, Leandro Mozzoni, Moldir Orazaly, Liliana Florez-Palacios, Andrea Acuna, Chengjun Wu, and Pengyin Chen. 2021. “Effect of Planting Date and Cultivar Maturity in Edamame Quality and Harvest Window.” *Frontiers in Plant Science* 11:585856.

Peltonen-Sainio, Pirjo, Lauri Jauhiainen, and Ilkka P. Laurila. 2009. “Cereal Yield Trends in Northern European Conditions: Changes in Yield Potential and Its Realisation.” *Field Crops Research* 110(1):85–90.

Rahaman, Md Matiur, Dijun Chen, Zeeshan Gillani, Christian Klukas, and Ming Chen. 2015. “Advanced Phenotyping and Phenotype Data Analysis for the Study of Plant Growth and Development.” *Frontiers in Plant Science* 6:619.

Ray, Deepak K., Nathaniel D. Mueller, Paul C. West, and Jonathan A. Foley. 2013. “Yield Trends Are Insufficient to Double Global Crop Production by 2050.” *PloS One* 8(6):e66428.

Reynolds, Matthew P., Maarten Van Ginkel, and Jean-Marcel Ribaut. 2000. “Avenues for Genetic Modification of Radiation Use Efficiency in Wheat.” *Journal of Experimental Botany* 51(suppl_1):459–73.

Rötter, R. P., F. Tao, J. G. Höhn, and T. Palosuo. 2015. “Use of Crop Simulation Modelling to Aid Ideotype Design of Future Cereal Cultivars.” *Journal of Experimental Botany* 66(12):3463–76.

Sabbagh, Manuel J., Sindhu Jagadamma, Lori A. Duncan, Forbes R. Walker, Jaehoon Lee, Michael E. Essington, Prakash Arelli, and Michael J. Buschermohle. 2020. “Cover Crop Diversity for Weed Suppression and Crop Yield in a Corn--Soybean Production System in Tennessee.” *Agrosystems, Geosciences \& Environment* 3(1):e20112.

Sagan, V, M. Maimaitijiang, P. Sidike, M. Maimaitiyiming, H. Erkbol, S. Hartling, K. T. Peterson, James Peterson, J. Burken, and FUAV Fritschi. 2019. “UAV/Satellite Multiscale Data Fusion for Crop Monitoring and Early Stress Detection.”

Sagan, Vasit, Maitiniyazi Maimaitijiang, Paheding Sidike, Kevin Eblimit, Kyle T. Peterson, Sean Hartling, Flavio Esposito, Kapil Khanal, Maria Newcomb, Duke Pauli, and others. 2019. “UAV-Based High Resolution Thermal Imaging for Vegetation Monitoring, and Plant Phenotyping Using ICI 8640 P, FLIR Vue Pro R 640, and Thermomap Cameras.” *Remote Sensing* 11(3):330.

Sarkar, Sayantan, Alexandre-Brice Cazenave, Joseph Oakes, David McCall, Wade Thomason, Lynn Abbot, and Maria Balota. 2020. “High-Throughput Measurement of Peanut Canopy

Height Using Digital Surface Models." *The Plant Phenome Journal* 3(1):e20003.

Schmidhuber, Josef, and Francesco N. Tubiello. 2007. "Global Food Security under Climate Change." *Proceedings of the National Academy of Sciences* 104(50):19703–8.

Singh, Keshav D., Hema S. N. Duddu, Sally Vail, Isobel Parkin, and Steve J. Shirtliffe. 2021. "UAV-Based Hyperspectral Imaging Technique to Estimate Canola (Brassica Napus L.) Seedpods Maturity." *Canadian Journal of Remote Sensing* 47(1):33–47.

Su, Zhenzhu, Chu Zhang, Tianying Yan, Jianan Zhu, Yulan Zeng, Xuanjun Lu, Pan Gao, Lei Feng, Linhai He, and Lihui Fan. 2021. "Application of Hyperspectral Imaging for Maturity and Soluble Solids Content Determination of Strawberry with Deep Learning Approaches." *Frontiers in Plant Science* 1897.

Walter, Achim, Bruno Studer, and Roland Kölliker. 2012. "Advanced Phenotyping Offers Opportunities for Improved Breeding of Forage and Turf Species." *Annals of Botany* 110(6):1271–79.

Wilcox, J. R., and Tuneo Sediyama. 1981. "Interrelationships among Height, Lodging and Yield in Determinate and Indeterminate Soybeans." *Euphytica* 30(2):323–26.

Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. "Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine Max." *Genetics* 206(2):1081–89. doi: 10.1534/genetics.116.198713.

Yu, Dajun, Tiantian Lin, Kemper Sutton, Nick Lord, Renata Carneiro, Qing Jin, Bo Zhang, Thomas Kuhar, Steven Rideout, Jeremy Ross, and others. 2021. "Chemical Compositions of Edamame Genotypes Grown in Different Locations in the US." *Frontiers in Sustainable Food Systems* 5:620426.

Zandonadi, Rodrigo, T. Coolong, and T. Pfeiffer. 2010. "Mechanical Harvesting of Edamame." *SARE Proj Final Rep. Available Online at: Https://Www. Uky. Edu/Ccd/Sites/Www. Uky. Edu. Ccd/Files/Edamame_mechanical_harvest. Pdf (Accessed February 7, 2021)*.

Zou, Sheng, Yu-Chien Tseng, Alina Zare, Diane L. Rowland, Barry L. Tillman, and Seung-Chul Yoon. 2019. "Peanut Maturity Classification Using Hyperspectral Imagery." *Biosystems Engineering* 188:165–77.

CHAPTER II

BRANCHING PATTERN, CANOPY COVER, AND POD LOCATION STUDY IN EDAMAME USING HIGH THROUGHPUT PHENOTYPING

Chapter II Section 1: Analysis of Shoot Architecture Traits in Edamame Reveals Potential Strategies to Improve Harvest Efficiency

Kshitiz Dhakal1, Qian Zhu1, Bo Zhang1, Mao Li2* and Song Li1*

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

[2]Donald Danforth Plant Science Center, St. Louis, MO, United States Department of Computer Science, Virginia Tech

Note: This first look version of this manuscript was published in Frontiers of Plant Science on 3rd March 2021 in Crop and Product Physiology section DOI: https://doi.org/10.3389/fpls.2021.614926

Citation: Dhakal, K., Zhu, Q., Zhang, B., Li, M., & Li, S. (2021). Analysis of shoot architecture traits in edamame reveals potential strategies to improve harvest efficiency. Frontiers in Plant Science, 12, 614926. Doi: https://doi.org/10.3389/fpls.2021.614926

ABSTRACT

Edamame [Glycine max (L.) Merr.] is a type of green, vegetable soybean and improving shoot architecture traits for edamame is important for breeding of high-yield varieties by decreasing potential loss due to harvesting. In this study, we use digital imaging technology and computer vision algorithms to automatically characterize major traits of shoot architecture for edamame. Using a population of edamame PIs, we seek to identify underlying genetic control of different shoot architecture traits. We found significant variations in the shoot architecture of the edamame lines including long-skinny and candle stick-like structures. To quantify the similarity and differences of branching patterns between these edamame varieties, we applied a topological measurement called persistent homology. Persistent homology uses algebraic geometry algorithms to measure the structural similarities between complex shapes. We found intriguing relationships between the topological features of branching networks and pod numbers in our plant population, suggesting combination of multiple topological features contribute to the overall pod numbers on a plant. We also identified potential candidate genes including a lateral organ boundary gene family protein and a MADS-box gene that are associated with the pod numbers. This research provides insight into the genetic regulation of shoot architecture traits and can be used to further develop edamame varieties that are better adapted to mechanical harvesting.

INTRODUCTION

Edamame is a type of green, vegetable soybean which has become a popular food ingredient in many countries because it is a nutritious food source of protein, isoflavones, and vitamins (Lee et al. 2018; Mahoussi et al. 2020; Mentreddy et al. 2002). Edamame has been cultivated in east Asian countries for more than 2000 years and documented edamame varieties have been mainly originated from this area (Shurtleff, Aoyagi, and others 2009). In recent years,

production and breeding of locally adapted edamame varieties have been reported in North and South America, Europe, and Africa (Konovsky, Lumpkin, and McClary 1994). The yield components of soybeans have been studied and include plant density, number of pods and number of seeds per pod and seed size (Liu et al. 2010; Ulloa et al. 2010). However, little is known about how these components affect edamame yield, because the yield is evaluated when the seeds are at an immature stage.

There are several major differences between edamame and grain soybeans. First, edamame is harvested when the pods are fully filled while beans are still green with high level of moisture and sugar content (Shanmugasundaram et al. 1991). In contrast, grain soybeans for feed and oil are typically harvested when the pods and beans are dry. Second, due to consumer preference, edamame seeds are much larger than grain soybean seeds (Carson et al. 2010). Because of these key differences, grain soybean varieties cannot be directly used for edamame production and optimization of additional traits are needed to produce new edamame varieties that are better accepted by the producers and the consumers. In the United States, despite being a major producer of grain soybeans, most frozen edamame products were imported from Asia. The main obstacles for commercial production of edamame in the United States are the efficiency of mechanic harvest and the cost of hand harvesting where manual harvesting is still a common practice for small farmer (Lord, 2019; Mebrahtu and Mullins 2007).

A number of studies have been performed to test commercial harvesters on edamame. For example, a common bean harvester, Oxbo BH100 were tested to harvest edamame and the results were compared to hand harvesting (Mebrahtu and Mullins 2007). It was found that hand harvesting generated twice as many pods as compared to mechanical harvest. However, mechanical harvest has generated cleaner products. Mechanical harvest was found to give the best results for plants of

55-66 cm in height. Harvest efficiency of the same type of harvester was tested on three edamame varieties and the harvest efficiency is between 54% and 85% (Zandonadi, Coolong, and Pfeiffer 2010). The speed of harvester does not affect the harvest efficiency when it was below 2 miles per hour. In a more recent study, four cultivars of edamame were used to study the optimal plant density of edamame (Dhaliwal and Williams 2020) and these varieties were harvested by the same bean harvester. This research showed, with higher plant density, the number of branches and pod mass/vegetative mass ratio decrease whereas height and leaf area index increase for all varieties tested. In particular, for the same variety, the main stem branch changes from 6 to 1 with increasing plant density. Using the mechanical harvester, it was found that 86%-95% marketable pods can be harvested (Dhaliwal and Williams 2020).

A number of environmental factors are known to affect pod numbers and plant architecture in soybeans and edamame. In a comparison of determinant and indeterminant varieties (Egli 1993) , it was found that 85% of pods were initiated before stage R5. R5 stage is one of the reproductive stages of the edamame when seeds begin to develop (Fehr and Caviness 1977). At this stage, the seed is 3mm in size, which develops inside a pod at one of the four uppermost nodes on the main stem with a fully developed leaf. Indeterminant varieties have longer pod production period for approximately 50 days. In a test of maturity of soybeans, late mature groups have more nods and more pods per plant as compared to early mature groups (Zhang and Kyei-Boahen, 2007). The photoperiod is a major factor that affects the number of pods in soybeans where the long photoperiod mainly affects pod number during the R3-R6 stage (Kantolic and Slafer 2007). Long days also delay flower to pod transition and seed filling, but it does not affect pod elongation (Nico et al. 2016). In addition to photoperiod, higher temperature can also contribute to higher number of flowers and pods, but these flowers may fail to produce mature pods and cause a reduction of

yield (Kim et al. 2020). A multi-year study of edamame breeding lines shows that there are significant trait variations between years, including changes in pod yield and plant height, suggesting environmental variation is a key factor for edamame development (Jiang, Rutto, and Ren 2018).

At molecular and genetic level, many key genes in soybeans related to the shoot architecture traits have been identified. Soybeans can be categorized into three types of stem growth habits: determinate, indeterminate and semi-determinate growth. Two genes, Dt1 and Dt2, are known to regulate this process in soybean (Liu et al. 2010; Ping et al. 2014; Tian et al. 2010; Zhang et al. 2019). GmDt1 is a homolog to Arabidopsis terminal flower 1 (TFL1) and GmDt1 in cultivated soybeans confers determinate growth habit. GmDt2 is a MADS-box transcription factor which represses GmDt1 expression and confers semi-determinate growth (Ping et al. 2014) (Ping et al., 2014). In addition to transcription factors, microRNAs, in particular, gmmiR156b has been shown to regulate soybean shoot architecture. Over expression of this microRNA lead to a 100% increase of branches without changing plant height. Pod per plant is also increased more than 30% without affecting seed protein and oil content (Sun et al. 2019). Using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), double mutant of GmSPL9a/b, the target gene of gmmiR156b, showed similar phenotype of increased branch numbers as found in gmmiR156b over expression lines (Bao et al. 2019). Besides these well characterized molecular pathways that regulate soybean shoot architecture, genome wide association studies have also identified many Genome-Wide Association Studies (GWAS), Quantitative Trait Locus (QTLs)that are associated with shoot architecture traits such as plant heights, branch numbers and pod numbers (Fang et al. 2017; Hao et al. 2012). Using pod number as one example, these published studies have identified 26 QTL loci and 9 candidate genes that are associated with pod number variations in soybeans.

The soybean research and breeding community have accumulated large amounts of genomics resources including more than 20,000 plant introductions (PIs) that were genotyped by 50K SNP array (Song et al. 2013) and over 3,000 PIs with full genome sequences available (Liu et al. 2020). To leverage this genetic diversity and genomic resources in edamame breeding to improve harvest efficiency, the major challenge lies in phenotyping. High throughput phenotyping in soybean has been used to study leaf shape (Chen and Nelson, 2004), root architecture (Fenta et al. 2014) and canopy cover (Xavier et al. 2017). In this work, we develop a phenotyping pipeline to collect images for edamame at harvest stage of R6 to R7 and to quantify major shoot architecture traits related to harvest efficiency including plant height, branching patterns, pod numbers and pod locations. We also collected canopy cover data over the growth season to quantify and correlation of canopy cover with other shoot architecture traits. We applied a topological approach called persistent homology (Li et al. 2019; Li Mao, Duncan Keith, Topp Christopher N. 2017) to quantify the shoot architecture in topological space and correlate the topological traits with geometric traits of edamame shoots. Using a mini-core collection of edamame, we explore the correlation between geometric traits and topological traits and test whether known markers for pod numbers are associated with the traits observed in our data. Our results provide a scalable pipeline of shoot architecture phenotyping and provide novel candidate markers and genes for improving shoot architecture traits in edamame.

METHODS

Plant materials and shoot image collection

A total of 151 soybean PIs with > 20g/100 seeds that are potential parental lines for developing edamame varieties (referred as edamame PIs) were sown in 3 meters row and 0.75 m row spacing (with a seeding rate of ~70,000 plants per hectare) arranged in a complete randomized

design with two to four replications in Kentland farm at Blacksburg, VA in 2019. We selected these 151 PIs in our collection and two to four replicates (plants) per PI (540 plants) were harvested by cutting them from the soil line using a bypass looper (large secateur). The leaves and petioles were taken off of the plants before they were taken to the imaging station. The imaging station consisted of a black background, inch tapes at the borders, a camera tripod, and a digital single-lens reflex (DSLR) camera. The entry names and sample numbers of the plants were printed as a barcode on an iPad and captured by the camera. Images were captured from both sides of the plants (see details in next section). We have generated 1202 images for all plants that were harvested. Based on a preliminary analysis of all images, we selected 178 images from 24 edamame PIs for this study because these images showed diverse phenotypic traits such as plant heights and branching patterns and all 24 varieties have been genotyped using 50K SNP array. These 24 PIs showed diverse heights from dwarf to tall. The branching pattern was diverse ranging from one branch to several branches and the shape was also varying from candle shaped to one straight branched. A list of 24 selected PIs and traits measured in our analysis is provided as supplementary table 1 (Table S1).

Drone image collection and analysis

A DJI Phantom 4-Advanced was used for canopy cover study during the 2019 growth season from May 2019 to September 2019. A total of 1853 images were collected during this growth season with an average of 120 drone images collected for each flight day. Drone flight waypoints were generated using an iPad app, DroneDeploy. Flight height was set to 30.5 meters (~100feet) above ground level. Side overlap and front overlap were set at 75% with padding. Ground control points (GCPs) were used and the precise GPS location of GCPs were determined using a Real-time kinematic (RTK) GPS. Orthorectified drone maps were generated using AgiSoft

Metashape professional addition (Version 1.6) and subplot extraction were done manually using ArcGIS pro software. Canopy cover was extracted and average across replicates to generate a growth curve for each variety and the results were compared and correlated with other shoot architecture traits.

Image analysis and characterization of topological and geometric traits

Two to eight images from each PIs were used for image annotation using ImageJ (Rueden et al. 2017) and ImgLabel (https://github.com/tzutalin/labelImg) Software. For each plant two images were taken from both sides of the plant. Because of the bilateral symmetry of edamame plants, for each PI, a plant is placed on a black background with branches laying on a flat surface to take the first photo and the plant is flipped to take another photo (Figure 2.1.S1A). These images were analyzed using the ImageJ program with a custom plugin script to label all the branches (Figure 2.1.S1B). The branches were then analyzed using script developed in Matlab to convert the labeled images into a network of branches with vertices representing locations of landmarks used in the labeling processes. The main branch and side branches were labeled separately which allows post processing to calculate the branch length separately and to identify internodes in the branch networks. The correlation between geometric parameters measured in the photos was tested using cor.test function in R to test for Pearson's product moment correlation coefficient based on fisher's Z transformation. Pods in each image were also labeled manually using ImgLabel, which generated an Extensible Markup Language (XML) file for each labeled image and the XML file contains all the x and y coordinates for the labeled pod locations (Figure 2.1.S1C). The top and bottom of each plant were also labeled using ImgLabel program. A python script was developed to process the XML files to extract traits including pod numbers, pod locations, plant height and pixel per center meter from the images. Each plant was imaged and labeled twice, and the results

were averaged for the final analysis. The primary branches and the main branch for each plant were also detected using Matlab script and manual curation (Figure 2.1.S1D). Each primary branch was represented by a path (a sequence of edges which join a sequence of vertices). The primary branch length was calculated by adding up the length of all the edges of this path. Density plots were generated using plot density function in ggplot2 package in R.

To calculate the topological similarities between different branching networks, we first calculated the geodesic distance from all the vertices on the branches to the bottom of the main branch. A persistence barcode was generated for each image following the published approach(Li et al. 2019; Li Mao, Duncan Keith, Topp Christopher N. 2017). Pairwise distance between different barcodes were calculated using bottleneck distance (Cohen-Steiner, Edelsbrunner, and Harer 2007) and multi-dimensional scaling (MDS) were used to perform dimension reduction in this pair-wise similarity space to obtain the coordinates of the first three MDS dimensions. Only the first three dimensions were used in our analysis and other dimensions were ignored in this analysis because these lower rank dimensions provide limited variation regarding the overall similarity between different branching networks. We used Euclidean MDS-PCA space to approximate the non-linear topological space. The percentage numbers calculated from variation of PCA from the MDS results are the estimation of the variation. Correlations between different traits and heatmap were generated using R programming language and pheatmap package (https://cran.r-project.org/web/packages/pheatmap/). Matlab codes are provided in our GitHub repository (https://github.com/maoli0923/Edmame-Shoot-Architecture).

Terminology used for shoot architecture analysis

Although a few excellent review papers have described the shoot architecture of many plant species (Alonge et al. 2020; Benlloch et al. 2015; Teichmann and Muhr 2015; Wang and Jiao

2018), there is no commonly accepted terminologies for edamame and soybean plants in order to provide detailed description of the branching patterns that we are aiming to study. Therefore, we provide a schematic diagram to illustrate the terminology used in our analysis (Figure 2.1.S2). Those terminologies are as follows: plant height (PH), pod number (PN), first pod height (FPH), multidimensional scaling 1 (MDS1), multidimensional scaling 2 (MDS2), multidimensional scaling 3 (MDS3), average primary branch length (APBL), main branch length (MBL), total branch length (TBL), first node height (FNH), pod number above 10 cm from ground (PN10), height above ground for 5% pods (P5H), height above ground for 1% pods (P1H), first internode length (FIL), second internode length (SIL), third internode length (TIL), number of primary branch (NPB). In the diagram, lines with arrowheads represent branches and circles represent flowers/pods (Figure 2.1.S2A). The main branch in our terminology is sometime called main stem in other publications. All edamame varieties in our study first generated several primary branches on the main branch before producing pods on the main branch. Primary branches are the side branches that directly emerged from the main branch and secondary branches are those initiated from the primary branches. In our data, only a few varieties generated secondary branches such that we did not include secondary branches in our analysis. Detailed descriptions of the terminology are included in the schematic diagram in Figure 2.1.S2.

Genetic data analysis

GWAS QTL markers were downloaded from soybase.org (Grant et al. 2009). Three published GWAS studies (Fang et al. 2017; Hao et al. 2012; Zhang et al. 2015) analyzed the shoot architecture traits in soybeans and these publications provided the marker names or candidate genes for these markers. Only statistically significant markers from these publications were used for our analysis. Because different studies used different genotyping approaches, we try to match

the markers used in our study (50K SNP array) to markers used in other publications by determining the genomic locations of these markers on the same reference genome. For the known markers that are associated with pod numbers, we first identified their location in a recent soybean genome release (Wm82.a2.v1). We then identify those 50K SNP array markers that are closest to these published markers (within 50kb). In most cases, we can find associated markers within 10kb from the published markers and in some cases, there are multiple markers located within our predefined genomic range. Marker data were downloaded from soybase.org as a Variant Call Format (VCF) file and the genotypes were summarized per plant based on whether the plants were having pod numbers higher than average or lower than average (Table 2.1.1). The association of markers with the pod number trait were tested using fisher's exact test (p value < 0.05). Candidate genes were identified as those genes that are most close to the significant Single Nucleotide Polymorphism (SNP) markers. In case the marker is located in a gene dense region, the gene functions were manually selected based on their homology to other plant genes that are more likely to be involved in regulating pod numbers.

RESULTS

Overview of phenomics analysis pipeline

To understand the genetic control of shoot architecture in edamame plants, we used a mini-core collection of 151 edamame PIs with maturity group (MG) IV and V that are adapted to the growth conditions in Virginia as our model population (Figure 2.1.1). Maturity group zones are defined as the areas where a cultivar is best adapted. MG IV and V are best adapted to the growth conditions of most of the southern states and in Virginia (Egli 1993; Mourtzinis and Conley 2017). For each of these 151 PIs, we have collected two types of image data. To study the shoot architecture and pod locations, we harvested 2-4 plants per PI and removed all leaves and petioles

before imaging (Figure 2.1.1, step 2). Because the shoot of edamame is bilateral-symmetrically distributed, we only need two photos for the "front" and "back" of each plant to capture the variations in the branching patterns. We also collected drone images to study edamame canopy coverage over the growth season (Figure 2.1.1, step 3). From these 151 PIs, we selected 24 varieties for detailed characterization because of their diversity in the genomic sequences as well as phenotypic variations. These images were manually labeled to identify the location of each pod with ImgLabel software, and also to trace the branches using a modified ImageJ plugin (Rueden et al. 2017). For each plant, 20 phenotypic traits, including length of the main branch, number of primary branches and number of pods were measured. The terminologies used here are described in detail in the material and method section and in supplementary figure 2.1.S2. We further translated the branching patterns using a topological approach called persistent homology and projected the topological pattern into lower dimensional space (Li Mao, Duncan Keith, Topp Christopher N. 2017). Finally, we studied the trait correlations between the visible traits measured on images with topological traits (Figure 2.1.1, step 5). To understand the potential genetic control of these traits, we analyzed the published SNP map of these 24 varieties and studied whether some known major QTLs control shoot architectures are candidate regulatory regions in these varieties.

Correlation of shoot architecture parameters between technical replicates

The shoot architecture and phyllotaxis of edamame has not been extensively documented before. Our approach of phenotyping is to take images of the whole plant on a flat surface from two sides. We hypothesize that this approach can capture most variations of branching patterns. To test this hypothesis, we manually analyzed 178 photos (89 pairs of images) from 24 varieties of edamame and measured 12 geometric traits related to the branching patterns and shoot architecture (Figure 2.1.2A). Among these parameters, we found that five parameters showed high

correlation between the images taken on both sides of the same plant. These parameters include plant height (PH, Figure 2B), main branch length (MBL, length of longest branch in cm), pod numbers (PN, total number of pods), total branch length (TBL, the sum of lengths of all branches) and average primary branch length (APBL, the sum of lengths of all branches divided by the total number of branches). The high correlation of these parameters between two images of the same plant is expected and suggested that only taking one image on one side of the plant is sufficient to capture the variations of these parameters in our plant population. Here technical replicates represent the two images of the same plant taken from two different sides as the edamame plant is roughly bilateral symmetry. First pod height (FPH) is a parameter that is related to harvester efficiency (Mebrahtu and Mullins 2007), and this parameter showed lower correlation (0.86) than plant height. This is likely due to the fact that branches are flexible and when flipping the plants while taking the images, some branches can change their position. Parameters such as branch length will not be changed but the location of the pod relative to the bottom of the plants will be affected.

Some other parameters, such as first internode length (FIL, Figure 2C) and second internode length (SIL) showed lower correlation of 0.63 and 0.60 respectively, but the correlations are still statistically significant. The only parameter that is not significantly correlated between the two images of the same plant is the third internode length (TIL, Figure 2A). Interestingly, these parameters are not affected by the position of the branches. There are two reasons that might explain these lower correlations. First, some outlier observations (Figure 2C) could reduce the overall correlation. Second, some internodes are very short and the precise location where the primary branches connected to the main branch could be blocked on one side of the plant but more visible on the other side of the plant. These results suggest that for the majority of shoot

architecture parameters, our approach of image analysis can provide accurate measurements. Cautions should be taken when trying to interpret results from FIL, SIL and TIL which showed variable results which were affected by the image analysis process.

Distribution of shoot architecture parameters

To understand the shoot architecture of edamame plants, we analyzed the distribution of shoot architecture parameters of our plant population (Figure 2.1.3). We first focused on parameters related to plant height and branch length (Figure 2.1.3A), and we found that the distribution of PH is almost identical to the MBL. There is a small shift towards longer length when measuring MBL as compared to PH. This is expected because we measure plant height by measuring the distance from the ground to the top of the main branch. For some edamame varieties, main branches may have small angles at each internode, therefore overall length of the main branch is highly similar to the overall height of the plant with some varieties have longer MBL than PH. The average MBL and PH are 55.7 and 54.4 cm, respectively. The average PH is shorter than data generated by measuring PH (68-81 cm) in the field conditions (Jiang et al. 2018; Zhang et al. 2015) and this difference is likely since we removed all petioles from the plants before the measurement. Petioles of edamame are very long (~30 cm) and contribute significantly to the height of the plant if the measurement is taken at field when leaves are still green and before plant reaches full maturity. Another possibility is that different soybean varieties were used in published studies. The average primary branch length is approximately half of the main branch length (Figure 2.1.3A, see method section). This result shows that the major contributor of the plant height for edamame is the main branch length and other primary branches are shorter than the main branches on average.

To understand where primary branches emerged from the main branch, we have measured four parameters: first node height, first internode length, second internode length and third

internode length (Figure 2.1.3B). The first node height is the length from ground to node where first primary branch meets the main branch. The first internode length is the distance from where the first primary branch meets the main branch to where the second primary branch meets the main branch. Second and third internode length are similarly defined. There are cases where we found very short internodes and such short internodes were ignored in our analysis but were included as one feature in our phenotypic analysis (Supplementary Table 2.1. S1). Such a short internode is a known feature for some soybean plants and whether such a short internode is genetically controlled is still not well understood (Yoshikawa et al. 2013).

Our results (Figure 2.1.3B) show that the average first node height is 5.47 cm, which is almost twice the length of average first internode length (2.75 cm). The second and third internodes are relatively shorter than the first internode but are similar to each other with average lengths of 2.08 cm and 2.10 cm respectively. Once the first primary branch has emerged, the second and third primary branches will emerge subsequently after similar intervals. Note that there is a large variation in the height of first node, with 2.97 cm and 7.47 cm at first quartile and third quartile, respectively. This interquartile length is 5.8 times larger than the interquartile length of the first internode, showing a large variation on the development of the main branch before transition into producing primary branches.

To understand how pod production is correlated with length of plant branches, we generated the density distributions of three parameters: first node height, first pod height and 5% pod height. The 5% pod height is defined as the distance from ground when 5% of all pods were observed on a plant. We chose to measure 5% pod height because the lower 5% of pods are more likely to be lost during the harvesting process than other pods. The average first pod height is 9.72 cm whereas the average heigh of first internode is 5.47 cm. In fact, the average first pod height is

higher than the sum of first node and first internode (7.55 cm) and slightly lower than the second internode (10.30 cm). Interestingly, the density distribution of the first node height and first pod height both seem to show two peaks (red and green curves in Figure 2.1.3C). The separation of two peaks in the first pod height distribution is clear with one summit of the distribution at ~7 cm and the second summit at ~19 cm. Although these two distinct peaks were not clearly visible in the 5% pod height distribution, a wide distribution of this parameter is noticeable. These results suggest that plants in this study can be approximately categorized into two types according to there the first pod were produced.

We also analyzed the distribution of total pod number and compared the pod number above 10 cm from ground (Figure 2.1.3D). As expected, the two distributions are highly overlapping, with the average pod number above 10 cm from ground (green histogram) slightly lower than total pod number (red histogram), suggesting some varieties have pod located below 10cm from ground level. These close-to-ground pods are difficult to pick up by mechanical harvesting. Finally, we analyzed the change of canopy cover of these edamame varieties during the growth season with drone images from 35 days after planting (DAP) to 81 days after planting. The average canopy cover showed a steady increase over the growth season as expected. We selected 61DAP and 81DAP data from canopy cover data for downstream analysis. These dates were selected because 61DAP represents one of the early dates of canopy expansion and 81DAP represents one of the late dates of canopy expansion, respectively.

Persistent homology of shoot architecture uncovers hidden connection between branching patterns and plant productivity

Because different varieties of edamame have different numbers of branches and the directions of these branches can also vary, comprehensive comparison of the branching patterns between different varieties is challenging. To solve this problem, we applied a mathematical

approach called persistent homology (Carlsson 2009; Edelsbrunner and Harer 2010.; Verri et al. 1993) to convert the complex patterns on a 2D image into a topological space where branching patterns are comparable between different samples. Images of edamame plants were manually skeletonized and labeled as main branch and primary branches (see materials and methods section). Secondary branches were not included in this analysis. For each plant image, the branching skeleton was translated into a network representation and the geodesic distance (also the shortest path length) from vertices on each branch to the ground was calculated. A persistence barcode summarizing the branching topological information was generated (Li et al. 2019; Li Mao, Duncan Keith, Topp Christopher N. 2017) for each plant and the distance between different plants were calculated using bottleneck distance (Cohen-Steiner et al. 2007). Multidimensional scaling was used to project the distance between different branching patterns into low dimension space and only the first three dimensions were included in this analysis (MDS1, MDS2 and MDS3, see method section). These three dimensions explained 54. 3% of total variations. The top 30 dimensions explain 85% of total variations, however, the fourth dimension and above each explain a small fraction of the total variations such that they were not included in our analysis.

To understand the relationship among the low dimension projection of the persistent homology and other traits, we performed pair-wise correlation analysis with hierarchical clustering (Figure 2.1.4A). The clustering heatmap shows that the first pod height (FPH) is highly correlated with the height of 1% and 5% of total pods (P1H PCC=0.981 and P5H PCC=0.921, p value < 2.2e-16, Figure 2.1.S3A and 2.1.S3B). We also found that the canopy cover at 61DAP is highly correlated (Pearson correlation 0.87, p value < 2.2e-16, Figure 2.1.S3C) with canopy cover at 81DAP, which is consistent with field observations. However, canopy cover traits do not show strong correlation with any other single trait, suggesting combination of multiple shoot architecture

traits or other traits (petiole length or leaf surface area) that are not measured in our study may contribute to the canopy cover. Another pair of high correlation was found between total branch length (TBL) and average primary branch length (APBL, PCC = 0.550, p value < 1.72e-15, Figure 2.1.S3D), because the APBL equals TBL divided by number of branches.

With regard to topological traits (MDS1, MDS2 and MDS3), we found strong correlation between MDS1, plant height (PCC=0.936, p value < 2.2e-16, Figure 2.1.S3E) and main branch length (PCC=0.940, p value < 2.2e-16), suggesting that the major variation in the topological space is related to plant height. To confirm this correlation, we plotted the plant height on the two-dimension MDS plot (Figure 2.1.4B), and we indeed found that higher MDS1 corresponds to taller plant and lower MDS1 corresponds to shorter plants. Surprisingly, we found that MDS2 is positively correlated with pod number (PN, PCC=0.293, p-value < 7.03e-05, Figure 2.1.S3F). This is an intriguing observation because in our data processing pipeline, pods and branches are labeled separately (with different software, ImgLabel for pods and ImageJ for branches). In another word, the data were analyzed independently, but the analysis showed a positive correlation between these two traits. To confirm this relationship, we plotted the pod number on the two-dimension MDS plot (Figure 2.1.4C), and we found that small MDS2 indeed correlated with small number of pods and large MDS2 tends to have higher number of pods. However, some plants with the highest number of pods (PN = 140-160) do not have high MDS2, suggesting that additional variation cannot be explained by MDS2. Further analysis of the correlation map shows that MDS3 does not seem to explain this additional variation, but MSD3 is positively correlated with first node height (FNH, PCC=0.363, p-value < 6.42e-07, Figure 2.1.S3G), which is another branch length related trait.

A central question of this work is to investigate the distribution of pod on the plant

branches. This trait has been studied in soybeans (Illipronti et al. 2000; Ning et al. 2018; Liu et al. 2010) and several other crop species (Decoteau and Graham 1986; Kigel et al. n.d.). There are five traits related to pod distribution and yield, which include pod number (PN), pod number above 10cm (PN10), first pod height (FPH), P1H and P5H. H. Interestingly, pod numbers are negatively correlated with the other four traits related to pod distribution on the plant. Pod distribution is related to the harvest efficiency, and specifically, the first pod height is positively correlated with harvest efficiency using combine harvester (Beiküfner et al. 2019; Ramteke, Singh, and Murlidharan 2012).

The negative correlation (average PCC = -0.601, Figure 2.1.S3H) between pod number and first pod height indicates that the lower the pods are produced, the more pods a plant can produce. However, because of this negative correlation, a challenge is to increase the first pod height (thus improve harvest efficiency), without reducing total pod number per plant.

Pods that are close to the ground are more likely to be lost due to harvest than those that are away from the ground. PN also has negative correlations with plant height (PCC=-0.18, p-value < 0.021, Figure 2.1.S3I), first node height (PCC=-0.417, p-value < 6.68e-09) and several other traits that are related to plant statues. These results show there is a trend where taller plants tend not to produce as many pods as shorter plants in this edamame population under our experimental condition including plant density and local climate.

Genetic control for edamame shoot architecture

To investigate whether the shoot architecture traits are genetically controlled, we first calculated the heritability using linear mixed effect model (Nyquist and Baker 1991) with lme4 (Bates et al. 2014) package (Table 2.1.1) with correction for unbalanced number of replicates. We found high heritability (>70%) for most traits measured in our study. For example, plant height is

a trait that has been studied in many prior reports and the estimated heritability of plant height is 84%, which is similar to what has been reported in other soybeans (85%) and edamame (79%) populations (Chang et al. 2018; Jiang et al. 2018). Other traits not in other published work but analyzed in our study also showed high heritability. In particular, pod number and first pod height have the highest heritability of 91%. Interestingly, the topological traits such as MDS1 and MDS2 also have high heritability. MDS1 is highly correlated with plant height and has the same heritability as plant height. MDS2 is highly correlated with pod numbers and has slightly lower heritability as compared to pod numbers. These results suggest, under our experimental environment and plant density, there is evidence of genetic control of pod numbers and first pod height in our selected varieties of edamame.

Because producing large numbers for fresh pod is a major goal for edamame breeding, to further investigate the candidate genes that control the pod numbers, we collected known GWAS QTLs that are associated with pod numbers in soybeans. These known QTLs were downloaded from the Soybase and the SNPs that are associated with pod numbers were provided from three studies (Table S2). There are 26 SNPs/genomic locations that are associated with pod number from chromosome 1, 2, 5, 6, 9, 11, 13, 15, 17, 18, and 19. Nine candidate genes were provided by one of the publications and the other publications did not provide candidate genes. Because different studies have different SNP markers and we used the 50K SNP array data for our selected edamame lines, we identified the SNP markers in our marker data that are closely localized to the published SNP markers. We found 114 SNP markers in our genotyping data that are with 50Kb from these published markers. Using fisher's exact test, we determine the association of these SNP markers to the pod number traits (Table 2.1.2).

We found eight SNP markers in our population showed statistically significant association with pod numbers. For example, ss715609881 appears as the same as the reference allele in 12 varieties with low pod numbers (Table 2.1.2, marked by $), and in 1 variety that has high pod number (Table 2.1.2, marked by #). Low and high pod numbers are defined as pod numbers below or above average, respectively. A candidate gene, Glyma.11G164800, is located within 5 kb from this SNP marker, and this gene encodes a LOB (lateral organ boundary) protein. Although the function of this particular gene has not been characterized, members of this gene family have been shown to be related to flower and embryo development in Arabidopsis (Borghi, Bureau, and Simon 2007; Chalfun-Junior et al. 2005), maize (Evans 2007) and rice (Li et al. 2008). These results support a potential role of this candidate gene in regulating pod numbers in edamame.

Another example marker is ss715582578, which appears as the same as the non-reference alleles in 12 varieties with low pod number and in 6 varieties with high pod number (p value = 0.014). There are two candidate genes that are located within 10kb of this SNP marker (Table 2.1.2). One of the candidate genes downstream of this SNP marker (Glyma.02G216600) encodes homologous gene to AGAMOUS-like 16. Genes in the AGAMOUS gene family are well known for their functional role in floral development in plants including Arabidopsis (Mizukamil and Ma2 1997) and soybeans (Chi et al. 2017). These results suggest that Glyma.02G216600 might be a shoot architecture-related gene. The upstream gene (Glyma.02G216500) is a TRAF-type zinc finger-related transcription factor which is poorly characterized but can also be considered as a candidate gene because of its potential role of expression regulation. We analyzed published gene expression data comparing shoot apical meristems with leaves in soybeans (Wong, Singh, and Bhalla 2013). We found six genes are highly expressed in shoot apical meristems with $\log_2$ fold change higher than 1 (Table 2.1.S3), which indicates more than 2-fold up regulation of these genes

in the tissue type that is related to pod formation. These results further support the potential roles of these genes in regulating pod number in edamame.

DISCUSSION

With the decreasing cost of sequencing, many soybean varieties have been either genotyped using SNP arrays (Song et al. 2013) or whole-genome re-sequencing (Zhou et al. 2015), which provides a rich resource of genetic markers and potential functional genetic variations. Therefore, to fully utilize such genetic resources, one approach is to perform association studies to identify SNP markers for traits of interest. A bottleneck for such association study is the ability to collected phenotypic data for a large population with different genetic makeup at field scale. In soybean research, a large number of GWAS studies have been published including those studies canopy cover (Xavier et al. 2017) and shoot architecture traits (Fang et al. 2017; Zhang et al. 2015). With the increasing use of drones in field research, measuring canopy cover and plant height from drone images has become a preferred approach and has led to the discovery of many GWAS QTL that are associated with these traits (Mogili and Deepak 2018).

However, in our study, we are interested in the distribution of pods on the plant and how this trait is related to other shoot architecture traits. The phenotyping task for this study is challenging because the shoot structure is covered by leaves when edamame was harvested. Manually removing all leaves is the major time limiting step in our analysis. One alternative approach is to collect image data after all leaves are dropped (for example see ref (Sun et al. 2018)), which will be explored in future study. Another time-consuming step in our analysis is to cut the plants and lay the plants on a flat surface for imaging. An alternative approach would be to use 3D imaging or LiDAR to collect data in the field without cutting down the plants (Dhami et al. 2019; Sun et al. 2018) . Another major obstacle to extending this study is that manually labelling

all the branches and pod in each image is also a time and labor-intensive process. How to automatically detect the pod locations using machine learning such as YOLO algorithms (Redmon and Farhadi, 2018) will be tested using data collected in this work. Finally, with the image data collected in this study, we can test whether machine learning methods can generate sematic labels (Adams et al. 2020; Barth et al. 2019) such as the location of the first internode or to determine the main branch and primary branches. Unlike the object detection task, such as detecting pods in an image, determining the difference between the main branch and primary branches requires computer algorithms to understand how branches are organized in the image. This is a more challenging task for machine learning than simply detecting objects such as pods in an image.

Although the population used in our study has been genotyped using 50K SNP array, some of the known markers published by other studies are not represented in our SNP data. For example, the well-known Dt1 alleles (Liu et al. 2010; Tian et al. 2010) that regulate stem growth habits are not represented in our SNP data and we can only use a SNP that is closely linked to Dt1 locus as a proxy to estimate the genotype of the individual plants in our population. Although all of our selected individuals (PIs) are homozygous at both Dt1 (and Dt2) neighboring loci, these data cannot rule out the possibility that there are mutations in Dt1 and Dt2 loci that are associated with the observed variations in phenotypes in our population. Several recent studies have shown that genomic variations at regulatory sequences for major QTL genes can be important in explaining the missing heritability problem that is commonly observed in GWAS studies (Alonge et al. 2020; Liu et al. 2020; Zhou et al. 2019). Therefore, a potential next step for our work is to generate the genomic sequences of the key genes or resequencing the entire genome to determine whether additional genetic variations existed in these key regulators of plant architectures and whether those variations are functionally connected to the observed phenotypes.

In our phenotypic analysis, we used a topological approach to understand plant shoot architecture. As compared to simple geometric features, such as branch length and internode length, topological features take into account the overall structure of the branching patterns and have been shown to provide more informative features for trait association studies (Li et al. 2019). In our case we found that the first dimension of the topological distance (MDS1) is highly correlated with plant height, which suggests that topological analysis did capture important plant architecture parameters. More interestingly, we found that the MDS2, but not other branch geometric parameters, is correlated with pod numbers. This is an important insight from the topological analysis where the hidden features of branch patterns can be associated with the number of pods produced on the whole plant. Additional analysis could help to dissect the connection between MDS2 and pod number. For example, from the clustering dendrogram in Figure 2.1.4A, we found that MDS2 and pod number are in the same clade as four other geometric parameters including number of primary branches, short internode, first internode length and third internode length. The other fourteen traits are clustered in a separate clade as compared to these six traits. However, the correlations between each of the four geometric traits with pod number are close to zero. This result suggests that combinations or interactions of the four geometric traits could be related to total pod numbers. This can be observed in our data, for example, more primary branches can lead to more pod formations for a plant. However, when there is short internode in a plant, the two branches sometimes are underdeveloped and do not produce as many pods as those found in typical branches. Additional quantitative analysis can further help to elucidate such relationships between different traits.

CONCLUSION

In conclusion, we performed phenotypic analysis of a small collection of edamame varieties and identified intriguing correlations between geometric traits and topological traits in these varieties. Using known genetic markers and genes that are associated with pod numbers, we found several novel candidate/putative genes that might be related to the pod numbers. We found a negative correlation between pod number and first pod height, which suggests that breeding for new varieties of edamame to optimize the distribution of pods on plants will require a balance between these two traits but choosing plants with higher first pod height will yield better harvest efficiency. In the future, a larger population of edamame varieties would be required to perform GWAS study to identify more markers and candidate genes using our analytical pipelines developed in this work.

REFERENCES

Adams, Jason, Yumou Qiu, Yuhang Xu, and James C. Schnable. 2020. "Plant Segmentation by Supervised Machine Learning Methods." *Plant Phenome Journal* 3(1). doi: 10.1002/ppj2.20001.

Alonge, Michael, Xingang Wang, Matthias Benoit, Sebastian Soyk, Lara Pereira, Lei Zhang, Hamsini Suresh, Srividya Ramakrishnan, Florian Maumus, Danielle Ciren, Yuval Levy, Tom Hai Harel, Gili Shalev-Schlosser, Ziva Amsellem, Hamid Razifard, Ana L. Caicedo, Denise M. Tieman, Harry Klee, Melanie Kirsche, Sergey Aganezov, T. Rhyker Ranallo-Benavidez, Zachary H. Lemmon, Jennifer Kim, Gina Robitaille, Melissa Kramer, Sara Goodwin, W. Richard McCombie, Samuel Hutton, Joyce Van Eck, Jesse Gillis, Yuval Eshed, Fritz J. Sedlazeck, Esther van der Knaap, Michael C. Schatz, and Zachary B. Lippman. 2020. "Major Impacts of Widespread Structural Variation on Gene Expression

and Crop Improvement in Tomato." *Cell* 182(1):145-161.e23. doi: 10.1016/j.cell.2020.05.021.

Bao, Aili, Haifeng Chen, Limiao Chen, Shuilian Chen, Qingnan Hao, Wei Guo, Dezhen Qiu, Zhihui Shan, Zhonglu Yang, Songli Yuan, Chanjuan Zhang, Xiaojuan Zhang, Baohui Liu, Fanjiang Kong, Xia Li, Xinan Zhou, Lam Son Phan Tran, and Dong Cao. 2019. "CRISPR/Cas9-Mediated Targeted Mutagenesis of GmSPL9 Genes Alters Plant Architecture in Soybean." *BMC Plant Biology* 19(1). doi: 10.1186/s12870-019-1746-6.

Barth, R., J. IJsselmuiden, J. Hemming, and E. J. Van Henten. 2019. "Synthetic Bootstrapping of Convolutional Neural Networks for Semantic Plant Part Segmentation." *Computers and Electronics in Agriculture* 161:291–304. doi: 10.1016/j.compag.2017.11.040.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. "Fitting Linear Mixed-Effects Models Using Lme4."

Beiküfner, Mareike, Bianka Hüsing, Dieter Trautz, and Insa Kühling. 2019. "Comparative Harvest Efficiency of Soybeans between Cropping Systems Affected by First Pod Height and Plant Length." *Organic Farming* 5(1):3–13. doi: 10.12924/of2019.05010003.

Benlloch, Reyes, Ana Berbel, Latifeh Ali, Gholamreza Gohari, Teresa Millán, and Francisco Madueño. 2015. "Genetic Control of Inflorescence Architecture in Legumes." *Frontiers in Plant Science* 6(JULY):1–14.

Borghi, Lorenzo, Marina Bureau, and Rüdiger Simon. 2007. "Arabidopsis JAGGED LATERAL ORGANS Is Expressed in Boundaries and Coordinates KNOX and PIN Activity." *Plant Cell* 19(6):1795–1808. doi: 10.1105/tpc.106.047159.

Carlsson, Gunnar. 2009. *TOPOLOGY AND DATA*. Vol. 46.

Carson, Luther C., Josh H. Freeman, James R. Harris, and Gregory E. Welbaum. 2010. *Cultivation and Nutritional Constituents of Virginia Grown Edamame*.

Chalfun-Junior, Antonio, John Franken, Jurriaan J. Mes, Nayelli Marsch-Martinez, Andy Pereira, and Gerco C. Angenent. 2005. "ASYMMETRIC LEAVES2-LIKE1 Gene, a Member of the AS2/LOB Family, Controls Proximal-Distal Patterning in Arabidopsis Petals." *Plant Molecular Biology* 57(4):559–75. doi: 10.1007/s11103-005-0698-4.

Chang, Fangguo, Chengyu Guo, Fengluan Sun, Jishun Zhang, Zili Wang, Jiejie Kong, Qingyuan He, Ripa A. Sharmin, and Tuanjie Zhao. 2018. "Genome-Wide Association Studies for Dynamic Plant Height and Number of Nodes on the Main Stem in Summer Sowing Soybeans." *Frontiers in Plant Science* 9. doi: 10.3389/fpls.2018.01184.

Chen, Yiwu, and Randall L. Nelson. n.d. *Evaluation and Classification of Leaflet Shape and Size in Wild Soybean*.

Chi, Yingjun, Tingting Wang, Guangli Xu, Hui Yang, Xuanrui Zeng, Yixin Shen, Deyue Yu, and Fang Huang. 2017. "GmAGL1, a MADS-Box Gene from Soybean, Is Involved in Floral Organ Identity and Fruit Dehiscence." *Frontiers in Plant Science* 8. doi: 10.3389/fpls.2017.00175.

Cohen-Steiner, David, Herbert Edelsbrunner, and John Harer. 2007. "Stability of Persistence Diagrams." *Discrete and Computational Geometry* 37(1):103–20. doi: 10.1007/s00454-006-1276-5.

Decoteau, Dennis R., and Heather A. Hatt Graham. 1986. *Plant Spatial Arrangement Affects Growth, Yield, and Pod Distribution of Cayenne Peppers*.

Dhaliwal, Daljeet S., and Martin M. Williams. 2020. "Economically Optimal Plant Density for

Machine-Harvested Edamame." *HortScience* 55(3):368–73. doi: 10.21273/HORTSCI14642-19.

Dhami, H., K. Yu, T. Xu, Q. Zhu, K. Dhakal, J. Friel, S. Li, and P. Tokekar. 2019. "Crop Height and Plot Estimation from Unmanned Aerial Vehicles Using 3d Lidar." *ArXiv*.

Donahue, Jeff, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. n.d. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*.

Edelsbrunner, Herbert, and John Harer. n.d. *COMPUTATIONAL TOPOLOGY AN INTRODUCTION*.

Egli, D. B. 1993. *Cultivar Maturity and Potential Yield of Soybean*. Vol. 32.

Evans, Matthew M. S. 2007. "The Indeterminate Gametophyte1 Gene of Maize Encodes a LOB Domain Protein Required for Embryo Sac and Leaf Development." *Plant Cell* 19(1):46–62. doi: 10.1105/tpc.106.047506.

Fang, Chao, Yanming Ma, Shiwen Wu, Zhi Liu, Zheng Wang, Rui Yang, Guanghui Hu, Zhengkui Zhou, Hong Yu, Min Zhang, Yi Pan, Guoan Zhou, Haixiang Ren, Weiguang Du, Hongrui Yan, Yanping Wang, Dezhi Han, Yanting Shen, Shulin Liu, Tengfei Liu, Jixiang Zhang, Hao Qin, Jia Yuan, Xiaohui Yuan, Fanjiang Kong, Baohui Liu, Jiayang Li, Zhiwu Zhang, Guodong Wang, Baoge Zhu, and Zhixi Tian. 2017. "Genome-Wide Association Studies Dissect the Genetic Networks Underlying Agronomical Traits in Soybean." *Genome Biology* 18(1). doi: 10.1186/s13059-017-1289-9.

Fehr, W. R., and C. E. Caviness. 1977. "Stages of Soybean Development. Spec. Rep. 80. Iowa Agric." *Home Econ. Exp. Stn., Iowa State Univ., Ames*.

Fenta, Berhanu A., Stephen E. Beebe, Karl J. Kunert, James D. Burridge, Kathryn M. Barlow, Jonathan P. Lynch, and Christine H. Foyer. 2014. "Field Phenotyping of Soybean Roots for Drought Stress Tolerance." *Agronomy* 4(3):418–35. doi: 10.3390/agronomy4030418.

Grant, David, Rex T. Nelson, Steven B. Cannon, and Randy C. Shoemaker. 2009. "SoyBase, the USDA-ARS Soybean Genetics and Genomics Database." *Nucleic Acids Research* 38(SUPPL.1):843–46. doi: 10.1093/nar/gkp798.

Hao, Derong, Hao Cheng, Zhitong Yin, Shiyou Cui, Dan Zhang, Hui Wang, and Deyue Yu. 2012. "Identification of Single Nucleotide Polymorphisms and Haplotypes Associated with Yield and Yield Components in Soybean (Glycine Max) Landraces across Multiple Environments." *Theoretical and Applied Genetics* 124(3):447–58. doi: 10.1007/s00122-011-1719-0.

Illipronti, R. A., W. J. M. Lommen, C. J. Langerak, and P. C. Struik. 2000. "Time of Pod Set and Seed Position on the Plant Contribute to Variation in Quality of Seeds within Soybean Seed Lots." *Netherlands Journal of Agricultural Science* 48(2):165–80. doi: 10.1016/S1573-5214(00)80012-3.

Jiang, Guo Liang, Laban K. Rutto, and Shuxin Ren. 2018. "Evaluation of Soybean Lines for Edamame Yield Traits and Trait Genetic Correlation." *HortScience* 53(12):1732–36. doi: 10.21273/HORTSCI13448-18.

Kantolic, Adriana G., and Gustavo A. Slafer. 2007. "Development and Seed Number in Indeterminate Soybean as Affected by Timing and Duration of Exposure to Long Photoperiods after Flowering." *Annals of Botany* 99(5):925–33. doi: 10.1093/aob/mcm033.

Kigel, Jaime, Irit Konsens, Micha Ofir, and Can J. Plant Sci. n.d. *Branching, Flowering and*

Pod-Set Patterns in Snap-Bean (Phaseolus Vulgaris L.) as Affected by Temperature L99l]l T233. Vol. 71.

Kim, Yean Uk, Doug Hwan Choi, Ho Young Ban, Beom Seok Seo, Junhwan Kim, and Byun Woo Lee. 2020. "Temporal Patterns of Flowering and Pod Set of Determinate Soybean in Response to High Temperature." *Agronomy* 10(3). doi: 10.3390/agronomy10030414.

Konovsky, J., T. A. Lumpkin, and D. McClary. 1994. "Edamame: The Vegetable Soybean. Understanding the Japanese Food and Agrimarket: A Multifaceted Opportunity."

Lee, Ji Yong, Michael P. Popp, Elijah J. Wolfe, Rodolfo M. Nayga, Jennie S. Popp, Pengyin Chen, and Han Seok Seo. 2018. "Information and Order of Information Effects on Consumers' Acceptance and Valuation for Genetically Modified Edamame Soybean." *PLoS ONE* 13(10). doi: 10.1371/journal.pone.0206300.

Li, A., Y. Zhang, X. Wu, W. Tang, R. Wu, Z. Dai, G. Liu, H. Zhang, C. Wu, G. Chen, and X. Pan. 2008. "DH1, a LOB Domain-like Protein Required for Glume Formation in Rice." *Plant Molecular Biology* 66(5):491–502. doi: 10.1007/s11103-007-9283-3.

Li Mao, Duncan Keith, Topp Christopher N., Chitwood Daniel H. 2017. "Persistent Homology and the Branching Topologies of Plants." *American Journal of Botany* 104(3):349–53. doi: 10.3732/ajb.1700046.

Li, Mao, Laura L. Klein, Keith E. Duncan, Ni Jiang, Daniel H. Chitwood, Jason P. Londo, Allison J. Miller, and Christopher N. Topp. 2019. "Characterizing 3D Inflorescence Architecture in Grapevine Using X-Ray Imaging and Advanced Morphometrics: Implications for Understanding Cluster Density." *Journal of Experimental Botany* 70(21):6261–76. doi: 10.1093/jxb/erz394.

Liu, Baohui, Satoshi Watanabe, Tomoo Uchiyama, Fanjiang Kong, Akira Kanazawa, Zhengjun Xia, Atsushi Nagamatsu, Maiko Arai, Tetsuya Yamada, Keisuke Kitamura, Chikara Masuta, Kyuya Harada, and Jun Abe. 2010. "The Soybean Stem Growth Habit Gene Dt1 Is an Ortholog of Arabidopsis TERMINAL FLOWER1." *Plant Physiology* 153(1):198–210. doi: 10.1104/pp.109.150607.

Liu, Yucheng, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo An Zhou, Haikuan Zhang, Zhi Liu, Miao Shi, Xuehui Huang, Yan Li, Min Zhang, Zheng Wang, Baoge Zhu, Bin Han, Chengzhi Liang, and Zhixi Tian. 2020. "Pan-Genome of Wild and Cultivated Soybeans." *Cell* 182(1):162-176.e13. doi: 10.1016/j.cell.2020.05.023.

Lord, Nick. n.d. *Administrator, 1890 Extension Program*.

Mahoussi, Kadoukpe Arnaud Djanta, Etchikinto Agoyi Eric, Agbahoungba Symphorien, Jean-Baptiste Quenum Florent, Josiane Chadare Flora, Ephrem Assogbadjo Achille, Agbangla Clement, and Sinsin Brice. 2020. "Vegetable Soybean, Edamame: Research, Production, Utilization and Analysis of Its Adoption in Sub-Saharan Africa." *Journal of Horticulture and Forestry* 12(1):1–12. doi: 10.5897/jhf2019.0604.

Mebrahtu, Tadesse, and Chris Mullins. 2007. *Efficiency of Mechanical Harvest for Immature Vegetable Soybean Pods 1*. Vol. 58.

Mentreddy, S. R., A. I. Mohamed, N. Joshee, A. K. Yadav, and others. 2002. "Edamame: A Nutritious Vegetable Crop." Pp. 432–38 in *Trends in new crops and new uses. Proceedings of the Fifth National Symposium, Atlanta, Georgia, USA, 10-13 November, 2001*.

Mizukamil, Yukiko, and Hong Ma2. 1997. *Determination of Arabidopsis Floral Meristem Ldentity by AGA MOUS*. Vol. 9. American Society of Plant Physiologists.

Mogili, Um Rao, and B. B. V. L. Deepak. 2018. “Review on Application of Drone Systems in Precision Agriculture.” Pp. 502–9 in *Procedia Computer Science*. Vol. 133. Elsevier B.V.

Mourtzinis, Spyridon, and Shawn P. Conley. 2017. “Delineating Soybean Maturity Groups across the United States.” *Agronomy Journal* 109(4):1397–1403. doi: 10.2134/agronj2016.10.0581.

Nico, Magalí, Anita I. Mantese, Daniel J. Miralles, and Adriana G. Kantolic. 2016. “Soybean Fruit Development and Set at the Node Level under Combined Photoperiod and Radiation Conditions.” *Journal of Experimental Botany* 67(1):365–77. doi: 10.1093/jxb/erv475.

Ning, Hailong, Jiaqi Yuan, Quanzhong Dong, Wenbin Li, Hong Xue, Yanshu Wang, Yu Tian, and Wen Xia Li. 2018. “Identification of QTLs Related to the Vertical Distribution and Seed-Set of Pod Number in Soybean [Glycine Max (L.) Merri].” *PLoS ONE* 13(4). doi: 10.1371/journal.pone.0195830.

Nyquist, Wyman E., and R. J. Baker. 1991. “Estimation of Heritability and Prediction of Selection Response in Plant Populations.” *Critical Reviews in Plant Sciences* 10(3):235–322.

Ping, Jieqing, Yunfeng Liu, Lianjun Sun, Meixia Zhao, Yinghui Li, Maoyun She, Yi Sui, Feng Lin, Xiaodong Liu, Zongxiang Tang, Hanh Nguyen, Lijuan Qiu, Zhixi Tian, Randall L. Nelson, Thomas E. Clemente, James E. Specht, and Jianxin Ma. 2014. “Dt2 Is a Gain-of-Function MADS-Domain Factor Gene That Specifies Semideterminacy in Soybean.” *Plant Cell* 26(7):2831–42. doi: 10.1105/tpc.114.126938.

Ramteke, Rajkumar, Devvrat Singh, and Pooja Murlidharan. 2012. *Selecting Soybean (Glycine Max) Genotypes for Insertion Height of the Lowest Pod, the Useful Trait for Combine*

Harvester. Vol. 82.

Redmon, Joseph, and Ali Farhadi. n.d. “YOLO9000: Better, Faster, Stronger.”

Rueden, Curtis T., Johannes Schindelin, Mark C. Hiner, Barry E. DeZonia, Alison E. Walter, Ellen T. Arena, and Kevin W. Eliceiri. 2017. “ImageJ2: ImageJ for the next Generation of Scientific Image Data.” *BMC Bioinformatics* 18(1). doi: 10.1186/s12859-017-1934-z.

Shanmugasundaram, S., Taiwan. Nong lin ting., Xing zheng yuan nong ye wei yuan hui (China), and Asian Vegetable Research and Development Center. 1991. *Vegetable Soybean fl Research Needs for Production and Quality I Held at Kenting, Taiwan, 29 April--2 May 1991*. [Center].

Shurtleff, William, Akiko Aoyagi, and others. 2009. “History of Edamame, Green Vegetable Soybeans, and Vegetable-Type Soybeans (1275-2009)[Electronic Resource].”

Song, Qijian, David L. Hyten, Gaofeng Jia, Charles V. Quigley, Edward W. Fickus, Randall L. Nelson, and Perry B. Cregan. 2013. “Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean.” *PLoS ONE* 8(1). doi: 10.1371/journal.pone.0054985.

Sun, Shangpeng, Changying Li, Andrew H. Paterson, Yu Jiang, Rui Xu, Jon S. Robertson, John L. Snider, and Peng W. Chee. 2018. “In-Field High Throughput Phenotyping and Cotton Plant Growth Analysis Using LiDAR.” *Frontiers in Plant Science* 9. doi: 10.3389/fpls.2018.00016.

Sun, Zhengxi, Chao Su, Jinxia Yun, Qiong Jiang, Lixiang Wang, Youning Wang, Dong Cao, Fang Zhao, Qingsong Zhao, Mengchen Zhang, Bin Zhou, Lei Zhang, Fanjiang Kong, Baohui Liu, Yiping Tong, and Xia Li. 2019. “Genetic Improvement of the Shoot

Architecture and Yield in Soya Bean Plants via the Manipulation of GmmiR156b." *Plant Biotechnology Journal* 17(1):50–62. doi: 10.1111/pbi.12946.

Teichmann, Thomas, and Merlin Muhr. 2015. "Shaping Plant Architecture." *Frontiers in Plant Science* 6(APR).

Tian, Zhixi, Xiaobo Wang, Rian Lee, Yinghui Li, James E. Specht, Randall L. Nelson, Phillip E. McClean, Lijuan Qiu, and Jianxin Ma. 2010. "Artificial Selection for Determinate Growth Habit in Soybean." *Proceedings of the National Academy of Sciences of the United States of America* 107(19):8563–68. doi: 10.1073/pnas.1000088107.

Ulloa, Santiago M., Avishek Datta, Goran Malidza, Robert Leskovsek, and Stevan Z. Knezevic. 2010. "Yield and Yield Components of Soybean [Glycine Max (L.) Merr.] Are Influenced by the Timing of Broadcast Flaming." *Field Crops Research* 119(2–3):348–54. doi: 10.1016/j.fcr.2010.08.006.

Verri, A., C. Uras, P. Frosini, and M. Ferri. 1993. *On the Use of Size Functions for Shape Analysis*. Vol. 70. Springer-Verlag.

Wang, Ying, and Yuling Jiao. 2018. "Axillary Meristem Initiation — a Way to Branch Out." *Current Opinion in Plant Biology* 41:61–66. doi: 10.1016/j.pbi.2017.09.001.

Wong, Chui E., Mohan B. Singh, and Prem L. Bhalla. 2013. "The Dynamics of Soybean Leaf and Shoot Apical Meristem Transcriptome Undergoing Floral Initiation Process." *PLoS ONE* 8(6). doi: 10.1371/journal.pone.0065319.

Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. "Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine Max." *Genetics* 206(2):1081–89. doi: 10.1534/genetics.116.198713.

Yoshikawa, Takanori, Suguru Ozawa, Naoki Sentoku, Jun Ichi Itoh, Yasuo Nagato, and Shuji Yokoi. 2013. “Change of Shoot Architecture during Juvenile-to-Adult Phase Transition in Soybean.” *Planta* 238(1):229–37. doi: 10.1007/s00425-013-1895-z.

Zandonadi, Rodrigo, T. Coolong, and T. Pfeiffer. 2010. “Mechanical Harvesting of Edamame.” *SARE Proj Final Rep. Available Online at: Https://Www. Uky. Edu/Ccd/Sites/Www. Uky. Edu. Ccd/Files/Edamame_mechanical_harvest. Pdf (Accessed February 7, 2021)*.

Zhang, Dajian, Xutong Wang, Shuo Li, Chaofan Wang, Michael J. Gosney, Michael V. Mickelbart, and Jianxin Ma. 2019. “A Post-Domestication Mutation, Dt2, Triggers Systemic Modification of Divergent and Convergent Pathways Modulating Multiple Agronomic Traits in Soybean.” *Molecular Plant* 12(10):1366–82. doi: 10.1016/j.molp.2019.05.010.

Zhang, Jiaoping, Qijian Song, Perry B. Cregan, Randall L. Nelson, Xianzhi Wang, Jixiang Wu, and Guo Liang Jiang. 2015. “Genome-Wide Association Study for Flowering Time, Maturity Dates and Plant Height in Early Maturing Soybean (Glycine Max) Germplasm.” *BMC Genomics* 16(1). doi: 10.1186/s12864-015-1441-4.

Zhang, Lingxiao, and S. Kyei-Boahen. n.d. *Growth and Yield of Vegetable Soybean (Edamame) in Mississippi*.

Zhou, Peng, Candice N. Hirsch, Steven P. Briggs, and Nathan M. Springer. 2019. “Dynamic Patterns of Gene Expression Additivity and Regulatory Variation throughout Maize Development.” *Molecular Plant* 12(3):410–25. doi: 10.1016/j.molp.2018.12.015.

Zhou, Zhengkui, Yu Jiang, Zheng Wang, Zhiheng Gou, Jun Lyu, Weiyu Li, Yanjun Yu, Liping Shu, Yingjun Zhao, Yanming Ma, Chao Fang, Yanting Shen, Tengfei Liu, Congcong Li,

Qing Li, Mian Wu, Min Wang, Yunshuai Wu, Yang Dong, Wenting Wan, Xiao Wang, Zhaoli Ding, Yuedong Gao, Hui Xiang, Baoge Zhu, Suk Ha Lee, Wen Wang, and Zhixi Tian. 2015. "Resequencing 302 Wild and Cultivated Accessions Identifies Genes Related to Domestication and Improvement in Soybean." *Nature Biotechnology* 33(4):408–14. doi: 10.1038/nbt.3096

Tables and Figures

Table 2.1. 1 Heritability of Plant Traits.

Traits	**Heritability**
Pod Number (PN)	0.91
First Pod Height (FPH)	0.91
MDS2	0.88
Average Primary Branch Length (APBL)	0.85
Main Branch Length (MBL)	0.85
Total Branch Length (TBL)	0.85
MDS1	0.84
Plant Height (PH)	0.84
First Node Height (FNH)	0.83
Pod Height 10% (PP10CM)	0.83
MDS_3	0.75
Number of Primary Branch (NPB)	0.73
Second Internode Length (SIL)	0.70
Short Internode (SIN)	0.56
Third Internode Length (TIL)	0.48
First Internode Length (FIL)	0.30

Table 2.1.2 Candidate genes associated with pod number.

NR: non-reference allele. Ref: reference allele. The numbers of NR and Ref indicate the number of varieties out of 24 PIs. Low and High are the pod numbers and are defined as pod numbers below or above average, respectively. The p values were calculated using fisher's exact test. Candidate genes are selected with 50Kb window of enriched SNP based on published 50K SNP data. References: 1. Zhang et al., 2015. 2. Fang et al., 2017. 3. Hao et al., 2012. SNP id ss715622826-7 means ss715622826 and ss715622827. SNP id ss715610851-6 means ss715610851, ss715610852, ss715610853, ss715610854, ss715610855, and ss715610856. The genes marked with asterisk (*) have log_2fold change > 1 indicating the candidate genes are expressed more than 2-fold higher in shoot apical meristem than in leaves. # and $ mark the numbers described in the main text.

Pod Number			**High**		**Low**					
Published QTL Marker	**Ref.**	**SNP id**	**NR**	**Ref**	**NR**	**Ref**	**p value**	**Candidate Gene**	**Fold Change**	**Gene Function**
Chr11:15649090	1	ss715609881	11	1[#]	0	12[$]	9.60E-06	Glyma.11G164800	3.17*	LOB domain-containing protein 4
Chr15:5252046	1	ss715622826-7	3	9	11	1	2.80E-03	Glyma.15G068500	8.6*	60S ribosomal protein L26-1-like
Chr18:55626229	2	ss715632229	5	7	12	0	4.60E-03	Glyma.18G274000	16.83*	RING finger protein 165-like

Chr11:5338661	3	ss715610851-6	8	4	1	11	9.40E-03	Glyma.11G071000	1.28*	auxilin-like protein 1-like
Chr02:40368201	3	ss715582578	6	6	12	0	1.40E-02	Glyma.02G216500	2.44*	TRAF-type zinc finger-related
Chr02:40368201	3	ss715582578	6	6	12	0	1.40E-02	Glyma.02G216600	1.39*	AGAMOUS-like 16
Chr11:33902439	1	ss715610584	8	4	2	10	3.60E-02	Glyma.11G245300	0.72	63 kDa inner membrane protein
Chr18:55662445	2	ss715632230	7	5	12	0	3.70E-02	Glyma.18G274200	0.08	Pollen Ole e1 extension family protein

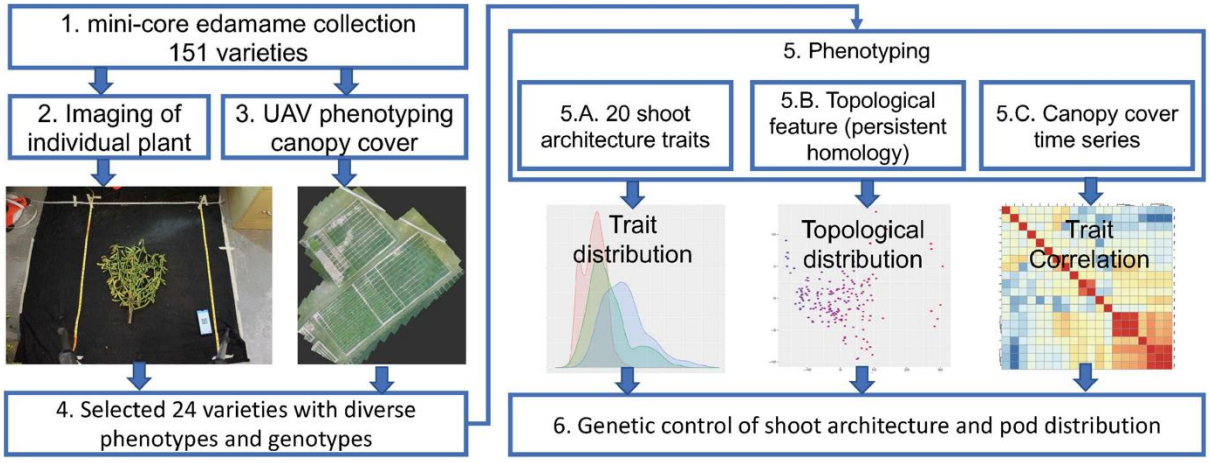


Figure 2.1.1 Workflow of phenomics analysis of edamame shoot architecture and canopy cover.

Step 1. Data from a mini-core collection of edamame varieties was used for this study. Step 2. Individual plants were imaged twice on a black background. Step 3. Unmanned aerial vehicle was used to collect canopy cover data over the growth season. Step 4. Selected varieties were used for detailed characterization of the shoot architecture. Step 5. Phenotypic data analyses were performed. Step 6. Potential genetic control points of shoot architecture were analyzed.

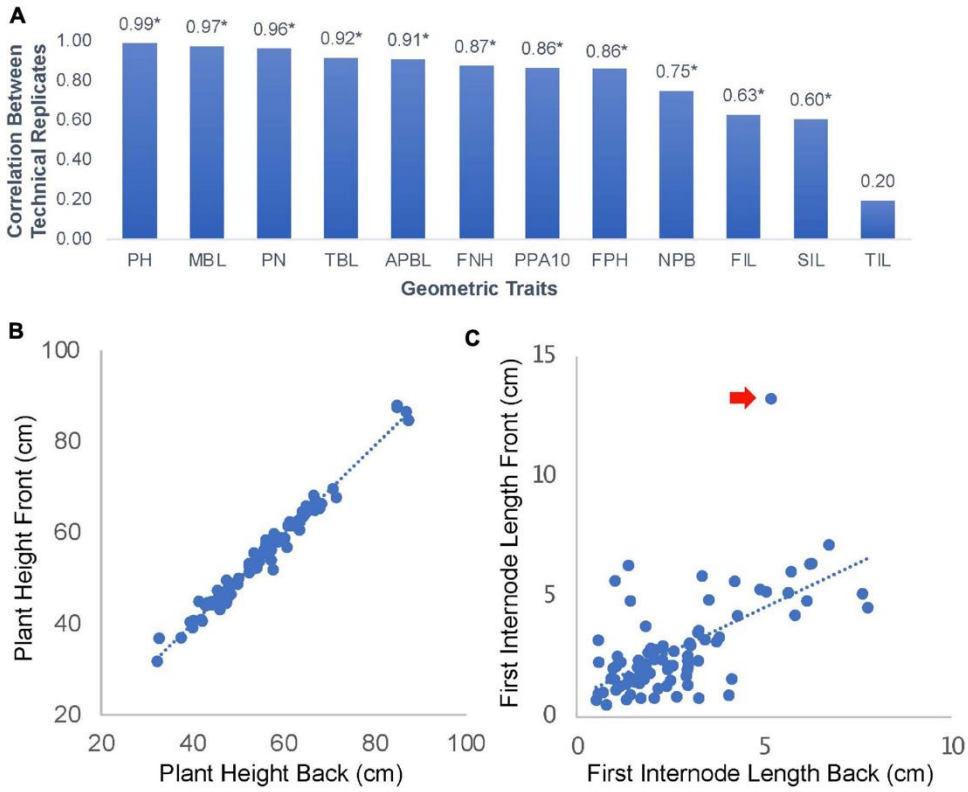


Figure 2.1.2 Parameter correlations for images of edamame shoot architecture.

(A) 12 geometric traits were measured in this study. PH: plant height. MBL: main branch length. PN: pod number. TBL: total branch length. APBL: average primary branch length. FNH: first node height. PPA10: percent of pod above 10 cm from ground. FPH: first pod height. NPB: number of primary branches. FIL: first internode length. SIL: second internode length. TIL: third internode length. Numbers on top of the bars are Pearson Correlation Coefficient (PCC). * Indicates statistical significance (p value < 0.01). (B) scatter plot of plant height measurement between technical replications (front and back images of the same plant). (C) scatter plot of first internode length. Red arrow indicates an outlier data point.

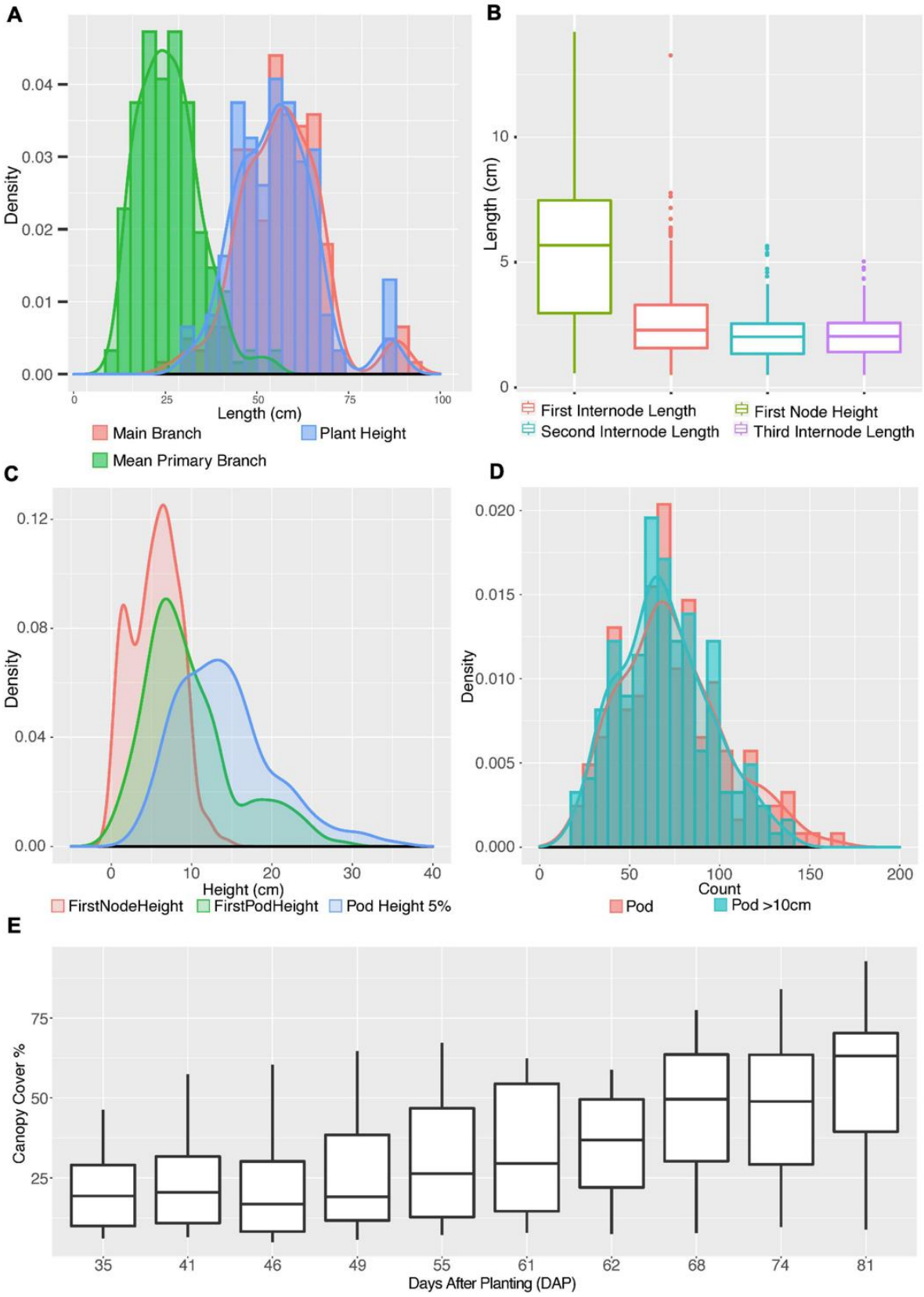

A
Density
Length (cm)
Main Branch
Plant Height
Mean Primary Branch
B
Length (cm)
First Internode Length
First Node Height
Second Internode Length
Third Internode Length
C
Density
Height (cm)
FirstNodeHeight
FirstPodHeight
Pod Height 5%
D
Density
Count
Pod
Pod >10cm
E
Canopy Cover %
Days After Planting (DAP)

Figure 2.1.3 Distribution of shoot architecture parameters in edamame plants.

(A) histograms and density plots for main branch length, plant height and average primary branch length. (B) Boxplot shows the distribution of first pod height, first, second and third internode length. (C) Density plot compares first node height with first pod height and 5% pod height. (D) histograms and density plots for number of pods per plant and number of pods above 10cm from ground. (E) Boxplot of canopy cover change in the growth season.

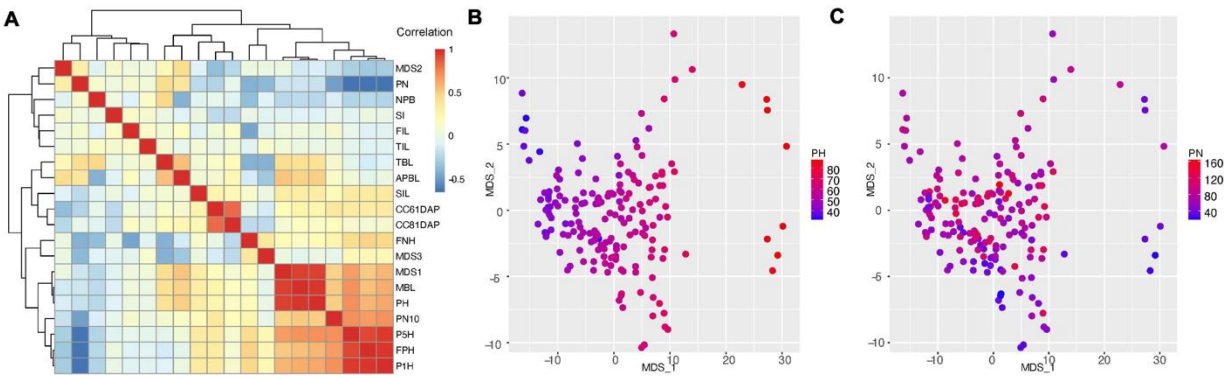


Figure 2.1.4 Correlation analysis of traits characterized in this study.

(A) correlation between edamame shoot architecture traits with topological traits (MDS1, 2, and 3). Trait names are described in figure 2 and main text. PN: pod number. NPB: number of primary branches. SI: short internode. FIL: first internode length. SIL: second internode length. TIL: third internode length. TBL: total branch length. APBL: average primary branch length. CC61DAP: canopy cover 61 days after planting. CC81DAP: canopy cover 81 days after planting. FNH: first node height. MBL: main branch length. PH: plant height. PN10: pod number above 10 cm from ground. P5H: height above ground for 5% pods. P1H: height above ground for 1% pods. FPH: first pod height. (B) comparison of MDS1 and MDS2 with plant height. (C) comparison of MDS1 and MDS2 with pod number.

Supplementary Figures

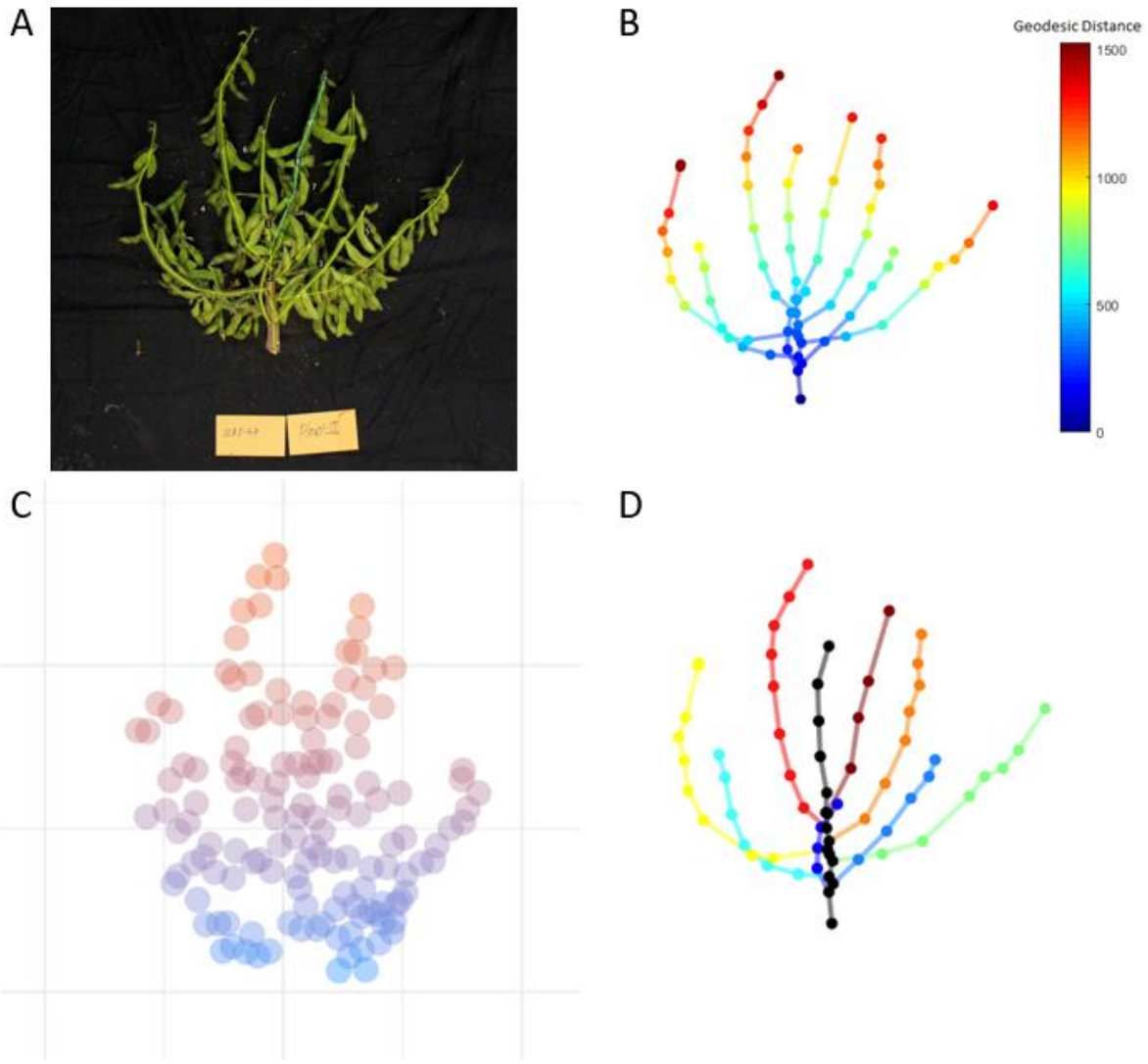


Figure 2.1.S.1

(A) Image of an edamame plant. All leaves and petioles were removed before imaging. Branches and pods were labeled. (B) Geodesic distance shows the distance of tip of each branch to the ground. (C) Pod location distribution is also analyzed. (D) Color coding for each branch with the main branch colored black.

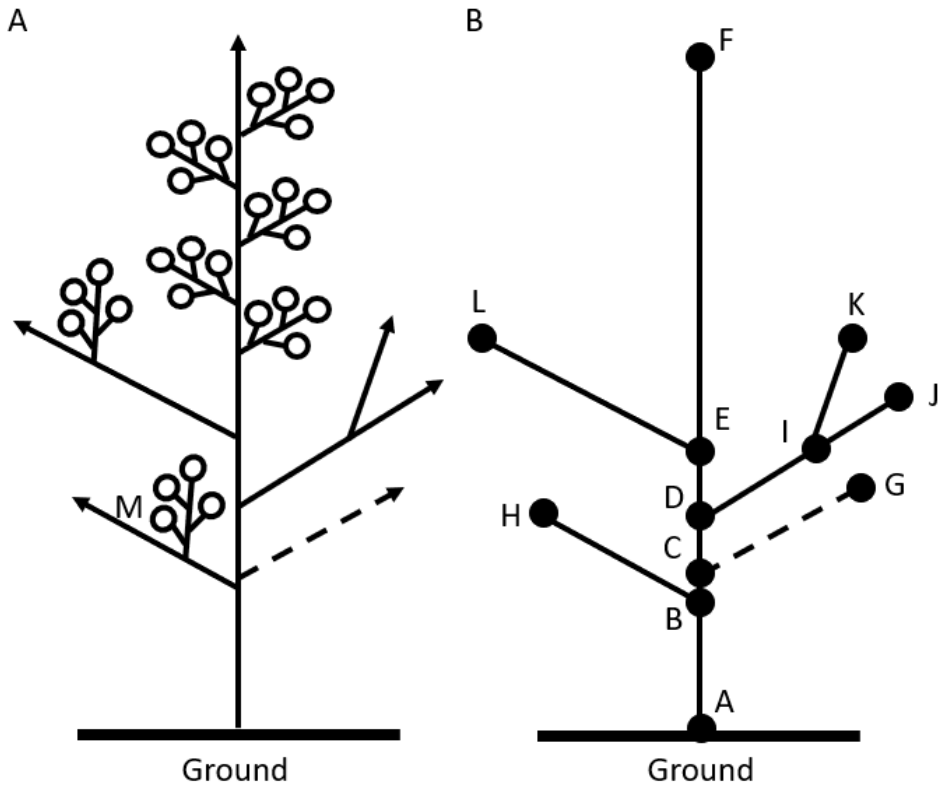


Figure 2.1.S.2

(A) a schematic illustration of edamame inflorescence. Arrows indicate branches and circles indicate flowers/pods. M marks the first pod location. The first pod height (FPH) is the distance from the first pod to the ground. (B) Main branch is defined as the longest branch (Branch AF). Primary branches are the branches grow from the main branches (BH, CG, DJ, and EL). If CG is present, we call CB as a short internode (SI) because CB is much shorter than other internodes. With CG, the number of primary branches (NPB) is 4. The first internode length is determined by DB, because we ignore the short internode. Total branch length (TBL) is the sum of all branches. APBL is the average primary branch length. First node height (FNH) is AB. Main branch length (MBL) and plant height (PH) are both AF in this illustration but may differ in real situations because the main branch may not be straight.

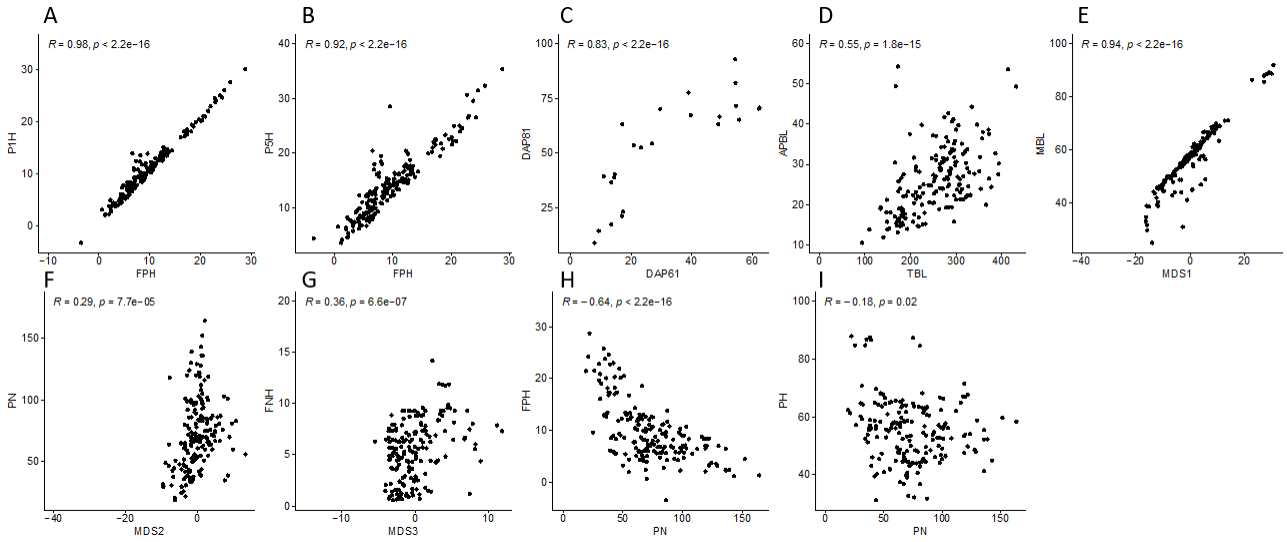


Figure 2.1.S.3 Pair-wise correlation plot for selected pairs of traits in main text figure 4.

Chapter II Section 2: Genome Wide Association Studies (GWASs) of Canopy Cover in Edamame

ABSTRACT

Edamame popularly known as vegetable soybean, is a ripening or immature fruit and seed stage of soybean. Canopy cover is one of the factors that contribute to the yield. Canopy cover naturally suppresses the weed and reduces the cost of application of herbicides. In this study, we use UAVs and Genome Wide Association Studies (GWAS) to find the genetic markers related to canopy cover for edamame. Using a population of edamame PIs, we identified SNPs with significant p value in GWAS study of canopy cover. We found significant variations in canopy covers over time. This research provides insight into the genetic regulation of canopy cover and can be used to further develop edamame varieties that have rapid canopy closure.

INTRODUCTION

Edamame [Glycine max (L.) Merr.] is a type of green, vegetable soybean which is a source of different nutrients such as protein, isoflavones, and vitamins(Lee et al. 2018; Mahoussi et al. 2020; Mentreddy et al. 2002). Edamame has become a popular snack food in the United States and many countries although it been cultivated in east Asian countries for more than 2,000 years and documented edamame varieties have been originated from this area (Shurtleff, Aoyagi, and others 2009). Although, the yield components (plant density, number of pods and number of seeds per pod and seed size) of soybeans have been studied (Liu et al. 2010; Ulloa et al. 2010) but the yield components of edamame have not been extensively done as it is harvested when the pods are still green. Although the United States is a major producer of grain soybeans, but most frozen edamame products consumed in the United States were imported from Asia. The main obstacles for commercial production of edamame in the United States is the cost of production where growers

must allot a big sum of money for weed control and the growth monitoring which is difficult for larger fields. Canopy Coverage (CC), which is the amount of ground area enclosed by the plant, is correlated with canopy light interception. Light interception is a trait that is positively correlated with grain yield but is difficult to measure (Purcell 2000). Canopy cover is the natural remedy to control and suppress the weed growth in the edamame field. The major challenge is canopy cover breeding is non-destructive phenotyping at the field scale.

High-throughput phenotyping platforms permit capturing a trait information in a non-destructive manner. It also permits the collection of time-series data at low cost which helps to track the plants' growth and development. High throughput phenotyping has been used to study leaf shape (Chen and Nelson, 2004), root architecture (Fenta et al. 2014) and canopy cover (Xavier et al. 2017) in soybeans. CC data have been successfully collected via UASs in multiple legumes (Cazenave et al. 2019; Sarkar et al. 2020; Xavier et al. 2017). Xavier (2017) phenotyped time-series CC among several recombinant inbred lines (RILs) of a soybean nested association mapping (SoyNAM) population and identified QTLs using Genome-wide association studies (GWASs).

Genome-wide association studies (GWASs) are widely used in plants to identify genetic markers associated with variation in a diverse array of plant traits (Liu and Yan 2019). Time-series data of a trait, along with GWASs, can dissect the dynamic regulation of plant phenotypes. However, very few studies have combined GWASs and time-series traits to dissect the changes in crop phenotypes. Field-based phenotyping systems including UASs have been used to decode several causal loci in maize (Wang et al. 2019; Anderson et al. 2020; Adak et al. 2021) and wheat (Lyra et al. 2020) for plant height at different growth stages. The soybean research and breeding community have a collection of substantial amounts of genomics resources including more than 20,000 plant introductions (PIs) that were genotyped by 50K SNP array (Song et al. 2013) and

over 3,000 PIs with full genome sequences available (Liu et al. 2020). These published studies have identified key genes in soybeans associated with canopy closure.

In this study, we combined the high throughput of UASs, and the diversity of edamame natural populations to dissect the changes in CC over time by collecting time-series CC data. In addition, the changing growth curves and rates of canopy closure discovered by time-series CC in this diversity panel have potential applications in breeding programs to develop improved edamame cultivars. In this work, we use genetic diversity and genomic resources in edamame breeding to improve canopy cover by developing a phenotyping pipeline to collect canopy cover data over two growth seasons. Our results provide an accessible pipeline of canopy cover quantification using drone phenotyping and provide novel candidate markers and genes for improving rapid canopy closure in edamame.

METHODS

Plant materials and phenotyping

A total of 269 soybean PIs with > 20g/100 seeds that are potential parental lines for developing edamame varieties (referred as edamame PIs) were sown in 3 meters two-row plots and 0.76 m row spacing (with a seeding rate of ~70,000 plants per hectare) arranged in a complete randomized design with two to four replications in Kentland farm at Blacksburg, VA in 2020 and 2021. Plots were organized in a randomized complete block design (RCBD) having two blocks. Because of limited seed quantity, we merged block replicates prior to seed processing. Following a similar planting method, 272 soybean PIs were sown in 2021.

Drone image collection and analysis

In 2020, we collected aerial images from 13 flights using a DJI Phantom 4-Advanced drone at an altitude of 30.5 meters (100feet) above ground level. 4,414 individual images were collected

during this growth season, with an average of 340 drone images collected for each flight day. In 2021, drone images from 5 time points (days) were collected from DJI Phantom 4-Advanced. MicaSense RedEdge (multispectral (MS) camera: 5 bands) mounted on DJI Inspire-2 was also used to collect drone images from seven different time points during this growth period. 675 and 8,230 (set of 1646) images were collected at 40 meters (~150 feet) above ground level using DJI Phantom 4-Advanced and MicaSense RedEdge camera respectively this year. The side overlap and front overlap for all the flights were set at 75% with padding. Drone flight waypoints were generated using an iPad app, DroneDeploy for DJI Phantom and DJI Go along with Atlas flight for DJI Inspire-2 mounted MicaSense RedEdge camera. The precise GPS location of the Ground control points (GCPs) were determined using a Real-time kinematic (RTK) GPS. Orthorectified drone maps were generated using AgiSoft Metashape professional addition (Version 1.6). During the orthomosaic generation procedures, the raw RGB images were aligned to detect the "interest points" or "key points" or "feature points." The selected points were used to determine the correspondences between interest points in images. The coordinates of Ground Control Points (GCPs) or markers placed in the field were imported and matched with their corresponding location in the image. This enabled the precise positioning of the GCPs in the orthomosaic. The subplot extraction was done manually using a tool called zonal statistics as table in ArcGIS Pro. Canopy cover data was imported in excel and averaged across replicates to generate a growth curve for each variety and the results were compared with other canopy cover data obtained at different dates from different indices via different cameras/drones for two years.

Phenotypic Data

PI-wise ExG (Excess Greenness) Index, was calculated by formula (ExG = 2 * Green - Red - Blue) for the PIs in 2020 and 2021 from the drone images from Phantom 4 Advanced. PI-

wise ExG index and Normalized Difference Vegetation Index (NDVI), was calculated by formula (NDVI = (NIR - Red) / (NIR + Red)) for the PIs in 2021 only from the drone images from MicaSense RedEdge camera counted on DJI Inspire-2.

Genotypic Data

We downloaded the SNP marker data of the 269 (in 2020) and 272 (in 2021) accessions from the SoySNP50K SNPs public data repository (Song et al., 2015). We filtered out a total of 42,509 initial SNPs by low minor allele frequency (MAF < 0.05) and missing genotypes, which resulted in 36,076 SNPs and these SNPs were used for population structure and GWAS study.

Population Structure

Population structure was carried out using the an R package, adegenet (Jombart 2008), with discriminant analysis of principal components (DAPC) to identify clusters belonging to genetically related individuals (Jombart, Devillard, and Balloux 2010). We used the function find.clusters with maximum clusters as k = 40 for successive k-means clustering. We retained 200 principal components (PCs), and used Bayesian information criterion (BIC) to identify the optimal number of clusters. We then used the function dapc by retaining an optimal number of PCs and we retained all discriminant functions and eigenvalues.

Genome-wide association study (GWAS) and candidate gene discovery

A genome-wide association study was conducted using rrBLUP (R) package. The single marker regression (SMR), GWAS statistical model was used for the GWAS. We used a modified Sidak correction for multiple testing to identify significant associations between markers. We used the poolr (R) package with the Li and Ji method (Li and Ji 2005) to calculate the effective number of markers (Meff) which was 4632. Meff replaced m (total number of markers). The adjusted significance threshold was 4.956 where α was set as 5%, and the suggestive threshold was 3.967

α was set as 25%. QQ and Manhattan plots were used to visualize results with the qqman package (Turner 2014).

Three published GWASs studies in legumes (Cazenave et al. 2019; Sarkar et al. 2020; Xavier et al. 2017) have analyzed the CC traits in soybeans and they have provided the candidate markers and genes for CC. Since different studies used differing genotyping methods, we compared the markers used in our study (50K SNP array) to the markers used in other publications by determining the genomic locations of these markers on the same reference genome. For the markers known to be associated with canopy cover, we first identified their location in a recent soybean genome release (Wm82.a2. v1). We then identify those 50K SNP array markers that are closest to these published markers (within 10-20kb) (Qin et al. 2019; Xie et al. 2018). In most cases, we can find associated markers within 10kb from the published markers and in some cases, multiple markers located within the predefined genomic range were found. Marker data was downloaded from soybase.org as a Variant Call Format (VCF) file. Candidate genes were identified as those genes that are closest to the significant Single Nucleotide Polymorphism (SNP) markers. In case the marker is in a gene dense region, the gene functions were manually selected based on their homology to other plant genes that are more likely to be involved in controlling CC. We reported the gene descriptions from gene homolog descriptions from TAIR for *Arabidopsis thaliana* (Berardini et al. 2015). The gene descriptions were reported from either PANTHER or GO databases (Anon 2021; Ashburner et al. 2000; Mi, Muruganujan, and Thomas 2012), whenever TAIR homologs were not available.

RESULTS

2020 Results

Phenotypic

Since DJI Phantom 4 Advanced was used to collect the canopy cover data in 2020, we used the Excess Greenness (EXG) index to perform GWAS. The EXG canopy cover data obtained from RGB images in 2020 seems to be increasing rapidly from 18th of June until 17th July, where it drops down and increases again with slow rate until 29th July and starts decreasing from 2nd August as shown in Figure 2.2.1.

Population Structure

We retained 200 principal components that accounted for 90% of cumulative variance through DAPC and with the smallest BIC, and the optimal number of clusters, k= 6 was determined (Figure 2.2.1). Clusters 1, 3 and 6 were closer to each other, but were well separated from clusters 2, 4 and 5. Also the 2nd and 4th clusters were near to each other but were well separated from the 5th cluster as shown by Figure 2.2.5.

GWAS on EXG index from RGB Images of Canopy cover data from 2020

Ten SNPs (ss715578480, ss715578590, ss715578606, ss715578448, ss715578586, ss715581587, ss715598441, ss715603992, ss715611125, and ss715611120) displayed significant associations as shown by Figure 2.2.6. The remaining 31 SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold. Chromosome (Chr) 1 had the most associations (five significant, six suggestive), followed by Chr 11 (two significant, three suggestive), Chr 7 (one significant, three suggestive), Chr 9 (one significant, one suggestive), Chr 2 (one significant), Chr 4,6,14,16, and 19 (three suggestive), and Chr 15, 17, and 20 (one suggestive).

Candidate Genes

On July 17, six candidate genes (Glyma07g07410, Glyma01g10320, Glyma01g13430, Glyma01g13654, Glyma01g13930, and Glyma01g09610) were found to be associated with CC, where the $-\log_{10}(P)$ of SNPs were more than 4.956. Three candidate genes (Glyma11g10480, Glyma11g10370, and Glyma02g24071) were associated with CC on July 29. One candidate gene (Glyma07g07393) was found to be associated with CC on July 21 and July 25. One candidate gene (Glyma09g31460) associated with CC on August 2. There were no common candidate genes on different dates.

2021 Results

Phenotypic

Both DJI Phantom 4 Advanced and Micasense RedEdge camera were used to collect RGB and Multispectral (MS) images in 2021, we calculated the percent canopy cover using ExG index and NDVI from all the PIs to perform GWAS.

The NDVI canopy cover data obtained from MS images in 2021 seems to be increasing overtime with varying rate as shown in Figure 2.2.2. Similarly, the EXG index canopy cover data obtained from RGB images in 2021 seems to be increasing rapidly from 13th of June until 21st July, where it increases again with slow rate until 31st July as shown in Figure 2.2.3. Also, the EXG index canopy cover data obtained from MS images in 2021 seems to be increasing rapidly from 13th of June until 6th July, where it increases again with slow rate until 31st July as shown in Figure 2.2.4.

Population Structure

We retained 150 principal components that accounted for 80% of cumulative variance through DAPC and with the smallest BIC, and the optimal number of clusters, k= 6 was determined (Figure 2.2.3). All six clusters were well separated from each other as shown by Figure 2.2.7.

GWAS on EXG index from RGB Images of Canopy cover data from 2021

Three SNPs (ss715618478, ss715585755, and ss715635643) displayed significant associations as shown in Figure 2.2.9. The remaining 17 SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold. Chromosome (Chr) 14 and 19 contained the two associations (one significant, one suggestive), followed by Chr 3 (one significant), Chr 7 (six suggestive), Chr 6 (five suggestive), and Chr 1, 2, 12, 13, and 17 (one suggestive).

Candidate Genes

On July 31 only, three candidate genes Glyma03g30900, Glyma14g05110, and Glyma19g40750) were associated with CC, where the –log10(P) of SNPs were over 3.967. There were not any candidate genes found on other dates on EXG index from RGB images of CC.

GWAS on NDVI from MS Images of Canopy cover data from 2021

Seven SNPs (ss715631071, ss715583614, ss715583635, ss715584302, ss715583646, ss715582577, and ss715583581) displayed significant associations. The remaining 39 SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold. Chromosome (Chr) 2 contained the most associations (six significant, eighteen suggestive), followed by Chr 18 (one significant), Chr 17 (six suggestive), Chr 1 (five suggestive), Chr 11 (four suggestive), Chr 4 (three suggestive), Chr 8 (two suggestive), and Chr 6, 15, and 20 (one suggestive) as shown in Figure 2.2.9.

Candidate Genes

On July 31, four candidate genes (Glyma02g00760, Glyma02g11660, Glyma02g00640, and Glyma02g00730,) were found to be associated with CC, where the $-\log_{10}(P)$ of SNPs were more than 4.956. One candidate gene (Glyma02g11660) was found to be associated with CC on July 22. One candidate gene (Glyma18g01010) was found to be associated with CC on July 14.

GWAS on EXG from MS Images of Canopy cover data from 2021

Three SNPs (ss715596520, ss715609766, and ss715613725) displayed significant associations. The remaining 30 SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold. Chromosome (Chr) 13 contained the most associations (one significant, three suggestive), followed by 7 (one significant, two suggestive), Chr 11 (one significant), Chr 16 (eight suggestive), Chr 1 (six suggestive), Chr 4, and 6 (three suggestive), Chr 3, and 18 (two suggestive), and Chr 5, and 15 (one suggestive) as shown in Figure 2.2.10.

Candidate Genes

On July 31, two candidate genes (Glyma11g03850, and Glyma13g01360) were associated with CC, where the –log10(P) of SNPs were more than 4.956. One candidate gene (Glyma07g16400) was found to be associated with CC on July 14.

DISCUSSION

Edamame CCs are significant for plant breeders as rapid CC naturally suppresses weeds and increases photosynthetic activity. Because of this, several GWAS studies have been published in canopy cover studies in soybean. The high-density marker set available in the SoySNP50K repository is a powerful tool to unlock the genetic potential of several traits as suggested by Lee (2019) and Jarquin (2016). In this study, we identified novel associations for CC in edamame using the diverse edamame accessions via GWASs.

A drawback for such association study is the capacity to collect the phenotypic data for a genetically diverse population at field scale. With the increasing use of UAVs in field research, measuring canopy cover from drone images has become an ideal approach and many GWAS QTL that are associated with CC have been discovered in legumes (Cazenave et al. 2019; Sarkar et al. 2020; Xavier et al. 2017).

We found 23 candidate genes from two different indices, from two different drones/cameras in two years for canopy cover in edamame. We found 11 candidate genes from EXG index from phantom drone having RGB camera in 2020 viz. Glyma07g07410, Glyma01g10320, Glyma01g13430, Glyma01g13654, Glyma01g13930, Glyma01g09610, Glyma11g10480, Glyma11g10370, Glyma02g24071, Glyma07g07393, and Glyma09g31460.

We found 3 candidate genes from EXG index from Inspire-2 drone having MS (multispectral) camera in 2021 viz. Glyma11g03850, and Glyma13g01360, and Glyma07g16400. Also, in the same year and using same drone/camera but with NDVI index we found following 6 candidate genes namely, Glyma02g00760, Glyma02g11660, Glyma02g00640, Glyma02g00730, Glyma02g11660, and Glyma18g01010. Using EXG index from phantom drone having RGB camera in 2021 we found 3 candidate genes for canopy cover. Those three candidate genes are Glyma03g30900, Glyma14g05110, and Glyma19g40750.

We did not find any common candidate genes that were found for both years. This could be related to environmental changes between different years. Even in the same year, 2021 with different indices and drones/cameras, we also did not find common candidate genes. This could mean that different drones and different indices captured different aspects of canopy cover. Alternatively, it could be that some SNPs are borderline significant in one index but not in the other.

CONCLUSION

In conclusion, we performed phenotypic analysis of a collection of edamame varieties over two years using RGB and MS images collected via two different drones. Using known genetic markers and genes that are associated with canopy cover, we found 23 novel candidate genes that might be related to the canopy cover.

REFERENCES

Adak, Alper, Seth C. Murray, Steven L. Anderson, Sorin C. Popescu, Lonesome Malambo, M. Cinta Romay, and Natalia de Leon. 2021. "Unoccupied Aerial Systems Discovered Overlooked Loci Capturing the Variation of Entire Growing Period in Maize." *The Plant Genome* 14(2):e20102.

Anderson, Steven L., Seth C. Murray, Yuanyuan Chen, Lonesome Malambo, Anjin Chang, Sorin Popescu, Dale Cope, and Jinha Jung. 2020. "Unoccupied Aerial System Enabled Functional Modeling of Maize Height Reveals Dynamic Expression of Loci." *Plant Direct* 4(5):e00223.

Anon. 2021. "The Gene Ontology Resource: Enriching a GOld Mine." *Nucleic Acids Research* 49(D1):D325--D334.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and others. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25(1):25–29.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait, and E. Huala. 2015. "The Arabidopsis Information Resource: Making and Mining the 'Gold Standard' Annotated Reference Plant Genome. Genesis 53."

Cazenave, Alexandre-Brice, Kushendra Shah, Tresa Trammell, Michael Komp, Justin Hoffman, Christy M. Motes, and Maria J. Monteros. 2019. "High-Throughput Approaches for Phenotyping Alfalfa Germplasm under Abiotic Stress in the Field." *The Plant Phenome Journal* 2(1):1–13.

Chen, Yiwu, and Randall L. Nelson. n.d. *Evaluation and Classification of Leaflet Shape and Size in Wild Soybean*.

Fenta, Berhanu A., Stephen E. Beebe, Karl J. Kunert, James D. Burridge, Kathryn M. Barlow, Jonathan P. Lynch, and Christine H. Foyer. 2014. "Field Phenotyping of Soybean Roots for Drought Stress Tolerance." *Agronomy* 4(3):418–35. doi: 10.3390/agronomy4030418.

Jarquin, Diego, James Specht, and Aaron Lorenz. 2016. "Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions." *G3: Genes, Genomes, Genetics* 6(8):2329–41. doi: 10.1534/g3.116.031443.

Jombart, Thibaut. 2008. "Analyses Multivariées de Marqueurs Génétiques: Développements Méthodologiques, Applications et Extensions." Lyon 1.

Jombart, Thibaut, Sébastien Devillard, and François Balloux. 2010. "Discriminant Analysis of Principal Components: A New Method for the Analysis of Genetically Structured Populations." *BMC Genetics* 11(1):1–15.

Lee, Ji Yong, Michael P. Popp, Elijah J. Wolfe, Rodolfo M. Nayga, Jennie S. Popp, Pengyin Chen, and Han Seok Seo. 2018. "Information and Order of Information Effects on Consumers' Acceptance and Valuation for Genetically Modified Edamame Soybean." *PLoS ONE* 13(10). doi: 10.1371/journal.pone.0206300.

Lee, Sungwoo, Kyujung Van, Mikyung Sung, Randall Nelson, Jonathan LaMantia, Leah K. McHale, and M. A. Rou. Mian. 2019. "Genome-Wide Association Study of Seed Protein, Oil and Amino Acid Contents in Soybean from Maturity Groups I to IV." *Theoretical and Applied Genetics* 132(6):1639–59. doi: 10.1007/s00122-019-03304-5.

Li, J., and L. Ji. 2005. “Adjusting Multiple Testing in Multilocus Analyses Using the Eigenvalues of a Correlation Matrix.” *Heredity* 95(3):221–27.

Liu, Baohui, Satoshi Watanabe, Tomoo Uchiyama, Fanjiang Kong, Akira Kanazawa, Zhengjun Xia, Atsushi Nagamatsu, Maiko Arai, Tetsuya Yamada, Keisuke Kitamura, Chikara Masuta, Kyuya Harada, and Jun Abe. 2010. “The Soybean Stem Growth Habit Gene Dt1 Is an Ortholog of Arabidopsis TERMINAL FLOWER1.” *Plant Physiology* 153(1):198–210. doi: 10.1104/pp.109.150607.

Liu, Hai-Jun, and Jianbing Yan. 2019. “Crop Genome-Wide Association Study: A Harvest of Biological Relevance.” *The Plant Journal* 97(1):8–18.

Liu, Yucheng, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo An Zhou, Haikuan Zhang, Zhi Liu, Miao Shi, Xuehui Huang, Yan Li, Min Zhang, Zheng Wang, Baoge Zhu, Bin Han, Chengzhi Liang, and Zhixi Tian. 2020. “Pan-Genome of Wild and Cultivated Soybeans.” *Cell* 182(1):162-176.e13. doi: 10.1016/j.cell.2020.05.023.

Lyra, Danilo H., Nicolas Virlet, Pouria Sadeghi-Tehran, Kirsty L. Hassall, Luzie U. Wingen, Simon Orford, Simon Griffiths, Malcolm J. Hawkesford, and Gancho T. Slavov. 2020. “Functional QTL Mapping and Genomic Prediction of Canopy Height in Wheat Measured Using a Robotic Field Phenotyping Platform.” *Journal of Experimental Botany* 71(6):1885–98.

Mahoussi, Kadoukpe Arnaud Djanta, Etchikinto Agoyi Eric, Agbahoungba Symphorien, Jean-Baptiste Quenum Florent, Josiane Chadare Flora, Ephrem Assogbadjo Achille, Agbangla Clement, and Sinsin Brice. 2020. “Vegetable Soybean, Edamame: Research, Production, Utilization and Analysis of Its Adoption in Sub-Saharan Africa.” *Journal of Horticulture*

and Forestry 12(1):1–12. doi: 10.5897/jhf2019.0604.

Mentreddy, S. R., A. I. Mohamed, N. Joshee, A. K. Yadav, and others. 2002. “Edamame: A Nutritious Vegetable Crop.” Pp. 432–38 in *Trends in new crops and new uses. Proceedings of the Fifth National Symposium, Atlanta, Georgia, USA, 10-13 November, 2001*.

Mi, Huaiyu, Anushya Muruganujan, and Paul D. Thomas. 2012. “PANTHER in 2013: Modeling the Evolution of Gene Function, and Other Gene Attributes, in the Context of Phylogenetic Trees.” *Nucleic Acids Research* 41(D1):D377--D386.

Purcell, Larry C. 2000. “Soybean Canopy Coverage and Light Interception Measurements Using Digital Imagery.” *Crop Science* 40(3):834–37.

Qin, Jun, Ainong Shi, Qijian Song, Song Li, Fengmin Wang, Yinghao Cao, Waltram Ravelombola, Qi Song, Chunyan Yang, and Mengchen Zhang. 2019. “Genome Wide Association Study and Genomic Selection of Amino Acid Concentrations in Soybean Seeds.” *Frontiers in Plant Science* 10:1445.

Sarkar, Sayantan, Alexandre-Brice Cazenave, Joseph Oakes, David McCall, Wade Thomason, Lynn Abbot, and Maria Balota. 2020. “High-Throughput Measurement of Peanut Canopy Height Using Digital Surface Models.” *The Plant Phenome Journal* 3(1):e20003.

Shurtleff, William, Akiko Aoyagi, and others. 2009. “History of Edamame, Green Vegetable Soybeans, and Vegetable-Type Soybeans (1275-2009)[Electronic Resource].”

Song, Qijian, David L. Hyten, Gaofeng Jia, Charles V. Quigley, Edward W. Fickus, Randall L. Nelson, and Perry B. Cregan. 2013. “Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean.” *PLoS ONE* 8(1). doi: 10.1371/journal.pone.0054985.

Turner, Stephen D. 2014. “Qqman: An R Package for Visualizing GWAS Results Using QQ and Manhattan Plots.” *Biorxiv* 5165.

Ulloa, Santiago M., Avishek Datta, Goran Malidza, Robert Leskovsek, and Stevan Z. Knezevic. 2010. “Yield and Yield Components of Soybean [Glycine Max (L.) Merr.] Are Influenced by the Timing of Broadcast Flaming.” *Field Crops Research* 119(2–3):348–54. doi: 10.1016/j.fcr.2010.08.006.

Wang, Xiaqing, Ruyang Zhang, Wei Song, Liang Han, Xiaolei Liu, Xuan Sun, Meijie Luo, Kuan Chen, Yunxia Zhang, Hao Yang, and others. 2019. “Dynamic Plant Height QTL Revealed in Maize through Remote Sensing Phenotyping Using a High-Throughput Unmanned Aerial Vehicle (UAV).” *Scientific Reports* 9(1):1–10.

Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. “Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine Max.” *Genetics* 206(2):1081–89. doi: 10.1534/genetics.116.198713.

Xie, Dongwei, Zhigang Dai, Zemao Yang, Jian Sun, Debao Zhao, Xue Yang, Liguo Zhang, Qing Tang, and Jianguang Su. 2018. “Genome-Wide Association Study Identifying Candidate Genes Influencing Important Agronomic Traits of Flax (Linum Usitatissimum L.) Using SLAF-Seq.” *Frontiers in Plant Science* 8:2232.

Tables and Figures

Table 2.2.1 Candidate gene and descriptions of the significantly associated SNPs for CC obtained from EXG index from RGB Images in 2020 using Wm82.a2.v1.

SNP (BP)	**Corresponding Gene ID**	**Location**	**P**	**Date**	**Gene Function Description**
ss715611125 (7494953)	Glyma11g10480	Gm11:7494447: 7495419	6.08	July29	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD. Peptidyl-prolyl cis-trans isomerase activity
ss715611120 (7444642)	Glyma11g10370	Gm11:7444199: 7444981	6.06	July29	N/A
ss715611120 (7444642)	Glyma11g10380	Gm11:7446107: 7447692	6.06	July29	Chalcone and stilbene synthases, N-terminal domain Transferase activity, transferring acyl groups
ss715581587 (24460388)	Glyma02g24071	Gm02:24448880: 24450074	5.98	July29	Importin beta binding domain Protein transporter activity AAA-type ATPase family protein
ss715603992 (38168295)	Glyma09g31460	Gm09:38159423: 38163907	5.32	Aug02	Acetylglucosaminyltransferase activity Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
ss715598441 (6108702)	Glyma07g07410	Gm07:6107440: 6109415	5.47	July17	Putative membrane lipoprotein Embryo Sac Development Arrest 8
ss715578480 (13374753)	Glyma01g10320	Gm01:13365717: 13365849	5.13	July17	N/A
ss715578590 (16841932)	Glyma01g13430	Gm01:16827663: 16847408	5.13	July17	E3 ubiquitin ligase involved in syntaxin degradation Zinc finger, C3HC4 type (RING finger)

ss715578590 (16841932)	Glyma01g13654	Gm01:16827428: 16868761	5.13	July17	Ring Finger Protein-related E3 ubiquitin-protein ligase BRE1 involved in syntaxin degradation Zinc finger, C3HC4 type (RING finger) Histone mono-ubiquitination 1

Base Pair (BP); Probability (P)

Gene function as described in TAIR, PANTHER, or GO annotation.

Table 2.2.2 Candidate gene and descriptions of the significantly associated SNPs for CC obtained from EXG index from RGB Images in 2021 using Wm82.a2.v1.

SNP (BP)	**Corresponding Gene ID**	**Location**	**P**	**Date**	**Gene Function Description**
ss715585755 (38751677)	Glyma03g30900	Gm03:38741322: 38744676	6.18	July31	F-box family protein
ss715618478 (3544969)	Glyma14g05110	Gm14:3543800: 3545480	5.50	July31	Protein of unknown function (DUF861)
ss715635643 (47085574)	Glyma19g40750	Gm19:47086840: 47087409	5.07	July31	Short-chain Dehydrogenases/Reductase

Base Pair (BP); Probability (P)

Gene function as described in TAIR, PANTHER, or GO annotation.

Table 2.2.3 Candidate gene and descriptions of the significantly associated SNPs for CC obtained from NDVI from MS Images in 2021 using Wm82.a2.v1.

SNP (BP)	**Corresponding Gene ID**	**Location**	**P**	**Date**	**Gene Function Description**
ss715583614 (549828)	Glyma02g00760	Gm02:564299: 568839	5.41	July31	Plant pleckstrin homology-like region
ss715583635 (562974)	Glyma02g00760	Gm02:564214: 569593	5.31	July31	Auxin canalization/ Plant protein of unknown function (DUF828) with plant pleckstrin homology-like region)
/ss715584302 (9860461)	Glyma02g11660	Gm02:9860461: 9860461	5.31	July22	UDP- Glucosyl/UDP-Glucouronosyl Transferases
ss715582577 (434129)	Glyma02g00640	Gm02:434129: 434129	5.22	July31	Xenotropic and Polytropic Retrovirus Receptor 1 (Protein SYG1 homolog)/Predicted small molecule transporter SPX domain/Integral component of membrane)/ Phosphate 1
ss715583581 (520296)	Glyma02g00730	Gm02:520056: 524702	5.03	July31	Protein coding neurogenic locus notch-like protein
ss715631071 (499478)	Glyma18g01010	Gm18:495174: 501637	4.96	July14	Meiosis specific protein HOP1/Mitotic nuclear division DNA-binding HORMA family protein

Base Pair (BP); Probability (P)

Gene function as described in TAIR, PANTHER, or GO annotation.

Table 2.2.4 Candidate gene and descriptions of the significantly associated SNPs for ExG Index from MS Images in 2021 using Wm82.a2. v1.

SNP (BP)	**Corresponding Gene ID**	**Location**	**P**	**Date**	**Gene Function Description**
ss715596520 (16078614)	Glyma07g16400	Gm07:16066951: 16078950	6.94	July14	Regulator of chromosome condensation (RCC1)
ss715609766 (2575660)	Glyma11g03850	Gm11:2574275: 2576090	5.45	July31	Transcription factor HEX, contains HOX and HALZ domains/Homeobox associated leucine zipper DNA binding
ss715613725 (1047028)	Glyma13g01360	Gm13:1047028: 1047028	5.13	July31	Diacylglycerol kinase 7 accessory domain

Base Pair (BP); Probability (P)

Gene function as described in TAIR, PANTHER, or GO annotation.

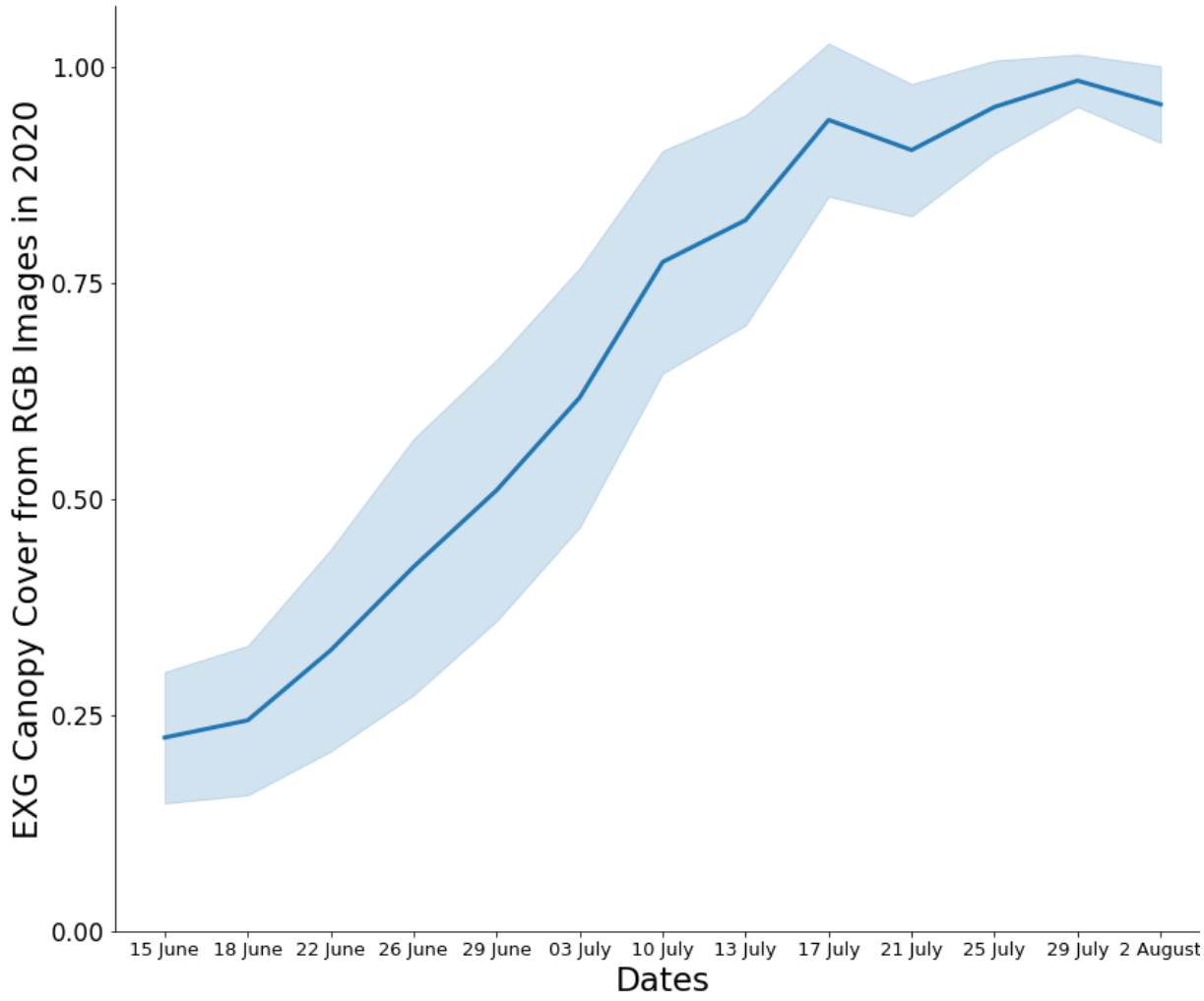


Figure 2.2.1 Line plot showing the percentage of canopy cover using EXG index obtained from RGB images over time in 2020. The bold line in the figure indicates the average value while the shaded area indicates the standard deviation canopy cover over the growing season.

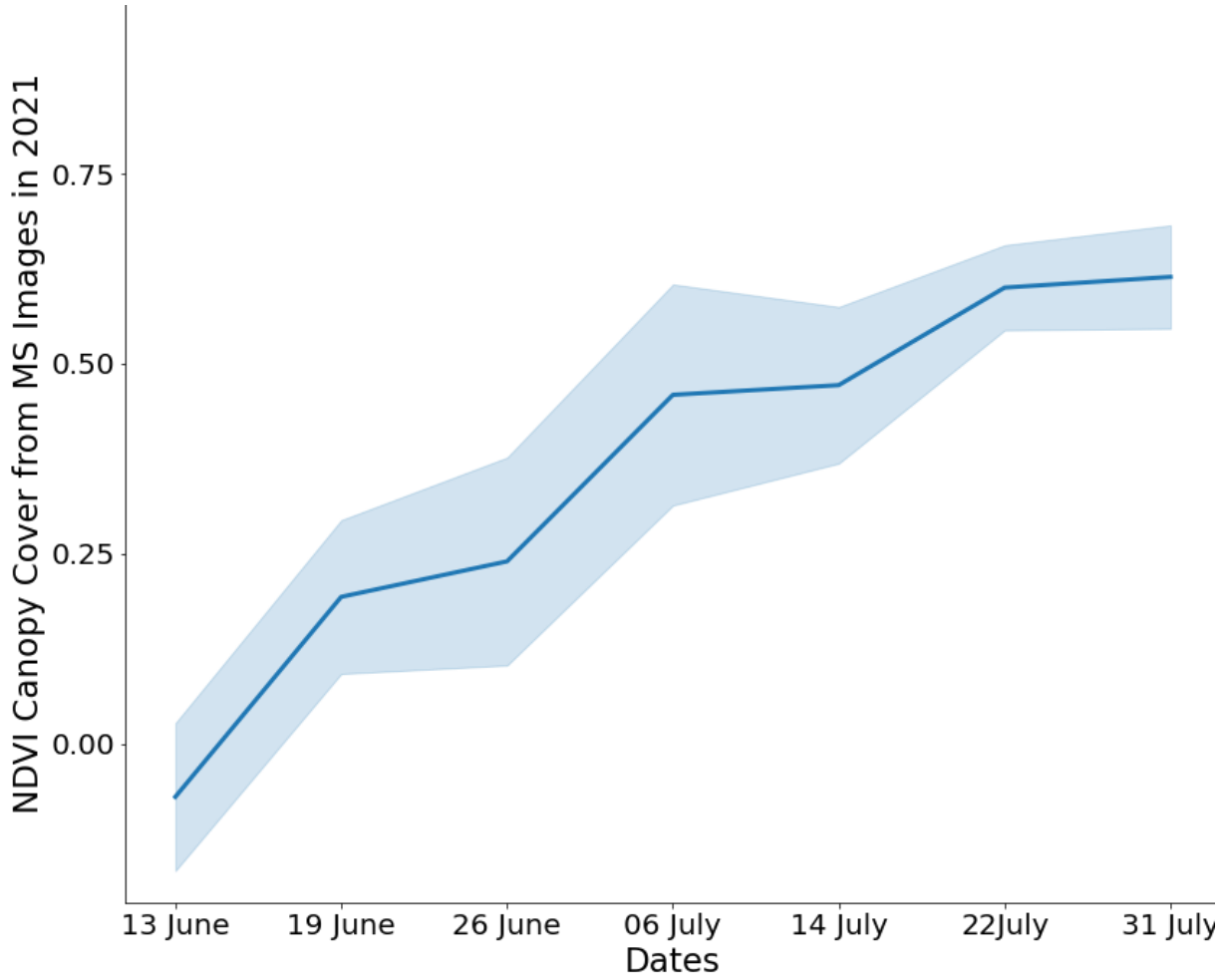


Figure 2.2.2 Line plot showing the percentage of canopy cover using NDVI index obtained from MS images over time in 2021. The bold line in the figure indicates the average value while the shaded area indicates the standard deviation canopy cover over the growing season.

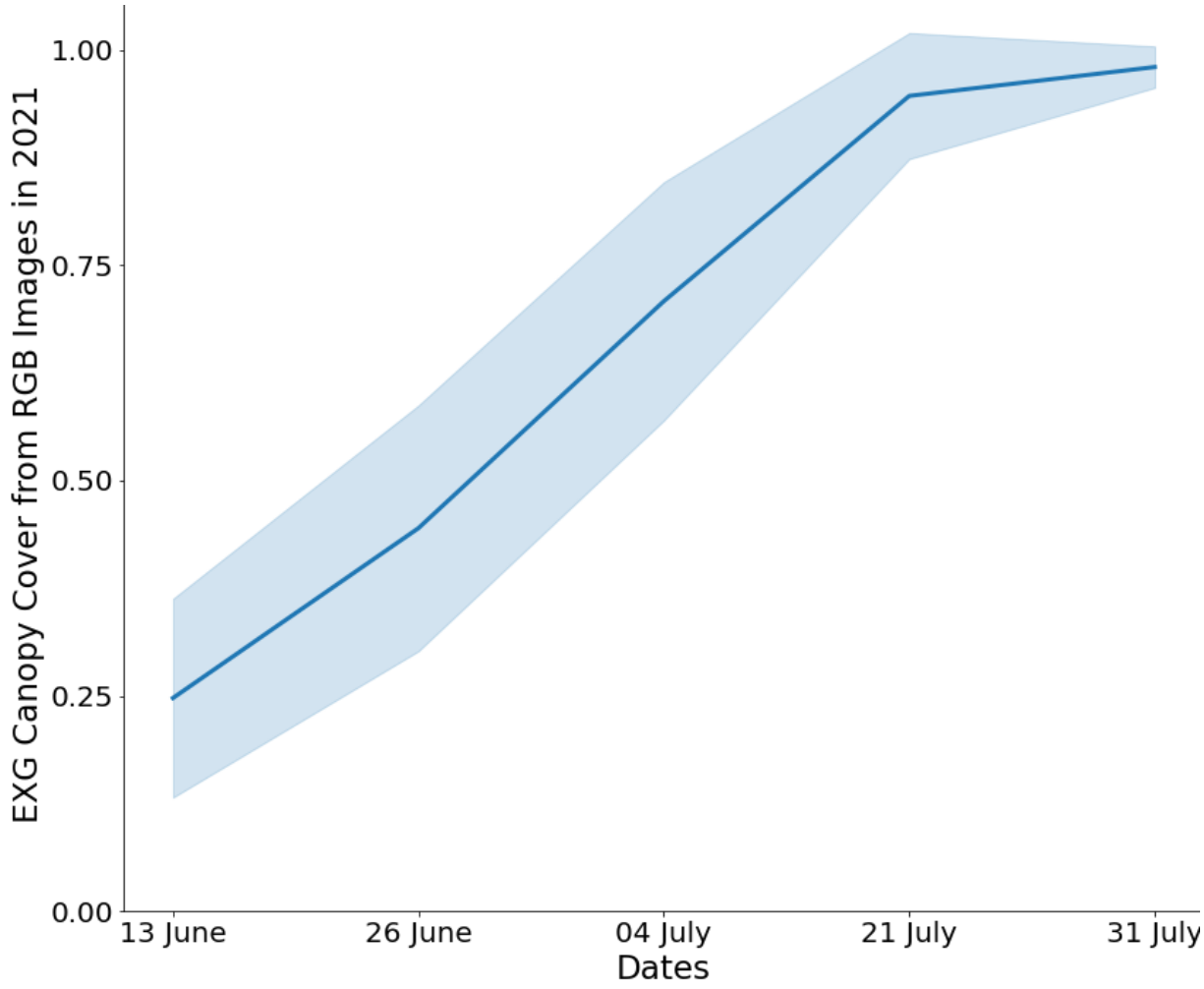


Figure 2.2.3 Line plot showing the percentage of canopy cover using EXG index obtained from RGB images over time in 2021. The bold line in the figure indicates the average value while the shaded area indicates the standard deviation canopy cover over the growing season.

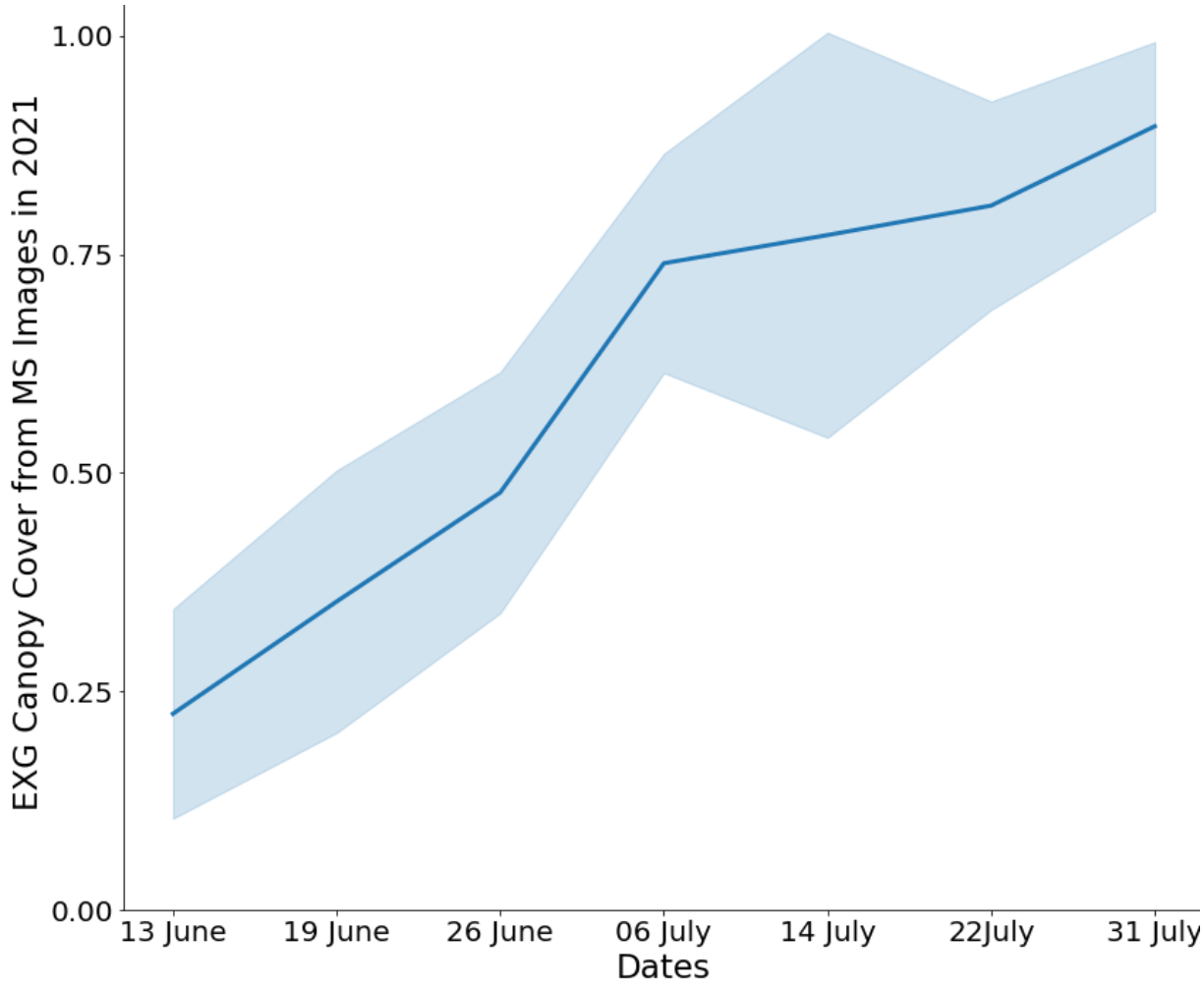
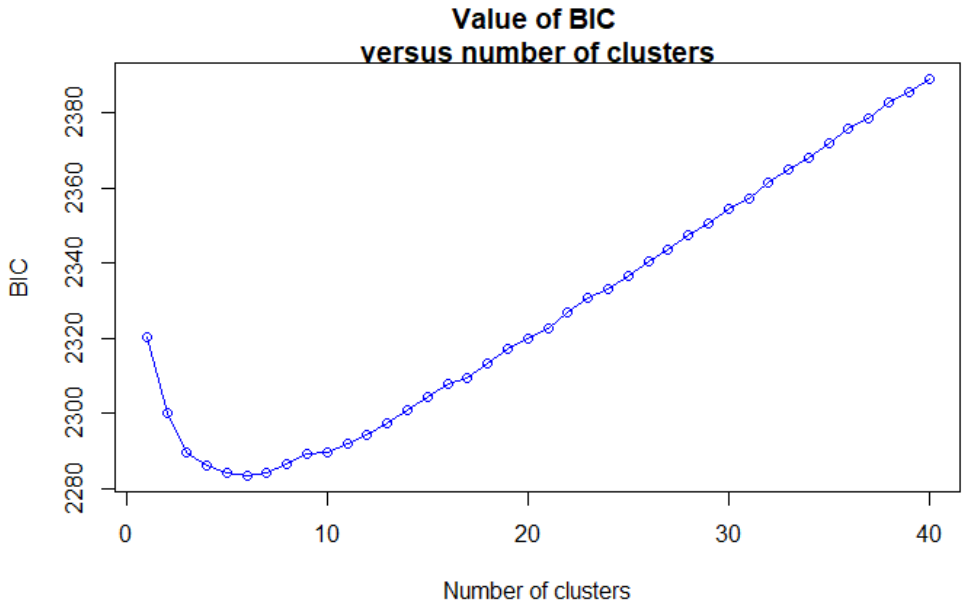


Figure 2.2.4 Line plot showing the percentage of canopy cover using EXG index obtained from MS images over time in 2021. The bold line in the figure indicates the average value while the shaded area indicates the standard deviation canopy cover over the growing season.



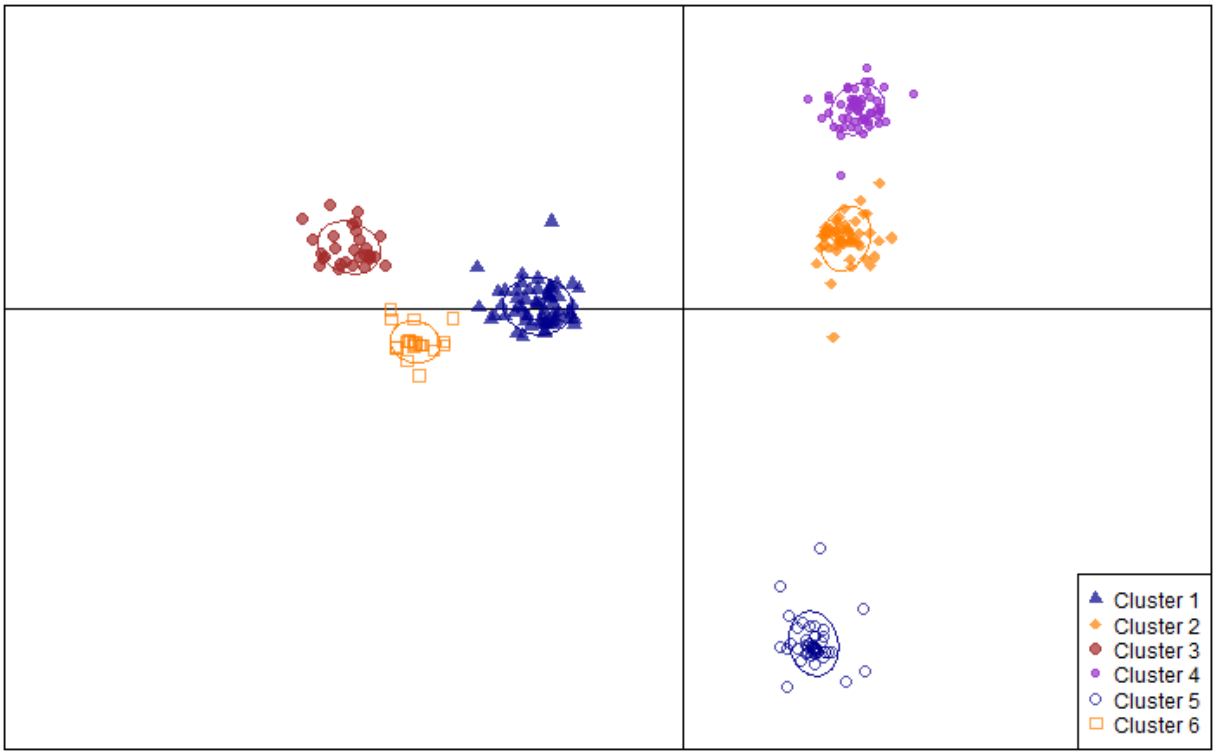
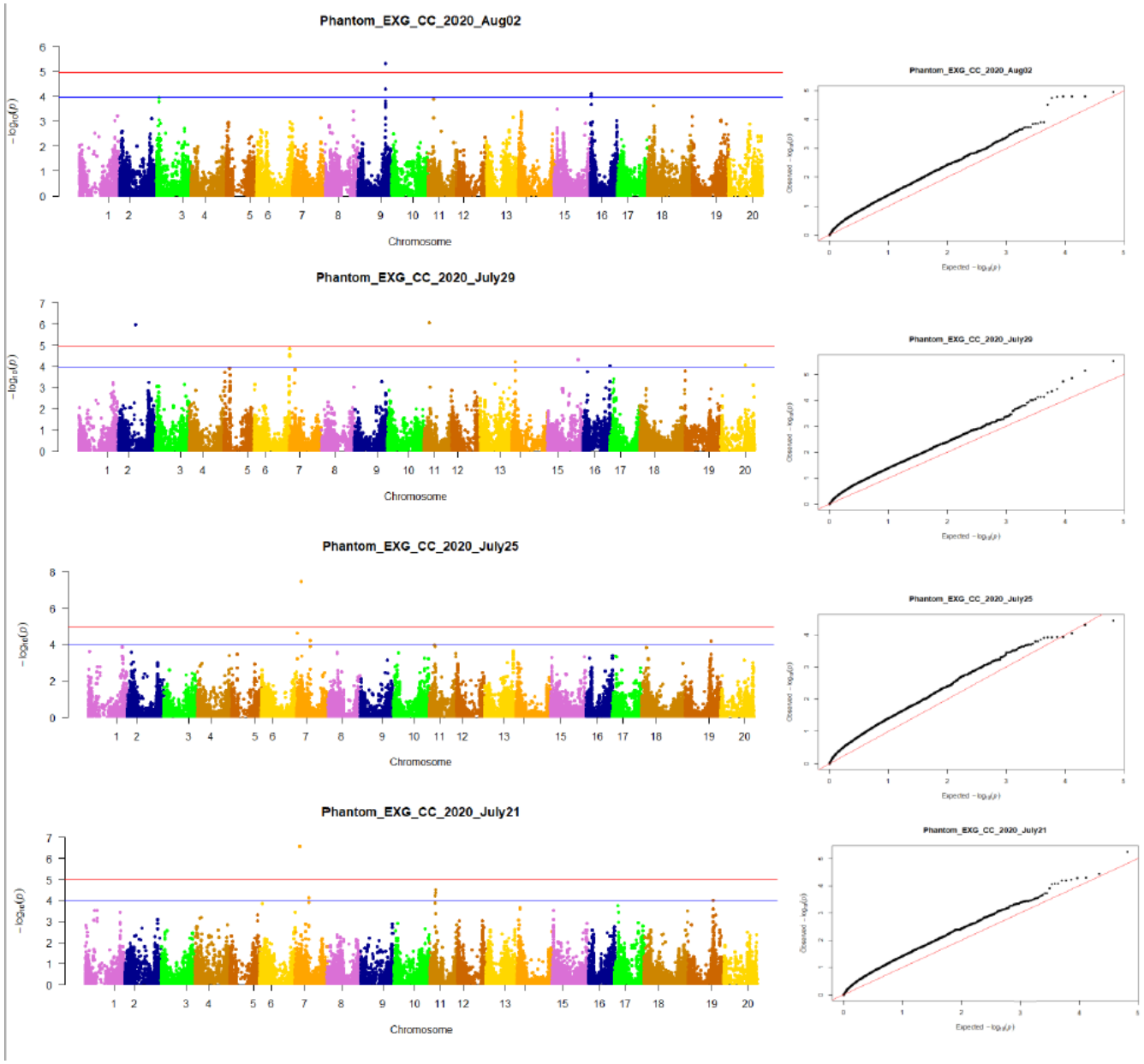
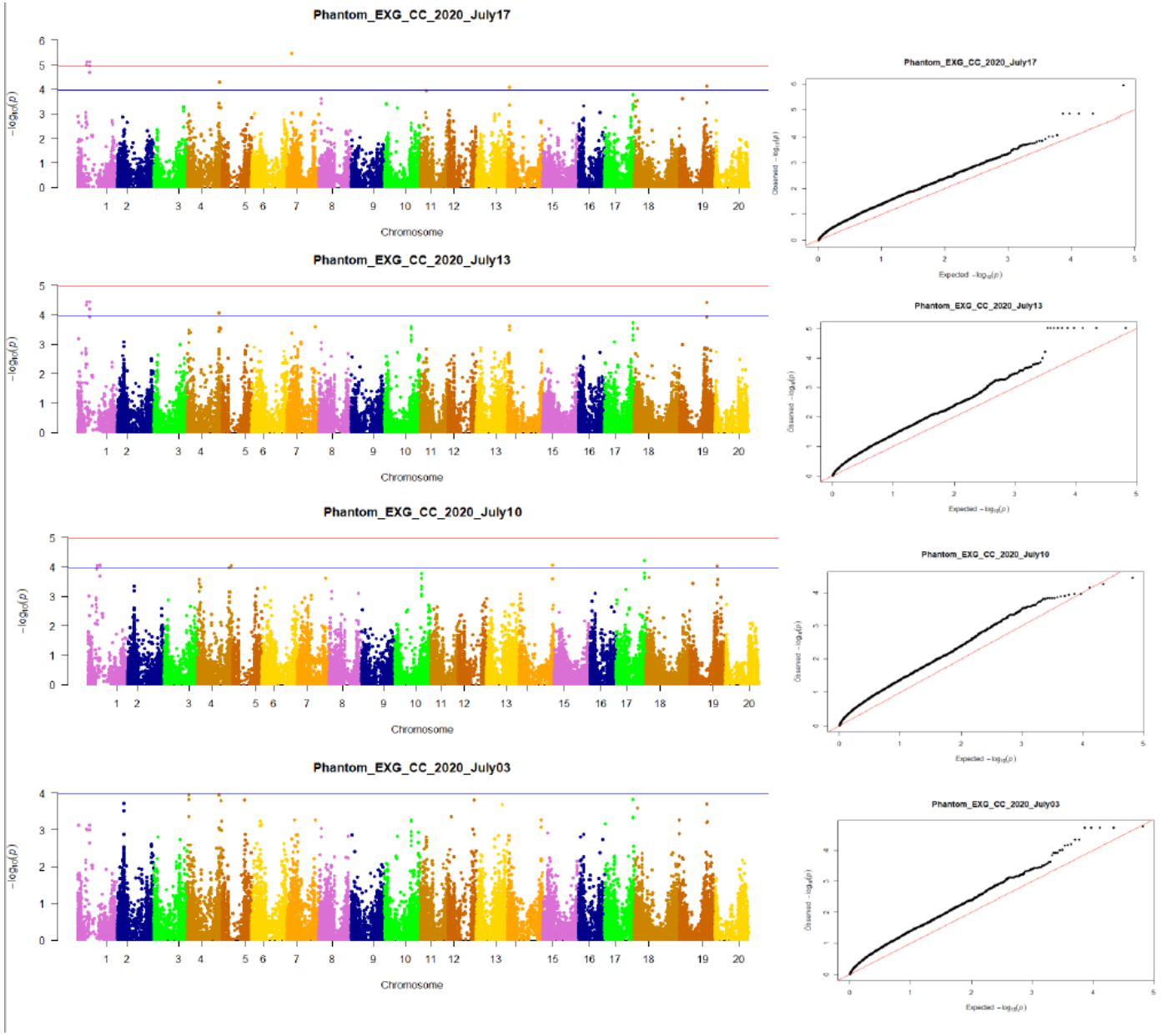
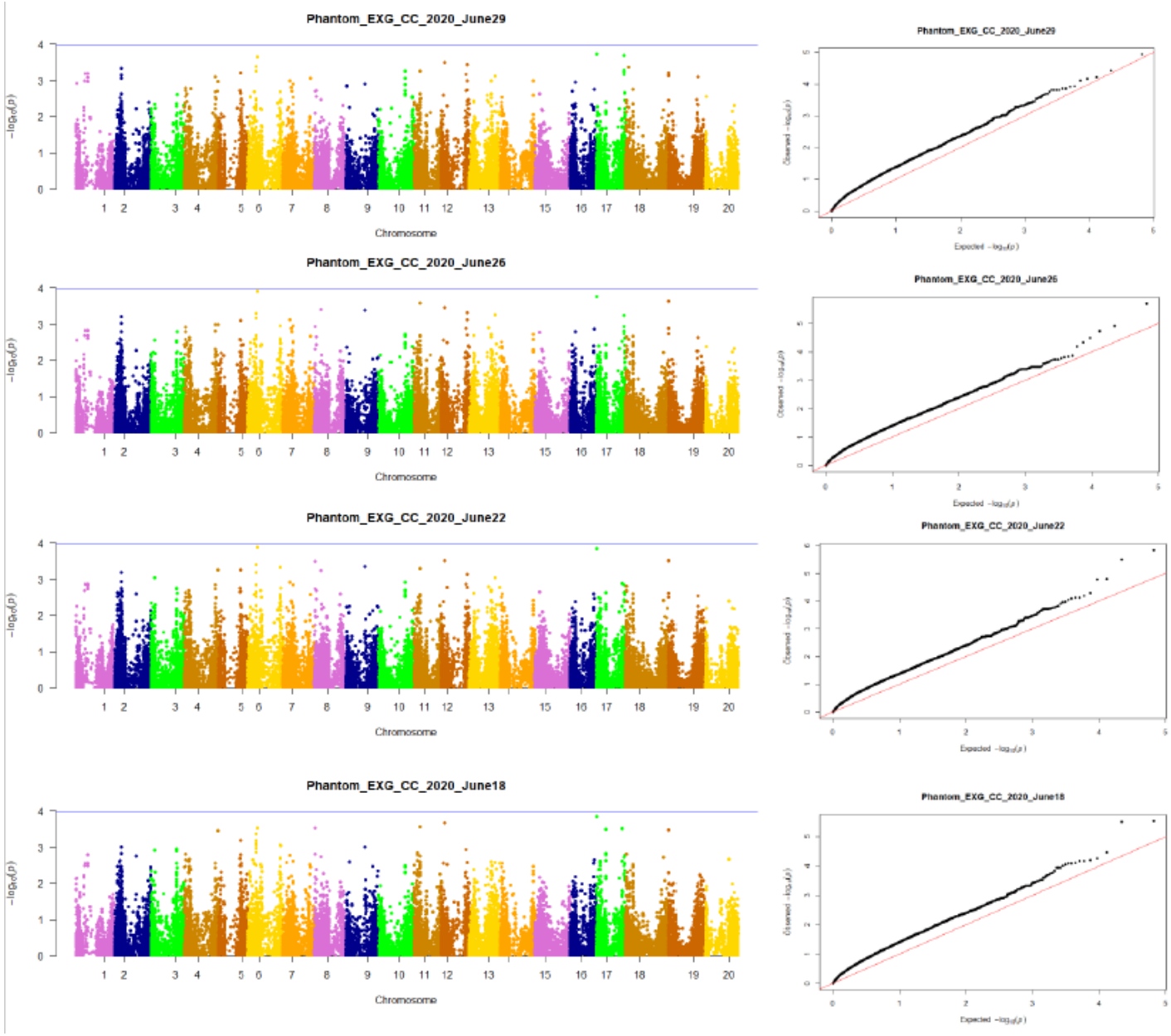


Figure 2.2.5 (a) Bayesian Information Criterion (BIC) used to select the optimal number of clusters. (b) A scatter plot showing the 6 clusters (k=3) identified as likely subpopulations within the 269 accessions in 2020.


Phantom_EXG_CC_2020_Aug02
Chromosome
Phantom_EXG_CC_2020_July29
Chromosome
Phantom_EXG_CC_2020_July25
Chromosome
Phantom_EXG_CC_2020_July21
Chromosome


Phantom_EXG_CC_2020_July17
Chromosome
Phantom_EXG_CC_2020_July17
Phantom_EXG_CC_2020_July13
Chromosome
Phantom_EXG_CC_2020_July13
Phantom_EXG_CC_2020_July10
Chromosome
Phantom_EXG_CC_2020_July10
Phantom_EXG_CC_2020_July03
Chromosome
Phantom_EXG_CC_2020_July03


Phantom_EXG_CC_2020_June29
Chromosome
Phantom_EXG_CC_2020_June29
Phantom_EXG_CC_2020_June26
Chromosome
Phantom_EXG_CC_2020_June26
Phantom_EXG_CC_2020_June22
Chromosome
Phantom_EXG_CC_2020_June22
Phantom_EXG_CC_2020_June18
Chromosome
Phantom_EXG_CC_2020_June18

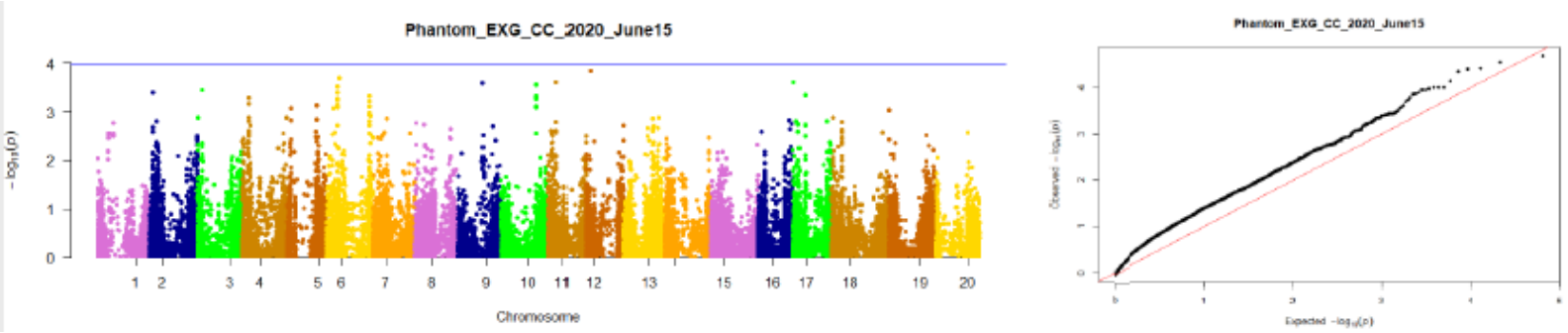


Figure 2.2.6 Manhattan plots and QQ-plots of CC obtained from EXG index from RGB Images in 2020.

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as $-\log_{10}(P) > 4.956$ as represented by red line and suggestive threshold as $-\log_{10}(P) > 3.967$ as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.

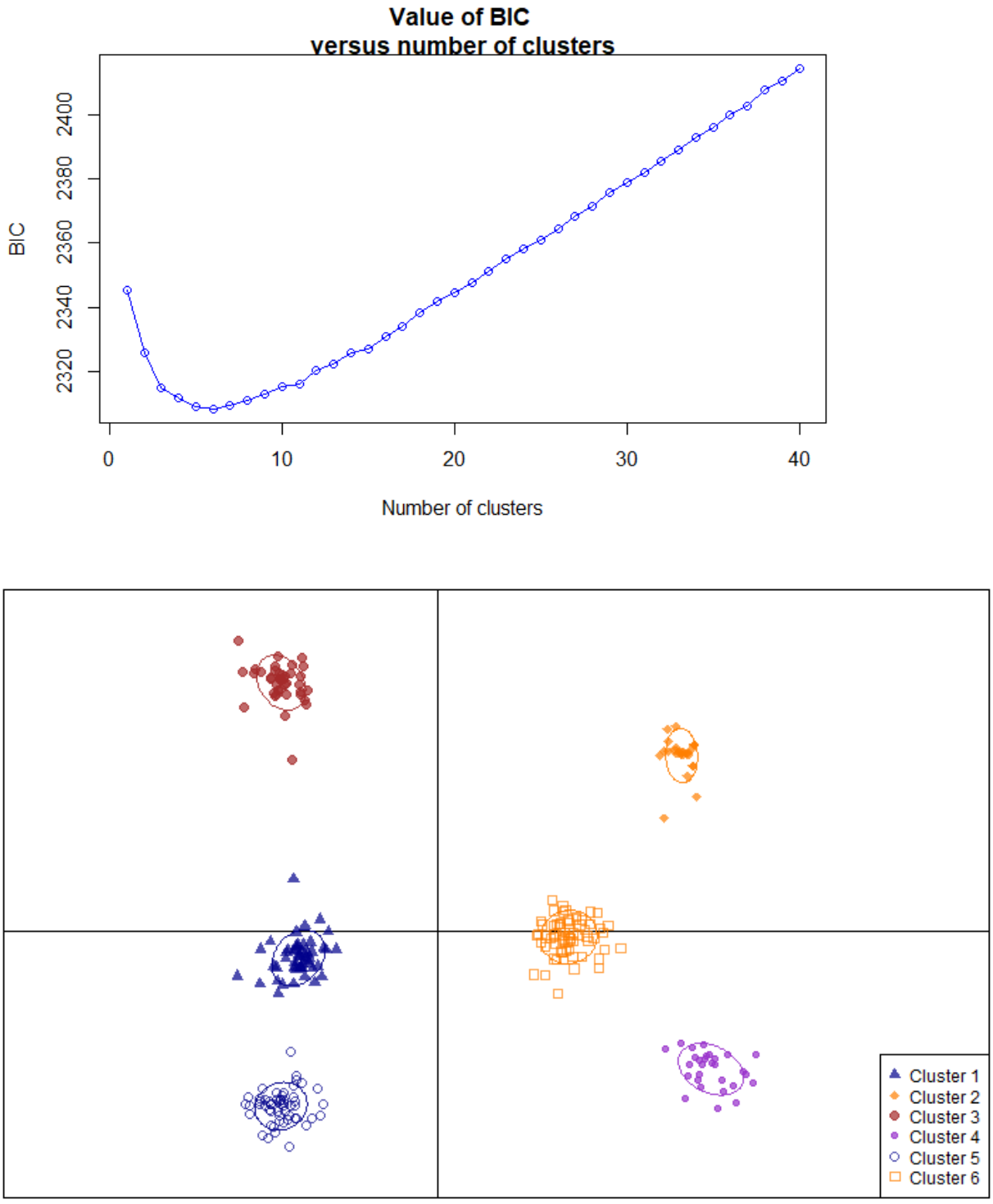


Figure 2.2.7 (a) Bayesian Information Criterion (BIC) used to select the optimal number of clusters in 2021. (b) A scatter plot showing the 6 clusters (k=6) identified as likely subpopulations within the 272 accessions in 2021.

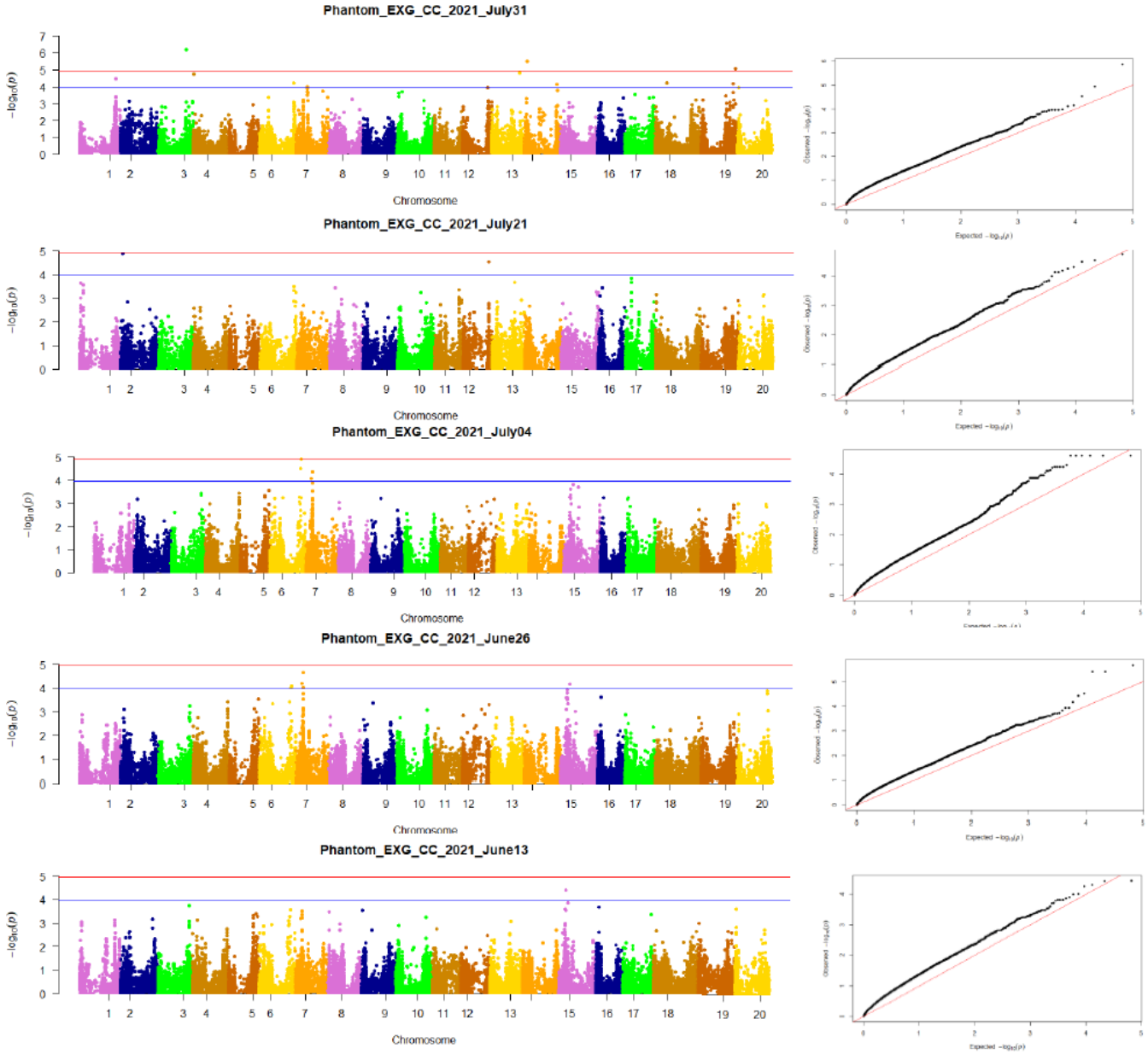
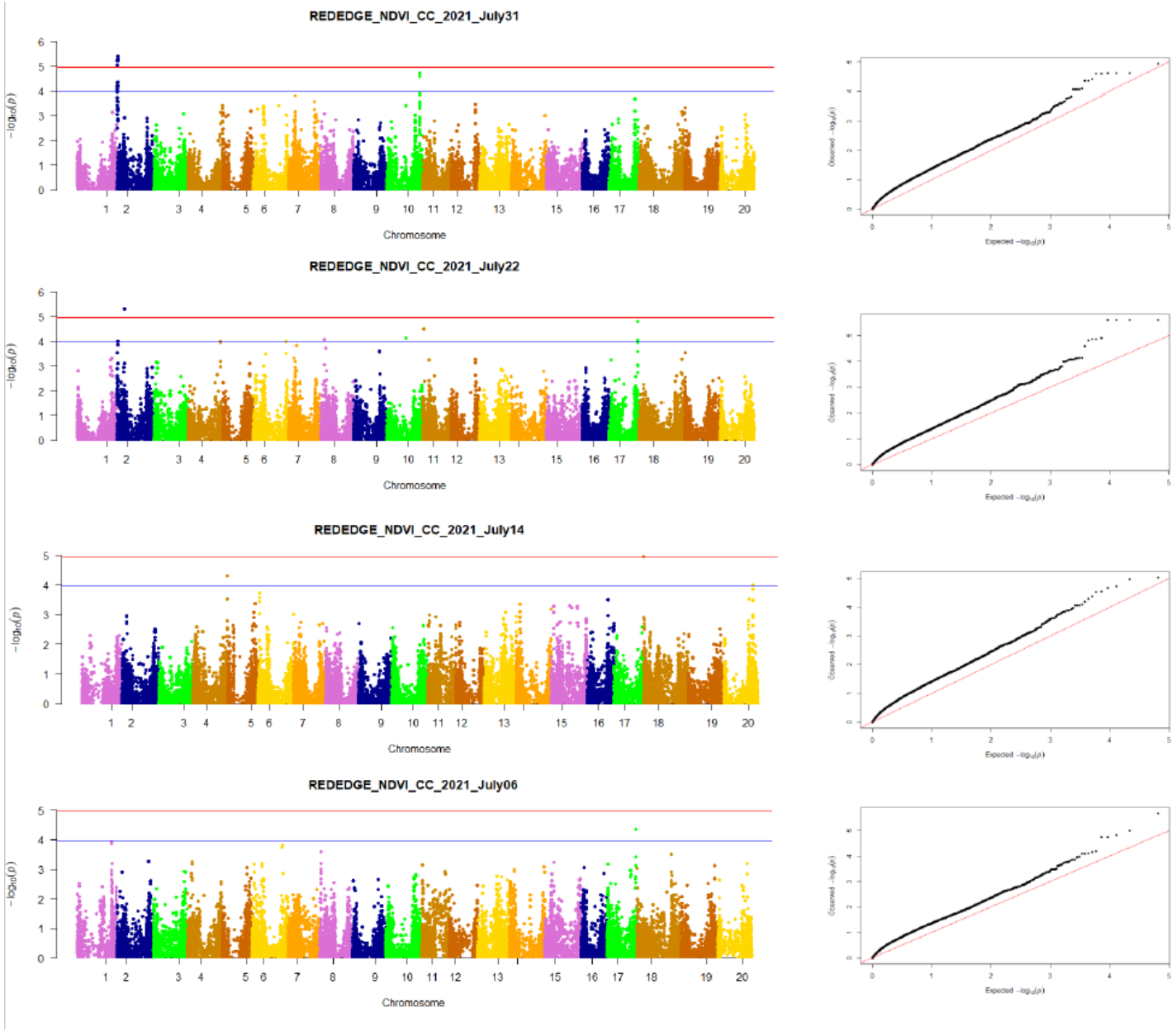

Phantom_EXG_CC_2021_July31
Chromosome
Phantom_EXG_CC_2021_July21
Chromosome
Phantom_EXG_CC_2021_July04
Chromosome
Phantom_EXG_CC_2021_June26
Chromosome
Phantom_EXG_CC_2021_June13
Chromosome

Figure 2.2.8 Manhattan plots and QQ-plots of CC obtained from EXG index from RGB Images in 2021.

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as $-\log_{10}(P) > 4.956$ as represented by red line and suggestive threshold as $-\log_{10}(P) > 3.967$ as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.


REDEDGE_NDVI_CC_2021_July31
REDEDGE_NDVI_CC_2021_July22
REDEDGE_NDVI_CC_2021_July14
REDEDGE_NDVI_CC_2021_July06
Chromosome

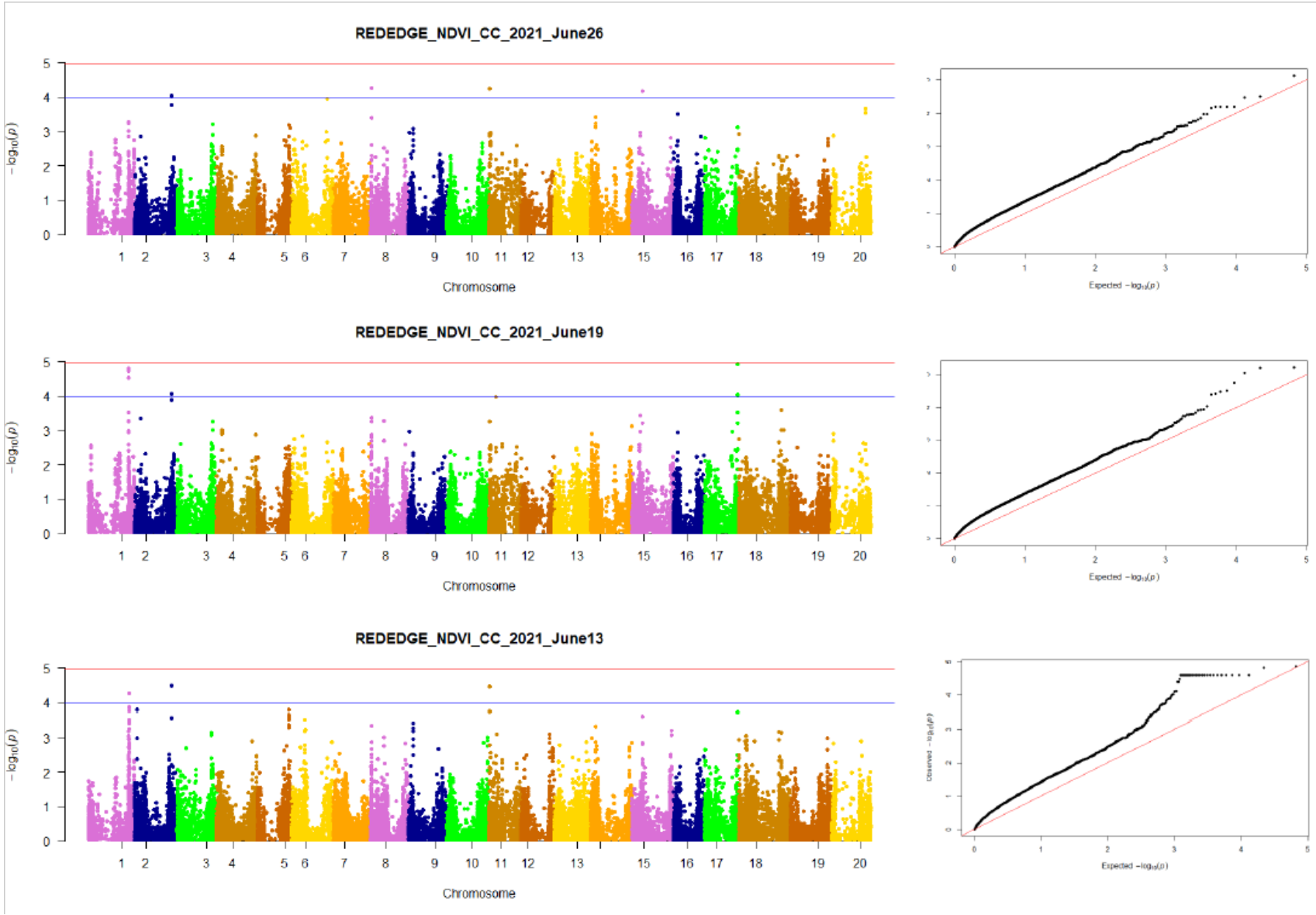
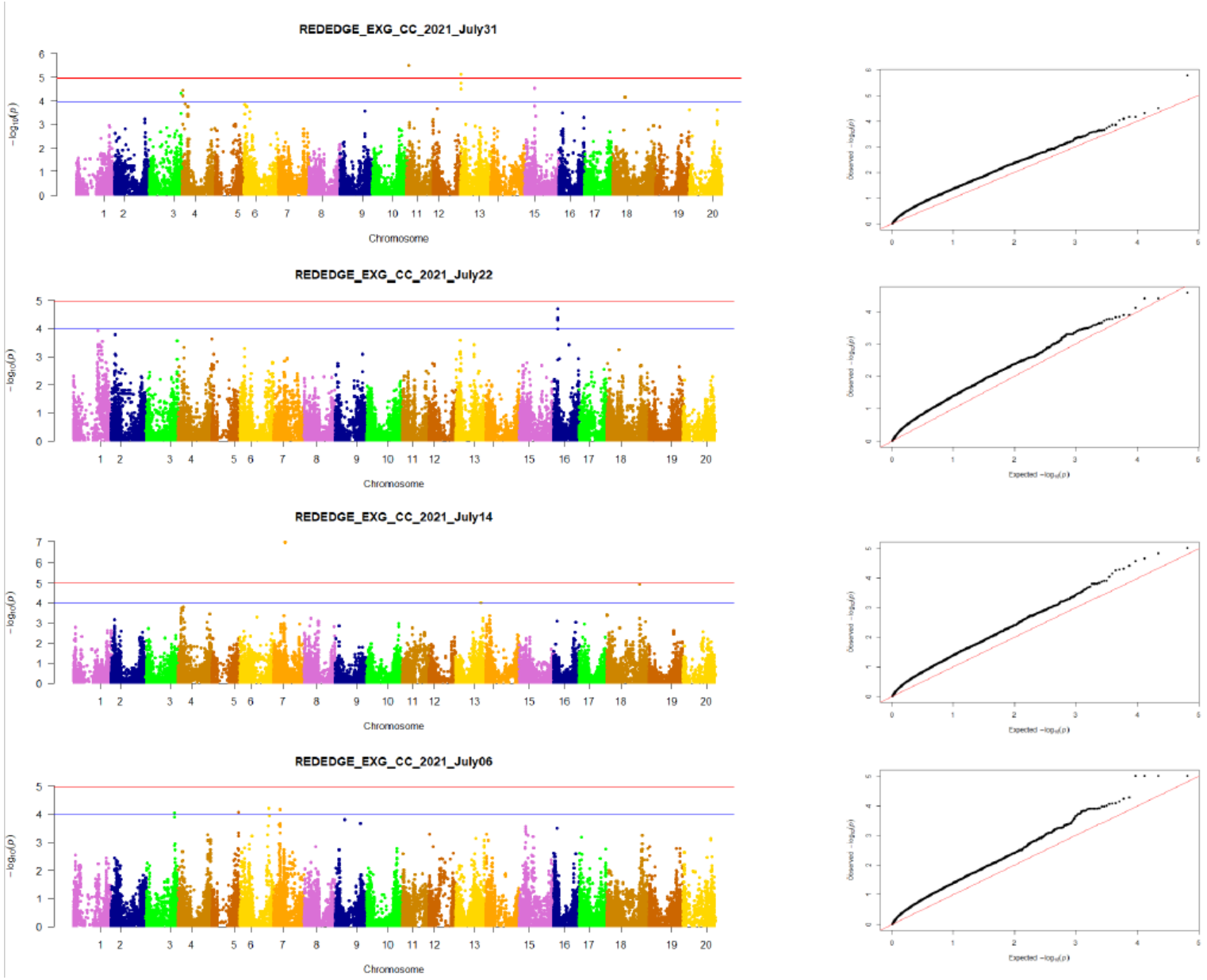

REDEDGE_NDVI_CC_2021_June26
REDEDGE_NDVI_CC_2021_June19
REDEDGE_NDVI_CC_2021_June13
Chromosome
Expected

Figure 2.2.9 Manhattan plots and QQ-plots of CC obtained from NDVI from MS Images in 2021.

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as $-\log_{10}(P) > 4.956$ as represented by red line and suggestive threshold as $-\log_{10}(P) > 3.967$ as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.


REDEDGE_EXG_CC_2021_July31
REDEDGE_EXG_CC_2021_July22
REDEDGE_EXG_CC_2021_July14
REDEDGE_EXG_CC_2021_July06
Chromosome
Expected

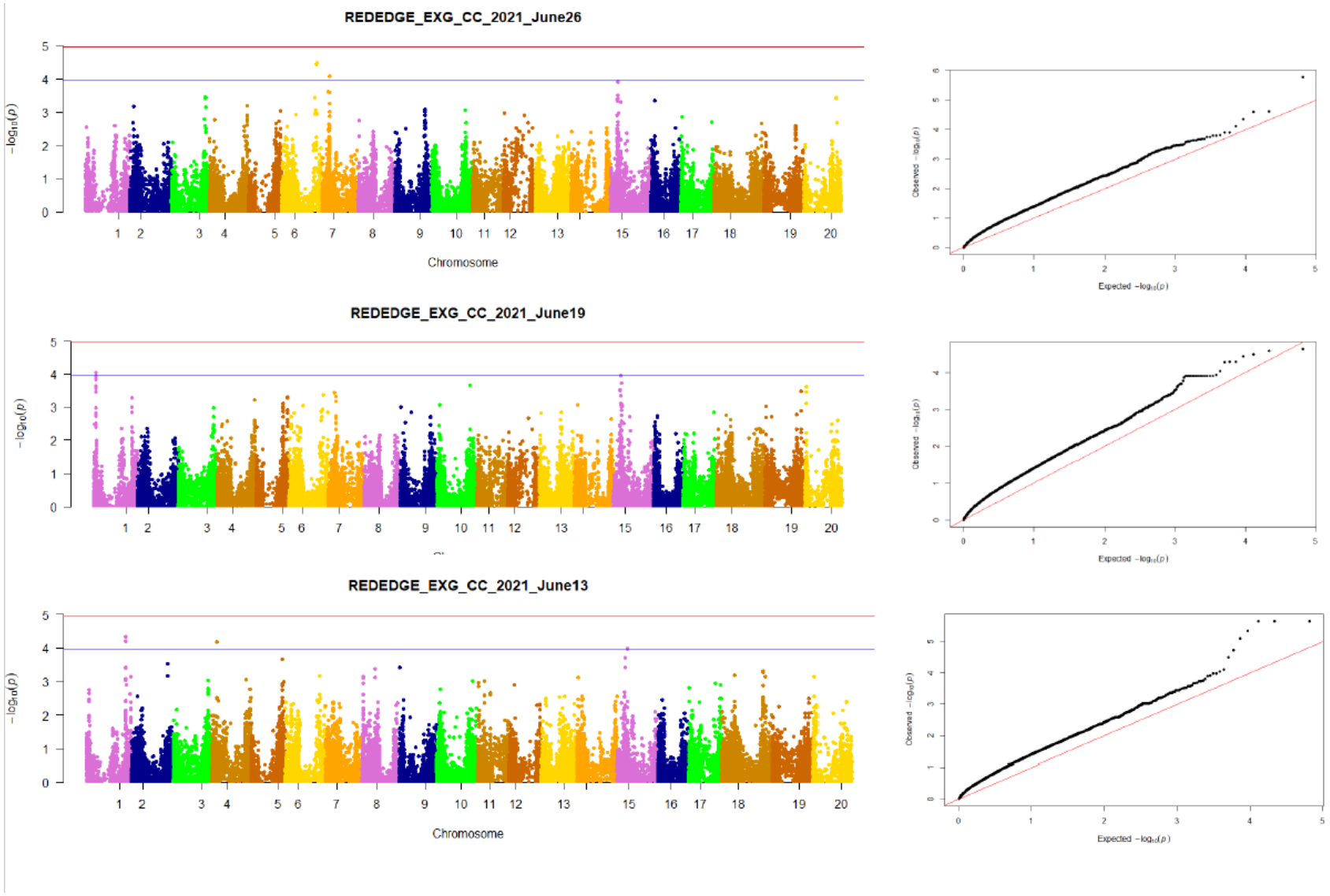


Figure 2.2.10 Manhattan plots and QQ-plots of CC obtained from ExG Index from MS Images in 2021.

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as $-\log_{10}(P) > 4.956$ as represented by red line and suggestive threshold as $-\log_{10}(P) > 3.967$ as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.

Chapter II Section 3: Pod Location and Branching Pattern Study in Edamame to Improve Harvest Efficiency

ABSTRACT

Improving pod location and branching pattern traits of edamame plant is important for breeding of high-yield varieties with increased harvest efficiency. In this study, we use digital imaging technology and computer vision algorithms to characterize major pod location and branching pattern traits automatically and efficiently for edamame. Using a diverse population of edamame PIs, we performed Genome Wide Association Studies (GWAS) to identify the genetic control of different pod location and branching pattern traits in edamame. We found significant variations in the pod locations and branching patterns of the edamame lines and this information can be used to further develop edamame varieties that have better pod location and branching pattern for better yield and increased harvest efficiency. This research provides a pipeline to quantify pod location and branching pattern using computer vision in edamame and the genetic regulation of pod location and branching pattern traits. The results from this research can be used to further develop edamame varieties that are better adapted to better mechanical harvesting.

INTRODUCTION

The immature fruit or the seed stage of edamame also known as vegetable soybean or green soybean is popularly consumed as a snack food. Edamame is constituted with protein, isoflavones, and vitamins. Hand harvesting is a common method of harvesting edamame among small growers, which increases the cost of edamame production (Lord et al., 2019; Mebrahtu and Mullins 2007). A study was performed to test commercial harvesters on edamame and it was found that the harvest efficiency of the harvester tested on three edamame varieties was in between 54 and 85% (Zandonadi, Coolong, and Pfeiffer 2010). There is a need to improve mechanical harvesting

efficiency for edamame. To address this need, the pod location and branching pattern traits have to be studied. It is labor intensive and challenging to physically measure pod location and branching pattern traits in edamame because the shoot structure is covered by leaves when the pods are harvested.

Plant phenotyping is a collection of methods that can extract useful information from data collected by imaging devices or sensors from plants. Plant phenotyping can be a potential solution to solve the challenge of measuring pod location and branching pattern traits in an automatic and consistent manner. High throughput phenotyping has been found to be used by Chen and Nelson (2004) in study leaf shape , by Xavier et al. (2017) in canopy cover and by Fenta et al. (2014) in root architecture. A topological approach called persistent homology (Li et al. 2019) was used to quantify the shoot architecture in topological space and correlate the topological traits with geometric traits of edamame shoots (Dhakal et al. 2021). Falk (2020) used high throughput phenotyping platform in soybean roots to quantify root traits and rank genotypes in a common environment. Trait measurements was carried out for a large plant population of sorghum plants that were grown inside in the greenhouse (Miao et al. 2020). Canopy-scale phenotyping of barley and wheat was carried out using PhenoTrac 4, a mobile platform for phenotyping under field conditions (Barmeier and Schmidhalter 2017). A research was carried out in Australia using, the Phenomobile system, a mobile platform to examine agronomically important traits, such as stay-green in wheat (Rebetzke et al. 2016).

Image-based plant phenotyping and computer vision methods can be utilized to obtain the quantitative data from large number of plant images. For example, the state-of-the-art computer vision systems based on deep convolutional neural networks can deal with various environments to robustly recognize complex objects in both indoor and outdoor environments (Bargoti and

Underwood 2017; Sa et al. 2016). Instance segmentation can detect several instances of objects in an image. To automatically detect the pod locations using deep learning such as Mask-R-CNN and YOLO algorithms (Redmon and Farhadi 2018) can be tested using the data collected using high throughput phenotyping methods.

In this study, we combined the computer vision and the diversity of edamame populations to dissect the regulation of pod location and branching pattern by collecting pod location and branching pattern data from images from diverse edamame PIs. Pod location and branching pattern traits quantified in an automatic and consistent manner from a diverse population can increase the statistical power of GWASs. In addition, the pod location and branching pattern revealed by the data in this study, have potential uses in breeding programs to develop improved edamame cultivars for better machine harvesting. Our results provide a pipeline of pod location and branching pattern quantification using computer vision and provide novel candidate markers and genes for improving harvest efficiency in edamame. From this study try to highlight the usefulness of computer vision and machine learning for removing bottlenecks in plant breeding studies.

METHODS

Plant materials and image collection platform.

A total of 174 (in 2020) and 150 (in 2021) soybean PIs with > 20g/100 seeds that are potential parental lines for developing edamame varieties (referred as edamame PIs) were sown in 3 meters row and 0.75 m row spacing (with a seeding rate of ~70,000 plants per hectare) arranged in a complete randomized design with two to four replications in Kentland farm at Blacksburg, VA in 2020 and 2021. We selected these 174 and 150 PIs in our collection and two to four replicates (plants) per PI were harvested by cutting them from the soil line using a bypass looper (large secateurs). The leaves and petioles were taken off the plants before being taken to the

imaging station. The imaging station consisted of a black background, two 9.1m rulers at the borders, a camera tripod, and a digital single-lens reflex (DSLR) camera. The entry names and sample numbers of the plants were printed as a barcode on an iPad and captured by the camera. Images were captured from one side of the plants. We have generated 324 images from 174 edamame PIs in 2020 and 294 images from 150 edamame PIs in 2020 for this study(available in Ag Data Commons website at https://data.nal.usda.gov/user/login?destination=node/579359). Only 150 PIs were obtained in 2021 out of 174 in 2020 because the other remaining 24 PIs were not planted either because enough seeds were not available, or they had inferior traits like low germination percentage. All 174 varieties have been genotyped using 50K SNP array so genotypic information was obtained from SNP marker data (www.soybase.org).

Phenotypic data collection

Since pod location and branching pattern traits were collected in 2020 and 2021, we calculated the first branch height, average internode length, first pod height, plant height, plant length, and number of branches in 2020 and the plant length, number of branches, number of pods, and first pod height in 2021. We used a web-based annotation tool called zooniverse (https://www.zooniverse.org/) to label the various parts of the plants in plant images. We also labeled the rulers and barcode printed on an iPad. The plant parts that were labeled included the bottom of the plant, top of the plant, the points from where branches emerge from the main stem, the position of the first pod from the bottom of the plant. These labels are as shown in figure 2.3.A.1 (Appendix: https://docs.google.com/spreadsheets/d/1nRonmFFMBP8ky-_t8ErCkO_D6pcHFoHG/edit#gid=1387736013). The length of trait was used as a reference for the other distance related traits extracted from these images.

To measure the first branch height, the distance between the bottom of the plant and the position of the first branch (the one from bottom of the plant) on the main stem was measured. To calculate the average internode length, the internode lengths (the distance between two nodes) were averaged by the number of nodes/branches. The first pod height was calculated by measuring the distance between the bottom of the plant and the position of the first pod from the bottom of the plant. Plant height was measured by measuring the distance between the top and bottom of the plant. Because all the plants were not straight, we marked several points in the plant main stem and add the distance between all those points, which we termed as the plant length. To get the number of branches, the points from where branches emerge from the main stem were counted. Similar techniques were followed in 2021 to measure the pod location and branching pattern traits.

A customized python script for extracting these all traits from the output (csv format) obtained from Zooniverse was written and the analysis was carried out in Jupyter notebook which are present in the GitHub repository (https://github.com/kshitiz52/Pod-location-and-branching-pattern-analysis-in-edamame). The collected data are available in USDA's AgDataCommons website (https://data.nal.usda.gov/dataset/edamame-images-shoot-architecture-traits)

Mask-R-CNN for pod detection in edamame plant images

For Mask-R-CNN, we used a dataset containing 132 RGB images (each image of size 6000*4000). The dataset was split into a training set containing 120 images (6655 pods), a test set composed of 12 images (551 pods). The objects (pods) in the images were labeled using VGG Image Annotator (VIA). The annotation files were saved in json format. The different instances of pods were detected using publicly available Keras/TensorFlow-based implementation for Mask RCNN by Matterport, Inc. (Abdulla, 2017), by pre-training with the COCO dataset (Lin et al., 2014). Matterport's implementation of Mask R-CNN for patch-based processing was personalized

for counting the number of pods in our plant images. The model was evaluated on 274 images and for each image, the model determined the number of pods.

RESULTS

Phenotypic traits

Plant length and height measured in 2020 had mean values of 103.84 cm and 97.3 cm respectively with some outliers, and the plants in 2021 had mean length of 98.55 cm in 2021 as shown in Figure 2.3.A.2 (Appendix). Similarly, the first branch heights in 2020 and 2021 were similar with an average of 13.03 cm and 13.93 cm, respectively. Also, the average first pod height in 2020 was 13.92 cm and that of average internode length was 6.91 cm as shown in Figure 2.3.A.3 (Appendix). And the average number of branches in 2020 was 5.95 cm and that in 2021 was 5.44 cm as shown in Figure 2.3.A.4 (Appendix). The average number of pods in 2021 predicted using Mask-R-CNN was 39.55 as shown in Figure 2.3.A.4 (Appendix). The detailed data for all these traits in both the years are shown in Table 2.3.A.1 (Appendix).

Number of pods using Mask-R-CNN results

The mean average precision(mAP) for pod detection in the plant images was 0.43. From the ROIs obtained from Mask-R-CNN result, we used to count the number of pods in the prediction dataset as shown in figure 2.3.S.1, 2.3.S.2, and 2.3.S.3.

Population Structure

We retained 150 principal components that accounted for 80% of cumulative variance through DAPC and with the smallest BIC, and the optimal number of clusters, k= 6 was determined (Figure 2.3.3). All six clusters were well separated from each other.

2020 Results

GWAS on pod location and branching pattern traits from 2020

Two SNPs (ss715586704, and ss715606829) displayed significant associations. The remaining 35 SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold. Chromosomes (Chr) 14 contained the eight suggestive associations, followed by Chr 4 with seven suggestive associations, Chr 7, 10, 13, and 18 with four suggestive associations, Chr 2 with three suggestive associations, Chr 6 with two suggestive associations, and Chr 3, 9, and 17 with one suggestive association.

Candidate Genes

Two candidate genes (Glyma03g04710, and Glyma10g30530) were associated with number of branches (Brcnt20), where the –log10(P) of SNPs were over 4.96. The remaining Sixteen candidate genes (Glyma04g40710, Glyma04g40720, Glyma04g40730, Glyma04g40740, Glyma04g40750, Glyma04g40800, Glyma13g30530, Glyma13g15590, Glyma14g07310, Glyma14g07320, Glyma14g07340, Glyma14g07360, Glyma14g07380, Glyma18g01280, Glyma18g01280, and Glyma18g01290) were associated with number of branches (Brcnt20), where the –log10(P) of SNPs were over 3.967. There were nine candidate genes (Glyma02g11480, Glyma02g11490, Glyma02g114470, Glyma07g07830, Glyma13g26760, Glyma18g11051, Glyma07g03320, Glyma07g03350, and Glyma09g06740) that were associated with first pod height (Fph20), where the –log10(P) of SNPs were over 3.967. There were seven candidate genes Glyma06g08340, Glyma06g08345, Glyma07g36530, Glyma10g31071, Glyma10g31225, Glyma17g17100, and Glyma18g45895) that were associated with plant height (Ph20), where the –log10(P) of SNPs were over 3.967. There were two candidate genes (Glyma14g11505, and Glyma13g07781) that were associated with average internode length (Ail20), where the –log10(P)

of SNPs were over 3.967. There was one candidate gene (Glyma10g06500) that was associated with plant length (Pl20), where the –log10(P) of SNPs were over 3.967.

2021 Results

GWAS on pod location and branching pattern traits from 2021

Only one SNPs (ss715583821) displayed significant association in 2021, which was associated with the number of pods. The remaining twenty-eight (28) SNPs had $-\log_{10}(P) > 3.967$, which were above the suggestive threshold for all the traits measured in 2021. Chromosome (Chr)2 contained the eleven suggestive associations, followed by Chr 6 with four suggestive associations, Chr 4, 8, and 15 with three suggestive associations, Chr 5 with two suggestive associations, and Chr 9, 10, and 20 with one suggestive association.

Candidate Genes

One candidate gene (Glyma02g00940) was associated with number of pods (Podcnt21), where the –log10(P) of SNPs were over 4.96. The remaining ten candidate genes (Glyma02g00780, Glyma02g00790, Glyma02g00471, Glyma05g12770, Glyma02g00850, Glyma02g00860, Glyma02g00730, Glyma02g00760, Glyma08g03710, and Glyma02g00500) were associated with number of pods (Podcnt21), where the –log10(P) of SNPs were over 3.967. Eleven candidate genes (Glyma04g17310, Glyma04g36200, Glyma04g39455, Glyma06g36730, Glyma06g36840, Glyma06g38076, Glyma06g38470, Glyma10g02010, Glyma15g15400, Glyma15g15410, and Glyma15g15431) were associated with number of branches (Brcnt21), where the –log10(P) of SNPs were over 3.967. There were five candidate genes (Glyma05g07580, Glyma08g19880, Glyma08g08410, Glyma09g39700, and Glyma20g24820) that were associated with plant length (Pl21), where the –log10(P) of SNPs were over 3.967. There were none of the candidate genes for first pod height, where the –log10(P) of associated SNPs were over 3.967.

DISCUSSION

Harvest loss during the mechanical harvesting is one of the major reasons that the growers tend to use traditional hand harvesting of edamame pods. There is a need to improve mechanical harvesting efficiency in edamame by changing the shoot architecture. It is labor intensive and challenging to physically measure the traits in edamame because the shoot structure is covered by leaves when the pods are harvested. High throughput phenotyping combined with computer vision make the trait quantification in plants more effective as compared to manual trait extraction.

Combining high throughput phenotyping, computer vision, and high-density marker set from the SoySNP50K repository via GWASs can effectively and more accurately find the genetic control of several traits (Lee et al., 2019; Jarquin et al., 2016). In this study, we identified novel associations for branching pattern and pod location traits in edamame using the diverse edamame accessions via GWASs.

We found 35 candidate genes and 25 SNPs for edamame pod location and branching pattern traits in 2020. We found 27 candidate genes and 29 SNPs for edamame pod location and branching pattern traits in 2021. No common SNPs and genes were found between the traits of 2020 and 2021. This could be due to small additive effects, where the SNPs could not cross the threshold to be regarded as significant SNPs.

CONCLUSION

In conclusion, we performed phenotypic analysis of a collection of edamame varieties and identified several putative genes for pod location and branching pattern traits in these varieties. Using these new genetic markers, the targeted traits can be potentially improved. For example, marker-assisted selection (MAS) can be carried out for better harvest efficiency in edamame breeding.

REFERENCES

Bargoti, Suchet, and James Underwood. 2017. "Deep Fruit Detection in Orchards." *Proceedings - IEEE International Conference on Robotics and Automation* 3626–33. doi: 10.1109/ICRA.2017.7989417.

Barmeier, Gero, and Urs Schmidhalter. 2017. "High-Throughput Field Phenotyping of Leaves, Leaf Sheaths, Culms and Ears of Spring Barley Cultivars at Anthesis and Dough Ripeness." *Frontiers in Plant Science* 8:1920.

Chen, Yiwu, and Randall L. Nelson. n.d. *Evaluation and Classification of Leaflet Shape and Size in Wild Soybean*.

Dhakal, K., Q. Zhu, B. Zhang, M. Li, and S. Li. 2021. "Analysis of Shoot Architecture Traits in Edamame Reveals Potential Strategies to Improve Harvest Efficiency." *Frontiers in Plant Science* 12. doi: 10.3389/fpls.2021.614926.

Falk, Kevin G., Talukder Z. Jubery, Seyed V Mirnezami, Kyle A. Parmley, Soumik Sarkar, Arti Singh, Baskar Ganapathysubramanian, and Asheesh K. Singh. 2020. "Computer Vision and Machine Learning Enabled Soybean Root Phenotyping Pipeline." *Plant Methods* 16(1):1–19.

Fang, Chao, Yanming Ma, Shiwen Wu, Zhi Liu, Zheng Wang, Rui Yang, Guanghui Hu, Zhengkui Zhou, Hong Yu, Min Zhang, Yi Pan, Guoan Zhou, Haixiang Ren, Weiguang Du, Hongrui Yan, Yanping Wang, Dezhi Han, Yanting Shen, Shulin Liu, Tengfei Liu, Jixiang Zhang, Hao Qin, Jia Yuan, Xiaohui Yuan, Fanjiang Kong, Baohui Liu, Jiayang Li, Zhiwu Zhang, Guodong Wang, Baoge Zhu, and Zhixi Tian. 2017. "Genome-Wide Association Studies Dissect the Genetic Networks Underlying Agronomical Traits in Soybean."

Genome Biology 18(1). doi: 10.1186/s13059-017-1289-9.

Fenta, Berhanu A., Stephen E. Beebe, Karl J. Kunert, James D. Burridge, Kathryn M. Barlow, Jonathan P. Lynch, and Christine H. Foyer. 2014. "Field Phenotyping of Soybean Roots for Drought Stress Tolerance." *Agronomy* 4(3):418–35. doi: 10.3390/agronomy4030418.

Jarquin, Diego, James Specht, and Aaron Lorenz. 2016. "Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions." *G3: Genes, Genomes, Genetics* 6(8):2329–41. doi: 10.1534/g3.116.031443.

Lee, Sungwoo, Kyujung Van, Mikyung Sung, Randall Nelson, Jonathan LaMantia, Leah K. McHale, and M. A. Rou. Mian. 2019. "Genome-Wide Association Study of Seed Protein, Oil and Amino Acid Contents in Soybean from Maturity Groups I to IV." *Theoretical and Applied Genetics* 132(6):1639–59. doi: 10.1007/s00122-019-03304-5.

Li, Mao, Laura L. Klein, Keith E. Duncan, Ni Jiang, Daniel H. Chitwood, Jason P. Londo, Allison J. Miller, and Christopher N. Topp. 2019. "Characterizing 3D Inflorescence Architecture in Grapevine Using X-Ray Imaging and Advanced Morphometrics: Implications for Understanding Cluster Density." *Journal of Experimental Botany* 70(21):6261–76. doi: 10.1093/jxb/erz394.

Liu, Hai-Jun, and Jianbing Yan. 2019. "Crop Genome-Wide Association Study: A Harvest of Biological Relevance." *The Plant Journal* 97(1):8–18.

Lord, Nick. n.d. *Administrator, 1890 Extension Program*.

Mebrahtu, Tadesse, and Chris Mullins. 2007. *Efficiency of Mechanical Harvest for Immature Vegetable Soybean Pods 1*. Vol. 58.

Miao, Chenyong, Zheng Xu, Eric Rodene, Jinliang Yang, James C. Schnable, and others. 2020. "Semantic Segmentation of Sorghum Using Hyperspectral Data Identifies Genetic Associations." *Plant Phenomics* 2020.

Rebetzke, Greg J., Jose A. Jimenez-Berni, William D. Bovill, David M. Deery, and Richard A. James. 2016. "High-Throughput Phenotyping Technologies Allow Accurate Selection of Stay-Green." *Journal of Experimental Botany* 67(17):4919–24.

Redmon, Joseph, and Ali Farhadi. 2018. "YOLOv3: An Incremental Improvement."

Sa, Inkyu, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. 2016. "Deepfruits: A Fruit Detection System Using Deep Neural Networks." *Sensors (Switzerland)* 16(8). doi: 10.3390/s16081222.

Song, Qijian, David L. Hyten, Gaofeng Jia, Charles V. Quigley, Edward W. Fickus, Randall L. Nelson, and Perry B. Cregan. 2013. "Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean." *PLoS ONE* 8(1). doi: 10.1371/journal.pone.0054985.

Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. "Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine Max." *Genetics* 206(2):1081–89. doi: 10.1534/genetics.116.198713.

Zandonadi, Rodrigo, T. Coolong, and T. Pfeiffer. 2010. "Mechanical Harvesting of Edamame." *SARE Proj Final Rep. Available Online at: Https://Www. Uky. Edu/Ccd/Sites/Www. Uky. Edu. Ccd/Files/Edamame_mechanical_harvest. Pdf (Accessed February 7, 2021)*.

Tables and Figures

Table 2.3.1 Candidate gene and descriptions of the significantly associated SNPs for SA in 2020 using Wm82.a2.v1.

SNP (BP)	**Corresponding Gene ID**	**Location**	**P**	**Traits**	**Gene Function Description**
ss715586704 (4903317)	Glyma03g04710	Gm03:4905288: 4908506	4.96	Brcnt20	Peroxidase activity
ss715606829 (39152386)	Glyma10g30530	Gm10:39150279: 39152709	4.95	Brcnt20	Prenylated RAB Acceptor-1 related
ss715617321 (10063083)	Glyma14g11505	Gm14:10067048: 10068453	4.82	Ail20	Cyclic nucleotide gated channel
ss715619685 (5500948)	Glyma14g07320	Gm14:5486471: 5487638	4.54	Brcnt20	DVL family/ Rotundifolia like 8
ss715619689 (5530899)	Glyma14g07340	Gm14:5564711: 5569297	4.54	Brcnt20	Actin and related protein 5
ss715608238 (5208391)	Glyma10g06500	Gm10:5207574: 5213118	4.53	Pl20	Sigma-70 region 2/DNA binding/RNA polymerase sigma-subunit F
ss715619693 (5603904)	Glyma14g07380	Gm14:5602042: 5610923	4.51	Brcnt20	Adapter-related protein complex 3, beta subunit/ Vesicle coat complex AP-3/Adaptin N terminal region/Intracellular protein transport/Protein affected trafficking 2
ss715619735 (6102984)	Glyma14g08070	Gm14:6101226: 6105979	4.49	Brcnt20	Predicted transmembrane transporter/Sugar transporter
ss715584253 (9666401)	Glyma02g11480	Gm02:9665139: 9669328	4.43	Fbh20	WDSAM1 Protein/ Plant U-box 26 domain/ Ubiquitin-protein transferase activity
ss715605156 (5480440)	Glyma09g06740	Gm09:5478944: 5481327	4.38	Fph20	Chloroplast import apparatus 2

ss715619688 (5519891)	Glyma14g07310	Gm14:5472292: 5474060	4.36	Brcnt20	Aspartyl Proteases/Aspartic-type endopeptidase activity
ss715588770 (46700121)	Glyma04g40750	Gm04:46699395: 46705586	4.35	Brcnt20	Splicing factor 3b, subunit 4/ Ataxin-2 C-terminal region/ Nucleic acid binding/ CTC-interacting domain 11
ss715588764 (46674438)	Glyma04g40720	Gm04:46675344: 46677019	4.33	Brcnt20	60s Acidic ribosomal protein
ss715596844 (2341725)	Glyma07g03320	Gm07:2337398: 2341853	4.32	Fph20	Protein binding/ Ubiquitin system component Cue protein
ss715588764 (46690093)	Glyma04g40730	Gm04:46680911: 46690250	4.31	Brcnt20	GDP-fucose protein O-fucosyl-transferase family protein
ss715631679 (55585929)	Glyma18g45895	Gm18:55582645: 55584957	4.27	Ph20	Common central domain of Tyrosinase/ Oxidoreductase activity
ss715584258 (9689986)	Glyma02g11490	Gm02:9687274: 9689620	4.26	Fbh20	N/A
ss715614002 (19554349)	Glyma13g15590	Gm13:19468523: 19477472	4.24	Brcnt20	Apoptotic ATPase/ Leucine Rich Repeat Protein binding/ Disease resistance protein (TIR-NBS-LRR class)
ss715588766 (46697167)	Glyma04g40740	Gm04:46690890: 46696289	4.23	Brcnt20	Splicing factor, Arginine/Serine-rich protein 34A/ Alternative splicing factor ASF/SF2 / RNA recognition motif/ Nucleic acid binding
ss715617215 (7984096)	Glyma13g07781	Gm13:7983344: 7994553	4.23	Ail20	Predicted transmembrane transporter / Sugar transporter/Plastidic GLC translocator
ss715606929 (39782183)	Glyma10g31225	Gm10:39784601: 39786275	4.21	Ph20	Member of the RIN4-like/NOI family; upregulated after spider mite feeding
ss715598063 (41892346)	Glyma07g36530	Gm07:41893233: 41895665	4.21	Ph20	Desiccation-like protein
ss715588765 (46677734)	Glyma04g40720	Gm04:46675344: 46677019	4.20	Brcnt20	60s Acidic ribosomal protein
ss715632553 (671304)	Glyma18g01290	Gm18:671566: 677448	4.18	Brcnt20	DNA polymerase alpha-primase complex/polymerase-associated subunit B

ss715632553 (671304)	Glyma18g01280	Gm18:667004: 671294	4.18	Brcnt20	Mitochondrial/plastidial beta-ketoacyl-ACP reductase/ Short chain dehydrogenase/ Oxidoreductase activity/ NAD(P)-binding Rossmann-fold superfamily protein
ss715581168 (13056993)	Glyma02g14470	Gm02:13050653: 13055733	4.17	Fbh20	ATP binding Cassette Transporter/ ABC-2 type transporter family protein
ss715615098 (29961821)	Glyma13g26760	Gm13:29969387: 29977878	4.17	Fbh20	H+/oligopeptide transporter-related activity/ Major facilitator superfamily protein
ss715626129 (13932633)	Glyma17g17100	Gm17:13926316: 13926822	4.15	Ph20	bZIP transcription factor/ Sequence-specific DNA binding transcription factor activity/ Basic leucine-zipper 5
ss715595392 (6094090)	Glyma06g08345	Gm06:6095037: 6095758	4.11	Ph20	N/A
ss715595392 (6094090)	Glyma06g08340	Gm06:6089833: 6090919	4.11	Ph20	Auxin responsive protein/ SAUR-like auxin-responsive protein family
ss715596849 (2361283)	Glyma07g03350	Gm07:2360805: 2362862	4.10	Fph20	Phosphatidic acid phosphatase-related / PAP2-related
ss715632988 (9890070)	Glyma18g11051	Gm18:9897112: 9899291	4.09	Fbh20	RING finger Protein 11/Zinc finger, C3HC4 type (RING finger)/ RING/U-box superfamily protein
ss715619692 (5587345)	Glyma14g07360	Gm14:5586404: 5589414	4.06	Brcnt20	Rare lipoprotein A (RlpA)-like double-psi beta-barrel
ss715619694 (5606320)	Glyma14g07380	Gm14:5602042: 5610923	4.06	Brcnt20	Adapter-related protein complex 3, beta subunit/Vesicle coat complex AP-3/Adaptin N terminal region/Intracellular protein transport/ Protein affected trafficking 2
ss715588779 (46750415)	Glyma04g40800	Gm04:46749495: 4675174	4.01	Brcnt20	Serine-Threonine Protein Kinase (Plant-type)
ss715606902 (39630441)	Glyma10g31071	Gm10:39629998: 39632630	4.00	Ph20	Low protein: ammonium transporter 1-like protein

ss715598496 (6491906)	Glyma07g07830	Gm07:6489496: 6492623	3.99	Fbh20	Bestrophin-like protein, RFP-TM, chloride channel

Base Pair (BP); Probability (P)
Gene function as described in TAIR, PANTHER, or GO annotation.

Table 2.3.2 Candidate gene and descriptions of the significantly associated SNPs for SA in 2021 using Wm82.a2. v1.

SNP (BP)	Corresponding Gene ID	Location	P	Traits	Gene Function Description
ss715583821 (707483)	Glyma02g00940	Gm02:707755: 708586	5.20	Podcnt21	Protein coding/Hypothetical protein
ss715605620 (1417351)	Glyma10g02010	Gm10:1419210: 1420028	4.48	Brcnt21	Mitochondrial career protein-related
ss715602622 (6031132)	Glyma08g08410	Gm08:6032627: 6034317	4.46	Pl21	Leucine Rich Repeat receptor like-protein Kinase/Protein tyrosine kinase
ss715583655 (573751)	Glyma02g00780	Gm02:574680: 579321	4.36	Podcnt21	Aldo/keto reductase /Voltage-gated shaker-like K+ channel/(NAD(P)-linked oxidoreductase superfamily
ss715583662 (578329)	Glyma02g00780	Gm02:574680: 579321	4.36	Podcnt21	Aldo/keto reductase /Voltage-gated shaker-like K+ channel/(NAD(P)-linked oxidoreductase superfamily
ss715583675 (587325)	Glyma02g00790	Gm02:583937: 591332	4.36	Podcnt21	FYVE zinc finger/metal ion binding/Regulator of chromosome condensation (RCC1) family protein
ss715583679 (588453)	Glyma02g00790	Gm02:583937: 591332	4.36	Podcnt21	FYVE zinc finger/metal ion binding/Regulator of chromosome condensation (RCC1) family protein
ss715581649 (266635)	Glyma02g00471	Gm02:262113: 268667	4.30	Podcnt21	Unknown function
ss715604839 (44732111)	Glyma09g39700	Gm09:44731480: 44733775	4.30	Pl21	KDEL (LYS-ASP-GLU-LEU) containing - related/Glycosyl transferase family 90
ss715637603 (34462359)	Glyma20g24820	Gm20:34463068: 34471228	4.30	Pl21	U4/U6-associated splicing factor PRP4/Protein kinase domain/Non-specific serine/threonine protein kinase/
ss715589872 (12920089)	Glyma05g12770	Gm05:12981571: 12984704	4.28	Podcnt21	Oxidoreductase activity/Iron/ascorbate family
ss715583751 (651317)	Glyma02g00850	Gm02:653251: 659169	4.25	Podcnt21	Serine/threonine specific protein phosphatase PP1, catalytic subunit/hydrolase activity
ss715583761 (662422)	Glyma02g00860	Gm02:662122: 663031	4.25	Podcnt21	Unknown function
ss715594347 (39102348)	Glyma06g36730	Gm06:39104286: 39107048	4.21	Brcnt21	Serine-Threonine Protein Kinase (Plant-type)

ss715594350 (39210184)	Glyma06g36840	Gm06:39220750: 39222745	4.21	Brcnt21	26S proteasome regulatory complex, subunit PSMD10/Ankyrin repeat family protein
ss715594390 (40849730)	Glyma06g38076	Gm06:40854187: 40854915	4.21	Brcnt21	FAD binding domain/UDP-N-acetylmuramate dehydrogenase activity/FAD-binding Berberine family
ss715620359 (11827535)	Glyma15g15400	Gm15:11823043: 11827476	4.19	Brcnt21	Transcription factor: MYB like DNA-binding protein/Myb proto-oncogene protein, plant
ss715620360 (11827535)	Glyma15g15410	Gm15:11828643: 11834079	4.19	Brcnt21	Phosphatidylinositol N-Acetylglucosaminyltransferase subunit-P
ss715620362 (11844860)	Glyma15g15431	Gm15:11851858: 11852191	4.19	Brcnt21	Zinc knuckle/Nucleic acid binding/Zinc ion binding, nucleic acid binding
ss715583581 (520296)	Glyma02g00730	Gm02:520056: 524702	4.09	Podcnt21	Protein coding/Neurogenic locus notch-like protein
ss715583646 (567618)	Glyma02g00760	Gm02:564214: 569593	4.09	Podcnt21	Auxin canalization/Plant protein of unknown function (DUF828) with plant pleckstrin homology-like region
ss715601228 (2636067)	Glyma08g03710	Gm08:2630927: 2633769	4.07	Podcnt21	IQ calmodulin-binding motif/Protein binding
ss715587289 (18662050)	Glyma04g17310	Gm04:18779621: 18787193	4.06	Podcnt21	Cell redox homeostasis/Thioredoxin
ss715588240 (42739828)	Glyma04g36200	Gm04:42731238: 42734117	4.02	Brcnt21	UDP-glucuronosyl and UDP-glucosyl transferase/Transferring hexosyl groups
ss715594409 (41416032)	Glyma06g38470	Gm06:41409188: 41415428	4.02	Brcnt21	Peptidyl-Prolyl Cis-Trans-Isomerase/Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD
ss715581700 (279925)	Glyma02g00500	Gm02:276998: 277312	4.01	Podcnt21	Protein coding/Ribosomal L28 family
ss71559247 (7628516)	Glyma05g07580	Gm05:7624076: 7629009	4.00	Pl21	Serine/Threonine Protein Kinase 3/TGF-beta stimulated factor
ss715588652 (45649110)	Glyma04g39455	Gm04:45648605: 45652897	3.99	Brcnt21	NYN domain/Putative endonuclease or glycosyl hydrolase)
ss715599646 (15013072)	Glyma08g19880	Gm08:15014296: 15015244	3.98	Pl21	Unknown function

Base Pair (BP); Probability (P)
Gene function as described in TAIR, PANTHER, or GO annotation.

Table 2.3.3 Pod location and branching pattern traits measured in 2020 and 2021.

S.N.	Traits	Year	Abbreviation
1	First branch height	2020	Fbh20
2	Average internode length	2020	Ail20
3	First pod height	2020	Fph20
4	Plant height	2020	Ph20
5	Plant length	2020	Pl20
6	Number of branches/Branches count	2020	Brcnt20
7	Plant length	2021	Pl21
8	Number of branches/Branches count	2021	Brcnt21
9	First pod height	2021	Fph21
10	Number of pods/Pods count	2021	Podcnt21

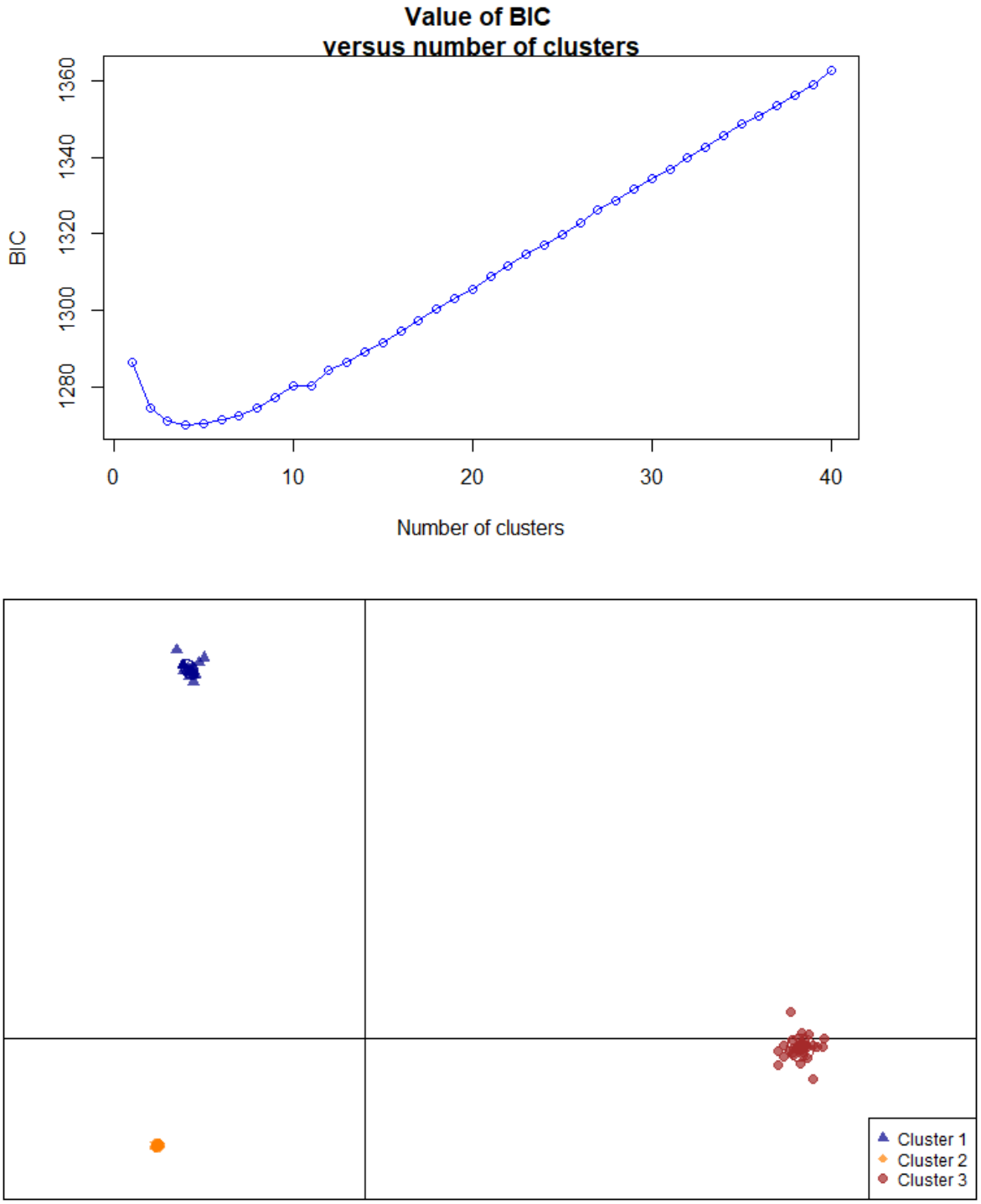
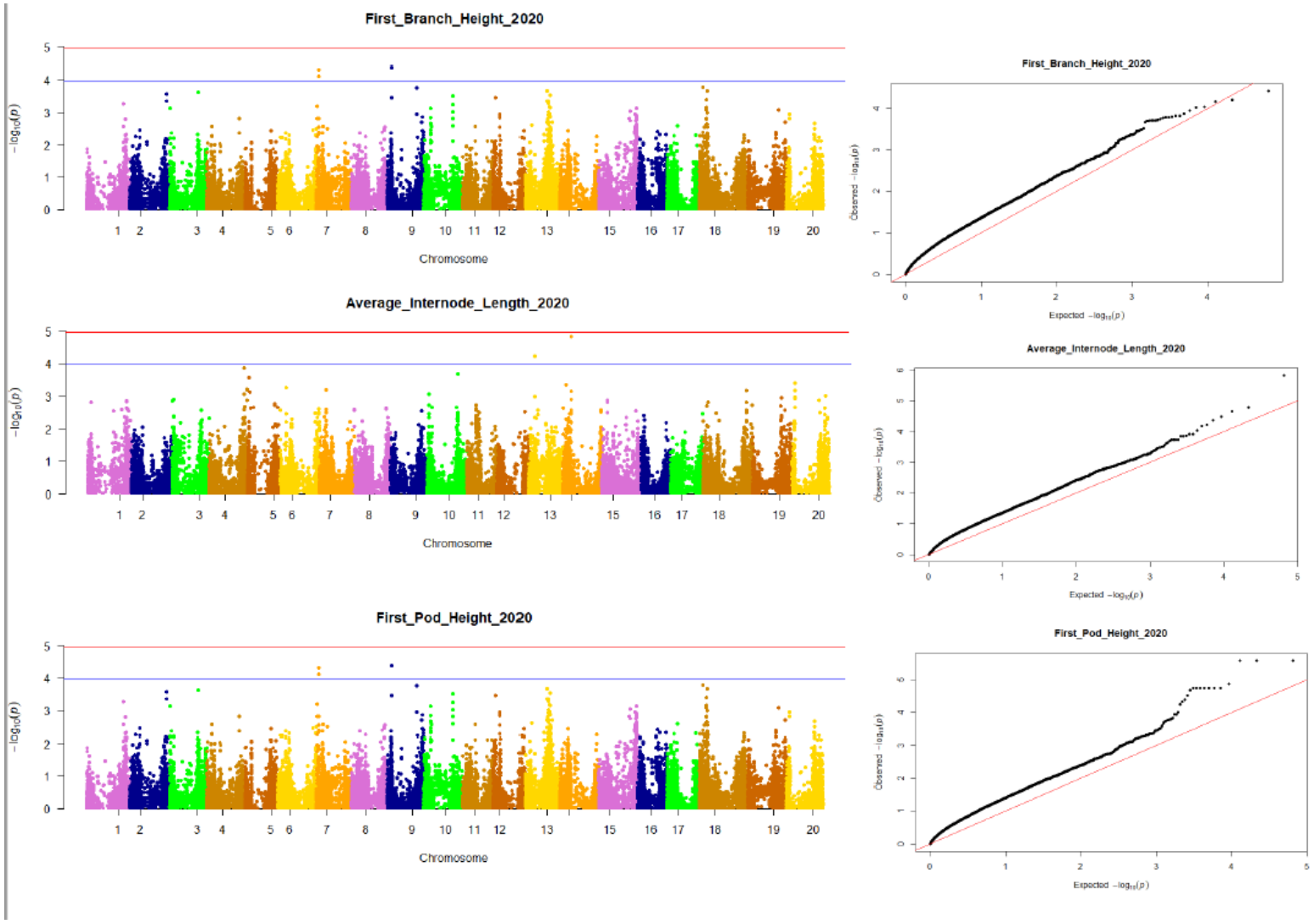
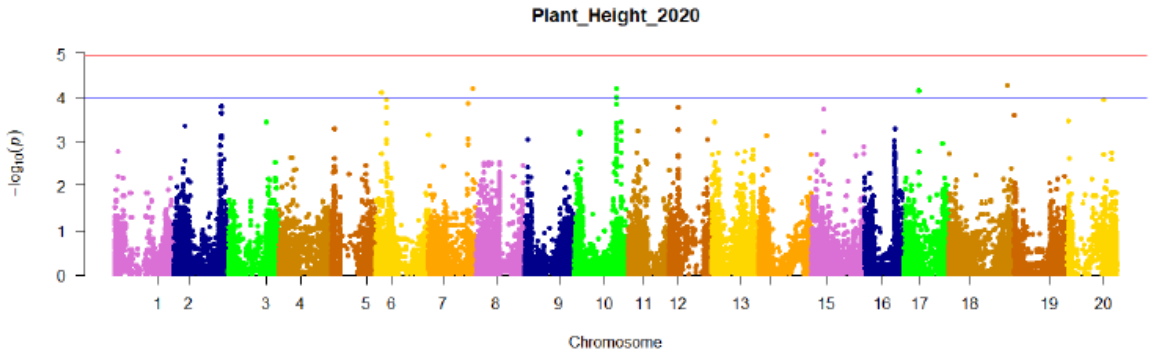
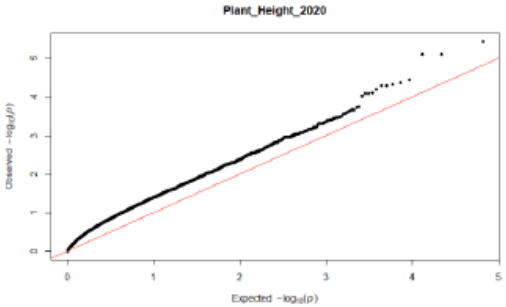
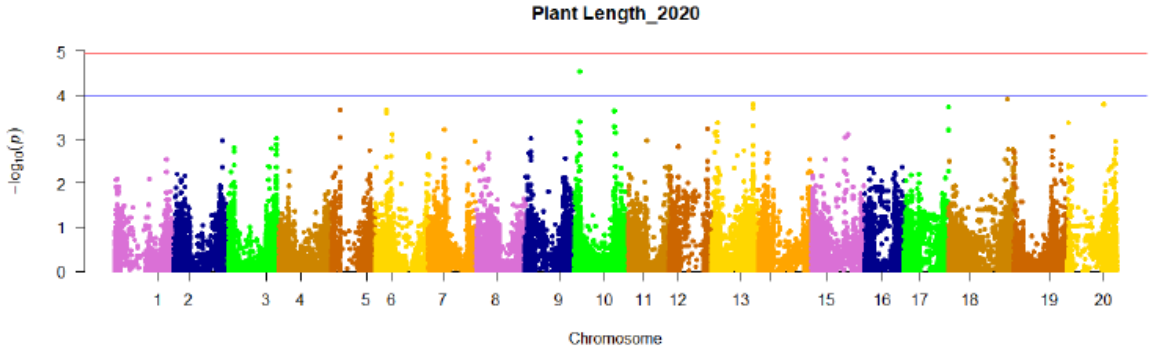


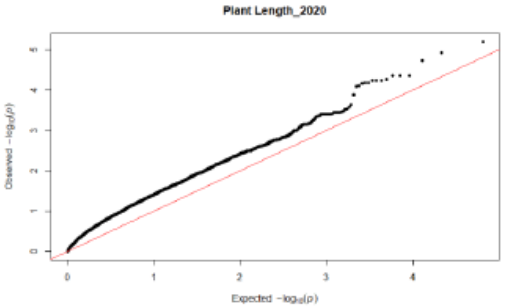
Figure 2.3.1.(a) Bayesian Information Criterion (BIC) used to select the optimal number of clusters. (b) A scatter plot showing the 3 clusters (k=3) identified as likely subpopulations within the 151 accessions.

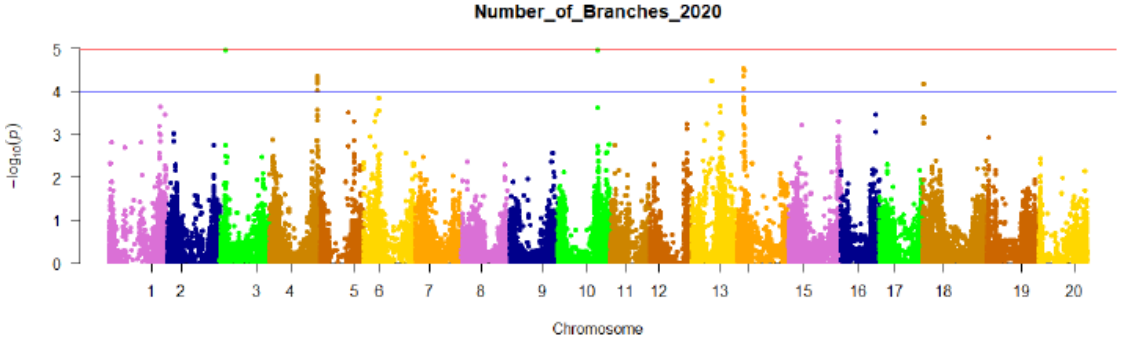

First_Branch_Height_2020
Chromosome
First_Branch_Height_2020
Expected
Average_Internode_Length_2020
Chromosome
Average_Internode_Length_2020
Expected
First_Pod_Height_2020
Chromosome
First_Pod_Height_2020
Expected


Plant_Height_2020
−log10(p)
Chromosome


Plant_Height_2020
Observed −log10(p)
Expected −log10(p)


Plant Length_2020
−log10(p)
Chromosome


Plant Length_2020
Observed −log10(p)
Expected −log10(p)


Number_of_Branches_2020
−log10(p)
Chromosome

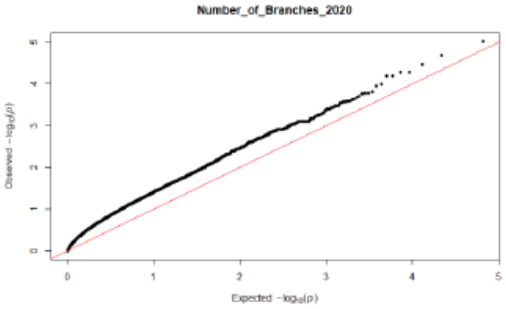

Number_of_Branches_2020
Observed −log10(p)
Expected −log10(p)

Figure 2.3.2 Manhattan plots and QQ-plots of CC obtained from pod location and branching pattern traits data from 2020.

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as –log10(P) > 4.96 as represented by red line and suggestive threshold as –log10(P) > 3.97 as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.

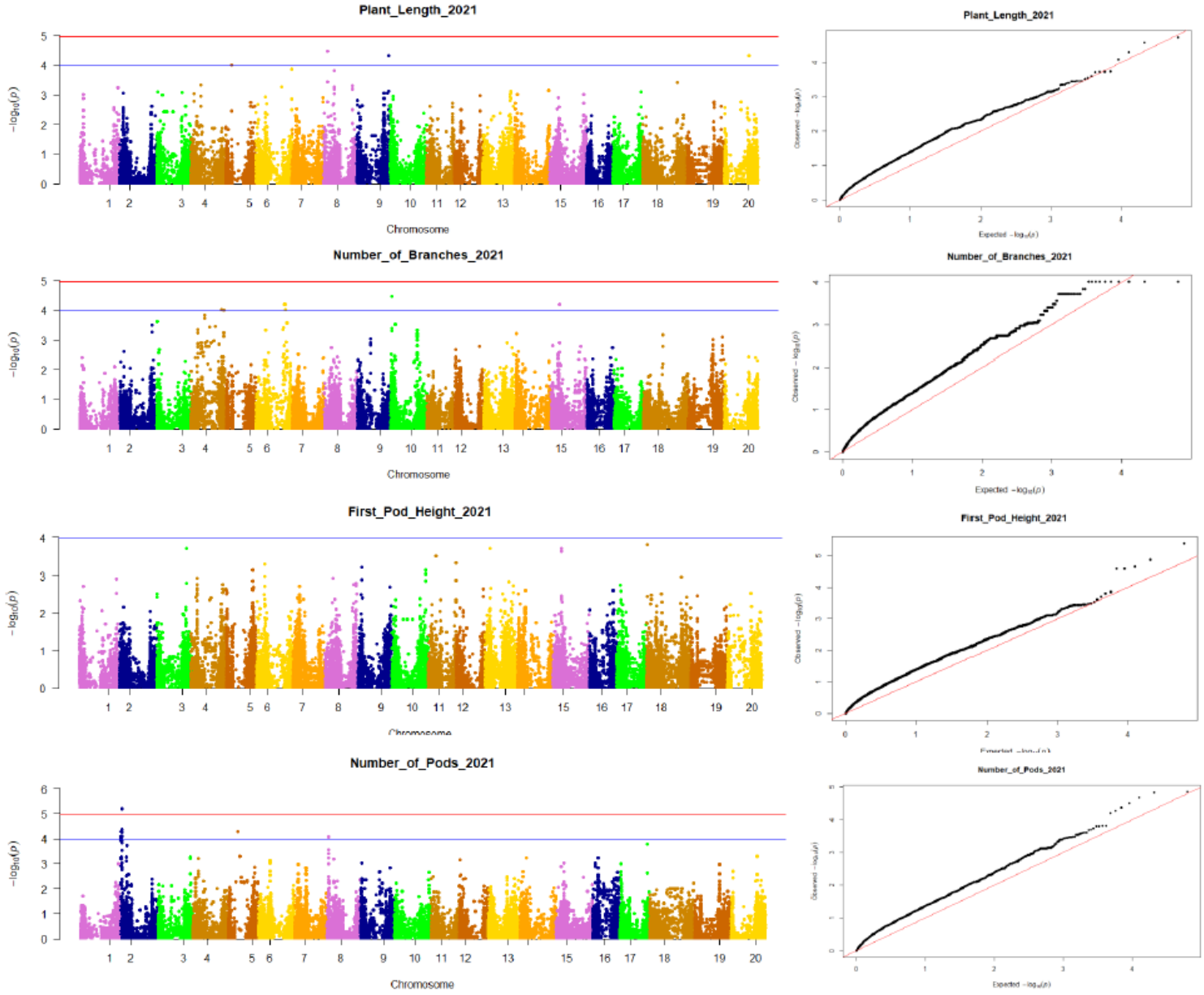

Plant_Length_2021
Plant_Length_2021
Number_of_Branches_2021
Number_of_Branches_2021
First_Pod_Height_2021
First_Pod_Height_2021
Number_of_Pods_2021
Number_of_Pods_2021
Chromosome
Chromosome
Chromosome

Figure 2.3.3 Manhattan plots and QQ-plots of CC obtained from pod location and branching pattern traits data from 2021

The x-axis of each Manhattan plot are the chromosome numbers, whereas the y-axis are the LOD (-log10(p-value)). The manhattan plot are color coded based on chromosome number. The significance threshold was set as $-\log_{10}(P) > 4.956$ as represented by red line and suggestive threshold as $-\log_{10}(P) > 3.967$ as represented by blue line. The x-axis and y-axis of each QQ-plot are the expected and observed -log10(p-value) respectively.

Supplementary Figures

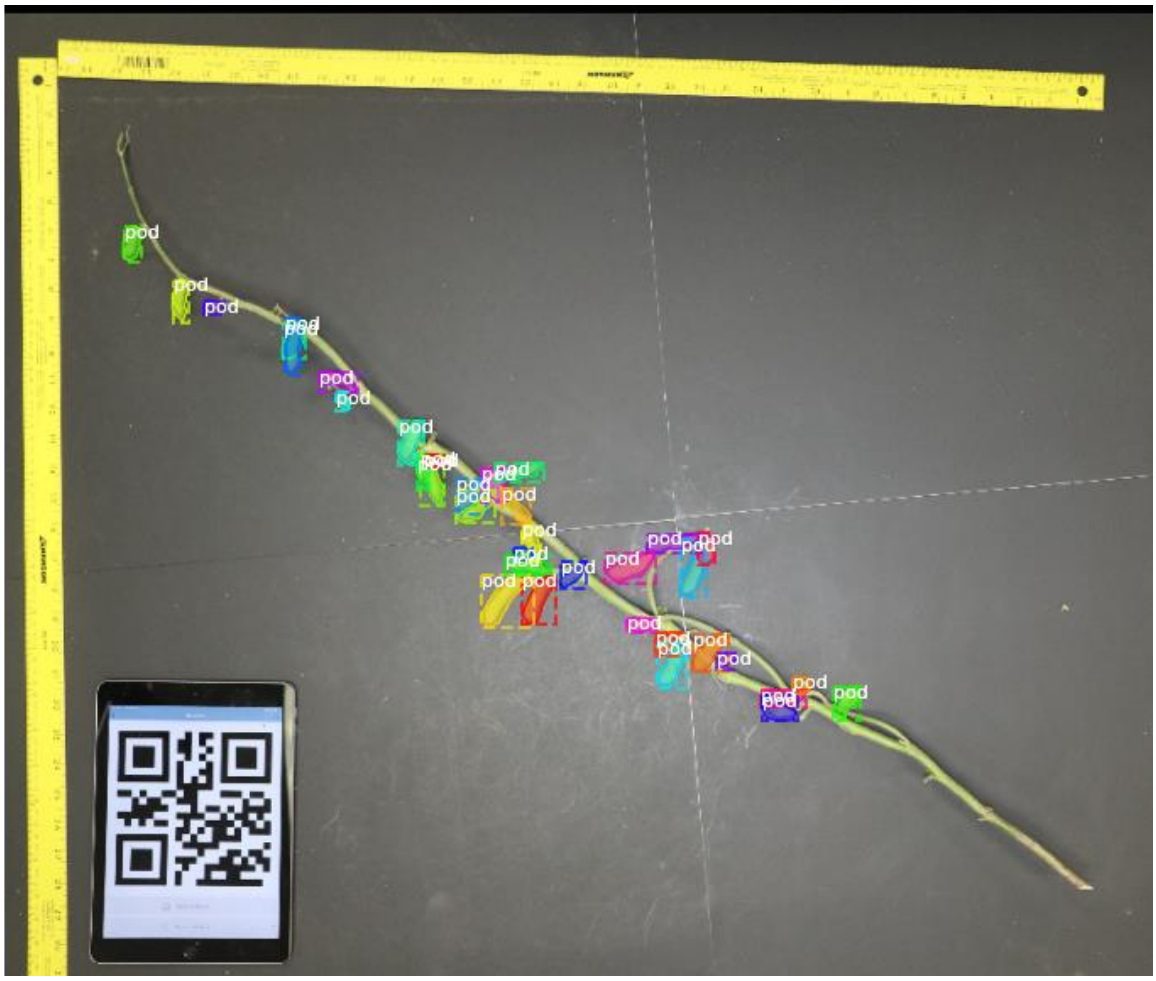


Figure 2.3.S.1 Ground Truth of segmentation on Edamame plant images.

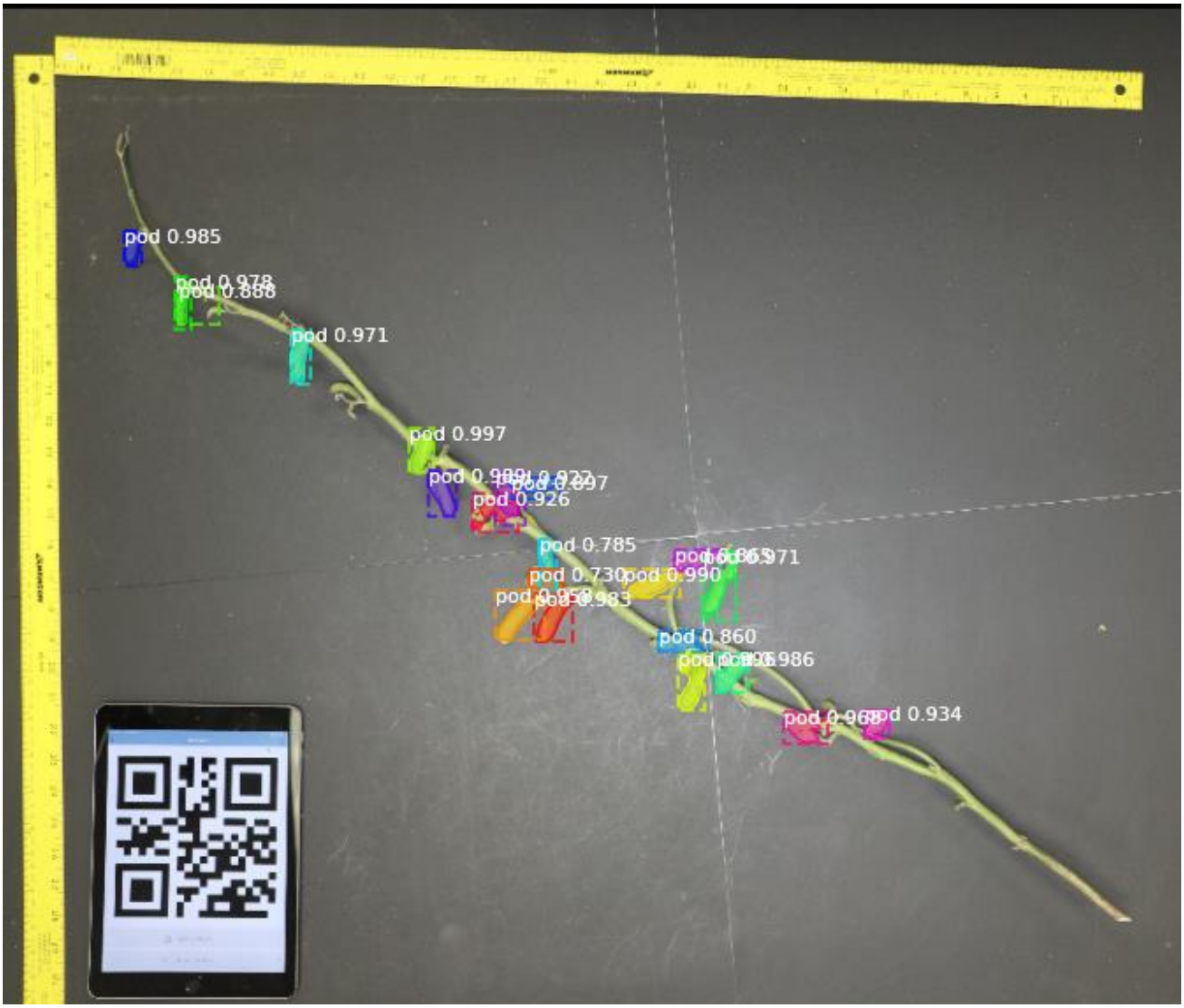


Figure 2.3.S.2 Mask-R-CNN Predicted segmentation of Edamame plant images.

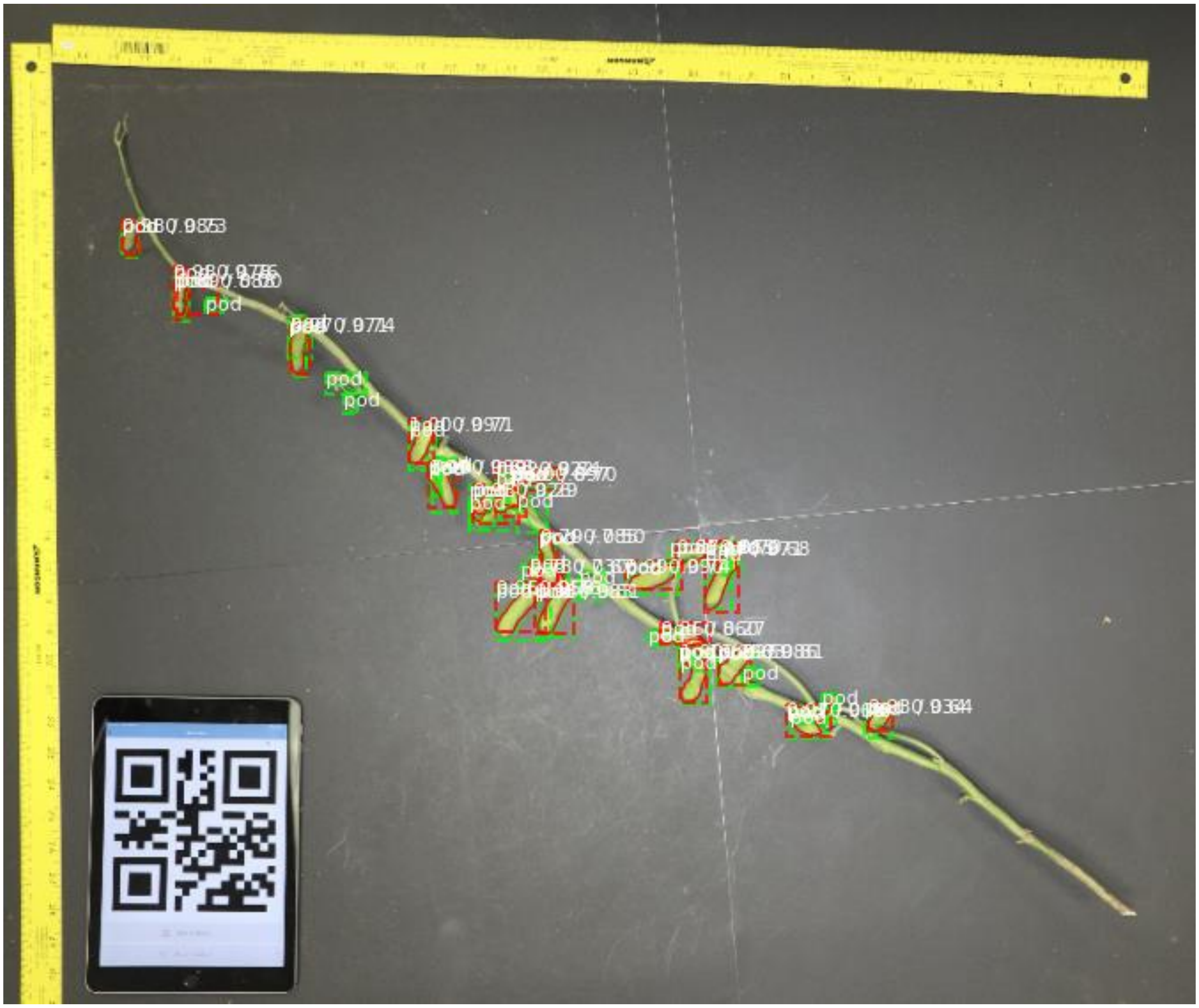


Figure 2.3.S.3 Mask-R-CNN results of IOU of Ground Truth & Predictions of Edamame plant images.

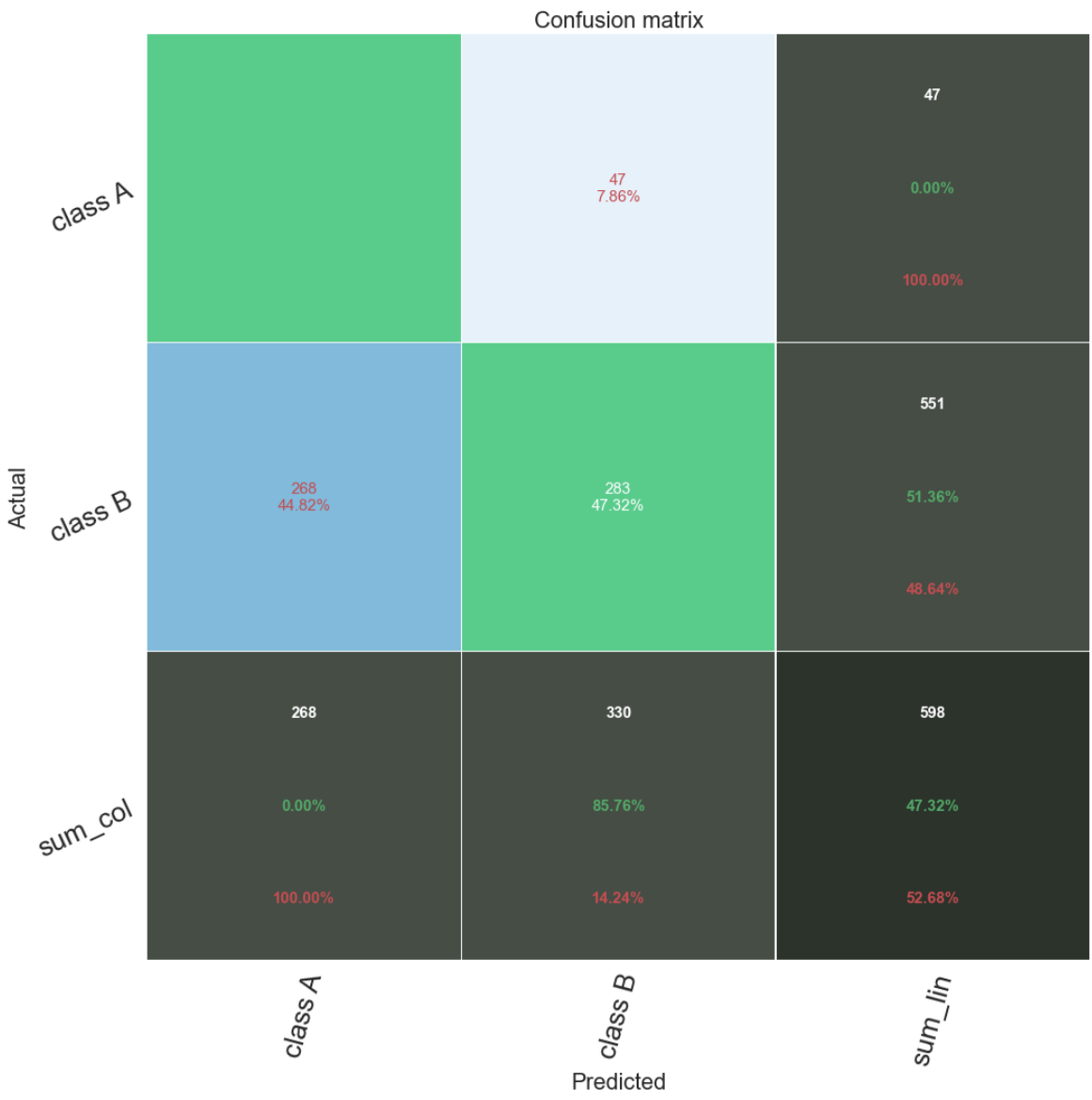


Figure 2.3.S.4 Confusion matrix of Mask-R-CNN results obtained from Edamame plant images.

Here class A is background, which is being counted to cover the cases when the model misses (detect background instead of pods or detect pods instead of background) and class B is pod, sum_col and sum_lin are the sum of columns and lines, respectively. The numbers in white are the number of instances of backgrounds and pods. The column at the far right contains the precision (green colored percentages) and false discovery rate (red colored percentages). The row at the

bottom shows t recall (green colored percentages) and the false negative rate (red colored percentages). The diagonal cells correspond to the observations correctly classified while other cells correspond to incorrectly classified predictions.

There are 330 pods in the evaluation dataset out of which 47 pods are correctly classified. 551 pods are classified out of which 268 are misclassified as background. The mean average precision(mAP) for original RGB images was 0.54, which is low, and the masks were not precise. The mAP was calculated with IoU threshold 0.5. IoU is measured for all classes and averaged over all classes and the mean of the averages of the IoU values is calculated. The mean IoU obtained is 47.32.

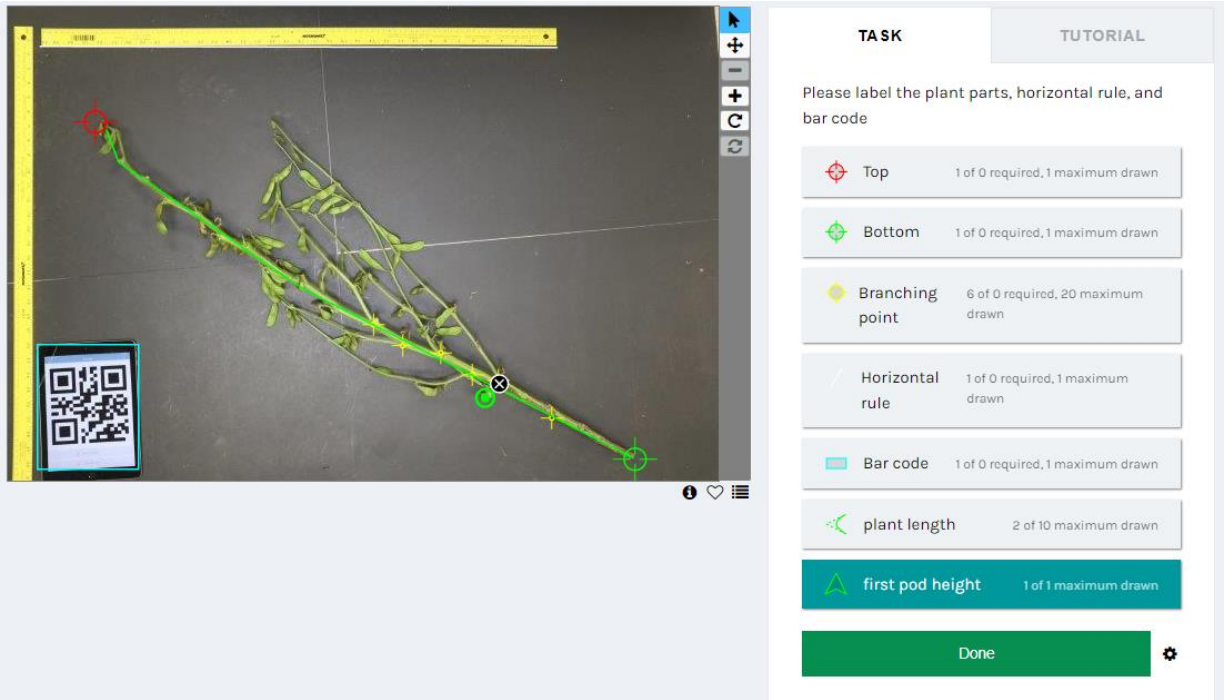


Figure 2.3.A.1 Zooniverse platform for plant parts labelling

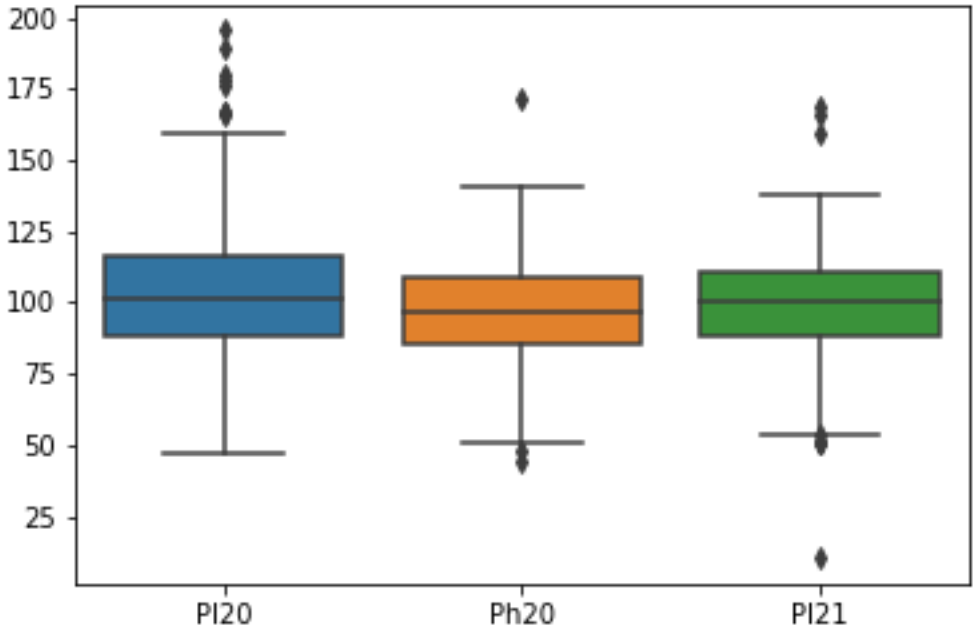


Figure 2.3.A.2 Distribution of plant length and plant heights traits in 2020 and 2021.

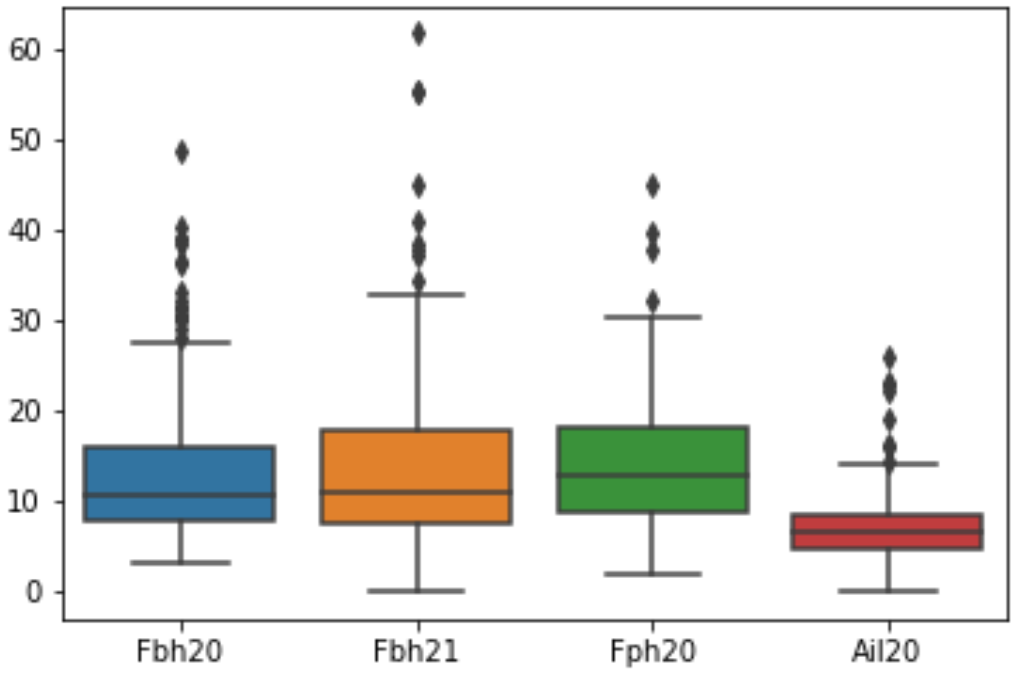


Figure 2.3.A.3 Distribution of pod location and branching pattern traits in 2020 and 2021.

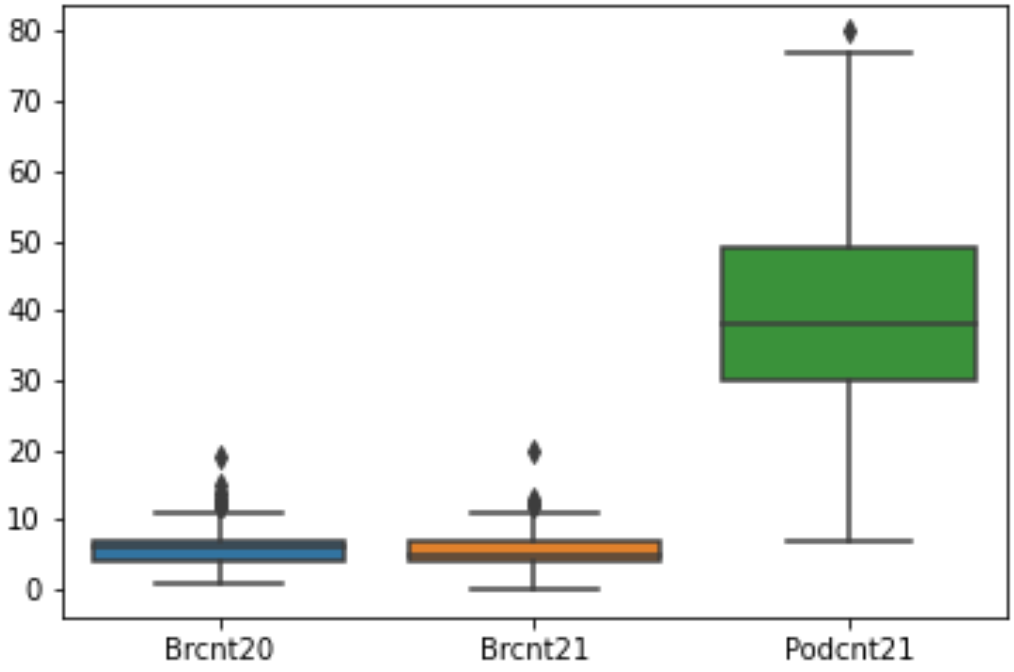


Figure 2.3.A.4 Distribution of number of branches in 2020 and 2021 and number of pods in 2021.

REFERENCES

Barré, Pierre, Ben C. Stöver, Kai F. Müller, and Volker Steinhage. 2017. "LeafNet: A Computer Vision System for Automatic Plant Species Identification." *Ecological Informatics* 40:50–56.

Bian, Yali, John Wenskovitch, and Chris North. 2020. "Deepva: Bridging Cognition and Computation through Semantic Interaction and Deep Learning." *ArXiv Preprint ArXiv:2007.15800*.

Bojarski, Mariusz, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J. Ackel, Urs Muller, Phil Yeres, and Karol Zieba. 2018. "Visualbackprop: Efficient Visualization of Cnns for Autonomous Driving." Pp. 4701–8 in *2018 IEEE International Conference on Robotics and Automation (ICRA)*.

Cox, Michael A. A., and Trevor F. Cox. 2008. "Multidimensional Scaling." Pp. 315–47 in *Handbook of data visualization*. Springer.

Denker, John, W. Gardner, Hans Graf, Donnie Henderson, R. Howard, W. Hubbard, Lawrence D. Jackel, Henry Baird, and Isabelle Guyon. 1988. "Neural Network Recognizer for Hand-Written Zip Code Digits." *Advances in Neural Information Processing Systems* 1.

Gil, J., and R. Kimmel. 2000. "Efficient Dilation, Erosion, Opening and Closing Algorithms in Mathematical Morphology and Its Applications to Image and Signal Processing." *V, J. Goutsias, L. Vincent, and D. Bloomberg, Eds. Palo-Alto, USA* 301–10.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." Pp. 770–78 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1(4):541–51.

Meyer, George E., and Joao Camargo Neto. 2008. "Verification of Color Vegetation Indices for Automated Crop Imaging Applications." *Computers and Electronics in Agriculture* 63(2):282–93.

Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. 2016. "Using Deep Learning for Image-Based Plant Disease Detection." *Frontiers in Plant Science* 7:1419.

Raid, A. M., W. M. Khedr, M. A. El-Dosuky, and Mona Aoud. 2014. "Image Restoration Based on Morphological Operations." *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)* 4(3):9–21.

Schaetti, Nils. 2018. "Character-Based Convolutional Neural Network and Resnet18 for Twitter Authorprofiling." in *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*. Vol. 2125.

Self, Jessica Zeitz, Michelle Dowling, John Wenskovitch, Ian Crandell, Ming Wang, Leanna House, Scotland Leman, and Chris North. 2018. "Observation-Level and Parametric Interaction for High-Dimensional Data Analysis." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8(2):1–36.

Wei Tan, Jing, Siow-Wee Chang, Sameem Abdul-Kareem, Hwa Jen Yap, and Kien-Thai Yong. 2018. “Deep Learning for Plant Species Classification Using Leaf Vein Morphometric.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17(1):82–90.

Yu, D., N. Lord, J. Polk, K. Dhakal, S. Li, Y. Yin, S. E. Duncan, H. Wang, B. Zhang, and H. Huang. 2022. “Physical and Chemical Properties of Edamame during Bean Development and Application of Spectroscopy-Based Machine Learning Methods to Predict Optimal Harvest Time.” *Food Chemistry* 368. doi: 10.1016/j.foodchem.2021.130799.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. “A Comprehensive Survey on Transfer Learning.” *Proceedings of the IEEE* 109(1):43–76.

CHAPTER III

PHYSICAL AND CHEMICAL PROPERTIES OF EDAMAME DURING BEAN DEVELOPMENT AND APPLICATION OF SPECTROSCOPY-BASED MACHINE LEARNING METHODS TO PREDICT OPTIMAL HARVEST TIME

Dajun Yu[a], Nick Lord[b], Justin Polk[b], Kshitiz Dhakal[b], Song Li[b], Yun Yin[a], Susan E. Duncan[a], Hengjian Wang[a], Bo Zhang[b], Haibo Huang[a]

[a] Department of Food Science and Technology, Virginia Tech, Blacksburg, VA, United States

[b] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

Note: This manuscript was published in Food Chemistry on 30 January 2022. DOI: https://doi.org/10.1016/j.foodchem.2021.130799. . My contribution to this work includes performing the data analysis under the guidance of Dr. Song Li, writing, and revising the manuscript.

Citation: Yu, D., Lord, N., Polk, J., Dhakal, K., Li, S., Yin, Y., ... & Huang, H. (2022). Physical and chemical properties of edamame during bean development and application of spectroscopy-based machine learning methods to predict optimal harvest time. Food Chemistry, 368, 130799.

Doi: https://doi.org/10.1016/j.foodchem.2021.130799

ABSTRACT

This study aims to investigate the changes in physical and chemical properties of edamame during bean development and apply a spectroscopy-based machine learning (ML) technique to determine optimal harvest time. The edamame harvested at R5 (beginning seed), R6 (full seed), and R7 (beginning maturity) growth stages were characterized for physical and chemical properties, and pods were measured for spectral reflectance (360–740 nm) using a handheld spectrophotometer. The samples were categorized into 'early,' 'ready,' and 'late' based on the characterized properties. The results showed that pod/bean weight and pod thickness peaked at R6 and remained stable thereafter. Sugar, starch, alanine, and glycine also peaked at R6 but proceeded to decline. The ML method (random forest classification) using pods' spectral reflectance had a high accuracy of 0.95 for classifying 'early' and 'late' samples and 0.87 for classifying 'early' and 'ready' samples. Therefore, this method can determine the optimal harvest time of edamame.

INTRODUCTION

Edamame, also called vegetable soybean [Glycine max (L.) Merr.], has been widely consumed in East Asia for centuries and is a trending soy product in the US. It is highly nutritious due to the content of high-quality protein with isoflavones, vitamins (C and E), monounsaturated fatty acids, minerals, and dietary fiber (Johnson, Wang, and Suzuki 2000; Mentreddy et al. 2002; Zeipi\cna, Alsi\cna, and Lepse 2017). The quality of edamame is mainly determined by morphological quality (pod size, color, and weight), eating quality, and nutrition (Zeipi\cna et al. 2017) and these quality parameters change over bean development.

Harvesting edamame at an appropriate time ensures peak morphological and eating quality, which offers the edamame high marketability and consumer acceptability (Konovsky, Lumpkin, and McClary 2020; Zeipi\cna et al. 2017). Moreover, edamame with consistent quality also eases

post-harvest processing. Edamame should ideally be harvested sometime between R6 and R7 growth stages, just before pods beginning to turn yellow and when moisture and seed weight approach their maximum levels (Moseley et al. 2021; Yu et al. 2021). Given the dynamic nature between the R6 and R7 stages of soybean growth, harvesting edamame outside of its optimal harvest time can potentially jeopardize the marketability of the pods or seeds. For example, harvesting too early can lead to reduced yield, sweetness, and size of seeds, while harvesting too late leads to fibrous and yellow seeds (Carson 2010). Further complicating edamame harvest is the narrow harvest window for growers (about one week) once they reach their optimal harvest time before yellowing begins (Carson et al. 2011). A plant at the R6 growth stage can be distinguished by pods with beans that have filled the pod cavity, while plants at R7 can be distinguished by at least one pod having reached mature pod color (Licht 2014). Accompanying the transition from R6 to R7 is a series of physiological changes that signal the culmination of reproductive growth and the subsequent initiation of senescence towards full maturity (R8). These changes include growing beans occupying approximately 85–90% of pod space, the color of leaves, pods, and beans changing from green to yellow, and sugars and other chemical constituents accumulating in the immature beans.

Studies have been conducted to investigate the physical properties and chemical compositions of edamame or soybean during seed development. Xu et al. (2016) studied the effects of edamame development on the physical, chemical, and anti-nutritional properties of edamame seeds. Saldivar et al. (2011) investigated chemical composition changes including protein, oil, starch, and soluble saccharides and the seed length changes during soybean development from R1 to R8. Yazdi-Samadi et al. (1977) investigated the oil, protein, sugars, starch, organic acids, and amino acid changes in developing soybean seeds. Lowell and Kuo (1989) studied the metabolism

and accumulation of oligosaccharides during the development of soybeans. All these studies provided a clear picture of how the physical and chemical attributes of edamame or soybean seeds change throughout reproductive development. However, exploiting these differences to predict optimal harvest time was rarely studied on edamame.

Current methods for determining optimal harvest time of edamame rely on the ability of experienced edamame growers to detect these changes visually, by touch, or by taste. These determination methods can be quite subjective; they can pose a major obstacle for relatively inexperienced or new edamame growers and cause significant economic losses due to the reduced quality of edamame harvested outside of the optimal window. Therefore, more rapid, consistent, and standardized methods for determining optimal harvest time are desired. Recently, spectroscopic techniques have been used to determine the optimal harvest time of strawberries (Gao et al. 2020; Shen et al. 2018), cherry tomatoes (Yang 2011), and apples (Bertone et al. 2012). However, to our best knowledge, no literature has reported the application of spectroscopic methods to identify the optimal harvest time of edamame. Moreover, using a handheld spectroscopy instrument is promising because it offers a fast way (usually a few seconds) for in-field determining the optimal harvest time of edamame compared to the lengthy chemical analysis. In addition, spectroscopic analysis can identify slight changes of pod color over edamame development, which usually cannot be captured by naked eyes. However, the spectroscopy-based analysis is a secondary method requiring calibration against a reference method for identifying the optimal harvest time of edamame. Calibration is usually conducted using multivariate regression analysis; nevertheless, it sometimes cannot deliver satisfactory results due to the complexity of the spectra (Cortés et al. 2019). Fortunately, the recent developments of machine learning techniques provide an opportunity to analyze the complex spectroscopic dataset and provide accurate and

reliable calibration (Singh et al. 2016; Singh et al. 2018). Random Forest (RF) is an ensemble learning technique and it has received increasing attention due to the excellent classification results and the speed of processing (Belgiu and Dr\uagu\ct 2016) RF has been widely applied to classify different types of food using multispectral and hyperspectral data. For example, RF was successfully applied to classify the adulterated and authentic nutmeg using infrared spectroscopy and the RF presented superior performance than other classification methods with Partial least-squares discriminant analysis (PLS-DA) and Soft independent modeling of class analogy (SIMCA) (de Santana et al., 2019). In another study, Piedad et al. (2018) applied three widely used machine learning methods (i.e., artificial neural network, RF, support vector machines) to classify bananas into various categories based on their measured qualities. The results demonstrated that RF achieved the highest classification accuracy among the three machine learning methods.

The objective of this study is to investigate the changes in the physical and chemical properties of edamame during seed development and apply the spectroscopy-based machine learning technique to determine the appropriate harvest time. In this study, physical properties and chemical compositions were quantified for the edamame harvested from R5 to R7 stages. Physical properties included pod weight, 20-seed weight, pod dimensions (width, length, and thickness), and color. Chemical compositions included soluble sugars and free amino acids (sucrose, fructose, glucose, alanine, and glycine), oligosaccharides (raffinose and stachyose), the moisture of fresh seeds, protein, starch, fat, neutral detergent fiber (NDF), and ash. The quantified physical and chemical properties of harvested seeds were used to determine the optimal stage for edamame harvesting as well as identifying the seeds that were harvested ‘too early’ and ‘too late.’ Meanwhile, the spectral reflectance between 360 and 740 nm was measured on the harvested edamame pods using a handheld, portable spectrophotometer. Using the measured spectra, a

machine learning approach was used to determine the readiness of edamame harvesting based on the collected spectral reflectance. This work would provide a platform technology for developing rapid and accurate prediction of optimum harvest time of edamame, which is essential to ensure consistent and high-quality edamame for the market.

METHODS

Spectral reflectance measurement on edamame pods

Before the weight of 20 seeds was measured, ten pods of each edamame sample were also measured for spectral reflectance between 360 and 740 nm using the portable Konica Minolta CM-700d spectrophotometer (Konica Minolta Sensing Americas, Inc, NJ, USA) with a measurement area of 8 mm, equipped with a pulsed xenon lamp and a diffuse illumination/8º viewing system. Three images were taken of each pod and the spectral reflectance data were averaged by the Konica Minolta software.

Preprocessing of the data

Based on the physical and chemical data obtained, the condition of the edamame samples was evaluated, and all 54 samples were classified into three categories and labeled by “early class”, “ready class” and “late class”. To match with the spectral reflectance dataset (n = 10 for each sample), the same class label was assigned to each of the 10 pods for every sample. Therefore, the dataset consists of 540 spectral reflectance of edamame pods. Among these data, 220 observations were from the late class, 180 were from the early class, and 140 were from the ready class. Both the raw spectral data (primary spectral) and the first-order derivatives (FOD) of each spectral data were calculated and used as a feature matrix for training and testing the random forest (RF) classifier. FOD transformations of the spectral curve are a commonly applied technique used to increase classification quality by enhancing spectral features and minimizing random noise.

Random forest classification

The random forest analysis method was adapted from Heim et al. (2018) using the R programming language and VSURF and caret packages. The dataset was split 80:20 into training and test data subsets, and 10-fold repeated cross-validation was applied to the training data. This process was repeated 100 times and the mean accuracy over these repetitions was calculated. RF classifiers were trained to assign each spectrum to one of the three classes (i.e., early, ready, and late) or two classes (e.g., early vs late) and the model performance as measured by classifier accuracy were compared based on cross-validation results. Accuracy was calculated as the following equation:

$$Accuracy\ (\%) = \frac{True\ Positve + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \times 100$$

All three classes were classified based on 39 predictor variables (waveband at 10nm resolution). The final model parameters were tuned to mtry = 30 and n-tree = 2000 after the best classifier was identified using the training data. Feature selection was performed using VSURF package, and selected spectral bands were used to determine the prediction accuracy.

Statistical analysis

All measurements were performed on biological triplicates and results were presented as means ± standard deviation (n = 3). One-way ANOVA was performed to observe the significant effect of harvest time on physical and chemical properties of edamame seeds and pods, followed by Tukey's Honestly Significant Different (HSD) to compare differences among groups using SPSS (22.0.0.0, IBM Corporation, Armonk, NY, USA). Principle component analysis (PCA) with trajectory analysis was also conducted by SPSS.

RESULTS

Spectral reflectance curves

The spectral data from all pods in early, ready, and late stages were averaged and plotted in Fig. 1. Overall, the spectral reflectance curves increase gradually from 360 to 490 nm, then start to increase rapidly and reach the first peak at approximately 565 nm. Afterward, the reflectance curves decrease and reach the local minimum between 670 and 690 nm before increasing again. The color spectrometer used in this study has a maximum detection wavelength at approximately 700 nm. The shape of the reflectance curves follows the typical observation of vegetation, particularly the green leaves, which are consistent with the observation of the green color of edamame pods when they are ready to be harvested. Although the overall trends are the same for all three categories of edamame pods, there is a clear separation between the average spectral curves among these categories. The pods that are too early to harvest have the lowest average reflectance compared to the pods that are ready and too late to harvest. The increase of reflectance in this spectral range is similar to those observed on senescence leaves (Carter 1993).

Three-class classification using full wavelengths

To determine, quantitatively, whether we can classify edamame pods from different categories using spectral data, we performed the classification using the RF algorithm (Fig. 2). At first, three-class RF classification models were trained using the primary (Prim) spectral data and first-order derivative (FOD) spectral data at full wavelengths, respectively (Fig. 2a). The trained models were tested with a new set of data and the accuracies were 0.69 (Prim) and 0.71 (FOD). Although the accuracies of 0.69–0.71 were not ideal, they indicate that the classification of ‘early’, ‘ready’, and ‘late’ pod stages may be realized using reflective spectra coupled with the machine learning technique, and a further improvement in accuracy is needed. A previous study used

similar spectroscopy-based techniques to predict harvest time for apples (Bertone et al. 2012). In their study, UV–Vis and near-infrared spectroscopies coupled with partial least square regression were used to monitor the chlorophyll content (the green color) of the skin of the red apple, which indicates the apple ripeness. The major difference is that the response variable in Bertone et al.'s study is continuous whereas, in our situation, it is categorical. Therefore, classification using machine learning methods such as RF is feasible for predicting the optimal harvest time for edamame and other vegetables/fruits.

Three-class classification using selected wavelengths

Using many spectral bands might create challenges for collecting data in the field and further lead to redundant information. Massive data generated by spectroscopic technique is also a challenge for data analysis. Therefore, selecting important wavelengths that carry the most useful information with minimal redundancy can help improve both the data collection and the data analysis efficiency. To determine the performances of models using selected important spectral bands, 12 wavelengths were selected from the analysis using the Prim spectral data and 9 wavelengths were selected from the analysis using the FOD spectral data. The number of wavelengths (12 and 9) was automatically selected through a machine learning process called feature selection in the RF method. Three-class RF classification models were trained again using the Prim and FOD spectral data of these selected wavelengths and their performances were shown in Fig. 3a, Prim/Prim, FOD/FOD, FOD/Prim, and Prim/FOD. Interestingly, the models built from selected important wavelengths showed similar accuracies (0.65 to 0.73) compared to the models built from full wavelengths. A previous study has compared the performances of support vector machine (SVM) models built on full spectra and selected optimal spectra of hyperspectral image systems on evaluating the ripeness of strawberries (Zhang et al. 2016). They found that SVM

models built on selected optimal spectra showed acceptable results compared to models built on full spectra. The difference is that in their study, the selected optimal spectra were from two different spectral ranges. Satisfactory results were observed on selected optimal wavelengths from 441.1 to 1013.97 nm and worse results were found on selected optimal wavelengths of 941.46 to 1578.13 nm. Thus, our study, together with others, indicates that the reduction of wavelength number can be realized without compromising the model accuracies for classifying edamame with different maturity stages.

Overall, the model trained using FOD spectral data performs better than using the Prim spectral data. This finding was not surprising because FOD spectra are generally better at resolving overlapping wavebands and reducing random noise. In a detailed investigation of spectral classification techniques, Ghiyamat et al. (2013)showed that FOD-based approaches showed the least improvement (over primary spectra) in complex datasets and the most improvement in less complex datasets. In this study, it was found that classification accuracy increased when using an RF classifier combined with the FOD spectra. Classification accuracy using primary spectra and a random forest classifier could still be considered substantial. Taken together with the results from other studies, both the classification method and the number of classes used in the classification influence whether the FOD spectra can improve the classification accuracy. However, visual inspection of FOD spectral data does not show a clear separation of the curves (Fig. 1).

Two-class classification

After analyzing the confusion matrix, we noticed that the model performed poorly when classifying spectral data of the “ready” category; however, it performed better when separating “early” with “late” categories (Fig. 3). With this observation, the data was further analyzed using two-class classification between every two categories separately with only primary spectral data.

We did not use the FOD because the visual inspection of FOD does not suggest a clear separation of “early” and “late” categories (data not shown). The model accuracy increases substantially to 0.95 for separating “early” and “late” categories and to 0.87 for separating “early” and “ready” categories (Fig. 2b). The model accuracy for separating “late” and “ready” remains as low as 0.68. Feature selection was performed, and 12 wavelengths were selected as the important ones (430, 450, 490, 500, 540, 550, 560, 600, 630, 640, 680, and 700 nm). The model using selected wavelengths provides a similar accuracy for the primary spectral data of edamame pods. The higher model accuracies for two-class classifications agreed with Cen et al. (2016) ’s study, which applied a hyperspectral imaging technique to the detection of chilling injury in cucumber fruit. The overall accuracies for the two-class classifications (i.e., normal and chilling) were 100%, while the overall accuracies for the three-class classifications (i.e., normal, lightly chilling, and severely chilling) were lower at 91.6%.

Overall, it was demonstrated that the RF classification method can identify the early, late, and ready stages of edamame harvest using the spectra collected by a handheld, portable spectrometer. This is practically important because edamame has a short harvest window (only about one week) for producing edamame with high marketability and consumer acceptability. The portable spectrometer coupled with the machine learning technique will allow for rapidly and accurately determining optimal harvest time in the field, ensuring peak morphological and eating quality of edamame, and mitigating the heavy reliance on experienced edamame growers through touch, taste, and observation to determine the harvest time. In real edamame production, it often happens that those pods of different maturity stages are on the same plant at the same time. For the small-scale edamame production where edamame is manually harvested, the technique will provide critical information about which pods are ready to harvest. For the large-scale edamame

production where edamame is harvested by large combines, it is not realistic to only harvest mature pods and leave young ones. However, the technique can still help decide the best harvesting time when most pods in plants are “ready to harvest.”

In the future, it may be necessary to refine spectral sets of data down to some level where a single waveband can be considered unique for the system under investigation. Further investigation would be needed to confirm or disprove this suggestion. The successful discrimination between spectral signatures is only the first step towards using spectral approaches to determine the proper harvesting stage in the edamame industry. The results of the present study represent a proof of concept for incorporating a spectral approach into a precision farming tool used for the edamame.

CONCLUSIONS

This study investigated the changes in physical and chemical properties of edamame over bean development and applied a spectroscopy-based machine learning method to identify the optimal harvest time of edamame. Pod weight, bean weight, and pod thickness reach the peak values at stage R6. The color of edamame becomes lighter, more yellow, and less green as the beans develop. All genotypes have similar chemical composition changes from R5 to R7. The sucrose, alanine, glycine, and starch contents are highest at R6 when the edamame has the 7 highest sweetness. Oppositely, the fat, NDF, and ash contents are relatively low at this stage. Considering all physical properties and chemical composition changes over the bean development, the early R6 (R6-1) stage was determined as the optimal time to harvest edamame. However, if a longer harvest window is needed, R6-2 is acceptable and better than R5 and R7. The machine learning method based on the pods’ spectral reflectance had a high accuracy of 0.95 for classifying “early” and “late” samples and 0.87 for classifying “early” and “ready” samples. However, a low accuracy of

0.68 was obtained for classifying “late” and “ready” samples. Overall, this study demonstrated that the machine learning method based on the pods’ spectra reflectance can identify the optimal harvest time of edamame.

REFERENCES

Belgiu, Mariana, and Lucian Dr\uagu\ct. 2016. “Random Forest in Remote Sensing: A Review of Applications and Future Directions.” *ISPRS Journal of Photogrammetry and Remote Sensing* 114:24–31.

Bertone, E., A. Venturello, R. Leardi, and F. Geobaldo. 2012. “Prediction of the Optimum Harvest Time of ‘Scarlet’Apples Using DR-UV--Vis and NIR Spectroscopy.” *Postharvest Biology and Technology* 69:15–23.

Carson, Luther C. 2010. “Cultivation and Nutritional Constituents of Virginia Grown Edamame.” Virginia Tech.

Carson, Luther C., Joshua H. Freeman, Kequan Zhou, Gregory Welbaum, and Mark Reiter. 2011. “Cultivar Evaluation and Lipid and Protein Contents of Virginia-Grown Edamame.” *HortTechnology* 21(1):131–35.

Carter, Gregory A. 1993. “Responses of Leaf Spectral Reflectance to Plant Stress.” *American Journal of Botany* 80(3):239–43.

Cen, Haiyan, Renfu Lu, Qibing Zhu, and Fernando Mendoza. 2016. “Nondestructive Detection of Chilling Injury in Cucumber Fruit Using Hyperspectral Imaging with Feature Selection and Supervised Classification.” *Postharvest Biology and Technology* 111:352–61.

Cortés, Victoria, José Blasco, Nuria Aleixos, Sergio Cubero, and Pau Talens. 2019. “Monitoring

Strategies for Quality Control of Agricultural Products Using Visible and Near-Infrared Spectroscopy: A Review." *Trends in Food Science \& Technology* 85:138–48.

Gao, Zongmei, Yuanyuan Shao, Guantao Xuan, Yongxian Wang, Yi Liu, and Xiang Han. 2020. "Real-Time Hyperspectral Imaging for the in-Field Estimation of Strawberry Ripeness with Deep Learning." *Artificial Intelligence in Agriculture* 4:31–38.

Ghiyamat, Azadeh, Helmi Zulhaidi M. Shafri, Ghafour Amouzad Mahdiraji, Abdul Rashid M. Shariff, and Shattri Mansor. 2013. "Hyperspectral Discrimination of Tree Species with Different Classifications Using Single-and Multiple-Endmember." *International Journal of Applied Earth Observation and Geoinformation* 23:177–91.

Heim, R. H. J., I. J. Wright, H. C. Chang, A. J. Carnegie, G. S. Pegg, E. K. Lancaster, D. S. Falster, and J. Oldeland. 2018. "Detecting Myrtle Rust (Austropuccinia Psidii) on Lemon Myrtle Trees Using Spectral Signatures and Machine Learning." *Plant Pathology* 67(5):1114–21. doi: 10.1111/ppa.12830.

Johnson, Duane, Shaoke Wang, and Akio Suzuki. 2000. "Edamame: A Vegetable Soybean for Colorado." *Energy (Kcal)* 582:573.

Konovsky, John, Thomas A. Lumpkin, and Dean McClary. 2020. "Edamame: The Vegetable Soybean." Pp. 173–81 in *Understanding the Japanese Food and Agrimarket*. CRC Press.

Licht, Mark. 2014. "Soybean Growth and Development." *Iowa State University Extension and Outreach. Retrieved from Website*.

Lowell, Cadance A., and Tsung Min Kuo. 1989. "Oligosaccharide Metabolism and Accumulation in Developing Soybean Seeds." *Crop Science* 29(2):459–65.

Mentreddy, S. R., A. I. Mohamed, N. Joshee, A. K. Yadav, and others. 2002. “Edamame: A Nutritious Vegetable Crop.” Pp. 432–38 in *Trends in new crops and new uses. Proceedings of the Fifth National Symposium, Atlanta, Georgia, USA, 10-13 November, 2001*.

Moseley, David, Marcos Paulo Da Silva, Leandro Mozzoni, Moldir Orazaly, Liliana Florez-Palacios, Andrea Acuna, Chengjun Wu, and Pengyin Chen. 2021. “Effect of Planting Date and Cultivar Maturity in Edamame Quality and Harvest Window.” *Frontiers in Plant Science* 11:585856.

Piedad, Eduardo Jr, Julaiza I. Larada, Glydel J. Pojas, and Laura Vithalie V Ferrer. 2018. “Postharvest Classification of Banana (Musa Acuminata) Using Tier-Based Machine Learning.” *Postharvest Biology and Technology* 145:93–100.

Saldivar, Xiaoyu, Ya-Jane Wang, Pengying Chen, and Anfu Hou. 2011. “Changes in Chemical Composition during Soybean Seed Development.” *Food Chemistry* 124(4):1369–75.

Shen, Fei, Bin Zhang, Chongjiang Cao, and Xuesong Jiang. 2018. “On-Line Discrimination of Storage Shelf-Life and Prediction of Post-Harvest Quality for Strawberry Fruit by Visible and near Infrared Spectroscopy.” *Journal of Food Process Engineering* 41(7):e12866.

Singh, Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh, and Soumik Sarkar. 2016. “Machine Learning for High-Throughput Stress Phenotyping in Plants.” *Trends in Plant Science* 21(2):110–24.

Singh, Asheesh Kumar, Baskar Ganapathysubramanian, Soumik Sarkar, and Arti Singh. 2018. “Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives.” *Trends in Plant Science* 23(10):883–98.

Xu, Yixiang, Arrieyana Cartier, Daniel Kibet, Krystal Jordan, Ivy Hakala, Stephanie Davis, Edward Sismour, Maru Kering, and Laban Rutto. 2016. “Physical and Nutritional Properties of Edamame Seeds as Influenced by Stage of Development.” *Journal of Food Measurement and Characterization* 10(2):193–200.

Yang, Hai Qing. 2011. “Nondestructive Prediction of Optimal Harvest Time of Cherry Tomatoes Using VIS-NIR Spectroscopy and PLSR Calibration.” Pp. 92–96 in *Advanced Engineering Forum*. Vol. 1.

Yazdi-Samadi, Bahman, R. W. Rinne, and R. D. Seif. 1977. “Components of Developing Soybean Seeds: Oil, Protein, Sugars, Starch, Organic Acids, and Amino Acids 1.” *Agronomy Journal* 69(3):481–86.

Yu, Dajun, Tiantian Lin, Kemper Sutton, Nick Lord, Renata Carneiro, Qing Jin, Bo Zhang, Thomas Kuhar, Steven Rideout, Jeremy Ross, and others. 2021. “Chemical Compositions of Edamame Genotypes Grown in Different Locations in the US.” *Frontiers in Sustainable Food Systems* 5:620426.

Zeipi\cna, Solvita, Ina Alsi\cna, and L\=\iga Lepse. 2017. “Insight in Edamame Yield and Quality Parameters: A Review.” *Research for Rural Development* 2:40–45.

Zhang, Chu, Chentong Guo, Fei Liu, Wenwen Kong, Yong He, and Binggan Lou. 2016. “Hyperspectral Imaging Analysis for Ripeness Evaluation of Strawberry with Support Vector Machine.” *Journal of Food Engineering* 179:11–18. doi: 10.1016/j.jfoodeng.2016.01.002.

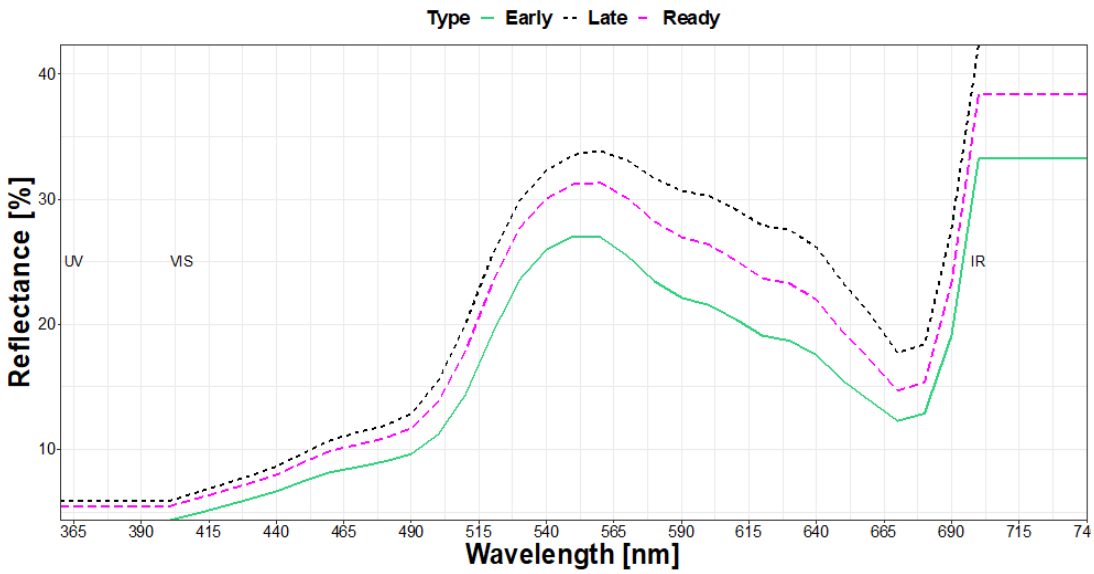


Figure 3.2.1 Analysis of spectral reflectance using machine learning.

Average spectral reflectance of edamame pods that are in the three categories: early, late, and ready to harvest.

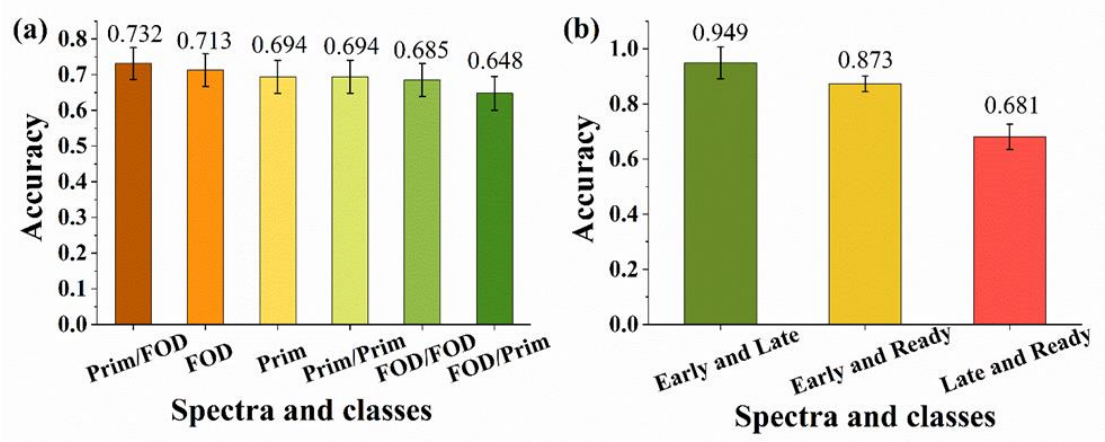


Figure 3.2.2 Comparison of model accuracy among different classification methods.

(A) Classification accuracy of three categories. (B) Classification accuracy of two categories. FOD: first-order derivative of spectral data. Prim: primary spectral data. FOD/FOD: use FOD selected wavelengths and FOD data for classification. FOD/Prim: use FOD selected spectral wavelengths and Prim spectral data for classification. Prim/Prim: use Prim selected spectral wavelengths and Prim spectral data for classification. Prim/FOD: use Prim selected spectral wavelengths and FOD spectral data for classification.

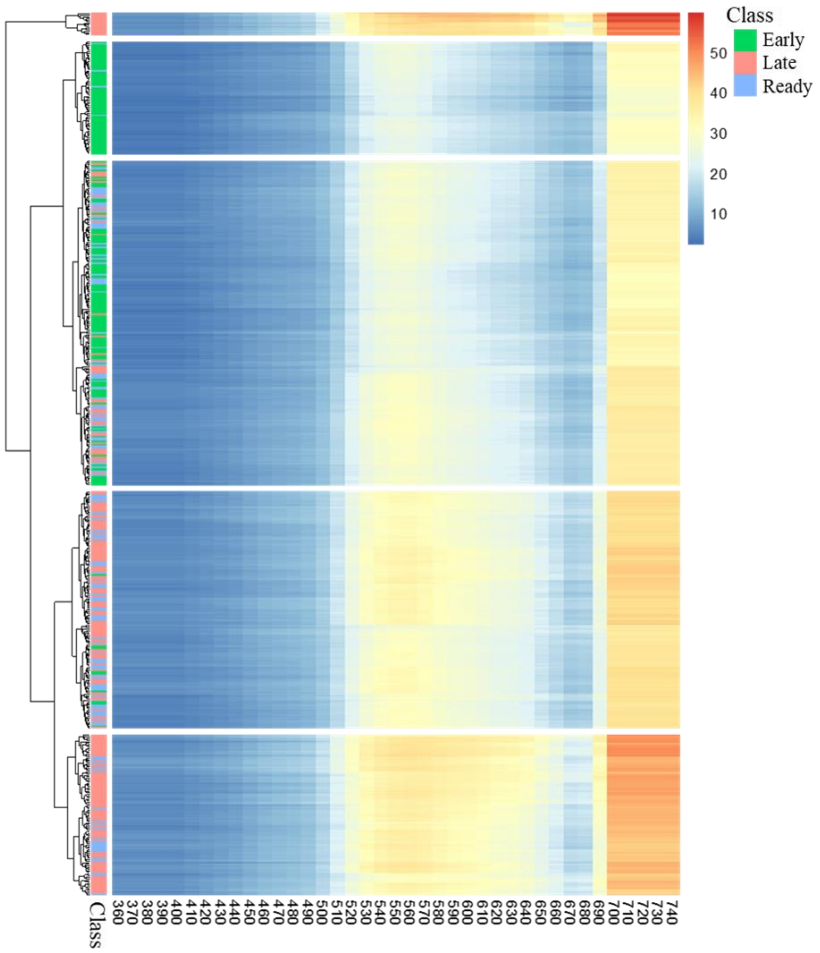


Figure 3.2.3 Hierarchical clustering of spectral reflectance across all spectral data of three categories.

The clustering results are separated into five groups based on the branch height in the dendrogram. More or fewer groups can be generated but the main observation remains the same.

CHAPTER IV

HYPERSPECTRAL IMAGE ANALYSIS OF WHEAT KERNELS FOR DEOXYNIVALENOL QUANTIFICATION USING MACHINE LEARNING

ABSTRACT

Hyperspectral imaging (HSI) has been used widely to perform early disease detection with promising results in recent years. However, most previously published data contained limited spatial or spectral information. In this study, we used Fusarium head blight (FHB) a fungal disease in grain kernels caused by Fusarium *graminearum*, as a model system to evaluate the ability of HSI for the disease classification and quantification of Deoxynivalenol (DON), a mycotoxin accumulated in kernels. Experiments were carried out to determine which machine learning methods had the best accuracy to classify different classes of kernels based on the DON content obtained from Gas Chromatography–Mass Spectrometry (GC-MS). G-Boost, an ensemble method is the most accurate method with 97% accuracy to classify wheat kernels into different severity levels. Mask-R-CNN, a state-of-the-art method of object detection can be used to segment the wheat kernels from HSI data. The regions of interest (ROIs) obtained from Mask-R-CNN achieved a high mAP of 0.98. Results from MaskRCNN, when combined with best classification methods, can be used to correlate the HSI data with the DON content in small grains with R^2 =0.7497. These results show the applicability of HS images to quantify the DON content in wheat kernel.

INTRODUCTION

Wheat (*Triticum aestivum*) is one of the most widely grown crops in the world (Shewry, 2009) and it is an important source of carbohydrates and a leading source of protein in human food. World trade in wheat is greater than all other crops combined. In 2020, wheat was the second most-produced cereal after maize, having world production of 761 million tons (Production &

Statistics, 2020). The global demand for wheat is increasing because of the increased consumption of processed wheat because of the worldwide industrialization process and change in diet.

Fusarium head blight (FHB) is one of the most devastating diseases of wheat worldwide. FHB is caused by a fungal plant pathogen, *Fusarium graminearum*. This fungus can infect wheat heads, resulting in significant yield loss. The major symptom of FHB is a bleaching of heads occurring shortly after flowering (Kleczewski, 2014). The fungal infection spreads into the rachis and nearby spikelet. The fusarium-infected kernels are shriveled, wrinkled, lightweight and dull grayish or pinkish. These kernels sometimes are referred to as "tombstones" because of their chalky, and lifeless appearance (Mcmullen et al., 1997). The kernels may be normal in size with slight discoloration yet still harbor mycotoxins even if there is infection later stage of kernel development (Friskop et al., 2018). This fungus can lead to severe accumulation of the mycotoxin Deoxynivalenol (DON) in wheat and barley kernels. Strict limitations are imposed for contaminated small grain products for both domestic consumption and export (Dohlman, 2003).

Visual disease-rating methods are based on the perception of specific colors that are in the visible light spectral range. Recent advancements of digital technologies, in particular, HSI technologies overcome this limitation of spectral resolution of human vision and have seen a boost in their usage in agriculture, especially in plant phenotyping for disease detection and assess chemical content in food and agriculture products (Abdulridha et al., 2020; Gold, Townsend, Chlus, et al., 2020; Nagasubramanian et al., 2018). The way that HSI works is that a light beam is projected onto the sample and the reflectance of the light is collected by a HSI camera. Human vision and typical digital cameras can see three types of color bands (red, 560-580nm, green, 535-545nm and blue, 420-440 nm) in the visible range (VIS spectrum). Commonly used hyperspectral systems can measure changes in reflectance with in a spectral range of 350 to 2500 nm and see

more narrow bands (~7-10nm bandwidth). A typical hyperspectral scan can generate hundreds of bands of reflectance data. The ability to see extra bands using HSI can identify signatures of chemicals such as leaf chemical composition (Pandey et al., 2016), because different chemical molecules have different reflectance characteristics. However, this large volume of high-dimensional data also created challenges in data analysis such as how to identify the wavelengths related to plant health or chemical composition.

Hyperspectral imaging has been used for the early identification of diseases in several plants. One of the notable usages is the early identification of anthracnose and gray mold in strawberries. In this research, six machine learning methods were developed, and their classification performance were evaluated and compared (Jiang et al., 2021) . A hyperspectral analysis method based on generative adversarial nets (GAN) was used for the early detection of the tomato spotted wilt virus and the results showed that the plant level classification accuracy was 96.25% before the symptoms were visible (Wang et al., 2019). Research on Sugarbeet for early detection and differentiation of Cercospora leaf spot, leaf rust and powdery mildew diseases based on Support Vector Machines and spectral vegetation indices showed that the early differentiation between healthy and inoculated plants as well as among specific diseases can be achieved by a Support Vector Machine (Rumpf et al., 2010). A study to diagnose charcoal rot in soybean was carried out using hyperspectral imaging and deep learning model and the model achieved a classification accuracy of 95.73% and an infected class F1 score of 0.87 (Nagasubramanian et al., 2019).

The VIS-NIR (400-900nm) HSI camera can be used to perform a large-scale screen for kernel and flour toxins for small grain producers. This approach has been used in research (Barbedo et al., 2018) and it has been demonstrated that HSI can classify kernels and flour

regarding whether DON toxin is above the export limit or not. The wheat kernels also can be classified as high and low DON content using HSI. A line-scan Raman hyperspectral imaging (RHI) for simultaneous detection of three potential chemical adulterants in wheat flour was carried out by applying spectral angle mapping (SAM) to distinguish adulterants' pixels from the flour background. The results demonstrated that RHI in combination with SAM performed well for distinguishing the noninvasive quality of powdered foods (Lohumi et al., 2019).

For dimensional reduction of hyperspectral data, feature extraction and band selection methods are being used. Band selection methods help to project the high dimensional datainto low dimension via subset selection of wavelengths from hyperspectral data. Several band selection algorithms have been used for plant disease identification, such as instance-based Relief-F algorithm (Mahlein et al., 2013), genetic algorithms (Nagasubramanian et al., 2018), partial least square (Gold et al., 2020; Gold et al., 2020), and random forest (Heim et al., 2018). In the past several years, machine learning (ML) methods have been increasingly used in crop production systems, especially for plant disease detection (Ganesh Babu & Chellaswamy, 2022; Poornappriya & Gopinath, 2020; Tomar et al., 2021). Machine learning are computational algorithms that can learn from data and perform classification or clustering tasks, which are suitable in finding the patterns of substantial amounts of data such as hyperspectral data. The scikit-learn library provides several functions for different machine learning approaches, dimensional reduction techniques, and feature selection methods. We used these methods in our current research project.

As compared to industrial automation, automation in agriculture is difficult due to the uncertainty regarding plant structures and field conditions. Proper detection and localization for yield components such as determining the distribution of fruits or pods on a plant are important aspects for crop monitoring, and for developing robotics and autonomous systems for agriculture

applications (Duckett et al., 2018). Fruit counting and yield estimation are useful not only to farmers but also to breeders and researchers (Kicherer et al., 2017; Rahim et al., 2021). Affordable RGB cameras plus easy to implement computer vision tools can be used to remove the bottlenecks of fruit counting, as well as other aspects of plant phenotyping. For example, Mask-RCNN (He et al., 2017) (Region-based Convolutional Neural Network) is a convolutional framework for instance segmentation that is simple to train and generalizes well (Liu et al., 2020). YOLO (Redmon & Farhadi, 2016), is a single-stage network that can detect objects without a previous region-proposal stage (Huang et al., 2017). We tested MaskRCNN on hyperspectral data to evaluate whether using the computer vision methods can be used to segment the kernels from the background or not.

The first goal of this manuscript is to determine whether we can classify infected and healthy wheat kernels by their spectral reflectance. We tested multiple machine learning methods and then evaluated the classification results. Second, we performed regression analysis to study the correlation between percentage of infected kernels and pixels with the DON content. Finally, we tested the Mask-R-CNN to segment individual kernels, classify those kernels based on percentage of infected kernel pixels, count the number of kernels in a sample and correlate that with DON content of sample obtained from GCMS.

MATERIALS AND METHODS

Plant Materials & Experimental Design

One hundred and twenty-nine wheat cultivars were grown in research plots at the Virginia Tech Eastern Virginia Agricultural Research and Extension Center (EVAREC) in Warsaw, Virginia during the 2020-2021 growing season. The soil type at this location is Kempsville sandy loam. Meteorological conditions throughout the growing season are presented in Table 4.1. These

129 cultivars were planted in a randomized complete block design. Plots were planted in a conventionally tilled field using a Hege plot planter and were managed using the management practices outlined in Table 4.2. Cultivars were grown under normal conditions and inoculated with *F. graminearum,* causal agent of FHB. Plots were harvested using a Wintersteiger Classic combine. Approximately 100 grams from each plot were used in this analysis.

DON Content Rating and Categorization

The FDA has established advisory levels for DON. The maximum allowable DON level is 1 µg/mL for finished grain products for human consumption; 5 µg/mL in swine (not to exceed 20 percent of ration) and all animal species (except cattle and poultry); 10 µg/mL in ruminating beef and feedlot cattle older than 4 months (providing grain at that level doesn't exceed 50 percent of diet) and poultry (providing grain at that level doesn't exceed 50 percent of diet); and 5 µg/mL in all other animals (providing grains don't exceed 40 percent of diet) (Kleczewski, 2014; Paul et al., 2005)

DON Content Quantification

Samples were ground to homogeneity with a coffee grinder (Hamilton Beach). DON quantification by Gas Chromatography–Mass Spectrometry (GC-MS) was based on methods described by Fuentes et al. (2005) . For GC-MS analysis, an Agilent 7890B/5977B system was used to operate in Selected Ion Monitoring (SIM) mode. An autosampler in split less mode injected 1 µL of each sample to detect DON onto an HP-5MS column (0.25 mm inner-diameter, 0.25 µm film thickness, 30 m length). The inlet temperature was set at 280 ˚C with a column flow rate of 1.2 mL/min using helium. The initial column temperature was detained at 150 ˚C for 1 min, increased to 280 ˚C at a rate of 30 ˚C/min, and held constant for 3.5 min. A post-run of 325 ˚C for 2.5 min was used to clean the column. DON was detected in SIM mode at a mass/charge ratio of

512.3 and had reference ions at 422.4 and 497.3. Mirex (hexachloropentadiene dimer) was used as an internal standard to check the quantitative precision of the instrument and was detected in SIM mode at a mass/charge ratio of 271.8 and had a reference ion of 275.8. A quadratic regression model was used to quantify a seven-point curve of DON with standards (Romer Labs) at concentrations ranging from 0.05 to 5.0 µg/mL.

Image collection and data collection

Two hundred wheat kernels, two kernels in each well were placed in a 3D printed well plate containment one hundred wells. The dimensions of each well were 7mm*3mm*3mm. The well plate containing two hundred kernels were scanned using a benchtop hyperspectral imaging system (Fig. 4.S.5), Pika L 2.4 (Resonon Inc., Bozeman MT, USA) having a 23 mm lens of of 380–1020 nm spectral range, with 281 spectral channels, 15.3º field of view, and 2.1 nm spectral resolution. The system also consists of a linear stage assembly which is moved by a stage motor. There are regulated lights placed above the linear stage to create optimal conditions for performing the scans. The hyperspectral imaging system was placed such that the distance from the lens to the linear stage was 0.5 m. The lights were also at the same level as the lens on a parallel plane. All the scans were performed using the Spectronon Pro (Resonon Inc., Bozeman MT, USA) software, connected to the camera system using a USB cable. We removed the dark current noise, before performing the kernel scans using the software and then calibrated the camera using a white tile (reflectance reference), provided by the manufacturer of the camera system. The white tile was placed in the same conditions as in where the kernel scans were performed. After each scan, the spectral data of the kernels were collected using a post-processing data analysis software (Spectronon Pro, Resonon Inc., Bozeman MT, USA). Several areas containing every class of objects were selected using the selection tool and the mean spectrum was generated. The pixels

were chosen manually by randomly selecting from five spectral scans of each class of objects (background, low DON containing kernels (DON content less than 0.5 µg/mL), and high DON containing kernels (DON content more than 1.5 µg/mL)) to avoid any bias. The reflectance data were exported from the software in the form of excel sheets using the export option (Abdulridha et al., 2020). Also following the method of estimating the percentage of Fusarium-damaged kernels (FDK) developed by Ackerman (2022), we collected the kernel damage percentage. This method quantifies visual symptoms with the help of human observation.

Data Analysis Pipeline

Classification of kernels into healthy and infected using Machine Learning methods

Background and kernel (areas that were healthy-looking and areas that were infected-looking with *F. graminearum*) pixels were selected from hyperspectral images and their reflectance values were imported from Spectronon Pro as input for the analysis pipeline. The reflectance values were normalized, and the average reflectance curve was plotted. Nine different ML methods were deployed on the dataset to compare the classification between different classes of pixels. The nine machine learning methods tested are abbreviated as follows: NB = Gaussian Naive Bayes; KNN = K-nearest neighbors; LDA = Linear discriminant analysis; MLPNN = Multi-layer perceptron neural network; RF = Random forest; SVML = Support vector machine with linear kernel; SVMR = Support vector machine with radial basis function kernel; GBoost = Gradient boosting; PLSDA= Partial Least Squares Discriminant Analysis. We performed stratified 10-fold cross-validation (CV) by repeating the data points three times. Two-class and three-class classification were performed, and the full analysis pipeline is shown in Figure 4.1.A starting from data collection to contour detection.

For two-classification, the data points from background, and foreground (kernels) were taken, split into training and testing set in the ratio of 9:1. The ML methods were trained on the training set and the predictions were tested on the testing set. The ML methods were evaluated using accuracy, F-1 score, precision, and recall. One of the best ML methods was used to select 200 random kernel pixels from each image that was considered healthy (less than 0.5 µg/mL), mild (0.5-1.5 µg/mL), and severe (more than 1.5 µg/mL) sample as obtained from of GC-MS results. There were 10 images of each category and 6000 pixels (2000 kernels pixels from healthy images, 2000 mild images, and 2000 kernel pixels from severe images) were used to train eight machine learning models (except PLSDA among the nine because of having more than two classes) for classification of those pixels into healthy, medium, and severe classes.

For the three-class classification, the same ML methods were deployed to compare the classification between the three classes (background, healthy areas and infected areas with *F. graminearum*). The best ML method was used to select 200 random healthy-looking pixels from each image that was considered healthy (less than 0.5 µg/mL), and 200 random infected-looking pixels from each image that was considered as severe (more than 1.5 µg/mL) sample. The background pixels were not further used in this case. There were 10 images of each category and 4000 pixels (2000 from healthy looking pixels from healthy mages, and 2000 from infected looking pixels from severe images) were used to train eight machine learning models for classification of those pixels into healthy-looking, and infected-looking areas. The evaluation report from one of the best ML methods among the eight methods was taken and the mis-classified and correctly classified pixels were taken for further evaluation. The healthy-looking pixels were taken as positives and the infected looking kernels were considered as negatives to evaluate the data points that were true positives, false positives, false positives, and false negatives. These data

points were further classified as positives and negatives using the same ML models to evaluate the performance of those eight ML models to classify the data points into TP, TN, FP, and FN. The healthy pixels predicted as healthy are referred to as true positives or TP, the severe pixels predicted as healthy are referred to as false positives or FP, the severe pixels predicted as severe are referred to as true negatives or TN, the healthy pixels predicted as severe are referred to as false negatives or FN.

Regression of percent infected kernels over total kernels with GCMS DON content

A regression analysis was performed to study the correlation between the percentage of pixels classified as severe and the DON content obtained from GCMS and the FDK estimates for each image. Finally, a deep neural network method, Mask-R-CNN was implemented to segment individual kernels, classify those kernels based on percentage of infected kernel pixels, count the number of kernels in a sample and correlate that with DON content of sample obtained from GCMS. The preprocessing and the analysis that are performed for this chapter were done in Python programming language (version 3.8.6). The jupyter notebooks are provided GitHub repository (https://github.com/LiLabAtVT/WheatHyperSpectral) that were used for the analysis in this work.

For Mask-R-CNN, we used a dataset containing forty RGB images obtained by converting the HSI data into RGB images (each image of size ~ 800*1600). The dataset was split into a training set containing 30 images (60,000 kernels), a test set composed of 5 images (1000 kernels) and the prediction dataset containing 5 images. Another dataset consisted of ten RGB images obtained by cutting the top two rows of 10 HSI data and converting them into RGB images (each image of size ~ 800*200). This dataset was split into a training set containing 8 images (28*8 kernels), a test set composed of 2 images (28*2 kernels). The objects (kernels) in the images were labeled using VGG Image Annotator (VIA). The annotation files were saved in json format. The

different instances of kernels were detected using publicly available Keras/TensorFlow-based implementation for Mask RCNN by Matterport, Inc. (Abdulla, 2017), by pre-training with the COCO dataset (Lin et al., 2014). Matterport's implementation of Mask R-CNN for patch-based processing was personalized for our kernel region of interest (ROIs) extraction analysis. Thirty HS images (ten spectral scans of each class of samples (healthy (DON content less than 0.5 µg/mL), mild (DON content in between 0.5 and 1.5 µg/mL) and severe (DON content more than 1.5 µg/mL)) were further cut into seven parts to make the kernel detection easier and precise in Mask-R-CNN analysis. The final dataset contained 210 HS images that were saved as RGB images to be used in the prediction dataset of Mask-R-CNN to get the ROIs. The extracted ROIs were saved as json file. The same 210 HS images were further used in classifying the pixels into healthy and infected using the similar protocol described earlier in three class classification. Now, the ROI information from all the images were imported in the classification results where the kernel pixels were classified into healthy and infected. The number of kernels that surpassed a threshold percentage of infected kernel pixels were counted as infected kernels. The final count (number of infected kernels) was correlated with DON content obtained from the GCMS results and the FDK estimates (as shown in Table 4.3).

RESULTS AND DISCUSSION

Machine learning of spectral wavelength can separate background and foreground (kernel)

We first compared nine different machine learning methods in their accuracy of classifying pixels from background and foreground (kernels). From the results, we found a clear separation between the average reflectance of background and foreground (Figure 4.2.A). This result is visualized in figure 4.1.A as well. The average reflectance (as measured in terms of normalized intensity) curve for background remains flat for all the wavelength, while that of foreground

showed a small peak at 410 nm followed by a small drop and kept increasing until 600nm. The average reflectance increases at a slower rate from 600nm to 700nm and then increases faster until it starts decreasing from 900nm onwards.

Most of the ML methods except LDA and NB did the perfect classification of classifying the kernel HSI into background and foreground and there were not any significant differences between each other (Figure 4.2.B). LDA and NB were statistically different from each other and the other remaining seven methods. The accuracy of NB was (0.99±0.02) which is very close to the methods with perfect accuracies and the accuracy of LDA was 89% with 7% variation.

To further separate the kernel images into healthy, mild, and severe class and compare the results with DON content obtained from HPLC, 200 random kernels pixels were picked from each class of images. Since the accuracy of classification of SVML in classifying the kernels pixels from background was perfect, we used SVML to select pixels from kernels first (Figure 4.2.B). Six thousand datapoints were used to calculate the average reflectance curves of the healthy, mild, and severe classes of images. The average reflectance for each category is different, in particular, at 500nm to 600nm wavelength, the healthy pixels had the highest average reflectance followed by severe and then mild pixels had the least average reflectance among the three classes of randomly selected pixels (Figure 4.2.C). The severe category has a lower average reflectance at 720 to 1050 nm range than the mild category, whereas the mild category has a higher average reflectance below 720 nm. After 900nm wavelength, the average reflectance of healthy pixels remained similar and had smaller decrement eventually, while that of mild and severe it decreased at faster rate. However, the standard deviation of the curves is very large and there was no clear separation between them except between 500nm to 600nm for the healthy kernels (Figure 4.2.C).

There were significant differences in accuracies between the eight ML methods of classification to classify the pixels as healthy, mild, and severe pixels (Figure 4.2.D). G-Boost (0.78±0.02) has the highest accuracy of classification and is significantly different than the other methods. NB (0.59±0.02), has the significantly lowest accuracy of classification. The accuracy of classification of healthy, mild, and severe pixels using G-Boost were 0.87, 0.69, and 0.77, respectively. This is consistent with the observation that healthy pixels have more distinct average reflectance (Figure 4.2.C). The confusion matrix of G-Boost results showed that 5349 pixels were correctly classified, and 651 pixels were misclassified. The results suggest that G-Boost can be used to classify the kernel pixels into healthy, medium, and severe classes with moderate level of accuracy (0.78).

Three-class classification: Background, healthy looking area, and infected looking area

In the previous section, we tried to directly classify randomly selected pixels according to the sample DON content (healthy, mild, and severe). However, by observing the images collected for different samples, we found that not all kernels look the same in the same sample. In severely infected samples, we do see a lot of kernels appear to be damaged, but we also have a fraction of pixels that appear to be healthy (Figure 4.S.6). In healthy samples, although most pixels are healthy looking but there is a small fraction of pixels that looks like infected (Figure 4.S.7). Because of this observation, we ask whether can we count the foreground pixels in each image and use this as a proxy for kernel DON content.

The spectral data collected from background, healthy, and infected areas were used to compare the accuracy of classification of different machine learning methods. From the results, we found that the background and the kernels curves are well separated (Figure 4.3A). The average reflectance curve for background remains similar for all the wavelengths, while that of kernels

pixels seems to be increasing from 350 nm to 900 nm and decreases afterwards. The reflectance curve of infected area goes up to 400 nm and keeps decreasing until 410 nm and keeps increasing until 900nm and then eventually decreases. The reflectance curve of healthy-looking area goes up to 400nm and keeps decreasing to 410nm and keeps increasing exponentially to 650nm. The average reflectance of healthy-looking areas decreases from 650nm to 750nm and then increases faster until it starts decreasing from 900nm onwards. These results show the applicability of machine learning methods to separate pixels from healthy and diseased in the same image.

To quantify the performance of different models and to pick the best performing model, we compared the accuracy of eight different machine learning models. There were significant differences in accuracies in the eight ML methods of classification to classify the background**,** healthy looking area, and infected looking area (Figure 4.3.B). Five ML methods had more than 92% accuracy to classify the kernel HSI into three classes and there were not any significant differences between MLPNN, RF, SVML, and G-Boost, and these methods had significantly highest accuracies than the other methods. The highest accuracy was 0.96±0.03 from SVML and the lowest is of LDA (0.80±0.09). These results suggest SVML can be further used to classify the areas of HSI into classify the background**,** healthy area, and infected areas.

Since the data points are already classified as background, healthy area, and infected areas, we wanted to have a matrix that can correlate with the GC-MS DON content. For this, we further picked some random pixels from each image that were considered healthy (less than 0.5 µg/mL), and severe (more than 1.5 µg/mL) and trained machine learning models to classify those pixels into healthy, and infected areas. After getting the number of pixels of healthy, and infected areas, we developed a matrix that was the percentage of severe pixels and correlated with GC-MS DON content. We did not pick the random pixels from image samples that were considered as mild (less

than 0.5 µg/mL and more than 1.5 µg/mL) because we were interested in the extreme classes (severe and healthy not the middle class (mild).

The best ML method was used to randomly select 200 healthy-looking pixels from each image that was considered healthy (less than 0.5 µg/mL), and 200 random infected-looking pixels from each image that was considered as severe (more than 1.5 µg/mL) sample. There were 10 images of each category used in this analysis, and 4000 pixels (2000 from healthy looking pixels from healthy mages, and 2000 from infected looking pixels from severe images) were used to train eight machine learning models for classification of those pixels into healthy, and infected areas. The evaluation report from one of the best ML methods among the eight methods was taken and the mis-classified and correctly classified. The healthy-looking pixels were taken as positives and the infected looking kernels were considered as negatives to get the data points that were true positives, false positives, false positives, and false negatives. These data points were further classified using the same ML models to evaluate the performance of those eight ML models to classify the data points into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The healthy pixels predicted as healthy are referred to as true positives or TP, the severe pixels predicted as healthy are referred to as false positives or FP, the severe pixels predicted as severe are referred to as true negatives or TN, the healthy pixels predicted as severe are referred to as false negatives or FN.

Correlation of DON content with pixel classification results

We further tested the hypothesis that ML can be used to correlate with DON content from GCMS. To test this hypothesis, we used the SVML method to pick 200 random kernels pixels from each class of images. We used SVML for this purpose because SVML performed the best

among other ML methods to classify between three categories. Two hundred pixels from each category were picked for dataset size increment. The average reflectance curve of the healthy, and severe pixels shows there is clear separation between them except until 410nm (Figure 4.3C). The average reflectance curve of healthy pixels is higher than that of severe pixels over the entire wavelength range after 410nm. There were significant differences in accuracies in between the eight ML methods of classification to classify the pixels as healthy, and severe pixels (Figure 4.3D). G-Boost (0.93±0.01) has the highest accuracy of classification and is significantly different than the other methods while, NB (0.72±0.02), has the significantly lowest accuracy of classification. The accuracy of classification of healthy and severe pixels using G-Boost were 0.86 and 0.99, respectively. The confusion matrix of G-Boost results showed that 3851 pixels were correctly classified, and 149 pixels were misclassified.

The evaluation report from G-Boost, the best ML method was taken, and the mis-classified and correctly classified data points were separated. The healthy-looking pixels were taken as positives and the infected looking kernels were considered as negatives to get the data points that were true positives, false positives, false positives, and false negatives. The average reflectance curve of those data points was plotted against the wavelengths and the results showed no clear separation between the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) pixels (Figure 4.3E). The average reflectance curve of true positive pixels is higher than that of TN, and FP over the entire wavelength range after 410nm.

Regression of GCMS DON content over other estimates

We correlated GCMS DON content with other estimates such as Fusarium-damaged kernels (FDK)estimate, percent of severe pixels over all kernel pixels as obtained by ML methods and number of infected kernels as obtained by Mask-R-CNN method. The mean average

precision(mAP) for original RGB images was 0.013, which is low, and the masks were not precise. The mAP was calculated with IoU threshold 0.5. IoU is measured for all classes and averaged over all classes and the mean of the averages of the IoU values is calculated. The mean IoU obtained is 0.95 as shown in Figure 4.S.3. To improve the performance of this analysis, we cropped the images into smaller image, and the cropped images showed substantially improved precision with mAP 0.97 with IoU threshold 0.5. The mean IoU obtained in this case is 97.62 as shown in Figure 4.S.4. Using these cropped images for ML and extracting ROIs, we classified a kernel as infected using different threshold of percentage infected pixels over all pixels within a kernel. For each sample, the number of infected kernels were obtained and those numbers were normalized to be used in correlation analysis with GCMS DON content. Although positive correlation was seen between the percentage of pixel that were predicted as infected over healthy, the FDK estimate and the DON content obtained from GCMS, the correlation coefficient was not more than 0.45 (Figure 4.4). The correlation between the GCMS DON content and the number of infected kernels obtained via Mask-R-CNN ROIs and ML method were better as compared to the correlation between the GCMS DON content and FDK estimate and percent severe pixels as shown in Figure 4.4, 4.5, and 4.6. The different pixel threshold to classify a kernel as infected showed varying correlation coefficient (as shown in Table 4.3) and the 70% thresholding gave the best result with $R^2 = 0.7497$. This shows the applicability of HS images to quantify the DON content in small grains.

CONCLUSION

This paper showed a study on the use of NIR hyperspectral images to detect the DON content in wheat kernels. G-Boost, an ensemble method, is the most accurate method to classify background and foreground pixels and can more accurately classify the foreground(kernel) pixels into healthy and severe classes based on DON contamination level. Current State-of-the-Art

methods of object detection combined with best classification methods can be used to quantify the DON content in small grains.

REFERENCES

Abdulla, W. (2017). *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*.

Abdulridha, J., Ampatzidis, Y., Kakarla, S. C., & Roberts, P. (2020). Detection of target spot and bacterial spot diseases in tomato using UAV-based and benchtop-based hyperspectral imaging techniques. *Precision Agriculture*, *21*(5), 955–978. https://doi.org/10.1007/s11119-019-09703-4

Ackerman, A. J., Holmes, R., Gaskins, E., Jordan, K. E., Hicks, D. S., Fitzgerald, J., Griffey, C. A., Mason, R. E., Harrison, S. A., Murphy, J. P., Cowger, C., & Boyles, R. E. (2022). Evaluation of Methods for Measuring Fusarium-Damaged Kernels Wheat. *Agronomy*, *12*(2). https://doi.org/10.3390/agronomy12020532

Barbedo, J. G. A., Guarienti, E. M., & Tibola, C. S. (2018). Detection of sprout damage in wheat kernels using NIR hyperspectral imaging. *Biosystems Engineering*, *175*, 124–132. https://doi.org/10.1016/j.biosystemseng.2018.09.012

de la Fuente, E., Martínez-Castro, I., & Sanz, J. (2005). Characterization of Spanish unifloral honeys by solid phase microextraction and gas chromatography-mass spectrometry. *Journal of Separation Science*, *28*(9–10), 1093–1100. https://doi.org/10.1002/jssc.200500018

Dohlman, E. (2003). Mycotoxin Hazards and Regulations. *International Trade and Food Safety: Economic Theory and Case Studies*, 97.

Duckett, T., Pearson, S., Blackmore, S., Grieve, B., Chen, W.-H., Cielniak, G., Cleaversmith, J., Dai, J., Davis, S., Fox, C., From, P., Georgilas, I., Gill, R., Gould, I., Hanheide, M., Hunter, A., Iida, F., Mihalyova, L., Nefti-Meziani, S., … Yang, G.-Z. (2018). *Agricultural Robotics: The Future of Robotic Agriculture*. http://arxiv.org/abs/1806.06762

Friskop, A., Shaobin, Z., & Brueggeman, R. (2018). Fusarium head blight (scab) of small grains. *PP804*.

Ganesh Babu, R., & Chellaswamy, C. (2022). Different stages of disease detection in squash plant based on machine learning. *Journal of Biosciences*, *47*(1). https://doi.org/10.1007/s12038-021-00241-8

Gold, K. M., Townsend, P. A., Chlus, A., Herrmann, I., Couture, J. J., Larson, E. R., & Gevens, A. J. (2020). Hyperspectral measurements enable pre-symptomatic detection and differentiation of contrasting physiological effects of late blight and early blight in potato. *Remote Sensing*, *12*(2). https://doi.org/10.3390/rs12020286

Gold, K. M., Townsend, P. A., Larson, E. R., Herrmann, I., & Gevens, A. J. (2020). Contact reflectance spectroscopy for rapid, accurate, and nondestructive phytophthora infestans clonal lineage discrimination. *Phytopathology*, *110*(4), 851–862. https://doi.org/10.1094/PHYTO-08-19-0294-R

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Heim, R. H. J., Wright, I. J., Chang, H. C., Carnegie, A. J., Pegg, G. S., Lancaster, E. K., Falster, D. S., & Oldeland, J. (2018). Detecting myrtle rust (Austropuccinia psidii) on lemon myrtle

trees using spectral signatures and machine learning. *Plant Pathology*, *67*(5), 1114–1121. https://doi.org/10.1111/ppa.12830

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., & others. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7310–7311.

Jiang, Q., Wu, G., Tian, C., Li, N., Yang, H., Bai, Y., & Zhang, B. (2021). Hyperspectral imaging for early identification of strawberry leaves diseases with machine learning and spectral fingerprint features. *Infrared Physics \& Technology*, *118*, 103898.

Kicherer, A., Herzog, K., Bendel, N., Klück, H. C., Backhaus, A., Wieland, M., Rose, J. C., Klingbeil, L., Läbe, T., Hohl, C., Petry, W., Kuhlmann, H., Seiffert, U., & Töpfer, R. (2017). Phenoliner: A new field phenotyping platform for grapevine research. *Sensors (Switzerland)*, *17*(7). https://doi.org/10.3390/s17071625

Kleczewski, N. (2014). *Fusarium head blight management in wheat*. Cooperative.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 740–755.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, *128*(2), 261–318.

Lohumi, S., Lee, H., Kim, M. S., Qin, J., & Cho, B. K. (2019). Raman hyperspectral imaging and

spectral similarity analysis for quantitative detection of multiple adulterants in wheat flour. *Biosystems Engineering*, *181*, 103–113. https://doi.org/10.1016/j.biosystemseng.2019.03.006

Mahlein, A. K., Rumpf, T., Welke, P., Dehne, H. W., Plümer, L., Steiner, U., & Oerke, E. C. (2013). Development of spectral indices for detecting and identifying plant diseases. *Remote Sensing of Environment*, *128*, 21–30. https://doi.org/10.1016/j.rse.2012.09.019

Mcmullen, M., Jones, R., Gallenberg, D., & America, S. (1997). 3Cab of 7Heat and " Arley ! 2E Emerging $ Isease of $ Evastating) Mpact. *Plant Disease*, *81*(12).

Nagasubramanian, K., Jones, S., Sarkar, S., Singh, A. K., Singh, A., & Ganapathysubramanian, B. (2018). Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods*, *14*(1), 1–13. https://doi.org/10.1186/s13007-018-0349-9

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., & Ganapathysubramanian, B. (2019). Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods*, *15*(1). https://doi.org/10.1186/s13007-019-0479-8

Pandey, P., Ge, Y., Stoerger, V., & Schnable, J. C. (2017). High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. *Frontiers in Plant Science*, *8*(August), 1–12. https://doi.org/10.3389/fpls.2017.01348

Paul, P. A., Lipps, P. E., & Madden, L. V. (2005). Relationship between visual estimates of Fusarium head blight intensity and deoxynivalenol accumulation in harvested wheat grain: A meta-analysis. *Phytopathology*, *95*(10), 1225–1236. https://doi.org/10.1094/PHYTO-95-

1225

Poornappriya, T. S., & Gopinath, R. (2020). Article ID: IJEET_11_10_050 Artificial Intelligence Approaches. *International Journal of Electrical Engineering and Technology (IJEET)*, *11*(10), 392–402. https://doi.org/10.34218/IJEET.11.10.2020.050

Production, F., & Statistics, T. (2020). Available online: http://www. fao. org/faostat/en/\# data. *QC/Visualize (Accessed on 30 November 2022)*.

Rahim, U. F., Utsumi, T., & Mineno, H. (2021). *Comparison of grape flower counting using patch-based instance segmentation and density-based estimation with convolutional neural networks*. 72. https://doi.org/10.1117/12.2605670

Redmon, J., & Farhadi, A. (2016). *YOLO9000: Better, Faster, Stronger*. http://arxiv.org/abs/1612.08242

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., & Plümer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, *74*(1), 91–99.

Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany*, *60*(6), 1537–1553. https://doi.org/10.1093/jxb/erp058

Tomar, A., Malik, H., Kumar, P., & Iqbal, A. (2021). Machine Learning , Advances in Computing , Renewable Energy and Communication. In *Proceedings of MARC 2020* (Vol. 768).

Wang, D., Vinson, R., Holmes, M., Seibel, G., Bechar, A., Nof, S., & Tao, Y. (2019). Early Detection of Tomato Spotted Wilt Virus by Hyperspectral Imaging and Outlier Removal

Auxiliary Classifier Generative Adversarial Nets (OR-AC-GAN). *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-40066-y

Tables and Figures

Table 4.1 Meteorological conditions for 2020-2021 growing season

Month	**Avg High (F)**	**Avg Low (F)**	**Rainfall (in.)**
October	71	51	6.97
November	65	47	5.62
December	51	32	5.99
January	46	30	2.79
February	44	32	2.34
March	62	39	4.55
April	69	45	2.42
May	77	52	4.87
June	85	64	6.70

Date	Activity
10/7/2020	Applied 30-60-60-12
10/8/2020	Applied 1 Ton/Acre Lime
10/24/2020	Planted
12/10/2020	Applied 25 lbs./Acre 12-0-0-1.5
1/30/2021	Applied 25 lbs./Acre 12-0-0-1.5
3/27/2021	Applied 60 lbs./Acre 24-0-0-3
6/17/2021	Harvested

Table 4.2 Field management practices for growing wheat

Table 4.3 Coefficient of determination (R2) of different thresholding percentage of severe kernel pixels over total kernel pixels to classify a kernel as severe and the GCMS DON content.

Threshold %	**Coefficient of Determination (R^2)**
5	0.23
10	0.31
15	0.36
20	0.44
30	0.57
40	0.62
50	0.73
60	0.74
70	0.75
80	0.74

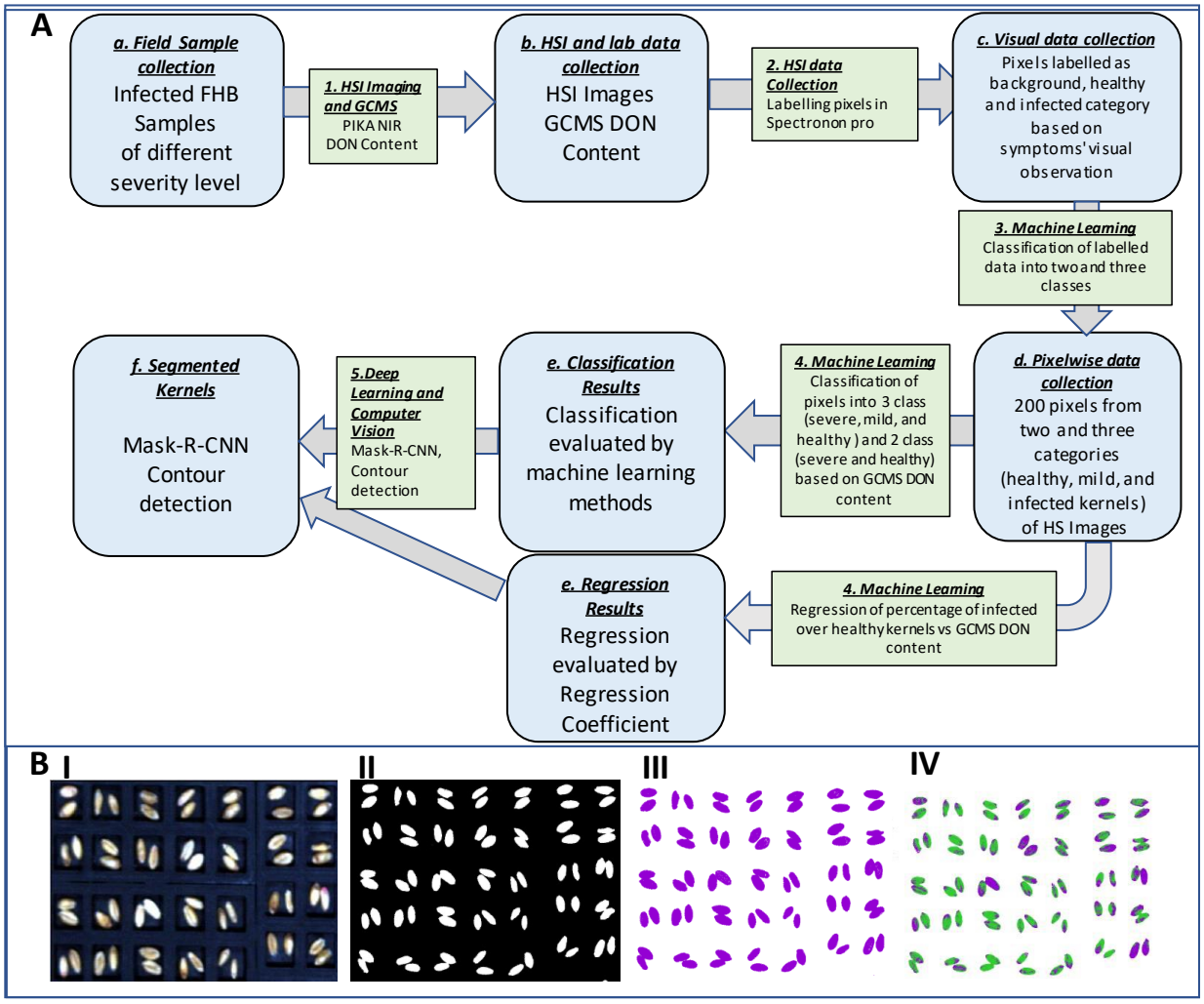


Figure 4.1.A Data analysis pipeline to select wavelengths for classifying healthy wheat kernels and kernels infected with Fusarium graminearum. B. I. RGB representation of HSI II. Binary Image of RGB Image III. Classification of HSI into foreground(purple) and background (white) pixels IV. Classification of HSI into infected (purple), healthy (green), and background (white) pixels.

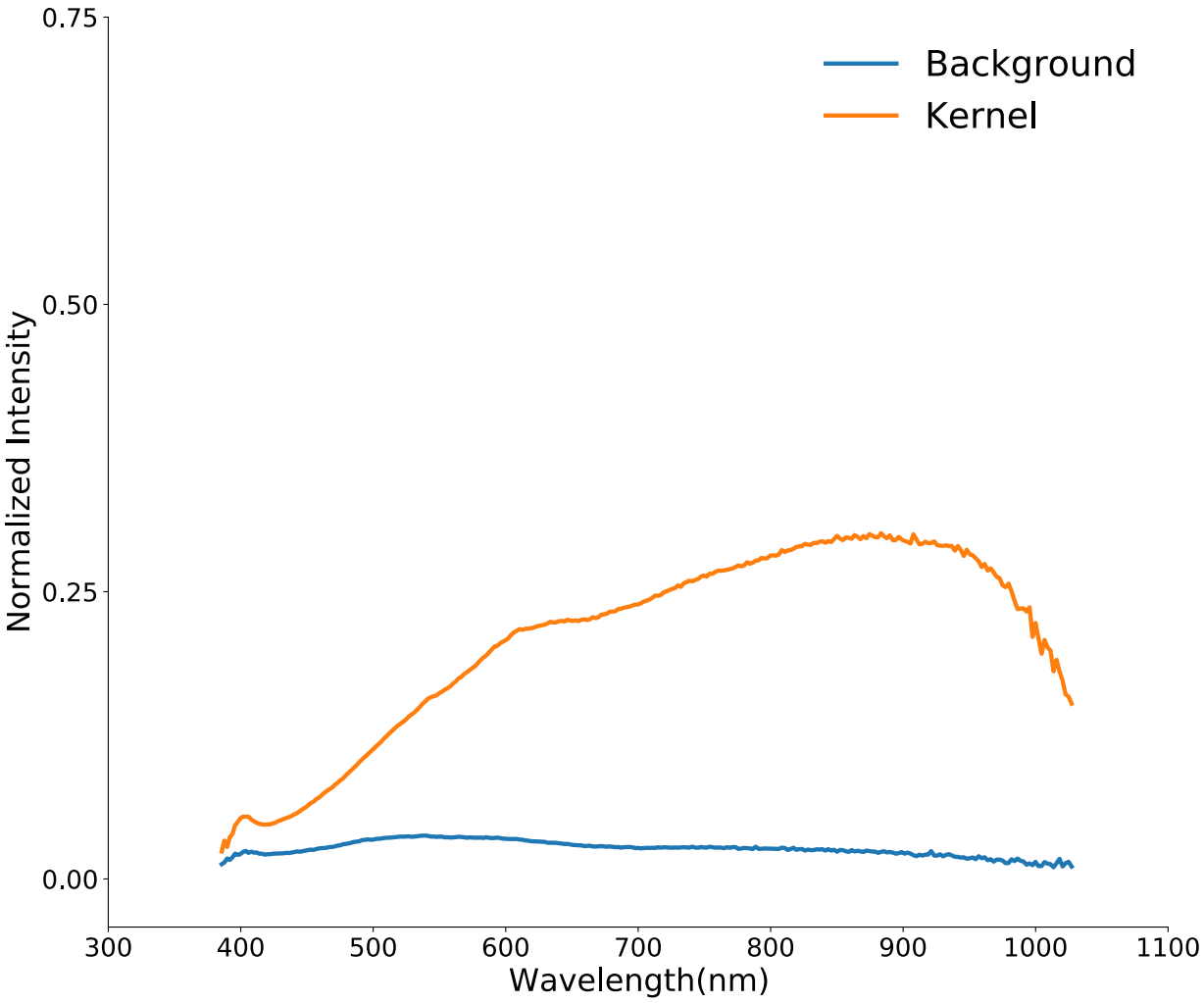


Figure 4.2. A Spectral profile of Background, and Kernels data points.

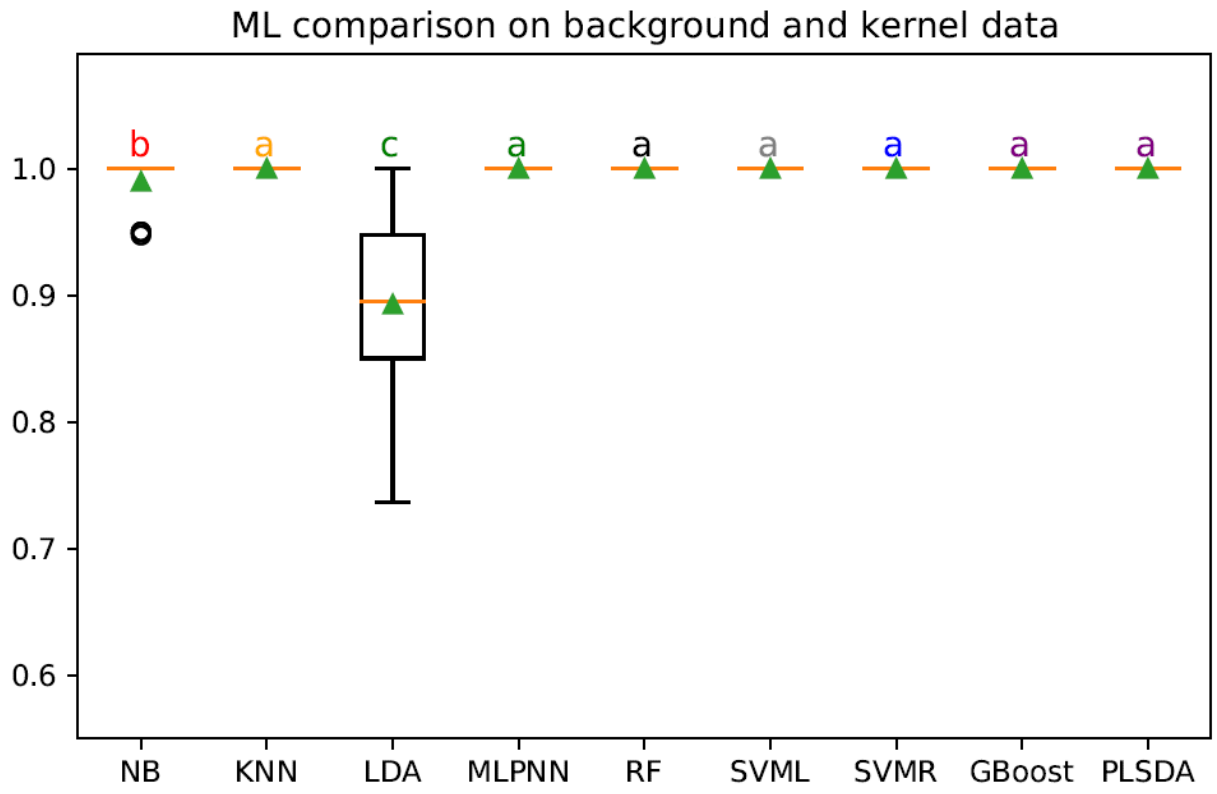


Figure 4.3. B Performance of nine machine learning methods compared to classify data points into background and kernel classes.

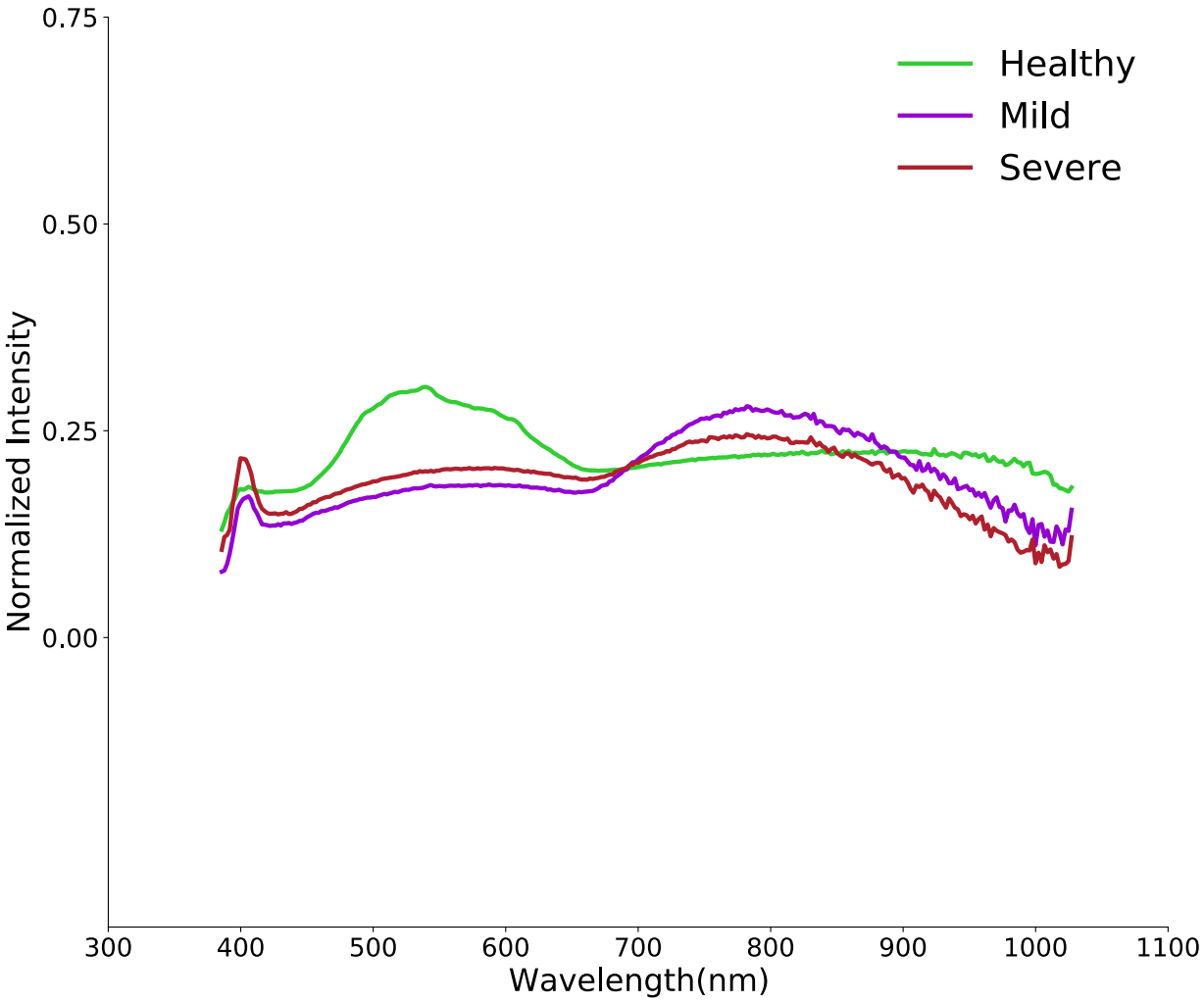


Figure 4.4**.** C Spectral profiles of Healthy, Mild, and Severe pixels.

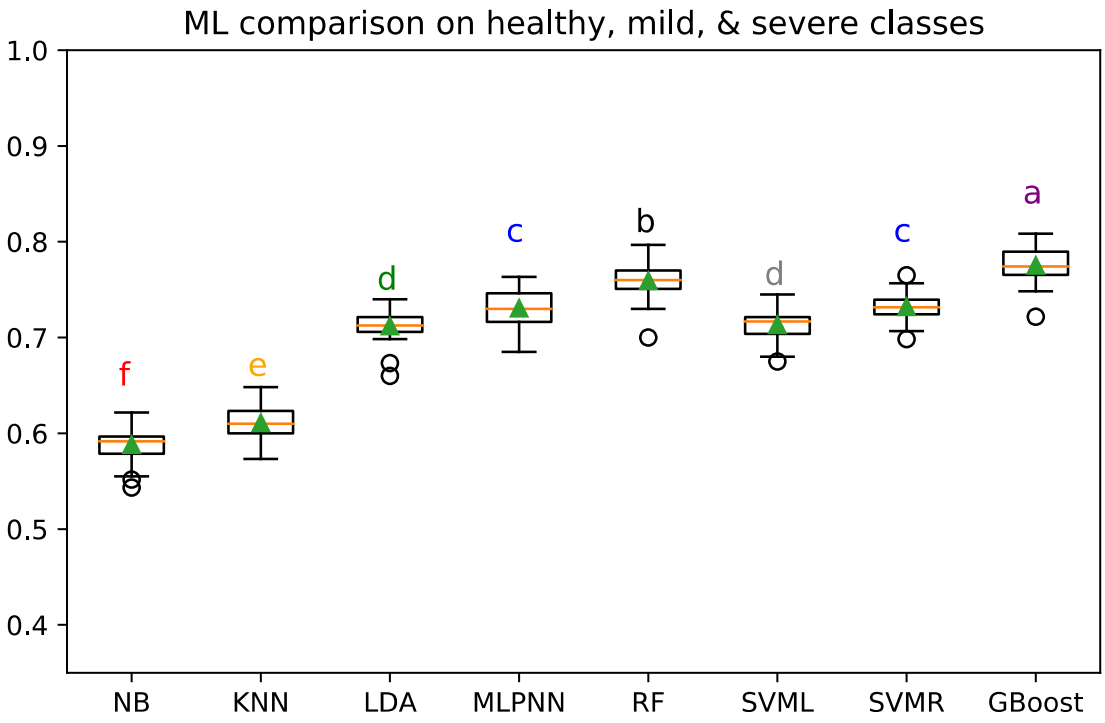


Figure 4.5**.** D Performance of eight machine learning methods compared to classify pixels into Healthy, Mild, and Severe classes.

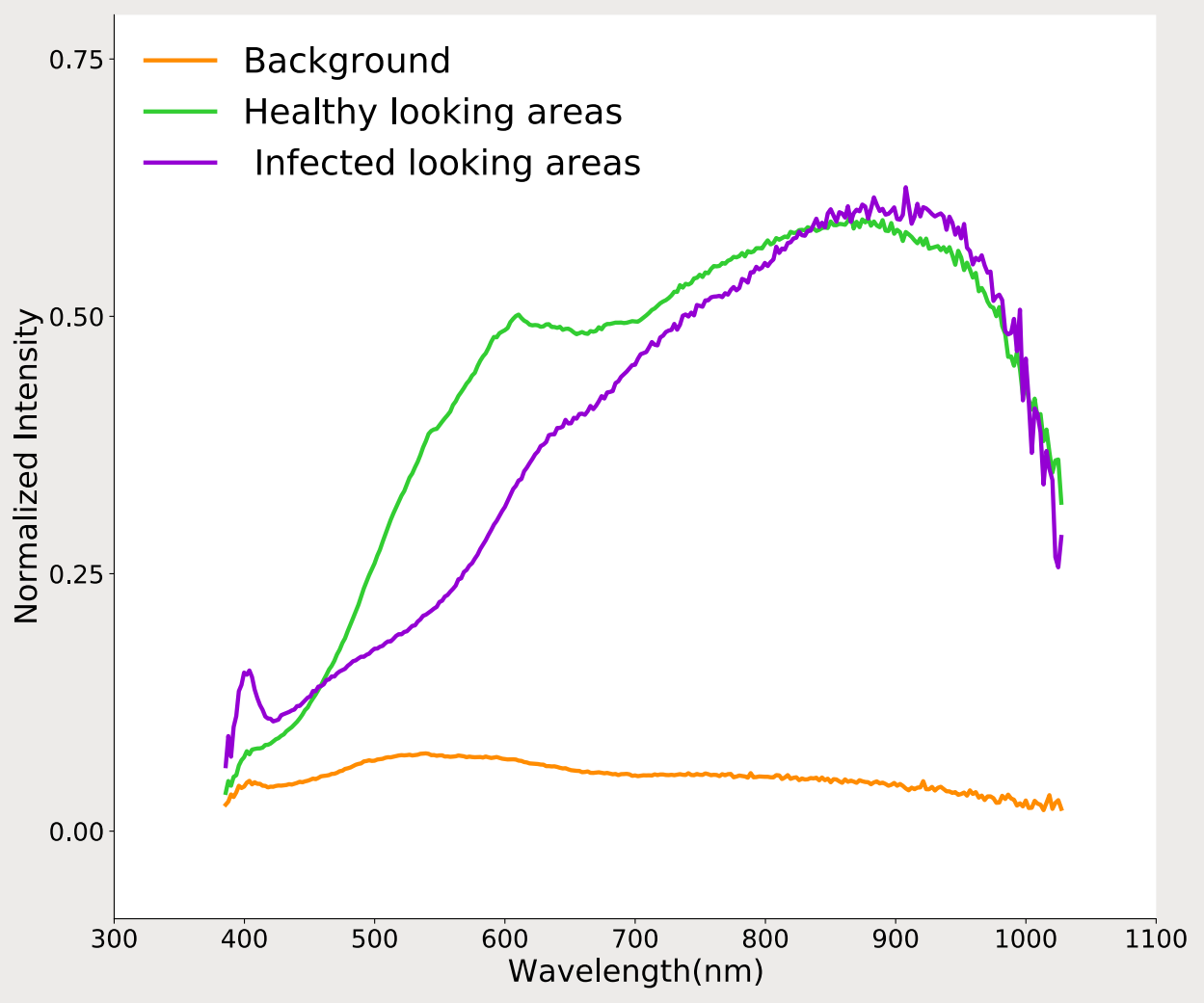


Figure 4.6. A Spectral profile of Background, Healthy-looking areas and Infected-looking healthy areas of wheat kernel HIS.

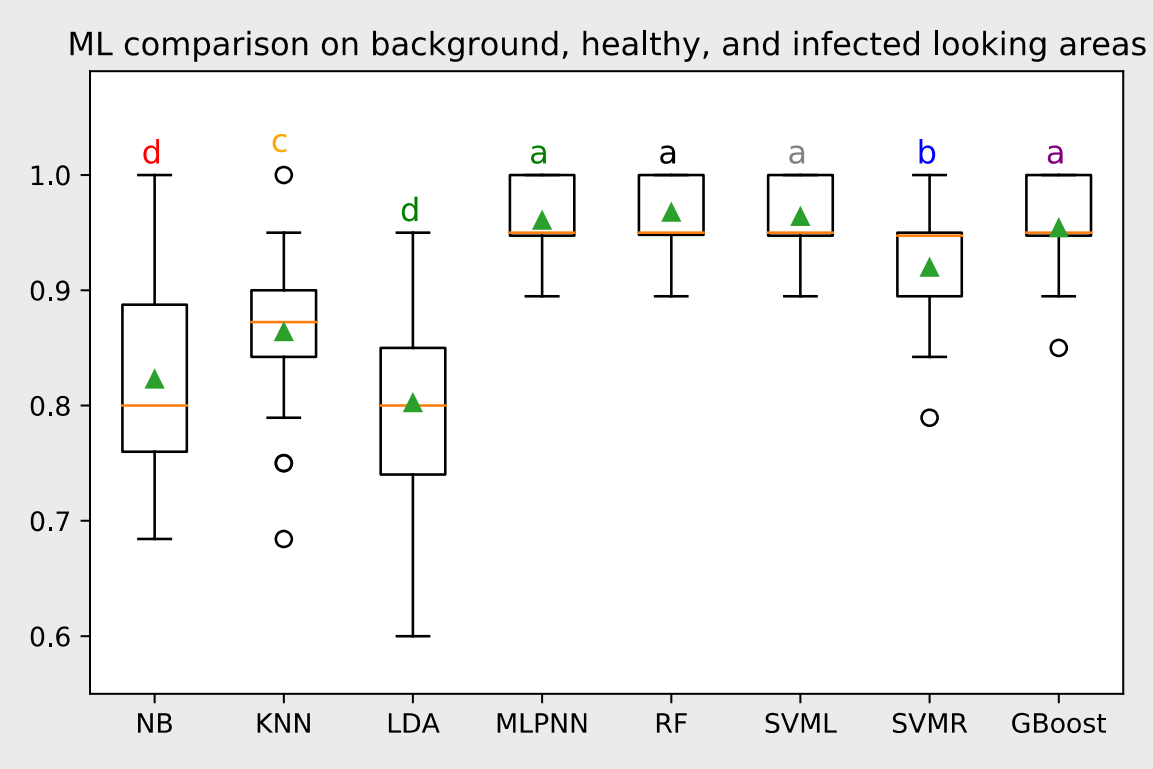


Figure 4.7**.** B Performance of eight machine learning methods compared to classify data points into Background, Healthy-looking areas and Infected-looking healthy areas classes.

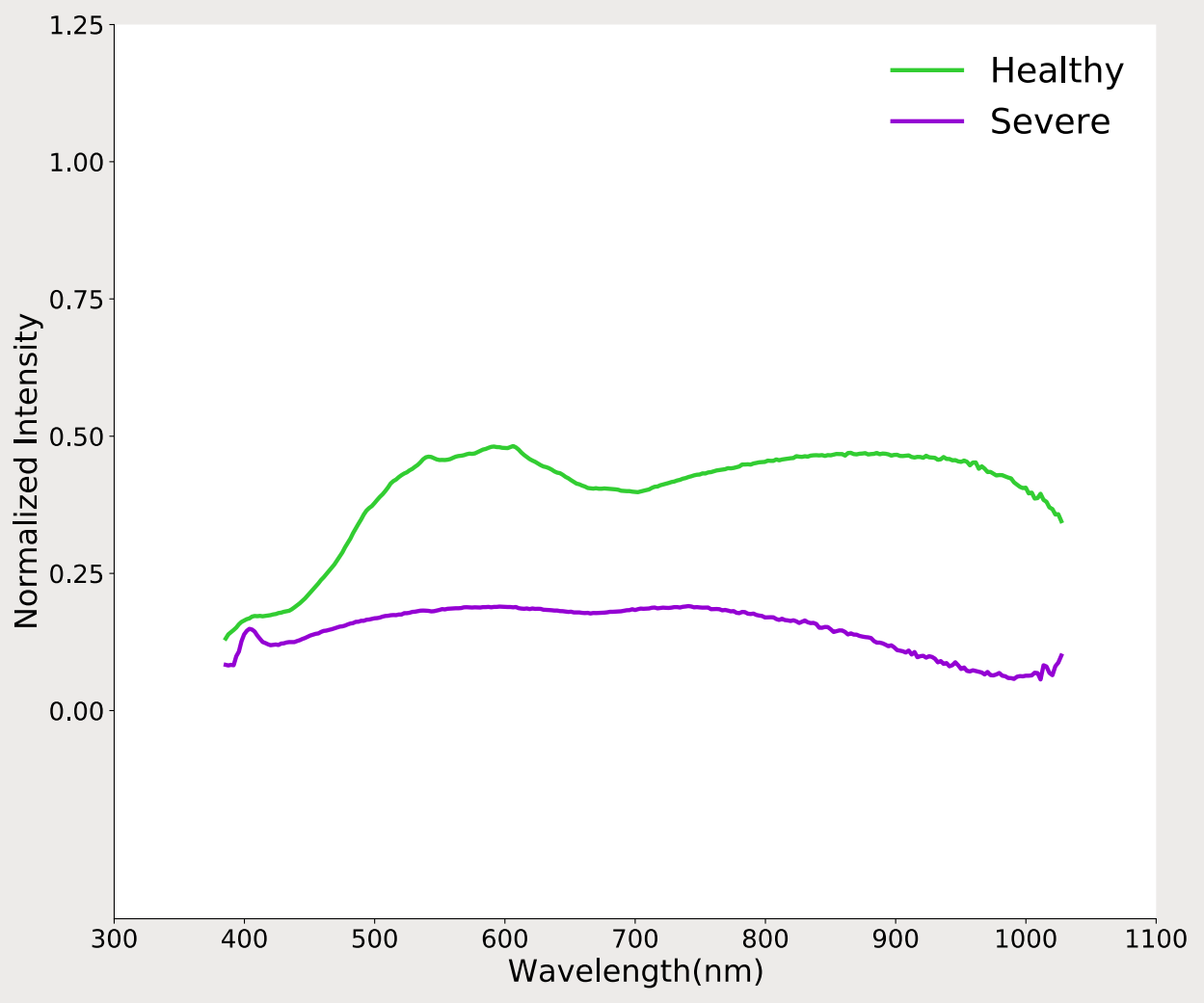


Figure 4.8. C Spectral profiles of Healthy, and Severe pixels.

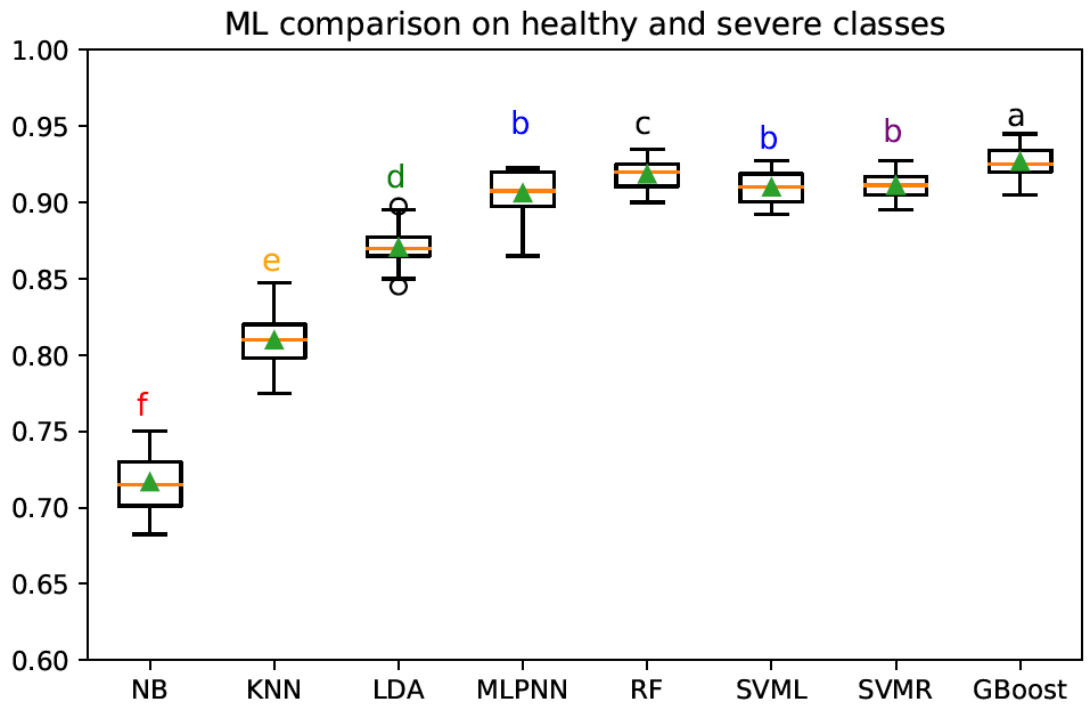


Figure 4.9. D Performance of eight machine learning methods compared to classify data points into Healthy, and Severe classes.

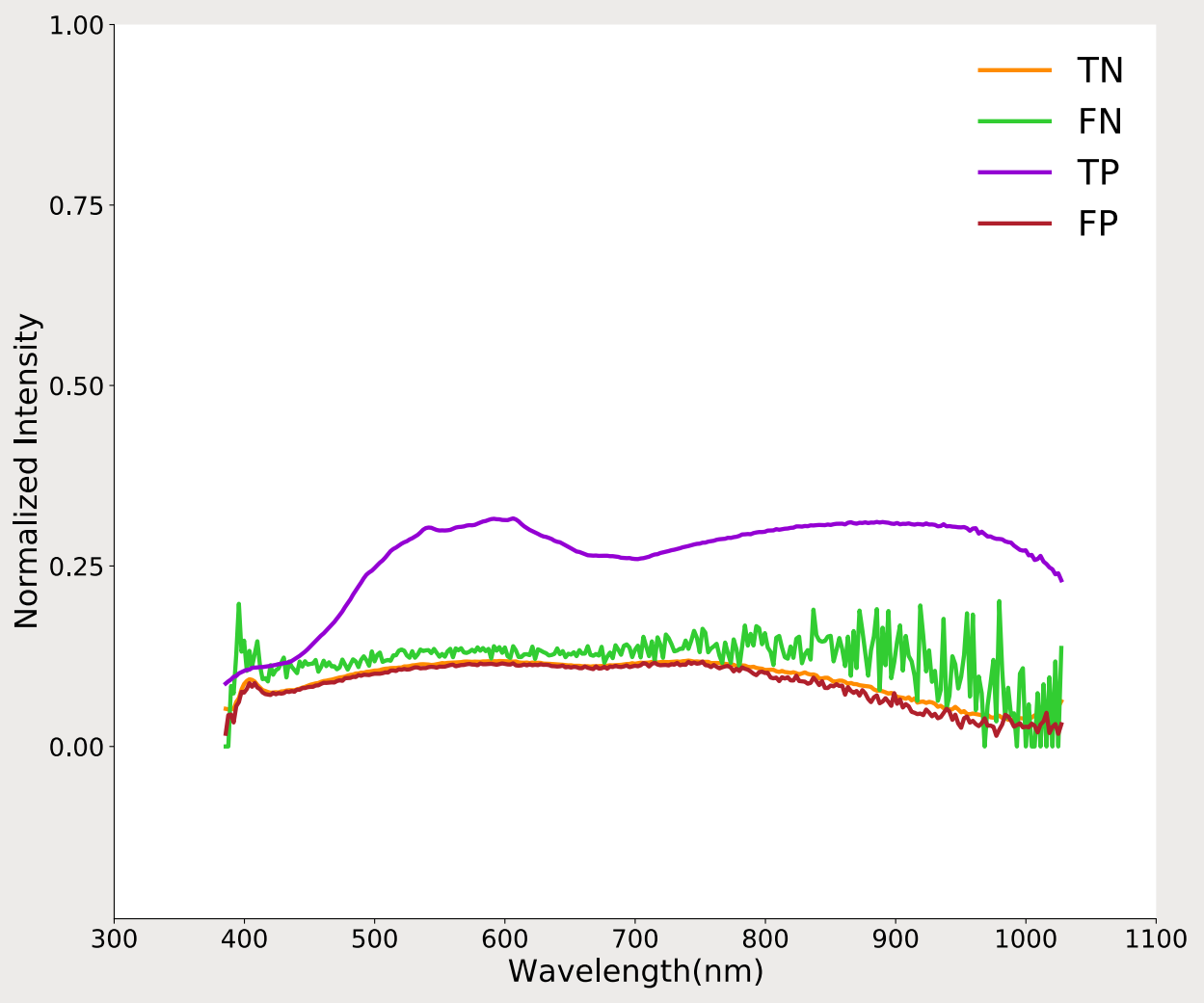


Figure 4.10. E Spectral profiles of TN, FN, TP, and FP pixels.

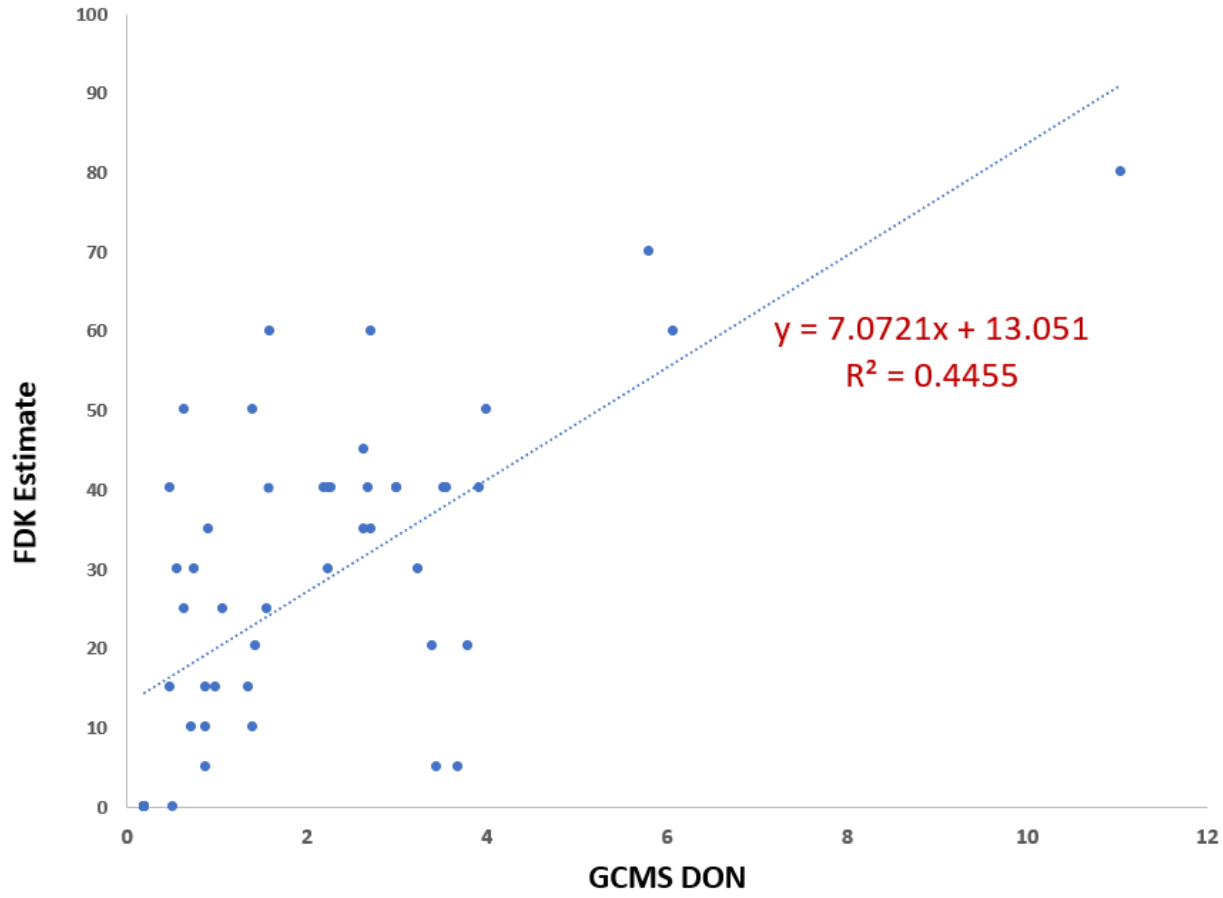


Figure 4.11 Correlation results between GC-MS DON content and FDK estimate.

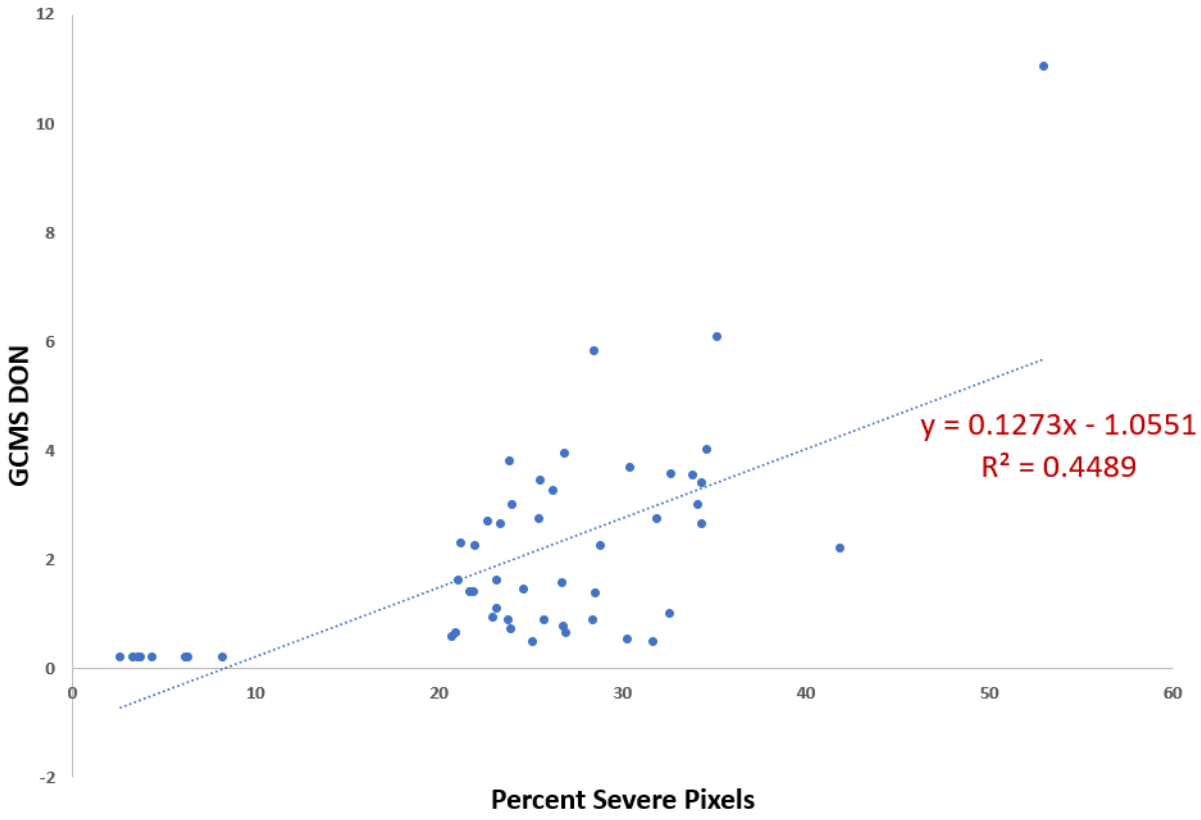


Figure 4.12 Correlation results between Percent Severe Pixels and GC-MS DON content.

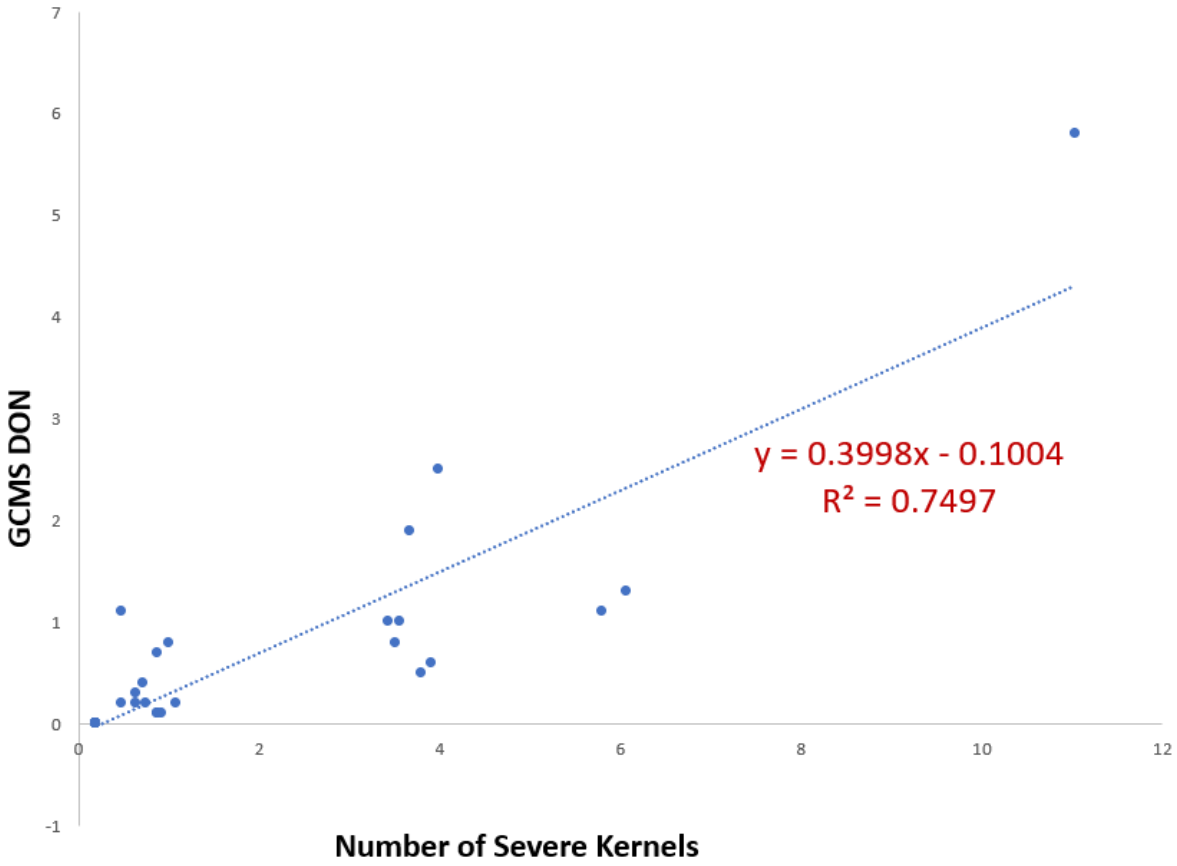


Figure 4.13 Correlation results between GC-MS DON and Number of Severe Kernels (with 70% threshold).

Supplementary Figures

Figure 4.S.1 Mask-R-CNN results (A. Ground Truth B. Predictions C. IOU of Ground Truth & Predictions) on original RGB images obtained from Wheat Kernels' HS Images.

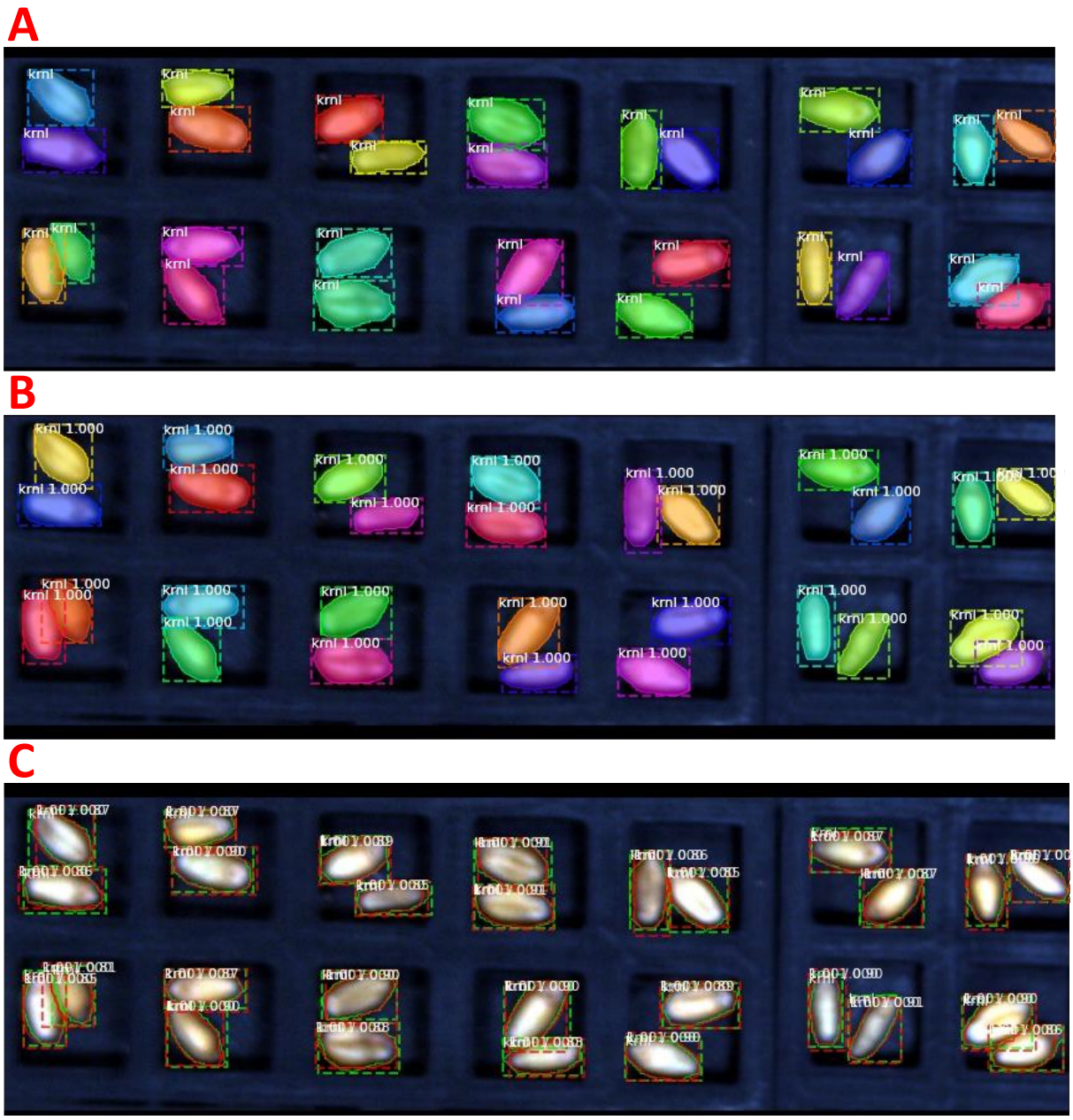


Figure 4.S.2. Mask-R-CNN results (A. Ground Truth B. Predictions C. IOU of Ground Truth & Predictions) on cropped RGB images obtained from Wheat Kernels' HS Images.

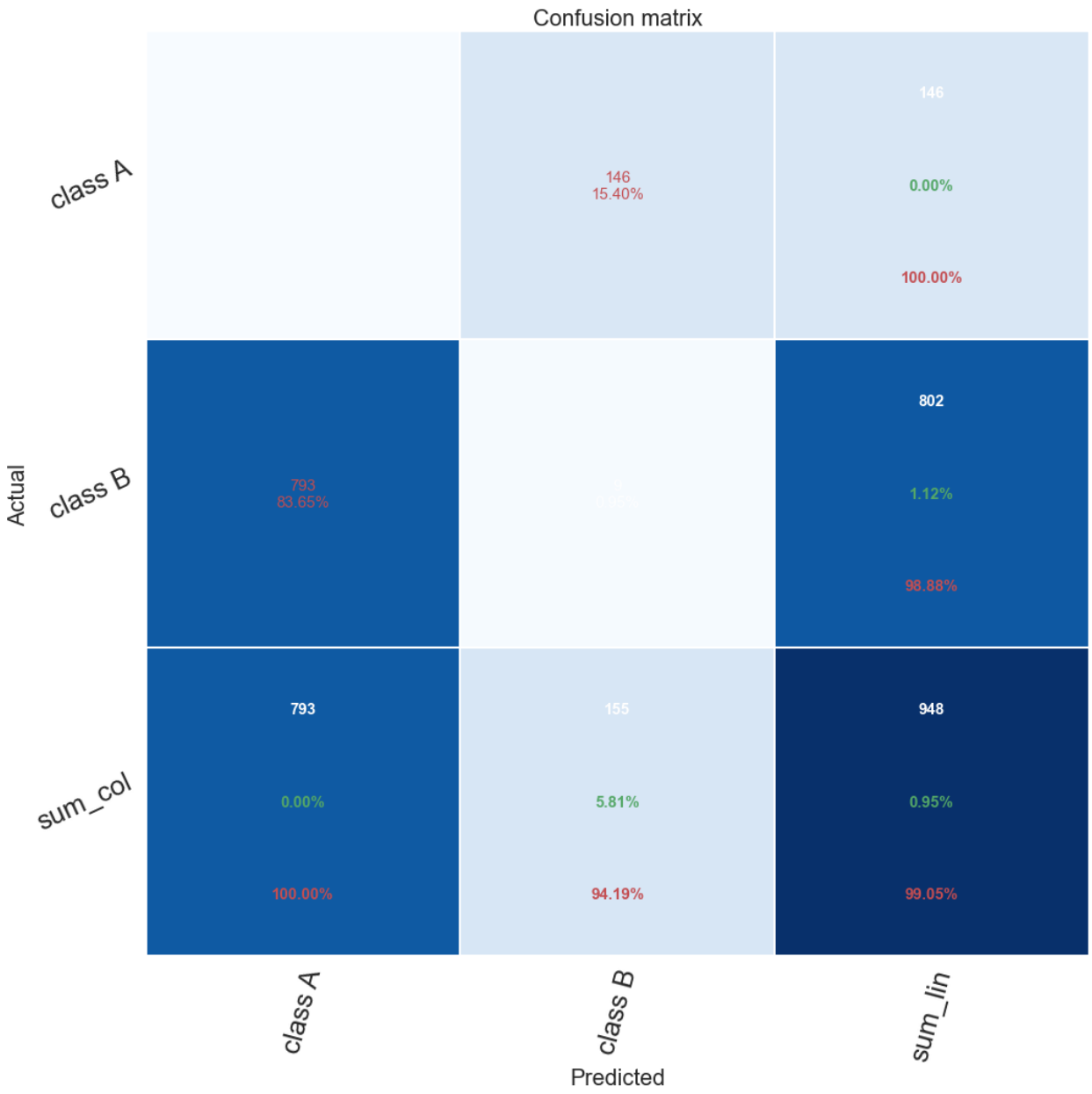


Figure 4.S.3. Confusion matrix of Mask-R-CNN results obtained from original RGB images obtained from Wheat Kernels' HS Images.

Here class A is background, which is being counted to cover the cases when the model miss (detect background instead of kernels or detect kernels instead of background) and class B is kernel, sum_col and sum_lin are the sum of columns and lines, respectively. The numbers in white are the

number of instances of backgrounds and kernels. The column at the far right contains the precision (green colored percentages) and false discovery rate (red colored percentages). The row at the bottom shows t recall (green colored percentages) and the false negative rate (red colored percentages). The diagonal cells correspond to the observations correctly classified while other cells correspond to incorrectly classified predictions.

There are 155 kernels in the evaluation dataset out of which 9 kernels are correctly classified. 802 kernels are classified out of which 793 are misclassified as background.

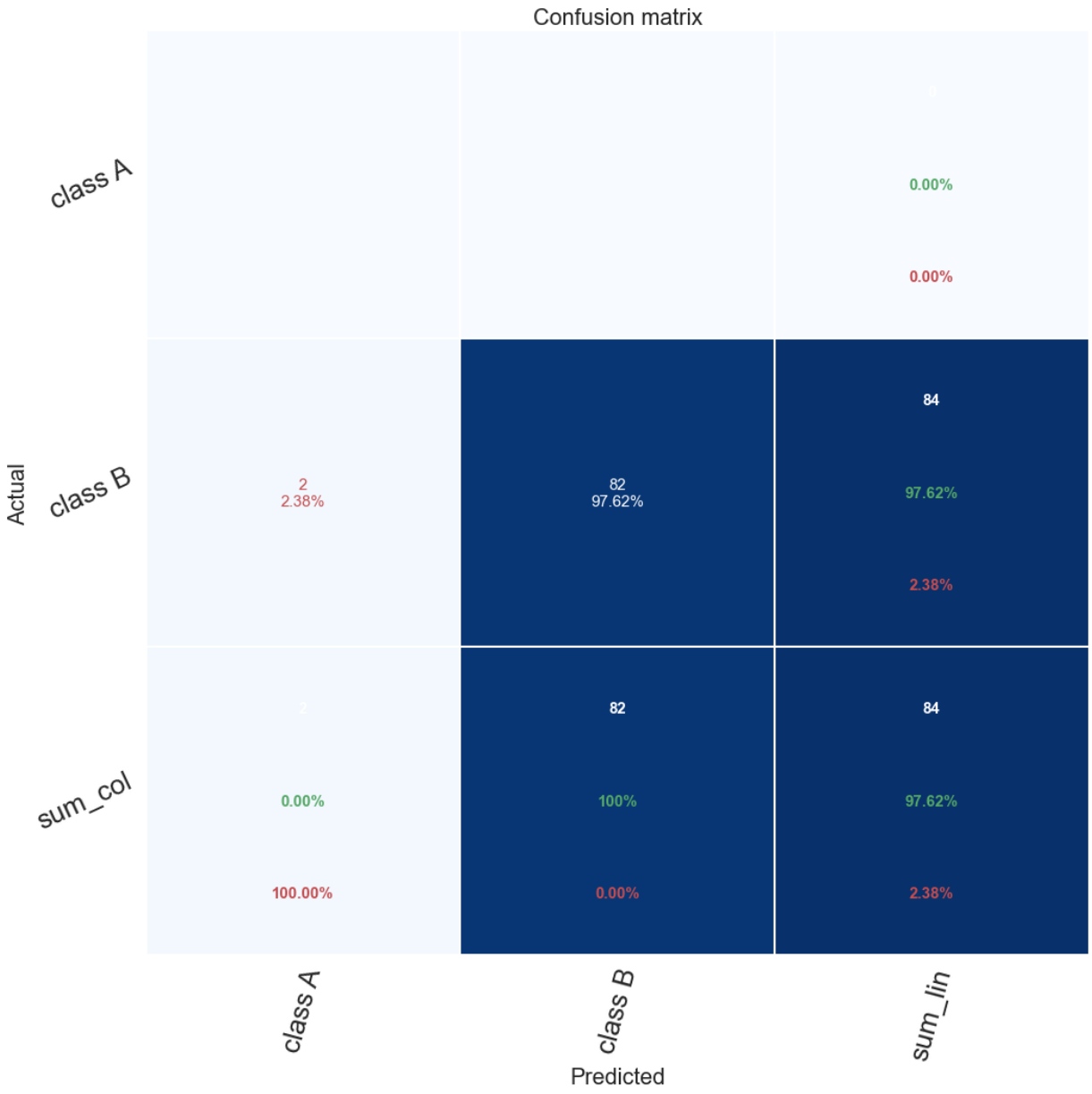


Figure 4.S.4. Confusion matrix of Mask-R-CNN results obtained from cropped RGB images obtained from Wheat Kernels' HS Images.

Here class A is background, which is being counted to cover the cases when the model miss (detect background instead of kernels or detect kernels instead of background) and class B is kernel,

sum_col and sum_lin are the sum of columns and lines, respectively. The numbers in white are the number of instances of backgrounds and kernels. The column at the far right contains the precision (green colored percentages) and false discovery rate (red colored percentages). The row at the bottom shows t recall (green colored percentages) and the false negative rate (red colored percentages). The diagonal cells correspond to the observations correctly classified while other cells correspond to incorrectly classified predictions.

There are 82 kernels in the evaluation dataset and all 82 kernels are correctly classified. 84 kernels are classified out of which only 2 are misclassified as background.

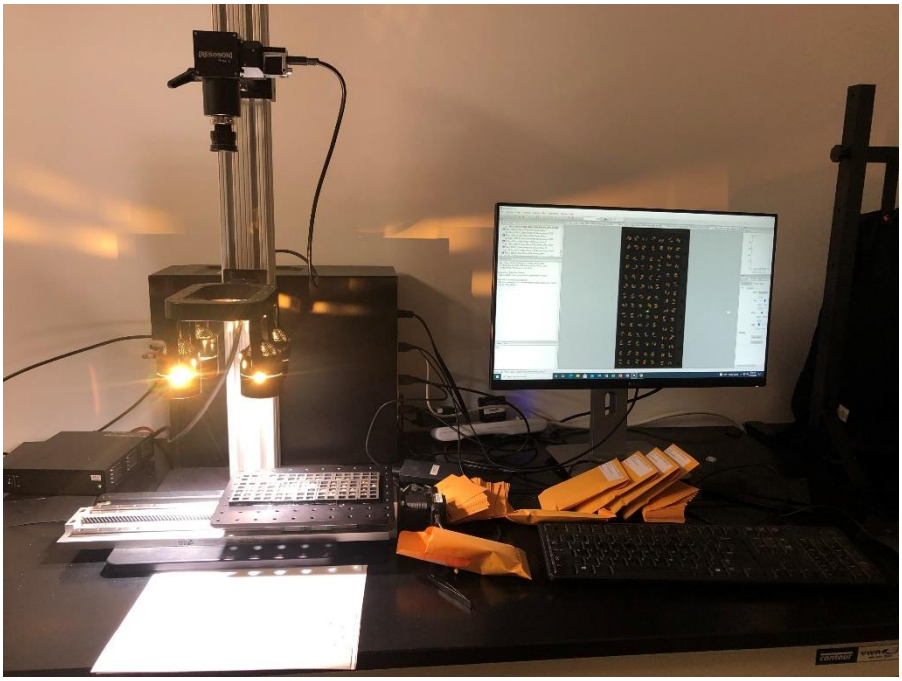


Figure 4.S. 5 Hyperspectral imaging platform at Li Lab.

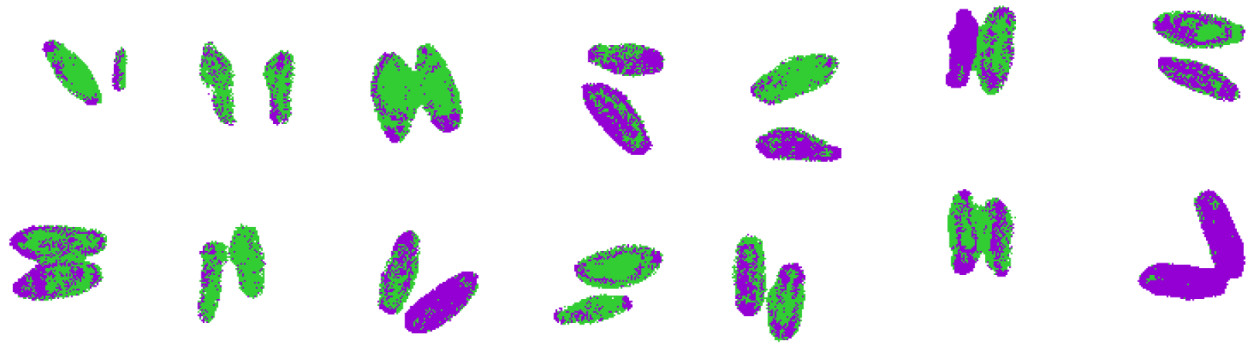


Figure 4.S. 6 Classified image of a severely infected samples, there most of the kernel pixels are infected (as shown in purple color) and a fraction are healthy (as shown in green color).

Figure 4.S. 7 Classified image of a healthy samples, there most of the kernel pixels are healthy (as shown in green color) and a fraction are infected (as shown in purple color).

CHAPTER V

CONCLUSION

The digital imaging technology and computer vision algorithms can be used to automatically characterize major traits of shoot architecture for edamame. Persistent homology can quantify the similarity and differences of branching patterns between these edamame varieties. We found and identified intriguing correlations between geometric traits and topological traits, suggesting combination of multiple topological features contribute to the overall pod numbers on a plant.

Using tools for spatial analysis and computer vision methods, we extracted traits related branching pattern, canopy cover, and pod location in edamame and performed genome-wide association study, to identify many single nucleotide polymorphisms (SNPs) that were associated with those traits. These SNPs could be used in marker-assisted selection to further develop edamame varieties that are better adapted to mechanical harvesting and higher yield. The specific SNPs located in the coding sequence of genes could be the key to understanding physiological mechanisms for better shoot architecture traits and better yield.

We re-projected the edamame pod images into different groupings based on maturity and disease using an interactive system called Andromeda and identified important visual features from the pixels highlighted by the model. We applied a spectroscopy-based machine learning method to identify the optimal harvest time of edamame. The machine learning method based on the pods' spectral reflectance had a high accuracy of 0.95 for classifying "early" and "late" samples and 0.87 for classifying "early" and "ready" samples. However, a low accuracy of 0.68 was obtained for classifying "late" and "ready" samples. Hence, we found that the machine learning method based on the pods' spectra reflectance can identify the optimal harvest time of edamame.

We showed that hyperspectral imaging combined with computer vision and machine learning methods can be used to quantify the levels of DON in wheat kernels in high through put fashion. G-Boost, an ensemble method, gives the highest (97%) accuracy to classify wheat kernels into different classes of severity levels based on FHB symptoms. Mask-R-CNN, method of object detection can segment out the wheat kernels from HS images, which then can be used to correlate the HS images with the DON content in small grains.