Evaluation of Wild and Cultivated Soybean Whole Genome Resequencing for Genotype Identity and Candidate Gene Discovery

Nicole Mihelich
Advised by Dr. Robert Stupar
9/25/2017

Introduction

Our world is facing unprecedented problems due to rapid population growth and climate change. Food security for the future will require not only more crop production, but crops that are adapted and resilient in a variety of ecological conditions which can be grown sustainably. This will require leveraging genetic information from entire germplasm collections all over the world to identify, introgress, and deploy beneficial alleles in future germplasm. Recent technological advancements in genomics and bioinformatics allow us to analyze such large and complex genetic data sets. Within the past decade, DNA sequencing costs have fallen and computational power has increased exponentially, allowing for many new plant genome sequencing and resequencing efforts. These rapidly evolving technologies are crucial in generating and exploiting big data to create superior crops to sustain our growing world.

Soybean (*Glycine max*) is one of the most valuable crops cultivated in Minnesota, the United States, and around the world, and thus, is a subject to interest for genomic research. Soybean is the largest source of protein for animal feed, and the second largest source of vegetable oil in the world. The narrow genetic base of soybean limits the identification of genetic loci responsible for traits of interest to breeders. In many crop species, wild relatives have been successfully used to reintroduce important crop traits, including fruit morphology (Tanksley *et al.,* 1996), pest and disease resistance (Steffenson *et al.,* 2007), and abiotic stress tolerance (Hajjar and Hodgkin, 2007; Nevo and Chen, 2010). While some effort has been made to exploit soybean landraces in modern breeding, the potentially rich reservoir of genes in *Glycine soja,* soybean's nearest wild relative, has been underutilized.

G. soja spans a wide range of environmental conditions across East Asia and exhibits local environmental adaptations. The USDA soybean germplasm collection including almost 20,000 accessions (1,168 of which are *G. soja*), has been recently genotyped with over 50,000 single nucleotide polymorphism (SNP) markers for use in association mapping (Song *et al.*, 2015). A variation of association mapping, in which marker information is coupled with environmental data from collection sites instead of phenotypic data, was recently used to detect genetic variants that are strongly associated to environmental conditions and potential abiotic stress tolerances (Anderson *et al.,* 2016). Ten candidate *G. soja* abiotic stress tolerance loci were identified in association with variables such as soil composition, precipitation, and temperature, and were deemed potentially useful targets for improving soybean abiotic stress tolerance.

To find the genetic basis of these potential abiotic stress tolerance loci, the genomic sequence must be scanned for polymorphisms linked with the associated marker. If the polymorphism appears to have a significant effect on a gene, it can be declared a candidate gene. Although the original genotypic data provided high marker density, this scanning of multiple loci requires more comprehensive genotypic information such as Whole Genome Sequencing (WGS). Once these specific polymorphisms and candidate genes are identified, they can be validated and precisely incorporated into elite germplasm using genome engineering approaches such as CRISPR/Cas9, which can circumvent linkage drag and the large time commitment of backcrossing. Zhou *et al*. (2015) published the most comprehensive soybean resequencing effort thus far with over 300 accessions of *G. soja* and *G. max*. This resequencing data can be publicly accessed to investigate the potential abiotic stress tolerance loci for candidate genes. However, discrepancies found in an initial per marker comparison of genotype accessions from the USA SNP chip and the Chinese WGS have raised concerns about genotype identity between the two data sets.

Objectives

The purpose of this study is to evaluate wild and cultivated soybean whole genome resequencing for candidate gene discovery and genotype identity. The first objective is to scan environmental association loci from Anderson *et al*. (2016) for candidate genes. The second objective is to compare genotype identity across international soybean collections.

Methods

Objective 1: 27 *G. soja* were determined to have climatic data at the collection site, as well as both USA genotyping SNP calls and WGS from the Chinese resequencing group. These 27 paired-end whole genome sequence reads were downloaded from the publicly accessible National Center for Biotechnology Information Sequence Read Archive. The reads from each of these accessions then underwent quality control to trim any bad quality and non-genomic sequences. The trimmed reads were then mapped to the reference genome 'Williams 82' to determine their position in the genome. From these positions, sequence variation between the *G. soja* genomes in the form of SNPs, insertions, or deletions were identified and filtered based on quality. The 10 top marker associations from the Anderson *et al*. (2016) were then investigated on a per marker basis for SNP consistency between the USA SNP and the Chinese WGS-derived SNP datasets. Markers that consistently matched across the 27 accessions will be scanned for causative sequence polymorphisms and candidate genes underlying the environmental association and potential abiotic stress tolerance.

Objective 2: Upon comparing the USA SNP calls with Chinese WGS-derived SNP calls at the 10 markers of interest, some markers were found to be inconsistent, even though they should be the same accession. This prompted a broader analysis of the two SNP data sets. All of the 302 *G. soja* and *G. max* whole genome sequences underwent the same pipeline as Objective 1 accessions, but instead of looking at the consistency of SNPs between the USA SNP and the Chinese WGS-derived SNP data on a per marker basis, all high-quality calls (up to about 50,000) will be evaluated across the genomes.

Summary

Scanning the regions around the markers significantly associated with potential abiotic stress tolerance adaptations can demonstrate how *G. soja* can be utilized for soybean improvement. However, this analysis is only possible when making the assumption that genotype identity for a given accession is the same across collections and data sets. Infrastructure issues such as these are crucial to prevent analyses rife with false positives or negatives. Through my research, I seek to investigate what genetic resources were left behind during domestication, and how in the future we can rediscover these resources, as well as work to curate our international resources for high quality research.

Anderson, J.E., T.J.Y. Kono, R.M. Stupar, M.B. Kantar, and P.L. Morrell. 2016. Environmental association analyses identify candidates for abiotic stress tolerance in glycine soja, the wild progenitor of cultivated soybeans. G3: Genes, Genomes, Genet. 6: 835–843.

Nevo, E., G.X. Chen. 2010. Drought and salt tolerances in wild relatives for wheat and barley improvement. Plant Cell Environ. 33:670-685

Song, Q., D.L. Hyten, G. Jia, C. V Quigley, E.W. Fickus, R.L. Nelson, and P.B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, Genomes, Genet. 5:1999–2006.

Steffenson BJ, Olivera P, Roy JK, Jin Y, Smith KP, Muehlbauer GJ. 2007. A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. Aus J Agric Res 58: 1–13.

Tanksley, S.D., Grandillo, S., Fulton, T.M., Zamir, D., Eshed, Y., Petiard, V., Lopez, J., and Beck-Bunn, T. 1996. Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative L. pimpinellifolium. Theor. Appl. Genet. 92, 213–224.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., Wan, W., Wang, X., Ding, Z., Gao, Y., Xiang, H., Zhu, B., Lee, S.H., Wang, W. and Tian, Z. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33, 408–414.