Environmental association analyses identify candidates for abiotic stress tolerance in *Glycine soja*, the wild progenitor of cultivated soybeans

Justin E. Anderson*, Thomas J. Y. Kono*, Robert M. Stupar*, Michael B. Kantar*,§, Peter L. Morrell*

*Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN,

§Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, BC

Abstract

Natural populations across a species range demonstrate population structure owing to neutral processes such as localized origins of mutations and migration limitations. Selection acts on a subset of loci, contributing to local adaptation. An understanding of the genetic basis of adaptation to local environmental conditions is a fundamental goal in basic biological research. When applied to crop wild relatives, this same research provides the opportunity to identify adaptive genetic variation that may be used to breed for crops better adapted to novel or changing environments. The present study explores the USDA germplasm collection, an *ex situ* conservation collection, of *Glycine soja,* the wild progenitor of *Glycine max* (soybean). The collection was genotyped at 32,416 SNPs to identify population structure and test for associations with bioclimatic and biophysical conditions variables. Candidate loci were detected that putatively contribute to adaptation to abiotic stresses. The identification of potentially adaptive variants in *ex situ* collection may permit a more targeted use of germplasm collections.

Methods

Public databases

USDA GRIN Soybean Germplasm Collection — www.ars-grin.gov

WorldClim (Global Climate Data) — www.worldclim.org

Soybase — www.Soybase.org

ISRIC (World Soils Database) — www.isric.org/

Data

533 *Glycine soja* lines
- Genotyped with SoySNP50K
- Latitude and Longitude
- 1 Km grid soils data: pH, % sand, % silt, % clay, bulk density, cation exchange capacity, and organic matter
- 1 Km grid climate data: temperature and precipitation yearly, quarterly, monthly and altitude

Analyses
- Population Structure
- Environmental Association (Genome wide association analysis using environmental data as phenotype) (Yoder et al., 2014)
- Spatial Ancestry Analysis (SPA)
- F_{ST} Outliers

Figure 1. Utilizing public databases to explore local adaptation in *Glycine soja*. Spatial Ancestry Analysis (SPA) was calculated according to Yang et al., 2012.

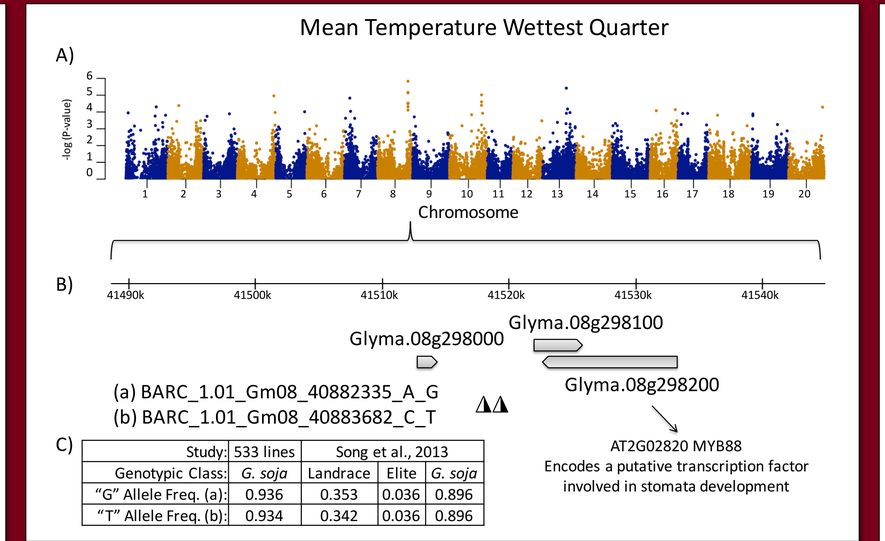
Results

Figure 2. Population distribution and clustering in *G. soja.* A) Results of STRUCTURE analysis and the sampling location of each accession. Colors correspond to the STRUCTURE assignment of each accession, Green: Mainland South; Blue: Mainland North; Red: Island. Assignment of samples into three genetic clusters generally accords with geography. B) Principle component analysis (PCA) of the genetic data. Samples are colored by STRUCTURE assignment with shapes for each country of origin. The first PC explained 5.1% of the variation and separated samples in an east to west gradient. The second PC explained 2.7% of the variation and corresponded more generally to a north to south separation.

Table 1. Diversity summary statistics within assigned clusters of *Glycine soja* sampled.

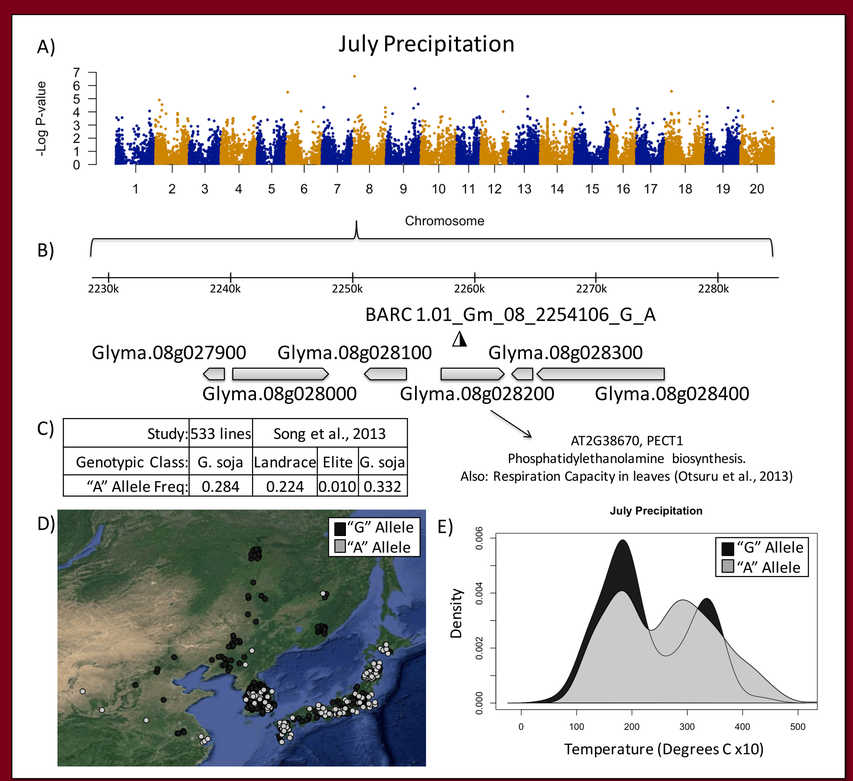
Population	Sample Size	Segregating sites	Private allelic richness	Percent pairwise difference
Island	216	31,698	0.025 (0.011)	0.340
Mainland South	275	32,360	0.009 (0.005)	0.337
Mainland North	42	23,797	0.001 (0.0001)	0.306
Mainland South + Island	492	32,416	0.25 (0.16)	0.349
Mainland North + Island	258	32,350	0.006 (0.002)	0.345
Mainland South + Mainland North	317	32,360	0.045 (0.029)	0.338

Env. Association



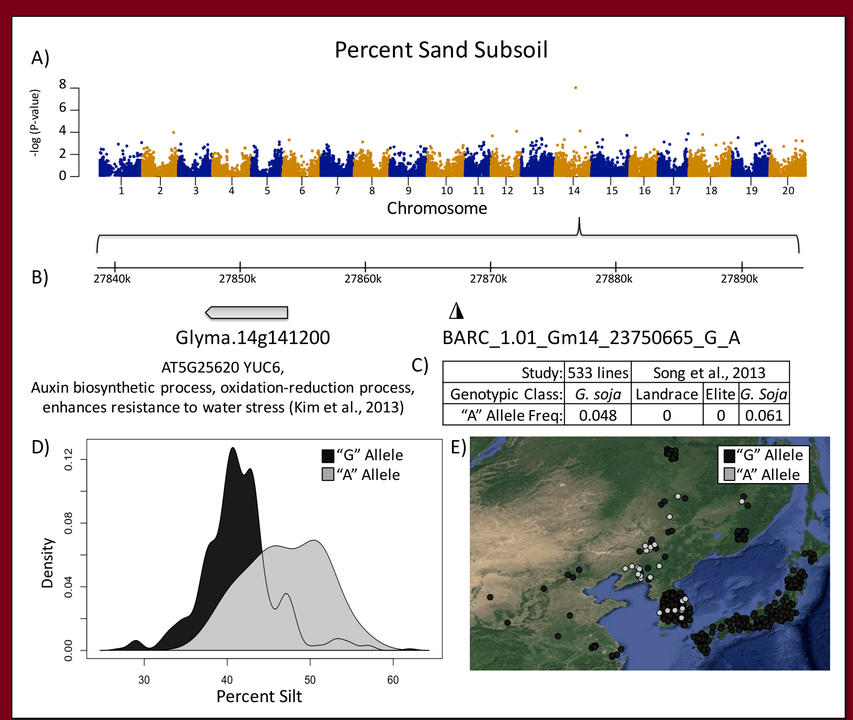
Study:	533 lines	Song et al., 2013		
Genotypic Class:	*G. soja*	Landrace	Elite	*G. soja*
"G" Allele Freq. (a):	0.936	0.353	0.036	0.896
"T" Allele Freq. (b):	0.934	0.342	0.036	0.896

Figure 3. Genome-wide associations with Mean Temperature Wettest Quarter. A) Manhattan plot of –log(*P*). B) Zoom in on 60 kb region around the two most significant markers, showing nearby genes. The Arabidopsis homolog for a near gene, Glyma.08g298200, is MYB88, a gene associated stomata development. C) The frequency of non-reference "G" and "T" alleles is high in *G. soja* and rare in a previous study of landrace and elite lines.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	G. soja	Landrace	Elite	G. soja
"A" Allele Freq:	0.284	0.224	0.010	0.332

Figure 4. Genome-wide association significant marker for July Precipitation and Precipitation Wettest Quarter. A) Manhattan plot of July Precipitation association results. B) Zoom in on 60 kb region around the significant marker BARC_1.01_Gm_08_2254106_G_A. The Arabidopsis homolog for the nearest gene, Glyma.08g028200, is AT2G38670, PECT1, involved in Phosphatidylethanolamine biosynthesis but also implicated in respiration capacity in leaves (Otsuru et al., 2013). C) The "A" allele is common in *G. soja* and landraces, but rare in elite lines (Song et al., 2013). D) Geographic location of individuals with the allele "A" (light gray) or reference allele "G" (dark gray) with jitter added to show overlapping samples. Individuals with missing genotyping data are not shown. E) Density plot of allele frequency distribution for July Precipitation. The reference allele "G" individuals are shaded in dark gray overlaid with the non-reference allele "A" individuals in light gray.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	*G. soja*	Landrace	Elite	*G. Soja*
"A" Allele Freq:	0.048	0	0	0.061

Figure 5. Genome-wide association results of percent sand and percent silt. A) Genome wide view of association results for Percent Sand Subsoil. B) Zoom in on 60 kb region around the most significant marker for topsoil and subsoil percent sand, and topsoil and subsoil percent silt. C) The "A" allele is rare in our sample and found to be rare or not present in a previous screen of soybean genotypic classes (Song et al., 2013). D) Density plot of allele frequency distribution for Percent Silt. The individuals with the "G" allele are shaded in dark gray overlaid with the "A" allele individuals in light gray. E) Geographic location of individuals with the "G" allele (Dark gray) or "A" allele (light gray) with jitter added to show overlapping samples.

SPA and F_{ST}

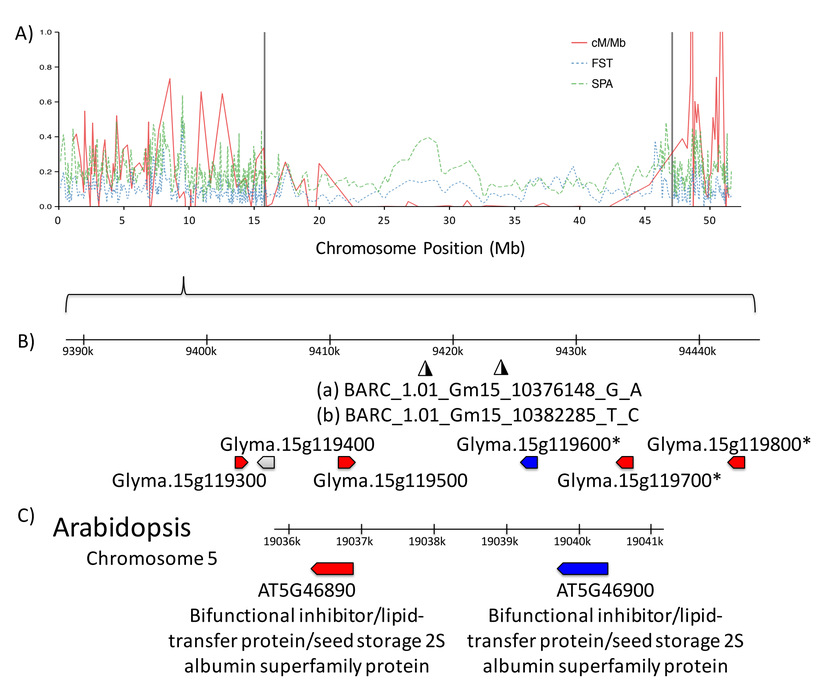


Figure 6. SPA, F_{ST}, and recombination rate in the *G .soja* genome. A) Sliding window means of these values plotted on chromosome 15. Recombination rate decreases dramatically through the pericentromeric region, denoted by the the vertical gray lines. B) Zoom in on 60 kb region around the two SNPs showing the highest SPA score, a region with notably low recombination rate and high F_{ST}. Three genes in this region (denoted with asterisks) were previously found to be duplicated or deleted in elite soybean lines (Anderson et al., 2014), and appear to be part of a Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein family. The Arabidopsis top hit for the genes denoted in red is AT5G46890 (C) and the Arabidopsis top hit for Glyma.15g119600, denoted in blue, is AT5G46900 (C). The implications of structural variation relating to F_{ST}, SPA hits, or recombination are not yet clear.

Conclusions

- Isolation by distance is the primary driver of population structure in *Glycine soja.*
- Environmental association readily detects loci putatively related to climatic adaptation. These loci often contain homologs already implicated in abiotic stress.
- Identifying loci associated with local adaptation in crop wild relatives has potential to address issues related to crop improvement; or issues likely to be exacerbated by a changing global climate

Implications

- Genotyped germplasm collections and large public databases present an opportunity for population genetics techniques exploring local adaptation and detecting putatively beneficial alleles.
- *Specific loci detected require further phenotypic validation but can readily be introgressed into elite germplasm for evaluation and pre-breeding.
- This method of targeted germplasm evaluation could prove useful in collaboration with recent initiatives to categorize and evaluate the world's germplasm collections (www.DivSeek.org, McCouch et al., 2013; Dempewolf et al., 2014).

Acknowledgements

This work was supported by the United Soybean Board, the Minnesota Varietal Development Fund, and the MnDRIVE 2014 Global Food Ventures Fellowship program in support of TJYK. This work was carried out in part using hardware and software provided by the University of Minnesota Supercomputing Institute.

References

Anderson, J. E., M. B. Kantar, T. Y. Kono, F. Fu, A. O. Stec *et al.*, 2014 A roadmap for functional structural variants in the soybean genome. G3 4: 1307–1318.

Dempewolf, H., R. J. Eastwood, L. Guarino, C. K. Khoury, J. V. Müller *et al.*, 2014 Adapting Agriculture to Climate Change: A Global Initiative to Collect, Conserve, and Use Crop Wild Relatives. Agroecol. Sustain. Food Syst. 38: 369–377.

Kim, J. I., D. Baek, H. C. Park, H. J. Chun, D.-H. Oh *et al.*, 2013 Overexpression of Arabidopsis YUCCA6 in potato results in high-auxin developmental phenotypes and enhanced resistance to water deficit. Mol. Plant 6: 337–49.

McCouch, S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting *et al.*, 2013 Agriculture: Feeding the future. Nature 499: 23–4.

Otsuru, M., Y. Yu, J. Mizoi, M. Kawamoto-Fujioka, J. Wang *et al.*, 2013 Mitochondrial phosphatidylethanolamine level modulates Cyt c oxidase activity to maintain respiration capacity in Arabidopsis thaliana rosette leaves. Plant Cell Physiol. 54: 1612–9.

Song, Q., D. L. Hyten, G. Jia, C. V Quigley, E. W. Fickus *et al.*, 2013 Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8: e54985.

Song, Q., D. L. Hyten, G. Jia, C. V Quigley, E. W. Fickus *et al.*, 2015 Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 5: 1999–2006.

Yang, W.-Y., J. Novembre, E. Eskin, and E. Halperin, 2012 A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. 44: 725–731.

Yoder, J. B., J. Stanton-Geddes, P. Zhou, R. Briskine, N. D. Young *et al.*, 2014 Genomic signature of adaptation to climate in Medicago truncatula. Genetics 196: 1263–75.