# Into the wild: The soybean genome meets its undomesticated relative

**Robert M. Stupar**[1]
*Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108*

Soybean (*Glycine max*) is one of the most widely grown crop species in the world. One of the major agricultural challenges of the 21st century will be to increase the yield of soybean and other major crop species to feed a growing population on a finite amount of farmland. Soybean breeding and improvement is hindered by a narrow domesticated germplasm relative to other crop species (1). Despite its importance, many outstanding questions remain regarding important aspects of soybean germplasm, including the extent of genomic variation within the domesticated germplasm and among domesticated and wild relatives. *Glycine soja* is the closest extant wild relative of soybean and is generally considered to be the undomesticated progenitor of the domesticated soybean. *G. max* and *G. soja* are phenotypically disparate in many ways, but they readily cross with one another and give rise to fertile hybrids, thus making *G. soja* a promising source of novel genes and alleles for soybean breeding and improvement.

The genome sequence of domesticated soybean was published earlier this year (2), bringing in a new era for soybean functional and comparative genomics. Comparative sequencing of soybean domesticates and wild relatives will substantially increase our understanding of the limitations of the domesticated germplasm and the potential to use wild relatives for crop breeding and improvement. The PNAS report by Kim et al. (3) focuses on the resequencing of wild soybean *G. soja* (accession no. IT182932) and the subsequent comparative genomic analysis with the reference *G. max* genome (2). Kim et al. (3) cataloged a wide range of nucleotide and structural variations between wild and domesticated soybean. A summary of the types and frequencies of the different gene variation classes identified in the Kim et al. analysis (3) are shown in Fig. 1. Nucleotide variants, such as base substitutions and small insertions and deletions, occurred at a frequency of 0.31% across the *G. max* and *G. soja* genomes. These types of alterations may affect the function and/or protein structure of more than 10,000 putative protein-encoding genes. Furthermore, many structural genomic differences were also apparent between *G. max* and *G. soja*, such as large insertions, deletions, and inversions. Deletions in the *G. soja* genome ranging from 100 bp to 100 kb may explain much of this structural variation. In total, approximately 1,000 genes were identified within regions of structural variation between *G. max* and *G. soja*. Although it is difficult to estimate what portion of these genes may in fact be located within respective sequence gaps, these results are corroborated by the recent resequencing of another *G. soja* accession (W05), which exhibited a similar number of gene content variants compared with *G. max* (4).

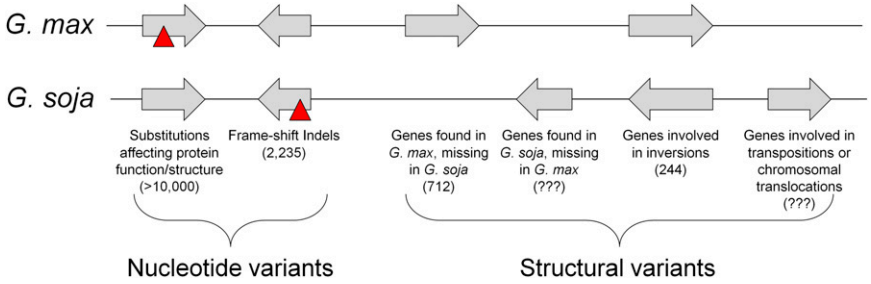


**Fig. 1.** Nucleotide and structural variation identified between domesticated soybean (*G. max*) and wild soybean (*G. soja*) in gene coding regions. The arrows indicate gene positions along a hypothetical chromosomal region. The numbers of genes exhibiting variation between *G. max* and *G. soja* for each type of variation are shown in parentheses. Nucleotide variants that influence protein function or structure, such as base substitutions and small frame-shift mutations, are shown on the left (red triangles represent sites of nucleotide differences). Genomic structural variations, such as inversions, deletions, insertions, and translocations, are shown on the right. The number of genes that are found in *G. soja* and missing in *G. max* is ambiguous because of sequence gaps in the reference sequence; however, several examples of *G. soja*-specific genes were validated. The methodology used in this study (3) was unable to resolve chromosomal translocations, so the number of genes in this category remains unknown.

The exact timeline of soybean domestication remains a matter of dispute. Most estimates approximate that domestication occurred somewhere between 3,100 and 9,000 y ago (5, 6). Kim et al. (3) used their comparative sequence data to estimate the time of divergence between *G. max* and the *G. soja* accession they sequenced. Surprisingly, they estimated that the split occurred approximately 270,000 y ago, substantially predating soybean domestication. The authors concede that this may be an overestimate, as human selection may have increased the frequency of variation in seemingly neutral genes. However, the discrepancy between the timing of the *G. max*/*G. soja* split and the timing of domestication suggests that the domestication process may have been more complicated than has been thought, perhaps occurring in a lineage that split long ago from the *G. soja* accession sequenced by Kim et al. (3).

Furthermore, the comparative sequence data may be significant for understanding the genetic mechanisms of soybean domestication. The domestication syndrome that distinguishes *G. max* and *G. soja* is vast, including differences in plant architecture, flowering time, pod dehiscence, seed size, and other characteristics. Quantitative trait loci (QTLs) have been genetically mapped for several soybean domestication traits (7), but only one gene associated with domestication has been characterized to date (8). The *G. soja* sequence will be an important resource for identifying candidate domestication genes. This type of analysis will be particularly powerful when combined with population-level comparative sequencing, allowing for the identification of regions of conserved divergence between the domesticated and wild accessions (4).

Perhaps the most important use of the *G. soja* sequence will be to identify genes and alleles from the wild germplasm that may have potential applications for use in soybean cultivar improvement. Several major crop species, including tomato, barley, and wheat, have made substantial use of their wild relatives to expand their gene pools and incorporate novel traits, particularly pest and disease

Author contributions: R.M.S. wrote the paper.
The author declares no conflict of interest.
See companion article on page 22032.
[1]E-mail: rstupar@umn.edu.

resistance (9–11). Soybean breeding has had less success at incorporating wild introgressions into elite cultivars for a variety of reasons (5). However, the influence of wild introgressions on the soybean germplasm may be underestimated, as recent studies indicated that *G. soja* introgressions are found in some soybean accessions and breeding lines (4, 12).

The allelic diversity in *G. soja* is greater than that of soybean (4). The sequence comparisons by Kim et al. (3) and Lam et al. (4) have identified a subset of genes and alleles in *G. soja* that are not found in the soybean reference sequence. Efforts have been and are being made to assess the impact of wild *G. soja* genetic introgressions on soybean phenotypes; in fact, the soybean breeding community has identified several QTLs for which the *G. soja* locus is more favorable than the *G. max* locus for specific qualitative and quantitative traits of interest (13–18). Linking the molecular sequence variation to the phenotypic variation between *G. max* and *G. soja* is clearly the next challenge. From a practical standpoint, the *G. max*/*G. soja* nucleotide and structural diversity revealed by Kim et al. (3) will be useful as a reference for the genetic mapping of *G. soja* introgressions in soybean populations and the identification of novel candidate genes and alleles in the *G. soja* sequence that may underlie QTLs conferring superior phenotypes. The identification of structural variants may also be useful for understanding why certain regions of the *G. soja* genome may be recalcitrant to stable introgression.

At present, it is clear that a revolution in the comparative sequencing of major crop species, like maize (19), rice (20), and soybean (3, 4), is well under way. The number of sequenced accessions available to the public will continue to grow rapidly, offering new opportunities and challenges for breeders, population geneticists, and molecular biologists. It is easy to recognize the potential of these resources from a research standpoint, but the real challenge may be in translating this new knowledge into advances in crop productivity and stability.


1. Hyten DL, et al. (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci USA* 103:16666–16671.
2. Schmutz J, et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183.
3. Kim MY, et al. (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc Natl Acad Sci USA* 107:22032–22037.
4. Lam HM, et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet*, in press.
5. Carter TE, Jr, Nelson R, Sneller CH, Cui Z (2004) Genetic diversity in soybean. *Soybeans: Improvement, Production and Uses*, eds Boerma HR, Specht JE (Am Soc Agronomy, Madison, WI), pp 303–416.
6. Hymowitz T, Shurtleff WR (2005) Debunking soybean myths and legends in the historical and popular literature. *Crop Sci* 45:473–476.
7. Liu B, et al. (2007) QTL mapping of domestication-related traits in soybean (Glycine max). *Ann Bot (Lond)* 100:1027–1038.
8. Tian Z, et al. (2010) Artificial selection for determinate growth habit in soybean. *Proc Natl Acad Sci USA* 107: 8563–8568.
9. Friebe B, Jiang J, Raupp WJ, McIntosh RA, Gill BS (1996) Characterization of wheat-alien translocations conferring resistance to diseases and pests: Current status. *Euphytica* 91:59–87.
10. Bai Y, Lindhout P (2007) Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? *Ann Bot (Lond)* 100:1085–1094.
11. Steffenson BJ, et al. (2007) A walk on the wild side: Mining wild wheat and barley collections for rust resistance genes. *Aust J Agric Res* 58:532–544.
12. Li YH, et al. (2010) Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytol* 188:242–253.
13. Concibido VC, et al. (2003) Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. *Theor Appl Genet* 106:575–582.
14. Kabelka EA, Carlson SR, Diers BW (2006) Glycine soja PI 468916 SCN resistance loci's associated effects on soybean seed yield and other agronomic traits. *Crop Sci* 46:622–629.
15. Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Sci* 40:1438–1444.
16. Nichols DM, Glover KD, Carlson SR, Specht JE, Diers BW (2006) Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. *Crop Sci* 46:834–839.
17. Lee JD, Shannon JG, Vuong TD, Nguyen HT (2009) Inheritance of salt tolerance in wild soybean (Glycine soja Sieb. and Zucc.) accession PI483463. *J Hered* 100: 798–801.
18. Li DD, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for yield and yield components associated with Glycine soja alleles. *Crop Sci* 48:571–581.
19. Lai J, et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42:1027–1030.
20. Huang X, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–967.