Evolution of the soybean genome and comparison to the *Phaseolus vulgaris* genome

Scott A. Jackson¹, Jer-Young Lin, Jessica Schlueter, Navdeep Gill, Jeffery Doyle, Christian Hans

¹ Department of Agronomy, Purdue University, West Lafayette, IN Email sjackson@purdue.edu

Abstract

Soybean has undergone at least two genome-wide duplication events as evidenced by chromosome number, RFLP hybridization and analyses of duplicated ESTs. We provide cytogenetic evidence that the most recent event may have been an allopolyploid event given that two centromere sequences mark subsets of chromosomes. Previous work has shown the existence of high levels of synteny between duplicated chromosome segments; we show that chromosomes carrying high levels of shared duplicated segments also carry differing centromeric repeats that presumably correspond to the allopolyploid progenitors. In addition to this work, we have also created a physical map of *Phaseolus vulgaris* (common bean) using fingerprinted BACs. These BAC have been end sequenced (BESs) and anchored to genetic maps. Hybridization markers from soybean have also been placed on the *Phaseolus* physical map to provide a network of linkages between the soybean and *Phaseolus* genomes. We will discuss the use of *Phaseolus* to understand the evolution of duplicated region in soybean and how this map will be useful for the agronomic improvement of common bean and other related Phaseoloid legumes.

Media Summary

Exploring genome structure in soybean and common bean will help lead to better genetic tools for improvement and an understanding of the evolution and domestication of these important crops.

Key Words

Genomics, common bean, soybean, evolution

Introduction

Soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*) are both legumes that diverged from a common ancestor approximately 19 MYA (). Both are important food and feed crops worldwide. The soybean genome is highly duplicated and approximately twice the size of the *Phaseolus* genome; thus, we anticipated that the 'diploid' *Phaseolus* genome could be used to help decipher duplication events in the soybean genome that occurred post-divergence. There are many lines of evidence indicating that soybean is paleopolyploid (or anciently duplicated) including RFLP mapping (Shoemaker et al., 1996), EST analyses (Blanc and Wolfe, 2004) and recent sequencing efforts (Schlueter et
al., 2007). Many lines of evidence also support the lack of significant duplication in *Phaseolus* including chromosome number and genetic mapping.

Analysis of genome duplication events are difficult as they often are rearranged, due to diploidization, over time and it can be challenging to find ‘blocks’ of duplicated genes. This is not the case in soybean where there is extensive colinearity between duplicated segments of the genome; however, evolutionary analysis is complicated even in soybean as it is difficult to determine ancestral states through the comparison of two duplicated segments. Thus, the inclusion of a relative that does not share the duplications is essential. This is the basis for our detailed analysis of the *Phaseolus* genome in addition to the importance of developing genomic tools for common bean crop improvement.

**Methods**

BACs were selected for DNA sequencing by library hybridization screening and BLAST searches of BAC end sequence (BES) databases. BESs were aligned to soybean genome scaffolds via BLAT and graphic examples were generated via SyMap (Soderlund et al., 2006).

**Results**

*Library construction, sequence tag connectors and physical map for Phaseolus and Glycine species*

BAC libraries and physical maps for both soybean and *Phaseolus* have been constructed (Schlueter et al., 2008) and are publicly available ([http://www.soybase.org](http://www.soybase.org); [http://phaseolus.genomics.purdue.edu](http://phaseolus.genomics.purdue.edu)). The physical maps were constructed by fingerprinting of BACs followed by assembly of BAC-based contigs using FPC software (Soderlund et al., 1997). BAC-end sequences were also generated for all BAC libraries used to generate FPC maps; these have proven useful for the whole genome shotgun sequencing of soybean and, in the case of *Phaseolus*, for anchoring the physical map to the soybean genome sequence via alignment of BESs.

Random shotgun sequences were generated for two *Glycine tomentella* species (a diploid and an allotetraploid) and *Glycine syndetika* to begin to determine how the genomes have diverged and how useful these soybean relatives might be for functional genomics for soybean. Small insert (~3-4 kb) libraries were made for all three species and one 384 plate was sequenced from both ends. The resulting sequences were analyzed for repeat content and then aligned to the soybean genome scaffolds. Using a legume specific repeat database (Gill et al., unpublished results), the sequences were repeat masked and aligned using BLAST; 64.7% of 693 reads for *G. tomentella* (T2, tetraploid), 65.3% of 694 reads for *G. syndetika* and 71.3% of 638 reads from *G. tomentella* (diploid) had good single alignments to the *G. max* genome. 13% of the *P. vulgaris* BESs have good single alignments to *G. max* and can be used to align *Phaseolus* FPC contigs.
Alignment of Phaseolus BESs and Glycine relatives to the soybean genome scaffolds

BESs from the *Phaseolus* BACs were aligned to the soybean draft genome sequence; the resulting alignments were then used to align FPC contigs to the soybean genome. The result of this alignment is that 45% of the 950 MB soybean draft sequence is covered physically by FPC contigs from *Phaseolus*. Of the 89,017 BESs generated for *Phaseolus*, 58,823 were nonrepetitive and generated 20,303 hits in 9,045 blocks with 55% aligning to at most two positions in the soybean genome; this is complicated by the fact that genes in soybean almost always exist in multiple copies due to the duplication events. Figure 1 is a screen-shot of the Symap generated alignment between a *Phaseolus* contig and the soybean genome. Alignment of the shotgun sequences from both *G. tomentellus* and *G. syndetika* revealed that a number of the sequences aligned to single best hits on the soybean genome scaffolds. The relative ease of aligning these sequences indicates that should BAC libraries and BESs be developed for these species that they would align relatively easily and be useful for mapping and introgressing traits between *Glycine* species.

Comparative sequencing of an orthologous region between soybean and Phaseolus

BACs surrounding soybean RFLP marker A711 were sequenced from two genomic locations in soybean and an orthologous region of *Phaseolus*. BACs were annotated and then compared to each other to assess levels of conservation and rearrangement (evolution) amongst all three loci. Surprisingly, given the ~19 MY of divergence between *Phaseolus* and *G. max*, the conservation of genes and gene order between either of the duplicated segments in soybean and *Phaseolus* was quite high. Moreover, *Phaseolus* can be used to help define gene structures and boundaries as there is little conservation of sequences outside of genes.

Conclusion

That the soybean genome is highly duplicated is no surprise, but the level of conservation between duplicated regions is surprisingly high, though there is evidence of local rearrangements (Schlueter et al., 2007) that might be consistent with post-polyplodization diploidization (restructuring of chromosomes). We have begun to investigate the nature of genome dynamics within a smaller set of legumes including soybean and its close relatives, such as the Glycine perennials, and other more distantly related legumes such as common bean (*Phaseolus*). The date from these analyses, thus far, indicates that there is considerable conservation of synteny and microcolinearity between the *Glycines* and *Phaseolus*. This observation bodes well for the use of soybean as a reference genome for the Phaseoloid legumes as well as for the transfer of information from other legumes back to soybean.

References


Figure 1. Screen-shot from Symap generated alignment of Phaseolus BAC-end sequences (purple lines), overgo based markers (green lines) and their representative BAC contigs (blue vertical lines) and their relationship to a portion of the soybean genome sequence (large vertical block on the right).