Supplemental Methods

Details on exome resequencing PAV genes

We analyzed exome resequencing data from the Archer, Minsoy, Noir1 and Wm82 individuals to determine the gene content variation among the cultivars. Presence-absence variants (PAV) were determined to be genes with high read counts (>30) in at least one cultivar and zero read counts in at least one other cultivar. The locations of the PAV along each chromosome are shown as red spots in Figure S1.

In total, 133 genes make up the high confidence list of presence-absence variants (PAV). The gene models and presence-absence profiles of these 133 genes are shown in Table S2 and the distribution of “absent” genes among the four cultivars is shown in Figure S5. Wm82 exhibited 14 absent genes, which may seem counterintuitive because the gene model list is derived from the reference Williams 82 sequence (Schmutz et al., 2010). However, this finding is not surprising because it has been previously established that the individual Wm82-ISU-01 has some genomic regions that are polymorphic to the reference Williams 82 sequence (Haun et al., 2011); 13 of the 14 Wm82 absent genes are located within such known regions. The frequency of “absent” genes from the Archer, Minsoy and Noir1 cultivars is similar (ranging from 54 to 79). Archer has a slightly lower number of “absent” genes than was found in Minsoy and Noir1, possibly because Williams 82 was the *Phytophthora* root rot resistance donor (Rps_1^k) in the Archer pedigree. This may account for the lack of structural variation between Wm82 and Archer at the end of chromosome 3 and likely elsewhere.

The 133 PAV genes identified represents a high confidence list, but almost certainly underestimates the number of genes that have full or partial gene content variation among the tested cultivars. There are several factors that could contribute to an underestimate. For instance, many genes did not meet the minimum requirement of 30 read counts. Also, the exon capture reads were required have a single unique match within the reference gene models, which may reduce the number of reads that map to moderately or highly duplicated gene classes. Additionally, the list of 133 genes only includes the gene content variants that were entirely absent across all exons. However, there were an additional 215 gene models that exhibited exon-specific content variation, in which the mapped reads for a cultivar indicated the presence of some of the exons but the absence of at least one exon relative to the other cultivars (data not shown). Given these factors, we estimate that the true rate of gene content and/or exon PAV among the cultivars is much greater than our high confidence list.

LITERATURE CITED

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP, et al (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol **155**: 645–655

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183