B-109

Characterizing and comparing international germplasm collections: finding unique gems and (useful) redundancy

Anne Brown*, Department of Agronomy, USDA-ARS, Iowa, USA
Nathan Weeks, USDA-ARS, Iowa, USA
David Grant, USDA-ARS, Iowa, USA
Steven Cannon, USDA-ARS, Iowa, USA

There are over 230,000 soybean accessions in germplasm repositories worldwide. Identifying essential genetic material among these many accessions is difficult. High throughput genotyping costs have dropped sufficiently to enable dense genotyping of large germplasm collections.

Nevertheless, large challenges remain due to data volumes and complexities. Comparisons are complicated, for example, by lack of common markers among data sets, differences in accession names, and by incompatible data formats. We are developing methods to merge and integrate genotyping data across projects. Essentially, the approach is to identify common and distinct genetic regions among genotyped collections, imputing missing data and determining soybean lineage/ancestry where possible. Using genotyping datasets from published U.S., Chinese, and Korean soybean experiments we identified common loci between the datasets. We integrated data from U.S GRIN accessions genotyped with the SoySNP50k chip and several other resequencing projects from international germplasms. Among the SNP arrays, we found that overlap of SNP positions is as high as 85% between the arrays. We merged VCF files based on common SNP positions, and determined overlap between the datasets on the basis of percent similarity. We find some groups of accessions that are identical genotypically, and some accessions common between experiments that appear to be different genotypically. We also find that measurements of similarity are quite dependent on methods and parameter choices.