UPOV

International Union for the Protection of New Varieties of Plants

E

Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular

BMT/16/26

Sixteenth Session
La Rochelle, France, November 7 to 10, 2017

Original: English
Date: October 30, 2017

DETERMINING THE PARAMETERS TO CHARACTERIZE SOYBEAN VARIETIES USING SINGLE NUCLEOTIDE POLYMORPHISMS

Document prepared by experts from Seed Association of the Americas (SAA)

Disclaimer: this document does not represent UPOV policies or guidance

INTRODUCTION

1. UPOV uses morphologically and physiologically expressed characteristics to evaluate the Distinctness, Uniformity, and Stability (DUS) criteria that are required to be met for the grant of Plant Variety Protection (PVP). With the rapid evolution and increase in cost effective use of molecular markers, these data are increasingly playing a role in identifying varieties both post their grant of protection, and in the granting of protection as surrogates of morphological characteristics. In addition, molecular markers are also being used to help manage reference collections that have become very difficult to use in some species due to the resources needed to accommodate the increasing numbers of varieties that make-up those collections (UPOV, 2011).

2. The BMT was specifically instituted to "study DNA profiling in connection with plant breeders' rights and to coordinate the development and harmonization of DNA analysis in the UPOV member States" (UPOV, 1993a). Key aspects of DNA analysis include the number of markers, the degree of map coverage, and the variety sampling method that is employed. Furthermore, as marker systems have evolved during the past 30 years to comprise from tens to thousands of loci, and especially as individual base pairs can now be routinely assayed, then inevitably many more markers are available for use in variety characterization compared to the number of morphological characteristics that are, or could be used in the DUS process. It is important to understand that very large increases in the numbers of markers used to characterize varieties also increases the sensitivity to detect residual heterozygosity and the presence of marker ideotypes, including in varieties that are, and have been considered acceptably distinct, uniform, stable, and thus protectable according to PVP. For example, Fasoula and Boerma (2007) reported the derivation of sub-lines by selection within publicly available soybean varieties. However, to maintain the level of protection afforded by PVP, it is a critical that distinctness be defined in relation to residual heterozygosity precisely to prevent the independent protection of sub-lines by a second breeder without any further cycles of crossing and selection. Understanding residual heterozygosity is a critical aspect when assessing the utility of markers to support declaration of Distinctness (D) and must also be considered to establish Uniformity (U) and Stability (S).

3. Previous experience examining varieties of the same species using different marker systems indicated that distance measurements between varieties are dependent upon the type of marker system that is used (Smith et al., 1997). We therefore chose to use Single Nucleotide Polymorphisms (SNPs) because these are the most numerous markers, provide the greatest degree of genomic coverage, provide a high degree of repeatability, and are rapidly becoming widespread and regularly used in soybean breeding and research. We address three important issues relevant to the potential use of SNPs in the characterization of, and comparisons among soybean varieties: 1) using single seed assays to measure the incidence of heterozygosity, and 2) using both single seed and various sized single seed bulks to measure residual heterozygosity (ideotypes) within soybean varieties. These results are then analyzed to determine a) how many seeds are required to obtain a reliable SNP profile for a soybean variety, and b) if bulk samples can be used, then how many seeds comprise the bulk.

MATERIALS

4. Five publicly available soybean varieties with expired PVP status were selected that had been developed in the United States by two proprietary breeding programs (Table 1). Each of these varieties had been granted PVP certificates by the US Plant Variety Protection Office using established morphological criteria. Consequently, each seed lot had been determined to be sufficiently Distinct, Uniform, and Stable for the granting of Intellectual Property Rights in the form of PVP.

Table 1.

Variety Name	**PI Accession Number**	**PVP Number**	**PVP Applicant**	**Pedigree as stated in PVP**
9171	PI 542056	9000181	Pioneer Hi-Bred International, Inc. (DuPont Pioneer)	9271 x A3127
9221	PI 542058	9000183	Pioneer Hi-Bred International, Inc. (DuPont Pioneer)	1677 X Franklin
9551	PI 550733	9100185	Pioneer Hi-Bred International, Inc. (DuPont Pioneer)	J74-45/1095C42. J74-45 is a full sister to Bedford. 1095C42=Essex/D66-5566. D66-5566=D49-2491*4/Hawkeye.
A2396	PI 540454	9000146	Asgrow Seed Company (Monsanto)	CM214 X A2943
A2835	PI 561718	9200223	Asgrow Seed Company (Monsanto)	X1972 X A3935 (X1972 = A2 X Vickery)

METHODS

Individual Seed and Bulk Seed Sampling Regimes

5. The BARCSoy6K panel (Lee et al, 2015) sourced from Illumina® (Illumina, Inc., San Diego, CA, USA) was used to profile each variety. Soybean varieties 9171, 9221, and 9551 were assayed at the DuPont Pioneer Lab in Johnston, IA. USA. Soybean varieties A2396 and A2835 were assayed at the Monsanto Lab in St. Louis, MO. USA.

6. The sampling regimes used varied slightly by lab, but maintained the essential components of individual seed assays and subsequent bulk seed assays from the same set of individuals.

DuPont Pioneer Sampling Regime

7. Nineteen individual seeds were grown out and each individual seed sampled and assayed. Then different size bulk regimes were sampled and assayed as follows; seeds 1-3 were bulk sampled, then seeds 1-5, 1-7, 1-9, 1-11, 1-13, 1-15, 1-17, & 1-19 made up the remaining bulk samples. Each sampling regime was repeated twice for each variety, i.e. 19 seeds were selected for 1 replicate, and 19 additional seeds were selected for the second replicate.

Monsanto Sampling Regime

8. Seventeen individual seeds were grown out and each individual seed sampled and assayed. Then different size bulk regimes were sampled and assayed from the same individual seeds as; seeds 1-5, 1-7, 1-9, 1-11, 1-13, 1-15, and 1-17. The sampling regime was not repeated.

Evaluation of Heterozygosity and Residual Heterozygosity (Presence of SNP Ideotypes Within a Variety)

9. The profiles of the BARCSoy6K panel generated from individual seeds were observed to determine the number of individuals and markers that were recorded as heterozygotes for each variety. SNP profiles of individual seeds were used to compute observed allele frequencies, which would also be the expected allele frequencies for the bulk samples. Heterogeneity can result from two sources: A) presence of heterozygous SNP individuals and B) individuals comprising the bulk that were reported as homozygous for alternate SNP alleles (Different Ideotypes). Minor allele frequencies were calculated from the results of single seed assays

and these would be the expected minor allele frequencies for the bulked samples of those individual seeds. Minor allele frequencies were calculated to determine if allele frequencies appeared to be associated with differences between observed compared to expected reports of heterogeneity in bulk samples.

Determining the Number of Seeds to Assay to Characterize a Soybean Variety

10. To determine the number of seeds to be sampled to adequately characterize a variety, we calculated the measurement uncertainty (MU – ISO, 1995) associated to the heterogeneity rate. This is derived as follows: Let M_i be a "heterogeneous SNP" with Minor Allele Frequency (MAF) equal to MAF_i. Given k seeds are sampled in the lot, then the probability p_i that there is at least one heterogeneous seed in the sample for this SNP is equal to:

$$p_i = 1-(1-MAF_i)^k$$ [1].

11. We assume that n SNP out of N characterizes heterogeneity for a given variety. Let X_i be the random variable having value 1 if there is at least one heterogeneous seed in the k seeds sampled for marker i, 0 otherwise. Then the distribution of $Y = \sum_{i=1}^{n} X_i$ is Poisson binomial with mean $\mu = \sum_{i=1}^{n} p_i$ and variance $\sigma^2 = \sum_{i=1}^{n} p_i (1-p_i)$. The heterogeneity rate h_k is therefore given by:

$$h_k = \frac{\mu}{N}$$

and its measurement uncertainty by:

$$MU = \frac{\sqrt{\sigma^2}}{N}$$

12. MUs were then graphed to review the precision using Monte Carlo simulation: for each number of seeds sampled, a random MAF value within the MAF range is generated for each heterogenous SNP 10,000 times and the mean of the MUs computed and displayed.

RESULTS

Incidence of Heterozygosity

13. Heterozygous SNPs ranged from 0.02 – 0.15% across all varieties in the experiment, so were not a major contributor to the level of heterogeneity observed in the soybean varieties when sampled as bulks (Table 2). The identity of heterozygous SNPs varied with each variety and where reported, were present in > 50% of individual seeds and consistently also using bulk.

Incidence of Residual Heterozygosity (ideotypes) within Soybean Varieties

14. Percentages of SNPs per variety scored as heterogeneous ranged from 0.16% to 4.27% (adding together % SNPs scored in both individual seed and bulk assays with those scored in individual seeds but not in bulks).

15. Comparison of levels of residual heterozygosity when measured using bulks to those known and expected from single seed assays if bulk sampling data were reported with 100% fidelity ranged from 91.67 to 97.1% fidelity.

16. Table 2 provides the levels of occurrence of MAF. The number of SNP ideotypes rather than presence of heterozygous individuals was the primary factor contributing to heterogeneity in each variety.

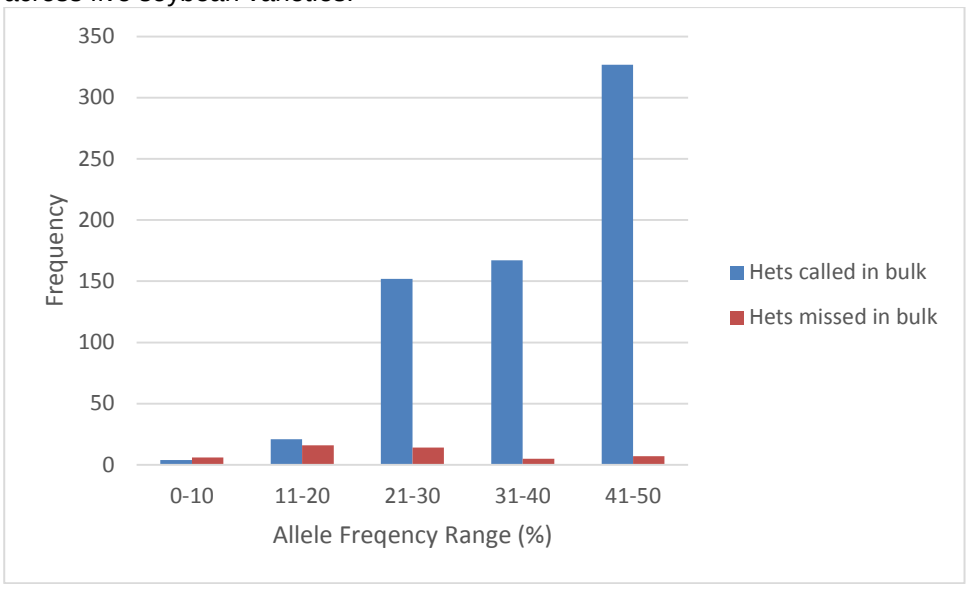
[1] It is assumed here that there is no heterozygous individuals for SNP i. Otherwise, $p_i \geq 1-(1-MAF_i)^k$.

Table 2. Heterozygosity, residual heterozygosity and expected versus observed heterogeneity (heterozygosity + residual heterozygosity) among individuals of the largest bulk sample size observed in the experiment for each variety and replicate.

Variety & Rep	Bulk Size	Total SNPs	# SNPs scored with > 50% individuals Heterozygous (All Heterozygous loci comprised > 50% of all individuals)	% SNPs scored with > 50% individuals Heterozygous (All Heterozygous loci comprised > 50% of all individuals)	# SNPs Scored with ideotypes in both single seed and bulk samples	% SNPs Scored with ideotypes in both single seed and bulk samples	# SNPs not scored as ideotypes in bulk, but ideotypes present in individuals	% SNPs not scored as ideotypes in bulk, but ideotypes present in individuals	Fidelity = % SNPs Heterozygotes + Ideotypes observed in bulks / % SNPs Heterozygotes + Ideotypes Expected in Individuals
9171 Rep 1	19	5902	1	0.02%	66	1.12%	4	0.07%	94.37%
9171 Rep 2	19	5902	1	0.02%	66	1.12%	2	0.03%	97.10%
9551 Rep 1	19	5902	3	0.05%	8	0.14%	1	0.02%	91.67%
9551 Rep 2	19	5902	3	0.05%	8	0.14%	1	0.02%	91.67%
9221 Rep 1	19	5896	1	0.02%	118	2.00%	5	0.08%	95.97%
9221 Rep 2	18	5896	1	0.02%	115	1.95%	8	0.14%	93.55%
A2835	17	4541	7	0.15%	178	3.92%	16	0.35%	92.04%
A2396	17	4541	5	0.11%	158	3.48%	11	0.24%	93.68%

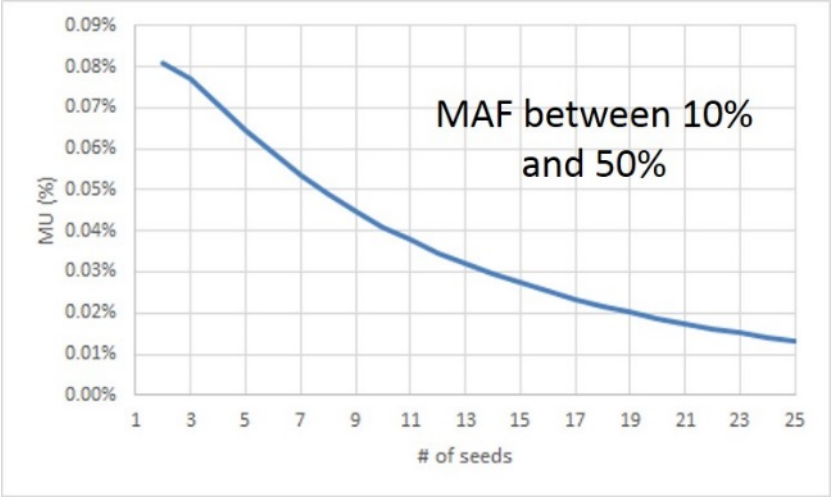
17. The number of SNPs reporting ideotypes in a variety ranged from 8-178. The overwhelming proportion of ideotypes were present with MAF greater than 21% (Figure 1) and the vast majority of these ideotypes were also reported using bulk assays (Table 2). Ideotypes scored when sampled as individual seeds, but not reported in bulks were distributed across MAFs and not abnormally higher at low MAFs (Figure 1). Low MAFs (<21%) comparisons between ideotypes reported versus not reported were proportionally higher, but there are so few in these frequency categories, the impact was very small. There were no reports of heterozygosity for a variety using bulk samples where individual single seed assays had reported no ideotypes.

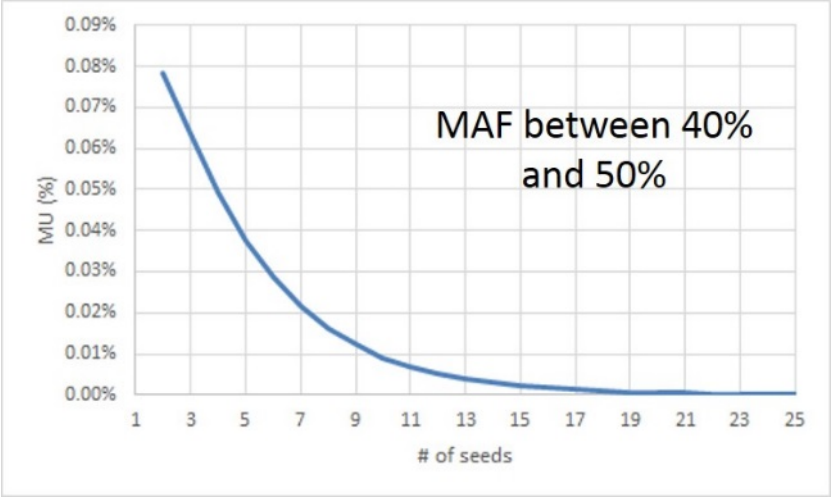
Figure 1. Allele frequency comparison between ideotypes reported and not reported by allele frequency across five soybean varieties.



What Size Bulks to Use?

Figure 2. Measurement uncertainty response for 90 heterogeneous SNPs out of a total of 5560 SNPs while varying MAF between 10% and 50% and between 40% and 50%.





18. Figure 2 shows the response for MUs at different MAF. With a MAF down to 10%, the MUs were 0.04% for 10 seeds and 0.02% for 20 seeds. At MAF down to 40% MAF, the MUs were 0.01% for 10 seeds and .0001% for 20 seeds.

19. The number of heterozygotes present in the soybean varieties was very low (0.02-0.15%). Presence of different ideotypes was, by far, the primary contributor to heterogeneity in the soybean variety assays (0.14-3.92%). These data show that for the vast majority of SNP loci we examined (MAF greater than 20%) that precision or a lower measure of uncertainty improves, but is fractionally very small for every additional seed assayed above 10. Additional evidence from experimental data where up to 19 individual seeds were assayed per variety confirmed this conclusion. Bulk sampling of 10 seeds is 10x more efficient in its use of resources than use of individual assays and maintains a high level of fidelity between results from bulk assays compared to results using those same individuals scored as individual seeds (91.7-97.1%). We therefore propose that a bulk sample of 10 seeds be used to provide the basis for a characteristic SNP profile for varieties soybean.

REFERENCES

Fasoula VA, Boerma HR (2007). Intra-cultivar variation for seed weight and other agronomic traits within three elite soybean cultivars. Crop Sci 47:367-373.

International Organization for Standardization (ISO) (1995). ISO Guide to the Expression of Uncertainty in Measurement, Geneva, Switzerland.

Lee S, Freewalt KR, McHale LK, Song Q, Jun T-H, Michel AP, Dorrance AE, Mian MAR (2015). A high resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6k. Mol Breed. doi:10.1007/s11032-015-0209-5

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S and Ziegle J (1997). An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L.): Comparisons with data from RFLPs and pedigree. Theor Appl. Genet 95:163-173.

UPOV. 1993a. Report adopted by the technical committee from the 28th 582 session, October 21-23,1992 in Geneva. TC/28/6. UPOV, Geneva, Switzerland.

UPOV. 2011. Possible use of molecular markers in the examination of Distinctness, Uniformity, and Stability (DUS). UPOV/INF/18//1 with 4 annexes. UPOV, Geneva, Switzerland. p.26.

Barry Nelson[1], Fred Achard[2], Jean-Louis Laffont[3], Stephen Smith[4] et al.

[1]DuPont Pioneer, Johnston, IA. USA
[2]Monsanto, St. Louis, MO. USA
[3]DuPont Pioneer, Aussonne, FR
[4]Seed Science Center, Iowa State University, Ames, IA. USA

[End of document]