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/17/18

Seventeenth Session
Montevideo, Uruguay, September 10 to 13, 2018

Original: English
Date: September 5, 2018

APPROACHES TO VARIETY IDENTIFICATION AND SIMILARITY MARKER SELECTION IN SOYBEANS USING SNPS: HOW FEW SNPS CAN DO THE JOB?

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

Disclaimer: this document does not represent UPOV policies or guidance

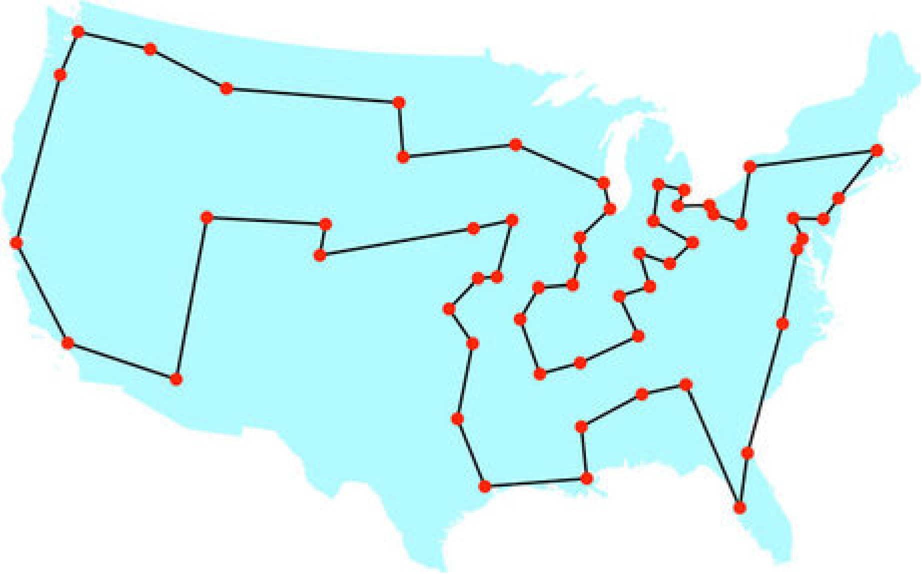
INTRODUCTION

1. Molecular markers are widely used by seed companies globally. In particular, Single Nucleotide Polymorphisms (SNPs) because they accurately reflect the genotype, are highly discriminative, reliably scored, and cost effective (Gale et al). SNP profiles can be used to assure varietal identity, genetic purity, and support intellectual property protection. The application of SNPs to these uses can be categorized as follows; variety identification (ID) and similarity or genetic distance (1-similarity) comparisons. A variety ID panel is comprised of a small number of markers where molecular codes are able to uniquely identify varieties. Use of a small number of markers is well suited to allow the assay of a large number of samples at very low cost. However, the variety ID approach should not be confused with or mistaken for as suitable for calculation of genetic distances, or similarities between varieties. For use of relatively few markers can lead to a loss of precision when genetic distances are calculated. Therefore, we need to be clear that variety ID and genetic similarity are very different issues that must be approached differently. Showing difference compared to determining the degree of difference, or alternately similarity, are not synonymous and require different approaches. The goal of this analysis is 1) to demonstrate useful methods to select low numbers of SNPs for variety ID and 2) to find a minimum number of SNPs that is useful for a similarity comparison in soybeans.

VARIETY IDENTIFICATION

2. Previous work has demonstrated that small sets of SNPs can be very effective to distinguish soybean varieties when SNPs are selected largely based on informativeness i.e. abilities to individually discriminate among varieties while collectively allowing most, if not all chromosomes to be sampled (Yoon et al, Liu et al). Here a different approach to SNP selection for variety ID is investigated using an algorithm to select SNPs based solely on their discrimination abilities regardless of known map positions. A public soybean database (Soybase.org) of genotyped varieties consisting of 322 varieties regarded as important to North American soybean breeding was used as the basis of the study (Grant et al). The varieties are curated and were genotyped by the USDA using BARCSoySNP50k SNPs (Song et al). The marker set was sub-set to match that of the BARCSoySNP6K (6k) marker set. Then a computer algorithm was used based on the Traveling Salesman Problem to select SNP markers (Lawler et al). The Traveling Salesman Problem is described as the most efficient city-to-city route a salesperson should travel while visiting each city exactly one time (Figure 1). This concept was applied to determine the minimum number of SNPs needed to differentiate all soybean varieties in the database. The process begins with the algorithm targeting just 2 SNPs then iteratively sampling 2 markers that differentiate the most varieties. Then the process was repeated systematically increasing the number of SNPs until all varieties are differentiated. Once the SNPs were selected, they were then tested for robustness with simulated missing data levels from 0, 10, 20, 30, 40, & 50%.

Figure 1. The Traveling Salesman Problem as demonstrated for the most efficient route a salesperson should travel while visiting each city exactly one time in the United States of America.



3. The database of 322 varieties was evaluated for heterozygotes and all varieties >1% heterozygotes were removed leaving 276 varieties for assessment. Using the Travelling Salesman algorithm, 14 highly informative SNP markers were selected and found to be sufficient to differentiate all 276 soybean varieties (Figure 2). Adding markers beyond 5 SNPs iteratively up to 10 SNPs had the greatest impact on number of varieties differentiated.

4. The initial 6K dataset used had a missing data rate of only 0.007%. This low level of missing data was very helpful for not only selecting the 14 SNPs, but also lends itself to an accurate assessment of robustness. However, other datasets may have various levels of missing data. Thus, this dataset proved to be a reliable vehicle to allow the robustness of variety ID to be tested with various levels of missing data. The results of robustness testing in the face of 0 to 50% missing data are shown in Figure 3. Missing data had little impact on the performance of the 14 selected SNPs where missing data levels as high as 40%, still maintained >90% differentiation of the pairs of varieties.

Figure 2. Differentiating power of successive SNPs selected using the Traveling Salesman Algorithm.

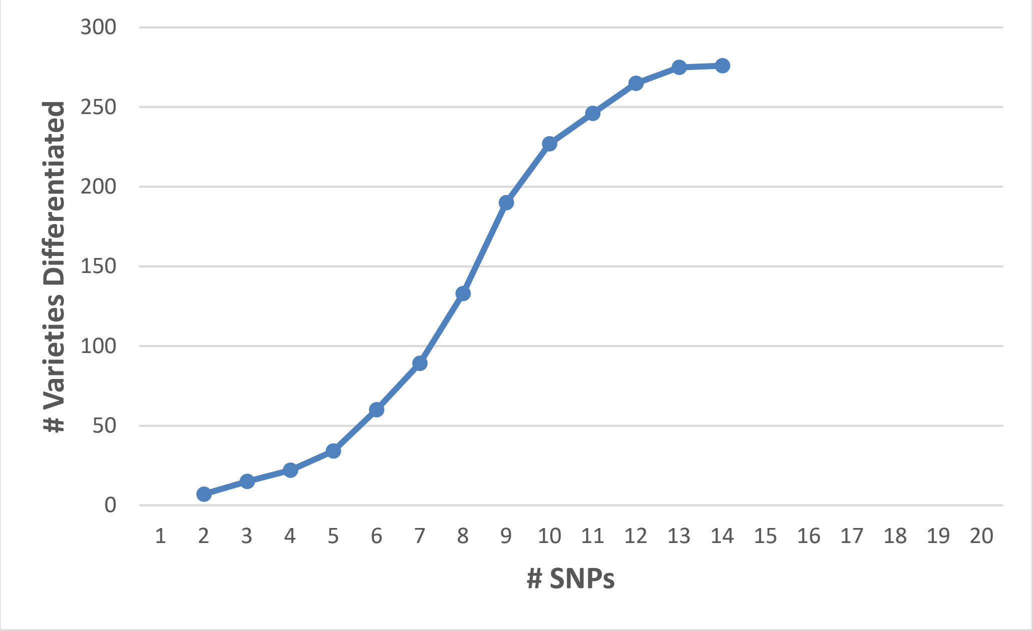
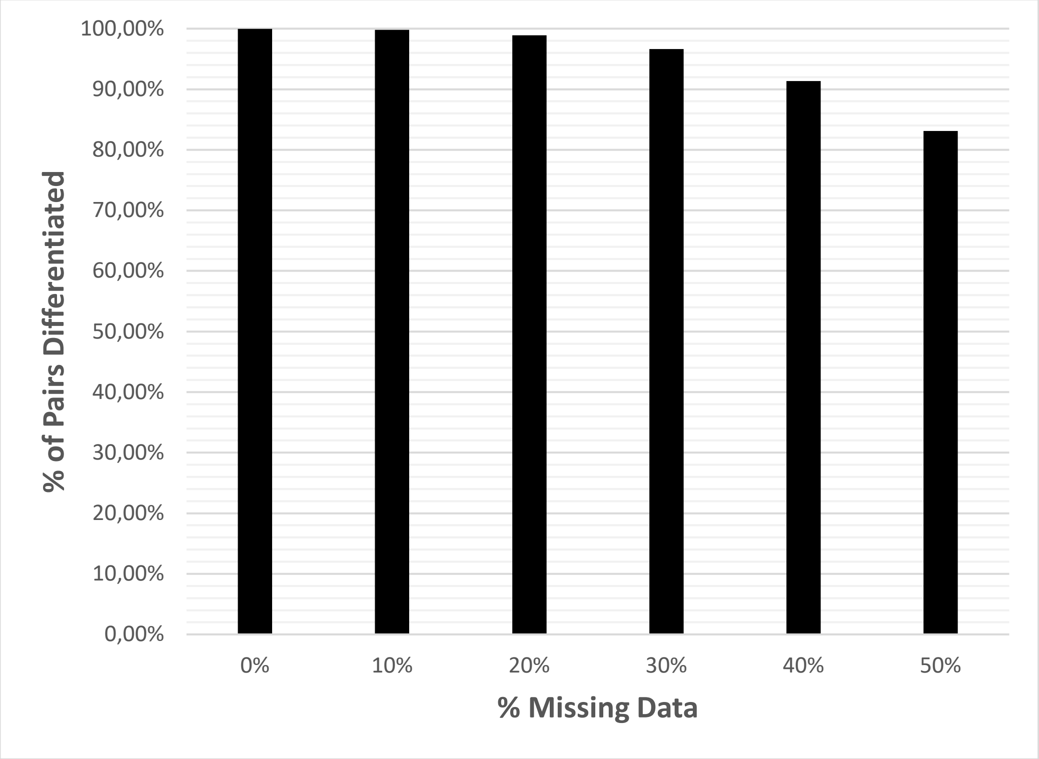


Figure 3. Performance of Traveling Salesman Algorithm selected SNP panel when subjected to various levels of simulated missing data.



5. The Traveling Salesman Algorithm performed very well in selecting 14 SNPs to differentiate all 276 soybean varieties. The 14 SNPs were robust in the face of missing data, even up to 40%. A key point in the analysis is that it's assumed that all the varieties deemed necessary to differentiate are contained in the database used. Therefore, if varieties or genetics change among the varieties one desires to identify, the algorithm and SNP selection process may need to be repeated to reflect these varietal changes.

SIMILARITY COMPARISONS

6. The same 6K SNP markers and 322 varieties from Soybase were used and expected heterozygosity values calculated for each marker (Nei 1987). Both expected heterozygosity and genetic map positions of individual SNPS were then used as the basis for sub-setting the 6K SNPs into smaller marker sets to test how smaller SNP sets perform in similarity comparisons. The full set of markers comprised 5346 markers, then sub-sets of 2673, 1336, 668, 334, and 167 markers were selected as the subsets. The average expected heterozygosity value for the full 5346 data set was 0.375. Care was taken to maintain this level of expected heterozygosity across each marker subset. Where markers shared the same genetic map position, the marker with the lower expected heterozygosity value was removed. As markers were removed, an attempt was made to maintain similar distances between markers across the genome ensuring marker coverage at the ends of each chromosome. After marker selection, pair-wise similarities were computed for all pairs of varieties using a simple matching routine at the allele level for each data set(Sokal and Michener). Pair-wise similarities for each data set were then regressed against the 5346 pair-wise similarities to determine how well they correlated, thus how far a marker set can be reduced, yet still be effective for similarity comparisons. Lastly, average difference and standard deviations were computed for each marker subset as compared to the pair-wise similarities of the full 5346 SNP set.

Table 1. Average, minimum, and maximum distance between SNP markers as marker number is reduced for each data set.

	Number of SNP Markers					
	5346	**2673**	**1336**	**668**	**334**	**167**
Avg	0.50	0.81	2.01	4.05	8.16	16.63
Min	0	0	0.01	0.01	0.01	0.01
Max	6.09	6.09	8.51	14.05	33.81	59.3

7. The strategy to maintain expected heterozygosity at 0.357 across marker sets, yet maintain consistent genome distribution is summarized in Table 1. The average distance between markers essentially doubles at the 2673 SNP set and each successive set below that point as the marker sets are cut in half. Although intervals increased between markers as SNP density decreased, genome coverage was still maintained at 99% across all marker sets.

Table 2. The full 322 variety and >80% pair-wise similarity R-square values for each marker set graphed against the full 5346 SNP pair-wise similarities while holding expected heterozygosity constant at 0.357.

	Number of SNP Markers				
	2673	**1336**	**668**	**334**	**167**
Full Data Set	0.969	0.953	0.93	0.879	0.82
>80% Similarity	0.888	0.878	0.825	0.767	0.719

8. R-square values remained relatively high (82% and up) across all marker sets even down to 167 SNPs (Table 2.) Generally, R-square values began dropping more rapidly moving down from the 668 to 334 SNP sets with a difference of 0.055. Given the majority of interest in similarity comparisons will be at the higher similarity levels in practice, the pair-wise similarities were sub-set to >80% and each marker set re-assessed against the full 5436 SNP set. R-square values changed little down to 1336 SNPs with a difference from 2673 SNPs to 1336 SNPs of 0.01. Reducing SNP numbers lower increased the rate of decline in R-square below 668 SNPs.

Figure 4. Average difference between pairs of each marker set as compared to the same pair-wise similarities of the full 5346 marker set.

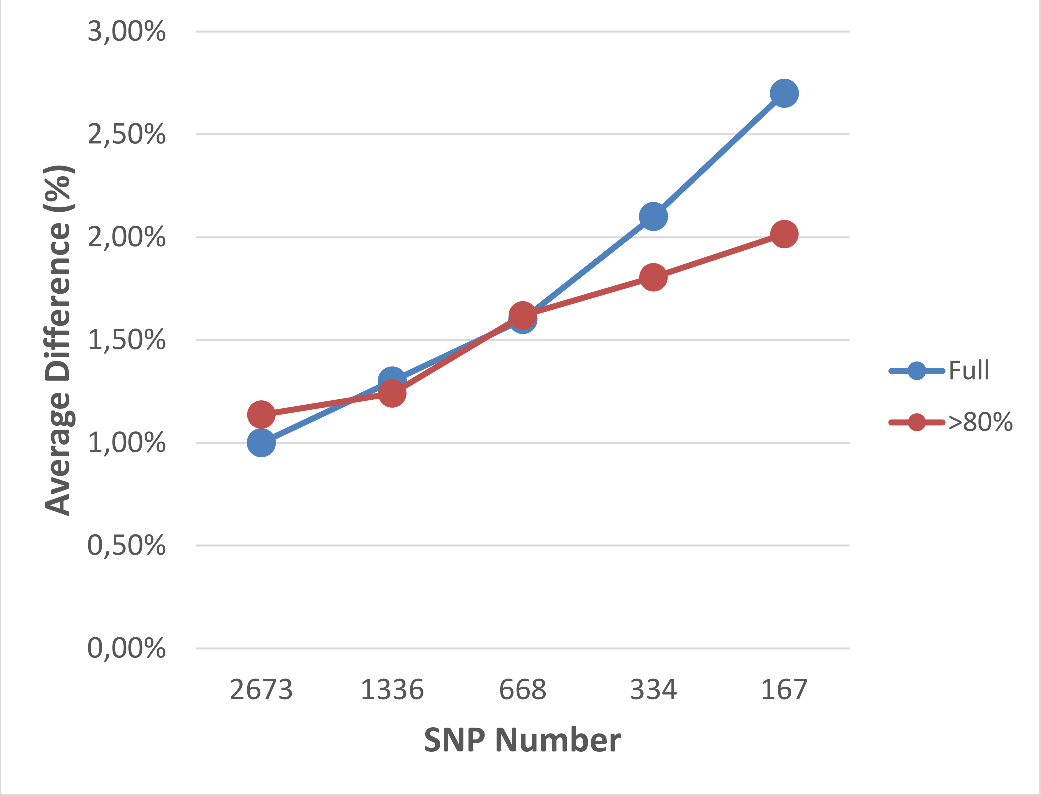
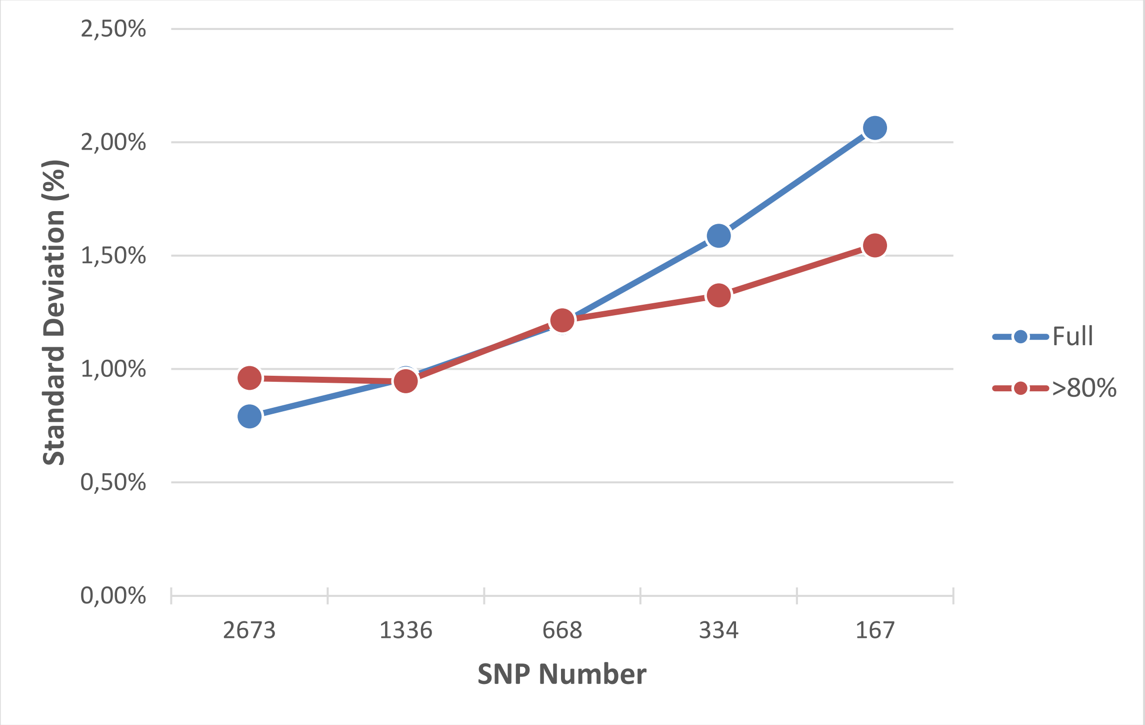


Figure 5. Standard deviations of each marker set as compared to the same pair-wise similarities of the full 5346 marker set.



9. Average difference from the full 5346 SNP set and standard deviations of SNP subsets are presented in Figure 4 & 5. SNP set sizes of 2673 and 1336 markers had an average difference between 1 and 1.5% while maintaining a <1% standard deviation for both the full set of pairs and the pairs >80%. The 668 SNP set is the point where average differences began to escalate just over 1.5% and standard deviations were >1% suggesting the interval of SNP numbers between 1336 to 668 may be a reasonable level for a minimum number of SNPs necessary for accurate similarity comparisons. All marker sets maintained an average similarity of 64% across all pairs for the full set and between 84 and 85% for the >80% similarity pair data.

10. Based on the analysis and the need to maintain high R-square values for accurate similarity comparisons, the levels of average similarity difference compared to the full 5346 SNP set, and standard deviations support a conservative minimum range of 1336 to 668 SNPs are necessary when basing a similarity comparison using the 6K SNPs. The approach maintaining genome coverage while focusing on keeping expected heterozygosity constant when sub-setting smaller SNP sets is an effective means for a more efficient number of markers for similarity comparisons.

REFERENCES

- Gale, Z. Jiang, H., Westcott M. 2005. An optimization method for the identification of minimal sets of discriminating gene markers: Application to cultivar indentification in wheat. Jour. Bioinformatics and Computational Biology. 3:269-279.
- Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database.
  Nucl. Acids Res. (2010) 38 (suppl 1): D843-D846.
- Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., and Shmoys, D. B. 1985. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* John Wiley and Sons, New York.
- Liu, Z., Li, J., Fan, X., Htwe, N. M. P. S., Wang, S., Huang, W., ... Qiu, L. (2017). Assessing the numbers of SNPs needed to establish molecular IDs and characterize the genetic diversity of soybean cultivars derived from Tokachi nagaha. *Crop Journal*, *5*(4), 326-336.
- Nei, Masatoshi. Molecular evolutionary genetics. 1987. Columbia University Press New York Chichester, West Sussex Copyright© 1987 Columbia University Press.
- Sokal, R., Michener, C. (1958). A statistical method for evaluating systematic relationships. In: Science bulletin, 38(22), The University of Kansas.

- Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, et al. (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE 8(1): e54985. https://doi.org/10.1371/journal.pone.0054985
- Yoon, M.S., Q.J. Song, I.Y. Choi, J.E. Specht, D.L. Hyten, P.B. Cregan. 2007. BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. Theor. Appl. Genet. 114, 885-899.

[End of document]