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/16/14 Add.

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

Original: English
Date: November 2, 2017

ADDENDUM TO THE USE OF REFERENCE VARIETY SIMILARITIES IN VARIETAL DISTINCTNESS II: REFERENCE VARIETY SELECTION

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

Disclaimer: this document does not represent UPOV policies or guidance

INTRODUCTION

1. At the fourteenth session of the Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular (BMT), Nelson et al. (2014) produced the idea of using genetic similarity coefficients with a set of reference varieties to determine varietal distinctness. At the fifteenth session of the BMT, Mr. Kees van Ettekoven (2016) expanded this model to fit within the context of a UPOV characteristic. Here we address reference variety selection, a key feature in application of the reference variety model. We explore the problem theoretically and empirically. Challenges and opportunities are addressed.

REFERENCE VARIETY SELECTION

2. The reference variety model utilizes genetic similarity coefficients between PVP candidate subject varieties and a set of pre-determined reference varieties. Reference varieties are selected from among varieties of common knowledge. Consider n varieties of common knowledge. Using DNA markers, an $n \times n$ similarity matrix can be computed. The task of reference variety selection is to determine which subset of k varieties should be chosen whereby distinctness can be assessed across the entire set of n varieties. The 'best' set of references will provide a measurement of relationship which most closely mirrors the true genetic relationship, as approximated by the DNA marker-based similarity coefficients.

Theoretical Example

3. In Figure 1, a fictitious set of 22 varieties of common knowledge are plotted in a two-dimensional space. One minus the Euclidian distance between these points can be used as a representation of genetic similarity, and is given in Table 1 as a 22×22 similarity matrix. Consider five selections of reference varieties: 1) subset from the periphery *[a, c, e, g, i, k]*, 2) subset from the intermediary *[m, n, o, p, q, r]*, 3) subset from the center *[s, t, u, z]*, 4) even-sampled subset *[a, e, i, n, p, r, z]*, and 5) all varieties (Table 2). The all variety selection is included as a point of reference. Using the previously calculated genetic similarities, we can recalculate five new 22×22 similarity matrices, one corresponding to each of the selections. The cosine measure of similarity was used for similarity calculation.

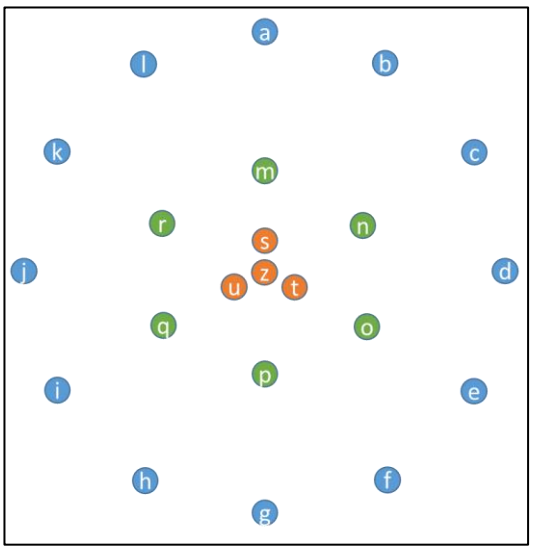


Figure 1 Representation of varieties of common knowledge.

4. In Figure 2, the five selection-based similarities are plotted against the original, true, similarities. Similarities from subset (1), with selections from the periphery, have the greatest correlation to the true similarities, equaling that of the all-variety selection. The intermediary-sampled subset provides the next-best approximation, followed by the even-sampled subset and lastly the center-sampled subset. From this theoretical example, we are led to surmise that the optimal selection of reference varieties will be from the outer periphery of varieties of common knowledge – or in other words, the selection of varieties which minimize both the similarity with the cohort and the similarity among themselves. In fact, when we compute average similarities between subgroup members and the cohort, and between subgroup members per se, we find that the varieties on the outer periphery do minimize both these values, at 0.42/0.38 respectively, where average similarities for the intermediary, center, and evenly-sampled varieties are 0.60/0.71, 0.65/0.93, and 0.53/0.56, respectively.

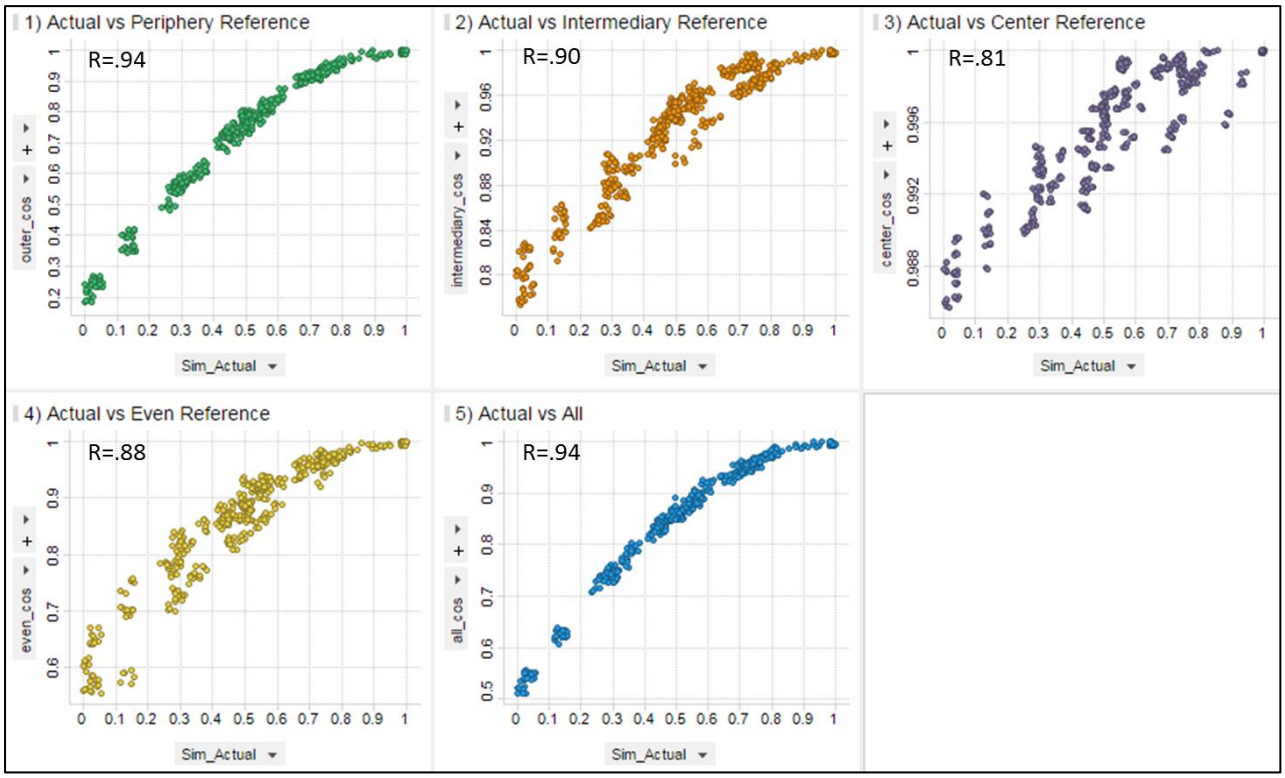


Figure 2: selection-based similarity plotted against actual similarity for each of the five reference selections.

Reference Variety Selection among Maize Inbreds

5. A set of 621 publicly available maize inbred lines were genotyped using the International Seed Federation (ISF) set of 3072 SNP markers (ISF 2014). Genetic similarity between all pairs of lines was calculated using a simple matching coefficient. The 621 lines can be classified generally into four groups: open pollinated varieties (OPV), public university inbreds (public), expired PVP inbreds (PVP), or temperate-adapted varieties from the United Stated Department of Agriculture – Agricultural Research Service (USDA-ARS) germplasm enhancement of maize program (GEM). Figure 3 gives a SNP similarity-based principal component plot of the varieties.

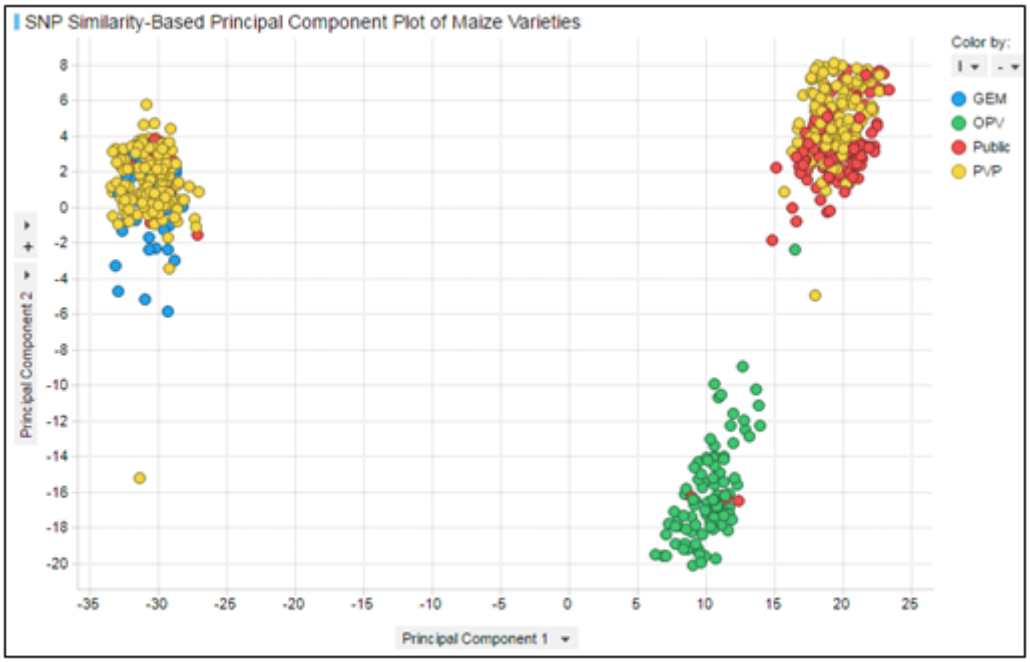


Figure 3: SNP marker-based principal component plot of 621 public varieties.

6. Two approaches were used to select 20 reference lines out of the 621. First, based on the outcome of the theoretical example above, we selected the 20 lines which satisfied the conditions of minimum similarity to the cohort and minimum similarity between themselves (Figure 4a).

7. For the second approach, we applied an optimization algorithm which balanced the conditions of maximum genetic variance and eliteness. A similar approach is used in financial investment strategy where an investor is seeking a balance of diversification and performance (Markowitz et al., 2000). Eliteness is an approximation of genetic improvement and by including this condition the model will increase selection of newer varieties which are more relevant to the incoming class of PVP applicant varieties (Figure 4b). The similarity matrix is used to compute the genetic variance for a set of reference lines. The selection of 20 reference lines out of 621 is a combinatorial problem, and there could be approximately 30^{20} such possible sets which could be created of 20 reference lines out of 621. Each of the sets could be unique in terms of the measure of genetic variance and eliteness. We have used techniques from the discrete optimization theory to obtain the optimal set of references out of the 30^{20} possible sets, which would maximize the variance and eliteness.

8. Figure 5 shows the outcome of these two reference selection methods, where the reference-based similarities are plotted against actual SNP-based similarities. Overall correlation is very good for both models, R=0.93 and R=0.92, meaning that for most bi-variety comparisons, the similarity approximation using reference varieties is accurately reflecting true genetic similarity.

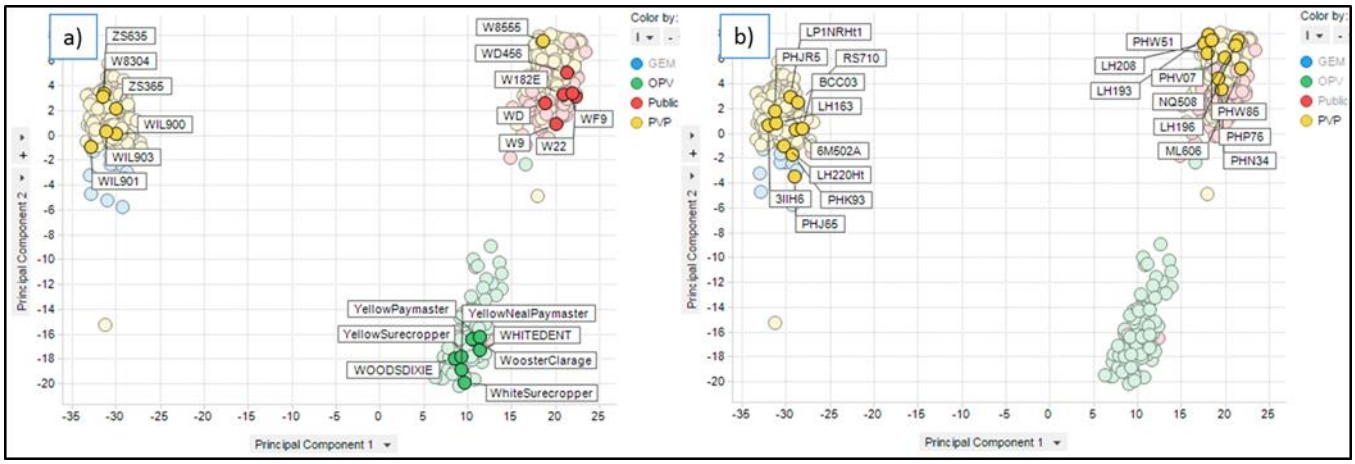


Figure 3 Selected reference varieties from two selection methods: a) maximizing diversity sampling and b) balancing diversity and eliteness.

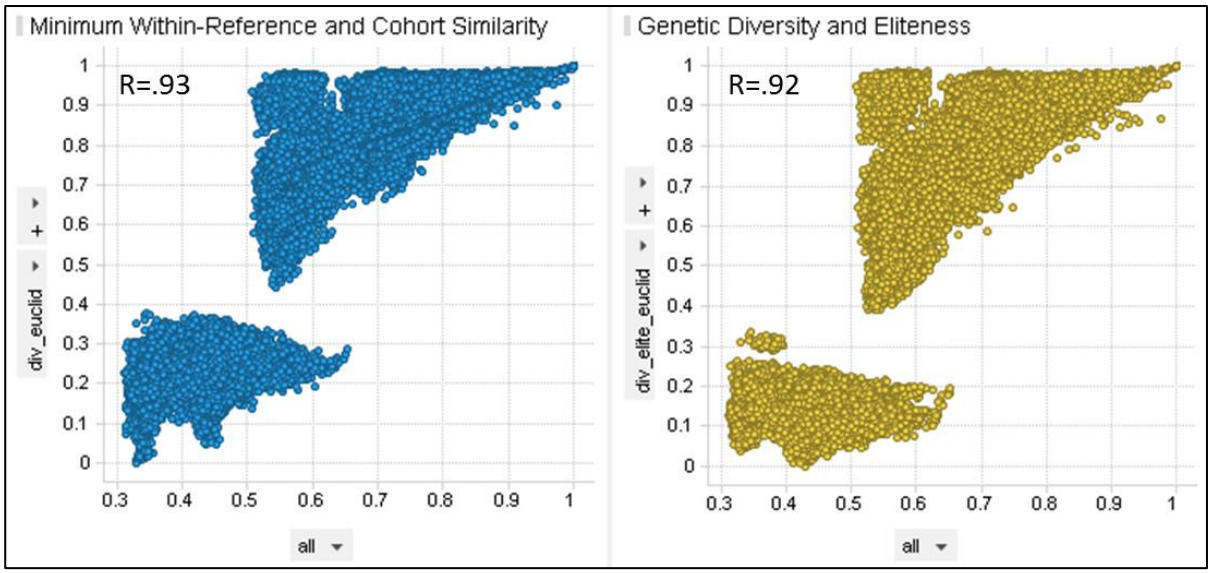
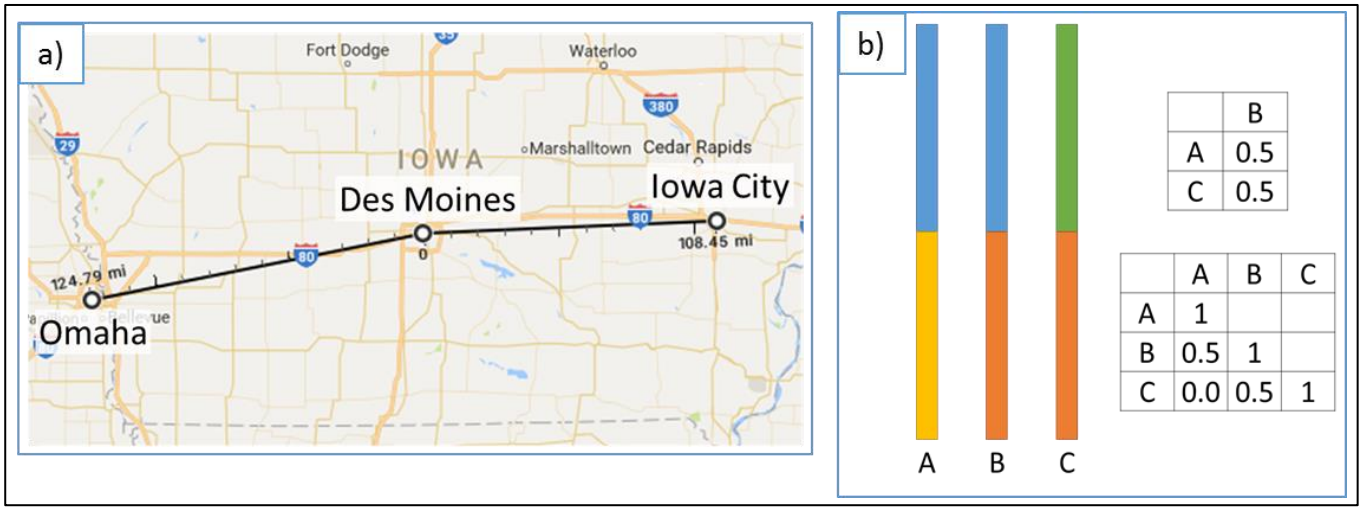


Figure 4: reference variety-based similarity plotted against actual similarity for two methods of variety selection – left: minimizing similarity with references and with cohort, right: balancing genetic diversity and eliteness.

9. There is cause for some concern over the shape of the cloud of points in Figure 4, particularly in the upper-similarity cloud which resembles a right triangle. Variety comparisons which fall in the upper left vertex of the triangle are being heavily biased toward a high similarity. This is particularly concerning since the purpose of this analysis is to determine distinctness between pairs of varieties, yet many are appearing artificially similar.

10. There is a simple genetic explanation for this similarity bias. First, let us reflect on the city analogy which was shared by Nelson et al (2014) at the fourteenth session of the BMT. Consider now three US Corn Belt cities: Omaha, Des Moines, and Iowa City (Figure 6a). Omaha and Iowa City are approximately equidistance from Des Moines, 125 and 108 miles, respectively. Based on this information alone, one might conclude that Omaha and Iowa City are near each other, where in fact they are situated in opposite directions from Des Moines.

11. Now consider one chromosome from each of three varieties, A, B and C, as pictured in Figure 6b, where matching colors represent genetic identity. If only the relationship to variety B is considered, one may conclude that varieties A and C are not distinct, each having a similarity of 0.5 to variety B, where in actuality, A and C have a genetic similarity of 0.0.



	B
A	0.5
C	0.5

	A	B	C
A	1		
B	0.5	1	
C	0.0	0.5	1

Figure 5: illustration of similarity bias with reference comparison. a) reference city example revisited. b) chromosomal region similarity between three varieties. (Map Image Credit: Google Maps)

CONCLUSION

12. Proper selection of reference varieties is an important part of the reference variety model. The optimization methods shown here can be useful in reference variety selection, but more can be done to refine the methods used and to explore alternate algorithms. In the process of establishing distinctness, the practitioner will give close attention to the upper tail of the similarity distribution, but unfortunately this is the area most confounded the similarity bias illustrated in this study. Further analysis should focus on minimizing this type of similarity bias by considering not only which varieties to select, but also how many should be selected – something not addressed here.

Table 1: 22 x 22 matrix of similarities between the fictitious varieties from Figure 1.

Variety	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	z
a	1	0.74	0.50	0.30	0.14	0.04	0.01	0.04	0.14	0.30	0.50	0.74	0.72	0.54	0.34	0.29	0.34	0.54	0.57	0.47	0.47	0.50
b	0.74	1	0.74	0.50	0.29	0.13	0.04	0.00	0.04	0.14	0.29	0.50	0.67	0.66	0.44	0.30	0.27	0.42	0.55	0.50	0.44	0.50
c	0.50	0.74	1	0.74	0.50	0.29	0.14	0.04	0.00	0.04	0.13	0.29	0.56	0.73	0.57	0.36	0.25	0.33	0.53	0.53	0.43	0.50
d	0.30	0.50	0.74	1	0.74	0.50	0.30	0.14	0.04	0.01	0.04	0.14	0.46	0.71	0.69	0.46	0.28	0.28	0.50	0.56	0.44	0.50
e	0.14	0.29	0.50	0.74	1	0.74	0.50	0.29	0.13	0.04	0.00	0.04	0.36	0.60	0.75	0.56	0.34	0.26	0.47	0.57	0.46	0.50
f	0.04	0.13	0.29	0.50	0.74	1	0.74	0.50	0.29	0.14	0.04	0.00	0.30	0.47	0.69	0.67	0.44	0.30	0.45	0.56	0.49	0.50
g	0.01	0.04	0.14	0.30	0.50	0.74	1	0.74	0.50	0.30	0.14	0.04	0.29	0.37	0.57	0.72	0.57	0.37	0.44	0.53	0.53	0.50
h	0.04	0.00	0.04	0.14	0.29	0.50	0.74	1	0.74	0.50	0.29	0.13	0.30	0.30	0.44	0.67	0.69	0.47	0.45	0.49	0.56	0.50
i	0.14	0.04	0.00	0.04	0.13	0.29	0.50	0.74	1	0.74	0.50	0.29	0.36	0.26	0.34	0.56	0.75	0.60	0.47	0.46	0.57	0.50
j	0.30	0.14	0.04	0.01	0.04	0.14	0.30	0.50	0.74	1	0.74	0.50	0.46	0.28	0.28	0.46	0.69	0.71	0.50	0.44	0.56	0.50
k	0.50	0.29	0.13	0.04	0.00	0.04	0.14	0.29	0.50	0.74	1	0.74	0.56	0.33	0.25	0.36	0.57	0.73	0.53	0.43	0.53	0.50
l	0.74	0.50	0.29	0.14	0.04	0.00	0.04	0.13	0.29	0.50	0.74	1	0.67	0.42	0.27	0.30	0.44	0.66	0.55	0.44	0.50	0.50
m	0.72	0.67	0.56	0.46	0.36	0.30	0.29	0.30	0.36	0.46	0.56	0.67	1	0.75	0.60	0.57	0.60	0.75	0.84	0.74	0.74	0.78
n	0.54	0.66	0.73	0.71	0.60	0.47	0.37	0.30	0.26	0.28	0.33	0.42	0.75	1	0.78	0.62	0.51	0.57	0.78	0.80	0.69	0.76
o	0.34	0.44	0.57	0.69	0.75	0.69	0.57	0.44	0.34	0.28	0.25	0.27	0.60	0.78	1	0.76	0.57	0.51	0.71	0.82	0.71	0.75
p	0.29	0.30	0.36	0.46	0.56	0.67	0.72	0.67	0.56	0.46	0.36	0.30	0.57	0.62	0.76	1	0.76	0.62	0.72	0.80	0.80	0.78
q	0.34	0.27	0.25	0.28	0.34	0.44	0.57	0.69	0.75	0.69	0.57	0.44	0.60	0.51	0.57	0.76	1	0.78	0.71	0.71	0.82	0.75
r	0.54	0.42	0.33	0.28	0.26	0.30	0.37	0.47	0.60	0.71	0.73	0.66	0.75	0.57	0.51	0.62	0.78	1	0.78	0.69	0.80	0.76
s	0.57	0.55	0.53	0.50	0.47	0.45	0.44	0.45	0.47	0.50	0.53	0.55	0.84	0.78	0.71	0.72	0.71	0.78	1	0.89	0.89	0.94
t	0.47	0.50	0.53	0.56	0.57	0.56	0.53	0.49	0.46	0.44	0.43	0.44	0.74	0.80	0.82	0.80	0.71	0.69	0.89	1	0.88	0.93
u	0.47	0.44	0.43	0.44	0.46	0.49	0.53	0.56	0.57	0.56	0.53	0.50	0.74	0.69	0.71	0.80	0.82	0.80	0.89	0.88	1	0.93
z	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.78	0.76	0.75	0.78	0.75	0.76	0.94	0.93	0.93	1

Table 2: Four subsets of reference varieties:

	1) Periphery Reference						2) Intermediary Reference						3) Center Reference				4) Even-sampled Reference						
Variety	a	c	e	g	i	k	m	n	o	p	q	r	s	t	u	z	a	e	i	n	p	r	z
a	1	0.50	0.14	0.01	0.14	0.50	0.72	0.54	0.34	0.29	0.34	0.54	0.57	0.47	0.47	0.50	1	0.14	0.14	0.54	0.29	0.54	0.50
b	0.74	0.74	0.29	0.04	0.04	0.29	0.67	0.66	0.44	0.30	0.27	0.42	0.55	0.50	0.44	0.50	0.74	0.29	0.04	0.66	0.30	0.42	0.50
c	0.50	1	0.50	0.14	0.00	0.13	0.56	0.73	0.57	0.36	0.25	0.33	0.53	0.53	0.43	0.50	0.50	0.50	0.00	0.73	0.36	0.33	0.50
d	0.30	0.74	0.74	0.30	0.04	0.04	0.46	0.71	0.69	0.46	0.28	0.28	0.50	0.56	0.44	0.50	0.30	0.74	0.04	0.71	0.46	0.28	0.50
e	0.14	0.50	1	0.50	0.13	0.00	0.36	0.60	0.75	0.56	0.34	0.26	0.47	0.57	0.46	0.50	0.14	1	0.13	0.60	0.56	0.26	0.50
f	0.04	0.29	0.74	0.74	0.29	0.04	0.30	0.47	0.69	0.67	0.44	0.30	0.45	0.56	0.49	0.50	0.04	0.74	0.29	0.47	0.67	0.30	0.50
g	0.01	0.14	0.50	1	0.50	0.14	0.29	0.37	0.57	0.72	0.57	0.37	0.44	0.53	0.53	0.50	0.01	0.50	0.50	0.37	0.72	0.37	0.50
h	0.04	0.04	0.29	0.74	0.74	0.29	0.30	0.30	0.44	0.67	0.69	0.47	0.45	0.49	0.56	0.50	0.04	0.29	0.74	0.30	0.67	0.47	0.50
i	0.14	0.00	0.13	0.50	1	0.50	0.36	0.26	0.34	0.56	0.75	0.60	0.47	0.46	0.57	0.50	0.14	0.13	1	0.26	0.56	0.60	0.50
j	0.30	0.04	0.04	0.30	0.74	0.74	0.46	0.28	0.28	0.46	0.69	0.71	0.50	0.44	0.56	0.50	0.30	0.04	0.74	0.28	0.46	0.71	0.50
k	0.50	0.13	0.00	0.14	0.50	1	0.56	0.33	0.25	0.36	0.57	0.73	0.53	0.43	0.53	0.50	0.50	0.00	0.50	0.33	0.36	0.73	0.50
l	0.74	0.29	0.04	0.04	0.29	0.74	0.67	0.42	0.27	0.30	0.44	0.66	0.55	0.44	0.50	0.50	0.74	0.04	0.29	0.42	0.30	0.66	0.50
m	0.72	0.56	0.36	0.29	0.36	0.56	1	0.75	0.60	0.57	0.60	0.75	0.84	0.74	0.74	0.78	0.72	0.36	0.36	0.75	0.57	0.75	0.78
n	0.54	0.73	0.60	0.37	0.26	0.33	0.75	1	0.78	0.62	0.51	0.57	0.78	0.80	0.69	0.76	0.54	0.60	0.26	1	0.62	0.57	0.76
o	0.34	0.57	0.75	0.57	0.34	0.25	0.60	0.78	1	0.76	0.57	0.51	0.71	0.82	0.71	0.75	0.34	0.75	0.34	0.78	0.76	0.51	0.75
p	0.29	0.36	0.56	0.72	0.56	0.36	0.57	0.62	0.76	1	0.76	0.62	0.72	0.80	0.80	0.78	0.29	0.56	0.56	0.62	1	0.62	0.78
q	0.34	0.25	0.34	0.57	0.75	0.57	0.60	0.51	0.57	0.76	1	0.78	0.71	0.71	0.82	0.75	0.34	0.34	0.75	0.51	0.76	0.78	0.75
r	0.54	0.33	0.26	0.37	0.60	0.73	0.75	0.57	0.51	0.62	0.78	1	0.78	0.69	0.80	0.76	0.54	0.26	0.60	0.57	0.62	1	0.76
s	0.57	0.53	0.47	0.44	0.47	0.53	0.84	0.78	0.71	0.72	0.71	0.78	1	0.89	0.89	0.94	0.57	0.47	0.47	0.78	0.72	0.78	0.94
t	0.47	0.53	0.57	0.53	0.46	0.43	0.74	0.80	0.82	0.80	0.71	0.69	0.89	1	0.88	0.93	0.47	0.57	0.46	0.80	0.80	0.69	0.93
u	0.47	0.43	0.46	0.53	0.57	0.53	0.74	0.69	0.71	0.80	0.82	0.80	0.89	0.88	1	0.93	0.47	0.46	0.57	0.69	0.80	0.80	0.93
z	0.50	0.50	0.50	0.50	0.50	0.50	0.78	0.76	0.75	0.78	0.75	0.76	0.94	0.93	0.93	1	0.50	0.50	0.50	0.76	0.78	0.76	1

REFERENCES

ISF Guidelines for Handling Disputes on Essential Derivation of Maize Lines. (2014). Retrieved October 10, 2017, from:
http://www.worldseed.org/wp-content/uploads/2015/10/ISF_Guidelines_Disputes_EDV_Maize_2014.pdf

Nelson, P.T., F. Achard, M. Butruille, and S. Madjarac. 2014. The use of Reference Varieties in Varietal Distinctness: an Approach under Investigation in the United States of America for Potential Application in Plant Variety Protection. Retrieved October 10, 2017, from
http://www.upov.int/meetings/en/doc_details.jsp?meeting_id=34524&doc_id=287358

Van Ettekoven K., 2016. Can Molecular Distance be used as Characteristic. Retrieved October 10, 2017, from http://www.upov.int/meetings/en/doc_details.jsp?meeting_id=39504&doc_id=341958

Markowitz, H. M., Todd, G. P., & Sharpe, W. F. (2000). *Mean-variance analysis in portfolio choice and capital markets* (Vol. 66). John Wiley & Sons.

Paul T. Nelson and Sambarta Dasgupta
Monsanto Company

[End of document]