E

UPOV

BMT/12/14
ORIGINAL: English
DATE: April 14, 2010

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA PROFILING IN PARTICULAR

Twelfth Session
Ottawa, Canada, May 11 to 13, 2010

STANDARDS FOR HELPING TO DETERMINE EDV STATUS IN MAIZE (ZEA MAYS L.) USING SSR'S AND FUTURE PROSPECTS USING SNP'S

Document prepared by experts from the United States of America

STANDARDS FOR HELPING TO DETERMINE EDV STATUS IN MAIZE (ZEA MAYS L.) USING SSRS AND FUTURE PROSPECTS USING SNPS

Barry Nelson[1], Elizabeth Jones[1], Alex Kahler[2], Jonathan Kahler[2], Steve Thompson[3], Ron Ferriss[4], Mark Mikel[5] and Stephen Smith[1]

[1]Pioneer Hi-Bred, 7300 NW62nd Ave., Johnston, Iowa, 50131
[2]Biogenetic Services, Inc., 801 32nd Ave., Brookings, SD 57006
[3]Dow AgroSciences LLC, 9330 Zionsville Road, Indianapolis, IN 46268-1054
[4]Syngenta, 7500 Olson Memorial Highway, Golden Valley, MN 55427
[5]Dept. of Crop Sciences and Roy J. Carver Biotechnology Center, 2608 Institute for Genomic Biology, University of Illinois, 1206 W. Gregory Dr., Urbana, IL 61801

Update: Selection and evaluation of panels of Simple Sequence Repeat (SSR) loci

1. The American Seed Trade Association (ASTA) has identified a set of simple sequence repeat (SSR) loci that can help in the determination of essentially derived varieties (EDV) status in maize (*http://crop.scijournals.org/content/vol50/issue2_Pages 486-503.*)

2. A final set of 285 SSR loci was selected based on informativeness (expected heterozygosity), genome distribution, and scoring ease of the markers that can be run on an inexpensive agarose gel based system. In order to gain efficiency in use of the SSR panel a subset of 150 SSR loci was selected using the same criteria to select the 285 set. The 150 SSR set was designated core set 1 and the additional 135 could be used if necessary. The complete 285 SSR panel is available on the ASTA web site (*http://www.amseed.org/news_srr.asp*).

3. The French Maize Breeders (*Union Française des Semenciers* (UFS), formerly SEPROMA) has also designated a core set of 163 SSRs to help determine EDV status. These SSRs have been selected to be run on a more discriminative capillary sequencing system.

Evaluation of Single Nucleotide Polymorphisms (SNPs) for genetic similarity in maize (*Zea mays L.)*

Introduction

4. Laboratories are now moving to the use of bi-allelic SNP markers in maize. It is therefore necessary to begin an evaluation of SNP loci for their potential utility in helping to determine EDV status. Initial issues to be addressed include the numerical range of numbers of SNPs required, and selection criteria such as level of informativeness and genomic coverage. We report upon those here.

Materials and Methods

5. Aliquots of the same DNA from the 98 inbreds profiled with the 285 ASTA SSRs were profiled with a 768 set of public SNPs previously reported on by Jones *et al.*, (2009). Quality control (QC) was performed on the data set and markers with >10% heterozygotes or scored in <70% of the inbreds were removed. Inbreds were removed if they contained

>10% heterozygotes or had <70% of the SNP loci scored. These QC steps resulted in an initial SNP set of 601 loci with 80 of the 98 ASTA inbreds and 26 of the 30 SEPROMA inbreds. Further subsets of the 601 SNP loci were selected to evaluate the impact of SNP number of coefficient of variation and similarity. The basis for SNP selection was to retain a high level of informativeness (expected heterozygosity) and also to retain even genome coverage. The second subset removed any loci not mapped or not informative resulting in a set of 447 loci among the 80 inbreds. A 306 subset was selected from the 447 set by removing markers mapped to the same location while retaining the most informative markers. Subsequent sets of 204, 83, and 42 loci were selected while maintaining or increasing mean expected heterozygosity and, although with increased mean genetic distance between loci, maintaining even genome coverage.

6. We utilized the coefficient of variation (CV) comparison reported by Van Inghelandt *et al.* (2010) for comparison of marker type and number. For a further basis of comparison to the SNP sets selected for informativeness and genome coverage, random SNP sets of the same size were selected beginning with the 601 set. Each marker set was bootstrapped 1,000 times using the Resample module in NTSYSpc version 2.21 (Rohlf, 2009). Roger’s distances were calculated in the Genetic Distance module, mean and standard deviation of bootstrapped Roger’s distances were computed in the Summary module, then extracted and CVs computed for each inbred pair. CVs of each inbred pair were then averaged across each marker set for comparison.

7. To assess correlations between marker data sets and genetic conformity, Roger’s distance matrices were converted to inbred pair format to compare relationships between pedigree relatedness, SSR, and SNP distance data sets. Cluster analysis on each data set was performed using the un-weighted pair group method with arithmetic average (UPGMA) in the SAHN module of NTSYSpc.

Results

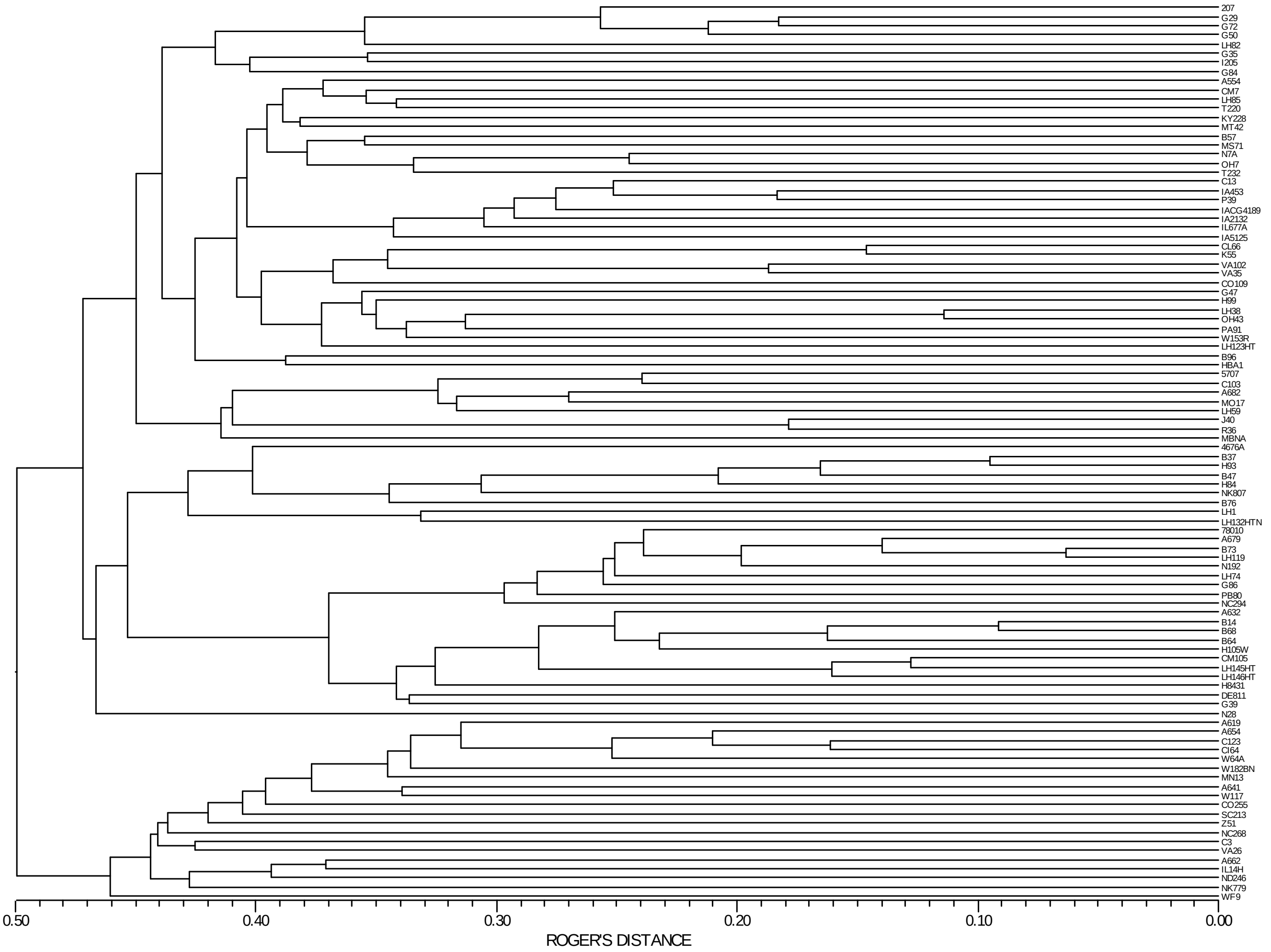
8. Expected heterozygosities generally increased from 0.35 for the 601 SNP set to 0.48 for the 83 and 42 sets (Table 1). Results were similar for the 26 subset of inbreds previously selected by SEPROMA for the SSR study (Table 1).

Table 1. Expected heterozygosities among 80 inbreds and the 26 inbred subset.

Marker Set	Average (range) expected heterozygosity for 80 inbred set	Average (range) expected heterozygosity for 26 inbred sub-set
150 ASTA SSR	0.53 (0.03 – 0.79)	0.49 (0.00 – 0.80)
163 SEPROMA SSR	NA	0.65 (0.00 – 0.88)
601 SNP	0.35 (0.00 – 0.50)	0.34 (0.00 – 0.50)
447 SNP	0.38 (0.03 - 0.50)	0.37 (0.00 - 0.50)
306 SNP	0.41 (0.03 – 0.50)	0.39 (0.00 – 0.50)
204 SNP	0.41 (0.03 – 0.50)	0.40 (0.00 – 0.50)
83 SNP	0.48 (0.30 – 0.5)	
42 SNP	0.48 (0.34 – 0.5)	

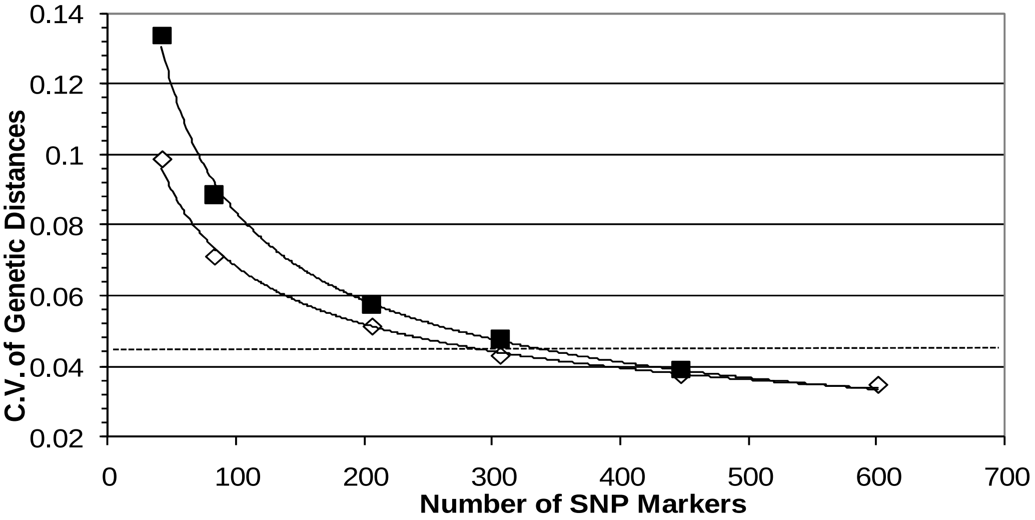
9. Cluster analysis with SSR sets showed associations consistent with known pedigrees (Kahler, *et al*., 2010). With the 306 SNP set, all inbreds were clearly distinguished and revealed similar associations (Figure 1).

Figure 1. Associations among inbred lines revealed the following multivariate analysis of Roger's distances computed from the 306 SNP subset.



10. CV values for each SNP set are shown in Figure 2. As the number of SNPs in a set is increased, then, with these 80 inbreds, CVs decreased with the lowest value at 601 SNPs. Subsets of SNPs were selected for both high informativeness and even genome coverage. These subsets resulted in lower CVs than the randomly chosen sets of SNPs; differences became more marked as the number of SNPs further decreased. Results were very similar using the 26 inbred subset (data not shown).

Figure 2. CVs of genetic distances for sets of SNPs with sub-sets being selected for genome distribution and high informativeness (open diamonds) or randomly selected (closed squares). CV values for two standard SSR sets of markers are represented by dashed lines: the ASTA set of SSRs (CV of 0.057) with short dashes, and the SEPROMA set of SSRs (CV of 0.048) with long dashes.



11. Correlations between Roger’s distances of the various SSR and SNP sets are shown in Table 2. Data are only reported for the data sets that had CV values performing at or better than those of the SSR sets, thus only the 204 and greater sets were examined for correlations. We focused on more related materials by limiting pairs to those with at least a 0.25 coefficient of pedigree relatedness (maximum = 1.0). The highest correlation among the 80 inbreds for the ASTA 150 core set of SSRs was with the 306 SNP set. The highest correlations among the 26 inbred subset for the ASTA core set of SSRs was with the 204 SNP set, although very similar to the 306 set. The SEPROMA 163 SSR core set was best correlated with the 204 SNP set. In general, correlations were consistently higher for the SEPROMA SSR core set than for the ASTA SSR core set.

Table 2. R^2 (coefficient of determination) values for correlations of inbred-inbred genetic distances calculated with different marker sets. Only inbred pairs with > 0.25 pedigree relatedness (Malecot's) were included in the analysis. R^2 values for the 80 inbred analyses are in the upper right of the matrix and for the 26 inbred sub-set are in the lower left of the matrix. Only the 26 inbred sub-set was profiled with the 163 SEPROMA SSR set; information is therefore not available for correlations involving 80 inbreds for this set of SSRs.

	MALECOT	150 ASTA SSR	163 SEPROMA SSR	601 SNP	447 SNP	306 SNP	204 SNP
MALECOT		0.38	NA	0.44	0.46	0.46	0.46
150 ASTA SSR	0.42		NA	0.55	0.58	0.61	0.59
163 SEPROMA SSR	0.66	0.80		NA	NA	NA	NA
601 SNP	0.49	0.56	0.75		0.98	0.96	0.92
447 SNP	0.52	0.61	0.80	0.98		0.98	0.95
306 SNP	0.53	0.65	0.82	0.97	0.99		0.97
204 SNP	0.56	0.66	0.82	0.95	0.97	0.98	

Discussion

12. ASTA and SEPROMA have designated core sets of SSRs that can be used to help determine EDV status.

13. Laboratories are now moving to the use of bi-allelic SNPs. It is therefore necessary to begin an evaluation of SNP loci for their potential utility in helping to determine EDV status. Initial issues include determining a numerical range of numbers of SNPs that will be required and criteria such as level of informativeness and genomic coverage. Only then can the next issue of determining SNP based thresholds of genetic similarity be appropriate to help determine EDV status be usefully examined.

14. In order to help contribute to the first issue we have demonstrated that SNP profiles in maize meet two key criteria required for a marker system to be able to identify and to usefully characterize inbred lines. These are 1) high discrimination ability and 2) the ability to show associations which reflect known pedigree backgrounds.

15. An initial basis for comparison were genetic similarities and associations among inbred lines shown using the SSR sets designated by ASTA and by SEPROMA. The SNP sets that correlated most closely to either of these SSR sets were the 204 and 306 sets of SNPs. We also examined CV of genetic similarities in relation to the number of SNP loci that were used in those computations. We found that CV values decreased as the number of SNPs increased. The 150 ASTA and 163 SEPROMA SSR sets had Roger's distance CV values of 0.057 and 0.048 respectively, which were equivalent to distance CVs found for 204 and 306 SNP sets. In contrast, Van Inghelandt *et al.* found CV values based on modified Roger's distances using 150 SSRs to be lower at roughly 0.03, which was equivalent to genetic distance CVs around 800 SNPs. The difference in our observations and Van Inghelandt *et al.* could be related to informativeness as selection criteria of the SNP sets. Our observations show that SNP sets selected on the basis of both high informativeness and maintaining even genome coverage

showed an increasing advantage in CV value over randomly selected SNPs as SNP number per set declined from the 601 set. The overall utility and efficiency (in terms of laboratory use) of a SNP set improve when consideration is given to both high informativeness (expected heterozygosity) and maintaining even genome coverage as components of selection.

Future Experiments

16. Illumina has developed a SNP chip of 60,000 public SNPs. This chip provides very high density resolution in comparisons of inbreds. We anticipate that future studies to evaluate the role of genetic similarity data obtained from comparisons of SNP profiles to help determine EDV status in maize will utilize this SNP chip. These studies will likely include to compare SNP profiles of inbred lines already used in EDV studies, additional sets of very closely related inbred lines, and comparisons with SSR and pedigree data.

References

Jones, E.S., W.-C. Chu, M. Ayele, J. Ho, E. Bruggeman, K. Yourstone, A. Rafalski, O.S. Smith, M.D. McMullen, C. Bezawada, J. Warren, J. Babayev, S. Basu and S. Smith. 2009. Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. Mol Breeding. 24:165-176.

Kahler, A.L., J.L. Kahler, S.A. Thompson, R.S. Ferriss, E.S. Jones, B.K. Nelson, M.A. Mikel, and S. Smith. 2010. North American Study on Essential Derivation in Maize: II. Selection and Evaluation of a Panel of Simple Sequence Repeat Loci. Crop Sci 50: 486-503.

Rohlf, F. J. 2009. NTSYS-pc, Numerical Taxonomy and Multivariate Analysis System, version 2.2 Exeter software, Setauket, New York, USA.

Van Inghelandt, D., A.E. Melchinger, C. Lebreton, and B. Stich. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet. 2010 Jan 10; [Epub ahead of print].

[End of document]