E

UPOV

BMT/12/15
ORIGINAL: English
DATE: April 15, 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

VARIETAL IDENTIFICATION IN MAIZE: ARE SIXTEEN SNP MARKERS SUFFICIENT?

Document prepared by experts from the United States of America

VARIETAL IDENTIFICATION IN MAIZE: ARE SIXTEEN SNP MARKERS SUFFICIENT?

Liz Jones, Steve Wall, Barry Nelson, Stephen Smith
Pioneer Hi-Bred, 7300 NW62nd Ave., Johnston, Iowa, 50131

Abstract

1. Molecular marker profiles can help assure varietal identity, to monitor genetic purity and to assist in obtaining intellectual property protection. A small number of highly informative markers that are amenable to high-throughput laboratory conditions would have practical advantages in providing a means of varietal identification that is both cost-effective and rapid. We have identified a set of 16 single nucleotide polymorphism (SNP) loci that provides marker profiles that are highly discriminative amongst proprietary and competitor maize inbred lines and which are adapted for use in the United States of America (US) or in Europe. The discriminative power of these SNP loci is an order of magnitude higher than that previously possible using data obtained from isozyme loci. The discriminative power of the 16 SNP loci was robust in the face of up to 50% missing data or 25% mis-scored data. If a marker platform could be established with high levels of data coverage (i.e. low levels of missing data), 8 markers may well be quite adequate for the purposes of variety identification where the goal is to rapidly identify the vast majority of the germplasm being tested. However, if very closely related inbreds need to be routinely separated, much larger marker numbers are needed, in the range of at least 100. This large increase in marker numbers produces just small gains in the efficiency of discrimination. In the proprietary inbred set tested, 16 SNP markers could identify 99.96% of the inbreds, 42 SNP markers could identify 99.99% and 165 SNP markers could uniquely identify 100% of inbreds.

Introduction

2. Molecular markers are now widely used to identify plant varieties (ISF, 2009: Guiard, 2007; UPOV BMT 2008a; 2008b; 2008c;2008d) and to monitor genetic purity (Staub, 1999; Nandakumar et al., 2004; Liu et al., 2007; Rana et al., 2007; Tsukazaki et al., 2008). It is important to utilize marker systems that are reflective of genotype, highly discriminative, reliably scorable, amenable to high-throughput analysis and cost-effective (Gale et al., 2005).

3. The first widely used marker systems were provided by the electrophoretic separation of isozymes or seed storage proteins (Smith and Wych, 1986). Stuber and Goodman (1983) reported that 73% of 406 maize (*Zea mays* L.) inbreds could be uniquely identified by the use of 21 isozymic loci. Greater efficiency in varietal identification can be gained by identifying a small set of highly polymorphic markers that collectively provide great discrimination. For example, Galli et al. (2005) identified 4 SSR loci that collectively could be used to uniquely identify 66 commercial apple (*Malus domestica* Borkh.) varieties (except for somatic mutants of these varieties). Likewise, Prasad et al. (2000) uniquely identified 48 wheat (*Triticum aestivum* L.) varieties using 12 SSR loci, Coombs et al., (2004) uniquely identified 17 potato (*Solanum tuberosum* L.) varieties using combinations of 4 SSR primer pairs, and Reid and Kerr (2007) identified 6 SSR loci that could differentiate over 400 potato varieties, including those on the United Kingdom National List. Song et al. (1999) identified 13 SSR loci that could be used to uniquely identify 66 elite North American soybean (*Glycine max* [L.] Merr.) varieties, including several varieties that had identical maturity, morphological, or pigmentation traits. Govan et al., (2008) identified 10 SSR loci that collectively could provide unique discrimination among 60 genotypes of strawberry (*Fragaria* x*ananassa*), including among sibling varieties.

4. A greater number of SNP loci are likely to be required to provide equivalent levels of discrimination among varieties because of their bi-allelic nature. Nonetheless, impressive powers of unique identification can be reached using SNPs. For example, Gale et al., (2005) note that "In theory, as few as 12 such markers can separate up to 4006 (=2 power12) possible genotypes." However, they further note that in practice markers may not always segregate independently; consequently, in practice the theoretical power of discrimination may not be achieved. Nonetheless, Yoon et al., (2007) were able to identify a set of 23 SNP loci that can be used to provide an equivalent power of unique identification as did 13 SSR loci. Likewise, Shirasawa et al. (2006) were able to identify 8 SNP loci that could be used to uniquely identify 43 Japanese rice varieties.

5. Different approaches have then been used to identify a minimum set of loci that collectively can provide unique identification for a large number of varieties. For example, Gale et al., (2005) describe an integer linear approach and Song et al (1999) describe a multivariate approach. In contrast, we have chosen to evaluate yet another approach: We utilized a genetic algorithm that was initially developed to solve the "travelling salesman problem" (Kruskal, 1956). The objectives of this paper, therefore, are: 1) to report upon the results obtained from using the genetic algorithm approach; 2) to determine robustness in the face of missing or mis-scored data; and 3) to determine the gains achieved with using more than a small number of markers.

Materials and Results

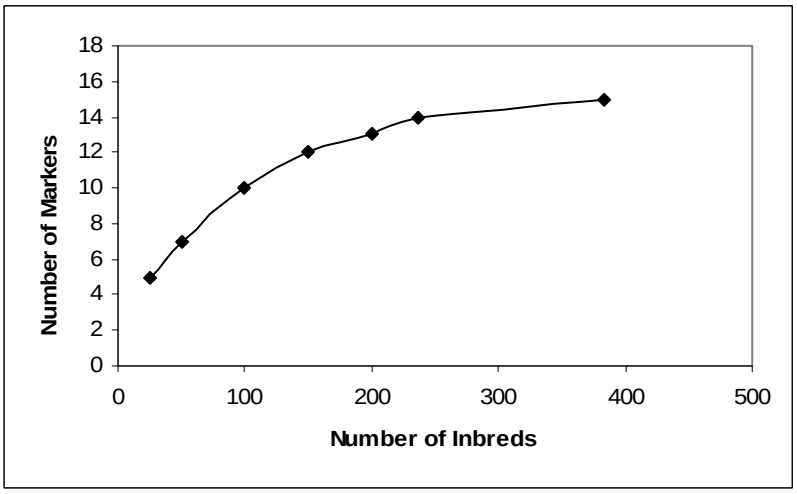
a) Determining the minimum number of SNPs needed to distinguish among 383 proprietary maize inbreds

6. A genetic algorithm was used to find the minimum number of markers required to uniquely distinguish among a set of inbreds. The genetic algorithm was developed to solve the 'Travelling salesman problem', where for a given set of cities, each city needs to be visited once and the overall travelling distance minimized. The algorithm essentially randomly places a marker into a set and then determines whether discrimination power has been improved, or not, when compared to the previous best combination of markers. The process is repeated thousands of times.

7. From an initial pool of 491 proprietary SNP markers and 383 Pioneer-proprietary maize inbred lines that are individually widely used in hybrid production in the US or Europe (primarily France, Germany and Italy), the genetic algorithm was run on 25, 50, 100, 150 and 200 randomly selected inbreds as well as on the complete set of 383 inbreds.

8. For 100 inbreds, 10 SNPs could uniquely identify each inbred, and for 200 inbreds 13 SNPs could uniquely identify each inbred (Figure 1). Beyond 200 inbreds, the number of SNPs required for unique identification leveled out so that for 383 inbreds, 15 loci were sufficient and, by extrapolation, no more than 16 SNP loci would be needed to discriminate among 500 inbreds.

Figure 1. The minimum number of SNP markers (from a set of 491) that together could uniquely identify a given set of inbreds.



9. Multiple combinations of 16 SNPs were found that could be used to completely discriminate each of 383 inbreds. Sets of 16 SNPs were therefore selected to enable some flexibility in the markers incorporated into the set, and because 16 SNPs allows for efficiency within the micro-titer plate formats required for high-throughput laboratory conditions. Consequently, six sets of 16 SNPs were selected based on the criteria of generating marker profiles that were completely discriminative for each of the 383 inbreds and also that they, collectively, allowed each maize chromosome to be sampled by at least one SNP. These sets of 16 SNPS were then tested under high-throughput conditions and the set that gave the highest data quality was selected for further use.

b) Comparison of SNPs to isozymes in real-life purity studies

10. The 16 SNP markers were compared with 15 isozymic loci (Acp1, Adh1, Idh1, Idh2, Mdh1, Mdh2, Mdh3, Mdh4, Mdh5, Mmm, Pgd1, Pgd2, Pgm1, Pgm2 and Phi1) in real-life genetic purity laboratory tests on a range of inbreds and hybrids. These isozyme loci were chosen because they have been routinely used within the US maize seed industry (Smith and Wych, 1986) to assay for genetic purity in both inbreds and hybrids.

11. Replicated individual plant samples (between 15 and 143 replicates) for 10 inbreds were analyzed in a side-by-side study with 15 isozymes and 16 SNP markers. SNPs were found to have a higher level of missing data at 2% compared with isozymes at 0.8%. The profiles were compared to 212 inbreds that had complete data for both the 16 SNPs and the 15 isozymes. A resolution score was assigned to each inbred with a score of 1 indicating complete resolution i.e. the only matching profile is to itself, and decreasing values indicate decreasing resolution power. The overall resolution score for SNPs was found to be 16 times that for isozymes.

Table 1. Overall information for real-life purity tests with marker sets.

Marker Type	Total Data	Missing Data	% missing data	Overall Resolution Score
SNPs	9120	185	2%	0.96
Isozymes	8550	7	0.8%	0.06

c) Analysis of competitor inbreds with expired PVP protection using the 16 SNP markers

12. SNP data was also collected for 58 competitor inbreds for which PVP protection had expired. Out of 1711 pairs, 1702 (99.5%) could be distinguished. The 9 pairs that could not be distinguished were also profiled with at least 400 SNPs that covered the whole genome and with these had an average of 84% similarity (range 78-92% similar). The pairs could be separated into 4 groups: (1) Cargill 2369 and Holden's LH149 had the same 16 SNP profile and both have B73 as a parent (2) Novartis inbreds 807 and 778 were both developed from W117 / B37 Ht (3) Holden's LH143 and LH145 were both developed from A632 Ht (4) Holden's LH51, LH54, LH65 and Novartis S8326 all could not be separated from each other and all had Mo17 in their pedigrees.

d) Testing for robustness with missing and mis-typed SNP data

13. A set of 438 inbred lines was chosen on the basis that they had been awarded PVP protection in the US and/or Europe. Collectively, the inbred set encompassed a broad and representative array of maize germplasm that is in use today in maize breeding and agriculture in the US or in Europe. Complete isozymic data for 15 loci was available for these inbreds and no heterozygotes were recorded. In contrast, for the 16 SNPs, there were some missing data (average of 3% on an inbred basis, range of 0 to 31%) and heterozygous data (average of 1% on an inbred basis, range of 0 to 38%). A set of 8 candidate inbred lines was selected to compare the distinguishing power and robustness of the 15 isozymic and 16 SNP profiles. The candidate inbreds were selected to represent the broad array of germplasm currently used in the US and because they had homozygous allele data recorded at each SNP and isozymic locus. Their pedigree backgrounds (numbers of inbreds in parentheses) were as follows: Stiff Stalk (3); Non Stiff-Stalk Iodent (2); Non Stiff-Stalk and non Iodent (2).

14. Robustness of the SNP profile data for each of the candidate inbred lines was evaluated by simulating missing data and mis-scored data for each candidate inbred line for all possible combinations of 1, 2, 4, and 8 SNP loci before computing profile matches to the validation set of 438 inbreds. We also computed the mean Malecot Coefficient of Parentage value for all inbreds that matched at each level of missing or mis-scored data for each candidate inbred line. Evaluation of discrimination power and robustness of isozymic data were performed in a similar fashion except that simulations of missing and mis-scored data were made with all possible combinations of 1, 2, 4, and 6 isozymic loci.

15. As either missing or mis-scored data were added to the profiles of candidate inbred lines then the number of inbreds with matching profiles in the validate set inflated rapidly for isozymes, especially above 2 (15% loci), and most especially for mis-scored data (Figure 2). The SNP16 marker set collectively exhibited a high ability to discriminate the candidate inbred from among the 437 other inbred lines in the validation set, even in the face of either missing or mis-scored data at up to 4 (25%) of these SNP loci (Figure 2).

16. Examination of results presented in Figure 3 shows robustness of the SNP16 and isozymic data in respect of the average degree of similarity by pedigree of matching inbreds. At each level of error, mis-scored SNP data are less likely to result in false matches, and those false matches are more related by pedigree to the candidate inbred, compared to mis-scored isozyme data.

Figure 2. Number of inbreds matching to candidate inbreds in the face of missing or mis-typed data for isozyme and for SNP16.

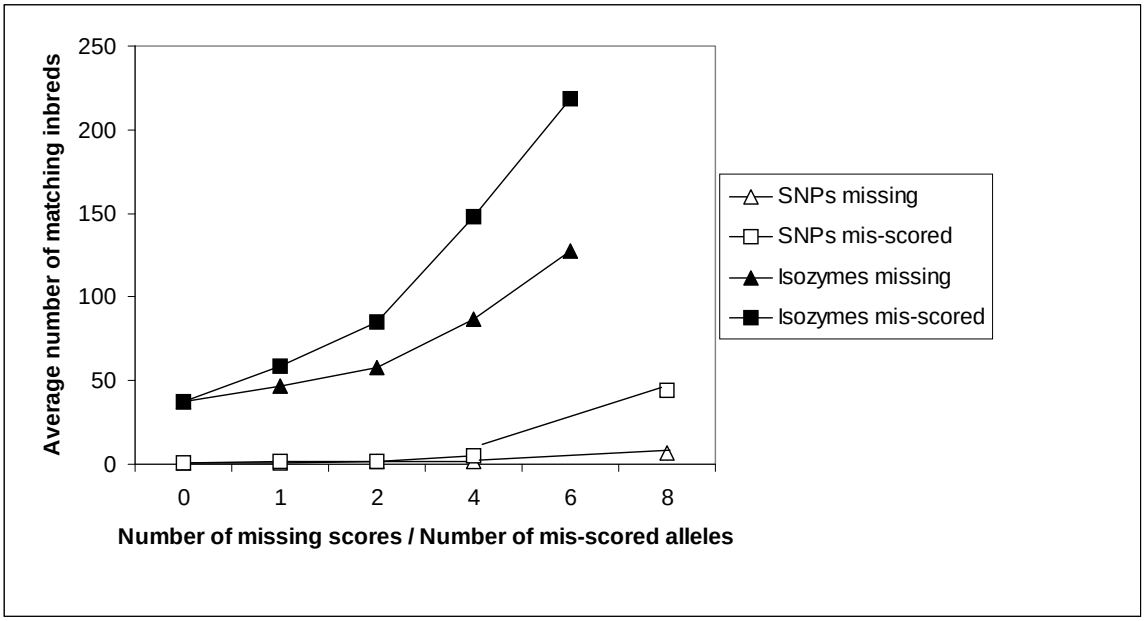
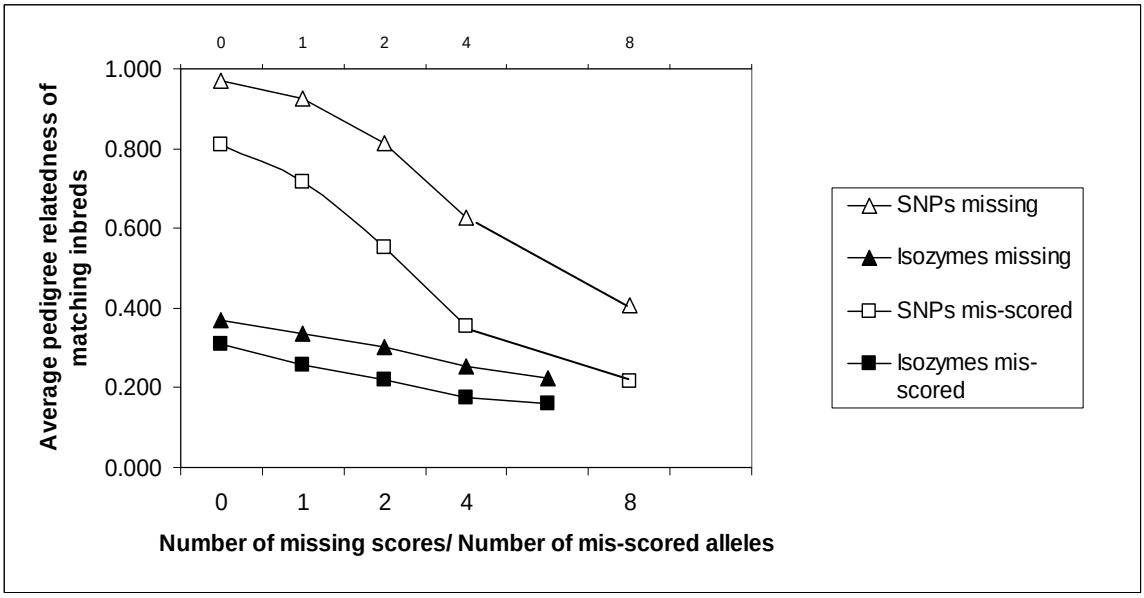


Figure 3. Mean Malecot Coefficient of Parentage for inbreds that match a candidate inbred line in the face of missing or mis-typed data from isozymic or for SNP16 profiles.



17. False matches that occurred as a result of errors in SNP scoring occurred with inbred lines that were far more related by pedigree to the true candidate inbred line than was the case for matches shown by isozymic data. The SNP16 loci had a polymorphic information content (PIC) of 0.47 whereas that for the isozymic loci was 0.17. Thus, although isozymes can reveal a large number of alleles (>20 for some loci e.g. B-GLU, PGM2) when observed across a diverse range of germplasm (e.g. races of maize from Mexico), the array of isozymic allelic diversity that is expressed across a more restricted germplasm diversity (e.g. North American inbred lines) is more often restricted to 2-3 alleles where a single allele is often found in greatest abundance. In contrast, SNP loci can be found in greater abundance than can isozymic loci thereby facilitating the selection of loci with relatively high PIC from within

even a relatively restricted array of maize germplasm. Second, the greater numerical array of SNP loci also facilitates selection of a set of those loci that collectively allow a more thorough sampling of the genome of maize than do the available isozymic loci that are also relatively high for their PIC.

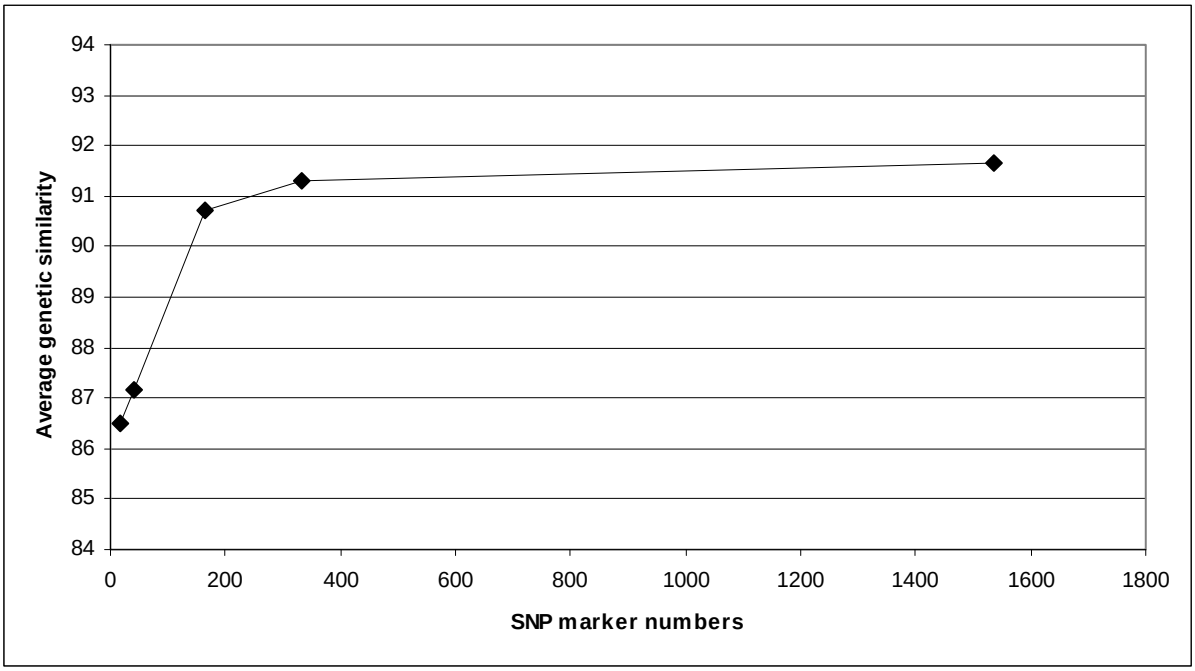
e) Gains in discrimination using larger numbers of SNP markers.

18. A total of 248 Pioneer inbreds with PVP protection were examined. The number of inbred pairs that could not be distinguished with different numbers of SNPs was determined. Using 16 SNP markers, 22 / 61256 pairs of inbreds could not be distinguished i.e. 99.96% pairs could be distinguished. Using 42 SNP markers, only 2 pairs could not be distinguished i.e. 99.99% pairs could be distinguished. Using 165 SNP markers, 100% of the inbreds could be distinguished.

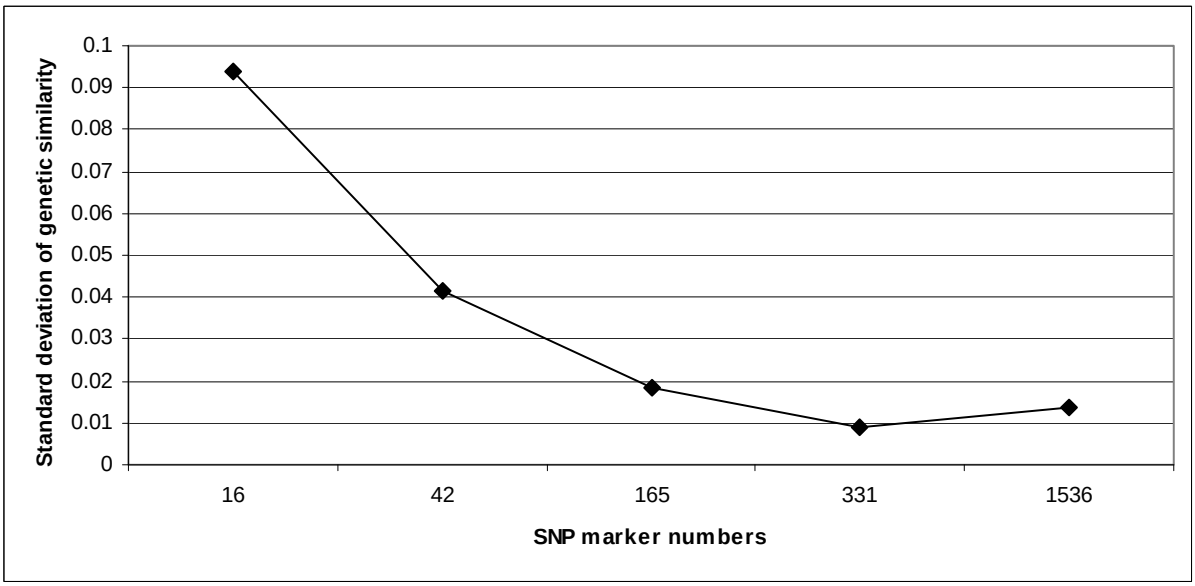
19. To assess how accurately small numbers of SNP markers could assess similarity among closely related inbreds, a set of 331 SNP markers selected to give good genome coverage was used to assess genetic similarities among the 248 Pioneer inbreds. Related inbred pairs with genetic similarities >90% and >95% as assessed with the 331 SNP set were then compared with genetic similarities with SNP sets of different sizes; 1450, 165, 42, plus the 16 SNPs selected for genetic purity. Only 2 pairs of inbreds had genetic similarities >95% with 331 SNPs and these were also found to be >95% similar with 1450 SNPs, and 165 SNPs. However with the 42 and 16 SNP sets, genetic similarities fell below the 95% threshold, with these lower SNP numbers being unable to pick up the smaller chromosomal differences that separated these inbreds. For the inbred pairs >90% similar with 331 SNPs, mean similarities were fairly consistent for the 331, 1450 and 165 SNP sets, but were underestimated for the 42 and 16 SNP sets. Standard deviations were also much higher for the 42 and 16 SNP sets. The exercise was repeated comparing pedigree distances with SNP sets of different sizes and the same trend was found (data not shown). Therefore, smaller SNP sets will tend to underestimate similarity measures and are not ideal for separating closely related individuals, as small chromosomal differences will not be detected.

Figure 4. (a) Average genetic similarities and (b) standard deviations determined with different sets of SNPs as compared with genetic similarities determined with 331 SNP markers.

(a)



(b)



References

Coombs JJ, Frank LM, Douches DS (2004) An applied fingerprinting system for cultivated potato using simple sequence repeats Amer. J. of Potato Res. 81:243-250.

Gale K, Jiang H, Westcott M (2005) An optimization method for the identification of minimal sets of discriminating gene markers: Application to cultivar identification in wheat. Jour. Bioinformatics and Computational Biology 3:269-279.

Galli Z, Halasz G, Kiss E, Heszky L, Dobranszki J (2005) Molecular identification of commercial apple cultivars with microsatellite markers. HortScience 40:1974-1977.

Govan CL, Simpson DW, Johnson AW, Tobutt KR, Sargent DJ (2008) A reliable multiplexed microsatellite set for genotyping *Fragaria* and its use in a survey of 60 *F.* x *ananassa* varieties Mol. Breeding 22:649-661.

Guiard (2007) Identification methods for protected plant material. GEVES La Miniere France 9 pp.

ISF (2009) ISF view on intellectual property ISF Nyon, Switzerland 15 pp. *http://www.worldseed.org/cms/medias/file/PositionPapers/OnIntellectualProperty/View_on_Intellectual_Property_2009.pdf*

Kruskal JB (1956) On the shortest spanning subtree of a graph and the travelling salesman problem. Proc. Amer. Math. Soc. 7:48-50.

Liu L-W, Wang Y, Gong Y-Q, Zhao T-M, Liu G, Li X-Y, Yu F-M (2007) Assessment of genetic purity of tomato (Lycopersicon esculentum L.) hybrid using molecular markers. Scientia Horticulturae 115:7-12.

Nandakumar N, Siugh AK, Sharma RK, Mohapatra T, Prabhu KV, Zaman FU (2004) Molecular fingerprinting of hybrids and assessment of genetic purity of hybrid seeds in rice using microsatellite markers Euphytica 136:257-264.

Prasad M, Varshney RK, Roy JK, Balyan HS, Gupta PK (2000) The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. Theor. Appl. Genet. 100:584-592.

Rana MK, Singh S, Bhat KV (2007) RAPD, STMS, and ISSR markers for genetic diversity and hybrid seed purity testing in cotton Seed Sci. and Technol. 35:709-721.

Reid A, Kerr EM (2007) A rapid and simple sequence repeat (SSR)-based identification method for potato cultivars. Plant Genetic Resources: Characterisation and Utilisation 5:7-13.

Shirasawa K, Shiokai S, Yamaguchi M, Kishitani S, Nishio T (2006) Dot-blot-SNP analysis for practical plant breeding and cultivar identification in rice. Theor. Appl. Genet. 113:147-155.

Smith JSC, Wych (1986) The identification of female selfs in hybrid maize: A comparison using electrophoresis and morphology Seed Sci. and Technol. 14:1-8.

Song QJ, Quigley CV, Nelson RL, Cater TE, Boerma HR, Strachen JL, Cregan PB (1999) A selected set of trinucleotide simple sequence repeat markers for soybean cultivar identification. Plant Varieties and Seeds 12:207-220.

Staub JE (1999) Intellectual property rights, genetic markers, and hybrid seed production Jour. of New Seeds 1:39-64.

Stuber CW, Goodman MM (1983) Allozyme genotypes for popular and historically important inbred lines of corn, *Zea mays* L. USDA-ARS, ARR-S-16. U.S. Gov. Print. Office, Washington, DC.

Tsukazaki H, Nunome T, Fukuoka H, Ohara T, Song YS, Yamashita K, Wako T, Kojima A, Kanamori H, Kono I (2008) Applications of DNA marker technology in Japanese bunching onion breeding Acta Horticulturae 770:153-158.

Document BMT/11/2 ‘Report on Developments in UPOV Concerning Biochemical and Molecular Techniques’. Eleventh session of the Working Group on Biochemical and Molecular Techniques, and DNA Profiling in Particular, held in Madrid from September 16 to 18, 2008.

Document BMT/11/13 ‘The Spanish Experience (GESLIVE-IRTA) on the Enforcement of Plant Variety Rights: DNA-Fingerprinting’. Eleventh session of the Working Group on Biochemical and Molecular Techniques, and DNA Profiling in Particular, held in Madrid from September 16 to 18, 2008.

Document BMT/11/15 ‘Preparation of Guideline for Method Validation of DNA Identification for the Enforcement of Plant Breeder’s Rights in Japan’. Eleventh session of the Working Group on Biochemical and Molecular Techniques, and DNA Profiling in Particular, held in Madrid from September 16 to 18, 2008.

Document BMT/11/20 ‘A Practical Example of the Possible Use of Molecular Techniques in Variety Identification’. Eleventh session of the Working Group on Biochemical and Molecular Techniques, and DNA Profiling in Particular, held in Madrid from September 16 to 18, 2008.

Yoon MS, Song QJ, Choi IY, Specht JE, Hyten DL, Cregan PB (2007) BARCSoySNP23: A panel of 23 selected SNPs for soybean cultivar identification.

[End of document]