E

BMT-TWA/Maize/2/5
ORIGINAL: English
DATE: November 9, 2007

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

AD HOC CROP SUBGROUP ON MOLECULAR TECHNIQUES FOR MAIZE

Second Session
Chicago, United States of America, December 3, 2007

SELECTION OF SIXTEEN SINGLE NUCLEOTIDE POLYMORPHISM (SNP) MARKERS FOR VARIETAL IDENTIFICATION USING A GENETIC ALGORITHM APPROACH IN MAIZE INBREDS

Document prepared by experts from Pioneer Hi-Bred International

Abstract

We used a genetic algorithm to select a combination of markers that could uniquely identify a given set of inbreds. Working with an initial set of approximately 500 high quality SNP markers, we determined that just 15-16 SNPs could uniquely identify a set of 383 Pioneer inbreds. In side-by-side studies, we found that 16 SNPs were 16 times more informative than 15 isozyme markers.

Data were collected on 309 inbreds from the United States of America (US) and 192 from European Pioneer inbreds, with Plant Variety Protection (PVP). The 16 SNP markers were found to uniquely identify over 99% of the inbred pairs. We conclude that a small number of carefully selected markers can together create a highly informative genetic fingerprint. Such a marker set could have great utility in the identification of varieties and for the management of germplasm collections.

Selection of Sixteen Single Nucleotide Polymorphism (SNP) markers for Varietal Identification Using a Genetic Algorithm Approach in Maize Inbreds

Liz Jones, Ken Yourstone, Jennifer Jaqueth, Dave Spaulding, Don Cerwick, Todd Krone,
Dinakar Bhattramakkii, Barry Nelson, Stephen Smith
Pioneer Hi-Bred International, A DuPont Company, Johnston, Iowa 50131

Introduction

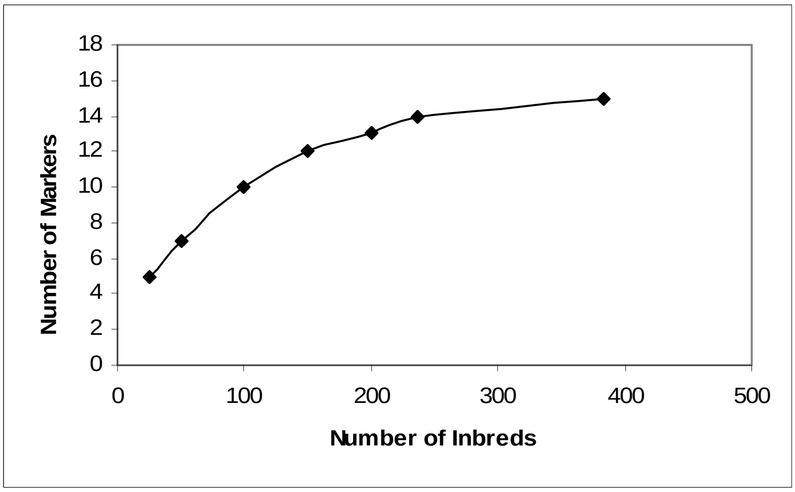
1. A genetic algorithm is a search technique used to find exact or approximate solutions to problems. Such an algorithm can be used to select the most informative set of markers that together can best distinguish between a given set of varieties. One key question is how many markers are adequate for variety identification? Many highly informative markers distributed across the whole genome would undoubtedly provide the highest level of distinction, but the cost to produce such a data set may be prohibitive. Single nucleotide polymorphism (SNP) markers are rapidly becoming the marker of choice for maize. We used the genetic algorithm approach to determine the minimum number of SNPs that would provide unique identification for the maximum number of varieties in a germplasm set.

Materials and Results

(a) How many SNP markers can uniquely identify inbred sets?

2. SNP markers were selected that had high distinguishing power among US and European maize germplasm and that produced good quality data under high throughput conditions. SNP data was initially available for 491 markers and 237 diverse US and European Pioneer inbreds. Subsequently, data became available for an additional 146 inbreds to give a total of 383 available inbreds. To determine how many markers could distinguish between different sub-sets of these inbreds, the genetic algorithm was run on 25, 50, 100, 150 and 200 randomly selected inbreds as well as the 237 and 383 inbreds. The minimum number of markers that could uniquely identify each inbred in the set was determined (Figure 1). For 383 inbreds, a set of 15 markers were found that could uniquely identify each individual.

Figure 1. The number of SNP markers [from a set of 491] that together uniquely identified a given set of inbreds.



3. Sets of 16 SNPs were subsequently selected and further tested. Selecting sets of 16 SNPs (rather than 15) gave us many markers sets to choose from and enabled a certain amount of redundancy due to help assuage the effects of missing data. Six sets of 16 markers with good distribution across the genome were tested for performance under high throughput production conditions and the best set was selected.

(b) How do SNP markers compare with isozyme markers for variety identification?

4. In order to accurately assess how SNP markers compared with isozyme markers for inbred identification, 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%. To determine whether the missing data for SNPs affected their ability to distinguish among inbreds, 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 profile defined as 1/the number of matching profiles. A score of 1 indicates complete resolution ie the only matching profile is to itself, and decreasing values indicate decreasing resolution power. Scores were accumulated across all samples for each inbred (Table 1). The overall resolution score for isozymes was 0.06 and for SNPs was 0.96; a 16-fold difference in power.

Table 1. Resolution scores for 10 inbreds with isozymes and SNPs

Inbred	Number of samples	Overall resolution 15 isozymes	Overall resolution 16 SNPs
A	145	0.05	0.94
B	20	0.05	0.91
C	21	0.08	1
D	16	0.07	0.94
E	23	0.05	1
F	20	0.07	1
G	16	0.03	1
H	15	0.03	1
I	48	0.17	0.98
J	48	0.17	0.98
Overall	387	0.06	0.96

(c) Analysis of PVPd inbreds with 16 SNP markers

5. SNP data were collected on 309 US and 192 European inbreds that had received Plant Variety Protection (PVP). Pairwise comparisons of the inbreds determined that the 16 markers could uniquely identify 47,292/47,542 (99.9%) of the US inbred pairs and 18,319/18,336 (99.9%) of the European inbred pairs. As there were some missing data in these datasets, it is possible that a complete genetic profile (i.e. data for all 16 SNP markers) would have given an even higher level of resolution.

Conclusion

6. We conclude that a small number of carefully selected SNP markers can together create a highly informative and powerful genetic fingerprint. Such a marker profile would be inexpensive to generate and could have great utility in the identification of varieties and for the management of germplasm collections.

[End of document]