View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DigitalCommons@University of Nebraska

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Agronomy & Horticulture -- Faculty Publications

Agronomy and Horticulture Department

2007

BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification

M. S. Yoon
Soybean Genomics and Improvement Lab, Beltsville Agricultural Research Center, USDA-ARS, Beltsville, MD 20705

Q.J. Song
University of Maryland - College Park, qijian.song@ars.usda.gov

I. Y. Choi
Soybean Genomics and Improvement Lab, Beltsville Agricultural Research Center, USDA-ARS, Beltsville, MD 20705

James E. Specht
University of Nebraska-Lincoln, jspecht1@unl.edu

D. L. Hyten
Soybean Genomics and Improvement Laboratory, USDA Agricultural Research Service, Beltsville, Maryland, david.hyten@unl.edu

Follow this and additional works at: https://digitalcommons.unl.edu/agronomyfacpub

See next page for additional authors

Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences Commons, Botany Commons, Horticulture Commons, Other Plant Sciences Commons, and the Plant Biology Commons

Yoon, M. S.; Song, Q.J.; Choi, I. Y.; Specht, James E.; Hyten, D. L.; and Cregan, P. B., "BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification" (2007). *Agronomy & Horticulture -- Faculty Publications*. 787.
https://digitalcommons.unl.edu/agronomyfacpub/787

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Agronomy & Horticulture -- Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Authors

M. S. Yoon, Q.J. Song, I. Y. Choi, James E. Specht, D. L. Hyten, and P. B. Cregan

This article is available at DigitalCommons@University of Nebraska - Lincoln: https://digitalcommons.unl.edu/agronomyfacpub/787

Theor Appl Genet (2007) 114:885–899
DOI 10.1007/s00122-006-0487-8

This article is a U.S. government work, and is not subject to copyright in the United States.

ORIGINAL PAPER

BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification

M. S. Yoon · Q. J. Song · I. Y. Choi · J. E. Specht · D. L. Hyten · P. B. Cregan

Received: 30 December 2005 / Accepted: 16 December 2006 / Published online: 12 January 2007
© Springer-Verlag 2007

Abstract This report describes a set of 23 informative SNPs (BARCSoySNP23) distributed on 19 of the 20 soybean linkage groups that can be used for soybean cultivar identification. Selection of the SNPs to include in this set was made based upon the information provided by each SNP for distinguishing a diverse set of soybean genotypes as well as the linkage map position of each SNP. The genotypes included the ancestors of North American cultivars, modern North American cultivars and a group of Korean cultivars. The procedure used to identify this subset of highly informative SNP markers resulted in a significant increase in the power of identification versus any other randomly selected set of equal number. This conclusion was supported by a simulation which indicated that the 23-SNP panel can uniquely distinguish 2,200 soybean cultivars, whereas sets of randomly selected 23-SNP panels allowed the unique identification of only about 50 cultivars. The 23-SNP panel can efficiently distinguish each of the genotypes within four maturity group sets of additional cultivars/lines that have identical classical pigmentation and morphological traits. Comparatively, the 13 trinucleotide SSR set published earlier (BARCSoySSR13) has more power on a per locus basis because of the multi-allelic nature of SSRs. However, the assay of bi-allelic SNP loci can be multi-plexed using non-gel based techniques allowing for rapid determination of the SNP alleles present in soybean genotypes, thereby compensating for their relatively low information content. Both BARCSoySNP23 and BARCSoySSR13 were highly congruent relative to identifying genotypes and for estimating population genetic differences.

Communicated by M. Bohn.

M. S. Yoon
Genetic Resources Division,
National Institute of Agricultural Biotechnology,
Rural Development Administration,
Suwon 441-707, South Korea

M. S. Yoon · Q. J. Song · I. Y. Choi · D. L. Hyten · P. B. Cregan (✉)
Soybean Genomics and Improvement Lab,
Beltsville Agricultural Research Center,
USDA-ARS, Beltsville, MD 20705, USA
e-mail: creganp@ba.ars.usda.gov

Q. J. Song
Department of Natural Resource Sciences
and Landscape Architecture, University of Maryland,
College Park, MD 20742, USA

J. E. Specht
Department of Agronomy and Horticulture,
University of Nebraska, Lincoln,
NE 68583-0915, USA

Introduction

Morphological, physiological, pigmentation and biochemical traits (isozymes) have been used to identify and differentiate among soybean cultivars for decades. However, as the number of cultivars has increased, so have the circumstances, whereby new cultivars are no longer distinguishable from existing ones based solely on these traits.

The advent of DNA-based marker techniques, such as restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP),

simple sequence repeat polymorphism (SSR) and single nucleotide polymorphism (SNP) has provided an alternative approach for characterizing and distinguishing cultivars. This has diminished the need to conduct the field trials needed for the time-consuming task of gathering trait-comparison data.

Of the DNA-based marker techniques, SSR or microsatellite markers have been especially useful. Highly polymorphic microsatellite markers have been used quite extensively for forensic and paternity analysis in humans (Balamurugan et al. 2001; Bashiardes et al. 2001; Gangitano et al. 2002; Melendez et al. 2004; Syn et al. 2005), parental and progeny testing in animals (Glowatzki-Mullis et al. 1995; Heyen et al. 1997; Luikart et al. 1999; Usha et al. 1995; Williams et al. 1997) and genotype identification in plants (Schueler et al. 2003; Song et al. 1999). However, SNPs are more abundant than SSRs in human, animal and plant genomes. The total number of SNPs in cultivated soybean is estimated to be in the range of 4–5 million based on the rate of 280 SNPs observed in 76.3 kbp of sequence in 25 diverse genotypes as reported by Zhu et al. (2003). The mutation rate of SNPs is low, 10^{-8} (Kondrashov 2003; Nachman and Crowell 2000) versus 10^{-3} in SSRs (Brinkmann et al. 1998). SNP analysis is generally more robust (Krawczak 1999), even for the analysis of highly degraded DNA (Petkovski et al. 2005). Furthermore, SNPs are more suitable for the development of high-throughput, easy-to-automate genotyping methods (Alifrangis et al. 2005; Faruqi et al. 2001; Hou et al. 2004; Olivier et al. 2002; Ranade et al. 2001) because most SNPs have only two alleles, thereby simplifying the genotyping approaches and analysis.

The objectives of the work reported here were to select a core set of soybean SNPs for efficiently identifying soybean genotypes, to estimate the potential utility of these markers in soybean identification and to compare the congruency of SNPs versus SSRs not only for cultivar identification but also for the estimation of genetic distance.

Materials and methods

Soybean plant material and DNA isolation

Core cultivar set

A core set of 96 soybean cultivars was used in the analysis of SNP and SSR variability. These 96 cultivars included (a) 21 modern Korean Elite Cultivars developed and released by Korean soybean breeders between 1980 and 1990 and chosen to represent the range of cultivars grown in Korea (Table 1); (b) 59 modern N. American cultivars developed and released by public institutions in the US and Canada between 1978 and 1988 and selected to represent the range of publicly developed cultivars grown in the US and Canada (Table 1) and (c) 16 N. American Ancestral cultivars that, based upon pedigree analysis (Gizlice et al. 1994), were deemed to represent more than 85% of the allelic variation present in North American cultivated soybean germplasm (Table 1). Seeds of each of the 96 cultivars were obtained from the USDA Soybean Germplasm Collection, courtesy of Dr. Randall Nelson (USDA-ARS, University of Illinois, Urbana, IL).

Four additional cultivar sets

For the purpose of creating arbitrary sets of cultivars with identical morphological, pigmentation and growth habit characteristics, 36 cultivars were selected from the soybean cultivar database maintained by the Plant Variety Protection Office, USDA (Table 1). The 36 cultivars fell into four Maturity Groups (MG): 10 in MG I, 7 in MG II, 10 in MG IV and 9 in MG VI. Within each group, cultivars were seemingly identical based upon maturity, seed coat color, hilum color, cotyledon color, leaflet shape, flower color, pod color, pubescence color and plant habit. A detailed description of each cultivar and sources of seeds were reported by Diwan and Cregan (1997).

DNA isolation

DNA was extracted from bulked leaf tissue of 30–50 plants of each cultivar using the method described by Keim et al. (1988).

SNP marker discovery

The DNA sequences of unigenes and ESTs were obtained from GenBank, and primers were designed with the goal of amplifying fragments of 300–600 bp in length. Genomic DNA of a set of soybean cultivars, i.e., Minsoy, Noir 1, Archer, Peking, Evans and PI209332 as described by Zhu et al (2003), was amplified using the PCR primers. After the initial determination via agarose gel electrophoresis that PCR primers appeared to produce a single amplicon from genomic DNA, the PCR products were treated with shrimp alkaline phosphatase (SAP) and exonuclease I (*Exo*I) to degrade excess PCR primers and dNTPs and were directly sequenced using one of the

Table 1 List of 132 genotypes used for SNP analysis including 21 modern Korean Elite cultivars, 59 N. American cultivars, 16 N. American Ancestral cultivars and 36 additional cultivars from four different maturity groups

Korean Elite cultivars			
Baegunkong	Duyoukong	Keunolkong	Samnamkong
Baekchun	Hwanggeumkong	Kwangankong	Sinpaldalkong
Bangsakong	Hwaseonputkong	Kwangkyo	Tankyongkong
Bokwangkong	Jangsukong	Namcheonkong	
Bukwangkong	Jangyeobkong	Namhaekong	
Deogyukong	Keomjeongkong	Paldalkong	
North American Elite cultivars			
Agassiz	Gasoy 17	Lawrence	Perrin
Bay	Glacier	Lloyd	Pershing
Benning	Glenwood	Logan	Preston
Braxton	Gordon	Macon	Ripley
Brim	Graham	Manokin	Savoy
Burlison	Hack	McCall	Sibley
Century	Harlon	Maple Donovan	Sprite
Cisne	Haskell	Maple Glen	Sturdy
Conrad	Hoyt	Maple Presto	Thomas
Cook	Hutchson	Maple Ridge	Toano
Dassel	Iroquois	Narow	Weber
Dawson	Johnston	OAC Arie	Williams
Dillon	Kershaw	Ozzie	Young
Evans	KS4694	Parker	Zane
Gail	Lambert	Pennyrile	
N. American Ancestral cultivars			
AK (Harrow)	Dunfield	Mandarin (Ottawa)	Perry
Anderson	Illini	Manitoba Brown	Richland
Capital	Jackson	Mukden	Roanoke
CNS	Lincoln	Ogden	S-100
Four additional sets of cultivars			
Maturity Group I	Maturity Group II	Maturity Group IV	Maturity Group I
Hardin91	A2187	A4715	60400
CM182	Amcor89	Bronson	A6961
BT1790	CM205	CX411	Hartz 507
3172	CM274	DSR-440	Hartz5050
DSR-138	HS2812	FFR 464	Hartz 608
DSR-189	J220	Hartz 4464	HSC623
B1420	Pavone	Nile	Hartz922
Pioneer9141		Pioneer9443	Pioneer9584
S16-60		Pioneer9444	Pioneer9692
B117		Pioneer9472	

PCR primers in a labeling reaction with BigDye Terminators v. 3.1 (Applied Biosystems, Foster City, CA). Sequences were analyzed on an ABI3730xl DNA Analyzer (Applied Biosystems, Foster City, CA). The sequence data from each amplicon were analyzed with PolyBayes SNP discovery software as described by Zhu et al. (2003). SNPs were mapped in the Minsoy × Noir 1 recombinant inbred line soybean mapping population (Song et al. 2004). The single base extension (SBE) method was used for the detection of SNPs (see below). A total of 58 SNPs on the linkage map was selected (Table 2). The SNPs were spaced at intervals of approximately 40 cM across the 20 consensus soybean linkage groups (Song et al. 2004).

SNP allele assay protocol

The SNP allele(s) present in a genotype were determined using a single base extension (SBE) assay of the DNA of the set of 96 cultivars and the set of 36 additional cultivars (Table 1). The SBE protocol used here was essentially that described by Chen et al. (2000) which involves (a) an initial PCR amplification of the SNP-containing genomic fragment, before (b) using a

Table 2 SNP name, linkage group, SNP alleles, PCR primers and SBE capture probe sequences of 58 SNPs

SNP name	Linkage group	Allele	PCR annealing temperature	Forward PCR primer	Reverse PCR primer	SBE Capture Probe (5′–3′) (ZIPCODE in bold)
BARC-014287-01306	A1	A/G	58	CGGAGCCCCAAAAATAATAGTGA	GCCAGCCCCACACGAGAAACTT	**CAA GCT TGG TTC GCG GAC CGT** TTACAATTCCTTGACCGACAC
BARC-017329-02265	A1	C/G	58	GCGGTCTGCTTATCTTGCTGAGTCA	GCGTGGAGGTAATGAGTGTTTGTGAA	**GGG AAA CTC CGC ACC GCC ACG** CTCAAAAAAAGATGGTAAGATGTAAAT
BARC-014639-01604	A1	A/T	58	GCGGAACGGAAAAATAAGATAT	GGCTTGGAGGTCCTTGTGACAC	**CTT ACC GCA CCT CGC AGT CGT** CCCATCTACTGATGCCTCA
BARC-018023-02499	A2	A/G	58	GGGCACCGAAAGAAATCCAAACTAAG	GGCAGCAGGCAATTCCATTCAG	**TCC ATG ACA AAC AGG GAG TCG** ATTAATTTGGAAACAGTTGGCAT
BARC-014665-01613	A2	A/G	60	CGCCGAGGGGAAGAAAAATTATG	CCGTCCAAGAACAGATCATCAG	**GCG GGC TTG GTA CGT TTG GGG** TCTTTGTTGGCTAGGGGG
BARC-018147-02532	A2	A/G	58	GCCGCAGCAGCAGGAGGTAAACT	CGCTTGGTGCCAGGACAAAAATGTG	**CAC AAC ATC CGG CCG TGA GAG** CTACAAACAGGCAGATAAATGAA
BARC-018869-03031	B1	A/G	58	AGAACCTACCCTCGCTAACAACC	GGTTCGTGAAATTGCTCAGGACTT	**GCA TGA ATA CAC CAG TAC TCG** AGTCTGCAACAAAAGCACA
BARC-018557-03202	B1	A/G	58	GGCACAATTAAGCGTTCAACACAA	GAGGGTTCCATGCTCACTTTCG	**TAC TGA CGA CAG CGG ACT TAC** ACAAGTTTAGTCATCATGAATAATACA
BARC-021459-04106	B1	A/T	58	ACCTTAATCGTCATCCTTCATCTCC	ACACCATGTACTTGCCGTCTTC	**GCC GGC GGA TAC GCG CTG GGG** TTTACAACTTTTACCTCAATGAGC
BARC-025703-04999	B1	C/G	58	GGCAACTATCAGAGGTTTCTCTT	CAAACACTATTTGACTTTGAGGG	**TGT TGA GCG TGA AGA CCG CCG** TCAGTTCTCGTCTAAATCTACTG
BARC-024295-04827	B2	A/G	58	GAAGGGAAATTTTAATGTTGGAC	CAAGAGACACGTGCATAACACTA	**AGG GAC AAC CTG TCG CGA GAT** CCCTCCGATGATATGAAGAAAA
BARC-013927-01275	B2	A/T	58	CGGCAGCGGCTCAAGAAATAAGAT	GCGTTGCTTGGAAATGACATGAAGA	**ACC CTT CCG CTG GAG ATT TAC** GGCTTGACAGATCCTGTTG
BARC-012953-00413	B2	C/T	54	GCCTTTCCAGAACAGCCCCATTT	GCCCTCCACGATGCGAAGC	**CGG TGA GCA GTT CAA AGC CGG** GCCAGTCCCAGATTGACAC
BARC-017211-02251	B2	C/T	58	GGGGCAATCAATTTCAGTCATTTAG	GGGTGCGATTTGATGTAAGTAAG	**CCT ACG AAT GCT CAG TGC ACG** ATTTTCTCACATGATCTGGAAGT
BARC-025861-05129	C1	A/T	58	CTTCTGGTATCCTTTGTCCTCTT	CATACAGGTTTATTTTCCACCAA	**GCG GCA CGT CGT CGA CCG** CAATAGAATTTTCGCTGCTCC

Table 2 continued

SNP name	Linkage group	Allele	PCR annealing temperature	Forward PCR primer	Reverse PCR primer	SBE Capture Probe (5′–3′) (ZIPCODE in bold)
BARC-019015-03051	C1	C/T	58	AGTTCCTTACACAATATAGTCTTCGA	CGAGCCTCCAGAAGACGAAT	**CAA CCT TGC GCA TAC GAC GTG** AGTCACAAATCAAATGCTGAAAA
BARC-014557-01577	C2	C/G	58	CGCCACGGAATCCAAAACACAAC	GGGGGTCAGGGGACAATAATCTT	**TGG GGA TCT GTA GAC CCA GCC** GTATGTGTAGTAGGAACCCTTAATATT
BARC-014491-01561	C2	A/G	58	GCGGGTTTTTCATTTTGTGTCTGTT	GCGGTTGGCTGCTCAGTTGTTTTGT	**CGC TGC AGT CTG CGT TCA CTC** ACAAGGTGGATTTGATATTATTGG
BARC-016027-02038	D1A	A/G	58	GCTCCTGAACAAAAATAGAAAACAAA	GGACCCAAACACCCGCCTTCATAATG	**CGA TCG GGG AAC CCA GTA AAG** CATTCACTCCTGTATATGCTAAAAT
BARC-017943-02462	D1A	A/T	58	GCGTCCGATTGACCGATCTATACAAA	GCGGGAATGGAGCGAAGGAACTTATC	**ATC TGT AGC GCC TCT TCG AGC** CTAGACCTTCCTTCTTCATGTC
BARC-021531-04136	D1A	C/T	58	ATCCAAACATCTTCCGTGCAGTATC	TCTAGAACCGCTCGACGCTAATAA	**AGG ACG GCG CTG GCA CGT TGA** TGCTAAAGCCTGTGTAGCTTA
BARC-018835-03260	D1A	A/G	58	GCGTATTTGGAAATTTGTTCAGAATG	AATCACACTCTATCCACTCCACAA	**AGT GTC TCG TCG ATC ATC** CTTTTTCACAGCCTATTCTAAGAG
BARC-017963-02483	D1B	A/T	54	CGGGTGCCATGAGTGACTAAAAT	CCCCTGCTGTAACAGTCATAGTATTC	**TGG GGA TCT GTA GAC CCA GCC** GCAATAAAGGAGCATTATGGC
BARC-017895-02427	D1B	C/G	58	GCGGAAAGAGGGAAAAGAGAATAAAA	GGGCATGGTCAAAATGGATAGATGA	**ACC CTT CCG CTG GAG ATT TAC** GGCACCTTCATTGACTGAGT
BARC-020293-04543	D1B	A/T	58	TCCAAGTGGGCTACATAGTCTAA	AGACTCGTCAAATTTAATGCAAA	**ATC GTC TGG GCG GTC TCA AAG** CGGGTAAGTCATGTATTAGTATTGT
BARC-014389-01344	D2	C/T	58	CGGTGGGGCAATGAACTCATCAAG	GCGAAGCGGGAAAAATATGAAGAG	**GTG GTA GCG GCA GTC GTG GTG** CGGGAGATGGCGCATTAA
BARC-024449-04894	D2	A/G	58	GAGGAAGAGGAAGATGATGAAGT	GCAAAAACGAAATTAAGGAGAAT	**CCA CCG TTC GTC GCT CGG GGT** CCAAAAATATTACAAGGTTAAGGC
BARC-024481-04933	E	C/T	58	AAGAGCCGTGCTATAATAGGAAT	CACTTGATTCATTCCACTTTGTC	**CGA GTC GGG CTG CTG TTG CAA** ACCTTGTTGATCATGAAATTAAGA
BARC-016037-02043	E	A/G	58	GCGTCGGCTCAAGAAAAGACAATGAT	GCGTGGGAAGGAAAATGGAGAAACA	**TCG ACT GGC CTC AAC CGA TTG** ATTATTTGGAGAGAGGAAAGAAGA
BARC-013501-00504	E	A/G	58	GGGGCTTGGCAAATTGAACTACAG	CGGGCTTGAGCAAAATGTGAAGAT	**CCA ACT TGA CAC GTC GCA AGG** AGTGGAAAAGAATTAAAAGACTAGATA

Table 2 continued

SNP name	Linkage group	Allele	PCR annealing temperature	Forward PCR primer	Reverse PCR primer	SBE Capture Probe (5′–3′) (ZIPCODE in bold)
BARC-020171-04491	E	A/C	58	AGTTCGATGATTGAAGTGGTAAA	AAGTTTGGTATCTTTTGCTTTCC	**GAG CCC GTA TCG CCG GAG TCA** AAAGAAAGTATACCAGCACGA
BARC-014411-01355	F	A/G	58	GCGAAACTAATGGCACAGATGATAC	GCGAAAACAGGCAATAAAAATAACA	**CTG GCG TTT GCC CGC TCG CTC** CATGAGTCACTTGAACAAGA
BARC-013633-01184	F	C/T	58	CCCCCACGAGACGAGAAGAGACAC	CCGTTCCAGCAGGCAAGCGATGTG	**CCA ACT TGA CAC GTC GCA AGG** AATGTTCGTCTTTCATGGGG
BARC-018741-03001	F	A/G	58	AACTAAGGATGGTACACACTGGTCTA	CTGGCGGATCTGGATCTGTCTC	**TCG ATC GCA GCC CAT CCC GGG** GGGCCCAGAGAGTTTATTCAT
BARC-020139-04480	G	C/T	54	TGAAGATGAGAAGAAGAAGCAAG	ATCTAGTTTCCAAAAGTGTGGGT	**CGG TCT GGA AGG GCG GCG AGG** ATCAACCCTGAGAACCCAAT
BARC-013509-00507	G	C/T	58	GCGAGGAGGAGGAGGAATCAAACACA	GCGTGGGCAGAATTGTTGAATACAT	**GCT GGC GCA GAC CTT TGT CTC** ACAAACTCCAAGGGCAGG
BARC-021603-04153	G	A/G	58	AACCATATTCAAACGGCAAACCAAA	CGTTGCGGGTTACGATAGTGAA	**GGT GGG ATT GTC ACT GCT GCT** AATCTGGTTCAGTATCAAATGAAAA
BARC-017095-02197	H	A/G	58	CGGCGGAACCAGGTGTTGAATG	CGGAGGGCGTTTCGTGATGATG	**AGA CGC CCG ATA GTT CGA GCC** AGGAAGGCCAGGCACAAT
BARC-021659-04168	H	A/G	58	CTAGGGCTGCAAATGCGACAA	AAGGTGAAGGATTCGTTCTTAGGAA	**ATT CTC GGT GTC CGC GGG CGA** ACTCATCATATATGGTAAGTTTGTT
BARC-017181-02241	H	A/G	58	GCCCACCTGGTGCTTCACTTAA	GCCAAGCTGTTGCTGAAATGATGT	**TGC TCA CGG CAG ACG AAT GAG** ACCCTTCCTGTCAGACATATATA
BARC-025709-05013	H	A/G	58	TATATGTTATTTCAGGGTCACCG	GAATACACATGTAGGGATAGGCA	**TCG ACT GGC CTC AAC CGA TTG** ACAGATGCTCCCCTATTTTCT
BARC-013583-01166	I	A/G	58	GCCAAGCCACGAAGCCAGAGAGTG	GCGAAAGGAGGTTGATGAACAGATGA	**ACG CCA TCG CGG TGC TAA AGC** CTCAGACACAAGGTCATTCAT
BARC-014559-01579	I	A/C	58	GGGCTCCATTGTGAGTCACTACTT	GGCACCGAGGATTTGTTTCTTGAC	**ATC CGT TCG GTG TTG CGT AGT** TGTTTCAATAAGTGACTGGTCTC
BARC-016029-02040	J	A/G	60	GCCTCGCCACAAACAATGGAGTAGTG	GCCTCACAGCCTCTTACAGAAACTT	**CCT ACG AAT GCT CAG TGC ACG** TTGGCCAGTGGAGGAAGA
BARC-013651-01220	J	A/G	58	GGCAAGGGGGTTAAAAAGAAATGATA	GGCGATGGGCAACACCTCAACTC	**ACG CAC CCC AAC CTG TCC GGA** AAGTGATTTACTGCCTATTCACT

Table 2 continued

SNP name	Linkage group	Allele	PCR annealing temperature	Forward PCR primer	Reverse PCR primer	SBE Capture Probe (5′–3′) (ZIPCODE in bold)
BARC-020031-04407	J	A/G	58	TGCAAGATATGGGTAATAATTCG	TAGCAGGTTCCACATAGTTCCTA	**CGG TGA GCA GTT CAA AGC CGG** ATAATGAAATGAGGGGAAAGAAATAA
BARC-014795-01662	J	A/C	58	GCGGGGTTGCTCCTGAATGAT	GCGTTCCAAAAGGCTTACTACAA	**GCT GGC GCA GAC CTT TGT CTC** AAAACCTTCAAGTACAAAACAAAAG
BARC-024229-04809	J	C/T	58	CGTATTTATAAAGTGAAACCCCC	AACCCAAAATTTATTTGAAGGAA	**ATT CTC GGT GTC CGC GGG CGA** GTTAGCAATATTACAGATTGTGGA
BARC-014659-01609	K	C/T	58	GCCGTGAACCTAAAGTATGAGACAT	GCGTCCCAACCCTCTTCTTCTTCTT	**TGT GCC AGC CGT CGG TGC CAT** CTCTGTCGTTTGTTTCCTTAGTA
BARC-016563-02119	K	C/G	58	GCGACCCGAACAAGAACAAGTT	GCGAGGATCTTCACATTCAATA	**AGG ACG GCG CTG GCA CGT TGA** AGGCAAACTCACCTCACTG
BARC-024345-04854	L	C/G	58	AACATTCATTCTCGAACCCTCT	CATATATATTGCCAAAGTTCCCC	**CGA GTC GGG CTG CTG TTG CAA** TTGGTTCAATCATGGTGGA
BARC-014655-01607	L	A/C	58	GGCAAGAATAAAGGGAATGTGTCAAT	GGCACCGCACAAGGAAGAACAGAC	**TCA GTG ACG CGC TCA GCG TTG** ACTTTGTATTACGCAGAGTACAG
BARC-021879-04229	M	A/G	58	CCGCAGGCACTTCATGTTCTAA	GCTGGTTCAGATGGTCAGAAGATAT	**TTC GTG GCC ATG GTG ACC GCT** ACTGCCTCTTACAATTGGAAATA
BARC-022289-04309	M	C/G	58	AGAACAGCAGAACACATATAGCATT	GAGACGGCAGTGATCGGGAA	**ATT CTC GGT GTC CGC GGG CGA** GGCACATCGATTAAAACCAC
BARC-013905-01267	M	A/T	60	CGCCATCCTAAAAATCTTCTGTGTG	CGGGACCAAGCAATAGATAGAG	**AAG CTT TAC GCC AGC GCC GAA** ACTGCTGACGAAACAGAAAC
BARC-024329-04849	N	A/G	58	TTCAAGGTCATCTTCTACTTGGA	AAACCACATCCTAATGATACCCT	**ACG CAC CCC AAC CTG TCC GGA** GTATTAATAAGATAATAAATTT AAGAACTC
BARC-018101-02517	O	A/G	54	GGGGCAGACGAAAGGCTACTCTAAGA	GGGTCGCACAATGTATGAAGAACTGA	**GCT GGC GCA GAC CTT TGT CTC** CGTCTATTGCCAGAAACAGT
BARC-018693-02992	O	C/T	58	ATATCCCTGAGAGACACATTGATTT	GGATGACAGCGTTCGCAATAT	**ACG GGA GCT CAA GTC CGT GCC** GATCACTTGTGCCAAGTTATG

SNP-specific SBE oligonucleotide capture probe designed to anneal next to the SNP, (c) using a biotin-labeled dideoxy terminator in the extension step, and finally (d) using microsphere-based flow cytometry for SNP allele identification. Details of each step follow:

Step (a) PCR amplification of the SNP-containing fragment was performed on a MBS 384 Thermo Hybaid thermocycler (Thermo Electron Corporation, Somerset, NJ). The reaction mixture contained 0.1 µl AccuPrime DNA Taq polymerase (Invitrogen, Carlsbad, Calif.), 0.5 µl 10× AccuPrime Taq DNA polymerase buffer, 1.25 µM of each primer and 30 ng template DNA in a total volume of 5 µl. Cycling conditions included an initial denaturation at 94°C for 2 min, then 30 cycles of denaturation at 94°C for 30 s, annealing temperatures ranging from 54 to 60°C (Table 3) for 30 s and extension at 68°C for 1 min, followed by an additional extension at 72°C for 10 min.

Step (b) A total of 0.5 µl of each of the ten PCR products from ten different SNP-containing loci were treated with 1 U of SAP and *Exo*I to degrade excess PCR primers and dNTPs. The ten-plex reaction solution was mixed thoroughly and incubated at 37°C for 1 h, followed by 15 min at 75°C to inactivate the enzymes. A total of 1.25 µl aliquot of the SAP/*Exo*I-treated PCR products was added to 2.5 µl SBE reaction mixture containing 0.332 µl 10× buffer, 0.000094 U Thermo Sequenase (USB, Cleveland, Ohio), 3 mM $MgCl_2$, 0.12 µM of each SBE capture probe primer, 0.4 µM allele-specific biotin-labeled ddNTP and 0.4 µM of each of the other three unlabeled ddNTPs. The thermocycling conditions involved an initial denaturation at 90°C for 1 min, then 79 cycles of denaturation at 90°C for 30 s, annealing at 50°C for 20 s and extension at 68°C for 15 s. The resulting SBE products were then held at 4°C. The SBE products were then precipitated in ethanol at a final concentration of 60% and were incubated in the dark at room temperature for 30 min, pelleted and dried.

Step (c) Each SBE capture probe contains an additional 18–21 nucleotide sequence (Zipcode) at its 5′-end which allows hybridization to a complementary sequence (called a cZipCode) attached to a fluorescently color-coded carboxylated polystyrene LabMAP™ microsphere (Luminex Corporation, distributed through MiraiBio Inc., Alameda, CA, USA). The hybridization of the capture probe to the microsphere was conducted as described by Chen et al. (2000). Hybridization procedures for binding the ten extended SBE capture probes (with their ZipCode) to their ten specific microspheres (with the corresponding anti-ZipCode) were carried out in a 50 µl total reaction volume, including 49.4 µl 1× TMAC [3 M tetramethylammonium chloride, 50 mM Tris–HCl (pH 8.0), 4 mM EDTA (pH 8.0), 0.1% Sarkosyl] and 0.06 µl microsphere with 3,000 microspheres of each type in the reaction mix. After hybridization at 54°C for 30 min, the biotinalated products were conjugated with streptavidin by adding 10 µl of 1× TMAC and 200 µg streptavidin R-phycoerythrin at 54°C for 5 min (60 µl of total reaction volume).

Step (d) A Luminex 100 flow cytometer equipped with a Luminex XY Platform plate reader was used firstly to identify the specific microsphere (and thus the SNP locus) based on microsphere color followed by the detection of the presence or absence of the streptavidin–biotin conjugate (indicative of the presence or absence of the specific SNP allele). The fluorescence on the surface of the microspheres resulting from the streptavidin label was converted to a mean fluorescence intensity (MFI) value based on a minimum of 100 microspheres of each of the ten microsphere types in the ten-plex.

SNP locus mapping

The genetic map position of the SNP loci was determined by applying the above SNP allele assay procedures to the parents and members of a Minsoy × Noir 1 recombinant inbred line (RIL) mapping population. This is one of the five mapping populations recently used by Song et al. (2004) to derive a consensus soybean genetic linkage map comprised of SSR, RFLP and other markers. A preliminary positioning of the SNPs relative to the known soybean map was deemed useful to ensure that the final selection of SNPs would not be closely linked and would generally span as much of the soybean genome as possible.

Table 3 Linkage groups, fluorescent dye labels and SSR diversity measured in 96 cultivars of a selected set of 13 simple sequence repeat loci, BARCSoySSR13

Locus	Linkage group	Fluorescent dye label	Number of alleles	SSR diversity
Satt009	N	NED	17	0.82
Satt038	G	FAM	5	0.60
Satt114	F	HEX	5	0.75
Satt147	D1a	NED	6	0.73
Satt177	A2	NED	6	0.66
Satt191	G	FAM	6	0.63
Satt242	K	NED	7	0.76
Satt243	O	NED	5	0.66
Satt294	C1	FAM	8	0.70
Satt308	M	FAM	6	0.72
Satt373	L	NED	13	0.83
Satt414	J	FAM	9	0.83
Satt534	B2	NED	9	0.78

Cultivar assay with SSRs

Thirteen highly informative tri-nucleotide soybean simple sequence repeat markers, known as the BARCSoySSR13 set (Table 3), had been previously identified by Song et al. (1999) as a marker-based means for distinguishing cultivars. For comparative purposes, this 13-SSR set was used to assay the same set of cultivars (Table 1) that were assayed with the SNPs. Protocols for PCR and fluorescent dye labeling described by Song et al. (1999) were followed. SSR alleles were separated on an ABI 3730xl DNA Analyzer and analyzed using GeneMapper V3.0 (Foster City, CA).

Selection of a core set of SNPs

First selection criterion

SNPs which were developed in the Soybean Genomics and Improvement Laboratory, Beltsville Agricultural Research Center and genetically mapped in the Department of Agronomy and Horticulture, University of Nebraska, Lincoln were used in the present study. An initial set of 58 SNP loci from each of the 20 soybean linkage groups was selected for testing (Table 2). When selecting these SNP loci, an attempt was made to sample all linkage groups and within any linkage group, to avoid SNP loci that were closely linked.

Second selection criterion

Based on the SNP assays of 96 Korean, N. American and Ancestral cultivars only those SNPs with SNP diversity greater than 0.2 in the populations were retained for the next stage of screening.

Third selection criterion

SNPs that because of their physical linkage provided similar information for purposes of distinguishing cultivars were generally eliminated via the application of the first selection criterion. However, it was also necessary to detect genetically unlinked SNPs that did not segregate independently as a result of genome-wide linkage disequilibrium. In order to identify a subset of the remaining SNP loci that would be small in number yet still be maximally efficient for distinguishing cultivars, pairwise similarity among the remaining SNPs were calculated (see Statistical analysis) and a cluster analysis was performed based upon the resulting similarity matrix. Because SNPs in the same cluster are (empirically at least) expected to provide similar information, one SNP was selected from each cluster. To maximize the number of linkage groups in the final selected set (without appreciably reducing the power of the SNP set to distinguish genotypes), the selection of SNP from clusters with multiple SNPs was made so as to maximize the number of linkage groups represented in the selected set.

Statistical analysis

SNP and SSR diversity

The relative informativeness of a marker was calculated as: $1 - \Sigma p_{ij}^2$ where p_{ij} is the frequency of the jth genotype summed across all genotypes at the ith locus (Weir 1990). This value is referred to as gene diversity.

SNP association and cluster analysis

To measure the association between any two SNP loci, the Hill and Robertson (1968) measure as later used by Awadalla et al. (1999) was used: $r^2 = (p_{AB}p_{ab} - p_{aB}p_{Ab})^2/(p_A(1 - p_A)p_B(1 - p_B))$, where A and a represent alleles at one locus, B and b represent alleles at the other and p_{AB}, p_{ab}, p_{aB} and p_{Ab} are the genotype frequencies in the population of 96 cultivars. Because r^2 ranges from 0 to 1 (i.e., from complete disassociation to complete association), the distance (d_{ij}) between each pair of loci was calculated as $1 - r_{ij}^2$. A pairwise distance matrix among SNPs was used to place SNPs in clusters according to their power to discern cultivar differences. The unweighted pairwise group with arithmetic averaging (UPGMA) was used for clustering. This was followed by the application of the third selection criterion as described above.

Estimation of the power to distinguish cultivars in soybean populations

To evaluate the power of BARCSoySNP23 versus randomly selected sets of 23 SNP loci to identify individuals in a population, computer simulations were performed. The observed allele frequency of each selected SNP in the population of 96 genotypes was used for this simulation. Simulated datasets with varying numbers of simulated cultivars were generated, i.e., cultivar population sizes of 50, 100, 500, 1,000, 1,500, 2,000, 2,200 and 2,500. At the beginning of each computational cycle, a random set of 23 SNPs was drawn from the set of 58 SNPs, the actual allele frequencies of each selected SNP in the 96 cultivars were used as the

probability to generate a set of 50, 100, 500, etc. individuals. A total of 1,000 sets of simulated cultivars with each population size (50, 100, 500, etc.) were generated 1,000 times for each random set of 23 loci and analyzed for identical allelic matches at all 23 loci. The process was repeated for each of the 1,000 random sets of 23 SNPs and the average number of uniquely identified simulated individuals was calculated. The maximum number of individuals that could be uniquely identified was defined as the average number of individuals that were uniquely identified in 1,000 × 1,000 computations. A smaller number of indistinguishable genotypes is an indicator of greater power to distinguish genotypes.

Congruence of the BARCSoySSR13 and BARCSoySNP23 distance matrices

The pair-wise distance matrices among cultivars derived from the SSR allele size data and the SNP allele data were measured by the proportion of loci that differed in each pairwise comparison at the locus. Thus, a pair of genotypes could be scored 0 at a locus if they possess two identical alleles, 0.5 if they possess one identical and one dissimilar allele or 1 if they possess no identical alleles. The pair-wise distance was calculated between each pair of cultivars by summing scores across all loci. The Mantel test (Mantel 1967) was used to determine the significance of the correlation between the 96 cultivar distance matrices derived from the SSR and SNP allelic profiles.

Calculation of average genetic distance within each population

To calculate the average genetic distance within the Korean cultivars, the N. American cultivars, the N. American Ancestral cultivars and the 36 maturity group I, II, IV and VI cultivar populations, a statistic similar to π (Nei and Li 1979) was calculated: $\pi = (2\Sigma\Pi_{ij}/(n(n-1)))/L$, where $\Pi_{ij}(i < j)$ is the number of allele differences between the ith and jth genotypes, the term $n(n-1)/2$ is the number of possible genotype pairs within the group, and the denominator L is the number of loci assayed.

Informative power of SNPs versus BARCSoySSR13

A cumulative SNP or SSR diversity was calculated based on the formula $1 - \Pi(1 - h_L)$ (Chakraborty et al. 1988), where L is the total number of loci and h_L is the SNP diversity of the Lth locus. Due to the variation of h_L for each locus, the average SNP diversity of the SNPs of 23 BARCSoySNP23 was used as h_L for each locus in order to estimate the number of SNPs required to match the power of BARCSoySSR13 panel.

Analysis software

SAS (1999) software was used for all of the above computations.

Results

Selection of a core set of SNPs

Based on the observation of the 96 cultivars, informativeness (SNP diversity) of the 58 SNPs varied from 0.02 to 0.57 with an average diversity of 0.34. Soybean cultivars are ordinarily homozygous, but when developed often trace to more than one plant or to a single heterozygous plant, and thus can be heterogeneous for non-observable markers despite careful selection for phenotypic homogeneity. In the present study, the number of cultivars heterogeneous for two alleles at a SNP locus ranged from 1 to 8 at 44 of the 58 SNP loci. Because of the SNP allele heterogeneity in a few cultivars, some SNPs had diversity values in excess of 0.5. Of the 58 SNPs, 40 had a diversity greater than 0.20 (Fig. 1). These 40 SNPs were further analyzed to create a dissimilarity matrix for cluster analysis (Fig. 2). The cluster analysis indicated that BARC-025861-05129 and BARC-013927-01275 provided the most unique information of the 40 loci analyzed. BARC-018557-03202 and BARC-025703-04999 provided the most similar information in identifying genotypes. The cluster analysis can be used to choose a subset of loci that provide

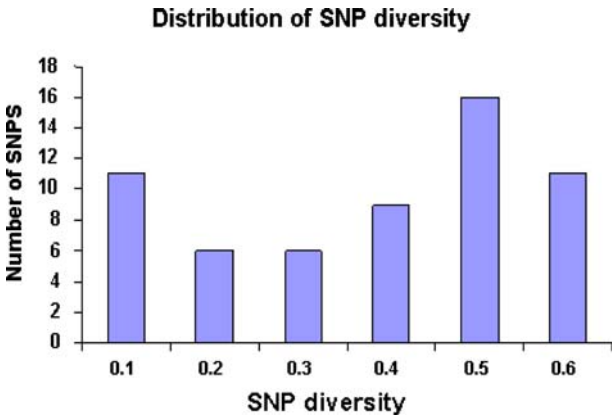

Fig. 1 SNP diversity of 58 SNPs based on the analysis of 96 Korean, N. American and N. American Ancestral cultivars

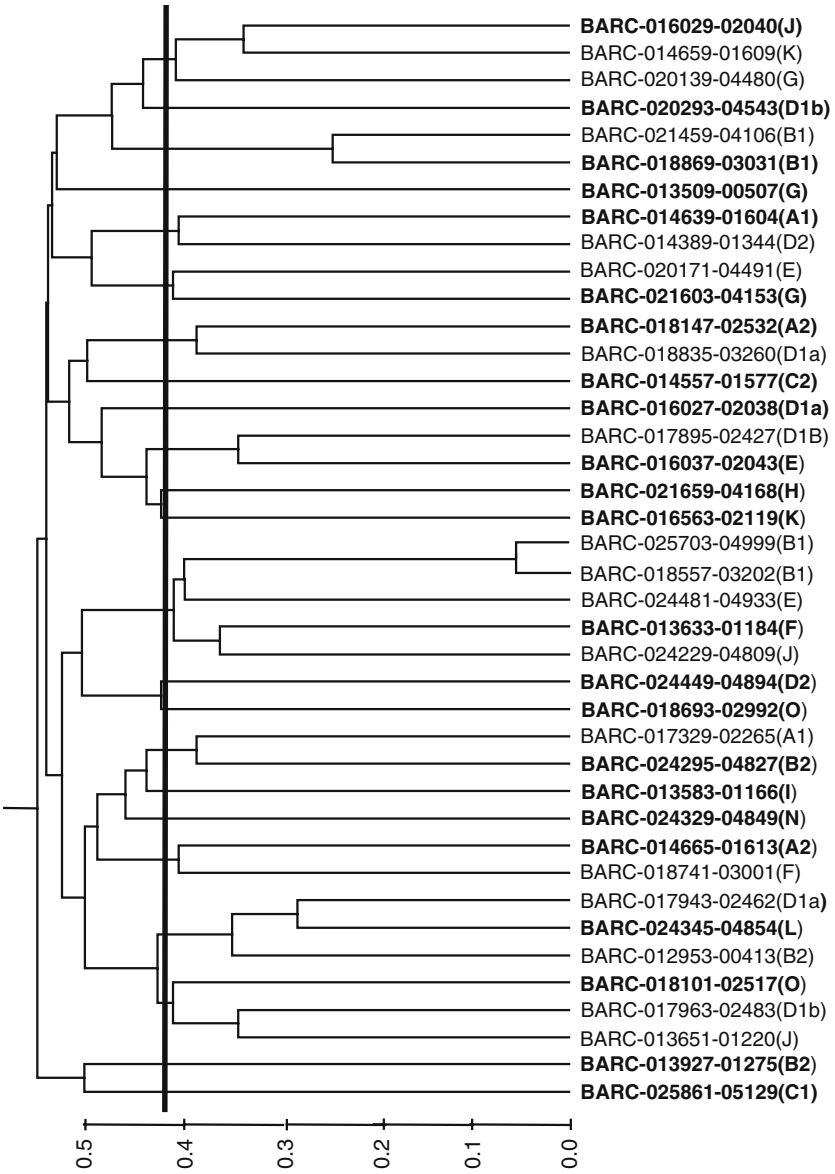

Fig. 2 Cluster analysis of 40 SNP loci from 19 of the 20 soybean consensus linkage groups based upon the alleles present in 96 Korean, N. American and N. American Ancestral cultivars. A total of 23 SNP loci (in *bold font*) that most efficiently distinguished the 96 cultivars were identified based upon the position of the bold intersecting line. The linkage group in which each SNP marker resides is indicated in parenthesis following the SNP name

maximum information for distinguishing the 96 genotypes. The number of SNPs to include such a subset is arbitrary and depends on the efficiency of markers to distinguish genotypes and the size and diversity of the test population. In the present case, 23 SNPs were selected based on the vertical criterion line in Fig. 2. One SNP was selected from each of the 23 sub-clusters, and for those sub-clusters with two or more SNPs within a sub-cluster, the one SNP selected was the one that maximized the number of linkage groups represented by the 23-SNP set (Fig. 2). The 23 selected loci were distributed on 19 of the 20 soybean linkage groups. Two SNPs were selected from each of the four linkage groups (A2, B2, G and O) while none was selected from linkage group M. The distances between the two SNPs on A2, B2, G and O were 56.1, 51.4, 23.0 and 87.9 cM, respectively. The SNP diversity of the 23 SNPs ranged from 0.22 to 0.56, and 19 of the 23 SNPs had SNP diversity values greater than 0.40. The frequency of genotypes homozygous for the minor allele at the 23 SNP loci (including those heterogeneous for the minor allele) ranged from 12.5 to 47.9% (Table 4).

BARCSoySNP23 to distinguish cultivars with identical traits

To examine the effectiveness of the BARCSoySNP23 panel to reliably distinguish morphologically identical individuals, we extended our analysis to a group of 36 genotypes from each of the four maturity groups selected for similarity of pigmentation and morphological characteristics within each maturity group

Table 4 Loci, linkage group and SNP diversity of a set of 23 selected SNPs based on the analysis of 96 Korean, N. American and N. American Ancestral cultivars

Locus	Linkage group	SNP diversity	Percentage of minor allele + heterogenous genotypes
BARC-014639-01604	A1	0.48	37.5
BARC-014665-01613	A2	0.31	18.8
BARC-018147-02532	A2	0.51	38.5
BARC-018869-03031	B1	0.52	42.7
BARC-013927-01275	B2	0.48	35.4
BARC-024295-04827	B2	0.52	46.9
BARC-025861-05129	C1	0.31	18.8
BARC-014557-01577	C2	0.51	39.6
BARC-016027-02038	D1a	0.40	26.1
BARC-020293-04543	D1b	0.37	22.9
BARC-024449-04894	D2	0.37	24.0
BARC-016037-02043	E	0.5	43.8
BARC-013633-01184	F	0.48	37.5
BARC-013509-00507	G	0.56	47.9
BARC-021603-04153	G	0.48	34.4
BARC-017095-04168	H	0.50	45.8
BARC-013583-01166	I	0.48	35.4
BARC-016029-02040	J	0.46	32.3
BARC-016563-02119	K	0.53	40.6
BARC-024345-04854	L	0.49	43.8
BARC-024329-04849	N	0.42	29.2
BARC-018101-02517	O	0.51	38.5
BARC-018693-02992	O	0.22	12.5

(Table 1). The assay showed that all SNPs, except BARC-018693-02992 (on LG-O), were polymorphic among the 36 genotypes. The SNP allele data were used to calculate similarity coefficients between the 36 cultivars within each of the MG I, II, IV and VI groups. The average similarity among all cultivars was 0.63; the most similar cultivars had common alleles at 20 of 23 SNP loci. In the case of the MG I group, the cultivars DSR138 and DSR189 as well as BT1790 and Pioneer 9141 were the most similar, having similarity coefficients of 0.87. In MG II, IV and VI, the most similar cultivars had similarity coefficients of 0.83, 0.87 and 0.80, respectively. The group of 23 SNP loci was adequate to distinguish all cultivars in the four MG sets that were not distinguishable using common classical traits. Distinct allelic profiles were generated for each cultivar (Table 5).

Maximum number of genotypes distinguishable by the BARCSoySNP23 panel

It was of interest to estimate the probability that any two hypothetical cultivars would have identical SNP allelic profiles using the selected 23 SNP loci. Although there are 2^{23} possible allelic combinations of 23 bi-allelic SNPs, the combinations that are empirically probable are much lower and dependent on the frequencies of the alleles at each SNP locus. Using the observed allele frequencies in the 96 cultivar set, an analysis of the simulation data indicated that the BARCSoySNP23 panel could uniquely identify each of 2,200 soybean cultivars. In order to compare the efficiency of the selected BARCSoySNP23 panel versus a random set of SNPs, sets of 23 SNPs were randomly drawn from the 58 SNPs and the maximum number of simulated cultivars uniquely distinguishable was counted at each level of simulated cultivar size. As indicated in Table 6, the average maximum number of genotypes that could be uniquely distinguished by any randomly chosen set of 23 loci was only 50. Thus, as would be anticipated, the selected panel of 23 SNPs substantively increased the power of cultivar identification over any set of 23 randomly selected SNP loci. This was not an unanticipated result given the careful selection of the BARCSoySNP23 set, but nonetheless, confirmed the power of this set of loci for cultivar identification.

Table 5 Mean and range of cultivar similarity coefficients within each Maturity Group of the four additional sets of cultivars listed in Table 1

Maturity Group	Mean	Minimum	Maximum
I	0.67	0.48	0.87
II	0.70	0.57	0.83
IV	0.74	0.54	0.87
VI	0.69	0.57	0.80
All 36 cultivars	0.63	0.35	0.87

Comparison of the consistency of BARCSoySSR13 versus BARCSoySNP23

The average pairwise distances among the 96 cultivars based on a BARCSoySSR13 analysis versus one based on a BARCSoySNP23 analysis were 0.7034 and 0.4275, respectively. The Z statistic calculated from the two distance matrices was 0.31, thus, the Mantel's test indicated a highly significant association between the two ($P < 0.00001$). The Mantel test was also used to test the congruency of distances determined by a BARCSoySSR13 versus a BARCSoySNP23 analysis applied to the 36 additional Maturity Group I, II, IV and VI cultivars. The correlation ($Z = 0.37$) was significant ($P < 0.0001$). Thus, the set of 23 SNPs and the set of 13 SSRs distinguish amongst these 36 cultivars in a highly congruent manner.

The magnitude of average genetic distance within the four populations (Korean cultivars, N. American cultivars, N. American Ancestral cultivars and the 36

Table 6. Number of genotypes theoretically distinguishable by the BARCSoySNP23 panel and by random sets of 23 SNPs

No. of genotypes generated	BARCSoySNP23			Sets of 23 randomly selected SNPs		
	No. of genotypes distinguishable	No. of indistinguishable genotypes	Frequency of indistinguishable genotypes (%)	Number of genotypes distinguishable	Number of indistinguishable genotypes	Frequency of indistinguishable genotypes (%)
50	50	0	0	50	0	0
100	100	0	0	99	1	1.01
500	500	0	0	494	6	1.21
1,000	1,000	0	0	981	19	1.90
1,500	1,500	0	0	1,452	48	3.20
2,000	2,000	0	0	1,888	112	5.60
2,200	2,200	0	0	2,075	125	5.70
2,500	2,497	3	0.12	2,355	145	5.81

MG I-VI cultivars) based on BARCSoySNP23 was generally consistent with that of BARCSoySSR13. The average genetic distances of genotypes within Korean cultivars, N. American cultivars, N. American Ancestral cultivars and 36 Maturity Group I–IV cultivars were 0.699 ± 0.152, 0.704 ± 0.144, 0.710 ± 0.186 and 0.663 ± 0.136 respectively, as evaluated by BARCSoySSR13, and were 0.391 ± 0.109, 0.398 ± 0.109, 0.465 ± 0.129 and 0.370 ± 0.108 respectively, as evaluated by BARCSoySNP23 panels. Using both marker sets, the highest values were observed within the N. American Ancestral cultivar population and the lowest in the 36 Maturity Group I–VI cultivars.

Discussion

Criteria for marker selection and the efficiency of the BARCSoySNP23 panel

Use of molecular markers for cultivar identification has been reported in numerous crops. Esselink et al. (2003) developed 35 microsatellite markers from enriched libraries of *Rosa hybrida* L, after discarding eleven loci due to their poor amplification, a total of 24 primer pairs was used to characterize 46 hybrid tea varieties and 30 rootstock varieties belonging to different species. Aranzana et al. (2003) used 16 SSRs to identify 212 peach and nectarine cultivars, 7 of which were used without prior knowledge of their level of variability in peach. Using random sets of markers or markers associated with genes for variety identification was reported by Oganisian et al. (1996) in potato, Sobotka et al. (2004) in oilseed rape, Singh and Ahuja (2006) in tea and Shirasawa et al. (2006) in rice.

A selected set of highly informative markers is expected to increase the efficiency of cultivar identification; however, the criteria for selecting such a panel of markers have not been well defined in the aforementioned studies involving crop plants. In previous reports, the allele frequency and genome location of markers were the most frequently used criteria. Candidate markers were selected based on high gene diversity (Lee et al. 2005; Krawczak 1999; Heaton et al. 2002), minor allele frequency greater than 0.1 (Werner et al. 2004), the absence of genetic linkage (Lee et al. 2005; Hochberg et al. 2003; Werner et al. 2004) and/or the necessity that alleles do not deviate from Mendelian inheritance (Heaton et al. 2002). In the present case, wherein 58 SNP loci were analyzed in 96 cultivars, a cluster analysis revealed clusters of SNP loci, whereby SNPs within a cluster possessed a similar ability to distinguish cultivars despite the fact that SNP loci within clusters are not physically linked. The elimination of SNPs that provide similar information in terms of distinguishing cultivars improves the efficiency of the core set, since the power for distinguishing cultivars will remain high, but with fewer SNPs to evaluate. Thus, in addition to selecting markers on the basis of allelic diversity and linkage group coverage, the 23 SNP marker loci chosen here involved a pairwise dissimilarity (based upon r^2, the widely used measure of linkage disequilibrium) criterion, to eliminate SNPs which are in sufficiently strong genome-wide LD. The selection was based on the cluster tree of SNPs, with one SNP selected from each of the sub-clusters, so that SNPs selected in this way would be the most effective set for cultivar identification. A similar strategy was successfully used for the identification of a core set of unique SSRs in soybean (Song et al. 1999), although the calculation of the distance matrix for SSRs was slightly different than that used here for the identification of maximally informative SNPs. This difference was the result of the multi-allelic nature of SSRs versus the bi-allelic SNPs used in the current study.

Single nucleotide polymorphisms are recognized as the most common source of soybean DNA diversity, and thus represent a virtually unlimited source of

molecular markers that can be used to distinguish genotypes. Because the SNP panel is currently optimized for the North American and Korean populations, it is likely to require modification for optimal utility in other populations. As greater ability to distinguish genotypes is required, additional SNPs can readily be added to the panel.

Comparison of the SSR and SNP panels for identifying cultivars

BARCSoySSR13 is a core set of SSRs selected by Song et al. (1999) for cultivar identification. Both BARCSoySSR13 and BARCSoySNP23 were used to characterize the 96 cultivars and the 36 genotypes from four different Maturity Groups. The average number of alleles per SSR locus observed in the 96 genotypes was 7.8 (range from 5 to 17), the mean SSR diversity was 0.73 (range from 0.60 to 0.83). The average genetic distance among genotypes was 0.75. In contrast, the average number of alleles per SNP locus, average diversity, and average pairwise distance determined by BARCSoySNP23 set were 2, 0.45 (range from 0.22 to 0.56) and 0.43, respectively. A similar trend of variation determined by SNPs and SSRs was also observed in the 36 MG I, II, IV and VI genotypes. It is clear from these data that the diversity and power of SSRs are higher than that of SNPs on a per locus basis, primarily because multi-allelism is a powerful determinant of informativeness. However, the lower informativeness of SNPs can be readily overcome by use of a larger number of SNP markers. The power of the BARCSoySSR13 to discriminate genotypes can be obtained by utilizing approximately 31 SNPs, assuming each of the SNPs had an average diversity of the SNPs in the BARCSoySNP23 panel. The cost of SNP assays varies with SNP assay methods. Lee et al. (2004) estimated the cost of reagents per simplex (per data point) reaction among four SNP genotyping assays to be $0.069–0.104. As SNPs continue to gain prominence in plant and animal genetic analysis, the cost effectiveness of SNP detection assays is likely to improve.

Acknowledgments The authors wish to express their thanks for the excellent technical assistance of Charles Quigley and Mike Livingston. This work was supported in part by a grant from the United Soybean Board whose support is gratefully acknowledged.

References

Alifrangis M, Enosse S, Pearce R, Drakeley C, Roper C, Khalil IF, Nkya WM, Ronn AM, Theander TG, Bygbjerg IC (2005) A simple, high-throughput method to detect *Plasmodium falciparum* single nucleotide polymorphisms in the dihydrofolate reductase, dihydropteroate synthase, and *P. falciparum* chloroquine resistance transporter genes using polymerase chain reaction- and enzyme-linked immunosorbent assay-based technology. Am J Trop Med Hyg 72:155–162

Aranzana MJ, Carbo J, Arus P (2003) Microsatellite variability in peach [Prunus persica (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. Theor Appl Genet 106:1341–1352

Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524–2525

Balamurugan K, Prabakaran N, Duncan G, Budowle B, Tahir M, Tracey M (2001) Allele frequencies of 13 STR loci and the D1S80 locus in a Tamil population from Madras, India. J Forensic Sci 46:1515–1517

Bashiardes E, Manoli P, Budowle B, Cariolou MA (2001) Data on nine STR loci used for forensic and paternity testing in the Greek Cypriot population of Cyprus. Forensic Sci Int 123:225–226

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408–1415

Chakraborty R, Meagher TR, Smouse PE (1988) Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genet. 118:527–536

Chen J, Iannone MA, Li MS, Taylor JD, Rivers P, Nelson AJ, Slentz-Kesler KA, Roses A, Weiner MP (2000) A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. Genome Res 10:549–557

Diwan N, Cregan PB (1997) Automated sizing of fluorescent labelled simple sequence repeat markers to assay genetic variation in soybean. Theor Appl Genet 95:723–733

Esselink GD, Smulders MJ, Vosman B (2003) Identification of cut rose (Rosa hybrida) and rootstock varieties using robust sequence tagged microsatellite site markers. Theor Appl Genet 106:277–286

Faruqi AF, Hosono S, Driscoll MD, Dean FB, Alsmadi O, Bandaru R, Kumar G, Grimwade B, Zong Q, Sun Z, Du Y, Kingsmore S, Knott T, Lasken RS (2001) High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. BMC Genomics 2:4

Gangitano DA, Garofalo MG, Juvenal GJ, Budowle B, Lorente JA, Padula RA (2002) STR data for the PowerPlex 16 loci in Buenos Aires population (Argentina). J Forensic Sci 47:418–420

Gizlice Z, Carter TE, Burton JW (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci 34:1143 1151

Glowatzki-Mullis ML, Gaillard C, Wigger G, Fries R (1995) Microsatellite-based parentage control in cattle. Anim Genet 26:7–12

Heaton MP, Harhay GP, Bennett GL, Stone RT, Grosse WM, Casas E, Keele JW, Smith TP, Chitko-McKown CG, Laegreid WW (2002) Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle. Mamm Genome 13:272–281

Heyen DW, Beever JE, Da Y, Evert RE, Green C, Bates SR, Ziegle JS, Lewin HA (1997) Exclusion probabilities of 22 bovine microsatellite markers in fluorescent multiplexes for semiautomated parentage testing. Anim Genet 28:21–27

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 33:54–78

Hochberg EP, Miklos DB, Neuberg D, Eichner DA, McLaughlin SF, Mattes-Ritz A, Alyea EP, Antin JH, Soiffer RJ, Ritz J (2003) A novel rapid single nucleotide polymorphism (SNP)-based method for assessment of hematopoietic chimerism after allogeneic stem cell transplantation. Blood 101:363–369

Hou P, Ji M, Li S, Lu Z (2004) Microarray-based approach for high-throughput genotyping of single-nucleotide polymorphisms with layer-by-layer dual-color fluorescence hybridization. Clin Chem 50:1955–1957

Keim P, Olson T, Shoemaker R (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newslett 15:150–152

Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21:12–27

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis 20:1676–1681

Lee SH, Walker DR, Cregan PB, Boerma HR (2004) Comparison of four flow cytometric SNP detection assays and their use in plant improvement. Theor Appl Genet 110:167–174

Lee HY, Park MJ, Yoo JE, Chung U, Han GR, Shin KJ (2005) Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. Forensic Sci Int 148:107–112

Luikart G, Biju-Duval MP, Ertugrul O, Zagdsuren Y, Maudet C, Taberlet P (1999) Power of 22 microsatellite markers in fluorescent multiplexes for parentage testing in goats (Capra hircus). Anim Genet 30:431–438

Mantel N (1967) The detecting of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Melendez E, Martinez-Espin E, Karlson IS, Lorente JA, Budowle B (2004) Population data on 15 STR loci (PowerPlex 16 kit) in a Costa Rica (Central America) sample population. J Forensic Sci 49:170–172

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Oganisian AS, Kochieva EZ, Ryskov AP (1996) [Fingerprinting of potato species and cultivar using RAPD-PCR]. Genetika 32:448–451

Olivier M, Chuang LM, Chang MS, Chen YT, Pei D, Ranade K, de Witte A, Allen J, Tran N, Curb D, Pratt R, Neefs H, de Arruda Indig M, Law S, Neri B, Wang L, Cox DR (2002) High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. Nucleic Acids Res 30:e53

Petkovski E, Keyser-Tracqui C, Hienne R, Ludes B (2005) SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. J Forensic Sci 50:535–541

Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, Pesich R, Hebert J, Chen YD, Dzau VJ, Curb D, Olshen R, Risch N, Cox DR, Botstein D (2001) High-throughput genotyping with single nucleotide polymorphisms. Genome Res 11:1262–1268

SAS Institute (1999) SAS/STAT User's Guide. Version 8. SAS Institute, Inc., Cary, NC

Schueler S, Tusch A, Schuster M, Ziegenhagen B (2003) Characterization of microsatellites in wild and sweet cherry (Prunus avium L.) markers for individual identification and reproductive processes. Genome 46:95–102

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

Singh D, Ahuja PS (2006) 5S rDNA gene diversity in tea (Camellia sinensis (L.) O. Kuntze) and its use for variety identification. Genome 49:91–96

Sobotka R, Dolanska L, Curn V, Ovesna J (2004) Fluorescence-based AFLPs occur as the most suitable marker system for oilseed rape cultivar identification. J Appl Genet 45:161–173

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

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109:122–128

Syn CK, Chuah SY, Ang HC, Lim SE, Tan-Siew WF, Chow ST, Budowle B (2005) Genetic data for the 13 CODIS STR loci in Singapore Chinese. Forensic Sci Int 152:285–288

Usha AP, Simpson SP, Williams JL (1995) Probability of random sire exclusion using microsatellite markers for parentage verification. Anim Genet 26:155–161

Weir BS (1990) Genetic data analysis methods for discrete genetic data. Sinauer Association, Sunderland

Werner FA, Durstewitz G, Habermann FA, Thaller G, Kramer W, Kollers S, Buitkamp J, Georges M, Brem G, Mosner J, Fries R (2004) Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. Anim Genet 35:44–49

Williams JL, Usha AP, Urquhart BG, Kilroy M (1997) Verification of the identity of bovine semen using DNA microsatellite markers. Vet Rec 140:446–449

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134