Identification and validation of a core set of informative genic SSR and SNP markers for assaying functional diversity in barley

R. K. Varshney · T. Thiel · T. Sretenovic-Rajicic · M. Baum · J. Valkoun · P. Guo · S. Grando · S. Ceccarelli · A. Graner

Received: 21 June 2007 / Accepted: 3 December 2007 / Published online: 28 December 2007

Abstract A 'core set' of 28 simple sequence repeat (SSR) and 28 single nucleotide polymorphism (SNP) markers for barley was developed after screening six diverse genotypes (DGs) representing six countries (Afghanistan, Pakistan, Algeria, Egypt, Jordan and Syria) with 50 SSR and 50 SNP markers derived from expressed sequence tags (ESTs). The markers of the core set are single locus with very high quality amplifications, high polymorphism information content (PIC) and are distributed across the barley genome. PIC values for the selected SSR and SNP markers ranged between 0.32–0.72 (average 0.58) and 0.28–0.50 (average 0.42), respectively. To make the SNP genotyping cost effective, CAPS (cleaved amplified polymorphic sequence) and indel assays were developed for 23 markers and the remaining 5 SNP markers were optimized for pyrosequencing. A high coefficient of correlations ($r = 0.96$, $P < 0.005$) between the genetic similarity matrices of SSR and SNP genotyping data of the core set on diverse genotypes (DGs) and their similar groupings according to the geographical distribution in both SSR and SNP phenograms with high bootstrap values underline the utility and reliability of the core set. A comparative allelic and sequence diversity for SSR and SNP markers between the DGs and six elite parental genotypes (PGs) of mapping populations showed comparable diverse nature of two germplasm sets. However, unique SNPs and indels were observed in both germplasm sets providing more datapoints for analysing haplotypes in a better way for the corresponding SNP marker.

Keywords Allele specific sequencing · CAPS · Genetic diversity · Nucleotide diversity · Polymorphism Information Content (PIC) · Pyrosequencing

Introduction

Detection and utilization of the genetic variation in crop plant genomes has been one of the most important tasks for plant geneticists and breeders

Electronic supplementary material The online version of this article (doi:10.1007/s11032-007-9151-5) contains supplementary material, which is available to authorized users.

R. K. Varshney · T. Thiel · T. Sretenovic-Rajicic · A. Graner
Leibniz Institute of Plant Genetics & Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany

Present Address:
R. K. Varshney (✉)
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502 324 Greater Hyderabad, AP, India
e-mail: r.k.varshney@cgiar.org

M. Baum · J. Valkoun · P. Guo · S. Grando · S. Ceccarelli
International Centre for Agricultural Research in Dry Areas (ICARDA), P.O. Box 5466, Aleppo, Syria

for understanding the genome architecture and also to devise strategies for crop improvement. The development and widespread adoption of molecular markers for genetical studies has provided a foundation for linking the phenotype to the genotype (see Lörz and Wenzel 2004). Over the past years many genetic diversity studies were performed in barley (*Hordeum vulgare* L.) (e.g., Powell et al. 1996; Russell et al. 1997, 2004; Fernandez et al. 2002; Matus and Hayes 2002). However, the different data sets are hardly comparable because of a lack of common core set of reference genotypes and the use of different marker systems. As to the latter, major problems arise from the comparison of complex banding patterns generated by generic marker assays such as random amplification of polymorphic DNA (RAPD) (Williams et al. 1990) and amplified fragment length polymorphism (AFLP) (Vos et al. 1995). On the other hand the integration of DNA fingerprinting data generated by microsatellite or simple sequence repeat (SSRs) (Tautz 1989) and single nucleotide polymorphism (SNP) (Coryell et al. 1999) markers entails less problems due to the simple banding patterns generated by SSRs and the unequivocal sequence information resulting from SNP analysis.

SSR markers are multiallelic and co-dominant in nature and therefore they have been developed in large number for all major crop plant species (Gupta and Varshney 2000). On the other hand SNPs are bi-allelic markers and represent the smallest units of genetic variation in genomes (Rafalski 2002). Although the development and genetic mapping of SNP markers is still underway in several crop plant species, these markers have already been successfully used for genetic diversity studies in barley (Kanazin et al. 2002; Bundock et al. 2003; Bundock and Henry 2004; Chiapparino et al. 2004; Russell et al. 2004).

The availability of a large set of expressed sequence tags (ESTs) for barley provides a resource for the systematic development of molecular markers including SSRs (see Varshney et al. 2005) and SNPs (Kota et al. 2001b, 2003, 2007). EST-derived SSR and SNP markers are a useful resource for assaying the functional genetic variation (Eujayl et al. 2001; Russell et al. 2004). It is, however, important to note that not all SSR and SNP markers are equal in terms of quality as well as information for genetic diversity studies. For instance, they can vary in robustness, quality of amplification products, amplification of single or multiple loci and also they may have lower information content.

The present study was undertaken to identify a core set of genic SSR and SNP markers for genotyping barley germplasm originating from semi-arid regions of the world. After screening a total of 100 SSR and SNP markers (derived from ESTs) on six genotypes, 28 SSR and 28 SNP markers, that were of highest quality, robust, highly informative (with high PIC values) and distributed across the barley genome, were selected for the core set. To make the application of selected SNP markers broader, cost effective CAPS assays were developed. In addition, sequence diversity of examined diverse genotypes was compared with the parental genotypes of three mapping populations of barley.

Materials and methods

Plant material

A set of six diverse barley (*Hordeum vulgare* subsp. '*vulgare*' L.) genotypes (DGs) obtained from the International Center for Agricultural Research in the Dry Areas (ICARDA) was used for screening with 50 SSR and 50 SNP markers. These diverse genotypes (DGs) included: IG28088 (Afghanistan, AFG), IG28159 (Pakistan, PAK), IG128170 (Algeria, DZA), IG128173 (Syria, SYR), IG128200 (Jordan, JOR) and IG128204 (Egypt, EGY). DNA was isolated from these genotypes as described by Thiel et al. (2003).

For comparing the sequence diversity between the DGs and parental genotypes (PGs) of three mapping populations i.e. Igri × Franka (Graner et al. 1991), Steptoe × Morex (Kleinhofs et al. 1993) and OWB$_{\mathrm{Dom}}$ × OWB$_{\mathrm{Rec}}$ (Costa et al. 2001), data generated earlier on PGs (Kota et al. 2001b; Kota et al. 2007) were included for analysis.

Marker analyses

A set of 50 SSR and 50 SNP markers derived from ESTs or genes were selected from the transcript map of barley (Stein et al. 2007). About three to four evenly spaced SSR and SNP markers were selected from each linkage group of barley (Table ESM 1).

SSR analysis

Amplification of microsatellite loci with fluorescent-dye labeled primer pairs was carried out as given in Thiel et al. (2003). Amplification products were separated on an ABI377 fragment analyzer and evaluated using GenoTyper 3.7 (Applied Biosystems, Foster City, CA, USA).

SNP analysis

EST-based SNP markers were used to amplify corresponding loci in DGs and allele-specific sequencing as well as SNP analysis was carried out according to Kota et al. (2001a, b; 2007).

Conversion of SNP markers into CAPS assay

Selected SNP markers of the core set were converted to CAPS markers by relating the SNP position to presence/absence of a restriction site in the panel of the six DGs examined by using "SNP2CAPS" tool (http://pgrc.ipk-gatersleben.de/snp2caps/; Thiel et al. 2004). Subsequently, corresponding restriction enzymes were tested on SNP-marker amplicons as described earlier (Thiel et al. 2004; Varshney et al. 2007a).

Optimization of pyrosequencing assay

For SNP genotyping by pyrosequencing, three primers are required: two PCR primers for PCR amplification of a SNP containing region and one sequencing primer for pyrosequencing the SNP containing (about 20-bp long) DNA fragment (Nyrén 2006). Primer pairs flanking SNPs were designed using software Assay Design (Biotage AB, Uppsala, Sweden). Depending on the nature of sequencing primer (for pyrosequencing in forward or reverse direction), one of the PCR primers was biotinylated at the 5′ end (Table ESM 2).

Amplification of SNP containing region in genome, optimization of pyrosequencing assay and pyrosequencing for five markers were performed on Pyrosequencer PSQ HS96 following instructions of manufacturer (Biotage AB, Uppsala, Sweden).

Diversity analysis

Polymorphism information content (PIC), nucleotide diversity (π)

The PIC values of SSR markers were calculated as given in Thiel et al. (2003). For SNP markers, the calculated nucleotide diversity (π), number of haplotypes, PIC of haplotype, and PIC of SNPs were calculated as described in Kota et al. (2003, 2007) and Thiel et al. (2004).

Phenogram preparation, bootstrap analysis and correlations of matrices

The profiles produced by SSR and SNP (including CAPS and pyrosequencing assays) markers were scored manually: each allele was scored as present (1) or absent (0) for each of the SSR and SNP loci.

Genetic similarities (GSs) were calculated for each pair of markers using the Jaccard's similarity coefficient with the help of NTSYS-pc 2.11 software package (Biostatistics Inc., USA, Rohlf 1998). SAHN clustering was employed for construction of UPGMA (Unweighted Pair Group Method of Arithmetic Average) phenograms. Bootstrapping was carried out using 10,000 iterations or re-sampling on PAUP* 4.0 Version 4.0b10 (for McIntosh) to evaluate the reproducibility of nodes of phenograms. Correlations between SSR and SNP matrices were calculated using Mantel test (Mantel 1967) after 10,000 random iterations with the help of Mental Nonparametric Test Calculator (Mantel version 2.0).

Results

Marker analyses

SSR-based allelic diversity

Out of the 50 markers used, only 47 markers showed polymorphism among six genotypes. The remaining three markers i.e. GBM1036 (2H), GBM1404 (6H) and GBM1456 (6H) were monomorphic in the six genotypes analysed. The 47 polymorphic SSRs yielded 2–4 alleles (average 2.7 allele per marker) and displayed PIC values between 0.13 and 0.52 (average 0.34).

SNP-based sequence diversity

Screening of 217–798 bp (average 412.2 bp) sequence data per marker yielded a total of 308 SNPs in 18,549 bp of non redundant sequence (Table 1). Thus the SNP frequency in the genotypes studied amounts to 1/60.2 bp. In addition to occurrence of SNPs, seven SNP markers namely GBS0154, GBS0182, GBS0400, GBS0214, GBS0535, GBS0554 and GBS0692 yielded a total of nine indels.

Of the 50 markers analyzed, 45 showed SNP-based polymorphism, while five markers namely GBS0131 (1H), GBS0613 (5H), GBS0360 (7H), GBS0693 (7H) and GBS0697 (3H) were monomorphic among six DGs examined (Table 1). Polymorphic markers detected 1 (GBS0582—1H, GBS0524—2H, GBS0214—3H, GBS0712—5H and GBS0537—7H) to 28 SNPs (GBS0535—2H) with an average of 6.69 SNPs per marker. The PIC value for individual SNPs varied between 0.27 and 0.50 with a mean value of 0.38.

The calculated nucleotide diversity index (π value) for these markers was observed in the range of 0.16×10^{-2} to 4.02×10^{-2} with a mean π value of 1.03×10^{-2}. Details on SNP diversity are presented in Table 1.

Comparative diversity between two germplasm sets

The SNP and SSR markers used in the present study were developed after mapping them in one of the three barley mapping populations i.e. Igri × Franka, Steptoe × Morex and $\text{OWB}_{\text{Dom}} \times \text{OWB}_{\text{Rec}}$ (Kota et al. 2001b, 2007; Varshney et al. 2006). Therefore, availability of genotyping and sequence data of these six parental genotypes (PGs) of mapping populations from the earlier studies allowed us to compare the diversity between PGs and DGs.

A comparison of sequence data of DGs with PGs for individual SNP markers revealed all the three possible cases: (i) the SNPs of PGs (similar) were retained in DGs, (ii) some SNPs of PGs were lost in DGs and (iii) some novel SNPs were observed in DGs that were not present in PGs. A summary on comparison of SNPs between PGs and DGs with 50 markers is given in Table 2. In total, 331 SNPs and 8 indels were obtained in PGs by 50 SNP markers while only 45 SNP markers (5 markers were monomorphic) revealed 308 SNPs and 9 indels. Between these two sets (DGs and PGs), only 231 (69.8%) SNPs and 8 indels were similar. A total of 100 (30.2%) SNPs of PGs were lost in DGs. However,

Table 1 Features of SNP diversity examined in diverse genotypes (DGs)[a]

	1H	2H	3H	4H	5H	6H	7H	Total
Used markers	7	7	8	7	6	7	8	50
Polymorphic markers	6	7	7	7	5	7	6	45
Sequence length examined (bp)	293–538 (372)	293–647 (453.71)	279–798 (437.62)	306–824 (440.57)	261–683 (419.66)	326–487 (394.57)	273–615 (397.62)	18,549 (416.52)
Average Pi (π)	0.0042−0.0325 (0.0189)	0.0042−0.0402 (0.0143)	0.0050−0.0303 (0.0113)	0.0042−0.0345 (0.0096)	0.0016−0.0225 (0.0104)	0.0066−0.0102 (0.0065)	0.0034–0.0199 (0.0093)	(0.0103)
Number of SNPs detected per markers	1−14 (7.5)	1−29 (8.42)	1−11 (5.74)	2−7 (3.42)	1−11 (6.4)	1−9 (3.5)	4−15 (7.2)	308 (6.69)
PIC range of SNPs	0.27−0.44 (0.33)	0.27−0.50 (0.40)	0.27−0.39 (0.33)	0.27–0.44 (0.39)	0.39–0.48 (0.45)	0.27−0.44 (0.38)	0.27−0.44 (0.36)	(0.38)
Number haplotypes obtained	2–6 (4.28)	2−6 (5)	2−6 (4.28)	2−6 (4)	2−6 (4)	3−6 (4)	2−6 (3.5)	(3.8)
PIC range of haplotypes	0.66–0.83 (0.76)	0.44−0.83 (0.66)	0.27−0.83 (0.67)	0.44−0.80 (0.66)	0.44−0.78 (0.67)	0.66−0.80 (0.65)	0.48−0.63 (0.55)	(0.62)

[a] Figures in parenthesis represent the mean/average value for the corresponding feature

Table 2 Comparison of occurrence of SNPs and indels between diverse genotypes (DGs) and parental genotypes (PGs) of mapping populations

Linkage group	No. of markers analyzed	Occurrences of SNPs and indels in diverse genotypes (DGs)		Occurrences of SNPs and indels in elite parental genotypes (PGs)		Similar SNPs and indels in PGs and DGs		Occurrences of new SNPs and indels in DGs		Loss of PGs' SNPs and indels		Net effect[a]	
		SNPs	indels	SNPs	indels	SNPs	indels	SNPs	indels	SNPs	indels	SNPs	indels
1H	7	44	2	33	1	25	1	19	1	8	0	+11	+1
2H	7	66	4	63	4	53	4	13	0	10	0	+3	0
3H	8	43	1	62	1	30	1	13	0	32	0	−19	0
4H	7	27	1	29	1	22	1	5	0	7	0	−2	0
5H	6	38	0	43	0	30	0	8	0	13	0	−5	0
6H	7	34	0	33	0	21	0	13	0	12	0	+1	0
7H	8	56	1	68	1	50	1	6	0	18	0	−12	0
Total	50	308	9	331	8	231	8	77	1	100	0	−23	+1

[a] Net gain and loss of SNPs in DGs are shown by '+' and '−', respectively

77 (25%) novel SNPs and one new indel was observed in DGs that were not present in PGs. Taken these observations together, a total of 23 SNPs were lost and 1 indel was gained. In terms of linkage groups, the SNP markers of 1H group gained the maximum (11) SNPs and markers of 3H group lost the maximum (19) SNPs (Table 2).

In addition to the above observation, PGs as compared to DGs showed a higher SSR allelic diversity, SNP frequency and haplotype diversity (Table 3). Nevertheless, the PIC value of SNPs and sequence diversity were slightly higher in DGs.

Comparison of SSR and SNP analyses for diversity analysis

In order to compare the potential of SSR and SNP markers for phenetic analysis, allelic data obtained for 47 SSR and 45 SNP markers respectively, were used to prepare the phenograms of six DGs. Both phenograms classified the examined DGs in almost similar way as three DGs (IG128088, IG128173 and IG128200) were grouped in one cluster, two DGs (IG128170 and IG128 204) in another cluster while one DG (IG128159) was distant to the above clusters (Fig. 1a, b). Bootstrap analyses (10,000 iterations) revealed comparatively higher level of confidence obtained for the branches of the SNP phenogram (Fig. 1b).

A comparison of two genetic similarity matrices obtained by SSR and SNP markers showed a highly significant correlation ($r = 0.98$, $P < 0.005$, 2.575, $g = 3.6921$) suggesting the principal equivalency of two marker assays for the phenetic analysis.

Core set of informative genic markers

As shown above, both types of markers are equally suitable for detection of genetic variation. Therefore on the basis of the above data, a total of 28 SSR and

Table 3 Comparative allelic and sequence diversity in two germplasm sets

Feature	Parental genotypes (PGs)	Diverse genotypes (DGs)
SSR diversity		
Number of alleles	2−5 (2.8)	2−4 (2.7)
PIC value	0.24−0.78 (0.62)	0.13−0.52 (0.34)
SNP diversity		
SNP frequency (bp)	1/59.1	1/60.7
Number of SNPs	331	308
Number of indels	8	9
Specific SNPs	100	77
PIC value of SNPs	0.24−0.50 (0.36)	0.27–0.50 (0.38)
Number of haplotypes	2–7 (3.89)	2–6 (3.80)
PIC value of haplotypes	0.24−0.85 (0.65)	0.27−0.83 (0.62)
Sequence diversity	0.0011−0.0395 (0.0087)	0.0016−0.0402 (0.0103)

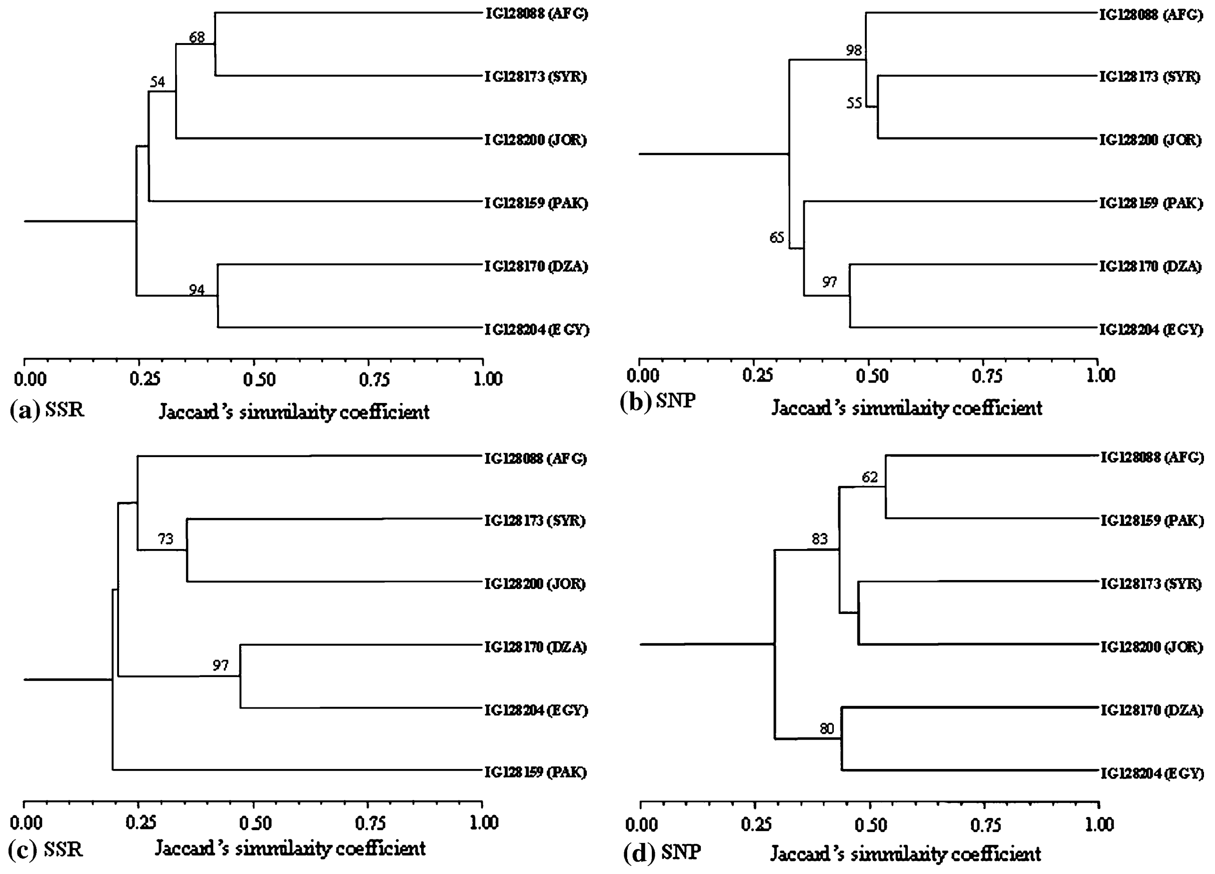


Fig. 1 Comparison of SSR and SNP phenograms for diversity analysis. A comparison of SSR and SNP phenograms of 6 DGs obtained by using 47 SSR and 45 SNP markers is shown in (a) and (b), respectively, while the phenograms of 6 DGs obtained by 28 SSR and 28 SNP markers (of core set) are shown in (c) and (d), respectively. Significant bootstrap values (>50) after resampling data for 10,000 times are shown on the nodes of phenograms

28 SNP markers were selected as a core set of informative gene-derived molecular markers for diversity studies in barley. While selecting markers for the core set, following criteria were considered: (1) they are single locus and provide good quality amplification, which are (2) distributed across the barley genome, and (3) exhibit reasonably high PIC values. The selected SSR and SNP markers of the core set along with number of alleles, PIC value and SNP assay optimized are given in Table 4. The genetic mapping position, as per consensus map of barley (Stein et al. 2007), the primer sequences, wherever possible and a putative function, deduced based on BLASTX analysis for the selected SSR and SNP markers are given in Table ESM 3 and Table ESM 4, respectively.

Selected SSR markers exhibit 2–4 (average 2.96) alleles with a PIC value of 0.32 to 0.72 (average 0.58) in the analysed set of 6 DGs (Table 4). The PIC values of identified SNP markers ranged from 0.28 to 0.50 with an average of 0.42. The selected SNP markers yield 1 to 29 SNPs (average 7.6) with 2 to 6 haplotypes (average 4.1) per marker. The haplotype based PIC values for these markers varied from 0.44 to 0.83 with an average of 0.67. Nucleotide diversity index (π value) for each of the selected SNP markers is in the range of 0.27×10^{-2} to 2.34×10^{-2} with a mean π value of 0.91×10^{-2}.

Development of CAPS assay for selected SNP markers

To make the SNP genotyping cost-effective in a large germplasm collection, targeted SNPs (with the higher PIC value) were investigated for presence of the

Table 4 Details on identified markers of the core set

Linkage group	SSR markers			SNP markers		
	Marker name	Alleles	PIC value	Marker name	Assay optimized[a]	PIC value
1H	GBM1007 (1HS)	4	0.72	GBS0546 (1HS)	*Sml*I	0.44
	GBM1029 (1HS)	2	0.50	GBS0554 (1HL)	*Hha*I	0.44
	GBM1013 (1HL)	2	0.44	GBS0361 (1HL)	*Hha*I	0.49
	GBM1461 (1HL)	4	0.72	GBS0528 (1HL)	*Hpy*CH4IV	0.49
2H	GBM1035 (2HS)	2	0.48	GBS0182 (2HS)	indel	0.41
	GBM1459 (2HS)	3	0.56	GBS0535 (2HS)	*Mse*I	0.50
	GBM1047 (2HL)	3	0.64	GBS0400 (2HL)	indel	0.28
	GBM1208 (2HL)	4	0.70	GBS0705 (2HL)	PS	0.44
3H	GBM1031 (3HS)	3	0.56	GBS0555 (3HS)	*Spe*I	0.48
	GBM1413 (3HS)	2	0.48	GBS0667 (3HS)	*Cac*8I	0.48
	GBM1059 (3HL)	4	0.72	GBS0431 (3HL)	*Rsa*I	0.49
	GBM1405 (3HL)	4	0.67	GBS0526 (3HL)	*Psi*I	0.49
4H	GBM1221 (4HS)	4	0.72	GBS0192 (4HS)	*Rsa*I	0.44
	GBM1323 (4HS)	3	0.61	GBS0692 (4HL)	indel	0.28
	GBM1003 (4HL)	4	0.72	GBS0288 (4HL)	*Hha*I	0.44
	GBM1015 (4HL)	3	0.56	GBS0461 (4HL)	PS_pos1_C/T	0.44
					PS_pos2_G/C	0.28
5H	GBM1176 (5HS)	2	0.48	GBS0527 (5HS)	*Eco*RV	0.44
	GBM1054 (5HL)	2	0.50	GBS0577 (5HS)	*Dde*I	0.50
	GBM1064 (5HL)	3	0.56	GBS0712 (5HL)	*Ava*II	0.28
	GBM1483 (5HL)	2	0.32	GBS0576 (5HL)	PS_pos1_G/T	0.49
					PS_pos2_C/T	0.49
					PS_pos3_G/C	0.44
6H	GBM1021 (6HS)	4	0.72	GBS0136 (6HS)	*Taq*I	0.44
	GBM1212 (6HS)	2	0.53	GBS0157 (6HS)	*Sal*I	0.44
	GBM1008 (6HL)	3	0.64	GBS0369 (6HL)	*Hae*III	0.44
	GBM1063 (6HL)	2	0.44	GBS0708 (6HL)	PS	0.28
7H	GBM1326 (7HS)	2	0.50	GBS0591 (7HS)	PS	0.44
	GBM1464 (7HS)	4	0.62	GBS0154 (7HS)	indel	0.28
	GBM1516 (7HS)	3	0.64	GBS0317 (7HL)	*Hha*I	0.28
	GBM1419 (7HL)	3	0.44	GBS0291 (7HL)	*Hinf*I	0.48

[a] Name of restriction enzymes for CAPS and PS for Pyrosequencing assays are given. In case of PS assays, pos1, pos2, pos3 represent different positions of SNPs, that were targeted in pyrosequencing assay

restriction enzyme recognition site with the help of SNP2CAPS tool (Thiel et al. 2004; http://pgrc.ipk-gatersleben.de/snp2caps/). Infact, four SNP markers (GBS0154, GBS0182, GBS0400 and GBS0692) showed occurrence of indel that made it possible to analyse these markers on standard agarose gels. The multiple sequence alignments (amplicon sequences for 6 DGs) for the remaining 24 selected SNP markers were subjected to identify potential restriction enzymes for assaying the SNPs. As a result, a total of 21 (87.5%) out of 24 alignments displayed at least one potential CAPS candidate; the remaining 3 alignments (for GBS0576, GBS0591 and GBS0705) did not provide any restriction enzyme recognition site or indel. Upon digestion of the corresponding PCR fragments, for 19 (90.5%) out of 21 marker-

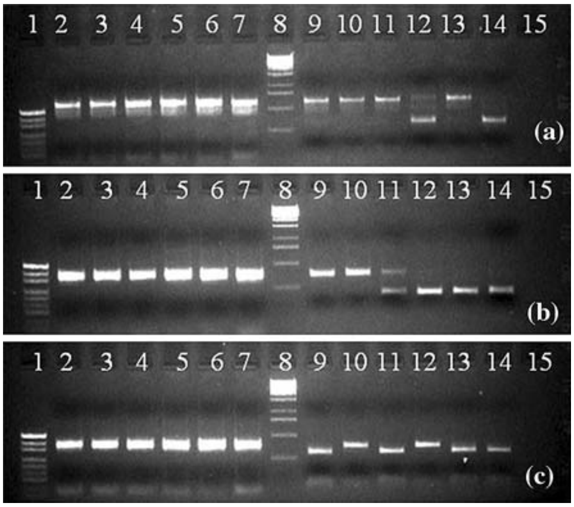


Fig. 2 Conversion of SNP markers into CAPS assays. Gel electrophoresis separation of cleaved amplicons has been shown for three markers: (**a**) GBS0554—*Hha*I, (**b**) GBS0361—*Hha*I and (**c**) GBS0288—*Hha*I. In all three panels (**a**, **b** and **c**) the gel lanes 1 and 8 contain DNA standards (size markers) as puC19/Msp23 and 1 kb DNA ladder, respectively and the other lanes contain DGs in following order: lanes 2, 9 = IG128088, lanes 3, 10 = IG128159, lanes 4, 11 = IG128170, lanes 5, 12 = IG128173, lanes 6, 13 = IG128200, lanes 7, 14 = 128204 and lane 15 = water. In each panel, lanes 2–7 contain the PCR amplicons and lanes 9–14 contain *Hha*I-digested/cleaved PCR amplicons of DGs obtained with corresponding markers

enzyme pairs the predicted restriction pattern could be revealed (Fig. 2). A complicated or unexpected banding pattern, however, was observed in remaining two markers i.e. GBS0461 (4H) and GBS0708 (6H). Thus it was possible to assay 23 SNP markers on agarose gel (19 CAPS and 4 indel assays).

Optimization of pyrosequencing assay

Pyrosequencing assay were optimized for five remaining SNP markers (GBS0461, GBS0576, GBS0591, GBS0705, GBS0708) (Table ESM 2). For GBS0461 and GBS0576, two and three SNPs (close to each other in the range of pyrosequencing), respectively while one SNP each for markers GBS0591, GBS0705 and GBS0708 were assayed in pyrosequencing. An example of assaying more than one SNP for one marker using pyrosequencing has been shown in case of GBS0461 (Fig. 3). The PIC values of two SNPs for the marker GBS0461 were 0.44 and 0.28 and of three SNPs for GBS0576 were 0.49, 0.49 and 0.44, respectively. The average PIC values for assayed SNPs for GBS0461 and GBS0576, however, were calculated as 0.36 and 0.47, respectively.

Evaluation of the core set

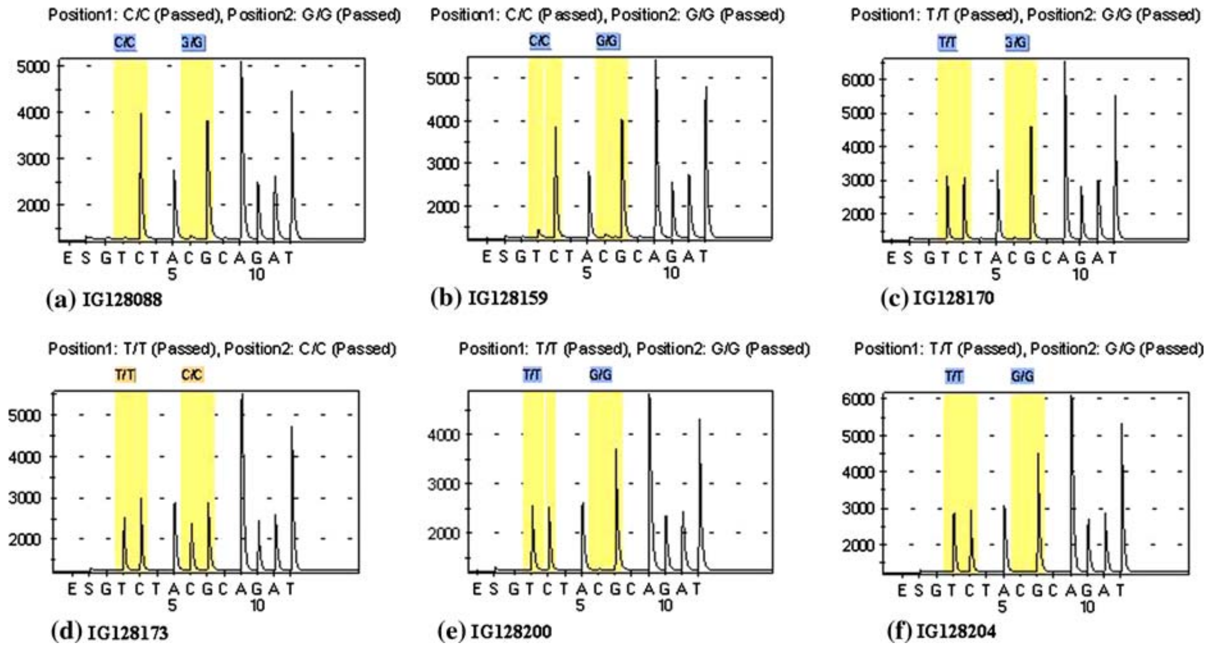
To compare the results of developed CAPS, indel and pyrosequencing assays for the selected SNP markers with the SSR markers of core set, genetic similarity matrices for both marker sets were compared that showed a high correlation ($r = 0.96$) and statistically significant ($P < 0.005$, 2.575, $g = 3.61$). Two phenograms prepared by using 83 and 62 datapoints obtained for 28 SSR and 28 SNP markers, respectively, were comparable to each other (Fig. 1c, d). While comparing these phenograms with the earlier mentioned phenograms (Fig. 1a, b), a minor change was noticed in both SSR and SNP phenograms. The SSR phenogram based on 28 markers (Fig. 1c) is similar to the earlier one prepared by using 47 SSR marker data (Fig. 1a) except the interchange of the positions of IG128088 and IG128200. While comparing two SNP phenograms, two DGs (IG128088 and IG128159), which were far apart earlier in the phenogram of 45 SNP markers (Fig. 1b), could be grouped together in one cluster in the phenogram of 28 SNP markers (Fig. 1d). The remaining two clusters in both SNP phenograms remained similar. Bootstrap values were still very high for the majority of the branches of dendrograms.

Discussion

Polymorphism and sequence diversity

After screening 50 SSR and 50 SNP markers, only 47 (94%) SSR and 45 (90%) SNP markers detected polymorphism among 6 DGs. However, it is noteworthy that the monomorphic markers in DGs were polymorphic in PGs of mapping populations as they were genetically mapped in at least one mapping population (Varshney et al. 2006; Kota et al. 2007).

A comparison of the sequence diversity between DGs and PGs revealed a reduced number of SNPs and indels in the DG set. Still both sets showed a



(a) IG128088 (b) IG128159 (c) IG128170

(d) IG128173 (e) IG128200 (f) IG128204

Fig. 3 Pyrosequencing assay for GBS0461 marker. Pyrograms of two SNP positions (C/T and G/C) for the marker GBS0461 are shown for (**a**) IG128088, (**b**) IG128159, (**c**) IG128170, (**d**) IG128173, (**e**) IG128200 and (**f**) IG128204. At the targeted two SNP positions, shown as highlighted regions, two genotypes i.e. IG128088 (**a**) and IG128159 (**b**) show the C, G alleles, three genotypes i.e. IG128170 (**c**), IG128200 (**e**) and IG128204 (**f**) show the T, G alleles, and the remaining genotype—IG128173 (**d**) shows the T and C alleles, respectively

larger number of unique SNPs. Identification of novel SNPs for the given markers are of importance for targeting them when the other SNPs are not successful for assaying for a given marker such as in primer extension assays. The analysis suggested a slightly lower SNP frequency in DGs (1/60.2 bp) as compared to PGs (1/59.1) for the assayed markers. The SNP frequency in the present study is relatively higher as compared to earlier reports ranging from 1/78 bp to 1/131 bp (Kanazin et al. 2002; Bundock et al. 2003; Russell et al. 2004). However it is important to note that the SNP frequency in a given species depends on the nature of marker/gene examined as well as the genotypes surveyed. The SNP markers used in the present study do not represent random set of genes, rather these were selected from a total resource of 220 SNP markers (Kota et al. 2007) based on high information content (e.g. SNP frequency, sequence diversity and PIC value) and hence showed higher SNP frequency.

The nucleotide diversity (as measured by π) ranged from 0.16×10^{-2} to 4.02×10^{-2} (mean 1.03×10^{-2}) in DGs. A wide range of sequence diversity has been observed in a recent study in barley (Russell et al. 2004) where diversity ranges from 0.21×10^{-2} to 1.89×10^{-2} in 24 diverse genotypes. Similar kinds of varying ranges in sequence diversity were observed in other plant species like *Arabidopsis* (Purugganan and Suddith 1999; Miyasahita et al. 1999), sugarbeet (Schneider et al. 2001), soybean (Zhu et al. 2003), rye (Varshney et al. 2007a), etc.

Although the SSR allelic diversity, SNP frequency, unique SNPs and number of haplotypes are slightly higher in PGs when compared to DGs, the nucleotide diversity and PIC values of SNPs are slightly higher in case of DGs. These analyses indicate that there is not much difference in the overall diversity between two germplasm sets, though we expected a more diversity in DGs examined in the present study as they were sampled from different geographical origins. However, it is noteworthy that two genotypes namely $\mathrm{OWB}_{\mathrm{Dom}}$ and $\mathrm{OWB}_{\mathrm{Rec}}$ present in PGs are quite diverse than usual cultivated barley genotypes (e.g. Igri, Franka, Steptoe and

Morex) as these represent dominant and recessive morphological marker spring barley stocks (Wolfe and Francowiak 1991). Nevertheless, the higher PIC value of SNPs and nucleotide diversity suggests that though DGs have relative less number of SNPs, these SNPs were more variant across the six genotypes. While the PGs had higher number of SNPs, these were specific to only one genotype ($\mathrm{OWB}_{\mathrm{Dom}}$ or $\mathrm{OWB}_{\mathrm{Rec}}$).

Core set of markers and their utility

Based on diversity analysis and information contents, a core set of informative SSR and SNP markers, representing all chromosome arms, was defined. The PIC values for the selected SSR markers of the core set ranged from 0.32 to 0.72 (average 0.58), which are a bit lower than the markers of 'genotyping set' of Macaulay et al. (2001), where they varied in the range of 0.08–0.94 (average 0.64). However, it should be noted here that (i) the selected SSR markers of the core set in present study are derived from ESTs and therefore they are generally expected to exhibit a lower PIC value as compared to genomic SSRs (Leigh et al. 2003; Varshney et al. 2005), and (ii) the present set of markers was selected based on the analysis of 6 genotypes while Macaulay et al. (2001) screened 24 genotypes. The PIC values for SNP markers of the core set were high in the range of 0.28–0.50 with an average of 0.42. Similarly, the nucleotide diversity (average 0.91×10^{-2}) for the selected SNP markers is quite reasonable. Furthermore, both SSR and SNP markers of the core set are gene-derived markers and a putative function is known for majority of these markers (Table ESM 3, 4). Indeed, it has been shown the different kinds of gene-based markers often yield similar and/or comparable results (Kota et al. 2001a; Russell et al. 2004; Varshney et al. 2007a), however this was not the case with anonymous markers such as RAPDs, AFLPs or genomic SSRs (Russell et al. 1997; Nybom 2004; Woodhead et al. 2005).

Reliability of the selected markers of the core set in the present study is reflected in two ways. Firstly, the coefficient of correlation between genetic similarity matrices of 28 SSR and 28 SNP markers is very high and highly significant ($r = 0.96$, $P < 0.005$). Infact as compared to the correlation ($r = 0.98$, $P < 0.005$) of similarity matrices of 46 SSR and 45 SNP markers, the coefficient of correlation is slightly lower. While comparing the correlation between different marker systems, the observed correlation between SSR and SNP markers in the present study is certainly higher in comparison to earlier reports (e.g., Russell et al. 1997). A possible reason for this is that two marker types used in the present study, as mentioned above, are derived from the expressed portion of the genome. However, majority of earlier studies have been based on marker types that may show a bias regarding sampling the expressed and the non expressed part of the genome. Secondly, comparison of phenograms based on markers of the core set (28 SSR and 28 SNP) and complete set (46 SSR and 45 SNP) showed comparable grouping of genotypes examined. Although only a small number of genotypes were analyzed in the present study, the results clearly demonstrate the potential of the core set. Further, in the present study, the majority of the branches of dendrograms were supported by higher bootstrap values, however this was not the case when a random and even comparatively larger random SSR and SNP marker datasets were used for preparing the dendrograms of seven barley genotypes (Kota et al. 2001a).

In terms of cluster analysis, although a small number of genotypes were examined, majority of genotypes were grouped according to their geographical distribution. For instance, the Fig. 1a and b show one stable cluster containing two African genotypes (IG128170, IG128204) and another cluster containing three genotypes, one each from Middle East (IG128173), South Asia (IG128088) and Africa (IG128200), while the remaining South Asian (Pakistan) genotype (IG128159) is a solitary genotype between two clusters. Similarly, the Fig. 1c shows similar clusters mentioned above i.e. one containing two African genotypes (IG128170, IG128204) and another one containing two genotypes from Middle East (IG128173, IG128200) and one South Asian genotype (IG128088) and the remaining South Asian (Pakistan) genotype (IG128159) remains aloof than these clusters. However, SNP phenogram shown in the Fig. 1d yields three distinct clusters containing two South Asian genotypes (IG128088, IG128159), two African genotypes (IG128170, IG128204) and remaining two genotypes (IG128173, IG128200) of Middle East.

Development of CAPS, indels and pyrosequencing assays

In order to allow for a broader application of the selected SNP markers, a set of 19 SNP markers was converted into CAPS assays and indel assay was applicable for four markers. For the remaining 5 markers, the pyrosequencing assays were optimised in the present study.

Although the optimisation of pyrosequencing assay initially requires technical skills, pyrosequencing is superior to other systems such as allele-specific sequencing, DHPLC (Kota et al. 2001b), microarray-based SNP assay (Kanazin et al. 2002), and SNaPshot for SNP analysis, because of its linear dose-response curve and high level of automation (Pettersson et al. 2003). Furthermore, pyrosequencing, like allele-specific sequencing, allows haplotype analysis for the corresponding marker/gene. Indeed, haplotype analysis as compared to individual SNP analysis is more informative as they exhibit higher PIC value and also useful for linkage disequilibrium studies (Ching et al. 2002). Occurrence of 7.6 SNPs per marker in pyrosequencing assays provides opportunities to analyse these markers for haplotype analysis, if required, by using pyrosequencing assay. In this way, the information content per assay can be substantially enhanced for germplasm analysis. However, pyrosequencing assay requires specialized instrumentation that is not available in most laboratories. In those situations, these five SNP markers can be assayed by some other SNP assays e.g. SNaPshot, allele specific PCR, etc.

In summary, the identified SSR and SNP markers provide a highly informative set of molecular markers, which are robust, easy to use, and easy to interpret and record. On one hand the SNP markers especially after converting them into CAPS assays offer a valuable source for genotyping genebank material as their results are recorded in the most easiest format (alphanumeric matrix) and amenable for storing in databases (Varshney et al. 2007b). The SSR markers, on the other hand will continue to be used in near future for genetic diversity studies because of their higher information content over the SNP markers.

Acknowledgements We are thankful to U. Beier for her excellent technical assistance and to Nils Stein for useful discussions. This work was financially supported by a GTZ project (no. 2002.7860.6-001.00 and Contract no. 1060503) sponsored by BMZ (Bundesministerium für Wirtschaftliche Zusammenarbeit und Entwicklung), Germany.

References

Bundock PC, Henry RJ (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. Theor Appl Genet 109:543–551

Bundock PC, Christopher JT, Eggler P, Ablett G, Henry RJ, Holton TA (2003) Single nucleotide polymorphisms in cytochrome P450 genes from barley. Theor Appl Genet 106:676–682

Chiapparino E, Lee D, Donini P (2004) Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. Genome 47:414–420

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, halpotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Coryell VH, Jessen H, Schupp JM, Webb D, Keim P 1999 Allele-specific hybridization markers for soybean. Theor Appl Genet 98:690–696

Costa JM, Corey A, Hayes PM, Jobet C, Kleinhofs A, Kopisch-Obusch A, Kramer SF, Kudrna D, Li M, Riera-Lizarazu O, Sato K, Szucs P, Toojinda T, Vales MI, Wolfe RI (2001) Molecular mapping of the Oregon Wolfe Barleys: a phenotypically polymorphic doubled-haploid population. Theor Appl Genet 103:415–424

Fernandez ME, Figueiras AM, Benito C (2002) The use of ISSR and RAPD markers for detecting DNA polymorphism, genotype identification and genetic diversity among barley cultivars with known origin. Theor Appl Genet 104:845–851

Eujayl I, Sorrells M, Baum M, Wolters P, Powell W (2001) Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. Euphytica 119:39–43

Graner A, Jahoor A, Schondelmaier H, Siedler K, Pillen K, Wenzel G, Herrmann RG (1991) Construction of an RFLP map of barley. Theor Appl Genet 83:250–256

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163–185

Kanazin V, Talbert H, See D, DeCamp P, Nevo E, Blake T (2002) Discovery and assay of single nucleotide polymorphsims in barley (*Hordeum vulgare*). Plant Mol Biol 48:529–537

Kleinhofs A, Kilian A, Saghai Maroof M, Biyashev R, Hayes P, Chen FQ, Lapitan N, Fenwick A, Blake TK, Kanazin V, Ananiev E, Dahleen L, Kudrna D, Bollinger J, Knapp SJ, Liu B, Sorrells M, Heun M, Franckowiak JD, Hoffman D, Skadsen R, Steffenson BJ (1993) A molecular isozyme and morphological map of barley (*Hordeum vulgare*) genome. Theor Appl Genet 86:705–712

Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001a) Generation and comparison of EST-derived SSRs and

SNPs in barley (*Hordeum vulgare* L.). Hereditas 135: 145–151

Kota R, Wolf M, Michalek W, Graner A (2001b) Application of DHPLC for mapping of single nucleotide polymorphisms (SNPs) in barley (*Hordeum vulgare* L.). Genome 44:523–528

Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Mol Genet Genom 270:224–233

Kota R, Varshney RK, Prasad M, Zhang H, Stein N, Graner A (2007) EST-derived single nucleotide polymorphism (SNP) markers for assembling genetic and physical maps of the barley genome. Funct Integr Genom. doi:10.1007/s10142-007-0060-9

Leigh F, Lea V, Law J, Wolters P, Powell W, Donini O (2003) Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. Euphytica 133:359–366

Lörz H, Wenzel G (2004) Molecular marker systems in plant breeding and crop improvement. Springer Verlag, Germany, pp 476

Macaulay M, Ramsay L, Powell W, Waugh R (2001) A representative, highly informative ‘genotyping set’ of barley SSRs. Theor Appl Genet 102:801–809

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

Matus IA, Hayes PM (2002) Genetic diversity in three groups of barley germplasm assessed by simple sequence repeats. Genome 45:1095–1106

Miyashita NT, Kawabe A, Innan H (1999) DNA variation in the wild plant Arabidopsis thaliana revealed by amplified fragment length polymorphism analysis. Genetics 152: 1723–1731

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol 13:1143–1155

Nyrén P (2006) The history of pyrosequencing. Methods Mol Biol 373:1–14

Pettersson M, Bylund M, Alderborn A (2003) Molecular haplotype determination using allele-specific PCR and pyrosequencing. Genomics 82:390–396

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Purugganan MD, Suddith JI (1999) Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics 151:839–848

Rafalski A (2002) Application of single nucleotide polymorphsims in crop genetics. Curr Opin Plant Biol 5:94–100

Rohlf FJ (1998) NTSYS-pc numerical taxonomy and multivariate analysis system. Version 2.02. Exeter Publications Setauket, New York

Russell JR, Fuller JD, Macaulay M, Hatz BG, Jahoor A, Powell W, Waugh R (1997) Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theor Appl Genet 95:714–722

Russell J, Booth A, Fuller J, Harrower B, Hedley P, Machray G, Powell W (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley. Genome 47: 389–398

Schneider K, Weisshaar B, Borchardt DC, Salamini F (2001) SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. Mol Breed 8:63–74

Stein N, Prasad M, Scholz U, Thiel T, Zhang H, Wolf M, Kota R, Varshney RK, Perovic D, Grosse I, Graner A (2007) A 1000 loci transcript map of the barley genome- new anchoring points for integrative grass genomics. Theor Appl Genet 114:823–839

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA. Nucleic Acids Res 17:6463

Tenaillon MI, Sawkins MC, Long AD, Gaut B, Doebley JF, Brandon S (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp mays L.). Proc Natl Acad Sci (USA) 98:9161–9166

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422

Thiel T, Kota R, Grosse I, Stein N, Graner A (2004) SNP2CAPS: a SNP and InDel analysis tool for CAPS marker development. Nucleic Acids Res 32(1):e5

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Varshney RK, Prasad M, Graner A (2004) Molecular marker maps of barley: a resource for intra- and interspecific genomics. In: Lörz H, Wenzel G (eds) Molecular markers in improvement of agriculture and forestry. Springer Verlag, Germany, pp 229–243

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers: features and applications. Trends Biotechnol 23:48–55

Varshney RK, Grosse I, Hahnel U, Thiel T, Rudd S, Zhang H, Prasad M, Stein N, Langridge P, Graner A (2006) Genetic mapping and physical mapping (BAC-identification) of EST-derived microsatellite markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 113:239–250

Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A (2007a) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. Theor Appl Genet 114:1105–1116

Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A (2007b) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Sci 173:638–649

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Fritjers A, Pot J, Peleman J, Kupier M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection in maize. Nature 398:236–239

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Wolfe RI, Franckowiak JD (1991) Multiple dominant and recessive genetic marker stocks in spring barley. Barley Genet Newsl 20:117–121

Woodhead M, Russell J, Squirrell J, Hollingsworth PM, Mackenzie K, Gibby M, Powell W (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Mol Ecol 14:1681–1695

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