Funct Integr Genomics (2008) 8:223–233
DOI 10.1007/s10142-007-0060-9

ORIGINAL PAPER

EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome

R. Kota · R. K. Varshney · M. Prasad · H. Zhang · N. Stein · A. Graner

Received: 3 July 2007 / Revised: 17 August 2007 / Accepted: 15 September 2007 / Published online: 30 October 2007

Abstract In a panel of seven genotypes, 437 expressed sequence tag (EST)-derived DNA fragments were sequenced. Single nucleotide polymorphisms (SNPs) that were polymorphic between the parents of three mapping populations were mapped by heteroduplex analysis and a genome-wide consensus map comprising 216 EST-derived SNPs and 4 *InDel (insertion/deletion)* markers was constructed. The average frequency of SNPs amounted to 1/130 bp and 1/107.8 bp for a set of randomly selected and a set of mapped ESTs, respectively. The calculated nucleotide diversities (π) ranged from 0 to 40.0×10^{-3} (average 3.1×10^{-3}) and 0.52×10^{-3} to 39.51×10^{-3} (average 4.37×10^{-3}) for random and mapped ESTs, respectively. The polymorphism information content value for mapped SNPs ranged from 0.24 to 0.50 with an average of 0.34. As expected, combination of SNPs present in an amplicon (haplotype) exhibited a higher information content ranging from 0.24 to 0.85 with an average of 0.50. Cleaved amplified polymorphic sequence assays (including *InDels*) were designed for a total of 87 (39.5%) SNP markers. The high abundance of SNPs in the barley genome provides avenues for the systematic development of saturated genetic maps and their integration with physical maps.

Keywords Molecular markers · SNPs · Haplotype diversity · Nucleotide diversity · Genetic map

Both R. Kota and R.K. Varshney contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10142-007-0060-9) contains supplementary material, which is available to authorized users.

R. Kota
Plant Disease Resistance Group, CSIRO–Plant Industry,
P.O. Box 1600, Canberra ACT 2601, Australia

R. K. Varshney
International Crops Research Institute for the
Semi-Arid Tropics (ICRISAT),
Patancheru 502324 (A.P.), India

M. Prasad
National Institute for Plant Genome Research (NIPGR),
New Delhi 110061, India

H. Zhang
Laboratory of Molecular Plant Physiology, University of Florida,
P.O. Box 110300, Gainesville 32611-300, USA

R. Kota · R. K. Varshney · M. Prasad · H. Zhang ·
N. Stein · A. Graner (✉)
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
Corrensstrasse 3, D-06466 Gatersleben, Germany
e-mail: graner@ipk-gatersleben.de

Introduction

Detection of genetic variation in crop plant genomes is an important prerequisite for understanding the genome architecture and to devise strategies for crop improvement. In this context, molecular markers represent an important tool and, hence, have been developed for all major crop plant species. In barley (*Hordeum vulgare* L.), the full spectrum of molecular markers is available; however, taking into consideration its large genome size (~5000 Mb), marker resources still need to be enhanced (reviewed by Varshney et al. 2004). The availability of a large set of expressed sequence tags (ESTs; Zhang et al. 2004) provides an opportunity for the systematic development of gene-based molecular markers to further saturate the genetic maps of barley.

Single nucleotide polymorphisms (SNPs) are the most common class and the smallest unit of genetic variation present in genomes (Cho et al. 1999; Picoult-Newberg et al. 1999; Rafalski 2002). Marker technologies exploiting the potential of SNPs provide the possibility of constructing genetic maps at 100-fold-higher marker density than by other types of DNA polymorphisms. Given the availability of the complete genome sequence information from more than a single genotype, SNP marker density can be determined at a kilobase scale as was shown in case of human (Sachidanandam et al. 2001), *Arabidopsis* (Schmid et al. 2003; Torjek et al. 2003), and rice (Nasu et al. 2002; Feltus et al. 2004). In contrast, the current genetic maps of crop species like barley and wheat provide a resolution only at the megabase level, and the availability of complete genome sequence data for such species is not in sight in the near future. Because of their high density/frequency and their lower mutation rate compared to microsatellite markers, SNP markers provide a powerful resource for genomewide linkage disequilibrium and association genetics studies, for studying genetic diversity, and for their deployment in marker-assisted breeding (Rafalski 2002).

Rapid advances in genotyping technologies make SNP markers an ideal tool for high throughput applications in plant genetics and breeding. As a consequence, the identification and mapping of SNPs has been initiated recently for crop species like rice (Nasu et al. 2002; Feltus et al. 2004), maize (Tenaillon et al. 2001; Ching et al. 2002; Batley et al. 2003; http://www.cerealsdb.uk.net/maize_snips), wheat (Somers et al. 2003, http://wheat.pw.usda.gov/SNP/), soybean (Zhu et al. 2003; Van et al. 2004), sugarbeet (Möhring et al. 2004), and sorghum (Hamblin et al. 2004). Also, in barley, SNP discovery and their application in genotyping of germplasm collections (Kota et al. 2001b; Kanazin et al. 2002; Bundock et al. 2003; Bundock and Henry 2004; Russell et al. 2004) and a SNP map based on abiotic stress responsive genes (Rostoks et al. 2005) have been reported. However, to be most effective, especially for genomewide association studies, the availability of a larger number of SNP markers evenly distributed throughout the whole genome is a prerequisite.

In the present study, we investigated the SNP frequency in the barley transcriptome and developed a genomewide set of >200 SNP markers for barley by relying on allele-specific sequencing and *in silico* SNP mining in EST databases. About 40% of the mapped SNP markers were converted into cleaved amplified polymorphic sequence (CAPS) markers, providing a cost-effective marker resource more or less independent of sophisticated laboratory equipment or expensive consumables.

Materials and methods

Plant materials

In the present study, all parental genotypes of three doubled-haploid (DH) mapping populations, i.e., Igri × Franka (IF, Graner et al. 1991), Steptoe × Morex (SM, Kleinhofs et al. 1993), and OWB_{Rec} × OWB_{Dom} (Oregon Wolfe Barley, OWB; Costa et al. 2001), were employed (70 DH lines of IF, 94 DH lines each of SM, and OWB respectively) together with cultivar Barke. DNA was prepared as described previously (Graner et al. 1991).

DNA amplification

To amplify genomic DNA by polymerase chain reaction (PCR), primer pairs were designed using EST sequence of *H. vulgare* cv. Barke, which are available from the CR-EST database (http://pgrc.ipk-gatersleben.de/cr-est/) as input to the software Primer Express (Applied Biosystems, Foster City, CA, USA). PCR was conducted with genomic DNA of the seven barley genotypes listed above, which primarily resulted in PCR fragments of 350–450 bp of average length. PCR was done in 20 μl reactions as described earlier (Kota et al. 2001b).

Detection and mapping of SNPs

For identification of SNPs, PCR products were sequenced in both forward and reverse orientation on an ABI 377XL automated sequencer using big dye-terminator chemistry (Applied Biosystems, Foster City, CA, USA). Base calling was carried out using Phred (Ewing et al. 1998). EST sequences were quality trimmed (sliding windows of 50 bp with a minimal average Phred score of 20) and filtered for a minimum length of 100 bp. In the first instance, after completion of sequence data check for sequencing error, the software "Sequencher" (Gene Codes, Ann Arbor, MI, USA) was used to generate contigs from forward and reverse sequence of each genotype under the following parameters: minimum match percentage, 85; minimum of overlap, 20 bases; and assembly algorithm, dirty data. Doubtful base calls were visually inspected by checking the sequence trace file. Subsequently, contigs for all the seven genotypes were aligned using either GCG pileup or ClustalW (Gribskov et al. 1984; Thompson et al. 1994) and checked manually to identify SNPs. Sequence alignments and marker information are available at the website http://pgrc.ipk-gatersleben.de/barley_snp/. Polymorphisms observed between the parental genotypes of any mapping population were evaluated and mapped by utilizing denaturing high-performance liquid chromatography (DHPLC) assays as previously described (Kota et al. 2001b).

Polymorphism information content and nucleotide diversity index (π)

Polymorphism information content (PIC) value or expected heterozygosity was calculated as described by Nei (1987) using the algorithm

$$PIC = l - \sum_{i=l}^{m} p_i^2$$

where m denotes the total number of alleles and p the frequency of the ith allele at a genetic locus.

Genetic variability in DNA sequences was measured by the nucleotide diversity index (π), with $\pi = K/L$. K is defined by pairwise sequence comparisons as the average number of differing nucleotide sites in a DNA sequence of length L (in bp; Nei and Li 1979). The standard deviation of π was calculated according to Hartl and Clark (1997).

Linkage mapping and nomenclature of SNPs markers

Linkage analysis was performed in one of the three mapping populations listed above. One hundred seventeen BIN or anchor markers available on the genetic maps (Kleinhofs and Graner 2001), as well as additional restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) anchor markers recently developed in our lab (Stein et al. 2007; Varshney et al. 2006) were used to prepare a consensus map using JoinMap ver. 3.0 programme using a logarithm of the odds (LOD) score of 3.0 (Stam 1993).

Mapped markers are coded as Gatersleben Barley SNP (GBS) followed by a four-digit numerical code as locus identifier. Additional information linked to each SNP marker includes the corresponding EST (EMBL accession ID), SNP position, presence of *insertions or deletions (InDels)*, etc., following the recommendations of the Nomenclature Working Group for Human Gene Mutations (Beutler et al. 1996; Antonarakis et al. 1998; den Dunnen and Antonarakis 2000) with some modifications (see ESM Table 1). A brief description on nomenclature of SNP markers derived from ESTs for barley is given below:

1. The presence of a SNP in a given genomic DNA sequence (in the absence of intronic regions) after amplification by using a defined primer set is given by including marker ID (laboratory specific) followed by EST ID (as per public domain, EMBL/GenBank/DDBJ databases) and position and type of nucleotide change in relation to the sequence data of the EST. For example, GBS0001_AL509356.485G>A represents a 'G>A' SNP present at nucleotide position 485 in the EST AL509356 from which marker GBS0001 has been developed.
2. Two or more SNPs in the same locus are listed within brackets, separated by a semicolon, e.g., GBS0031_AL503315. [137C>G; 143T>A; …].
3. Deletions or insertions of a few basepairs are designated by 'del' and 'ins,' respectively, preceded by the indication of their basepair position in relation to EST sequence. Their length is written in subscript, e.g., GBS0530_AL510162.282del$_1$C represents a 1-bp deletion (bp position 282) relative to EST AL510162; and GBS0177_AL499652.271–272ins$_3$AAG represents a 3-bp long 'AAG' insertion between bp 271 and 272 of EST AL499652. If no sequence information is available on deletions and insertions, they can be specified by a question mark, e.g., GBS0179_AL450603.?delins.
4. Introns are designated as intron variable sequence (IVS) and can be specified as mentioned above in case of insertions (see '3'). SNPs in intronic regions are designated by their position in intronic regions in a similar way as mentioned above in '1' and '3'; e.g. GBS0008_AL509087.404–405IVS$_{107}$ 55C>T, refers to the presence of a 107-bp-long intron starting at bp 404 position and '55C>T' is the SNP at basepair position 55 in this intron. If a particular sequence contains SNPs in intronic, as well as in exonic regions, the SNP is designated as GBS0132_AL503243.[103–104IVS$_{291}$102C>A; 156T>C; 239A>T], where 103–104IVS$_{291}$102C>A refers to a 'C>A' SNP at basepair position 102 in an intron of 291 bp present at nucleotide 103 in EST AL503243, 156T>C represents a 'T>C' SNPs at basepair position 156 and 239A>T is a 'A>T' SNPs present at basepair position 239.

Functional annotation

Mapped SNP-containing sequences (SNP-ESTs) were compared to the NR-PEP protein database of June 2005 (Refseq-release 11) at the Husar, DKFZ, Heidelberg, using the Blastx2 program (Altschul et al. 1990), using a threshold value <1E-10 (for details see http://genome.dkfz-heidelberg.de/).

Conversion of mapped SNP markers into CAPS assays

Mapped SNP markers were converted to CAPS markers by relating the SNP position to the presence/absence of a restriction site in amplicons derived from the panel of seven genotypes examined. To achieve this, sequence alignments from the seven genotypes obtained by the program ClustalW (Thompson et al. 1994) were loaded in Fasta format into the 'SNP2CAPS' tool (http://pgrc.ipk-gatersleben.de/snp2caps/; Thiel et al. 2004), which employed the Rebase database (version 304, March 24, 2003)

containing the recognition sequence information of a total of 235 non-isoschizomeric and commercially available restriction enzymes. Subsequently, a set of 45 restriction enzymes was tested on SNPs-marker amplicons as described earlier (Thiel et al. 2004).

Results

SNP discovery and frequency

To obtain sequence information for seven genotypes, PCR primer pairs were developed from a set of 710 unigene EST sequences derived from cultivar Barke (Zhang et al. 2004). Four hundred thirty-seven (62%) yielded a single amplicon and could be sequenced. Of the remaining, 92 (13%) primer pairs failed to produce any PCR product, and 182 (26%) showed either multiple or weak amplicons and, thus, were dropped from further analysis without further attempts to optimize PCR conditions. A comparison of the amplified DNA sequences to the consensus EST-based sequence from cultivar Barke revealed that 143 (33%) amplicons exhibited greater than the predicted size because of the presence of intron(s) in the target genomic sequence.

In total, 163,828 bp of non-redundant sequence data were scanned leading to the identification of 1,257 SNPs with an overall SNP frequency of 1 SNP per 130 bp (Table 1). Among these, 1,125 (89%) were rated as common SNPs (occurring in more than one genotype). In the total set of SNPs, transitions accounted for 717 (57%) and transversions for 540 (43%), respectively. This difference is statistically significant (χ^2=19.18, p<0.001, df=1). The relative value of the characterized SNPs, as it was determined by calculating the PIC of haplotypes based on the seven genotypes, ranged from 0 to 0.85 with a mean PIC value of 0.34. The nucleotide diversity index (π) ranged from 0 to 40×10^{-3} with a mean of 3.10×10^{-3} (SD=±0.006; Table 1, Fig. 1).

Considering only the mapped set of markers (see below), the average SNP frequency was 1/107.8 bp in the seven genotypes analyzed. Similarly, the nucleotide diversity index (π) for these SNPs ranged from 0.05×10^{-2} to 39.51×10^{-3} with a mean of 4.37×10^{-3} (Fig. 1, ESM Table 1). Among the total of 942 SNPs identified in mapped markers, 554 (58.8%) were caused by transitions and 388 (41.2%) to transversions.

Table 1 Summary of SNP discovery in barley

Parameter	Complete set	Mapped markers
Number of loci screened	437	220
Total length of sequence analyzed (in basepairs)	163,828	101,483
Number of SNPs identified	1,257	942
Transition/transversion ratio	1.33	1.43
Frequency of SNPs	1/130 bp	1/107.8 bp
Average nucleotide diversity	3.10×10^{-3}	4.73×10^{-3}
Average PIC value of haplotypes	0.34	0.50

Genetic mapping

In the process of genetic mapping of the newly developed SNP markers, the highest level of polymorphism was revealed between the parental genotypes of the OWB population, i.e., 193 of the 437 EST (44%) were polymorphic. Between Steptoe and Morex, 158 of the 437 ESTs (36%) and, between Igri and Franka, 74 of the 437 ESTs (17%) could be potentially mapped. Markers polymorphic in more than one mapping populations were mapped in only one population and, thus, a total of 13, 78, and 129 markers were mapped in the IF, SM, and OWB population, respectively.

The marker segregation data from all three mapping populations (I × F, S × M, OWB) were subsequently used to prepare a consensus map. In total, 220 SNP markers were mapped to the seven linkage groups spanning an overall genetic distance of 1,136 cM (Fig. 2, ESM Table 1). Linkage group 5H exhibited the highest number of markers (41), whereas linkage group 4H exhibited the lowest number of markers (22; Table 2; ESM Table 1).

Expected heterozygosity

SNP markers are mainly biallelic, and therefore, their information content (PIC) can not exceed 0.50. However, if combinations of SNPs present within an amplicon are considered as a haplotype, higher PIC values can be expected as the result of the presence of multiple alleles. In the present study, two to seven (average of 3.0) haplotypes were observed per amplicon, giving rise to a mapped marker (ESM Table 1). The haplotype-based PIC values ranged from 0.24 to 0.85 (average 0.50), whereas the PIC values for the individually mapped SNPs were in the range of 0.24 to 0.50 with an average of 0.34. Thus, the analysis of haplotypes instead of individual SNPs would be more informative for genetic diagnostics. In fact, approximately 44.1% of the amplicons yielded a haplotype-PIC of >0.50, whereas only four SNPs reached the optimal PIC value of 0.50 (ESM Table 1). As an example for marker GBS0546, the detection of 18 SNPs resulted in ten different haplotypes for the seven genotypes included in the analysis (Table 3). The PIC value for the individual SNPs, however, was in the range of 0.25 to 0.49 (average of 0.33), as compared to 0.85 at the haplotype level. Keeping in view the importance of informativeness of haplotype analysis, a

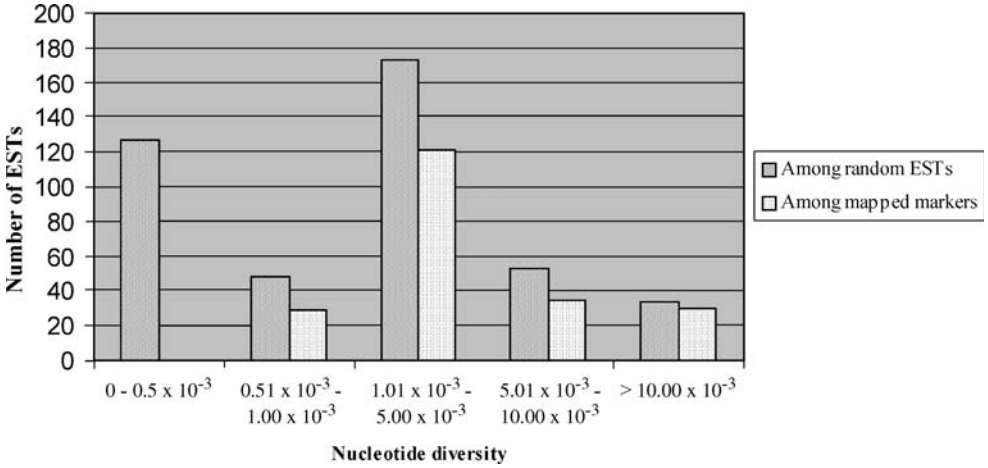

Fig. 1 Distribution of nucleotide diversity (π) in barley ESTs. Random ESTs used for allele-specific sequencing in seven barley genotypes (Igri, Franka, Steptoe, Morex, OWB_{Dom}, OWB_{Rec}, and Barke) are represented as *dark*, whereas mapped ESTs are shown as *white* columns

set of 28 SNP markers, randomly distributed on all the linkage groups (generally representing each chromosome arm), was identified that provides a maximal information content (ESM Table 2).

Functional annotation

Since the presented SNP markers were derived from EST, a putative function may be assigned to the underlying genes based on a comparison to a protein sequence database. After Blastx analysis to the non-redundant protein (NR-PEP) database of GenBank (National Center for Biotechnology Information, NCBI), a putative function was deduced for 171 (77.8%) markers (ESM Table 1). Among them, 96 (56.1%) markers showed homology to known proteins, 55 markers (32.2%) to putative proteins, 18 (10.5%) to unknown/unnamed proteins, and 2 (1.2%) to hypothetical proteins. The remaining 49 (22.2%) markers did not show a homology to any protein sequence represented by the database.

Development of CAPS assays for the SNP markers

Many SNP detection and genotyping platforms currently depend or rely on expensive equipment or consumables and may result in considerable costs per data point. To allow for a broader application of the presented SNP markers, a set of SNP markers was converted into CAPS assays after identification of restriction enzyme recognition sites. Multiple sequence alignments (amplicon sequences for seven genotypes) for all the mapped SNP markers were subjected to identify potential restriction enzymes for assaying the SNPs. A total of 203 (91.8%) out of 220 alignments displayed at least one potential CAPS candidate when the set of 235 commercially available non-isoschizomeric restriction enzymes was applied to the data set (ESM Table 1). As expected, the number of potential CAPS candidates decreased to 128 (57.9%) when only 45 common enzymes (common and relatively less expensive) were taken into account. Subsequently, all the CAPS candidates identified with the 45 restriction enzymes underwent experimental verification, and for 82 (64.1%), the predicted and unequivocal restriction pattern could be revealed (Fig. 3). In addition, five markers namely GBS0179, GBS0182, GBS0214, GBS0318, and GBS0539 could be assayed as *InDel* markers, and thus, a total of 87 SNP markers can be assayed as CAPS or *InDel* markers on agarose gel. These 87 SNP markers were distributed over all the linkage groups and chromosome arms ranging from 10 (2H and 4H) to 14 (5H and 7H) markers per chromosome (Table 2). An informative or core set of 28 SNP/CAPS markers exhibiting high PIC values was identified, which is randomly distributed on all linkage groups, and represents most of the chromosome arms (ESM Table 2).

Discussion

The present study was undertaken to develop a resource of mapped SNP markers for barley. This was initiated by using the sequences from an existing set of 20,000 unigenes for barley derived from 20 different complementary DNA (cDNA) libraries (Zhang et al. 2004). Of the 216 SNP and 4 *InDel* markers placed onto the barley genetic map, 83 markers were converted into CAPS and *InDel* markers.

Characteristics and features of barley SNPs

The majority of the SNPs were identified by a comparative sequencing approach of randomly selected ESTs. However, to increase the efficiency of SNP discovery by pre-selecting polymorphic ESTs, a database mining approach was used in case of 25 markers (marked in ESM Table 1) by using the SNiPpER algorithm (Kota et al. 2003).

By using both of the above approaches, the SNP frequency in barley amounted to 1/130 bp. In different sets

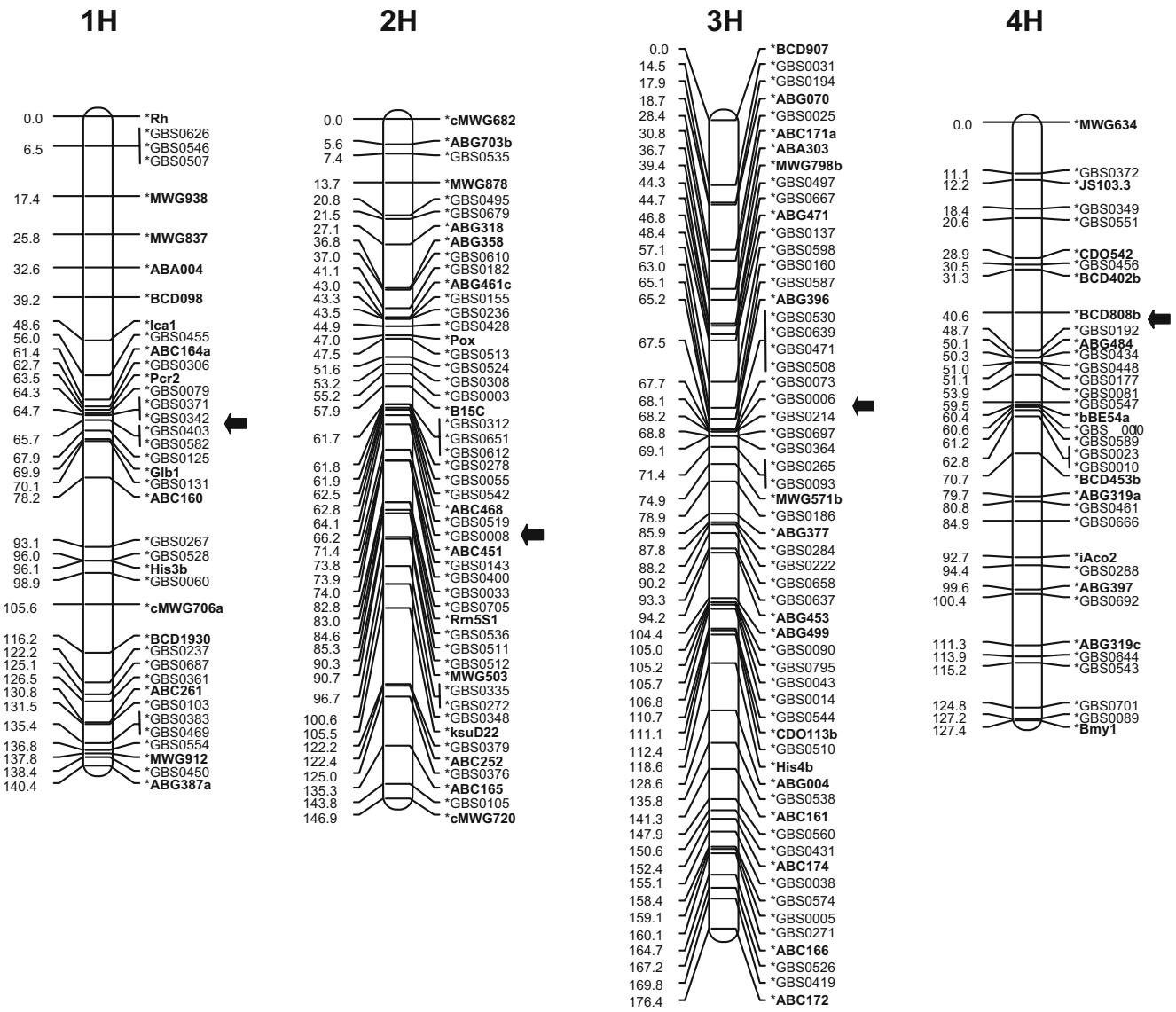

Fig. 2 A consensus SNP map of barley. A total of 220 EST-derived SNP and InDel markers that were mapped in IF, SM, or OWB mapping population were used together with the BIN markers (indicated in *bold*). Centromeres, determined by Kleinhofs and Graner (2001), are indicated by *arrowheads*. Maps are represented with the *short arm on top*

of barley germplasm and across various loci, estimates on SNP frequency varied from 1/27 bp (Bundock and Henry 2004), 1/78 bp (Russell et al. 2004), 1/131 bp (Bundock et al. 2003), 1/189 bp (Kanazin et al. 2002), and 1/200 bp (Rostoks et al. 2005). As expected, the selection of the germplasm affects the observed SNP frequency, as higher frequencies were observed in studies involving a large number of landrace and wild barley accessions (Bundock and Henry 2004; Russell et al. 2004) as compared to those dealing with a smaller selection of cultivated germplasm (Bundock et al. 2003; Kanazin et al. 2002). In this context, up to twofold differences in SNP frequency were observed between the OWB (1/291 bp) and the I × F (to 1/600 bp) population. A similar variation in SNP frequencies was reported in two different sets of germplasm of maize (Tenaillon et al. 2001; Ching et al. 2002). Furthermore, if we consider only the mapped EST loci, a higher SNP frequency (1/107.8 bp) and a higher mean nucleotide diversity ($\pi=4.37\times10^{-3}$) was observed in comparison to the total set of analyzed ESTs (SNP frequency=1/130 bp, $\pi=3.1\times10^{-3}$). Even within the mapped EST loci, about a twofold difference in SNP frequency was observed in pre-selected polymorphic ESTs by using the database-mining approach (1/60.4 bp) compared to randomly selected ESTs (1/130 bp). This increment was statistically significant ($\chi^2=94.30$, $p<0.001$, $df=1$). Similarly pre-selected ESTs had a higher mean π value (12×10^{-3}) than randomly selected ESTs (4.19×10^{-3}; $p<0.001$, two-tailed or U test). On the one hand, these data provide evidence that *in silico* pre-selection of potentially polymorphic ESTs enhances

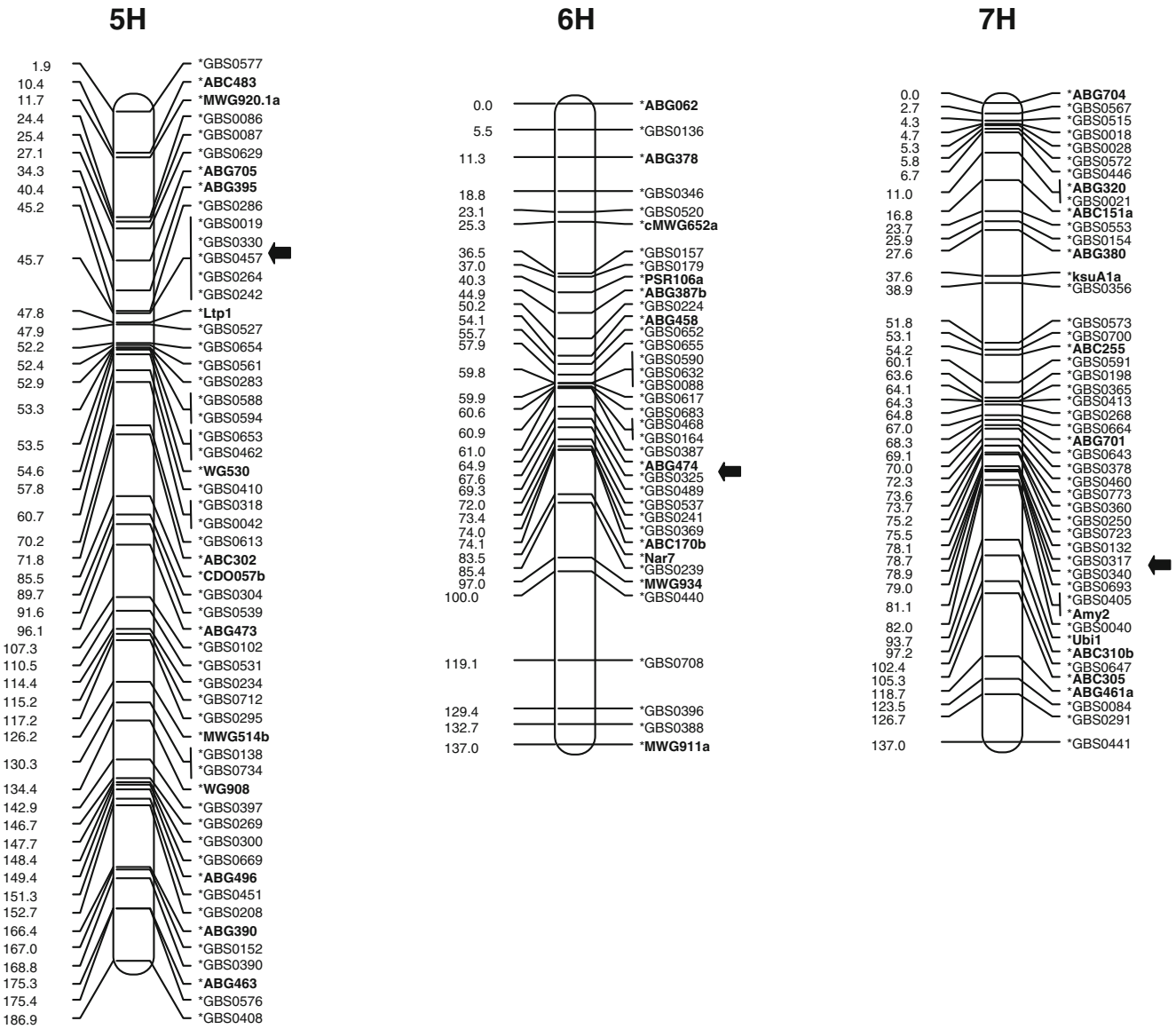

Fig. 2 (continued)

SNP identification efficiency and leads to generating markers with high nucleotide diversity (Kota et al. 2003). On the other hand, SNP frequency and nucleotide diversity estimates also depend on both the selection of germplasm and the nature of EST/gene loci used for SNP discovery and data analysis.

While comparing the SNP frequency in barley with the estimated SNP frequencies in other crop species where comparable datasets are available, it can be seen that SNP frequency in cultivated barley (1/130 bp) is equivalent or higher than that of soybean (1/278 bp, Van et al. 2004), sugarbeet (1/130 bp, Schneider et al. 2001), wheat (1/540 bp, Somers et al. 2003), equal to sorghum (1/123 bp, Hamblin et al. 2004) but lower than in maize (1/104 bp, Tenaillon et al. 2001; 1/60.8 bp, Ching et al. 2002). In line with this, the mean nucleotide diversity in barley (3.10×10^{-3}) was higher compared to soybean (0.97×10^{-3}, Zhu et al. 2003; 0.70×10^{-3}, Van et al. 2004) similar to sorghum (2.25×10^{-3}, Hamblin et al. 2004) and lower than in maize (9.6×10^{-3}, Tenaillon et al. 2001, 6.3×10^{-3}, Ching et al. 2002). The fact that despite lower SNP frequencies, higher nucleotide diversities were observed in wheat (6.9×10^{-3}, Somers et al. 2003), and sugarbeet (7.6×10^{-3}, Schneider et al. 2001) is the result of a more even distribution of the SNP alleles in the corresponding populations.

Development of functional SNP markers

Although a variety of molecular markers, mainly RFLP and SSR markers, are already available for barley (Varshney et al. 2004), SNP markers provide additional options because of their abundance and amenability to high throughput

Table 2 Summary of SNP marker development and characterization

	Linkage group							Total/ overall
	1H	2H	3H	4H	5H	6H	7H	
Mapping population								
IF	1	–	5	–	1	2	3	13
SM	7	16	12	8	10	10	16	78
OWB	15	17	23	14	30	14	16	129
Total	23	33	40	22	41	26	35	220
CAPS assay optimized[a]	10	10 (*1*)	17 (*1*)	10	14 (*2*)	12 (*1*)	14	87 (*5*)
Marker features								
Nucleotide diversity index	0.0008–0.0164 (0.0045)	0.0007–0.0282 (0.0047)	0.0005–0.0395 (0.0055)	0.0006–0.0113 (0.0044)	0.0006–0.0390 (0.0054)	0.0005–0.0076 (0.0029)	0.0007–0.0226 (0.0055)	0.0005–0.0395 (0.0047)
Number of SNPs per amplicon	1–18 (4.3)	1–41 (4.6)	1–22 (4.8)	1–13 (3.84)	1–23 (4.6)	1–10 (3.0)	1–28 (4.9)	1–41 (4.3)
PIC value of mapped SNPs	0.24–0.49 (0.34)	0.24–0.50 (0.34)	0.24–0.49 (0.35)	0.24–0.50 (0.35)	0.24–0.50 (0.36)	0.24–0.49 (0.31)	0.24–0.49 (0.32)	0.24–0.50 (0.34)
Number of haplotypes per marker	2–7 (3.3)	2–6 (2.7)	2–6 (3.0)	2–6 (2.9)	2–6 (2.9)	2–5 (2.9)	2–6 (3.1)	2–7 (3.0)
PIC value of haplotypes	0.24–0.85 (0.53)	0.24–0.82 (0.45)	0.24–0.83 (0.51)	0.24–0.82 (0.51)	0.24–0.83 (0.50)	0.24–0.78 (0.48)	0.24–0.82 (0.49)	0.24–0.85 (0.50)

[a] Markers represented in *italic* font can be assayed as InDels.

approaches. In recent years, SNP markers were employed for estimating the SNP frequency or genotyping germplasm collections in barley (Kanazin et al. 2002; Bundock et al. 2003; Bundock and Henry 2004; Russell et al. 2004; Chiapparino et al. 2004) and also a SNP map based on abiotic stress responsive genes was constructed (Rostoks et al. 2005).

In the present study, the linkage groups 5H and 3H contain the highest number of mapped loci, suggesting the presence of more genes on these two chromosomes. This is in accordance with previous studies in wheat and barley where the highest number of EST-derived markers were mapped on chromosome 3H (Varshney et al. 2006) and the homoeologous linkage group 3 (http://wheat.pw.usda.gov/cgi-bin/westsql/map_locus.cgi; Qi et al. 2004), respectively.

Evaluation of the developed SNP markers on the basis of allelic frequencies of mapped SNPs in the analyzed genotypes showed an average PIC value of 0.34. In comparison, EST-derived SSR markers showed an average PIC value of 0.45 (Thiel et al. 2003; Varshney et al. 2006). Nevertheless, it should be noted that the PIC values calculated on the basis of haplotypes rather than individual SNPs were about 1.5 times as high (results not shown). Thus, the information content of SNP haplotypes observed in the present study is comparable to the information content of EST-derived SSR markers. In this regard, the utilization of the core set of highly informative markers (average haplotype PIC=0.74) should prove useful for diversity studies and other applications. However, the complete exploitation of the haplotype information would require the development of assays that are able to interrogate all SNPs contributing to a haplotype. Using the technology presently available would increase the cost of genotyping, relative to the analysis of single SNPs, as each haplotype was defined by 2–12 SNPs. However, in the light of the ongoing advancement of DNA sequencing technologies, re-sequencing is expected to get increasingly cost efficient to recover haplotype information in the future even from large number of accessions.

Furthermore, the present set of EST-based SNP markers represents a useful resource to be deployed in related cereal species. In this regard, 48 ESTs from the present set were utilized for SNPs discovery, genetic mapping and diversity assessment in rye (Varshney et al. 2007).

Practical utility of SNP markers in barley genetics and breeding

Originally, before large parts of the transcriptome of crop species became accessible in the form of ESTs, molecular markers, e.g., RFLP, RAPD, SSR, or AFLP, were developed from anonymous genomic DNA (summarized in Varshney et al. 2004). Results from such molecular markers obtained

Table 3 Haplotype diversity for marker GBS0546

Genotype	Group	Position in basepairs[a]																	
		119	173	188	209	214	215	231	272	305	343	363	376	380	382	390	401	424	442
Igri	1	C	C	C	G	G	C	A	C	C	G	T	G	G	A	C	C	C	G
Steptoe	2	C	C	C	G	G	C	A	C	C	G	T	C	A	A	C	C	T	C
OWB_{Dom}	3	C	C	C	A	G	C	G	C	C	T	A	G	G	G	T	C	C	G
Barke	4	C	C	C	A	G	C	G	C	G	G	A	G	G	G	T	C	C	G
OWB_{Rec}	5	T	C	C	G	T	T	A	C	C	G	T	G	G	A	C	C	C	C
Morex	6	C	T	C	G	G	C	A	G	C	G	A	G	G	G	C	C	C	G
Franka	7	C	T	T	G	G	C	G	C	C	G	A	G	G	G	T	T	C	G
PIC of SNPs		0.25	0.40	0.25	0.40	0.25	0.25	0.49	0.25	0.25	0.25	0.49	0.25	0.25	0.49	0.49	0.25	0.25	0.40

[a] Position number refers to the sequence of barley EST BU988993 corresponding to the marker GBS0546.

independently for diversity analyses were not essentially similar (Russell et al. 1997). In contrast, the different classes of functional gene-based molecular markers yield similar or comparable results and, for example, reveal similar groupings of the genotypes in germplasm screenings (Kota et al. 2001a; Graner et al. 2004; Russell et al. 2004). Because results obtained from SNP markers can be described in an alphanumeric manner according to the four nucleotides (generally in binary fashion), their documentation is simpler and more straightforward than for any other type of markers. With the present state of knowledge, SNP markers seem the best for meeting the requirements for marker-assisted management of genetic resources in gene-banks, as well as for diversity studies and marker-assisted selection in breeding programs. Furthermore, SNP markers, depending on the assay, can also be used for the quantitative assessment of allele frequencies in populations.

One of the limitations regarding many applications of SNP markers in plant genetics and breeding is that most of the presently available SNP assays rely on expensive, specialized equipment and chemicals. The conversion of SNPs to CAPS markers provides, as shown in this study, an opportunity for widespread applications also in laboratories equipped with simple infrastructure facilities.

Although the employed SNP2CAPS tool (Thiel et al. 2004) suggested a putative restriction recognition site for 91.9% (203) of the markers developed in the present study, only 128 markers were finally selected for verification in *wet lab* experiments because, for the remaining possible cases, only rarely available and/or relatively expensive restriction enzymes were predicted to be used. Out of these 128 markers, only for 82 markers (~64.1%), unequivocal and mappable restriction patterns were observed. This decreased success rate can be attributed to mainly three different issues: (1) Depending on the presence of sequence ambiguities (recorded as "N" in sequence alignments), the SNP2CAPS tool may erroneously predict a non-existing restriction site (Thiel et al. 2004), which reached in the present study about 10% of false positive CAPS candidates. Taking this into account, out of 128 marker–enzyme pairs,

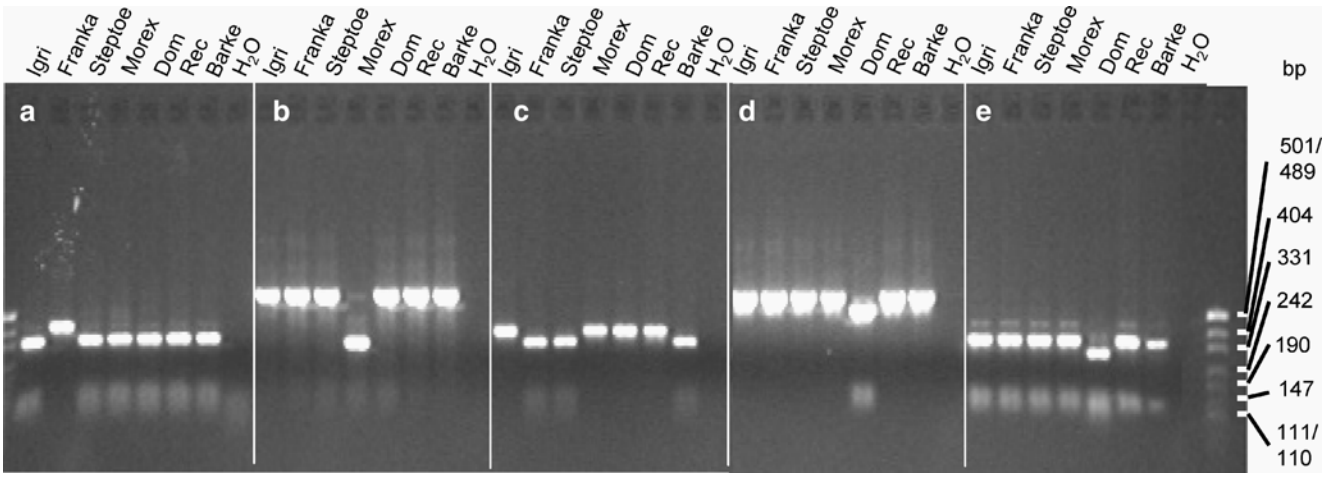

Fig. 3 Conversion of SNPs into CAPS markers. Gel electrophoretic separation on 1.2% agarose gels of five CAPS markers: **a** GBS0546–*Hha*I, **b** GBS0554–*Hha*I, **c** GBS0589–*Hha*I, **d** GBS0667–*Dde*I, and **e** GBS0295–*Cac8*I. The sizes (in basepairs, bp) of PUC19/*Msp*I restriction fragments are indicated on the right

we expected to get good prediction for about 115 markers. (2) Furthermore, ten markers (7.8%) displayed restriction patterns too complex for unequivocal differentiation of alleles. (3) In the remaining cases (17.8%), either the critical restriction site was located too closely to the borders of the PCR fragment, or it was too close to a second restriction site not allowing for satisfactory resolution of the polymorphic DNA fragments on agarose gels. Nevertheless, the successful design of 87 marker assays (CAPS and *InDel*) points at the feasibility to exploit highly informative SNP markers (average PIC=0.34) at relatively low cost for low to medium throughput analysis, reaching from diversity studies to genetic mapping and marker-assisted selection in breeding programs. Hence, the relative abundance of SNPs in the barley genome and the availability of a comprehensive collection of ESTs generally offer the possibility for constructing a saturated SNP map that will significantly improve the marker based access to the barley genome.

Integration of genetic and physical maps

In addition to being used for diversity studies, trait mapping and marker-assisted selection, EST-derived SNPs markers will represent a crucial resource for the alignment of BAC contigs and genetic maps. Given the uneven distribution of genes in the barley genome, PCR-based screening of BAC libraries using the available EST-derived marker resources provides a possibility to sample the gene space (Varshney et al. 2006). If these markers were previously mapped, the corresponding BACs are automatically connected to the genetic map, thus, establishing the link between sequence and trait information. Evidently, many more EST-based SNP markers will be required for systematic sampling of the gene space. Therefore, efforts are underway to significantly enlarge resource of mapped SNPs (Rostoks et al. 2006).

Acknowledgment We are grateful to Patrick Hayes for providing the DH lines of barley mapping populations "Steptoe" × "Morex" and "Oregon Wolfe$_{Dom}$" × "Oregon Wolfe$_{Rec}$" and Ulrike Beier for the excellent technical assistance. The assistance of Thomas Thiel in using the SNP2CAPS software is greatly appreciated. We thank Uwe Scholz for establishing the website featuring the supplemental data. This work was funded by the German Federal Ministry of Education and Research in conjunction with the GABI program (BMBF grants 0312270/4, 0312271A, and 0312278C).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Antonarakis SE (1998) Recommendations for a nomenclature system for human gene mutations. Human Mutat 11:1–3

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Beutler E, McKusick VA, Motulsky AG, Scriver CR, Hutchinson F (1996) Mutation nomenclature: nicknames, systematic names, and unique identifiers. Human Mutat 8:203–206

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

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

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) SNPs frequency, halpotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3:19

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Sapolsky RJ, Anderson M, Drenkard E, Dewdney L, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nature Genet 23:203–207

Costa JM, Corey A, Hayes PM, Jobet C, Kleihofs A et al (2001) Molecular mapping of the Oregon Wolfe Barleys: a phenotypically polymorphic doubled-haploid population. Theor Appl Genet 103:415–424

den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Theor Appl Genet 15:7–12

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNPs resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. Genome Res 14:1812–1819

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

Graner A, Dehmer KJ, Thiel T, Börner A (2004) Plant genetic resources: benefits and implications of using molecular markers. In: Carmen de Vincente M (ed) Issues in genetic resources no. 11. IPGRI, Rome, pp 26–32

Gribskov M, Devereux J, Burgess RR (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res 12:539–549

Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland, USA

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167:471–483

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

Kleinhofs A, Graner A (2001) An integrated map of the barley genome. In: Phillips RL, Vasil IK (ed) DNA markers in plants. Kluwer, Dordrecht, The Netherlands, pp 187–199

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 K-J, Graner A (2001a) Generation and comparison of EST-derived SSR and SNPs markers in barley (*Hordeum vulgare* L.). Hereditas 135: 141–151

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 Genomics 270:24–33

Möhring S, Salamini F, Schneider K (2004) Multiplexed, linkage group-specific marker sets for rapid genetic mapping and fingerprinting of sugar beet (*Beta vulgaris* L.). Mol Breed 14: 475–488

Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N, Monna L, Minobe Y (2002) Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNPs markers. DNA Res 9:163–171

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, NewYork USA

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

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168:701–712

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

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome-wide SNPs discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274:515–527

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Svensson JT, Prasanna B, Stein N, Varshney RK, Marshall D, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci 103:18656–18661

Russell J, Fuller J, 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 9:714–722

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. Genome 47:389–398

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933

Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 13:1250–1257

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

Somers DJ, Kirkpatrick R, Moniwa M, Walsh A (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 46:431–437

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

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

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 U S A 98:9161–9166

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development of cDNA derived microsatellite 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 SNPs 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

Torjek O, Berger D, Meyer RC, Mussig C, Schmid KJ, Rosleff Sorensen T, Weisshaar B, Mitchell-Olds T, Altmann T (2003) Establishment of a high-efficiency SNPs-based framework marker set for *Arabidopsis*. Plant J 36:122–140

Van K, Hwang E-Y, Young Kim M, Kim Y-H, Cho Y-I, Cregan PB, Lee S-H (2004) Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica 139:147–157

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

Varshney RK, Grosse I, Hahnel U, Siefken R, Prasad M, Stein N, Langridge P, Altschmied L, Graner A (2006) Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. Theor Appl Genet 113:239–250

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

Zhang H, Sreenivasulu N, Weschke W, Stein N, Rudd S, Radchuk V, Potokina E, Scholz U, Schweizer P, Zierold U, Langridge P, Varshney RK, Wobus U, Graner A (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. Plant J 40:276–290

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