Published online January 2, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 1 ***e5***
DOI: 10.1093/nar/gnh006

SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development

Thomas Thiel, Raja Kota, Ivo Grosse, Nils Stein and Andreas Graner*

Institute for Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany

Received August 12, 2003; Revised and Accepted November 14, 2003

ABSTRACT

With the influx of various SNP genotyping assays in recent years, there has been a need for an assay that is robust, yet cost effective, and could be performed using standard gel-based procedures. In this context, CAPS markers have been shown to meet these criteria. However, converting SNPs to CAPS markers can be a difficult process if done manually. In order to address this problem, we describe a computer program, SNP2CAPS, that facilitates the computational conversion of SNP markers into CAPS markers. 413 multiple aligned sequences derived from barley ESTs were analysed for the presence of polymorphisms in 235 distinct restriction sites. 282 (90%) of 314 alignments that contain sequence variation due to SNPs and InDels revealed at least one polymorphic restriction site. After reducing the number of restriction enzymes from 235 to 10, 31% of the polymorphic sites could still be detected. In order to demonstrate the usefulness of this tool for marker development, we experimentally validated some of the results predicted by SNP2CAPS.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA variation in the genome (1). Because of their abundance, the exploitation of SNPs for marker assays has the potential to provide answers to a large number of important biological, genetic, pharmacological and medical questions. The identification of SNPs has progressed remarkably over the last several years and multiple assays have been devised (2–5). However, most of these assays require expensive and specialized equipment and chemicals for analysis.

Hence, there is a need for simple and accurate genotyping assays that can be implemented in laboratories that do not have access to sophisticated equipment. A solution to this problem is the detection of a SNP site by an appropriate restriction endonuclease whose recognition sequence has been altered or introduced by the SNP. In combination with a PCR assay, the corresponding SNP can be analysed as a cleaved amplified polymorphic sequence (CAPS) marker (6). The costs of a CAPS assay is generally low, especially when it relies on commonly used restriction enzymes.

In order to facilitate a computational conversion of SNPs into CAPS markers, a program called dCAPS Finder 2.0 was previously developed (7). The dCAPS Finder program works on the principle of designing mismatched PCR primers that would create or remove a restriction recognition site in the analysed SNP. The conversion of SNP sites into CAPS markers by the artificial introduction of restriction sites involves the creation of mismatched primers, whose successful application is not always trivial depending on the number, positions and types of mismatches.

We present a computer program named SNP2CAPS that works in a different manner. A simple algorithm involves the screening of multiply aligned sequences for restriction sites followed by a selection pipeline that allows the deduction of CAPS candidates by the identification of putative alternative restriction patterns. It should be noted that in this algorithm any primer pair flanking the SNP site may be suited for CAPS marker analysis.

In order to evaluate the efficacy of SNP2CAPS, a set of 3045 sequences derived from eight barley accessions based on 413 expressed sequence tags (ESTs) was used and analysed in terms of (i) the potential number of SNP markers that can be converted into CAPS markers taking into account all commercially available restriction enzymes and (ii) the number of CAPS markers that can be typed if only the 10 most commonly used enzymes are considered. To investigate the accuracy of this tool, 14 EST-based SNP markers have been experimentally validated.

MATERIALS AND METHODS

Description of SNP2CAPS

Two input files that contain data about the sequence alignments and the restriction enzymes are required for SNP2CAPS. The first input file is a modified FASTA formatted file that stores one or more multiple alignments of sequences of different accessions. In order to ensure compatibility with existing alignment tools, 15 additional multiple alignment formats (e.g. ClustalW, MSF and MEME) can be imported using the AlignIO handler of the BioPerl module v1.2 (http://bioperl.org/). The second input file contains data on the restriction enzymes that can be downloaded in different

*To whom correspondence should be addressed. Tel: +49 39482 5521; Fax: +49 39482 5155; Email: graner@ipk-gatersleben.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

formats from the restriction enzyme database REBASE (http://rebase.neb.com/).

For the analysis process, each of the aligned sequences is screened for the corresponding restriction enzyme recognition sites using the standard regular expression syntax. In contrast to common tools for restriction analysis, an assessment is made as to whether a single restriction enzyme is suitable for CAPS analysis. Only those enzymes that display a restriction site polymorphism among all of the aligned sequences are of interest. In this respect, analogous recognition sites are identified by identical positions within the alignment. This is important, because the selection of restriction enzymes as CAPS candidates should be avoided in cases where the respective restriction patterns only reflect the sequence size variation of different accessions due to insertions or deletions (InDels) outside or between putative restriction sites.

The analysis process continues with an assessment of each marker–enzyme pair. In principle, CAPS candidates fall into two categories: alternative restriction patterns may occur because of either SNPs or InDels. In the case of SNPs, the algorithm decides to assign the recognition site to one of the four classes as exemplified in Figure 1. The first class represents one or more altered non-ambiguous bases (A, C, G or T) for at least one of the sequences. For example, the sequence illustrated in Figure 1 contains a SNP that does not match the enzyme recognition sequence, indicating a restriction site polymorphism. The second class contains sequences that display an additional ambiguous base (N) within one of the predicted CAPS sites and therefore indicates a lack of sequence information. Class three contains sequences designated 'false positives': because of the presence of an ambiguous base (N), the program indicates that more information is needed for analysis of the potential SNP site. The fourth class comprises sequences that either contain the recognition sequence at all given loci (Fig. 1d) or do not contain a recognition sequence at all. Figure 1e and f, illustrates the presence of insertions or deletions within recognition sequences of CAPS candidates.

The outcome of the analysis can be viewed individually in accordance with the four classes described above. Results can be displayed for marker–enzyme pairs as groups of sequences that (i) show the same restriction pattern or (ii) share the same sequence pattern at the recognition sites.

SNP2CAPS is available as a Perl5 script, which can be either executed from the command line or as a graphical user interface (GUI) application using the Perl/Tk tool kit (Fig. 2). SNP2CAPS is freely available and can be downloaded from http://pgrc.ipk-gatersleben.de/snp2caps/.

Data sets

Sequencing efforts resulted in a set of 413 partially sequenced genes spanning a total of 153 kb. On average, each locus had a length of 370 bp. 3045 sequences were obtained by sequencing three to eight barley accessions per gene locus, resulting in a total of 1.13 Mb. Sequences were aligned using the ClustalW program (8).

For computational restriction analysis the GCG format data file from the REBASE database (version 304, March 24, 2003) was used, comprising information on 645 type II restriction enzymes. In the present study, diagnostic restriction was investigated using (i) a total of 235 enzymes that are non-isoschizomeric and commercially available and (ii) a subset of the following 10 enzymes that are widely used in daily bench work: the 4 bp cutters AluI, HpaII, MseI and RsaI and the 6 bp cutters BamHI, DraI, EcoRI, EcoRV, HindIII and XbaI.

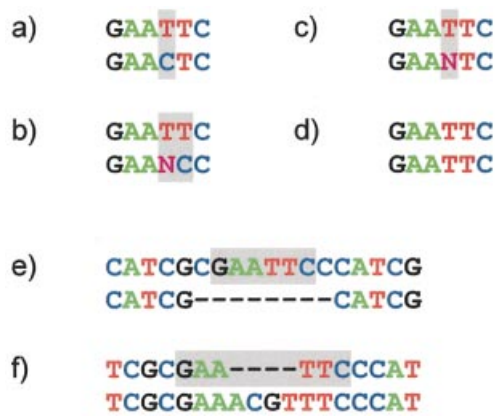

Figure 1. Illustration of possible scenarios at an EcoRI recognition site (G↑AATTC) between two aligned sequences. (a–d) Four different types that are recognized by the program algorithm: (**a**) class I, CAPS candidates; (**b**) class II, CAPS candidates containing N; (**c**) class III, false positive candidates; (**d**) class IV, no restriction site polymorphisms (due to no or uniform restriction patterns). (e and f) The role of insertions/deletions for the analysis of CAPS marker candidates: (**e**) deletion of the restriction site; (**f**) insertion at a restriction site.

Plant material

For experimental verification of the SNP2CAPS results a genotype set of seven barley (*Hordeum vulgare* ssp. *vulgare*) accessions was used, namely the winter barley cultivars 'Igri' and 'Franka', the spring barley cultivars 'Steptoe', 'Morex' and 'Barke', the genetic stocks 'Oregon Wolfe Rec' and 'Oregon Wolfe Dom' and accession HOR11508 of wild barley (*H.vulgare* ssp. *spontaneum*). For these eight accessions genomic DNA was extracted from leaf material as described in Graner *et al*. (9).

Design of primers, PCR and sequence analysis

Primer pairs were designed on the basis of ESTs generated from the barley accession 'Barke' using the computer program Primer Express (Applied Biosystems, Foster City, CA). PCR was performed with genomic DNA of the above mentioned eight barley accessions resulting in PCR fragments of 350–450 bp in length. PCR was done in 20 µl reactions as described in Kota *et al*. (10). To identify SNPs, PCR products amplified from genomic DNA templates were sequenced in both directions on an ABI 377XL automated sequencer using Big dye terminator chemistry (Applied Biosystems). DNA sequence data was checked manually for sequencing errors using the Sequencher computer program (Gene Codes Corp., Ann Arbor, MI).

Restriction endonuclease digestion and agarose gel electrophoresis

Twenty microlitres of the PCR product were digested in 1× buffer with 1 U of one of eight restriction enzymes (BamHI,

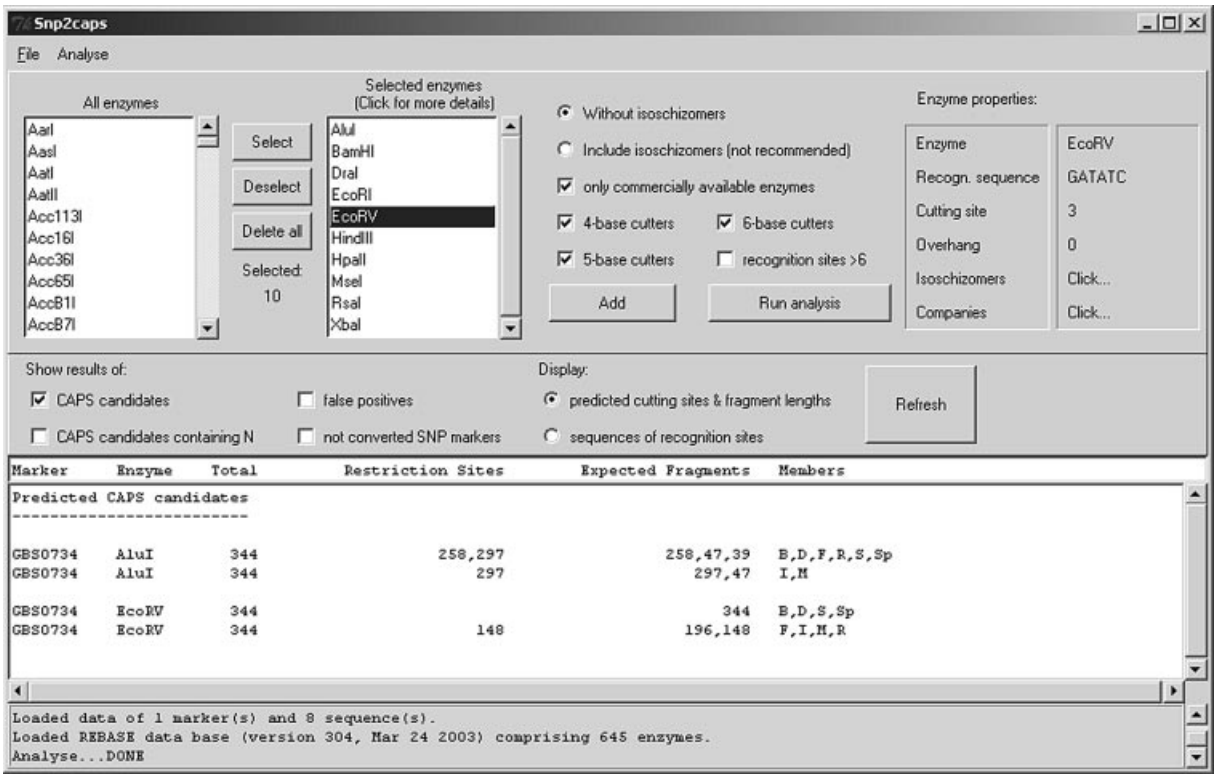

Figure 2. Screenshot of SNP2CAPS. The screenshot shows the output of the results after the screening of SNP marker GBS0734 with 10 selected restriction enzymes.

Table 1. Conversion of SNPs to putative CAPS markers

		235 enzymes	10 enzymes
(I)	CAPS candidates	281	94
(II)	CAPS candidates (+N)	2	3
(III)	False positives	6	11
(IV)	No restriction site polymorphisms	25	206

The numbers are based on 314 polymorphic sequence alignments screening for all commercially available, non-isoschizomeric restriction enzymes represented in the REBASE database as well as for a subset of 10 commonly used enzymes.

DraI, EcoRV, HaeII, HaeIII, HindIII, PstI and XbaI) (New England Biolabs, Beverly, MA) for 3 h in a 37°C water bath, followed by electrophoretic separation (1.5% agarose, 1× TBE buffer pH 8.0, 1 h, 110 V).

RESULTS AND DISCUSSION

A total of 413 multiple sequence alignments were subjected to the identification of CAPS candidates using SNP2CAPS. Out of these, 314 alignments contained distinct types of sequence variation within the genotype set that can be summarized as follows: in 204 cases, sequence variation was caused by SNPs, in 19 alignments it was caused by InDels and in the remaining 91 alignments SNPs were observed together with InDels.

A total of 282 (90%) out of 314 alignments displayed at least one potential CAPS candidate with a single restriction enzyme (class I and II of Fig. 1) when all 235 commercially available, non-isoschizomeric restriction enzymes were applied to the data set. As shown in Table 1, the remaining 10% of the sequence polymorphic alignments contributed to false positives (class III) or did not display restriction site polymorphisms among the barley accessions (class IV). As expected, the number of potential CAPS candidates decreased when only 10 enzymes are taken into account. However, a total of 97 (31%) alignments still contained SNPs or InDels that could be converted into potential CAPS markers. Within these alignments polymorphic restriction patterns of one or more enzymes were caused by the presence of SNPs in 103 cases, whereas in 25 cases (20 deletions and 5 insertions) the polymorphism was caused by the presence of InDel sites within the recognition sequences. The running time of SNP2CAPS scales linearly with the total sequence length and the database: it took ~25 s to analyse the set of 413 multiple aligned sequences with 235 restriction enzymes on a 2.4 GHz Pentium 4 PC.

For experimental verification we randomly selected 14 EST-derived SNP markers after computational analysis with a set of common restriction enzymes (Table 2). Upon digestion of the corresponding PCR fragments, all 14 EST–restriction

Table 2. Experimental validation of 14 randomly chosen EST–restriction enzyme pairs

SNP marker	Restriction enzyme	PCR product size	Restriction site	Confirmed fragment sizes	Accessions
GBS0003	HaeII	335	131	204, 131	B,F,I,R,S
		335	–	335	D
GBS0028	BamHI	244	–	244	B,D,F,I,R,S
		244	95	149, 95	M
GBS0131	BamHI	286	–	286	B,D,F,I,M,Sp,S
		286	241	241, 45	R
GBS0153	PstI	580	–	580	B,D,F,I,M,R
		580	507	507, 73	S
GBS0344	HindIII	452	177	275, 177	B,F,M,R,Sp,S
		452	–	452	D
GBS0379	DraI	289	–	289	B,D,M
		287	125	162, 125	F,I,R,S
GBS0400	XbaI	790	–	790	B,D,R,Sp,S
		953	607	607, 346	M
GBS0419	HaeIII	343	–	343	B,D,F
		343	143	200, 143	I,M,R,Sp,S
GBS0468	HaeII	330	–	330	B
		330	134	196, 134	D,F,I,M,R,Sp,S
GBS0535	HindIII	624	61	563, 61	B,F,I,S
		638	–	638	D,R
		627	62	565, 62	M
GBS0560	HaeIII	351	88, 194	157, 106, 88	B,F,I,M
		345	188	188, 157	D,R,Sp
GBS0560	PstI	351	304	304, 47	B,F,I,M
		345	140, 298	158, 140, 47	D,R,Sp
GBS0734	EcoRV	319	–	319	B,D,Sp,S
		319	123	196, 123	F,I,M,R
GBS0347	HindIII	690	–	690	B,F,I,M,R,Sp,S
		691	54	637, 54	D

In all cases the predictions of SNP2CAPS could be confirmed. 'Igri' (I), 'Franka' (F), 'Steptoe' (S), 'Morex' (M), 'Dom' (D), 'Rec' (R), 'Barke' (B) and *H.vulgare* ssp. *spontaneum* (Sp) are different barley accessions used in this study.

enzyme pairs that were tested revealed the predicted restriction pattern. An example of a sequence alignment and the corresponding restriction pattern for a selected marker (GBS0734) is shown in Figure 3.

It should be mentioned that the current version of SNP2CAPS does not evaluate the technical usefulness of the suggested restriction patterns. This means that in a few cases the diagnostic restriction sites are (i) too close to the borders of the PCR fragments or (ii) too close to a second restriction site to allow resolution of the size of polymorphic DNA fragments in agarose gels. However, the experimenter may easily resolve the usefulness of the prediction by visual inspection of the computed results.

Recently, several tools have been developed for the computational analysis of DNA sequences in order to develop markers in different organisms, e.g. MISA or SPUTNIK for the computational identification of sequence-derived microsatellite markers of barley or pepper (11,12). Regarding the development of SNP-based markers, the SNP2CAPS program described here represents a generic tool for the computational identification of polymorphic restriction sites that can be readily analysed as CAPS. SNP2CAPS is of use in cases where laboratories are not equipped with high throughput instrumentation for performing SNP analysis.

Another aspect is the relatively inexpensive cost of CAPS assays, making this procedure reasonable for low to medium throughput analysis, reaching from diversity studies to genetic mapping and marker-assisted selection in breeding programmes. Since the nature of the SNP that underlies the corresponding CAPS is known, the resulting information is fully compatible with any other assay of the same SNP (e.g. MALDI-TOF, heteroduplex analysis, etc.).

In this case study, 90% of EST-derived SNP markers could be converted into CAPS markers, although some of the restriction enzymes predicted for use are rare and can be quite expensive. Employing only the most commonly used 10 enzymes, 31% of the polymorphic alignments still contain potential CAPS. Given the steadily increasing number of EST-derived SNPs that are being identified for barley and for many other organisms, a combination of the SNP2CAPS program with other software tools for computer-assisted identification of SNPs from EST databases (13) will lead to the generation of a comprehensive resource of CAPS markers.

ACKNOWLEDGEMENTS

We thank Ulrike Beier and Jacqueline Pohl for their technical assistance and Uwe Scholz for valuable discussions. This work was funded by the German Federal Ministry of Education and Research (BMBF grants 0312706A, 0312270/4, 0312271A and 0312278C).

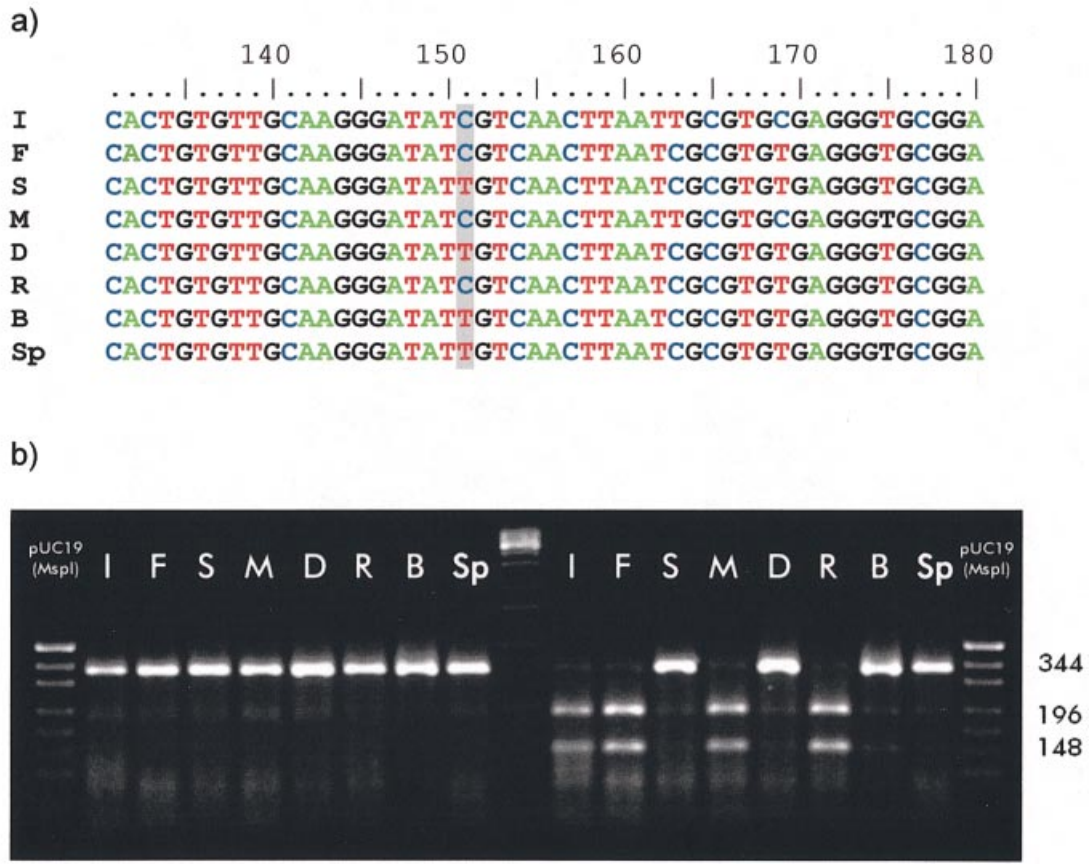

Figure 3. Conversion of the barley SNP marker GBS0734 into an EcoRV CAPS marker. (**a**) Relevant part of the multiple sequence alignment of SNP marker GBS0734. The recognition site of EcoRV (GAT↑ATC) is affected by one SNP at position 151 (C→T transition). 'Igri' (I), 'Franka' (F), 'Steptoe' (S), 'Morex' (M), 'Dom' (D), 'Rec' (R), 'Barke' (B) and *H.vulgare* ssp. *spontaneum* (Sp) are different barley accessions used in this study. (**b**) Gel electrophoretic separation of PCR products (left undigested, right after EcoRV digestion). As predicted by SNP2CAPS, the restriction enzyme EcoRV cuts PCR fragments of 'Igri', 'Franka', 'Morex' and 'Rec' into two fragments of the predicted sizes (148 and 196 bp), whereas 'Steptoe', 'Dom', 'Barke' and *H.vulgare* ssp. *spontaneum* display the undigested PCR product (344 bp). Relevant fragment sizes (bp) are denoted on the right side.

REFERENCES

1. Brookes,A.J. (1998) The essence of SNPs. *Gene*, **234**, 177–186.
2. Shi,M.M. (2001) Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clin. Chem.*, **47**, 164–172.
3. Gut,I.G. (2001) Automation in genotyping of single nucleotide polymorphisms. *Hum. Mutat.*, **17**, 475–492.
4. Kwok,P.Y. (2000) High-throughput genotyping assays approaches. *Pharmacogenomics*, **1**, 95–100.
5. Rafalski,J.A. (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.*, **162**, 329–333.
6. Parsons,B.L. and Heflich,R.H. (1997) Genotypic selection methods for the direct analysis of point mutations. *Mutat. Res.*, **387**, 97–121.
7. Neff,M.M., Turk,E. and Kalishman,M. (2002) Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.*, **18**, 613–615.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.
9. Graner,A., Jahoor,A., Schondelmaier,H., Siedler,K., Pillen,K., Wenzel,G. and Herrmann,R.G. (1991) Construction of an RFLP map of barley. *Theor. Appl. Genet.*, **83**, 250–256.
10. Kota,R., Wolf,M., Michalek,W. and Graner,A. (2001) Application of DHPLC for mapping single nucleotide polymorphisms (SNPs) in barley (*Hordeum vulgare* L.). *Genome*, **44**, 523–528.
11. Thiel,T., Michalek,W., Varshney,R.K. and 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.
12. Sanwen,H., Baoxi,Z., Milbourne,D., Cardle,L., Guimei,Y. and Jiazhen,G. (2000) Development of pepper SSR markers from sequence databases. *Euphytica*, **117**, 163–167.
13. Kota,R., Rudd,S., Facius,A., Kolesov,G., Thiel,T., Zhang,H., Stein,N., Mayer,K. and Graner,A. (2004) SNiPping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol. Gen. Genomics*, in press.