Theor Appl Genet (2006) 113:239–250
DOI 10.1007/s00122-006-0289-z

ORIGINAL PAPER

Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome

R. K. Varshney · I. Grosse · U. Hähnel · R. Siefken · M. Prasad · N. Stein · P. Langridge · L. Altschmied · A. Graner

Received: 16 December 2005 / Accepted: 4 April 2006 / Published online: 1 June 2006

Abstract A set of 111,090 barley expressed sequence tags (ESTs) was searched for the presence of microsatellite motifs [simple sequence repeat (SSRs)] and yielded 2,823 non-redundant SSR-containing ESTs (SSR–ESTs). From this, a set of 754 primer pairs was designed of which 525 primer pairs yielded an amplicon and as a result, 185 EST-derived microsatellite loci (EST–SSRs) were placed onto a genetic map of barley. The markers show a uniform distribution along all seven linkage groups ranging from 21 (7H) to 35 (3H) markers. Polymorphism information content values ranged from of 0.24 to 0.78 (average 0.48). To further investigate the physical distribution of the EST–SSRs in the barley genome, a bacterial artificial chromosomes (BAC) library was screened. Out of 129 markers tested, BAC addresses were obtained for 127 EST–SSR markers. Twenty-seven BACs, forming eight contigs, were hit by two or three EST–SSRs each. This unexpectedly high incidence of EST–SSRs physically linked at the sub-megabase level provides additional evidence of an uneven distribution of genes and the segmentation of the barley genome in gene-rich and gene-poor regions.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-006-0289-z and is accessible for authorized users.

Communicated by T.Sasaki

Primer sequences for developed SSR markers are available upon request from the corresponding author (A. Graner).

R. K. Varshney · I. Grosse · U. Hähnel · R. Siefken · M. Prasad · N. Stein · L. Altschmied · A. Graner (✉)
Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany
e-mail: graner@ipk-gatersleben.de

P. Langridge
Australian Centre for Plant Functional Genomics, University of Adelaide, Waite Campus PMB1, Glen Osmond 5064 SA, Australia

Present address: R. K. Varshney
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324 AP, India

Present address: R. Siefken
TECAN Deutschland GmbH, Theodor-Storm-Strasse 17, 74564 Crailsheim, Germany

Present address: M. Prasad
National Centre for Plant Genome Research (NCPGR), New Delhi 110061, India

Introduction

The cereal species assigned to the *Triticeae* tribe comprise important staple crops including wheat, barley and rye. They are characterized by large genomes, ranging in size from 5.6×10^9 bp for barley up to 1.5×10^{10} bp for wheat (Bennett and Leitch 2003). More than 80% of their genomes consist of repetitive DNA, which in turn mainly consists of transposable elements (Schulman et al. 2004). The large content of repetitive DNA forms the major obstacle for sequencing the *Triticeae* genomes resulting in only 6.1 Mb of genomic sequence available in the public domain as of July, 2005 (http://www.ncbi.nlm.nih.gov/). Notwithstanding these limitations the available data may shed some light on the organization of cereal

genomes at the sequence level in general and on the distribution of genes along the chromosomes in particular. In several instances, gene islands have been identified, which are characterized by a relatively high density of genes spaced between 5 and 10 kb (for references see Keller and Feuillet 2000). Gene islands are contrasted by gene-free regions, which may extend over several hundred kilobases, and which are mainly composed of repetitive DNA and frequently show reduced recombination or no recombination at all (Wicker et al. 2001, 2005). Similar findings have been obtained at a larger resolution provided by genetic and cytogenetic maps. Here, about 50% of the single and low-copy markers from a genome wide map of barley could be assigned to only 5% of the physical genome complement indicating the presence of a distinct gene space (Künzel et al. 2000). Similar observations have been reported from physical mapping studies in wheat using deletion lines (for reference see Gill 2004; Erayman et al. 2004).

Due to the abundance of repetitive DNA, sequencing of expressed sequence tags (ESTs) has been the key approach for systematic gene identification in *Triticeae* species. In the case of barley 419,146 ESTs (*Hordeum vulgare* ssp. *vulgare* and *H. vulgare* ssp. *spontaneum*) are deposited at present in dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, dbEST release 071505). These are expected to cover a significant portion of the gene repertoire of barley and provide the groundwork to understand the organization of the barley transcriptome (Zhang et al. 2004). For structural genomics ESTs provide a valuable resource for the development of functional molecular markers to be deployed in comparative mapping studies (Anderson and Lübberstedt 2003; Perovic et al. 2004). Among the most important and popular molecular markers that can be developed from ESTs are simple sequence repeats (SSRs) or microsatellite markers (see Varshney et al. 2005a).

Microsatellite markers have been deployed in a variety of applications in plant genetics and breeding. In the cereals, dense microsatellite maps comprising more than 2,000 loci are available in maize and rice and, similarly, about 2,000 microsatellites have been mapped in wheat (reviewed by Varshney et al. 2004). In barley, about 400 SSR loci have been mapped (Ramsay et al. 2000; Pillen et al. 2000; Li et al. 2003). While most of the available SSR markers were developed from genomic DNA libraries based on experimental approaches, the availability of ESTs facilitated the systematic identification of SSRs and corresponding marker development based on computer-assisted analytical approaches (Varshney et al. 2002; Thiel et al. 2003).

In this study we have searched a set of 111,090 barley ESTs for the presence of SSRs. The corresponding information was used for the development and genetic mapping of a non-redundant set of 185 genic microsatellite markers. The distribution of a randomly chosen set of 129 out of the 185 SSR markers in a large-insert genomic [bacterial artificial chromosomes (BAC)] library provided evidence on the non-random distribution of genes within the barley genome.

Materials and methods

Plant material

Simple sequence repeat polymorphisms were screened in a set of five barley (*H. vulgare* L.) cultivars comprising Barke, Igri, Franka, Steptoe and Morex, and two genetics stocks of Oregon Wolfe barley (OWB), OWB_{Dom} and OWB_{Rec}. Barke was included as a standard because this cultivar was used for the construction of most EST libraries at Institute of Plant Genetics and Crop Plant Research (IPK), whereas the other six genotypes represent the parents of three doubled haploid (DH) mapping populations (Igri × Franka, Steptoe × Morex, OWB_{Dom} × OWB_{Rec}). Genomic DNA isolation was carried out as given in Thiel et al. (2003).

EST-database analysis

A total of 111,090 barley IPK-EST sequences corresponding to approximately 55.9 Mb were screened for microsatellites using the MISA software (Thiel et al. 2003 and available at http://pgrc.ipk-gatersleben.de/misa/). ESTs comprising both 5′ and 3′ sequences were developed from 22 cDNA libraries representing different tissues or developmental stages with the vast majority of sequences being derived from the cultivar Barke (Zhang et al. 2004). To improve the efficiency of the identification of polymorphic SSRs, an additional set of 207,449 barley ESTs (corresponding to 107.6 Mb) developed from cultivars other than Barke (Fig. 1) was included from the EMBL database.

Marker development and map construction

In order to minimise redundancy, a cluster analysis was performed on ESTs containing SSRs (SSR–ESTs) using the StackPACK2.1 software (Miller et al. 1999). Primer pairs for non-redundant microsatellites were designed using the PRIMER3 software as described earlier (Varshney et al. 2002; Thiel et al. 2003).

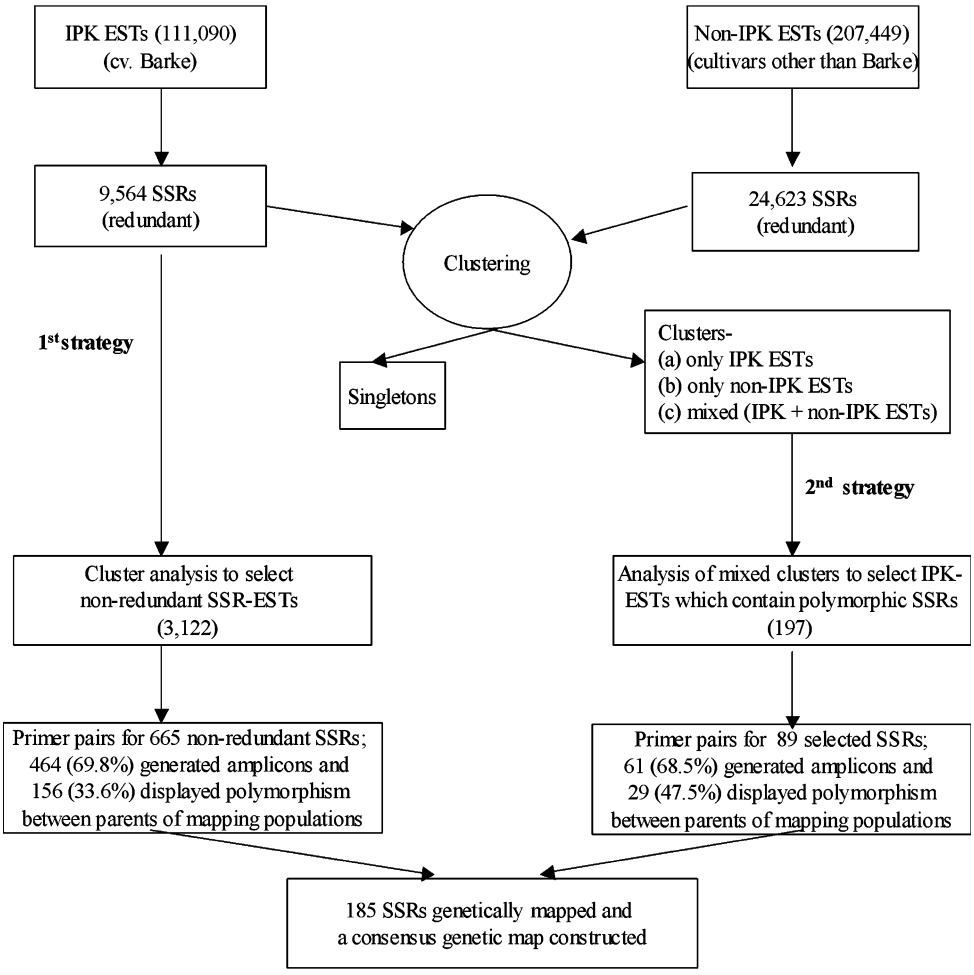

Fig. 1 Systematic EST–SSR marker development in barley. In the *first strategy*, only the IPK dataset comprising 111,090 ESTs was used to identify and develop SSR markers. In the *second strategy*, all available ESTs, IPK ESTs (developed from cv. Barke) and non-IPK ESTs (developed from cultivars other than Barke), were used for identification of SSRs and subsequent clustering. Mixed clusters, containing ESTs derived from more than one genotype were analyzed for the presence of polymorphic SSRs. Polymorphic markers were mapped in one of three mapping populations of which a consensus map was constructed

Polymerase chain reaction amplification of microsatellite loci, gel electrophoresis, visualization and linkage mapping were performed as described earlier (Thiel et al. 2003). Mapped markers are coded as Gatersleben Barley Microsatellite followed by a four-digit numerical code as locus identifier. Generally, one marker was mapped in one of the three mapping population listed above. However, 14 markers were mapped in more than one mapping population. In addition to these common markers, other common RFLP markers (data not shown) and anchor or BIN markers (Kleinhofs and Graner 2001) available on the genetic maps of these three populations were used to prepare a consensus map using the JoinMap v 2.0 software (Stam 1993).

Polymorphism information content

The polymorphism information content (PIC) of individual EST–SSR markers was calculated by using the standard formula (Anderson et al. 1993):

$$\mathrm{PIC} = 1 - \sum_{i=1}^{k} P_i^2.$$

Here, k is the total number of alleles detected for a microsatellite, and P_i is the frequency of the ith allele in the set of the investigated barley accessions.

Functional annotation

Simple sequence repeat–ESTs were compared to the NR-PEP protein database (RefSeq-release 8 November 2004) at the DKFZ Heidelberg by the BLASTX2 software (Altschul et al. 1990) using a threshold value of 1E-10 (for details see http://genome.dkfz-heidelberg.de/).

PCR-based screening of the barley BAC library

A four-step PCR-based screening protocol was established for identifying gene-containing clones in an ordered BAC library of barley with more than 300,000

clones (Yu et al. 2000) utilizing the same primer pairs that were used for genetic mapping of SSR markers. Amplification was achieved in a total volume of 20 µl (buffer: 10 mM Tris–HCl (pH 9.0), 50 mM KCl, 0.1% Triton X-100, 2 mM $MgCl_2$), 0.2 mM each of dATP, dCTP, dGTP and dTTP, 100 pmol of each primer and approximately 1 U of *Taq* polymerase) using a touchdown PCR protocol (95°C, 3 min/10 cycles: 95°C, 1 min; 65–0.5°C/cycle, 1 min; 72°C, 1 min/25 cycles: 95°C, 1 min; 60°C, 1 min; 72°C, 1 min/72°C, 3 min/15°C). The first round of screening was performed on 90 so-called super pools of BAC-DNA each comprising 3,456 clones from 9 consecutive 384-well microtiter plates of the library. Thus a total of 311,040 clones out of 313,344 clones of the Morex BAC library were used for screening. Amplicons with a typical size of 100–500 bp were analysed on 3% agarose gels. For those super pools that yielded a fragment of the same length as genomic DNA from *H. vulgare* cv. Morex, all nine individual plate pools of BAC DNA were examined during the second round of screening. The third round of screening was performed on 16 row and 24 column pools running through a positive plate. These row and column pools were derived from rows and columns running through a rectangular arrangement of 24×34 microtiter plates of the BAC library to minimize the number of DNA preparations. From the glycerol stock of a clone present at the intersections of positive row and column pools within a positive plate, a frozen bit of bacterial culture was obtained and grown for 20 h at 37°C in 200 µl LB medium (1 l: 5 g yeast extract, 10 g NaCl, 10 g tryptone, pH 7.0) containing Chloramphenicol (30 µg/ml). Five microlitre of such an overnight culture was used for PCR verification of the results from the pool screenings. On average, this strategy should require 390 PCRs (without control reactions) for the identification of all BAC clones in the library containing a single copy sequence, assuming a six-fold genome coverage of the library.

Preparation of pools of BAC DNA

A total of 1,930 pools of BAC DNA were prepared for the PCR-based screening strategy: 810 plate pools, each from the 384 BAC cultures of a single microtiter plate, and 576 column and 544 row pools. Super pools were assembled after the preparation of DNA from nine plate pools. Using a liquid handling robot equipped with a plate storage device, all bacterial cultures belonging to a pool were collected into a 96-well microtiter plate by placing 4–6 cultures (10 µl each) in a single well. After growth for 20 h at 37°C in a total volume of 200 µl LB medium per well, the whole contents of this microtiter plate was transferred to 460 ml of liquid medium (two portions of 230 ml in a 1 l Erlenmeyer flask each) and grown for another 16 h at 37°C. DNA was prepared from these cultures using Qiagen Maxi Prep kits (Qiagen, Hilden Germany) as recommended by the manufacturer. On average, 110 µg DNA were obtained. Approximately 25 ng of pool DNA was used as template for PCR for screening.

Testing the null hypothesis about uniform distribution of genes

The BAC library comprising 311,040 BAC clones was screened with a total of 129 markers. Two markers did not hit any BAC clone, and a total of 311 BAC clones were hit 318 times by 127 markers, meaning that some BAC clones were hit by more than one marker. The total number of collisions, C, was defined as follows: for each $i = 1, 2,...,311{,}040$, let N_i denote the number of BAC hits on BAC i, define by $C_i = N_i - 1$ the number of collisions on BAC i, and define the total number of collisions by $C = \sum_i C_i$, where the sum runs over all indices i for which C_i is positive. The null hypothesis that the 318 BAC hits were distributed uniformly among the 311,040 BACs of the Morex library used for screening was tested by using C as test statistic, and the probability of finding C or more than C collisions under the null hypothesis was estimated by the following simulation.

Simulation of the screening process leading to gene-containing BAC clones

A hypothetical genome was divided into 311,040 overlapping segments of identical size to represent the 311,040 clones of the BAC library. Neighbouring segments were made to overlap by 5/6 of their length to simulate the sixfold genome coverage of the barley BAC library.

In each simulation run 129 markers (or genes) were placed at random positions in the hypothetical genome. All six segments overlapping at the position of a gene were labelled. Each of the labelled segments was selected at random with a probability of 2.465/6 to account for the fact that on average 2.465 = 318/129 BACs per gene were identified in the screening process. From these selected segments the number of collisions was calculated.

Each simulation run was repeated 10^8 times to obtain reliable estimates of the probability of the occurrence of 7 or more than 7 collisions down to a P value of 10^{-7}. In order to test the reliability of the P value

estimates and to obtain their 95% confidence intervals, the batch of 10^8 simulation runs was repeated 200 times.

Results

Development of microsatellite markers

Database mining

A dataset of 111,090 IPK-barley ESTs was screened using the MIcroSAtellite (MISA) software (Thiel et al. 2003). This identified 9,564 (8.6%) redundant SSRs in 8,766 (7.9%) ESTs. Cluster analysis of these SSR-ESTs yielded a final number of 3,122 (2.8%) non-redundant SSRs present in 2,823 ESTs. As expected, trimeric SSRs constitute the major portion at 52.6 and 63.4% of the total SSRs identified in non-redundant and redundant SSR–EST sets, respectively. Pentameric and hexameric microsatellites were present at less than 1% of all the SSRs searched. The SSR motif AG/CT among dimeric SSRs and the motif CCG/CGG among trimeric SSRs were the most abundant. The most frequent tetrameric microsatellite motif was ACGT/ATGC. No specific trends were observed for pentameric and hexameric SSR motifs.

SSR polymorphism

For the development of EST–SSR markers, SSR–ESTs were selected by using two strategies (Fig. 1). In the first approach, non-redundant SSR-ESTs were selected from the set of 111,090 IPK-ESTs for designing primer pairs. A total of 665 primer pairs (including 311 primer pairs reported earlier in Thiel et al. 2003) were employed to amplify the corresponding SSR loci in the set of seven genotypes. Of the 464 primer pairs (69.8%) which yielded amplicons in the analysed genotypes, 156 primer pairs (33.6%) displayed polymorphisms between the parents of at least one mapping population.

To enhance the level of polymorphism in the genotypes of interest, a second strategy, based on additional non-IPK barley ESTs from the public domain, was adopted. While the IPK-ESTs were derived from the cultivar Barke, non-IPK ESTs from the public domain were developed from a diverse series of cultivars (Kota et al. 2003). Hence, a comparison of Barke to non-Barke ESTs would allow a pre-selection of polymorphic SSRs. The 207,449 non-IPK barley ESTs were screened for the presence of SSRs. This resulted in the identification of 18,041 SSR–ESTs containing 24,623 SSRs. In the combined set of 318,539 ESTs, a total of 26,807 ESTs were found to contain 34,187 redundant SSRs and 7,438 non-redundant SSRs. This corresponds to one SSR every 4.78 kb in the transcriptome of barley based on 163.5 Mb of EST sequence analysed.

The combined, total set of redundant SSR–ESTs was subjected to cluster analysis. As a result three types of clusters containing SSR–ESTs were observed from (a) only IPK ESTs, (b) only non-IPK ESTs, and (c) mixed clusters containing IPK and non-IPK ESTs (Fig. 1). Mixed clusters, were further analyzed after preparing the consensus sequence of IPK ESTs and non-IPK ESTs, separately. These two consensus sequences of the mixed clusters were compared to detect variation in SSR length. Altogether a total of 197 mixed clusters containing SSR–ESTs that showed variation in SSR length between IPK and non-IPK ESTs were identified. Of this set, 89 IPK SSR–ESTs were selected for amplification of the corresponding microsatellite loci in the set of seven accessions, and amplicons were obtained with 61 (68.5%) of the primer pairs. Of these, 38 (62.3%) SSR–ESTs showed polymorphisms between two or more of the seven genotypes used in the present study, and 29 detected polymorphisms that could be mapped in the corresponding mapping populations. Thus the level of polymorphism detected in parental genotypes of at least one mapping population (29/61) was increased by 13.9%, and this higher level of polymorphism, compared to the first strategy, was statistically significant (χ^2 test, $P < 0.01$).

Using both strategies a total of 754 primer pairs were analysed on the set of seven genotypes. A total of 525 (69.6%) primer pairs yielded amplicons of which 185 (35.2%) primer pairs detected polymorphism between parents of at least one mapping population (Fig. 1).

Genetic mapping of microsatellite loci

A total of 129 of the 185 polymorphic markers were mapped in the $OWB_{Rec} \times OWB_{Dom}$, 47 in the Steptoe × Morex and 23 in the Igri × Franka population. Twelve SSRs were mapped in both Igri × Franka and OWB while two SSRs were mapped in Steptoe × Morex and OWB. In addition to these common SSR markers available anchor markers were used to construct a consensus map of all three mapping populations. The EST–SSR markers were fairly evenly distributed with numbers ranging from 21 (7H) to 35 (3H) at an average of 27 per chromosome (Fig. 2, Tables 1 and ESM 1), with some gaps (around 25 cM)

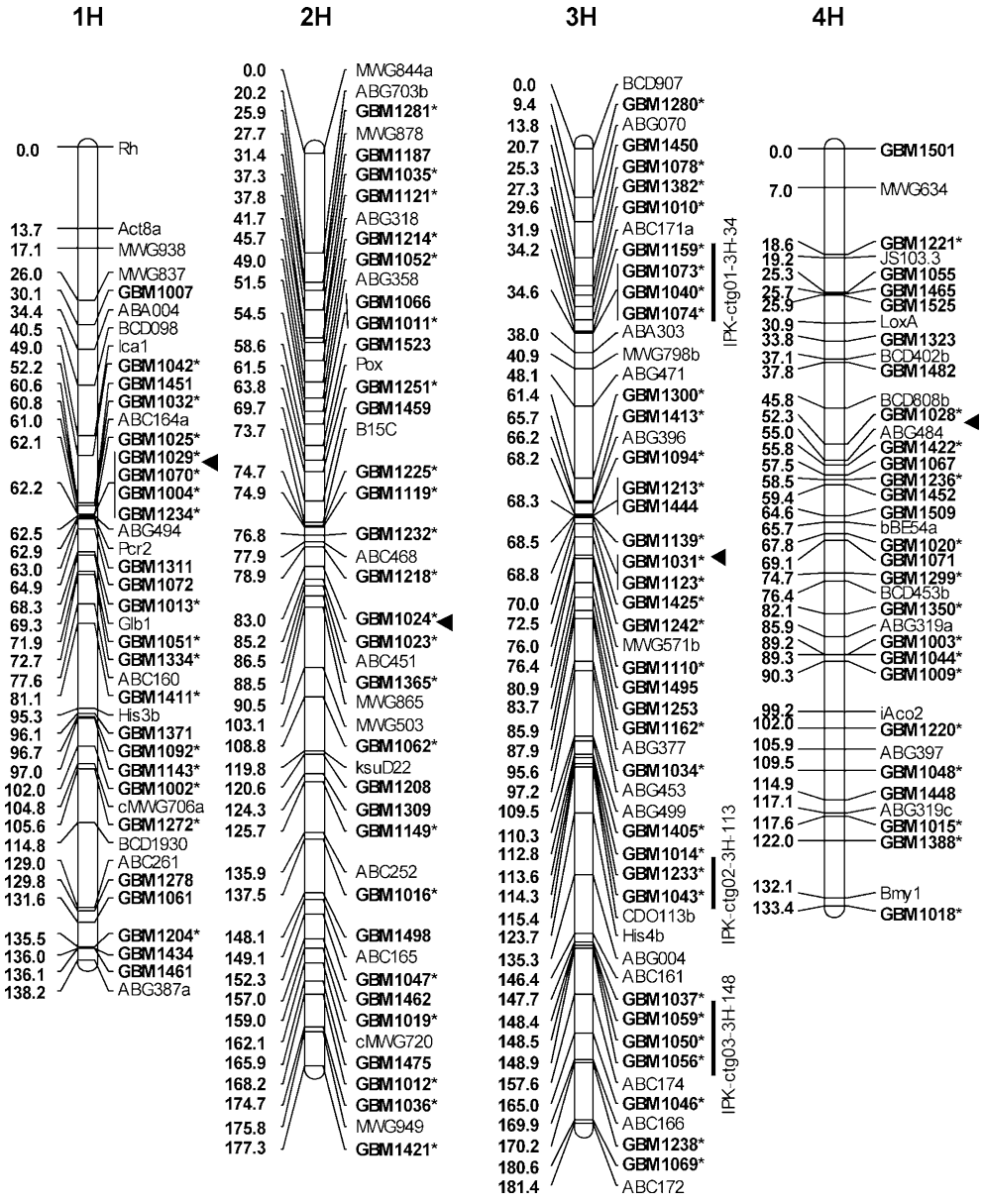

Fig. 2 Consensus genetic map of barley based on EST-derived SSR markers. Linkage groups are oriented with the short chromosome arms at the *top*. The genic-SSR loci are shown in *bold letters*. To assist the alignment with existing maps, BIN markers (shown in *normal font*) from the consensus map of Kleinhofs and Graner (2001) were included. *Arrowheads* indicate the approximate position of centromeres (according to Künzel et al. 2000). Markers, for which BAC addresses are available, are indicated by an *asterisk*. Eight contigs as given in Table 2 are shown as *vertical bars* along the genetic map

on chromosome 6H and short arm of chromosomes 1 and 5H.

Polymorphism information content

For all mapped markers, the PIC value was calculated on the basis of observed alleles in six (with 76 markers GBM1001–GBM1076) or seven genotypes (with remaining markers). The mapped markers detected 2–5 alleles with an average of 2.7 alleles per locus. Their corresponding PIC values ranged from 0.24 to 0.78 with an average of 0.48. About half of the markers (95) displayed PIC values greater than 0.50 (Table ESM 1). Markers derived from 3′ESTs showed higher PIC values than those derived from 5′ESTs. For instance, 33.3% of the markers from 3′ESTs and 25.2% of the markers from 5′ESTs had a PIC value greater than 0.60 (Table ESM 1). Regarding their SSR motifs, 36.9% of the dimeric, 19.5% of the trimeric and 25% of the tetrameric microsatellites had PIC values greater than 0.60.

A putative gene function could be assigned to 103 (55.7%) mapped SSR–EST markers based on a comparison to the NR-PEP protein sequence database. Among these markers, 69 showed homology with known proteins, 21 with putative proteins, six with hypothetical proteins and 7 with (presently) unknown proteins (Table ESM 1). The remaining 82 SSR–EST markers did not show any significant homology to a known protein.

Fig. 2 continued

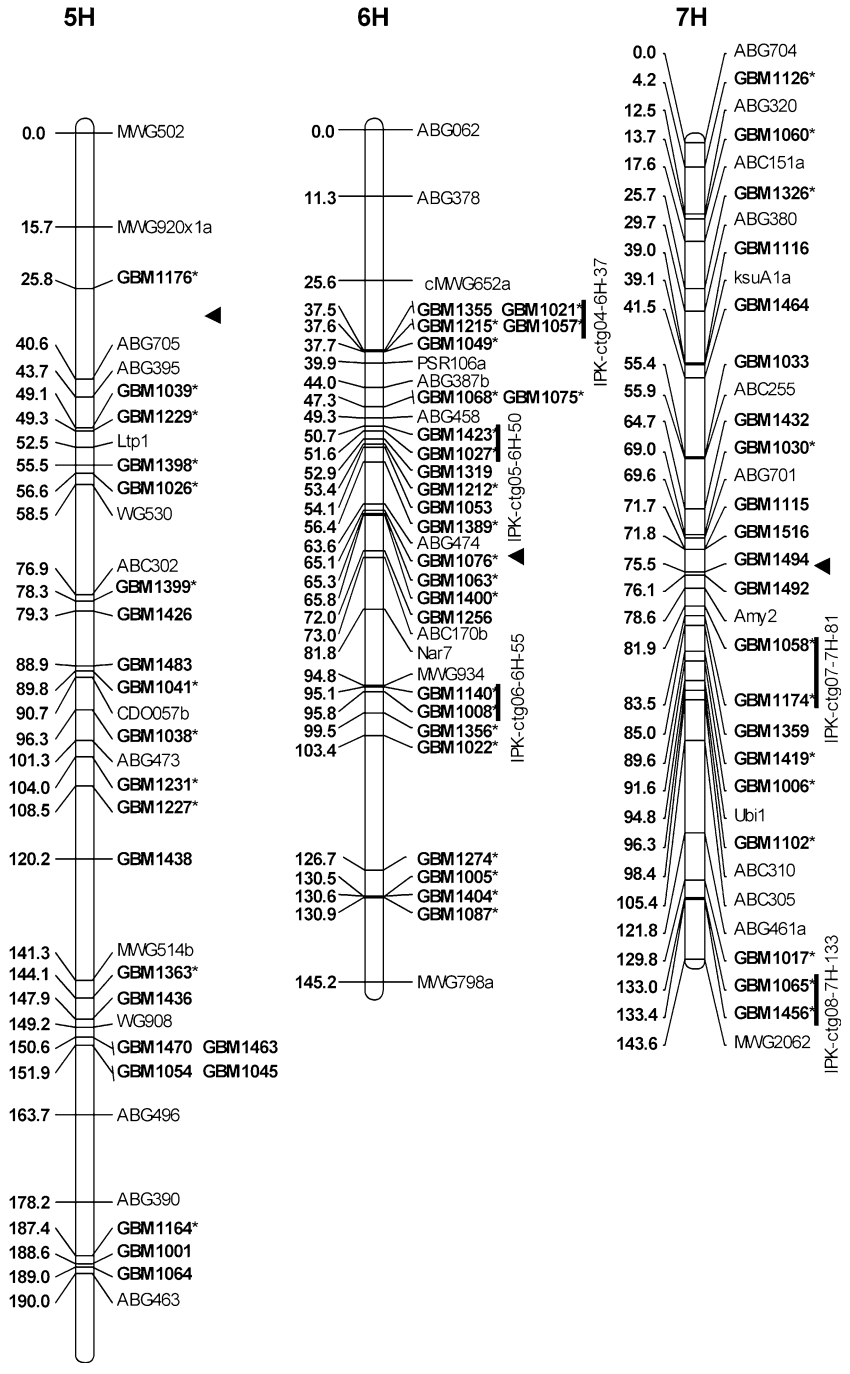

Table 1 Summary on EST-SSR markers

Number of markers	1H	2H	3H	4H	5H	6H	7H	Total
Genetically mapped	25	31	35	26	22	25	21	185
BAC addresses obtained	16	22	31	15	12	21	12	129
Functional annotation	12	18	23	13	11	12	14	103
Range of PIC value	0.23–0 .73	0.24–0.69	0.24–0.78	0.24–0.69	0.28–0.69	0.28–0.72	0.24–0.69	0.24–0.78 (mean 0.48)

Assignment of SSR markers to BAC clones

Gene-containing BAC clones

A genomic BAC library of barley was screened using 129 EST-based SSR markers. These were randomly selected from the set of 185 mapped SSRs to create anchor points between the genetic map and a 'future' physical map of barley. A four-step PCR-based screening strategy was developed employing DNA of BAC pools and a confirmation step at the level of individual clones. Using that strategy the BAC clone addresses were obtained for 127 (98%) of the SSR markers assayed (Table ESM 1).

Almost the complete (> 99%; 311,040 clones of the available 313,344 clones) BAC library was screened during the initial screening step performed on 90 super-pools each containing the DNA of 3,456 BAC clones. In order to save time and costs typically three super-pools were then chosen to continue the screening process. 311 individual BAC clones could be identified from a total of 384 positive super pools selected in this way. Their plate addresses are given in Table ESM 1. For 15 primer pairs, 2 to 5 BAC clones were identified on a single microtiter plate of the ordered BAC library. This frequency is almost 40 times greater than the expected value. Although these BAC clones do not occupy neighbouring wells in all cases, we regarded them as cross-contaminations, which may have occurred during construction, transport or copying of the library and counted them as one BAC clone in the statistical analysis outlined below. Previous testing of the used copy of the library via hybridization of radiolabelled gene probes to colonies spotted onto nylon membranes had not revealed this problem, which may have been uncovered only due to the higher sensitivity of PCR. At present, we observed cross-contaminations in 15 microtiter plates out of 254 from which we obtained individual clones during this study. Therefore, users of the BAC clone information should be aware that testing of individual colonies from an address of the ordered library is required to confirm the given information.

Identification of gene-rich BAC clones

Three positive super pools were selected at random for each marker to be subsequently resolved down to the level of individual BAC clones. Thus only about 50% of the positive pools were analysed (because in a 6× library six hits would be expected on average). While in the majority of cases BACs were hit by a single marker only, six BAC clones harbouring at least two markers were identified. EST cluster information (see project 'g03' at http://pgrc.ipk-gatersleben.de/cr-est/ on 135,031 barley ESTs that includes 111,090 barley ESTs, searched for SSRs in the present study) and the BLASTX analysis showed that these markers were derived from independent genes.

In one case, two closely linked markers (GBM1058 and GBM1174) separated by a map distance of 1.6 cM hit the same BAC clone (804L09). To further elucidate whether this was due to coincidence, artefact or real physical linkage, all markers mapping closer than 2 cM to each other were tested by PCR on those BAC clones previously identified by the neighbouring marker. As a result a total of eight groups of markers were identified on three different chromosomes (3H, 6H and 7H), which contain one or more BAC clones on which at least two markers are located (Table 2). GBM1159 and GBM1073 in the contig IPK-ctg01-3H-34 (containing total four markers) and GBM1059 and GBM1050 in the contig IPK-ctg03-3H-148 (containing total three markers) were derived from the genes which contained two different SSR motifs, respectively. Cluster analysis of the corresponding EST sequences confirmed that independent sequences were amplified by the respective markers. To rule out the possibility that during the clustering non-overlapping ESTs derived from one gene might have been assigned to two different contigs, the corresponding consensus sequences were blasted against the rice genome. All of the consensus sequences detected independent loci in rice demonstrating that they belong to different genes.

In addition, re-screening of the BAC clones with neighbouring markers (linked at less than 2 cM) provided the BAC addresses for two additional SSR markers. As a result, the BAC addresses are now available for 129 markers (Table 1).

Computational analysis on gene-distribution

Screening of the BAC library yielded eight BAC clones hit by at least two markers. Two BACs (312D09 and 526J23) were hit by GBM1073 and GBM1159 representing the identical gene. However, the markers for the remaining 6 BAC clones [(194B18, 582L01 (hit by GBM1056 and GBM1059), 486M04 (hit by GBM1050 and GBM1056), 499N10 (hit by GBM1049 and GBM1057), 516M05 (hit by GBM1021, GBM1049 and GBM1057) and 804L09 (hit by GBM1058 and GBM1174)] represent different genes. Thus 310,729 BAC clones were not hit by any marker, 305 BAC clones were hit by exactly one marker, 5 BAC clones were hit by exactly two markers, 1 BAC clone was hit by exactly three markers, and no BAC was hit by more

Table 2 Bacterial artificial chromosomes contigs identified by EST–SSR markers

Contig ID[a]	Marker ID	Position on genetic map	BAC clones[b]
IPK-ctg01-3H-34 (BPMD-ctg2012)	GBM1159[c]	3H-34.2	**064A24**, **312D09**, **526J23**, **579B12**
	GBM1073[c]	3H-34.6	**064A24**, **312D09**, **526J23**, **579B12**
	GBM1040	3H-34.6	**064A24**, **169D02**, **312D09**, **526J23**, **579B12**
	GBM1074	3H-34.6	**169D02**
IPK-ctg02-3H-113 (BPMD-ctg1199)	GBM1233	3H-113.6	**41C24**, 93E12, 305H14
	GBM1043	3H-114.3	**41C24**, 274P02, 559M03
IPK-ctg03-3H-148 (BPMD-ctg2068)	GBM1059[d]	3H-148.4	**194B18**, **486M04**, **509D02**, **537M01**, **582L01**, **679O07**
	GBM1050[d]	3H-148.5	**486M04**, **537M01**, **679O07**
	GBM1056	3H-148.9	**194B18**, **486M04**, **509D02**, **537M01**, **582L01**, **679O07**
IPK-ctg04-6H-37 (BPMD-ctg1122)	GBM1021	6H-37.5	301H19, **516M05**, **676O18**
	GBM1057	6H-37.6	**334A06**, **499N10**, **516M05**, **676O18**
	GBM1049	6H-37.7	**334A06**, **499N10**, **516M05**, **676O18**
IPK-ctg05-6H-50 (BPMD-ctg1887)	GBM1423	6H-50.7	**214C05**, **256M21**, 256M22, 317A19, 676C09
	GBM1212	6H-53.4	**214C05**, 230C05, **256M21**, 579A07
IPK-ctg06-6H-95	GBM1140	6H-95.1	113D01, 113F01, 306D14, 351F23, **810K09**
	GBM1008	6H-95.8	793I12, **810K09**
IPK-ctg07-7H-81 (BPMD-ctg619)	GBM1058	7H-81.9	**290K01**, **310J18**, **474G04**, **804L09**
	GBM1174	7H-83.5	**290K01**, **310J18**, **474G04**, **804L09**
IPK-ctg08-7H-133 (BPMD-ctg534)	GBM1065	7H-133.0	**111C05**, **274E14**, **536G12**
	GBM1456	7H-133.4	**111C05**, **274E14**, **536G12**

[a]Contigs prepared at IPK are designated according to the position of their markers on the genetic map. Contig IDs mentioned in parenthesis are obtained from the US Barley Physical Mapping Database (BPMD, http://phymap.ucdavis.edu:8080/barley/index.jsp)

[b]BAC clones typed in bold font were hit by more than one marker, underlined BAC clones present in one micro-titre plate, in neighbouring wells

[c]Markers derived from two ESTs of the same consensus sequence

[d]Markers derived from the same EST representing different SSR motifs

than three markers. This yielded a total of $305 \times 1 + 5 \times 2 + 1 \times 3 = 318$ BAC hits, and a total number of $C = 5 \times 1 + 1 \times 2 = 7$ collisions (because there are 5 BACs with $C_i = 1$, and there is one BAC with $C_i = 2$).

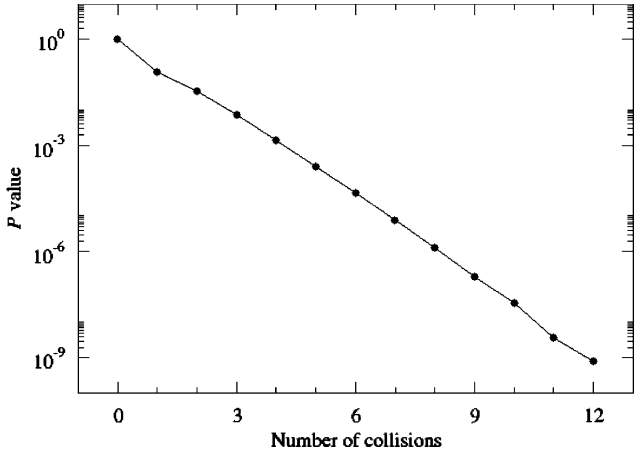

Fig. 3 Probability P_C of finding C, or more than C, collisions versus C. *Circles* represent the mean values of P_C for C ranging from 0 to 12 based on 2×10^{10} simulations. For $C < 11$ the 95% confidence intervals of P_C are smaller than the size of the *circles*. The probability of finding 7, or more than 7, collisions is $p_7 = 7.6 \times 10^{-6} \pm 0.03 \times 10^{-6}$, stating that the null hypothesis of a uniform distribution of the 129 markers across the 311,040 BACs must be rejected

The identification of BAC clones hit by more than one marker (gene) or in other terms, the identification of seven collisions (C) provided a clue about the non-random distribution of genes in the barley genome. Under the null hypothesis that the 129 markers are distributed uniformly across the 311,040 BACs, the probability of finding seven or more than seven collisions is 7.6×10^{-6} (Fig. 3). As this P value is smaller than 0.05, the null hypothesis that the distribution of markers across the BACs is uniform in the barley genome must be rejected.

Discussion

In the present study, 109 novel EST-derived SSR markers were developed further increasing the number of EST–SSRs on our genetic consensus map to 185. The assignment of 129 SSRs to the corresponding clones of the Morex BAC library will provide a first resource of connecting points between existing genetic maps and upcoming local physical maps (http://phymap.ucdavis.edu:8080/barley/index.jsp) of the barley genome.

Frequency of microsatellites in the barley transcriptome

Over the past 5 years, large-scale genome and EST sequencing projects were initiated in several plant species including cereals. The data generated from these projects was utilized for studying the frequency, distribution and organization of microsatellites in the expressed portion of the genome, and in some cases also in the whole genome (reviewed by Varshney et al. 2005a). In the present study, a total of 9,564 (8.6%) redundant and 3,122 (2.8%) non-redundant SSRs were identified in a dataset of 111,090 ESTs. The overall trend in the frequency and distribution of different classes of SSRs agreed with results obtained in earlier studies (Varshney et al. 2002, 2005a, 2005b Thiel et al. 2003).

Level of polymorphism of EST–SSRs

For the development of microsatellite markers, two strategies were adopted. Using the first strategy, primers were developed flanking SSR-motifs in a set of randomly selected ESTs. Here 33.6% of the functional primer pairs displayed a polymorphism between the parents of the mapping populations. In the second strategy, which was based on the in silico pre-selection of SSR-containing ESTs that were polymorphic between the cultivars Barke and varieties other than Barke, 47.5% of the functional primer pairs could be mapped in at least one of the populations studied. The application of the second strategy resulted in a significant increase in the detection rate of polymorphisms. In 38 out of 41 SSR–ESTs, in silico polymorphism was confirmed on the experimental level. In three cases, however, a 2 bp polymorphism in silico predicted between Barke and Morex SSR–ESTs could not be confirmed. This may be due to the limited resolving power of the polyacrylamide gel system used. The remaining 20 SSR–ESTs showed polymorphisms between Barke and Japanese genotypes (Haruna Nijo, Akashinriki and H602). While these SSRs were monomorphic in the genotypes analyzed in this study, 11 (55%) out of these 20 SSR–ESTs showed polymorphisms in the parental genotypes of two other mapping populations and were successfully integrated into genetic maps (R. Niks, Wageningen, personal communication). Hence, the computational pre-selection of polymorphic SSRs presents an efficient strategy to increase the success in the development of informative SSR markers. Similar results were obtained from database mining for SNPs, where the computational selection of polymorphic SNPs in the EST-database increased the likelihood of detecting polymorphism in a given set of germplasm (Kota et al. 2003).

Approximately 70% of the analysed primer pairs were functional of which 35.2% were polymorphic in the parents of the mapping populations. Failure of amplification in the remaining 30% primer pairs may have been due to primer mismatch, the extension of primers across a splice site or the presence of large introns in the genomic DNA fragment to be amplified. Moreover, only one standard amplification protocol was applied to all markers, and no efforts were undertaken to optimize amplification conditions for unsuccessful primer pairs (Thiel et al. 2003). As reflected by an average PIC value of 0.48, the polymorphism of EST-derived microsatellites is lower than that of genomic DNA-derived microsatellites (Ramsay et al. 2000; Li et al. 2003).

Development of functional microsatellite markers

Genetic mapping of 185 microsatellite markers showed no obvious clustering around the centromere as was observed for microsatellites derived from genomic DNA (Ramsay et al. 2000; Li et al. 2003). The observation of a higher number of markers on chromosome 3H (18.8%) and chromosome 2H (17.2%) suggests the presence of more genes on these two chromosomes as has also been observed in case of a transcript map containing more than 1,000 genes (unpublished results). Physical mapping of ESTs using deletion lines of wheat also revealed the highest number of EST loci on homologous group 3 (16.3%) followed by group 2 (15.9%) (http://wheat.pw.usda.gov/cgi-bin/westsql/map_locus.cgi, Qi et al. 2004).

Although it seems that further markers are required to cover the distal portions of chromosome arms 1HS, 5HS, 6HS and 6HL, some of these gaps correspond to gaps on other published maps (Kleinhofs et al. 1993; Ramsay et al. 2000). Synteny between rice and barley may facilitate filling up these gaps, if required, by using bioinformatics analyses in combination with available barley ESTs and the completed rice genome sequence, as it has been shown recently (Perovic et al. 2004; Varshney et al. 2005b).

Physical anchoring of mapped SSR–ESTs and identification of gene-rich BAC clones

The availability of genome-wide BAC contigs has been an invaluable resource for sequencing the genomes of Arabidopsis and rice (Sasaki and Burr 2000). However, progress in contig-based physical mapping of the

barley genome is slow because of the large genome size and the limited efforts that have been initiated to date. To further circumvent the complexity of large genomes, novel approaches such as methylation and C_0t filtration are being investigated to reduce the representation of heterochromatic regions in BAC libraries and to confine the construction of physical maps to the gene space (Rabinowicz et al. 2003; Peterson et al. 2002). Another approach to enrich for coding regions may be the pre-selection of gene-containing BACs.

Bacterial artificial chromosomes screening and genetic mapping revealed eight groups of markers which selected one or more BAC clones with at least two genes. A comparison of the presented data with the US Barley Physical Mapping Database (BPMD; http://phymap.ucdavis.edu:8080/barley/index.jsp) confirms the existence of seven contigs. BAC clones for the remaining contig (IPK-ctg06-6H-95) were not present in the BPMD at the time of this analysis.

The fact that markers are spaced at 2.7 cM and yet co-locate on a single BAC could be indicative of a recombination hotspot. On the other hand the observed map distance may have been overestimated as a result of merging the segregation data of the three different populations employed in this study into a consensus map. Inaccurate estimates of small map intervals are almost unavoidable if neighboring markers were mapped in different populations (Somers et al. 2004), as is the case with markers GBM1423 and GBM1212 in contig IPK-ctg05-6H-50, that were mapped in the OWB and the Steptoe × Morex populations, respectively.

Non-random distribution of genes in the genome

The statistical analysis of the distribution of the SSR markers across the BAC library provided evidence that the distribution of genes across the BACs is not uniform on a genome-wide scale. A non-uniform gene distribution in large genomes of cereal crops including barley was suggested earlier by buoyancy density gradient methods (Carels et al. 1995; Barakat et al. 1997) or by a comparison of genetic with physical maps based on cytogenetic stocks (Künzel et al. 2000; Erayman et al. 2004; Gill 2004). The results of the present study provide evidence that this non-uniform gene distribution on the physical level extends to a resolution of about 100 kb (the average size of a BAC clone). Sequence analysis of DNA contigs in wheat and barley has already pointed at the presence of clusters of closely linked genes forming gene islands that are separated by large stretches of repetitive DNA (e.g. Rostocks et al. 2002; Gill 2004; Wicker et al. 2001, 2005). The majority of the corresponding data are biased towards disease resistance genes, raising the possibility that their pattern of distribution is not representative for the whole genome. In the present study, functional annotation of genes mapping to a single BAC does not show any bias towards a specific functional category. However, two contigs comprised markers representing different members of gene families (GBM1040, GBM1073, GBM1074; GBM1065, GBM1456). Since there is evidence that about 50% of the barley genes represent members of gene families, this result is expected (Zhang et al. 2004). Although unlikely, we can not completely rule out that the findings of the present study pertain only to genes containing SSRs. BAC selection by EST-derived probes devoid of SSRs will help to address this issue.

Acknowledgements We are grateful to Timothy J. Close (University of California, Riverside, USA) for his valuable suggestions on the physical mapping data. We thank Uwe Scholz and Christian Künne (IPK) for performing cluster analysis of SSR-ESTs of IPK and non-IPK ESTs, and Paul Krapivsky (Boston University, Boston, USA), Stefan Posch (Martin Luther University Halle-Wittenberg, Halle, Germany) and Roland Schnee (IPK) for helpful discussions. We also thank Christine Künzel, Anita Czech, Brigitte Schmidt and Ingelore Dommers for technical assistance. The present work was funded by grants from the Grain Research and Development Corporation, Australia (GRDC, UA476), the Federal Ministry of Education and Research (BMBF, GABI-PLANT 312271A,B,C) and BMBF Bioinformatics Centre, Gatersleben/Halle 0312706A).

References

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

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Andersen JR, Lübberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554–560

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci USA 94:6857–6861

Bennett MD, Leitch IJ (2003) Angiosperm DNA C-values database (release 4.0, January 2003).

Carels N, Barakat A, Bernardi G (1995) The gene distribution of the maize genome. Proc Natl Acad Sci USA 92:11057–11060

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32:3546–3565

Gao LF, Jing RL,Hu NX, Li XP, Zhou RH, Chang XP, Tang JF, Ma ZY, Jia JZ (2004) One hundred and one new microsatellite loci derived from ESTs (EST–SSRs) in bread wheat. Theor Appl Genet 108:1392–1400

Gill KS (2004) Gene distribution in cereal genomes. In: Gupta PK, Varshney RK (eds) Cereal genomics. Kluwer, Dordrecht, pp 361–384

Holton TA, Christopher JT, McClure L, Harker N, Henry RJ (2002) Identification and mapping of polymorphic SSR

markers from expressed gene sequences of barley and wheat. Mol Breed 9:63–71
Keller B, Feuillet C (2000) Colinearity and gene density in grass genomes. Trends Plant Sci 5:246–251
Kleinhofs A, Graner A (2001) An integrated map of the barley genome. In: Phillips RL, Vasil IK (eds) DNA markers in plants. Kluwer, Dordrecht, pp 187–199
Kleinhofs A, Kilian A, Saghai-Maroof MA, Biyashev RM, Hayes PM (1993) A molecular, isozyme and morphological map of barley (*Hordeum vulgare*) genome. Theor Appl Genet 86:705–712
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 Gen Genome 270:224–233
Künzel G, Korzun L, Meister A (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. Genetics 154:397–412
Li JZ, Sjakste TG, Röder MS, Ganal MW (2003) Development and genetic mapping of 127 new microsatellite markers in barley. Theor Appl Genet 107:1021–1027
Miller RT, Christoffels AG, Gopalkrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res 9:1143–1155
Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002) Integration of cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807
Perovic D, Stein N, Zhang H, Drescher A, Prasad M, Kota R, Kopahnke D, Graner A (2004) An integrated approach for comparative mapping in rice and barley based on genomic resources reveals a large number of syntenic markers but no candidate gene for the *Rph16* resistance locus. Funct Integr Genomics 4:74–83
Pillen K, Binder A, Kreuzkam B, Ramsay L, Waugh R, Förster J, Leon J (2000) Mapping new EMBL-derived barley microsatellites and their use in differentiating German barley cultivars. Theor Appl Genet 101:652–660
Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J 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
Rabinowicz PD, McCombie WR, Martienssen RA (2003) Gene enrichment in plant genomic shotgun libraries. Curr Opin Plant Biol 6:150–156
Ramsay L, Macaulay M, Ivanissevich DS,MacLean K, Cardle L, Fuller J, Edwards KJ, Tuvesson S, Morgante M, Massari A, Maestri E, Marmiroli N, Sjakste T, Ganal M, Powell W, Waugh R (2000) A simple sequence repeat-based linkage map of barley. Genetics 156:1997–2005
Rostoks N, Park YJ, Ramakrishna W, Ma J, Druka A, Shiloff BA, SanMiguel PJ, Jiang Z, Brueggeman R, Sandhu D, Gill K, Bennetzen JL, Kleinhofs A. (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. Funct Integr Genomics 2:51–59
Sasaki T, Burr B (2000) International rice genome sequencing project: the effort to completely sequence the rice genome. Curr Opin Plant Biol 3:138–141
Schulman AH, Gupta PK, Varshney RK (2004) Organization of retrotransposons and microsatellites in cereal genomes. In: Gupta PK, Varshney RK (eds) Cereal genomics. Kluwer, Dordrecht, pp 83–118
Somers DJ, Isaac P, Edwards K (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). Theor Appl Genet 109:1105–1114
Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–774
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
Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7:537–546
Varshney RK, Korzun V, Börner A (2004) Molecular maps in cereals: methodology and progress. In: Gupta PK, Varshney RK (eds) Cereal genomics. Kluwer, Dordrecht, pp 35–82
Varshney RK, Graner A, Sorrells ME (2005a) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55
Varshney RK, Sigmund R, Borner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A (2005b) Interspecific transferability and comparative mapping of barley EST–SSR markers in wheat, rye and rice. Plant Sci 168:195–202
Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contigous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316
Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, Graner A, Stein N (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-eif4e* locus: recombination, re-arrangements, and repeats. Plant J 41:184–194
Yu Y, Tmkins JP, Waugh R, Frisch A, Kleinhofs A, Brueggeman RS, Muehlbauer GJ, Wise RP, Wing RA (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. Theor Appl Genet 101:1093–1099
Zhang H, Sreenivasulu N, Weschke W, Stein N, Rudd S, Radchuk V, Potokina E, Scholz U, Schweizer P, Zierold U, Langridge P, Varshney R K, Wobus U, Graner A (2004). Large-scale analysis of the barley transcriptome based on expressed sequence tags. Plant J 40:276–290