Funct Integr Genomics (2002) 2:51–59
DOI 10.1007/s10142-002-0055-5

ORIGINAL PAPER

**Nils Rostoks · Yong-Jin Park · Wusirika Ramakrishna
Jianxin Ma · Arnis Druka · Bryan A. Shiloff
Phillip J. SanMiguel · Zeyu Jiang
Robert Brueggeman · Devinder Sandhu
Kulvinder Gill · Jeffrey L. Bennetzen
Andris Kleinhofs**

# Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley

Received: 7 January 2002 / Accepted: 7 March 2002 / Published online: 25 April 2002


**Abstract** Barley (*Hordeum vulgare* L.) is one of the most important large-genome cereals with extensive genetic resources available in the public sector. Studies of genome organization in barley have been limited primarily to genetic markers and sparse sequence data. Here we report sequence analysis of 417.5 kb DNA from four BAC clones from different genomic locations. Sequences were analyzed with respect to gene content, the arrangement of repetitive sequences and the relationship of gene density to recombination frequencies. Gene densities ranged from 1 gene per 12 kb to 1 gene per 103 kb with an average of 1 gene per 21 kb. In general, genes were organized into islands separated by large blocks of nested retrotransposons. Single genes in apparent isolation were also found. Genes occupied 11% of the total sequence, LTR retrotransposons and other repeated elements accounted for 51.9% and the remaining 37.1% could not be annotated.

**Keywords** Barley · Genomic sequence · Gene density · Repeated sequences

N. Rostoks · A. Druka · R. Brueggeman
Department of Crop and Soil Sciences,
Washington State University, Pullman, WA 99164, USA

Y.-J. Park · W. Ramakrishna · J. Ma · P.J. SanMiguel
J.L. Bennetzen
Department of Biological Sciences, Purdue University,
West Lafayette, IN 47907, USA

B.A. Shiloff · Z. Jiang
National Center for Genome Resources,
2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA

D. Sandhu
G302 Agronomy Hall, Iowa State University, Ames,
IA 50011–1010, USA

K. Gill
Department of Agronomy, University of Nebraska,
Lincoln, NE 68583, USA

A. Kleinhofs (✉)
School of Molecular Biosciences
and Department of Crop and Soil Sciences,
Washington State University, Pullman, WA 99164, USA
e-mail: andyk@wsu.edu
Tel.: +1-509-3354389, Fax: +1-509-3358674

## Introduction

Barley is an important crop throughout the world that has been extensively used for genetic studies. There are numerous detailed barley genetic maps available (Graner et al. 1991; Heun et al. 1991; Kleinhofs et al. 1993; Becker et al. 1995; Waugh et al. 1997; Qi et al. 1998; Kunzel et al. 2000; Ramsay et al. 2000; Kleinhofs and Graner 2001; Mano et al. 2001). An RFLP map based on a Steptoe × Morex cross is regularly updated and available on-line at http://barleygenomics.wsu.edu. The barley cv. Morex has been used as the source of a 6.3× genome coverage, bacterial artificial chromosome (BAC) library and serves as a valuable resource for the research community (Yu et al. 2000). In addition, over 140,000 barley-expressed sequence tags (EST) have been deposited and are available at (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

During the past 5 years a significant increase in our understanding of the structure and dynamics of plant genomes has emerged as more genomic DNA sequences became available. A remarkable example was the study by SanMiguel et al. (1996) of a 280-kb DNA sequence flanking the maize *adh1* gene. This study suggested that at least 50% of the maize genome is comprised of different long terminal repeat (LTR) retrotransposons arranged primarily as nested insertions. Comparison of orthologous *adh* regions from maize and sorghum revealed largely conserved gene content and order, but significant differences in the intergenic space (Tikhonov et al. 1999). While the intergenic space in sorghum contained several

repeated sequences, it lacked LTR retrotransposons like those abundant in the maize genome. This supported a previous study suggesting that the maize genome had expanded in size at least twofold in the last 5 million years due to retrotransposon insertions (SanMiguel et al. 1998). These studies also provided evidence that local gene densities in maize and sorghum were often significantly higher than the genome average. Studies in diploid wheat (*Triticum monococcum* L.) found three genes within a 31-kb cluster separated from other genes by extensive and often nested retrotransposon insertions (Wicker et al. 2001).

In barley, genomic sequences of large contiguous regions were not available until recently. Previous to this study, only three barley genomic sequences larger than 50 kb had been reported: the chromosome 5(1H) WG644 locus (102 kb; Dubcovsky et al. 2001), the chromosome 2(2H) *Rar1* locus (66 kb; Shirasu et al. 2000) and the chromosome 4(4H) *Mlo* locus (60 kb; Panstruga et al. 1998). In addition, 240 kb of the *Mla* gene cluster on chromosome 5(1H) has been characterized in significant detail, although not completely sequenced (Wei et al. 1999).

Even this limited amount of sequence information provided significant insights into barley genome organization. All barley genome studies found similar gene densities of about 1 gene per 20 kb. This is about tenfold higher than the average gene density (1 gene per 200 kb) calculated for barley assuming a genome size of 5,000 Mb (Arumuganathan and Earle 1991) and the number of genes similar to *Arabidopsis thaliana* (25,498 genes; The Arabidopsis Genome Initiative 2000). Most of the genes were clustered in "gene islands", although single genes were encountered as well. The intergenic space consisted primarily of *Ty1-copia* and *Ty3-gypsy* group LTR retrotransposons. The most abundant of those was *BARE-1*, the major component of barley genome repetitive DNA (Manninen and Schulman 1993; Suoniemi et al. 1996). *BARE-1* is actively transcribed, forming virus-like particles in barley cells (Jaaskelainen et al. 1999; Vicient et al. 2001). Several novel retrotransposons were also identified, including *BAGY-1* (Panstruga et al. 1998), *BAGY-2*, *Sabrina*, *Nikita* and *Sukkula* (Shirasu et al. 2000). Similar to maize, these elements were often arranged as nested retrotransposons. All studies found several full-length retrotransposons, solo LTRs and fragments of retrotransposons. Intraelement recombination between LTRs or unequal crossing over were implicated as possible mechanisms to counteract constant genome expansion due to retrotransposon propagation (Shirasu et al. 2000). Recently, Wicker et al. (2001) suggested that even larger genome regions consisting of different types of repetitive elements may be deleted, restricting the expansion of the *T. monococcum* L. genome, although a possible molecular mechanism was not demonstrated.

Meiotic recombination is not distributed uniformly along chromosomes, but appears to occur primarily within genes (Thuriaux 1977; Lichten and Goldman 1995; Puchta and Hohn 1996; Schnable et al. 1998). Recombination in the *bz* locus in maize, for example, is about 100-fold higher than in an average segment of the maize genome (Dooner et al. 1985; Dooner and Martinez-Ferez 1997) or in adjacent retrotransposon regions (Fu et al. 2002). The *bz* locus is also extremely gene-rich containing ten active genes within 32 kb (Fu et al. 2001). High recombination rates were also reported for the *B*, *A1*, *R*, and *Wx* loci (Eggleston et al. 1995; Patterson et al. 1995; Xu et al. 1995; Okagaki and Weil 1997). Evidence that recombination was associated with gene-rich regions has also been reported in wheat (Gill et al. 1996a, b; Faris et al. 2000; Sandhu and Gill 2002). Most of the recombination events detected by presumed gene, cDNA and *Pst*I, probes on the wheat chromosome 1 short arm were localized to only 15% of the physical length of the chromosome (Sandhu et al. 2001).

In barley, the average physical to genetic distance was estimated to be about 4 Mb/cM (Kleinhofs et al. 1993), however, a 0.1–298 Mb/cM range was calculated using a translocation breakpoint physical map (Kunzel et al. 2000). The high (less than 1 Mb/cM) recombination regions occupied about 4.5% of the genome and contained approximately 60% of all mapped cDNA and *Pst*I genomic probes (most likely representing genes). The ratio of physical to genetic distance in the *Mla* cluster on chromosome 5(1H) varied from 0.18 to 5 Mb/cM (Wei et al. 1999). The region on barley chromosome 1(7H) bin01 containing the barley homologues of the maize *Rp1-D* gene (Ayliffe et al. 2000) had a physical to genetic distance ratio of 525 kb/cM (Rostoks et al., in press). The association of high recombination rates with gene-rich regions suggests that focusing on these regions could permit efficient positional cloning of genes in large genome species.

In this study we present sequence analysis of barley BAC clones from four different chromosome locations with a combined length of 417.5 kb. Genes were often clustered in "gene islands" and occupied 11% of the sequence. Most of the sequence was due to LTR retrotransposons (40.1%) and other repetitive elements (11.4%).

## Materials and methods

### BAC clone selection and restriction mapping

Barley (BCD135, BCD1434 and *Wx*) and rice (RZ567) cDNAs were used as probes to select BAC clones for sequencing. BCD135 and RZ567 are anchor probes for comparative mapping of the grass genomes (Van Deynze et al. 1998). The *Wx* (pcWX27) and BCD1434 probes are described by Rohde et al. (1988) and Heun et al. (1991), respectively. A complete list of the barley BAC clones selected by each probe can be found at the Barley Genomics Laboratory Web page (http://barleygenomics.wsu.edu/db3/db3.html). Map locations of the probes and the addresses of BAC clones selected for sequencing are in Table 1.

BAC clones corresponding to each of the probes were identified by standard DNA hybridization to the arrayed barley BAC library derived from cv. Morex (Yu et al. 2000). Individual BAC clone DNA was *Hin*dIII-digested, electrophoresed in agarose gels, transferred to nylon filters and hybridized with the probe to deter-

**Table 1** Characteristics of the sequenced barley BAC clones

| BAC clone (length in bp) | 011o09 (77,136) | 259i16 (124,038) | 745c13 (102,842) | 773k14 (113,510) | Average values |
|---|---|---|---|---|---|
| GenBank no. | AF474982 | AF474373 | AF474071 | AF474072 | N/A[a] |
| Selection probe | BCD1434 | *Wx* | RZ567 | BCD135 | N/A |
| Map location | Unknown[b] | 1(7H) bin 02 | 6(6H) bin 07 | 2(2H) bin 13 | N/A |
| Base composition | A =27.1%; T=28.3%; C=22.7%; G=21.9% | A =27.9%; T=27.1%; C=22.0%; G1=22.9% | A =28.3%; T=26.8%; C=22.4%; G=22.5% | A =28.8%; T=27.6%; C=21.9%; G=21.6% | A =28.1%; T=27.4%; C=22.2%; G=22.3% |
| Predicted genes | 5 | 10 | 1 | 4 | 5 |
| Gene density | 15.4 kb | 12.4 kb | 102.8 kb | 28.4 kb | 20.8 kb |
| Gene regions (as %) | 22.9 | 16.7 | 3.4 | 3.9 | 11.0 |
| LTR retrotransposons (as %) | 51.7 | 21.1 | 56.8 | 38.9 | 40.4 |
| Other repetitive elements (as %) | 4.7 | 21.6 | 9.4 | 7.0 | 11.5 |
| Non-annotated sequences (as %) | 20.7 | 40.6 | 30.4 | 50.2 | 37.1 |

[a] *N/A* not applicable
[b] Map location of the *bcd1434* locus is 5(1H) bin 02. Actual map location of the 011o09 BAC clone is unknown, because it contains a paralog of the BCD1434 cDNA

mine that the same size band was present in all BAC clones and, hence, that they belonged to the same contiguous series of clones (contig). In order to select the BAC clone with the marker as centrally located as possible, restriction enzyme maps of the BAC contigs were constructed using BAC DNA digested with *Not*I in combination with several other eight-base recognition restriction enzymes. The fragments were separated by pulsed-field gel electrophoresis, blotted and hybridized with the probe to reveal its location within the map.

### BAC clone sequencing

BAC sequencing and sequence assembly were as described in Dubcovsky et al. (2001) except that sequence assembly was done at around 10× sequence coverage.

### Sequence annotation

BAC sequences were annotated using the Comparative Genomics System (CGS) tools (http://www.ncgr.org/cgs/) provided by the National Center for Genome Resources (NCGR). The CGS tools include integrated BLASTN, BLASTX and BLASTP searches of the non-redundant EST database and other non-redundant databases using the entire BAC clone sequences, as well as GENSCAN version 1.0 (Burge and Karlin 1997) and GeneMark.hmm version 2.2a (Lukashin and Borodovsky 1998) gene prediction programs. These programs were run locally at the NCGR. Different BLAST searches were used to identify expressed genes and putative gene products. Three gene prediction programs were used to confirm known genes and detect other possible genes: GENSCAN with maize parameters, GeneMark.hmm with rice parameters and FGENESH version 1.0 with maize parameters. The FGENESH predictions were run at http://www.softberry.com/.

BAC genomic regions were defined as genes if they found nearly identical or similar (BLASTN score >100) ESTs and did not have homology to known repetitive elements. BLASTX and BLASTP searches were performed to assign the predicted genes a putative function. If no significant matches were found, the predicted gene was postulated to encode an unknown protein. In case there were no corresponding cDNAs or homologous protein sequences, a particular sequence was only assigned as a hypothetical gene if at least two gene prediction programs indicated the presence of a gene. ESTs from barley or closely related species and full-length cDNAs, if available, were used to predict the intron/exon structure of a gene. If there were no matching cDNA sequences in GenBank, BLASTX data in combination with gene prediction programs were used to predict intron/exon structure. In cases, where two gene prediction programs indicated the presence of a gene but the predicted intron/exon structure was somewhat different, an attempt was made to integrate the two predictions based on ORFs and manual inspection of potential splice sites. In general, the prediction that yielded the largest coding sequence was accepted as most likely.

Repeated regions within BAC sequences were visualized using the two-dimensional dot-plot program Dotter (Sonnhammer and Durbin 1995). Positions of transposable elements were determined by combination of BLASTN and BLASTX searches against the GenBank non-redundant database and selected known transposable elements. The presence of large tandemly arranged DNA sequences, as determined by the two-dimensional dot-plot analysis, often indicated retrotransposon LTRs. The rice repeat database at TIGR (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml) was also used to find repetitive sequences. The presence of tRNAs was checked by the tRNAscan-SE program at http://www.genetics.wustl.edu/eddy/tRNAscan-SE/ (Lowe and Eddy 1997). Simple sequence repeats were identified by RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html (Smit and Green, unpublished results).

### Genetic mapping

The 150-line Steptoe × Morex doubled haploid population developed by the North American Barley Genome Project was used for genetic mapping (Kleinhofs et al. 1993).

## Results

### Selection of BAC clones and sequencing

Genetic markers detecting a single major locus per genome were used to select BAC clones from different regions of the barley genome for sequencing. Importantly, these markers have homologous counterparts in genomes of other grasses that could be used for comparative genome analysis. Because these probes hybridized to one major band per haploid genome, the BAC clones isolated from different grass genomes should contain orthologous genome regions. The BAC clone with the probe near the center of the sequence, as determined by restriction mapping, was selected for sequencing in order to facilitate

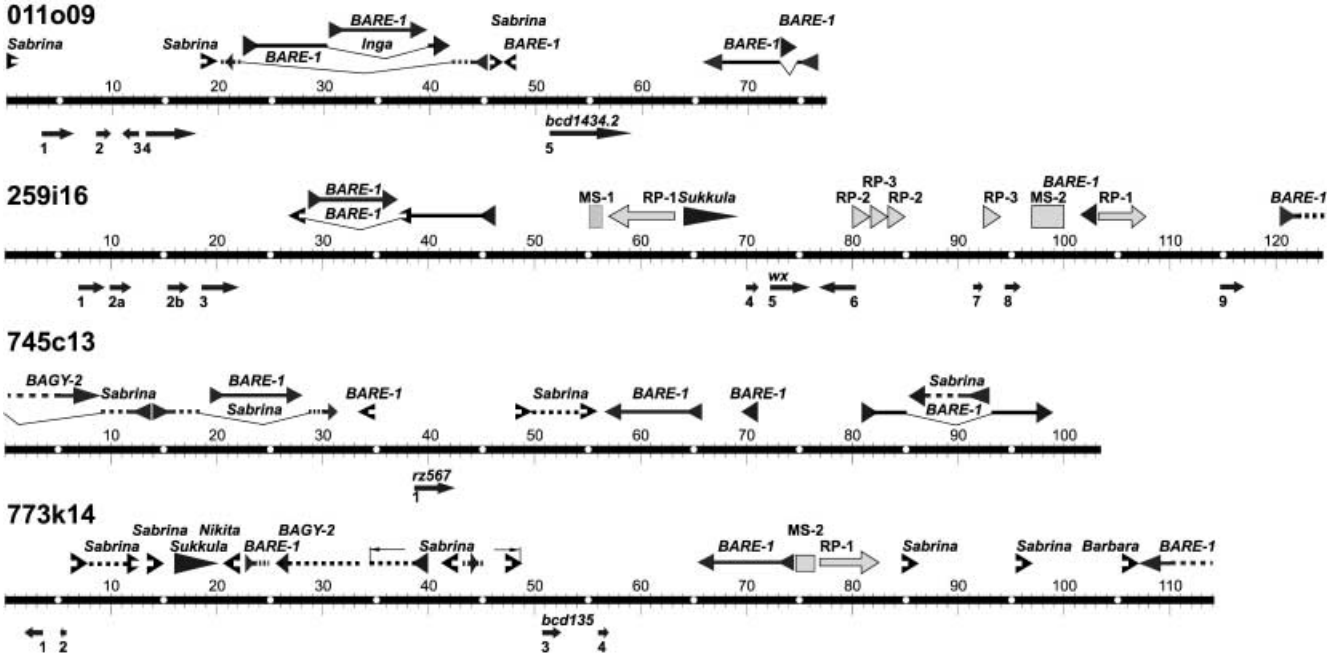


**Fig. 1** Predicted genes (*below BAC line*) and organization of repetitive elements (*above BAC line*) of four barley cv. Morex BAC clones. *Single-ended black arrows* predicted genes (exon / intron structure is not shown); *double-ended black arrow* full-length retrotransposon; *black arrowhead* LTR; *partially filled black arrowheads* partial LTRs; *black arrowheads on dotted lines* incomplete elements; *gray-shaded block* mini satellite block; *gray-shaded arrows and arrowheads* unidentified repeats

future comparisons with orthologous regions of other grass species. Sequencing and assembly was done as described in Dubcovsky et al. (2001) except that the assembly was done at around 10× sequence coverage. The predicted error rate in assembled sequence was lower than 0.1 in 100 kb based on PHRED quality score and sequence coverage. The assembly was checked by comparison with the pattern of restriction enzyme digestion. The GenBank accession numbers and general characteristics of the barley BAC sequences are given in Table 1.

The BAC clone designated 011o09 was identified with the cDNA probe BCD1434 encoding a protein homologous to the Mei-2-like protein BAA2237 from *A. thaliana* (Fig. 1; Tables 1 and 2). However, the 1.5 kb sequence of the probe matched the genomic sequence for only about 200 bp at 78% homology indicating that the BAC clone contained a paralog of the original BCD1434 probe. This region was designated BCD1434.2 and it is a part of gene 5 also with homology to the Mei-2 protein. Several searches of the 6.3× barley cv. Morex BAC library failed to identify a clone matching the original BCD1434 probe suggesting that it does not exist in the library. The genetic map location of the BCD1434.2 BAC clone 011o09 was not identified due to the lack of polymorphism.

## Analysis of gene regions

Predicted genes (Table 2) were numbered consecutively from the 5′ end of the BAC sequence with the exception of duplicated genes 2a and 2b from the BAC clone 259i16. Nine of the 20 predicted genes identified matching Triticeae (barley and wheat) ESTs indicating an expressed gene. Four of the nine were genes matching the cDNA probes used for selection of the BAC clones including the known function gene *waxy* (*Wx*; Rohde et al. 1988). Another four genes identified homologous, but not identical, ESTs from barley or other grass species. Five of the predicted genes were identified only by the gene-finder programs. Two of predicted genes had matches in the BLASTP search, but no similar ESTs. A putative function was identified for 8 of the 20 predicted genes by the BLASTP searches and 1, *Wx*, is known to encode a granule-bound glycogen-starch synthase (Rohde et al. 1988; Table 2).

The average gene density for all barley BAC sequences analyzed was 1 gene per 21 kb with a range from 1 per 12.4 kb for the 259i16 (*Wx*) BAC to 1 per 102.8 kb for the 745c13 (RZ567) BAC (Table 1). The gene content of the BAC clones varied from 22.9% for the 011o09 (BCD1434) BAC to 3.4% for the 745c13 BAC. Genes were located mostly in “gene islands” with an average gene density within the islands of about 1 gene per 4 kb. On the other hand, we observed regions ranging from 33.0 to 61.1 kb that lacked any genes.

The gene *Wx* from BAC 259i16 encodes a starch-granule-bound UDP glucose-starch glucosyltransferase (EC 2.4.1.11). There were three nucleotide differences in the coding region between our sequence from cv. Morex and the published sequence from cv. Vogelsanger Gold (Rohde et al. 1988; X07931). Only one of these would lead to an amino acid change of serine to arginine at po-

**Table 2** Predicted genes in the sequenced barley BAC clones

| Gene ID[a] | Gene boundaries / protein length / exon number | Predicted function[b] | Supporting evidence[c] |
|---|---|---|---|
| 011o09 BAC (77,136 bp) selected with BCD1434 probe | | | |
| Gene 1 (+) | 2,563–5,843 / 381 aa / 8 exons | Similar to *Arabidopsis thaliana* hypothetical protein T17F15.140 (198; E =6e–50) | ESTs AV932197, AV837000, FSH, GMK, GSN, BLASTP |
| Gene 2 (+) | 8,513–9,942 / 183 aa / 2 exons | Hypothetical protein | GMK, GSN |
| Gene 3 (–) | 11,490–12,037 / 101 aa / 2 exons | Hypothetical protein | GMK, GSN |
| Gene 4 (+) | 13,234–18,258 / 427 aa / 8 exons | Unknown protein | EST BG299549, FSH, GMK |
| Gene 5 (+) | 51,305–58,727 / 961 aa / 13 exons | Similar to *A. thaliana* Mei-2-like protein BAA22374 (577; E = e–163). Contains two pfam00076 RNA recognition motifs (51; E =3e–7 and 48; E =2e–6) | EST BE194208, FSH, GMK, GSN, BLASTP |
| 259i16 BAC (124,038 bp) selected with *Wx* probe | | | |
| Gene 1 (+) | 6,912–8,975 / 170 aa / 7 exons | Similar to *A. thaliana* putative leucine aminopeptidase encoded by AC005967 (160; E =5e–39). Contains pfam00883 peptidase M17 domain from cytosol aminopeptidase family (112; E =2e–26) | Similar barley EST AV915606 (238; E =5e–61), maize EST AI795470 (159; E =1e–35), FSH, GMK, GSN, BLASTP |
| Gene 2a (+) | 10,203–12,408 / 219 aa / 2 exons | Unknown protein | EST BI779618, FSH |
| Gene 2b (+) | 15,600–16,939 / 152 aa / 2 exons | Unknown protein | EST BI779618, FSH, GMK |
| Gene 3 (+) | 18,744–22,075 / 256 aa / 2 exons | Hypothetical protein | FSH, GMK |
| Gene 4 (+) | 70,375–70,971 / 198 aa / 1 exon | Hypothetical protein | FSH, GMK, GSN |
| Gene 5 (*wx1*) (+) | 72,092–75,664 / 603 aa / 12 exons | Granule bound glycogen-starch synthase precursor | Rohde et al. 1988, mRNA accession X07932 |
| Gene 6 (–) | 77,035–80,247 / 245 aa / 5 exons | Similar to *A. thaliana* unknown protein AF344326 (233; E =8e–61) | EST BE215794, FSH, GMK, GSN, BLASTP |
| Gene 7 (+) | 91,421–92,153 / 171 aa / 2 exons | Unknown protein | *Aegilops speltoides* EST BF292644 (170; E =4e–41), FSH, GMK, GSN |
| Gene 8 (+) | 94,451–96,002 / 216 aa / 2 exons | Unknown protein | Similar barley EST BG34392 6 (545; E =1e–152), FSH, GMK, GSN |
| Gene 9 (+) | 114,883–116,944 / 418 aa / 5 exons | Unknown protein | Wheat EST BE404465 (262; E =3e–68), FSH, GMK |
| 745c13 BAC (102,842 bp) selected with RZ567 probe | | | |
| Gene 1 (+) | 38,259–41,747 / 708 aa / 3 exons | Similar to *A. thaliana* protein BAB01483 containing similarity to kinesin light chain (686; E =0.0) | ESTs AW927148, BE230967, BE230963, FSH, GMK, GSN, BLASTP |
| 773k14 BAC (113,510 bp) selected with BCD135 probe | | | |
| Gene 1 (–) | 1,943–3,636 / 426 aa / 5 exons | Similar to *A. thaliana* putative receptor protein kinase NP_176010 (314; E =1e–84). Contains pfam00069 protein kinase domain (164; E =7e–42) | FSH, GMK, BLASTP |
| Gene 2 (–) | 6,312–6,743 / 143 aa / 1 exon | Hypothetical protein | FSH, GMK |
| Gene 3 (+) | 50,890–52,560 / 556 aa / 1 exon | Similar to *A. thaliana* putative protein NP_197560 (90; E =5e–17) | BCD135 cDNA, FSH, GMK, GSN, BLASTP |
| Gene 4 (+) | 56,119–56,703 / 194 aa / 1 exon | Similar to *A. thaliana* putative RING zinc finger protein NP_179593 (89; E =2e–17). Contains smart00184 RING finger domain (34; E =0.007) | FSH, GMK, GSN, BLASTP |

[a] Genes are numbered starting from the 5′ end of the DNA sequence as deposited in GenBank. Localization of predicted gene on plus or minus strand is indicated in parentheses

[b] Protein function is based on the BLASTP result with exception of *Wx* for which a reference is provided. BLASTP score and E value if available are given in parentheses

[c] Evidence supporting the existence of gene is as follows: *EST* – gene transcripts have been identified as EST sequences nearly identical to genomic sequence (only the GenBank accession given) or similar EST sequences (GenBank accession provided and the BLASTN score and E value given in parentheses); *FSH* – FGENESH prediction; *GSN* – GENSCAN prediction; *GMK* – GeneMark prediction

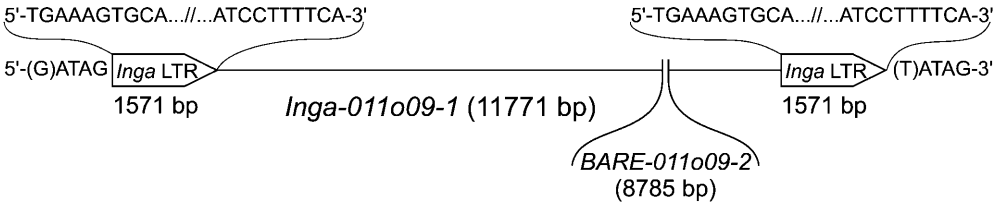


**Fig. 2** Structure of the novel barley LTR retrotransposon *Inga-011o09–1*. Inverted repeat sequences at the ends of the LTRs are indicated *above the LTR arrows*. The *Inga-011o09–1* insertion created 5-bp imperfect direct repeats. The 8,785-bp *BARE-011o09–2* element insertion is indicated *below*

sition 70 of the predicted amino acid sequence. There were nine other nucleotide changes in the 5′ and 3′ non-coding sequences, including a 4-bp insertion in the 5′ non-coding region plus 2-bp and 1-bp insertions in the 3′ non-coding region. Comparison with the 5.1-kb published genomic sequence revealed multiple single nucleotide polymorphisms (SNP) and several insertions/deletions in the non-transcribed regions and introns, as well as a microsatellite length polymorphism in the sixth intron. The Morex sequence had an $(AT)_{13}$ repeat, while the Vogelsanger Gold cultivar had $(AT)_9$.

The 259i16 BAC clone regions from 57.5 to 62.9 kb and 101.1 to 108.2 kb represented highly homologous (ca. 97% identity) inverted repeats. All three gene prediction programs indicated these regions as putative genes. The deduced protein sequence showed high homology to a putative TNP-like transposable element from rice (AC079852). To gain more information, we designed probes from the predicted exons (104,455–105,673 bp and 107,431–107,665 bp) and performed gel blot hybridization to barley genomic DNA filters. Both probes hybridized to multiple bands. Strong hybridization was detected after a short (30 min) exposure to X-ray film indicating that the sequences homologous to the probes were abundant in the barley genome and therefore represent a repetitive element. This was supported by finding a homologous region in the sequence of the unrelated BAC clone 773k14 (BCD135). This element did not exhibit characteristics of known barley repetitive elements such as LTRs or regions homologous to a retrotransposon polyprotein, therefore it was designated as unknown barley repeat RP-1 (Fig. 1).

## Analysis of repetitive elements

LTR retrotransposons, consisting of full-length elements, nested retrotransposons and solo LTRs, accounted for 167 kb or 40% of the DNA sequenced (Fig. 1; Table 1). In individual BAC clones the range was from 21% for the 259i16 BAC to 57% for the 745c13 BAC. This retroelement density was in inverse proportion to the gene space predicted for these BAC clones. Based on homology with the LTRs, most of the retroelements were of the *BARE-1* type. However, other elements with homology to *BAGY-2*, *Sabrina*, *Nikita* and *Sukkula* were also identified (Fig. 1).

The polyprotein region of one retrotransposon in BAC clone 011o09 exhibited highest homology (BLASTX score =2,078; E =0.0) to the maize *Opie-2* LTR retrotransposon polyprotein region (*Ty1-copia* group; accession number – T04112), but its LTRs showed no homology to any known retroelements in BLASTN, therefore it was annotated as a novel LTR retrotransposon *Inga-011o09–1*. Both LTR sequences were 1,571 bp long and each was flanked by conserved sequences (Fig. 2). It contained an insertion of a full-length *BARE-011o09–2* element between the polyprotein gene and the 3′ LTR. Excluding the insertion of the 8,785-bp *BARE-011o09–2* element the total length of the *Inga-011o09–1* element was 11,771 bp. This *Inga-011o09–1*element is probably non-functional because of the *BARE-011o09–2* element insertion and several stop codons in the polyprotein gene. A BLASTX search using the sequence between the two LTRs (without *BARE-011o09–2*) confirmed that the *Inga-011o09–1* element belongs to the *Ty1-copia* group of LTR retrotransposons because of the conserved order of aspartic protease, integrase, reverse transcriptase and RNaseH. Homology to the *Opie-2* polyprotein started from position 26,131 in the BAC sequence and extended to position 29,425, covering the entire length of the *Opie-2* polyprotein. The region corresponding to the putative *gag* gene in the *Inga-011o09–1*element showed a 318-aa ORF (from 23,521 to 24,474 bp) homologous to the maize *Opie-2 gag* protein. The 2,675-bp sequence located between the putative polyprotein and the 3′ LTR did not show significant homology in GenBank searches. In other plant retrotransposons, it has been argued that this region might encode an *env* function indicative of a retroviral function and/or origin of this element (reviewed in Kumar and Bennetzen 1999, 2000).

Other repeated elements were homologous to non-LTR retrotransposon sequences, although full-length non-LTR retrotransposons were not found. In addition to the RP-1 element (discussed in the Analysis of gene regions section), two other putative repeats, RP-2 and RP-3, were identified by two-dimensional dot-plot analysis. Three minisatellite regions were also identified (Fig. 1) as well as microsatellites with eight or more repeat units (see GenBank entries for details). Repetitive elements other than LTR retrotransposons made up approximately 11.5% of the total sequence (Table 1). Ribosomal RNA or transfer RNA sequences were not found on the sequenced BAC clones. The rest of the BAC sequences (37.1%) could not be assigned a function.

## Discussion

We analyzed 417.5 kb of barley genomic sequences and identified 20 putative genes with an average density of 1 per 21 kb. The gene density is in agreement with the 1 gene per 20 kb estimated in previous studies analyzing a total of 228 kb of barley genomic sequences (Panstruga et al. 1998; Shirasu et al. 2000; Dubcovsky et al. 2001). The value of 1 gene per 21 kb is 10 times higher than the 1 gene per 200 kb genome average calculated for the 5,000 Mb barley genome and assuming a similar number of genes (25,498) as predicted for the *A. thaliana* genome (Arumuganathan and Earle 1991; The Arabidopsis Genome Initiative 2000). The obvious bias in previous barley genome sequencing studies, including ours, was that the genomic regions to be sequenced were selected using genes as probes, thus insuring that at least one gene was present. Kunzel et al. (2000) subdivided the barley genome into high (less than 1 Mb/cM), medium (1–4 Mb/cM), and low (greater than 4 Mb/cM) recombination regions based on a physical map developed using translocation breakpoints and PCR mapping of low-copy-number probes on isolated chromosomes. On the hypothesis that recombination rates are related to gene density, we analyzed the genomic location of the sequenced BAC clones in relation to the Kunzel et al. (2000) predicted recombination regions. The *bcd135* locus mapped to a high recombination region and the corresponding BAC clone contained four genes. The *Wx* locus mapped to a medium recombination region and the *Wx* BAC clone had ten predicted genes. On the other hand, the *rz567* locus mapped between medium and low recombination regions and the BAC clone had no other predicted genes besides the *rz567* homologue itself. Unfortunately, the *bcd1434.2* locus was not mapped and thus could not be positioned on the physical map. The previously published barley gene density value of 1 per 20 kb was also derived from BAC clones originating from the high and medium recombination regions of the barley genome. The number of BAC clones sequenced is still small and no randomly selected BAC clones have been sequenced, but there appears to be a general correspondence between the high and medium recombination and gene-rich regions in the barley genome.

Even within the putative gene-rich regions (excluding the RZ567 BAC clone), variation in gene densities ranged from 1 per 12.4 kb to 1 per 28.5 kb suggesting an uneven distribution. Assuming an overall average of 1 gene per 20 kb in the gene-rich regions and approximately 25,000 genes (based on *A. thaliana*), the gene-rich regions would occupy approximately 10% of the barley genome. This is in fair agreement with the 12% gene-rich barley genome portion calculated by Barakat et al. (1997). The RZ567 BAC clone with a single gene per 103 kb suggested that there are genes in regions with much lower overall gene density and that there must be very large regions with few if any genes. A 140-kb region devoid of genes was recently reported in a 211-kb *T. monococcum* genome sequence (Wicker et al. 2001). On the other hand, 47 apparently expressed genes were found in the centromeric regions of *A. thaliana* encoding different types of proteins (The Arabidopsis Genome Initiative 2000).

Even though the average gene density was 1 per 21 kb, we found that genes within BAC clones were not distributed evenly. In most cases the genes were clustered with gene density as high as 1 per 4 kb. We also observed apparently isolated single genes, e.g. gene 5 on the 011o09 BAC, gene 9 on the 259i16 BAC and the RZ567 homologue on the 745c13 BAC (Fig. 1).

Studies of the structure of predicted genes revealed that the average exon number per gene was 4.2, a number that is lower than the 5.2 exons per gene predicted for the *A. thaliana* genome (The Arabidopsis Genome Initiative 2000). The five genes encoding hypothetical proteins averaged only 1.6 exons as opposed to an average of 5.4 exons for genes having matching or similar ESTs. Hence, it is likely that we have missed exons for some hypothetical genes (especially exons that do not encode a part of the final peptide). The average exon size was 244 bp, similar to the 251-bp average reported for *A. thaliana*. The average intron size in predicted genes was 398 bp, range 72 to 2,562(Table 3), significantly larger than the 170 bp for the *A. thaliana* genome (The Arabidopsis Genome Initiative 2000). This difference in intron size may be due to the gene prediction programs that may ignore very small exons and thus increase the apparent intron size. However, it has been observed that small genome species like *A. thaliana* have somewhat smaller introns than large genome species (Dubcovsky et al. 2001).

The EST sequences proved to be very useful for identification of genes and also to determine their intron/exon structure. About a half of the predicted genes (9 out of 20) were likely to be transcribed as evidenced by matching nearly identical Triticeae ESTs (based on 148,600 barley + 78,900 wheat as of March 1, 2002). If similar ESTs from barley and from other grass species were included, the number of predicted genes with ESTs increased to 13 out of 20 (65%). This is close to the approximately 60% of genes with ESTs reported for *A. thaliana* (The Arabidopsis Genome Initiative 2000). A probable or known function could be assigned to 9 out of the 20 predicted genes based on BLASTP homology and literature data (see Table 2 for predicted function and homology scores). The average protein length showed significant variation ranging from 708 aa for the 745c13 BAC to 265 aa for the 259i16 BAC, although the value for the 745c13 BAC was from a single gene. The average predicted protein length was 336 aa (Table 3).

The sequences with no known function constituted 37.1% of the 417.5 kb of sequences analyzed. Most of these sequences could be unidentified repetitive elements that did not have any homologues in GenBank, while some could be DNA sequences surrounding the coding sequences providing for 5′- and 3′-untranslated regions and regulatory elements (Table 1).

The average GC content of the barley BAC clone sequences (44.5%; Table 1) fitted within the range of the

**Table 3** Characteristics of the sequenced BAC clone gene regions

| BAC clone (length in bp) | 011o09 (77,136) | 259i16 (124,038) | 745c13 (102,842) | 773k14 (113,510) | Average values |
|---|---|---|---|---|---|
| Average gene length in bp | 3,533 | 2,071 | 3,494 | 1,107 | 2,315 |
| Average protein length in aa | 411 | 265 | 708 | 330 | 336 |
| Average exon number | 6.6 | 4.0 | 3.0 | 2.0 | 4.2 |
| Exon size: average / minimum / maximum (in bp) | 186 / 14 / 636 | 199 / 31 / 596 | 804 / 200 / 1,998 | 495 / 142 / 1,670 | 244 |
| Intron size: average / minimum / maximum (in bp) | 412 / 72 / 1815 | 424 / 77 / 2,562 | 423 / 231 / 852 | 104 / 89 / 111 | 398 |

barley average base composition (42.8–45.6%; Melzer and Kleinhofs 1987). It was also very close to the value of another barley BAC clone 635P2 (44.2%; Dubcovsky et al. 2001), but higher than the range determined for the barley "gene space" (42.6–43.4%; calculated from CsCl buoyant density in Barakat et al. 1997 using the formula from Mandel et al. 1968). The BAC clones 259i16 (ten genes; 16.7% gene space) and 745c13 (one gene; 3.4% gene space) had the same GC content (44.9%) indicating that the average base composition in these genome regions is not indicative of gene content.

Analysis of the largest barley genomic DNA sequence to date revealed the presence of gene islands and single genes separated by large regions of nested LTR retrotransposons. These results are in agreement with previous studies suggesting that this arrangement may be a common feature of the barley genome. Analysis of gene density and recombination frequencies supported the theory that high-recombination regions constitute gene-rich regions. High recombination regions constitute a relatively small portion of the barley genome and it may be possible to manipulate them efficiently for map-based cloning of agriculturally important genes.

## Note added in proof

After submitting the manuscript, we mapped a single-copy probe from 011o09 BAC clone gene 5 to the barley chromosome 2(2H) bin 008 locus *bcd1434.2*. Mapping was with the Oregon Wolf Barley Dominant by Recessive doubled-haploid population (Costa et al. 2001). An attempt to relate the *bcd1434.2* locus to the Kunzel et al. (2000) translocation breakpoint physical map was complicated due to the lack of common probes between the mapping populations. Thus, precise positioning within the region with defined recombination rate was not possible, although more distant probes suggested that 011o09 BAC clone could be located within the high-recombination region, in line with the high gene density found in the BAC clone.

**Acknowledgements** This is Scientific Paper No. 0212–11 from the College of Agriculture and Home Economics Research Center, Washington State University, Pullman, WA; project 0196. Research was supported by NSF Grant No. DBI-9975796.

## References

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 25:208–218

Ayliffe MA, Collins NC, Ellis JG, Pryor A (2000) The maize *rp1* rust resistance gene identifies homologues in barley that have been subjected to diversifying selection. Theor Appl Genet 100:1144–1154

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

Becker J, Vos P, Kuiper M, Salamini F, Heun M (1995) Combined mapping of AFLP and RFLP markers in barley. Mol Gen Genet 249:65–73

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

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

Dooner HK, Martinez-Ferez IM (1997) Recombination occurs uniformly within the *bronze* gene, a meiotic recombination hotspot in the maize genome. Plant Cell 9:1633–1646

Dooner HK, Weck E, Adams S, Ralston E, Favreau M, English J (1985) A molecular genetic analysis of insertion mutations in the *bronze* locus in maize. Mol Gen Genet 200:240–246

Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol 125:1342–1353

Eggleston WB, Allerman M, Kermicle JL (1995) Molecular organization and germinal instability of *R-stippled* maize. Genetics 141:347–360

Faris JD, Haen KM, Gill BS (2000) Saturation mapping of a gene-rich recombination hot spot region in wheat. Genetics 154:823–835

Fu H, Park W, Yan X, Zheng Z, Shen B, Dooner HK (2001) The highly recombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome. Proc Natl Acad Sci USA 98: 8903–8908

Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99:1082–1087

Gill KS, Gill BS, Endo TR, Boyko EV (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics 143:1001–1012

Gill KS, Gill BS, Endo TR, Taylor T (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics 144:1883–1891

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

Heun M, Kennedy AE, Anderson JA, Lapitan NLV, Sorrells ME, Tanksley SD (1991) Construction of a restriction fragment length polymorphism map for barley (*Hordeum vulgare*). Genome 34:437–447

Jaaskelainen M, Mykkanen AH, Arna T, Vicient CM, Suoniemi A, Kalendar R, Savilahti H, Schulman AH (1999) Retrotransposon *BARE-1*: expression of encoded proteins and formation of virus-like particles in barley cells. Plant J 20:413–422

Kleinhofs A, Graner A (2001) An integrated map of the barley genome. In: Phillips RL, Vasil IK (eds) DNA-based markers in plants, 2nd edn. Kluwer, Dordrecht, pp 187–199

Kleinhofs A, Kilian A, Saghai Maroof MA, Biyashev RM, 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 the barley (*Hordeum vulgare*) genome. Theor Appl Genet 86:705–712

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Kumar A, Bennetzen JL (2000) Retrotransposons: central players in the structure, evolution and function of plant genomes. Trends Plant Sci 5:509–510

Kunzel 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

Lichten M, Goldman AS (1995) Meiotic recombination hotspots. Annu Rev Genet 29:445–476

Lowe T, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Mandel M, Schildkraut CL, Marmur J (1968) Use of CsCl density gradient analysis for determining the guanine plus cytosine content of DNA. In: Grossman L, Moldave K (eds) Methods in enzymology, 12B. Academic Press, New York, pp 184–195

Manninen I, Schulman A (1993) *BARE-1*, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22:829–846

Mano Y, Kawasaki S, Takaiwa F, Komatsuda T (2001) Construction of a genetic map of barley (*Hordeum vulgare* L.) cross 'Azumamugi' × 'Kanto Nakate Gold' using a simple and efficient amplified fragment-length polymorphism system. Genome 44:284–292

Melzer JM, Kleinhofs A (1987) Molecular genetics of barley. In: Yasuda S, Konishi T (eds) Proceedings of the fifth international barley genetics symposium. Sanyo Press, Okayama, pp 481–491

Okagaki RJ, Weil CF (1997) Analysis of recombination sites within the maize *waxy* locus. Genetics 147:815–821

Panstruga R, Buschges P, Schulze-Lefert P (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res 26:1056–1062

Patterson GI, Kubo KM, Shroyer T, Chandler VL (1995) Sequences required for paramutation of the maize *b* gene map to a region containing the promoter and upstream sequences. Genetics 140:1389–1406

Puchta H, Hohn B (1996) From centiMorgans to base pairs: homologous recombination in plants. Trends Plant Sci 1:340–348

Qi X, Stam P, Lindhout P (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. Theor Appl Genet 96:376–384

Ramsay L, Macaulay M, degli Ivanissevich S, 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

Rohde W, Becker D, Salamini F (1988) Structural analysis of the *waxy* locus from *Hordeum vulgare*. Nucleic Acids Res 16: 7185–7186

Rostoks N, Zale JM, Soule J, Brueggeman R, Druka A, Kudrna D, Steffenson B, Kleinhofs A (in press) A barley gene family homologous to the maize rust resistance gene *Rp1-D*. Theor Appl Genet

Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803-811

Sandhu D, Champoux JA, Bondareva SN, Gill KS (2001) Identification and physical localization of useful genes and markers to a major gene-rich region on wheat group 1S chromosomes. Genetics 157:1735–1747

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:737–738

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schnable PS, Hsia AP, Nikolau BJ (1998) Genetic recombination in plants. Curr Opin Plant Biol 1:123–129

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10:908–915

Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-G10

Suoniemi A, Anamthawat-Jonsson K, Arna T, Schulman AH (1996) Retrotransposon *BARE-1* is a major, dispersed component of the barley (*Hordeum vulgare* L.) genome. Plant Mol Biol 30:1321–1329

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Thuriaux P (1977) Is recombination confined to structural genes on the eukaryotic genome? Nature 268:460–462

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Van Deynze AE, Sorrells ME, Park WD, Ayres NM, Fu H, Cartinhour SW, Paul E, McCouch SR (1998) Anchor probes for comparative mapping of grass genera. Theor Appl Genet 97:356–369

Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125:1283–1292

Waugh R, Bonar N, Baird E, Thomas B, Graner A, Hayes P, Powell W (1997) Homology of AFLP products in three mapping populations of barley. Mol Gen Genet 255: 311–321

Wei F, Gobelman-Werner K, Morroll SM, Kurth J, Mao L, Wing R, Leister D, Schulze-Lefert P, Wise RP (1999) The *Mla* (powdery mildew) resistance cluster is associated with three NBS-LRR gene families and suppressed recombination within a 240-kb DNA interval on chromosome 5S (1HS) of barley. Genetics 153:1929–1948

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316

Xu X, Hsia AP, Zhang L, Nikolau BJ, Schnable PS (1995) Meiotic recombination break points resolve at high rates at the 5′ end of a maize coding sequence. Plant Cell 7:2151–2161

Yu Y, Tomkins JP, Waugh R, Frisch DA, Kudrna D, 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