Theor Appl Genet (2002) 104:97–103

W. Michalek · W. Weschke · K.-P. Pleissner
A. Graner

EST analysis in barley defines a unigene set comprising 4,000 genes

Received: 15 March 2001 / Accepted: 18 April 2001

Abstract We report the generation of 13,109 EST (Expressed Sequence Tag) sequences from barley as a first step towards the generation of a unigene set for this organism. Sequences were generated from three libraries encompassing 7,568 cDNA clones. Comparisons to nucleic acid and protein sequence databases enabled the assignment of putative functions to the mRNAs. The results of the searches against protein databases were parsed and built into a regularly updated database, available over the World Wide Web. The Stack_Pack clustering system has been applied to survey the level of redundancy, which was calculated to amount to 69%, thus we identified 4,000 different barley genes. To prove the usability of the results of the clustering process for further experiments, we subjected alignments with sequences similar to elongation factor 1 alpha to additional analysis. These sequences represented the largest group with identical putative functions (228 members) and clustering based on the analysis of 3´ sequences subdivided the group into five different assemblies. Alignments of the consensus sequences facilitated the development of PCR assays suitable for genetic mapping of four of the different gene-family members, which reside on chromosomes 2H, 4H and 5H, thus demonstrating the suitability of the cluster-results as a basis for in-depth analyses of barley gene families.

Keywords Barley · EST

Communicated by G. Wenzel

W. Michalek (✉) · W. Weschke · K.-P. Pleissner · A. Graner
Institute of Plant Genetics and Crop Plant Research (IPK),
Corrensstrasse 3, 06466 Gatersleben, Germany
e-mail: michalek@ipk-gatersleben.de;
Tel.:+49-39482-5522, Fax: +49-39482-5595

Present address:
W. Michalek, PLANTA GmbH, Grimsehlstr. 31, 37555 Einbeck, Germany
email: w.michalek@kws.de
Tel.: +49-5561-311723, Fax: +49-5561-311243

Introduction

Since EST-sequencing was introduced for the analysis of the human genome the generation of single pass, partial sequences from cDNA clones has become an extensively used strategy for gene discovery and mapping in a wide range of organisms. The availability of rapidly growing sequence databases allows the detection of regions showing sequence similarity in functionally related gene products from distantly related organisms. Thus, it is increasingly possible to assign putative functions for a large proportion of anonymous cDNA clones. For many plants, which are important for human nutrition, comprehensive sets of EST sequences are already available (http://www.ncbi.nlm.nih.gov). These have proved useful in a number of ways: EST clones were used extensively as molecular markers for the construction of high-density genetic linkage maps of rice and maize (Harushima et al. 1998; Davis et al. 1999) and for a physical map of rice (Kurata et al.1997). Furthermore the sequencing data can be used to study gene families (Cooke et al. 1997; Epple et al. 1997) and they form a basis for SNP development (Cho et al. 1999). Apart from applications in the field of genetic and physical mapping, ESTs are the central resource for the analysis of gene expression with the help of high-density arrays, as demonstrated for *Arabidopsis* (Schena et al. 1995; Girke et al. 2000; Schenk et al. 2000).

Barley (*Hordeum vulgare* L.) is an important cereal species grown in temperate climates and ranks no. 4 in world crop production. In the recent past 68,658 EST sequences accumulated in dbEST (release 011901) emerging from several sequencing efforts. Taking into account that less than 1,000 protein sequences for the genus *Hordeum* are stored in the databases, the potential role for gene discovery together with the long-term value for genome analysis within the agriculturally important grasses is obvious. Here we describe a set of 13,109 barley EST sequences, their organization into clusters to generate a first “unigene”-set of barley and the integration of these data into a database.

Materials and methods

cDNA libraries

cDNA libraries were constructed with tissue from the barley variety 'Barke', a spring barley cultivar, which is used for malting. For this sequencing project three libraries were constructed, one from etiolated seedlings (HK), one from developing caryopsis (HY) and one from roots (HW). For the HK library, plants were grown on wet filterpaper for 6 days at 25°C in the dark, and cDNA was made with the Superscript system (GIBCO) according to the manufacturer's instructions. The cDNA was ligated directionally (*Sal*I/*Not*I) into pBluescriptSK vector and transformed into XL1Blue cells (Stratagene). For the HY library, mRNA was made from developing caryopsis (1–15 days after flowering), and for the HW library roots were grown for 2 days on filter paper at room temperature. Both libraries were constructed in λ-ZAP Express (Stratagene) according to the instructions of the manufacturer. After in vivo excision the cDNA inserts were present directionally (*Eco*RI/*Xho*I) within the vector pBK-CMV in *Escherichia coli* XLOLR cells.

Sequencing

Plasmid DNA was isolated with the 96-well Turbo Plasmid prep system (Qiagen, Hilden) on a Qiagen robot. The quality of the plasmid preparation was controlled on agarose gels and used directly for sequencing. Sequencing was performed on ABI377XL sequencers using dRhod- (mainly for 3′-ends) and BigDye-terminator chemistry (mainly for 5′-ends) respectively (ABI Perkin Elmer, Weiterstadt).

Data processing

Vector sequences and sequence ends were trimmed immediately after sequencing using Sequencher (v.3.1) Software (Gene Codes, Ann Arbor). Sequences were trimmed from the 5′- and 3′-ends until a 50-bp window contains less than two ambiguities. The maximum length was set to 700 bp. In a second step, CrossMatch (http://www.genome.washington.edu/uwgc/analysistools/swat.htm) was used to detect remaining vector artifacts. Only sequences longer than 100 bp after this process were included in the dataset.

Prior to clustering, polyT and polyA stretches were removed. Redundancy was analyzed with the Stack_Pack clustering system using an alpha version under Solaris and version 2.0 under Linux, respectively (Miller et al. 1999). The default parameters were used unless otherwise stated. The process includes subsequent steps of masking, clustering, assembly, alignment analysis, and consensus partitioning. The masking step employed CrossMatch to mask remaining vector artifacts and repetitive sequences as simple repeats and plant sequences from RepBase5.02 (Jurka 1998). Until the last step EST sequences (3′ and 5′) were handled individually, and finally the package used clone information (shared clone ID of two EST sequences) to join clusters and to produce so-called "clonelinks".

For clone linking we used the parameter LINK=1, to link clusters which share at least one clone ID. Since singleton sequences were not linked by the program, for the calculation of the number of genes we performed this step with Excel spreadsheet operations.

Sequence-similarity searches against databases were conducted with HUSAR (Heidelberg Unix Sequence Analysis Resource) using BlastX2 and BlastN2 from release 2.0.9 of the Blast2 (www.ncbi.nlm.nih.gov/; Altschul et al. 1990) suite of programs. These programs use filtering tools by default (SEG for BlastX2 and DUST for BlastN2). Searches were performed using the default parameters. Sequence-similarity searches against sets of *Arabidopsis*, maize, and rice ESTs were carried out locally. EST datasets were downloaded from public databases and the WU-Blast 2.0 programs were employed (W. Gish 1996, unpublished) using the default values with no filtering. Multiple alignments were performed with ClustalX (Thompson et al. 1997).

Genetic mapping

Specific primers were designed manually based on alignment information (see Fig. 2). Standard PCR reactions were performed at 60°C annealing temperature, for 35 cycles. Products from all parents were sequenced directly using the PCR primers after primer removal with QiaQuick columns and ABI Dye Terminator sequencing chemistry.

Genetic mapping was attempted on three mapping populations. 'Igri'×'Franka' (Graner et al. 1993), 'Steptoe'×'Morex' (Kleinhofs et al. 1993) and the Oregon-Wolfe barley stock (http://www.css.orst.edu/barley/WOLFEBAR/WOLFNEW.HTM).

Results

cDNA libraries and sequencing

Based on restriction digestion of plasmids, or PCR amplification of insert sequences, the average insert size of the libraries is in the range between 1,000 and 1,100 bp (178 clones tested). The average sequence length is 550 bp. For most of the clones 5′- and 3′-sequencing was attempted, resulting in 13,109 sequences, which have been submitted to EMBL (AL450491–450982, AL499630–500520, AL500542–512267). The overall sequencing success rate was approximately 80%, calculated after removal of poor quality and vector sequences. This dataset is based on 7,568 different cDNA clones and contains approximately equal numbers of 5′- and 3′-sequence data (Table 1).

To estimate the frequency of inverted clones the set of 5′-sequences was visually inspected for the appearance of poly-T sequences. Within 6,505 sequences we found 381 putative 3′-sequences (5.9%). Since the sequence identifiers indicate that the inverted clones are randomly distributed throughout the dataset, wrong assignments of complete 96-well sequencing runs (or reactions) can be excluded. Comparing the Blast hits from 300 clones, which were sequenced from both ends, we estimate that 3% of the clones are chimaeric. At this stage of the analysis we assume the clones of both categories to represent cloning artifacts.

Comparison to databases

Comparisons to amino acid- and nucleotide-sequence databases

To estimate the proportion of clones from the mitochondrial and chloroplast genomes, Blast searches were performed with the complete chloroplast sequence from rice (emb: X15901.1) and the mitochondrial sequence from

Table 1 Summary of sequencing results

Library	Number of sequences		Σ
	3′	5′	
HK	891	492	1,383
HW	2,643	2,764	5,407
HY	3,070	3,249	6,319
Σ	6,604	6,505	13,109

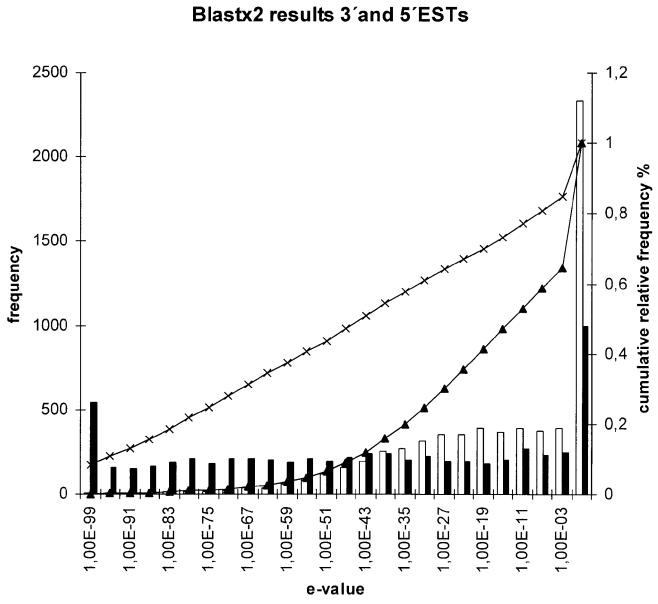


Fig. 1 Quality of BlastX hits obtained after comparison of 3′ and 5′-sequences to SWISSPIRPLUS. Black bars: frequency of hits with 5′-sequences; *open bars:* frequency of hits with 3′-sequences; *line+cross:* cumulative frequency of hits with 5′-sequences; *line+triangle:* cumulative frequency of hits with 3′-sequences

Arabidopsis (emb: Y08501.1; Y08502.1), respectively. The proportion of clones with similarity to mitochondrial sequences was 0.6%; similarities to chloroplast sequences were detected with 1.1% of the clones (WU-Blast 2.0, *p*-value <1.0 E–05). In both cases most of these clones (80%) originated from the caryopsis library.

According to the number of hits to *E. coli* sequences in the public databases the amount of clones with inserts from microorganisms was calculated to be <1%. Since thresholds for "contaminating" sequences are a critical issue and due to the small proportion of this kind of data in the complete set, these sequences were not omitted from the submission to dbEST.

Using the BlastX2 program, all sequences were compared to SWISSPIRPLUS a database containing entries from SWISSPROT, PIR (which are not in SWISSPROT) and translated EMBL (SP-TREMBL, for details see: http://genome.dkfz-heidelberg.de/). The quality of the hits is illustrated in Fig. 1. Not unexpectedly, 3′-ends give lower e-values than 5′-sequences and the percentage of hits among the non-redundant 5′-sequences is 61% (assuming a BlastX e-value <1E-20 as a hit).

To make the search results available to the scientific community we established a SQL database, which contains the EST-identifier, descriptions of the hits, the scores, and e-values. BlastX results are updated on a regular scale (http://www.ipk-gatersleben.de).

Comparisons to other plant EST-sequences

To address the question, which proportion of the barley EST sequences is already represented by EST sets from other plants, WU-BlastN searches were conducted against *Arabidopsis*, maize, and rice ESTs (46,706, 42,391, and 45,157 entries). The analysis was performed with the non-redundant set of 5′-sequences to exclude effects due to multiple sequences of the same gene. As expected, more hits and higher score values were obtained comparing the barley sequences to the rice and maize ESTs as to *Arabidopsis* ESTs. However, 30% (1,054) of the non-redundant sequences certainly had no hit (p-value >1E-05) against the rice dataset. This value decreased to 23% if all three EST sets (rice, maize, Arabidopsis) were taken into account. To avoid bias, possibly introduced by different codon usage, we analyzed the data with TBlastX and the value changed to 76%, but only due to weak similarities with *p*-values between 1E-05 and 1E-10. No difference in the number of hits generated by nucleotide or amino-acid comparisons was observed, if hits with a *p*-value <1E-10 were taken into consideration.

For the analysis of the relationship between already known *Hordeum* sequences within the public databases and barley ESTs, all known protein sequences (Query: organism=*Hordeum*) were downloaded from SWALL (http://srs6.ebi.ac.uk). This dataset, containing 750 entries, was used to perform WU-BlastX searches against the consensi, and singletons from the non-redundant set of 5′-EST sequences (3,518 consensus and singleton sequences). Assuming, that hits with a *p*-value >1E–05 certainly do not represent significant similarities, 69% (2,433) of this set of sequences represent genes yet unknown for the genus *Hordeum*. From the 750 *Hordeum* protein sequences only a subset of 231 was hit with EST sequences at a *p*-value <1E–20 (assuming this is a significant degree of similarity).

Even though only a small number of libraries was used, the EST data cover a major part of the known barley genes. In this context a nearly five-fold coverage of each hit with putative "non-redundant" sequences was observed. Apart from sequencing artifacts, accounting for approximately 10% of this effect, this may be explained by two considerations. (1) Different 5′-sequences may derive from several truncated cDNA clones belonging to the same gene. (2) The program package used to determine redundancy (see below) is more sensitive to differentiate members of gene families than the Blast algorithm with default settings, with the score or e-value as the only criterion for identity. One example are the sequences HY01A03T and HY07K17V, belonging to different clusters, but coding for the same protein (beta-amylase). When these sequences were "blasted" against each other, an e-value of 0.0 was calculated on the basis of 86% identical bases over 664 bp, but closer inspection revealed two different genes (data not shown).

Redundancy within the EST collection

Redundancy analysis, using the Stack_Pack software package, linked clusters to produce 467 so-called clone-

links, 1,210 clusters remained unlinked (no shared clone IDs) and the analysis of the singletons revealed 2,323 different clones. The average length of the consensus sequences is 760 bp. In summary, we estimate to have identified 4,000 genes. The level of redundancy based on cDNA clones as it was calculated from the data generated by Stack_Pack is 69%. From the results of the clustering library specific clusters could be identified, i.e. clusters containing sequences of one library only (Table 2). Only clusters with ≥5 sequences were considered for this analysis. Apart from genes with known tissue-specific expression, for example sucrose synthase or thionins (which are expressed in developing caryopsis and etiolated leaves, respectively) up to 22% of the library specific clusters gave no hit against protein databases (assuming a BlastX e-value <E-20 as a hit).

Table 2 Library specific clusters

Library	Etiolated leaves	Roots	Caryopsis
No. of specific clusters[a]	11	49	77
No. of sequences	91	329	575
No. of no-hit clusters	1 (9%)	11 (22%)	12 (16%)

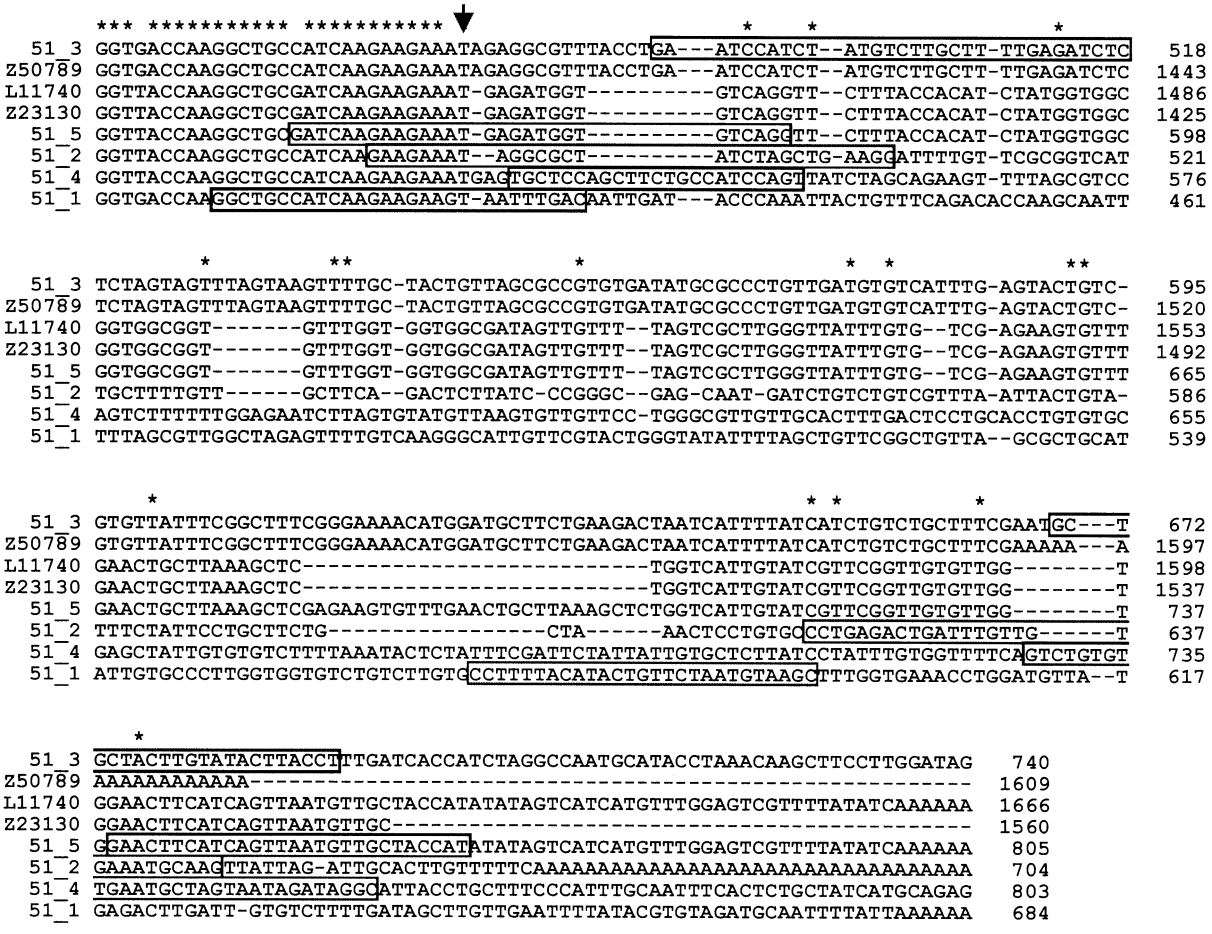
[a] Only clusters ≥5 members are considered

When these criteria were applied to the complete dataset, 85% of the groups with ≥5 sequences contained a sequence encoding a known protein, whereas only 44% of singletons or sequences in the smaller groups yielded BlastX hits. This result points to the fact that abundantly expressed genes tend to be more likely represented in databases.

Analysis of the eEF1A gene family

As an example for the utilization of the result of the clustering process, we analyzed the most-abundant class of transcripts in this study, transcription elongation factor 1-alpha (eEF1A). Overall, the BlastX2 analysis identified 228 sequences to be similar to eEF1A, derived from 143 cDNA clones. The Stack_Pack analysis resulted in five clusters (51_1 to 51_5) comprising 275 DNA-sequences. The formation of individual clusters was based on 103 3′ sequences (51_1, 4; 51_2, 16; 51_3, 20; 51_4, 39; 51_5, 24); all possible artifacts as well as a so-called secondary consensus in cluster 51_5 were excluded.

Fig. 2 Multiple alignment (ClustalX) of consensus sequences from clusters with 3′ sequences together with published sequences (EMBL Acc.: L11740, Z23130, Z50789). *Boxes* mark the regions of primer design. The *arrow* indicates the stop codon



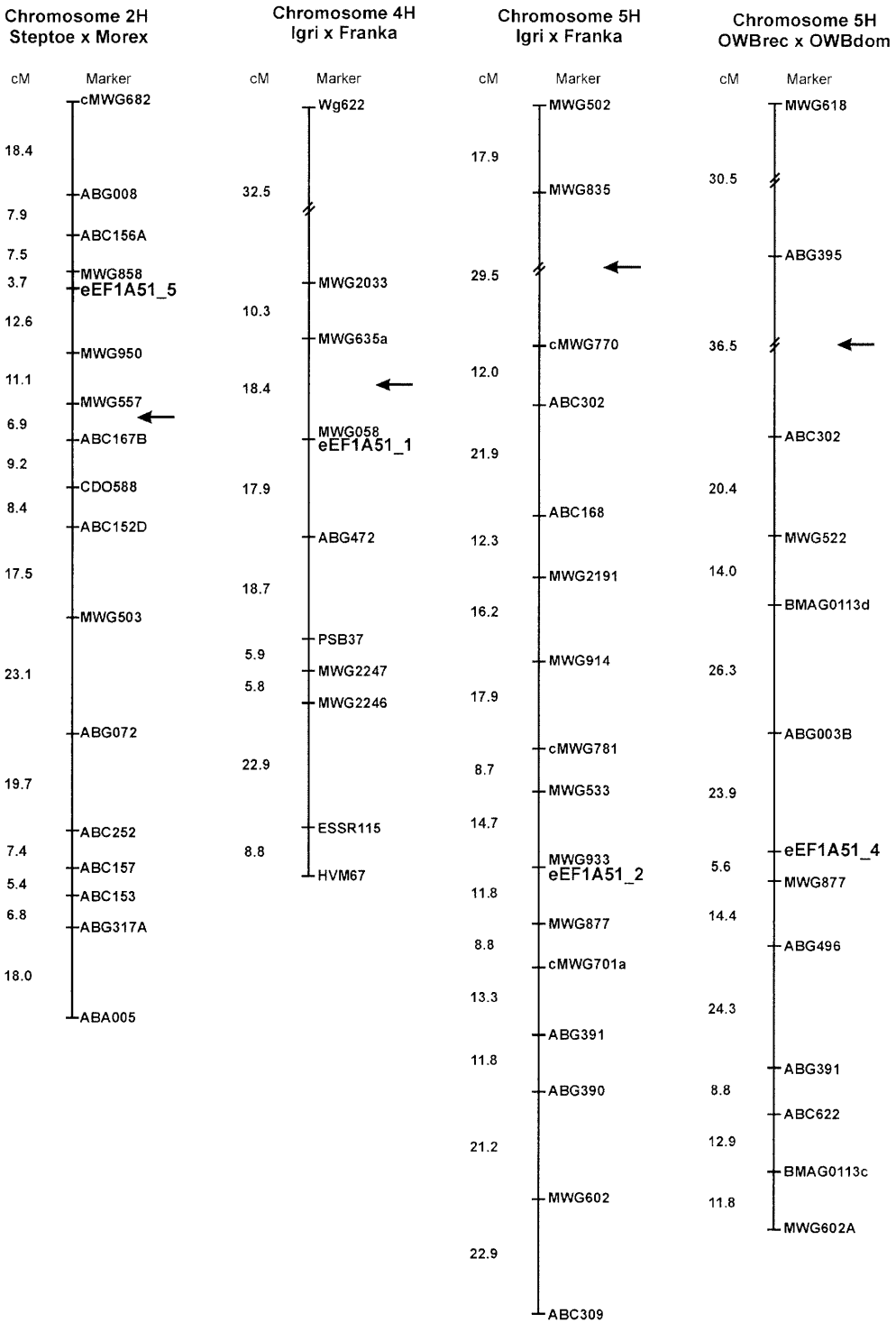


Fig. 3 Molecular mapping of eEF1A genes on barley chromosomes 2H, 4H and 5H. Chromosomes are oriented with the short arm on top. The approximate position of centromeres has been indicated by an *arrow*. Marker distances are given in centiMorgan (cM). The eEF1A genes are denoted eEF1A with the cluster ID (51_1, 51_2, 51_4, 51_5) added

Since the 3′-sequences within a cluster were highly conserved, a multiple alignment using ClustalX software was attempted and primers were defined in order to enable specific amplification of the fragments from genomic barley DNA (Fig. 2). In all cases a single product was obtained and it was possible to detect a polymorphism in one of the three mapping populations tested for all fragments, except 51_3. Fragments based on clusters 51_1 and 51_2 revealed point mutations between the cultivars ‘Igri’ and ‘Franka’, the amplification products with 51_4 uncovered a 11 bp deletion in parent ‘OWBrec’, and with primers designed from the 51_5 cluster data a 27-bp duplication in ‘Morex’ was detected (visible as an insertion/deletion between 51_5 and Z23130 in Fig. 2). Based on these polymorphisms the corresponding ESTs were integrated into the genetic maps of chromosomes 2H, 4H and 5H (Fig. 3). For the monomorphic fragment from 51_3 we used wheat-barley addition lines (Islam et al.

1981) to localize the the amplicon on chromosome 4H (data not shown).

Discussion

As a first step towards the construction of a "unigene set" we generated 13,109 barley EST sequences derived from 7,568 clones of three cDNA libraries derived from leaves, roots and the young developing caryopsis. The sequencing success rate and the other quality parameters used to characterize the present data are comparable to the human EST sequencing project (Hillier et al. 1996). A critical step is the determination of the number of different genes identified. For this purpose the availability of 3´ sequences proved to be essential in this study. Only a limited number of clones of a standard cDNA library are full length, and may contain 5´ non-translated sequences. Therefore, contrary to the rather conserved 5´ sequences, which were mainly derived from the coding region, the DNA variability of the 3´ untranslated region allowed an efficient dissection of gene families. Based on the cluster analysis the number of tentative unique cDNAs was estimated to be 4,000.

The quality of the clustering process was examined by closer inspection of the most-abundant transcripts found in this set of ESTs, namely elongation factor 1 alpha (eEF1A). Despite extensive studies carried out on this multifunctional protein, its role in the living cell is still not completely understood (Browning 1996). In *Arabidopsis thaliana*, eEF1A proteins are encoded by a multigene family of four members located on two loci (Tremousaygue et al. 1997). In barley, eEF1A has been mapped to four different loci on chromosomes 2H, 4H, 6H and 5H (Nielsen et al. 1997). Our data indicate, that the (at least) five members of the eEF1A family are present in this set of EST-data and have been correctly differentiated by the software used. As shown in Fig. 2 some sequences from non-translated 3´-regions show perfect homology to previously published barley eEF1A sequences. The present EST collection does not contain all members of the gene family, since none of the eEF1A members mapped to chromosome 6H. Nielsen et al. (1997) used a 10-DAP endosperm library for screening with a 20-DAP endosperm mRNA as a negative and mRNA from the pericarp fraction as a positive probe. Re-screening and analysis of 3´-sequences from seven clones revealed sequence identity. In contrast, the EST data show the expression of different eEF1A genes in the cDNA library from developing caryopses. Nielsen and coworkers possibly identified an aleurone-specific eEF1A gene, and genetic mapping of the clone uncovered the loci of other members of the gene family due to cross hybridization. The occurrence of cross-hybridization is not surprising, since the high conservation of the coding sequence of eEF1A is well-documented (Browning 1996). We did not analyze the full-length genes or the complete ESTs, but clustering of all of the 5´-sequences of the clones in two contigs (data not shown) illustrates the high similarity in the coding regions and the importance of 3´-sequence data for dissecting gene families. Since the consensus sequence of cluster 51_3 is identical to the sequence published by Nielsen et al. (1997) it may be speculated that this eEF1A gene resides on chromosome 4H.

While 3´ sequence information has proven essential for the separation of the individual members of gene families, the sequence information obtained from the 5´ end has proven to be superior regarding gene discovery (Fig. 1). Based on the unigene set 61% of the non-redundant barley ESTs showed significant homology to known genes. Clearly, this value is transient and will improve with the augmenting information deposited in databases. Therefore, it is not surprising that the gene discovery rate in previous studies was lower, as 28% was reported for rice (Sasaki et al. 1994). Interestingly 23% of the present ESTs did not find a homologue in the EST collections of rice, maize and *Arabidopsis*. This may be taken as an indication that the libraries used in the present study may contain a significant number of plant genes that are not represented in the above-mentioned EST collections. On the other hand, this might mean that a considerable portion of barley genes has evolved rapidly enough to be unique within the set of species compared.

The results of the clustering process gave hints towards differential gene expression within libraries as well. In order to identify only reliable differences, clusters with less than five members (sequence information from three clones minimum) were excluded (Ewing et al. 1999). Although the dataset is limited with regard to the number of libraries under consideration, we observed differential gene expression. From the 442 clusters with ≥5 sequences, 137 (31%) showed library specific expression and more than half of them (56%) were observed in the caryopsis library (Table 2). Inspection of the BlastX data revealed that the root library might contain slightly more unknown transcripts than the other libraries. This tendency is also documented in the data published by Yamamoto and Sasaki (1997). One explanation for the fact that BlastX hits occur more frequently with redundant sequences there may be a tendency for the protein database to contain more highly expressed genes than those still waiting for discovery and, therefore, are more likely to be represented by an EST.

Based on the identification of singletons and distinct gene families, calculations revealed a total of 11,601 protein types for the *Arabidopsis* genome (The Arabidopsis Genome Initiative 2000). Thus, 4,000 tentative singletons and the distinct gene families identified in the present study may account for 25–30% of the genes representing a plants proteome. Therefore, the dataset already represents a valuable resource for a series of applications, including the identification and development of novel microsatellite markers which occur at a frequency of more than 1% (Thiel, unpublished), the development of SNP markers, especially using 3´ sequence information, and the development of cDNA arrays to be used for RNA profiling (Brown and Botstein 1999). In the area of

comparative genomics, the deployment of the present EST data allows the rapid transfer of mapping data from rice via the identification of the barley homologue based on sequence similarity.

In order to make this EST resource, and the information attached to it, available, a comprehensive database has been set up, containing sequences, putative functions, and information about redundancy. This information is regularly updated and can be reached through the internet, thus forming a valuable tool for the utilization of the barley EST data set.

Acknowledgments We gratefully acknowledge the technical assistance of C. Künzel and B. Brückner. We thank Winston Hide for the Stack_Pack software and Tania Broveak-Hide for helpful advice during installation and use of the program. The SQL database was constructed by Ulf Willscher. The work was funded under Sachsen-Anhalt grant No. 2686A/0087B.

References

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

Brown PO, and Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nature Genet 21:33–37

Browning KS (1996) The plant translational apparatus. Plant Mol Biol 32:107–144

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

Cooke R, Raynal M, Laudie M, Delseny M (1997) Identification of members of gene families in *Arabidopsis thaliana* by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. Plant J 11:1127–1140

Davis GL, McMullen MD, Baysdorfer C, Musket T, Grant D, Staebell M, Xu G, Polacco M, Koster L, Melia-Hancock S, Houchins K, Chao S, Coe EH (1999) A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged site ESTs in a 1,736-locus map. Genetics 152:1137–1172

Epple P, Apel K, Bohlmann H (1997) ESTs reveal a multigene family for plant defensins in *Arabidopsis thaliana*. FEBS Lett 400:168–172

Ewing RM, Ben Kahla A, Poirot O, Lopez F, Audic S, Claverie J-M (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9:950–959

Girke T, Todd J, Ruuska S, White J, Benning C, Ohlrogge J (2000) Microarray analysis of developing *Arabidopsis* seeds. Plant Physiol 124:1570–1581

Graner A, Bauer E, Kellermann A, Kirchner S, Muraya JK, Jahoor A, Wenzel G (1993) Progress of RFLP-map construction in winter barley. Barley Genet Newslett 23:53–59

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin S, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics 148:479–494

Hillier L, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Bento Soares M, Tan F, Thierry-Meg J, Trevaskis E, Underwood K, Wohldman P, Waterston R, Wilson R, Marra M (1996) Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6:807–828

Islam AKMR, Shepherd KW, Sparrow DHB (1981) Isolation and characterization of euplasmic wheat-barley chromosome addition lines. Heredity 46:161–174

Jurka J (1998) Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol 8:333–337

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

Kurata N, Umehara Y, Tanoue H, Sasaki T (1997) Physical mapping of the rice genome with YAC clones. Plant Mol Biol 35:101–113

Miller RT, Christoffels AG, Gopalakrishnan 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

Nielson PS, Kleinhofs A, Olsen O-A (1997) Barley elongation factor 1 alpha: genomic organization, DNA sequence and phylogenetic implications. Genome 40:559–565

Sasaki T, SongJ, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Mirayama-Kayano E, Takiguchi T, Takasuga A, Niki T, Ishimaru K, Ikeda H, Yamamoto Y, Mukai Y, Ohta I, Miyadera N, Havukkala I, Minobe Y (1994) Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J 6:615–624

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, Manners JM (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. Proc Natl Acad Sci USA 97:11655–11660

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

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Tremousaygue D, Bardet C, Dabos P, Regad F, Pelese F, Nazer R, Gander E, Lescure B (1997) Genome DNA sequencing around the EF-1 alpha multigene locus of *Arabidopsis thaliana* indicates a high gene density and a shuffling of noncoding regions. Genome Res 7:198–209

Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35:135–144