CHAPTER 1

Expressed Sequence Tags (ESTs) and cDNA arrays as tools for global expression analysis in barley

1.1 AN INTRODUCTION TO EXPRESSED SEQUENCE TAGS

The discovery, characterization, and exploitation of agriculturally important genes is critical to further increase productivity and to meet the food security needs of mankind because of the ever increasing population and the hardship being faced by the agriculture. Prime targets are the genes of crop plants like rice, wheat, maize, barley and sorghum, which belong to the ten most important crop plants worldwide. Traditionally, gene discovery programs followed a "one gene at a time" approach, which is both costly and time consuming. In the present era of genomics, scientists are taking global approaches such as genomic sequencing. Among cereals, rice has the smallest genome with a size of only 430 Mbp. Consequently, the complete set of genes and their genomic locations could be identified *via* genomic sequencing with an acceptable investment. The genomes of the other species are considerably larger, sorghum 800 Mbp, maize 2,500 Mbp, barley 5,500 Mbp and wheat 16,000 Mbp, which precludes this approach. As an alternative to a genomic sequencing program, an EST based approach, which is an unedited sequence generated from single-pass sequencing read of a cDNA clone chosen randomly from a library at all stages of plant growth and life cycle allows fast and affordable gene identification at a large scale (Adams *et al.,* 1992; Rounsley *et al.,* 1996). This approach greatly assists in the identification and isolation of economically important genes among cereals. Large EST programs for the grasses and other crop species are currently under way in many research groups worldwide. The ESTs from wheat, barley, maize, sorghum, or other closely related Triticeae species are being produced to maximize the access to all genes in the cereal genomes (Table 1). Currently, the EST database (http://www.ncbi.nlm.nih.gov/dbEST) contains 684,838 EST entries from monocotyledonous

plants, out of which 163,282 are reported from wheat, 155,288 are reported from maize, 155,287 from barley, 104,880 from rice, and 107,278 from different species of sorghum.

Table 1: Expressed sequence tags of major cereals in dbEST

species	ESTs	cDNA libraries	low quality	≤100 b/≥800 b	*E. coli*
Triticum aestivum	163,283	38	4,068	82 / 1,793	198
Zea mays	155,288	31	3,850	186 / 1,352	16
Hordeum vulgare	155,287	31	5,043	637 / 24,916	178
Oryza sativa	104,880	27	8,889	140 / 1,464	289
Sorghum bicolor	84,712	10	132	349 / 18	132
Sorghum propinquum	21,387	2	41	10 / -	31
Sorghum halepense	1,179	1	-	- / -	10

For the major cereals the number of entries in dbEST (05-2002) and the number of cDNA libraries from which more than 500 ESTs were derived is listed. Critical quality parameters include the number of ESTs containing low quality segments (≥3 ambiguities/25 bases), short (≤100 bases) and overly long ESTs (≥800 bases) as well as contaminations, e.g. sequences showing homology to *E.coli* sequences (>100 bases with ≥95% identity).

The large-scale EST projects provided an extensive reservoir of sequences in cereals. To accomplish further biological knowledge the available sequence information of respective genes has to be converted into biologically significant knowledge with respect to putative identification of functional role of genes and relative abundance of transcripts belonging to different cells, tissues, developmental stages and stress treatments. Proper annotation of EST data is crucial to integrate the various kinds of data into a higher level of biological knowledge. The EST approach is inexpensive and efficient in gene-discovery that can be used to identify novel cDNAs encoding enzymes of specific plant metabolic pathways. Collections of ESTs from metabolically active tissues during different developmental stages of plant growth and seed set provide a platform for quantitative estimates of gene expression levels and thus to unravel plant metabolic and regulatory networks.

1.1.1 EST-based gene discovery - its merits and inherent limitations

Gene discovery *via* ESTs is comprised of four steps which include (i) the construction of cDNA libraries, (ii) single-pass sequencing of (randomly) selected clones and EST quality

check, (iii) the alignment of ESTs to identify the number of genes represented and (iv) the annotation of these partial sequences or genes which are available thereof.

(i) cDNA library generation

The production of ESTs starts with the construction of cDNA libraries. Within a certain tissue of defined developmental and physiological status, only a specific fraction of the entire set of genes of an organism is expressed and the level of abundance of mRNAs for different genes varies widely. This makes it less likely to identify low expressed genes and leads to redundant sequencing of the ones that are highly expressed. In addition to the construction of several cDNA libraries to cover a wider spectrum of expressed genes, various strategies have been applied to circumvent or minimize redundant sequencing. cDNA libraries can be normalized either during their synthesis by subtractive hybridization or related approaches (Kohchi *et al.,* 1995) or afterwards by techniques such as oligonucleotide fingerprinting (Guerasimova *et al.,* 2001). The identification and exclusion of already sequenced cDNAs or even complete libraries when redundant sequencing exceeds a certain limit, provides another valid alternative to minimize the cost of uncovering new genes. Table 1 provides an overview for the number of relevant cDNA libraries employed in these programs. Despite these efforts it can be shown for species with completely sequenced genomes that the number of genes represented by ESTs is significantly smaller as the number of predicted genes. For instance more than 113,000 ESTs from *Arabidopsis* represent less than 16,200 genes out of the 25,556 genes, which are predicted in the genome.

(ii) EST sequencing and quality check

After the isolation of cDNA clones, plasmid preparation and single-pass sequencing, several quality issues have to be addressed. Vector and low quality sequences as well as bacterial sequences or other contaminations need to be removed from the non-processed sequence data. No generally accepted standards exist for these procedures so that the quality of submitted sequences does depend on the submitting laboratory. Wrong bases as well as small insertions and deletions (indels) go undetected in single-pass sequences. Especially indels occur frequently at short homo polymer stretches at greater read length. For that reason, sequences should be trimmed at a certain read length. This has not been done for many database entries, as can be seen by the large number of ESTs with more than 800 bases (Table 1). Furthermore, handling errors or lane tracking problems in gel-based sequence analysis lead to wrong assignments of clones and sequences. Such errors can not be recognized in databases, but will

become apparent when the cDNA clones have to be used, e.g. for the construction of cDNA arrays (see below).

(iii) EST clustering / Gene content

The assembly of gene sequences or parts thereof from a collection of ESTs to determine the number of represented genes is a non-trivial task. Above-mentioned problems with sequence quality and possible sequence errors together present huge challenges for EST clustering. Special program packages such as the Phred/Phrap/Consed system (http://www.phrap.org/), UniGene (Boguski *et al.,* 1995), Genexpres Index (Houlgatte *et al.,* 1995), TIGR_ASSEMBLER (ftp://ftp.tigr.org/pub/software/TIGR_assembler/), STACK_PACK (Christoffels *et al.,* 1999; Miller *et al.,* 1999), CAP3 (Huang and Madan, 1999), PCP/CAP4 (www.paracel.com/products), HarvESTer (http://mips.gsf.de/proj/gabi/news/bioinformatics.html) and others have been and continue to be developed for the assembly of large EST collections. The result of the assembly process can be divided in so-called singletons, sequences which do not assemble with any other sequence, and groups of assembled sequences which might be called clusters, contigs, tentative consensus, tentative genes, unique genes (unigenes), etc.

Several institutions provide pre-calculated assemblies of ESTs, sometimes including completely sequenced cDNA clones and genomic sequences to improve the results. Prominent examples are the gene indexes at The Institute of Genomic Research (TIGR; http://www.tigr.org), which provide an overview of gene indices of various species. Even so certain quality issues of ESTs are addressed by TIGR, one should keep in mind that the number of unique sequences should not be interpreted as the number of genes identified in a certain species.

(iv) Employing bioinformatic tools for annotation of ESTs

In addition to the number of genes represented by ESTs, it is important to collect information about their (potential) function and to associate this information with the respective clones. This process, called annotation will help to identify promising targets for further research and to interpret results of downstream applications which employ these clones, respectively their sequences, e.g. global expression analysis. The annotation process has to face the same difficulties as the annotation of unknown genes in genomic sequences (except splice site prediction), but is further complicated by the partial information and the high, yet undefined

error content of ESTs. To minimize these problems, consensus sequences of aligned ESTs should be used whenever available, because they contain more information of increased reliability with respect to individual ESTs. The primary question, which needs to be addressed from the annotation point of view, is if the EST is identical or similar to a known gene. The possible approach is comparing its sequence with appropriate databases using Blast or FASTA programs. Comparisons at the nucleotide level will identify closely related database entries, whereas comparisons at the amino acid level, after translation of the EST in all (meaningful) reading frames, can be used to uncover less related genes. The public availability of databases and the Blast (Altschul *et al.,* 1997) and FASTA (Stoesser *et al.,* 2002) programs as well as the low price of high computing power make it feasible to run many thousand comparisons at low costs within a moderate time. Yet, the incomplete sequence information with respect to the cDNA clone itself and with respect to the gene content of the genome usually precludes a precise answer. Usually the description and references contained in a database entry related to an EST provide a quick access to the relevant information, but several problems are associated with this approach. Mainly as a result of genomic sequencing, many hypothetical genes will be encountered for which no functions could be assigned. The description of a database entry might be outdated or even worse it may propagate annotation errors. To obtain a higher level of confidence specialized databases, which are curated and providing more detailed information can be used for sequence comparisons, e.g. SwissProt (Bairoch and Apweiler, 2000), TRANSFAC for transcription factors (Wingender *et al.,* 2000), BRENDA for enzymes (Schomburg *et al.,* 2002; Schoof *et al.,* 2002).

In case no related genes could be identified for an EST or if the related gene does not provide information with respect to function, attempts shall be made to identify functional motifs, which may guide further investigations. The identification of protein patterns from the PROSITE database (Falquet *et al.,* 2002), Pfam (Bateman *et al.,* 2002) and other databases, the prediction of targeting signals and transmembrane helices as well as the prediction of open reading frames provide several opportunities. In general computational annotation of ESTs is still in its infancy (Table 2). Software tools have to be improved significantly to meet the challenges provided by a rapidly increasing number of ESTs and to cope with their specific problems. Especially for cereals with large genomes EST development will be important because complete genomic sequences are not expected to be available in the near future.

Table 2: Web sites useful for EST annotation

programs	purpose	URL
BLAST	sequence comparison	http://www.ncbi.nlm.nih.gov/BLAST/
FASTA	sequence comparison	http://www.ebi.ac.uk/fasta33
SWISSPROT	protein sequence comparison	http://www.expasy.org/sprot/
PFAM	protein sequence comparison	http://www.sanger.ac.uk/Software/Pfam/
PROSITE	protein pattern findings	http://www.expasy.ch/prosite/
TRANSFAC	transcription factor detection	http://transfac.gbf.de/TRANSFAC/
BRENDA	enzyme functional data collection	http://www.brenda.uni-koeln.de/
TMPRED	trans membrane prediction	http://www.ch.embnet.org/software/TMPRED_form.html
TMHMM	trans membrane helice prediction	http://www.cbs.dtu.dk/krogh/TMHMM/
FRAMED	GC content	http://www.toulouse.inra.fr/FrameD/cgi-bin/FD
GENEMARK	prediction of ORF	http://genemark.biology.gatech.edu/GeneMark/
GENESCAN	prediction of ORF	http://202.41.10.146/
BESTORF	prediction of ORF	http://genomic.sanger.ac.uk/gf/gf.html

The table presents the tools, which are useful for the annotation of ESTs. Some of the publicly available tools are listed, which might be used to annotate translated ESTs with respect to functional motifs, but none of them has been designed or adjusted to handle ESTs specifically and to take care of associated problems.

1.1.2 High throughput transcript profiling by EST arrays

A popular new approach for the examination of global changes in gene expression is the use of high-density cDNA / EST arrays (PCR amplified inserts of full-length or partial sequence cDNAs), which allow to study genome-wide expression levels in parallel (Schena *et al.,* 1995). ESTs provide the main resource for the construction of cDNA arrays in cereals, because genomic sequences are not available, except for rice. The rapidly growing EST databases allow the detection of regions showing sequence homology in functionally related gene products even from distantly related organisms. Thus, it is increasingly possible to assign putative functions for a large proportion of anonymous cDNA clones/ ESTs. Such type of ESTs, once annotated by BLAST search, are being used as resources for the analysis of gene expression with the help of high-density arrays as demonstrated in *Arabidopsis* (Schena *et al.,* 1995; Girke *et al.,* 2000). It is also important to note that array-based results identify novel genes most worthy of detailed characterization. It is often interesting to look into genes belonging to different metabolic pathway that show a dramatic induction or repression in their expression, which in turn provide an integrative view of physiological information of a plant's response during developmental studies.

The construction and use of such EST arrays for high-throughput transcript profiling can be divided into four general steps, which are depicted in Figure 1. These steps comprise (i) the identification of a non-redundant set of cDNA clones, (ii) the synthesis and deposition of hybridization targets on an appropriate surface, (iii) preparation of mRNA from the tissue of interest, labelling of the hybridization probe and hybridization of the array and (iv) data acquisition and evaluation.

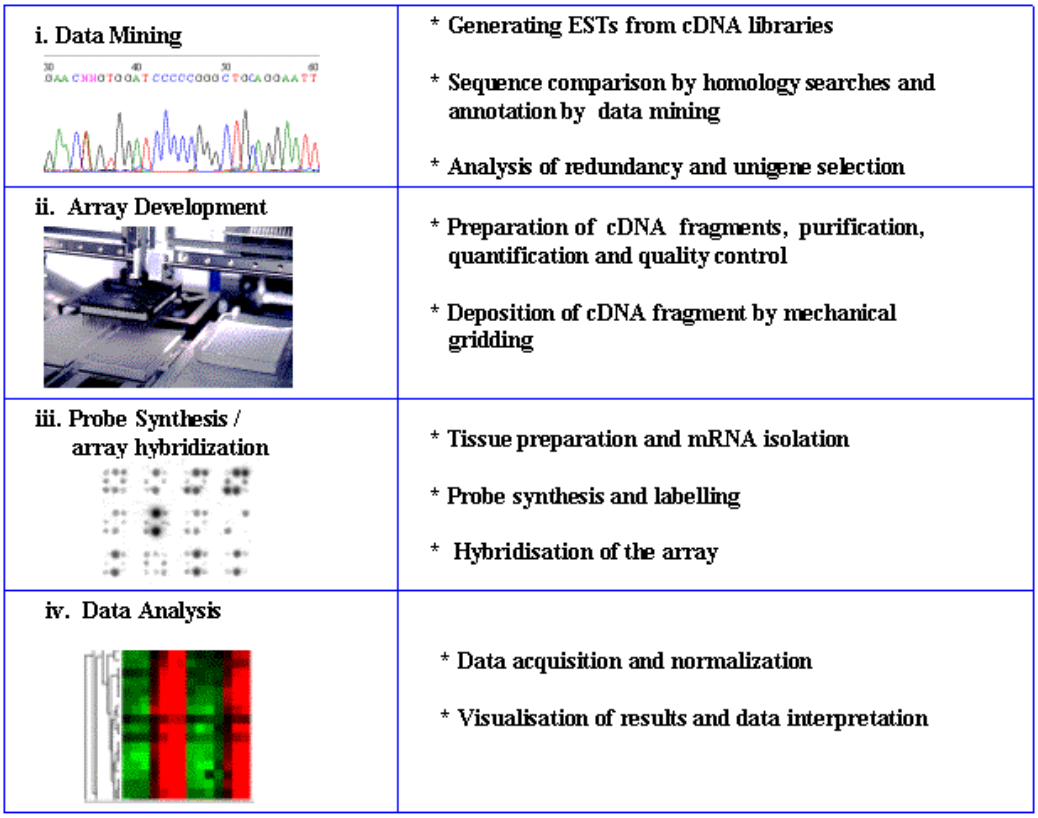

Fig. 1 A diagrammatic representation of EST-array technique.
Four major steps involved in EST-array production technology are i. Database mining; ii. Array development; iii. Probe synthesis/array hybridization; iv. Data analysis. The sub-steps followed in every major step have been provided with a star mark on the right side of the diagram.

(i) *Data mining*: The development of a non-redundant unigene set from ESTs has been covered in the above section. It serves the purpose to minimize the number of samples on a cDNA array mainly for technical reasons, even so a low degree of redundancy will provide data for quality control (Herwig *et al.,* 2001).

(ii) *Array development*: Several different approaches, which are summarized in Table 3, could be taken for the construction of a cDNA array. The least expensive approach is the PCR amplification of cDNA fragments using vector primers and their spotting on Nylon membranes or chemically modified glass or plastic surfaces (for review specifically on plant cDNA arrays see Richmond and Sommerville, 2000). For that purpose the cDNA clones from the EST project have to be available and all handling errors with respect to the clones will be reflected on the array. The second approach uses long oligonucleotides (50 – 80mers), which can be synthesized and spotted instead of cDNA fragments. The advantage of this approach is that oligonucleotides can be designed to distinguish members of gene families, that cDNA clones need not to be available and that handling errors with respect to the clones will not affect the array. The third approach is the on-chip synthesis of short oligonucleotides (25mers), which is offered by Affymetrix (http://www.affymetrix.com/). Set-up costs are high; furthermore, the array design is rather static with respect to the gene content, because a new design would require a completely new set-up. Therefore, construction of these types of arrays is thought to be useful, if a genomic sequence is available to identify most of the genes or parts thereof with a high degree of reliability. Except for Affymetix arrays, the oligonucleotides or cDNA fragments need to be transferred and permanently attached to the array surface. Usually this is accomplished by solid or slit pins which pick-up the samples from microtiter plate wells and transfer them to the target locations on the array. Spot distances on the order of 100 to 400 μm, up to several thousand spots per array and transferred volumes in the picolitre-range require high precision, high speed moving devices which perform this task in an environment with precisely controlled temperature and humidity. For the permanent bonding of cDNA fragments gene products are immobilized on to the solid support such as nylon or nitrocellulose membrane defined as macroarrays or onto glass surface usually called microarrays. Array designs used for expression analysis differ widely with respect to the hybridization targets, the solid support, the method of application of hybridization targets and their density, as well as the label which is used to detect hybridization intensities.

Table 3: Design principles of arrays used for expression analysis

target on array	array surface	target application	features cm^{-2}	label
cDNA fragments	Nylon membrane	spotting	100	^{33}P
cDNA fragments	modified glass or plastic	spotting	4,000	fluorescent dye
Oligonucleotides (50 – 80mers)		spotting	4,000	fluorescent dye
oligonucleotides (25 mers)		on-chip synthesis	300,000	fluorescent dye

(iii) *Probe synthesis / hybridization*: The next step in cDNA array analysis involves the isolation of mRNA, probe synthesis and labelling as well as the hybridization with the array. To synthesize a labelled hybridization probe various protocols are available (Gupta *et al.,* 1999). Generally, ^{33}P-labelled nucleotides are employed when membrane based macroarrays are hybridized, because incorporation rates are high and sensitive phoshoimagers can be used for signal detection. Radioactive labels cannot be used for any kind of microarray, because the spatial resolution of phosphoimagers is not sufficient to separate signals of neighboring spots. Usually, fluorescent dyes are incorporated either directly using dye modified nucleotides (CyDye™ fluorescent dyes: Amersham/Pharmacia) or indirectly *via* aminoallyl-modified dUTP (Molecular Probes, Stratagene). Alternative strategies employ for example the incorporation of biotinylated nucleotides and labelling with phycoerythrin-conjugated streptavidin after the hybridization was performed (Affymetrix). Hybridizations are performed under the most stringent conditions possible to prevent cross-hybridization.

(iv) *Data analysis*: Afterwards signals are detected using specialized scanners for microarrays and phosphoimagers for macroarrays. Resulting images are processed with software for automatic spot detection to derive a list of signal intensities for all features on array. This raw data has to be processed to gain biological knowledge. Important steps include (a) the critical assessment of data reliability and normalization to allow the comparison of different experiments as well as (b) the categorizing of gene expression profiles and their biological interpretation.

(a) Depending on the type of experiment, various procedures can be employed to normalize raw data for comparison with a series of other experiments. These procedures range from

mathematical methods, which assume that the intensity distribution of signals does not change between experiments to the use of reference signals, which are derived from housekeeping genes or foreign mRNAs included in probe synthesis. The choice of a method will often influence the experimental design and has to be made before an array is constructed. Based on the comparative results with macroarray experiments and Northern blot controls for many differentially expressed genes lead to the conclusion that mathematical methods are sufficiently accurate (Sreenivasulu *et al.,* 2002; Potokina *et al.,* 2002). Equally important is a careful evaluation of signal and array quality. Most often the initial dataset will be reduced to

Table 4: Analytical tools with application to gene expression and worldwide web addresses of software's for array data analysis from the public domain as well as the private sector.

Organization	**Primary function**	**URL**
Academic software's:		
Array Viewer	Multi experiment viewer,	http://www.tigr.org/softlab/
Image/J	Image processing	http://rsb.info.nih.gov/ij/
Spot finder	Spot detection	http://www.tigr.org/softlab/
Scan Alyze	Spot detection	http://rana.lbl.gov/EisenSoftware.htm
Cluster	Data filtering/ clustering	http://rana.lbl.gov/EisenSoftware.htm
Tree View	Cluster visualisation	http://rana.lbl.gov/EisenSoftware.htm
Xcluster	Clustering, visualisation	http://genome-www.stanford.edu/~sherlock/cluster.html
J-Express	Clustering, visualisation	http://www.ii.uib.no/~bjarted/jexpress/
Genesis	Clustering, visualisation	http://genome.tugraz.at
Amanda	Clustering, visualization	http://xialab.hku.hk/software
Data explorer	Data flow visual program	http://www.opendx.org/
The R language	Comprehensive statistical Analysis, clustering, etc	http://cran.us.r-project.org/
Cyber T	t-test variants for gene expression datasets	http://genomics.biochem.uci.edu/genex/cybert/
Commercial softwares:		
Array-Pro	Spot detection	http://www.mediacy.com/arraypro.htm
Array Vision	Image visualization, Spot detection	http://imaging.brocku.ca/products/Arrayvision.htm
Array Explorer	Clustering and visualization	http://www.spotfire.net/
Expressionist	Clustering, visualisation	http://www.genedata.com/products/expressionist/
Gene Maths	Clustering, visualisation	http://www.applied-maths.com/ge/ge.htm
Gene Sight	Clustering, visualization	http://www.biodiscovery.com/products/genesight/ genesight.html
Gene Spring	Clustering, visualisation and normalization	http://www.sigenetics.com/cgi/SiG.cgi/index.smf
JMA Viewer	calls KEGG, BLAST,	http://sequence.aecom.yu.edu:8000/jmaviewer/
Partek	Clustering, visualisation 3D gene expression data	http://www.partek.com/

a much smaller dataset of differentially expressed genes within this selected dataset. Experimental artifacts, which lead to large differences in signal intensity, will specifically accumulate and cause misleading interpretations. In addition, the biological variability will significantly influence the data and it is good practice to repeat each experiment with hybridization probes from independently obtained tissue samples. It seems to be very difficult or even impossible to control all environmental variables to such an extent that no significant variation in gene expression is observed in such repeats.

(b) As a consequence of the large number of data points obtained from just a few moderately sized experiments, evaluation of the data has to be supported by computational methods. For these purposes several software packages are available commercially and in the public domain. An overview is given in Table 4. To categorize expression profiles, several methods from multivariate statistics can be employed, such as hierachical clustering (Eisen *et al.,* 1998), K-mean clustering (Tavazoie *et al.,* 1999), principal component analysis, self-organizing maps (Tamayo *et al.,* 1999) and others. If they are used on a carefully controlled reliable dataset, they will yield similar, but not identical results.

1.1.3 Biological interpretation of expression data

Finally, expression data are expected to yield insights into metabolic and regulatory processes during plant development. To reach that goal, it is necessary to compare the preprocessed array data with known models of metabolic and regulatory networks as depicted in KEGG (Goto *et al.,* 1997, http://www.genome.ad.jp/kegg/metabolism.html), the Boehringer biochemical pathway database (Michal, 1993; http://www.expasy.ch/cgi-bin/search-biochem-index) or the general literature and to confirm or reject specific hypotheses. Many successful examples have been provided already, e.g. the analysis of seed development (White *et al.,* 2000, Ohlrogge and Benning, 2000) or phytochrome A signalling (Teppermann *et al.,* 2001) in Arabidopsis and the analysis of salt stress in rice (Kawasaki *et al.,* 2001).

Until now, most of this interpretation process is a manual task, which requires the simultaneous integration of many different information resources. Software tools to support this complicated process are still in their infancy. Implementation of powerful interactive simulation environments for metabolic and regulatory networks, such as Metabolika (Hofestädt and Scholz, 1998), with integrated access to the information about related genes, proteins and metabolites as well as the actual expression data will be a next important step.

Until such tools are available the development of new hypotheses from the data of expression analysis will continue to depend on human ingenuity.

1.2 RESULTS AND DISCUSSION

1.2.1 EST generation from developing caryopses library (0-15 DAF)

A program aimed at the functional genomics of barley seed development was started with the synthesis of cDNA libraries from developing caryopses. In the Institute of Plant Genetics and Crop Plant Research (IPK) cDNA libraries from developing caryopses (0-15 DAF) were constructed and cloned into λ-ZAP Express (Stratagene) according to the manufacturers instructions (W. Weschke). In total 6,319 ESTs were generated from developing caryopses libraries either from 3' or 5' ends. Sequence cleaning and quality check has been performed under high-stringent conditions. Comparisons to other plant EST-sequences and redundancy within the EST collection has also been performed (Michalek *et al.,* 2002). The EST sequence of all clones along with clustering information is available at our web site http://pgrc.ipk-gatersleben.de.

1.2.2 Annotation and functional classification of barley ESTs from developing caryopses

We examined the cDNA clones associated with pre-storage and initial storage phase of developing barley caryopses by EST approach. Clones were selected preferentially from a cDNA library of developing caryopses (1235 clones) and smaller numbers were chosen from etiolated seedlings (70) and roots (104) library. ESTs were annotated with reference to gene function using the results of BlastX2 comparisons with the SwissProt protein database. SwissProt was used instead of TrEMBL to prevent the occurrence of a large number of functionally non-informative database matches with putative or hypothetical proteins from genomic sequencing projects. Information regarding score, length of the aligned sequence segment and other parameters were extracted from the results using a custom made Perl script. EST sequences were grouped in three categories based on the score and the length of the aligned sequence segment with the top database hit after BlastX2 comparison with SwissProt. Two straight lines, which separate the three categories, were defined on a scatter plot of score versus aligned length by manual annotation of approximately 700 sequences.

These lines run through a common point defined by the minimal alignment length of 12 amino acids (aa) and a corresponding score of 27 bits. The "Secure" and "potential" assignments were separated by a straight line with the slope of 1.36 bits/aa and "potential" and "unassigned" sequences were separated at 0.62 bits/aa. In case, 5'- and 3'-end sequences were available the highest category was assigned to the cDNA clone of top hit. All cDNA clones on our array were categorized by using these criteria. Out of 1421 cDNA fragments 1309 unique ESTs were identified based on BlastX2 assignment. ESTs were grouped in three categories, called "secure" (509 clones, 38.9%), "potential" (308 clones, 23.5%) and "unassigned" (492 clones, 37.6%) (Fig. 3a), based on the ratio between score and length of the aligned sequence segment as described above.

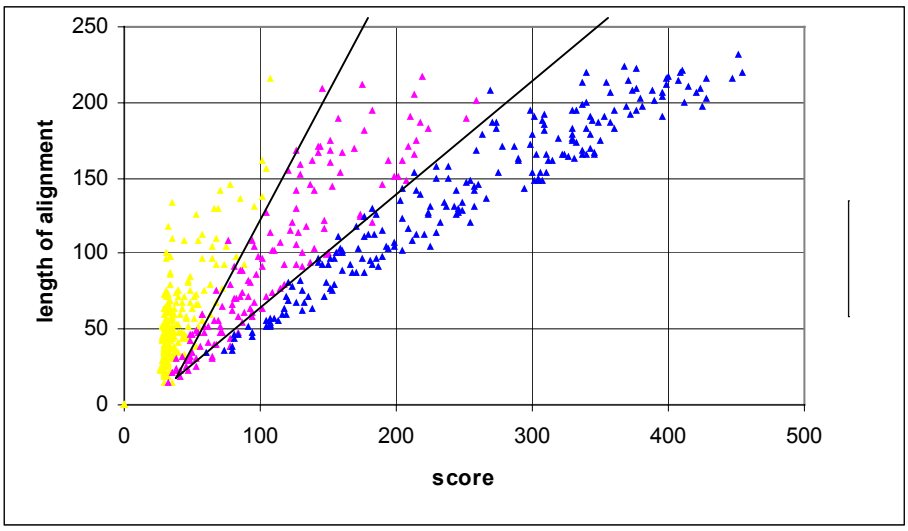

Fig. 2 Scatter plot representation of EST annotation data

Plotting of score value and length of alignment of 700 ESTs on X and Y axis respectively. 'Secure' class is represented by blue colour, 'potential' class by pink colour and 'unassigned' class by yellow colour.

A second, independent approach was taken to estimate the number of genes represented on the cDNA array, which is independent of known genes in databases. For that purpose sequences of a larger set of ESTs (Michalek *et al.,* 2002) from which clones on the array had been selected were clustered using StackPack. Depending on the use of 5'- or 3'-end sequence data, the ESTs on the array represent between 1176 and 1199 consensus sequences and singeltons. Of those approximately 410 (404 [5'], 426 [3']) belong to the "secure", 300 (300 [5'], 299 [3']) to the "potential" and 470 (472 [5'], 474 [3']) to the functionally unassigned group of ESTs.

To allow the placement of EST encoded genes on metabolic pathway charts, as provided by KEGG (www.tokyo-center.genome.ad.jp/kegg), EC-numbers were extracted from the description line of a matching SwissProt entry for "secure" and "potential" assignments (62.4% of the clones present on our array). The remaining 37.6% with no assignment were placed in non-significant homology section (Fig. 3a). In the total cDNA set, sequences assigned to carbohydrate metabolism (No. 1 in Fig. 3b; 6.94%), amino acid metabolism (No. 5; 6.85%) and genes involved in energy metabolism (No. 3; 6%) dominate, followed by groups of metabolism of miscellaneous substances (No. 8, 2.3%), cell division and cell cycle genes (No. 13, 1.57%) and genes involved in transcription (No. 14; 3.05%) and translation (No. 16; 4.6%). The largest group (No. 17; 11.8%) contains non-classified genes (Fig. 3b). The complete list of genes and their classification along with sub-classes for all unique clones on our cDNA-macroarray is available on-line at http://pgrc.ipk-gatersleben.de/sreeni.

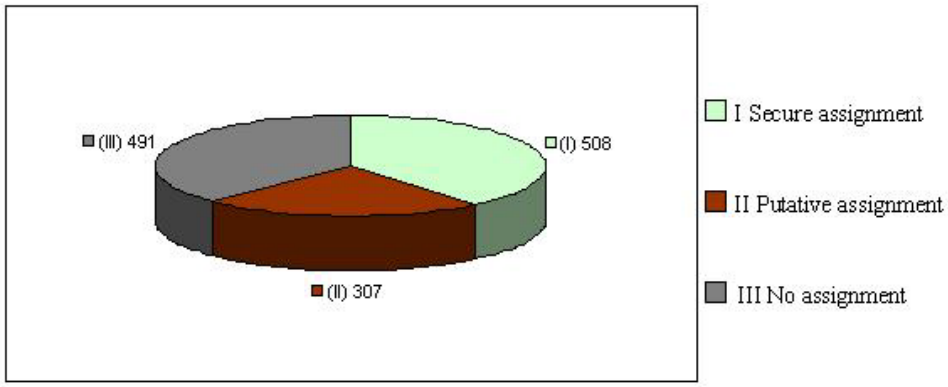

Fig. 3a Annotation of 1400 ESTs from developing caryopses (0-12 DAF)

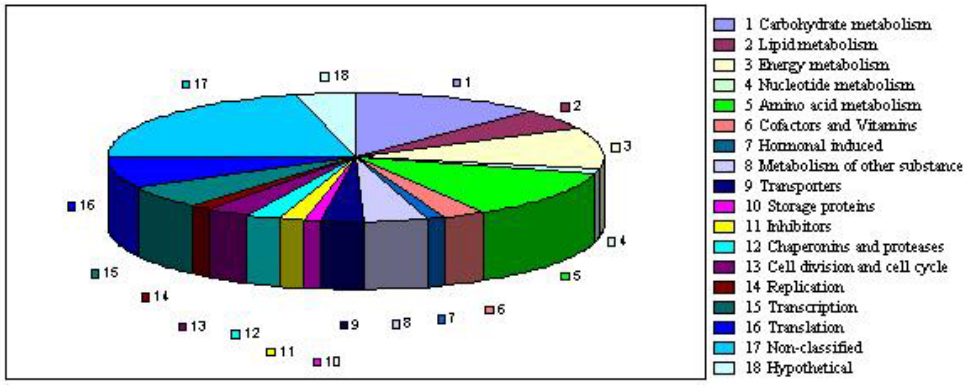

Fig. 3b Functional classification of ESTs from developing caryopses (0 to 12 DAF)

1.2.3 Preparation of an EST macroarray

DNA array technology is an attractive and ideal tool to investigate expression profiles in developmental studies in a large-scale fashion (Tanaka *et al.,* 2000). In comparison, among the available array techniques, the use of nylon membranes and radioactively labelled probes seems to be especially reliable (Herwig *et al.,* 2001). In this study, high-density nylon arrays together with a ^{33}P radioactive probe based hybridization technique have been employed. During the early phase of this program 711 clones representing more than 620 unique genes were selected to construct a cDNA array. Among them 517 clones from a cDNA library of developing caryopses, 70 clones from etiolated seedlings and 104 clones from roots were selected. To produce a larger array, inserts of 1412 cDNA clones containing 1184 unique clones and additionally, some internal control cDNAs were amplified. The same EST amplified independently or different ESTs representing the same gene were used as controls. A complete list of these clones as well as BlastX2 results and other data relevant to this chapter are available from our WWW-server (http://pgrc.ipk-gatersleben.de/sreeni). Based on current sequence and clustering data these clones represent more than 1184 unique genes and therefore comprise the largest collection used for expression analysis of barley reported so far. The cDNA inserts of all clones used for array preparation were amplified with vector specific primers, purified, analyzed on agarose gels, adjusted to concentrations between 2.0 and 1.8 µg/µl, and spotted in duplicate onto nylon membranes as described in Materials and methods. The resulting 711-cDNA array (5 x 9 cm) consists of 10 x 18 subarrays with square of nine spots (see Fig. 4). The 1412-cDNA array (8 x 12 cm) consists of 16 x 24 subarrays, each being a square of nine spots. The central spot of each subarray provides a blank control, while the remaining eight spots contain four different amplification products, each of them represented twice. After hybridization with ^{33}P-labelled second strand cDNA and three washing steps under highly stringent conditions the signals on the array were detected using a phosphoimager. Resulting images were processed with a specialized software package for spot detection, and data files were exported to a standard spreadsheet program. To allow the comparison of data sets from different experiments, signals were normalized with respect to the total amount of radioactivity bound to the array after background subtraction in case of 711-cDNA array. To allow comparison of signal intensities across experiments the median of the logarithmically scaled intensity distribution for each experiment was set to zero in case of 1412-cDNA array (median centering of arrays, Eisen *et al.,* 1998).

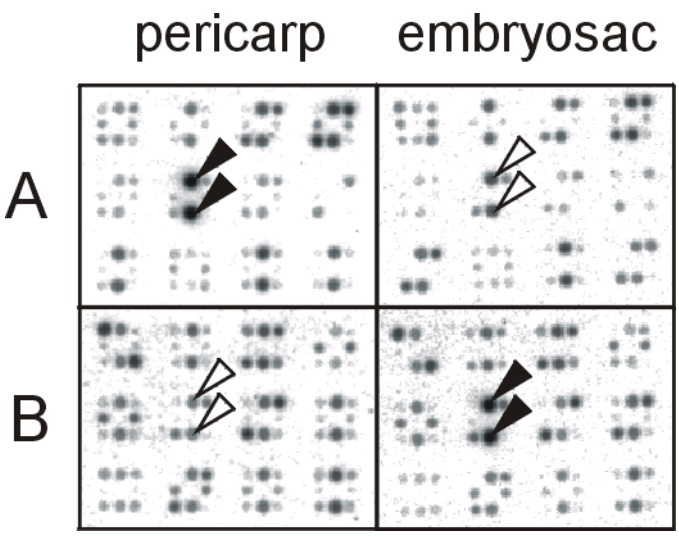

Fig. 4 Segment of a cDNA macroarray

A cDNA macroarray containing 711 clones was hybridized with ^{33}P-labelled second-strand cDNAs derived from pericarp and embryo sac tissues of the developing barley grain 1-7 DAF. Each panel shows 12 subarrays in a 4 x 3 arrangement, which are made up of a blank spot in their center and eight surrounding spots representing four different cDNA-fragments, spotted in duplicate. Hybridization signals for the cDNA clones HY05K19 (A) and HY09L21 (B) which were used for Northern analysis and *in situ* hybridization are marked. The filled triangles indicate strong signals, the open triangles, weaker signals.

1.2.4 Performance of an EST macroarray containing 711 clones

It is important to note here that most of the technical aspects of array preparation and its performance has been dealt with two different cDNA arrays, one with 711 ESTs (620 unique genes) and second one with 1412 (1184 unique genes). In order to get primary insights into pericarp and embryo sac tissue specific expression and to look into the technical details of performance of macroarray, we pooled pericarp (0-7 DAF) and embryo sac probes (0-7 DAF), labelled the probes and hybridized to the macroarray containing 711 ESTs. The probes were synthesized from two completely independent preparations of pericarp and embryo sac tissues (tissue preparations 1 and 2) and used for hybridization first with array 1. In addition, tissue preparation 2 was hybridized to a second membrane (array 2) to check the consistency of results between different arrays. A comparison of the results is shown in a scatter plot (Fig. 5A) which clearly demonstrates that relevant deviations between the two arrays occur only at low signal intensity when the accuracy of the spot finding algorithm diminishes and the influence of background noise increases considerably. Fig. 5B shows the plotted results of a representative experiment (tissue 2/array 1) in which the membrane was hybridized first with cDNA from pericarp and then, after probe removal, with cDNA from embryo sac tissue. cDNAs with a more than two-fold difference in signal intensity between the two tissues can be identified as being outside of the two parallel lines in Fig. 5B. In this experiment, 48 cDNAs appeared to be expressed preferentially in the pericarp and 42 genes were more highly expressed in the embryo sac that gave a signal intensity above 5 arbitrary units (au) in at least

one of the two tissues examined and at levels at least two-fold higher than in the other tissue. If all three experiments (tissue 1 / array 1; tissue 2 / array 1; tissue 2 / array 2) are taken into consideration, 38 clones, representing 34 different genes, consistently showed a more than two-fold difference between the two tissues (Tables 5 and 6). The ratio between the highest and the lowest signal, defined as the average background intensity plus three standard deviations, was used to approximate the dynamic range of our array experiments. With background values ranging from 0.05 – 0.25 au (arbitrary units; standard deviation 0.08 – 0.12 au) and the most intense signals between 450 and 1000 au, the dynamic range has been greater than 1000 in all our experiments. As a consequence of the weak influence of intense signals on neighboring spots (data not shown), we did not fully exploit this dynamic range, but rather restricted our interpretations to clones which had a signal intensity above 5 au in at least one of the two tissues examined (see Fig. 5B).

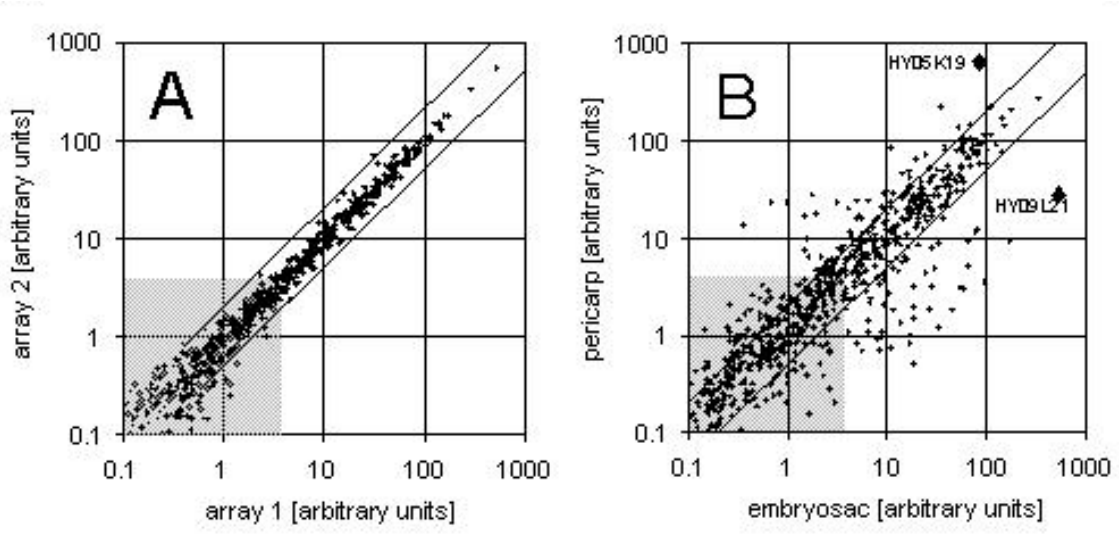

Fig. 5 Comparison of the normalized signal intensities obtained from two independently spotted arrays hybridized with the same labelled cDNA (A) and from one array hybridized successively with labelled cDNA from embryo sac and pericarp tissues of the developing barley grain 1-7 DAF (B).

Signals outside the diagonal lines differ by more than a factor of two between the two hybridization experiments. The cDNA clones HY05K19 and HY09L21 (enlarged symbols in B) were used for Northern analysis and *in situ* hybridization. Signals within the shaded areas were excluded from further evaluation because of their low signal intensities.

Internal controls of genes represented twice on the array but amplified independently (for example, in Table 5, HK03G06) or derived from different cDNA clones of the same gene (HY03B06 and HY10J06; HK03G06 and HW01G04 in Table 5) showed that expression

ratios between pericarp and embryo sac can be reproduced with considerable accuracy within an array. In contrast to the ratios, the signal intensities should not be used as reliable indicator of the expression level of a certain gene, because their values depend strongly on the amount and quality of the spotted amplification product. Reproducibility from array to array was evaluated by hybridizing two independently spotted arrays with the same labelled cDNA. In addition, several cDNAs that showed wide variability between hybridizations with tissue preparations 1 and 2 have been observed, e.g. clone HY10J06 and HY03B06, both are derived from the one gene, pericarp/embryo sac ratios vary from 11 to 40, see Table 5) between tissue preparations. We confirmed that this variability is not a result of hybridization artifacts, but rather a consequence of differences between the two tissue preparations (for further explanation, see below).

Table 5 cDNA clones that are preferentially expressed in pericarp

Clone ID	Tissue 1 /array 1			Tissue 2 / Array1			Tissue 2 / Array 2			BlastX2 result	Score
	peri	emb	ratio	peri	emb	ratio	peri	emb	ratio		[bits]
HY10J06	21	1.9	**11**	24	0.7	**34**	20	0.5	**38**	filamentous flower protein FIL, arabidopsis	68
HY03B06	15	1.4	**11**	14	0.3	**40**	12	0.4	**34**	filamentous flower protein FIL, arabidopsis	133
HY01A03	17	3.0	**5.8**	23	1.0	**23**	25	1.1	**22**	β-amylase	446
HY02E15	20	3.5	**5.8**	17	1.9	**9.0**	29	2.1	**14**	lipoxygenase 2	416
HY05B22	12	2.9	**4.1**	17	1.5	**12**	18	1.7	**10**	probable NADP-dependent	205
HK04H17	19	3.2	**6.0**	28	1.8	**16**	21	2.0	**10**	nonspecific lipid-transfer protein	80
HY05O10	223	44	**5.1**	219	34	**6.4**	270	41	**6.7**	fructokinase	254
HY05K19	459	126	**3.6**	644	84	**7.6**	549	87	**6.3**	methyltransferase/methionine synthase	357
HY03J19	19	5.0	**3.8**	23	3.1	**7.4**	20	3.3	**6.0**	(hypothetical protein T8P19.200,	166
HK03G06	15	3.7	**4.0**	10	1.6	**6.2**	8.7	1.6	**5.4**	cysteine proteinase 1	114
HK03G06	8.9	2.0	**4.4**	7.4	1.5	**5.0**	6.4	1.3	**5.1**	cysteine proteinase 1	114
HY03G16	7.8	2.1	**3.8**	6.5	1.6	**4.1**	7.7	1.9	**4.1**	(hypothetical protein F26G5.50,	122
HW01K18	9.1	3.9	**2.3**	8.0	3.5	**2.3**	9.0	2.5	**3.5**	(rna polymerase beta subunit, virus strain)	40
HW02F11	28	8.3	**3.4**	20	4.1	**5.0**	18	5.3	**3.4**	vacuolar invertase	247
HY08P04	14	4.6	**3.0**	11.5	4.2	**2.7**	7.5	2.3	**3.2**	(oncogene protein, chicken)	30
HY04F24	47	16	**3.0**	72	24	**2.9**	67	22	**3.0**	(p-selectin precursor, mouse)	37
HY04H09	31	11	**2.8**	30	9.5	**3.1**	43	15	**3.0**	auxin-responsive protein	184
HY04N15	57	23	**2.5**	65	20	**3.2**	67	24	**2.9**	probable NADP-dependent	102
HY07C03	37	12	**3.0**	33.8	17.4	**2.0**	40	16	**2.6**	glucan endo-1,3-beta-glucosidase	83
HY03C16	63	18	**3.5**	45	14	**3.1**	49	15	**2.6**	acyl-coa-binding protein	117
HW01G04	48	21	**2.3**	29	12	**2.4**	22	9.3	**2.3**	cysteine proteinase 1	282
HW01F04	11	4.2	**2.7**	5.2	2.1	**2.4**	5.1	2.2	**2.3**	UDP-glucose 4-epimerase	86
HY04F14	20	7.6	**2.6**	16	5.0	**3.2**	17	7.4	**2.4**	glycine dehydrogenase	287
HY10D17	32	8.4	**3.8**	21	8.7	**2.4**	31	14	**2.2**	(web1 protein, baker‘s yeast)	28
HY08K19	18	8.9	**2.0**	16	6.0	**2.8**	17	7.7	**2.2**	(growth factor receptor-bound protein 7,	32
HY09N04	50	23	**2.2**	49	17	**2.9**	29	13	**2.2**	(hypothetical protein, arabidopsis)	46

Table 6 cDNA clones preferentially expressed in the embryo sac

Clone ID	Tissue 1 / Array			Tissue 2 / Array 1			Tissue 2 / Array 2			BlastX2 result	Score
	peri	emb	ratio	peri	emb	ratio	peri	emb	ratio		[bits]
HY03M02	1.3	5.4	**4.2**	3.5	96	**27**	4.2	114	**27**	α-hordothionin	272
HY09L21	100	932	**9.4**	28	541	**19**	20	518	**26**	(nuclear transition protein 2, pig)	38
HY06E14	4.8	16	**3.4**	1.8	46	**26**	2.0	50	**24**	flower-specific gamma-thionin	70
HY09N16	14	34	**2.5**	9.4	174	**18**	11	169	**16**	sucrose synthase 2	331
HY04N22	6.1	17	**2.8**	2.9	58	**20**	4.0	63	**15**	(none)	0
HY04E07	19	73	**3.9**	1.8	26	**14**	1.5	23	**15**	vacuolar processing enzyme	211
HY01O19	6.5	21	**3.2**	3.1	54	**17**	3.3	48	**15**	α-amylase / subtilisin inhibitor	69
HY07E21	25	141	**5.7**	9.1	64	**7.0**	8.2	67	**8.2**	replication factor C 38kD subunit	142
HY03O04	1.4	9.4	**6.6**	1.0	8.6	**8.9**	1.5	11	**7.2**	(monocarboxylate transporter 8, human)	29
HY09L18	22	110	**4.9**	12	77	**6.6**	12	65	**5.5**	probable aspartic proteinase	296
HY02B16	6.8	35	**5.1**	1.8	11	**6.4**	2.4	11	**4.6**	serine carboxypeptidase I precursor	390
HY02F04	57	122	**2.1**	12	82	**6.8**	27	84	**3.2**	(hypothetical protein F3A4.230,	65

Table 5: Clones that are preferentially expressed in the pericarp are defined as those that give pericarp/embryo sac signal intensity ratios larger than 2 and absolute signal intensities greater than 5 au in three array experiments. Normalized signal intensities for pericarp (Peri) and embryo sac (Emb) tissues, as well as the corresponding Peri/Emb (P/E) ratios are listed for three experiments involving two different tissue preparations (Tissues 1 and 2) for probe synthesis, and two different array filters (Arrays 1 and 2). Top scores from a BlastX2 search against the protein databases Swissprot and PIR provide hints as to the potential functions of the respective genes.

Table 6: Clones preferentially expressed in the embryo sac are defined as those with embryo sac to pericarp signal intensity ratios larger than 2, and absolute signal intensities greater than 5 au in three array experiments. Normalized signal intensities for embryo sac (Emb) and pericarp (Peri) tissues, as well as the corresponding Emb/Peri (E/P) ratios are listed for three experiments. For further details see legend of Table 5.

1.2.5 Performance of an EST macroarray containing 1412 clones

The 1412-cDNA array (8 x 12 cm) represent approximately 1184 unique genes of which 63% can be annotated with respect to gene function using the available partial sequences. To gain biological knowledge of pericarp and embryo sac development during pre-storage and storage phase we used cDNA array of 1412 clones. Details of the results were presented in Chapter 2. To allow comparison of signal intensities across experiments in case of 1412-cDNA array, the median of the logarithmically scaled intensity distribution for each experiment was set to zero (median centering of arrays, Eisen *et al.,* 1998). We would like point out that technical detail of array performance has been performed also with cDNA array containing 1412 (1184 unique) clones. The reproducibility of expression patterns of internal controls of some genes represented twice on the array (amplified independently) was evaluated on 1412-cDNA macroarray during pericarp and embryo sac development. The same EST clones amplified

independently showed similar expression patterns and reproduced with considerable accuracy e.g. (Table 7: pericarp-specific expression HK03G06, HY05B22, HY03G16) and (Table 10: embryo sac specific expression pattern during initial storage phase HY07C09, HY04I11, HY10G16, HY05G12, HY02B16, HY03H23). The EST derived from different cDNA clones represents the same gene (Table 7: HY07K19 and HW01G04 represent for cysteine proteinase 1) showed pericarp-specific expression pattern. Further the EST clones (Table 10: HY10G10, HY07L04, HY01H24, HY07F09, HY10F20, HW08D05, HY03P12 represent sucrose synthase 1; HY09L14, HY09N16, HY07C09, HY06C06, HY09N15, HY07H01, HY10D14, HY04D17, HY03J05 represent sucrose synthase 2; HY01D18 and HY04D04 represent ADP-glucose pyrophosphorylase large subunit 1; HY10G16, HY08N11 and HK03F04 represent ADP-glucose pyrophosphorylase small subunit; HY06L03 and HY02N15 represent aspartate aminotransferase, cytoplasmic; HY08O12, HW06P12 and HW03P06 represent 6-phosphofructokinase beta subunit) showed embryo sac specific expression pattern during initial storage phase (6-12 DAF).

In case of gene families, it is expected that homologous sequences will lead to cross-hybridization which may obscure the data for individual members. Even so the overall sequence identity might be as low as 70%, more highly conserved segments do result in considerable cross-hybridization signals (Girke *et al.,* 2000). For that reason we do not know how cross-hybridization influences our data set in general and in special cases where we have clearly identified gene families, e.g. the two different members of the sucrose synthase (SUS) family present on our cDNA array (Table 10: sucrose synthase 1 and sucrose synthase 2). On the other hand based on expression data sucrose synthase 1 and 2 groups are well resolved into different clusters (compare Fig. 10 clusters 4_1 and 4_2).

1.2.6 Expression analyses of selected genes

Based on cDNA macroarray results the methionine synthase (HY05K19) and the unknown 'Nucpro' gene (HY09L21), depict abundant expression in pericarp and embryo sac tissues, respectively (Fig. 5 A, B). To validate the array data, expression of HY05K19 and HY09L21 ESTs was monitored by *in situ* hybridizations. In accordance with expression data, HY05K19 expression was localized mainly in the outer part of the pericarp, especially in the micropylar region but also in endospermal transfer cells (Sreenivasulu *et al.,* 2002). Clone HY09L21 localization by *in situ* hybridization revealed expression exclusively in the cells of nucellar projection (Sreenivasulu *et al.,* 2002), leading to the provisional name 'Nucpro'. This

unexpected result can be explained by the tissue organization of the developing caryopsis, which leads to the adherence of maternal nucellar projection cells to the filial embryo sac (see Sreenivasulu *et al.,* 2002). The sequence of HY09L21 did not exhibit significant homology to any gene of known function, i.e. we regard the BLASTX2 result "nuclear transition protein 2" (score bits 38) as fairly low. The *in situ* localization confirms tissue-specificity seen on the macroarray for the two selected clones. It also points to the importance of using techniques with high spatial resolution, such as *in situ* hybridization, in gene expression studies.

To further validate the macroarray data we selected additional 4 clones for RNA-gel blot analyses. HY03B06 (FIL) shows the highest ratio of pericarp to embryo sac expression (see Table 5), HY09L21 (Nucpro) shows one of the highest ratio of embryo sac to pericarp expression, HY09N16 (HvSUS2) shows unexpected major differences in the ratios between the two independent experiments (see Table 6) and HW02F11 (vacuolar invertase, HvVCINV) shows a rather low expression level but a fairly constant pericarp to embryo sac ratio in all three experiments (see Table 5). Among the genes with low transcript levels but highly specific expression in the pericarp is one represented by two ESTs, HY10J06 and HY03B06 (see Table 5). The sequence is related to that of an *Arabidopsis* gene coding for the transcription factor FIL (*FILAMENTOUS FLOWER*). FIL shows homology to the *CRABS CLAW* genes, founding members of the *Arabidospsis YABBY* gene family (Bowman and Symth, 1999). Northern analysis with an HY03B06 probe (see Fig. 6) in principle confirmed the array data. Northern analysis verified specific expression of HW02F11 in the maternal pericarp with especially high levels at 0 DAF and very low levels at 8 DAF (Fig. 6, 7A). Northern blot analysis of HvSUS2 (HY09N16) expression during early grain development (Fig. 6, 7B) revealed that mRNA levels in the embryo sac tissues rose from undetectable (2DAF) to relatively low (4 and 6 DAF) and eventually very high levels (8 DAF). This increase can be correlated with the enzyme activity profile described earlier (Weschke *et al.,* 2000). A comparison of the mRNA profile of Fig. 7B with the different values obtained for HvSUS2 expression in cDNA array experiments 1 (ratio 2.4) and 2/3 (ratio 18/16, see Table 6) points to a problem in our 711 array analysis which becomes evident for genes like sucrose synthase 2. We studied early seed development from 1 to 7 DAF. Since HvSUS2 mRNA levels rise dramatically at around 7 DAF (see Fig. 6, 7B), the signal intensity in the array is critically dependent on the exact developmental stage of the material collected for analysis. Generally, day 7 after flowering marks the beginning of an exponential increase in sucrose

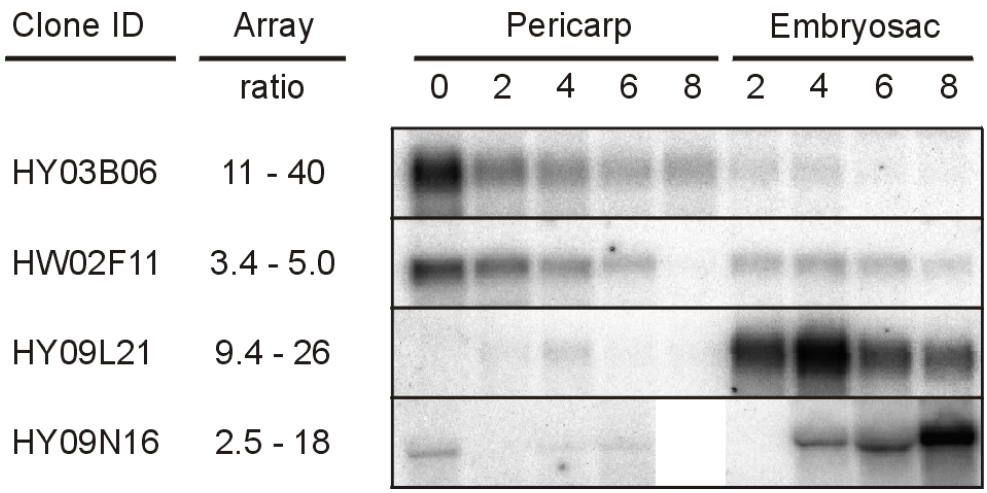

Fig. 6 Levels of transcripts differentially accumulated in pericarp and embryo sac of developing caryopses measured by northern analysis

Differentially expressed clones pericarp to embryo sac ratios (HY03B06 and HW02F11) or embryo sac to pericarp ratios of expression (HY09L21 and HY09N16) as determined by cDNA array analysis were compared by northern blot analysis. The numbers 0, 2, 4, 6 and 8 indicate the time (in days after flowering) at which tissues samples were taken for the isolation of total RNA. HY03B06: FIL-related transcription factor; HW02F11: vacuolar invertase (Hv VCINV); HY09L21: gene of unknown function termed Nucpro; HY09N16: sucrose synthase isoform 2 (HvSUS2)

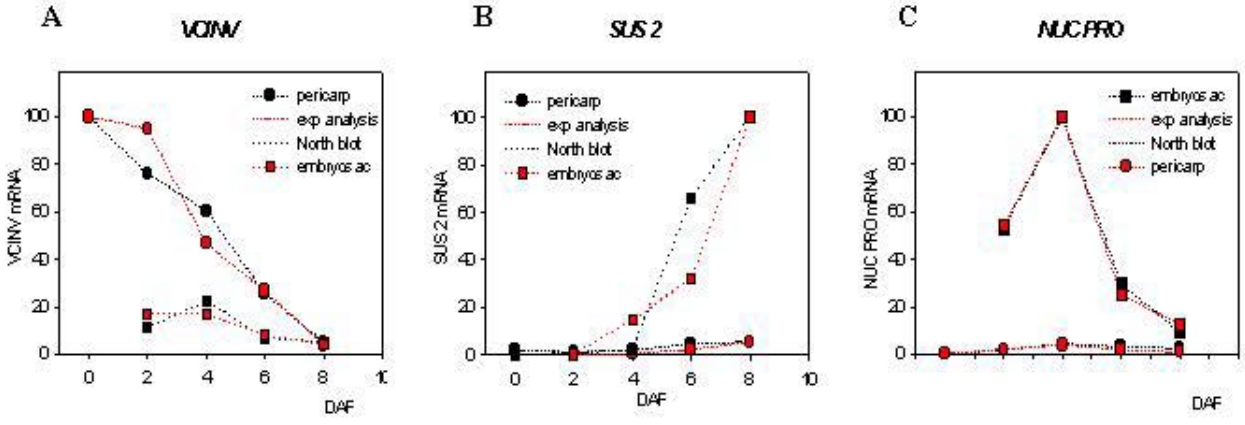

Fig. 7 Comparison of expression levels of selected genes by cDNA array (containing 1412 clones) and northern blot experiments

Transcripts differentially expressed in pericarp and embryo sac tissues of developing caryopses (0 to 12 DAF) were selected based on expression data of 1412-cDNA array and verified expression data by northern blot analysis. For every gene, the highest signal intensity value is considered to be 100% across the developmental scale. The intensity of northern blot

signals was quantified by using the BioImage Analyser and the signal intensity is given in percentage of the most intensive signal (100%). The numbers 0, 2, 4, 6 and 8 indicate the time (in days after flowering) at which tissue samples were taken for the isolation of total RNA. HW02F11: vacuolar invertase (Hv VCINV); HY09L21: gene of unknown function termed Nucpro; HY09N16: sucrose synthase isoform 2 (HvSUS2)

concentration in the whole caryopsis, a remarkable increase in the expression of the caryopsis-specific sucrose transporter HvSUT1 and a linear increase in sucrose synthase activity (Weschke *et al.,* 2000). All these parameters indicate the beginning of starch accumulation in the starchy endosperm. Therefore, small age difference between the caryopses used for pericarp and embryo sac tissue preparation will result in large differences in expression levels observed for genes related to carbohydrate metabolism, as was found for HvSUS2 (see above). Further, we demonstrated the nicely correlated expression patterns of SUS2 during 0 to 8 DAF between array (containing 1412 cDNA clones) and northern blot experiments.

1.3 CONCLUSIONS

In our first attempt we annotated ESTs generated from early stages of developing caryopses and used these resources to establish the cDNA macroarray technique to analyse the complexity of developing processes in the early stages of barley grain. Performance and evaluation of macroarray results originating from replicated experiments enhanced our ability to confirm the standardisations. Further, our results provide correlative evidence for expression data of selected gene candidates by using independent methods such as Northern blotting and *in situ* hybridization. Results presented in Chapter 1 can be used as starting point for isolation of tissue specific promoters. Given the drive to the more focused analysis, we used larger array for detailed studies of the timing of expression patterns in seed development (see Chapter 2) and, further, *seg8* mutant analysis during seed development (see Chapter 3).