ANALYSIS OF EXPRESSED SEQUENCE TAGS (ESTS) FROM SORGHUM BICOLOR GERMINATING EMBRYOS

by

ST PATRICK MARINHO REID

(Under the direction of Lee Pratt)

ABSTRACT

Sorghum is a major source of nutrition in developing countries. Additionally, among cereal grains sorghum has one of the smallest genomes, making it an excellent model system for cereal grains. In this study an Expressed Sequence Tag (EST) database from *Sorghum bicolor* embryos germinated for a period of 24 hr was constructed. High-throughput sequencing produced a total of 5,405 and 5,126 ESTs from the 5′ and 3′ ends of the cDNA clones respectively. High-throughput bioinformatic analysis revealed that although a small percentage of contamination was observed most of the ESTs were derived from the cDNA library. Although expression analysis offered insight into the numerous processes taking place in the germinating embryo, a great majority of the ESTs were not able to be identified by BLAST searches. The results indicate that many of the embryo ESTs is not represented in the public database.

INDEX WORDS: ESTs, High-throughput sequencing, Bioinformatics, Sorghum, Germination

ANALYSIS OF EXPRESSED SEQUENCE TAGS (ESTS) FROM SORGHUM BICOLOR GERMINATING EMBRYOS

by

ST PATRICK MARINHO REID

B.S., The University of Rochester, 1999

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2002

ANALYSIS OF EXPRESSED SEQUENCE TAGS (ESTS) FROM SORGHUM BICOLOR GERMINATING EMBRYOS

by

ST PATRICK MARINHO REID

Approved:

Major Professor: Lee Pratt

Committee: Gary Kochert
Russell Malmberg

Electronic Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
August 2002

ACKNOWLEDGMENTS

I would like to thank god for allowing me to beat the odds and for giving me the strength to carry the burden associated with a life within the light. I would also like to thank my family, especially my Segenie and Clinton Reid, Alpha and friends for all their support throughout this arduous process.

I would very much like to thank my advisor Dr. Lee Pratt for all his advice and support throughout the years. I would like to thank Karen Webb and Dr. Beth Olivares for this opportunity. I would like to thank Dr. Marie-Michéle Cordonnier-Pratt, Dr. Alain Gingle, Christine Marsala-Helfgott, Marc Sudman, Aynsely Eastman, Vickie Wentzel, Manish Shah, Dr. Chun Liang, Bob Sullivan and the rest of the Pratt-lab family. I would like to thank my committee members Dr. Russell Malmberg and Dr. Gary Kochert for all their advice and support and for passing me so I could graduate. I would like to thank the Botany/Plant biology staff.

The work completed here represents another of many stepping stones that support our rise to even greater scientific knowledge in this particular area of study. It also serves to open the door for further and perhaps more detailed research to be conducted. It is always necessary every once in awhile, to refocus the lens with which you use to appreciate life in order to not lose sight of the big picture. Research however indirect it maybe should always be focused on providing a building block from which even more beneficial work can be created, thus contributing to bettering life on this planet. It is with that in mind that for now I offer this, my contribution.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Laying the groundwork for EST technology

The traditional foundation of scientific investigation has been the hypothesis-driven approach. That is, one formulates a hypothesis following which, with the aid of any number of a variety of techniques from various disciplines, one sets out to test its validity. Although we know that this paradigm has obviously been a tremendous asset, we must understand that it does not stand alone as a means of furthering scientific knowledge. A data-driven approach represents another way to address questions in order to learn more about a given area of research. When using this approach, one creates and studies data with general goals in mind, then from that base formulates and tests any number of specific hypotheses.

With respect to biology, the roots of a data-driven approach lie in two converging paths (Zweiger and Scott, 1997). One is heralded by the discovery of the structure of DNA in 1953 by Watson and Crick, the other by the development of the semiconductor in 1959 by Hoerni and Noyce (Zweiger and Scott, 1997). With the discovery of the structure of DNA, and later the development of the central dogma of molecular biology, the foundation was laid for the development of Expressed Sequence Tag (EST) technology. With the advancement of semiconductor technology it was finally possible to produce, store and process the massive amounts of biological information required for

a data-driven approach. The work described here represents an example of this newer data-driven approach in biology.

The origins and uses of EST technology

The genomic sequences of higher organisms are only to a minor extent represented by protein-coding sequences (Sterky and Lundeberg, 2000). The vast majority of the information content of an organism's genome, however, is thought to reside in protein-coding sequences and sequences flanking them. Reverse transcription of messenger RNA (mRNA) to complementary DNA (cDNA) represents a direct way to obtain these sequences of interest. As a result, single-pass cDNA sequencing has become well established. In addition, the advent of high-throughput automated sequencing made it possible to select cDNA clones from plasmid cDNA libraries and determine quickly and at relatively low cost the DNA sequence of several hundred bases from one or both ends (Gerhold and Caskey, 1996). These partial cDNA sequences were called expressed sequence tags (ESTs) because each represents a sequence tag of a gene that is being expressed.

The EST approach has its roots in the early 1980s when it was observed that short stretches of cDNA sequences can be used to identify genes (Putney et al., 1983). The approach gained prominence in the early 1990s (Adams et al., 1991; Okubo et al., 1992), at which time it became the subject of controversy when the National Institutes of Health (NIH) tried to patent EST sequences. The attempt fueled a debate on the patentability of genes (Roberts, 1992). In addition, the EST approach was seriously challenged by some in the academic community, partially due to its simplicity (Zweiger and Scott, 1997).

Critics thought it lacked intellectual appeal because it required little foresight. However, since its development EST technology has proven to be a rapid, powerful and inexpensive route to gene discovery (Adams et al., 1991, 1992; Okubo et al., 1991; Matsubara and Okubo, 1993; Vasmatzis et al., 1998), thereby validating the enormous value of ESTs as a simple yet powerful tool.

EST sequencing can be used to complement genome sequencing in order to decode an organism's genetic information. Genome sequencing alone can be highly effective at gene identification when working with organisms such as *Mycoplasma genitalium*, whose genome is ~88% coding sequence (Fraser et al., 1995). When working with larger eukaryotic genomes, however, noncoding sequences account for a majority of the genome. Although noncoding sequences are known to contain regulatory information (Nicoloso et al., 1996; Filipowicz et al., 1999; Filipowicz, 2000) and serve a variety of other functions (Koop and Hood, 1994; Moore, 1996; Marcand et al., 1997), if one's aim is gene identification, coding sequence information is the primary tool.

Predictive algorithms can be used to aid in gene identification with genomic DNA. In general, predictive algorithms attempt to identify coding regions based on a variety of statistical analyses and also by detection of landmark sequences such as transcriptional signals (*e.g.*, TATA-boxes and transcription start sites). However, problems such as distinguishing pseudo genes from working genes and dealing with genes that overlap each other, appearing in different reading frames and on different strands can yield misleading results. Because they derive from mRNA, EST sequences represent a direct route for obtaining coding sequence information. A drawback to using EST sequences for genome-wide analysis, however, is that it is difficult to identify every

gene being expressed in all cell types and at all developmental stages for a given organism, and under every possible environmental condition. An alternative is to utilize both EST and genome sequencing. For example, EST sequences can be used to aid in defining intron-exon boundaries, which can help in finding coding regions in genomic sequences. Additionally, ESTs can be used along with predictive algorithms for finding coding regions in genomic DNA. Thus ESTs can be used to complement genome sequencing to identify an organism's genetic composition.

The EST approach has proven to be a valuable tool in several aspects other than gene identification. EST sequence data can be used to search for homologous genes in different organisms, identify chromosomal locations of genes, and analyze alternative splicing (Pandey and Lewitter, 1999; Ohlrogge and Benning, 2000). For example, ESTs proved to be a useful tool for identifying potential human homologues of *Drosophila* genes (Banfi et al., 1996). Comparative analysis of Arabidopsis and rice ESTs has identified gene families common to the two species (Ewing et al., 2000). More recently, using a doubled haploid mapping population, rice ESTs were mapped to chromosomal regions containing genetically defined resistance genes (Wang et al., 2001). Alternative splicing of pre-mRNA is a fairly common event, allowing for tissue-specific or temporal expression of a novel gene form, in some cases as a function of some environmental signal or stress. Alternative splicing allows one pre-mRNA to be processed into different mature forms in a cell (Brett et al., 2000). In some cases the small difference in transcript sizes are almost impossible to detect by conventional northern blotting (Pandey and Lewitter, 1999). EST data can provide instant sequence information about a splicing

event. Gibbs et al. (1998), for example, assembled several EST sequences in order to detect successfully alternative transcripts of a gene that encodes mouse heme oxygenase 2. In addition EST sequences have been used to estimate that the transcripts from ~35% of human genes are alternatively spliced (Hanke et al., 1999; Mironov et al., 1999).

On a larger scale, greater numbers of ESTs permit electronic transcriptional profiling, which can be performed by counting the number of ESTs for a given gene within an EST population (Pandey and Lewitter, 1999; Prade et al. 2001). This application for EST technology has proven to be a useful tool, for example, in identifying novel enzymes in specific plant metabolic pathways (Ohlrogge and Benning, 2000). Additionally, cDNA-based microarray technology is a novel expression-profiling tool that combines the two relatively young technologies of microarray and EST sequencing. The advantage of construction from cDNA clones for which ESTs have been generated, rather than from anonymous clones, is that it can yield more informative gene expression patterns (Richmond and Somerville, 2000).

ESTs can also provide information about 5′ and 3′ untranslated regions (5′ UTR and 3′ UTR). Among the non-coding regions, the 5′ UTR and 3′ UTR have been demonstrated experimentally to contain sequence elements crucial for many aspects of gene regulation and expression (Singer, 1992; Klausner et al., 1993; Wilhelm and Vale, 1993; Decker and Parker, 1994; Kaufman, 1994; McCarthy and Kollmus, 1995; Bashirullah et al., 1998). In pea seedlings, for example, the 5′ UTR of transcripts encoding light-harvesting chlorophyll-binding proteins is involved in high-fluence blue light-induced mRNA destabilization (Anderson et al., 1999). As a second example, Curie

and McCormick (1997) showed that the 5′ UTR of the pollen-specific LAT59 gene of tomato dramatically reduces mRNA accumulation without affecting mRNA stability.

Although the EST field got off to a rocky start owing to distractions related to patent issues and to skeptics who doubted its appeal, EST sequencing has become a widely used tool. The EST database (dbEST) created by GenBank (Boguski et al., 1993) has been one of GenBank's fastest growing divisions and further validates the usefulness of EST technology in the scientific community. EST sequencing is thus a very versatile and powerful tool that can be used in both large and small-scale gene analysis efforts.

Managing EST data

As we transit from the genomics era into what is being called by some the post-genomics era, we must be ever aware of the luggage we bring in the form of sequence data, which continues to be deposited into public databases at an enormous rate. In the past few years, public nucleotide databases have increased from millions to billions of bases per year (Benton, 1996). With that in mind, data management and analysis become of paramount importance, notably to researchers involved in their respective sequencing projects. Both proper software and hardware are needed for efficient data processing, assembly, annotation (Sterky and Lundeberg, 2000) and quality control. An effective database system is at the core of ensuring that those processes are handled efficiently.

For this research project we used a system for data processing, assembly and annotation based on an Oracle relational database management system (RDBMS). A database management system facilitates the storage of data in a database; additionally, it also facilitates the modification and extraction of the information contained in the

database. The advantages of a computerized management system over a manual filing system are its speed, accuracy and accessibility. The selection of the right database management system is dependent on the way one would like their data to be organized internally. The internal organization can affect how quickly and flexibly information can be obtained. For example, a flat-file database system is relatively simple in that an entire database can be contained within a single table; in contrast, a relational database system uses multiple tables to store information, using relationships among the tables to link data together (Fig. 1).

A relational database can be thought of as comprehensive tables of data. Each table is formally described and organized in such a way that the data contained in each table can be easily accessed. Additionally, because few assumptions are required about how data is related and how it can be extracted, the same database can be viewed in a diverse number of ways. A table in a relational database contains data fitted into predefined categories (Fig. 1). Each table (sometimes referred to as a relation) has one or more data categories in columns. In Figure 1, for example, this would be the attributes of libraries in one table (*i.e.*, species, library information) or of clones (*i.e.,* clone identification, clone annotation, library information) in the other. Each row in the table will contain a unique piece of information about a library or clone for the category defined by the column. An added advantage of relational databases is that they are easy to extend. A new data category can be added without requiring the modification of existing categories. The key feature of a RDBMS is its ability to store data in two or more tables and enable the user to define relationships among tables. The link between the tables is based on one or more field values common to both tables (*e.g.,* 'Library

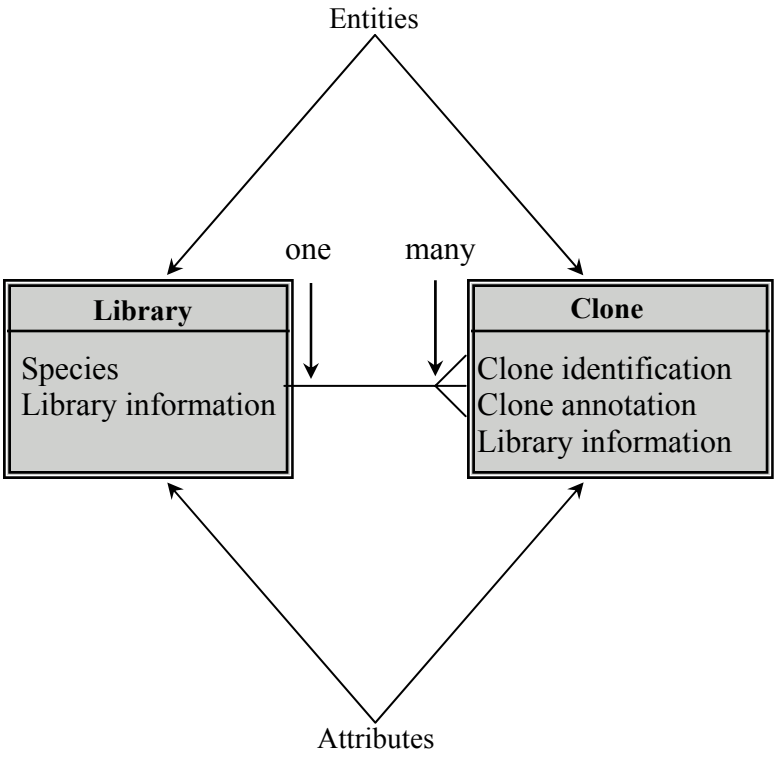
information' in Fig. 1); the link defines the relationships among tables. Internal organization of data in this way provides a user with rapid access and flexibility when extracting data.

One of the main goals when working with large quantities of data is to store and organize it in a way that allows for maximum accessibility for analysis. The use of multiple related tables gives an RDBMS the ability to efficiently store data while requiring few assumptions about how the data is related or how it will be extracted from the database. Storage in this form allows for flexibility when posing database queries.

Bioinformatics

Bioinformatics is an interdisciplinary field that broadly speaking, describes any use of computers to handle biological information. In relation to this research project, the definition of bioinformatics can be specified to mean the integration of computer science with EST sequencing, with the aim of obtaining answers to biological questions. The term 'bioinformatics' was coined in the early 1990s (Boguski, 1998). It is important to note that although the term bioinformatics is fairly new, scientists have been building databases, developing algorithms and making biological discoveries by sequence analysis since the 1960s (Dayhoff, 1969; Hirschberg, 1975;Hunt and Szymanski, 1977). The current concept of bioinformatics though, can be described as the convergence of two technological revolutions: the explosive growth in biotechnology, paralleled with the explosive growth in information technology (Murray-Rust, 1994). Advances in DNA sequencing led to the production of massive quantities of data, greatly increasing the size of databases such as GenBank (Benson et al., 2000). A similar increase was seen in

Figure 1. An example of the tables in which a relational database stores data. Relationships 'connect' or link the data in one table to the data in another. Each table is composed of entities and attributes. Entities are the primary data objects about which information is to be collected. An entity can be thought of as the subject to be covered; a relationship thus represents an association between entities. In this example, the entities for these tables are 'library' and 'clone. Attributes describe the entity with which they are associated. In this example, attributes are 'species', 'library information', 'clone identification', 'clone annotation', and 'library information'. They represent the data that is to be kept about entities and are thus arranged as columns in a table; each row in a table contains unique information about the entity. The relationship here would be: libraries have clones, or clones are generated from libraries. The lines in between the tables indicate the relationship. The most common relationship type, one-to-many, is shown. Here one record in one table is related to many records in another table. For example, one library will have many clones. The end of the relationship line that branches out denotes the many end of a one-to-many relationship.


Entities
one
many
Library
Species
Library information
Clone
Clone identification
Clone annotation
Library information
Attributes

microprocessors as the number of transistors per chip increased, thus obeying Moore's Law of the exponential growth rate of computing power. Along with those advances came the development of bioinformatic tools, which helps one to extract meaningful biological information from massive amounts of data.

Bioinformatic tools are required for efficient processing, assembly, and annotation of data. Base-calling and screening programs are tools that can be used to process raw sequence data. Phred (Ewing et al., 1998) is commonly used for calling bases. Phred also assigns to each called base a quality value, which corresponds to an error probability; the lower the error probability the higher the quality value. Screening of vector sequence is also an important component of data processing. The program CrossMatch (http://www.phrap.org) is a vector-screening program often used to analyze sequences generated by Phred. CrossMatch compares the generated sequences to vector sequences, 'masking' regions of the generated sequences that correspond to vector sequence.

Processed sequences can then be clustered or assembled using any of a number of programs, including CAP3 (Huang, 1996), TIGR assembler (Sutton et al., 1995) and Phrap (http://www.phrap.org). Phrap works well in concert with Phred and CrossMatch. Phrap assembles overlapping sequences into contigs. A contig is defined here as a set of overlapping sequences from which a consensus sequence can be derived. In contrast, singletons are sequences that appear to share no similarity to other sequences in the database. Additionally, Phrap can also generate what is known as contigs of one, or singleton contigs. These contigs of one are individual sequences that matched other sequences but could not be consistently assembled together with them. Phrap also works

well with the graphical interface Consed (Gordon et al., 1998), which is a useful visualization tool for analysis of Phrap assemblies.

Sequence annotation from a bioinformatics standpoint involves the use of sequence similarity searches. Using these searches for annotation assumes that similar sequences will have similar function. The most widely used programs for sequence similarity searches include the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990), FASTA (Pearson and Lipman, 1988) and the Smith-Waterman algorithm (Smith and Waterman, 1981). The BLAST programs (Table I), which are fastest among the three in terms of search speed, have been written to compare protein or DNA queries with protein or DNA databases in any combination (Altschul et al., 1997). The original intended use for BLAST was for individual comparisons of a query sequence to a single database. Although BLAST is still used for individual searches, it is also now used for comparing large numbers of sequences to multiple databases. High-throughput BLAST (HT-BLAST) (http://www.sgi.com/solutions/sciences/chembio/resources/papers/HTBlast/HT_Whitepaper.html) is a modified version of BLAST optimized for searching multiple query sequences against multiple databases.

Although the focus here was on a few of the bioinformatics tools that can be used for efficient data processing, assembly and annotation, there are a variety of others available. The selection of the proper bioinformatics tool depends on the purpose one would want it to serve.

Table I. A listing of the BLAST programs. The selection of a particular program depends upon two factors, the type of sequence used to query the database and the type of database one would like to query.

Program	Query type	Database
BLASTN	DNA	DNA
BLASTP	Protein	Protein
BLASTX	trDNA[a]	Protein
TBLASTN	Protein	trDNA
TBLASTX	trDNA	trDNA

[a]- translated DNA

Sorghum bicolor

Sorghum (*Sorghum bicolor*) is a member of the tribe Andropogoneae within the grass family Poaceae (Clayton, 1987). The Poaceae is comprised of greater than 10,000 species, and collectively accounts for 60% of the world's food production (http://www.wri.org/biodiv/foodcrop.html), making it economically the most important plant family.

As a member of the Poaceae, sorghum itself is a major source of nutrition in developing countries, ranking fifth in importance among the world's grain crops (Doggett, 1988). The ability of sorghum to exhibit unusual tolerance to adverse environments (Doggett, 1988; Ludlow and Muchow, 1990), along with its small genome (760 Mbp; Arumuganathan and Earle, 1991) relative to most grasses, with the exception of rice (*Oryza sativa,* 466 Mbp; Yu et al., 2002), makes it an excellent model system for genome-scale investigations. In addition high-resolution genetic and physical maps have been generated (Hulbert et al., 1990; Chittenden et al., 1994; Pereira et al., 1994; Ragab et al., 1994; Xu et al., 1994; Peng et al., 1999; Draye et al., 2001).

Within the family Poaceae, genome organization and size varies greatly. Chromosome numbers vary from 2n=4 in the species *Zineria biebersteiniana* (Bennett et al., 1995), to 2n=266 for the polyploid *Poa litorosa* (Hair and Beuzenberg, 1961). Their genome size also varies greatly. For example, the genomes of sorghum (7.6×10^8 bp) and bread wheat (1.7×10^{10} bp) (Arumuganathan and Earle, 1991) differ roughly by a factor of 23. Comparative genomics then, using these examples, represents an excellent tool for analysis of genome structure. With its comparatively small genome sorghum can, for

example, be used as a reference for gene isolation from larger genomes. In addition, the synteny among the cereals further enhances the utility of sorghum for gene isolation.

The utility of investigating smaller genomes for discovery of genes in much larger genomes has been demonstrated in plants (Dunford et al., 1995; Bennetzen and Freeling, 1997; Guimaraes et al., 1997). As an example, the fact that the smaller and simpler genome size of sorghum as compared to maize does not correlate with morphological and physiological complexities of the organisms (Laurie and Bennett, 1985; Bennetzen et al., 1994; SanMiguel and Bennetzen, 1998), as well as the observed co-linearity between the maize and sorghum genomes (Hulbert et al., 1990; Berhan et al., 1993; Avramova et al., 1996), suggest sorghum would be an excellent model system for characterization of the maize genome. A high level of homology and synteny has also been observed between sorghum and rice. Comparisons of genetic maps have also revealed co-linearity between sorghum and other grasses (Moore et al., 1995; Devos and Gale, 1997; Keller and Feuillet, 2000). Physical analysis of the sorghum genome will, therefore, facilitate the cloning of genes associated with many aspects of plant domestication and crop productivity in the larger genomes of other grasses (Draye et al., 2001).

Sorghum germinating embryos

I have chosen to examine gene expression in the embryo of sorghum seeds germinated for 24 hr. In addition to the embryo, a seed also generally includes a seed coat and endosperm. The seed coat is formed from the integument of the ovule and hardens to form a protective shell around the embryo and endosperm once the seed separates from the parent plant. The endosperm functions as a nutritive tissue for the

developing embryo, accumulating nutrients such as starch and fats that serve as food reserve during germination. The endosperm and seed coat generally play supportive roles during development. The embryo, however, being the part of the seed that will develop into the plant, expresses the necessary information leading directly to the development of an adult plant. Thus the embryo in this case is of primary importance when analyzing gene expression during early plant development.

Germination is a crucial phase in the life cycle of plants. It is characterized as the period in development beginning with imbibition of the quiescent dry seed, and ending with the elongation of the embryonic axis (Bewley and Black, 1994). Visually, germination is complete when the radicle emerges from the seed coat. During this period, a seed must adapt its metabolic and developmental program to prevailing environmental conditions (Holdsworth et al., 1999). Germination is thus a carefully coordinated complex of events (Tschining et al., 2000) that includes among other things the expression of genes involved in respiration and macromolecular synthesis. Although the aforementioned events are also common to other developmental stages, it has been suggested that many of the genes expressed during germination are distinct from those expressed during later development (Hughes and Galau, 1989; Kermode, 1990; Berry and Bewley, 1991). Germination-specific genes associated with reserve mobilization during seedling growth (Jacobson et al., 1995; Kermode, 1995) are an example of such genes.

All the cellular machinery for resumption of metabolic activity is already present in the dry quiescent embryo (Sen et al., 1975; Spiegal and Marcus, 1975; Delseny et al., 1977; Cuming and Lane, 1978, 1979; Harris and Dure, 1978; Sanchez and Aquilar, 1984). Therefore, upon imbibition metabolic activity rapidly resumes, becoming

detectable within only minutes after the onset of germination. In addition, preformed mRNAs are also present within the embryo (Bewley, 1997). Some of these are residual messages associated with previous developmental processes (Comai and Harada, 1990; Lane, 1991). Examples include messages encoding proteins that are important during seed maturation and drying, such as Late Embryogenesis Abundant (LEA) proteins. As germination proceeds, however, the residual messages are degraded (Jiang and Kermode, 1994; Han et al., 1997) and replaced with others encoding proteins required for continued germination. Although some of the mRNA being transcribed encodes proteins that are essential for the support of normal cellular metabolism, the production of germination-specific proteins also occurs during this period (Hughes and Galau, 1989; Goldmark et al., 1992; Johnson et al., 1995). Thus, monitoring gene expression after a 24-hr germination period will permit the detection of germination-specific genes as well as common or constitutively expressed genes.

EST sequences generated from sorghum embryos germinated for 24 hr permit analysis of the transcripts present early in germination. During this period of development, cell enlargement and division are initiated in the embryo, as transcripts are produced that influence its organization and development into an adult plant. Utilizing EST technology together with a variety of bioinformatic tools will permit the analysis of the transcripts present and thus indirectly the proteins required for these events to occur. In particular, bioinformatic tools for sequence processing, assembly and annotation are key in data analysis. The combination of EST sequencing technology and bioinformatic tools offers both a relatively fast and efficient way of identifying transcripts present early

in germination and the identity of the proteins they encode, and also discovering potential heretofore unknown gene products involved in embryo development.

Project goals

To create the database and address specific areas, the project was divided into two overlapping phases. The first phase, or wet-lab portion, refers to the laboratory bench work. It spanned the range of activities beginning with the isolation of embryos through sequencing the cDNA clones derived from them. The wet-lab portion then gave way to the bioinformatics portion, which required a wholly computational approach for data processing, storage, access and analysis. Following production of ESTs for an embryo library, there were five specific areas chosen for exploration. Those areas are:

1. Database quality control.
2. Annotation and classification
3. Sequence clustering
4. Identification of the relative abundance of full-length coding cDNAs
5. Identification of possible alternative splicing events within the data set

CHAPTER 2

MATERIALS AND METHODS

Embryo Preparation

Sorghum (*Sorghum bicolor* (L.) Moench, BTx623) seeds were collected from a greenhouse and germinated in Petri dishes (100 x 15 mm) on white filter-paper discs (Whatman No. 29; 9-cm diameter). The discs were evenly wetted with distilled water. Germinating seeds were maintained in darkness for 24 hr in an incubator set at 25°C and 100% humidity, following which they were milled with a Quaker model 4-E plate mill (Clinton Separators Inc., Warminster, PA). Milled seeds were dropped directly into liquid nitrogen in order to minimize RNase activity. Once ground, frozen endosperm and embryo were easily separated from one another using a sieve (0.84-mm opening) kept in liquid nitrogen. The endosperm, which was ground into a fine powder, passed easily through the sieve while the intact embryo remained. The embryos were then collected and ground in liquid nitrogen using a mortar and pestle. The resulting fine powder was placed into a labeled specimen container (Fisher Scientific, Pittsburgh, PA) and stored at -80°C.

RNA Isolation

Total RNA was prepared from embryos using a phenol/chloroform method. RNA was extracted into TLES buffer, which consisted of 10 mM Tris-Cl (pH 8.0, 22°C), 10 mM LiCl, 1 mM EDTA (pH 8.0), and 1% SDS. Buffer and buffer-saturated phenol were

first combined together at a ratio of 0.8 ml each per 0.25 g of embryos, yielding a total mixture of 1.6 ml per 0.25 g of embryos. Embryos were then added to the combination of buffer and phenol and vortexed for 2 min at room temperature, after which they were homogenized for 1 min using an Ultra-Turrax (IKA Works, Inc., Wilmington, NC). To the extract, 0.75 ml of chloroform was added followed by vortexing for another 2 min. The mixture was then transferred to 1.5-ml microfuge tubes at a volume of 0.5 ml per tube and centrifuged at 16,000 *g* for 10 min at 4°C. Supernatants were transferred to new microfuge tubes and an equal volume of 4 M LiCl was added. The contents were mixed by inversion and placed on ice overnight at 4°C. The following day the tubes were centrifuged at 16,000 *g* for 30 min at 4°C, supernatants were discarded, and pellets were air dried on the bench top. Each dried pellet was dissolved in 250 µl of H_2O that had been treated with DEPC (diethyl pyrocarbonate). The dissolved pellet was then vortexed for 30 sec, after which 40 µl of 2.5 M Na-acetate was added, followed by the addition of 0.6 ml of 100% ethanol. The samples were mixed by inversion and incubated for 1 hr at -20°C. Following incubation, tubes were centrifuged at 16,000 *g* for 20 min at 4°C after which supernatants were carefully decanted. To each pellet 0.5 ml of 70% ethanol was added followed by vortexing for 30 sec. Resuspended samples were centrifuged at 16,000 *g* for 7 min at 4°C, supernatants were discarded, and tubes were inverted onto paper towels in an RNase free area and allowed to air dry. Subsequently, the content of each tube was dissolved in 45 µl DEPC-treated H_2O. The samples were stored at -20°C.

All samples were evaluated by electrophoresis in a 1% formaldehyde gel to verify the integrity of the RNA that had been isolated. RNA concentrations were determined with the use of a Hewlett-Packard 8452 diode array spectrophotometer. RNA from 10

preparations were combined and a total amount of 1.15 mg of RNA was shipped to Stratagene for cDNA library construction. The concentration of each sample was determined by using the formula: μg/ml = A_{260} x dilution x 40 μg/ml, where A_{260} is the absorbance at 260 nm for a 1-cm optical path. Samples were diluted with DEPC-treated H_2O (3:500 v/v). Forty represents the conversion factor for RNA at 260 nm.

Mass Excision

An unamplified, unidirectional Uni-ZAP XR cDNA library using *Eco*RI and *Xho*I cloning sites was commercially prepared from 1.15 mg of total RNA (Stratagene, La Jolla, CA). Stratagene used a drip column containing Sepharose CL-2B gel filtration medium for size fractionation of the cDNAs. Two separate library fractions were received. The titers were 2.1 x 10^6 plaque-forming units/ml for fraction 1, and 1.4 x 10^7 plaque-forming units/ml for fraction 2. A total of 12 clones were randomly selected by Stratagene to determine the insert size range for fraction 1. The average insert was 2.0 kb, with a range of from 1.4 to 2.8 kb. Fraction 2 contained smaller inserts of unidentified size. An amplified fraction of the library was also included, but was not used, as it would be expected to preferentially represent those clones that replicate faster. Instead, for analysis the presumably less biased, unamplified fraction 1 was utilized.

Plasmid-containing bacterial colonies were obtained using a modified version of Stratagene's Uni-Zap mass excision protocol. Separate cultures of XL1-Blue MRF′ and SOLR *Escherichia coli* cells were grown overnight in LB broth supplemented with 2% (w/v) maltose, 10 mM $MgSO_4$, and antibiotic. Kanamycin at a final concentration of 50 μg/ml was added to the SOLR cells, while tetracycline at a final concentration of 15

μg/ml was added to the MRF′ culture. The cultures were grown overnight in a shaker at 250 rev/min for 16 hr at 37°C. Cultures were diluted to an OD_{600} between 0.1 and 0.4 with 50 ml LB broth containing the same supplements as mentioned above, including antibiotic. Diluted cells were incubated under the previously stated overnight growth conditions with periodic measurement of OD_{600} until it reached 1.0. Optical density readings were obtained from a 3-ml sample in a 1-cm light-path cuvette and measured using a Varian Techtron UV-Vis spectrophotometer (Varian Inc., Palo Alto, CA). In sterile Oak Ridge centrifuge tubes 5 ml of each culture was centrifuged at 1000 *g* for 5 min. Supernatants were discarded and the pellets resuspended in 5 ml of 10 mM $MgSO_4$.

In a 2.0-ml screw-cap microfuge tube a mixture was prepared consisting of MRF′ cells, unamplified phage library and ExAssist helper phage, which aids in the infection process. The mixture was prepared at a ratio of 10^4 ExAssist plaque forming units (pfu) to 10^3 MRF′ cells to 1 library pfu. The mixture was incubated for 15 min in a water bath at 37°C to allow for transfection, after which an additional 1 ml of LB medium was added. The samples were incubated in a shaker at 250 rev/min for 2.5 hr at 37°C. Immediately following the incubation, the samples were placed in a water bath at 68°C for 20 min. The tubes were then centrifuged at 1000 *g* for 10 min and the resulting supernatants, which contained excised phagemids, were transferred to sterile micro-centrifuge tubes.

In a sterile microfuge tube 2 μl of phagemids were mixed with 200 μl of SOLR cells and incubated in a water bath at 37°C for 15 min. For plating, 30 μl of these transfected SOLR cells was combined with 270 μl of LB medium. From that mixture 100 μl was placed onto NZY agar plates (Fisher Scientific, cat no. 12-565-296)

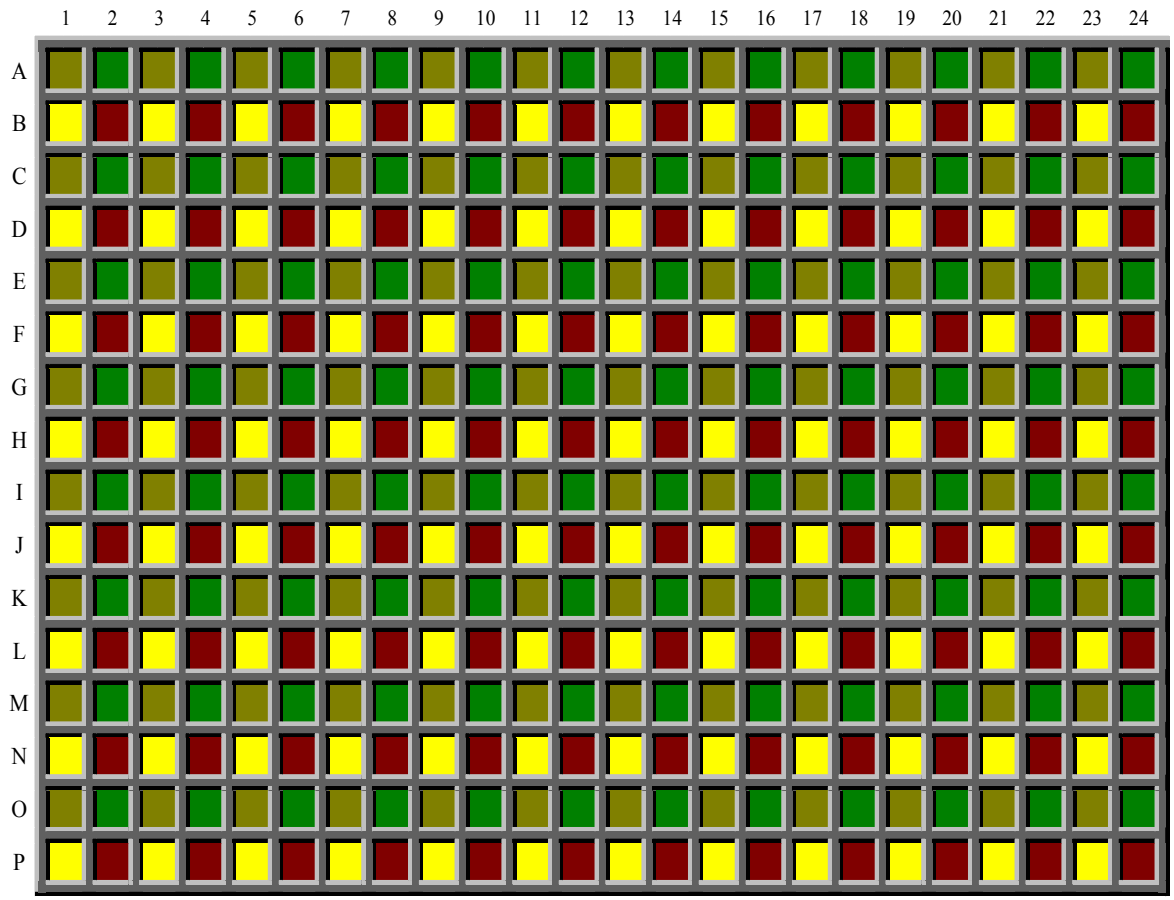
containing ampicillin at a concentration of 150 μg/ml. The plates were incubated overnight at 37°C.

Picking, Inoculating and Archiving

Using an 8-channel pipette (Finnpipette digital MCP, Franklin, MA), 100 μl of freezing medium was placed into each well of 384-well plates (Fisher Scientific, cat. no. 12-565-294). Freezing medium consisted of 180 ml TB medium, 20 ml salts (0.17 M KH_2PO_4, 0.72 M K_2HPO_4), 14 ml of 100% glycerol and 321 μl of 100 mg/ml ampicillin. Bacterial colonies from the NZY agar plates were picked with sterile toothpicks that were then used to inoculate wells of the 384-well plates. The inoculated plates (Marsh, cat. no. AB-0718) were put into a HiGro shaking incubator (Gene Machines, San Carlos, CA) at 520 rev/min for 20 hr at 37°C. After the 20-hr incubation the 384-well plates were stored at -80°C until they were ready to be used for plasmid isolation.

Frozen 384-well plates were thawed on a Titer Plate Shaker (Lab-Line Instruments Inc., model # 4625, Dubuque, IA) at a speed of 5 for 30 min. While the plates were thawing, deep 96-well blocks (Beckman cat. no. 140504) and 96-well v-bottom plates (Dynex cat. no. 62402-914) were filled with medium. Using a Repeator pipette (Eppendorf Repeator Plus, Westbury, NY) a total of 1.5 ml of TB with added salts (0.15 mg/ml salts) was added to each well of a deep-well block. With an 8-channel pipette a total of 150 μl of freezing media was added to each well of a v-bottom plate. Bacterial cells from the thawed 384-well plates were transferred to the deep-well blocks and v-bottom plates. Transferring bacterial cells from a 384-well format to a 96-well format involved dividing the 384-well plate into four quadrants (Fig. 2); thus, a total of

Figure 2. Diagram of a 384 well-plate. Each shaded square represents a well on the plate. The color of each square designates the quadrant to which it belongs (*e.g.*, A1& C1 belong to quadrant 1, A2 & C2 to quadrant 2, B1 & D1 to quadrant 3, and B2 & D2 to quadrant 4). As a result a single 384 well-plate can be subdivided into four 96-well plates, as each quadrant has a total of 96-wells. Thus, when using a Hydra96 for liquid transfer, samples from each quadrant can be transferred to separate 96 well-plates in just four operations, one per quadrant.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P

four deep-well blocks (one for each quadrant) and eight v-bottom plates (two for each quadrant) were needed for each 384-well plate. Using an 8-channel pipette 1 µl of bacterial cells from each well in each quadrant was transferred to the corresponding well of one deep-well block and two v-bottom plates. V-bottom plates were covered with gas-permeable seals and grown in a HiGro at 520 rev/min for 16 hr at 37°C. Following incubation the gas-permeable seals were replaced with foil seals, and the duplicate plates were placed in separate -80°C freezers for permanent, archival storage. After inoculation deep-well blocks were also covered with gas-permeable seals but grown in a shaking incubator at 250 rpm for 20 hr at 37°C. Bacterial cells grown in the deep-well blocks were used for plasmid isolation.

Plasmid Isolation

Plasmids were isolated using a modified version of an alkaline lysis method (http://www.genome.ou.edu/dsisol_seq.html). Deep-well blocks containing bacterial cultures were centrifuged (Jouan, CR422, Winchester, VA) at 1740 *g* for 5 min at 4°C. Following centrifugation the supernatants were carefully discarded. The blocks were placed inverted on paper towels for 2 min. With the use of a Hydra96 (Robbins Scientific, Sunnyvale, CA), which is a 96-channel robotic liquid transfer system (290 µl syringe size), 200 µl of TE-RNaseA consisting of 50 mM Tris-Cl (pH 7.6, 22°C), 10 mM EDTA (pH 8.0) and 20 µg/ml RNaseA, was added to each. The blocks were covered with acetate plate seals (Dynex Technologies, Chantilly, VA) and placed on a Titer Plate Shaker at a setting of 7 for 45 min. Cells were lysed by adding to each well 200 µl of 0.2 N NaOH and 1% SDS. The blocks were re-sealed and again placed on the Titer Plate

Shaker at a setting of 7 for 45 min. Subsequently, 200 µl of 3 M K-acetate (pH 5.0) was added. The blocks, which were covered with acetate seals, were held firmly by hand on a Fisher Vortex Genie 2 (Fisher Scientific cat. no. 12-812) and vortexed vigorously for 3 min at maximum speed. The blocks were then placed in a shaking incubator at 250 rpm for 20 min at 37°C, and afterward in a -80°C freezer overnight. The following day the blocks were placed under a laminar flow hood and thawed at room temperature. Once they were fully thawed (2.5-3 hr), they were centrifuged at 2870 *g* for 45 min at 4°C. Using a Hydra, 200 µl of supernatant from each well was transferred from the deep 96-well Beckman blocks to corresponding wells of deep 96-well Costar blocks (Fisher, cat. no. 09-761-117). Plasmid DNA was precipitated by addition of 500 µl of 95% ethanol to each well using a Repeator pipette. After thorough mixing, blocks were centrifuged at 2870 *g* for 30 min at 4°C. Supernatants were decanted and discarded, after which the blocks were set inverted for 1 min on paper towels. Next, 250 µl of 70% ethanol was added to each pellet and blocks were again centrifuged at 2870 *g* for 15 min at 4°C. Supernatants were decanted and the blocks set topside up to dry in an incubator for 30 min at 37°C. Each sample was dissolved by adding 150 µl of distilled water and mixing with the titer plate shaker for 15 min at a setting of 6. Samples were then transferred in their entirety to v-bottom plates and stored at -20°C.

Sequencing

Plasmid DNA was cycle sequenced from both the forward (3′) and reverse (5′) ends using a thermal cycler (Applied Biosystems, GeneAmp 9700, Foster City, CA). Using a Hydra, 2 µl of double-distilled H_2O was added to each well of two 384-well

cycle sequence plates, after which 2 µl of DNA from 4 v-bottom plates was transferred into each well of the cycle sequence plates (Perkin Elmer-Applied Biosystems, cat. no. 4305505). The Hydra was programmed to follow the quadrant system described in Figure 2, with each v-bottom plate corresponding to a quadrant on the 384-well cycle sequence plate. A $1/12^{th}$ cycle sequencing master mix was prepared by combining 214 µl double-distilled H_2O, 130 µl of DMSO, 532 µl of 5x buffer [400 mM Tris-Cl (pH 9.0, 22°C), 10 mM $MgCl_2$], 56 µl of primer and 268 µl of Big Dye Terminator Cycle Sequence Ready Reaction version 2.0 (Perkin Elmer-Applied Biosystems, Foster City, CA). Master mix preparations were primer specific. For 5′ sequencing reactions, 56 µl of 300 pmol/µl JenRev primer (5′-CAGGAAACAGCTATGACC-3′) was added. For 3′ sequencing reactions, 56 µl of 450 pmol/µl of anchored polyT primer (5′-TTTTTTTTTTTTTTTTTTTT(C/G/A)-3′) was added. In a few cases 56 µl of 150 pmol/µl of T7 primer (5′-TAATACGACTCACTATAGGG-3′) was used for 3′ sequencing. Using a Repeator pipette, 3 µl of the master mix was added to each template DNA for thermal cycling, yielding a final volume of 7 µl. The thermal cycling program was as follows: 99 cycles of 96°C for 10 sec, 50°C for 5 sec, and 60°C for 4 min, followed by a hold at 4°C.

Durapore 96-well filtration plates (Millipore, cat. no. MADV N65 50) containing hydrated Sephadex G-50 (Amersham Pharmacia, cat. no. 17-0043-02) were used for removing unincorporated nucleotides and other low molecular weight contaminants. Fully hydrated G-50 plates were centrifuged at 960 *g* for 2 min at 4°C. One hundred microliters of distilled autoclaved water was added to each well and plates were centrifuged again at 960 g for 2 min at 4°. While the G-50 plates were being centrifuged

10 µl of distilled autoclaved H_2O was added to each well of the 384-well cycle sequence plates. Using a Hydra, the contents from the cycle sequence plates were then transferred to the G-50 plates. With the use of plate adapters, a new 96-well MicroAmp optical reaction plate (Perkin Elmer-Applied Biosystems, Foster City, CA) was positioned directly under the G-50 plates. The plates were centrifuged again at 960 *g* for 2 min at 4°C, such that sequencing reaction products were transferred to corresponding wells in the new MicroAmp plates. The samples were lyophilized in a SpeedVac Concentrator (Savant Instruments Inc., Holbrook, NY) for 3 hr. Upon completion the plates were covered with foil seals and stored at -20°C.

Before placement in an ABI Prism 3700 sequencer (Perkin Elmer-Applied Biosystems, Foster City, CA) the contents of a set of four 96-well plates that had previously been subdivided for G-50 cleanup were recombined into a 384-well plate. Prior to the samples being recombined, a 12-channel stepper pipette (Finnpipette digital MCP, Franklin, MA) was used to add 20 µl of distilled autoclaved H_2O to each well of a MicroAmp plate containing the lyophilized samples. The plate was then centrifuged for 30 sec at 960 *g*, after which it was placed on a Titer Plate Shaker at a setting of 6 for 15 min. The samples were then recombined from a 96-well format to a 384-well format. Using a Hydra, 10 µl of sample was transferred from four 96-well plates to one 384-well plate using the quadrant system format in Figure 2. The 384-well plate was covered with a foil seal and centrifuged for 30 sec at 960 *g*, after which it was placed into the ABI Prism 3700 for sequencing. For sequencing a few parameters were changed for optimization. The cuvette temperature of the ABI Prism 3700 was decreased from the default value of 40°C to 30°C, the injection time was increased from 30 to 60 sec, and the

run time was reduced from 115 min to 105 min. All other sequencing parameters were set at the default values.

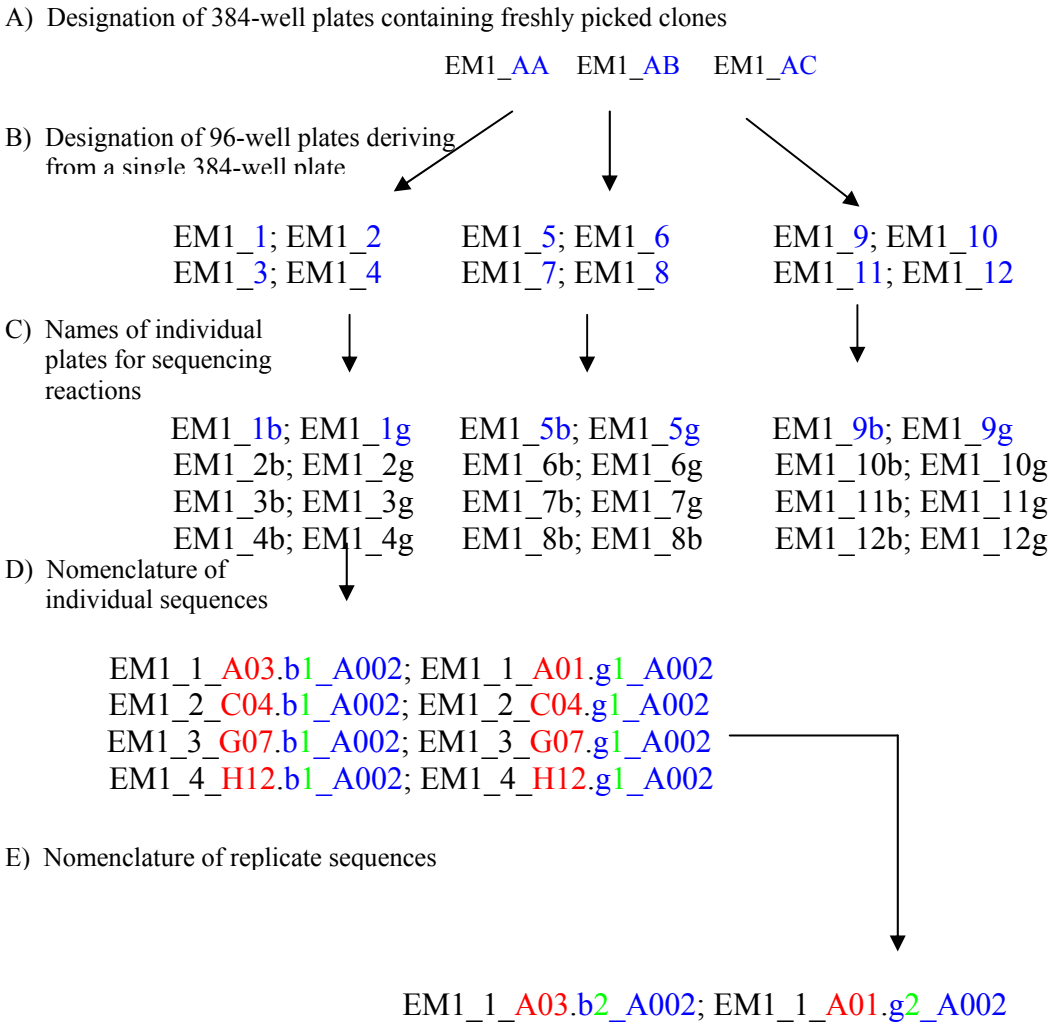
Clone Nomenclature

All picked clones were placed into labeled 384-well plates (Fig. 3A). The label denoted the library (EM1 = the first embryo library) and the order in which the plates were picked (*e.g.*, AA, AB, and so on). Plates selected for plasmid isolation were subdivided into four 96-well plates as detailed in Figure 2. These 96-well plates were labeled numerically in sets of four, providing first the library name (EM1), followed by an underscore and the plate number (Fig. 3B). Once plasmid isolation was complete, the plates to be thermal cycled had an additional descriptor added (Fig. 3C). Sequences obtained using a reverse primer were normally 5′ ESTs, while those obtained with a forward primer were normally 3′ ESTs, identified as 'b' or 'g,' respectively. To identify individual sequences, the well location was placed following the plate number and directly before the indication of sequence direction (Fig. 3D). The number following the indication of sequence direction was used to distinguish replicate sequences from the same plate (Fig. 3D). The code (here A002) placed directly after the numeral indicating replicate number identifies the organism from which the sequenced cDNA came.

Sample Tracking Forms

A number of paper records were kept for each block prepared during the plasmid purification process. The records provided needed quality control, serving as a uniform way of monitoring the dates of initiation and completion of a particular procedure. In

Figure 3. The figure illustrates how clone identities were tracked. EM1 refers to the embryo library. (A) When colonies were picked and placed into 384-well plates, the identifier of each plate was alphabetical (*e.g.*, AA, AB, AC). Hence, plates were labeled EM1_AA, EM1_AB, and so on. (B) For the plasmid isolation procedure, and as illustrated in Figure 2, each 384-well plate was subdivided into four 96-well plates that were identified numerically. For example, EM1_AA was used to produce plates EM1_1, EM1_2, EM1_3 and EM1_4. (C) Each plate of plasmid DNA was used to prepare two plates for sequencing, one of which was labeled 'b,' for sequences from the reverse or 5′ end (*e.g.*, EM1_1b), and the other 'g,' for sequences from the forward or 3′ end (*e.g.*, EM1_1g). (D) Individual sequences were identified by the well from which they were derived (shown here in red) followed by a letter (b or g) indicating from which end the sequence was obtained. (E) In addition, individual sequences were labeled to identify replicate sequences from the same plate (shown in green) in the event a plate was sequenced more than once. Thus, if EM1_1_A03 were sequenced a second time, the resulting sequences would be EM1_1_A03.b2_A002 and EM1_1_A01.g2_A002 as indicated. The letter followed by three numerals (in this case A002) added at the end of the sequence identifies the species (here *Sorghum bicolor*) from which the sequence was derived.


A) Designation of 384-well plates containing freshly picked clones
EM1_AA EM1_AB EM1_AC
B) Designation of 96-well plates deriving from a single 384-well plate
EM1_1; EM1_2
EM1_3; EM1_4
EM1_5; EM1_6
EM1_7; EM1_8
EM1_9; EM1_10
EM1_11; EM1_12
C) Names of individual plates for sequencing reactions
EM1_1b; EM1_1g
EM1_2b; EM1_2g
EM1_3b; EM1_3g
EM1_4b; EM1_4g
EM1_5b; EM1_5g
EM1_6b; EM1_6g
EM1_7b; EM1_7g
EM1_8b; EM1_8b
EM1_9b; EM1_9g
EM1_10b; EM1_10g
EM1_11b; EM1_11g
EM1_12b; EM1_12g
D) Nomenclature of individual sequences
EM1_1_A03.b1_A002; EM1_1_A01.g1_A002
EM1_2_C04.b1_A002; EM1_2_C04.g1_A002
EM1_3_G07.b1_A002; EM1_3_G07.g1_A002
EM1_4_H12.b1_A002; EM1_4_H12.g1_A002
E) Nomenclature of replicate sequences
EM1_1_A03.b2_A002; EM1_1_A01.g2_A002

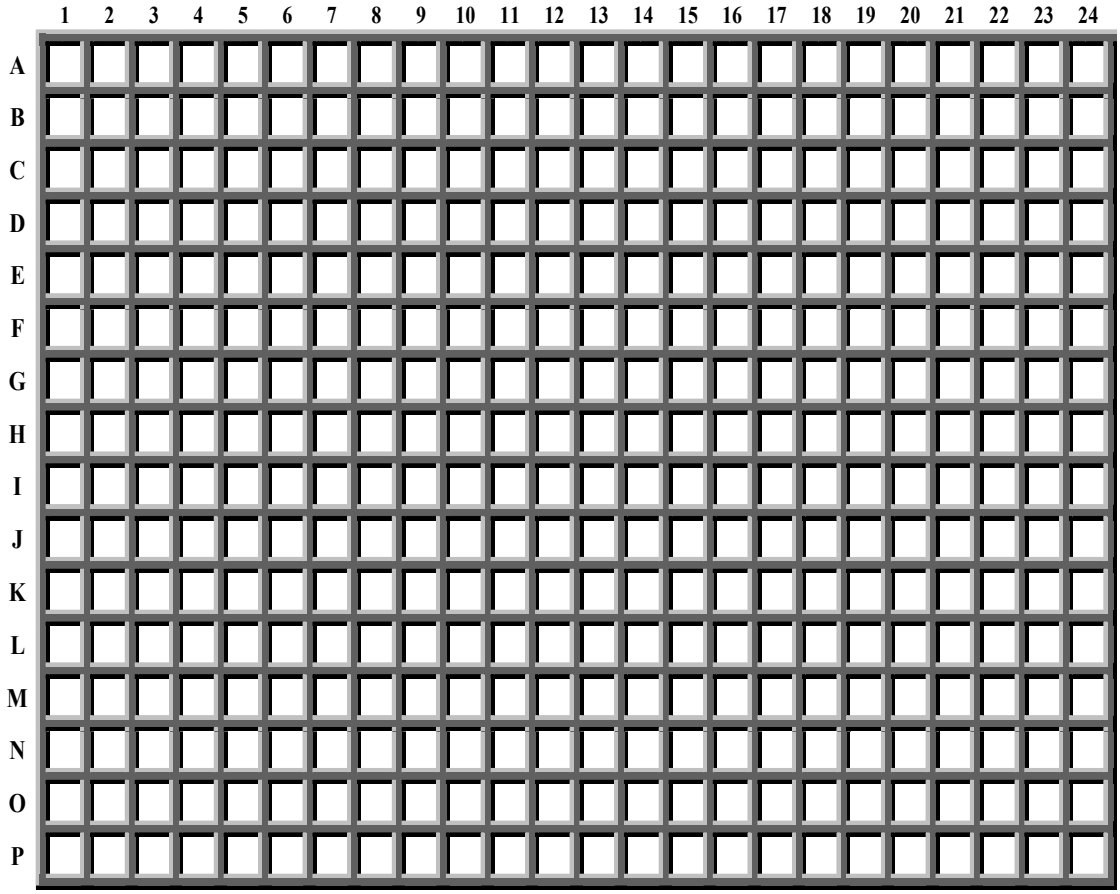
addition, one had the opportunity to track the progress of each step, noting for example when a mishap could have occurred. The first form monitored the growth of bacterial cells picked into 384-well plates; wells in which there were little or no growth could be identified and their locations denoted on the form (Fig. 4). A plasmid preparation form monitored the steps from the inoculation of the deep 96-well blocks and 96-well v-bottom plates to the last plasmid isolation step (Fig. 5). A cycle-sequence form was used to monitor the thermal cycling and G-50 cleanup procedures (Fig. 6). Each cycle-sequence form was primer-specific and allowed one to enter the date thermal cycling was initiated, to identify the thermal cycler used, and to note the thermal cycle initiation and completion times; additionally, the initiation and completion times of the G-50 cleanup procedure was also recorded. A final summary form (not shown) monitored for each block the date and name of the individual that completed each step, from plasmid preparation to importing the sequence information into the database as described in the following section.

Data Processing and Analysis

Data was processed by a variety of programs and stored in an Oracle relational database. Data flow is illustrated in Figure 7. Upon completion of sequencing, trace files created by the ABI sequencer were moved via an upload interface (accessed at a local workstation) into a UNIX server (3c in Fig. 7). A web server was utilized as a 'central conduit' for data movement and to provide protection to the database. Data processing took place in the UNIX server automatically (on a nightly basis via a 'cron job'). Phred (Ewing et al., 1998) was used to call bases; it also uses trace parameters to assign an error

Figure 4. An example of the form used to monitor overnight growth of bacterial cells in 384-well plates. Each form required that the date of inoculation, initials of the individual inoculating the plate and library description be listed. The library description is the library name (*e.g.*, EM1) followed by the alphabetical identifier of the plate (see Fig. 3). Each square represents a well on the 384-well plate. After overnight growth each well was carefully analyzed. If no growth, poor growth or possible contamination was observed in a well, the corresponding symbol (as it appears in the key) was placed in that square.

Date _____ Initials _____ Library Des._____



Key: X = no growth; ? = poor or no growth; C = possibly contaminated

Figure 5. An example of the forms used to track plasmid isolation. (A) The first page tracks the inoculation of bacteria into deep 96-well blocks and 96-well v-bottom plates, archival storage of those v-bottom plates, and centrifugation of the deep 96-well blocks to obtain bacterial pellets. The grid at the upper left represents a deep 96-well block. After inoculation and overnight growth of bacteria in the deep-well blocks an X was placed in the squares that correspond to wells with no visible bacterial growth. B) The second page tracks the lysis, supernatant transfer and DNA precipitation steps. As an added quality control measure on the second page, during the resuspension/lysis step reagent lot numbers were included. On both pages vital information, such as the name of the individual performing the different plasmid isolation steps, date of initiation, as well as start and end times for certain steps were also required.

A)

Library

Block Name:

	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C												
D												
E												
F												
G												
H												

Comments:

Please score bacterial growth: *Wells that have not grown are marked with an X*

Please Record Below the Operations that You Are Performing:

Bacteria inoculation: inoculate from (pre-picked) 384-well plates frozen in -80°C

Label on 384-well plate:_____

Inoculate one Beckman Deep-well (TB/SALTS/AMP)

Beckman Deep-well; 1.5ml Terrific broth/TB salts/Ampicillin

Incubation: 37°C@250rpm; bacteria should grow 20hrs

STARTS @:_____ ENDS @:_____ _____ Total# of hours: ____

Expected time Exact time

Inoculate two Dynex V-bottom (TB/SALTS/AMP/GLYCEROL)

Dynex V-bottom; Dynex V-bottom; 150µl freezing media

Dynex V-bottom; Dynex V-bottom; 150µl freezing media

STARTS @:_____ ENDS @:_____ _____ Total# of hours: ____

Expected time Exact time

V-bottom permanent storage

V-bottom 1 - cover with foil and place in permanent storage (-80)

V-bottom 2 - cover with foil and place in permanent storage (-80)

Bacteria are pelleted

Centrifuge for **5 min @ 3000 rpm** (PRGM 31)

Pour off supernatant in bacterial waste box

Keeping block inverted transfer to paper towels et on saran wrap

Drain DW block for a couple of minutes and gently tap on paper

Transfer to a clean place on towel, let drain some more

Block is ready to be

Resuspended/lysed or,

covered with transparent film and frozen. *Write* on top of film *Date* and *Name of Block*

Autoclave waste!

B)

Resuspension/lysis (HYDRA) Date:_____ Operator's Name: _____

Add 200ul TE-Rnase (lot#_____)
Place on shaker (speed 8) for 30 min. (check for complete resuspension)
Add 200ul SDS/NaOH (lot#_____)
Place on shaker (speed 8) for 45-60 min. (check for complete lysis)
One block at a time, add 200ul KOAc (lot#_____)
Mix on the vortexer such that ALL 96 wells turn "black" (i.e. are vortexing properly)
Maintain vortex for 3 min. inverting block after 1.5 min
Put on Titertek shaker @ speed 4 until all blocks have been vortexed
Place in incubator for 20 min. (37C @ 350 rpm)
Put in the –80 freezer for a min of 8 hr. STARTS @:_____ ENDS @:_____

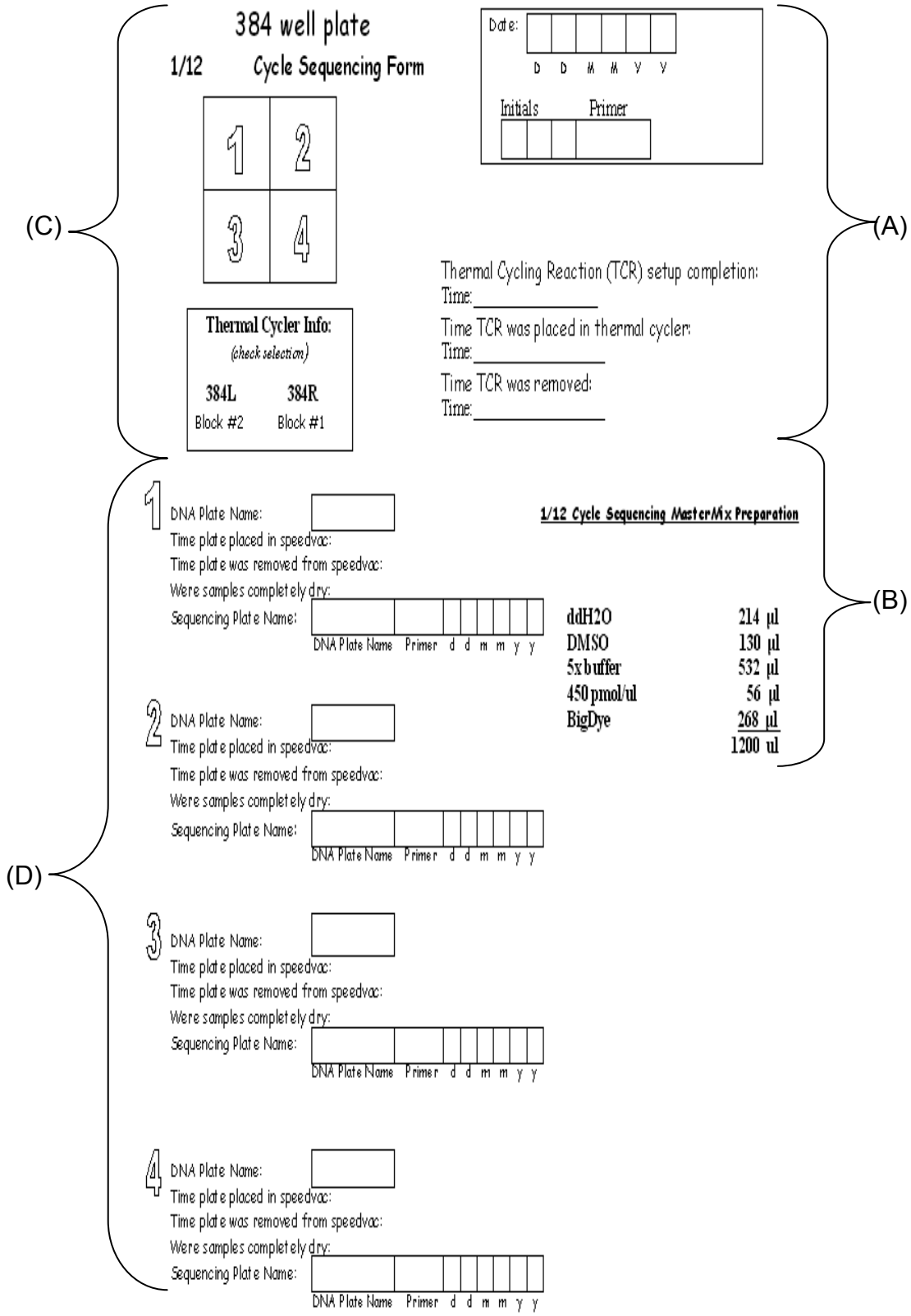
Supernatant transfer Date:_____ Operator's Name: _____

Thaw on the counter top for ca. 2 hours. (check for thorough thaw!)
Centrifuge for **45 min @ 3850 rpm (PRGM 35)**
Transfer 200ul of the supernatant into clean Costar Deep-well block
Clean bacterial debris out of Beckman block **immediately**

DNA Precipitation:

Add **500 µl** of **95% EtOH**
Immediately centrifuge for **30 min @ 3850 rpm (PRGM 34)**
Gently decant supernatant into sink and keeping blocks inverted set them gently one at a time on super-absorbent paper towel. Let block drain for 3 minutes.
Add **250 µl** of **70% EtOH**
Immediately centrifuge **15 min @ 3850 rpm (PRGM 33).** Gently decant supernatant into sink and keeping blocks inverted set them gently one at a time on super-absorbent paper towel for 5 min.
Dry blocks upside up @ 37°C for at least 20 min or until dry (small incubator).
Add **150 µl ddH_2O** to each well and put on the Titertek shaker at speed 7 for 10 min.
Transfer 150 µl to V-bottom plate and cover with foil cover. Store @ -20°C.

Figure 6. An example of the form used to monitor thermal cycling, elimination of unincorporated nucleotides, and lyophilization of sequencing reactions. (A) Pertinent information such as the date of thermal cycling, initials of the individual preparing the reactions, and the primer used were included in the form. Additionally, the time reactions were prepared, and the times when plates were put into and removed from the thermal cycler, were recorded. (B) A recipe for the reagents used (the master mix), including their volumes, was also included on the form. Reactions were labeled 1/12 because only $1/12^{th}$ of the standard volume of the ABI BigDye Ready Reaction reagent was used. (C) The box on the upper left side subdivided into quadrants 1 through 4 designates the four quadrants of a 384-well plate. The box directly below it allowed a user to keep track of which block in which thermal cycler was used. Following thermal cycling, the contents of each quadrant were transferred to a 96-well plate for the G-50 procedure. After the products of the G-50 cleanup reaction were collected, the contents of the plates were lyophilized in a SpeedVac. (D) The name of each plate, as well as the time it was put into and removed from the SpeedVac was recorded.


384 well plate
1/12 Cycle Sequencing Form
1
2
3
4
Date:
D D M M Y Y
Initials
Primer
Thermal Cycler Info:
(check selection)
384L
Block #2
384R
Block #1
Thermal Cycling Reaction (TCR) setup completion:
Time:
Time TCR was placed in thermal cycler:
Time:
Time TCR was removed:
Time:
1
DNA Plate Name:
Time plate placed in speedvac:
Time plate was removed from speedvac:
Were samples completely dry:
Sequencing Plate Name:
DNA Plate Name Primer d d m m y y
1/12 Cycle Sequencing MasterMix Preparation
ddH2O 214 µl
DMSO 130 µl
5x buffer 532 µl
450 pmol/ul 56 µl
BigDye 268 µl
1200 ul
2
DNA Plate Name:
Time plate placed in speedvac:
Time plate was removed from speedvac:
Were samples completely dry:
Sequencing Plate Name:
DNA Plate Name Primer d d m m y y
3
DNA Plate Name:
Time plate placed in speedvac:
Time plate was removed from speedvac:
Were samples completely dry:
Sequencing Plate Name:
DNA Plate Name Primer d d m m y y
4
DNA Plate Name:
Time plate placed in speedvac:
Time plate was removed from speedvac:
Were samples completely dry:
Sequencing Plate Name:
DNA Plate Name Primer d d m m y y
(A)
(B)
(C)
(D)

probability (also called quality values) to each base. CrossMatch was used to identify bacterial and vector sequences. The assembler Phrap then combed through the sequences called by Phred and assembled similar sequences into contigs. Phrap uses the assigned error probabilities along with the sequence alignment to attach an error probability to each base of an inferred consensus sequence of an assembled contig (Gordon et al., 1998). Sequences not assembled into contigs were classified as singletons. For direct GenBank submission an in-house script in the UNIX server was used. Only those high quality ESTs with regions having a quality score of at least 16 for a minimum of 100 continuous bases after removal of vector, *E. coli* and when present poly T, were submitted. A quality score of 16 corresponds to a probability of 97.5% to have called a base correctly. Similarity comparisons of all high quality ESTs were carried out with HT-BLAST. BLAST (Altschul et al., 1990) searches were performed against dbEST (Boguski et al., 1993) and the GenEMBL database. The latter is a combination of the GenBank (Benson et al., 2000) and EMBL (Baker et al., 2000) databases. Additionally, BLAST searches were performed against SWISS-PROT (Bairoch and Apweiler, 1997). Processed EST sequence data and BLAST data were stored in the relational database, where access was available via the web server.

Quality Control

Quality control was done on two separate levels. On one level, analysis was conducted to verify that plates were labeled correctly. On the other level the library quality was evaluated by determining whether *E. coli* contamination had occurred.

A. Verifying samples were labeled correctly

Prior analysis by Dr. Alan Gingle determined for each plate sequenced the total number of clones that provided overlapping forward and reverse sequences. Clones from plates that yielded very few overlapping sequence pairs were chosen for analysis. Analysis was conducted to assess whether the low number of clones with overlapping forward and reverse sequences was due to sample tracking errors, to short sequences that failed to overlap as a result of poor sequence quality, or to the existence of too few clones that had produced acceptable data.

A subset of clones from the plates with only a few overlapping pairs were chosen at random for analysis. Analysis consisted first of selecting the 3′ sequences from each clone. Using an in-house interface that allows a user to view 3′ contig clusters, each 3′ sequence and the sequences present in its cluster were identified and recorded. ESTs derived from other libraries sequenced in house (Table II) were included in the clusters. From the same subset of randomly selected clones the 5′ sequences were identified. Each 5′ sequence was used to perform BLASTN searches against the EST database. The results of each BLAST search was analyzed to identify database ESTs that originated from libraries sequenced in house. The results of the BLAST searches were recorded. Sequences present in each 5′ EST BLAST search were compared to the sequences present in the corresponding cluster of 3′ sequences.

B. Determining the extent of *E. coli* contamination

HT-BLAST was used to perform BLASTN searches of embryo ESTs against GenEMBL. BLASTN search results were recorded and stored as output files. The

Figure 7. Schematic flow of sequence data. Arrows indicate the direction and path of information exchange. Sequencing with an ABI3700 is preceded by 'wet lab' preparation, such as RNA and plasmid isolation (1). Upon completion of sequencing, trace files (chromatograms) produced by the sequencer are transferred to a local workstation (2). At the local workstation, data relating to plate identification, template preparation, thermal cycling, and so on, are entered. These data are transferred to the database via the web server (3a, 5a) and the trace files are renamed and information about them uploaded via the same path into the web server. In a subsequent step, the trace files are copied to a UNIX server via FTP (3c). Data processing in the UNIX server was an automated procedure under control of the database, via the web server (5b, 4a). Phred calls each base and to each assigns a quality value. Crossmatch identifies and tags bases originating from vector or *E. coli*. Phrap then assembles similar sequences into contigs. All output from the UNIX server reaches database tables via the web server (4b, 5a). Additionally, processed high quality sequences are submitted to GenBank via an email server (not shown) within the UNIX server. The latter requests and obtains the necessary information from the database via the web server (4, 5). High quality sequences are defined as those with regions having a quality score of at least 16 for a minimum of 100 continuous bases after removal of vector and, when present, polyT. Laboratory access to the sequence information contained in the RDBMS is accomplished from a local workstation communicating through the web server (3, 5). Similarly, public access passes through the web server (6, 5).

Bioinformatics Pipeline

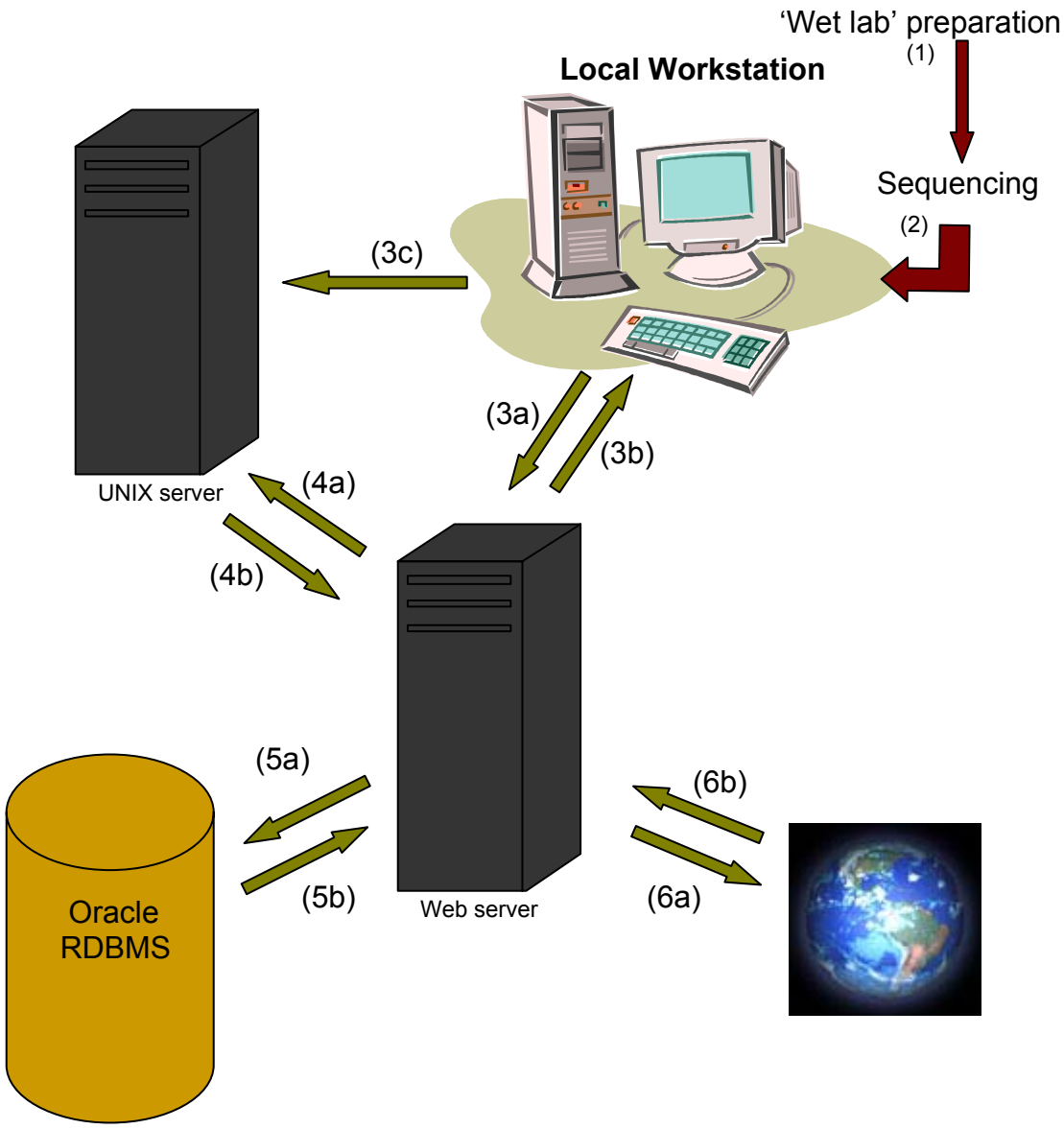


Table II. A listing of abbreviations and the cDNA libraries to which they refer.

Abbreviations	Library
EM1	Embryo
DG1	Dark-grown seedling
FM1	Floral meristem
IP1	Immature panicle
LG1	Light-grown seedling
OV1 & OV2	Ovary
PI1	Pathogen-induced plant
RHIZ1	Rhizome
WS1	Water-stressed seedlings

output files were parsed using a SQL query written by Dr. Alan Gingle. The resulting output was entered into the relational database. A second SQL query written by Dr. Chun Liang enabled a table containing the selected BLAST hit for each sequence result, was exported to the Spotfire DecisionSite program (http://www.spotfire.com). Spotfire is a commercially available visualization and analytical tool for examining data. In Spotfire, the ESTs were queried to reveal only those that returned hits to *E. coli* sequences. A new table containing only the query results was created. The new table was transferred to a spreadsheet for manual analysis. Manual analysis consisted of selecting only the highest scoring match per EST sequence; this was a necessary step, because a number of ESTs returned multiple hits to *E. coli* sequences. Additionally, manual analysis consisted of identifying ESTs that returned significant similarity (bit scores > 80) to *E. coli* sequences.

Annotation and Classification

HT-BLAST was used to perform BLASTX searches against SWISS-PROT. The BLAST search results were recorded and stored as output files. The output files were parsed using a SQL query written by Dr. Alan Gingle; the resulting data was entered into the relational database. A second SQL query, written by Dr. Chun Liang enabled a table containing the BLAST results to be transferred to a spreadsheet for manual analysis. Manual analysis consisted of selecting only the highest scoring match per EST sequence. Similar to the preceding section, this was a necessary step as a majority of ESTs returned multiple hits from the database. A new table containing the result of the manual analysis

was exported to Spotfire. In Spotfire the ESTs were queried to reveal only those that returned significant similarity (bit scores >80) to database sequences.

Sequence Clustering

A. Analysis of singleton ESTs and contigs-of-one

An initial SQL query, written by Dr. Chun Liang and using the EM1-only contig assemblies, revealed all the EM1 ESTs determined to be singletons by Phrap. The query result was manually compared against the result of another SQL query written by Dr. Liang. The other query revealed the ESTs identified by Phrap as singletons following the clustering of ESTs from other libraries, including the embryo (Table II). Manual comparison between the two query results was conducted to identify the EM1 singletons that remained singletons after the clustering with other libraries. Of the embryo ESTs that remained singletons, 50 were selected for analysis.

Contigs-of-one were also selected for analysis. Contigs-of-one are sequences that show substantial similarity to other sequences such that Phrap attempts to include them into a contig. These sequences, however, eventually failed to cluster with any other because they fell below the clustering parameters that were required by Phrap for inclusion. A total of 20 contigs-of-one were selected for analysis, in addition to the 50 singletons.

BLAST searches were conducted to identify the selected singleton and contig-of-one sequences. The procedure for assigning identities is shown in figure 8A. Matches were considered to be significant when a bit score >80 was observed, using BLASTN and BLASTX with all parameters set at the defaults. Sequences with significant similarities

to known genes after BLASTN searches against the non-redundant nucleotide database were assigned identities as illustrated in the path down the right side of Figure 8. Sequences with significant similarity to genomic DNA sequences, as well as sequences that failed to show significant similarity to public database sequences with BLASTN were further searched by BLASTX. Sequences that failed to show significant similarity with BLASTX were further searched by BLASTN against the EST database (Fig. 8B). The result of each BLAST search was manually analyzed to determine the identity of the gene, protein or EST match. The likely origin of the sequences was determined based on the identity of the database sequence.

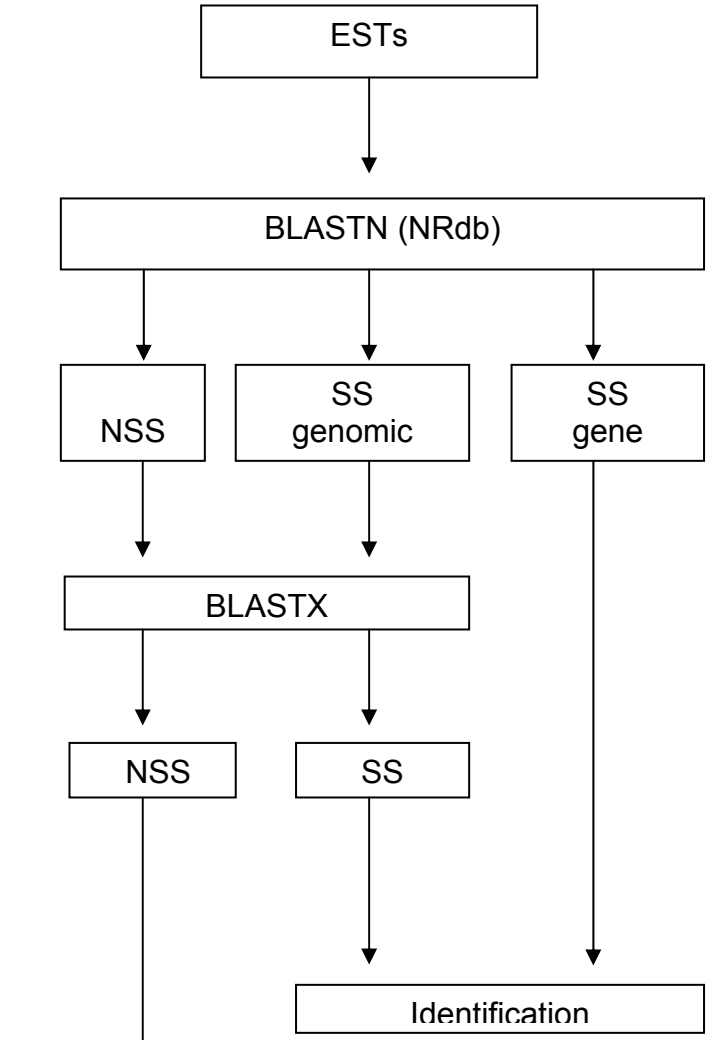
B. Assembly of a virtual unigene set

To identify a minimal set of sequences in which each represents a different gene in the genome, 3′ ESTs were assembled into contigs or determined to be singletons using Phrap. For the contigs a consensus sequence was determined from their alignment. To increase assembly stringency the Phrap minscore parameter was adjusted. The minscore is the minimum alignment score (based on Smith-Waterman algorithm) required for assembly. The parameter was increased from its default value of 30 to 80; all other parameters were left at their default values.

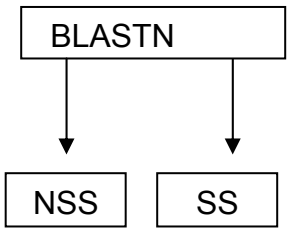
The graphical interface Consed was used for viewing and editing the contig assemblies. The 'high quality matches elsewhere' option in Consed allows the user to identify similar sequences that were assembled into different contigs. Once similar sequences were identified they were compared to each other through BLAST analysis. One of the two similar sequences was used in a BLASTN search against the EST

Figure 8. (A) Schematic presentation of the procedure used for analyzing singleton and contig-of-one sequences using BLASTN and BLASTX. Matches were considered significant when a bit score >80 was observed. The sequences were used to perform BLASTN searches against a non-redundant nucleotide database (NRdb). Sequences that returned one or more significant hits to a genomic DNA sequence annotated as a gene (SS gene) provided immediate identification. Those that either yielded no significant similarity (NSS) to an annotated gene or significant similarity to apparently non-genic genomic DNA (SS genomic DNA) were evaluated further by BLASTX against a non-redundant protein database. (B) Sequences with no significant similarity after both BLASTN and BLASTX searches were evaluated by BLASTN against the EST database (dbEST). Sequences showing significant similarity to EST sequences in the database were binned into an SS group, while sequences with no significant similarity were binned into an NSS group.

A)


ESTs
BLASTN (NRdb)
NSS
SS genomic
SS gene
BLASTX
NSS
SS
Identification

B)


BLASTN
NSS
SS

database, which contained the other. Following sequence comparisons based on the BLAST search results, the similar sequences could be aligned and temporarily assembled into a new contig in Consed. The temporary contigs were visually inspected for alignment inconsistencies (*i.e.,* insertion/deletion events and sequence misalignments). Sequences were permanently assembled into new contigs when these inconsistencies were not observed.

C. Identification of embryo-specific genes

Phrap was used to assemble into contigs the 3′ ESTs from the various libraries (Table II). The resulting contig assemblies were entered into the relational database. The contig assemblies were viewed using an in-house contig visualization interface. The interface enabled a user to view the contig assemblies in a number of ways. Of interest here was the ability to identify the sequences that comprised a contig, and to identify the composition of each contig with respect to contributions from the different libraries (Fig. 9). The assemblies were analyzed to identify and record contigs comprised of embryo-only ESTs. A representative from each of these contigs (the longest sequence) was used to perform BLAST searches against the non-redundant protein database. ESTs that returned significant similarities (bit score >80) were identified based on the description of the database protein.

Identification of the relative abundance of full-length-coding cDNAs

HT-BLAST was used to perform BLASTX searches against SWISS-PROT. The BLAST search results were written to output files. Using a script written by Dr. Chun

Liang, the output files were parsed and data corresponding to the highest bit score for each 5′ sequence was entered into the relational database. A new table composed of the query results was exported to Spotfire. The exported table contained a scoring report for each 5′ EST sequence and its corresponding database hit (Table III). A modified version of the table was created in Spotfire. The modified table contained only 5′ ESTs that aligned to their corresponding database hits beginning at the first published amino acid residue (denoted S1) with a bit score >100. The bit score parameter is the normalized alignment score taking into account the scoring system used. Creation of the modified table allowed for 3-dimensional visualization of the quality of the BLAST hits obtained with each EST sequence using the bit score, Expect value and percentage identity values assigned by BLAST (Fig. 10). The Expect value refers to the probability that an alignment will by chance be the same as that reported.

Analysis of ESTs potentially spliced alternatively

The visualization tool Consed was used to identify not only similar ESTs that were clustered into separate contigs, but also sequences within contigs that display alignment inconsistencies. In Consed the 'high quality matches elsewhere' option facilitated identification of these sequences. Only those sequences that were observed to align with 100% identity outside of the differential region were selected for further analysis. The term differential region is used here to identify a region of DNA that is present on one EST sequence, but absent from the other. Further analysis involved performing BLAST searches with each sequence in the pair against the nucleotide, protein and EST databases. Additionally, to identify the open reading frame, selected sequences were translated in the three forward frames using Map (GCG, Madison, WI).

Figure 9. An example of the contig analysis interface. The interface enables a user to identify all the sequences present in a particular contig. The libraries in the interface are color coded. The upper window shows a more compressed view of contigs than does the lower window. Here only embryo ESTs are visible in the upper window. A region selected in the upper window will appear in an expanded view in the lower window, where ESTs from all libraries are visible. When a contig is selected in the lower window the sequences that comprise the contig can be viewed. The X-axis of both windows indicates the contig number. The Y-axis indicates the percentage composition of each contig, in terms of the library from which a sequence is derived. In this particular example, the upper window is displaying the percentage of embryo sequences in each contig. Of interest are those contigs that contain preferentially or exclusively embryo sequences.

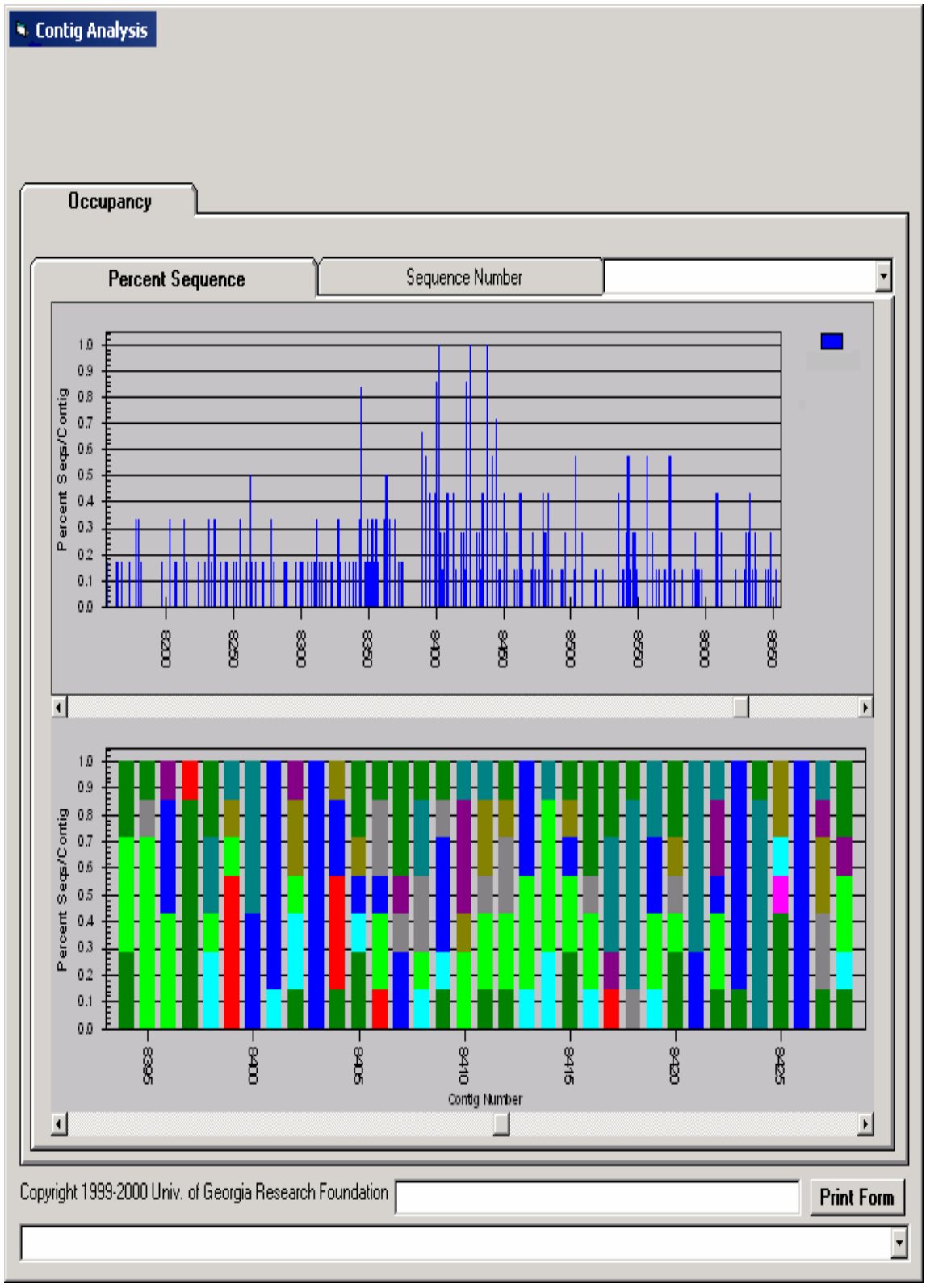
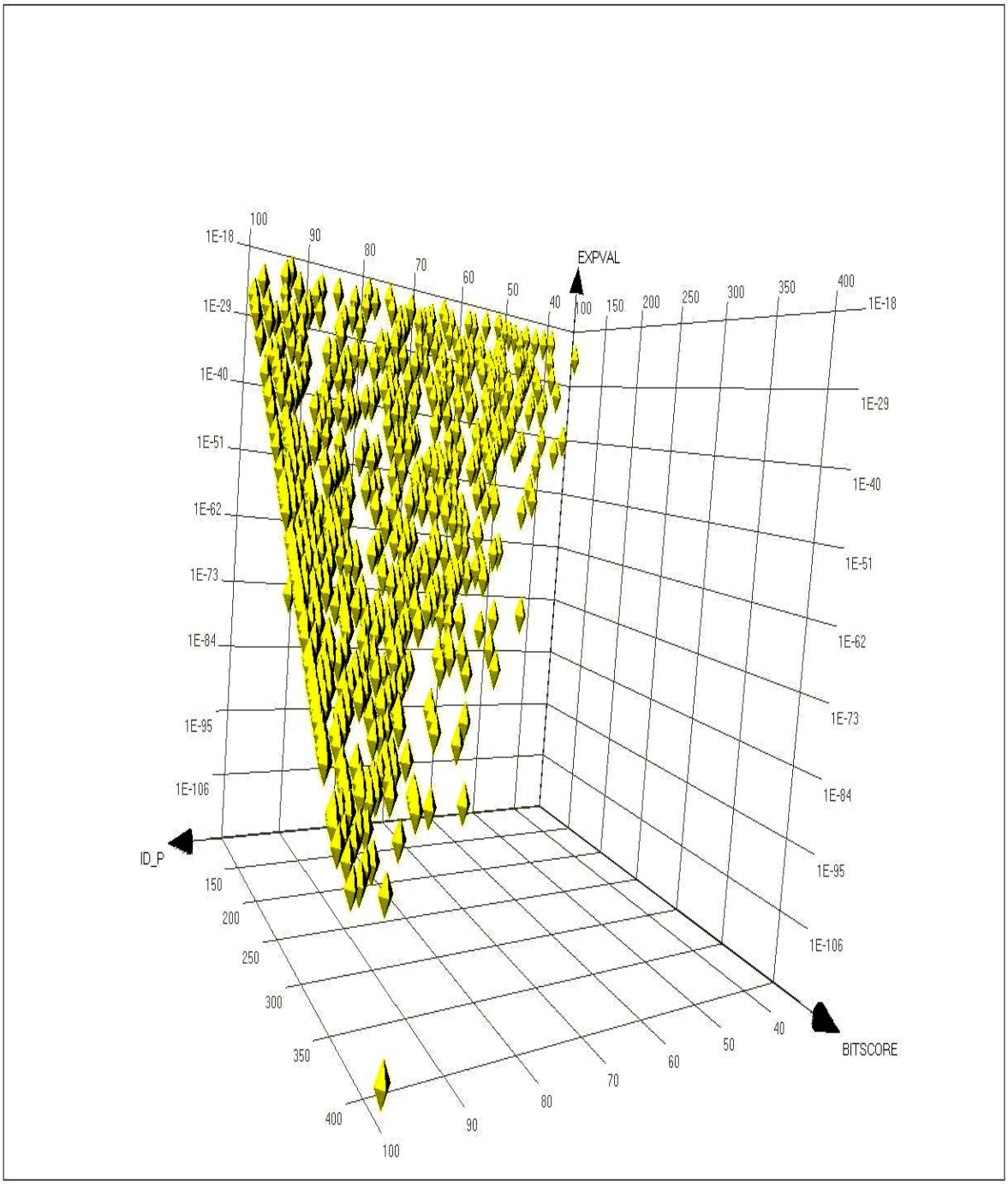

Contig Analysis
Occupancy
Percent Sequence
Sequence Number
Percent Segs/Contig
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
8200
8250
8300
8350
8400
8450
8500
8550
8600
8650
Percent Segs/Contig
8395
8400
8405
8410
8415
8420
8425
Contig Number
Copyright 1999-2000 Univ. of Georgia Research Foundation
Print Form

Table III. An example of the information contained in the original table exported to Spotfire.

QUERY_NAME[a]	SBJCT_GB_NUM[b]	S1[c]	S2[d]	Q1[e]	Q2[f]	EXPVAL[g]	CLONE_ID[h]	ID_P[i]	BITSCORE[J]	LENGTH[k]	CONTIG_S[l]
EM1_13_D07.b1_A002	Q9XI07	77	242	29	523	5.00E-45	EM1_13_D07	54	181	166	SINGLETON
EM1_13_D08.b1_A002	Q9SY69	232	404	18	536	9.00E-56	EM1_13_D08	59	217	173	SINGLETON
EM1_13_D09.b1_A002	Q9SWB4	3	159	33	563	5.00E-36	EM1_13_D09	44	151	177	266_763
EM1_13_D10.b1_A002	Q9ST79	77	158	240	482	9.00E-31	EM1_13_D10	78	134	82	3090_556
EM1_13_D11.b1_A002	P95164	240	267	208	291	0.14	EM1_13_D11	64	37.4	28	9675_683
EM1_13_D12.b1_A002	Q9C6D2	6	145	100	519	6.00E-54	EM1_13_D12	67	211	140	5943_718
EM1_13_E01.b1_A002	Q9MAS5	1	66	167	364	8.00E-30	EM1_13_E01	86	130	66	8113_654
EM1_13_E03.b1_A002	Q9SH49	1	83	87	338	6.00E-31	EM1_13_E03	78	134	84	7159_713
EM1_13_E04.b1_A002	Q9CAJ5	707	876	9	518	8.00E-54	EM1_13_E04	58	210	170	2516_640
EM1_13_E05.b1_A002	Q9CAA7	382	507	97	465	2.00E-09	EM1_13_E05	32	63.2	126	8155_753

[a] QUERY_NAME - the identity of an EST query sequence.
[b] SBJCT_GB_NUM - the unique identifier given to a subject sequence by GenBank.
[c] S1 – the first amino acid residue in the subject sequence that aligns with the EST query sequence.
[d] S2 – the last amino acid residue in the subject sequence that aligns with the EST query sequence.
[e] Q1 – the first nucleotide in the EST query sequence that aligns with the subject sequence.
[f] Q2 – the last nucleotide in the EST query sequence that aligns with the subject sequence.
[g] EXPVAL - Expect value assigned by BLAST to the alignment.
[h] CLONE_ID - the identity of the clone from which the EST sequence was generated.
[i] ID_P - the percentage identity of the alignment.
[j] BITSCORE - the alignment score assigned by BLAST.
[k] Length – the length of the match in amino acid residues
[l] Contig_S - indicates whether the query sequence is a member of a contig (by including the contig number) or is a singleton.

Figure 10. A 3-D scatter plot of all EST sequences used for full-length analysis. A total of 833 EST sequences were selected after a BLAST search to SWISS PROT. Expect values (EXPVAL) were plotted on the Z-axis. Expect values are the proportion of alignments with scores equivalent to or better than the reported score that are expected to occur by chance. Percentage identities (ID_P) and bit scores were plotted on the X- and Y-axes, respectively. Bit scores are the alignment scores assigned by BLAST, normalized for the particular scoring matrix used.


EXPVAL
1E-18
1E-29
1E-40
1E-51
1E-62
1E-73
1E-84
1E-95
1E-106
100
90
80
70
60
50
40
ID_P
100
150
200
250
300
350
400
BITSCORE
100
150
200
250
300
350
400

CHAPTER 3

RESULTS

Sequence Statistics

The sequencing goal was to obtain a minimum of 5,000 high quality ESTs from both 3′ and 5′ ends of embryo-derived cDNA clones. High quality is defined here as sequences that after removal of vector and polyT have a region with overall quality greater than a Phred value of 16 for at least 100 nt. A total of 7,296 3′ and 7,104 5′ sequence reactions were evaluated in order to obtain 5,405 and 5,126 high quality 5′ and 3′ sequences, respectively (Table IV).

Quality Control

Quality control was conducted on two different levels. On one level, sample tracking was evaluated to determine whether identification errors occurred during EST production. Such errors would lead to an inability to correlate sequences with the cDNA clones from which they were obtained. On the other level, the library quality was evaluated to determine the extent of *E. coli* contamination.

A. Verifying samples were labeled correctly

Plates containing a relatively small number of overlapping forward and reverse sequences were selected for analysis. Analysis was conducted to determine whether the small number of overlapping 3′ and 5′ sequences observed for these plates was the result of poor sequence quality, and thus short sequence reads, or to errors such as mislabeling. A

subset of clones was analyzed from each of these plates. For the subset of cDNA clones analyzed, the sequences present in the corresponding 3′ cluster for the most part matched the corresponding sequences that were returned by the 5′ BLAST analysis (Table V). For example, in Table V the 3′ sequence EM1_10_F07.g1 clustered with two other sequences EM1_45_E10.g1 and OV2_35_C07.g1. Using the 5′ sequence from the same template (EM1_10_F07.b1) for BLAST analysis, the corresponding 5′ sequences EM1_45_E10.b1 and OV2_35_c7.b1 were returned. The result verified that no labeling error had occurred, because the 3′ sequence EM1_10_F07.g1 and the 5′ sequence EM1_10_F07.b1 originated from the same cDNA template. There were some cases in which a sequence would be present in the 3′ cluster, but its corresponding 5′ sequence did not appear in the BLAST results. Further analysis revealed that these sequences were of very low quality and thus were not deposited in the public database.

Analyses of plates with a small, average and large number of high quality (submittable) sequences establishes that the number of overlapping sequence pairs is directly related to this parameter (Fig. 11). Similarly, the number of overlapping sequence pairs is also directly related to average sequence read length (Fig. 12).

B. Determining the extent of *E. coli* contamination

BLASTN searches revealed a total of 30 ESTs that matched *E. coli* sequences (Table VI). Of the 30, 12 were identified as having significant similarity to an *E. coli* sequence (Table VI, shown in red). Thus, it was determined that 12 embryo clones were actually contaminants by *E. coli* DNA and were not detected during the *E. coli* screening process. All 12 ESTs were sequenced from what would have been

Table IV. Sequence statistics for the embryo library.

End of cDNA that was sequenced	5′	3′
Total sequences	7104	7296
High quality sequences[a]	5405	5126
Proportion good	76%	70%
Average high quality length	521	508

[a] defined as sequences that after removal of vector and polyT, have a region with overall quality greater than a value of 16 for at least 100 nt.

Table V. Results from the analysis of plates with very few overlapping 3′ and 5′ sequences. The first column lists the 3′ (*.g1) and corresponding 5′ (*.b1) sequences used for analysis. The second column lists the results of the 3′ clustering, while the third column lists the results of the 5′ BLAST analysis. When sequences returned from the 5′ BLAST come from the same DNA templates as those in the 3′ cluster, the results verify that the 3′ and 5′ sequences in the first column do come from the same well of the same plate.

Sequence identification	Length (nt)	Sequences present in 3′ cluster	Sequences present in 5′ BLAST results
EM1_10			
EM1_10_B10.g1	564		
EM1_10_B10.b1 [a]	523		
EM1_10_E05.g1	609	DG1_2_C12.g1	DG1_2_C12.b1
EM1_10_E05.b1	471	DG1_71_H05.g1	DG1_71_H05.b1
		DG1_84_H01.g1	DG1_84_H01.b1
		EM1_43_C08.g1	EM1_43_C08.b1
		LG1_242_B10.g1	OV2_5_F01.b1
		OV2_5_F01.g1	
EM1_10_F07.g1	544	EM1_45_E10.g1	EM1_45_E10.b1
EM1_10_F07.b1	510	OV2_35_C07.g1	OV2_35_C07.b1
EM1_11			
EM1_11_A06.g1	644	IP1_52_D02.g1	IP1_52_D02.b1
EM1_11_A06.b1	475	IP1_52_D03.g1	IP1_52_D03.b1
		PI1_11_D08.g1	PI1_29_B07.b1
		PI1_29_B07.g1	
		PI1_86_G01.g1	
EM1_11_B09.g1	647	DG1_47_F01.g1	EM1_16_D07.b1
EM1_11_B09.b1	502	EM1_16_D07.g1	EM1_23_A09.b1

		EM1_23_A09.g1	EM1_46_C05.b1
		EM1_46_C05.g1	FM1_8_G12.b1
		EM1_82_H05.g1	FM1_10_G06.b1
		FM1_1_H03.g1	PI1_35_H10.b1
		FM1_8_G12.g1	
		FM1_10_G06.g1	
		FM1_56_F08.g1	
		PI1_35_H10.g1	
EM1_11_F03.g1	542	DG1_5_B11.g1	DG1_5_B11.b1
EM1_11_F03.b1	450	DG1_53_G09.g1	EM1_36_F03.b1
		DG1_53_G09.g2	EM1_42_H01.b1
		EM1_7_G02.g1	EM1_78_G08.b1
		EM1_36_F03.g1	LG1_353_H05.b1
		EM1_42_H01.g1	OV2_22_H01.b1
		EM1_54_C03.g1	PI1_65_A07.b1
		EM1_78_G08.g1	PI1_92_E08.b1
		LG1_353_H05.g1	
		OV2_22_H01.g1	
		PI1_65_A07.g1	
		PI1_92_E08.g1	
		WS1_3_H03.g1	
EM1_77[b]			
EM1_9			
EM1_9_A08.g1	576	EM1_76_G09.g1	EM1_76_G09.b1
EM1_9_A08.b1	472	FM1_8_E08.g1	FM1_8_E08.g1
		FM1_29_D12.g1	FM1_29_D12.g1
		FM1_37_G11.g1	FM1_37_G11.g1
		PI1_10_F04.g1	PI1_10_F04.g1
		PI1_88_F04.g1	PI1_88_F04.g1
		PI1_91_D10.g1	PI1_91_D10.g1
EM1_9_E03.g1	463	DG1_31_A03.g1	DG1_31_A03.b1
EM1_9_E03.b1	385	EM1_35_G09.g1	EM1_35_G09.b1
		EM1_49_H09.g1	EM1_49_H09.b1
		EM1_68_E11.g1	OV1_21_B07.b1
		OV1_21_B07.g1	OV1_23_D05.b1
		OV1_23_D05.g1	RHIZ2_41_B05.b1
		OV1_27_B03.g1	
		OV1_32_B12.g1	
		RHIZ2_41_B05.g1	
EM1_9_F06.g1	184	EM1_3_C12.g1	EM1_3_C12.b1
EM1_9_F06.b1	260	EM1_16_H12.g1	EM1_16_H12.g1
		EM1_17_E04.g1	EM1_17_E04.b1
		EM1_21_B12.g1	EM1_21_B12.b1
		EM1_22_B09.g1	EM1_22_B09.b1
		EM1_34_H10.g1	EM1_74_B12.b1
		EM1_35_G12.g1	
		EM1_39_B04.g1	
		EM1_57_G06.g1	
		EM1_66_D01.g1	
		EM1_74_B12.g1	

EM1_12			
EM1_12_B05.g1	482	EM1_14_D08.g1	EM1_14_D08.b1
EM1_12_B05.b1	483	EM1_14_H09.g1	EM1_14_H09.b1
		EM1_18_E05.g1	LG1_352_D02.b1
		LG1_352_D02.g1	WS1_39_E11.b1
		WS1_31_G06.g1	
		WS1_34_C02.g1	
		WS1_39_E11.g1	
		WS1_51_H10.g1	
EM1_12_C07.g1	570	LG1_223_H10.g1	LG1_349_H12.b1
EM1_12_C07.b1	504	LG1_235_B11.g1	OV2_22_G01.b1
		LG1_349_H12.g1	PI1_83_E01.b1
		OV2_22_G01.g1	PI1_87_G01.b1
		PI1_83_E01.g1	RHIZ2_65_A06.b1
		PI1_87_G01.g1	
		RHIZ2_65_A06.g1	
EM1_12_E06.g1	659	EM1_4_C02.g1	EM1_4_C02.b1
EM1_12_E06.b1	512	EM1_36_D02.g1	OV2_29_G02.b1
		EM1_42_A02.g1	PI1_26_G10.b1
		OV2_19_E01.g1	PI1_51_E04.b1
		OV2_29_G02.g1	PI1_88_C01.b1
		PI1_26_G10.g1	PI1_95_E03.b1
		PI1_26_G10.g2	
		PI1_51_E04.g1	
		PI1_83_D12.g1	
		PI1_88_C01.g1	
		PI1_95_E03.g1	

[a] EM1_10_B10.b1 contained the entire cDNA insert sequence. A BLAST search against the EST database using EM1_10_B06.b1 revealed its corresponding 3′ sequence EM1_10_B06.g1, thus verifying that the two sequences originated from the same clone.

[b] Block EM1_77 contained only 15 submittable 5′ sequences. Of these, only 3 were submittable 3′ sequences. Of the three, one was a singleton, and two were in clusters of greater than 60 sequences. Hence, no information was entered for EM1_77 in order to save space.

Figure 11. The number of overlapping sequence pairs observed as a function of the number of high-quality (submittable) sequences obtained from a 96-well plate. The number of high-quality sequences, presented at the top of each bar, is the average of the number of forward and reverse sequences submitted to GenBank. The number of overlapping sequence pairs is directly correlated with the number of high-quality sequences.

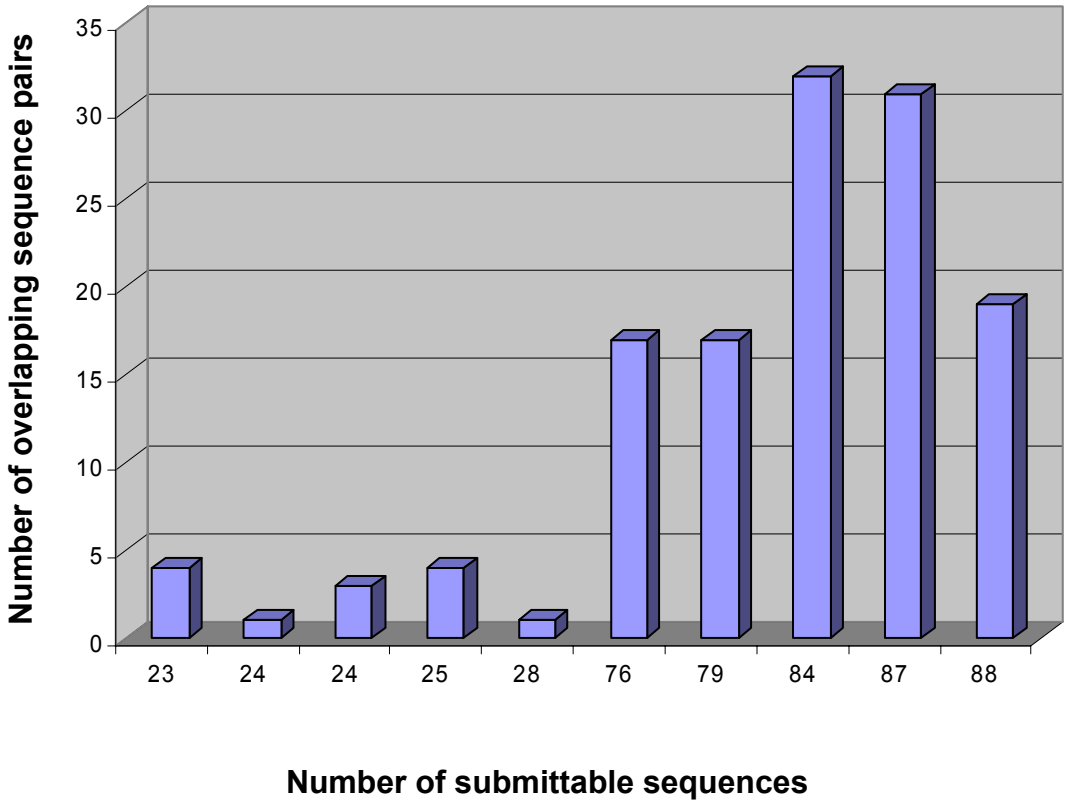

Number of overlapping sequence pairs
35
30
25
20
15
10
5
0
23
24
24
25
28
76
79
84
87
88
Number of submittable sequences

Figure 12. The number of overlapping sequence pairs observed as a function of the average quality 16 sequence length. A quality 16 value corresponds to a 97.5% probability that a base was called correctly. The quality 16 lengths were determined by taking the average of all sequences exceeding a length of 100 nt for a 96-well plate.

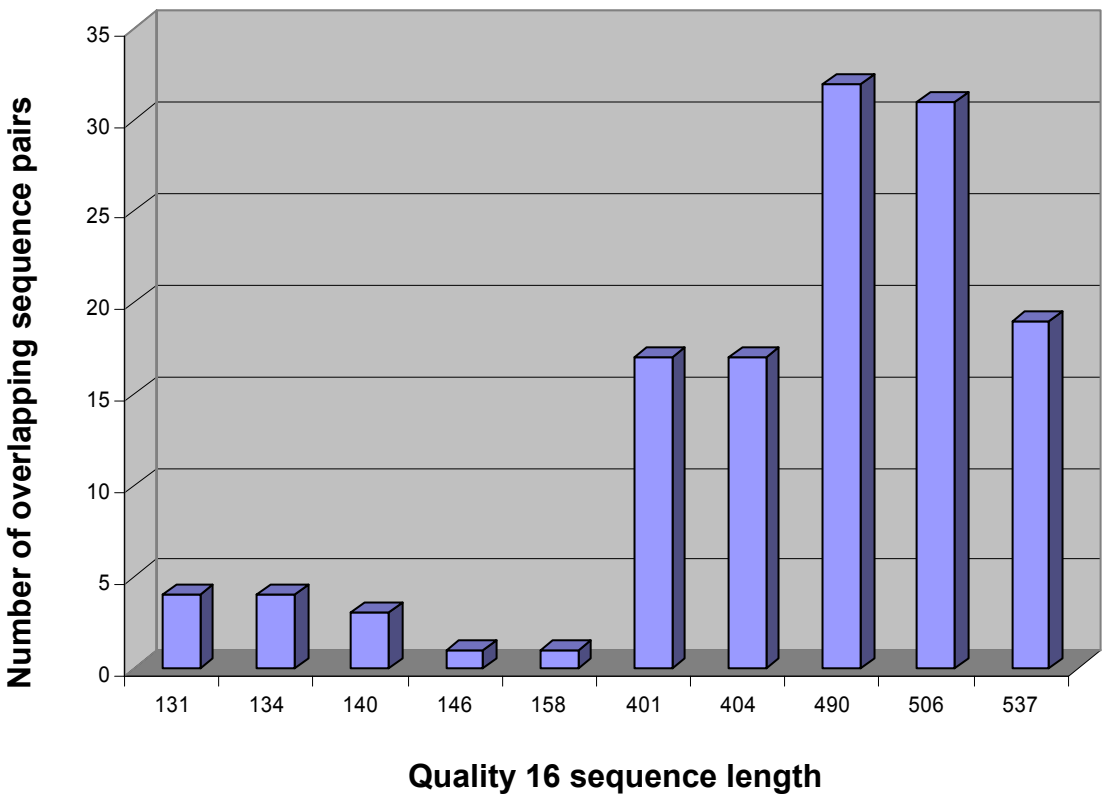

Number of overlapping sequence pairs
35
30
25
20
15
10
5
0
131
134
140
146
158
401
404
490
506
537
Quality 16 sequence length

Table VI. A listing of ESTs that match *E. coli* sequences in a BLASTN search against GenEMBL. Matches with significant similarities, defined as those with bit scores greater than 80, are shown in red.

QUERY_NAME[a]	LENGTH[b]	SBJCT_INFO[c]	BIT SCORE[d]	EXPVAL[e]	IDENTITY(%)[f]
EM1_1_A01.b1_A002	444	E.coli genomic DNA, Kohara clone #443(59.8-60.2min.)	42.1	0.51	21/21 (100%)
EM1_3_C01.b1_A002	587	Escherichia coli genomic DNA. (23.5 - 23.8 min)	993	0	501/501 (100%)
EM1_3_D01.b1_A002	569	Escherichia coli genomic DNA. (23.5 - 23.8 min)	957	0	483/483 (100%)
EM1_7_G04.g1_A002	526	E. coli chromosomal region from 87.2 to 89.2 minutes	40.1	2.4	20/20 (100%)
EM1_12_D01.b1_A002	365	Escherichia coli K12 MG1655 section 95 of 400 of the complete genome	553	e-155	279/279 (100%)
EM1_14_E08.b1_A002	546	Escherichia coli ECOR 31 anthranilate isomerase(trpC), complete cds	42.1	0.63	21/21 (100%)
EM1_15_A08.b1_A002	517	Escherichia coli ECOR 31 anthranilate isomerase(trpC), complete cds	42.1	0.59	21/21 (100%)
EM1_18_A11.b1_A002	216	Escherichia coli 0157:H7 EDL933 genome, contig 290	428	e-117	216/216 (100%)
EM1_19_B09.b1_A002	617	Escherichia coli K12	731	0	371/372 (99%)
EM1_33_A11.b1_A002	490	E.coli broad-host-range IncP-beta plasmidR751 traD, traC-2, traC-3, traC-4 genes	44.1	1.40E-01	22/22 (100%)
EM1_34_G12.b1_A002	150	Escherichia coli genomic DNA. (23.5 - 23.8 min)	99.6	8.00E-19	134/146 (91%)
EM1_35_A11.g1_A002	156	E.coli genomic DNA, Kohara clone #316	40.1	0.64	20/20 (100%)
EM1_36_A11.b1_A002	579	Escherichia coli K12 MG1655 section 110 of400 of the complete genome	1148	0	579/579 (100%)
EM1_36_D09.g1_A002	197	E.coli genomic DNA, Kohara clone #445(60.2-60.6min.)	40.1	0.84	20/20 (100%)
EM1_36_G06.b1_A002	519	Escherichia coli K12 MG1655 section 110 of400 of the complete genome	1029	0	519/519 (100%)
EM1_40_F04.g1_A002	512	Escherichia coli O157:H7 DNA, complete genome, section 1/20	40.1	2.3	20/20 (100%)
EM1_45_C08.b1_A002	547	E.coli genomic DNA, 5'flanking region of rrn Hgene	46.1	0.04	23/23 (100%)
EM1_48_H11.g1_A002	696	Escherichia coli genomic sequence of minutes 9 to12	44.1	0.2	22/22 (100%)
EM1_51_B11.b1_A002	365	Escherichia coli K-12 genome, approximately 57minutes	42.1	0.41	24/25 (96%)

EM1_51_D07.b1_A002	405	E.coli genomic DNA, Kohara clone #305(34.7-35.1min.)	44.1	0.12	22/22 (100%)
EM1_53_A09.b1_A002	418	E.coli genomic DNA, Kohara clone #420(54.9-55.2min.)	46.1	0.031	26/27 (96%)
EM1_53_F11.b1_A002	525	E.coli genomic DNA, Kohara clone #373(49.5-49.9min.)	44.1	0.15	25/26 (96%)
EM1_56_E09.g1_A002	546	Synthetic E.coli alkaline phosphatase gene, partial cds	42.1	0.63	21/21 (100%)
EM1_56_F09.b1_A002	416	E. coli plasmid RP4 traF (5'end), traG, traH, traI, traJ, traK, traL and traM genes	44.1	0.12	22/22 (100%)
EM1_58_D12.b1_A002	478	Escherichia coli genomic DNA. (23.5 - 23.8 min)	777	0	392/392 (100%)
EM1_59_G09.b1_A002	314	Escherichia coli genomic DNA. (17.6 - 18.0 min)	607	e-171	309/310 (99%)
EM1_60_C12.b1_A002	278	Escherichia coli genomic DNA. (23.5 - 23.8 min)	535	e-150	276/278 (99%)
EM1_68_G02.g1_A002	478	Escherichia coli YgaC (ygaC) gene, completecds, and H-NSB (stpA) gene, complete cds.	40.1	2.2	20/20 (100%)
EM1_79_D08.b1_A002	386	Escherichia coli genomic DNA. (17.1 - 17.4min)	488	e-135	246/246 (100%)
EM1_79_G07.b1_A002	431	E. coli plasmid RP4 traF (5'end), traG, traH, traI, traJ, traK, traL and traM genes	44.1	0.12	22/22 (100%)

[a] QUERY_NAME- The identity of the EST query sequence.
[b] LENGTH- The length in bases of the EST query sequence.
[c] SBJCT_INFO- The description of the subject sequence.
[d] BITSCORE- The alignment score assigned by BLAST.
[e]EXPVAL- Expect value assigned by BLAST to the alignment.
[f] IDENTITY- The length of the alignment. The numerator is the length of the aligned EST query sequence. The denominator is the length of the subject sequence in which the alignment occurred.

the 5′ end of the insert if it were a cDNA clone. Of the 12 ESTs with significant similarities, only one was derived from a clone that gave high quality sequences from both ends. This observation was expected since a polyT primer was used to sequence the clone inserts from the 3′ direction. Consequently, a contaminant (presumably genomic DNA) should not have a polyA tract with which the primer could anneal, and thus could not yield a presumably 3′ sequence. Interestingly enough, while the exception EM1_12_D01.b1_A002 showed significant similarity to an *E. coli* genomic DNA sequence, the sequence from the other end of the insert showed significant similarity to a ribosomal protein from *Zea mays*. These results indicate EM1_12_D01 is likely a chimeric clone.

Annotation and classification

A total of 10,531 ESTs (5,405 5′ESTs and 5126 3′ESTs) were compared by BLASTX against SWISS-PROT. Of the 10,531 ESTs, 3,927 had significant similarity to sequences in the protein database. The remaining 6,604 ESTs could not be identified. Based on the description of the database sequences the ESTs were grouped into 17 functional categories. These categories were adapted from the MIPS functional classification applied to *Arabidopsis* genes (The Arabidopsis Initiative, 2000; Table VII).

Analysis of the distribution of ESTs into the 17 functional categories revealed that ESTs derived from genes involved in various metabolic processes were most abundant (Fig. 13A). However, the unknown, hypothetical and unclassified categories when combined, account for 29% of the total ESTs (Fig. 13B).

EST datasets can also provide gene expression information (Ohlrogge and Benning, 2000). Based on BLAST comparisons the seven most abundant ESTs were identified (Table VIII). The genes present in the group of seven represent a variety of functional categories. In this embryo dataset, ESTs similar to transcription factors (91 ESTs), heat shock proteins (90 ESTs) and elongation factors (83 ESTs) were also abundant; however, since each of these gene identities represents a gene family rather than a unigene, they were not included in this table.

It is important to note that a number of ESTs showed similarity to genes that can be placed in numerous functional categories. For example, while heat shock proteins play a role as molecular chaperones, there is biochemical and genetic evidence to support that some heat shock proteins are involved in signal transduction (Pratt, 1993; Rutherford and Zuker, 1994; Nathan and Lindquist, 1995). In these instances the ESTs were grouped to the most notable and widely recognized functional category. Thus, using this example, heat shock proteins would be grouped in the protein fate functional category.

Sequence Clustering

A. Analysis of singleton ESTs and contigs-of-one

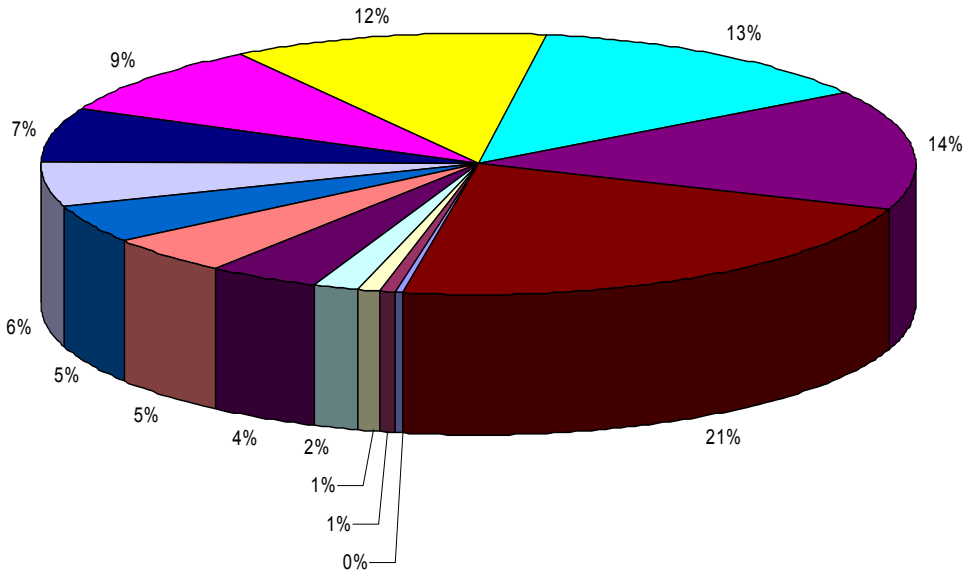
Singleton ESTs are those that do not correspond to any other EST coming from the same dataset. Singletons could be derived either from genes that are rarely expressed, and thus sampled only once, or from contaminating DNA in the EM1 cDNA library. Similarly, contigs-of-one are sequences selected by Phrap for clustering with other sequences; however, they were not clustered successfully such that each was made the only member of its contig. BLAST comparisons were conducted in order to assess

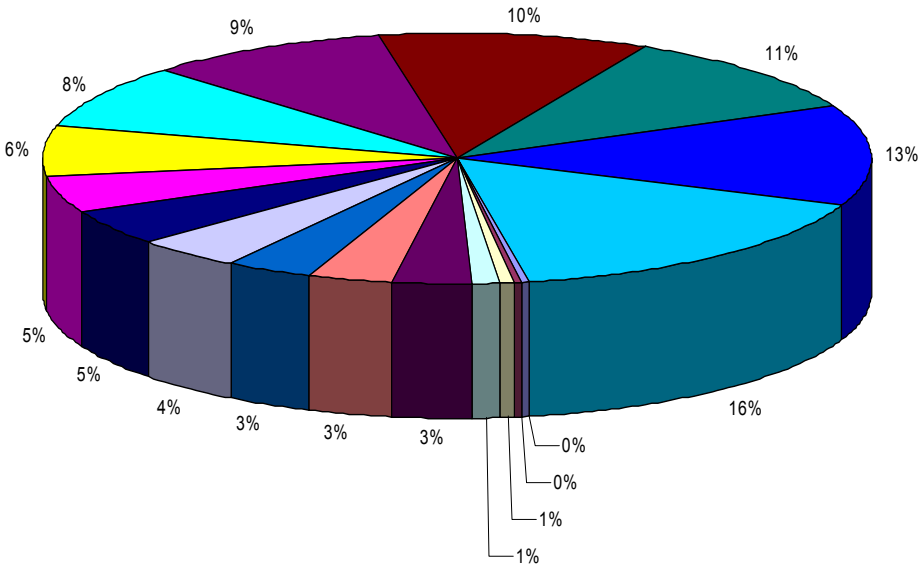
Table VII. A listing of functional categories and their descriptions. EST sequences were divided into one categories listed. The third column lists the total number of ESTs grouped to each category.

Functional category	Category description	No. of ESTs
Unknown	A gene without significant homology to any protein, but with EST homology.	183
Hypothetical	A gene predicted with a gene prediction program	512
Metabolism	Amino acid, carbohydrate, lipid, nitrogen and secondary metabolism	629
Energy	Glycolysis, TCA cycle, electron transport, glyoxylate cycle, respiration, pentose phosphate pathway, fermentation, glyoxylate cycle and oxidation of fatty acids	370
Protein fate	Protein folding, modification, degradation	412
Protein synthesis	Ribosomal proteins, translation (*e.g.,* initiation and elongation factors) and t-RNA synthases	331
Transcription/post transcription	mRNA synthesis, splicing, stabilization and degradation	186
Cellular transport and transport mechanisms	Transport ATPases, Ion channels, amino acid, lipid, carbohydrate and metal transporters	250
Signal transduction	Receptor proteins, second messagers, G-proteins, kinases and phosphatases	120
Cell cycle and DNA processing	Cell growth, cell cycle control, mitosis, cytokinesis, DNA synthesis and replication	46
Cell structure	Plasma membrane, cell wall and cytoskeletal proteins	131
Defense	Environmental stimuli responses (*e.g.,* stress), detoxification, DNA repair and degradation of foreign compounds	163
Storage	Storage protein (*e.g.,* globulin)	22
Mobile elements	Transposable elements	15

Intracellular trafficking	Vesicular transport	8
Development	Proteins influential in embryonic development (*e.g.,* Phytochrome, Gibberellin oxidase)	131
Unclassified	1)Generic description (*e.g.,*serine rich protein), Genomic DNA and uncharacterized arabidopsis, drosophila, mouse and human sequences listing only an identification number 2) Functional identity could not be determined	418

Figure 13. The pie charts summarize the functional categories into which the embryo ESTs were grouped. A) Distribution of ESTs into functional categories. B) Same as A, but also including unknown, hypothetical and unclassified categories. The nonfunctional categories combined are the most abundant category of ESTs in the dataset.


12%
13%
9%
7%
14%
6%
5%
5%
4%
2%
1%
1%
0%
21%
Intracellular trafficking (0.3%)
Mobile elements (1%)
Storage (1%)
Cell cycle and DNA processing (2%)
Signal Transduction (4%)
Cell structure (5%)
Development (5%)
Defense (6%)
Transcription/post transcription (7%)
Cellular transport and transport mechanisms (9%)
Protein synthesis (12%)
Energy (13%)
Protein fate (14%)
Metabolism (21%)



- Intracellular trafficking (0.4%)
- Mobile elements (0.6%)
- Storage (1%)
- Cell cycle and DNA processing (1%)
- Signal transduction (3%)
- Cell structure (3%)
- Development (4%)
- Defense (4%)
- Unknown (5%)
- Transcription/post transcription (5%)
- Cellular transport and transport mechanisms (6%)
- Protein synthesis (8%)
- Energy (9%)
- Protein fate (10%)
- Unclassified (11%)
- Hypothetical (13%)
- Metabolism (16%)

Table VIII. A listing of the seven most abundant ESTs.

Gene identification	Functional category	No. of ESTs
60S ribosomal protein	Protein synthesis	53
Phytochrome A	Development	37
40S ribosomal protein	Protein synthesis	36
Polyubiquitin	Protein fate	36
ADP/ATP carrier protein	Transport	32
Cytochrome P450	Metabolism	29
S-adenosylmethionine decarboxylase	Metabolism	28

whether or not singletons and contigs-of-one originated from contaminants. A total of 50 singletons and 20 contigs-of-one was chosen randomly for detailed investigation.

The 50 singletons were compared by BLAST to non-redundant nucleotide, protein and EST databases (Fig. 14A). Of the 50 singletons, 7 displayed significant similarity to genes in the nucleotide database (Table IX), and 15 displayed significant similarity to protein sequences (Table X). Of the remaining 28 singletons, 12 were identified as having significant similarity to numerous plant ESTs in the database. The remaining 16 singletons failed to show significant similarity after BLAST searches to nucleotide, protein or EST databases. These singletons were considered unknown sequences, perhaps representing novel genes.

Similar to the singleton analysis, 20 contigs-of-one were compared by BLAST to nucleotide, protein and EST databases (Fig. 14B). Of the 20, 16 displayed significant similarity to a gene present in the nucleotide database (Table XI), and three displayed significant similarity to a protein sequence. The lone remaining sequence was identified as having significant similarity to numerous plant ESTs in the database.

The identity of the database sequence to which a singleton or contig-of-one showed significant similarity offers insight into the origin of that singleton or contig-of-one. Of the seven singletons with significant similarity to sequences in the nucleotide database (Table IX, Fig. 14A), four recorded significant similarity to either a sorghum sequence or a sequence from a related species (*i.e.*, *Zea mays* or *Lolium perenne*). Additionally, of the remaining three singletons, two recorded similarity to sequences from other plant species (*i.e.*, *Brassica juncea* and *Cucumis sativus*). One singleton was found to show significant similarity to a sequence derived from *Botrytis cinerea*, a known

plant pathogen. Although the singleton sequence does not need to be derived from a *Botrytis* cDNA, similarity to the pathogen indicates that it is likely the singleton was not derived from the embryo cDNA library. Of the 15 singletons with significant similarity to the sequences in the protein database (Table X, Fig. 14A), 14 recorded similarities to either *Zea mays, Oryza sativa* or *Arabidopsis thaliana* protein sequences. One singleton, however, showed significant similarity to a sequence derived from *Saccharomyces cerevisiae.* The singleton displayed similarity to a dihydroxy-acid dehydratase. The protein is an enzyme involved in the synthesis of valine and isoleucine. The singleton sequence was found to encode an amino acid sequence 57% identical to the yeast protein. A more detailed analysis of the yeast protein revealed a homologue in *Arabidopsis*. BLAST 2 Sequence (Tatusova and Madden, 1999), a tool for aligning two sequences, showed that the yeast protein was 58% identical to the *Arabidopsis* homologue. Thus the *Arabidopsis* homologue and the EST singleton revealed similar identities to the yeast protein, 58% and 57% respectively. The similarity suggested alignment to the yeast protein was plausible for an EST derived from the embryo cDNA library.

A total of 16 contigs-of-one were similar to sequences in the nucleotide database (Table XI, Fig. 14B). All 16 were similar to sequences either from sorghum or from a related species (*i.e.*, *Zea mays* and *Oryza sativa*). Of the four contigs-of-one remaining from the original 20 examined, two were similar to an *Arabidopsis thaliana* protein, and one to a fungal protein sequence. The contig-of-one sequence with similarity to a fungal protein displayed significant similarity (70% identical) to an aspartic protease from *Aspergillus fumigatus.* Further analysis of the *fumigatus* protein revealed 23 plant homologues. Each plant homologue was ~30% identical to the *fumigatis* protein; therefore the contig-of-

one's 70% identity was rather high. Thus, it was concluded that this contig-of-one was a contaminant.

In both the singleton and contig-of-one analyses, sequences that showed no significant similarity in BLAST comparisons to both the non-redundant nucleotide and protein databases were searched against the EST database. Of the remaining 28 singletons, 12 displayed significant similarity to multiple plant ESTs. The only remaining contig-of-one also displayed similarity to multiple plants ESTs (Fig. 14).

B. Assembly of a virtual unigene set

Initial assembly of the 5,126 3′ sequences resulted in 3,481 sequences clustered into a total of 1,136 contigs. The remaining 1,645 sequences remained as singletons, not identical to any other ESTs in this dataset. Of the 1,136 contigs, 167 were contigs-of-one. Contigs-of-one are sequences that show substantial similarity to other sequences such that Phrap attempts to include them in larger contigs. These sequences, however, eventually failed to cluster with any other because they fell below the clustering parameters that were required by Phrap for inclusion.

Analysis and editing of the Phrap assembly in Consed resulted in the merger of three contig pairs, reducing the contig count from 1,136 to 1,133. Each pair of contigs was merged after alignment of the two in a BLAST search gave high comparison scores (score > 500, E-value = 0, 99-100% identity), and after analysis of those alignments in Consed verified that the mergers were valid. Five of the six contigs in the three contig pairs were contigs-of-one. Thus merger of these contigs, in addition to reducing the total contig count, also reduced the number of contigs-of-one. The number of contigs-of-one

Figure 14. A summary of the results of the singleton and contigs-of-one analysis. A) Results of the singleton analysis. B) Results of the contigs-of-one analysis. SS denotes significant similarity, NSS no significant similarity. Also listed are the identities of the organisms with each the ESTs shared significant similarity.

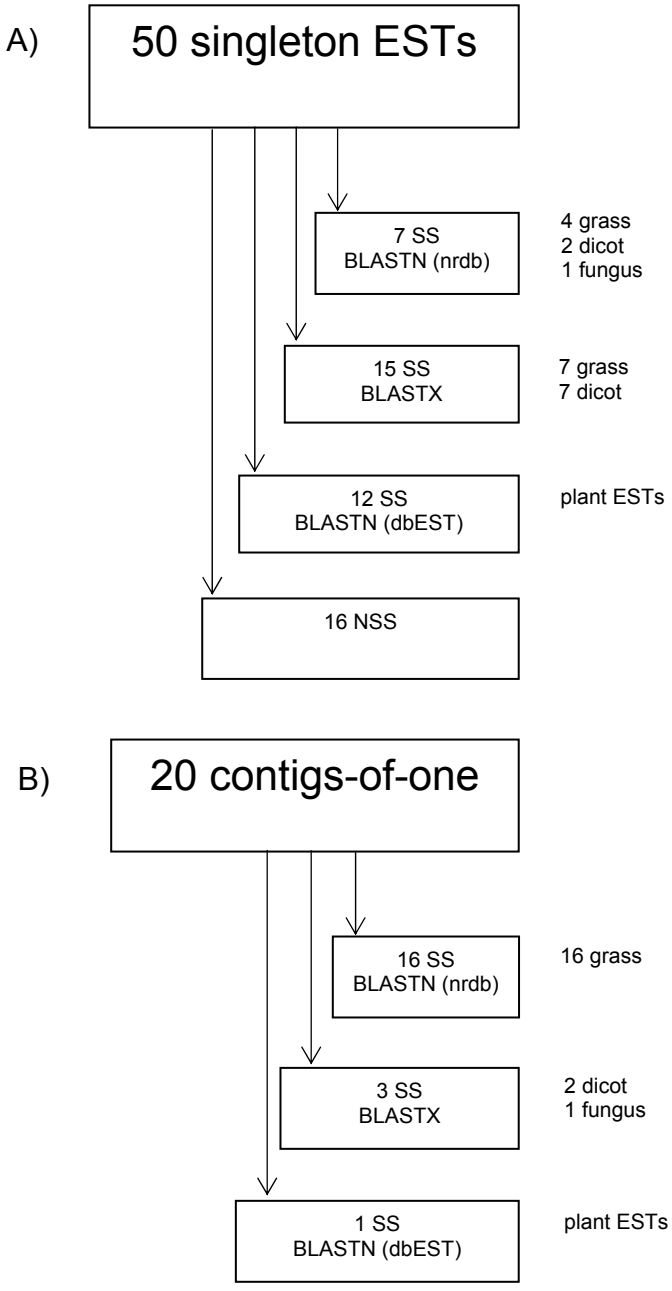

A)
50 singleton ESTs
7 SS
BLASTN (nrdb)
4 grass
2 dicot
1 fungus
15 SS
BLASTX
7 grass
7 dicot
12 SS
BLASTN (dbEST)
plant ESTs
16 NSS
B)
20 contigs-of-one
16 SS
BLASTN (nrdb)
16 grass
3 SS
BLASTX
2 dicot
1 fungus
1 SS
BLASTN (dbEST)
plant ESTs

Table IX. BLASTN search results for the seven singleton sequences with significant similarity to known genes. BLASTN searches were conducted against a non-redundant nucleotide database.

Sequence ID[a]	Length (bp)[b]	Accession no.[c]	Description[d]	Bit score[e]	E-value[f]	Identity(%)[g]
EM1-15-C10.g1	664	AY014281.1	Lolium perenne clone 8 gibberellin 20-oxidase mRNA	109	3.0E-21	97/111(87%)
EM1-16-F07.g1	617	AF109695	Brassica juncea monodehydroascorbate reductase, mRNA	90	3.0E-15	126/153 (82%)
EM1-1-A05.g1	200	AJ278666.1	Zea mays mRNA for putative Rop family GTPase, ROP4 (rop4 gene)	317	2.0E-84	187/196 (95%)
EM1-2-C05.g1	567	AL112323.1	Botrytis cinerea strain T4 cDNA library under conditions of nitrogen deprivation	86	4.0E-14	139/171 (81%)
EM1-40-H08.g1	424	AB049102.1	Cucumis sativus CsPHR mRNA for CPD photolyase	82	5.0E-13	71/81 (87%)
EM1-40-H11.g1[h]	578	AF023165.1	Zea mays leucine-rich repeat transmembrane protein kinase 2 (ltk2) mRNA	581	1E-163	376/403 (96%)
EM1-76-E08.g1	502	AF061282.1	Sorghum bicolor 22 kDa kafirin cluster	137	1.0E-29	108/121 (89%)

[a] Sequence ID - The identity of the EST query sequence.
[b] Length - The length in bases of EST query sequence.
[c] Accession no. - The unique identifier given to the database sequence.
[d] Description - The identity of the database sequence.
[e] Bit score - The bit (alignment quality) score assigned by BLAST.
[f] E-value - Expect value assigned by BLAST to the alignment.
[g] Identity - The length of the alignment. The numerator is the length of the aligned EST query sequence. The denominator is the length of the subject sequence in which the alignment occurred.
[h] The alignment of EM1_40_H11.g1 to the leucine-rich repeat transmembrane protein kinase contain an eight nucleotide gap

Table X. BLASTX search results for the 15 singleton sequences that showed significant similarity to database proteins. BLASTX searches were conducted against a non-redundant protein database.

Sequence ID[a]	Length[b]	Accession no.[c]	Description[d]	Bit score[e]	E-value[f]	Identity(%)[g]
EM1-2-A07.g1	591	NP_197475.1	(NM_121979)peptidase-like protein [Oryza sativa]	144	4.00E-34	75/153 (45%)
EM1-3-B04.g1	685	AAL59021.1	putative cell cycle regulatory protein [Oryza sativa]	150	1.00E-35	75/112 (66%)
EM1-3-B06.g1	440	NP_012550.1	catalyzes third step in common pathway leading to biosynthesis of branched-chain amino acids; Ilv3p [Saccharomyces cerevisiae].	167	2.00E-41	78/136 (57%)
EM1-8-A01.g1	403	CAB51835.1	(AJ243961) l1332.6 [Oryza sativa]	89	1.00E-17	44/59 (74%)
EM1-9-G04.g1	581	NP_566737.1	expressed protein [Arabidopsis thaliana]	122	3.00E-27	62/89 (69%)
EM1-19-F11.g1	654	NP_199933.1	(NM_124499) putative protein [Arabidopsis thaliana]	145	3.00E-34	86/218 (39%)
EM1-19-H04.g1	447	AAL01177.1	(AC079843) Hypothetical protein [Oryza sativa]	152	1.00E-36	78/141 (58%)
EM1-21-A02.g1	578	NP_191055.1	(NM_115352) putative protein [Arabidopsis thaliana]	107	4.00E-27	53/90 (58%)
EM1-22-B02.g1	521	NP_563893.1	ATP-dependent Clp protease proteolytic subunit (ClpP6)[Arabidopsis thaliana]	79	2.00E-14	34/48 (70%)
EM1-22-B07.g1	428	NP_197391.1	putative protein [Arabidopsis thaliana]	136	4.00E-32	58/75 (77%)
EM1-25-A11.g1	431	NP_197094.1	(NM_121595) putative protein [Oryza sativa]	112	8.00E-25	58/135 (42%)
EM1-26-F04.g1	351	BAA94529.2	Similar to Zea mays S-domain receptor-like protein kinase	209	3.00E-54	98/115 (85%)
EM1-41-D11.g1	473	T01258	hypothetical protein F16M14.20 - Arabidopsis thaliana	91	3.00E-18	40/50 (80%)
EM1-41-E02.g1	333	BAB64689.1	unknown protein [Oryza sativa]	98	1.00E-20	54/87 (62%)
EM1-82-D09.g1	685	NP_181526.1	putative peroxisomal membrane carrier protein [Arabidopsis thaliana]	189	2.00E-47	98/135 (72%)

[a] Sequence ID - The identity of the EST query sequence.
[b] Length - The length in bases of the EST query sequence.
[c] Accession no. - The unique identifier given to the database sequence.
[d] Description - The identity of the database sequence.
[e] Bit score - The alignment score assigned by BLAST.
[f] E-value - Expect value assigned by BLAST to the alignment.
[g] Identity - The length of the alignment in amino acid residues. The numerator is the length of the aligned EST query sequence. The denominator is the length of the subject sequence in which the alignment occurred.

Table XI. BLASTN search results for the contig-of-one sequences that showed significant similarity to database sequences. BLASTN searches were conducted against the non-redundant nucleotide database.

Sequence ID[a]	Length[b]	Description[c]	Bit score[d]	E-value[e]	Identity[f]
EM1_2_B05.b1	467	Zea mays PCO084551 mRNA sequence	611	e-172	421/459 (91%)
EM1_2_C04.g1	574	Zea mays H+-pyrophosphatase mRNA	319	2.00E-84	176/181 (97%)
EM1_2_D06.b1	551	Oryza sativa GTP-binding protein mRNA	359	2.00E-96	205/213 (96%)
EM1_2_H03.b1	501	Oryza sativa (japonica cultivar-group) chromosome 10 clone OJ1014H12	248	5.00E-63	164/177 (92%)
EM1_2_H12.g1	656	Zea mays PCO098705 mRNA sequence	381	e-103	249/268 (92%)
EM1_3_E08.b1	485	Zea mays PCO101905 mRNA sequence	757	0	445/465 (95%)
EM1_4_C04.g1	532	Zea mays CL33183_1 mRNA sequence	281	4.00E-73	249/284 (87%)
EM1_4_E05.g1	567	Zea mays CL1830_1 mRNA sequence	450	e-124	348/387 (89%)
EM1_7_D01.b1	218	Zea mays mRNA for aldehyde oxidase	331	2.00E-88	200/211 (94%)
EM1_7_H03.g1	370	Sorghum bicolor cytochrome P450-like protein	141	6.00E-31	113/127 (88%)
EM1_9_B03.b1	322	Oryza sativa chromosome 3 BAC OSJNBa0013M12	157	9.00E-36	130/147 (88%)
EM1_42_G09.g1	338	Z.mays gene for cyclophilin	339	1.00E-90	256/282 (90%)
EM1_44_C03.g1	489	Zea mays PCO099302 mRNA sequence	521	e-145	364/395 (92%)
EM1_54_A05.b1	435	Zea mays PCO117022 mRNA sequence	460	e-127	283/300 (94%)
EM1_73_A10.g1	152	Z.mays mRNA for porin	137	4.00E-30	133/149 (89%)
EM1_73_F06.b1	352	Zea mays PCO103251 mRNA sequence	186	1.00E-44	100/102 (98%)

[a] Sequence ID - The identity of the EST query sequence.
[b] Length - The length in bases of the EST query sequence.
[c] Description - The identity of the database sequence.
[d] Bit score - The alignment score assigned by BLAST.
[e] E-value - Expect value assigned by BLAST to the alignment.
[f] Identity - The length of the alignment in amino acid residues. The numerator is the length of the aligned EST query sequence. The denominator is the length of the subject sequence in which the alignment occurred.

were reduced from 167 to 162. The number of sequences in a given contig ranged from 1 to 58 (Fig. 15). With the inclusion of contigs-of-one, sequences clustered into contigs accounted for 68% of the unigene set. The value of 68% can also be viewed as a redundancy value, indicating there to be a 68% chance that the next sequence introduced into the unigene set will already be represented in the contig assembly. Assuming that 97% of the singleton ESTs (33 of 34 that displayed significant similarity; Fig. 14A) and 95% of the contig-of-one ESTs (19 of the 20 that displayed significant similarity; Fig. 14B) derive from sorghum transcripts, the embryo unigene set contains 2,721 members.

C. Identification of embryo-specific genes

Using the contig analysis interface, contigs comprised of embryo-only ESTs were selected. For each embryo-only contig a representative sequence was selected for BLAST comparison searches. The result of each search was recorded and used to annotate the contigs.

A total of 441 ESTs were assembled into 131 embryo-only contigs. The number of ESTs present in each contig ranged from 2 to 12. Of all the contigs analyzed, 37 (28%) contained ESTs that returned significant similarity to known database sequences. Table XII lists the identity of database sequences preferentially expressed in the embryo. The 94 remaining contigs returned either no significant similarity or similarity to an unknown or hypothetical protein.

Identification of the relative abundance of full-length coding cDNAs

Full length coding sequences were determined by identifying EST query sequences that encode the first published amino acid residue of the subject sequence (denoted S1=1). The first published amino acid residue identifies the presumptive translation start site. Based on the Spotfire analysis, 833 of the 5,405 5′ ESTs were found to align with a corresponding top scoring protein hit where the alignment began at the first published amino acid residue. Of the total 5′ ESTs (5,405), only 3,103 were used for full length analysis; this number represents the number of ESTs with significant matches to the protein database. Thus, by extrapolation it can be estimated that 27% (833/3103) of the 5′ ESTs are derived from full-length cDNA clones. In all cases the EST query sequences aligned to the first published amino acid residue with more than 6 nt of query sequence remaining upstream of the alignment. As many as 372 nt were found upstream of the first codon.

Analysis of ESTs potentially spliced alternatively

Five pairs of ESTs were selected from the dataset for analysis. These five pairs were selected because, apart from an extension of one member relative to the other at each end, each aligned with 100% identity over their entire length, except that one of the two sequences had an insert relative to the other (Fig. 16). This additional sequence was in each of the five cases of high quality and at or near the center of the consensus sequence for the pair. These were the only five pairs detected that met these criteria. The first pair of sequences, EM1_6_H05.b1 and EM1_23_B02.b1, constituted a contig of two members. EM1_6_H05.b1 had a 101-nt insertion when compared to the other (Fig. 17).

Figure 15. Distribution of 3′ ESTs clustered into contigs. The number of sequences in a contig shown on the X-axis ranged from 1 (contigs of one) to 58. The frequency of ESTs in a particular contig, which is shown on the Y-axis, was determined by dividing the number of sequences in contigs of a particular size (number of sequences in a contig multiplied by the number of contigs of that size) by the total number of high quality 3′ sequences (5,126).

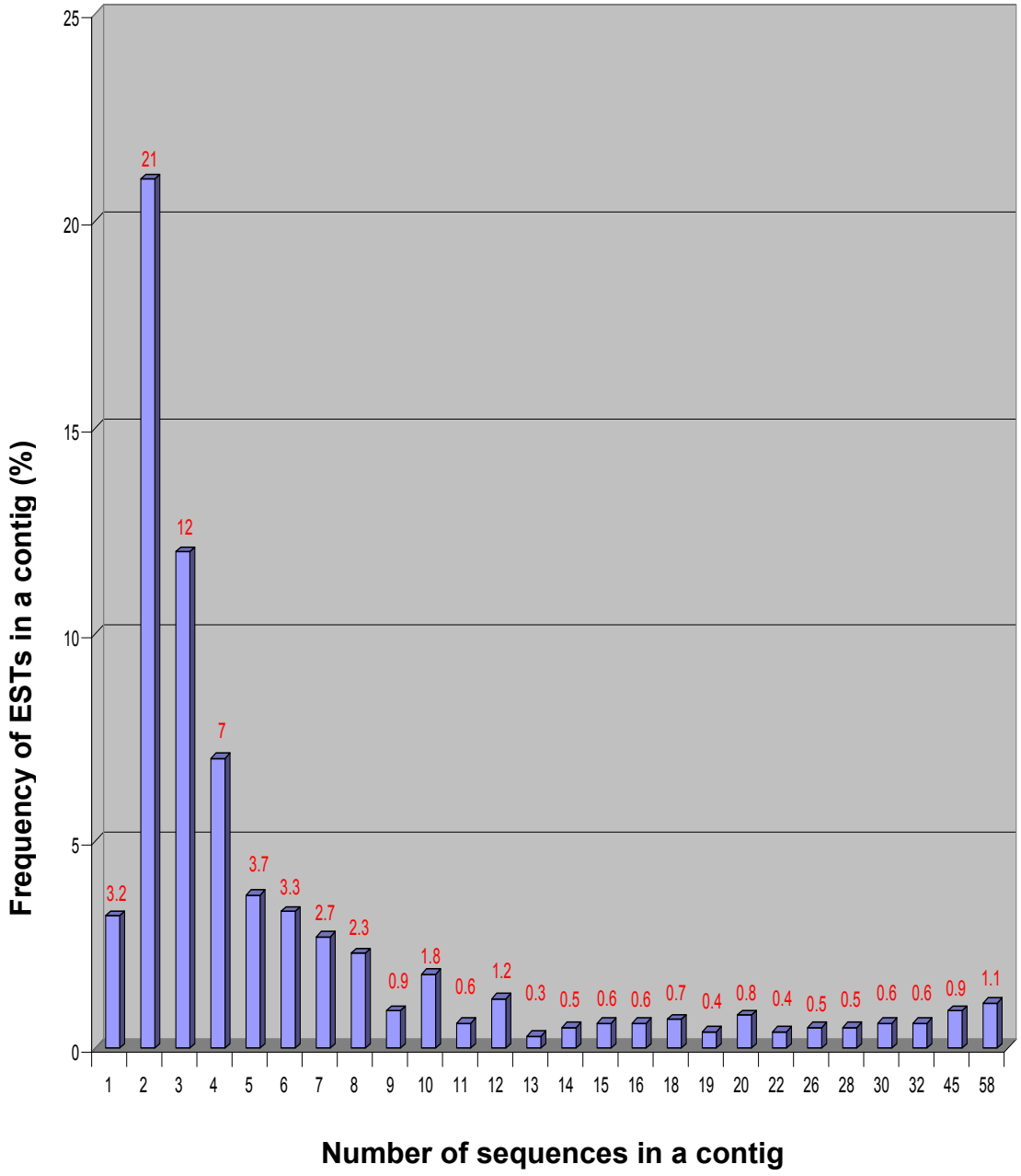

Frequency of ESTs in a contig (%)
25
20
15
10
5
0
3.2
21
12
7
3.7
3.3
2.7
2.3
0.9
1.8
0.6
1.2
0.3
0.5
0.6
0.6
0.7
0.4
0.8
0.4
0.5
0.5
0.6
0.6
0.9
1.1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 19 20 22 26 28 30 32 45 58
Number of sequences in a contig

Table XII. Description of genes to which the embryo-only contigs displayed significant similarity.

Description	Number of ESTs in contig
S-adenosylmethionine decarboxylase 2	12
Oleosin ZM-I	8
Cytochrome P450	8
Caleosin	7
3-methyl-2-oxobutanoate hydroxy-methyl-transferase	7
seed maturation protein	7
steroleosin	6
C13 cysteine proteinase precursor	6
Globulin-2 precursor - maize	5
17.5 kd class II heat shock protein	5
glucose and ribitol dehydrogenase homolog	4
Heat shock 70 kDa protein	4
oil body protein	4
Cysteine proteinase inhibitor	4
alpha expansin	4
nucleotide excision repair protein XP-D	4
root cap protein 1	3
Putative CER1	3
plastid division protein ftsZ	3
11-beta-hydroxysteroid dehydrogenase	3
casein kinase II beta subunit CKB1	3
phytochrome A	3
Putative splicing factor	3
Glutamyl-tRNA synthetase	3
Putative protein kinase Xa21	3
Putative Ruv DNA-helicase	3
Putative cycloartenol synthase	3
ferredoxin--nitrite reductase	3
Putative electron transfer oxidoreductase	3
zinc finger protein-like	2
kinesin heavy chain	2
initiation factor 3d	2
probable ubiquitin activating enzyme 2	2
Putative MAP kinase	2
serine carboxypeptidase II,	2
succinate dehydrogenase subunit 3	2
glycine hydroxymethyltransferase	2

Analysis of EM1_6_H05.b1 revealed that the first two nucleotides on both ends of the insert matched the highly conserved intronic sequences (GT at the 5′ donor site and AG at the 3′ acceptor site) normally present at splice junctions. Additionally, Figure 17 shows that the sequences around the junction match the consensus splice junction sequences in plants (Luehrsen et al., 1994): a splicing donor site of AGGTA and splicing acceptor site of TGCAGG. Both sequences when compared by BLAST to protein, mRNA and EST databases returned the same subject sequences. Protein and mRNA BLAST comparisons in particular, returned matches to an *Oryza sativa* alpha amylase. Both sequences aligned near the 5′ end of the sequence, covering a portion of the 5′UTR as well as some coding sequence. The additional 101-nt of EM1_6_H05.b1 were, however, not identified in any database except with reference to itself. The results indicate that the 101 nt is likely an unspliced intronic region. For further analysis both sequences were translated in the three forward reading frames (Map, GCG, Madision, WI) to analyze the insertion. In the open reading frame of EM1_6_H05.b1 a stop codon was detected within the presumptive intron.

Comparison of EM1_14_G09.b1 to EM1_37_F03.b1, both contigs-of-one, revealed that EM1_37_F03.b1 contained a 94-nt insertion when compared to EM1_14_G09.b1 (Fig. 18). Similar to the previous case, intronic donor and acceptor sequences were observed. Figure 18 also demonstrates that the junction sites of the insertion matches the consensus splice junction sequences in plants. BLASTN searches to the nucleotide database in both cases revealed no significant similarities. Additionally, both sequences only matched each other in searches against the EST database. However, in searches to the protein database both sequences showed significant similarity to an

unknown protein which also lacked the insertion present in EM1_37_F03.b1. For further analysis both sequences were translated in the three forward reading frames. Similar to the prior case, each frame was compared against the BLASTX result to identify the open reading frame. A stop codon was detected in the insertion in EM1_37_F03.b1 in what was determined to be the open reading frame (Fig. 18). The results received here, similar to the previous case, indicate that the 94-nt insertion represents an unspliced intronic region.

Analysis of the remaining three pairs of sequences revealed the reverse of the two previous cases. In these three cases, comparisons to public databases indicate that there appears to have been a deletion rather than an insertion. For the three pairs, the deleted region averaged 142 nt. In each of the three cases, BLAST comparisons to the non-redundant nucleotide database and the EST database showed the additional region to be present in all significantly similar sequences. Thus the additional region appeared to be a normal part of the sequence read and not an intron as in the previous two cases. The absence of the additional region suggests these sequences may be the product of an alternatively spliced transcript in which a portion of the coding region is lost (Fig. 19). In all three cases BLAST searches against the protein database revealed no significant similarities. Additionally, in all three cases an open reading frame could not be successfully identified as stop codons were present sporadically in all three frames. In these last three cases additional information is required in order to determine whether each pair represents alternatively spliced transcripts.

Figure 16. An illustration of the difference identified between each sequence pair. EST A represents a sequence with an insertion relative to EST B. With the exception of an extension at one or both ends (indicated by the blue arrows) and the insertion in EST A, the two sequences aligned with 100% identity.

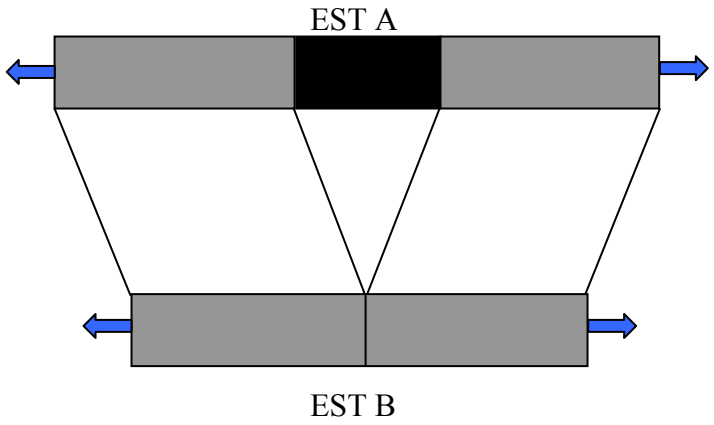

EST A
EST B

Figure 17. Sequence comparison of EM1_6_H05.b1 to EM1_23_B02.b1. The highly conserved splice site donor and acceptor motifs (GT and AG, respectively) are in red. Corresponding junction consensus sequences are underlined. The sequence alignment indicates that the region retained in EM1_6_H05.b1 is an intron. The open reading frame was identified after translation of the sequences in the three forward reading frames. The amino acids shown on blue are those present in the insertion region. The star within the insertion region identifies a stop codon in the open reading frame.

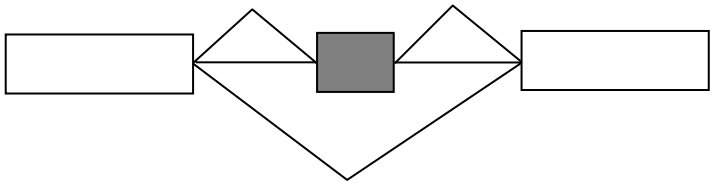
```
EM1_6_H05.b1      CCACCACCCA ATCTGAAATA CAGCAACTCG CAGTCGATCG AGACATGAAG

EM1_23_B02.b1     CCACCACCCA ATCTGAAATA CAGCAACTCG CAGTCGATCG AGACATGAAG
                                                             M  K


EM1_6_H05.b1      CACTCGAGCA GCTTGTGTTT GCTCTTCCTC GTGGCGCTCT GCATCAGCCT
EM1_23_B02.b1     CACTCGAGCA GCTTGTGTTT GCTCTTCCTC GTGGCGCTCT GCATCAGCCT
                  H  S  S  L   L  C  L   L  F  L   V  A  L  C   I  S  L

EM1_6_H05.b1      GGCCTGCGGC CTGGTTCAGG CACAGGTCCT CTTCCAGGTA CGAGTACGAC
EM1_23_B02.b1     GGCCTGCGGC CTGGTTCAGG CACAGGTCCT CTTCCAG... ..........
                   A  C  G   L  V  Q  A   Q  V  L   F  Q  V   R  V  R  L

EM1_6_H05.b1      TACGAAAGCT GTACACGTCA ATGTTGTCGT ACGTACGATG ACGATCTCGT
EM1_23_B02.b1     .......... .......... .......... .......... ..........
                    R  K  L   Y  T  S   M  L  S  Y   V  R  *   R  S  R

EM1_6_H05.b1      CAAACTTATC ATGCCACAAT CTGTGCACGT ACGTGCAGGG GTTTAACTGG
EM1_23_B02.b1     .......... .......... .......... ........GG GTTTAACTGG
                  Q  T  H  Y   A  T  I   C  A  R   T  C  R  G   F  N  W

EM1_6_H05.b1      GAGTCGTGCA AGCAGCCTGG AGGCTGGTAC AACAGGCTCA AGGCCCAGGT
EM1_23_B02.b1     GAGTCGTGCA AGCAGCCTGG AGGCTGGTAC AACAGGCTCA AGGCCCAGGT
                  E  S  C  K   Q  P  G   G  W  Y   N  R  L  K   A  Q  V

EM1_6_H05.b1      CGACGACATC GCCAACGCCG GCGTCACGCA CGTCTGGCTG CCTCCACCCT
EM1_23_B02.b1     CGACGACATC GCCAACGCCG GCGTCACGCA CGTCTGGCTG CCTCCACCCT
                   D  D  I   A  N  A  G   V  T  H   V  W  L   P  P  P  S

EM1_6_H05.b1      CGCACTCCGT CTCGCCACAA GGGTACATGC CGGGGCGCCT ATACGACCTG
EM1_23_B02.b1     CGCACTCCGT CTCGCCACAA GGGTACATGC CGGGGCGCCT ATACGACCTG
                    H  S  V   S  P  Q   G  Y  M  P   G  R  L   Y  D  L

EM1_6_H05.b1      GACGCGTCCA AGTACGGCAC GGCGGCGGAG CTCAGG
EM1_23_B02.b1     GACGCGTTCA AGTACGGCAC GGCGGCGGAG CTCAGG
                  D  A  F  K   Y  G  T   A  A  E   L  R
```

Figure 18. Sequence comparison of EM1_37_F03.b1 to EM1_14_G09.b1. The consensus splice donor and acceptor sites are shown in the red, additionally the consensus junction sequences are underlined. Shown in blue are the amino acid sequences present in the open reading frame of the insertion region. The blue star present at the start of the insertion region identifies a stop codon found in the open reading frame.

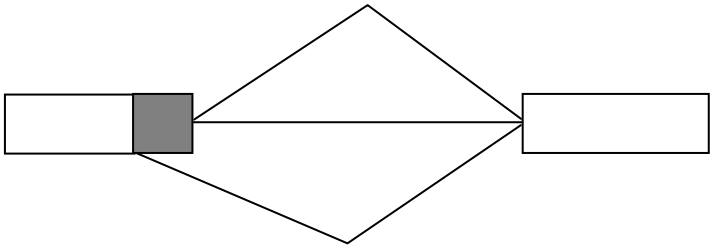
```
EM1_37_F03.b1     GCACGAGGCT CACCCCTGCT TCCTCGGGCG GCGGCGCCGC CGCCGTTTGG
EM1_14_G09.b1     GCACGAGGCT CACCCCTGCT TCCTCGGGCG GCGGCGCCGC CGCCGTTTGG
                  A  R  G  S   P  L  L   P  R  A   A  A  P  P   P  F  G

EM1_37_F03.b1     CGCACGCGCG ATGCCGGCGA CGGCGGAGGA CTCCCCGGCG ATACGGAAGC
EM1_14_G09.b1     CGCACGCGCG ATGCCGGCGA CGGCGGAGGA CTCCCCGGCG ATACGGAAGC
                   A  R  A   M  P  A  T   A  E  D   S  P  A   I  R  K  L

EM1_37_F03.b1     TGGGGCAGCT CTTCCGGCTG ACCGAAGTAT ACCTCTGGTA AGTCTCCTGT
EM1_14_G09.b1     TGGGGCAGCT CTTCCGGCTG ACCGAAGTAT ACCTCTG... ..........
                    G  Q  L   F  R  L   T  E  V  Y   L  W  *   V  S  C

EM1_37_F03.b1     CTCCCCTACC CGCTTCTCAG CGTAGGGTTT GCCATTTCGC GAGGAGTGTC
EM1_14_G09.b1     .......... .......... .......... .......... ..........
                  L  P  Y  P   L  L  S   V  G  F   A  I  S  R   G  V  S

EM1_37_F03.b1     CCGTCTCCTC AAACTGCGCT CTCTCCCTGC AGGGACGATT CGTATGGCGC
EM1_14_G09.b1     .......... .......... .......... ..GGACGATT CGTATGGCGC
                   R  L  L   K  L  R  S   L  P  A   G  T  I   R  M  A  L

EM1_37_F03.b1     TGGACCTCAC GATGGACGGA AGAACTGGCG CTCGGCGGAG GCTGCTCTCG
EM1_14_G09.b1     TGGACCTCAC GATGGACGGA AGAACTGGCG CTCGGCGGAG GCTGCTCTCG
                    D  L  T   M  D  G   R  T  G  A   R  R  R   L  L  S

EM1_37_F03.b1     TGGACTCTCA TACAGATAAA ACATGCAATG AAGCATCCAA TGGCACTGAC
EM1_14_G09.b1     TGGACTCTCA TACAGATAAA ACATGCAATG AAGCATCCAA TGGCACTGAC
                  W  T  L  I   Q  I  K   H  A  M   K  H  P  M   A  L  T

EM1_37_F03.b1     AAAGGCCATT TGTTTGTTGA AGATTTGG
EM1_14_G09.b1     AAAGGCCATT TGTTTGTTGA AGATTTGG
                   K  A  I   C  L  L  K   I  W
```

Figure 19. Types of alternative splicing that could have resulted in the loss of a portion of the coding region. A) Exon inclusion/exclusion. In the upper portion two flanking introns are spliced while the exon in the middle is retained. In contrast, the lower portion shows the removal of both introns together with the middle exon. B) Alternative 5′ site. In this example the lower portion shows an alternative 5′ splice site being used as opposed to the normal 5′ splice site shown in the upper portion. C) Alternative 3′ site. In this example the lower portion shows an alternative 3′splice site in contrast to the normal 3′ splice site used in the upper portion.

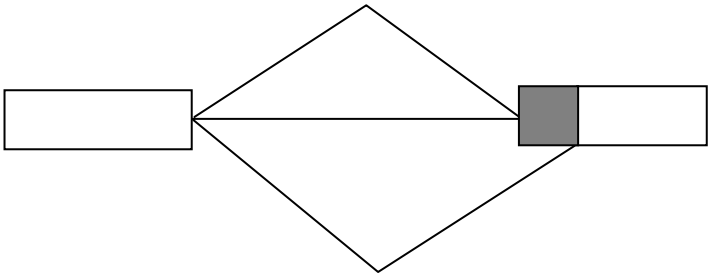
A)



B)



C)



CHAPTER 4

DISCUSSION

In this study, a total of 5,126 and 5,405 (3′ and 5′, respectively) high quality ESTs were produced from a *Sorghum* germinating embryo cDNA library. The sequences were used to conduct a number of analyses in order to attain both quality control and expression information.

Quality control

A. Verifying samples were labeled correctly

Accurate sample tracking minimizes the possibility of identification errors in a library, and thus allows for proper sequence analysis. Moreover, for our purposes there must be correspondence between a sequence submitted to GenBank and a clone in permanent storage, because all clones are made available for outside use. To ensure the accuracy of sample tracking a subset of plates were analyzed for identification errors. These errors could lead to discrepancies between the archived clones and the ESTs derived from them.

Plates containing relatively few overlapping forward and reverse sequences were analyzed. These plates were selected because identification errors would result in such plates. For example, if a plate containing ESTs sequenced from the 3′ end and the corresponding plate containing ESTs sequenced from the 5′ end were mislabeled during the thermal cycle step of sequence production, it is likely that by chance none or only a few sequences would overlap. Aside from identification error, a low number of

overlapping forward and reverse sequences could also be due to a plate containing short sequences as a result of poor sequence quality, or having too few clones that produce acceptable data.

The analysis presented here revealed that identification error was not the cause of the low number of overlapping forward and reverse sequences. The 3′ sequences examined, for the most part, grouped into the same cluster as their corresponding 5′ sequences (Table V). Comparative analysis using plates with a relatively low, average and high number of overlapping forward and reverse sequences revealed that the number of overlapping forward and reverse sequences obtained from a plate was correlated with the number of submittable sequences from that same plate (Fig. 11). Not surprisingly, the number of overlapping forward and reverse sequences was also directly correlated with the average sequence read length for a plate (Fig. 12). At no time was identification error observed in this analysis. Thus, the relatively low number of overlapping forward and reverse sequences observed in the plates analyzed was attributed to poor sequence quality, rather than poor quality control.

B. Library quality

To evaluate the extent of *E. coli* contamination embryo EST sequences were used to perform BLASTN searches against the non-redundant nucleotide database. A total of 12 EST sequences were identified with a bit score of at least 99 and with at least 91% sequence identity to *E. coli* DNA sequences (Table VI). All significant similarities observed were to genomic DNA, a likely source of contamination. Although low this

finding indicates that the *E. coli* screening that occurred during sequence processing was not fully effective.

Among the 12 EST sequences with significant similarities to *E. coli* genomic DNA, one was observed in which the two ends of the same cDNA were significantly similar to sequences from very different organisms. This was the first and only detection of a chimeric cDNA clone in the EM1 library. An analysis using the other 11 contaminant ESTs and 25 other ESTs selected randomly did not detect another chimeric cDNA clone. A more detailed study will, however, be required to reach a more definitive conclusion regarding the frequency of chimeric cDNA clones.

Annotation and Classification

A total of 3,927 ESTs returned significant similarities to sequences in SWISS-PROT. The ESTs were grouped into functional categories based on the description of the database sequence (Table VII). Among the functional categories, metabolism contained the most abundant EST sequences. However, when combined the nonspecific categories (*i.e.*, unknown, hypothetical and unclassified) account for the greatest number of ESTs.

Generally, ESTs were similar to a wide variety of database sequences. This was expected because a number of processes are underway in a germinating embryo. These include normal cellular processes such a respiration, which can be detected within minutes after imbibition (Bewley, 1997). Enzymes necessary for pathways such as glycolysis and the pentose phosphate shunt, as well as the Krebs cycle, become activated (Nicolás and Aldasoro, 1979; Salon et al., 1988). Protein synthesizing complexes are formed to translate newly transcribed mRNAs as well as mRNAs that were present in

dormant embryos. Thus, proteins such as ribosomal proteins and translation initiation and elongation factors are mobilized. Lipids such as triacylglycerols, which are the primary energy reserves supporting germination (Boothe et al., 1997) are metabolized; thus, proteins involved in lipid metabolism are also present.

As can be expected, the most abundant ESTs within most functional categories were derived from 'housekeeping' genes. Among genes present in the protein synthesis functional category, ribosomal proteins and translational initiation and elongation factors were most abundant. Among genes that play a role in protein fate, heat shock proteins and other molecular chaperones were most abundant. Among proteins involved in metabolism, there were a number of genes such as amylogenin and hydroxylase, which play a role in carbohydrate and lipid metabolism, respectively. Additionally genes that play a role in development and defense such as phytochrome and dehydrin were also detected.

Overall, a great majority (73%) of the EST sequences in the embryo library could not be characterized because they either did not exhibit significant similarity to an annotated database sequence, or the sequence to which they were similar was not characterized. This finding still leaves somewhat of a shroud over the canvas that is gene expression in the embryo library. Taking into account the observation that embryo ESTs had average high quality sequence lengths of 508 nt and 521 nt (3′ and 5′ ends, respectively) it is evident that a majority of the embryo sequences are not represented in the public database. Only in time, as more sequences are characterized and deposited in the public database will the mystery surrounding the origin of a majority of ESTs in the embryo dataset be revealed.

Sequence Clustering

A. Analysis of singleton ESTs and contigs-of-one

Singleton ESTs and contigs-of-one were analyzed in order to determine their likely origin. Singleton ESTs are sequences that do not correspond to any other EST in the same dataset. Similarly, contigs-of-one are sequences selected by Phrap for clustering with other sequences; however, they failed to cluster successfully and as result each became the only member in its contig. Similarity comparisons were used to identify the likely origin of both singletons and contigs of one.

The sequences with no significant similarity to known genes after the BLASTN searches were further analyzed by BLASTX searches. BLASTX searches have the advantage of being more sensitive than BLASTN searches due to degeneracy at the DNA level. Different codons can encode the same amino acid. This means two identical protein sequences can differ substantially at the DNA level. This enhanced sensitivity was validated as BLASTX searches revealed 15 singleton sequences with significant similarity to database proteins. Analysis of the BLASTX results revealed that of the singleton sequences that had significant similarity to proteins in the database, a great majority of them were to proteins with unknown functions from model systems such as *Arabidopsis thaliana* and *Oryza sativa* (Table X). Similarity to these plant species suggests that the singleton sequences more than likely originated from the *Sorghum* embryo cDNA library and not from a bacterial or fungal contaminant. Singleton sequences with no significant similarities after searches to the nucleotide and protein databases were searched against the EST database. Searches to the EST database were conducted to determine whether there were similar uncharacterized ESTs elsewhere. The

presence of ESTs in the database from plant cDNA libraries similar to 12 embryo singletons analyzed strengthens the case that this particular group of singletons originated from the *Sorghum* embryo cDNA library. The remaining 16 singletons returned no significant similarity and perhaps represent novel genes.

In the overall singleton analysis two ESTs were found to have significant similarity to fungal sequences. In one case, the singleton sequence was found to encode an amino acid sequence 57% identical to a yeast protein. An *Arabidopsis* homologue of the yeast protein was observed to be 58% identical. The similar identity suggested that the EST sequence might also be a homologue of the yeast protein; thus, it was concluded that the singleton sequence originated from the Sorghum embryo cDNA library. The other case involved a singleton that shared significant similarity to a cDNA sequence derived from *Botrytis cinerea.* Similarity to the plant pathogen indicates that the singleton sequence is likely a contaminant.

Contigs of one were analyzed in the same manner as singletons. Of the 20 contigs of one, 16 displayed significant similarity to sequences in the nucleotide database. All 16 database sequences were derived from grasses. As a result the 16 contigs of one were concluded to have been derived from the *Sorghum* embryo cDNA library. Of the four searched against the protein database, three returned significant similarities to database sequences, two of which were to *Arabidopsis* sequences. One returned similarity to a fungal sequence and was determined to be a contaminant. The lone contig-of-one the remained was searched against the EST database where it was observed to similar to many plant ESTs.

In total, 97% of the singletons and 95% of the contigs-of-one sequence analyzed were concluded to have originated from the *Sorghum* embryo cDNA library. Analysis of this random subset of sequences suggests that the vast majority of singletons and contigs-of-one in the dataset originated from the *Sorghum* embryo cDNA library. These sequences thus represent excellent candidates for analysis of transcripts present at very low levels in the library.

B. Assembly of a virtual unigene set

A unigene set is comprised of sequences assembled into contigs, with each contig ideally representing a unique gene. Sequences derived from the 3′ end of the insert were used for unigene assembly. The 3′ sequences were selected because cDNA inserts were created with an oligo dT primer, which barring events such as alternative priming or polyadenylation at different sites, should have begun reverse transcription at the same point. Conversely, the starting point from the 5′ ends of similar inserts can vary based on differing points of detachment by the reverse transcriptase during first-strand cNDA synthesis. Consequently, sequencing from the 5′ end will yield ESTs from the same transcript at different starting points (Rounsley et al., 1996). Therefore, the use of 5′ sequence assemblies to define a unigene set would result in possible overestimation of the number of unigenes. Using 3′ sequences to define a unigene set is also not without flaws, however. Events such as internal priming (Hillier et al., 1996) during cDNA library construction, alternative splicing or even the presence of multiple polyadenylation sites for the same gene (Ewing and Green, 2000) can also lead to overestimation of the

number of unigenes. In spite of the events mentioned, the 3′ end still represents the better of the two ends with which to assemble a unigene set.

Analysis of 3′ ESTs from sorghum germinating embryos permitted the identification of putatively 2,721 uniquely expressed transcripts. It is important to note that 2,721 does not represent all the genes expressed in the embryo; rather, it represents a unigene set based on the limited number of high quality 3′ EST sequences produced from the embryo library. Continued sequencing of the embryo library would certainly increase the unigene set number. However, as the maximum number of unique transcripts present in the library is approached, the unigene set number would level off.

The sequences analyzed here were derived from a non-normalized and unamplified library. Since no prior adjustments were made, all high quality sequences were considered; thus our EST dataset represents an approximate, albeit incomplete, snapshot of the mRNA population in the embryo tissue.

An EST unigene set provides an invaluable resource for identification and isolation of genes in the embryo involved in germination. While the unigene set will undoubtedly contain housekeeping genes expressed throughout development, many of the genes expressed during germination are distinct from those expressed during subsequent development (Hughes and Galau, 1989; Kermode, 1990; Berry and Bewley, 1991). In addition, in many species it is known that plant growth hormones, notably gibberellins, are required to stimulate germination by inducing the expression of germination-specific genes (Hilhorst and Karssen, 1992; Jacobsen et al., 1995). Embryo genes involved in germination are likely to play a variety of roles in assuring proper development, including assistance in the weakening of the endosperm coat to allow for radicle emergence

(Watkins and Cantliffe, 1983; Groot and Karssen, 1897; Ni and Bradford, 1993; Sánchez and de Miguel, 1997). A unigene dataset is thus an important investigative tool that can be used, among other things, to identify probable embryo gene transcripts involved in germination.

C. Identification of embryo-specific genes

Embryo ESTs were assembled into contigs along with ESTs from a variety of other cDNA libraries (Table II). The result of the assembly was analyzed in order to identify gene transcripts preferentially expressed in the embryo. The analysis revealed that 441 embryo ESTs were assembled into 131 embryo-only contigs. BLAST analysis was conducted to reveal the identity of these embryo ESTs. The BLAST analysis revealed that of 131 contigs analyzed, 94 (72%) returned either no significant similarity, or similarity to an unknown or hypothetical protein. Table XII lists the results of the 37 contigs that displayed significant similarity to known proteins. A variety of proteins were identified that play a role in germination and plant development. Collectively, oil body related proteins such as oleosin, caleosin, steroleosin and an oil body protein were the most abundant. Oil bodies are widely distributed throughout plant cells, but most abundant in seeds. In seeds, oil bodies are present in the embryo (Wu et al., 1998) and aleurone layers. They store triacylglcerols, the primary energy reserve used during germination (Boothe et al., 1997). Structural proteins such as oleosin are found on the surface of oil bodies (Tzen et al., 1992) providing structural support during dessication.

The largest contig assembly displayed similarity to S-adenosylmethionine decarboxylase 2, an enzyme involved in polyamine biosynthesis. Seed storage proteins such as globulins were also identified. Globulins are synthesized during seed maturation.

During germination, globulins are broken down by enzymes such as serine carboxypeptidase (also identified) to serve as an initial food source (Bewley and Black, 1994; Shewry et al., 1995). Surprisingly, two separate heat shock proteins were identified in the analysis. Heat shock proteins participate in diverse cellular processes by acting as molecular chaperones (Hong and Vierling, 2000). The presence of the proteins preferentially in the embryo library suggests that they perform a protective function specifically during germination, perhaps similar to pea seed germination (DeRocher and Vierling, 1995). A variety of other proteins including phytochrome A, which is involved in development, and alpha expansin, which is involved in radicle protrusion, were also identified.

The proteins identified represent only minor sampling of the transcripts preferentially expressed in the embryo. A majority of the contigs analyzed (72%) were not able to be identified by BLAST comparison. Contigs containing up to 12 ESTs were among those that returned no significant similarity. Although not quantified, a majority of the contigs that returned no significant similarity did not simply report low similarity scores; rather, they were not even remotely similar to any sequence in the database. This observation suggests that a number of embryo-only contigs might represent rare transcripts that have yet to be seen in databases.

Identification of the relative abundance of full-length coding cDNAs

ESTs from the 5′ end of cDNA inserts were used to determine the relative abundance of full-length coding sequences in the EM1 library. Analysis was conducted by performing BLAST comparisons against SWISS-PROT. Full length coding

sequences were determined by identifying EST query sequences that encode the first published amino acid residue of the subject sequence.

Of the 5,405 high quality 5′ EST sequences, 3,103 displayed significant similarity to database sequences. Of these 3,103 sequences 833 or 27% were identified as containing full-length coding sequences. It was previously observed that conventional library construction normally yields fewer then 20% full-length cDNA clones (Kato et al., 1994). Thus, 27% stands as an above average result.

Only 60% (3103/5405) of the total high quality 5′ EST sequences returned significant similarities. It is likely that as more sequences are deposited into SWISS-PROT, the number of EST sequences with significant similarity to database sequences will increase. As the number of ESTs with significant similarity increases, the number of full-length inserts identified in the library is expected to increase as well.

Analysis of EST sequences potentially spliced alternatively

A preliminary bioinformatic analysis of the EM1 ESTs was conducted to identify cDNA clones derived from putative alternatively spliced transcripts. Comparative analysis identified five pairs of EST sequences with differences between them that could have arisen from alternative splicing. Of the five pair of sequences analyzed, intron retention was observed in two and probable exon exclusion had occurred in three. The study reported here, similar to other bioinformatic analyses of alternative splicing (Brett et al., 2000; Modrek et al., 2000; International Human Genome Consortium, 2001), relied on identifying EST sequences that presumably came from the same gene, and then looking for insertion/deletion differences between them. This particular study differs in

that genomic sequences (Mironov et al., 1999; Kan et al., 2001) or annotated intronic sequences (Croft et al., 2000) were not available; instead, similarity comparisons to dbEST and the non-redundant nucleotide and protein databases were the means utilized for comparative analysis. Additionally, translation in the three forward frames in order to identify an open reading frame was utilized when necessary.

Conclusion

EST sequence analysis provided valuable information about the *Sorghum* germinating embryo cDNA library. Evaluation of the library quality revealed that although a small percentage of contamination was observed, most of the ESTs were derived from the cDNA library. Expression analysis offered insight into the numerous processes taking place in the germinating embryo as it prepares to emerge from the seed coat. In all, much information was gained from the analyses, yet at the same time there remains a great deal more to learn. Expression analysis in particular, was akin to finding a treasure chest known to contain riches that upon opening revealed a smaller chest.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252:** 1651-1656

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2,375 human brain genes. Nature **355:** 632-634

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

Altshcul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389-3402

Anderson MB, Folta K, Warpeha KM, Gibbons J, Gao J, Kaufman LS (1999) Blue light-directed destabilization of the pea Lhcb1*4 transcript depends on sequences within the 5′ untranslated region. Plant Cell **11:** 1579-1590

Arumuganathan E, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9:** 208-218

Avramova Z, Tikhonov A, SanMiguel P, Jin YK, Liu C, Woo SS, Wing RA, Bennetzen JL (1996) Gene identification in a complex chromosomal continuum by local genomic cross-referencing. Plant J. **10:** 1163-1168

Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli M (2000) The EMBL nucleotide sequence database. Nucleic Acids Res. **28:** 19-23

Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J. Mol. Med. **75:** 312-316

Banfi S, Borsani G, Rossi E, Bernard L, Guffanti A, Rubboli F, Marchitiello A, Giglio S, Coluccia E, Zollo M, Zuffardi O, Ballabio A (1996) Identification and mapping of human cDNAs homologous to *Drosophila* mutant genes through EST database searching. Nat. Genet. **13:** 167-174

Bashirullah A, Cooperstock RL, Lipshitz HD (1998) RNA localization in development. Ann. Rev. Biochem. **67:** 335–394

Bennett ST, Leitch IJ, Bennett MD (1995) Chromosome identification and mapping in the grass *Zingeria biebersteiniana* (2n=4) using fluorochromes. Chromosome Res. **3:** 101-108

Bennetzen JL, Freeling M (1997) The unified grass genome: synergy in synteny. Genome Res. **7:** 301-306

Bennetzen JL, Schrick K, Springer PS, Brown WE, SanMiguel P (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome **37:** 565-576

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) GenBank. Nucleic Acids Res. **28:** 15-18

Benton D (1996) Bioinformatics - principles and potential of a new multidisciplinary tool. Trends Biotechnol. **14:** 261-272

Berhan AM, Hulbert SH, Butler LG, Bennetzen JL (1993) Structure and evolution of the genomes of *Sorghum bicolor* and *Zea mays*. Theor. Appl. Genet. **86:** 598-604

Berry T, Bewley JD (1991) Seeds of tomato (*Lycopersicon esculentum* Mill) which develop in a fully hydrated environment in the fruit switch from a developmental to a germinative mode without a requirement for desiccation. Planta **186:** 27-34

Bewley JD (1997) Seed germination and dormancy. Plant Cell **9:** 1055-1066

Bewley JD, Black M (1994) Seeds: physiology of development and germination. New York: Plenum Press.

Boguski MS (1998) Bioinformatics - a new era. Trends Bioinformatics **suppl:** 1-3

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". Nat. Genet. **4:** 332-333

Boothe J, Saponja J, Parmenter D (1997). Molecular farming in plants: oilseeds as vehicles for the production of pharmaceutical proteins. Drug Development Res. **42:** 172-181

Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, Krueger S, Reich J, Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. **474:** 83-86

Chittenden LM, Schertz KF, Lin Y-R, Wing RA, Paterson AH (1994) A detailed RFLP map of *Sorghum bicolor* x *S. Propinquum*, suitable for high-density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. Theor. Appl. Genet. **87:** 925-933

Clayton WD (1987) Andropogoneae. *In* TR Soderstrom, KW Hilu, CS Campbell, ME Barkworth, eds, Grasses Systematics and Evolution. Smithsonian Institution Press, Washington, DC, pp 307-309

Comai L, Harada JJ (1990) Transcriptional activities in dry seed nuclei indicate the timing of transition from embryogeny to germination. Proc. Natl. Acad. Sci. USA **87:** 2671-2674

Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick J (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. **24:** 340-341

Cuming AC, Lane BG (1978) Wheat embryo ribonucleates XI. Conserved mRNA in dry wheat embryos and its relation to protein synthesis during early inhibition. Can. J. Biochem. **56**: 365-369

Cuming AC, Lane BG (1979) Protein synthesis in imbibing wheat embryos. Eur. J. Biochem. **99:** 217-224

Curie C, McCormick S (1997) A strong inhibitor of gene expression in the 5′ untranslated region of pollen-specific LAT59 gene of tomato. Plant Cell **9:** 2025-2036

Dayhoff MO (1969) Computer analysis of protein evolution. Sci. Amer. **221:** 86-95

Decker CJ, Parker R (1994) Mechanisms of mRNA degradation in eukaryotes. Trends Biochem. Sci. **19:** 336-340

Delsney M, Aspart L, Guitton Y (1977) Disappearance of stored polyadenylic acid and mRNA during early germination of radish (*Raphanus sativus* L.*)* embryo axis. Planta **135:** 125-138

DeRocher A, Vierling E (1995) Cytoplasmic HSP70 homologues of pea: differential expression in vegetative and embryonic organs. Plant Mol. Biol. **27:** 441-456

Devos KM, Gale MD (1997) Comparative genetics in the grasses. Plant Mol. Biol. **35:** 3-15

Doggett H (1988) Sorghum, Ed 2. John Wiley and Sons, Inc., New York

Draye X, Lin YR, Qian X, Bowers E J, Burrow BG, Morrell LP, Peterson GD, Presting GG, Ren S, Wing R, Paterson AH (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiol. **125:** 1325-1341

Dunford RP, Kurata N, Laurie DA, Money TA, Minobe Y, Moore G (1995) Conservation of fine-scale DNA marker order in the genomes of rice and the Triticeae. Nucleic Acids Res. **23:** 2724-2728

Ewing B, Green P (2000) Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. **25:** 232-234

Ewing B, Hillier L, Wendl M, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175-185

Ewing R, Poirot O, Claverie JM (2000) Comparative analysis of *Arabidopsis* and rice expressed sequence tag (EST) sets. In Silico Biol. **4:** 197-213

Filipowicz W (2000) Imprinted expression of small nucleolar RNAs in brain: time for RNomics. Proc. Natl. Acad. Sci. USA **97:** 14035-14037

Filipowicz W, Pelczar P, Pogacic V, Dragon F (1999) Structure and biogenesis of small nucleolar RNAs acting as guides for ribosomal RNA modification. Acta Biochim. Pol. **46**: 377-389

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb J-F, Dougherty BA, Bott KF, Hu P-C, Lucier TS, Peterson SN, Smith HO, Hutchinson III CA, Venter JC (1995) The *Mycoplasma genitalium* genome sequence reveals a minimal gene complement. Science **270:** 397-403

Gerhold D, Caskey TC (1996) It's the gene! EST access to human genome content. BioEssays **18:** 973-981

Gibbs L, Willis D, Morgan MJ (1998) The identification and expression of heme oxygenase-2 alternative transcripts in the mouse. Gene **221:** 171-177

Goldmark PJ, Curry J, Morris CF, Walker-Simmons MK (1992) Cloning and expression of an embryo-specific mRNA upregulated in hydrated dormant seeds. Plant Mol. Biol. **19:** 433-441

Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. Genome Res. **8:** 195-202

Groot SPC, Karssen CM (1987) Gibberellin regulates seed germination in tomato by endosperm weakening: a study with gibberellin-deficient mutants. Planta **171:** 525-531

Guimaraes CT, Sills GR, Sobral BWS (1997) Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. Proc. Natl. Acad. Sci. USA **94:** 14261-14266

Hair J, Beuzenberg E, (1961) High polyploidy in a New Zealand Poa. Nature **189:** 160

Han B, Hughes DW, Galau GA, Bewley JD, Kermode AR (1997) Changes in late embryogenesis abundant (LEA) messenger RNAs and dehydrins during maturation and premature drying of *Ricinus communis* L. seeds. Planta **201:** 27-35

Hanke J, Brett D, Zastrow I, Aydin A, Delbrück S, Lehmann G, Luft F, Reich J, Bork P (1999) Alternative splicing of human genes: more the rule than the exception? Trends Genet. **15:** 389-390

Harris B, Dure III L (1978) Developmental regulation in cotton seed germination: polyadenylation of stored messenger RNA. Biochemistry **17:** 3250-3256

Hillier L, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chiose S, Dietrich N, Dubuque T, Favello A, Gish W (1996) Generation and analysis of 280,000 human expressed sequences tags. Genome Res. **6:** 807-828

Hilhorst HMW, Karssen CM (1992) Seed dormancy and germination: the role of abscisic acid and gibberellins and the importance of hormone mutants. Plant Growth Regul. **11:** 225-238

Hirschberg DS (1975) A linear space algorithm for computing longest common subsequences. Commun. ACM **18:** 341-343

Holdsworth M, Kurup S, McKibbin R (1999) Molecular and genetic mechanisms regulating the transition from embryo development to germination. Trends Plant Sci. **4:** 275-280

Hong SW, Vierling E (2000) Mutants of *Arabidopsis thaliana* defective in the acquisition of tolerance to high temperature stress. Proc. Natl. Acad. Sci. USA **97:** 4392-4397

Huang X (1996) An improved sequence assembly program. Genomics **33:** 21-31

Hughes DW, Galau GA (1989) Temporally modular gene expression during cotyledon development. Genes Dev. **3:** 358-369

Hulbert SH, Richter TE, Axtell JD, Bennetzen JL (1990) Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. Proc. Natl. Acad. Sci. USA **87:** 4251-4255

Hunt JW, Szymanski TG (1977) A fast algorithm for computing longest common subsequences. Commun. ACM **20:** 350-353

International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860-921

Jacobsen JV, Gubler F, Chandler PM (1995) Gibberellin action in germinated cereal grains. In PJ Davies, ed., Plant Hormones: Physiology, Biochemistry and Molecular Biology. Kluwer Academic Publishers, Dordrecht, The Netherlands, 246-271

Jiang L, Kermode AR (1994) Role of desiccation in the termination of expression of genes for storage proteins. Seed Sci. Res. **4:** 149-173

Johnson RR, Cranston HJ, Chaverra ME, Dyer WE (1995) Characterization of cDNA clones for differentially expressed genes in embryos of dormant and nondormant *Avena fatua* L. caryopses. Plant Mol. Biol. **28:** 113-122

Kan Z, Rouchka E, Gish W, States D (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. **11:** 889-900

Kato S, Sekine S, Oh S–W, Kim N–S, Umezawa Y, Abe N, Yokoyama-Kobayashi M, Aoki T (1994) Construction of a human full-length cDNA bank. Gene **150:** 243-250

Kaufman RJ (1994) Control of gene expression at the level of translation initiation. Curr. Opin. Biotechnol. **5:** 550–557

Keller B, Feuillet C (2000) Colinearity and gene density in grass genomes. Trends Plant Sci. **5:** 246-251

Kermode AR (1990) Regulatory mechanisms involved in the transition from seed development to germination. Cri. Rev. Plant Sci. **9:** 155-195

Kermode AR (1995) Regulatory mechanisms in the transition from seed development to germination: Interactions between the embryo and the seed environment. In: Seed

Development and Germination. Galili G, Kigel J, eds. Marcel Dekker, Inc., New York, 273-332

Klausner RD, Roualt TA, Harford JB (1993) Regulating the fate of mRNA: the control of cellular iron metabolism. Cell **72:** 19-28

Koop BF, Hood L (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. Nat. Genet. **7:** 48-53

Lane BG (1991) Cellular desiccation and hydration: Developmentally regulated proteins and the maturation and germination of seed embryos. FASEB J. **5:** 2893-2901

Laurie DA, Bennett MD (1985) Nuclear DNA content in the genera zea and sorghum. Intergeneric, interspecific and intraspecific variation. Heredity **55:** 307-313

Ludlow MM, Muchow RC (1990) A critical evaluation of traits for improving crop yields in water-limited environments. Adv. Agron. **43:** 107-153

Luehrsen KR, Taha S, Walbot V (1994) Nuclear pre-mRNA processing in higher plants. Prog. Nucleic Acid Res. Mol. Biol. **47:** 149-193

Marcand S, Gilson E, Shore D (1997) A protein-counting mechanism for telomere length regulation in yeast. Science **275:** 986-990

Matsubara K, Okubo K (1993) Identification of new genes by systematic analysis of cDNAs and database construction. Curr. Opin. Biotechnol. **4:** 672-677

McCarthy JE, Kollmus H (1995) Cytoplasmic mRNA-protein interactions in eukaryotic gene expression. Trends Biochem. Sci. **20:** 191–197

Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res. **9:** 1288-1293

Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide analysis of alternative splicing using human expressed sequence data. Nucleic Acids Res. **29:** 2850-2859

Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution: Grasses, line up and form a circle. Curr. Biol. **5:** 737-739

Moore MJ (1996) Gene expression. When the junk isn't junk. Nature **379:** 402-403

Murray-Rust P (1994) Bioinformatics and drug discovery. Curr. Opin. Biotechnol. **5:** 648-653

Nathan DF, Lindquist S (1995) Mutational analysis of Hsp90 function: interactions with a steroid receptor and a protein kinase. Mol. Cell Biol. **15:** 3917-3925

Ni BR, Bradford KJ (1993) Germination and dormancy of abscisic acid and gibberellin-deficient mutant tomato (*Lycopersicon esculentum*) seeds: sensitivity of germination to abscisic acid, gibberellin, and water potential. Plant Physiol. **101:** 607-617

Nicolás G, Aldasoro JJ (1979) Activity of the pentose phosphate pathway and changes in nicotinamide nucleotide content during germination of seeds of *Cicer arientinum* L. J. Exp. Bot. **30:** 1163-1170

Nicoloso M, Qu LH, Michot B, Bachellerie JP (1996) Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2′-O-ribose methylation of rRNAs. J. Mol. Biol. **260:** 178-195

Ohlrogge J, Benning C (2000) Unraveling plant metabolism by EST analysis. Curr. Opin. Plant Biol. **3:** 224-228

Okubo K, Hori N, Matoba R, Niiyama T, Matsubara K (1991) A novel system for large-scale sequencing of cDNA by PCR amplification. DNA Seq. **2:** 137-144

Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet. **2:** 173-179

Pandey A, Lewitter F (1999) Nucleotide sequence databases: a gold mine for biologists. Trends Biochem. Sci. **24:** 276-280

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:** 2444-2448

Peng Y, Schertz KF, Cartinhour S, Hart GE (1999) Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breed. **188:** 225-235.

Pereira MG, Lee M, Bramel-Cox P, Woodman W, Doebley J, Whitkus R (1994) Construction of an RFLP map in sorghum and comparative mapping in maize. Genome **37:** 236-243

Prade RA, Ayoubi P, Krishnan S, Macwana S, Russell H (2001) Accumulation of stress and inducer-dependent plant-cell-wall-degrading enzymes during asexual development in *Aspergillus nidulans*. Genetics **157:** 957-967

Pratt WB (1993) The role of heat shock proteins in regulating the function, folding and trafficking of the glucocorticoid receptor. J. Biol. Chem. **268:** 21455-21458

Putney SD, Herlihy WC, Schimmel P (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature **302:** 718-721

Ragab RA, Dronavalli S, Saghai-Maroof MA, Yu YG (1994) Construction of a sorghum RFLP linkage map using sorghum and maize DNA probes. Genome **37:** 590-594

Richmond T, Somerville S (2000) Chasing the dream: plant EST microarrays. Curr. Opin. Plant Biol. **3:** 108-116

Roberts L (1992) NIH gene patents, round two. Science **255:** 912-913

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of Arabidopsis expressed sequence tag assemblies. A new resource to facilitate gene identification. Plant Physiol. **112:** 1177-1183

Rutherford SL, Zuker CS (1994) Protein folding and the regulation of signaling pathways. Cell **79:** 1129-1132

Salon C, Raymond P, Pradet A (1988) Quantification of carbon fluxes through the tricarboxylic acid cycle in early germinating lettuce embryos. J. Biol. Chem. **263:** 12278-12287

Sanchez JE, Aguilar R (1984) Protein synthesis patterns. Relevance of old and new messenger RNA in germinating maize embryos. Plant Physiol. **75:** 231-234

Sánchez RA, de Miguel L (1997) Phytochrome promotion of mannan-degrading enzyme activities in the micropylar endosperm of *Datura ferox* seeds, its relationship with germination. J. Exp. Bot. **37:** 1574-1580

SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. (London) **82:** 37-44

Sen S, Payne PI, Osborne DJ (1975) Early ribonucleic acid synthesis during germination of rye (*Secale cereale*) embryos and the relationship to early protein synthesis. Biochem. J. **148:** 381-387

Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structures and biosynthesis. Plant Cell **7:** 945-956

Singer RH (1992) The cytoskeleton and mRNA localization. Curr. Opin. Cell Biol. **4:** 15-19

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J. Mol. Biol. **147**: 195-197

Spiegal S, Marcus A (1975) Polyribosome formation in early wheat germination independent of either transcription or polyadenylation. Nature **256:** 228-230

Sterky F, Lundeberg J (2000) Sequence analysis of genes and genomes. J. Biotechnol. **76:** 1-31

Sutton G, White O, Adams MD, Kerlavage AR (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci. Technol. **1:** 9-19

Tatusova T, Madden T (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174:** 247-250

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

Tschining C, Meng-Chiang K, Kay-Hooi K, Sadako I, Yasuo I (2000) Developmentally regulated expression of a peptide: *N*-glycanase during germination of rice seeds (*Oryza sativa*) and its purification and characterization. J. Biol. Chem. **275:** 129-134

Tzen JT, Lie GC, Huang AH (1992) Characterization of the charged components and their topology on the surface of plant seed oil bodies. J. Biol. Chem. **267:** 15626-15634

Vasmatzis G, Essand M, Brinkmann U, Lee B, Pastan I (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. Proc. Natl. Acad. Sci. USA **95:** 300-304

Wang Z, Taramino G, Yang D, Liu G, Tingey SV, Miao GH, Wang GL (2001) Rice ESTs with disease-resistance gene- or defense-response gene-like sequences mapped to regions containing major resistance genes or QTLs. Mol. Genet. Genomics **265:** 302-310

Watkins JT Cantliffe DJ (1983) Mechanical resistance of the seed coat and endosperm during germination of *Capsicum annuum* at low temperature. Plant Physiol. **72:** 146-150

Wilhelm JE, Vale RD (1993) RNA on the move: the mRNA localization pathway. J. Cell Biol. **123:** 269-274

Wu L, Wang LD, Chen PW, Chen LJ, Tzen J (1998) Genomic cloning of 18 kDa oleosin and detection of triacylglycerols and oleosin isoforms in maturing rice and postgerminative seedlings. J. Biochem. **123:** 386-391

Xu GW, Magill CW, Schertz KF, Hart GE (1994) An RFLP linkage map of *Sorghum bicolor* (L.) Monech. Theor. Appl. Genet. **89:** 139-145

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X,

Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). Science **296:** 79-92

Zweiger G, Scott WR (1997) From expressed sequence tags to 'epigenomics': an understanding of disease processes. Curr. Opin. Biotechnol. **8:** 684-687