


**Comparative EST Analyses in Plant Systems**

*By Qunfeng Dong, Lori Kroiss, Fredrick D. Oakley, Bing-Bing Wang, and Volker Brendel*

**Abstract**

Expressed sequence tag (EST) data are a major contributor to the known plant sequence space. Organization of the data into non-redundant clusters representing tentative unique genes provides snapshots of the gene repertoires of a species. This chapter reviews availability of sequences and sequence analysis results and describes several resources and tools that should facilitate broad-based utilization of EST data for gene structure annotation, gene discovery, and comparative genomics.

**Introduction**

Expressed sequence tags (ESTs) are generated by high-throughput single-pass sequencing of complementary DNA (cDNA) clones (Adams et al., 1991). A cDNA library is a snapshot representation of the messenger RNA (mRNA) population within a given cell type, tissue, or organism, under a certain set of conditions at a particular moment. Over the past decade, ESTs have accumulated at an exponential rate and have become a major source of plant sequence data. Today more than 2 million plant-derived ESTs from various species are available at public databases. These data have provided a rich resource for gene discovery and annotation (Rudd, 2003). Additional value can be leveraged from these collections by comparing ESTs from multiple species (Fulton et al., 2002; Vincentz et al., 2004). Despite this potential, the resources available for comparative EST analyses in plants remain fairly limited and dispersed. The PlantGDB database (Dong et al., 2004; http://www.plantgdb.org/) was developed to provide informatics tools for comparative plant genomics. Drawing on our
experiences with this project, we discuss various data sources and tools for plant EST analyses and then describe a few case studies on ways EST data may be mined.

Getting the Data

Data Sources

dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) (Boguski et al., 1993), a division of GenBank, is a central repository for all the publicly available EST sequences. Users can search ESTs from plants (as well as other kingdoms) with NCBI’s Entrez system. For example, all the maize ESTs can be retrieved by typing “Zea mays [ORGN]” into the search text box on the dbEST front page. The resulting sequences can subsequently be downloaded in a variety of formats (e.g., FASTA) for users to analyze. The complete Entrez syntax is described at http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html. A convenient shortcut to the data from large-scale plant EST sequencing projects is provided at the NCBI Plant Genomes Central home page (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html). In addition to web retrieval through Entrez, EST data can be obtained from NCBI by anonymous ftp (ftp://ftp.ncbi.nih.gov/repository/dbEST/). Unfortunately, all ESTs in the dbEST ftp repository are bundled, irrespective of sequence origin. Customized parsing of the bulk files is required if users are only interested in a subset of plant species. An alternative data resource is the PlantGDB download site (http://www.plantgdb.org/download.php), which provides EST sequences in FASTA format organized by species. As discussed later in this chapter, PlantGDB also displays EST assembly and annotation results based on the regularly downloaded sets from dbEST. Because of the complexity of that analysis, there will typically be a time lag between the most up-to-date dbEST collections and the subsets of those sequences that were processed at PlantGDB.

In addition to the sequence data, for some applications it is critical to have the quality scores associated with the sequence entry. A quality score is a number between 1 and 100 that gives the relative confidence that a particular base was determined correctly. The scores are generated by the sequencing centers at the time the bases are called. Because of the single-pass nature of EST sequencing, the sequencing error rate is relatively high, estimated in one study at about 3% (Hillier et al., 1996). Sequencing accuracy is probably higher with modern sequencing machines; however, for applications such as single nucleotide polymorphism (SNP) analysis (Garg et al., 1999), the accuracy of each base is critical and must be assessed
from the quality scores. The NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) intends to be a central repository to store trace data generated by sequencing centers. At this point, the vast majority of the trace data posted are from genomic sequencing rather than ESTs. Therefore, users who want to obtain EST trace files will generally have to contact the individual sequence providers for availability. For some plant species, their community database stores trace files. For example, the EST trace files from the Maize Gene Discovery Project (Lunde et al., 2003) are available through MaizeGDB (http://www.maizegdb.org) (Lawrence et al., 2003).

Filtering Out Contaminants

Although various sequence processing suites are widely available to prepare sequences for submission to the databases (Chou and Holmes, 2001; Parkinson et al., 2004; Scheetz et al., 2003), our advice is to still screen the EST sequences deposited at dbEST for contaminating nonnative sequences before conducting survey studies using the ESTs. Some contaminants, such as sequences from plant-associated fungi, may be integral to an EST library preparation because they are not separable at the time of RNA extraction. More typically, contaminations are artifacts of the sequencing process itself. This type of contamination includes cloning vectors, bacterial host, and even human sequences presumably derived from the library preparation team. Such contaminants must be removed before sequence analysis because their presence poses various problems. For instance, if a contaminated sequence was used for primer design, the primer would not allow the intended target region to be amplified. Contaminant sequences found in multiple species might suggest an evolutionary relationship that does not exist. Approaches to identify contaminants for removal are based on matching EST sequences of interest to an annotated database of common contaminants. For example, the UniVec database at NCBI (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) collects a large number of commonly used cloning vector, adapters, and primers.

In addition to the nonnative sequences discussed earlier, analysis of EST sets may be confounded if the ESTs contain sequences that are derived from native repetitive elements. The TIGR Plant Repeat Database (http://www.tigr.org/tdb/e2k1/plant.repeats/) (Ouyang and Buell, 2004) is commonly used to screen plant EST sequences for the presence of sequences derived from retrotransposons and other repetitive elements. To compare ESTs with repeat databases, similarity-search–based programs such as BLAST (Altschul et al., 1997) and cross_match (http://www.phrap.org/phredphrapconsed.html) may be used. Vmatch (Abouelhoda et al., 2004;
(http://www.vmatch.de/) is a less known but highly versatile and efficient sequence comparison program that implements a suffix-array algorithm for fast string comparison. As part of the PlantGDB EST processing pipeline, we apply Vmatch routinely to remove vector contaminations and identify repetitive elements. The percentage of contaminated sequences overall is not large; for example, when processing the maize ESTs at PlantGDB, less than 0.4% of the sequences were found to contain vector sequences. However, because more than 400,000 sequences are in the maize collection, it is still a matter of thousands of sequences, which, if not removed, could confound a whole set analysis. Even when an EST dataset does not match databases of known contaminants or repeats, this does not mean the sequences are free of these types of sequences, because the screening databases, for example, the TIGR Plant Repeat Database, are being added to continually, and will not anytime soon consist of a comprehensive list of all repetitive plant sequences.

Use of Library Information

Information about the source of ESTs (e.g., the tissue type, the stage of development of the plant, and the cultivar) allows the sequence information to be more fully used. For example, for a researcher interested in the initiation of flowering in maize, it would be desirable to compare ESTs derived from floral meristematic tissue in maize to ESTs from other types of tissue. In this way, candidate flowering-specific genes might be identified. Making the ability to perform these kinds of queries easily accessible to plant biologists is an ongoing challenge. Although library information is embedded in the GenBank EST entries, it is very complicated for a user to retrieve sets of ESTs from multiple libraries with specific characteristics from the NCBI site. However, a user can easily do such a search at PlantGDB. This is made possible by organizing the library information (taken from GenBank) into a relational database in PlantGDB, which allows more specific searches than are possible with NCBI-Entrez. For example, even though the requisite information is part of the GenBank records, the execution of a query like “display all the root ESTs generated from maize a particular inbred line” is currently impossible at GenBank. The TableMaker tool at PlantGDB makes this kind of query straightforward (http://www.plantgdb.org/TableMaker.php).

Another consideration when using ESTs to infer expression under specific conditions is how the library was generated. Some libraries are constructed to reflect the relative abundance of the mRNA under certain conditions, at a particular stage, and/or tissue type. Other libraries, referred to as normalized libraries, attempt to capture a higher percentage of lower
copy mRNA by filtering out redundant copies of more abundant mRNAs before constructing the library. If downloaded EST sequence data are used to explore expression profiles, expression level ought only to be inferred from nonnormalized expression libraries. When multiple EST libraries made under different, well-documented conditions exist, they can be used to make inferences about expression differences among the represented tissue sources or developmental stages. For example, Ronning et al. (2003) surveyed the potato transcriptome using ESTs from diverse tissues, in particular tissues challenged with late-blight pathogen. The authors were able to identify a number of candidate genes that were specifically expressed during incompatible interactions with the late-blight pathogen. Results gleaned from comparison between multiple libraries must be statistically supported. Stekel et al. (2000) describe a likelihood ratio method for comparing the level of gene expression from multiple EST libraries to identify differentially expressed genes. The author also provided a nice review of alternative methods for comparing expression levels between two EST libraries.

EST Clustering and Unigene Assembly

Because ESTs usually correspond to only partial cDNA sequences, and because EST samples typically are highly redundant, ESTs are commonly clustered to derive a set of unique putative genes. During this clustering process, ESTs are grouped based on mutual percent identity over a minimum number of overlapping bases. Consensus sequences are derived from the multiple-sequence alignment from each group to provide a tentative “unigene.” EST clustering is computationally intensive given a large redundant set of ESTs, and for this reason specialized databases, listed in Table I, have evolved to make the resulting clusters available to the public.

<table>
<thead>
<tr>
<th>Organization</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Génaplante</td>
<td><a href="http://genoplante-info.infobiogen.fr/">http://genoplante-info.infobiogen.fr/</a></td>
</tr>
<tr>
<td>NCBI dbEST</td>
<td><a href="http://www.ncbi.nlm.nih.gov/dbEST/">http://www.ncbi.nlm.nih.gov/dbEST/</a></td>
</tr>
<tr>
<td>PlantGDB</td>
<td><a href="http://www.plantgdb.org">http://www.plantgdb.org</a></td>
</tr>
<tr>
<td>Sputnik</td>
<td><a href="http://sputnik.btk.fi">http://sputnik.btk.fi</a></td>
</tr>
<tr>
<td>TIGR</td>
<td><a href="http://www.tigr.org/tdb/tgi/plant.shtml">http://www.tigr.org/tdb/tgi/plant.shtml</a></td>
</tr>
</tbody>
</table>
PlantGDB is an example of this type of resource. At PlantGDB, for each species, ESTs are organized in terms of tentative unique genes (TUGs), which are either contigs (EST clusters of two or more sequences; see Fig. 1 for a sample display of an EST contig) or singlets (ESTs that are not significantly similar to any other ESTs). This is achieved by a routine data analysis pipeline. After processing EST sequences with Vmatch to remove contaminants and repeats, the ESTs are clustered with the PaCE program (Kalyanaraman et al., 2003). PaCE employs parallel computing technology to allow fast processing of huge datasets. Therefore, the TUGs can be updated in a timely fashion to reflect the current status of known expressed genes. The program groups sequences that overlap to a certain degree. Then a consensus sequence for each cluster is generated with CAP3 (Huang and Madan, 1999).

In addition to PlantGDB, other databases provide collections of clustered ESTs using different software. The Sputnik project (http://sputnik.btk.fi/) aims to place paralogous sequences in separate clusters using proprietary software. Each EST cluster and the associated peptide are functionally and structurally annotated in an automatic pipeline (Rudd et al., 2003). The TIGR gene indices (http://www.tigr.org/tdb/tgi/plant.shtml) are constructed with the TIGR assembler (Pertea et al., 2003; Quakenbush et al., 2001). In addition to the raw EST data available at NCBI or EBI, used by PlantGDB and Sputnik, the TIGR gene indices are constructed from NCBI gene sequences and expressed transcript (ET) sequences from the TIGR Expressed Gene Anatomy Database (http://www.tigr.org/tdb/egad/egad.shtml). A third collection is the Unigene set maintained by NCBI at GenBank, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene. This collection differs significantly from that of PlantGDB, the Sputnik project, and the TIGR gene indices, as the Unigene clusters are not assembled. Furthermore, all Unigene clusters include sequences providing evidence of the 3’ end of the transcription unit. As a result, not all of the dbEST sequences are represented by a Unigene (Pontius et al., 2003).

Different assembly programs or parameters used for assembly largely explain the differences between the EST clusters provided by different databases, in addition to the differences due to the scope of ESTs analyzed. Efforts to compare different assemblies are currently under way (http://www.phytome.org/). For species with large amounts of genome sequence available, the alignment between ESTs and genomic data can be used as a “gold standard” to calibrate the assembly parameters, that is, ESTs aligned to the same locus should have been clustered (Zhu et al., 2003). However, users of Unigene sets should realize that the parameters being used for assembly were usually developed by trial and error. Thus, when looking at
EST Contig

Sequence ID: ZMtu02-12-23.10040
Title: Tentative Unique Contig assembled from Zea mays ESTs
Type: EST Contig
Length: 1034 bp
Organism: Zea mays

Similar to proteins:
- ID: Q41028
  Description: Rab-like small GTP-binding protein
- ID: Q41332
  Description: Small GTP-binding protein
- ID: Q41028
  Description: RAB1A

Diagram of ZMtu02-12-23.10040 with its component ESTs

Distribution of member ESTs among libraries:

<table>
<thead>
<tr>
<th>Library Name</th>
<th>ESTs</th>
<th># of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>902 - RMD tissue from Weloce Lab (Escherichia coli)</td>
<td>17970010, 17921096, 17921884</td>
<td>3</td>
</tr>
<tr>
<td>1000 - Unigene from Maize Genome Project</td>
<td>17968372</td>
<td>1</td>
</tr>
<tr>
<td>1111 - Unigene from Maize Genome Project</td>
<td>52403420</td>
<td>2</td>
</tr>
<tr>
<td>10447 - N. rhynchoriana</td>
<td>14249329, 14249229</td>
<td>2</td>
</tr>
<tr>
<td>114 - root cDNA library from Welbot Lab</td>
<td>5649295, 5947014, 5296361</td>
<td>3</td>
</tr>
</tbody>
</table>
the EST clusters provided by the databases, they should be considered
heuristic results. Although reclustering the entire EST dataset is typically
out of the computational capabilities of general biologists, users are encour-
aged to recluster subsets of their ESTs of interest with different assembly
programs or parameters to evaluate the robustness of results for themselves.

EST Annotation

If a genome were to be completely annotated, the function, gene
structure, location, and evolutionary history of each gene would have to
be determined. Gene structure annotation, via the spliced alignment of
ESTs to genomic DNA, is covered in the last section of this chapter. To
facilitate physical mapping in other species, their ESTs were compared to
the *Arabidopsis* genome and results displayed on maps of the *Arabidopsis*
chromosomes (http://genopole.toulouse.inra.fr/bioinfo/Iccare/) (Muller
et al., 2004). Similar attempts to leverage whole genome information for
model species can be anticipated as more genomes are being
finished.

As a step toward the elusive goal of whole genome annotation, ESTs
themselves can be functionally annotated, and this is commonly done to
give a general idea of what kinds of genes are expressed by an organism.
Ideally all annotation would be either done or checked manually by expert
curators. Because it is impossible to functionally annotate tens of
thousands of ESTs this way, major EST databases rely on automated
BLAST searches to assign putative function based on sequence similarity
to gene products that have previously been functionally annotated. For
example, at PlantGDB, contig consensus sequences and singleton se-
quenues are compared to the protein database GenPept at NCBI using
BLASTX. The top three hits below the default E-20 cutoff value are
assigned as putative functional homologues of the protein products coded
for in the EST contig. The NCBI Unigene collection is annotated by
providing protein similarity data for one representative sequence from
each cluster to proteins from a group of eight model organisms (including
*Arabidopsis*). Obviously, such functional annotation provided at any EST

---

**Fig. 1.** A screenshot of an expressed sequence tag (EST) contig page displayed at PlantGDB. On the left from top to bottom: contig identification, links to similar proteins that were assigned by BLASTX searches against protein databases, and a schematic diagram of the EST assembly into the contig. Each member EST is linked to its sequence and information about the library in which the EST was generated. On the right are links for downloading either the consensus sequence of the contig or all the sequences of the component ESTs, as well as links to analysis tools such as BLAST@PlantGDB.
database, often retrieved through general text search tools, is just a suggestion, suffering from potential inaccurate transfer of functional assignments and cascading of annotation errors. However, automated annotation provides a starting point, and users can and should always reannotate their sequences of interest. Different programs or databases can be used, parameters can be changed, and alignments can be manually examined. This may reduce the propagation of errors due to the limitations of automated annotation. At PlantGDB, end-user functional annotation is facilitated by links to analysis tools from each record display page (Fig. 1).

A potential barrier to the custom annotation of user-derived EST datasets is the cumbersome nature of the BLAST output generated from running searches for functional annotation purposes. Several programs can be used to process BLAST output, such as BEAUTY (Worley et al., 1998), PhyloBLAST (Brinkman et al., 2001), MuSeqBox (Xing and Brendel, 2001), and Zerg (Paquola et al., 2003). MuSeqBox was designed specifically to meet the need for a convenient way of examining multiquery sequence BLAST output during large-scale EST annotation projects. It parses the informative parameters of BLAST hits into a tabular form, from which subsets of BLAST hits can be selected according to user-specified criteria. For example, MuSeqBox can screen the BLASTX output from an EST versus protein database search to indicate sequences that potentially represent full-length coding sequences. Extensive tutorials for the use of MuSeqBox are available at http://bioinformatics.iastate.edu/bioinformatics2go/mb/MuSeqBox.html.

Often biologists want to derive open reading frames (ORFs) from ESTs so that protein sequences can be used for further analysis (e.g., domain analysis at protein motif databases). In addition to using alignments with protein sequence to determine ORFs, ab initio programs can be used. For example, ESTscan2.0 (Lottaz et al., 2003) can be used to identify coding regions in EST contigs. The efficacy of ESTscan2.0 and similar programs for translation initiation site determination has been assessed (Nadershahi et al., 2004). For domain analysis, there are also programs using nucleotide sequences directly to search for protein motifs. For example, InterProScan (http://www.ebi.ac.uk/interpro/) will read the input nucleotide sequences, translate them into protein sequences internally, and then search the domain databases (Zdobnov and Apweiler, 2001).

Tools for Comparative EST Analyses

A number of tools are available for comparative genomics; however, most of the tools are not used by a wide audience, either because of barriers, such as an absence of a user-friendly interface, or because of a lack of
understanding for how to distinguish the underlying algorithms. The most popular tools remain some of the most “simple,” such as BLAST. Here, we discuss a few tools for comparative EST analysis. First is the multispecies BLAST at PlantGDB. Although BLAST is available on many servers, the user is generally limited to searching against either very general groups or one species or data type at a time. The unique feature of BLAST@PlantGDB (http://www.plantgdb.org/cgi-bin/PlantGDBblast) is the ability for the user to search against any combination of species and EST or other sequence types. Searches can be conducted against user-designated single or multiple BLAST databases simultaneously (e.g., only rice or maize ESTs or both; or all monocot ESTs; or rice ESTs and all cereal EST contigs). The selection of specific databases for comparison can be critical to correctly assess statistical significance of observed sequence matches.

Another common task is to align ESTs to genomic DNA to determine exon–intron gene structure. BLAST is not the most suitable program for this task, because it was not designed to perform the spliced alignments necessary to overlay exons, as represented by the EST sequence, onto genomic DNA. Several spliced-alignment programs have been developed, such as sim4 (Florea et al., 1998), Spidey (Wheelan et al., 2001), BLAT (Kent, 2002), and GeneSeqer (Brendel et al., 2004; Usuka et al., 2000). GeneSeqer is perhaps the most versatile tool for annotating plant genomic DNA. It can produce plant gene structure models based on spliced alignment to genomic sequences of both native and homologous ESTs, cDNAs, and protein sequences (Usuka and Brendel, 2000). The GeneSeqer@PlantGDB server (Schlueter et al., 2003) allows users to run the GeneSeqer program over the web. As with BLAST@PlantGDB, users can select any combination of EST species collections from the PlantGDB database, or they can supply their EST sequences. Figure 2 displays a typical spliced alignment generated from GeneSeqer@PlantGDB.

In addition to using analysis programs such as BLAST and GeneSeqer, users may want to selectively retrieve information from databases for more detailed analyses and comparisons. Simple text search capabilities and fixed table reporting, typically implemented at biological databases, do not always meet the needs of biological researchers. The PlantGDB TableMaker tool (http://www.plantgdb.org/TableMaker.php) allows for criteria to be specified for different columns and for the selection of columns for inclusion in a tabular report format. Executing queries such as “Show me all the maize ESTs, library name, tissue type, and EST contigs they belong to” can be accomplished by filling out web forms.
Using EST Collections

Plant ESTs can be used for various investigations. The primary usage of ESTs is for gene discovery. A typical biologist is usually interested in questions such as whether the gene of interest exists and is expressed in certain organisms or tissues. Such questions are typically answered by simply searching EST databases using the gene of interest as a query in a BLAST run. Here we discuss the application of plant ESTs for more complex tasks: identification of gene families that are conserved among plant species.
species considered, discovery of splicing variants, and delineating gene structure via spliced alignments.

**Identification of Gene Families Conserved across Species**

Gene families conserved across plants can be used in phylogenetic studies and in the identification of conserved noncoding regions or to make comparative maps of major crop species. Although the amount of genetic data available for plants is increasing exponentially, most of the work is being done in just a few species. The identification of a set of gene families common to a large variety of plants (e.g., all angiosperms) would allow researchers studying less well-characterized plants to capitalize more fully on gains being made in *Arabidopsis*, rice, *Medicago*, and other model or reference plant species. Because ESTs give an idea of the genes expressed in an organism, and because EST data are abundant for a variety of species, ESTs are ideal starting material for the identification of genes conserved among species. There are two general approaches. The first is aligning ESTs or protein sequences from one species to the genomic DNA of another species. Fulton et al. (2002) developed a Conserved Ortholog Set (COS) of 1025 single-copy markers by matching tomato EST-derived Unigene sequences to the *Arabidopsis* genomic sequence with BLAST. A second common approach is to compare ESTs from one species to ESTs from another species, usually also based on BLAST-type sequence comparisons. For example, Lee et al. (2002) identified Tentative Ortholog Groups (TOGs) by selecting reciprocal BLASTN best-hit pairs.

**Identification of Alternative Splicing**

There are usually two approaches to detect alternative splicing by EST alignments. For species that have nearly completed genome sequences, such as *Arabidopsis* and rice, all EST and cDNA sequences can be aligned against the genome sequence using spliced alignment programs. Alternative splice cases can be inferred from the coordinates of predicted introns and exons. This method has been widely used in animal system including *Caenorhabditis elegans*, mouse, and human (Kan and Gish, 2002; Kent and Zahler, 2000; Thanaraj et al., 2003). For species lacking genome sequences, ESTs can be aligned against themselves. Any reliably predicted insertions in one sequence relative to another would indicate alternative splicing cases. This method is applicable only when the insertions are long enough to be distinguished from polymorphisms and sequencing errors. Different types of alternative splicing, however, cannot always be reliably distinguished from the EST–EST alignments. A study by Kan et al. (2004)
revealed a possible third approach: aligning ESTs of one species to the genome sequence of close relatives.

Because *Arabidopsis* and rice have nearly completed genome sequences and large EST and cDNA collections, we deployed the first approach to detect alternative splicing in plants (see also Zhu et al., 2003). A total of 4161 *Arabidopsis* genes and 5378 rice genes showed five types of alternative splicing: alternative donor sites, alternative acceptor sites, alternative position (both donor and acceptor sites are different), exon skipping, and intron retention. Updates and details about these cases are available at the Alternative Splicing In Plants (ASIP) web site, http://www.plantgdb.org/prj/SiP/ASIP/EnterDB.php.

An example from the ASIP database is shown in Fig. 3. At3g01150 (atPTB2a) is one of the three polypyrimidine tract binding proteins in *Arabidopsis* (Wang and Brendel, 2004). Its human homologue functions as a splicing regulator by binding to the polypyrimidine tract of introns (Lin

![Fig. 3. Visualization of expressed sequence tag (EST)/complementary DNA (cDNA) alignments and alternative splicing. Arabidopsis ESTs and cDNAs were aligned against the genome sequence using GeneSeqer. The black scale on top of the picture indicates the chromosome three coordinates of the alignments. Thick horizontal bars represent exons, and thin lines represent introns. Red indicates EST/cDNAs alignments. Blue indicates GenBank annotation. The green and red triangles above the blue bars indicate the translation start and stop positions, respectively. For alternative splicing, the green box on the red lines indicates an exon-skipping case. For the alternative donor site in the last intron sites, a black vertical bar indicates the most prevalent splice site, and a green vertical bar represents the alternative splice site.](image-url)
and Patton, 1995). As shown in the figure, a 102-bp exon was skipped in three full-length cDNAs (gi3395937, gi42466316, and gi42464393). The skipped exon is indicated by a green box. Inclusion of the exon in two full-length cDNAs (gi20260637, gi42463960) and ESTs (gi7613127, gi42532553) will introduce an in-frame premature stop codon. Thus, the exon skipping transcript is likely the wild-type splice form. In addition, the last intron of the gene can also be alternatively spliced. As shown in Fig. 3, six ESTs/cDNAs use a canonical donor site, which is marked by a black bar. Three EST/cDNAs, however, use an alternative donor site located 47 bp upstream of the canonical site, marked by a green bar. Both the constitutive donor site and the upstream donor site have conserved splice signals, as indicated by high scores in the GeneSequer alignment. Detailed alignment of representative ESTs/cDNAs is displayed in Fig. 4. Use of the upstream donor site introduces a frameshift and premature stop codon in the transcript.

**Genome Annotation**

The annotation accompanying the whole genome sequences of *Arabidopsis* and rice as stored at NCBI was mostly generated by TIGR (http://www.tigr.org/). Much of their structural annotation was based on *ab initio* gene prediction. Individual biologists could reannotate their regions of interest using spliced alignment programs to align ESTs to the sequence. A recommended starting point is to visit genome browsers provided by major databases. At PlantGDB, for example, for *Arabidopsis thaliana*, the most recent EST and full-length cDNA collections are periodically threaded onto the model organism’s five established chromosome sequences. The AtGDB genome browser (http://www.plantgdb.org/AtGDB/) provides a means for the visual assessment of the resulting gene structure annotation (Zhu *et al.*, 2003). Specifically, at any chosen region of the *Arabidopsis* genome, users can view the original annotation downloaded from NCBI (provided by TIGR), spliced alignments of all ESTs and full-length cDNAs, and exon–intron structural annotation based on the spliced alignment results. Currently 206,672 EST sequences, 58,679 full-length cDNA sequences, and 28,952 TIGR-predicted gene models are incorporated into AtGDB. Approximately 60% of the predicted gene models (protein coding regions) are fully supported by EST or cDNA evidence, whereas 25% are based solely on computational gene structure prediction. In addition, 70% of the predicted gene models are found in a genomic context so that corresponding EST and cDNA alignments display some level of incongruence, including incompletely annotated noncoding regions, alternative splicing, and erroneous gene predictions.
A similar browser has been created for rice (http://www.plantgdb.org/OsGDB), although the framework here is the complete rice bacterial artificial chromosome (BAC) sequences, rather than pseudo-chromosomes, the assembly of which is still in flux. Currently, 284,006 EST sequences, 32,128 full-length cDNA, 3572 gene models defined as GenBank file features, and 59,712 transcription unit (TU) models defined by TIGR (http://www.tigr.org/tdb/e2k1/osa1) are on display at OsGDB. Of the more current TU gene model predictions, 43% are fully confirmed by EST and cDNA evidence, whereas 48% are based on computational gene prediction alone. In addition, 61% of the TUs display some level of incongruence with local EST and cDNA alignments, as discussed for Arabidopsis.

At both AtGDB and OsGDB, users can contribute updated annotations of their own to a shared community annotation system through the
use of web-based annotation tools (Schlueter et al., 2005). These tools were
developed to allow a user to easily access ab initio predictions, native and
homologous sequence alignments, ORF estimations, and other useful ana-
lyses for gene structure determination. The results of these individual
analyses are incorporated into the annotation system so a user can simply
select the custom gene structure they desire. This “User-Contributed
the community.

Concluding Remarks

Plant ESTs are a rich resource. The examples discussed here might
indicate how this resource can still be used more effectively by a broader
community. Both for genome annotation in species with whole genome
sequencing efforts and for gene discovery in species without substantial
genome sequencing, EST sequencing seems to be a highly cost-effective
approach.

Acknowledgments

This work was supported in part by NSF Plant Genome Research Projects grants DBI-
0110189 and DBI-0321600. Some of the work was initiated while L. K. was a participant of the
NIH-NSF 2003 Summer Institute in Bioinformatics and Computational Biology at ISU
(supported by NSF grant EEC-0234102).

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H.,
Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and
human genome project. Science 252, 159–164.
Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman,
alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20,
1157–1169.
Bioinformatics 17, 1093–1104.


Abstract

Few plant species are dioecious and only a small fraction of these species are known to have sex chromosomes. Considerable efforts to isolate sex-linked genes from dioecious Silene latifolia (Caryophyllaceae) have resulted in the isolation of surprisingly few sex-linked genes, suggesting that the methods used previously were not efficient in plants. This chapter analyzes the methods that have been and can be used for isolation of genes from plant sex chromosomes. The most successful method used for the isolation of Y-linked genes included the screening of a male complementary DNA (cDNA) library with the probe obtained by degenerate oligonucleotide-primed polymerase chain reaction (PCR) of the microdissected Y chromosomes. However, chromosome microdissection requires sophisticated equipment and is difficult to apply to species with cytologically indistinguishable sex chromosomes. Genome and cDNA library subtraction methods were surprisingly unsuccessful, probably because of low divergence between the homologous X- and Y-linked genes in plants. Segregation testing and genomics-based methods are increasingly popular and are the most promising approaches for isolation of multiple genes from plant sex chromosomes.

Introduction

Sex chromosomes are probably quite a rare phenomenon in plants. Only about 50% of plant species are dioecious (having separate male and female individuals), and only a fraction of these dioecious species possess morphologically distinguishable sex chromosomes (Westergaard, 1958;