The Construction of Arabidopsis Expressed Sequence Tag Assemblies

A New Resource to Facilitate Gene Identification

Steven D. Rounsley*, Anna Glodek, Granger Sutton, Mark D. Adams, Chris R. Somerville[1], J. Craig Venter, and Anthony R. Kerlavage

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850 (S.D.R., A.G., G.S., M.D.A., J.C.V., A.R.K.); and Carnegie Institution of Washington, Department of Plant Biology, 290 Panama Street, Stanford, California 94305–4101 (C.R.S.)

The generation of large numbers of partial cDNA sequences, or expressed sequence tags (ESTs), has provided a method with which to sample a large number of genes from an organism. More than 25,000 *Arabidopsis thaliana* ESTs have been deposited in public databases, producing the largest collection of ESTs for any plant species. We describe here the application of a method of reducing redundancy and increasing information content in this collection by grouping overlapping ESTs representing the same gene into a "contig" or assembly. The increased information content of these assemblies allows more putative identifications to be assigned based on the results of similarity searches with nucleotide and protein databases. The results of this analysis indicate that sequence information is available for approximately 12,600 nonoverlapping ESTs from Arabidopsis. Comparison of the assemblies with 953 Arabidopsis coding sequences indicates that up to 57% of all Arabidopsis genes are represented by an EST. Clustering analysis of these sequences suggests that between 300 and 700 gene families are represented by between 700 and 2000 sequences in the EST database. A database of the assembled sequences, their putative identifications, and cellular roles is available through the World Wide Web.

It is frequently possible to infer the probable function of an otherwise anonymous gene solely on the basis of partial nucleotide or deduced amino acid sequence homology to genes or gene products of known function (Adams et al., 1991, 1992). In many cases, extended regions of sequence homology are found in functionally related gene products from phylogenetically distant organisms such as bacteria and humans. Thus, it is frequently possible to assign function to a large proportion of randomly chosen anonymous cDNA clones by obtaining partial nucleotide sequence information (Adams et al., 1991, 1992). These partial cDNA sequences are referred to as ESTs. At the time of this writing, more than 25,000 individual *Arabidopsis thaliana* EST sequences have been deposited in dbEST, a public database for EST sequences maintained at the National Center for Biotechnology Information (Boguski et al., 1993). Approximately 6100 of these sequences have been produced by a consortium of French laboratories (Höfte et al., 1993), and the remaining 19,500 have been produced by the Arabidopsis cDNA-sequencing project at Michigan State University (Newman et al., 1994). Combined with clone availability, EST databases can offer a researcher a convenient path to the full sequence of many Arabidopsis genes, given an appropriate query. Existing methods of accessing the Arabidopsis EST sequences via the Internet have been summarized elsewhere (Newman et al., 1994; Swope et al., 1995; Rounsley et al., 1996).

Despite the power of EST data, there are inherent limitations to the approach. The single-pass sequencing technique used to produce ESTs results in sequence ambiguities that confound sequence comparisons. Also, the length of each sequence is limited to what can be read from a single sequencing run, usually about 400 bp. These two factors can affect the usefulness of EST data for gene identification. In addition, the fact that EST sequence data are usually obtained from randomly chosen clones from cDNA libraries results in highly expressed genes being sequenced multiple times. This redundancy can be reduced by using normalized libraries in which the frequency of highly expressed genes is reduced by subtractive hybridization or a related approach (Sankhavaram et al., 1991; Kohchi et al., 1995). However, in practice, redundancy is never removed completely. The redundancy can hinder effective searching of the databases by producing multiple hits for what is, in fact, the same gene represented by many ESTs.

Here we describe a database in which the redundancy in the Arabidopsis EST entries present in dbEST has been reduced by grouping together sequences derived from the same gene to form a "contig" or assembly of EST sequences. These assembled EST sequences are used to form a single TC sequence, which represents the sum of the sequence information contained in all of the individual ESTs. The TC sequence is usually longer than any of the individual ESTs

[1] C.R.S. was supported in part by a grant from the National Science Foundation (grant no. BIR9313751).

* Corresponding author; e-mail rounsley@tigr.org; fax 1–301–838–0208.

Abbreviations: EGAD, expressed gene anatomy database; EST, expressed sequence tag; TC, tentative consensus; TIGR, The Institute for Genomic Research; WWW, World Wide Web.

and thus increases the information content and consequently the chances of successful identification of the transcript. The presence of several overlapping ESTs also allows resolution of sequence ambiguities that are present in a small subset of the contributing sequences. A similar approach has been previously reported for the analysis of more than 170,000 human ESTs (Adams et al., 1995).

MATERIALS AND METHODS

Computer Programs

Computer programs mentioned here often have been developed to interact specifically with database schema internal to TIGR. They have been previously described in more detail elsewhere (Kerlavage et al., 1993; Adams et al., 1995; Kerlavage et al., 1995). These tools will be made available to academic researchers on request. Inquiries should be made by e-mail to tools@tdb.tigr.org.

EST Sequence Preparation

To begin the assembly process, *Arabidopsis thaliana* EST sequences present in dbEST were downloaded into a Sybase relational database called the Non-Human cDNA Database, along with source and library information. Each EST sequence was then run through a quality control step that used GRASTA, a modified FASTA (Pearson and Lipman, 1988) search program that searches both strands of the sequence, to search for the presence of common vector sequences, and poly(A), (T), or (CT) tracts. Where these stretches of sequence were at either end, they were trimmed away. In other cases, such as where long stretches of vector were present in the middle of a sequence, the EST was flagged and removed from the data set. Next, checks were made for ambiguous bases (Ns) that were common in EST sequences. A program called NCOUNTER was used to trim away the 3′ end of the sequence entries until the level of Ns was below 4% of the total nucleotides. Finally, any sequences that were shorter than 100 bp after these trimming procedures were flagged and removed from the data set. The "cleaned" EST sequences were used in the assembly process.

Nonredundant Arabidopsis Transcripts

To identify and group together ESTs that correspond to previously cloned Arabidopsis genes, a set of nonredundant transcript sequences from Arabidopsis was constructed. All Arabidopsis coding sequences present in the plant division of the Genome Sequence Database (Keen et al., 1996) were extracted and searched against each other using the Blast algorithm (Altschul et al., 1990). This identified multiple accessions encoding the same gene. In these cases the entry containing the longest transcript was stored in EGAD at TIGR, and the others were stored simply as related accession numbers. These sequences are available through a WWW interface to EGAD at universal resource locator (URL, http://www.tigr.org/tdb/egad/egad.html) and are also available as a multiple FASTA file via file transfer protocol (FTP, ftp://ftp.tigr.org/pub/data/a_thaliana/at.egad). Periodically, the nonredundant set of transcripts is updated with new sequences using a similar process to screen for nonredundant sequences. The set of nonredundant transcripts from Arabidopsis stored in EGAD were added to the EST data set for the assembly step.

Initial Assembly

All of the cleaned EST sequences and the nonredundant transcripts were combined into a multiple FASTA file and used as input to TIGR Assembler, a program designed to assemble large sets of sequence data (Sutton et al., 1995). This program has been successfully used to assemble three complete bacterial genomes and more than 180,000 EST sequences from human tissues (Adams et al., 1995; Fleischmann et al., 1995; Fraser et al., 1995; Bult et al., 1996). The algorithm constructs a table of 10-mer content for each EST and finds a candidate overlapping sequence by comparison of their 10-mer content. Starting with a seed sequence, the program searches for the best candidate for an overlapping sequence from the data set and then attempts to align it using a modified Smith-Waterman algorithm. The candidate is added to the assembly only if the overlap is at least 95% identical over a minimum of 40 bp, with a maximum of 25 bp of unmatched sequence at either end. The multiple-aligned sequences for an assembly are used to produce the TC sequence for that group of ESTs. The consensus sequence selection is governed by a set of simple rules (Sutton et al., 1995). Uppercase bases are used where that base occurs in greater than two-thirds of the aligned sequences; otherwise lowercase bases, lowercase n, or two-base ambiguity codes are used, depending on the relative frequencies of bases at that position.

Searching and Assigning Putative Identifications

Assemblies that exhibited greater than 98% sequence identity to one of the known Arabidopsis transcripts from EGAD were assigned the name of the cloned gene. All other assemblies were searched against all nucleotide sequences in GenBank and against a nonredundant protein database constructed from GenPept, Protein Identification Resource (PIR), and SwissProt. Nucleotide searches used the Blast-to-Graze procedure, which combines the speed of the Blast algorithm (Altschul et al., 1990), to identify potential matches and the accurate alignments of Graze (Kerlavage et al., 1995), a modified Smith-Waterman algorithm (Smith and Waterman, 1981). Protein searches used Blaze (Intelligenetics, Mountain View, CA [Brutlag et al., 1993]) and a script, MBLZT, to combine search results from each frame into a single output, incorporating frame shifts that would otherwise destroy the alignment.

The results of these searches were viewed and assessed using BYOB (Kerlavage et al., 1995). This program presents the results of both the nucleotide and protein searches in a graphical manner, allows the user to assess the significance of all matches based on the length of the matches and the levels of similarity, and then, where appropriate, records a suitable putative identification in the Non-Human cDNA

Database. Each assembly with a putative identification was assigned to one of four classes. Matches against known Arabidopsis genes were either exact (class 1) or nonexact (class 2). Matches against non-Arabidopsis genes were assigned to class 3 and also assigned to one of two subdivisions, plant and non-plant genes. Class 4 was used for contamination of various kinds such as *Escherichia coli* DNA sequences, rRNA, and other forms of nonprotein coding sequences such as tRNA genes.

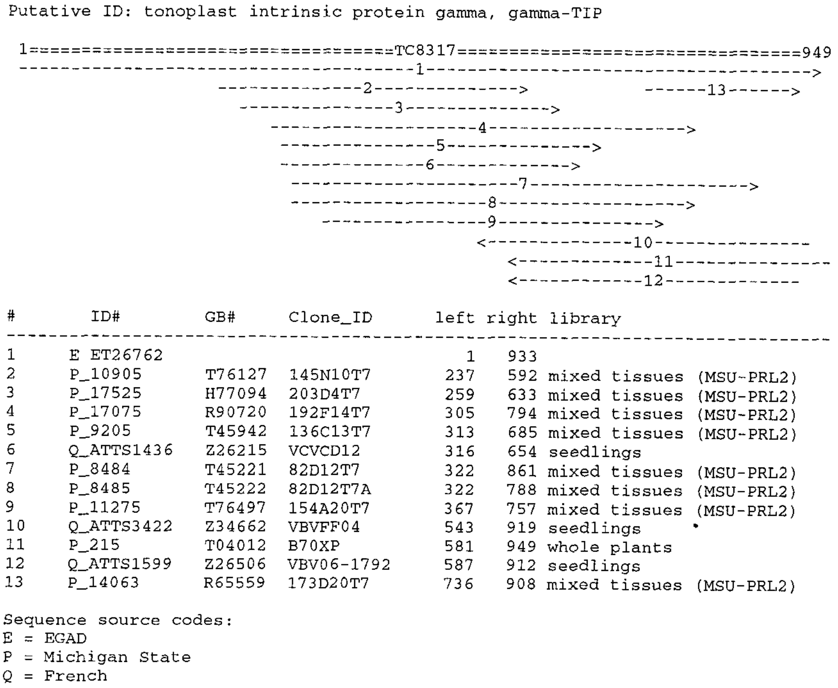
ESTs that did not assemble with others, known as singletons, were not manually assigned identifications since the results of database searches were already available in the dbEST database. A script was written to extract the top-scoring protein match from these Blast results and record that in the Non-Human cDNA Database as the putative identification for that EST if the Blastx score was greater than 90.

The sequences of the assembled and singleton ESTs, along with the putative name identifications, are available in the TIGR AT database. This database is available through a WWW interface at the URL (http://www.tigr.org/tdb/at/at.html). This interface to the database allows searching at several levels: keyword searches of the putative identifications, nucleotide and amino acid level sequence searching, and simple report retrieval using unique identifiers for each sequence, such as the GenBank accession number (Rounsley et al., 1996). An example of how an assembly is presented in the database is shown in Figure 1.

The putative identifications for each assembly and each singleton were grouped according to general role categories of the function of the closest protein match. The main categories used were cell division, cell signaling/communication, gene/protein expression, metabolism, plant-specific functions, and structural proteins, with subdivisions within each category. There is an additional category of unassigned function, which includes those ESTs that match hypothetical proteins, as well as matches that could not easily be assigned to one of the above categories. Two versions of the list of putative identifications, with and without role assignments, are available for browsing and downloading from the TIGR AT database. Figure 2 illustrates the format of this list of putative identifications.

Updating of Assemblies

As more Arabidopsis sequences, both ESTs and non-ESTs, are deposited in the databases, the assemblies need to be updated. Periodic updates are achieved without reassembling the whole data set by the use of a clustering procedure. New Arabidopsis transcripts from EGAD and new EST entries from the Non-Human cDNA Database go through initial quality control checks as described above. All of the new sequences are then compared with each other and with the current data set, which consists of existing assemblies, singletons, and known Arabidopsis transcripts in EGAD. Any new sequence that matches and is completely contained within an existing assembly is linked to that assembly in the database, without reassembly. All other matches are then used to define clusters of sequences that are sufficiently closely related to warrant reassembly. Assemblies in these clusters are separated into their constituent ESTs for the reassembly process. The new assemblies produced with this process are assigned new TC names, since they now have a different consensus sequence. The previous TC name is retained and can be used to retrieve the most current assembly. This suite of proce-



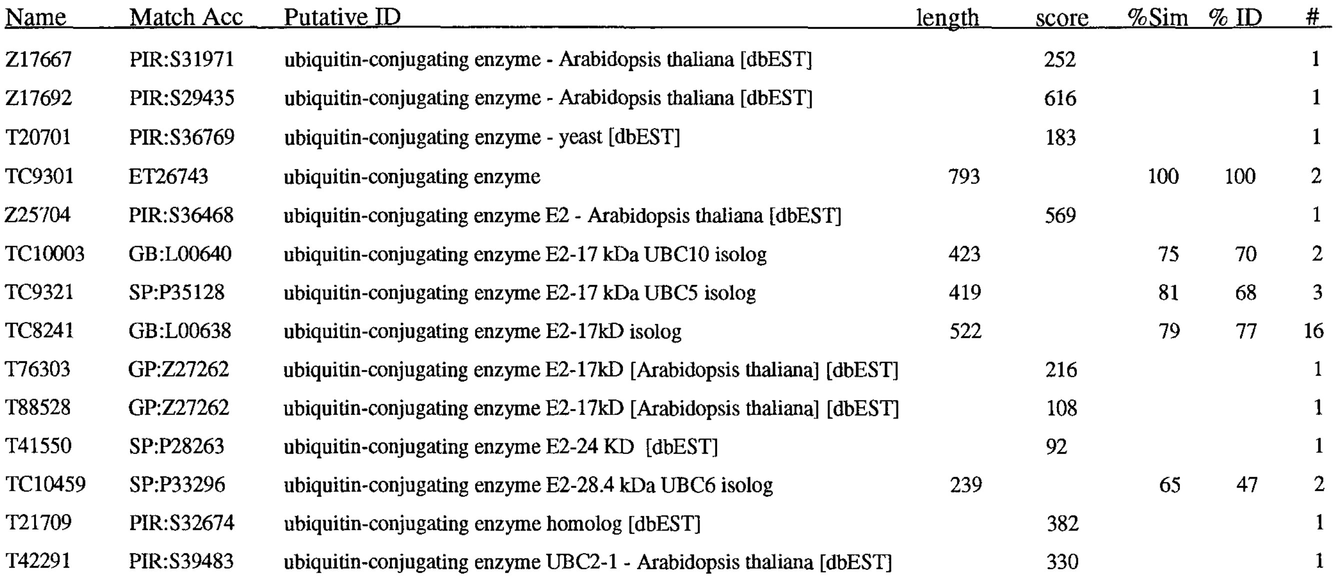
```
Putative ID: tonoplast intrinsic protein gamma, gamma-TIP

 1===================================TC8317=================================949
 --------------------------------------1-------------------------------------->
                   --------------2-------------->          ------13------>
                     ---------------3-------------->
                        --------------------4-------------------->
                         ---------------5--------------->
                         --------------6------------->
                          -----------------------7---------------------->
                          --------------------8------------------->
                             -----------------9---------------->
                                            <--------------10---------------
                                               <-------------11---------------
                                               <------------12-------------

#      ID#          GB#      Clone_ID       left right library
--------------------------------------------------------------------------------
1      E ET26762                               1   933
2      P_10905      T76127   145N10T7        237   592 mixed tissues (MSU-PRL2)
3      P_17525      H77094   203D4T7         259   633 mixed tissues (MSU-PRL2)
4      P_17075      R90720   192F14T7        305   794 mixed tissues (MSU-PRL2)
5      P_9205       T45942   136C13T7        313   685 mixed tissues (MSU-PRL2)
6      Q_ATTS1436   Z26215   VCVCD12         316   654 seedlings
7      P_8484       T45221   82D12T7         322   861 mixed tissues (MSU-PRL2)
8      P_8485       T45222   82D12T7A        322   788 mixed tissues (MSU-PRL2)
9      P_11275      T76497   154A20T7        367   757 mixed tissues (MSU-PRL2)
10     Q_ATTS3422   Z34662   VBVFF04         543   919 seedlings
11     P_215        T04012   B70XP           581   949 whole plants
12     Q_ATTS1599   Z26506   VBV06-1792      587   912 seedlings
13     P_14063      R65559   173D20T7        736   908 mixed tissues (MSU-PRL2)

Sequence source codes:
E = EGAD
P = Michigan State
Q = French
```

Figure 1. An example EST assembly for a γ-TIP gene containing 12 EST sequences and a known transcript contained in EGAD. The schematic diagram indicates the relative arrangement of the individual sequences within the assembly. The individual sequences are listed below the diagram and, when viewed from within the TIGR AT database, are linked to other community databases. GB#, GenBank accession no. MSU, Michigan State University.



Name	Match Acc	Putative ID	length	score	%Sim	% ID	#
Z17667	PIR:S31971	ubiquitin-conjugating enzyme - Arabidopsis thaliana [dbEST]		252			1
Z17692	PIR:S29435	ubiquitin-conjugating enzyme - Arabidopsis thaliana [dbEST]		616			1
T20701	PIR:S36769	ubiquitin-conjugating enzyme - yeast [dbEST]		183			1
TC9301	ET26743	ubiquitin-conjugating enzyme	793		100	100	2
Z25704	PIR:S36468	ubiquitin-conjugating enzyme E2 - Arabidopsis thaliana [dbEST]		569			1
TC10003	GB:L00640	ubiquitin-conjugating enzyme E2-17 kDa UBC10 isolog	423		75	70	2
TC9321	SP:P35128	ubiquitin-conjugating enzyme E2-17 kDa UBC5 isolog	419		81	68	3
TC8241	GB:L00638	ubiquitin-conjugating enzyme E2-17kD isolog	522		79	77	16
T76303	GP:Z27262	ubiquitin-conjugating enzyme E2-17kD [Arabidopsis thaliana] [dbEST]		216			1
T88528	GP:Z27262	ubiquitin-conjugating enzyme E2-17kD [Arabidopsis thaliana] [dbEST]		108			1
T41550	SP:P28263	ubiquitin-conjugating enzyme E2-24 KD [dbEST]		92			1
TC10459	SP:P33296	ubiquitin-conjugating enzyme E2-28.4 kDa UBC6 isolog	239		65	47	2
T21709	PIR:S32674	ubiquitin-conjugating enzyme homolog [dbEST]		382			1
T42291	PIR:S39483	ubiquitin-conjugating enzyme UBC2-1 - Arabidopsis thaliana [dbEST]		330			1

Figure 2. For each TC sequence that has an identification, the values listed are the TC number, the database, and accession (Acc) number of the most similar entry from non-EST public databases, the putative identification, the length of the sequence match, the level of similarity (Sim) and identity (ID), and the number of sequences in this assembly. For a TC sequence that contains an EGAD sequence, the length indicates the length of the known transcript, and the identity is shown as being 100%. For singletons, the Blastx score is listed.

dures enables easy update of the assembly database, without unnecessary reassembly of the complete data set.

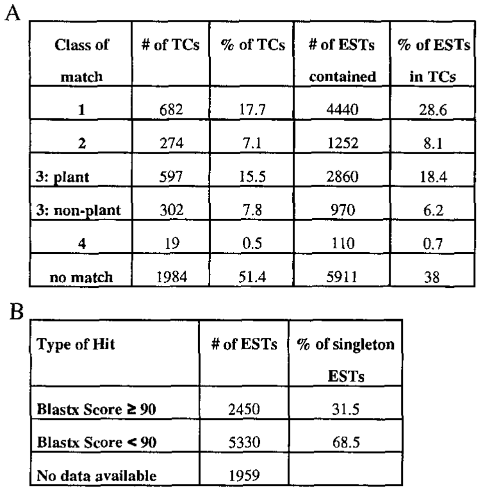
Putative identifications also need to be updated periodically. This is of use when the previous identification was associated with a more distantly related species, but, in the period since that identification was made, new Arabidopsis genes have been reported and characterized, allowing more accurate identification. The other obvious need for regular updating is when sequences have no significant database matches. As new genes are reported, identification of these previously unknown sequences may be possible. Therefore, we have an automated procedure that searches the unknown sequences against weekly updates of GenBank and reports possible matches for manual inspection and updating of the TIGR AT database.

RESULTS AND DISCUSSION

Release 1.1 of the TIGR AT database contains the results of assembling 25,662 Arabidopsis EST sequences contained in dbEST prior to April 1996. These sequences were reduced to 25,262 after the quality control process was performed. More than 1,500 database entries for Arabidopsis sequences were analyzed and used to produce a set of 953 nonredundant transcripts. These were combined with the cleaned ESTs and assembled using the TIGR Assembler software (Sutton et al., 1995).

The results show that 15,543 ESTs (61.5%) were grouped into only 3,858 assemblies, which is approximately a 4-fold reduction of redundancy. The average number of ESTs per assembly was 4.0, with the largest containing 221 EST sequences. Approximately one-half of the 3,858 assemblies, containing 62% of the ESTs, were given a putative identification by comparison with sequences in the public nucleotide and protein databases. These data are summarized in Figure 3. The depth of assemblies in the different identification classes are not uniform. Assemblies with a class 1 match have an average of 6.5 ESTs each, whereas assemblies with no significant match have only 3.0. Class 2 and 3 assemblies have an average of 4.6 and 4.3 ESTs each. The large number of ESTs in class 1 assemblies is due mainly to the inclusion of full-length coding sequences in many of these assemblies, which bring together ESTs that would otherwise not overlap. It is also skewed by the few large assemblies of ESTs from highly expressed genes such as the Rubisco subunits and proteins of the light-harvesting complexes, many of which have more than 100 ESTs present in dbEST. It is also possible that those Arabidopsis genes that have already been cloned and sequenced tend to be more highly expressed than those that have yet to be found and, therefore, are more likely to be represented by an EST. In addition, the correlation between small assemblies and lack of putative identification could be due to the tendency for these assemblies to have shorter consensus sequences and therefore be less likely to find a match in the databases, particularly if the ESTs in question come from the 5′ or 3′ untranslated regions of a gene. A similar explanation could explain the fact that a singleton EST, which is therefore shorter than an assembly, is much less likely to have an identification (31.5%) than one that is part of an assembly (62%), although direct comparison may not be possible, since the method of assigning these identifications is not the same for the two groups.

Not all assemblies without a putative identification contain a small number of ESTs. There are many that contain more than 15 ESTs that have no significant match to known proteins. These genes may be particularly interesting be-



A

Class of match	# of TCs	% of TCs	# of ESTs contained	% of ESTs in TCs
1	682	17.7	4440	28.6
2	274	7.1	1252	8.1
3: plant	597	15.5	2860	18.4
3: non-plant	302	7.8	970	6.2
4	19	0.5	110	0.7
no match	1984	51.4	5911	38

B

Type of Hit	# of ESTs	% of singleton ESTs
Blastx Score ≥ 90	2450	31.5
Blastx Score < 90	5330	68.5
No data available	1959	

Figure 3. A, Putative identifications for TCs have been assigned to one of four classes based on the source of the sequence to which the most significant match was found. Class 1 TCs represent previously cloned Arabidopsis genes, and class 2 TCs are close homologs of such genes, whereas class 3 TCs match genes from other species. Class 4 TCs are noncoding such as rRNA genes or bacterial sequence contamination. The number of TCs and the number of ESTs contained within them are given as numbers or as percentages of the total number. B, Singleton ESTs are assigned an identification from homology data stored in dbEST if the top-scoring protein match scores 90 or greater in a Blastx search, otherwise no assignment is made. For some of the more recently produced ESTs, no homology information is available in dbEST.

cause it is apparent that they are expressed at reasonably high levels in the cDNA libraries used to produce the ESTs. It would be interesting to know what role these genes play in Arabidopsis and whether they encode a plant-specific or even an Arabidopsis-specific function. A complete list of the names and sizes of assemblies that do not have significant database matches is available from the TIGR AT database.

One of the goals of EST projects is to identify as many genes as possible, so it is important to estimate how many genes are represented in a given data set. The approach we have taken partially answers that question by reducing the redundancy in the data, but it is still not possible to get absolute numbers. The assembly process requires overlap of sufficient length and similarity. In cases in which ESTs from the same gene have small or nonexistent overlaps, or marginal sequence quality in an overlap region, not all will assemble. In contrast, there are a few situations in which ESTs from very similar but distinct genes will assemble together, because of regions of near identity in the overlap regions, but these cases are the exception rather than the rule. The parameters of the assembly algorithm require any nonidentical overhang regions at the ends of aligned sequences to be very short, and this will cause some otherwise very similar sequences to not assemble. This parameter is necessary to prevent different sequences that share a common domain from being assembled, but as a result, some redundancy may remain.

To address the question of how much redundancy remains in our database, we searched all assemblies and singletons against each other at the nucleotide level using the Blastn algorithm. A pair of sequences that shared 95% sequence identity over a given window size were, for the purpose of this analysis, considered to represent the same gene. These pairs were used to produce clusters of sequences so that, if two pairs, AB and BC, were given, a cluster ABC would be produced. The results indicate that when window sizes from 50 to 200 bp are used between 10 and 13% of the sequences can be grouped into such clusters, or conversely 87 to 90% of the sequences are unique (data not shown). However, there are differences between sequences in a cluster that cause them not to assemble together, and these differences could be real or could be due to sequencing errors. Without sequencing these clones further or analyzing the original electropherograms, it is difficult to make judgments in these cases. If these sequence differences are artifactual, the number of nonoverlapping sequences, and therefore the upper limit for a gene number estimate, is between 12,400 and 12,800.

Another factor affecting estimates of gene number is that some cDNA clones, in particular from the French laboratories, are sequenced from both ends (Cooke et al., 1996). These are present as different sequences in the database, but the clone information allows them to be linked together. In release 1.1 of the database, there are approximately 1800 such clones with sequences that may be singletons or parts of assemblies. Where both sequences are part of the same assembly, gene number estimates would not be affected. In other cases, the linking information allows us to reduce any gene number estimate by about 1,300 sequences, altering the above range for an upper limit to between 11,100 and 11,500. Methods to display such linkage in the database are being developed.

Given this range of numbers as an upper limit for how many genes are represented by ESTs, it is interesting to examine what proportion of Arabidopsis genes that have previously been cloned are represented by an EST. Of the 953 Arabidopsis genes included in the assembly process, 544 assembled with ESTs, thus approximately 57% of previously cloned Arabidopsis genes, are represented by an EST. If the Arabidopsis coding sequences so far reported are a suitably random subset of all Arabidopsis coding genes, then it may be appropriate to extrapolate and suggest that the 57% of all genes are represented by an EST. Given the current estimates of 20,000 genes in the Arabidopsis genome (Gibson and Somerville, 1993; Meyerowitz, 1994), we would expect to have ESTs from 11,400 genes. This is not too far from the estimates given above. There are, however, some reasons to doubt the randomness of the genes so far cloned, such as the ease of cloning when message abundance is high, but this may be counteracted by the fact that many Arabidopsis genes have been cloned via genetic screens.

It is possible to make estimates about the numbers and sizes of gene families represented in the EST data using the same search results that were used earlier to assess the remaining redundancy. Figure 4 shows the results of such analyses, using different levels of nucleotide identity as cutoffs, along with two different window sizes. Clustering

A

Window Size	%ID	# Sequences	# Clusters
200bp	65	1988	727
	75	1322	497
	85	727	310
100bp	65	3372	1110
	75	1964	715
	85	1097	494

B

Cluster Size	Freq.
2	346
3	96
4	19
5	13
6	6
7	5
8	3
9	2
10	3
>10	4

Figure 4. A, Pairs of sequences that are less than 95% identical but greater than the specified percentage identity over the given window size were clustered. The number of sequences in each cluster and the number of clusters are given. B, The distribution of the cluster sizes for sequences sharing between 75 and 95% sequence identity are shown.

was performed in the same manner, but all but one representative from the same gene clusters identified above were removed. This method is a very straightforward and a somewhat naive way of identifying gene families, because it relies only on nucleotide comparisons and uses the Blastn search algorithm, which is not very sensitive, and does not insert gaps to extend a match. However, it is a useful first-glance analysis that suggests that there are relatively few highly similar isoforms of genes in Arabidopsis. This is in contrast to reports that suggest that almost half of all genes in tomato are members of multigene families (Bernatzky and Tanksley, 1986). Further detailed analyses using Grail-predicted reading frames and more sensitive amino acid alignments will be more revealing.

Although space limitations prevent the printing of all of the putative identifications, this list is available on our web server. Having this available in a browsable form is a useful alternative to the searchable database, which requires the formulation of a specific query. It is hoped that viewing such a list will be thought-provoking in terms of what types of genes have been found and what types have not been found via an EST approach. It is important to bear in mind that the assignments to role categories are designed more for ease of presentation and browsing than for defining the biological function. The assignment does not indicate the role of the Arabidopsis EST sequence but of the protein to which the EST has greatest similarity. In addition, the categories are such that many proteins could be assigned multiple roles, and in these cases the assignments are somewhat arbitrary. Also, it is not intended to be a static document. We encourage feedback and hope that this list will continue to be improved as suggestions are received from colleagues.

Ultimately, a list of gene names to which Arabidopsis ESTs have been found to match may be interesting to browse but can be little more than a small stepping stone toward exploring the wealth of information contained within the Arabidopsis genome. However, the EST databases will be an important tool in identifying coding regions as we move toward the upcoming phase of large-scale genomic sequencing of Arabidopsis.

ACKNOWLEDGMENTS

We would like to acknowledge the work of Tom Newman of Michigan State University (East Lansing) and Ernie Retzel of the University of Minnesota (Minneapolis), who together with the consortium of French laboratories, have produced the Arabidopsis EST sequences that made this work possible. We also thank Lisa Fitzgerald and Robert Shirley for their assistance with database and WWW aspects of this project.

Received May 20, 1996; accepted August 2, 1996.
Copyright Clearance Center: 0032–0889/96/112/1177/07.

LITERATURE CITED

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

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 the human genome project. Science **252:** 1651–1656

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JA, Brandon RC, Chiu M-W, Clayton RA, Cline RT, Cotton MD, Earle-Hughes J, Fine LD, Fitzgerald LM, FitzHugh WM, Fritchman JL, Geoghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle Jr PS, Kelley JM, Klimek KM, Kelley JC, Liu L-I, Marmaros SM, Merrick JM, Moreno-Palanques RF, McDonald LA, Nguyen DT, Pellegrino SM, Phillips CA, Ryder SE, Scott JL, Saudek DM, Shirley R, Small KV, Spriggs TA, Utterback TR, Weidman JF, Li Y, Barthlow R, Bednarik DP, Cao L, Cepeda MA, Coleman TA, Collins E-J, Dimke D, Feng P, Ferrie A, Fischer C, Hastings GA, He W-W, Hu J-S, Huddleston KA, Greene JM, Gruber J, Hudson P, Kim A, Kozak DL, Kunsch C, Ji H, Li H, Meissner PS, Olsen H, Raymond L, Wei Y-F, Wing J, Xu C, Y G-L, Ruben SM, Dillon PJ, Fannon MR, Rosen CA, Haseltine WA, Fields C, Fraser CM, Venter JC (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature Suppl **377:** 3–174

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

Bernatzky R, Tanksley SD (1986) Majority of random cDNA clones correspond to single loci in the tomato genome. Mol Gen Genet **203:** 8–14

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

Brutlag DL, Dautricourt JP, Diaz R, Fier J, Moxon B, Stamm R (1993) Blaze (Tm)—an implementation of the Smith-Waterman sequence comparison algorithm on a massively parallel computer. Comput Chem **17:** 203–207

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb J-F, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann J, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow P, Hanna M, Cotton M, Roberts K, Hurst M, Kaine BP, Klenk H-P, Fraser CM, Smith HO, Woese CR, Venter JC (1996) Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii*. Science **273:** 1058–1073

Cooke R, Raynal M, Laudié M, Grellet F, Delseny M, Morris P-C, Guerrier D, Giraudat J, Quigley F, Clabault G, Li Y-F, Mache R, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clément B, Philipps G, Hervé C, Bardet C, Tremousaygue D, Lescure B, Lacomme C, Roby D, Jourjon M-F, Chabrier P, Charpenteau J-L, Desprez T, Amselem J, Chiapello H, Höfte H (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. Plant J **9:** 101–124

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton G, Fitzhugh W, Fields C, Jocayne JD, Scott J, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton. MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:** 496–512

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, Hutchison CA III, Venter JC (1995) The Mycoplasma genitalium genome sequence reveals a minimal gene complement. Science **270:** 397–403

Gibson S, Somerville CR (1993) Isolating plant genes. Trends Biotechnol **11:** 306–313

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerrier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D, Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana.* Plant J **4:** 1051–1061

Keen G, Burton J, Crowley D, Dickinson E, Espinosa-Lujan A, Franks E, Harger C, Manning M, March S, McLeod M, O'Neill J, Power A, Pumilia M, Reinert R, Rider D, Rohrlich J, Schwertfeger J, Smyth L, Thayer N, Troup C, Fields C (1996) The Genome Sequence DataBase (GSDB): meeting the challenge of genomic sequencing. Nucleic Acids Res **24:** 13–16

Kerlavage AR, Adams MD, Kelley JC, Dubnick M, Powell J, Shanmugam P, Venter JC, Fields C (1993) Analysis and management of data from high-throughput expressed sequence tag projects. *In* TN Mudge, V Milutinov, L Hunter, eds, Proceedings of the 26th Hawaii International Symposium on System Sciences, Hawaii. Institute of Electrical and Electronics Engineers Computer Society Press, Los Alamitos, CA, pp 585–594

Kerlavage AR, FitzHugh W, Glodek A, Kelley JC, Scott J, Shirley R, Sutton G, Wai-Chiu M, White O, Adams MD (1995) Data management and analysis for high-throughput DNA sequencing projects. IEEE Eng Med Biol **14:** 710–717

Kohchi T, Fujishige K, Ohyama K (1995) Construction of an equalized cDNA library from *Arabidopsis thaliana.* Plant J **8:** 771–776

Meyerowitz E (1994) Structure and organization of the *Arabidopsis thaliana* nuclear genome. *In* E Meyerowitz, CR Somerville, eds, Arabidopsis. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp 21–36

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E, Somerville CR (1994) Genes galore. A summary of the methods for accessing the results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol **106:** 1241–1255

Pearson W, Lipman D (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA **85:** 2444–2448

Rounsley SD, Glodek A, Sutton G, Shirley R, Adams MD, Venter JC, Kerlavage AR (1996) *Arabidopsis* EST analysis at TIGR. Weeds World **3**(i): 26–33

Sankhavaram RP, Parimoo S, Weissman SM (1991) Construction of a uniform abundance (normalized) cDNA library. Proc Natl Acad Sci USA **89:** 1943–1947

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

Sutton GS, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: a new tool for assemblying large shotgun sequencing projects. Genome Sci Technol **1:** 9–19

Swope KL, Newman TC, Shoop E, Bieganski P, Chi E, Holt O, Carlis J, Riedl J, Retzel EF (1995) Everything you wanted to know about the University of Minnesota's analysis of Arabidopsis ESTs but were afraid to ask. Weeds World **2**(ii): 21–26