Published in final edited form as:

Curr Protoc Bioinformatics. 2010 March ; 0 1: Unit1.6.1–Unit1.636. doi:10.1002/0471250953.bi0106s29.

Using the DFCI Gene Index Databases for Biological Discovery

Corina Antonescu¹, Valentin Antonescu¹, Razvan Sultana¹, and John Quackenbush¹

¹Dana-Farber Cancer Institute, Boston, Massachusetts

Abstract

The DFCI Gene Index Web pages provide access to analyses of ESTs and gene sequences for nearly 114 species, as well as a number of resources derived from these. Each species-specific database is presented using a common format with a home page. A variety of methods exist that allow users to search each species-specific database. Methods implemented currently include nucleotide or protein sequence queries using WU-BLAST, text-based searches using various sequence identifiers, searches by gene, tissue and library name, and searches using functional classes through Gene Ontology assignments. This protocol provides guidance for using the Gene Index Databases to extract information.

Keywords

gene index database; gene index; databases; DFCI

INTRODUCTION

The DFCI Gene Index Web pages (<http://compbio.dfci.harvard.edu/tgi/tgipage.html>; Fig. 1.6.1) provide access to analyses of Expressed Sequence Tags (ESTs) and gene sequences for over 114 species, as well as a number of resources derived from these. A summary of the species currently represented can be found in the Appendix at the end of this unit; additional species are regularly added to the collection based on the availability of EST data and user requests. Each species-specific database is presented using a common format with a home page similar to that shown in Figure 1.6.2. A variety of methods exist, listed immediately below the heading “Search the Index by,” that allow users to search each species-specific database. Methods implemented currently include searching of nucleotide or protein sequences using WU-BLAST (see Basic Protocol 1), text-based searches using various sequence identifiers (GenBank Accessions and Tentative Consensus (TC) identifiers), searches by tissue and library names and gene names, and searches using functional classes through Gene Ontology assignments (*UNIT 7.2*). In addition, a comprehensive annotation of all ESTs in the database, based on the annotation of the TCs in which they are contained, is provided.

The Eukaryotic Gene Ortholog database (EGO; see Basic Protocol 3), which uses DNA sequence-based comparisons to identify tentative ortholog pairs by linking across the various Gene Index databases, also provide a means of entry. In addition to providing for sequence-, accession-, and gene name-based searches, the DFCI Gene Index is also cross-referenced to the Online Mendelian Inheritance in Man (OMIM) database (*UNIT 1.2*),

allowing users to link to Tentative Ortholog Groups (TOGs), and from there to representative sequences in the individual gene index databases. RESOURCERER, designed to annotate and cross-reference mammalian orthologs, as well as the Genome Viewers, also provide means of entry to the databases.

The Gene Index Databases are constructed within a species-specific framework, and users should keep this in mind while using this resource. Although some general search utilities exist (such as BLAST searches; see Basic Protocol 1), most searches begin with a selection of a target species (see Alternate Protocols 1 to 5). Each species has a distinct home page that can be reached through a URL of the form <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=xxxx>, where xxxx is the appropriate code from the “Common name” column in Table 1.6.2 (see the Appendix at the end of this unit). Within the Gene Index for each species, the primary resources available are detailed reports for each of the component sequences, including the assembled TCs and the individual ESTs, as well as expressed transcripts (ETs), which are typically annotated CDS features in GenBank records. In most of the following protocols, the Maize Gene Index will be used as an example; similar tools and pages exist for the other databases, although the appropriate gene index name must be substituted in the queries (see Table 1.6.2 for the full list).

The completion of a number of eukaryotic genomes provides the opportunity to search the Gene Index databases by their physical location. A list of available genomes can be found by following the Genomic Maps link on the DFCI home page to the mapping page, <http://compbio.dfci.harvard.edu/tgi/map.html>. A detailed guide to doing such searches is provided below (see Basic Protocol 2).

TCs from one species can also be found through the mapping of possible orthologs. The Eukaryotic Gene Ortholog (EGO) database catalogs tentative ortholog groups based on shared DNA sequence using pairwise reciprocal best matches between species. Details on using this resource are also included in the unit (see Basic Protocol 3).

The protocols below provide examples of ways to use the Gene Index Databases to extract and explore the information they provide. The examples are not meant to be exhaustive, but rather illustrative. Users should note that new features and new species are continuously being added and that updated versions of these databases are released every four months (February 1, June 1, and October 1).

BASIC PROTOCOL 1: IDENTIFYING A TENTATIVE CONSENSUS (TC) REPRESENTING A SPECIFIC SEQUENCE WITH BLAST

If one has either nucleotide or amino acid sequences, WU-BLAST 2.0 can be used to search the collection of TCs, singleton ESTs, and singleton ETs from each species.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

Files—FASTA-formatted sequence (APPENDIX 1B)

1. Open the BLAST search page (Fig. 1.6.3) in the DFCI Gene Indices Web site by one of the following methods.
 - a. Connect to the Gene Indices home page (<http://compbio.dfci.harvard.edu/tgi/tgipage.html>) and select the BLAST hyperlink from the top menu bar under the Gene Indices pull-down menu (Fig. 1.6.1).
 - b. Click on the BLAST link under the “Sequence Similarity Search” heading on the DFCI Maize Gene Index home page (<http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=maize>; Fig. 1.6.2) or the corresponding home page for another species.
 - c. Directly enter the BLAST search URL <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/Blast/index.cgi>.
2. From the Program pull-down menu, select the search program to run: BLASTN (UNIT 3.3) for a nucleotide query sequence or TBLASTN (UNIT 3.4) for a protein query, which will be searched against the six-frame translation of the appropriate TGI nucleotide database.

A SAGE tag is a short nucleotide sequence (typically 10 or 14 bp) that has been found within an mRNA through the construction and sequencing of a Serial Analysis of Gene Expression (SAGE) library (Velculescu et al., 1995).

SAGE10 and SAGE14, also included in the Program pull-down menu, are BLASTN searches using parameters optimized to search SAGE tags 10 and 14 nucleotides in length, respectively.

3. From the Database pull-down menu, select an appropriate target database; one or more databases can be specified at each time by holding down the Control key while clicking within the list.
4. Scroll down to the middle of the page. Enter an appropriate FASTA-formatted sequence either by uploading a file containing a single sequence using the Browse button or pasting it directly into the text box.

The TGI BLAST server does not presently allow multiple sequences to be searched simultaneously, although such a utility is under development.

Note that although there is no a priori limit on sequence length, some browsers may time out during searches of long sequences.

5. Users may also select the options other than the defaults for various parameters, including Alignments (using the pull-down menu right below the Program pull-down menu), and Matrix, Filter, Expect, Cutoff, Strand, Descriptions, Wordlength, Echofilter, Graphical Overview, and Ignore Hypotheticals (using the pull-down menus near the bottom of the page).

Descriptions for these options can be found by clicking on each button.
Further discussion of the parameters can be found in UNIT 3.3.

6. Users are also provided with the option of supplying an e-mail address where they will be notified when the search is completed.

Although most searches are completed quickly, search time depends on the sequence length and databases selected, as well as machine use. Search results are held for 48 hr and then discarded.

7. Standard BLAST search results are returned with alignments. Hyperlinks have been added to each of the identified target sequences. Target sequence names are specified in one of three formats depending on their source:

For TC:

species|TCxxxxx; 'THC' for human

For EST:

species|est name

For ET:

species|NP[ET]xxxxxx|GBnucleotide accession| GBprotein accession.

8. Click on the name of the target sequence to retrieve the corresponding TC report. Review the selected TC report (see Guidelines for Understanding Results).

ALTERNATE PROTOCOL 1: SEARCHING BY TENTATIVE CONSENSUS, EXPRESSED TRANSCRIPTS, EXPRESSED SEQUENCE TAG, OR GENBANK IDENTIFIER

The TCs within each gene index can be searched using a variety of accessioned identifiers that users may get from a variety of sources, including publications or other database searches. Identifiers that can be used include the TC identifiers, GenBank accession numbers, EST IDs, and Expressed Transcripts (ETs/NPs).

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Starting from a species home page (e.g., Fig. 1.6.2), click on the “Identifiers or Keywords” link to open the search page.

For each species, the appropriate URL is of the form <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gireport.pl?gudb=xxxx>, where xxxx is the common name from Table 1.6.2. Figure 1.6.4 shows the search page for maize.

2. For the identifier chosen, complete the appropriate entry on the form and click the GO button. Be aware that each of the three types of identifier has a slightly different specification:

For TC:

TC#, the TC identifier, can be either THCxxxxxx (for human) or TCxxxxxx (any other species), or just the numerical part of the TC number, xxxxxx.

For ET:

GB# can be either the GenBank accession of a sequence containing an annotated CDS, or the corresponding GenPept protein sequence accession.

NP#, the DFCI accession for each CDS feature parsed from GenBank records, can be of the form NPxxxxx where the HT/ET designators are used to maintain continuity with the DFCI qcGene database (<http://compbio.dfci.harvard.edu/tgi/qcGene.html>).

For EST:

GB# is the GenBank accession of an EST sequence.

EST ID is the EST number in each dbEST record.

CLONE Name is a cDNA clone identifier, such as an IMAGE ID, associated with a particular sequence.

3. For TC number searches, the standard TC report (see Guidelines for Understanding Results) is returned. For ET and EST searches, the search provides a sequence report page, with links to the relevant TC report if the ET or EST sequence is not a singleton.

Unlike accessions in some other databases, the TGI TC numbers are retired with each build and a new set of accessions is provided. However, a significant effort has been made to track TC identifiers from one release to the next, and the header line for each TC FASTA sequence contains the history of that assembly. Because this information is stored in a relational table, users can search the database using an “expired” TC number and get the current incarnation.

4. To search keyword(s) in annotations, enter the name to be searched as keyword(s) or a Boolean expression and hit the GO button. Keep in mind that gene name searches can be inaccurate, as many genes have multiple names and aliases and that the gene names in the TGI databases are not curated.

When an exact name search does not yield the expected result, more general terms related to the target or alternative names should be tried. As trusted databases with curated gene names become available, these will be used to update the annotation in TGI.

The search returns a table with information about the query sequence, including links to the TC in which that sequence is contained.

ALTERNATE PROTOCOL 2: SEARCHING BY GENE ONTOLOGY FUNCTIONAL CLASSIFICATION

Gene Ontology (GO; Ashburner et al., 2000; *UNIT 7.2*) terms provide classification for proteins based on three classes: Molecular Function, Biological Process, and Cellular Component. GO terms and Enzyme Commission (EC) numbers are assigned to TCs by using BLASTX (*UNIT 3.4*) to compare the sequence to the SwissProt database and then using a SwissProt-to-GO translation table provided by the GO Consortium (<http://www.geneontology.org>). For inexact matches, the DFCI Gene Indices are conservative and assign more general terms so as to avoid misclassification. It should be noted that because GO is evolving, many genes and TCs have not as yet been assigned precise classifications.

The GO terms can be used within any species to find those TCs likely to have a specific function or to be involved in particular processes.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Starting from a species home page (e.g., Fig. 1.6.2), click on the Gene Ontology link under the “Functional Annotation and Analysis” heading to open the Gene Ontology Assignments page.

Each of the species represented in the TGI has assigned GO terms (UNIT 7.2), and the GO assignments are summarized on a page accessible at a URL of the form <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/GObrowser.pl?species=Maize&gidir=zmg> (see Fig. 1.6.5). This page lists the number of TCs with each class and a bar graph shows the fraction of all TCs and those with GO assignments falling into each class.

2. Clicking on the functional category of interest, such as “reproduction” brings up a GO browser (Fig. 1.6.6), which shows both of the subclasses that fall into that category. Each line includes the current level, child ids, child GO terms, the number of TCs at this level, and the number of subtree TCs. Clicking on underlined entries in the last two columns brings up a list of TCs within that classification along with EC numbers linked to the KEGG metabolic pathway database (Kanehisa and Goto, 2000).

ALTERNATE PROTOCOL 3: SEARCHING BY RADIATION HYBRID MAP LOCATION (FOR HUMAN, MOUSE, AND RAT ONLY)

The TCs for human, mouse, and rat have been mapped to their corresponding genomes using the corresponding radiation hybrid (RH) maps that are available. Although genome sequence is rapidly becoming available for these species, the RH maps remain useful because many of the markers have also been placed on linkage maps, and as such provide a useful resource for candidate gene identification in genetic mapping studies. To produce the

mapped transcript set, the ECs are mapped to the appropriate genome using e-PCR (Schuler, 1997) and the marker data available from a variety of sources (see <http://compngen.rutgers.edu/EnhancedMaps/Default.aspx> for a summary). The following method will help find the TCs related to a specific map location.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Open the URL http://compbio.dfci.harvard.edu/tgi/gi/xxxx/searching/rh_map.html where *xxxx* is *hgi* for human, *mgi* for mouse, or *rgi* for rat.

Figure 1.6.7 shows the RH page for mouse (http://compbio.dfci.harvard.edu/tgi/gi/mgi/searching/xpress_search.html).

2. Select the chromosome to view and set the number of records to be displayed on each page.
3. Figure 1.6.8 shows the resulting table containing columns for TC#, Marker, 5' marker position in TC, 3' marker position in TC, Panel, Chromosome location, and P-value (from the RH map).

Here, the 5' and 3' positions refer to where the mapped RH marker falls within the mapped TC. Users will be most interested in examining the RH map location, which provides relative coordinates on the chromosome.

ALTERNATE PROTOCOL 4: SEARCH GENE EXPRESSION BY LIBRARY ANNOTATION

TCs can be identified based on patterns of gene expression determined using the annotation of the libraries from which the component ESTs were derived (Smith et al., 2001). It should be noted that EST library information in dbEST is not curated, and as such may not be correct or may be represented using nonstandard language. While an attempt has been made to correct some inconsistencies in the representation of the tissues from which libraries were derived, many remain.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

- 1a Access expression information through a URL of the form http://compbio.dfci.harvard.edu/tgi/gi/xxxx/searching/xpress_search.html, where *xxxx* is replaced by *gi_symbol* (Table 1.6.2) representing the species of interest.
- 1b Alternatively, starting from a species home page (e.g., Fig. 1.6.2), click on the “Libraries” link under the “Sequence Reports” heading. Figure 1.6.9 shows an example from maize.

- 2a** *To identify TCs in a given tissue:* In the top section of the page (see the “Search for Tissue Specific Transcripts” section in Fig. 1.6.9), specify a tissue or organ of interest and a minimum percentage for representation of ESTs from that organ within a TC.

For example, specifying “root” and 50% will return all TCs in which more than 50% of the ESTs are from root. Clicking on the Search button returns a table formatted to include:

1st column: the TC number for each TC satisfying the specified criteria, linked to a TC report.

2nd column: the number of ESTs from specified tissue or organ and the total number of ESTs within that TC.

3rd column: the fractional representation of the specified tissue or organ among ESTs in that TC.

4th column: the library catalog numbers (cat#s) corresponding to the tissue or organ of interest with links to the appropriate library report.

5th column: the number of ESTs from each specific library within this TC.

6th column: the number of ESTs from component libraries for all TCs.

7th column: the number of EST singletons from component libraries.

- 2b** *To identify TCs associated with a keyword:* In the upper middle section of the page (see the “Search cDNA Libraries by Keyword” section in Fig. 1.6.9), enter one or more keywords.

A list of all libraries annotated with those terms is returned, with links to the appropriate library reports.

- 2c** *To identify TCs associated with library identifiers:* In the lower middle section of the page (see the “Search cDNA Libraries by Library Identifier” section in Fig. 1.6.9), enter the library identifier.

Users can also retrieve library reports by searching the Gene Index databases using the appropriate library identifier parsed from GenBank EST records. These are the “dbEST lib id” fields from GenBank, and it should be noted that as these are not curated, some inconsistencies do exist in the annotation. Users are provided with a list of all TCs linked to the appropriate TC report containing one or more ESTs annotated as coming from a particular library.

- 2d** *To compare TCs expressed in two different tissues or organisms:* Compare patterns of gene expression based on library annotation and identify TCs that are statistically significantly differentially expressed in any one library relative to others by clicking “Scan a list of TCs by Library Expression” (Fig. 1.6.9) at the bottom of the page.

This produces a list of libraries from which the user can select those of interest (Fig. 1.6.10).

Users should note that tissue designations come from the library annotation provided in GenBank records, and, as such, the same tissue may be represented by different tissue terms. Users can therefore select multiple tissues for each of the two groups they wish to compare. Clicking on the “Get Expression” button returns a graphical matrix representation of expression in which each row represents a TC and columns represents the R stat, TC#, the number of ESTs in that TC, the number of ESTs found in libraries selected in group A, and the number of ESTs found in libraries selected in group B (Fig. 1.6.11).

The results illustrated in Figure 1.6.11 were obtained by selecting tissue type “aerial, root, whole plant” for Group A, and “aleurone layer” for group B.

Significant differential expression is identified using the “R statistic” (Stekel et al., 2000); a large R for a TC indicates that there is a significant bias toward one or more libraries in that TC.

A–B contains all TCs with more than one EST in A and zero or one EST in B.

ALTERNATE PROTOCOL 5: SEARCHING BY METABOLIC PATHWAY

The Gene Index databases can also be searched by means of metabolic pathway maps.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Starting from the appropriate Gene Index home page (e.g., Fig. 1.6.2), select “Metabolic Pathways” link under the “Functional Annotation and Analysis” heading to produce a graphical representation of a number of metabolic pathways.
2. Select an appropriate pathway.

A list of TCs corresponding to elements in that pathway is returned. These can be used to bring up TC reports corresponding to the individual pathway elements.

BASIC PROTOCOL 2: USING THE GENOMIC MAPS WITH THE DFCI GENE INDICES

Completed or draft genome sequences are now available for a number of eukaryotic species, including *Anopheles gambiae*, *Bos taurus*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Gallus gallus*, *Macaca mulatta*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Oryza sativa* (rice BACs from the

Japonica cultivar), *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. In addition, alignments of rice TC sequences for both the Indica and Japonica cultivars are mapped to the Indica contigs from the draft of that genome (Yu et al., 2002). For all maps, TCs are approximately localized within relevant genomes using MegaBLAST or BLAT, with final alignments performed using gap2, which incorporates splicing rules and is optimized for transcript-to-genome alignments. Mapping information is stored in a relational database and used to create user-friendly Web displays. Table 1.6.1 lists the genomes currently represented and the Gene Indices that are mapped to each genome.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. . Open the Genomic Maps page by one of the following methods.
 - a. Connect to the Gene Indices home page (<http://compbio.dfci.harvard.edu/tgi/tgipage.html>) and select the “Genomic Maps” hyperlink from the bar under the DFCI Gene Indices header (Fig. 1.6.1).
 - b. Directly enter the mapping page URL, <http://compbio.dfci.harvard.edu/tgi/map.html>.
2. Select the genome to explore by clicking on the appropriate icon.
3. Select an individual chromosome or BAC, as appropriate, for which mapping information is to be examined.
4. Examine the map.

A representative genome mapping display for Arabidopsis thaliana chromosome 1 is shown in Figure 1.6.12. The display is divided into two frames: the upper frame includes navigational and display tools, the lower shows a graphical representation of individual alignments TC alignments with the genome; putative exons are represented as colored boxes, introns as dashed lines, and unmatched regions of the TC as open boxes. To aid in navigating the display, individual species are distinguished by the color of the mapped TCs. Wherever available, the putative annotation of the genome is displayed at the top of the lower panel; in the case of human and mouse, this is the current EnSEMBL annotation. Additional markers may be added to these displays in the future, including genetic markers.

5. A region of the target chromosome can be selected either by clicking on the approximate position in the upper left corner of the top panel, or by entering approximate 3' and 5' coordinates. Placing the mouse over a TC in the lower panel returns information about that TC in the upper panel. At the bottom of the upper panel, the putative annotation is displayed and on the right hand side details of the alignment of each putative exon is provided.

BASIC PROTOCOL 3: USING EGO TO IDENTIFY ORTHOLOGOUS GROUPS

The Eukaryotic Gene Ortholog (EGO) database provides putative links between putatively orthologous TCs, as well as an indexed list of TCs linked to disease-associated human genes through the Online Mendelian Inheritance in Man (OMIM; *UNIT 1.2*) database. EGO is based on the results of high-stringency pairwise sequence comparisons between the TCs and singleton ETs from all TGI databases. Tentative Ortholog Groups (TOGs) are constructed using a transitive, reflexive closure process based on the assumption of parsimony to associate sequence-specific best hits, with the requirement that three sequences from separate species must be represented. Some TCs may belong to multiple TOGs, although TOGs containing significant overlap in their membership are merged. The result is that in some instances paralogous sequences appear in the same TOG, particularly if a sequence from a primitive eukaryote such as yeast is represented in the TOG. Each TOG is assigned a unique accession number (TOG #) that can be used to reference the collection of sequences. EGO has been a valuable tool for identifying orthologs of known genes as well as those existing only as uncharacterized ESTs.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Open the EGO page by one of the following methods.
 - a. Connect to the Gene Indices home page (<http://compbio.dfci.harvard.edu/tgi/tgipage.html>) and select the EGO hyperlink from the bar under the DFCI Gene Indices header (Fig. 1.6.1).
 - b. Directly enter the EGO URL, <http://compbio.dfci.harvard.edu/tgi/ego/>.

The main EGO page is returned (Fig. 1.6.13). On the EGO main page, there are links to two search functions, Search the Ortholog Database and Orthologs of Human Disease Genes.

2. Clicking on the Search the Ortholog Database button brings up a page that allows searches to be done using nucleotide or protein searches through BLAST (*UNITS 3.3 & 3.4*), using TOG numbers, using gene names, or using TCs from any of the species within EGO. The next page will be a list of orthologs. The title (Tentative Ortholog xxxx) is a link to a more detailed report. A representative TOG report for a putative transcription factor gene is shown in Figure 1.6.14A,B.
3. The Orthologs of Human Disease Genes link from the EGO home page allows searches by Omim Identifier, OMIM Locus ID, Gene name (such as CDK2, cyclin-dependent kinase 2) GenBank Accession number, DFCI Accession Number (for human only), or EGO Identifier.
4. TOG reports have three main parts as shown in Figure 1.6.14 (A,B). At the top is a table listing the component TCs with putative annotation and links to the component TC reports. There is also a graphic representing the connections

between the component sequences used for constructing the TOG. Below the TOG is a table listing the results of all pair-wise searches contributing to the TOG, with percent identity, match length, *p*-value, and asterisks marking reciprocal best hits. At the bottom of each TOG report is a ClustalW alignment showing the relationship between the aligned DNA sequences; this alignment can also be viewed using JalView (<http://www.jalview.org/>).

BASIC PROTOCOL 4: USING RESOURCERER

RESOURCERER is a microarray resource annotation and cross-reference database—i.e., a resource for microarray experiments that both provides annotation for widely used platforms and makes it possible to compare gene expression from experiments in one species with expression patterns discerned in the same or another species. Annotation for any microarray platform or clone set included in RESOURCERER is provided through the appropriate Gene Index database. Comparisons between different resources for one species are provided through TGI, and comparisons across species are derived from EGO. At present, only human, mouse, rat, zebrafish, xenopus, cattle, *C.elegans*, and rice are represented in RESOURCERER, but other species will be added as standard platforms come into widespread use. Microarray resources represented in RESOURCERER include cDNA clone sets, long oligo sets from Operon/Qiagen and Compugen/Sigma, and the Affymetrix GeneChips for representative species.

Necessary Resources

Hardware—Computer with Internet access

Software—Web browser

1. Open the RESOURCERER page by one of the following methods.
 - a. Connect to the Gene Indices home page (<http://compbio.dfci.harvard.edu/tgi/>) and select the RESOURCERER hyperlink from the bar under the DFCI Gene Indices header (Fig. 1.6.1).
 - b. Directly enter the RESOURCERER URL, <http://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl>.

The main RESOURCERER page is returned (Fig. 1.6.15) with summary instructions on its use and links to a more extensive README.

2. To obtain annotation for a single microarray resource already in the database, select resource using the drop-down menu, then Submit. On the next page, you can select annotation fields of your interest. Clicking the Get Table button returns annotations similar to that shown in Figure 1.6.16, including:

The clone name associated with each element, if available

Either the Rearray ID assigned by the clone set developer or the Affymetrix Probe ID, as appropriate

A representative GenBank Accession number

UniGene IDs

Locus Link IDs

Physical map location based on alignments of the DFCI TCs to the appropriate draft genome sequence

The TC number for the appropriate species

TC numbers from the other mammalian species

Assigned GO Terms based on the TC assignment

Putative annotation

For mouse, a corresponding Mouse Genome Informatics (MGI) database accession.

Where appropriate, elements in the table are hot-linked to an appropriate database, including the NetAffx database for Affymetrix probe_ids.

3. To compare the elements represented in two microarray resources, on the main RESOURCERER page (Fig. 1.6.15), simply select a first resource as Data Set A and the “Compare to another resource” radio button. On the next page, you will select the second source—Data Set B. Also, select whether the comparisons should be made through EGO (and the TGI), which returns valid comparisons either within a single species or between species, or UniGene IDs, which is only valid for comparisons within a single species. Finally, select the type of comparison: whether the search should return those elements in common to both Data Set A and Data Set B (Intersection), those unique to Data Set A (A unique), or those unique to Data Set B (B unique).

Clicking the Get Table button returns a cross-reference table that contains annotations similar to that shown in Figure 1.6.17, again with appropriate links to other databases.

GUIDELINES FOR UNDERSTANDING RESULTS

Examining a TC Report

The TC sequences are the central elements in the DFCI Gene Index databases. The TCs are assembled from EST and gene sequences and represent likely transcripts encoded within a particular genome. In that sense, the TCs are distinct from clusters in other approaches such as UniGene in that alternative splice forms and gene family members are more likely to be represented by separate objects in our databases. This has some advantages and disadvantages based on one's application of interest. In principle, with a large-enough collection of ESTs, sequences representing a wide variety of tissues, developmental stages, and disease states, the Gene Index databases would reconstruct the entire transcriptome of a particular organism.

Figure 1.6.18A,B,C shows a representative TC with the annotation and features provided in each Gene Index. TCs are indexed by an accession of the form TCyyyy, where yyyy is a number that is simply a sequential identifier assigned each time each database is rebuilt. For each species-specific database, TC numbers are never reused. However, TC numbers are tracked through subsequent builds so that users with a TC number from a previous release of the database can get the current representation of that particular transcript.

TC reports can be accessed in a variety of ways, many of which are detailed below. However, users can link directly to any TC report by entering a URL of the form <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/tcreport.pl?species=xxx&tc=yyyy>, where xxx is the common species name (typed exactly as in the table, without spaces) from the first column of Table 1.6.2 (no spaces) and yyyy is the TC number of interest. If a TC number from a previous build is used, the corresponding TC from the current build is provided. Note that for human, *TC* is replaced by *THC*.

Each TC report contains the following features:

1. At the top of Figure 1.6.18A is a FASTA-formatted sequence representing the consensus assembly of its component EST and gene sequences. The FASTA header includes the current TC number assigned to that assembly, as well as previous TC numbers associated with it; this allows users to track TC numbers through various rebuilds of the database. Wherever possible, predicted polyadenylation signals are identified and highlighted in red within the sequence.
2. Immediately below the FASTA sequence is a graphical representation of the TC with putative open reading frames (ORFs) predicted using NCBI's ORF Finder, ESTScan, FRAMEFINDER, and DIANAEST (Iseli et al., 1999; Hatzigeorgiou et al., 2001). ORF Finder scans each of the six potential reading frames looking for ORFs; the remaining programs use a variety of approaches to identify and correct reading frame errors and to select the most likely ORF for each TC. The bars representing the ORFs are active; clicking on them takes the user to a page from which they can explore the properties of the predicted protein-coding sequence (Fig. 1.6.18A).
3. Below the predicted ORFs is a map representing the individual sequences that comprise the TC, showing their approximate position in the TC and their relative lengths. Each sequence is represented by an arrow showing orientation, and paired reads from the same clone are linked by a dotted line. Annotated mRNA sequences are highlighted in pink. All sequences are numbered and indexed to a table of linked identifiers, which immediately follows the map (Fig. 1.6.18A).
4. A table lists the individual sequences comprising the TC, indexed by numbers appearing in the sequence map (Fig. 1.6.18B). Each row in the table represents a particular EST or gene sequence and these are annotated with a source laboratory (wherever possible), a sequence ID, a GenBank accession, clone name, the 5' position in the TC, 3' position in the TC, and source library annotation. Wherever possible, these entries are linked to other databases or sources of information. The sequence ID is linked to an EST report or ET report page at DFCI, the GenBank

accession is linked to a sequence record at NCBI, and the clone name, wherever possible, is linked to a public clone repository. Immediately following the list is a key to the clone source codes used, showing the laboratories from which the clone sequences were generated; all ET sequences are coded as ETG. Links to laboratories contributing a significant number of ESTs from a particular species can be found on the home page for each species-specific Gene Index.

5. Assembling the TCs can produce consensus sequences with arbitrary orientation. Using annotated information about the component sequences, including the presence of mRNA sequences and the 3' and 5' orientation of the ESTs, one attempts to identify the appropriate orientation of the TC and provide the evidence used for that determination (Fig. 1.6.18B).
6. Alternative splice forms, identified through alignment of TCs within each TGI database, can be found by clicking on the "Alternative Splice Forms" button (when info is available).
7. An expression summary, based on the libraries from which the ESTs were derived, can be found by clicking on the Expression Summary button (Fig. 1.6.18C).
8. Putative gene identification is made using a variety of methods. First, TCs are annotated using the names associated with any mRNA sequences they contain; this is listed as the Putative ID for each TC. The consensus sequences are also searched against a nonredundant protein database; the top five hits are listed and a controlled vocabulary is used with these to assign a name to each (Fig. 1.6.18C).
9. The "GO annotation" lists assignments based on the Gene Ontology project's classifications (<http://www.geneontology.org>; UNIT 7.2). TCs are searched against SwissProt and SwissProt-to-GO tables to provide conservative assignments based on the level of sequence homology (Fig. 1.6.18C).
10. Potential orthologs are identified through the EGO database. A detailed description of the EGO database is provided in Basic Protocol 3.
11. TC sequences are also mapped to a variety of completed eukaryotic genomes. At the bottom of each TC report is a "Maps to" section with links to alignments with draft or completed genomic sequences from model organisms.
12. TC reports may also contain buttons providing links to single-nucleotide polymorphisms (SNPs) identified in the TC sequence, as well as predicted 70-mer oligos for microarray projects (Fig. 1.6.18C).

COMMENTARY

Background Information

The goal of any genome project is the identification and functional characterization of the entire catalog of the genes encoded within a particular genome. Although genome sequencing projects in human, mouse, *Arabidopsis*, and other eukaryotic species have generated a wealth of data, identification of the genes encoded in the sequence and assignment of function to these remains a significant challenge. Nowhere is that more

apparent than in the two completed drafts of the human genome (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001), where an independent analysis of the competing annotations has found that many of the gene predictions, other than for previously known genes, are disjoint (Hogenesch et al., 2001). Indeed, it is becoming increasingly clear that the completion of a genome sequence is only a starting point and that significant additional analysis is required before one can declare its annotation, and the genome itself, complete.

The sequencing of ESTs continues to supply important insight into the transcribed genes in a wide variety of species and has become a widely used approach to gene discovery and the analysis of gene expression. ESTs are the most extensive available survey of the transcribed portion of the eukaryotic genomes; there are currently more than 10,000,000 ESTs in GenBank, nearly 45% of which are human and 75% of which represent higher mammals (human, mouse, rat, cattle, and pig; http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). For many species, ESTs remain the primary source of gene sequence data and provide a basic survey of gene expression in various tissues, as well as in various developmental and disease states. ESTs have also proven their value in genome annotation as they provide experimental evidence for the presence of the genes, their genomic structure, and patterns of expression.

However, analysis of ESTs presents a number of challenges as each sequence typically represents only a partial gene sequence and EST projects generally produce very large numbers of redundant sequences. The DFCI Gene Indices (TGI; Quackenbush et al., 2001; http://compbio.dfci.harvard.edu/tgi/publications/NAR_GeneIndex2001.pdf) attempt to avoid these limitations by first clustering, then assembling ESTs to reconstruct the original gene transcripts (mRNAs) as high-fidelity, virtual transcripts. While there are many other projects that cluster ESTs, including UniGene (Boguski and Schuler, 1995) and IMAGEne (Cariaso et al., 1999), and others that assemble EST clusters such as STACK (Christoffels et al., 2001) and DoTS (<http://www.allgenes.org>), the DFCI Gene Indices have distinguished themselves by producing high-fidelity EST assemblies for over 60 species (see Table 1.6.2). The indices provide annotation and other ancillary information about the genes, their structure, genomic localization (Quackenbush et al., 2000, 2001), and potential orthologs and paralogs (Lee, 2002), and serve as a resource for comparative sequence analysis (Tsai, 2001).

Obtaining data from the DFCI Gene Index databases through FTP—As an alternative to using the Web site, flat-file versions of all of the DFCI databases are available through FTP links on each Gene Index home page and the EGO page; RESOURCER flat files can be downloaded through the Web site. The Gene Index download files include:

1. A multi FASTA file with TC sequences (annotation in the defline) and singleton sequences.
2. A FASTA file, containing the complete set of TC sequences for that species with TC identifiers from previous builds in the definition line.
3. A tab-delimited file containing the TC identifiers and the ESTs that comprise them.

The FASTA files can be used to create local BLAST databases, or used for other purposes. The file that includes TC numbers and the list of ESTs can be used for linking ESTs to the TCs that contain them.

4. A multi-FASTA file indexed by TC; information includes GO ID, GO Term, E.C. Number, GO category.
5. Oligomer data is available for some gene indices.

Assembly of the Gene Index databases—The DFCI Gene Indices are assembled independently for each species using a “divide and conquer” approach in which ESTs are first placed in clusters based on sequence similarity and then assembled on a cluster-by-cluster basis to produce Tentative Consensus (TC) sequences (Liang et al., 2000; Quackenbush et al., 2001). A schematic overview of this process is shown in Figure 1.6.19 and a software implementation of the clustering and assembly tools used, TGICL, is freely available (Pertea et al., 2003; <http://compbio.dfci.harvard.edu/tgi/software/>). TGICL is an open-source pipeline for analysis of large EST and mRNA databases, in which sequences are first clustered based on pairwise sequence similarity and then assembled to produce the TC sequences.

Briefly, ESTs and coding gene sequences are first downloaded from dbEST and parsed from GenBank records. The annotated CDS features in GenBank records are assigned NP (for nucleotide-protein) identifiers to provide a unique accession for each coding DNA sequence; some GenBank records have multiple annotated coding features. Sequences are trimmed to remove vector, poly(A/T) tails, adaptor sequences, and contaminating bacterial sequences. Clustering begins by indexing a multi-FASTA-formatted sequence database and performing all-versus-all pairwise similarity searches. The authors use mgbblast, a modified version of the megablast program (Zhang et al., 2000), for this purpose. The mgbblast program differs from the original megablast program in that it produces a simple tab-delimited output, uses specific output filtering options such as minimum overlap length and identity, and allows the use of a dynamic offset within the database when performing incremental searches of portions (slices) of the database against itself. Each line in the mg-blast output represents one identified overlap between two sequences in the database. The search results are sorted in order by decreasing pairwise alignment score. The sequence overlaps are filtered using user-defined criteria: the minimum overlap length (default 40 base-pairs), the minimum percent identity for the overlap (default 95%), and the maximum mismatched overhang allowed around the overlap (dynamically adjusted for long sequences and long overlaps; the default value starts at 30 nucleotides). Based on the results of these similarity searches, sequences are grouped into clusters using a transitive closure approach and a graph representation in which the sequences are the graph nodes and the alignments represent edges (Pertea et al., 2003); the resulting clusters represent the connected subgraphs within the dataset.

This clustering stage is an important step if one then wants to assemble the expressed sequences to reconstruct the transcripts they represent. Most sequence-assembly programs were developed for genomic applications and face particular difficulty in dealing with the challenges presented by ESTs, including extremely deep and uneven coverage from diverse

biological sources, low-quality sequences often without quality scores, relatively frequent chimerism, and a moderately high rate of vector and adapter contamination. Further, while many DNA sequence assembly programs assemble contigs from large numbers of sequences, they can easily be overwhelmed by a very large unpartitioned dataset and produce incorrect chimeric assemblies (Liang et al., 2000).

A systematic analysis of the performance of various sequence-assembly programs (Liang et al., 2000) led the authors of this unit to select the Paracel Transcript Assembler (PTA), an improved version of CAP3 (Huang and Madan, 1999), to independently assemble each cluster. The assembly process produces a collection of Tentative Consensus sequences (TCs) and a set of unassembled singletons. The TCs are annotated in preparation for release on the DFCI Gene Index Web site. First, TCs are searched against a variety of DNA and protein databases and high-scoring hits are used to provide putative functional annotation using a controlled vocabulary. Hits to SwissProt records are used to assign Gene Ontology (GO) terms and Enzyme Commission (EC) Numbers using a SwissProt to GO translation table provided by the GO consortium (<http://www.geneontology.org>; UNIT 7.2). Open reading frames in each sequence are assigned using NCBI's ORF Finder, ESTScan, FRAMEFINDER, and DIANAEST (Iseli et al., 1999; Hatzigeorgiou et al., 2001). The orientation of each TC is determined using a consensus-based approach that uses the orientation and identity of its component sequences. Additional information and annotation for each sequence is provided through links to the EGO database (see below), to completed genomic sequences, to other maps where available, and to other appropriate annotation databases including the Mouse Genome Database at The Jackson Laboratory (<http://www.jax.org>). The TGI home page is shown in Figure 1.6.1 and a representative TC report in Figure 1.6.18A,B,C.

Evaluation of orthologous genes—Cross-referencing the available genomic data has a number of important applications, including the identification of homologous genes in eukaryotes. Gene homologs can be separated into two classes, orthologs and paralogs (Fitch, 1970). Orthologs are genes that are related by direct evolutionary descent while paralogs are homologous genes that are the result of a duplication event within the same lineage. The identification of orthologs is particularly important since these genes should play similar developmental or physiological roles and should therefore share conserved functional and regulatory domain. Further, the study of these genes in one organism can provide insight into their function in others.

Makalowski and Boguski (1998) conducted what was at the time the most comprehensive survey of eukaryotic orthologs available. Their dataset contained 1880 rodent-human ortholog pairs and 470 sequences shared by all three species. Their analysis of both the coding and noncoding regions indicated that not only are both the DNA and protein coding regions highly conserved in mammals, but, more surprisingly, that the flanking 5' and 3' noncoding regions are extremely well conserved and that the evolutionary distance estimated for the 5' and 3' UTRs are similar and generally indistinguishable from that for synonymous coding sites. This suggested to the authors of this unit that EST sequences, derived primarily from the 3' UTR, could be used to identify orthologs in closely related species. Based on this observation, and the fact that the TC sequences within the DFCI Gene

Index databases represented the most comprehensive survey of eukaryotic gene sequences available at the time, the authors began construction of the Eukaryotic Gene Ortholog (EGO; Lee et al., 2002) database in 1999. EGO has allowed identification and cataloging of more than 86,630 tentative orthologous groups in eukaryotes and it provides a tool for cross-referencing other genomic resources, including commonly used resources for DNA microarrays (Tsai et al., 2001).

Identification of Tentative Ortholog Groups (TOGs)—Tentative Consensus sequences (TCs) and the singleton Expressed Transcripts (sETs) from each of the DFCI Gene Indices are concatenated into a single multiFASTA database, which is partitioned and used in all-versus-all pairwise searches using mgbblast. Matches scoring better than a maximum e-value of 10^{-10} are recorded. Reciprocal best hits, defined as pairs of sequences from separate species that independently identify each other as a best match in their respective species, are identified, and a transitive closure process using these pairs and requiring sequences in three or more species is used to identify tentative orthologs (TOGs). Multiple alignments of each of the TOG sequences are preformed using ClustalW (Thompson et al., 1994; *UNIT 2.3*) and are displayed at <http://compbio.dfci.harvard.edu/tgi/ego/> with links to the individual TC reports; alignments can also be viewed using JalView (go to <http://www.jalview.org/>). The individual sequences in EGO can be searched by BLAST (*UNITS 3.3 & 3.4*) and all of the orthologous genes are cross-referenced to the Online Mendelian Inheritance in Man (OMIM; *UNIT 1.2*) database. A representative TOG is shown in Figure 1.6.14A,B.

Annotation of mammalian microarray resources—DNA microarray analysis (Schena et al., 1995) has emerged as one of the most widely used techniques for assessment of gene expression on a genomic scale, allowing tens of thousands of genes to be assayed in a single experiment. However, the widespread use of this technique has resulted in a proliferation of experimental platforms and reagents, making a comparison of results from different experimental groups a significant challenge. An additional and possibly more important need is the ability to make comparisons of gene expression patterns between species. Analysis of expression in model organisms, particularly mouse and rat, has become a fundamental tool for the study of human development and disease. Effective use of these animal models with microarray assays requires the development of a convenient means of identifying corresponding array elements between species and platforms. To address these issues, the authors of this unit developed RESOURCERER (<http://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/p1.pl>), a utility designed to provide annotation for and comparisons between widely used microarray platforms. RESOURCERER provides information for the most widely used microarray mammalian gene resources, including the Research Genetics Sequence Verified Human cDNA clone set, the BMAP and NIA mouse clone sets, the DFCI Rat Gene Index cDNA collection, human and mouse 70-mer oligo sets from Operon, and the Affymetrix Human, Mouse, and Rat GeneChip sets.

Literature Cited

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE,

- Ringwald M, Rubin GM, Sherlock G. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25:25–29. [PubMed: 10802651]
- Boguski MS, Schuler GD. Establishing a human transcript map. *Nat Genet.* 1995; 10:369–371. [PubMed: 7670480]
- Cariaso M, Folta P, Wagner M, Kuczmarski T, Lennon G. IMAGEne I: Clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes. *Bioinformatics.* 1999; 15:965–973. [PubMed: 10745985]
- Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W. STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.* 2001; 29:234–238. [PubMed: 11125101]
- Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool.* 1970; 19:99–113. [PubMed: 5449325]
- Hatzigeorgiou AG, Fizev P, Reczko M. DIANA-EST: A statistical analysis. *Bioinformatics.* 2001; 17:913–919. [PubMed: 11673235]
- Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, Schultz PG, Cooke MP. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell.* 2001; 106:413–415. [PubMed: 11534548]
- Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res.* 1999; 9:868–877. [PubMed: 10508846]
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409:860–921. [PubMed: 11237011]
- Iseli, C.; Jongeneel, CV.; Bucher, P. ESTScan: A program for detecting, evaluating and reconstructing potential coding regions in EST sequences. ISMB '99 (Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology); Menlo Park, Calif: AAAI Press; 1999. p. 138-148.
- Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28:27–30. [PubMed: 10592173]
- Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai T, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* 2002; 12:493–502. [PubMed: 11875039]
- Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* 2000; 28:3657–3665. [PubMed: 10982889]
- Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA.* 1998; 95:9407–9412. [PubMed: 9689093]
- Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics.* 2003; 19:651–652. [PubMed: 12651724]
- Quackenbush J, Liang F, Holt I, Pertea G, Upton J. The TIGR gene indices: Reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* 2000; 28:141–145. [PubMed: 10592205]
- Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* 2001; 29:159–164. [PubMed: 11125077]
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science.* 1995; 270:467–470. [PubMed: 7569999]
- Schuler GD. Sequence mapping by electronic PCR. *Genome Res.* 1997; 7:541–550. [PubMed: 9149949]
- Smith TP, Grosse WM, Freking BA, Roberts AJ, Stone RT, Casas E, Wray JE, White J, Cho J, Fahrenkrug SC, Bennett GL, Heaton MP, Laegreid WW, Rohrer GA, Chitko-McKown CG, Pertea G, Holt I, Karamycheva S, Liang F, Quackenbush J, Keele JW. Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. *Genome Res.* 2001; 11:626–630. [PubMed: 11282978]

- Stekel DJ, Git Y, Falciani F. The comparison of gene expression from multiple cDNA libraries. *Genome Res.* 2000; 10:2055–2061. [PubMed: 11116099]
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; 22:4673–4680. [PubMed: 7984417]
- Tsai J, Sultana R, Lee Y, Perteu G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J. RESOURCERER: A database for annotating and linking microarray resources within and across species. *Genome Biol.* 2001; 2:software0002.1–software0002.4. [PubMed: 16173164]
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science.* 1995; 270:484–487. [PubMed: 7570003]
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science.* 2001; 291:1304–1351. [PubMed: 11181995]
- 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. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science.* 2002; 296:79–92. [PubMed: 11935017]
- Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 2000; 7:203–214. [PubMed: 10890397]

APPENDIX: DFCI GENE INDICES

DFCI Gene Indices (Table 1.6.2) are a collection of species-based databases that assemble the Expressed Sequence Tags (ESTs) and the Expressed Transcripts (ETs) into Tentative Consensus (TC) sequences. Singletons (sET and sEST) are ET/EST sequences that are not incorporated into any of the TCs during assembly. TCs, sETs, and sESTs represent potentially unique sequences in TGI. As of June 2003, there were 61 species represented by a Gene Index database. Each line in the table provides information about a single database and includes a common name, species name, gene index name and version, the total number of TCs in the current release, and the number of singleton ETs and singleton ESTs. For some of the Gene Indices, ESTs were pooled from dbEST for the genus, not a single species. The table is broken into four groups representing animals (42 species), plants (47), fungi (10) and protists (15).

Table 1.6.2

Summary of DFCI Gene Indices (TGI), June 2009 Release

Common name	Species name	GI name	Release	TCs	sETs	sESTs
<i>Animal (42 species)</i>						
A.aegypti	<i>Aedes aegypti</i>	AeGI	5.0	25627	110	14880
A.burtoni	<i>Astatotilapia burtoni</i>	AbGI	2.1	1284	51	6675
A.salmon	<i>Salmo salar</i>	AsGI	4.0	49630	369	40458
A.variegatum	<i>Amblyomma variegatum</i>	AvGI	2.0	490	1	1661
B.malayi	<i>Brugia malayi</i>	BmGI	5.1	2565	52	7477
B.microplus	<i>Boophilus microplus</i>	BmiGI	2.1	9851	39	4696
Bear	<i>Ursus americanus</i>	UaGI	4.1	4925	29	12719
Black_tick	<i>Ixodes scapularis</i>	IsGI	3.0	20932	23	17437
C.elegans	<i>Caenorhabditis elegans</i>	CeGI	9.0	17951	5035	7933
C.intestinalis	<i>Ciona intestinalis</i>	CinGI	5.0	31571	147	16349
Catfish	<i>Ictalurus punctatus</i>	CfGI	7.0	5342	310	19908
Cattle	<i>Bos taurus</i>	BtGI	12.0	90392	491	110291
Chicken	<i>Gallus gallus</i>	GgGI	11.0	75408	860	112983
Cricket	<i>Laupala kohalensis</i>	LkGI	2.0	2562	0	6013
Dog	<i>Canis familiaris</i>	DogGI	7.0	32481	570	32481
Drosophila	<i>Drosophila melanogaster</i>	DGI	11.0	27100	1314	14124
Fathead_minnow	<i>Pimephales promelas</i>	PpGI	1.0	27048	0	29623
Frog	<i>Xenopus laevis</i>	XGI	10.1	56494	406	58231
Fugu	<i>Takifugu rubripes</i>	FGI	3.0	3961	550	7812
H.chilotes	<i>Haplochromis chilotes</i>	HchGI	1.1	2291	8	4140
H.red_tail_sheller	<i>Haplochromis sp red tail sheller</i>	HsGI	1.1	1942	0	4562
Honeybee	<i>Apis mellifera</i>	AMGI	5.0	12167	3202	9640
Human	<i>Homo sapiens</i>	HGI	17.0	328301	19585	736049
Hydra	<i>Hydra magnipapillata</i>	HmGI	1.0	15510	13	22276
Killifish	<i>Fundulus heteroclitus</i>	FhGI	4.0	9251	26	26933

Common name	Species name	GI name	Release	TCs	sETs	sESTs
Locust	<i>Locusta migratoria</i>	LomiGI	1.0	4355	19	7625
Macaca_cynomolgus	<i>Macaca fascicularis</i>	MfGI	1.0	12613	851	68561
Medaka	<i>Oryzias latipes</i>	OIGI	8.0	37198	230	30997
Mosquito	<i>Anopheles gambiae</i>	AgGI	9.0	22557	9603	18782
Mouse	<i>Mus musculus</i>	MGI	16.0	210249	10684	769704
O.volvulus	<i>Onchocerca volvulus</i>	OvGI	4.1	1205	30	3283
Pea_aphid	<i>Acyrtosiphon pisum</i>	AcpiGI	1.0	17251	8	17704
Pig	<i>Sus scrofa</i>	SsGI	13.0	104293	819	132636
R.appendiculatus	<i>Rhipicephalus appendiculatus</i>	RaGI	2.1	2642	24	4917
R.trout	<i>Oncorhynchus mykiss</i>	RtGI	7.0	40320	291	49408
Rat	<i>Rattus norvegicus</i>	RGI	14.0	76570	2497	105867
Red_flour_beetle	<i>Tribolium castaneum</i>	TrcaGI	1.0	8594	4514	15273
S.mansoni	<i>Schistosoma mansoni</i>	SmGI	7.0	19291	80	28026
Sheep	<i>Ovis aries</i>	OaGI	1.0	22305	311	28783
X.tropicalis	<i>Xenopus tropicalis</i>	XtGI	3.1	69590	87	81625
Zebra_finch	<i>Taeniopygia guttata</i>	TaguGI	1.0	14384	36	19443
Zebrafish	<i>Danio rerio</i>	ZGI	17.0	63667	829	85826
<i>Plant (47 species)</i>						
A.cepa	<i>Allium cepa</i>	OnGI	2.0	4063	27	8155
Apple	<i>Malus x domestica</i>	MdGI	2.0	31789	38	26448
Aquilegia	<i>Aquilegia</i>	AqGI	2.1	13556	111	7278
Arabidopsis	<i>Arabidopsis thaliana</i>	AtGI	13.0	34155	8632	39039
Barley	<i>Hordeum vulgare</i>	HvGI	10.0	41206	172	39345
Bean	<i>Phaseolus vulgaris</i>	PhvGI	3.0	11940	142	9415
Beet	<i>Beta vulgaris</i>	BvGI	2.0	4784	132	12235
C.reinhardtii	<i>Chlamydomonas reinhardtii</i>	ChrGI	6.0	15554	119	26535
Clementine	<i>Citrus clementina</i>	CiclGI	2.0	32287	2	10229
Cocoa	<i>Theobroma cacao</i>	TcaGI	3.0	17424	24	31514
Cotton	<i>Gossypium</i>	CGI	10.0	50069	80	66367
Cotton_raimondii	<i>Gossypium raimondii</i>	GoraGI	1.0	9508	0	15383
Grape	<i>Vitis vinifera</i>	VvGI	6.0	33638	14825	30513
Ice_plant	<i>Mesembryanthemum crystallinum</i>	McGI	5.0	3627	66	6706
L.japonicus	<i>Lotus japonicus</i>	LjGI	5.0	21367	39	20996
Leafy_spurge	<i>Euphorbia esula</i>	EuesGI	1.0	10727	8	15761
Lettuce	<i>Lactuca sativa</i>	LsGI	3.0	12505	71	17309
Maize	<i>Zea mays</i>	ZmGI	19.0	112156	310	202621
Medicago	<i>Medicago truncatula</i>	MtGI	9.0	29273	11494	26696
Morning_glory	<i>Ipomoea nil</i>	IpniGI	1.0	11754	39	9721
Moss	<i>Physcomitrella patens subsp.patens</i>	PpspGI	2.0	30695	16670	19787
N.benthamiana	<i>Nicotiana benthamiana</i>	NbGI	3.0	5861	106	10160
Oilseed_rape	<i>Brassica napus</i>	BnGI	3.1	47634	59	42634
Orange	<i>Citrus sinensis</i>	CsGI	1.0	26081	26	72791

Common name	Species name	GI name	Release	TCs	sETs	sESTs
Peach	<i>Prunus persica</i>	PrpeGI	2.0	9633	42	16412
Pepper	<i>Capsicum annuum</i>	CaGI	4.0	14747	104	17568
Petunia	<i>Petunia hybrida</i>	PhGI	2.0	2230	42	6457
Pine	<i>Pinus</i>	PGI	7.0	34181	145	27538
Poplar	<i>Populus</i>	PplGI	4.0	49638	249	49764
Potato	<i>Solanum tuberosum</i>	StGI	12.0	31567	186	29619
Prickly_lettuce	<i>Lactuca serriola</i>	LaseGI	1.0	8047	0	13958
Rice	<i>Oryza sativa</i>	OsGI	17.0	77158	19426	85212
Robusta_coffee	<i>Coffea canephora</i>	CocaGI	1.0	7420	6	10206
Rye	<i>Secale cereale</i>	RyeGI	4.0	1471	78	4038
Scarlet bean	<i>Phaseolus coccineus</i>	PcGI	1.0	22518	1	50410
Sorghum	<i>Sorghum bicolor</i>	SbGI	9.0	23442	257	22326
Soybean	<i>Glycine max</i>	GmGI	14.0	70880	161	62508
Spruce	<i>Picea</i>	Sgi	3.0	42051	39	38404
Sugarcane	<i>Saccharum officinarum</i>	SoGI	2.2	40016	43	76529
Sunflower	<i>Helianthus annuus</i>	HaGI	6.0	20130	269	32717
Switchgrass	<i>Panicum virgatum</i>	PaviGI	1.0	52936	0	32286
T.versicolor	<i>Triphysaria versicolor</i>	TverGI	2.0	7165	3	5644
Tall_fescue	<i>Festuca arundinacea</i>	FaGI	2.0	6686	10	13241
Tobacco	<i>Nicotiana tabacum</i>	NtGI	5.0	37223	237	63781
Tomato	<i>Solanum lycopersicum</i>	LeGI	12.0	25764	201	20884
Triphysaria	<i>Triphysaria</i>	TriphGI	1.0	17442	0	17043
Wheat	<i>Triticum aestivum</i>	TaGI	11.0	91464	256	124732
<i>Protist (15 species)</i>						
C.parvum	<i>Cryptosporidium parvum</i>	CpGI	5.1	3833	146	48
D.discoideum	<i>Dictyostelium discoideum</i>	DdGI	5.1	13819	520	3627
E.tenella	<i>Eimeria tenella</i>	EtGI	5.0	2992	201	5191
Leishmania	<i>Leishmania</i>	LshGI	6.0	16048	3702	5538
N.caninum	<i>Neospora caninum</i>	NcGI	5.1	2131	5	3878
P.berghei	<i>Plasmodium berghei</i>	PbGI	6.0	13433	202	10899
P.falciparum	<i>Plasmodium falciparum</i>	PfGI	9.0	8919	1773	4536
P.vivax	<i>Plasmodium vivax</i>	PvGI	2.1	4387	2839	2481
P.yoelii	<i>Plasmodium yoelii</i>	PyGI	5.1	7727	29	2770
S.neurona	<i>Sarcocystis neurona</i>	SnGI	6.0	1053	0	2596
T.brucei	<i>Trypanosoma brucei</i>	TbGI	5.1	5203	3930	1367
T.cruzi	<i>Trypanosoma cruzi</i>	TcGI	6.0	12319	186	2742
T.gondii	<i>Toxoplasma gondii</i>	TgGI	9.0	10184	197	15498
T.thermophila	<i>Tetrahymena thermophila</i>	TtGI	5.0	12363	15594	6709
T.vaginalis	<i>Trichomonas vaginalis</i>	TvGI	2.1	4740	24227	461
<i>Fungi (10 species)</i>						
A.flavus	<i>Aspergillus flavus</i>	AfGI	5.0	4026	41	4070
A.nidulans	<i>Aspergillus nidulans</i>	AnGI	5.0	3662	6788	3106

Common name	Species name	GI name	Release	TCs	sETs	sESTs
C.posadasii	<i>Coccidioides posadasii</i>	CpoGI	2.1	6893	5522	2288
Cryptococcus	<i>Cryptococcus neoformans</i>	CrGI	8.0	8430	157	1447
F.verticilliioides	<i>Fusarium verticilliioides</i>	FvGI	8.0	8510	33	4807
M.grisea	<i>Magnaporthe grisea</i>	MgGI	6.0	13984	5964	10760
N.crassa	<i>Neurospora crassa</i>	NcrGI	4.1	10927	2092	1477
Potato_late_blight	<i>Phytophthora infestans</i>	PhinGI	1.0	12149	1	24321
S.cerevisiae	<i>Saccharomyces cerevisiae</i>	ScGI	4.0	4382	1443	197
S.pombe	<i>Schizosaccharomyces pombe</i>	SpGI	3.0	2449	2974	510

Figure 1.6.1.

The DFCI Gene Index home page at <http://compbio.dfci.harvard.edu/tgi/tgipage.html> has links to the 114 species-specific databases currently available. Other resources available include the Eukaryotic Gene Ortholog (EGO) database, the RESOURCERER utility for annotating and cross-referencing mammalian microarray resources, and maps of the TCs to completed genome sequences.

Figure 1.6.2.
The home page for the Maize Gene Index.

Computational Biology and Functional Genomics Laboratory
The Gene Index Project

Home GCOD Resource Gene Index Genomic Maps EGO

Select a program and database to search against:

Program: Database:

Alignments:

Upload a file containing a sequence OR paste it into the textbox:
(Note: If both are entered, the file will be ignored.)

Enter the name of the file containing a sequence in FASTA or raw format:

Enter your sequence in FASTA or raw format:

Sequence identifier (for definition line if raw sequence entered):

Options:

Matrix: Filter: Expect:

Cutoff: Strand: Description:

Wordlength (for blastn only): Echofilter: ☐

Graphical Overview: ☒ Ignore Hypotheticals: ☐

The email option allows the user to be notified by email when the blast search is completed. The email will contain a link to the results. Use this option when running a lengthy search. To activate this option, check the box to the left and fill in an email address in the box to the right.

☐ Send a link to results to email address:

Figure 1.6.3.

The BLAST search page allows users to query any of the DFCI Gene Index databases, as well as the EGO and RESOURCERER databases, using protein or DNA sequences.

Computational Biology and Functional Genomics Laboratory
The Gene Index Project

Home GCOB Resources Gene Index Genome Maps EGO

ZmGI Reports

Search TCs
TC#: e.g. TC373709

Search ETs
NP#: e.g. NP072361
GB#: e.g. AY103599

Search ESTs
EST ID: e.g. 949077C01.y2
GB#: e.g. B075039
Clone Name: e.g. cml00481

Search Keyword(s) in annotations
Keyword: e.g. histone

Batch Search
Identifier: ☐ GB# ☐ Clone name ☐ EST ID ☐ TC# ☐ NP#
upload file: email:

Search Maize

BLAST

Identifiers or keywords

TC Annotator

EST Annotator

Literatures

CAT# download

Alternative Splice Forms

EST Expression

Gene Ontology

Metabolic Pathways

Oligomer Prediction

Figure 1.6.4.

The main search page for the Maize Gene Index allows users to search the database using a variety of accession numbers, including DFCI TC number, a Transcript Identifier, GenBank Accessions, and clone identifiers.

Figure 1.6.5. Gene Ontology (GO) terms and Enzyme Commission (EC) identifiers are assigned to the TCs to provide functional annotation and to provide links to metabolic pathway databases.

Figure 1.6.6.
The GO browser shows the hierarchy of functional assignments for TCs identified as members of a particular functional class.

Figure 1.6.7.

For humans, mouse, and rat, TCs are mapped to their respective genomes using the available radiation hybrid maps.

Figure 1.6.8.
RH Mapping Data. A snippet of Mouse TCs containing markers mapped to chromosome 1.

Computational Biology and Functional Genomics Laboratory
The Gene Index Project

Home GCOB Resources Gene Index Genomic Maps EGO

The Maize Gene Index (ZmGI)

Gene Expression Data

Search for Library Specific Assemblies

This query allows you to search for assemblies that are expressed exclusively or predominantly in libraries related to a specific tissue or organ. For example, to find assemblies which have 50% or more ESTs from libraries derived from *Maize root* cDNA library, enter "root" into the Anatomy tag box (without quotes) and 50 into the Percent box. To find predominantly expressed transcripts, you may enter a higher ratio, such as 90. Currently, only a single term can be searched at a time.

Anatomy tag (tissue/organ): Percent: Search

Entries displayed per page:

Search cDNA Libraries by Keyword

All text associated with the represented cDNA libraries can be searched. This search engine uses a flexible name search syntax based on Boolean operations. For example, to find all libraries from root, type "root".

Enter search keywords: Submit

Search cDNA Libraries by Library Identifier

If you know the library identifier, also referred to as catalog number or cat#, you may enter it here to retrieve a Library Report. The Library Report returns detailed information about the library including the total number of ESTs sequenced from the library and the number of TCs containing ESTs from the library.

Cat#: Submit

Scan a List of TC by Library Expression

To scan a list of TCs by library expression data click on the link below. The R Statistics calculated is based on libraries selected by the user. R > 9 is shown. Comparison can be made by choosing tissues or libraries in selection boxes A and B.

[Scan List of TC by Library Expression Data](#)

R Statistics

Stokel, G., and Falciani (2000)
The comparison of gene expression from multiple cDNA libraries.
Genome Research 10:2055-2061.

Figure 1.6.9.

The expression summary page allows each Gene Index database to be explored using information on the libraries from which the ESTs were derived.

Computational Biology and Functional Genomics Laboratory
The Gene Index Project

Home GCOB Resources Gene Index Genomic Maps EGO

The Zea mays Gene Index (ZmGI)

ZmGI Library Expression Search:

Please select a time type and/or library for expression information.
For comparison, select both A and B.

Library Description

A:

- aerial, root, whole plant
- aleurone layer
- anthers
- apex
- bms tissue

B:

- aerial, root, whole plant
- aleurone layer
- anthers
- apex
- bms tissue

OR

- #74C
- #74D
- #8GP
- #8MJ
- #8MM

Only display libraries with more than 50 EST sequences.

[Get Expression](#) [Reset](#)

Comments and suggestions : [Contact Us](#)

Figure 1.6.10.

The Expression Search page allows the frequency of ESTs from various libraries to be compared in order to identify differentially expressed genes based on the sources of libraries from which the ESTs were derived.

Figure 1.6.11.
An example of a library-based expression comparison. The relative abundance of ESTs is depicted using a hot/cold (red/blue) color map and significant differences between classes of ESTs are denoted by the associated *R* statistic (Stekel et al., 2000). For the color version of this figure go to <http://www.currentprotocols.com/protocol/bi0106>.

Figure 1.6.12.
Gbrowse. ESTs from the various plant Gene Index databases are aligned to the *Arabidopsis thaliana* genome sequence.

Figure 1.6.13.
The home page for the Eukaryotic Gene Ortholog (EGO) database.

Figure 1.6.14. A TOG alignment from the EGO database showing alignments of a possible transcription factor from *A. Salmon*, *C.posadaii*, cattle, dog, *Medicago*, oilseed rape, and Trout. **(A)** Shows a table with all TC components of the group and their putative function. The next table shows the blast results. **(B)** Shows a snippet of the sequence alignments.

Figure 1.6.15.

The RESOURCERER home page allows users to select a variety of widely used microarray resources for human, mouse, and rat for annotation or cross-platform and cross-species comparisons. Users can also enter their own microarray platform for annotation by providing GenBank accession numbers.

Figure 1.6.16. Annotation for the Affymetrix HG U95Av2 provided by RESOURCER includes Affymetrix Probe IDs, Clone names (when available), GenBank accessions, UniGene identifiers, DFCI TC numbers for human identified though EGO, GO terms, and annotated function, Physical map location based on alignments of the DFCI THCs, with links to the appropriate databases.

Figure 1.6.17. RESOURCERER also allows microarray platforms to be compared. Here, annotations for Affymetrix HG U95Av2 and HG U95C human GeneChip are compared through EGO. Only elements in common to both datasets are shown (intersection). The annotation includes Affymetrix Probe IDs, Clone names when available, GenBank IDs with links to NCBI, the TGI TC numbers for Human (THCs).

B

#	Src	EST Id	GB#	Clone/Protein	left	right	Library/Protein name
1	100A	AEM-F29_A11	DW714525		1	808	Aedes aegypti adult female salivary gland
2a	TIGR	NABN020TRB	DV294033	NABN020	38	711	Aedes aegypti infected with Brugia malayi
2b	TIGR	NABN020TF	DV294029	NABN020	92	797	Aedes aegypti infected with Brugia malayi
2c	TIGR	NABN020TR	DV294031	NABN020	103	706	Aedes aegypti infected with Brugia malayi
3a	TIGR	NABQ052TRB	DV314940	NABQ052	38	774	Aedes aegypti infected with Brugia malayi
3b	TIGR	NABQ052TF	DV314939	NABQ052	305	611	Aedes aegypti infected with Brugia malayi
4a	TIGR	NAAGJ21TF	DV283134	NAAGJ25	39	772	Aedes aegypti - Fat Bodies Normalized
4b	TIGR	NAAGJ25TR	DV283135	NAAGJ25	39	693	Aedes aegypti - Fat Bodies Normalized
5	TIGR	NABRD21TF	DV320754	NABRD21	39	797	Aedes aegypti infected with Plasmodium falciparum
6a	TIGR	NACND34TRB	DV377096	NACND34	39	759	Aedes aegypti infected with Plasmodium falciparum
6b	TIGR	NACND34TF	DV377095	NACND34	300	772	Aedes aegypti infected with Plasmodium falciparum
7a	TIGR	NAAIE59TR	DV285309	NAAIE59	40	698	Aedes aegypti - Fat Bodies Normalized
7b	TIGR	NAAIE59TF	DV285308	NAAIE59	92	667	Aedes aegypti - Fat Bodies Normalized
8a	TIGR	NABQ068TR	DV301182	NABQ068	40	772	Aedes aegypti infected with Brugia malayi
8b	TIGR	NABQ068TF	DV301184	NABQ068	108	656	Aedes aegypti infected with Brugia malayi
9a	TIGR	NABT431TR	DV330720	NABT431	41	772	Aedes aegypti infected with Plasmodium falciparum
9b	TIGR	NABT431TF	DV330693	NABT431	44	797	Aedes aegypti infected with Plasmodium falciparum
10	100A	AEM-PLATE44_H08	DW715910		48	799	Aedes aegypti adult female salivary gland
11	TIGR	AEDA212TO	DV261942	AEDA212	50	764	Aedes re-sequencing
12a	TIGR	NABRX02TO	DV325538	NABRX02	56	737	Aedes aegypti infected with Plasmodium falciparum
12b	TIGR	NABRX02TF	DV325536	NABRX02	305	673	Aedes aegypti infected with Plasmodium falciparum
13a	TIGR	NABNU01TR	DV295739	NABNU01	92	747	Aedes aegypti infected with Brugia malayi
13b	TIGR	NABNU01TR	DV295742	NABNU01	144	346	Aedes aegypti infected with Brugia malayi
14a	TIGR	NABFW17TR	DV299819	NABFW17	92	799	Aedes aegypti infected with Brugia malayi
14b	TIGR	NABFW17TR	DV299821	NABFW17	92	700	Aedes aegypti infected with Brugia malayi
15	TIGR	NABW002TR	DV346010	NABW002	92	797	Aedes aegypti infected with Brugia malayi
16a	TIGR	NAAF087TR	DV271585	NAAF087	140	785	Aedes aegypti - Fat Bodies Normalized
16b	TIGR	NAAF087TF	DV271610	NAAF087	147	772	Aedes aegypti - Fat Bodies Normalized
17a	TIGR	NAAF989TR	DV271283	NAAF989	153	799	Aedes aegypti - Fat Bodies Normalized
17b	TIGR	NAAF989TR	DV271284	NAAF989	160	799	Aedes aegypti - Fat Bodies Normalized
18	100A	AEM-PLATE44_D09	DW715875		397	550	Aedes aegypti adult female salivary gland
19	100C	EST38613	DW222355	AEVR-LS1-059-C	518	785	Larval Stage 1
20	100C	EST08391	DW192121	AEVR-LS1-010-A	519	799	Larval Stage 1
21	100C	EST79838	EG007990	AEMR-LS1-142-C	544	785	Larval Stage 1
22	100C	EST06052	DW189782	AEGR-SG1-053-E	566	797	Salivary Gland 1
23	100C	EST19065	DW203601	AEMR-SG1-030-B	595	784	Salivary Gland 1
24	100C	EST18288	DW202018	AEMR-LS1-061-D	633	797	Larval Stage 1
25	100C	EST31511	DW215253	AEMR-LS1-109-C	633	797	Larval Stage 1

Sequence source codes:

100C Instituto de Quimica Universidade de Sao Paulo, Departamento de Bioquimica
TIGR The Institute for Genomic Research, TIGR
100A National Institute of Allergy and Infectious Diseases, Laboratory of Malaria and Vector Research

TC sequence strand: coding strand

	Evidence for coding strand	Evidence for reverse complement	Inconclusive evidence
ESTs	5' ESTs (used as +)	7 (used as -)	-
	3' ESTs (used as -)	- (used as +)	-
	total	7	31
Expressed transcripts	-	-	-
Predicted genes	-	-	-
Protein hits	-	5	-
Trimmed polyA/T	1	-	-

Expression Summary

Figure 1.6.18.

A sample TC report for *Aedes Aegypti* TC57832. **(A)** At the top of each record is a FASTA-formatted sequence representing the consensus produced by the clustering and assembly process. Immediately following that are predicted open reading frames, a graphical representation of the EST, and gene sequences that comprise the TC. **(B)** Shows a table with links to a variety of resources including GenBank records, source laboratory etc; it also shows a prediction of the coding strand and the evidence used to support the assignment. **(C)** Buttons provide links to expression summaries based on the libraries represented in each TC assembly, SNPs identified in the TC, and predicted 70-mers oligos. Links to the top 5 results of the searches against a protein database, GO term and EC number assignments, and links to Metabolic Pathways in KEGG, are also given.

Figure 1.6.19.

A schematic overview of the Gene Index Assembly process. For each species represented, EST sequences are downloaded from the dbEST database at the NCBI (<http://www.ncbi.nlm.nih.gov/dbEST>). Sequences are cleaned to remove contaminating vector, adapter, mitochondrial, ribosomal, and other sequences wherever possible. Coding sequences (annotated CDS regions) representing genes are parsed from GenBank records. All EST and gene sequences are compared pairwise using megaBLAST and grouped based on shared sequence similarity. Each cluster is then assembled at high stringency to produce Tentative Consensus (TC) sequences, which are annotated by sequence similarity search against a local copy of UNIPROT, and released through the DFCI Web site.

Table 1.6.1

Summary of the Gene Index Databases Mapped to Completed and Draft Genomes

Genome	Gene indices mapped to that genome
Human	HGI, MGI, RGI, BtGI, ScGI
Mouse	HGI, MGI, RGI, BtGI, SsGI, DGI, CeGI, AtGI, ScGI
Rat	HGI, MGI, RGI, BtGI, SsGI, DGI, CeGI, AtGI, ScGI
Fly	DGI, HGI, CeGI, AtGI, ScGI
Worm	DGI, HGI, CeGI, AtGI, ScGI
Mosquito	AgGI, HGI, DGI, CeGI
Fugu	HGI, MGI, RGI, OIGI, XGI, ZGI
<i>Arabidopsis</i>	CGI, AtGI, LGI, StGI, GmGI, MtGI, McGI, OGI, ZnGI, TaGI, SbGI, HvGI
Yeast	ScGI, SpGI, CrGI, NcrGI, AnGI, DGI, HGI, CeGI, AtGI