EST sequencing and usage

Introduction

Expressed sequence tags, or ESTs, are single DNA sequencing reads made from complementary DNA (cDNA) clone libraries constructed from a known tissue source. Sequencing a large number of these clones from such a library allows one to sample the set of expressed genes, or transcripts, in that particular tissue and experimental state, thus providing a snapshot of the tissue's active genes under those defined conditions.

ESTs provide an excellent resource both for novel gene discovery, and for confirmation of in silco gene predictions from genomic DNA sequence. The physical cDNA clones themselves are useful resources in many types of functional biology studies, including gene inhibition experiments, their potential uses being dependent upon the type of cloning vector employed to constuct them.

Biological material

cDNA clones are the biological material from which EST sequences and finished cDNA sequences are produced. A set (or library) of cDNA clones are made from a biological tissue (or pooled tissues) by using routine molecular biological techniques. Cellular mRNA is extracted from the tissue, it is reverse transcribed to produce DNA complementary to the initial mRNA, and its incorporated into a plasmid, with appropriate vectors and linkers. The plamids are then grown up in a suitable host such as *E. coli.*

Ideally the cDNA clones produced contain a single DNA insert derived from a gene transcript (mRNA) that was being expressed by the cell at the time the tissue was harvested. The tissue is generally selected to sample a particular cell type or organ of interest, and the time the tissue is taken is chosen to correspond to a developmental stage, and could be from a 'normal' individual/organism, one in disease state, or one having been subjected to experimental manipulation by: drug, environmental stress, or other means.

Use of EST sequences

Large scale EST sequencing has become a comparatively inexpensive method for novel gene discovery. Prior to elucidating the full genome sequence of an organism, it provides a shortcut to the transcribed portions of the genome, allowing one to rapidly screen for novel genes, based on the various types of bioinformatic sequence searches available.

Consequently millions of human ESTs have been sequenced, as well as many from mouse, zebrafish and other species. However, with the ever-falling cost of automated DNA sequencing, many organisms have now had their whole genome sequenced. ESTs from these organisms remain useful, providing a means to locate the genes within the genome, by sequence similarity search, and also to lend supporting evidence to the genes predicted by other bioinformatic methods, in fact, they play a vital role in decoding and annotating each genome sequence as

it becomes available to the scientific community.

Analysis process and pipeline

Process outline

The est_db analysis process typically consists of:

- Loading cDNA library and sequencing primer information
- Batch-loading of EST sequences
- Public databank sequence submission and tracking
- EST clustering
- Similarity search analysis
- Protein sequence prediction
- Viewing and extracting results

Loading cDNA library and sequencing primer information

At the initiation of a sequencing project, information describing the cDNA libraries being sequenced needs to be loaded into est_db, this includes library names, descriptions, tissue source, developmental stage, etc. The sequencing primers to be used in the project also need to be specified.

Batch-loading of EST sequences

The EST sequences themselves are loaded into the est_db following base-calling and appropriate base-quality and vector clipping. Input is accepted as suitably-formatted FASTA files, and data is validated to confirm the library, clone name and sequencing primers and direction are recognised for each of the sequences.

Public databank sequence submission and tracking

est_db incorporates an automated system for submission of processed EST sequences to public databanks, tracking of the submission process, and subsequent updates to the databanks should they prove necessary. A final quality control (QC) check is generally performed on all EST sequences prior to their potential submission. This takes the form of one or more project-specific BLASTN similarity searches for any vector, adaptor, host, or contaminating organism sequences that might possibly be encountered in the laboratory environment(s) where the cDNA library construction and sequencing took place. Results of the searches are used to mark each cDNA clone stored in the database as clean or contaminated. EST reads from contaminated clones are not subsequently be submitted.

During the process of sequencing of a particular library, it is likely that some clones will be sequenced more than once, potentially duplicating EST reads. This might occur following resequencing of plate, following partial failure of the initial sequencing reads from the plate, or as a consequence of steps carried out to optimise the sequencing reaction conditions, etc. est_db incorporates the facility to remove this redundancy, that might be introduced with every additional batch of ESTs sequenced and loaded from a particular library, ensuring the best set of ESTs (based on length and other controllable criteria) are selected for submission,

and thus subsequently made available to the public.

Using this battery of checks, and other user chosen parameters, the submission system then decides to submit a new entry, perform a sequence update, or withdraw an existing entry for each EST sequence. Tracking of the sequences is by their clone and library names, accession numbers received from the public databanks, and sequence checksum.

EST clustering and assembly

Clustering of EST sequences, followed by their assembly is absolutely necessary to make sense of EST sequence data, removing the considerable redundancy always present in EST libraries, and allowing long transcripts to be reconstructed *in silico*. Rather than (re)implement a system to carry out these tasks, we chose to design the est_db API and relational database schema to store, retrieve, and manipulate EST clustering, and contig assembly results in a generic fashion, independent of their source. Currently est_db directly accepts results from the StackPACK analysis system, which has become a widely accepted package for EST clustering and assembly.

Similarity search

Similarity searches can be performed to help identify individual ESTs, and also consensus EST sequences derived from EST clustering and assembly. Any number of nucleotide and protein sequences databases can be installed and used by est_db in similarity searches with the WuBLAST family of programmes. Typically these databases would be specific to the aims of a particular project, based on species from which the cDNA library was contructed, the search for known or novel sequences, of a particular class, etc. est_db uses a concise format to store and reconstruct the gapped pairwise sequence alignments resulting from such searches, allowing very large numbers of alignments to be stored, should this prove necessary, though a number of options are available to prune the hits stored to those most relevant.

The est_db pipeline itself has features to allow lengthy analysis processes, such as some BLAST searches, to be split over many CPUs or machines, speeding their execution. These automated job creation and management features can reduce the time to completion for a given BLAST search from weeks to hours, and have been tested to more than three hundred machines simultaneously running BLAST analysis against a single est_db and its database server.

Protein sequence prediction

Protein sequences are predicted from consensus EST sequences as part of the analysis process. We have incorporated ESTScan into the est_db system for this purpose.

Viewing and extracting results

All results from the varied analyses are easily accessible through the est_db web interface, with both textual and graphic views being used to the render the

information. A single search form allows one to retrieve information pertaining to a particular cDNA clone, or EST read, as well as view EST clustering information pertaining to it, such as multiple sequence alignments of the cluster it belongs to, etc.

BLAST results can be searched by protein or gene name or description, with the BLAST hits, predicted protein sequence, etc being show together in a graphical panel, hyperlinked to display the gapped-alignments from the original searches. A step by step guide is available on the searching the est_db analysis page showing many of the available views.

Statistics derived from the analysis of a whole library can also be viewed. The total numbers of clones and ESTs sequenced from the library is displayed, with a breakdown available by sequencing direction, etc. A summary of EST clustering results is also included, allowing one to browse resultant clusters by size etc. A full description of the information available is provided on the Viewing cDNA library analysis statistics page

Viewing cDNA library analysis statistics

Introduction

Detailed statistics are available from the analysis of each of the cDNA libraries sequenced and loaded into est_db on the LibraryStatisticsView pages. These can be reached by clicking the library names on the est_db web SearchPage.

Three types of information are provided:

- Total numbers of clones and ESTs
- EST clustering summary
- Analysis of abundant transcripts

Total numbers of clones and ESTs

Counts of the numbers of Clones and ESTs sequenced, broken down by sequencing direction, etc, are the first statistics presented (see Figure 1).

Library name	TA11
Description	X. tropicalis ESTs from TNeu, TGas, TEgg & TTp clone libraries
Clones	160423
5' ESTs	153780
3' ESTs	79372
Clone-paired ESTs	72729
Duplicate reads	0
Total ESTs	233152

Figure 1. Clone and EST counts from the EST sequencing of a cDNA library, in this case over 230,000 *X. tropicalis* ESTs pooled from a number of embryonic, egg and

tadpole tissue sources.

Clones

The number of complementary DNA (cDNA) clones that were successfully sequenced from the library i.e. yielding at least one EST sequence, and potentially more if sequenced from both ends, or duplicate reads were made from clones.

5' ESTs and 3' ESTs

DNA sequencing technology allows one to read from one or both ends of a DNA sequence. Generally the method of cloning mRNA transcripts to produce cDNA sequences, then insert them into vectors to produce cDNA clones, is a directional one. Therefore when the clone is sequenced, one can infer from which end of the cellular mRNA the sequence is derived.

Thus 5' and 3' ESTs are counts of the number of ESTs sequenced from the library, corresponding to the expected 5' and 3' ends, respectively, of the original mRNA transcripts that were used as the library source material.

Clone-paired ESTs

The number of clones sequenced in the library that yielded both 5' and 3' ESTs.

Duplicate reads

The number of duplicate reads (ESTs from the same clone, with the same direction) within the library.

EST clustering summary

Statistics from EST clustering analysis of the library (if performed) are displayed (see Figure 2).

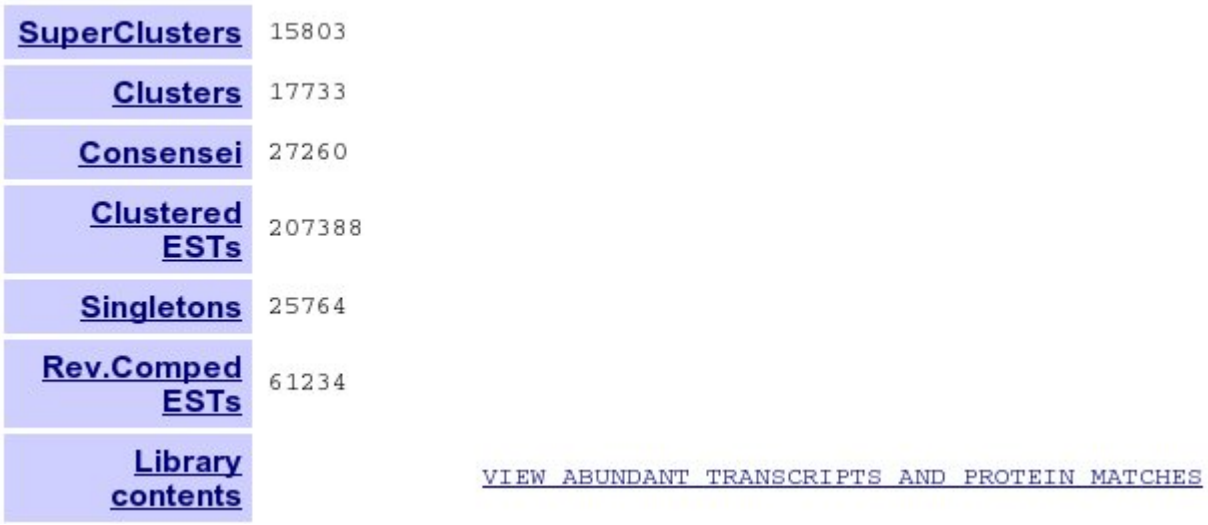

Figure 2. EST clustering summary for the library. Shown are the numbers of SuperClusters, Clusters, ESTs clustered, singletons, etc. Note in this example counts and programmes which produced them (d2_cluster, PHRAP, CRAW) are

components of the StackPACK package. See text for full description.

Superclusters

Superclusters are sequence clusters of similar nucleotide sequence produced by the 'loose' clustering programme d2_cluster. Each supercluster likely represents transcripts derived from the same gene, or several closely related genes. Named: library_S_x Where x is a number starting from 1. Example: TNeu_S_1

Clusters

Each supercluster is further processed by the PHRAP programme, potentially splitting the supercluster to produce one or more clusters. Named: library_C_x_y Where x is the supercluster number, and y is a number starting from 1. Example: TNeu_C_1_1

Consensus sequences (consenseni)

Clusters are analysed for sequence diversity with the CRAW programme. This results in the generation of one or more alignments, each with a consensus sequence. Multiple consensei may suggest alternative-splicing of transcripts. Named: Library_x_y_z, where x, y are the supercluster and cluster numbers from which the consensus is derived and z is number starting from 1. Example: TNeu_1_1_1

Clustered ESTs

The total number of ESTs clustered by d2_cluster, every one thus belonging to a supercluster. EST names are of the form clone_name.sp6 (sp6 being the sequencing primer)

Singletons

The total number of ESTs that were not clustered.

Rev.Comped ESTs

The number of ESTs that analysis suggested they should be included in clusters as the reverse complement of their original sequence. Usually these are 3' sequenced ESTs within the library. Occasionally they result from incorrect direction of insertion of the cDNA into the clone during library construction. Indicated by (R.C.) in alignments.

A breakdown of the results of the clustering analysis, counting how many superclusters and clusters were generated that fall in a certain size range is also given. Hotlinks allow the binned clusters to be reviewed (See Figure 3).

Cluster Sizes

SuperCluster	Count	Cluster	Count
2	5142	2	5740
3	2203	3	2534
4	1522	4	1717
5	1019	5	1145
6	751	6	845
7–8	1084	7–8	1214
9–10	708	9–10	789
11–20	1599	11–20	1809
21–30	671	21–30	734
31–40	334	31–40	365
41–50	218	41–50	233
51–60	104	51–60	116
61–70	75	61–70	89
71–80	70	71–80	74
81–90	49	81–90	49
91–100	26	91–100	45
101–200	157	101–200	164
201–300	35	201–300	35
301–400	12	301–400	12
401–500	1	401–500	10
501–1000	13	501–1000	11
1001+	10	1001+	3

Available Searches

Ensembl mouse (WuBLASTX) Unigene X. laevis (WuBLASTN) Swall (WuBLASTX) TIGR XGI (WuBLASTN) Xenopus sp. mRNA (WuBLASTN) Ensembl human (WuBLASTX)

Figure 3. Supercluster and cluster sizes summarised for the library.

One can notice that are a large number of both clusters and superclusters made up of only 2 ESTs (more than 5000 in each case), but considerably lower numbers of each as the absolute sizes increase, for example there are only 10 superclusters, which are composed of 1000 ESTs or more. Also shown are the available (BLAST) search results for the library.

Analysis of abundant transcripts

One can investigate the abundant transcripts in a particular library, together with their analysis results by following the 'VIEW ABUNDANT TRANSCRIPTS AND PROTEIN MATCHES' link provided on each of the LibraryStatisticsView pages (see Figure 2). These show the consensus sequences, sorted by transcript abundance, as assessed by the clustering analysis.

View results page: 1 2 3 4 5 6 7 8 9 10 11 >>]

Consensus name	TA11_1_1_1
Containing	8797 ESTs
Size rank	1
Best protein matches	No significant hits

Consensus name	TA11_7_300_1
Containing	1209 ESTs
Size rank	2

Best protein matches

Alignment	Ext DB Entry	Description	Score	P value
Q8AVK9	Q8AVK9	Similar to nuclease sensitive element binding prot	279	4.70e-23
P21573	P21573	Nuclease sensitive element binding protein 1 (Y bo	270	4.20e-22
Q8AVY9	Q8AVY9	Similar to nuclease sensitive element binding prot	270	4.20e-22
Q90376	Q90376	Y-box binding protein	264	1.90e-21
Q06066	Q06066	Nuclease sensitive element binding protein 1 (Y bo	260	4.60e-21

Consensus name	TA11_20_176_4
Containing	789 ESTs
Size rank	3

Best protein matches

Alignment	Ext DB Entry	Description	Score	P value
P16878	P16878	Keratin, type II cytoskeletal (XENCK55(5/6))	1481	1.30e-150
Q9I9P5	Q9I9P5	Inner-ear cytokeratin	1387	1.10e-140
O93532	O93532	Keratin, type II cytoskeletal cochleal (Cytokerati	1292	1.40e-130
Q8HZR5	Q8HZR5	Keratin 7	1269	3.90e-128
Q96CL4	Q96CL4	Similar to keratin 6A	1204	4.20e-127

Figure 4. Browsing abundant transcripts. Shown (in this case) are the three most abundant transcripts, as assessed from the EST clustering analysis. If a translated BLAST (WuBLASTX) search of the consensus sequence produced significant protein sequence hits, the five highest scoring ones are displayed. In this example a transcript similar to that for y box binding protein-1 is highly expressed (second most abundant) in this library of *X. tropicalis* ESTs.