EST Clustering

An expressed sequence tag or EST is a short sub-sequence of a transcribed cDNA sequence. They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination. An EST is produced by one-shot sequencing of a cloned mRNA (i.e. sequencing several hundred base pairs from an end of a cDNA clone taken from a cDNA library). The resulting sequence is a relatively low quality fragment whose length is limited by current technology to approximately 500 to 800 nucleotides. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be present in the database as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand.

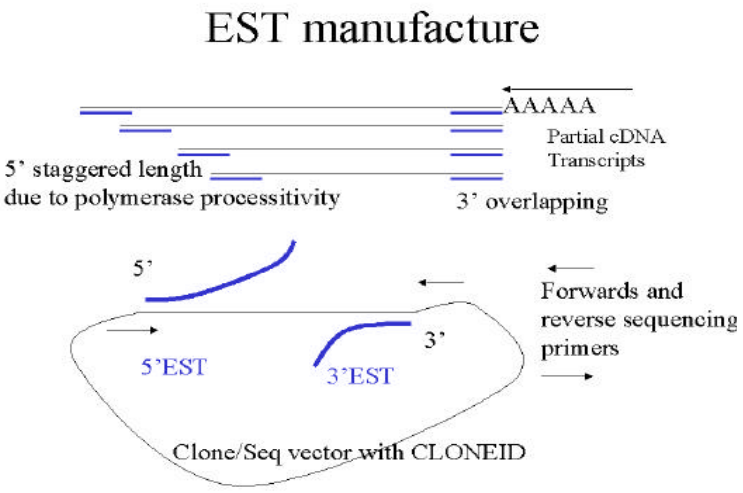


Fig1: Manufacture of EST

Overview of clustering and consensus generation

EST Clustering is performed as a process that utilizes clustering information that is less and less definitive. Initially sequence identity provides a good guide to cluster membership. Shared annotation provides joining information that can be of more variable quality. Thus the number of accurately clustered ESTs is heavily dependent on a strategy that can assign cluster membership based on verifiable criteria; sequence identity is currently the most useful of these. Clustering can be performed with or without sequence consensus generation. It is preferable, although more difficult, to manufacture a consensus sequence from each cluster. The clustering overview will briefly describe processes that result in consensus sequence generation.

What is an EST clustering

A cluster is fragmented, EST data (DNA or protein) and (if known) gene sequence data, consolidated, placed in correct context and indexed by gene such that all expressed data

concerning a single gene is in a single index class, and each index class contains the information for only one gene. The goal of the clustering process is to incorporate overlapping ESTs which tag the same transcript of the same gene in a single cluster. For clustering, we measure the similarity (distance) between any 2 sequences. The distance is then reduced to a simple binary value: accept or reject two sequences in the same cluster.

Similarity can be measured using different algorithms:

- *Pairwise alignment algorithms:*

  (a) Smith-Waterman is the most sensitive, but time consuming (ex. cross-match)
  (b) Heuristic algorithms, as BLAST and FASTA, trade some sensitivity for speed

- *Non-alignment based scoring methods:*

  d2 cluster algorithm: based on word comparison and composition (word identity and multiplicity) (*Burke et al.*, 99). No alignments are performed) fast.

- Pre-indexing methods.
- Purpose-built alignments based clustering methods.

Types of clustering

Loose and stringent clustering

ESTs by their nature have a degree of erroneous sequence data, complicated by short length and some mis-annotation. Stringent one-pass assembly methods tend to result in fewer, shorter consensus sequences. Looser systems for clustering result in larger, more 'sloppy' clusters, with various expressed forms being represented within each cluster. Each approach has its advantages and disadvantages. Stringent clustering provides greater initial fidelity, at a cost of lower coverage of expressed gene data and a lower inclusion rate of expressed gene forms. Loose clustering provides greater coverage, at a cost of possible inclusion of paralogous expressed genes, lower fidelity data, but at a gain of greater inclusion of alternate expressed forms.

(a) *Stringent clustering*:

- Greater initial fidelity
- One pass
- Lower coverage of expressed gene data
- Lower cluster inclusion of expressed gene forms
- Shorter consensi

(b) *Loose clustering*

- Lower initial fidelity
- Multi-pass
- Greater coverage of expressed gene data

- Greater cluster inclusion of alternate expressed forms
- Longer consensi
- Risk to include paralogs in the same gene index

Supervised and unsupervised EST clustering

- *Supervised clustering*

  ESTs are classified with respect to known reference sequences or "seeds" (full length mRNAs, exon constructs from genomic sequences, previously assembled EST cluster consensus).

- *Unsupervised clustering*

  ESTs are classified without any prior knowledge

The three major gene indices use different EST clustering methods:

- *TIGR Gene Index* uses a stringent and supervised clustering method, which generate shorter consensus sequences and separate splice variants.
- *STACK* uses a loose and unsupervised clustering method, producing longer consensus sequences and including splice variants in the same index.
- A combination of supervised and unsupervised methods with variable levels of stringency is used in *UniGene.* No consensus sequences are produced.

Importance for ESTs:

- ESTs represent the most extensive available survey of the transcribed portion of genomes.
- ESTs are indispensable for gene structure prediction, gene discovery and genomic mapping.
- Characterization of splice variants and alternative polyadenylation.
- *In silico* differential display and gene expression studies (specific tissue expression, normal/disease states).
- SNP data mining.
- High-volume and high-throughput data production at low cost.

Low data quality of ESTs:

- High error rates (_ 1=100) because of the sequence reading single-pass.
- Sequence compression and frame-shift errors due to the sequence reading single-pass.
- A single EST represents only a partial gene sequence.
- Not a defined gene/protein product.
- Not curated in a highly annotated form.
- High redundancy in the data) huge number of sequences to analyze.

Improving ESTs: Clustering, Assembling and Gene indices:

The value of ESTs is greatly enhanced by clustering and assembling. It can solve many problems associated with ESTs

- solving redundancy can help to correct errors
- longer and better annotated sequences
- easier association to mRNAs and proteins
- detection of splice variants
- fewer sequences to analyze

Gene indices: All expressed sequences (as ESTs) concerning a single gene are grouped in a single index class, and each index class contains the information for only one gene.

Different clustering/assembly procedures have been proposed with associated resulting databases (gene indices):

- UniGene (http://www.ncbi.nlm.nih.gov/UniGene)
- TIGR Gene Indices (http://www.tigr.org/tdb/tgi.shtml)
- STACK (http://www.sambi.ac.za/Dbases.html)

UniGene (http://www.ncbi.nlm.nih.gov/UniGene)

UniGene Gene Indices available for a number of organisms. UniGene clusters are produced with a supervised procedure: ESTs are clustered using GenBank CDSs and mRNAs data as "seed" sequences. There is no attempts to produce contigs or consensus sequences. UniGene uses pairwise sequence comparison at various levels of stringency to group related sequences, placing closely related and alternatively spliced transcripts into one cluster.

UniGene procedure:

(1) Screen for contaminants, repeats, and low-complexity regions in Embank:

(a) Low-complexity are detected using Dust.
(b) Contaminants (vector, linker, bacterial, mitochondrial, ribosomal sequences) are detected using pairwise alignment programs.
(c) Repeat masking of repeated regions (RepeatMasker).
(d) Only sequences with at least 100 informative bases are accepted.

Clustering procedure:

(a) Build clusters of genes and mRNAs (GenBank).
(b) Add ESTs to previous clusters (megablast).
(c) ESTs that join two clusters of genes/mRNAs are discarded.
(d) Any resulting cluster without a polyadenilation signal or at least two 3' ESTs is discarded.

(e) The resulting clusters are called anchored clusters since their 3' end is supposed known.
(f) Ensures 5' and 3' ESTs from the same cDNA clone belongs to the same cluster.
(g) ESTs that have not been clustered, are reprocessed with lower level of stringency. ESTs added during this step are called guest members.
(f) Clusters of size 1 (containing a single sequence) are compared against the rest of the clusters with a lower level of stringency and merged with the cluster containing the most similar sequence.
(j) For each build of the database, clusters IDs change if clusters are split or merged.

TIGR Gene Indices (http://www.tigr.org/tdb/tgi)

(a) TIGR produces Gene Indices for a number of organisms
(b) TIGR Gene Indices are produced using strict supervised clustering methods.
(c) Clusters are assembled in consensus sequences, called *tentative consensus* (TC) sequences,that represent the underlying mRNA transcripts.
(d) The TIGR Gene Indices building method tightly groups highly related sequences and discard under-represented, divergent, or noisy sequences.
(e) TC sequences can be used for genome annotation, genome mapping, and identification of orthologs/paralogs genes.
(f) TIGR Gene Indices characteristics:

- separate closely related genes into distinct consensus sequences
- separate splice variants into separate clusters
- low level of contamination

TIGR Gene Indices procedure:

(a) EST sequences recovered form dbEST (http://www.ncbi.nlm.nih.gov/dbEST)
(b) Sequences are trimmed to remove:
- Vectors
- polyA/T tails
- adaptor sequences
- bacterial sequences

(c) Get Tentative consensus and singletons from previous database build
(d) Supervised and strict clustering:
- Use ETs, TCs, and CDSs as template;
- Compare cleaned ESTs to the template using FLAST (a rapid pairwise comparison program).
- Sequences are grouped in the same cluster if both conditions are true:
  (a) they share _ 95% identity over 40 bases or longer regions
  (b) < 20 bases of mismatch at either end

(e) Each cluster is assembled using CAP3 assembling program to produce tentative consensus (TC) sequences.

- CAP3 can generate multiple consensus sequences for each cluster
- CAP3 rejects chimeric, low-quality and non-overlapping sequences.
- New TCs resulting from the joining or splitting of previous TCs, get a new TC ID.

(f) Builded TCs are loaded in the TIGR Gene Indices database and annotated using information from GenBank and/or protein homology.

(g) Track of the old TC IDs is maintained through a relational database.

STACK

Based on "loose" unsupervised clustering, followed by strict assembly procedure and analysis to identify and characterize sequence divergence (alternative splicing, etc). The "loose" clustering approach, *d2 cluster,* is not based on alignments, but performs comparisons via non-contextual assessment of the composition and multiplicity of words within each sequence. Because of the "loose clustering it produces longer consensus sequences than the TIGR Gene Indices. It also introduces ~30% more sequences than the UniGene, due to the "loose" clustering approach.

STACK Procedures

(a) *Sub-partioning*

- Select human ESTs from Embank
- Sequences are grouped in tissue-based categories ("bins") which will allow further specific tissue transcription exploaration.
- A "bin" is also created for sequences derived from disease-related tissues.

(b) *Masking*

Sequences are masked for repeats and contamination using cross-match.

- Human repeat sequences (RepBase)
- Vector sequences
- Ribosomal and Mitochondrial DNA, other contaminants.

(c) *"Loose" clustering using d2 cluster.*

- The algorithm looks for the co-occurrence of n-length words (n = 6) in a window of size 150 bases having at least 96% identity.
- Sequences shorter than 50 bases are excluded from the clustering process.
- Clusters highly related sequences.
- Clusters also sequences related by rearrangements or alternative splicing.
- Because d2 cluster weights sequences according to their information content, masking of low complexity regions is not required.

(d) *Assembly*

- The assembly step is performed using Phrap.
- STACK don't use quality information available from chromatograms.
- The lack of trace information is largely compensated by the redundancy of the ESTs data.
- Sequences that cannot be aligned with Phrap are extracted from the clusters (singletons) and processed later.

(e) Alignment analysis.

- The CRAW program is used in the first part of the alignment analysis.
- CRAW generates consensus sequence with maximized length.
- CRAW partitions a cluster in sub-ensembles if _ 50% of a 100 bases window differ from the rest of the sequences of the cluster.
- Rank the sub-ensembles according to the number of assigned sequences and number of called bases for each sub-ensemble (CONTIGPROC).
- Annotate polymorphic regions and alternative splicing.

(f) Linking.

- Joins clusters containing ESTs with shared clone ID.
- Add singletons produced by Phrap in respect to their clone ID.

(g) STACK update.

- New ESTs are searched against existing consensus and singletons using cross-match.
- Matching sequences are added to extend existing clusters and consensus.
- Non-matching sequences are processed using d2 cluster against the entire database and the new produces clusters are renamed)Gene Index ID change.

(h) STACK outputs.

- Primary consensus for each cluster in FASTA format.
- Alignments from Phrap in GDE (Genetic Data Environment) format.
- Sequence variations and sub-consensus (from CRAW processing).

Data source

The data sources for clustering can be in-house, proprietary, public database or a hybrid of this (chromatograms and/or sequence files).

Each EST must have the following information:

- A sequence ID (ex. sequence-run ID);
- Location in respect of the poly A (3' or 5');
- The CLONE ID from which the EST has been generated;

- Organism;
- Tissue and/or conditions;
- The sequence.

The steps suggested in EST clustering are as follows:

(1) Pre-processing

Sequences are masked for repeats and vector, and formatted for the clustering engine. Sequence quality is often assessed at this step. A minimum number of residues are accepted above a known quality threshold. All masked sequence data is accepted for clustering above 50bp in length.

EST pre-processing consists in a number of essential steps to minimize the chance to cluster unrelated sequences.

- Screening out low quality regions:
  - Low quality sequence readings are error prone.
  - Programs as Phred (*Ewig et al., 98*) read chromatograms and assesses a quality value to each nucleotide.
- Screening out contaminations.
- Screening out vector sequences (vector clipping).
- Screening out repeat sequences (repeats masking).
- Screening out low complexity sequences.

(2) Initial clustering

An initial clustering is performed based on a measure of high sequence identity.

(3) Assembly

Assembly is either part of the initial clustering (as used in TIGR_ASSEMBLER) or separated into clustering followed by assembly performed by a specialist assembly package such as PHRAP or CAP2 / 3[2]

(4) Alignment processing

Aligned clusters, particularly those generated by a loose clustering engine, need to be processed for errors and alternate forms of expressed sequences. Consensus generation may be a result of this step (as in STACK), or a consensus can be accepted directly from the assembly step.

(5) Cluster joining

Once clustered, clusters and/or cluster consensi can be further joined by available annotative approaches.

(a) Clone joining

The most powerful cluster joining method is clone-joining, which utilizes the physically shared clone id between 3' and 5' EST fragments sequenced from the same starting clone. A

different approach would be to link clone-related sequences in the pre-processing phase, but this may increase errors and processing time requirements at the clustering and assembly stages. By either approach, linking by clone annotation is an error-prone step in the EST consolidation process as it relies entirely on the accuracy of the sequence annotation and the uniqueness of clone IDs if data from disparate sources is to be used.

(b) Available parents

If a parent mRNA sequence is available (non-EST) it can be used to physically link EST cluster(s) via sequence comparison.

(6) Output

Tools Available for different stages of EST data analysis

- *Pre-processing:*
  Lucy, SeqClean, SeqTrim, Phred/Cross_match, RepMask, UniVec, RepeatMasker
- *Clustering:*
  CD-hit, d2_cluster
- *Assembly:*
  CAP3, GAP4, Phrap, Fak2
- *Mapping:*
  KOBAS, CGMAP, SIM

Resources for EST data

- dbEST at NCBI
  http://www.ncbi.nlm.nih.gov/dbEST/
- The TIGR Gene Indices
  http://www.tigr.org/tdb/tgi/
- UniGene database at NCBI
  www.ncbi.nlm.nih.gov/UniGene
- Plant Gene Research, Kazusa DNA Research Institute
  http://www.kazusa.or.jp/en/plant/database.html

There are number of online software is available for EST clustering:

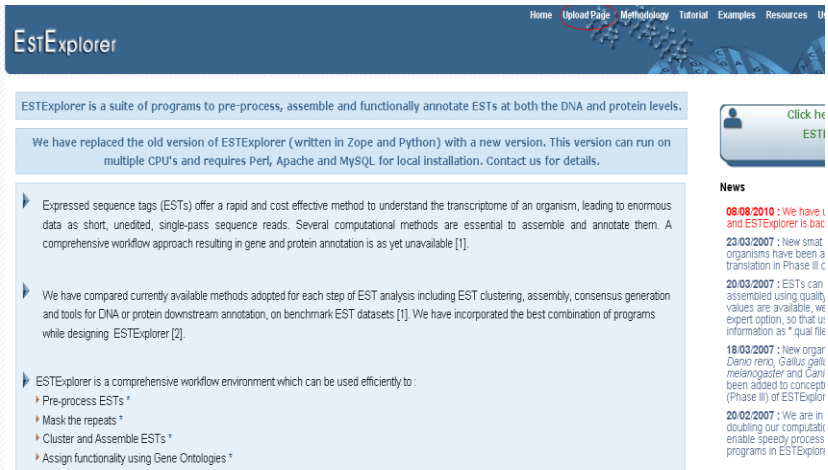
Software/Pipelines for EST Analysis

(a) *ESTExplorer:* It is a comprehensive workflow system for EST data management and analysis. The pipeline uses a 'distributed control approach' in which the most appropriate bioinformatics tools are implemented over different dedicated processors. Species-specific repeat masking and conceptual translation are in-built. ESTExplorer

accepts a set of ESTs in FASTA format which can be analysed using programs selected by the user. After pre-processing and assembly, the dataset is annotated at the nucleotide and protein levels, following conceptual translation. Users may optionally provide ESTExplorer with assembled contigs for annotation purposes. Functionally annotated contigs/ESTs can be analysed individually. The overall outputs are gene ontologies, protein functional identifications in terms of mapping to protein domains and metabolic pathways. ESTExplorer has been applied successfully to annotate large EST datasets from parasitic nematodes and to identify novel genes as potential targets for parasite intervention. ESTExplorer runs on a Linux cluster and is freely available for the academic community at http://137.111.41.149/index.php

ESTExplorer is a comprehensive workflow environment which can be used efficiently to:

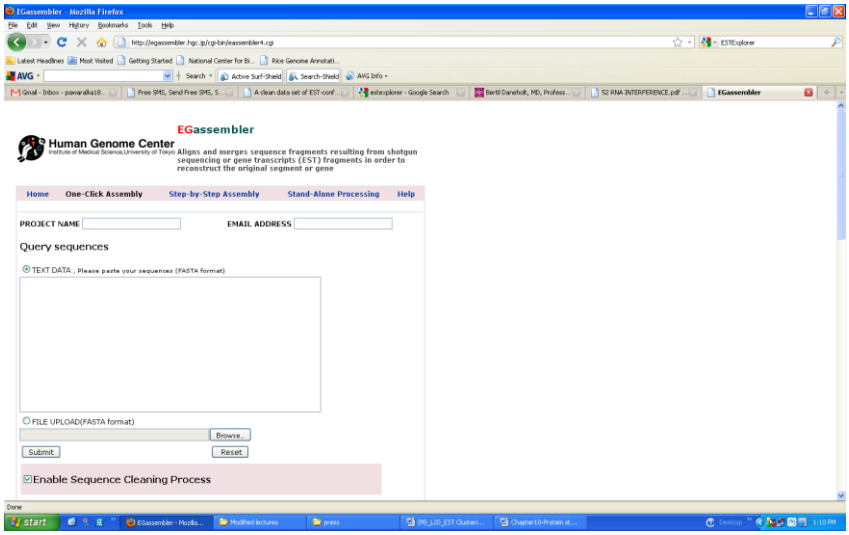
- Pre-process ESTs
- Mask the repeats
- Cluster and Assemble ESTs
- Assign functionality using Gene Ontologies
- Conceptually translate ESTs into peptides
- Mapping to protein domains and metabolic pathways



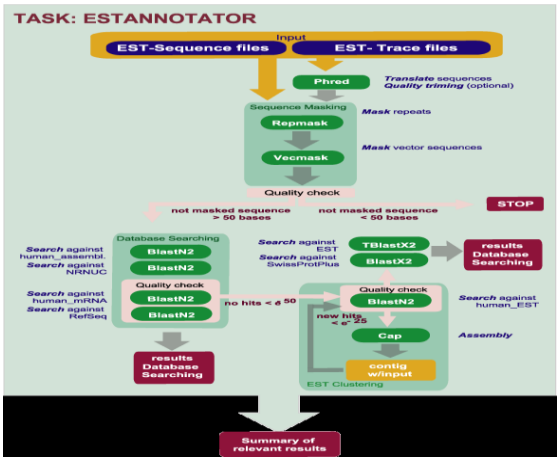
(b) *EGAssembler:* EGassembler is an online service, which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering and assembly of ESTs and genomic fragments. EGassembler consists of a pipeline of the following five components, each using highly reliable open-source tools and a non-redundant custom-made database of vectors and repeats covering almost all publicly available vectors and repeats databases.

EGassembler web interface has three sub-menus, each targeted for different users.

1. One-Click Assembly
2. Step-by-Step Assembly
3. Stand-Alone Processing



(c) *ESTAnnotator:* It is a tool for automatically analysis of EST sequences supporting the search of functional annotations of novel transcript sequences. n a first quality check step repeats, vector parts and low quality sequences are masked. Then successive steps of BLAST searching against suitable databases and EST clustering are performed. Already known transcripts present within mRNA and genomic DNA reference databases are identified. Subsequently, tools for the clustering of anonymous ESTs and for further database searches at the protein level are executed. ESTAnnotator was successfully applied for the systematic identification and characterization of novel human genes involved in cartilage/bone formation, growth, differentiation and homeostasis. The server is available at
http://genome.dkfz-heidelberg.de/menu/biounit/examples/estannotator/estannotator_result.html



(d) *ESTAP:* ESTAP (EST Analysis Pipeline) is a series of automated procedures that verify, store and analyze EST (Expressed Sequence Tag) data generated using high-throughput platforms. The Web-accessible ESTAP software provides EST analysis support to a distributed group of researchers working on a variety of target organisms.
ESTAP automatically cleanses raw sequence data by removing vector, low quality, and contaminating sequences. The cleansed sequences are compared to a set of DNA or

protein databases using the BLAST algorithms. ESTAP also clusters and assembles EST collections into singlet and contig (redundant) datasets using d2_cluster and CAP3.

The function of the singlets and contigs can be automatically annotated using the InterProScan program from the European Bioinformatics Institute (EBI). The raw and cleansed data and analysis results are stored in a relational database. ESTAP affords easy viewing of the original data, the cleansed data, and the analysis results via a Web browser. It also allows the data owner to automatically prepare and submit selected sequences to dbEST. Beyond the scope of EST projects, ESTAP is able to handle sequences from small-scale genomic DNA libraries. It can clean, BLAST and assemble the genomic DNA sequences. The server is available at http://staff.vbi.vt.edu/estap/

References and suggested readings:

Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis.

Nagaraj SH, Deshpande N, Gasser RB, Ranganathan S. (*2007).* ESTExplorer: an automated assembly and annotation platform to analyse expressed sequence tags (ESTs).*Nucleic Acids Res. Jul;35*

EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Res.* **34**:W459-462.

Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S. (2003). ESTAnnotator: A tool for high throughput EST annotation. *Nucleic Acids Res.* Jul **1;31(13):**3716-9.

Mao C, Cushman JC, May GD, Weller JW (2003). ESTAP - an automated system for the analysis of EST data. *Bioinformatics* **19:**1720-1722

Win Hide, Rob Miller, Andrey Ptitsyn, Janet Kelso, Chellapa Gopallakrishnan and Alan Christoffels (1999). NSANBI: EST Clustering Tutorial.