Transcripts Databases

Unigene (NCBI)

Website: http://www.ncbi.nlm.nih.gov/UniGene

Production of ESTs Clusters

Preliminary steps:

- Elimination of sequences of foreign origin (E. coli)
- Cleaning from cloning vector or artificial primers or linkers
- rRNA and mitochondrial sequences are eliminated
- Through the Trace Archive, base-level error probabilities are used to identify the highest quality segment of each sequence
- Repetitive sequences: simple repeats (low-complexity regions) are idendified with the alghorithm DUST and transposable repetitive elements are identified by comparison with a library of known repeats for each organism. These sequences are "soft-masked" in the sense that they are not used to initiate a sequence alignment
- For a sequence to be included in UniGene, the clone insert must have at least 100 base pairs that are of high quality and not repetitive

To be classified as coming from a single gene sequences must overlap and should form a near-perfect sequence-alignment:

- Alignment parameters are chosen by examining ratios of true to false clusterings in curated test sets.
- The resulting clusters may contain more than one alternative splice form.

UniGene clusters must be anchored at the 3' end of a transcription unit. This evidence can be either

- a polyadenilation signal
- or the presence of a polyA tail
- or the presence of at least two ESTs generated using 3' sequencing primer.

Alternative spliced terminal 3' exons will appear as disincts clusters until sequence that spans the distinct splice forms is submitted.

TIGR Gene Index

Website: http://compbio.dfci.harvard.edu/tgi/

Production of TIGR TCs

Similarly to what appens during UniGenes construction, sequences are cleaned from bacterial sequences, vector sequences, adaptors, polyA tails.

- ESTs are clustered using WU-BLAST and sequences with >= 95% identity over regions of at least 40 bp in length are collected
- The sequences comprising each cluster are assembled using the assembling software CAP3

Assembling has several advantages over simple clustering used by NCBI:

- It separates closely related genes into distinct consensus sequences (Tested on family members with high sequence homology).
- It separates splice variants (while for UniGene only alternative spliced terminal 3' exons will appear as disincts clusters).
- It produces longer representations of the underlying gene sequences.

During assembly error rates are modeled considering that lower sequence quality corresponding to starting and ending bases of the sequence.

Version history

Version	Tracking of changes	Name	Date
0.1	Initial version	Alberto	March 31, 2007