DOE Joint Genome Institute EST and cDNA Program

Erika Lindquist, Peter Brokstein and Paul Richardson

The US Department of Energy Joint Genome Institute has recently established an EST and cDNA sequencing and Analysis Program. The purpose of the program is to generate high quality cDNA libraries, optimize EST sequencing, present a comprehensive view of the sequence data, assist in genome assembly and annotation and in some cases full-length sequence cDNA clones.

Since its inception in mid 2003, the program has grown from processing 4 species to 27 species in a total of ~3 million sequences. Many of the organisms sequenced had few or no ESTs publicly available. JGI will establish EST projects to support DOE and JGI missions. EST projects have been and will be initiated for most large and small Eukaryotic genome projects both to support genome assembly and annotation as well as to provide transcript data for the scientific community. Additionally, EST projects include sequencing which does not include a corresponding genomic sequence to provide the scientific community with transcript data. All EST sequences will be submitted to Genbank.

It is preferred that cDNA libraries be created at the JGI although processing is also done on collaborator generated libraries. The quality and diversity of the library directly affects the decision regarding depth of sequencing for each library. Library quality is determined at various levels, both by PCR and analysis of the end sequence reads.

The EST Sequence Processing Pipeline is a computational tool to analyze, report, and display intra and inter library quality based on individual reads, clustering, assemblies and annotation. Summary reports and output files are generated through a web interface based on input data and user definable processing parameters.

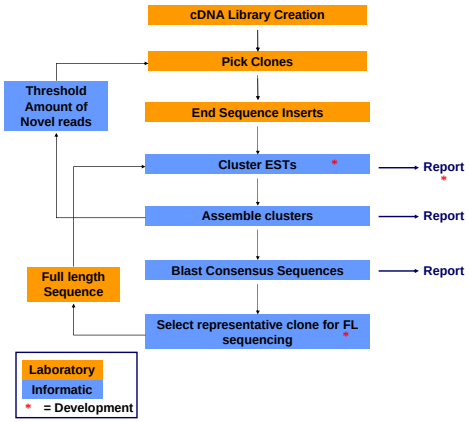
The importance of EST processing relates to the nature of EST sequences. ESTs are single pass often redundant sequences whose value is increased by clustering and assembly. ESTs represent a selective view of the transcribed portions of the genome therefore thus their usefulness for genome annotation, for validation of ab initio gene predictions, extending predicted genes (untranslated region), and for gene predictor training (translational start, intron-exon boundaries).

One organism of particular interest in the JGI collaboration with the Environmental Protection Agency (EPA), is the species Pimephales promelas commonly known as fathead minnow. The aim of the collaboration is to identify novel transcripts to assist in development of expression analysis tools, specifically microarrays. The ability to generate molecular data for Pimephales promelas in response to environmental exposure will aid in toxicity predictions and chemical risk assessments.

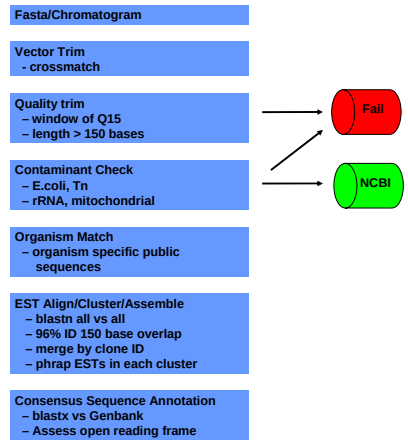
A sampling of DOE EST Projects

Alvinella pompejana	90,720	**Naegleria gruberi**	34,560
Artemisia annua	81,696	Nectria haematococca MPV1	41,472
Aspergillus niger	35,328	**Nematostella vectensis**	177,600
Branchiostoma floridae	97,536	Petromyzon marinus	38,400
Capitella sp.I	156,864	Physcomitrella patens subsp. patens	36,864
Chlamydomonas reinhardii	45,312	Pichia stipitis CBS6054	22,272
Compositae	302,880	Pimephales promelas	304,416
Daphnia pulex	101,376	Postia placenta Mad-698-R	44,544
Emiliania huxleyi CCMP1516	109,824	Reniera sp.	83,040
Glomus intraradices	23,424	Selaginella moellendorffii	77,568
Helobdella robusta	94,080	Spironucleus vortens	29,184
Karenia brevis	41,472	Sporobolomyces roseus	29,952
Laccaria bicolor	40,704	Thalassiosira pseudonana	57,216
Lottia gigantea	166,656	Trichoderma reesei	92,928
Micromonas pusilla NOUM17	36,864	Trichoplax adhaerens Red Sea Grell	31,872
		Xenopus tropicalis	714,432

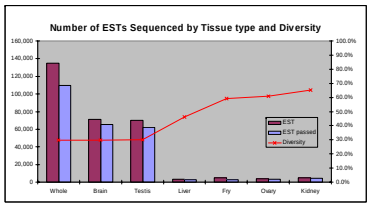
EST Program Schematic



EST Sequence Processing Pipeline

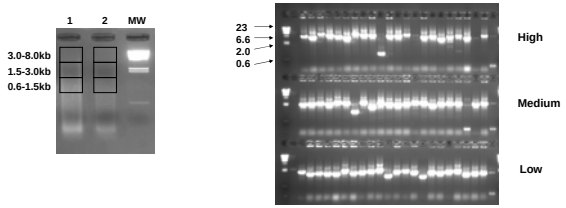


Pimephales promelas cDNA Libraries



Diversity (#clusters/#clones *100) decreases with increased number of ESTs sequenced.

cDNA Library Generation and PCR QC



- Isolate mRNA from Total RNA
- Convert mRNA to cDNA using reverse transcriptase [RT]
- Generate two or three size selected cDNA libraries
- Size select cDNA insert ranges using gel separation (left gel photo)
  - High ~3.0-8.0kb
  - Medium ~1.5-3.0kb
  - Low ~0.6-1.5kb
- Clone inserts into vector
- Determine insertless rate and size range of cDNA libraries
  - PCR and gel electrophoresis (right gel photo, provided by Dean Ng)
- Sequence both ends of insert to yield 5' and 3' Expressed Sequence Tags [ESTs]

Selected Full length Sequencing Results

FL length	Protein query hit within 3 bases of start
4227	Top2a protein [Danio rerio]
4143	PREDICTED: similar to laminin, beta 1 [Danio rerio]
3095	PREDICTED: similar to Cold autoinflammatory syndrome 1 protein (Cryopyrin)
3255	sorting nexin 13 [Takifugu rubripes]
2876	Zswim6 protein [Danio rerio]
3197	hyaluronan-mediated motility receptor [Danio rerio]
3024	PREDICTED: similar to calcium-activated chloride channel [Danio rerio]
2745	PREDICTED: similar to cytosolic phospholipase A2 epsilon [Danio rerio]
2344	muskelin [Danio rerio]
2572	suppressor of fused homolog [Danio rerio]
1757	suppression of tumorigenicity 7 [Danio rerio]
	Protein query hit within 60 bases of start
2991	signal transducer and activator of transcription 1 [Carassius auratus]
2178	novel protein similar to human transcription factor NRF [Danio rerio]
3258	PREDICTED: similar to trans-acting transcription factor 3 isoform 1 isoform 1 [Danio rerio]
3608	cancer susceptibility candidate 3 [Danio rerio]
2002	similar to serologically defined colon cancer antigen 10 [Danio rerio]

-1152 clones were selected for FL sequencing

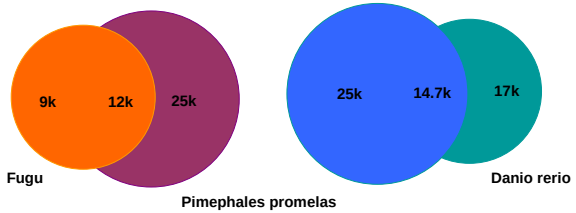
-Average insert length 2305 bases

-Range of insert length 4227 – 597 bases

-220 Clones aligned within 3 bases of the start of protein blast hit

-Full length sequencing done at the Stanford Human Genome Center

Overlap of Pimephales promelas with Danio rerio and Fugu



Blastx results of Pimephales promelas cluster consensus sequences overlapping the peptides contained in databases for both Danio rerio and Fugu downloaded from Ensemble Feb 2006. Blast expect value <e-9.

Correspondence: EALindquist@lbl.gov

LBNL-59258

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.