

The 7th international symposium on earthworm ecology · Cardiff · Wales · 2002

## The earthworm Expressed Sequence Tag project

Stephen R. Stürzenbaum<sup>1\*</sup>, John Parkinson<sup>2</sup>, Mark Blaxter<sup>2</sup>, A. John Morgan<sup>1</sup>, Peter Kille<sup>1</sup>  
and Oleg Georgiev<sup>3</sup>

<sup>1</sup> School of Biosciences, University of Wales, P.O. Box 915, Cardiff CF10 3TL, Wales, UK

<sup>2</sup> Institute of Cell, Animal and Population Biology, University of Edinburgh, UK

<sup>3</sup> Institute of Molecular Biology, Zurich University, Switzerland

Submitted September 6, 2002 · Accepted May 13, 2003

### Summary

This paper aims to provide a brief description of the earthworm Expressed Sequence Tag (EST) project. ESTs are short single read sequences randomly derived from cDNA libraries and provide the means to acquire large scale sequence information of coding DNA. The earthworm EST project is growing rapidly and the analysis of the first 577 sequences corresponded to ~ 400 different genes, with 79 represented by two or more ESTs. Significant sequence similarity to known proteins was observed in 76 % of cases and the remaining 24 % were classified as novel genes. Using a combination of bioinformatic tools the sequence information was used to build a relational database, Lumbribase, which can be queried via an internet interface by sequence similarity and key word searches (see <http://www.earthworms.org>).

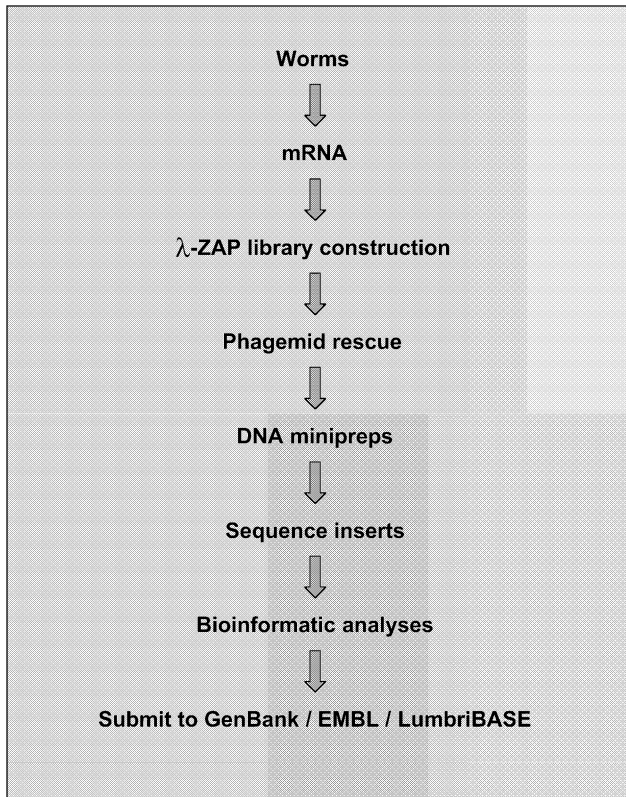
**Key words:** Expressed Sequence Tag, EST, sequencing project, earthworm

The genome of any given animal comprises 5 % to 25 % coding DNA that is transcribed into mRNA. *In vitro*, these mRNAs can be reverse transcribed, resulting in stable complementary DNAs (cDNAs) which in turn can be cloned into cDNA library vectors. Expressed Sequence Tags (ESTs) are short single read sequences derived from cDNA library clones selected at random (see Fig. 1 for schematic diagram). Assuming that the number of ESTs is a reflection of actual mRNA levels, the EST strategy provides information on mRNA expression and the metabolic activity, and highlights important processes. Unfortunately, most sequence data submitted to public databases show an extreme species bias, with the majority of sequences derived from a few “model” species. EST analysis is a rapid and effective way to redress this phylogenetic

bias by providing the tools to acquire genomic data for neglected but by no means unimportant species. Prior to the year 2000, the entire taxonomic group of Lumbricidae was represented by 60 sequence data depositions with the majority originating from *Lumbricus rubellus* (22), *L. terrestris* (21) and *Eisenia fetida* (12). For this reason it was considered important to initiate an earthworm EST project. Here we provide a primary description of the earthworm EST project, the analysis of the first 577 single pass sequences and a brief introduction to Lumbribase.

Adult earthworms were sampled from pastures near Dinas Powys, Wales and whole worm total RNA prepared using TRI reagent (Sigma) followed by mRNA isolation by oligo dT-cellulose chromatography (Pharmacia). A  $\lambda$ -ZAP Express cDNA library was con-

\*E-mail corresponding author: [SturzenbaumSR@Cardiff.ac.uk](mailto:SturzenbaumSR@Cardiff.ac.uk)



**Fig. 1.** Schematic diagram of the earthworm EST sequencing approach

structed with 5 µg mRNA and rescued according to the manufacturer's protocol (Stratagene). The cDNA library approximated  $0.5 \times 10^6$  primary recombinants with an average size of 1400 bp (Fig. 2 A). DNA was isolated by standard protocols and processed for fluorescent dye terminator sequencing. Single pass sequences were determined from the 5' end of each clone using an ABI Sequenator. DNAs and sequencing chromatograms were collected and archived. ABI sequence reads were processed to derive high quality sequence and clustered using customised software. The median processed sequence length was 650 bp (Fig. 2 B) and the "cap3" software was used to produce a single data file from all overlapping /duplicate sequences. All sequences were automatically submitted to the public EST database, dbEST.

The entire dataset was analysed using the bioinformatic tools provided by the National Center for Biotechnology Information (NCBI). In detail, sequences (consensi for the clusters, individual reads for the singletons) were analysed using six frame translated searches (BlastX against non-redundant database) and nucleotide searches (BlastN against non-redundant database and BlastN against dbEST). All sequences were clustered using an algorithm based on 'BLAST' and 257 reads could be grouped into 79 clus-

ters of two or more reads. The remaining 320 sequences (55%) were single copy sequences and thus designated as 'singletons' (Fig. 2 C). These results reflect other EST project outputs (e.g. Inaba et al. 2002) and thus are a typical representation of a eukaryotic cell. The most abundant transcripts identified were of mitochondrial, ribosomal or housekeeping origin (Fig. 2 D). Exceptional, maybe, was the high incidence of blood coagulating enzymes / fibrinolytic enzymes with a proportional representation of nearly 6%. No significant similarity to known proteins was observed in 24% of cases. The remaining clustered nicely into a functional classification loosely adapted from Lee et al. (1999) (Fig. 2 E). The abundance of novel genes is comparable with other EST projects for example in the human parasite *Brugia malayi* (Blaxter et al. 2002), in common carp head kidney cells (Savan and Sakai 2002) and *Bos taurus* mammary glands (Sonstegard et al. 2002).

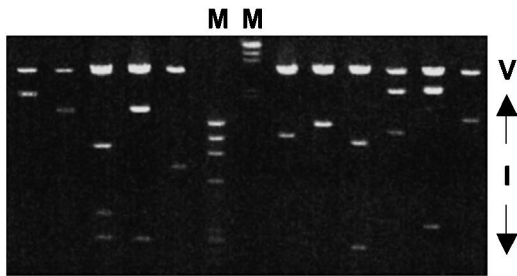
The resultant information was used to build a relational sequence database (LumbriBASE) that is freely accessible via the web (<http://www.earthworms.org>). LumbriBASE can be queried for sequence similarity (in house BLAST) and by annotation. Figure 3 shows a typical output of an annotation search. The input of a key word, in this case GAPDH (Fig. 3 A), provides information on identified clusters (Fig. 3 B), a cartoon of the sequence alignments (Fig. 3 C) and the final consensus sequence (Fig. 2 D). The BLAST menu (Fig. 3 E) provides information on sequence identity and similarity on a protein or nucleotide level (Fig. 2 F) and is identical to an NCBI BLAST output (Altschul et al. 1997).

Funds have been secured to start a follow-up project that will build on large scale amplification and sequencing processes developed and refined primarily for the nematode EST project (Parkinson et al. 2001). In future the earthworm EST project will therefore exploit fully automated high-throughput systems that use robotic facilities for plating and picking clones into high density bar coded micro-titre plates. Duplicate glycerol stocks will be generated and aliquots archived at  $-80^{\circ}\text{C}$  for future use in downstream post-genomic activities, which will include amongst others micro-array technology. In short, LumbriBASE is set to grow!

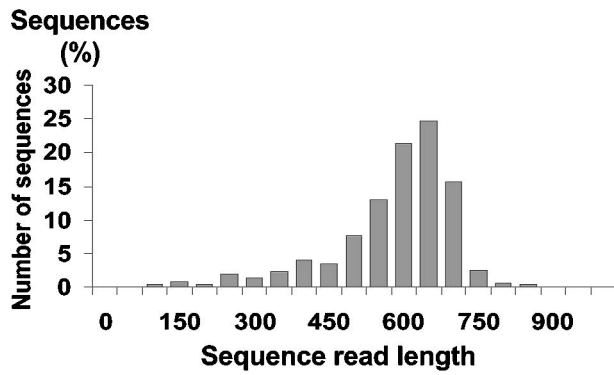
## References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389–3402.

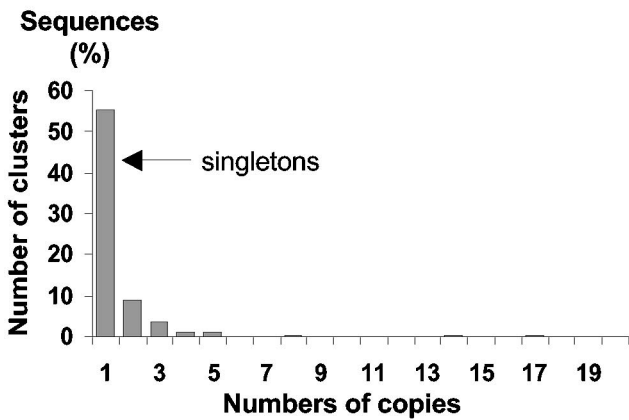
**A:**



**B:**



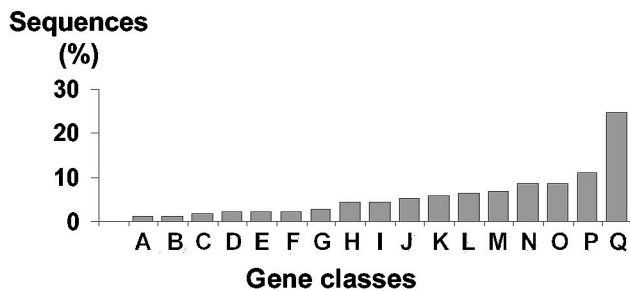
**C:**



**D:**

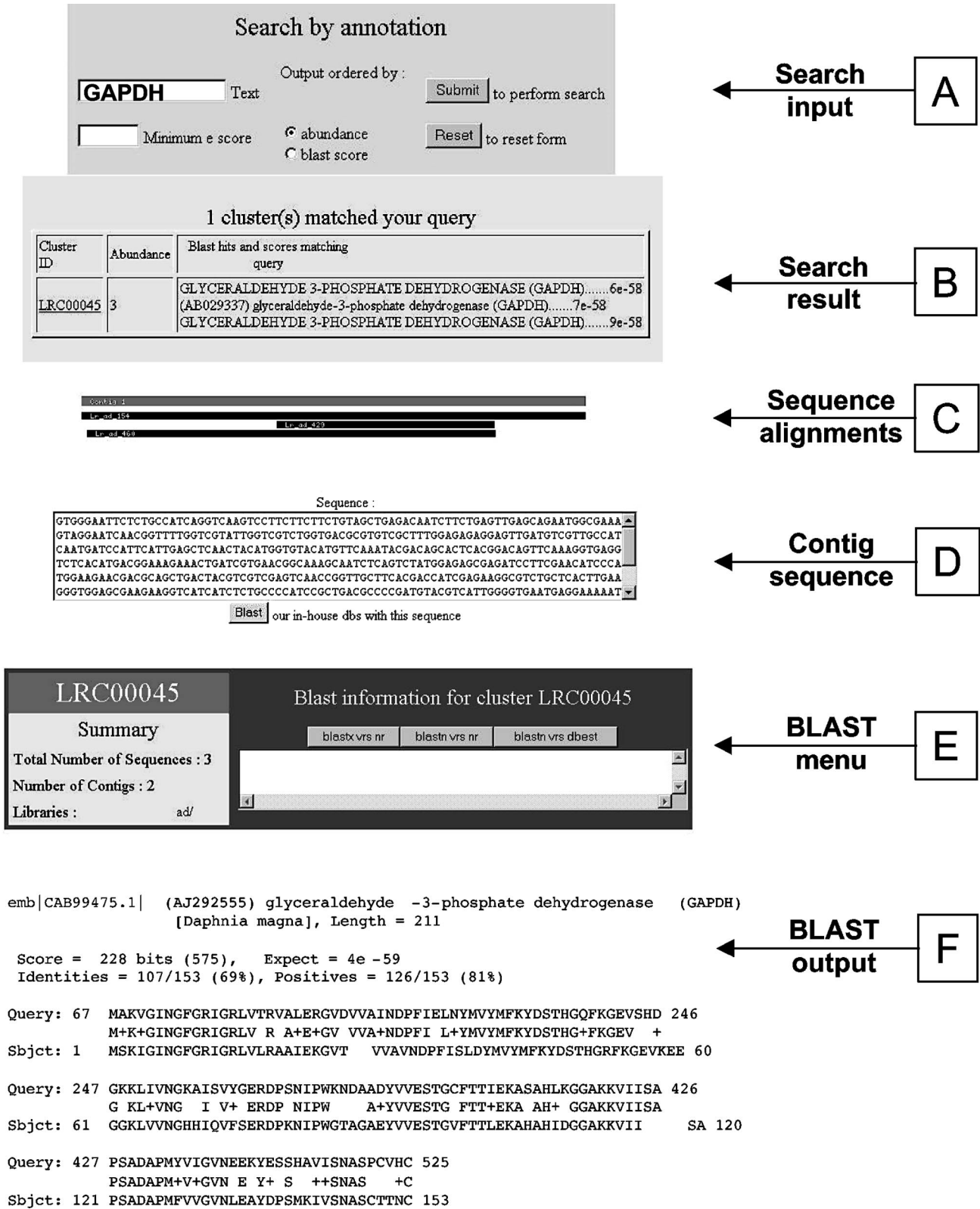
List of highly abundant cDNA clusters
mitochondrial large ribosomal subunit
fibrinolytic enzymes
troponins
cytochrome c oxidases
myosin
globins
translationally -controlled tumor protein
sialic acid -specific acetylsterase
60S ribosomal pro teins
actin

**E:**



- |                             |                                  |
|-----------------------------|----------------------------------|
| <b>A: Reproduction</b>      | <b>J: Catabolism</b>             |
| <b>B: Protein synthesis</b> | <b>K: Coagulation</b>            |
| <b>C: Stress response</b>   | <b>L: Ribosomal</b>              |
| <b>D: RNA binding</b>       | <b>M: Metabolism</b>             |
| <b>E: Transport</b>         | <b>N: Intermediate synthesis</b> |
| <b>F: Haemotic proteins</b> | <b>O: Muscular</b>               |
| <b>G: Housekeeping</b>      | <b>P: Others</b>                 |
| <b>H: Cell signaling</b>    | <b>Q: Novel</b>                  |
| <b>I: Metalloproteins</b>   |                                  |

**Fig. 2.** Analysis of the ESTs, indicating digested vectors (V), cDNA inserts (I) and molecular weight markers (M) (Panel A), sequence length (Panel B), number of contig clusters (Panel C), list of abundant genes (Panel D) and representation within gene classes (Panel E)



**Fig. 3.** A typical LumbriBASE search by annotation: after a key word input (Panel A) cluster match(es) will be displayed (Panel B) along with alignments (Panel C) and the resultant contig sequence (Panel D). Via the BLAST menu (Panel E) detailed information on sequence identity and similarity is provided (Panel F)

- Blaxter, M., Daub, J., Guiliano, D., Parkinson, J., Whitton, C. (2002) The *Brugia malayi* genome project: expressed sequence tags and gene discovery. Transactions of the Royal Society of Tropical Medicine and Hygiene 96, 7–17.
- Inaba, K., Padama, P., Satouh, Y., Shin-I, T., Kohara, Y., Satoh, N., Satou, Y. (2002) EST analysis of gene expression in testis of the ascidian *Ciona intestinalis*. Molecular Reproduction and Development 62, 431–445.
- Lee, Y.H., Huang, G.M., Cameron, R.A., Graham, G., Davidson, E.H., Hood, L., Britten, R.J. (1999) EST analysis of gene expression in early cleavage-stage sea urchin embryos. Development 126, 3857–3867.
- Parkinson, J., Whitton, C., Guiliano, D., Daub, J., Blaxter, M. (2001) 200 000 nematode expressed sequence tags on the net. Trends in Parasitology 17, 394–396.
- Savan, R., Sakai, M. (2002) Analysis of expressed sequence tags (ESTs) obtained from common carp, *Cyprinus carpio* L., head kidney cells after stimulation by two mitogens, lipopolysaccharide and concanavalin-A. Comparative Biochemistry and Physiology B 131, 71–82.
- Sonstegard, T. A., Capuco, A. V., White, J., Van Tassell, C. P., Connor, E. E., Cho, J., Sultana, R., Shade, L., Wray, J. E., Wells, K. D., Quackenbush, J. (2002) Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. Mammalian Genome 13, 373–379.