

# ESTree DB: a Tool for Peach Functional Genomics

Barbara Lazzari <sup>(1)</sup>, Luciano Milanesi <sup>(2)</sup>, Alessandra Stella <sup>(1)</sup>, Andrea Caprera <sup>(3)</sup>, Francesca Bianchi <sup>(4)</sup>, Alberto Vecchietti <sup>(1)</sup>, Carlo Pozzi <sup>(1)</sup>

<sup>(1)</sup> Parco Tecnologico Padano – CERSA, Via Haussmann, 7, 26900 Lodi - Italy  
carlo.pozzi@unimi.it

<sup>(2)</sup> Istituto Tecnologie Biomediche, Via Fratelli Cervi 93, 20090 Segrate (MI) – Italy

<sup>(3)</sup> CISI, Via Fratelli Cervi 93, 20090 Segrate (MI) – Italy

<sup>(4)</sup> Università degli Studi di Milano, Di.Pro.Ve. Via Celoria 2, 20133 Milano - Italy

**Keywords.** ESTree DB, Functional Genomics, Peach, Bioinformatics

## Introduction

A collection of about 8000 Expressed Sequence Tags (EST) sequences has been prepared starting from clones belonging to four cDNA peach libraries. Libraries have been prepared from *Prunus persica* mesocarps at four different developmental stages with the aim to collect data for deep investigation of the maturation process at the molecular level.

A fully automated pipeline (ESTree DB) has been prepared to process EST sequences using public software integrated by in-house developed Perl scripts and data have been collected in a MySQL database called ESTree available at this URL: <http://www.itb.cnr.it/ESTree>. These data are produced in the frame of the activities of the National Consortium for Peach Genomics (ESTree), involving also the Universities of Padova, Udine and other research Institutions.

## Materials and Methods

The program phred [1] has been used for base calling, producing three files for each sequence (electropherogram, text file and quality file) that are stored in the database. Sequence files and quality files have been processed with the program Lucy [2] in order to identify and remove vector contamination and low quality regions using default parameters. Vector-free high quality sequences have been submitted to the program CAP3 [3] to perform contig assembly. Stringency parameters have been modified (-p 95, -d 60) to prevent over assembly and help identify potential paralogs.

All the EST sequences and all the contig consensus sequences have been submitted to the BLASTx [4] program for annotation. BLASTx is run locally against the Genbank nr protein database. Blast output has been parsed with the MuSeqBox [5] BLAST parser program and MuSeqBox output is stored in a MySQL database. A number of accessory Perl programs has been integrated into the pipeline to allow data flow among the main public programs and to recover further information from intermediate elaboration steps and store it into the database.

A Unigene data set has been defined marking as Unigene all the singleton sequences and the longest sequence of each contig.

A php-based web interface has been prepared to surf and query the database.

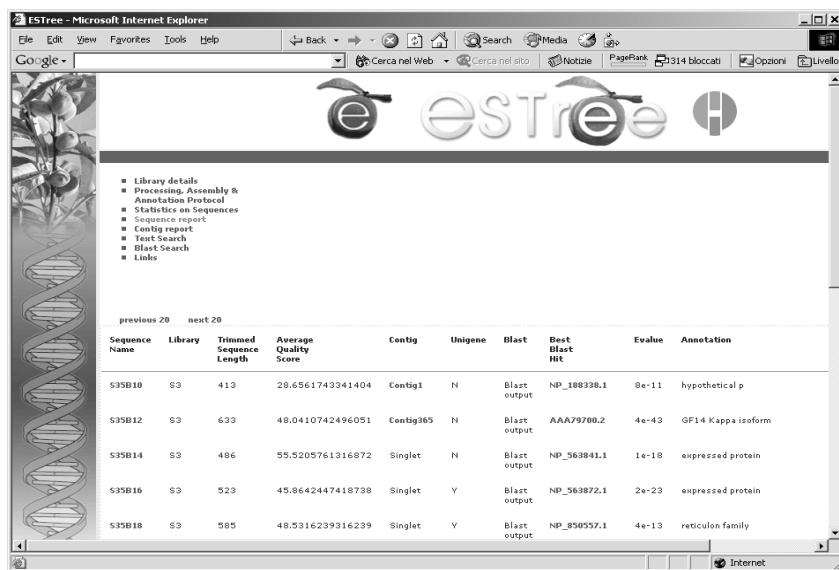
Multiple sequence tables are presented that contain links to single sequence pages and to every sequence-related information.

A text search utility is available and queries can be performed either on the sequence report table or on the contig report table exploring every database field. BLAST E-value intervals can also be selected by users.

Local BLASTn, BLASTp, BLASTx and tBLASTx programs are also available to perform BLAST and batch BLAST searches on the ESTree database.

## Results

The resulting database provides a collection of data for improving knowledge on peach functional genomics. EST sequences will be analysed for the detection of Single Nucleotide Polymorphisms (SNPs) with the aim to produce functional maps and microarray data will allow a time course analysis of the maturation events. Deeper data mining is also under progress to associate EST-derived putative proteins to metabolic pathways in order to make the ESTree database a complete collection of information concerning peach genomics and proteomics.



Sequence Name	Library	Trimmed Sequence Length	Average Quality Score	Contig	Unigene	Blast	Best Blast Hit	Evalue	Annotation
S35B10	S3	413	29.6561749341404	Contig1	N	Blast output	NP_188338.1	9e-11	hypothetical p
S35B12	S3	633	49.0410742496051	Contig365	N	Blast output	AAA79700.2	4e-43	GF14 Kappa isoform
S35B14	S3	486	55.5205761316872	Singlet	N	Blast output	NP_563841.1	1e-18	expressed protein
S35B16	S3	523	45.8642447418738	Singlet	Y	Blast output	NP_563872.1	2e-23	expressed protein
S35B18	S3	585	48.5316239316239	Singlet	Y	Blast output	NP_850357.1	4e-13	reticulon family

Fig. 1 The sequence report page in the ESTree DB web site

## Acknowledgements

This work is supported by Parco Tecnologico Padano, MIUR “Functional Genomics” 449/97, FIRB projects and the centre of excellence C.I.S.I. (Centre for Bio-molecular Interdisciplinary Studies and Industrial Applications).

## References

- [1] Ewing, B., Hiller, L., Wendl, M. and Green, P., Basecalling of automated sequence traces using phred. I. Accuracy assessment. *Genome Research* 8: 175-185, 1998.
- [2] Chou, H.-H. and Holmes, M.H., DNA sequence quality trimming and vector removal. *Bioinformatics*, 17:12, 1093-1104, 2001.
- [3] Huan, X. and Madan, A., CAP3: A DNA sequence assembly program. *Genome Research*, 9: 868-877, 1999.
- [4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410, 1990.
- [5] Xing, L. and Brendel, V., MuSeqBox: a program for multi-query sequence BLAST output examination. *Bioinformatics* 17: 744-745, 2000.