UNIVERSITA' DEGLI STUDI DI NAPOLI 'FEDERICO II'

FACOLTA' DI SCIENZE AGRARIE

Dipartimento di Scienze del Suolo, della Pianta, dell'Ambiente e delle Produzioni Animali

CORSO DI DOTTORATO DI RICERCA IN

AGROBIOLOGIA E AGROCHIMICA
indirizzo **MIGLIORAMENTO GENETICO ED ORTICOLTURA**
ciclo **XX**

BIOINFORMATICS ANALYSES FOR THE GENOMIC INVESTIGATION OF PLANTS OF AGRONOMIC IMPORTANCE

TUTOR

Prof. Luigi Frusciante

CO-TUTOR

Dott. Maria Luisa Chiusano

COORDINATORE DEL DOTTORATO

Prof. Antonio Violante

DOTTORANDO

Nunzio D 'Agostino

ANNO ACCADEMICO 2006/2007

Non vogliate negar l'esperienza
di retro al sol, del mondo sanza gente.
Considerate la vostra semenza
fatti non foste a viver come bruti
ma per seguir virtute e canoscenza

Dante Alighieri
Divina Commedia
Inferno, canto XXVI, 116-120

TABLE OF CONTENTS

1 BACKGROUND 6

1.1 EXPRESSED SEQUENCE TAGS (ESTS) 6

1.1.1 WHAT IS AN EST? 6
1.1.2 EST DATA QUALITY 7
1.1.3 IMPROVING EST DATA 8
1.1.4 EST RESOURCES 9
1.1.5 APPLICATION OF EST DATA 10

1.2 OVERVIEW ON EST APPLICATIONS: MANAGEMENT AND ANALYSIS OF PLANT EST COLLECTIONS TO ADDRESS DIFFERENT BIOLOGICAL QUESTIONS 13

1.2.1 COMPREHENSIVE ANALYSIS OF SOLANACEAE ESTS: GENE STRUCTURE PREDICTION AND COMPARATIVE GENOMICS 13
1.2.2 GENE-EXPRESSION PROFILES IN A *CROCUS SATIVUS* (SAFFRON) CDNA LIBRARY 16
1.2.3 CHARACTERIZATION OF A *CITRUS SINENSIS* GENE FAMILY BY EST SCREENING 17

2 METHODS 19

2.1 SET UP OF A PIPELINE FOR EST DATA ANALYSIS BASED ON PARALLEL COMPUTING 19

2.1.1 DATA SOURCES 19
2.1.2 REMOVAL OF THE OVER-REPRESENTED ESTS 20
2.1.3 PRE-PROCESSING: CHECKING FOR CONTAMINATIONS AND REPETITIVE ELEMENTS 20
2.1.4 CLUSTERING AND ASSEMBLING 21
2.1.5 FUNCTIONAL ANNOTATION 22
2.1.6 OVERVIEW OF THE GENE ONTOLOGIES 22
2.1.7 OVERVIEW OF THE ENZYME ASSIGNMENTS 22

2.2 THE EST DATABASE 24

2.2.1 IMPLEMENTATION AND ARCHITECTURE 24
2.2.2 WEB APPLICATION 25

2.3 EST-BASED GENE DISCOVERY AND GENE MODEL BUILDING 28

2.3.1 SETTING UP THE GENERIC GENOME BROWSER DATABASE 28

2.3.2 EST-TO-GENOME ALIGNMENTS 28
2.3.3 GENE MODELS FROM ESTS 28

2.4 COMPARATIVE ANALYSIS OF THE TOMATO AND POTATO TRANSCRIPTOMES 31
2.5 IDENTIFICATION OF NEW MEMBERS OF THE GLUTHATIONE S-TRANSFERASE SUPERFAMILY IN CITRUS SINENSIS 32

2.5.1 IDENTIFICATION OF ESTS ENCODING PUTATIVE GST PROTEINS 32
2.5.2 GST CLASS ASSIGNMENT 34
2.5.3 OPEN READING FRAME FINDING 35
2.5.4 MULTIPLE ALIGNMENTS GENERATION 35
2.5.5 TOTAL RNA EXTRACTION AND GENE EXPRESSION ANALYSIS BY SEMIQ RT–PCR 37

3 RESULTS 39

3.1 PARPEST EFFICIENCY 39

3.2 CHARACTERIZING THE TOMATO AND THE POTATO TRANSCRIPTOMES 41

3.2.1 THE TOMATO DATA-SET 41
3.2.2 BUILDING OF TOMATO UNIGENE SETS 41
3.2.3 FUNCTIONAL ANNOTATION OF THE TOMATO UNIGENE SETS 43
3.2.4 THE POTATO DATA-SET 45
3.2.5 BUILDING OF POTATO UNIGENE SETS 46
3.2.6 FUNCTIONAL ANNOTATION OF THE POTATO UNIGENE SETS 47

3.3 EST SURVEY OF OTHER SOLANACEAE TRANSCRIPTOMES 50

3.4 GENE HUNTING: ESTS AND GENE MODEL BUILDING 51

3.5 ARABIDOPSIS PROTEOME INFORMATION FOR INTERPRETING SEQUENCE CONSERVATION AND DIVERGENCE BETWEEN TOMATO AND POTATO 52

3.6 ISOL@: AN ITALIAN SOLANACEAE GENOMICS RESOURCE 59

3.7 SAFFRON GENES: AN EST DATABASE FROM SAFFRON STIGMAS 63

3.7.1 CONSTRUCTION AND FUNCTIONAL ANNOTATION OF SAFFRON UNIGENE SET 63
3.7.2 TCS COMPOSED OF MOST ABUNDANT ESTS 65

3.8 ON THE GLUTHATIONE S-TRANSFERASE GENE FAMILY IN CITRUS SINENSIS 69

3.8.1 *IN SILICO* IDENTIFICATION AND TISSUE EXPRESSION PROFILING OF GST ENCODING TRANSCRIPTS 70

4 DISCUSSION 74

4.1 THE SIGNIFICANCE OF EST IN THE 'OMICS' ERA 74

5 CONCLUSION 80

6 LITERATURE CITED 82

1 BACKGROUND

The growing availability of 'sequence data' combined with the 'high throughput' technologies makes bioinformatics essential in supporting the analysis of the structure and the function of biological molecules. To this end, bioinformatics plays a key role for data-mining and it has broad applications in the molecular characterization of an organism's gene and protein space (i.e. genomics and proteomics); in the genome-wide study of mRNA expression (transcriptomics); in the systematic study of the chemical fingerprints that specific cellular processes leave behind (metabolomics); in drug discovery; in the identification of biomarkers as biological indicators of disease, toxicity of pathogens or effectiveness of healing. The real bioinformatics challenge is the design of computational methods which should be suitable as to reveal the information that biological data still hide as to integrate the large amount of '-omics' data, in the attempt to approach a systems biology view. The long term goal is the creation of models for the simulation of biological systems' behaviour and their exploitation into biotechnology, plant and agricultural science applications. This work encompasses the design of methods and the implementation of algorithmic tools to facilitate the collection, the organization and the analysis of large amounts of plant 'sequence data' in the mold of Expressed Sequence Tags.

1.1 Expressed Sequence Tags (ESTs)

1.1.1 What is an EST?

Messenger RNA (mRNA) sequences represent expressed genes in the cell. The "reverse transcription" mechanism allows the genetic information contained in the mRNA to be converted into a double-stranded DNA form (i.e. complementary DNA or cDNA). The resultant cDNA can be inserted into an appropriated plasmid (cloning) so as to produce a cDNA clone. The collection of cDNA clones, isolated from an organism or a specific tissue, represents a cDNA library. All the cDNA clones in a library, can be sequenced using a large-scale approach. This means that they are randomly sequenced on a single strand yielding 5' and 3' Expressed Sequence Tags (ESTs).

A single sequencing read produces from 100 to 800 readable nucleotides. Thus, an EST provides a "tag level" association with an expressed gene sequence.

cDNA cloning and EST sequencing are schematically summarized in figure 1.

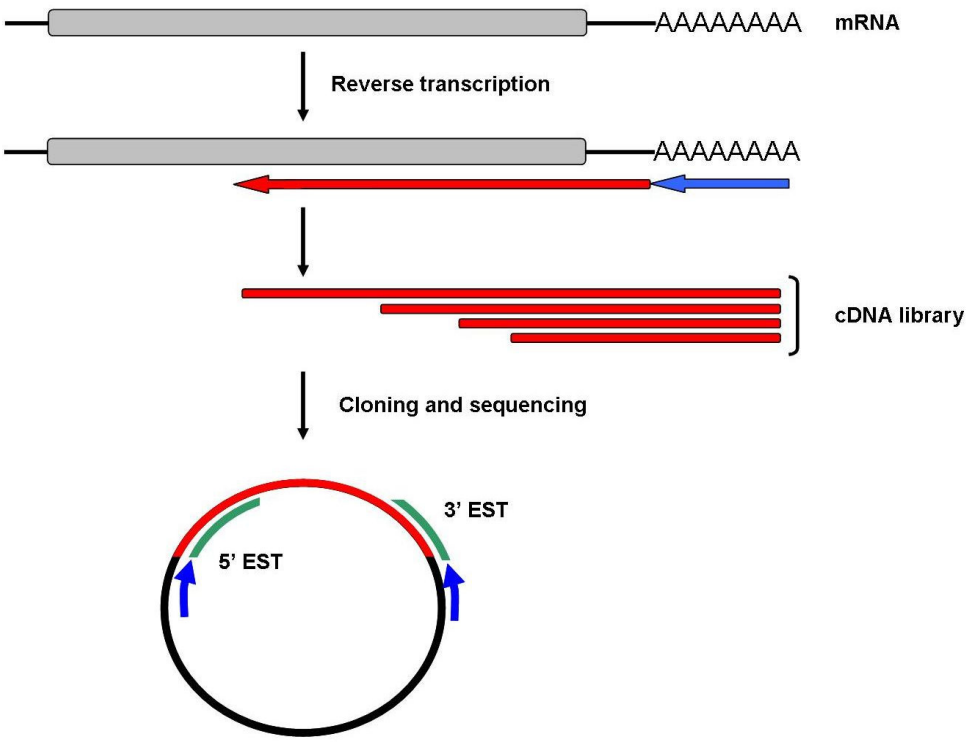


Figure 1. Summary of cDNA cloning and EST sequencing.
A cDNA population is reverse transcribed from the mRNA population. The 3' poly-A tail is used as a selective tag for mRNA selection. cDNA sequences are inserted into a cloning vector or plasmide. Then using *ad hoc* primers, the DNA sequence is read from the ends of the cDNA, yielding 5' and 3' ESTs.

1.1.2 EST data quality

EST data quality is highly variable and depends on the sequencing approach (e.g. partial, single-pass sequencing) and on the cDNA library construction (e.g. concatenated adaptors/linkers, chimeric genes or inversely inserted cDNAs). However, it is generally accepted that ESTs are highly error prone sequences, especially at the ends where base-calling and/or base-stuttering (repeated bases) errors are frequently observed. On the other hand the overall sequence quality is significantly better in the middle. Furthermore there can be possible contaminations, either at the end or rarely in the middle, from vector or linker/adaptor.

1.1.3 Improving EST data

ESTs represent the first truly high-throughput technology to have populated the biological databases and have made the rapid growth of advanced computational studies in biology inevitable.

Expressed Sequence Tags are generated and deposited in the public repository as redundant and un-annotated sequences, with negligible biological information content. The weak signal associated to an individual raw EST increases when a lot of ESTs are analysed together, so as to provide a snapshot of the transcriptome of a species.

Different strategies, which use different combinations of computational tools, have been developed for the analysis of large data-sets of ESTs (http://biolinfo.org/EST/). These strategies are originated from the need of making the analysis, the organization and the storage of EST data automatic.

A generic EST analysis pipeline should schedule the following steps:

1. raw EST sequences are screened for the identification and the removal of vector sequences. Then, repeats and low complexity sub-sequences are detected and masked;
2. the high quality EST data-set is then subjected to a clustering/assembling procedure in order to group overlapping ESTs (putatively derived from the same gene) and to generate consensus sequences which putatively represent the transcripts. This step permits hopefully the full-length transcript sequences to be reconstructed - by gathering information from several short EST sequences simultaneously - and the redundancy of the EST collection to be reduced;
3. DNA and/or protein database similarity searches are carried out to assign a putative function. The value of the functional annotations can be enhanced by performing protein domain and motif analysis as well as by including gene ontology assignments.

A detailed list of the main key tools used to accomplish the different tasks of the EST analysis is reported in table 1 while the available EST analysis pipelines are listed in table 2.

Programs for EST pre-processing			
Name	**Website**	**Description**	**Category***
Phred/Cross_Match	http://www.phrap.org/	Base caller/vector trimming and removal	F
SeqClean	http://compbio.dfci.harvard.edu/tgi/software/	trimming and validation of ESTs	F
VecScreen	http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html	Vector, linker and adapter identification	F
Vector cleaning	http://www.aborygen.com/products/biOpen/tools-for-biOpen/vector-cleaning.php	Vector cleaning	C
Paracel	http://www.paracel.com/	EST pre-processing package	C, W
Lucy2	http://www.complex.iastate.edu/download/Lucy2/index.html	Sequence trimming and visualization	F
Dust	ftp://ftp.ncbi.nih.gov/blast/	Low-complexity regions masked	F
MaskerAid	http://blast.wustl.edu/maskeraid/	Repeats masked	F
RepeatMasker	http://www.repeatmasker.org/	Repeats masked	F
Programs for EST clustering			
blastclust	ftp://ftp.ncbi.nih.gov/blast/	is part of the standalone **BLAST** package	F
megaBLAST	ftp://ftp.ncbi.nih.gov/blast/	is part of the standalone **BLAST** package	F, W
THC_BUILB	http://compbio.dfci.harvard.edu/tgi/software/	is part of the **TGICL** package	F
d2_cluster	http://www.sanbi.ac.za/Dbases.html#stackpack	is part of the **stackPACK** system	F
ESTate package	http://www.ebi.ac.uk/~guy/estate/		F
CLOBB	http://xyala.cap.ed.ac.uk/CLOBB/		F
PaCE	The source code and executables can be obtained by email		F
Programs for EST sequence assembly and consensus generation			
CAP3	http://genome.cs.mtu.edu/cap/cap3.html		F, W
TIGR_ASSEMBLER	http://www.tigr.org/software/assembler/		F
Phrap	http://www.phrap.org/		F
essem	http://alggen.lsi.upc.es/recerca/essem/frame-essem.html		F
miraEST	http://www.chevreux.org/projects_mira.html		F
Paracel Transcript Assembler	http://www.paracel.com/		C, W

Table 1. Programs used to accomplish the different tasks of the EST analysis (Modified from Nagaraj et al., 2006). *F= free for academic users; C= commercial package; W= web interface available.

EST analysis pipeline			
Pipeline Name	**Pre-processing**	**Clustering & Assembling**	**ORF prediction and EST functional annotation**
TGICL	**SeqClean & megaBLAST**	**CAP3 Paracel TranscriptAssembler**	**DIANA-EST, ESTscan & Framefinder**
ESTAP	**Phred & Cross_Match**	**D2_cluster & CAP3**	**BLASTX**
ESTIMA	**Information not available**	**BlastClust & CAP3**	**BLASTX**
ESTAnnotator	**Phred & RepMask & UniVec**	**CAP3**	**BLASTX**
PipeOnline	**Phred & Cross_Match**	**Phrap**	**BLASTX**
openSputnik	**Cross_Match**	**HPT2(Biomax informatics) & CAP3**	**ESTscan & BLASTX**

Table 2. Characteristics of the available EST analysis pipelines (Modified from Nagaraj et al., 2006).

1.1.4 EST resources

In 1993, a database called dbEST (Boguski et al. 1993) was established to serve as a collection point for ESTs. Since then they are distributed to the scientific community as the EST division of GenBank. This database represents the primary data source from which the EST sequences are recovered to be processed.

The quality limitations, the issues of redundancy as well as the less-than-full-length nature of ESTs, are the motivations of the development of automated analytical systems for the reconstruction and organization of expressed gene sequences.

Many public EST resources have been developed in the attempt to address these questions. The most widely known effort is UniGene (Pontius et al., 2003). It is a system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene. It should be noted that no assembly is available in UniGene, but only the longest sequence in each cluster is indicated.

The TIGR Gene Index Project (Lee et al. 2005) aims to integrate data from international EST sequencing projects, in order to produce species-specific sets of unique and high-

fidelity virtual transcripts which are called TCs from tentative consensus sequences. Each set is an inventory of likely genes for which additional information, concerning their functional roles, are provided.

The PlantGDB Project (Dong et al., 2005b) intends to gather information concerning all the major plant species from every GenBank release, providing an estimation of plant gene space. The PlantGDB-assembled Unique Transcripts (PUT) are an important resource for the plant research community.

The STACK Project (Christoffels et al., 2001) is comprised for the STACKdb™, a resource of virtual human transcripts, as well as stackPACK™, the tools to create the catalogue of the transcripts. The system organizes the ESTs into tissue/state context, then each tissue-grouped set is sent through a pipeline of clustering, assembling and consensus generation. STACK differs from other gene indexing projects because of the tissue-based and/or disease-related segmentation of the virtual human transcripts. In addition, the non-alignment-based clustering tool d2_cluster (Burke et al., 1999) is focused on the comprehensive capturing of transcript variants.

1.1.5 Application of EST data

Though intrinsic shortcomings due to contaminations and limited sequence quality, ESTs are a versatile data source and have multiple applications. In the absence of complete genome sequences, the cDNA (and its ESTs) remains the only link back to the genome (Richmond and Somerville, 2003).

EST sequences can be used as landmarks in the construction of physical genome maps. An EST sequence can be used as STS (i.e. Sequence-Tagged Site) assuming that it is operationally unique and has single occurrence in a genome.

Expressed sequence tags are widely used for gene location discovery and for gene structure prediction. Gene predictions are usually based on spliced-alignment of source-native ESTs onto the genomic sequences (Adams et al., 1991; Kan et al. 2001; Brendel et al., 2004).

A particular exciting aspect is the use of EST data aimed to investigate different types of mRNA transcription variants such as those due to alternative splicing, initiation, polyadenylation and intron retention. From a mixture of EST fragments, the most likely set of full-length isoforms can be reconstructed (Gautheret et al., 1998; Brett et al., 2000; Gupta et al., 2004; Galante et al., 2004).

The usefulness of EST data has been extended to the discovery and the characterization of the most common type of DNA sequence variation, the single nucleotide polymorphisms (SNPs). Considering that the majority of the EST libraries are obtained from different individuals, the assembly of overlapping sequences for the same region can lead to the identification of new SNPs (Picoult-Newberg et al., 1999).

A further relevant application of EST data is the study of gene expression. Digital gene expression profiles (i.e. digital Northern) can be successfully exploited to point out the expression levels of different genes. The strategy is based on the fact that the number of ESTs is reported to be proportional to the abundance of cognate transcripts in the tissue or cell type used to make the cDNA library (Audic and Claverie, 1997).

Large scale computer analyses of EST sequences can be used in the identification and in the analysis of co-expressed genes (Ewing et al., 1999; Wu et al., 2005; Faccioli et al., 2005). It is important to find genes with similar expression patterns (co-expressed genes) because there is evidence that many functionally related genes are co-expressed and because this co-expression may reveals more about the genes' regulatory systems. Similar analyses can be carried out in order to point out genes exhibiting tissue- or challenge-specific expression (Mégy et al. 2003).

EST sequences represent a valuable resource for designing oligonucleotide probes for array chip. With the advent of cDNA array-based methods, ESTs have become a key reagent within an experiment rather than the final product. In these arrays, a large collection of cDNAs is fixed to a substrate and an associated EST sequence provides the link between an experimental coordinate and a gene that might be up- or down-regulated. Array experiments allow massive, parallel investigation of gene expression from different tissues or under specific challenges, for example biotic or abiotic stress conditions.

Expressed sequence tags remain a dominant reference for the characterization of the protein-encoding portions of various genomes. The larger the EST collection to examine, the grater the possibility to generate all the theoretical protein coding regions expressed within a genome (defining of a virtual proteome). Therefore, ESTs have also become invaluable resources in the area of proteomics for peptide identification and proteome characterization, especially in the absence of complete genome sequence information (Lisacek et al., 2001).

Last but not least, the utilization of EST data for comparative genomics must be mentioned. Assuming that ESTs are a quick method of sampling an organisms'

transcriptome, large-scale analysis can be performed in order to encompass the evolution of genome function and structure; to assess sequence conservation and divergence between transcriptomes of different organisms and/or finally to illustrate the process of sequence divergence during speciation (Dong et al., 2005a; Caicedo and Purugganan 2005).

1.2 Overview on EST applications: management and analysis of plant EST collections to address different biological questions

1.2.1 Comprehensive analysis of Solanaceae ESTs: gene structure prediction and comparative genomics

The Solanaceae family (common name: Nightshade) comprises about 95 genera and at least 2,400 species. Many of these species have considerable economic importance as food (tomato, potato, eggplant, garden pepper), ornamental (petunia) and drug plants (tobacco). The Solanaceae species show a wide morphological variability and occupy various ecological niches though they share high genome conservation. The need of increasing the knowledge of the genetic mechanisms which determine Solanaceae diversification and adaptation, has brought the scientific efforts to be gathered into the International Solanaceae (SOL) Genome Project (Figure 2). The cultivated tomato, *Solanum lycopersicum*, is the plant chosen by the SOL initiative for a BAC-based genome sequencing. The long term goal is to exploit the information generated by the Tomato Genome Sequencing Project (Mueller at al., 2005b) for the analysis of the genome organization, of the functionality and the evolution of the entire Solanaceae family. Early studies of the Solanaceae genomes, in fact, revealed conservation of gene content among potato, tomato, tobacco, petunia and eggplant (Zamir and Tanksley, 1988). Furthermore, these Solanaceae species have a base chromosome number of twelve.

Figure 2. The International Solanaceae Genomics Project (SOL): Systems Approach to Diversity and Adaptation. Whitepaper at: http://www.sgn.cornell.edu/documents/solanaceae-project/docs/SOL_vision.pdf.

In order to address key questions risen by the SOL vision, large amounts of data from different "-*omics*" approaches are being generated. These data are going to enrich the existing sequence data, which year after year have been made available for the Solanaceae.

Since no full genome sequence of a member of the Solanaceae family is yet available, much of the existing worldwide sequence data consists of EST sequences. Their

availability has dramatically increased afterward the start-up of the International Tomato Genome Sequencing Project (Figure 3).

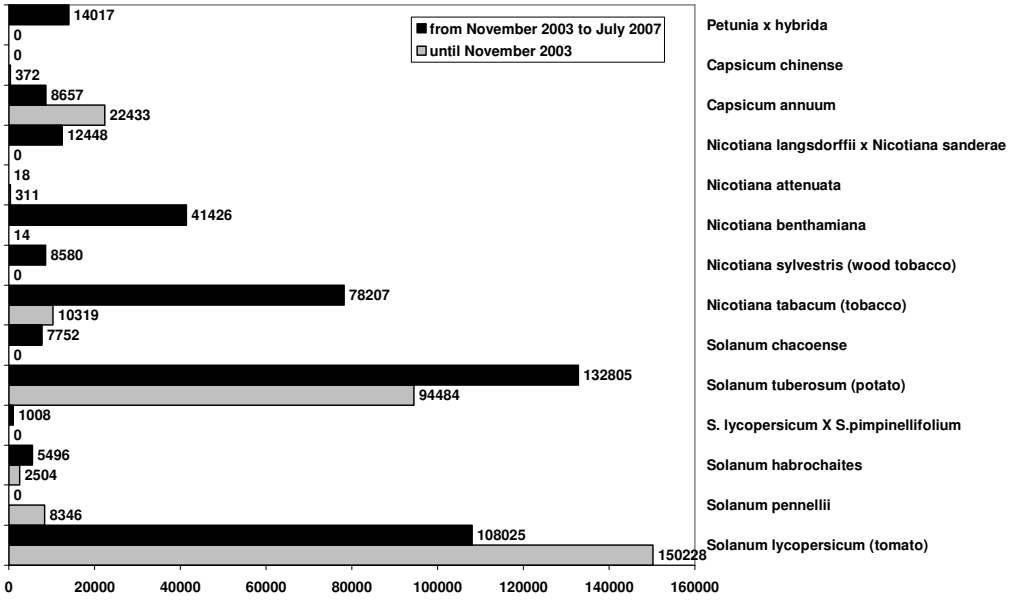


Figure 3. Distribution of EST data submission. The Solanaceae species with most available ESTs at dbEST are reported. The black bars indicate the number of EST sequences submitted to the dbEST repository since November 2003, the start-up of the International Tomato Genome Sequencing Project, until July 2007. The grey bars indicate the overall number of EST entries available at dbEST (release 072007).

It is not surprising because the screening of the Solanaceae EST collections represents a noteworthy and parallel contribution to the study of the tomato gene content as well as to the investigation on the Solanaceae family members.

Hereby the Solanaceae EST resources and repositories worldwide available are briefly discussed. The SOL Genomics Network (SGN; Mueller at al., 2005a), a website dedicated to the biology of the Solanaceae family, organizes and distributes ESTs and the corresponding unigene builds for tomato, potato, pepper, eggplant, and petunia.

Another sizeable effort is the TIGR collection of high-fidelity virtual TC sequences (Lee et al. 2005) constructed by clustering and assembling ESTs from pepper (release 2.0), potato (release 11.0), petunia (release 1.0), tobacco (release 3.0), *Nicotiana benthamiana* (release 2.0) and tomato (release 11.0).

The PlantGDB (Dong et al., 2005b) is a valuable resource which intends to provide an estimation of plants' gene space. It picks up EST collections from tomato, potato, petunia and different species of Nicotiana genus.

Furthermore, in order to address particular research interests, specific resource have been developed such as the Tomato Stress EST Database (TSED;

http://abrc.sinica.edu.tw/ibsdb/), which contains *S. lycopersicum* ESTs from more than 10 stress-treated subtractive cDNA libraries and the Micro-Tom Database (MiBASE; Yano et al., 2006), which distributes ESTs from a full-length cDNA library from the fruit of Micro-Tom (a miniature and dwarf tomato cultivar). The list of the cited Web resources is summarized in table 3.

RESOURCE	*WEB SITE*	*EST COLLECTION AVAILABLE*
Solanaceae Genomics Network (SGN)	http://www.sgn.cornell.edu/	tomato, potato, pepper, eggplant, and petunia
TIGR Plant Gene Indices	http://compbio.dfci.harvard.edu/tgi/plant.html	tomato, potato, pepper, petunia, tobacco, N. benthamiana
PlantGDB -Plant Genome DataBase	http://www.plantgdb.org/prj/ESTCluster/index.php	tomato, potato, petunia and different species of Nicotiana genus.
Tomato Stress EST Database (TSED)	http://ibs.sinica.edu.tw/ibsdb/app_all/index.php	tomato ESTs from stress-treated substractive cDNA libraries.
MiBASE- Micro-Tom Database	http://www.kazusa.or.jp/jsol/microtom/indexj.html	Micro-Tom EST libraries

Table 3. Summary of Solanaceae EST resources.

I will report on the analysis of EST sequences of 14 Solanaceae species available at dbEST and on the construction of the corresponding gene indices. Certainly, these ESTs are hardly useful as they stand and need to be converted into biological meaningful information. Therefore, bioinformatics approaches become pre-eminent, though the results might be far from being exhaustive and complete.

To this end, I implemented ParPEST (Parallel Processing of ESTs; D'Agostino et al., 2005), the pipeline that automatically executes the different steps the EST analysis requires, and in addition I designed the relational database where the processed data were stored. The need of creating a custom tool originates from the fact that only few efforts have been made to integrate all the consecutive steps for EST pre-processing, clustering/assembling and annotation into a single procedure (Table 2). Furthermore, the diversity of data sources, the quality of the annotations and the produced detailed information make our effort useful in the context of EST Solanaceae resources.

Once the gene indices have been constructed, they were used for gene discovery and gene structure predictions. In fact, my team - as part of the iTAG (international Tomato genome Annotation Group; http://www.ab.wur.nl/TomatoWiki) – has been committed, within the EU-SOL Project, to align Solanaceae expressed transcripts to the tomato draft genomic sequences (i.e. BAC sequences) released by the Tomato Genome Sequencing Consortium.

Finally, the EST resources have been exploited for a survey of the Solanaceae transcriptomes and for an Arabidopsis-based investigation to assess sequence conservation and divergence between tomato and potato data-sets.

1.2.2 Gene-expression profiles in a *Crocus sativus* (saffron) cDNA library

Saffron (*Crocus sativus L.*) is a triploid, sterile plant which has been propagated and used as a spice and as a medicinal plant in the Mediterranean area for thousands of years (Fernandez, 2004). It is likely that the domestication of saffron occurred in the Greek-Minoan civilization between 3.000 and 1.600 B.C. A fresco depicting saffron gatherers dating back to 1.600 B.C. has been unearthed in the island of Santorini, Greece. Saffron is commonly considered the most expensive spice on earth. Nowadays, the main producing countries are Iran, Greece, Spain, Italy, and India (Kashmir). Apart from the commercial and historical aspects, several other characteristics make saffron an interesting biological system: the spice is derived from the stigmas of the flower (Figure 4A), which are manually harvested and subjected to desiccation. The main colours of saffron, crocetin and crocetin glycosides, and the main flavours, picrocrocin and safranal, are derived from the oxidative cleavage of the carotenoid, zeaxanthin (Bouvier et al., 2003b; Moraga et al., 2004) (Figure 4B).

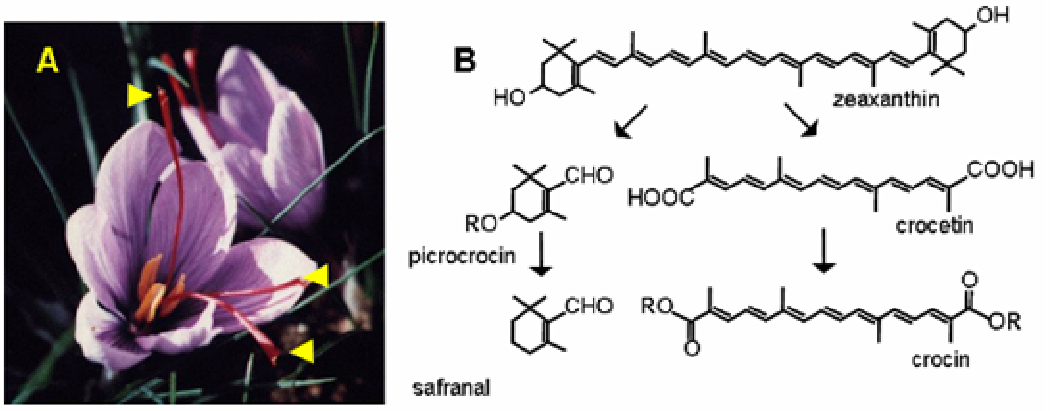


Figure 4. The saffron spice.
A. Crocus flowers. Arrowheads point to the stigmas which once are harvested and desiccated constitute the saffron spice.
B. Biosynthetic pathway of main saffron colour (crocin) and flavours (picrocrocin and safranal) (modified from Bouvier et al., 2003b).

Saffron belongs to the Iridaceae (Liliales, Monocots) with poorly described genomes. The characterization of the transcriptome of saffron stigmas is likely to shed light on

several important biological phenomena: the molecular basis of flavour and colour biogenesis in spices; the biology of the gynoecium; and the genomic organization of Iridaceae. For these reasons, the sequencing and the bioinformatics characterization of expressed sequence tags from saffron stigmas have been undertaken. The primary goal is to point out the expression levels of different genes, assuming that the number of ESTs is proportional to the abundance of cognate transcripts in the stigma tissue.
This research topic originates from a collaborative effort between my team and the one directed by Mr. Giovanni Giuliano, who is Senior Research Scientist at ENEA (Italian National Agency for New Technologies, Energy and the Environment) in Rome.

1.2.3 Characterization of a *Citrus sinensis* gene family by EST screening

This research originates in the frame of the AGRONANOTECH Project and is in collaboration with the team directed by Mr. Giuseppe Reforgiato Recupero, who is Senior Research Scientist at the ISAGRU Institute (CRA - Istituto Sperimentale per l' Agrumicoltura) in Acireale.
Mr Reforgiato's team is involved in the study of the anthocyanin biosynthetic pathway and of the molecular mechanisms that underpin the production and accumulation of anthocyanin pigments in the flesh of the blood orange fruits.
In a previous paper (Licciardello et al., 2007), they discussed on the identification of differentially expressed genes in the flesh of pigmented (Moro nucellare 58-8D-I) and non-pigmented (Blonde cadenera) orange genotypes. One of the genes, that have been detected by expression profiling and resulted up-regulated in the flesh of pigmented orange, encodes for a gluthatione S-transferase (GST), an enzyme of the anthocyanin pathway (Figure 5).
It has been shown that plant GSTs are important in binding secondary metabolites like anthocyanins and in their transfer from the site of synthesis in the cytoplasm into the vacuole, where they are permanently stored (Marrs 1996; Mueller et al. 2000). However, it is known that plant cells can express several GSTs belonging to different GST classes and grouped into a gene family (Frova, 2003). For this reason, we decided to characterize the GST gene family by screening a collection of 94.127 *Citrus sinensis (L.) Osbeck* expressed sequence tags. Tissue expression patterns of the putative full-length transcripts identified in this study were inferred by querying the dbEST database with respect to different tissues. Furthermore, Semi-Quantitative Reverse Transcription

Polymerase Chain Reaction (SemiQ RT-PCR) analyses were performed by Miss Licciardello in order to check the real presence in the cells of the *in silico* assembled GST transcripts and to assess their expression patterns in the albedo, flavedo, flesh, young and adult leaves and ovary. In addition, the experiments were also performed to compare the gene expression levels between the Blonde cadenera and Moro nucellare 58-8D-I genotypes.

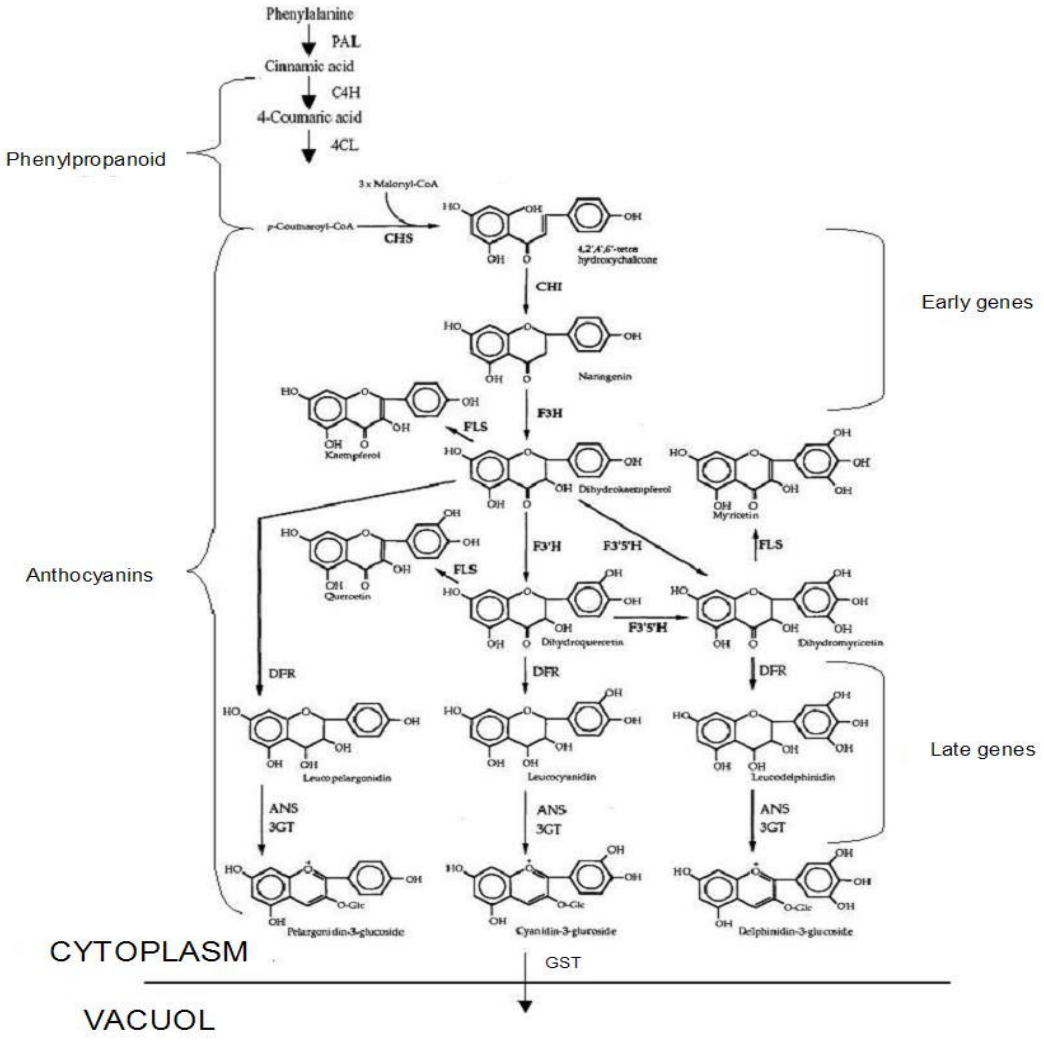


Figure 5. Schema of the anthocyanin pathway.

2 METHODS

2.1 Set up of a pipeline for EST data analysis based on parallel computing

The "analysis pipeline" ParPEST was implemented using public software integrated by in-house developed Perl scripts (D'Agostino et al., 2005). It currently operates on a "Beowulf class" cluster running the Fedora Linux Core 4 operating system and the OSCAR 4.0 toolkit (http://hpcs2003.ccs.usherbrooke.ca/papers/desLigneris_01.pdf) able to support the cluster management and the job scheduling and monitoring.

The schematic view of the ParPEST pipeline is shown in figure 6.

In succession, a brief description of each module of the analysis and the input/output resources are discussed.

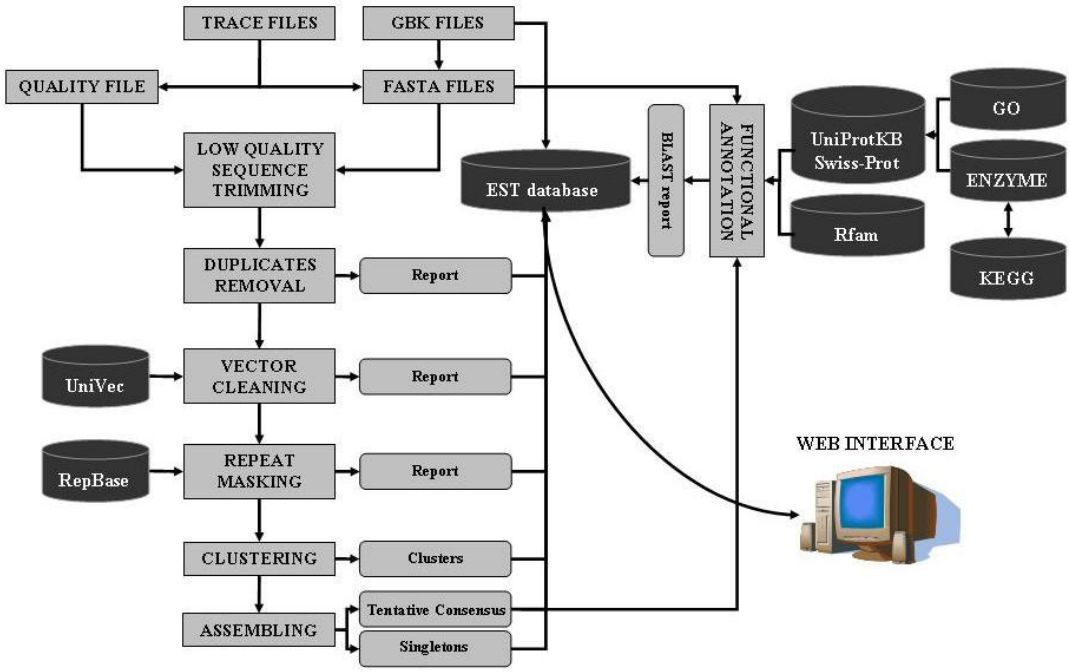


Figure 6. Schematic view of the ParPEST pipeline.

2.1.1 Data sources

EST sequences can be submitted to the ParPEST pipeline in the ABI (Applied Biosystems Inc.), GenBank or FASTA format. In case of submission in the ABI format, the first module in ParPEST is the base-caller PHRED (Ewing and Green, 1998) as it converts chromatogram files to bases and quality indices. The output sequences are

written in the standard FASTA format. Besides the input format, it is convenient that the following valid attributes are associated to each EST: the name of the library, the organism from which the library has been prepared, the organism strain or the plant cultivar, the tissue type, the developmental stage and a brief description of the library preparation method.

2.1.2 Removal of the over-represented ESTs

It is known that the dbEST division of the GenBank repository contains a large number of EST copies (i.e. over-represented ESTs). In order to reduce the EST analysis execution time, which slows down dramatically as the number of input sequences grows, it is necessary to remove these over-represented EST sequences so as to clip the original data-set and produce a non-redundant collection. To achieve this task, a parallel algorithm has been designed (C++ language). First, input ESTs are sorted according to their sequence length. Then, each input sequence is compared to each other considering also its reverse-complement counterpart. If a sequence X is contained in the sequence Y as X being a subsequence of Y, the sequence X is excluded from the subsequent step because its sequence information is already represented by Y. Therefore, it is defined as a “contained” sequence and the “contained” relationship between X and Y is marked so that it can be easily recorded in a database table. Indeed, for each “container” sequence (i.e. the larger “parent” sequence) can be traced the contribution of all the “contained” sequences (i.e. children sequences).

2.1.3 Pre-processing: checking for contaminations and repetitive elements

A division of the EST sequences deposited in dbEST can be contaminated by non-native sequences such as those derived from the cloning vectors or the bacterial host. In addition, ESTs can include repetitive elements as well as low complexity sub-sequences. This can prevents the correct and accurate generation of clusters and assemblies. The pipeline includes the RepeatMasker (http://www.repeatmasker.org/) tool to identify both contaminations and low complexity sub-sequences and/or repetitive elements. In particular, the NCBI UniVec database (ftp://ftp.ncbi.nih.gov/pub/UniVec/) was used for the identification and the masking of vector and bacterial contaminations. Once the masked nucleotides were trimmed off, the next step follows.

RepBase (Jurka, 2000) was used as the filtering database for masking low complexity sub-sequences and interspersed repeats. RepeatMasker is not designed for parallel computing. Therefore, in order to reduce its execution time, a specific utility was designed to submit an arbitrary number of serial jobs to the PBS (Portable Batch System) simply by creating a command file with one command per line. Such utility showed an excellent speedup over its sequential counterpart and its memory requirements are almost negligible making it suitable to run virtually on any data size.

2.1.4 Clustering and Assembling

The sequences were clustered by the PaCE (Parallel Clustering of ESTs) program (Kalyanaraman et al., 2003) which builds a distributed representation of the generalized suffix tree data structure in parallel. This data structure was used for the generation of groups of overlapping sequences. It is assumed that all sequences in a cluster represent the same gene, this is why each cluster is defined as a gene index. For the parallel execution PaCE requires an MPI (Message Passing Interface) environment.

CAP3 (Huang and Madan, 1999), with an overlapping window of 60 nucleotides and a minimum score of 85, was the program used to perform the assembling process. A specific utility was designed to bundle groups of CAP3 commands to be executed sequentially by each processor in order to speed up the execution of CAP3 and to avoid the overhead time consuming of PBS. All the EST sequences into a PaCE cluster, were assembled into tentative consensus sequences (TCs) which were generated from multiple sequence alignments of ESTs. Each EST which during the clustering process did not meet the match criteria to be clustered with any other EST, can be thought of as a cluster by itself and it is defined as singleton (sEST).

The EST set in a cluster can be assembled in one or multiple TCs. Indeed, since the clustering process is a simple “tentative closure” procedure, PaCE finds the overlaps among EST sequences not considering if they make sense all together. When sequences in a cluster cannot be all reconciled into a consistent multiple alignment during the much more rigorous assembly phase, they are accordingly split into multiple assemblies or TCs. Possible interpretations of multiple TCs from a cluster are: (i) alternative transcription, (ii) paralogy or (iii) protein domain sharing.

TCs/sESTs resulting from the CAP3 step were assumed to be putative transcripts.

2.1.5 Functional annotation

Both raw EST sequences and TCs were independently annotated by an automated module. Two different database searches were performed. The first annotation procedure was carried out by BLASTx (E-value $\leq 10^{-3}$) versus the UniProtKB/Swiss-Prot database (Release 27012006; The UniProt Consortium, 2007); the second one was performed by BLASTn (E-value $\leq 10^{-5}$) versus the Rfam database (version 7.0; Griffiths-Jones et al., 2005). This second annotation procedure allows to better refine the sequence function assignment. In fact, it is well known that mRNA-like non-coding RNAs (ncRNAs) can be present in EST collections (MacIntosh et al., 2001).

In case Affymetrix Gene Array probe-sets are available for the organism under investigation, a further BLASTn analysis (E-value $\leq 10^{-5}$) was carried out to establish correspondences between the EST sequence data-set and the Affymetrix probe-sets (http://www.affymetrix.com/products/arrays/index.affx).

2.1.6 Overview of the Gene Ontologies

For a standard and controlled classification of gene products, the protein annotation was switched to the Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000) in the event that the subject UniProt identifier is recorded in a local GO database.

Gene Ontology assignments were reduced using GO slim terms in order to give a broad overview of the ontology content of each transcriptome. GO slims are, in fact, cut-down versions of the gene ontologies and contain a subset of the terms in the whole GO. The map2slim.pl script, distributed as part of the go-perl package (version 0.04), was used to convert all the GO terms related to each transcript to the plant GO slim terms (http://www.geneontology.org/GO_slims/goslim_plant.obo).

2.1.7 Overview of the ENZYME assignments

ENZYME (Bairoch, 2000) is a repository which describes each type of characterized enzyme (which an EC number has been provided to). The protein annotation was switched to the ENZYME assignments in the event that the EC number was present in the description lines of the subject UniProt hit. The release of may 2006, which comprises 4037 entries, was considered.

The association to an EC number let the expressed sequence be associated to the KEGG (Kyoto Encyclopedia of Genes and Genomes; Kanehisa et al., 2006) metabolic pathways.

2.2 The EST database

2.2.1 Implementation and architecture

The EST database architecture consists of a main MySQL relational database, where all the data generated by ParPEST were deposited, and of two satellite databases myGO and myKEGG. The Entity-Relationship (ER) diagram, which illustrates the relationships between entities in the EST database, is shown in figure 7.

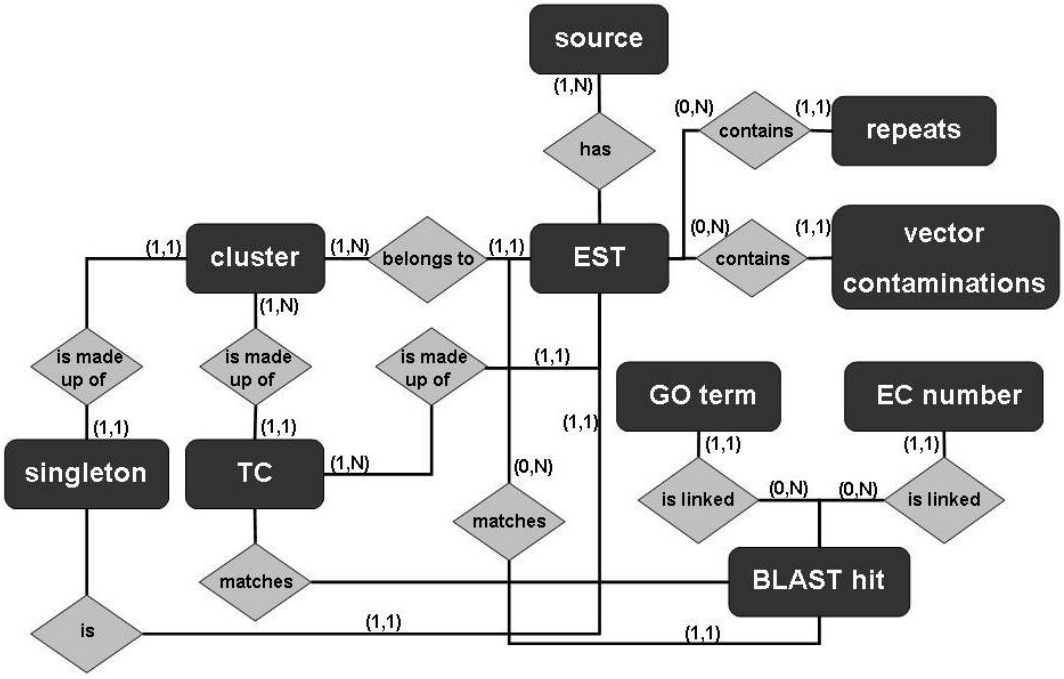


Figure 7. The Entity-Relationship (ER) diagram of the EST database.
The ER diagram is reported to show the database structure schema. The schema describes the entities and their relationships.

The myGO database, which comprises both ontology and annotation data, was built from the flat files available on the GO website (www.geneontology.org) which were downloaded in the mySQL format (go_200605-assocdb.data.gz and go_200605-assocdb.tables.tar.gz).

The myKEGG database was built by parsing the KEGG XML formatted files (ftp://ftp.genome.jp/pub/kegg/xml/map/) using the XML::XPath::XMLParser Perl module and by downloading the metabolic maps in the GIF format (ftp://ftp.genome.jp/pub/kegg/pathway/map/).

The web interface to the database was created using HTML and PHP scripts which dynamically execute MySQL queries. It operates under an Apache web server on a Fedora Linux Core 4 system.

2.2.2 Web application

The EST database web application was developed in order to support data retrieval through pre-defined query systems. Data can be inspected via three different HTML forms that allow distinctive queries on (i) EST sequences, (ii) clusters and (iii) putative transcripts (i.e. transcript indices).

The first HTML form produces an "*EST report page*" (Figure 8A) displaying each EST as a bar. The colour of the EST bar changes depending on the EST type. "Container" ESTs are blue, stand-alone ESTs are green while "contained" ESTs are not traced since their sequence information is already represented by the corresponding "container" sequence. Vector contaminations, low complexity sub-sequences and repeats are properly highlighted with colours (black, red). The EST bar is linked to the nucleotide sequence. Protein as well as ncRNA matching regions are drawn on the length of the query sequence as grey and/or brown bars respectively; each bar is linked to the details concerning the local alignments.

The second HTML form results in a "*clusters report page*" (Figure 8B) where data are presented in a summary table. In a row, are reported the cluster identifier, the number of TC(s) which the EST sequences in a cluster are split into, and the total number of the ESTs within the cluster. Via the cluster ID, the EST multiple sequence alignment constructed for each TC can be accessed. Each TC is represented as an orange bar along which the EST bars are drawn so as to reconstruct the assembly. The protein and the ncRNA matching regions are traced on the length of the query sequence (Figure 8C).

The third HTML form results in a "*transcript indices report page*" where data corresponding to the user-selected criteria are listed in a summary table. These data can be also investigated considering two different classes of objects: the enzymes and the metabolic pathways. Enzymes are classified into classes, subclasses and sub-subclasses according to the guidelines of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). They are listed as HTML-based tree menus (Figure 9A). For each enzyme are enumerated all the associated transcripts as determined by the functional annotation module.

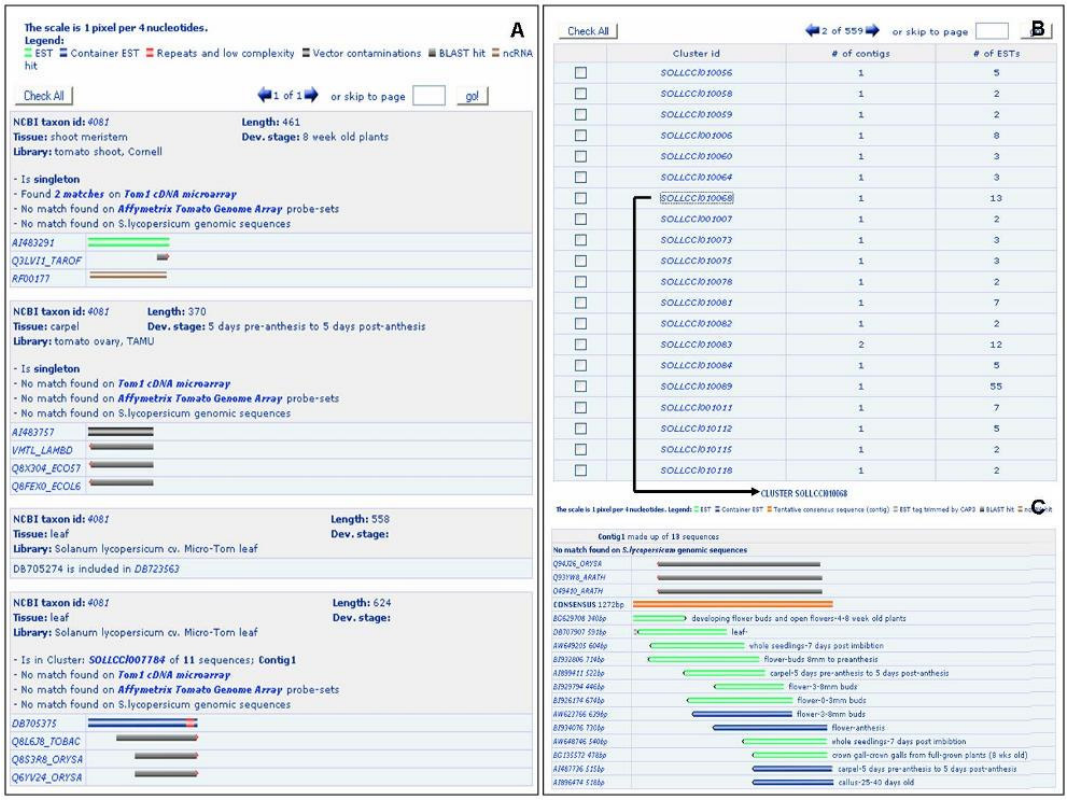


Figure 8. Screenshots of the EST database web interface.
A. 'EST report page'. EST sequences are displayed as bars whose colour depends on how the EST has been classified (the green bar corresponds to a stand-alone EST while the blue bar corresponds to a *container* EST). The black bar describes an EST whose nucleotide sequence is likely to represent a vector contamination. Low complexity sub-sequences and repeats are highlighted as red segments. Protein as well as ncRNA matching regions are traced along the EST query sequence as grey or brown bars respectively.
B. 'clusters report page'. Data are presented in a summary table where in a row are reported the cluster identifier, the number of TC(s) which the EST sequences in a cluster are split into, and the total number of the ESTs within the cluster.
C. Representation of the multiple alignment of EST reads generating a TC. TC is represented as an orange bar along which the EST bars (green or blue according to the EST type) are drawn so as to reconstruct the assembly. The protein (grey) and the ncRNA (brown) matching regions are traced on the length of the TC sequence.

Redundancy may occur because (i) more proteins are referenced in the ENZYME repository with the same EC number; (ii) different transcripts encode for different subunits of the same enzyme; (iii) different transcripts represent different segments of the same cDNA which have not been assembled because of the EST "tag" nature. Because one enzyme can contribute to more than one metabolic pathway, all the pathways which the enzyme belongs to are also enumerated in the tree menu. The class "*metabolic pathway*" is useful to investigate on a specific map and on its "*coverage*"; indeed, transcripts are mapped on-the-fly onto the pathways via the enzymes they are associated to. Each metabolic pathway represents the main node of the HTML-based tree menu; the number of the enzymes mapped as well as the number of the map-

specific enzymes are indicated. Nodes, which describe each enzyme and enumerate the transcripts associated to the enzyme itself, are added to the menu object (Figure 9B). Metabolic pathways can be always accessed as GIF images modelled as graphs where a node represents a compound and an edge represents the enzyme-catalyzed reaction. For each image the enzymes, which have been mapped on-the-fly, are highlighted in red (Figure 9C).

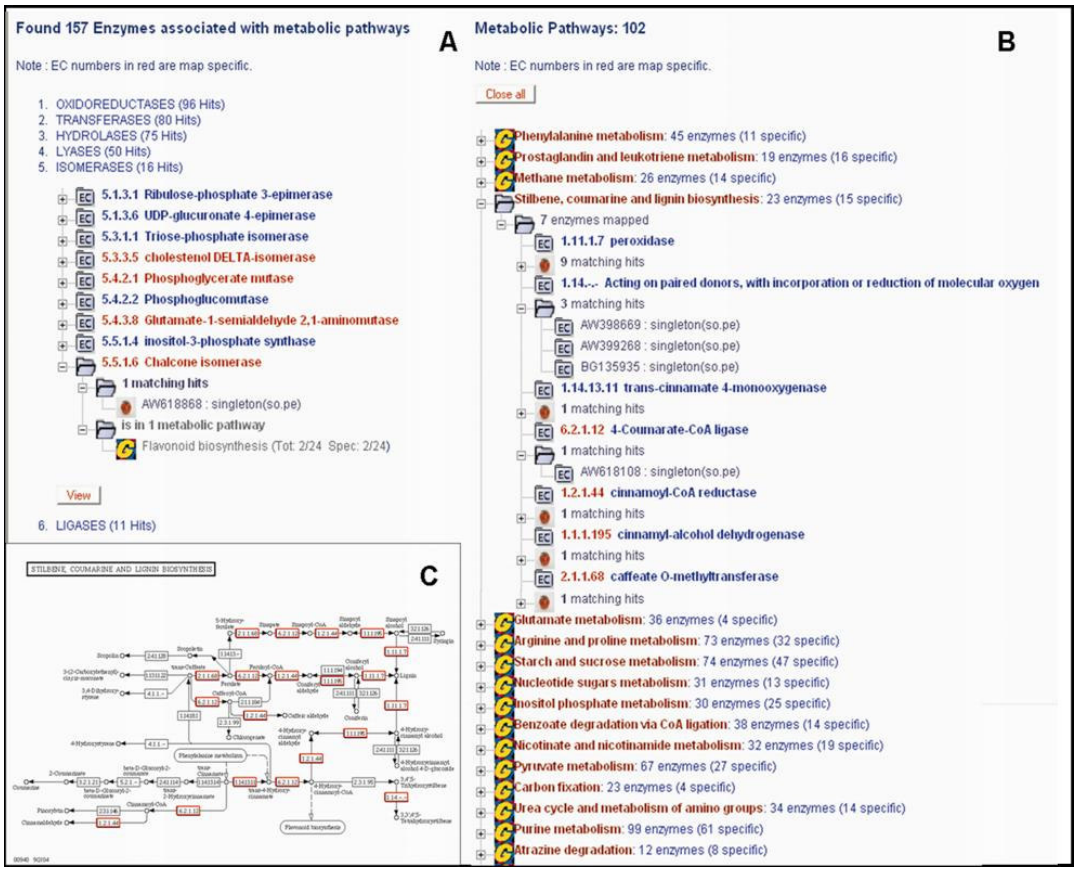


Figure 9. Screenshots of the EST database web interface.
A. An example of a tree menu listing all the metabolic enzymes annotated for a plant species is shown. The node corresponding to chalcone isomerase (EC 5.1.1.6) is expanded and shows the transcript(s) associated to the enzyme and the metabolic pathway(s) which include the enzyme.
B. An example of a tree menu listing all the metabolic pathways associated to the plant species is shown. The node corresponding to 'Stilbene, coumarine and lignine biosynthesis' is expanded.
C. The pathway image shows seven on-the-fly mapped enzymes highlighted in red.

2.3 EST-based gene discovery and gene model building

2.3.1 Setting up the Generic Genome Browser database

The Generic Genome Browser (GBrowse) is an open-source browser developed as part of the Generic Model Organism Database project (GMOD; Stein et al., 2002). It is a Web-based application for displaying DNA, protein, or other sequence features within the context of a reference sequence such as a chromosome, a BAC or a metacontig. The release "Gbrowse-1.62" was retrieved from the Sourceforge download page and installed onto a Fedora Linux Core 4 system.

GBrowse is based on the GFF file format which stands for "General Feature Format" (http://www.sanger.ac.uk/Software/formats/GFF). The GFF format is a flat tab-delimited file, each line of which corresponds to a feature (i.e. an annotation).

For smaller data-sets the GBrowse uses a file-based database (i.e. the 'in-memory' database) which allows it to run directly off text files. On the other hand, for larger data-sets the GBrowse requires a MySQL database management system. Then, because we deal with a large amount of data, we used the BioPerl utility bp_load_gff.pl to upload in the MySQL database a series of GFF and FASTA files.

2.3.2 EST-to-genome alignments

The program GenomeThreader (Gremme et al., 2005) (settings: coverage ≥ 80%; identity ≥ 90%) is used to produce EST/TC to genomic DNA spliced alignments. The alignment data were parsed and converted into the GFF format by a Perl script and subsequently uploaded into the MySQL database.

2.3.3 Gene models from ESTs

The GeneModelEST software (D'Agostino et al., 2007b) was used for defining a data-set of candidate gene models based exclusively on EST evidences. It requires two GFF formatted files which describes the *in silico* derived coordinates of EST- and TC-to-genome alignments.

The GFF files must include two features:

- the *match* feature: it indicates the full-length of the EST/TC-to-genome alignment from the start to the end coordinate.
- the *HSP* (High-Scoring Pairs) feature: it indicates the start and the end coordinates of a section of the *match* feature. In other words all the *HSPs*

belonging to a *match* feature describe all the consecutive elements representing exons. Instead, the genomic region between two consecutive *HSPs* corresponds to an intron region.

An NCBI BLAST report file, in which are recorded information on sequence similarities between TCs and proteins, is fed into GeneModelEST too.

Firstly, given the coordinates of each TC *match* feature, the TC-to-TC status is established as follows:

- overlapping TCs
- non-overlapping TCs

We need to define the HSPs of non-overlapping TCs as c-HSPs and all the EST HSPs as e-HSPs. Then, the status of each c-HSPs was established by comparing its coordinates to the ones of the e-HSPs aligned in the same genomic region. Possible results of the pair-wise comparisons are classified according to the instances shown in figure 10.

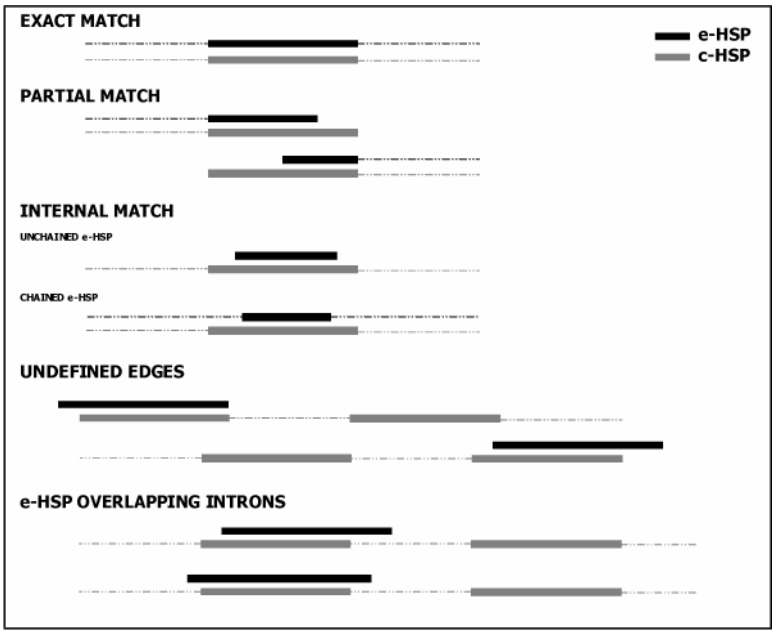


Figure 10. Representation of possible results from c-HSP and e-HSP pair-wise comparison.
i) ***exact match*****:** the start and the end positions of the e-HSP coincide with the ones of the c-HSP;
ii) ***partial match*****:** at least one of the edges of the e-HSP is exactly the same of the edges of the c-HSP. Therefore, because of EST length-limit, the other e-HSP edge is included in the region spanned by the c-HSP;
iii) ***internal match*****:** an e-HSP lies within a c-HSP. According to the e-HSP status this instance can be split in two cases: **a)** ***unchained e-HSP*****:** in case the e-HSP is completely included in the corresponding c-HSP or **b)** ***chained e-HSP*****:** in case the e-HSP is concatenated to the flanking e-HSPs;
iv) undefined edges: the e-HSP is overlapping a terminal c-HSP going beyond one of its edges. This implies that the e-HSP is describing a terminal exon longer than the c-HSP.
v) e-HSP overlapping introns: one or both edges of an internal e-HSP lie within one of the intron regions defined by aligning the TC to the genome sequence. This implies that the EST, which the e-HSP belongs to, is representing an intron retaining sequence or an alternatively spliced transcript of the same gene or the transcript of a gene which is overlapping the same locus.

The status of each c-HSP was defined as *confirmed*, *undefined* or *ambiguous* by combining all the possible instances enumerated with Boolean operators illustrated in table 4.

cHSP status	exact match		partial match		internal match				undefined edges		e-HSP overlapping introns
					Unchained e-HSP		Chained e-HSP				
Confirmed	X	**OR**	X	**OR**	X	**AND**	0	**AND**	0	**AND**	0
Undefined	X	**OR**	X	**OR**	X	**AND**	0	**AND**	X	**AND**	0
Ambiguous	X	**OR**	X	**OR**	X	**AND**	X	**OR**	X	**AND**	X

Table 4. c-HSP status determination. The determination of the c-HSP status originates from the result of all pair-wise comparisons versus all e-HSPs aligned in the same genomic region. X indicates at least one occurrence of the instance. Boolean operators have been used to define the c-HSP status.

GeneModelEST assigns a TC as a consequence of the evaluation of all its c-HSPs to one of the following classes: *optimal*, *acceptable* and *rejected*. Optimal are those TCs for which all the c-HSP are *confirmed*; acceptable are those TCs presenting at least one *undefined* c-HSP; rejected are those TCs with at least one *ambiguous* c-HSP.

Alternative gene structures must be avoided in the definition of candidate gene models. Therefore, GeneModelEST declares as candidate gene models those TCs which have been classified as *optimal* or *acceptable*. Indeed, *rejected* TCs are excluded because they represent either possible alternative splicing or intron retaining sequences, and therefore, they need a human-curated validation.

In order to assign a preliminary functional annotation to TCs, GeneModelEST evaluates the protein sequence coverage (%coverage) and the similarity threshold (%positives) of the highest scoring alignment described in the NCBI BLAST report file according to the following rules :

1. Complete TCs (coverage $\geq$ 95%) are classified as:
    a. identical to (similarity $\geq$ 90%)
    b. similar to (similarity < 90%)
2. Uncomplete TCs (50% $\leq$ coverage < 95%) are classified as:
    a. similar to (similarity $\geq$ 60%)
    b. low similarity to (similarity < 60%)
3. Undefined product: TCs with protein coverage < 50%.
4. Expressed product: TCs without BLAST matches.

2.4 Comparative analysis of the tomato and potato transcriptomes

The total protein complement of the *Arabidopsis thaliana* genome was obtained by the *.faa* files (NC_003070, NC_003071, NC_003074, NC_003075, NC_003076) retrieved from the genome session of the NCBI ftp site. In total 30,480 Arabidopsis protein sequences were collected. The available functional annotations was established by the TIGR and were reported in the description line of FASTA amino acid files.

The gene families and genes are displayed in the tab delimited file gene_family_tab_121906.txt which was downloaded from the ftp session of The Arabidopsis Information Resource (TAIR; Rhee et al., 2003). This file enumerates 996 gene families and 8,331 genes.

In order to collect all the protein-Refseq sequences which corresponded to the 8,331 genes, the file TAIR7_NCBI_mapping_prot was retrieved (ftp://ftp.arabidopsis.org/home/tair/Proteins/Id_conversions/TAIR7_NCBI_mapping_prot) and used to switch the AGI (Arabidopsis Genome Initiative) gene model IDs into the NCBI protein Refseq ID. Therefore, the Bio::DB::Query::GenBank BioPerl module was used to query GenBank in order to collect all protein sequences belonging to gene families in FASTA format. Initially, BLASTx searches were performed to identify tomato as well as potato transcripts with significant sequence similarities (e-value $< 10^{-5}$) to the Arabidopsis proteome. Then the same query sequence sets were searched against the Arabidopsis protein collection which entries are annotated and classified into families. BLASTx results were filtered for significant hits using an e-value cut-off $< 10^{-5}$. Finally tomato and potato transcripts which did not have a match in Arabidopsis were used for pair-wise comparisons (BLASTn with an e-value cut-off $< 10^{-10}$) in order to check for significant sequence similarities at the nucleotide level.

2.5 Identification of new members of the gluthatione S-transferase superfamily in *Citrus sinensis*

The schematic view of the step-by-step analysis was aimed to identify new members of the gluthatione S-transferase in *Citrus sinensis* by EST screening (Figure 11). In succession, a brief description of each module of the analysis is discussed.

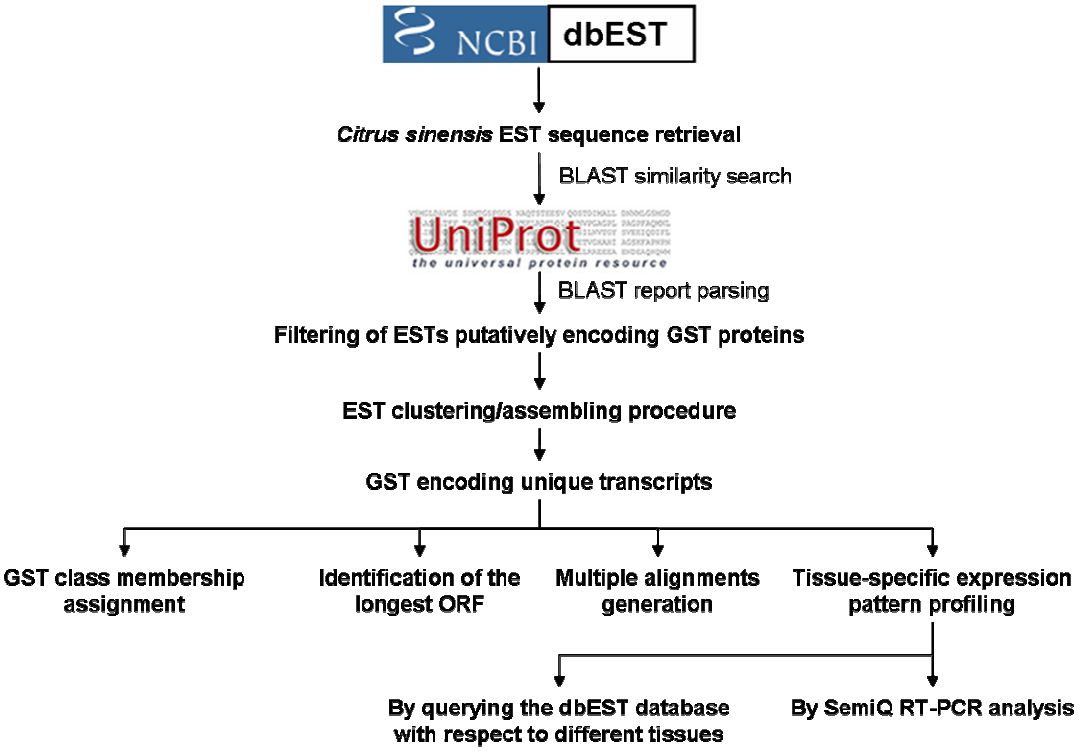


Figure 11. Step-by-step analysis procedure using EST sequences as primary data source.

2.5.1 Identification of ESTs encoding putative GST proteins

Members of the *Citrus sinensis* GST superfamily were identified by screening the EST collection retrieved from the dbEST divison of the GenBank repository (Release 01-11-06). A preliminary functional annotation was based on BLASTx comparisons of 94.127 EST sequences against the UniProtKB/Swiss-Prot database (Release 01-11-2006). The NCBI BLAST report file was parsed with an in-house Perl script in order to select orange ESTs that matched as best hit a GST protein. The original data-set was reduced to 370 putative GST encoding sequences. This collection was used to feed the clustering/assembling procedure.

PaCE (default parameter) is the EST clustering software (Kalyanaraman et al., 2003) that we used in order to group the ESTs putatively derived from the same gene. Instead, the assembly software CAP3 (Huang and Madan, 1999) (with an overlapping window of 30 nucleotides and a minimum score of 95) was used in order to assemble the short ESTs which were clustered in the same group into tentative consensus sequences (TCs). The clustering/assembling resulted in 62 putative distinct transcripts: 28 TCs and 34 singletons (sESTs) (Table 5).

GST CLASS	SEQ. ID	# ESTs	ESTs PER TISSUE	TISSUE
TAU	CITSI24:CX077363	1	1	callus
	CITSI13:CX300684	1	1	phloem
	CITSI60:BQ624512	1	1	entire seedling
	CITSI31:CN188035	1	1	pulp
	CITSI47:3	14	4	callus
			6	immature ovaries
			2	leaf, petiole, bark
			2	phloem
	CITSI47:1	20	3	callus
			4	flower
			2	flush leaves and stems
			3	immature ovaries
			2	rind
			2	seed
			4	shoot meristem
	CITSI47:2	2	1	flower
			1	flush leaves and stems
	CITSI17:CX675467	1	1	callus
	CITSI07:DY257387	1	1	mature fruit abscission zone C
	CITSI06:DY257487	1	1	mature fruit abscission zone C
	CITSI55:CB293075	1	1	rind containing flavedo and albedo
	CITSI57:1	4	2	rind containing flavedo and albedo
			1	entire seedling
			1	flush leaves and stems
	CITSI43:1	13	1	entire seedling
			2	flavedo
			2	flower
			1	flush leaves and stems
			1	leaf blade
			1	peel (flavedo)
			2	phloem
			2	pulp
			1	rind
	CITSI16:1	9	2	callus
			1	entire seedling
			1	flavedo
			2	flower
			1	immature ovaries
			2	rind containing flavedo and albedo
	CITSI53:1	10	2	entire seedling
			2	flavedo, albedo, some red scale
			1	flower
			1	immature ovaries
			1	leaves
			1	phloem
			2	rind containing flavedo and albedo
	CITSI15:DN618611	1	1	flavedo, albedo, some red scale
	CITSI41:1	5	2	entire seedling
			1	flavedo
			1	peel (flavedo)
			1	phloem
	CITSI59:1	2	1	entire seedling
			1	phloem
	CITSI61:BQ623696	1	1	entire seedling
	CITSI58:BQ624883	1	1	entire seedling
	CITSI40:CK701666	1	1	entire seedling
	CITSI25:1	5	3	entire seedling
			2	callus
	CITSI39:CK739807	1	1	entire seedling
	CITSI01:EG358290	1	1	flesh
	CITSI51:1	4	2	phloem
			1	leaves
			1	peel (flavedo)
	CITSI04:1	9	8	leaves
			1	phloem
	CITSI44:1	9	2	leaves
			1	callus
			1	entire seedling
			2	phloem
			1	rind
			2	shoot meristem
	CITSI36:1	8	2	flavedo
			2	flower
			1	immature ovaries
			1	leaves
			2	rind
	CITSI03:DY305803	1	1	leaves

GST CLASS	SEQ. ID	# ESTs	ESTs PER TISSUE	TISSUE
PHI	CITSI54:CB293267	1	1	rind containing flavedo and albedo
	CITSI52:1	94	20	callus
			9	entire seedling
			5	flavedo
			5	flavedo, albedo, some red scale
			11	flower
			2	flush leaves and stems
			4	immature ovaries
			3	leaf blade
			2	leaf, petiole, bark
			5	ovary
			1	peel (flavedo)
			6	phloem
			2	pulp
			6	rind
			3	rind containing flavedo and albedo
			7	seed
			4	shoot meristem
	CITSI38:CK934228	1	1	flower
	CITSI23:1	2	2	leaf, petiole, bark
	CITSI65:BQ623038	1	1	entire seedling
	CITSI34:CK939385	1	1	flower
	CITSI28:CX046491	1	1	flavedo
	CITSI27:CX070573	1	1	callus
	CITSI33:1	29	8	callus
			1	entire seedling
			1	flavedo
			2	flavedo, albedo, some red scale
			7	flower
			1	immature ovaries
			2	phloem
			3	pulp
			1	rind
			3	shoot meristem
	CITSI56:CB292998	1	1	rind containing flavedo and albedo
	CITSI05:DY257328	1	1	mature fruit abscission zone C
	CITSI22:CX672147	1	1	leaf, petiole, bark
	CITSI11:1	9	1	entire seedling
			1	leaf, petiole, bark
			6	phloem
			1	phloem
	CITSI02:1	11	11	flesh
	CITSI00:1	11	11	flesh
	CITSI29:1	4	2	flavedo
			1	flower
			1	callus
	CITSI35:CK936125	1	1	flower
	CITSI08:CV716584	1	1	callus
	CITSI18:1	6	6	callus
	CITSI26:1	2	2	callus
LAMBDA	CITSI63:BQ623555	1	1	entire seedling
	CITSI46:1	23	1	entire seedling
			4	flavedo, albedo, some red scale
			3	flower
			2	flush leaves and stems
			1	immature ovaries
			1	phloem
			4	pulp
			4	rind
			3	rind containing flavedo and albedo
	CITSI32:CN187483	1	1	pulp
	CITSI42:CK665050	1	1	entire seedling
	CITSI37:CK935114	1	1	flower
	CITSI30:CV884511	1	1	fower
THETA	CITSI49:CF504122	1	1	immature ovaries
	CITSI48:CF506057	1	1	immature ovaries
ZETA	CITSI21:1	3	2	leaf, petiole, bark
			1	entire seedling
	CITSI50:1	7	2	immature ovaries
			2	rind containing flavedo and albedo
			2	seed
			1	flavedo, albedo, some red scale
	CITSI62:BQ623695	1	1	entire seedling
MAPEG	CITSI20:1	11	1	entire seedling
			2	flavedo, albedo, some red scale
			4	flower
			2	leaf, petiole, bark
			1	pulp
			1	seed
	CITSI09:1	3	2	entire seedling
			1	callus

Table 5. List of the 62 GST-encoding putative transcripts. In a row are reported the GST class, the sequence ID, the total number of EST sequences which have been assembled to generate the transcript, the number of ESTs grouped per tissue which the ESTs have been derived from.

The collection of the 62 transcripts was compared against the GenBank non-redundant nucleotide database to establish if some of the 62 sequences could have been further

extended. This allowed to concatenate the TCs CITSI00:1 and CITSI02:1 into a unique transcript which corresponds to the sequence DQ198153 from GenBank (Lo Piero et al., 2006).

2.5.2 GST class assignment

Entrez query was carried out to retrieve all the *Arabidopsis thaliana* protein sequences belonging to the GST class Tau (resulting in 29 different sequences), class Phi (20 sequences), class Zeta (3 sequences), class theta (2 sequences), class Lambda (6 sequences) and class MAPEG (1; Membrane-Associated Proteins involved in Eicosanoid and Glutathione metabolism). All the protein sequences in each class were analysed by Block Maker (Henikoff and Henikoff, 1997), a tool for the identification of conserved blocks (i.e. segments corresponding to the most highly conserved regions of proteins) in a set of related sequences. An embedded consensus sequence for each of the GST classes was generated using COBBLER (COnsensus Biasing By Locally Embedding Residues; Henikoff and Henikoff, 1997). These COBBLER-embedded sequences were used as a reference to classify the putative *Citrus sinensis* GST sequences into specific GST classes.

2.5.3 Open Reading Frame finding

The EXPASY Translate tool (http://www.expasy.ch/tools/dna.html) was used to define the longest Open Reading Frame (ORF) for each GST putative encoding transcript.
The transcripts defined as full-length mRNAs (FL) are the ones showing a complete Open Reading Frame (ORF). The remaining transcripts exhibited partial ORFs. Those including the start triplet ATG but lacking the stop codon were classified as *5' fragments* (5F). On the other hand those lacking the initiating ATG but presenting a termination triplet were classified as *3' fragments* (3F). Finally, the transcripts which show interspersed stop codons were classified as "no good ORFs" (NGO)(Figure 12).

2.5.4 Multiple alignments generation

The ClustalW program (Larkin et al., 2007) was used to generate multiple alignments of the nucleotide (mRNA) sequences for each GST class. Transcripts, which no good ORFs are detected in, were also included in the multiple alignments. It happened because they share similarities with the remaining sequences in the corresponding class despite of insertions/deletions putatively due to sequencing errors. The alignment editor BioEdit (Hall, 1999) was used to edit multiple alignments in order to define the mRNA structure and to identify the mRNA segments (5' UTR, codig exons, 3'UTR).
Segment-to-segment DNA distances were calculated using the DNADIST program in PHYLIP (Felsenstein, 1993). This, in the attempt to further group sequences within a GST class.
The cut off of 20% divergence was used to define 2 segments as closely related. They show the same colour in figure 12. Indeed, a deeper analysis of the multiple alignment highlighted different subgroups of sequences per class. Each sequence in a subgroup still presented nucleotide heterogeneity that is hardly referred to sequencing errors.

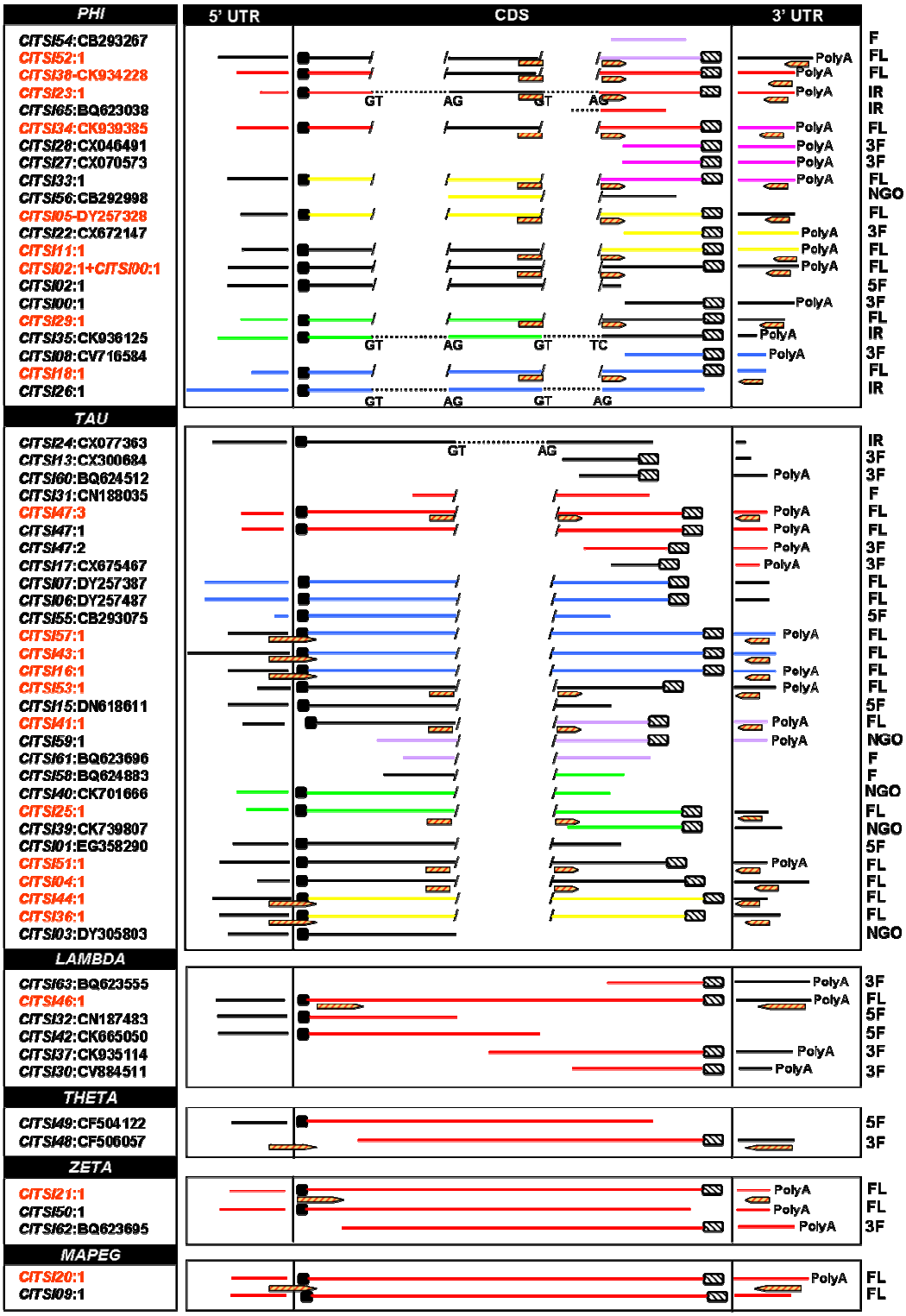


Figure 12. Schematic view of the 6 different multiple alignments generated by grouping the 62 transcripts according to GST class membership. The coding exons, the 5' and 3' UTRs (mRNA segments) are reported. Closely related segments (DNADist cut off of 20% divergence) are reported with the same colour. In red we marked the ID of those sequences to be analysed by SemiQ RT-PCR. Primers (zebra bars) and their localization along the corresponding GST transcript are shown. The last column describes each GST sequence as full length (**FL**), 5' fragment (**5F**), 3' fragment (**3F**), fragment (**F**), no good ORF (**NGO**), intron retaining (**IR**).

2.5.5 Total RNA extraction and gene expression analysis by SemiQ RT–PCR

This part of the analysis was performed by the Dr. Concetta Licciardello, from the ISAGRU Institute.

Tissue samples were collected from the Moro nucellare 58-8D-1 and the Blonde Cadenera cultivars. The TRIzol® LS Reagent (Invitrogen, Scotland UK) was used for extracting total RNA from 3g of flesh tissue while guanidine hypochlorite was used for the RNA extraction from 2g of albedo, flavedo, young and adult leaf tissues.

The RNeasy plant mini kit (Qiagen) was used to isolate total RNA from 0.1g of ovary.

The amount and quality of the total RNA were estimated by spectrophotometer readings and by agarose gel electrophoresis (0.8% agarose in 1x TAE). The electrophoresis gels were stained with ethidium bromide.

Semi-Quantitative Reverse Transcription Polymerase Chain Reaction (SemiQ RT-PCR) analyses were performed to assess the expression level of the putative GST genes in the albedo, flavedo, flesh, young and adult leaf tissues and ovary.

SuperScript III One-Step RT-PCR with Platinum Taq (end point) (Invitrogen) was used. The amplification of the Elongation Factor (EF) alpha chain (AY498567) RNA was used as control. The primers used to amplify the target regions are shown in table 6 where in a row are reported the sequence, the annealing temperature and the expected size of the amplicon. The oligonucleotide primers were designed according to different criteria in order to avoid non-specific amplifications. All the reverse primers were designed into the 3' UTR regions close to the polyA tail (Figure 12). Since the Phi class multiple alignment includes sequences that we classified as intron-retaining (marked as IR in Figure 12), the forward primers were selected straddle the second and the third coding exons.

Considering that an intron-retaining sequence was included in the Tau class multiple alignment too, some of the forward primers were designed straddle the first and the second coding exons (Figure 12).

Sequence ID	Primer features			
	Forward (5'-3')	Reverse (5'-3')	Ta (°C.)	Amplicon bp
CITSI52:1	GGCCTTCCTTTCTTTGAATCCATTC	TTTTGATAAACCCATTGGGACAGTCGT	60	835
CITSI38-CK934228	GTACCTCAAATTGCAGCCTTTCGGA	TTTTCTCCCAAGGCCCCAAGCATTT	63	659
CITSI23:1	ACGTTATACGGTAGAATCTCGAGCTATCA	TTTTGTCCTCCAAGGCCCCAAGCAT	63	610
CITSI34-CK939385	AGTACCTCAAATTGCAGCCTTTCGGT	TTTTGTCCTCCAAGGCCCCCAGCAT	64	655
CITSI33:1	TTTATACGAGTCGCGAGCTATCATGAGGT	TTTTAAAGCTCCAACTCCAAACAT	60	658
CITSI05-DY257328	ATTTTATACGAGTCGCGAGCTATCATGAGG	ACCCCTTATCCAAGGAACATTTCCCA	64	565
CITSI11:1	TGGGGATTTTACTCTATACGAATCGCGA	ATGGCGACAACAAGAAATCGCCGCA	63	640
CITSI29:1	TCTGAAGATCCAGCCCTTTGGCCAA	TGGGAAATTATTAGACCATGCCA	60	732
CITSI18:1	TCTTGCCAAGAATCCCTTCGGTCA	CATCAATGTAAAATCATCACGCAACCA	60	569
DQ198153(CITSI02:1+CITSI00:1)	TCGAGGGCAATCATAAGGTACTACGCAGC	GATAACAGTAATGACAGCCAGCCGAA	55	642
CITSI47:3	TCGCCAAGCCCATTTGTGATGAGGGCA	ACGAGACAGGCTGCTGCCTAGTCCGA	66	866
CITSI57:1	AGTAAGCTTCTGTAATAATGGCGGACGA	ACAATACCCTAAGATAACAGTCGGGGACA	64	879
CITSI43:1	TCTGTCACAATGGCGGACGAAGTGGT	AGCAGGCAGCACGATTGCGCTGCT	68	732
CITSI16:1	TCACTCGCCCTTAATTCTCAGTAAGGT	AGATTGACGCCACATAATATTCCCA	60	966
CITSI53:1	TGCTGGGTTACTGGGCAAGCCCCT	ACCTTCATGCATGGGCAACCGCTGA	66	686
CITSI41:1	GTTTACAGGGTGATTTGGGCTCTGA	ACCACTATGCTAGTCCCCCGAACT	63	757
CITSI25:1	GTTCATCGACGAAAAGCTGTTGGCA	ACACAGAGAGAGAGCTAACCCAATCA	63	449
CITSI51:1	TGGCCAAGCCCGTTTGTGTTTAGGGT	AGACTTTCCACACAACATCACACTAC	63	756
CITSI04:1	AGACGTGGTCAAGCCCCTTTGGT	TGGAATGGGAAAAGGGCAAAAGGA	62	889
CITSI44:1	GCAGAAGATTATGGCAACAAAAGTG	GAGCGTACAGAAAGGAGACACGTGCA	62	764
CITSI36:1	GCGAAAATAATATGGCCAAAGAAGTGACGCT	GTCATTACAACACACCACAACACCACCT	64	765
consensus CITSI48-CITSI49	TGGGTGGGCTAAAGAAAGGAAAATGAAGC	TGCGGAACATATAGGCAACATTGAAACCT	64	464
CITSI21:1	ATGCTGAAACTGTATTCATACTGGAGGAGT	TGCTGCTTATTGAGGGTCAACAAAGGCTG	64	880
CITSI46:1	CCTCCAAGATAGGCCCGCTTGGTAC	TCCAGCAACGTACACAAGCTCACATCGGCA	66	672
CITSI20:1	CGACTCGACTATGGCGGATGCAAC	CTATGAGCTTATGCTTGCGCCATGCAGC	66	638

Table 6. List of sequences and primers designed to perform Semi-Quantitative RT-PCR experiments.

This to ensure the amplification of mature GST transcripts. In the remaining cases, anyway, the forward primers were selected straddle the 5' UTR and the ATG initial codon because of high sequence variability in the 5'UTR regions (Figure 12). The same criterion was assumed for the selection of the forward primers of the Theta as well as the MAPEG class. For the amplification of Lambda and Zeta class sequences, the forward primers were selected immediately after the ATG initial codon (Figure 12).

For RT-PCR reactions, first-strand cDNA was synthesized from 200 ng of total RNA in a volume of 25 µl containing 1x PCR Reaction mix, 0.2 µM of each target-specific amplification primer, 1U of SuperScript III One-Step RT-PCR with Platinum Taq. Reverse transcription was performed at 50°C for 30 min., followed by PCR amplification: denaturation at 94°C for 6 min followed by 35 cycles of denaturation at 94°C for 30 s, annealing for 30 s, and extension at 72°C for 120 s, with a final extension at 72°C for 7 min, in a GeneAmp PCR system 9700 (Applied Biosystems, Foster City, CA, 94404). The amplified DNA samples were separated by agarose gel electrophoresis (1.5% agarose in 1x TAE) and stained with ethidium bromide.

3 RESULTS

3.1 ParPEST efficiency

The ParPEST pipeline analyses large amount of EST data and it is based on distributed computing. Jobs are scheduled according to the available resources.

The free release of PaCE (Kalyanaraman et al., 2003) that we experienced to be restricted to 30.000 sequences has been replaced with the updated version provided by the authors, who successfully tested the software with more than 200.000 sequences (personnel communication). Therefore, the sole constraining factor appears to be the memory requirements for data storage.

The ParPEST "analysis pipeline" has been tested on a cluster of 8 single processor computing nodes. Tests were performed considering different dataset dimensions (randomly selected ESTs) and different cluster configurations (4, 6 and 8 nodes).

Table 7 reports the global execution time as well as the ones calculated for each of the main steps of the ParPEST pipeline.

	#sequences	BLAST on ESTs	Pre-processing	Clustering	Assembling	BLAST on TCs	TOT
4 nodes	250	3712	441	15	201	501	4870
	500	7072	613	15	201	441	8342
	1000	13643	857	30	202	1474	16206
	5000	70490	2979	150	257	6806	80682
	10000	14559	6029	346	328	16045	168287
6 nodes	250	1992	441	15	201	350	2999
	500	3648	443	15	201	421	4728
	1000	6911	847	30	212	903	8903
	5000	35647	2834	136	268	4137	43022
	10000	72525	5483	240	357	7845	86450
8 nodes	250	1600	441	15	201	280	2537
	500	2517	443	15	202	461	3910
	1000	4704	797	30	212	733	6476
	5000	23819	2784	121	267	2853	29844
	10000	48700	5377	240	357	7845	62519

Table 7. ParPEST execution time (in seconds) calculated considering different dataset dimensions as well as different cluster configurations. Execution time is enumerated for each step of the pipeline: **1) BLAST on ESTs:** functional annotation of raw EST sequences; **2) Pre-processing:** vector contaminations cleaning and low complexity and interspersed repeat masking; **3) Clustering; 4) Assembling; 5) BLAST on TCs:** functional annotation of consensus sequences**. 6) TOT:** is the ParPEST global execution time.

As it is evident, the global execution time of the pipeline are strongly dependent on the MPI-BLAST analyses. As a consequence, ParPEST 's behaviour - in terms of scalability and performance - is greatly biased by BLAST searches both on single ESTs and on TCs.

As expected, the larger the dataset to be analysed (>1000 ESTs), the wider the execution time decrease observed at the increase of the number of computing nodes

(Figure 13). In case of dataset of small size, execution time continue to be the same despite the cluster configuration is different. This is due to the overhead time that the software spends for the job scheduling. A better evaluation of overhead time is reported in figure 14 where the average running time per sequence is reported at varying the cluster configuration. As the number of the ESTs increases, the timing profiles flatten out since the average response of the system becomes more stable due to the reduction of the overhead effect.

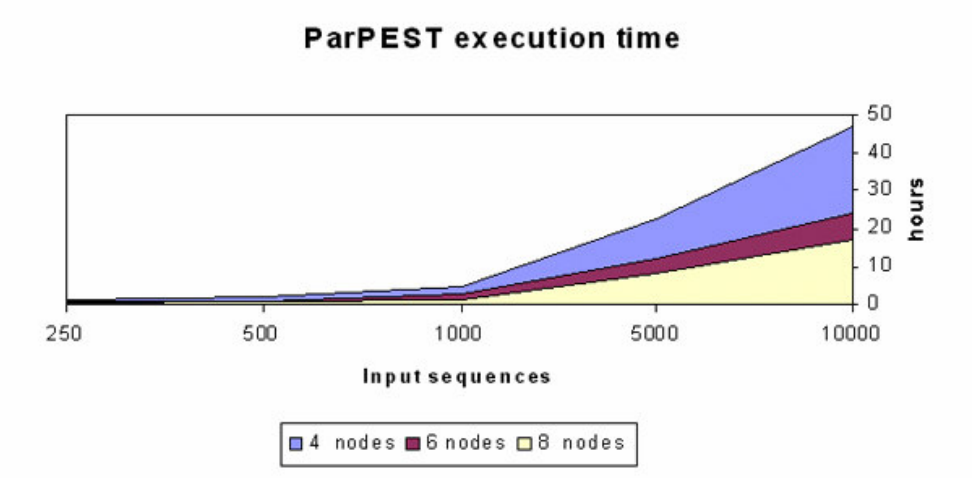


Figure 13. ParPEST global execution time. Execution time (in hours) is calculated considering different dataset dimensions and an increasing number of computing nodes (from 4 to 8).

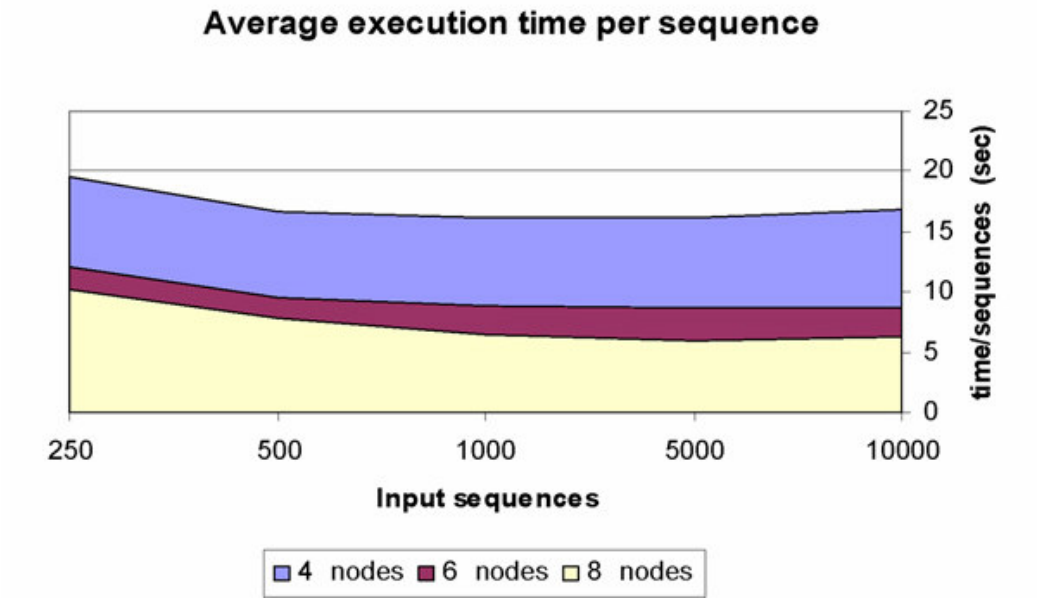


Figure 14. Average execution time per sequence. The average execution time (in seconds) is calculated considering different dataset dimensions and an increasing number of computing nodes (from 4 to 8)

3.2 Characterizing the tomato and the potato transcriptomes

3.2.1 The tomato data-set

In total, 267,906 tomato sequences were available into the dbEST session of the GenBank repository until January 2007. The most economically important species of the Lycopersicon subgenus, *Solanum lycopersicum,* was the source of the majority of the sequences having 250,552 ESTs derived from 102 cDNA libraries representing 29 distinct cultivars. The depth of sequencing of *S. lycopersicum* cDNA libraries ranged from 1 to 30,569 EST sequences (data not shown).

1,008 EST sequences from a crossbreeding between *S. lycopersicum* and *S. pimpinellifolium* (also known as the currant tomato, a species of small tomato native to South America) are collected. The remaining tomato species, *S. pennellii* and *S. habrochaites*, which are close wild relative of *S. lycopersicum* from South America, are represented by 16, 346 ESTs (data not shown).

These libraries represent 21 distinct tissue types covering both the sexual reproductive and the vegetative parts of the plant (Figure 15).

TomatEST (D'Agostino et al., 2007a) is the secondary database which collects and organizes as tomato primary data from dbEST as the ParPEST-processed data.

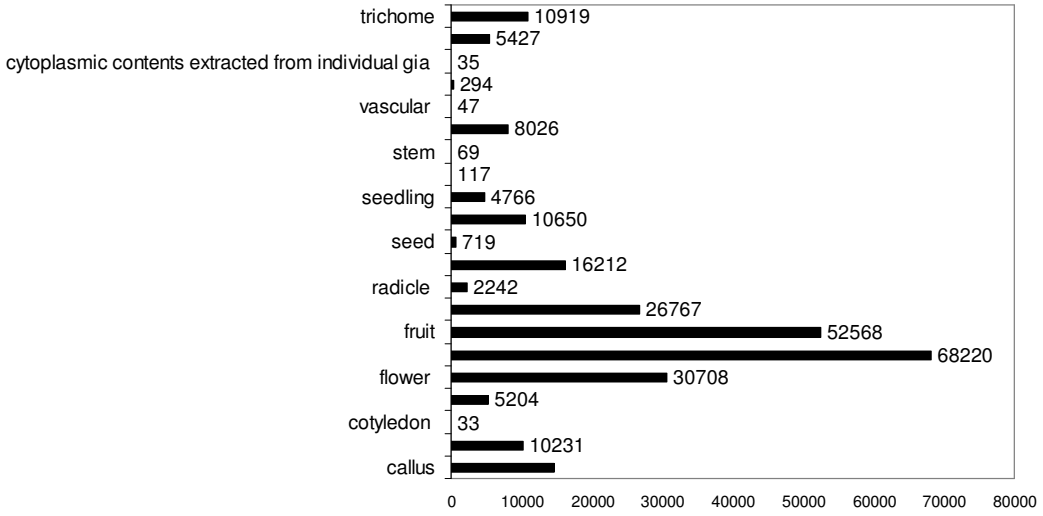


Figure 15. Tissue-based classification of the EST libraries. The number of EST sequences per tissue type is reported.

3.2.2 Building of tomato unigene sets

In order to generate an unigene set for each tomato species, ESTs were fed into the ParPEST. The first module of the procedure (see Methods 2.1.2) involves the

identification of over-represented ESTs and their removal from the original data-set. The *S. lycopersicum* EST collection was reduced of the 24% overall; the *S. pennellii* data-set was reduced of the 17.5% overall, while no significant decrease was observed in the remaining collections (Table 8). For each data-set, the number of sequences was further reduced because of the presence of non-native sequences such as vector contaminations (see Methods 2.1.3) (Table 8).

SOURCE	Total ESTs	nr ESTs	% decrease	container ESTs (long parent sequences)	contained ESTs (duplicates)	vector-clean ESTs
S. lycopersicum	250552	190763	23,86	22873	59789	190593
S. pennellii	8346	6888	17,47	442	1458	6887
S. habrochites	8000	7868	1,65	89	132	7859
S. lycopersicum X S. pimpinellifolium	1008	979	2,88	19	29	977

Table 8. Summary of quality control of tomato EST data-sets.
In a row, the number of EST sequences in the raw data-set; the number of non-redundant (nr) ESTs (after the duplicates removal module); the decrease as a percentage of the original EST data-set; the number of ESTs classified as *container* (i.e. long parent sequences); the number of ESTs classified as *contained* (i.e. duplicates or children sequences); the number of EST in clean data-set (after procedures detailed in the pre-processing module) are reported per each species.

Then, the vector-cleaned data-sets were submitted to the clustering/assembling module (see Methods 2.1.4) in order to incorporate overlapping ESTs that tag the same gene in a single cluster (i.e. gene index) and generate a tentative consensus sequence (TC) per putative transcript. ESTs which did not meet the criteria to be clustered with any other EST in the collection were classified as singleton ESTs (sESTs). A summary of the composition of each gene index is shown in table 9.

SOURCE	Gene indices	TCs	Average TC length	sESTs	Average singleton length	Total putative transcrpts
S. lycopersicum	44759	17629	913,10	28005	446,3	45634
S. pennellii	3863	730	679,84	3140	464,07	3870
S. habrochites	4101	907	890,48	3203	547	4110
S. lycopersicum X S.pimpinellifolium	744	94	475,45	650	345,78	744
	53467	19360		34998		54358

Table 9. Summary of the gene index for each tomato species.

Multiple TCs split from the same cluster. In particular, 658 *S. lycopersicum* clusters are assembled into multiple TCs, ranging in size from 2 to 25 members (Figure 16).Possible interpretations are: alternative transcripts, shared protein domains or paralogy.

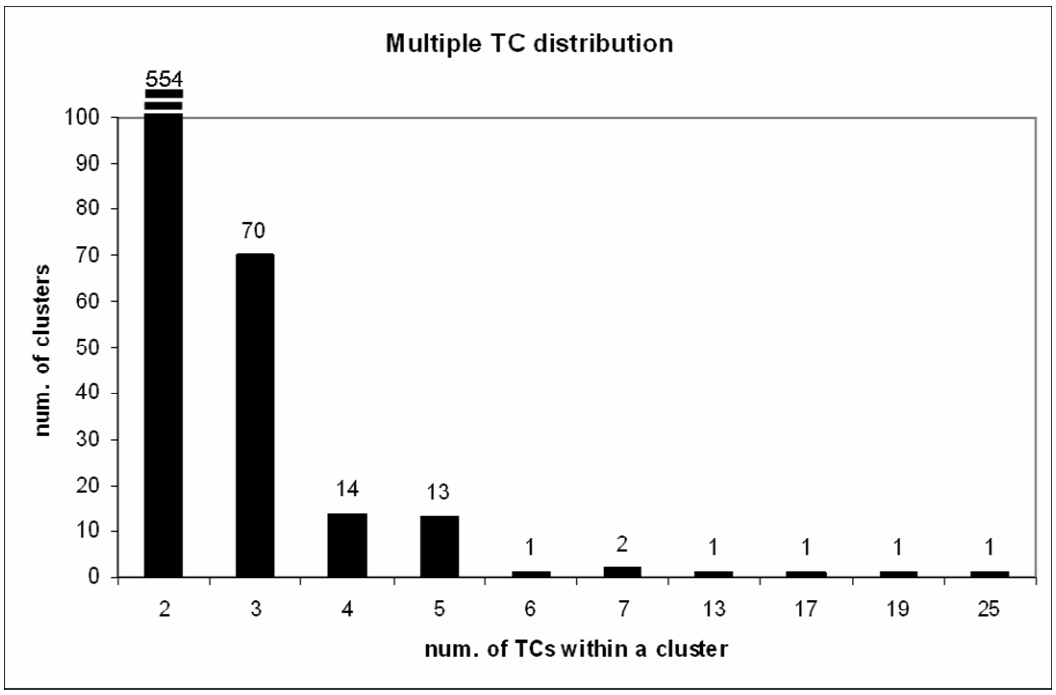


Figure 16. Number of *S. lycopersicum* clusters which are split into multiple TCs.

3.2.3 Functional annotation of the tomato unigene sets

Comparisons against the UniProtKB/Swiss-Prot database were performed using BLASTx, as part of the ParPEST pipeline, in order to assign a putative function to each transcript (see Methods 2.1.5).

About 73% (33,174 sequences) of the *S. lycopersicum* transcripts showed significant similarities to proteins in the UniProt database. 8,955 transcripts (27%) are similar to proteins which have been annotated as hypothetical, unknown or expressed proteins. This uninformative result is not surprising considering that a great number of sequences in the UniProt database represents uncharacterized proteins. BLASTx analysis for the remaining tomato species revealed similar trends as shown in table 10.

SOURCE	Total putative transcrpts	Transcripts showing similarity to UniProtKB proteins		Transcripts showing similarity to uninformative UniProtKB proteins	
			% to total transcripts		% to annotated transcripts
S. lycopersicum	45634	33174	72,69	8955	27,00
S. pennellii	3870	3042	78,60	659	21,66
S. habrochites	4110	3446	83,84	683	19,82
S. lycopersicum X S.pimpinellifolium	744	606	78,29	57	7,36

Table 10. Summary of the highest scoring BLAST hits against the UniProtKB/Swiss-Prot database. For each tomato species, uninformative BLAST hits are counted and the percentage to annotated transcripts is reported.

In addition, a more detailed functional annotation was provided by mapping transcripts to the GO hierarchy (see Methods 2.1.6). The Gene Ontology provides a controlled vocabulary to describe gene products. It is organized into three ontology areas which are considered independent from each other: molecular function, biological process and cellular component. For this reason, multiple gene ontology term assignments are possible. In total, 13,361 tomato transcripts are associated to one or more ontology terms. Thus, 312,920 assignments were made to the molecular function, 213,834 to the biological process and finally 157, 326 to the cellular component class. To give a broad overview of the ontology content without the detail of the specific fine grained terms, the entire set of the ontologies was mapped onto the plant GO slims terms.

The GO slim assignments for the four tomato species are shown in figure 17.

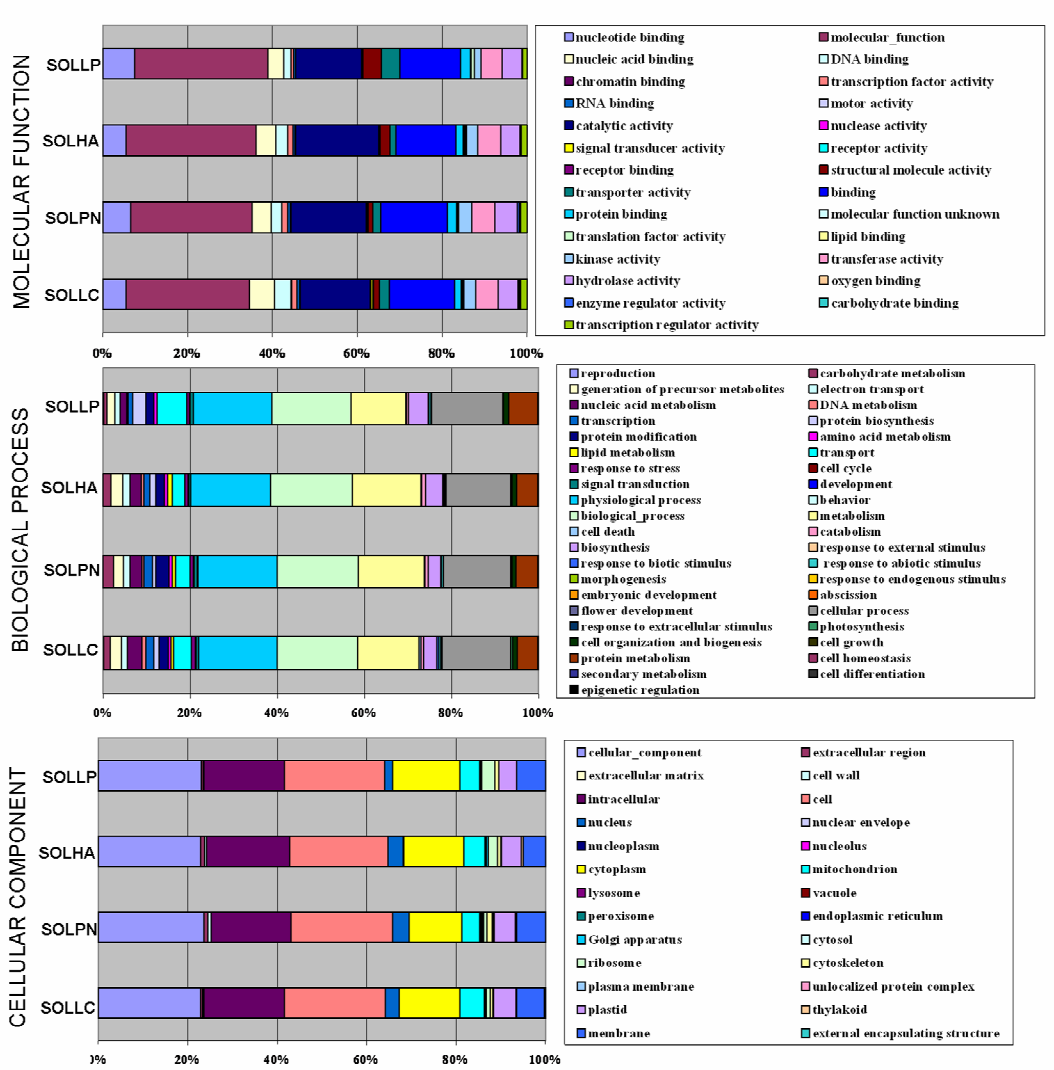


Figure 17. Assignment of Gene Ontology terms to tomato unique transcripts. Plant GOSlim terms were assigned to the four tomato species in the 3 GO indicated areas**. SOLLP:** *S. lycopersicum X S. pimpinellifolium*; **SOLHA:** *S. habrochaites*; **SOLPN:** *S. pennellii*; **SOLLC:** *S. lycopersicum*.

In the *molecular function* area, the largest functional categories are molecular function (28.5-31%), catalytic activity (15.5-19.5%), binding (14-15.5%), transferase (5-5.5%) and hydrolase activity (4.5-5.25%), while the remaining categories are less represented.
Considering the *biological process* area, the vast majority of the GO assignments correspond to the more generic physiological process (18%), biological process (18-19%), cellular process (15-16.4%) and metabolism (12.5-15.75%) categories. A large functional category is the protein metabolism (4.5-6.5%) too. The 2% of the GO annotations describes responses to biotic or abiotic stimula and to stress. This is not surprising since a division of the collected EST was derived from tissues/libraries responding to plant pathogen challenge or salt stressed.
Finally, if the cellular component, intracellular and cell categories are neglected, the remaining assignments for the *cellular component* area were to the nucleus (1.5-3.5%), cytoplasm (12-15%), mitochondrion (4-5.5%), plastid (3.5-5%) and membrane (5-6.5%).

3.2.4 The potato data-set

In total, 234,557 potato sequences are recovered from dbEST (release January 2007) to be processed. The most of the collected sequences are from *Solanum tuberosum*, which is the world's most widely grown tuber crop, and the fourth largest food crop in terms of fresh produce after rice, wheat and maize. They amount to 226,805 ESTs which are split into 68 cDNA libraries representing 18 cultivars. The depth of sequencing of *S. tuberosum* cDNA libraries ranged from 1 to 20,758. All the libraries represent 29 tissue types. Of course the more represented tissue is the tuber (~ 25% of the whole collection). Indeed the tissue sources used for library construction and sequencing largely reflected the various agronomic usages and research foci of the potato species (Figure 18).
The remaining 7,752 ESTs are from *Solanum chacoense* a wild species related to the cultivated potato (i.e. *Solanum tuberosum*). This species is indigenous to northern Argentina and the surrounding areas. It is of interest to plant breeders because some individuals produce foliar-specific leptine glycoalkaloids which seem to confer resistance to the Colorado potato beetle (Lorenzen et al., 2001).
All the processed potato EST sequences are compiled into the secondary database PotatEST (http://biosrv.cab.unina.it/potatestdb).

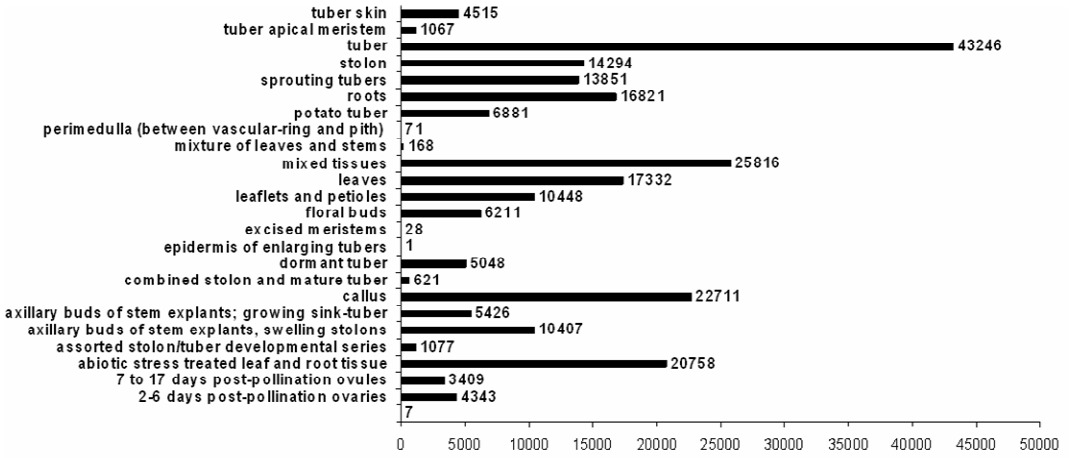


Figure 18. Tissue-based classification of the potato EST libraries. The number of EST sequences per tissue type is reported.

3.2.5 Building of potato unigene sets

Each EST collection was independently subjected to ParPEST. The first module of the analysis pipeline (see Methods 2.1.2), which has the job of identifying and removing over-represented ESTs from the original data-set, produced the results summarized in table 11. The *S. tuberosum* EST collection was reduced of the ~ 9% overall, while no significant decrease was observed in the *S. chacoense* collection (Table 11). The number of sequences in each data-set was further reduced by performing the vector cleaning step of the pre-processing module (see Methods 2.1.3) (Table 11).

SOURCE	Total ESTs	nr ESTs	% decrease	container ESTs (long parent sequences)	contained ESTs (duplicates)	vector-clean ESTs
S. tuberosum	226805	206696	8,87	14642	20109	206462
S. chacoense	7752	7750	0,03	2	2	7651

Table 11. Summary of quality control of potato EST data-sets.
In a row, the number of EST sequences in the raw data-set; the number of non-redundant (nr) ESTs (after the duplicates removal module); the decrease as a percentage of the original EST data-set; the number of ESTs classified as *container* (i.e. long parent sequences); the number of ESTs classified as *contained* (i.e. duplicates or children sequences); the number of EST in clean data-set (after procedures detailed in the pre-processing module) are reported per each species.

Then, the vector-cleaned data-sets were fed to the clustering/assembling module (see Methods 2.1.4) in order to incorporate fragmented copies (i.e. ESTs) of the same gene in a single cluster (i.e. gene index) and generate a tentative consensus sequence (TC) per putative transcript. A summary of the composition of each gene index is shown in table 12.

SOURCE	Gene indices	TCs	Average TC length	sESTs	Average singleton length	Total putative transcrpts
S. tuberosum	62752	19732	962,47	44138	565,34	63870
S. chacoense	7192	306	762,01	6886	819,1	7192
	69944	20038		51024		71062

Table 12. Summary of the gene index for each potato species.

Multiple TCs are assembled from the same cluster as a consequence of alternative transcription, shared protein domains or paralogy. In particular, 903 *S. tuberosum* clusters are split into multiple TCs ranging in size from 2 to 13 members (Figure 19).

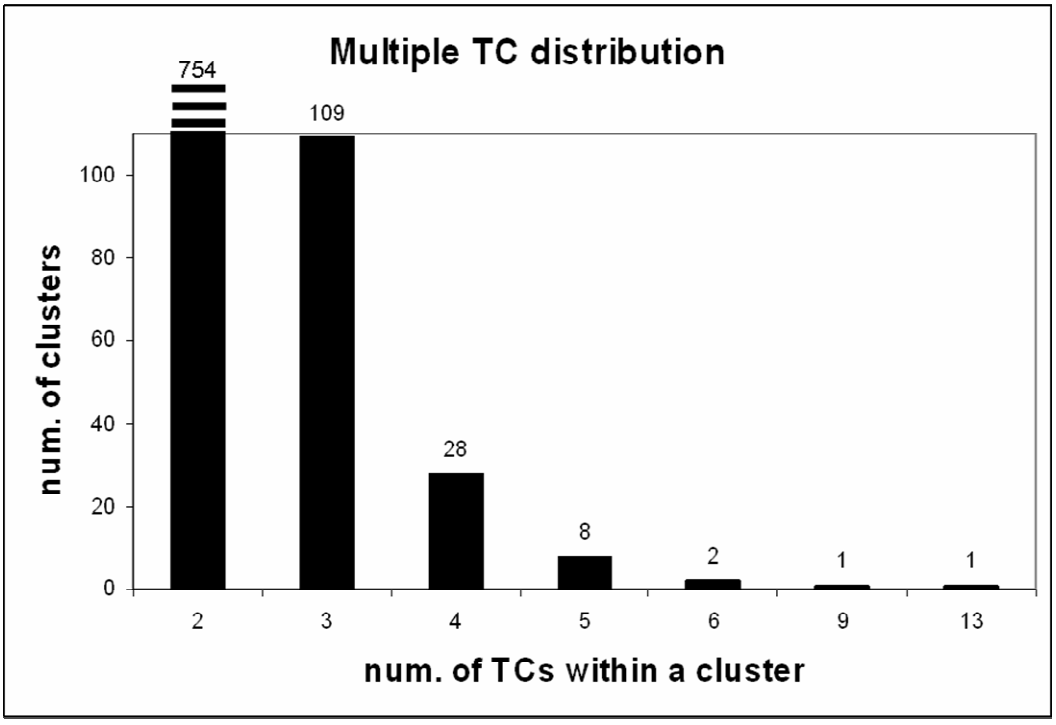


Figure 19. Number of *S. tuberosum* clusters which are split into multiple TCs.

3.2.6 Functional annotation of the potato unigene sets

The functional annotation module (see Methods 2.1.5 and 2.1.6),which is part of the ParPEST pipeline, assigned a putative function to each transcript. About 70% (44,547 sequences) of the *S. tuberosum* transcripts showed significant similarities to proteins in the UniProtKB/Swiss-Prot database. 11,392 transcripts (25%) are similar to proteins which have been annotated as hypothetical, unknown or expressed proteins. As already mentioned in the paragraph 3.2.3, these uninformative results are not surprising but highlight the still limited information available in the databases.

Furthermore, functional annotation analysis for the wild potato species *S. chacoense* revealed similar trends as shown in table 13.

SOURCE	Total putative transcrpts	Transcripts showing similarity to UniProtKB proteins		Transcripts showing similarity to uninformative UniProtKB proteins	
			% to total transcripts		% to annotated transcripts
S. tuberosum	63870	44547	69,75	11392	25,57
S. chacoense	7192	5603	77,91	1274	22,74

Table 13. Summary of the highest scoring BLAST hits against the UniProtKB/Swiss-Prot database. For both potato species, uninformative BLAST hits are counted and the percentage to annotated transcripts is reported.

In addition, in order to enhance the annotations, the biological functions associated to each transcript are converted into Plant GO slims terms (see Methods 2.1.6).

GO assignments for both potato species are shown in figure 20.

Considering the *molecular function* area, the largest functional categories, as already observed in case of the tomato GO assignments, are molecular function (29%), catalytic activity (16%), binding (10 and 15% respectively), transferase (5%) and hydrolase activity (5.3 and 4.3%). Moreover the nucleic acid binding (6%) and the nucleotide binding (6%) categories are also significantly represented

Also in the case of the *biological process* area the vast majority of the GO assignments reproduces the tomato behaviour. Therefore, physiological process (18%), biological process (18.4%), cellular process (15.7%), metabolism (14%) and protein metabolism (5-6%) are the larger functional categories. A similar trend for potato GO *cellular component* assignments is observed with respect to the tomato ones.

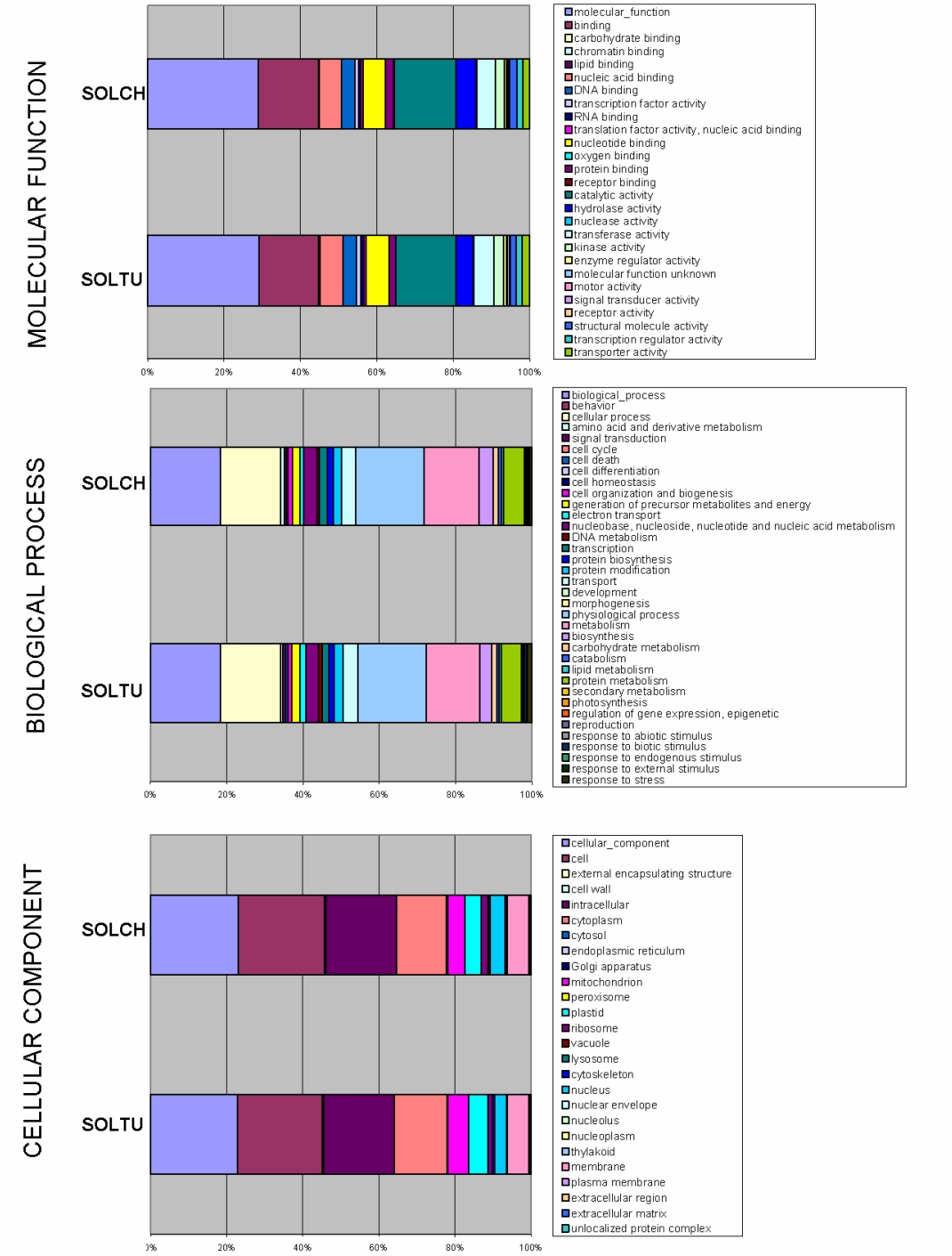


Figure 20. Assignment of Gene Ontology terms to potato unique transcripts. Plant GOSlim terms were assigned to both potato species in the 3 indicated GO areas.
SOLTU: *S. tuberosum*; **SOLCH:** *S. chacoense*.

3.3 EST survey of other Solanaceae transcriptomes

By querying dbEST, the most copious Solanaceae EST collections are those from tomato and potato. This is due to the wide interest for tomato (for fruit production) and potato (for tuber) with regards to cultivation for food consumption.

However, thousands of EST sequences for other Solanaceae species are deposited in the public database dbEST. We decided to gather the most representative collections and, then, to build, by performing ParPEST, a catalogue of preliminary annotated unique transcripts (i.e. gene indices) per each species. Table 14 resumes the composition of each gene indices. The EST sequence themselves, as well as the gene indices we built, are a valuable resource for gene discovery along tomato DNA stretches, whose sequencing is currently underway, and for the study of the Solanaceae family by means of comparative approaches.

family	genus	species	total EST	nr ESTs	gene indices	TC	sEST	total transcripts
Solanaceae	Nicotiana	TOBAC	74940	67745	37845	7529	30578	38107
		NICBE	27010	24784	9420	3206	6315	9521
		NICLS	12448	11749	6785	958	5840	6798
		NICSY	8580	8425	7534	512	7023	7535
		NICAT	329	324	312	11	301	312
	Capsicum	CAPAN	31089	28664	15703	3474	12262	15736
		CAPCH	372	372	343	11	332	343
	Petunia	PETHY	10670	10336	7004	1166	5842	7008
Rubiaceae	Coffea	COFCA	46907	38308	16121	4494	11713	16207
		COFAR	1071	1059	1007	42	965	1007

Table 14. Summary of gene indices of Nicotiana, Capsicum, Petunia and Coffea species.
TOBAC: *Nicotiana tabacum*; **NICBE:** *Nicotiana benthamiana*; **NICLS:** *Nicotiana langsdorffii x Nicotiana sanderae*; **NICSY:** *Nicotiana sylvestris*; **NICAT:** *Nicotiana attenuata*; **PETHY:** *Petunia x hybrida*; **CAPAN:** *Capsicum annuum*; **CAPCH:** *Capsicum chinense*; **COFCA:** *Coffea canephora*; **COFAR:** *Coffea arabica*; **nr EST:** non-redundant ESTs. It is the number of EST sequences in the collection after the removal of over-represented EST; **gene indices** are created by grouping overlapping EST sequences into clusters. Each cluster corresponds to an unique gene. **TC:** tentative consensus; TCs are generated from multiple sequence alignments of ESTs (assembling process). **sEST:** singleton EST. The total transcripts are created by combining the TCs and sESTs.

3.4 Gene hunting: ESTs and gene model building

The use of huge amount of ESTs to facilitate gene finding in genomic sequence is challenging. Several efforts are known to identify gene structure in long stretches of genomic sequences (Jiang and Jacob, 1998; Schlueter et al., 2003). In an effort to produce a reliable and considerable collection of gene models, we have developed GeneModelEST (D'Agostino et al., 2007b) a software that supports the process of gene model building by aligning source-native or non-native ESTs and TCs to genome sequences (see Methods 2.3.3). The definition of a good quality and representative data set of gene models is one of the tasks of the international Tomato Annotation Group (iTAG) which we are part of, and it is a preliminary requirement for the training of *ab initio* gene predictors for tomato. Tomato and potato ESTs/TCs mapped onto the 493 BAC (Bacterial Artificial Chromosome) sequences, released by the Tomato Genome Sequencing Consortium up to date (November 2007), are considered.

In figure 21 we report the number of TCs per each species that GeneModelEST classified as *optimal* or *acceptable,* and that have been selected as candidate gene models (see Methods 2.3.3). Different colours are used in the figure to discriminate distinct functional classes as they are determined as a consequence of comparisons between TCs and a protein database (see Methods 2.3.3). Only the candidate gene models which nearly cover the complete protein (green and yellow boxes) are considered "reliable" since they permit to describe the gene structure in a trustworthy way. The actual release accounts to 482 reliable gene models from native and non-native sources.

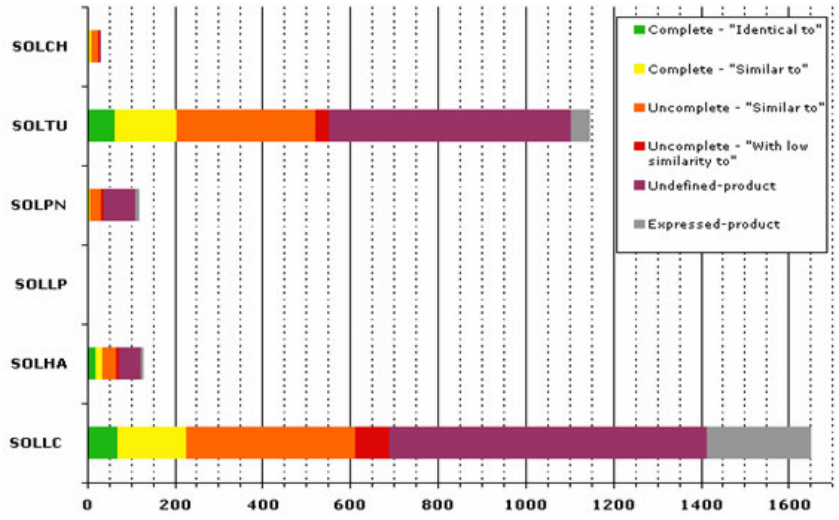


Figure 21. The bar chart describing the number of gene models per each species. Different colours are used to discriminate the functional classes which gene models are grouped into.

3.5 Arabidopsis proteome information for interpreting sequence conservation and divergence between tomato and potato

To gain an overall insight into how similar the *S. lycopersicum* and *S. tuberosum* transcriptomes are and to begin to investigate the extent to which the *S. lycopersicum* and *S. tuberosum* transcriptomes overlap, the tomato and potato unique transcripts were compared. A recent report of sequence conservation within the Solanaceae family (Ronning et al., 2003; Rensink et al., 2005), revealed high degree of sequence conservation between tomato and potato. The 78% of the tomato sequences had nucleotide sequence similarity with a sequence in the potato gene index, using a BLAST threshold of significance of e^{-10}.

We decided to go beyond a pair-wise sequence comparison strategy, by applying an Arabidopsis-based gene and gene family annotation.

The classification schema proposed in figure 22 is useful to build a backbone for Solanaceae comparative genomics studies. This approach permitted the tomato and potato transcriptomes to be compared to a reference plant species and to identify putative ortholog sequences in both of the Solanaceae species.

The protein complement of Arabidopsis thaliana (30,482 proteins) annotated by the TIGR was analysed to cluster related protein sequences according to their biological function(s).

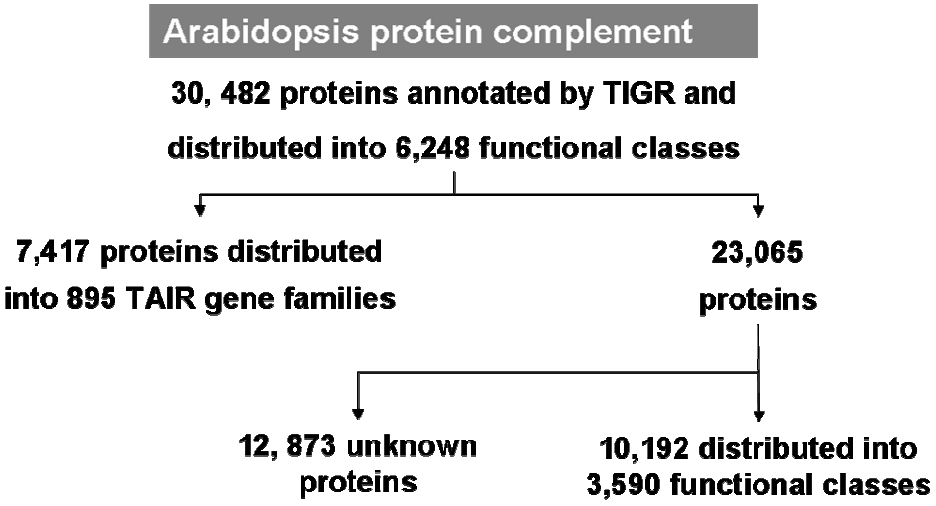


Figure 22. The Arabidopsis-based classification schema used as backbone for Solanaceae comparative genomics analysis.

The 30,482 proteins are grouped into 6,248 functional classes ranging from 1 to 547 protein members. The "Unknown protein" class, which amounts to a total of 12,873 proteins, is considered apart.

In addition, for a sub-group of these proteins, we can refine the functional annotation by exploiting the superfamily/family classification available at the TAIR (Rhee et al., 2003).

A nodal point, we ran against, was the redundancy of gene families at the TAIR due to the fact that researchers are independently curating specific gene family data. As a consequence, a manual inspection of the TAIR gene families is needed and the original 996 gene families were reduced to 895 ones.

As an example we report the case of WRKY trascription factor superfamily. On one hand, 74 members split in 8 distinct families have been deposited, on the other hand 72 members belonging to the same family have been submitted. As ground rule, we decided to discard the less informative classification.

Initially BLASTx searches (see Methods 2.4) were performed to identify tomato as well as potato TCs with significant sequence similarity to the Arabidopsis protein complement. The analysis was restricted to TC sequences, since both the sequence quality and possibly dubious origin (intronic, chimeric or contaminant) of singletons are impossible to assess. Therefore the TCs represent more stable group of sequences than the singletons.

The purpose of this large-scale sequence comparison analysis is to support an immediate identification of proteins or classes of proteins (grouped according to their biological function) which are common among the plant species considered, or, on the other hand, to estimate the magnitude of protein divergence among the same species. We are well aware that this is not an exhaustive strategy, since the limiting factor is how much of the tomato as well as of the potato transcriptome has been sampled by the EST-based analysis.

15,807 (89.6%) tomato TCs are mapped onto the Arabidopsis proteome while just 4,331(24.5%) have correspondences to proteins classified into families. On the other hand, 16,924 (85.75%) potato TCs match Arabidopsis proteins, a part of them consisting of 4,593 (23.3%) sequences is mapped onto Arabidopsis protein families.

However, to summarize thousands of BLAST comparisons we propose the visualization in figure 23. The figure is made up of 14 distinct panels, each of them representing the set of functional classes which are constituted by the exact number or a range of protein members indicated at the base of each panel. A panel can be viewed as an horizontal multiple bar histogram whose height is logarithmic-scaled and is proportional to the number of functional classes.

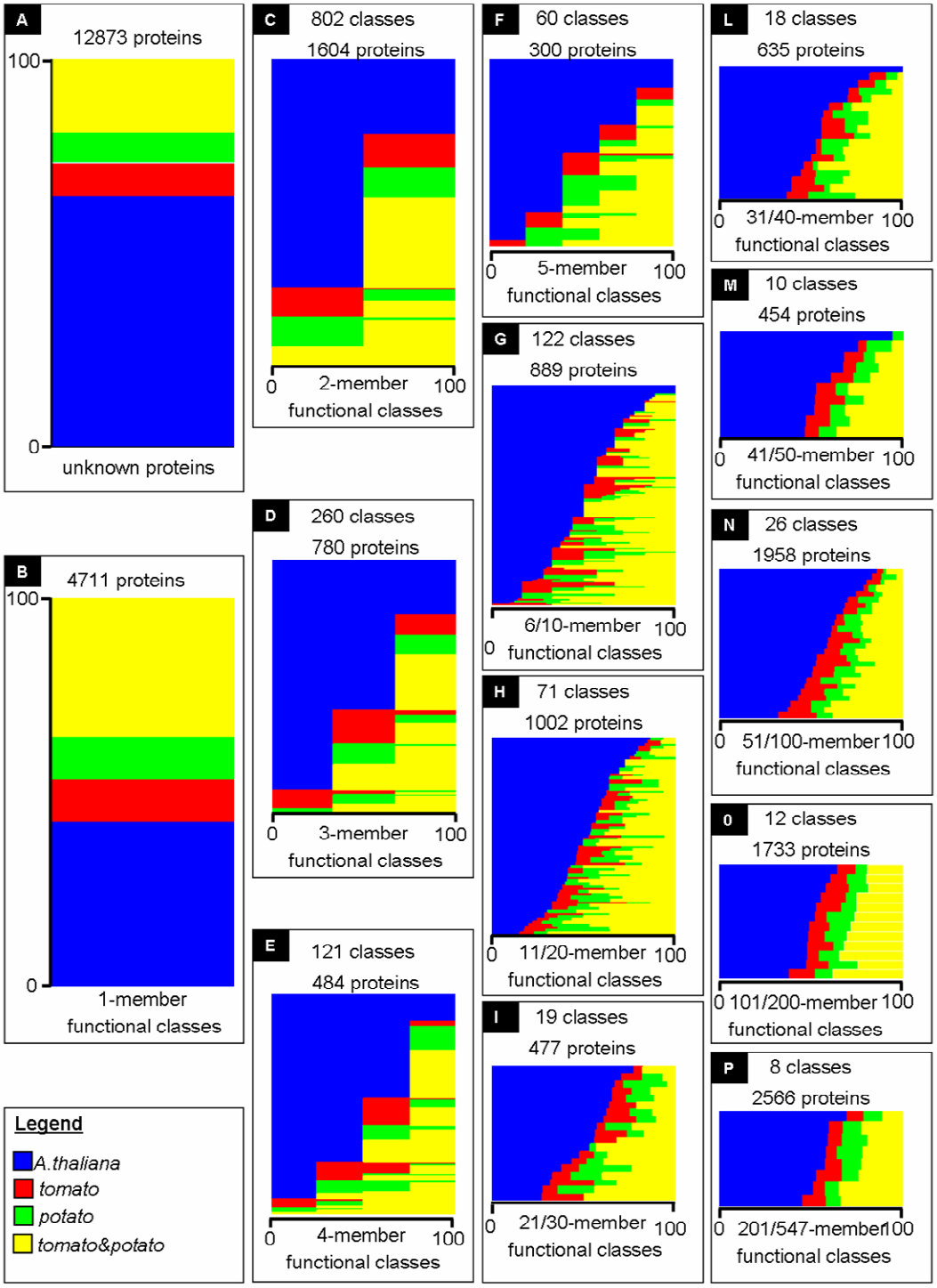
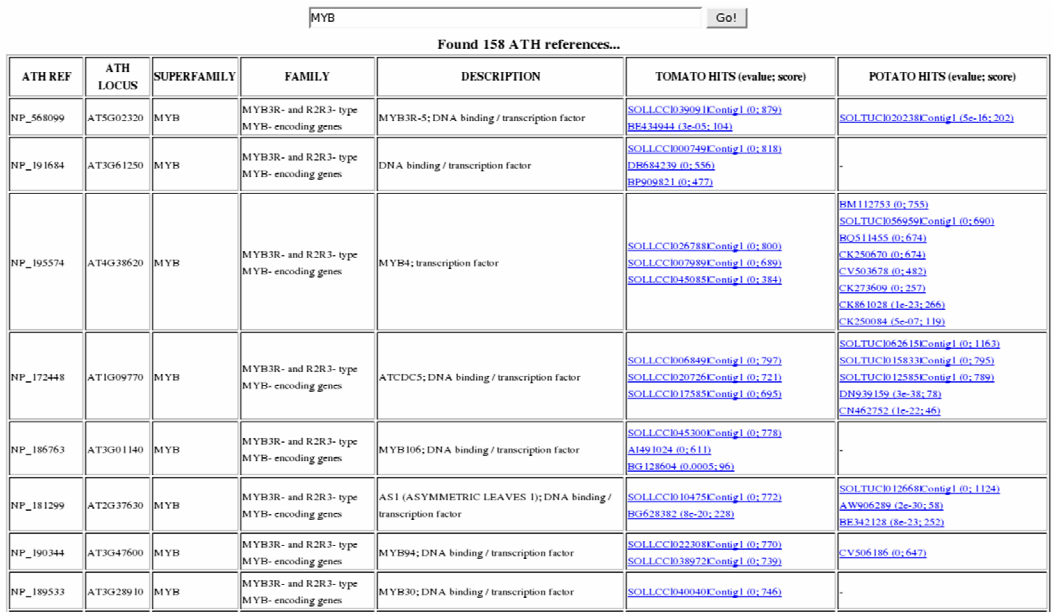


Figure 23. Summary of the large-scale sequence similarity analysis performed by comparing tomato and potato TCs against the Arabidopsis protein complement.

Each bar represents a functional class and is scaled to 100%. Each bar can be split at most into 4 segments which are coloured in different ways. Each coloured segment in a bar represents one of the four different types of BLAST results. The colour codes are the following: all the Arabidopsis proteins which match neither tomato nor potato TCs are indicated in blue; Arabidopsis proteins which match only tomato TCs are in red, while those Arabidopsis proteins which match only potato TCs are in green; in yellow are those Arabidopsis proteins which match both tomato and potato TCs.

The proposed plot show possible specificity within a functional class (i.e. bar). As an example, one-colour bars are representing functional classes which are identifying 100% correspondence to a given colour code. The overall view in figure 23 is helpful for investigators interested in specific protein functionalities.

In order to support the plant community, we set up a Web search engine, which allows users to browse the Arabidopsis-based annotations of the tomato and the potato transcriptomes (Figure 24). This work is still underway and improvements are still needed. However, it provides a quick route to decipher the function of tomato and potato protein products and to identify ortholog sequences.


MYB Go!

Found 158 ATH references...

ATH REF	ATH LOCUS	SUPERFAMILY	FAMILY	DESCRIPTION	TOMATO HITS (evalue; score)	POTATO HITS (evalue; score)
NP_568099	AT5G02320	MYB	MYB3R- and R2R3- type MYB- encoding genes	MYB3R-5; DNA binding / transcription factor	SOLLCC|039091|Contig1 (0; 879) BE434944 (3e-05; 104)	SOLTUC|020238|Contig1 (5e-16; 202)
NP_191684	AT3G61250	MYB	MYB3R- and R2R3- type MYB- encoding genes	DNA binding / transcription factor	SOLLCC|000749|Contig1 (0; 818) DR684239 (0; 556) BP909821 (0; 477)	-
NP_195574	AT4G38620	MYB	MYB3R- and R2R3- type MYB- encoding genes	MYB4; transcription factor	SOLLCC|026788|Contig1 (0; 800) SOLLCC|007989|Contig1 (0; 689) SOLLCC|045085|Contig1 (0; 384)	BM112753 (0; 755) SOLTUC|056959|Contig1 (0; 690) BQ511455 (0; 674) CK250670 (0; 674) CV503678 (0; 482) CK273609 (0; 257) CK861028 (1e-23; 266) CK250084 (5e-07; 119)
NP_172448	AT1G09770	MYB	MYB3R- and R2R3- type MYB- encoding genes	ATCDC5; DNA binding / transcription factor	SOLLCC|006849|Contig1 (0; 797) SOLLCC|020726|Contig1 (0; 721) SOLLCC|017585|Contig1 (0; 695)	SOLTUC|062615|Contig1 (0; 1163) SOLTUC|015833|Contig1 (0; 795) SOLTUC|012585|Contig1 (0; 789) DN939159 (3e-38; 178) CN462752 (1e-22; 146)
NP_186763	AT3G01140	MYB	MYB3R- and R2R3- type MYB- encoding genes	MYB106; DNA binding / transcription factor	SOLLCC|045300|Contig1 (0; 778) AI491024 (0; 611) BG128604 (0.0005; 96)	-
NP_181299	AT2G37630	MYB	MYB3R- and R2R3- type MYB- encoding genes	AS1 (ASYMMETRIC LEAVES 1); DNA binding / transcription factor	SOLLCC|010475|Contig1 (0; 772) BG628382 (8e-20; 228)	SOLTUC|012668|Contig1 (0; 1124) AW906289 (2e-30; 158) BE342128 (8e-23; 252)
NP_190344	AT3G47600	MYB	MYB3R- and R2R3- type MYB- encoding genes	MYB94; DNA binding / transcription factor	SOLLCC|022308|Contig1 (0; 770) SOLLCC|038972|Contig1 (0; 739)	CV506186 (0; 647)
NP_189533	AT3G28910	MYB	MYB3R- and R2R3- type MYB- encoding genes	MYB30; DNA binding / transcription factor	SOLLCC|040040|Contig1 (0; 746)	-

Figure 24. Snapshot of the search engine we developed to browse the results obtained by comparing tomato and potato transcriptomes to the model plant Arabidopsis. The search by using the key word "MYB" resulted into 158 *A. thaliana* references. In a row are shown the protein RefSeq accession number; the AGI (Arabidopsis Genome Initiative) gene model ID; the TAIR superfamily and family which the protein has been classified into; the TIGR functional class, the list of the tomato and potato unique transcripts (both TCs and sESTs) matching the Arabidopsis protein.

7,417 Arabidopsis proteins are also annotated into TAIR families as well as functional groups. This kind of classification-driven annotation is critical for a preliminary

organization of tomato and potato transcripts into protein families. To this end BLASTx searches (see Methods 2.4) were performed to associate tomato and potato TCs to those Arabidopsis proteins which have already been classified into families. Also in this case, it is needed to fall back upon a user-friendly graphical visualization in order to make sense of thousands of BLAST results. We thought that an easy way to represent our results is a rooted graph (Figure 25). Each small black node in the graph represents a superfamily which is, in its turn, connected to one or more nodes representing the protein families. The size of each family node is proportional to the number of proteins in the family, while the chromatic tones encodes the results of BLAST comparisons. So in green are those families whose protein members match mainly potato TCs; in yellow are those families whose protein members match fairly potato and tomato TCs; in red are those families whose protein members match mainly tomato TCs. Finally different black colour tones are used to identify those Arabidopsis families whose protein members do not match neither tomato nor potato TCs. This representation, though immediate, does not provide the fine grained protein terms. Therefore the interpretation of the chromatic scale, which passes from green through yellow to red, is referred to a protein family as a whole.

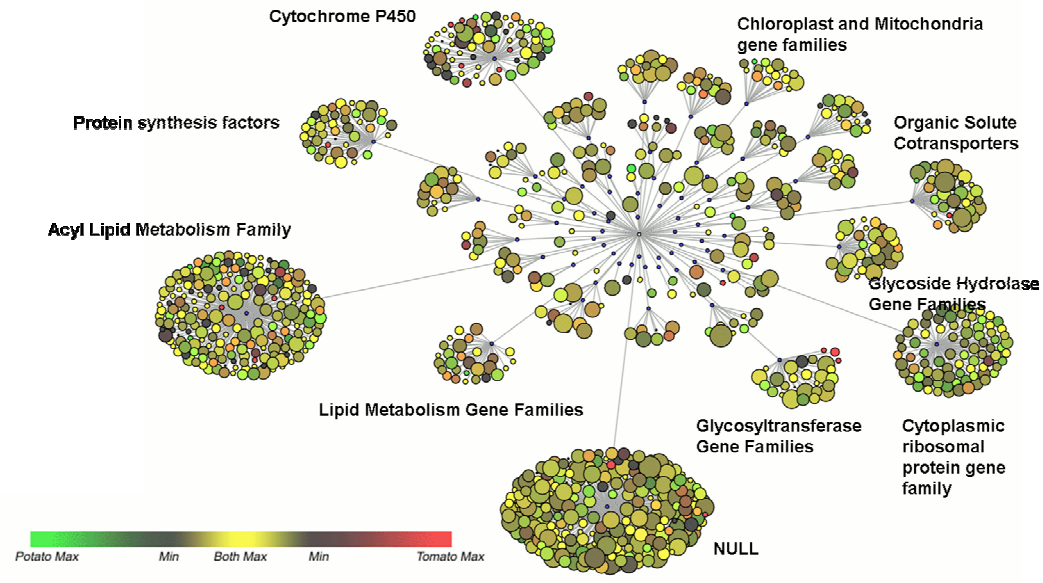


Figure 25. The rooted graph shows the Arabidopsis protein superfamily/family organization. Close to some of the external clouds the name of the superfamilies are indicated. The cloud "NULL" groups all the families not compiled into any superfamily. Small black nodes corresponding to superfamilies while the remaining nodes, whose size is variable and proportional to the number of proteins in the family, are coloured according to a chromatic scale, which passes from green through yellow to red. The interpretation of colours is strictly linked to the results of the comparisons.

All the tomato and potato tentative consensus sequences, which did not match Arabidopsis proteins, were filtered out and were used to perform an all-against-all comparison. In this step of the analysis we considered also the set of tomato and potato singleton ESTs with no match against the Arabidopsis proteins.

Results are summarized in table 15.

TOMATO				VS	POTATO						% aligned onto 498 tomato BACs
sESTs	3083	no match found vs UniProt	2558		no match found vs UniProt	2248	sESTs	1714			13.77
							TCs	534			11.24
					match found vs UniProt	310	sESTs	192			5.73
							TCs	118			9.32
		match found vs UniProt	525		no match found vs UniProt	142	sESTs	120			16.67
							TCs	22			9.09
					match found vs UniProt	383	sESTs	247	same UniProt sbjct	165	17.41
									different UniProt sbjct	82	
							TCs	136	same UniProt sbjct	63	15.44
									different UniProt sbjct	73	
TCs	881	no match found vs UniProt	671		no match found vs UniProt	618	sESTs	391			14.58
							TCs	227			19.82
					match found vs UniProt	53	sESTs	31			13.64
							TCs	22			16.13
		match found vs UniProt	210		no match found vs UniProt	49	sESTs	35			22.86
							TCs	14			28.57
					match found vs UniProt	161	sESTs	87	same UniProt sbjct	57	19.54
									different UniProt sbjct	30	
							TCs	74	same UniProt sbjct	53	24.32
									different UniProt sbjct	21	

Table 15. Summary of the results obtained by performing the tomato against potato comparison. The tomato as well as the potato collections are made up of both TCs and sESTs (singletons) and are cleaned out from those sequences that have been mapped onto the Arabidopsis proteome. Each set of sequences (i.e. TCs and sESTs) are split into different sub-groups according to the annotation type categories provided by both TomatEST and PotatEST databases.

The 9970 tomato sequences, that have neither similarity in Arabidopsis databases nor a potato counterparts, are made up of 1112 TCs and 8858 sESTs.

On the other hand, the 19826 potato sequences, that did not match any Arabidopsis proteins or did not have tomato counterparts, are made up of 2808 potato TCs and 17018 sESTs. Their function assignments are determined by tacking advantage of the sequence annotation provided by both TomatEST and PotatEST databases. Functional annotations are grouped into 10 arbitrary categories of data sources as indicated in figure 25. The reason of still finding Arabidopsis matches is because the EST annotation procedure, which I am referring to, was carried out by BLASTx with an e-value cut-off $\leq 10^{-3}$ (Methods 2.1.5). However, in case of tomato TCs and sESTs matching Arabidopsis proteins, the most of them are hypothetical proteins, transposase or retroelements polyproteins (data not shown).

As evident from the panel A of the figure 26, the TCs most represented (85%) are those that have no function assigned (i.e. NULL category). Among these 147 have been successfully aligned onto the 498 available tomato BACs.

The same trend is observed if tomato sESTs are considered. Indeed, the more sizeable slice of the cake graph is referred to sESTs that have no function assigned (87%) (Figure 26B). However, at least 914 out of 7582 are certainly no sequencing artefacts or chimeras since they were fully splice-aligned along the tomato genome sequences (Methods 2.3.2).

Comparable results are obtained if the 19826 potato sequences are taking into account. Once more the most represented category, evaluating both TC (Figure 26C) and sEST (Figure 26D) sub-sets, is the NULL one. Also in this case, 64 TCs and 224 sESTs with no function assigned have been fully mapped onto tomato genome sequences.

This facet provides insight into how still poorly annotated are the protein databases and how much important is the integration of data in order to identify molecular information and to focus on specific genes for which a reliable annotation is still needed.

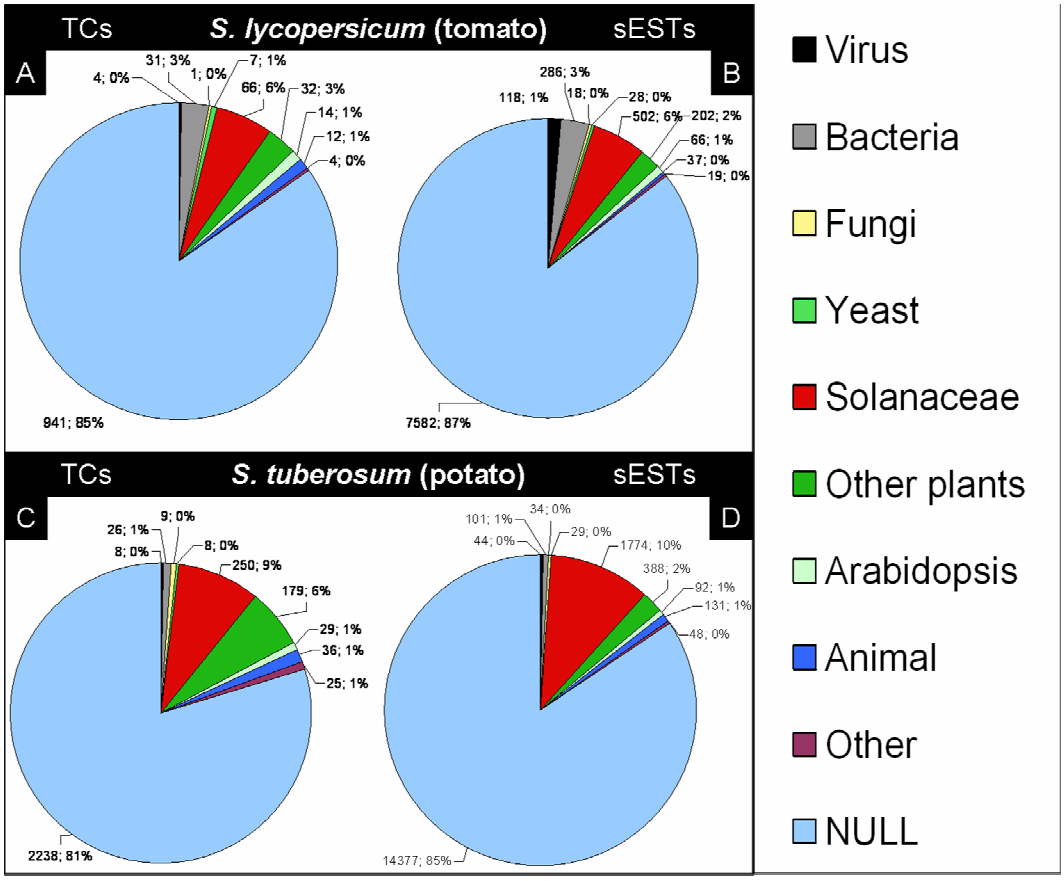


Figure 26. The figure shows four cake graphs, each of them summarizes the functional annotation results with respect to the data source of origin. A and C concern tomato and potato TCs respectively. Instead, the B and D panels are about sESTs.

3.6 ISOL@: an Italian SOLAnaceae genomics resource

Our effort of sampling and analysing the Solanaceae transcriptomes as well as of providing a structural and functional annotation of the tomato genome, drove us to develop the computational platform ISOL@ (Italian SOLAnaceae genomics resource; Chiusano et al., 2007). ISOL@ is originated from the idea to develop a suitable computational platform to gather, converge and integrate the overwhelming amounts of '-omics' data generated worldwide and useful to address key questions risen by the vision of the international Solanaceae Genomics Project. These data (genome sequences; information on gene expression; information on the cell metabolic status; and other…) represent multiple aspects of a biological system and need to be investigated for understanding the biological system as a whole, shedding light on the mechanisms which underpin the system functionality. To this end we conceived ISOL@ as a multi-level computational environment where the multi-level structure summarizes the semantics of the biological data entities (Figure 27).

ISOL@ currently consists of two main levels: the genome and the expression levels. The cornerstone of the genome level is represented by the *S. lycopersicum* genome draft sequences produced by the International Tomato Genome Sequencing Consortium. Instead, the basic element of the expression level is the transcriptome information from different Solanaceae species, in the form of species-specific comprehensive collections of expressed sequence tags. Each level can be independently accessed through specific Web applications which allow user-driven data investigation (Figure 28). The cross-talk between the genome and the expression level is based on the sharing of data sources and on tools classified as "basic" or "subsidiary". These tools enhance data quality, extract information content from the levels' under-parts and aim to produce valued-added biological knowledge. The existing multilevel environment has been designed to be extended to the proteome and metabolome levels (Figure 28), through pre-defined entry points, as soon as results in standard formats will be provided to the SOL community.

ISOL@ is daily accessed by scientists from different countries (Figure 29) because it provides a preliminary annotation of the tomato genome while awaiting for the official annotation by the international Tomato Annotation Group. Furthermore, the platform collects and distributes the Solanaceae transcripts, provides their functional annotation and classification, allows EST-based investigations on genome functionalities and supports expression pattern analysis.

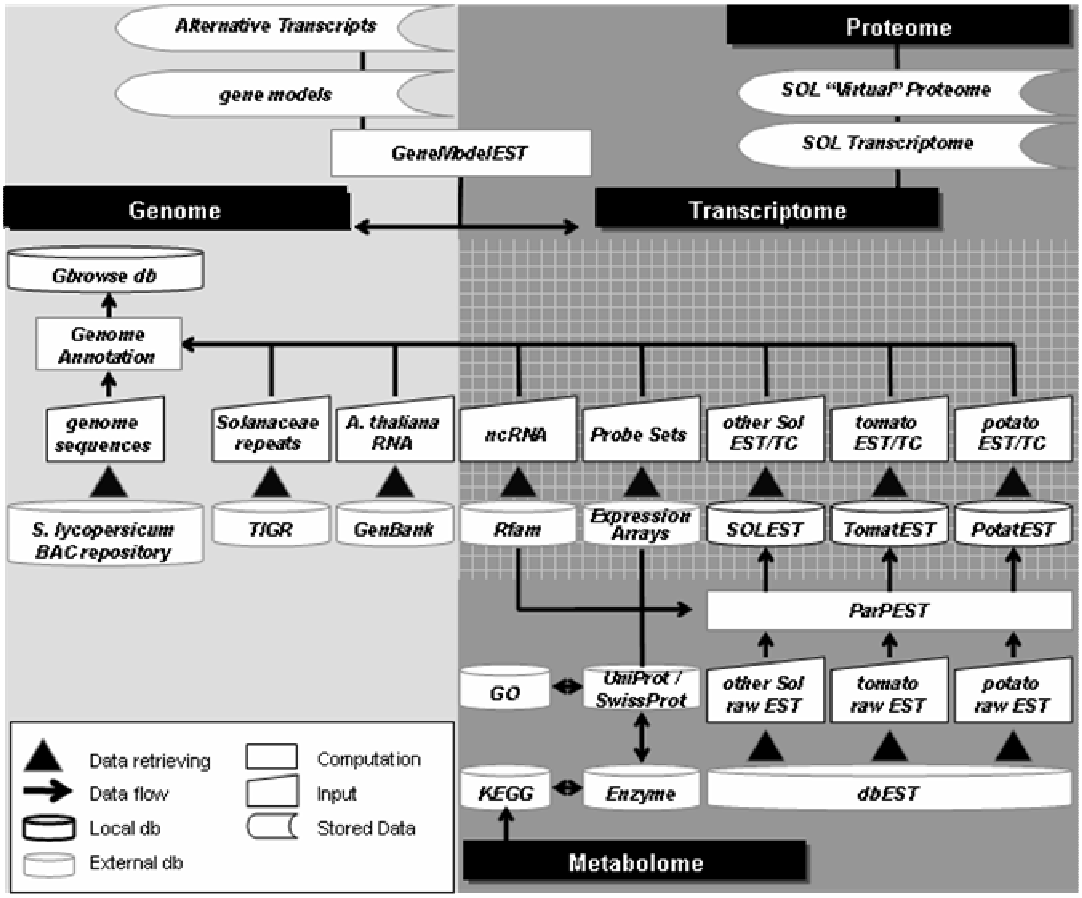


Figure 27. ISOL@ multi-level environment.
Data sources and tools are reported. The genome and the expression levels are respectively indicated in the light and dark grey backgrounds. Shared data are located in the gridded area.
Entry points for proteome and metabolome data are shown.

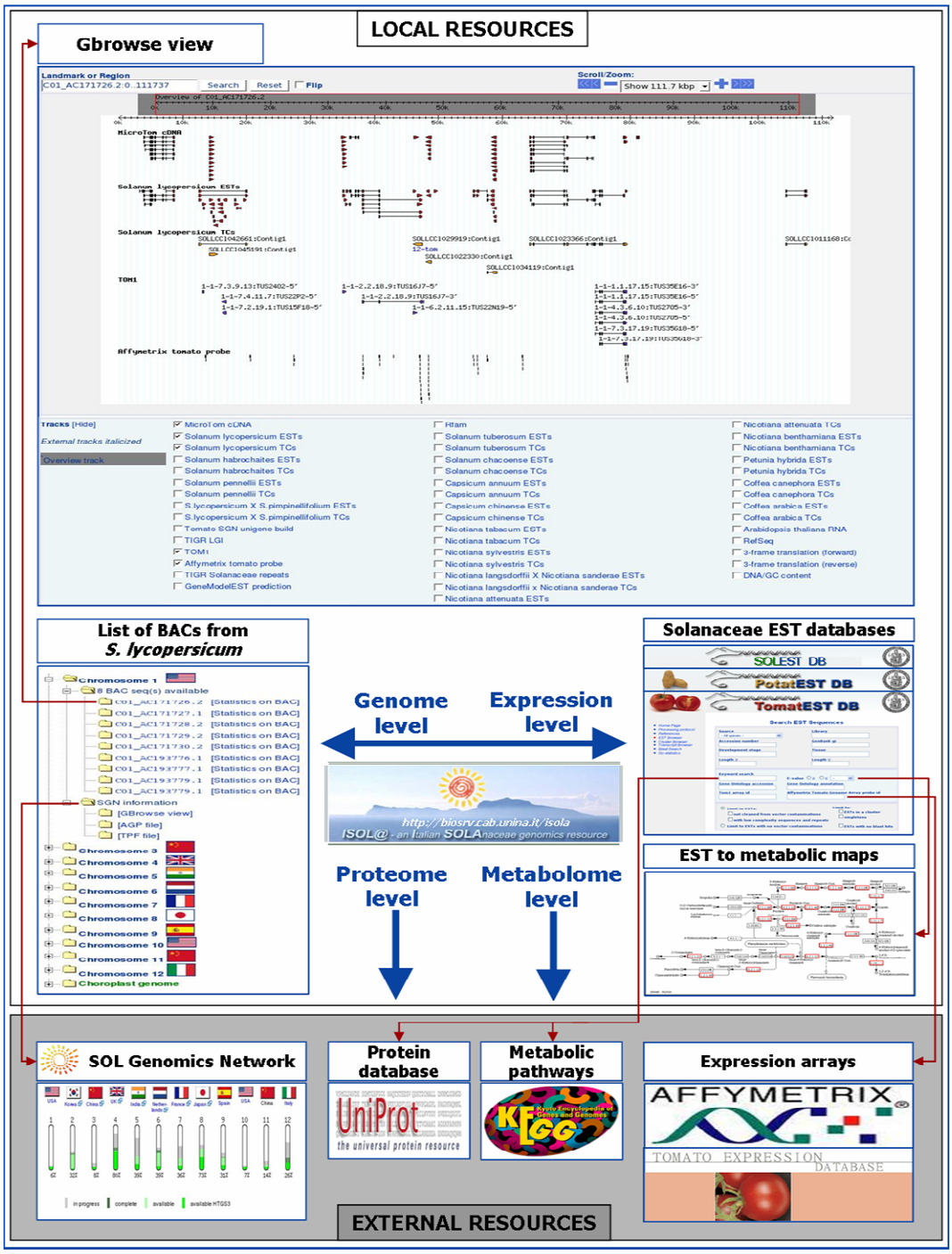


Figure 28. Snapshot of the Web-based application to navigate through ISOL@.
ISOLA is accessible through two different gateways. The Genome Browser gateway let the user explore the list of the tomato BAC sequences grouped by chromosome and visualize the tracks displayed along each BAC. Each track is cross-linked to other local or external resources. Cross-references to the tomato genome annotation pages at the SOL Genomics Network are part of the 'genome level' too. The Solanaceae EST database gateway let the user investigate Solanaceae transcriptomes as revealed by EST sampling. The EST database can be queried to identify functional annotations associated to a single EST or a TC. Cross-links to the UniProt external resource are established. In case an expressed sequence is associated to an enzyme function, a cross-link to the corresponding KEGG metabolic pathway(s) is provided. These links permit data from proteomics and metabolomics approaches to be integrated in the existing multi-level environment. In addition, the association of the tomato ESTs to the probes from the Affymetrix or from the TED database expands information concerning the 'expression level' and provides the opportunity to integrate data from expression profiling into the platform.

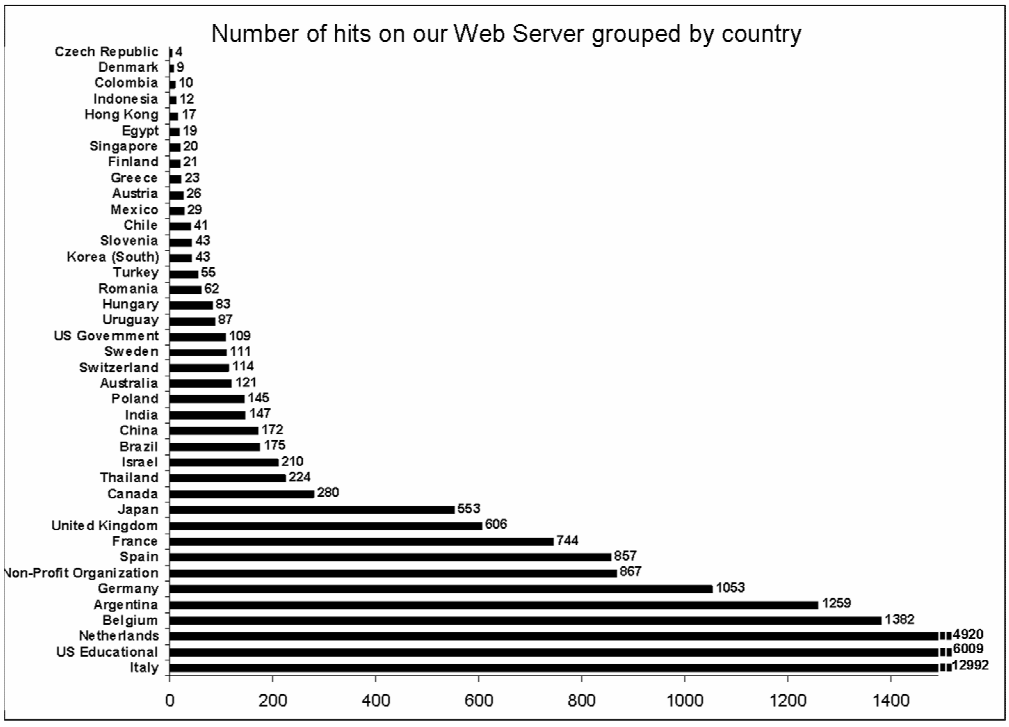


Figure 29. Number of hits on our Web server since December 2006 until November 2007. Every time a browser hits our Web sites it leaves a trail in our access log. The number of hits is grouped by country.

3.7 Saffron genes: an EST database from saffron stigmas

Saffron (*Crocus sativus L.*, Iridaceae) flowers have been used as a spice and medicinal plant ever since the Greek-Minoan civilization. The edible part – the stigmas - are commonly considered the most expensive spices in the world and are the site of a peculiar secondary metabolism responsible for the characteristic colour and flavour of saffron. The characterization of the transcriptome of saffron stigmas is likely to shed light on several important biological aspects: the molecular basis of flavour and colour biogenesis in spices, the biology of the gynoecium, and the genomic organization of Iridaceae. For these reasons, in the frame of a collaboration with the group directed by Mr. Giuliano, we have undertaken the sequencing and bioinformatics characterization of Expressed Sequence Tags from saffron stigmas.

3.7.1 Construction and functional annotation of saffron unigene set

9,769 electropherograms produced by Mr. Giuliano's team were analysed in order to remove low quality sequences from the 5' and 3' ends of EST reads (see Methods 2.1.1). The sequences were further processed to remove vector contaminations and to mask low complexity and/or repeat sub-sequences (see Methods 2.1.3). This process reduced the original dataset to 6,603 high-quality sequences longer than 60 nucleotides. Only 6,202 EST fragments, whose length is more than or equal to 100 nucleotides, were considered for the submission to the NCBI dbEST division. They are accessible under the accession numbers from EX142501 to EX148702.

The EST dataset was subjected to the ParPEST clustering/assembling module in order to group overlapping ESTs which tag the same gene in a single gene index. The total number of clusters generated are 1,893.

1,376 clusters are made up of a single EST and are therefore classified as singletons.

The remaining 517 clusters are made up of 5,324 ESTs, assembled into 534 distinct TCs (Table 16).

Singleton ESTs	**N. of sequences**	**1376**
	Avg. EST length (nt)	**239**
ESTs in TCs	**N. of sequences**	**5324**
	Avg. EST length (nt)	**427**
TCs	**N. of TCs**	**534**
	Avg. length (nt)	**552**

Table 16. Assembly statistics.

In 11 clusters, ESTs are assembled so that multiple TCs are defined (ranging from 2 to 6). Identification and assignment of function to each transcript was performed by the ParPEST annotation module. Of 1,910 transcripts, 1,158 (60.6%) have no protein similarities. This highlights the relevance of studying Iridaceae for the lacking of molecular information on this plant species. The remaining 752 (39.4%) have at least one significant match in the protein database. Within this latter set, 131 (6.9%) are described as hypothetical, unknown or expressed proteins. In this way, they are not confirming an effective functional role of the transcript products and consequently they are remarking how much still magre is the annotation of plant proteomes. The protein annotation can be switched to the GO terms for just 157 sequences. In many cases, multiple gene ontology terms could be assigned to the same sequence, resulting in 210 assignments to the *molecular function*, 944 to the *biological process* and finally 2,192 to the *cellular component* GO areas. The GO annotation were further reduced using plant GO-slim terms (Figure 28). In the *molecular function* ontology area, the most represented terms describe catalytic (33.3%) and hydrolase activity (20.0%) (Figure 30A). The remaining categories are less represented. Considering the *biological process* area, the vast majority of the GO assignments corresponds to the transport category (~78.8%) (Figure 30B). Finally, for the *cellular component* area the assignments were mainly given to the plastid (36%), mitochondrion (33%), and cytoplasmic membrane-bound vescicle (29%) components (Figure 30C).

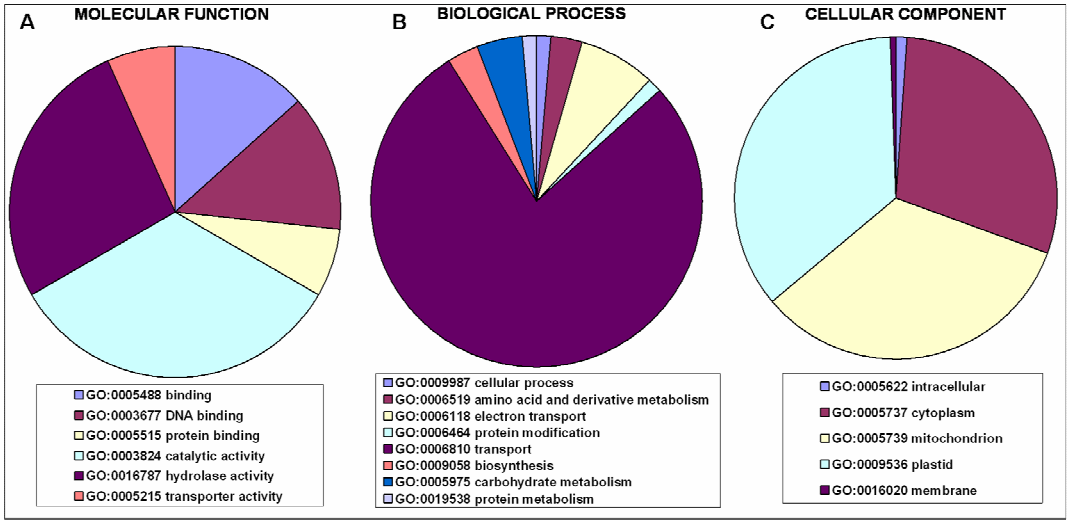


Figure 30. Assignments of Plant Gene Ontology terms to the Crocus putative transcripts. A. Molecular function; **B.** Biological process; **C.** Cellular component.

All the data reported were used to construct the Saffron Genes database accessible at http://saffrongenes.org and published on *BMC Plant Biology* (D'Agostino et al., 2007c).

3.7.2 TCs composed of most abundant ESTs

Analysis of EST abundance comprising a TC can provide insights with respect to gene expression levels occurring in the stigma tissue. Therefore to identify genes that were highly expressed, we detected those TCs that are composed of 20 ESTs or more. (Table 17). The most highly expressed TC, Cl000057:2 (547 ESTs), bears homology to short chain dehydrogenases (PF00106.12). This protein family comprises members involved in hormone biosynthesis, like the ABA2 gene of Arabidopsis which catalyzes the conversion of xanthoxin into ABA aldehyde (Gonzalez-Guzman et al., 2002), or in sexual organ identity, like the TASSELSEED2 (TS2) gene of maize (Figure 31). TS2 is expressed in pistil primordia cells of maize, where it activates a cell death process eliminating these cells from male reproductive organs (Calderon-Urrea and Dellaporta, 1999).

TC	# ESTs	length (nt)	BLASTx annotation	e-value
Cl000057:2	547	1242	Q7XL00_ORYSA -OJ000315_02.17 protein	0
Cl000837:2	122	1528	Q8VZY2_MUSAC -Cytochrome P450-1	0
Cl000799:2	114	711	-	-
Cl001953:2	109	755	O80821_ARATH -Hypothetical protein At2g41470	1,00E-16
Cl001114:3	104	770	HSP13_ARATH -18.2 kDa class I heat shock protein (HSP 18.2)	1,00E-32
Cl000299:1	104	570	Q9XHD5_IPOBA -B12D protein	2,00E-32
Cl000870:1	94	592	Q6ZX06_ORYSA -Lipid transfer protein	3,00E-26
Cl001582:1	61	600	-	-
Cl000209:1	61	1071	Q5G1M8_9POTV -Polyprotein (Fragment)	0
Cl001173:1	56	785	Q6H452_ORYSA -Putative monoglyceride lipase	0
Cl000220:1	55	831	Q94HY3_ORYSA -Putative gamma-lyase	0
Cl000348:1	54	955	Q9AVB7_9LILI -LhMyb protein	0
Cl001319:1	47	460	Q8RVT5_PANGI -Acyl-CoA-binding protein	1,00E-35
Cl001051:1	45	665	Q8H293_ANACO -Cytochrome b5	0
Cl000246:1	45	537	-	-
Cl000336:1	44	685	GPAT6_ARATH -Glycerol-3-phosphate acyltransferase 6 (EC 2.3.1.15) (AtGPAT6)	0
Cl000468:2	42	1021	Q70SZ8_9ASPA -Carboxyl methyltransferase	0
Cl000482:1	38	730	Q84P95_ORYSA -Disulfide isomerase	0
Cl000982:1	38	230	-	-
Cl001040:1	37	734	Q8GZR6_LYCES -GcpE	0
Cl001329:1	36	384	Q4LEZ4_ASPOF -MADS-box transcription factor	1,00E-29
Cl001815:1	34	992	BGAL_ASPOF -Beta-galactosidase precursor (EC 3.2.1.23) (Lactase)	0
Cl000113:1	33	634	Q6VAB3_STERE -UDP-glycosyltransferase 85A8	9,00E-16
Cl000687:1	33	782	Q9XGS6_PRUDU -Cytosolic class II low molecular weight heat shock protein	0
Cl000887:1	33	802	Q9FVZ7_ORYSA -Putative steroid membrane binding protein	0
Cl001463:1	32	605	Q9FE65_ARATH -60S ribosomal protein L34, putative	0
Cl000932:1	32	974	Q652L6_ORYSA -Putative monodehydroascorbate reductase	0
Cl001812:1	30	554	Q42338_ARATH -B12D-like protein	5,00E-32
Cl001134:1	29	569	Q8W453_ARATH -Hypothetical protein (DIR1 protein) (At5g48485)	7,00E-14
Cl001906:1	28	602	Q4TES1_TETNG -Chromosome undetermined SCAF5157, whole genome shotgun sequence.	9,00E-07
Cl001988:1	25	1446	Q8VX49_WHEAT -Cytochrome P450 reductase (EC 1.6.2.4)	0
Cl001107:1	24	783	Q9SGA5_ARATH -F1C9.14 protein (At3g02070)	0
Cl001447:1	24	453	Q5VS45_ORYSA -Hypothetical protein P0425F02.23	1,00E-12
Cl000515:1	24	506	Q6ZCF3_ORYSA -Putative copper chaperone	8,00E-15
Cl000762:1	24	247	-	-
Cl001114:2	23	748	HSP13_ARATH -18.2 kDa class I heat shock protein (HSP 18.2)	1,00E-32
Cl001894:1	23	312	-	-
Cl000057:1	23	740	TRXH1_ARATH -Thioredoxin H-type 1 (TRX-H-1)	1,00E-36
Cl001263:1	22	667	Q9XH76_ARATH -Zinc finger protein-like (PMZ)	0
Cl001010:1	21	1066	Q8H2A7_ANACO -PFE18 protein (Fragment)	0
Cl000300:1	21	506	Q93WW3_NARPS -Metallothionein-like protein type 2	6,00E-12
Cl000057:3	21	183	-	-
Cl000885:2	21	753	Q41067_PINSY -Polyubiquitin	0
Cl001397:1	20	798	Q9LSQ5_ARATH -1,4-benzoquinone reductase-like; Trp repressor binding protein-like	0
Cl001774:1	20	457	Q9SN96_ARATH -Hypothetical protein F18L15.150 (Hypothetical protein MTH12.17)	7,00E-19
Cl000185:1	20	397	Q84LB7_MALDO -Cysteine protease inhibitor cystatin (Fragment)	2,00E-12
Cl001935:1	20	673	SRP19_ARATH -Signal recognition particle 19 kDa protein (SRP19)	4,00E-38
Cl000333:1	20	418	Q7F6G0_ORYSA -Putative metallothionein-like protein	6,00E-20
Cl000594:1	20	1145	SUS1_TULGE -Sucrose synthase 1 (EC 2.4.1.13) (Sucrose-UDP glucosyltransferase 1)	0

Table 17. Saffron TCs composed of the most redundant ESTs.

Biochemical studies suggest that the TS2 protein is a hydroxysteroid dehydrogenase (Wu et al., 2007). It will be interesting to determine the function and substrate specificity of the saffron Cl000057:2 product.

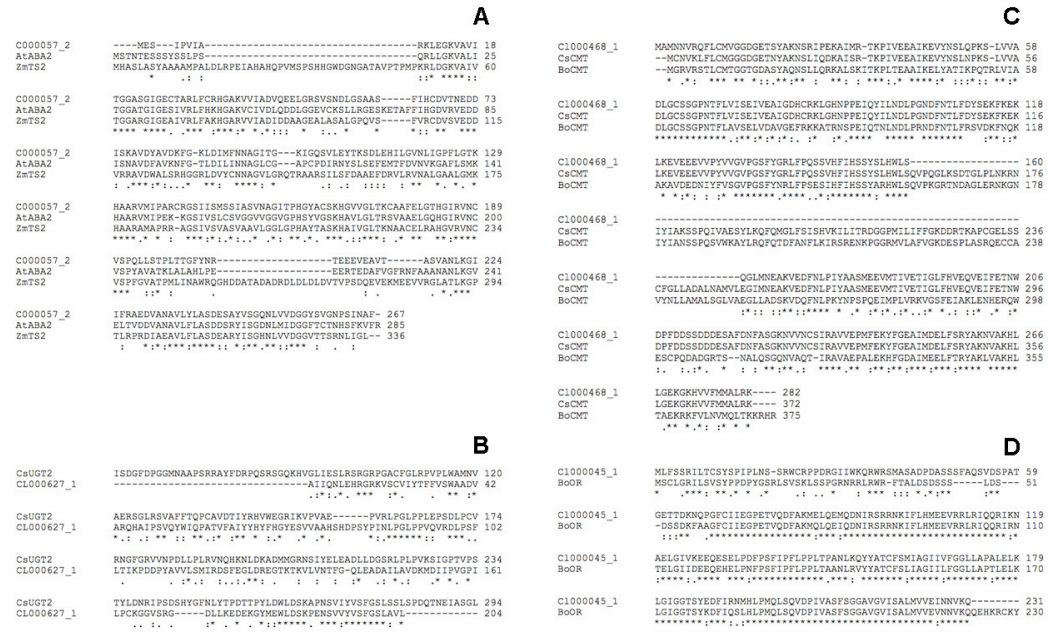


Figure 31. ClustalW alignments of deduced protein sequences expressed in Crocus stigmas.

A large number of Cytochrome P450 sequences are expressed in saffron stigmas, some of which at very high levels (Tables 17 and 18). Also, lipid metabolism seems to be very active, judging from the TCs encoding proteins involved in this process (Table 18). Several TCs encode putative carotenoid metabolism enzymes (Table 18): Cl000944:1 encodes non-heme -carotene hydroxylase (PF03897), which is highly expressed in saffron stigmas (Castillo et al., 2005). Cl000627:1 encodes a putative glucosyltransferase, very similar to UGTCs2, which is able to glycosylate crocetin in vitro (Moraga et al., 2004) (Figure 31). Cl001532:1 and Cl001032:1 also, encode putative isoprenoid GTases, one of which could represent the still missing enzyme responsible for the glycosylation of picrocrocin. Cl001432:1 encodes a protein similar to plastid terminal oxidase, involved in phytoene desaturation (Carol and Kuntz, 2001), while EST cr36_B21 encodes a protein similar to fibrillin, which is a carotenoid-binding protein in pepper chromoplasts (Deruere et al, 1994). Cl000468 encodes a carboxyl methyltransferase very similar to the one catalyzing the synthesis of bixin (Bouvier et al., 2003a) (Figure 31). This TC seems to encode a “short” form of the annatto and crocus methyltransferases from GenBank, possibly derived from alternative splicing (Figure 31).

Although a methyltransferase reaction has not been described in saffron stigmas, the biosynthesis of bixin and that of crocin share some features in common, since both pigments are derived from the oxidative cleavage of a carotenoid (Giuliano et al., 2003).

Transcript ID	length (nt)	# ESTs	BLASTx annotation	e-value
Cyt. P450				
Cl000837:2	1528	122	Q8VZY2_MUSAC -- Cytochrome P450-1	0
Cl001988:1	1446	25	Q8VX49_WHEAT -- Cytochrome P450 reductase (EC 1.6.2.4)	0
Cl000837:3	674	17	Q8L5Q2_CICAR -- Putative cytochrome P450 monooxygenase	2,00E-27
Cl000414:1	752	5	Q9AVM1_ASPOF -- Cytochrome P450	0
Cl000150:1	406	3	Q9ATU9_LOLRI -- Putative cytochrome P450	4,00E-17
Cl000166:1	710	3	Q6EP96_ORYSA -- Putative cytochrome P450	9,00E-16
Cl001887:1	248	2	Q6H516_ORYSA -- Putative cytochrome P450	0.0004
Cl000837:1	600	2	Q8VZY2_MUSAC -- Cytochrome P450-1	3,00E-16
cr13_O11	360	1	Q8S7S6_ORYSA -- Cytochrome P450-like protein	7,00E-35
cr21_F05	448	1	Q8S7S6_ORYSA -- Cytochrome P450-like protein	1,00E-37
cr28_M16	533	1	Q6Z0U4_ORYSA -- Putative cytochrome P450 reductase	0
cr34_J15	509	1	Q8S7S6_ORYSA -- Cytochrome P450-like protein	0
Lipid metabolism				
Cl000870:1	592	94	Q6ZX06_ORYSA -- Lipid transfer protein	3,00E-26
Cl001173:1	785	56	Q6H452_ORYSA -Putative monoglyceride lipase	0
Cl000787:1	743	10	Q94GF2_ORYSA -- Putative phospholipase	0
Cl001992:1	637	5	Q52RN7_LEOAR -- Non-specific lipid transfer protein-like	2,00E-28
Cl001009:1	667	5	O04439_ALLPO -- 3-ketoacyl carrier protein synthase III	0
Cl001749:1	635	5	Q9NCL8_DICDI -- Phosphatidylinositol transfer protein 1	5,00E-30
Cl000344:1	704	5	O49902_NICRU -- 1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase	0
Cl000816:1	677	2	Q6K7T9_ORYSA -- Peroxisomal fatty acid beta-oxidation multif. protein	0
Cl000294:1	707	2	Q84Z91_ORYSA -- Oxysterol-binding protein-like	0
Cl000741:1	734	2	STAD_ORYSA -- Acyl-(acyl-carrier-protein) desaturase, chloroplast precursor	0
cr13_F23	350	1	Q8S459_LYCES -- Putative sphingolipid delta 4 desaturase DES-1	0
cr15_P04	306	1	GPX4_MESCR -- Probable phospholipid hydroperoxide glutathione peroxidase	5,00E-16
cr27_P08	74	1	Q5N7U2_ORYSA -- Phospholipid/glycerol acyltransferase-like protein	4,00E-06
cr35_M17	437	1	GPX4_MESCR -- Probable phospholipid hydroperoxide glutathione peroxidase	1,00E-24
Carotenoid metabolism				
Cl000944:1	645	11	Q8VXP2_9ASPA -- Beta-carotene hydroxylase	4,00E-17
Cl001432:1	602	2	Q9FZ04_CAPAN -- Plastid terminal oxidase	0
Cl001532:1	420	7	GT_CITUN -- Limonoid UDP-glucosyltransferase	2,00E-06
Cl001032:1	426	2	5CD69_9MYRT -- Monoterpene glucosyltransferase	2,00E-08
Cl000627:1	611	2	69UF5_ORYSA -- Putative anthocyanin 5-O-glucosyltransferase	0
Cl000468:2	1021	42	Q70SZ8_9ASPA -- Carboxyl methyltransferase	0
Cl000468:1	767	6	70SZ8_9ASPA -- Carboxyl methyltransferase	0
cr9_J02	69	1	Q9FEC9_LYCES -- Plastid quinol oxidase (Plastid terminal oxidase)	1,00E-05
cr36_B21	706	1	PAP2_ORYSA -- Probable plastid-lipid associated protein 2, chloroplast precursor (Fibrillin-like protein 2)	0
Cl000045:1	746	14	Q9FKF4_ARATH -- Hypothetical protein At5g61670	0
Transcription factors				
Cl000348:1	955	54	Q9AVB7_9LILI -- LhMyb protein	0
Cl001329:1	384	36	Q4LEZ4_ASPOF -- MADS-box transcription factor	1,00E-29
Cl000348:2	669	6	Q70RD2_GERHY -- MYB8 protein	0
Cl000712:1	714	6	Q6Z8N9_ORYSA -- Putative AT-hook DNA-binding protein	0
Cl000359:1	593	5	O82115_ORYSA -- Zinc finger protein	5,00E-19
Cl000502:1	565	3	ULT1_ARATH -- Protein ULTRAPETALA1	4,00E-37
Cl000652:1	537	2	Q6ZG02_ORYSA -- Putative DNA-binding protein WRKY2	0
cr17_J15	567	1	Q6Q6W8_9ASPA -- Agamous MADS-box transcription factor 1a	0
cr26_B12	653	1	Q8LAP4_ARATH -- Contains similarity to MYB-related DNA-binding protein	2,00E-23
cr6_B13	312	1	Q9M7F3_MAIZE -- LIM transcription factor homolog	0

Table 18. Expressed sequences grouped by putative function.

Finally, Cl000045:1 encodes a protein highly similar to the cauliflower Or gene product, a plastid-associated protein with a cysteine-rich DnaJ domain. A dominant Or mutation induces -carotene accumulation in cauliflower inflorescences, suggesting that Or is somehow involved in the control of chromoplast differentiation (Lu et al., 2006).

Several TCs encode putative transcription factors (Table 18). The most abundantly expressed, Cl000348:1, encodes a Myb-like protein with high similarity to LhMyb (from Lilium, GenBank accession BAB40790) Myb8 (from Gerbera (Elaoma et al., 2003) – also showing similarity to Cl000348:2) and Myb305 (From Antirrhinium (Jackson et al., 1991)). All three factors are highly expressed in flowers. Also highly expressed is Cl001329:1, encoding a putative MADS box transcription factor. This protein shows high similarity to AODEF, a B-functional transcription factor from Asparagus expressed in stamens and inner tepals (Park et al., 2003) and to LMADS1, a lily protein whose ectopic expression in dominant negative form causes an ap3-like phenotype in Arabidopsis (Tzeng and Yang, 2001).

Finally, several TCs - Cl000209:1 (61 ESTs) Cl000582:1 (18 ESTs) Cl001827:1 (5 ESTs) and Cl000731(2 ESTs) - show similarity to potyviral sequences, indicating that the sequenced library likely derives from potyvirus-infected tissue. Potyviruses like Iris Mild Mosaic Virus are known to infect Crocus (Navalinskijene and Samuitiene, 2001). The sequences of these TCs will prove useful for diagnostic and phytosanitary purposes.

3.8 On the Gluthatione S-transferase gene family in *Citrus sinensis*

Glutathione *S* -transferases (GSTs; EC 2.5.1.18) are an ancient and ubiquitous gene family encoding ~ 25- to 29-kD proteins that form both homodimers and heterodimers *in vivo*.

Historically, GST enzymes were first discovered in animals in the 1960s for their importance in the metabolism and detoxification of drugs (Wilce and Parker, 1994). Their presence in plants was recognized shortly afterwards, in 1970, when a GST activity from maize was shown to be responsible for protecting the crop from injury by the chloro-S-triazine atrazine herbicide (Frear and Swanson 1970).

Thereby, GSTs were thought as detoxification enzymes which are liable for the inactivation of toxic chemical compounds by catalysing their conjugation to glutathione (GSH). GSTs recognize not only reactive electrophilic xenobiotic molecules (i.e. drugs or herbicides) but also compounds that are of endogenous origin. In plants, many secondary metabolites are phytotoxic, even for the cells that produce them, and thereby the targeting to the vacuole is crucial (Martinoia et al. 1993). Anthocyanin pigments, for example, require GSH conjugation for transport into vacuole since their cytoplasmic retention is toxic to the cells and prevents the synthesis of new anthocyanins.

This was demonstrated first in 1995 by Marrs et al. who suggested the maize gene *Bronze-2* to be a glutathione S-tranferase involved in vacuolar transfer of anthocyanins. Furthermore Mueller et al. (2000) evidenced that the glutathione S-transferase *AN9* from *Petunia hybrida* is a flavonoid-binding protein required for efficient anthocyanin export into the vacuole, where it is permanently stored.

Herein we characterized the *Citrus sinensis* GST gene family by screening a large collection of expressed sequence tags. We focused our attention on *Citrus sinensis* (L. Osbeck) glutathione S-transferase because of the role of this gene family on anthocyanins vacuolarization. The accumulation of anthocyanins confers to blood oranges (cultivar Tarocco, Moro, Sanguinello) the typical dark red colour (the most important characteristic of Sicilian and Italian oranges), softness and eyes good quality. In addition to these characteristics, the presence of anthocyanins in orange fruits have a deep impact on human health since these molecules act as scavenger of free radicals and prevents inflammatory and cardiovascular diseases.

3.8.1 *In silico* identification and tissue expression profiling of GST encoding transcripts

Different members of the *C. sinensis* GST gene family were identified by the *in silico* screening procedure described in Methods 2.5. Only 25 GST sequences, which we classified as full length mRNAs (see Methods 2.5.3 and Figure 12) were analysed by Semi-Quantitative RT-PCR experiments. First of all these experiments are valuable because they let the *in silico* defined GST transcripts be confirmed. In addition, they point out other findings such as differences in the gene expression levels between the blood (Moro nucellare) and the common (Blonde cadenera) orange and tissue specific expression profiles. These findings may be relevant for the comprehension and the characterization of phenotype differences between the orange cultivars that we are investigating on.

Tissue expression profiling was performed on 6 different tissues. Furthermore, DNA band patterns are obtained by analysing as a pigmented (blood) as a non-pigmented (blond) orange cultivars (Figure 32, 33 and 34). This in order to point out differences in the DNA band patterns with respect to the different tissue analysed, as well as substantial changes in gene expression level due to different genotypes we investigated.

In the RT- PCR panel which refers to the PCR amplification performed on the Phi class GST sequences, we point out that the CITSI29:1 seems to be a tissue specific GST because DNA band patterns are evident exclusively in the ‘young leaf’ and ‘adult leaf’ lanes (red box in Figure 32). However, this evidence contrasts whit the one inferred by evaluating EST-based tissue information. Indeed, the sequence CITSI29:1 is generated by assembling 4 ESTs (2 from flavedo, 1 from flower and 1 from callus) (see Methods 2.5.1 Table 5). In addition a consistent difference in the gene expression levels in the young leaf tissue is observed between the 2 genotypes.

Remarkable difference in gene expression levels between the two genotypes are also observed in the lane ‘flesh’ concerning the sequence CITSI02:1+CITSI00:1. This gene has a higher expression in the Moro Nucellare cultivar as shown in the figure 32 (green box).

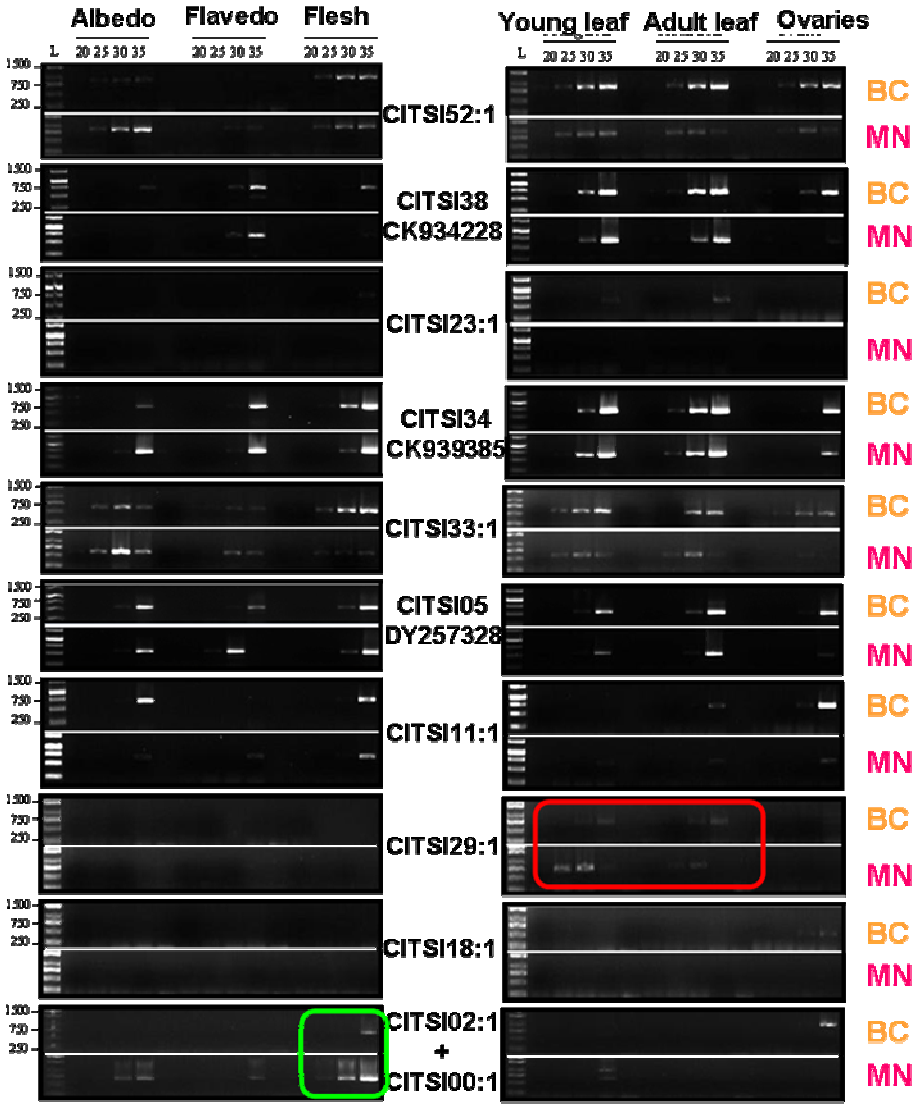


Figure 32. Tissue expression profile analysis on Phi class GST sequences.
SemiQ RT-PCR experiments are performed considering the 6 following tissues: fruit tissues (albedo, flavedo and flesh) in ripening time, young and adult leaf and ovaries. Two different orange genotypes are considered: Blond Cadenera (BC) and Moro Nucellare (MN). PCR amplifications are reported at cycle 20, 25, 30 and 35.

In the panel that reports the PCR amplification performed on the Tau class GST sequences (Figure 33), it is manifest an over-expression of the sequence CITSI51:1 in the genotype Blond Cadenera and in all the tissue considered apart the young leaf tissue where no amplification occurred in both of the genotypes (red box). In addition, in some cases, double bands can be observed in the lane “ovaries” (Figure 33 blue arrows). This is not surprising if we take into account the ovary as an organ rather than a tissue. Furthermore, some bands in the “ovaries” lanes are of a size longer than the expected ones. It is likely that these bands can be the results of PCR amplifications of intron

retaining mRNAs whose expression can be regulated during the development or in a tissue-specific manner (ovaries comprise multiple tissues).

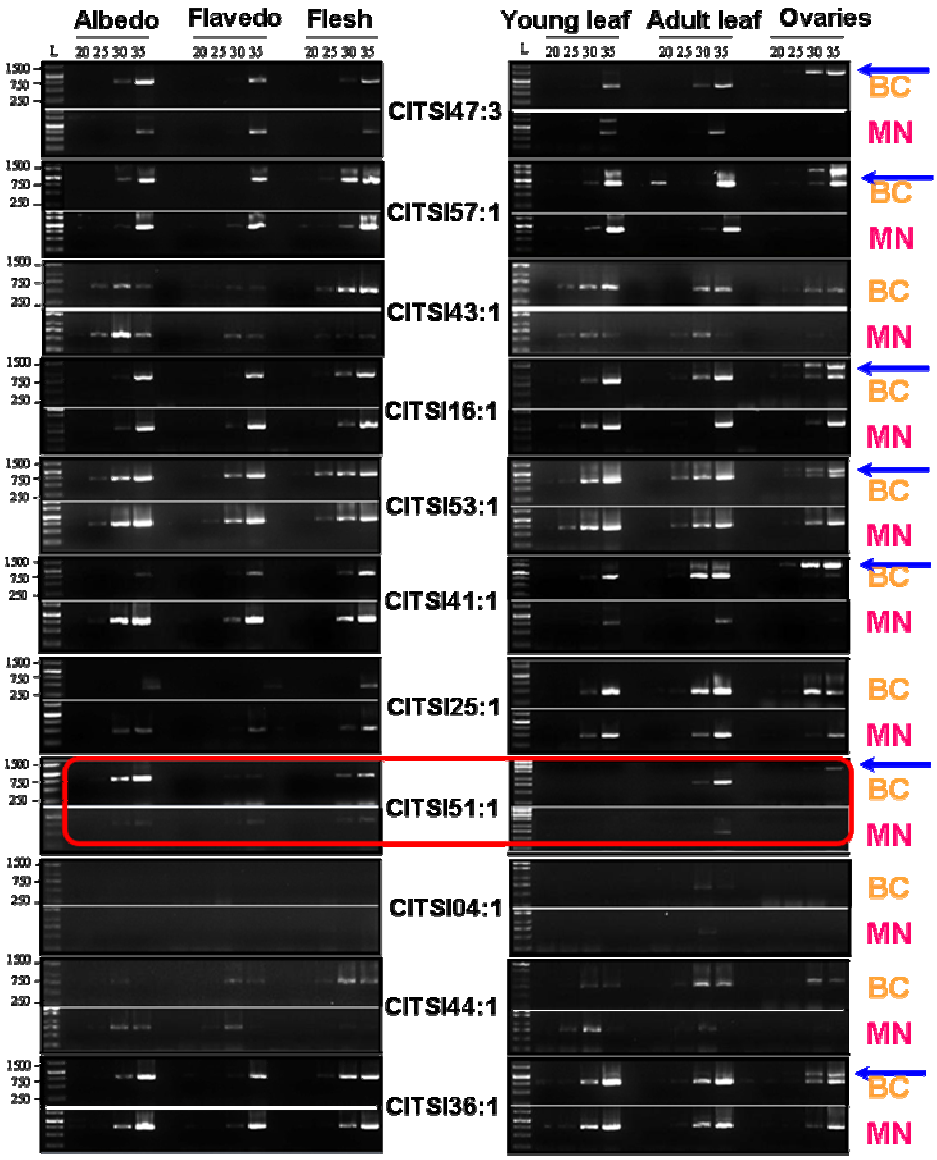


Figure 33. Tissue expression profile analysis on Tau class GST sequences.
SemiQ RT-PCR experiments are performed considering the 6 following tissues: fruit tissues (albedo, flavedo and flesh) in ripening time, young and adult leaf and ovaries. Two different orange genotypes are considered: Blond Cadenera (BC) and Moro Nucellare (MN). PCR amplifications are reported at cycle 20, 25, 30 and 35.

The PCR amplification results concerning the remaining GST sequences are reported in figure 34. Multiple bands are observed for the sequence zeta (red box) because of no specific primer selection.

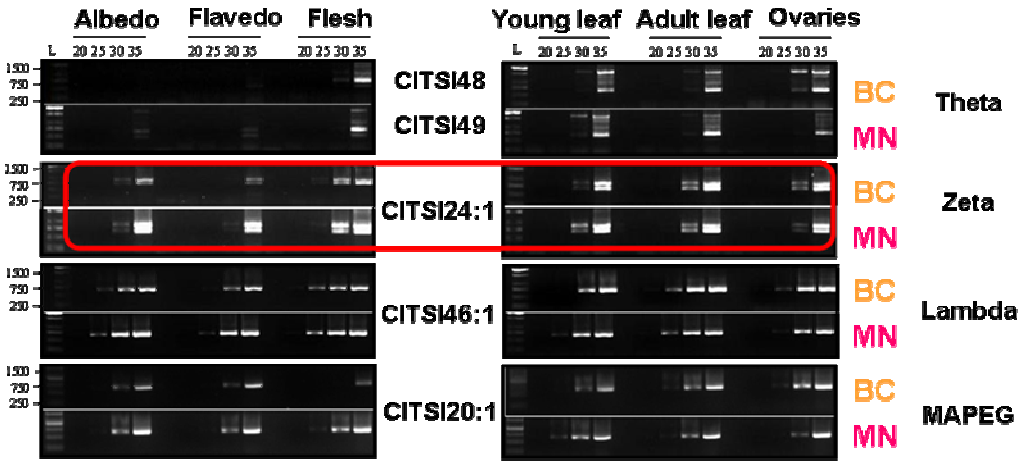


Figure 34. Tissue expression profile analysis on Theta, Zeta, Lambda, MAPEG class GST sequences.

SemiQ RT-PCR experiments are performed considering the 6 following tissues: fruit tissues (albedo, flavedo and flesh) in ripening time, young and adult leaf and ovaries. Two different orange genotypes are considered: Blond Cadenera (BC) and Moro Nucellare (MN). PCR amplifications are reported at cycle 20, 25, 30 and 35.

4 DISCUSSION

4.1 The significance of EST in the 'omics' era

Shotgun sequencing of genomes through Expressed Sequence Tags has proved to be a rapid method of identifying a significant proportion of genes of a target organism. EST sequencing, on the one hand certainly avoids the biggest problems associated with genome size and the accompanying repetitiveness, on the other hand does not yield sequences for all of the expressed genes of a target organism. Some genes, in fact, may not be expressed under the sampled conditions, others may be expressed at very low levels and missed through the random sampling that underlies the library design and the sequencing strategy. However the creation of EST libraries from a range of conditions such as different tissues, developmental stages or environmental exposures, supports a closer examination of the biology of the species under investigation.

As the ability of scientific investigations to produce large amounts of "EST sequence data" has become mainstream, the need to handle, gather, store, process and analyse them has made the role of bioinformatics essential for biological or plant science projects.

The notion of an “analysis pipeline” for processing and analysing large batches of EST sequences has become familiar. Appropriate individual bioinformatics tools and pipelines pertaining to EST analysis have been built and actively used (http://biolinfo.org/EST/).

During the first year of my PhD program, the “analysis pipeline” ParPEST (D’Agostino et al., 2005) was implemented to integrate all the consecutive steps of the analysis (i.e. EST pre-processing, clustering, assembling, consensus generation and tools for DNA and protein annotation) into a single procedure. It has a modular organization and each module corresponds to each step of the analysis.

ParPEST have been designed in a parallel environment because parallel computing is effective in reducing the execution time of the different steps of the EST analysis. Run-time efficiency, in fact, is a fundamental task considering that EST data are in continuous upgrading.

High-throughput analyses of ESTs often encounter data management challenges and the presentation of these data to the scientific community can be rather challenging. The first logical step is to define a database architecture in order to properly organize the biological data entities and relationships among biological data sources. Then, immediate and clear ways of visual representation of data are needed for the

dissemination of the information produced to the scientific community. To this end, the design and the building of a database-driven Web applications is strongly required.

To accomplish these requirements, I designed an entity-relationship diagram, a data modeling tool that is worthwhile in organizing the data into entities and in defining their relationships, in whatever EST project. It expresses the overall logical structure of each dedicated database that we built for data storage. In addition, a browsing data system has been intended to support non-expert users. The search engine underpins on Web-based application and the results returned for each query are displayed in a user-friendly manner. Different systems of classification are used to describe features and functions and categorize all information referred to each expressed sequence. HTML-based tree menus are the facilities we selected for graphical listing of enzymes as well as of metabolic pathways associated to each expressed sequence. Furthermore, we implemented the "on-the-fly" mapping of specific data, as they issued from the user-selected criteria, onto metabolic pathways which can be accessed as GIF images.

Bioinformatics analysis of Solanaceae sequence data and the leveraging of these data are the main goals of my PhD program. The large-scale production of ESTs from different Solanaceae species goes with the International Tomato Genome Sequencing Project. Since, complete sequencing of the tomato genome is ongoing, an affordable solution to study the Solanaceae biology is to develop Expressed Sequence Tag databases which provide a wealth of information in a relatively short time. The long term goal is to establish a well-characterized, non-redundant EST resource for the Solanaceae genomics community.

For this reason, during the three years of my PhD program, Expressed Sequence Tag databases for different Solanaceae species have been built. They will serve as the most abundant source of new coding sequences available today as well as a source of genes of value to agriculture. Furthermore they will support the study of Solanaceae biology and certainly provide a consistent resource for gene discovery, genome annotation, gene expression studies and comparative genomics. All EST sequence data from multiple tomato species are compiled in the TomatEST database (D'Agostino et al., 2007a). They are powerful tools in the hunting for known genes and can be used to help the identification of unknown genes and to map their positions within the tomato genome sequences. This facilitates the structural and the functional annotation of the tomato genome sequences, the international Tomato Genome Sequencing Consortium is releasing.

As members of the international Tomato Annotation Group (iTAG) we will provide the public with a high quality and homogeneous annotation of the whole tomato genome. We are committed, within the tomato genome annotation pipeline, to generate EST-to-genome alignments. We included in the analysis ESTs of non-native origin (i.e. EST data compiled in the PotatEST database and other Solanaceae ESTs) because they will improve the accuracy of gene annotation (genes, which lack source-native EST evidence, should remain otherwise undiscovered) and will help to identify orthologs conserved among different species.

Tracks showing EST/TC-to-genome alignments are released to the scientific community through the Gbrowse Web application (Stein et al., 2002) at http://biosrv.cab.unina.it/GBrowse/. Each track is cross-linked to local or external databases so as to associate the predicted gene structure to a preliminary biological function. In addition, the availability of large numbers of EST-to-genome sequence alignments represents a valuable source in the task of defining a consistent number of reliable gene models. A reference set of species-specific gene models is needed to train *ab initio* gene finder and is one of the primary goal of the iTAG. On one hand, the definition of a reliable set of gene models is performed manually by expert annotators who filter out all the possible conflicts and inconsistencies which could sidetrack the training of gene-finder tools. On the other hand, the vast amount of data to analyse represents a drawback for the human annotation. In this reference frame, we developed GeneModelEST (D'Agostino et al., 2007b), a "pipeline analysis" which aims to automatically build a reliable set of gene models. GeneModelEST permits the tentative consensus sequences of source-native as well as of non-native origin to be properly classified (see Methods 2.3.3) and, therefore, to be selected as candidate gene models.

EST data are helpful also to establish the viability of alternative transcripts such as alternative splicing, initiation, polyadenylation, and intron retention. Alternative transcription is an important mechanism of modulating gene expression and function as well as of expanding proteome diversity. To investigate how frequent alternative transcription is in each data-set, we checked on all the clusters which are assembled into multiple tentative consensus sequences (assemblies). Indeed, since the clustering process is a simple 'tentative closure' procedure, the clustering program will incorporate overlapping ESTs which tag the same gene in a single cluster, not considering if they make sense all together. When sequences in a cluster cannot be all reconciled into a consistent multiple alignment, during the much more rigorous

assembly phase, they are accordingly split into multiple assemblies. The foremost interpretation of multiple assemblies from a cluster is precisely alternative transcription. However, other possible interpretations could be paralogy or protein domain sharing. To corroborate this kind of evidence, we exploited the tomato genome sequences and evaluated if multiple assemblies from the same cluster resulted aligned in the same tomato genomic region. As an example we discussed the instance about the cluster SOLLCCl018639 from *S. lycopersicum* (Figure 35) This EST cluster is made up of 18 sequences which have been split into 3 distinct assemblies. The SOLLCCl018639:1 sequence is generated by assembling 13 ESTs from different tissue types (such as: root, flower, leaf...). On the contrary, the remaining 2 sequences, SOLLCCl018639:2 and SOLLCCl018639:3, seem to be tissue-specific transcripts since the first is generated by assembling 3 ESTs from the carpel tissue, while the second is generated by assembling 2 ESTs from the pericarp tissue. All the transcripts are mapped in the same genomic regions of a BAC anchored to the chromosome 10.

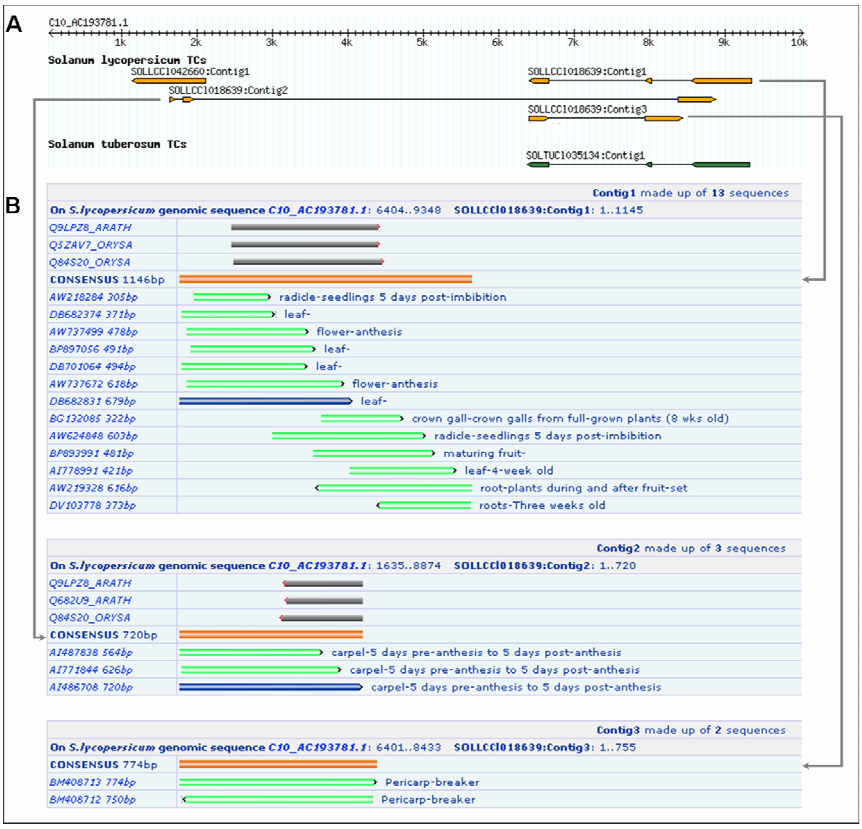


Figure 35. Example of alternative transcription established by *S.lycopersicum* EST data.
Panel **A** shows the three TCs (SOLLCCl018639:1; SOLLCCl018639:2 and SOLLCCl018639:3). from the cluster SOLLCCl018639 aligned in the same genomic region of a BAC anchored to tomato chromosome 10. The TC SOLLCCl018639:1 has a *S. tuberosum* counterpart represented by the sequence SOLTUCl035134:1.
Panel **B** shows the representation of the multiple alignments of the EST reads (green and blue bars) which generate the three TCs (orange bars). In correspondence to each EST the tissue/developmental stage origin is reported.

ESTs are helpful not only to identify a set of genes but also to gain an understanding of when, where, and how a gene is turned on. This process is commonly referred to as gene expression. Gene expression profiling holds tremendous promising for dissecting the regulatory mechanisms and transcriptional networks that underlie biological processes. Furthermore the identification of genes differentially expressed in different tissues, during development, during specific biotic or abiotic stress, is of foremost interest to both basic and breeder researches.

We used the so-called *in silico* transcriptional profiling, which is performed by counting the number of sequenced ESTs for a given gene within the whole sequenced EST population, to identify *Crocus sativus* genes that were highly expressed in the stigma tissue. By querying the Saffron genes database (D'Agostino et al., 2007c), tentative consensus sequences composed of most abundant ESTs can be retrieved. A series of interesting sequences, such as putative sex determination genes, lipid and carotenoid metabolism enzymes and transcription factors have been identified. They underlie the molecular biology of stigma biogenesis as well as biochemical functions occurring in saffron secondary metabolism.

Particularly important is the fact that this type of data-mining can be used to corroborate and extend upon the expression data obtained from micro-array experiments. For this reason, we established, for example, correspondences between *S. lycopersicum* EST data set and the Affymetrix Tomato Genome Array probe-sets. Forthcoming approaches will consider a comprehensive data mining of EST and expression arrays data.

A further relevant application of EST data concerns comparative genomics studies.

The enormous biological sequence data thus flooding into the EST databases necessitates the development of efficient tools for comparative genome sequence analysis as starting point to understand species diversification and evolution. Comparative genomic analysis, which involves the comparison of two complete genomes or sets of gene products from two different organisms, is the cornerstone of *in silico*-based approaches to understanding biological systems and processes across plant species. Since tomato and potato ESTs are the most number-consistent data-sets we collected in our Solanaceae repository, comparison of gene inventories (from EST collections) allows us to address a fundamental question about what makes tomato species different from potato and to investigate the evolutionary relationships between related Solanaceae species.

Here we presented the results of systematic analysis of the tomato and potato non-redundant EST sets so that gene families in common between the two species as well as species-specific genes could be identified. The protein complement of the model plant *Arabidopsis thaliana*, is used as first attempts to estimate common proteins (orthology) among the three species. It is likely that these sequences represent, according to their pattern of sequence similarity, plant-specific proteins. All the unique transcripts from tomato and potato with no match to the Arabidopsis proteome, are then analysed by a pair-wise comparison strategy. This two-step combined strategy permitted to estimate the extent to which the *S. lycopersicum* and *S. tuberosum* transcriptomes overlap and to isolate those sequences that are likely to be species-specific genes.

This facet can be explored in a meaningful way also considering the data from EST-to-genome mapping analysis, so that to evaluate if each tomato transcript have a potato counterpart mapped onto tomato genome.

Last but not least, we have developed a comprehensive expressed sequence tag database search method and used it for the identification of new members of the Gluthatione S-transferase superfamily in *Citrus sinensis*.

The research topic of “taking a group of related sequences and compare them to investigate on the enzymes that they encode and on their expression patterns”, expands the interface between bioinformatics and experimental biology and highlights that computation coupled with experiments will still provide the most reliable way of performing research. SemiQ RT-PCR experiments propped up results from bioinformatics analyses and showed the effectiveness as well as the mere existence of the *in silico* defined GST transcripts. In addition GST tissue-specific expression patterns, inferred by querying the dbEST database with respect to the different tissues, are comparable to those revealed by SemiQ RT-PCR. Thus, this is a clear example on how "united we stand, divided we fall", and on how the challenge of integration concerns not only large amount of data but also skills and expertise.

The most of the data-mining tools, which have been developed for the EST applications we discussed, are converged into the computational platform ISOL@ (Chiusano et al., 2007). ISOL@ meets the need to collect, integrate and explore high-throughput and heterogeneous biological data and aims to enhance the quality of the data gathered. Since it is designed as a multi-level computational environment, its is thought to be flexible and to easily evolve in consideration of the continuous production of new data and novel analysis methods. We believe that the need to investigate on the structure, the

function and the evolution of plant genomes represents a suitable test bench to challenge and expand this effort.

5 CONCLUSION

High-throughput EST analysis requires integrated and automated approaches enabling EST data mining. Furthermore *ad hoc* methods for data storage, data warehousing, data integration, data visualization and data modeling are fundamental. So bioinformatics becomes pre-eminent and is directly dependent on the efficiency of data integration and on the value added information which they produce. This is, in turn, determined by the diversity of data sources and by the quality of the annotation they are endowed with.

Suitable bioinformatics methods permits the undeniable value of ESTs to be exploited to address different and complex biological questions crossing the '-omics' barrier for "whole-istic biology" interpretations (Chong and Ray, 2002). The importance of exploring the data as a whole is recognized as the scope of contemporary science.

.

ACKNOWLEDGEMENTS

This thesis was improved by conversations with a large number of people who helped to debug it.

Thanks to Luigi Frusciante, who supervised my activities concerning the PhD project.

Particular thanks to Maria Luisa Chiusano, who directed my work and helped to develop the analysis.

Perceptive criticisms also came from Giovanni Giuliano.

I'm grateful to all the members of CAB group, who have played a decisive role in supporting me. These persons are: Maria Luisa Chiusano, Mario Aversano, Alessandra Traini, Enrico Raimondo and Concetta Licciardello. You were very patient and beared with my bad temper. Without your willingness, suggestions and job offers, this PhD thesis would not have been written.

Also to my parents, who are still waiting for me to quit 'fooling around with computers' and despite that they urge me in any way.

Finally, I would like to thank Marisa, whose backing has been very encouraging. During a critical period of my life, she forgot, neither for a moment, my feelings or my character.

6 LITERATURE CITED

Adams M.D., Kelley J.M., Gocayne J.D., Dubnick M., Polymeropoulos M.H., Xiao H., Merril C.R., Wu A., Olde B., Moreno R.F., et al. (1991) **Complementary DNA sequencing: Expressed sequence tags and human genome project.** Science. 252:1651-1656.

Audic S., Claverie J.M. (1997) **The significance of digital gene expression profiles.** Genome Research. 7(10):986-95.

Bairoch A. (2000) **The ENZYME database in 2000**. Nucleic Acids Research. 28:304-305.

Boguski M.S., Lowe T.M., Tolstoshev C.M. (1993) **dbEST-database for "expressed sequence tags".** Nature Genetics. 4:332–333.

Bouvier F., Dogbo O., Camara B. (2003a) **Biosynthesis of the food and cosmetic plant pigment bixin (annatto)**. Science. 300(5628):2089-2091.

Bouvier F., Suire C., Mutterer J., Camara B. (2003b) **Oxidative remodeling of chromoplast carotenoids: Identification of the carotenoid dioxygenase CsCCD and CsZCD genes Involved in Crocus secondary metabolite biogenesis.** Plant Cell. 15(1):47-62.

Brendel V., Xing L., Zhu W. (2004) **Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus.** Bioinformatics. 20(7):1157-1169.

Brett D., Hanke J., Lehmann G., Haase S., Delbrück S., Krueger S., Reich J., Bork P. (2000) **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** FEBS Letters. 474:83–6.

Burke J., Davidson,D. and Hide,W. (1999) **d2_cluster: A validated method for clustering EST and full-length cDNA.** Genome Research. 9:, 1135–1142.

Caicedo A. L. and Purugganan M.D. (2005) **Comparative Plant Genomics. Frontiers and Prospects.** Plant Physiology. 138(2): 545 – 547.

Calderon-Urrea A., Dellaporta S.L. (1999) **Cell death and cell protection genes determine the fate of pistils in maize**. Development. 126(3):435-441.

Carol P., Kuntz M. (2001) **A plastid terminal oxidase comes to light: implications for carotenoid biosynthesis and chlororespiration**. Trends Plant Science. 6(1):31-36.

Castillo R., Fernandez J.A., Gomez-Gomez L. (2005) **Implications of Carotenoid Biosynthetic Genes in Apocarotenoid Formation during the Stigma Development of Crocus sativus and Its Closer Relatives**. Plant Physiology. 139(2):674-689.

Chiusano M.L., D'Agostino N., Traini A., Licciardello C., Raimondo E., Aversano M., Frusiante L. (2007) **ISOL@: an Italian SOLAnaceae genomics resource.** BMC Bioinformatics, accepted

Chong, L., Ray, L.B. (2002) **Whole-istic Biology.** Science 295(1):1661.

Christoffels A., van Gelder A., Greyling G., Miller R., Hide T., Hide W. (2001) STACK: **Sequence Tag Alignment and Consensus Knowledgebase.** Nucleic Acids Research. 29:234–238.

D'Agostino N., Aversano M., Chiusano M.L. (2005) **ParPEST: a pipeline for EST data analysis based on parallel computing.** BMC Bioinformatics. 6(Suppl. 4):S9.

D'Agostino N., Aversano M., Frusciante L., Chiusano M.L. (2007a) **TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species.** Nucleic Acids Research. 35(Database issue):D901-5.

D'Agostino N., Traini A., Frusciante L., Chiusano M.L. (2007b) **Gene models from ESTs (GeneModelEST): an application on the *Solanum lycopersicum* genome.** BMC Bioinformatics. 8(Suppl 1)**:**S9

D'Agostino N., Pizzichini D., Chiusano M.L., Giuliano G. (2007c) **An EST database from saffron stigmas.** BMC Plant Biology.7(1):53

Deruere J., Romer S., d'Harlingue A., Backhaus R.A., Kuntz M., Camara B. (1994) **Fibril assembly and carotenoid overaccumulation in chromoplasts: a model for supramolecular lipoprotein structures**. Plant Cell. 6(1):119-133.

Dong Q., Kroiss L., Oakley F.D., Wang B.B., Brendel V. (2005a) **Comparative EST analyses in plant systems.** Methods Enzymology. 395:400-18.

Dong Q., Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V. (2005b) **Comparative Plant Genomics Resources at PlantGDB.** Plant Physiology. 139:610-618.

Elomaa P., Uimari A., Mehto M., Albert V.A., Laitinen R.A., Teeri T.H. (2003) **Activation of anthocyanin biosynthesis in Gerbera hybrida (Asteraceae) suggests conserved protein-protein and protein-promoter interactions between the anciently diverged monocots and eudicots**. Plant Physiology. 133(4):1831-1842.

Ewing B. and Green P. (1998) **Base-calling of automated sequencer traces using phred. II. Error probabilities.** Genome Research.8(3):186-94.

Ewing R.M., Ben Kahla A., Poirot O., Lopez F., Audic S., Claverie J.M. (1999) **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression**. Genome Research. 9:950–959.

Faccioli P., Provero P., Herrmann C., Stanca A.M., Morcia C., Terzi V. (2005) **From single genes to co-expression networks: extracting knowledge from barley functional genomics.** Plant Molecular Biology 58(5):739-50.

Felsenstein J. (1993) **PHYLIP (Phylogeny Inference Package) version 3.5c**. Seattle:University of Washington.

Fernandez J.A. (2004) **Biology, biotechnology and biomedicine of saffron.** Recent Research Development in Plant Science. 2:127-159.

Frear D.S., Swanson H.R. (1970) **Biosynthesis of S-(4 ethylamino 6-isopropylamino 2-s-triazino) glutathione; partial purification and properties of a glutathione S-transferase from corn.** Phytochemistry. 9: 2123–2132

Frova C. (2003) **The plant glutathione transferase gene family: genomic structure, functions, expression and evolution.** Physiologia Plantarum. 119: 469–4

Galante P.A., Sakabe N.J., Kirschbaum-Slager N., de Souza S.J. (2004) **Detection and evaluation of intron retention events in the human transcriptome.** RNA. 10(5):757-65.

Gautheret D., Poirot O., Lopez F., Audic S., Claverie J.M. (1998) **Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering.** Genome Research. 8(5):524-30.

Giuliano G., Rosati C,. Bramley P.M. (2003) **To dye or not to dye: biochemistry of annatto unveiled**. Trends in Biotechnology. 21(12):513-516.

Gonzalez-Guzman M., Apostolova N., Belles J.M., Barrero J.M., Piqueras P., Ponce M.R., Micol J.L., Serrano R., Rodriguez P.L. (2002) **The short-chain alcohol dehydrogenase ABA2 catalyzes**

the conversion of xanthoxin to abscisic aldehyde. Plant Cell. 14(8):1833-1846.

Gremme G., Brendel V., Sparks M.E. and Kurtz S. (2005) **Engineering a software tool for gene structure prediction in higher organisms.** Information and Software Technology, 47(15):965-978.

Griffiths-Jones S., Moxon S., Marshall M., Khanna A., Eddy S.R and Bateman A. (2005) **Rfam: annotating non-coding RNAs in complete genomes.** Nucleic Acids Research. 33:D121-D124.

Gupta S., Zink D., Korn B., Vingron M., Haas S.A. (2004) **Genome wide identification and classification of alternative splicing based on EST data.** Bioinformatics. 20(16):2579-85.

Hall T.A. (1999) **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT**. Nucleic Acids Symposium. Series. 41:95-98

Henikoff S., Henikoff J.G. (1997) **Embedding strategies for effective use of information from multiple sequence alignments.** Protein Science. 6:698-705.

Huang X. and Madan A. (1999) **CAP3: A DNA sequence assembly program.** Genome Research. 9:868-877.

Jackson D., Culianez-Macia F., Prescott A.G., Roberts K., Martin C. (1991) **Expression patterns of myb genes from Antirrhinum flowers**. Plant Cell. 3(2):115-125.

Jiang J., Jacob H.J. (1998) **EbEST: an automated tool using expressed sequence tags to delineate gene structure.** Genome Research 8(3):268-75.

Jurka J. (2000) **Repbase Update: a database and an electronic journal of repetitive elements.** Trends in Genetics. 9**:**418-420.

Kalyanaraman A., Aluru S., Kothari S., Brendel V. (2003) **Efficient clustering of large EST data sets on parallel computers**. Nucleic Acids Research. 31:2963-2974.

Kan Z., Rouchka E.C., Gish W.R., States D.J. (2001) **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** Genome Research. 11:889-900.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., and Hirakawa M. (2006) **From genomics to chemical genomics: new developments in KEGG.** Nucleic Acids Research. 34: D354-357.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. and Higgins D.G. (2007) **ClustalW and ClustalX version 2.** Bioinformatics. 23(21): 2947-2948.

Lee Y., Tsai J., Sunkara S., Karamycheva S., Pertea G., Sultana R., Antonescu V., Chan A., Cheung F., Quackenbush J. (2005) **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** Nucleic Acids Research. 33:D71–D74.

Licciardello C., Russo M.P., Valè G., Reforgiato G.R. (2007) **Identification of differentially expressed genes in the flesh of blood and common oranges.** Tree Genetics & Genomes. 1614-2942 (Print); 1614-2950 (Online)

Lisacek F.C., Traini M.D., Sexton D., Harry J.L., Wilkins M.R. (2001) **Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting.** Proteomics. 1(2):186-93.

Lo Piero A.R., Puglisi I., Petrone G. (2006) **Gene isolation, analysis of expression, and in vitro synthesis of glutathione S-transferase from orange fruit [Citrus sinensis L. (Osbeck)].** Journal of Agricultural and Food Chemistry. 54(24):9227-33.

Lorenzen J.H., Balbyshev N.F., Lafta A.M., Casper H., Tian X., Sagredo B. (2001) **Resistant potato selections contain leptine and inhibit development of the Colorado potato beetle (Coleoptera: Chrysomelidae).** Journal of Economic Entomology. 94(5):1260

Lu S., Van Eck J., Zhou X., Lopez A.B., O'Halloran D.M., Cosman K.M., Conlin B.J., Paolillo D.J., Garvin D.F., Vrebalov J., Kochian L.V., Kupper H., Earle E.D., Cao J., Li L. (2006) **The cauliflower Or gene encodes a DnaJ cysteine-rich domain-containing protein that mediates high-levels of {beta}-carotene accumulation**. Plant Cell. 18:3594-3605.

MacIntosh G.C., Wilkerson C., Green P.J. (2001) **Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs.** Plant Physiology. 127(3):765-76.

Martinoia E., Grill E., Tommasini R., Kreuz K., Amrhein N. (1993) **ATP-dependent glutathione S-conjugate 'export' pump in the vacuolar membrane of plants.** Nature. 364:247–249.

Mégy K., Audic S., Claverie J.M. (2003) **Heart-specific genes revealed by expressed sequence tag (EST) sampling.** Genome Biology.3(12):RESEARCH0074.

Moraga A.R., Nohales P.F., Perez J.A., Gomez-Gomez L. (2004) **Glucosylation of the saffron apocarotenoid crocetin by a glucosyltransferase isolated from Crocus sativus stigmas.** Planta. 219(6):955-966.

Marrs K.A. (1996) **The functions and regulation of glutathione S-transferases in plants.** Annual Review of Plant Physiology and Plant Molecular Biology. 47: 127–58.

Marrs K.A., Alfenito M.R., Lloyd A.M., Walbot V. (1995) **A glutathione S-transferase involved in vacuolar transfer encoded by the maize gene Bronze-2.** Nature.375(6530):397-400.

Mueller L.A., Goodman C.D., Silady R.A., Walbot V. (2000) **AN9, a petunia glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein.** Plant Physiology. 123: 1561–1570.

Mueller L.A., Solow T.H., Taylor N., Skwarecki B., Buels R., Binns J., Lin C., Wright M.H., Ahrens R., Wang Y., et al. (2005a) **The SOL Genomics Network: a comparative resource for *Solanaceae* biology and beyond.** Plant Physiology. 138:1310–131.

Mueller L.A., Tanksley S.D., Giovannoni J.J., van Eck J., Stack S., Choi D., Kim B.D., Chen M., Cheng Z., Li C., et al. (2005b) **The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL).** Comparative and Functional Genomics. **6:**153-158.

Nagaraj S.H., Gasser R.B., Ranganathan S. (2007) **A hitchhiker's guide to expressed sequence tag (EST) analysis.** Briefings in Bioinformatics 8(1):6-21.

Navalinskijene M., Samuitiene M. (2001) **Viruses affecting some bulb and corm flower crops**. Biologija. 4:40-42.

Park J.H., Ishikawa Y., Yoshida R., Kanno A., Kameya T. (2003) **Expression of AODEF, a B-functional MADS-box gene, in stamens and inner tepals of the dioecious species Asparagus officinalis L**. Plant Molecular Biology. 51(6):867-875.

Picoult-Newberg L., Ideker T.E., Pohl M.G., Taylor S.L., Donaldson M.A., Nickerson D.A., Boyce-

Jacino M. (1999) **Mining SNPs from EST databases.** Genome Research. 9(2):167-74.

Pontius J.U., Wagner L., Schuler G.D. (2003) **UniGene: a unified view of the transcriptome.** The NCBI Handbook.

Rensink W.A., Lee Y., Liu J., Iobst S., Ouyang S., Buell C.R. (2005) **Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts.** BMC Genomics. 146:124.

Rhee S.Y., Beavis W., Berardini T.Z., Chen G., Dixon D., Doyle A., Garcia-Hernandez M., Huala E., Lander G., Montoya M., Miller N. et al. (2003) **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.** Nucleic Acids Research. 31(1):224-8.

Richmond T. and Somerville S. (2000) **Chasing the dream: plant EST microarrays.** Current Opinion in Plant Biology. 3: 108-116.

Ronning C.M., Stegalkina S.S., Ascenzi R.A., Bougri O., Hart A.L., Utterbach T.R., Vanaken S.E., Riedmuller S.B., White J.A., Cho J. et al. (2003) **Comparative analyses of potato expressed sequence tag libraries.** Plant Physiology.131(2):419-29.

Schlueter S.D., Dong Q., Brendel V. (2003) **GeneSeqer@PlantGDB: Gene structure prediction in plant genomes.** Nucleic Acids Research. 31(13):3597-600.

Stein L.D., Mungall C., Shu S., Caudy M., Mangone M., Day A., Nickerson E., Stajich J.E., Harris T.W., Arva A., Lewis S. (2002) **The generic genome browser: a building block for a model organism system database.** Genome Research. 12(10):1599-610.

The Gene Ontology Consortium. (2000) **Gene Ontology: tool for the unification of biology.** Nature Genetics. 25: 25-29.

The UniProt Consortium. (2007) **The Universal Protein Resource (UniProt).** Nucleic Acids Research. 35: D193-197.

Tzeng T.Y., Yang C.H. (2001) **A MADS Box Gene from Lily (Lilium longiflorum) is Sufficient to Generate Dominant Negative Mutation by Interacting with PISTILLATA (PI) in Arabidopsis thaliana.** Plant Cell Physiology. 42(10):1156-1168.

Wilce M.C.J., Parker M.W. (1994) **Structure and function of glutathione S-transferases.** Biochimica et Biophysica acta. 1205: 1-18.

Wu X., Knapp S., Stamp A., Stammers D.K., Jornvall H., Dellaporta S.L., Oppermann U. (2007) **Biochemical characterization of TASSELSEED 2, an essential plant short-chain dehydrogenase/reductase with broad spectrum activities**. Febs Journal. 274(5):1172-1182.

Wu X., Walker M.G., Luo J., Wei L. (2005) **GBA server: EST-based digital gene expression profiling.** Nucleic Acids Research. 33(Web Server issue):W673-6.

Yano K., Watanabe M.,Yamamoto N., Tsugane T., Aoki K., Sakurai N., Shibata D. (2006) **MiBASE: A database of a miniature tomato cultivar Micro-Tom** Plant Biotechnology 23:195–198.

Zamir D. and Tanksley S.D. (1988) **Tomato genome is comprised largely of fast-evolving, low copy-number sequences.** Molecular & general genetics. 213:254-261.