MHR

RESEARCH ARTICLE

H.-T. DONG, ET AL. [2005] MED HYPOTHESES RES 2: 418-496.

LARGE-SCALE IDENTIFICATION OF ESTS FROM NICOTIANA TABACUM BY NORMALIZED CDNA LIBRARY SEQUENCING

HAI-TAO DONG, WENZHENG LI, SANDRA PÉREZ ALVAREZ, HONGMEI LUO, YE DENG, QINGCHAO JIN CHEN'EN DAI, DANIEL CABEZAS MONTERO, YONGQI FANG , JING SHAO, YICHUN LOU, WENQIN SHI, YAN SHAO,YONGPING LI, JACK CHEN AND DE-BAO LI*

BIOINFORMATICS AND GENE NETWORK RESEARCH GROUP, COLLEGE OF AGRICULTURE AND BIOTECHNOLOGY, ZHEJIANG UNIVERSITY, HANGZHOU, 310029, CHINA; THE TOBACCO SCIENCE RESEARCH INSTITUTE OF YUNNAN PROVINCE, YUXI, 653100, CHINA; THE UNIT OF BIOCOMPUTATION, NICHD/NIH, BETHESDA, MD 20892, USA; THE LIFE SCIENCE COLLEGE,HANGZHOU TEACHERS COLLEGE, HANGZHOU, 310012, CHINA

ABSTRACT. AN EXPRESSED SEQUENCE TAGS (EST) resource for tobacco plants (Nicotiana tabacum) was established using high-throughput sequencing of randomly selected clones from 1 cDNA library representing a range of plant organs (leaf, stem, root and root base). Over 5000 ESTs were generated from the 3'-ends of 8000 clones, analyzed by BLAST searches, and categorized functionally. Clustering of these ESTs identified a unigene set of 3600 genes that are represented in the EST collection. 696 unique sequences (19.33% in TUTs) were identified as highly significant matches ($E \leq 10^{-20}$) to the known gene sequences and were regarded as representation of known genes according to results from the BLASTX research of the available database. All annotation ESTs were classified in 18 functional categories, with unique transcripts involved in energy the largest group accounting for 32.32% in all annotation ESTs. After excluding 2450 non-significant TUTs in sequence similarity, 100 unique sequences (1.67% in total TUTs) were identified from N. tabacum database and the remaining 1050 TUTs showed partial homology to genes in other organisms. Through comparing the hybridization data, 359 high-quality ESTs from four tobacco varieties were generated and analyzed using the Cyber-T statistic program, 7 EST from this analysis may have a relation to TMV resistance in tobacco plants, particularly to the Hongda variety. The array results were confirmed using a Real-time quantitative RT-PCR.

*ADDRESS ALL CORRESPONDENCE TO: DR. DE-BAO LI, BIOTECHNOLOGY INSTITUTE OF ZHEJIANG UNIVERSITY, NO. 268, KAIXUAN ROAD, ZHEJIANG PROVINCE, CHINA 310029.
TELEPHONE: 86-571-86971183. FAX: 86-571-86961525. E-MAIL ADDRESS: xchen@helix.nih.gov

• MEDICAL HYPOTHESES AND RESEARCH • THE JOURNAL FOR INNOVATIVE IDEAS IN BIOMEDICAL RESEARCH •

1. INTRODUCTION

The use of the word tobacco "Nicotiana tabacum" is generally accepted as referring to the products of the tabacum species [1]. Tobacco holds an unparalleled position among crop plants in the world such as: it is one of the very few crops entering world trade entirely on a leaf basis; it is the most widely grown commercial non-food plant in the world; it holds a high importance in financial and economic policies in many countries. Controlling virus such as TMV in tobacco plants is an important component to maximizing tobacco yield and quality.

Although cultivated tobacco is of great economic significance, relatively little information exists on its genome structure and organization. Tobacco is studied as a plant model for genome structure and function because of its well-characterized amphidiploids genetic background [2]. Characterization of different expressed sequences tags (ESTs) provides information of agronomic importance on the tobacco genome. Because an EST is usually unique to a particular cDNA, and because cDNAs correspond to a particular gene in the genome [3], ESTs have been used in this study to help identify unknown genes, especially those related to TMV, and to map their position in the genome in tobacco plants.

ESTs sequencing is a rapid and cost-effective way to begin gene discovery in species where the coding regions comprise only a fraction of the whole genome. Partial cDNA sequencing to generate ESTs is being used at present for the fast and efficient obtainment of a detailed profile of genes expressed in various tissues, cell types, or developmental stages [3]. Large numbers of ESTs have now been generated from a variety of model organisms. In plants, substantial collections of ESTs are available for Arabidopsis and rice, in each case representing significant proportions of the estimated total numbers of genes [4]. Opperman et al. [5] constructed cDNA libraries from N. tabacum and N. benthamiana for EST sequencing and to date they have sequenced ~11,000 ESTs from each library; in the database of the Institute of Genomic Research (TIGR) there are over 9000 tobacco ESTs (released in June, 2004; http://www. tigr.org/tigr-scripts/tgi/T_index.cgi?species=tobacco) and the ultimate goal of the TIGR Gene Index projects is to represent a non-redundant view of all tobacco genes and data on their expression patterns, cellular roles, functions, and evolutionary relationships.

In this research, one cDNA library population that include leaves, stem, root and roots base of four varieties of tobacco plants "Nicotiana tabacum" have been generated to develop a databases of 5927 ESTs. The expression profiles of these genes were compared to discuss the possible roles of TBT012G08 and TBT087G01 in TMV disease resistance. The enrichment provided by cDNA cloning in combination with the efficiency of cDNA array was an effective approach used in this study towards identifying developmentally regulated genes in tobacco.

2. MATERIALS AND METHODS

2.1. PLANT MATERIALS

Four varieties of tobacco seeds from Yunnan province (Hongda, NC82, YH05 and YUN201) were sow on fertilizer soil until plants grew about 10 cm. Leaves, stems, roots and roots base were used for library construction.

2.2. CDNA LIBRARY SUBTRACTION

Leaves, stems, roots and roots base that were powdered in a mortar with liquid N_2 using TRIZOL® (Gibco-BRL) reagent with the Vendor's protocol (GIBCO-BRL) were used for total RNA extraction. Isolation of Poly A + mRNA from total RNA was done according to the standardized manufacturer's protocol (Oligotex mRNA Spin-Column). Double-stranded cDNAs were synthesized with the Universal RiboClone® cDNA Synthesis System (Promega).

2.3. TEMPLATE PREPARATION

cDNA libraries were planted on 1.5% agar solid LB medium (pH 7.0) with 100 μg/mL Ampicillin and inoculated in 37ºC overnight.

Randomly selected white bacterial colonies picked with toothpicks were immersed in each well of a 96-Square-Well block with 2YT that was prepared at pH 7.0, contained the appropriate selective agent. Cultures were allowed to grow overnight by shaking plates at 200 rpm at 37°C and then they were storage at –70ºC freezer, next plasmid DNA was prepared by a modified alkaline lysis protocol using plasmid DNA purification Kits (Millipore). The results of isolation were analyzed by agarose gel electrophoresis, and then they were used for sequence. Sequence reactions were run and analyzed from 3' end of each cDNA on MegaBACE^{TM} 1000 automated sequencer (MD) using DYEnamic ET Dye Terminator Cycle Sequencing Kit (Amersham Pharmacia biotech). The primer was standard M13 forward primer: 5'-CCCAGTCACGACG-TTGTAAAACG-3'.

2.4. ESTS DATA ANALYSIS

Classification of clones was based on their similarity to Genebank data-base entries. BLASTX and BLASTN programs [6] were used to search for similarity to proteins and nucleic acids, respectively. Default parameters were used for both programs. EST sequences from both pools also were compared with the non-redundant and the dbEST databases. In general, similarities with probabilities $<10^{-4}$ were considered significant. Gene sequences from both cDNA populations have been deposited in the following Genebank EST data-base:

(http://www.estarray.org/san/xest/index.asp?sp=tbt)

2.5. EST CLUSTERING

Every EST sequence subjected to BLAST searching was used to construct a BLAST database. The EST sequences were then considered as BLASTN queries against this database. ESTs that produced an alignment with another EST with an overlap of /100 bases with/95% identity were linked with that EST to produce a cluster.

2.6. PREPARATION OF CDNA-ARRAY

The unique cDNA clones were randomly collected and amplified by means of PCR in a 96-well format on the basis of primer pairs specific for the vector ends (M13 forward, 5'-cccagtcacgacgttgtaaaacg-3', and M13 reverse, 5'-agcggataacaatttcacacagg-3'), 100 μL of PCR mixture containing 10 μL of 10 × PCR buffer, 7 μL of 25 mM $MgCl_2$, 1 μL of 20 mM dNTP, 1 μL of 100 ng/μL M13 forward and M13 reverse respectively, 3 units of Tag DNA polymerase (Promega) and approximately 10 ng of plasmid template. The reactions were run on multi-block system (MBS) thermocycler (ThermoHybaid, Ashford, UK) using an amplification program of 5 min denaturation at 94ºC, then followed by 36 cycles of 30 s at 94ºC, 60 s at 58ºC, and 90 s at 72ºC, and terminated by a 10 min extension at 72ºC. The PCR products were precipitated by adding 100 μL of lsopropylalcohol and 10 μL of sodium acetate (3 M, pH 5.2) and centrifuging at 3 500 × g at 4ºC for 35 min. The DNA precipitate was washed with 70% (w/v) ethanol and centrifuged at 3500 × g for 20 min again. When dried, the precipitate was suspended in 15 μL of denatured solution (0.4 M NaOH and 10 mM EDTA), waiting for printing. Array was produced by using GeneTAC TM G3 arrayed (Genomic Solution), those spots (approximately 150 μm in diameter) printed on the filter (Hybrid N^+) were aligned in 24 ×16 large array with each 12 × 2 cDNA spot array. Two duplications of DNA spot were arrayed adjacently to verify the reproducibility of experiments.

2.7. PREPARATION OF LABELED CDNA

To generate probes, PCR products were labeled with radioactive dCT33P using random primed labeling. Each reaction (50 μL) consisted of 50 ng of total RNA, 5 μg of oligo(dT) 16 primers, 300 pmol of random hexamer, 30 mM each of dATP, dTTP, dGTP and 5 μL of α-33P-dCTP (10 mCi/mL, Amersham Biosciences), 10 μL of 5× reaction buffer provided with reverse transcriptase (Life Technologies), 0.5 mM DTT and 400 unit superscript II reverse transcriptase (Life Technologies). Total RNA and primers were heated to 70ºC for 5 min and quick cooled on ice before the remain reaction continue for 1 h at 37ºC pre-incubation followed by 2 h at 42ºC, and then 50 μL TEN

TABLE 1. ANALYSIS OF N. TOBACCO ESTS

EST SUMMARY	
Total ESTs	5927
Mean EST length(nt)	351
Number of TUTs	3600
Number of contigs	521
Redundancy(%)a	48.1
Number of singletons	3079
Cotigs sizes	
2-5 ESTs	426
6-10 ESTs	46
11-15 ESTs	20
16-20 ESTs	11
21-30 ESTs	9
≥31 ESTs	9
G+C content (%)	43.6

[a]Redundancy = ESTs assembled in contigs/total ESTs

TABLE 2. BLASTX RESULTS OF TUTS.

SEQUENCE MATCHESA	NUMBER OF TUTS	PERCENTAGE (%)
Significant matches High	696	19.33
Moderate	300	8.33
Weak	154	4.27
Not significant matches	2450	68.05
All the TUTs	3600	100

[a] The TUT sequence matches were divided into highly significant($E \leq 10^{-20}$), moderately significant($10^{-20} < E \leq 10^{-10}$, weakly significant($10^{-10} < E \leq 10^{-5}$ and not significant ($E > 10^{-5}$) classes according to the expect (E) values of BLASTX hit sequences.

buffer (40 mM Tris-Cl pH 7.5, 1 mM EDTA pH 8.0, 150 mM NaCl) were added. Each of the labeled cDNA products were purified using G-50 column (Amersham biosciences), the purified labeled cDNA products were incubated at 70ºC for 5 min and then cooled on ice before hybridization.

2.8. ARRAY HYBRIDIZATION AND DATA ANALYSIS

Hybridization was performed overnight at 60ºC in hybridization oven. Washing was done at 60ºC in 2× SSC and 0.1% SDS for 20 min and twice in 0.1× SSC and 0.1% SDS for 20 min. The signals of the array were absorbed by storage phosphor screen for 2 days, and the signal intensities were measured by scanning the storage phosphor screen with the Typhoon 9200 (Molecular Dynamics).

Spot finding and analysis of signal intensity were quantified based on software ArrayVision 6.0 (Amersham Pharmacia Biotech). Normalization was performed after adjusting the total signal intensity of the image (global normalization). To insure the results, three individual hybridization experiments were performed for each probe, and linear correlation analysis was used in pair-wise comparisons which correlation coefficients (R2) were between 0.89~0.94.

Inferring changes of every gene between two tissues were used *t*-test. We selected 0.01 as confidence level, and when t exceeded this threshold, the two populations were considered to be different.

Most of the sequences (>95%) had been submitted to the NCBI tobacco dbEST. Related resource and data analysis by UniGene on NCBI are also available at: http://www.estarray.org/san/xest/ index.asp?sp=tbt.

2.9. QUANTITATIVE REAL-TIME PCR

Real-time quantitative RT-PCR with gene-specific primers was used to confirm some of the results obtained with cDNA array. Contaminating genomic DNA in total RNA was removed by treatment with RNase-free DNase (Promega). Then RNA samples was reversed transcribed with the oligo(dT) 16 primer using SuperScriptTM II RNase H-Reverse Transcriptase (Invitrogen). Genes of interest were detected by using primers that had been optimized to generate a single amplicon of ~80 bp nucleotides. Real-time PCR was performed by using an iCycler (Bio-Rad) and iQTM SYBR

TABLE 3. SUMMARY OF THE EST ANALYSIS FROM N. TABACUM CDNA LIBRARY.

FUNCTIONAL CATEGORYA	NUMBER OF TUTS	NUMBER OF ESTS	REDUNDANCY FACTOR b	PERCENTAGE IN ANNOTATION ESTS (%)
Energy	196	831	4.24	32.32
Metabolism	126	572	4.54	22.25
Protein synthesis	219	369	1.68	14.35
Sub-cellular localization	137	261	1.90	10.15
Protein fate	83	136	1.64	5.27
Cellular transport and transport mechanism	63	113	1.79	4.39
Defence and virulence	34	106	3.12	4.12
Biogenesis of cellular components	40	85	2.12	3.31
Transcription	22	25	1.14	0.97
Protein with binding function or cofactor requirement (structural or catalytic)	12	24	2	0.93
Cell cycle and DNA processing	9	10	1.11	0.39
Cellular communication/signal transduction mechanism	5	7	1.4	0.27
Interaction with the cellular environment	4	7	1.75	0.27
Transport facilitation	4	5	1.25	0.19
Storage protein	2	2	1	0.08
Cell type differentiation	0	0	0	0
Development(Systemic)	0	0	0	0
Unclassified proteins	14	18	1.28	0.70
Total	970	2571	2.65	100.00

[a] Category were assigned according to MIPS Functional Catalogue Database by Gene Ontology annotation.
[b] Redundancy factor = number of ESTs/TUTs of a given category

Green Supermix. Reactions were performed in 96-well plates with optical sealing tape (Bio-Rad) and contained 25 μL total volumes. β-Actin primers were used in parallel for each run. The PCR amplification was carried out with an initial step at 93°C for 1 min followed by 50 cycles of 10 s at 93°C, 20 s at 60°C and 10 s at 72°C. The relative expression of the gene of interest was determined by calculating the ratio of the extrapolated concentration of that gene to the extrapolated concentration of β-actin.

3. RESULTS

3.1. EST SEQUENCING AND ASSEMBLY

A total of 5927 sequences were generated from the 3' ends of clones selected from the Nicotiana tabacum normalized cDNA library with a mean insert size of 1.2 kb. All ESTs with an average read length of 351 nt following vector and low-quality sequence (definition found in materials and methods) trimming were submitted to NCBI database. The library result for further analysis is summarizing in TABLE 1. From this dataset 521 contigs and 3079 singletons were formed after assembly of the 5927 ESTs.

Contig containing the biggest quantity of ESTs was CLS_1_contig_CL1Contig1 according to our summary standard (TABLE 4). The results showed 3600 TUTs including 3079 singletons (85.53% in TUTs) and 521contigs. The number of contigs with twenty or more ESTs presumably representing highly expressed genes was 9 (1.73%

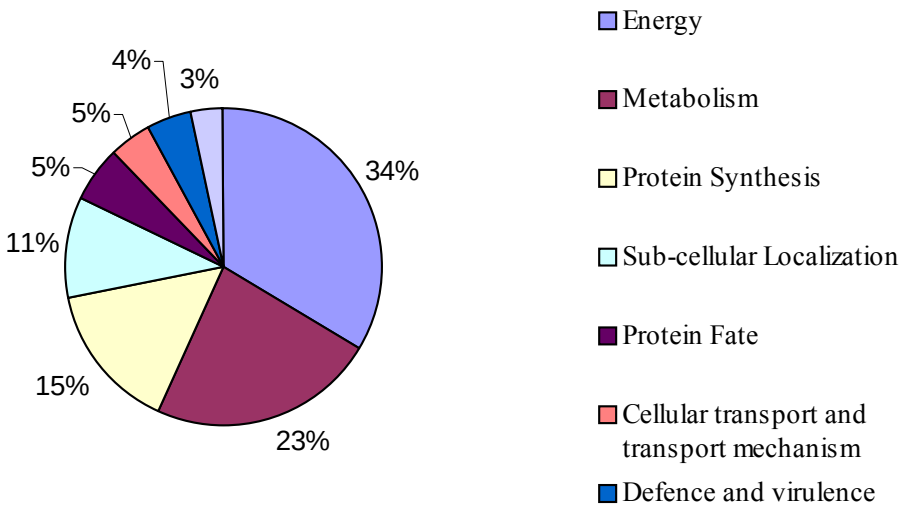

FIGURE 1. Distribution of functional categories among ESTs percent.

in contigs, 0.25% in TUTs respectively), and the remaining 512 contigs (98.28 % in contigs, 14.22% in TUTs respectively) were used to represent genes expressed moderately. G+C content was 43.6% in all assembled EST sequences, this content is important to know the nucleotide composition of the genes, also G+C represent the strongest sequence linkage.

A redundancy of 48.1% was found in this set of sequences (TABLE 1); these redundant sequences could be transcripts of the same gene or cognate genes. The redundancy inherent in cDNA libraries is beneficial in the generation of assemblies as overlapping ESTs from a single gene can be aligned and compiled to generate a consensus sequence. Comparison of multiple ESTs from a given assembly is useful for identifying the boundaries of open reading frames, alternative splicing, and strain-specific polymorphisms.

3.2. FUNCTIONAL ANNOTATION AND CLASSIFICATION

TABLE 2 shows 3600 non-redundant sequences that were examined based on BLASTX searches of the GenBank database. A number of 696 unique sequences (19.33% in TUTs) were identified highly significant matches ($E \leq 10^{-20}$) to the known gene sequences and was regarded as representation of known genes. A number of 2450 TUTs (68.05% in TUTs) had no significant matches in sequence similarity ($E > 10^{-5}$) according to result of BLASTX searches for available database.

Each sequence was searched against the non-redundant protein database of GenBank to obtain qualitative information about the ESTs. As a result, 696 ESTs (45.4%) encode putative amino acid sequences with high significant similarities to those registered in the NCBI.

A total of 2571 annotation ESTs were further given a putative functional assignment according to a MIPS (Munich Information center for Protein Sequences) classification system and the expression profile as revealed by EST analysis is summarize in TABLE 3. This annotation allows the classification of generated ESTs by function [7] with the aim to create universal vocabulary for consensus annotation [8].

Although several schemes of categorization may be equally valid, here proteins that share similar functional characteristics or cellular roles have been attempted to group together, rather than follow a strict biochemical classification. Roles were assigned according to the known or putative

TABLE 4. THE TUTS EXPRESSED PREDOMINANTLY AS MEASURED BY EST REDUNDANCY.

CONTIG	BLASTX ANNOTATIONA	ORGANISM	E VALUE	NUMBER OF ESTS
CLS_1_contig_CL1Contig1	Ribulose bisphosphate carboxylase small chain, chloroplast precursor (EC 4.1.1.39) (RuBisCO small subunit) (TSSU3-8)	Nicotiana sylvestris	0	246
CLS_32_contig_CL4Contig2	Plastocyanin, chloroplast precursor	Lycopersicon esculentum	0	77
CLS_33_contig_CL5Contig1	Translationally controlled tumor protein homologue (TCTP)	Nicotiana tabacum	0	70
CLS_2_contig_CL1Contig2	Ribulose bisphosphate carboxylase small chain, chloroplast precursor (EC 4.1.1.39) (RuBisCO small subunit) (TSSU3-8)	Nicotiana sylvestris	0	68
CLS_25_contig_CL3Contig1	Metallothionein-like protein type 2	Nicotiana glutinosa	2E-26	65
CLS_3_contig_CL1Contig3	Ribulose bisphosphate carboxylase small chain S41, chloroplast precursor (EC 4.1.1.39) (RuBisCO small subunit S41)	Nicotiana sylvestris	0	56
CLS_12_contig_CL2Contig8	Chlorophyll a-b binding protein 16, chloroplast precursor (LHCII type I CAB-16) (LHCP)	Nicotiana tabacum	0	48
CLS_34_contig_CL6Contig1	Photosystem II 10 kDa polypeptide, chloroplast precursor (PII10)	Nicotiana tabacum	0	37
CLS_43_contig_CL10Contig1	Basic endochitinase 2 precursor (EC 3.2.1.14)	Oryza sativa	1E-14	35
CLS_35_contig_CL6Contig2	Photosystem II 10 kDa polypeptide, chloroplast precursor (PII10)	Nicotiana tabacum	0	30
CLS_27_contig_CL3Contig3	Metallothionein-like protein type 2	Nicotiana glutinosa	1E-27	29
CLS_28_contig_CL3Contig4	Metallothionein-like protein type 2	Nicotiana glutinosa	4E-11	28
CLS_51_contig_CL14Contig1	Auxin-repressed 12.5 kDa protein	–	2E-26	25
CLS_66_contig_CL18Contig1	Oxygen-evolving enhancer protein 3, chloroplast precursor (OEE3) (16 kDa subunit of oxygen evolving system of photosystem II) (OEC 16 kDa subunit)	Spinacia oleraceae	0	23
CLS_38_contig_CL8Contig1	Photosystem II reaction center W protein, chloroplast precursor (PSII 6.1 kDa protein)	Spinacia oleraceae	1E-26	22
CLS_36_contig_CL7Contig1	Probable protease inhibitor P322 precursor	Solanum tuberosum	4E-15	21
CLS_42_contig_CL9Contig3	Ferredoxin I, chloroplast precursor	Lycopersicon esculentum	0	21
CLS_37_contig_CL7Contig2	Probable protease inhibitor P322 precursor	Solanum tuberosum	2E-13	20

[a] BLASTX annotation was assigned by Gene Ontology term according to results of the highest scoring BLASTX hit used to compare our consensus cluster sequences against UniProt database by the corresponding ESTs number more than 20, otherwise no BLASTX hit was list.

TABLE 5. PERCENTAGE OF N. TABACUM TUTS WITH HOMOLOGY TO SPECIES.

ORGANISM	NUMBER OF TUTS	PERCENTAGE (%)
Arabidopsis thaliana	249	6.92
Nicotiana tabacum	100	2.77
Lycopersicon esculentum	80	2.22
Solanum tuberosum	39	1.08
Nicotiana sylvestris	27	0.75
Other plants	308	8.55
Homo sapiens	26	0.72
Cyanobacteria	13	0.36
Other bacteria	13	0.36
Eukaryotes	6	0.16
Ascomycetes	12	0.33
Other organisms	17	0.47

involvement of a gene or a protein in a cellular process or pathway, as opposed to its participation in a specific binding or catalytic functions.

All annotation ESTs were classified in 18 functional categories. Unique transcripts involved in energy were largest group accounting for 831 (32.32%) in all annotation ESTs. The following category including more highly expressed genes were metabolism 572 ESTs (22.25%) where similar representation of ESTs from the cell metabolisms category was reported for other normalized cDNA libraries[9], protein synthesis 369 ESTs (14.35%), sub-cellular localization 261 ESTs (10.15%), protein fate 136 ESTs (5.27%), cellular transport and transport mechanism 113 ESTs (4.39%), defence and virulence 106 ESTs (4.12%), and biogenesis of cellular components 85 ESTs (3.31%) in turn (FIG. 1). All this result shows that the normalization step that took place in the construction of the cDNA libraries was efficient in diminishing the level of highly abundant transcripts equally represented in the different analyzed tissues.

The percentage in other classified function groups was less than 1%. The redundancy factor in metabolism was highest (4.54) compare with other categories and we calculate the average level of redundancy factor in all annotation TUTs for 2.65.

The most abundant contig (CLS_1_contig_CL1Contig1) including 246 ESTs (4.15% in ESTs) was found highly homologous to hydrophobin encoded by Ribulose bisphosphate carboxylase small chain, chloroplast precursor (EC 4.1.1.39) (RuBisCO small subunit) (TSSU3-8) gene of Nicotiana sylvestris . The second most abundant unique transcripts (1.3% in ESTs) was CLS_32_contig_CL4Contig2 encoding Plastocyanin, chloroplast precursor from Lycopersicon esculentum. There were high abundant genes homologous to Ribulose bisphosphate carboxylase small chain, chloroplast precursor (EC4.1.1.39) (RuBisCO small subunit) (TSSU3-8) from Nicotiana sylvestris, this included 68 and 56 ESTs (1.15, 0.94% in ESTs). The homologues of translationally controlled tumor protein homologue (TCTP), chlorophyll a-b binding protein 16, chloroplast precursor (LHCII type I CAB-16) (LHCP) and photosystem II 10 kDa polypeptide, chloroplast precursor (PII10) from Nicotiana tabacum could be found expressed abundantly in our library.

3.4. HOMOLOGOUS COMPARISON TO OTHER ORGANISM

After excluding 2450 non-significant TUTs in sequence similarity, 100 unique sequences (only 1.67% in total TUTs) were identified from N. tabacum database and the remaining 1050 TUTs showed partial homologous to the genes of the other organisms (TABLE 5). A number of 249 TUTs (6.92% in total TUTs) had most significant TUTs in sequence similarity to Arabidopsis thaliana. The second most match species excluding N. tabacum was Lycopersicon esculentum with 80 TUTs (2.22%) and only few significant match number with Nematodes and algae with 5 TUTs (0.14%) were found according to BLASTX match with available database.

TABLE 6. HIGH QUALITY ESTS ANALYSIS FROM ARRAY RESULTS.

VARIETIES	EST #	ANNOTATION$	E-VALUE
Hongda - NC82	TBT001C07	Pollen specific protein C13 precursor	3.00E-18
	TBT046G07	No annotation	0
	TBT106C01	Histone H4	8.00E-36
Hongda-YH05	TBT011B01	No annotation	0
	TBT108H08	ATP synthase 6 kDa subunit, mitochondrial	0
	TBT112A12	Metallothionein-like protein type 2	4.00E-11
Hongda - YUN201	TBT012G08	Basic form of pathogenesis-related protein 1 precursor (PRP 1)	0
	TBT087G01	Ubiquitin	1.00E-14
	TBT018C04	No annotation	0
NC82- YH05	TBT036E01	Ferredoxin I, chloroplast precursor	6.00E-17
	TBT070D01	Chlorophyll a-b binding protein CP29.2, chloroplast precursor	3.00E-25
	TBT115D05	No annotation	0
NC82- YUN201	TBT012G08	Basic form of pathogenesis-related protein 1 precursor	0
	TBT065F12	No annotation	0
	TBT070E11	Proteasome subunit alpha type 3	4.00E-26
YH05- YUN201	TBT050H06	Elongation factor 2 (EF-2)	4.00E-41
	TBT055D02	No annotation	0
	TBT084B03	Glycine-rich RNA-binding protein GRP1A	4.00E-11

[#] Accession number for cDNA used in this study deposited in our GenBank.
[$] Annotation of clones obtained by universal primer was subjected to homology search against the NCBI database.

3.5. ANALYSIS OF CDNA ARRAY DATA AND IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

To identify genes among of 5289 clones (http://www.estarray.org/san/xest/index.asp?sp=tbt) the amplified cDNAs were arrayed on nylon filters and were hybridized with ^{32}P-labeled mRNA probes from separate RNA isolations of four tobacco varieties (Hongda, NC82, Yun 201 and YH05).

After comparing the hybridization data 359 high quality ESTs were generated and analyzed using the Cyber-T statistic program (TABLE 6). This program was used to analyze the data set and examined gene expression levels and *t*-test was a filter method where two classes of gene expression datasets were compared, in this case four varieties between each other were compared, the chosen genes tend to be strongly expressed in one class and weakly in the other class. This regularized *t*-test approach is available for online use over the Internet at http:/www.igb.uci.edu. The identified genes were significantly differentially expressed if

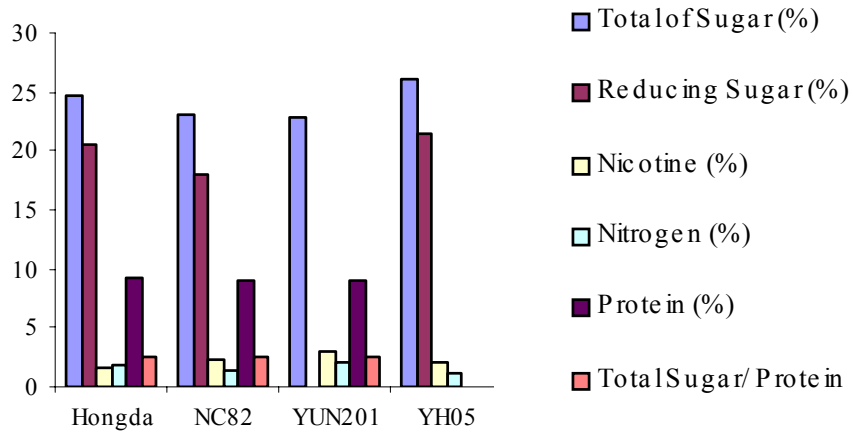

FIGURE 2. VARIETIES CHARACTERISTICS.

they had a confidence level $P < 0.01$. The average R2 was 0.93-0.95 between replication among the arrays so DNA array data at various time points provided a preliminary assessment of gene activity. These data served as a basis for subsequent experiments at a relevant time point.

In the array result two genes strongly related to tobacco mosaic virus (TMV) were obtained, basic form of pathogenesis-related protein 1 precursor (TBT012G08) and ubiquitin (TBT087G01) both of them were found in the variety Hongda. This variety has been reported as susceptible to TMV infection.

As we all know within the plant sciences, studies of TMV have contributed significantly to the understanding of plant-microbe interaction in general and in particular to our current knowledge of induced resistance phenomena.

TMV is a virus of angiosperm plants and found worldwide. It has been isolated from many species but most commonly from cultivated tobacco (N. tabacum). The virus is contagious and has no specific vector. It is considered to have a host range of intermediate size [10], being able to infect around 199 plant species in 30 families [11]. This host range size is presumably typical of most tobamoviruses.

The history of TMV illustrates how pragmatic strategies to control an economically important disease of tobacco have had unexpected and transforming effects across platforms that impinge on plant health and public health.

3.5.1. PROBES CHARACTERIZATION

Four varieties of N. tabacum were used as probes for the array. The characteristics of these varieties could be seeing in FIG. 2. The varieties Hongda and YH05 have a high content of total sugar (24.68% and 26.04%), reducing sugar (20. 65% and 21. 52%) and the low one of nicotine (1.67% and 2.06%) respectively. Maybe the sugar and nicotine content in these varieties, specially in Hongda, may have a relation with TMV resistant or susceptibility in tobacco plants.

Plants contain thousands of secondary metabolites which are not involved in the primary metabolic pathways. However, secondary metabolites have very specific functions. In tobacco plants nicotine and related alkaloids are important secondary metabolites which are produced in high concentrations in order to protect the plant from insect attack [12,13].

3.6. QUANTITATIVE REAL-TIME PCR ANALYSIS TO VALIDATE ARRAY RESULTS

Among the 359 genes on the arrays, several of them were selected for further analysis because of the potential physiological significance of changes

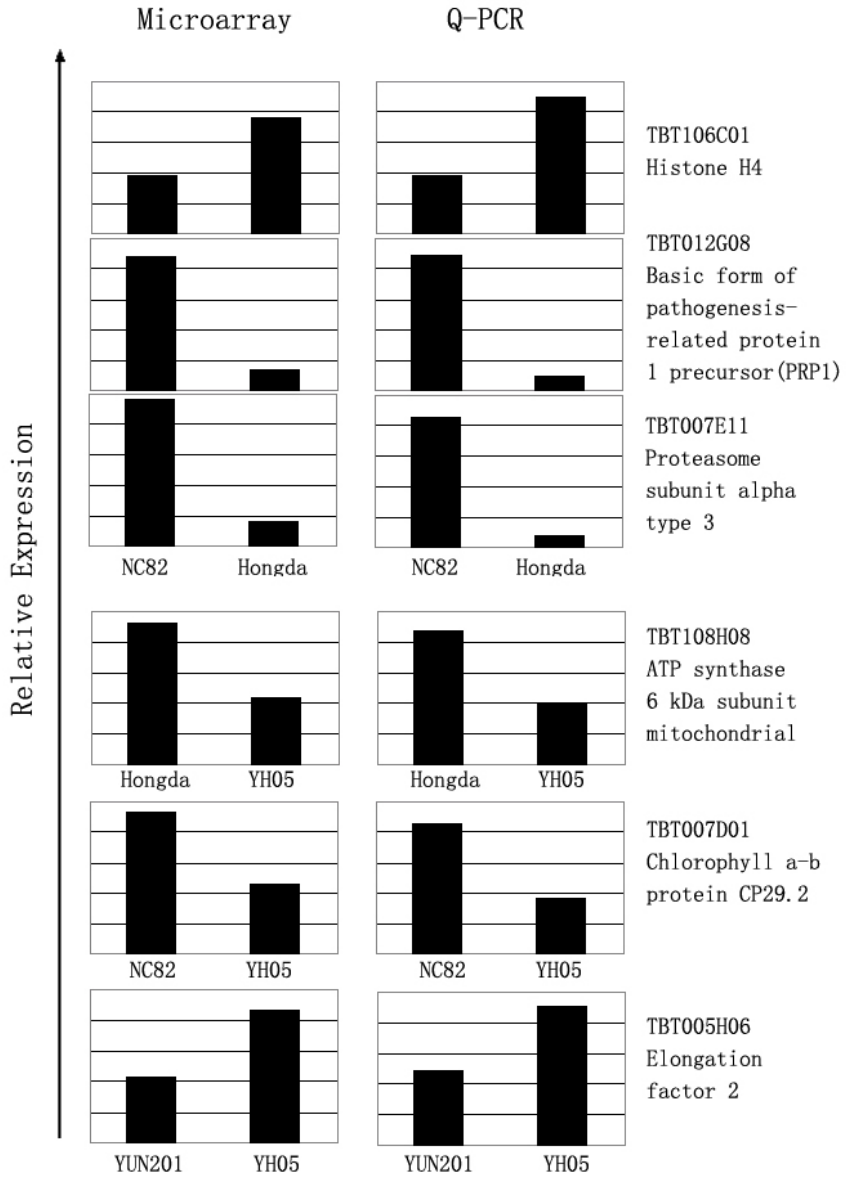

FIGURE 3. RT-PCR FOR ARRAY VALIDATION RESULT.
Left: Global normalized expression values from microarray experiments.
Right: Relative quantification of genes performed by using real-time detection of PCR amplification.

(or lack of changes) that were detected (FIG. 3). Total RNA was prepared and subsequently reverse transcribed using oligo(dT) primers. The cDNAs were then used as template for the PCRs for each set of primers. The sequences of the specific primers used for each set of reactions, as well as the size of each PCR product, were described in Materials and Methods. All parameters of the reaction, such as template concentration and number of cycles, were determined for each set of primers so that the reaction was within the linear range of the PCR [14]. Control used in this experiment was β-actin. It

has previously been shown that β-actin is a house-keeping gene and therefore was used to confirm that the cDNAs used for each reaction were of the same quantity.

4. DISCUSSION

4.1. EST SEQUENCING AND ASSEMBLY

This EST analysis provides a first picture of the numerous tobacco genes potentially involved in tobacco quality. Obviously, a large scale EST analysis is needed to elucidate gene expression patterns during tobacco plants development. It is clear that various factors can lead to misrepresentation of transcript abundance, in particular potential biases in cDNA cloning, efficiency of in vivo excision, or colony viability [15].

To gain insight into the tobacco transcriptoma, a generation of 5927 EST sequences from tobacco were report in this study. Because the inherent nature of EST sequencing results in the redundant sequencing of identical transcripts, the redundancy in the EST dataset was reduced by creating a gene index. Through a series of clustering and assembly processes, the approximately 5927 ESTs were reduced into 3600 unique sequences. In this research also some transcripts unique to single cDNA libraries that represent differentially expressed genes and sequences broadly expressed in all sampled libraries that represent "housekeeping" genes were identified.

The sequences reported through this study provide a significant improvement in the general understanding of tobacco plants because the ESTs generated in this research were derived from biologically and agronomically relevant tissues. Aerial tissue was represented by leaf: top, medium and lower leaves and one steam. Below-ground tissue was represented by root and root base libraries.

4.2. ESTS FUNCTIONAL ANNOTATION AND HIGHLY EXPRESSED GENES

The abundance of ESTs with predicted cellular roles in tobacco were categorized with respect to a MIPS classification system into 18 broad categories of biological roles (see TABLE 3). Although several schemes of categorization may be equally valid, the proteins that share similar functional characteristics or cellular roles were attempted to group together, rather than follow a strict biochemical classification. Roles were assigned according to the known or putative involvement of a gene or a protein in a cellular process or pathway, as opposed to its participation in a specific binding or catalytic function. Genes involved in energy (32.32%) constitute the most abundant class among ESTs. Genes related to metabolism (22.25%) and protein synthesis (14.35%) are also highly represented among ESTs (FIG. 1).

The comparison of the distributions of the functional categories between the ESTs reveals the cellular roles represented by the highly expressed genes (see TABLE 3). This is most obvious for the energy category, which includes genes encoding proteins involved in photosynthesis like Ribulose bisphosphate carboxylase Rubisco (4.15% of total ESTs) and Plastocyanin, chloroplast precursor gene (1.30% of total ESTs). This category of genes, along with those for protein synthesis are sampled more often by random sequencing of cDNA libraries which reflects their general house-keeping role [16].

EST programs from photosynthetic organisms normally result in the identification of numerous sequences encoding the small subunit of Rubisco, the principal enzyme in carbon fixation [17]. Rubisco catalyze two competing reactions that involve the carboxylation and oxygenation of ribulose-P2, and initiating the primary steps of photosynthetic C reduction and photorespiration [18]. RUBISCO is likely to be the world's must abundant protein comprising approximately 50 % of the soluble protein in green leaves [19].

Mazur and Chui [20] have cloned two RUBISCO SS genes from tobacco. These genes are not linked to each other, nor are they closely linked to other tobacco SS genes. Sequencing of one of these genes confirmed that it is closely related to the SS genes from other species, particularly those

from dicots. The tobacco SS transit sequence, however, is much less conserved relative to the SS transit sequences of other species. The tobacco promoter and transit sequences may therefore be the DNA of choice to use in gene fusions in which foreign genes are expressed in tobacco, and in fusions in which foreign proteins are directed to the chloroplast.

Many chemical reactions of metabolism are coupled oxidation-reductions that utilize a chain of electron transport molecules that specialize in oxidizing and reducing at specific energy levels to minimize energy loss in energy transfers. In the process of photosynthesis, the electron transport carriers are embedded in the thylakoid membranes. The electrons oxidized from chlorophyll are the source of initial reduction in the transfer chain. Molecules that specialize in energy transfers in the thylakoids are: NADP, plastoquinone, two cytochromes, plastocyanin and ferredoxin [21].

Plastocyanin is a soluble electron transfer protein found in most photosynthetic organisms. This blue copper protein, which functions as the oxidant of cytochrome f and the reductant of the P700 reaction center, has been the subject of studies ranging from kinetics [21] to molecular biology [22].

A total of 48 ESTs encoded Chlorophyll a-b binding proteins that constitute the antenna system of the photosynthetic apparatus [23]. Chlorophyll is the pigment involved in harvesting the light energy for photosynthesis and it is found in the chloroplasts of green plants. This process is the basis of the plant's ability to accumulate carbohydrates. There are actually 2 types of chlorophyll, named a and b. Both of these two chlorophylls are very effective photoreceptors because they contain a network of alternating single and double bonds, and the orbital can delocalise stabilizing the structure. Such delocalised polyenes have very strong absorption bands in the visible regions of the spectrum, allowing the plant to absorb the energy from sunlight [24]. Genes involved with Chlorophyll a-b binding protein represent 0.81% of a total ESTs.

Chlorophyll is one of the pigments involved in the light reactions, plastocyanin participate in the cyclic electron flow of the light reaction and Rubisco is the protein involved in the Calvin cycle.

4.3. DIFFERENTIALLY EXPRESSED GENES AFTER ARRAY ANALYSIS

In the last decade, the cloning of a number of plant genes that confer resistance to various pathogens has brought significant progress in our understanding of host–pathogen interactions. This array analysis shows that there are two genes TBT012G08 and TBT087G01 with relation to TMV in the Hongda variety. TBT012G08 is a basic form of pathogenesis-related protein 1 precursor and TBT087G01 represent ubiquitin.

TMV has had an illustrious history for more than 100 years, dating to Beijerinck's description of the mosaic disease of tobacco as a contagium vivum fluidum [25] and the modern usage of the word "virus." Since then, TMV has been acknowledged as a preferred didactic model and a symbolic model to illuminate the essential features that define a virus. TMV additionally emerged as a prototypic model to investigate the biology of host plants (tobacco). TMV also exemplifies how a model system furthers novel, and often unexpected, developments in biology and virology. Today, TMV is used as a tool to study host-pathogen interactions and cellular trafficking, and as a technology to express valuable pharmaceutical proteins in tobacco.

Pathogenesis-related protein was first discovered in tobacco after TMV infection, hence called PR proteins [26]. These proteins are induced in several plant species when they are infected with viruses, viroids, fungi or bacteria. The occurrence of these proteins is not pathogen-specific, but is determined by the type of reaction of the host plant. They form a protective barrier against pathogens by collecting at infection sites and act to decrease susceptibility of the plants. Although they are implicated in plant defense, they have not been identified because of their anti-pathogenic actions, but solely because of their accumulation in infected plants.

This inducible, extracellular protein has been detected in many plants species but has been characterized best in tobacco (N. tabacum). Tobacco PR proteins can be grouped into at least five functionally, biochemically, serologically, and genetically distinct groups or families called PR1 through PR5 [27]. The acidic PR proteins (16000 M) of tobacco were the subject of much of the earlier work in this area and continue to be intensively studied, but the function(s) of the PR1 proteins remain unknown. However, there is an association between the synthesis and/or TMV in tobacco [28].

Now we can ask: Are the pathogenesis-related proteins involved in the resistance of tobacco plants to TMV infection?

Numerous studies in N. tabacum have emphasized the relationship between the synthesis of PR1 proteins and the restriction of TMV replication during the hypersensitive response, as well as, the correlation between the systemic induction of the PR1 genes and the level of systemic acquired resistance to secondary infections by TMV. However, viral resistance in the absence of PR1 proteins synthesis has been reported in some instances. In addition, TMV infection induces PR genes believed to be involved in resistance to bacterial and fungal infections, and the synthesis of the PR1 proteins can be induced by nonviral pathogens [27].

Powell et al. [29] used the PRb transgenic plant to test the hypothesis that the PR1 proteins of tobacco are unique components of the defence mechanism directed against TMV infection. Constitutive expression of the PR1b gene caused no apparent delay in onset or alteration of systemic disease symptoms in Xanthi transformants inoculated with 1.0 μg/mL TMV. Lowering the virus inoculum to 0.1 μg/ mL also proved unsuccessful in demonstrating a low level viral resistance in contrast to that observed in tobacco plants engineered to express TMV coat protein.

Ubiquitin is a highly conserved 76-amino-acid eukaryotic protein that is covalently attached to proteins as monomers or lysine-linked chains of polyubiquitin [30]. Conjugation with ubiquitin can mark a protein down for destruction, particularly if the target protein is polyubiquitinated. However, monoubiquitination is associated with regulation of protein activity or the compartmentalization of proteins in animals and plants [30,31]. Speculatively, ubiqutination plays some role in either the in vivo assembly of virions or in their localization in the infected cell. However, the importance, if any, of H-protein remains unknown [11]. TMV was the first virus that was found to be ubiquitinated, although subsequently others have been found [32,33].

Becker et al.[34], found that a clone NF36 from TMV-infected Xanthi NN encodes tobacco ubiquitin and its expression in the TMV based vector likely triggers a necrotic response by a co-suppression mechanism. When tobacco plants are compromised in their ubiquitin systems due to the co-suppression effect of an ubiquitin transgene, spontaneous necrotic lesions appear. In addition, such plants have heightened resistance to TMV infection and display and altered pattern of PR gene expression.

After all these analysis is possible to said that the genes TBT012G08 and TBT087G01 have a high relation to TMV and further investigation may be required to demonstrated more clearly if these genes could be used by tobacco plants to defence itself against TMV infection.

4.4. USES FOR THE TOBACCO ESTS

The EST data from the N. tabacum cDNA library described here can initially be used to generate probes to isolate genomic DNA containing the corresponding genes and to provide markers for physical maps. Gene-expression studies may identify genes with cell-type-specific or symbiotically regulated expression patterns. Once isolated from genomic DNA, the promoters of such genes may provide valuable reagents for transgenic promoter-fusion experiments. Other genes described here may be useful as controls for constitutive expression.

Finally, the EST database may be of use to scientists who have biochemical purified proteins of interest from N. tabacum. The partial peptide

sequence of a purified protein could be compared against translated EST sequences. If present, related and more extensive cDNAs could then be readily identified and used as tools for additional studies [35].

Like N. tabacum, ESTs of many other plant species are being sequenced and annotated for use by plant researchers. Such plants as beet, soybean, cotton, wheat, and sorghum, among others, all have EST collections. Sequencing ESTs provide information about what genes are expressed at any given time in a given organ for individual plants, making them a very useful resource. Because EST collections provide information about which genes are turned on and in which tissues, they complement genomic sequencing efforts.

CONCLUSION

The array analyses realized through this research provide a picture of some genes potentially involved in plant defense and plant resistance against TMV. In this context, it is encouraging to note that from 359 high quality ESTs generated after comparing the hybridization data two genes TBT012G08 and TBT087G01 are related to TMV. This study, like others, shows that it is fruitful to characterize a number of different cDNA libraries, and thus the coordinated international effort on TMV is providing the community with valuable data for gene identification, protein prediction and DNA array production.

Study of genes related with plant defense and plant resistance against TMV is an important element in the outcome of a plant-pathogen interaction. Isolation of elicitor-inducible genes, characterization of activation profiles and identification of their putative functions make these genes good candidates for understanding early and late events in plant disease resistance responses.

REFERENCES

[1] AKEHURST BC [1968] Tobacco. London. pp. 2-11.

[2] GRIMMER M, MAUNDERS M AND EDWARDS K [2000] Genome analysis using expressed sequences markers. Plant and Animal Genomes VIII Conference. January 9-12, 2002, San Diego, California.

[3] ADAMS MD, KELLEY JM, GOCAYNE JD, DUBNICK M, POLYMERO-POULOS, MH, XIAO HC, MERRIL R, WU A, OLDE B, MORENO RF, KERLAVAGE AR, MCCOMBIE WR AND VENTER JC [1991] Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651-1656.

[4] EWING R, POIROT O AND CLAVERIE JM [1999] Comparative analysis of the Arabidopsis and rice Expressed Sequence Tag (EST) sets. Silico Biol 1: 197-213.

[5] OPPERMAN CH, LOMMEL SA, SOSINSKI B, BURKE M, LAKEY N, HE L, BRIERLEY R, SALSTEAD A, GADANI F AND HAYES A [2003] The tobacco genome initiative Plant & Animal Genomes XI Conference. January 11-15, 2003, Town & Country Convention Center, San Diego, CA.

[6] ALTSCHUL SF, MADDEN TL, SCHAFFER AA, ZHANG J, ZHANG Z, MILLER W AND LIPMAN DJ [1997] A new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

[7] OUZOUNIS CA, COULSON RMR, ENRIGHT AJ, KUNIN V AND PEREIRA-LEAL JB [2003] Classification schemes for protein structure and function. Nature Rev 4: 509-519.

[8] CAMON E, BARRELL D, BROOKSBANK C, MAGRANE M AND APWEILER R [2003] The Gene Ontology Annotation (GOA) project-application of GO in Swiss-Prot, TrEMBL and InterPro. Comp and Funct Genom 4: 71-74.

[9] ALI S, HOLLOWAY B AND TAYLOR WC [2002] Normalization of cereal endosperm EST libraries for structural and functional genomic analysis. Plant Mol Biol Rep 18: 123-132.

[10] SCHNEIDER WL AND ROOSSINCK MJ [2000] Evolutionarily Related Sindbis-Like Plant Viruses Maintain Different Levels of Population Diversity in a Common Host. J Virol 74: 3130-3134.

[11] ZAITLIN M [1999] Elucidation of the genome organization of tobacco mosaic virus. Philos Trans R Soc Lond B Biol Sci 29; 354: 587-91.

[12] WINK M [1988] Plant Breeding: Importance of plant secondary metabolites for protection against pathogens and herbivores. Theor Appl Genet 75: 225-233.

[13] WINK M AND ROBERTS MF [1998] Biochemistry, Ecology and Medicinal Application. New York: Plenum Press.

[14] FREEMAN WM, WALKER SJ AND VRANA KE [1999] Quantitative RT-PCR: Pitfalls and potential. BioTechniques 26: 112-125.

[15] JOURNET EP, VAN TUINEN D, GOUZY J, CRESPEU H, CARREAU V, FARMER MJ, NIEBEL A, SCHIEX T, JAILLON O, CHATAGNIER O, GODIARD L, MICHELI F, KAHN D,

GIANINAZZI-PEARSON V AND GAMAS P [2002] Exploring root symbiotic programs in the model legume Medicago truncatula using EST analysis. Nucl Acids Res 30: 5579-5592.

[16] SAWBRIDGE T, ONG E, BINNION CH, EMMERLING M, MEATH K, NUNAN K, O'NEILL M, O'TOOLE F, SIMMONDS J, WEARNE K, WINKWORTH A AND SPANGENBERG G [2003] Generation and analysis of expressed sequence tags in white clover (Trifolium repens L). Plant Sci 165: 1077-1087.

[17] HÖFTE H, DESPREZ T, AMSELM J, CHIAPELLO H, CABOCHE M, MOISAN A, JOURION MF, CHARPENTEAU JL, BERTHOMIEU P AND GUERRIER D [1993] An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA clones from Arabidopsis thaliana. Plant J 4: 1051-1061.

[18] READ BA AND TABITA FR [1992] A hybrid ribulose bisphosphate carboxylase/oxygenase enzyme exhibiting a substantial increase in substrate specificity factor. Biochemistry 31: 5553-5560.

[19] KING SD [1976] Science 191: 429-434.

[20] MAZUR BJ AND CHUI CF [1985] Sequence of a genomic DNA clone for the small subunit of ribulose bis-phosphate carboxylase-oxygenase from tobacco. Nucl Acids Res 13: 2374-2386.

[21] FARVER O AND PECHT I [1988] Preferred sites and pathways for electron transfer in blue copper proteins. Prog Clin Biol Res 274: 269-283.

[22] SMEEKENS S, BAUERLE C, HAGEMAN J, KEEGSTRA K AND WEISBEEK P [1986] The role of the transit peptide in the routing of precursors toward different chloroplast compartments. Cell 46: 365-375.

[23] JANSSON S [1999] A guide to the Lhc genes and their relatives in Arabidopsis. Trends Plant Sci 4: 236-240.

[24] STREITWEISER AND HEATHCOCK [1981] Introduction to Organic Chemistry. New York: MacMillan Press.

[25] BEIJERINCK MW [1898] Concerning a contagium vivum fluidum as cause of the spot disease of tobacco leaves. Verhandelingen der Koninklyke akademie van Wettenschappen te Amsterdam 65: 3-21

[26] VAN LOON LC AND VAN KAMMEN A [1970] Polyacrylamide disc electrophoresis of the soluble leaf proteins from Nicotiana tabacum var "Samsun" and "Samsun NN" II Changes in protein constitution after infection with tobacco mosaic virus. Virology 40: 199-211.

[27] CARR JP AND KLESSING DF [1989] The pathogenesis-related protein of plants. In: Genetic Engineering, Principles and Methods. Vol. 11. Setlow J (Editor), Plenum: New York. pp. 65-100.

[28] VAN LOON LC [1985] Pathogenesis-related proteins. Plant Mol Biol 4: 111-116.

[29] POWELL AP, NELSON RN, HOFFMANN N, ROGERS SG, FRALEY RT AND BEACHY RG [1986] Delay of disease development in transgenic plant that express the tobacco mosaic virus coat protein gene. Science 232: 738-743.

[30] WEISSMANN AM [2001] Themes and variations on ubiquitylation. Nature Rev Mol Cell Biol 2:169-178.

[31] BACHMIR A, NOVATCHKOVA M, POTUSHAK T AND EISENHABER F [2001] Ubiquitylation in plants: A post-genic look at post- translational modification. Trends Plant Sci 6: 463-470.

[32] DUNIGAN DD, DIETZGEN RG, SCHOELZ JS AND ZAITLIN M [1988] Tobacco mosaic virus particles contain ubiquitinated coat protein subunits. Virology 165: 310-312.

[33] HAZELWOOD D AND ZAITLIN M [1990] Ubiquitinated conjugates are found in preparations of several plant viruses. Virology 177: 352-356.

[34] BECKER F, BUSCHFELD E, SCELL J AND BACHMAIR A [1993] Altered response to viral infection by tobacco plants perturbed in ubiquitin system. Plant J 3:875-881.

[35] COVITZ AP, SMITH SL AND LONG RH [1998] Expressed Sequence Tags from a Root-Hair-Enriched Medicago truncatula cDNA Library. Plant Physiol 117: 1325-1332.

RECEIVED ON 5-8-2005.
ACCEPTED ON 6-18-2005.