Electronic Journal of Biotechnology ISSN: 0717-3458
© 2005 by Pontificia Universidad Católica de Valparaíso -- Chile

Vol. 8 No. 2, Issue of August 15, 2005
Received December 9, 2004 / Accepted April 27, 2005

RESEARCH ARTICLE

A legume genomics resource: The Chickpea Root Expressed Sequence Tag Database

Jayashree B.
Bioinformatics and Computational Biology Unit
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel: 91 40 23296161
Fax: 91 40 23241239
E-mail: b.jayashree@cgiar.org

Hutokshi K. Buhariwalla
MS Swaminathan Applied Genomics Lab
International Crops Research Institute for the Semi- Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel: 91 40 23296161
Fax: 91 40 23241239
E-mail: h.k.buhariwalla@cgiar.org

Sanjeev Shinde
Bioinformatics and Computational Biology Unit
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel: 91 40 23296161
Fax: 91 40 23241239
E-mail: s.shinde@cgiar.org

Jonathan H. Crouch *
M.S. Swaminathan Applied Genomics Lab
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Patancheru, Andhra Pradesh 502 324, India
Tel:91 40 23296161
Fax:91 40 23241239
E-mail: j.crouch@cgiar.org

Financial support: Earmarked unrestricted grants from the governments of UK, Japan, European Union and the Generation Challenge Program.

Keywords: cloning, data mining, drought avoidance, drought tolerance, EST database, root traits, stress.

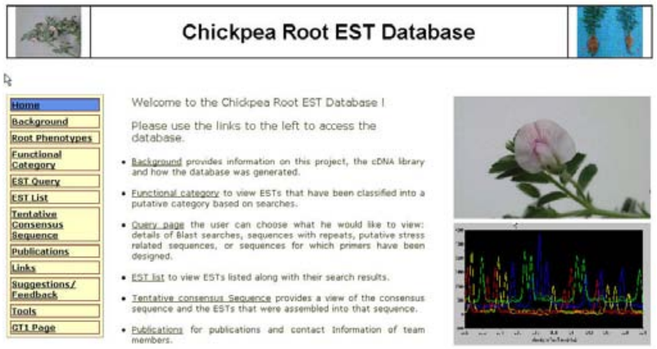
Abbreviations: ASP: active server page;
BLAST: Basic Local Alignment Search Tool;
dbEST ID: database for expressed sequence tags- Identity;
EST: expressed sequence tag;
GO: Genome ontology;
NCBI: National Centre for Biotechnology Information;
NSH: no significant homology;
RPS-BLAST: Reversed Position Specific BLAST;
SSH: subtractive suppressive hybridization;
SSR: simple sequence repeats;
TIGR: The Institute for Genomic Research.

Chickpea, a lesser-studied grain legume, is being investigated due to its taxonomic proximity with the model legume genome *Medicago truncatula* and its ability to endure and grow in relatively low soil water contents making it a model legume crop for the study of agronomic response to drought stress. Public databases currently contain very few sequences from chickpea associated with expression in root tissues. However, root traits are likely to be one of the most important components of drought tolerance in chickpea. Thus, we have generated a set of over 2800 chickpea expressed sequence tags (ESTs) from a library constructed after subtractive suppressive hybridization (SSH) of root tissue from two closely related chickpea genotypes possessing different sources of drought avoidance and tolerance (ICC4958 and Annigeri respectively). This database provides researchers in legume genomics with a major new resource for data mining associated with

* Corresponding author

root traits and drought tolerance. This report describes the development and utilization of the database and provides the tools we have developed to facilitate the bioinformatics pipeline used for analysis of the ESTs in this database. We also discuss applications that have already been achieved using this resource.

Chickpea (*Cicer arietinum* L) is one of the world's most important but lesser-studied leguminous food crops with nearly 10 M ha grown across the Americas, the Mediterranean basin, East Africa, the Middle East, Asia and Australia (FAOSTAT, 2004). While in the developed world it represents a valuable crop for export, in the developing world it provides a protein-rich supplement to cereal-based diets. Chickpea is generally grown without irrigation, planted in the post-rainy season, surviving through to harvest on progressively declining residual soil moisture. Thus, chickpea grows during the time of the year when many other legumes are rarely cropped, displaying considerable drought avoidance and/or tolerance. Efforts to identify genes underlying drought tolerance are mostly focused on model species and major cereal crops such as rice and maize (Bruce et al. 2002; Nguyen et al. 2004) with lesser attention being given to legumes even though they are known to possess high levels of drought tolerance (Turner et al. 2001).



Welcome to the Chickpea Root EST Database !

Please use the links to the left to access the database.

- Background provides information on this project, the cDNA library and how the database was generated.
- Functional category to view ESTs that have been classified into a putative category based on searches.
- Query page the user can choose what he would like to view: details of Blast searches, sequences with repeats, putative stress related sequences, or sequences for which primers have been designed.
- EST list to view ESTs listed along with their search results.
- Tentative consensus Sequence provides a view of the consensus sequence and the ESTs that were assembled into that sequence.
- Publications for publications and contact Information of team members.

Figure 1. The Chickpea Root EST Database main menu page. Accessible through URL: http://www.icrisat.org/gt1/cpest/home.asp. All subsequent pages provide links to viewing ESTs through the toolbars on the left of the screen.

Root traits are likely to be one of the most important components of drought tolerance in chickpea. However, little attention has been given to defining the molecular genetic determinants associated with drought avoidance root traits or indeed any other more general mechanisms of drought tolerance in chickpea. The merits of generating ESTs from chickpea roots are many, especially as the root system is a primary sensor of drought stress (Davies and Zhang, 1991). Moreover, chickpea is taxonomically one of the closest crops to *Medicago* and both are well adapted to dry environments. Thus, in an effort to study the mechanisms of drought tolerance in chickpea, ESTs were generated from a subtractive suppressive hybridization (SSH) library using the local landrace accession ICC 4958 and a local variety Annigeri, which are both considered as different sources of drought avoidance and tolerance (Saxena, 1993; Saxena, 2003). The unique ESTs post-clustering are now available with Genbank (CK148643-CK149150). All the ESTs can also be accessed from a relational database: http://www.icrisat.org/gt1/cpest/home.asp. This database is an important resource for chickpea genomics scientists and molecular breeders, and can provide a platform for studies on legume genomics and drought tolerance by the wider community.

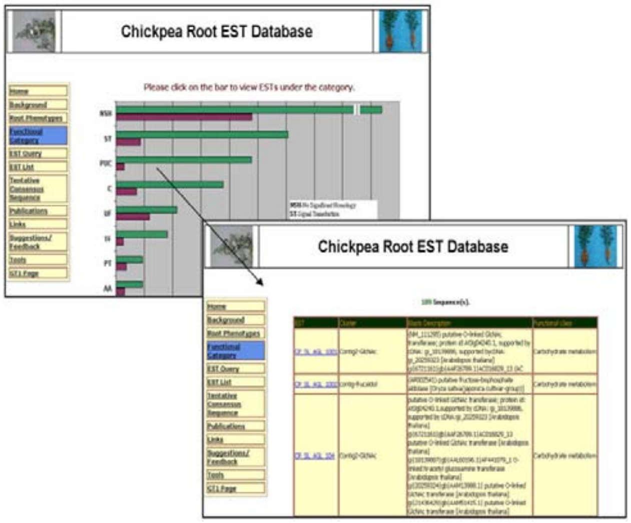

Chickpea Root EST Database
Please click on the bar to view ESTs under the category.
Home
Background
Root Phenotypes
Functional Category
EST Query
EST List
Tentative Consensus Sequence
Publications
Links
Suggestions/ Feedback
Tools
GT1 Page
NSH
ST
PMC
C
UF
TF
PT
AA
Chickpea Root EST Database

Figure 2. EST annotation data pages of the Chickpea Root EST Database. Summary of EST annotations according to their functional classification, clicking on a bar in the histogram will display a window with a list of all sequences that have been classified under that category, the individual sequences and associated data may also be retrieved directly.

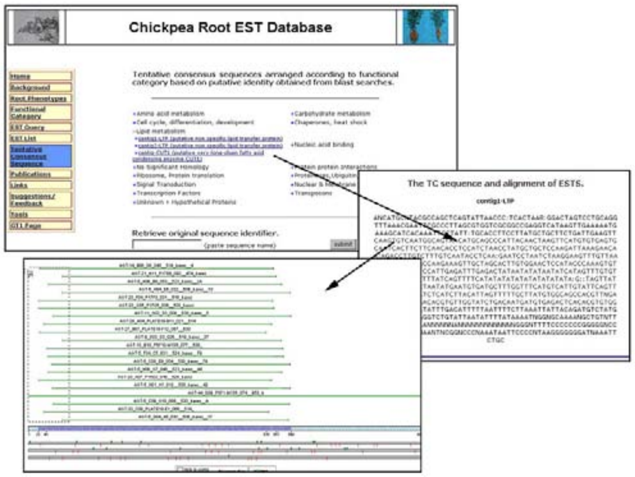
MATERIALS AND METHODS

Development of the EST library

Plants used for RNA extraction were grown under non-stressed conditions as the target of this study was the identification of genes conferring prolific root system development, which is known to be expressed in the target genotype under both stress and unstressed conditions (Saxena, 1993). The subtraction of the two genomes was achieved through a process of PCR-based suppression of genes common to both genotypes, and genes that are differentially expressed are enriched. Being PCR based, low copy number cDNA can be detected from the tester (ICC 4958). Full details of the methodologies used have been reported elsewhere (Buhariwalla et al. 2005).

Sequencing and annotation of ESTs

A total of 4000 reads were generated, only sequences of more than 170 bp with less than 5% ambiguity (2858 sequences, 71.5%) were processed. The average length of the ESTs was 492bp. All root ESTs were subjected to a processing pipeline consisting of base calling with the ABI DNA Sequence Analysis Software (v 3.7) or by Chromas v2.2 (Technelysium Pty Ltd. Australia) followed by vector screening using Sequencher v.4 (Gene codes, Ann arbor, MI). Quality was visually verified through the trace files. Processed ESTs were screened against multiple plant DNA databases (*A. thaliana*, *M. truncatula*, *G. max*, *Z. mays* and *O. sativa* in the TIGR-Unique Gene Indices, that represent clustered assemblies of EST sequences and dbEST) using BLASTn and tBlastx (Altschul et al. 1997). Sequences were clustered into contigs using Sequencher v.4, the 2858 ESTs analysed clustered into 210 contigs and 267 singletons. The consensus sequence from contigs was searched for conserved domains using RPS-BLAST. Amongst the 210 contigs, 77% (162 of 210) found homology, while 67% (178 of 267) of singletons found homology.



Tentative consensus sequences arranged according to functional category based on putative identity obtained from blast searches.

- Amino acid metabolism
- Cell cycle, differentiation, development
- Lipid metabolism
  - contig1-LTP (putative non specific lipid transfer protein)
  - contig2-LTP (putative non specific lipid transfer protein)
  - contig-CUT1 (putative very long-chain fatty acid condensing enzyme CUT1)
- No Significant Homology
- Ribosome, Protein translation
- Signal Transduction
- Transcription Factors
- Unknown + Hypothetical Proteins
- Carbohydrate metabolism
- Chaperones, heat shock
- Nucleic acid binding
- Protein protein interactions
- Proteases, Ubiquitin
- Nuclear & Membrane
- Transposons

Retrieve original sequence identifier.

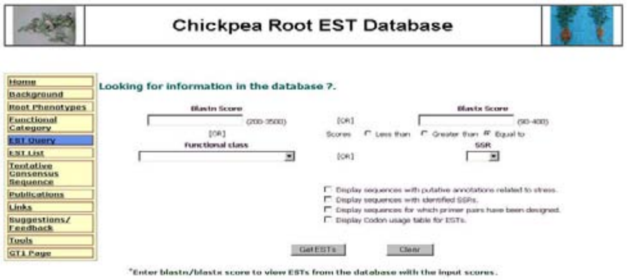
(paste sequence name)

The TC sequence and alignment of ESTS.

Figure 3. EST clustering information of the Chickpea Root EST Database. Description of tentative consensus sequences, selecting a contig from a particular functional class will display a window showing the consensus sequence and the alignments of individual ESTs within a contig.

Development of tools for EST analysis

Tools were developed to parse output files from NCBI and TIGR BLAST searches and the parsed results were processed into MS-Access data-tables. The BLAST output analyzer program reads BLASTn, RPS-BLAST, and tBLASTx output saved as text files from NCBI and TIGR blast searches. The program collects statistics relating to the top scoring hit alignment, which includes the Blast description, score, e-value, percentage, identity, length, strands, species of origin etc. The program creates tables to store extracted data in a MS-Access database. The database structure is included in the tool, which is Windows compatible. While parser tools such as Btab are available in the public domain, this tool differs in its ability to handle output files from varieties of BLAST searches, effects group processing of files, and outputs results directly into MS-Access tables, which the user can read, analyse and import to other sheets or databases.



Looking for information in the database ?.

Blastn Score □ (200-3500) [OR] Blastx Score □ (90-400)

[OR] Scores ○ Less than ○ Greater than ● Equal to

Functional class [OR] SSR

□ Display sequences with putative annotations related to stress.
□ Display sequences with identified SSRs.
□ Display sequences for which primer pairs have been designed.
□ Display Codon usage table for ESTs.

*Enter blastn/blastx score to view ESTs from the database with the input scores.

Figure 4. The query page from Chickpea Root EST Database. The database can be queried through a range of user-defined variables.

RESULTS

ICC 4958 has shown relatively high productivity under various drought stress field conditions and has been used as a source of drought tolerance by chickpea breeders across India and Australia. It has been shown that the prolific root system of ICC 4958 (expressed under both optimum and drought conditions) is the single most important mechanism for drought avoidance. ICC 4958 has a 30% higher root biomass than Annigeri (Saxena, 1993). Since it is known that ICC 4958 and Annigeri are very closely related, a subtractive approach was expected to identify only a moderate number of differentially expressed transcripts. Constitutive drought tolerance mechanisms (such as root system development) are a common target for the crop physiology community and are increasingly identified by genomics studies as important components of stress tolerance mechanisms.

A range of bioinformatic analyses was carried out on the ESTs generated from this SSH library. These included putative annotation of the sequences, grouping them into functional categories, clustering, searching for repeats and conserved domains, and primer design. To make these ESTs available as a user-friendly resource for the whole community, a structured database was considered essential (Figure 1). Thus, all the information available on the chickpea variety and the sequence data and bioinformatic analysis have been made available through the Chickpea Root EST Database in this first release. All of the data is available through web-interfaces; the database allows data mining through a set of query pages and users may also search their sequences against the EST database using the Blast option. The database architecture includes SQL-Server (v.7) database server at the backend, the Windows

2000 IIS web server and web pages have been written in ASP.

Database content

The database stores individual nucleic acid sequences, the results of the Blast analysis against all databases searched, the manual annotation as well as clustering information and a pictorial view of the EST-alignment within a cluster. The database also contains information on identified microsatellites, ESTs annotated for stress, sequences to which primers were designed and results on amplification from the laboratory. Putative functional annotations for the ESTs were obtained from the most significant match obtained from database searches while ESTs with no significant database match were labeled NSH (no significant homology) (Buhariwalla et al. 2005). The putative annotations were manually grouped under 12 general categories based on similarity of biochemical function and the categories used for the annotation of *Arabidopsis*.

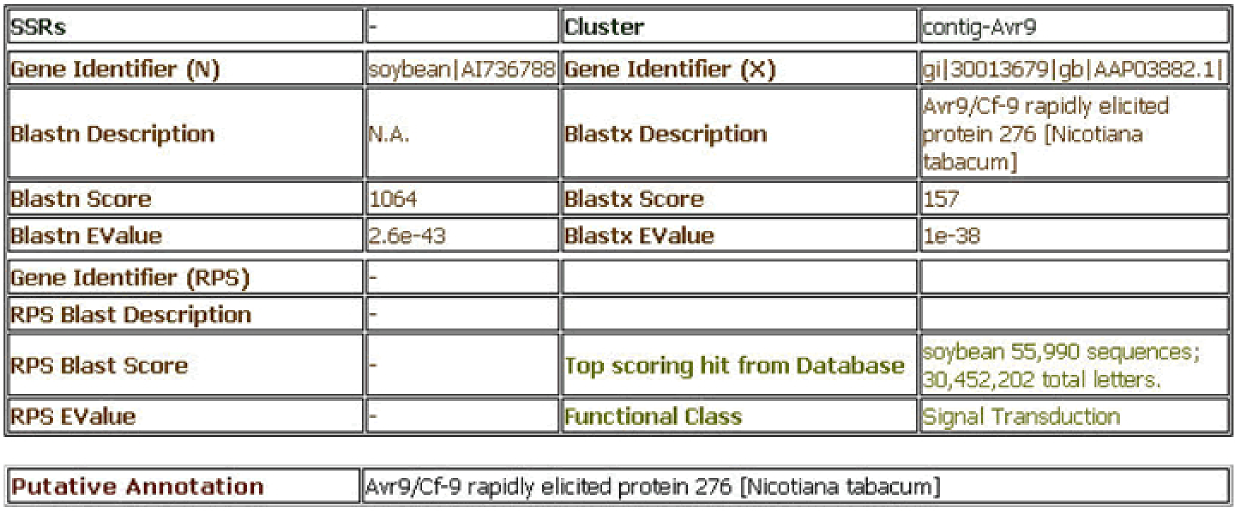
Access to annotated data

The functional class may be accessed in several ways: a) through the web-page containing the “Functional category” histogram: where the user may click on an individual functional class bar to access all the ESTs with their putative function (Figure 2); (b) through the “Tentative-consensus sequence” page, where clicking on the functional category displays the clusters of ESTs and their consensus sequences that have been classified as belonging to that functional class (Figure 3); (c) through the “query” page, where the user can view all ESTs after selecting a functional category (Figure 4). The query pages enable users to develop personalized queries through a combination of filters using Boolean options and/or selectors (less than, equal to, greater than). The user can structure a query to identify chickpea sequence homology reports with user defined cut-off values during the selection of query parameters. The database can be searched through database-specific identifications, Genbank accession numbers or dbEST ID. The query pages also permit listing and retrieval of sequences containing SSRs or sequences to which primer pairs have been designed. The user may also retrieve sequences that have been classified as putative stress transcripts and also obtain a codon usage table for chickpea. The query result output display is in a tabular format (Figure 5).

The ESTs were searched for microsatellite repeat motifs using the Tandem Repeat Finder program (Benson, 1999). For these searches, di, tri, tetra or penta repeats were only considered as potential simple sequence repeat (SSR) markers if they had an overall motif length of more than 12 bp. About 650 ESTs (50 TCs and singletons) contained SSRs. Most of them were dinucleotide repeats (TC and AG). Primers were designed for all unique ESTs containing SSRs (putative EST-SSRs) and also for some ESTs with promising putative annotations associated with stress response. Results from the laboratory indicating those EST-SSRs and unigene-ESTs that amplified in chickpea are also available along with the number of alleles and product size (Buhariwalla et al. 2005).

Tools available

The Blast output analyzer tool is available for download, including the database structure. The parsing program also has a database query facility. The program runs in manual or automatic modes (Figure 6). Parsed data can be retrieved based on nucleotide/protein scores, description, sequence identity and the query results may be saved to a new or existing table.



SSRs	-	Cluster	contig-Avr9
Gene Identifier (N)	soybean|AI736788	Gene Identifier (X)	gi|30013679|gb|AAP03882.1|
Blastn Description	N.A.	Blastx Description	Avr9/Cf-9 rapidly elicited protein 276 [Nicotiana tabacum]
Blastn Score	1064	Blastx Score	157
Blastn EValue	2.6e-43	Blastx EValue	1e-38
Gene Identifier (RPS)	-		
RPS Blast Description	-		
RPS Blast Score	-	Top scoring hit from Database	soybean 55,990 sequences; 30,452,202 total letters.
RPS EValue	-	Functional Class	Signal Transduction

Putative Annotation	Avr9/Cf-9 rapidly elicited protein 276 [Nicotiana tabacum]

Figure 5. Query result output display from the Chickpea Root EST Database. The table contains clustering information and statistics related to the top Blast hit alignment that led to the manual annotation.

Overview of EST Annotations

This database provides the first insight into the genes that may be associated with root development and abiotic stress tolerance in chickpea. We have isolated EST with putative relationships to signal transduction and transcriptional factors, proteases and lipid transfer proteins, transporters and heat-shock proteins, and, an enzyme involved in the synthesis of trehalose which is a known membrane protectant associated with dessication tolerance in resurrection plants. The EST sequences reported here have a range of putative relationships to genes thought to be involved in drought stress, as described in detail elsewhere (Buhariwalla et al. 2005). These annotated ESTs provide a useful resource as candidates for functional genomics and candidate gene mapping of drought tolerance, and for allele mining of drought avoidance and tolerance in cool season legumes such as chickpea.

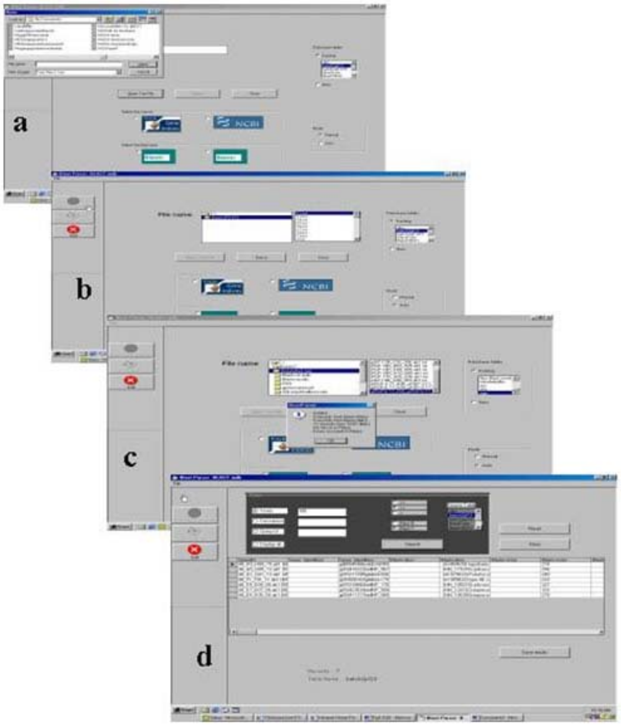


Figure 6. The Blast output analyzer tool from the Chickpea Root EST Database. The program runs in two modes:

(a) Manual or

(b) Automatic.

(c) The message box displays the number of records appended to the database and information on sequences that returned no hits in the homology searches.

(d) The query interface allows the user to retrieve parsed data and provides filtering options such as selecting records based on top scoring hit, user-defined scores, blast description etc.

Overview of EST marker development and use

Over two hundred EST markers have already been developed from these sequences representing diverse functional/stress annotations, a selection of which contained SSR motifs. In addition, primers were also designed from a random selection of sequences categorized as having ‘no significant homology' (NSH) and ‘unknown function' (UF) when compared with either nucleotide or protein sequences in public databases. Nearly 50 of these have been used for diversity analysis of chickpea germplasm representing 9 annual *Cicer* species as described in detail elsewhere (Buhariwalla et al. 2005). Gene-based markers have proven effective for diversity analysis in *Cicer* and may be useful in identifying promising candidates for interspecific hybridization programs. The levels of polymorphism detected by these markers suggest that they would be promising candidates for allele mining of germplasm collections for new sources of drought tolerance. The availability of gene-based markers together with highly polymorphic flanking SSR markers will also greatly assist in reducing linkage drag and increasing the speed and efficiency of subsequent introgression programs.

DISCUSSION

The Chickpea Root EST Database is a relational database system designed to provide on-line data mining of ICRISAT's chickpea root EST dataset. Users may sort through ESTs and contigs derived from them using pre-structured queries with user defined variables. The EST database developed in this study provides a profile of the expressed genes that have been sampled from chickpea roots associated with putative annotations related to stress tolerance and root development. The user can view the complete BLAST report that was used in the annotation of each sequence. Many candidate genes have been identified from chickpea in this study based on sequence similarity to known genes in model systems. In addition, many SSRs have been identified; mostly with dinucleotide repeat motifs (TC and AG). These will be of particular use for molecular breeding and diversity analysis applications in chickpea. We anticipate that the database and its associated web pages would provide a platform for data mining of these chickpea transcripts and that will facilitate comparative studies across other legumes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the programming support extended by Mr. P. Vinod Kumar and Ms. R. Rajalakshmi.

REFERENCES

ALTSCHUL, Stephen F.; MADDEN, Thomas L.; SCHAFFER, Alejandro A.; ZHANG, Jinghui; ZHANG, Zheng; MILLER, Webb and LIPMAN, David J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, September 1997, vol. 25, no.17, p.* 3389-3402.

BENSON, Gary. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research,* 1999, vol. 27, no. 2, p. 573-580.

BUHARIWALLA, Hutokshi K.; JAYASHREE, B. and CROUCH, Jonathan H. ESTs from chickpea roots with putative roles in drought tolerance. *Bio Med Central Plant Biology*, 2005. In press.

BRUCE, Wesley B.; EDMEADES, Gregory O. and BARKER, Thomas C. Molecular and physiological approaches to maize improvement for drought tolerance. *Journal of Experimental Botany*, January 2002, vol. 53, no 366, p 13-25.

DAVIES, William J. and ZHANG, Jianhua. Root signals and the regulation of growth and development of plants in drying soil. *Annual Review of Plant Physiology and Plant Molecular Biology*, June 1991, vol. 42, p. 55-76.

Food and Agriculture Organization of the United Nations (FAOSTAT), [cited 2004], 2004. Available at http://faostat.fao.org/.

NGUYEN, T.T.; KLUEVA, N.; CHAMARECK, V.; AARTI, A.; MAGPANTAY, G.; MILLENA, A.C.; PATHAN, M.S. and NGUYEN, H.T. Saturation mapping of QTL regions and identification of putative candidate genes for drought tolerance in rice. *Molecular Genetics and Genomics*, August 2004, vol. 272, no. 1, p. 35-46.

SAXENA, Narendra P. Management of drought in chickpea: a holistic approach. In: *Management of Agricultural drought – Agronomic and Genetic Options*. New Delhi, Oxford and IBH Publishers, 2003, Chapter 7, p. 103-122. ISBN 81-204-1529-9.

SAXENA, Narendra P.; KRISHNAMURTHY, Lakshmanan and JOHANSEN, Chris. Registration of a drought resistant chickpea germplasm. *Crop Science*, November 1993, vol. 33, p. 1424-1426.

TURNER, Neil C.; WRIGHT, Graeme C. and SIDDIQUE, K.H.M. Adaptation of grain legumes (pulses) to water-limited environments. *Advances in Agronomy*, 2001, vol. 71, p. 193-231.