BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 8(1)

The capsicum transcriptome DB: a "hot" tool for genomic research

Elsa Góngora-Castillo[1], Rubén Fajardo-Jaime[3], Araceli Fernández-Cortes[2], Alba E. Jofre-Garfias[1], Edmundo Lozoya-Gloria[1], Octavio Martínez[2], Neftalí Ochoa-Alejo[1], Rafael Rivera-Bustamante[1]*

[1]Departamento de Ingeniería Genética de Plantas, Centro de Investigación y de Estudios Avanzados del I.P.N (Cinvestav)-Unidad Irapuato, Km 9.6 Libramiento Norte, Carretera Irapuato-León, 36821-Irapuato, Gto., México; [2] Laboratorio Nacional de Genómica para la Biodiversidad (Langebio), Cinvestav-Irapuato, Km 9.6 Libramiento Norte, carretera Irapuato-León, 36821-Irapuato, Gto., México; [3] División de Ciencias Económicas y Administrativas, Universidad de Guanajuato, Fracc. El Establo. Campus Guanajuato. 36250. Guanajuato, Gto., México; Rafael Rivera-Bustamante – Email: rrivera@ira.cinvestav.mx; *Corresponding author

Received November 14, 2011; Accepted December 17, 2011; Published January 06, 2012

Abstract:

Chili pepper (*Capsicum annuum*) is an economically important crop with no available public genome sequence. We describe a genomic resource to facilitate *Capsicum annuum* research. A collection of Expressed Sequence Tags (ESTs) derived from five *C. annuum* organs (root, stem, leaf, flower and fruit) were sequenced using the Sanger method and multiple leaf transcriptomes were deeply sampled using with GS-pyrosequencing. A hybrid assembly of 1,324,516 raw reads yielded 32,314 high quality contigs as validated by coverage and identity analysis with existing pepper sequences. Overall, 75.5% of the contigs had significant sequence similarity to entries in nucleic acid and protein databases; 23% of the sequences have not been previously reported for *C. annuum* and expand sequence resources for this species. A MySQL database and a user-friendly Web interface were constructed with search-tools that permit queries of the ESTs including sequence, functional annotation, Gene Ontology classification, metabolic pathways, and assembly information. The Capsicum Transcriptome DB is free available from http://www.bioingenios.ira.cinvestav.mx:81/Joomla/

Background:

Chili pepper (*Capsicum annuum*, L.) constitutes one of the most important crops in Mexico. Mexico is the second largest chili pepper producer in the word and it has been suggested as a center of domestication of this species **[1],** which is reflected in the large number of pepper types found in the country. Sequence and analysis of Expressed Sequence Tags (ESTs) are primary tools for the discovery of novel genes in plants and other organisms. As the chili pepper genome is not currently available, transcriptome data can provide major insights into the genes and gene families involved in important biological processes. In this report, we present the Capsicum Transcriptome DB (Database), a web-based EST database. We constructed cDNA libraries derived from different organs of chili pepper plants including root, stem, leaf, flower and fruit. In some tissues, samples were collected after exposure to a variety of stress agents or during several developmental stages. The leaf transcriptome was deeply sequenced with both pyrosequencing and Sanger technologies, while cDNAs from the remaining tissues were sequenced solely with the Sanger platform. Sequences were assembled using a hybrid approach resulting in a reference transcriptome **(Figure 1**) of 32,314 contigs and 59,991 singletons. The Capsicum Transcriptome DB integrates comprehensive information including functional annotation, Gene Ontology (GO) **[2],** and metabolic pathway assignments **[3].** To provide public access to the sequence and annotation data, we developed and implemented a user-friendly, SQL query-builder tool into the Capsicum Transcriptome DB web site publicly available at http://www.bioingenios.ira.cinvestav.mx:81/Joomla/.

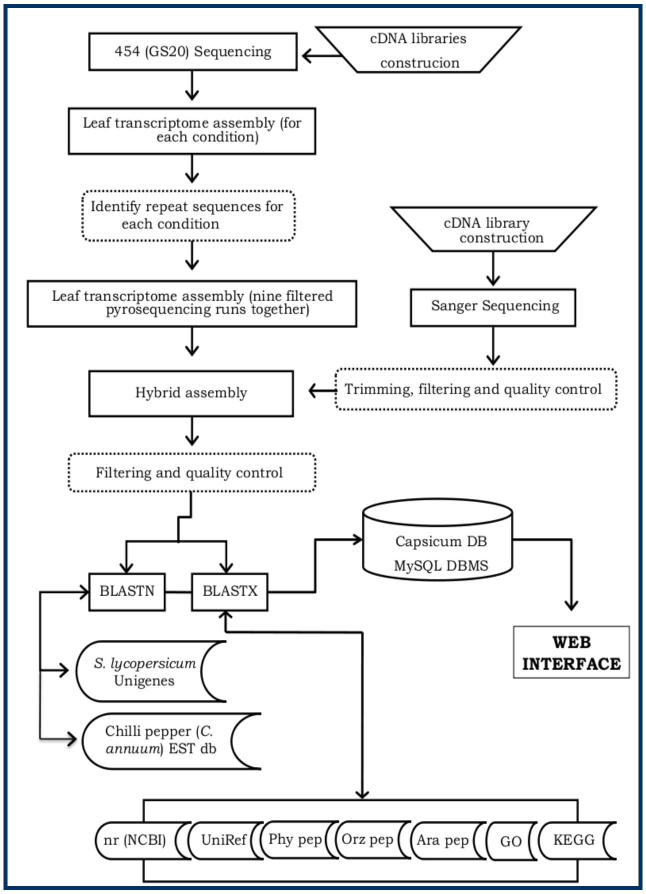

454 (GS20) Sequencing
cDNA libraries construcion
Leaf transcriptome assembly (for each condition)
Identify repeat sequences for each condition
cDNA library construction
Leaf transcriptome assembly (nine filtered pyrosequencing runs together)
Sanger Sequencing
Hybrid assembly
Trimming, filtering and quality control
Filtering and quality control
Capsicum DB MySQL DBMS
BLASTN
BLASTX
WEB INTERFACE
S. lycopersicum Unigenes
Chilli pepper (C. annuum) EST db
nr (NCBI)
UniRef
Phy pep
Orz pep
Ara pep
GO
KEGG

Figure 1: Schematic of the hybrid assembly process and database construction. The diagram represents the process followed to obtain the hybrid assembly and to construct the Capsicum Transcriptome database. The process included a sequential comparison with several genomic sequences databases (NCIB nr, UniRef100, *P. patents* peptides **[10],** *O. sativa* peptides **[9],** *A. thaliana* peptides **[8],** GO [2], KEGG **[3])** to annotate the contigs.

Methodology and Results:

Data Source

The Capsicum Transcriptome DB was established using sequence data from two *C. annuum* varieties, Serrano Tampiqueño 74 and Sonora Anaheim. For the Sanger-based data, we generated 83,116 ESTs from 30 cDNA libraries derived from root, stem, leaf, flower and fruit tissues **Table 1 (see supplementary material).** mRNA was isolated using Poly(A)+ capture and cloned using Gateway technology. After sequencing, Phred and custom Perl scripts were used to extract high-quality regions from the raw sequences, trim vector sequences, adaptors, polyA/T regions and eliminate short ESTs (< 90 bp) **Table 1 (see supplementary material).** A total of 70,743 high-quality Sanger sequences with an average length of 678.35 nucleotides (nt) were obtained. Three cDNA libraries constructed from DNA virus-infected and non-infected pepper leaf tissue (*C. annuum* cv. Sonora Anaheim) were sequenced using 454 pyrosequencing **[4] Table 2 (see supplementary material).** A total of 1,838,567 pyrosequencing reads were obtained with an average length of 99.89 nt with an average quality of 27.90 **Table 3 (see supplementary material).**

Capsicum reference transcriptome:

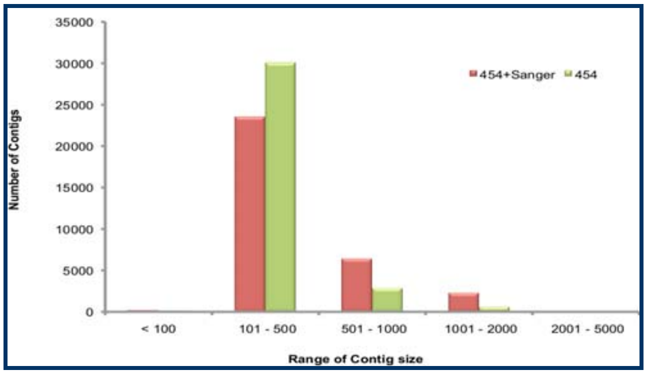


Figure 2: Comparison of contig sizes obtained with 454-only and hybrid assembly process. When the reads from 454 pyrosequencing are assembled separately, the majority of the contigs are in the 100-499 nt size range. When Sanger-derived sequences are included in the assembly, the number of contigs in the 101-500nt decreases almost 20% and the number of contigs in the next size ranges dramatically increase, especially the 1001 – 2000 size range.

Three rounds of assembly were performed using Newbler (v1.3) with default parameters. First, each library was assembled independently to identify sequences classified as "assembled reads" and "singletons" by Newbler **Figure 1, Table 2 (see supplementary material) .** Sequences classified as "repeats" were discarded due to over-representation, which are problematic in the assembly process. A total of 1,253,773 sequences from all 9 runs derived from the "assembled reads" and "singletons" were used for a second assembly **(Figure 1)** in which, 887,718 reads were assembled into 33,652 contigs with an average length of 251.7 nt **Table 3 (see supplementary material).** A third assembly was performed using a hybrid approach in which a total of 1,324,516 sequences (1,253,773 pyrosequencing-reads plus 70,743 ESTs) were assembled **(Figure 1, Table 3).** A total of 1,144,574 sequences were assembled into 32,538 contigs with 92,211 remaining as singletons **Table 3 (see supplementary material)**. Custom Perl scripts were used to filter polyA/T regions, low quality, and short contigs (<90 nt) obtaining 32,314 high-quality contigs and 51,118 singletons **Table 3 (see supplementary material).** The hybrid assembly compared to the 454-only strategy increased the contig length from 251.71 nt (454-only) to 388.5 nt (hybrid) **(Table 3)** and the number of large contigs (≥ 500 nt) increased from 3,438 to 8,792 with an average size increasing from 777 to 871 nt **(Figure 2).**

Assembly quality measurement

Using BLASTN **[5]** an analysis of alignment coverage was carried out against pepper ESTs database **[6].** The results revealed than 60% (19,388) of the contigs had coverage greater than 90% with average identities of 99%; 21% (6,575) of the contigs have 60-89% coverage and 99% identity; the remaining contigs (6,351) had coverage less than 60% with 98% identity. The assembled contigs were examined for similarity to pepper and other plant sequence databases. We found that 52.5% (17,090) of our contigs (E ≤1e-06, % identity ≥ 90) had high identity with sequences in the Pepper ESTs dataset **(Figure 3).** The remaining contigs (47.5%) were then compared against the

Tomato Unique Genes database **[7],** and 14.3% (4,654) of contigs had high similarity (E ≤1e-06) in this dataset (Figure 3). For the remaining contigs, a sequential BLASTX-based search **[5]** was extended to three protein sequence datasets, *Arabidopsis thaliana* peptides (TAIR) **[8],** *Oryza sativa* peptides (RefSeq) **[9]** and NCBI nr. Of the remaining contigs, ~7.5% (2,452) had high identity to Arabidopsis and Rice proteins (E ≤1e-05) and 1% (375) had matches to NCBI nr sequences (E ≤1e-04). In summary, 75.5% of our contigs share sequence similarity with transcripts, genes or proteins in public databases **(Figure 3)** and 7,481 of these sequences (23%) have not been previously reported for the species *Capsicum annuum*. A total of 7,743 (23.8%) contigs are novel transcripts that are specific to the *C. annuum*.

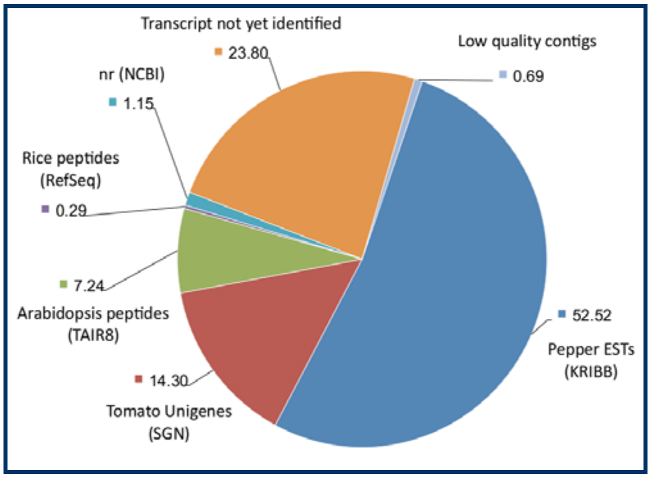
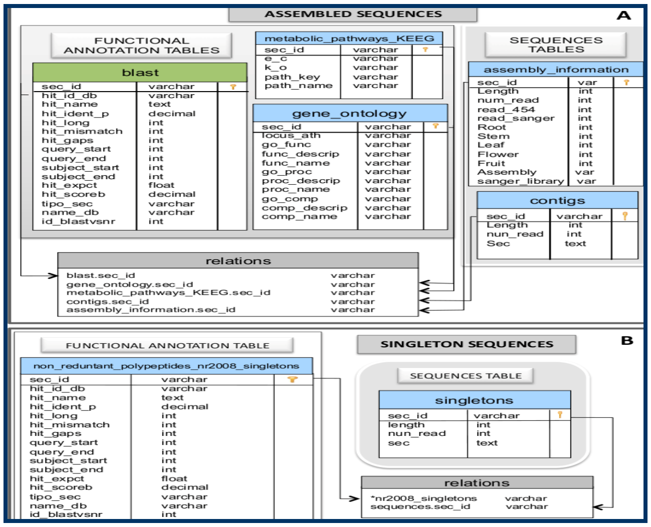


Figure 3: Sequential comparison of hybrid contigs with different plant databases. BLASTN searches of the contigs were performed against two Solanaceae databases, Pepper ESTs hosted at KRIBB **[6],** and Tomato Unique genes from the Sol Genomic Network [7]. We use an E-value threshold ≤ 1e-06 and identity rate ≥ 90% for Pepper ESTs, and E-value ≤ 1e-06 for Tomato Unigene. Sequences without a significant match were compared (BLASTX) against Arabidopsis and Rice proteome **[8, 9]** using an E-value ≤ 1e-05. Finally, the remaining contigs (no match in the previous four databases) were searched against the NCBI nr database (E-value cutoff ≤ 1e-04). The chart shows the percentage of contigs sequences identified in each database with 23.8% lacking homology to any of these databases.

Features of the Web database

A website and database were constructed using open source technologies with the Linux operating system (Ubuntu v9.1). MySQL Database Management System (v5.1) was used to store and manage the data. An Apache HTTP server (v2.2.4), PHP Hypertext Pre-processor (v5.3.1), JavaScript (v3.1.0) and HTML (v4) were used to create the query-builder module for connecting and querying the database. Custom Perl (v5.10) scripts were used to automatically parse the database and Joomla (v1.5) was used as content management system (CMS) for building the web site. Information stored in the database is divided into two main sections: assembled (contigs) and singleton sequences. Each section is then sub-divided into two categories, Functional Annotation and Sequences. **(Figure 4A)** shows the tables that store the data derived from assembled sequences (contigs) and their relationships. A total of eight tables store information related to functional annotation. The genomics databases used for functional annotation were: *S. lycopersicum* unigenes **[7],** *A. thaliana* peptides **[8],** *O. sativa* peptides **[9],** *P. patens* **[10],** NCBI nr and UniRef100 (UniProt).



blast		
sec_id	varchar	🔑
hit_id_db	varchar	
hit_name	text	
hit_ident_p	decimal	
hit_long	int	
hit_mismatch	int	
hit_gaps	int	
query_start	int	
query_end	int	
subject_start	int	
subject_end	int	
hit_expct	float	
hit_scoreb	decimal	
tipo_sec	varchar	
name_db	varchar	
id_blastvsnr	int	

metabolic_pathways_KEEG		
sec_id	varchar	🔑
e_c	varchar	
k_o	varchar	
path_key	varchar	
path_name	varchar	

gene_ontology		
sec_id	varchar	🔑
locus_ath	varchar	
go_func	varchar	
func_descrip	varchar	
func_name	varchar	
go_proc	varchar	
proc_descrip	varchar	
proc_name	varchar	
go_comp	varchar	
comp_descrip	varchar	
comp_name	varchar	

assembly_information		
sec_id	var	🔑
Length	int	
num_read	int	
read_454	int	
read_sanger	int	
Root	int	
Stem	int	
Leaf	int	
Flower	int	
Fruit	int	
Assembly	var	
sanger_library	var	

contigs		
sec_id	varchar	🔑
Length	int	
nun_read	int	
Sec	text	

relations	
blast.sec_id	varchar
gene_ontology.sec_id	varchar
metabolic_pathways_KEEG.sec_id	varchar
contigs.sec_id	varchar
assembly_information.sec_id	varchar

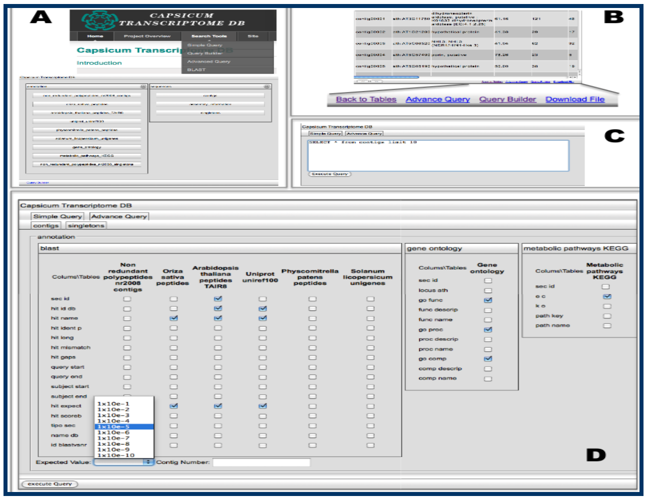
non_reduntant_polypeptides_nr2008_singletons		
sec_id	varchar	🔑
hit_id_db	varchar	
hit_name	text	
hit_ident_p	decimal	
hit_long	int	
hit_mismatch	int	
hit_gaps	int	
query_start	int	
query_end	int	
subject_start	int	
subject_end	int	
hit_expct	float	
hit_scoreb	decimal	
tipo_sec	varchar	
name_db	varchar	
id_blastvsnr	int	

singletons		
sec_id	varchar	🔑
length	int	
nun_read	int	
sec	text	

relations	
*nr2008_singletons	varchar
sequences.sec_id	varchar

Figure 4. Database tables. Several tables were created for the Capsicum database. **A)** Tables for the assembled sequences. The functional annotation category has eight tables in total. Six tables have the same structure; the "blast" which stores tabular outputs from BLAST alignments is representative of these six tables. Two other tables store Gene Ontology **[2]** and Metabolic Pathway information **[3].** The sequence category has two tables: "assembly_information" and "contigs", which store information of the assembly and the assembled sequences, respectively. **B)** Two tables store information for singleton sequences. The "functional annotation" table stores the tabular output from BLAST alignment and the "singletons" table stores the sequences. Each group (assembled sequences or singletons) has an additional table called "relations" showing the relations between the tables. Tables for each group are related using "sec_id" as primary key; however, assembled sequences and singletons tables are not related each other.

The GO terms **[2]** and KEGG metabolic pathway **[3]** annotations derived from the highest scoring BLASTX results from TAIR **[8].** The assembly and sequence data are stored in tables named "assembly information" and "contigs", respectively. The "assembly information" table contains the number of sequences (either Sanger or 454 reads) that were assembled into each contig, and the sequence origin (root, stem, leave, flower or fruit). The "Contigs" table stores the assembled sequences. Two additional tables were created to store information for singletons. The majority of the singletons (97.4%) are 454 pyrosequencing reads with an average length of ~100 nt. However, 1,349 are Sanger-derived ESTs with an average length of ~650 nt **Table 3 (see supplementary material).** We annotated all singletons using BLASTX against the NCBI nr database. Sequence and annotation are stored in separate tables **(Figure 4B)** using "seq_id" as a primary key to relate the two tables. A query-builder module, adapted as user-friendly Web interface, was developed. The module allows the user to explore the database in three different ways: i) Simple search, ii) Advanced query, and, iii) Query builder **(Figure 5).** Using the "Simple search" option, the user is able to access and download the full content of a specific table **(Figure 5A).** The searches can be refined using "search options". The advanced query section was designed for users with SQL knowledge in which searches can be performed through SQL **(Figure 5C**). The Query-builder permits the user to collect functional annotation and sequence information from different tables **(Figure 5D).** This module was designed to use checkboxes to make multiple attribute selections from a number of tables and columns and also provides search options to define a query **(Figure 5D).** In every result generated by the module, the user is able to download a file in CSV (comma separated value) format **(Figure 5B).**



contig20021	ath:AT3G11750	dihydroneopterin aldolase, putative / K01633 dihydroneopterin aldolase [EC:4.1.2.25]	61.16	121	48
contig20022	ath:AT1G21280	hypothetical protein	41.38	29	17
contig20023	ath:AT5G08520	NHL3; NHL3 (NDR1/HIN1-like 3)	41.04	62	32
contig20024	ath:AT5G57490	porin, putative	75.28	23	5
contig20025	ath:AT5G05190	hypothetical protein	52.00	26	19

```
SELECT * from contigs limit 10
```

Colums\Tables	Non redundant polypeptides nr2008 contigs	Oriza sativa peptides	Arabidopsis thaliana peptides TAIR8	Uniprot uniref100	Physcomitrella patens peptides	Solanum licopersicum unigenes
sec id	□	□	☑	□	□	□
hit id db	□	□	☑	☑	□	□
hit name	□	☑	☑	☑	□	□
hit ident p	□	□	□	□	□	□
hit long	□	□	□	□	□	□
hit mismatch	□	□	□	□	□	□
hit gaps	□	□	□	□	□	□
query start	□	□	□	□	□	□
query end	□	□	□	□	□	□
subject start	□	□	□	□	□	□
subject end		□	□	□	□	□
hit expect		☑	☑	☑	□	□
hit scoreb		□	□	□	□	□
tipo sec		□	□	□	□	□
name db		□	□	□	□	□
id blastvsnr		□	□	□	□	□

Colums\Tables	Gene ontology
sec id	□
locus ath	□
go func	☑
func descrip	□
func name	□
go proc	☑
proc descrip	□
proc name	□
go comp	☑
comp descrip	□
comp name	□

Colums\Tables	Metabolic pathways KEGG
sec id	□
e c	☑
k o	□
path key	□
path name	□

Figure 5: Examples of resources available in Capsicum Transcriptome DB. Access to different tools available in the module is demonstrated. **A**) The Simple Query module where a user can perform a table-specific searches is shown. Each button represents one table and search options are provided to define the query. **B)** Specific-table search. After the data retrieval by the module, the user is able to go back to the tables, redefine searches by using "Advanced query" or "Query Builder" or simply download the results file in cvs-format file. The results are displayed on the screen in groups of 30 rows. **C)** Using the Advanced Query the user can perform searches through SQL queries. **D**) The Query Builder module was designed with checkboxes to make multiple selections across different tables and columns where every box represents a different column or attribute from each table**.**

Utility:

Our user-friendly web interface is a straightforward tool, providing access to molecular data in a simple and dynamic manner without requisite training in bioinformatics. The user is allowed to perform queries to analyze a group of sequences or an individual sequence. The database contains 32,314 high-quality assembled contigs and 51,118 high-quality singletons. Functional annotation was assigned to 75% of the contigs including 21,744 sequences common to Solanaceae (Pepper ESTs and Tomato unique genes) and 7,481 novel sequences not previously reported for *Capsicum annuum*.

Future Developments:

We will continue sequencing several chill-pepper tissues and updating the database with new sequences and functional annotation. Comments and requests regarding the database should be sent to Dr. Rafael Rivera-Bustamante at capsicum@ira.cinvestav.mx

Conclusions:

The data presented in this study shows the advantages of using multiple sequencing technologies for *de novo* assembly of a transcriptome in the absence of a reference genome. With the hybrid assembly approach, we were able to improve multiple contig quality measures. A detailed coverage analysis showed the high quality of the assembly suggesting that the rate of contig artifacts is low. Using an in-depth annotation pipeline, we identified 75% of the contigs including 7,481 novel sequences not previously reported for *Capsicum annum*. These data expand our knowledge of gene expression across diverse pepper tissues and complement the data in existing databases **[6,7].** In summary, the bioinformatics methods applied to the reported data demonstrate that our Capsicum Reference transcriptome is a reliable resource and an important "hot" tool for downstream functional studies.

Acknowledgements:

This work was supported by Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA) and Consejo Nacional de Ciencia y Tecnologia (Conacyt) [11806] We acknowledge support from Conacyt-Mexico to Elsa Gongóra-Castillo (Ph. D. fellowship).

We acknowledge and thank Lisset Herrera-Isidrón, Rosa M. Adame-Alvarez, Cesar Aza-González, Martha G. Betancourt-Jiménez, Fernando Hernández-Godinez, Enrique Ibarra-Laclette, Hector G. Nuñez-Palenius and Harumi Shimada-Beltran for participating in the construction of cDNA libraries. We would like to specially thank to Eduardo Jaramillo-Olvera for his support during the database and website development, and Drs. Robin Buell and Rebecca Davidson for critical reading of the manuscript and helpful comments.

References:

[1] Aguilar-Melendez A *et al. American Journal of Botany*. 2010 **96**: 1190 [PMID: 21628269]
[2] Ashburner M *et al. Nat Genet*. 2000 **25**: 25 [PMID: 10802651]
[3] Kanehisa M & Goto S, *Nucleic Acids Res*. 2000 **28**: 27 [PMID: 10592173]
[4] Margulies M et al. Nature. 2005 **437**: 376 [PMID: 16056220]
[5] Altschul SF *et al. J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]
[6] Kim, HJ *et al. BMC Plant Biol*. 2008 **8**: 101 [PMID: 18844979]
[7] Mueller LA *et al. Plant Physiol*. 2005. **138**: 1310 [PMID: 16010005]
[8] Rhee SY *et al. Nucleic Acids Res*. 2003. **31**: 224 [PMID: 12519987]
[9] Tanaka T *et al. Nucleic Acids Res*. 2008 **36**: D1028 [PMID: 18089549]
[10] Rensing SA *et al. Science*. 2008 **319**: 64 [PMID: 21551031]

Edited by P Kangueane
Citation: **Góngora-Castillo *et al*.** Bioinformation 8(1): 043-047 (2012)

Supplementary Material:

Table 1: Summary of sequences contained in the Capsicum Transcriptome DB:

Sequencing Method	Sequence type	Tissue	No. of sequences [z]
Sanger	EST	Root [a]	6,063
Sanger	EST	Stem [a]	6,751
Sanger	EST	Leaf [a]	5,875
Sanger	EST	Flower [b]	11,842
Sanger	EST	Fruit [c]	23,218
Sanger	EST	Pericarp [c]	11,819
Sanger	EST	Placenta [c]	1,757
Sanger	EST	Sterile seedlings	1,698
Sanger	EST	Mixed tissues [d]	1,720
454	NGS	Leaf [e]	1,838,567
Total Reads			**1,909,310**

[a]Roots, stems and leaves were sampled at three different stages: seedlings with cotyledons, plants with 5 to 7 true leaves and branched plants; [b]Mix of flower buds. [c]Fruit was harvested at three different ripening stages: 20, 40 and 60 days post-anthesis; [d]Root, stem and leaf; [e]DNA virus-infected and healthy leaf tissues; [z]Number of obtained sequences after filtering by quality and size for Sanger-type sequences. Raw sequences for 454-type sequences.

Table 2: Statistics summary of the *de novo* assembly for nine runs of 454-pyrosequencing

Sample	Number of Runs	Total reads	Average size (nt)	Average quality	Assembled Reads (%)	Repeat[z] (%)	Singletons (%)	Contigs
Leaf1 [a]	1	222,558	99.34	27.8	144,418 (64.9)	37,388 (16.8)	40,752 (18.3)	9,863
Leaf2 [b]	5	865,103	98.64	28.14	520,467 (60,2)	268,861 (31)	75,775 (8.8)	30,301
Leaf3 [c]	3	750,906	101.7	27.53	410,685 (54.7)	278,545 (37.1)	61,676 (8.2)	30,846
	9	**1,838,567**			**1,075,570**		**178,203**	

[a]Non-infected tissue; [b]DNA virus-infected tissue in the symptom stage; [c]DNA virus-infected tissue in the recovery stage; [z]Numberof 454-reads identified as "repeat" sequences by Newbler assembler.

Table 3: Summary of the *de novo* assembly using a hybrid approach

Raw data	
Total bases	159,832,345
Filtered 454-reads	1,253,773
Filtered ESTs	70,743
Average ESTs length	678.35
Average 454-contig length [a]	251.71
Assembly results	
Assembled hybrid sequences	1,144,574
Total number of contigs	32,538
Number of high quality contigs	32,314
Average Hybrid-contig length	388.5
N50 contig size	631
Range contig length	100 - 3,033
Number of large contigs (≥ 500 bp)	8,792
Total number of singletons	92,211
Number of high quality singletons [b]	51,118

[a]Contigs derived from 454-only assembly; [b]Singletons elements contain 1,349 Sanger-type ESTs and 49,769 pyrosequencing-type reads.