



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



Accessing Sequence Resources



David Francis
The Ohio State
University, OARDC
SolCAP workshop

Supported by the AFRI Plant Breeding, Genetics, and Genomics
Program of USDA's National Institute of Food and Agriculture

This module Introduces resources for sequence data. After following this module, you should be able to:

Locate tomato sequence data in various databases.

Know how to retrieve sequence data, and even create specialized sets of sequence data.

Demonstrate knowledge of distributed resources for sequence analysis.

Where do I go for data?

National Center for Biotechnology Information (NCBI):
<http://www.ncbi.nlm.nih.gov/>

UniProt (SWISS-PROT): <http://www.uniprot.org/>

European Molecular Biology Laboratory (EMBL) nucleotide
sequence database <http://www.ebi.ac.uk/embl/>

Crop/family Specific databases

Solanaceae Genomics Network (SGN): <http://sgn.cornell.edu/>

Solanaceae Coordinated Agricultural Project:
<http://solcap.msu.edu/>

Gene indexes (Formerly TIGR):
<http://compbio.dfci.harvard.edu/tgi/>



National Center for
Biotechnology Information

Search All Databases ▾

Search

Clear

Resources

NCBI Home

All Resources (A-Z)

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.



|| 1 2 3 4

How To...

- [Determine conserved synteny between the genomes of two organisms](#)
- [Find a homolog for a gene in another organism](#)
- [Obtain the full text of an article](#)
- [Design PCR primers and check them for specificity](#)

Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

NCBI News

[NCBI Workshop: A Practical Guide to Genome-Scale Data](#)


03 Nov 2010

Presented at the ASHG Meeting on Thursday, November 4, 2010,

[New Find-in-sequence feature for Nucleotide and Protein databases](#)

‘<http://www.ncbi.nlm.nih.gov/>’





GenBank Overview

PubMedEntrezBLASTOMIMBooksTaxonomyStructure

SearchEntrezforGo

NCBI Home
NCBI Site Map
Submit to GenBank
Submit an update
Search GenBank
GenBank and RefSeq:
a comparison
BLAST

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2008 Jan 36\(Database issue\):D25-30](#)). There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank [record](#) may be viewed for a *Saccharomyces cerevisiae* gene.

In The News: 2009 H1N1 Flu Virus (Swine Flu)

The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the [NCBI Flu Resource](#)

▶ NLM/NCBI 2009 H1N1 Flu Resources:

H1N1 Flu Info

<http://www.ncbi.nlm.nih.gov/Genbank/index.html>



Lycopersicon esculentum SP6A (SP6A) gene, complete cds

Features Sequence

DEFINITION *Lycopersicon esculentum* SP6A (SP6A) gene, complete cds.

VERSION AY186737.1 GI:28200393

KEYWORDS

SOURCE *Solanum lycopersicum* (*Lycopersicon esculentum*)

ORGANISM *Solanum lycopersicum*

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
asterids; lamiids; Solanales; Solanaceae; Solanoideae; Solaneae;
Solanum; Lycopersicon.

REFERENCE 1 (bases 1 to 3463)

AUTHORS Carmel-Goren, L., Liu, Y.S., Lifschitz, E. and Zamir, D.

TITLE The SELF-PRUNING gene family in tomato

JOURNAL Plant Mol. Biol. 52 (6), 1215-1222 (2003)

PUBMED 14682620

REFERENCE 2 (bases 1 to 3463)

AUTHORS Carmel-Goren, L., Liu, Y.S., Lifschitz, E. and Zamir, D.

Change Region Shown

Customize View

Sequence Analysis Tools

► BLAST Sequence

► Pick Primers

Recent Activity

GenBank “flat file” format

lycopersicum[ORGN] AND
Se... (7)

Nucleotide

[All links from this record](#)

► Full text in PMC

Google

Search

Check

AutoFill

Bookmarks

Tools

http://www.ncbi.nlm.nih.gov/nucore/28200393?ordinalpos=1&itool=EntrezSystem2.PEntrez.Sequence.Sequence_ResultsP

Google

Nucleotide

Page

Done

Internet

100%

ORIGIN

1

ctagaaaaag

gaataatcga

cgaaaatddd

aatgtacact

ttttaactca

agctagtttt

61

ggcgaactga

atcataatcg

tttattttcgg

ttcttggcct

gaatgttgat

ttaatccgca

121

catgtgacac

ctctacttag

ggcgcatggc

ttcttgagtt

ggttttatac

atcttacaaa

181

gatcaacatc

acctaagaat

aatatagttt

catagtaaaa

gaataccata

gaaagggttat

241

agattctttg

acaccaataa

aatatcaaca

tagtaagttt

actactcata

taaatagcca

301

taccctttga

taaaaatddd

tatcattgca

ataacaatcg

aaagaagaaa

aaactaaaaa

361

atcctagcat

attgtcatat

tatatatatt

atatataaac

tttatattat

catgcctaga

421

gttgatccat

tgatagttgg

togtgtgata

ggtgaagttt

tagatccatt

cactaggtct

481

gttgatctta

gagttgttta

taataataga

gaagtgaaca

atgcatgtgt

gttgaaacct

541

tcacaagttg

ttatgcaacc

taagggttat

atcgagggg

acgatcttgc

caccttttac

601

actctggtaa

atdddtaatt

ttactcgctc

tgtcttgttc

tattgtattt

tatgtgacac

661

tattatatac

tatttgagga

gttatagagg

tttaatgttc

tgaattatgc

actattgttt

721

cttatattat

taattatagt

attdcttttt

tgttctctct

tattttatgt

gacgctatta

781

tcatttgagg

atcgtcacaa

aggtttaata

ttttttatct

aattttttgg

attattaatt

841

atatattata

tattcatcct

cacttgcat

atcttaacgt

gtactattac

tattctgaga

901

gttaaaactt

aggtttaatt

ttcaacgac

aagacgatt

aagatattat

tttaattgta

961

attdcatgag

ctattaacta

ctcttaagta

aaaaaaaaag

aagaaaattc

actcttttta

1021

tttattttat

gtgaagctat

aagagggtta

ttactctttt

ctaattgttt

tgttattatt

1081

aactattatc

gcttatatta

ttttttatat

attcgctcgt

tcatttttac

tttaggtgat

1141

aatattacta

ttctgagagt

taaacaaagg

cttcattttc

tattatcaag

acggtattag

1201

atatgttttc

actgttttca

tgagctattg

actattactt

ttctctgact

ctctcggttc

1261

tattttttgt

gaogctatta

atatttgagg

atagtattta

ctttctttta

attdttcttt

1321

attdatgttt

tgattacatt

gttggttgag

attcttttgt

aattgttttt

ttatgaagat

1381

tatggtggat

cctgatgctc

caagcccaag

caatcctaac

ttgagggagt

atctacactg

1441

gttagtattt

togatcttat

tagatcaaac

acgtaaaaat

tctttttttt

tttttgaatt

1501

aataatgacg

atgaaactcg

aatcaagaaa

cttggttagg

ctcataatgc

aaagtgaaaa

1561

cccacaattg

ataacataaa

atcataatgt

catgtgtact

gaaaatccta

ttactctagt

1621

tttcctaatt

atcatcatcg

togttattat

tgttattttt

attatctctc

gtatatttta

1681

ttcttcgatt

tttatggtaa

tacgttattt

ctttttctct

gtttattgta

ttgttttctc

1741

gattattttc

atcataaact

cttcaactact

atattcccta

ttcatacatg

ctacttaaat

1801

caagagtcta

tcaaaaaaaa

tttctctacc

ttcacaatgc

agatatatcg

tacgtaaata

1861

cacataaac

tctacctaaa

caccacttat

ataattatac

taaatatatt

attattatta

GenBank “Flat file” format (continued).

GenBank: AY186737.1

Lycopersicon esculentum SP6A (SP6A) gene, complete cds

>gi|28200393|gb|AY186737.1| Lycopersicon esculentum SP6A (SP6A) gene, complete cds

```
CTAGAAAAAGGAATAATCGACGAAAAATTTAATGTACACTTTTTAACTCAAGCTAGTTTTGGCGAACTGA
ATCATAATCGTTTTATTTCCGTTCTTGGCCTGAATGTTGATTTAATCCGCACATGTGACACCTCTACTTAG
GGCGCATGGCTTCTTGAGTTGGTTTTATACATCTTACAAAGATCAACATCACCTAAGAATAATATAGTTT
CATAGTAAAAGAATAACCATAGAAAGGTTATAGATTCTTTGACACCAATAAAATATCAACATAGTAAGTTT
ACTACTCATATAAATAGCCATACCCTTTGATAAAAAATTTTATCATTGCAATAACAATCGAAAGAAGAAA
AAACTAAAATATCCTAGCATATTGTCATATTATATATATTATATATAAACTTTATATTATCATGCCTAGA
GTTGATCCATTGATAGTTGGTCGTGTGATAGGTGAAGTTTTAGATCCATTCACTAGGTCTGTTGATCTTA
GAGTTGTTTATAATAATAGAGAAGTGAACAATGCATGTGTGTTGAAACCTTCACAAGTTGTTATGCAACC
TAAGGTTTTATATCGGAGGGGACGATCTTCGCACCTTTTACACTCTGGTAAATTTTAAATTTTACTCGCTC
TGCTTTGTTCTATTGTTTATGTTGACACTATTATATACTATTTGGAGAGTTATAGAGGTTTAAATGTTT
TGAATTATCGCATTTGTTTCTTATATTATTAATTATAGTATTTCTTTTTGTTCTCTCTTATTTTATGT
GACGCTATTATCATTTTGGAGATCGTCACAAAGGTTTAAATTTTTTATCTAAATTTTTGGATTATTAATT
ATATATTATATATTTCATCCTCACTTGCATTATCTTACGTGGTACTATTACTATTCTGAGAGTTAAACTT
AGGTTTAAATTTTCAACGATCAAGACGATTTAAGATATTATTTTAAATGTTAATTTTATGAGCTATTAAC
CTCTTAAGTAAAAAAGAAAGAAATTCACCTTTTTTATTTATTTTATGTTGAAGCTATAAGAGGTTTA
TTACTCTTTTCTAATGTTTTGTTATTATTAACATTATCGCTTATATTATTTTATATATTTCGCTCGT
TCATTTTTACTTTAGGTGATAATATTACTATTCTGAGAGTTAAACAAAGGCTTCATTTTCTATTATCAAG
ACGGTATTAGATATGTTTCACTGTTTTCATGAGCTATTGACTATTACTTTCTCTGACTCTCTCGGTTT
TATTTTTTGTGACGCTATTAATATTTGGAGATAGTATTTACTTTCTTTTAAATTTTTCTTTATTTATGTTT
TGATTACATTGTTGTTTGGAGATTCTTTGTAATTGTTTTTATGAAGATTATGGTGGATCCTGATGCTC
```

Change Region Shown

Customize View

Sequence Analysis Tools

- [BLAST Sequence](#)
- [Pick Primers](#)

Recent Activity

[Turn Off](#) [Clear](#)

- [lycopersicum\[ORGN\] AND se... \(7\)](#) Nucleotide
- [Le000035s_FL7600 Tomato Genome SNP, InDel Search](#)
- [FI855311 \(1\)](#) GSS
- [Lycopersicon esculentum SP6A \(SP6A\) gene, complete cds](#)
- [lycopersicum\[ORGN\] AND](#)

FASTA file format:

FASTA is the standard for sequence data format.

“>” is followed by a name/description of the sequence.

Everything following the first paragraph break is expected to be a sequence string of nucleotide or protein sequence.

Search across databases

GO

Clear

Help

'http://www.ncbi.nlm.nih.gov/sites/gquery'

Welcome to the Entrez cross-database search page



PubMed: biomedical literature citations and abstracts



PubMed Central: free, full text journal articles



Site Search: NCBI web and FTP sites



Books: online books



OMIM: online Mendelian Inheritance in Man



OMIA: online Mendelian Inheritance in Animals



Nucleotide: Core subset of nucleotide sequence records



EST: Expressed Sequence Tag records



GSS: Genome Survey Sequence records



Protein: sequence database



Genome: whole genome sequences



Structure: three-dimensional macromolecular structures



dbGaP: genotype and phenotype



UniGene: gene-oriented clusters of transcript sequences



CDD: conserved protein domain database



3D Domains: domains from Entrez Structure



UniSTS: markers and mapping data



PopSet: population study data sets



Descriptions of sequence databases:

Nucleotide – Contains high quality annotated sequences

EST – “Expressed Sequence Tag”. Derived from cDNA (mRNA) and therefore represents transcribed (expressed) sequences.

GSS – “Genomic short sequences”. Contains genomic sequence. For example, sequenced PCR products.

Unigene - Each UniGene entry is a set of transcript sequences that appear to come from the same locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location

<http://www.ncbi.nlm.nih.gov/unigene>

UniGene Home - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

Google

File Edit View Favorites Tools Help

Google Search

UniGene Home

My NCBI [Sign In] [Register]

NCBI

UniGene

ORGANIZED VIEW OF THE TRANSCRIPTOME

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals

Search UniGene for Go Clear

Limits Preview/Index History Clipboard Details

UniGene

Homepage

FAQs

Query Tips

Library Browser

DDD

Download UniGene

Related Databases

Gene

HomoloGene

dbEST

Trace Archive

UniGene: An Organized View of the Transcriptome.

Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location.

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus (cow)	43,448
Canis lupus familiaris (dog)	27,853

Internet

100%

Start

E.

2

2

B.

A.

M.

My Documents

Family

Desktop

10:54 AM

UniGene Home - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

Google

File Edit View Favorites Tools Help

Google Search

UniGene Home

Page Tools

Citrus clementina	9,123
Citrus sinensis (Valencia orange)	15,808
Glycine max (soybean)	31,395
Gossypium hirsutum (upland cotton)	21,738
Gossypium raimondii	3,297
Helianthus annuus (sunflower)	12,216
Lactuca sativa (garden lettuce)	7,940
Lotus japonicus	14,493
Malus x domestica (apple)	16,932
Medicago truncatula (barrel medic)	18,098
Nicotiana tabacum (tobacco)	19,753
Populus tremula x Populus tremuloides (hybrid aspen)	9,652
Populus trichocarpa (western balsam poplar)	14,965
Prunus persica (peach)	7,620
Raphanus raphanistrum (wild radish)	18,788
Raphanus sativus (radish)	17,649
Solanum lycopersicum (tomato)	18,228
Solanum tuberosum (potato)	18,784
Vigna unguicula See UniGene Summary for Solanum lycopersicum	15,740
Vitis vinifera (wine grape)	23,166

http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=4081

Internet 100%

Start E. 2. 2. B. A. M. My Documents Family Desktop 10:55 AM

Other databases:

The SWISS-PROT database contains high-quality annotation, is non-redundant and cross-referenced to many other databases in May 26, 2009, the SWISS-PROT database was merged into the UniProt database.

<http://www.uniprot.org/>

European Molecular Biology Laboratory (EMBL) nucleotide sequence database <http://www.ebi.ac.uk/embl/>



Search in

Query

Protein Knowledgebase (UniProtKB)

Search

Clear

Fields »

Search

Blast

Align

Retrieve

ID Mapping

WELCOME

The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	<p>Protein knowledgebase, consists of two sections:</p> <ul style="list-style-type: none"> ★ Swiss-Prot, which is manually annotated and reviewed. ★ TrEMBL, which is automatically annotated and is not reviewed.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.

NEWS



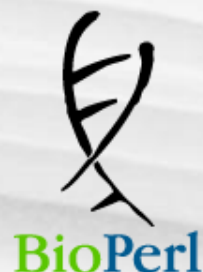
UniProt release 15.5 – Jul 7, 2009

New insights into drug development with Polyketide synthases · Cross-references to UCSC

- › Statistics for UniProtKB: [Swiss-Prot](#) · [TrEMBL](#)
- › Forthcoming changes
- › News archives

SITE TOUR





main links

- Main Page
- Getting Started
- Downloads
- Installation
- Recent changes
- Random page

documentation

- Quick Start
- FAQ
- HOWTOs
- Scrapbook
- BioPerl Tutorial
- Tutorials
- Deobfuscator
- Browse Modules

community

- News
- Mailing lists
- Supporting BioPerl
- BioPerl Media
- Hot Topics
- About this site

page discussion view source history

Swissprot sequence format

Description

The following is an example of [Swissprot](#) "flat" format, the only swissprot format flavor that [BioPerl](#) is capable of parsing as of version 1.5.1. The [Bio::SeqIO](#) system can parse these files with the [Bio::SeqIO::swiss](#) module.

[Swissprot](#) also releases an [XML](#) formatted database.

Example

```
ID  MA32 HUMAN      STANDARD;      PRT;      282 AA.
AC  Q07021;
DT  01-FEB-1995 (Rel. 31, Created)
DT  01-FEB-1995 (Rel. 31, Last sequence update)
DT  01-OCT-2000 (Rel. 40, Last annotation update)
DE  COMPLEMENT COMPONENT 1, Q SUBCOMPONENT BINDING PROTEIN, MITOCHONDRIAL
DE  PRECURSOR (GLYCOPROTEIN GC1QBP) (GC1Q-R PROTEIN) (HYALURONAN-BINDING
DE  PROTEIN 1) (PRE-MRNA SPLICING FACTOR SF2, P32 SUBUNIT) (P33).
GN  GC1QBP OR HABP1 OR SF2P32 OR C1QBP.
OS  Homo sapiens (Human).
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC  Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX  NCBI_TaxID=9606;
RN  [1]
RP  SEQUENCE FROM N.A., AND SEQUENCE OF 74; 76-93 AND 208-216.
RC  TISSUE=FIBROBLAST;
RX  MEDLINE=94085792; PubMed=8262387;
RA  Honore B., Madsen P., Rasmussen H.H., Vandekerckhove J., Celis J.E.,
RA  Leffers H.;
RT  "Cloning and expression of a cDNA covering the complete coding region
RT  of the P32 subunit of human and rDNA subunit factor SF2."
```

This ends an introduction to general on-line databases.

Next, a discussion of (1) family specific resources and (2) downloading “customized” data.

Questions?



Other databases:

Crop/family specific databases

e.g. Solanaceae Genomics Network (SGN)

<http://sgn.cornell.edu/>

e.g. The arabidopsis information resource (TAIR)

<http://www.arabidopsis.org/>

Gene indexes (Formerly TIGR)

<http://compbio.dfci.harvard.edu/tgi/>



The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a [database](#) of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.



TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation.



Breaking News

Synteny Viewer At TAIR [July 7, 2009]

A synteny viewer, comparing syntenic regions between *A. thaliana* and *A. lyrata*, is now available at TAIR. More genomes will be added soon.

TAIR9 Genome Release [June 19, 2009]

The TAIR9 genome release is now available at TAIR and NCBI, with 282 new loci, updates to 1254 gene structures and 739 new splice variants.

Tips for searching DNA stocks including vectors and amiRNA clones [May 20, 2009]

Are you searching for clones and vectors available from ABRC? Here are some tips for finding them using the TAIR



http://compbio.dfci.harvard.edu/tgi/tgipage.html



Links

- [The Gene Indices](#)
- [TGI Software](#)
- [What's new?](#)
- [Definitions](#)
- [TGI Publications](#)
- [TGI FAQ](#)
- [TGI Disclaimer](#)
- [Contact Us](#)

The Gene Index Project Overview

The promise of genome projects has been a complete catalogue of genes in a wide range of organisms. While genome projects have been successful in providing reference genome sequences, the problem of finding genes and their variants in genomic sequence remains an ongoing challenge.

The sequencing of Expressed Sequence Transcripts (ESTs), fragments of genes that have been copied from DNA to RNA, provides the most comprehensive evidence for the existence of genes and their structure.

The goal of The Gene Index Project is to use the available EST and gene sequences, along with the reference genomes wherever available, to provide an inventory of likely genes and their variants and to annotate these with information regarding the functional roles played by these genes and their products.

In addition, we are attempting to use these catalogues to find links between genes and pathways in different species and to provide lists of features within completed genomes that can aid in the understanding of how gene expression is regulated.

Computational Biology and Functional Genomics Laboratory

The Gene Index Project



Animals

Plants

Protist

Fungi



Apple 2.0
5-27-09



Aquilegia 2.1
6-6-08



Arabidopsis 14.0
6-25-09



Barley 10.0
6-3-08



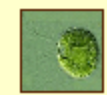
Bean 3.0
5-22-09



Beet 2.0
5-19-08



Brassica
napus 3.1
5-31-08



Chlamydomonas
reinhardtii 6.0
6-11-08



Clementine 2.0
5-22-09



Cocoa 3.0
5-21-09



Coffee 1.0
6-27-08



Cotton 10.0
5-26-09



Cotton
(raimondii) 1.0
7-2-08



Grape 6.0
7-30-08



Ice Plant 5.0
6-19-08



Leafy spurge 1.0
6-30-08

http://Solgenomics.net


Applications Places System Sat Oct 30, 6:38 AM David

Sol Genomics Network - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://solgenomics.net/ Google

Most Visited Getting Started Latest Headlines

 sol genomics network home | forum | contact | help

search maps genomes tools sol search

log in | new user

Maps & Markers

1
CT233
C015
C2_At4g15790

Genes

Phenotypes

Done

[FacMeeting092020...] [MeetingNotesREDi...] [bin - File Browser] [BioPerl-1.6.1 - File ...] Sol Genomics Netw...



sol genomics network

home | forum | contact | help

search

maps

genomes

tools

sol search

Browse

Tomato genome data

Projects

Solanaceae project (SOL)

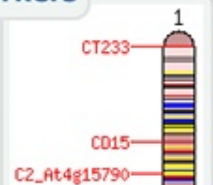
International Tomato Sequencing Project

Solanum pimpinellifolium Project (CSHL)

U.S. tomato sequencing project

log in | new user

Maps & Markers



Phenotypes



http://solgenomics.net/genomes/Solanum_lycopersicum/genome_data.pl

[FacMeeting092020...]

[MeetingNotesREDi...]

[bin - File Browser]


[BioPerl-1.6.1 - File ...]

Sol Genomics Netw...


Applications Places System Sat Oct 30, 6:57 AM David

Tomato Genome Data - Sol Genomics Network - Mozilla Firefox

File Edit View History Bookmarks Tools Help

[http://solgenomics.net/genomes/Solanum_lycopersicum/genome_data.pl](#)  Google

Most Visited Getting Started Latest Headlines

 sol genomics network

home | forum | contact | help

search maps genomes tools sol search

[log in](#) | [new user](#)

Tomato Genome Data

Tomato genome sequence builds

Release	Date	Description	Annotation	Download
1.00	Dec 2009	initial build, based on the Newbler assembler and containing only 454 sequencing data	ITAG1	scaffolds proteins cds
1.03	Jan 2010	like 1.00, but with additional 454 runs and improved contamination screen	Not annotated	scaffolds
cabog1.00	Mar 2010	All 454 data, bac end and fosmid end data, assembled using the CABOG assembler	Not annotated	scaffolds
1.50	Apr 2010	Includes all 454 data, bac ends, fosmid ends, polishing with Solexa and SOLiD data	Not annotated	scaffolds
2.00	Jun 2010	Release withdrawn.	Not annotated	-
2.10	Jun 2010	Additional scaffold merging using clone end sequences. Scaffolds placed and oriented using multiple physical maps, first release to	Not annotated	scaffolds , chromosomes

Done

[FacMeeting092020...] [MeetingNotesREDi...] [bin - File Browser] [BioPerl-1.6.1 - File ...] Tomato Genome Da...

SGN houses draft genome sequence for H1706 and LA1589

SolCAP's contribution:

GAI sequencing of transcribed sequences (transcriptomes)

S. lycopersicum (6 varieties and accessions)

OH9242

FL7600

NC84173

OH08-6405

PI 114490

PI 128216



http://solcap.msu.edu/

Applications Places System Sat Oct 30, 7:28 AM David

SolCAP Solanaceae Coordinated Agricultural Project - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://solcap.msu.edu/ Google

Most Visited Getting Started Latest Headlines

SolCAP



Solanaceae Coordinated Agricultural Project

- Home
- Potato
- Tomato
- Education/Outreach
- Protocols

The SolCAP project links together people from public institutions, private institutions and industries who are dedicated to the improvement of the Solanaceae crops: potato and tomato. Through innovative research, education and training the SolCAP project will focus on translating genomic advances to US tomato and potato breeding programs. This will lead to significantly improved varieties that benefit the processors, the consumer and the environment.

Done

[bin - File Browser] [BioPerl-1.6.1 - File ...] SolCAP Solanaceae ... [HCS806S2010 - Fil...] Extension - File Bro...

- Home
- Potato
- Tomato
- Education/Outreach
- Protocols
- Tools
- Publications
- Links
- Downloads
- Contacts

- Germplasm
- Phenotype Data
- GenotypeData

The SolCAP project links together people from public institutions, private institutions and industries who are dedicated to the improvement of the potato and tomato. Through innovative research, education and project will focus on translating genomic advances to US tomato programs. This will lead to significantly improved varieties that the consumer and the environment.



The SolCAP project is supported by the Agriculture and Food Research Initiative

[Home](#)[Potato](#)[Tomato](#)[Education/Outreach](#)[Protocols](#)[Tools](#)[Publications](#)[Links](#)[Downloads](#)[Contacts](#)

The SolCAP project is supported by the Agriculture and Food Research Initiative

The SolCAP project links together people from public institutions, private institutions and industries who are dedicated to the improvement of the Solanaceae crops: potato and tomato. Through innovative research, education and training the SolCAP project will focus on translating genomic advances to US tomato and potato breeding programs. This will lead to significantly improved varieties that benefit the processors, the consumer and the environment.

[Potato Intra SNPs](#)[Tomato Inter SNPs](#)[MSU Solanaceae SNP](#)[MSU Solanaceae SSR](#)



Solanaceae Coordinated Agricultural Project

[Home](#)[Potato](#)[Tomato](#)[Education/Outreach](#)[Protocols](#)[Tools](#)[Publications](#)

Tomato Intervarietal TA496 vs. Heinz1706 SNPs

This tool will give SNPS that were found between our TA496 EST assemblies and the Heinz1706 genomic sequence found in GenBank.

Search using any or all of the following criteria:

- Chromosome number: The chromosome number according to the GenBank record of the genomic sequence which the contig was mapped to.
- Genomic GenBank ID: The GI number of the genomic sequence which the contig was mapped to.
- Contig ID: The SolCAP contig ID for our Sanger-derived EST assemblies. The ID contains

Done

[bin - File Browser]

[BioPerl-1.6.1 - File ...]

SolCAP Solanaceae ...

[HCS806S2010 - Fil...]

Extension - File Bro...

Tomato

Education/Outreach

Protocols

Tools

Publications

Links

Downloads

Contacts

The SolCAP project is supported by the Agriculture and Food Research Initiative Applied Plant Genomics CAP Program of USDA's National Institute of Food and Agriculture.



United States
Department of
Agriculture
National Institute of
Food and
Agriculture

Search using any or all of the following criteria:

- Chromosome number: The chromosome number according to the GenBank record of the genomic sequence which the contig was mapped to.
- Genomic GenBank ID: The GI number of the genomic sequence which the contig was mapped to.
- Contig ID: The SolCAP contig ID for our Sanger-derived EST assemblies. The ID contains the species and variety type -- SolCAP_SPECIES_VARIETY_Contig#. More information regarding these assemblies can be found in the README found at the bottom of this page.
- Minimum EST Coverage: The minimum number of ESTs that were mapped to the genomic sequence at a given position for a SNP.

Chromosome number:

11

Genomic GenBank ID:

Contig ID:

Minimum EST Coverage:

2

Submit

Reset Form

Or download the data from our [FTP site](#):

- [Download TA496 EST assemblies](#)
- [README and tab-delimited text files with TA496 SNP results](#)

Recommendation:
Set to 3 or 4

Tomato

Education/Outreach

Protocols

Tools

Publications

Links

Downloads

Contacts

The SolCAP project is supported by the Agriculture and Food Research Initiative Applied Plant Genomics CAP Program of USDA's National Institute of Food and Agriculture.



United States
Department of
Agriculture
National Institute of
Food and
Agriculture

Search using any or all of the following criteria:

- Chromosome number: The chromosome number according to the GenBank record of the genomic sequence which the contig was mapped to.
- Genomic GenBank ID: The GI number of the genomic sequence which the contig was mapped to.
- Contig ID: The SolCAP contig ID for our Sanger-derived EST assemblies. The ID contains the species and variety type -- SolCAP_SPECIES_VARIETY_Contig#. More information regarding these assemblies can be found in the README found at the bottom of this page.
- Minimum EST Coverage: The minimum number of ESTs that were mapped to the genomic sequence at a given position for a SNP.

Chromosome number:

Genomic GenBank ID:

Contig ID:

Minimum EST Coverage:

Submit

Reset Form

Or download the data from our FTP site:

- [Download TA496 EST assemblies](#)
- [README and tab-delimited text files with TA496 SNP results](#)

SolCAP EST Contigs

Tomato Intervarietal TA496 vs. Heinz1706 SNPs

Column Descriptions:

- Variety-specific EST Contig Name: SolCAP assigned name for an assembly. Click to view sequence in a new window.
- Heinz1706 Genomic GenBank ID: Click to view the GI in GenBank.
- Chr. #: Chromosome number of the genomic sequence based on GenBank record information.
- SNP: reference base/alternate allele.
- Contig Pos: The position number in the contig.
- Genomic Pos: The position number in the genomic sequence.
- Genomic Base: The base of the genomic sequence at the specified position.
- EST Cov: The number of ESTs aligned to the genomic sequence at the specified position.

Downloads:

- README and tab-delimited text file with all TA496 vs. Heinz1706 SNP results
- Information regarding the SolCAP TA496 Assemblies

Results:

If you would like to recapitulate these results, you may take the contig and the genomic sequence and blast them using BLAST2. You can access the contig and genomic sequence by clicking on their respective IDs.

Sort the table by clicking the column headers

Variety-specific EST Contig Name	Heinz1706 Genomic Genbank ID	Chr. #	genomic/ EST	Contig Pos.	Genomic Pos.	Genomic Base	Ori.	#As	#Cs	#Gs	#Ts	EST Cov.
SolCAP_SL_TA496_Contig10047	100159197	11	T/G	341	118977	T	+	0	0	1	2	2
SolCAP_SL_TA496_Contig10680	100159189	11	G/A	303	148279	G	-	2	0	1	0	2
SolCAP_SL_TA496_Contig11163	100159194	11	A/C	403	21813	A	-	1	2	0	0	2
SolCAP_SL_TA496_Contig11549	81295488	11	T/C	313	61854	T	+	0	1	0	2	2
SolCAP_SL_TA496_Contig11741	100159189	11	T/A	643	52164	T	+	1	0	0	2	2

Visit us at <http://solcap.msu.edu/>

SolCAP



Solanaceae Coordinated Agricultural Project

Home

Project Description

Executive Committee

Project Participants

Calendar

Newsletters

Tools

Downloads

Resources

Meetings & Workshops

Extension

Contact

Candidate Genes

The SolCAP project links together people from public institutions, private institutions and industries who are dedicated to the improvement of the Solanaceae crops: potato and tomato. Through innovative research, education and training the SolCAP project will focus on translating genomic advances to US tomato and potato breeding programs. This will lead to significantly improved varieties that benefit the processors, the consumer and the environment.



Tools, →
Downloads

The SolCAP project is supported by the National Research Initiative Plant Genome Program of USDA's Cooperative State Research, Education and Extension Service.



Tools

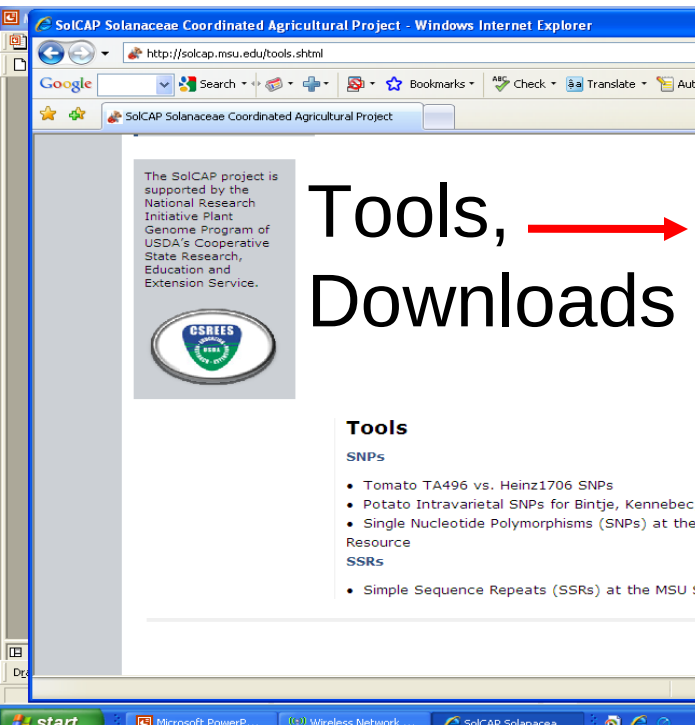
SNPs

- Tomato TA496 vs. Heinz1706 SNPs
- Potato Intravarietal SNPs for Bintje, Kennebec
- Single Nucleotide Polymorphisms (SNPs) at the MSU Solanaceae Genomics Resource

SSRs

- Simple Sequence Repeats (SSRs) at the MSU Solanaceae Genomics Resource

Website maintained by: Kelly Zarka
Web template provided by A. Viklund.



This ends an introduction to on-line databases.

Next, a discussion of downloading “customized” data.

Questions?



This module Introduces the ENTREZ search capability of the NCBI database.

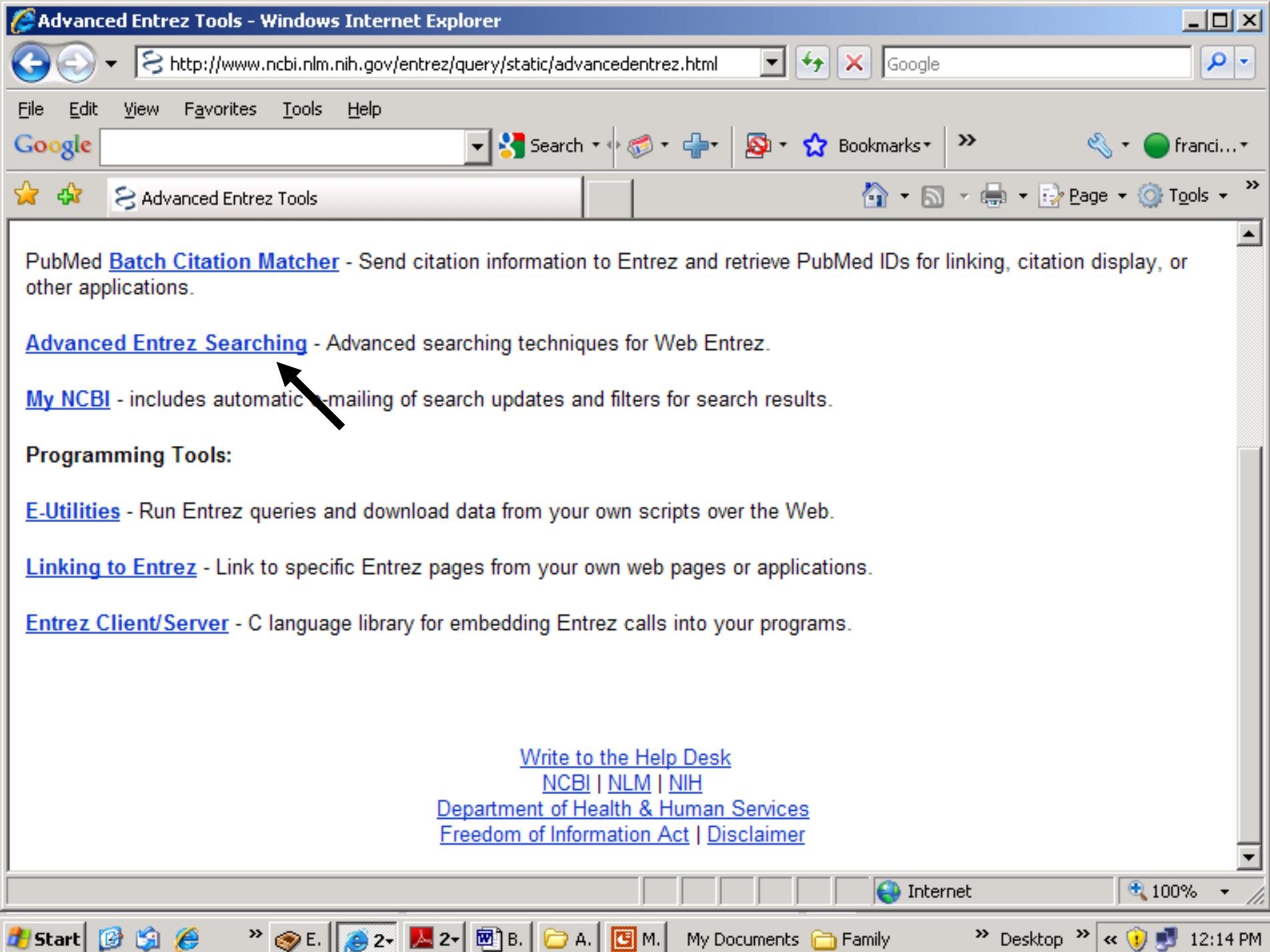
After following this module, you should be able to:

- Describe the different databases within NCBI

- Use ENTREZ to search for specific sequences

- Use ENTREZ advanced search options to refine and improve database searches

- Use ENTREZ to download specific sets of data



Entrez uses the following Boolean Operators:

AND: Entrez will find all documents that contain BOTH terms

OR: Entrez will find all documents that contain EITHER term.

NOT: Entrez will find all documents that contain search term 1 BUT NOT search term 2.

The Entrez search rules and syntax for using Boolean operators are:

1. Boolean operators AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements).
2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first as a unit and then incorporated into the overall strategy. For example, the search statement: g1p3 AND (response element OR promoter) is processed by Entrez by ORing the terms response element OR promoter first and then ANDing the resulting set of documents with g1p3.
3. The Details button shows how Entrez translated and executed your search strategy.
4. See Writing Advanced Search Statements for more information on using Boolean Operators and Entrez Search Field Qualifiers.

The use of parentheses can change your search results significantly. Compare the number of records retrieved in the Nucleotide database as of February 2006 in each case below.

Example

g1p3 AND (response element OR promoter) retrieves three records
g1p3 AND response element OR promoter retrieves 354,554

NCBI Home

NCBI Web Site

File Edit

Google

★ ☆

NCBI Logo

PubM

Search

SITE MAP

Alphabeti

Resource

About NCBI

An introduct

NCBI

GenBar

Sequence

submission

and software

Literature databases

All Databases

NCBI Web Site

PubMed

Protein

Nucleotide

EST

GSS

Structure

Genome

BioSystems

Books

CancerChromosomes

Conserved Domains

3D Domains

Gene

Genome Project

dbGaP

GENSAT

GEO Profiles

GEO Datasets

HomoloGene

Journals

MeSH

NLM Catalog

OMIA

OMIM

PMC

PopSet

Probe

Internet Explorer

http://www.ncbi.nlm.nih.gov/

Google

Search

Bookmarks

franci...

Home

Page

Tools

National Center for Biotechnology Information

[National Library of Medicine](#) [National Institutes of Health](#)

BLAST OMIM Books TaxBrowser Structure

for

What does NCBI do?

Established in 1988 as a national resource for biology information, NCBI creates databases, conducts research in computational biology, develops software for analyzing genome data, and disseminates biomedical information - all for the understanding of molecular and cellular processes affecting human health and disease. [More about NCBI...](#)

Hot Spots

- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools

NLM/NCBI H1N1 Flu Resources

Done

Internet 100%

NCBI HomePage - Windows Internet Explorer

← →

http://www.ncbi.nlm.nih.gov/

↺ ↻

✕

Google

🔍

File Edit View Favorites Tools Help

Google

🔍

Search

📁

+

🔒

★ Bookmarks

»

🔧

franci...

»

★

NCBI HomePage

🏠


📡

🖨

📄 Page

⚙ Tools

»

NCBI

National Center for Biotechnology Information

[National Library of Medicine](#)[National Institutes of Health](#)

PubMedAll DatabasesBLASTOMIMBooksTaxBrowserStructure

Search for

SITE MAP

Alphabetical List

Resource Guide

About NCBI

An introduction to NCBI

GenBank

Sequence submission support and software

Literature databases

▶ What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More about NCBI...](#)

▶ NLM/NCBI H1N1 Flu Resources

Hot Spots

▶ Clusters of orthologous groups

▶ Coffee Break, Genes & Disease, NCBI Handbook

▶ Electronic PCR

▶ Entrez Home

▶ Entrez Tools

Internet

100%

Start

📁 E.

📁 2

📁 2

📁 B.

📁 A.

📁 M.

My Documents

📁 Family

» Desktop

⏪ ⏩ 10:59 AM

lycopersicum [ORGN] AND Rio Grande - EST Results - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuclest&cmd=search&term=ly

Google

File Edit View Favorites Tools Help

Google Search

lycopersicum [ORGN] AND Rio Grande - EST Results

My NCBI [Sign In] [Re]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals

Search EST for lycopersicum [ORGN] AND Rio Grande Go Clear Save Search

Limits Preview/Index History Clipboard Details

Found 22063 nucleotide sequences. Nucleotide [90] EST [21973]

Display Summary Show 20 Sort By Send to

All: 21973 Bacteria: 0 mRNA: 21973

Items 1 - 20 of 21973 Page 1 of 1099 Next

1: GH622898 Reports Links
PCF0411x184 PAMP-elicited tomato leaf Solanum lycopersicum cDNA 5-, mRNA sequence
gi|221062735|gb|GH622898.1|[221062735]

2: GH622897 Reports Links

Recent Activity

Turn Off

lycopersicum[ORGN] AND Ri... (21973)

lycopersicum[ORGN] AND Ri (5)

Done Internet 100%

Start E. 2- 2- B. A. M. My Documents Family Desktop 11:01 AM

Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato - EST Results - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuclest&cmd=search&term=Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato

File Edit View Favorites Tools Help

Google Search

Phytophthora [ORGN] AND Judelson [AUTH] AND To...

My NCBI [Sign In] [Re...

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals

Search EST for Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato Go Clear Save Search

Limits Preview/Index History Clipboard Details

Found 3921 nucleotide sequences. EST [3921]

Display Summary Show 20 Sort By Send to

All: 3921 Bacteria: 0 mRNA: 3921

Items 1 - 20 of 3921 Page 1

Send to
Text
File
Printer
Clipboard

1: CV969339 Reports Links
PI010H11 infected tomato, outside of lesion 3 dpi Phytophthora infestans cDNA, mRNA sequence
gi|58159088|gb|CV969339.1|[58159088]

2: CV060338 Reports Links

Recent Activity
Turn Off
Phytophthora[ORGN] A Ju... (3921)
lycopersicum[ORGN] AND Ri... (21973)

Done Internet 100%

Start E. 2 2 B. A. M. My Documents Family Desktop 12:20 PM

Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato - EST Results - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucst&cmd=search&term=Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato - EST Results

Google

File Edit View Favorites Tools Help

Google Search

Connecting...

NCBI

All Databases PubM

Search EST

Limits Preview/Index

Found 3921 nucleotide sequences

Display Summary

All: 3921 Bacteria: 0

Items 1 - 20 of 3921

1: CV969339 Reports

PI010H11 infected tomato, outside of lesion 3 dpi Phytophthora infestans cDNA, mRNA sequence
gi|58159088|gb|CV969339.1|[58159088]

2: CV060338 Reports

File Download

Do you want to open or save this file?

Name: nucst_result.txt
Type: Text Document
From: www.ncbi.nlm.nih.gov

Open Save Cancel

While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. [What's the risk?](#)

My NCBI

[Sign In] [Re

OMIM PMC Journa

Clear Save Search

Recent Activity

Turn Off

Phytophthora[ORGN] A
Ju... (3921)

lycopersicum[ORGN]
AND Ri... (21973)

Done

Internet 100%

Start

E. 2- 2- B. A. M.

My Documents Family

Desktop

12:23 PM

Phytophthora

Summary

ASN.1

FASTA

XML

GenBank

GI List

TinySeq XML

INSDSeq XML

LinkOut

Assembly

Gene Links

EST Genome Project Links

GENSAT Links

GEO Profile Links

HomoloGene Links

OMIM Links

BioAssay Links

BioAssay by DNA target

BioAssay by RNA Target

PubChem Compound Links

PubChem Substance Links

PMC Links

PopSet Links

Probe Links

Protein Cluster Links

PubMed Links

PubMed (RefSeq) Links

SNP Links

Gene Genotype Links

Structure Links

20:

Items

Display

AND Tomato - EST Results - Windows Internet Explorer

es/entrez?db=nucst&cmd=search&term=Pt

Google

Search

Bookmarks

franci...

h [AUTH] AND To...

58159070]

Links

side of lesion 3 dpi *Phytophthora infestans* cDNA,

58159069]

Page 1 of 197 Next

Show 20 Sort By Send to

Write to the Help Desk

NCBI | NLM | NIH

Department of Health & Human Services

Privacy Statement | Freedom of Information Act | Disclaimer

Done

Internet

100%

NCBI Sequence Viewer v2.0 - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sviewer/viewer.fcgi?tool=portal&db=nucleotide

File Edit View Favorites Tools Help

Google Search

NCBI Sequence Viewer v2.0

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search EST for Phytophthora [ORGN] AND Judelson [AUTH] AND T Go Clear

Display FASTA Show 20 Send to

Item 1 - 20 of 3921 page 1 of 197 Previous Next

1: [CV969339](#). Reports PI010H11 infected...[gi:58159088] Links

Next sequence

>gi|58159088|gb|CV969339.1|CV969339 PI010H11 infected tomato, outside of lesion 3 dpi Phytophthora infestans cDNA, mRNA sequence

CGGGTCGACCCACGCGTCCGATTCATTTAATTGTTTATACACTTAAAATGGATAACGATTGGAACAGGAA
AAAGGTTCTTTCTAGAGTGGATCGTGAAATTTTTCGCTCGAAGCCTGGCGGACATTGAAATCTGTGCCAT
CTTCGGCGTCCGTATTCGTAGCCACCTTCAACTCAAGCTTGTAGATGCCATCAAAGAAGCTGGAAATTAA
ATATCATGAATTTATTCGAAGTAGCCAATATGGAAGAAGATATGTGGTTCAATGTCCCTTCAACACAAAC
CAATTCACATTCAAGATCGGTATTATTCTTAAATTCACGATGCACTAGTCAGATATGAAGAAAGTAAAGA
ATGAAACTTGCGGAAGGGTACTGAGTACAGAAGATCCATCTTATAAAAAAAAAAATATTGATTTGTTAAAA
AGTACATGTGGGCTTGATCCCACTTTAGAATCAGTGTTGGGCTACTTAAGACGGTATGTGACGAAAATAA
TTTTATGTTCTAGCCCTTATTAGTTTAACTAGTGTCTGGTGTCTAAGCTATTTTTTAAAATCATTAGAA

Done Internet 100%

Start My Documents Family Desktop 12:28 PM

Lycopersicum [ORGN] AND Rio Grande
21973

Lycopersicum [ORGN] AND Rio Fuego
171

Lycopersicum [ORGN] AND MicroTom
120462

Lycopersicum [ORGN] AND TA496
116711

Lycopersicum [ORGN] AND Moneymaker
833

Phytophthora [ORGN] AND Judelson [AUTH] AND Tomato
3921

This concludes a review of sequence resources for tomato, and how to download specific sets of data for marker development.

Questions?

