



Bioinformatic Web-Tools: The Basics

sol genomics network

presented by
Suzy Strickler
Mueller Lab
Rm 217

Slides can be found here: [ftp://ftp.solgenomics.net/
bioinfo_class/intro_2014/Bioinfo_interns_6_2014_partII.pdf](ftp://ftp.solgenomics.net/bioinfo_class/intro_2014/Bioinfo_interns_6_2014_partII.pdf)



Part I: Biological Databases.



Source: Contributing Organizations at GMOD

Biological Databases:

1- Types.

2- Public Repositories.

3- Community specific databases.

3.1- For species.

3.2- For specific datatypes.

4- Genomic Browsers.

There is 3 types of biological databases (Rhee SY. *et al.* 2006):

- Public repositories with massive data storage.
- Community-specific databases.
- Project-specific databases.



* Public repositories.

- Maintained by public agencies or public international consortiums.
- Massive data amounts (**quantity**).
- No curated or poorly curated data.
- Long term data storage.
- Examples: GenBank, Uniprot, Gene Expression Omnibus (GEO), ArrayExpress.



* Community-specific databases.

- Maintained by scientific groups, frequently associated with an specific project or a research line.
- Considerable data amount related with the community needs.
- Curated or highly curated data (**quality**).
- Long term data storage

* Community-specific databases.

- Different types:
 - Model Organism Databases (MODs):
TAIR, MaizeGDB
 - Clade Organism Databases (CODs):
Sol Genomics Network (SGN), Gramene.
 - Metabolic Pathways Databases:
MetaCyc, KEGG pathways.
 - Specific Vocabulary (Ontologies) Databases:
Gene Ontology, Plant Ontology.

* Project specific databases.

- Maintained by a group or a small consortium
- Low data amount.
- Variability for data curation (from poorly to highly).
- Limited lifespan generally associated with a project.
- Examples: Plant Genome Network (PGN)



Biological Databases:

1- Types.

2- Public Repositories.

3- Community specific databases.

3.1- For species.

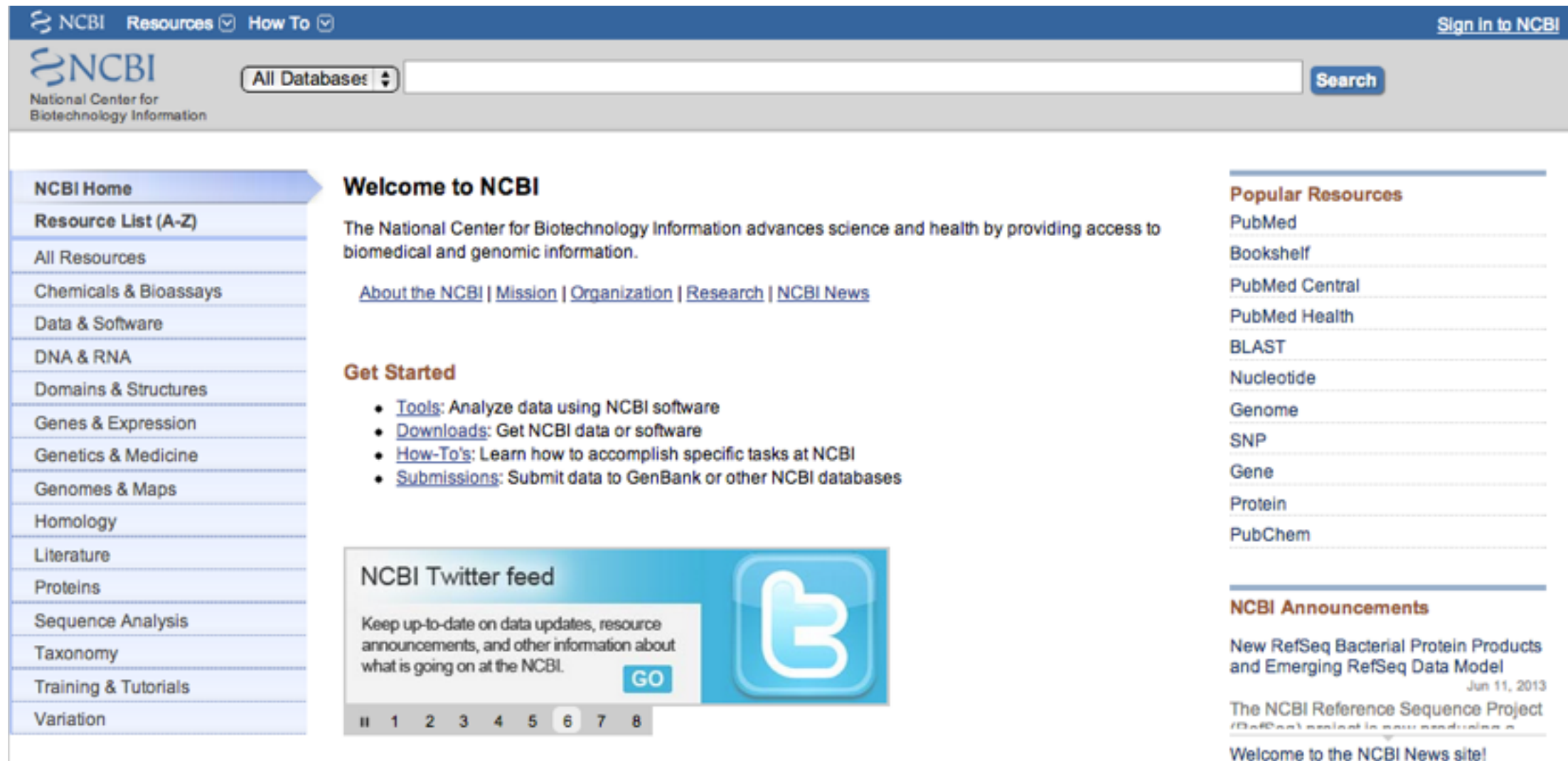
3.2- For specific datatypes.

4- Genomic Browsers.

2. Public Repositories.

NCBI (National Center for Biotechnology Information)

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



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, navigation links for Resources and How To, and a Sign In to NCBI link. Below the header is a search bar with a dropdown menu for 'All Databases' and a 'Search' button. The main content area is divided into three columns. The left column is a vertical menu with links to various resources. The middle column features a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-To's, and Submissions, and a 'NCBI Twitter feed' section. The right column lists 'Popular Resources' and 'NCBI Announcements'.

NCBI Resources How To Sign In to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home

Resource List (A-Z)

All Resources

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.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [NCBI News](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

NCBI Twitter feed

Keep up-to-date on data updates, resource announcements, and other information about what is going on at the NCBI.

GO

1 2 3 4 5 6 7 8

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

New RefSeq Bacterial Protein Products and Emerging RefSeq Data Model

Jun 11, 2013

The NCBI Reference Sequence Project (RefSeq) provides a comprehensive and accurate set of genomic sequences and annotations for the human genome.

Welcome to the NCBI News site!

2. Public Repositories.

NCBI (National Center for Biotechnology Information)

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

Highlights:

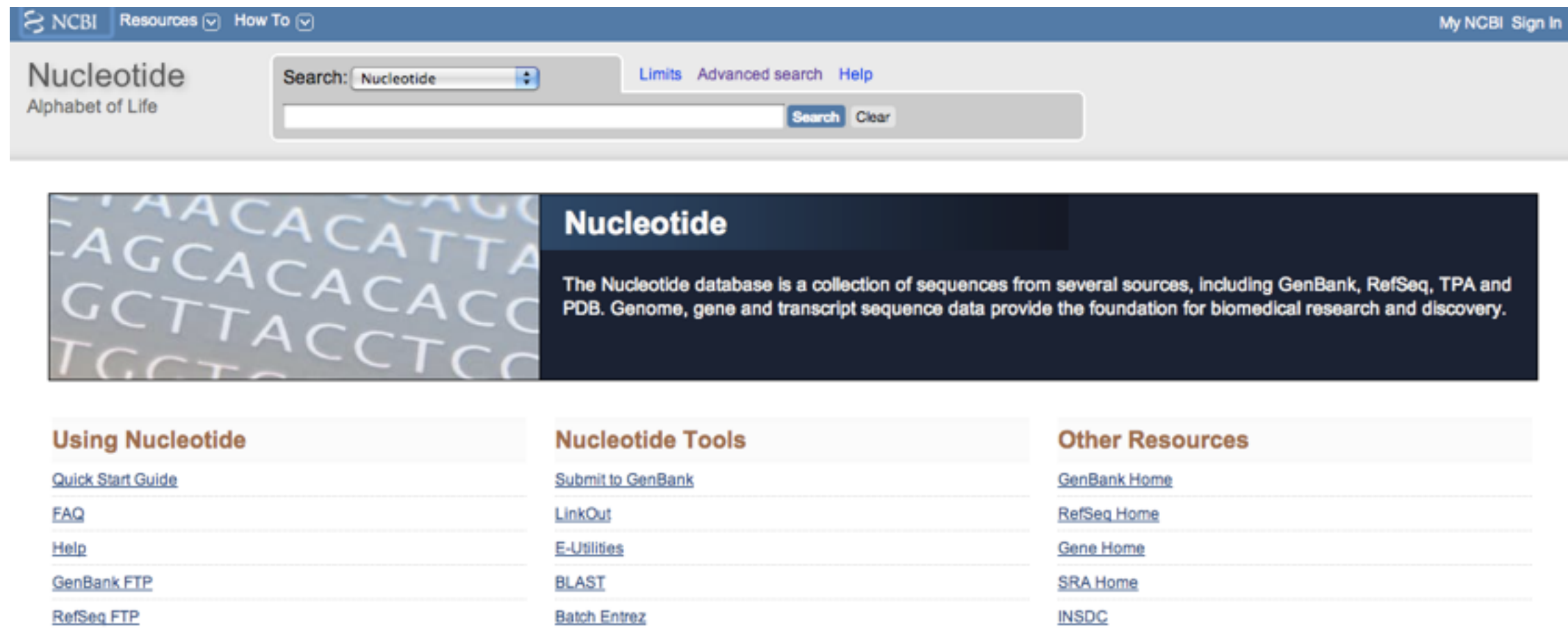
- GenBank.
- PubMed.
- Gene Expression Omnibus (GEO)
- Taxonomy

2. Public Repositories: NCBI

GenBank, NIH database for sequences, an annotated collection of ALL publicly available DNA sequences (Benson DA. *et al.* 2011).

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

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide>



The screenshot shows the NCBI Nucleotide database homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this, the 'Nucleotide' section is highlighted with the tagline 'Alphabet of Life'. A search bar is present with a dropdown menu set to 'Nucleotide' and buttons for 'Limits', 'Advanced search', and 'Help'. Below the search bar, there's a large banner with a background image of DNA sequences and the text: 'The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.' At the bottom, there are three columns of links: 'Using Nucleotide' (Quick Start Guide, FAQ, Help, GenBank FTP, RefSeq FTP), 'Nucleotide Tools' (Submit to GenBank, LinkOut, E-Utilities, BLAST, Batch Entrez), and 'Other Resources' (GenBank Home, RefSeq Home, Gene Home, SRA Home, INSDC).

NCBI Resources How To My NCBI Sign In

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide

- [Quick Start Guide](#)
- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

Nucleotide Tools

- [Submit to GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [BLAST](#)
- [Batch Entrez](#)

Other Resources

- [GenBank Home](#)
- [RefSeq Home](#)
- [Gene Home](#)
- [SRA Home](#)
- [INSDC](#)

2. Public Repositories: NCBI



GenBank:

1 → Search Section

NCBI Resources How To My NCBI Sign In

Nucleotide Alphabet of Life

Search: Nucleotide Save search Limits Advanced search Help

Drought Search Clear

Display Settings: Summary, 20 per page, Sorted by Default order Send to:

Found 770004 nucleotide sequences. Nucleotide (15269) EST (754652) GSS (83)

Results: 1 to 20 of 15269

1. [Arabidopsis thaliana chromosome 1, complete sequence](#)
30,427,671 bp linear DNA
Accession: CP002684.1 GI: 332189094
[GenBank](#) [FASTA](#) [Graphics](#)

2. [Gossypium hirsutum mitogen-activated protein kinase \(MAPK\) gene, promoter region](#)
1,622 bp linear DNA
Accession: HM150999.1 GI: 315258198
[GenBank](#) [FASTA](#) [Graphics](#)

3. [Gossypium hirsutum mitogen-activated protein kinase 16 \(MPK16\) gene, complete cds](#)
6,195 bp linear DNA
Accession: FJ966896.1 GI: 297748124
[GenBank](#) [FASTA](#) [Graphics](#)

Filter your results:

All (15269)
[Bacteria \(198\)](#)
[INSDC \(GenBank\) \(14860\)](#)
[mRNA \(9150\)](#)
[RefSeq \(401\)](#)
[Manage Filters](#)

Top Organisms [Tree]

Populus tremula x Populus alba (7835)
Oryza sativa (2155)
Oryza sativa Indica Group (1366)
Pinus taeda (577)
Oryza sativa Japonica Group (524)
All other taxa (4292)
More...

2 → Similar Results in other sections

3 → Sequence Type Filter

4 → Taxonomic Filter

2. Public Repositories: NCBI



GenBank:

1 → Filter application box

Nucleotide
Alphabet of Life

Search: Save search Limits Advanced search Help

Search Clear

Display Settings: ☐ Summary, 20 per page, Sorted by Default order Send to: ☐

Found 22653 nucleotide sequences. Nucleotide (559) EST (22094)

Results: 1 to 20 of 49 << First < Prev Page 1 of 3 Next > Last >>

- ☐ [Capsicum annuum chitinase class II \(CACHi2\) mRNA, complete cds](#)
1. 1,004 bp linear mRNA
Accession: AF091235.1 GI: 3641354
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- ☐ [Capsicum annuum stellacyanin-like protein CASLP1 precursor, mRNA, complete cds](#)
2. 937 bp linear mRNA
Accession: AF291179.1 GI: 9885805
[GenBank](#) [FASTA](#) [Graphics](#)
- ☐ [Nicotiana attenuata lipid transfer protein 1-like \(LTP1\) mRNA, partial sequence](#)
3. 672 bp linear mRNA
Accession: HM068895.1 GI: 298155395
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- ☐ [Nicotiana attenuata osmotin 1-like \(OSM1\) mRNA, complete sequence](#)
4. 958 bp linear mRNA

Filter your results:

- [All \(559\)](#)
- Bacteria (0)
- [INSDC \(GenBank\) \(559\)](#)
- mRNA (49)**
- RefSeq (0)

[Manage Filters](#)

▼ Taxonomic Groups [List]

- ☐ Solanaceae (49)
 - Nicotiana (27)
 - Capsicum (12)
 - Solanum (10)

Find related data

Database:

2. Public Repositories: NCBI



GenBank:

Tools Links

Nucleotide
Alphabet of Life

Search: Nucleotide

[Limits](#) [Advanced search](#) [Help](#)

[Display Settings:](#) ☒ GenBank [Send:](#) ☒

Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence

GenBank: HM068893.1
[FASTA](#) [Graphics](#)

[Go to:](#) ☒

LOCUS	HM068893	958 bp	mRNA	linear	PLN 28-DEC-2010
DEFINITION	Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence.				
ACCESSION	HM068893				
VERSION	HM068893.1 GI:298155393				
KEYWORDS	.				
SOURCE	Nicotiana attenuata				
ORGANISM	Nicotiana attenuata Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; asterids; lamiids; Solanales; Solanaceae; Nicotianoideae; Nicotianeae; Nicotiana.				
REFERENCE	1 (bases 1 to 958)				
AUTHORS	Re,D.A., Dezar,C.A., Chan,R.L., Baldwin,I.T. and Bonaventure,G.				
TITLE	Nicotiana attenuata NaHD20 plays a role in leaf ABA accumulation during water stress, benzylacetone emission from flowers, and the timing of bolting and flower transitions				
JOURNAL	J. Exp. Bot. 62 (1), 155-166 (2011)				
PUBMED	20713465				
REFERENCE	2 (bases 1 to 958)				
AUTHORS	Bonaventure,G., Re,D. and Baldwin,I.				
TITLE	Analysis of drought and ABA responsive genes in Nicotiana attenuata				
JOURNAL	Unpublished				

Analyze this sequence

- [Run BLAST](#)
- [Pick Primers](#)
- [Find in this Sequence](#)

LinkOut to external resources

[Gramene](#) [Gramene]

All links from this record

- [Full text in PMC](#)
- [PubMed](#)

Recent activity

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

Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence Nucleotide

2. Public Repositories: NCBI



GenBank:

Format

File Storage

Nucleotide
Alphabet of Life

Search: Nucleotide

[Limits](#) [Advanced search](#) [Help](#)

[Display Settings:](#) ☒ GenBank

[Send:](#) ☒

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

LinkOut to external resources

Gramene [Gramene]

All links from this record

Full text in PMC

PubMed

Recent activity

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

Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence Nucleotide

GenBank: HM068893.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS HM068893 958 bp mRNA linear PLN 28-DEC-2010

DEFINITION Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence.

ACCESSION HM068893

VERSION HM068893.1 GI:298155393

KEYWORDS .

SOURCE Nicotiana attenuata

ORGANISM [Nicotiana attenuata](#)

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; asterids; lamiids; Solanales; Solanaceae; Nicotianoideae; Nicotianeae; Nicotiana.

REFERENCE 1 (bases 1 to 958)

AUTHORS Re,D.A., Dezar,C.A., Chan,R.L., Baldwin,I.T. and Bonaventure,G.

TITLE Nicotiana attenuata NaHD20 plays a role in leaf ABA accumulation during water stress, benzylacetone emission from flowers, and the timing of bolting and flower transitions

JOURNAL J. Exp. Bot. 62 (1), 155-166 (2011)

PUBMED [20713465](#)

REFERENCE 2 (bases 1 to 958)

AUTHORS Bonaventure,G., Re,D. and Baldwin,I.

TITLE Analysis of drought and ABA responsive genes in Nicotiana attenuata

JOURNAL Unpublished

2. Public Repositories: NCBI



GenBank:

File Storage

Display Settings: ☒ Summary, 20 per page, Sorted by Default order

Found 21339 nucleotide sequences. Nucleotide (515) EST (20824)

Results: 10

- ☐ [Solanum lycopersicum mRNA for SIGRX1 protein, cultivar Hongbaoshi](#)
1. 1,003 bp linear mRNA
Accession: FN646220.1 GI: 308233000
[GenBank](#) [FASTA](#) [Graphics](#)
- ☐ [Lycopersicon esculentum ethylene-responsive late embryogenesis-like protein \(ER5\) mRNA, complete cds](#)
2. 748 bp linear mRNA
Accession: U77719.1 GI: 1684829
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- ☐ [Lycopersicon chilense proline-rich protein \(PRP13\) gene, complete cds](#)
3. 552 bp linear mRNA
Accession: U19098.1 GI: 1001952
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- ☐ [Lycopersicon esculentum non specific lipid transfer protein \(le16\) mRNA, complete cds](#)
4. 583 bp linear mRNA
Accession: U81996.1 GI: 1816534
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Send to: ☒ **Filter your results:**

Choose Destination

- ☒ File
- ☐ Clipboard
- ☐ Collections
- ☐ Analysis Tool

Download 10 items.

Format

- ☒ Summary
- ☐ GenBank
- ☐ GenBank (full)
- ☐ FASTA
- ☐ ASN.1
- ☐ XML
- ☐ INSDSeq XML
- ☐ TinySeq XML
- ☐ Feature Table
- ☐ Accession List
- ☐ GI List

[Manage Filters](#)

Taxonomic Groups [List]

- Solanum (20)

Analyze these sequences

[Run BLAST](#)

Find related data

Database:

[Find items](#)

2. Public Repositories: NCBI



PubMed, NIH database for scientific literature and publications.

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

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed
drought tomato

RSS Save search Limits Advanced search Help

Search Clear

Display Settings: ☒ Summary, 20 per page, Sorted by Recently Added

Send to: ☐

Results: 1 to 20 of 117

<< First < Prev Page 1 of 6 Next > Last >>

- ☐ [An insertional mutagenesis programme with an enhancer trap for the identification and tagging of genes involved in abiotic stress tolerance in the tomato wild-related species *Solanum pennellii*.](#)
Atarés A, Moyano E, Morales B, Schleicher P, García-Abellán JO, Antón T, García-Sogo B, Perez-Martin F, Lozano R, Flores FB, Moreno V, Del Carmen Bolarin M, Pineda B.
Plant Cell Rep. 2011 Jun 7. [Epub ahead of print]
PMID: 21647638 [PubMed - as supplied by publisher]
[Related citations](#)
- ☐ [Identification and expression pattern of one stress-responsive NAC gene from *Solanum lycopersicum*.](#)
Han Q, Zhang J, Li H, Luo Z, Ziaf K, Ouyang B, Wang T, Ye Z.
Mol Biol Rep. 2011 Jun 3. [Epub ahead of print]
PMID: 21637957 [PubMed - as supplied by publisher]
[Related citations](#)
- ☐ [Atypical epigenetic mark in an atypical location: cytosine methylation at asymmetric \(CNN\) sites within the body of a non-repetitive tomato gene.](#)
Gonzalez RM, Ricardi MM, Iusem ND.
BMC Plant Biol. 2011 May 20;11(1):94. [Epub ahead of print]
PMID: 21599976 [PubMed - as supplied by publisher] **Free Article**
[Related citations](#)

2. Public Repositories: NCBI

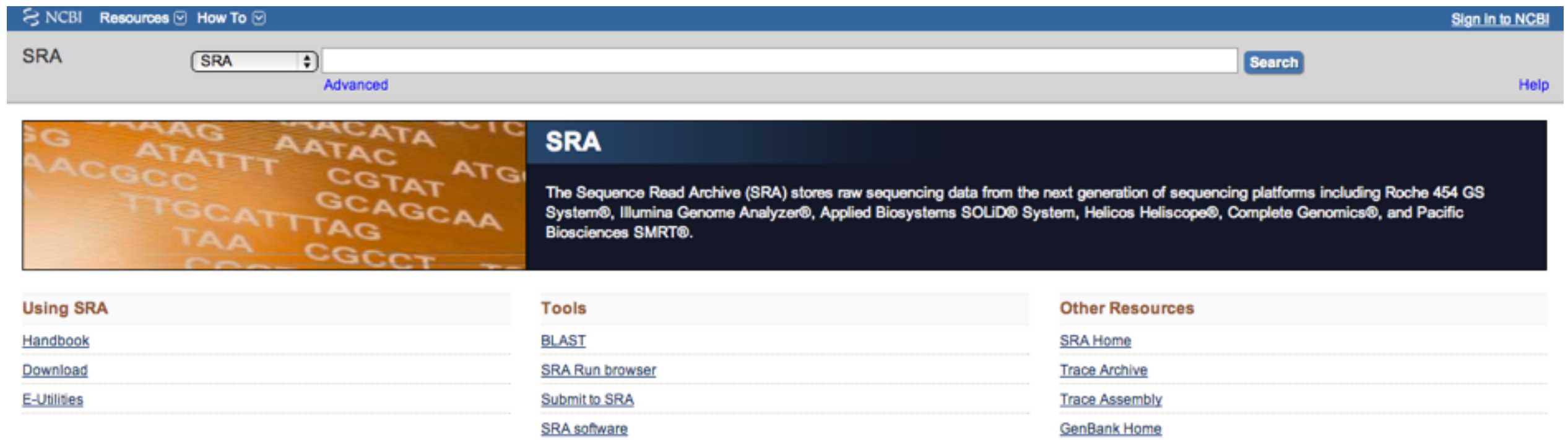
PubMed:

- Relatively updated (Gap between publication and loading in PubMed database around 1-2 days).
- It doesn't have all plant science related journals (for example: Theoretical Applied and Genetics or Crop Science).
(More information: <http://wwwcf.nlm.nih.gov/serials/journals/index.cfm>)
- There are no links between articles and genes, sequences, expression or other information contained in the publication.

2. Public Repositories: NCBI

Sequence Read Archive (SRA), Database to store sequences produced by NGS such as Illumina, 454, Solid, Helicos...

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



The screenshot shows the NCBI SRA website interface. At the top, there is a navigation bar with "NCBI", "Resources", and "How To" links, along with a "Sign in to NCBI" button. Below this is a search bar with "SRA" selected in the dropdown menu and a "Search" button. A "Help" link is also present. The main content area features a large banner with a DNA sequence background and the text: "SRA The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®." Below the banner, there are three columns of links: "Using SRA" (Handbook, Download, E-Utilities), "Tools" (BLAST, SRA Run browser, Submit to SRA, SRA software), and "Other Resources" (SRA Home, Trace Archive, Trace Assembly, GenBank Home).

NCBI Resources How To Sign in to NCBI

SRA SRA Search Help

SRA

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Using SRA

- [Handbook](#)
- [Download](#)
- [E-Utilities](#)

Tools

- [BLAST](#)
- [SRA Run browser](#)
- [Submit to SRA](#)
- [SRA software](#)

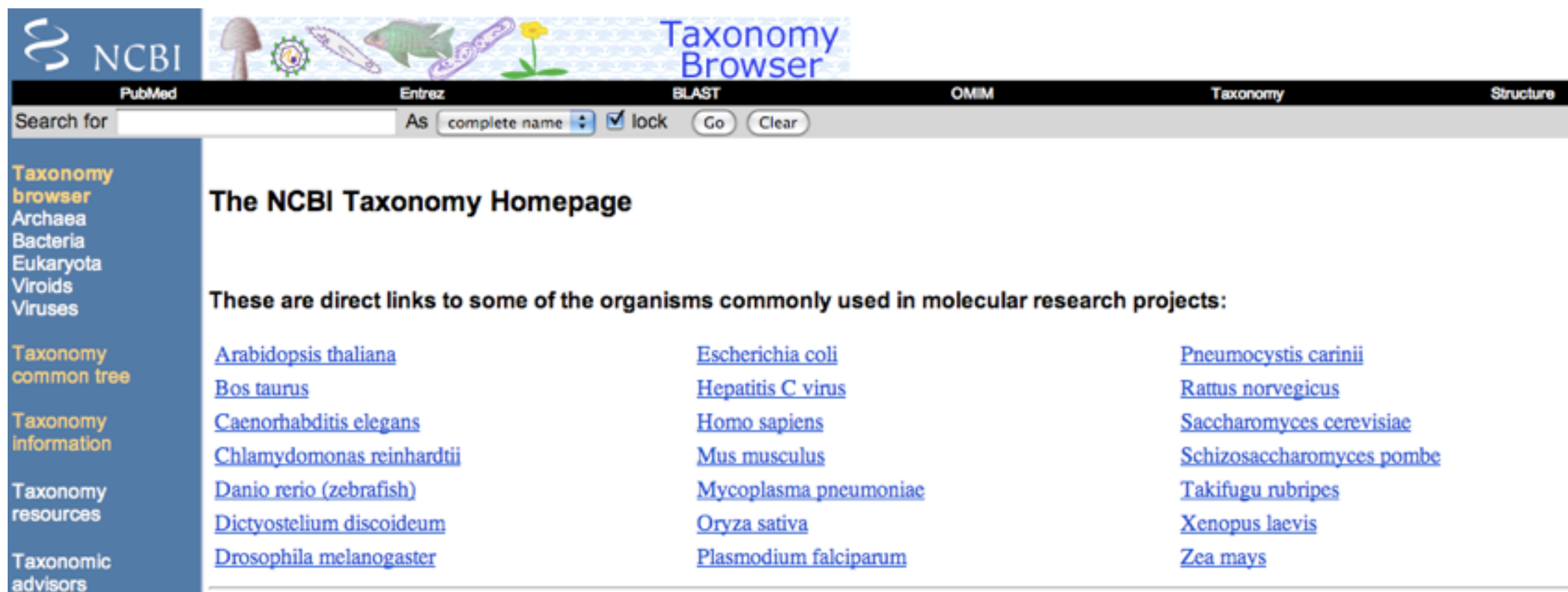
Other Resources

- [SRA Home](#)
- [Trace Archive](#)
- [Trace Assembly](#)
- [GenBank Home](#)

2. Public Repositories: NCBI

Taxonomy, database with taxonomic names and classifications for NCBI organisms.

<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>



The screenshot shows the NCBI Taxonomy Browser homepage. At the top, there is a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. Below this is a search bar with a dropdown menu set to 'AS' and a 'lock' checkbox. The main content area is titled 'The NCBI Taxonomy Homepage' and features a list of direct links to commonly used organisms in molecular research projects. On the left side, there is a sidebar with links to 'Taxonomy browser', 'Taxonomy common tree', 'Taxonomy information', 'Taxonomy resources', and 'Taxonomic advisors'.

NCBI Taxonomy Browser

Search for AS complete name ☒ lock

The NCBI Taxonomy Homepage

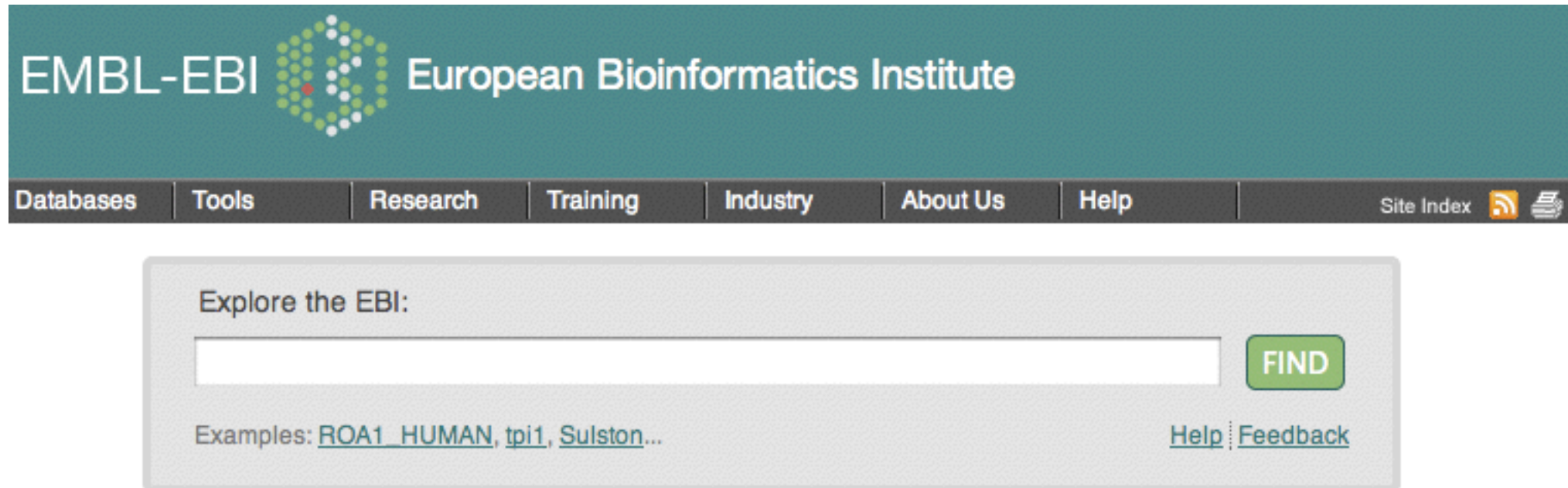
These are direct links to some of the organisms commonly used in molecular research projects:

Arabidopsis thaliana	Escherichia coli	Pneumocystis carinii
Bos taurus	Hepatitis C virus	Rattus norvegicus
Caenorhabditis elegans	Homo sapiens	Saccharomyces cerevisiae
Chlamydomonas reinhardtii	Mus musculus	Schizosaccharomyces pombe
Danio rerio (zebrafish)	Mycoplasma pneumoniae	Takifugu rubripes
Dictyostelium discoideum	Oryza sativa	Xenopus laevis
Drosophila melanogaster	Plasmodium falciparum	Zea mays

2. Public Repositories: EBI

EBI (European Bioinformatics Institute)

<http://www.ebi.ac.uk/>



EMBL-EBI  European Bioinformatics Institute

Databases Tools Research Training Industry About Us Help Site Index  

Explore the EBI:

FIND

Examples: [ROA1_HUMAN](#), [tpi1](#), [Sulston...](#) [Help](#) | [Feedback](#)

Data Resoures and Tools

- [ENA](#)
- [UniProt](#)
- [ArrayExpress](#)
- [Ensembl](#)
- [InterPro](#)
- [PDBe](#)
- [Genomes](#)
- [Nucleotide Sequences](#)
- [Protein Sequences](#)
- [Macromolecular Structures](#)
- [Small Molecules](#)
- [Gene Expression](#)
- [Molecular Interactions](#)
- [Reactions& Pathways](#)
- [Protein Families](#)
- [Enzymes](#)
- [Literature](#)
- [Taxonomy](#)
- [Ontologies](#)
- [Patent Resources](#)
- [Sequence Similarity & Analysis](#)
- [Pattern & Motif Searches](#)
- [Structure Analysis](#)
- [Text Mining](#)
- [Downloads](#)
- [Web Services](#)

2. Public Repositories: EBI

EBI (European Bioinformatics Institute)

<http://www.ebi.ac.uk/>

Highlights:

- ENA (European Nucleotide Archive).
- UniProt
- ArrayExpress
- Ensembl
- InterPro

2. Public Repositories: EBI

InterPro, protein domain database organized by superfamilies, families and subfamilies. It is frequently used for genome functional annotation, specially to link genes with gene ontologies associated with protein domains. (<http://www.ebi.ac.uk/interpro/>).

EBI > Databases > InterPro

InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: 32.0 18th April 2011 (see [Release Notes](#) for further details)

Search InterPro:

Do a sequence search of InterPro, via [InterProScan](#)

Extract large datasets by querying our [BioMart](#)

You can access our data programmatically, via [Web Services](#)

Use the updated [InterProScan Web Service](#)

If you have any questions or feedback please [contact us](#).



InterPro 32.0

Biological Databases:

1- Types.

2- Public Repositories.

3- Community specific databases.

3.1- For species.

3.2- For specific datatypes.

4- Genomic Browsers.

3. Community specific databases



Name	Species	Data	Link
The Arabidopsis Information Resource (TAIR)	<i>Arabidopsis</i>	Single Species Genomes, Genetic Markers, SNPs, Genes, Expression, Proteins, Ontologies, Metabolic Pathways, Publications	http:// www.arabidopsis.org/
Gramene	Monocots (Grape and Arabidopsis)*	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Metabolic Pathways, QTLs	http:// www.gramene.org/
Sol Genomics Network (SGN)	Solanaceae, Rubiaceae	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Expression*, Proteins, Ontologies, Metabolic Pathways, Publications, QTLs and Maps, Phenotypes	http://solgenomics.net/
Genome Database for Rosaceae (GDR)	Rosaceae	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Phenotypes, Unigenes	http:// www.rosaceae.org/
Dendrome	Trees	Multiple Species Genomes, Genes, SNPs, Genetic Markers, Genetic Maps, Phenotypes, Literature, Expression	http:// dendrome.ucdavis.edu/
Curcubit Genomics Database (ICuGI)	Cucurbitaceae	Multiple Sequence Genomes, Genes, Unigenes, Genetic Markers, Genetic Maps, Metabolic Pathways	http://www.icugi.org/

3. Community specific databases

Name	Species	Data	Link
The Compositae Genome Project	Compositaceae	Multiple Species Genomes, Genetic Markers, Genetic Maps, Genes	http://compgenomics.ucdavis.edu/
MaizeGDB	Zea mays (Maize)	Single Species Genomes, Genetic Markers, SNPs, Genes, QTLs, Cytogenetics, Phenotypes	http://www.maizegdb.org/
CottonDB	Gossypium spp.	Genetic Markers, Maps, Genes, Taxonomic. Data	http://cottondb.org/
SoyBase	Glycine Max (Soybean)	Multiple Species Genomes, Genetic Markers, SNPs, Genes, Proteins, Ontologies, Metabolic Pathways, QTLs	http://soybase.org/
Phytozome	Plants	Multiple Species Genomes	http://www.phytozome.net
Plant Genome Database (PlantGDB)	Plants	Multiple Species Genomes, Genes, Unigenes	http://www.plantgdb.org/

Biological Databases:

1- Types.

2- Public Repositories.

3- Community specific databases.

3.1- For species.

3.2- For specific datatypes.

4- Genomic Browsers.

3. Community specific databases

There are other community driven databases focused in a knowledge area:

Metabolic databases:

MetaCyc: <http://metacyc.org/>

KEGG: <http://www.genome.jp/kegg/>

Ontology databases:

Gene Ontology: <http://www.geneontology.org/>

Plant Ontology: <http://www.plantontology.org/>

Transcription Factors database:

TranscriptionFactorDB (DBD): www.transcriptionfactor.org

Biological Databases:

1- Types.

2- Public Repositories.

3- Community specific databases.

3.1- For species.

3.2- For specific datatypes.

4- Genomic Browsers.

4. Genomic Browsers

A **Genome Browser** is a graphical interface that shows aligned genomic data.

Each data type is in a **track**.

The tracks are hierarchically organized by track size. For example, the first track could be a *chromosome*, the second one a *region* and the third one, a *detailed region* with gene structures.



4. Genomic Browsers

Genome Browser most used:

- GBrowse (GMOD).
- UCSC Genome Browser.
- Ensembl Genome Browser.
- Vista Genome Browser.
- JBrowse (GMOD).



4. Genomic Browsers

Genome Browser most used:

- GBrowse (GMOD).
- UCSC Genome Browser.
- Ensembl Genome Browser.
- Vista Genome Browser.
- JBrowse (GMOD).



GBrowse (Generic Genome Browser):

It is one of the most popular tools from the GMOD project, used by databases such as FlyBase, Wormbase, SGN, Gramene or GDR.

It is a combination of database and interactive web pages designed for show and manage genome annotations.

It has three levels: *Overview*, *Region* and *Detail*. Each level has different **tracks** to show the an specific information type such as gene models or markers.



4. Genomic Browsers

GBrowse (Generic Genome Browser):

Highlights:

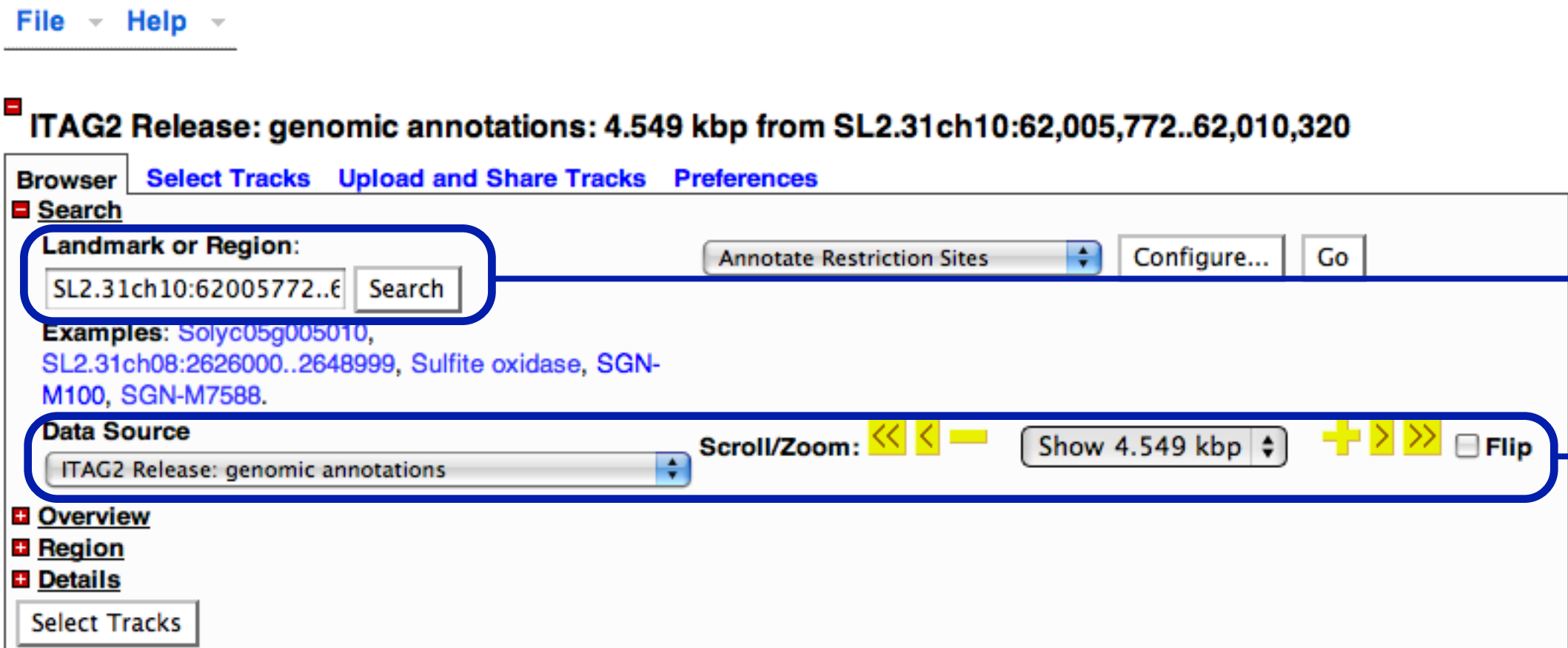
- Simultaneous bird's eye and detailed views of the genome.
- Scroll, zoom, center.
- Use a variety of premade glyphs or create your own.
- Attach arbitrary URLs to any annotation.
- Order and appearance of tracks are customizable by administrator and end-user.
- Search by annotation ID, name, or comment.
- Supports third party annotation using GFF formats.
- Settings persist across sessions.
- DNA and GFF dumps.
- Connectivity to different databases, including BioSQL and Chado.
- Multi-language support.
- Third-party feature loading.
- Customizable plug-in architecture

4. Genomic Browsers

Example: GBrowse at SGN.

2

1



File ▾ Help ▾

ITAG2 Release: genomic annotations: 4.549 kbp from SL2.31ch10:62,005,772..62,010,320

Browser Select Tracks Upload and Share Tracks Preferences

Search

Landmark or Region: SL2.31ch10:62005772..62010320 Search

Annotate Restriction Sites Configure... Go

Examples: Solyc05g005010, SL2.31ch08:2626000..2648999, Sulfite oxidase, SGN-M100, SGN-M7588.

Data Source ITAG2 Release: genomic annotations

Scroll/Zoom: << < - + > >> Show 4.549 kbp Flip

Overview Region Details

Select Tracks

Region

Source

Region:

Reference Sequence + “:” + Initial Coordinate + “..” + Final Coordinate.

Example:

SL2.40ch08:56810100..56812450

4. Genomic Browsers



Example: GBrowse at SGN.

File ▾ Help ▾

ITAG2.3 Release: genomic annotations: 2.351 kbp from SL2.40ch08:56,810,100..56,812,450

Browser [Select Tracks](#) [Upload and Share Tracks](#) [Preferences](#)

Search

Landmark or Region:

Examples: [Solyc05g005010](#), [SL2.40ch08:2626000..2648999](#), [Sulfite oxidase](#), [SGN-M100](#), [SGN-M7588](#).

Data Source

4 Scroll/Zoom: << < — + > >> ☐ Flip

1 Overview

SL2.40ch08

0M 10M 20M 30M 40M 50M 60M

2 Region

56800k 56900k

3 Details

SL2.40ch08: 2.351 kbp

500 bp

56811k 56812k

[Clear highlighting](#)

Three levels: *Overview*, *Region* and *Details*

4. Genomic Browsers



Example: GBrowse at SGN.

File ▾ Help ▾

ITAG2.3 Release: genomic annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274

Browser [Select Tracks](#) [Upload and Share Tracks](#) [Preferences](#)

Search

Landmark or Region:

Examples: [Solyc05g005010](#), [SL2.40ch08:2626000..2648999](#), [Sulfite oxidase](#), [SGN-M100](#), [SGN-M7588](#).

Data Source:

Zoom for Details

4 Scroll/Zoom: << < — + > >> ☐ Flip

Overview

SL2.40ch08

0M 10M 20M 30M 40M 50M 60M

Region

56800k 56900k

5 Details

SL2.40ch08: 10 kbp

56807k 56808k 56809k 56810k 56811k 56812k 56813k 56814k 56815k 56816k

[Clear highlighting](#)

Scroll Downstream and Upstream for Details region

4. Genomic Browsers



Example: GBrowse at SGN.

[File](#) ▾ [Help](#) ▾

■ **ITAG2.3 Release: genomic annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274**

Browser [Select Tracks](#) [Upload and Share Tracks](#) [Preferences](#)

■ **Search**

Landmark or Region:

Examples: [Solyc05g005010](#), [SL2.40ch08:2626000..2648999](#), [Sulfite oxidase](#), [SGN-M100](#), [SGN-M7588](#).

Data Source

Scroll/Zoom: << < — + > >> ☐ Flip

■ **Overview**
SL2.40ch08
0M 10M 20M 30M 40M 50M 60M

■ **Region**
56800k 56900k

■ **Details**
SL2.40ch08: 10 kbp
56807k 56808k 56809k 56810k 56811k 56812k 56813k 56814k 56815k 56816k

6

Tracks selection

4. Genomic Browsers



Example: GBrowse at SGN.

Browser Select Tracks Upload and Share Tracks Preferences

<< Back to Browser

Tracks

7 ☒ **Overview** ☐ All on ☐ All off

☐ Infernal ☐ Markers ☐ tRNAscanSE

8 ☒ **Region** ☐ All on ☐ All off

☐ GC content ☐ Infernal

☐ Genes ☐ Markers

9 ☒ **DNA** ☐ All on ☐ All off

☐ DNA/GC content ☐ 6-frame translation

☒ **Gene models** ☐ All on ☐ All off

☐ CDS - click to browse protein ☐ Gene models ☐ KD plot/peptides

☒ **Genetic loci** ☐ All on ☐ All off

☐ SGN Locus Sequences ☐ SGN markers

☒ **Genome data and reagents** ☐ All on ☐ All off

☐ SL2.40 assembly ☐ SL1.00 / ITAG1 scaffolds ☐ Tomato BACs

☐ ESTs and cDNAs - Other Solanaceae ☐ MicroTom full-length cDNAs

☐ ESTs and cDNAs - Tomato ☐ SGN unigenes

☒ **Other genomes** ☐ All on ☐ All off

☐ Potato BACs ☐ Tobacco genome contigs

☒ **Paired clone-end mapping** ☐ All on ☐ All off

☐ All clones ☐ All clones (linking)

Tracks selection

4. Genomic Browsers



Example: GBrowse at SGN.

File Help

ITAG2.3 Release: genomic annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274

Browser Select Tracks Upload and Share Tracks Preferences

Search

Landmark or Region:

SL2.40ch08:56806275..5 Search

Annotate Restriction Sites

Configure...

Go

Examples: Solyc05g005010, SL2.40ch08:2626000..2648999, Sulfite oxidase, SGN-M100, SGN-M7588.

Data Source

ITAG2.3 Release: genomic annotations

Scroll/Zoom:



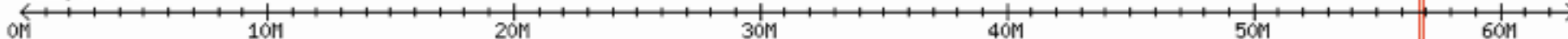
Show 10 kbp



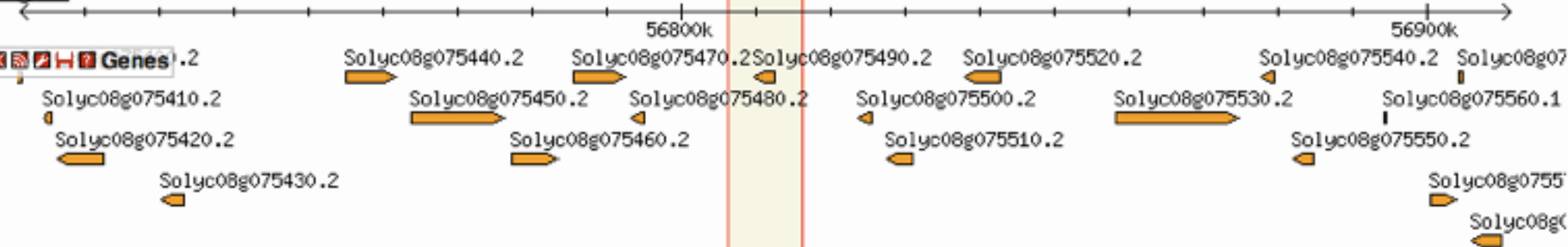
Flip

Overview

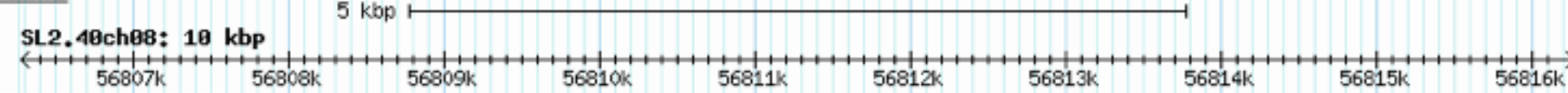
SL2.40ch08



Region



Details



Gene models

Solyc08g075490.2.1
Carotenoid cleavage dioxygenase 48

ESTs and cDNAs - Other Solanaceae

Select Tracks

Clear highlighting

4. Genomic Browsers



Example: GBrowse at SGN.

10

File Help

Bookmark this

Share these tracks

Export as...

Get chrom sizes

Reset to defaults

Examples: Solyc05g005010, SL oxidase, SGN-M100, SGN-M756

Data Source

ITAG2.3 Release: genomic annotations

Annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274

Annotate Restriction Sites

Configure...

Go

Scroll/Zoom: << < - Show 10 kbp + > >> Flip

Overview

SL2.40ch08

Markers

Region

Markers

Details

56800k

C2_At4g19170

SSR41

56900k

Data download in different formats

4. Genomic Browsers



Example: GBrowse at SGN.

10

File **Help**

Bookmark this
Share these tracks
Export as...
Get chrom sizes
Reset to defaults

examples: [Solyc05g005010](#), [SL2.40ch08:56,806,275..56,816,274](#)
oxidase, SGN-M100, SGN-M756

Data Source
ITAG2.3 Release: genomic annotations

Overview
SL2.40ch08
0M 10M 20M 56800k

Region
56800k

Details

annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274

...low-res PNG image
...editable SVG image
...GFF annotation table
..FASTA sequence file

Markers_region+Genes_region+mrna+sgn_unigenes+sgn_loci+Markers_over...

```
SL2.40ch08:56810100..56812450
TTATAGCTTCATAAGATCATTTTCACTCACGAAAAGTCCATGGAACCGTAAGGCACACG
ACGAGGCAATTTGACATTAGCCACAATGTCAAGATTAGGTGATGTTGCATCCATGACCAA
AAAATTTGATTCOCCTGTCTTCTCATTGTGTACGTATGACACCAGTAGCCATCGTCTTC
ATCAATCGAAGGATTTTAGGCACGAAAAATGGTTCACCACCATAGCAATCTTTTCCAAA
TATACGACATGCCAGATGCAATCGCGACGATCAATTTCTGCTACGGATACGTCTAATTT
TGCTATCCCTATTACTTTTGGCATAGGACCCCAATGGCTGCATATACGTACCTACACGA
AACAAAAAATTATTTTATATGTCTGTAAATTTTGGATTGACACGAAATTTAAATGATT
TTTTGTTGTATGGTCTCTGCTGAAATTAATCATGTGTTTATAAGTCATAAAGGACCTA
CACATGTCCATATGCTGCACTATAGATGGTCCAAACAAAAATCAATAAATTTATAATAGTG
TTATATACTAGTACTAGTACTGTCAATGTTTATTTATTTGGTGACACATTAATAAATGTTG
ATAGAATAAAGTTATAATTTTATTATATTATTTTATTTAAAAATTTGGTATATAAAAAAT
GCAAAATTTATTTTGGATTTTTCAAATCAAAGAGTAAAAATAGATACATATTTTTCAGTT
TAGTGGATTAGTAAAAATTAATGAGGAGAGTGAATTTGTTCTATAAAAGTACCGAGCTA
CTCCTTTGCTCAAATTAATTGCCATAATTTGTTTTCGAGATTGATTTAACTAATTTTA
CGATTAATATGTATGTATACCTATATTAATGAATATCAAGTAAGTTACTGAAGATAAAT
GGATATTTAAAAACGTAATTTGTTCTTTTCCAACTAAGCTGGATTGATGACTCCAAAA
TCAAGATTTCTGTAGAAATTTGGATGCTGCTCACCATTCTGTCTTCAAATTTATTTTC
ACTTTCTCAACACTTGATGTATCATGTCCATCTCTAGTGTATGTTCCACCGATAAT
ATATTCGGCGCATCAACACTATCGTATCACCTCCATCCCTCATCCACCGGTTTATCGCG
TGTAACAATATAAATCCTGATACATCGAACCACTCATTTCCGACTCATCTTTGGCGTAA
CGTGGTATCACACCAATTCGGGAATTTTCCCGAGTCAAGTACCCACCGGTGAACACCG
GTGAGTAAATCAATTGGATTCAATTTGATATGTCGAAAAATATCGCGATTTTTTTT
GTGATTGCGAAGTCGTGAAGAAATGAGGGGCGGGTCATTGAGAATATGGGTACATCTGG
GTTTTGTACCATCCGGTTGACCGGAAAGTAAGTTATAACCGAGGTATCGGGCGGTAA
CGGAAAGCAAAAGCCTCGTTAGTTTCGTGGTCGATTTTCGGATGTGCCGTCATGCTCATA
GAAAGCTTTCGGTTAAATCGTGACGGCGGAGGGTAAAAATGTCACCATTTGGGGCTATT
TTACCTCATATGGTAAATCAGATTACCGAGGGCGAAAAAGTTTACCTCCAAATAAAGCT
AAGCTTGTATTCGCAAGGCTATACCGTTTGGGGATTAACTGTCCCGCAATAGCTCGG
GCCGCGGTGAGTGCACACGCGCGCGGAGGCGGTAGACCGTTGAAACCGGAGAACACA
TTAGGAATAATGGGTGACCGGCTTCATTTTCGATGTTGTACTTGAAGTTTGGACAAAT
CGACTACATAACGTAGCTTTACCTTGGGAAATTTAATGGAGTGAAGCATTCATCACCG
TCGAATAGATGGTAAGGTCCACGTGGAAGATTTGAGGGTTAGGTCCATTTCTAATATAC
GCGCATCTAGGCAAGGTGGAAGGAGCCTTCCACTACTTCGATTTCGGTAGGAGGAAGT
TCGTCTACTGGAGCGAAGTTGTTAGAGAGAACATACCTTGGATCAACAGACGATTTCTTA
GGTGGATCAATAAAGTATTTACGAAATCATCGAATGCATTGAAGATAAAGAGGGAAAT
GATGGTTCAATAGGTTTCTTGGGGTATTGGTTCTCTTTTGGAGATGGTTTGGTGGT
GAAGGGGTTGATTTTCTTGGTGGTTTGTATAGTTGGTGGTTTGTGGTCTTTCTTCA
ATTCTAACAGAAAAATCTTTTAGAGTAGGAGAAATAGTAATGATAATTATTATTATTA
TTATAAGGAGAAAGAAGAGATTTAGGGTTTGTGATAATGTAGAAAGGAAAGTTGAAGAC
```

4. Genomic Browsers



Example: GBrowse at SGN.

10

File

Help

Bookmark this

Share these tracks

Export as...

Get chrom sizes

Reset to defaults

Examples: Solyc05g005010, SL2.40ch08, oxidase, SGN-M100, SGN-M756

Data Source

ITAG2.3 Release: genomic annotat

Overview

SL2.40ch08

Markers

Region

Markers

Details

annotations: 10 kbp from SL2.40ch08:56,806,275..56,816,274

...low-res PNG image

...editable SVG image

...GFF annotation table

...FASTA sequence file

Annotate Restriction Sites

Configure...

Go

```
Markers_region+Genes_region+mrna+sgn_unigenes+sgn_loci+Markers_overview+Markers_region+sgn_markers_SL2.40ch08-56810100..56812450.gff3
##gff-version 3
##date Sun Jun 12 18:54:23 2011
##source gbrowse gbff gff3 dumper
##genome-build unknown
##sequence-region SL2.40ch08:56810100..56812450
SL2.40ch08 ITAG_eugene gene 56809898 56812517 . - . Name=Solyc08g075490.2;ID=2441230;Alias=Solyc08g075490;from_BOGAS=1;length=2620
SL2.40ch08 ITAG_eugene mRNA 56809898 56812517 . - . Name=Solyc08g075490.2.1;Parent=2441230;ID=2441231;from_BOGAS=1;sifter_term=GO:
0005515;Note=Carotenoid 20c leavage 20c dioxygenase 20c 40;length=2620;ontology_term=GO:0005515;nb_exon=2
SL2.40ch08 ITAG_eugene exon 56809898 56810451 . - . Parent=2441231;ID=2441232;from_BOGAS=1
SL2.40ch08 ITAG_eugene three_prime_UTR 56809898 56810099 . - . Parent=2441231;ID=2441233;from_BOGAS=1
SL2.40ch08 ITAG_eugene CDS 56810100 56810451 . - 1 Parent=2441231;ID=2441234;from_BOGAS=1
SL2.40ch08 ITAG_eugene intron 56810452 56811017 . - . Parent=2441231;ID=2441235;from_BOGAS=1
SL2.40ch08 ITAG_eugene exon 56811018 56812517 . - 1 Parent=2441231;ID=2441236;from_BOGAS=1
SL2.40ch08 ITAG_eugene CDS 56811018 56812450 . - 0 Parent=2441231;ID=2441237;from_BOGAS=1
SL2.40ch08 ITAG_eugene five_prime_UTR 56812451 56812517 . - . Parent=2441231;ID=2441238;from_BOGAS=1
SL2.40ch08 ITAG_eugene mRNA 56809898 56812517 . - . Name=Solyc08g075490.2.1;ID=2441231;from_BOGAS=1;sifter_term=GO:0005515;Note=Carotenoid 20c leavage
20c dioxygenase 20c 40;length=2620;ontology_term=GO:0005515;nb_exon=2
SL2.40ch08 ITAG_eugene exon 56809898 56810451 . - . Parent=2441231;ID=2441232;from_BOGAS=1
SL2.40ch08 ITAG_eugene three_prime_UTR 56809898 56810099 . - . Parent=2441231;ID=2441233;from_BOGAS=1
SL2.40ch08 ITAG_eugene CDS 56810100 56810451 . - 1 Parent=2441231;ID=2441234;from_BOGAS=1
SL2.40ch08 ITAG_eugene intron 56810452 56811017 . - . Parent=2441231;ID=2441235;from_BOGAS=1
SL2.40ch08 ITAG_eugene exon 56811018 56812517 . - 1 Parent=2441231;ID=2441236;from_BOGAS=1
SL2.40ch08 ITAG_eugene CDS 56811018 56812450 . - 0 Parent=2441231;ID=2441237;from_BOGAS=1
SL2.40ch08 ITAG_eugene five_prime_UTR 56812451 56812517 . - . Parent=2441231;ID=2441238;from_BOGAS=1
SL2.40ch08 ITAG_sgn_markers watch 56809921 56811225 0.915 . . Name=C2_At4g19170;ID=6137688;Alias=SGN-M8299;Target=SGN-M8299;201;20749;20; ;Note=marker
20name(s):20C2_At4g19170;20C2;20SGN-M8299
SL2.40ch08 ITAG_sgn_markers watch_part 56809921 56810451 0.809 . . Parent=6137688;ID=6137689
SL2.40ch08 ITAG_sgn_markers watch_part 56811018 56811225 0.981 . . Parent=6137688;ID=6137695
SL2.40ch08 ITAG_sgn_unigenes watch 56809921 56812128 0.933 . . Name=SGN-U272543;ID=6137690;Target=SGN-U272543;201;201653;20;
SL2.40ch08 ITAG_sgn_unigenes watch_part 56809921 56810451 0.806 . . Parent=6137690;ID=6137691
SL2.40ch08 ITAG_sgn_unigenes watch_part 56811018 56812128 0.955 . . Parent=6137690;ID=6137696
SL2.40ch08 ITAG_sgn_unigenes watch 56809937 56812326 0.998 . . Name=SGN-U570287;ID=6137692;Target=SGN-U570287;201;201824;20;
SL2.40ch08 ITAG_sgn_unigenes watch_part 56809937 56810451 0.998 . . Parent=6137692;ID=6137693
SL2.40ch08 ITAG_sgn_unigenes watch_part 56811018 56812326 0.998 . . Parent=6137692;ID=6137697
SL2.40ch08 ITAG_sgn_loci watch 56810100 56812450 0.954545454545455 . . Name=carotenoid 20c leavage
20dioxygenase;ID=6137694;Target=Potato_SGNlocusID_1962_DQ206633;201;20410;20; ;blast_cdna_base_mismatches=19;cdna_seq_name=Solyc08g075490.1.1;blast_cdna_watch_pct=95.45;blast_cdna_t_gaps=0;blast_cdn
a_t_gap_bases=0;Alias=cd;DQ206633;Potato_SGNlocusID_1962_DQ206633;sgn_locus_id=1962;blast_cdna_q_gaps=0;blast_cdna_q_gap_bases=0
```

4. Genomic Browsers

GFF3 (<http://www.sequenceontology.org/gff3.shtml>)

Column 1:	seqid	example: SL2.40ch08
Column 2:	source	example: ITAG_eugene
Column 3:	type	example: mRNA
Columns 4 & 5:	start & end	example: 56809898 and 56812517
Column 6:	score	example: .
Column 7:	strand	example: -
Column 8:	phase	example: .
Column 9:	attributes	example: Name=Solyc08g075490.2.1;length=2620;

Exercise I



- I. You are a coffee researcher and want to understand more about caffeine synthesis. Using the tools we discussed, do the following analyses with caffeine synthase.
 1. Find some papers on caffeine synthase published since 2010.
 2. How many caffeine synthase protein sequences are in GenBank? How many are from *Coffea arabica*?
 3. How many organisms have a caffeine synthase homolog?
 4. Is caffeine synthase specific to the Gentianales clade or is it found elsewhere?
 5. Which of the homologs seem realistic? Download all *Coffea* sequences in fasta format and select full-length proteins. How many appear full-length?
 6. What reaction(s) does caffeine synthase catalyze?

Please save your results for the next exercise.



Exercise I Solutions

I. use pubmed (<http://www.ncbi.nlm.nih.gov/pubmed>)

Results: 6

Filters activated: Publication date from 2010/01/01 to 2014/01/01. [Clear all](#) to show 23 items.

- ☐ [Identification and isolation of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from guarana \(Paulinia cupana\).](#)
1. Figueirêdo LC, Faria-Campos AC, Astolfi-Filho S, Azevedo JL.
Genet Mol Res. 2011 Jun 21;10(2):1188-99. doi: 10.4238/vol10-2gmr1124.
PMID: 21732283 [PubMed - indexed for MEDLINE] [Free Article](#)
[Related citations](#)
- ☐ [Producing low-caffeine tea through post-transcriptional silencing of caffeine synthase mRNA.](#)
2. Mohanpuria P, Kumar V, Ahuja PS, Yadav SK.
Plant Mol Biol. 2011 Aug;76(6):523-34. doi: 10.1007/s11103-011-9785-x. Epub 2011 May 12.
PMID: 21562910 [PubMed - indexed for MEDLINE]
[Related citations](#)
- ☐ [Agrobacterium-mediated silencing of caffeine synthesis through root transformation in Camellia sinensis L.](#)
3. Mohanpuria P, Kumar V, Ahuja PS, Yadav SK.
Mol Biotechnol. 2011 Jul;48(3):235-43. doi: 10.1007/s12033-010-9364-4.
PMID: 21181507 [PubMed - indexed for MEDLINE]
[Related citations](#)
- ☐ [A transcriptomic approach highlights induction of secondary metabolism in citrus fruit in response to Penicillium digitatum infection.](#)
4. González-Candela L, Alamar S, Sánchez-Torres P, Zacarías L, Marcos JF.
BMC Plant Biol. 2010 Aug 31;10:194. doi: 10.1186/1471-2229-10-194.
PMID: 20807411 [PubMed - indexed for MEDLINE] [Free PMC Article](#)
[Related citations](#)
- ☐ [Essential region for 3-N methylation in N-methyltransferases involved in caffeine biosynthesis.](#)
5. Mizuno K, Kurosawa S, Yoshizawa Y, Kato M.
Z Naturforsch C. 2010 Mar-Apr;65(3-4):257-65.
PMID: 20469646 [PubMed - indexed for MEDLINE]
[Related citations](#)
- ☐ [Expression for caffeine biosynthesis and related enzymes in Camellia sinensis.](#)
6. Kato M, Kitao N, Ishida M, Morimoto H, Irino F, Mizuno K.
Z Naturforsch C. 2010 Mar-Apr;65(3-4):245-56.
PMID: 20469645 [PubMed - indexed for MEDLINE]
[Related citations](#)



Exercise I Solutions

2. 84 proteins total, 21 from *C. arabica* (<http://www.ncbi.nlm.nih.gov/protein>)

Protein Save search Advanced

Display Settings: Summary, 20 per page, Sorted by Default order Send to:

Results: 1 to 20 of 84 << First < Prev Page 1 of 5 Next > Last >>

- ☐ [TPA_exp: caffeine synthase \[Paullinia cupana var. sorbilis\]](#)
1. 360 aa protein
Accession: DAA64605.1 GI: 645085978
[GenPept](#) [FASTA](#) [Graphics](#)
- ☐ [caffeine synthase, partial \[Paullinia cupana var. sorbilis\]](#)
2. 351 aa protein
Accession: AHA44434.1 GI: 557942098
[GenPept](#) [FASTA](#) [Graphics](#)
- ☐ [caffeine synthase, partial \[Paullinia cupana var. sorbilis\]](#)
3. 134 aa protein
Accession: AHA44433.1 GI: 557942098
[GenPept](#) [FASTA](#) [Graphics](#)
- ☐ [caffeine synthase \[Camellia sinensis\]](#)
4. 369 aa protein
Accession: ABP98983.1 GI: 145952324
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

Exercise I Solutions

3. 15 species:

▼ Top Organisms [Tree]
Coffea arabica (21)
Camellia sinensis (14)
Coffea canephora (10)
Oryctolagus cuniculus (9)
Coffea eugenioides (8)
Citrus sinensis (5)
Triticum urartu (5)
Paullinia cupana var. sorbilis (3)
Glycine max (2)
Coffea benghalensis (2)
Saccharomyces cerevisiae S288c (2)
Fragaria vesca subsp. vesca (1)
Theobroma cacao (1)
Vitis vinifera (1)
Camellia sinensis var. sinensis (1)
Less...

*click on “Tree” for next question



Exercise I Solutions (cont'd)

4. Found in Gentianales, Ericales, Sapindales
Fabales, Rosales, Vitales, Malvales



*click on “List” for next answer



Exercise 1 Solutions (cont'd)

5. *Coffea arabica*, *Coffea canephora*, *Camellia sinensis*, *Theobroma cacao*. 18 *Coffea* sequences appear to be full-length.

The screenshot shows the Sol Genomics Network search results for the query 'caffeine synthase'. The left sidebar lists top organisms, with *Coffea arabica* having 21 sequences. The main results area shows 21 items, with the first three displayed: 1. 385 aa protein (Accession: BAB39213.1), 2. Probable caffeine synthase 4 (Accession: Q9AVL9.1), and 3. Probable caffeine synthase 3 (Accession: Q9AVK1.1). A 'Choose Destination' dialog box is open, showing options to download the 21 items in FASTA format. The 'Recent activity' section at the bottom right shows the current search query and the number of results (21).

▼ Top Organisms [Tree]

- Coffea arabica* (21)
- Camellia sinensis* (14)
- Coffea canephora* (10)
- Oryctolagus cuniculus* (9)
- Coffea eugenioides* (8)
- Citrus sinensis* (5)
- Triticum urartu* (5)
- Paullinia cupana* var. *sorbilis* (3)
- Glycine max* (2)
- Coffea benghalensis* (2)
- Saccharomyces cerevisiae* S288c (2)
- Fragaria vesca* subsp. *vesca* (1)
- Theobroma cacao* (1)
- Vitis vinifera* (1)
- Camellia sinensis* var. *sinensis* (1)
- Less...

Display Settings: ☑ Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 21

1. [caffeine synthase \[Coffea arabica\]](#)
385 aa protein
Accession: BAB39213.1 GI: 13365694
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

2. [RecName: Full=Probable caffeine synthase 4; Short=CtCS4; AltName=like 1; Short=CaMTL1 \[Coffea arabica\]](#)
385 aa protein
Accession: Q9AVL9.1 GI: 75168238
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

3. [RecName: Full=Probable caffeine synthase 3; Short=CtCS3; AltName: Full=Methyltransferase-like 2; Short=CaMTL2 \[Coffea arabica\]](#)
385 aa protein
Accession: Q9AVK1.1 GI: 75168231
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

4. [tentative caffeine synthase 7 \[Coffea arabica\]](#)

Choose Destination

☒ File ☐ Clipboard
☐ Collections

Download 21 items.

Format
FASTA

Sort by
Default order

Create File

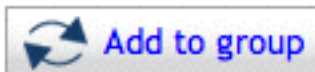
Search

Recent activity

(caffeine sy [orgn] (21))



7.) <http://metacyc.org/>



[MetaCyc](#) Enzyme: (

Gene: CaDXMT1

Accession Number: G-9084 (MetaCyc)

Species: [Coffea arabica](#)

Summary:

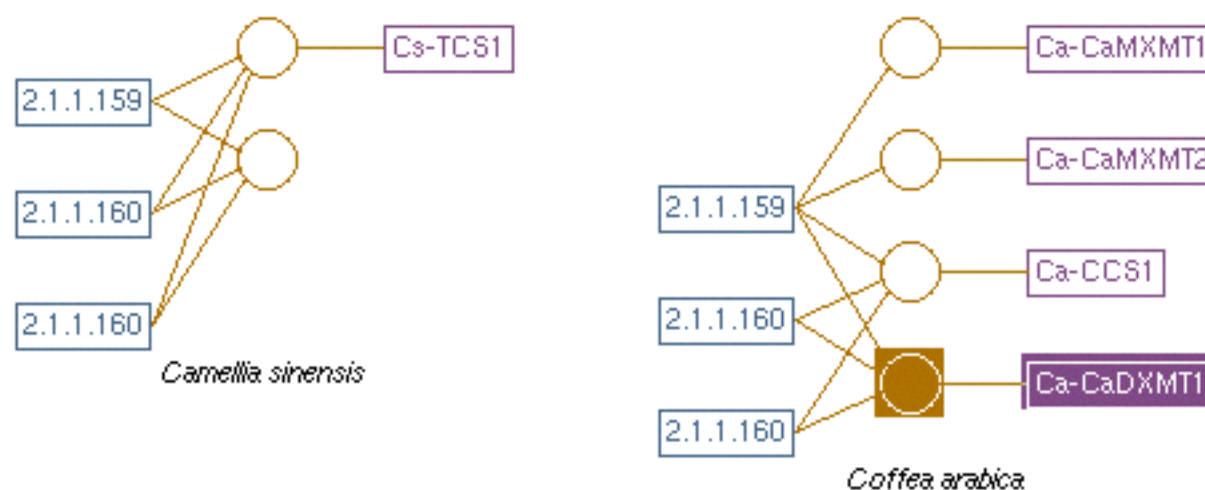
The recombinant CaDXMT1 catalyzes the conversion of 7-methylxansine to theobromine and theobromine to caffeine. In addition, Ca the most preferred substrate. XMP is not an effective substrate. CaDXMT1 is predominantly expressed in immature fruits of coffee.

Citations: [\[Uefuji03\]](#)

Molecular Weight of Polypeptide: 43.3 kD (from nucleotide sequence)

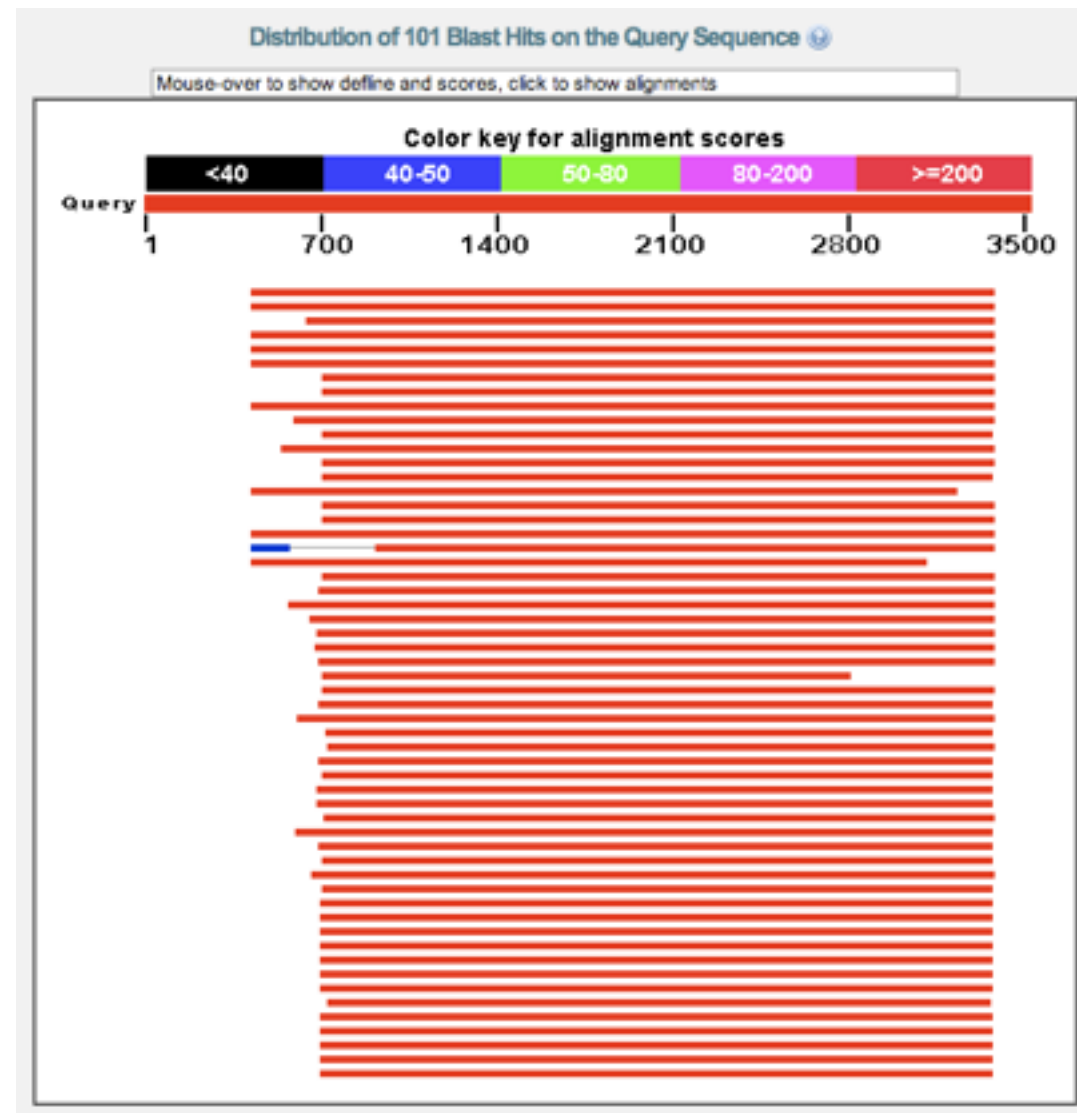
Relationship Links: [Entrez-Nucleotide:PART-OF:AB084125](#)

Gene-Reaction Schematic: ?





Part II: Web Tools



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

I. Search Tools

Text Searches:

One or more words are introduced in a box. The system use them to search coincidences with database fields or file sections such as genomic annotations.

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



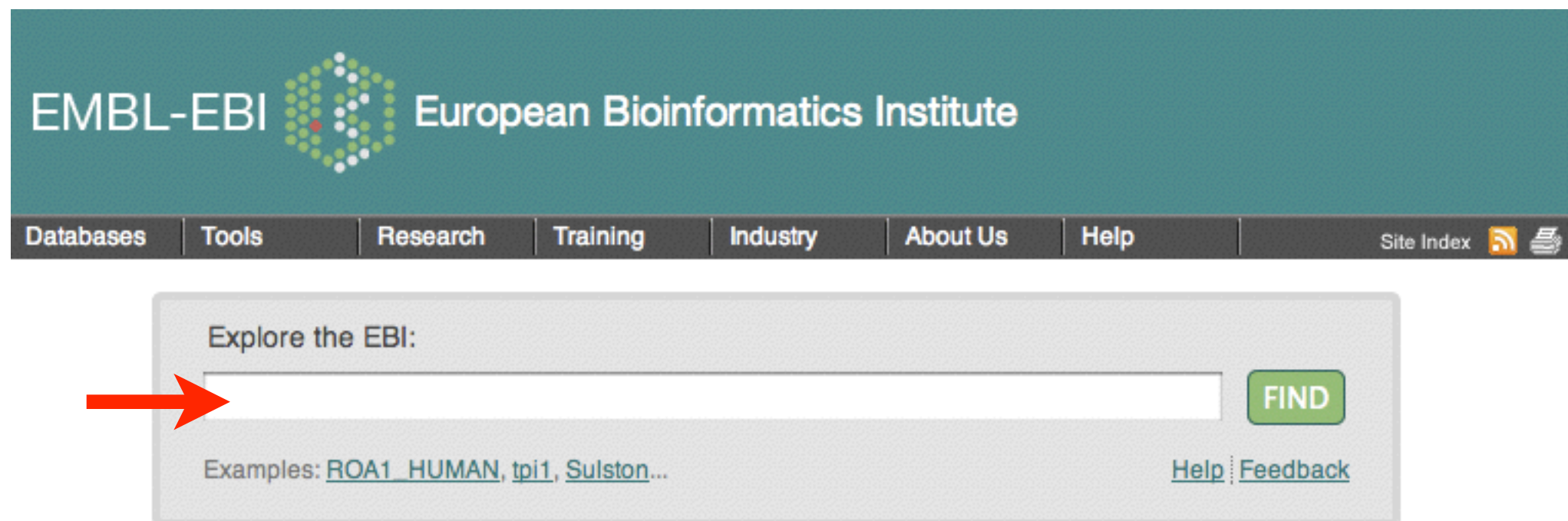
The screenshot shows the NCBI homepage. At the top, there is a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a large search bar with a dropdown menu set to 'All Databases'. A red arrow points to the search input field. To the left of the search bar is the NCBI logo and the text 'National Center for Biotechnology Information'. Below the search bar, the page is divided into three main sections: 'NCBI Home' (a sidebar with links like 'Site Map (A-Z)', 'All Resources', 'Chemicals & Bioassays', etc.), 'Welcome to NCBI' (a central section with a welcome message and links to 'About the NCBI', 'Mission', 'Organization', 'Research', and 'RSS Feeds'), and 'Popular Resources' (a right sidebar with links to 'BLAST', 'Bookshelf', 'Gene', 'Genome', 'Nucleotide', 'OMIM', 'Protein', 'PubChem', 'PubMed', 'PubMed Central', and 'SNP').

I. Search Tools



Text Searches:

EBI: <http://www.ebi.ac.uk/>



EMBL-EBI  European Bioinformatics Institute

Databases Tools Research Training Industry About Us Help Site Index  

Explore the EBI:



Examples: [ROA1_HUMAN](#), [tpi1](#), [Sulston...](#)

[Help](#) [Feedback](#)

FIND

Data Resources and Tools

- [ENA](#)
- [UniProt](#)
- [ArrayExpress](#)
- [Ensembl](#)
- [InterPro](#)
- [PDBe](#)
- [Genomes](#)
- [Nucleotide Sequences](#)
- [Protein Sequences](#)
- [Macromolecular Structures](#)
- [Small Molecules](#)
- [Gene Expression](#)
- [Molecular Interactions](#)
- [Reactions & Pathways](#)
- [Protein Families](#)
- [Enzymes](#)
- [Literature](#)
- [Taxonomy](#)
- [Ontologies](#)
- [Patent Resources](#)
- [Sequence Similarity & Analysis](#)
- [Pattern & Motif Searches](#)
- [Structure Analysis](#)
- [Text Mining](#)
- [Downloads](#)
- [Web Services](#)

I. Search Tools



sol genomics network

Text Searches:

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



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.

CARNEGIE INSTITUTION FOR SCIENCE TAIR is located at the [Carnegie Institution for Science Department of Plant Biology](#) and funded by the [National Science Foundation](#).

Updates on TAIR funding are available [here](#).

Breaking News

- [Subscribe to news feed](#)
- [Follow our Twitter feed](#)
- [Join our Facebook group](#)

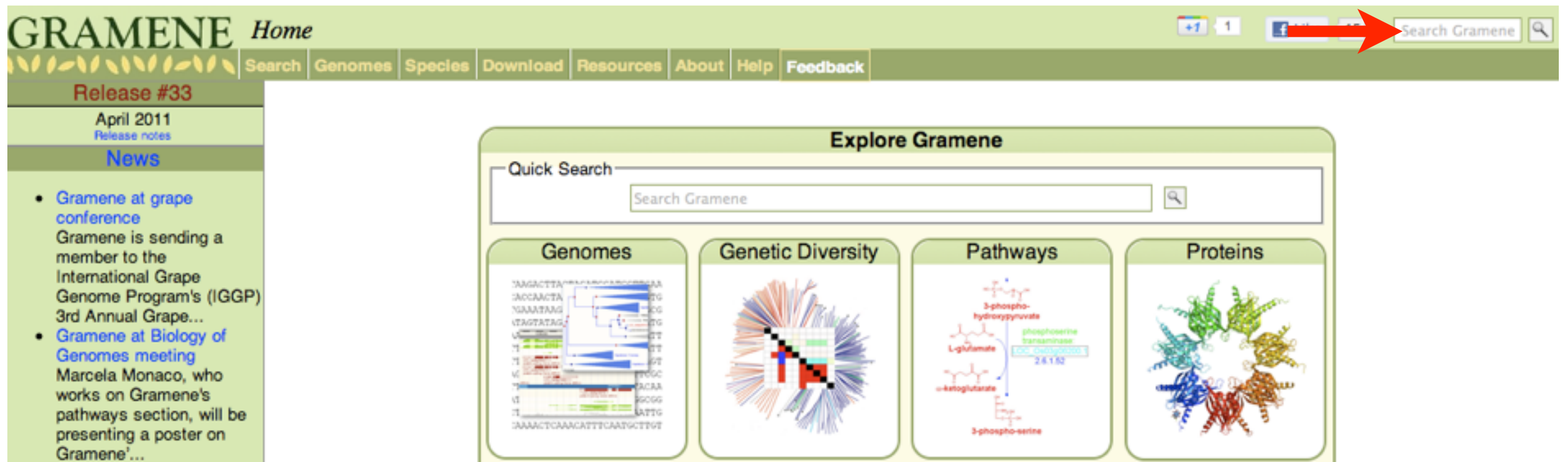
GBrowse now available for eight plant species at TAIR [May 19, 2011]

GBrowse instances for the following plants have been added to TAIR: *Arabidopsis lyrata*, *Brachypodium distachyon*, *Oryza sativa japonica*, *Oryza sativa indica*, *Populus trichocarpa*, *Physcomitrella patens*, *Sorghum bicolor*, *Vitis vinifera*, *Zea mays*.

I. Search Tools

Text Searches:

GRAMENE: <http://www.gramene.org/>



The screenshot shows the GRAMENE website homepage. The header includes the GRAMENE logo, a "Home" link, and a navigation bar with links: Search, Genomes, Species, Download, Resources, About, Help, and Feedback. A search bar is located in the top right corner, with a red arrow pointing to it. Below the navigation bar, there is a "Release #33" section dated April 2011, with a link to "Release notes". A "News" section follows, listing two items: "Gramene at grape conference" and "Gramene at Biology of Genomes meeting". The main content area is titled "Explore Gramene" and features a "Quick Search" bar. Below this, there are four panels: "Genomes" (showing a sequence alignment), "Genetic Diversity" (showing a phylogenetic tree), "Pathways" (showing a metabolic pathway diagram), and "Proteins" (showing a 3D protein structure).

I. Search Tools



sol genomics network

Text Searches:

SGN: <http://solgenomics.net/>



sol genomics network

[home](#) | [forum](#) | [contact](#) | [help](#)

[search](#)

[maps](#)

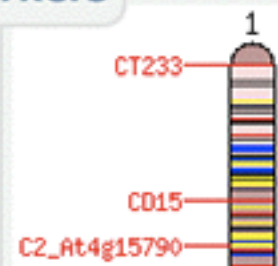
[genomes](#)

[tools](#)

[sol search](#)

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

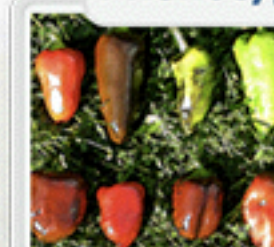
Maps & Markers



Genes



Phenotypes



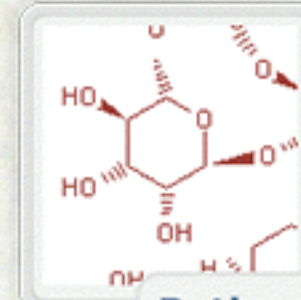
Breeders Toolbox



Genomes & Sequences



Pathways





Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

Ontologies in bioinformatics:

Controlled vocabulary of terms used to describe gene functions, localizations, plant organs... that allow the communication between different groups.

Example: *Size increase, Mass increase...*

Growth, The increase in size or mass of an entire organism, a part of an organism or a cell. (GO:0040007)

Ontology Types:

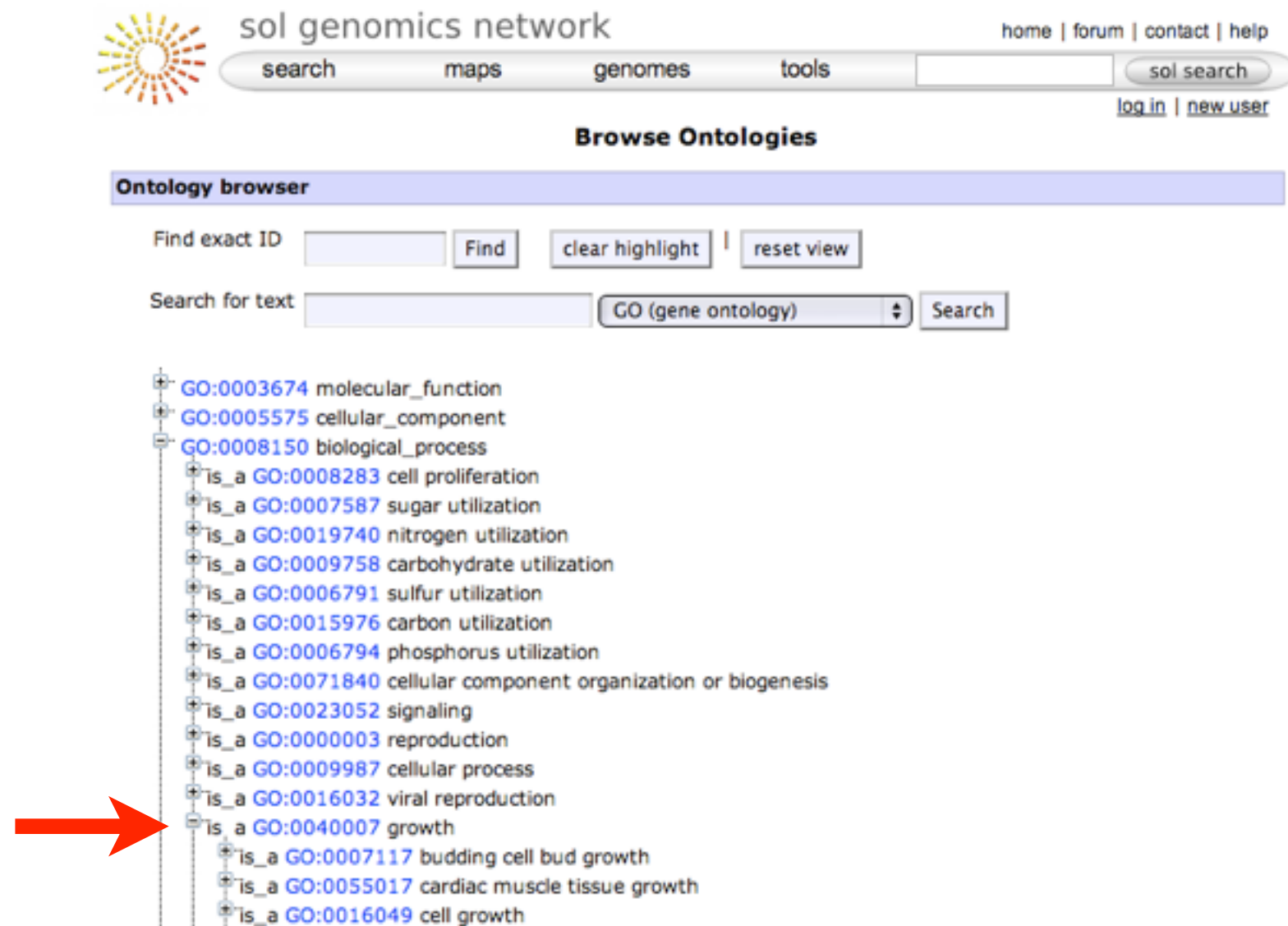
More used is *Gene Ontology* (<http://www.geneontology.org/>), to define biological process, cellular component and molecular function terms.

For plants, also is important *Plant Ontology* (<http://www.plantontology.org/>), to define plant parts such as organs and process associated to plants such as flowering or ripening.

A complete ontology list can be found at:

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

Searches by Ontology: Ontologies have parents-children relations. To optimize the search, it is common the use of tools such as the ontology browsers.



The screenshot shows the 'sol genomics network' website. At the top, there is a navigation bar with links for 'home', 'forum', 'contact', and 'help'. Below this is a search bar with a 'sol search' button. The main section is titled 'Browse Ontologies' and contains an 'Ontology browser' panel. This panel has a 'Find exact ID' field with a 'Find' button, a 'clear highlight' button, and a 'reset view' button. Below these is a 'Search for text' field with a dropdown menu set to 'GO (gene ontology)' and a 'Search' button. The search results are displayed as a tree structure. The root node is 'GO:0003674 molecular_function'. It has two children: 'GO:0005575 cellular_component' and 'GO:0008150 biological_process'. The 'biological_process' node is expanded, showing a list of its children. A red arrow points to the 'GO:0040007 growth' node, which is the first child in this list. The other children of 'growth' are 'GO:0007117 budding cell bud growth', 'GO:0055017 cardiac muscle tissue growth', and 'GO:0016049 cell growth'.

- GO:0003674 molecular_function
 - GO:0005575 cellular_component
 - GO:0008150 biological_process
 - is_a GO:0008283 cell proliferation
 - is_a GO:0007587 sugar utilization
 - is_a GO:0019740 nitrogen utilization
 - is_a GO:0009758 carbohydrate utilization
 - is_a GO:0006791 sulfur utilization
 - is_a GO:0015976 carbon utilization
 - is_a GO:0006794 phosphorus utilization
 - is_a GO:0071840 cellular component organization or biogenesis
 - is_a GO:0023052 signaling
 - is_a GO:0000003 reproduction
 - is_a GO:0009987 cellular process
 - is_a GO:0016032 viral reproduction
 - is_a GO:0040007 growth
 - is_a GO:0007117 budding cell bud growth
 - is_a GO:0055017 cardiac muscle tissue growth
 - is_a GO:0016049 cell growth

<http://solgenomics.net/tools/onto/index.pl>

Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

Sequence homology/similarity searches:

It is based in the sequence comparison through a pair sequence alignment using different algorithms (blast, uses an approach to the Smith-Waterman algorithm). Matched sequences (hits) with some statistical values are selected and returned as result.

Most used programs are:

- Blast: (faster) <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- Fasta (sensitive): <http://www.ebi.ac.uk/Tools/sss/fasta/>

More information at: http://en.wikipedia.org/wiki/Sequence_alignment_software

I. Search Tools



Sequence homology/similarity searches:

NCBI: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

The screenshot shows the NCBI BLAST homepage. At the top, there's a blue header with the BLAST logo and the text "Basic Local Alignment Search Tool". Below the header, there are navigation tabs: "Home", "Recent Results", "Saved Strategies", and "Help". The main content area has a section titled "NCBI BLAST Home" with a description: "BLAST finds regions of similarity between biological sequences. [more...](#)". Below this, there's a red box with the text "New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. [Go](#)".

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

☐ Human	☐ Oryza sativa	☐ Gallus gallus
☐ Mouse	☐ Bos taurus	☐ Pan troglodytes
☐ Rat	☐ Danio rerio	☐ Microbes
☐ Arabidopsis thaliana	☐ Drosophila melanogaster	☐ Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

I. Search Tools



Sequence homology/similarity searches:

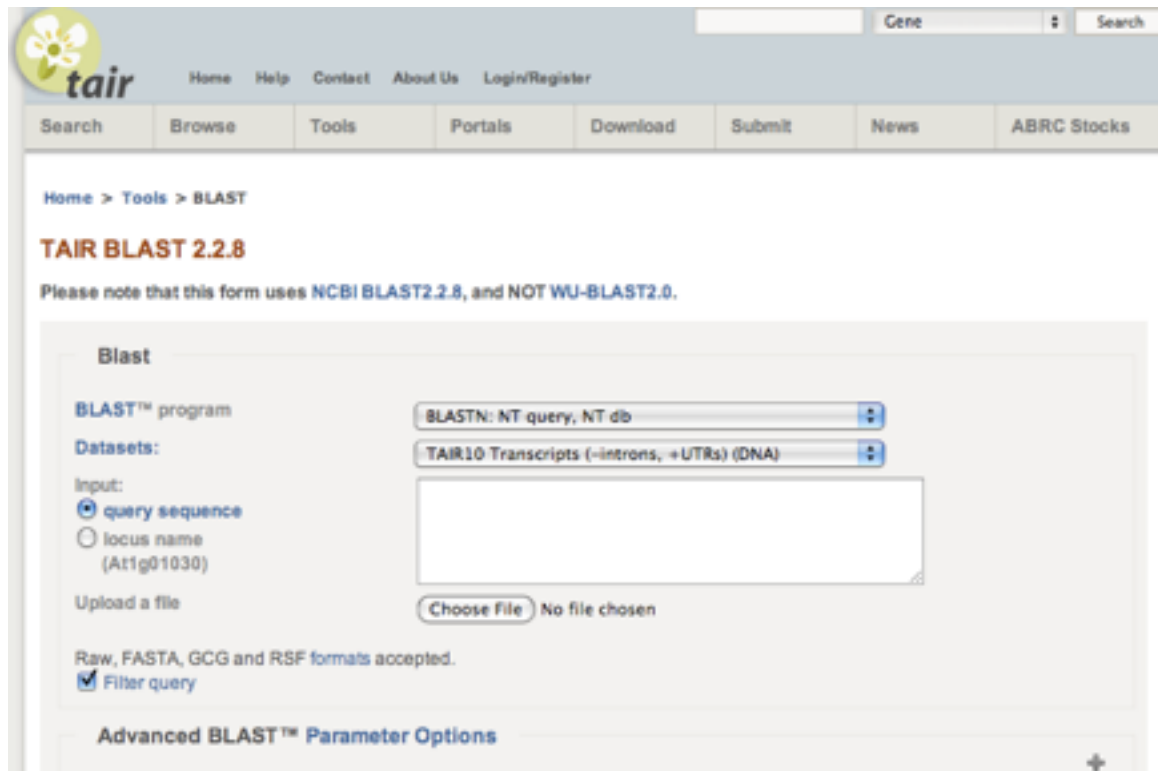
EBI: <http://www.ebi.ac.uk/Tools/sss/>

The screenshot shows the EMBL-EBI Sequence Similarity Searching (SSS) web interface. At the top, there is a navigation bar with the EMBL-EBI logo, a search bar with the text "Enter Text Here" and a "Find" button, and links for "Help" and "Feedback". Below the navigation bar, there are tabs for "Databases", "Tools", "Research", "Training", "Industry", "About Us", and "Help". The "Tools" tab is selected. On the left side, there is a sidebar with links: "Tools Home", "Tools A-Z", "Web Services", and "Download". The main content area is titled "Sequence Similarity Searching" and contains a list of tools under the heading "BLAST". The tools listed are: "NCBI BLAST" (NCBI BLAST Sequence Similarity Search using the NCBI BLAST (blastall) program), "WU-BLAST" (Sequence Similarity Search using the Washington University (WU) BLAST2 program), and "PSI-BLAST" (Position Specific Iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile is automatically constructed from the first set of BLAST alignments). Each tool has a "Launch" button. Below the "BLAST" section, there is a "FASTA" section with a "FASTA" tool (Sequence Similarity Search using the FASTA program). This tool also has a "Launch" button. Below the "FASTA" section, there is an "SSEARCH" section with an "SSEARCH" tool (Sequence Similarity Search using the SSEARCH program). This tool also has a "Launch" button. Below the "SSEARCH" section, there is a "PSI-Search" section with a "PSI-Search" tool (PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST (blastpgp) iterative profile construction strategy to find distantly related protein sequences). This tool also has a "Launch" button.

I. Search Tools

Sequence homology/similarity searches:

TAIR: <http://www.arabidopsis.org/Blast/index.jsp>
<http://www.arabidopsis.org/cgi-bin/fasta/nph-TAIRfasta.pl>



Home > Tools > BLAST

TAIR BLAST 2.2.8

Please note that this form uses NCBI BLAST2.2.8, and NOT WU-BLAST2.0.

Blast

BLAST™ program: BLASTN: NT query, NT db

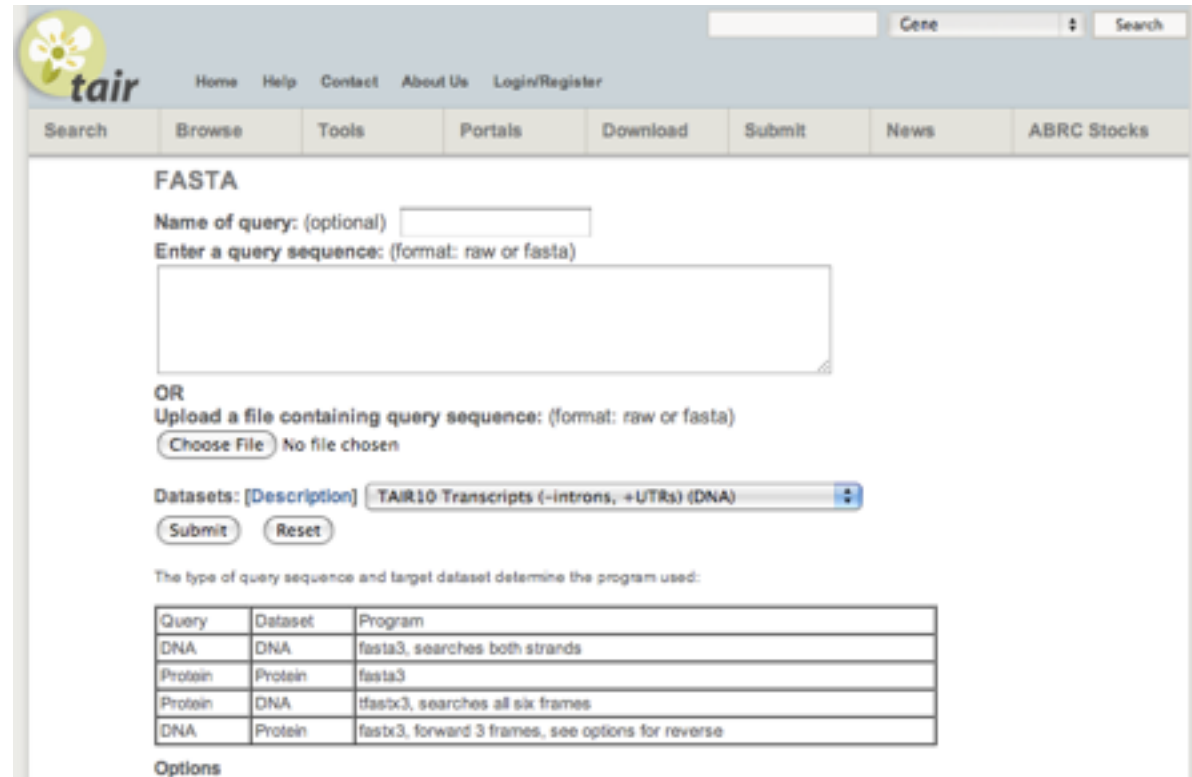
Datasets: TAIR10 Transcripts (-introns, +UTRs) (DNA)

Input:
☒ query sequence
☐ locus name (At1g01030)

Upload a file: Choose File No file chosen

Raw, FASTA, GCG and RSF formats accepted.
☒ Filter query

Advanced BLAST™ Parameter Options



Home > Tools > FASTA

FASTA

Name of query: (optional)

Enter a query sequence: (format: raw or fasta)

OR

Upload a file containing query sequence: (format: raw or fasta)
Choose File No file chosen

Datasets: [Description] TAIR10 Transcripts (-introns, +UTRs) (DNA)

Submit Reset

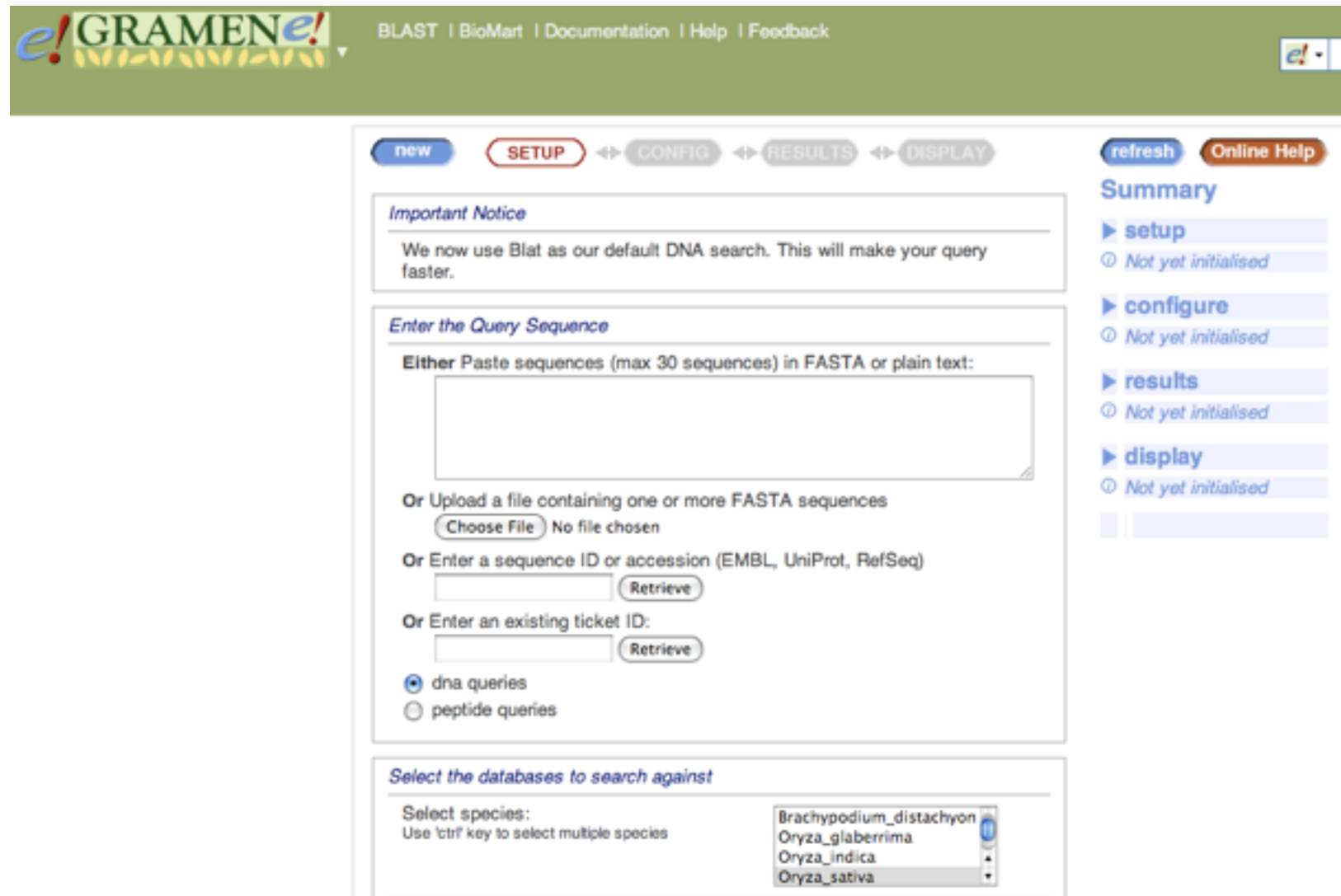
The type of query sequence and target dataset determine the program used:

Query	Dataset	Program
DNA	DNA	fasta3, searches both strands
Protein	Protein	fasta3
Protein	DNA	tfast3, searches all six frames
DNA	Protein	fastx3, forward 3 frames, see options for reverse

Options

Sequence homology/similarity searches:

GRAMENE: <http://www.gramene.org/multi/blastview>



The screenshot shows the GRAMENE multi-BLAST search interface. At the top, there is a green header with the GRAMENE logo and navigation links: BLAST | BioMart | Documentation | Help | Feedback. Below the header, there is a navigation bar with buttons: new, SETUP (highlighted), CONFIG, RESULTS, and DISPLAY. On the right side, there are buttons for refresh and Online Help, and a Summary section with links to setup, configure, results, and display, each with a status indicator "Not yet initialised".

Important Notice
We now use Blat as our default DNA search. This will make your query faster.

Enter the Query Sequence
Either Paste sequences (max 30 sequences) in FASTA or plain text:

Or Upload a file containing one or more FASTA sequences
 No file chosen

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)

Or Enter an existing ticket ID:

☒ dna queries
☐ peptide queries

Select the databases to search against
Select species:
Use 'ctrl' key to select multiple species

Brachypodium_distachyon
Oryza_glaberrima
Oryza_indica
Oryza_sativa

I. Search Tools

Sequence homology/similarity searches:

SGN: <http://solgenomics.net/tools/blast/index.pl>



sol genomics network

home | forum | contact | help

search maps genomes tools sol search

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

NCBI BLAST

Sequence Set [db details](#)

Program

Query sequence
single sequence only, use Advanced for multiple

Expect (e-value) Threshold

Substitution Matrix

Show Graphics

Max. hits to show

Blast:

It is a tool designed to find regions with local similarity for a sequence pair. It compare nucleotides or protein sequences and calculate the statistical significance.

Blast Programs:

		INPUT		
DATABASE		Nucleotide	Translated Nucleotide	Protein
	Nucleotide	BlastN	-	-
	Translated Nucleotide	-	TBlastX	TBlastN
	Protein	-	BlastX	BlastP



1.2 -Search by Sequence Homology

Blast uses:

- ▶ *Homologous gene search:*

BlastX (input=cDNA, database=proteins).

BlastP (input=protein, database=proteins).

TBlastN (input=proteins, database=cDNA)

- ▶ *Intron-Exon alignment:*

BlastN (input=cDNA, database=genomic DNA).

(better Blat or GeneWise)

- ▶ *SNP search:*

BlastN (input=cDNA,gDNA, database=cDNA,gDNA).

1.2 -Search by Sequence Homology

Blast terminology:

Query: Input sequence.

Subject: Sequence from the database

Query Coverage: Percentage of the input sequence cover by the database sequence.

E-value (expect value): Expected hits at random. It depends from the database size and it decrease exponentially with the sequence pair score.

% identity: Identity percentage for a sequence pair.



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

1.3 -Search by Sequence Location

Sequence Location Searches (By coordinates):

Based in the location at a chromosome region or a sequence. It has a start and end coordinates.

These searches generally uses Genome Browser as search software

More information at: http://en.wikipedia.org/wiki/Genome_browser



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

There are dozens of sequence manipulation tools with different licenses or for different operating systems.

+ Commercial package:

LaserGene (DNAS_{tar}) (<http://www.dnastar.com/t-products-lasergene.aspx>)

+ Free packages:

BioEdit (Windows) (<http://www.mbio.ncsu.edu/bioedit/bioedit.html>)

eBioTools (MacOS) (<http://www.ebioinformatics.org/>)

Some databases have programs with similar functions integrated with the database interface.



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment(Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

There are two tools types to find the right ORF for an expressed nucleotide sequence.

- Align with a mRNA or cDNA sequence, and select the longest ORF.
- Gene prediction based on the exon-intron structure

Tool types:

- ▶ Translators (DNA to proteins without exon-intron consideration, and analyzing all the possible ORFs). Use coding.
- ▶ *Gene Predictors* (DNA to CDS considering the intron-exon structure). They require software training with manually curated intron-exon structures.

Web-based translator programs:

- Translate Tool (ExPASy): <http://expasy.org/tools/dna.html>
- ORF Finder (NCBI): <http://www.ncbi.nlm.nih.gov/projects/gorf/>
- Transeq (EBI): <http://www.ebi.ac.uk/Tools/emboss/transeq/>
- RevTrans 1.4 Server (CBS): <http://www.cbs.dtu.dk/services/RevTrans/>
- Transeq (UMass): <http://biotools.umassmed.edu/cgi-bin/biobin/transeq>
- Dnatoprotein (JHI): <http://www.dnatoprotein.com/>
- EstScan (embnet): <http://www.ch.embnet.org/software/ESTScan2.html>

2.1 - Translators and Gene Predictors.



- **Transeq (EBI):** <http://www.ebi.ac.uk/Tools/emboss/transeq/>

EMBL-EBI

Enter Text Here Find Help Feedback

Databases Tools Research Training Industry About Us Help Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- EMBOSS-Transeq Help
- Emboss Programmatic Access

EBI > Tools > Sequence Analysis > EMBOSS

EMBOSS Transeq

Transeq translates nucleic acid sequences to the corresponding peptide sequence. It can translate in any of the 3 forward or three reverse sense frames, or in all three forward or reverse frames, or in all six frames.

Frame: 1 (selected), 2, 3, F, -1, -2, -3, R, 6

Table: Standard Code

Trim: No

Reverse: No

Colour: No

Enter or paste nucleic acid Sequence in any format: Help

Upload a file: Choose File No file chosen Run Reset

Web-based gene predictor programs:

- FGENESH (ULondon):

<http://mendel.cs.rhul.ac.uk/mendel.php?topic=fgen-file>

- GENESCAN (MIT):

<http://genes.mit.edu/GENSCAN.html>

- GeneMark.hmm (GaTech):

<http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi>

- Augustus:

<http://augustus.gobics.de/submission>



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

2.2 - Multiple Sequence Alignment

There are programs for multiple sequence alignment (nucleotide or protein) such as ClustalW or Muscle

Some of them, as ClustalW, can create simple phylogenetic trees based in simple algorithms such as *Neighbor-Joining*.

- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>
- Kalign (EBI): <http://www.ebi.ac.uk/Tools/msa/kalign>
- MAFFT (EBI): <http://www.ebi.ac.uk/Tools/msa/mafft>
- MUSCLE (EBI): <http://www.ebi.ac.uk/Tools/msa/muscle>
- T-Coffee (EBI): <http://www.ebi.ac.uk/Tools/msa/tcoffee>

2.2 - Multiple Sequence Alignment



- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>

```
sequence.fasta
>gil15225089|refINP_180705.1| F-box protein DOR [Arabidopsis thaliana]
MKSRRQNVSVARQITLGRDENFEPIIDLVIEIFSRSPVKSIARCRCVSKLWASILRLPYFTELYLTKSC
ARPRLLFACQKHRELFFFSTPQPHNPNESSSPLAASFHMKIPFDGRFNIISPIGGLVFVRYEQILKGRKT
PEFVSAICNPSTGQSLTLPKPKTRKRIWGTSHFGYDPIEKQFKVLSMNIGDGVYKEHYVLTGTENLSWR
RIECSIPHVHSGKGICINGVLYYRAKADMFSGTLMIVCDFVRFEKFSYIKILKPTTTLISYNGKLASLVW
EGPSYICGKRFEMWVLGDPEKHEWLKHTYELRPRWQNVLGEDLLIFAGMTGTNEIVLSPKYPSHPFYVY
YNLERNTIRRVEIQGMGAFKVNEDYIFLDHVEDVKLI

>gil297837229|refIXP_002886496.1| hypothetical protein ARALYDRAFT_893290 [Arabidopsis lyrata subsp. lyrata]
MKTERQNVSEDEVVVTERNKRAKTSNNGGEPIDLTVEICSRPAKSISRFRCVLKLGWSILRLPYFTE
LFLTRSLARPQLLFACHKDNHVFVFSSPQPNIDNNASSLLAANYHMKIPFYASSFERCSSVRGLVFFG
DERYSNGKEHKVSVICNPSTRQSLTLPKPKTRKRIWGTSHFGYDPIEKQFKVLSMTWGIYGTDRMDSEEH
QVLTGTRKPSWRMIECWIPHSLYHTYNNVCINGVLYYPAVNTSSKGFIIVSFDRSEEFRFVEDTDTSI
SSYYGPHLINYNGKLGSLGSGGFGGIGASCTSLRVLEDAEKHEWSEHIYVLPAAWKNIFGGECTVLSV
VGVTRTNEIVLSLRFSTPFYVFFYNTERNAIRRVEIQGQEAFFKDHVSFVFLDHVENVMKLEGF

>gil15239182|refINP_201386.1| F-box protein [Arabidopsis thaliana]
MRTLRRNVTENRLTISRRTKKTSPNKTEKSVQIPVDIIIEILLRLPAKSIATCRVSKLWISVICRQD
FTFLTRSLHRPQLLFCCCKDGNLFFSSPQLQNPYENSSAISLKNFSLCYKISRPVNGLICFKRKEMN
ETVTVICNPSTGHTLSLPKPMKTSIGPSRFFVYEPIQKQFKVLLSYKSEHQVLTGTGELSWRIECSM
PHILGMSEICINGVLYYPAINLSSGDYIIVCFDVRSEKFRFITVMEEFKAADGTLINYNGLASLVSE
RYCFVDGRSKSIELWVLQDAEKKEWSKHTYVLPAAWQHRIGTLNLRFGVTRTNEIMLSPCYQTVPFQVY
YFNIERKTMMVAIQGMEAFQGHVFTYLDHVENVKLLHNMF

>gil15229553|refINP_189038.1| putative F-box protein [Arabidopsis thaliana]
MRSRQLHNVSEDRETLSRRNKRSKTSLNHGIPIIDLLIEIFLKLVPKSIATCRVSKFWTYVLGRQDFTEL
FLTKSSSRPQLLFACANDNGYFFSSNPQNLDENSSPIAAYPLTHVPKSRDLGPPINGLVSLRGERILK
GRIRPVDVSIINPSTGESLTLPKTNMTRKKIYTVTSFLGYDPIEKQFKVLSMNMSYEKHPKCEGYQVLT
LGTGKLSWRMIKCLNYQHPLKNSEICINGVLYYLAAMVNGSSWPTRAVVCFDIRSEMFMFMEVYRELSYT
TTLINYNNGKLGMLMGQEAHKTISGICRSFELWVLEDTVKHEWSKHVYLLPPLWKDAVANTRYFAGMIG
TSEIVLFRPDEPLCVFYINIDRNTIKRVGIRGLEAFKYFRIFLNHVENVKLF

>gil297819588|refIXP_002877677.1| hypothetical protein ARALYDRAFT_906230 [Arabidopsis lyrata subsp. lyrata]
MSTMMKKRRHVSKEVALTISSSLGEYGENSGTLPMDLMVEILSRVPAKSAKFHCVSKNWNLSLLRSSY|
FTNLYLTRSPTRPRLITFQAEGKWSFFSSPEYLISDQNSNLVVVDNHMDVPKDYSFGVCEPVCGLLCTR
DEWVLSRKKDARMMICNPSTRQFQSLPKVRSRRNKVITYIGYDPIEKQFKVLCMTICERPVMFKAEEHQV
LTLGTGKLKWRMLKCFVEHFPKHKEICINGVLYYLAVKDETREDIIVCFHVKEKFQFILNKAPLSTLIN
YNGKLGIRHGFMEGGVAGYELWDLIEKEDWTRHIHILPPMWKQVVGETRYYVVGIMGTSEIVFSPFVK
SNPFYIFHLNIERNISITRVEIQGTGPLEGQVYTFINHIENVKLIM
```

Input: Set of 5 proteins

2.2 - Multiple Sequence Alignment



- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>

EBI > Tools > Multiple Sequence Alignment > ClustalW2

ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

```
subsp. Jyratal
MSTMMKKRRKRVSKEDVALTISSSLGEYGENSGTLPMDLMVEILSRVPAKSAKFHCVSKNWNLSLLRSSY
FTNLYLTRSPTRPRLLITFQAEGKWSFFSSPEYLISDQNSNLVVVDNHMDVPKDYSGVCEPVCGLLCTR
DEWVLSRKKDARMMICNPSTRQFQSLPKVRSRRNKVITYIGYDPIEKEYKVLCTICERPVMFKAEHQV
LTLGTGKLKWRMLKCFVEHFPHHKEICINGVLYYLAVKDETREDIIVCFHVKEKFQFILNKAPLSTLIN
YNGKLGIRHGFMEGGVAGYELWDLIEKEDWTRHHILPPMWKQVVGETRVYVVGMICTSEIVFSPFVK
SNPFYIFHLNIERNISITRVEIQGTGPLEGQQVYTFINHIENVKLIM
```

Or, upload a file: No file chosen

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ☒ Slow ☐ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Set your Multiple Sequence Alignment Options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 4 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

2.2 - Multiple Sequence Alignment



- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>

EBI > Tools > Multiple Sequence Alignment > ClustalW2

ClustalW2 Results

Alignments Result Summary Guide Tree Submission Details Submit Another Job

Alignment

Download Alignment File Show Colors

CLUSTAL 2.1 multiple sequence alignment

```
gi|15225089|ref|NP_180705.1|      ---MKSR-RQNVSVARQTLGRDE-----NFEPIPIDLVIEIFSR 36
gi|297837229|ref|XP_002886496.  ---MKTE-RQNVSEDVVVVTERNKRAKTSNN---GGEPIPFDLTVEICSR 43
gi|15239182|ref|NP_201386.1|    ---MRTL-RRNVTENRLTISRRTTEKKTSPNKTEKSVQIPVDIIIEILLR 46
gi|15229553|ref|NP_189038.1|    ---MRSRQLHNVSEDRETLSRRNKRSKTSLN----GHIPIDLLIEIFLK 42
gi|297819588|ref|XP_002877677.  MSTMMKKRKRHVSKEDVALTISSSLGEYGEN----SGTLPMDLMVEILSR 46
      * .   :::   .:               :*.*: :**  :

gi|15225089|ref|NP_180705.1|    SPVKSIARCRCVSKLWASILRLPYFTELYLTKSCARPRLLFACQKHREL 86
gi|297837229|ref|XP_002886496.  LPAKSISRFRCLVCLKWGSILRLPYFTEFLTRSLARPQLLFACHKDNHVF 93
gi|15239182|ref|NP_201386.1|    LPAKSIATCRCVSKLWISVICRQDFTEFLTRSLHRPQLLFCCCKDGNLF 96
gi|15229553|ref|NP_189038.1|    LPVKSIATCRSVSKFWTYVLGRQDFTEFLTKSSSRPQLLFACANDNGYF 92
gi|297819588|ref|XP_002877677.  VPAKSAAKFHCVSKNWNSLLRSSYFTNLYLTRSPTRPRLITFQAEKWS 96
      *.** :  :.* * *  :  :  **:*:*:*:  **:*:  .

gi|15225089|ref|NP_180705.1|    FFSTPQPHNP--NESSSPLAASFHMKIPFDG-RFNIISPIGGLVVFVRYEQ 133
gi|297837229|ref|XP_002886496.  VFSSPQPQNIDDNNASSLLAANYHMKIPFYASSFERCSSVRGLVFFGDER 143
gi|15239182|ref|NP_201386.1|    FFSSPQLQNPYENSSAISLKN-----FSLCYKISRPNGLICFK--- 135
gi|15229553|ref|NP_189038.1|    FFSSNQPNLDENSSPIAAYP-----LTHVPKSRDLGPPINGLVSLRGER 137
gi|297819588|ref|XP_002877677.  FFSSPEYLISDQNSNLVVVDN---HMDVPKDYSFGVCEPVCGLLCTRDEW 143
      .**: :      * .               .: **:

gi|15225089|ref|NP_180705.1|    ILKGRKTPEFVSAICNPSTGQSLTLPKPK-TRKRIW-GTSHFGYDPIEQ 181
gi|297837229|ref|XP_002886496.  YSNGKE--HKVSVICNPSTRQSLTLPKLK-TRKRIG-VRSYFGFEPIEQ 189
gi|15239182|ref|NP_201386.1|    ---RKEMNETVTVICNPSTGHTLSLPKP--MKTSIG-PSRFFVYEPIQK 179
gi|15229553|ref|NP_189038.1|    ILKGRIRPVDVSIINPSTGESLTLPKTNMTRKKIYTVTSFLGYDPIEQ 187
gi|297819588|ref|XP_002877677.  VLSRKK--DARMMICNPSTRQFQSLPKVRSRRNKVI---TYIGYDPIEKE 188
      :      * **** .  :***  :. :  .: :*:*:*
```

2.2 - Multiple Sequence Alignment



- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>

EBI > Tools > Multiple Sequence Alignment > ClustalW2

ClustalW2 Results

Alignments Result Summary **Guide Tree** Submission Details Submit Another Job

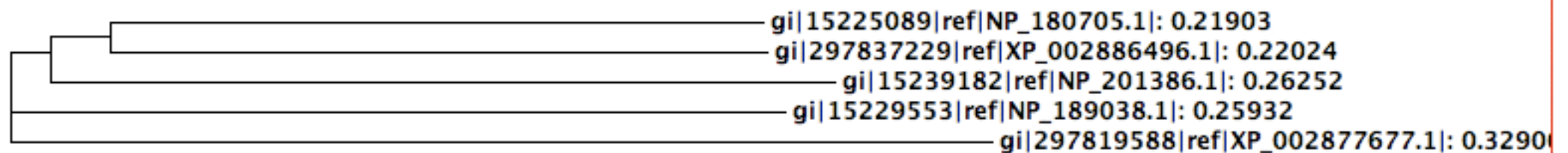
Guide Tree

Download Guide Tree File

```
(
(
(
gi|15225089|ref|NP_180705.1|:0.21903,
gi|297837229|ref|XP_002886496.1|:0.22024)
:0.01981,
gi|15239182|ref|NP_201386.1|:0.26252)
:0.01344,
gi|15229553|ref|NP_189038.1|:0.25932,
gi|297819588|ref|XP_002877677.1|:0.32906);
```

Phylogram

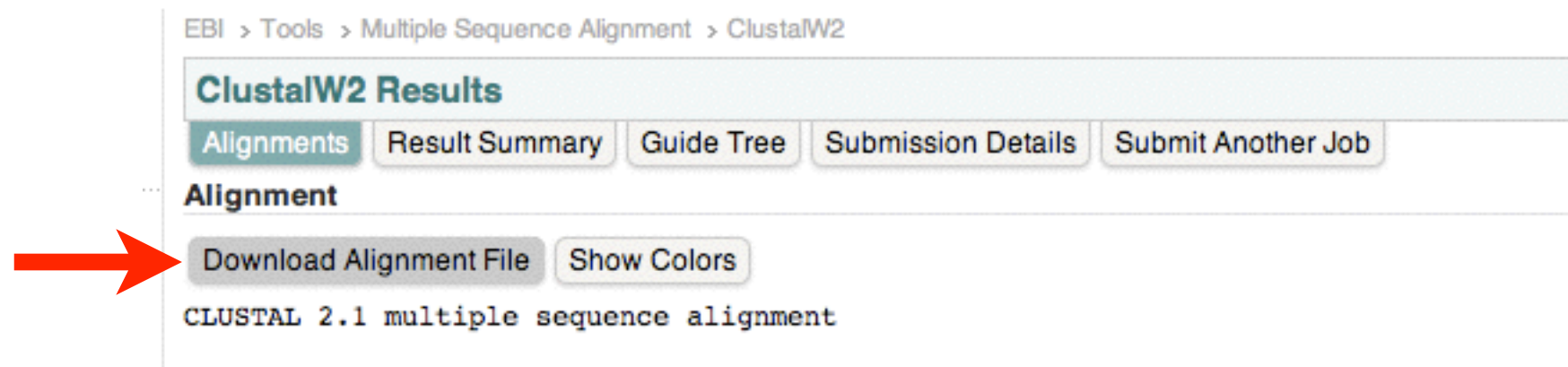
Show as Cladogram Tree Hide Distances



Right-click on the above tree to see display options.

2.2 - Multiple Sequence Alignment

- ClustalW (EBI): <http://www.ebi.ac.uk/Tools/msa/clustalw2/>



The alignment can be downloaded to be used by phylogenetic programs like Protpars (from Phylip package).

phylip 3.67: protpars

Protein Sequence Parsimony Method ?

Alignment File ? (use example data)

paste upload EDIT CLEAR

Enter your data below:

```
DIQGMCAFKV NEDYFELDHY EDVKU-----
DIQGMCAFKD HSYVTELDHY ENVMKLLG F
AIQGMCAEQG HSYVTELDHY ENVMKLLHNE -
GIRGLEAKY FR--IFUNHY ENVKLF-----
DIQGTGLEG QQVYTIINH ENVKUM-----
```

▣ Parsimony options

Use Threshold parsimony (T)

Threshold parsimony value

Genetic code for 'categories' model (C)

Protein parsimony algorithm, version 3.

One most parsimonious tree found:

```
      +--ArFbox2
+-----3
!      +--AtFbox1
+--2
! !      +--ARALY_9062
! +-----4
!      +--ARALY_8932
!
+-----AtDOR
```

remember: this is an unrooted tree!

requires a total of 1060.000

Web-based Phylip package:

<http://mobyli.pasteur.fr/cgi-bin/portal.py?#welcome>



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

2.3 -Protein Domain Analysis

Some of the functional annotations are made by homology search with conserved protein fragments or **domains**.

InterPro (<http://www.ebi.ac.uk/interpro/>) is an EBI resource with several protein domain databases such as *ProSite*, *Pfam* or *Superfamily*.



The tools used for functional domain search is InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>).

2.3 -Protein Domain Analysis



InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>).

EBI > Tools > Protein Functional Analysis > InterProScan

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package ([Readme file](#) or [FAQ's](#)), or the [InterPro user manual](#) or [help pages](#).

Use this tool

STEP 1 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

```
>gi|15225089|ref|NP_180705.1| F-box protein DOR [Arabidopsis thaliana]
MKSRRQNVSVARQTILGRDENFEPIIDLVEIFSRSPVKSIARCRCVSKLWASILRLPYFTELYLTKSC
ARPRLLFACQKHRELFFSTPQPHNPNESSSPLAASFHMKIPFDGRFNIISPIGGLVFRVEQILKGRKT
PEFVSAICNPSTGQSLTLPKPKTRKRIWGTSHFGYDPIEQFKVLSMNIGDGVYKEHYVLTGLENLSWR
RIECSIPHVHSGKICINGVLYYRAKADMFSGTLMIVCFDVRFEKFSYIKLKPTTTTISYNGKLASLVW
EGPSYICGKRFEMWVLGDPEKHEWLKHTYELRPRWQNVLGEDLLIFAGMTGTNEIVLSPKYPSPHPFYFY
YNLERNTIRRVEIQGMGAFKVNEDYIFLDHVEDVKLI
```

Or, upload a file: No file chosen

STEP 2 - Select the applications to run

Select All Clear All

☒ BlastProDom	☒ FPrintScan	☒ HMMPIR	☒ HMMPfam	☒ HMMSmart
☒ HMMTigr	☒ ProfileScan	☒ HAMAP	☒ PatternScan	☒ SuperFamily
☒ SignalPHMM	☒ TMHMM	☒ HMMPanther	☒ Gene3D	

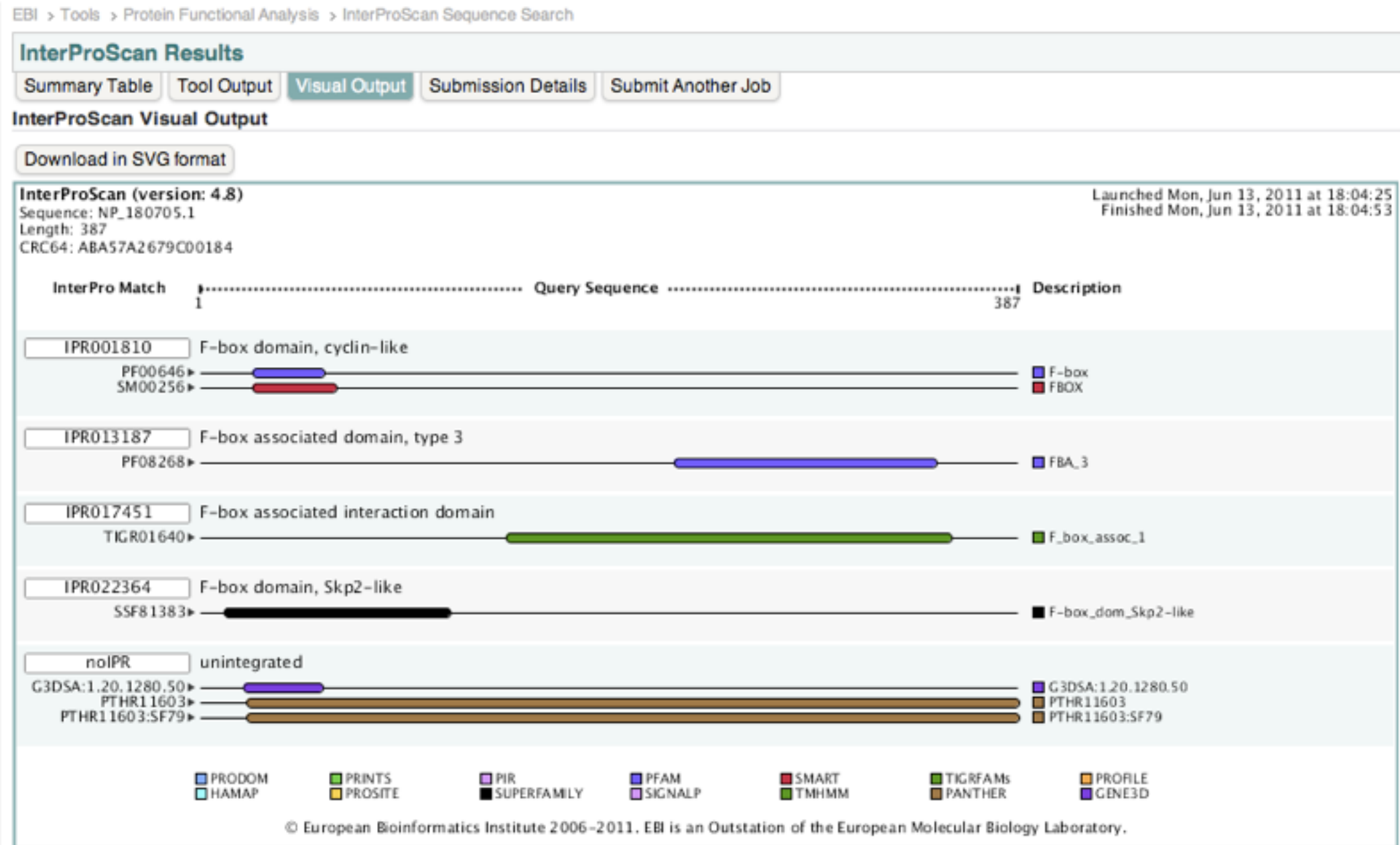
STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

2.3 -Protein Domain Analysis



InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>).



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

2.4 - Signal Peptide Analysis

A signal peptide is a short (3-60 amino acids long) peptide chain that directs the transport of a protein. Signal peptides may also be called targeting signals, signal sequences, transit peptides, or localization signals. (wikipedia).

Examples:

Transport to the nucleus (NLS)	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
Transport to the endoplasmic reticulum	H ₂ N-Met-Met-Ser-Phe-Val-Ser-Leu- Leu-Leu-Val-Gly-Ile-Leu-Phe- Trp-Ala-Thr-Glu-Ala-Glu-Gln- Leu-Thr-Lys-Cys-Glu-Val-Phe- Gln-
Retention to the endoplasmic reticulum	-Lys-Asp-Glu-Leu-COOH
Transport to the mitochondrial matrix	H ₂ N-Met-Leu-Ser-Leu-Arg-Gln-Ser- Ile-Arg-Phe-Phe-Lys-Pro-Ala- Thr-Arg-Thr-Leu-Cys-Ser-Ser- Arg-Tyr-Leu-Leu-
Transport to the peroxisome (PTS1)	-Ser-Lys-Leu-COOH
Transport to the peroxisome (PTS2)	H ₂ N-----Arg-Leu-X ₅ -His-Leu-

SignalP (<http://www.cbs.dtu.dk/services/SignalP/>) is a program to predict signal peptides.



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.



3.2 - Primer Design.

There are some web-based tools to design primers or to check the possible amplify fragments with the primers designed.

- Primer-Blast (NCBI) (design):

<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>

- Primer3 (design):

<http://frodo.wi.mit.edu/primer3/>

- In-Silico PCR (SGN) (fragment analysis):

<http://solgenomics.net/tools/insilicopcr/index.pl>

3.2 - Primer Design.



- Primer3 (design): <http://frodo.wi.mit.edu/primer3/>

Copy the downloaded sequence to Primer3.

Change min. size to 123 pb (intron size)

Change target to 200 (intron start), 123 (intron length)

Primer3 (v. 0.4.0) Pick primers from a DNA sequence.	Checks for mispriming in template.	disclaimer	Primer3 Home
	Primer3plus interface	cautions	FAQ/WIKI

Paste source sequence below (5'→3', string of ACGTNacgtt -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Mispriming Library \(repeat library\)](#):

```
AACGTCAATGAATAGATAGATGGCTGCCGCGGCAATCCAAAGTTCCCGGCTGCTTCCCG
CCACCACCACTTCCACCTCACCTGGTGGCTCATTACCAAAGTTCTTGAAATGATAATTA
CTCCCCATTTCACTAAACTCCTCAGTCCTCACACAATCCGCCTTCAAACCTCAGCTCTG
TTATTCAGAATCAAAAACCTACATATCAGATCAACAAGTTAATTCCTTCCCTTTGAA
CCTTTTTCCTTATCATACTGTTCAACCTTCACATAAATGTACATCTATTTACAAACACA
CAGTTAATTAAGCAAAATATACCTGCAAAGAGATCAAAAATCAATTTACAGCTAAAC
```

☒ Pick left primer, or use left primer below:	☐ Pick hybridization probe (internal oligo), or use oligo below:	☒ Pick right primer, or use right primer below (5' to 3' on opposite strand):

[Sequence Id:](#) A string to identify your output.

[Targets:](#) E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the [source sequence](#) with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

[Excluded Regions:](#) E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the [source sequence](#) with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

[Product Size Ranges](#)

Number To Return	5	Max 3' Stability	9.0
Max Repeat Mispriming	12.00	Pair Max Repeat Mispriming	24.00
Max Template Mispriming	12.00	Pair Max Template Mispriming	24.00

-
- | Response | Percentage |
|--|------------|
| Yes, the government is doing a good job | 68% |
| No, the government is not doing a good job | 32% |

Government	Percentage
Current government	85%
Previous government	15%



Bioinformatic Web Tools:

1 - Search Tools:

1.1 - By Ontology.

1.2 - By Sequence Homology/Similarity (Blast).

1.3 - By Sequence/Chromosome coordinates (GBrowse).

2 - Manipulation and Sequence Analysis Tools:

2.1 - Translators and *Gene Predictors*.

2.2 - Multiple Sequence Alignment (Clustalw).

2.3 - Protein Domain Analysis (InterProScan).

2.4 - Signal Peptide Analysis (SignalP).

3 - Other Tools:

3.1 - Linkage Map Viewers (CViewer).

3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.



4 - Web Pages with Multiple Tools.

Useful bioinformatic web-portals with classical bioinformatic tools *on-line*:

- EBI (European Bioinformatic Institute): Analysis of sequences.

<http://www.ebi.ac.uk/Tools/>

- Mobyle (Institut Pasteur): Phylogenetic analysis.

<http://mobyle.pasteur.fr/cgi-bin/portal.py?#welcome>

- ExPASy (SwissProt): Analysis of proteins and sequences.

<http://expasy.org/tools/>

- CBS (Center For Biological Sequence Analysis).

<http://www.cbs.dtu.dk/biotools/>

- Phylemon2: Molecular evolution analysis

<http://phylemon.bioinfo.cipf.es/evolutionary.html>

Exercise 2

1. Select CaDXMTI from exercise 1 part 5, what domains can be found?
2. Find the *Arabidopsis thaliana* best protein match to CaDXMTI.
3. Find the tomato best protein match to CaDXMTI
4. What sequences are upstream and downstream of the tomato match from part 2? How many introns does the match have?
5. Align all sequences from exercise 1.4 with the *Arabidopsis* and tomato protein matches.
6. Make a phylogenetic tree with the alignment from 5. Which sequences appear to be most closely related?



Exercise 1 Solutions (cont'd)

1. interpro scan (<http://www.ebi.ac.uk/interpro/>)
: SAM dependent carboxyl methyltransferase

Protein family membership

IPR005299 SAM dependent carboxyl methyltransferase (IPR005299)

Domains and repeats

None predicted.

Detailed signature matches

IPR005299	SAM dependent carboxyl methyltransferase	PF03492 (Methyltran...)
no IPR	Unintegrated signatures	G3DSA:3.40.50... PB014564 (Pfam-B_14564) PTHR31009 PTHR31009:SF0 SSF53335

Exercise 2 Solutions

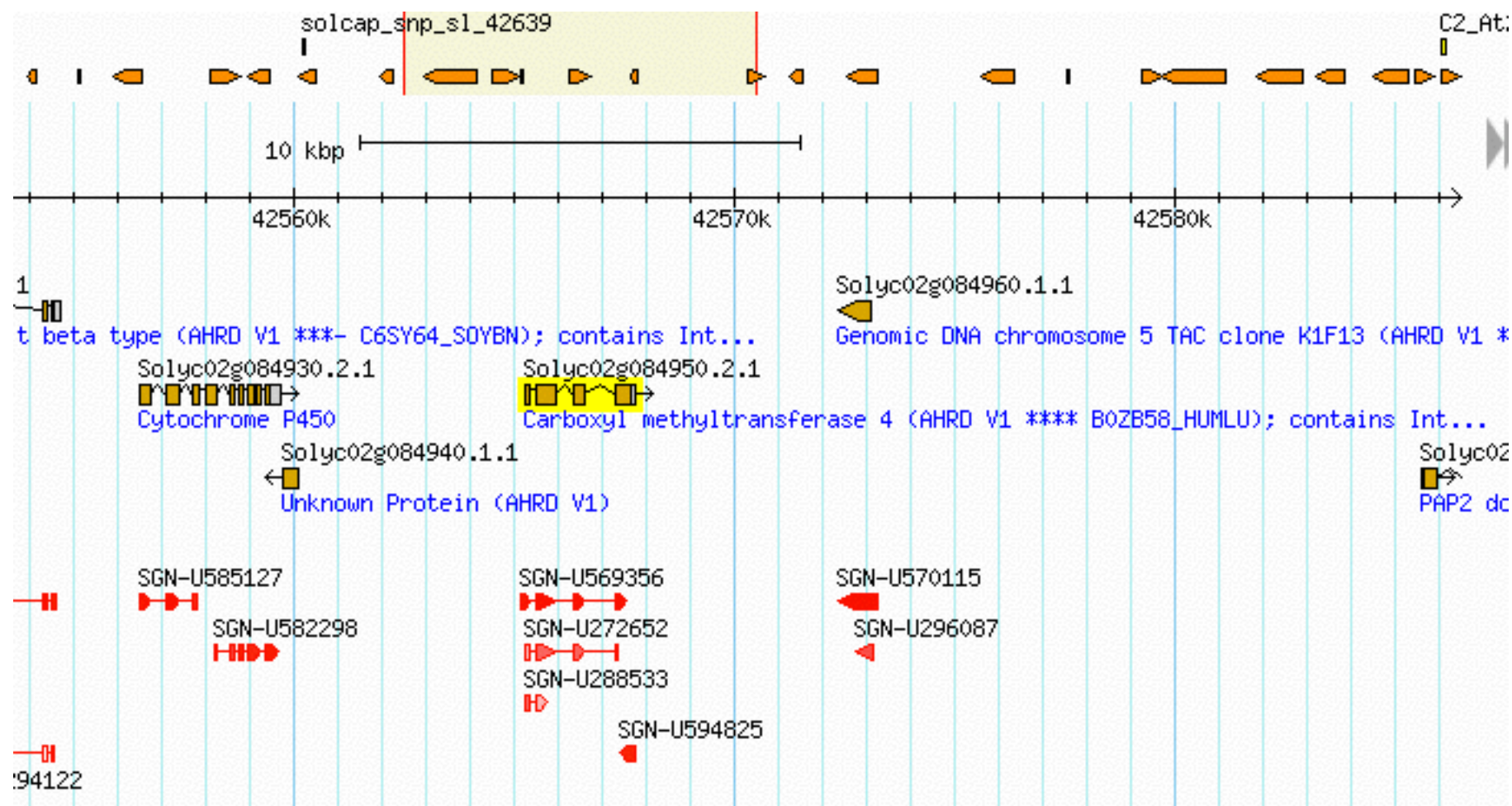
2. *At4g36470.1* (<http://arabidopsis.org/Blast/index.jsp>)

3. *Solyc02g084950.2.1* (<http://solgenomics.net/tools/blast/index.pl>)



Exercise 2 Solutions

4.



http://solgenomics.net/gbrowse/bin/gbrowse/ITAG2.3_genomic/

Exercise 2 Solutions

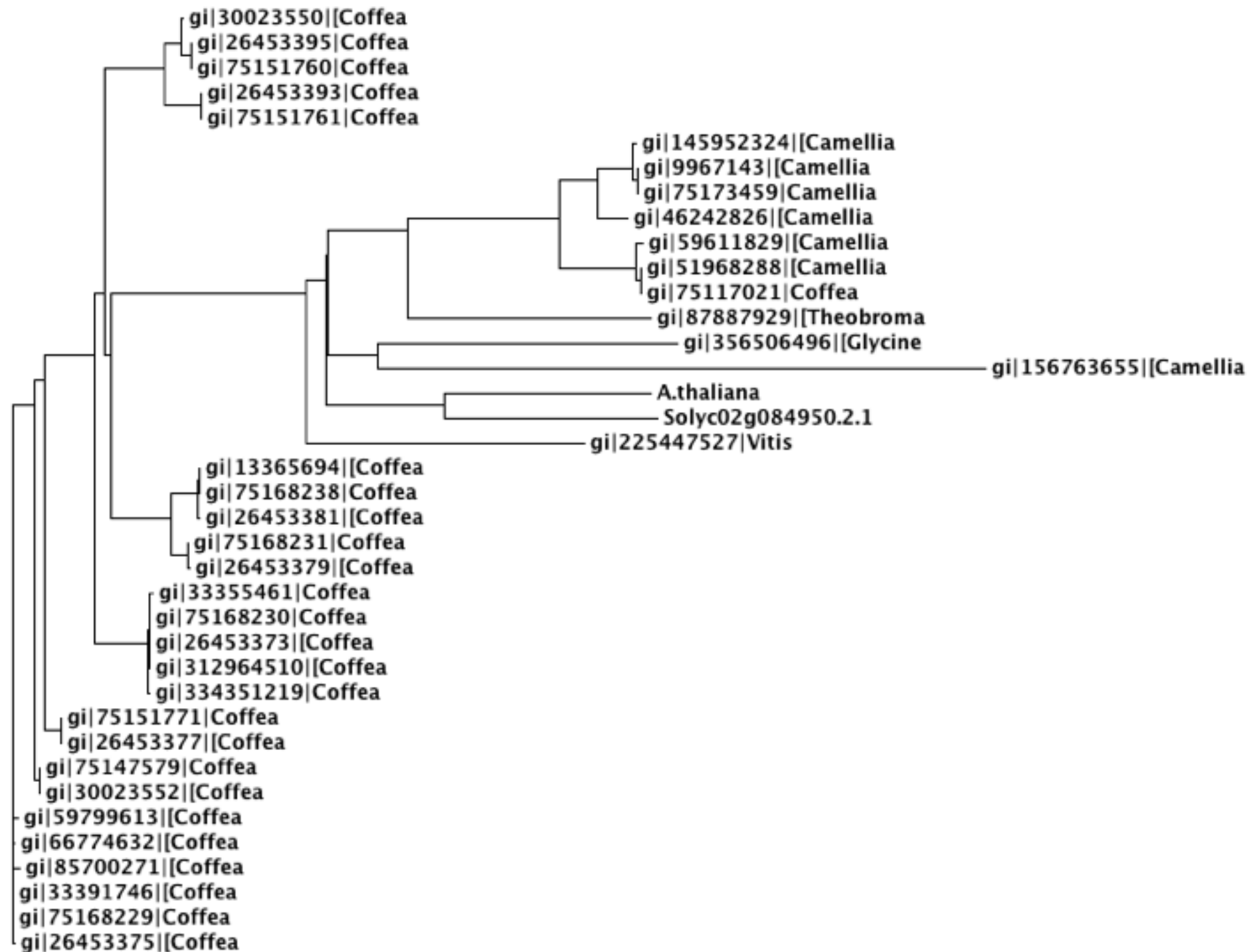
5.

```
gi|9967143|Camellia      ELLAMAIAELVSQGLIDEDK-----LDTFNIPSYFASLEEVKDIVE 282
gi|75173459|Camellia     ELLAMAIAELVSQGLIDEDK-----LDTFNIPSYFASLEEVKDIVE 282
gi|145952324|Camellia    ELLAMAIAELVSQGLIDEDK-----LDTFNIPSYFASLEEVKDIVE 282
gi|46242826|Camellia     -----
gi|51968288|Camellia     ELLAVAIAELVSQGLIDEDK-----LDTFNVPSYFPSLEEVKDIVE 278
gi|75117021|Coffea       ELLAVAIAELVSQGLIDEDK-----LDTFNVPSYFPSLEEVKDIVE 278
gi|59611829|Camellia     ELLAVAIAELVSQGLIDEDK-----LDTFNVPSYFPSLEEVKDIVE 278
gi|87887929|Theobroma    EVLADAISYQVSQGLIDEEK-----LNSFNVPYYIPSQEEVRDLVN 276
gi|26453395|Coffea       DLLEMSINDLVIEGHLEEEK-----LDSFNVPIYAPSTEEVKCIVE 279
gi|75151760|Coffea       DLLEMSINDLVIEGHLEEEK-----LDSFNVPIYAPSTEEVKCIVE 279
gi|30023550|Coffea       DLLEMSINDLVIEGHLEEEK-----LDSFNVPIYAPSTEEVKCIVE 279
gi|26453393|Coffea       DLLEMSINDLVIEGHLEEEK-----LDSFNVPIYAPSTEEVKRIVE 279
gi|75151761|Coffea       DLLEMSINDLVIEGHLEEEK-----LDSFNVPIYAPSTEEVKRIVE 279
gi|33391746|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|75168229|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|26453375|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|85700271|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|66774632|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|59799613|Coffea       DLLDMAINDLIVEGLLEEEK-----LDSFNIPFFTPSAEEVKCIVE 279
gi|75147579|Coffea       DLLDMAINDLIVEGHLEEEK-----LASFNLPPFFTPSAEEVKCIVE 279
gi|30023552|Coffea       DLLDMAINDLIVEGHLEEEK-----LASFNLPPFFTPSAEEVKCIVE 279
gi|75151771|Coffea       DLLDMAINDLIVEGHLEEEK-----LASFNLPPFFTPSAEEVKCIVE 279
gi|26453377|Coffea       DLLDMAINDLIVEGHLEEEK-----LASFNLPPFFTPSAEEVKCIVE 279
gi|75168230|Coffea       DLLEMAINDLVVEGHLEEEK-----LDSFNLPVYIPSAEEVKCIVE 280
gi|26453373|Coffea       DLLEMAINDLVVEGHLEEEK-----LDSFNLPVYIPSAEEVKCIVE 280
gi|312964510|Coffea      DLLEMAINDLVVEGHLEEEK-----LDSFNLPVYIPSAEEVKCIVE 280
gi|33355461|Coffea       DLLEMAINDLVVEGHLEEEK-----LDSFNLPVYIPSAEEVKCIVE 280
gi|334351219|Coffea      DLLEMAINDLVVEGHLEEEK-----LDSFNLPVYIPSAEEVKCIVE 280
gi|13365694|Coffea       DLLEMAINDLVAEGRLGEEK-----LDSFNVPIYTASVEEVKCMVE 280
gi|75168238|Coffea       DLLEMAINDLVAEGRLGEEK-----LDSFNVPIYTASVEEVKCMVE 280
gi|26453381|Coffea       DLLEMAINDLVAEGRLGEEK-----LDSFNVPIYTASVEEVKCMVE 280
gi|75168231|Coffea       DLLEMAINDLVVEGHLEEEK-----LDSFNVPIYAASVEELKCIVE 280
gi|26453379|Coffea       DLLEMAINDLVVEGHLEEEK-----LDSFNVPIYAASVEELKCIVE 280
gi|225447527|Vitis       ELISMTLKDMVTEGSIQESE-----LDSLNIPLFMPSPEQVRSVIQ 281
A.thaliana              ELLSRSIADLVAQGETEEEEK-----LDSYDMHFYAPSADEIEGEVD 285
Solyc02g084950.2.1      KILYQALTNLISKGEVEKEK-----LDSYEVHFFYAPCKEEIEKVAR 288
gi|356506496|Glycine     EVISMVLNDKVQEGLEEER-----LDFFNLRRYRHSAAEEIRQVVE 242
gi|156763655|Camellia    SIAERSGVPVIADGGISNSGXIVKALVLGASTVMMGSFLAGSNEAPGAYE 392
```



6.

Exercise 2 Solutions



When using web tools remember:

- 1.) Often not all program options are available
- 2.) Jobs are run on another server, large jobs may be better run locally

1. slides at ftp://ftp.solgenomics.net/bioinfo_class/intro_2014/Bioinfo_interns_6_2014_partII.pdf
2. If you are interested in hands on command line training, come to the advanced bioinformatics sessions:
 1. Advanced Course 1 -- based on VM (Noe), June 24
 1. VM installation
 2. Command line Linux
 2. Advanced Course 2 (Mueller), July 1
 1. Command line Linux
 2. Fasta, Genbank, etc. formats -
 3. running blastall
 3. Advanced Course 3 (Surya) July 8
 1. NextGen? seq
 2. fastq
 3. qc with fastqc
 4. Advanced RNASeq 4 (Naama) July 15
 1. align reads (bowtie2)
 2. file manipulation with samtools, bed tools
 5. Advanced 5 (Suzy) July 22
 1. Tuxedo suite
3. You will need to have a virtual machine installed prior to next class

<http://btiplantbioinfocourse.wordpress.com/3rd-bti-bioinformatic-course-2014/installing-the-virtual-machine/>