sol genomics network

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/interns/2015

Boyce Thompson Institute for Plant Research

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.

1. Types.

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

- Public repositories with massive data storage.

- Community-specific databases.

- Project-specific databases.

1. Types.

- ✴ 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.

1. Types.

sol genomics network

- 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

1. Types.

sol genomics network

- 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.

1. Types.

- ✳ 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.

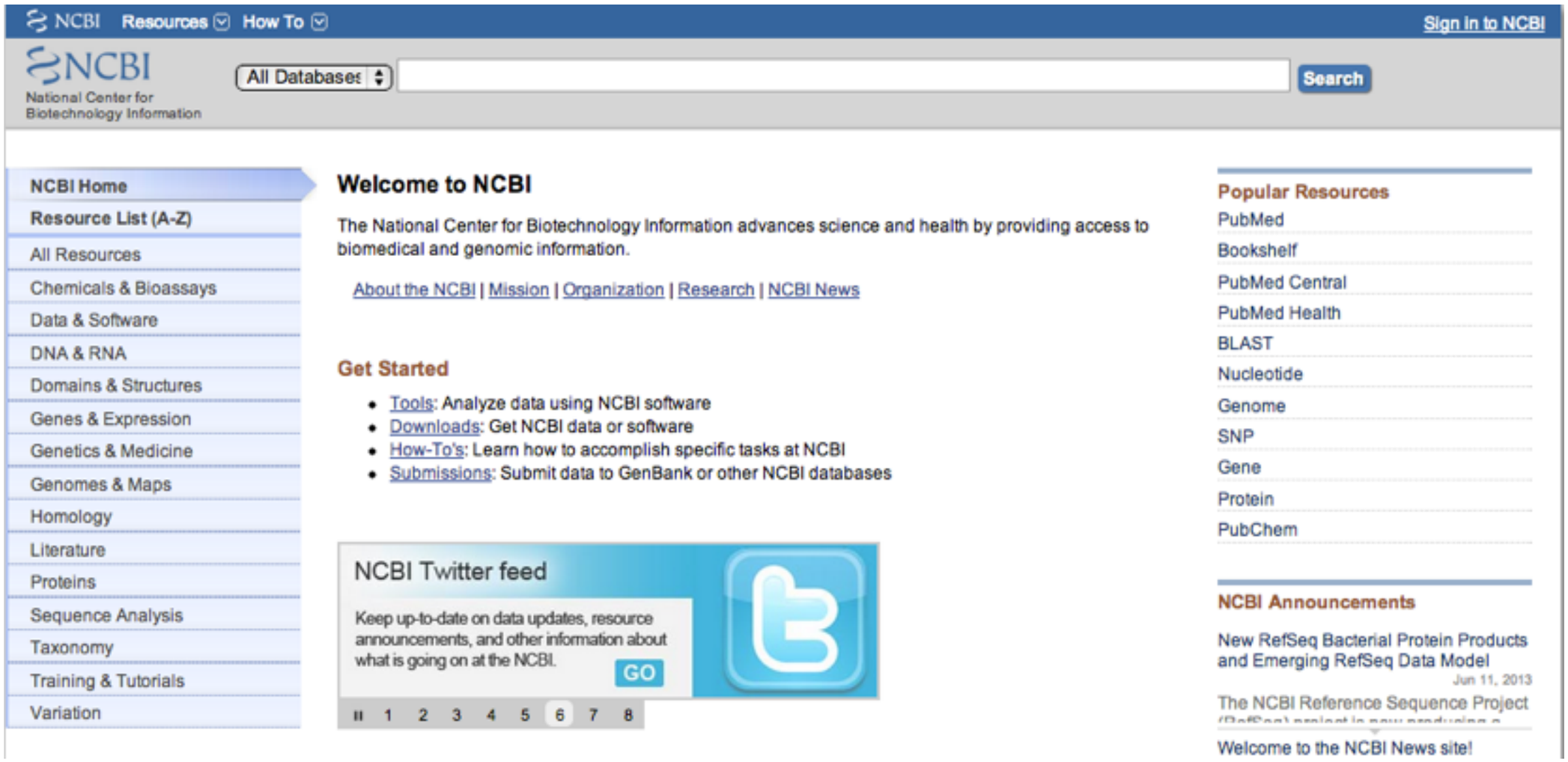
4- Genomic Browsers.

2. Public Repositories.

sol genomics network

NCBI (National Center for Biotechnology Information)

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



2. Public Repositories.

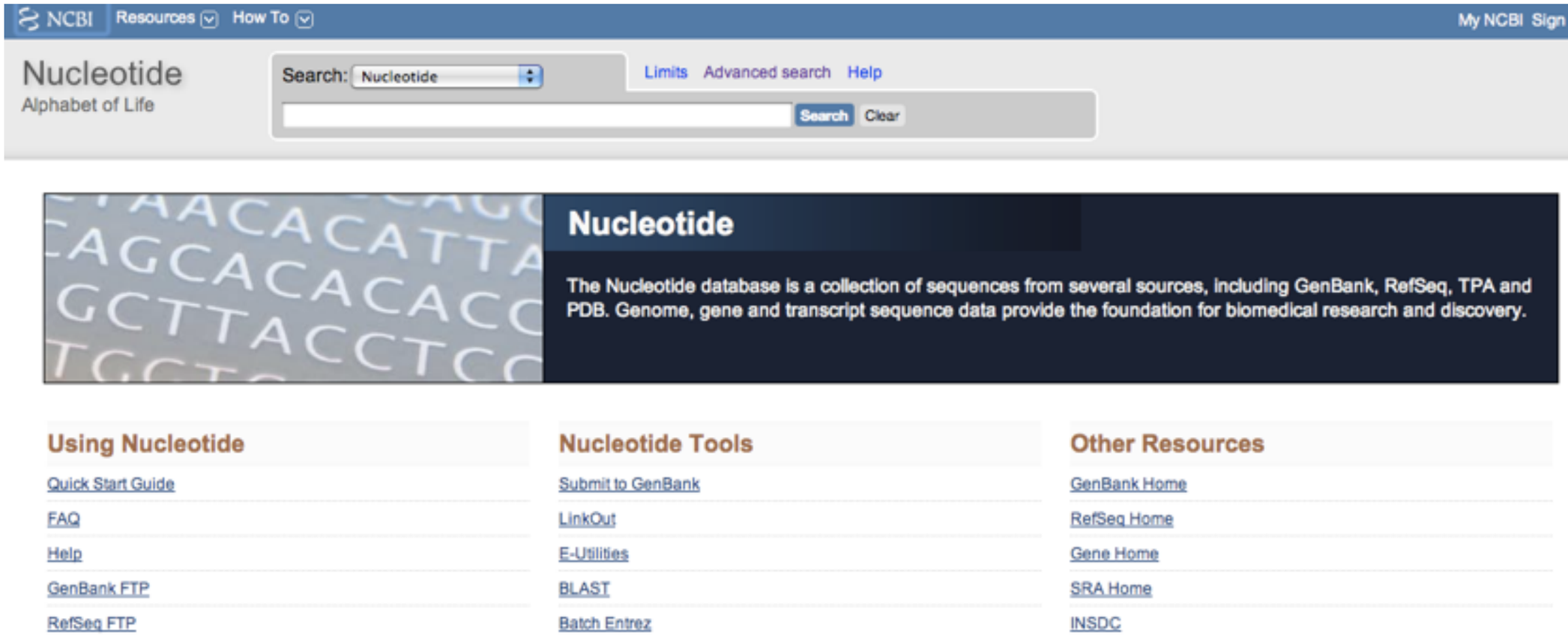
NCBI (National Center for Biotechnology Information)
http://www.ncbi.nlm.nih.gov/

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

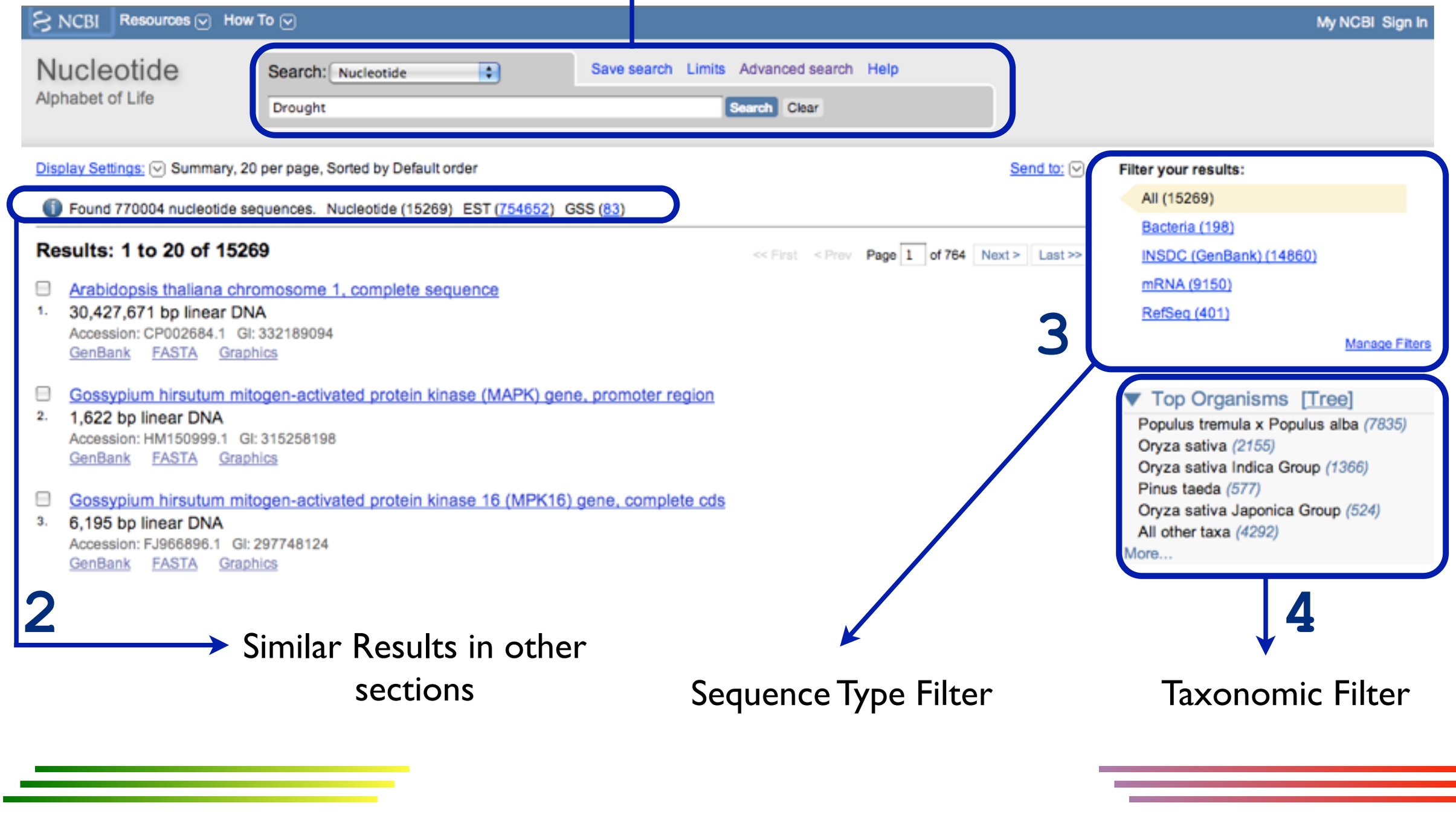


2. Public Repositories: NCBI

sol genomics network

GenBank:

1 → Search Section



2 → Similar Results in other sections

Sequence Type Filter

Taxonomic Filter

2. Public Repositories: NCBI

sol genomics network

GenBank:

1 → Filter application box

Nucleotide
Alphabet of Life

Search: Nucleotide | Save search | Limits | Advanced search | Help

((Drought) AND "seed plants"[porgn:__txid58024]) AND "Solanales"[po | 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 >>

1. Capsicum annuum chitinase class II (CAChi2) mRNA, complete cds
   1,004 bp linear mRNA
   Accession: AF091235.1 GI: 3641354
   GenBank FASTA Graphics Related Sequences

2. Capsicum annuum stellacyanin-like protein CASLP1 precursor, mRNA, complete cds
   937 bp linear mRNA
   Accession: AF291179.1 GI: 9885805
   GenBank FASTA Graphics

3. Nicotiana attenuata lipid transfer protein 1-like (LTP1) mRNA, partial sequence
   672 bp linear mRNA
   Accession: HM068895.1 GI: 298155395
   GenBank FASTA Graphics Related Sequences

4. Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence
   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: Select

Find items

2. Public Repositories: NCBI

GenBank:

Tools Links

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

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
```

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

2. Public Repositories: NCBI

GenBank:

Format

File Storage

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

Display Settings: GenBank

Send:

Change region shown

Customize view

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

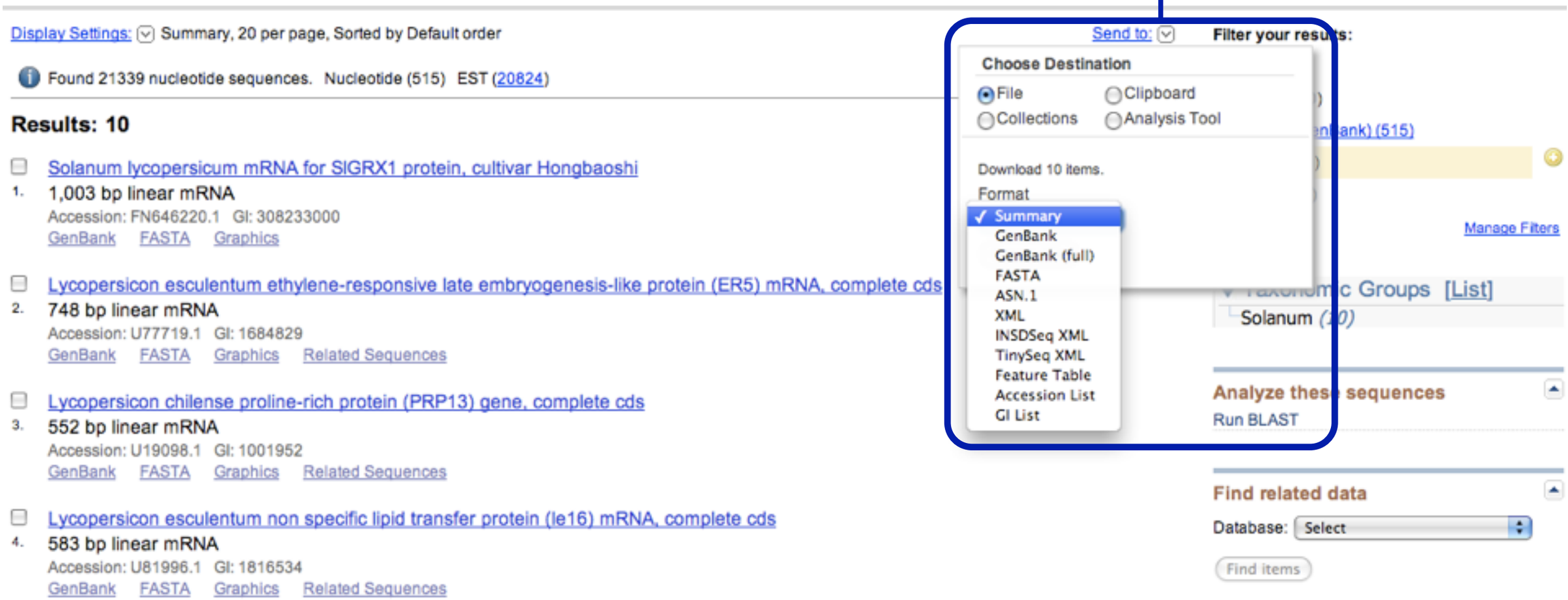
Turn Off Clear

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

2. Public Repositories: NCBI

GenBank:

File Storage



2. Public Repositories: NCBI

sol genomics network

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 | RSS | Save search | Limits | Advanced search | Help

drought tomato | 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 >>

1. 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

2. 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

3. 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).

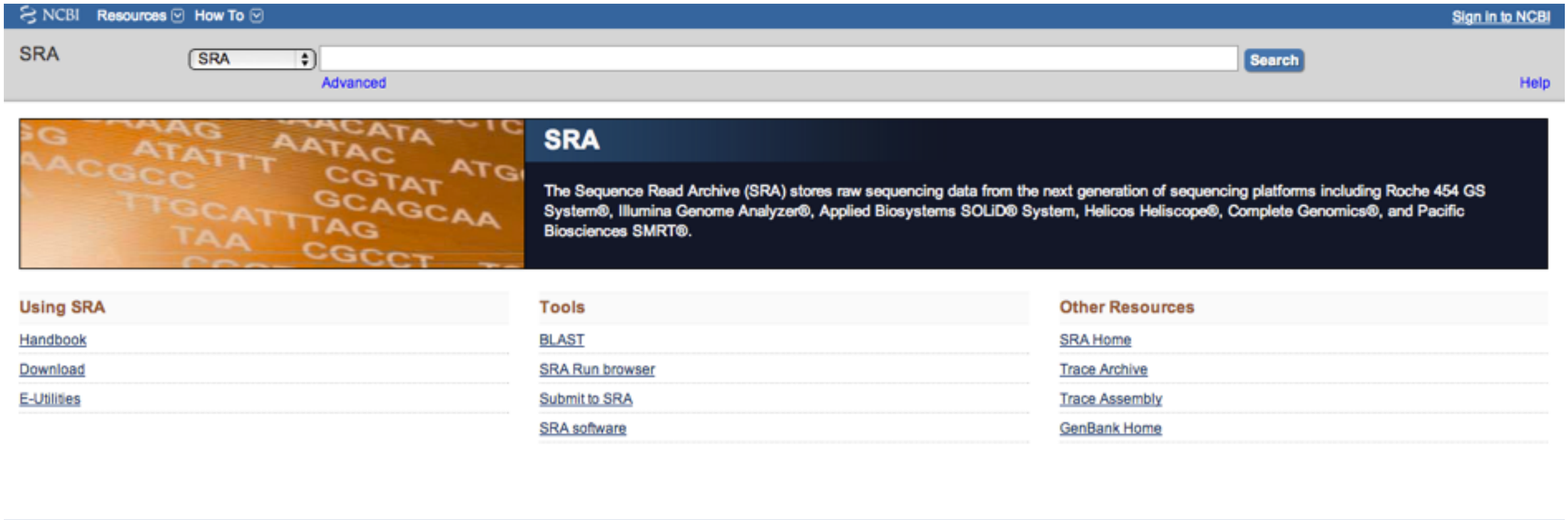
- It doesn't have all plant science related journals (for example: Theoretical Applied and Genetics or Crop Science).

  (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.

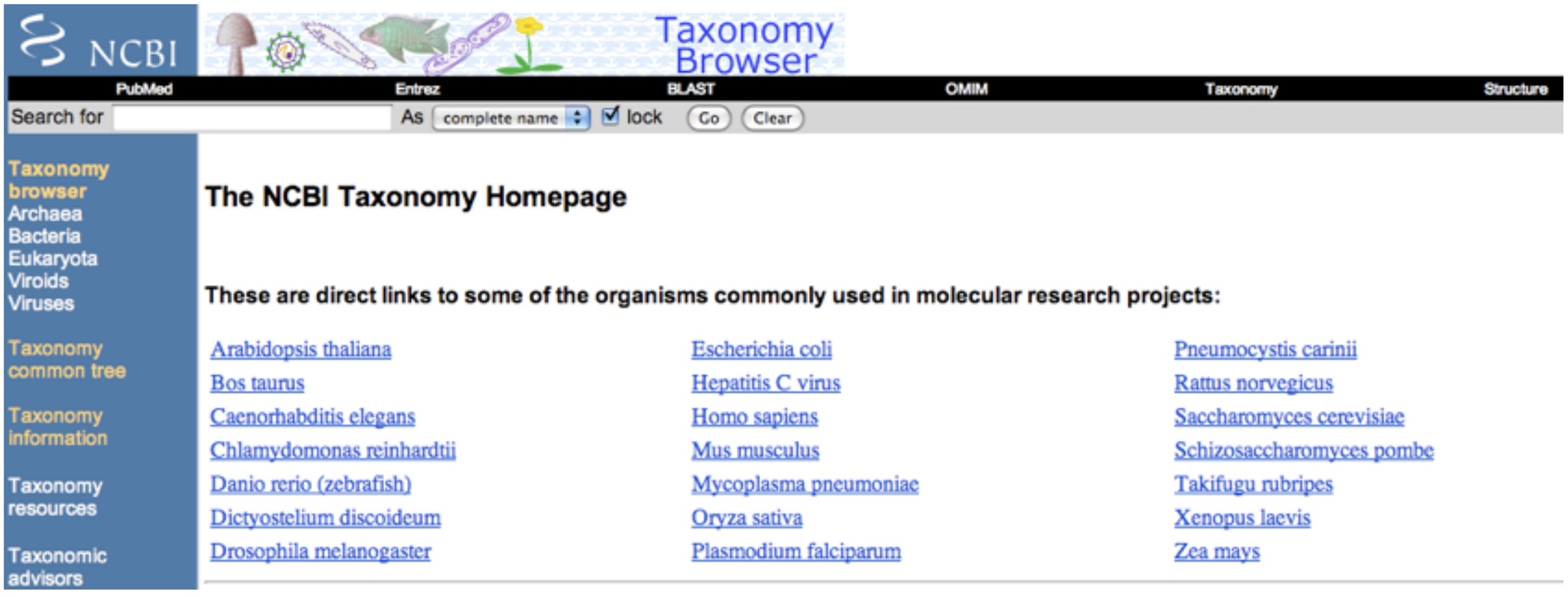
2. Public Repositories: NCBI

Sequence Read Archive (SRA), Database to store sequences produced by NGS such as Illumina, 454, Solid, Helicos...
http://www.ncbi.nlm.nih.gov/sra



2. Public Repositories: NCBI

Taxonomy, database with taxonomic names and classifications for NCBI organisms.
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

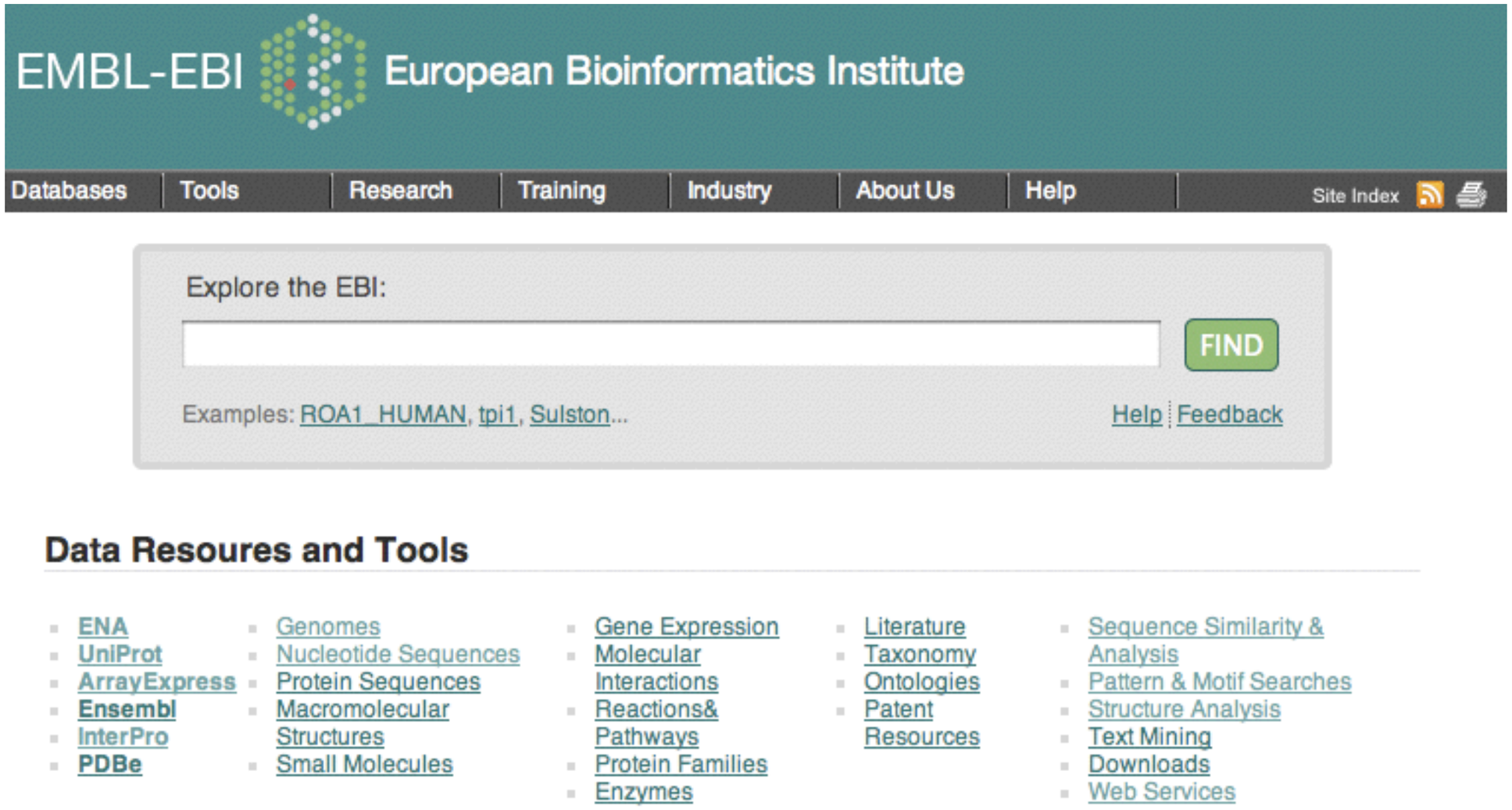


2. Public Repositories: EBI

sol genomics network

EBI (European Bioinformatics Institute)

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



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/).

sol genomics network

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

sol genomics network

Name	Species	Data	Link
The Arabidopsis Information Resource (TAIR)	*Arabidopsis*	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/
Curcurbit Genomics Database (ICuGI)	Cucurbitaceae	Multiple Sequence Genomes, Genes, Unigenes,Genetic Markers, Genetic Maps, Metabolic Pathways	http://www.icugi.org/

3. Community specific databases

sol genomics network

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/

sol genomics network

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

sol genomics network

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

sol genomics network

Genome Browser most used:

- JBrowse (GMOD).

- GBrowse (GMOD).

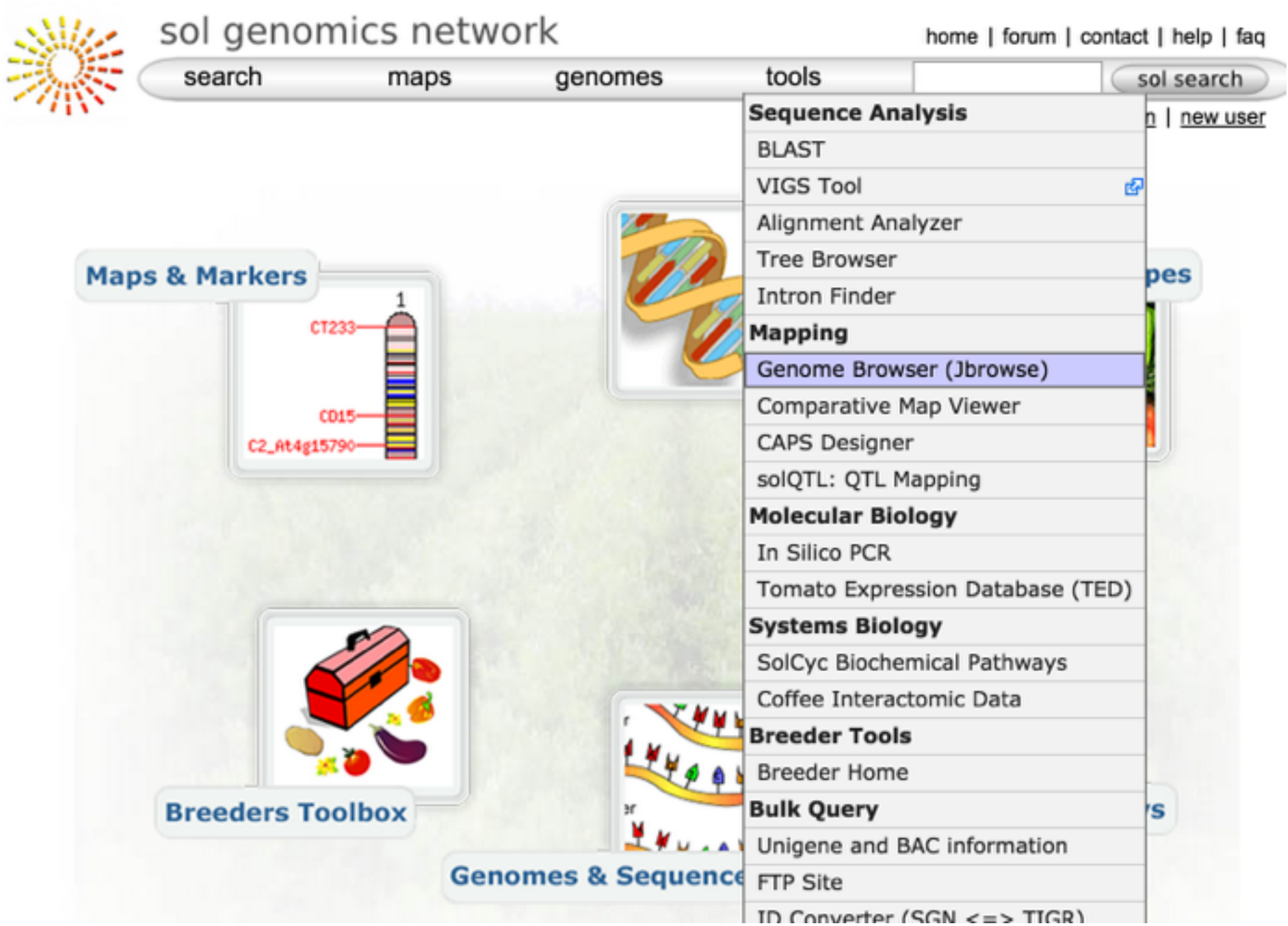
- UCSC Genome Browser.

- Emsembl Genome Browser.

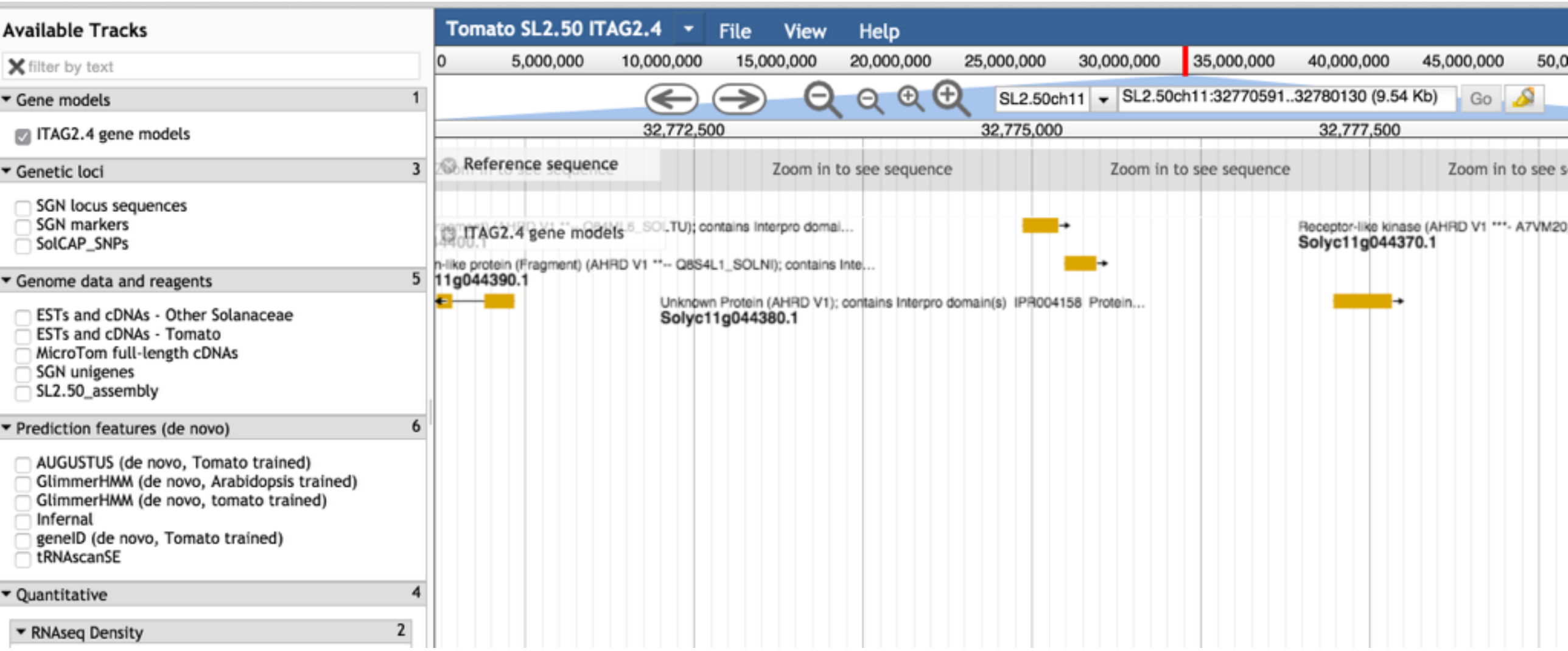
- Vista Genome Browser.

JBrowse

http://solgenomics.net/



JBrowse


Available Tracks
filter by text
Gene models 1
ITAG2.4 gene models
Genetic loci 3
SGN locus sequences
SGN markers
SolCAP_SNPs
Genome data and reagents 5
ESTs and cDNAs - Other Solanaceae
ESTs and cDNAs - Tomato
MicroTom full-length cDNAs
SGN unigenes
SL2.50_assembly
Prediction features (de novo) 6
AUGUSTUS (de novo, Tomato trained)
GlimmerHMM (de novo, Arabidopsis trained)
GlimmerHMM (de novo, tomato trained)
Infernal
geneID (de novo, Tomato trained)
tRNAscanSE
Quantitative 4
RNAseq Density 2
Tomato SL2.50 ITAG2.4
File
View
Help
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000
SL2.50ch11
SL2.50ch11:32770591..32780130 (9.54 Kb)
Go
32,772,500
32,775,000
32,777,500
Reference sequence
Zoom in to see sequence
Zoom in to see sequence
ITAG2.4 gene models
Receptor-like kinase (AHRD V1 ***- A7VM20
Solyc11g044370.1
11g044390.1
Unknown Protein (AHRD V1); contains Interpro domain(s) IPR004158 Protein...
Solyc11g044380.1

JBrowse

sol genomics network

home | forum | contact | help | faq

search | maps | genomes | tools | sol search

log in | new user

***Solanum lycopersicum* gene Solyc11g020230.1**

Gene

Name	Type	Organism
Solyc11g020230.1	Gene	*Solanum lycopersicum*

Synonyms / aliases	Length
Solyc11g020230	2,337

Description

Pto-like, Serine/threonine kinase protein, resistance protein

Location(s)

SL2.50ch11:10541737..10544073

GFF source

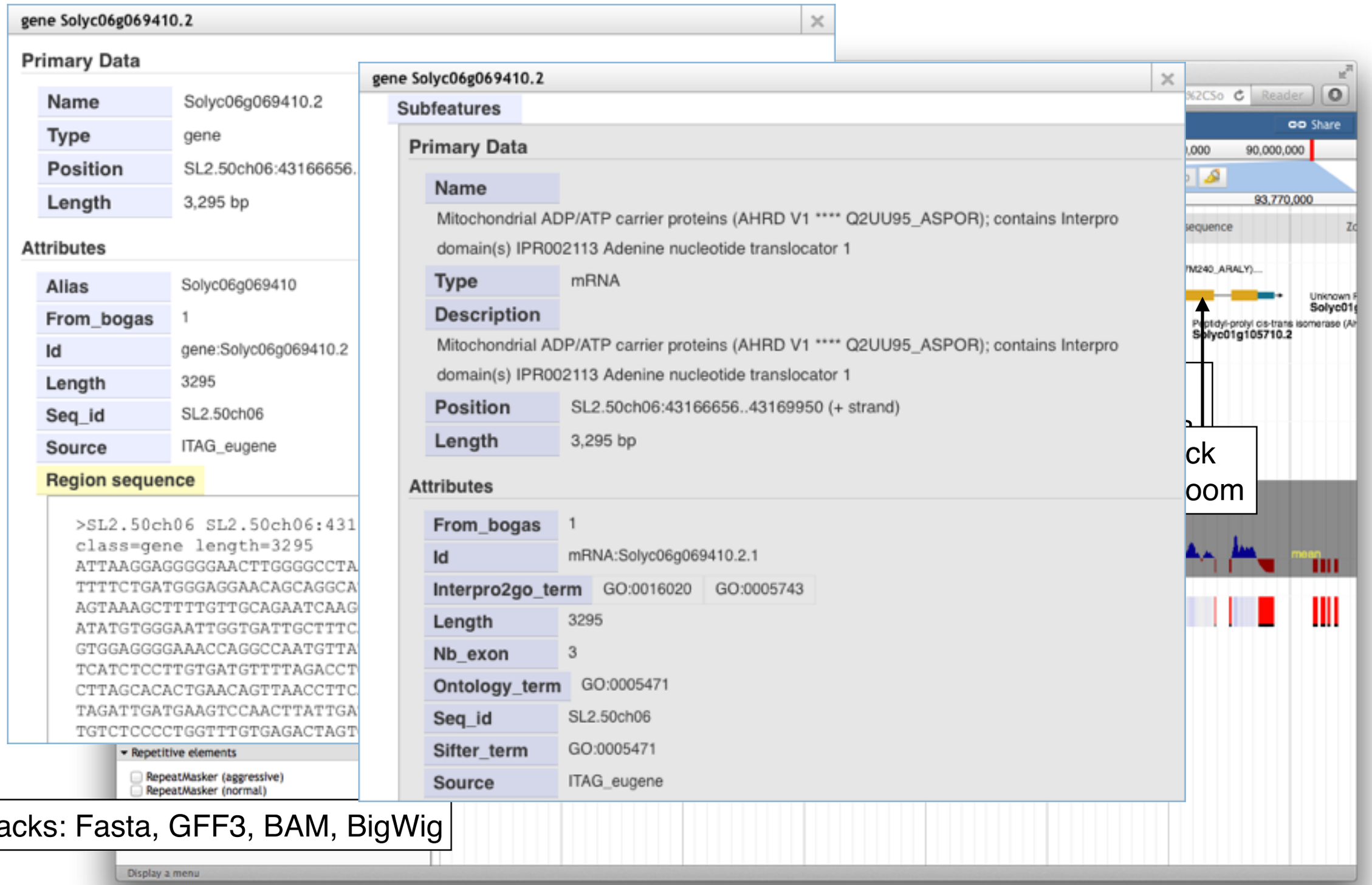
ITAG_eugene

From BOGAS	Length
1	2337

Genome Viewer

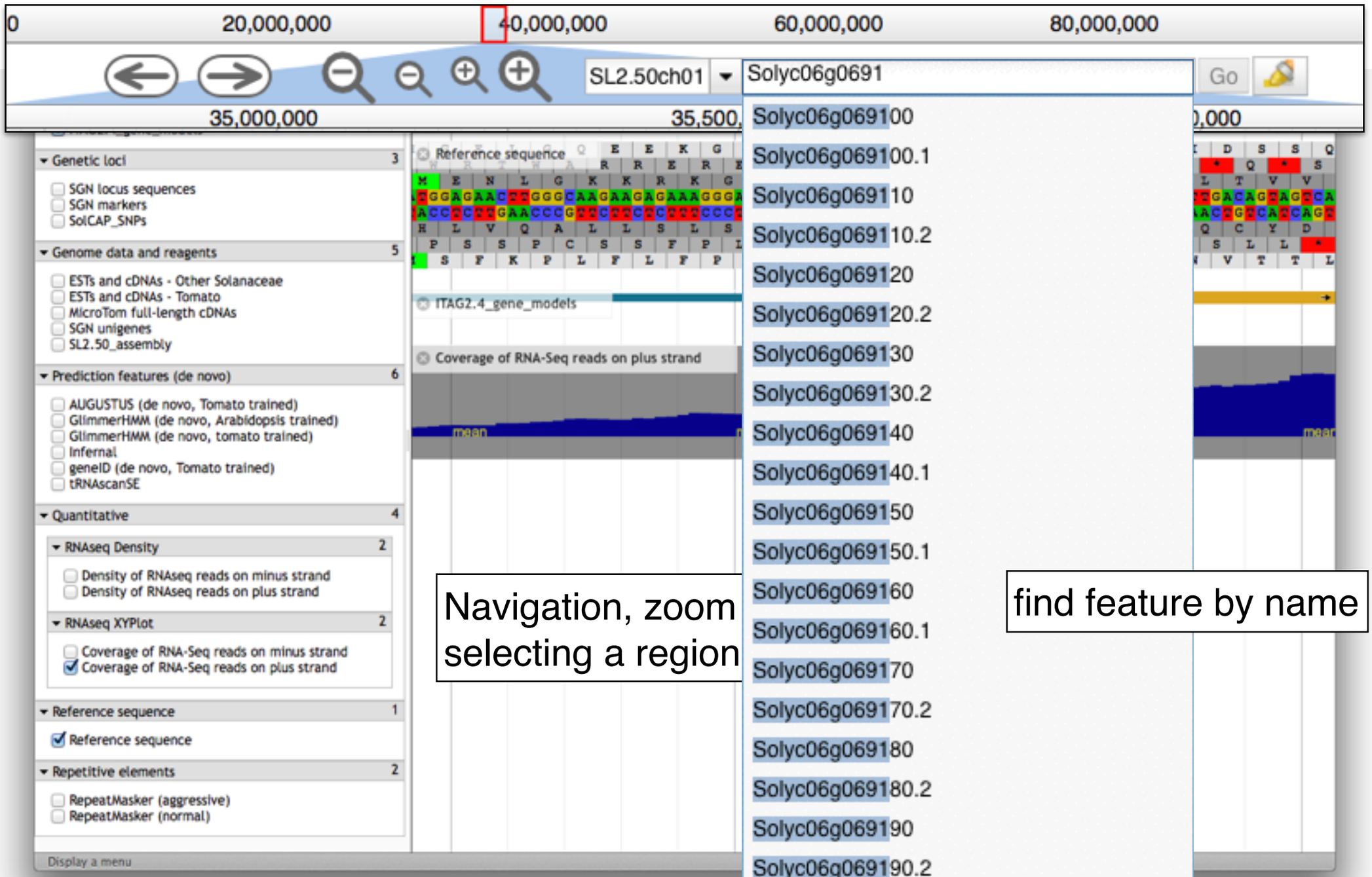
Tomato SL2.50 ITAG2.4 | File | View | Help | Full-screen view

JBrowse

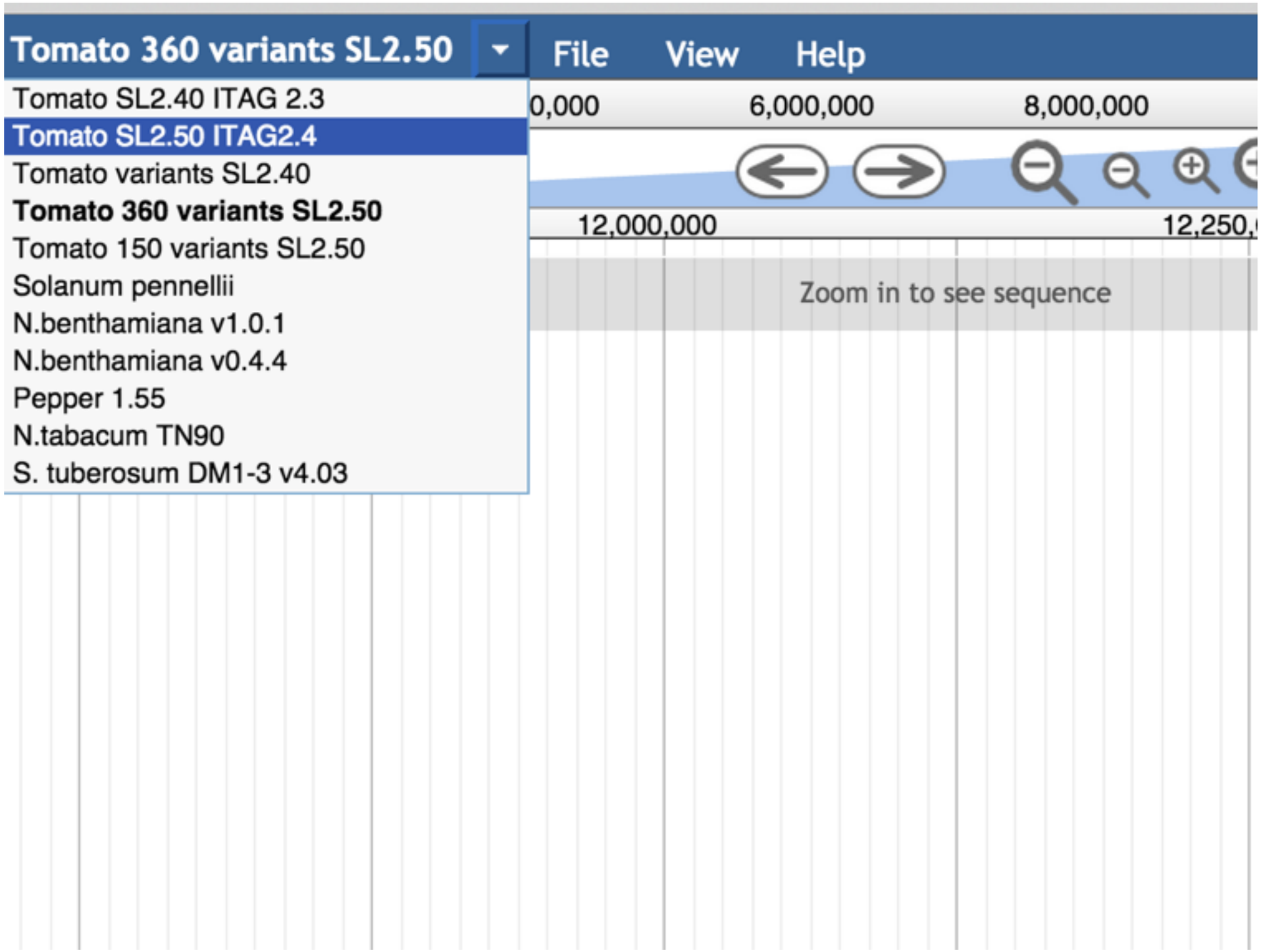


load tracks: Fasta, GFF3, BAM, BigWig

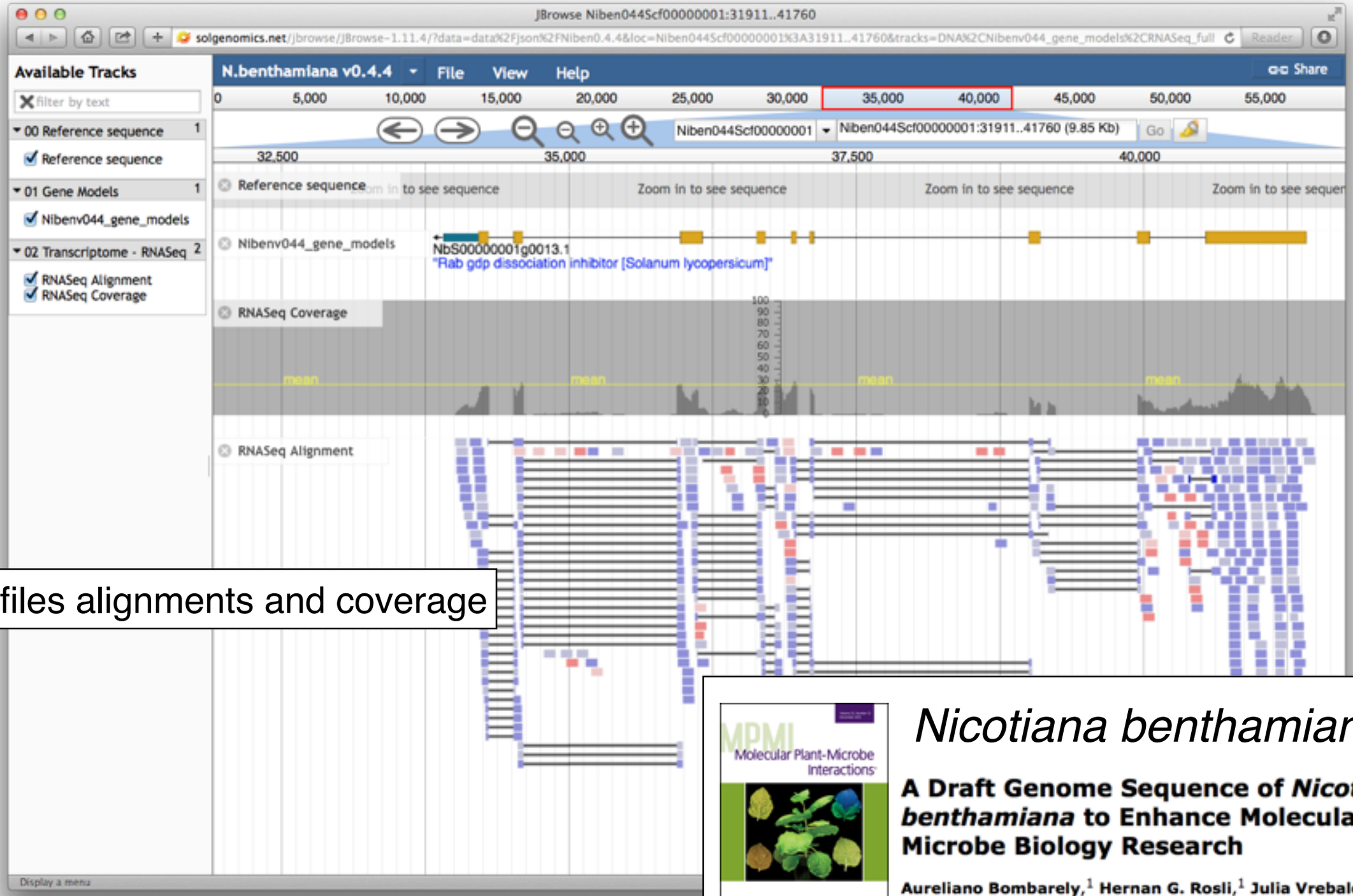
JBrowse


0
20,000,000
40,000,000
60,000,000
80,000,000
SL2.50ch01
Solyc06g0691
Go
35,000,000
Genetic loci 3
SGN locus sequences
SGN markers
SolCAP_SNPs
Genome data and reagents 5
ESTs and cDNAs - Other Solanaceae
ESTs and cDNAs - Tomato
MicroTom full-length cDNAs
SGN unigenes
SL2.50_assembly
Prediction features (de novo) 6
AUGUSTUS (de novo, Tomato trained)
GlimmerHMM (de novo, Arabidopsis trained)
GlimmerHMM (de novo, tomato trained)
Infernal
geneID (de novo, Tomato trained)
tRNAscanSE
Quantitative 4
RNAseq Density 2
Density of RNAseq reads on minus strand
Density of RNAseq reads on plus strand
RNAseq XYPlot 2
Coverage of RNA-Seq reads on minus strand
Coverage of RNA-Seq reads on plus strand
Reference sequence 1
Reference sequence
Repetitive elements 2
RepeatMasker (aggressive)
RepeatMasker (normal)
Display a menu
Reference sequence
ITAG2.4_gene_models
Coverage of RNA-Seq reads on plus strand
Solyc06g069100
Solyc06g069100.1
Solyc06g069110
Solyc06g069110.2
Solyc06g069120
Solyc06g069120.2
Solyc06g069130
Solyc06g069130.2
Solyc06g069140
Solyc06g069140.1
Solyc06g069150
Solyc06g069150.1
Solyc06g069160
Solyc06g069160.1
Solyc06g069170
Solyc06g069170.2
Solyc06g069180
Solyc06g069180.2
Solyc06g069190
Solyc06g069190.2
Navigation, zoom selecting a region
find feature by name

JBrowse


Tomato 360 variants SL2.50
File
View
Help
Tomato SL2.40 ITAG 2.3
Tomato SL2.50 ITAG2.4
Tomato variants SL2.40
Tomato 360 variants SL2.50
Tomato 150 variants SL2.50
Solanum pennellii
N.benthamiana v1.0.1
N.benthamiana v0.4.4
Pepper 1.55
N.tabacum TN90
S. tuberosum DM1-3 v4.03
0,000
6,000,000
8,000,000
12,000,000
12,250,
Zoom in to see sequence

JBrowse: *Nicotiana benthamiana*



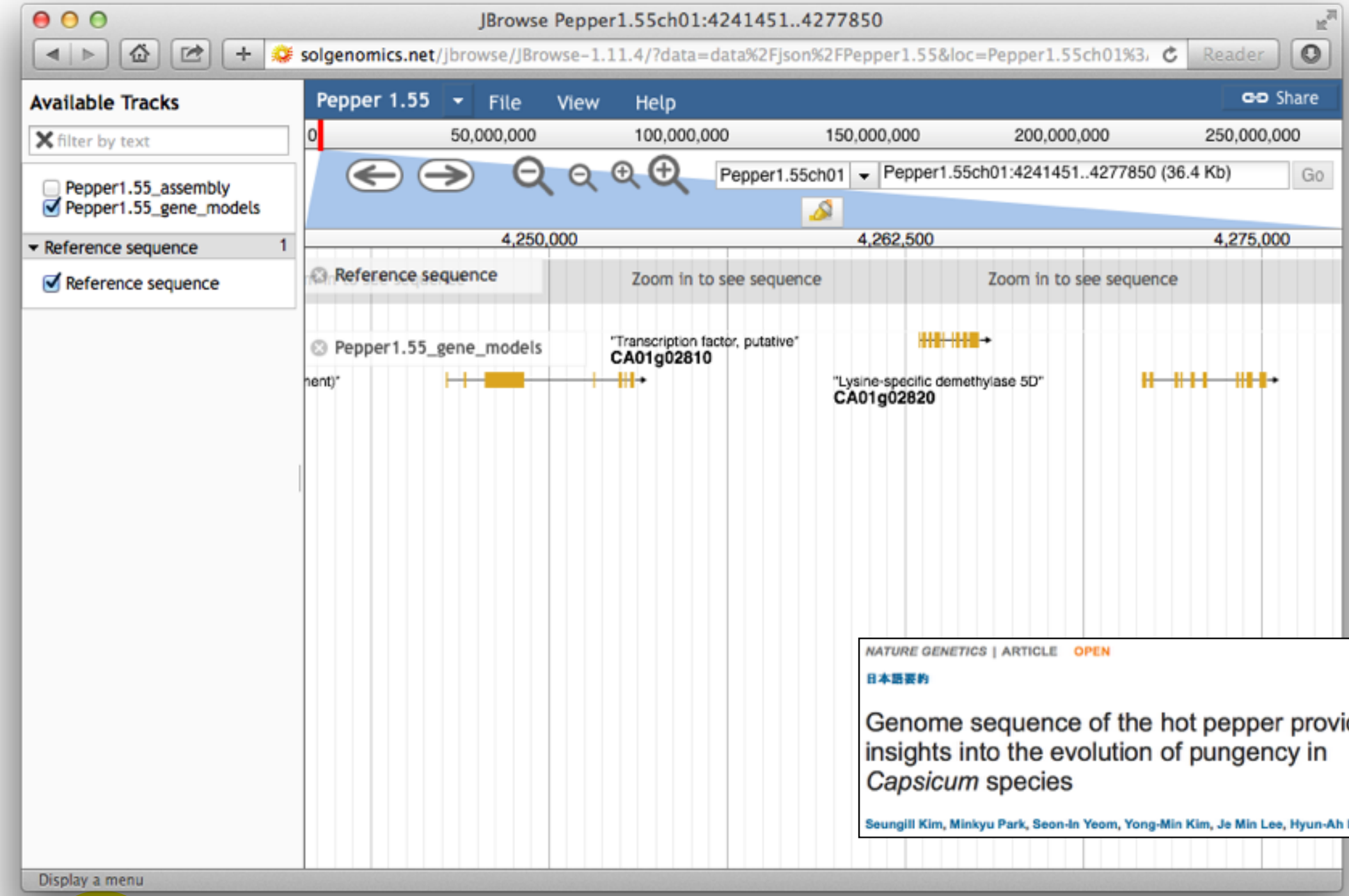
bam files alignments and coverage

Nicotiana benthamiana

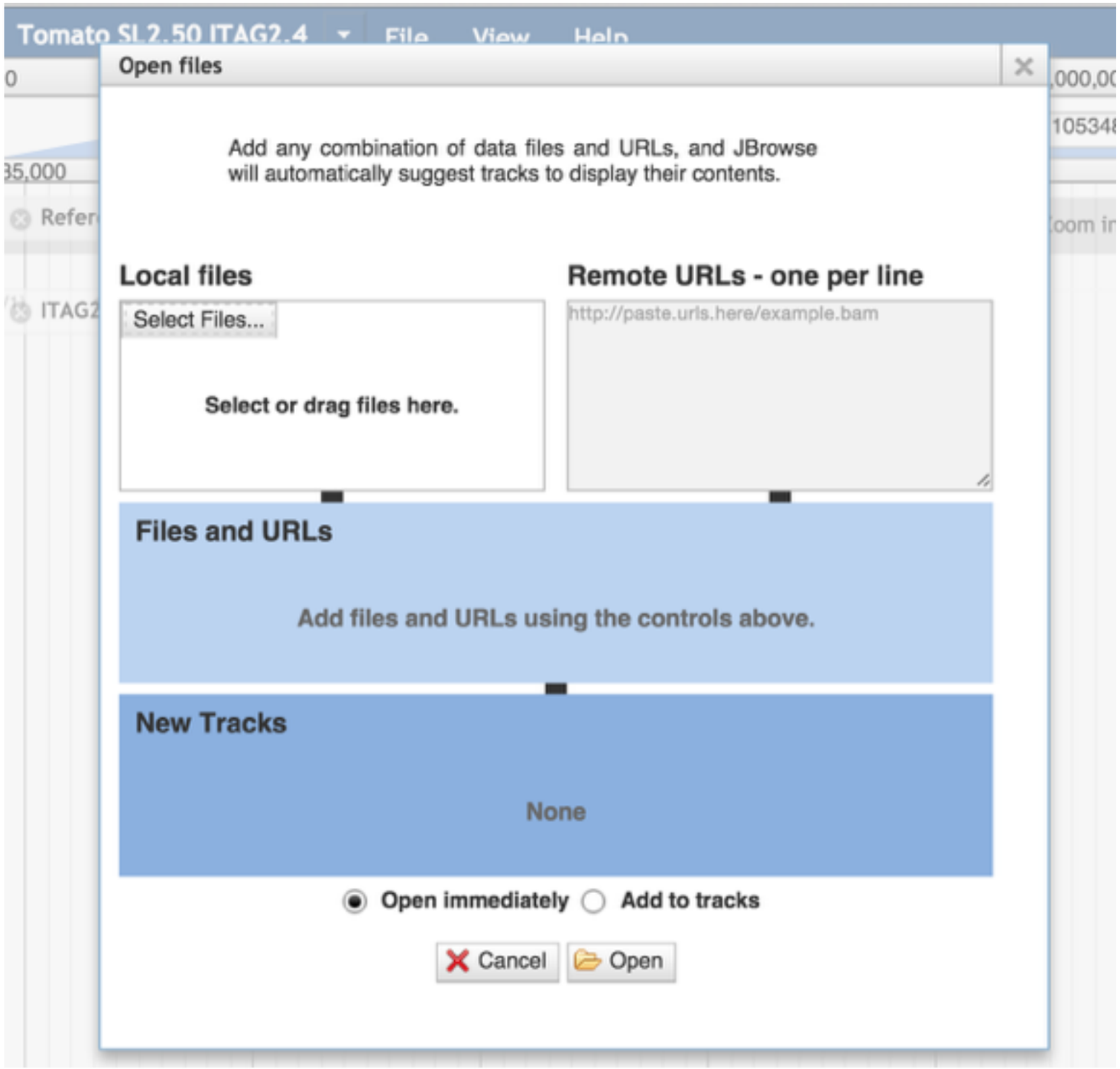
A Draft Genome Sequence of *Nicotiana benthamiana* to Enhance Molecular Plant-Microbe Biology Research

Aureliano Bombarely,[1] Hernan G. Rosli,[1] Julia Vrebalov,[1] Peter Moffett,[1,2] Lukas A. Mueller,[1] and Gregory B. Martin[1,3,4]

JBrowse: Pepper genome



JBrowse


Tomato SL 2.50 ITAG2.4
File
View
Help
Open files
Add any combination of data files and URLs, and JBrowse will automatically suggest tracks to display their contents.
Local files
Select Files...
Select or drag files here.
Remote URLs - one per line
http://paste.urls.here/example.bam
Files and URLs
Add files and URLs using the controls above.
New Tracks
None
Open immediately
Add to tracks
Cancel
Open

4. Genomic Browsers

sol genomics network

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 1

sol genomics network

1. 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 species 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* homolog 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 1 Solutions

1. 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.

1. Identification and isolation of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from guarana (Paullinia cupana).
   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

2. Producing low-caffeine tea through post-transcriptional silencing of **caffeine synthase** mRNA.
   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

3. Agrobacterium-mediated silencing of caffeine synthesis through root transformation in Camellia sinensis L.
   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

4. A transcriptomic approach highlights induction of secondary metabolism in citrus fruit in response to Penicillium digitatum infection.
   González-Candelas 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

5. Essential region for 3-N methylation in N-methyltransferases involved in caffeine biosynthesis.
   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

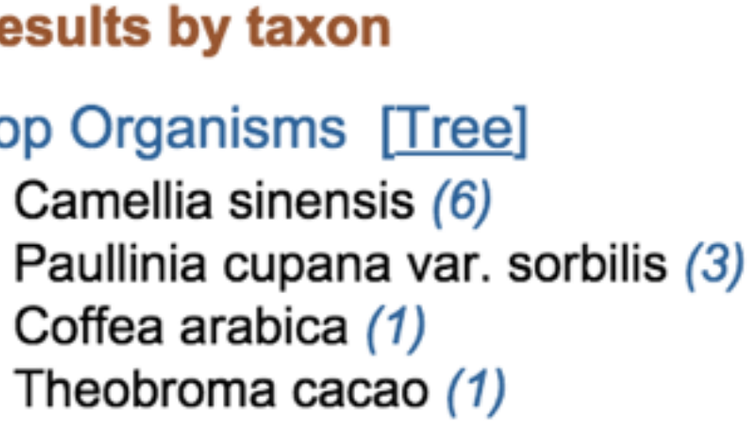
6. Expression for caffeine biosynthesis and related enzymes in Camellia sinensis.
   Kato M, Kitao N, Ishida M, Morimoto H, Irino F, Mizuno K.
   Z Naturforsch C. 2010 Mar-Apr;65(3-4):245-56.
   PMID: 20469645 [PubMed - indexed for MEDLINE]
   Related citations

Exercise 1 Solutions

2. 11 proteins total, 1 from *C. arabica* (http://www.ncbi.nlm.nih.gov/protein)

Exercise 1 Solutions

3. 4 species:



*click on “Tree” for next question

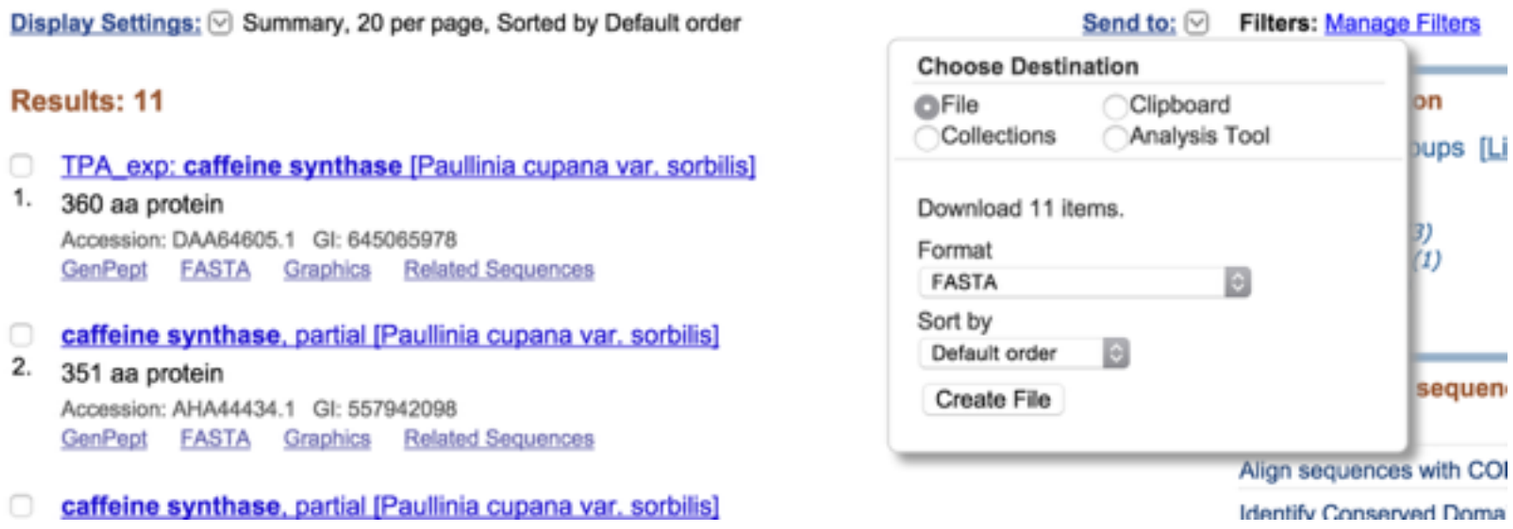
Exercise 1 Solutions (cont'd)

4. Found in Gentianales, Ericales, Sapindales, Malvales

*click on "List" for next answer

Exercise 1 Solutions (cont'd)

5. *Coffea arabica, Coffea canephora, Camellia sinensis, Theobroma cacao, Paullinia.* 7 sequences appear to be full-length.



7.) http://metacyc.org/

Add to group

MetaCyc Enzyme:

Gene: CaDXMT1 Accession Number: G-9084 (MetaCyc)

Species: Coffea arabica

Summary:

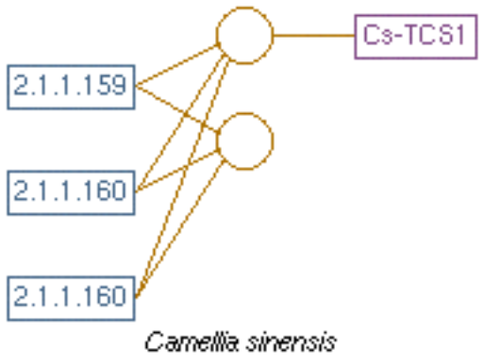
The recombinant CaDXMT1 catalyzes the conversion of 7-methylxansine to theobromine and theobromine to caffeine. In addition, Ca
the most preferred substrate. XMP is not an effective substrate. CaDXMT1 is predominantly expressed in immature fruits of coffee.

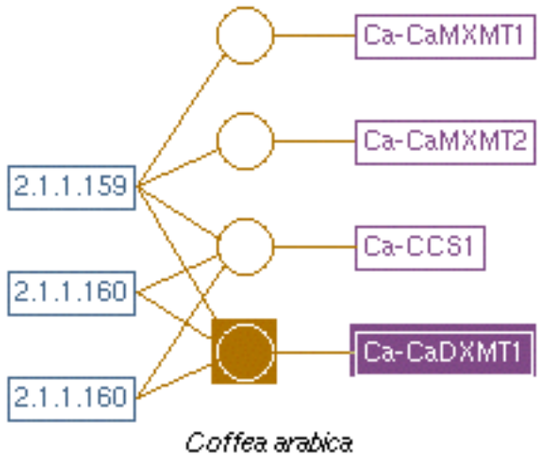
Citations: [Uefuji03]

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

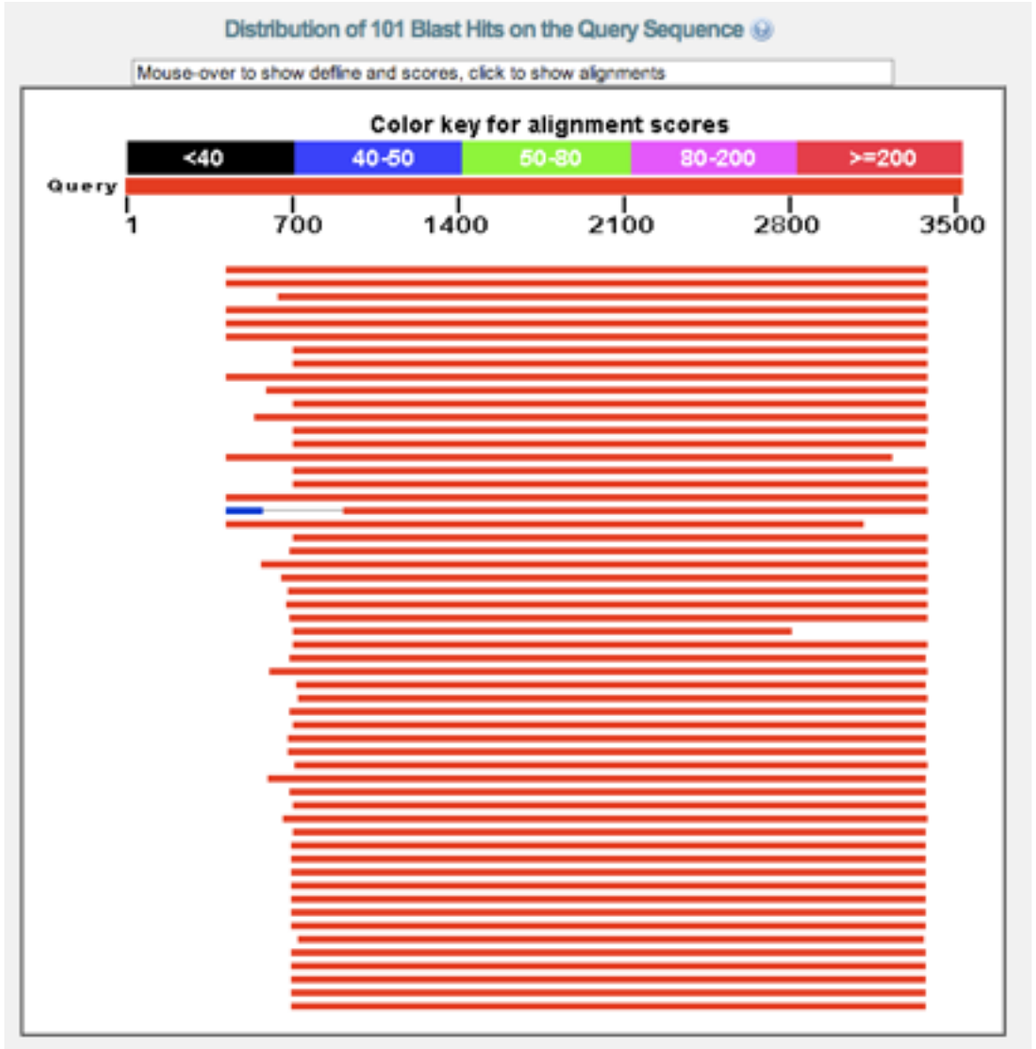
Relationship Links: Entrez-Nucleotide:PART-OF:AB084125

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.

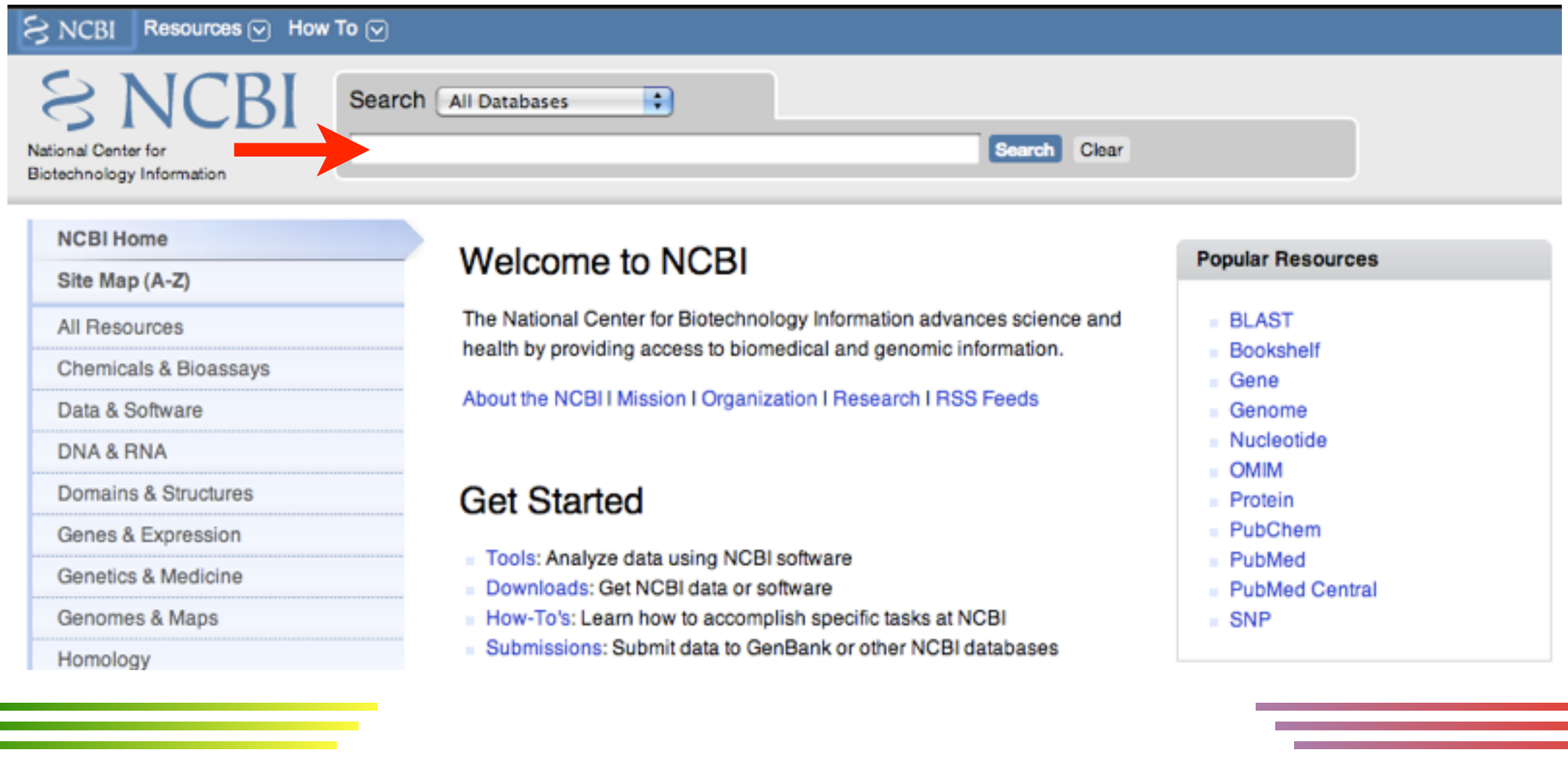
1. Search Tools

sol genomics network

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/

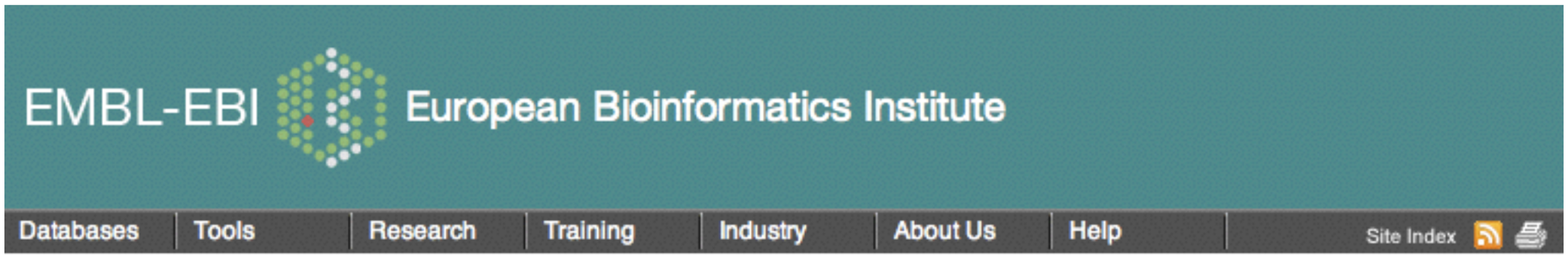


1. Search Tools

sol genomics network

Text Searches:

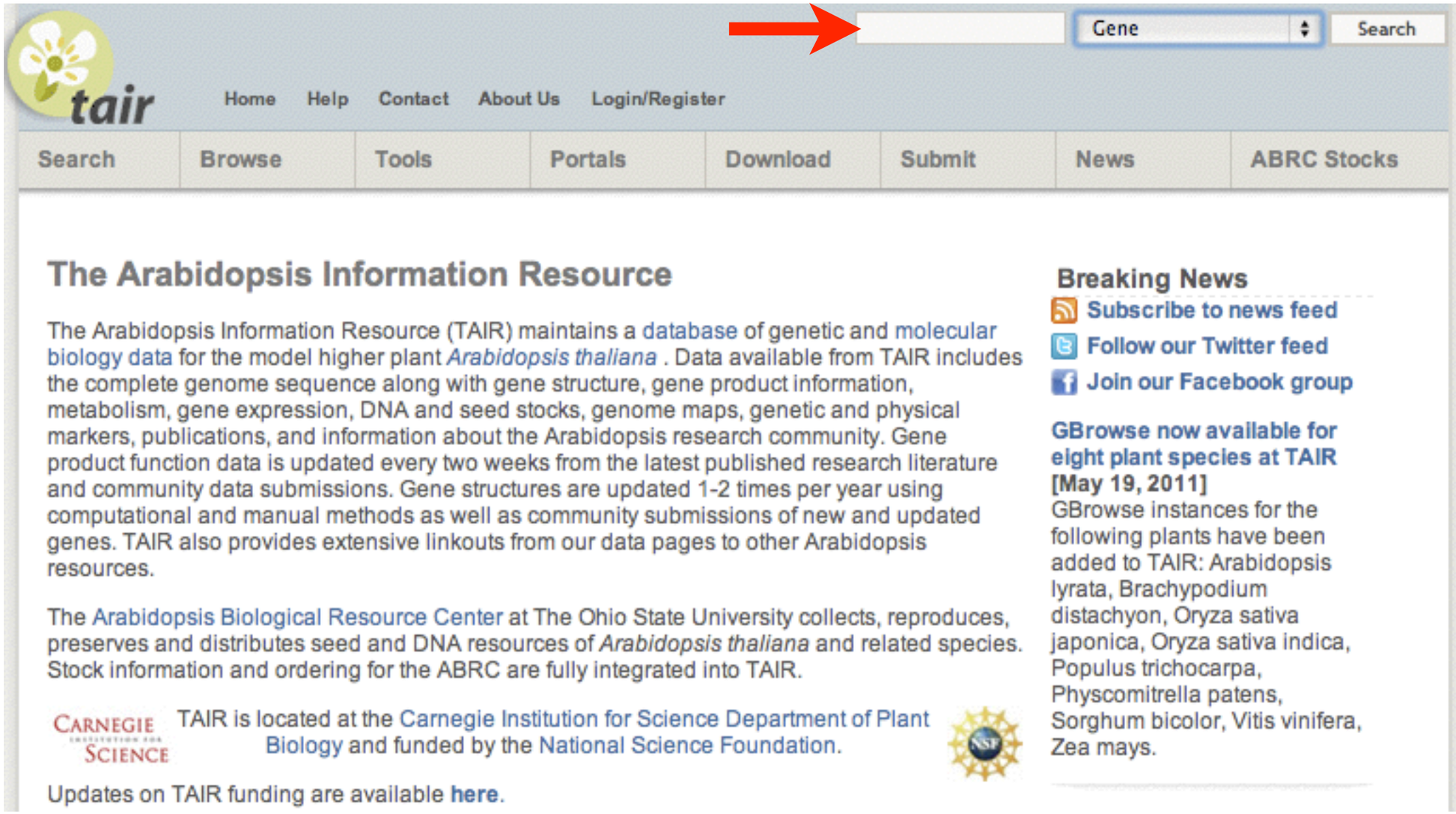
EBI: http://www.ebi.ac.uk/



1. Search Tools

Text Searches:

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

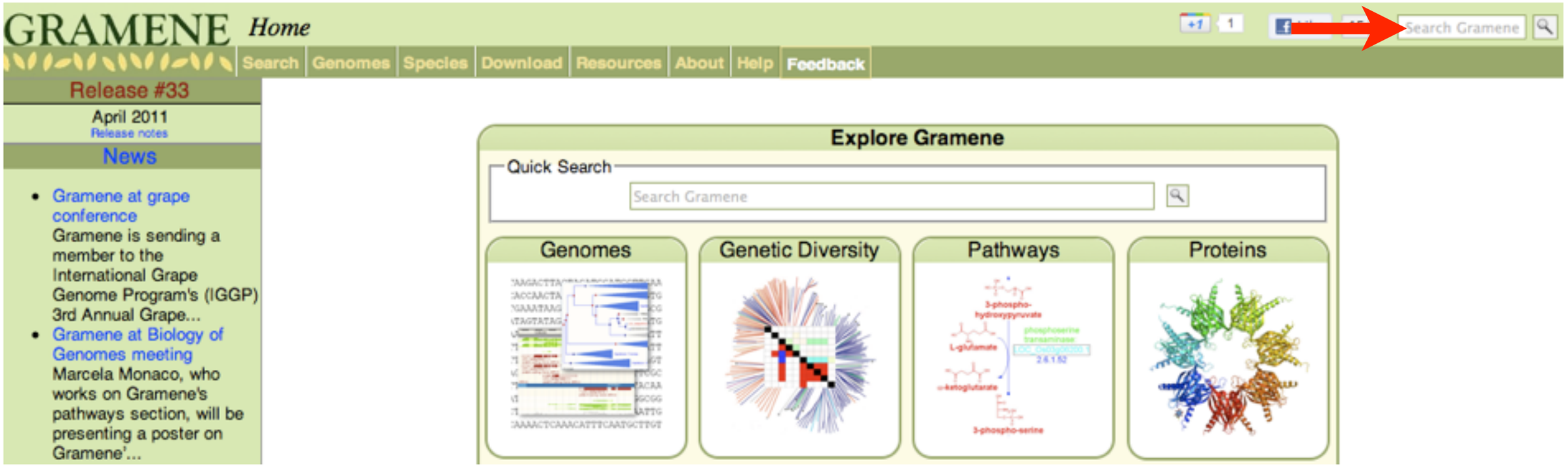


1. Search Tools

sol genomics network

Text Searches:

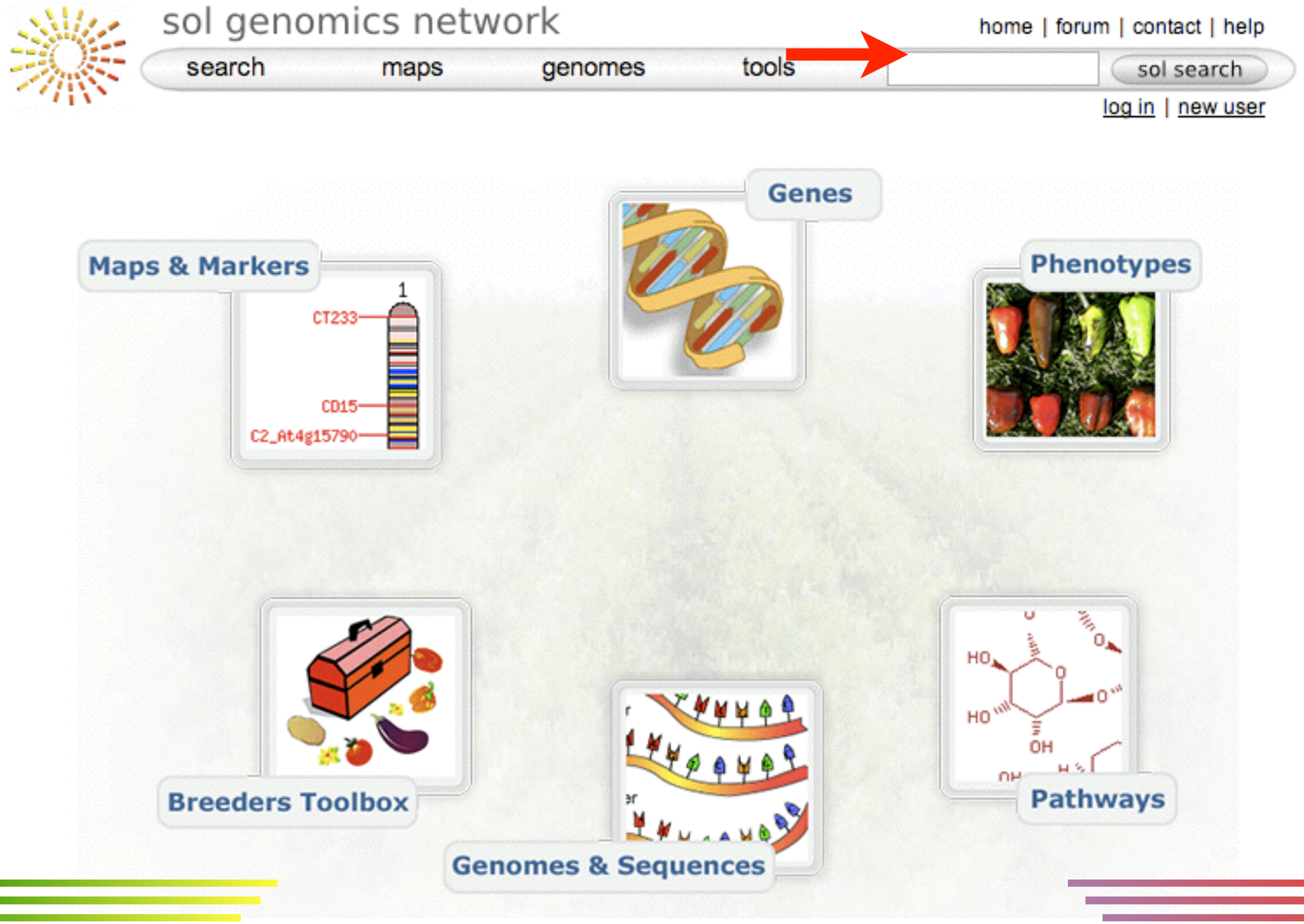
GRAMENE: http://www.gramene.org/



1. Search Tools

Text Searches:

SGN: http://solgenomics.net/



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.

1.1.Searches by Ontology

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.

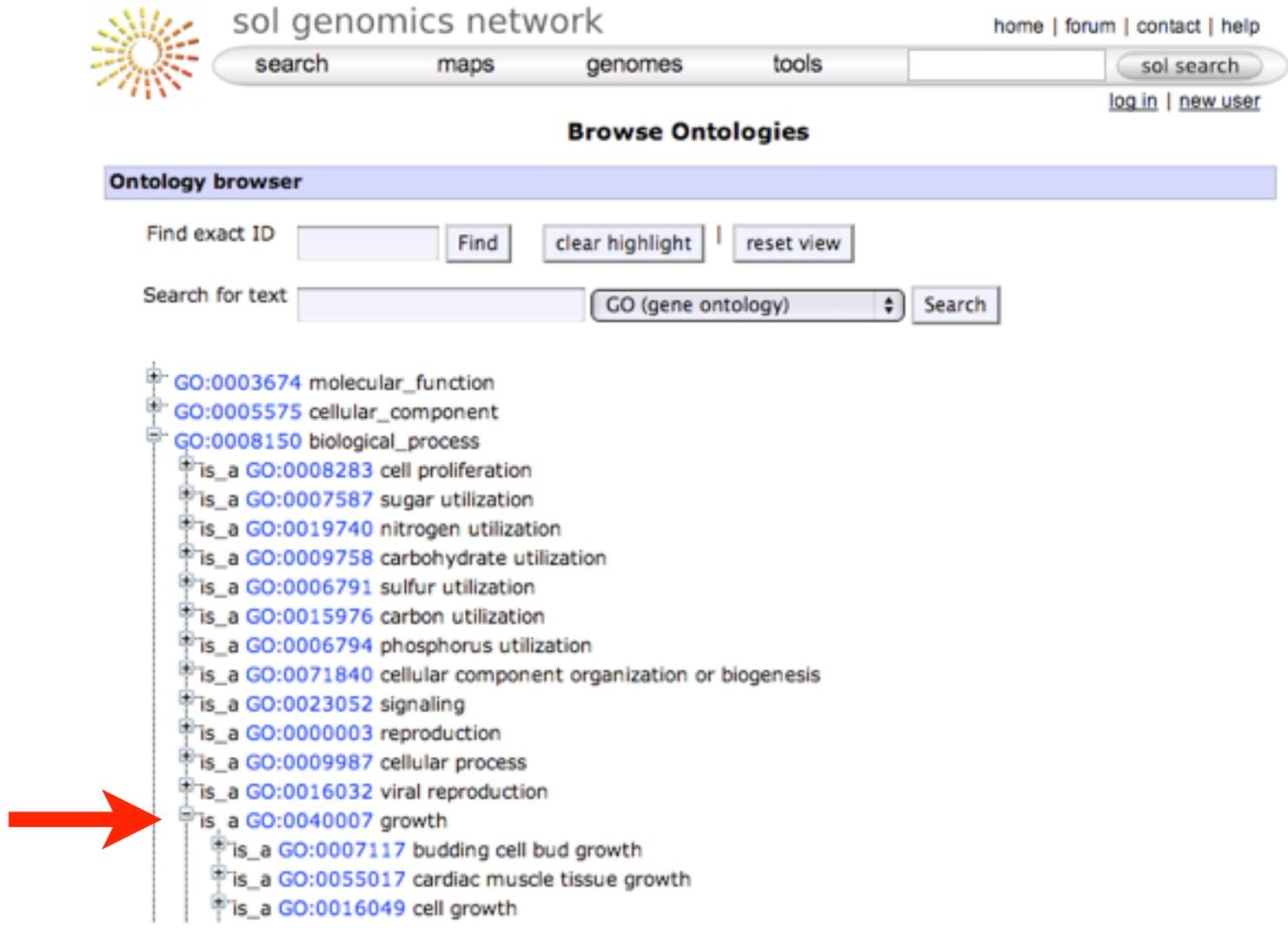
For plants, also is important *Plant Ontology* (http://www.plantontology.org/), to define plant parts such as organs and process associated to plants such as flowering or ripening.

A complete ontology list can be found at:

http://www.obofoundry.org/

1.1.Searches by Ontology

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



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.

1. Search Tools

sol genomics network

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/

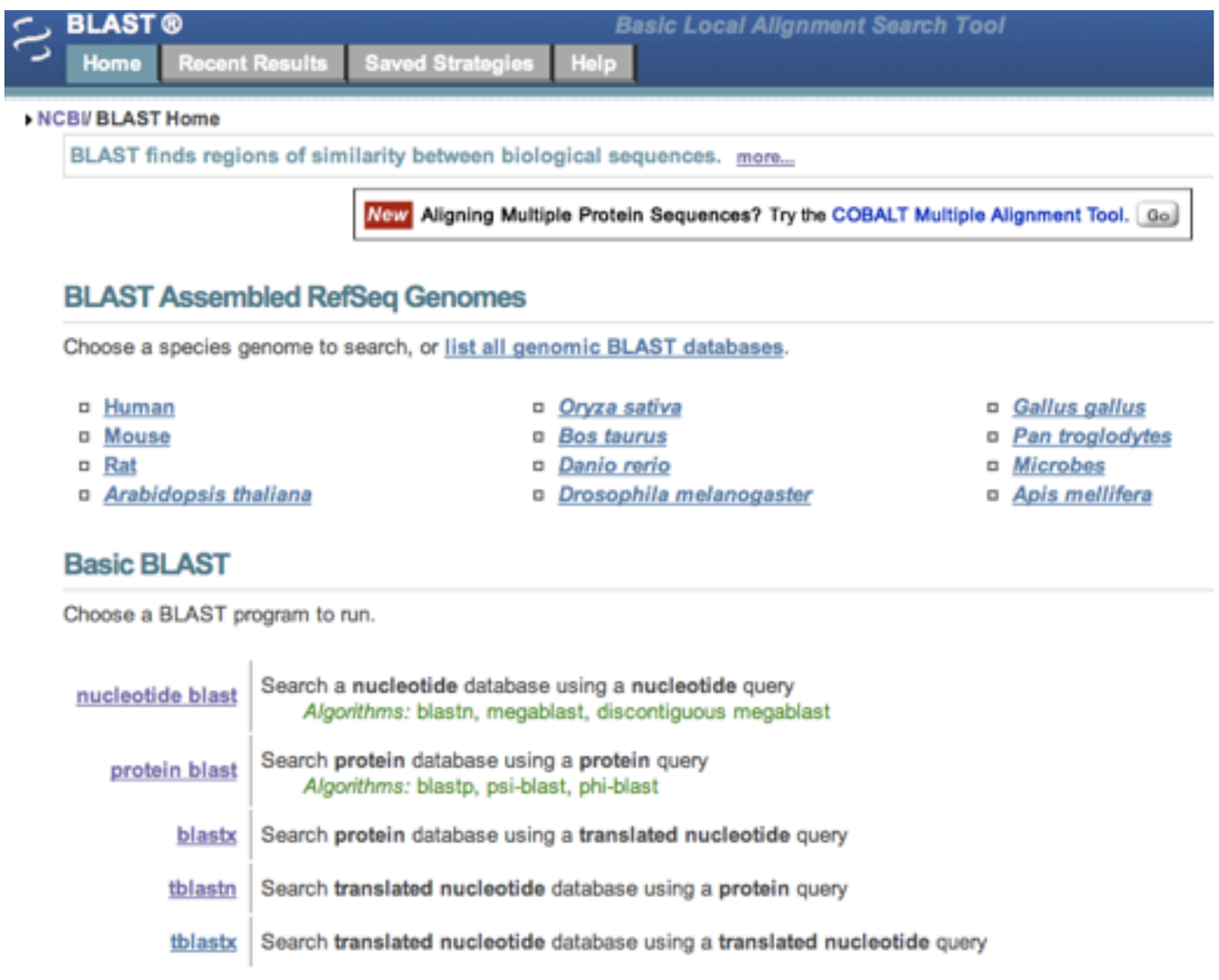
More information at: http://en.wikipedia.org/wiki/Sequence_alignment_software

1. Search Tools

sol genomics network

Sequence homology/similarity searches:

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

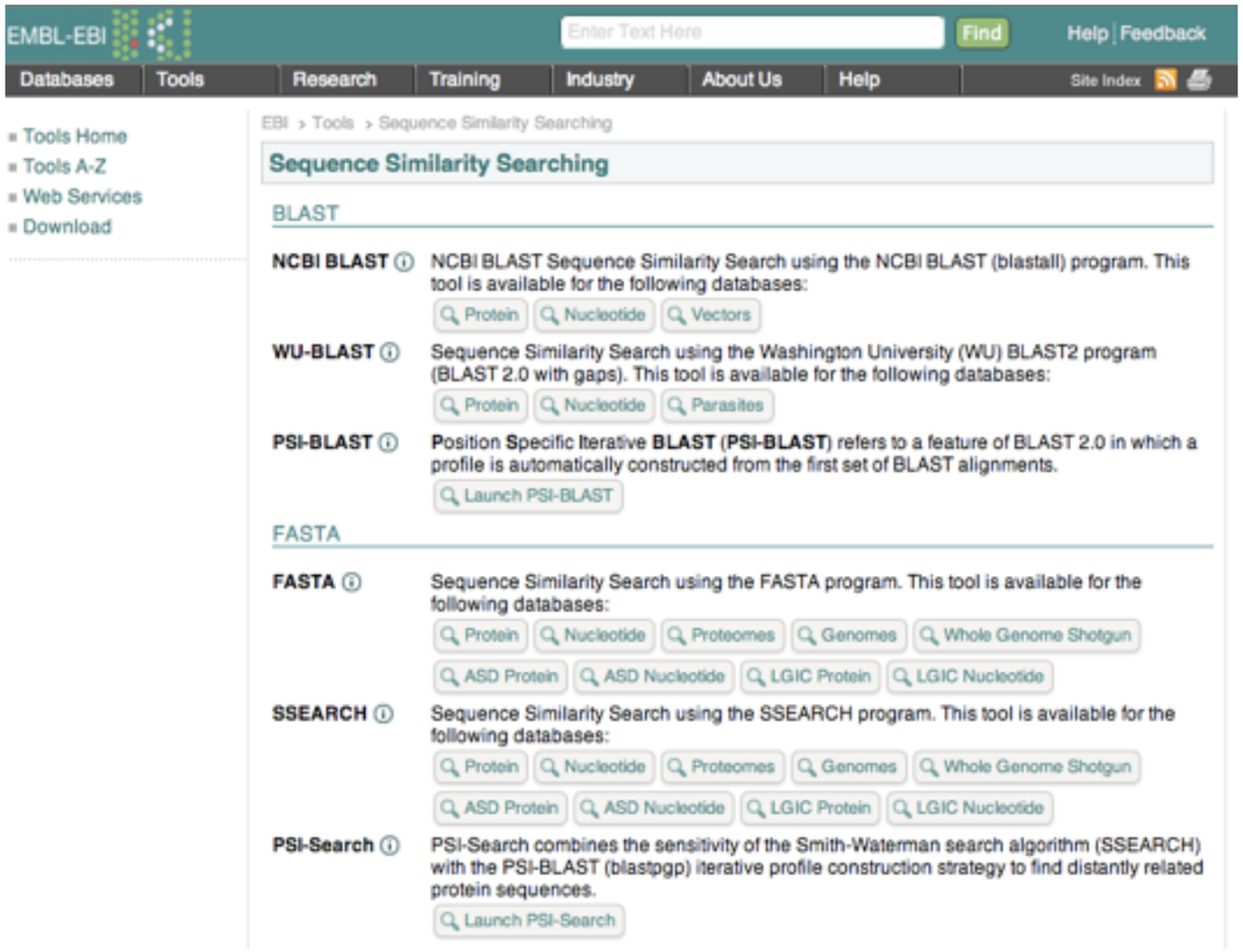


1. Search Tools

sol genomics network

Sequence homology/similarity searches:

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

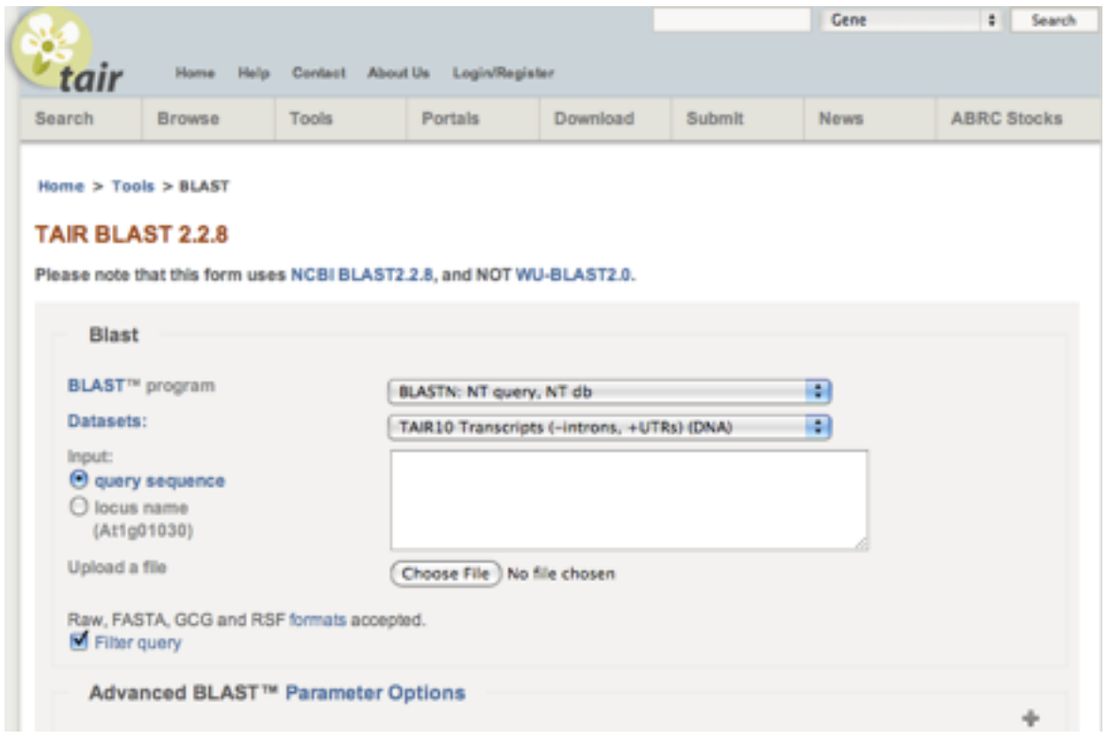


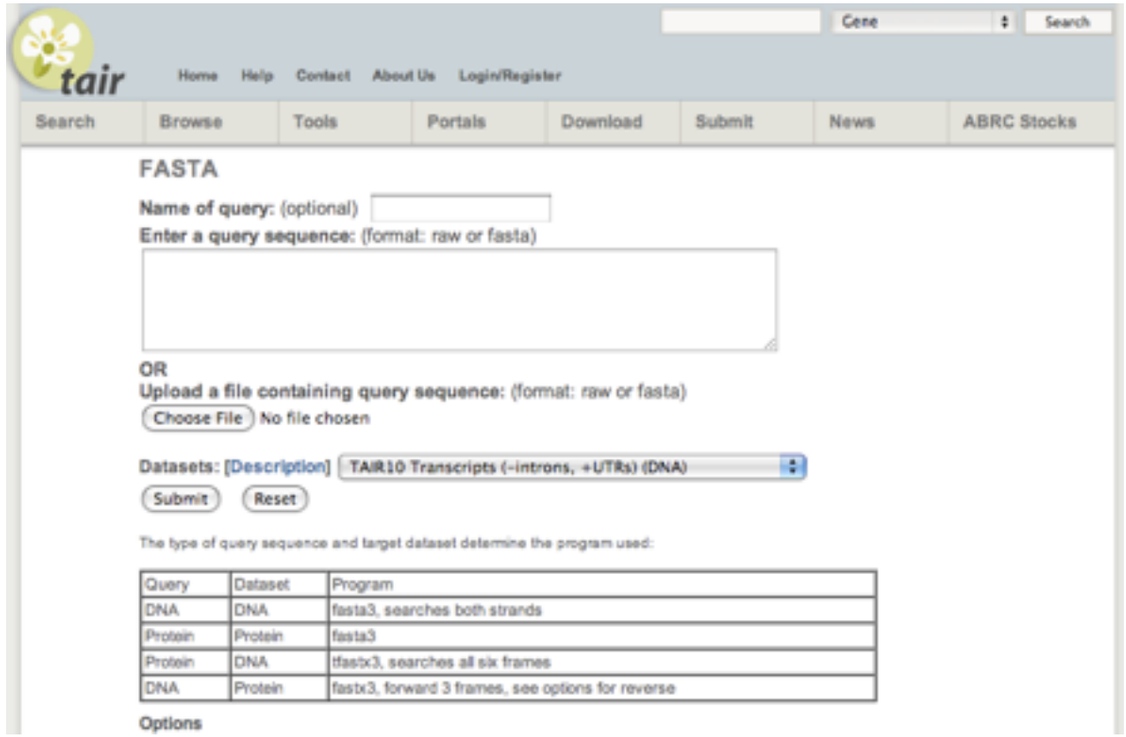
1. Search Tools

sol genomics network

Sequence homology/similarity searches:

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



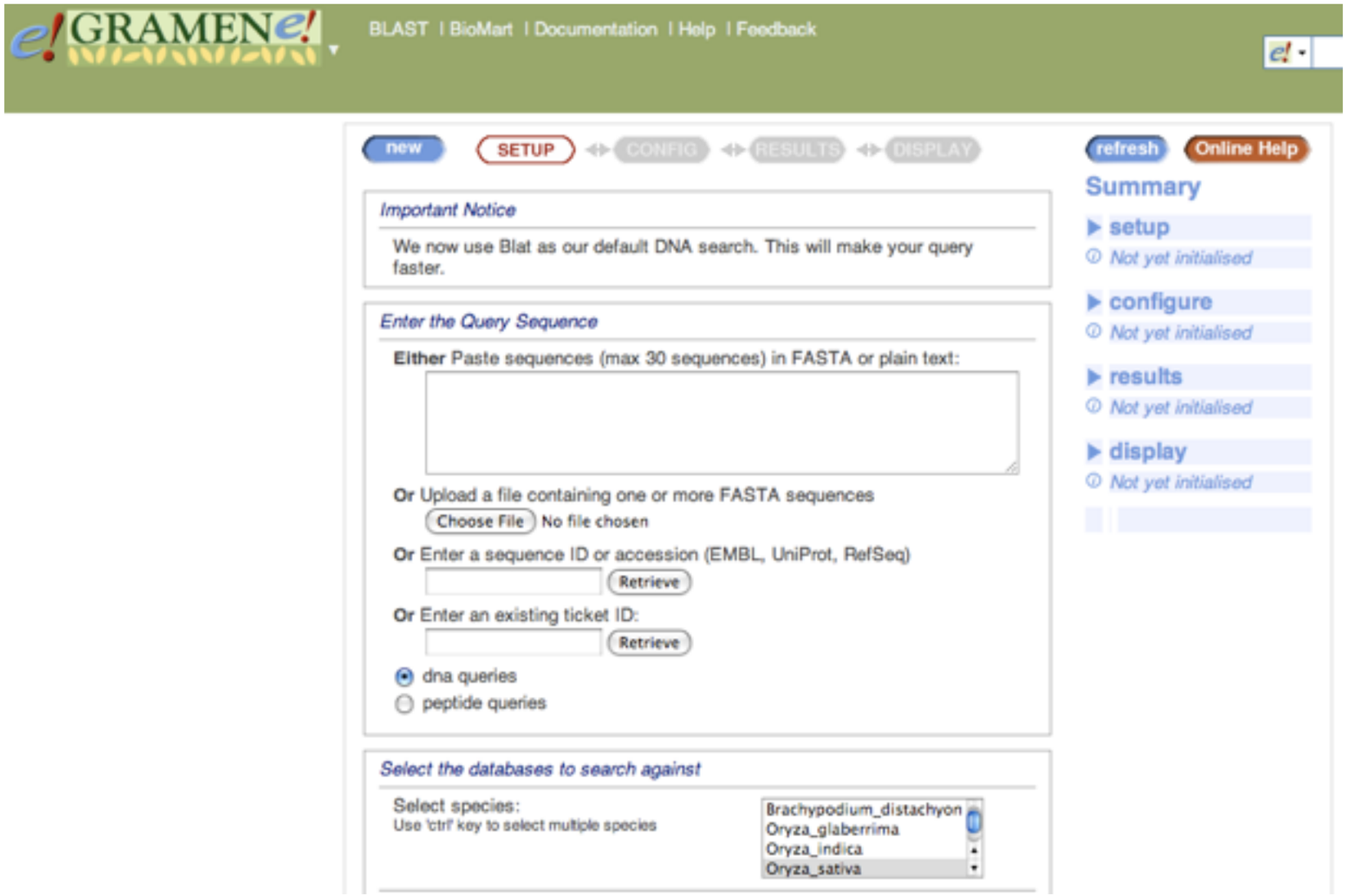


1. Search Tools

sol genomics network

Sequence homology/similarity searches:

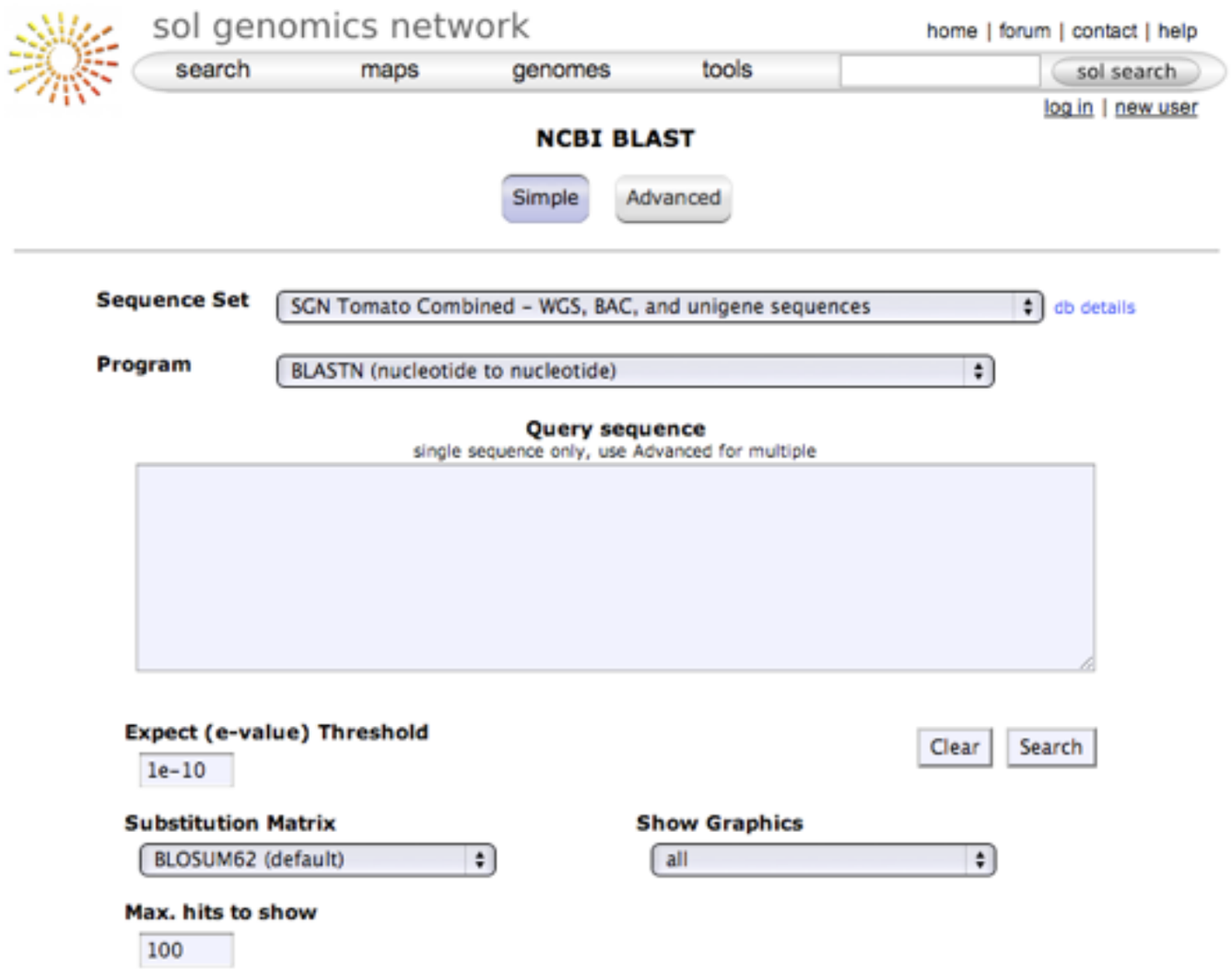
GRAMENE: http://www.gramene.org/multi/blastview



1. Search Tools

Sequence homology/similarity searches:

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



1.2 -Search by Sequence Homology

sol genomics network

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		
		Nucleotide	Translated Nucleotide	Protein
DATABASE	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.

sol genomics network

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

sol genomics network

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

sol genomics network

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 -Manipulation and Sequence Analysis Tools

sol genomics network

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

+ Commercial package:
  LaserGene (DNAStar) (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.

2.1 - Translators and *Gene Predictors*.

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.

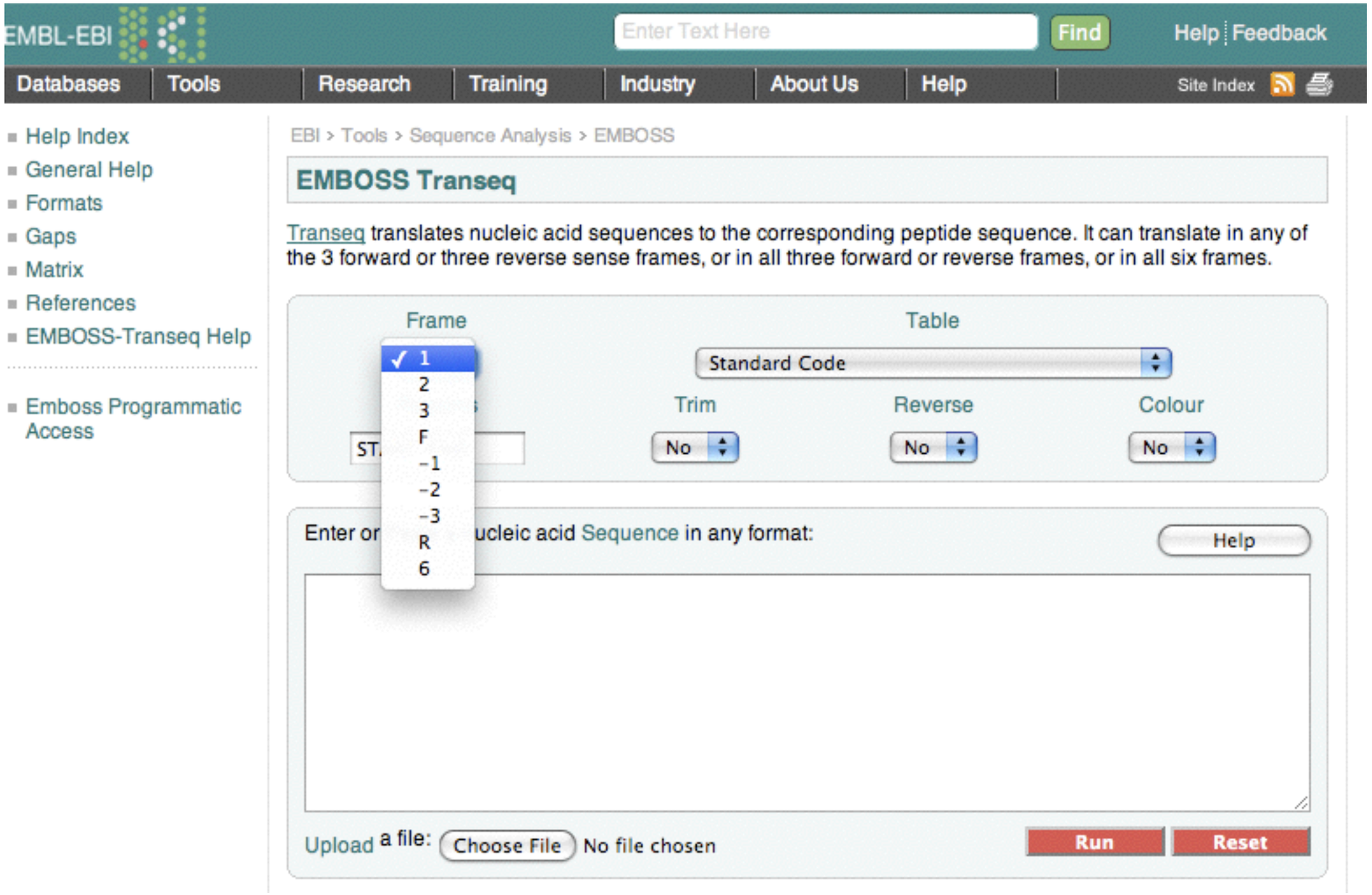
sol genomics network

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/



sol genomics network

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

sol genomics network

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/

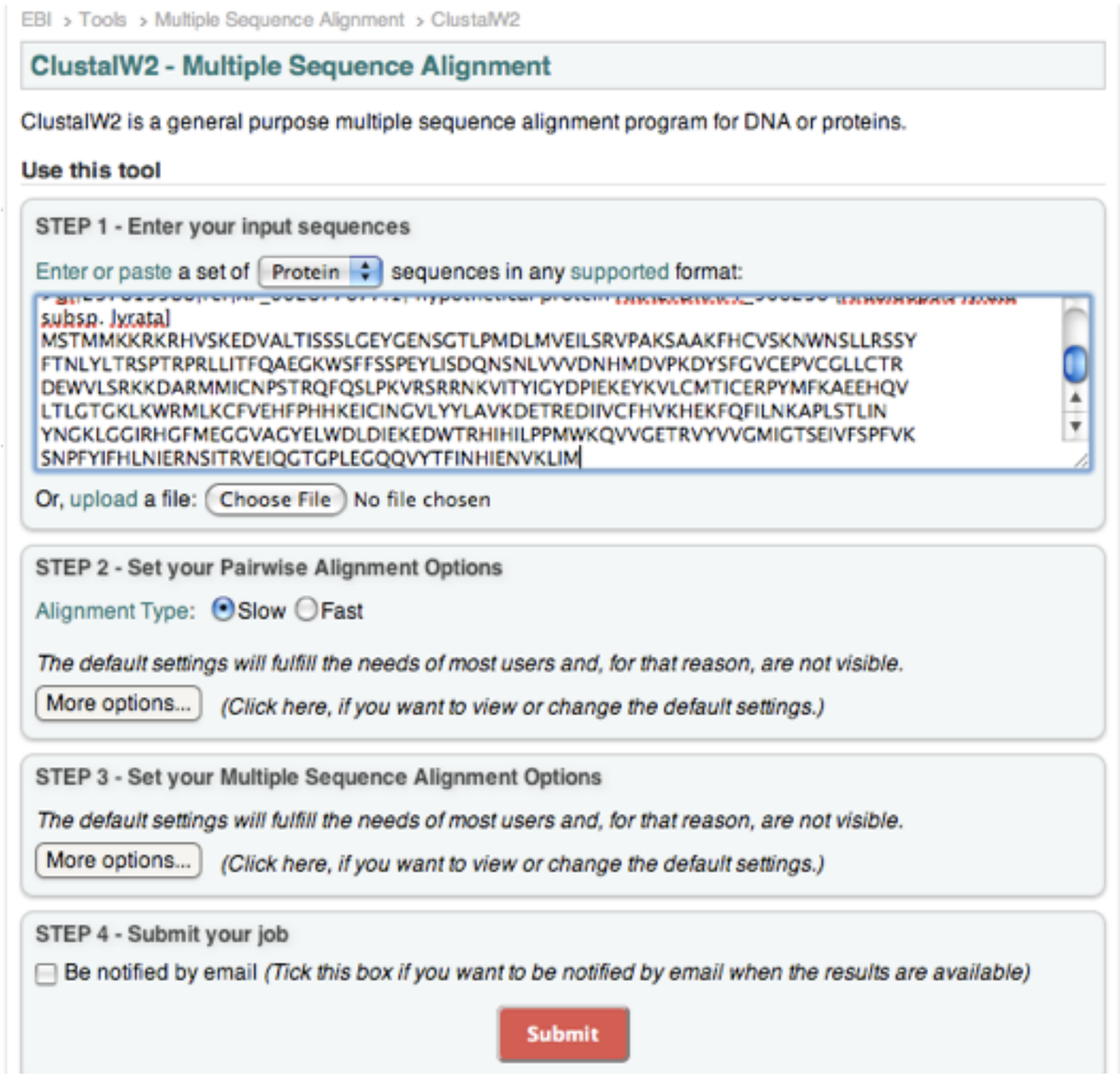
sequence.fasta

```
>gi|15225089|ref|NP_180705.1| F-box protein DOR [Arabidopsis thaliana]
MKSRRQNVSVARQTILGRDENFEPIPIDLVIEIFSRSPVKSIARCRCVSKLWASILRLPYFTELYLTKSC
ARPRLLFACQKHRELFFFSTPQPHNPNESSSPLAASFHMKIPFDGRFNIISPIGGLVFVRYEQILKGRKT
PEFVSAICNPSTGQSLTLPKPKTRKRIWGTSHFGYDPIEKQFKVLSMNIGDGVYKEHYVLTLGTENLSWR
RIECSIPHVHGSKGICINGVLYYRAKADMFSGTLMIVCFDVRFEKFSYIKILKPTTTLISYNGKLASLVW
EGPSYICGKRFEMWVLGDPEKHEWLKHTYELRPRWQNVLGEDLLIFAGMTGTNEIVLSPKYPSHPFYVFY
YNLERNTIRRVEIQGMGAFKVNEDYIFLDHVEDVKLI

>gi|297837229|ref|XP_002886496.1| hypothetical protein ARALYDRAFT_893290 [Arabidopsis lyrata subsp. lyrata]
MKTERQNVSEDVVVVTERNKRAKTSNNGGEPIPFDLTVEICSRLPAKSISRFRCVLKLWGSILRLPYFTE
LFLTRSLARPQLLFACHKDNHVFVFSSPQPQNIDDNNASSLLAANYHMKIPFYASSFERCSSVRGLVFFG
DERYSNGKEHKVSVICNPSTRQSLTLPKLKTRKRIGVRSYFGFEPIEKQYKVLSMTWGIYGTRDMDSEEH
QVLTLGTRKPSWRMIECWIPHSLYHTYNNVCINGVLYYPAVNTSSKGFIIVSFDFRSEEFRFVEDTDTSI
SSYYGPHLINYNGKLGSLGSGGFGGIGASCTSITLRVLEDAEKHEWSEHIYVLPAWWKNIFGGECTVLSV
VGVTRTNEIVLSLRFPSTPFYVFYYNTERNAIRRVEIQGQEAFKDHSVYTFLDHVENVNMKLLEGF

>gi|15239182|ref|NP_201386.1| F-box protein [Arabidopsis thaliana]
MRTLRRNVTENRLTISRRRTEKKTSPNKTEKSVQIPVDIIIEILLRLPAKSIATCRCVSKLWISVICRQD
FTELFLTRSLHRPQLLFCCKKDGNLFFFSSPQLQNPYENSSAISLKNFSLCYKISRPVNGLICFKRKEMN
ETVTVICNPSTGHTLSLPKPMKTSIGPSRFFVYEPIQKQFKVLLSYKSDEHQVLTLGTGELSWRIIECSM
PHILGMSEICINGVLYYPAINLSSGDYIIVCFDVRSEKFRFITVMEEFIKAAHDGTLINYNGKLASLVSE
RYCFVDGRSKSIELWVLQDAEKKEWSKHTYVLPAWWQHRIGTLNLRFVGVTRTNEIMLSPCYQTVPFDVY
YFNIERKTMMSVAIQGMEAFQGHLVFTYLDHVENVKLLHNMF

>gi|15229553|ref|NP_189038.1| putative F-box protein [Arabidopsis thaliana]
MRSRQLHNVSEDRETLSRRNKRSKTSLNGHIPIDLLIEIFLKLPVKSIATCRSVSKFWTYVLGRQDFTEL
FLTKSSSRPQLLFACANDNGYFFFSSNQPQNLDENSSPIAAYPLTHVPKSRDLGPPINGLVSLRGERILK
GRIRPVDVSIIYNPSTGESLTLPKTNMTRKKIYTVTSFLGYDPIEKQYKVLSMNMSYEKHPKCEGYQVLT
LGTGKLSWRMIKCCLNYQHPLKNSEICINGVLYYLAMVNGSSWPTRAVVCFDIRSEMFNFMEVYRELSYT
TTLINYNNGKLGMLMGQEAHKTISGICRSFELWVLEDTVKHEWSKHVYLLPPLWKDAVANTRLYFAGMIG
TSEIVLFRPDEPLCVFYYNIDRNTIKRVGIRGLEAFKYFRIFLNHVENVKLF

>gi|297819588|ref|XP_002877677.1| hypothetical protein ARALYDRAFT_906230 [Arabidopsis lyrata subsp. lyrata]
MSTMMKKRKRHVSKEDVALTISSSLGEYGENSGTLPMDLMVEILSRVPAKSAAKFHCVSKNWNSLLRSSY
FTNLYLTRSPTRPRLLITFQAEGKWSFFSSPEYLISDQNSNLVVVDNHMDVPKDYSFGVCEPVCGLLCTR
DEWVLSRKKDARMMICNPSTRQFQSLPKVRSRRNKVITYIGYDPIEKEYKVLCMTICERPYMFKAEEHQV
LTLGTGKLKWRMLKCFVEHFPHHKEICINGVLYYLAVKDETREDIIVCFHVKHEKFQFILNKAPLSTLIN
YNGKLGGIRHGFMEGGVAGYELWDLDIEKEDWTRHIHILPPMWKQVVGETRVYVVGMIGTSEIVFSPFVK
SNPFYIFHLNIERNSITRVEIQGTGPLEGQQVYTFINHIENVKLIM
```

Input: Set of 5 proteins

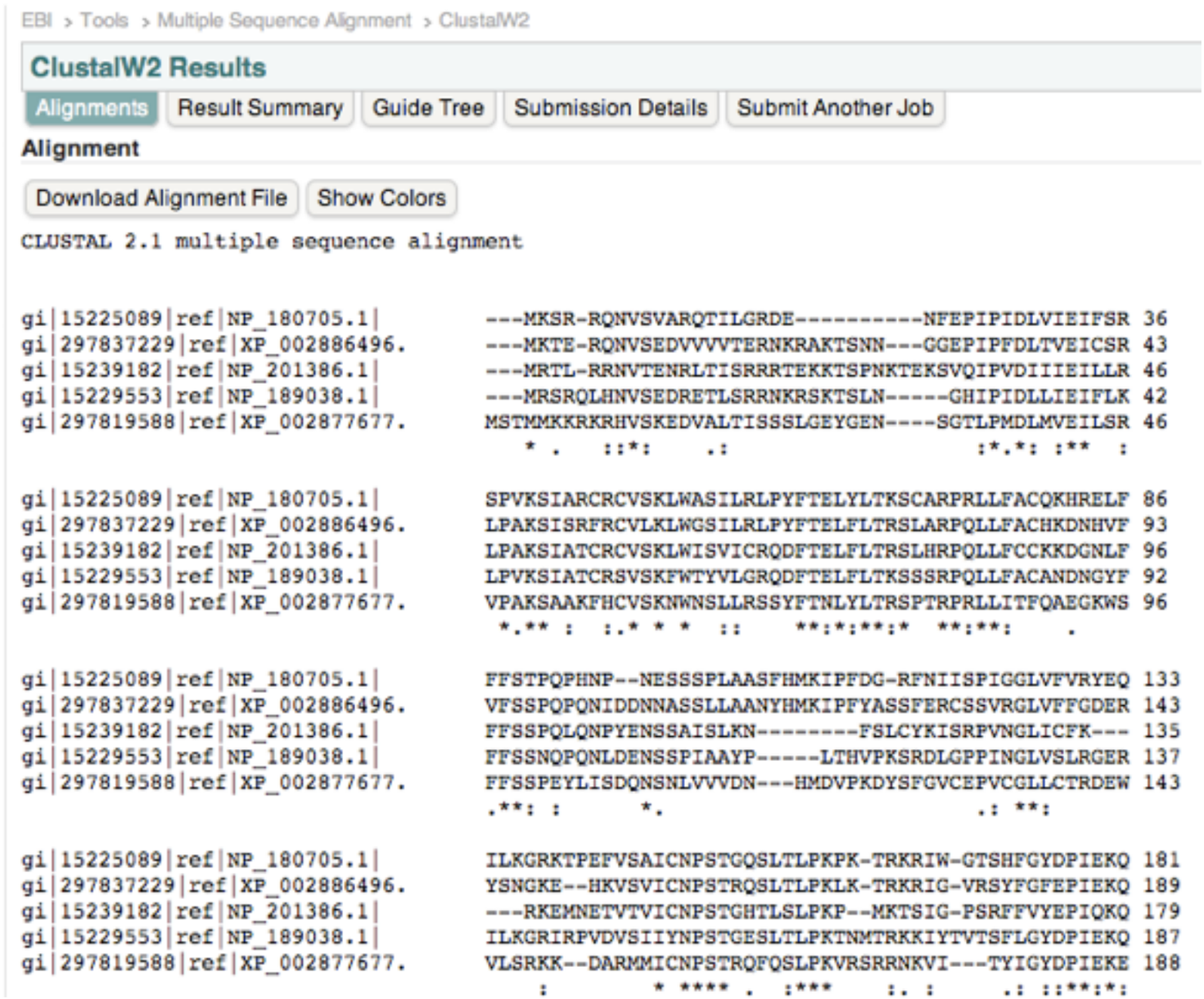
2.2 - Multiple Sequence Alignment

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



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-RQNVSVARQTILGRDE----------NFEPIPIDLVIEIFSR 36
gi|297837229|ref|XP_002886496.      ---MKTE-RQNVSEDVVVVTERNKRAKTSNN---GGEPIPFDLTVEICSR 43
gi|15239182|ref|NP_201386.1|        ---MRTL-RRNVTENRLTISRRRTEKKTSPNKTEKSVQIPVDIIIEILLR 46
gi|15229553|ref|NP_189038.1|        ---MRSRQLHNVSEDRETLSRRNKRSKTSLN-----GHIPIDLLIEIFLK 42
gi|297819588|ref|XP_002877677.      MSTMMKKRKRHVSKEDVALTISSSLGEYGEN----SGTLPMDLMVEILSR 46
                                       * .   ::*:    .:                   :*.*: :**  :

gi|15225089|ref|NP_180705.1|        SPVKSIARCRCVSKLWASILRLPYFTELYLTKSCARPRLLFACQKHRELF 86
gi|297837229|ref|XP_002886496.      LPAKSISRFRCVLKLWGSILRLPYFTELFLTRSLARPQLLFACHKDNHVF 93
gi|15239182|ref|NP_201386.1|        LPAKSIATCRCVSKLWISVICRQDFTELFLTRSLHRPQLLFCCKKDGNLF 96
gi|15229553|ref|NP_189038.1|        LPVKSIATCRSVSKFWTYVLGRQDFTELFLTKSSSRPQLLFACANDNGYF 92
gi|297819588|ref|XP_002877677.      VPAKSAAKFHCVSKNWNSLLRSSYFTNLYLTRSPTRPRLLITFQAEGKWS 96
                                     *.** :  :.* * *  ::    **:*:**:*  **:**:    .

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

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

2.2 - Multiple Sequence Alignment

sol genomics network

- 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

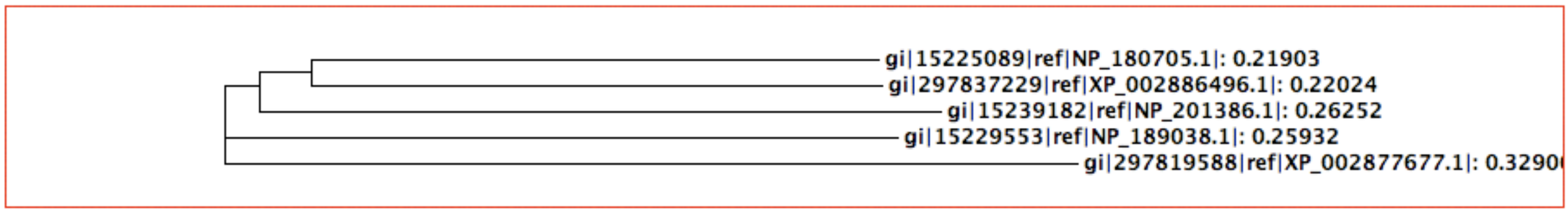
Guide Tree

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

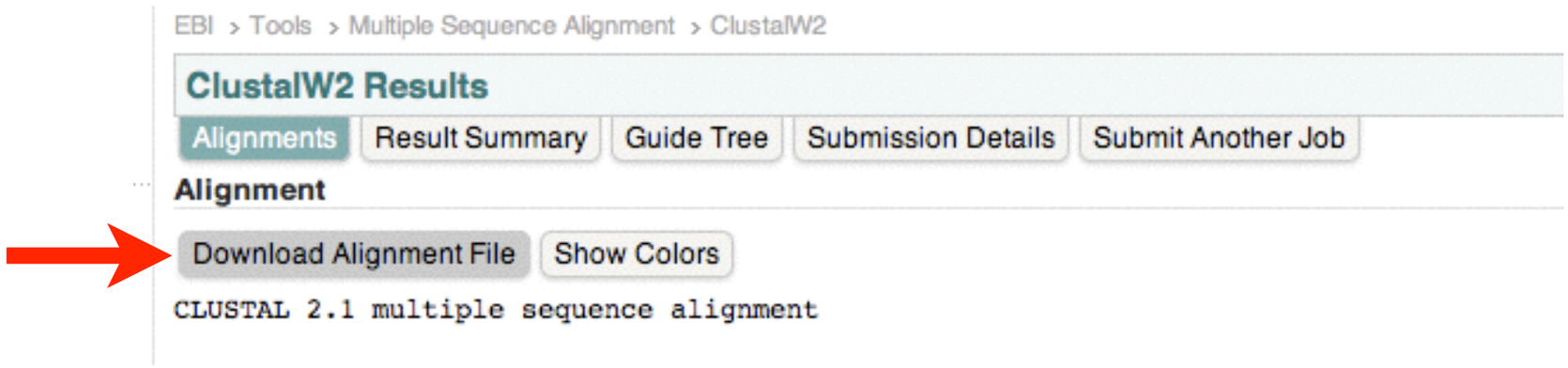
Show as Cladogram Tree | Hide Distances



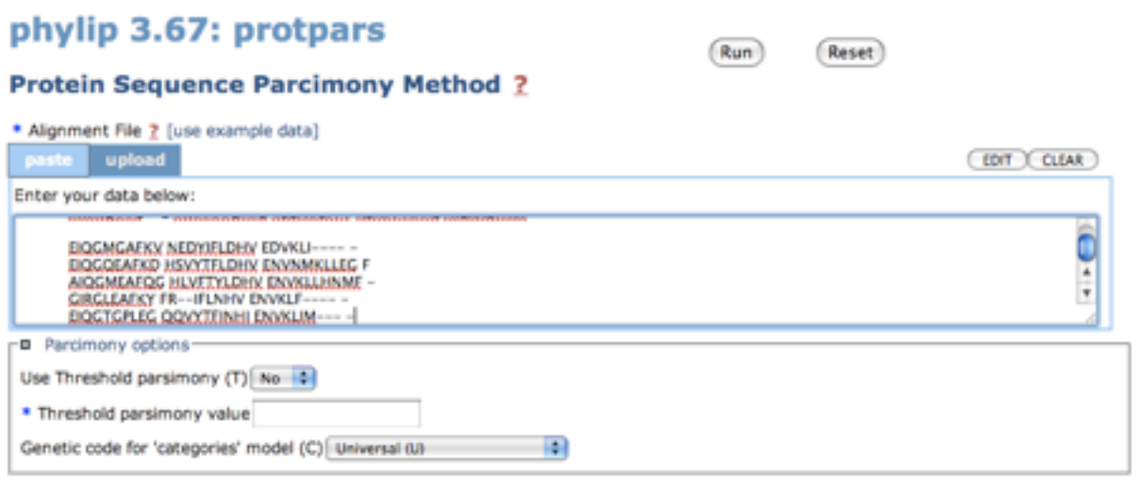
Right-click on the above tree to see display options.

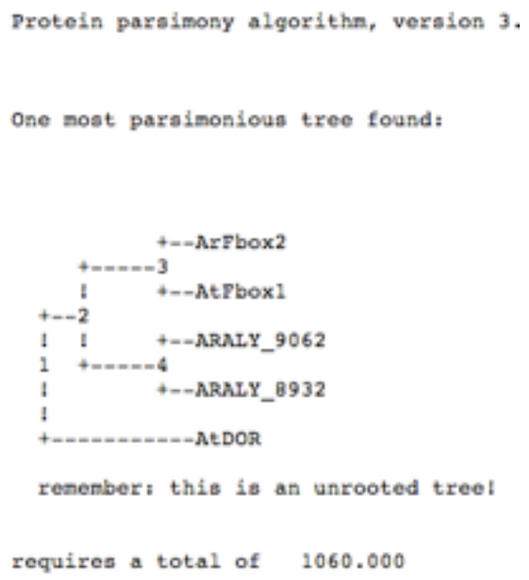
2.2 - Multiple Sequence Alignment

- 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).





Web-based Phylip package:
http://mobyle.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*.

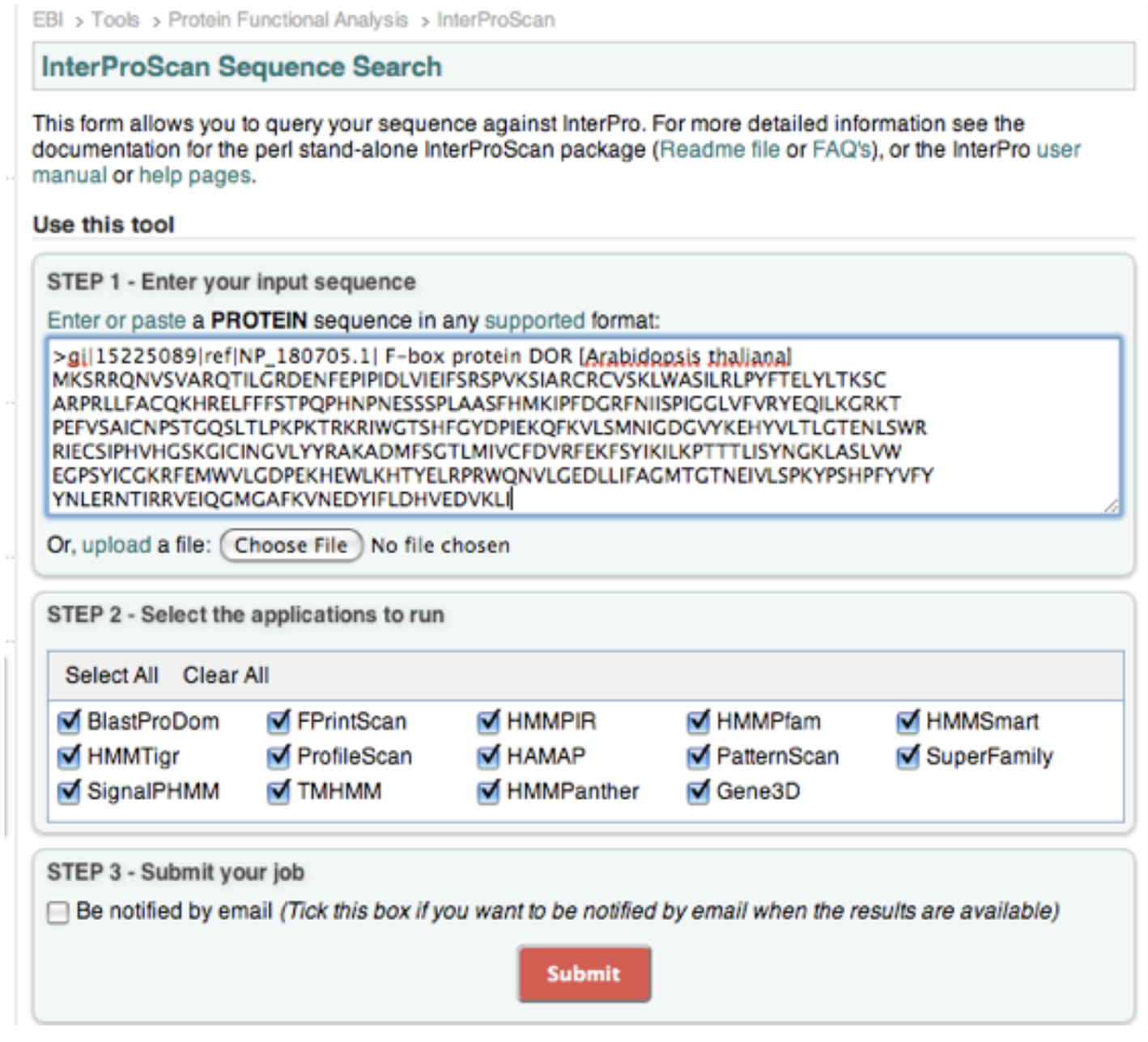
InterPro 32.0

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

2.3 -Protein Domain Analysis

sol genomics network

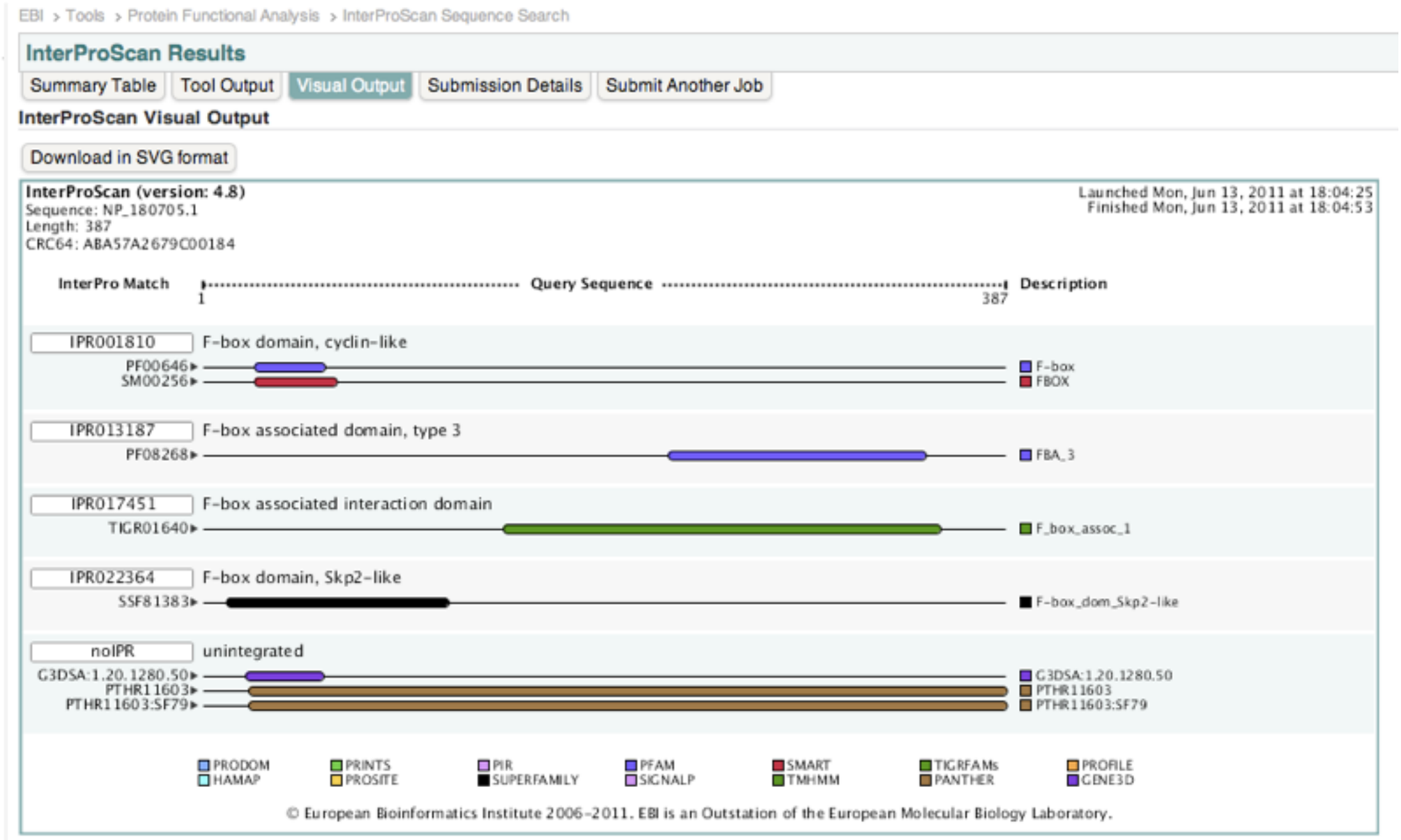
InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/).



2.3 -Protein Domain Analysis

sol genomics network

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



sol genomics network

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_2N-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_2N-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_2N-----Arg-Leu-X_5-His-Leu-

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

sol genomics network

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 to123 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 ACGTNacgtn -- 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): NONE

```
AACGTCAATGAATAGATAGATGGCTGCCGCCGGCAATCCAAAGTTCCCCGGCTGCTTCCCG
CCACCACCACTTCCACCCTCACCTGGTGGCTCATTACCAAAGTTCTTGAAATGATAATTA
CTCCCCATTTCACTAAAACTCCTCAGTCCTCACACAATCCGCCTTCAAACCTCAGCTCTG
TTATTCAAGAATCACAAAACCTACATATCAGATCAACAAGTTAATTCCCTTCCCTTTGAA
CCTTTTTCCTTATCATACTGTTCAACCCTTCACATAAATGTACATCTATTTACAAACACA
CAGTTAATTAAAAGCAAAATATACCTGGAAAGAGATCAAAAATCAATTTACAGCTAAAAC
```

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):

Pick Primers Reset Form

Sequence Id: A string to identify your output.

Targets: 200,123 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 123-223

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

Pick Primers Reset Form

3.2 - Primer Design.

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

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any   3'  seq
LEFT PRIMER        157   19   60.20    52.63  3.00  3.00 ATCCGCCTTCAAACCTCAG
RIGHT PRIMER       373   21   59.51    47.62  2.00  2.00 AAGGGGTTGGTGAGTTTTAGC
SEQUENCE SIZE: 524
INCLUDED REGION SIZE: 524

PRODUCT SIZE: 217, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00
TARGETS (start, len)*: 200,123

    1 AACGTCAATGAATAGATAGATGGCTGCCGCGGCAATCCAAAGTTCCCCGGCTGCTTCCCG

   61 CCACCACCACTTCCACCCTCACCTGGTGGCTCATTACCAAAGTTCTTGAAATGATAATTA

  121 CTCCCCATTTCACTAAAACTCCTCAGTCCTCACACAATCCGCCTTCAAACCTCAGCTCTG
                                          >>>>>>>>>>>>>>>>>>>

  181 TTATTCAAGAATCACAAAACCTACATATCAGATCAACAAGTTAATTCCCTTCCCTTTGAA
                        ****************************************

  241 CCTTTTTCCTTATCATACTGTTCAACCCTTCACATAAATGTACATCTATTTACAAACACA
      ************************************************************

  301 CAGTTAATTAAAAGCAAAATATACCTGGAAAGAGATCAAAAATCAATTTACAGCTAAAAC
      **********************                                <<<<<<<<

  361 TCACCAACCCCTTATCAATAAAATCATCAAAAACAAATCCTATTTGAAATTCACTTCATT
      <<<<<<<<<<<<

  421 CAACTAAATTGACTGCATTTTCAGTTCACACCAAGAACCCCCAAAACACAACTTCCCCAC

  481 CCACCAATCCAATAAAGAACACACCTTTTGACCTTCAAATACAC

KEYS (in order of precedence):
****** target
>>>>>> left primer
<<<<<< right primer
```

sol genomics network

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 (Instituto 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 a protein from exercise 1 part 5, what domains can be found?
2. Find the *Arabidopsis thaliana* best protein match to the protein.
3. Find the tomato best protein match to the protein
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 2 Solutions (cont'd)

1.1 used the gene from *C. arabica*: interpro scan (http://www.ebi.ac.uk/interpro/) : SAM

Protein family membership

SAM dependent carboxyl methyltransferase (IPR005299)

Domains and repeats

None predicted.

Detailed signature matches

IPR005299 — SAM dependent carboxyl methyltransferase

no IPR — Unintegrated signatures

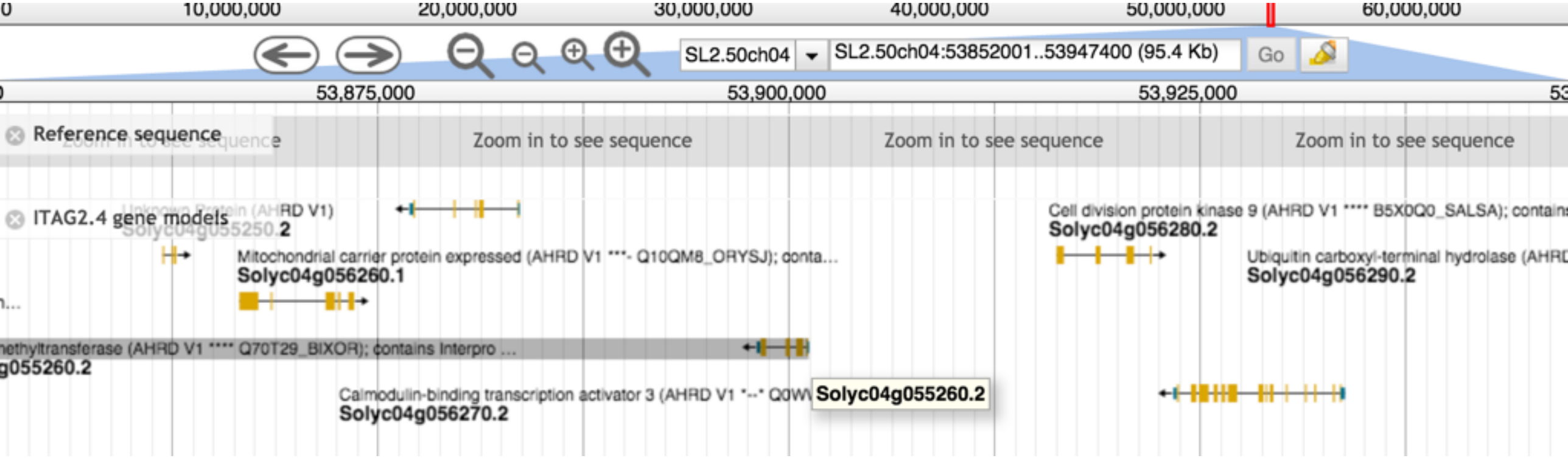
Exercise 2 Solutions

2. At5g04380 (http://arabidopsis.org/Blast/index.jsp)

3. Solyc04g05560 (http://solgenomics.net/tools/blast/index.pl)

Exercise 2 Solutions

4. 3 introns



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

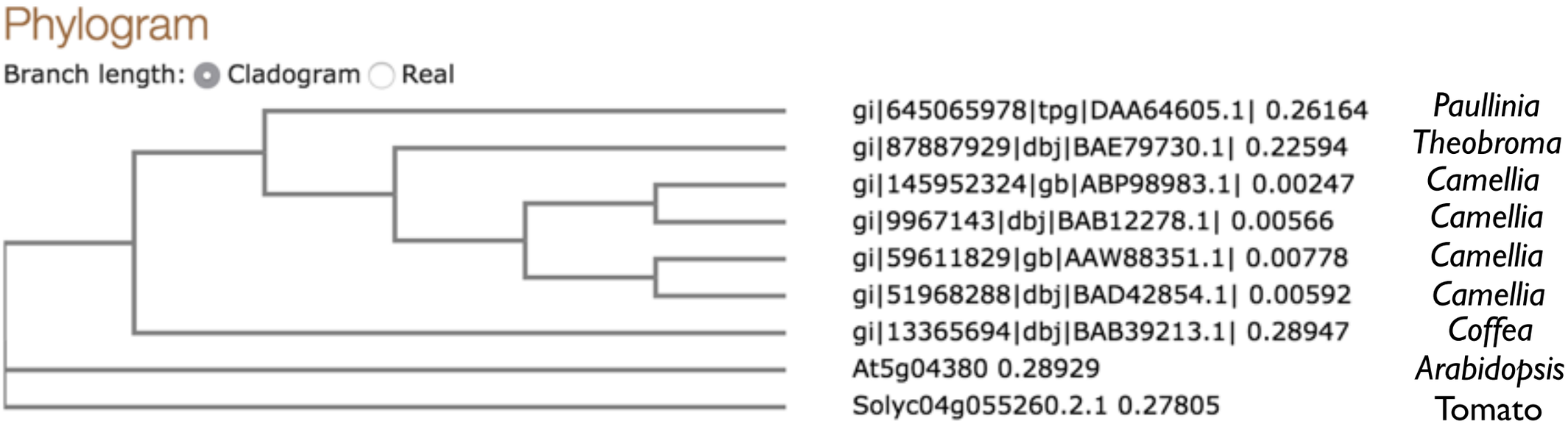
Exercise 2 Solutions

5.

```
CLUSTAL O(1.2.1) multiple sequence alignment


gi|645065978|tpg|DAA64605.1|      -------------------------MDMKDVLCMNTGEGESSYLLNSKFTNVTAIKSIPT
gi|87887929|dbj|BAE79730.1|       -------------------------MEVKEMLFMNKGDGENSYVKTSGYTQKVAAVTQPV
gi|145952324|gb|ABP98983.1|       -------------------MELATAGKVNEVLFMNRGEGESSYAQNSSFTQQVASMAQPA
gi|9967143|dbj|BAB12278.1|        -------------------MELATAGKVNEVLFMNRGEGESSYAQNSSFTQQVASMAQPA
gi|59611829|gb|AAW88351.1|        ------------------------MKEVKEALFMNKGEGESSYAQNSSFTQTVTSMTMPV
gi|51968288|dbj|BAD42854.1|       ------------------------MKEVKEALFMNKGEGESSYAQNSSFTQTVTSMTMPV
gi|13365694|dbj|BAB39213.1|       -------------------------MELQEVLHMNGGEGEASYAKNSSFNQLVLAKVKPV
At5g04380                        MSLCLILCRCDCKSEYKVDEERSSKYPFVGALCMNGGDVDNSYTTKSLLQKRVLSITNPI
Solyc04g055260.2.1               -------------------------MEVTKVLHMNGGMGDASYAKNSLLQQKVILMTKSI
                                                            .   * ** *  : **  .*   : .

gi|645065978|tpg|DAA64605.1|     LKRAIESLFKEESPPFEHLLNVADLGCASGSTSNTIMPTVVQTVVNKCRE--LNHKIPEF
gi|87887929|dbj|BAE79730.1|      VYRAAQSLFTGRNSCSYQVLNVADLGCSSGPNTFTVMSTVIESTRDKCSE--LNWQMPEI
gi|145952324|gb|ABP98983.1|      LENAVETLFSR-DFHL-QALNAADLGCAAGPNTFAVISTIKRMMEKKCRE--LNCQTLEL
gi|9967143|dbj|BAB12278.1|       LENAVETLFSR-DFHL-QALNAADLGCAAGPNTFAVISTIKRMMEKKCRE--LNCQTLEL
gi|59611829|gb|AAW88351.1|       LENAVETLFSK-DFHLLQALNAVDLGCAAGPTTFTVISTIKRMVEKKCRE--LNCQTLEL
gi|51968288|dbj|BAD42854.1|      LENAVETLFSK-DFHLLQALNAVDLGCAAGPTTFTVISTIKRMMEKKCRE--LNCQTLEL
gi|13365694|dbj|BAB39213.1|      LEQCVRELLRANLPNINKCIKVADLGCASGPNTLLTVWDTVQSIDKVKQEMKNELERPTI
At5g04380                        LVKNTEEMLTN--LDFPKCIKVADLGCSSGQNTFLAMSEIVNTINVLCQK--WNQSRPEI
Solyc04g055260.2.1               TDEAISSLYNN--LSSRETICIADLGCSSGPNTFLSVSQFIQTIDKERKKK-GRHKAPEF
                                   .    :         . :  .****::* .:   :    .      :   . .   :
```

http://www.ebi.ac.uk/Tools/msa/clustalw2/

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

Additional Bioinformatics Classes

1. slides at ftp://ftp.solgenomics.net/bioinfo_class/interns/2015/

2. If you are interested in hands on command line training, come to the advanced bioinformatics sessions:
    1. Introduction to Unix Command-line - Noe Fernandez (6/25)
    2. Mapping NGS Data - Aimin Yan (7/2/15)
    3. SNP calling from NGS Data - Naama Menda (7/9)
3. Sign-up: email srs57@cornell.edu
    I[
4. **You will need to have a virtual machine installed prior to next class**

https://btiplantbioinfocourse.wordpress.com/how-to/set-a-virtual-machine-using-vm-virtualbox/