sol genomics network

Bioinformatics Intro and Web-Tools

sol genomics network

presented by
Suzy Strickler
Rm 217

Slides can be found here: ftp://ftp.solgenomics.net/bioinfo_class/interns/2017/

Boyce Thompson Institute
for Plant Research

What is bioinformatics?

WIKIPEDIA
The Free Encyclopedia

- Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

- Bioinformatics organizes and analyzes basic biological data, whereas computational biology builds theoretical models of biological systems, just as mathematical biology does with mathematical models.

Bioinformatics can...

- Identify similar sequences
- Provide a putative function for a sequence
- Assemble sequences (genomes, transcriptomes)
- Annotate genomes
- Identify differentially expressed genes
- Build networks of genes or metabolites
- Determine phylogenetic relationships
- Mine literature for biological information
- Uncover differences between two genomes
- Calculate how a protein folds

Slide credit: Lukas Mueller

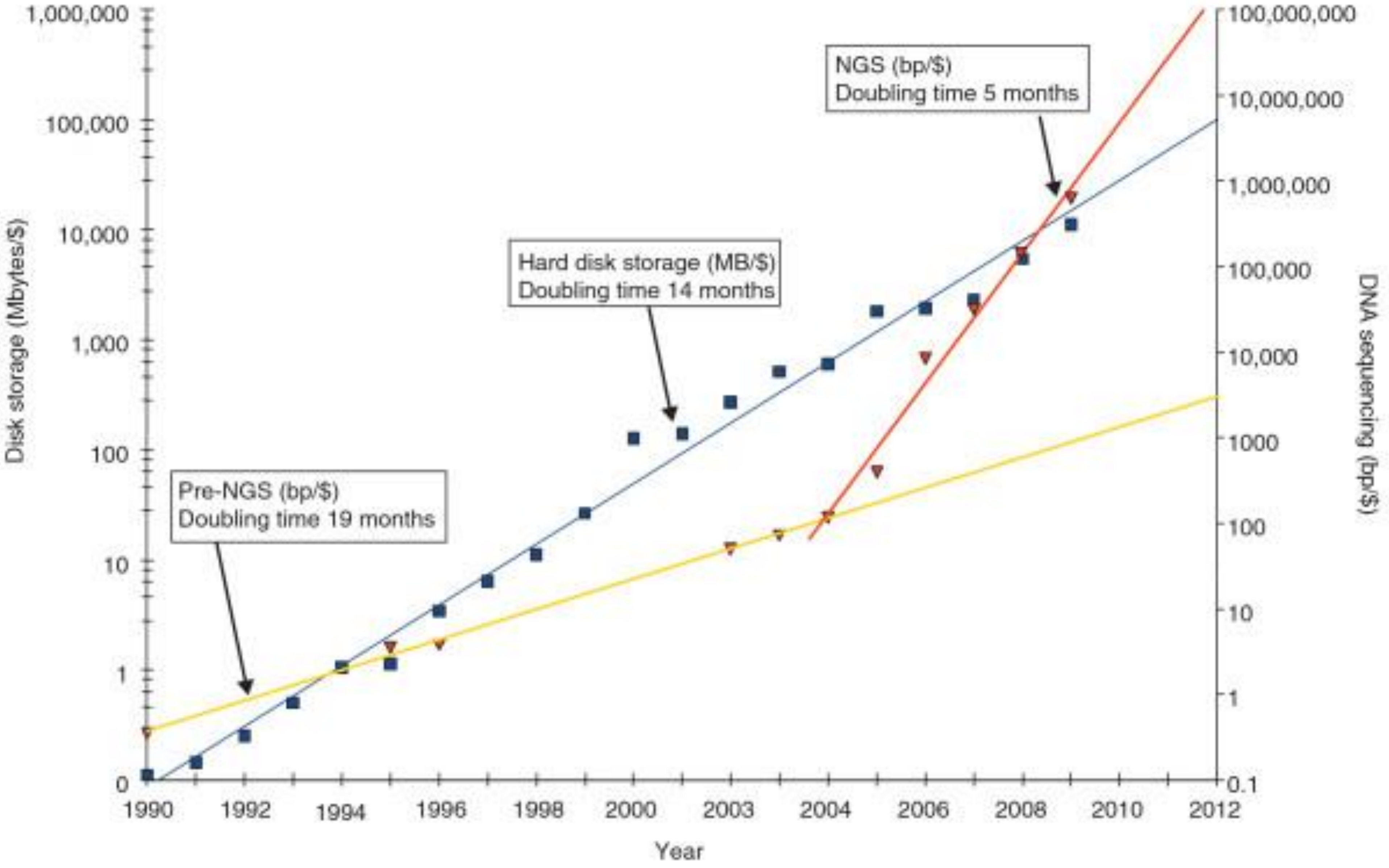
What can bioinformatics do for me?

- Majority of projects involve large datasets
- Speed up your research
- Enable you to ask new questions
- Basic knowledge of bioinformatics needed
    - Extract information
    - Transform information
    - Run analyses
    - Build hypotheses, etc.

Slide credit: Lukas Mueller

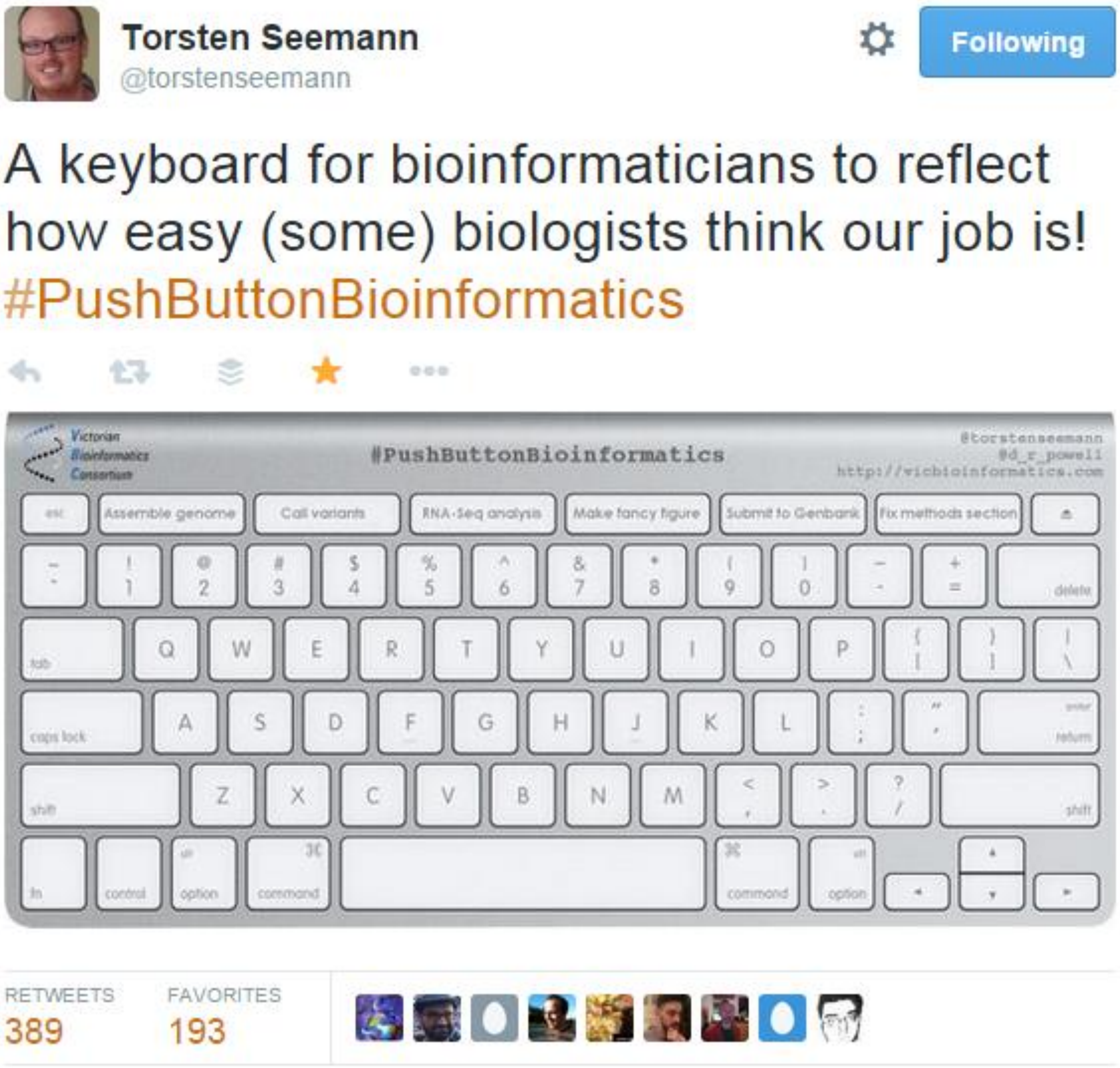
Why do we need bioinformatics?

Increase in Sequencing Data



L. Stein, Genome Biology, 2010

Slide credit: Lukas Mueller



Torsten Seemann
@torstenseemann

A keyboard for bioinformaticians to reflect how easy (some) biologists think our job is! #PushButtonBioinformatics

RETWEETS 389
FAVORITES 193

slide credit: Surya Saha

Linux

- UNIX-based, free and open source operating system
- Very stable, easy to use
- Created by Linus Torvalds in 1990s as a student
- Adopted for most bioinformatics work
  - Also: installed on cell phones, laptops, desktops, clusters, supercomputers
- Can run on your computer!
  - Virtualized or native

http://www.linux-netbook.com/linux/distributions/

More on this next week!

Slide credit: Lukas Mueller

Scripting

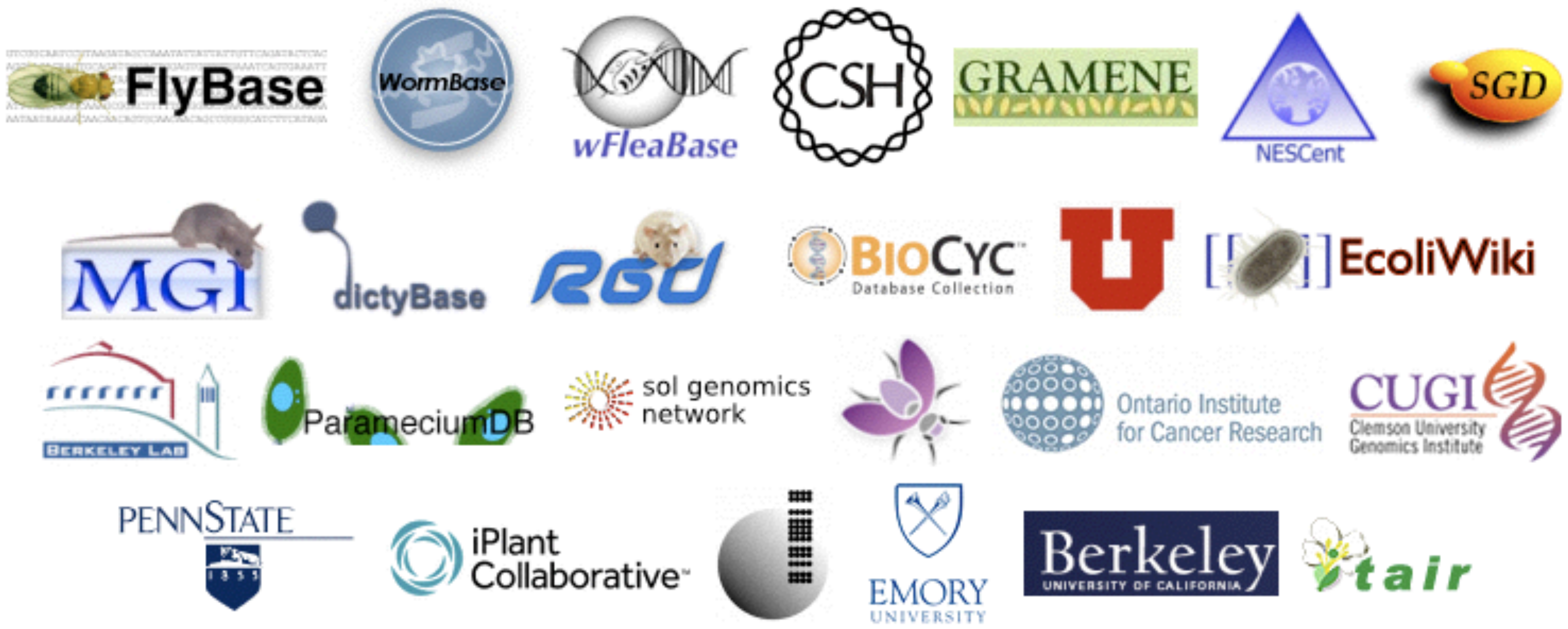
- Scripts: Small programs written by the end-user that control the execution of other programs or perform a simple algorithm
- Extremely flexible
- Written in Shell, Perl, Python Also R
- You can write them yourself!!!

Slide credit: Lukas Mueller

Web-based bioinformatics

- Many databases and tool are accessible through a graphical user interface (GUI).
- We will focus on these today.

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.

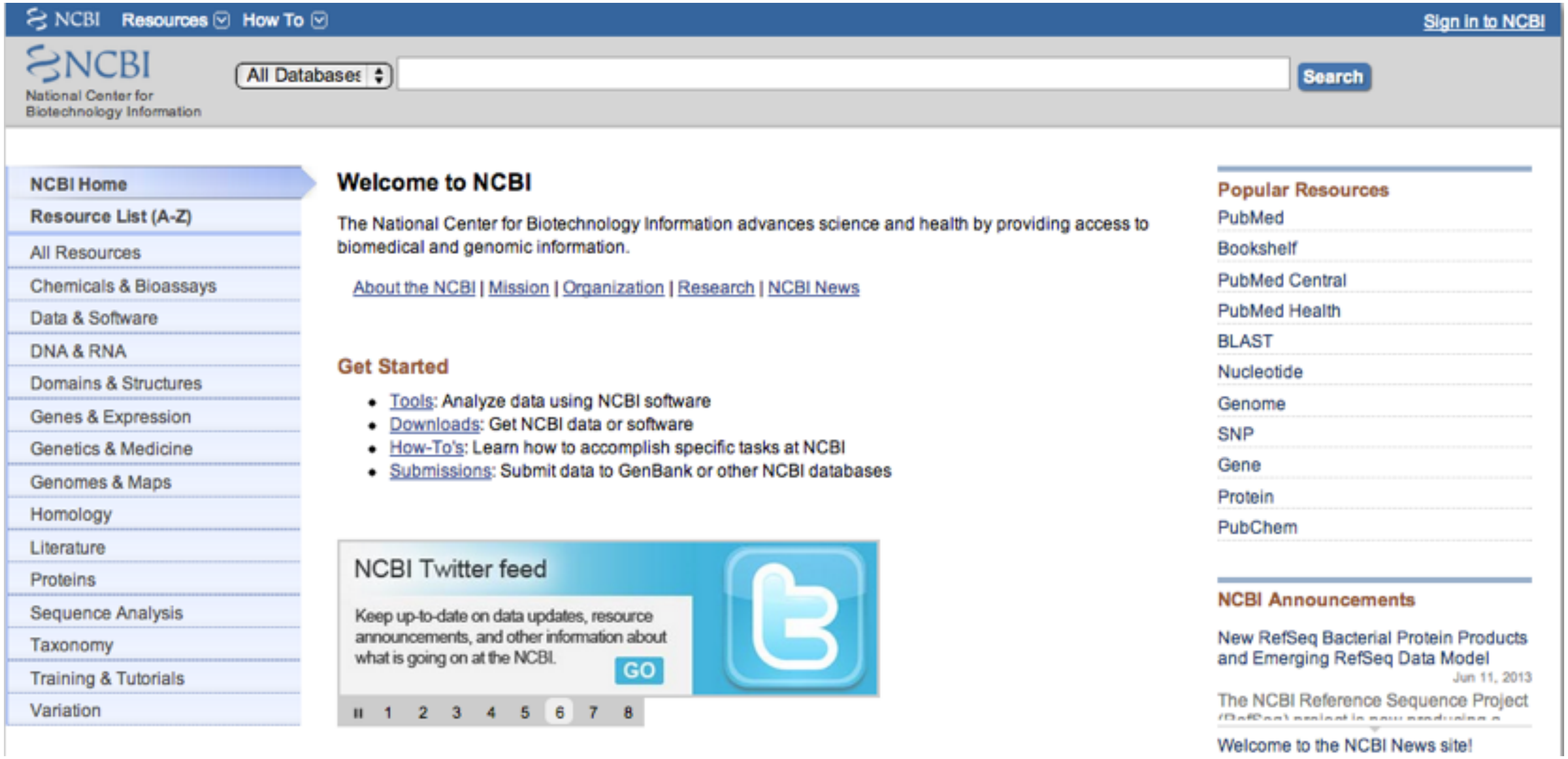
1. Types.

- ✳ Public repositories.
    - Maintained by public agencies or public international consortiums.
    - Massive data amounts (**quantity**).
    - No curated or poorly curated data.
    - Long term data storage.

2. Public Repositories.

NCBI (National Center for Biotechnology Information)

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



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

II 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) project is now producing a

Welcome to the NCBI News site!

2. Public Repositories.

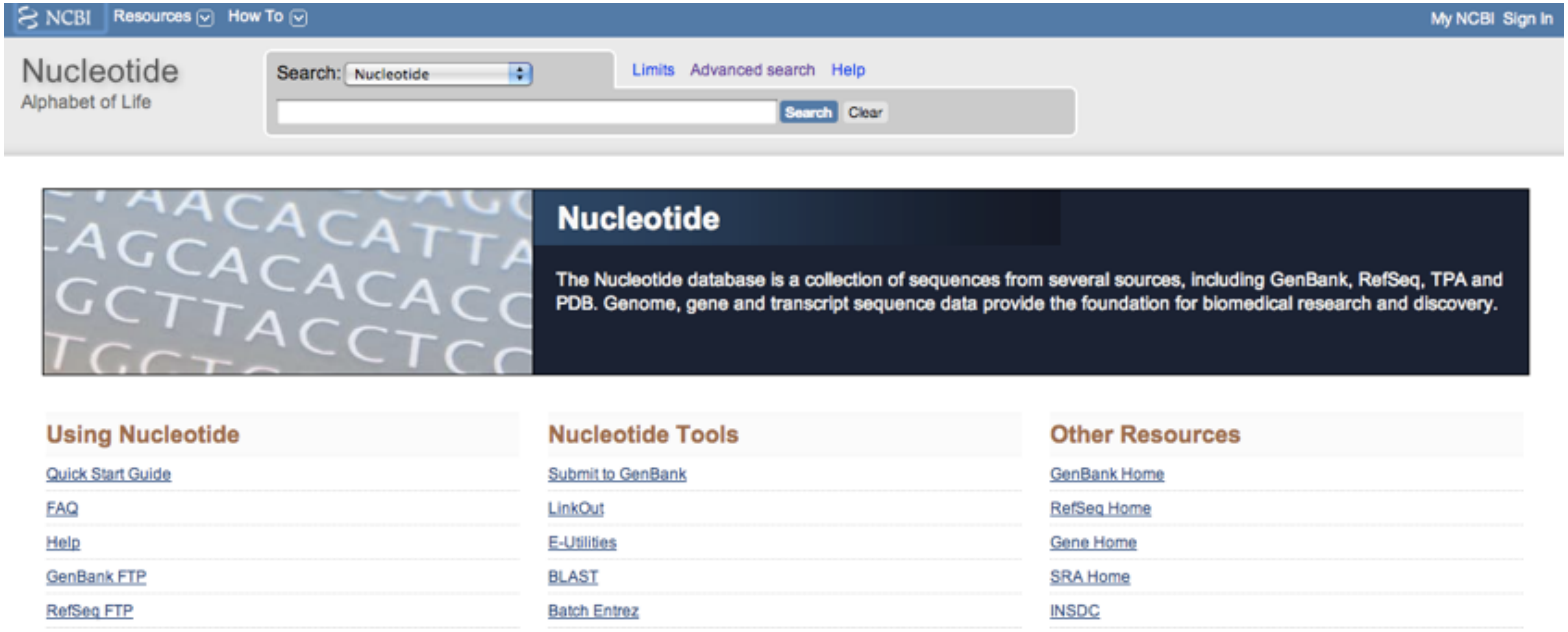
NCBI (National Center for Biotechnology Information)
http://www.ncbi.nlm.nih.gov/

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

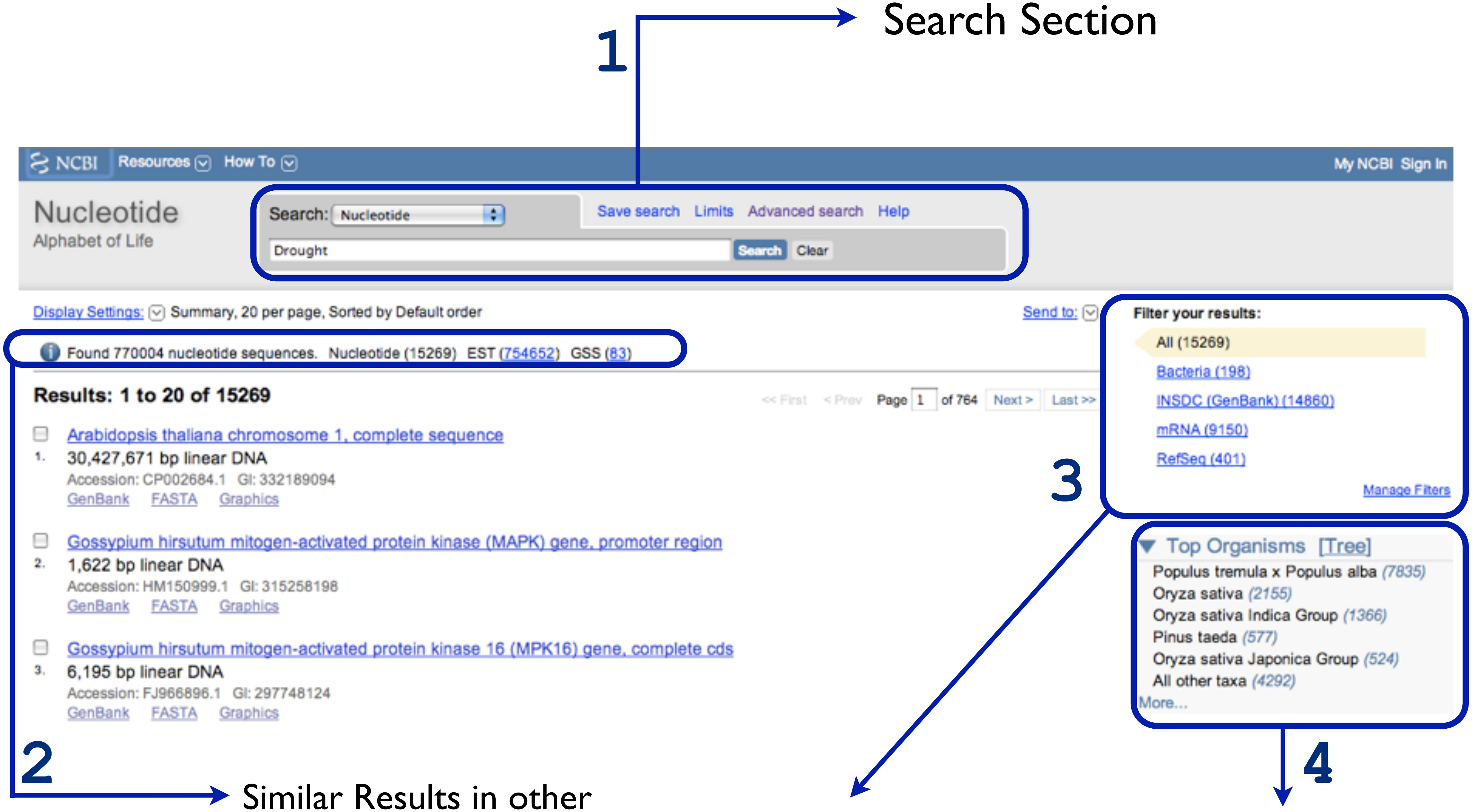


Nucleotide

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.

2. Public Repositories: NCBI

GenBank:



Nucleotide
Alphabet of Life

Search: Nucleotide

Drought

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

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

Results: 1 to 20 of 15269

Page 1 of 764

□ Arabidopsis thaliana chromosome 1, complete sequence
1. 30,427,671 bp linear DNA
Accession: CP002684.1 GI: 332189094
GenBank FASTA Graphics

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

□ Gossypium hirsutum mitogen-activated protein kinase 16 (MPK16) gene, complete cds
3. 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)

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)*

Search Section

Similar Results in other sections

Sequence Type Filter

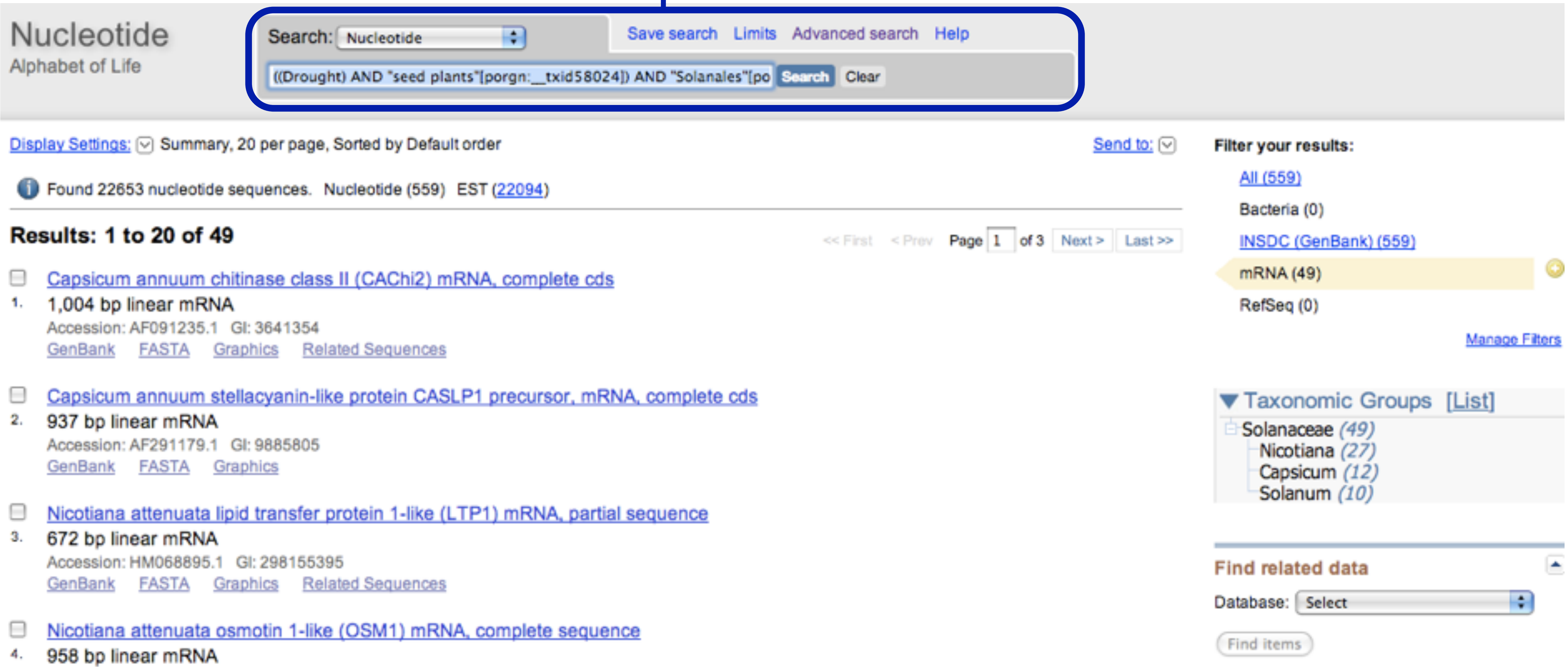
Taxonomic Filter

2. Public Repositories: NCBI

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 >>

□ 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

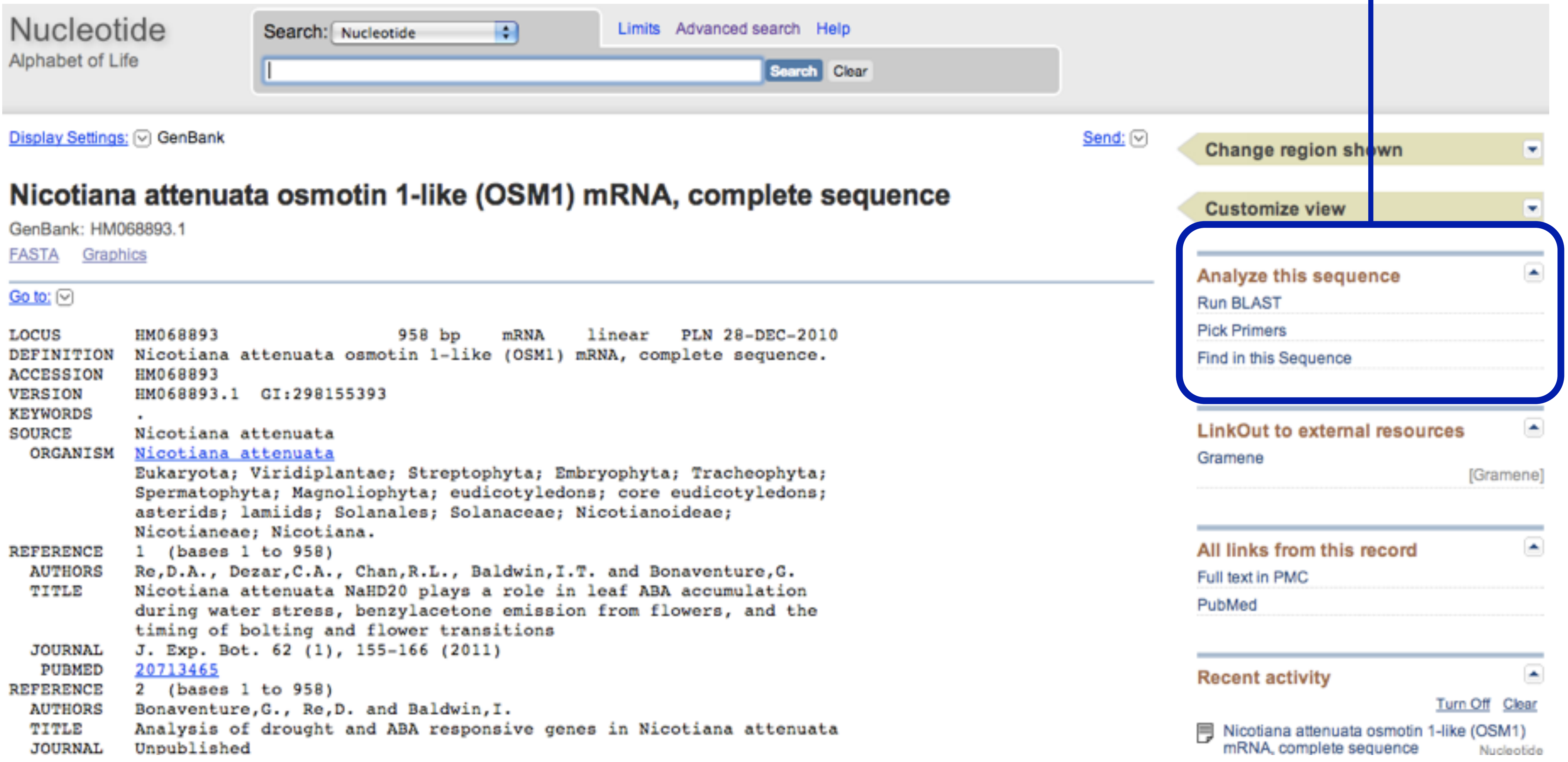
Taxonomic Groups [List]
Solanaceae (49)
Nicotiana (27)
Capsicum (12)
Solanum (10)

Find related data
Database: Select
Find items

2. Public Repositories: NCBI

GenBank:

Tools Links



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

GenBank: HM068893.1

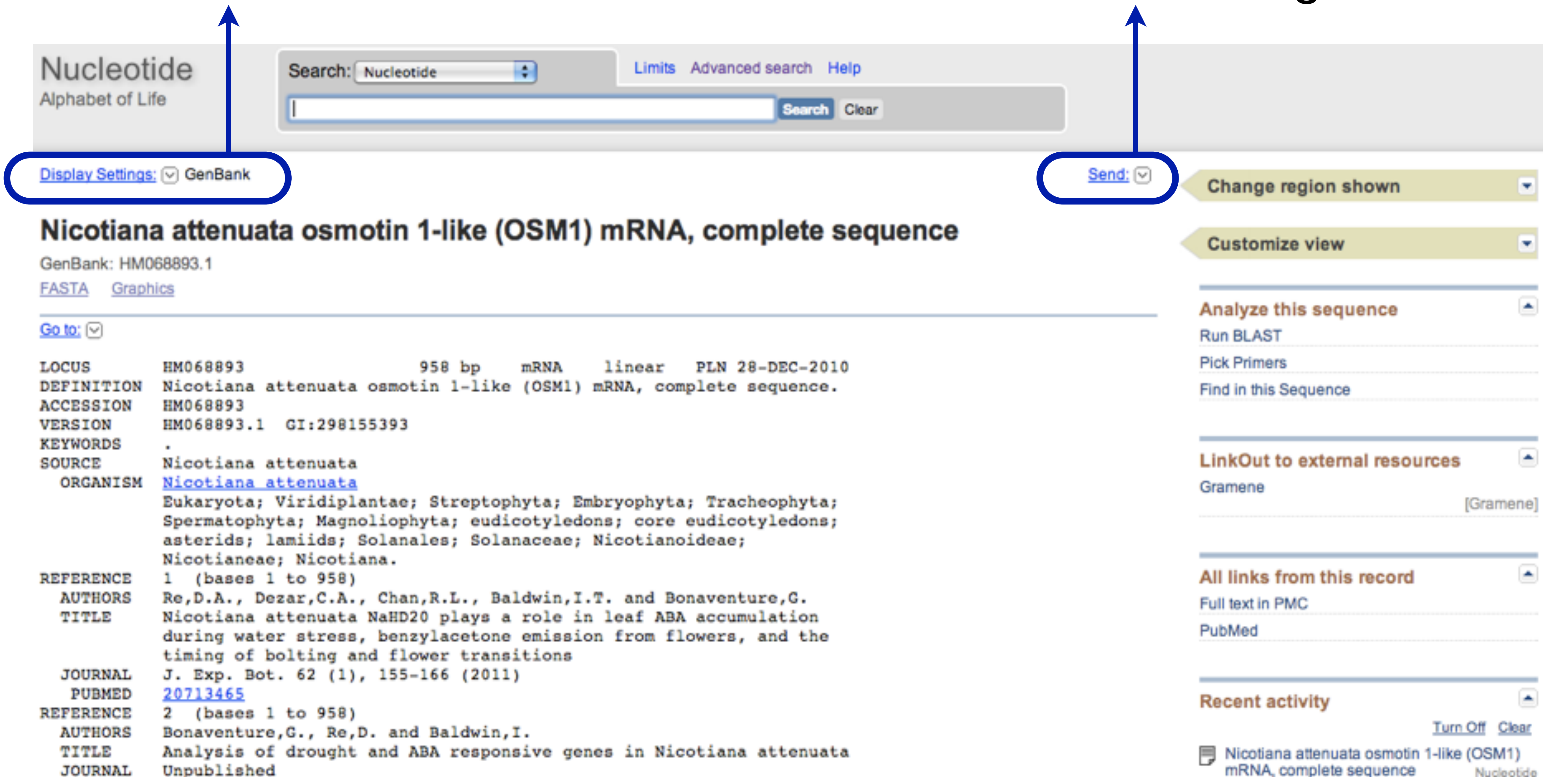
```
LOCUS       HM068893                 958 bp    mRNA    linear   PLN 28-DEC-2010
DEFINITION  Nicotiana attenuata osmotin 1-like (OSM1) mRNA, complete sequence.
ACCESSION   HM068893
VERSION     HM068893.1  GI:298155393
KEYWORDS    .
SOURCE      Nicotiana attenuata
  ORGANISM  Nicotiana attenuata
            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:

Format

File Storage



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

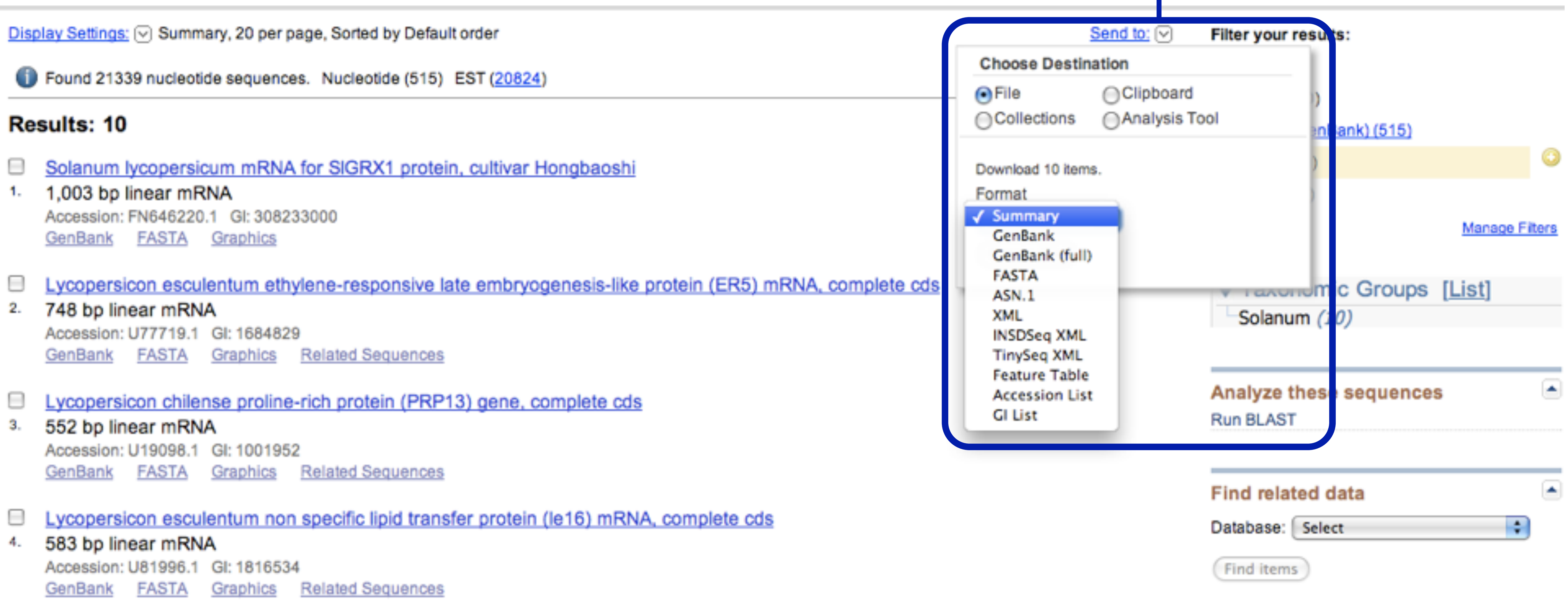
GenBank: HM068893.1

```
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

Send to:

Filter your results:

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

Results: 10

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 10 items.

Format

1. □ Solanum lycopersicum mRNA for SlGRX1 protein, cultivar Hongbaoshi
   1,003 bp linear mRNA
   Accession: FN646220.1 GI: 308233000
   GenBank FASTA Graphics

2. □ Lycopersicon esculentum ethylene-responsive late embryogenesis-like protein (ER5) mRNA, complete cds
   748 bp linear mRNA
   Accession: U77719.1 GI: 1684829
   GenBank FASTA Graphics Related Sequences

3. □ Lycopersicon chilense proline-rich protein (PRP13) gene, complete cds
   552 bp linear mRNA
   Accession: U19098.1 GI: 1001952
   GenBank FASTA Graphics Related Sequences

4. □ Lycopersicon esculentum non specific lipid transfer protein (le16) mRNA, complete cds
   583 bp linear mRNA
   Accession: U81996.1 GI: 1816534
   GenBank FASTA Graphics Related Sequences

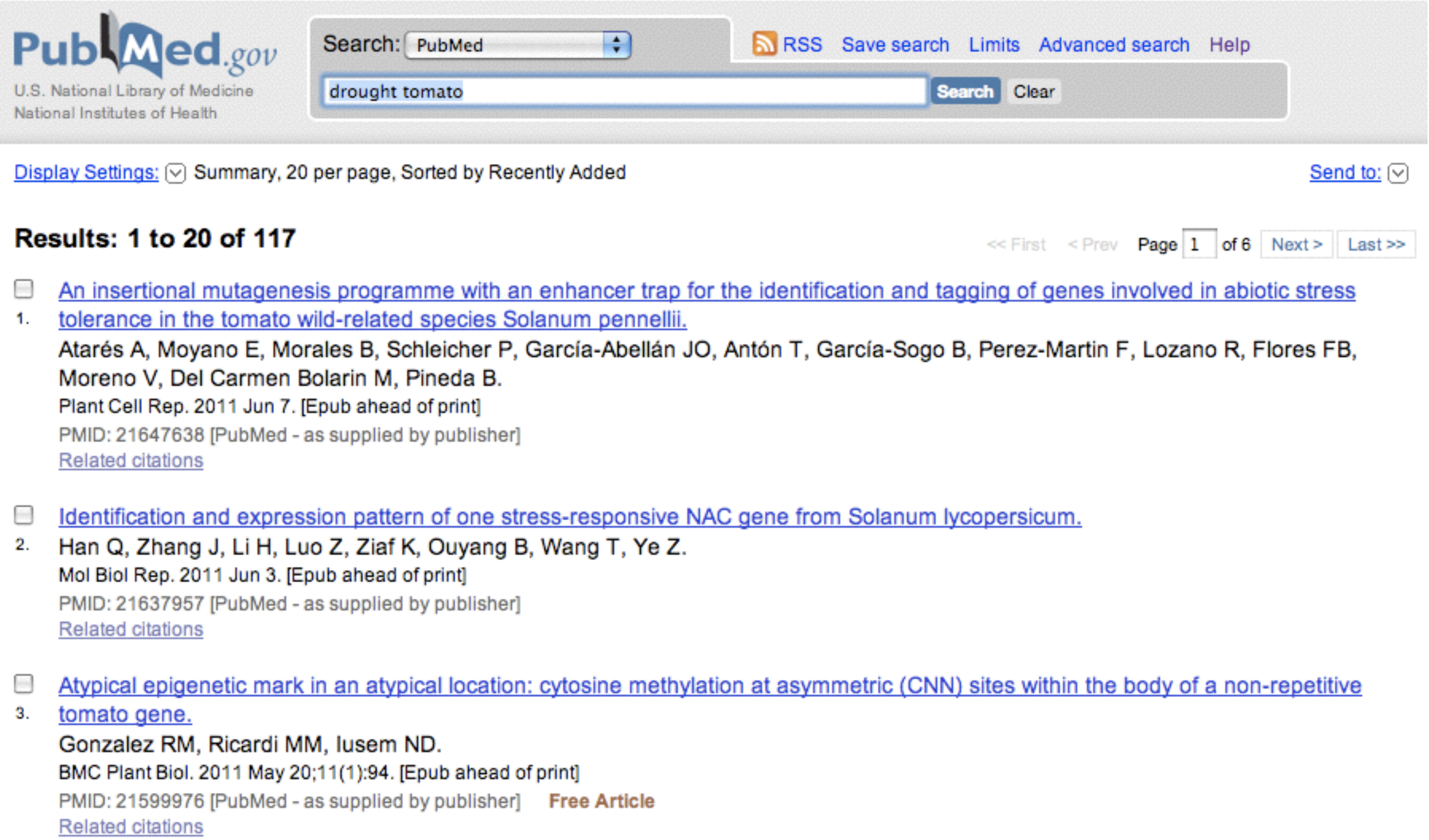
Analyze these sequences
Run BLAST

Find related data
Database: Select
Find items

2. Public Repositories: NCBI

PubMed, NIH database for scientific literature and publications.

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



drought tomato

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

Results: 1 to 20 of 117

Page 1 of 6

□ 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).

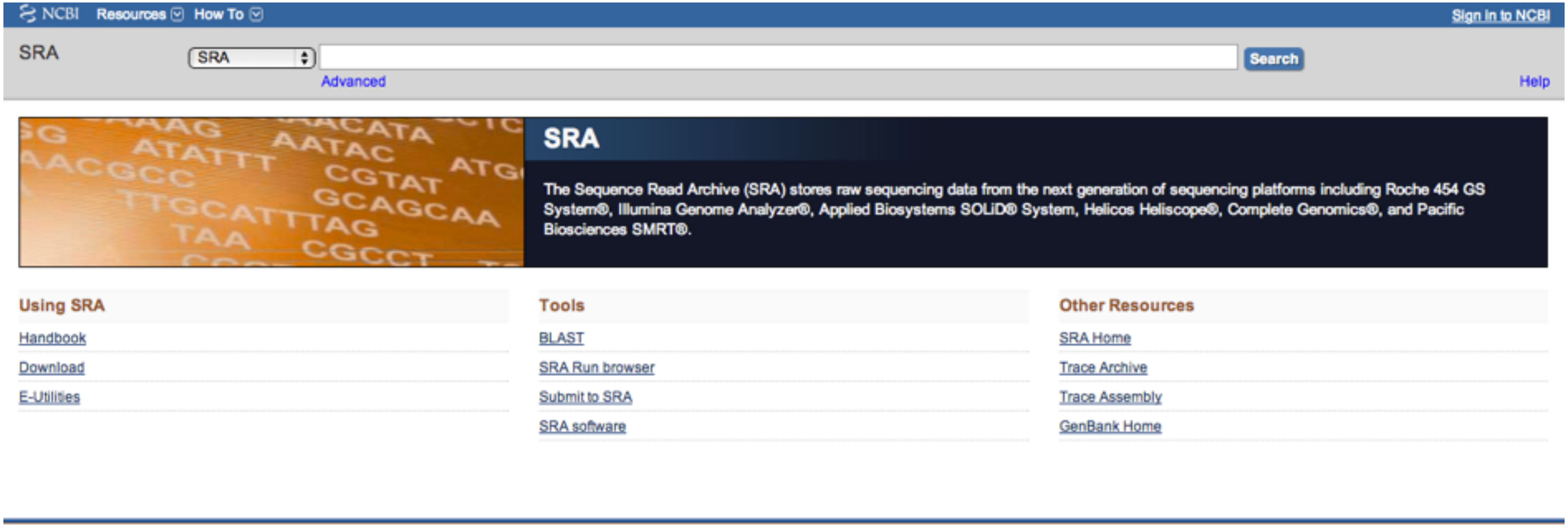
- It doesn't have all plant science related journals (for example: Theoretical Applied and Genetics or Crop Science).

  (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



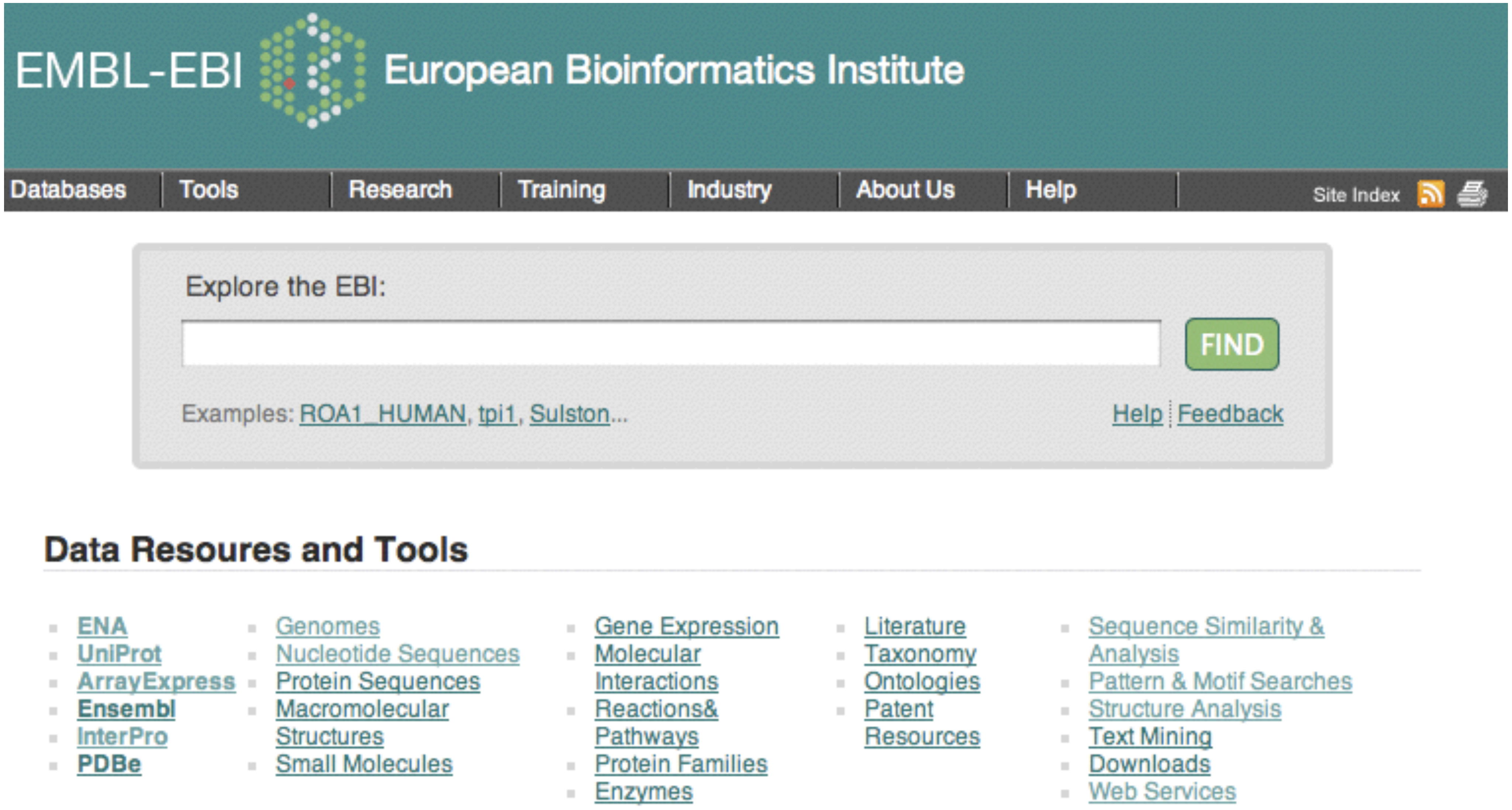
The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Using SRA	Tools	Other Resources
Handbook	BLAST	SRA Home
Download	SRA Run browser	Trace Archive
E-Utilities	Submit to SRA	Trace Assembly
	SRA software	GenBank Home

2. Public Repositories: EBI

EBI (European Bioinformatics Institute)

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



Explore the EBI:

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

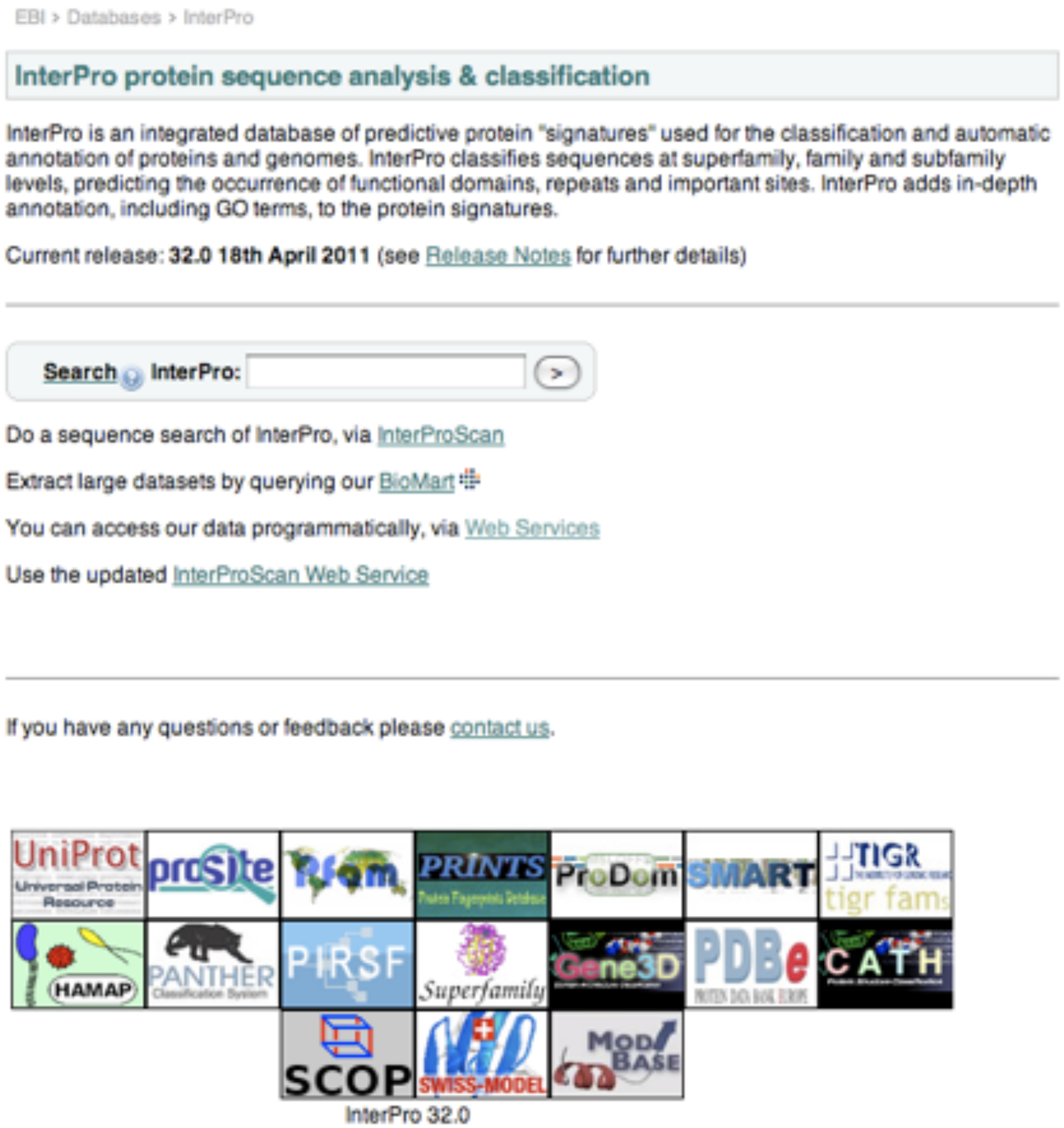
EBI (European Bioinformatics Institute)
http://www.ebi.ac.uk/

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)

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

1. Types.

- 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

3. Community specific databases

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/
Phytozome	Plants	Multiple Species Genomes	http://www.phytozome.net
Plant Genome Database (PlantGDB)	Plants	Multiple Species Genomes, Genes, Unigenes	http://www.plantgdb.org/

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

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)

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.

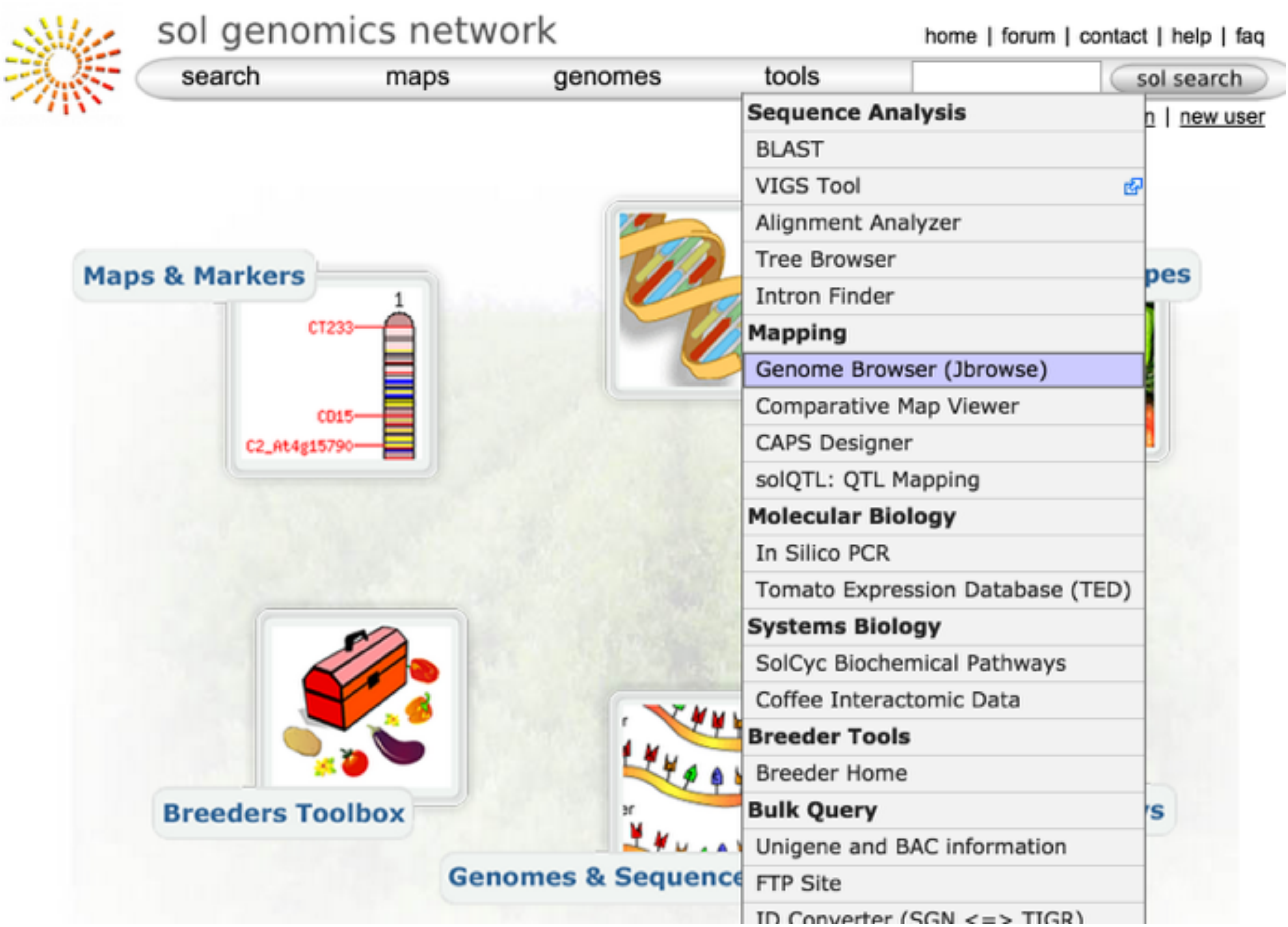
4. Genomic Browsers

Genome Browser most used:

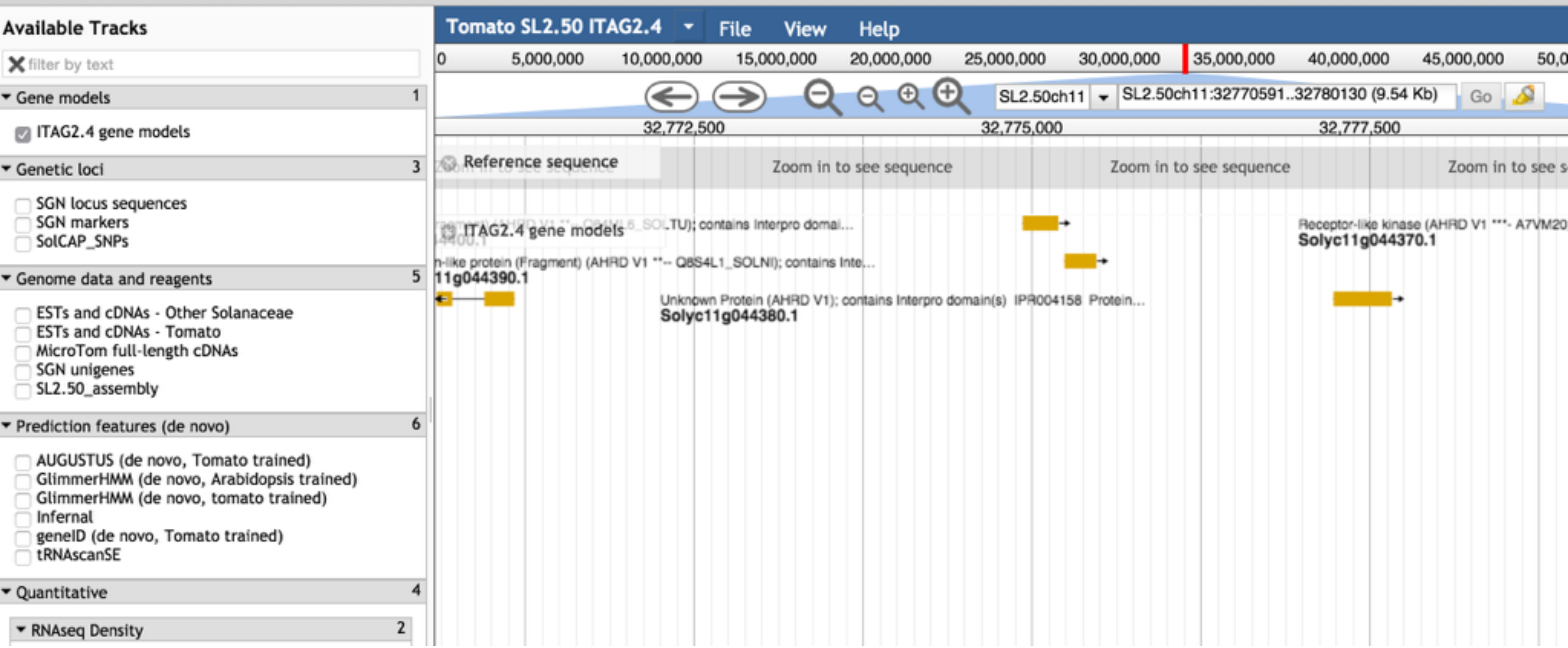
- JBrowse (GMOD).
- GBrowse (GMOD).
- UCSC Genome Browser.
- Emsembl Genome Browser.
- Vista Genome Browser.

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

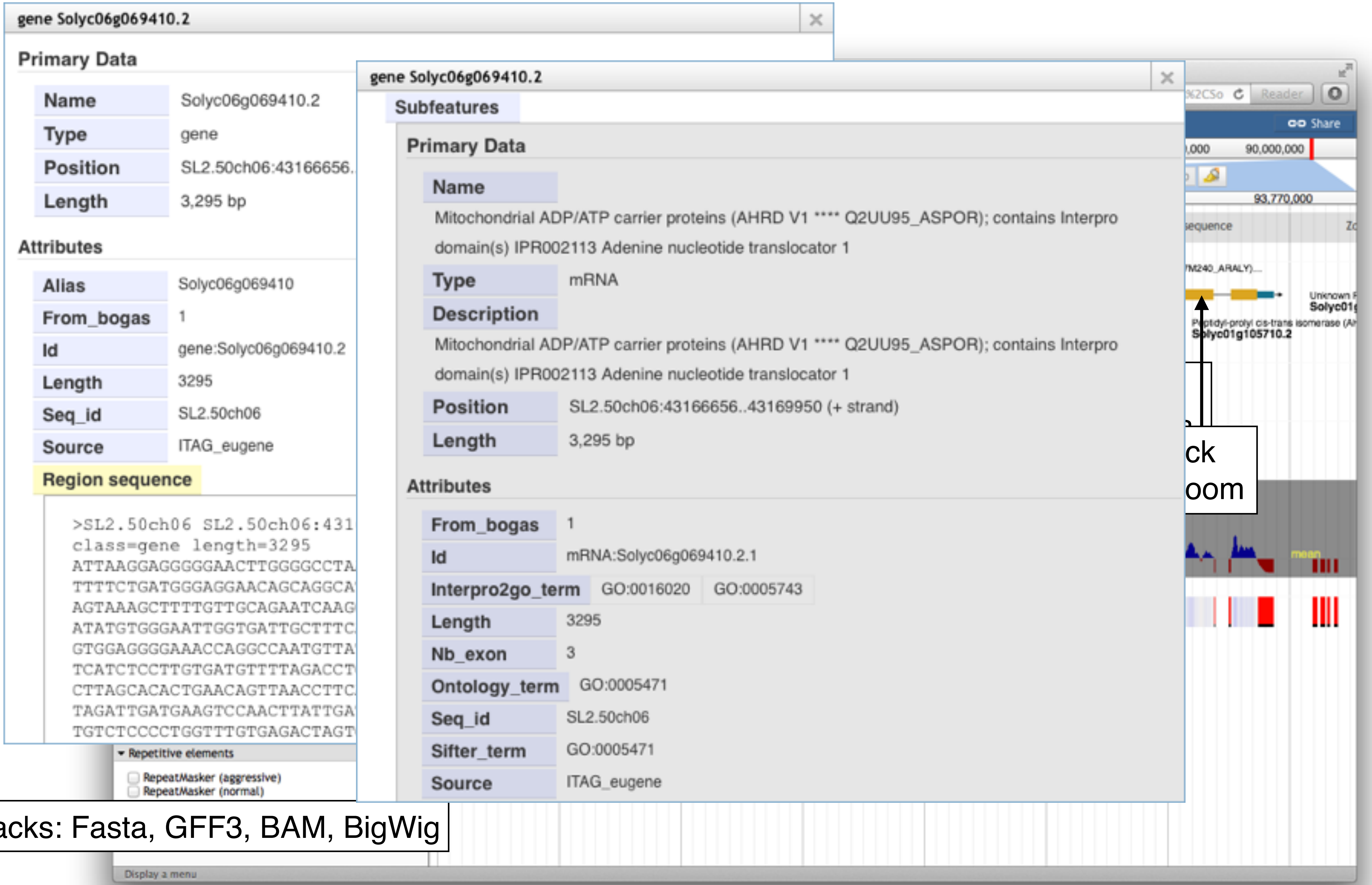
☐ AUGUSTUS (de novo, Tomato trained)
☐ GlimmerHMM (de novo, Arabidopsis trained)
☐ GlimmerHMM (de novo, tomato trained)
☐ Infernal
☐ geneID (de novo, Tomato trained)
☐ tRNAscanSE

▾ Quantitative 4

▾ RNAseq Density 2

SL2.50ch11 ▾ SL2.50ch11:32770591..32780130 (9.54 Kb) Go

JBrowse



gene Solyc06g069410.2

Primary Data

Name	Solyc06g069410.2
Type	gene
Position	SL2.50ch06:43166656.
Length	3,295 bp

Attributes

Alias	Solyc06g069410
From_bogas	1
Id	gene:Solyc06g069410.2
Length	3295
Seq_id	SL2.50ch06
Source	ITAG_eugene

Region sequence

```
>SL2.50ch06 SL2.50ch06:431
class=gene length=3295
ATTAAGGAGGGGGAACTTGGGGCCTA
TTTTCTGATGGGAGGAACAGCAGGCA
AGTAAAGCTTTTGTTGCAGAATCAAG
ATATGTGGGAATTGGTGATTGCTTTC
GTGGAGGGGAAACCAGGCCAATGTTA
TCATCTCCTTGTGATGTTTTAGACCT
CTTAGCACACTGAACAGTTAACCTTC
TAGATTGATGAAGTCCAACTTATTGA
TGTCTCCCCTGGTTTGTGAGACTAGT
```

gene Solyc06g069410.2

Subfeatures

Primary Data

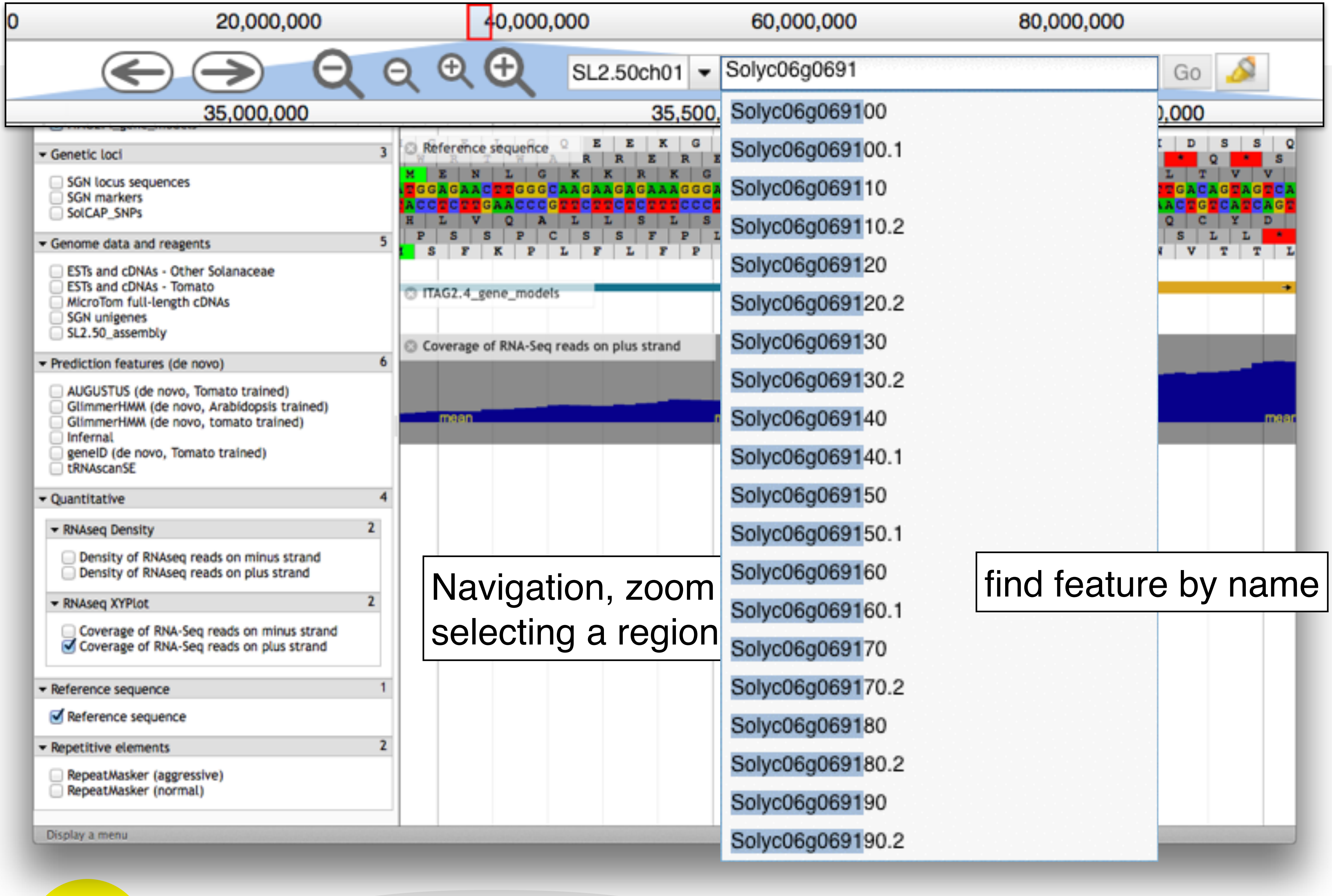
Name	Mitochondrial ADP/ATP carrier proteins (AHRD V1 **** Q2UU95_ASPOR); contains Interpro domain(s) IPR002113 Adenine nucleotide translocator 1
Type	mRNA
Description	Mitochondrial ADP/ATP carrier proteins (AHRD V1 **** Q2UU95_ASPOR); contains Interpro domain(s) IPR002113 Adenine nucleotide translocator 1
Position	SL2.50ch06:43166656..43169950 (+ strand)
Length	3,295 bp

Attributes

From_bogas	1	
Id	mRNA:Solyc06g069410.2.1	
Interpro2go_term	GO:0016020	GO:0005743
Length	3295	
Nb_exon	3	
Ontology_term	GO:0005471	
Seq_id	SL2.50ch06	
Sifter_term	GO:0005471	
Source	ITAG_eugene	

load tracks: Fasta, GFF3, BAM, BigWig

JBrowse



Solyc06g0691

- 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

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

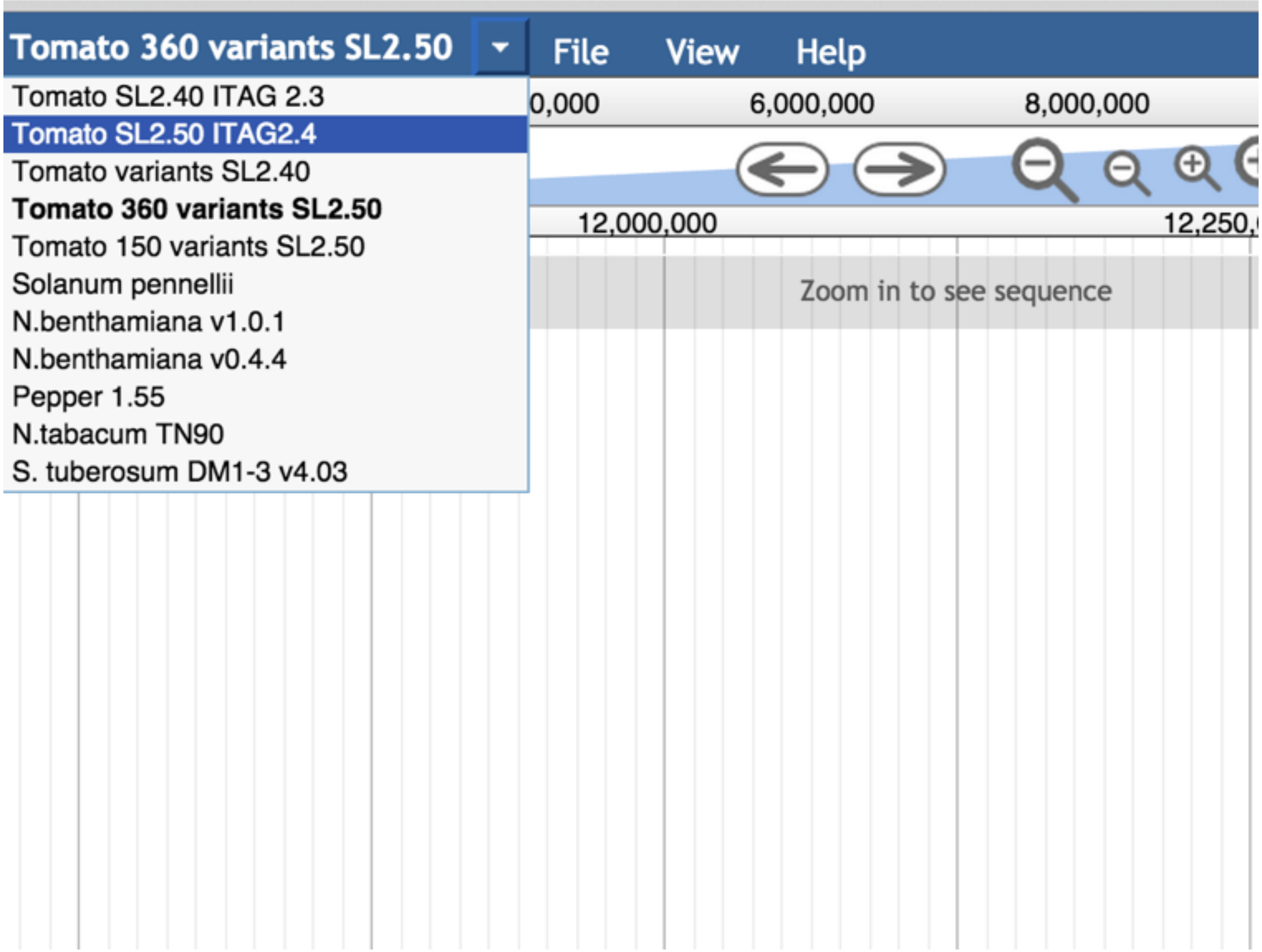
Repetitive elements 2

□ RepeatMasker (aggressive)
□ RepeatMasker (normal)

Navigation, zoom selecting a region

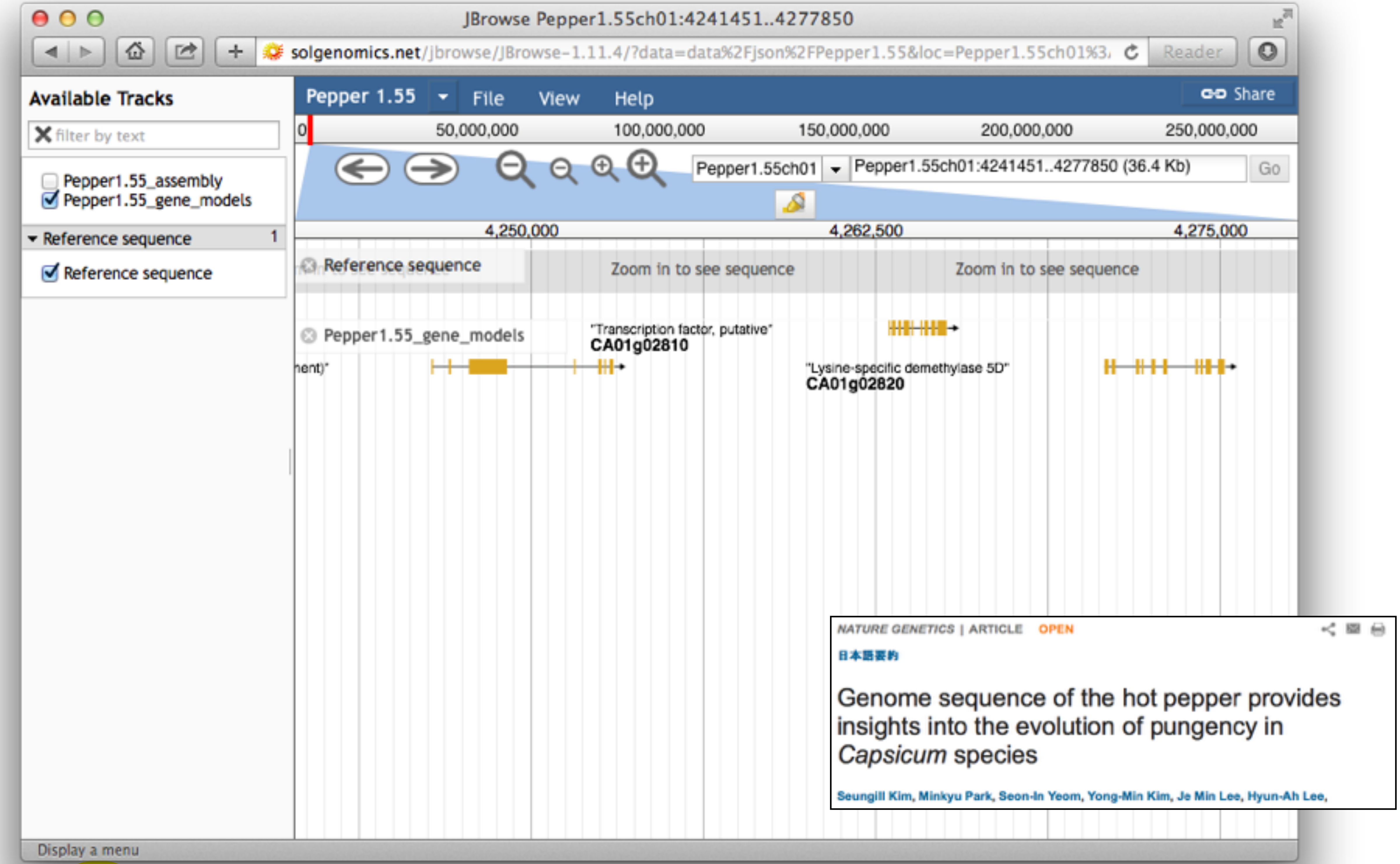
find feature by name

JBrowse



- 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

JBrowse: Pepper genome



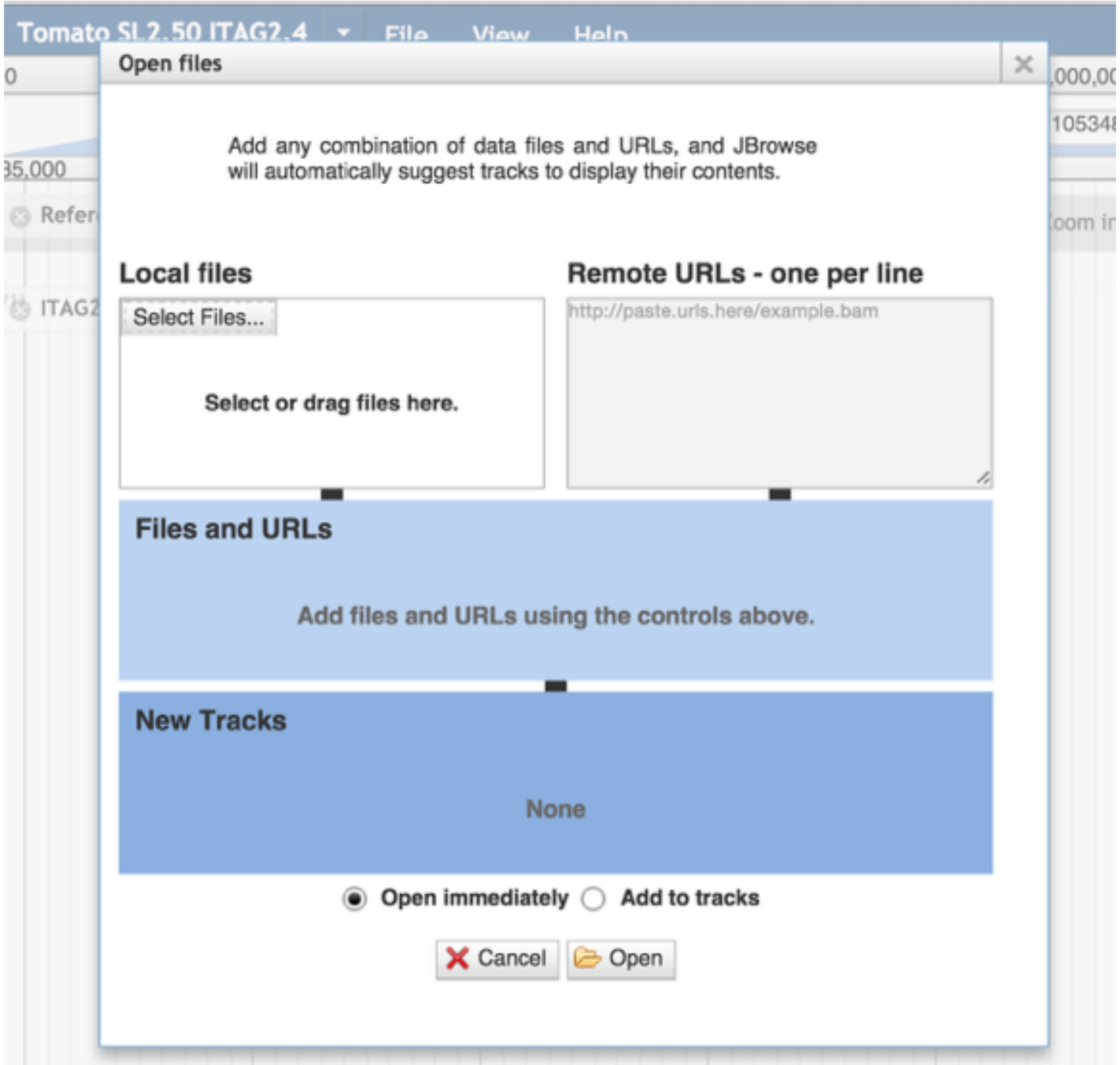
NATURE GENETICS | ARTICLE OPEN

日本語要約

Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species

Seungill Kim, Minkyu Park, Seon-In Yeom, Yong-Min Kim, Je Min Lee, Hyun-Ah Lee,

JBrowse



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

Exercise 1

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 plant caffeine synthase protein sequences are in GenBank? How many are from *Coffea arabica*?

    3. How many species have a caffeine synthase homolog?

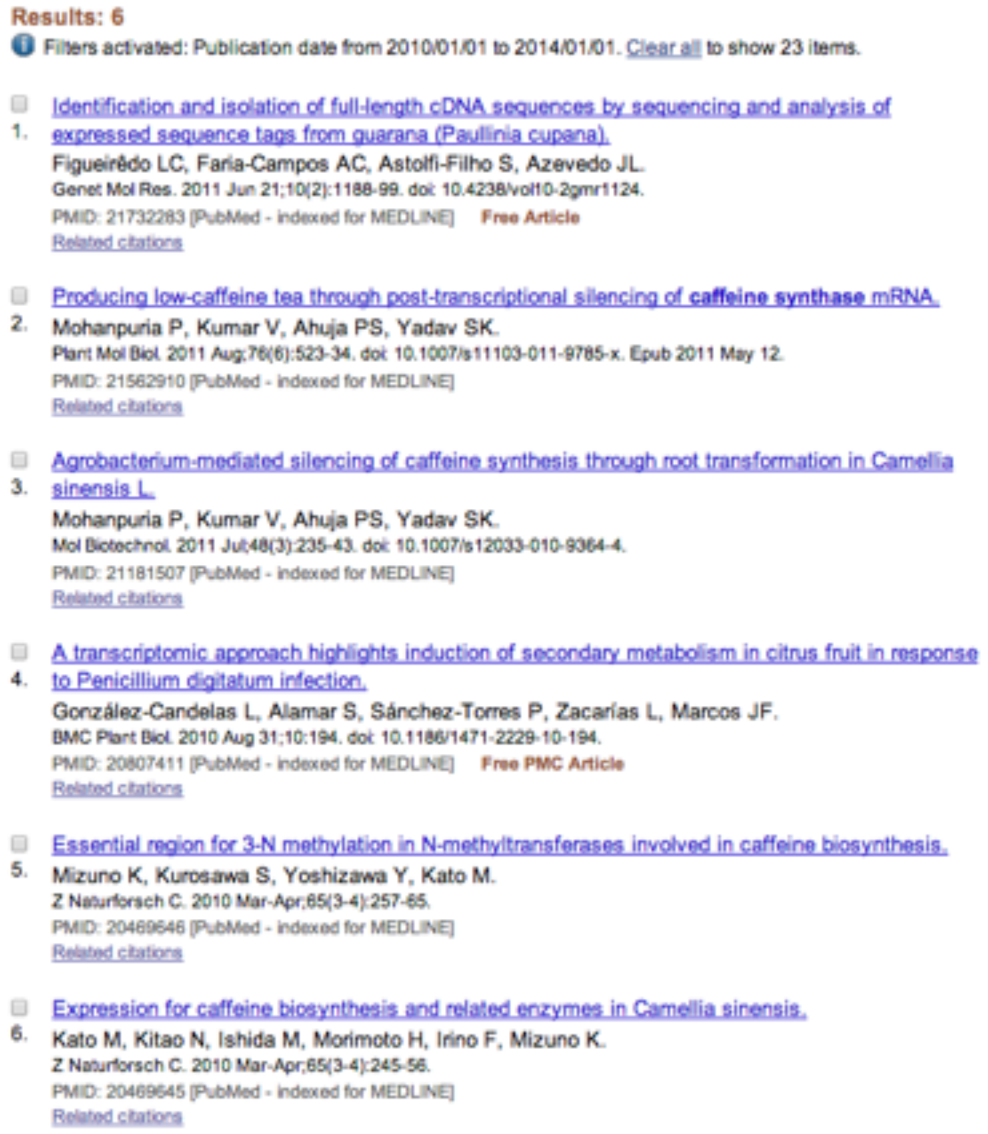
    4. Is caffeine synthase specific to the Gentianales clade or is it found elsewhere?

    5. Which of the homologs seem realistic? Download all *Coffea canephora* homolog sequences in fasta format and select full-length proteins. How many appear full-length?

Please save your results for the next exercise.

Exercise 1 Solutions

1. Find some papers on caffeine synthase published since 2010.
   - 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

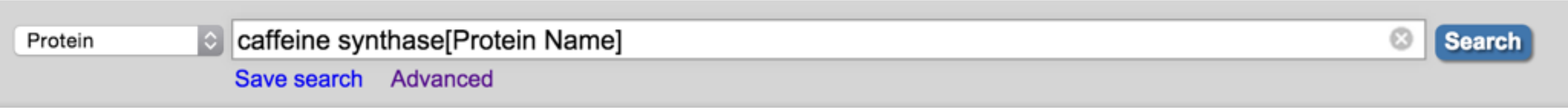
□ 4. A transcriptomic approach highlights induction of secondary metabolism in citrus fruit in response to Penicillium digitatum infection.
González-Candelas L, Alamar S, Sánchez-Torres P, Zacarías L, Marcos JF.
BMC Plant Biol. 2010 Aug 31;10:194. doi: 10.1186/1471-2229-10-194.
PMID: 20807411 [PubMed - 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

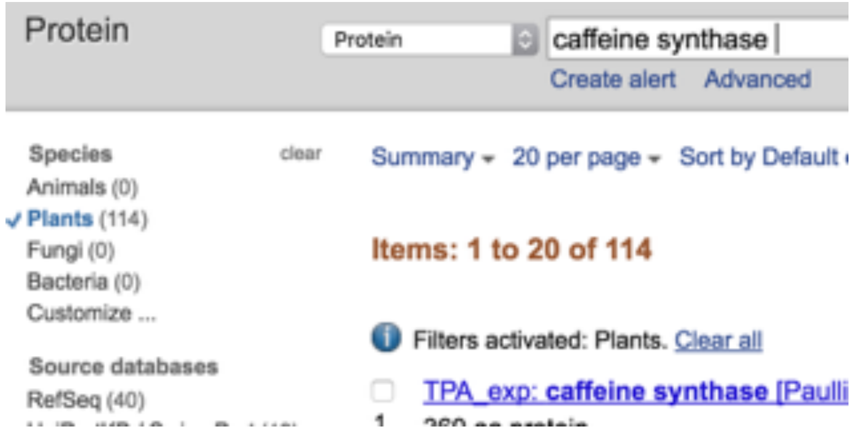
Exercise 1 Solutions

2. How many plant caffeine synthase protein sequences are in GenBank? How many are from *Coffea arabica*?
   - 114 proteins total are from plants, 21 from *C. arabica* (http://www.ncbi.nlm.nih.gov/protein)



Protein

caffeine synthase[Protein Name]



Protein

Protein

caffeine synthase

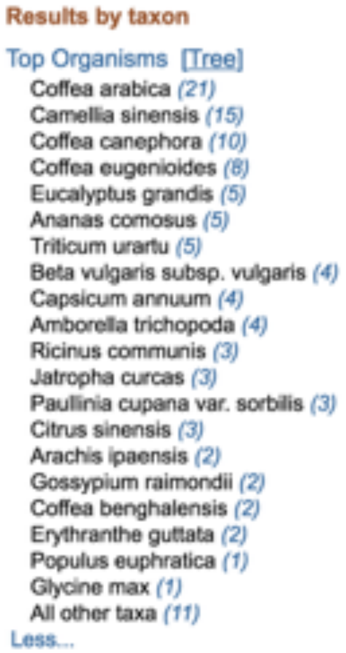
Species clear
Animals (0)
✓ Plants (114)
Fungi (0)
Bacteria (0)
Customize ...

Source databases
RefSeq (40)

Items: 1 to 20 of 114

Filters activated: Plants. Clear all

□ TPA_exp: caffeine synthase [Paulli
1 360 aa protein



Results by taxon

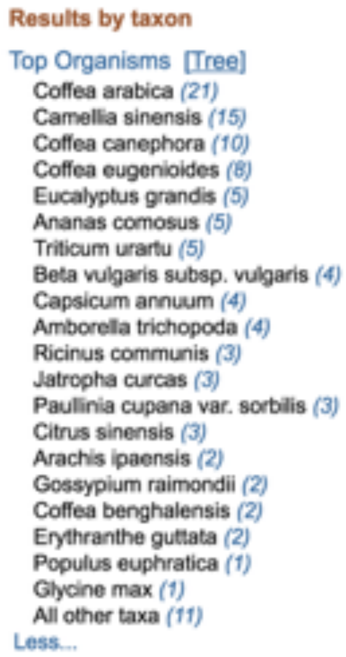
Top Organisms [Tree]

Coffea arabica	(21)
Camellia sinensis	(15)
Coffea canephora	(10)
Coffea eugenioides	(8)
Eucalyptus grandis	(5)
Ananas comosus	(5)
Triticum urartu	(5)
Beta vulgaris subsp. vulgaris	(4)
Capsicum annuum	(4)
Amborella trichopoda	(4)
Ricinus communis	(3)
Jatropha curcas	(3)
Paullinia cupana var. sorbilis	(3)
Citrus sinensis	(3)
Arachis ipaensis	(2)
Gossypium raimondii	(2)
Coffea benghalensis	(2)
Erythranthe guttata	(2)
Populus euphratica	(1)
Glycine max	(1)
All other taxa	(11)

Less...

Exercise 1 Solutions

3. 31 species:



Results by taxon

Top Organisms [Tree]

Coffea arabica (21)
Camellia sinensis (15)
Coffea canephora (10)
Coffea eugenioides (8)
Eucalyptus grandis (5)
Ananas comosus (5)
Triticum urartu (5)
Beta vulgaris subsp. vulgaris (4)
Capsicum annuum (4)
Amborella trichopoda (4)
Ricinus communis (3)
Jatropha curcas (3)
Paullinia cupana var. sorbilis (3)
Citrus sinensis (3)
Arachis ipaensis (2)
Gossypium raimondii (2)
Coffea benghalensis (2)
Erythranthe guttata (2)
Populus euphratica (1)
Glycine max (1)
All other taxa (11)

Less...

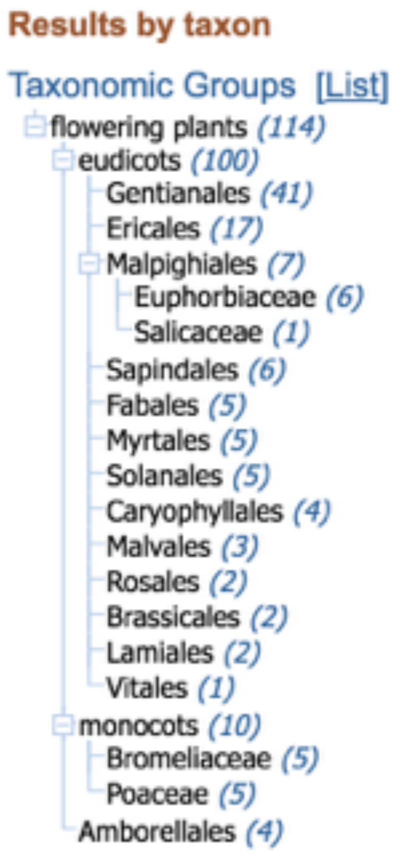
*click on "Tree" for next question

Exercise 1 Solutions (cont'd)

4. How many species have a caffeine synthase homolog?

- Found in Gentianales, Ericales, Sapindales, Malvales, etc

*click on "List" for next answer



Results by taxon

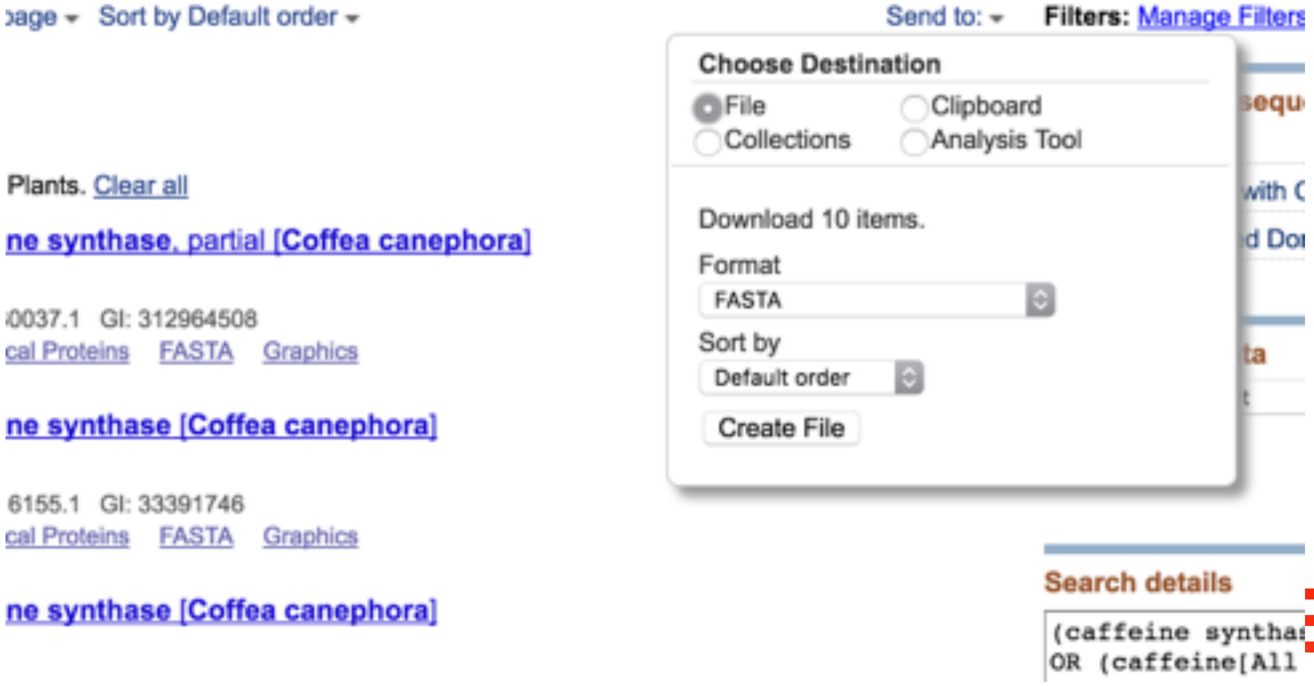
Taxonomic Groups [List]

- flowering plants (114)
  - eudicots (100)
    - Gentianales (41)
    - Ericales (17)
    - Malpighiales (7)
      - Euphorbiaceae (6)
      - Salicaceae (1)
    - Sapindales (6)
    - Fabales (5)
    - Myrtales (5)
    - Solanales (5)
    - Caryophyllales (4)
    - Malvales (3)
    - Rosales (2)
    - Brassicales (2)
    - Lamiales (2)
    - Vitales (1)
  - monocots (10)
    - Bromeliaceae (5)
    - Poaceae (5)
  - Amborellales (4)

Exercise 1 Solutions (cont'd)

5. Which of the homologs seem realistic? Download all *Coffea canephora* homolog sequences in fasta format and select full-length proteins. How many appear full-length?

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



›age ▾ Sort by Default order ▾

Send to: ▾ Filters: Manage Filters

Plants. Clear all

ne synthase, partial [Coffea canephora]

›0037.1 GI: 312964508

cal Proteins FASTA Graphics

ne synthase [Coffea canephora]

6155.1 GI: 33391746

cal Proteins FASTA Graphics

ne synthase [Coffea canephora]

Choose Destination

● File ○ Clipboard

○ Collections ○ Analysis Tool

Download 10 items.

Format: FASTA

Sort by: Default order

Create File

sequ

with C

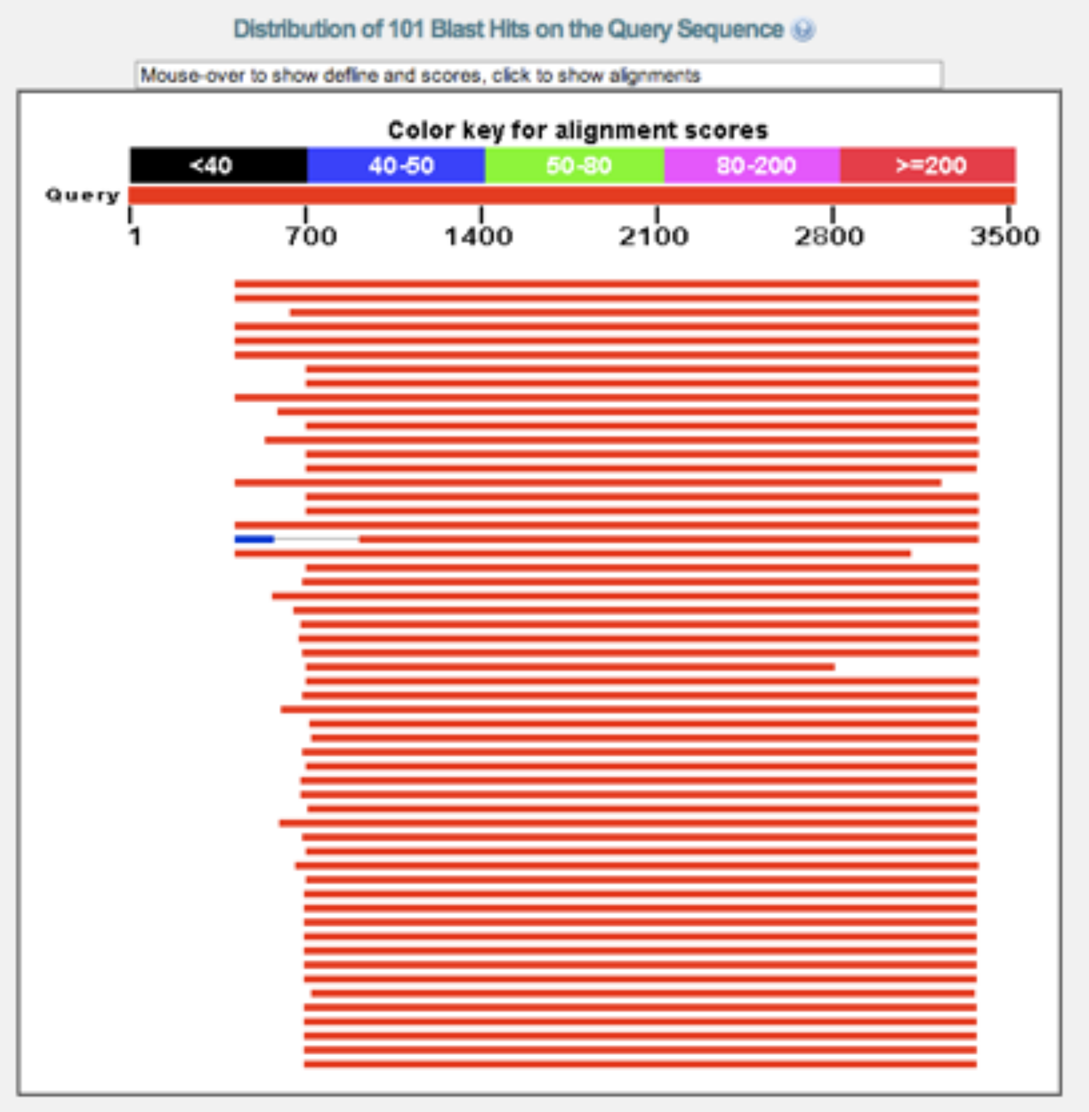
d Do

ta

Search details

```
(caffeine synthas
OR (caffeine[All
```

Part II: Web Tools


Distribution of 101 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
700
1400
2100
2800
3500

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).

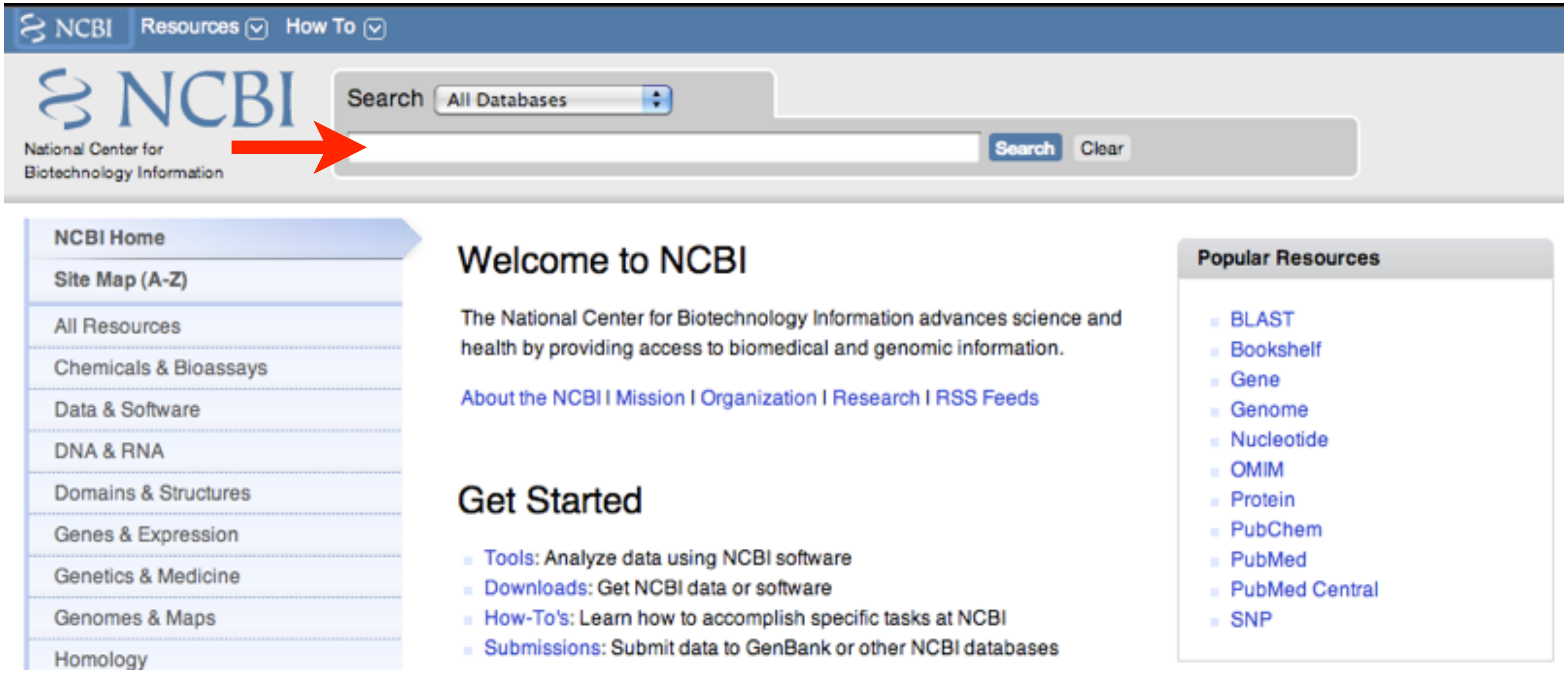
  3.2 - Primer Design (Primer3).

4 - Web Pages with Multiple Tools.

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/



NCBI Home
Site Map (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology

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 | RSS Feeds

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

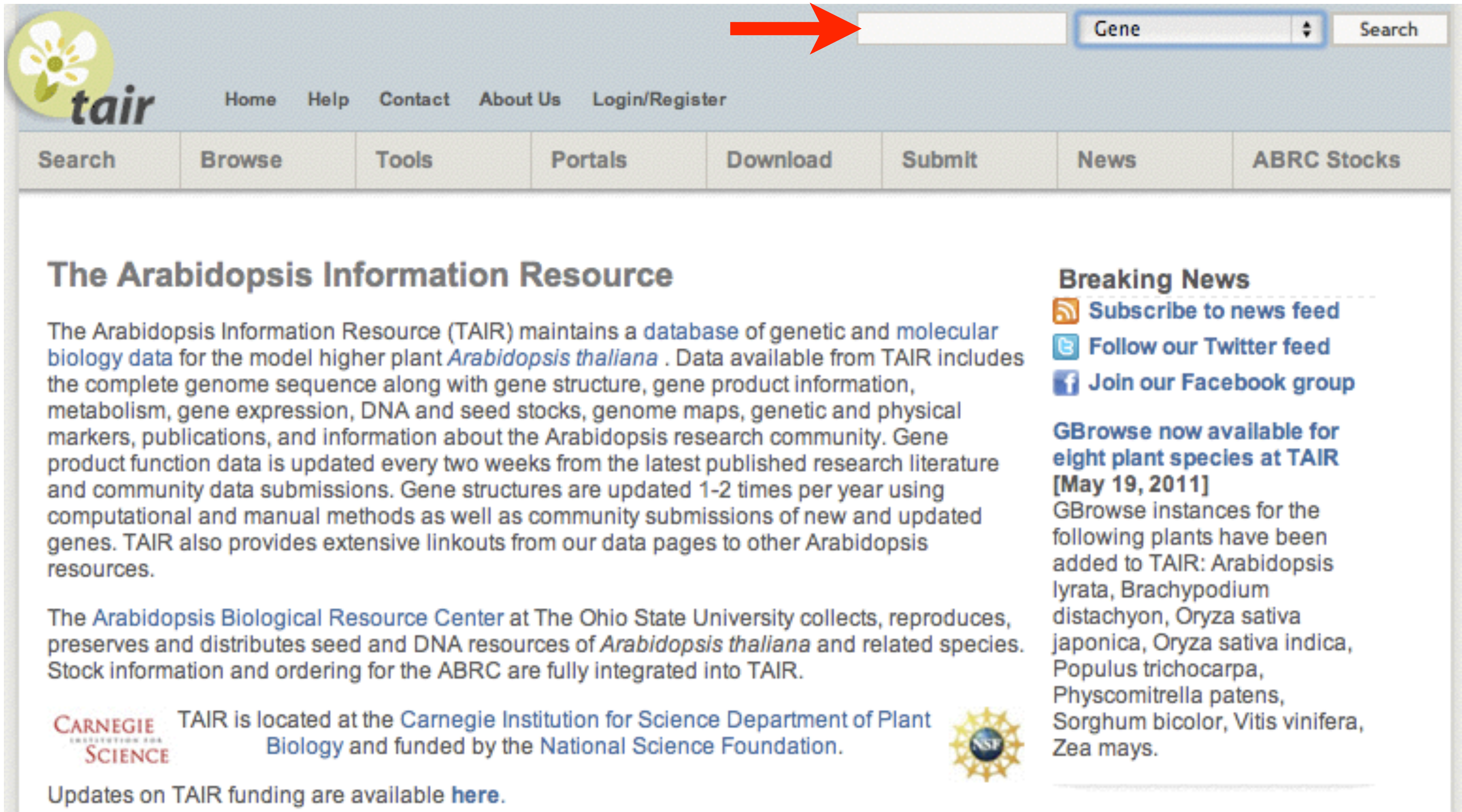
Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

1. Search Tools

Text Searches:

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



Gene
Search
tair
Home Help Contact About Us Login/Register
Search Browse Tools Portals Download Submit News ABRC Stocks

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 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.

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

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:

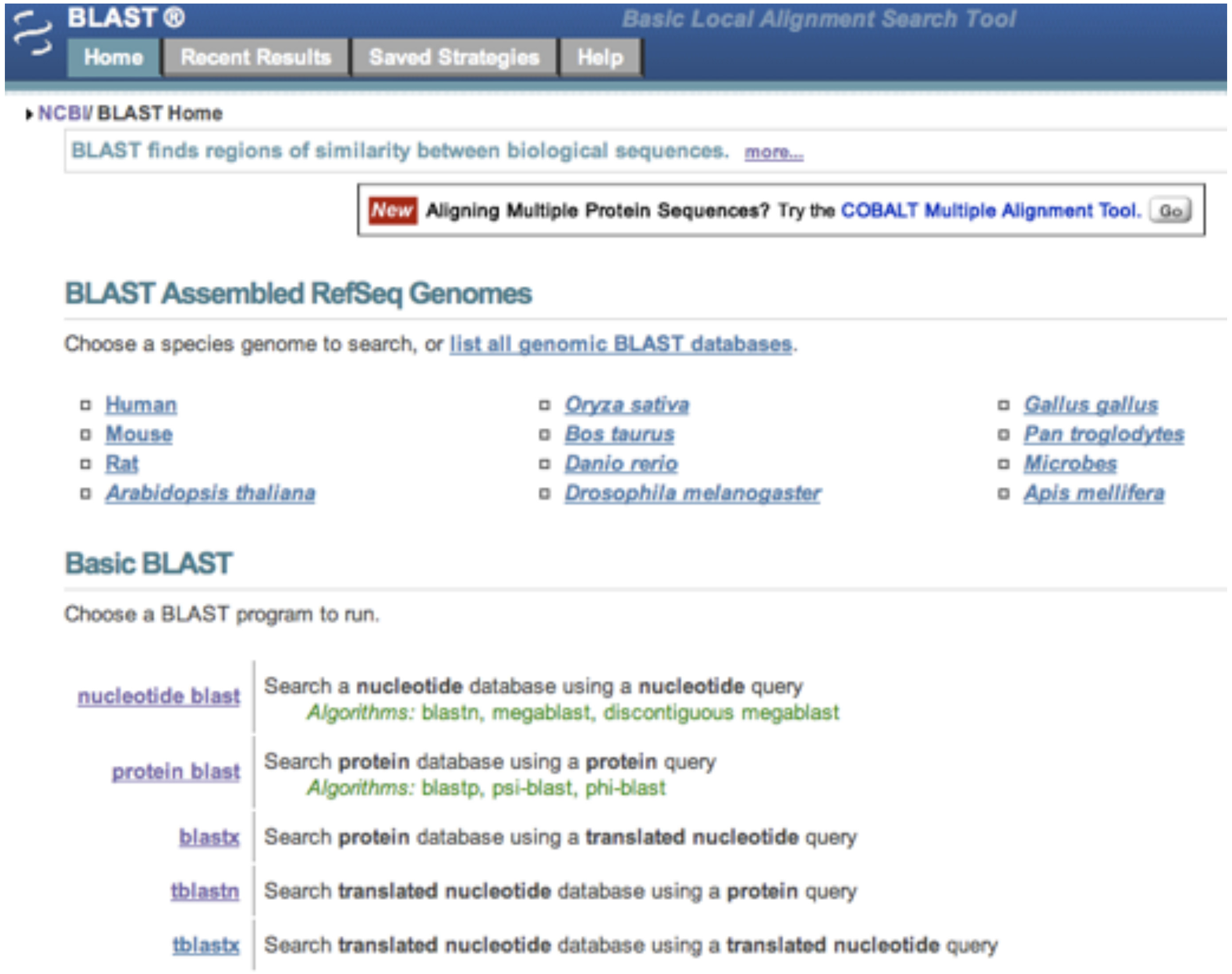
- Blast: (faster) http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Fasta (sensitive): http://www.ebi.ac.uk/Tools/sss/fasta/

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

1. Search Tools

Sequence homology/similarity searches:

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



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

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

Basic BLAST

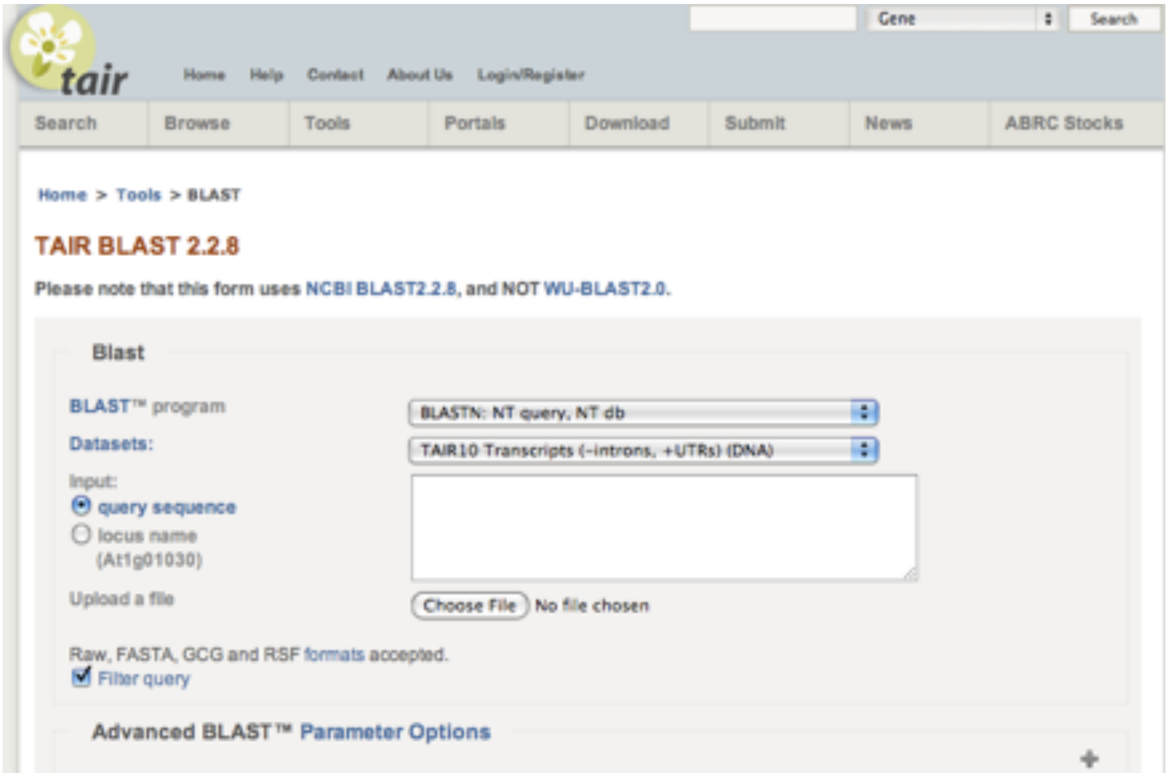
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
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

1. 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

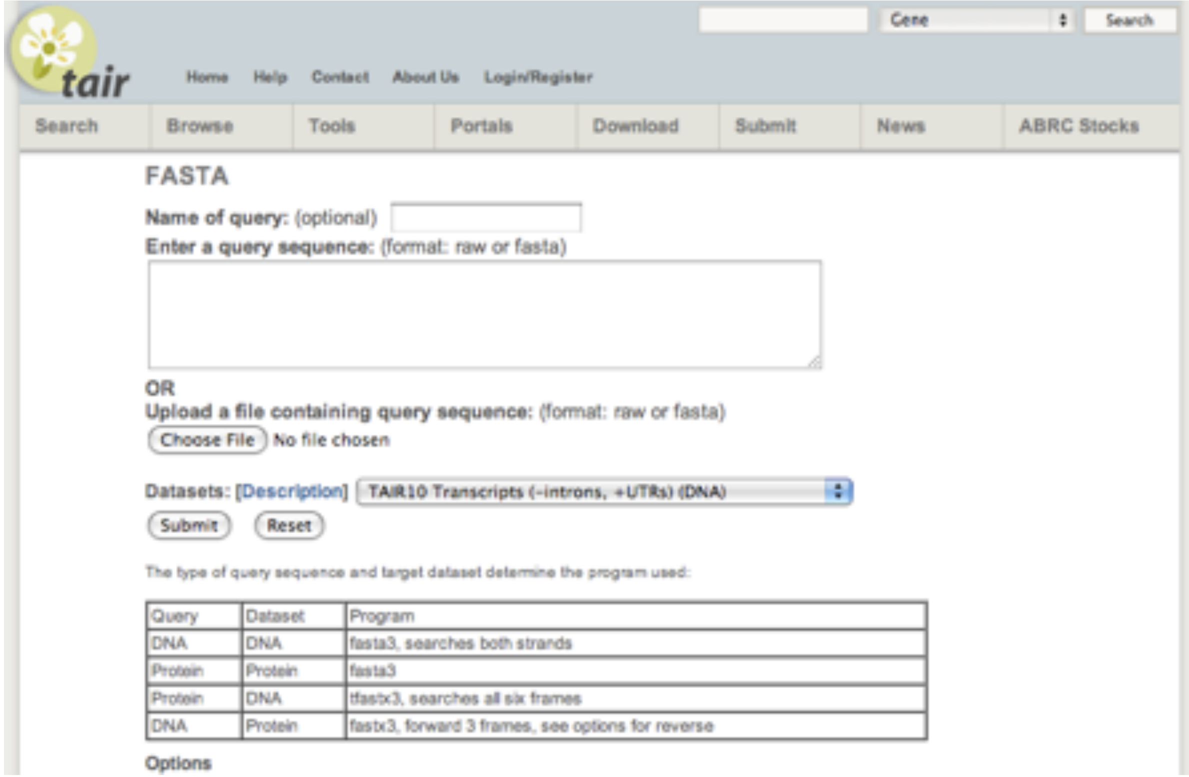
Datasets: TAIR10 Transcripts (-introns, +UTRs) (DNA)

Input:
● query sequence
○ locus name (At1g01030)

Upload a file: No file chosen

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

Advanced BLAST™ Parameter Options



FASTA

Name of query: (optional)
Enter a query sequence: (format: raw or fasta)

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

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

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

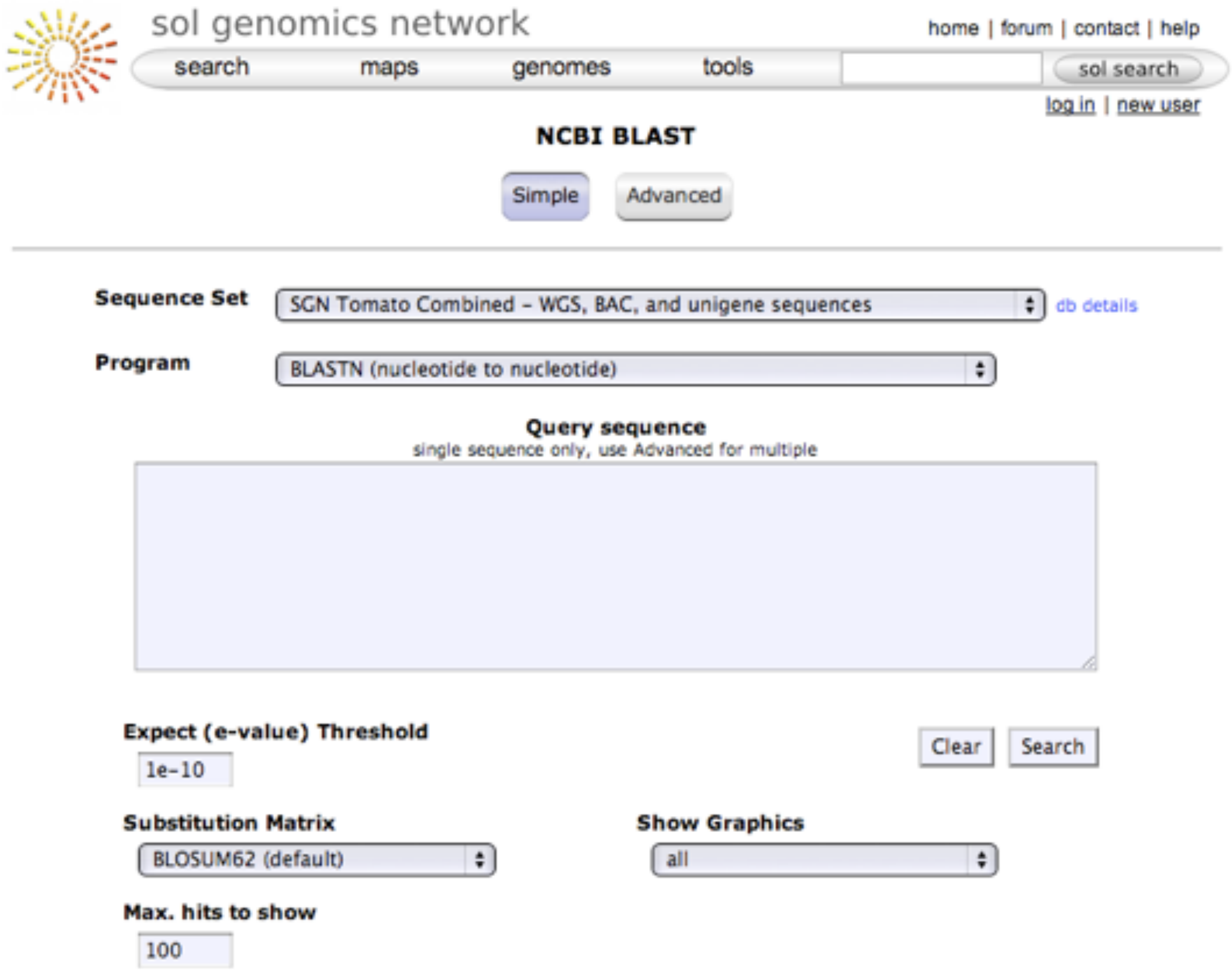
Query	Dataset	Program
DNA	DNA	fasta3, searches both strands
Protein	Protein	fasta3
Protein	DNA	tfastx3, searches all six frames
DNA	Protein	fastx3, forward 3 frames, see options for reverse

Options

1. Search Tools

Sequence homology/similarity searches:

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



NCBI BLAST

Sequence Set: SGN Tomato Combined - WGS, BAC, and unigene sequences

Program: BLASTN (nucleotide to nucleotide)

Query sequence
single sequence only, use Advanced for multiple

Expect (e-value) Threshold: 1e-10

Substitution Matrix: BLOSUM62 (default)

Show Graphics: all

Max. hits to show: 100

1.2 -Search by Sequence Homology

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.

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

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/)
  Mega (Win/OSX) (http://www.megasoftware.net/)

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.

- Select the longest ORF.
- Gene prediction based on the exon-intron structure

Tool types:

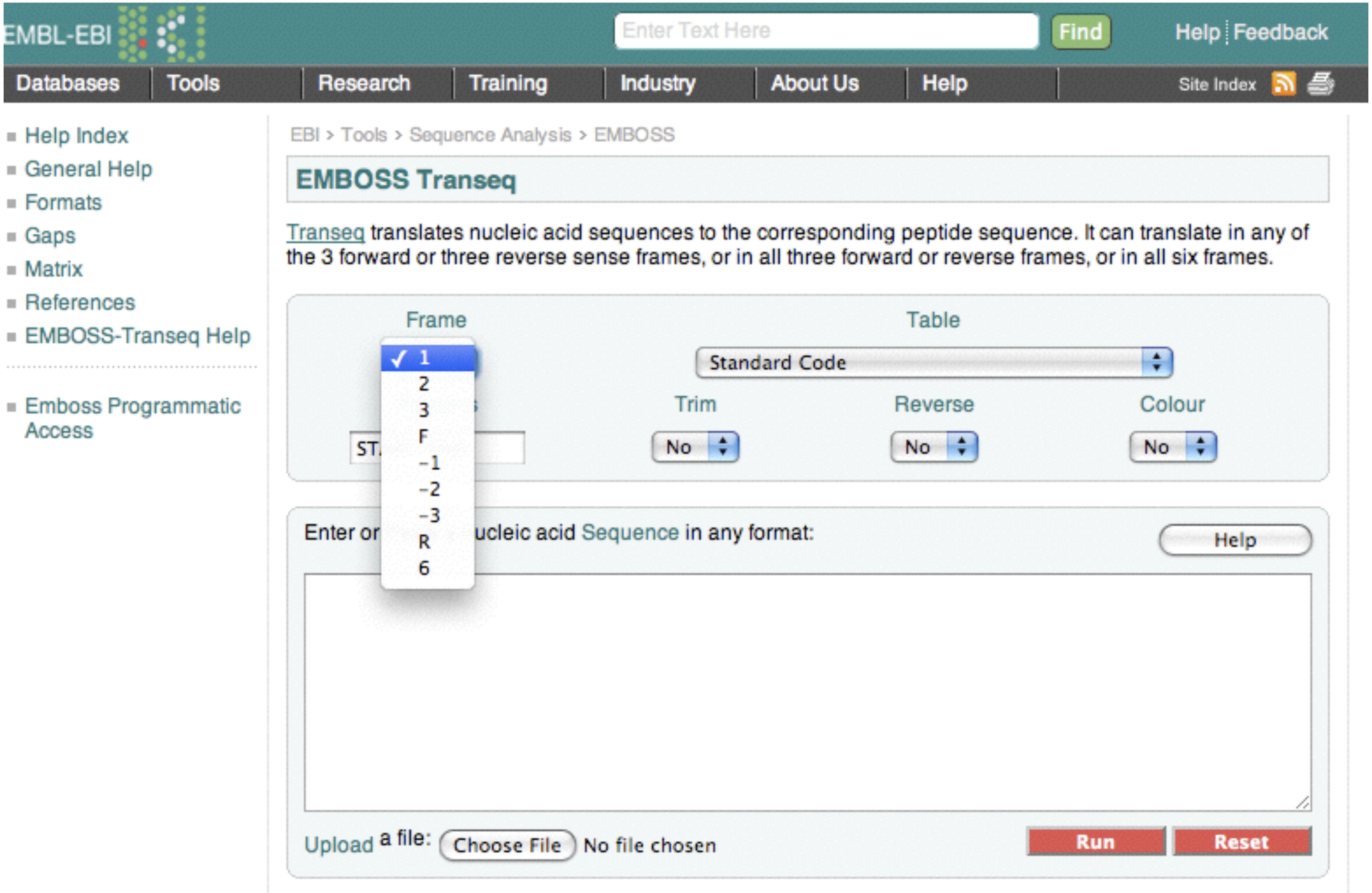
- Translators (DNA to proteins without exon-intron consideration, and analyzing all the possible ORFs). Use coding.
- *Gene Predictors* (DNA to CDS considering the intron-exon structure). They require software training with manually curated intron-exon structures.

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/



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
Table: Standard Code
Trim: No
Reverse: No
Colour: No

Enter or ucleic acid Sequence in any format:

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

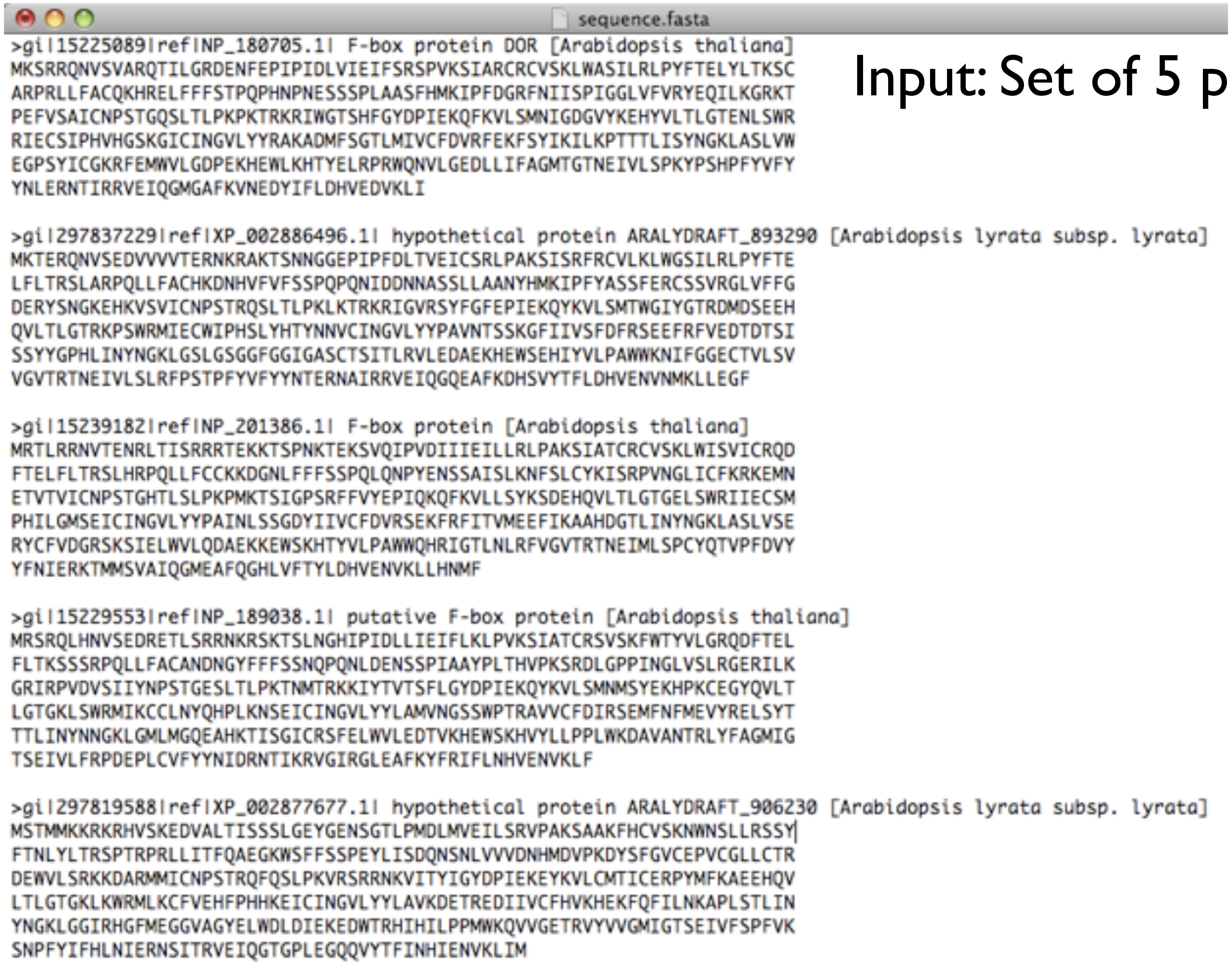
There are programs for multiple sequence alignment (nucleotide or protein) such as ClustalW or Muscle

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/clustalo/
- 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/clustalo/

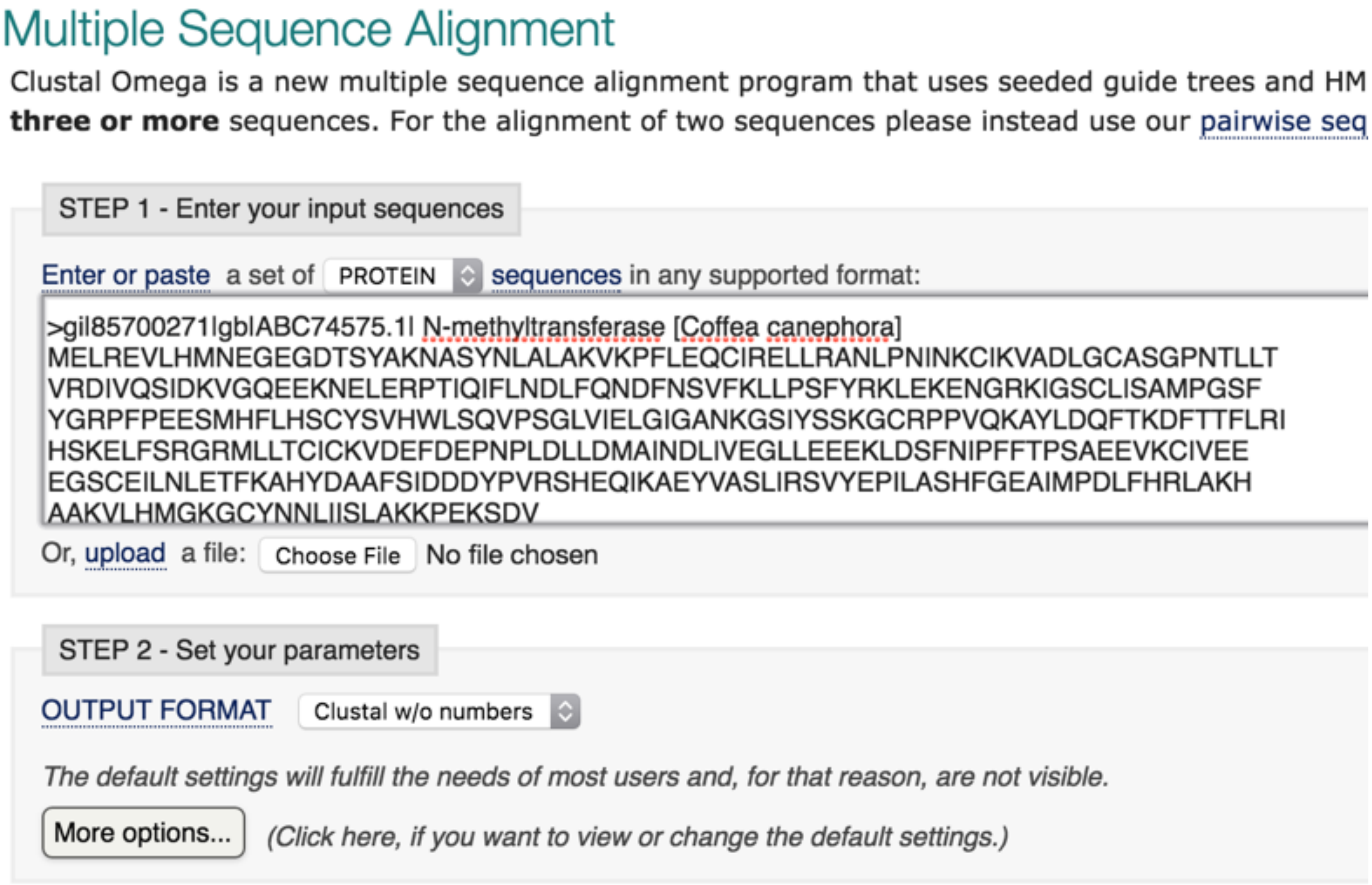


Input: Set of 5 proteins

```
>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
```

2.2 - Multiple Sequence Alignment

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



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HM
three or more sequences. For the alignment of two sequences please instead use our pairwise seq

STEP 1 - Enter your input sequences

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

```
>gi|85700271|gb|ABC74575.1| N-methyltransferase [Coffea canephora]
MELREVLHMNEGEGDTSYAKNASYNLALAKVKPFLEQCIRELLRANLPNINKCIKVADLGCASGPNTLLT
VRDIVQSIDKVGQEEKNELERPTIQIFLNDLFQNDFNSVFKLLPSFYRKLEKENGRKIGSCLISAMPGSF
YGRPFPEESMHFLHSCYSVHWLSQVPSGLVIELGIGANKGSIYSSKGCRPPVQKAYLDQFTKDFTTFLRI
HSKELFSRGRMLLTCICKVDEFDEPNPLDLLDMAINDLIVEGLLEEEKLDSFNIPFFTPSAEEVKCIVEE
EGSCEILNLETFKAHYDAAFSIDDDYPVRSHEQIKAEYVASLIRSVYEPILASHFGEAIMPDLFHRLAKH
AAKVLHMGKGCYNNLIISLAKKPEKSDV
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your parameters

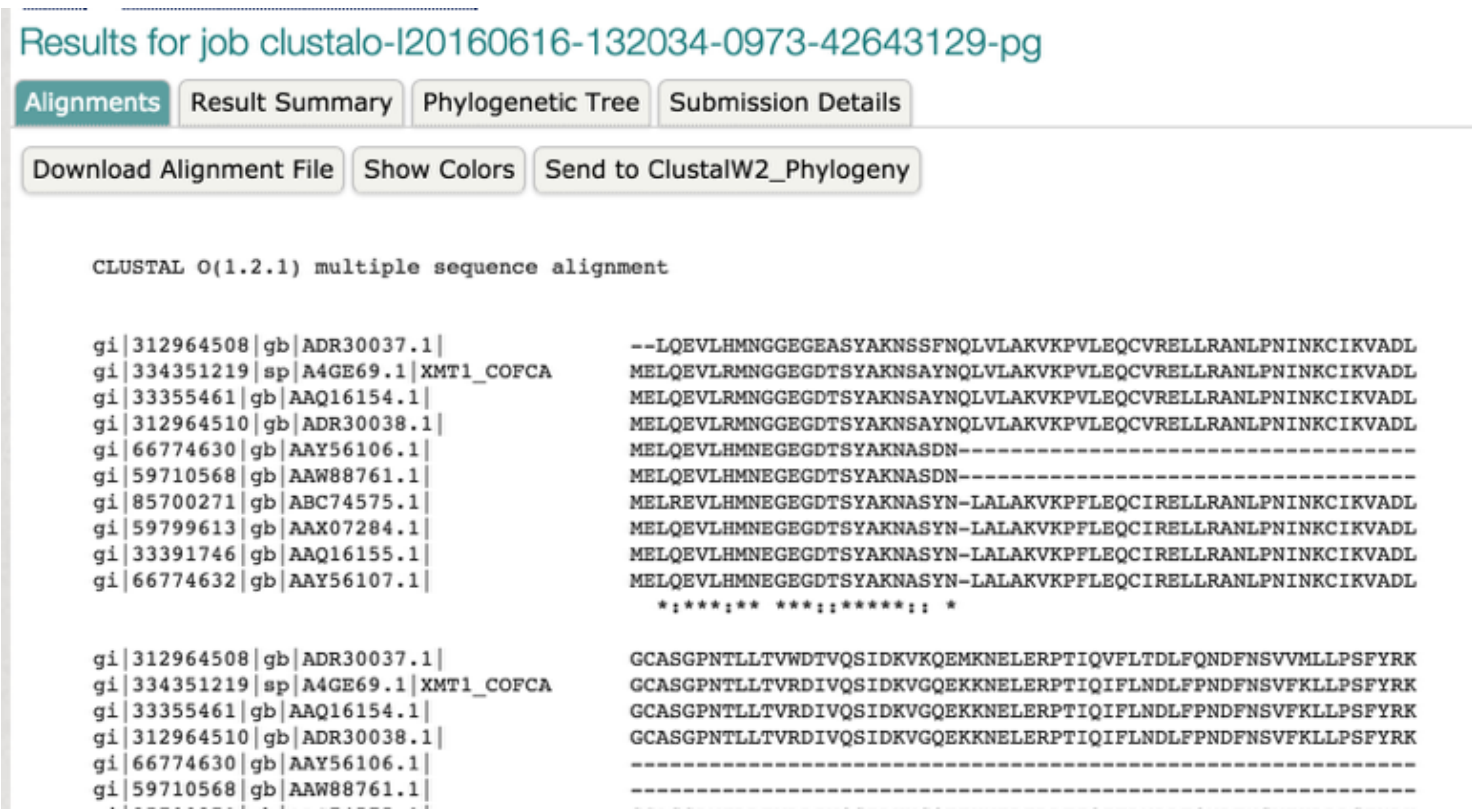
OUTPUT FORMAT Clustal w/o numbers

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

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

2.2 - Multiple Sequence Alignment

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

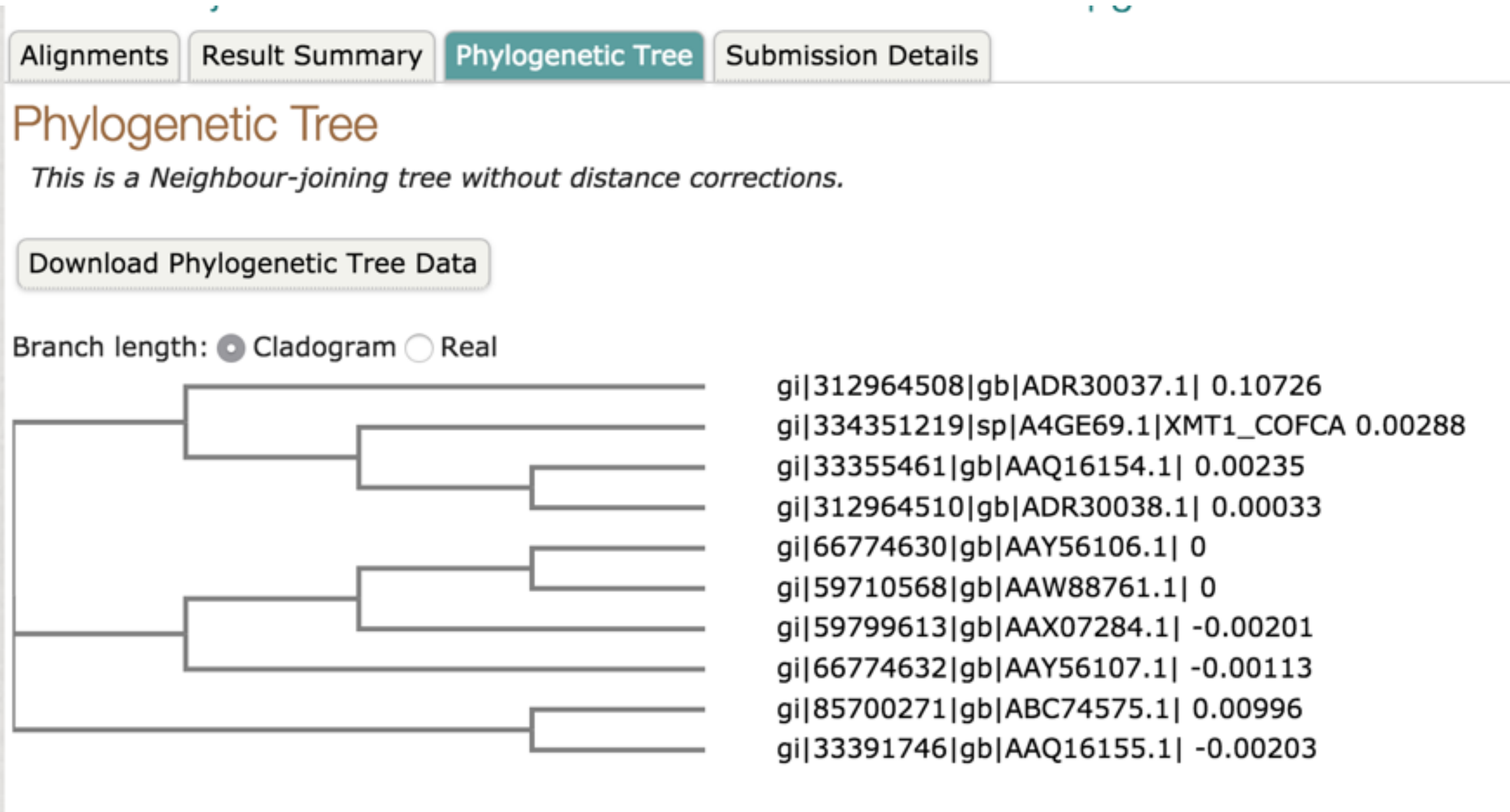


Results for job clustalo-I20160616-132034-0973-42643129-pg

```
CLUSTAL O(1.2.1) multiple sequence alignment


gi|312964508|gb|ADR30037.1|              --LQEVLHMNGGEGEASYAKNSSFNQLVLAKVKPVLEQCVRELLRANLPNINKCIKVADL
gi|334351219|sp|A4GE69.1|XMT1_COFCA      MELQEVLRMNGGEGDTSYAKNSAYNQLVLAKVKPVLEQCVRELLRANLPNINKCIKVADL
gi|33355461|gb|AAQ16154.1|               MELQEVLRMNGGEGDTSYAKNSAYNQLVLAKVKPVLEQCVRELLRANLPNINKCIKVADL
gi|312964510|gb|ADR30038.1|              MELQEVLRMNGGEGDTSYAKNSAYNQLVLAKVKPVLEQCVRELLRANLPNINKCIKVADL
gi|66774630|gb|AAY56106.1|               MELQEVLHMNEGEGDTSYAKNASDN-----------------------------------
gi|59710568|gb|AAW88761.1|               MELQEVLHMNEGEGDTSYAKNASDN-----------------------------------
gi|85700271|gb|ABC74575.1|               MELREVLHMNEGEGDTSYAKNASYN-LALAKVKPFLEQCIRELLRANLPNINKCIKVADL
gi|59799613|gb|AAX07284.1|               MELQEVLHMNEGEGDTSYAKNASYN-LALAKVKPFLEQCIRELLRANLPNINKCIKVADL
gi|33391746|gb|AAQ16155.1|               MELQEVLHMNEGEGDTSYAKNASYN-LALAKVKPFLEQCIRELLRANLPNINKCIKVADL
gi|66774632|gb|AAY56107.1|               MELQEVLHMNEGEGDTSYAKNASYN-LALAKVKPFLEQCIRELLRANLPNINKCIKVADL
                                           *:***:** ***::*****:: *

gi|312964508|gb|ADR30037.1|              GCASGPNTLLTVWDTVQSIDKVKQEMKNELERPTIQVFLTDLFQNDFNSVVMLLPSFYRK
gi|334351219|sp|A4GE69.1|XMT1_COFCA      GCASGPNTLLTVRDIVQSIDKVGQEKKNELERPTIQIFLNDLFPNDFNSVFKLLPSFYRK
gi|33355461|gb|AAQ16154.1|               GCASGPNTLLTVRDIVQSIDKVGQEKKNELERPTIQIFLNDLFPNDFNSVFKLLPSFYRK
gi|312964510|gb|ADR30038.1|              GCASGPNTLLTVRDIVQSIDKVGQEKKNELERPTIQIFLNDLFPNDFNSVFKLLPSFYRK
gi|66774630|gb|AAY56106.1|               ------------------------------------------------------------
gi|59710568|gb|AAW88761.1|               ------------------------------------------------------------
```

2.2 - Multiple Sequence Alignment

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



Phylogenetic Tree

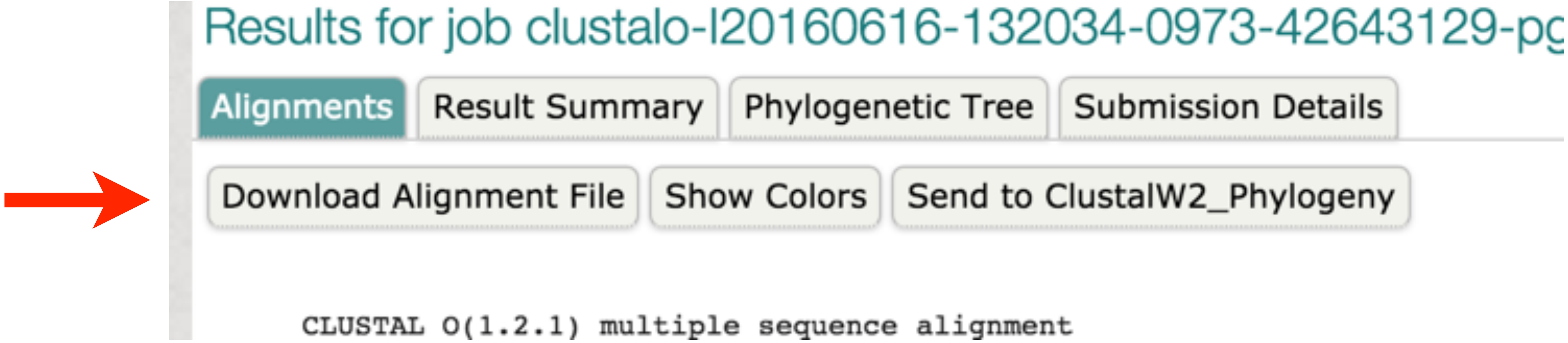
This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

```
gi|312964508|gb|ADR30037.1| 0.10726
gi|334351219|sp|A4GE69.1|XMT1_COFCA 0.00288
gi|33355461|gb|AAQ16154.1| 0.00235
gi|312964510|gb|ADR30038.1| 0.00033
gi|66774630|gb|AAY56106.1| 0
gi|59710568|gb|AAW88761.1| 0
gi|59799613|gb|AAX07284.1| -0.00201
gi|66774632|gb|AAY56107.1| -0.00113
gi|85700271|gb|ABC74575.1| 0.00996
gi|33391746|gb|AAQ16155.1| -0.00203
```

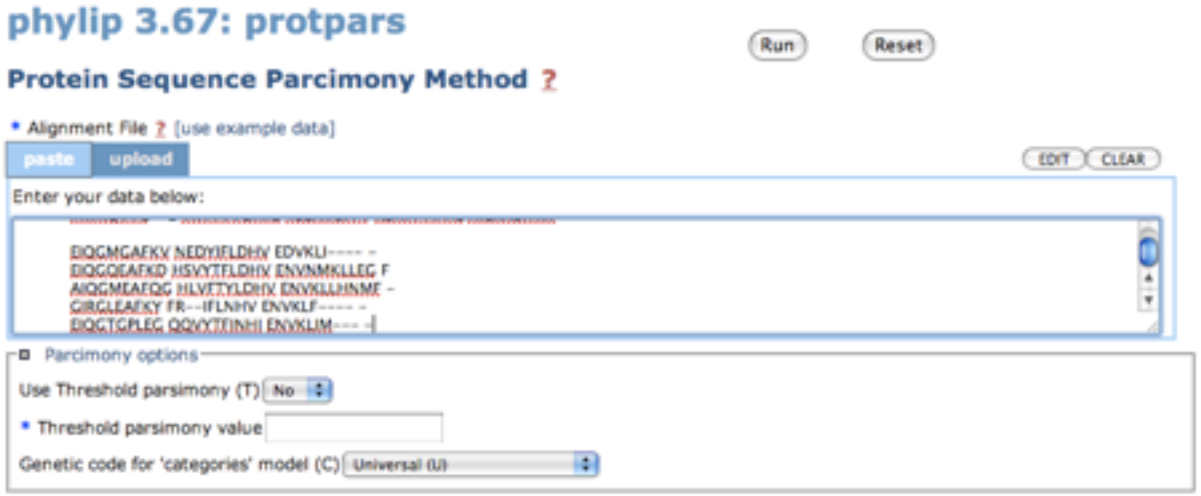
2.2 - Multiple Sequence Alignment

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



```
CLUSTAL O(1.2.1) multiple sequence alignment
```

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

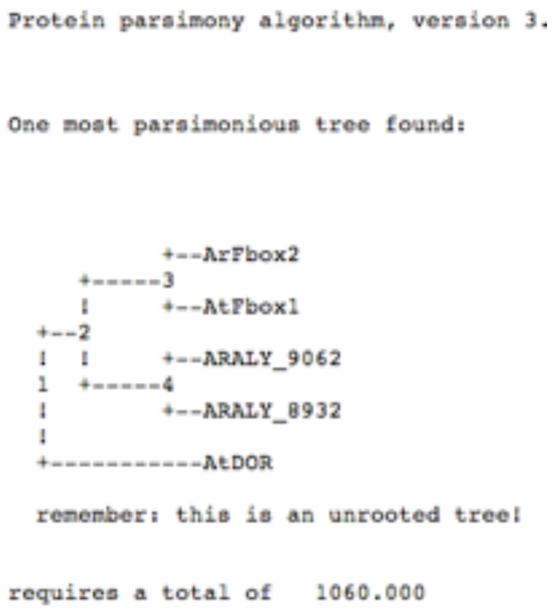


```
DIQGMGAFKV NEDYIFLDHV EDVKLI---- -
EIQGQEAFKD HSVYTFLDHV ENVNMKLLEG F
AIQGMEAFQG HLVFTYLDHV ENVKLLHNME -
GIRGLEAFKY FR--IFLNHV ENVKLF---- -
EIQGTGPLEG QQVYTFINHI ENVKLIM--- -
```

Use Threshold parsimony (T): No

* Threshold parsimony value:

Genetic code for 'categories' model (C): Universal (U)



```
Protein parsimony algorithm, version 3.


One most parsimonious tree found:



             +--ArFbox2
        +-----3
        !    +--AtFbox1
     +--2
     !  !    +--ARALY_9062
     1  +-----4
     !       +--ARALY_8932
     !
     +----------AtDOR

  remember: this is an unrooted tree!


requires a total of   1060.000
```

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.**

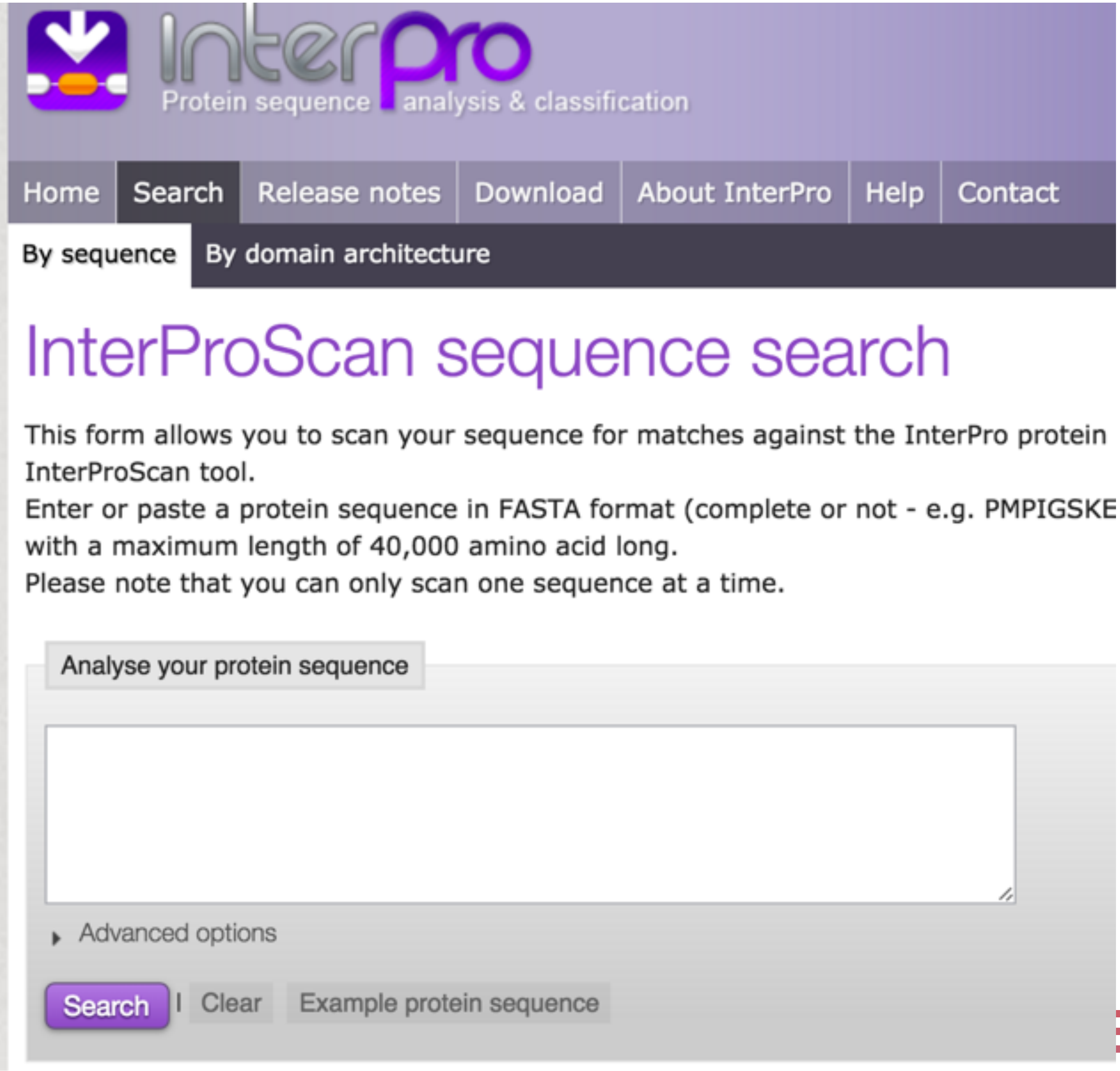
InterPro (http://www.ebi.ac.uk/interpro/) is an EBI resource with several protein domain databases such as *ProSite*, *Pfam* or *Superfamily*.

InterPro 32.0

The tools used for functional domain search is InterProScan (http://www.ebi.ac.uk/interpro/search/sequence-search).

2.3 -Protein Domain Analysis

InterProScan (http://www.ebi.ac.uk/interpro/search/sequence-search).

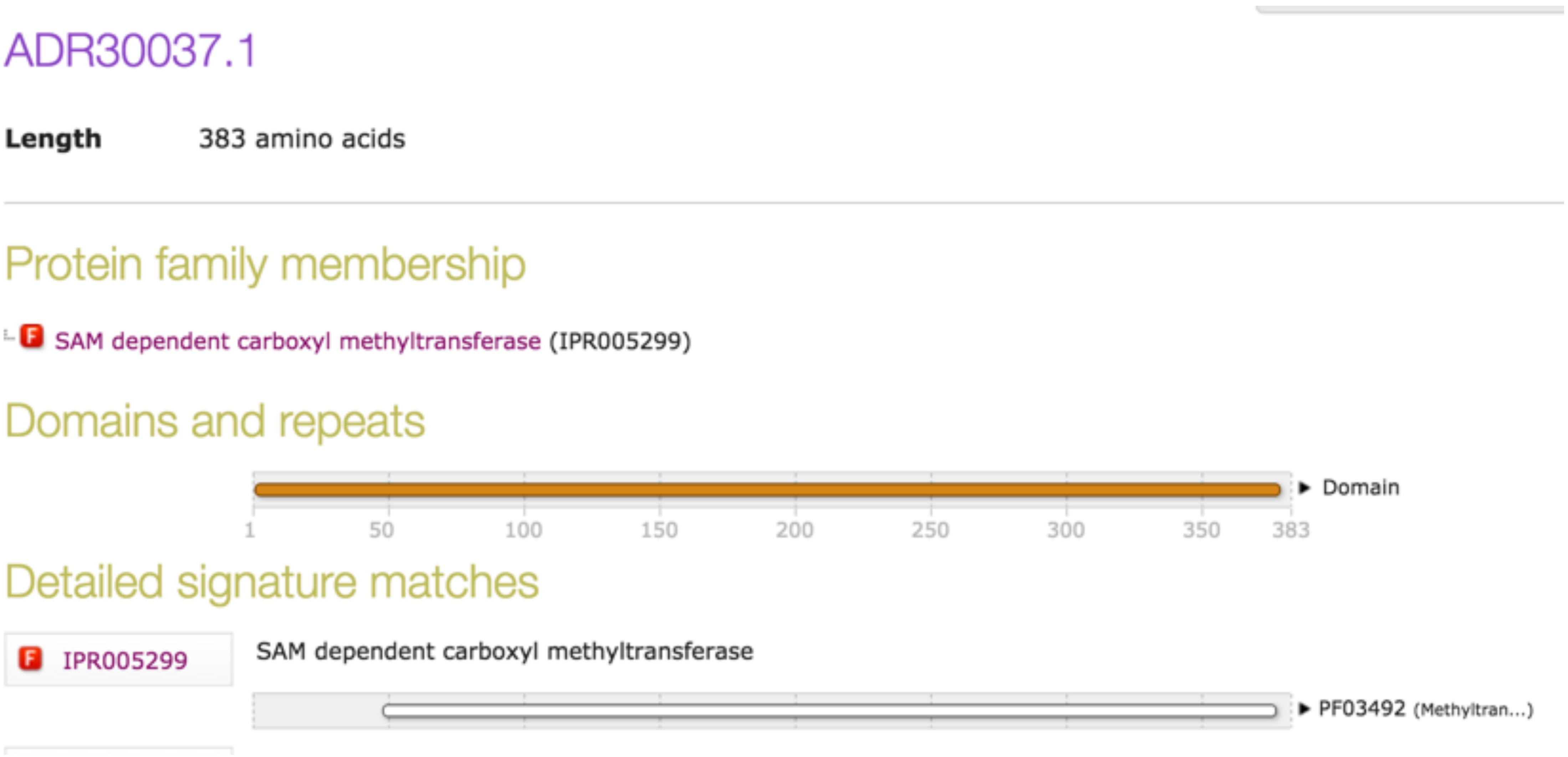


InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein InterProScan tool.
Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKE with a maximum length of 40,000 amino acid long.
Please note that you can only scan one sequence at a time.

2.3 -Protein Domain Analysis

InterProScan (http://www.ebi.ac.uk/interpro/search/sequence-search).



ADR30037.1

Length 383 amino acids

Protein family membership

F SAM dependent carboxyl methyltransferase (IPR005299)

Domains and repeats

Detailed signature matches

F IPR005299 SAM dependent carboxyl methyltransferase

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).

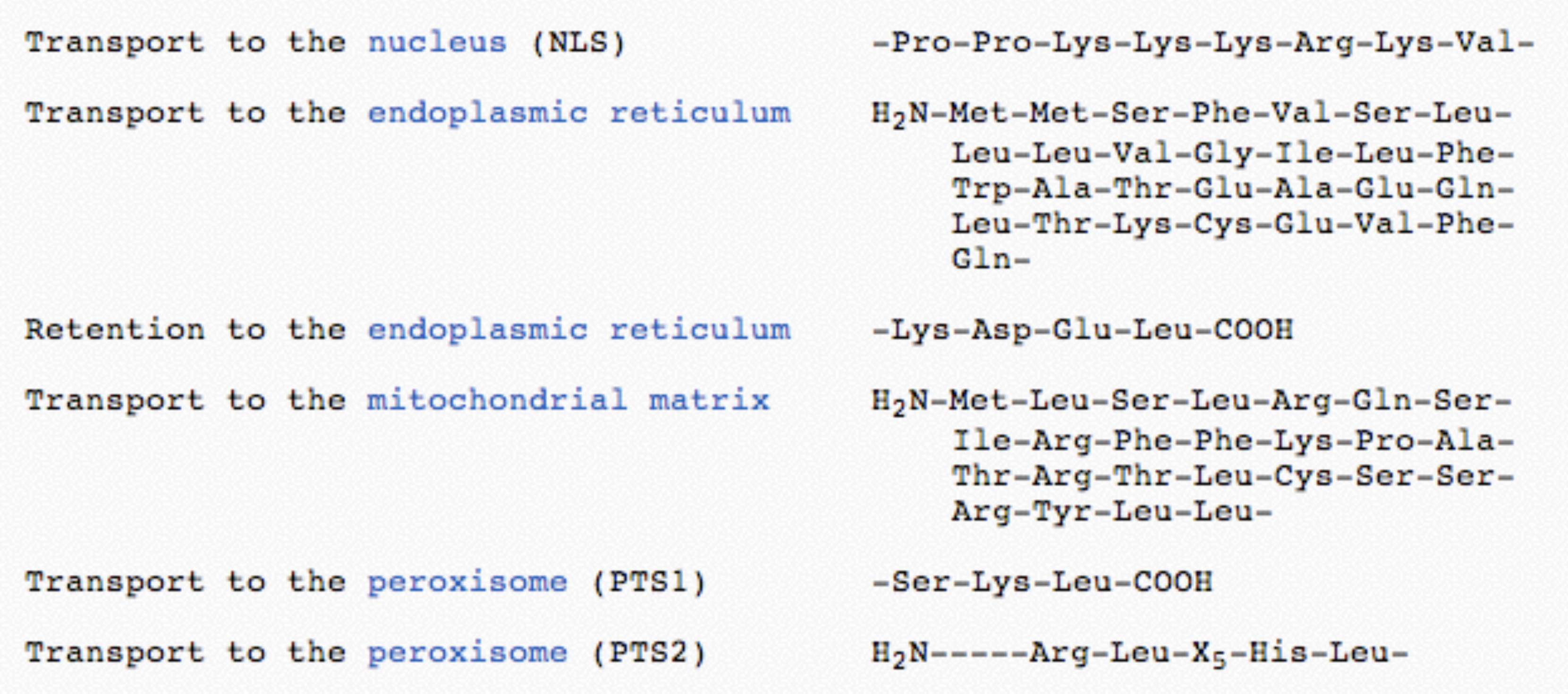
3.2 - Primer Design (Primer3).

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.

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/

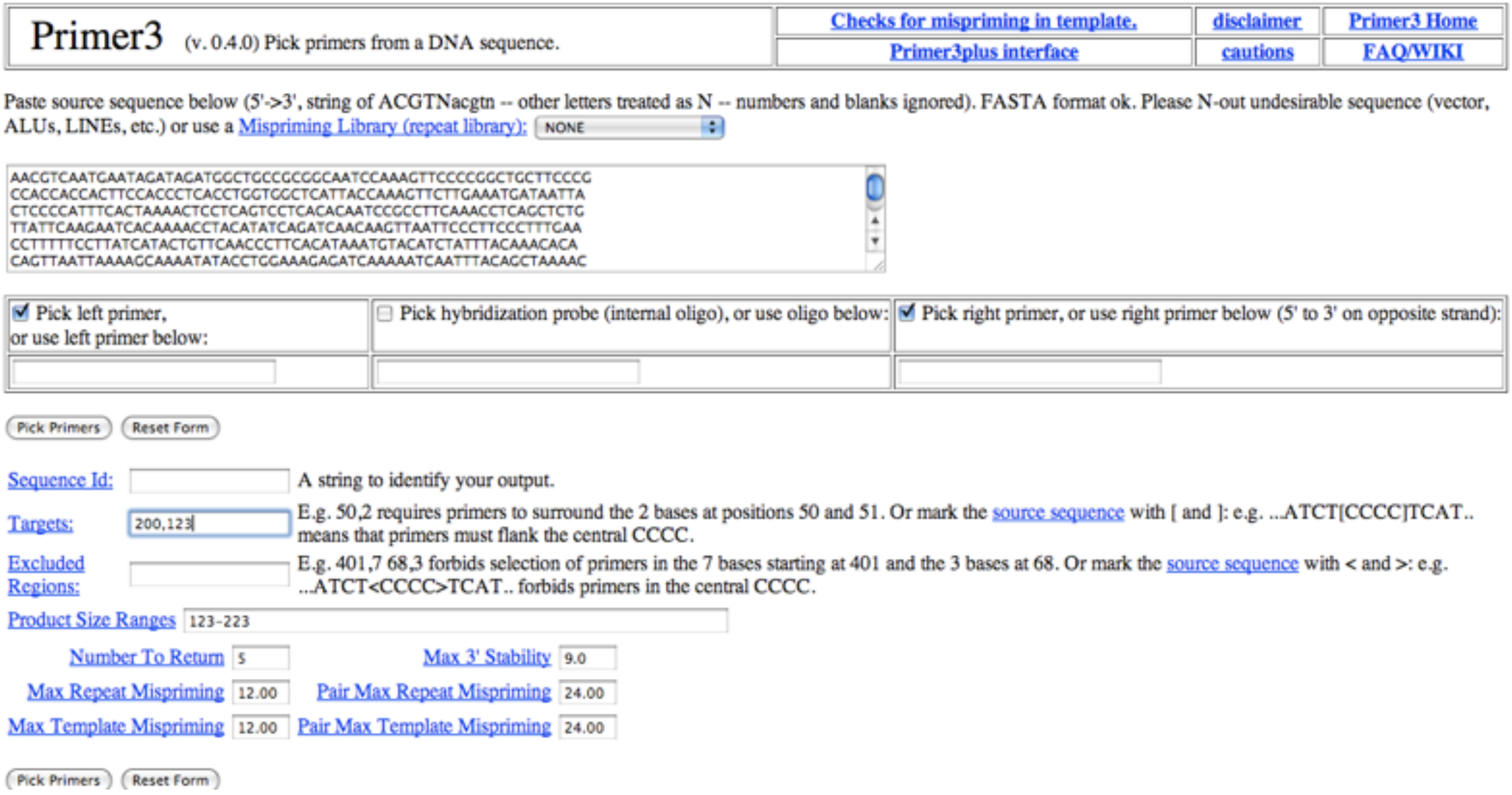
- In-Silico PCR (SGN) (fragment analysis):

  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.

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

```
AACGTCAATGAATAGATAGATGGCTGCCGCGGCAATCCAAAGTTCCCCGGCTGCTTCCCG
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):

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.

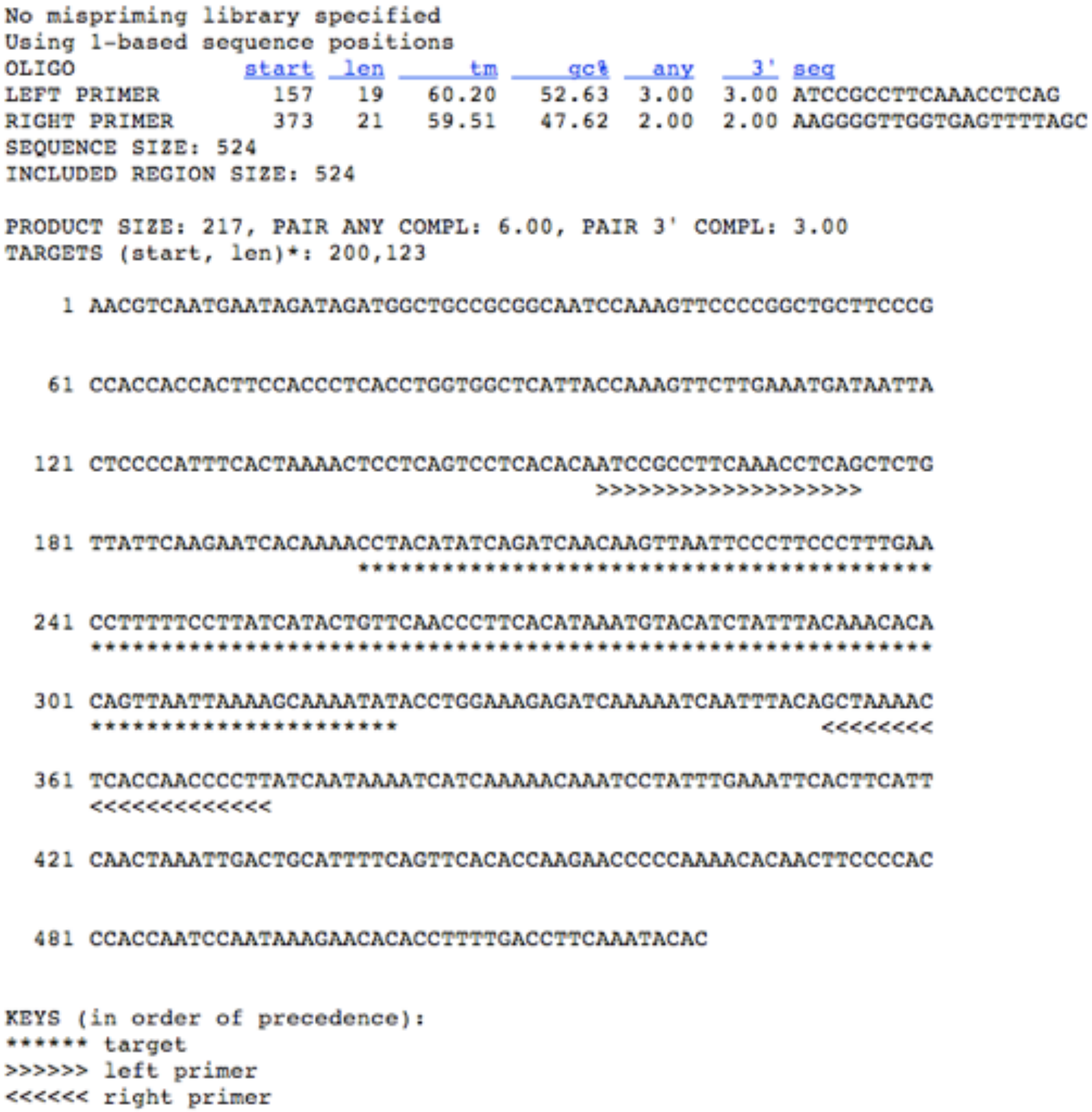
Excluded Regions: E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the source sequence with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

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

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
```

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?

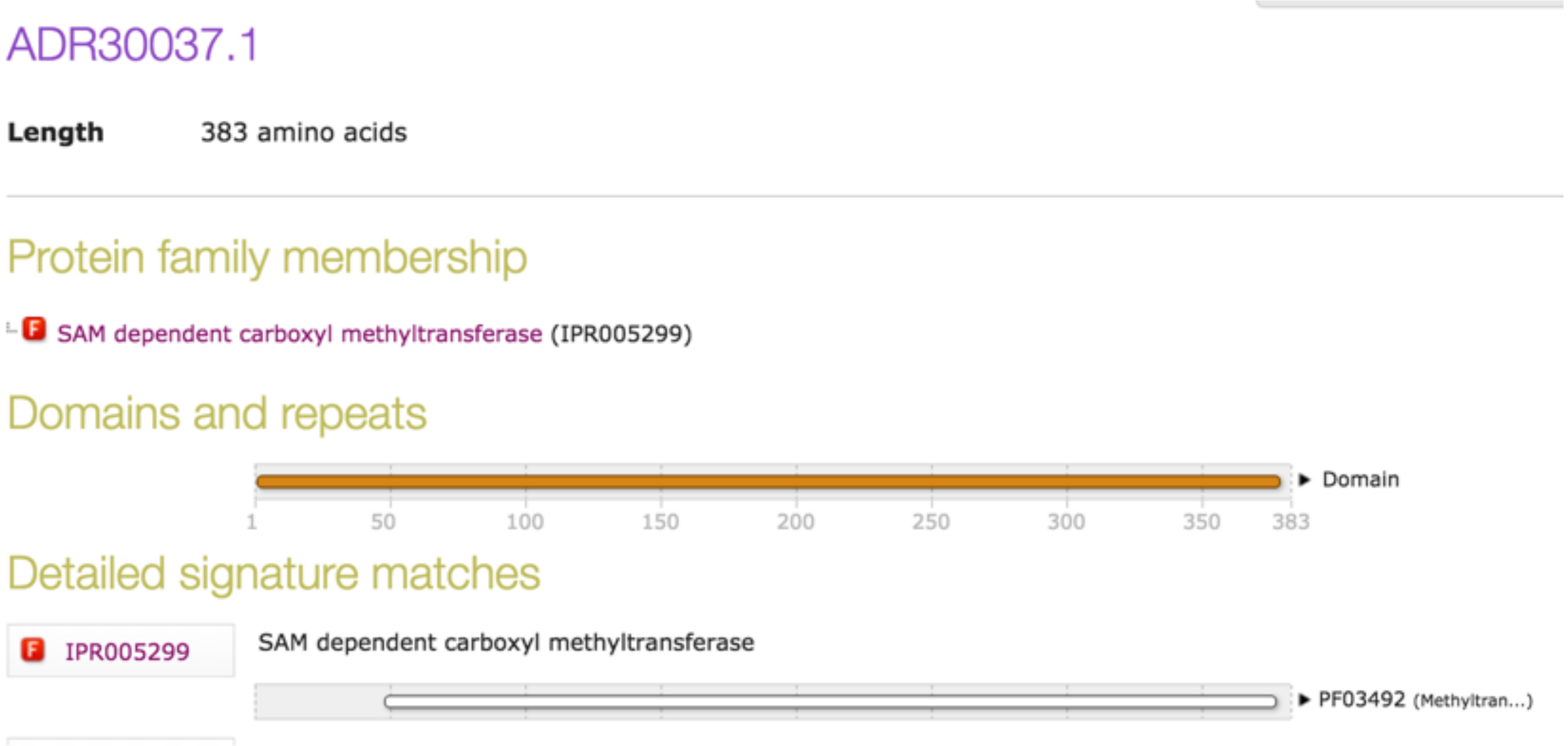
5. Align all sequences from exercise 1.4 with the *Arabidopsis* and tomato protein matches.

6. Make a phylogenetic tree with the alignment from 5. Which sequences appear to be most closely related?

Exercise 2 Solutions (cont'd)

1. Select a protein from exercise 1 part 5, what domains can be found?

http://www.ebi.ac.uk/interpro/search/sequence-search



ADR30037.1

Length 383 amino acids

Protein family membership

F SAM dependent carboxyl methyltransferase (IPR005299)

Domains and repeats

Detailed signature matches

F IPR005299 SAM dependent carboxyl methyltransferase

Exercise 2 Solutions

2. Find the *Arabidopsis thaliana* best protein match to the protein.

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

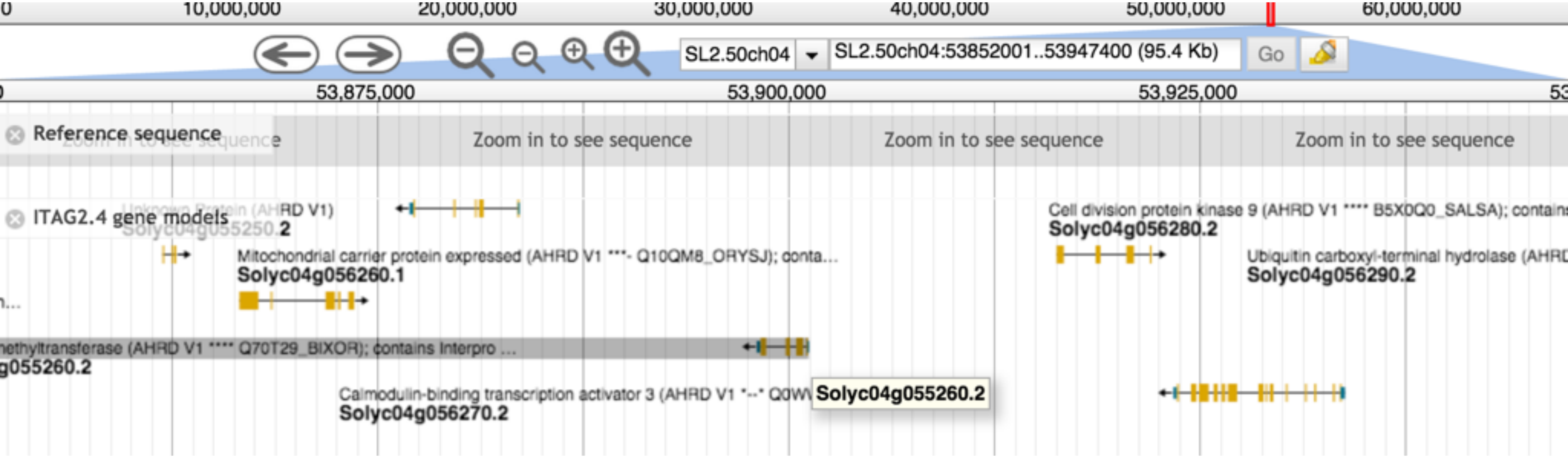
3. Find the tomato best protein match to the protein

Solyc04g055260 (http://solgenomics.net/tools/blast/index.pl)

Exercise 2 Solutions

4. What sequences are upstream and downstream of the tomato match from part 2? How many introns does the match have?

3 introns



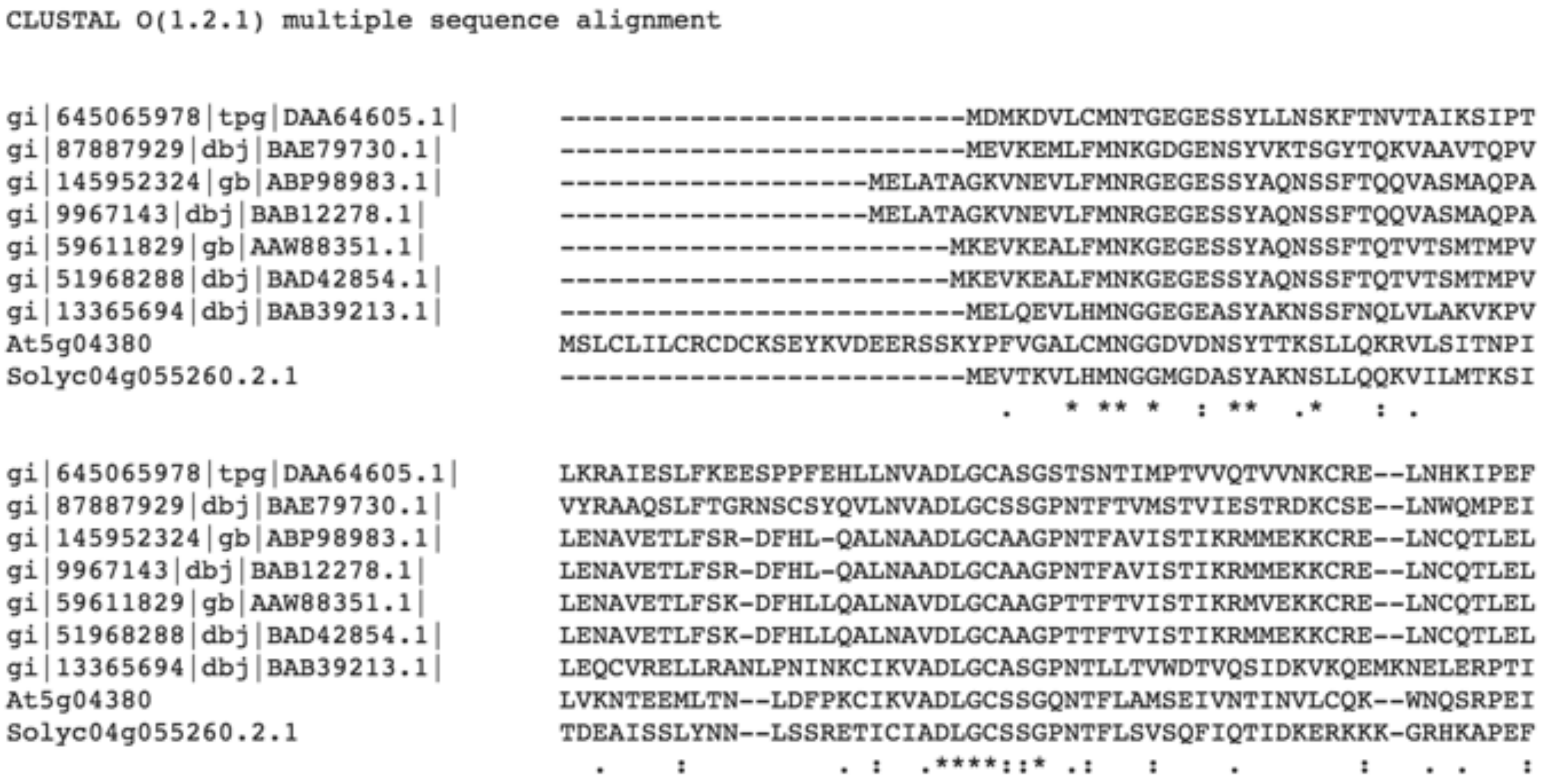
Reference sequence
Zoom in to see sequence
Zoom in to see sequence
Zoom in to see sequence
Zoom in to see sequence

ITAG2.4 gene models
Unknown Protein (AHRD V1)
Solyc04g055250.2
Cell division protein kinase 9 (AHRD V1 **** B5X0Q0_SALSA); contains
Solyc04g056280.2
Mitochondrial carrier protein expressed (AHRD V1 ***- Q10QM8_ORYSJ); conta...
Solyc04g056260.1
Ubiquitin carboxyl-terminal hydrolase (AHRD
Solyc04g056290.2
n...
methyltransferase (AHRD V1 **** Q70T29_BIXOR); contains Interpro ...
g055260.2
Calmodulin-binding transcription activator 3 (AHRD V1 *--* Q0W
Solyc04g055260.2
Solyc04g056270.2

https://solgenomics.net/jbrowse_solgenomics/

Exercise 2 Solutions

5.Align all sequences from exercise 1.4 with the *Arabidopsis* and tomato protein matches.

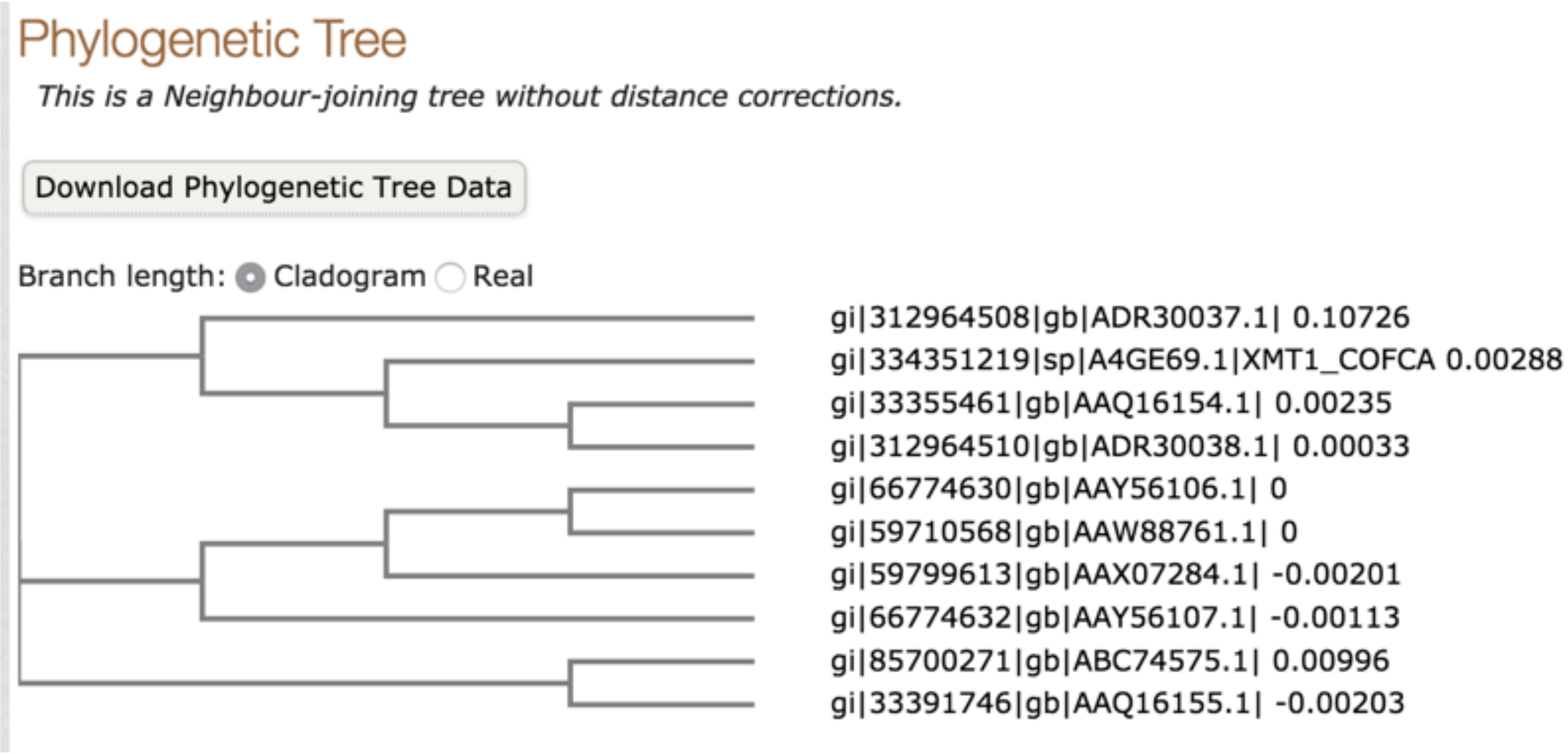


```
CLUSTAL O(1.2.1) multiple sequence alignment


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--LSSRETICIADLGCSSGPNTFLSVSQFIQTIDKERKK-GRHKAPEF
                                      .    :         . :  .****::* .:   :    .       :   . .   :
```

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

Exercise 2 Solutions

6. Make a phylogenetic tree with the alignment from 5. Which sequences appear to be most closely related?



Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

gi|312964508|gb|ADR30037.1| 0.10726
gi|334351219|sp|A4GE69.1|XMT1_COFCA 0.00288
gi|33355461|gb|AAQ16154.1| 0.00235
gi|312964510|gb|ADR30038.1| 0.00033
gi|66774630|gb|AAY56106.1| 0
gi|59710568|gb|AAW88761.1| 0
gi|59799613|gb|AAX07284.1| -0.00201
gi|66774632|gb|AAY56107.1| -0.00113
gi|85700271|gb|ABC74575.1| 0.00996
gi|33391746|gb|AAQ16155.1| -0.00203

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

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. Next class will give hands on command line training,
   - Linux Basics - Bryan Ellerbrock (6/22)
2. Following courses are optional:
   - Intro to commandline tools: Adrian Powell (6/29)
   - Next Generation Sequence Data - Surya Saha (7/6)
   - Intro to R - Nick Morales (7/13)
3. Sign-up for optional courses: email srs57@cornell.edu
4. **You will need to have a virtual machine installed prior to next class. Instructions are here:**

https://btiplantbioinfocourse.wordpress.com/how-to/installing-the-virtual-machine/

PLEASE STOP BY BIOINFORMATICS HOUR WEDNESDAY 1 - 2 PM IN THE RESOURCE CENTER TO SHOW US YOUR WORKING VIRTUAL MACHINE.

*** You will not be able to participate in the next class exercises without it.***