Introduction to Bioinformatics

By Ursula Frei, Walter Suza, Thomas Lübberstedt, Madan Bhattacharyya (ISU)

Introduction

A biological sequence database is a collection of molecular data organized in a manner that allows easy access, management, and update of the data. Biological sequence databases serve an important role of providing access to sequence information to the research community. The databases contain molecular information of multiple organisms and are constantly being updated and re-designed to allow more robust data query and analysis. Examples of biological databases include European Molecular Biology Laboratory (EMBL), GenBank, the National Center for Biotechnology Information (NCBI), and the DNA Databank of Japan (DDBJ). Every sequence submitted to the database has a unique number assigned to it, called the Accession number. Even if the same gene has been submitted several times by different investigators each will have a different accession number.

Database Types

Databases can be classified in to primary (archival), secondary (curated), and composite databases.

- A **primary database** (e.g. EMBL/DDBJ/GenBank for nucleic acids) contains information of the sequence or structure alone, for example, DNA, RNA, or protein sequences.
- A **secondary database** (e.g. eMOTIF at Stanford University, PROSITE of Swiss Institute of Bioinformatics) contains information derived from the primary databases and represent sequences that are consensus of a population, for example, conserved features and motifs of a sequence.
- A **composite database** contains a variety of different primary databases and provides multiple options for database search (e.g. NCBI, MaizeGDB). New tools are continuously developed to make both submission and access to sequence databases more efficient.

The lesson includes practical examples of using database tools. It is recommended that you use “try this” questions to become familiar with sequence databases.

Objectives

- To introduce some of the most commonly used databases in molecular plant breeding
- To help the student learn the tools for accessing and manipulating biological databases
- To help develop proficiency in the use of biological databases

Access and Use of Sequence Databases

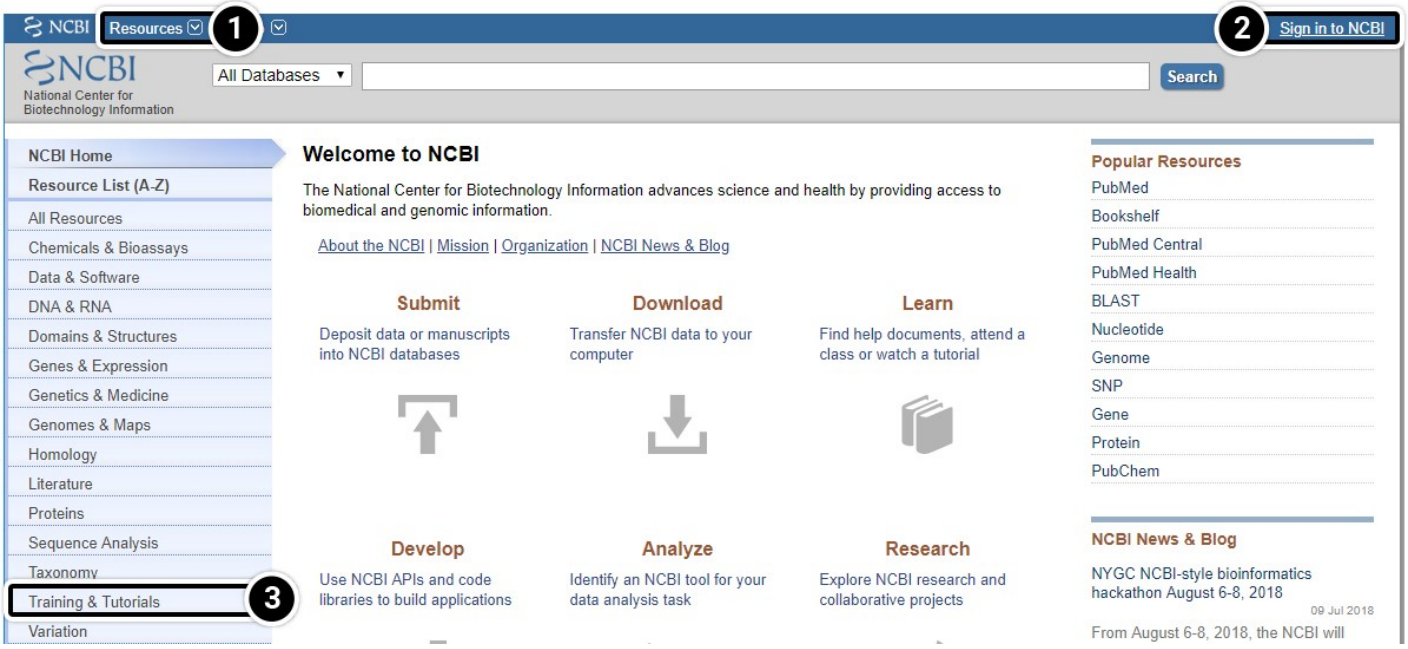
Once a new sequence has been determined a common step in its analysis is to compare the sequence with related genes that have already been sequenced, often from other organisms. A few things to keep in mind about database searches and sequence databases in general:

1. Do not assume that if a sequence is in the database it must be correct. Databases are full of errors!
2. Similarity with a known protein or gene does not necessarily mean the query is the same gene as the one it has similarity with.
3. Two nucleotide sequences may have low similarity yet code for proteins that are functionally related.
4. Protein sequences may also have low similarity yet still be functionally or structurally related.

About NCBI

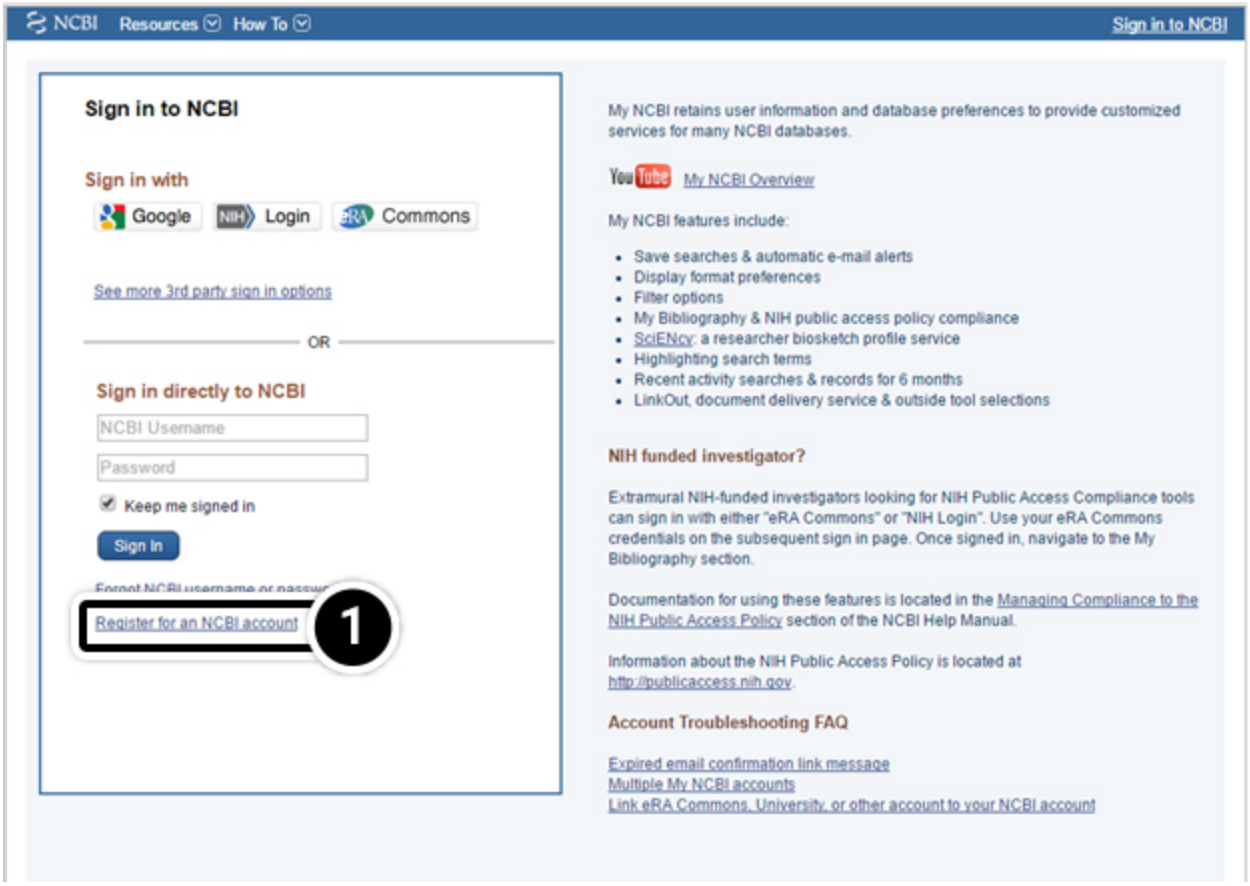
NCBI was created in 1988 as a division of the US National Library of Medicine at the National Institute of Health. The role of NCBI is to create automated systems for storing and analyzing sequence information.

1. To access various resources available through NCBI select **Resources**.
2. We recommend that you set up an account with NCBI to allow you the option of saving your results. Click the **Sign in** link to do so
3. Video tutorials are available under the **Training & Tutorials** link to enhance learning.



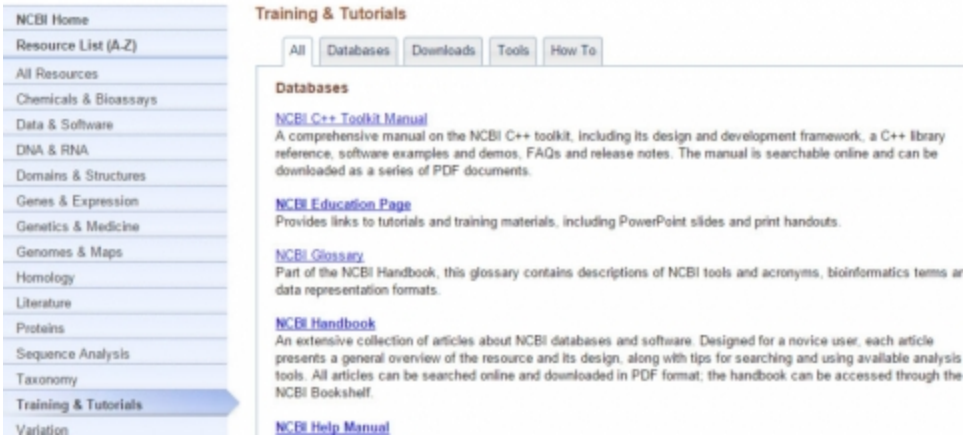
Sign Up for NCBI

1. Click **Register** to set up a new NCBI account.



NCBI Training

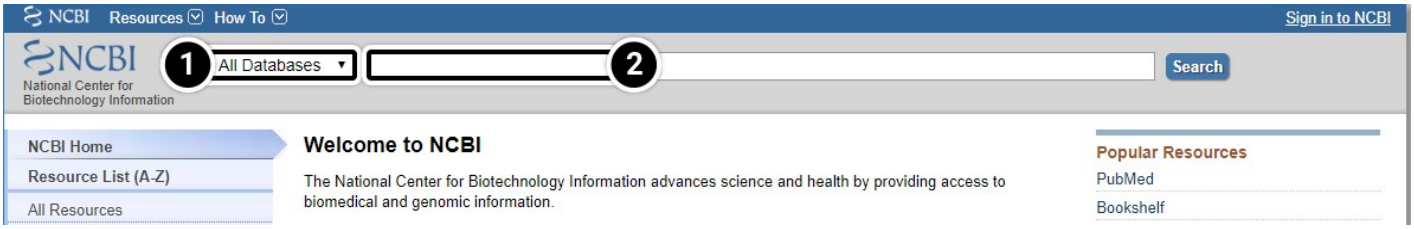
NCBI was created in 1988 as a division of the US National Library of Medicine at the National Institute of Health. The role of NCBI is to create automated system for storing and analyzing sequence information.



Information Retrieval from NCBI

One of the most widely used interfaces for the retrieval of sequence information from biological databases is the **NCBI Entrez system**. Entrez relies on preexisting, logical relationships between the individual sequences (data points) available in various public databases.

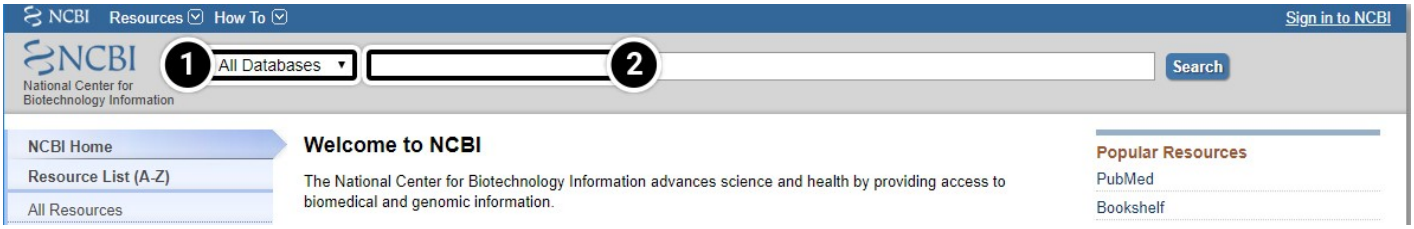
1. Searching all databases is often a good starting point to get an overview of the state of your research field.
2. Searches are based on keywords.



Searching NCBI by Keywords

Searches can be restricted to a single database or expanded to include all other databases. The simplest way to query is through the use of individual search terms, coupled by Boolean operators such as AND, OR, or NOT. A Boolean operator is a variable that can have only a true or false value.

1. Select individual databases, or search them all.
2. **AND**: To 'AND' two search terms together instructs Entrez to find all documents that contain BOTH terms
   OR: To 'OR' two search terms together instructs Entrez to find all documents that contain EITHER term.
   NOT: To 'NOT' two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2.



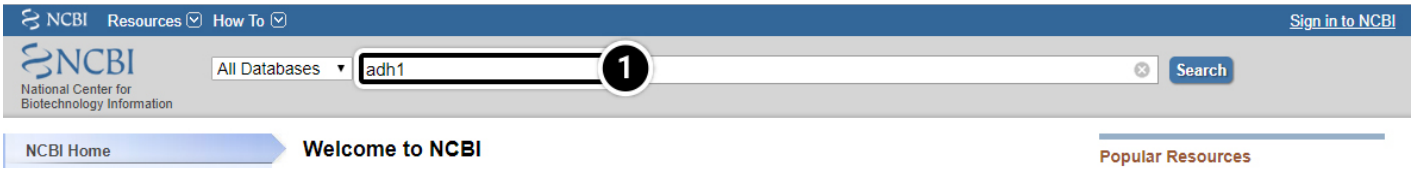
Try This: Access and Use of Sequence Databases

This activity consists of the following pages:

Try This: Access and Use of Sequence DBs (Compare the sequences) - 1

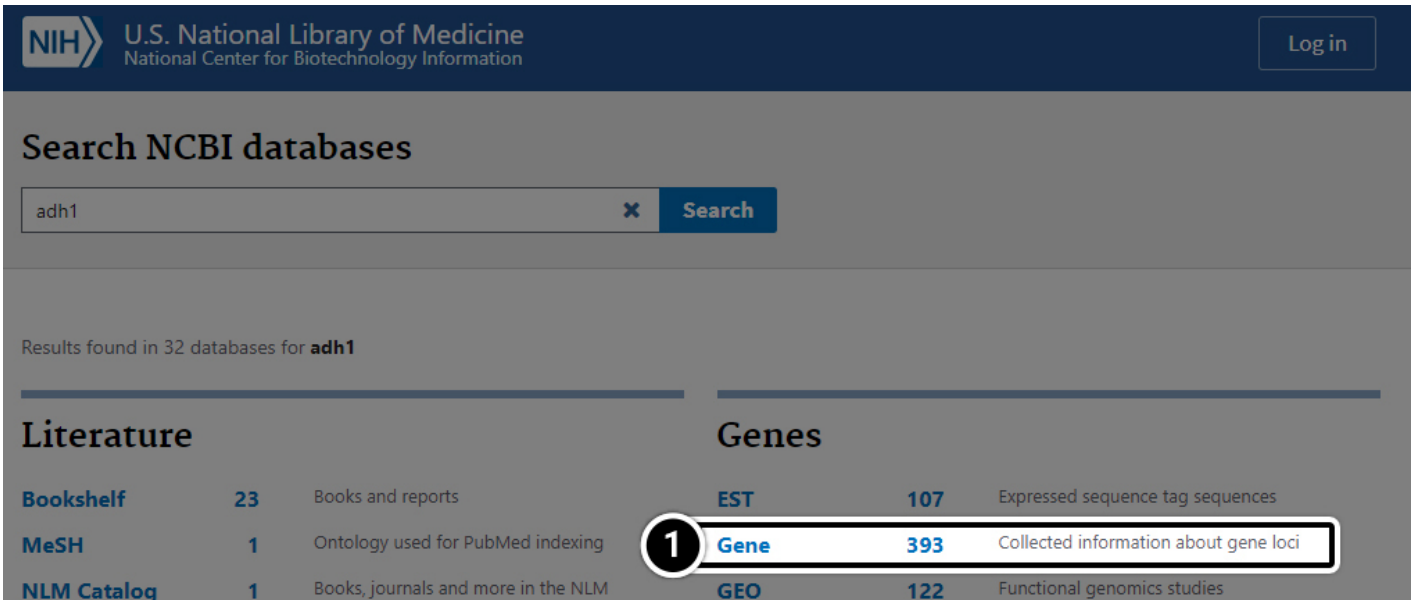
Compare the sequences for the adh1 gene in maize and sorghum. Navigate to the **NCBI site**.

1. Enter adh1 in the "search across databases" window. How many adh1 candidates did your search find?



Try This: Access and Use of Sequence DBs (Results of a search) - 2

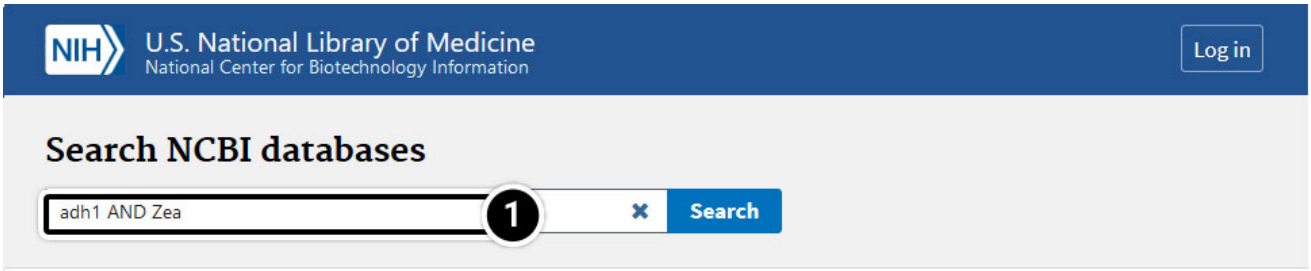
Results of a search for “adh1” across all databases:



Try This: Access and Use of Sequence DBs (Compare the results) - 3

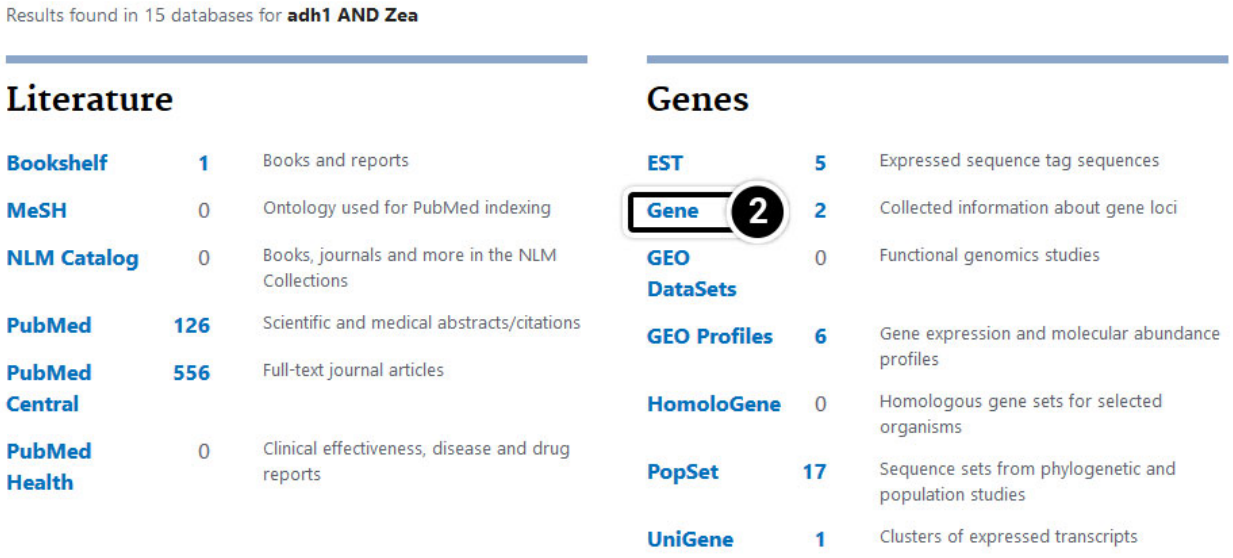
Compare the sequences for the adh1 gene in maize and sorghum:

1. Enter **adh1 AND Zea** in the search window.



Compare the results in the **Gene** category.

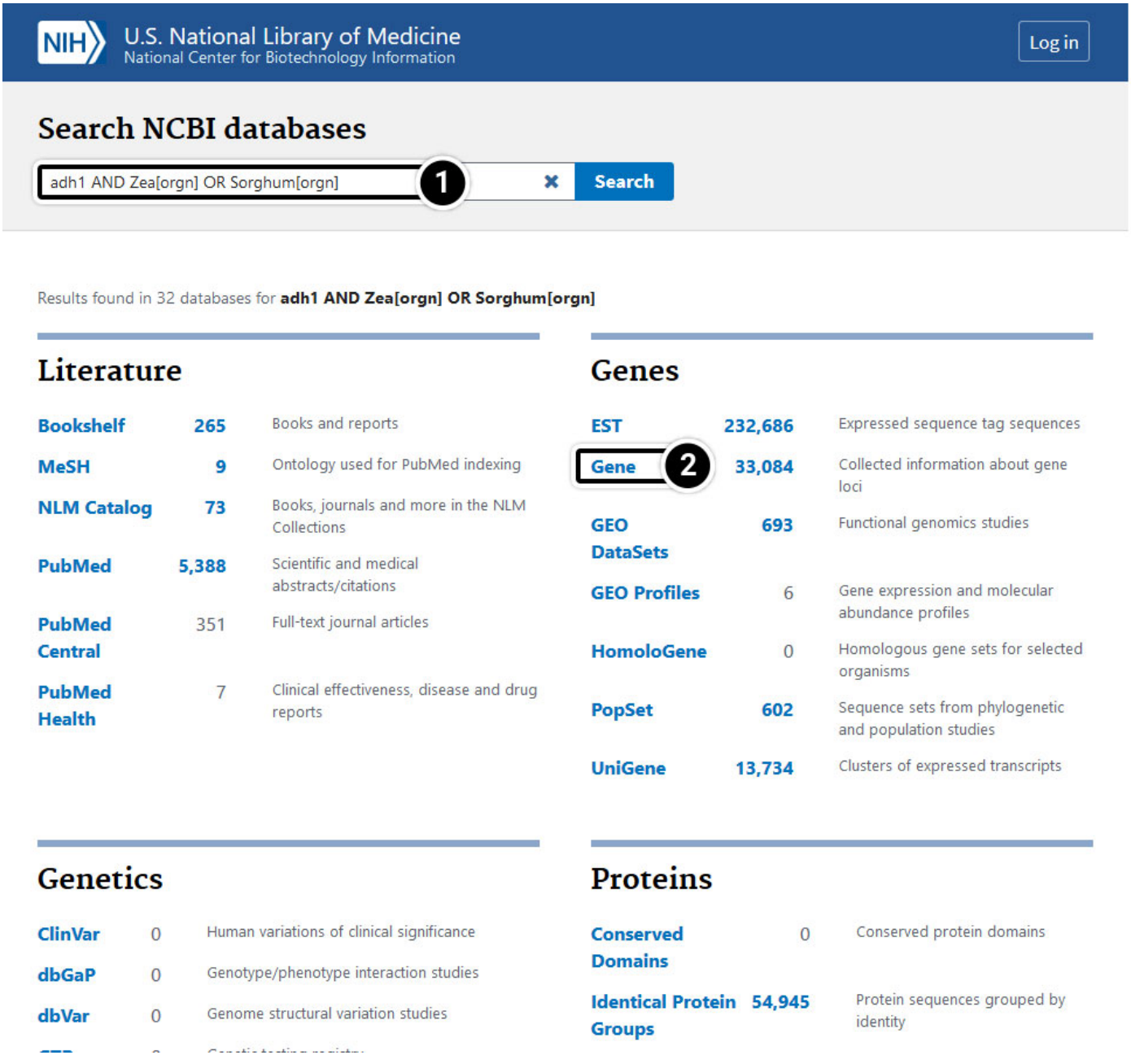
2. Boolean operators can be used to restrict a search and allow users to obtain specific information about their organism of interest.



Try This: Access and Use of Sequence DBs (Operators) - 4

Now try these operators.

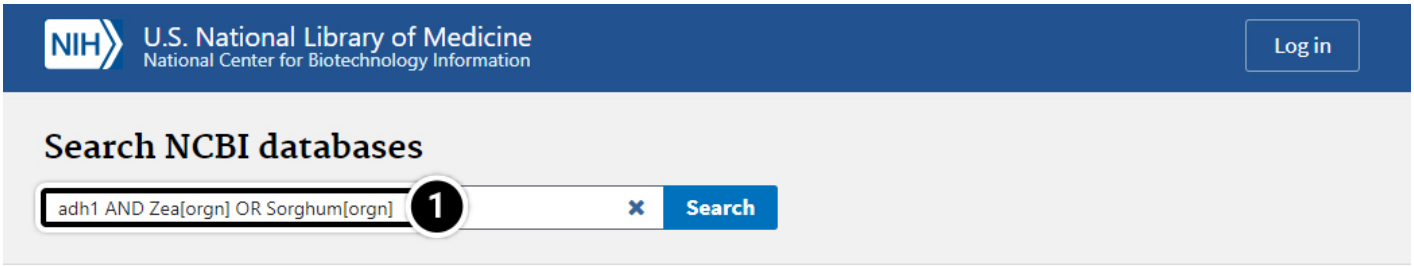
1. Enter **adh1 AND Zea[orgn] OR Sorghum[orgn]** in the search window.
2. Results



Try This: Access and Use of Sequence DBs (Operators) - 5

Now try these operators.

1. Enter **adh1 AND Zea[orgn] OR Sorghum[orgn]** in the search window.



What stands out when you compare results when using the search terms "*adh1 AND Zea[orgn] OR Sorghum[orgn]*" and "*adh1 AND (Zea[orgn] OR Sorghum[orgn])*? Can you identify any differences among the results obtained from the following sets of search terms?

- *"adh1 AND Zea[orgn]"* and *"adh1 AND Zea[orgn] OR Sorghum[orgn]"*
- *"adh1 AND Zea[orgn]"* and *"adh1 AND (Zea[orgn] OR Sorghum[orgn])"*

1. Enter **adh1 AND (Zea[orgn] OR Sorghum[orgn])** in the search window.
2. Compare the results in the **Gene** category.

U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI databases

adh1 AND (Zea[orgn] OR Sorghum[orgn]) 1

Search

Results found in 15 databases for **adh1 AND (Zea[orgn] OR Sorghum[orgn])**

Literature

Bookshelf	1	Books and reports
MeSH	0	Ontology used for PubMed indexing
NLM Catalog	0	Books, journals and more in the NLM Collections
PubMed	131	Scientific and medical abstracts/citations
PubMed Central	351	Full-text journal articles
PubMed Health	0	Clinical effectiveness, disease and drug reports

Genes

EST	5	Expressed sequence tag sequences
Gene 2	2	Collected information about gene loci
GEO DataSets	0	Functional genomics studies
GEO Profiles	6	Gene expression and molecular abundance profiles
HomoloGene	0	Homologous gene sets for selected organisms
PopSet	19	Sequence sets from phylogenetic and population studies
UniGene	1	Clusters of expressed transcripts

Genetics

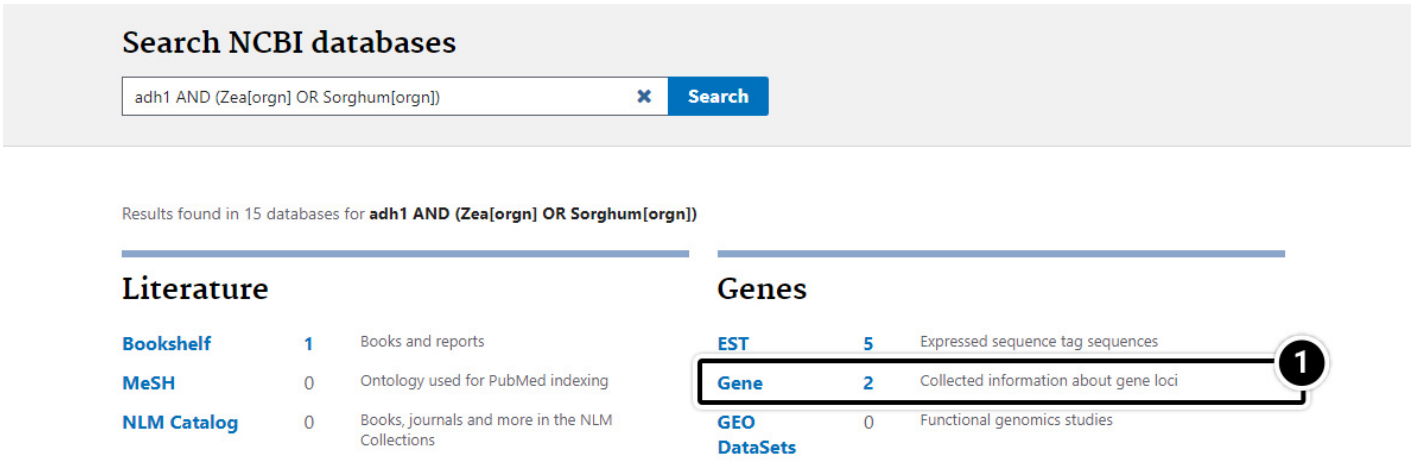
ClinVar	0	Human variations of clinical significance
dbGaP	0	Genotype/phenotype interaction studies

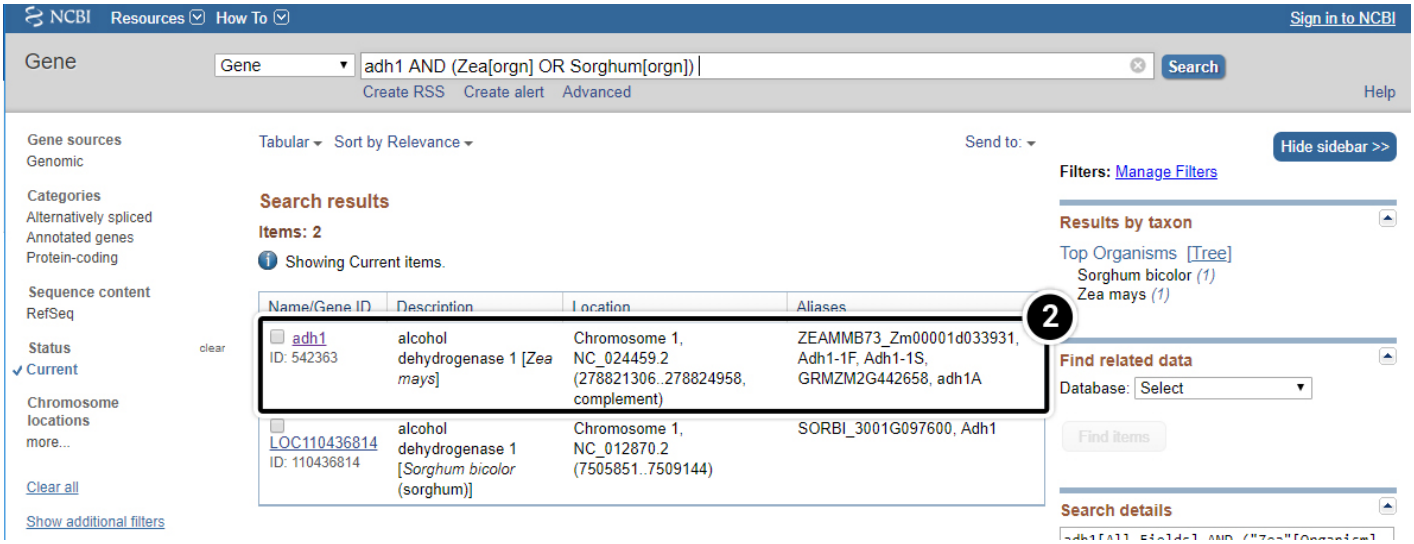
Proteins

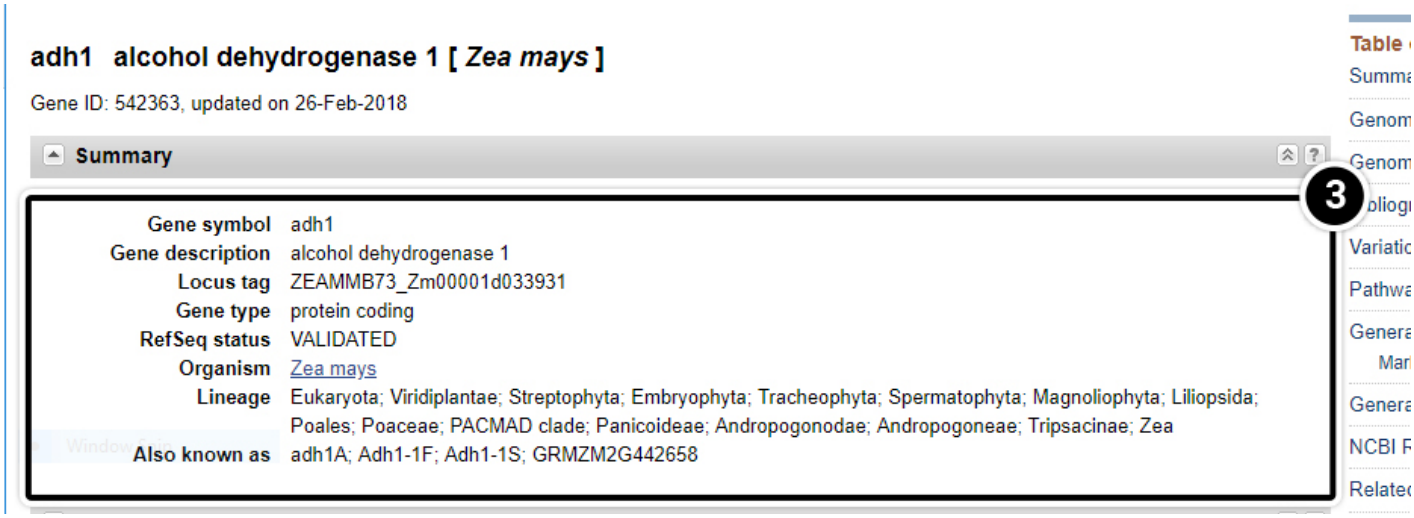
Conserved Domains	0	Conserved protein domains

Try This: Access and Use of Sequence DBs (Gene-centered Info) - 6

1. Click on "**Gene**" to get gene-centered information on the output in the last screen results (also shown here).
2. Click on the first "**adh1".**
3. Review the output window.



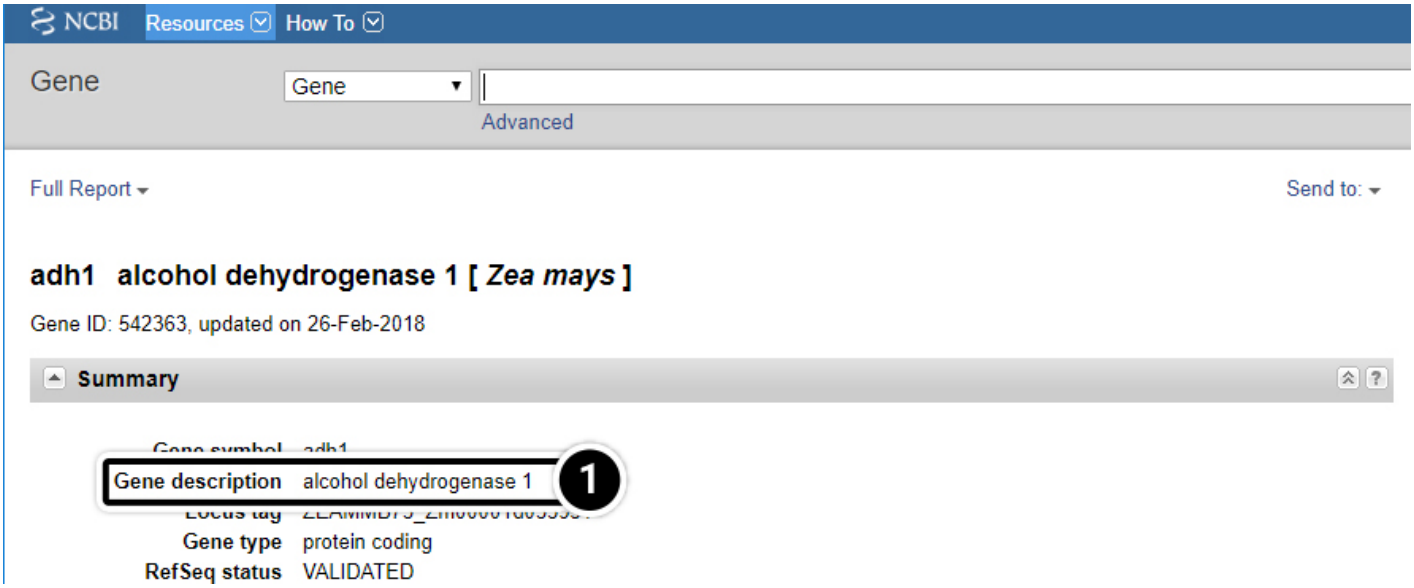




Try This: Access and Use of Sequence DBs (adh1) - 7

What is the function of adh1?

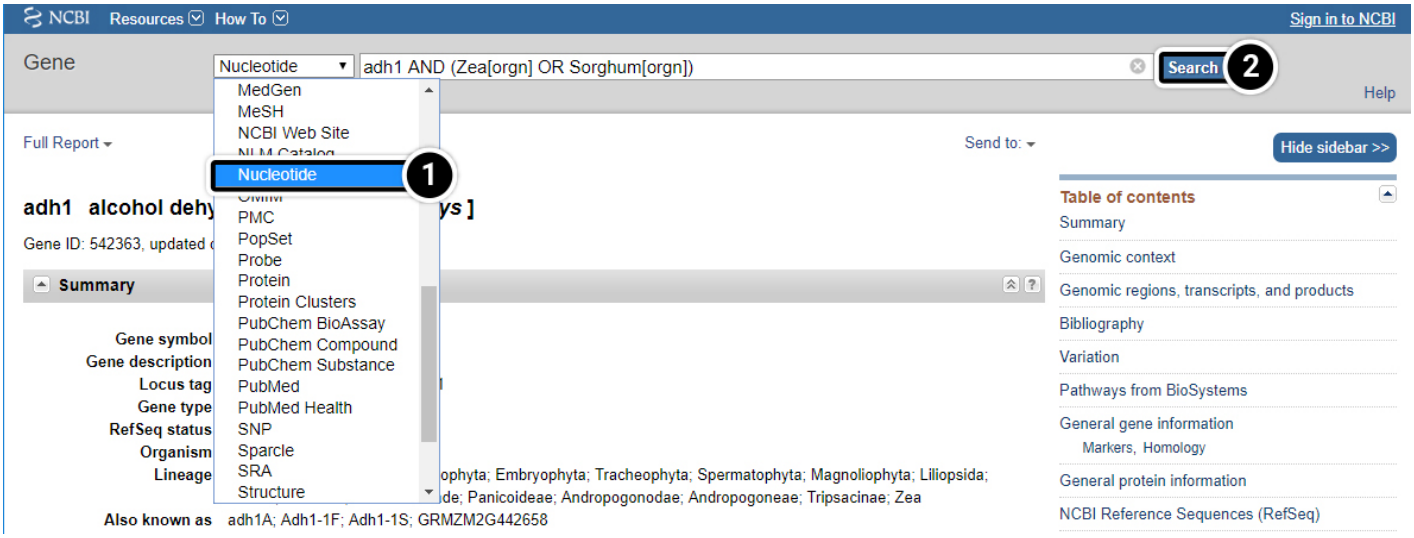
1. Answer found below.



Try This: Access and Use of Sequence DBs (Nucleotide results) - 8

Let's examine the Nucleotide results.

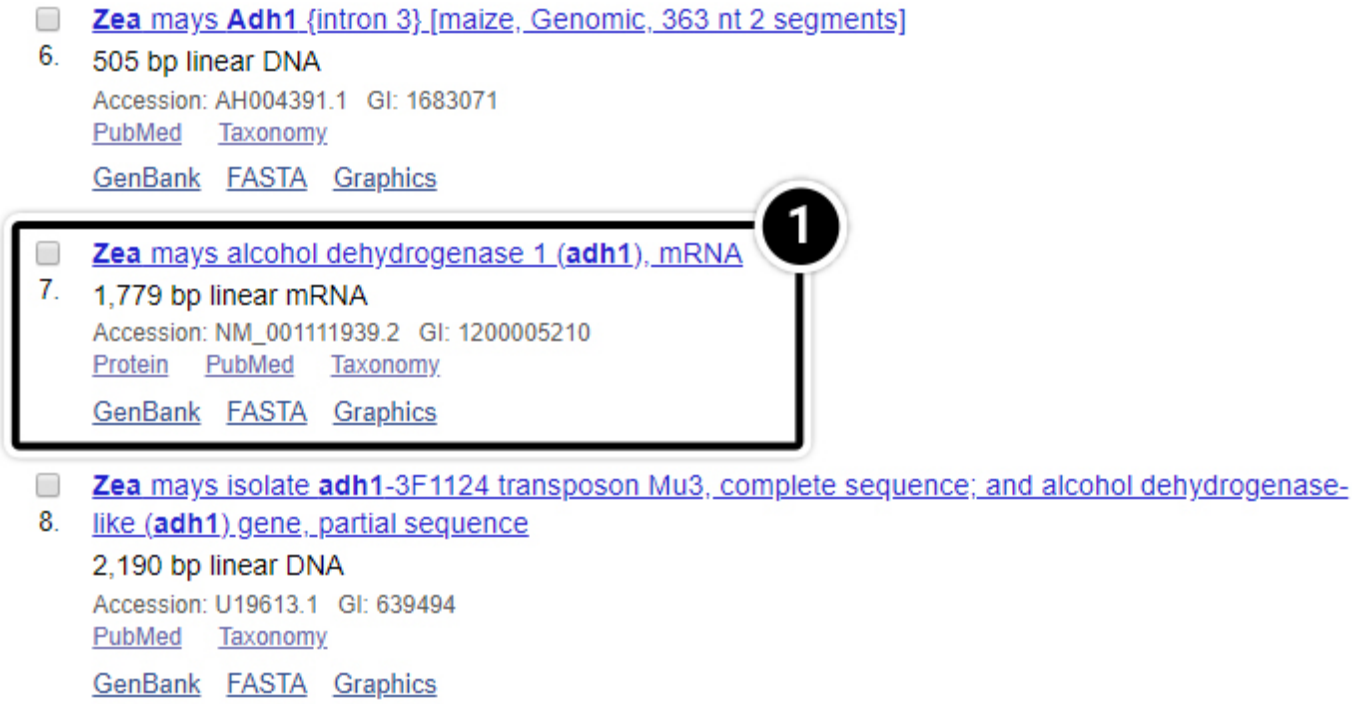
1. Click this pull-down menu for more information about this gene and select **Nucleotide**.
2. Click **Search**. Your search will result in 100s of hits.



Try This: Access and Use of Sequence DBs (Nucleotide results) - 9

Let's examine the Nucleotide results.

1. After selecting the **nucleotide** option as in previous screen, click on the **adh1 mRNA** as indicated. You may have to scroll down to find it.

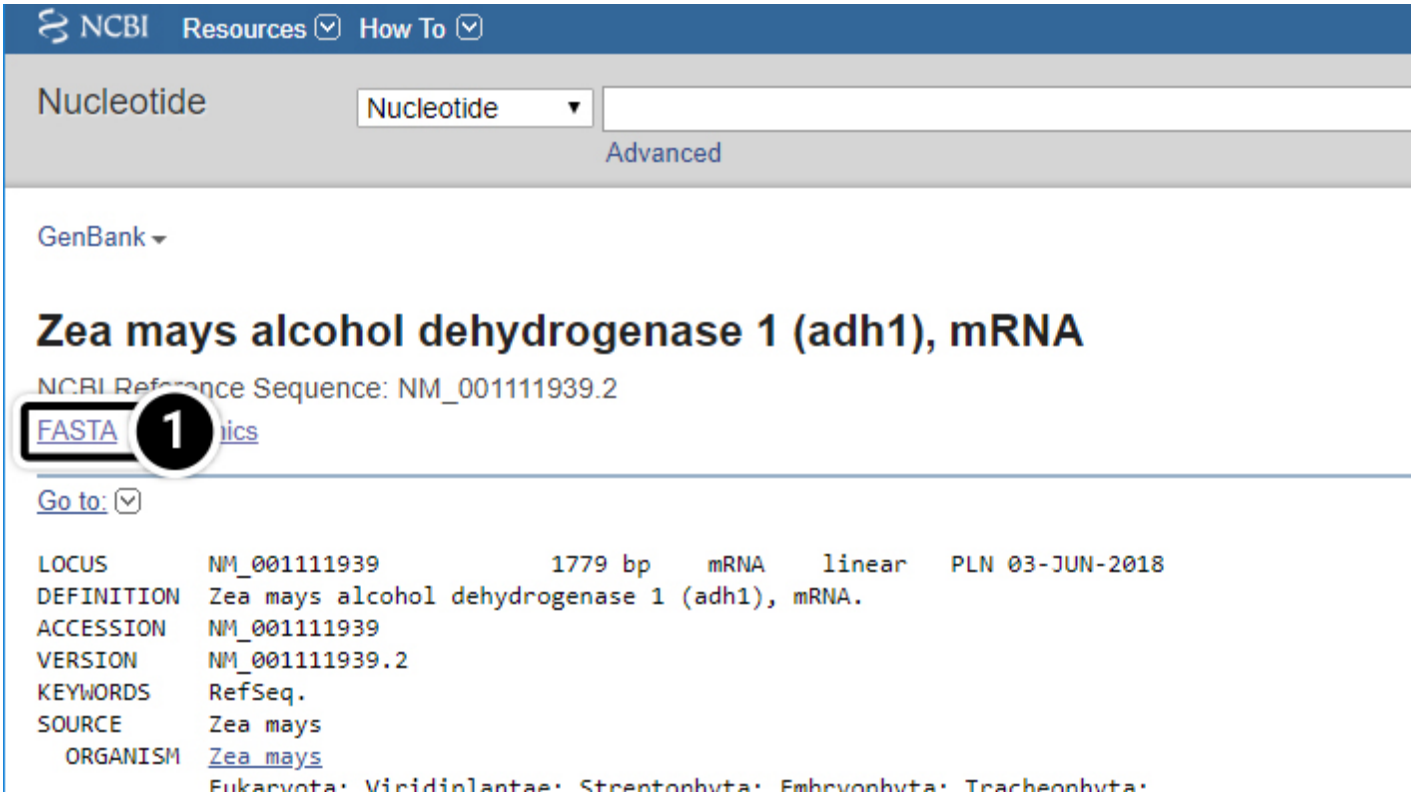


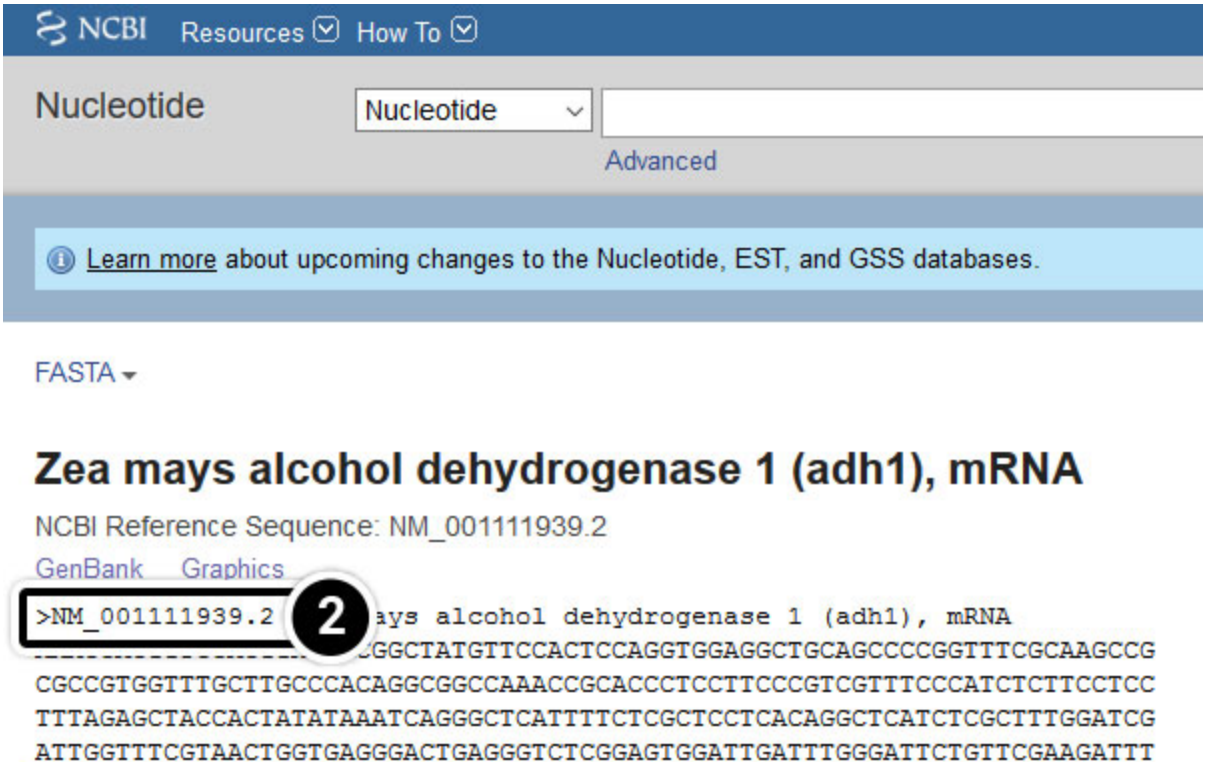
Try This: Access and Use of Sequence DBs

Let's examine the Nucleotide results.

1. Click the **FASTA** link
2. Reference sequences are accessed through GenBank to provide non-redundant curated data derived from experimental knowledge of known genes.

Additional information about RefSeq can be found **here**.




NCBI
Resources
How To
Nucleotide
Nucleotide
Advanced
Learn more about upcoming changes to the Nucleotide, EST, and GSS databases.
FASTA
Zea mays alcohol dehydrogenase 1 (adh1), mRNA
NCBI Reference Sequence: NM_001111939.2
GenBank
Graphics
>NM_001111939.2
2
ays alcohol dehydrogenase 1 (adh1), mRNA
GGCTATGTTCCACTCCAGGTGGAGGCTGCAGCCCCGGTTTCGCAAGCCG
CGCCGTGGTTTGCTTGCCCACAGGCGGCCAAACCGCACCCTCCTTCCCGTCGTTTCCCATCTCTTCCTCC
TTTAGAGCTACCACTATATAAATCAGGGCTCATTTTCTCGCTCCTCACAGGCTCATCTCGCTTTGGATCG
ATTGGTTTCGTAACTGGTGAGGGACTGAGGGTCTCGGAGTGGATTGATTTGGGATTCTGTTCGAAGATTT

Try This: Access and Use of Sequence DBs

Zea mays alcohol dehydrogenase 1 (adh1), mRNA

NCBI Reference Sequence: NM_001111939.2

GenBank Graphics

```
>NM_001111939.2 Zea mays alcohol dehydrogenase 1 (adh1), mRNA
AAACCACGGTCCACGGACCACGGCTATGTTCCACTCCAGGTGGAGGCTGCAGCCCCGGTTTCGCAAGCCG
CGCCGTGGTTTGCTTGCCCACAGGCGGCCAAACCGCACCCTCCTTCCCGTCGTTTCCCATCTCTTCCTCC
TTTAGAGCTACCACTATATAAATCAGGGCTCATTTTCTCGCTCCTCACAGGCTCATCTCGCTTTGGATCG
ATTGGTTTCGTAACTGGTGAGGGACTGAGGGTCTCGGAGTGGATTGATTTGGGATTCTGTTCGAAGATTT
GCGGAGGGGGGCAATGGCGACCGCGGGGAAGGTGATCAAGTGCAAAGCTGCGGTGGCATGGGAGGCCGGC
AAGCCACTGTCGATCGAGGAGGTGGAGGTAGCGCCTCCGCAGGCCATGGAGGTGCGCGTCAAGATCCTCT
TCACCTCGCTCTGCCACACCGACGTCTACTTCTGGGAGGCCAAGGGGCAGACTCCCGTGTTCCCTCGGAT
CTTTGGCCACGAGGCTGGAGGTATCATAGAGAGTGTTGGAGAGGGTGTGACTGACGTAGCTCCGGGCGAC
CATGTCCTTCCTGTGTTCACTGGGGAGTGCAAGGAGTGTGCCCACTGCAAGTCGGCAGAGAGCAACATGT
GTGATCTGCTCAGGATCAACACCGACCGCGGTGTGATGATTGCCGATGGCAAGTCGCGGTTTTCAATCAA
TGGGAAGCCTATCTACCACTTTGTTGGGACTTCCACCTTCAGCGAGTACACCGTCATGCATGTGGGTTGT
GTTGCAAAGATCAACCCTCAGGCTCCCCTTGATAAAGTTTGCGTCCTTAGCTGTGGTATTTCTACCGGTC
TTGGTGCATCAATTAATGTTGCAAAACCTCCGAAGGGTTCGACAGTGGCTGTTTTCGGTTTAGGAGCCGT
TGGTCTTGCCGCTGCAGAAGGTGCAAGGATTGCTGGAGCGTCAAGGATCATTGGTGTCGACCTGAACCCC
AGCAGATTCGAAGAAGCTAGGAAGTTCGGTTGCACTGAATTTGTGAACCCAAAAGACCACAACAAGCCAG
TGCAGGAGGTACTTGCTGAGATGACCAACGGAGGGGTCGACCGCAGTGTGGAATGCACTGGCAACATTAA
TGCTATGATCCAAGCTTTCGAATGTGTTCATGATGGCTGGGGTGTTGCTGTGCTGGTGGGTGTGCCACAT
AAGGACGCTGAGTTCAAGACCCACCCGATGAACTTCCTGAACGAAAGGACCCTGAAGGGGACCTTCTTTG
GCAACTATAAGCCACGCACTGATCTGCCAAATGTGGTGGAGCTGTACATGAAAAAGGAGCTGGAGGTGGA
GAAGTTCATCACGCACAGCGTCCCGTTCGCCGAGATCAACAAGGCGTTCGACCTGATGGCCAAGGGGGAG
GGCATCCGCTGCATCATCCGCATGGAGAACTAGATTTCGCTGTCTAGTTTGTGATCTGGCCTGGGCTTGG
GGTTAATAAAAGAGGCAATGCTAGCCTGCCCTTTCGATGAGGAGGTACATACACGCTGGCGATGGACCGC
GCTTGTGTGTCGCGTTCAGTTTGGCTTTTGCCAAGCAGTAGGGTAGCTTCCCGTGTCGGTAATTATATGG
TATGAACCATCACCTTTTGGCTCTACATGGTATGAACGTAAGATACAAATTCCAACTACCTCTAGCTCGC
TTGTGTGGTATCTGTATCAGTATTCATGTGTTTGTTTGCTTATGTGTTTGTTTGCTTGTATTTGCTGGTG
CTTGTATCGCGGGATGCAATGAGTTGTTG
```

After clicking the FASTA link, what kind of information do you get?
Does the entire mRNA sequence for adh1 you obtained code for a protein product? If not, how would you identify the coding sequence?
Now that you have found the adh1 coding sequence, what is its estimated length?

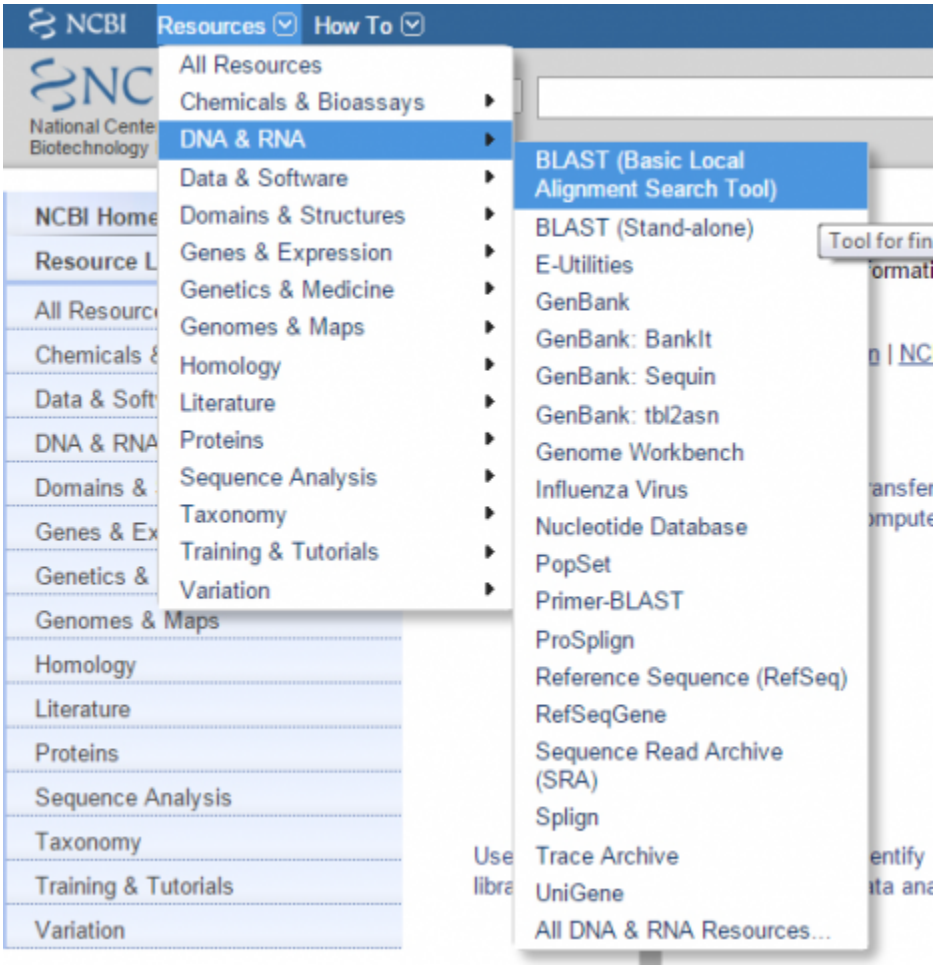
Enter your answer here.

Show Answer

NCBI BLAST

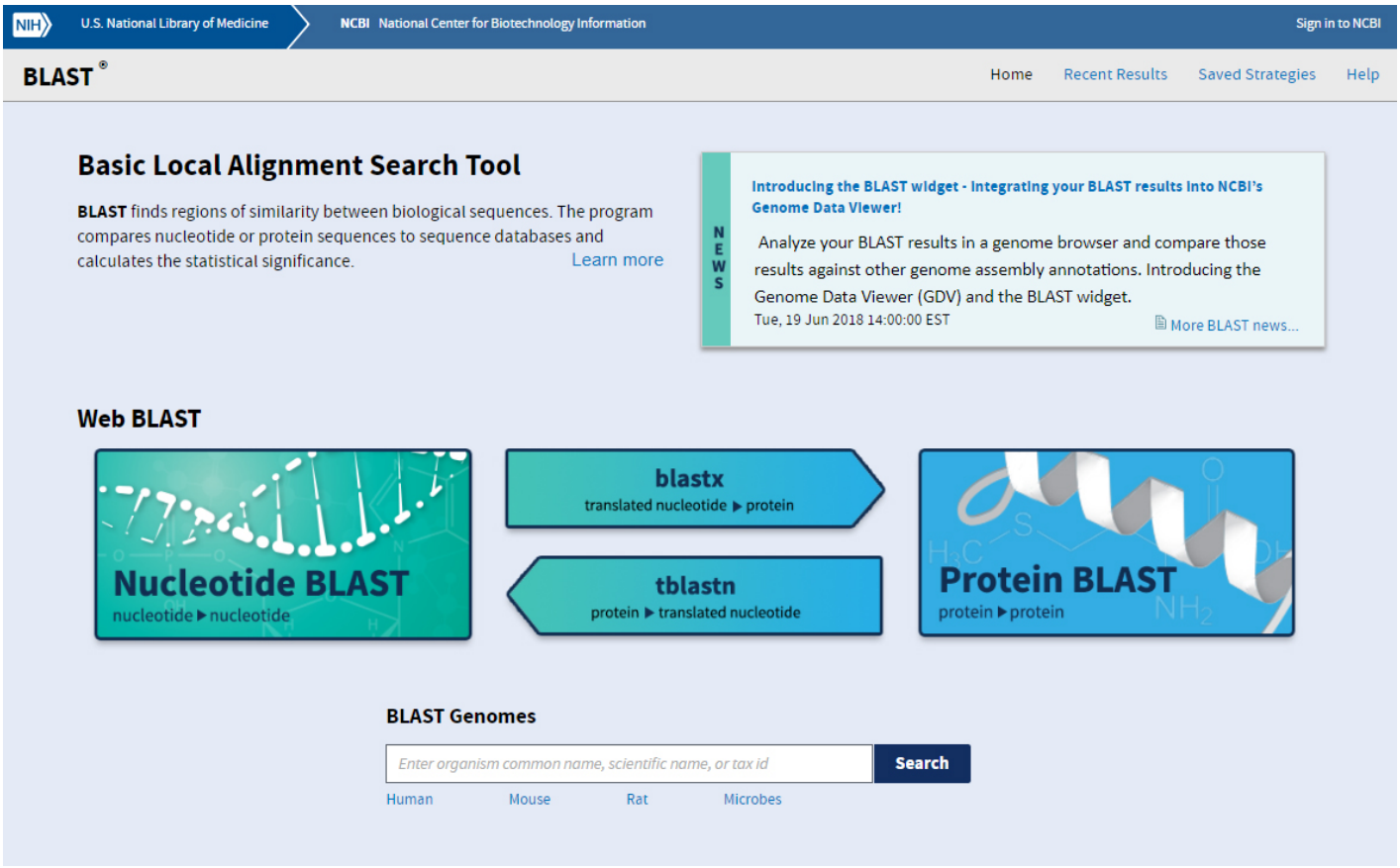
NCBI Basic Local Alignment Search Tool (BLAST)

Not only keywords can be used to search sequence databases. Sequences can also be used to perform a BLAST search, making BLAST probably the most important tool in any sequence database. BLAST allows the comparison of sequence data using an algorithm developed by Altschul et al. (1990). The algorithm attempts to detect high-scoring segment pairs, which are pairs of sequences that can be aligned with one another and, when aligned, meet the certain scoring and statistical criteria.



BLAST Interface

Here the user can restrict searches to a specific species and to the assembled reference sequences for that species. For a plant researcher, it may not be necessary to restrict a search except for those working with rice and Arabidopsis. For all other plant species reference sequences are not fully developed.



BLAST Features

1. Basic BLAST features include blastn, blastp, blastx, tblastn, and tblastx.
2. Specialized features include "**Global Align**" for sequence alignment.

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

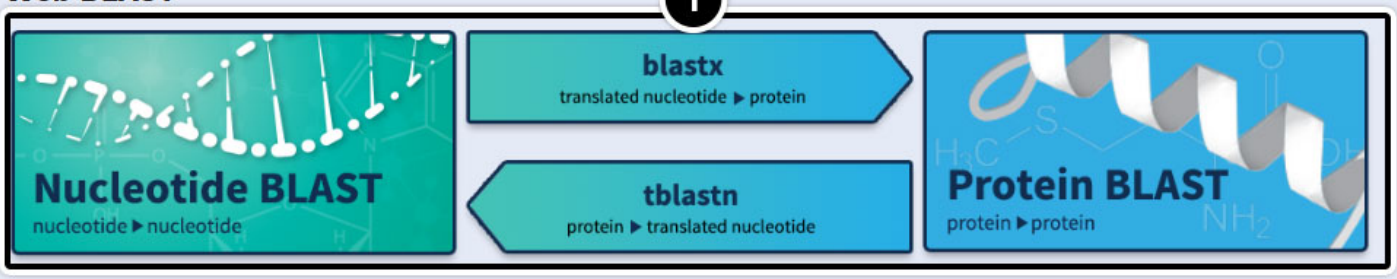
A new version (1.4.0) of the BLAST RNA-seq mapping tool, Magic-BLAST, is now available

Tue, 21 Aug 2018 16:00:00 EST

More BLAST news...

Web BLAST

1



BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human | Mouse | Rat | Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

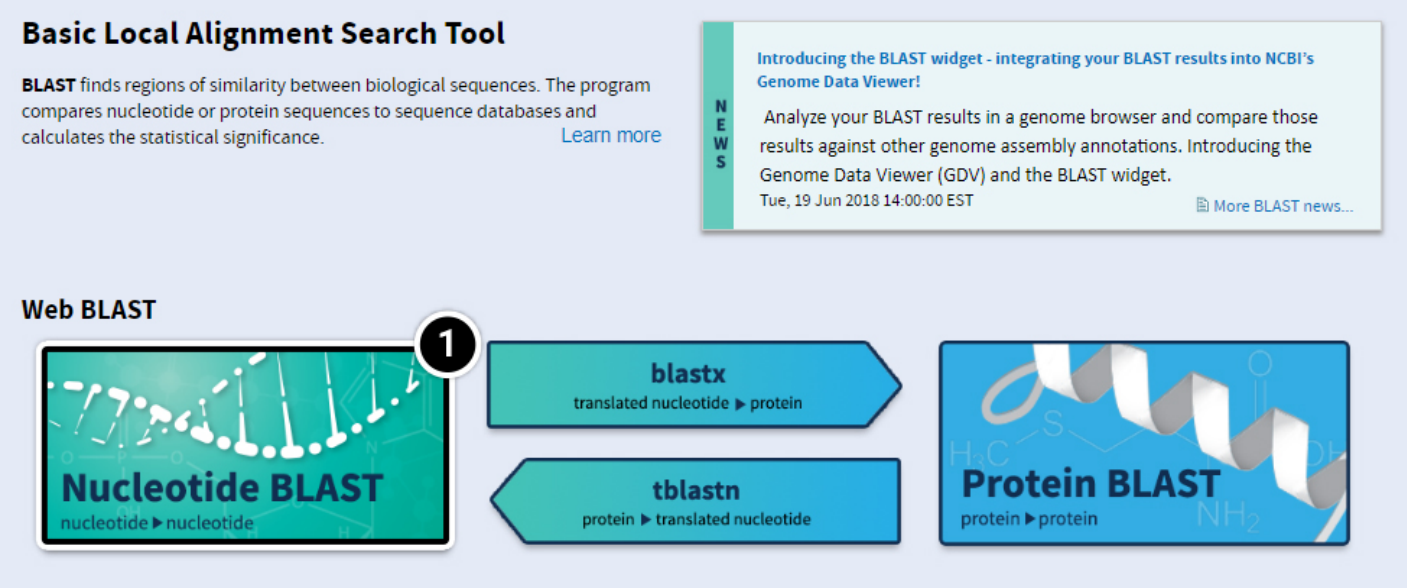
2

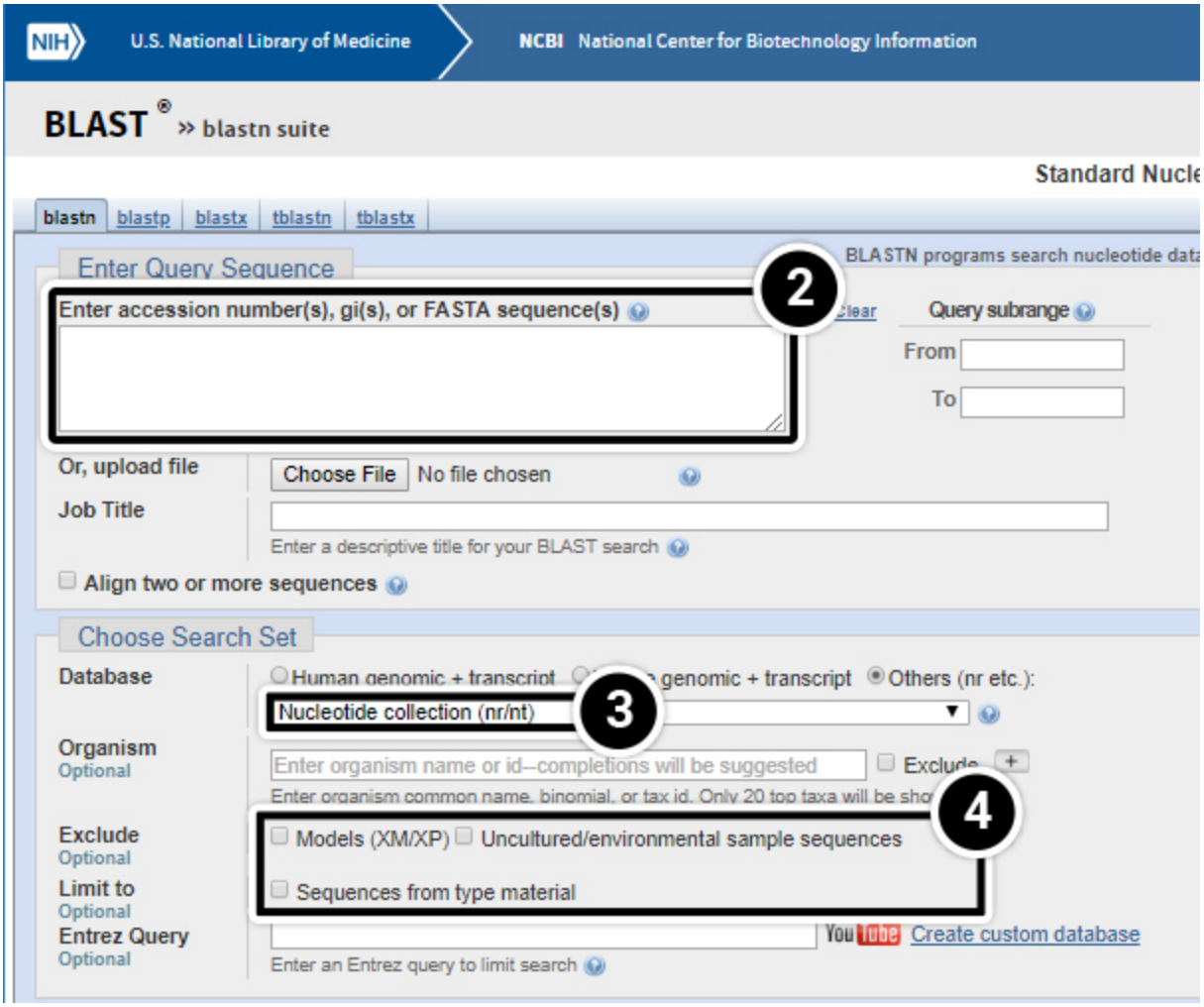
SmartBLAST	Primer-BLAST	Global Align	CD-search
Find proteins highly similar to your query	Design primers specific to your PCR template	Compare two sequences across their entire span (Needleman-Wunsch)	Find conserved domains in your sequence

Try This: Using NCBI BLAST

Try This: Using NCBI BLAST

1. Within the **Basic BLAST** window, click on **Nucleotide BLAST**. A new window appears asking you to setup your search options.
2. This is where your query sequence will go.
3. This selects the D**atabase** you want to search.
4. Other parameters you may want to set different from the standard settings.




U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST ® » blastn suite
Standard Nucle
blastn blastp blastx tblastn tblastx
Enter Query Sequence
BLASTN programs search nucleotide data
2
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
3
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be sho
4
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
YouTube
Create custom database
Enter an Entrez query to limit search

Try This: Using NCBI BLAST

You have various options of entering your query sequence: copy and paste or uploading a saved sequence from your computer.

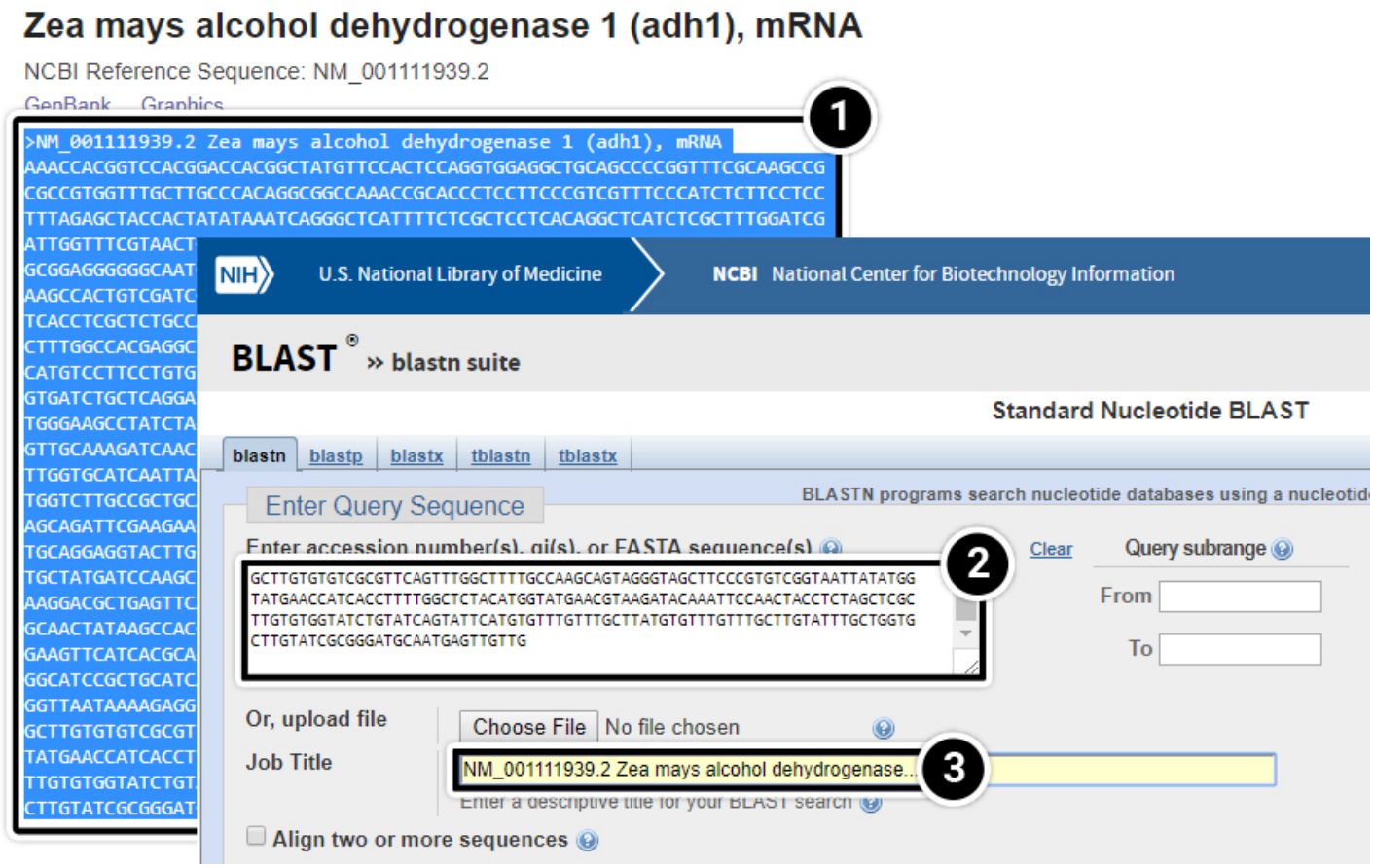
Your query sequence has to be annotated in FASTA format. FASTA is a text-based format consisting of a definition line followed by the sequence data in single letter code. The definition line starts with the character ">", followed by a sequence name, and ends with a return or newline. Everything that follows until the next ">" will be considered as the sequence data. It is possible to save multiple sequences in one FASTA file.

1. In the screenshot below,
   Definition line starts with "**>**" character,
   gi stands for GenBank identification, followed by GenBank ID number,
   ref stands for reference sequence, followed by the accession number.
   Both GenBank ID and reference sequence numbers can be used to enter a query sequence into BLAST.

Try This: Using NCBI BLAST

You may enter your query (adh1) as a sequence in FASTA format.

1. To do that, copy the entire adh1 sequence
2. Paste it in the **Enter accession numbers(s), gi(s), or FASTA sequence(s)** window.
3. Note that the **Job Title** filled automatically.



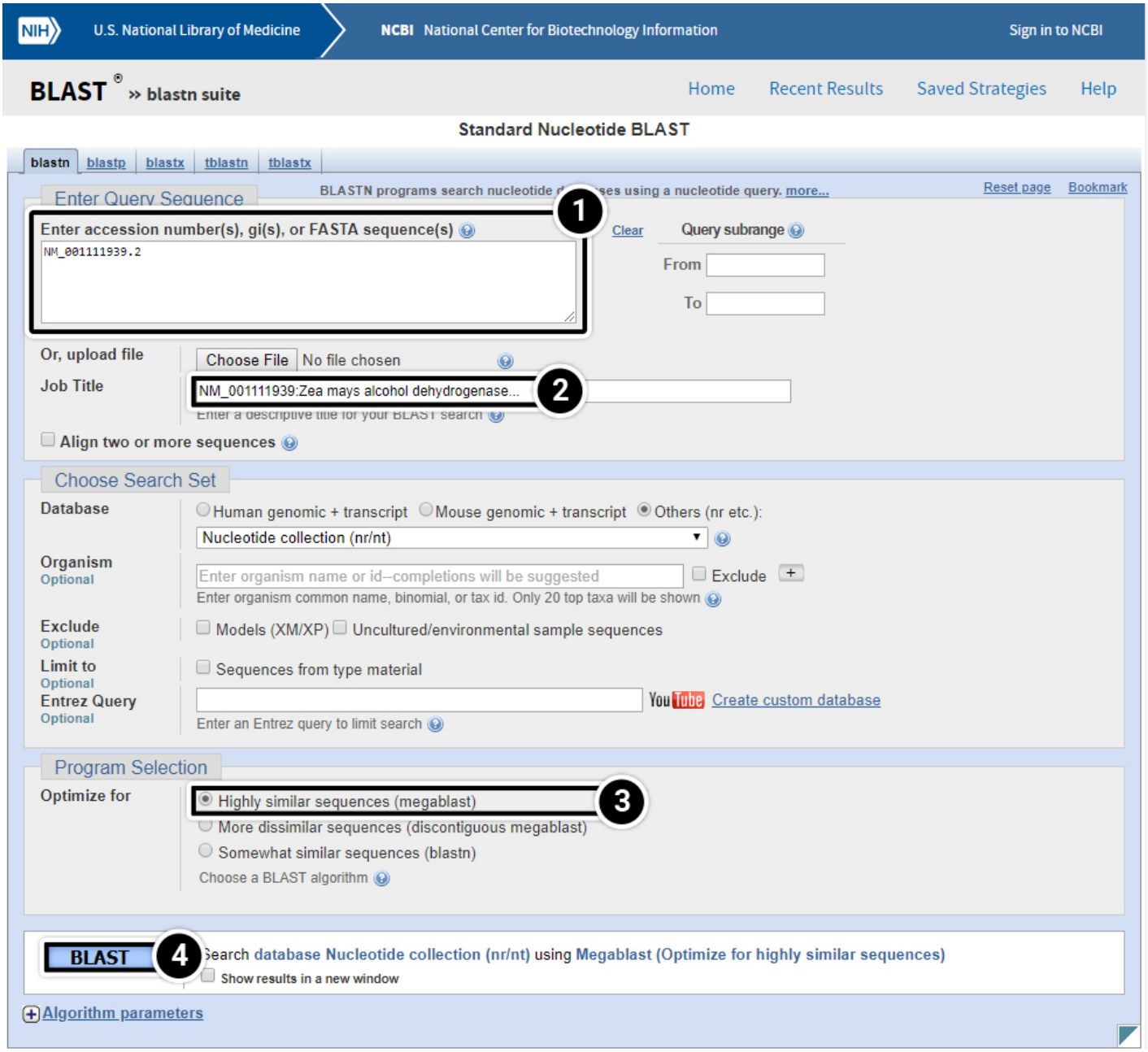
Try This: Using NCBI BLAST

1. Alternatively, you can query your sequence using the **Run BLAST** command. Click "Run BLAST" to query the sequence from the FASTA display screen.

Try This: Using NCBI BLAST

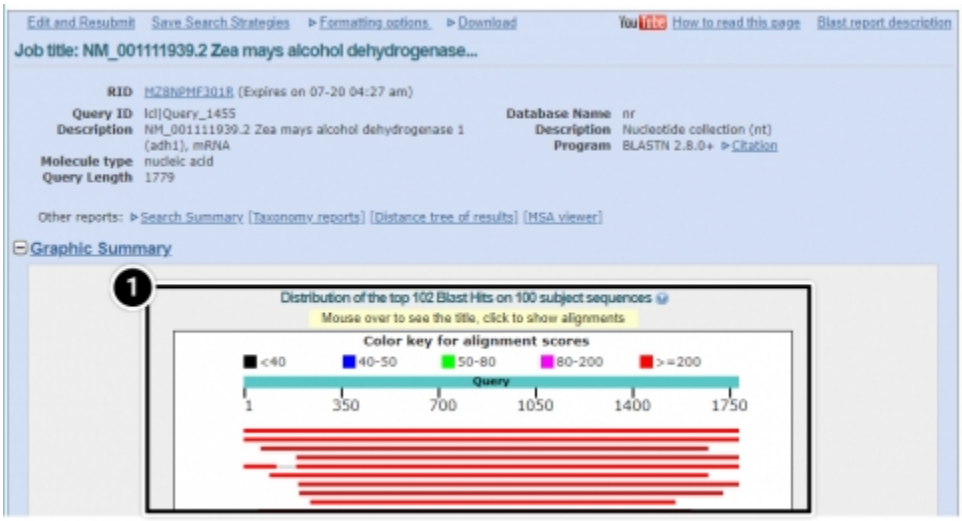
Clicking on the **Run BLAST** command will lead you to this window.

1. **Accession number** of adh1 will automatically fill in
2. **Job Title** should automatically fill in, if it does not you can click in the **Job Title** field and it should appear automatically.
3. Optimize your search to **megablast** to identify highly similar sequences.
4. Finally, select the **BLAST** button.



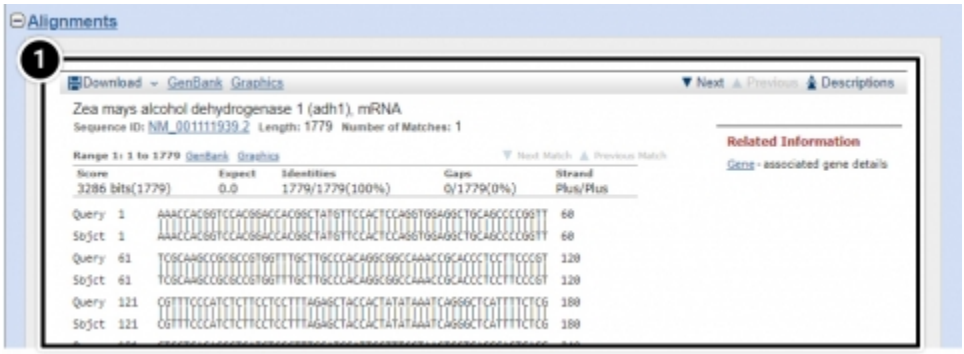
Try This: Using NCBI BLAST

1. **Graphic Summary:** BLAST results that are summarized in a graphic form.



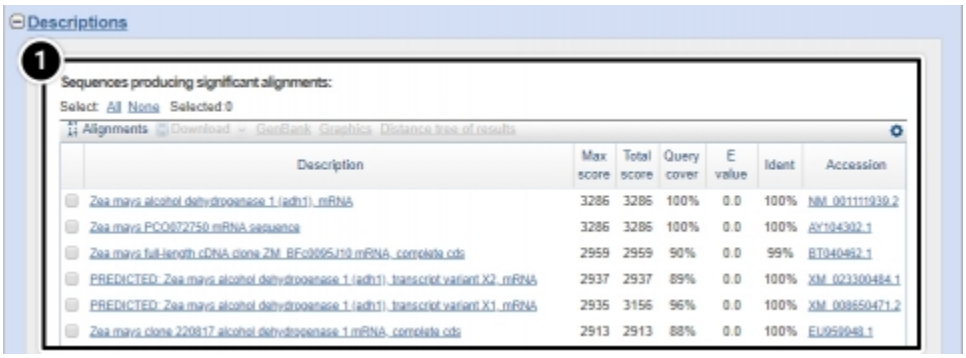
Try This: Using NCBI BLAST

1. **Alignments:** BLAST results that contain sequence alignment information.



Try This: Using NCBI BLAST

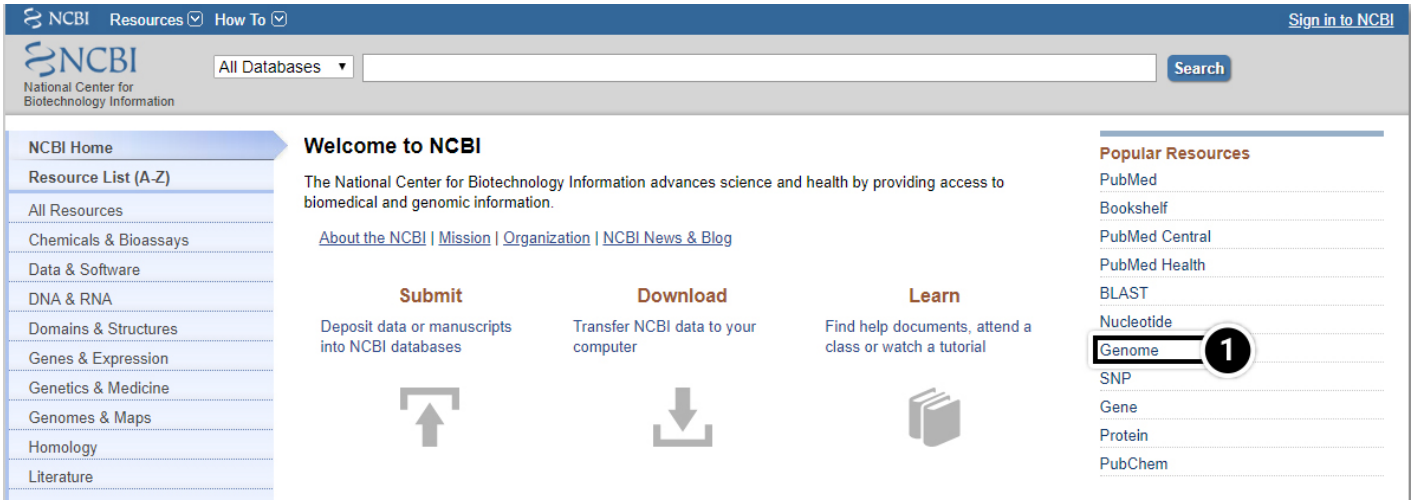
1. **Descriptions:** Accession number and source organism information is provided for sequences producing high alignment scores.



Try This: Using NCBI BLAST

Step 4: Locating adh1 on a chromosome

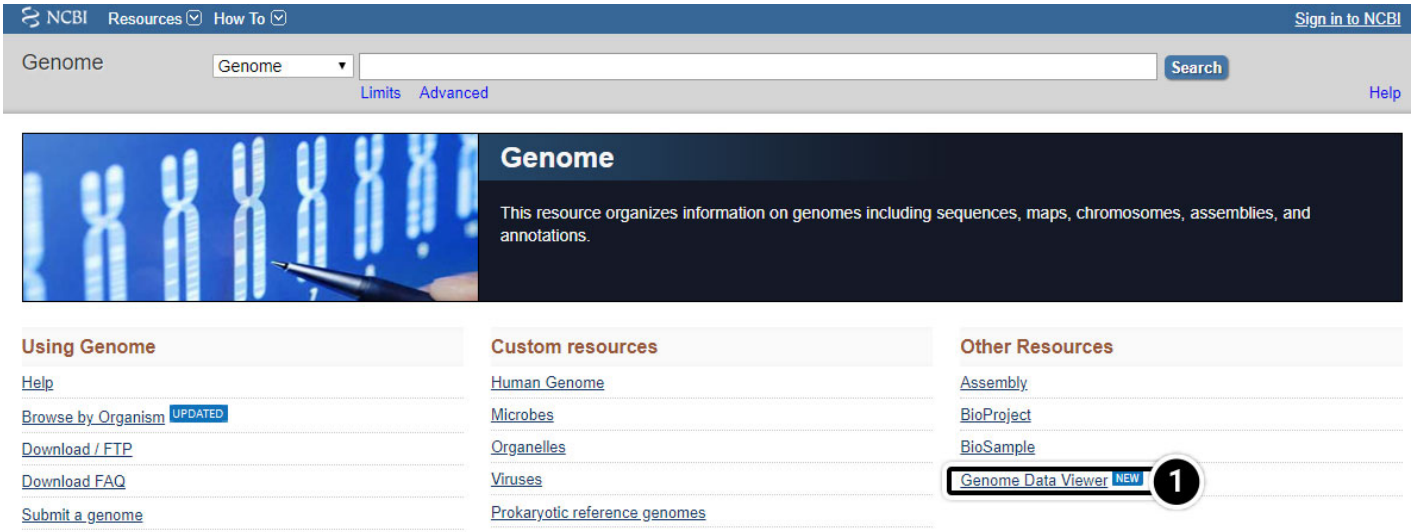
1. From the NCBI home page, select **Genome**.



Try This: Using NCBI BLAST

Locating adh1 on a chromosome

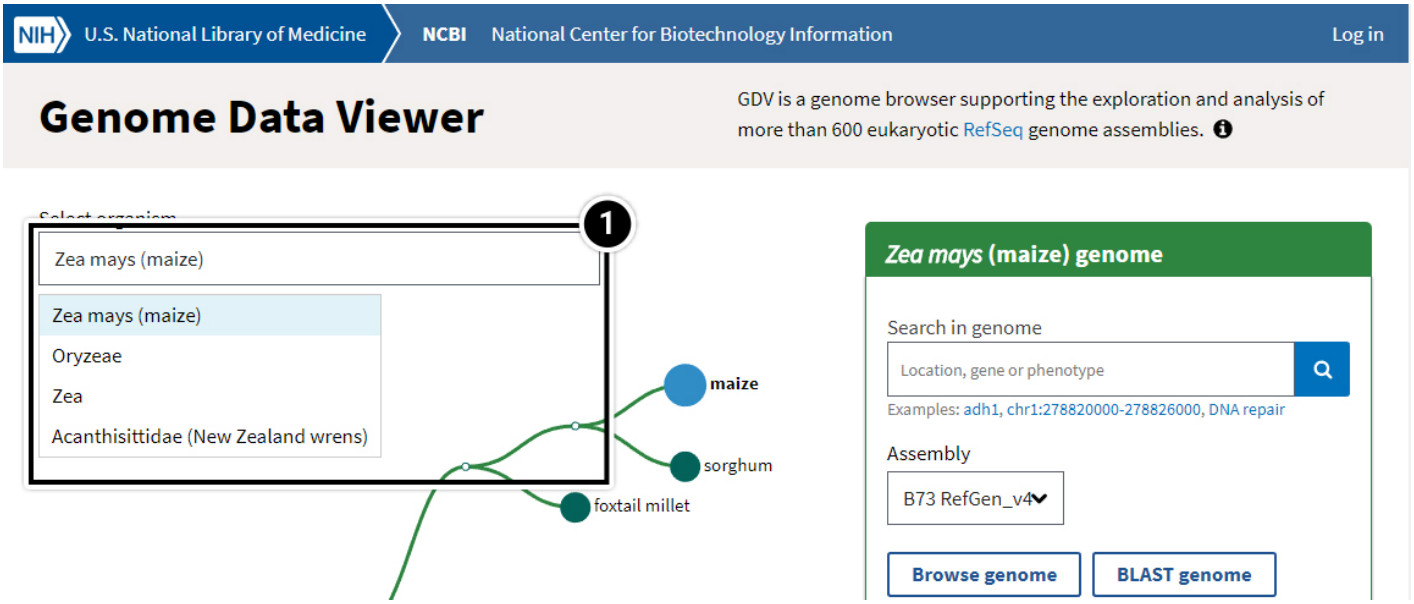
1. From the genome page, select **Genome Data Viewer (previously known as Map Viewer)**.



Try This: Using NCBI BLAST

Locating adh1 on a chromosome

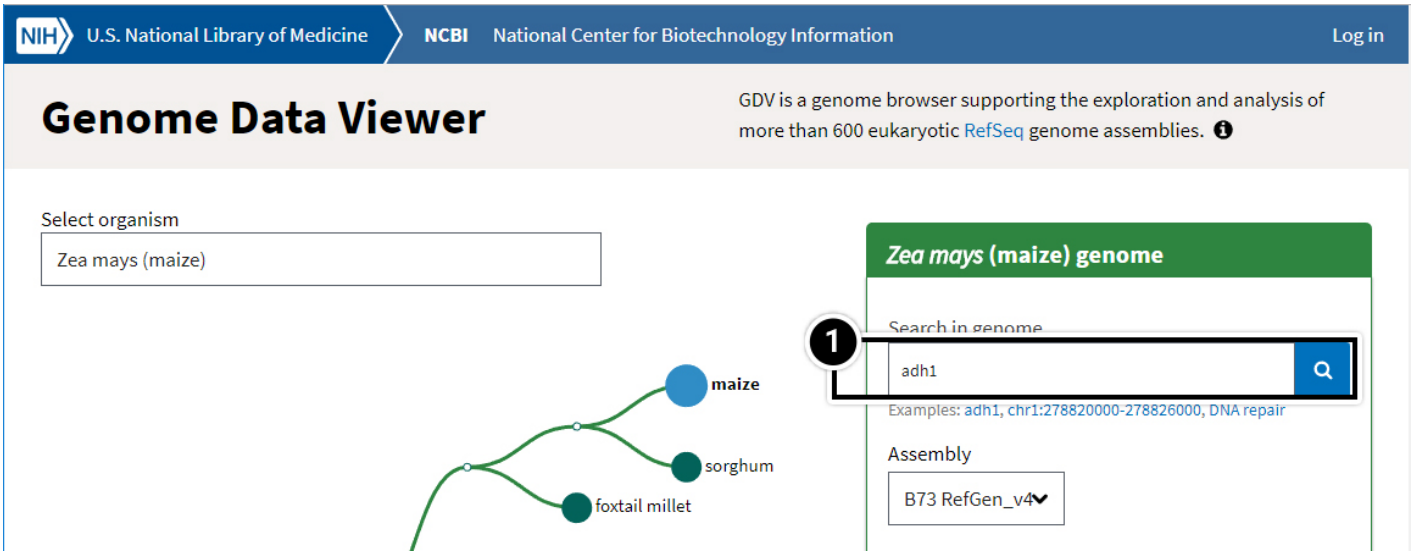
1. Within **Genome Data Viewer** home you can select your organism or species.



Try This: Using NCBI BLAST

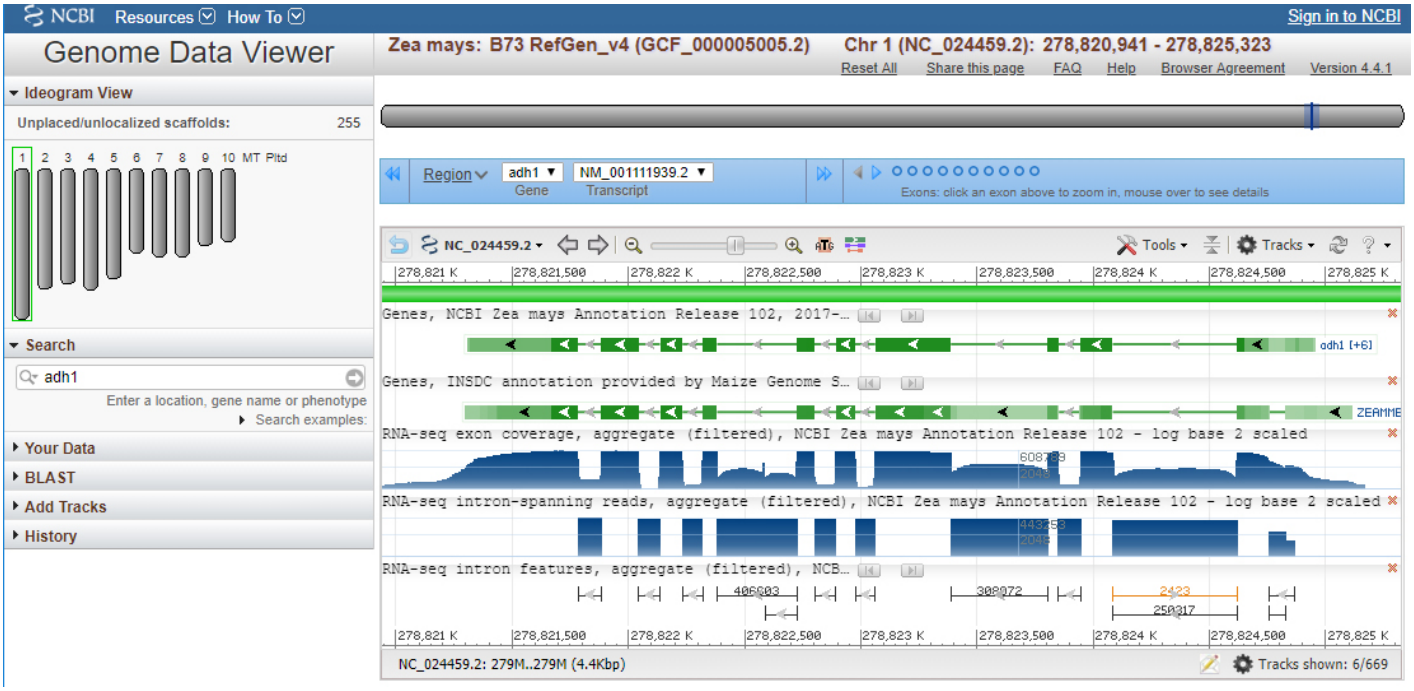
Locating adh1 on a chromosome

1. Try searching the Zea mays genome for the adh1 gene.



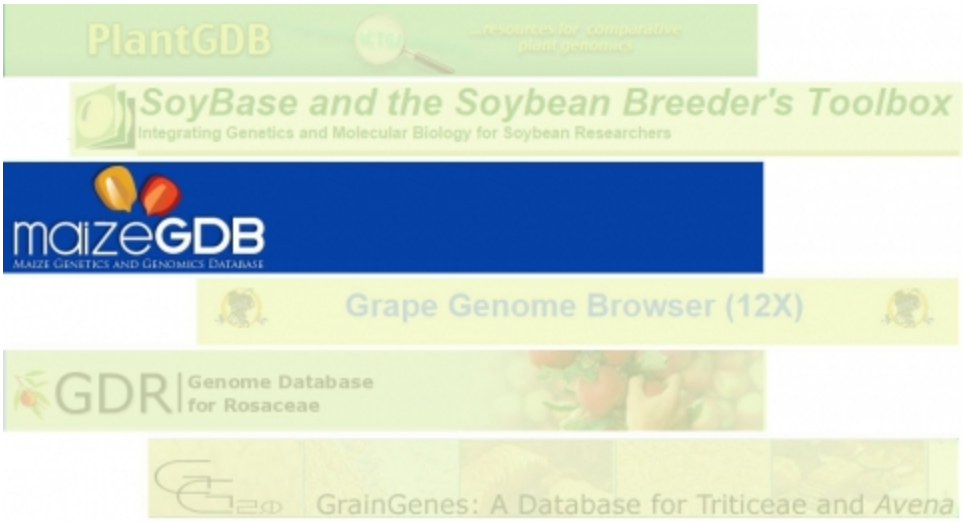
Try This: Using NCBI BLAST

The NCBI Map View search for adh1 on the maize genome produces these results. "Ideogram view" highlights chromosome 1 to show that the adh1 gene is located on chromosome 1.



Plant Species Sequence Databases

The advent of genomics has resulted in a number of plant species specific sequence databases. For this lesson, Maize Genetics and Genomics Database (MaizeGDB) will be the focus.



MaizeGDB

MaizeGDB was first released in 1991 (as MaizeDB) and has transitioned from a focus on curation of genetic maps and stocks to the handling of reference maize genome sequence, multiple maize genomes, and sequence-based gene expression data. MaizeGDB relies on the research community for data and on expertise distributed across the USA. We recommend the use of an internet browser other than Internet Explorer (e.g. Google Chrome) to access the [MaizeGDB site](#).

1. Tutorials are available by clicking the video icon.

MaizeGDB: Tutorials

Useful tutorials are available to help the user familiarize with MaizeGDB.

Try This: Using MaizeGDB

Perform a Basic Search

Similar to NCBI, the MaizeGDB is a composite database allowing you to search broadly among databases or to restrict your query to a single database.

1. Open your web browser and go to **https://www.maizegdb.org**.
2. Enter **adh1** into the search box.
3. Press **Enter** or click the Search icon to search within all available data.

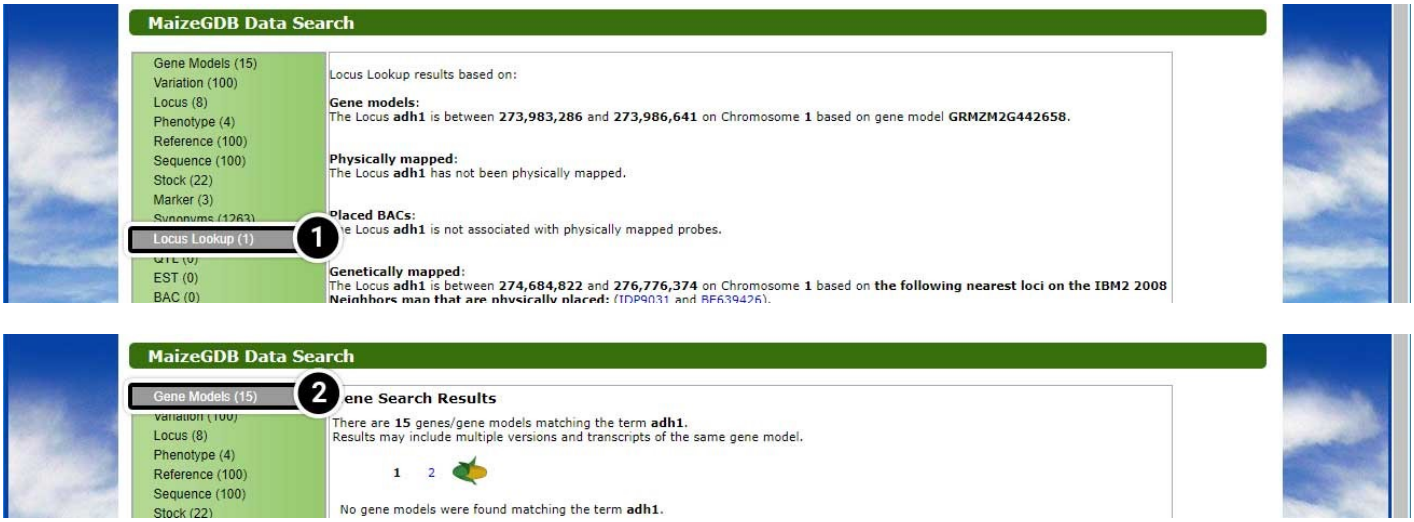
Click the image below to see a larger version.

Explore the Search Results

This search will lead you to a window containing various options.

1. Click on **Locus Lookup (1)** in the left-hand menu.
2. Click on **Gene Models (15)** in the left-hand menu.

Explore the other data available to you by clicking the links in the green box. Click the image below to see a larger version.



Access the Genome Browser

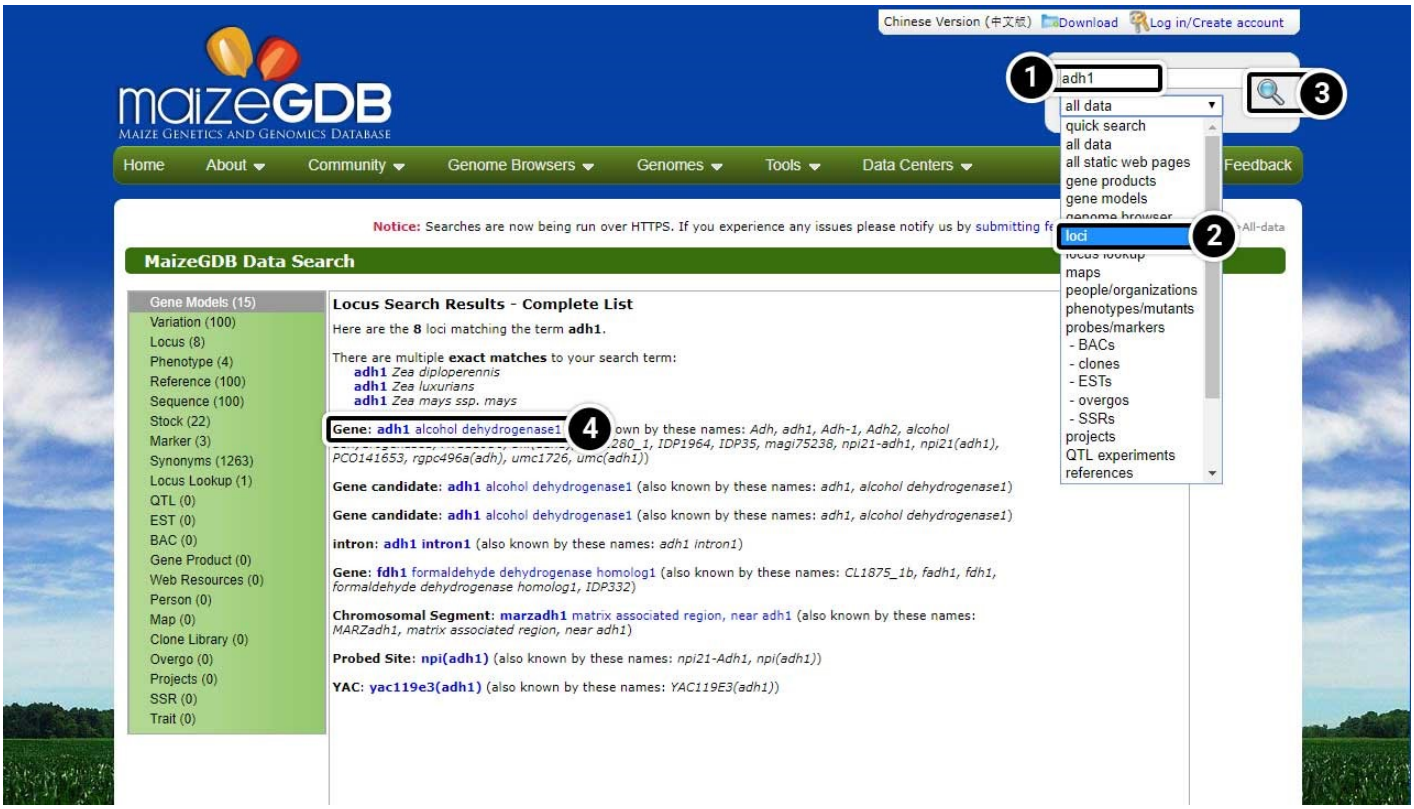
Access the genome browser to obtain information about maize adh1.

1. Type **adh1** into the Search bar.
2. Select **loci** from the Search options.
3. Click the Search icon.

This time when the results load, you will see only the loci associated with adh1.

4. Click the link for **adh1 alcohol dehydrogenase1** to take a closer look at the gene.

Click the image below to see a larger version.



Explore the Gene Record

The locus record screen provides detailed information on the adh1 gene. Explore the genetic information for **adh1 alcohol dehydrogenase1**.

1. Click on **Chromosome Coordinates** when ready to proceed.

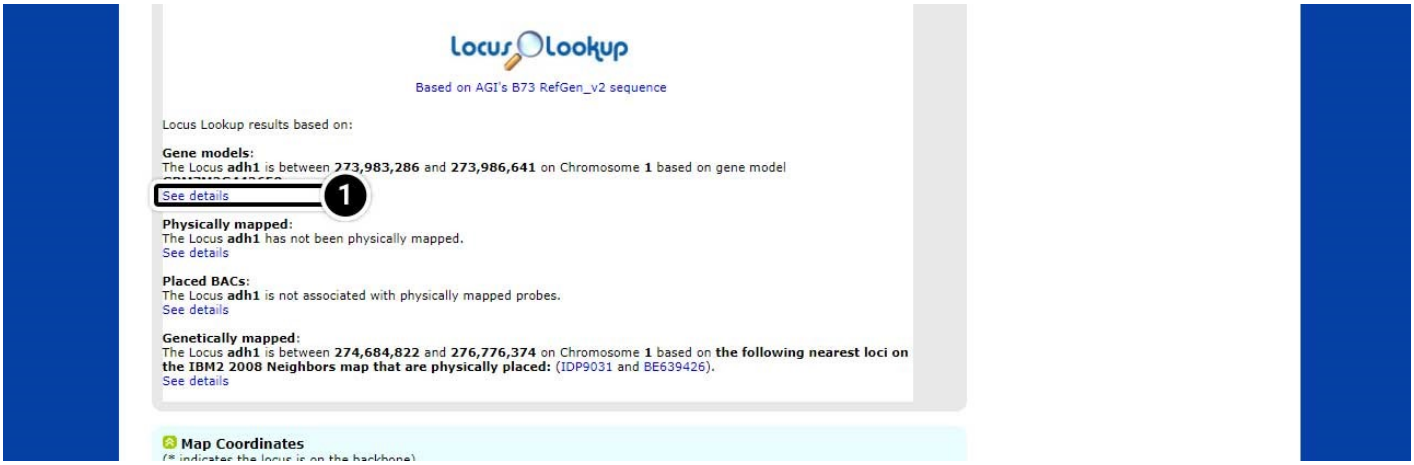
Click the image below to see a larger version.

See Details in Locus Lookup

The page will scroll down to the Locus Lookup section.

1. Click on **Show details** to expand this section of the results.

Click the image below to see a larger version.



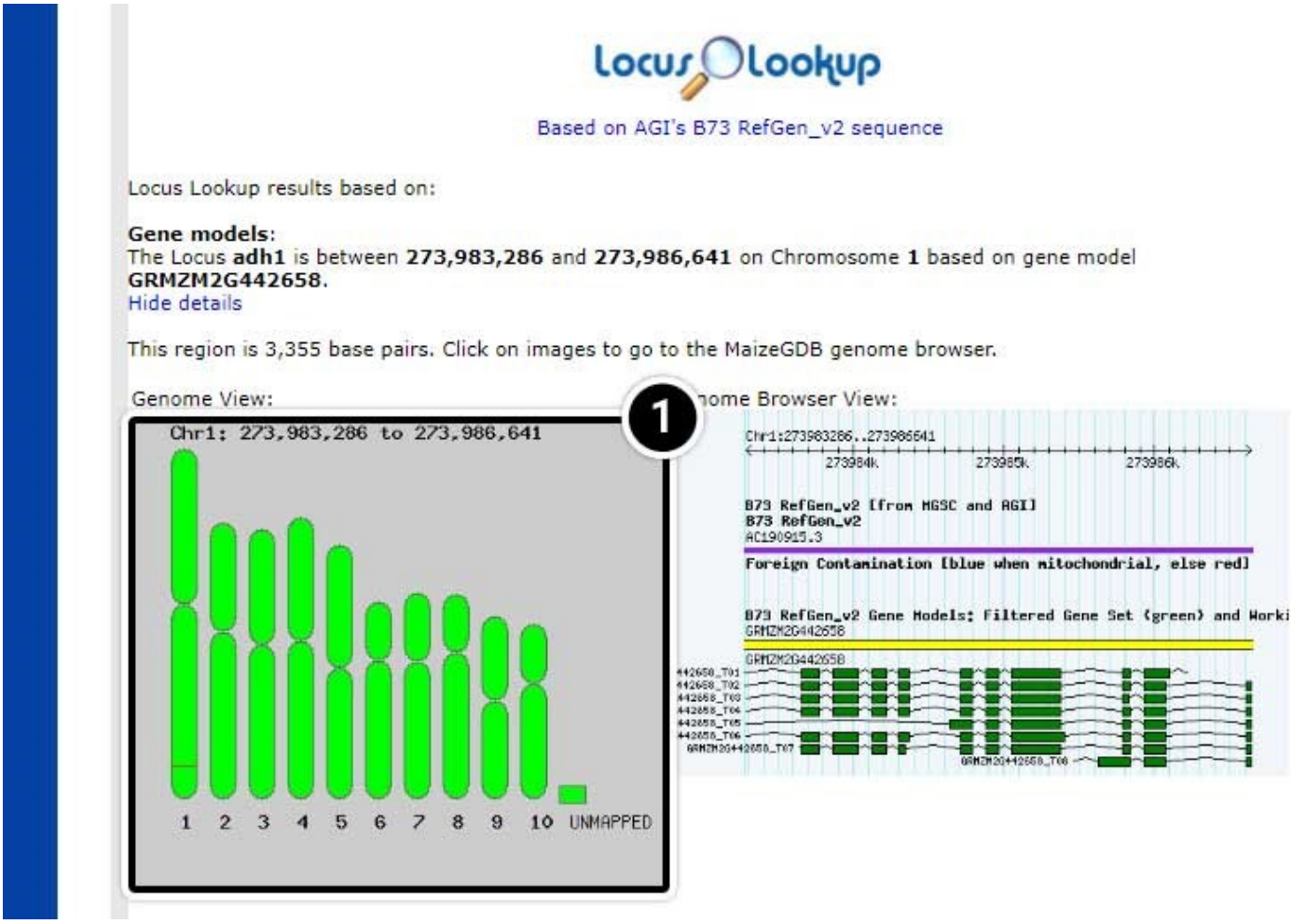
Expanded Details in Locus Lookup

When the details have loaded, explore the available information.

Note the position of adh1 based on “AGIs B73 RefGen_v2 sequence” (adh1 is located between 273,983,286 and position 273,986,641 on chromosome 1.

1. Click on the map image to launch the MaizeGDB genome browser.

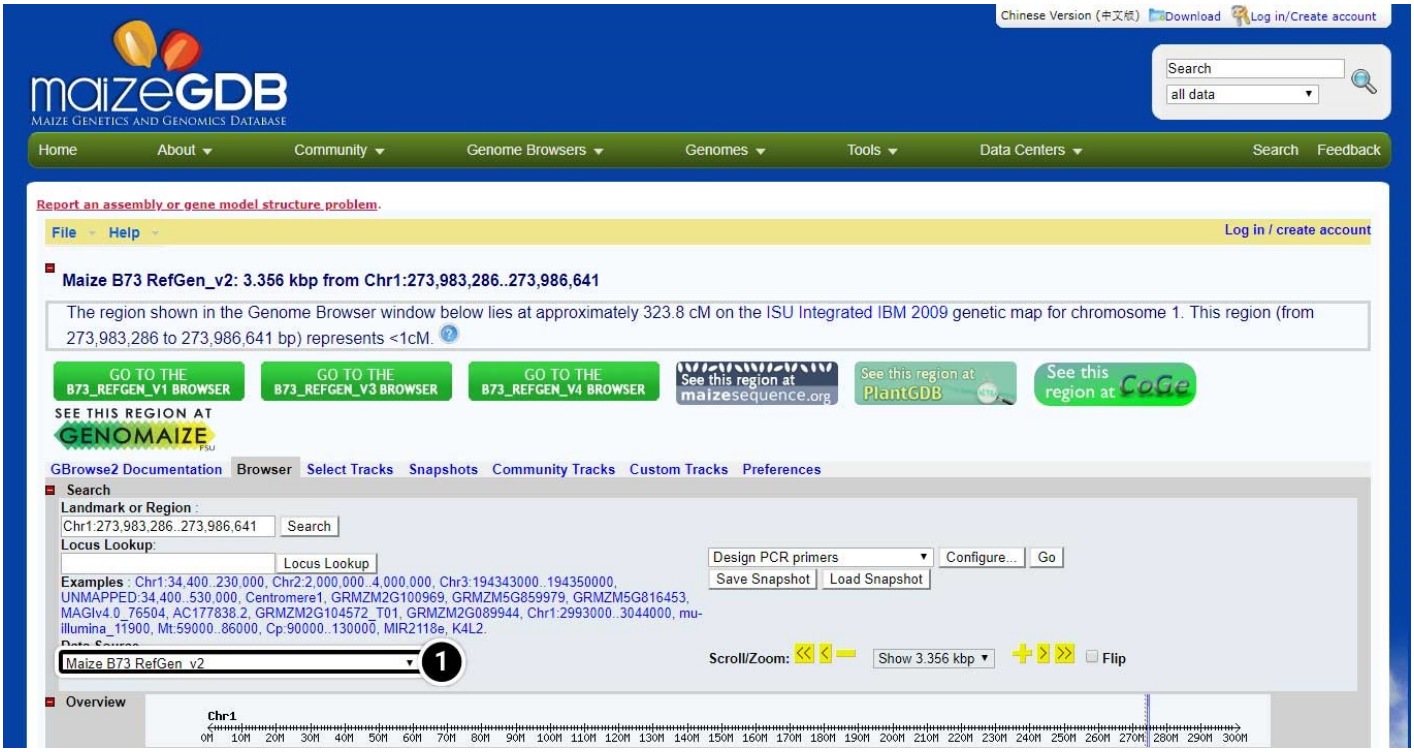
Click the image below to see a larger version.



View Datasets in the MaizeGDB Genome Browser

The MaizeGDB Genome Browser is displayed. Click the image below to see a larger version.

1. Here you can use the other datasets available in MaizeGDB including **B73 RefGen_v1 sequence**, **B73 RefGen_v3 sequence**, **B73 RefGen_v4 sequence**, and **BAC-based genome assembly**.



MaizeGDB Study Questions

Find the position of adh1 on chromosome 1 in "B73 RefGen_v1 sequence" and "BAC-based genome assembly."

B73 RefGen_v2 sequence: Between 273,983,286 and 273,986,641

B73 RefGen_v1 sequence: Between [ ] and [ ]

BAC-based genome assembly: Between [ ] and [ ]

Check

What might be the reason for the discrepancy in adh1 position among the three datasets?

Enter your answer here.

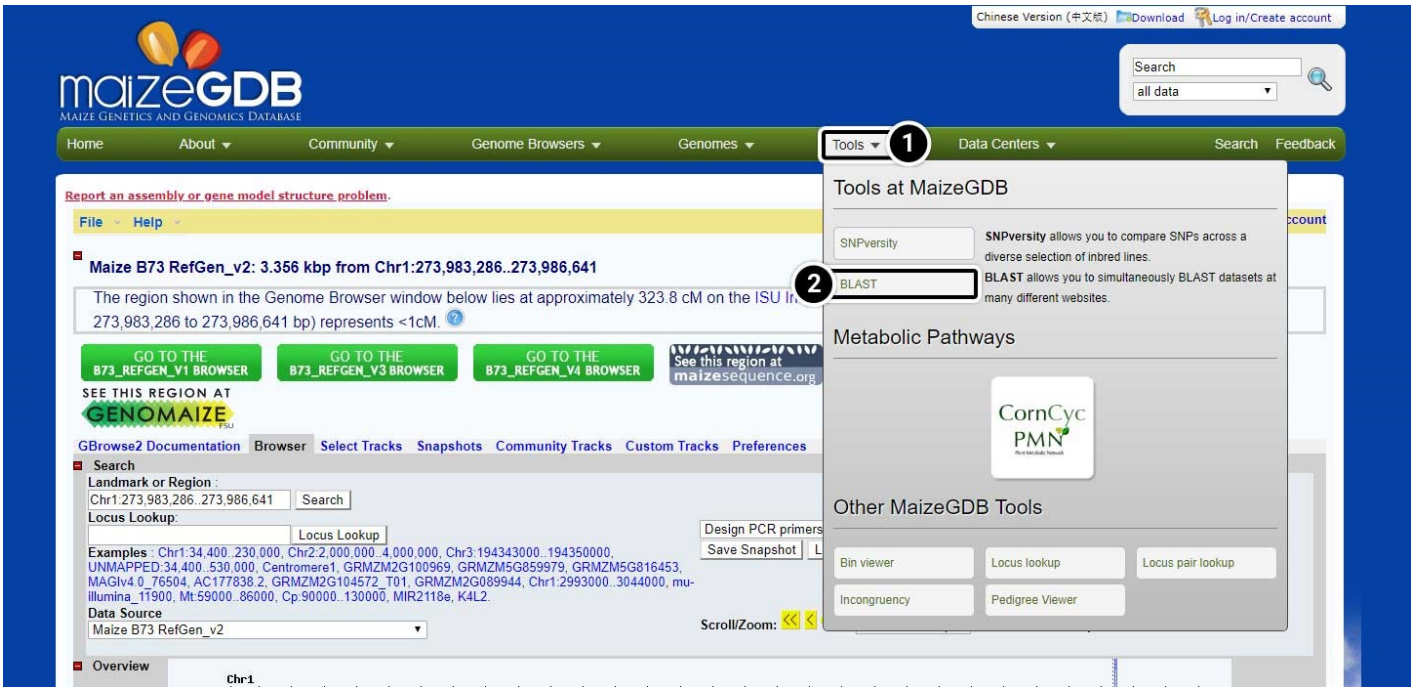
Show Answer

Navigate to BLAST

Next, we will conduct a BLAST search for adh1 in maize GDB using adh1 mRNA from GenBank.

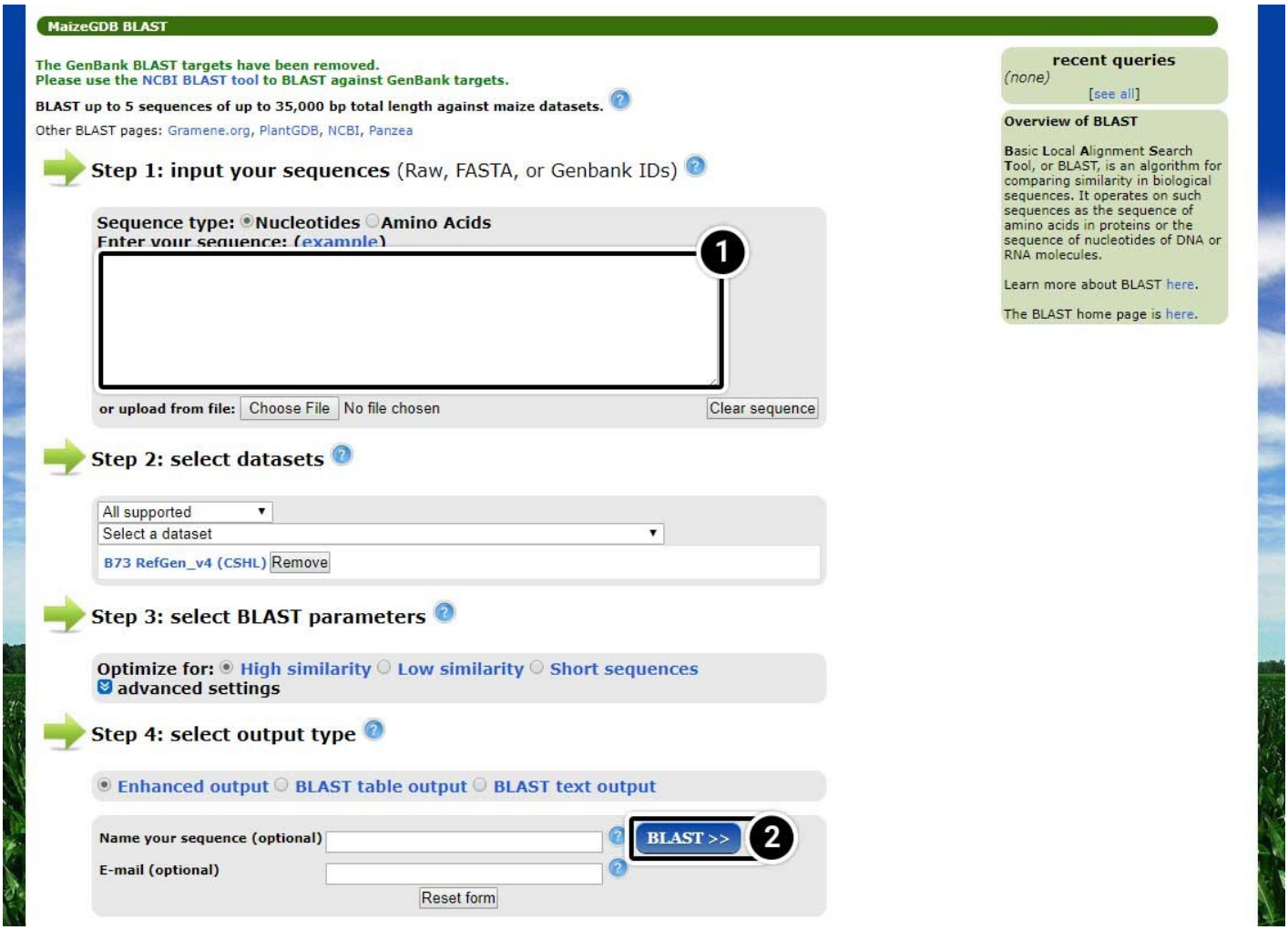
1. In the navigation bar, hover over **Tools**
2. Then click the **BLAST** button.

Click the image below to see a larger version.



Input the BLAST Parameters

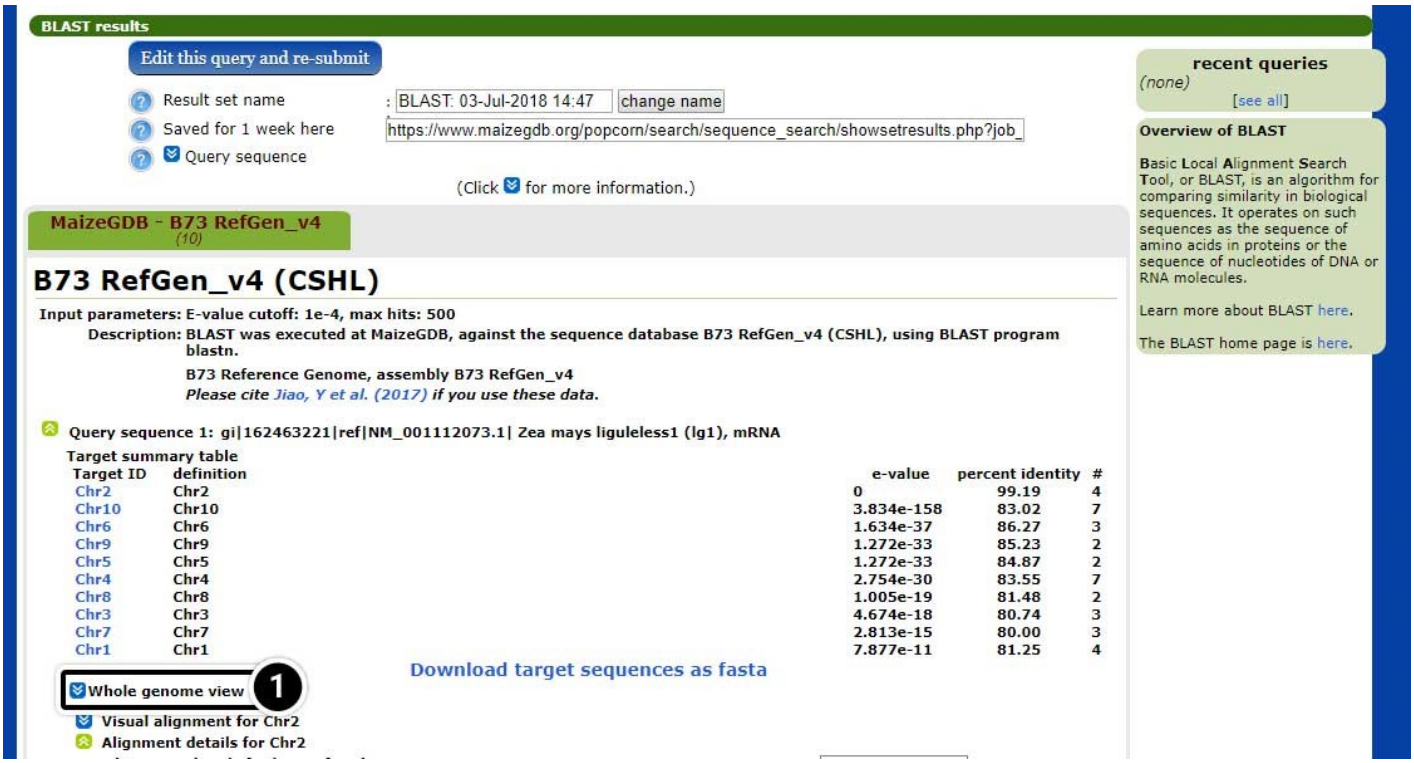
1. Enter the adh1 mRNA sequence in FASTA format in the box.
2. Use the default parameters to search for adh1 and click the BLAST button.



View the BLAST Results

The table of BLAST results includes information on chromosomes, probability values, sequence identity, and the number of likely candidates (hits). Also, you can view a representation of the entire genome in the context of where adh1 may be located.

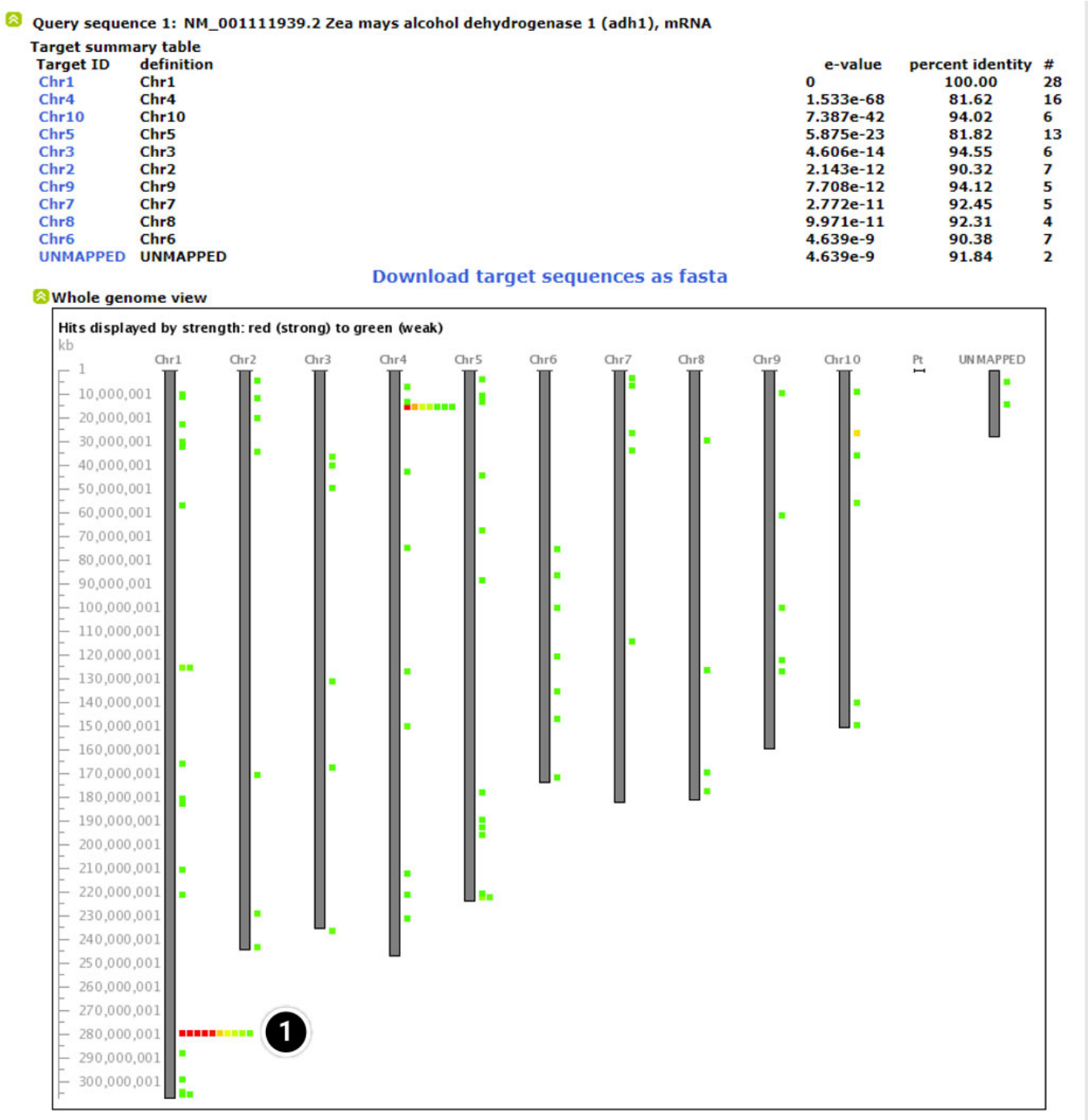
1. Click the arrow next to “Whole genome view” to see the entire genome in context of where adh1 may be located.



Explore the Whole Genome View

The whole genome view allows visualization of the 10 chromosomes of maize including, the predicted position match the adh1 sequence.

1. Click on "Chr1" corresponding to the red box on Chromosome 1 (E-value = 0).
2. Now, click on "View at MaizeGDB", next to the hit on Chr1.



Visual alignment for Chr1

Alignment details for Chr1

Alignment details for hit #1 for Chr1

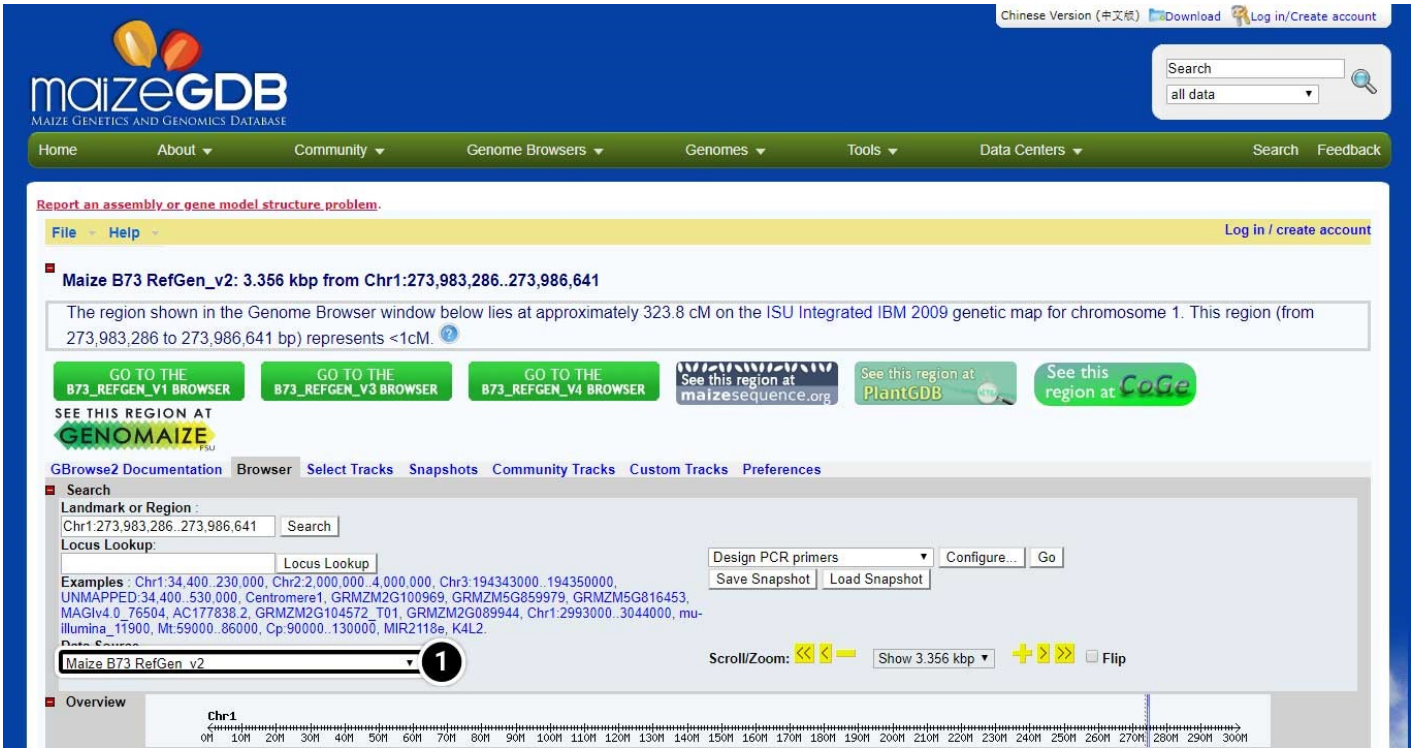
2 View at MaizeGDB

```
Score =  465 bits (859.812),  Expect = 0
Identities = 465/465 (1.0000%), Gaps = 0 (0.0000%)
Strand = Plus / Minus

Query  1315       AGGAGCTGGAGGTGGAGAAGTTCATCACGCACAGCGTCCCGTTCGCCGAGATCAACAAGG  1374
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821785  AGGAGCTGGAGGTGGAGAAGTTCATCACGCACAGCGTCCCGTTCGCCGAGATCAACAAGG  278821726

Query  1375       CGTTCGACCTGATGGCCAAGGGGGAGGGCATCCGCTGCATCATCCGCATGGAGAACTAGA  1434
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821725  CGTTCGACCTGATGGCCAAGGGGGAGGGCATCCGCTGCATCATCCGCATGGAGAACTAGA  278821666

Query  1435       TTTCGCTGTCTAGTTTGTGATCTGGCCTGGGCTTGGGGTTAATAAAAGAGGCAATGCTAG  1494
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821665  TTTCGCTGTCTAGTTTGTGATCTGGCCTGGGCTTGGGGTTAATAAAAGAGGCAATGCTAG  278821606

Query  1495       CCTGCCCTTTCGATGAGGAGGTACATACACGCTGGCGATGGACCGCGCTTGTGTGTCGCG  1554
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821605  CCTGCCCTTTCGATGAGGAGGTACATACACGCTGGCGATGGACCGCGCTTGTGTGTCGCG  278821546

Query  1555       TTCAGTTTGGCTTTTGCCAAGCAGTAGGGTAGCTTCCCGTGTCGGTAATTATATGGTATG  1614
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821545  TTCAGTTTGGCTTTTGCCAAGCAGTAGGGTAGCTTCCCGTGTCGGTAATTATATGGTATG  278821486

Query  1615       AACCATCACCTTTTGGCTCTACATGGTATGAACGTAAGATACAAATTCCAACTACCTCTA  1674
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821485  AACCATCACCTTTTGGCTCTACATGGTATGAACGTAAGATACAAATTCCAACTACCTCTA  278821426

Query  1675       GCTCGCTTGTGTGGTATCTGTATCAGTATTCATGTGTTTGTTTGCTTATGTGTTTGTTTG  1734
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821425  GCTCGCTTGTGTGGTATCTGTATCAGTATTCATGTGTTTGTTTGCTTATGTGTTTGTTTG  278821366

Query  1735       CTTGTATTTGCTGGTGCTTGTATCGCGGGATGCAATGAGTTGTTG  1779
                  |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  278821365  CTTGTATTTGCTGGTGCTTGTATCGCGGGATGCAATGAGTTGTTG  278821321
```

Change the Data Source

The selection you made on the last screen will open a new window containing information on the position of adh1 and data sources.

1. Click on the pull-down menu of “data source” (arrow) to explore other data sets.



BLAST Study Questions

Find the position of adh1 using BLAST results and all versions of datasets. Compare BLAST results with maize genome browse results. If the results are not similar, what might be the reason for the differences?
Position of the adh1 locus in:

B73 RefGen_v2 sequence

Maize genome browse result: Between 273,983,286 and 273,983,641

BLAST: Between [ ] and [ ]

B73 RefGen_v1 sequence

Maize genome browse result: Between 272,905,082 and 272,905,552

BLAST: Between [ ] and [ ]

BAC based genome assembly

Maize genome browse result: Between 262,399,900 and 262,556,700

BLAST: Since the introduction of the pseudomolecule, the BAC based genome assembly is no longer supported.

Check

Possible Explanation for BLAST Study Question Results

One reason for discrepancies might be that there are in this genomic region several copies of the gene (eventually ancient duplication no longer actively transcribed due to mutations or whatever). Depending on the origin of your query sequence you use to find the gene, they might show different hit scores from these versions of the gene. As for the version2 pseudo- molecule the location seems to be quite similar…

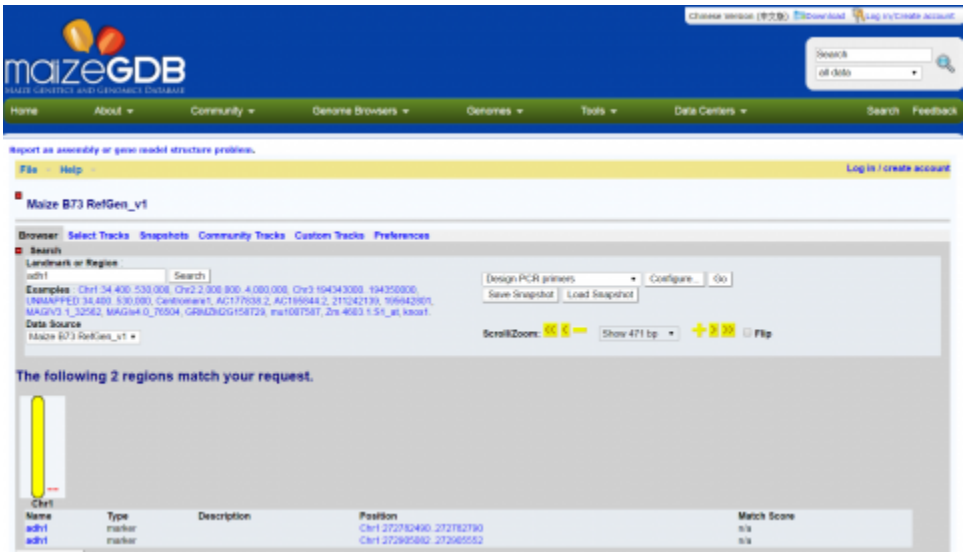


Fig. 1 Screenshot of the BLAST search output page.

Multiple Sequence Alignment

Some of the key steps in building a multiple alignment include:

1. Obtaining the sequences to align by database searching
2. Running the multiple alignment program and,
3. Identifying the residues that differ or are conserved among the sequences (finding polymorphisms)

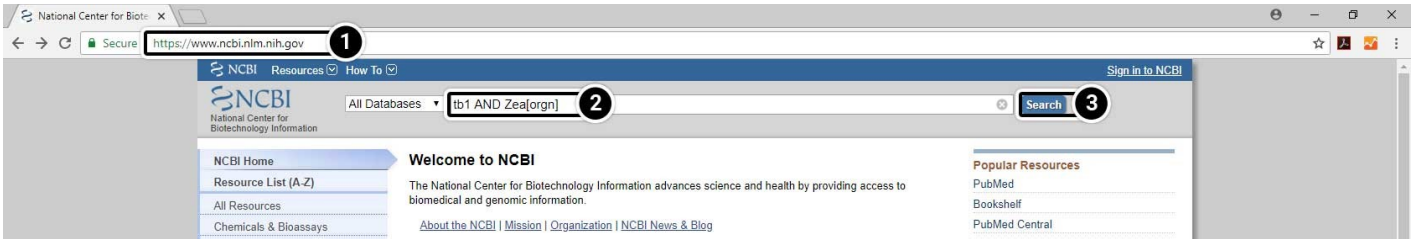
Enter the NCBI site and use the following steps to guide your activity.

Try This: Multiple Sequence Alignment

Search the NCBI Website for the Allelic Sequences

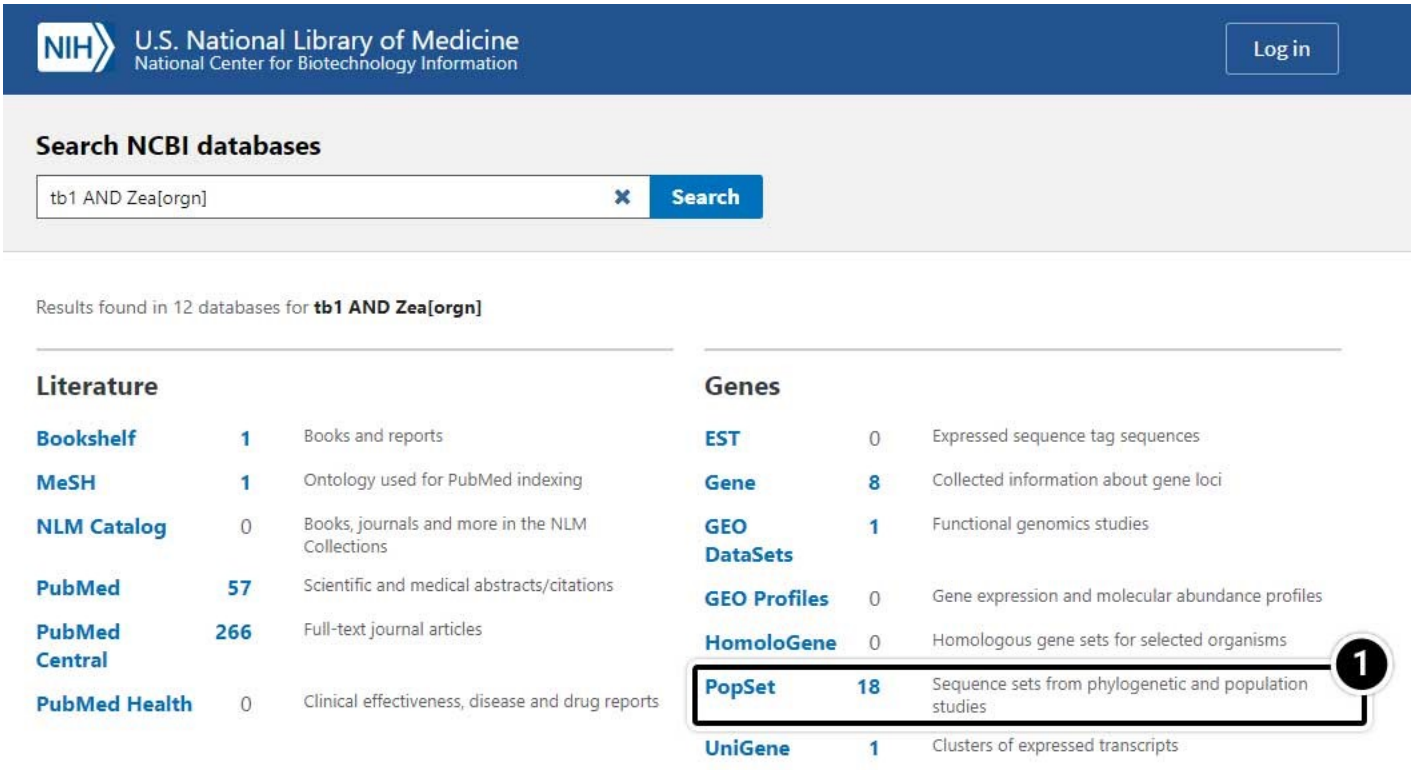
Find the allelic sequences for a maize gene. Here we will use teosinte branched1 (tb1) gene from maize as an example.

1. Open your web browser and go to https://www.ncbi.nlm.nih.gov/.
2. Enter **tb1 AND Zea[orgn]** into the search box.
3. Press **Enter** or click **Search**.



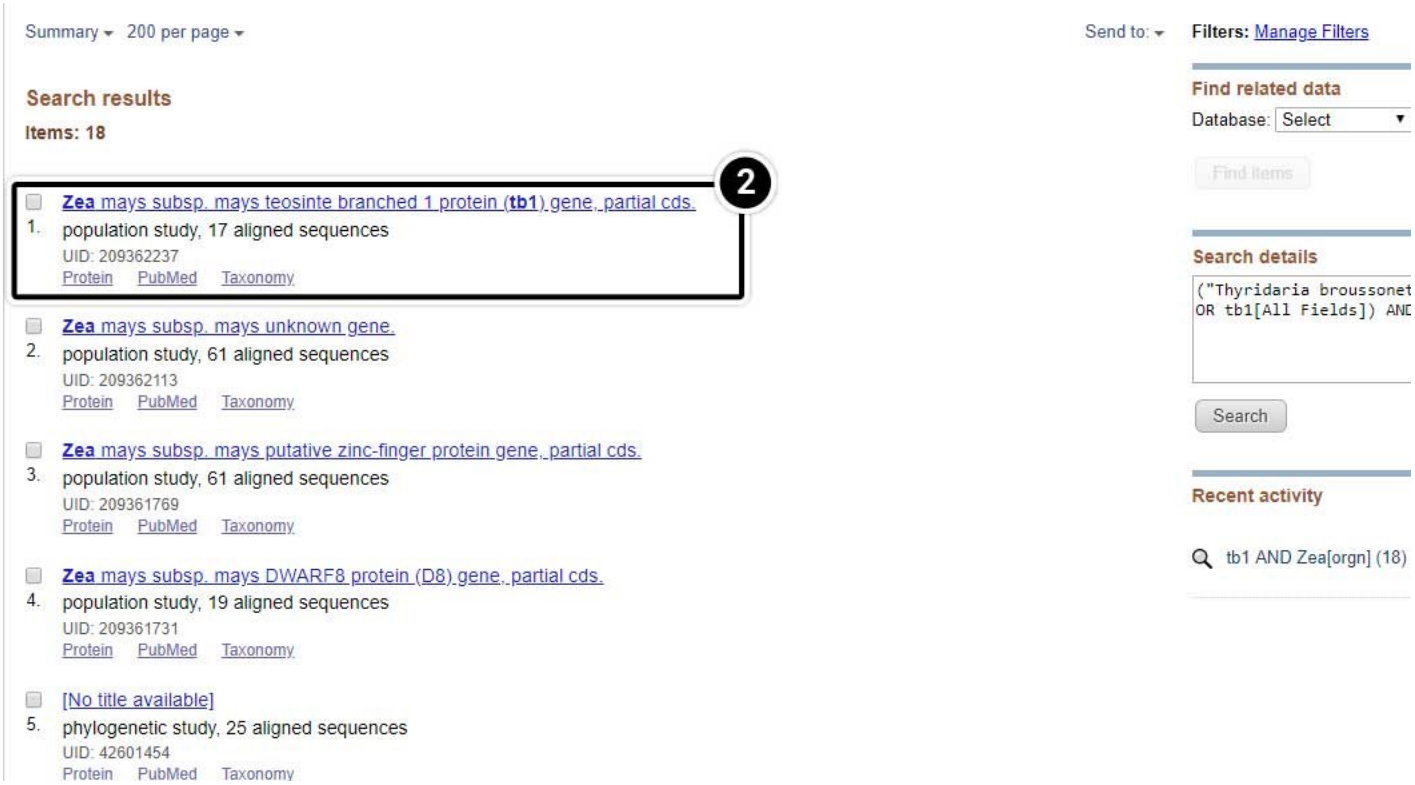
Narrow the Search Results

1. Select PopSet (population data sets).



Choose the Specific Search Result

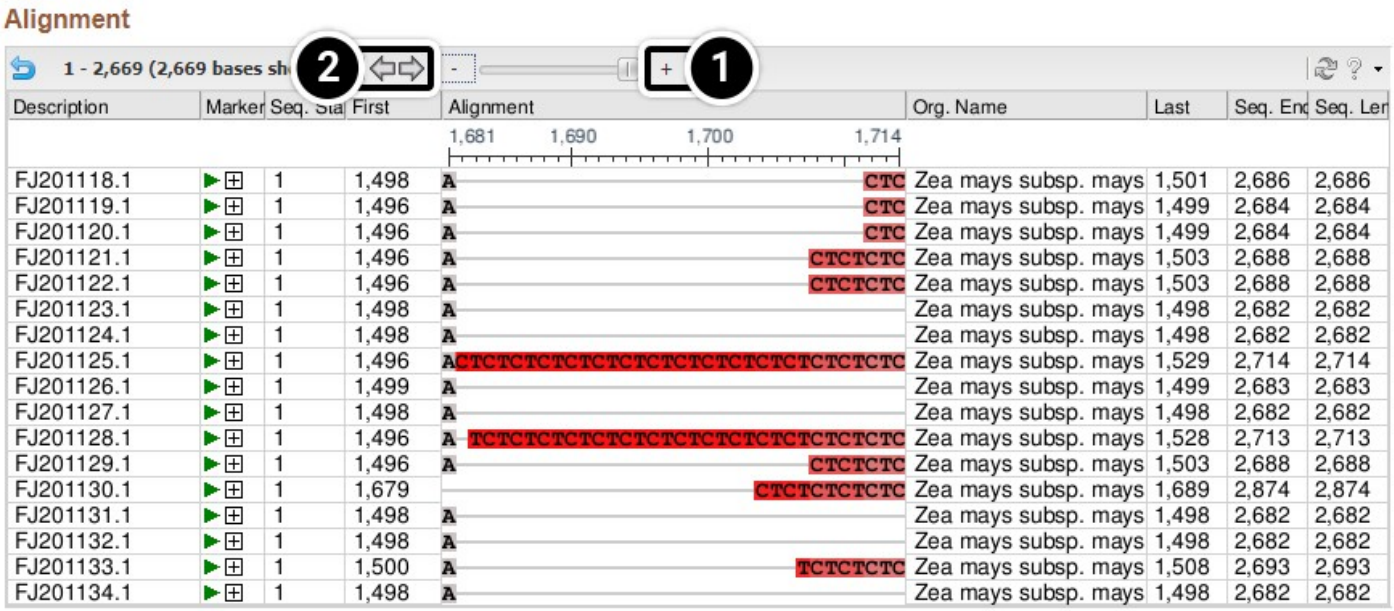
2. Select the result containing 17 aligned sequences of tb1 partial cds from a population study. (UID 209362237)



Explore the Alignment

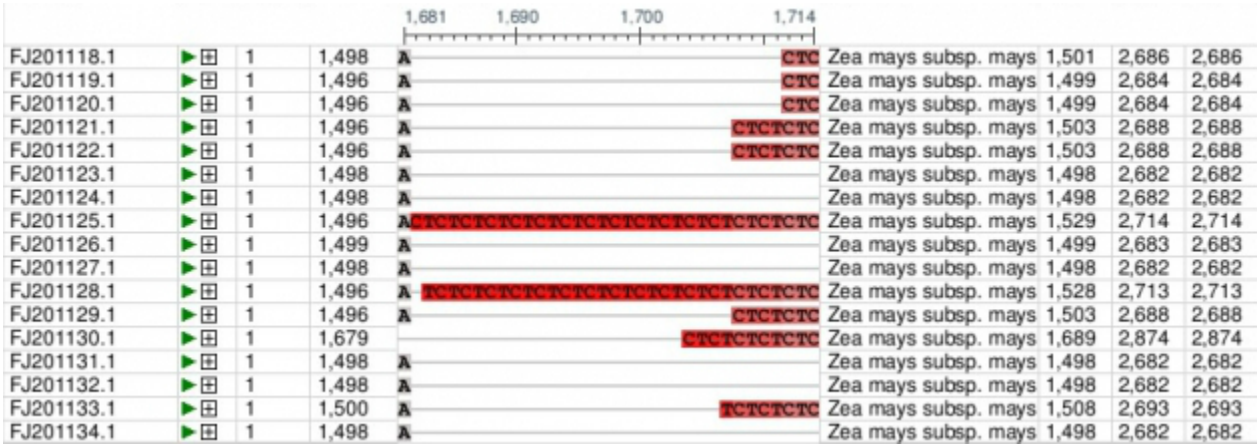
Scroll down to see the alignment of the 17 tb1 partial cds.

1. Click the + sign until you can see the nucleotides.
2. Click the arrow to pan right in the sequences until you can see the region between 1500 and 1530.



Multiple Sequence Alignment Study Questions

Review your output from this activity.



FJ201118.1	1	1,498	A CTC	Zea mays subsp. mays	1,501	2,686	2,686
FJ201119.1	1	1,496	A CTC	Zea mays subsp. mays	1,499	2,684	2,684
FJ201120.1	1	1,496	A CTC	Zea mays subsp. mays	1,499	2,684	2,684
FJ201121.1	1	1,496	A CTCTCTC	Zea mays subsp. mays	1,503	2,688	2,688
FJ201122.1	1	1,496	A CTCTCTC	Zea mays subsp. mays	1,503	2,688	2,688
FJ201123.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682
FJ201124.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682
FJ201125.1	1	1,496	ACTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC	Zea mays subsp. mays	1,529	2,714	2,714
FJ201126.1	1	1,499	A	Zea mays subsp. mays	1,499	2,683	2,683
FJ201127.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682
FJ201128.1	1	1,496	A TCTCTCTCTCTCTCTCTCTCTCTCTCTCTC	Zea mays subsp. mays	1,528	2,713	2,713
FJ201129.1	1	1,496	A CTCTCTC	Zea mays subsp. mays	1,503	2,688	2,688
FJ201130.1	1	1,679	CTCTCTCTCTC	Zea mays subsp. mays	1,689	2,874	2,874
FJ201131.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682
FJ201132.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682
FJ201133.1	1	1,500	A TCTCTCTC	Zea mays subsp. mays	1,508	2,693	2,693
FJ201134.1	1	1,498	A	Zea mays subsp. mays	1,498	2,682	2,682

Is the sequence present between position 1,500 and 1,530 suitable as a marker?

- ○ No
- ○ Yes

Check

What type of polymorphism is this?

Check

What detection strategy would you use to detect the polymorphism?

Check

Finding Polymorphisms

Using Clustal Omega

To detect polymorphisms in a set of candidate genes requires a program that aligns multiple sequences. Clustal Omega is one of the commonly used programs. Clustal Omega is a hierarchical multiple alignment program that combines a robust method for multiple sequence alignment with a user-friendly interface. There are different webservers that provide access to Clustal Omega. For this lesson we will use the European Bioinformatics Institute webserver. Clustal Omega can also be downloaded to a personal computer for more routine use. The following is an example of how to use Clustal Omega to detect polymorphisms.

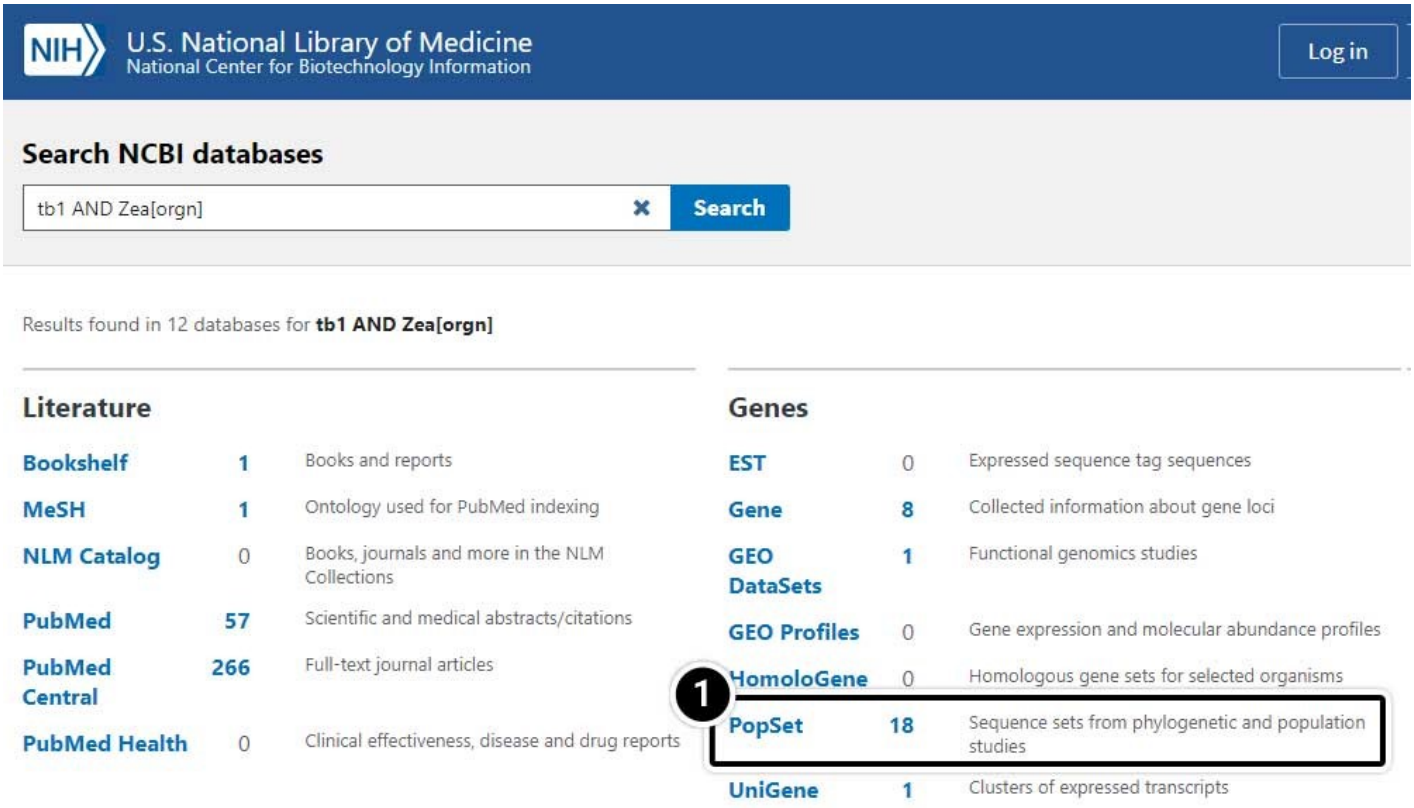
Try This: Using Clustal Omega

Search the NCBI Website

1. Go to [the NCBI website] and search for **tb1 AND Zea[orgn]**.
2. Click **Search**

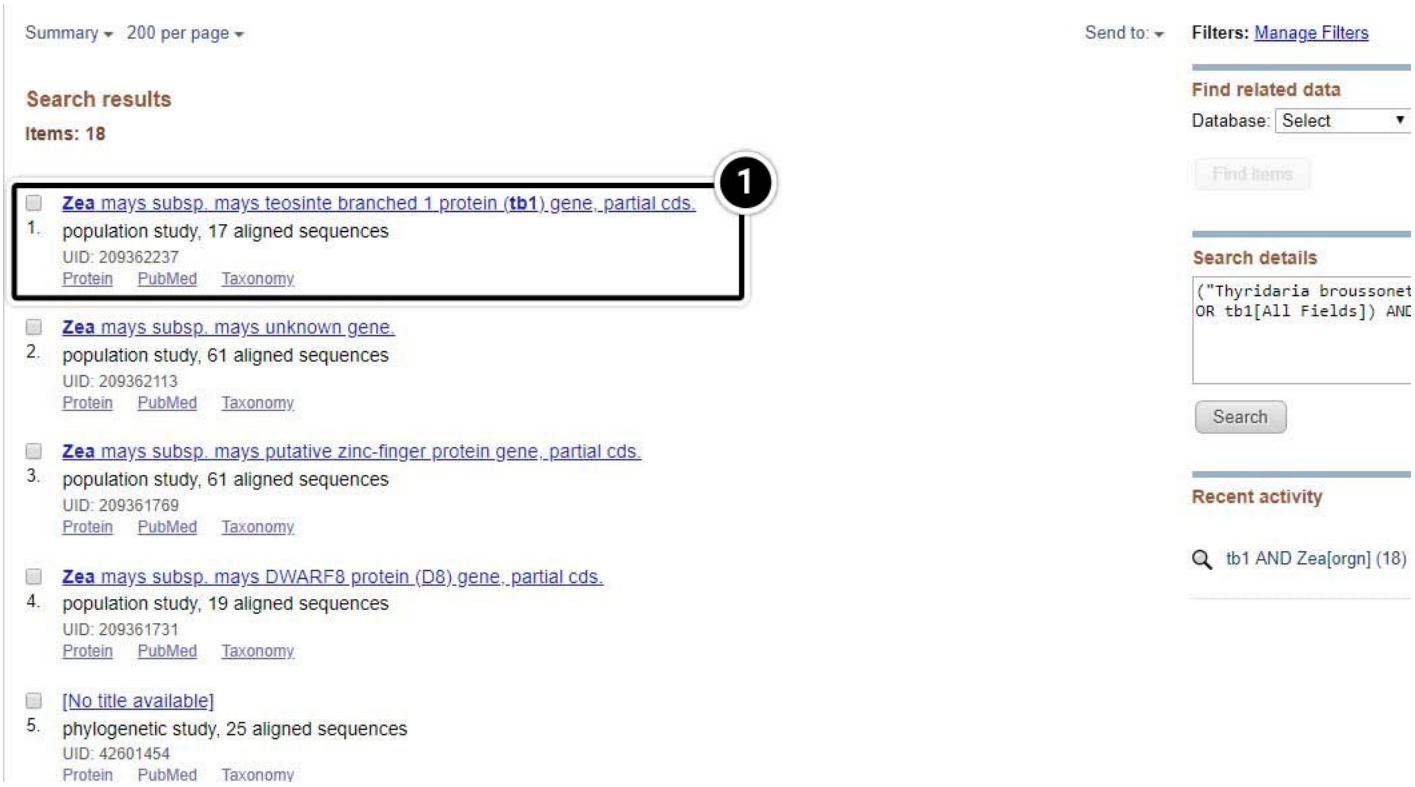
Explore the Search Results

1. Select PopSet (population data sets).



Select the Population Set

1. Click on the population set we studied earlier (UID 209362237)

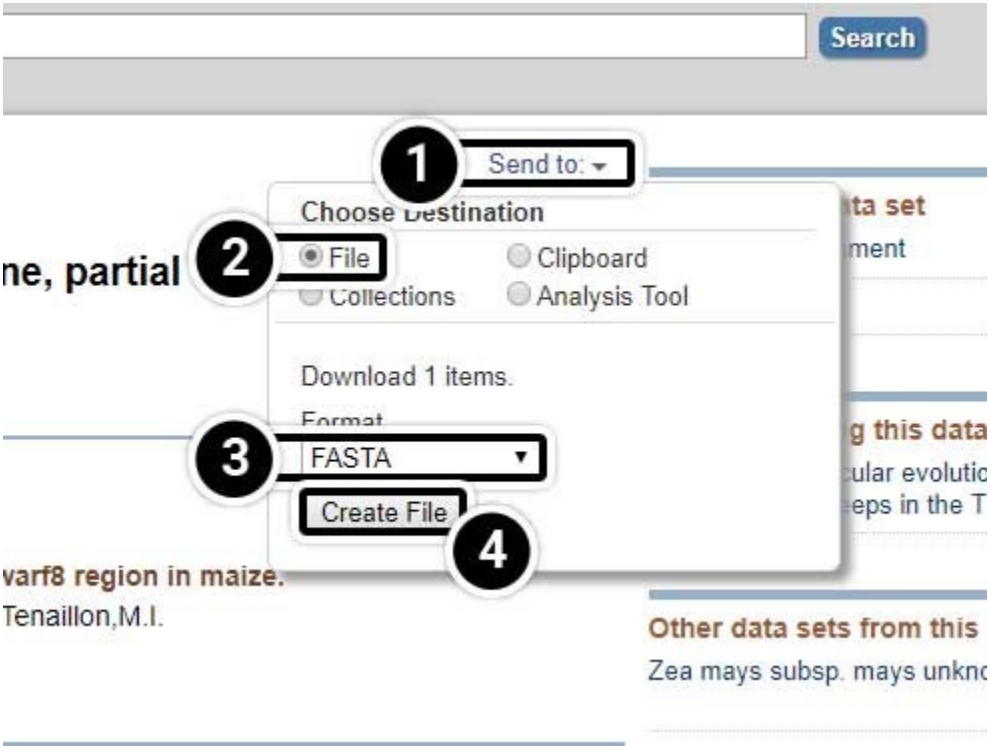


Add new comment

Create a FASTA File

Create a FASTA file of the 17 tb1 sequences.

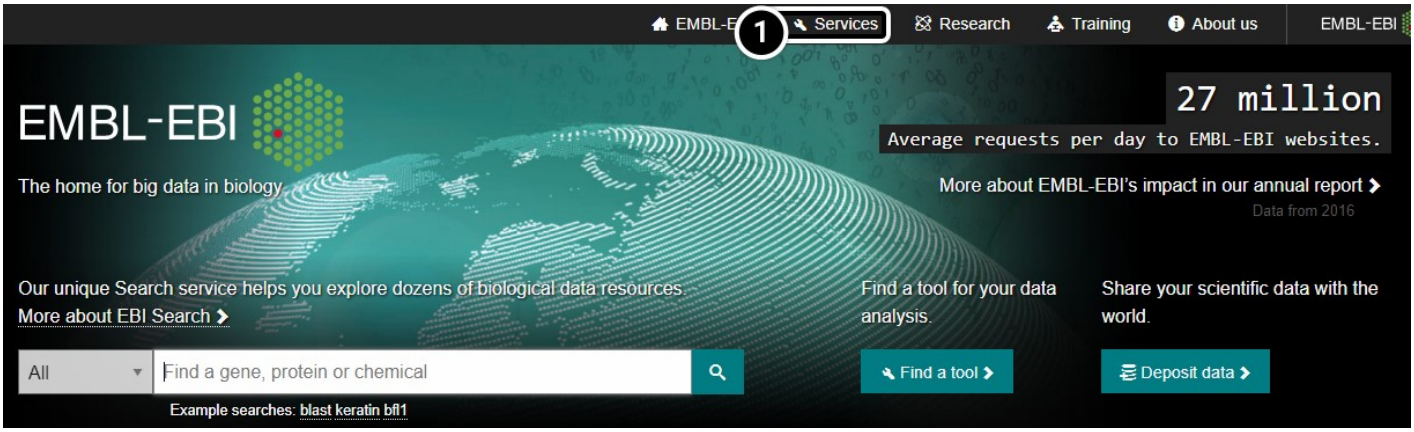
1. Click the pull-down menu **Send to**: at the top right of the screen
2. In the menu that appears, select **File** for the destination
3. Select the **FASTA** format
4. Finally, click **Create File**.

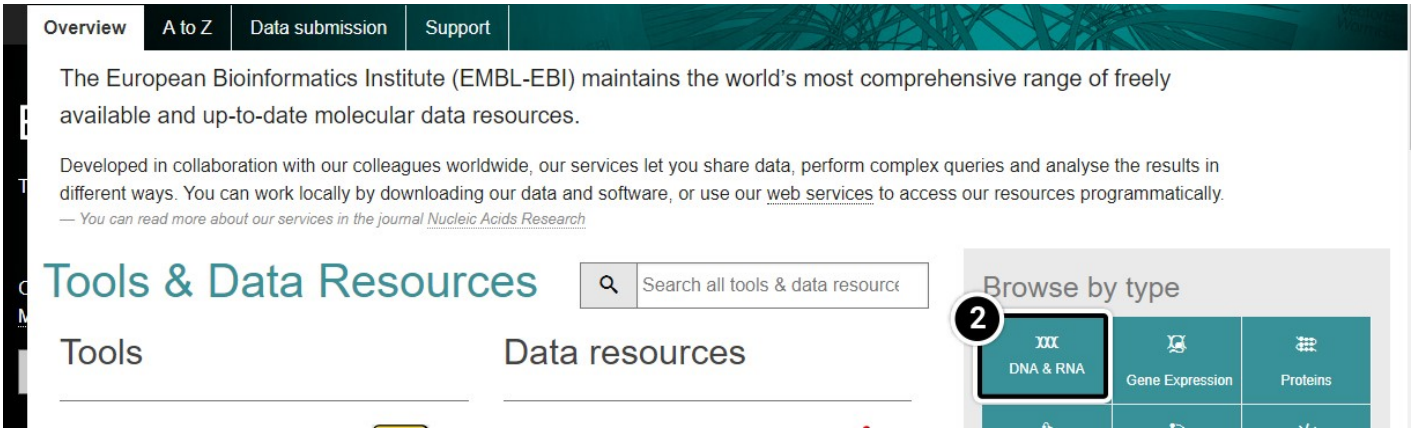


Access Clustal Omega

Access the Clustal Omega program through **EMBL-EBI**.

1. Click the Services link
2. Under Browse by type, click **DNA & RNA**

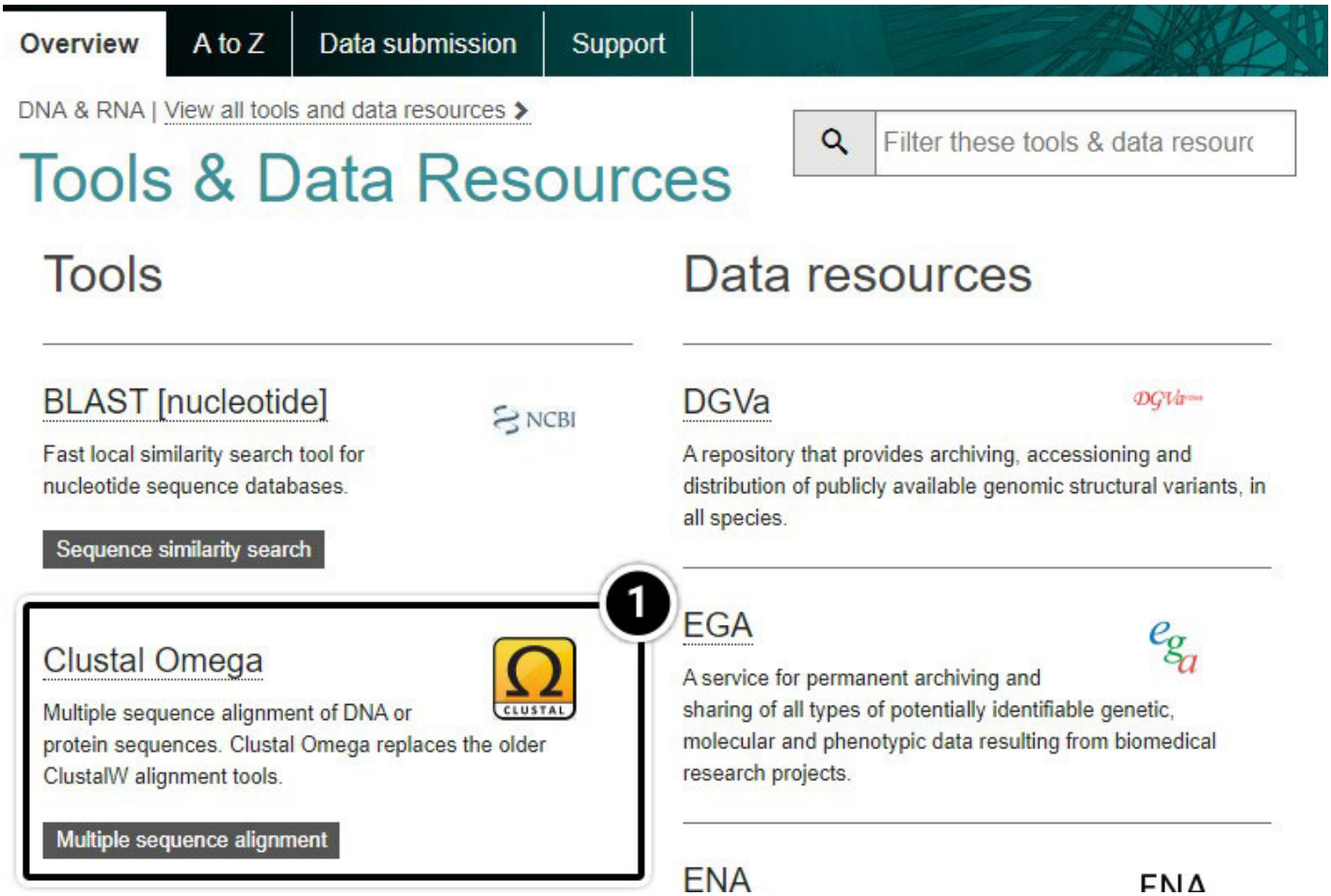




Perform Alignment

Perform alignment of tb1 partial cds using Clustal Omega. Within the Clustal Omega window you have the option of pasting sequences, or uploading files containing your sequences in FASTA format. We will upload the FASTA file you created in Step 2. As you may notice in this window, the default is set as "PROTEIN." Since you wish to align tb1 DNA sequences, you must change this parameter. Upload your file and click **Submit**.

1. Click **Clustal Omega**
2. Select **DNA** from the dropdown
3. Click **Choose File** to browse for the file you created.
4. Click **Submit**



STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

PROTEIN

DNA 2

RNA

Or, upload a file: Choose File 3 chosen

See example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

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

STEP 3 - Submit your job

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

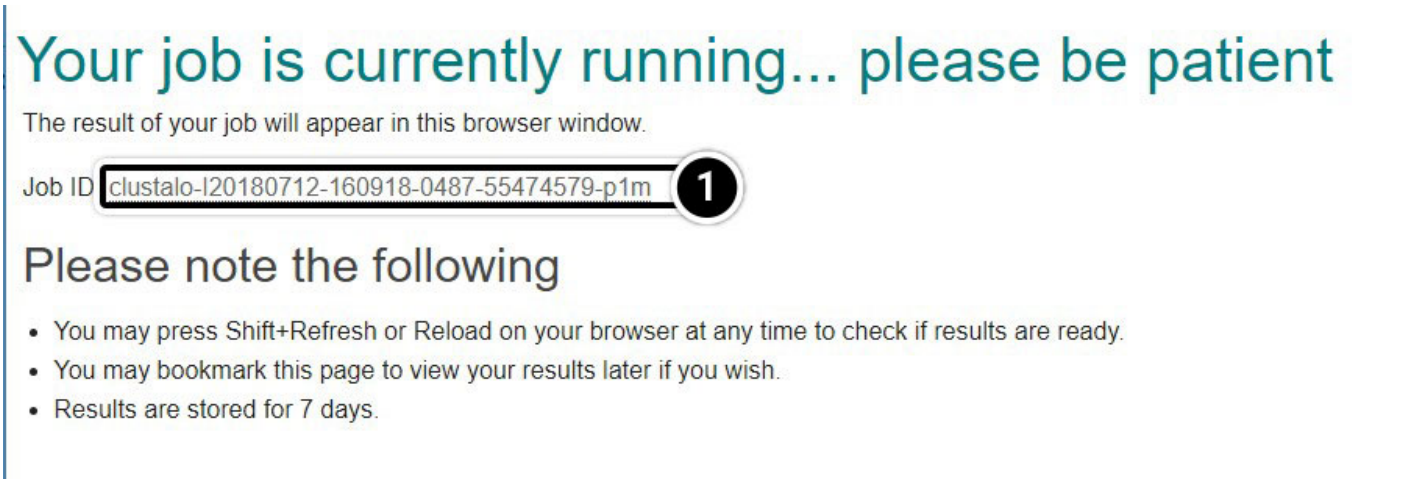
4

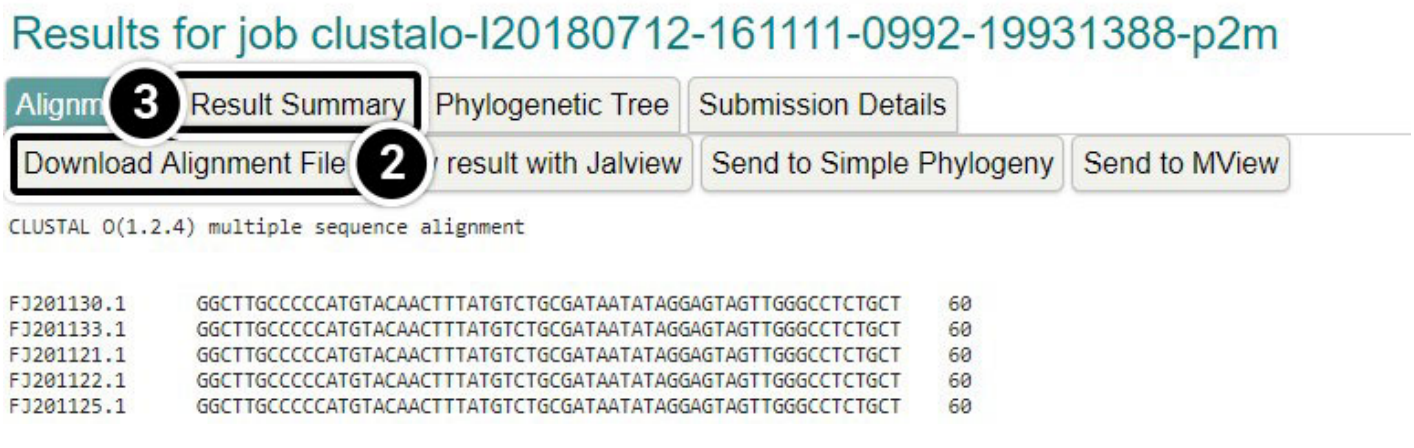
Submit

Explore the Output

It will take a moment before you obtain a report of your job request. You can click and save the "Your Job Output" URL to view your results for up to seven days.

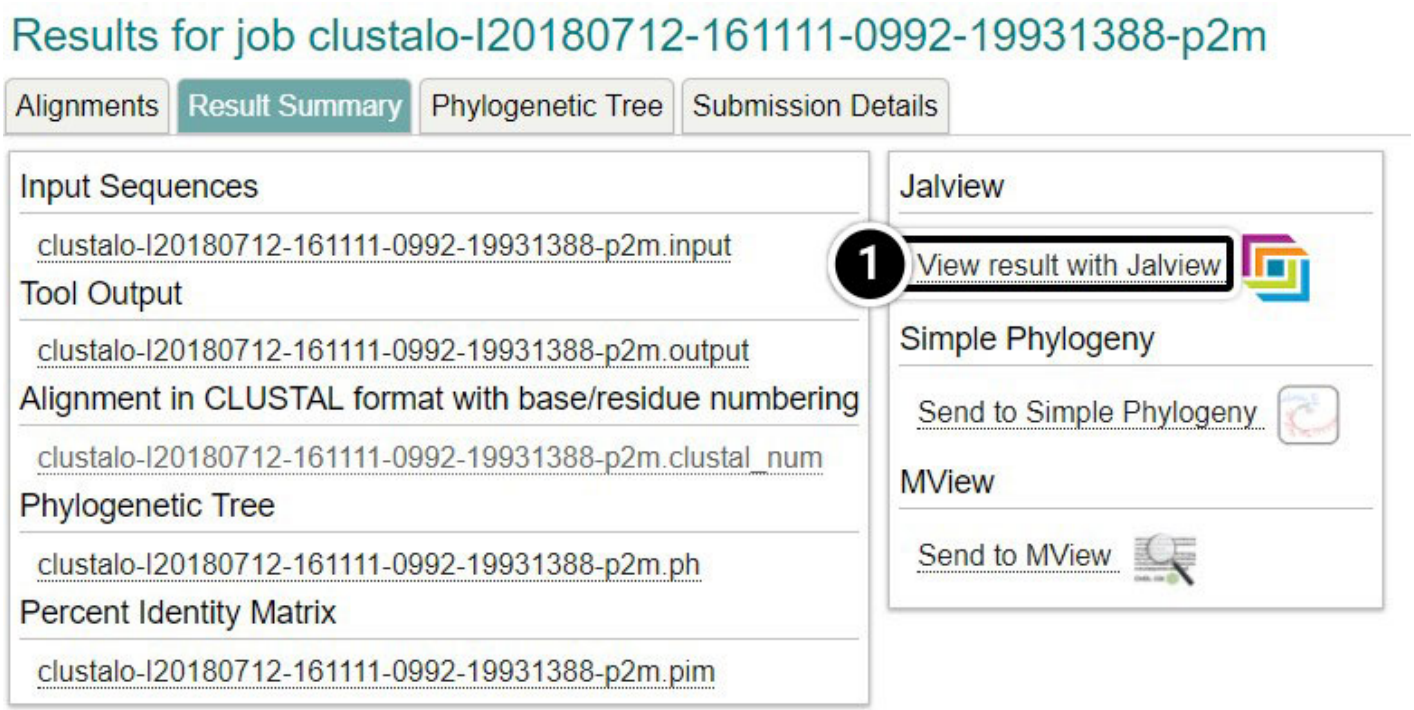
1. Click the **Job ID** link
2. You can click the **Download Alignment File** but that is not necessary for this activity
3. Click **Result Summary**

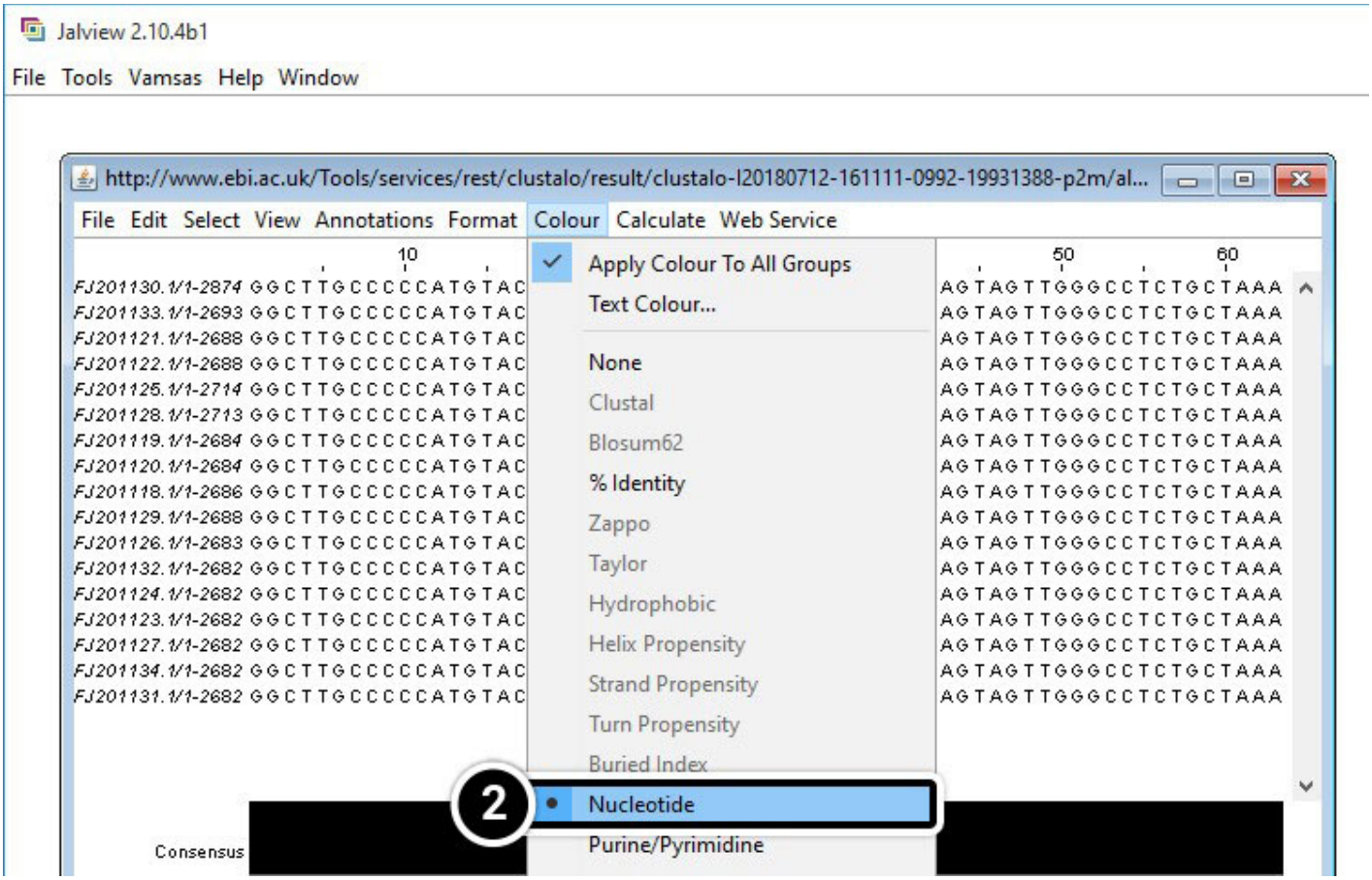


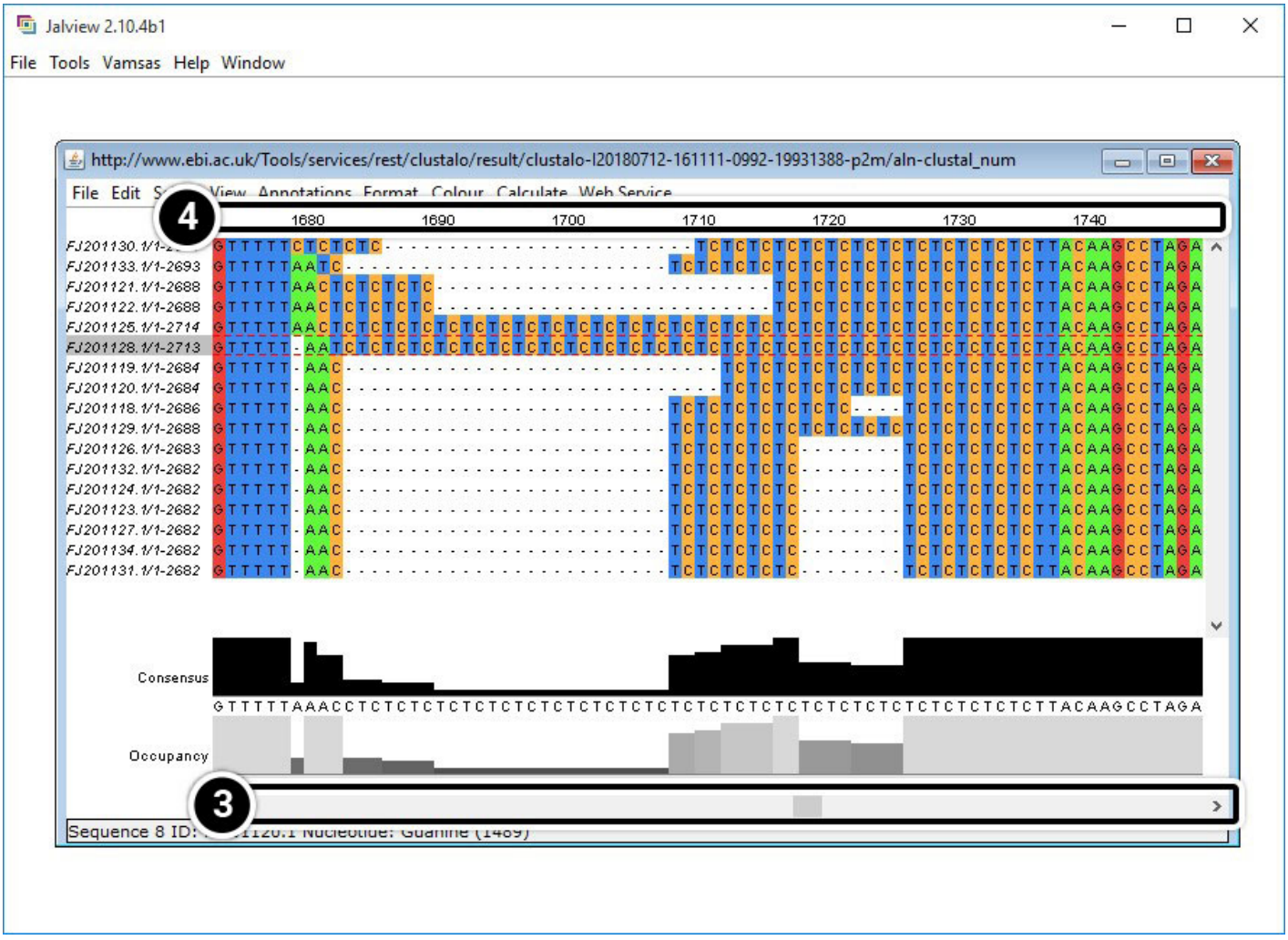


View Result with Jalview

1. Click **View result with Jalview**
2. Once Jalview opens, click **Colour** then **Nucleotide**
3. Use the scroll bar to navigate to the alignment.
4. Scroll to align the region from 1680 to 1740.



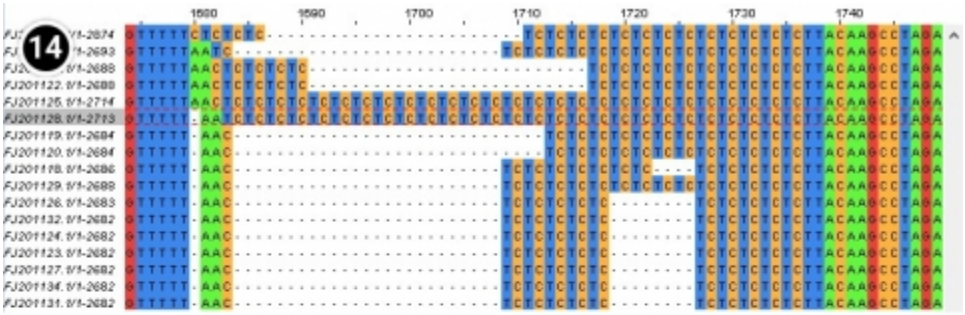

Jalview 2.10.4b1
File Tools Vamsas Help Window
http://www.ebi.ac.uk/Tools/services/rest/clustalo/result/clustalo-I20180712-161111-0992-19931388-p2m/al...
File Edit Select View Annotations Format Colour Calculate Web Service
10
50
60
FJ201130.1/1-2874 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201133.1/1-2693 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201121.1/1-2688 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201122.1/1-2688 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201125.1/1-2714 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201128.1/1-2713 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201119.1/1-2684 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201120.1/1-2684 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201118.1/1-2686 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201129.1/1-2688 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201126.1/1-2683 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201132.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201124.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201123.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201127.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201134.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
FJ201131.1/1-2682 GGCTTGCCCCCATGTAC AGTAGTTGGGCCTCTGCTAAA
Apply Colour To All Groups
Text Colour...
None
Clustal
Blosum62
% Identity
Zappo
Taylor
Hydrophobic
Helix Propensity
Strand Propensity
Turn Propensity
Buried Index
Nucleotide
Purine/Pyrimidine
2
Consensus

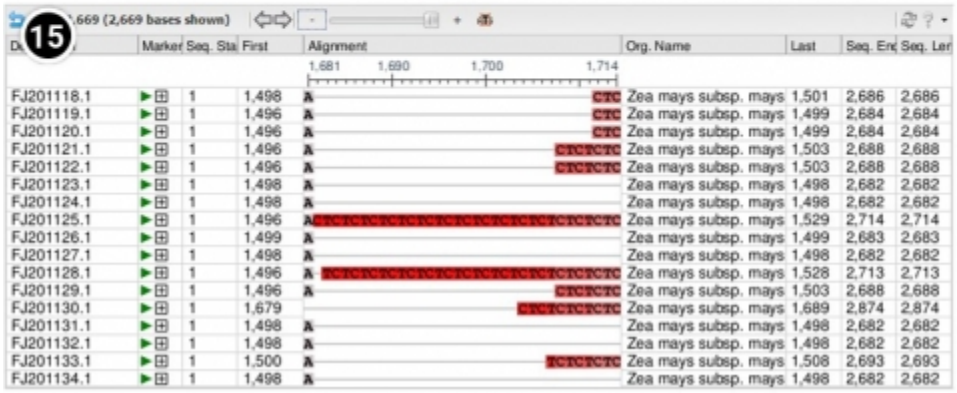

Jalview 2.10.4b1
File Tools Vamsas Help Window
http://www.ebi.ac.uk/Tools/services/rest/clustalo/result/clustalo-I20180712-161111-0992-19931388-p2m/aln-clustal_num
File Edit
View Annotations Format Colour Calculate Web Service
4
1680
1690
1700
1710
1720
1730
1740
FJ201130.1/1-
FJ201133.1/1-2693
FJ201121.1/1-2688
FJ201122.1/1-2688
FJ201125.1/1-2714
FJ201128.1/1-2713
FJ201119.1/1-2684
FJ201120.1/1-2684
FJ201118.1/1-2686
FJ201129.1/1-2688
FJ201126.1/1-2683
FJ201132.1/1-2682
FJ201124.1/1-2682
FJ201123.1/1-2682
FJ201127.1/1-2682
FJ201134.1/1-2682
FJ201131.1/1-2682
Consensus
GTTTTTAAACCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTTACAAGCCTAGA
Occupancy
3
Sequence 8 ID:
Nucleotide: Guanine (1469)

Compare Results from JalView and NCBI BLAST

Analyze region 1680 to 1740 of your JalView results (below). What is unique about this region? How does it compare with the region between positions 1500 and 1530 in the NCBI BLAST?

14. JalView
15. NCBI BLAST





Developing Marker Assays

Recall in Module 2 you learned how SSR and SNP can be analyzed by PCR and restriction enzymes. In lesson 8 of this course, you will learn additional strategies to detect DNA polymorphisms for marker development.

Summary

Biological sequence databases serve an important role of providing access to sequence information to the research community. Searches can be restricted to a single database or expanded to include all other databases. Whole genomes can be explored to predict positions that match a specific sequence. To detect polymorphisms in a set of candidate genes a program that aligns multiple sequences is required. The detected polymorphisms can be used to develop markers to assist in selection.

Reflection

The **Module Reflection** appears as the last "task" in each module. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the module and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this module.
2. What is the most valuable concept that you learned from the module? Why is this concept valuable to you?
3. What concepts in the module are still unclear/the least clear to you?

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Molecular Plant Breeding Introduction to Bioinformatics Author: Ursula Frei, Walter Suza, Thomas Lübberstedt, and Madan Bhattacharyya (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Frei, U., W. Suza, T. Lübberstedt, and M. Bhattacharyya. 2016. Introduction to Bioinformatics. *In* Molecular Plant Breeding, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.