NCBI

Data retrieval using GenBank

CENTURION UNIVERSITY

Shaping Lives...
Empowering Communities!

What is NCBI?

On November 4, 1988 that President Ronald Reagan signed the Health Omnibus Extension Act to create The National Center for Biotechnology Information as part of National Library of Medicine at NIH.

- Create automated systems for knowledge about **molecular biology**, biochemistry, and genetics.
- Perform research into advanced methods of analyzing and interpreting **molecular biology** data.
- Enable **biotechnology** researchers and medical care personnel to use the systems and methods developed.

CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!

GenBank Sequence Format

- To search GenBank effectively using the text-based method requires an understanding of the GenBank sequence format.
- GenBank is a relational database. However, the search output for sequence files is produced as flat files for easy reading.
- The resulting flat files contain three sections; Header, Features, and Sequence entry.
- There are many fields in the Header and Features sections. Each field has an unique identifier for easy indexing by computer software.
- Understanding the structure of the GenBank files helps in designing effective search strategies.

CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!

Figure 2.3: NCBI GenBank/GenPept format showing the three major components of a sequence file.

AIM OF THE EXPERIMENT

Data retrieval using GenBank, NCBI.

Procedure

GOOGLE NCBI

↓

NCBI HOME PAGE

↓

IN POP-UP GO TO
NUCLEOTIDE

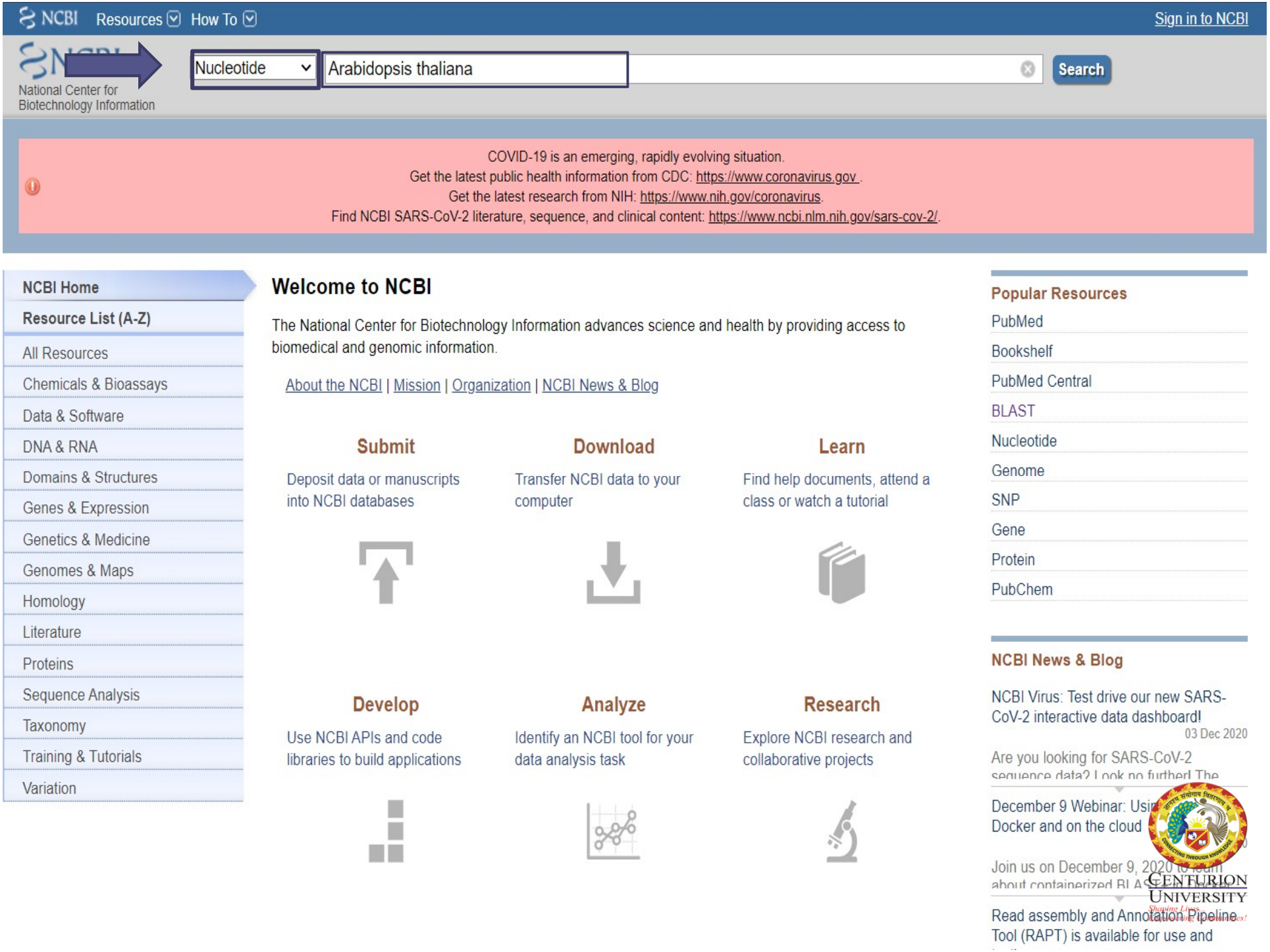
↓

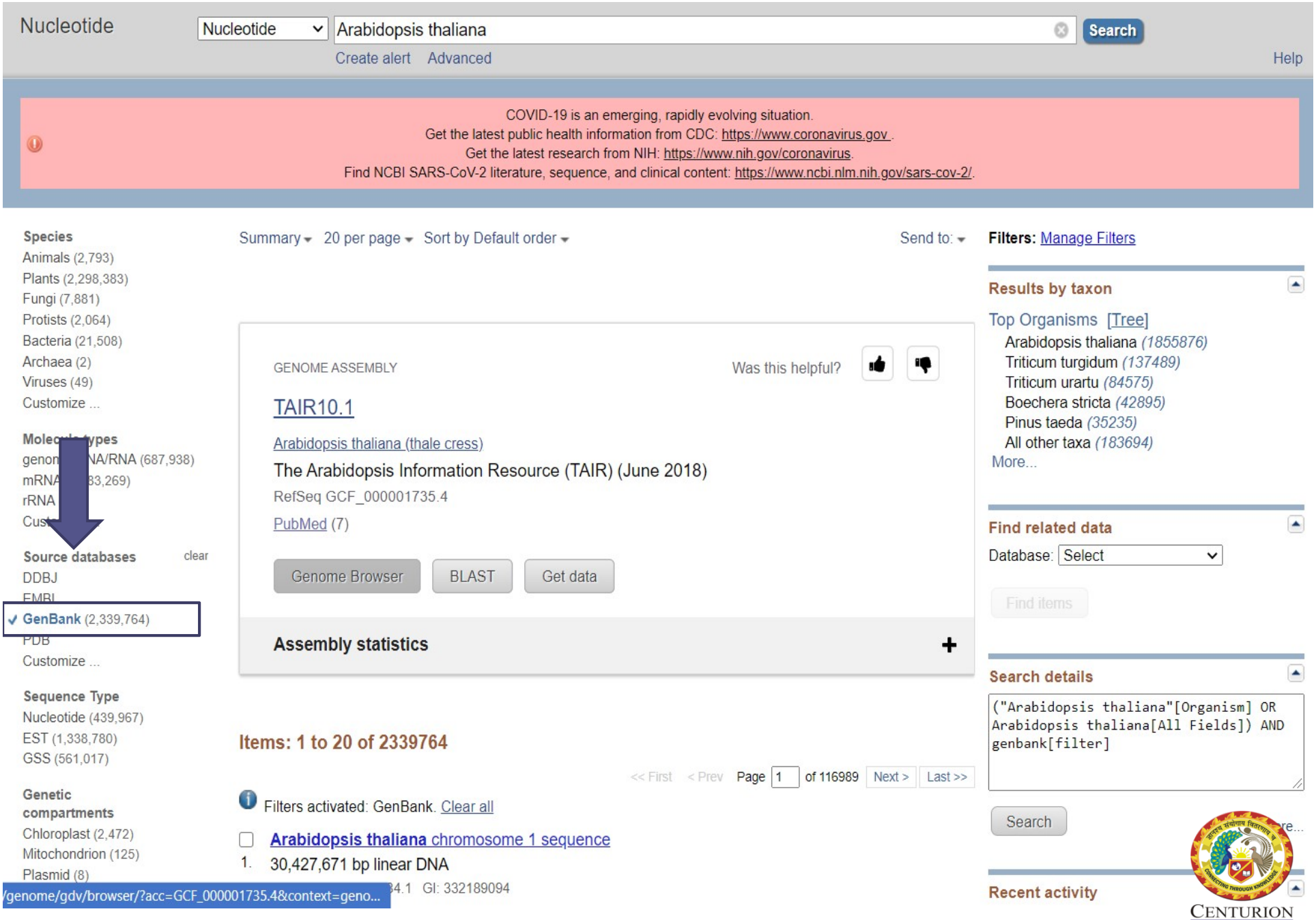
ENTER THE NAME OF THE ORGANISM

↓

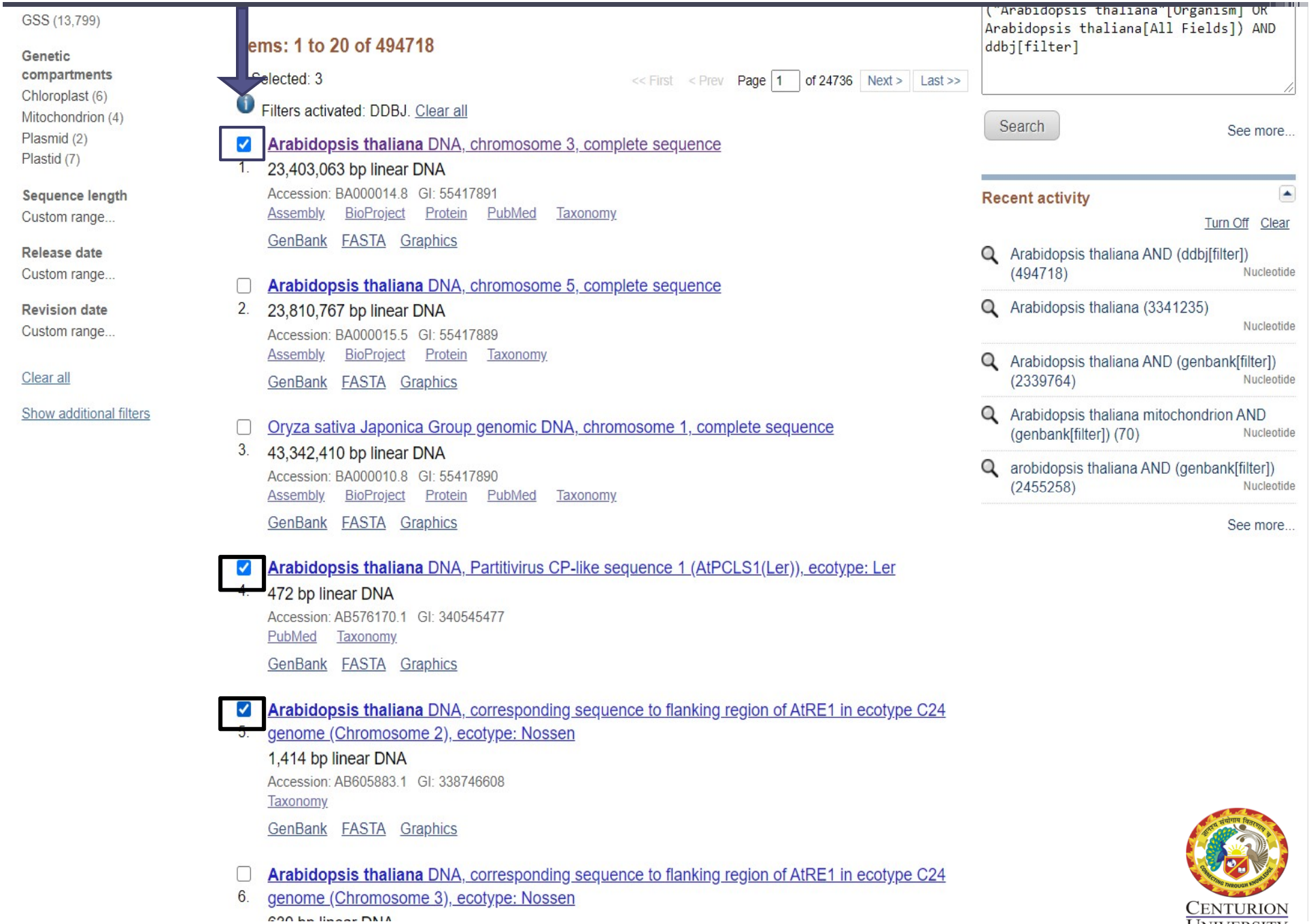
DOWNLOAD THE REQUIRED
SEQUENCES
CHOSING DESIRED FORMAT

CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!


NCBI Resources How To
Sign in to NCBI
National Center for Biotechnology Information
Nucleotide
Arabidopsis thaliana
Search
COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.
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 | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
NCBI Virus: Test drive our new SARS-CoV-2 interactive data dashboard!
03 Dec 2020
Are you looking for SARS-CoV-2
December 9 Webinar: Usi
Docker and on the cloud
Join us on December 9, 2020 to learn
Read assembly and Annotation Pipeline
Tool (RAPT) is available for use and

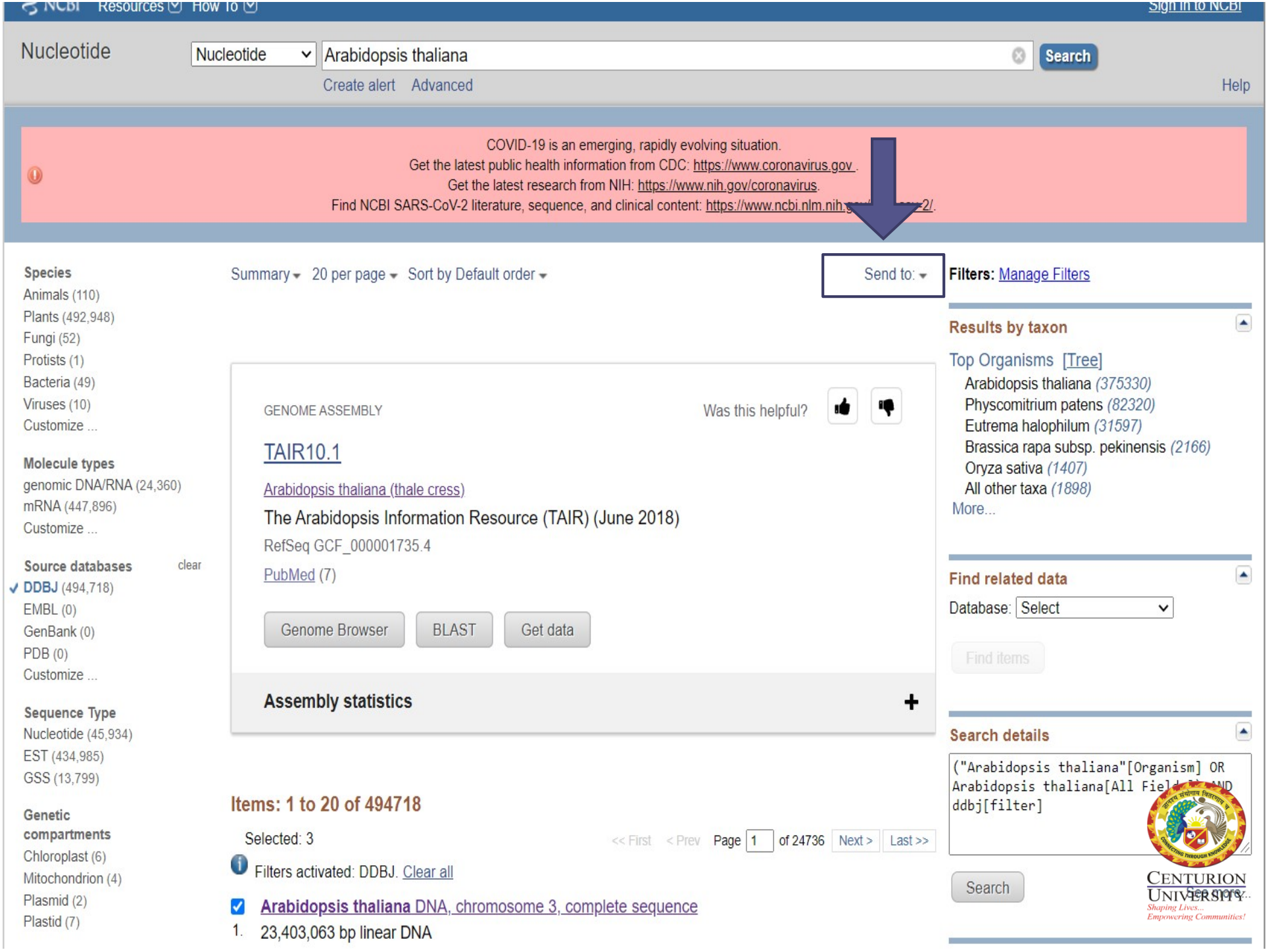


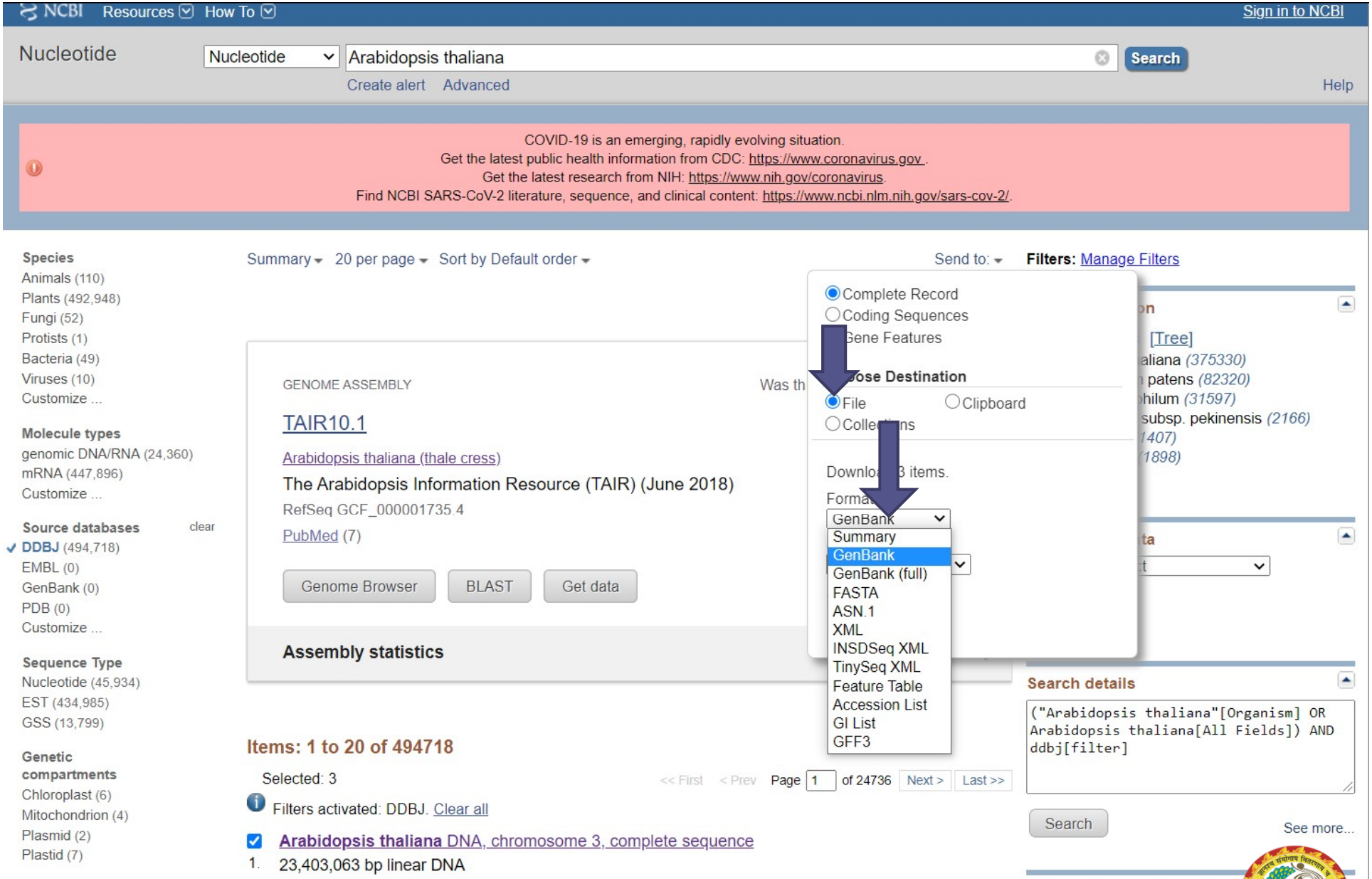
Here in source databases we can filter by choosing GenBank,for retrieving genbank data only.



Mark the desired sequence which you want to download.

CENTURION UNIVERSITY
Shaping Lives... Empowering Communities!


Resources
How To
Sign in to NCBI
Nucleotide
Nucleotide
Arabidopsis thaliana
Search
Create alert
Advanced
Help
COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.g
Species
Animals (110)
Plants (492,948)
Fungi (52)
Protists (1)
Bacteria (49)
Viruses (10)
Customize ...
Molecule types
genomic DNA/RNA (24,360)
mRNA (447,896)
Customize ...
Source databases
clear
DDBJ (494,718)
EMBL (0)
GenBank (0)
PDB (0)
Customize ...
Sequence Type
Nucleotide (45,934)
EST (434,985)
GSS (13,799)
Genetic compartments
Chloroplast (6)
Mitochondrion (4)
Plasmid (2)
Plastid (7)
Summary
20 per page
Sort by Default order
Send to:
Filters: Manage Filters
GENOME ASSEMBLY
Was this helpful?
TAIR10.1
Arabidopsis thaliana (thale cress)
The Arabidopsis Information Resource (TAIR) (June 2018)
RefSeq GCF_000001735.4
PubMed (7)
Genome Browser
BLAST
Get data
Assembly statistics
Items: 1 to 20 of 494718
Selected: 3
<< First
< Prev
Page 1 of 24736
Next >
Last >>
Filters activated: DDBJ. Clear all
Arabidopsis thaliana DNA, chromosome 3, complete sequence
1. 23,403,063 bp linear DNA
Results by taxon
Top Organisms [Tree]
Arabidopsis thaliana (375330)
Physcomitrium patens (82320)
Eutrema halophilum (31597)
Brassica rapa subsp. pekinensis (2166)
Oryza sativa (1407)
All other taxa (1898)
More...
Find related data
Database: Select
Find items
Search details
("Arabidopsis thaliana"[Organism] OR
Arabidopsis thaliana[All Field
AND
ddbj[filter]
Search
See more...
CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!



- In choose destination choose **file**
- In the format section chose any format you want to- **here I am choosing genbank for getting all the details.**
- To retrieve only the sequences we can download in **fasta format.**

CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!

Results- in GenBank format

sequence (2).gb - Notepad

File Edit Format View Help

```
LOCUS       BA000014            23403063 bp    DNA     linear   CON 19-MAY-2007
DEFINITION  Arabidopsis thaliana DNA, chromosome 3, complete sequence.
ACCESSION   BA000014
VERSION     BA000014.8
DBLINK      BioProject: PRJNA13190
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1
  AUTHORS   European Union Chromosome 3 Arabidopsis Genome Sequencing
            Consortium, The Institute for Genomic Research and Kazusa DNA
            Research Institute.
  TITLE     Sequence and analysis of chromosome 3 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 408, 820-822 (2000)
COMMENT     On Nov 5, 2004 this sequence version replaced BA000014.7.
            Constructed (26-Dec-2000) by DDBJ.
FEATURES             Location/Qualifiers
     source          1..23403063
                     /organism="Arabidopsis thaliana"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:3702"
                     /chromosome="3"
                     /ecotype="Columbia"
CONTIG      join(AC067753.4:1..4580,AC008261.5:1145..93735,
            AC010676.6:16803..58448,AC010870.6:41653..86324,
            AC009325.8:5472..105543,AC010797.4:28249..89154,
            AC011664.10:39082..103960,AC009755.7:28826..94369,
            AC068900.3:5412..22087,AC021640.7:7198..103904,
            AC018363.8:2513..100559,AC012328.6:2096..96540,
            AC009895.4:2933..82289,AC009327.8:28249..89934,
            AC009540.6:10210..101410,AC011698.6:13672..63654,
            AC016829.7:17115..110804,AC022287.9:5541..71616,
            AC011437.6:5133..95310,AC009465.7:5248..93234,
            AC009177.7:946..140064,AC009606.4:15146..91924,
            AC011620.8:14598..100887,AC012393.5:12002..87045,
            AC013454.6:24997..62271,AC068073.1:5033..24745,
            AC018907.5:5358..93478,AC011623.5:21013..86022,
            AC020580.11:634..94704,AC036106.6:16992..51646,
```

- Then open the downloaded file on notepad.

If you download the file in Fasta format

sequence.fasta - Notepad

File Edit Format View Help

```
>AB576170.1 Arabidopsis thaliana DNA, Partitivirus CP-like sequence 1 (AtPCLS1(Ler)), ecotype: Ler
CCTCGAAAAGATCTCTTCCTTTCTCCACATGAAACACAGCATGGCTTGGTTCAATCAAGTAAAAGGAGTT
GCGGATGATGTTGCTGCATCCTTTGAAGGATCCGGCACCCTTGCTGACTGCTCTCCACATGGGTTGGTTG
CTAACCAAGTCATGGTCGTTCTCTCTACCCCGGAACATCTTCCAGAATCACCAAATTGTATCGCTGATAA
GCGTGCAACCTATGAGTTCGGTTACCAGCTCAAAAGCACAGTTCGCAACCTCCCTCCCCTTGCGGAAGCA
CTTGCCGCCTTCTCTCAGACATATATCAGGATGTTTCCTAACCATCCCTTCTTCGGGACGTTTGGGTCGA
AAACTCTTGATCATGGCCCGTTCTGGAAAATAAGGCCTATTGGCTCCAGCCTTACCGATAACAGCTCCTA
CCTCACTATCCCATCGATTGTCAAGCAAGCATTCAAAGGCCCAGCCACGTAG

>AB605883.1 Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype C24 genome (Chromosome 2), ecotype: Nossen
TTTTAGGTTTTTGCTGTTTGTCTTTTGTCTTTTTCAAGAATAGAGTCAGATCTGGCCGATCCATCCAAGG
CAGAAGCGTTTCTATCCAAATTTTCGCCTCAGACTTCTGAAATTCCACCTGAAGATCCCGATCTCGTTCT
CTCCTTCATGAGGCTCTCACTCCTTTGGGTCTCTGGATGATGGTTTTATCTCATCGCCCATTCGTGTTTA
GGTTGAGTTAAGTGAGGCTTCTAGTCTCAGATCTATCAGTTTATCATTGTTGTTTTAGACTTTGTATATT
TCGTTTGAACTCTTGGATTTTGGGTCTGGGAAGTACTTGATCAGATGTGATGTAATTTGTTAGATCGGTT
TTTTTCCCCTTTGTTTCATTTAAATATGTTTTTAGTTCCTGGAAGGTCCCATGTACTCTCCGGCTTTTAG
AAGAGTGTTTCTCTTGTCAGCTCTTGTACTCAGATCTTCTATCAATGAAATCTATATTTCCAGTGAAAAA
AAAAATAGTTATAGATAGATGTTGGTCTGAGACTCTCCGACTCATTAGACCATCCTCATCGAATAATTTT
TCATAGATGTATTTCATATAAAATAATATTAAAGGTATAGATGTCTTAAATGTTTTAGGAAGATATAATT
TTCATTTTAATGACCCTCATATTTTTTATTTGTTTTACTCTTTTATATTTAAAAATAGCCATCTAAAAAC
ACTCATAAAAATAACCAATGTGGATACTCTTAAATCAGTTTGGACATAACCAACAAAAAGGAAAGAATAA
AGAAGTAGGTGGTCGCAAAATCACTAGGATTTGATAGATTCAAGCAGAAACTAAAAGGAGCGAGATTTCT
GGGAAATATCAGAGAAGCAGACGAGGTCAAGTCCTATAAATCTGATGGAGAGTAATGTATAGAGGCATCA
ATTGTACACTAGAAAATCCCTAATATTACTTACTCATCTAGCTCTAAGAAGTAAGAACATCTCAGCGAAA
TCTTTGTAACTTTATTTCGGTATATAAAAGTCCAGTTTGATTCTTAAAATTGTGATCATGTTTGATTCGA
TGGAGGACTAGTTATGCGATCTTGCATCCACAATAGAGTAATATTGACCAGAAAGAAATTACAAACAAAG
TTACAATAGTATGACATGGAAAAATACCAAATCCAAAGGATCATCAACGTTGGAATGTAAAAATTTTGAT
TTGTGTAACTCATAGTCATAGGATGTGTTCGAGGAGTGAATAATCTCCCATGCTTAATACTAGTATATAT
CTGGATGCGAAATTCGCAATTTTTTTATTTTTGTAGCGAAATGGTAAAAAGAGATTCAAAACATTCATCC
AGCAATCACTTTAATTAATCCGAATGATTGCAAATGAAAATAGATTCTTTATTGAACACATATGTTTACA
AAACAAGAATAGTA

>AB605847.1 Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype Columbia genome, ecotype: Noss
AAGACTTCATATTATCCGTTTGGTCTCTTTGCTCTTCTTCTATGAGTGCTAAAGAAAAAAAGTTCAAATT
GTCTTCTGAGTTCCCTCAAAGCGCAGGACGTCTTTAGGGTGTATGCATCTTCCAACTGGTCACCCTTCTT
CGCCATGGCCATGTAGAACCCGAAACAAAATGCTACATGGCCTGCGTTTTGGATGCACCACTGTTTGCAC
CAGGCACCAAGCTTCTCCCTCGTGTCCA
```

Conclusion

- GenBank is the most complete collection of annotated nucleic acid sequence data for almost every organism.
- The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms.
- There is also a GenPept database for protein sequences, the majority of which are conceptual translations from DNA sequences, although a small number of the amino acid sequences are derived using peptide sequencing techniques.

Thank You

CENTURION UNIVERSITY
Shaping Lives...
Empowering Communities!