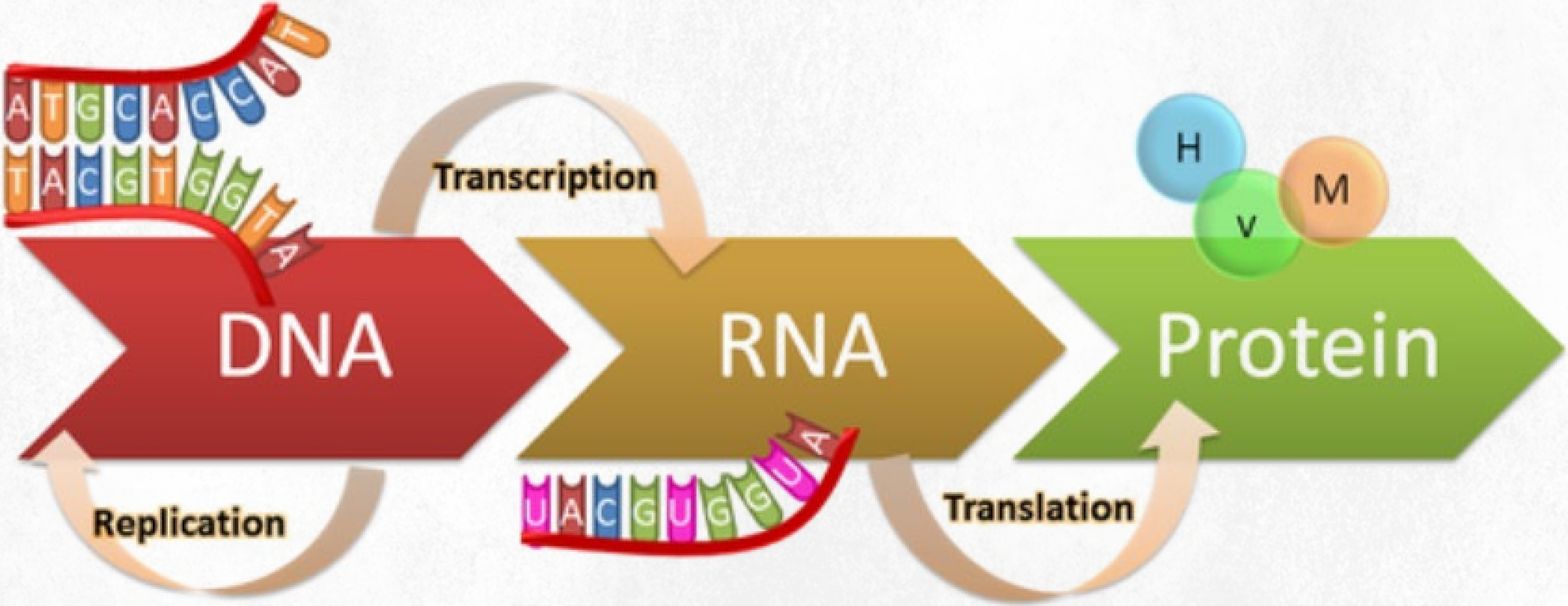
Identifying disease-causing organisms in soil: Primer to public databases

NOURISH THE FUTURE

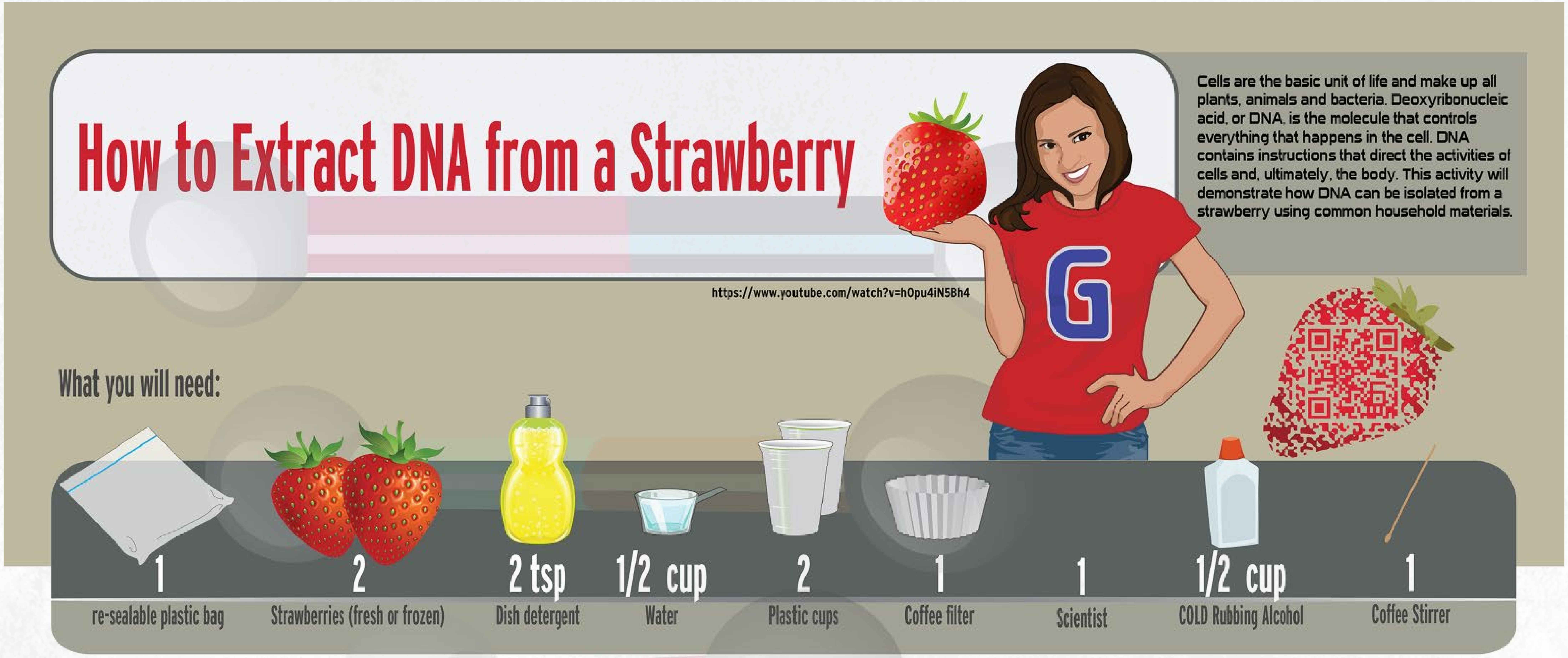
Tomorrow's science is looking for leaders

Central dogma of molecular biology



genius.com/Biology-genius-the-central-dogma-annotated

Extractions help students visualize DNA



How to Extract DNA from a Strawberry

Cells are the basic unit of life and make up all plants, animals and bacteria. Deoxyribonucleic acid, or DNA, is the molecule that controls everything that happens in the cell. DNA contains instructions that direct the activities of cells and, ultimately, the body. This activity will demonstrate how DNA can be isolated from a strawberry using common household materials.

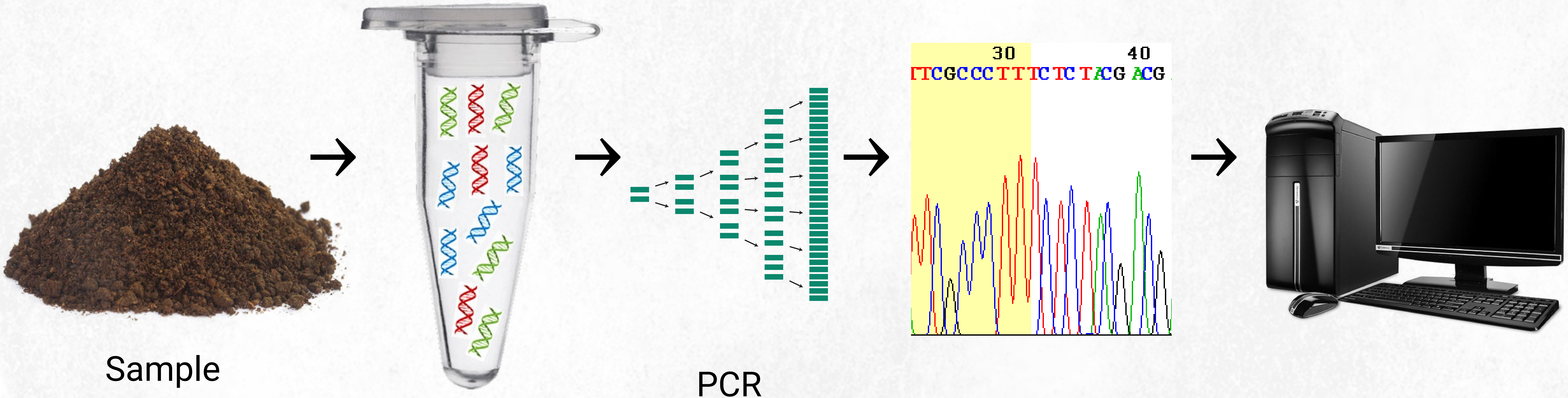
https://www.youtube.com/watch?v=hOpu4iN5Bh4

What you will need:

1	re-sealable plastic bag
2	Strawberries (fresh or frozen)
2 tsp	Dish detergent
1/2 cup	Water
2	Plastic cups
1	Coffee filter
1	Scientist
1/2 cup	COLD Rubbing Alcohol
1	Coffee Stirrer

genome.gov/Pages/Education/Modules/StrawberryExtractionInstructions.pdf

The process: Sample to DNA analysis



Sample collection → DNA extraction → PCR → Sequencing → Bioinformatics

The process: Sample to DNA analysis

Soil sample

→

DNA extraction

Sample collection

↓

DNA extraction

↓

PCR

↓

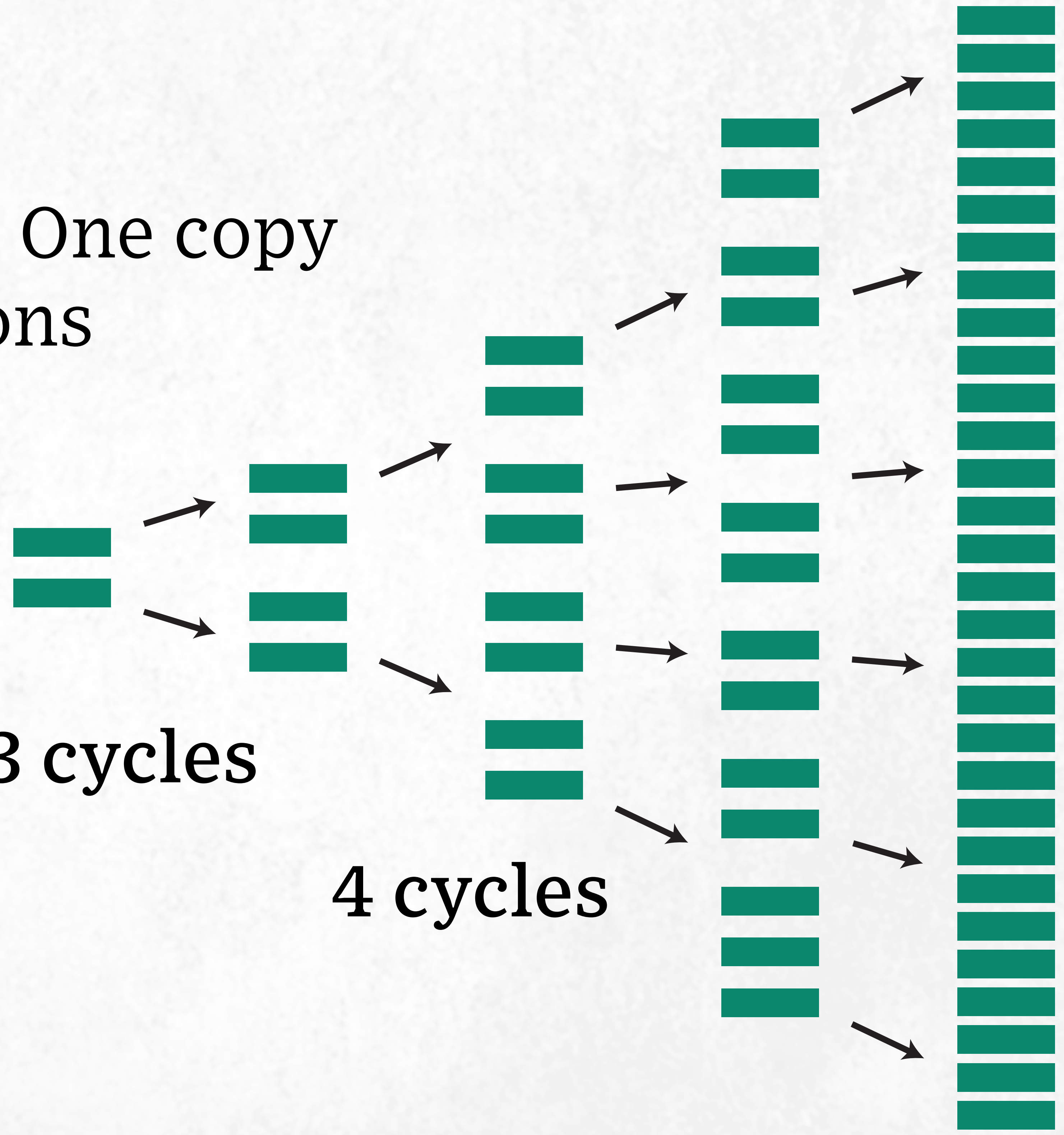
Sequencing

↓

Bioinformatics

The process: Sample to DNA analysis

Polymerase chain reaction (PCR): Cycling process to artificially make copies (amplify) specific pieces of DNA in the genome

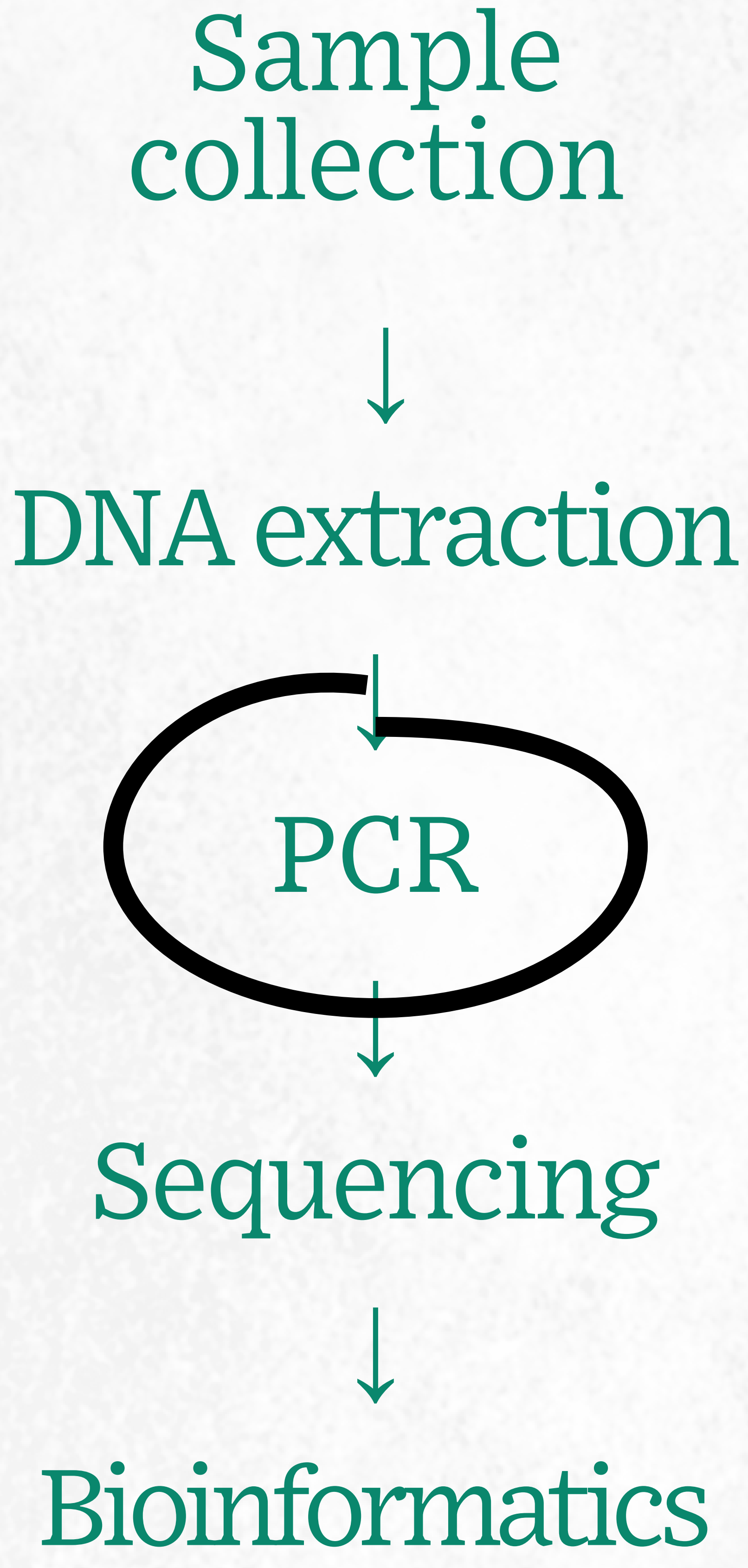


Result: One copy to billions

3 cycles

4 cycles

30 cycles



Sample collection

↓

DNA extraction

↓

PCR

↓

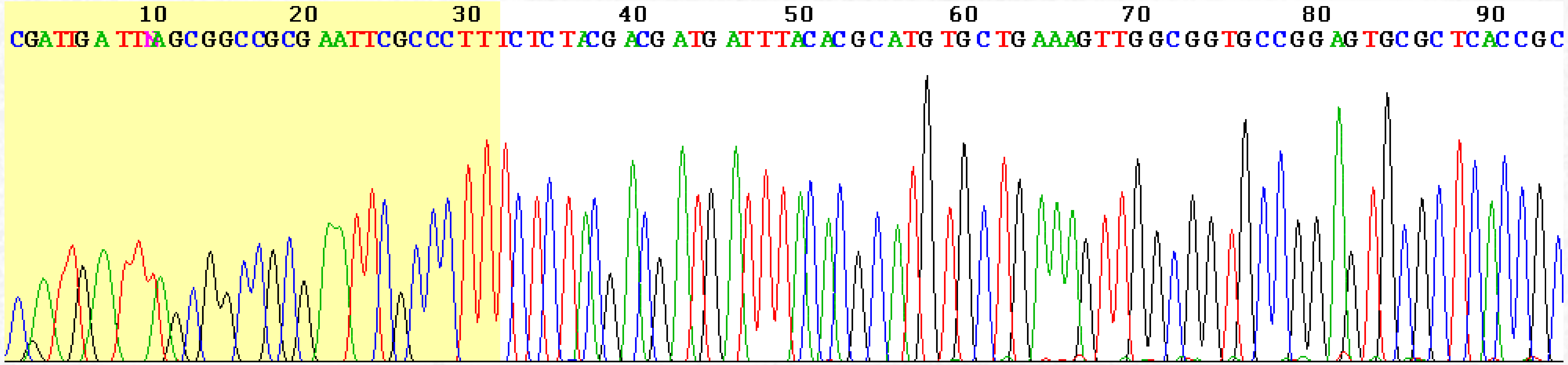
Sequencing

↓

Bioinformatics

The process: Sample to DNA analysis

DNA sequencing: Amplified PCR products are digitized into their nucleotide sequence



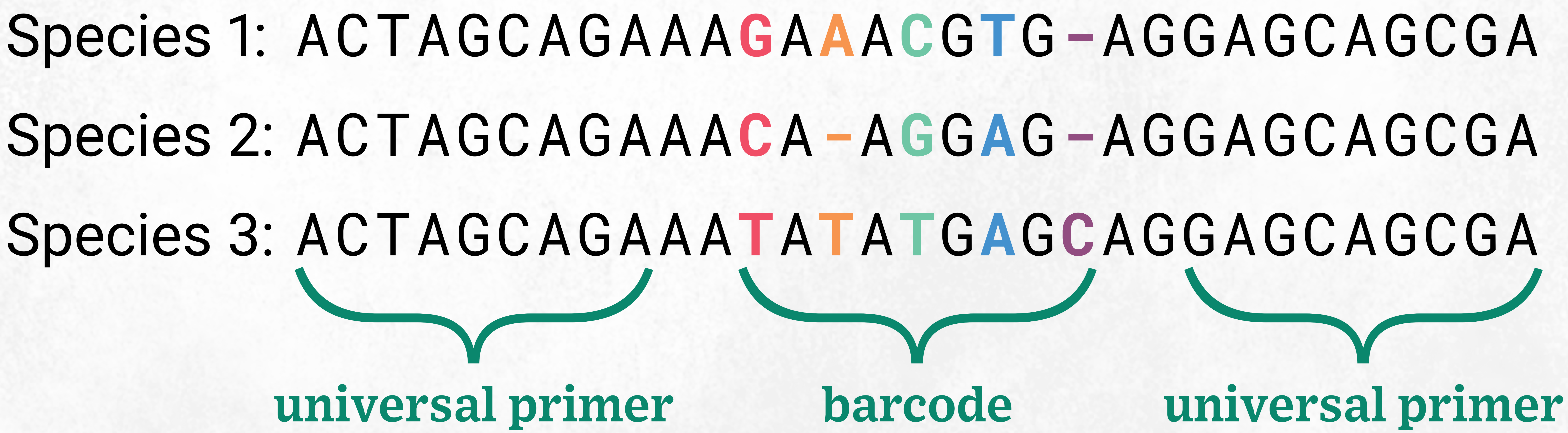
Sample collection
↓
DNA extraction
↓
PCR
↓
Sequencing
↓
Bioinformatics

What is amplified by PCR?

Barcodes: regions of DNA that differ between species.

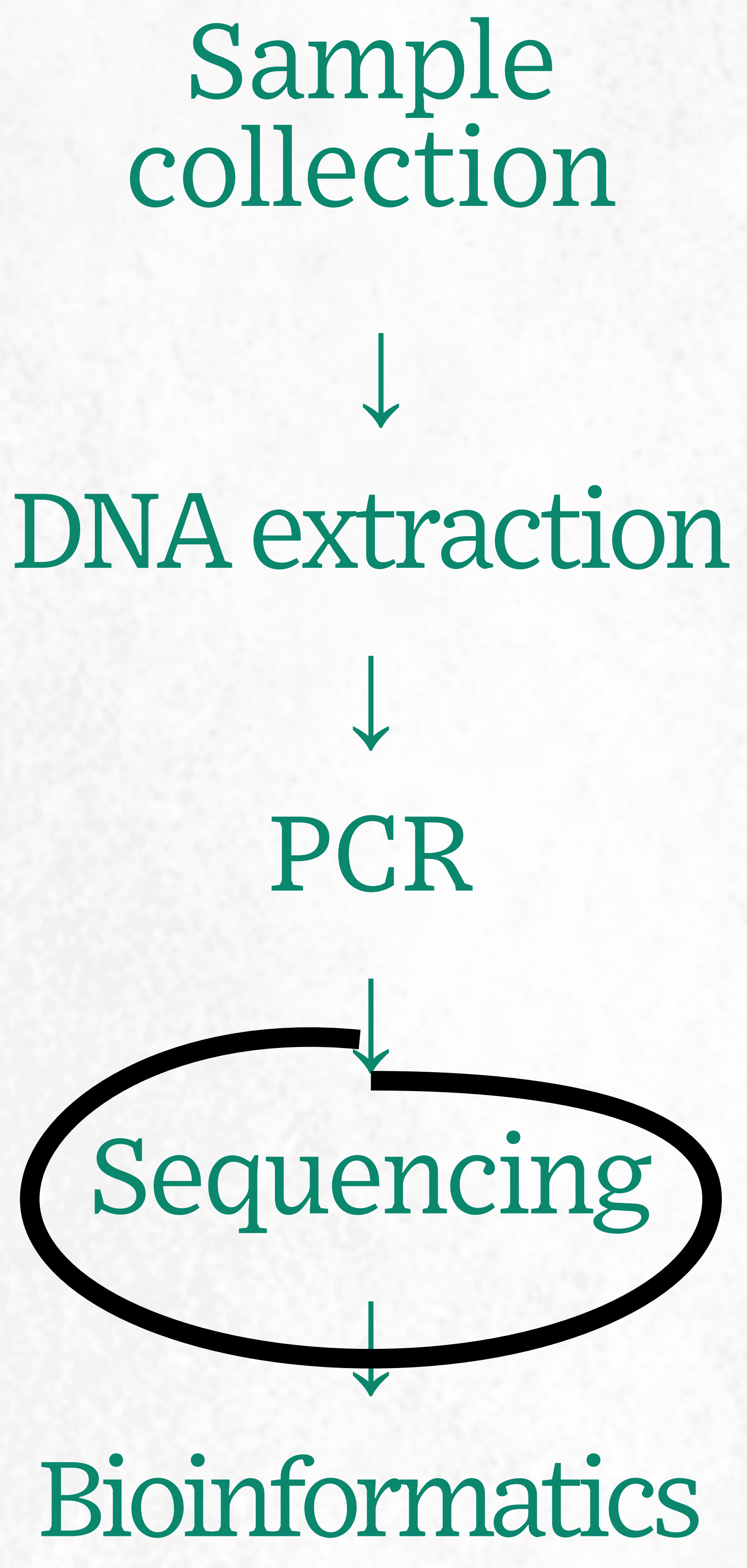
Scientists use a set of **universal primers** in conserved regions that are upstream and downstream for PCR to amplify the barcode region. DNA sequencing and bioinformatics are then used to identify those differences.

DNA sequence alignment



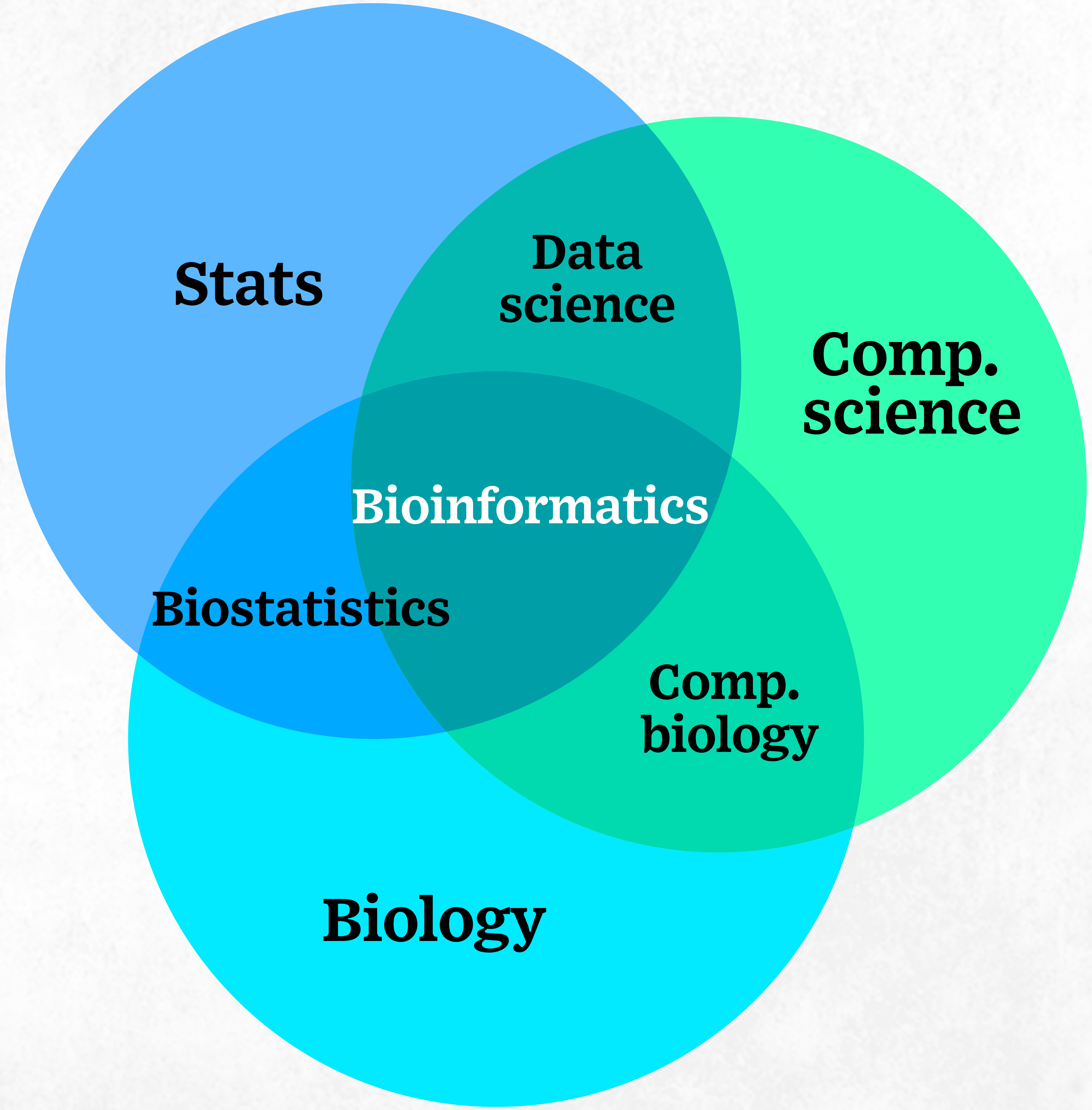
Species 1: ACTAGCAGAAAGAAACGTG-AGGAGCAGCGA
Species 2: ACTAGCAGAAACA-AGGAG-AGGAGCAGCGA
Species 3: ACTAGCAGAAATATATGAGCAGGAGCAGCGA

universal primer
barcode
universal primer



Sample collection
↓
DNA extraction
↓
PCR
↓
Sequencing
↓
Bioinformatics

The process: Sample to DNA analysis



Sample collection

↓

DNA extraction

↓

PCR

↓

Sequencing

↓

Bioinformatics

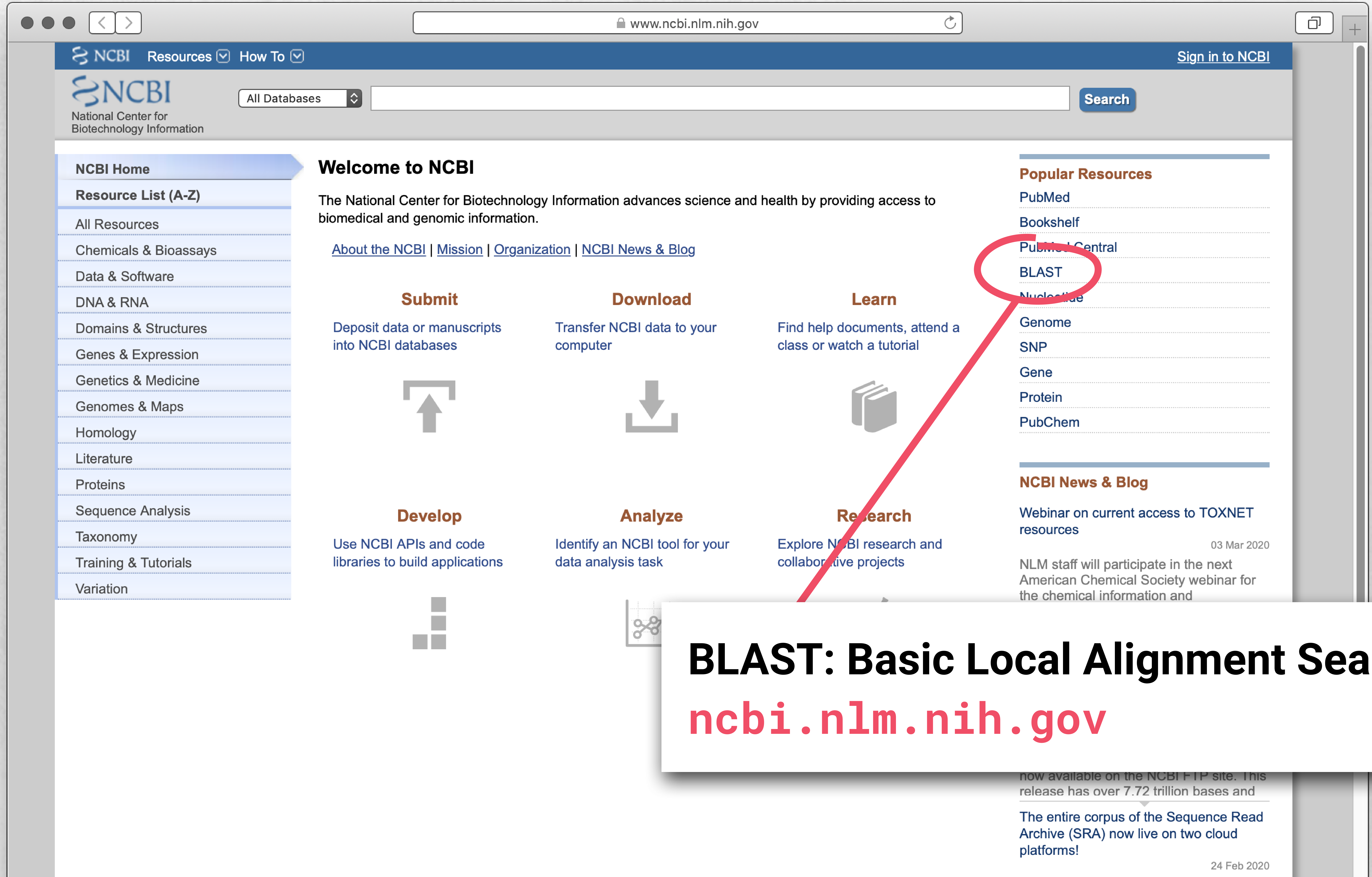
Applications of bioinformatics in food sustainability

1. Genome sequencing and analysis
2. Protein structure analysis
3. Gene discovery:
    - Insect and disease resistance
    - Improved nutritional quality
    - Drought resistance
4. Disease discovery and control

Farm story

Shadybrook Farm is having trouble with production of soybeans and has noticed increasing levels of disease, despite the use of fungicides. The farm manager elects for a soil test from a laboratory to identify potential disease-causing organisms. The laboratory produced a report that contained the top two most common DNA sequences from the soil.

Your job as a bioinformatician is to help the farm manager identify the species, using the given sequences of DNA, through the use of public databases and basic bioinformatic tools.


www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI
National Center for
Biotechnology Information

All Databases Search

- **NCBI Home**
- **Resource List (A-Z)**
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | 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

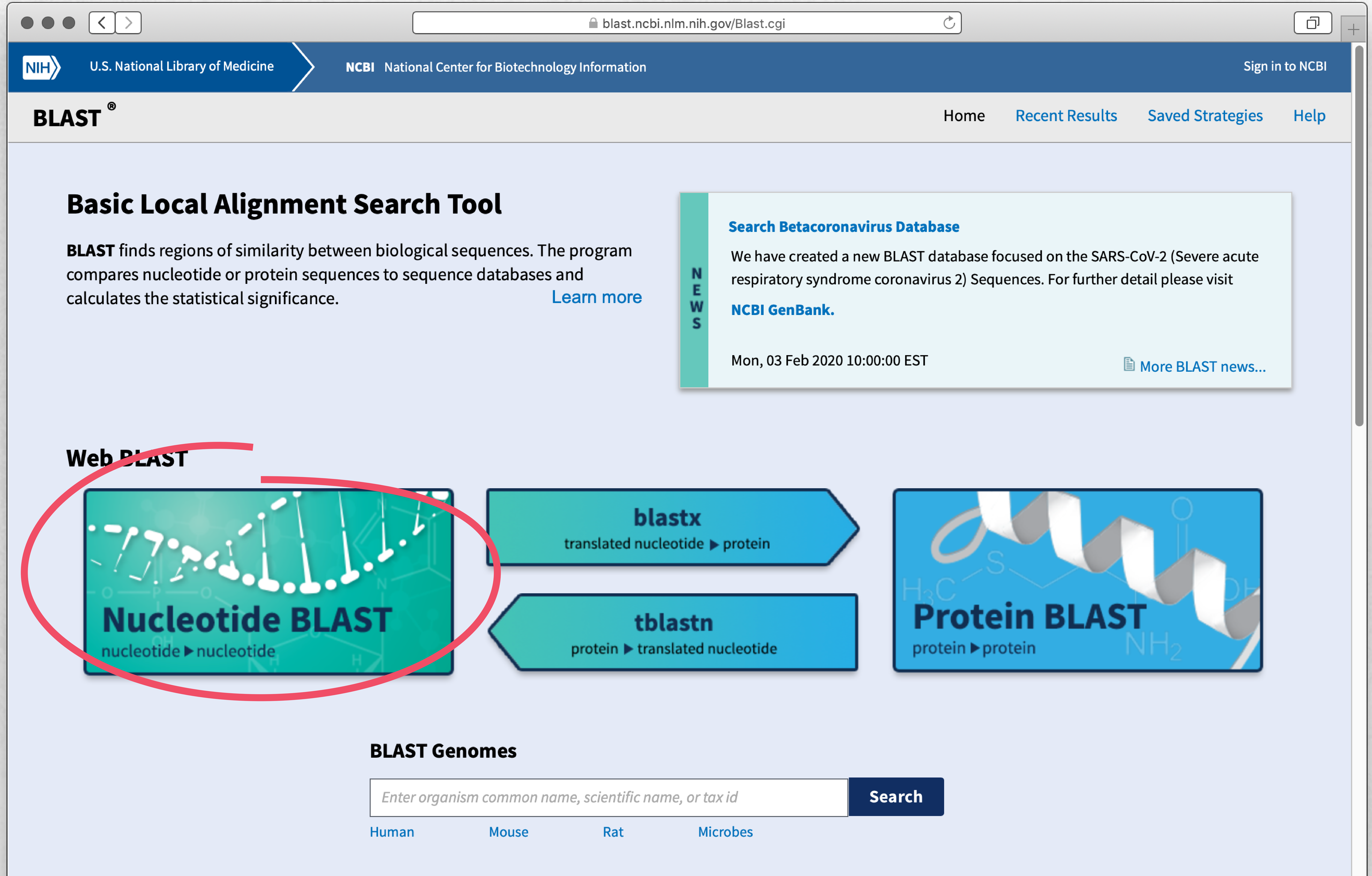
Webinar on current access to TOXNET resources
03 Mar 2020
NLM staff will participate in the next American Chemical Society webinar for the chemical information and

now available on the NCBI FTP site. This release has over 7.72 trillion bases and

The entire corpus of the Sequence Read Archive (SRA) now live on two cloud platforms!
24 Feb 2020

BLAST: Basic Local Alignment Search

ncbi.nlm.nih.gov



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

Search Betacoronavirus Database

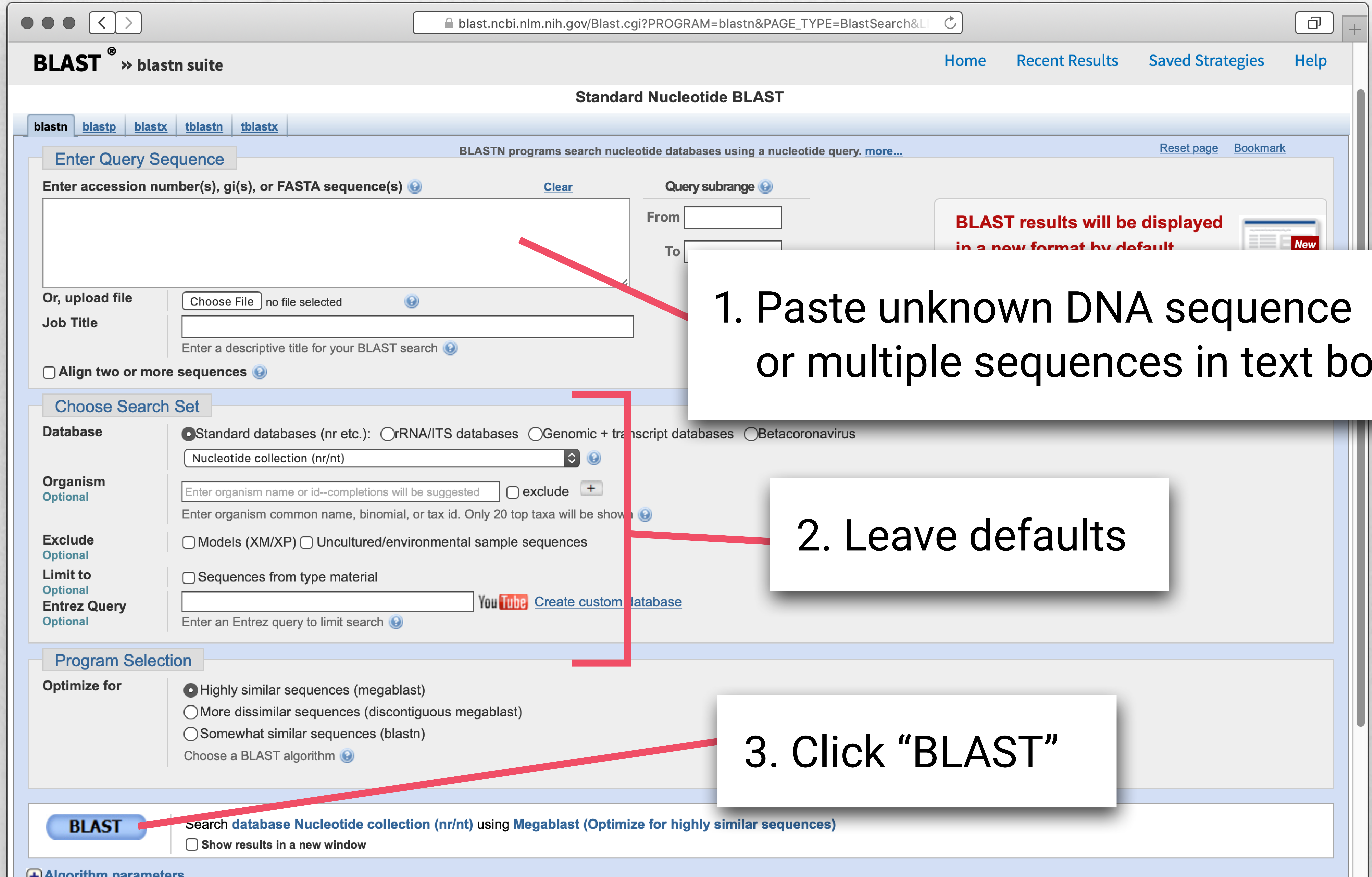
We have created a new BLAST database focused on the SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences. For further detail please visit **NCBI GenBank.**

Mon, 03 Feb 2020 10:00:00 EST

More BLAST news...

BLAST Genomes

Enter organism common name, scientific name, or tax id



blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&L

BLAST® » **blastn suite**

Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

BLAST results will be displayed in a new format by default

New

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

☐ **Align two or more sequences**

Choose Search Set

Database ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

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 shown

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

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

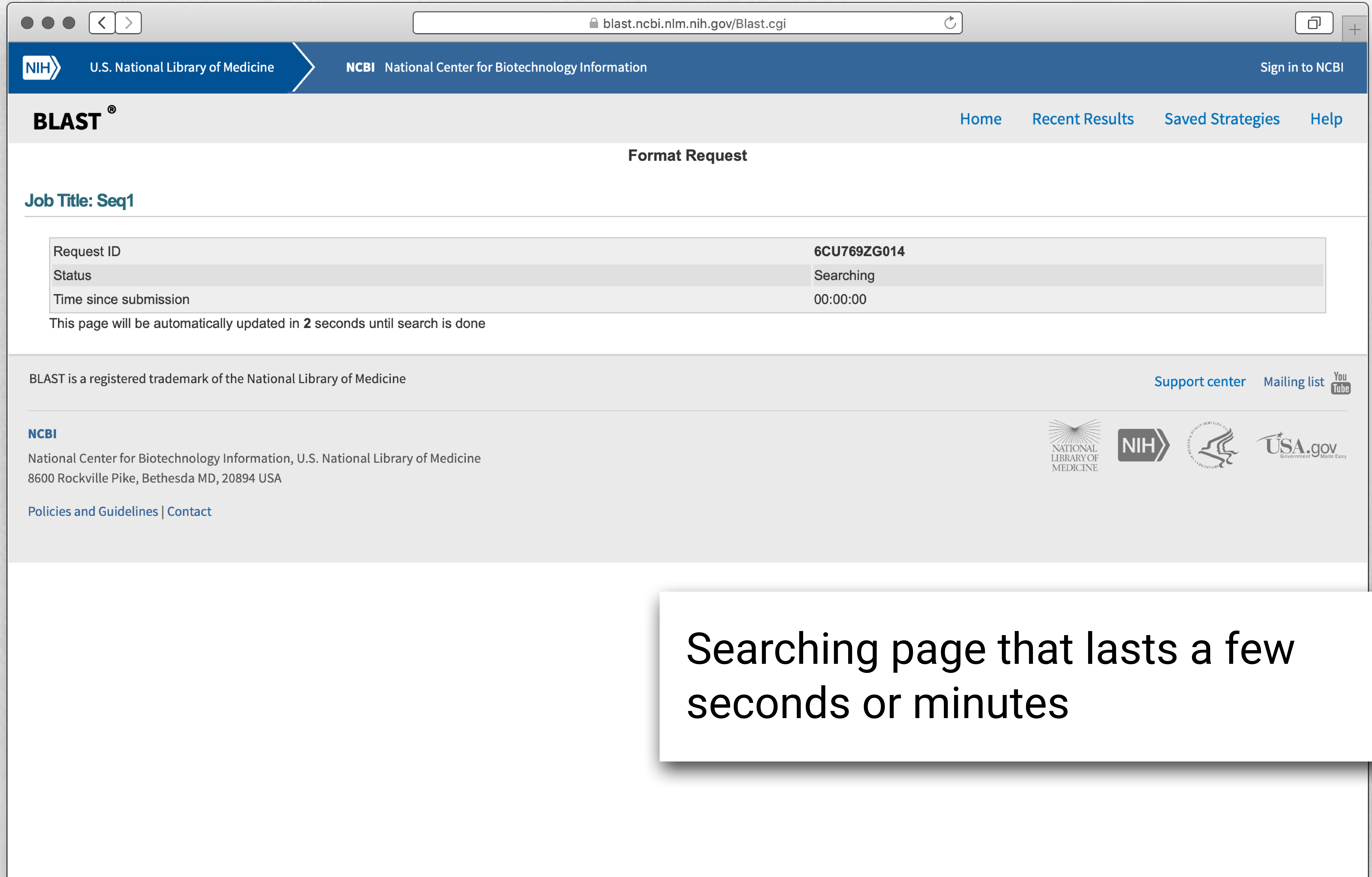
Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ **Show results in a new window**

Algorithm parameters

1. Paste unknown DNA sequence or multiple sequences in text box
2. Leave defaults
3. Click “BLAST”



Format Request

Job Title: Seq1

Request ID	**6CU769ZG014**
Status	Searching
Time since submission	00:00:00

This page will be automatically updated in **2** seconds until search is done

BLAST is a registered trademark of the National Library of Medicine

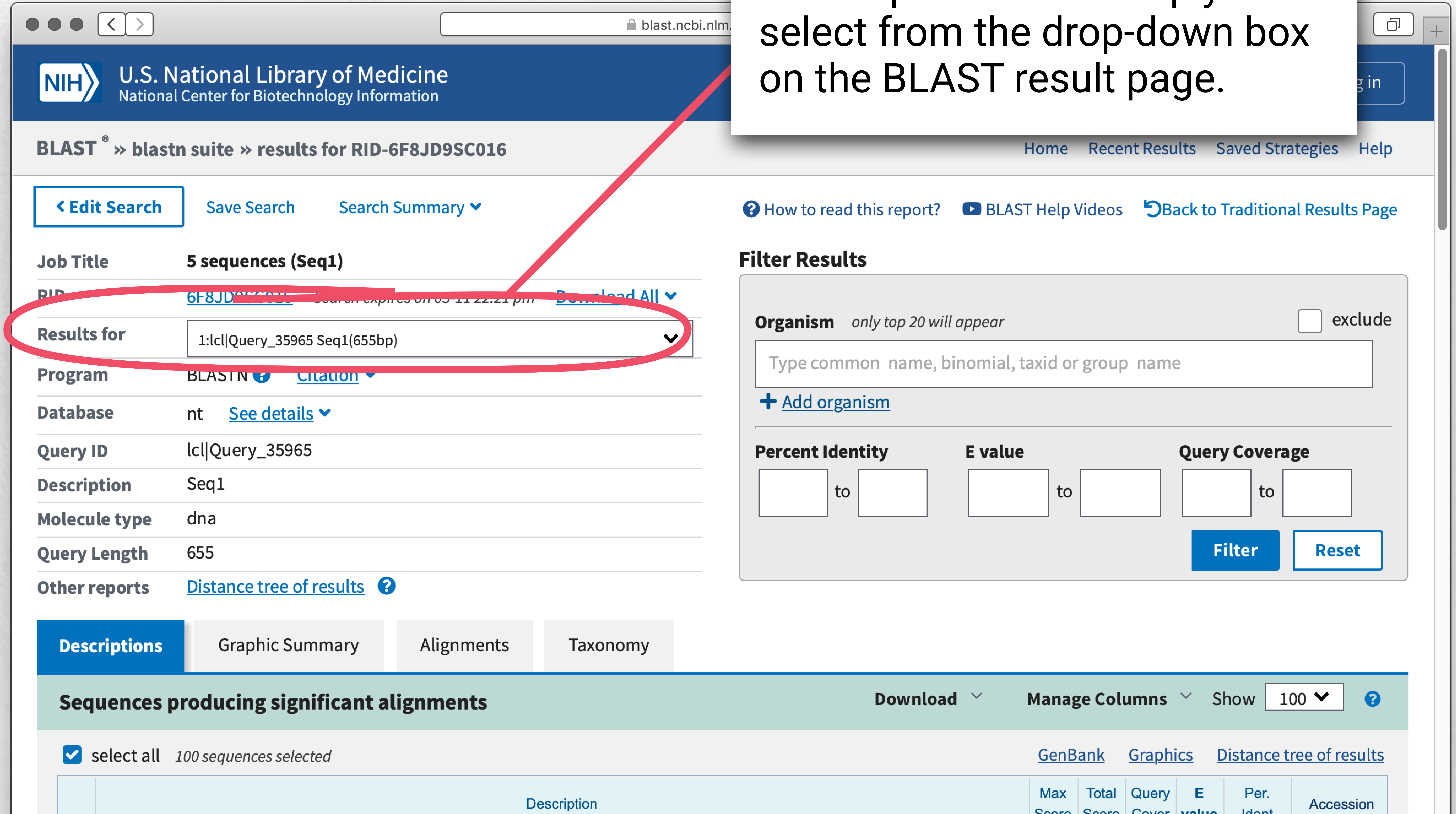
NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Searching page that lasts a few seconds or minutes

BLAST results

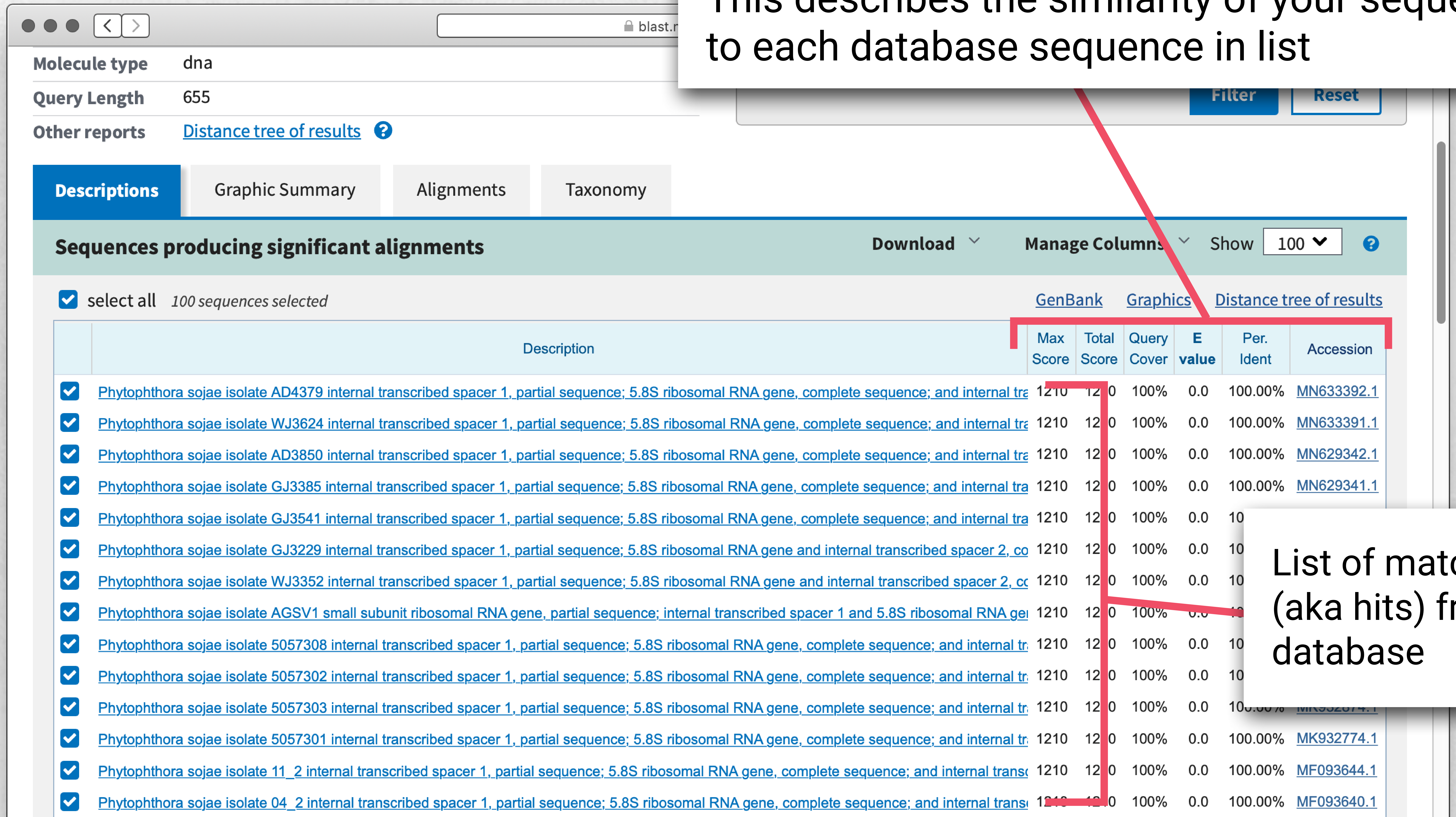
Multiple sequence searches can be performed. Simply select from the drop-down box on the BLAST result page.



Job Title	5 sequences (Seq1)
RID	6F8JD9SC016 Search expires on 03-11 22:21 pm Download All
Results for	1:lcl|Query_35965 Seq1(655bp)
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_35965
Description	Seq1
Molecule type	dna
Query Length	655
Other reports	Distance tree of results

BLAST results

This describes the similarity of your sequence to each database sequence in list



Molecule type	dna
Query Length	655
Other reports	Distance tree of results

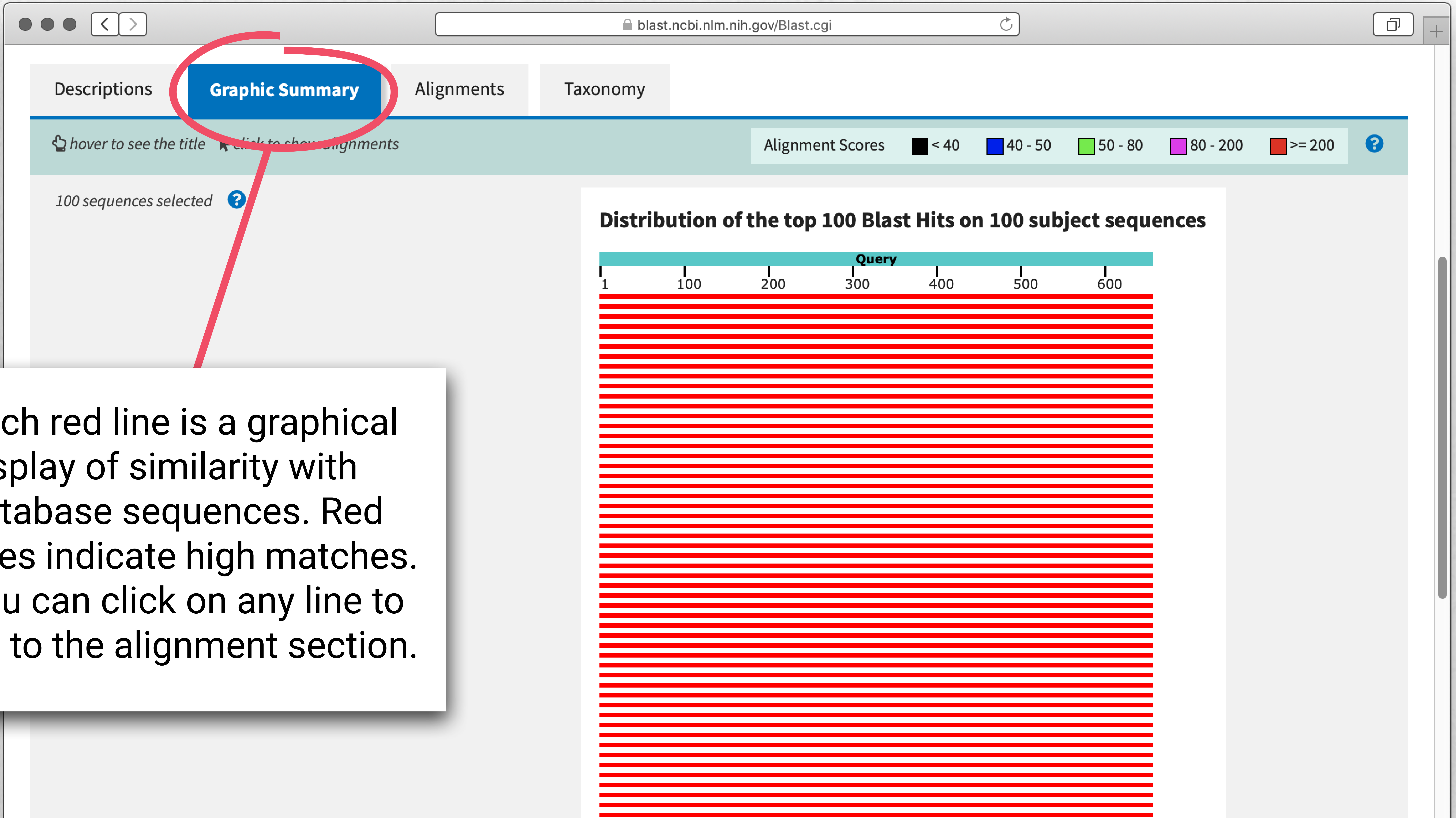
Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Phytophthora sojae isolate AD4379 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tra	1210	[illegible]	100%	0.0	100.00%	MN633392.1
☑	Phytophthora sojae isolate WJ3624 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tra	1210	[illegible]	100%	0.0	100.00%	MN633391.1
☑	Phytophthora sojae isolate AD3850 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tra	1210	[illegible]	100%	0.0	100.00%	MN629342.1
☑	Phytophthora sojae isolate GJ3385 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tra	1210	[illegible]	100%	0.0	100.00%	MN629341.1
☑	Phytophthora sojae isolate GJ3541 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tra	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate GJ3229 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, co	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate WJ3352 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, c	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate AGSV1 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1 and 5.8S ribosomal RNA ge	1210	[illegible]	100%	[illegible]	[illegible]	[illegible]
☑	Phytophthora sojae isolate 5057308 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tr	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate 5057302 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tr	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate 5057303 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tr	1210	[illegible]	100%	0.0	[illegible]	[illegible]
☑	Phytophthora sojae isolate 5057301 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal tr	1210	[illegible]	100%	0.0	100.00%	MK932774.1
☑	Phytophthora sojae isolate 11_2 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal trans	1210	[illegible]	100%	0.0	100.00%	MF093644.1
☑	Phytophthora sojae isolate 04_2 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal trans	[illegible]	[illegible]	100%	0.0	100.00%	MF093640.1

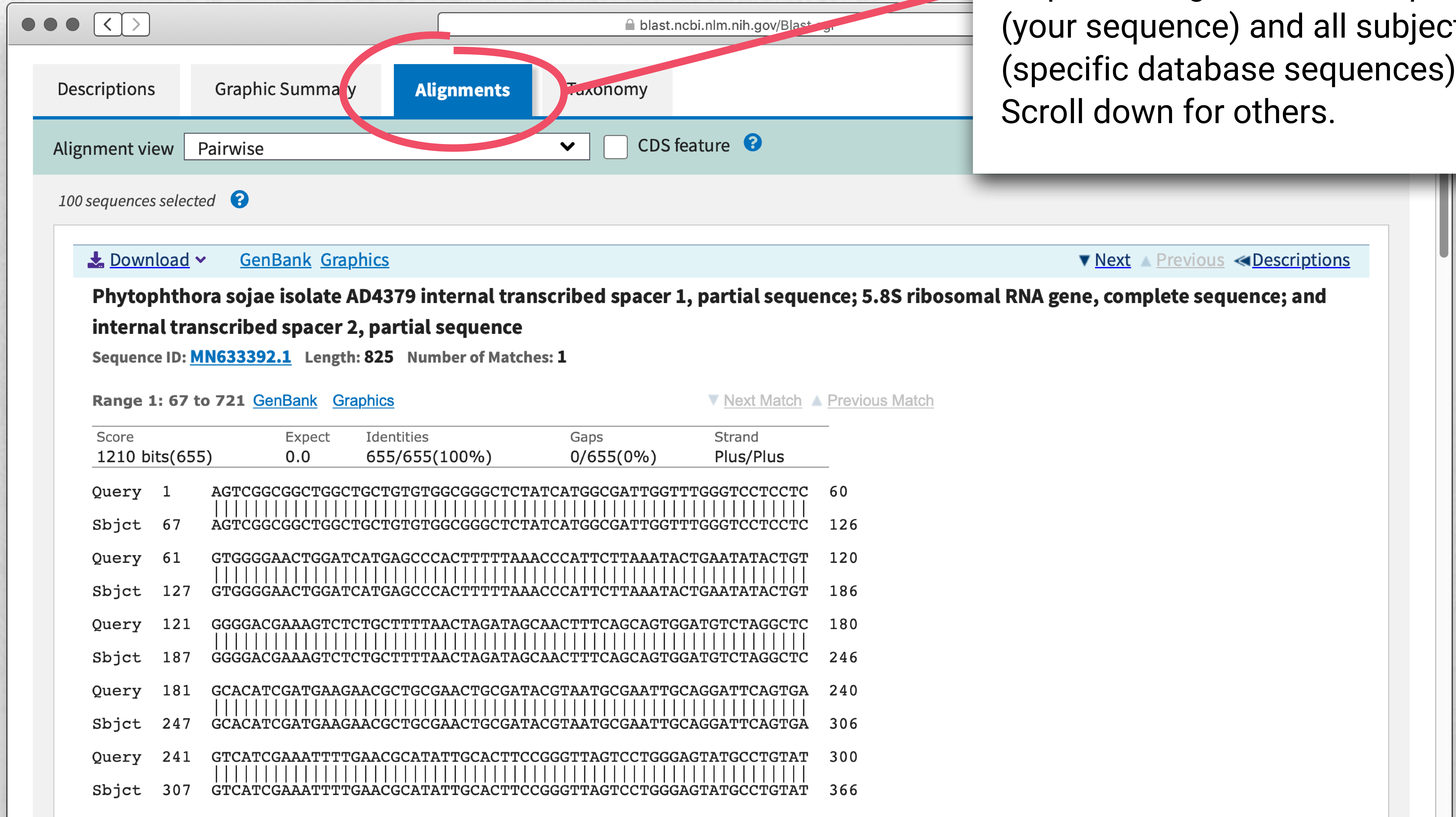
List of matches (aka hits) from database

BLAST results: graphical summary



Each red line is a graphical display of similarity with database sequences. Red lines indicate high matches. You can click on any line to go to the alignment section.

BLAST results: alignment



Alignment view: Pairwise

☐ CDS feature

100 sequences selected

Phytophthora sojae isolate AD4379 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence

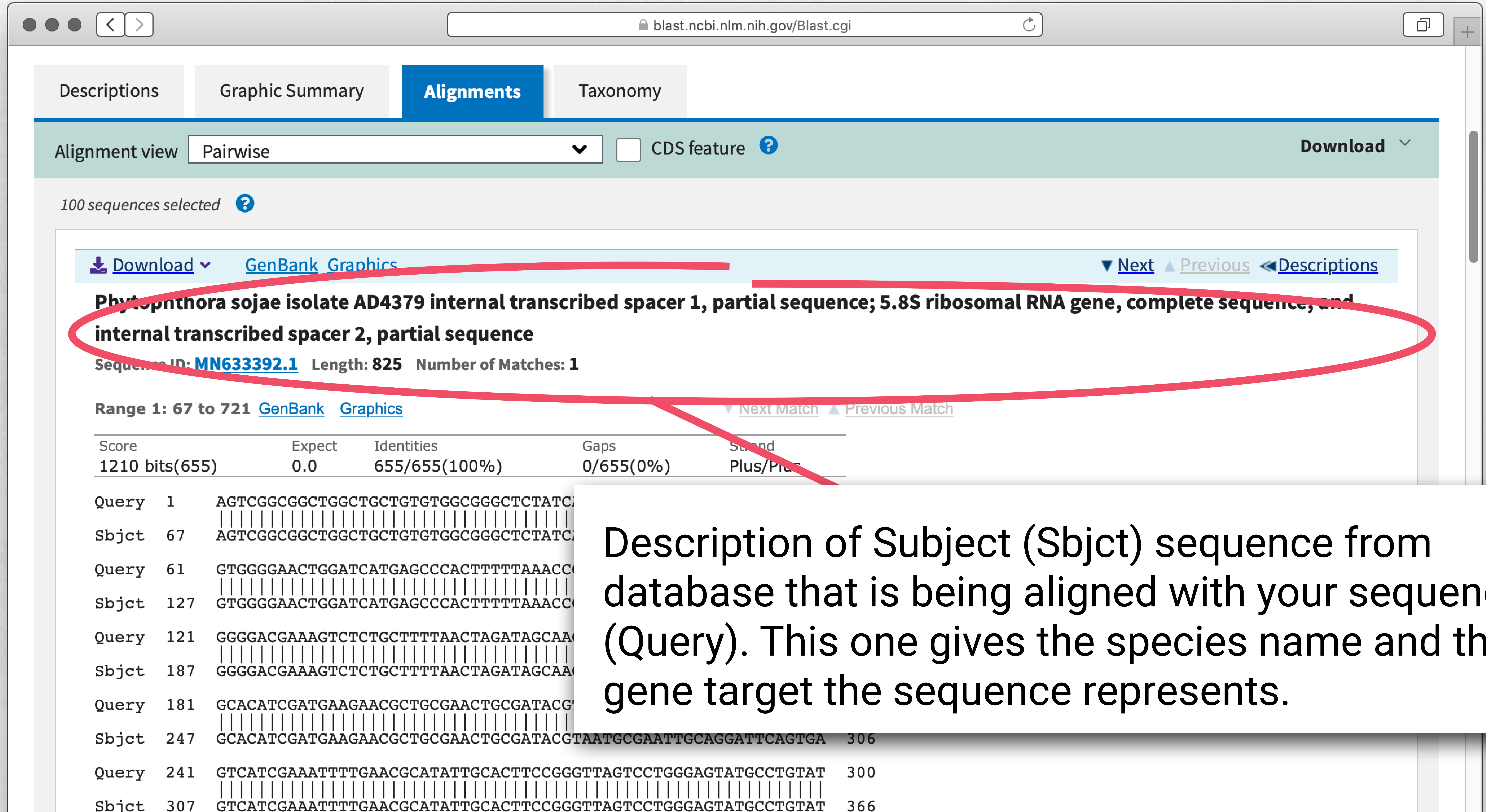
Sequence ID: MN633392.1 **Length:** 825 **Number of Matches:** 1

Range 1: 67 to 721 GenBank Graphics

```
Score            Expect   Identities         Gaps          Strand
1210 bits(655)   0.0      655/655(100%)      0/655(0%)     Plus/Plus

Query  1    AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATCATGGCGATTGGTTTGGGTCCTCCTC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  67   AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATCATGGCGATTGGTTTGGGTCCTCCTC  126

Query  61   GTGGGGAACTGGATCATGAGCCCACTTTTTAAACCCATTCTTAAATACTGAATATACTGT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  127  GTGGGGAACTGGATCATGAGCCCACTTTTTAAACCCATTCTTAAATACTGAATATACTGT  186

Query  121  GGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAACTTTCAGCAGTGGATGTCTAGGCTC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  187  GGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAACTTTCAGCAGTGGATGTCTAGGCTC  246

Query  181  GCACATCGATGAAGAACGCTGCGAACTGCGATACGTAATGCGAATTGCAGGATTCAGTGA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  247  GCACATCGATGAAGAACGCTGCGAACTGCGATACGTAATGCGAATTGCAGGATTCAGTGA  306

Query  241  GTCATCGAAATTTTGAACGCATATTGCACTTCCGGGTTAGTCCTGGGAGTATGCCTGTAT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  307  GTCATCGAAATTTTGAACGCATATTGCACTTCCGGGTTAGTCCTGGGAGTATGCCTGTAT  366
```

Alignment tab displays individual sequence alignments with query (your sequence) and all subjects (specific database sequences). Scroll down for others.

Sequence alignment (100% match)



100 sequences selected

Phytophthora sojae isolate AD4379 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence

Sequence ID: MN633392.1 **Length:** 825 **Number of Matches:** 1

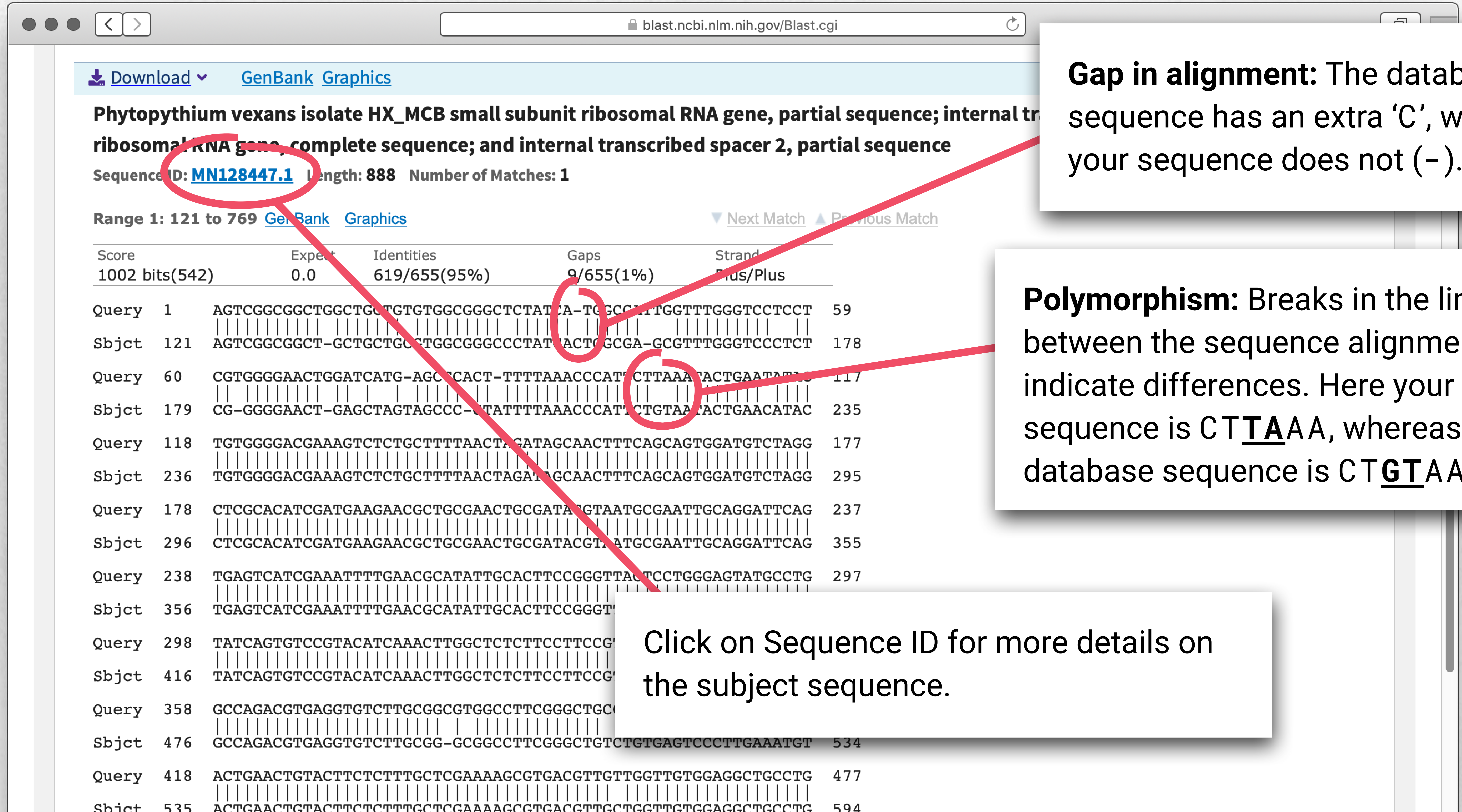
Range 1: 67 to 721

Score	Expect	Identities	Gaps	Strand
1210 bits(655)	0.0	655/655(100%)	0/655(0%)	Plus/Plus

```
Query  1    AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATC
            |||||||||||||||||||||||||||||||||||
Sbjct  67   AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATC

Query  61   GTGGGGAACTGGATCATGAGCCCACTTTTTAAACC
            |||||||||||||||||||||||||||||||||||
Sbjct  127  GTGGGGAACTGGATCATGAGCCCACTTTTTAAACC

Query  121  GGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAA
            |||||||||||||||||||||||||||||||||||
Sbjct  187  GGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAA

Query  181  GCACATCGATGAAGAACGCTGCGAACTGCGATACG
            |||||||||||||||||||||||||||||||||||
Sbjct  247  GCACATCGATGAAGAACGCTGCGAACTGCGATACGTAATGCGAATTGCAGGATTCAGTGA  306

Query  241  GTCATCGAAATTTTGAACGCATATTGCACTTCCGGGTTAGTCCTGGGAGTATGCCTGTAT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  307  GTCATCGAAATTTTGAACGCATATTGCACTTCCGGGTTAGTCCTGGGAGTATGCCTGTAT  366
```

Description of Subject (Sbjct) sequence from database that is being aligned with your sequence (Query). This one gives the species name and the gene target the sequence represents.

Sequence alignment (95% match)



Phytopythium vexans isolate HX_MCB small subunit ribosomal RNA gene, partial sequence; internal tr
ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence

Sequence ID: MN128447.1 Length: 888 Number of Matches: 1

Range 1: 121 to 769 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1002 bits(542)	0.0	619/655(95%)	9/655(1%)	Plus/Plus

```
Query  1    AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATCA-TGGCCATTGGTTTGGGTCCTCCT  59
            ||||||||||| |||||||  |||||||| |||||| |||| |   ||||||||||  ||
Sbjct  121  AGTCGGCGGCT-GCTGCTGCCTGGCGGGCCCTATCACTGGCGA-GCGTTTGGGTCCCTCT  178

Query  60   CGTGGGGAACTGGATCATG-AGCCCACT-TTTTAAACCCATTCTTAAATACTGAATATAC  117
            || |||||||| || |  | ||||| || ||| || || ||  ||||  ||||||||| ||||
Sbjct  179  CG-GGGGAACT-GAGCTAGTAGCCC-CTATTTAAACCCATTTCTGTAATACTGAACATAC  235

Query  118  TGTGGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAACTTTCAGCAGTGGATGTCTAGG  177
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  236  TGTGGGGACGAAAGTCTCTGCTTTTAACTAGATAGCAACTTTCAGCAGTGGATGTCTAGG  295

Query  178  CTCGCACATCGATGAAGAACGCTGCGAACTGCGATACGTAATGCGAATTGCAGGATTCAG  237
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  296  CTCGCACATCGATGAAGAACGCTGCGAACTGCGATACGTAATGCGAATTGCAGGATTCAG  355

Query  238  TGAGTCATCGAAATTTTGAACGCATATTGCACTTCCGGGTTACTCCTGGGAGTATGCCTG  297
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  356  TGAGTCATCGAAATTTTGAACGCATATTGCACTTCCGGGT

Query  298  TATCAGTGTCCGTACATCAAACTTGGCTCTCTTCCTTCCG
            ||||||||||||||||||||||||||||||||||||||||
Sbjct  416  TATCAGTGTCCGTACATCAAACTTGGCTCTCTTCCTTCCG

Query  358  GCCAGACGTGAGGTGTCTTGCGGCGTGGCCTTCGGGCTGC
            ||||||||||||||||||||||| | |||||||||||||
Sbjct  476  GCCAGACGTGAGGTGTCTTGCGG-GCGGCCTTCGGGCTGTCTGTGAGTCCCTTGAAATGT  534

Query  418  ACTGAACTGTACTTCTCTTTGCTCGAAAAGCGTGACGTTGTTGGTTGTGGAGGCTGCCTG  477
            |||||||||||||||||||||||||||||||||||||||| |||||||||||||||||||
Sbjct  535  ACTGAACTGTACTTCTCTTTGCTCGAAAAGCGTGACGTTGCTGGTTGTGGAGGCTGCCTG  594
```

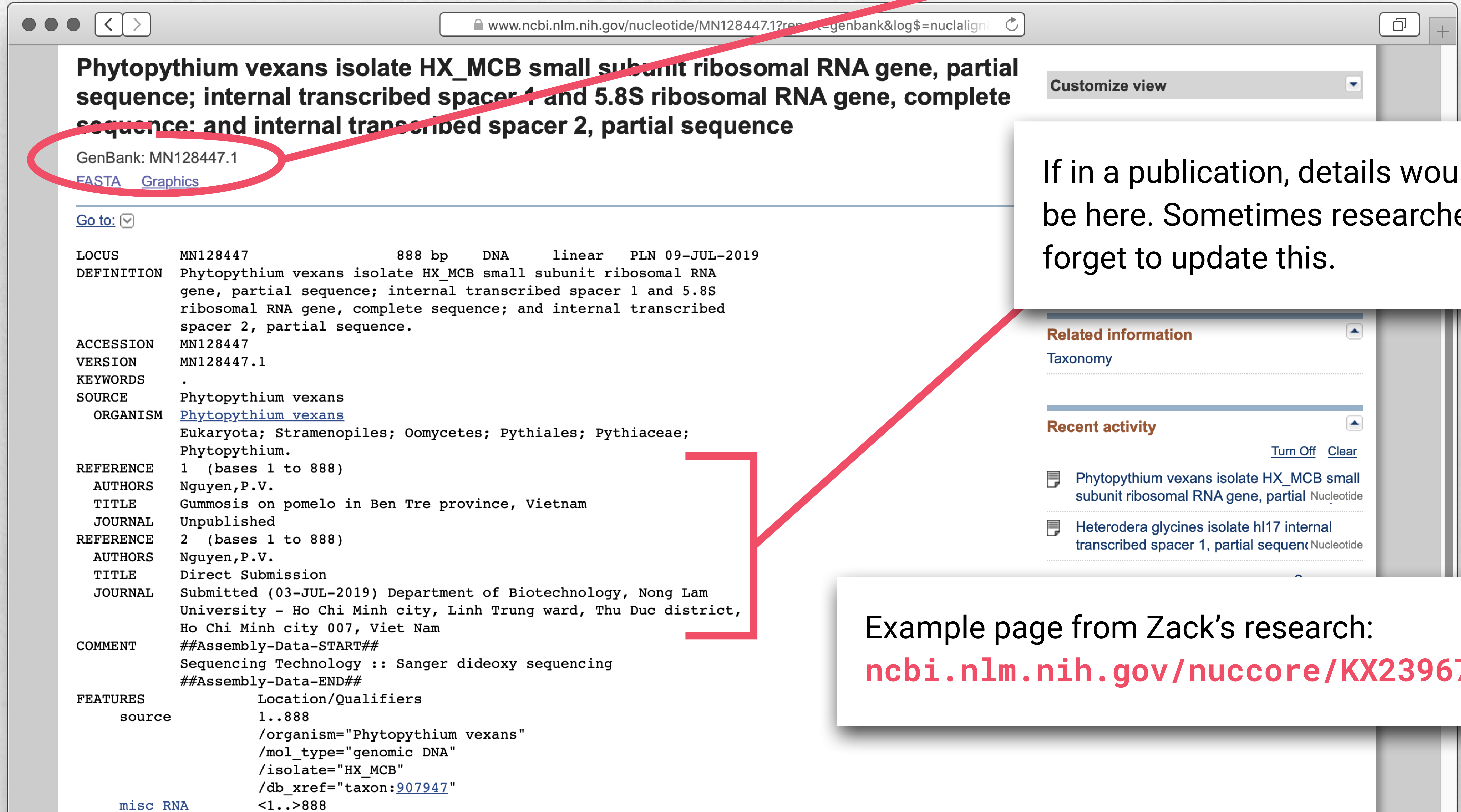
Gap in alignment: The database sequence has an extra ‘C’, where your sequence does not (–).

Polymorphism: Breaks in the lines between the sequence alignments indicate differences. Here your sequence is CT**TA**AA, whereas the database sequence is CT**GT**AA.

Click on Sequence ID for more details on the subject sequence.

Subject sequence page

Sequence ID number



Phytopythium vexans isolate HX_MCB small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1 and 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence

GenBank: MN128447.1

FASTA Graphics

Go to:

```
LOCUS       MN128447                 888 bp    DNA     linear   PLN 09-JUL-2019
DEFINITION  Phytopythium vexans isolate HX_MCB small subunit ribosomal RNA
            gene, partial sequence; internal transcribed spacer 1 and 5.8S
            ribosomal RNA gene, complete sequence; and internal transcribed
            spacer 2, partial sequence.
ACCESSION   MN128447
VERSION     MN128447.1
KEYWORDS    .
SOURCE      Phytopythium vexans
  ORGANISM  Phytopythium vexans
            Eukaryota; Stramenopiles; Oomycetes; Pythiales; Pythiaceae;
            Phytopythium.
REFERENCE   1  (bases 1 to 888)
  AUTHORS   Nguyen,P.V.
  TITLE     Gummosis on pomelo in Ben Tre province, Vietnam
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 888)
  AUTHORS   Nguyen,P.V.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-JUL-2019) Department of Biotechnology, Nong Lam
            University - Ho Chi Minh city, Linh Trung ward, Thu Duc district,
            Ho Chi Minh city 007, Viet Nam
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..888
                     /organism="Phytopythium vexans"
                     /mol_type="genomic DNA"
                     /isolate="HX_MCB"
                     /db_xref="taxon:907947"
     misc_RNA        <1..>888
```

Customize view

Related information

Taxonomy

Recent activity

Turn Off Clear

Phytopythium vexans isolate HX_MCB small subunit ribosomal RNA gene, partial Nucleotide

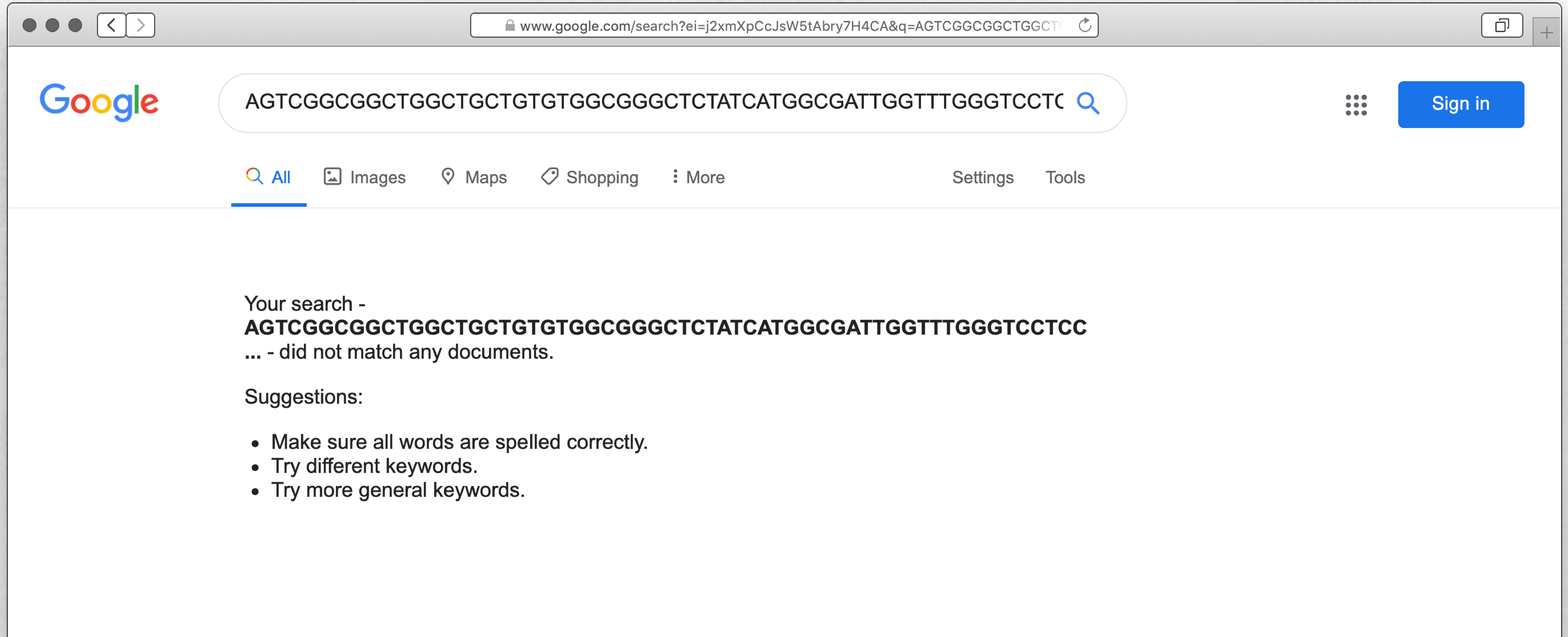
Heterodera glycines isolate hl17 internal transcribed spacer 1, partial sequen Nucleotide

If in a publication, details would be here. Sometimes researchers forget to update this.

Example page from Zack's research:
ncbi.nlm.nih.gov/nuccore/KX239675.1

Additional ideas for teaching DNA databases

Have your students try to identify the sequences before instruction on databases.



Your search - **AGTCGGCGGCTGGCTGCTGTGTGGCGGGCTCTATCATGGCGATTGGTTTGGGTCCTCC ...** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

Reflection and feedback

- Under what unit in your curriculum would you incorporate bioinformatics?
- What are some other potential genetic topics or DNA analysis of interest?
- If time allows, try searching NCBI for genes or proteins that interest you. We can help!

Thank you!

Come visit us to learn more!

@NAGClab **@educationproj**

zack.bateson@genotypingcenter.com

jane@educationprojects.org