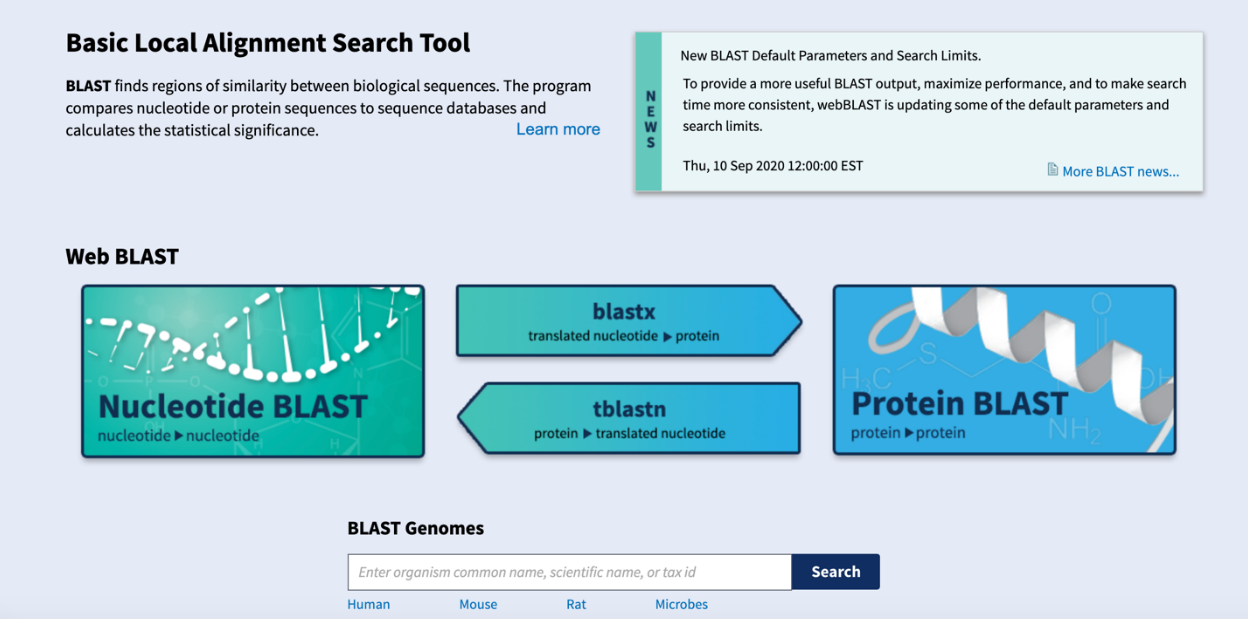
# -Introduction to the BLAST Suite and BLASTN-

An important goal of genomics is to determine if a particular sequence is like another sequence. This is accomplished by comparing the new sequence with sequences that have already been reported and stored in the databases. So, if you have a new sequence with unknown function, you need to align your sequence with the available sequences in the database. This analysis can tell you two things:

1) Whether public databases contain any sequence that can be a potential homolog of your newly derived sequence,
2) and whether your sequence contains some non-functional motif present in other protein families.

Basic local alignment search tool (BLAST) is the most widely used local alignment tool in the world. BLAST uses a robust statistical framework that can determine if the alignment between a query sequence and the target sequence found in the database is statistically significant.

**This the web interface of BLAST. There are four different types of BLAST, which one you are going to use depends on what sequence you have and what your aim is.**



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.

New BLAST Default Parameters and Search Limits.

To provide a more useful BLAST output, maximize performance, and to make search time more consistent, webBLAST is updating some of the default parameters and search limits.

Thu, 10 Sep 2020 12:00:00 EST

**In this lab, you are presented with two exercises that involves unknown nucleotide sequences and to determine what is requested you must perform Nucleotide BLAST by comparing it with the existing nucleotide sequences in the database.**

## Exercise 1: Biofilm analysis

Public water supply lines are immersed in water for decades and a community of microorganisms thrives on these wet surfaces. These slippery coatings are referred to as biofilms and the bacterial makeup is generally unknown because scientists are unable to culture and study the vast majority of these organisms in the laboratory. In 2003, Schmeisser and colleagues published a study where they collected and sequenced the DNA from bacteria growing on pipe valves of a drinking water network in Northern Germany. Through **sequence similarity**, they were able to classify a large number of these organisms as belonging to certain **species or groups**. In this process they identified many new species.

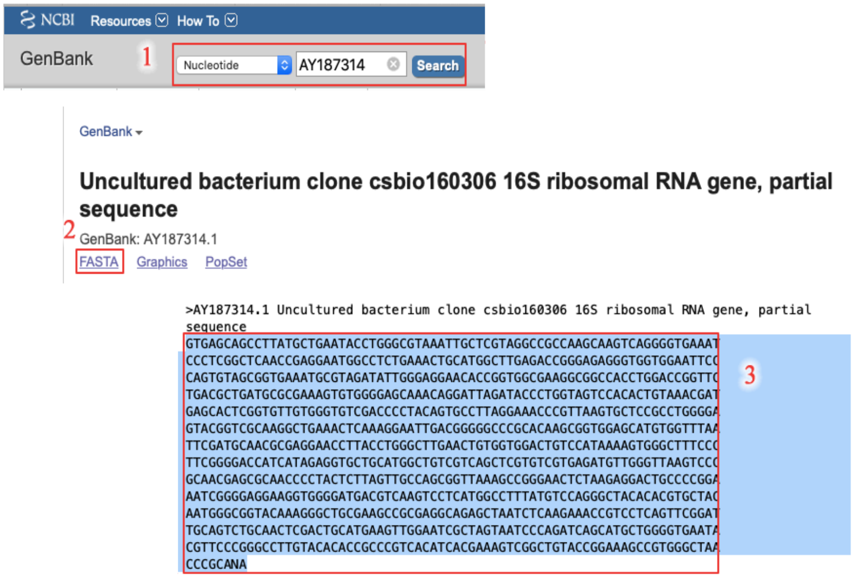
In this exercise, you are asked to use **BLASTN** to repeat some of their analysis and identify the makeup of these biofilms. Below is a list of 5 sequence accession numbers from their study. You are going to use the **NCBI BLASTN** web form to search for sequence similarities to try to identify the bacteria growing within these biofilms.

**AY187314 - AY187315 - AY187316 - AY187317 - AY187318**

## As an example on how to run BLASTN, the results of the first sequence accession number (AY187314) is explained:

**First,** you will need to retrieve the sequence from the **NCBI GenBank (Nucleotide database)** then copy the **FASTA format** of the sequence.

*Based on the annotation of this sequence record, the gene was identified as*:



Uncultured bacterium clone csbio160306 16S ribosomal RNA gene, partial sequence

GenBank: AY187314.1

```
>AY187314.1 Uncultured bacterium clone csbio160306 16S ribosomal RNA gene, partial
sequence
GTGAGCAGCCTTATGCTGAATACCTGGGCGTAAATTGCTCGTAGGCCGCCAAGCAAGTCAGGGGTGAAAT
CCCTCGGCTCAACCGAGGAATGGCCTCTGAAACTGCATGGCTTGAGACCGGGAGAGGGTGGTGGAATTCC
CAGTGTAGCGGTGAAATGCGTAGATATTGGGAGGAACACCGGTGGCGAAGGCGGCCACCTGGACCGGTTC
TGACGCTGATGCGCGAAAGTGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACACTGTAAACGAT
GAGCACTCGGTGTTGTGGGTGTCGACCCCTACAGTGCCTTAGGAAACCCGTTAAGTGCTCCGCCTGGGGA
GTACGGTCGCAAGGCTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAA
TTCGATGCAACGCGAGGAACCTTACCTGGGCTTGAACTGTGGTGGACTGTCCATAAAAGTGGGCTTTCCC
TTCGGGGACCATCATAGAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCC
GCAACGAGCGCAACCCCTACTCTTAGTTGCCAGCGGTTAAAGCCGGGAACTCTAAGAGGACTGCCCCGGA
AATCGGGGAGGAAGGTGGGGATGACGTCAAGTCCTCATGGCCTTTATGTCCAGGGCTACACACGTGCTAC
AATGGGCGGTACAAAGGGCTGCGAAGCCGCGAGGCAGAGCTAATCTCAAGAAACCGTCCTCAGTTCGGAT
TGCAGTCTGCAACTCGACTGCATGAAGTTGGAATCGCTAGTAATCCCAGATCAGCATGCTGGGGTGAATA
CGTTCCCGGGCCTTGTACACACCGCCCGTCACATCACGAAAGTCGGCTGTACCGGAAAGCCGTGGGCTAA
CCCGCANA
```

**Note:**

- **16S ribosomal RNA gene** is a DNA sequence not yet to be transcribed, meaning it is NOT a RNA sequence.
- In case of any confusion, you can check the (LOCUS) in the GeneBank page → in this example it is noted as a **linear DNA**

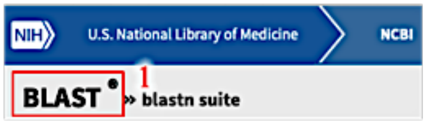
## To conduct the sequence alignment:

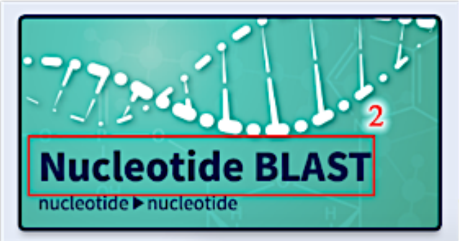
1) Navigate to the **NCBI BLASTN** Web form **OR** from the right menu of NCBI Nucleotide database choose <u>RUN BLAST</u>.

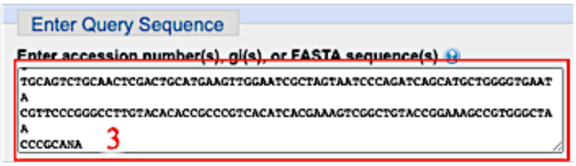
2) Paste the FASTA format of the DNA sequence into the Query window. You can also BLAST the sequence using the accession number.

3) In BLAST:

- *Since the search sequence is on a DNA sequence*, choose the "**Nucleotide collection (nr/nt)**" as the database to be searched.
- *To save lots of time for your searches*, restrict your search to "**bacteria (taxid:2)**" in the **organism field**.
- Optimize your search for "**Somewhat similar sequences (blastn)**" as the program to be used in the search. *This will show the similar sequences and will not restrict your search to the identical or different sequences.*
- Launch the search by clicking on the "**BLAST**" button.



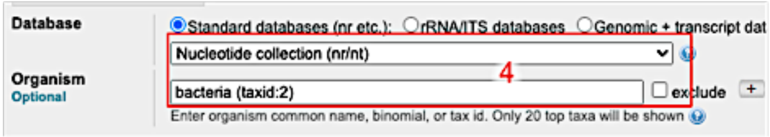




Enter Query Sequence

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

```
TGCAGTCTGCAACTCGACTGCATGAAGTTGGAATCGCTAGTAATCCCAGATCAGCATGCTGGGGTGAAT
A
CGTTCCCGGGCCTTGTACACACCGCCCGTCACATCACGAAAGTCGGCTGTACCGGAAAGCCGTGGGCTA
A
CCCGCANA
```



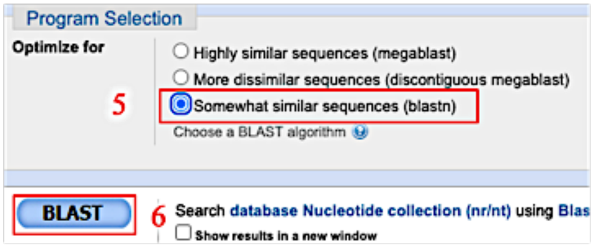
Database ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript dat

Nucleotide collection (nr/nt)

Organism Optional

bacteria (taxid:2) □ exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown



Program Selection

Optimize for

○ Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

● Somewhat similar sequences (blastn)

Choose a BLAST algorithm

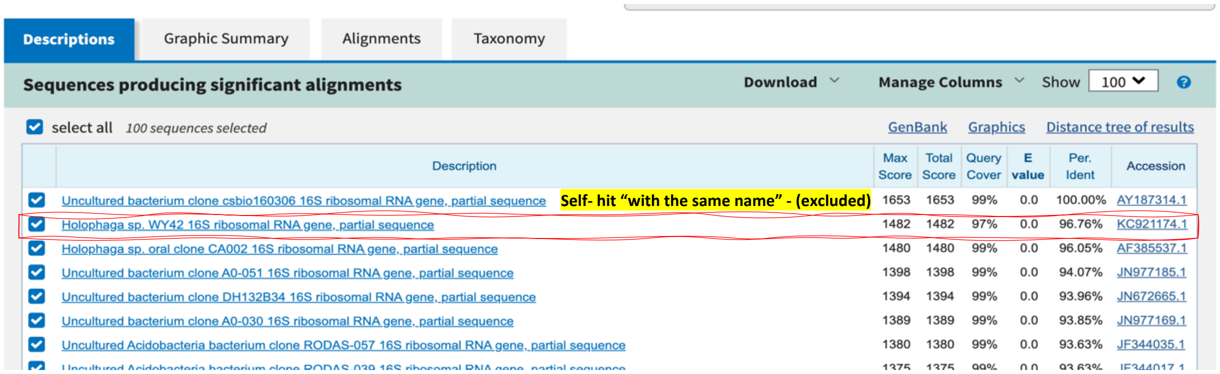
BLAST Search database Nucleotide collection (nr/nt) using Blas

□ Show results in a new window

## To draw conclusions as to what kind of bacteria the sequence came from:

Survey the results **graphic**, **table**, and **alignments** to assign the unknown sequence to an organism. You may not find 100% identity between your query and the hits, except for the **self-hit**.

- Note that the first hit may also be an unknown so you should examine all the hits before excluding any results.



**Sequences producing significant alignments**

☑ select all *100 sequences selected*

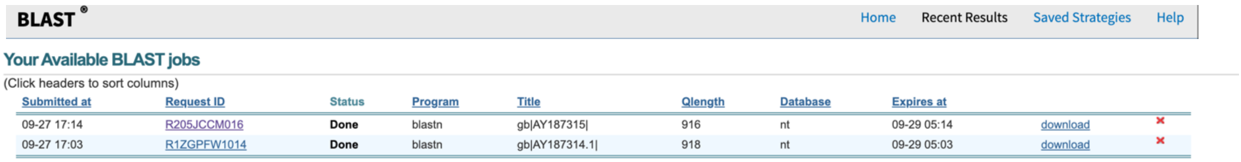
| | Description | Max Score | Total Score | Query Cover | E value | Per. Ident | Accession |
|---|---|---|---|---|---|---|---|
| ☑ | Uncultured bacterium clone csbio160306 16S ribosomal RNA gene, partial sequence **Self- hit "with the same name" - (excluded)** | 1653 | 1653 | 99% | 0.0 | 100.00% | AY187314.1 |
| ☑ | Holophaga sp. WY42 16S ribosomal RNA gene, partial sequence | 1482 | 1482 | 97% | 0.0 | 96.76% | KC921174.1 |
| ☑ | Holophaga sp. oral clone CA002 16S ribosomal RNA gene, partial sequence | 1480 | 1480 | 99% | 0.0 | 96.05% | AF385537.1 |
| ☑ | Uncultured bacterium clone A0-051 16S ribosomal RNA gene, partial sequence | 1398 | 1398 | 99% | 0.0 | 94.07% | JN977185.1 |
| ☑ | Uncultured bacterium clone DH132B34 16S ribosomal RNA gene, partial sequence | 1394 | 1394 | 99% | 0.0 | 93.96% | JN672665.1 |
| ☑ | Uncultured bacterium clone A0-030 16S ribosomal RNA gene, partial sequence | 1389 | 1389 | 99% | 0.0 | 93.85% | JN977169.1 |
| ☑ | Uncultured Acidobacteria bacterium clone RODAS-057 16S ribosomal RNA gene, partial sequence | 1380 | 1380 | 99% | 0.0 | 93.63% | JF344035.1 |
| ☑ | Uncultured Acidobacteria bacterium clone RODAS-039 16S ribosomal RNA gene, partial sequence | 1375 | 1375 | 99% | 0.0 | 93.63% | JF344017.1 |

→ **The results are sorted by percent identity (Per. Ident) from the highest percentage to the lowest. So, according to the table of results above, the next highest identity percentage with a known bacterial strain is our answer. For this accession number → it is *Holophaga.***

**Note:**

✓ For the remaining four accession numbers, open up additional Internet browser windows and launch the other searches, individual windows of results will be returned within a few minutes. Be sure to stay organized and record your conclusions of the bacterial strain for each accession number.

✓ In case you needed to view recent searches, Check the **Recent Results** hyperlink at the top of the page.



**Your Available BLAST jobs**

(Click headers to sort columns)

| Submitted at | Request ID | Status | Program | Title | Qlength | Database | Expires at | | |
|---|---|---|---|---|---|---|---|---|---|
| 09-27 17:14 | R205JCCM016 | **Done** | blastn | gb\|AY187315\| | 916 | nt | 09-29 05:14 | download | ✖ |
| 09-27 17:03 | R1ZGPFW1014 | **Done** | blastn | gb\|AY187314.1\| | 918 | nt | 09-29 05:03 | download | ✖ |

## Exercise 2: RuBisCO

It is often said that ribulose bisphosphate carboxylase (RuBisCO) is the most abundant protein on the planet. This enzyme is part of the Calvin cycle and is the key enzyme in the incorporation of carbon from carbon dioxide into living organisms. It is part of an enzyme complex found in plants, terrestrial or aquatic, and most probably played an important role in the development of our atmosphere and life on earth.

**Arabidopsis thaliana**, a member of the mustard family, is an important model system for higher plants. It is easily cultivated in the laboratory, undergoes rapid development, and produces a large number of seeds, making it amenable to genetic studies. Although not important agronomically, Arabidopsis has provided fundamental knowledge of plant biology and it was the first plant genome to be sequenced in 2000.

In this exercise, you will identify members of the RuBisCO gene family in Arabidopsis by running BLASTN for **Arabidopsis RuBisCO small chain subunit 1b mRNA using its accession number (NM_123204).**

**First,** you will need to retrieve the sequence from the **NCBI GenBank (Nucleotide database)** then copy the **FASTA format** of the sequence and paste it the sequence into the Query window.

**Or**

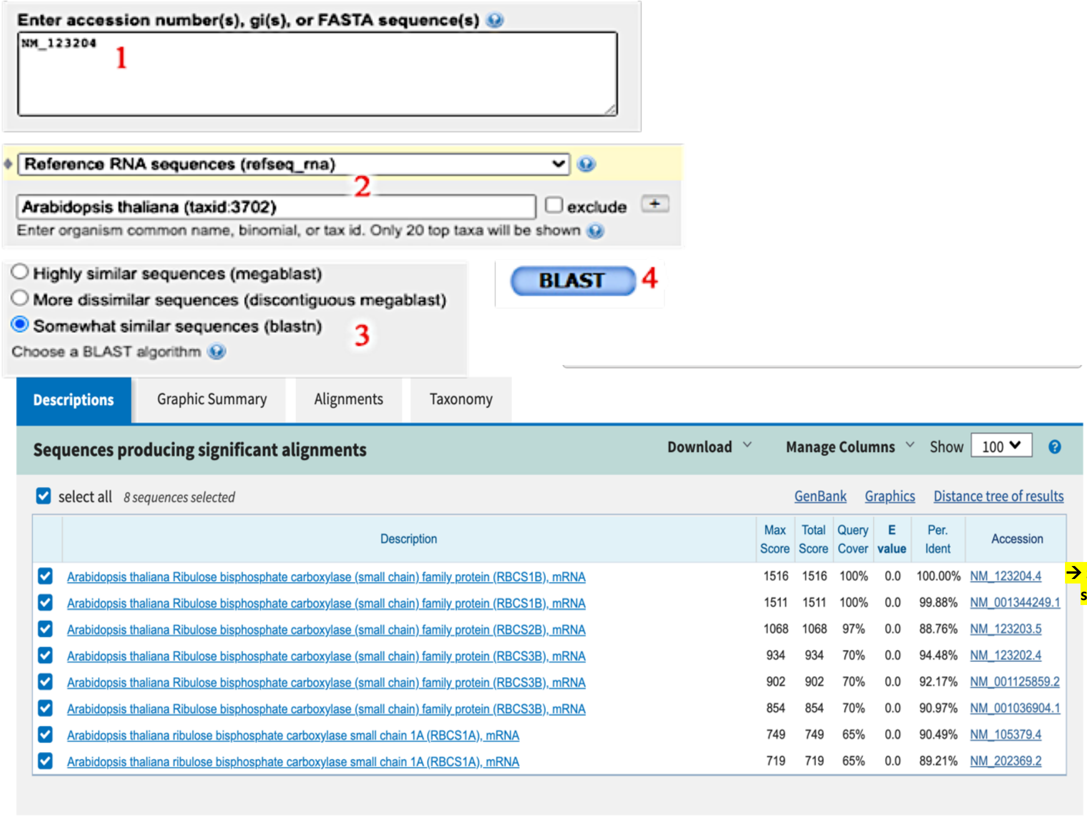
BLAST the sequence using the accession number.

**Note:**

- From the (LOCUS) in the GeneBank page → in this example the accession number is noted as a **mRNA linear**

**To distinguish members of the RuBisCO gene family in Arabidopsis:**

- *Since the search sequence is on an RNA sequence,* set the database to "**Reference RNA sequences (refseq_rna)**" and restrict the organism to "**Arabidopsis thaliana (taxid:3702)**."
- Set the program selection to "**Somewhat similar sequences (blastn)**" and click on the "BLAST" button to launch the search.
- When the results are returned, you should now utilize the graphic, table, and alignments to identify the **family members.**



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

NM_123204

Reference RNA sequences (refseq_rna)

Arabidopsis thaliana (taxid:3702) □ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)
Choose a BLAST algorithm

Sequences producing significant alignments

☑ select all *8 sequences selected*

| | Description | Max Score | Total Score | Query Cover | E value | Per. Ident | Accession |
|---|---|---|---|---|---|---|---|
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS1B), mRNA | 1516 | 1516 | 100% | 0.0 | 100.00% | NM_123204.4 |
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS1B), mRNA | 1511 | 1511 | 100% | 0.0 | 99.88% | NM_001344249.1 |
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS2B), mRNA | 1068 | 1068 | 97% | 0.0 | 88.76% | NM_123203.5 |
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS3B), mRNA | 934 | 934 | 70% | 0.0 | 94.48% | NM_123202.4 |
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS3B), mRNA | 902 | 902 | 70% | 0.0 | 92.17% | NM_001125859.2 |
| ☑ | Arabidopsis thaliana Ribulose bisphosphate carboxylase (small chain) family protein (RBCS3B), mRNA | 854 | 854 | 70% | 0.0 | 90.97% | NM_001036904.1 |
| ☑ | Arabidopsis thaliana ribulose bisphosphate carboxylase small chain 1A (RBCS1A), mRNA | 749 | 749 | 65% | 0.0 | 90.49% | NM_105379.4 |
| ☑ | Arabidopsis thaliana ribulose bisphosphate carboxylase small chain 1A (RBCS1A), mRNA | 719 | 719 | 65% | 0.0 | 89.21% | NM_202369.2 |

**→ Self- hit "with the same name" - (NOT excluded since all family members should be identified)**

The Reference RNA database should not have any redundancy but three family members **(RBCS1B RBCS3B, RBCS1A)** are repeated due to alternatively spliced mRNAs.

**→So, there are 8 family members (shown in the picture above)**

- To describe and understand the major differences between these **family member transcripts**. Create a table with a listing of the **names** of **family member transcripts** and their **accession numbers**, their **mRNA length**, and the coordinates of the coding regions (**CDS**).

| Name of transcript | Accession number | mRNA length | CDS |
|---|---|---|---|
| (RBCS1B), mRNA | NM_123204.4 | 840 bp | 30..575 |
| (RBCS1B), mRNA<br>⋮ | NM_001344249.1 | 1894 bp | 166..711 |