Opening Thought Questions – 2/28

What are the consequences for phytoplankton in the following scenarios?
i.e., when do we see phytoplankton blooms and which phytoplankton do we usually see first, second, third, etc?

- Tropical mixed layer is nutrient limited year-round
- Temperate mixed layer is nutrient limited in late-summer and fall
- Polar mixed layer develops in summer and is nutrient rich

Tropical mixed layer is nutrient limited year-round

- blooms are rare
- vertical stratification of species composition

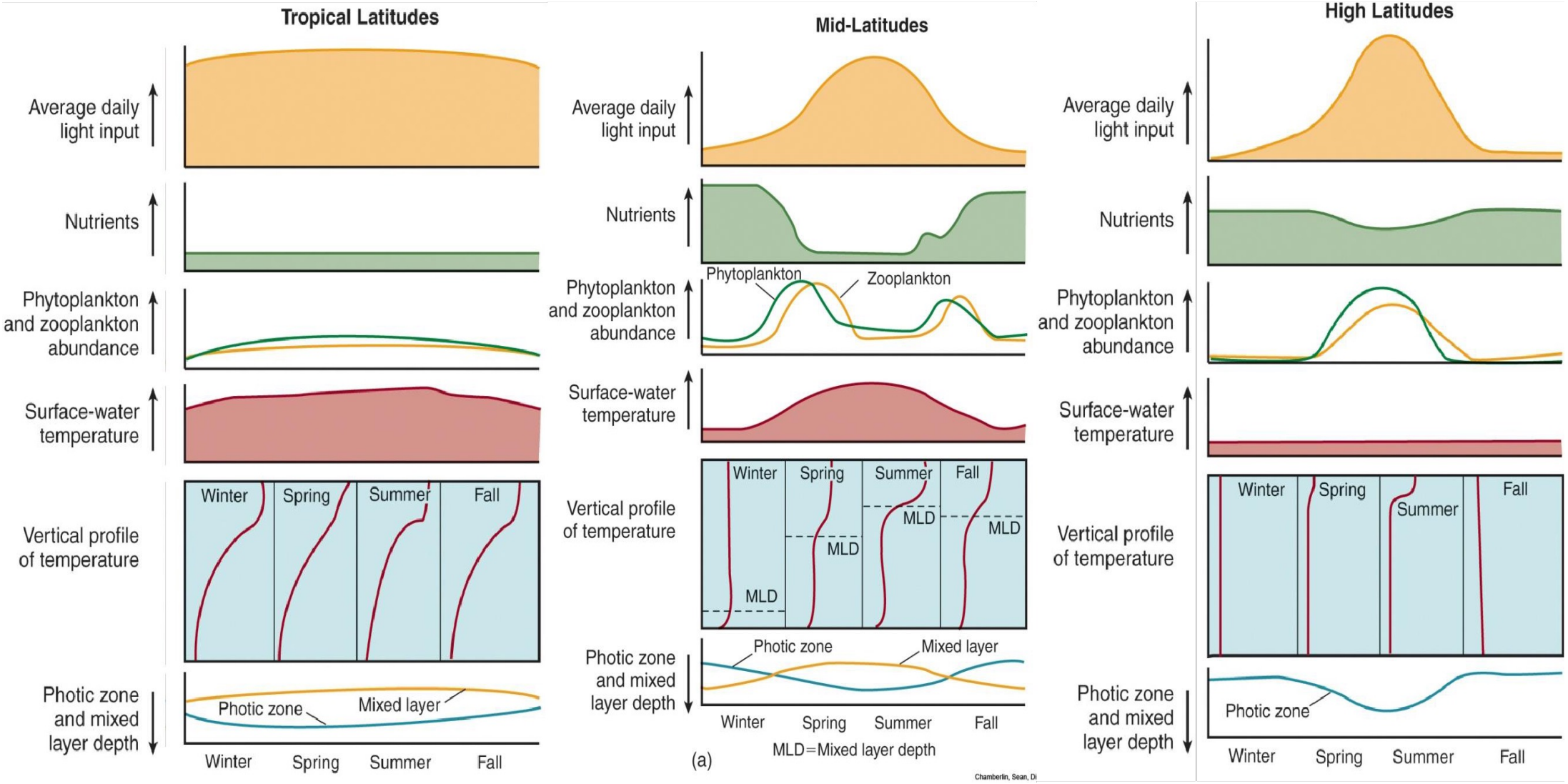
Temperate mixed layer is nutrient limited in late-summer and fall

- spring bloom, very predictable
- occasional fall bloom when strength of mixed layer decreased and critical depth is still deep
- Seasonal patterns of species “succession”

Polar mixed layer develops in summer and is nutrient rich

- more light limited
- one large summer bloom

Seasonal Changes at different latitudes

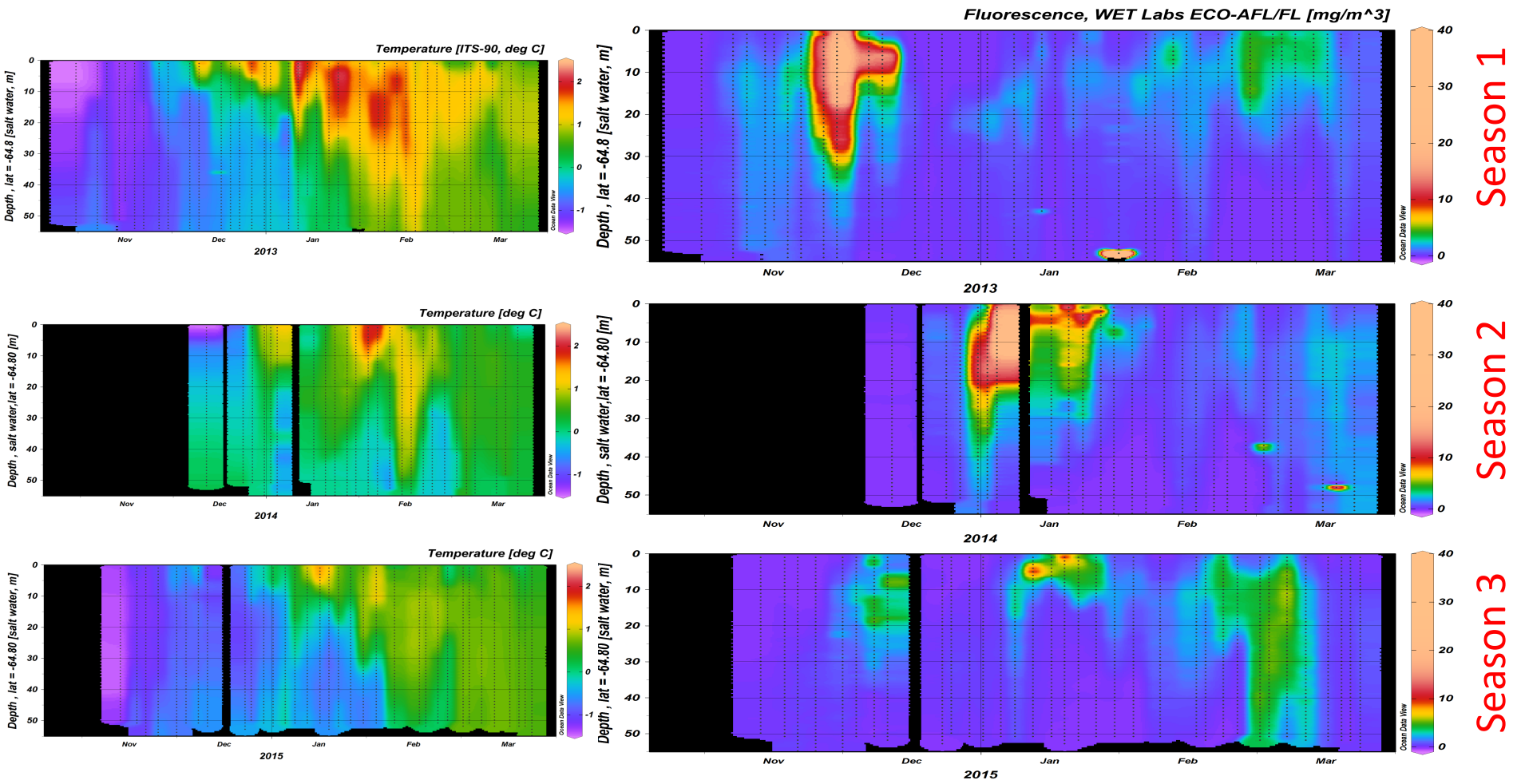


Remember these graphs from Phil's Antarctica work?

- Given your knowledge about the mixed layer (e.g., thermocline) and phytoplankton blooms, can you explain what is going on in these graphs and why?

Temperature

Fluorescence (proxy for biomass)

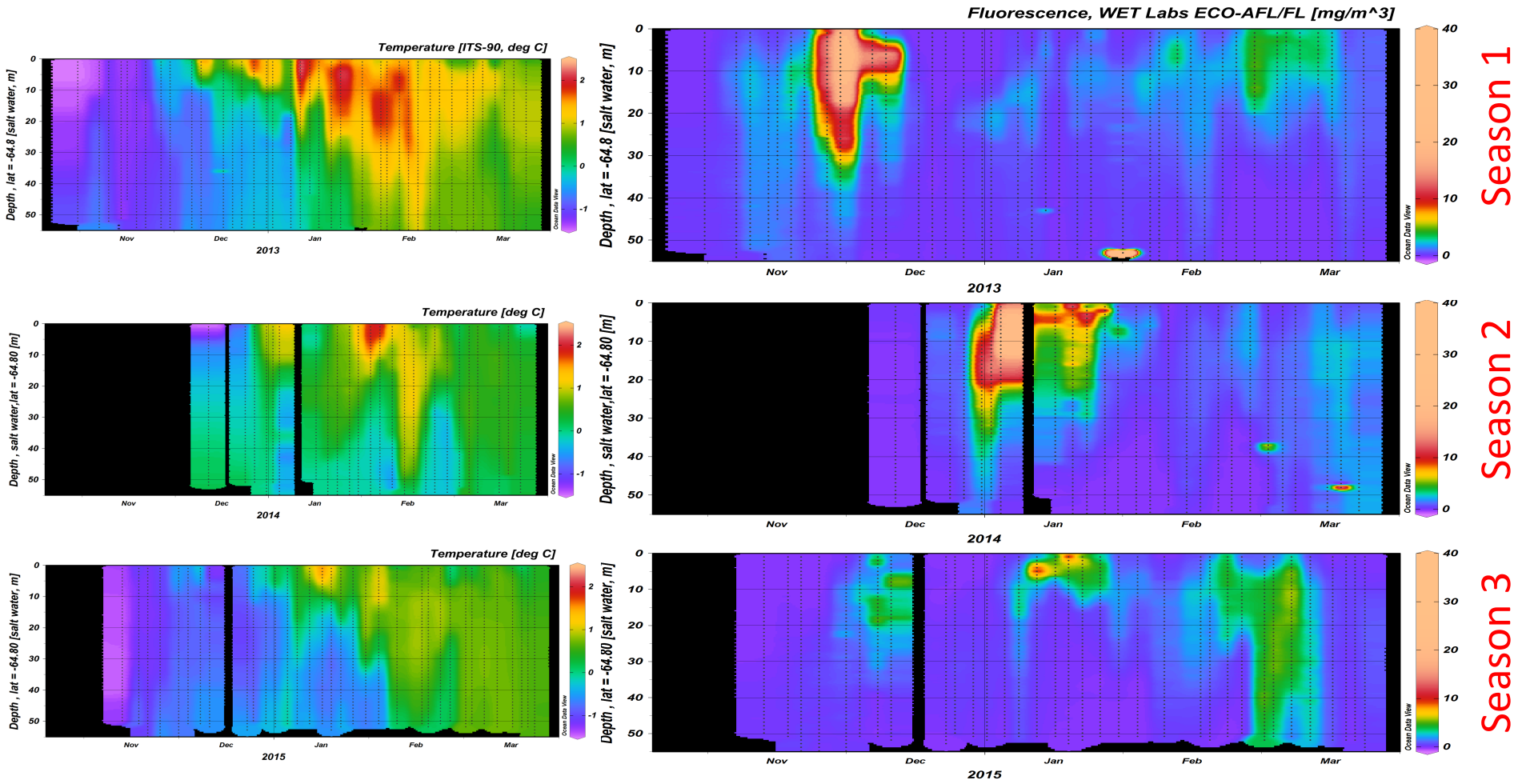


Polar mixed layer develops in summer and is nutrient rich

- one large summer bloom…unless climate is disrupted

Temperature

Fluorescence (proxy for biomass)



Thinking back to Dr. Sarah Smith's lecture, what are some key discoveries revolutionized the way we think about biological oceanography?

List up to three.

Marine Bioinformatics
Spring 2023
Lab and Homework Review

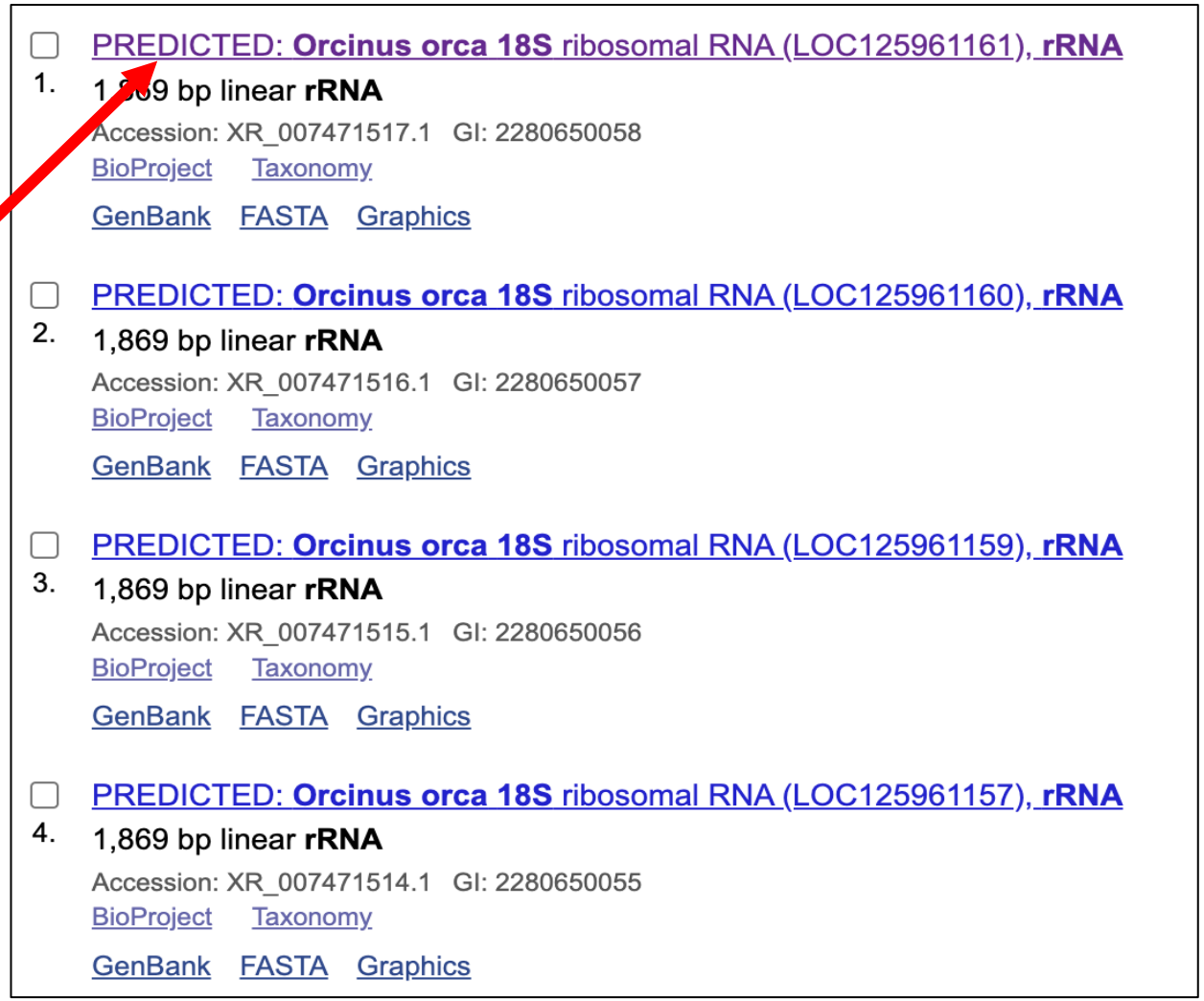
Exam: Tuesday March 7

Lab 1

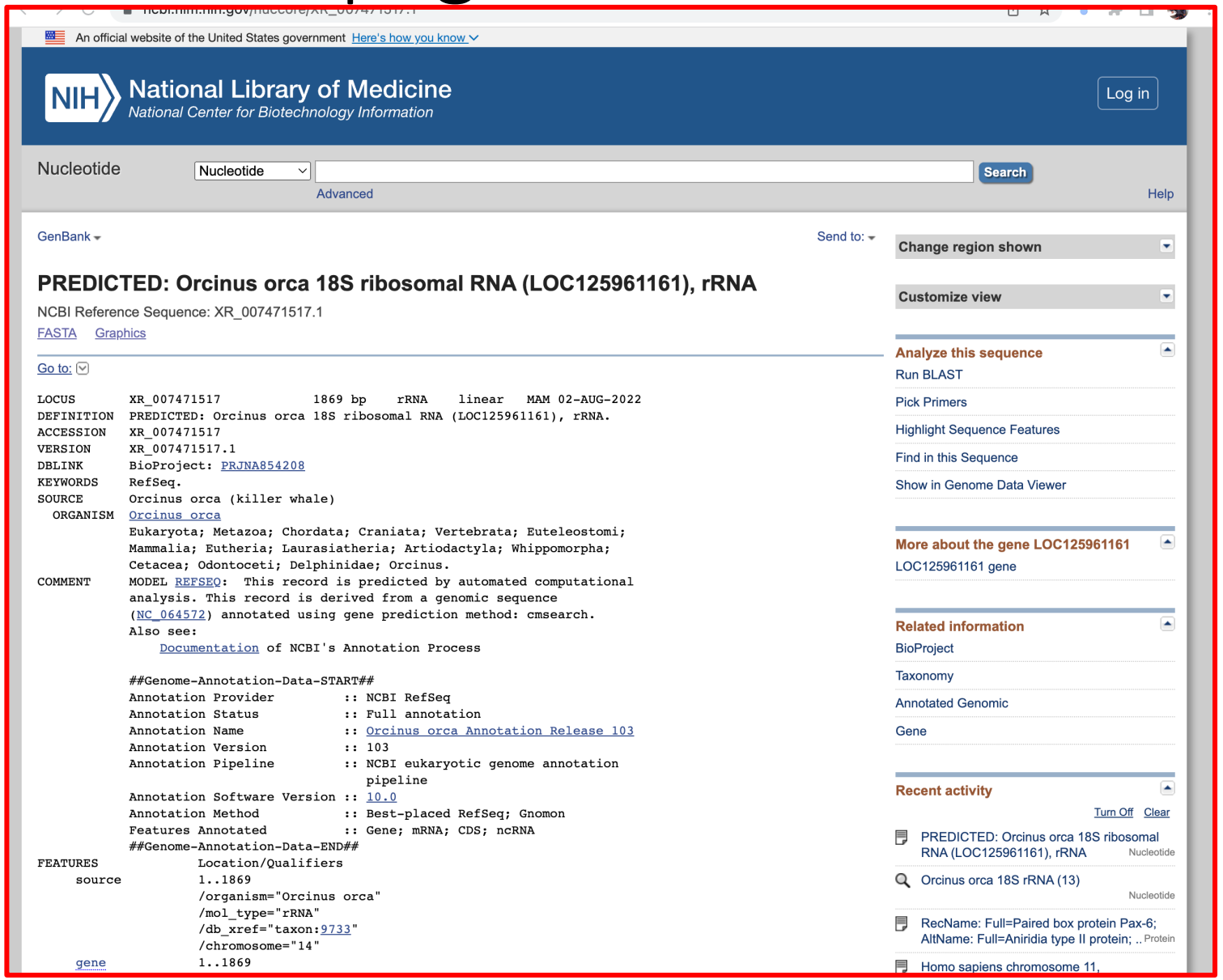
- The big question: What marine organisms are most closely related to orcas?
- The skills:
  - BLAST
    - Doing it
    - Interpreting the results
  - GP pages

(1) … browse to https://www.ncbi.nlm.nih.gov/ and type “Orcinus orca 18S rRNA” into the search field

CLICK



(2) Look at the GP page



Where's the sequence?

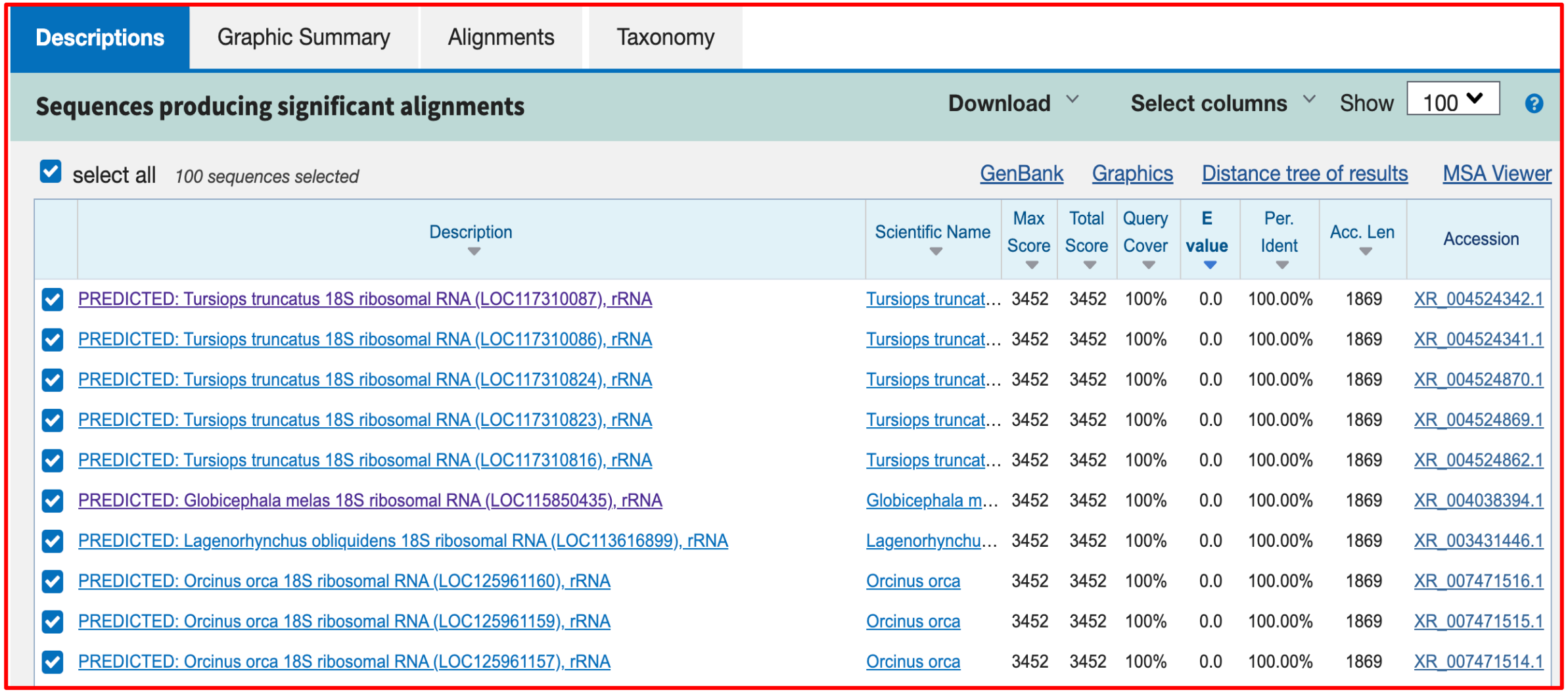
End of page

```
ORIGIN
        1 tacctggttg atcctgccag tagcatatgc ttgtctcaaa gattaagcca tgcatgtcta
       61 agtacgcacg gccggtacag tgaaactgcg aatggctcat taaatcagtt atggttcctt
      121 tggtcgctcg ctcctctcct acttggataa ctgtggtaat tctagagcta atacatgccg
      181 acgggcgctg acccccttcg cgggggggat gcgtgcattt atcagatcaa aaccaacccg
      241 gtcagcctcc ctccggcccc ggccgggggt cgggcgccgg cggctttggt gactctagat
      301 aacctcgggc cgatcgcacg ccccccgtgg cggcgacgac ccattcgaac gtctgcccta
      361 tcaactttcg atggtagtcg ccgtgcctac catggtgacc acgggtgacg gggaatcagg
      421 gttcgattcc ggagagggag cctgagaaac ggctaccaca tccaaggaag gcagcaggcg
      481 cgcaaattac ccactcccga cccggggagg tagtgacgaa aaataacaat acaggactct
      541 ttcgaggccc tgtaattgga atgagtccac tttaaatcct ttcgcgagga tccattggag
      601 ggcaagtctg gtgccagcag ccgcggtaat tccagctcca atagcgtata ttaaagttgc
      661 tgcagttaaa aagctcgtag ttggatcttg ggagcgggcg ggcggtccgc cgcgaggcga
      721 gccaccgccc gtccccgccc cttgcctctc ggcgcccccc cgatgctctt agctgagtgt
      781 cccgcggggc ccgaagcgtt tactttgaaa aaattagagt gttcaaagca ggcccgagcc
      841 gcctggatac cgcagctagg aataatggaa taggaccgcg gttctatttt gttggttttc
      901 ggaactgagg ccatgattaa gagggacggc cgggggcatt cgtattgcgc cgctagaggt
      961 gaaattcttg gaccggcgca agacggacca gagcgaaagc atttgccaag aatgttttca
     1021 ttaatcaaga acgaaagtcg gaggttcgaa gacgatcaga taccgtcgta gttccgacca
     1081 taaacgatgc cgactggcga tgcggcggcg ttattcccat gacccgccgg gcagcttccg
     1141 ggaaaccaaa gtctttgggt tccgggggga gtatggttgc aaagctgaaa cttaaaggaa
     1201 ttgacggaag ggcaccacca ggagtggagc ctgcggctta atttgactca acacgggaaa
     1261 cctcacccgg cccggacacg gacaggattg acagattgat agctctttct cgattccgtg
     1321 ggtggtggtg catggccgtt cttagttggt ggagcgattt gtctggttaa ttccgataac
     1381 gaacgagact ctggcatgct aactagttac gcgacccccg agcggtcggc gtcccccaac
     1441 ttcttagagg gacaagtggc gttcagccac ccgagattga gcaataacag gtctgtgatg
     1501 cccttagatg tccggggctg cacgcgcgct acactgactg gctcagcgtg tgcctaccct
     1561 acgccggcag gcgcgggtaa cccgttgaac cccattcgtg atggggatcg gggattgcaa
     1621 ttattcccca tgaacgagga attcccagta agtgcgggtc ataagcttgc gttgattaag
     1681 tccctgccct ttgtacacac cgcccgtcgc tactaccgat tggatggttt agtgaggccc
     1741 tcggatcggc cccgccgggg tcggcccacg gccctggcgg agcgctgaga agacggtcga
     1801 acttgactat ctagaggaag taaaagtcgt aacaaggttt ccgtaggtga acctgcggaa
     1861 ggatcatta
//
```

Additional information

```
LOCUS       XR_007471517            1869 bp    rRNA    linear   MAM 02-AUG-2022
DEFINITION  PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961161), rRNA.
ACCESSION   XR_007471517
VERSION     XR_007471517.1
DBLINK      BioProject: PRJNA854208
KEYWORDS    RefSeq.
SOURCE      Orcinus orca (killer whale)
  ORGANISM  Orcinus orca
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Whippomorpha;
            Cetacea; Odontoceti; Delphinidae; Orcinus.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NC_064572) annotated using gene prediction method: cmsearch.
```

(3) Nucleotide BLAST the query



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310087), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524342.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310086), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524341.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310824), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524870.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310823), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524869.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310816), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524862.1
PREDICTED: Globicephala melas 18S ribosomal RNA (LOC115850435), rRNA	Globicephala m…	3452	3452	100%	0.0	100.00%	1869	XR_004038394.1
PREDICTED: Lagenorhynchus obliquidens 18S ribosomal RNA (LOC113616899), rRNA	Lagenorhynchu…	3452	3452	100%	0.0	100.00%	1869	XR_003431446.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961160), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471516.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961159), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471515.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961157), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471514.1

(4) Range of E-values: all zero point zero (remember what that means)

(5) Top non-Tursiops hit is probably *Globicephala melas*

Descriptions | Graphic Summary | Alignments | Taxonomy

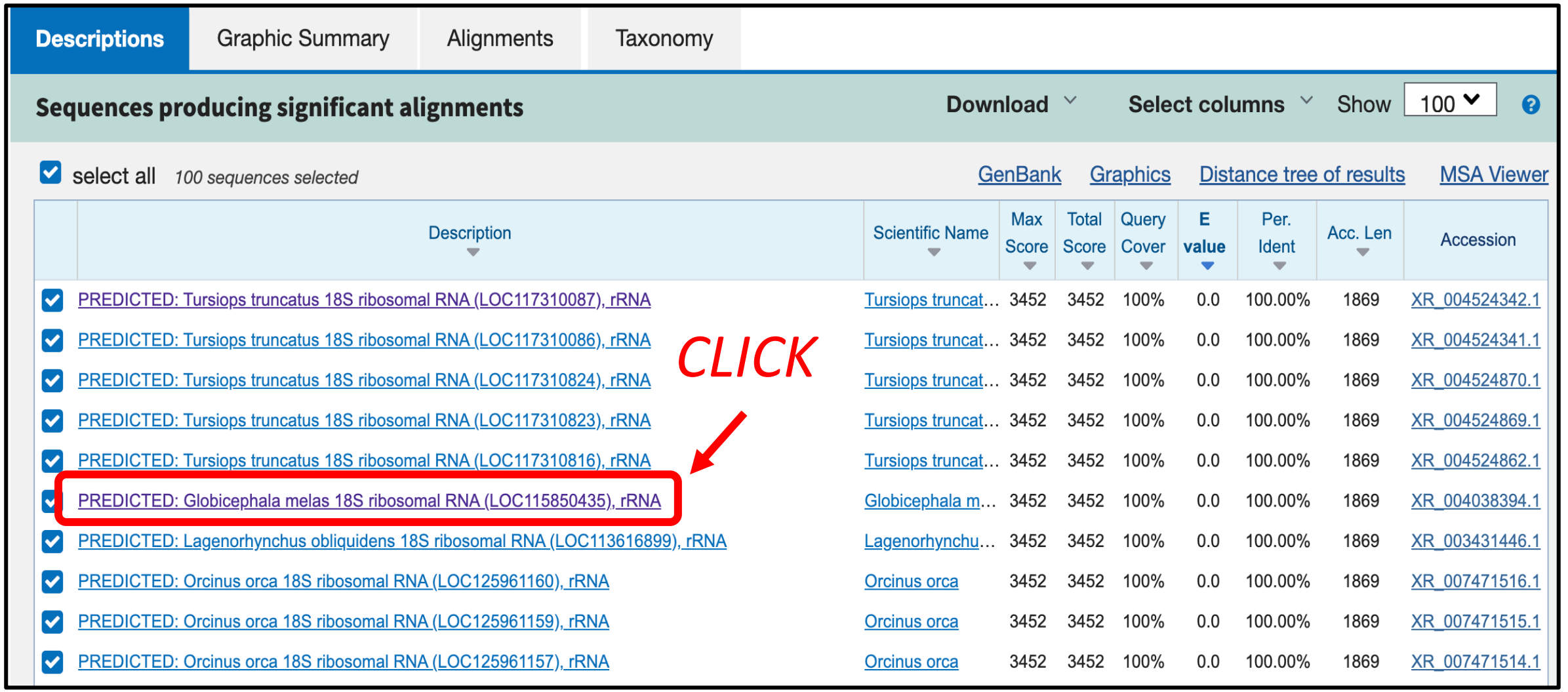
Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310087), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524342.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310086), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524341.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310824), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524870.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310823), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524869.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310816), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524862.1
PREDICTED: Globicephala melas 18S ribosomal RNA (LOC115850435), rRNA	Globicephala m…	3452	3452	100%	0.0	100.00%	1869	XR_004038394.1
PREDICTED: Lagenorhynchus obliquidens 18S ribosomal RNA (LOC113616899), rRNA	Lagenorhynchu…	3452	3452	100%	0.0	100.00%	1869	XR_003431446.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961160), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471516.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961159), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471515.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961157), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471514.1

Globicephala melas: the long-finned pilot whale

(6) The *Tursiops/Globicephala alignment*



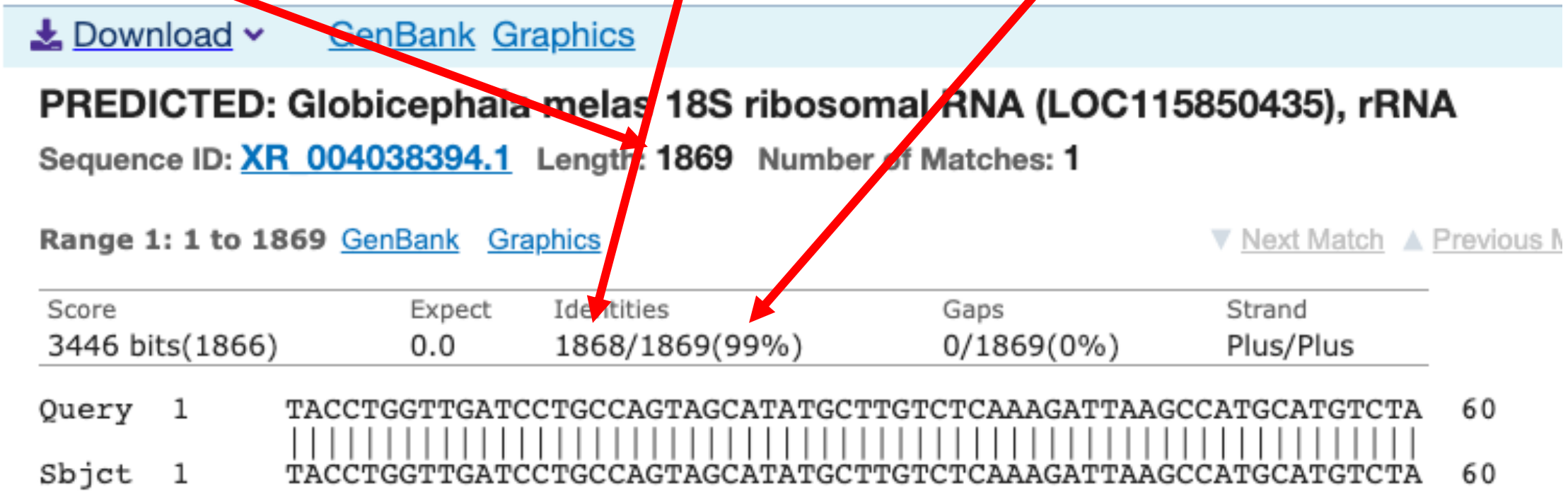
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310087), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524342.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310086), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524341.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310824), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524870.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310823), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524869.1
PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310816), rRNA	Tursiops truncat…	3452	3452	100%	0.0	100.00%	1869	XR_004524862.1
PREDICTED: Globicephala melas 18S ribosomal RNA (LOC115850435), rRNA	Globicephala m…	3452	3452	100%	0.0	100.00%	1869	XR_004038394.1
PREDICTED: Lagenorhynchus obliquidens 18S ribosomal RNA (LOC113616899), rRNA	Lagenorhynchu…	3452	3452	100%	0.0	100.00%	1869	XR_003431446.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961160), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471516.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961159), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471515.1
PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961157), rRNA	Orcinus orca	3452	3452	100%	0.0	100.00%	1869	XR_007471514.1

(6) The *Tursiops/Globicephala alignment*

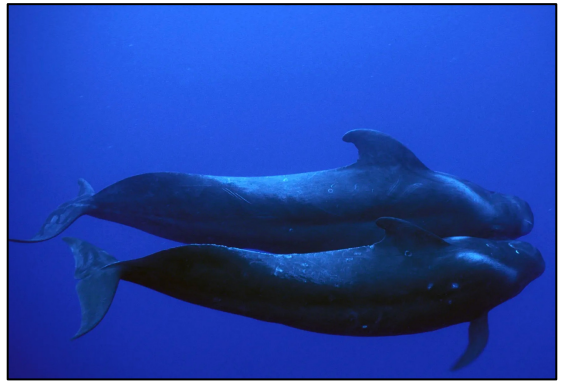
Length of subject sequence

Identity columns

% identity



(7) Back to “Descriptions”: non-query subjects (i.e. not *Orcinus orca)*



Globicephala melas
long-finned pilot whale

Lagenorhynchus obliquidens
Pacific white-sided dolphin

Cavia porcellus
Guinea pig

Ovis aries
Sheep

Elephas maximus indicus
Indian elephant

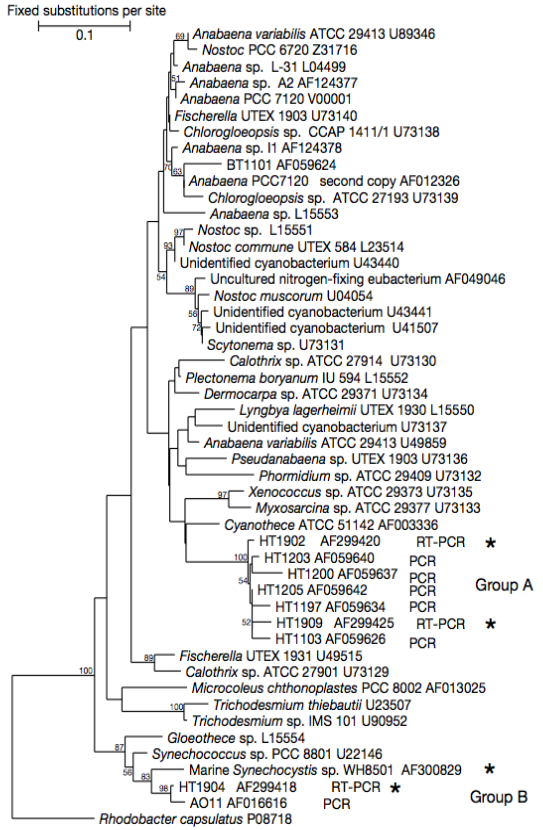
(8) Was that tedious? ➔ I thought it was

(9) Which pix surprised me?

If information is impossible to believe,
question the source and your interpretation.

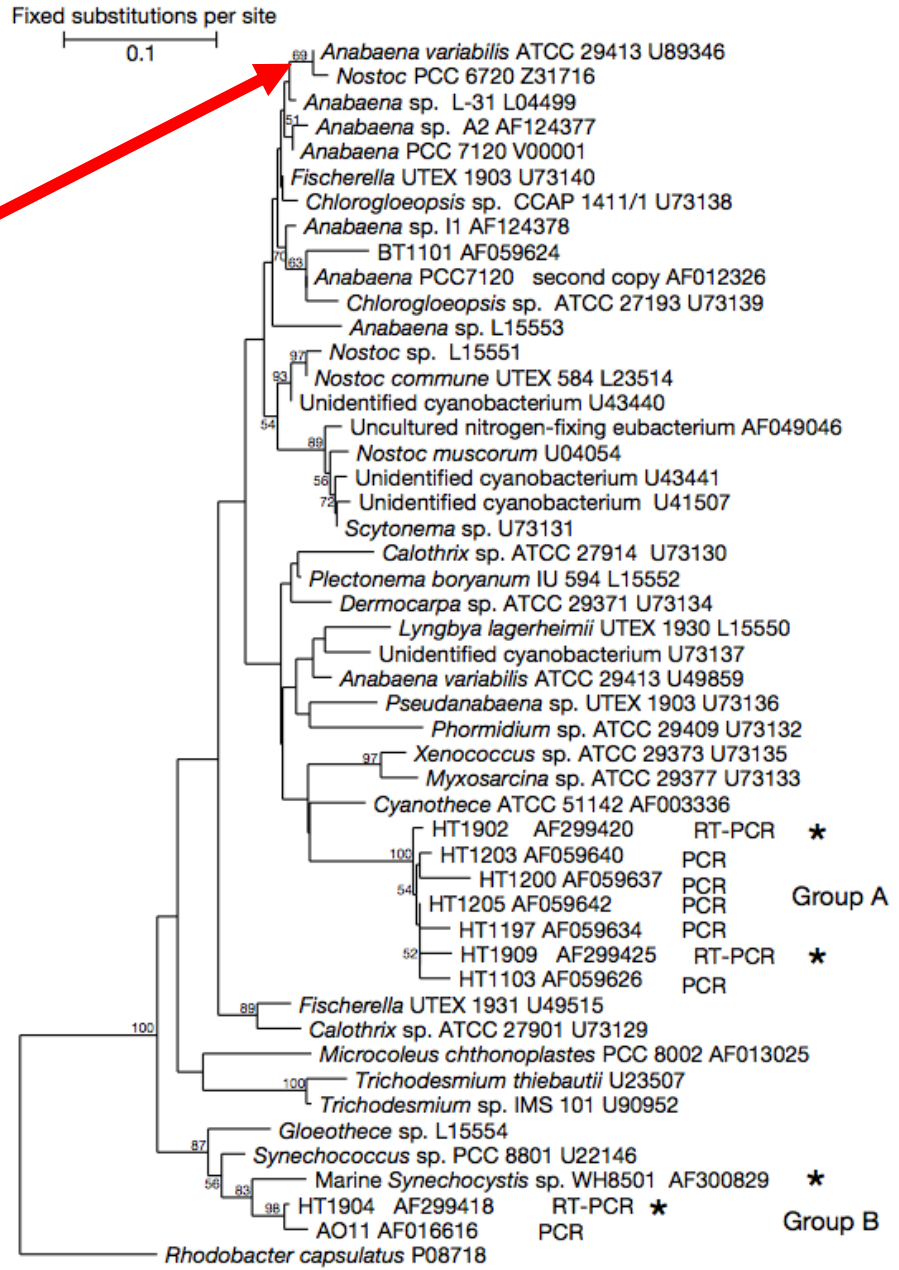
Lab 2

- The big question: Is the original Zehr nifH tree accurate?
- The skills:
  - Editing a fasta file
  - Multiple Sequence Alignment (MSA)
  - Contructing a tree



(1) Look at the tree

Top OTU is *Anabaena viriabilis*,
Accession # U89346



The GP page

Title

Anabaena variabilis dinitrogenase reductase (nifH) gene, complete cds

GenBank: U89346.1

FASTA Graphics

Go to:

```
LOCUS       AVU89346                1450 bp    DNA     linear   BCT 15-AUG-1997
DEFINITION  Anabaena variabilis dinitrogenase reductase (nifH) gene, complete
            cds.
```

Record date

Scientific name matches

```
KEYWORDS    .
SOURCE      Trichormus variabilis ATCC 29413 (Nostoc sp. PCC 7937)
  ORGANISM  Trichormus variabilis ATCC 29413
            Bacteria; Cyanobacteriota; Cyanophyceae; Nostocales; Nostocaceae;
            Trichormus.
REFERENCE   1  (bases 1 to 1450)
  AUTHORS   Thiel,T., Lyons,E.M. and Erker,J.C.
  TITLE     Characterization of genes for a second Mo-dependent nitrogenase in
            the cyanobacterium Anabaena variabilis
  JOURNAL   J. Bacteriol. 179 (16), 5222-5225 (1997)
   PUBMED   9260968
REFERENCE   2  (bases 1 to 1450)
  AUTHORS   Thiel,T.
```

(2) The fasta file

Defline is accurate

```
>U89346.1 Anabaena variabilis dinitrogenase reductase (nifH) gene, complete cds
CGATATTGTCAAAGTAGTACTGCAAGGCGCGTGTGGCTCCTGTTCTAGTAGTACAGCCACCTTGAAAATA
…
```

Sequences match

GP Page

```
ORIGIN
        1 cgatattgtc aaagtagtac tgcaaggcgc gtgtggctcc tgttctagta gtacagccac
       61 cttgaaaata gcgattgaat ccagattacg cgatcgcatt aatcccagcc tagtagtaga
      121 agcagtttag tcattagtca ttagtcatta gtcaatggtc attagtcaac agtgaaaaat
```

Adding the missing records: have to remove the numbers and blank spaces

```
  1 tctactcgct tgatccttaa ctgtaaagcg catgtcacag ttctacactt agccgcagaa
 61 cggggttctg ttgaagatat agaactcgaa gatgtactgc tcacagggtt tgaagacatc
121 aaatgcgtag aatcaggtgg tcctgaacct ggcgtaggat gcgctggtcg tgggattatc
181 actgccatca acttccttga agaagaagga gcttacgaag acatagattt cgtatcctac
241 gacgtattag gggacgttgt ctgcggtggt ttcgctatgc ctatccgtga aggaaaagca
301 caagaaatct acatcgtaac ctct
```

>L15554.1 Gloeothece sp. PCC 6909 nitrogen fixation protein (nifH) gene, partial cds
TCTACTCGCTTGATCCTTAACTGTAAAGCGCATGTCACAGTTCTACACTTAGCCGCAGAACGGGGTTCTG
TTGAAGATATAGAACTCGAAGATGTACTGCTCACAGGGTTTGAAGACATCAAATGCGTAGAATCAGGTGG
TCCTGAACCTGGCGTAGGATGCGCTGGTCGTGGGATTATCACTGCCATCAACTTCCTTGAAGAAGAAGGA
GCTTACGAAGACATAGATTTCGTATCCTACGACGTATTAGGGGACGTTGTCTGCGGTGGTTTCGCTATGC
CTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGTAACCTCT.

No, I didn't enjoy editing it!

(3) Aligning the fasta using CLUSTAL

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

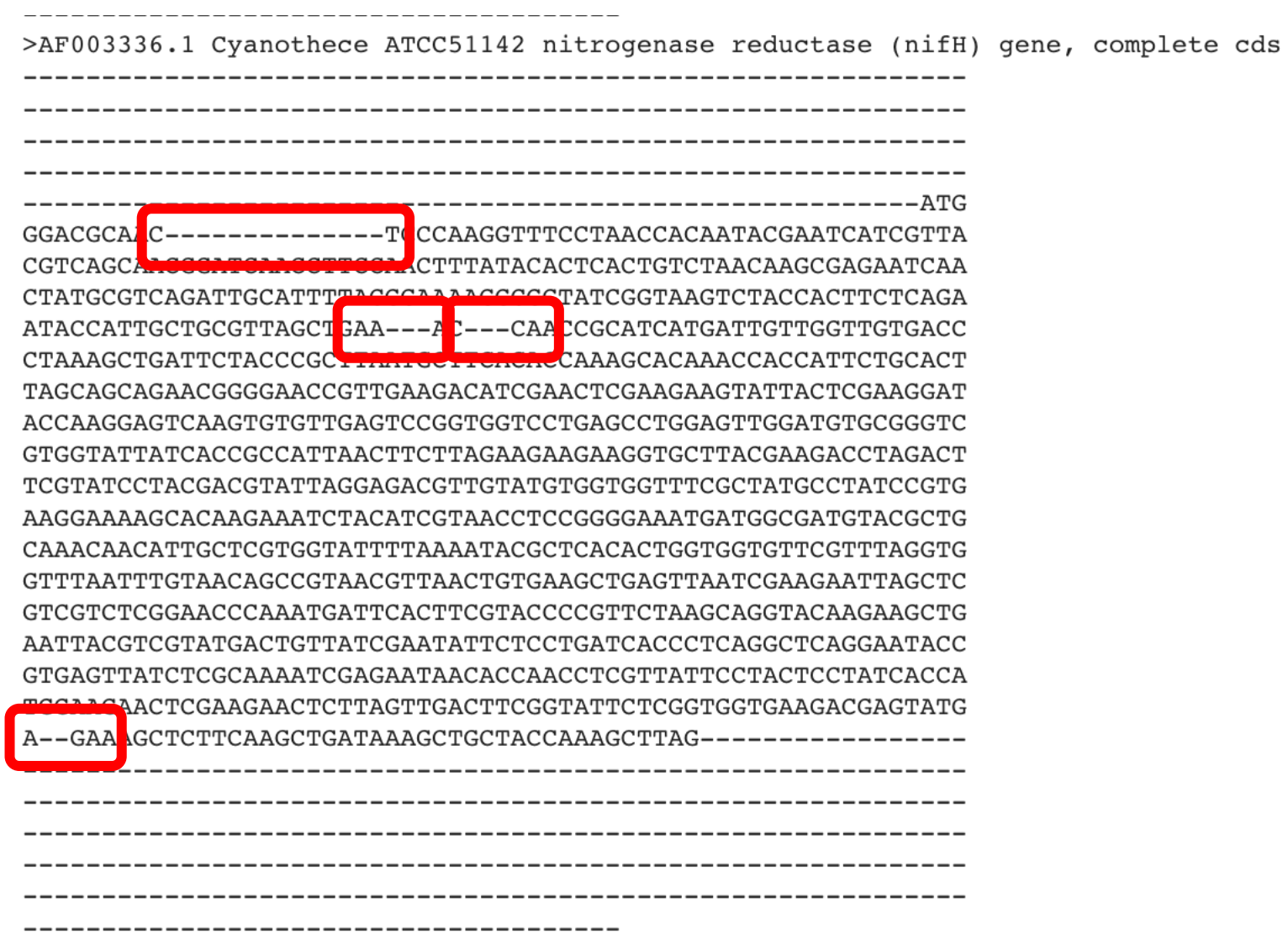
The default settings will fulfill the needs of most users.

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

An aligned sequence with no internal indels

```
>U73132.1 Phormidium sp. dinitrogenase reductase (nifH) gene, partial cds
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
----------TCCACTCGCCTGATGCTCCACAGCAAAGCCCAAACCTCCGTACTACAAC
TGGCTGCCGAACTCGGTGCCGTTGAAGATGTCGAACTTGACCAGGTGCTGCAAGAAGGCT
ATCGCGGCGTTAAGTGCGTCGAGTCCGGTGGTCCTGAGCCCGGCGTCGGCTGTGCCGGTC
GCGGCATTATCACTGCCATCAACTTCCTGGAAGAAGAGGGCGCTTACGAAGATCTTGATT
TCGTCTCCTACGACGTACTCGGTGACGTAGTTTGCGGCGGTTTCGCCATGCCCATTCGGG
AAGGCAAAGCCCAAGAAATCTACATTGTTGTCTCC----------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
```

An aligned sequence with internal indels



Can you tell, by looking at the MSA, which sequences are most closely related/similar to one another?

No, TMI, and the information is at a low level

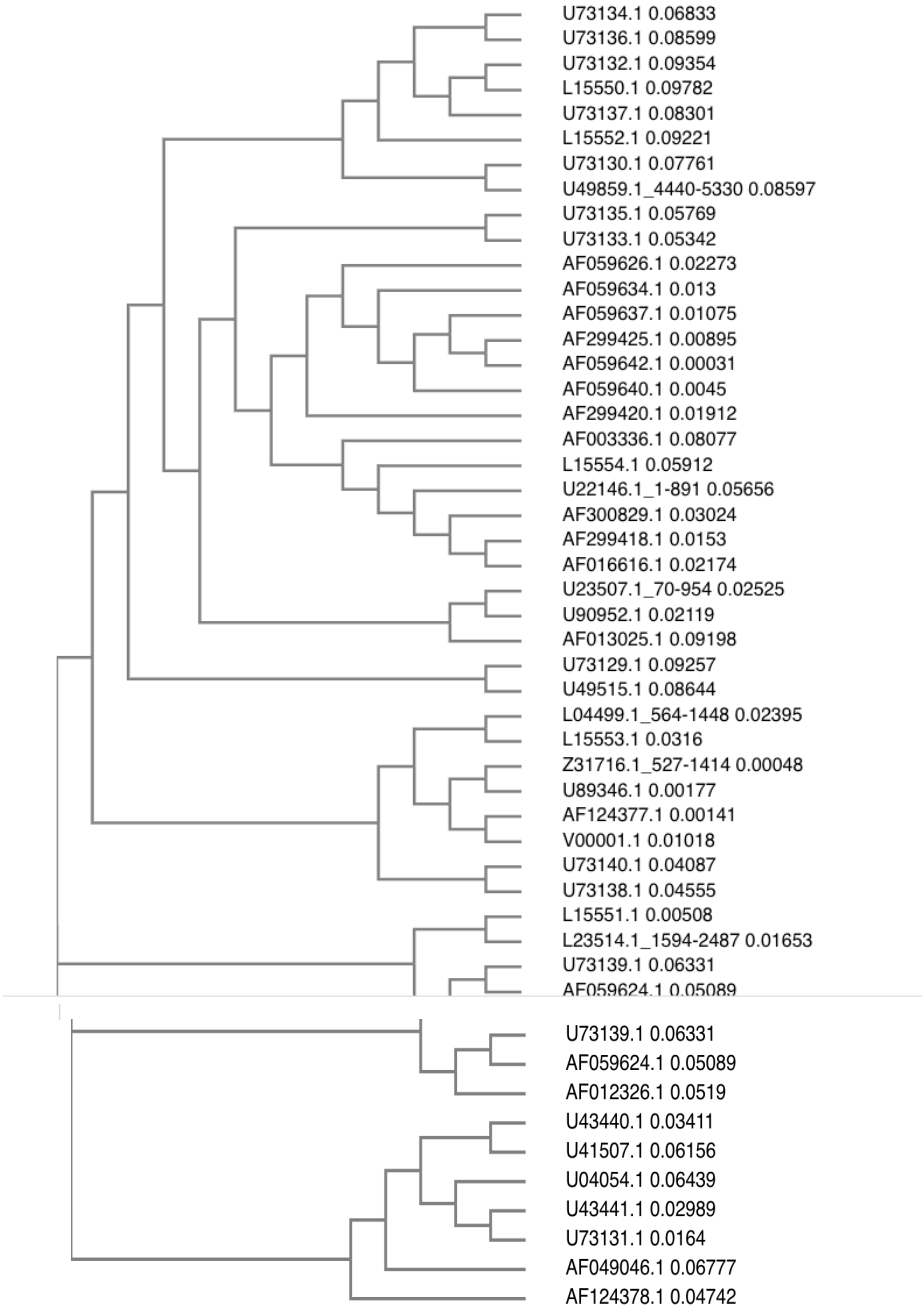
My tree

What are the accession numbers of the first 3 sequences in the MSA?

U73134.1, U73136.1, U73132.1

Are they the same as the first 3 OTUs in the Zehr tree?

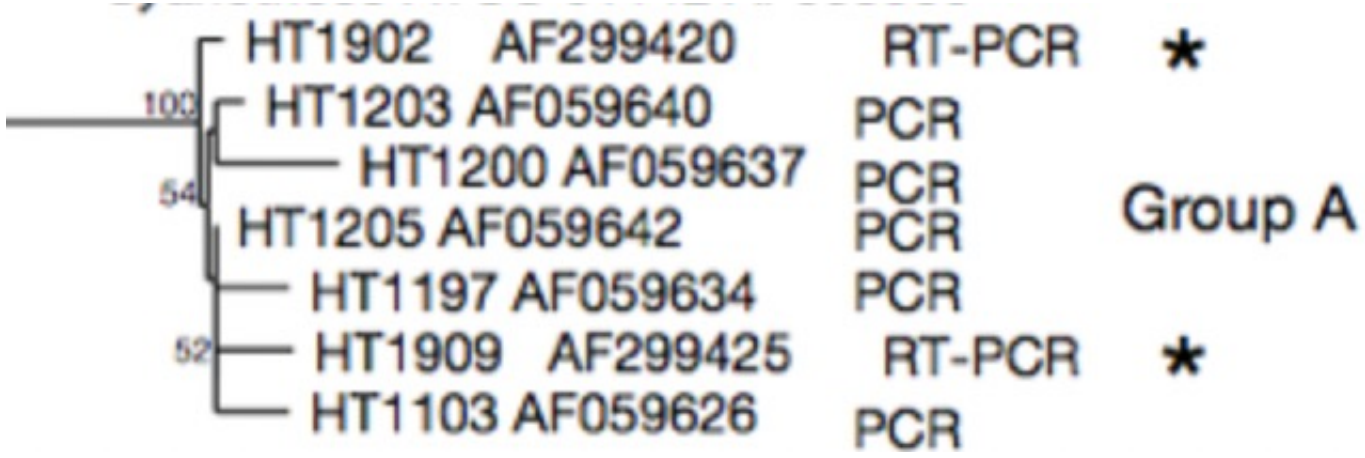
No, but that doesn't mean the trees are different



The 7 Original Group A sequences

What are the accession numbers of the 7 Group A sequences?
AF299420, AF059640, AF059637, AF059642, AF059634, AF299425, AF059626

Do they form a monophyletic group in the Clustal tree? Yes



(3) Expand the tree

Results: ARBitrator uses a two-step process composed of a broad collection of potential homologues followed by screening with a best hit strategy to conserved domains. 34 420 nifH sequences were identified in GenBank as of November 20, 2012. The false-positive rate is ~0.033%. ARBitrator rapidly updates a public nifH sequence database, and we show that it can be adapted for other genes.

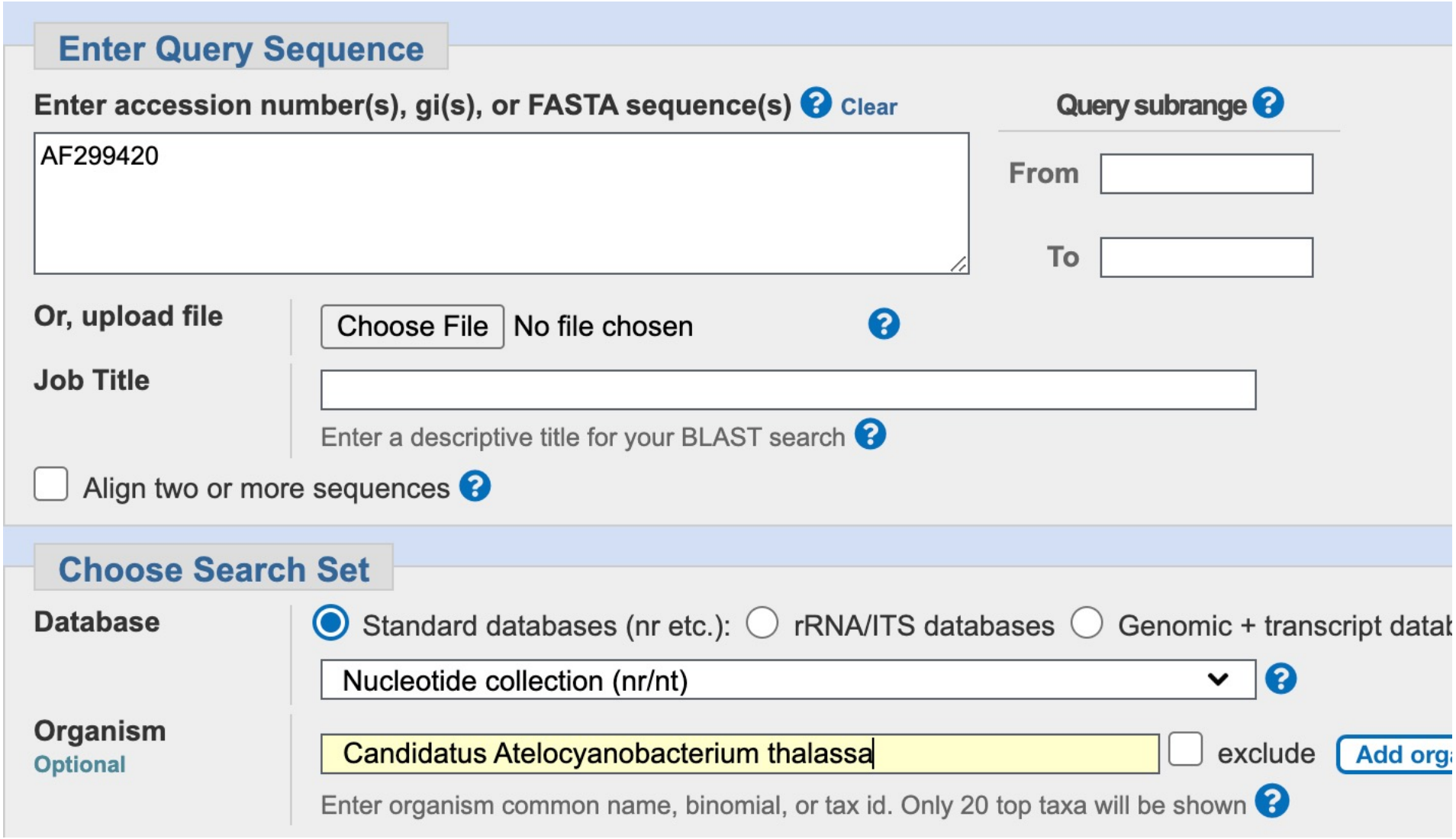
ation and paralogues; moreover, GenBank's structure and tools are not conducive to searching solely by function. For some genes, such as the nifH gene commonly used to assess community potential for N_2 fixation, manual collection and curation are becoming intractable be-

How can you gain confidence that UCYN-A is monophyletic?

- Monophyletic means:
  - All UCYN-As are in a single taxon
  - Nothing else is in that taxon

- SO
  - Collect some more UCYN-A *nifH* sequences
  - Collect some non-UCYN-A *nifH* sequences
  - Build the tree.
  - Is there still 1 monophyletic UCYN-A clade?

Finding more UCYN-A *nifH* sequences using BLAST



BLAST results: choose 5 where description is “Candidatus Atelocyanobacterium Thalassa…”, E-value is < 1E-100, and length is ~300-600

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Cyanobacterium endosymbiont of Braarudosphaera bigelowii CPSB-1 DNA, complete genome	cyanobacterium…	532	532	100%	3e-152	96.30%	1491611	AP024987.1
Candidatus Atelocyanobacterium thalassa isolate SIO0111C12A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	532	532	100%	3e-152	96.30%	359	KF806612.1
Candidatus Atelocyanobacterium thalassa isolate SIO0111C09A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	532	532	100%	3e-152	96.30%	359	KF806610.1
Candidatus Atelocyanobacterium thalassa isolate SIO0910A05A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	532	532	100%	3e-152	96.30%	359	KF806607.1
Candidatus Atelocyanobacterium thalassa isolate SIO1210B03A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	532	532	100%	3e-152	96.30%	359	KF806605.1
Candidatus Atelocyanobacterium thalassa isolate SIO0511E01A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	532	532	100%	3e-152	96.30%	359	KF806604.1
Candidatus Atelocyanobacterium thalassa isolate SIO0111C10A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	527	527	100%	2e-150	95.99%	359	KF806611.1
Candidatus Atelocyanobacterium thalassa isolate SIO1210B08A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	527	527	100%	2e-150	95.99%	359	KF806609.1
Candidatus Atelocyanobacterium thalassa isolate SIO0111C06A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	525	525	100%	5e-150	95.68%	359	KF806608.1
Candidatus Atelocyanobacterium thalassa isolate SIO1210B04A_T7 dinitrogenase reductase (nifH) gen…	Candidatus Atel…	521	521	100%	7e-149	95.68%	359	KF806606.1
Candidatus Atelocyanobacterium thalassa clone 1-20 nitrogenase iron protein NifH (nifH) gene, partial cds	Candidatus Atel…	510	510	100%	2e-145	95.06%	325	MH144433.1
Candidatus Atelocyanobacterium thalassa isolate ALOHA, complete genome	Candidatus Atel…	510	510	100%	2e-145	95.06%	1443806	CP001842.1
Candidatus Atelocyanobacterium thalassa SAG_AD-638_J10 dinitrogenase reductase (nifH) gene, parti…	Candidatus Atel…	311	311	52%	2e-85	99.42%	184	MH815013.1

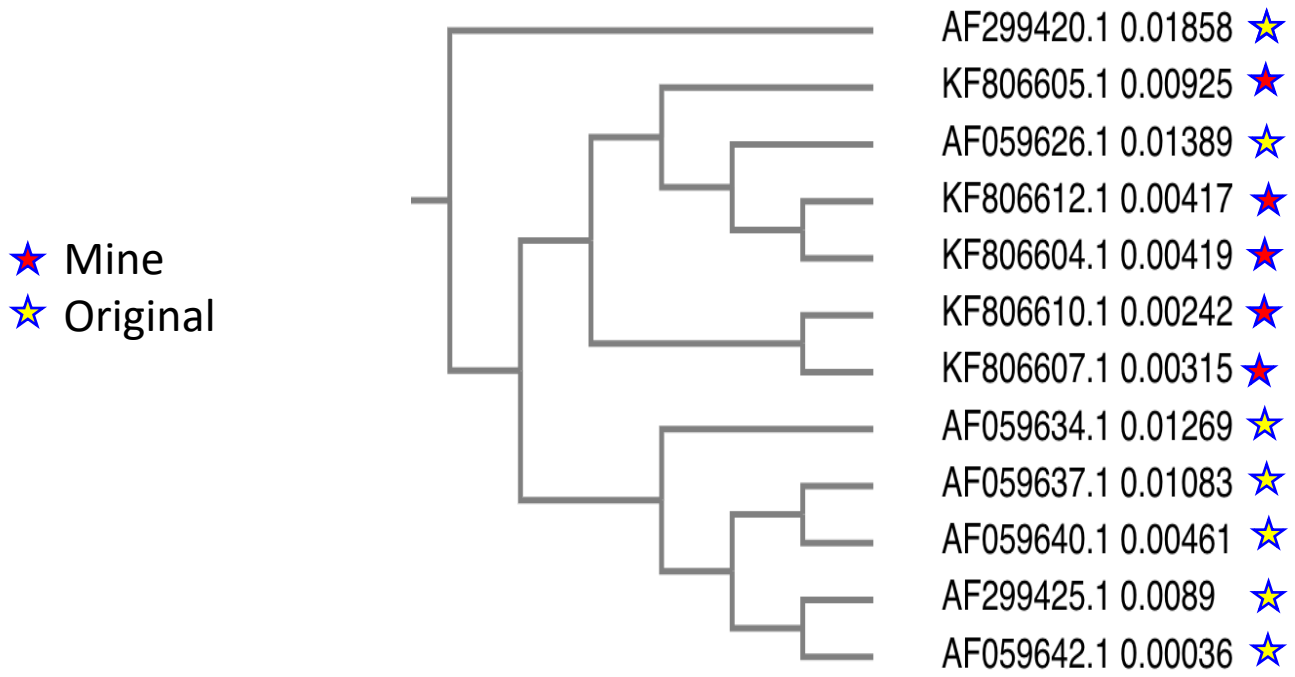
Collect non-*nifH* sequences

- Download the big database file
- Choose any 5 sequences where
  - The length is 300-700 bp
  - The accession number isn't already in your fasta file
  - The scientific name is given (so don’t choose something like “uncultured archaeon”), and the name isn't *Candidatus Atelocyanobacterium thalassa*.

Yours will vary. These are mine:

Accession #	Organism (formatted correctly!)
ABD73338.1	*Rhizobium tropici*
ABX57800.1	*Mesorhizobium amorphae*
ABX57804.1	*Mesorhizobium tianshanense*
SDE59660.1	*Fontibacillus panacisegetis*
SDE65010.1	*Eubacterium pyruvativorans*

A piece of my new tree. *Supports* but doesn't *prove* that UCYN-A *nifH* is monophyletic



HOMEWORK 1

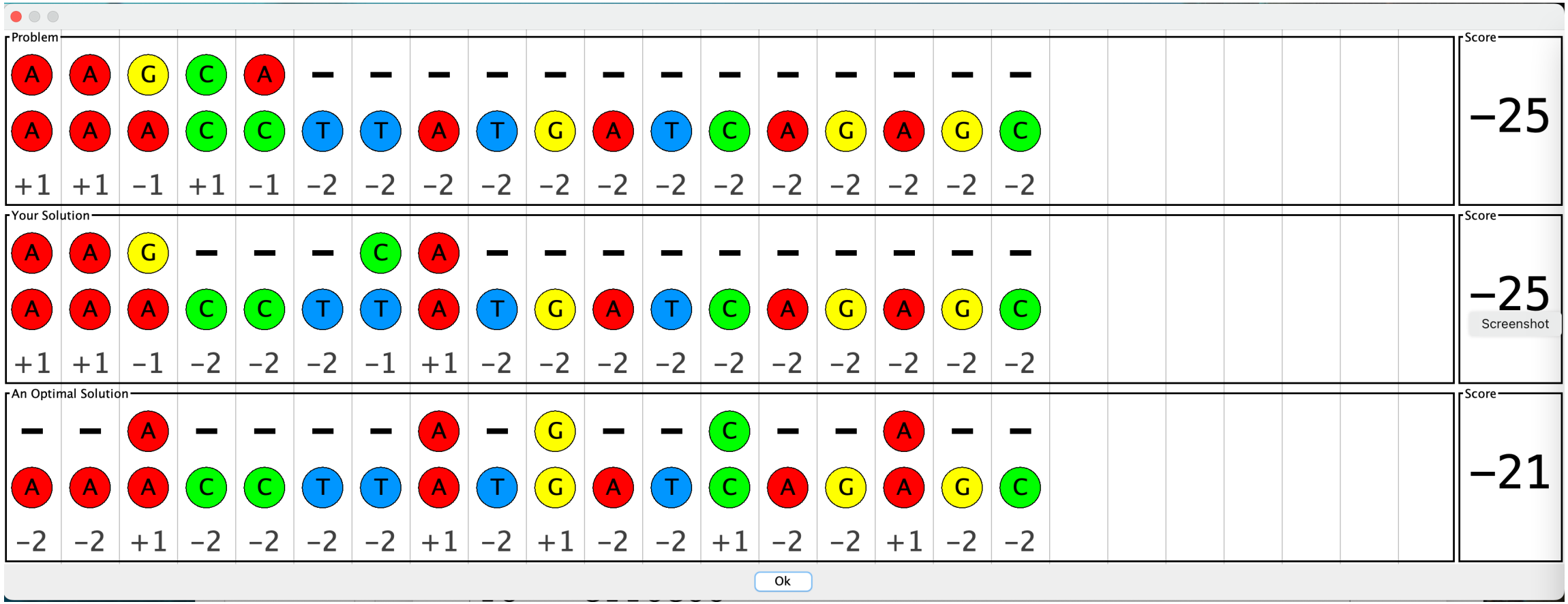
1: (10 points) What is the score of this alignment? Show all your work (zero points if you don't). Use the standard scoring scheme: +1 for a match, -1 for a mismatch, -2 for an indel.

```
ATCGGC---GCA
TC---GTTCGCC
```

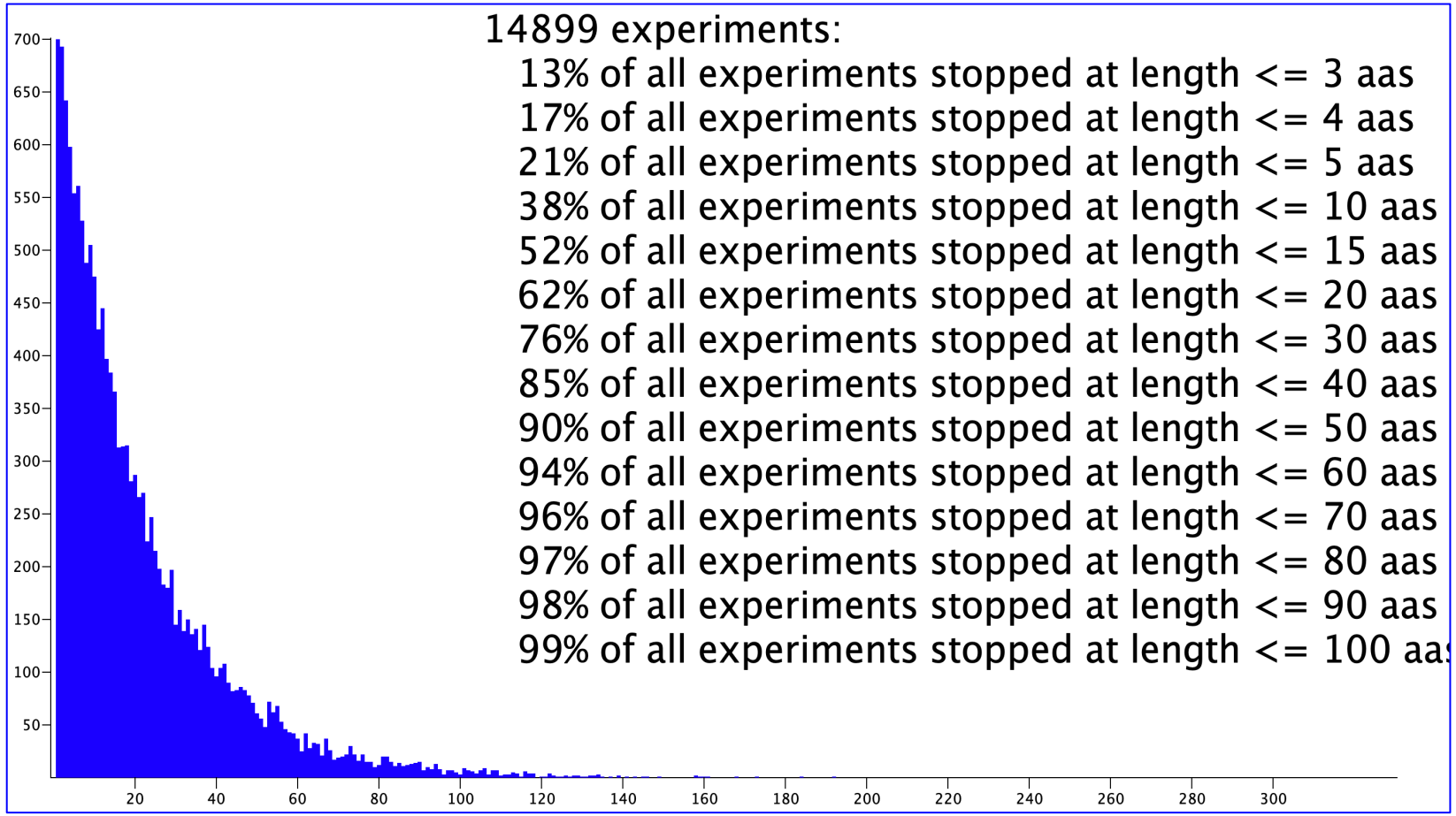
2 matches, 4 mismatches, 6 indels ➔ 2*1 – 4*1 - 6*2 = -14

2: (10 points) An indel lowers an alignment score by twice as much as a mismatch. Explain why in terms of evolution. An indel causes a frame shift, which with high probability causes a premature STOP codon. The resulting protein could be non-functional. So an indel mutation is probably disadvantageous and unlikely to be commonly found.

3: The pairwise alignment app



4: Premature-STOP app



(5) What is the score of this global alignment? Why would local alignment be a much better option for this situation?

Each flanking region: -2 * 1000
Middle: 10 matches
➔ -4000 + 10 = -3990

Local alignment score would be 10, ignoring the flanking regions And highlighting the perfect identity

(6) More marine animals. Lots of different answers