

# 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

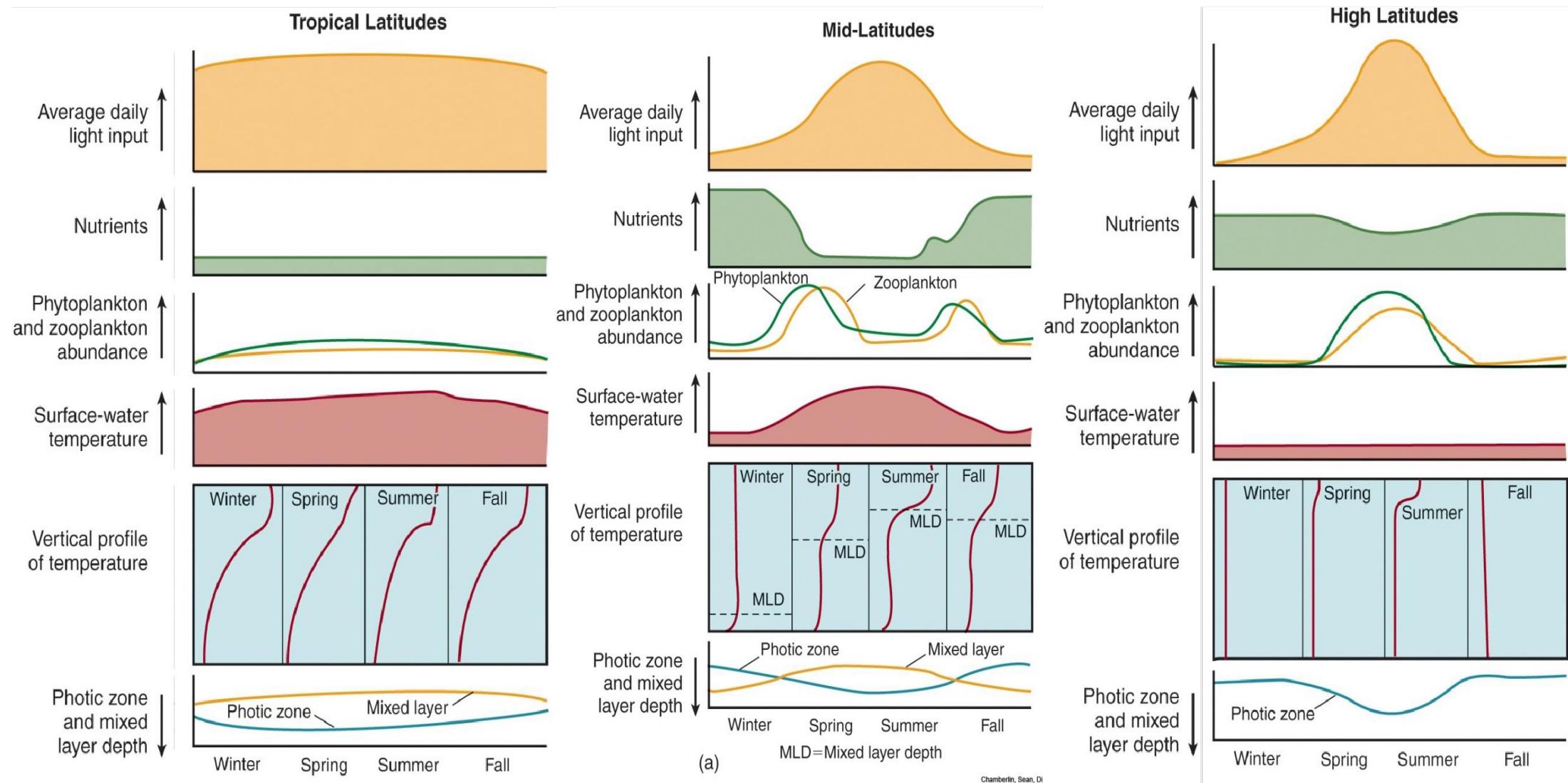
Temperate mixed layer is nutrient limited in late-summer and fall

- 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

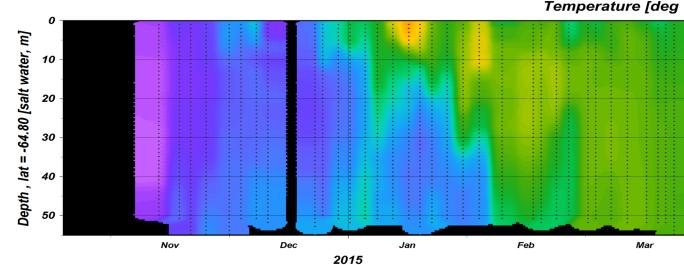
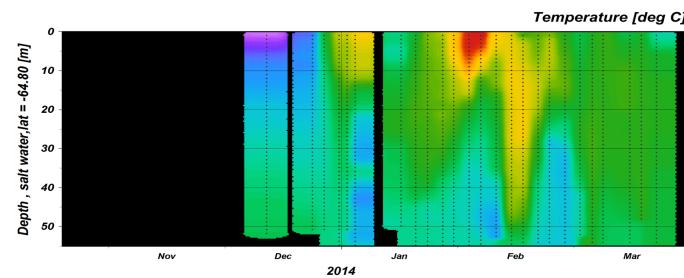
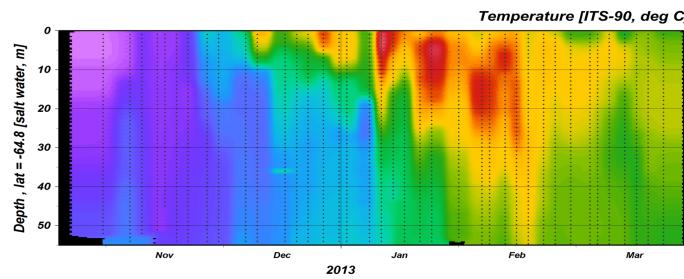
# Seasonal Changes at different latitudes



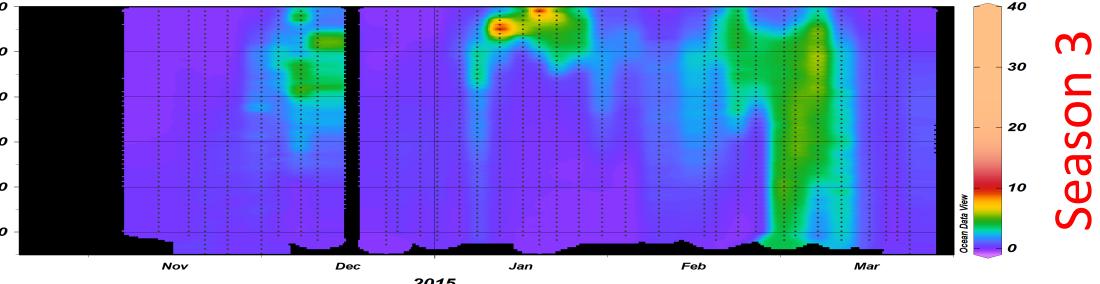
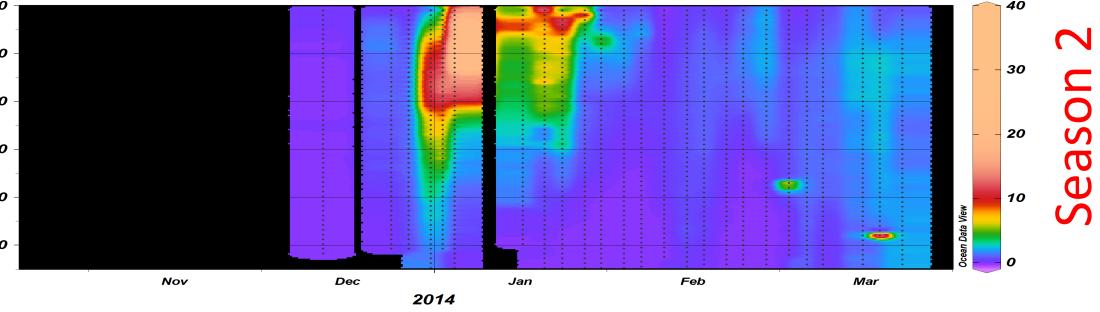
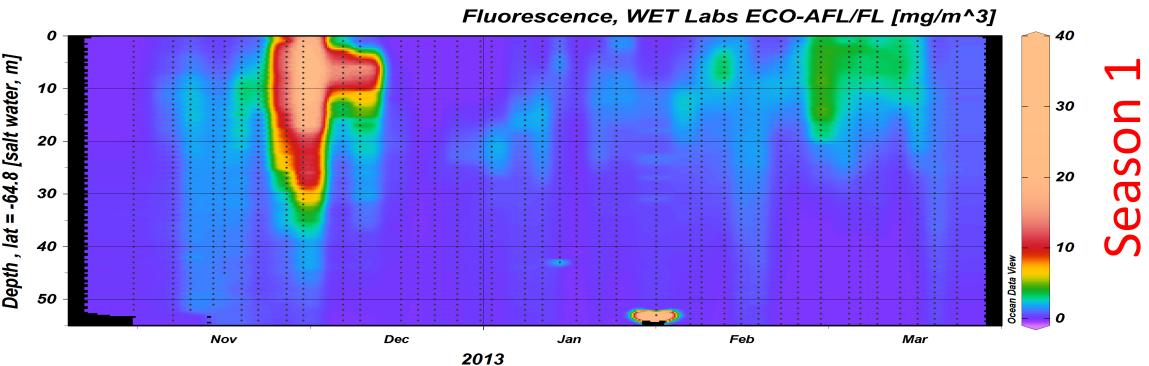
# Remember these graphs from Phil's Antarctica work?

- 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)



Season 1

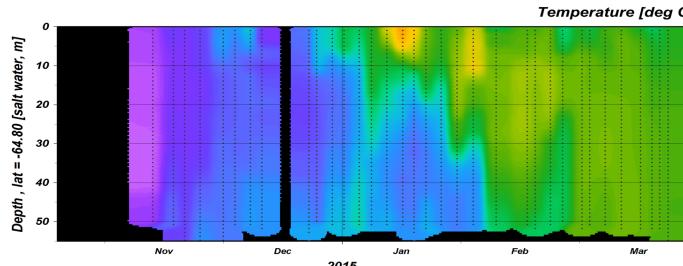
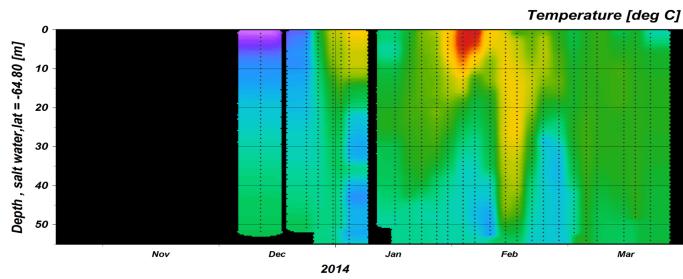
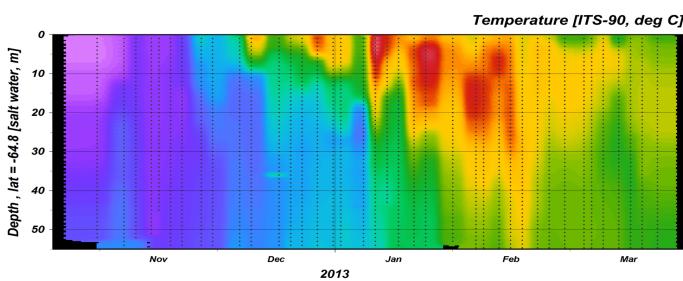
Season 2

Season 3

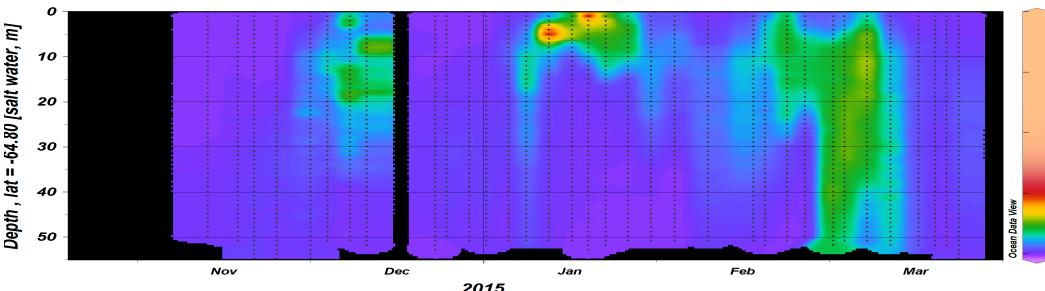
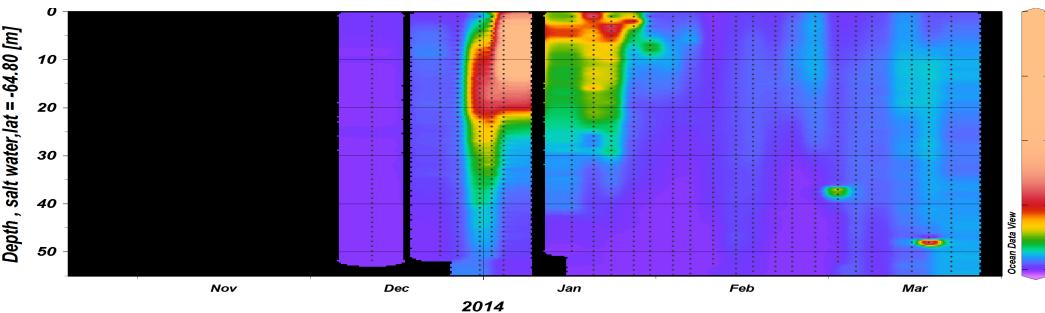
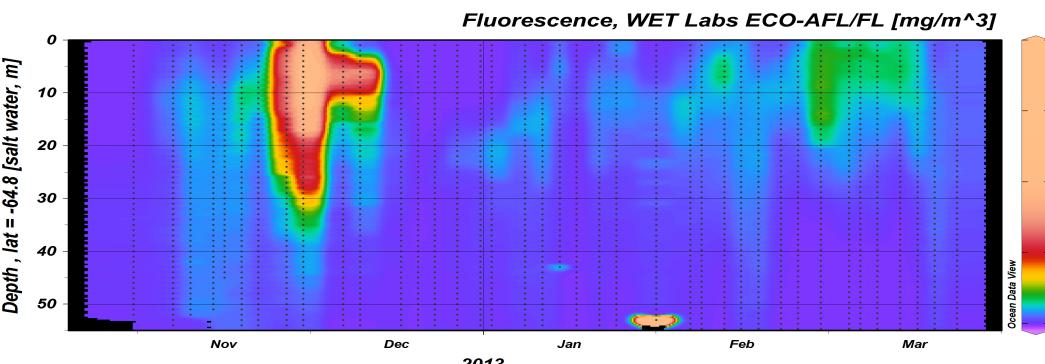
# Polar mixed layer develops in summer and is nutrient rich

- one large summer bloom...unless climate is disrupted

## Temperature



## Fluorescence (proxy for biomass)



Season 1

Season 2

Season 3

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

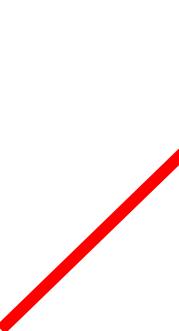
Exam: Tuesday March 7

# 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**



[PREDICTED: Orcinus orca 18S ribosomal RNA \(LOC125961161\), rRNA](#)

1. 1,569 bp linear rRNA

Accession: XR\_007471517.1 GI: 2280650058

[BioProject](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

[PREDICTED: Orcinus orca 18S ribosomal RNA \(LOC125961160\), rRNA](#)

2. 1,869 bp linear rRNA

Accession: XR\_007471516.1 GI: 2280650057

[BioProject](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

[PREDICTED: Orcinus orca 18S ribosomal RNA \(LOC125961159\), rRNA](#)

3. 1,869 bp linear rRNA

Accession: XR\_007471515.1 GI: 2280650056

[BioProject](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

[PREDICTED: Orcinus orca 18S ribosomal RNA \(LOC125961157\), rRNA](#)

4. 1,869 bp linear rRNA

Accession: XR\_007471514.1 GI: 2280650055

[BioProject](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

# (2) Look at the GP page

The screenshot shows the NCBI Nucleotide search results for the NCBI Reference Sequence XR\_007471517.1. The page title is "PREDICTED: Orcinus orca 18S ribosomal RNA (LOC125961161), rRNA". The sequence details include:

- NCBI Reference Sequence: XR\_007471517.1
- FASTA and Graphics links
- Sequence details:
  - 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.
  - Also see: Documentation of NCBI's Annotation Process
- Annotation details:
  - Annotation Provider :: NCBI RefSeq
  - Annotation Status :: Full annotation
  - Annotation Name :: Orcinus orca Annotation Release 103
  - Annotation Version :: 103
  - Annotation Pipeline :: NCBI eukaryotic genome annotation pipeline
  - Annotation Software Version :: 10.0
  - Annotation Method :: Best-placed RefSeq; Gnomon
  - Features Annotated :: Gene; mRNA; CDS; ncRNA
- FEATURES:
  - source 1..1869 /organism="Orcinus orca" /mol\_type="rRNA" /db\_xref="taxon:9733" /chromosome="14"
  - gene 1..1869

The right sidebar contains links for analysis (Run BLAST, Pick Primers, Highlight Sequence Features, Find in this Sequence, Show in Genome Data Viewer), related information (BioProject, Taxonomy, Annotated Genomic, Gene), and recent activity (Recent activity, Turn Off, Clear).

# Where's the sequence?

End of page

## ORIGIN

```
1 tacctggttg atcctgccag tagcatatgc ttgtctcaa gattaagcca tgcatagtcta  
61 agtacgcacg gccggtagac tgaaactgcg aatggctcat taaatcagtt atggccctt  
121 tggtcgctcg ctcctctcct acttggataa ctgtggtaat tctagagcta atacatgccg  
181 acgggcgcctg acccccctcg cggggggat gcgtgcattt atcagatcaa aaccaacccg  
241 gtcagcctcc ctccggcccc ggccgggggt cgggcgcgg cggttttgt gactctagat  
301 aacctcgggc cgatgcacg ccccccgtgg cggcgacgac ccattcgaac gtctgcccata  
361 tcaactttcg atggtagtcg ccgtgcctac catggtgacc acgggtgacg gggaaatcagg  
421 gttcgattcc ggagagggag cctgagaaac ggctaccaca tccaaggaag gcagcaggcg  
481 cgcaaattac ccactcccga cccggggagg tagtgacgaa aaataacaat acaggactct  
541 ttcgaggccc tctaatttggaa atgagtccac tttaaatctt ttcgcgagga tccattggag  
601 ggcaagtctg gtgccagcag ccgcggtaat tccagctcca atagcgtata ttaaagtgtc  
661 tgcagttaaa aagctcgtag ttggatcttgg gggcggcccg ggcggccgc cgcgaggcga  
721 gccaccggcc gtcggccccc ctgcctctc ggccggccctt cgatgctttt agctgagtgt  
781 cccgggggc ccgaagcggt tactttgaaa aaatttaggt gttcaaagca ggcccgagcc  
841 gcctggatac cgcaagctagg aataatggaa taggaccgcg gttctatttt gttggtttgc  
901 ggaactgagg ccatgattaa gagggacggc cggggcatt cgtattgcgc cgctagaggt  
961 gaaattcttgc gaccggcgca agacggacca gagcggaaagc atttgc当地 aatgtttca  
1021 ttaatcaaga acgaaagtgc gaggttcgaa gacgatcaga taccgtcgta gttccgacca  
1081 taaacgatgc cgactggcgta tgccggcccg ttattcccat gaccggcccg gcagcttccg  
1141 ggaaaccaaa gtctttgggt tccggggggaa gtatggttgc aaagctgaaa cttaaaggaa  
1201 ttgacggaaag ggaccacca ggagtggagc ctgcggctt atttactca acacggggaaa  
1261 cctcacccgg cccggacacg gacaggattt acagatttgc agctcttctt cgattccgtg  
1321 ggtgggtggc catggccgtt cttttttttt ggagcgattt gtctggtaa ttccgataac  
1381 gaacgagact ctggcatgct aacttagtttac ggcaccccg agcggtcgcc gtcggccaaac  
1441 ttcttagagg gacaagtggc gttcagccac ccggatttgc gcaataacag gtctgtgtat  
1501 cccttagatgc tccggggctg cacgcgcgtt acactgactt gctcagcgtt tgcctaccct  
1561 acgcggcag ggcgggttac cccgttgcac cccattcgtt atggggatcg gggattgcaa  
1621 ttattccca tgaacgagga attccatgtt gtcggggcataaagcttgc gttgattaag  
1681 tccctggccct ttgtacacac cggccgtcgac tactaccat tggatggttt agtgaggccc  
1741 tcggatcgcc cccggccggg tcggccacgc gcccctggccgg agcgctgaga agacggcga  
1801 acttgactat ctagaggaag taaaagtcgt aacaaggattt ccgttaggttga acctgcggaa  
1861 ggatcatta
```

//

# Additional information

LOCUS	XR_007471517	1869 bp	rRNA	linear	MAM	02-AUG-2022
DEFINITION	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961161), rRNA.					
ACCESSION	XR_007471517					
VERSION	XR_007471517.1					
DBLINK	BioProject: <a href="#">PRJNA854208</a>					
KEYWORDS	RefSeq.					
SOURCE	<i>Orcinus orca</i> (killer whale)					
ORGANISM	<i>Orcinus orca</i> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Whippomorpha; Cetacea; Odontoceti; Delphinidae; <i>Orcinus</i> .					
COMMENT	MODEL <a href="#">REFSEQ</a> : This record is predicted by automated computational analysis. This record is derived from a genomic sequence ( <a href="#">NC_064572</a> ) 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	?	
<input checked="" type="checkbox"/> 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			
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310087), rRNA	<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524342.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310086), rRNA	<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524341.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310824), rRNA	<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524870.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310823), rRNA	<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524869.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310816), rRNA	<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524862.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Globicephala melas</i> 18S ribosomal RNA (LOC115850435), rRNA	<i>Globicephala m...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004038394.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Lagenorhynchus obliquidens</i> 18S ribosomal RNA (LOC113616899), rRNA	<i>Lagenorhynchu...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_003431446.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961160), rRNA	<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471516.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961159), rRNA	<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471515.1</a>			
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961157), rRNA	<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471514.1</a>			

(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	?	
							GenBank	Graphics	Distance tree of results			MSA Viewer				
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession							
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310087), rRNA	<a href="#">Tursiops truncat...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524342.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310086), rRNA	<a href="#">Tursiops truncat...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524341.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310824), rRNA	<a href="#">Tursiops truncat...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524870.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310823), rRNA	<a href="#">Tursiops truncat...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524869.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310816), rRNA	<a href="#">Tursiops truncat...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524862.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Globicephala melas</i> 18S ribosomal RNA (LOC115850435), rRNA	<a href="#">Globicephala m...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004038394.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Lagenorhynchus obliquidens</i> 18S ribosomal RNA (LOC113616899), rRNA	<a href="#">Lagenorhynchu...</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_003431446.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961160), rRNA	<a href="#">Orcinus orca</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471516.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961159), rRNA	<a href="#">Orcinus orca</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471515.1</a>							
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961157), rRNA	<a href="#">Orcinus orca</a>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471514.1</a>							

# *Globicephala melas*: the long-finned pilot whale



# (6) The *Tursiops/Globicephala* alignment

Descriptions	Graphic Summary	Alignments	Taxonomy									
Sequences producing significant alignments				Download		Select columns		Show 100	?			
				GenBank	Graphics	Distance tree of results		MSA Viewer				
	Description			Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310087), rRNA			<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524342.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310086), rRNA			<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524341.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310824), rRNA			<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524870.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310823), rRNA			<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524869.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Tursiops truncatus</i> 18S ribosomal RNA (LOC117310816), rRNA			<i>Tursiops truncat...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004524862.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Globicephala melas</i> 18S ribosomal RNA (LOC115850435), rRNA			<i>Globicephala mel...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_004038394.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Lagenorhynchus obliquidens</i> 18S ribosomal RNA (LOC113616899), rRNA			<i>Lagenorhynchu...</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_003431446.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961160), rRNA			<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471516.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961159), rRNA			<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471515.1</a>	
<input checked="" type="checkbox"/>	PREDICTED: <i>Orcinus orca</i> 18S ribosomal RNA (LOC125961157), rRNA			<i>Orcinus orca</i>	3452	3452	100%	0.0	100.00%	1869	<a href="#">XR_007471514.1</a>	

CLICK



# (6) The *Tursiops/Globicephala* alignment

*Length of subject sequence*      *Identity columns*      *% identity*

[Download](#) [▼ GenBank Graphics](#)

**PREDICTED: Globicephala melas 18S ribosomal RNA (LOC115850435), rRNA**

Sequence ID: [XR\\_004038394.1](#) Length: 1869 Number of Matches: 1

Range 1: 1 to 1869 [GenBank](#) [Graphics](#) [▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Identities	Gaps	Strand
3446 bits(1866)	0.0	1868/1869(99%)	0/1869(0%)	Plus/Plus

Query	1	TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTA	60
Sbjct	1	TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTA	60

# (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

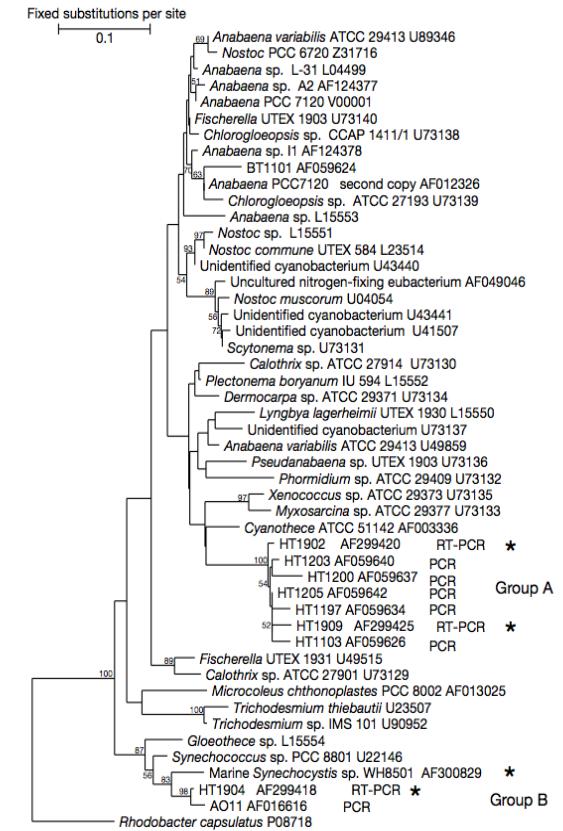
(9) Which pix surprised me?



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

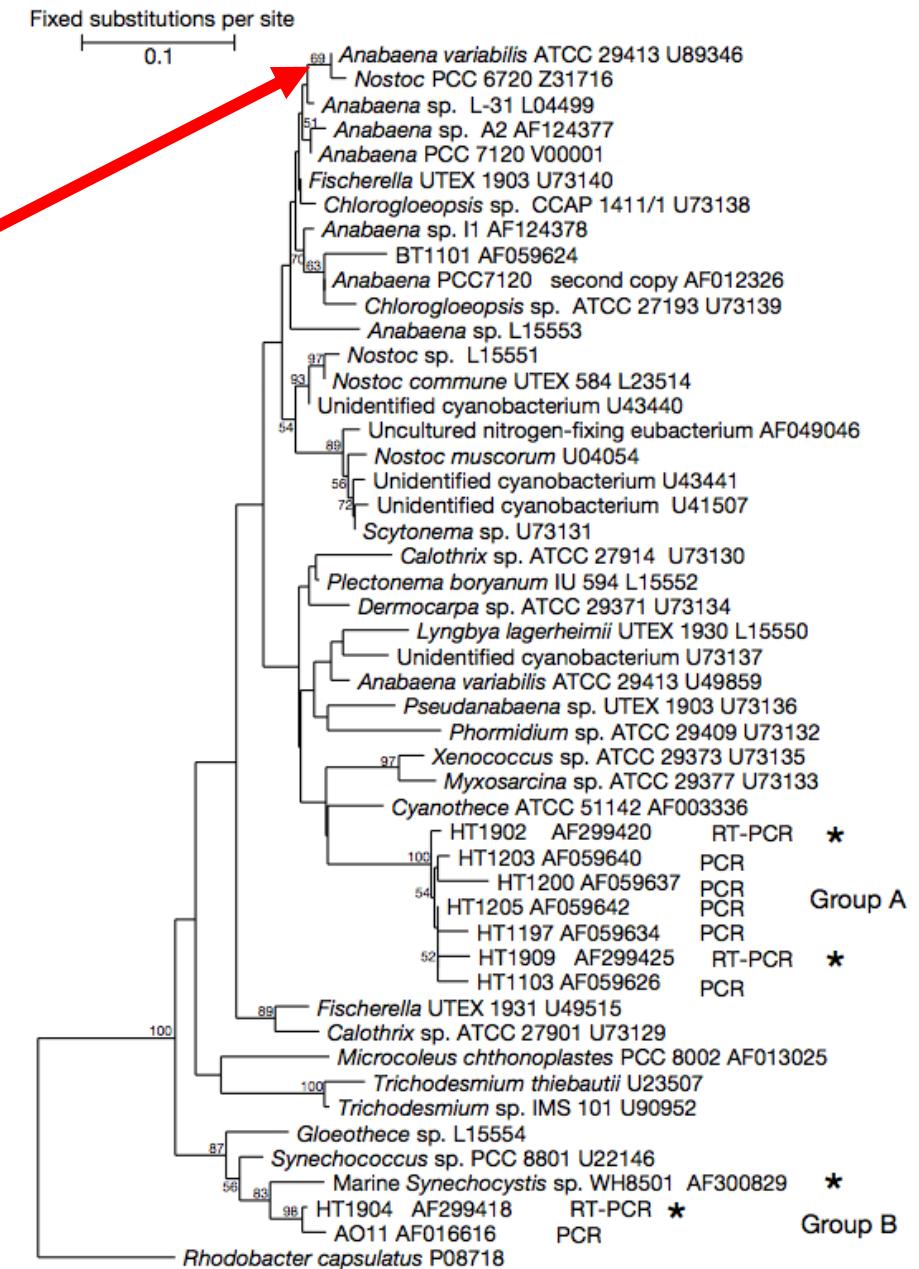
# Lab 2

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



# (1) Look at the tree

Top OTU is *Anabaena variabilis*,  
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

CGATATTGTCAAAGTAGTACTGCAAGGCGCGTGGCTCCTGTTCTAGTAGTACAGGCCACCTTGAAAAATA

...

*Sequences match*

ORIGIN

1 c~~gatattgtc~~ aaagttagtac tgcaaggcgc gtgtggctcc t~~gttctagta~~ gtacagccac  
61 cttgaaaata gcgattgaat ccagattacg cgatcgatt aatcccagcc tagtagtaga  
121 agcagtttag tcattagtca tttagtcatta gtcaatggtc attagtcaac agtgaaaaaat

GP Page

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

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



```
>L15554.1 Gloeothece sp. PCC 6909 nitrogen fixation protein (nifH) gene, partial cds
TCTACTCGCTTGATCCTTAAGTAAAGCGCATGTCACAGTTCTACACTAGCCGCAGAACGGGGTTCTG
TTGAAGATATAGAACTCGAAGATGTACTGCTCACAGGGTTGAAGACATCAAATGCGTAGAATCAGGTGG
TCCTGAACCTGGCGTAGGATGCGCTGGTCGTGGGATTATCACTGCCATCAACTCCTGAAGAAGAAGGA
GCTTACGAAGACATAGATTCGTATCCTACGACGTATTAGGGGACGTTGTCTCGGGTGGTTCGCTATGC
CTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGAACCTCT.
```

*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:  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.*

*(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

-----TCCACTCGCCTGATGCTCCACAGCAAAGCCAAACCTCCGTACTACAAC  
TGGCTGCCGAACTCGGTGCGTTGAAGATGTCGAACCTGACCAGGTGCTGCAAGAAGGGCT  
ATCGCGGCCTTAAGTGCCTGAGTCCGGTGGTCTGAGCCCCGGCGTCGGCTGTGCCGGTC  
GCGGCATTATCACTGCCATCAACTTCCCTGGAAGAAGAGGGCGCTTACGAAGATCTGATT  
TCGTCTCCTACGACGTACTCGGTGACGTAGTTGCGGCGGTTCGCCATGCCATTGGG  
AAGGCAAAGCCAAGAAATCTACATTGTTGTCTCC-----

# An aligned sequence with internal indels

```
>AF003336.1 Cyanothece ATCC51142 nitrogenase reductase (nifH) gene, complete cds
```

```
-----ATG-----  
GGACGCAJC-----TCCAAGGTTCTAACCAACAAATCGAATCATCGTTA  
CGTCAGCAACGCTCTACCTTACACTCACTGTCTAACAAAGCGAGAACAA  
CTATGCAGATTGCATTTCAGATTGCTACCTCTCAGA  
ATACCATTGCTGCCTAGCTGAA---AC---CAACCGCATCATGATTGTTGGTTGTGACC  
CTAAAGCTGATTCTACCCGCTAAACGCTAACAAAGCACAAACCACCTCTGCACT  
TAGCAGCAGAACGGGAACCGTTGAAGACATCGAACTCGAAGAAGTATTACTCGAAGGAT  
ACCAAGGAGTCAGTGTGTTGAGTCCGGTGGCTGAGCCTGGAGTTGGATGTGCAGGGTC  
GTGGTATTATCACCGCCATTAACCTCTTAGAAGAAAGAAGGTGCTACGAAGACCTAGACT  
TCGTATCCTACGACGTATTAGGAGACGTTGTATGTGGTGGTTCGCTATGCCATCCGT  
AAGGAAAAGCACAAGAAATCTACATCGTAACCTCCGGGAAATGATGGCGATGTACGCTG  
CAAACAAACATTGCTCGTGGTATTTAAAATACGCTCACACTGGTGGTGGCTAGGTG  
GTTTAATTGTAACAGCGTAACGTTAAGCTGAGCTGAGTTAACGAAAGAATTAGCTC  
GTCGTCTCGGAACCCAAATGATTCACTCGTACCCGTTCAAGCAGGTACAAGAAGCTG  
AATTACGTCGTATGACTGTTATCGAATATTCTCCTGATCACCCCTCAGGCTCAGGAATACC  
GTGAGTTATCTCGAAAATCGAGAATAACACCAACCTCGTTATTCCCTACTCCTATCACCA  
TCCGAAACTCGAAGAACTCTTAGTTGACTTCGGTATTCTCGGTGGTGAAGACGAGTATG  
A--GAAAGCTCTTCAAGCTGATAAAGCTGCTACCAAAGCTTAG-----
```

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*

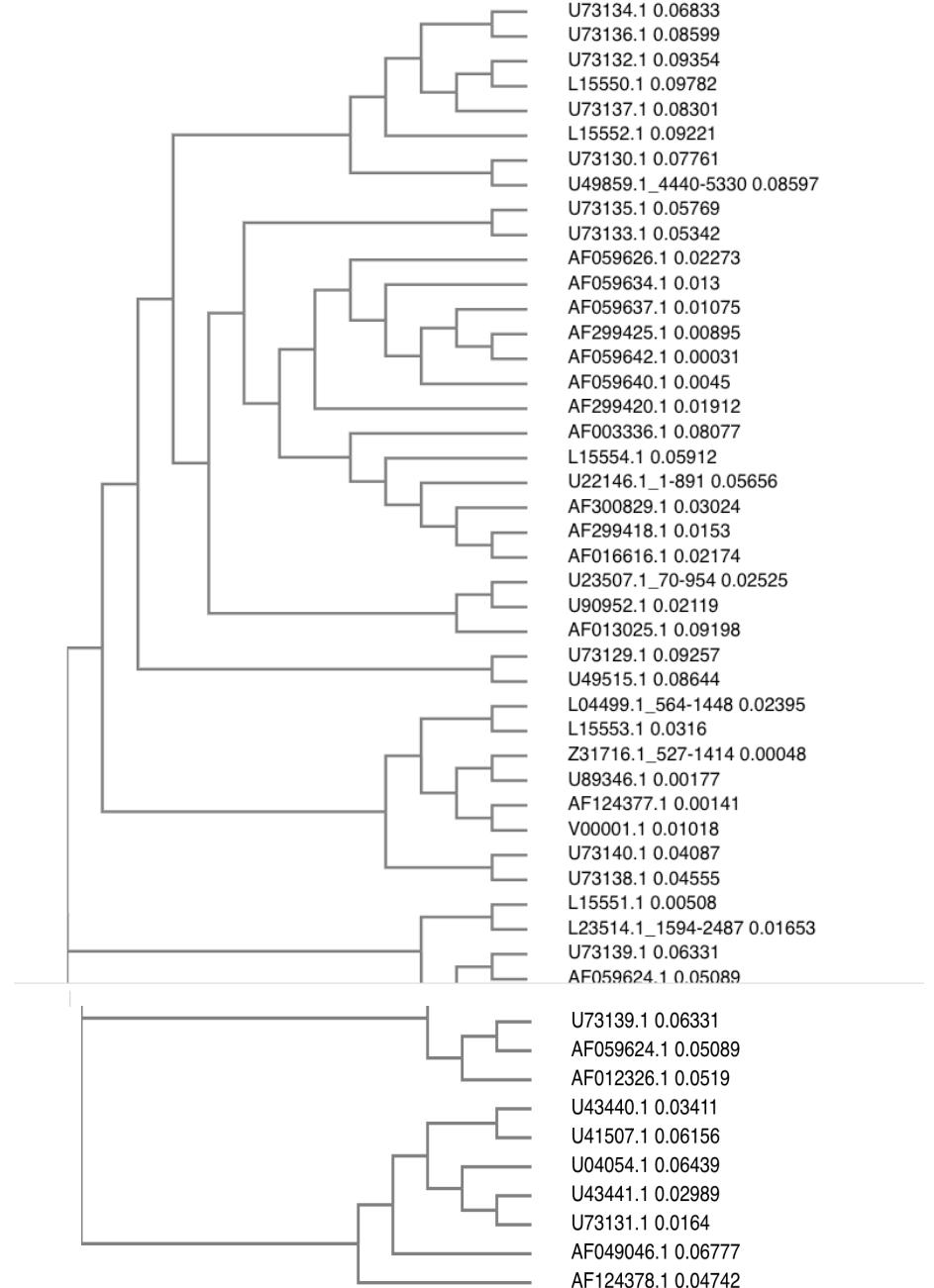
# My tree

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**

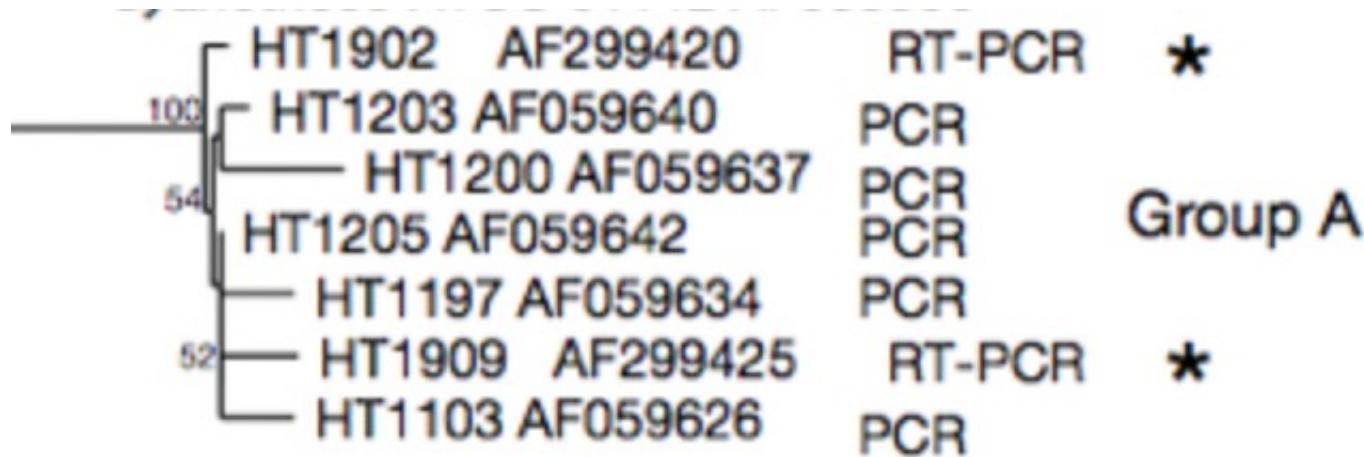


# The 7 Original Group A sequences

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<sub>2</sub> fixation, manual collection and curation are becoming intractable because of the large number of sequences in GenBank and the large

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

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file  No file chosen [?](#)

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 database

Nucleotide collection (nr/nt)  [?](#)

Organism **Optional**   exclude [Add org...](#) [?](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

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
<input checked="" type="checkbox"/>	<a href="#">Cyanobacterium endosymbiont of Braarudosphaera bigelowii CPSB-1 DNA, complete genome</a>	<a href="#">cyanobacterium...</a>	532	532	100%	3e-152	96.30%	1491611	<a href="#">AP024987.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0111C12A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	532	532	100%	3e-152	96.30%	359	<a href="#">KF806612.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0111C09A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	532	532	100%	3e-152	96.30%	359	<a href="#">KF806610.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0910A05A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	532	532	100%	3e-152	96.30%	359	<a href="#">KF806607.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO1210B03A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	532	532	100%	3e-152	96.30%	359	<a href="#">KF806605.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0511E01A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	532	532	100%	3e-152	96.30%	359	<a href="#">KF806604.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0111C10A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	527	527	100%	2e-150	95.99%	359	<a href="#">KF806611.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO1210B08A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	527	527	100%	2e-150	95.99%	359	<a href="#">KF806609.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO0111C06A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	525	525	100%	5e-150	95.68%	359	<a href="#">KF806608.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate SIO1210B04A_T7 dinitrogenase reductase (nifH).gene</a>	<a href="#">Candidatus Atel...</a>	521	521	100%	7e-149	95.68%	359	<a href="#">KF806606.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa clone 1-20 nitrogenase iron protein NifH (nifH).gene, partial cds</a>	<a href="#">Candidatus Atel...</a>	510	510	100%	2e-145	95.06%	325	<a href="#">MH144433.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa isolate ALOHA, complete genome</a>	<a href="#">Candidatus Atel...</a>	510	510	100%	2e-145	95.06%	1443806	<a href="#">CP001842.1</a>
<input checked="" type="checkbox"/>	<a href="#">Candidatus Atelocyanobacterium thalassa SAG_AD-638_J10 dinitrogenase reductase (nifH).gene, parti...</a>	<a href="#">Candidatus Atel...</a>	311	311	52%	2e-85	99.42%	184	<a href="#">MH815013.1</a>

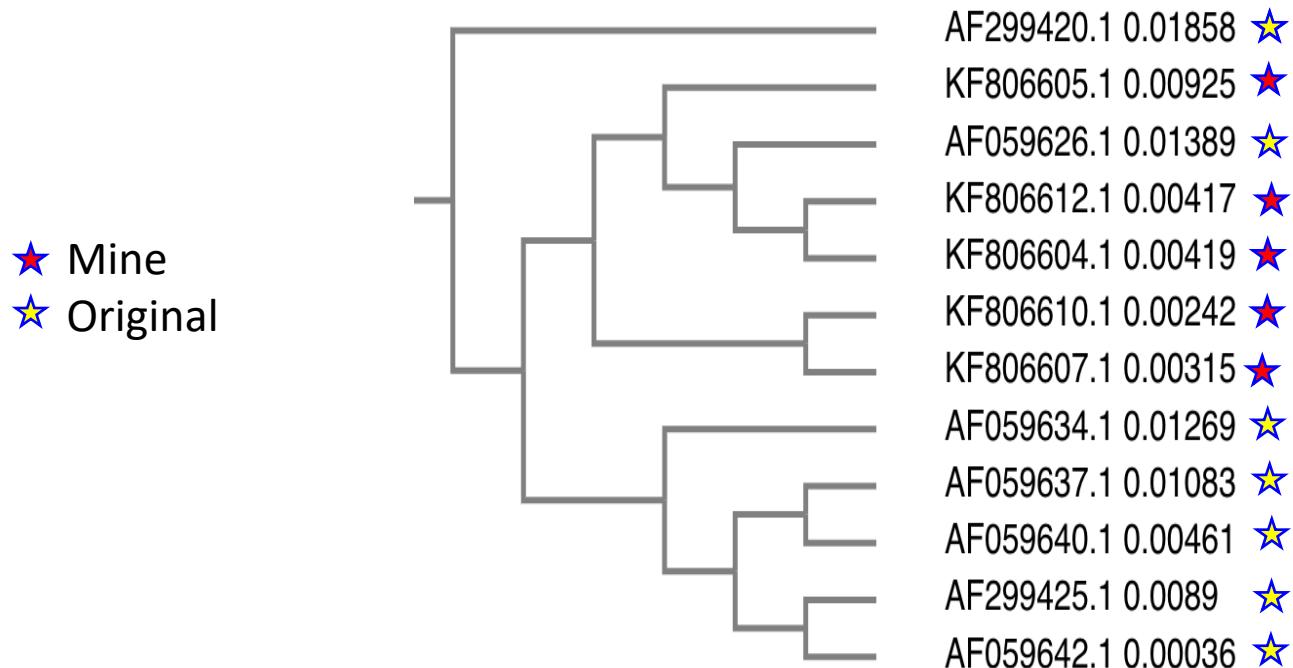
# Collect non-*nifH* sequences

- 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	<i>Rhizobium tropici</i>
ABX57800.1	<i>Mesorhizobium amorphae</i>
ABX57804.1	<i>Mesorhizobium tianshanense</i>
SDE59660.1	<i>Fontibacillus panacisegetis</i>
SDE65010.1	<i>Eubacterium pyruvativorans</i>

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

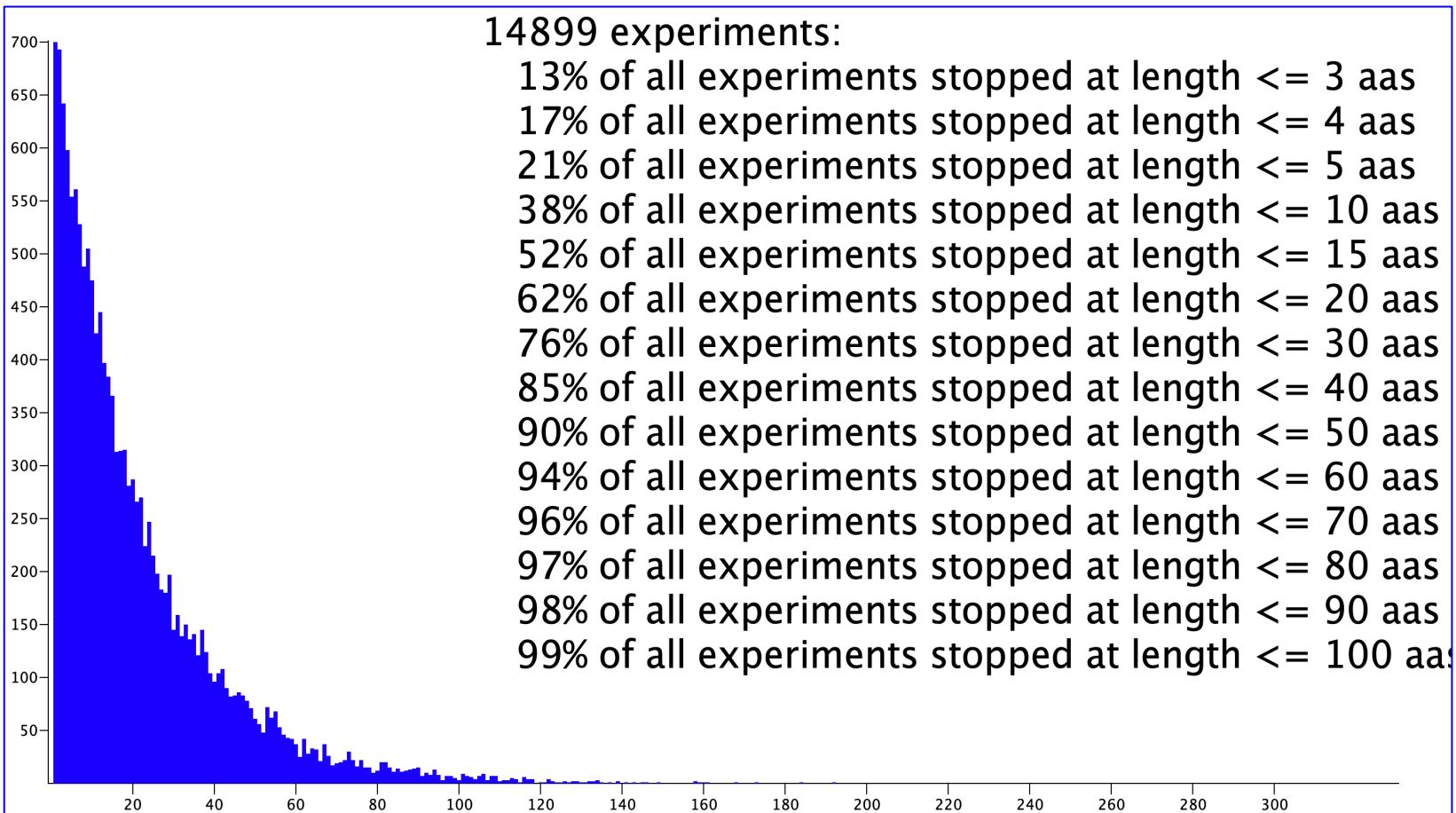
TC---GTTGCC

2 matches, 4 mismatches, 6 indels →  $2*1 - 4*1 - 6*2 = -14$

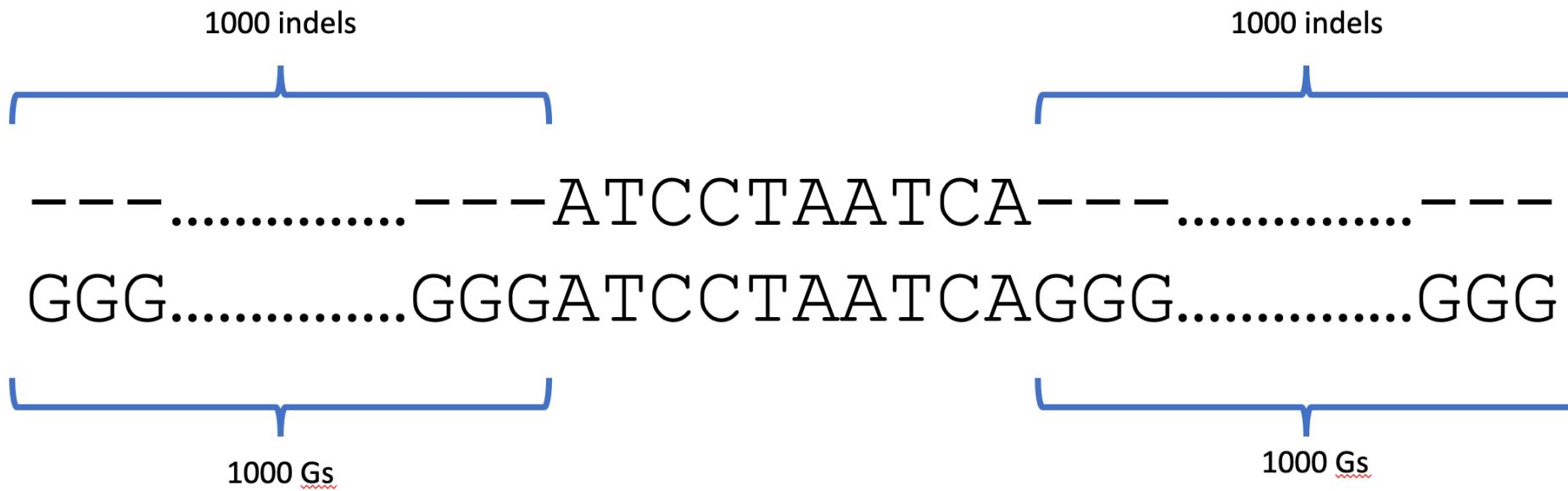
**2: (10 points)** An indel lowers an alignment score by twice as much as a mismatch. Explain why in terms of evolution. An indel causes a frame shift, which with high probability causes a premature STOP codon. The resulting protein could be non-functional. So an indel mutation is probably disadvantageous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



# 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