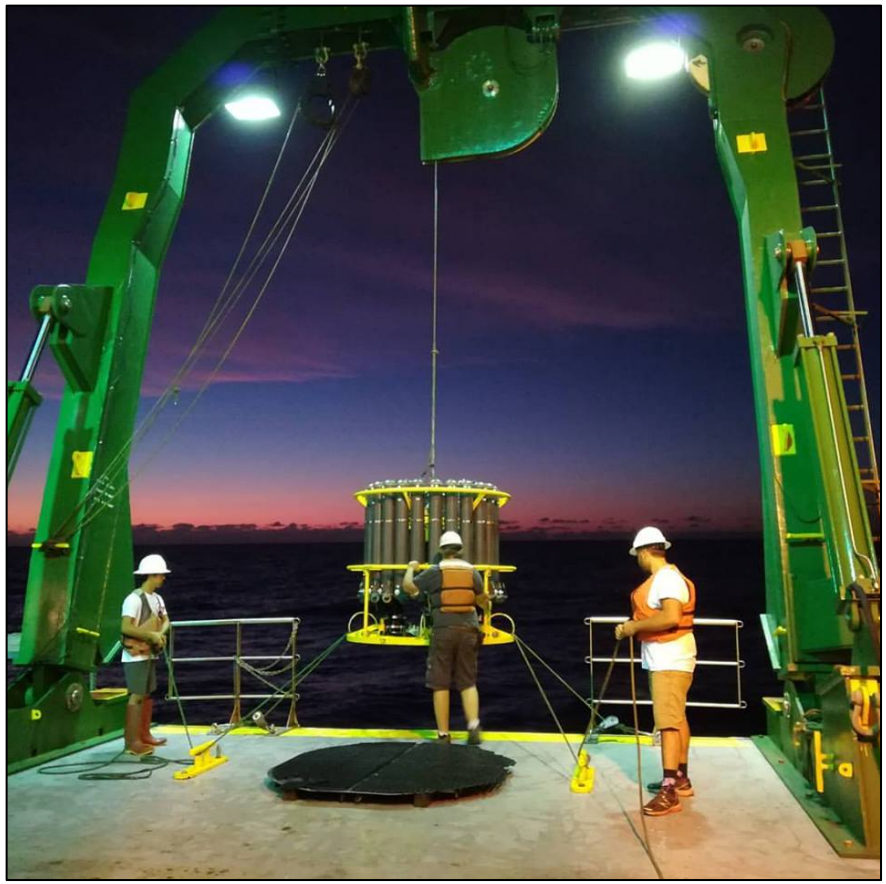
BIO/CS 123B
Part 6: Metagenomics and UCYN-A

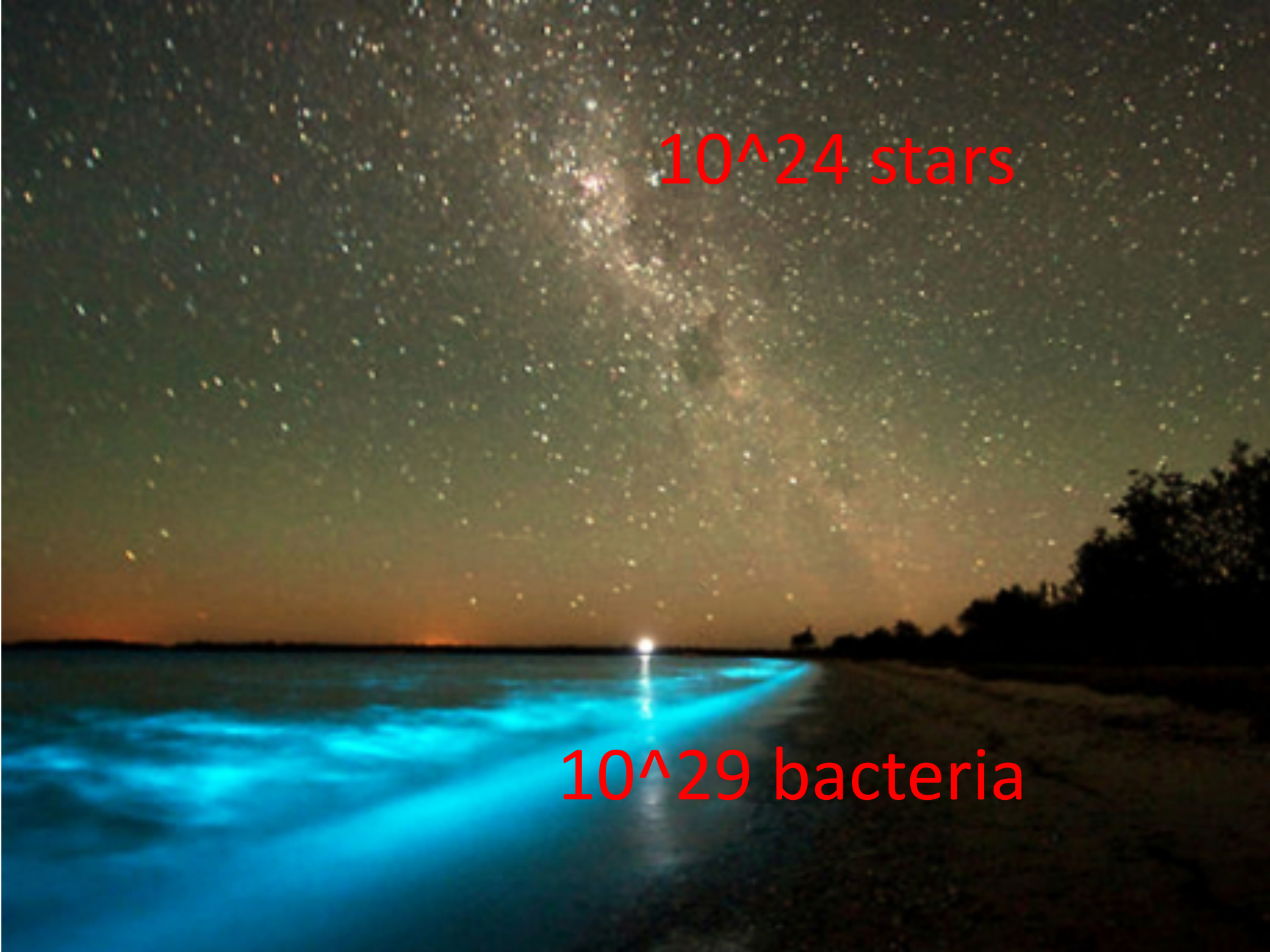


Spring 2021

Phil Heller

Meta

- Literally means “beyond”
  - Metaphysics
  - Metamorphosis
  - Metastasis
  - Metadata
- But in metagenomics/metatranscriptomics/ metaprote-omics, it means “community”
  - Usually a microbial community
  - Impossible to identify individuals visually
  - Impossible to count individuals visually


10^24 stars
10^29 bacteria

Bacteria in the ocean

- Average concentration: 10^6 cells / ml (Whitman et al. 1998)
- Higher near land (nutrient runoff), near surface (photosynthesis)
- Lower at depth (no sunlight), in open ocean

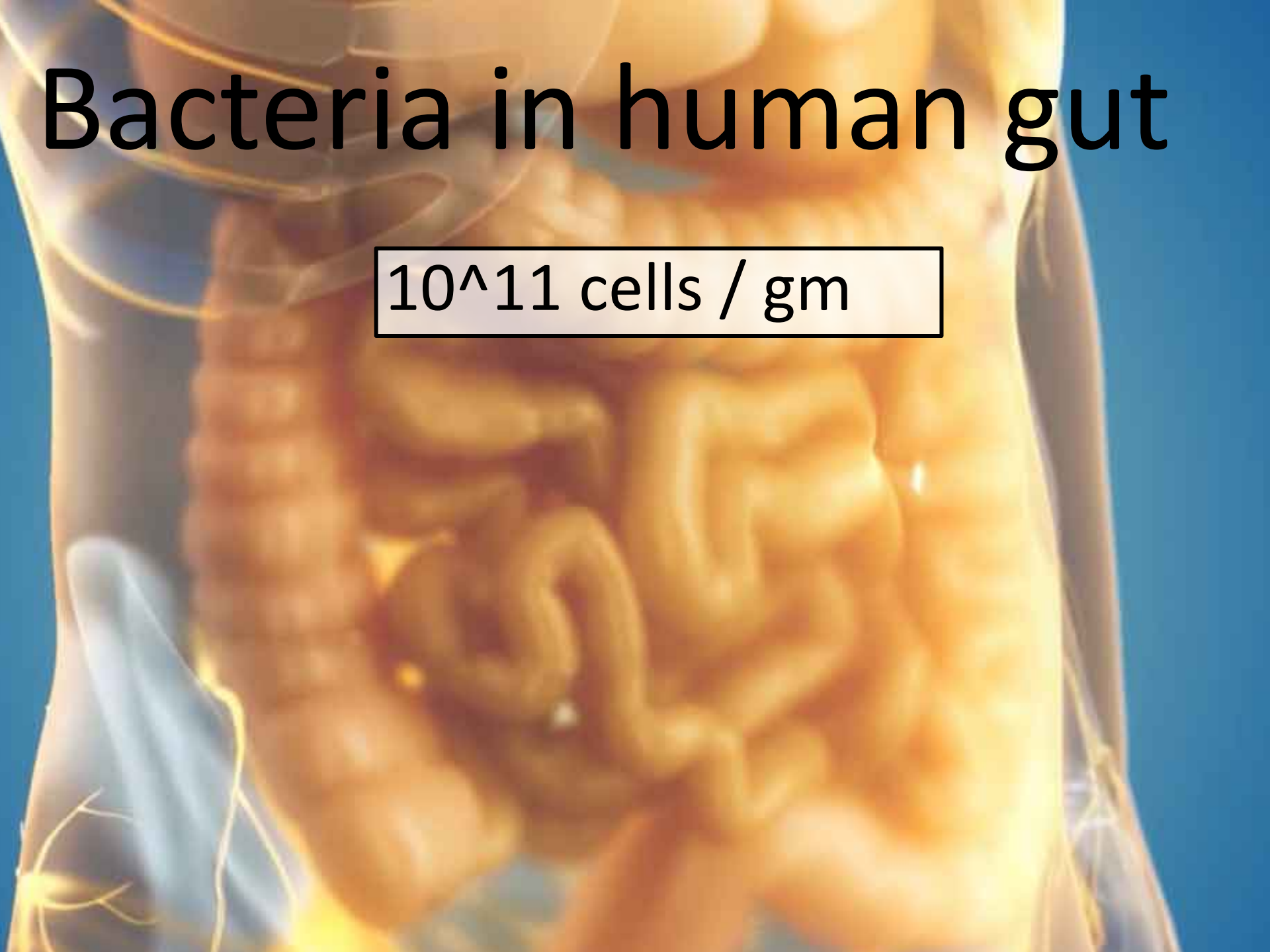
"The moment you sample ocean water, you have a big-data problem"

- PH


Bacteria in soil
10^8 cells / gm

Bacteria in human gut

10^11 cells / gm



Bacteria in human saliva

5 x 10^5 cells / gm

Questions that a metagenomic study can answer

- Taxonomic identification: What species are present in the community, and in what proportions?
- Functional identification: What is the *genetic potential* of the community?
  - What genes are present, regardless of what species own them
  - E.g. presence of *nifH* means the community might fix nitrogen … probably does, but not definitively proved

Metatranscriptomics

- Better than genetic potential, for a price
- What genes were actually being expressed at the moment you sampled?
- Expression volume can fluctuate over a 24-hour cycle: "diel" expression → have to sample n times over 24 hours
  - Example: photosynthesis genes: half-life < 12 hours
- Technology:
  - RNA-Seq
  - cDNA ("*complementary* DNA")

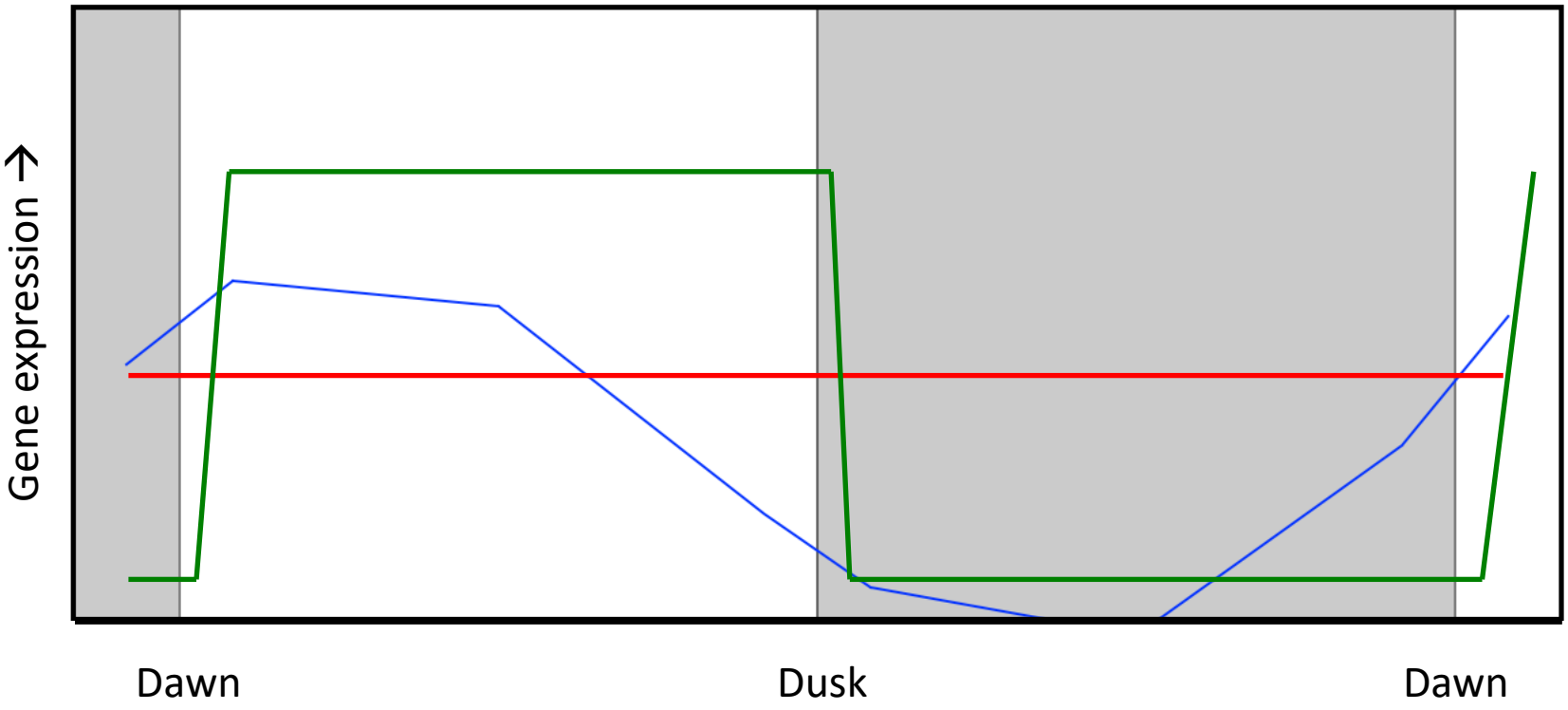
cDNA

- It’s hard to sequence single mRNA strands
- It’s easy to sequence DNA
- Add As/Cs/Gs/Ts ➔ mRNA recruits its own complement
- Result is almost like DNA
- U—A instead of T—A
- Amplification and sequencing technologies don’t notice the difference
- Just make cDNA and amplify/sequence as if it were DNA

The kinds of thing we can learn from metatranscriptomics

- Cyanobacteria = the phylum of photosynthetic bacteria.
- Ecologically very important to marine ecologists.
- Photosystem II genes (which code for light-harvesting proteins) are costly and don't last very long.
- What strategy do cyanobacteria use to optimize resource use?

Possible strategies for photosystem gene expression timing



- Worst: constant expression
- Better: trigger by light level
- Best: just-in-time manufacturing

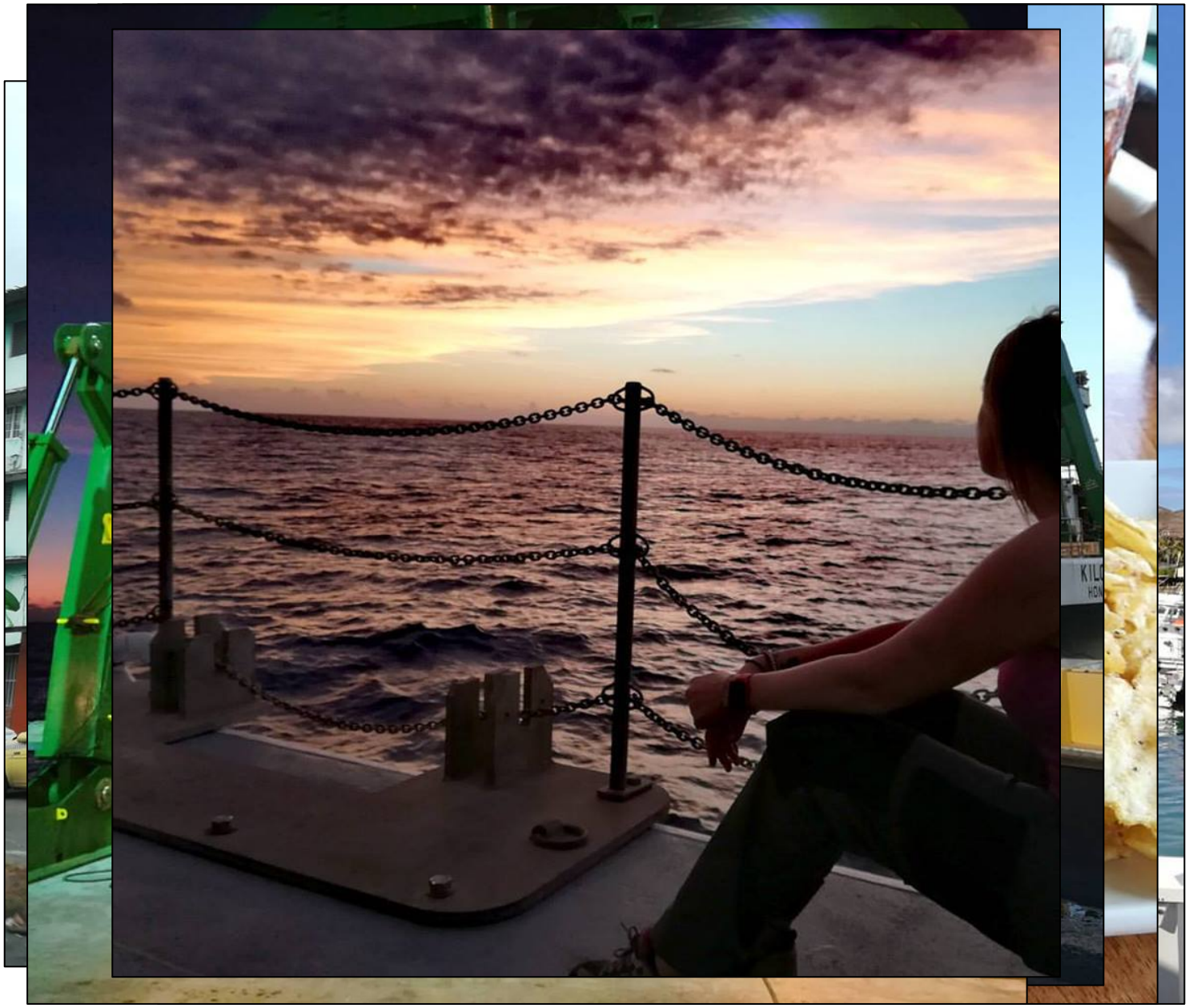
The steps of a metagenomic pipeline

1) Collect
2) Sequence
3) Merge
4) Quality trim
5) Identify
6) Analyze

Step 1: Collect

- As a bioinformatician, you probably won't get to participate.
  - Expeditions are expensive.
- Collection might include size filtering
  - Eukaryotes, prokaryotes, and viruses all exist at different size scales

My colleague's recent Facebook photos

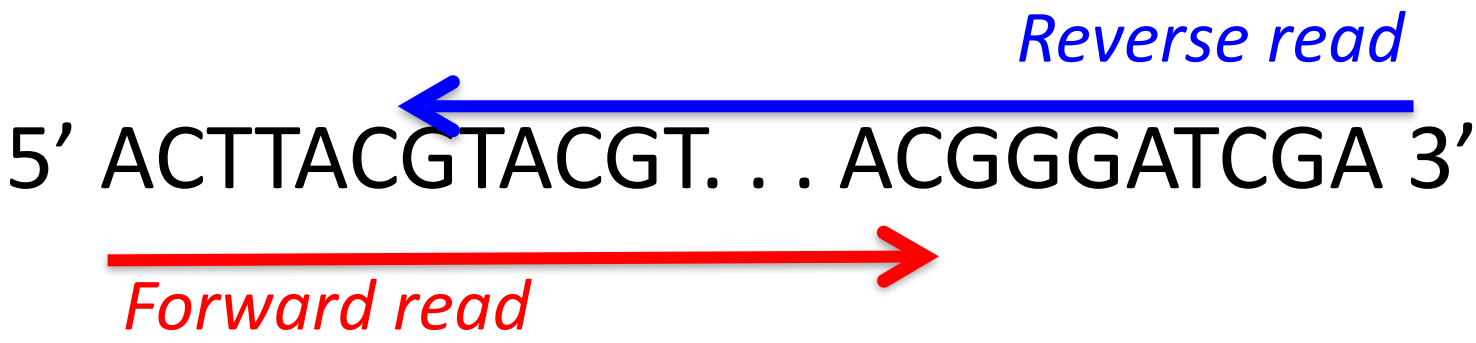


The steps of a metagenomic pipeline

1) Collect 
2) Sequence ←
3) Merge
4) Quality trim
5) Identify
6) Analyze

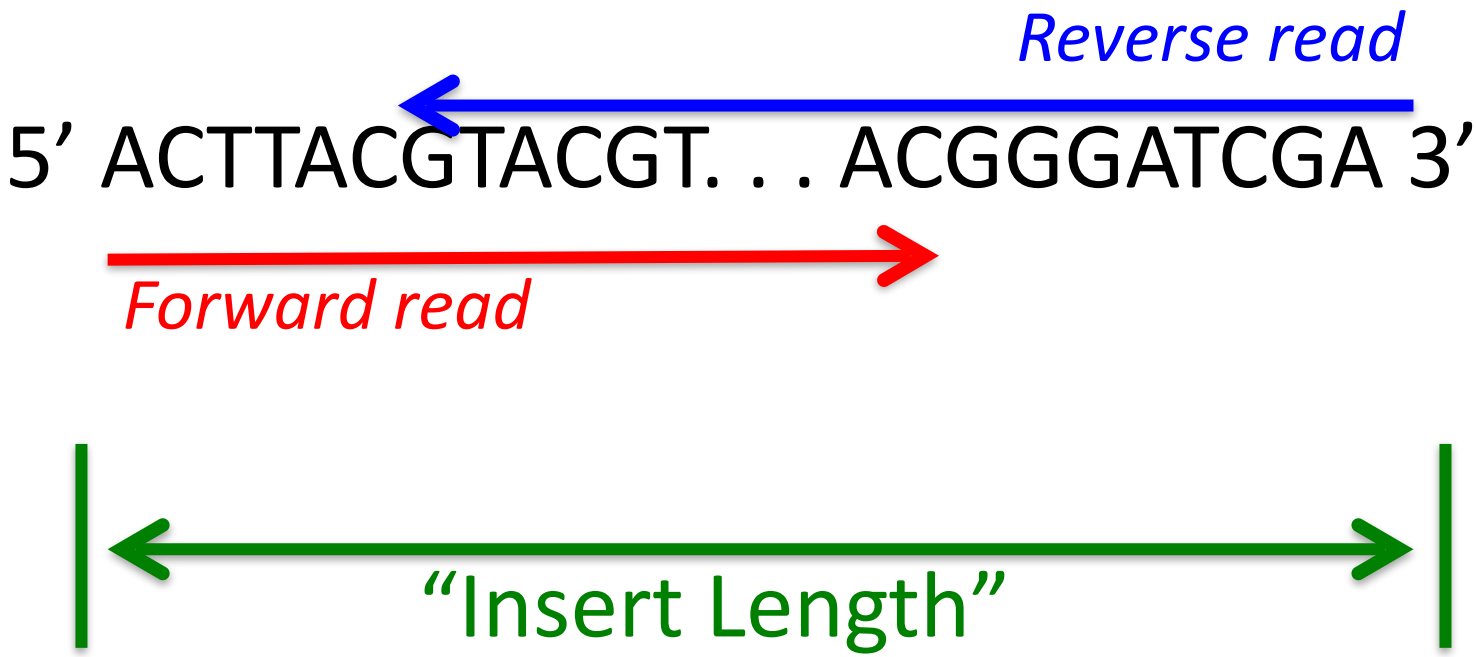
Step 2: Sequence

- Result is fastq file or 2 files
- Paired-end sequencing
  - Next-Gen sequencing (pyrosequencing) can only reliably read 200-800 bases from 5’ of a fragment
  - Best quality is near 5’ end of fragment, gets progressively worse in 3’ direction
  - Therefore quality near 3’ can be pretty bad
  - So sequence from both ends



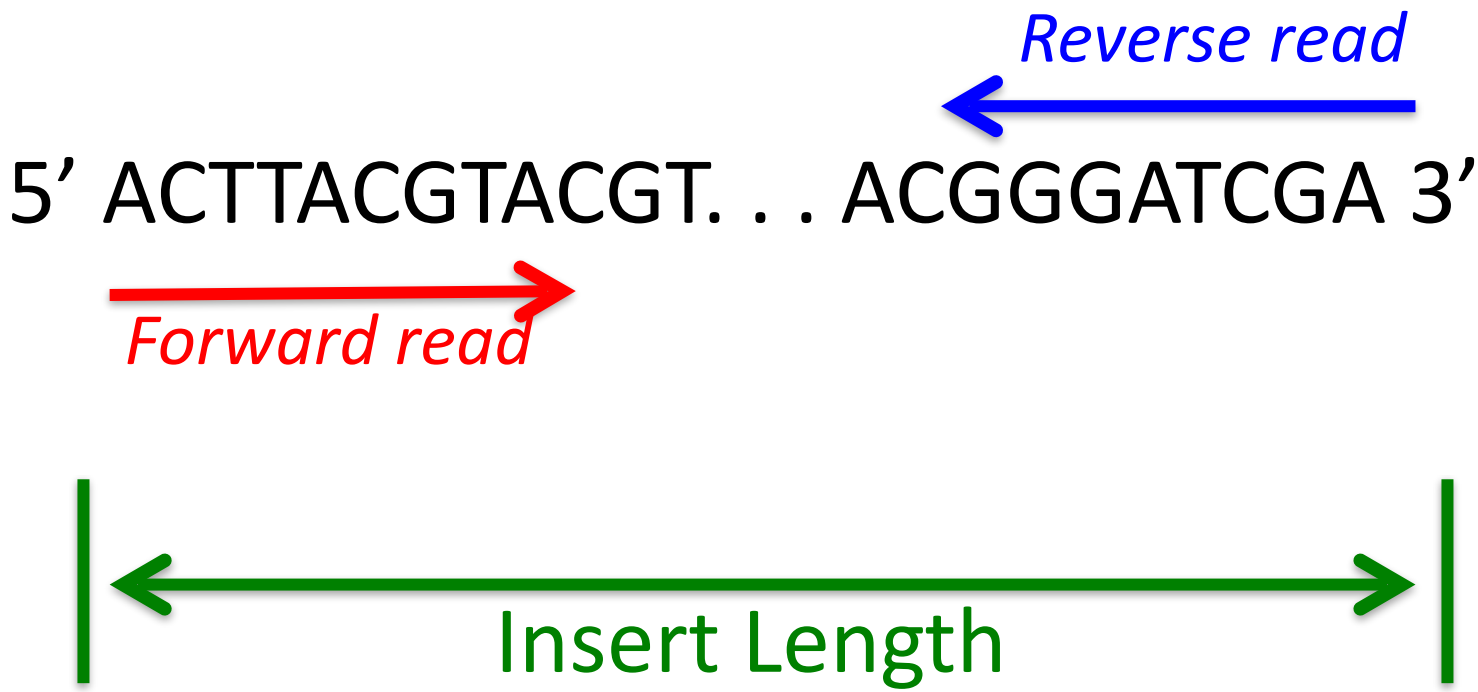
Paired-end sequencing: bonus information

- The “insert length” is known, and roughly constant for all reads

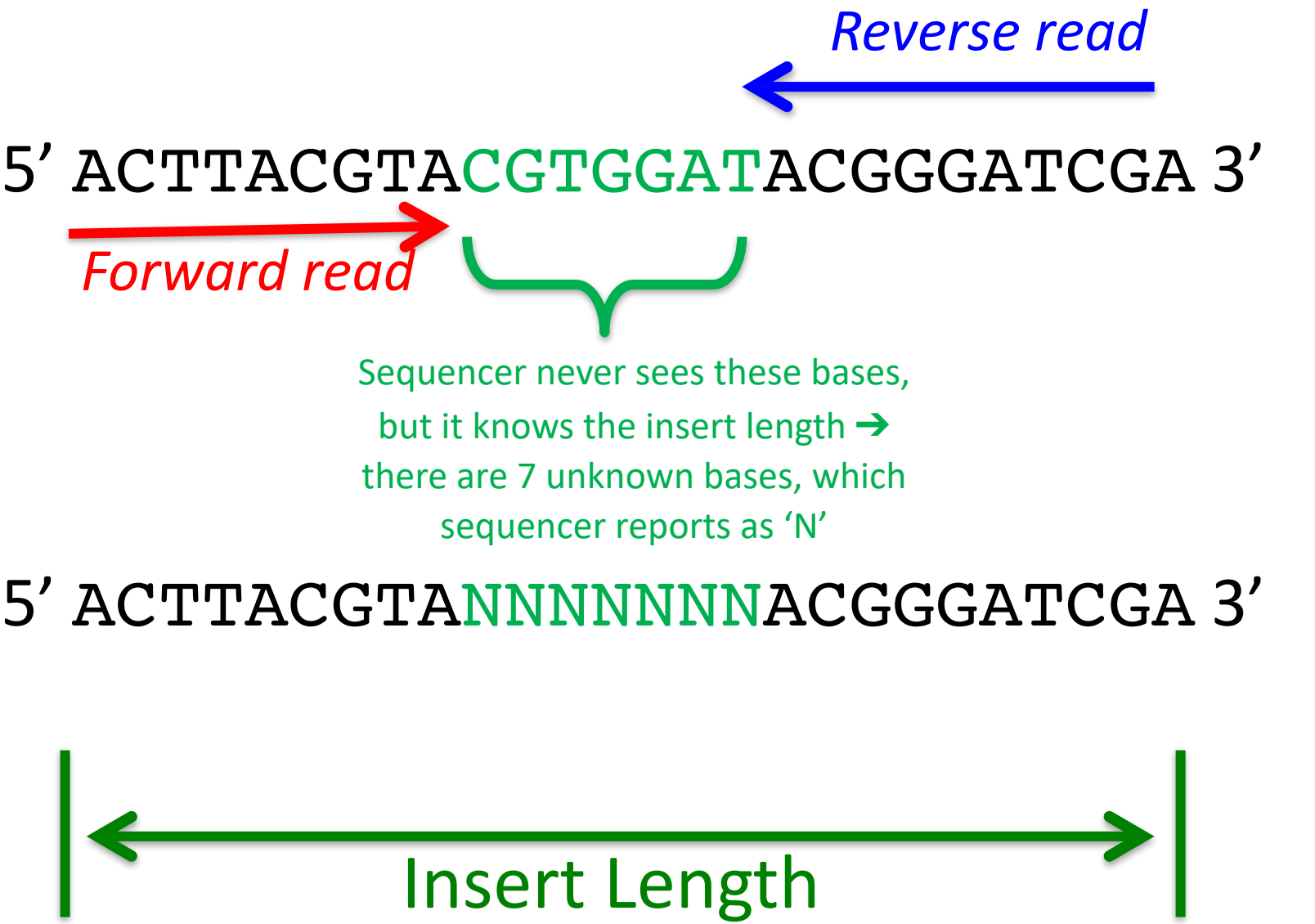


Paired-end non-overlapping reads

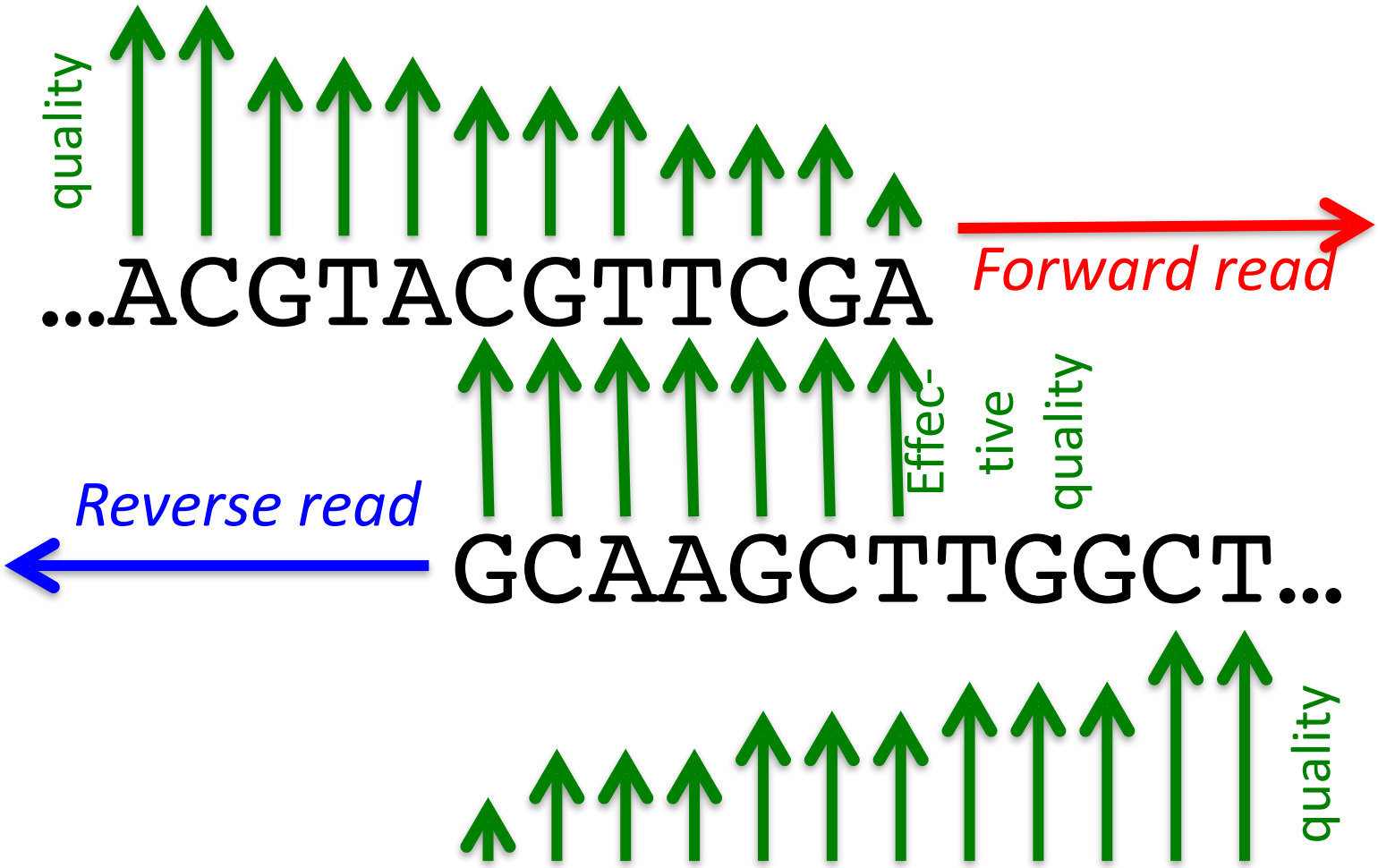
- If insert length > forward read length + reverse read length
- Relative positions of the pair of reads is valuable information for assemblers



Paired-end non-overlapping reads



Paired-end overlapping reads: bonus quality where you need it most



Paired-end overlapping reads

- 3’ ends of reads have poorest quality
- But they overlap, so it’s meaningful if they agree
- If 2 unreliable witnesses independently report the same event, P(event really happened) is high

Paired-end fastq files

- Fastq format includes sequence and base-by-base quality scores
- 2 files per sample (“library”)
  - Forward fastq
  - Reverse fastq
  - Nth record in each file matches Nth record of the other file

The steps of a metagenomic pipeline

1) Collect 
2) Sequence 
3) Merge ←
4) Quality trim
5) Identify
6) Analyze

Step 3: Merge

- Here's where the Bioinformatics begins.
- 2 fastqs of relatively short reads become 1 fastq of longer sequences.
- Non-overlapping reads: no information about the bases in the middle.
  - Insert "N"s between forward and reverse read
  - Make total length = insert length
  - This is why you sometimes see runs of "N"s in GenBank
  - Translate to protein ➔ sometimes see B/J/O/U/X/Z
- Overlapping reads
  - Quality scores of overlap positions could be better or worse than in original fastas
  - Better if both reads agree
  - Worse if they disagree

The steps of a metagenomic pipeline

1) Collect 
2) Sequence 
3) Merge 
4) Quality trim ←
5) Identify
6) Analyze

Step 4: Trim

- Also called “Quality Trimming”
- Choose a quality threshold, below which you don’t trust a read.
- Throw away any reads where < 95% of bases meet or exceed the threshold
- Convert the fastq input to fasta format
  - Lose quality codes

The steps of a metagenomic pipeline

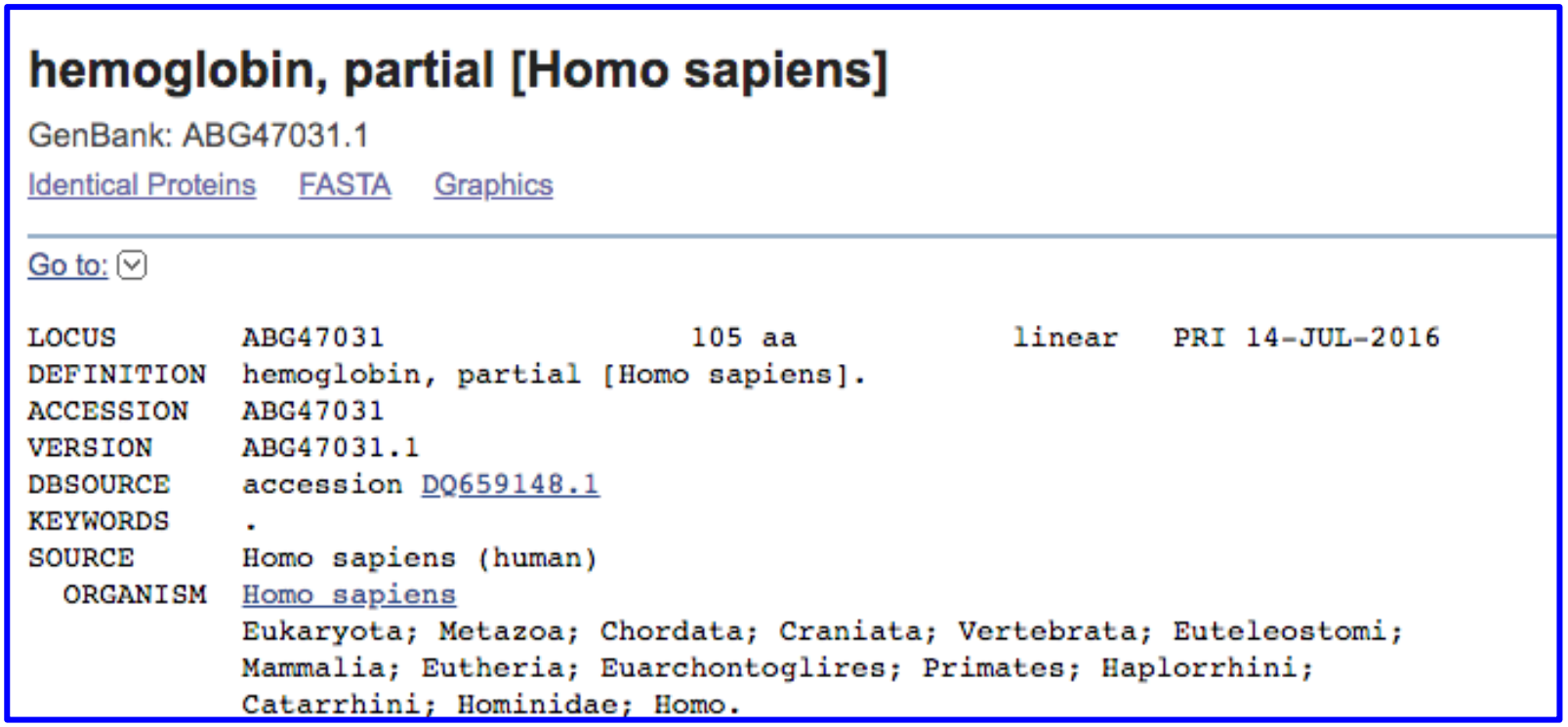
1) Collect 
2) Sequence 
3) Merge 
4) Quality trim 
5) Identify ←
6) Analyze

Step 5: Identify

- Identification approaches are alignment-based
  - Blastn each read against a “reference database”
    - Usually GenBank
    - Custom vouchered reference database

Blasting against GenBank

- GenBank contains organism and function annotations



hemoglobin, partial [Homo sapiens]

GenBank: ABG47031.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       ABG47031                 105 aa            linear   PRI 14-JUL-2016
DEFINITION  hemoglobin, partial [Homo sapiens].
ACCESSION   ABG47031
VERSION     ABG47031.1
DBSOURCE    accession DQ659148.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

Blasting against GenBank

- GenBank contains organism and function annotations
  - FWIW!
  - Whether you care about organism or function, you have the information you need
  - Information can be hard to deal with (or even wrong!)
    - E.g. there are 220 synonyms for COI, some due to spelling errors

Some of the 220 annotations of COI

- Cytochrome c oxidase subunit I
- Cytochrome c oxidase I subunit
- COI
- CO1
- COX1
- ctochrome oxidase I (oops)
- ctyochrome oxidase subunit 1 (oops)
- cyotchrome c oxidase 1 (oops)
- cyotochrome oxidase subunit I (oops)
- cytchrome oxidase subunit I (oops)
- 210 others

Sidebar

- Collecting all records of a gene of interest from GenBank is *hard*.
- You've seen FunGene … non-seed sequences are suspect.
- Collecting based on software inspection of gene name annotation is too error-prone.
  - Correctly identified genes often have misspelled names.
  - Too many genes are incorrectly annotated.
- The ARBitrator algorithm (2014) is promising.

A sharper tool than GenBank: Custom voucher-based databases

- Expensive.
- Best bioinformatics practices are not yet developed.

Voucher-based studies

- An expert identifies an organism
- Extract tissue from the organism
- Sequence some of the tissue
- Cold-store remaining tissue: the “voucher”
  - In case of controversy, the voucher *vouches for* the identification
  - “You say that’s vampire squid DNA? Prove it!”
  - Can’t do that with GenBank records

Aquarium of the Pacific

Yes, vampire squid is a thing.

Voucher-based studies

- Sample an environment
- Compare sampled sequences against vouchered database
  - Blast reads against custom database
  - Usually E-value of best hit is << (much much better than) other hits
  - Custom database, so need to develop new intuitions about range of E-values that mean strong hits
- Advantages over GenBank:
  - *Much* higher confidence in identity of subjects
  - Faster blast
- Disadvantages:
  - You can't identify anything that isn't in your database

Voucher-based study example: Invertebrate invasive species in Monterey Bay

- Causes
  - Ballast water dumping
  - Climate change ➔ Northward incursion of cold-intolerant species
- Effects
  - Hardy invaders displace traditional members of ecological niches (like weeds)
  - Predator/prey relationships are disrupted
  - Fisheries are impacted
    - Ecology affects economy

Why we need metagenomics to study invasion

Conservation Biology

Explore this journal >

Decline of a Native Mussel Masked by Sibling Species Invasion

Jonathan B. Geller

First published: June 1999 Full publication history

Cryptic species

- "Cryptic" means can't be visually distinguished from a different species.
- Invasive species are often hard to distinguish from natives.
- Example: mussels

Mytilus trossulus
Native to California coast

Mytilus galloprovincialis
Invaded southern California
? 19th century ? Early 20th ?

Voucher-based study example: Monterey Bay ARMS

- A.R.M.S. = Autonomous Reef Monitoring Structure
- = A bunch of cheap plastic, bolted together
- 10 ea 1’ x 1’ PVC squares, 1” separation

Autonomous Reef Monitoring System: A.R.M.S.

Voucher-based studies can be *really* gross

- Separate the plates
- Remove macro-scale individuals
- Scrape slime & goo into a bowl
  - Includes *M. trossulus* and *M. galloprovincialis slime & goo*
- Emulsify with a hand blender
  - Like this, but it's not raspberries!
- Extract DNA
- Use primers & PCR to amplify COI
- Sequence
- Blast results against vouchered database
  - Check ratio of *M. trossulus* to *M. galloprovincialis* sequences

The steps of a metagenomic pipeline

1) Collect ✓
2) Sequence ✓
3) Merge ✓
4) Quality trim ✓
5) Identify ✓
6) Analyze ←

The steps of a metagenomic pipeline

1) Collect
2) Sequence
3) Merge
4) Quality trim
5) Identify
6) Analyze


Case Study: UCYN-A


10^24 stars
10^29 bacteria

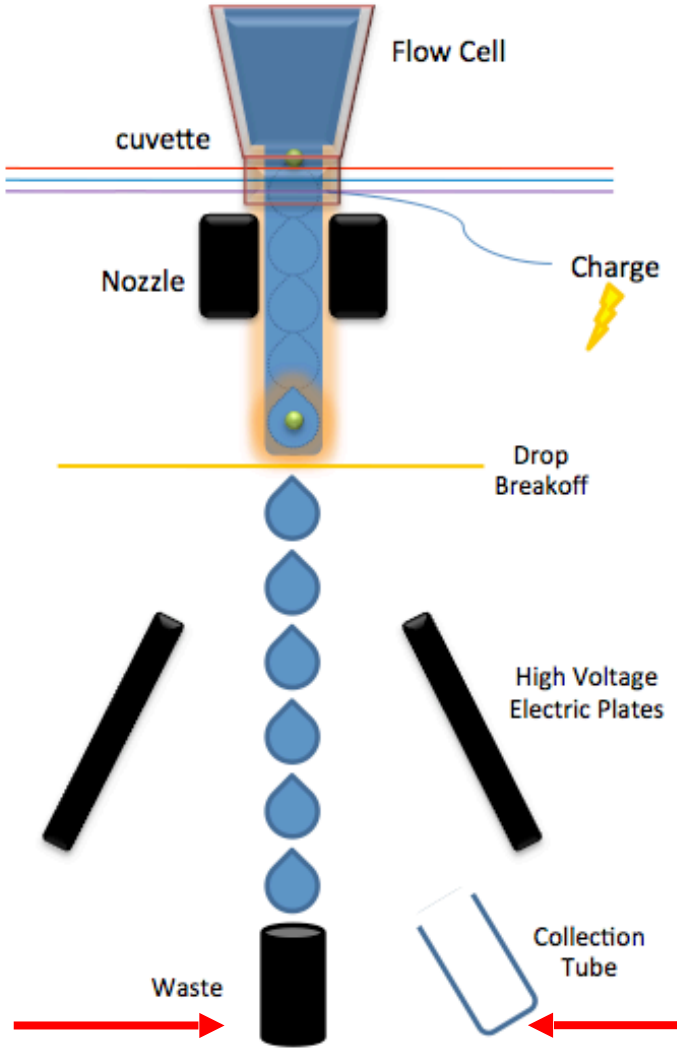
But this is relatively new knowledge

- The old belief:
  - Green / turquoise / light blue = chlorophyll = life
    - Only in shallow water, near land
  - Dark blue = desert
    - The open ocean

Sallie (Penny) Chisholm

https://www.ted.com/talks/penny_chisholm_the_tiny_creature_that_secretly_powers_the_planet#t-985653

Flow Cytometers don't just *analyze*, they can *separate* (cell sorters)

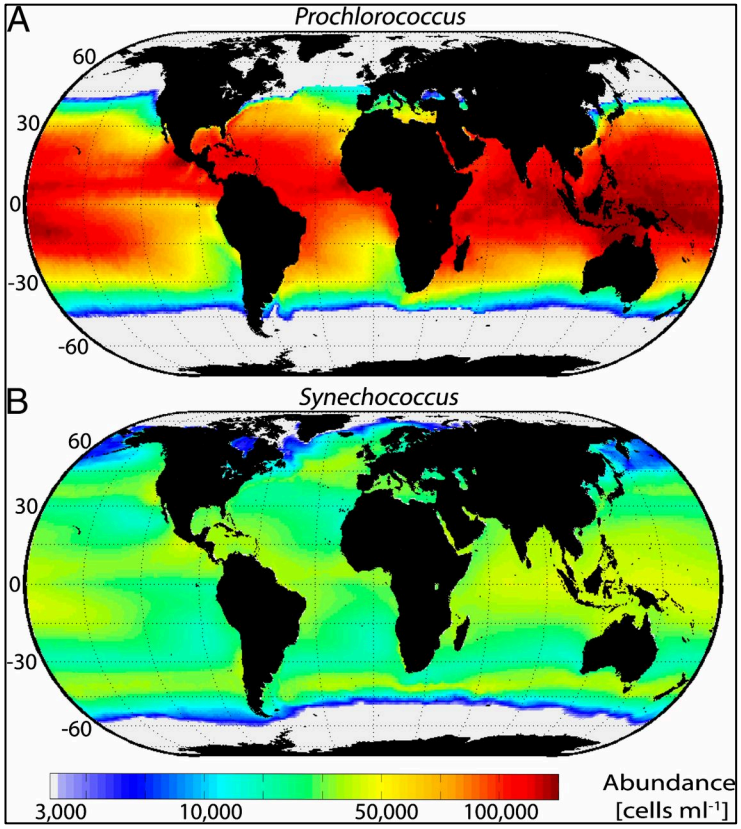


What you don't want

What you want

Prof. Chisholm's great discovery

- Flow cytometers work on ships
- Genus *Prochlorococcus*
- A tiny cyanobacterium (photosynthesizer)
- Found in much of the open ocean except high northern and southern latitudes
- (Synechococcus: the previously known open-ocean tiny cyano … note much lower abundance)



Flombaum et al. PNAS 2013

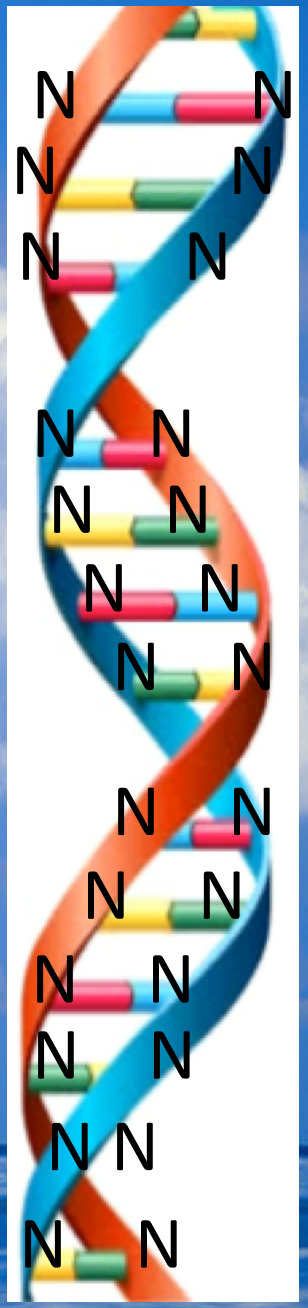
Link to article in note for this slide

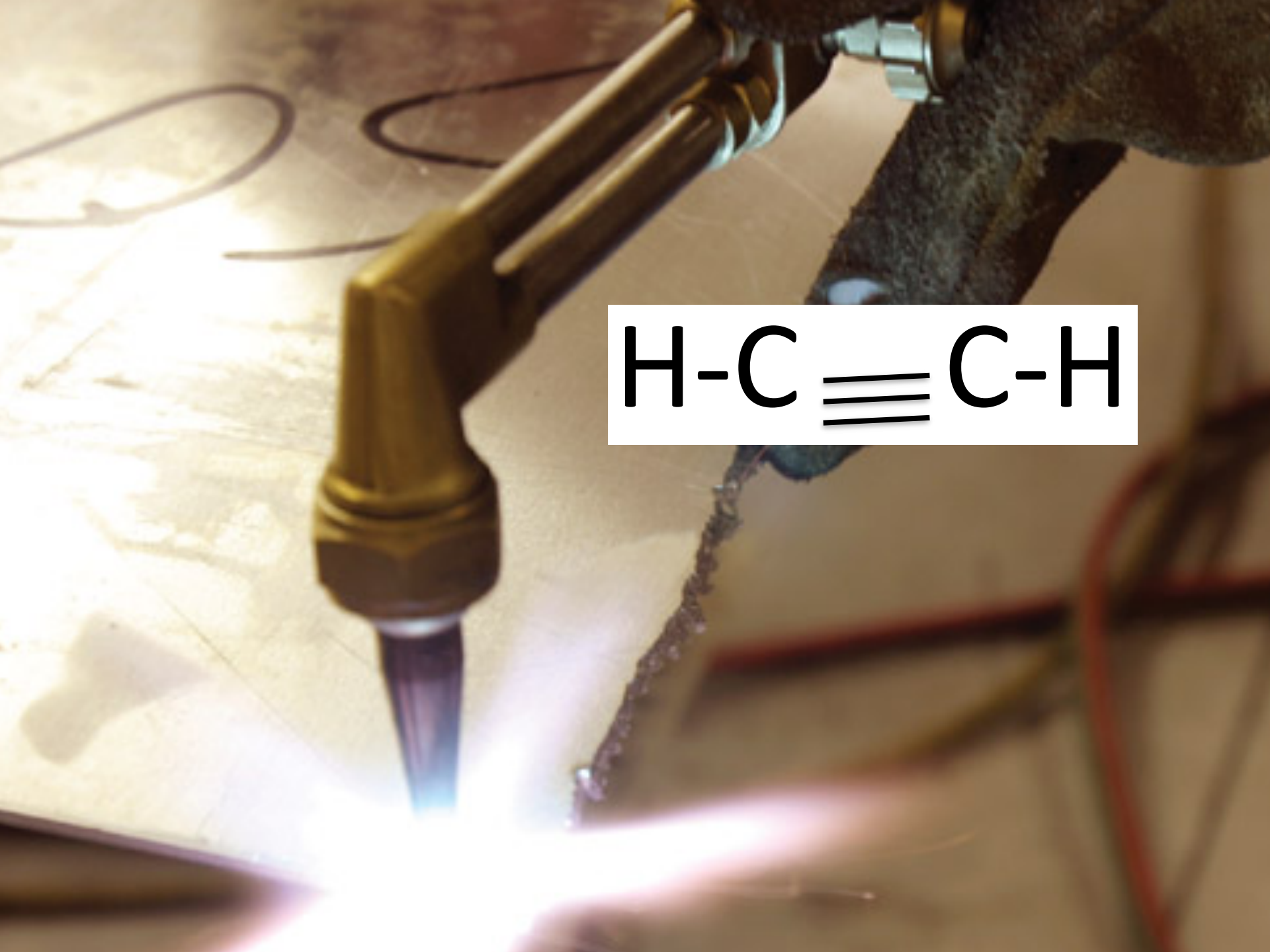
So for a long time

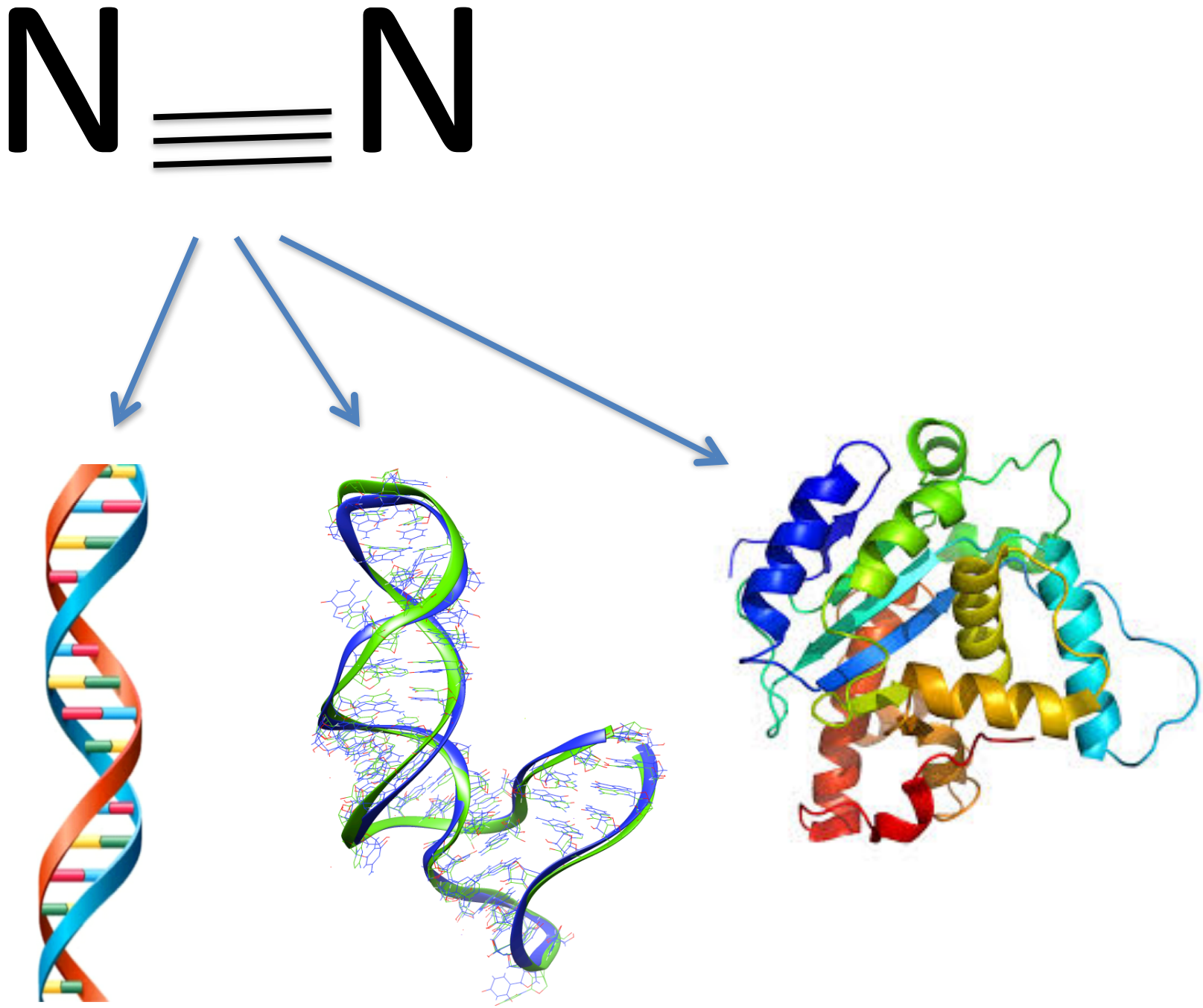
- Prochlorococcus and Synechococcus were believed to be the main open-ocean bacteria.
- But the ecological budget didn’t balance.

Which brings us to nitrogen

$N \equiv N$




H-C≡C-H

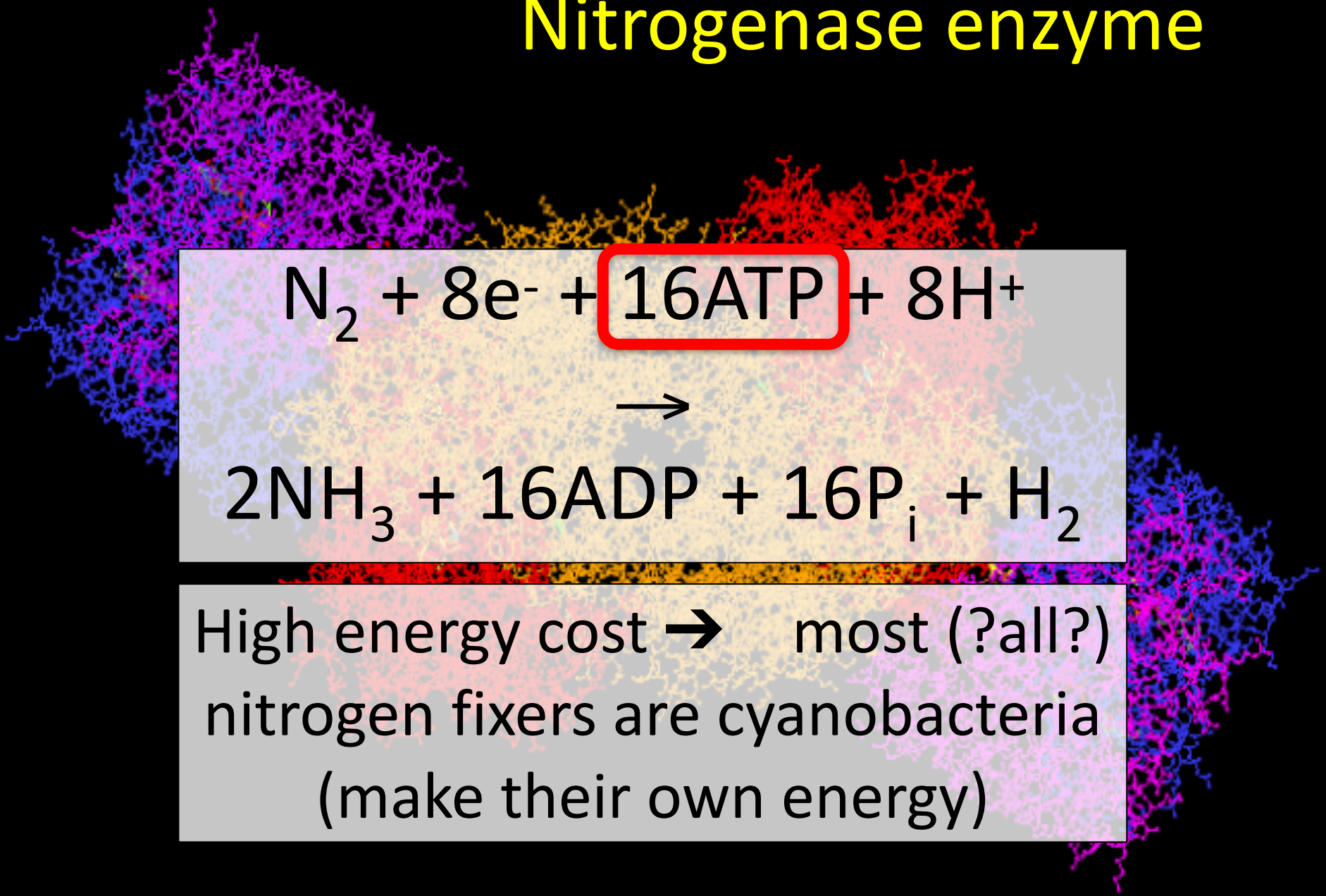

N
N

Nitrogenase enzyme

$$N_2 + 8e^- + 16ATP + 8H^+ \rightarrow 2NH_3 + 16ADP + 16P_i + H_2$$

High energy cost → most (?all?) nitrogen fixers are cyanobacteria (make their own energy_

Nitrogenase enzyme

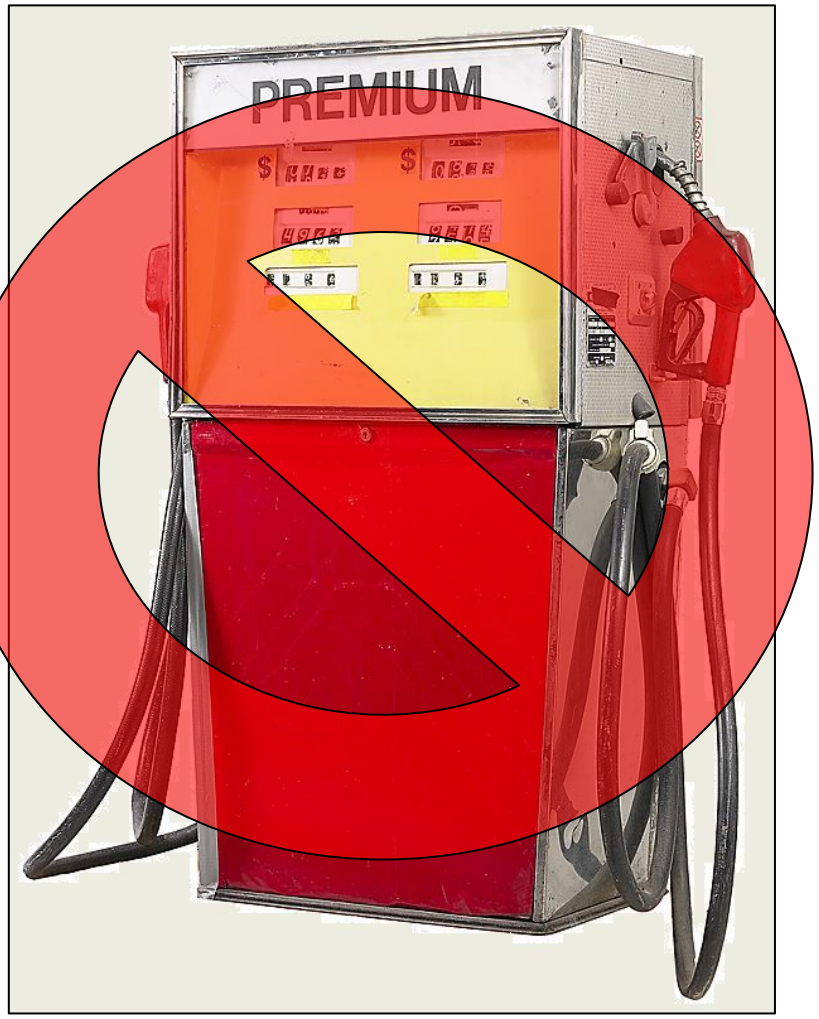


$$N_2 + 8e^- + 16ATP + 8H^+ \rightarrow 2NH_3 + 16ADP + 16P_i + H_2$$

High energy cost → most (?all?) nitrogen fixers are cyanobacteria (make their own energy)

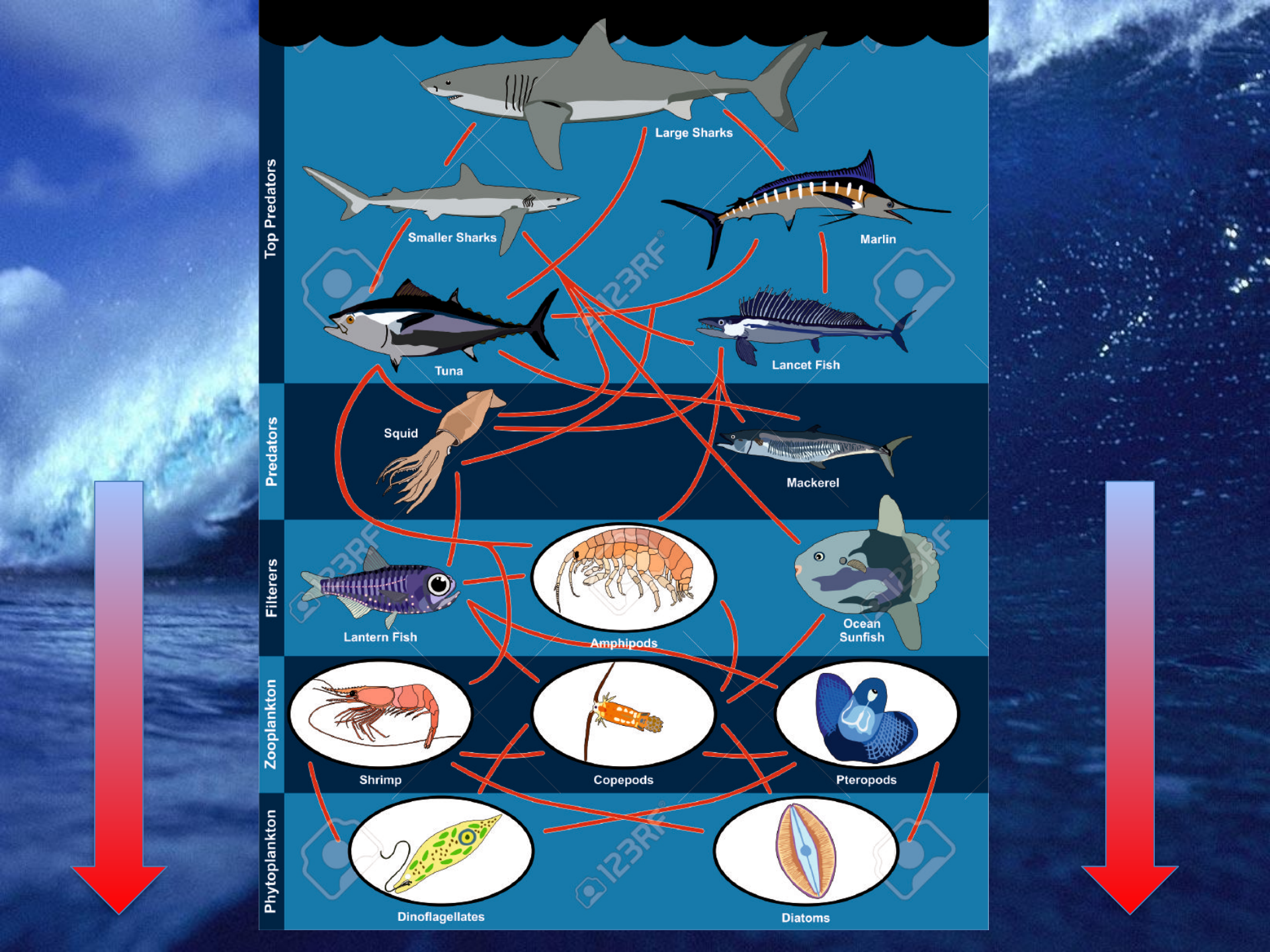
Marine nitrogen fixers drive a biogeochemical cycle that keeps us alive

- The “carbon pump”




PREMIUM


SCIENCEphotoLIBRARY


Top Predators
Large Sharks
Smaller Sharks
Marlin
Tuna
Lancet Fish
Predators
Squid
Mackerel
Filterers
Lantern Fish
Amphipods
Ocean Sunfish
Zooplankton
Shrimp
Copepods
Pteropods
Phytoplankton
Dinoflagellates
Diatoms

- Until 15 years ago, believed to be the only significant marine nitrogen fixer
- Easy to see

Trichodesmium alone doesn't account for all marine nitrogen fixation

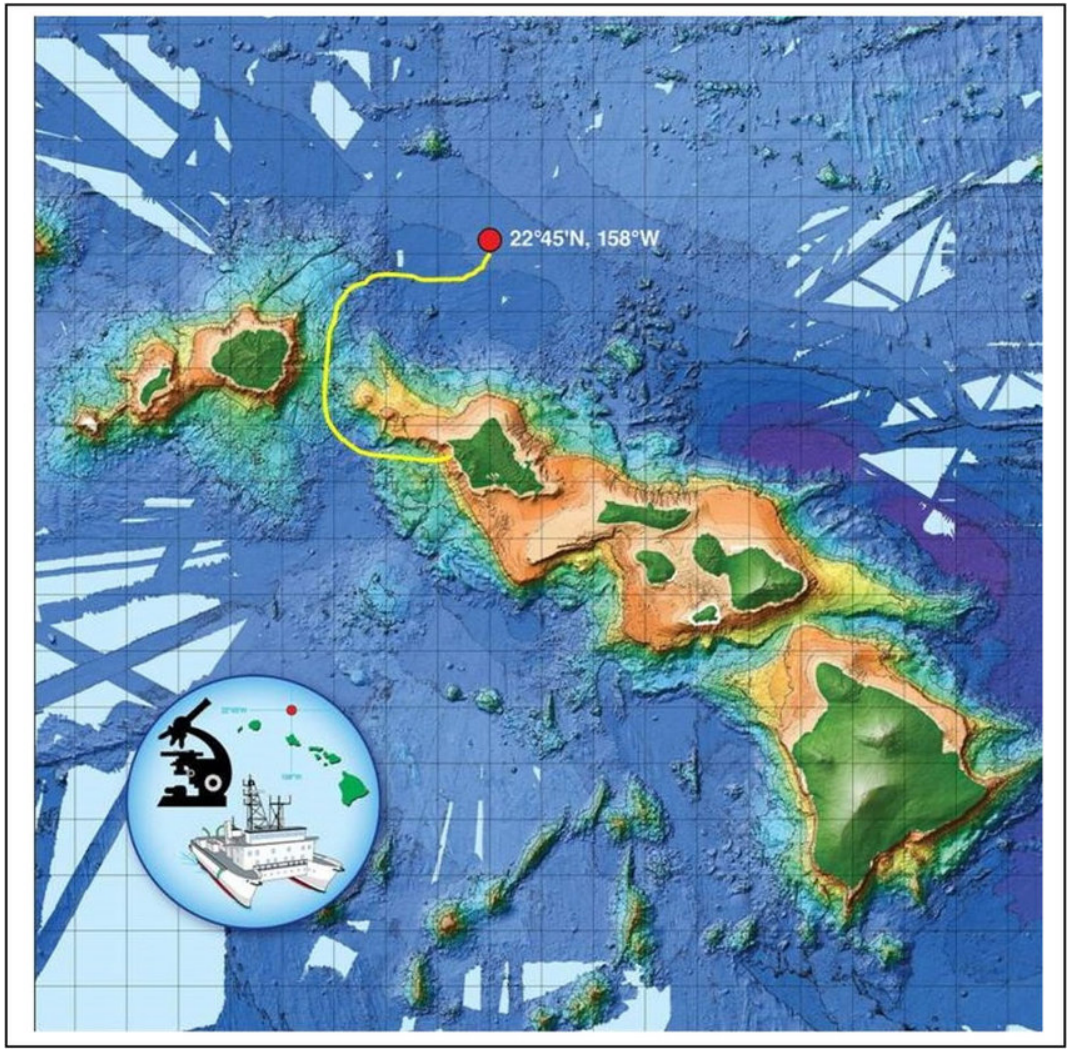
- "Direct estimates of N_2 fixation, largely of … *Trichodesmium*, can account for only a quarter to one-half of the geochemically derived rates of N_2 fixation in various ocean basins."

  - Mahaffey et al., 2005, Am.J. of Sci.

"Geochemically derived" = from computer models

- What else in the oceans is fixing 70 – 105 Tg of nitrogen annually?
  - ~= 1/10,000 of the atmosphere

A metagenomic search for marine nitrogen fixers at Station Aloha



A metagenomic search for marine nitrogen fixers at Station Aloha

- Sample the open ocean
- PCR with degenerate nifH primers
  - Wild-card search for nifH genes among all that DNA

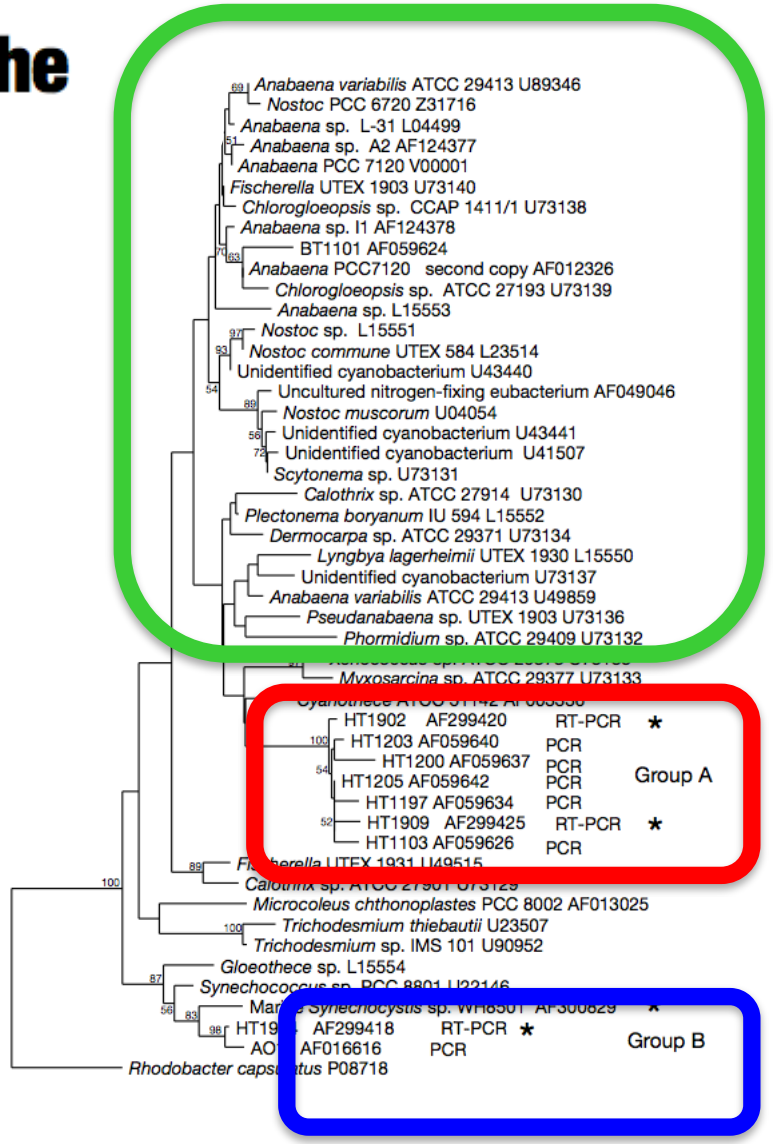
Discovery of unicellular marine nitrogen fixers

Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean

Jonathan P. Zehr*, John B. Waterbury†, Patricia J. Turner*, Joseph P. Montoya‡, Enoma Omoregie*, Grieg F. Steward*, Andrew Hansen§ & David M. Karl§

-- Nature, 2001

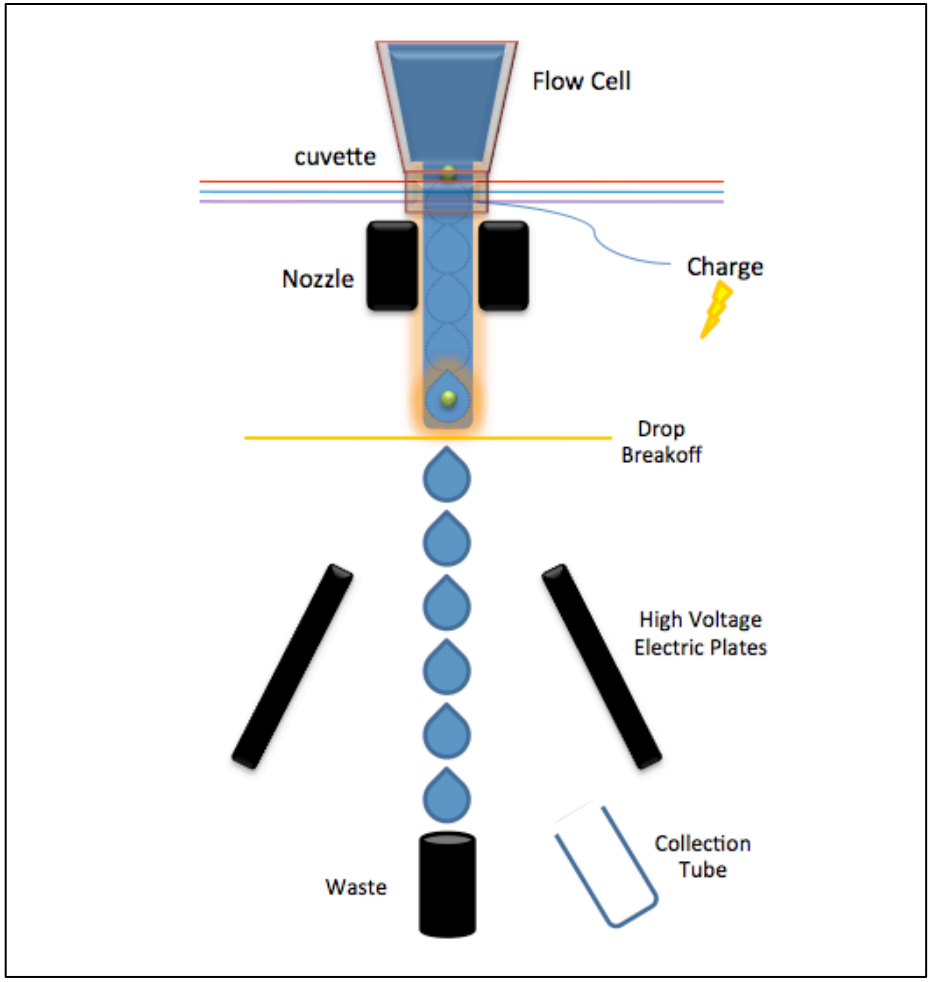
UCYN-A



Group B = *Crocosphaera watsonii*

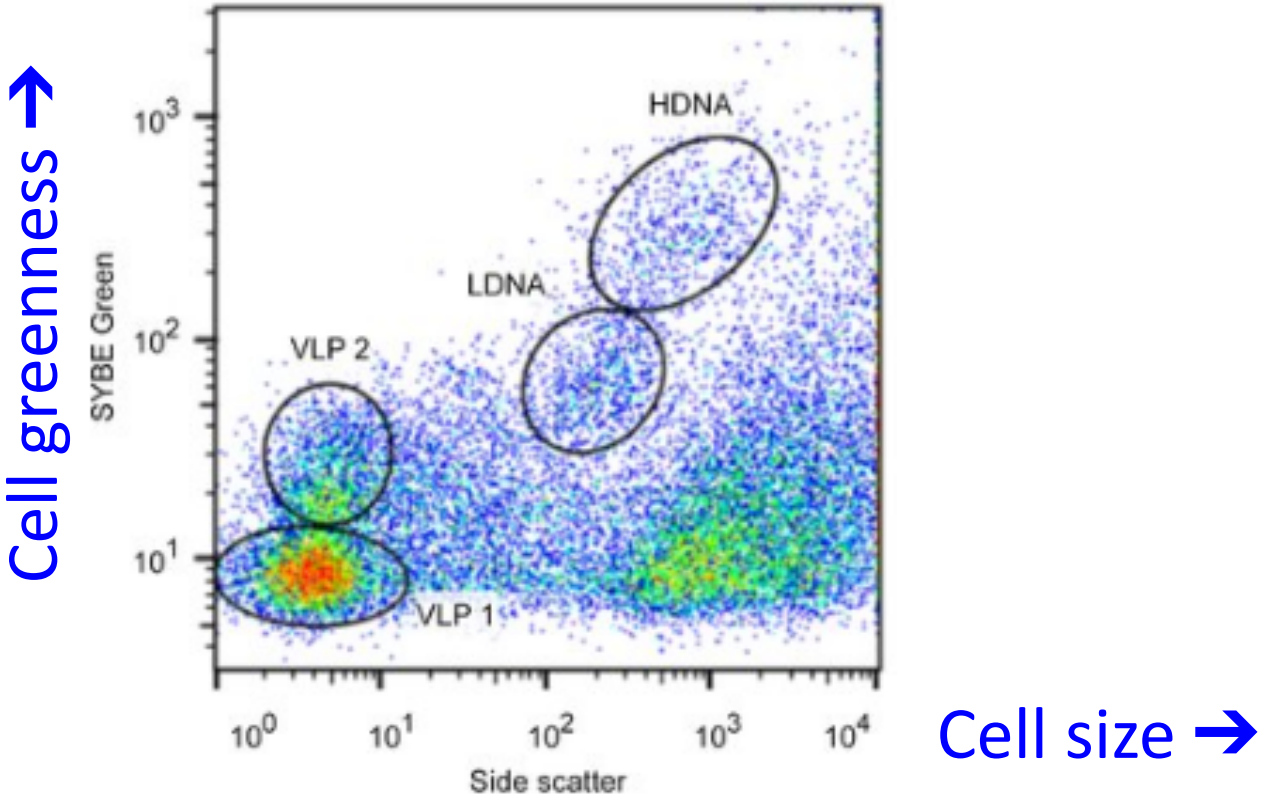
The Jonathan Zehr lab at U.C. Santa Cruz investigated UCYN-A using a flow cytometer

Remember Flow Cytometers?



Cytogram

- A heatmap of cells seen by a flow cytometer, by size (horizontal axis) and green intensity (vertical axis).
- Cytometer can be programmed to collect cells within a specified range of (size, color).

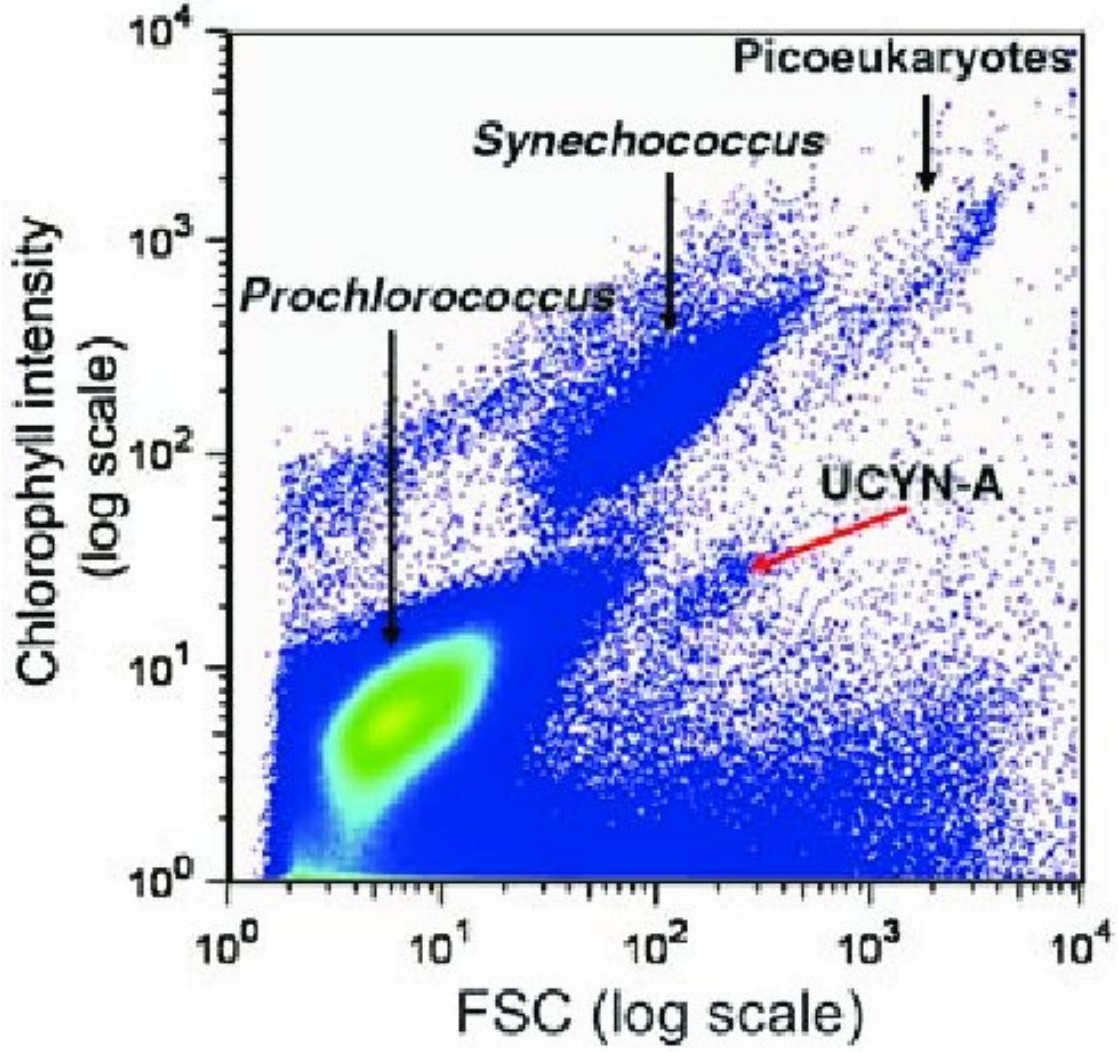


Exhaustive search of (size, greenness) settings

1. Sort water sample (1 cell at a time, many hours)
2. Extract DNA
3. PCR to amplify nifH
4. Sequence
5. Look for sequences that aren't known species
6. If 5 fails, change settings and go to 1

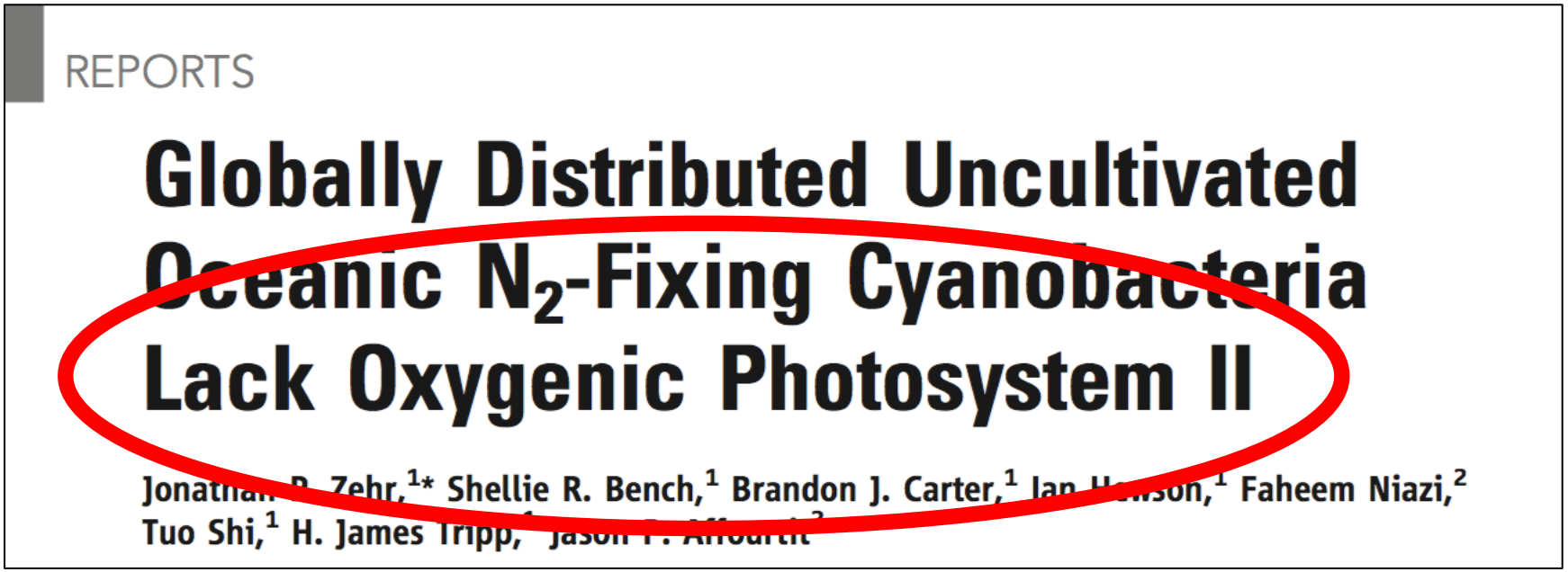
Heroic effort by Brandon Carter

Heroic effort by Brandon Carter



DNA from the UCYN-A sort region was sequenced and assembled

Nature 2008



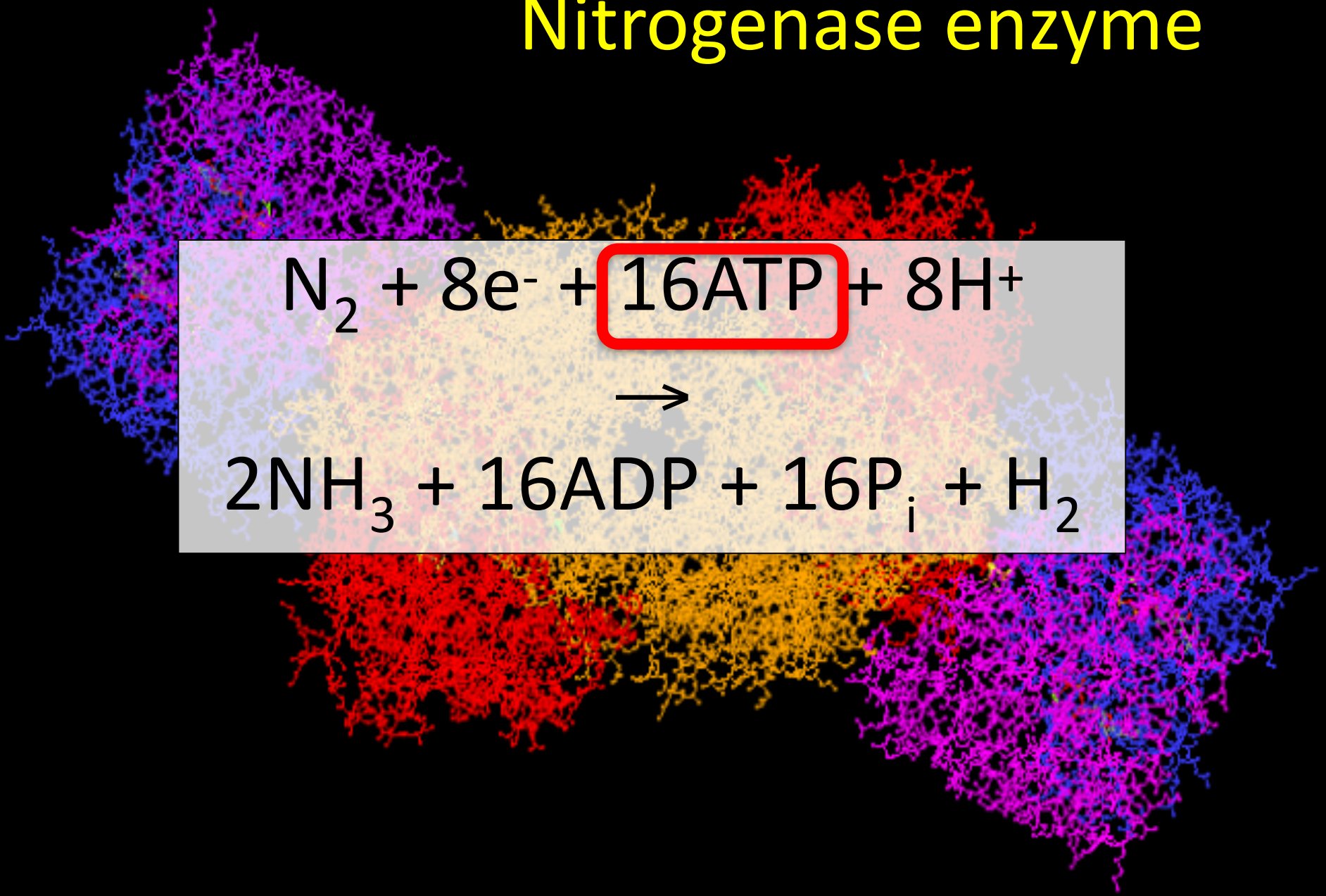
REPORTS

Globally Distributed Uncultivated Oceanic N_2-Fixing Cyanobacteria Lack Oxygenic Photosystem II

Jonathan P. Zehr,[1*] Shellie R. Bench,[1] Brandon J. Carter,[1] Ian Hewson,[1] Faheem Niazi,[2] Tuo Shi,[1] H. James Tripp,[1] Jason F. Affourtit[2]

How can it get enough energy to reduce nitrogen?

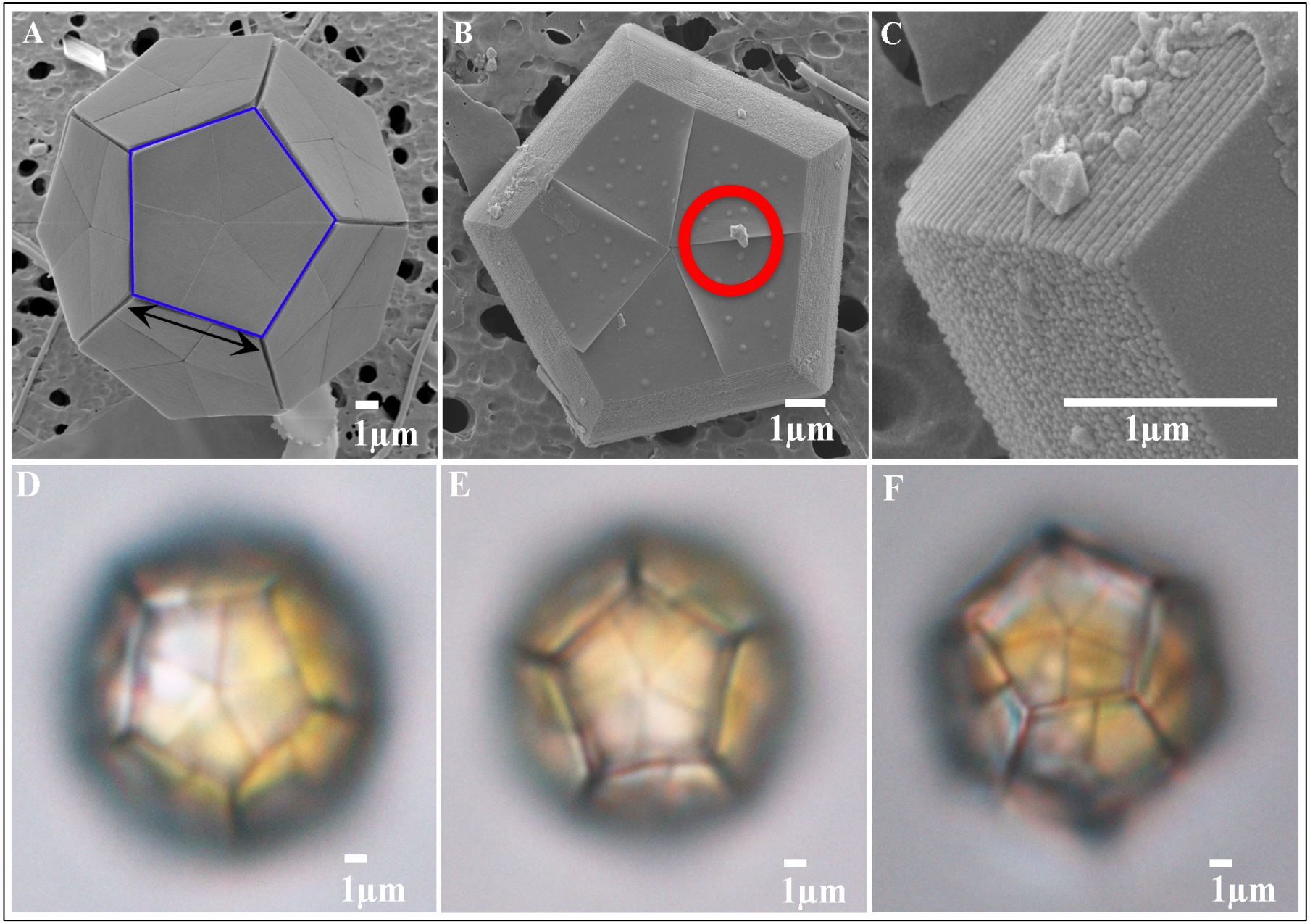
Nitrogenase enzyme



$$N_2 + 8e^- + 16ATP + 8H^+ \rightarrow 2NH_3 + 16ADP + 16P_i + H_2$$

Zehr's Insight: It's a symbiont

- Partner provides photosynthesis product (= ATP) to UCYN-A.
- UCYN-A provides fixed nitrogen to partner.
- Maye that's why UCYN-A is so hard to find: it's nestled inside its partner.



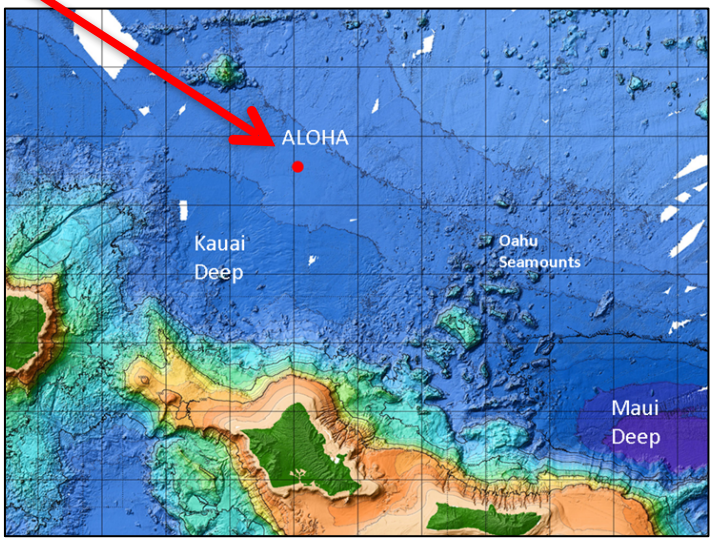
Source: Hagino et al., PLoS ONE 2013

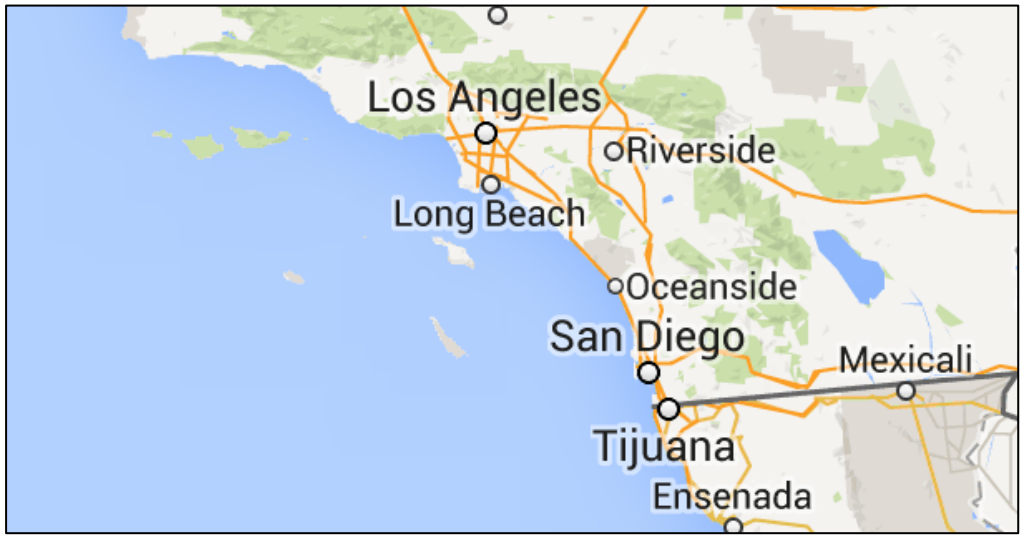
Our story so far …

- Something undiscovered is doing most of the marine nitrogen fixation on this planet.
- Discovery of UCYN-A in subtropical open ocean.
- Is UCYN-A the one? How would we know?
  - If it's so productive, it must be widespread
  - Let's look somewhere that *isn't* the subtropical open ocean

Does UCYN-A live in California coastal waters?

- Original assembly was from Station ALOHA.





- UCYN-A also observed in Santa Monica basin.
- Search using PCR primers were unproductive.
- What if there's a different strain, just divergent enough to escape PCR detection?

Discovery of a second strain ("UCYN-A2") in a radically different environment would suggest that UCYN-A is adaptable ➔ maybe widespread enough to be *the* mystery nitrogen fixer

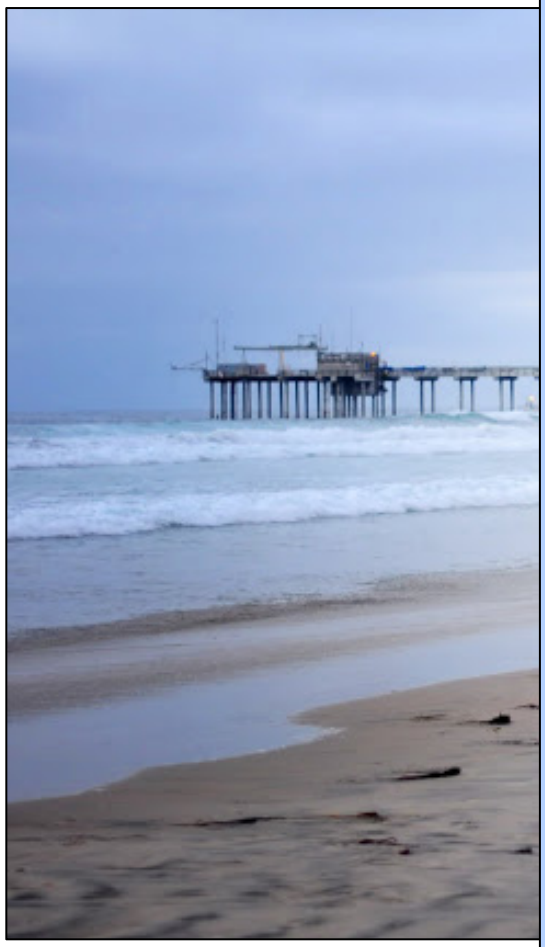
Station Aloha

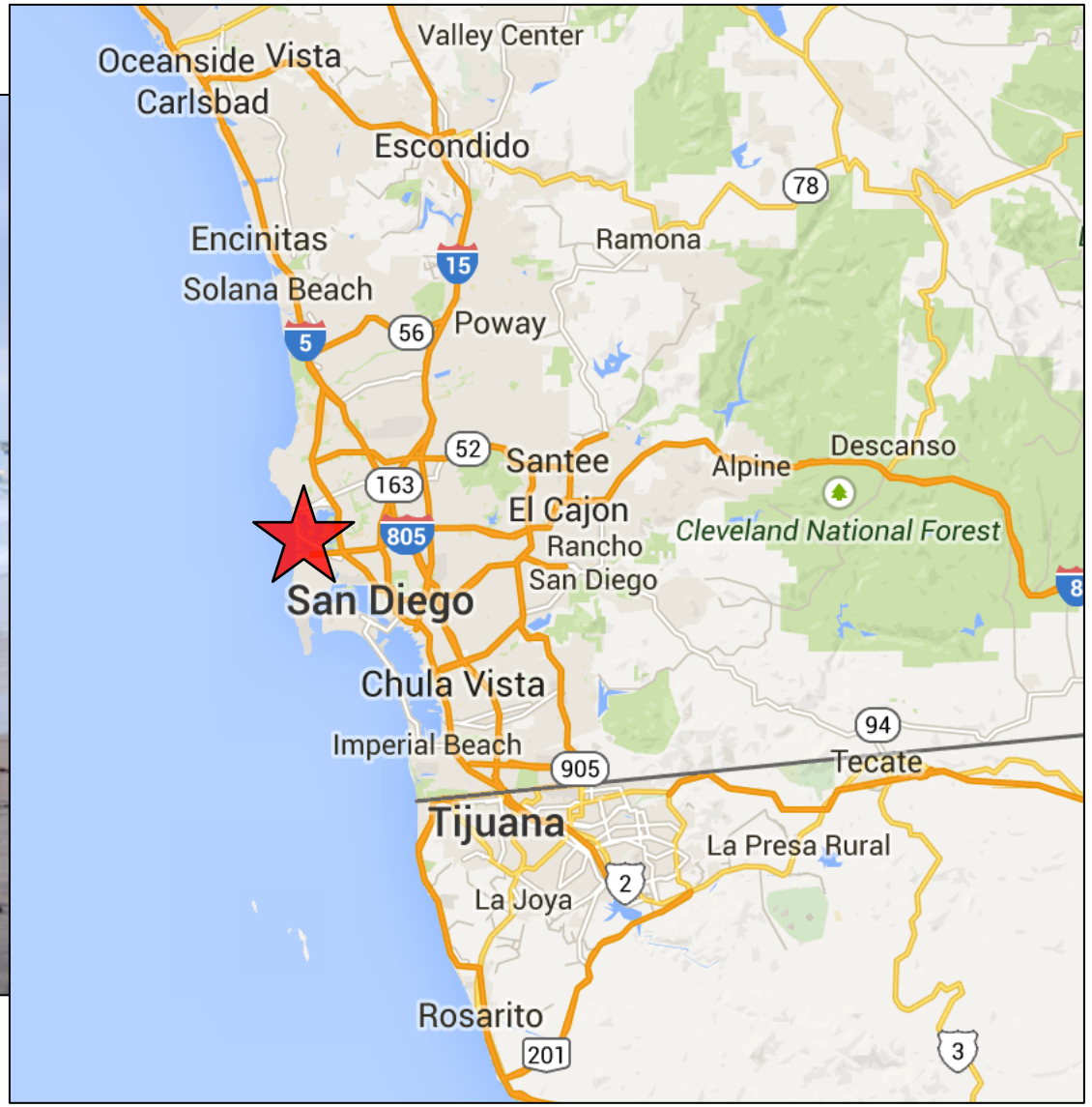
- Open ocean
- Dark blue water
- Few predators
- Chemically simple

La Jolla

- Coastal
- Lighter blue water
- More life, incl. predators
- Chemically complicated (runoff from land)

Coastal Sampling Site: Scripps Pier, La Jolla, CA





The ISME Journal (2014) 8, 2530–2542

www.nature.com/ismej

ORIGINAL ARTICLE

Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria

Deniz Bombar[1,4,5], Philip Heller[2,4], Patricia Sanchez-Baracaldo[3], Brandon J Carter[1] and Jonathan P Zehr[1]

[1]*Ocean Sciences Department, University of California, Santa Cruz, CA, USA;* [2]*Biomolecular Engineering Department, University of California, Santa Cruz, CA, USA and* [3]*Schools of Biological and Geographical Sciences, University of Bristol, Bristol, UK*

strains

"UCYN-A2"

Today: Many successful searches for known and new UCYN-A strains

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 April 2018
doi: 10.3389/fmicb.2018.00554

Check for updates

Distributions and Abundances of Sublineages of the N_2-Fixing Cyanobacterium *Candidatus* Atelocyanobacterium thalassa (UCYN-A) in the New Caledonian Coral Lagoon

Britt A. Henke[1], Kendra A. Turk-Kubo[1], Sophie Bonnet[2] and Jonathan P. Zehr[1]*

[1] Department of Ocean Sciences, University of California, Santa Cruz, Santa Cruz, CA, United States, [2] IRD, MIO, UM 110 – IRD Centre of Noumea, Aix-Marseille University, University of South Toulon Var, CNRS/INSU, Noumea, France

The UCYN-A Team
(Zehr Lab, UCSC, 2011)