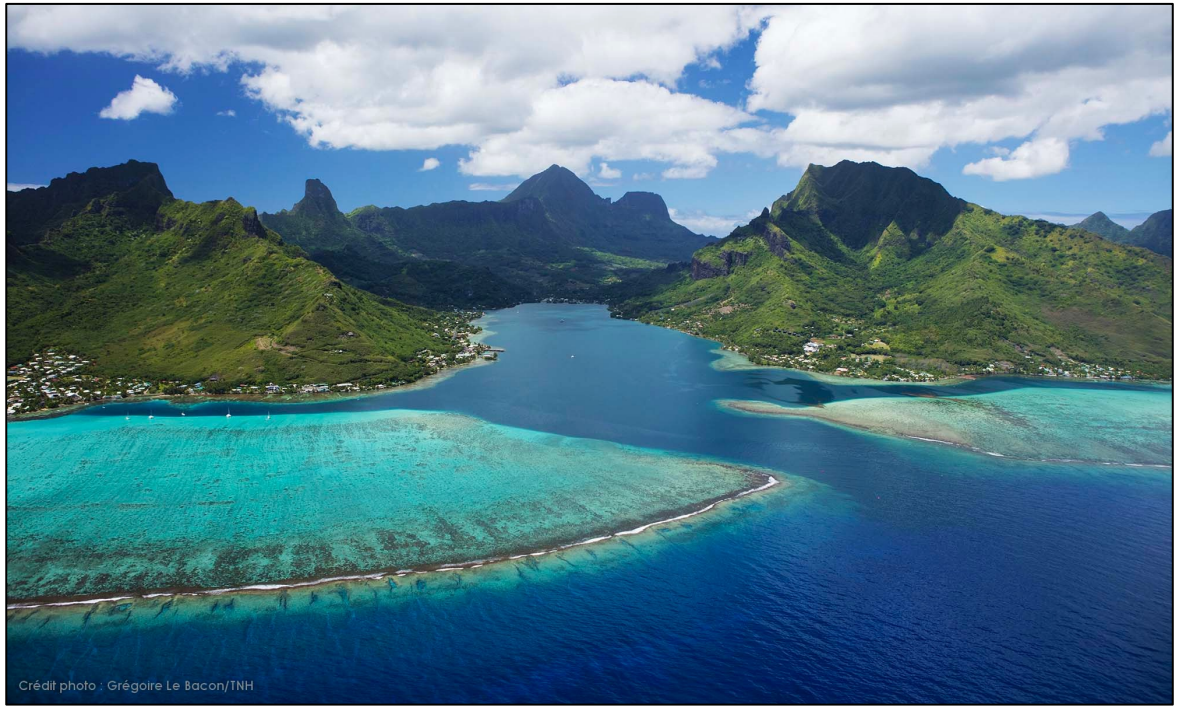
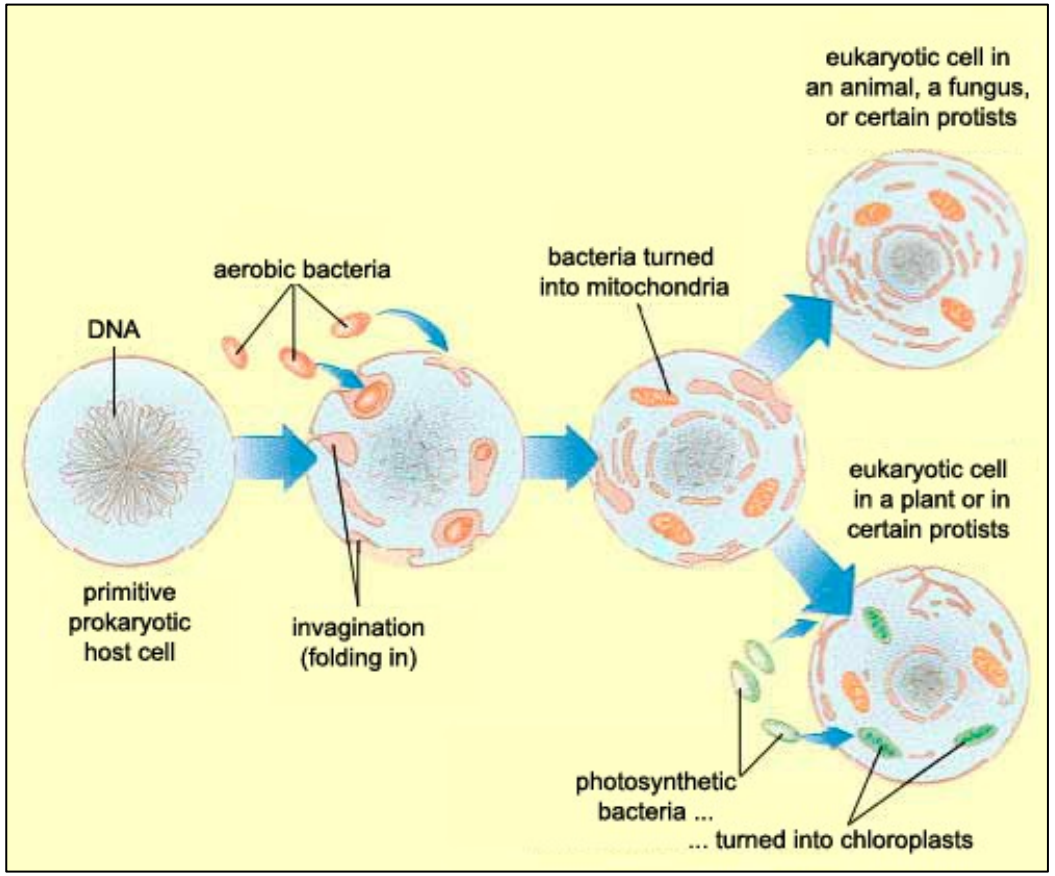
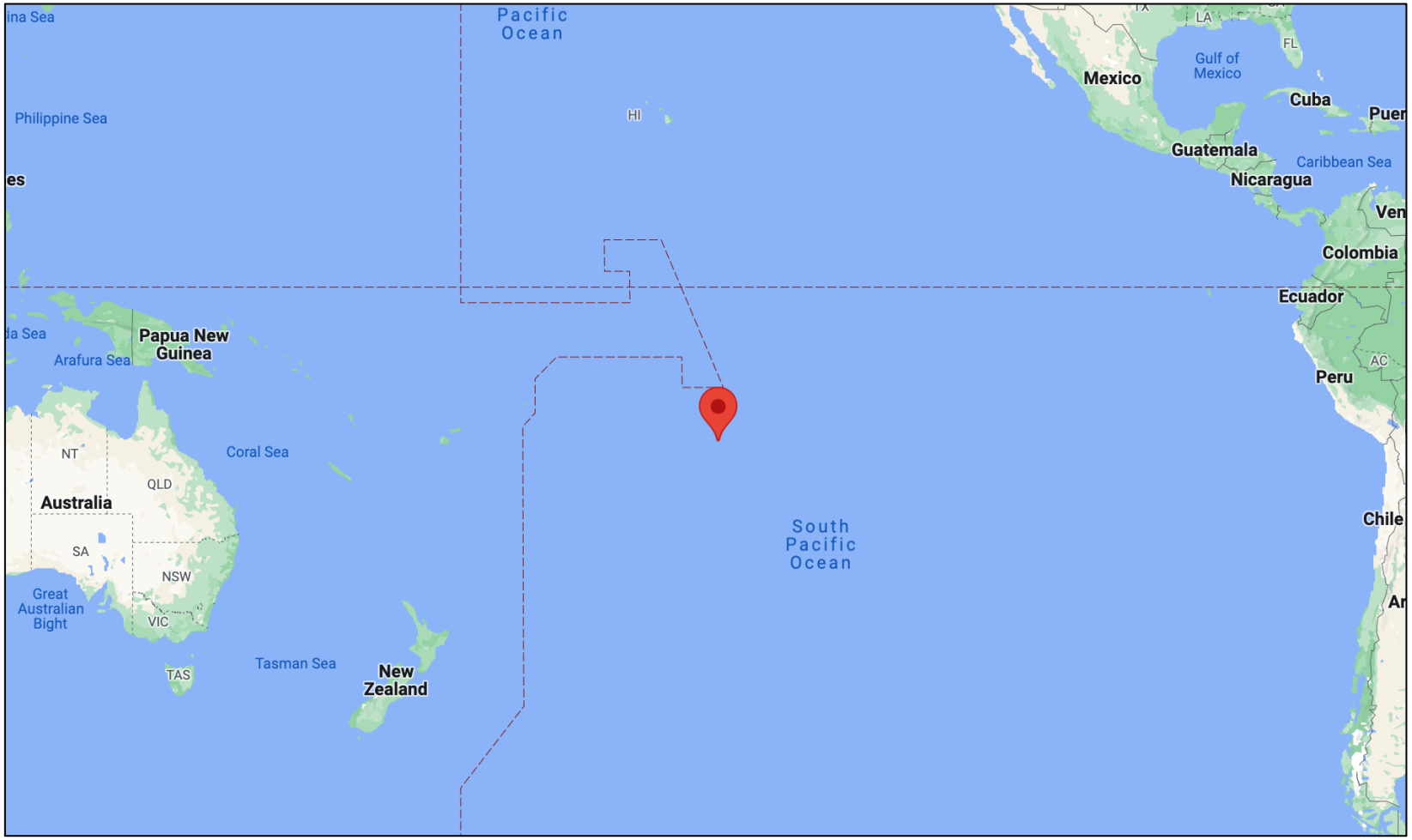
Barcoding

March 14, 2023

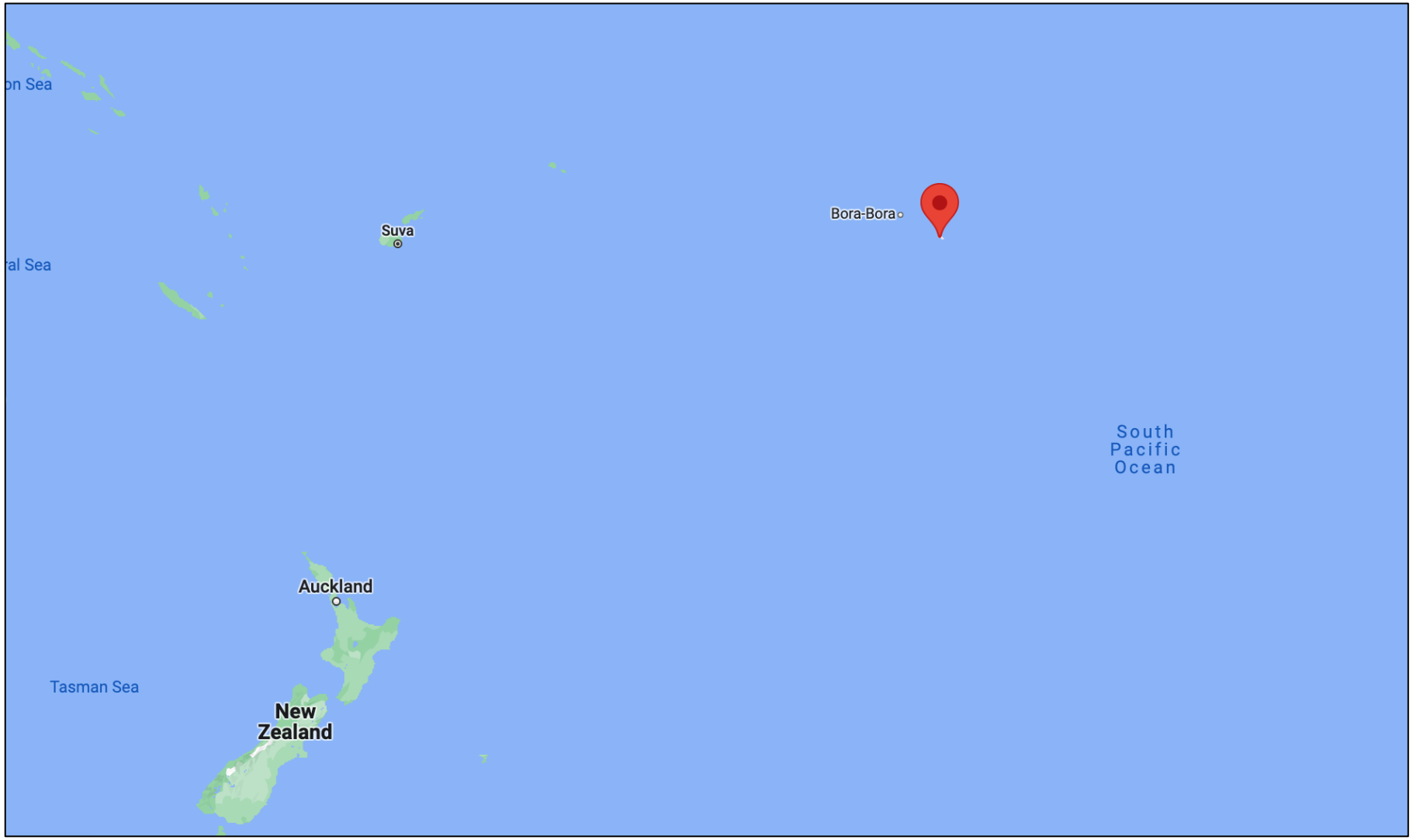




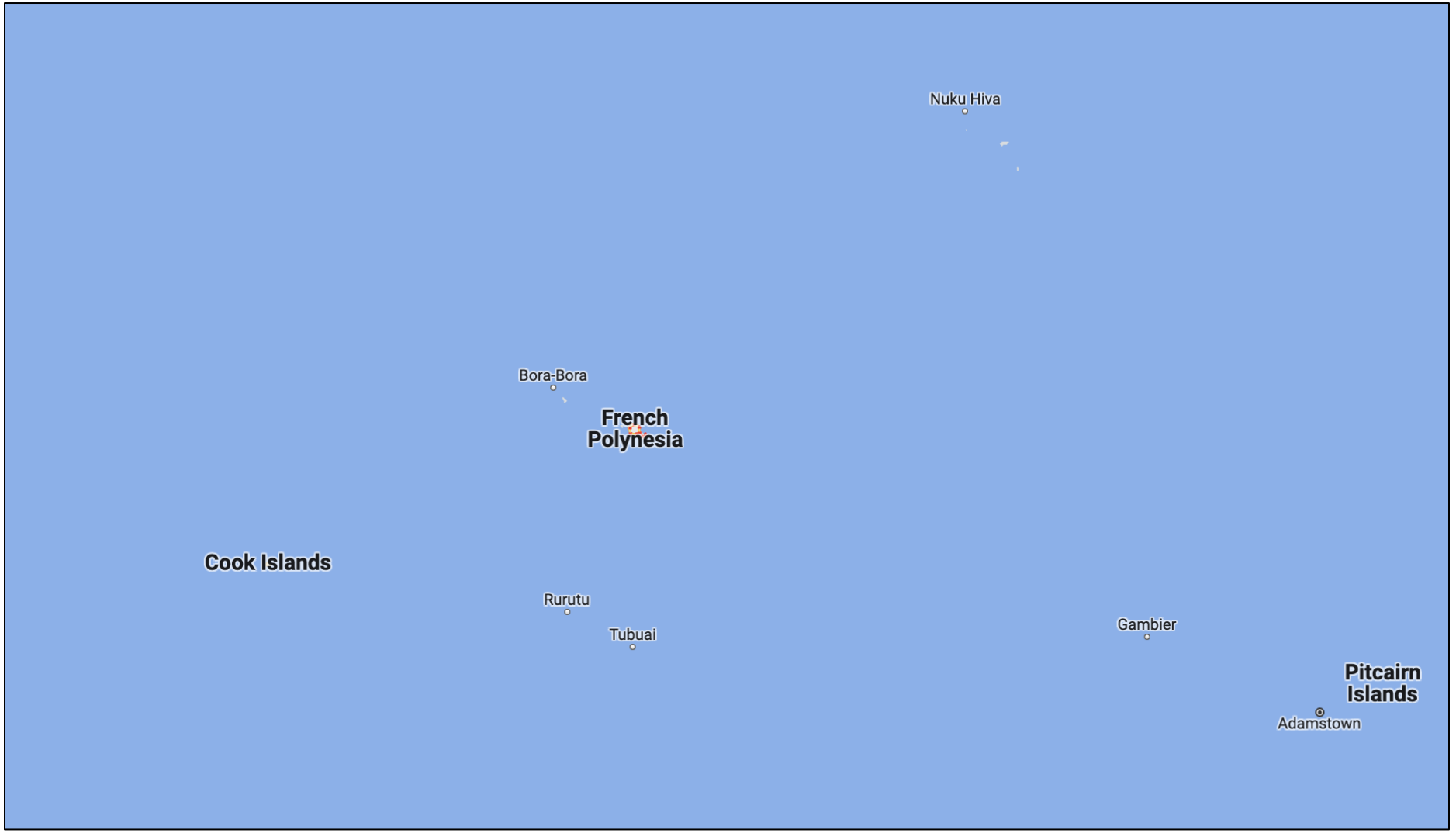
The Moorea Biocode Project



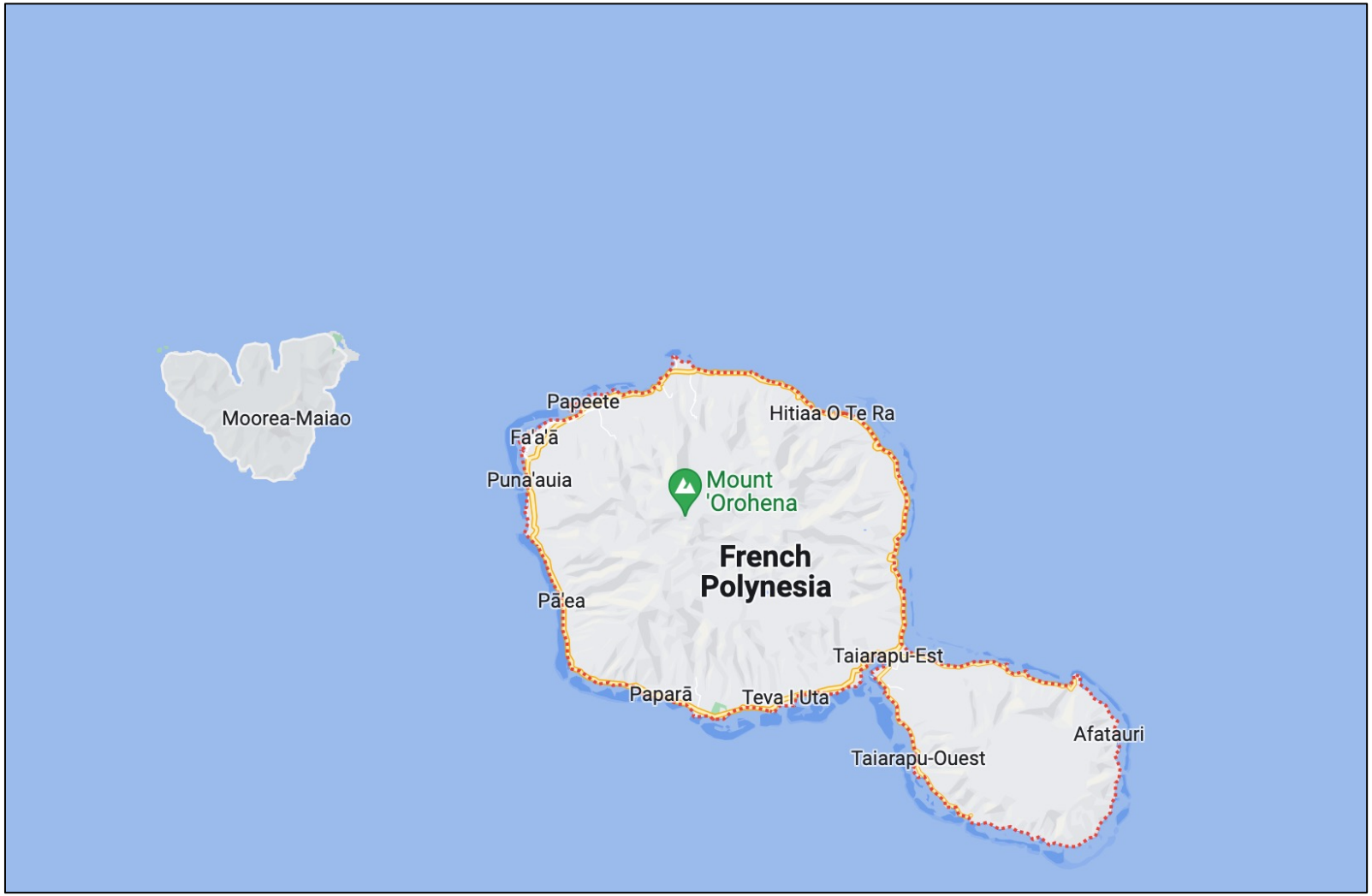
The Moorea Biocode Project



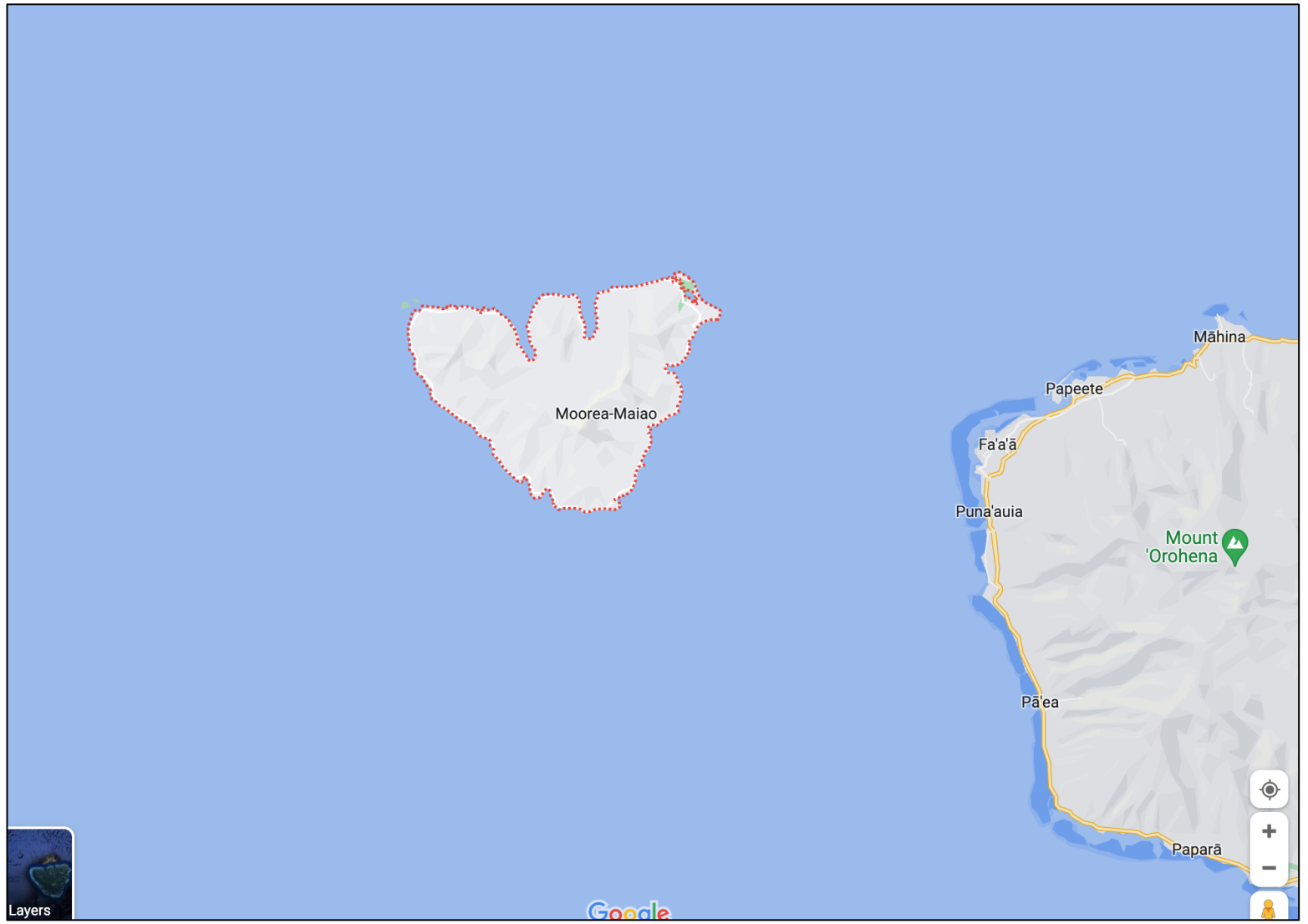
The Moorea Biocode Project



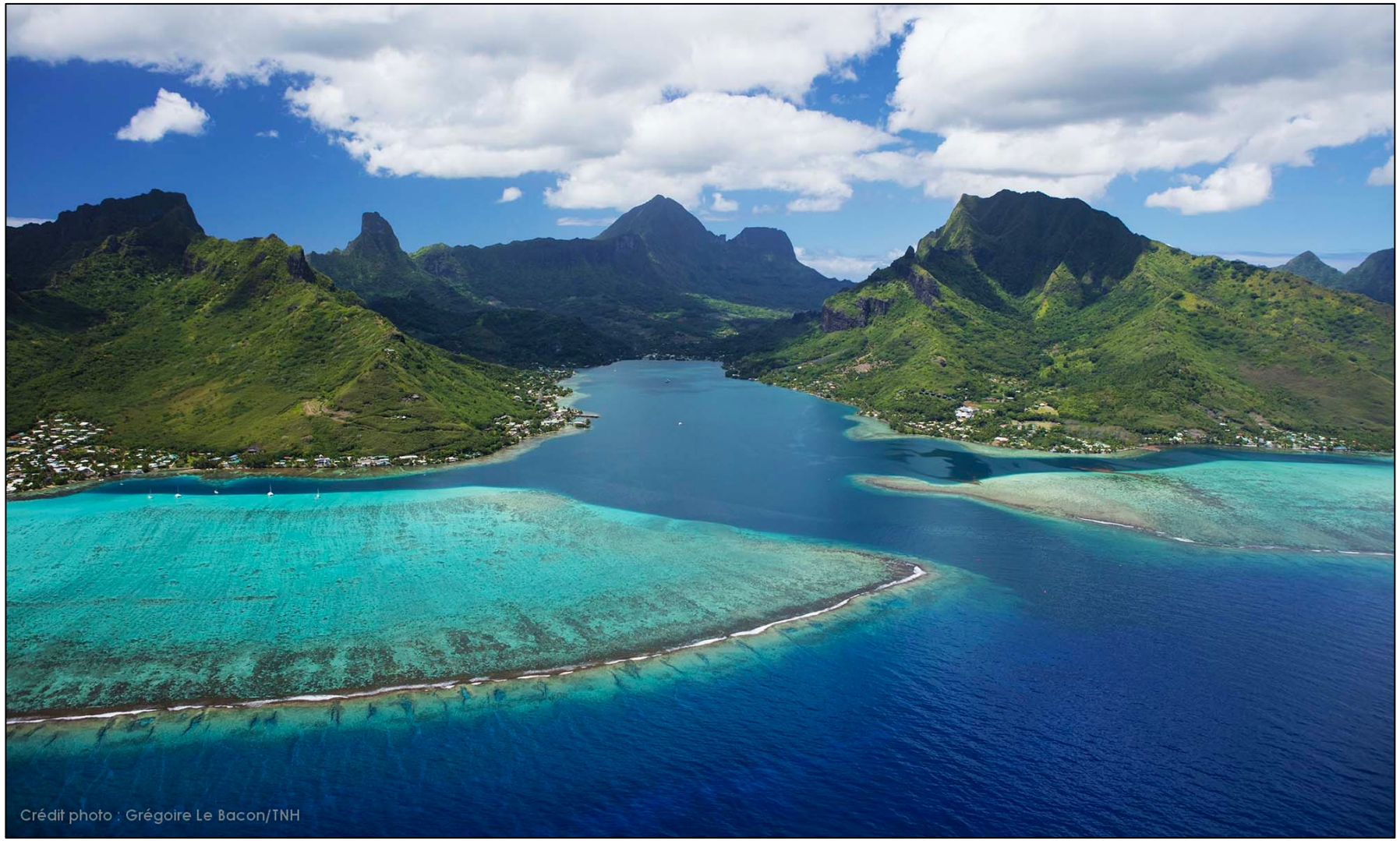
The Moorea Biocode Project



The Moorea Biocode Project



The Moorea Biocode Project



https://ocean.si.edu/ecosystems/coral-reefs/welcome-moorea-biocode-project

DNA Barcoding

- Analogy to product barcodes
- Scanning a barcode sticker is easier than scanning an apple
- A database of barcodes provides lookup table for easy identification
- 2 de-facto standard genes: 16S/18S rRNA and COI

Requirements for a barcode gene

1. Universal
2. Easy to extract and sequence
3. Unique in each species
4. Bonus points: a *barcode gap*

The 2 most common barcode genes

- 16S / 18S ribosomal RNA
  - Better for single-cell and very small protist metagenomics
- COI (“See-Oh-Won”)

Advantages of COI over rRNA (other than in metagenomic studies)

- Shorter than rRNA
  - Fewer chances for sequencing errors, easier bioinformatic analysis
- More reliable amplification than rRNA
  - Especially of degraded samples, e.g. museum specimens
- More variable than rRNA
  - Less possibility of 2 species with same barcode

COI: The de-facto standard barcoding gene for animal life

- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial ➔ no introns ➔ easy to align
  - Indel mutations are rare ➔ easy to align
  - High substitution rate in 3rd codon position ➔ lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

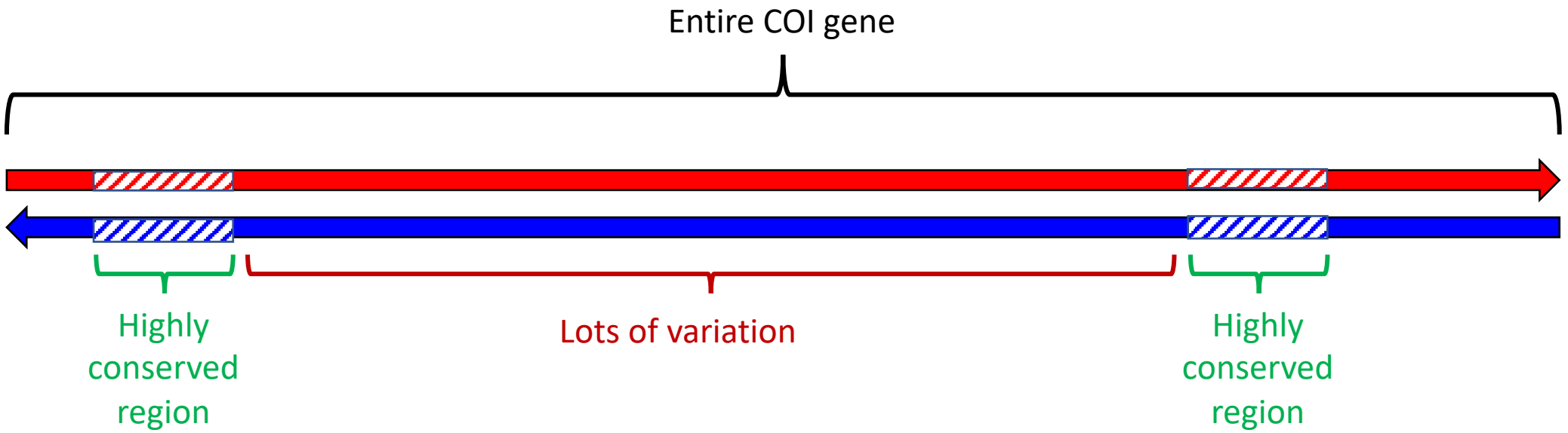
COI: The de-facto standard barcoding gene for animal life

- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial ➔ no introns ➔ easy to align
  - Indel mutations are rare ➔ easy to align
  - High substitution rate in 3rd codon position ➔ lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

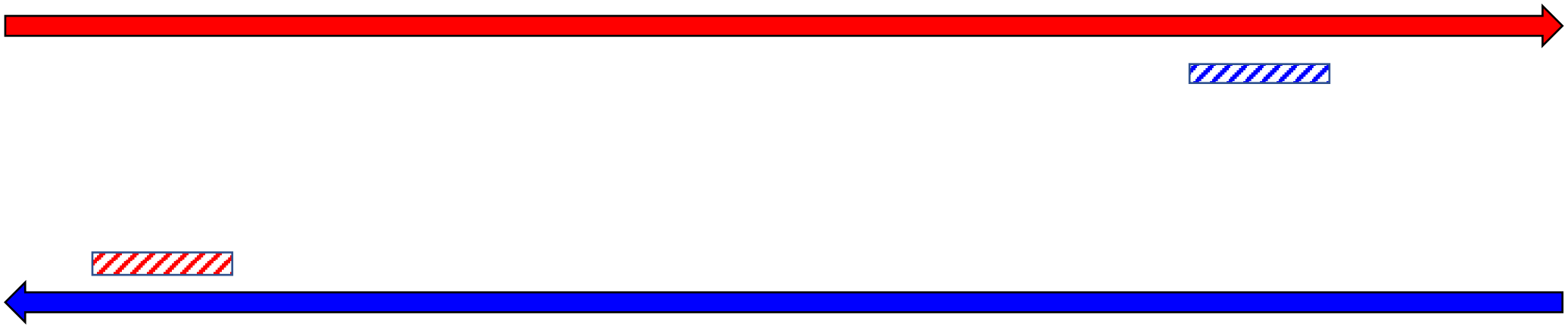
COI: The de-facto standard barcoding gene for animal life

- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial → no introns → easy to align
  - Indel mutations are rare → easy to align
  - High substitution rate in 3rd codon position → lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

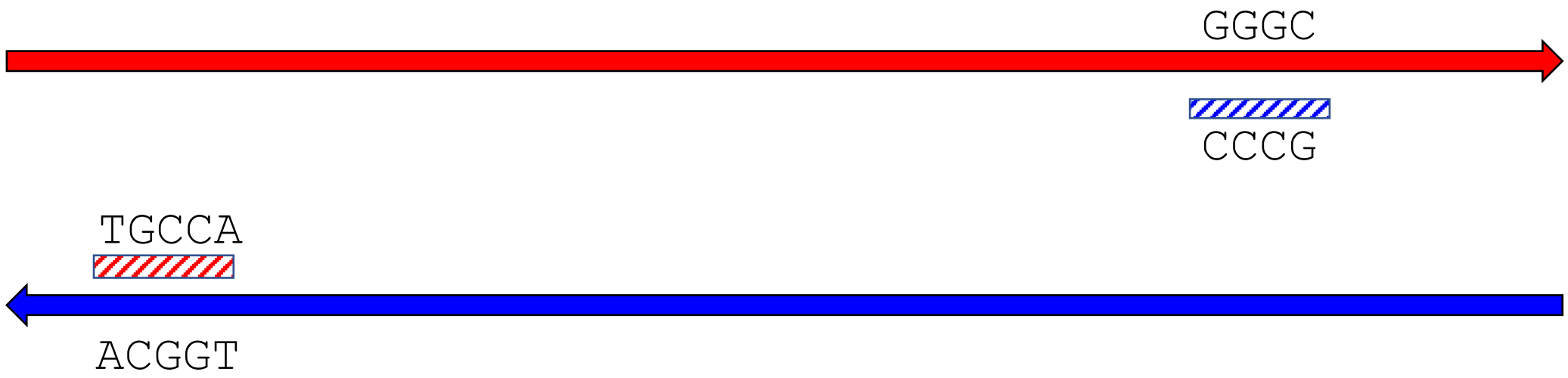
There are good PCR primers for COI



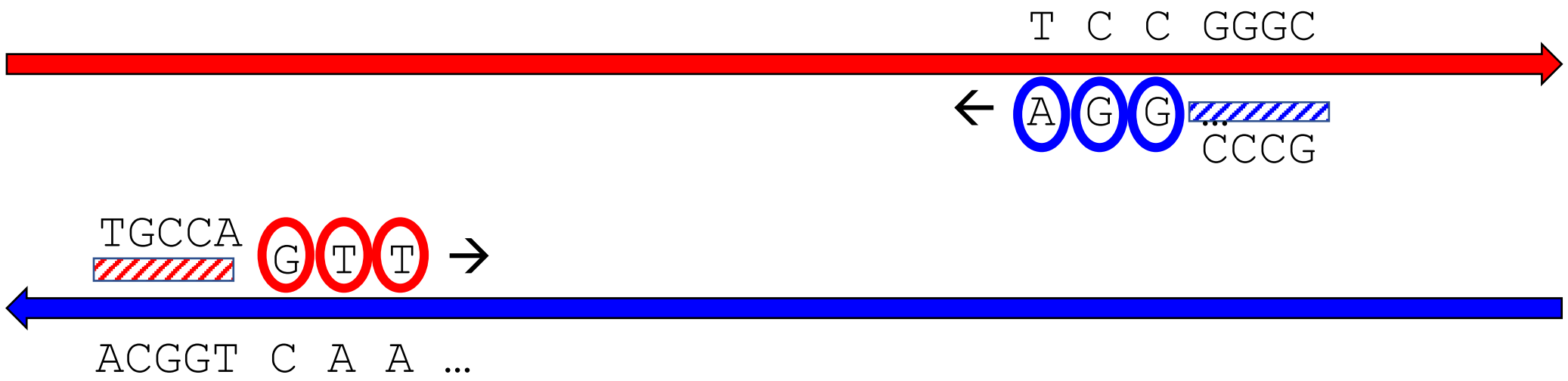
- Manufacture short strands (“primers”) that match the conserved regions.
- Separate the two strands of DNA
- Primers will attach to their reverse complements



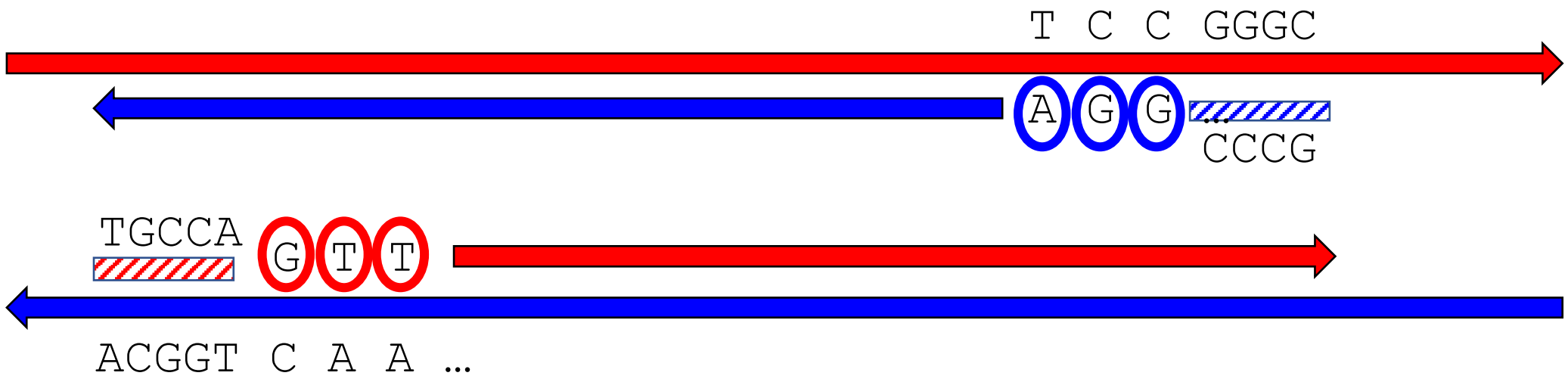
- Manufacture short strands (“primers”) that match the conserved regions.
- Separate the two strands of DNA
- Primers will attach to their reverse complements
- With help from *polymerase*, primers will extend by recruiting free-floating individual nucleotides



- Manufacture short strands (“primers”) that match the conserved regions.
- Separate the two strands of DNA
- Primers will attach to their reverse complements
- With help from *polymerase*, primers will extend by recruiting free-floating individual nucleotides



- Manufacture short strands (“primers”) that match the conserved regions.
- Separate the two strands of DNA
- Primers will attach to their reverse complements
- With help from *polymerase*, primers will extend by recruiting free-floating individual nucleotides

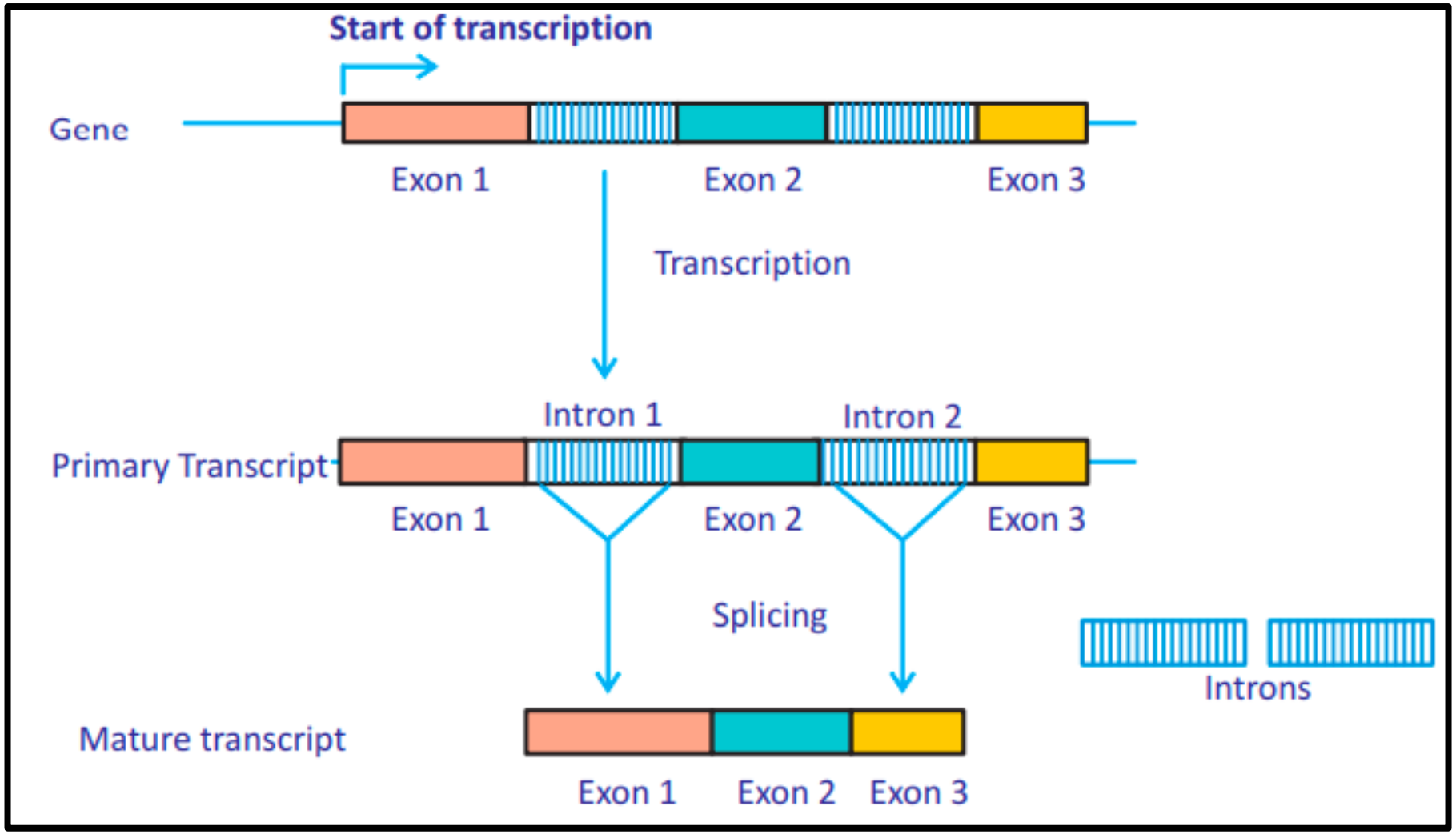


- Eventually each strand is bound to its own reverse-complement
- So original strand between the primers has been duplicated
- Do it again ➔ 2 strands become 4
- Do it again ➔ 4 become 8

COI: The de-facto standard barcoding gene for animal life

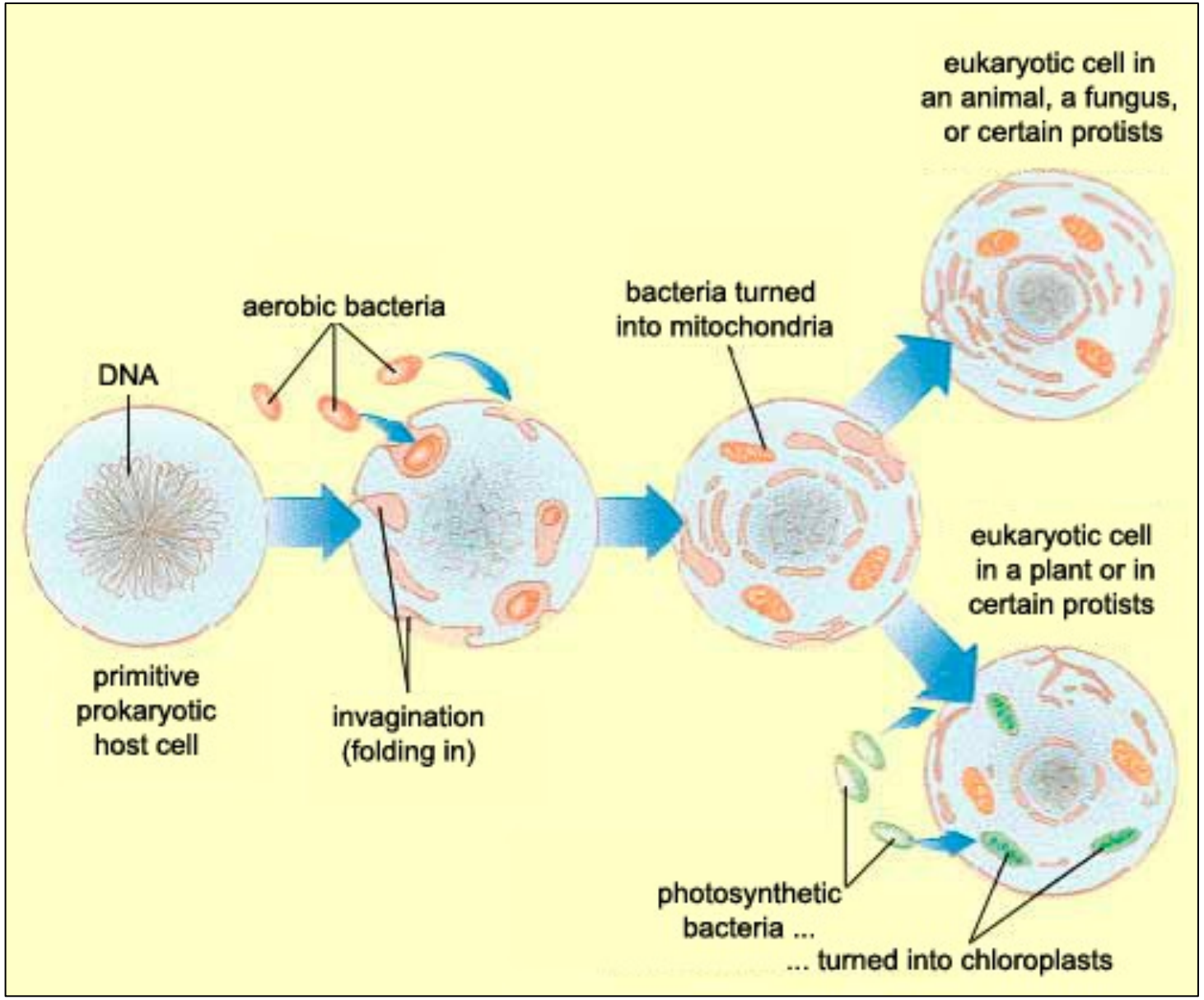
- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial → no introns → easy to align
  - Indel mutations are rare → easy to align
  - High substitution rate in 3rd codon position → lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

Eukaryote DNA has *introns* and *exons.* These are spliced out of the mRNA transcript before translation to protein



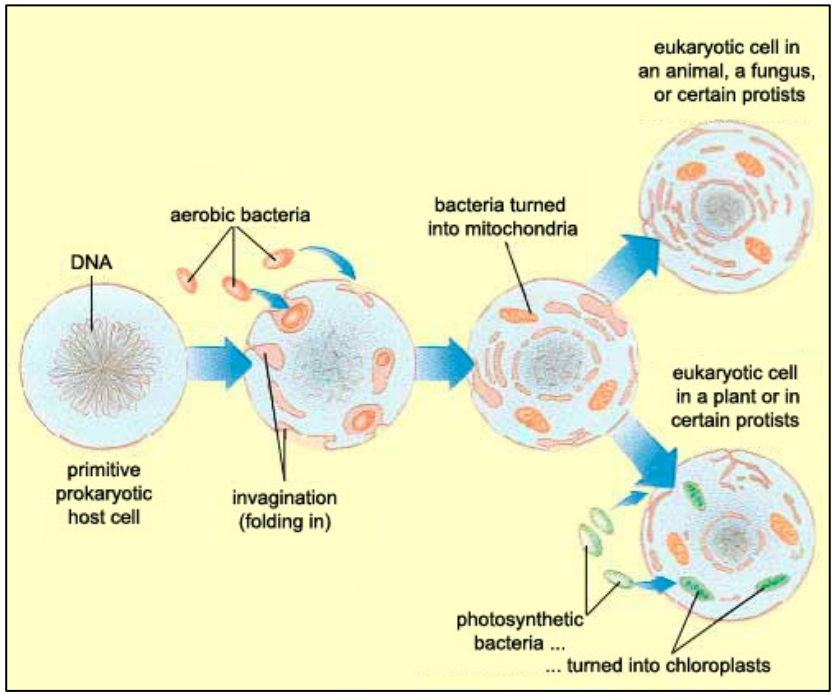
COI is mitochondrial

We got it through endosymbiosis



Endosymbiosis

- Much later: host organism has evolved to become a eukaryote, with introns/exons in its DNA.
- The aerobic bacterium has become a mitochondrion. Its own DNA never acquires introns/exons.
- When prokaryote cells divide, mitochondria reproduce so each daughter cell gets some.
- Sperm cells don't donate their mitochondria to the zygote → all our mtDNA descends from egg cell's mitochondria



So we animals really have 2 genomes

- The *nuclear* genome
  - In the cell nucleus
  - Half from each parent
  - Introns, exons
- The *mitochondrial* genome
  - In each mitochondrion
  - All from mother, none from father
  - No introns/exons

So we animals really have 2 genomes

The *nuclear* genome

- In the cell nucleus
- Half from each parent
- Introns, exons

The *mitochondrial* genome

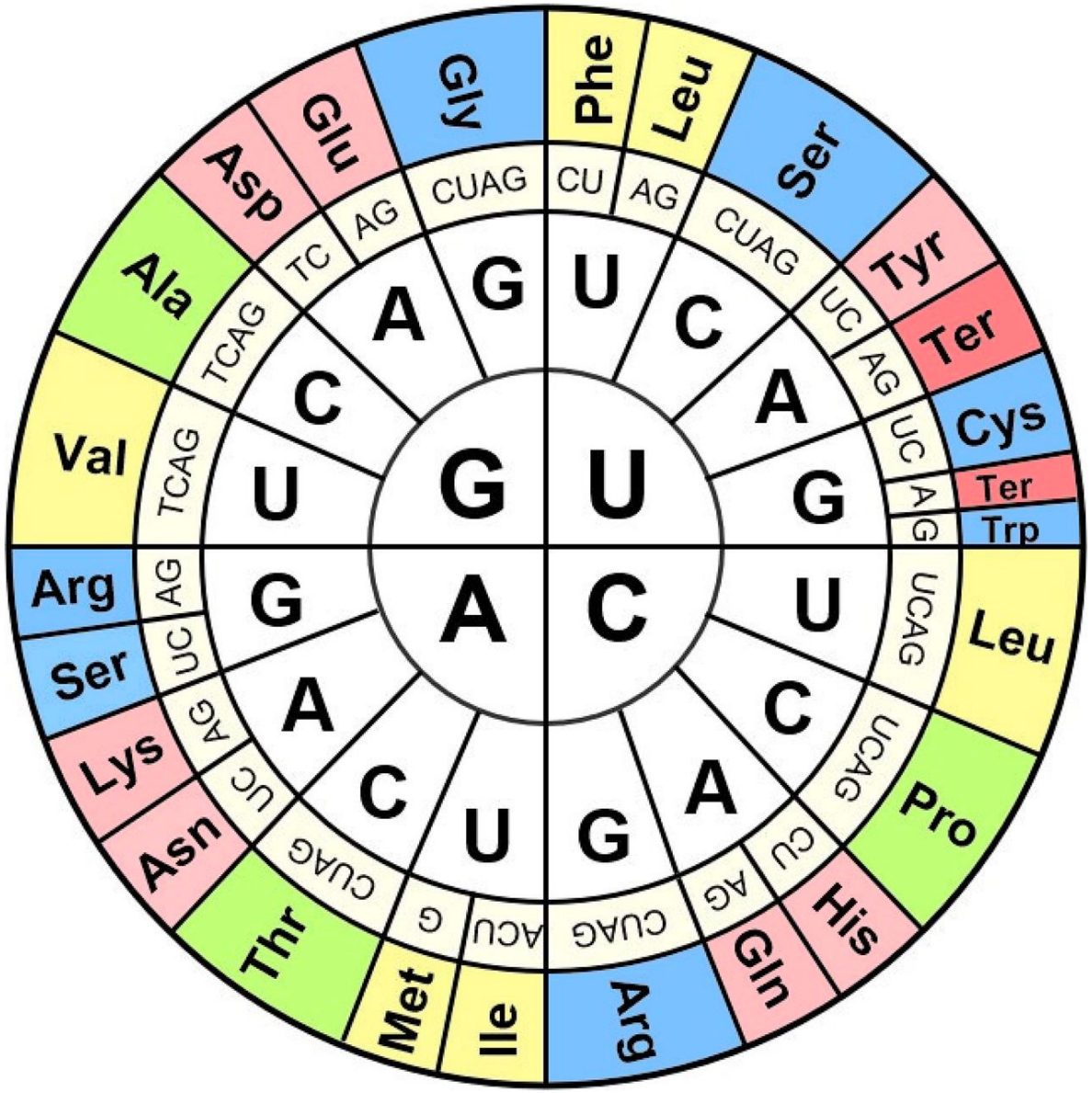
- In each mitochondrion
- All from mother, none from father
- No introns/exons

Introns are long, highly variable regions that mess with alignment algorithms.
➔A barcode gene without introns is an advantage.
COI gene is mitochondrial.

COI: The de-facto standard barcoding gene for animal life

- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial → no introns → easy to align
  - Indel mutations are rare → easy to align
  - High substitution rate in 3rd codon position → lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

High substitution rate in the 3rd codon position



- G-G-Anything ➔ Glycine
- U-C-Anything ➔ Serine
- G-U-Anything ➔ Valine
- Etc
- Substitutions in the 3rd codon position often have no effect on the resulting amino acid (<u>silent mutations</u>)
- Think of this in terms of evolution

Evolution: 3rd vs 2nd codon position

- 3rd position mutation
  - If silent, produces same amino acid as wild-type.
  - Therefore same protein as wild-type.
  - No competitive advantage or disadvantage.
  - Eventually might be seen in significant fraction of population ➔ diversity.
- 2nd position mutation
  - Most mutations here are missense mutations.
  - A different amino acid.
  - Which might have different chemistry from wild-type.
  - Which results in different protein shape.
  - Which results in different protein function.
  - Often fatal to the cell ➔ mutation is never passed on.

Are there silent mutations in the 2nd codon position?

PH

are there silent mutations in the 2nd nucleotide position of a codon

Yes, there are some instances where mutations at the second nucleotide position of a codon can be silent, meaning they do not result in a change to the amino acid sequence of the protein.

This is because the genetic code is redundant, with multiple codons often encoding the same amino acid. For example, the amino acid alanine can be encoded by the codons GCU, GCC, GCA, or GCG. If a mutation occurs in the second nucleotide of a codon that changes it to a different codon that still encodes for the same amino acid, then the resulting change will be silent.

PH example of silent mutation in the 2nd nucleotide position of a codon

An example of a silent mutation in the second nucleotide position of a codon is a change from CTC to CTT in the gene that codes for the amino acid leucine. Both CTC and CTT code for leucine, so a mutation that changes the second nucleotide from C to T will not result in a change to the amino acid sequence of the protein.

PH

still wrong position

I apologize for my previous mistakes. Here is another example of a silent mutation in the 2nd position of a codon:

Original DNA sequence: ATG CTC CGA TCA GAA TTA
Original mRNA sequence (after transcription): AUG CUC CGA UCA GAA UUA
Original amino acid sequence (after translation): Met Leu Arg Ser Glu Leu

Let's say a silent mutation occurs in the 2nd nucleotide position of the fifth codon, changing the T to a C:

Mutated DNA sequence: ATG CTC CGA TCA GAC TTA

GAA → GAC

Via T→C in 3rd position

????

PH

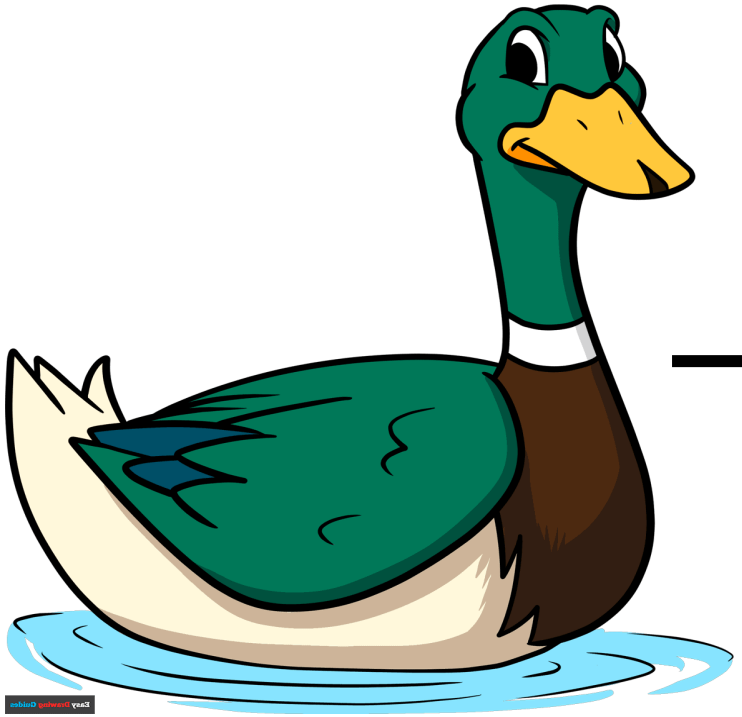
it is disappointing that you can't count to 2

I apologize for my previous mistakes. Thank you for bringing it to my attention. I strive to provide accurate information, and I will ensure to be more careful and accurate in the future.

COI: The de-facto standard barcoding gene for animal life

- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial → no introns → easy to align
  - Indel mutations are rare → easy to align
  - High substitution rate in 3rd codon position → lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

Vouchered sequence databases



tissue sample

Sequencing

> Duck
ACGGCTTACAT…

Database

“voucher”

-80C freezer

If identification or sequence are challenged, retrieve the voucher and sequence again.

COI: The de-facto standard barcoding gene for animal life

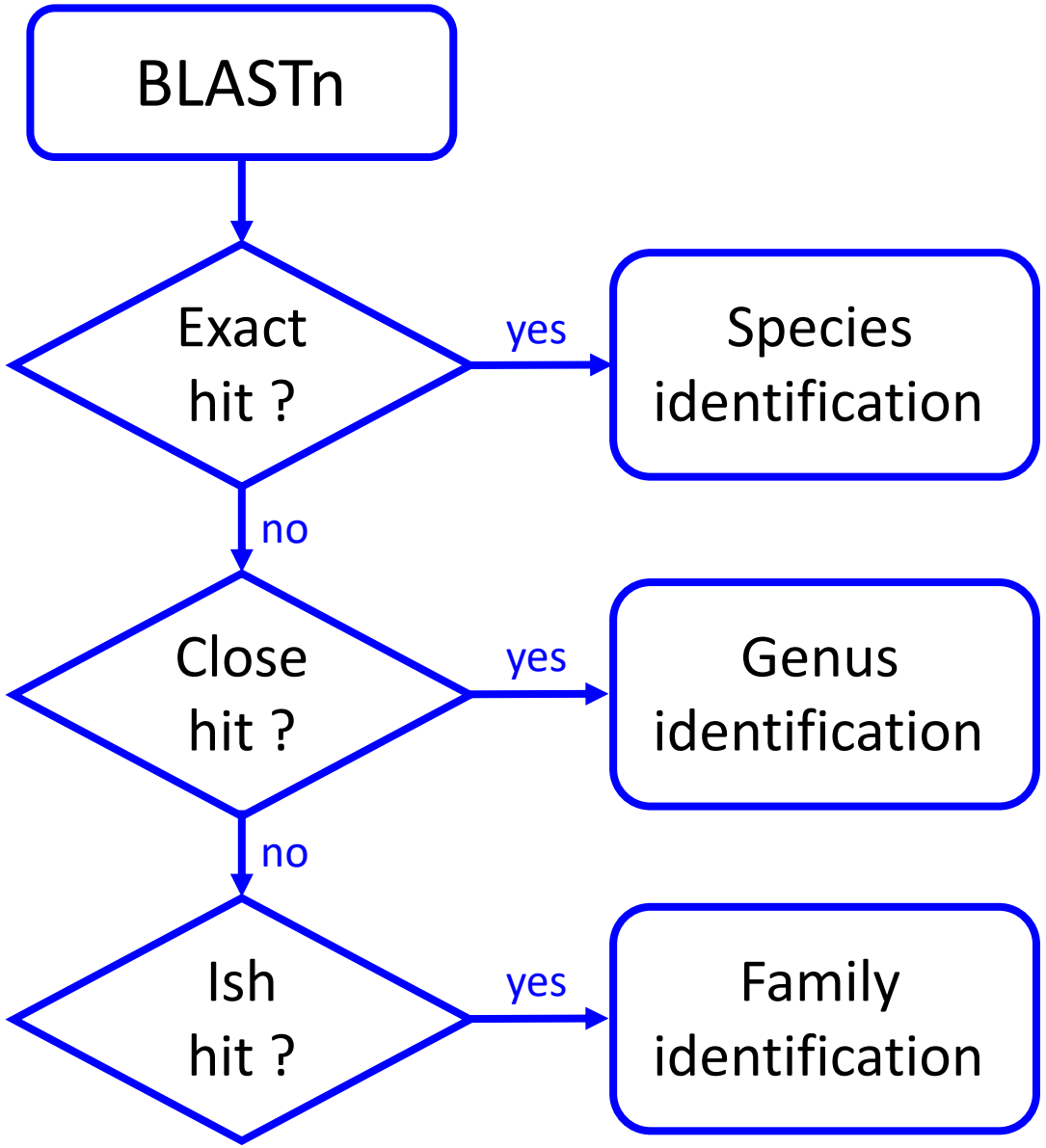
- 2002: Hebert &al propose COI as barcoding standard
  - All animals have it
  - Good primers
  - Mitochondrial ➔ no introns ➔ easy to align
  - Indel mutations are rare ➔ easy to align
  - High substitution rate in 3rd codon position ➔ lots of diversity
- BOLD database
  - "Barcode Of Life Database"
  - Vouchered, expensive, highly COI-specific
- CO-ARBitrator algorithm & database: software curated from GenBank, equally specific, more sensitive

The BOLD database is highly COI specific

- Specific: (nearly) everything in the database really is COI.
- Not very sensitive: there are lots of COI sequences in the world that aren't in BOLD.
- This isn't surprising, and isn't a problem.
- BOLD's stringent submission requirements, including voucher requirement, make submission slow and expensive.

The trouble with COI: the “barcode gap”

- A reasonable expectation: given any 2 samples, COI nucleotide sequence similarity reflects species similarity
- Therefore if you find a novel species, blasting its COI against BOLD should tell you the right genus
- Et cetera



The reasonable expectation requires a “Barcode gap”

- Given any clade, every member should be more similar to every other member, than it is to every non-member
- d(x, y) means evolutionary “distance” between x and y, e.g. # of mutations





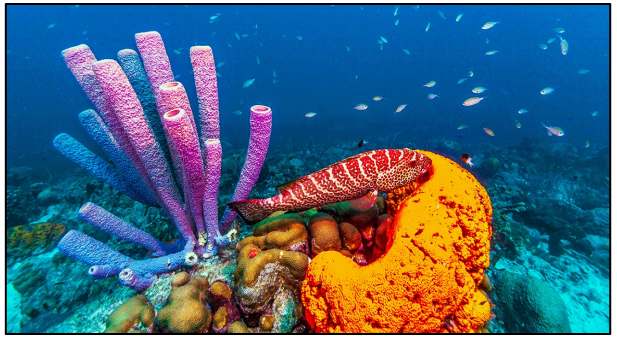

d(Panther, Panda)
should be >
d(Panther, Lion)

Genus *Ursus*

Genus *Pantera*

But the barcode gap is not universal

- Sep 30, 2022: downloaded all BOLD metazoan COI sequences.
- Converted to blastable.
- BLASTed every sequence against this database, ignoring hits to species of query.
- Simulates considering each species as novel, therefore not yet in BOLD.
- Best hit not always to different species in correct genus.
- Within phylum *Porifera*, 38% of species identified with wrong genus.
- Within phylum *Nematoda*, 39% of species identified with wrong genus.



Dos and Don’ts of identifying a mystery sequence using COI barcoding

- BLAST the mystery against a COI database
- *Do* trust any exact hit
- *Don’t* use a close but inexact hit to infer genus