# Marine Ecological Genetics

## 12. DNA barcoding | Computer practical

- Extract DNA barcodes from Sanger reads
- Identify samples with reference databases (BOLD, GenBank)
- Evaluate genetic distances and accuracy of identification

Martin Helmkampf

# Access files for practical

Download files from GitHub:

https://github.com/mhelmkampf/meg24.git (Code | Download ZIP)

Alternatively, run git from terminal:

```
git clone https://github.com/mhelmkampf/meg24.git
```

Open script called **MarEcolGen_barcoding.sh** in a text editor

# #fischdetektive

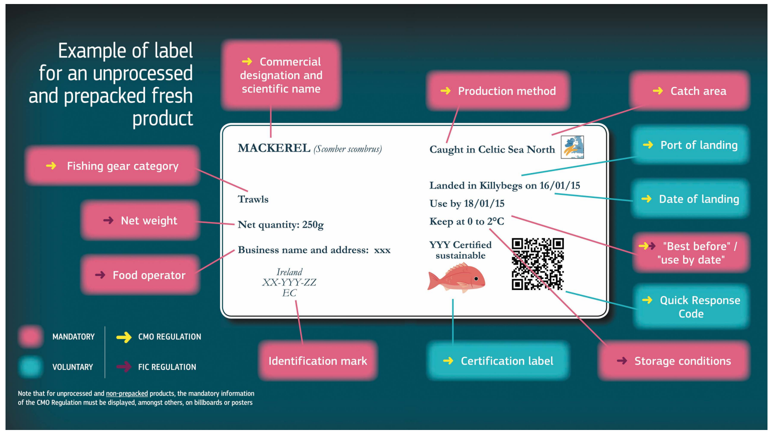
Citizen science project

at GEOMAR (2017) with over

700 participants (10–14 years)

Where does our seafood come

from, and is it labeled correctly?



Note that for unprocessed and non-prepacked products, the mandatory information of the CMO Regulation must be displayed, amongst others, on billboards or posters

Thorsten Reusch, GEOMAR

Sample collection

↓

DNA extraction

↓

PCR (COI)

↓

Sequencing

←

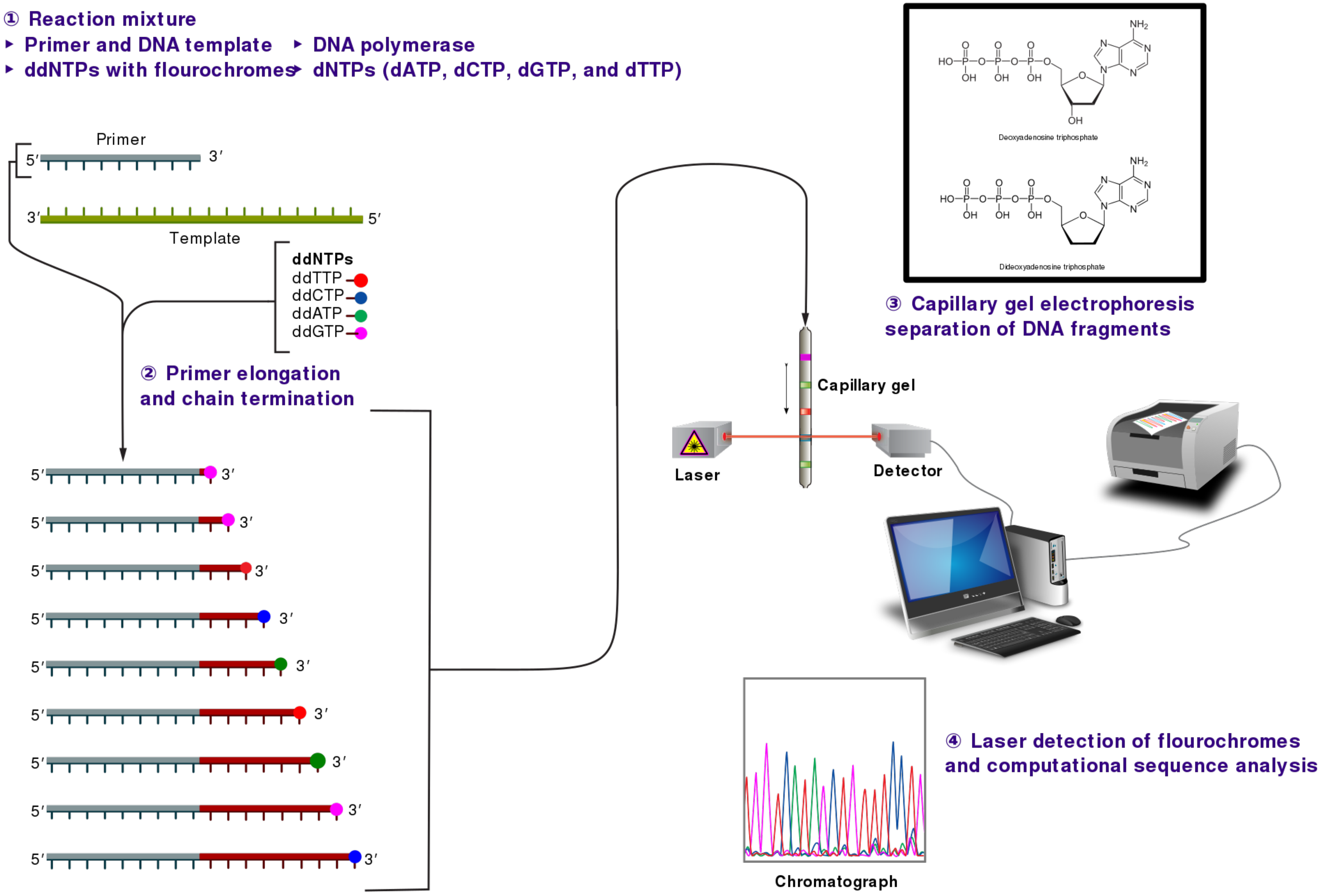
ID

# COI barcode

Approx. 650 bp in 5’ region of cytochrome c oxidase subunit I (COI)

```
>MN604318.1 Oncorhynchus keta cytochrome c oxidase subunit I gene, complete cds; mitochondrial
GTGGCAATCACACGATGATTCTTCTCAACCAACCACAAAGACATTGGCACCCTCTATTTAGTATTTGGTGCCTGAGCCGGGATAGTAGGCACCGCCCTG
AGCCTACTAATTCGGGCAGAACTAAGCCAGCCAGGCGCTCTTCTAGGGGATGACCAGATCTACAATGTAATCGTTACAGCCCATGCCTTCGTTATAATT
TTCTTTATAGTCATACCAATTATAATCGGAGGCTTTGGAAACTGATTAATCCCCCTAATGATCGGGGCACCAGATATAGCATTCCCACGAATAAATAAC
ATAAGCTTCTGACTCCTACCTCCGTCCTTCCTCCTCCTCCTTTCTTCATCTGGAGTTGAAGCCGGCGCTGGTACCGGGTGGACAGTTTATCCCCCTCTA
GCCGGAAACCTTGCCCACGCAGGAGCATCTGTCGACTTAACCATCTTCTCCCTCCATTTAGCTGGAATCTCCTCAATTTTGGGGGCCATTAATTTTATT
ACGACCATTATCAACATAAAACCCCCAGCTATTTCTCAGTACCAAACCCCGCTTTTTGTCTGAGCTGTACTAATCACTGCTGTACTTCTACTATTATCA
CTCCCCGTTCTGGCAGCAGGTATTACTATGTTGCTCACAGATCGAAATTTAAACACCACTTTCTTTGACCCGGCGGGTGGCGGAGATCCAATTTTATAC
CAACACCTCTTTTGATTCTTCGGTCACCCAGAGGTCTATATTCTGATCCTCCCAGGCTTTGGTATAATTTCACATATCGTTGCATATTACTCTGGTAAG
AAAGAACCTTTCGGGTACATAGGAATAGTGTGAGCTATAATAGCCATCGGCTTGTTAGGATTTATCGTTTGAGCCCACCACATATTTACTGTCGGGATG
GACGTGGACACTCGTGCCTACTTTACATCTGCCACCATAATTATCGCTATCCCCACAGGAGTAAAAGTATTTAGCTGACTAGCTACACTGCACGGAGGC
TCGATCAAATGAGAGACACCACTTCTCTGAGCCCTAGGATTTATCTTCCTATTTACAGTGGGCGGATTAACGGGCATCGTCCTTGCTAACTCCTCATTA
GACATTGTTTTACATGACACTTATTACGTAGTCGCCCATTTCCACTACGTACTCTCAATAGGAGCTGTATTTGCCATTATGGGCGCTTTCGTACACTGA
TTCCCCCTATTCACAGGGTACACCCTTCACAGCACATGAACCAAAATCCATTTTGGAATTATATTTATCGGTGTAAATTTAACCTTTTTCCCACAGCAT
TTCCTAGGCCTCGCAGGGATACCACGACGGTACTCTGACTACCCGGACGCCTACACGCTATGAAACACTGTATCCTCAATCGGATCCCTTGTCTCCTTA
GTAGCTGTAATTATGTTCCTATTTATTCTTTGAGAGGCTTTTGCTGCCAAACGAGAAGTAGCATCAATCGAAATAACTTCAACAAACGTAGAATGACTA
CACGGATGCCCCCCACCCTACCACACATTCGAGGAACCAGCATTTGTCCAAGTACGAACGTACTAA
```

# Sanger sequencing



① Reaction mixture

- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

② Primer elongation and chain termination

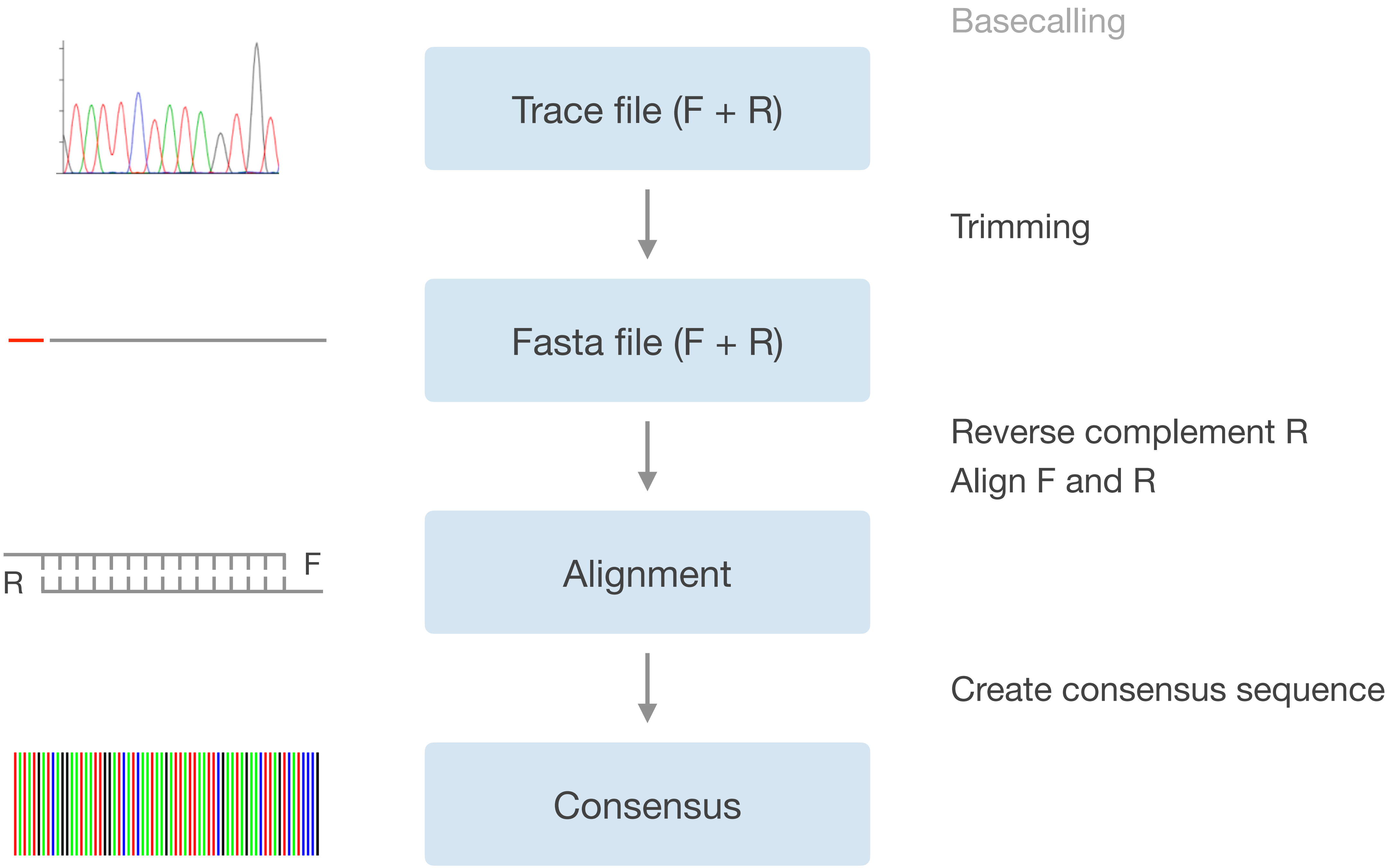
③ Capillary gel electrophoresis separation of DNA fragments

④ Laser detection of flourochromes and computational sequence analysis

Estevezj, CC BY-SA 3.0

# Sanger read processing

Exercise 1



Basecalling

Trace file (F + R)

Trimming

Fasta file (F + R)

Reverse complement R
Align F and R

Alignment

Create consensus sequence

Consensus

# Sequence alignment

Exercise 1

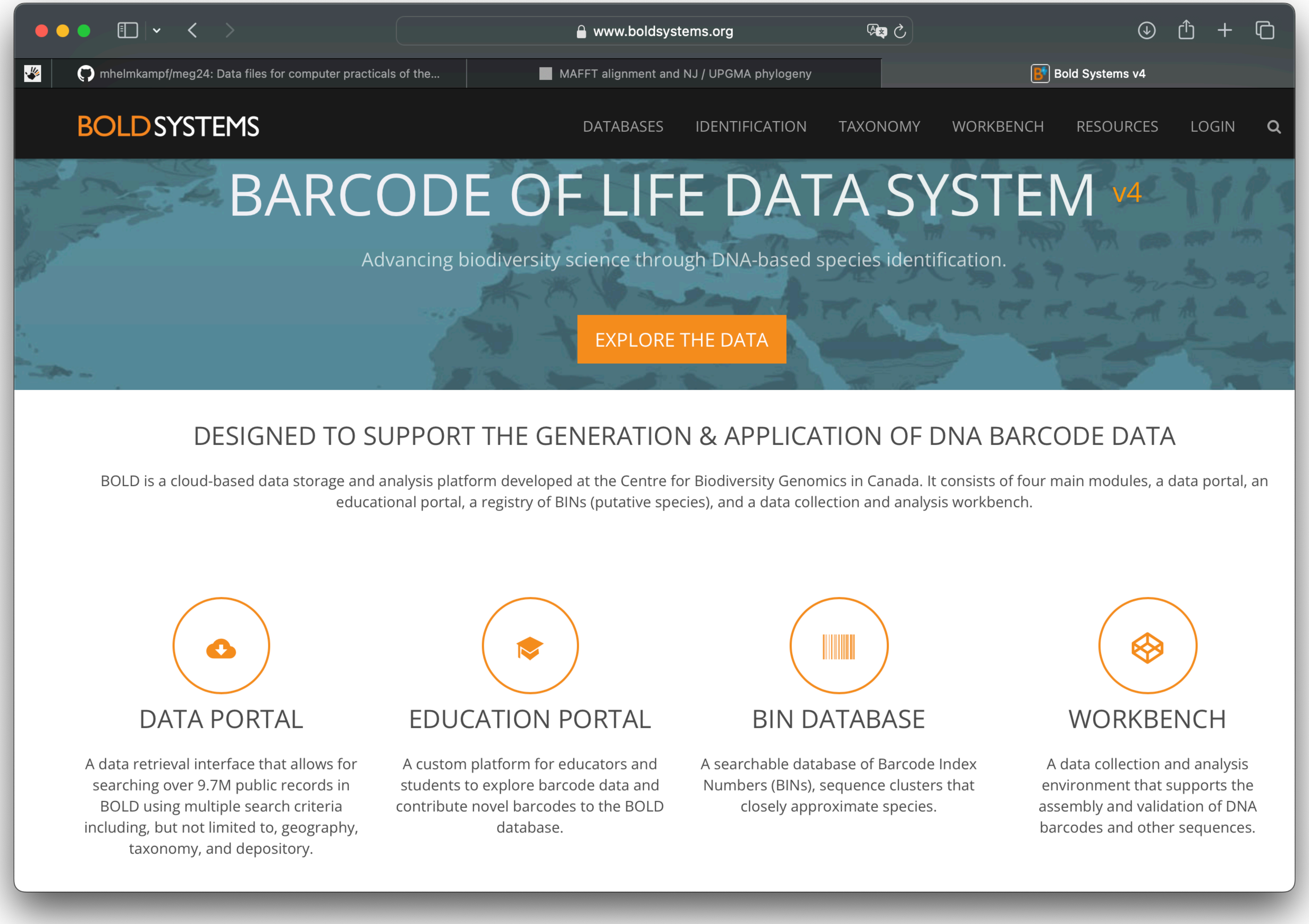
```
G A T G T T C G A A
G A T C – – – G A A
G A C C – T C G – T
```

Arranges nucleotide or amino acid sequences so that the number of mismatches and gaps are minimized

- Multiple sequence alignments can be constructed progressively from pairwise alignments
- Computationally complex, often requires heuristic solutions
- Key to identify evolutionary relationships between sequences (e.g. homology)

# BOLD

Exercise 2



BARCODE OF LIFE DATA SYSTEM v4

Advancing biodiversity science through DNA-based species identification.

DESIGNED TO SUPPORT THE GENERATION & APPLICATION OF DNA BARCODE DATA

BOLD is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada. It consists of four main modules, a data portal, an educational portal, a registry of BINs (putative species), and a data collection and analysis workbench.

DATA PORTAL

A data retrieval interface that allows for searching over 9.7M public records in BOLD using multiple search criteria including, but not limited to, geography, taxonomy, and depository.

EDUCATION PORTAL

A custom platform for educators and students to explore barcode data and contribute novel barcodes to the BOLD database.

BIN DATABASE

A searchable database of Barcode Index Numbers (BINs), sequence clusters that closely approximate species.

WORKBENCH

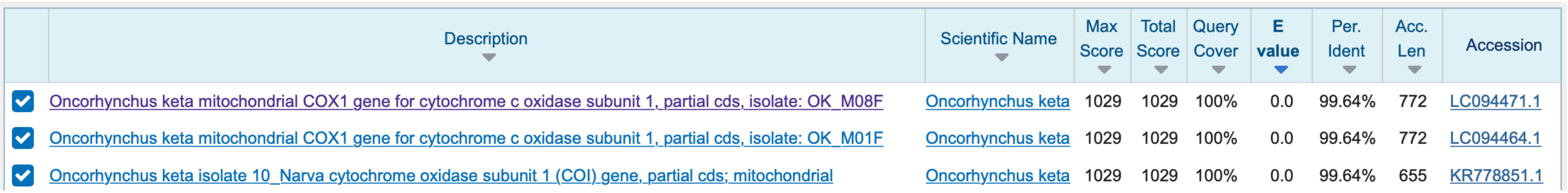
A data collection and analysis environment that supports the assembly and validation of DNA barcodes and other sequences.

# Matching sequences to a database with BLAST

Exercise 2

Algorithm overview

- Split query into very short segments (*k*-mers or words)
- Find exact matches between words and sequences in database (seeds)
- Extend matches to local alignments (HSP; stops once too many mismatches occur)
- Evaluate statistical significance of each HSP (e-value)



| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M08F | Oncorhynchus keta | 1029 | 1029 | 100% | 0.0 | 99.64% | 772 | LC094471.1 |
| ☑ | Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M01F | Oncorhynchus keta | 1029 | 1029 | 100% | 0.0 | 99.64% | 772 | LC094464.1 |
| ☑ | Oncorhynchus keta isolate 10_Narva cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial | Oncorhynchus keta | 1029 | 1029 | 100% | 0.0 | 99.64% | 655 | KR778851.1 |

…

# Genetic distances

Exercise 2

```
AAGCCAGCCAGGCGCTCTTCTAGGGGATGACCAGATCTACAATGTAATCG   # 50 positions total
AAGTCAACCTGGTGCACTTCTTGGTGATGATCAAATTTATAATGTGATCG
***.**.** **.** ***** ** ** **.**.**.**.*****.****   # 13 differences
```

Uncorrected distance

*p* = 13 / 50 = 0.26

K2P distance (Kimura 1980)

*K* = 0.33

$$K = -\frac{1}{2}\ln((1 - 2p - q)\sqrt{1 - 2q})$$

*p*: proportion of transitions (A↔G, C↔T)

8 / 50 = 0.16

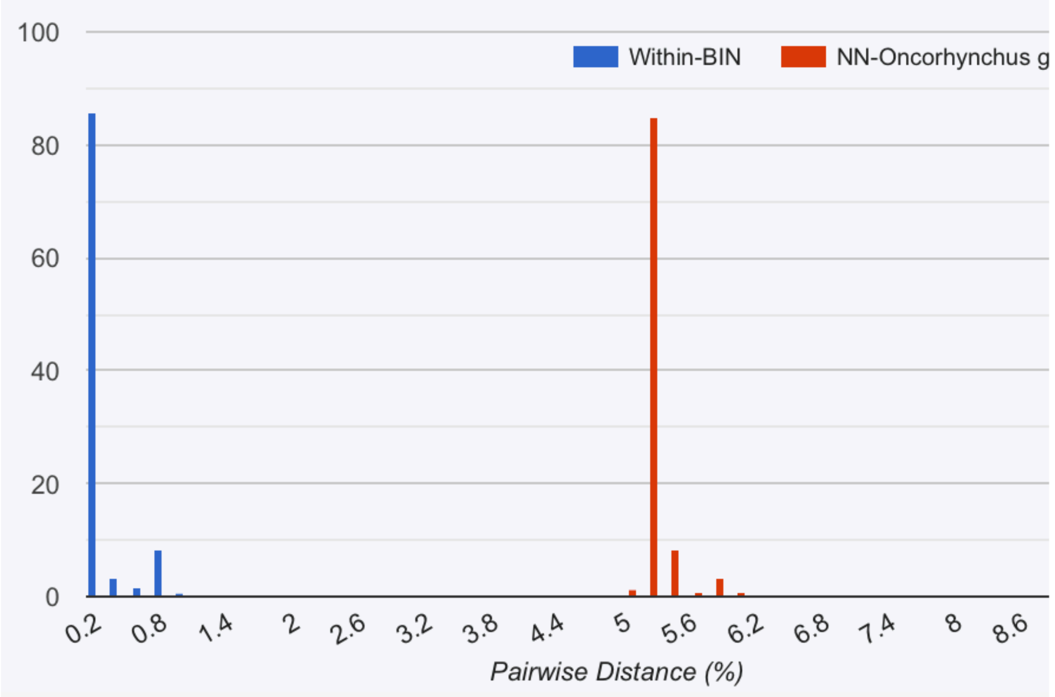
*q*: proportion of transversions

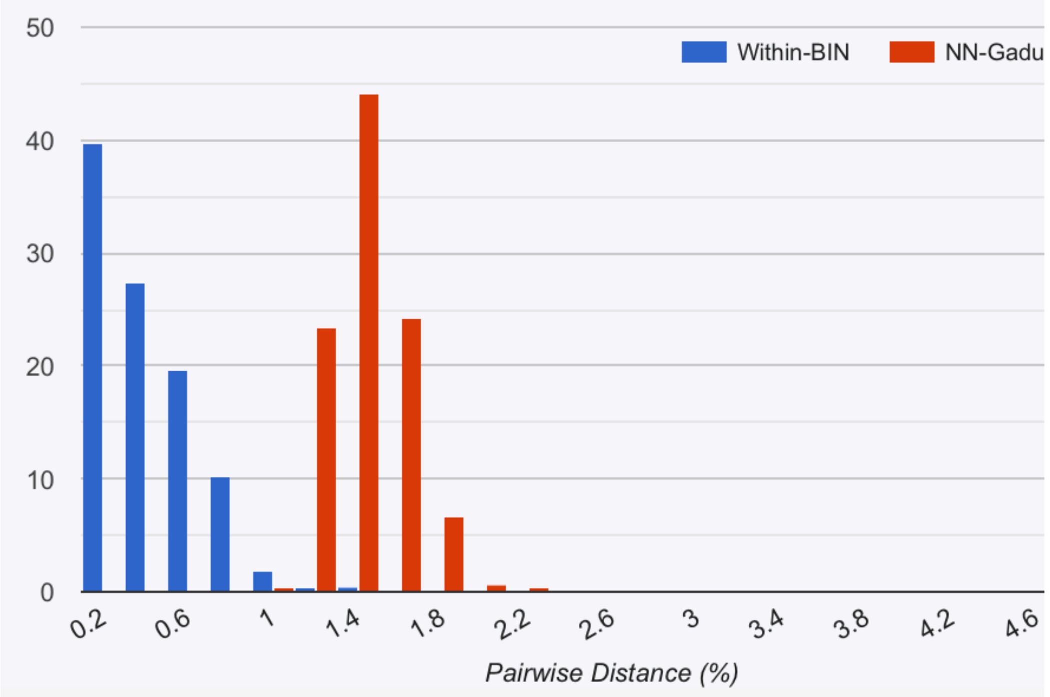
5 / 50 = 0.1

# Barcode gap

Exercise 2

Comparing genetic distances within BIN and to nearest neighbor (NN) BIN





Large barcode gap

*Oncorhynchus keta*

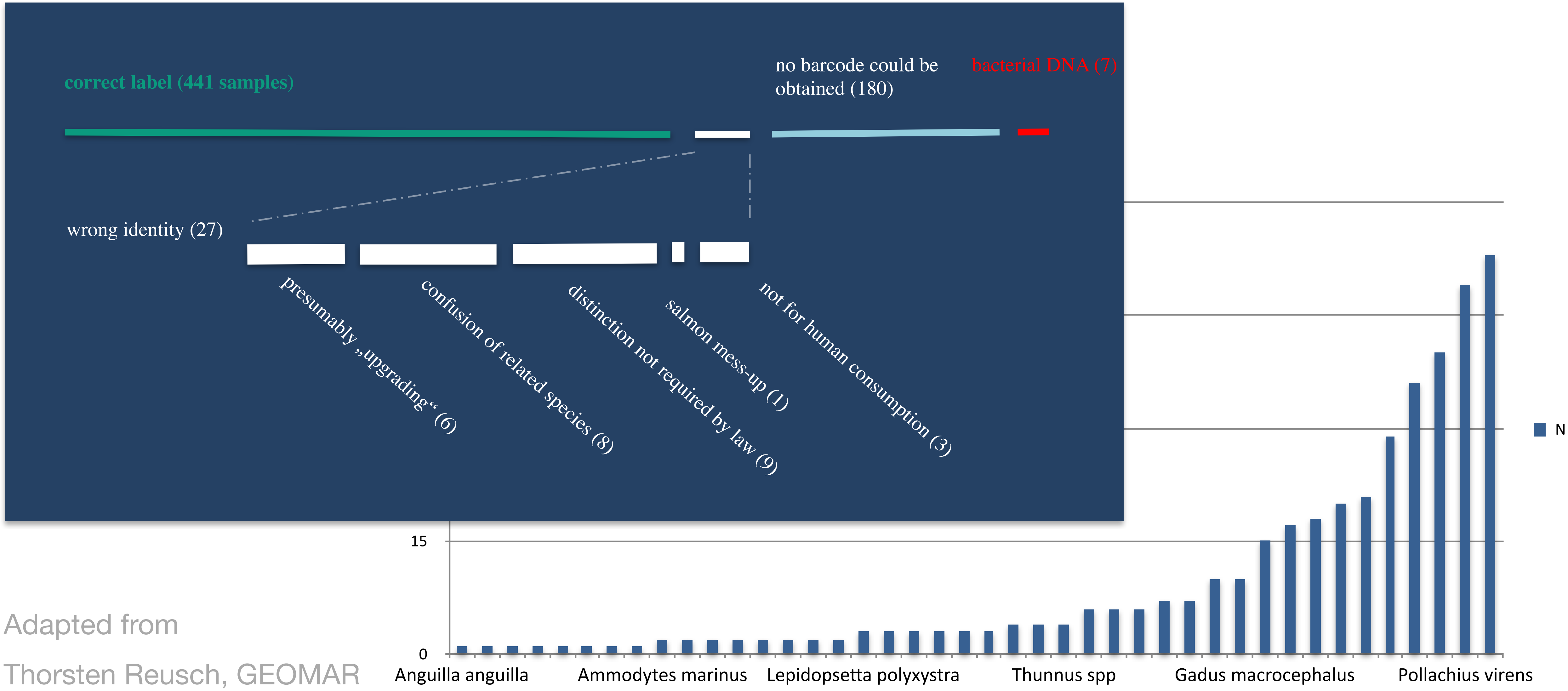
Small barcode gap

*Gadus chalcogrammus*

# #fischdetektive results

Mislabeling seems to be only a moderate problem in Germany in frozen and fresh fish

(but may be higher in Sushi-grade fish and processed fish products)



correct label (441 samples)

no barcode could be obtained (180)

bacterial DNA (7)

wrong identity (27)

presumably „upgrading“ (6)

confusion of related species (8)

distinction not required by law (9)

salmon mess-up (1)

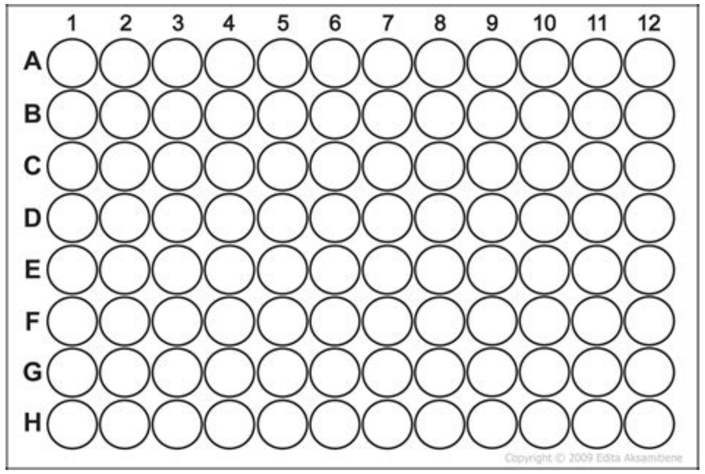
not for human consumption (3)

Adapted from

Thorsten Reusch, GEOMAR

# Portable 3rd gen sequencing

Oxford Nanotechnologies MinION

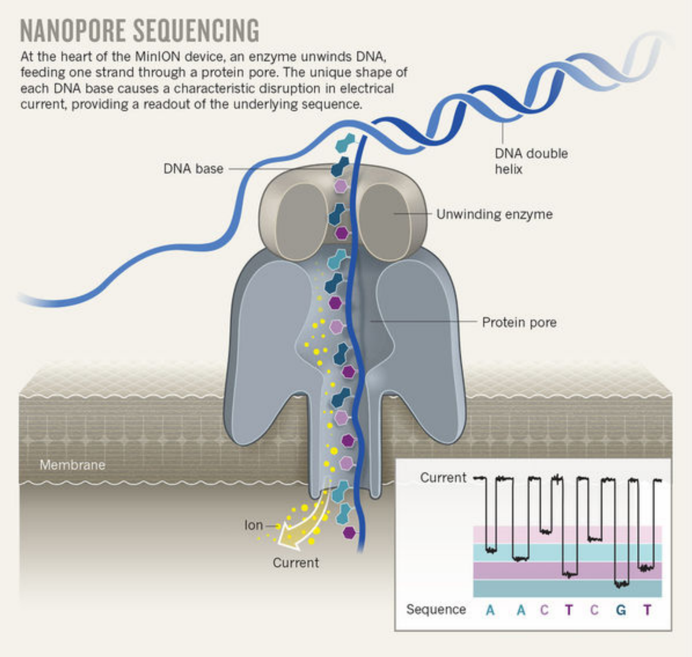


96-well plate

Indexing during PCR

Library preparation

whatech.com



NANOPORE SEQUENCING

At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.

blogs.nature

# Take-home messages

- DNA barcoding can be a powerful tool to identify species from tissue or environmental samples

- It is not foolproof, relies on the quality of the reference database, and does not replace taxonomic expertise

- For animals, the most common marker is cytochrome *c* oxidase I (COI)

- Important databases are BOLD (curated) and GenBank (most comprehensive)