Marine Ecological Genetics

12. DNA barcoding | Computer practical

- Extract DNA barcodes from Sanger reads
- Identify samples with the Barcode of Life Data System
- Evaluate genetic distances and id quality

Martin Helmkampf

Summer 2024

Carl von Ossietzky
Universität
Oldenburg

Access files for practical

Download files from GitHub:

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

Alternatively, run git from terminal:

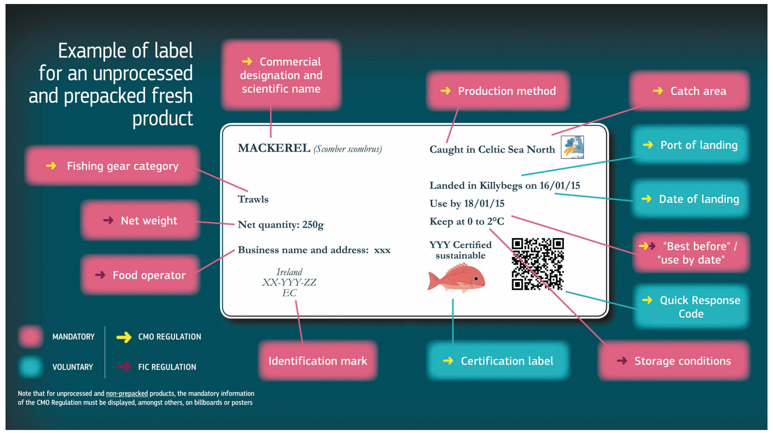
```
git clone https://github.com/mhelmkampf/meg24.git
```

Open R script called **Mar_Ecol_Gen_week2.R** in RStudio

#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?



Example of label for an unprocessed and prepacked fresh product

→ Commercial designation and scientific name

→ Production method

→ Catch area

→ Fishing gear category

→ Port of landing

→ Net weight

→ Date of landing

→ Food operator

→ "Best before" / "use by date"

→ Quick Response Code

→ Identification mark

→ Certification label

→ Storage conditions

MACKEREL (*Scomber scombrus*)

Caught in Celtic Sea North

Trawls

Landed in Killybegs on 16/01/15

Use by 18/01/15

Keep at 0 to 2°C

Net quantity: 250g

Business name and address: xxx

YYY Certified sustainable

Ireland
XX-YYY-ZZ
EC

🟥 MANDATORY	→ CMO REGULATION
🟦 VOLUNTARY	→ FIC REGULATION

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

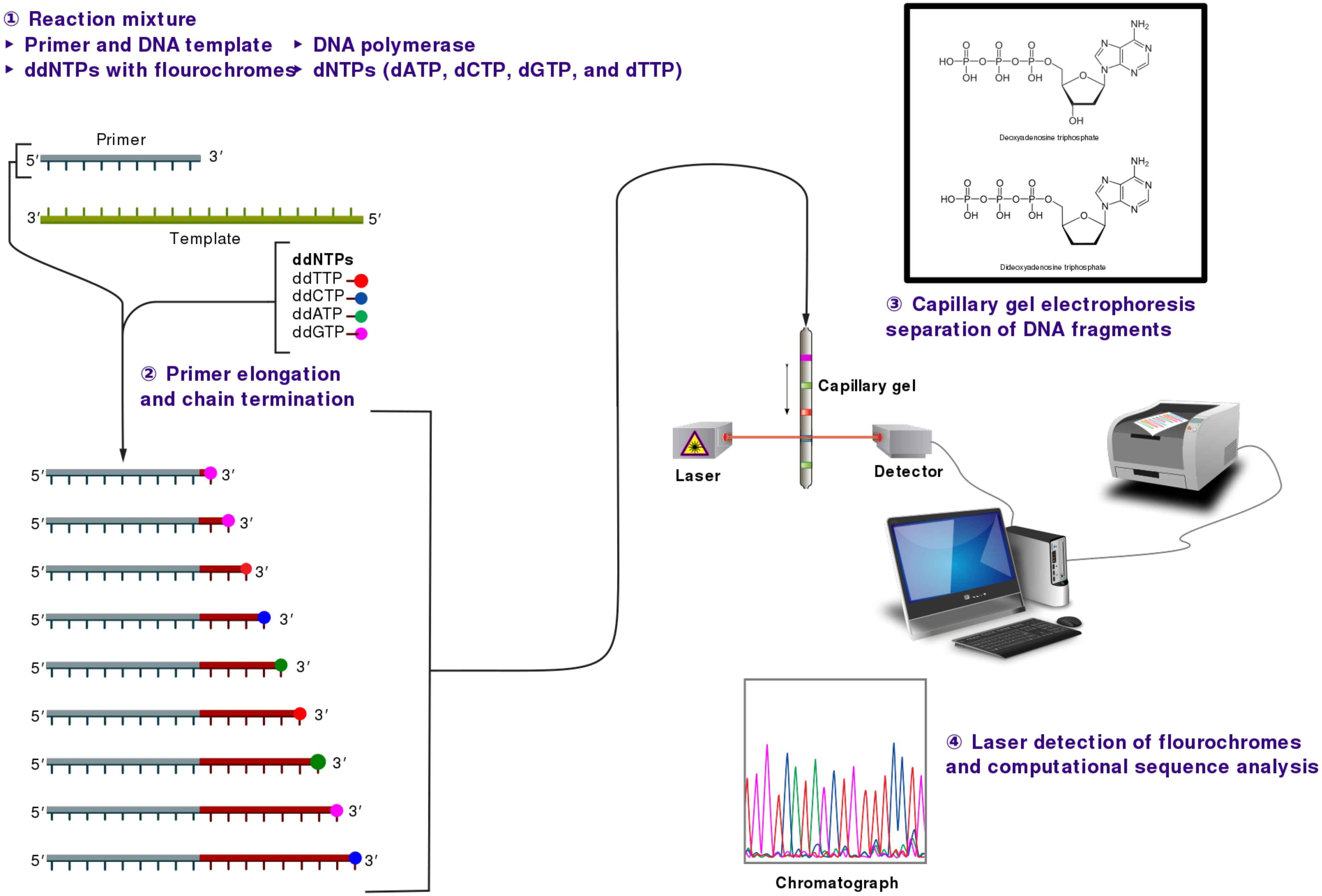
Sample collection → DNA extraction → PCR (COI) → Sequencing → ID

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

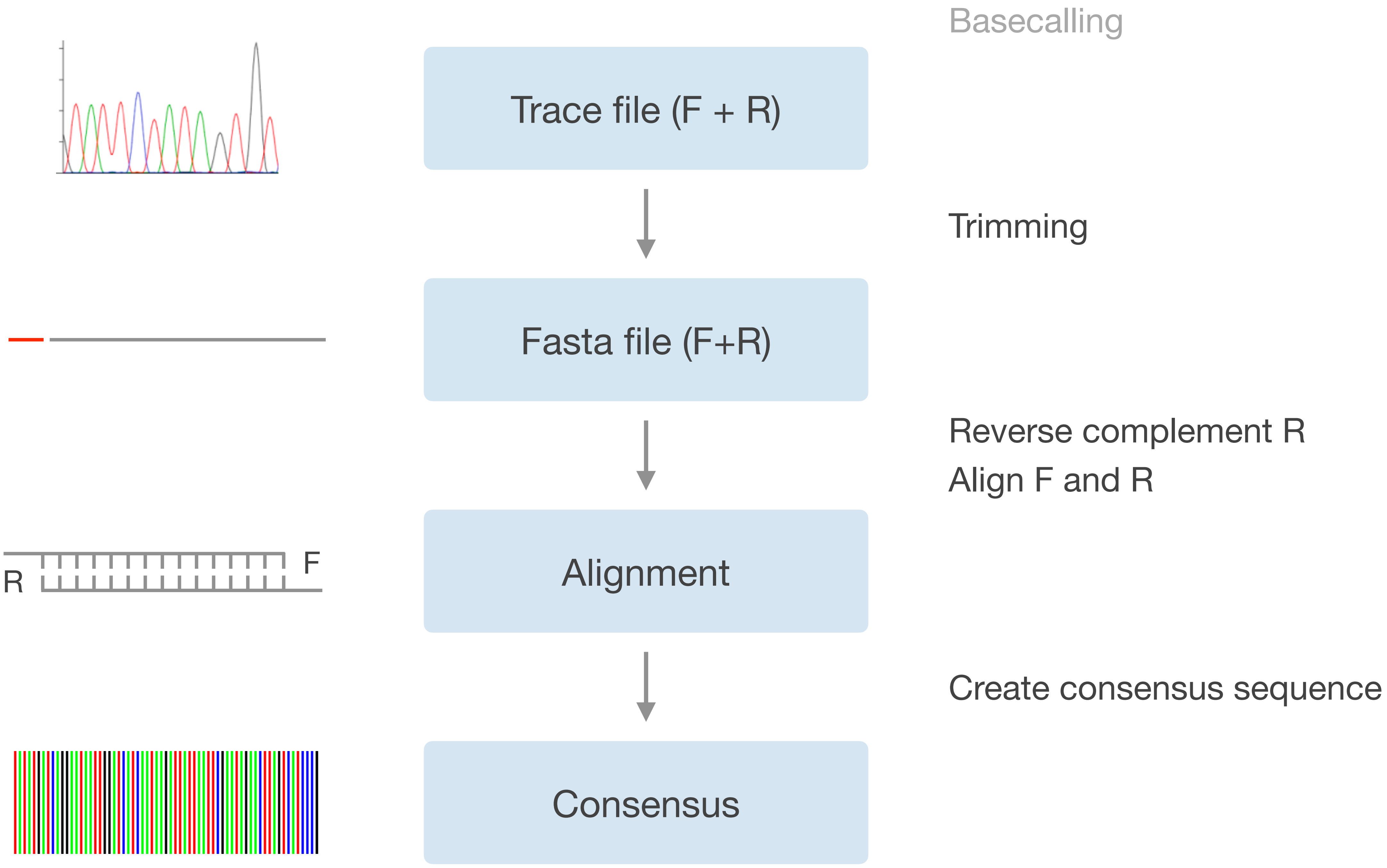
③ Capillary gel electrophoresis separation of DNA fragments

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)

Genetic distances

Exercise 2

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

Uncorrected distance

p = 13 / 50 = 0.26

K2P distance (Kimura 1980)

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

K = 0.33

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

8 / 50 = 0.16

q: proportion of transversions

5 / 50 = 0.1

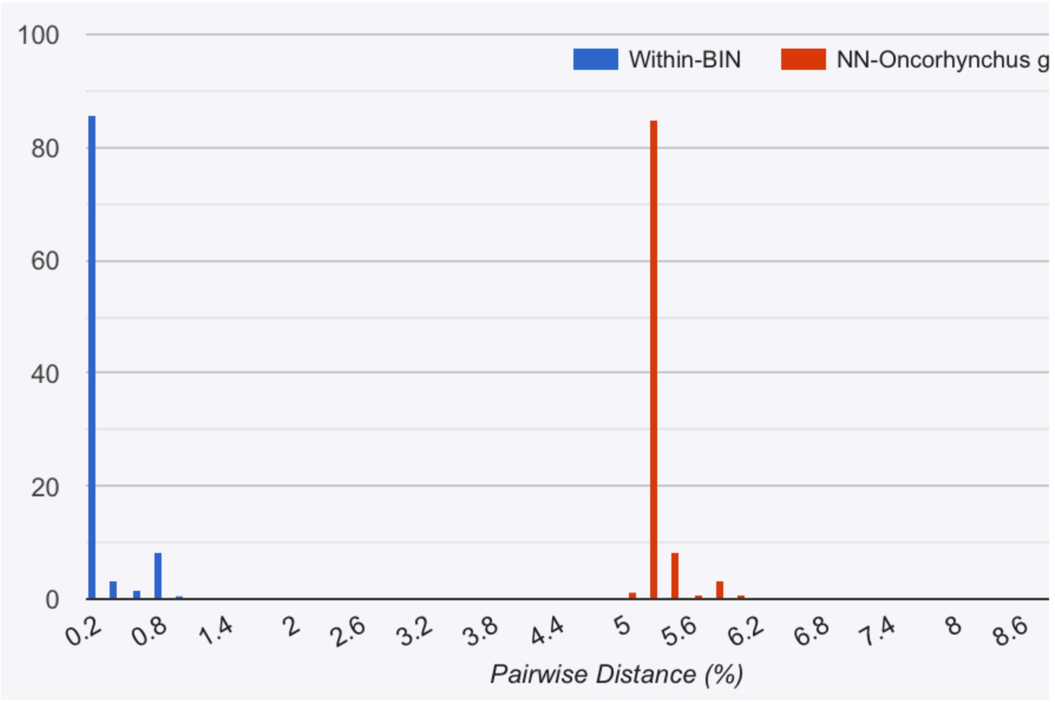
BOLD

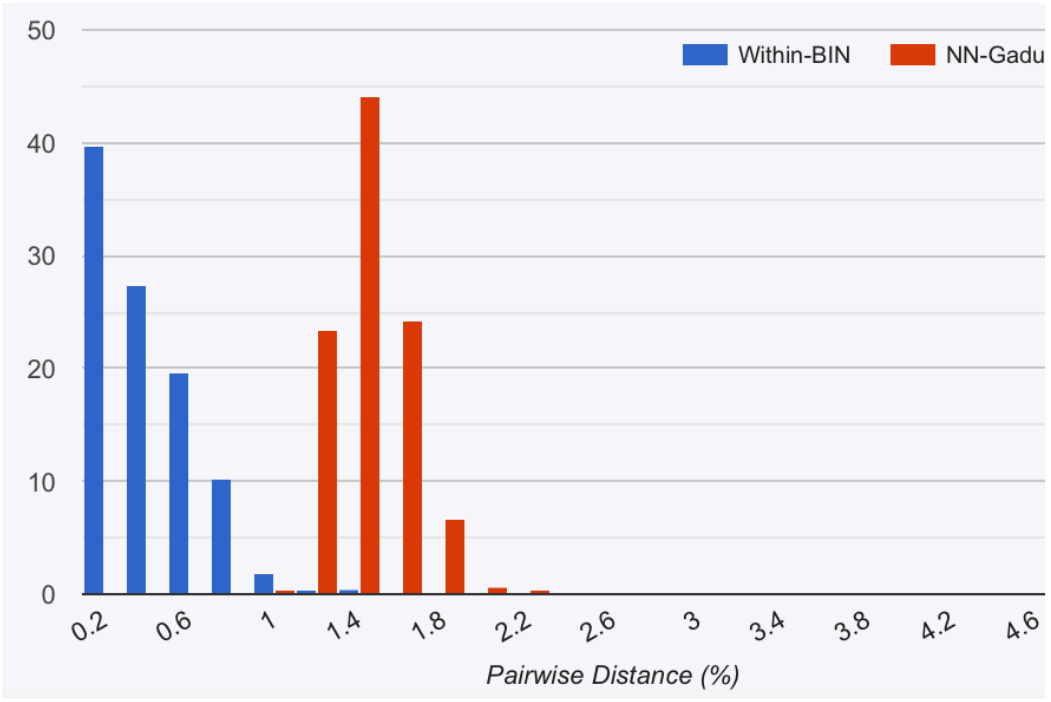
Exercise 2

Barcode gap

Exercise 2

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





Large barcode gap

Oncorhynchus keta

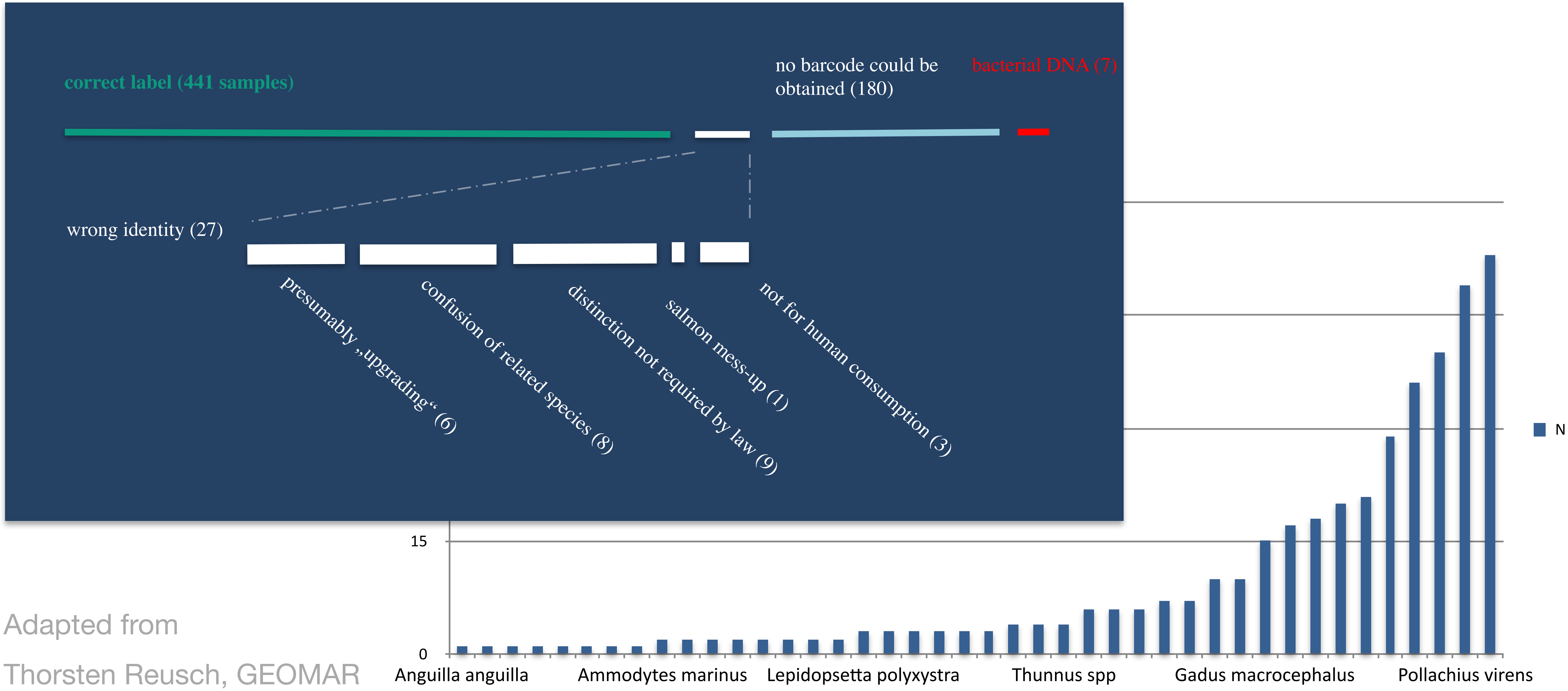
Small barcode gap

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)

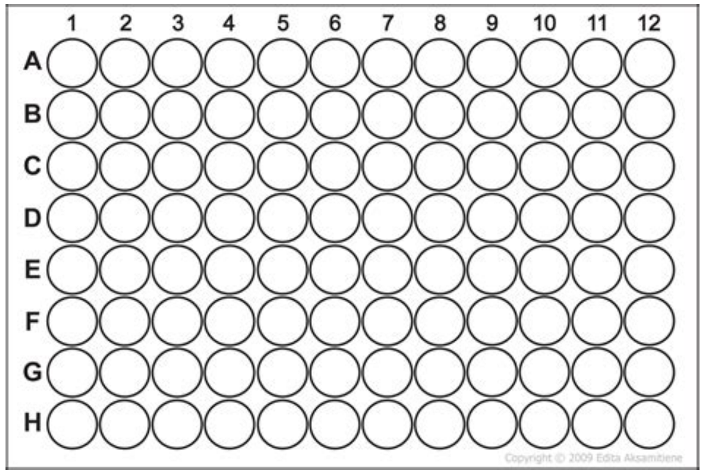


Adapted from

Thorsten Reusch, GEOMAR

Portable 3rd gen sequencing

Oxford Nanotechnologies MinION



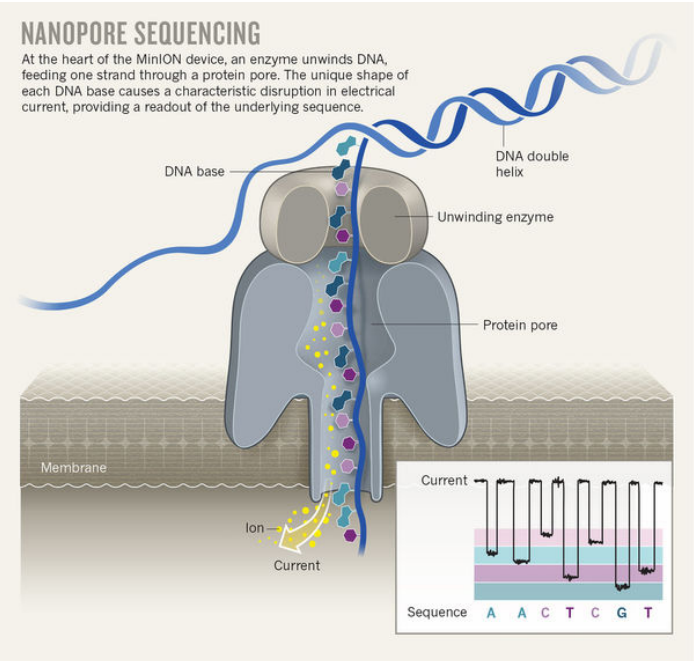
96-well plate

→

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

- Extract DNA barcodes from Sanger reads
- Identify samples with the Barcode of Life Data System
- Evaluate genetic distances and id quality