Exercises in Marine Ecological Genetics

09. DNA barcoding

- Extract barcodes from Sanger reads
- Match sequence to BOLD database
- Evaluate id quality using genetic distances

Martin Helmkampf

Foto: dpa

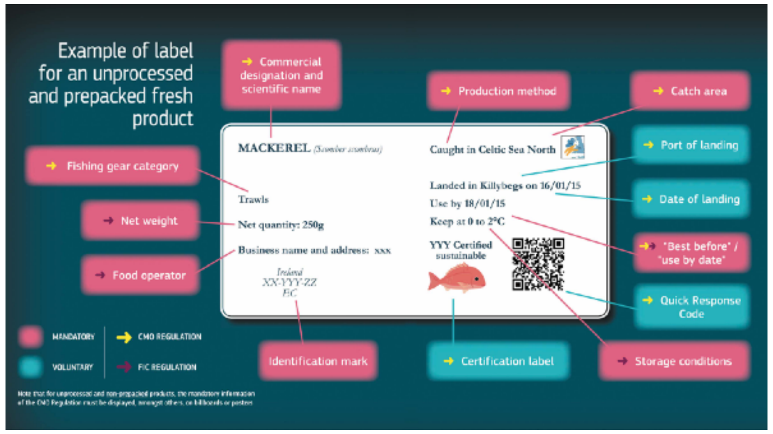
https://github.com/mhelmkampf/meg25

#fischdetektive

Where does our seafood come from, and is it labeled correctly?

Citizen science project at GEOMAR (2017) with over 700 participants (10–14 years)

Rotbarsch Filet Nordatlantik Schleppnetz 100g 2,49
Steinbeisser-Katfischfilet Nordostatlantik Schleppnetz Frisch 100g 2,45
Welsfilet Aqua kultur Deutschland Frisch 100g 1,69
Frisches Lachsfilet Zucht Norwegen



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)

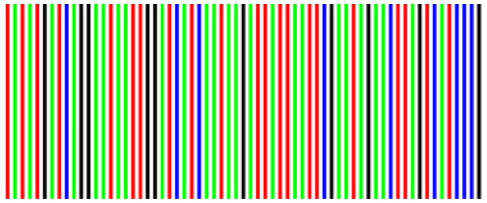
↓

Sequencing

←

Matching

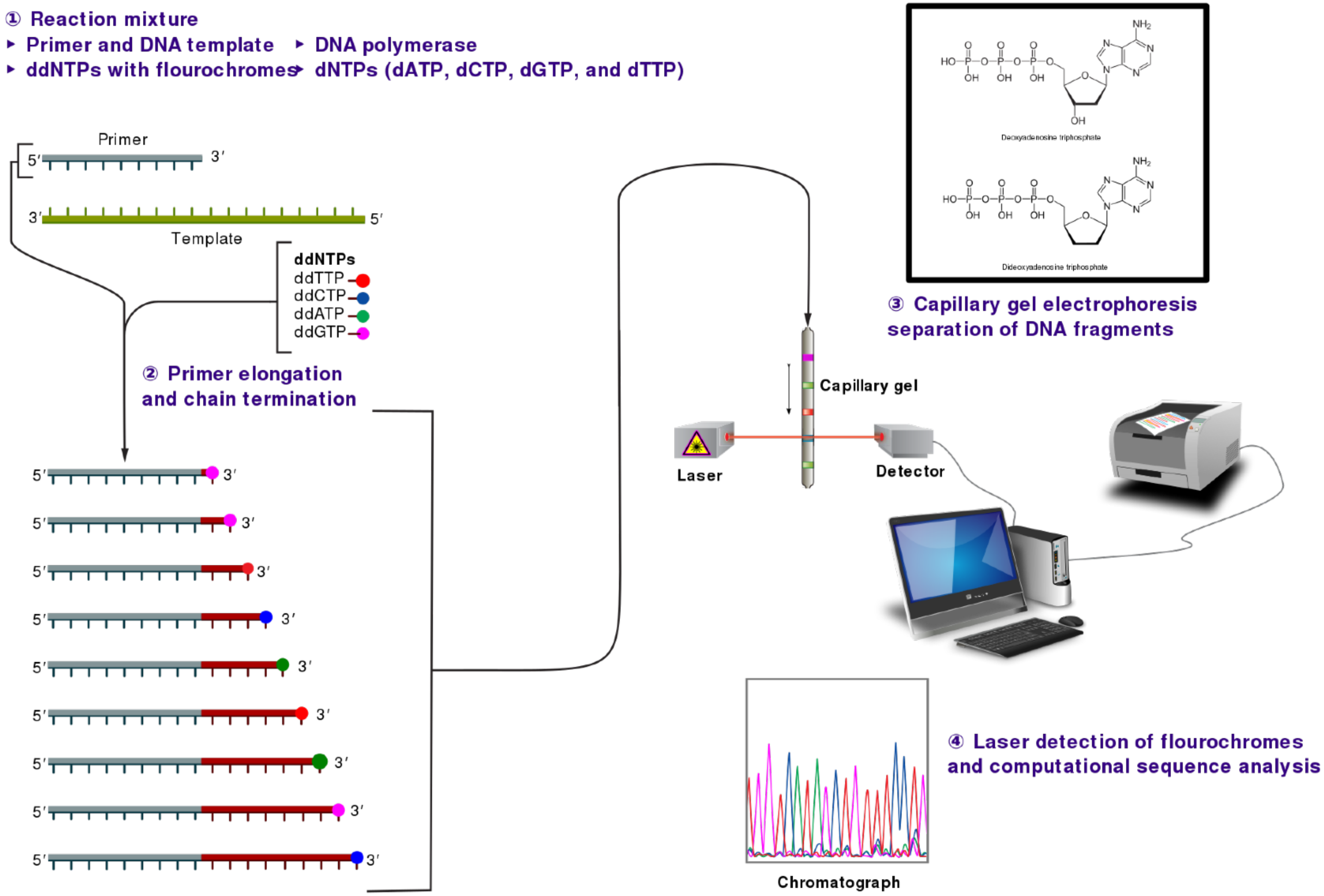
COI barcode



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
ATAAGCTTCTGACTCCTACCTCCGTCCTTCCTCCTCCTTTCTTCATCTGGAGTTGAAGCCGGCGCTGGTACCGGGTGGACAGTTTATCCCCCTCTA
GCCGGAAACCTTGCCCACGCAGGAGCATCTGTCGACTTAACCATCTTCTCCCTCCATTTAGCTGGAATCTCCTCAATTTTGGGGGCCATTAATTTTATT
ACGACCATTATCAACATAAAACCCCCAGCTATTTCTCAGTACCAAACCCCGCTTTTTGTCTGAGCTGTACTAATCACTGCTGTACTTCTACTATTATCA
CTCCCCGTTCTGGCAGCAGGTATTACTATGTTGCTCACAGATCGAAATTTAAACACCACTTTCTTTGACCCGGCGGGTGGCGGAGATCCAATTTTATAC
CAACACCTCTTTTGATTCTTCGGTCACCCAGAGGTCTATATTCTGATCCTCCCAGGCTTTGGTATAATTTCACATATCGTTGCATATTACTCTGGTAAG
AAAGAACCTTTCGGGTACATAGGAATAGTGTGAGCTATAATAGCCATCGGCTTGTTAGGATTTATCGTTTGAGCCCACCACATATTTACTGTCGGGATG
GACGTGGACACTCGTGCCTACTTTACATCTGCCACCATAATTATCGCTATCCCCACAGGAGTAAAAGTATTTAGCTGACTAGCTACACTGCACGGAGGC
TCGATCAAATGAGAGACACCACTTCTCTGAGCCCTAGGATTTATCTTCCTATTTACAGTGGGCGGATTAACGGGCATCGTCCTTGCTAACTCCTCATTA
GACATTGTTTTACATGACACTTATTACGTAGTCGCCCATTTCCACTACGTACTCTCAATAGGAGCTGTATTTGCCATTATGGGCGCTTTCGTACACTGA
TTCCCCCTATTCACAGGGTACACCCTTCACAGCACATGAACCAAAATCCATTTTGGAATTATATTTATCGGTGTAAATTTAACCTTTTTCCCACAGCAT
TTCCTAGGCCTCGCAGGGATACCACGACGGTACTCTGACTACCCGGACGCCTACACGCTATGAAACACTGTATCCTCAATCGGATCCCTTGTCTCCTTA
GTAGCTGTAATTATGTTCCTATTTATTCTTTGAGAGGCTTTTGCTGCCAAACGAGAAGTAGCATCAATCGAAATAACTTCAACAAACGTAGAATGACTA
CACGGATGCCCCCCACCCTACCACACATTCGAGGAACCAGCATTTGTCCAAGTACGAACGTACTAA
```

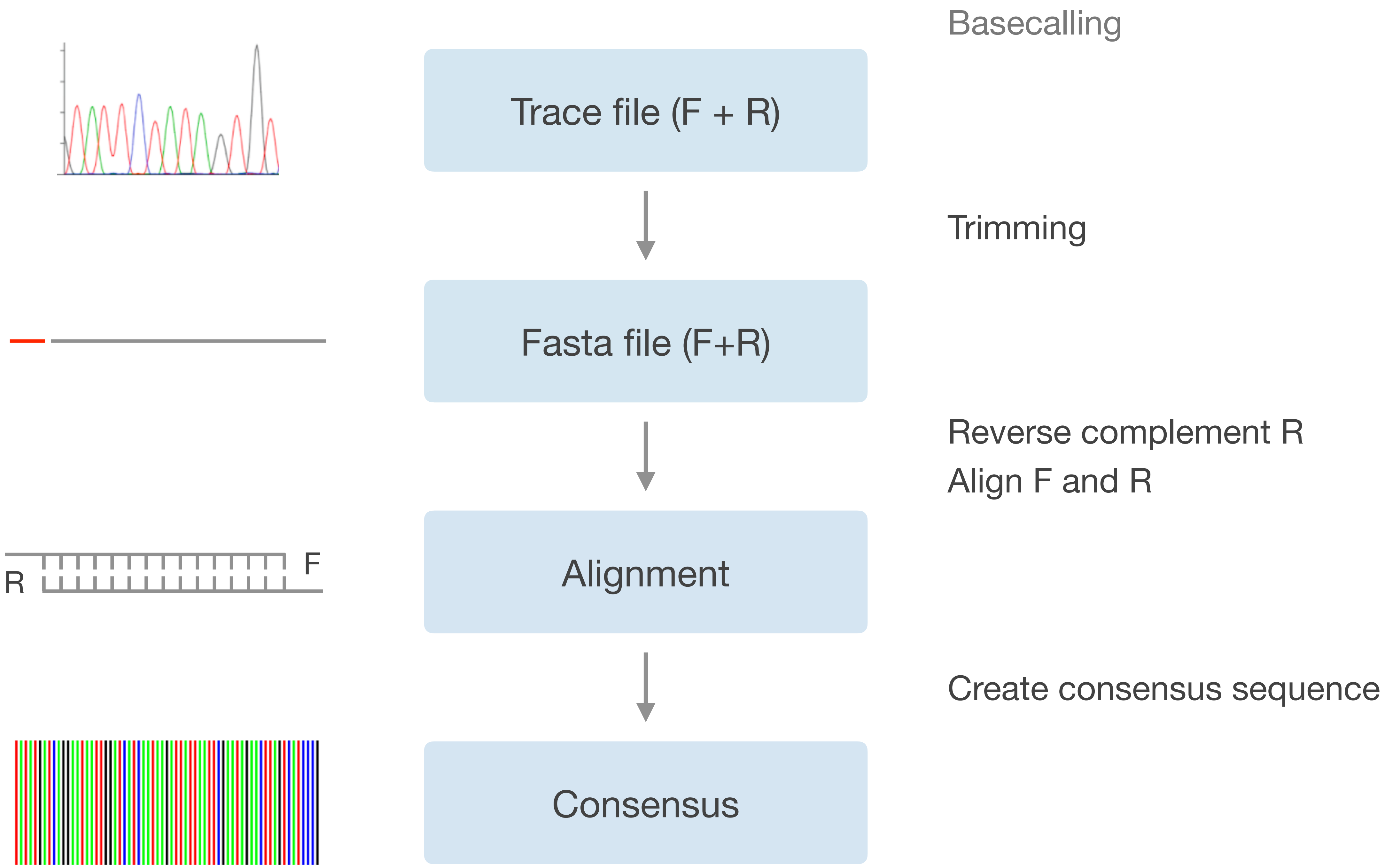
Sanger sequencing


① Reaction mixture
▸ Primer and DNA template
▸ DNA polymerase
▸ ddNTPs with flourochromes
▸ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
5′
3′
3′
5′
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
Deoxyadenosine triphosphate
Dideoxyadenosine triphosphate
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
Chromatograph
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 are minimized

- Accomplished by introducing gaps (–), which represent insertions or deletions (**indels**) and account for sequence length differences due to mutations over time
- Computationally complex, often requires heuristic solutions
- **Reveals evolutionary or functional relationships** between sequences (e.g. homology)
- Key to variant calling, sequence assembly, species identification and phylogenetics

Barcode Index Number (BIN)

Exercise 3

- **Clusters of similar COI barcodes**
- Clustering is based on genetic similarity, independent of formal taxonomy
- Represent operational taxonomic units (OTUs) that often correspond to species
- Enable species identification and biodiversity assessments

Genetic distance

Exercise 3

Uncorrected or *p*-distance:

Proportion of nucleotides at which two sequences differ

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

$p = 13 / 50 = 0.26$

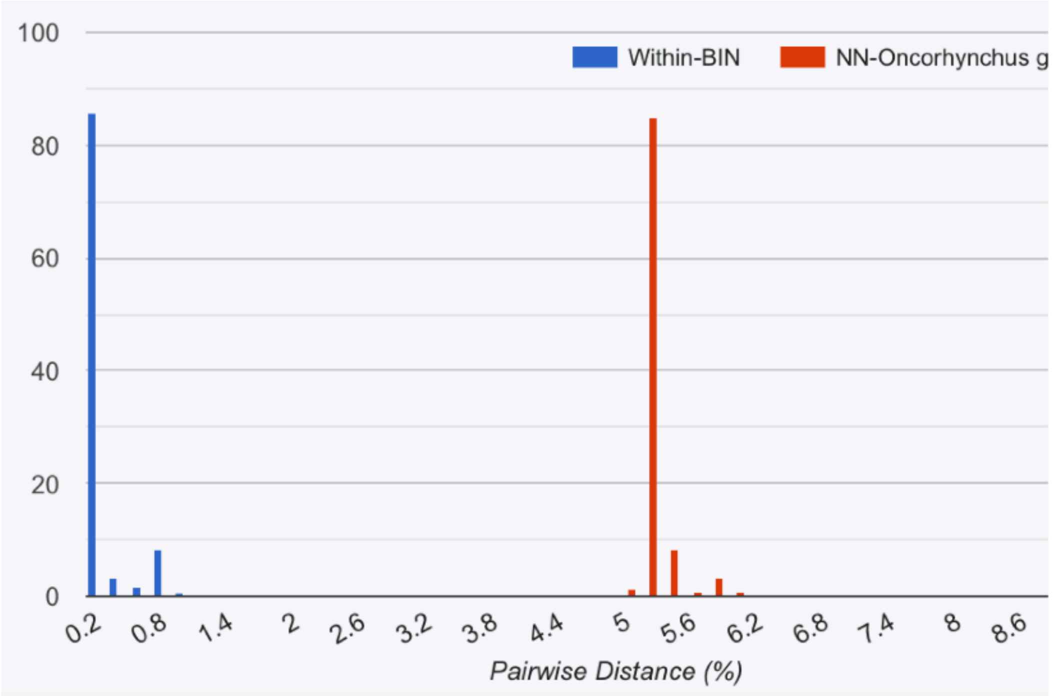
Does not correct for multiple substitutions (repeated mutations of the same site)

or transition / transversion bias (> K2P distance, Kimura 1980)

Barcode gap

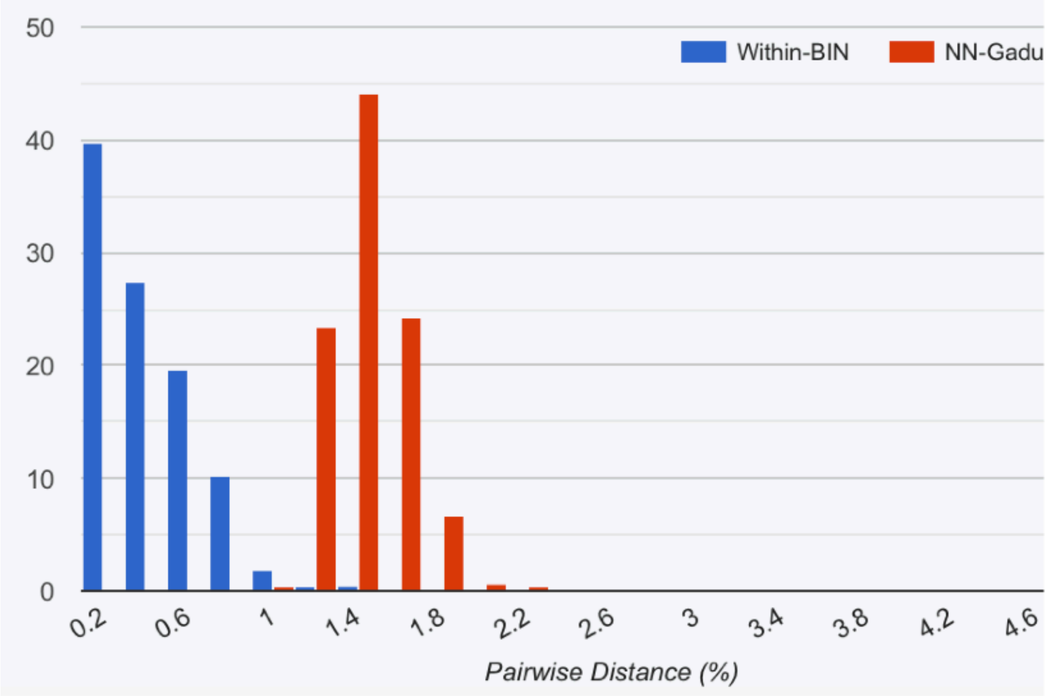
Comparing genetic distances within BIN and to nearest neighbor (NN) BIN

Comparison	Typical p-distance (COI)
Within species	< 2 %
Between species	> 2–3 %
Between genera	10–20 %



Large barcode gap

Oncorhynchus keta



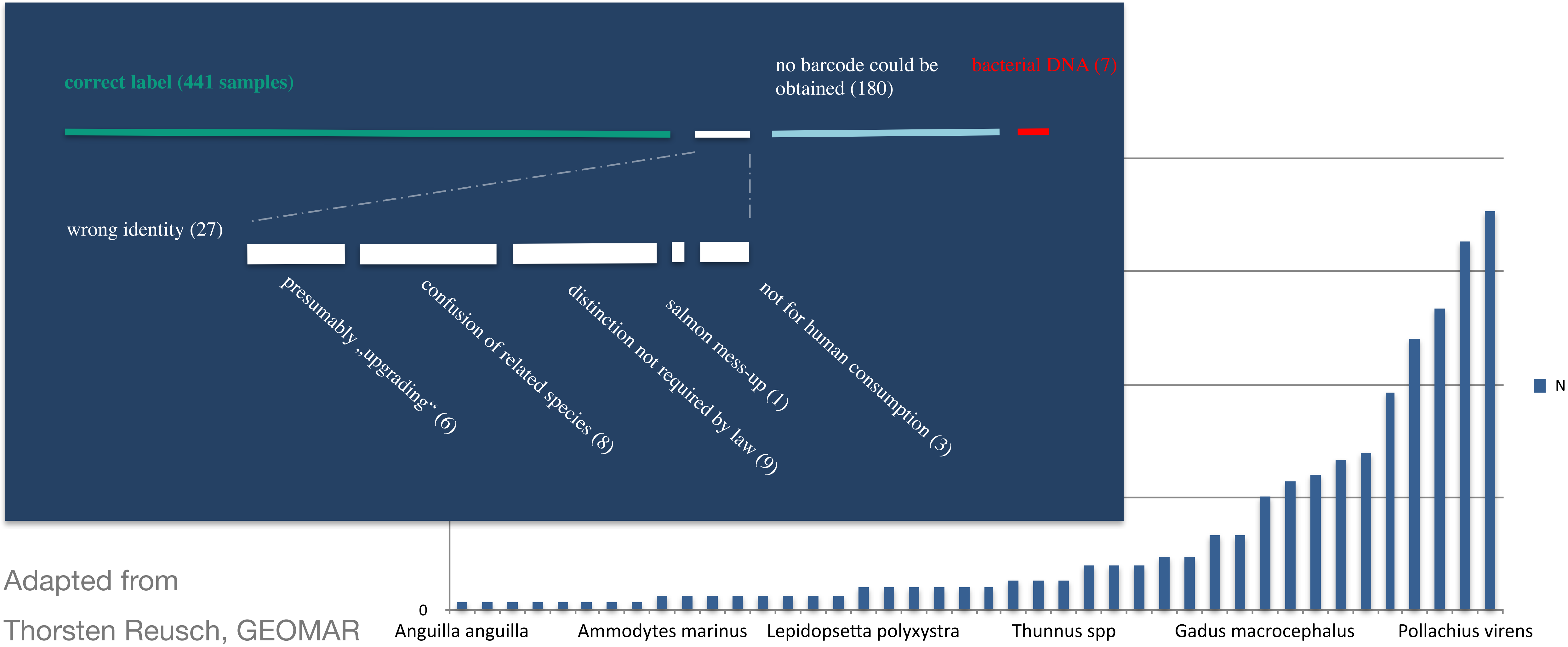
Small barcode gap

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)

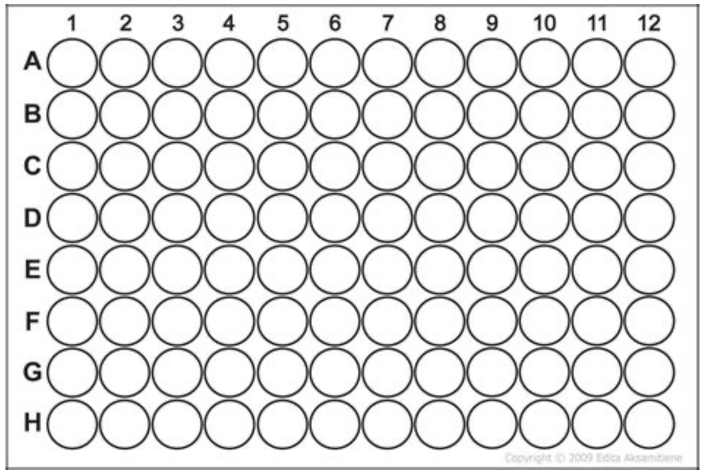


Adapted from

Thorsten Reusch, GEOMAR

Portable 3rd gen sequencing

Oxford Nanotechnologies MinION



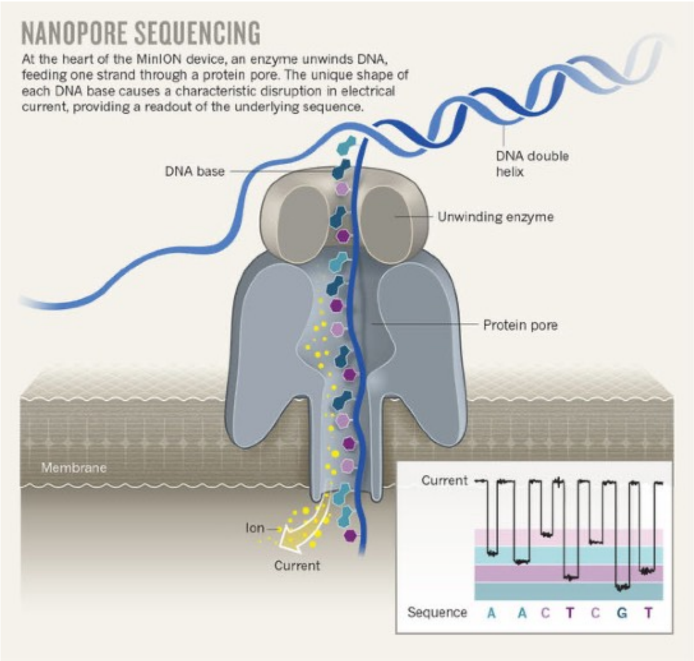
96-well plate

→

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

DNA barcoding

Summary

```
seqkit subseq -r start:end input.fas > output.fas

seqkit seq -r -p input.fas > output.fas

merger -asequence forward.fas -bsequence reverse.fas … -outseq consensus.fas
```

- Accurate barcodes like COI can be extracted from high-quality Sanger reads after **careful processing**, including trimming and consensus formation
- **Matching to a reference database** like BOLD allows to assign samples to species, but results depend on sequence quality and database completeness
- **Identification reliability** may be assessed using sequence similarity and genetic distance to nearest-neighbor (“barcode gap”)

Evaluation — please participate!

https://elearning.uni-oldenburg.de/plugins.php/unizensusplugin/show?cid=3660d16e8eb3daf479389cf8233c12fb