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

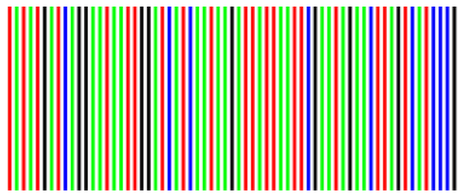
https://github.com/mhelmkampf/meg25

Course outline

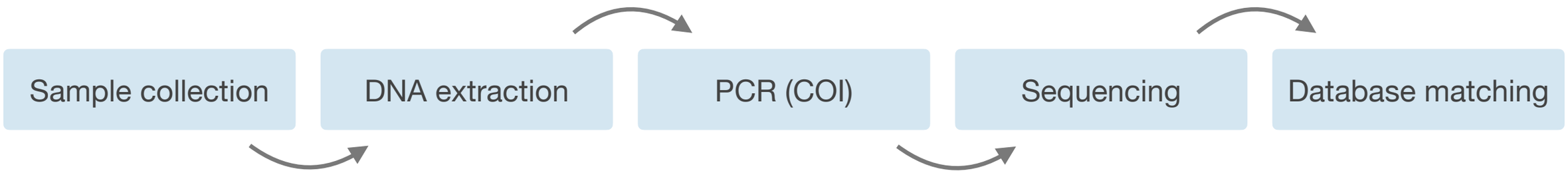
Updated, but may be subject to change

Class	Date	Topics	Script
01	Apr 11	Introduction, setup	01_intro.R
–	Apr 18	Good Friday	
02	Apr 25	Hardy-Weinberg equilibrium	02_hwe.R
03	May 02	Population structure I	03_popst.R
04	May 09	Genome sequencing and assembly	04_asm.sh
05	May 16	Variant calling and SNPs	05_vcal.sh
06	May 23	Population genomics and genetic diversity	06_gdiv.sh
–	May 30	Himmelfahrt break	
07	Jun 06	Population structure II	07_clust.sh
–	Jun 13	Selection	08_sel.R
08	Jun 20	Student presentations – no exercises	
09	Jun 27	DNA barcoding	09_barcode.sh
10	Jul 04	Metabarcoding / eDNA	
11	Jul 11	Introduction to phylogenetics	

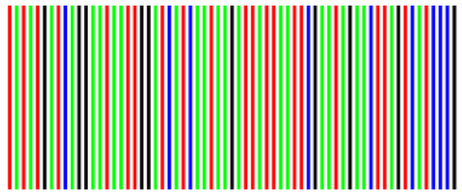
DNA barcoding



- Method of **species identification** using short, standardized gene region (COI, cytb, 16S …)
- Barcode is compared to **reference sequence database** to find closest match
- **Applications**
    - Identify species from parts or life stages (detect food fraud, wildlife trafficking)
    - Determine species boundaries (complement traditional taxonomy)
    - Analyze whole communities (biodiversity surveys, microbial ecology: metabarcoding)



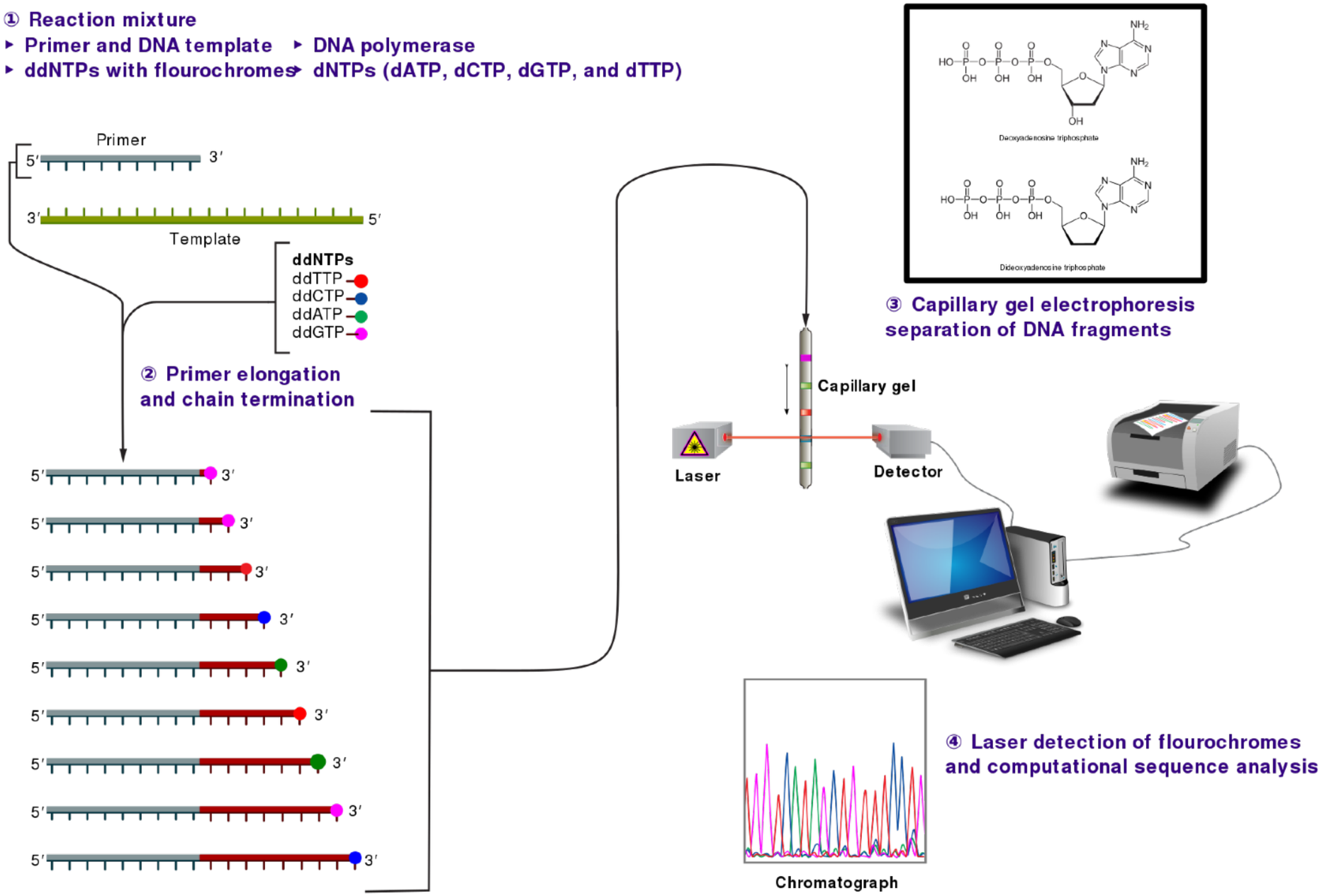
COI barcode



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

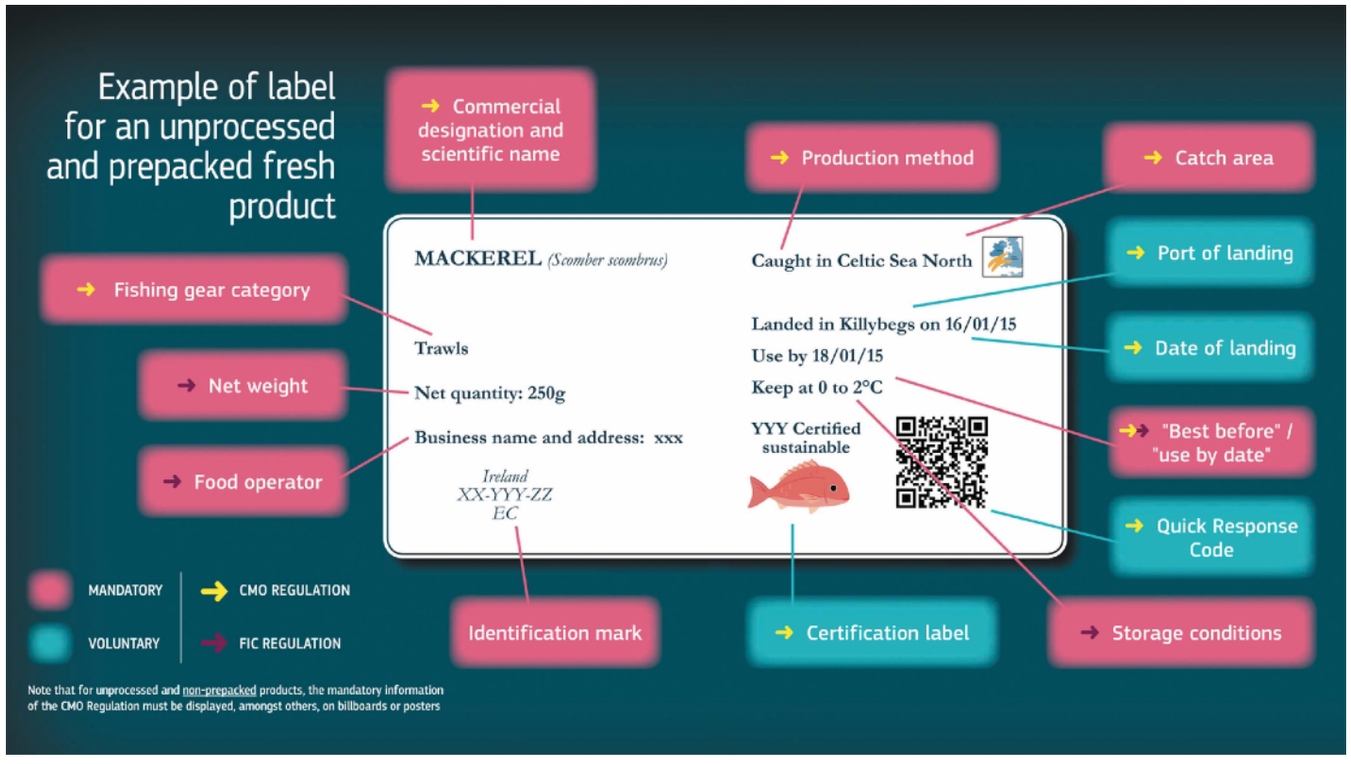


#fischdetektive

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

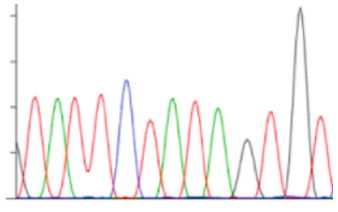
Citizen science project at GEOMAR in 2017 with over 700 participants (10–14 years)

Thorsten Reusch, GEOMAR



Sanger read processing

Exercise 1



Trace file (F + R)

Basecalling

Trimming

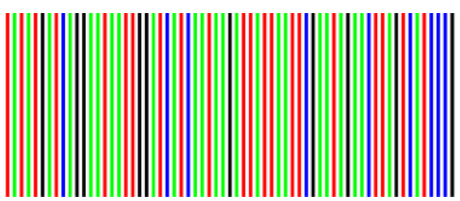
Fasta file (F+R)

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**) / account for sequence length differences due to mutations over time
- **Reveals evolutionary or functional relationships** between sequences (e.g. homology)
- Key to variant calling, sequence assembly, species identification and phylogenetics
- Computationally complex, often requires heuristic solutions

Barcode Index Number (BIN)

Exercise 2

- **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 2

Uncorrected or *p*-distance:

Proportion of nucleotides at which two sequences differ

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

p = 13 / 50 = 0.26

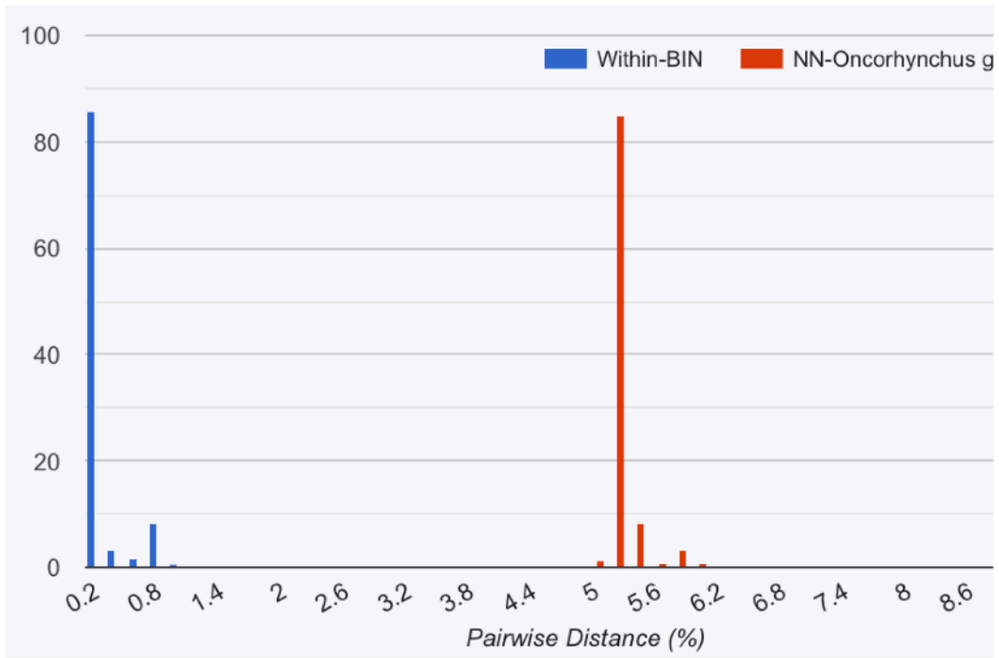
Does not correct for multiple substitutions (repeated mutations of the same site)

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

Barcode gap

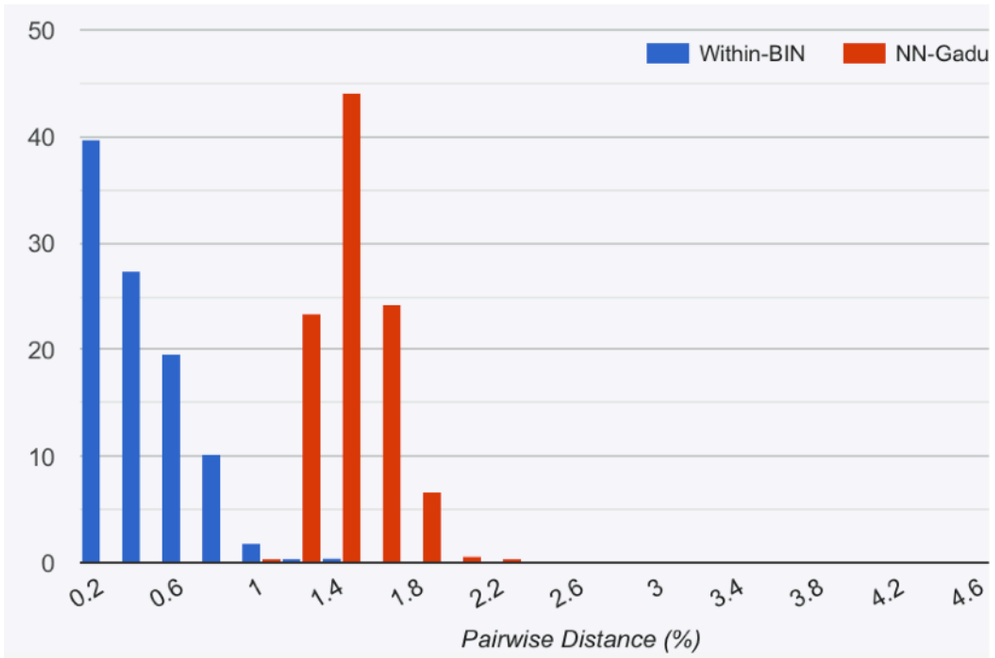
Comparing genetic distances within BIN and to nearest neighbor (NN) BIN

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



Large barcode gap

Oncorhynchus keta



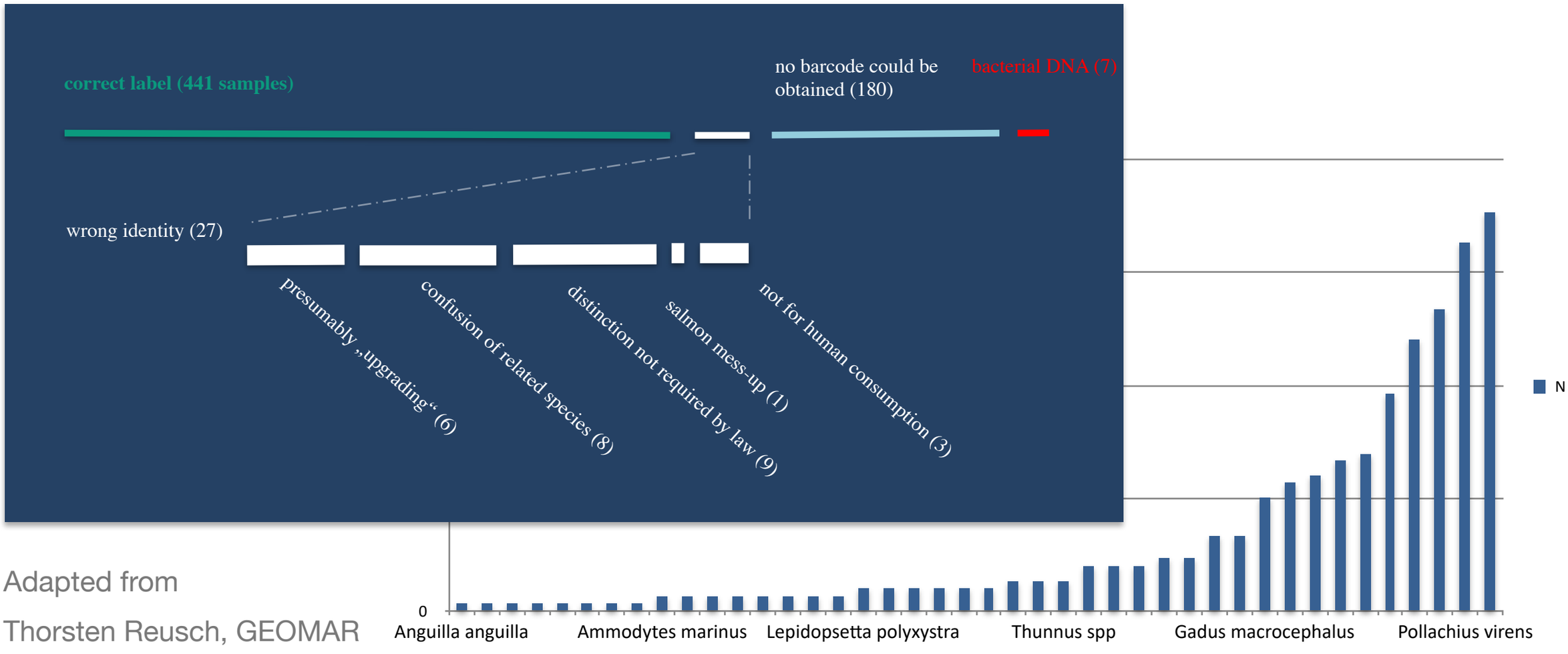
Small barcode gap

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

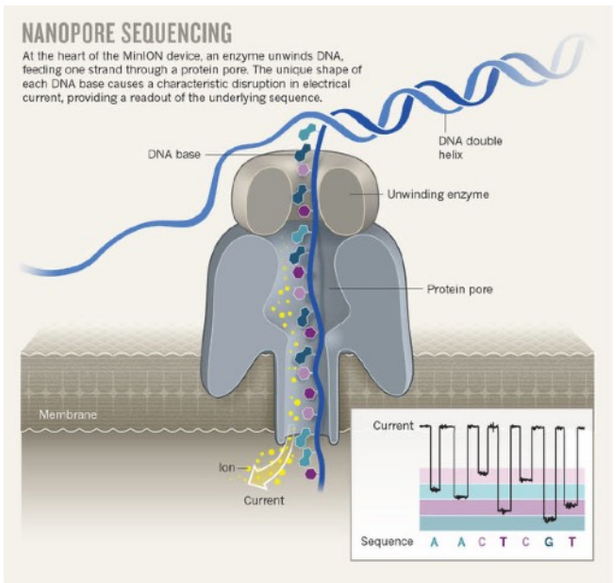
Portable 3rd gen sequencing

Cretus Joseph Mtonga

PhD project: Build database of the fish fauna in the Western Indian Ocean to enable biodiversity surveys / fisheries monitoring using eDNA

Oxford Nanotechnologies MinION

whatech.com



blogs.nature.com

DNA barcoding

Summary

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

seqkit seq -r -p input.fas > output.fas                            # reverse complement

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

- Accurate barcodes like COI can be obtained 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, database completeness, and species

- **Identification reliability** may be assessed using sequence similarity and genetic distance to nearest-neighbor (“barcode gap”)

Course evaluation — please participate!

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