

# Exercises in Marine Ecological Genetics

## 10. DNA barcoding

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

Martin Helmkamp

# Download course materials using git

Go to project directory

```
cd dir          # e.g. Documents/meg23_exercises  
ls -l           # view directory contents, long format
```

Update course repository

```
cd meg23_repo  
git pull
```



In case of an error message

```
cd ..                # go back to project directory  
rm -rf meg23_repo    # delete old repository  
git clone https://github.com/mhelmkampf/meg23\_repo.git
```

# Avoiding version conflict

Please do not save over files in the course repository. Instead, save your own scripts to the local subdirectory (including copies of course scripts you would like to edit), e.g with

```
cp code/10_barcode.sh ../local/10_barcode_lc.sh      # cp [source] [destination]
```

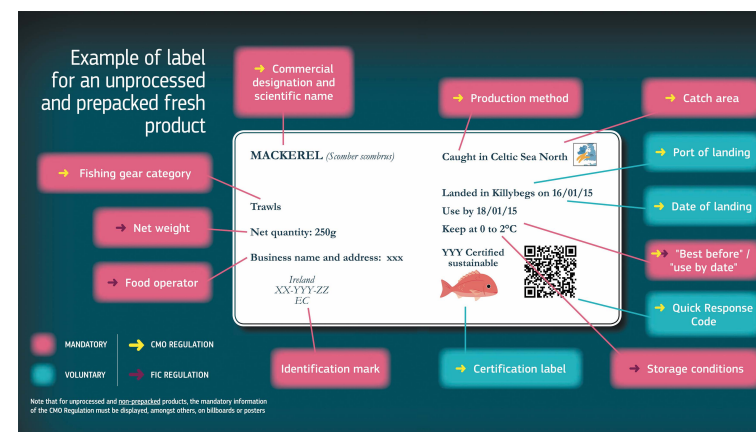
# #fischdetektive

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

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



Thorsten Reusch, GEOMAR



Sample collection



DNA extraction



PCR (COI)

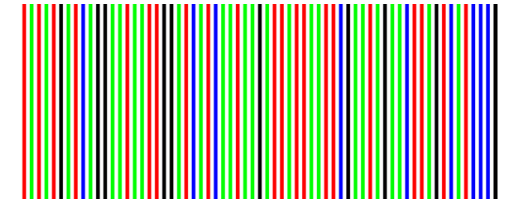


Sequencing



ID

# COI barcode



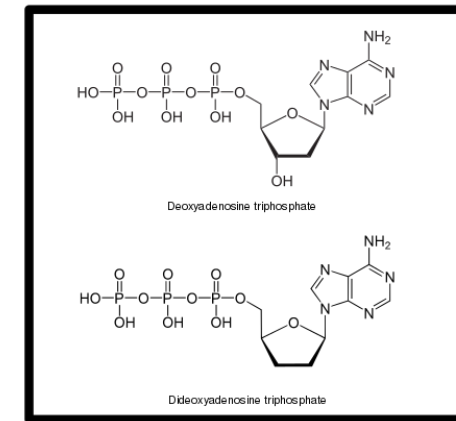
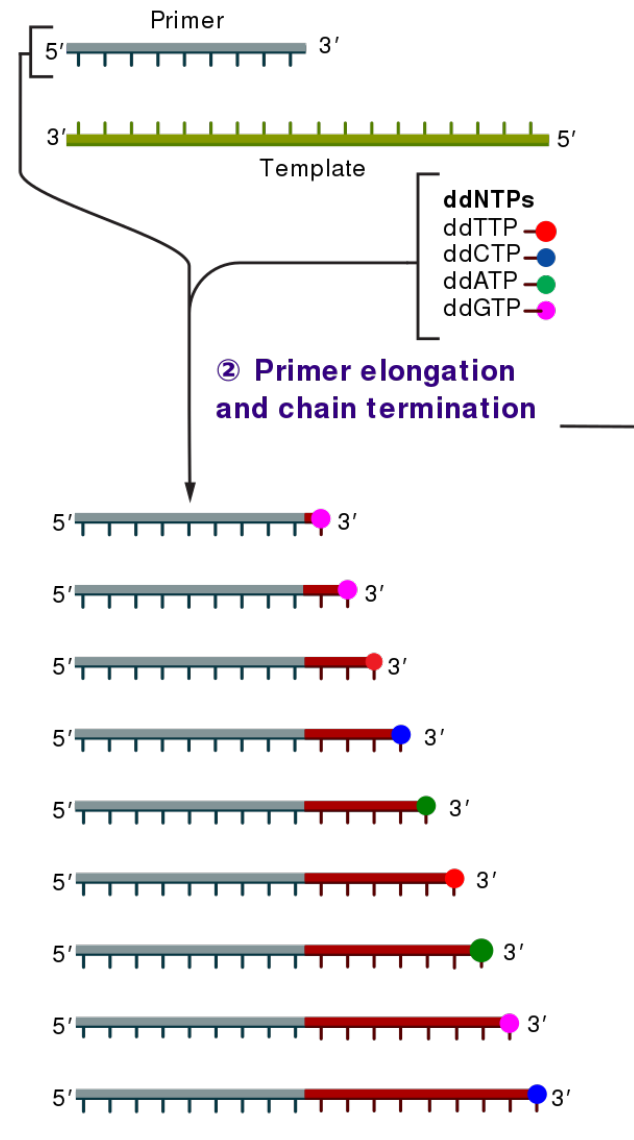
Approx. 650 bp in 5' region of cytochrome c oxidase subunit I (COI)

```
>MN604318.1 Oncorhynchus keta cytochrome c oxidase subunit I gene, complete cds; mitochondrial
GTGGCAATCACACGATGATTCTTCTCAACCAACCACAAAGACATTGGCACCTCTATTTAGTATTTGGTGCCTGAGCCGGGATAGTAGGCACCGCCCTG
AGCCTACTAATTCGGGCAGAACTAAGCCAGCCAGGCGCTCTTCTAGGGGATGACCAGATCTACAATGTAATCGTTACAGCCCATGCCTTCGTTATAATT
TTCTTTATAGTCATACCAATTATAATCGGAGGCTTTGGAACTGATTAATCCCCCTAATGATCGGGGCACCAGATATAGCATTCCCACGAATAAATAAC
ATAAGCTTCTGACTCCTACCTCCGTCCTTCCTCCTCCTCTTCTTCATCTGGAGTTGAAGCCGGCGCTGGTACCGGGTGGACAGTTTATCCCCCTCTA
GCCGGAACCTTGCCACGCAGGAGCATCTGTCGACTTAACCATCTTCTCCCTCCATTTAGCTGGAATCTCCTCAATTTTGGGGGCCATTAATTTTATT
ACGACCATTATCAACATAAAACCCCCAGCTATTTCTCAGTACCAAACCCCGCTTTTTGTCTGAGCTGTACTAATCACTGCTGTACTTCTACTATTATCA
CTCCCCGTTCTGGCAGCAGGTATTACTATGTTGCTCACAGATCGAAATTTAAACACCACTTTCTTTGACCCGGCGGGTGGCGGAGATCCAATTTTATAC
CAACACCTCTTTTGATTCTTCGGTCACCCAGAGGTCTATATTCTGATCCTCCCAGGCTTTGGTATAATTTACATATCGTTGCATATTACTCTGGTAAG
AAAGAACCTTTTCGGGTACATAGGAATAGTGTGAGCTATAATAGCCATCGGCTTGTTAGGATTTATCGTTTGAGCCCACCACATATTTACTGTGCGGATG
GACGTGGACACTCGTGCCTACTTTACATCTGCCACCATAATTATCGCTATCCCCACAGGAGTAAAAGTATTTAGCTGACTAGCTACACTGCACGGAGGC
TCGATCAAATGAGAGACACCACTTCTCTGAGCCCTAGGATTTATCTTCCTATTTACAGTGGGCGGATTAACGGGCATCGTCCTTGCTAACTCCTCATT
GACATTGTTTTACATGACACTTATTACGTAGTCGCCCATTTCCTACTACGTACTCTCAATAGGAGCTGTATTTGCCATTATGGGCGCTTTCGTACACTGA
TTCCCCCTATTACAGGGTACACCCTTCACAGCACATGAACCAAATCCATTTTGAATTATATTTATCGGTGTAAATTTAACCTTTTTCCACAGCAT
TTCCTAGGCCTCGCAGGGATACCACGACGGTACTCTGACTACCCGGACGCCTACACGCTATGAAACACTGTATCCTCAATCGGATCCCTTGTCTCCTTA
GTAGCTGTAATTATGTTCTATTTATTCTTTGAGAGGCTTTTGCTGCCAAACGAGAAGTAGCATCAATCGAAATAACTTCAACAAACGTAGAATGACTA
CACGGATGCCCCCACCCTACCACACATTCGAGGAACCAGCATTTGTCCAAGTACGAACGTACTAA
```

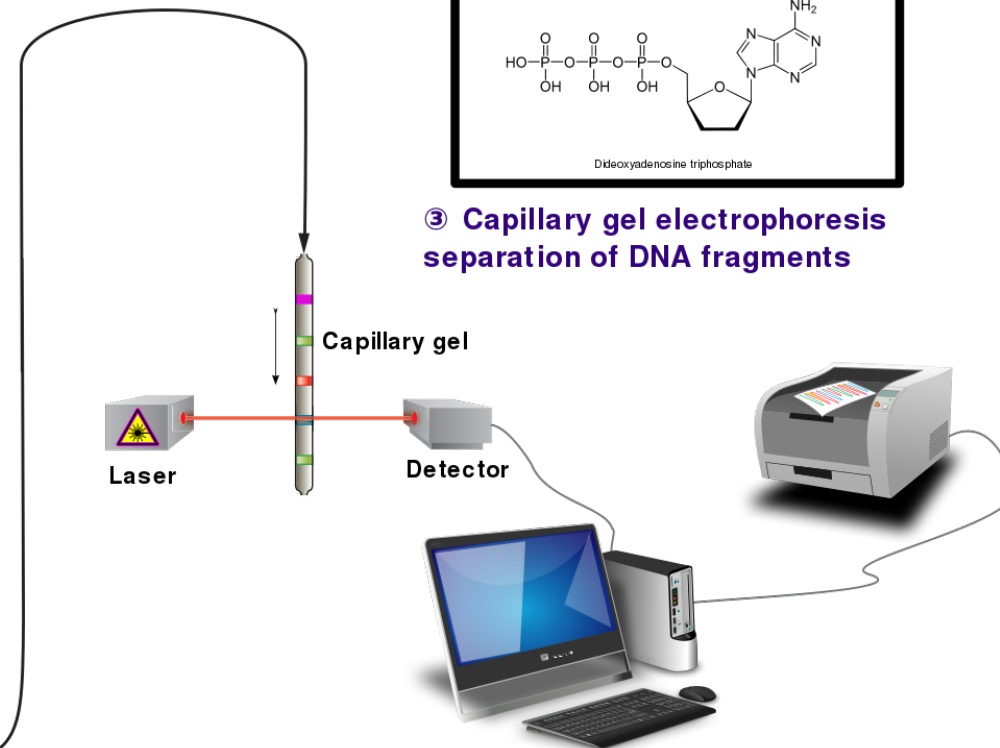
# Sanger sequencing

## ① Reaction mixture

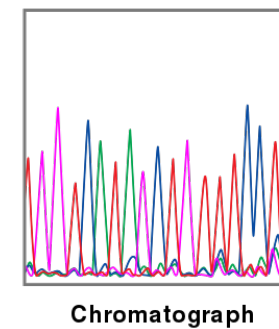
- ▶ Primer and DNA template ▶ DNA polymerase
- ▶ ddNTPs with flourochromes▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



## ③ Capillary gel electrophoresis separation of DNA fragments



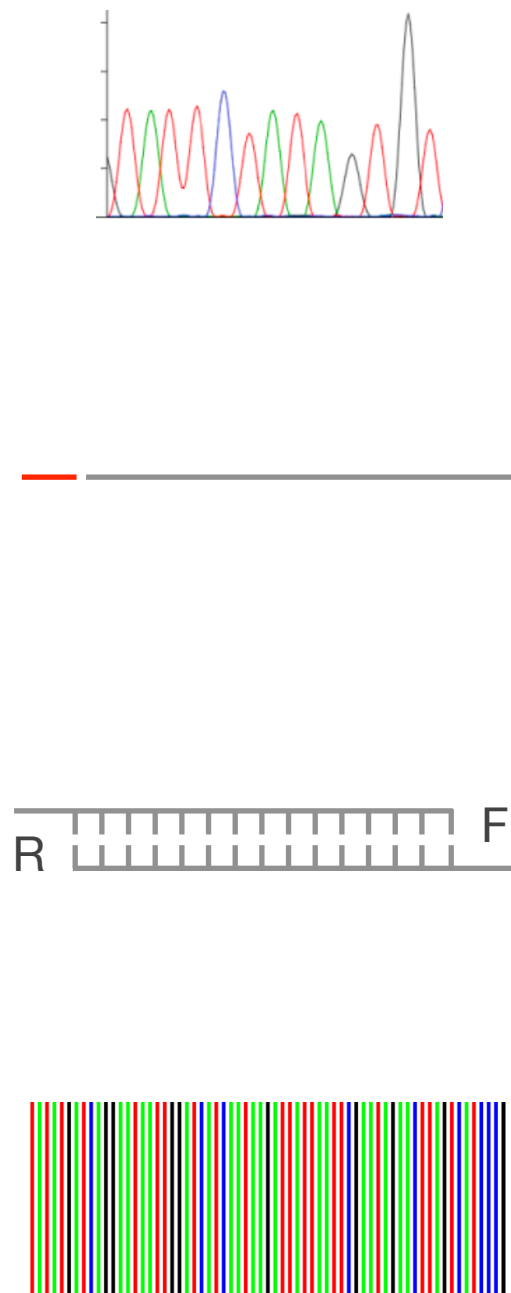
## ④ Laser detection of flourochromes and computational sequence analysis



Estevezj, CC BY-SA 3.0

# Sanger read processing

## Exercise 1



Trace file (F + R)



Fasta file (F+R)



Alignment



Consensus

Basecalling

Trimming

Reverse complement R  
Align F and R

Create consensus sequence

# 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)



# Matching sequences to 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
✓	<a href="#">Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M08F</a>	<a href="#">Oncorhynchus keta</a>	1029	1029	100%	0.0	99.64%	772	<a href="#">LC094471.1</a>
✓	<a href="#">Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M01F</a>	<a href="#">Oncorhynchus keta</a>	1029	1029	100%	0.0	99.64%	772	<a href="#">LC094464.1</a>
✓	<a href="#">Oncorhynchus keta isolate 10_Narva cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	<a href="#">Oncorhynchus keta</a>	1029	1029	100%	0.0	99.64%	655	<a href="#">KR778851.1</a>

...

# Genetic distances

## Exercise 2

```
AAGCCAGCCAGGCGCTCTTCTAGGGGATGACCAGATCTACAATGTAATCG    # 50 positions total
AAGTCAACCTGGTGCACCTTCTTGGTGATGATCAAATTTATAATGTGATCG
***.**.*.*** **.*.*** **.*.*** **.*.*** **.*.*** **.*.*** # 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)

$$8 / 50 = 0.16$$

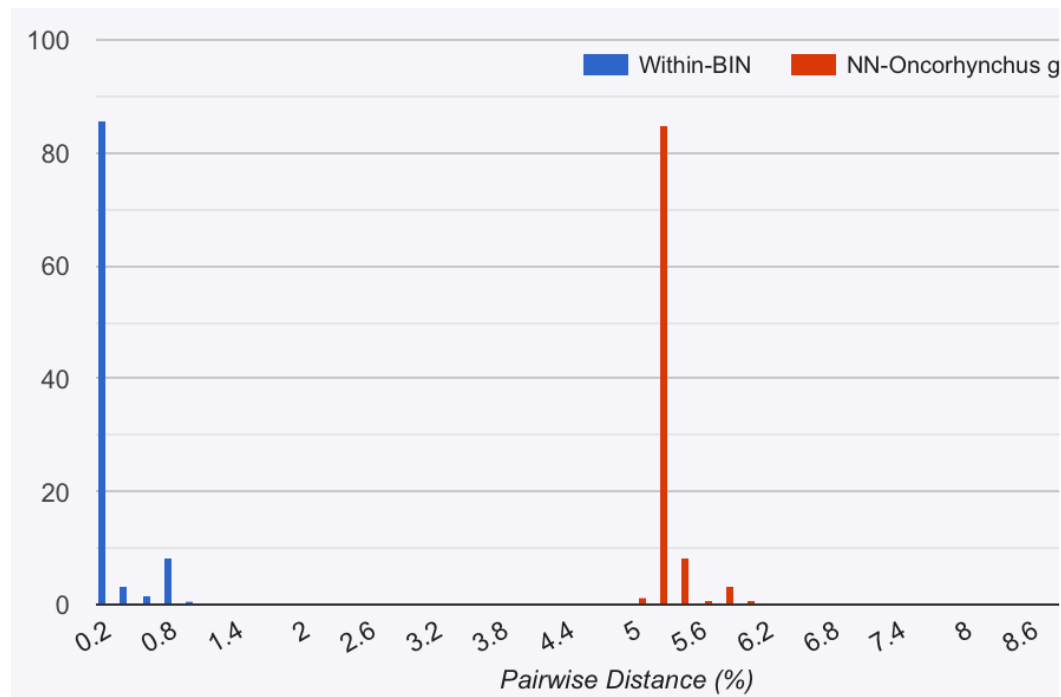
$q$ : proportion of transversions

$$5 / 50 = 0.1$$

# Barcode gap

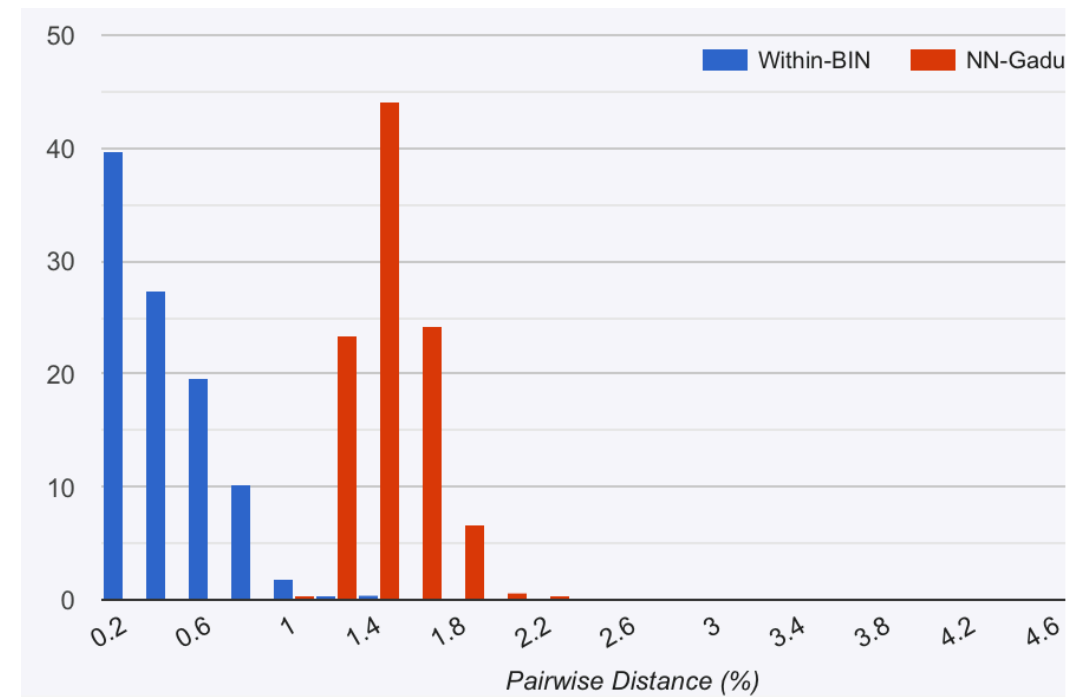
## Exercise 2

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



Large barcode gap

*Oncorhynchus keta*

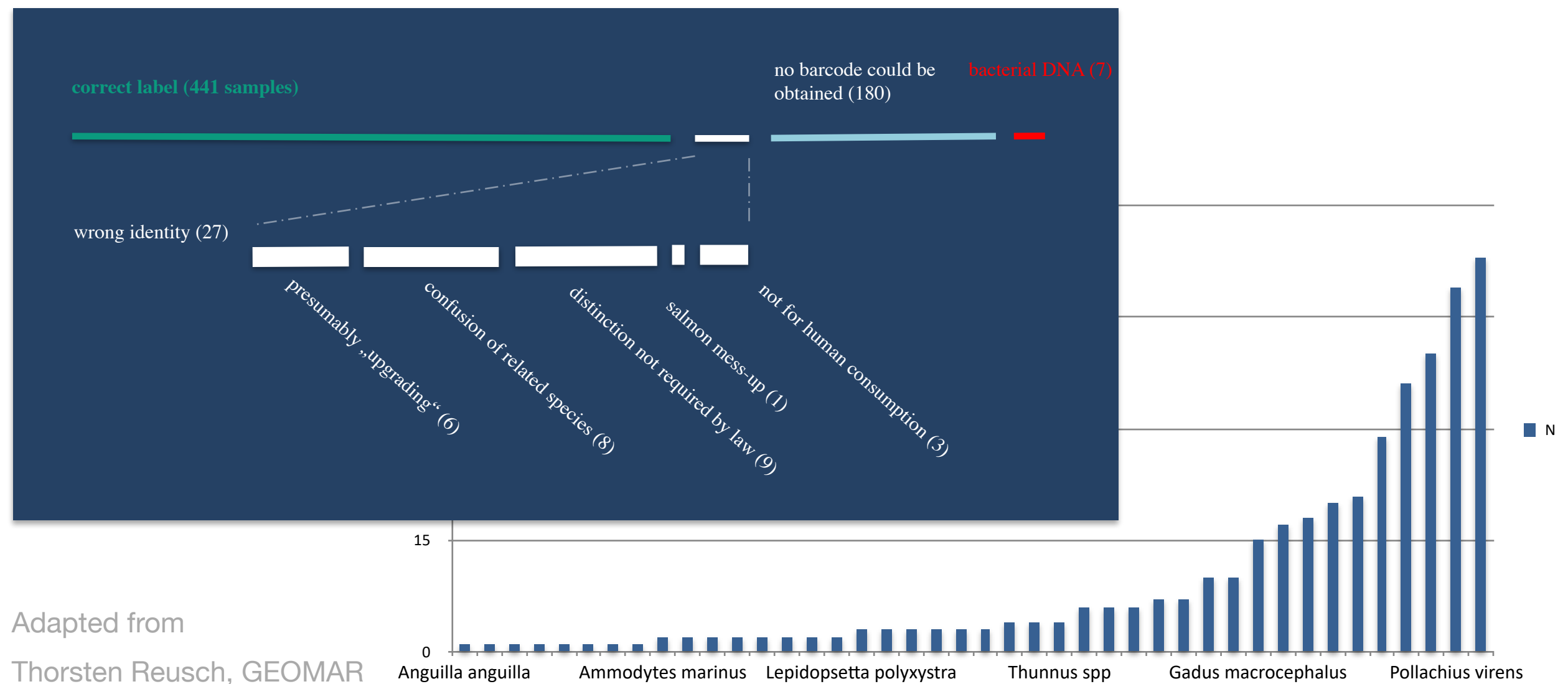


Small barcode gap

*Gadus chalcogrammus*

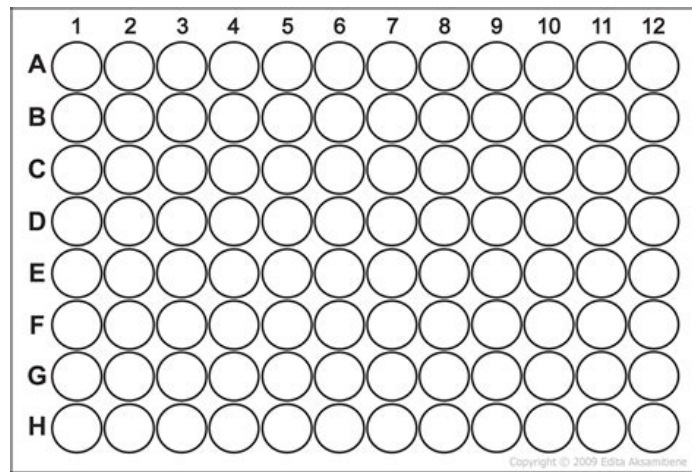
# #fischdetektive results

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)



# Portable 3<sup>rd</sup> gen sequencing

Oxford Nanotechnologies MinION



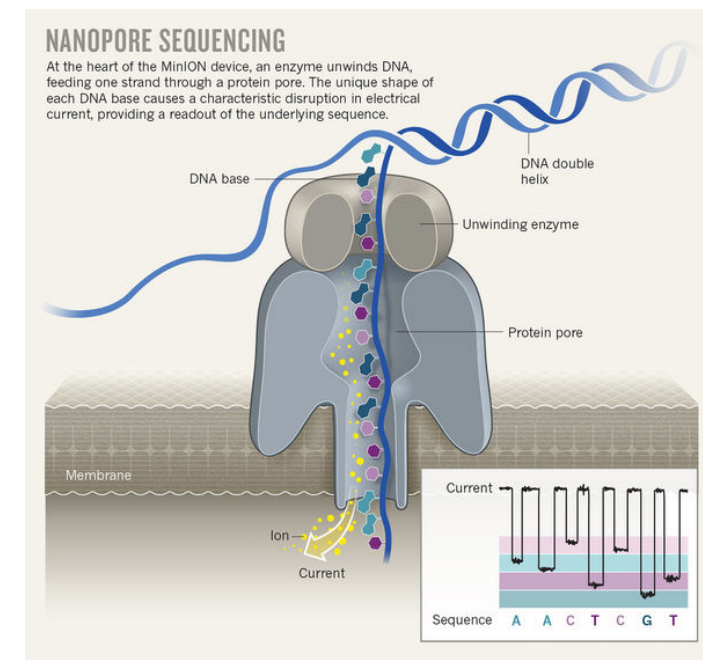
96-well plate



Indexing during PCR  
Library preparation



[whatech.com](http://whatech.com)



[blogs.nature](http://blogs.nature)

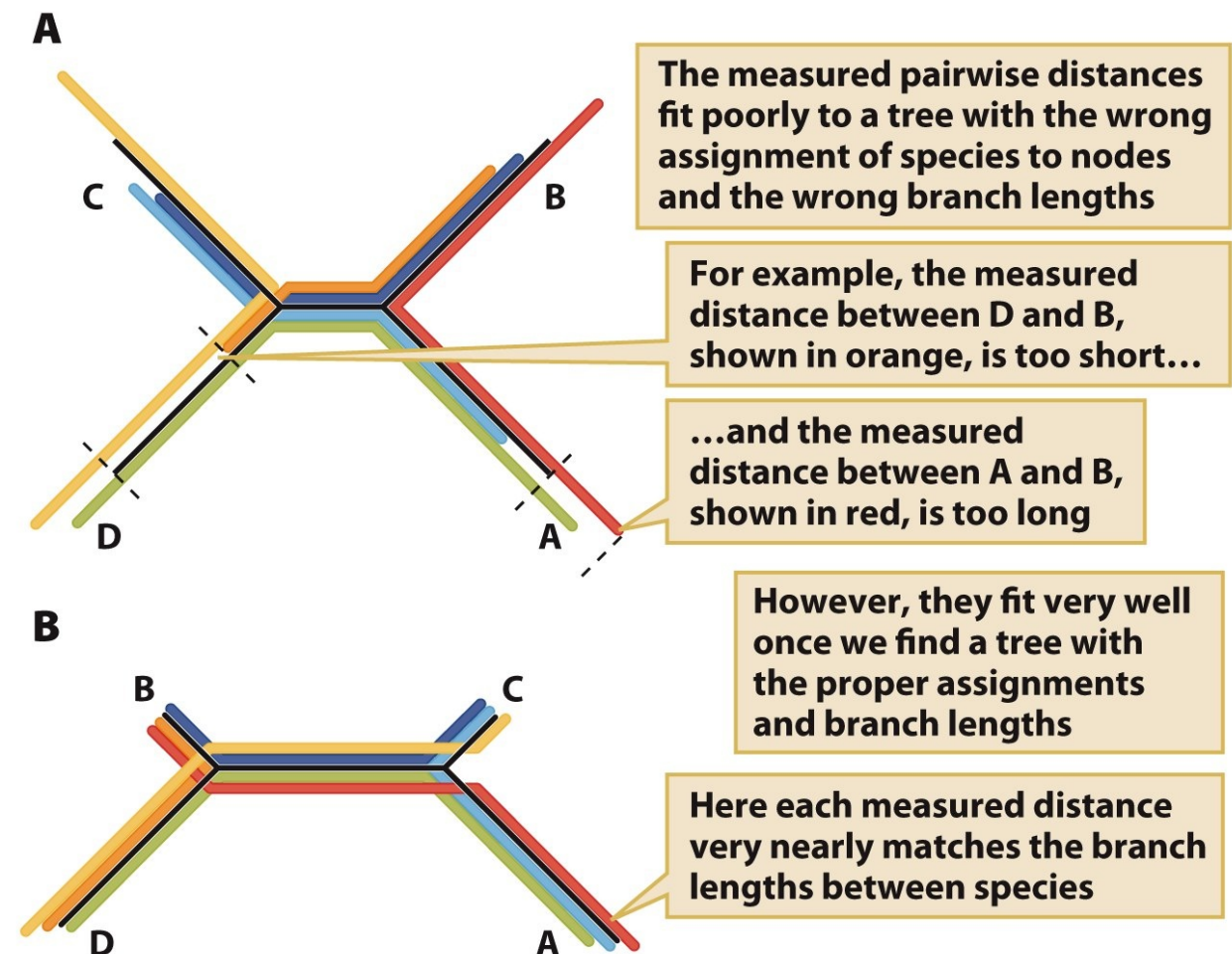
# Course outline

May be subject to change

Class	Date	Topics	Script
01	Apr 14	Introduction, software installation	01_intro.R
02	Apr 21	Hardy-Weinberg equilibrium	02_hwe.R
03	Apr 28	Genetic drift and effective population size	03_drift.R
04	May 05	Population structure and gene flow	04_structure.R
05	May 12	Isolation by distance (lecture online, exercises in person)	05_ibd.R
–	May 19	<a href="#">Himmelfahrt break</a>	–
06	May 26	Genome sequencing and assembly	06_genseq.sh
07	Jun 02	Genotyping, SNPs and population genomics	07_snps.sh
08	Jun 09	Recombination and linkage disequilibrium	08_recomb.R
–	Jun 16	<a href="#">Student presentations</a>	–
09	Jun 23	Selection and mutation	09_sel.R
10	Jun 30	DNA barcoding	10_barcode.sh
11	Jul 07	Metabarcoding I	
12	Jul 14	Metabarcoding II	

# Distance-based phylogeny

- Calculate pairwise distances
- Compile distance matrix
- Find tree that minimizes total distance



Bergstrom and Dugatkin, Evolution (Norton & Co., 2012)