

Sommersemester 2025

Vertiefte Bestimmungsübungen an Tieren (MEES003/C3)

## **Environmental DNA (eDNA) Metabarcoding Analysis**

**Day 1, 2**

Shumpei Yamakawa



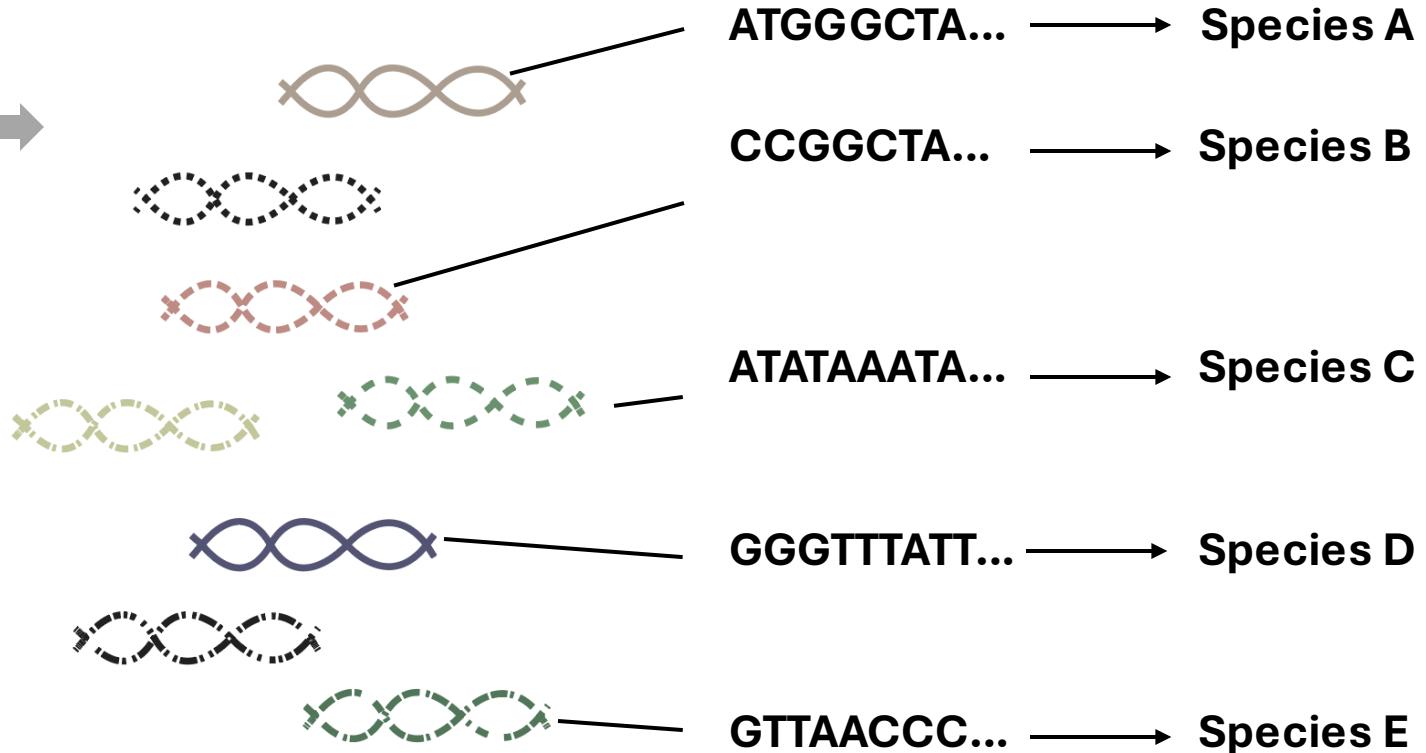
**Day 1, 2:**  
**eDNA Metabarcoding for identifying**  
**animals living in a pond near Halle**





## 2. Sequencing

## 3. Annotation

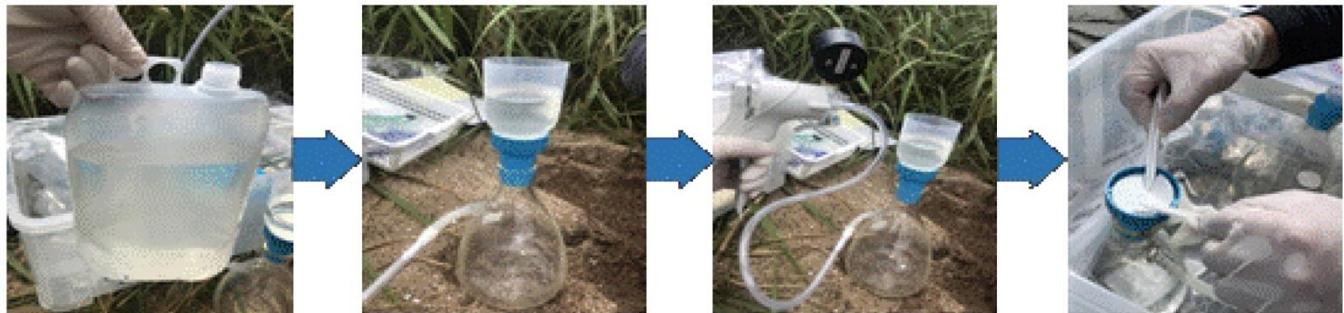


# 1. Sampling



**500 mL sample**

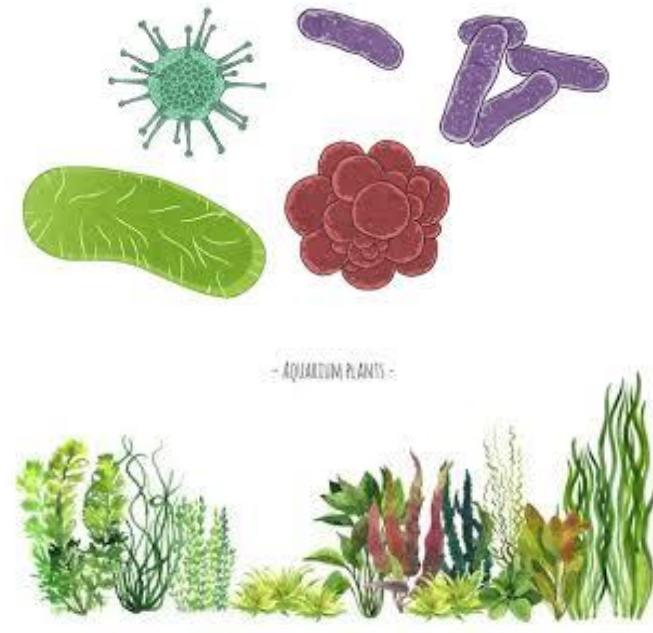
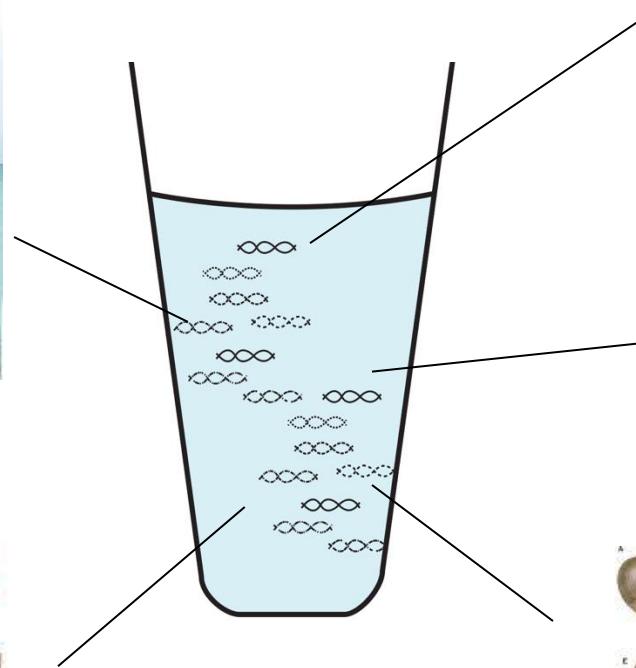
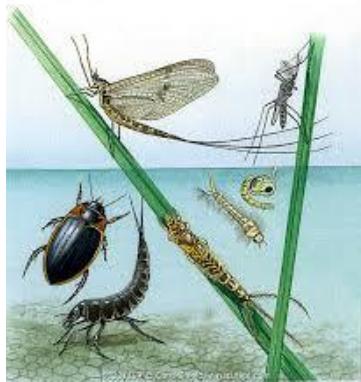
Field  
Sampling



Kim et al., 2019

# 1. Sampling

The sample contains DNA from various organisms



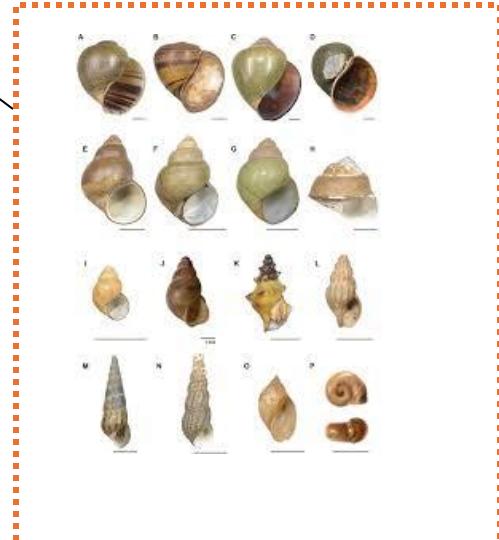
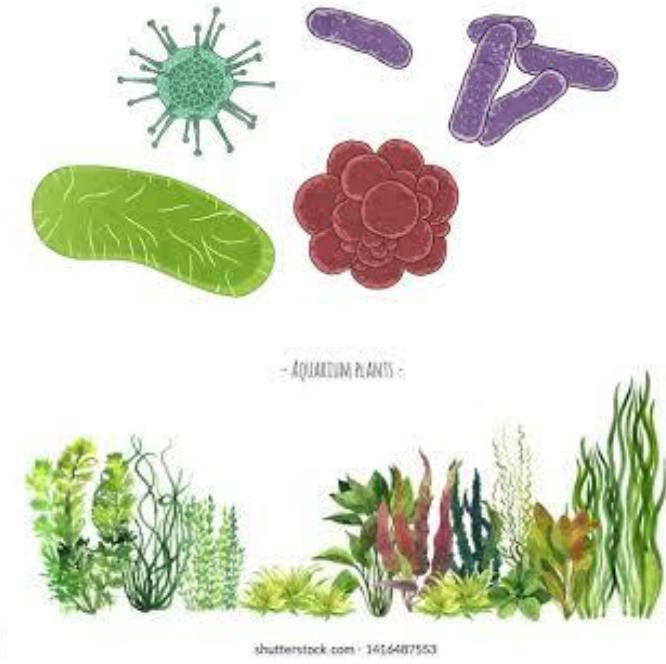
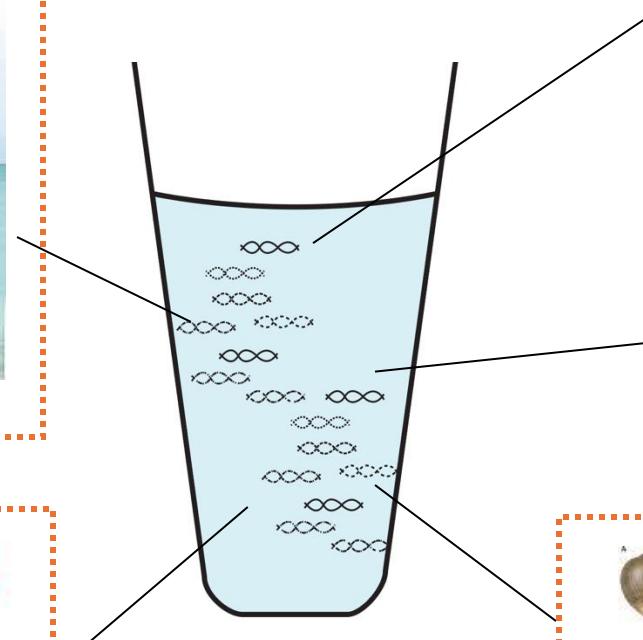
- AQUARIUM PLANTS -

shutterstock.com - 1416487553



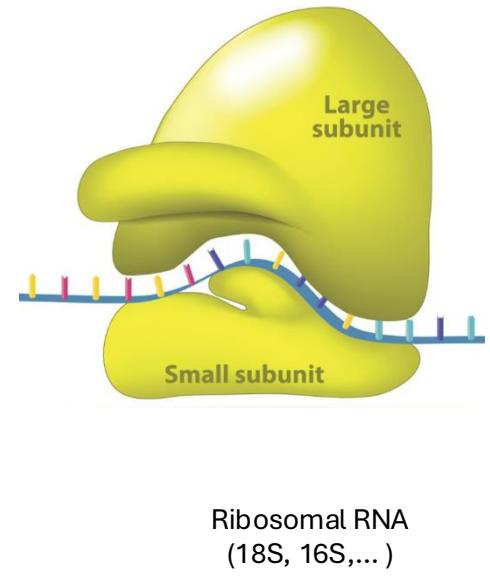
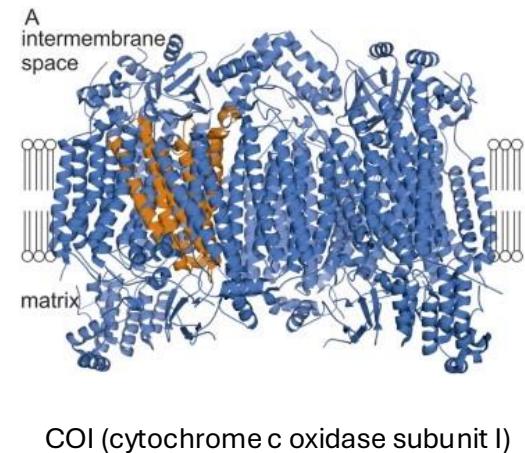
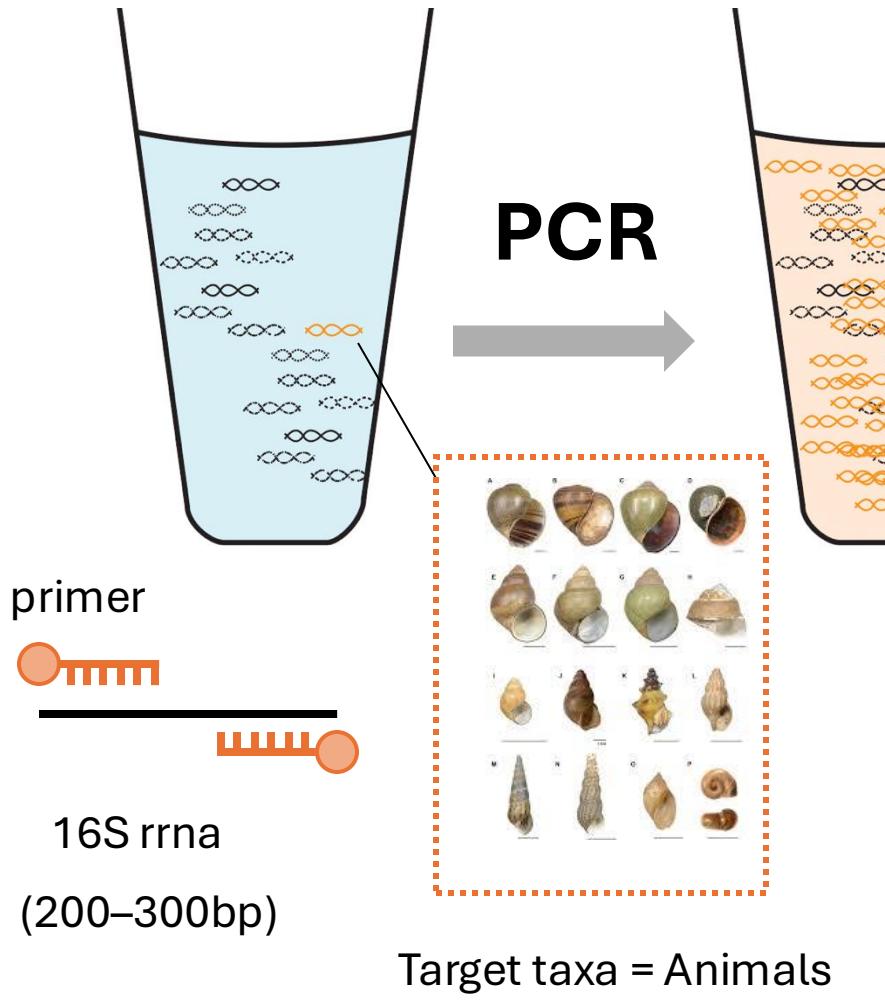
# 1. Sampling

The sample contains DNA from various organisms



# 1. Sampling

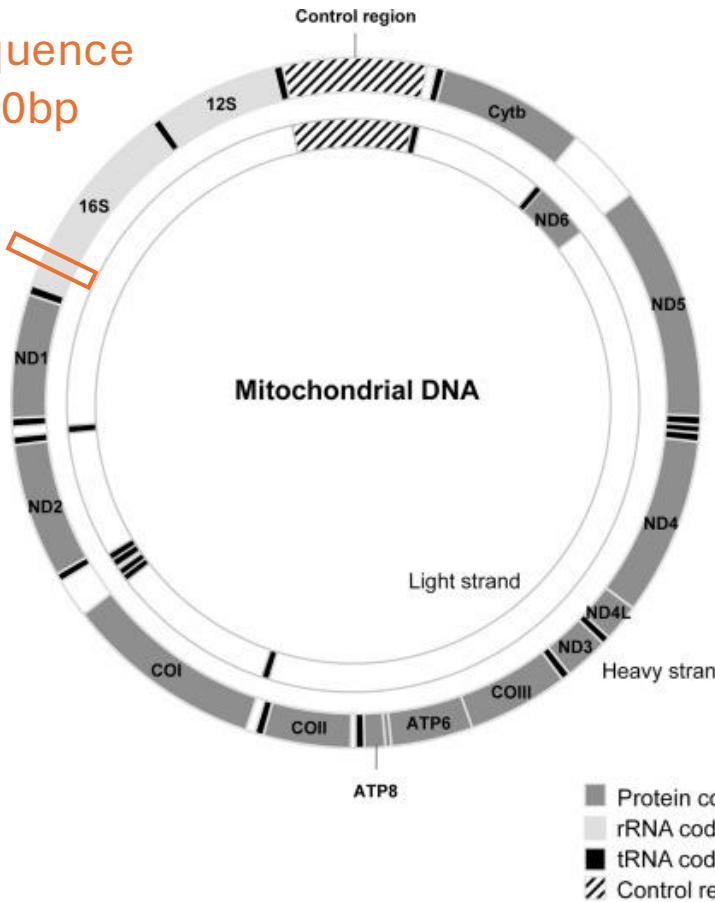
Amplification of specific sequences



# 1. Sampling

## Universal primers for animals

Target sequence  
200–300bp



PLOS One

Publish About Browse Search advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Katy E. Klymus, Nathaniel T. Marshall, Carol A. Stepien

Published: May 18, 2017 • <https://doi.org/10.1371/journal.pone.0177643>

463 Save 148 Citation

17,607 View 16 Share

Download PDF

Print Share

Check for updates

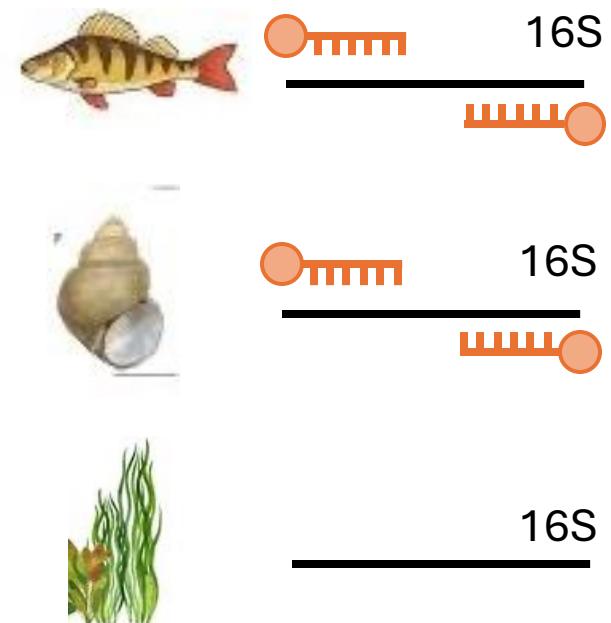
### Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes

Katy E. Klymus, Nathaniel T. Marshall, Carol A. Stepien

Published: May 18, 2017 • <https://doi.org/10.1371/journal.pone.0177643>

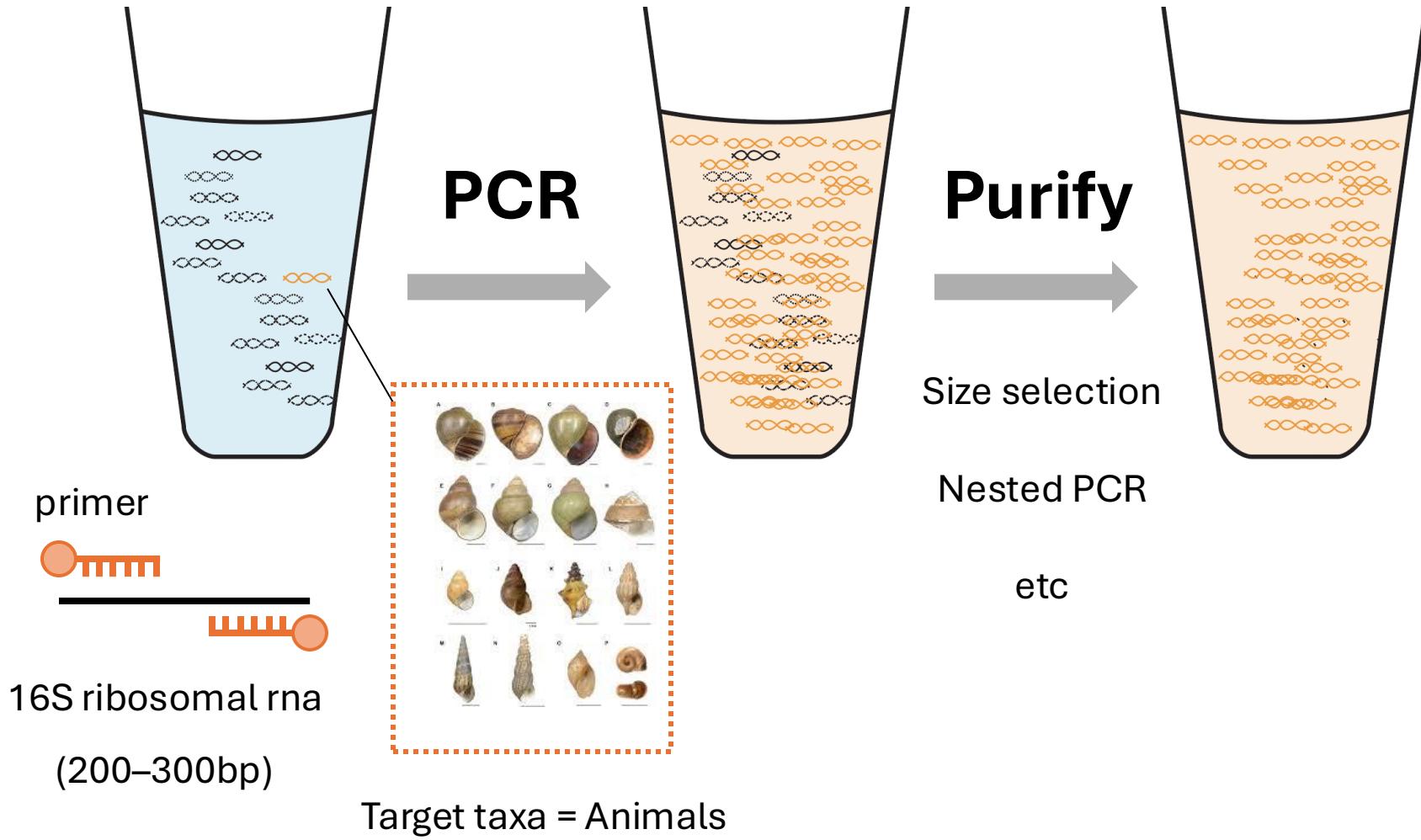
Article	Authors	Metrics	Comments	Media Coverage
Abstract	Introduction Materials and methods Results Discussion Conclusions Supporting information	Abstract	Describing and monitoring biodiversity comprise integral parts of ecosystem management. Recent research coupling metabarcoding and environmental DNA (eDNA) demonstrate that these methods can serve as important tools for surveying biodiversity, while significantly decreasing the time, expense and resources spent on traditional survey methods. The literature emphasizes the importance of genetic marker development, as the markers dictate the applicability, sensitivity and resolution ability of an eDNA assay. The present study developed two metabarcoding eDNA assays using the mtDNA 16S RNA gene with Illumina MiSeq platform to detect invertebrate fauna in the Laurentian Great Lakes and surrounding waterways, with a focus for use on invasive bivalve and gastropod species monitoring. We employed careful	

ADVERTISEMENT



# 1. Sampling

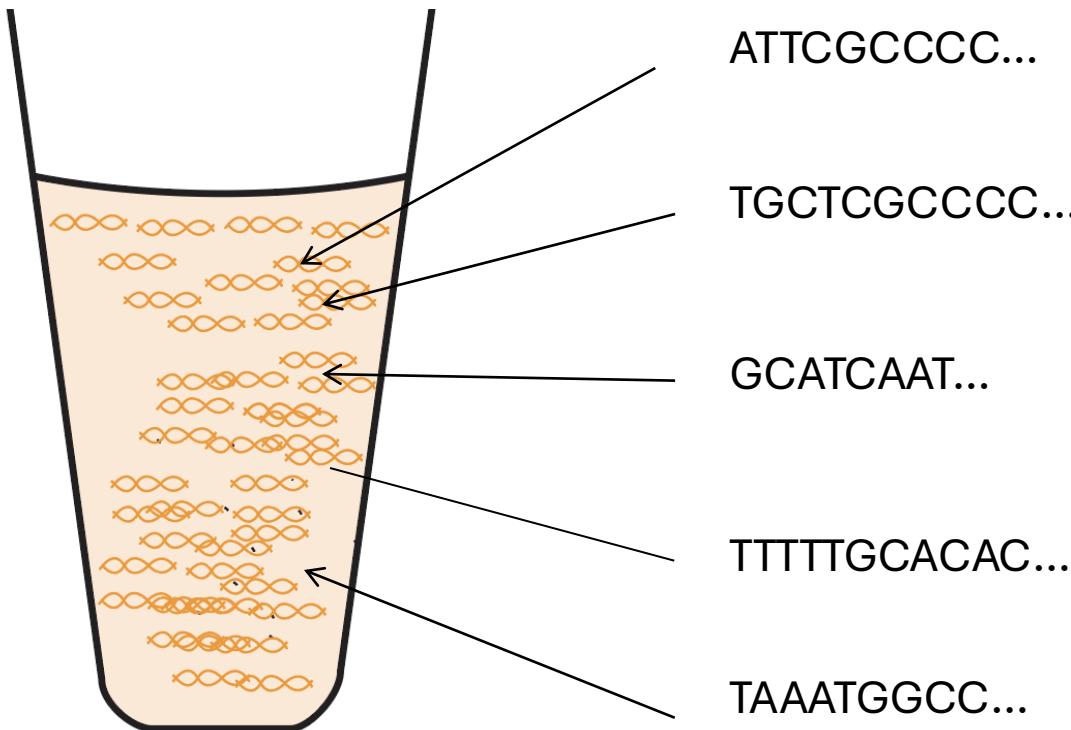
Amplification of specific sequences



## 2. Sequencing

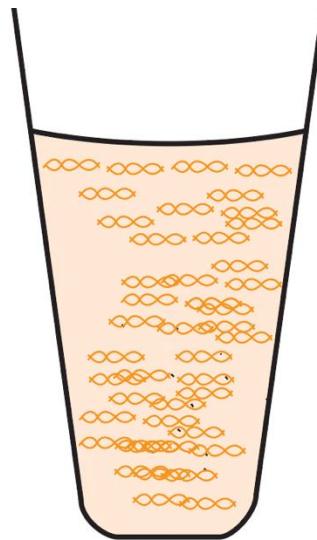
Sequencing of various DNA fragments

# How?



## 2. Sequencing

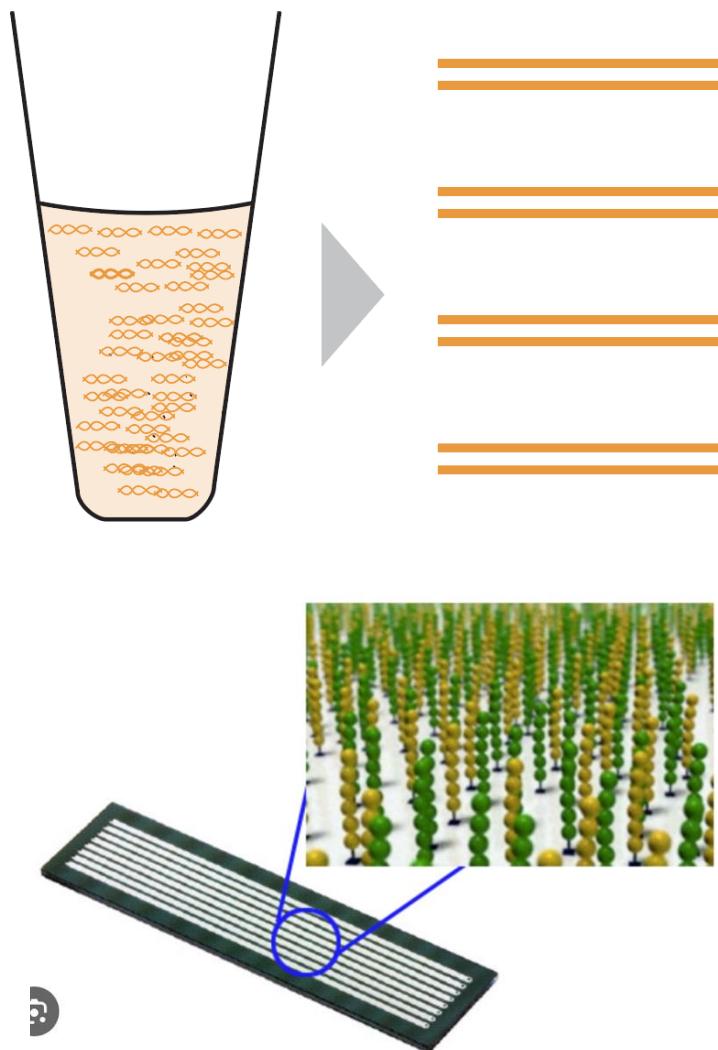
High-throughput sequencing using Next Generation Sequencer



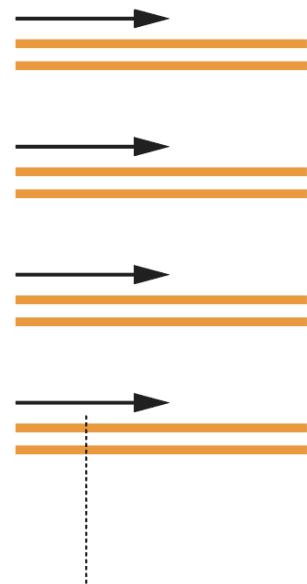
**NGS**



ATTCGCC...  
TGCTCGCC...  
GCATCAAT...  
TTTTGCACAC...  
TAAATGGCC...  
⋮



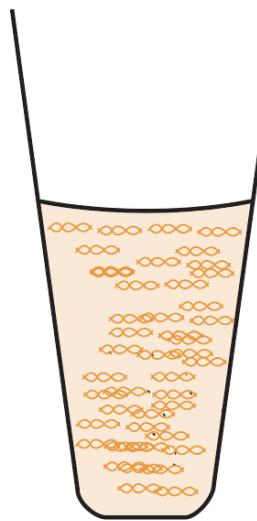
## Sequencing of each fragment



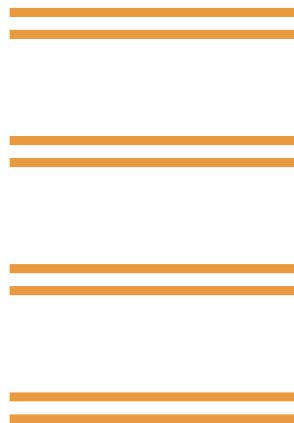
Short read sequencing  
(50–300 bp)

**Read**

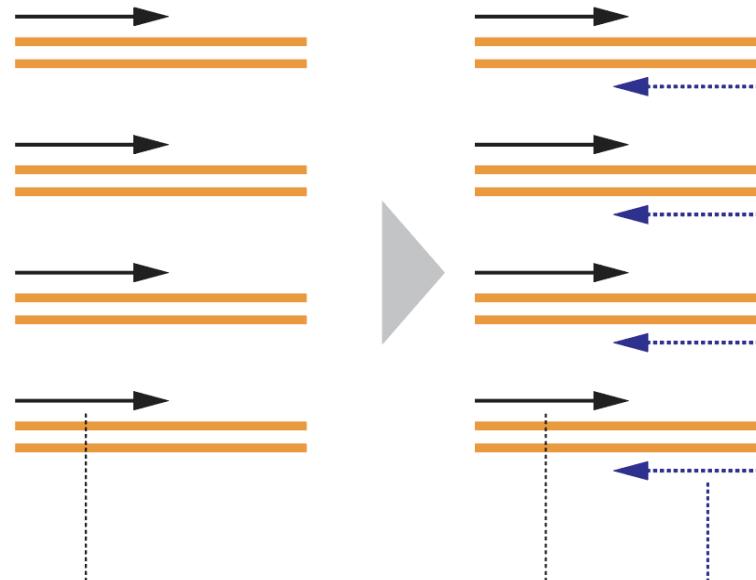
AAAAAATTTTCCCC



## Separation



## Sequencing of each fragment



AAAAAATTTTCCCC

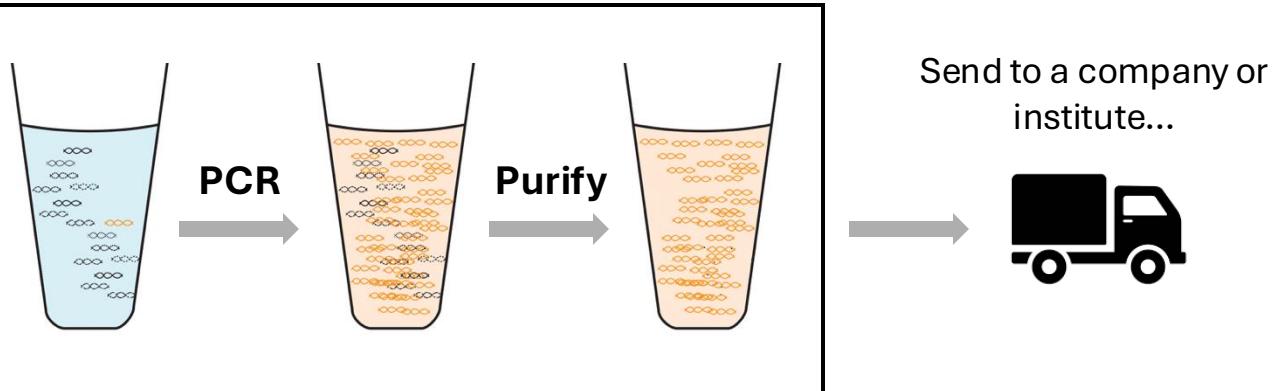
AAAAAATTTTCCCC

TTTTTAAAAGGGG

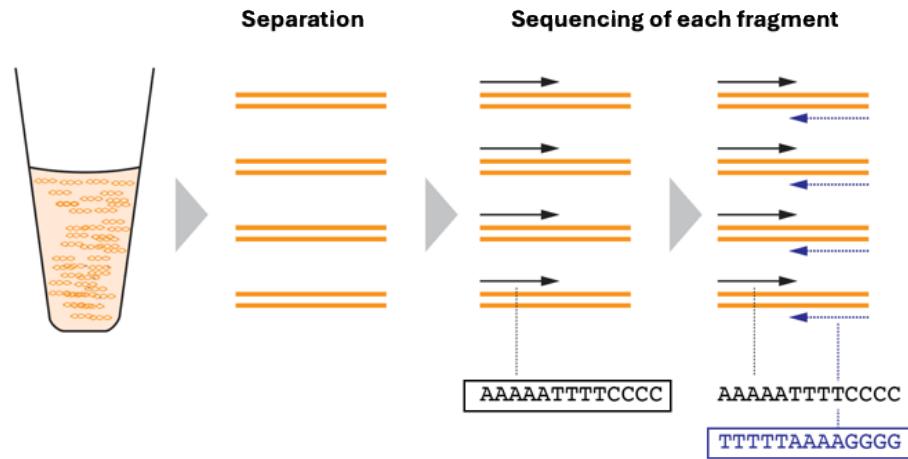
**Forward Read**

**Reverse Read**

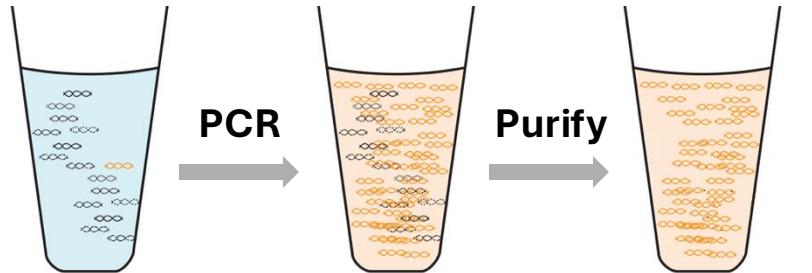
**Paired end sequencing**



## NGS analysis



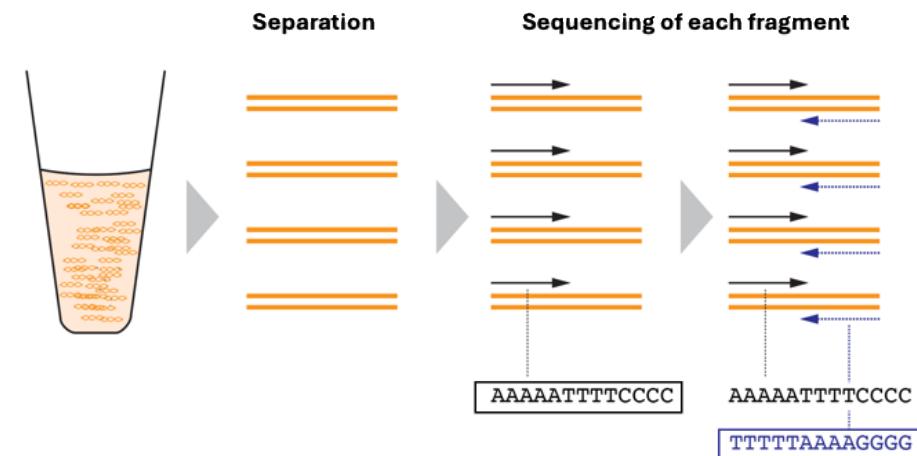
Receive the sequencing data



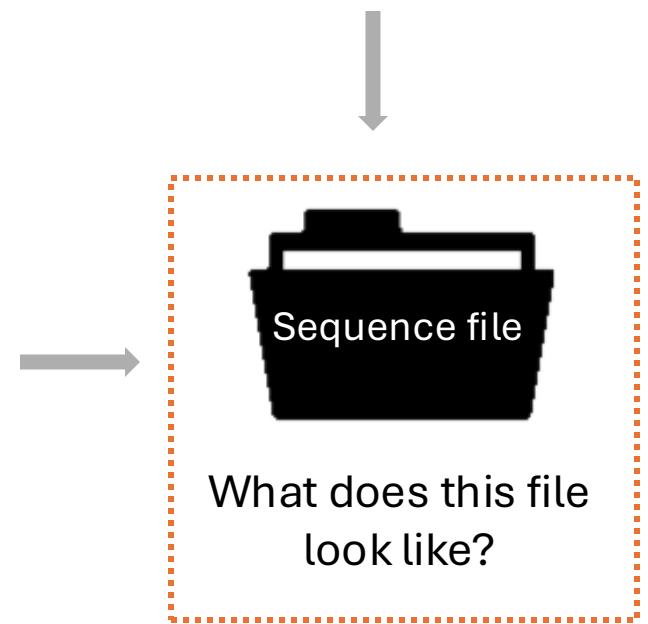
**NGS analysis  
on your laptop!**



## NGS analysis



**NGS analysis  
on your  
laptop!**



# Practice: Viewing Raw Datasets (fasta and fastq data)

## Day1

Download the raw read sequences

```
cd test_meta
mkdir Sample5
#Any names are okay
cd Sample5

wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample5_L001_R1_001.fast
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample5_L001_R2_001.fast
```

Sample5\_L001\_R1\_001.fastq.gz      **Forward Read sequences**

Sample5\_L001\_R2\_001.fastq.gz      **Reverse Read sequences**

## Practice: Viewing Raw Datasets (fasta and fastq data)

### Viewing the raw data

```
gunzip Sample5_L001_R1_001.fastq.gz  
#check the file contents  
head Sample5_L001_R1_001.fastq  
less Sample5_L001_R1_001.fastq
```



## Sample5\_L001\_R1\_001.fastq

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_L001_R1_001.fastq

@M02319:279:000000000-L8DJ7:1:1101:22276:1917 1:N:0:AACCAACG+GCGTAAGA
ACGAGAAGACCCGTGGACTTAATTTATCGTATCTAAACTCTGCCATTAAATTGTATG
GTGCTACTGAGTAAACATATTACTATATTACTTTAGATTATCTATTCTTAACTAT
TTTAGAACACTACTTGGGATAACAGGGTTAGTGTTCGGGTTTCCTATCGATGAAC
AATTACGACCTCGATGTTGGATCGGAAGAGCACACGTCTGACTCCCGTCACAACC
CGATCTCGTCTCCGTCTTGCTATAAAAAACATTCTCTTTTTTTT
+
?CCCCGGGGGGGGFGGGGGGGF,CEFGGF@E,,CE9,<@EEFFEF8FC,,,<CF,C,C,
,,6,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@FE9,6,,5CE5CEE,,+,@=9,,9,:,@,,:CFG>=:7++8A@ED?FGG?A,,,4,8
,4CE9E++@>FGC+@,6@,6,@,@F9D@6++,4,311@C8EF7E:@E**04;*@,41@*,3*4/*45*;*971*)2:>?5C8BEA*)))))//++/,)./.)).)).)).)).)))*1)/(
@M02319:279:000000000-L8DJ7:1:1101:20377:1978 1:N:0:AACCAACG+GCGTAAGA
ACGAGAAGACCCGTGGAGCTTAATTTATCGTAGCTAAACTCTGCCATTAAATTGTATG
GGGACTACTGAGTAAACATATGACTTATGTTACTTTAGATTGATCTATTCTTAACTAT
TTGAGAACACTTGGGATAACAGGGTAGTGTTCGGGAGTCCTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACCTCAGTCACAAC
CGATCTCGTATGCCGTCTTGCTGAAATAAAAAACTCTTTCTACTTCTTTT
+
CCCCCGGFFGGGGGGFGGGGGGGGFGGGGGEFG8,CFD9<FGFF9FGDGGG,9EFGGFAFE
88:FGGGGG?DCF9FCFFGCGA<FEG,CEFGDFGFFGCC@FG?,C@F9FFAFGGGEG??AE
FF,9<,9EEGGFFFG,,7:,CDFFG9=CF+@A9ECGGGGCEC++@?<FFGGGGGG9D,C,@
,>DGGDFGDGGGGGG, @EA,, @AFGGGGG+3,,5,6@CECEFF,=EFGDC4CF7B@9CEG=
DGF7*9C4C4CCAF3AD@FEFFF@A3).)))7+6A20)+-))))))--)****)
@M02319:279:000000000-L8DJ7:1:1101:24081:1998 1:N:0:AACCAACG+GCGTAAGA
ACGAGAAGACCCGTGGAGCTTAATTTATCGTAGCTAGACTCTACCATTAAATTGTATG
GGGACTACTGAGTAAACAAATGACTTATATTACTTTAGATTGATCTATTCTTAACTAT
TTGAGAACACTTGGGATAACAGGGTAGTGTTCGGGAGTCCTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACCTCAGTCACAAC
CGATCTCGTATGCCGTCTTGCTGACAAAAAAATTCTCTCTTTCT
```

## Fastq format

A standard format for sequence data that includes base calling quality information

## Sample5\_L001\_R1\_001.fastq

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_L001_R1_001.fastq
```

```
@M02319:279:00000000-L8DJ7:1:1101:22276:1917 1:N:0:ACCAACG+  
GGTAAGA
```

```
ACGAGAACCGCTGGAACTTAATTTATCGTATCTAAACTCTGCCATTAAATTGTATG  
GTGTCTACTGAGTAAACATATTACTTATATTACTTTAGATTATCTATTCCCTTAACAT  
TTAGAACATCTACTTGGGGATAACAGGGTTAGTGTTCGGGGTTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGGATCGGAAGAGCACACGTCTGACTCCGTACAACC  
CGATCTCGTCTGCCGTCTGCTGCTATAAAAAACATTCCCTTTTTTTTT
```

```
+
```

```
?CCCCGGGGGGGGFGGGGGGF, CEEFGGF@E,, CE9,<@EEFFEF8FC,,, <CF,C,C,  
,,6,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@  
FE9,6,,,5CE5CEE,,+,@=9,,,9,:@,,:CFFG>=:7++8A@ED?FGG?A,,,4,8  
,4CE9E++@>FGC+@,6@,6,@, @F9D@6++,,4,311@C8EF7E:@E**04;*@,41@*,  
3*4/*45*;*971*2:>?5C8BEA*)))))/++/.)./.)).))).)))*)1)/((
```

```
@M02319:279:00000000-L8DJ7:1:1101:20377:1978 1:N:0:ACCAACG+  
GGTAAGA
```

```
ACGAGAACCGCTGGAGCTTAATTTATCGTAGCTAAACTCTGCCATTAAATTGTATG  
GGGACTACTGAGTAAACATATGACTTATGTTACTTTAGATTGATCTATTCCCTTAACAT  
TTGAGAACGCTACTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAAC  
CGATCTCGTATGCCGTCTGCTGCTAAATAAAACTTCTTTCTACTTCTTTT
```

```
+
```

```
CCCCCGGFFGGGGGGGGGGGGGGGGGGGGFEG8, CFD9<FGFF9FGDG, 9EFGGFAFE  
88:FGGGGG?DCF9FCFFGCGA<FEG, CEFGDFGFFGCC@FG?, C@F9FFAFGGGEG?@AE  
FF, 9<, 9EEGGFFF, , 7:, CDFFG9=CF+@A9ECGGGGCEC++@?<FFGGGGGG9D, C, @  
>DGGDFGDGGGGGG, @EA,, @AFGGGGG+3,, 5, 6@CECEFF, =EFGDC4CF7B@9CEG=  
DGF7*9C4C4CCAF3AD@FEFFFEA3). )) )7+6A20+-)))))--*****)
```

```
@M02319:279:00000000-L8DJ7:1:1101:24081:1998 1:N:0:ACCAACG+  
GGTAAGA
```

```
ACGAGAACCGCTGGAGCTTAATTTATCGTAGCTAGACTCTACCATTAAATTGTATG  
GGGACTACTGAGTAAACAAATGACTTATATTACTTTAGATTGATCTATTCCCTTAATTAT  
TTGAGAACGCTACTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAAC  
CGATCTCGTATGCCGTCTGCTGCTGCTAAATAAAACTTCTCTCTTTCTTT
```

≡ sequence id

(ex. read1)

≡ sequence id

(ex. read2)

≡ sequence id

(ex. read3)

# Sample5\_L001\_R1\_001.fastq

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_L001_R1_001.fastq
```

```
@M02319:279:00000000-L8DJ7:1:1101:22276:1917 1:N:0:ACCAACG+  
GGTAAGA
```

≡ sequence id (ex. read1)

```
ACGAGAACCTGTGGACTTAATTTATCGTATCTAAACTCTGCCATTAAATTGTATG  
GTGTCTACTGAGTAAACATATTACTTATATTACTTTAGATTATCTATTCCCTTAACAT  
TTAGAACATCTACTTGGGGATAACAGGGTTAGTGTTCGGGGTTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGATCGGAAGAGCACACGTCTGACTCCGTACAACC  
CGATCTCGTCTGCCGTCTGCTGCTATAAAAAACATTCCCTTTTTTTT  
+
```

sequence information

```
?CCCCGGGGGGGGFGGGGGGF, CEEFGGF@E,, CE9,<@EEFFEF8FC,,, <CF,C,C,  
,,6,,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@  
FE9,6,,,5CE5CEE,,+,@=9,,,9,:@,,:CFFG>=:7++8A@ED?FGG?A,,,4,8  
,4CE9E++@>FGC+@,6@,6,@, @F9D@6++,4,311@C8EF7E:@E**04;*@,41@*,  
3*4/*45*;*971*2:>?5C8BEA*)))))/++/.)./.)).))).)))*)1)/((
```

```
@M02319:279:00000000-L8DJ7:1:1101:20377:1978 1:N:0:ACCAACG+  
GGTAAGA
```

≡ sequence id (ex. read2)

```
ACGAGAACCCCGTGGAGCTTAATTTATCGTAGCTAAACTCTGCCATTAAATTGTATG  
GGGACTACTGAGTAAACATATGACTTATGTTACTTTAGATTGATCTATTCCCTTAACAT  
TTGAGAACCTACTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAAC  
CGATCTCGTATGCCGTCTCTGCTGAAATAAAAAACTTCTTTCTACTTCTTTT  
+
```

sequence information

```
CCCCCGGFFGGGGFGGGGGGGGFGGGGGEFG8, CFD9<FGFF9FGDGGG, 9EFGGFAFE  
88:FGGGGG?DCF9FCFFGCGA<FEG, CEFGDFGFFGCC@FG?, C@F9FFAFGGGEG?@AE  
FF, 9<, 9EEGGFFF, , 7:, CDFFG9=CF+@A9ECGGGGCEC++@?<FFGGGGGG9D, C, @  
>DGGDFGDGGGGGG, @EA,, @AFGGGGG+3,, 5, 6@CECEFF, =EFGDC4CF7B@9CEG=  
DGF7*9C4C4CCAF3AD@FEFFFEA3). )) )7+6A20+-))))) )-- **** )
```

```
@M02319:279:00000000-L8DJ7:1:1101:24081:1998 1:N:0:ACCAACG+  
GGTAAGA
```

≡ sequence id (ex. read3)

```
ACGAGAACCCCGTGGAGCTTAATTTATCGTAGACTCTACCATTAAATTGTATG  
GGGACTACTGAGTAAACAAATGACTTATATTACTTTAGATTGATCTATTCCCTTAATTAT  
TTGAGAACCTACTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAAC  
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAAC  
CGATCTCGTATGCCGTCTCTGCTGCTGAAATAAAAAACTTCTCTCTTTCTCTT
```

sequence information

# Sample5\_L001\_R1\_001.fastq

## Sequence

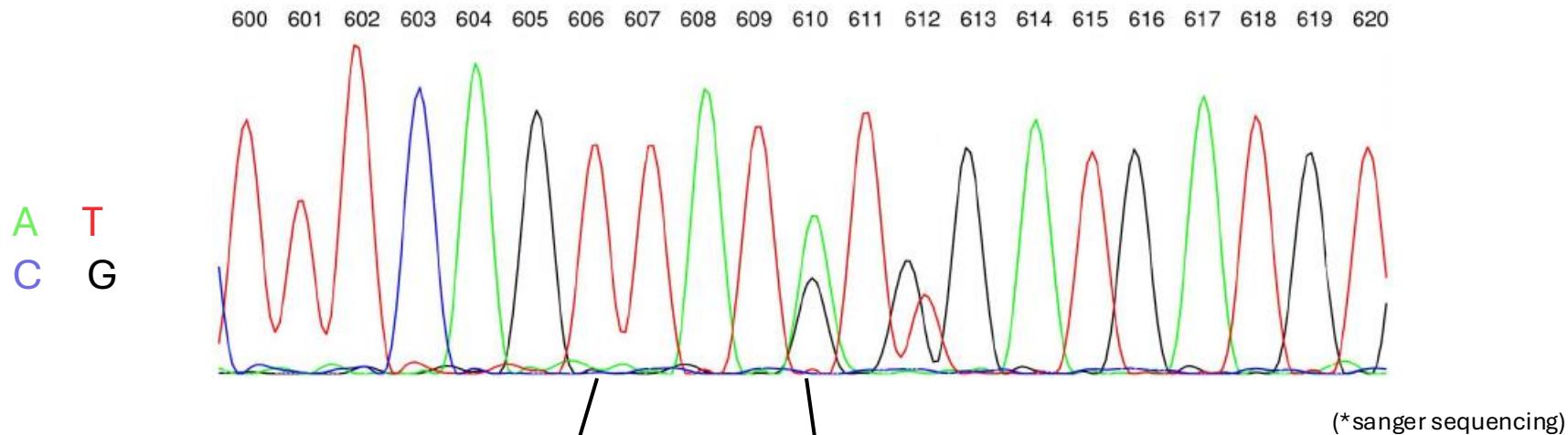
```
@M02319:279:00000000-L8DJ7:1:1101:22276:1917 1:N:0: AACCAACG+  
GCGTAAGA
```

```
ACGAGAAGACCCTGTGGAACTTAATTTATCGTATCTAAACTCTGCCATTAAATTGTATG  
GTGTCTACTGAGTAAACATATTACTTATATTACTTTAGATTATCTATTCCCTTAACATAT  
TTTAGAACATCTACTTGGGGATAAACAGGGTTAGTGTTCGGGTTCTATCGATGAACTC  
AATTACGACCTCGATGTTGATCGGAAGAGCACACGTCTGTACTCCGTACAACCAT  
CGATCTCGTCTGCCGTCTGCTATAAAAAACATTCCCTTTTTTTT
```

```
+
```

```
?CCCCGGGGGGGGFGGGGGGGF, CEEFGGF@E,, CE9,<@EEFFEF8FC,,, <CF,C,C,  
,,6,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@  
FE9,6,,,5CE5CEE,,,+,@=9,,,9,:,@,,:CFFG>=:7++8A@ED?FGG?A,,,4,8  
,4CE9E++@>FGC+@,6@,6,@,0F9D@6++,,4,311@C8EF7E:@E**04;*@,41@*,  
3*4/*45*;*971*)2:>?5C8BEA*)))))) /++/, ) ./) .))) .))) *1) / ((
```

## Sequence “quality”



Base calling

Probability of error (P)

Quality score (Q)

$$Q = -10 \times \log_{10}(P)$$

	T	A
0.001	0.1	
30	10	

42	*	74	J
43	+	75	K
44	,	76	L
45	-	77	M
46	.	78	N

(+33)

ASCII

?

+

60	<	92	\
61	=	93	]
62	>	94	^
63	?	95	_

## Sample5\_L001\_R1\_001.fastq

```
@M02319:279:00000000-L8DJ7:1:1101:22276:1917 1:N:0:ACCAACG+  
GCGTAAGA
```

```
ACGAGAAGACCCTGTGGAACCTAACATTATCGTATCTAAACTCTGCCATTAAATTGTATG  
GTGTCTACTGAGTAAACATATTACTTATATTACTTTAGATTTATCTATTCCCTTAACATAT  
TTTAGAATCTACTTGGGGATAAACAGGGTTAGTGTTGGGGTTTCCTTATCGATGAACTC  
AATTACGACCTCGATGTTGTTGATCGGAAGAGCACACGTCTGTACTCCGTCACAACCAT  
CGATCTCGTCTGCCGTCTTGCTATAAAAAACATTCCTTTTTTTT
```

```
+
```

```
?CCCCGGGGGGGGFGGGGGGGF, CEEFGGF@E,, CE9,<@EEFFEF8FC,,, <CF,C,C,  
,,6,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@  
FE9,6,,,5CE5CEE,,,+,@=9,,,9,:,:,@,,,:CFG>=:7++8A@ED?FGG?A,,,4,8  
,4CE9E++@>FGC+@,6@,6,@,@F9D@6++,,4,311@C8EF7E:@E**04;*@,41@*,  
3*4/*45*;*971*)2:>?5C8BEA*) )) ) /+/+,) ./) .) ) ) .) ) ) *1) / ((
```

?

High quality

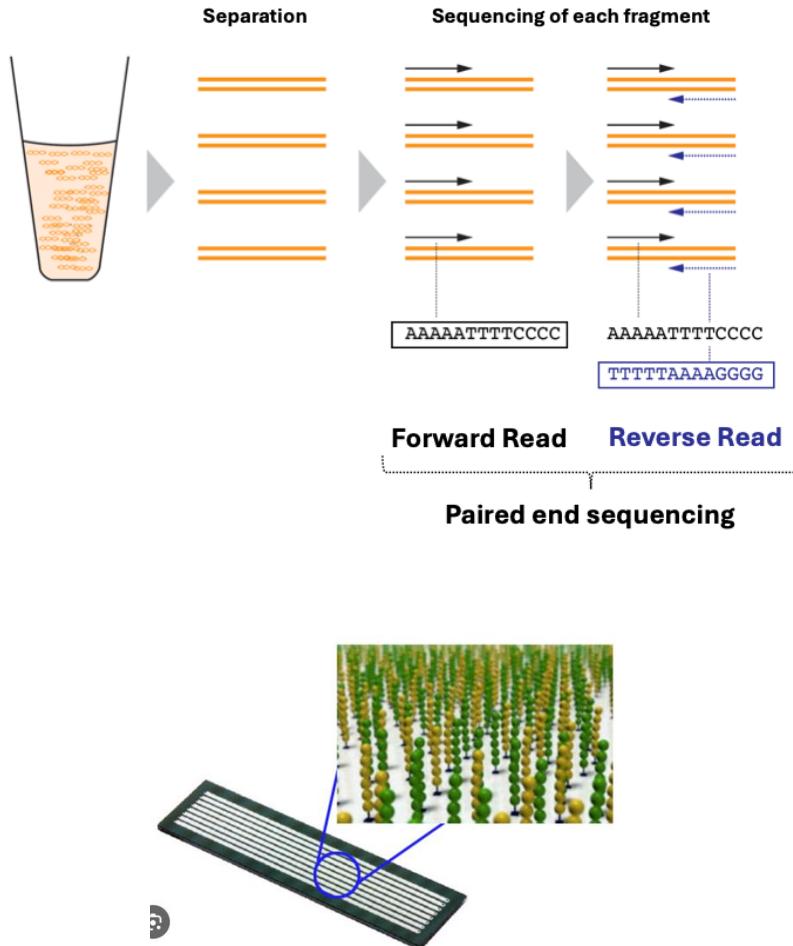
+

Base calling may be wrong...

```
gunzip Sample5_L001_R1_001.fastq.gz
#check the file contents
head Sample5_L001_R1_001.fastq
less Sample5_L001_R1_001.fastq
#less can be terminated by typing "q"

gzip Sample5_L001_R1_001.fastq ←
###Do not forget to compress the fastq file!!!
```

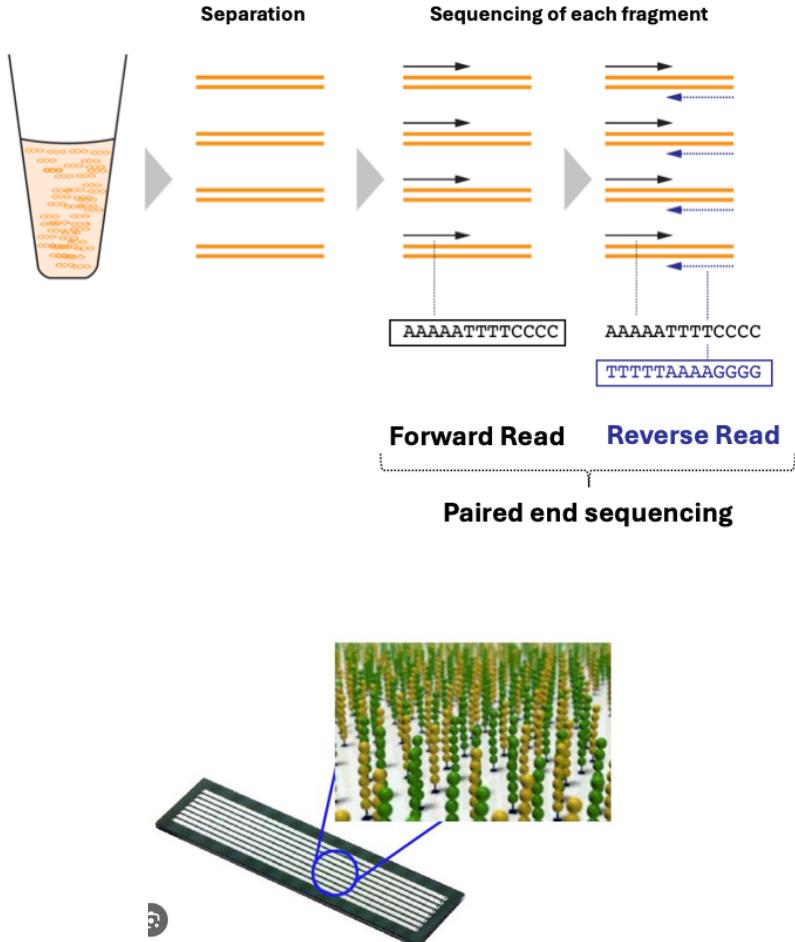
## **Before determining species using the dataset...**



Sample5\_L001\_R1\_001.fastq

**50,685  
reads**

## **Before determining species using the dataset...**



Sample5\_L001\_R1\_001.fastq

**50,685  
reads**



**50,685  
species**

Read 1      ATGCGATCGTA

Read 2      GCGATCGTAAA

Read 3      AAAATCGATCG

Read 4      TGCTAAATATTG

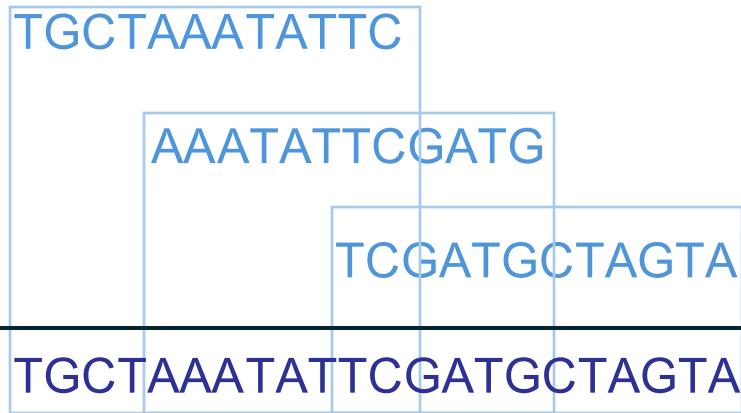
Read 5      AAATATTGATG

Read 6      TCGATGCTAGTA

**6 species?**

### Species A

Read 1	ATGCGATCGTA
Read 2	GCGATCGTAAA
Read 3	AAAATCGATCG
Read 4	TGCTAAATATTG
Read 5	AAATATTGATG
Read 6	TCGATGCTAGTA



### Species B

Checkpoint 1: Multiple reads originate from the same species sequence

Read 1 ATGCGATCGTA **ATG**

Read 2 ATGCGATCGTA **CCC**

Read 3 ATGCGATCGTA **TTT**

Read 4 ATGCGATCGTA **TGA**

Read 5 ATGCGATCGTA **GGG**

Read 6 ATGCGATCGTA **ATG**

**6 species?**

Read 1	ATGCGATCGTA <b>ATG</b>
Read 2	ATGCGATCGTA <b>CCC</b>
Read 3	ATGCGATCGTA <b>TTT</b>
Read 4	ATGCGATCGTA <b>TGA</b>
Read 5	ATGCGATCGTA <b>GGG</b>
Read 6	ATGCGATCGTA <b>ATG</b>

@M02319:279:00000000-L8DJ7:1:1101:22276:1917 1:N:0:AACCAACG+  
 GCGTAAGA  
 ACGAGAAAGACCCCTGTGGAACCTAACATTATCCTATCTAAACTCTGCCATTAAATTGTATG  
 GTGTCTACTGAGTAAACATATTACTTATATTACTTTAGATTATCTATTCTTAACTAT  
 TTTAGAACATCTACTTTGGGGATAACAGGGTTAGTGTTCGGGTTCCCTTATCGATGAACCT  
 AATTACGACCTCGATGTTGTTGATCGGAAGAGCACACGGTCTGTACTCCCGTCACAACCCT  
 CGATCTCGTCTGCCGTCTCTGCTATAAAAAACATTCCCTTTTTTTTT  
 +  
 ?CCCCGGGGGGGGGGFGGGGGGGF, CEEFGGF@E,,CE9,<@EEFFEF8FC,,,<CF,C,C,  
 ,,6,;C9CF9,,C,<,CAF9F,,<CF,C,CC,,C@E@,8,C@,,C@F,FE<6CFF,6CC<@  
 FE9,6,,,5CE5CEE,,,+,@=9,,,9,:,@,,:CFGF>=:7++8A@ED?FGG?A,,,4,8  
 ,4CE9E++@>FGC+@,6@,6,@,6@F9D@6++,,4,311@C8EF7E:@E\*\*04;\*@,41@\*,  
 3\*4/\*45\*;\*971\*)2:>?5C8BEA\*)))))/+/,./).))))..))))\*1)/((

**Sequence**

**Sequence “quality”**

Read 1

ATGCGATCGTA **ATG**

000... **XXX**

← Quality (ex. good or bad)

Read 2

ATGCGATCGTA **CCC**

000... **XXX**

Read 3

ATGCGATCGTA **TTT**

000... **XXX**

Read 4

ATGCGATCGTA **TGA**

000... **XXX**

Read 5

ATGCGATCGTA **GGG**

000... **XXX**

Read 6

ATGCGATCGTA **ATG**

000... **XXX**

**Variant sequences are likely due to technical errors**



Sequencing errors often occur near the 3' end

## Denoising



Check point 2: NGS data is noisy.  
remove low-quality reads and sequences to ensure accuracy.

# Data processing for species determination

Sample5\_L001\_R1\_001.fastaq

50,685  
reads



## Denoising

## Extraction of the accurate sequences

# Amplicon Sequence Variants (ASVs)

**ATGCGATCGTT**

ATGCGATCGCA

ATGCGAAAGTA

**ATGCGATTAA**

ATACGATCGTA



## Annotation

# Data processing for species determination

## Sample5\_L001\_R1\_001.fastaq

reads

# DADA2 analysis

# Denoising

# Extraction of the accurate sequences

## Amplicon Sequence Variants (ASVs)

**ATGCGATCGTT**

ATGCGATCGCA

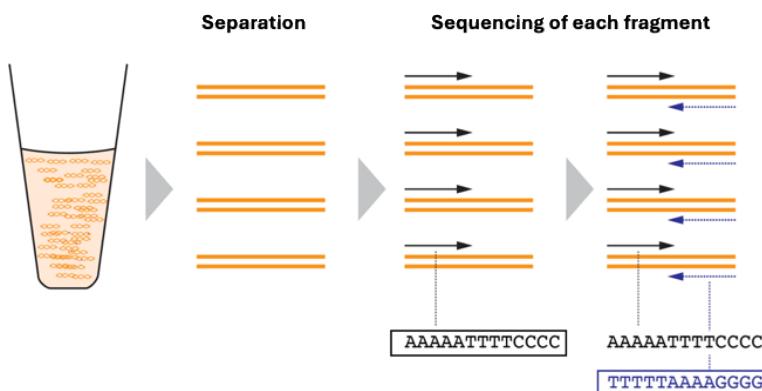
ATGCGAAAGTA

**ATGCGATTAA**

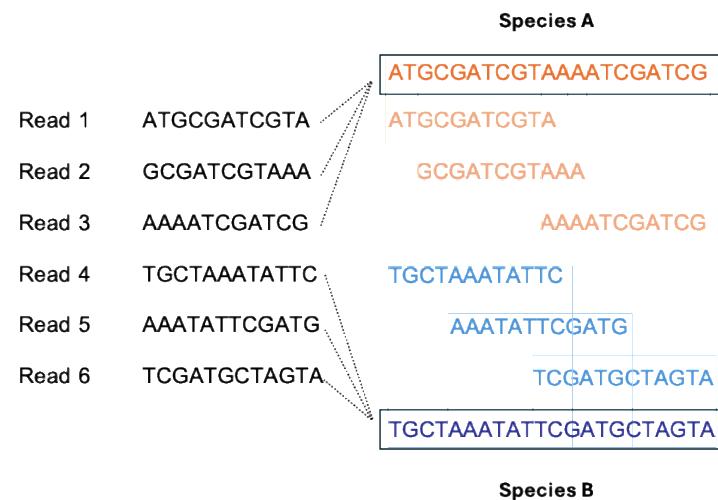
**ATACGATCGTA**

## Annotation

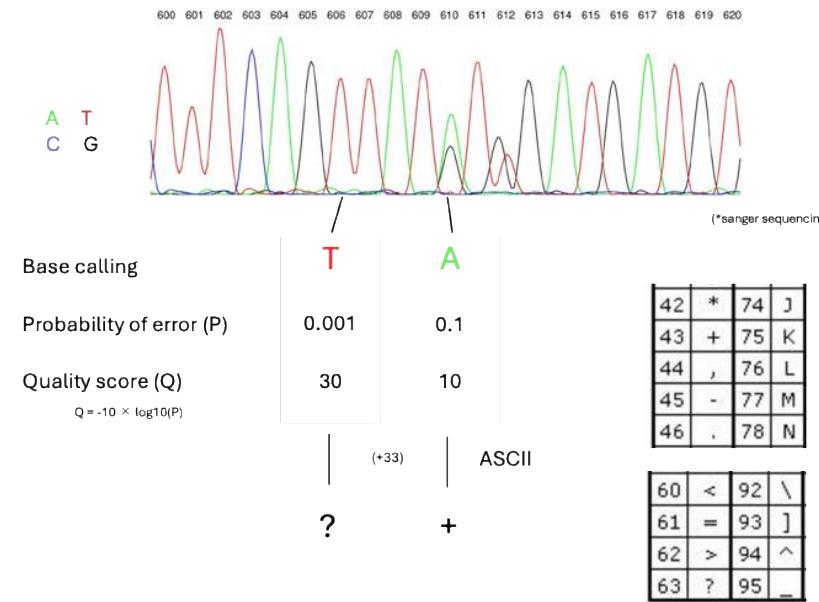
## Paired end sequencing



## Sequencing variation



## Sequencing accuracy



## Practice 2: Metabarcoding Data Processing

dada2

Install

Tutorial

Big Data

Documentation ▾

Evaluation ▾



### DADA2: Fast and accurate sample inference from amplicon data with single-nucleotide resolution



The DADA2 1.26 release is live, with native support for ARM architectures such as the Apple M1/M2 chips! [Release notes.](#)

#### Installation

Binaries for the current release version of DADA2 (1.26) are available from Bioconductor. Note that you must have R 4.2.0 or newer, and [Bioconductor version 3.16](#), to install the most current release from Bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("dada2", version = "3.16")
```

If you wish to install the latest and greatest development version, or to install to earlier versions of R, see our [from-source installation instructions](#).

#### Tutorials

## R set up

```
sudo apt update -qq
sudo apt install --no-install-recommends software-properties-common dirmngr
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc | sudo tee -a /etc/apt/trusted.gpg.d/cran-archive-keyring.gpg
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_release -cs)-cran40/"
sudo apt install --no-install-recommends r-base
```

## Quality check

```
R
#open R console
```

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ R

R version 4.5.0 (2025-04-11) -- "How About a Twenty-Six"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

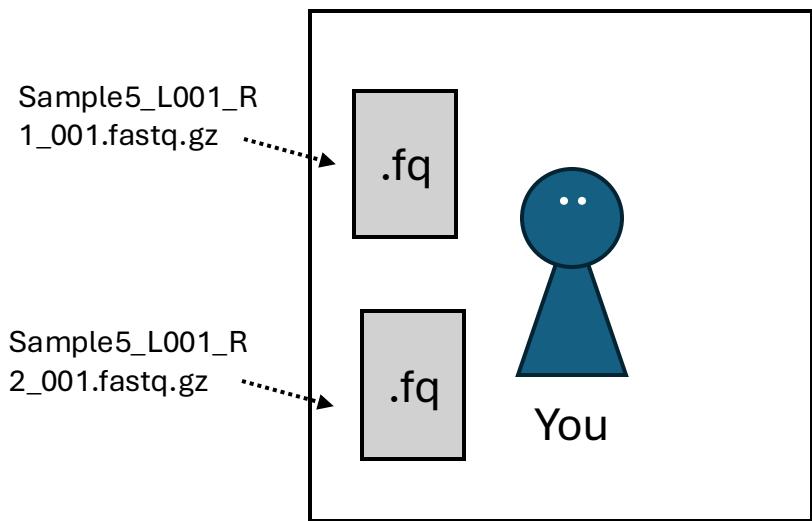
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

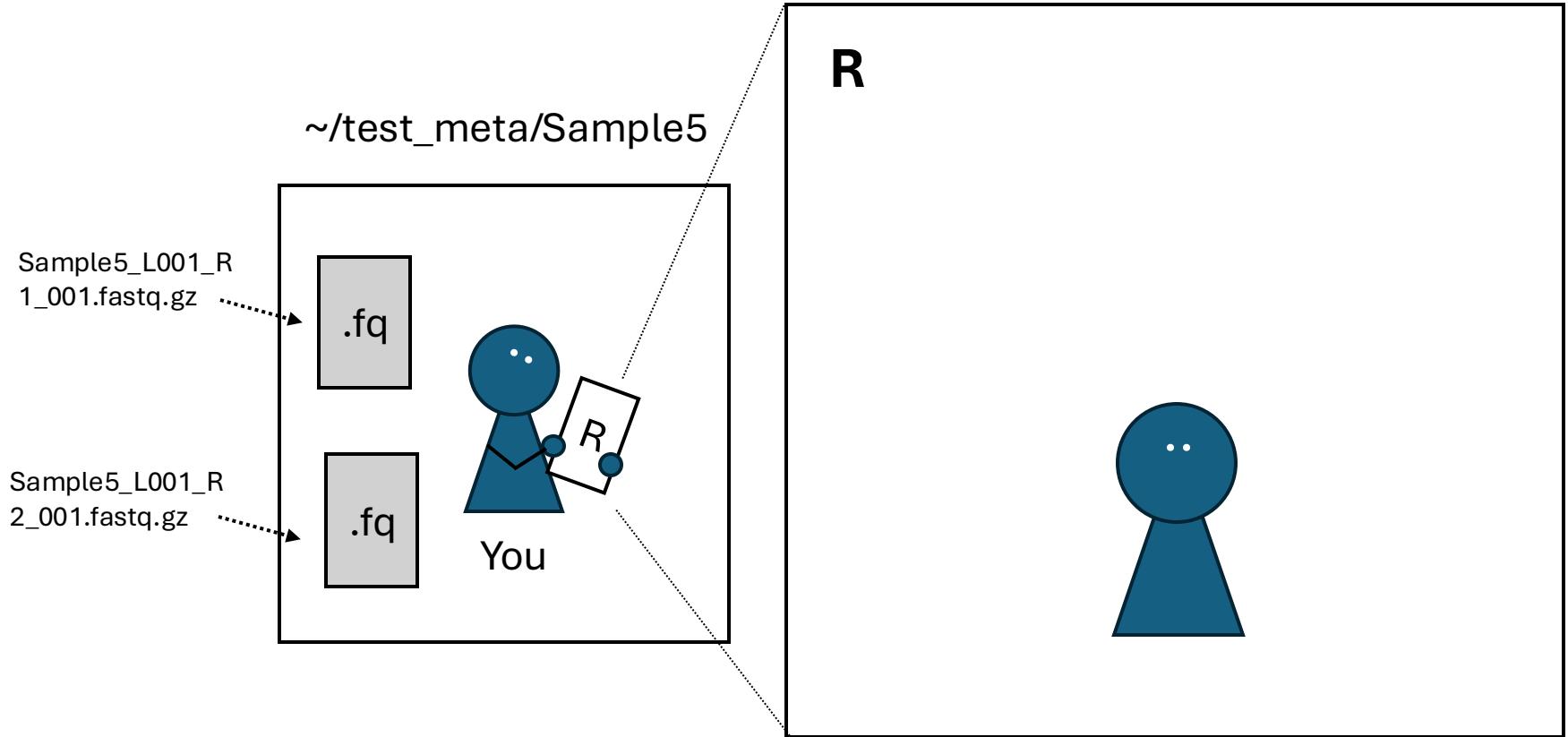
> []
```

## Dada2 set up and data installing

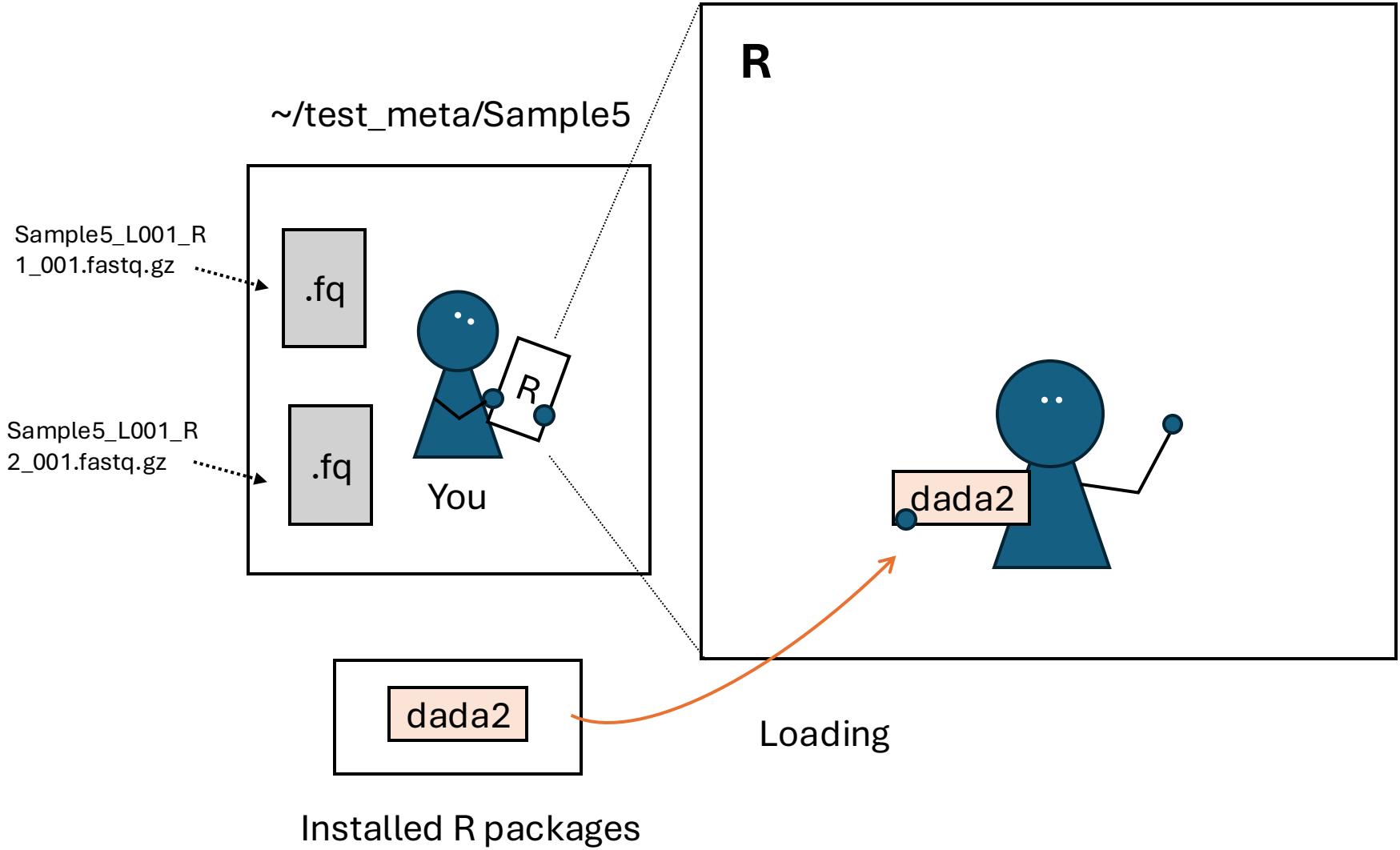
~/test\_meta/Sample5



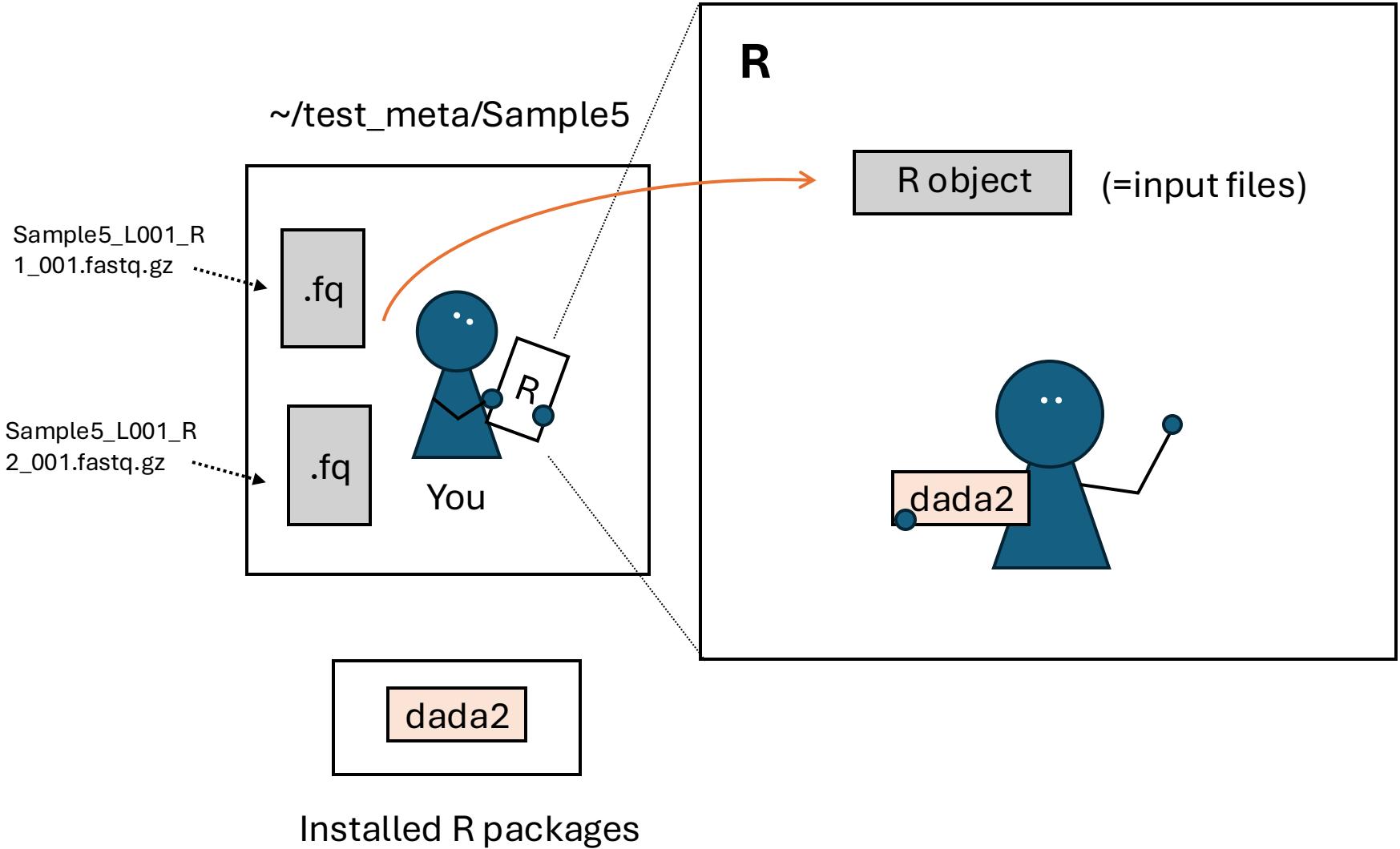
## Dada2 set up and data installing



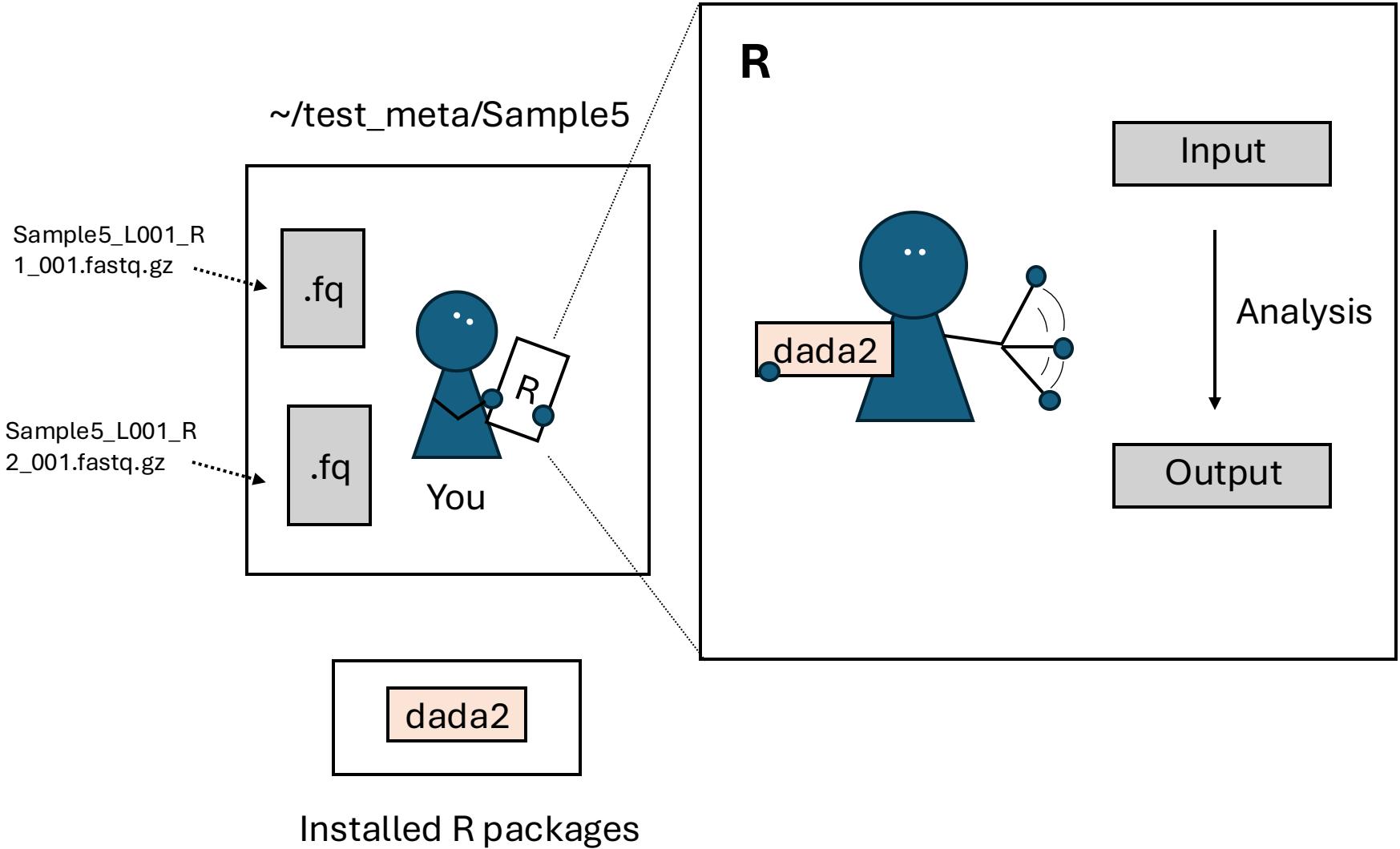
## Dada2 set up and data installing



## Dada2 set up and data installing



## Dada2 set up and data installing



## Dada2 set up and data installing

Installed  
R packages

DADA2

ggplot2

⋮

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ R

R version 4.5.0 (2025-04-11) -- "How About a Twenty-Six"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(dada2)
Loading required package: Rcpp
> []
```

library() is a function to load an installed R package into the current session

# Dada2 set up and data installing

```
fnFs <- c("Sample5_L001_R1_001.fastq.gz")
fnRs <- c("Sample5_L001_R2_001.fastq.gz")
sample.names <- c("Sample5")
```

An R object is any data structure or piece of data that exists in R's memory — essentially, **everything** you work with in R is an object.

## 📦 Common Types of R Objects

Object Type	Example	Description
Vector	<code>x &lt;- c(1, 2, 3)</code>	A series of numbers (or characters, etc.)
Matrix	<code>matrix(1:4, nrow=2)</code>	2D array of same-type data
List	<code>list(name="A", age=25)</code>	Collection of objects (can be different types)
Data Frame	<code>data.frame(name="A", age=25)</code>	Table-like structure (like Excel sheet)
Function	<code>myfun &lt;- function(x) x^2</code>	Functions are also objects
Factor	<code>factor(c("low", "high"))</code>	Categorical data
S4 / S3 object	from packages like Seurat	More advanced, used in bioinformatics etc.

## Quality check

Functions for plotting\*\*

```
png("plot_fnFs.png", width=800, height=600)
plotQualityProfile(fnFs)
dev.off()
```

Function for quality check

\*\*If you are using your local PC, the png() functions are unnecessary. Enter plotQualityProfile(fnFs), and the plot will appear in a new window.

Cloud Shell Editor

File Edit Selection View Go Run Terminal Help

Search

Click “Open file...”

Open a file, get code suggestions as you type, and press **tab** to accept

Press **ctrl+i** to ask Gemini to create or modify code

Select code in the

Layout: U.S.

cloudshell

```
patterns = "R2_001.fastq",
full.names = TRUE))

sample.names <- sapply(strsplit(basename(fnFs), " "), `[`, 1)
[1] "Sample5_L001_R1_001.fastq.gz" "Sample5_L001_R2_001.fastq.gz"
> sample.names
[1] "Sample5"
> fnFs
[1] "./Sample5_L001_R1_001.fastq.gz"
> fnFs <- c("Sample5_L001_R1_001.fastq.gz")
fnRs <- c("Sample5_L001_R2_001.fastq.gz")

sample.names <- c("Sample5")
> fnFs
[1] "Sample5_L001_R1_001.fastq.gz"
> png("plot_fnRs.png", width=800, height=600)
plotQualityProfile(fnRs[1:2])
dev.off()
null device
  1
> png("plot_fnFs.png", width=800, height=600)
plotQualityProfile(fnFs[1:2])
dev.off()
null device
  1
> fnFs[1:2]
[1] "Sample5_L001_R1_001.fastq.gz" NA
> plotQualityProfile(fnFs)
> png("plot_fnFs.png", width=800, height=600)
plotQualityProfile(fnFs)
dev.off()
null device
  1
```

## Cloud Shell Editor

File Edit Selection View Go Run Terminal Help



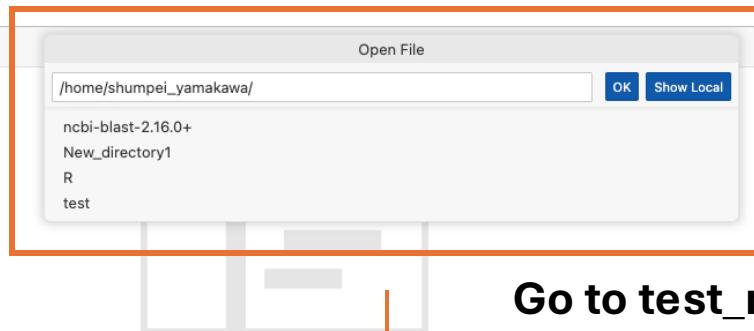
X 0 △ 0 Cloud Code - No Project

```
cloudshell + 

pattern="R2_001.fastq",
full.names = TRUE)

sample.names <- sapply(strsplit(basename(fnFs), " "), `[`, 1)
[1] "Sample5_L001_R1_001.fastq.gz" "Sample5_L001_R2_001.fastq.gz"
> sample.names
[1] "Sample5"
> fnFs
[1] "./Sample5_L001_R1_001.fastq.gz"
> fnFs <- c("Sample5_L001_R1_001.fastq.gz")
fnRs <- c("Sample5_L001_R2_001.fastq.gz")

sample.names <- c("Sample5")
> fnFs
[1] "Sample5_L001_R1_001.fastq.gz"
> png("plot_fnFs.png", width=800, height=600)
plotQualityProfile(fnRs[1:2])
dev.off()
null device
  1
> png("plot_fnRs.png", width=800, height=600)
plotQualityProfile(fnFs[1:2])
dev.off()
null device
  1
> fnFs[1:2]
```

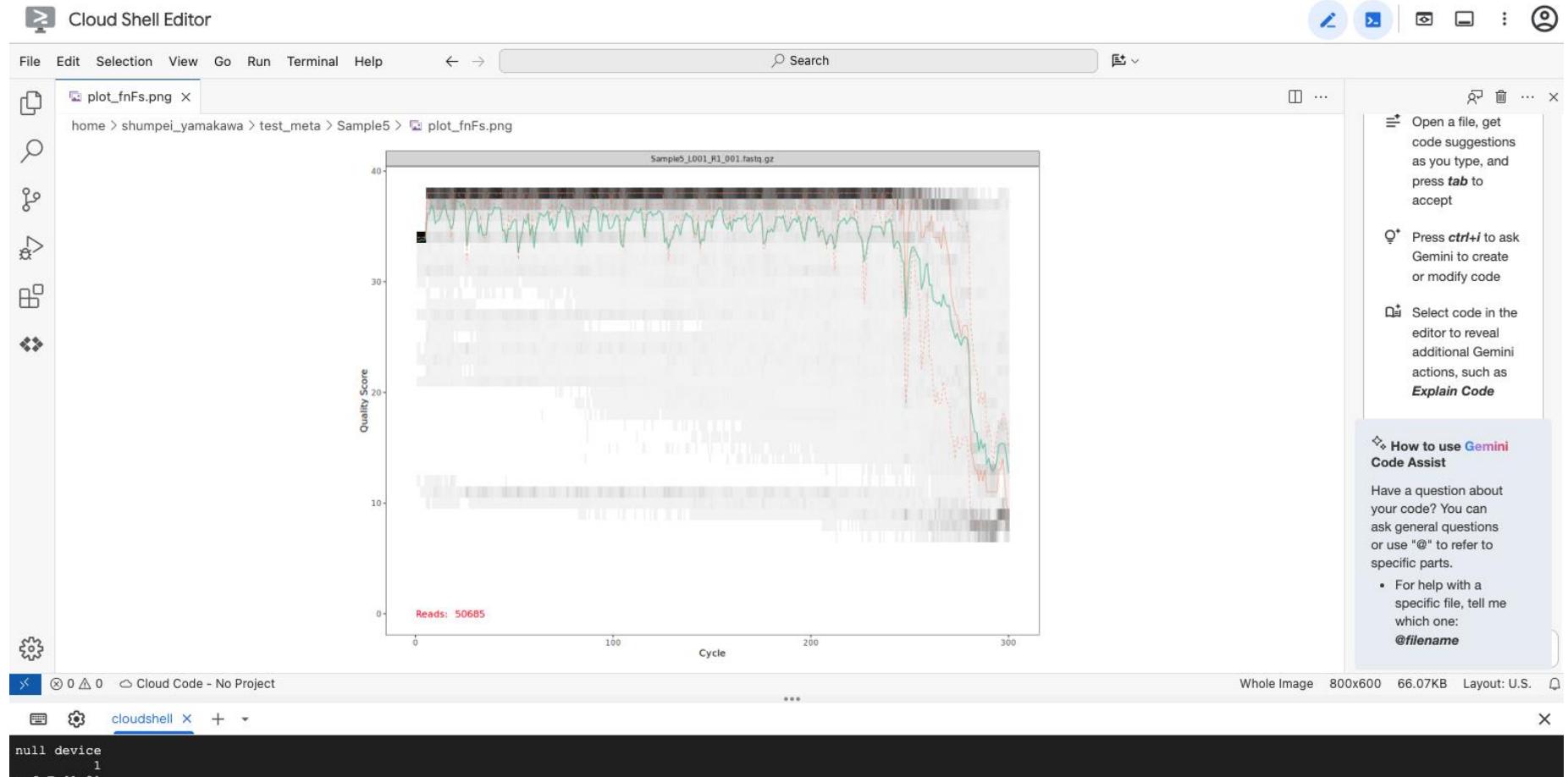


Go to test\_meta/Sample5



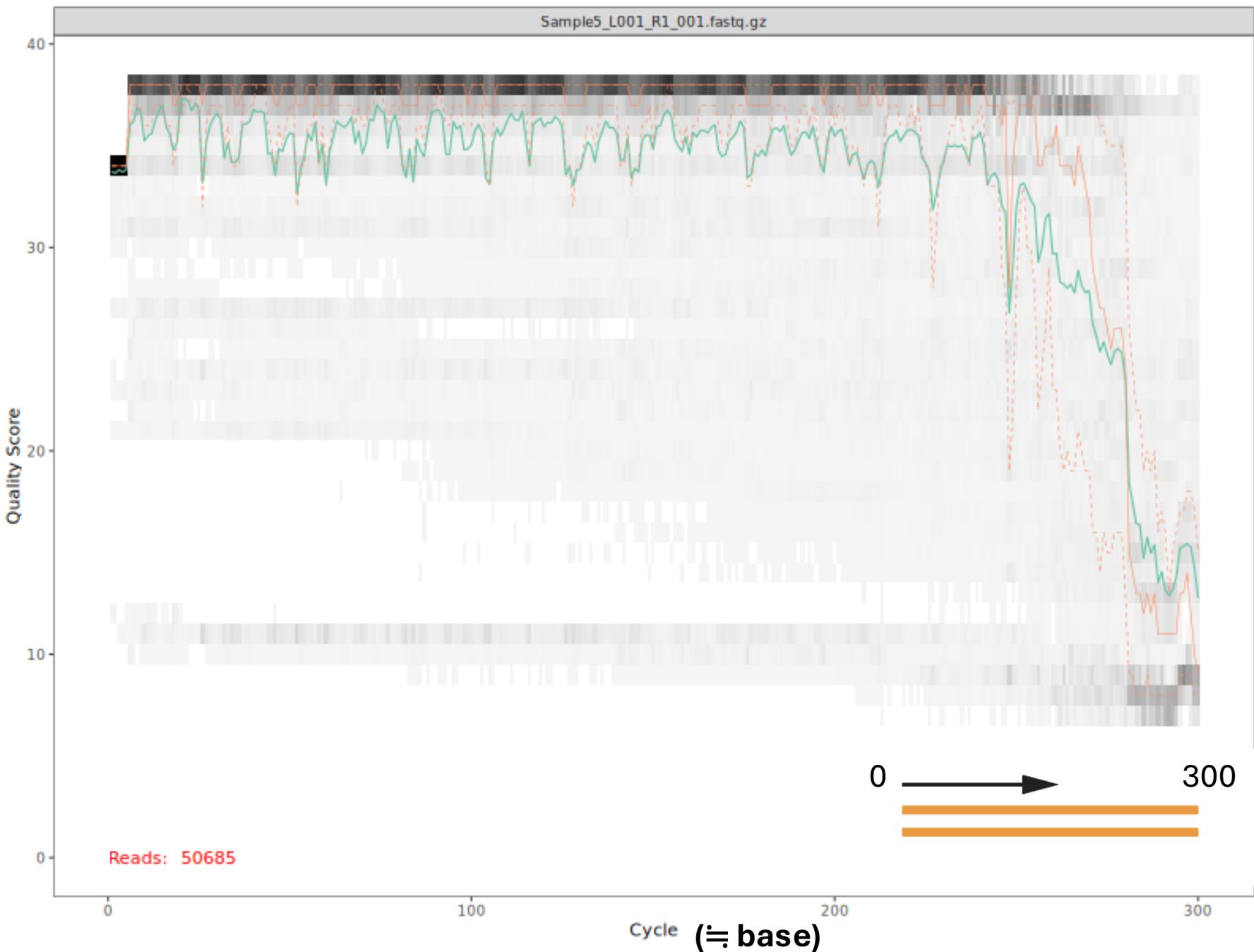
Find “plot\_fnFs.png”

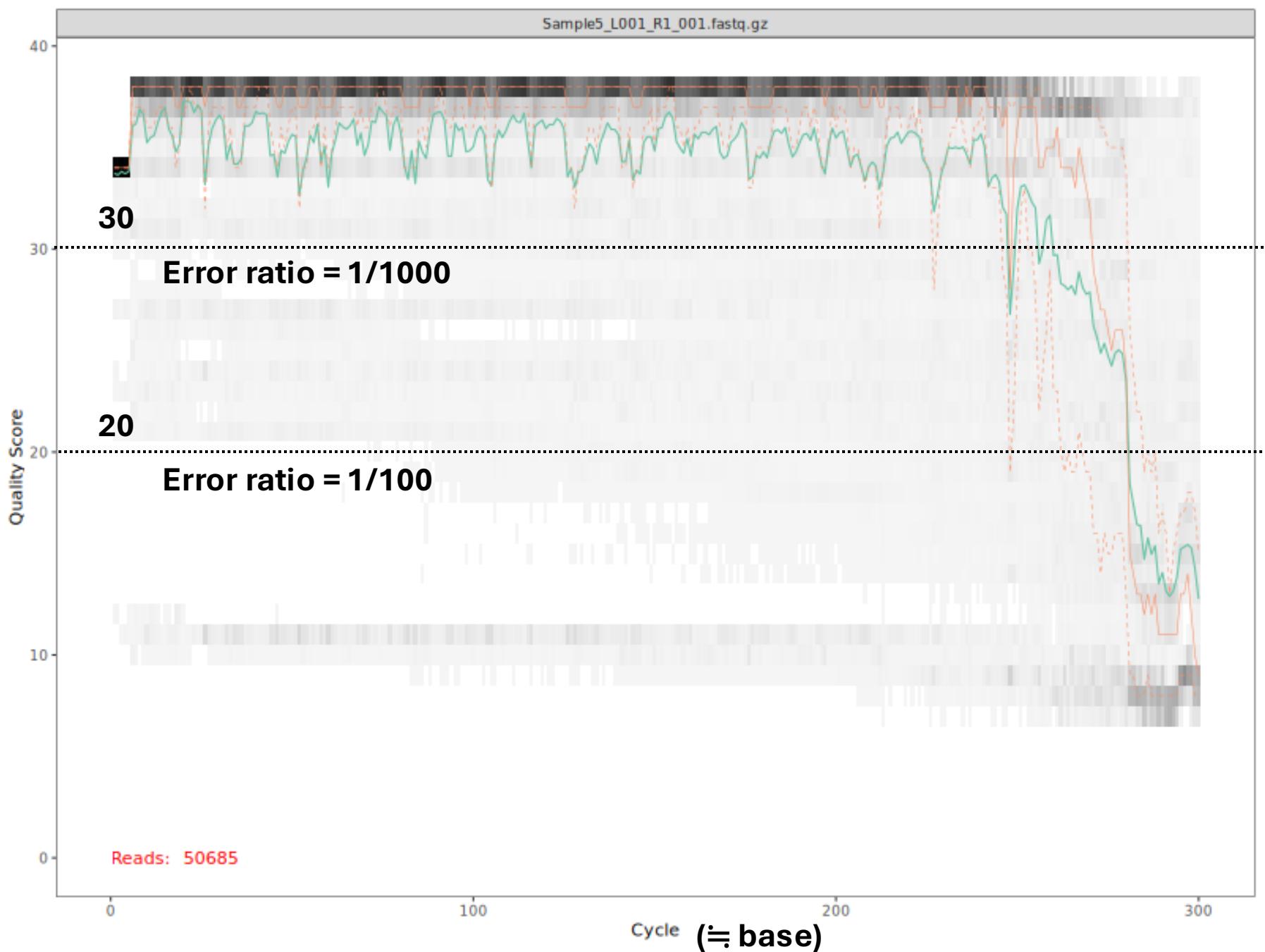
# “plot\_fnFs.png” opens above the shell

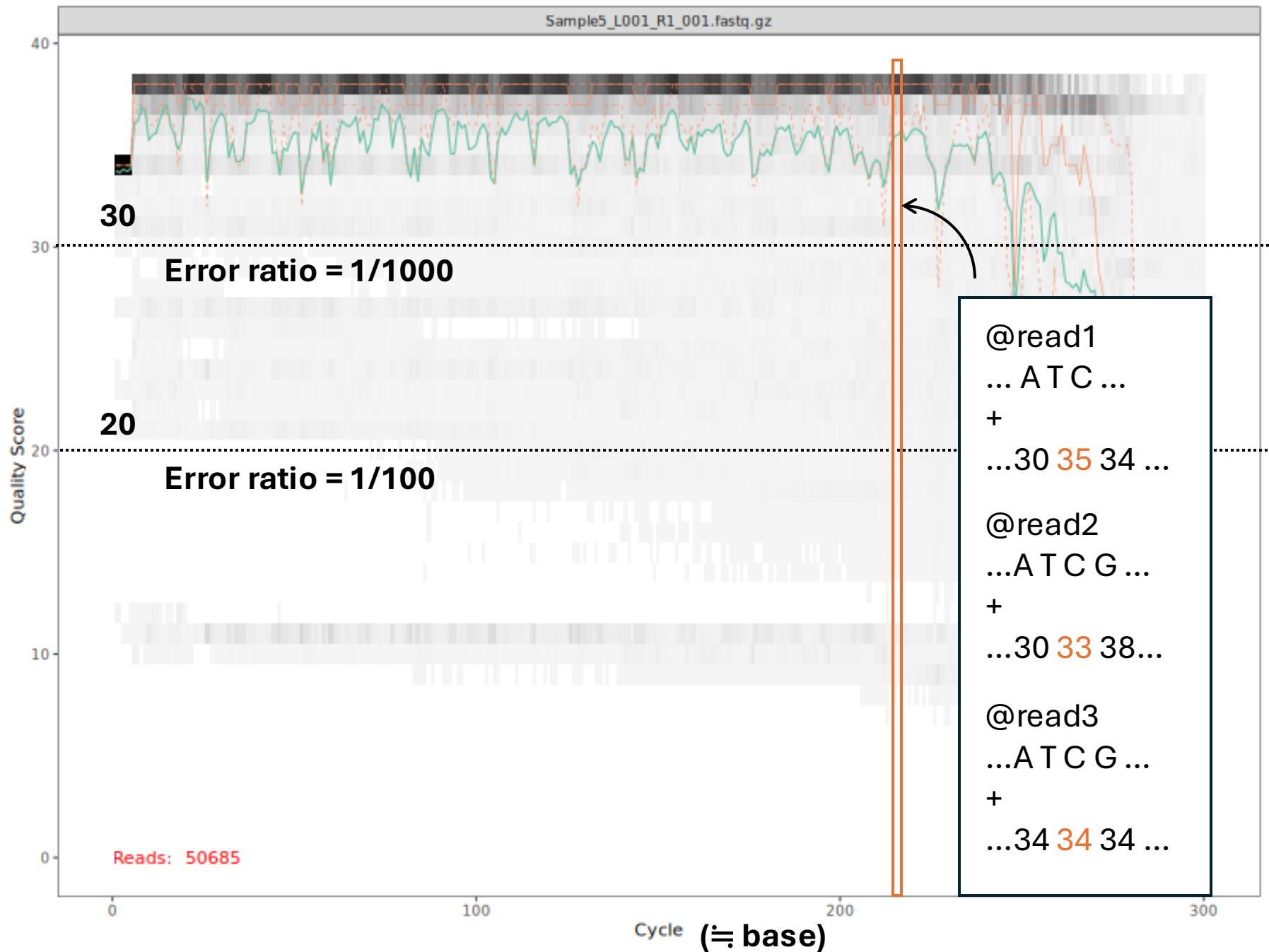




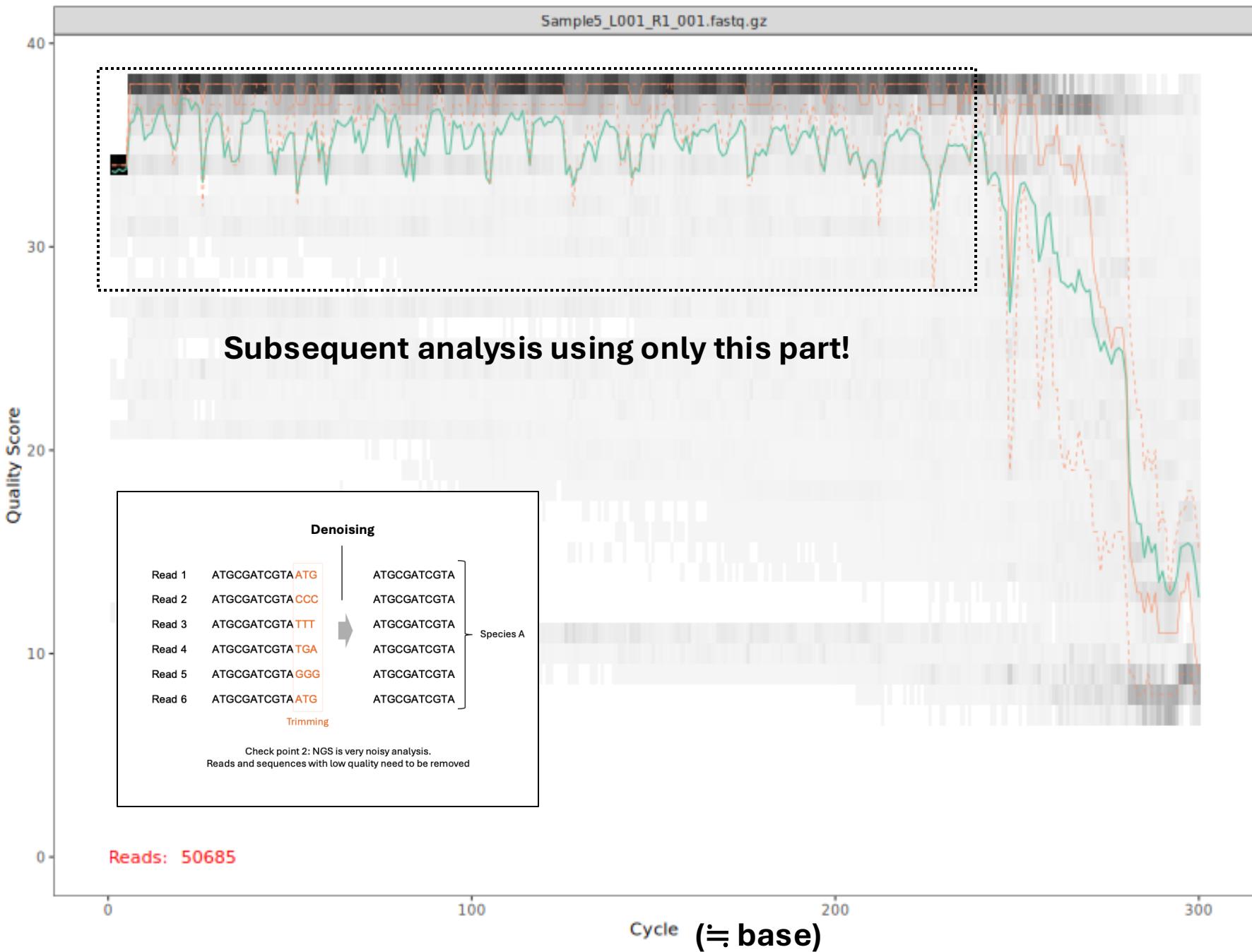
Sample5\_L001\_R1\_001.fastq.gz











# Day1.md

## Day1

You will learn how to process raw read data and perform annotation using eDNA metabarcoding data from a pond near Halle as test data.

Download the raw read sequences

```
cd test_meta  
mkdir Sample5  
#Any names are okay  
cd Sample5  
  
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample5_L001_R1_001.fq  
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample5_L001_R2_001.fq
```

Viewing the raw data

```
gunzip Sample5_L001_R1_001.fastq.gz  
#check the file contents  
head Sample5_L001_R1_001.fastq  
less Sample5_L001_R1_001.fastq  
  
#change fastaq to fasta  
cat Sample5_L001_R1_001.fastq | seqkit fq2fa | head  
cat Sample5_L001_R1_001.fastq | seqkit fq2fa | less  
#less can be terminated by typing "q"  
  
gzip Sample5_L001_R1_001.fastq  
##Do not forget to compress the fastq file!!!
```

## Day2

R set up

```
sudo apt update -qq  
sudo apt install --no-install-recommends software-properties-common dirmngr  
wget -qO - https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc | sudo tee -a /etc/apt/trusted.gpg.d/cran  
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu ${lsb_release -cs}-cran40/"  
sudo apt install --no-install-recommends r-base
```

Quality check

```
R  
#open R console  
  
library(dada2)  
  
# Assign the file path to the R objects
```

```
# Assign the file path to the R objects  
fnFs <- c("Sample5_L001_R1_001.fastq.gz")  
fnRs <- c("Sample5_L001_R2_001.fastq.gz")  
sample.names <- c("Sample5")
```

```
#quality check  
png("plot_fnFs.png", width=800, height=600)  
plotQualityProfile(fnFs)  
dev.off()  
  
png("plot_fnRs.png", width=800, height=600)  
plotQualityProfile(fnRs)  
dev.off()
```

Filtering

```
filtFs <- c("../filtered/Sample5_L001_R1_filtered_001.fastq.gz")  
filtRs <- c("../filtered/Sample5_L001_R2_filtered_001.fastq.gz")  
names(filtFs) <- sample.names  
names(filtRs) <- sample.names  
  
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),  
                      maxEE=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,  
                      compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE  
  
png("plot_fnFs_filtered.png", width=800, height=600)  
plotQualityProfile(filtFs)  
dev.off()  
  
png("plot_fnRs_filtered.png", width=800, height=600)  
plotQualityProfile(filtRs)  
dev.off()
```

Error rate estimate

```
errF <- learnErrors(filtFs, multithread=TRUE)  
errR <- learnErrors(filtRs, multithread=TRUE)  
  
#you can visualize the error rate using the following commands  
#png("plot_error_filtFs.png", width=800, height=600)  
#plotErrors(errF, nominal=TRUE)  
#dev.off()  
#png("plot_error_filtRs.png", width=800, height=600)
```

## Filtering

```
filtFs <- c("../filtered/Sample5_L001_R1_filtered_001.fastq.gz")
filtRs <- c("../filtered/Sample5_L001_R2_filtered_001.fastq.gz")
names(filtFs) <- sample.names
names(filtRs) <- sample.names

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),
                      maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
                      compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE

png("plot_fnFs_filtered.png", width=800, height=600)
plotQualityProfile(filtFs)
dev.off()

png("plot_fnRs_filtered.png", width=800, height=600)
plotQualityProfile(filtRs)
dev.off()
```

```
filtFs <- c("../filtered/Sample5_L001_R1_filtered_001.fastq.gz")  
filtRs <- c("../filtered/Sample5_L001_R2_filtered_001.fastq.gz")
```



Specify the path and file name where the reads to be filtered in subsequent analysis will be stored

```
names(filtFs) <- sample.names  
names(filtRs) <- sample.names
```



Use the same sample names

Input file (F)      Output file (F)      Input file (R)      output file (R)

```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),
                      maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
                      compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```

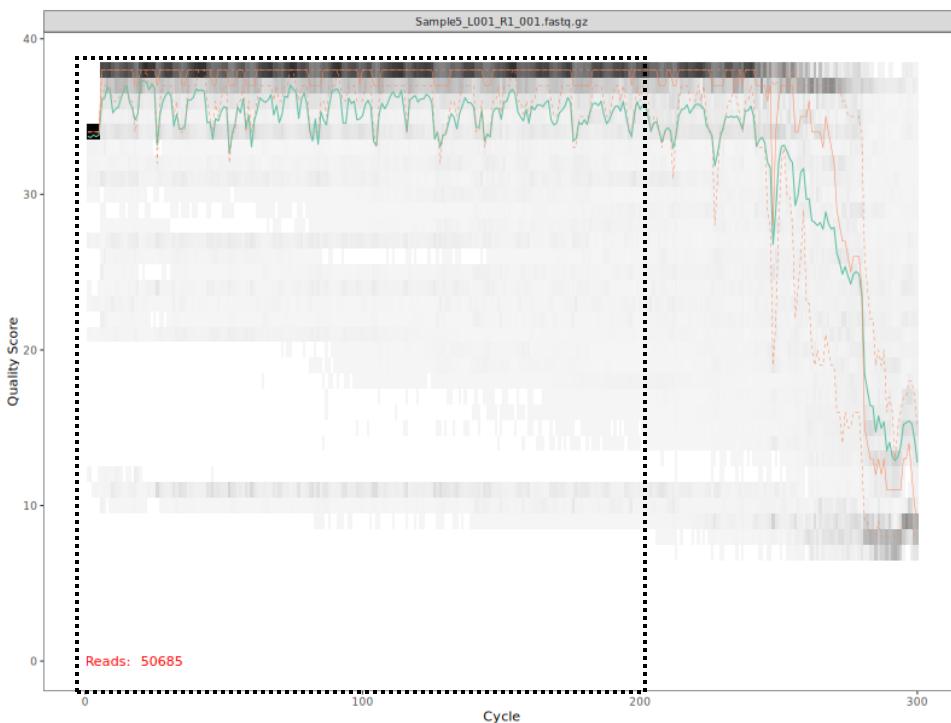
Function for filtering  
and trimming

Conditions/parameters  
for filtering

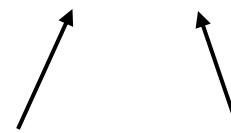
```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),  
                     maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,  
                     compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```



Truncates the forward and reverse reads to 200 bases each



truncLen = c( XXX, YYY)



Length of F reads    R reads

Ex) truncLen = c( 100, 200)

truncLen = c( 150, 200)

truncLen = c( 200, 100)

```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),  
                     maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,  
                     compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```



maxN

Maximum number of ambiguous bases (Ns) allowed per read. Reads with more are discarded.

maxEE

Maximum expected errors allowed per read (forward and reverse). Reads exceeding are discarded.

truncQ

Quality score threshold to truncate reads at the first base with quality  $\leq$  this value.

rm.phix

Whether to remove reads matching the PhiX genome (TRUE = remove).

compress

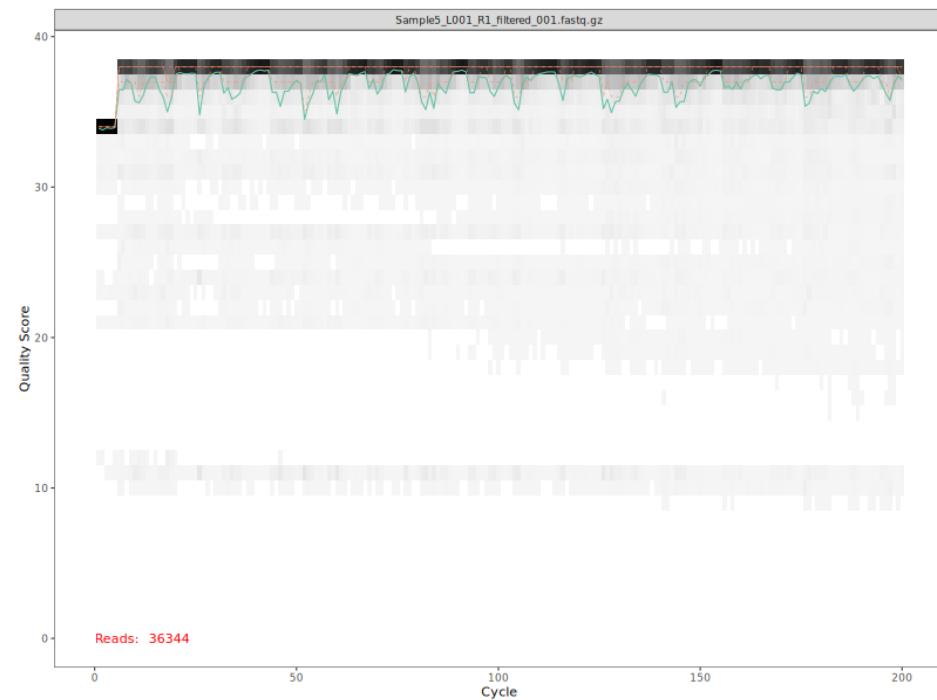
Whether to save filtered reads as compressed .gz files (TRUE = compressed).

```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),
                      maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
                      compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE

>
> out
          reads.in  reads.out
Sample5 L001 R1 001.fastq.gz    50685     36344
```



**Before**  
**(plot\_fnFs.png)**



**After**  
**(plot\_fnFs\_filtered.png)**

## Error rate estimate

```
errF <- learnErrors(filtFs, multithread=TRUE)
errR <- learnErrors(filtRs, multithread=TRUE)

#you can visualize the error rate using the following commands
#png("plot_error_filtFs.png", width=800, height=600)
#plotErrors(errF, nominalQ=TRUE)
#dev.off()
#png("plot_error_filtRs.png", width=800, height=600)
#plotErrors(errR, nominalQ=TRUE)
#dev.off()
```



### Ex) Estimate the errors

TAT**C**TATC    TAT**G**TATC  
TAT**C**TATC    TAT**G**TATC  
TAT**C**TATC    TAT**G**TATC

...

5,000 reads    2,800 reads

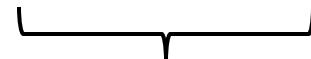


**Biological variation**

TAT**C**TATC    TAT**G**TATC  
TAT**C**TATC  
TAT**C**TATC

...

10,000 reads    1 reads



**Technical errors (C -> G)**

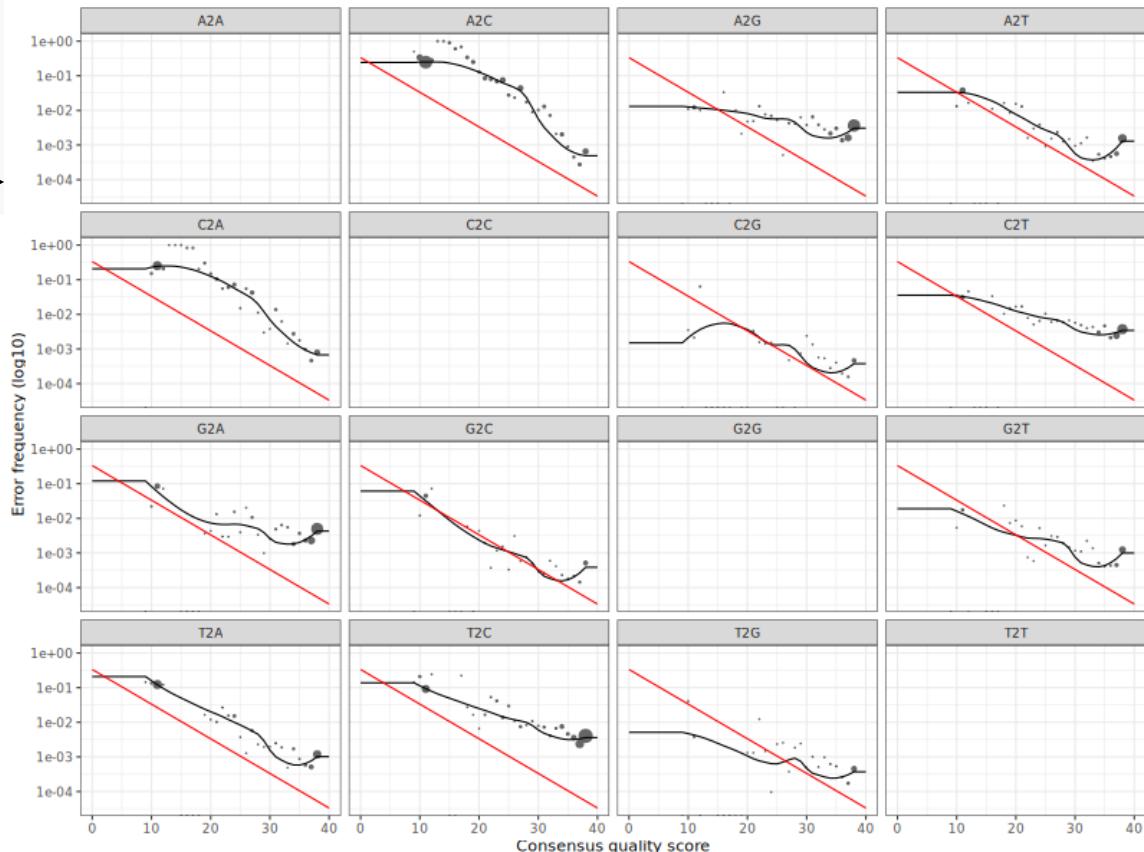
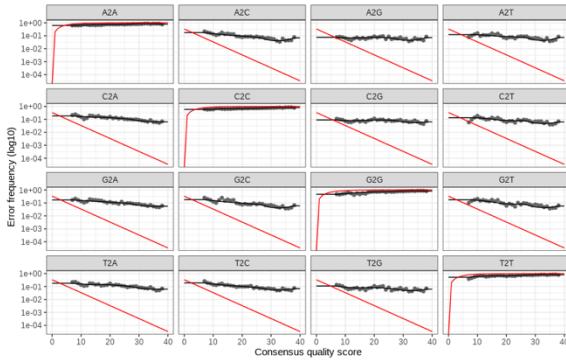
```
errF <- learnErrors(filtFs, multithread=TRUE)
errR <- learnErrors(filtRs, multithread=TRUE)
```

```
> errF <- learnErrors(filtFs, multithread=TRUE)
7268800 total bases in 36344 reads from 1 samples will be used for learning the error rates.
> errR <- learnErrors(filtRs, multithread=TRUE)
7268800 total bases in 36344 reads from 1 samples will be used for learning the error rates.
```

```
#you can visualize the error rate using the following commands
#png("plot_error_filtFs.png", width=800, height=600)
#plotErrors(errF, nominalQ=TRUE)
#dev.off()
#png("plot_error_filtRs.png", width=800, height=600)
#plotErrors(errR, nominalQ=TRUE)
#dev.off()
```



bad example

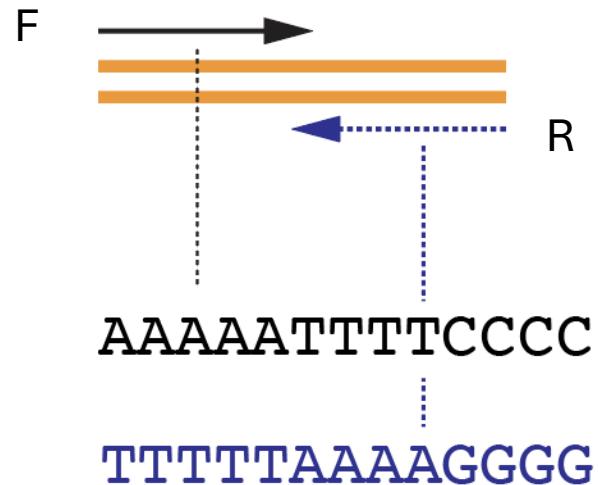


## Merge

```
dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
dadaRs <- dada(filtRs, err=errR, multithread=TRUE)

dadaFs
dadaRs

mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
# Inspect the merger data.frame from the first sample
head(mergers[[1]])
```



## Detection of variant sequences

```
dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
```

Filtered reads (F)

Estimated error rate

```
> dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
Sample 1 - 36344 reads in 15805 unique sequences.
> dadaFs
dada-class: object describing DADA2 denoising results
169 sequence variants were inferred from 15805 input unique sequences.
Key parameters: OMEGA A = 1e-40, OMEGA C = 1e-40, BAND SIZE = 16
```

Raw data	50,685
After filtering	36,444
Unique sequences	15,805
Sequence variant	169

Sample5\_L001\_R1\_001.fastaq

A thick, solid gray arrow pointing to the right, indicating the direction of the next section.

# Denoising

## Extraction of the accurate sequences

# AmpliSeq Sequence Variants (ASVs)

**ATGCGATCGTT**

**ATGCGATCGCA**

ATGCGAAAGTA

**ATGCGATTAA**

**ATACGATCGTA**

# Forward: 169

# Reverse: 116

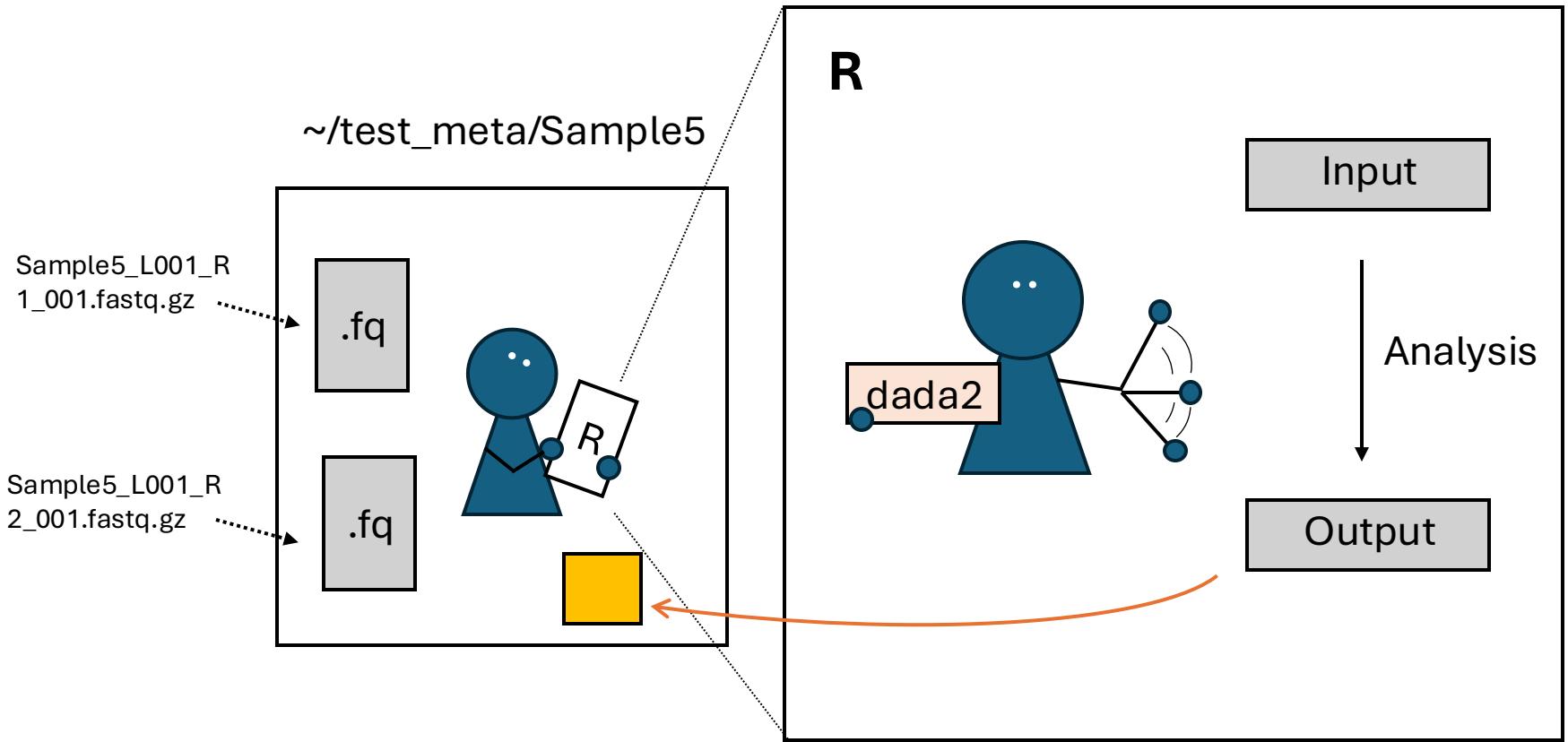
```
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
31661 paired-reads (in 126 unique pairings) successfully merged out of
35355 (in 586 pairings) input.
```

```
> head(mergers[[1]])
[1] "ACGAGAACCCGTGGAGCTAATTTATCGTAGCTAAACTCTGCCATTAATTGTATGGGACTACTGAGTAACATATGACTTATATTACTTTAGATTGATCTATTCTTAACTATTGAGAAGCTACTTGGGATAACAGGGTAGTCCTT
ATCGATGACACAATTACGACCTCGATGTTGGA"
[2] "ACGAGAACCCGTGGAGCTAATTTATCGTAGCTAAACTCTGCCATTAATTGTATGGGACTACTGAGTAACATATGACTTATATTACTTTAGATTGATCTATTCTTAACTATTGAGAAGCTACTTGGGATAACAGGGTAGTCCTT
ATCGATGACACAATTACGACCTCGATGTTGGA"
[3] "ACGAGAACCCGTGGAGCTAATTTATCGTAGCTAAACTCTGCCATTAATTGTATGGGACTACTGAGTAACAAATGACTTATATTACTTTATATTGATCTATTCTTAAATTATTGAGAAGCTACTTGGGATAACAGGGTAGTCCTT
ATCGATGACACAATTACGACCTCGATGTTGGA"
[4] "ACGAGAACCCGTGGAGCTAATTTATCGTAGCTAAACTCTGCCATTAATTGTATGGGACTACTGAGTAACAAATGACTTATATTACTTTATATTGATCTATTCTTAAATTATTGAGAAGCTACTTGGGATAACAGGGTAGTCCTT
ATCGATGACACAATTACGACCTCGATGTTGGA"
[5] "CGAGCCTCTCCGATCTACGAGAACCCGTGGAGCTTACTTTAAGGTTAAGCCTACAAGTTAAATGGGAACCTTGAGGAACAAAATACCTCTGCCACCTACTCTCTATTGAGAAGCTACTTGGGATAACAGGTAATACAGCTGGGAGCACCTATCG
ATAGTTGTTCTACGACCTCGATGTTGAGATCGGAAGAGCACAC"
[6] "ACGAGAACCCGTGGAGCTAATTTACGCGCTTTATACCTATGCCGGTACAAATTACATGGGACTGTTGAGTGAGAAAATAATTGAAACTCTTACTATTGATCTAACCTTAATTATCGAAAAGCTACTTGGGATAACAGGGTAGCGAGTGGA
GAGTTCTTATCGATACTGCAATTACGACCTCGATGTTGGA"
```

**Forward: 169**

**Reverse: 116**

**Merge -> 126**



## Output to shell

```
write.table(mergers, file = "Sample5_asv.csv", row.names=FALSE, sep=",")
```

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ ls
a          mergers_table  plot_error_filtFs.png  plot_fnFs_filtered.png  plot_fnRs_filtered.png  Rplots.pdf
filtered   out.RDS        plot_error_filtRs.png  plot_fnFs.png          plot_fnRs.png          Sample5_asv.csv
```

Sample5\_asv.csv



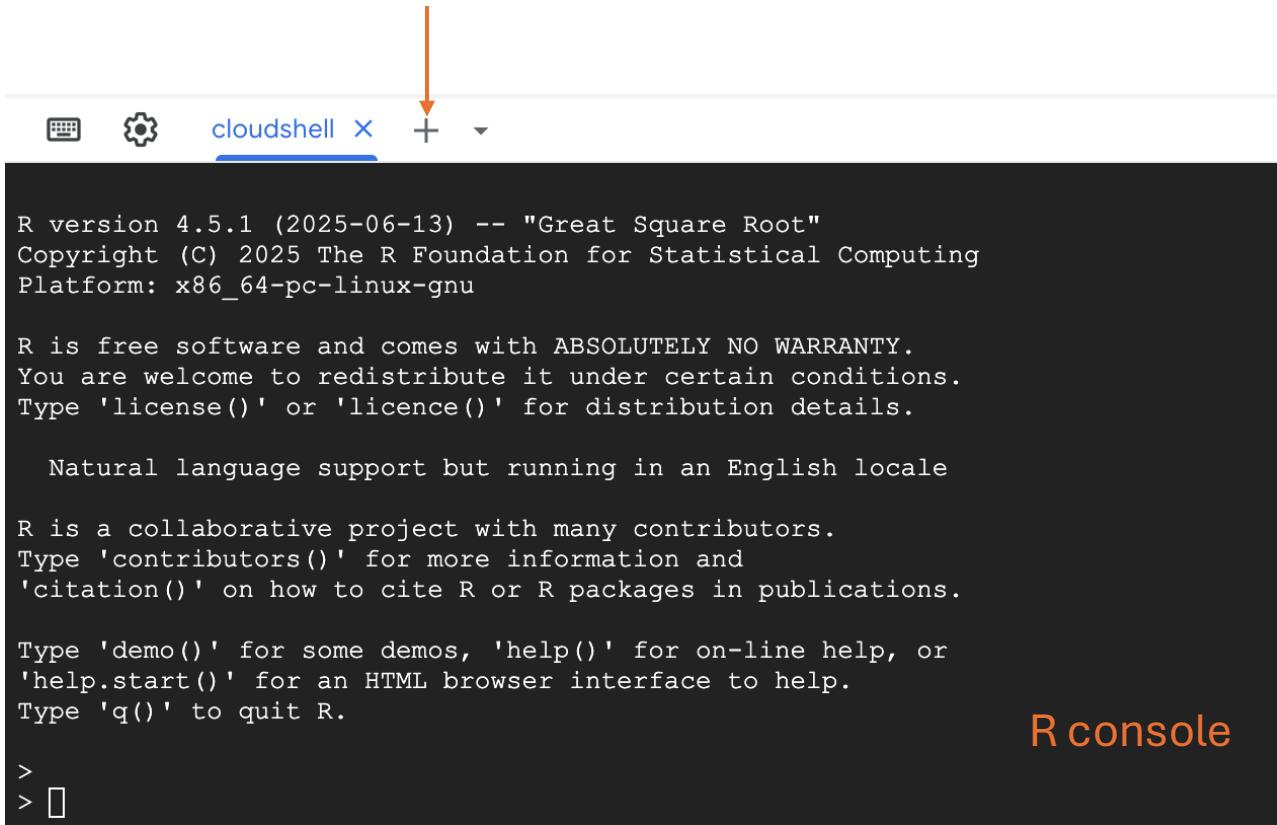
ASV table which was generated using R

```
Sample5$ head Sample5_asv.csv
```

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_asv.csv
"sequence","abundance","forward","reverse","nmatch","nmismatch","nindel","prefer","accept"
"ACGAGAAAGCCCTGTGGAGCTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACATATGACTTATTTAGATTGATCTATCCTTAACATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",1095,2,2,196,0,0,1,TRUE
"ACGAGAAAGCCCTATGGAGCTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACATATGACTTATTTAGATTGATCTATCCTTAACATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",1094,1,4,196,0,0,1,TRUE
"ACGAGAAAGCCCTGTGGAGCTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACAAATGACTTATTTATATTGATCTATCCTTAATTATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",903,3,3,196,0,0,1,TRUE
"ACGAGAAAGCCCTATGGAGCTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACAAATGACTTATTTATATTGATCTATCCTTAATTATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",836,5,16,196,0,0,1,TRUE
"CGACGCTCTCCGATCTACGAGAAAGACCTATGGAGCTTACTTTAAGGTTAACGCTACAAGTTAAATGGGACTTTGAGGAACAAATACCTCTGCCACCTACTCTCTATTGAGAACGCTACTTGGGATAACAGGTAATACAGCTGGGAGCACTATCGATAG
TTGTCTTACGACCTCGATGTTGATGATCGGAAGAGCACAC",816,6,52,183,0,0,1,TRUE
"ACGAGAAAGCCCTATGGAGCTTAATTACGCCTGTTTATACCTATGCCCGTATCAAATTACATGGGACTGTTGAGTGAGAAAATAATTGAAACTCTTACTATTGATCTAACTCTTAATTATCTGAAAGCTACTTGGGATAACAGGGTAAGCGAGTGGAGAGT
TCTTATCGATACTTGCATTACGACCTCGATGTTGA",804,4,8,188,0,0,1,TRUE
"ACGAGAAAGCCCTGTGGAACTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACAAATGACTTATTTATATTGATCTATCCTTAATTATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",802,13,15,196,0,0,1,TRUE
"ACGAGAAAGACCCGTGGAGCTTAATTACGCCTGTTTATACCTATGCCCGTATCAAATTACATGGGACTGTTGAGTGAGAAAATAATTGAAACTCTTACTATTGATCTAACTCTTAATTATCTGAAAGCTACTTGGGATAACAGGGTAAGCGAGTGGAGAGT
TCTTATCGATACTTGCATTACGACCTCGATGTTGA",790,10,9,188,0,0,1,TRUE
"ACGAGAAAGCCCTGTGGAACTTAATTATCGTAGCTAAACTCTGCCATTAAATTGTATGGGACTACTGAGTAAACATATGACTTATTTAGATTGATCTATCCTTAACATTGAGAACGCTACTTGGGATAACAGGGTGTAGTGTTCGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGA",765,7,5,196,0,0,1,TRUE
```

## Tips for the analysis

### Open a new tab



R version 4.5.1 (2025-06-13) -- "Great Square Root"  
Copyright (C) 2025 The R Foundation for Statistical Computing  
Platform: x86\_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

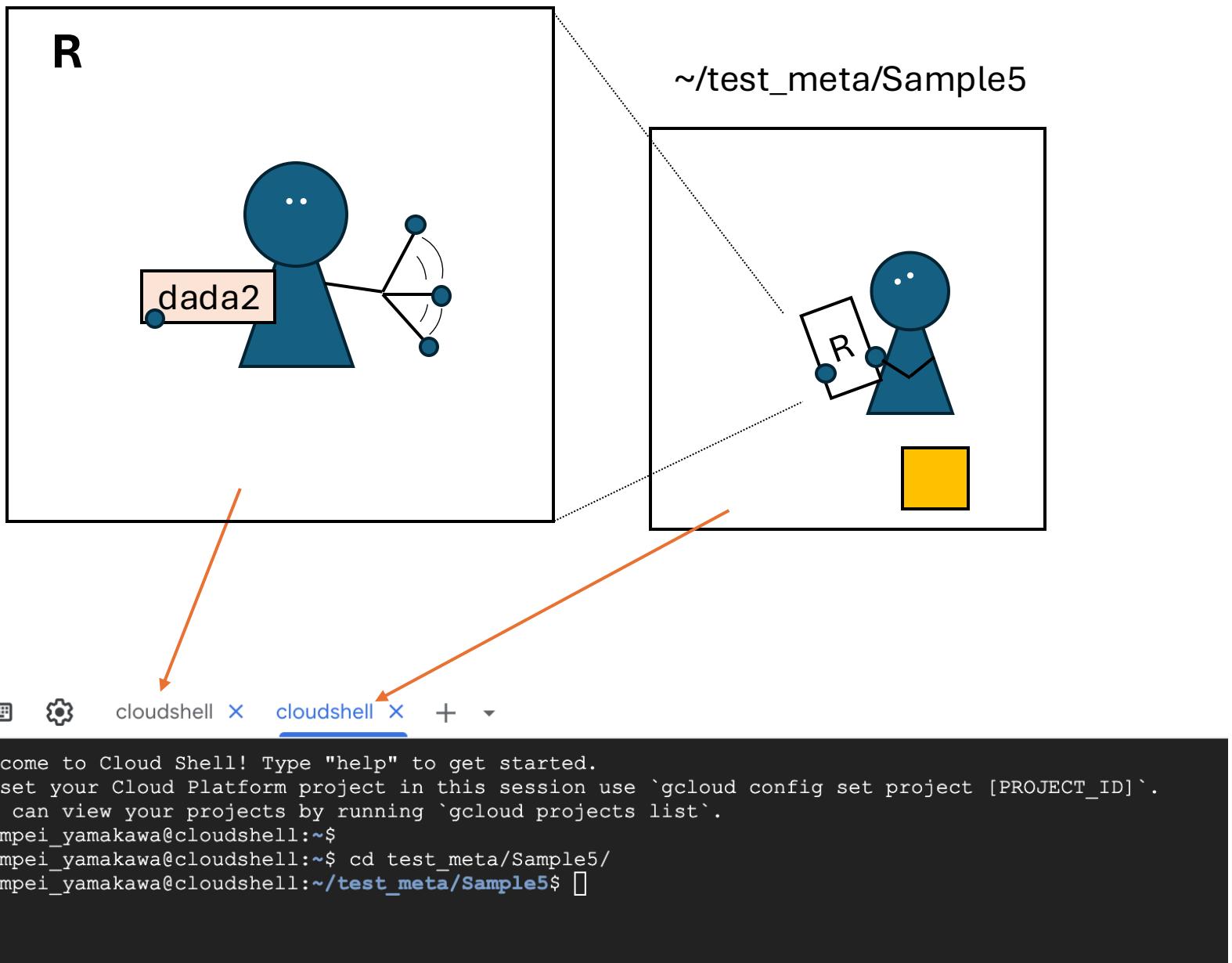
Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

>  
> □

R console



Sample5\_L001\_R1\_001.fastaq

50,685 reads

A large, solid grey arrow pointing to the right, indicating the direction of the next section.

# Denoising

## Extraction of the accurate sequences

# AmpliSeq Sequence Variants (ASVs)

**ATGCGATCGTT**

**ATGCGATCGCA**

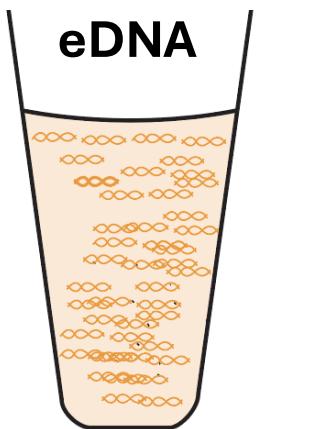
ATGCGAAAGTA

**ATGCGATTAA**

ATACGATCGTA

126

# variants



### Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

ATGCGATCGCA

ATGCGAAAGTA

ATGCGATTNTTA

ATACGATCGTA

**126**

**variants**

→Annotation

-11:30

## 1. Filtering

## 2. Error rate calculation

## 3. Merging

## 4. Creating output file

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ ls
a          mergers_table  plot_error_filtFs.png  plot_fnFs_filtered.png  plot_fnRs_filtered.png  Rplots.pdf
filtered   out.RDS        plot_error_filtRs.png  plot_fnFs.png           plot_fnRs.png           Sample5_asv.csv
```

# Practice 3: Annotation of Metabarcoding Data

## GenBank Overview

### What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan 41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample [GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

### Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

### GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the copying, or distribution of the information contained in GenBank.

# Genbank

# Database

## Query

ASVs

ATGCGATCGTT

ATGCGATCGCA

ATGCAGAAAGTA



National Library of Medicine  
National Center for Biotechnology Information

GenBank Nucleotide

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

GenBank Resources

GenBank Home  
Submission Types  
Submission Tools  
Search GenBank  
Update GenBank Records

## BLAST search

National Library of Medicine  
National Center for Biotechnology Information

BLAST®

Important update  
Effective August 2025, the ClusteredNR database will become the default Protein BLAST database. Learn more about ClusteredNR

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

News, 17 Mar 2025  
Improvements include upgrading to GCP Artifact Registry and better handling of job completion status in Kubernetes version 1.30+.  
ElasticBLAST 1.4.0 is now available! [More BLAST news...](#)

Web BLAST

Nucleotide BLAST (nucleotide ➤ nucleotide)

blastx (translated nucleotide ➤ protein)

tblastn (protein ➤ translated nucleotide)

Protein BLAST (protein ➤ protein)

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

 National Library of Medicine  
National Center for Biotechnology Information

[Log in](#)

BLAST®

**Important update**  
Effective August 2025, the **ClusteredNR** database will become the default Protein BLAST database. [Learn more about ClusteredNR](#)

**Basic Local Alignment Search Tool**

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

Mon, 17 Mar 2025  
Improvements include upgrading to GCP Artifact Registry and better handling of job completion status in kubernetes version 1.30+.  
ElasticBLAST 1.4.0 is now available! [More BLAST news...](#)

**Web BLAST**

**Nucleotide BLAST**  
nucleotide ► nucleotide

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Protein BLAST**  
protein ► protein

**BLAST Genomes**

Enter organism common name, scientific name, or tax id

**Search**

```
Sample5$ head Sample5_asv.csv
```

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_asv.csv
"sequence","abundance","forward","reverse","nmatch","nindel","prefer","accept"
"ACGAGAACGGCTGTGGAGCTTAATTTATCGTAGCTAAACTCTGCCATTAAATTGATGGGGACTTGAGTAAACATATGACTTATTTAGATTGATCTATTCTTAACTATTGAGAAGCTACTTGGGATAACAGGGTAGTGTTCGGGAGTCCTTATCG
ATGAACACATTACGACCTCGATGTTGGA",1095,2,2,196,0,0,1,TRUE
```

copy & paste

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

[Reset page](#) [Bookmark](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From  To

Or, upload file  No file chosen [?](#)

Job Title   
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  Standard databases (nr etc.)  rRNA/ITS databases  Genomic + transcript databases  Betacoronavirus  Experimental databases

Core nucleotide database (core\_nt) [?](#)

Organism **Optional**  
Enter organism name or id—completions will be suggested   exclude [Add Organism](#)  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude **Optional**  
 Models (XM/XP)  Uncultured/environmental sample sequences

Limit to **Optional**  
 Sequences from type material

Entrez Query **Optional**  
 [YouTube](#) [Create custom database](#)  
Enter an Entrez query to limit search [?](#)

[Feedback](#)

## Standard Nucleotide BLAST

blastn    blastp    blastx    tblastn    tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

[Reset page](#)    [Bookmark](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From   
To

Or, upload file  Choose file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  Standard databases (nr etc.)  rRNA/ITS databases  Genomic + transcript databases  Betacoronavirus  Experimental databases

[?](#)

Organism [Optional](#)

Enter organism name or id--completions will be suggested   exclude [Add Organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [Optional](#)

Models (XM/XP)  Uncultured environmental sample sequences

Limit to [Optional](#)

Sequences from type material

[YouTube](#) [Create custom database](#)

Entrez Query [Optional](#)

Enter an Entrez query to limit search [?](#)

[Feedback](#)

Core nucleotide database:  
“All” nucleotide sequences which are deposited in ncbi

**Standard Nucleotide BLAST**

blastn    blastp    blastx    tblastn    tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

[Reset page](#)    [Bookmark](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From  To

Or, upload file [Choose file](#) Sample5\_asv.fa [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  Standard databases (nr etc.)  rRNA/ITS databases  Genomic + transcript databases  Betacoronavirus  Experimental databases

Core nucleotide database (core\_nt) [?](#)

Organism [Optional](#)

Enter organism name or id--completions will be suggested   exclude [Add Organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [Optional](#)

Models (XM/XP)  Uncultured/environmental sample sequences

Limit to [Optional](#)

Sequences from type material

Entrez Query [Optional](#)

[YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

**Program Selection**

Optimize for  Highly similar sequences (megablast)  More dissimilar sequences (discontiguous megablast)  Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

**BLAST** [Search database core\\_nt using Megablast \(Optimize for highly similar sequences\)](#)

Show results in a new window

**+ Algorithm parameters**

[Feedback](#)

Click here

# Query sequence (ex. seq1)

Job Title	seq1
RID	4GJMWD8K016 Search expires on 06-11 22:09 pm <a href="#">Download All</a> ▾
Results for	1:lcl Query_5674738 seq1(204bp) ▾
Program	BLASTN ? <a href="#">Citation</a> ▾
Database	core_nt <a href="#">See details</a> ▾
Query ID	lcl Query_5674738
Description	seq1
Molecule type	dna
Query Length	204
Other reports	<a href="#">Distance tree of results</a> <a href="#">MSA viewer</a> ?

**Filter Results**

**Organism** only top 20 will appear  exclude  
 Type common name, binomial, taxid or group name  
[+ Add organism](#)

Percent Identity	E value	Query Coverage
<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>

[Filter](#) [Reset](#)

**Descriptions** [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

**Sequences producing significant alignments** [Download](#) ▾ [Select columns](#) ▾ [Show](#) 100 ▾ ?

select all 72 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">Brachionus angularis chromosome I mitochondrion, complete sequence</a>	<a href="#">Brachionus angu...</a>	102	102	40%	2e-17	89.16%	10764	<a href="#">MT875425.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW10 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	380	<a href="#">GQ203164.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus urceolaris voucher S. H. Cheng 008 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus urce...</a>	95.3	95.3	40%	3e-15	87.80%	379	<a href="#">FJ426637.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW6 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	379	<a href="#">GQ203212.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW11 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	380	<a href="#">GQ203198.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW15 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	380	<a href="#">GQ203169.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW1 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	380	<a href="#">GQ203163.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW10 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	379	<a href="#">GQ203197.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW11 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	45%	3e-15	85.87%	380	<a href="#">GQ203165.1</a>
<input checked="" type="checkbox"/>	<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-LW12 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	<a href="#">Brachionus calyc...</a>	95.3	95.3	40%	3e-15	87.80%	379	<a href="#">GQ203181.1</a>

“Top hit” sequence



Des	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len
<a href="#">Brachionus angularis chromosome I mitochondrion, cox1</a>	102	102	40%	2e-17	89.16%	10764
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW10_16S</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus urceolaris voucher S. H. Cheng_008_16S</a>	95.3	95.3	40%	3e-15	87.80%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW6_16S</a>	95.3	95.3	45%	3e-15	85.87%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW11_16S</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW15_16S</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW1_16S</a>	95.3	95.3	45%	3e-15	85.87%	380

Des	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len
<a href="#">Brachionus angularis chromosome I mitochondrion, complete sequence</a>	102	102	40%	2e-17	89.16%	10764
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW10 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus urceolaris voucher S. H. Cheng 008 16S rRNA</a>	95.3	95.3	40%	3e-15	87.80%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW6 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW11 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW15 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW1 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380

### Brachionus angularis chromosome I mitochondrion, complete sequence

Sequence ID: [MT875425.1](#) Length: 10764 Number of Matches: 1

Range 1: 10324 to 10405 [GenBank](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Identities	Gaps	Strand
102 bits(55)	2e-17	74/83(89%)	2/83(2%)	Plus/Plus
Query 123		TTGAGAAAGCTACTTGGGGATAACAGGGTAG-TGTTGGGGAGTCCTTATCGATGAAC		181
Sbjct 10324		TTGAGAAAGCTACTTGGGGATAACAGGGTG-AGATGTTGGAGAGTCCTTATCGATAAGC		10382
Query 182		ACAATTACGACCTCGATGTTGGA	204	
Sbjct 10383		ATAGTTACTACCTCGATGTTGGA	10405	

Des	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len
<a href="#">Brachionus angularis chromosome I mitochondrion, complete sequence</a>	102	102	40%	2e-17	89.16%	10764
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW10 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus urceolaris voucher S. H. Cheng 008 16S rRNA</a>	95.3	95.3	40%	3e-15	87.80%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW6 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	379
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-TW11 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW15 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380
<a href="#">Brachionus calyciflorus voucher AHNU-Rotifer-JW1 16S rRNA</a>	95.3	95.3	45%	3e-15	85.87%	380

### Brachionus angularis chromosome I mitochondrion, complete sequence

Sequence ID: [MT875425.1](#) Length: 10764 Number of Matches: 1

Range 1: 10324 to 10405 [GenBank](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

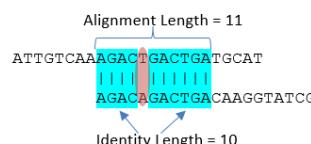
Score 102 bits(55)	Expect 2e-17	Identities 74/83(89%)	Gaps 2/83(2%)	Strand Plus/Plus
-----------------------	-----------------	--------------------------	------------------	---------------------

Query 123 TTGAGAAGCTACTTGGGGATAACAGGGTAG-TGTTCGGGAGTCCTTATCGATGAAC 181

Sbjct 10324 TTGAGAAGCTACTTGGGGATAACAGGGTG-AGATGTGAGAGTCCTTATCGATAAGC 10382

Query 182 ACAATTACGACCTCGATGTTGGA 204

Sbjct 10383 ATAGTTACTACCTCGATGTTGGA 10405



$$\text{Alignment Identity \%} = \frac{\text{Identity Length}}{\text{Alignment Length}} = \frac{10}{11}$$

$$\text{Query Identity \%} = \frac{\text{Identity Length}}{\text{Query Length}} = \frac{10}{25}$$

$$\text{Query Coverage \%} = \frac{\text{Alignment Length}}{\text{Query Length}} = \frac{11}{25}$$

$$\text{Subject Identity \%} = \frac{\text{Identity Length}}{\text{Subject Length}} = \frac{10}{21}$$

$$\text{Subject Coverage \%} = \frac{\text{Alignment Length}}{\text{Subject Length}} = \frac{11}{21}$$

## Seq1

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<a href="#">Brachionus angularis chromosome I mitochondrion, complete sequence</a>	Brachionus angu...	102	102	40%	2e-17	89.16%	10764	MT875425.1



89.16%

Rotifer (*Brachionus angularis*??)

## Seq100

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<a href="#">Bufo bufo isolate Toad_3537 16S ribosomal RNA gene, partial sequence; mitochondrial</a>	Bufo bufo	553	553	100%	7e-153	99.67%	565	MF461235.1



99.67%

Toad (*Bufo bufo*)

## Seq126

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<a href="#">Myocastor coypus mitochondrion, complete genome</a>	Myocastor coypus	508	508	100%	1e-139	99.64%	16874	MH182628.1



99.64%

Nutria (*Myocastor coypus*)

National Library of Medicine  
National Center for Biotechnology Information

Log in

Taxonomy Taxonomy Limits Advanced Search Help



## Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

### Using Taxonomy

[Quick Start Guide](#)  
[FAQ](#)  
[Handbook](#)  
[Taxonomy FTP](#)  
[Important Update: Phyla Changing](#)  
[Important Update: New Flu species Names](#)  
[Important Update: New ranks replaced superkingdom](#)

### Taxonomy Tools

[Browser](#)  
[Common Tree](#)  
[Statistics](#)

### Other Resources

[GenBank](#)  
[LinkOut](#)  
[E-Utilities](#)

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC

Search for  as   lock

Display  levels using filter:

**Bufo bufo**

Taxonomy ID: 8384 (for references in articles please use NCBI:txid8384)

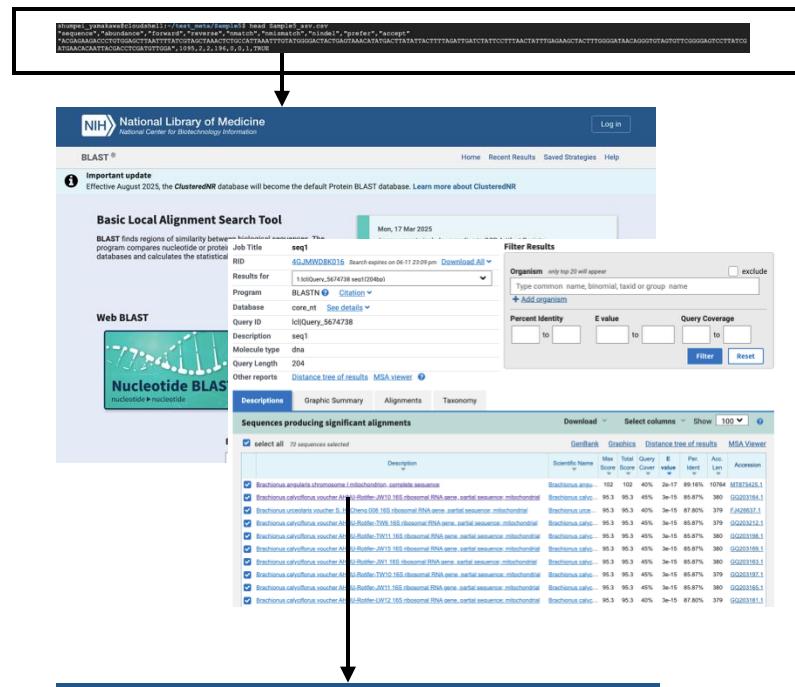
current name  
**Bufo bufo** (Linnaeus, 1758)  
[basionym: **Rana bufo** Linnaeus, 1758]

Genbank common name: common toad  
NCBI BLAST name: frogs & toads  
Rank: species  
Genetic code: [Translation table 1 \(Standard\)](#)  
Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)  
Other names:  
common name(s)  
**European toad, common European toad**

[Lineage \(full\)](#)  
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amphibia; Batrachia; Anura; Neobatrachia; Hyloidea; Bufonidae; Bufo

## Command line

```
shupsei_yamatowa@cloudshell:~/test_mate$ sample5_head Sample5.csv
"sequence","abundance","order","reverse","name","misslabel","bind1","prefer","accept"
"ACGAGACGACCTCGATCTTATTTATCAGTCAGTAACATCTAAATTATGATGGCTACTGAGTAACATATGTTATATTCTTGTAGTGATCTTAACTATTTGAGAGCTACTTGGGATAACAGGGTAGTGCTGGGGAGCTTAA
```



**National Library of Medicine**  
National Center for Biotechnology Information

Taxonomy Taxonomy Search Help

**Taxonomy**

**Taxonomy**

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

**Using Taxonomy**

Quick Start Guide  
FAQ  
Handbook  
Taxonomy FTP  
Important Update: Phyla Chango  
Important Update: New Floc species Names  
Important Update: New ranks replaced super kingdom

**Taxonomy Tools**

Browser  
Common Tree  
Statistics  
Microbial Taxonomy

**Search** Enter:  **Pubmed** **Prokaryotes** **Nucleotide** **Protein** **Genome** **Structure** **PMC**

Display: [3] levels using filter: none  compact name  lock

**Bufo bufo**

Taxonomy ID: 8384 (for references in articles please use NCBI税ID8384)  
current name  
**Bufo bufo** (Linnaeus, 1758)

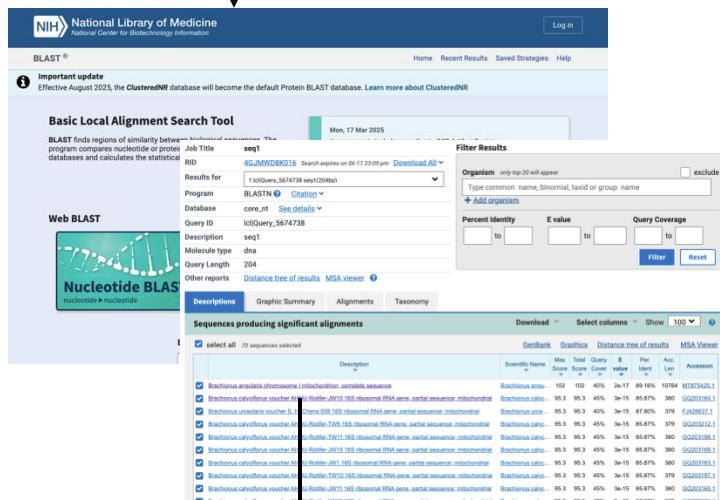
Gruber, common name: common toad  
NCBI BLAST name: Frogs & toads  
Rati, species:  
Mitochondrial: [Translation table 1 \(Standard\)](#)  
Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)  
Other names (synonyms/names):  
**European toad, common European toad**

LncRNA (full)

Saccharomyces cerevisiae: Eukaryota: Opisthokonta: Mucoromycota: Entomogastrida: Deuterostomia: Ciliophora: Choanoflagellata: Teleostei: Euchlorostomi: Sacculinomycetidae: Saccromycetidae: Teredidae: Amphibia: Batrachia: Bilateria: Neobrachia: Hylodidae: Bafidae: Bufidae

## Command line

```
blmpej_1_yanakal@cloudshell:~/Desktop$ samtools head Sample3.csv.vcf
```



The screenshot shows the homepage of the National Library of Medicine's Taxonomy Database. At the top, there is a navigation bar with links for NIH, National Library of Medicine, and National Center for Biotechnology Information, along with a "Log in" button. Below the navigation bar is a search bar with dropdown menus for "Taxonomy" and "Search". The main content area features a large image of butterflies on the left and a central section titled "Taxonomy" with a sub-section "The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet." On the left sidebar, there are sections for "Using Taxonomy" (Quick Start Guide, FAQ, Handbooks, Taxonomy FTP), "Taxonomy Tools" (Browser, Common Tree, Statistics), and "Other Resources" (GenBank, LinkOut, Utilities). A footer at the bottom includes the NCBI logo, links for Taxonomy Browser, Entrez, PubMed, Nucleotide, Structure, and Genomes, and a "PAC" link.

## Command line

```
shumpel_yamakawa@cloudshell:~/test_mate$ samples$ head Samples_asv.csv
"sequence","abundance","match","mismatch","index","prefer","accept"
"ACGAGAAGACCTTGTGACCTTAATTTATCGTACTGTCATCCTGCAATTAAATTGTATGGGAGTACTGAGTAAACATACTTAACTTATCTTGTGATCTTACCTTGTGAGCTACTTGCGGAAACACGGGTGTAGTGTGGGGAGTCTTAT
```

# Shell scripts

## BLAST and taxonomy search

# Day2

I am providing a customized script for performing a BLAST search against 35,267 metazoan 16S rRNA sequences and extracting taxonomic information (NCBI taxonomy, including phylum and class). Please note that the script is roughly designed and not formatted for publication-level use, so you may encounter missing or blank fields in some cases.

## 1. Set up

```
cd ~/test_meta
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/blast_annot_setup.sh
sh blast_annot_setup.sh
```

## 2. Copy the csv file to "16s\_metazoa\_rrna\_blast\_annot" directory

```
cp ~/test_meta/Sample5/Sample5_asv.csv ~/test_meta/16s_metazoa_rrna_blast_annot/
```

## 3. BLAST and extract taxonomy information

```
#move to the analysis directory
cd ~/test_meta/16s_metazoa_rrna_blast_annot/
ls
#Check if the csv file that you want to analyze is in the directory!!

sh 16s_metazoa_rrna_blast_annot.sh
## You will see the message "Enter an ASV csv file"
## Type the name of csv file that you want to analyze

## Then you will see the message "Enter a threshold of species determination (identity of blast tophit):"
## Type the number (ex. 95)

## Then the analysis will start

=====
## The following is an example of the results...

#Enter an ASV csv file: Sample5_asv.csv
#Enter a threshold of species determination (identity of blast tophit): 95
###
#Setting

#blastn seqrhc / database: 35267 seqs of 16S ribosomal RNA from NCBI Genbank
#Extract taxonomic information from NCBI Taxonomy (using taxonkit: doi 10.1016/j.jgg.2021.03.006)

###
#Analysis

#Create a fasta file
```

See Day2.md!

## 1. Set up

```
cd ~/test_meta  
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/blast_annot_setup.sh  
sh blast_annot_setup.sh
```

## 2. Copy the csv file to "16s\_metazoa\_rrna\_blast\_annot" directory

```
cp ~/test_meta/Sample5/Sample5_asv.csv ~/test_meta/16s_metazoa_rrna_blast_annot/
```

~/test\_meta/

~/test\_meta/Sample5

Sample5\_asv.csv



You

~/test\_meta/16s\_metazoa..



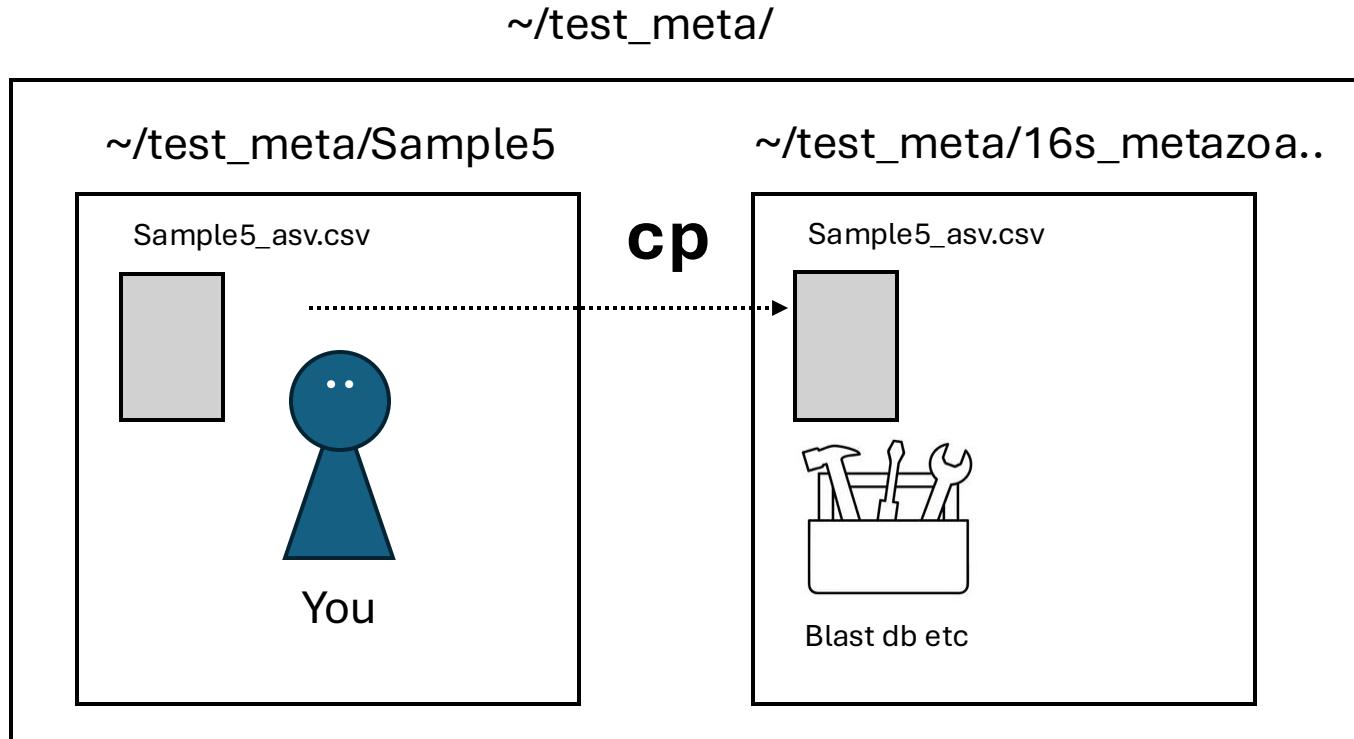
Blast db etc

## 1. Set up

```
cd ~/test_meta  
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/blast_annot_setup.sh  
sh blast_annot_setup.sh
```

## 2. Copy the csv file to "16s\_metazoa\_rrna\_blast\_annot" directory

```
cp ~/test_meta/Sample5/Sample5_asv.csv ~/test_meta/16s_metazoa_rrna_blast_annot/
```



### 3. BLAST and extract taxonomy information

```
#move to the analysis directory  
cd ~/test_meta/16s_metazoa_rrna_blast_annot/  
ls  
#Check if the csv file that you want to analyze is in the directory!!
```

~/test\_meta/

~/test\_meta/Sample5

Sample5\_asv.csv



You

~/test\_meta/16s\_metazoa..

Sample5\_asv.csv

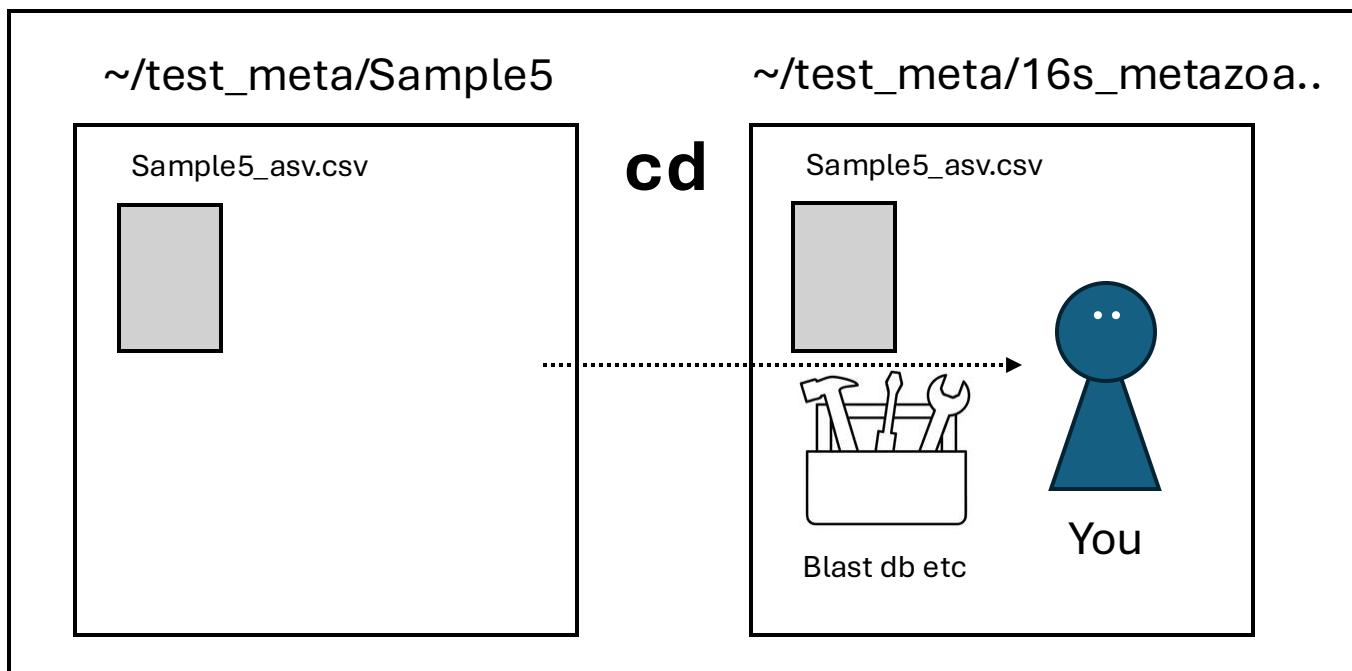


Blast db etc

### 3. BLAST and extract taxonomy information

```
#move to the analysis directory  
cd ~/test_meta/16s_metazoa_rrna_blast_annot/  
ls  
#Check if the csv file that you want to analyze is in the directory!!
```

~/test\_meta/



### 3. BLAST and extract taxonomy information

```
#move to the analysis directory  
cd ~/test_meta/16s_metazoa_rrna_blast_annot/
```

```
ls
```

```
#Check if the csv
```

```
ot$ ls  
16s_metazoa_rrna_blast_annot.sh  ncbi_s16.db.nhr  ncbi_s16.db.nto  
dehnodes.dmp                  ncbi_s16.db.nin  nodes.dmp  
list                           ncbi_s16.db.njs  
merged.dmp                     ncbi_s16.db.not  
names.dmp                      ncbi_s16.db.nsq  
ncbi_s16.db.ndb                ncbi_s16.db.ntf
```

~/test\_meta/

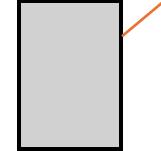
~/test\_meta/Sample5

Sample5\_asv.csv



~/test\_meta/16s\_metazoa..

Sample5\_asv.csv



Blast db etc



You

Sample5\_asv.csv  
**taxonkit**

taxonkit

```

sh 16s_metazoa_rrna_blast_annot.sh
## You will see the message "Enter an ASV csv file"
## Type the name of csv file that you want to analyze

## Then you will see the message "Enter a threshold of species determination (identity of blast tophit):"
## Type the number (ex. 95)

## Then the analysis will start

```

#####  
Results

Abundance: total 31661 reads

ASVs: 126

16 species was identified

Phylum	Class	Species	blast_tophit	blast_ident	abundance	seqs	
Arthropoda	Branchiopoda	Daphnia longispina	JN874595.1	97.521	128	seq63	
Arthropoda	Branchiopoda	Eubosmina cf.	EU650685.1	98.347	67	seq93	
Arthropoda	Branchiopoda	Scapholeberis mucronata	EF189615.1	98.326	33	seq114	
Arthropoda	Branchiopoda	Simocephalus vetulus	LC382447.1	97.531	36	seq113	
Arthropoda	Insecta	Cloeon dipterum	LC801945.1	96.680	43	seq109	
Bryozoa	Phylactolaemata	Plumatella repens	DQ305341.1	98.438	31	seq115	
Chordata	Actinopteri	Carassius auratus	DQ868870.1	99.674	59	seq101	
Chordata	Actinopteri	Gasterosteus aculeatus	DQ027919.1	99.340	135	seq59	
Chordata	Actinopteri	Leucaspis delineatus	NC_020357.1	99.342	68	seq92	
Chordata	Actinopteri	Pseudorasbora interrupta	MN175390.1	99.342	5	seq125	
Chordata	Amphibia	Bufo bufo	JN647011.1	99.669	202	seq54, seq100	
Chordata	Amphibia	Pelophylax lessonae	MH105105.1	99.656	16	seq118	
Chordata	Amphibia	Rana temporaria	KC977158.1	100.000	96	seq80	
Chordata	Aves	Gallinula chloropus	DQ485864.1	98.635	38	seq112	
Chordata	Mammalia	Myocastor coypus	AF422886.1	99.281	4	seq126	
Rotifera	Eurotatoria	Keratella quadrata	AF499046.1	99.010	4439	seq6, seq8, seq18, seq2	

Output files are in Sample5\_asv.csv\_95

Output files are in Sample5\_asv.csv\_95

## Fasta.file

```
>seq1
ACGAGAAAGACCCCTGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACATGACTTATATTACTTTAGATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq2
ACGAGAAAGACCCATGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACATGACTTATATTACTTTAGATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq3
ACGAGAAAGACCCCTGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTATATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq4
ACGAGAAAGACCCATGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTATATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq5
CGACGCCTTCGATCTACGAGAAAGACCTATGGAGCTTACTTAAGGTTAACCTACA
AGTTAAATGGAACTTTGGAGAACAAAATACCTCGCCGACCTACTCTTATTCG
GAGAAGACTACTTGGGATAACAGGGTAATACAGCTGGGAGCACTTATCGATAGTTGTT
CTTACGACCTCGATGTTGATGTCGAGACAC
>seq6
ACGAGAAAGACCCCTGGAGCTTAATTTCGCGCTGTTTATACCTATGCCCGTATACAA
ATTACATGGGACTGTTGAGTGAGAAATAATTGAAACTCTTACTATTGATCTAAC
TCTTAATTATCTGAAAGCTACTTGGGATAACAGGGTGAGCGAGTGGAGAGTTCTTA
TCGATACTCTGCAATTACGACCTCGATGTTGGA
>seq7
ACGAGAAAGACCCCTGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTATATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq8
ACGAGAAAGACCCCTGGAGCTTAATTTCGCGCTGTTTATACCTATGCCCGTATACAA
ATTACATGGGACTGTTGAGTGAGAAATAATTGAAACTCTTACTATTGATCTAAC
TCTTAATTATCTGAAAGCTACTTGGGATAACAGGGTGAGCGAGTGGAGAGTTCTTA
TCGATACTCTGCAATTACGACCTCGATGTTGGA
>seq9
ACGAGAAAGACCCCTGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACATGACTTATATTACTTTAGATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq10
CGACGCCTTCGATCTACGAGAAAGACCCCTGGAGCTTACTTAAGGTTAACCTACA
AGTTAAATGGAACTTTGGAGAACAAAATACCTCGCCGACCTACTCTTATTCG
GAGAAGACTACTTGGGATAACAGGGTAATACAGCTGGGAGCACTTATCGATAGTTGTT
CTTACGACCTCGATGTTGATGTCGAGACAC
>seq11
ACGAGAAAGACCCATGGAGCTTAATTTCGAGCTAAACTCTGCCATTAATTGTAT
GGGGACTACTGAGTAAACATGACTTATATTACTTTAGATTGATCTTCCCTTAAC
ATTGAGAAAGCTACTTGGGATAACAGGGTGAGTGTTCGGGGAGTCCTTATCGATGAA
```

## Blast results

```
# BLASTN 2.15.0+
# Query: seq1
# Database: ncbi_s16.db
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, eval, bit score
# 121 hits found
seq1 GQ283163.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283164.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283165.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283166.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283167.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283168.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283169.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283170.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283171.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283172.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283173.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283174.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283175.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283176.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283177.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283181.1 87.805 82 8 2 124 284 288 368 1.4e-18 95.3
seq1 GQ283196.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283197.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283198.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283208.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283208.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283202.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283202.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
seq1 GQ283205.1 85.870 92 10 3 114 284 288 369 1.4e-18 95.3
```

## Species name

AF499046.1 *Keratella quadrata*  
AF422886.1 *Myocastor coypus*  
LCB01945.1 *Cloeon dipterum*  
MH105105.1 *Pelophylax lessonae*  
LC382447.1 *Simocephalus vetulus*  
KC977158.1 *Rana temporaria*  
JN874595.1 *Daphnia longispina*  
JN647011.1 *Bufo bufo*  
DQ027919.1 *Gasterosteus aculeatus*  
EU650685.1 *Eubosmina cf.*  
EF189615.1 *Scapholeberis mucronata*  
DQ868870.1 *Carassius auratus*  
DQ485864.1 *Gallinula chloropus*  
DQ305341.1 *Plumatella repens*  
NC\_020357.1 *Leucaspis delineatus*  
MN175390.1 *Pseudorasbora interrupta*

Etc...





*Myocastor coypus*



*Gallinula chloropus*



*Cloeon dipteron*



*Bufo bufo*



*Daphnia longispina*



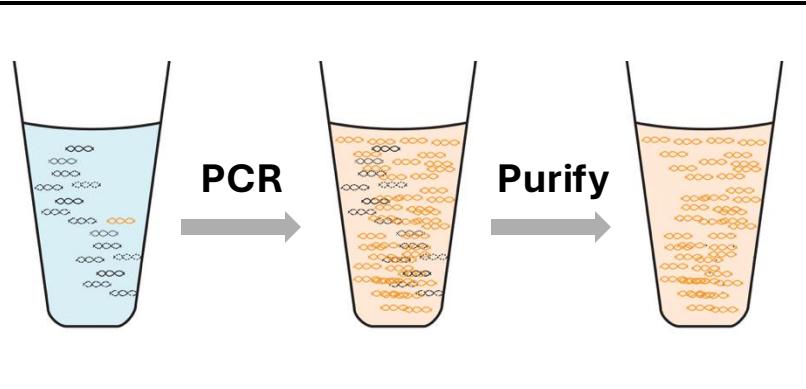
*Scapholeberis mucronata*



*Plumatella repens*



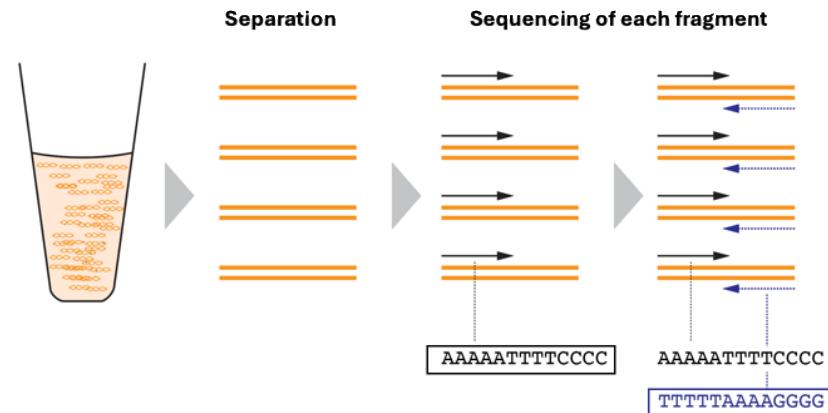
*Gasterosteus aculeatus*



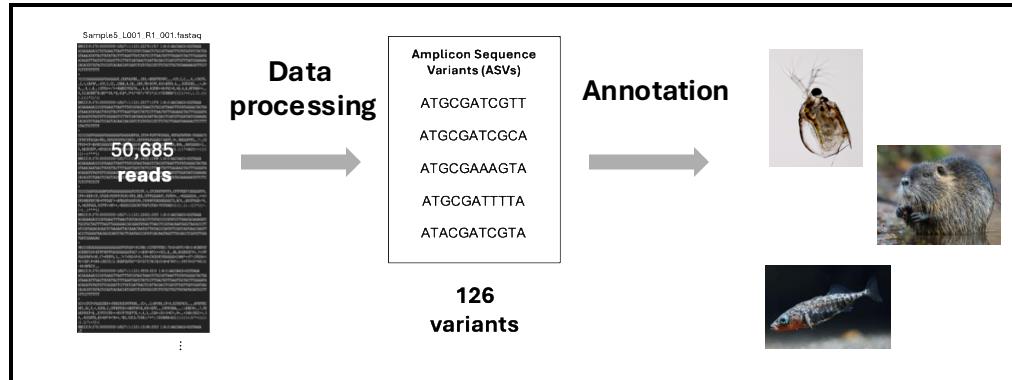
Send to the company or  
institute...



## NGS analysis



Receive the sequencing data



-12:30

## 1. Annotation of test data (Sample 5)

## 2. Analysis of another dataset (Sample 6)

### Tasks

Please profile the animal species in the following two freshwater datasets (Sample 6 and Sample 7). The sampling and sequencing methods are the same as those used for the test data (Sample 5). Follow the scripts you used on Day 1 and Day 2.

1. Create a new directory for each dataset under the test\_meta directory (e.g., Sample6 and Sample7).
2. Move to each directory and download the sequence datasets (see details below).
3. Load the data into R and check the quality using dada2.
4. Generate and save an ASV file for each dataset using dada2.
5. Copy the ASV files to the 16s\_metazoa... directory.
6. Annotate the sequences.
7. Try processing another sample.

```
#Sample6
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample6_L001_R1_001.fq
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample6_L001_R2_001.fq
```

```
#Sample7
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample7_L001_R1_001.fq
wget https://github.com/ShumpeiYamakawa/FSUJENA_2025_species_determination/raw/refs/heads/main/Sample7_L001_R2_001.fq
```

(Day2.md)