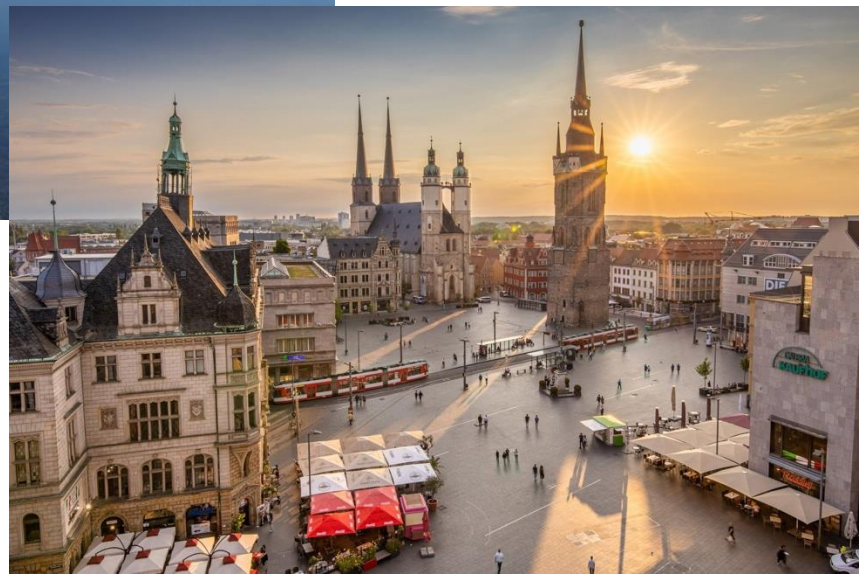
Sommersemester 2025
Vertiefte Bestimmungsübungen an Tieren (MEES003/C3)

Environmental DNA (eDNA) Metabarcoding Analysis

Day 1

Shumpei Yamakawa



Day 1 Task:

eDNA Metabarcoding for identifying animals living in a pond near Halle

2. Sequencing

3. Annotation

ATGGGCTA... → Species A

CCGGCTA... → Species B

ATATAAATA... → Species C

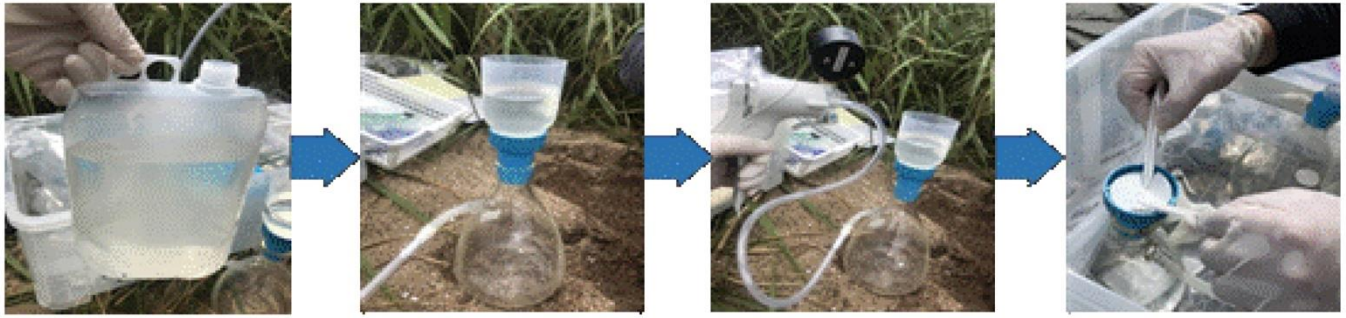
GGGTTTATT... → Species D

GTTAACCC... → Species E

1. Sampling

500 mL sample

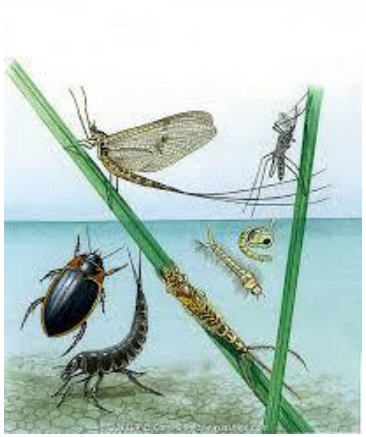
Field Sampling

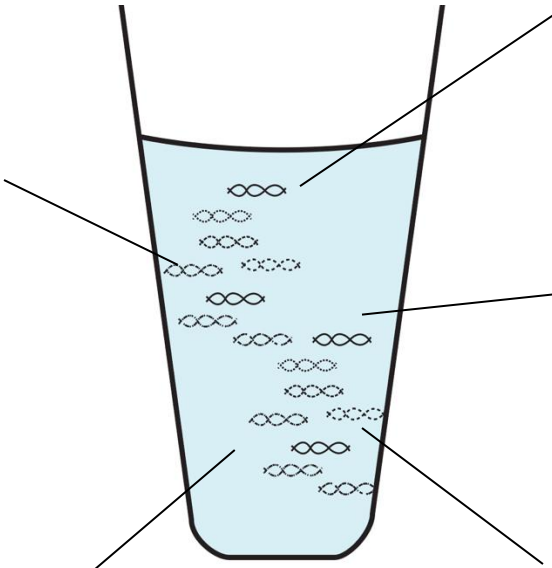


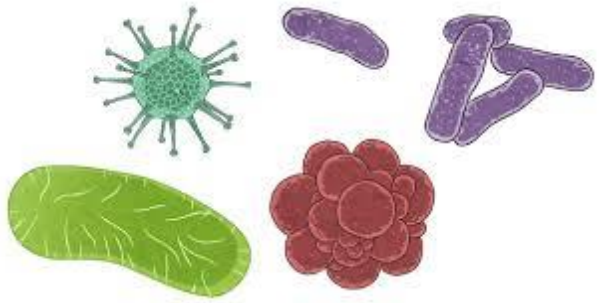
Kim et al., 2019

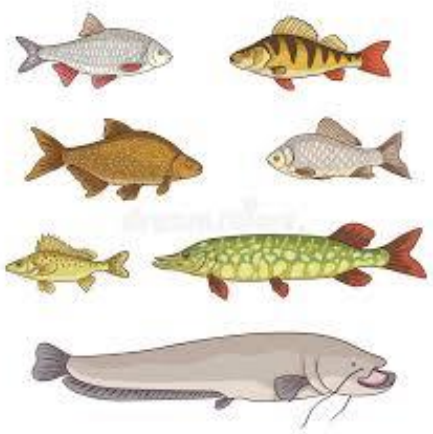
1. Sampling

The sample contains DNA from various organisms



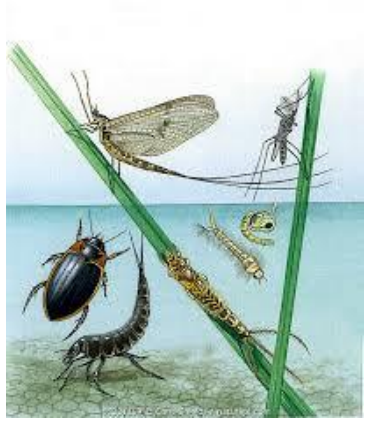


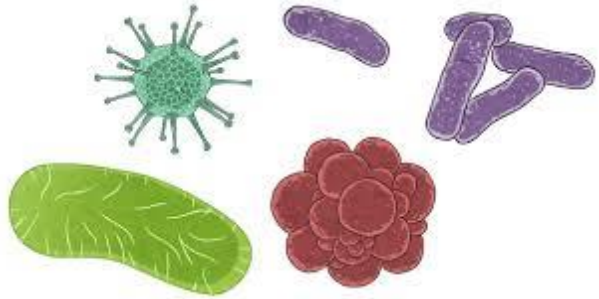


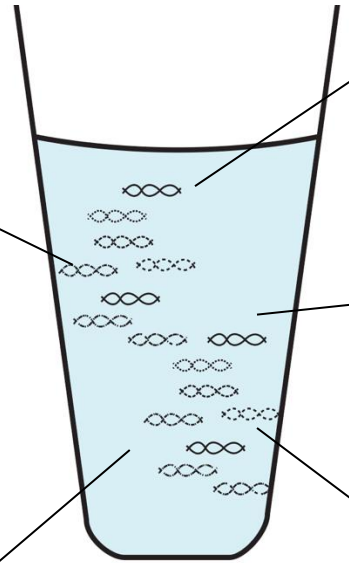


1. Sampling

The sample contains DNA from various organisms

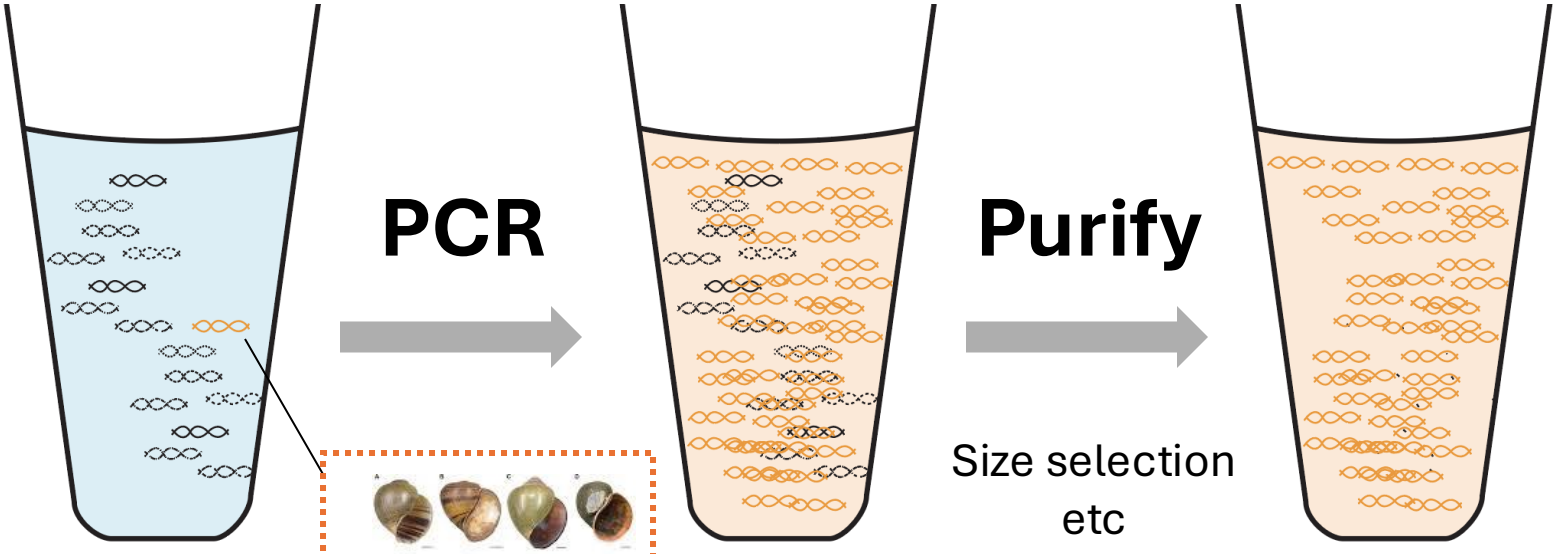






1. Sampling

Amplification of specific sequences



primer

16S rrna

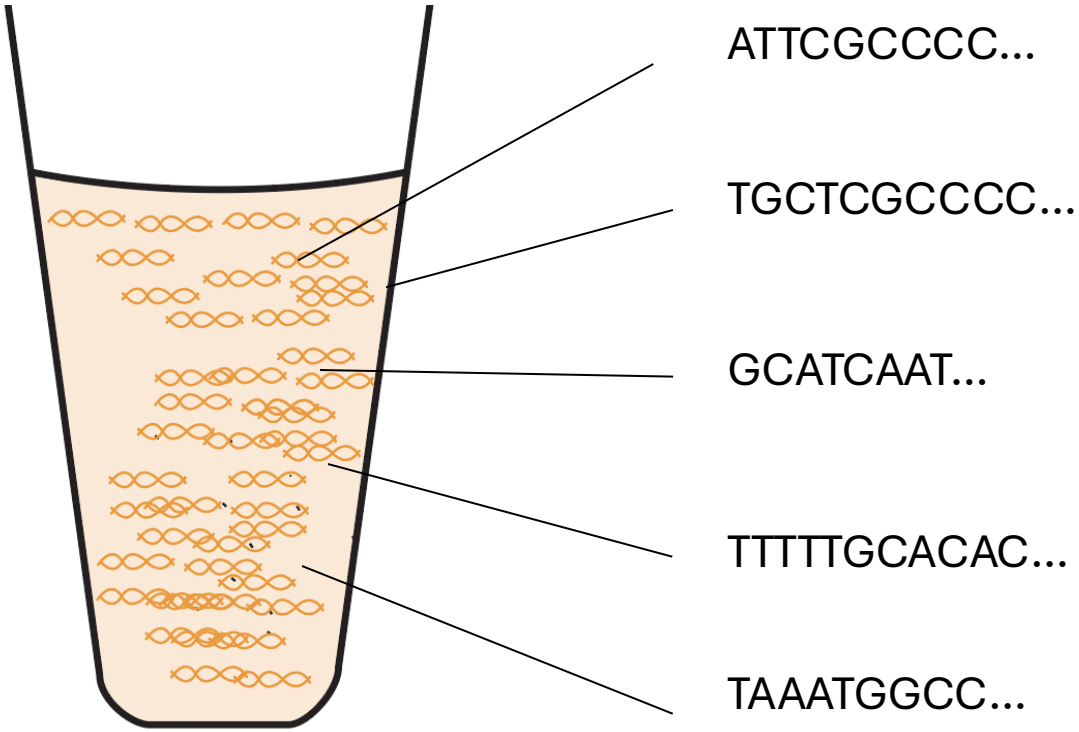
(200–300bp)

Target taxa = Animals

2. Sequencing

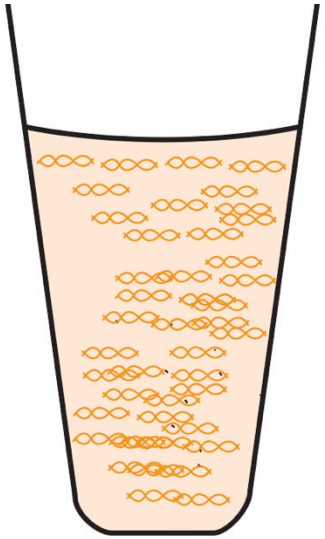
Sequencing of various DNA fragments

How?



2. Sequencing

High-throughput sequencing using Next Generation Sequencer



NGS

ATTCGCCCC…

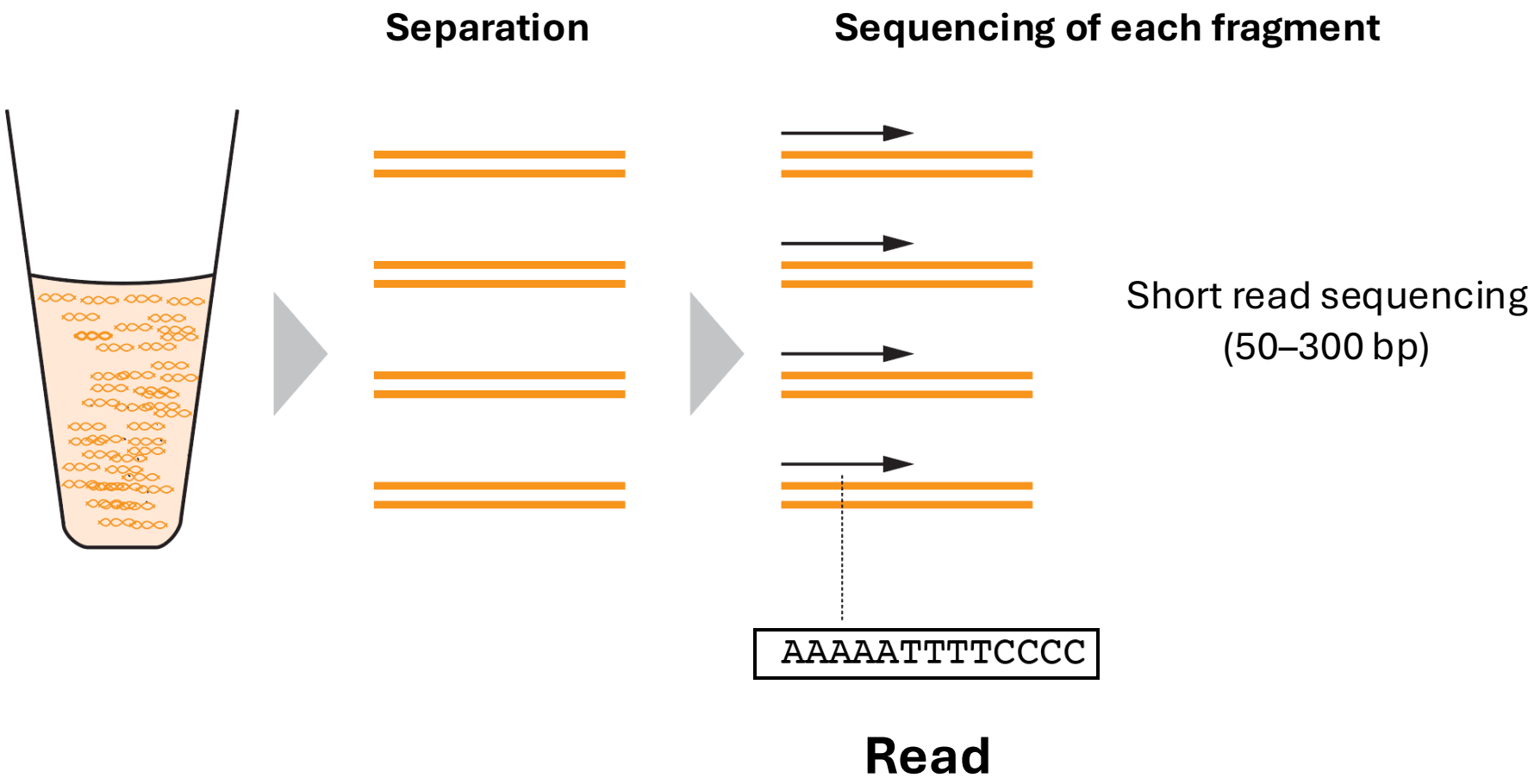
TGCTCGCCCC…

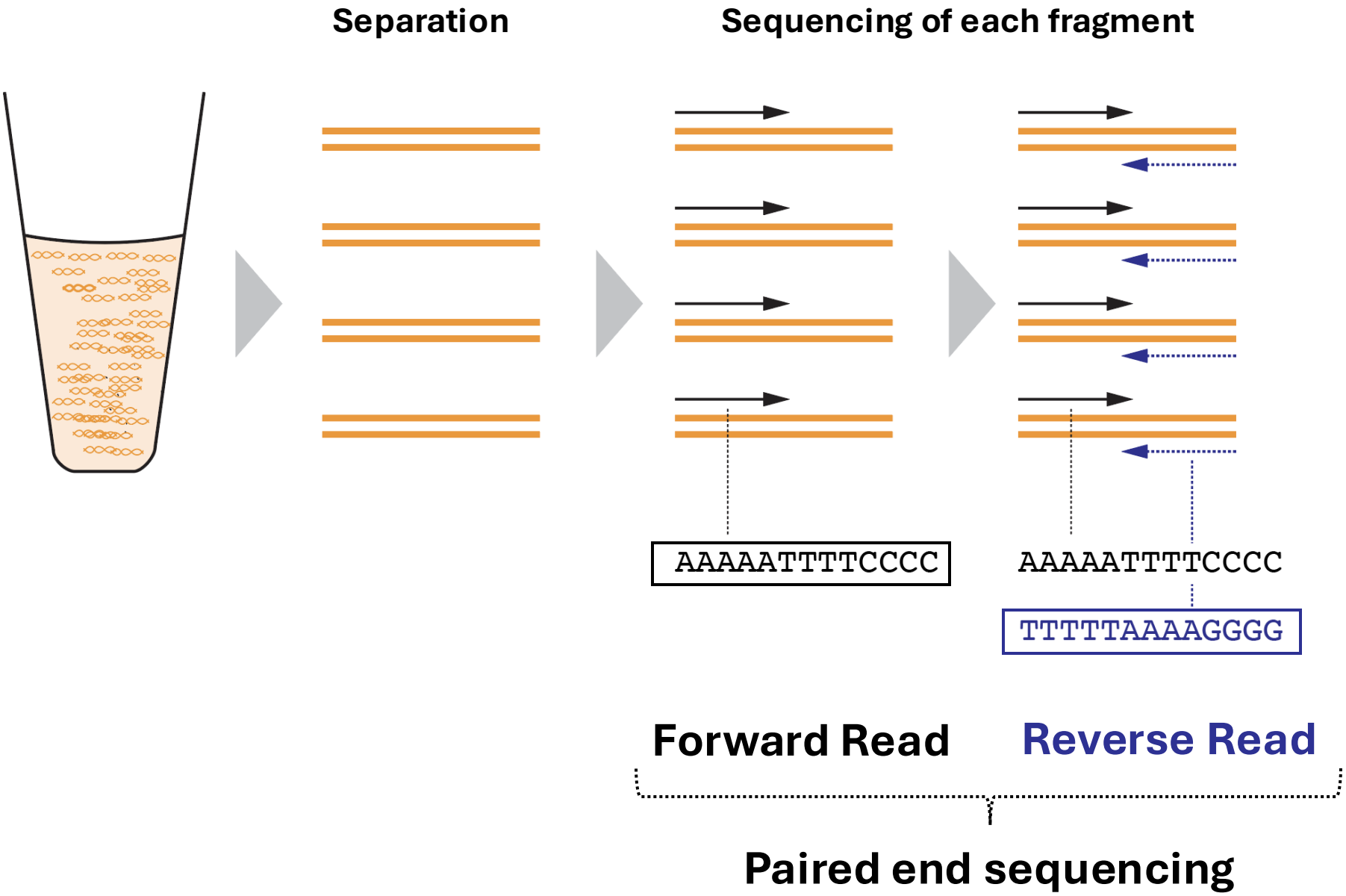
GCATCAAT…

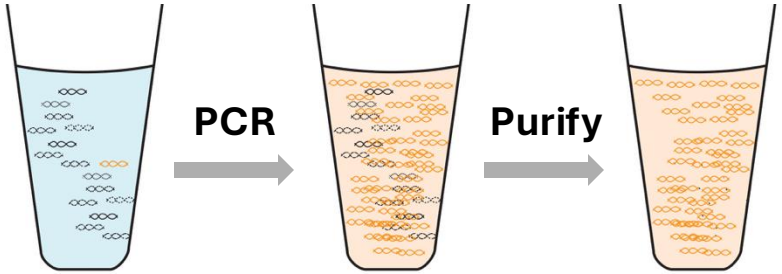
TTTTTGCACAC…

TAAATGGCC…

⋮

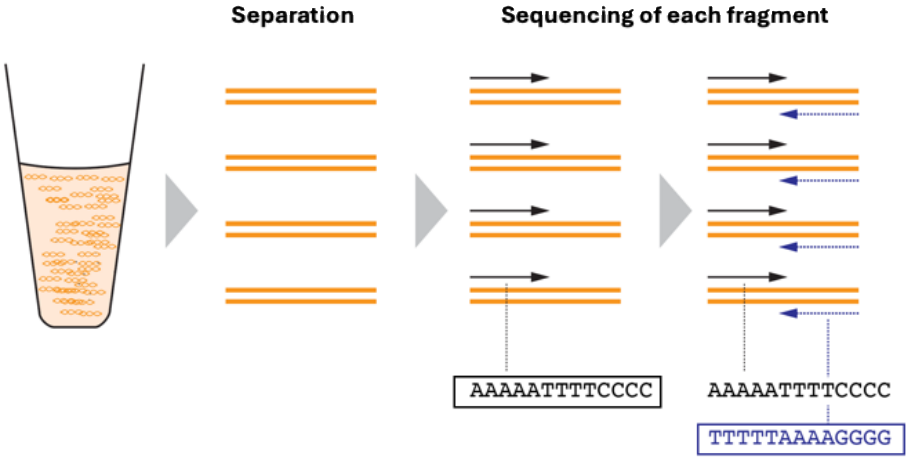

Separation
Sequencing of each fragment
AAAAATTTTCCCC
Read
Short read sequencing
(50–300 bp)


Separation
Sequencing of each fragment
AAAAATTTTCCCC
AAAAATTTTCCCC
TTTTTAAAAGGGG
Forward Read
Reverse Read
Paired end sequencing

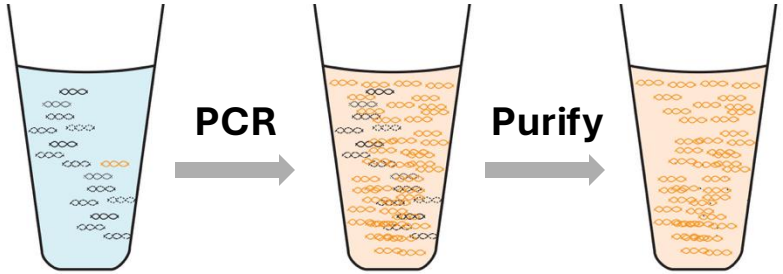


Send to a company or institute...

NGS analysis

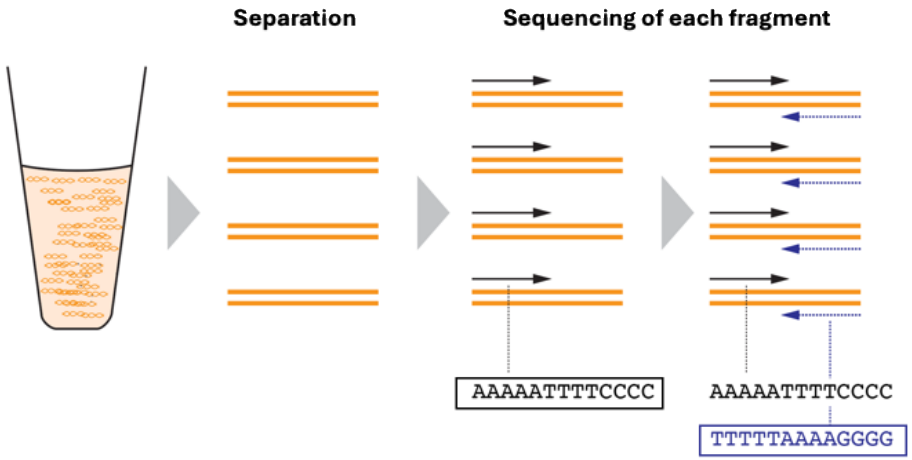


Receive the sequencing data



NGS analysis on your laptop!

NGS analysis



NGS analysis on your laptop!

Sequence file

What does this file look like?

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
ACGAGAAGACCCTGTGGAACTTAATTTTATCGTATCTAAACTCTGCCATTAAATTTGTATG
GTGTCTACTGAGTAAACATATTACTTATATTACTTTTAGATTTATCTATTCCTTTAACTAT
TTTAGAATCTACTTTGGGGATAACAGGGTTTAGTGTTCGGGGTTTCCTTATCGATGAACTC
AATTACGACCTCGATGTTGTTTGATCGGAAGAGCACACGTCTGTACTCCCGTCACAACCAT
CGATCTCGTCTGCCGTCTTCTGCTTGCTATAAAAAACATTTCCTTCTTTTTTTTTT
+
?CCCCGGGGGGGGGFGGGGGGGF,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)/((
@M02319:279:000000000-L8DJ7:1:1101:20377:1978 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATG
GGGACTACTGAGTAAACATATGACTTATGTTACTTTTAGATTGATCTATTCCTTTAACTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGAAATAAAAAACTTCTTTTCTACTTCTTTTT
+
CCCCCGGFFGGGGGFGGGGGGGGFGGGGGGEFG8,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@FEFFFEA3).)))7+6A20)+-)))))))--)****))
@M02319:279:000000000-L8DJ7:1:1101:24081:1998 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAGACTCTACCATTAAATTTGTATG
GGGACTACTGAGTAAACAAATGACTTATATTACTTTTAGATTGATCTATTCCTTTAATTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGTACAAAAAAATATTCTTCTCTCTTTCTCTT
```

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

@M02319:279:000000000-L8DJ7:1:1101:22276:1917 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCTGTGGAACTTAATTTTATCGTATCTAAACTCTGCCATTAAATTTGTATG
GTGTCTACTGAGTAAACATATTACTTATATTACTTTTAGATTTATCTATTCCTTTAACTAT
TTTAGAATCTACTTTGGGGATAACAGGGTTTAGTGTTCGGGGTTTCCTTATCGATGAACTC
AATTACGACCTCGATGTTGTTTGATCGGAAGAGCACACGTCTGTACTCCCGTCACAACCAT
CGATCTCGTCTGCCGTCTTCTGCTTGCTATAAAAAACATTTCCTTCTTTTTTTTTT
+
?CCCCGGGGGGGGGFGGGGGGGF,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)/((
@M02319:279:000000000-L8DJ7:1:1101:20377:1978 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATG
GGGACTACTGAGTAAACATATGACTTATGTTACTTTTAGATTGATCTATTCCTTTAACTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGAAATAAAAAACTTCTTTTCTACTTCTTTTT
+
CCCCCGGFFGGGGGFGGGGGGGGFGGGGGGEFG8,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@FEFFFEA3).)))7+6A20)+-)))))))--)****))
@M02319:279:000000000-L8DJ7:1:1101:24081:1998 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAGACTCTACCATTAAATTTGTATG
GGGACTACTGAGTAAACAAATGACTTATATTACTTTTAGATTGATCTATTCCTTTAATTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGTACAAAAAAATATTCTTCTCTCTTTCTCTT
```

≒ 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

@M02319:279:000000000-L8DJ7:1:1101:22276:1917 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCTGTGGAACTTAATTTTATCGTATCTAAACTCTGCCATTAAATTTGTATG
GTGTCTACTGAGTAAACATATTACTTATATTACTTTTAGATTTATCTATTCCTTTAACTAT
TTTAGAATCTACTTTGGGGATAACAGGGTTTAGTGTTCGGGGTTTCCTTATCGATGAACTC
AATTACGACCTCGATGTTGTTTGATCGGAAGAGCACACGTCTGTACTCCCGTCACAACCAT
CGATCTCGTCTGCCGTCTTCTGCTTGCTATAAAAAACATTTCCTTCTTTTTTTTTT
+
?CCCCGGGGGGGGGFGGGGGGGF,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)/((
@M02319:279:000000000-L8DJ7:1:1101:20377:1978 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATG
GGGACTACTGAGTAAACATATGACTTATGTTACTTTTAGATTGATCTATTCCTTTAACTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGAAATAAAAAACTTCTTTTCTACTTCTTTTT
+
CCCCCGGFFGGGGGFGGGGGGGGFGGGGGGEFG8,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@FEFFFEA3).)))7+6A20)+-)))))))--)****))
@M02319:279:000000000-L8DJ7:1:1101:24081:1998 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCCGTGGAGCTTAATTTTATCGTAGCTAGACTCTACCATTAAATTTGTATG
GGGACTACTGAGTAAACAAATGACTTATATTACTTTTAGATTGATCTATTCCTTTAATTAT
TTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAACAC
AATTACGACCTCGATGTTGGATGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAA
CGATCTCGTATGCCGTCTTCTGCTTGTACAAAAAAATATTCTTCTCTCTTTCTCTT
```

≒ sequence id (ex. read1)

sequence information

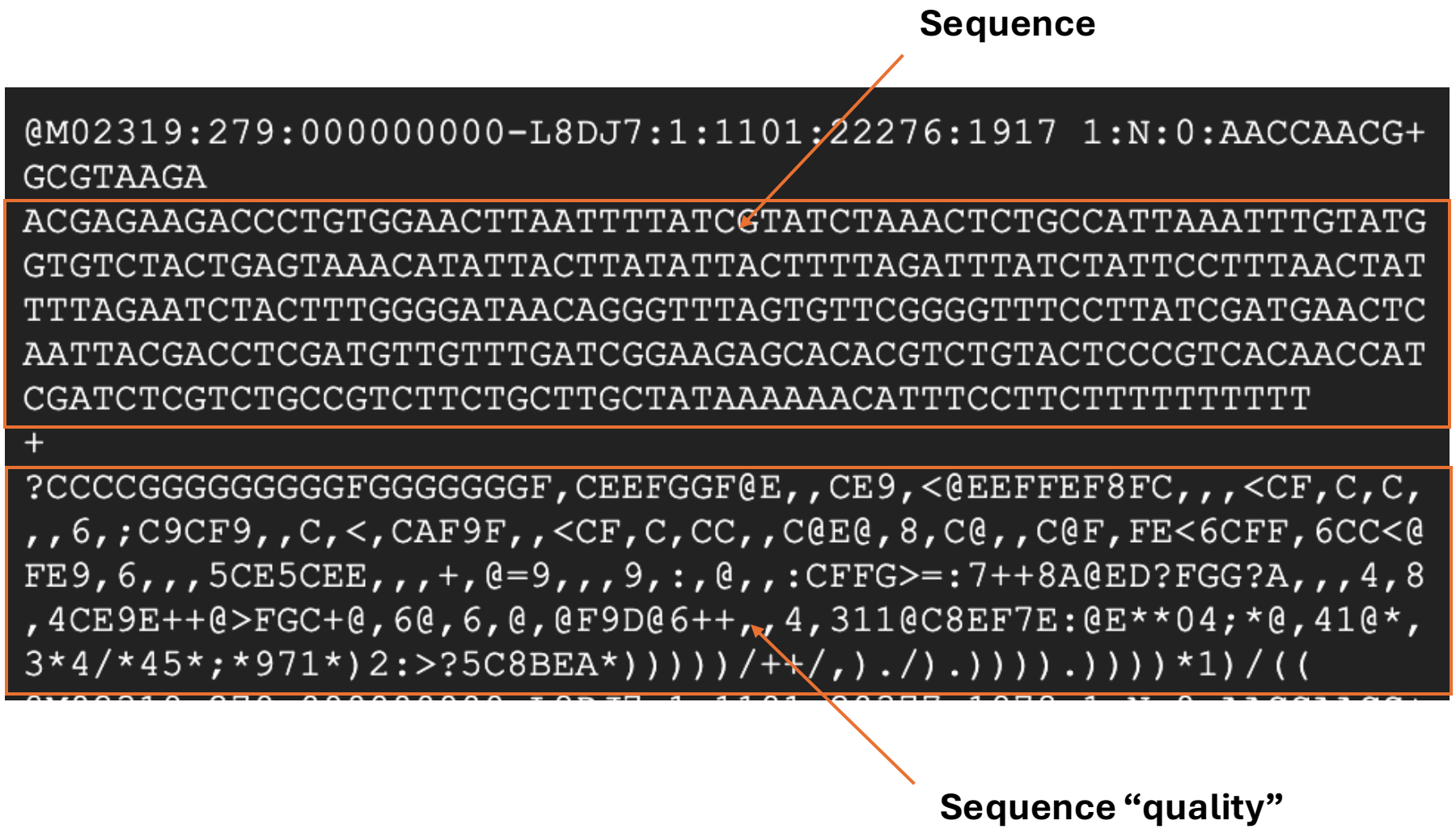
≒ sequence id (ex. read2)

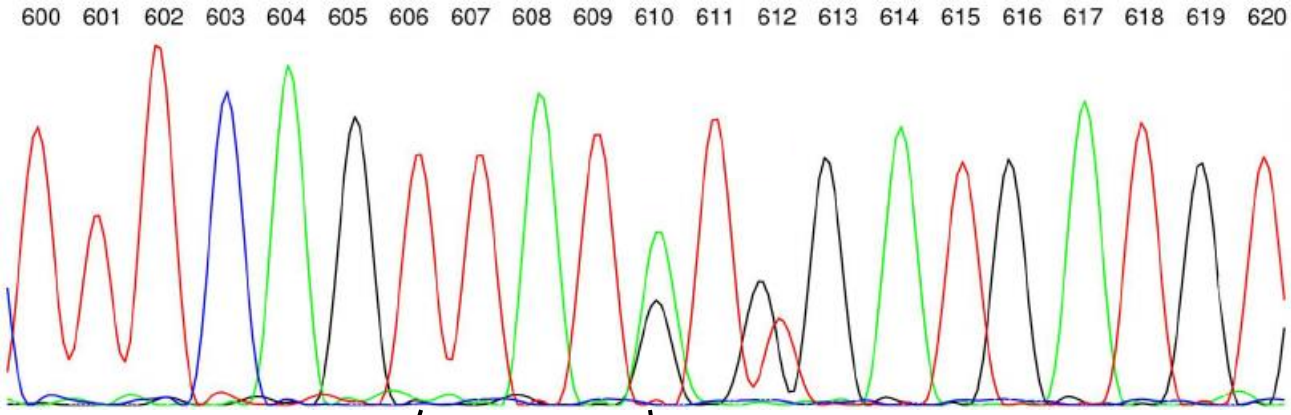
sequence information

≒ sequence id (ex. read3)

sequence information

Sample5_L001_R1_001.**fastq**





(*sanger sequencing)

	T	A
Base calling		
Probability of error (P)	0.001	0.1
Quality score (Q)	30	10

Q = -10 × log10(P)

(+33) ASCII

? +

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

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

Sample5_L001_R1_001.**fastq**

```
@M02319:279:000000000-L8DJ7:1:1101:22276:1917 1:N:0:AACCAACG+
GCGTAAGA
ACGAGAAGACCCTGTGGAACTTAATTTTATCGTATCTAAACTCTGCCATTAAATTTGTATG
GTGTCTACTGAGTAAACATATTACTTATATTACTTTTAGATTTATCTATTCCTTTAACTAT
TTTAGAATCTACTTTGGGGATAACAGGGTTTAGTGTTCGGGGTTTCCTTATCGATGAACTC
AATTACGACCTCGATGTTGTTTGATCGGAAGAGCACACGTCTGTACTCCCGTCACAACCAT
CGATCTCGTCTGCCGTCTTCTGCTTGCTATAAAAAACATTTCCTTCTTTTTTTTTT
+
?CCCCGGGGGGGGGFGGGGGGGF,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)/((
```

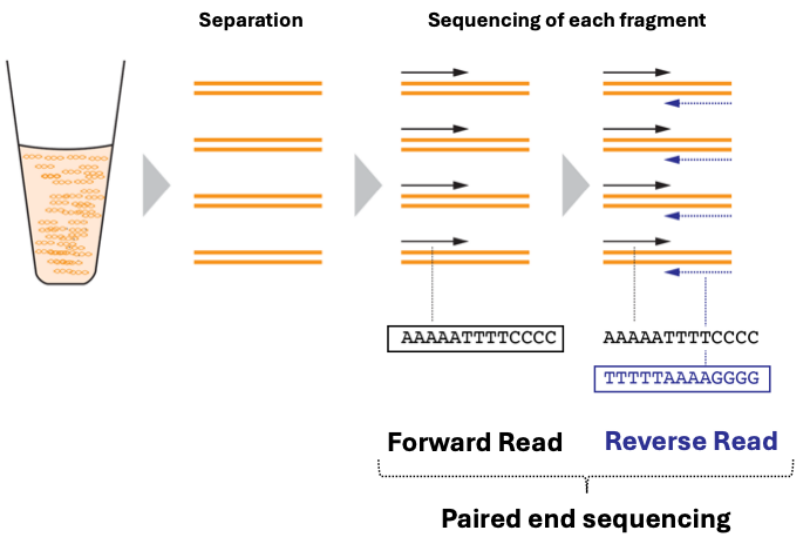
?

High quality

+

Base calling may be wrong…

Before determining species using the dataset...

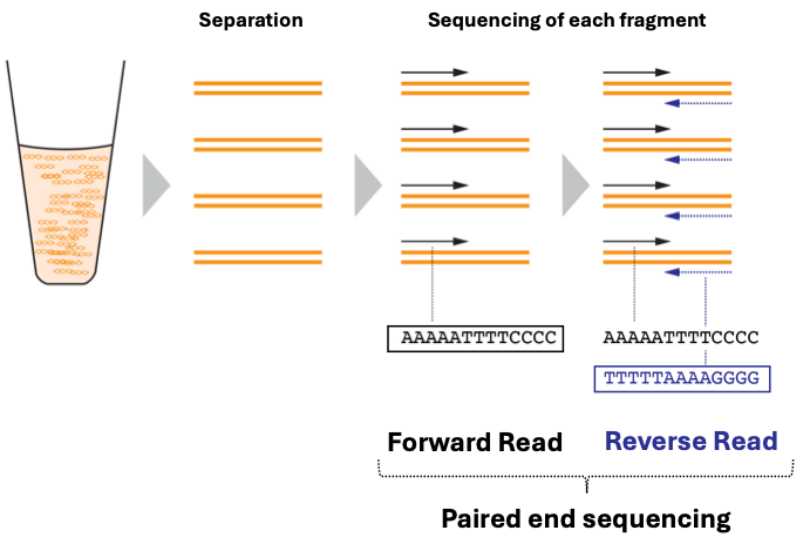


Sample5_L001_R1_001.fastq

50,685
reads

⋮

Before determining species using the dataset...



Sample5_L001_R1_001.fastq

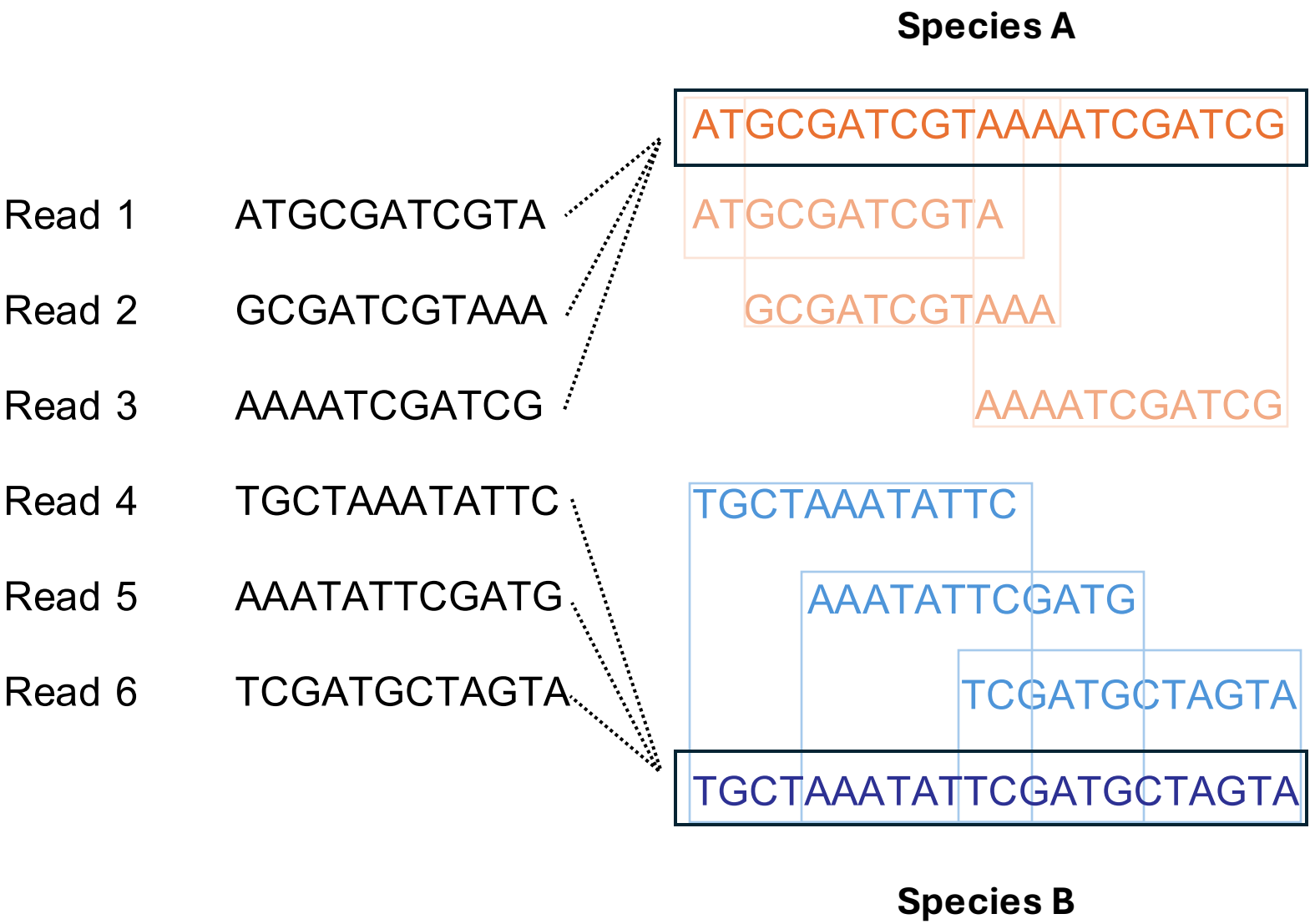
50,685 **reads**

≠

50,685 **species**

Read 1	ATGCGATCGTA
Read 2	GCGATCGTAAA
Read 3	AAAATCGATCG
Read 4	TGCTAAATATTC
Read 5	AAATATTCGATG
Read 6	TCGATGCTAGTA

6 species?

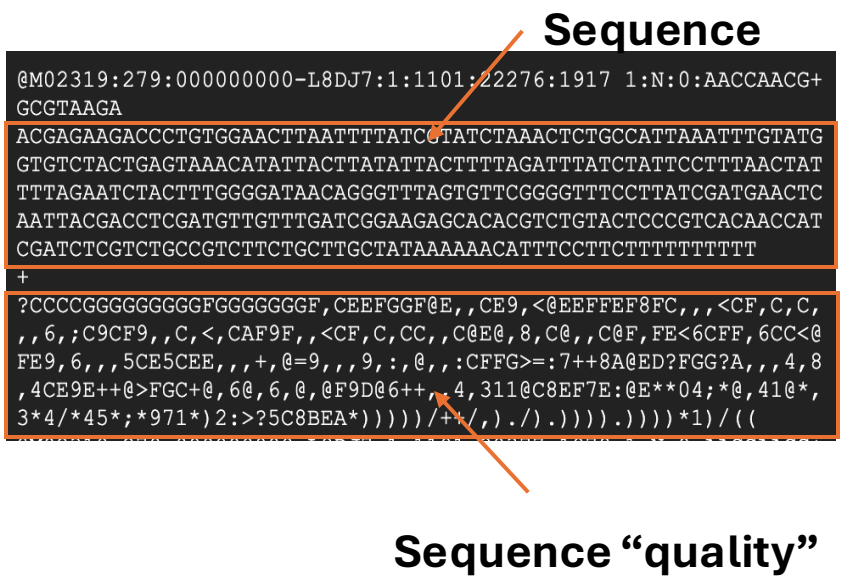


Checkpoint 1: Multiple reads originate from the same species sequence

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?

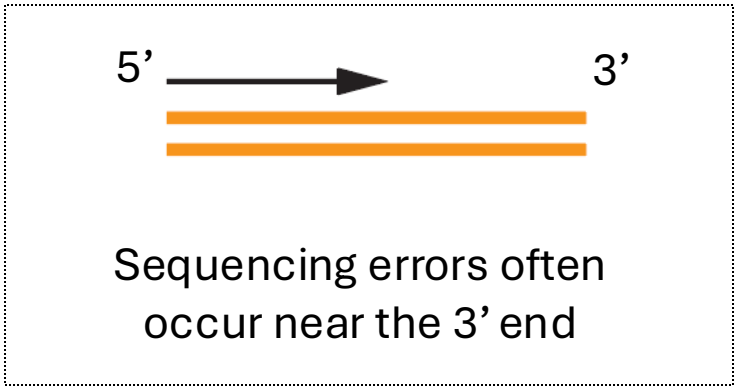
Read 1	ATGCGATCGTA ATG
Read 2	ATGCGATCGTA CCC
Read 3	ATGCGATCGTA TTT
Read 4	ATGCGATCGTA TGA
Read 5	ATGCGATCGTA GGG
Read 6	ATGCGATCGTA ATG

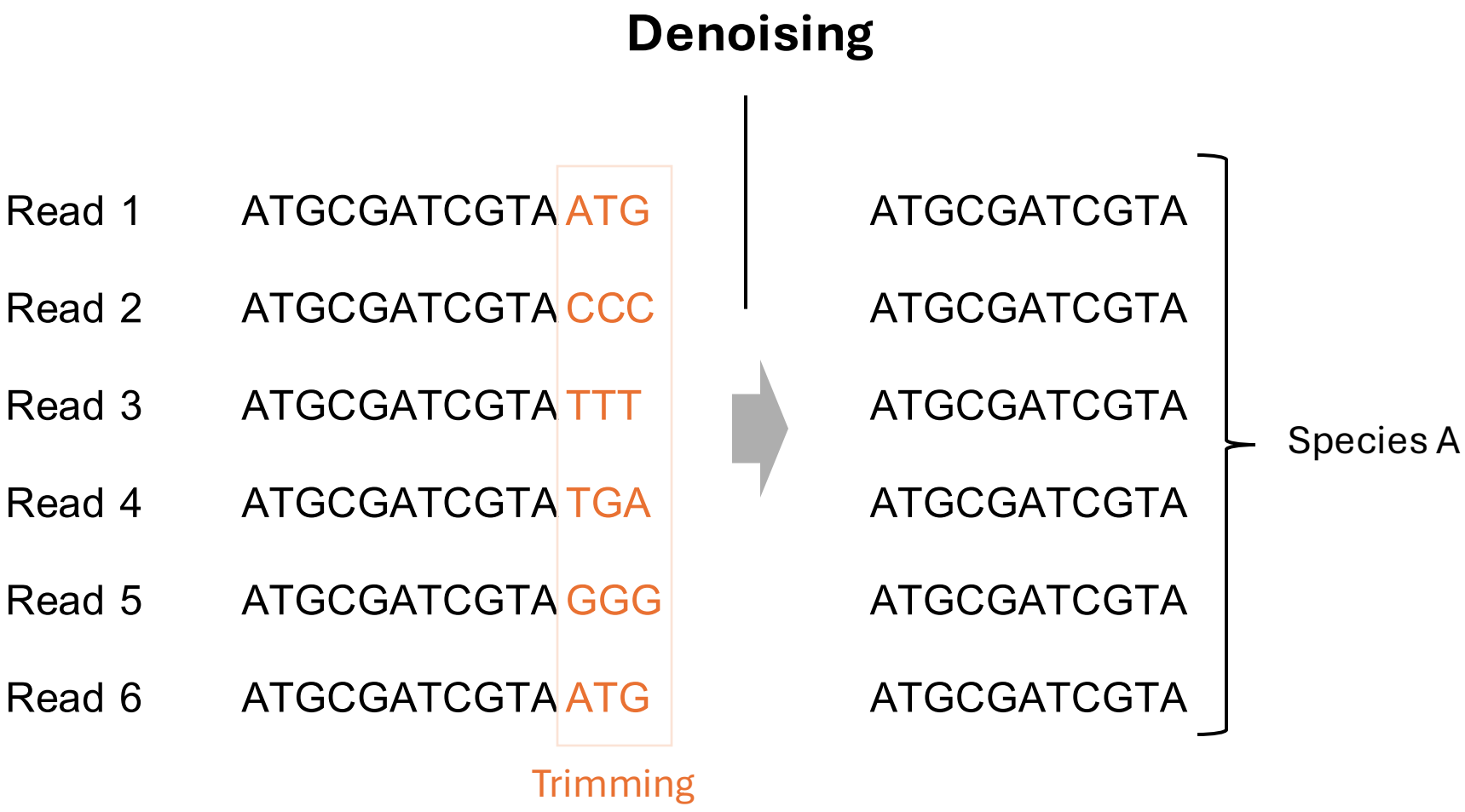


Read 1	ATGCGATCGTA	ATG
	OOO…	XXX
Read 2	ATGCGATCGTA	CCC
	OOO…	XXX
Read 3	ATGCGATCGTA	TTT
	OOO…	XXX
Read 4	ATGCGATCGTA	TGA
	OOO…	XXX
Read 5	ATGCGATCGTA	GGG
	OOO…	XXX
Read 6	ATGCGATCGTA	ATG
	OOO…	XXX

Quality (ex. good or bad)

Variant sequences are likely due to technical errors





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

Data processing for species determination

Sample5_L001_R1_001.fastaq

Denoising

Extraction of the accurate sequences

Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

ATGCGATCGCA

ATGCGAAAGTA

ATGCGATTTTA

ATACGATCGTA

Annotation

Data processing for species determination

Sample5_L001_R1_001.fastaq

DADA2
analysis

Denoising

Extraction of the accurate sequences

Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

ATGCGATCGCA

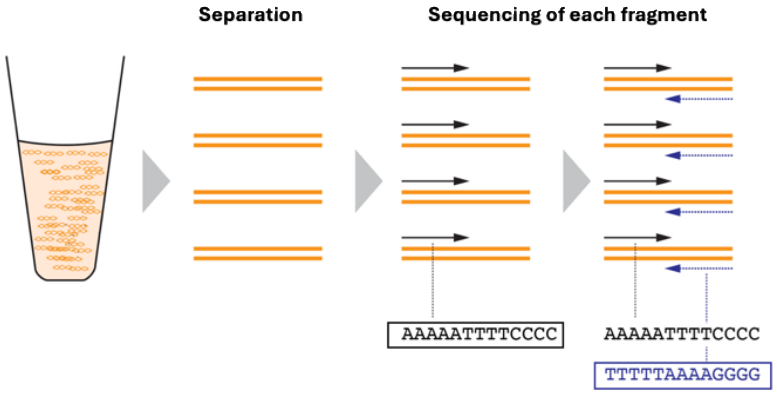
ATGCGAAAGTA

ATGCGATTTTA

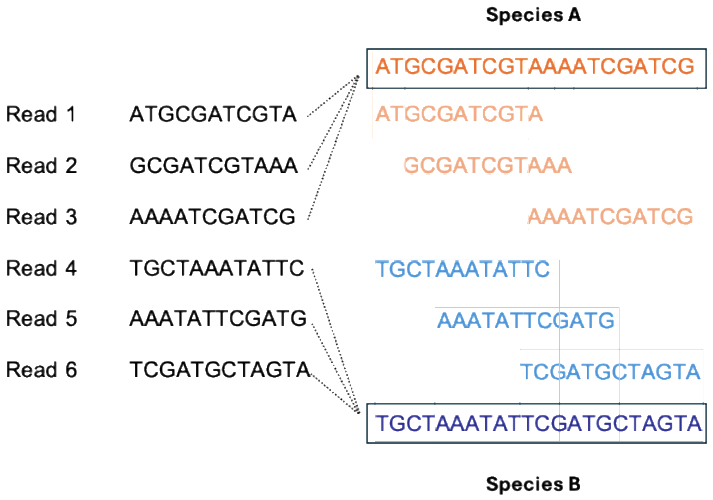
ATACGATCGTA

Annotation

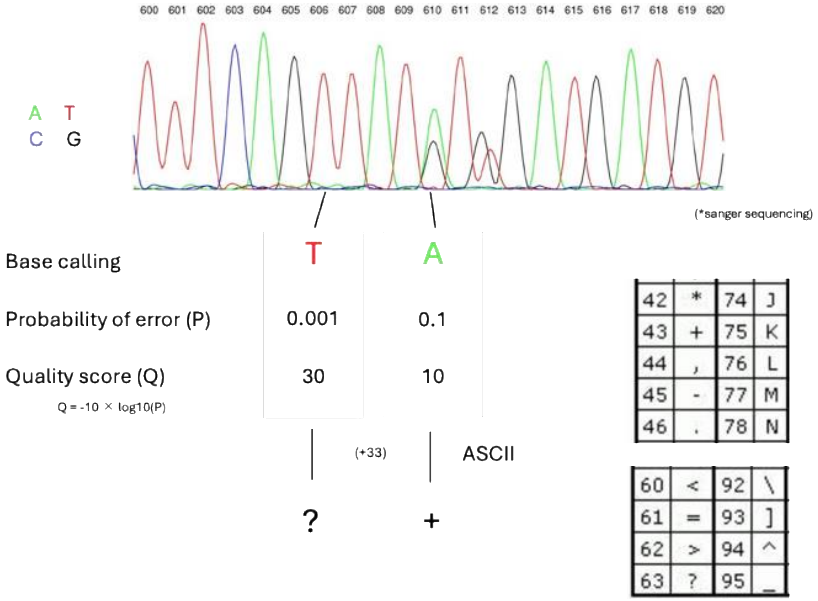
Paired end sequencing



Sequencing variation



Sequencing accuracy



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

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

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.

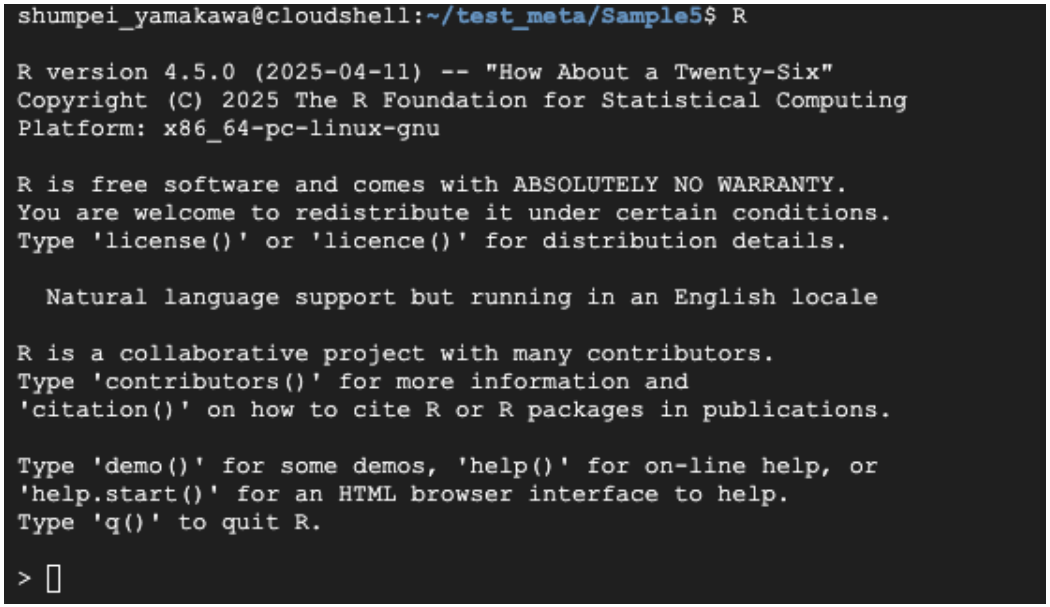
Tutorials

R set up

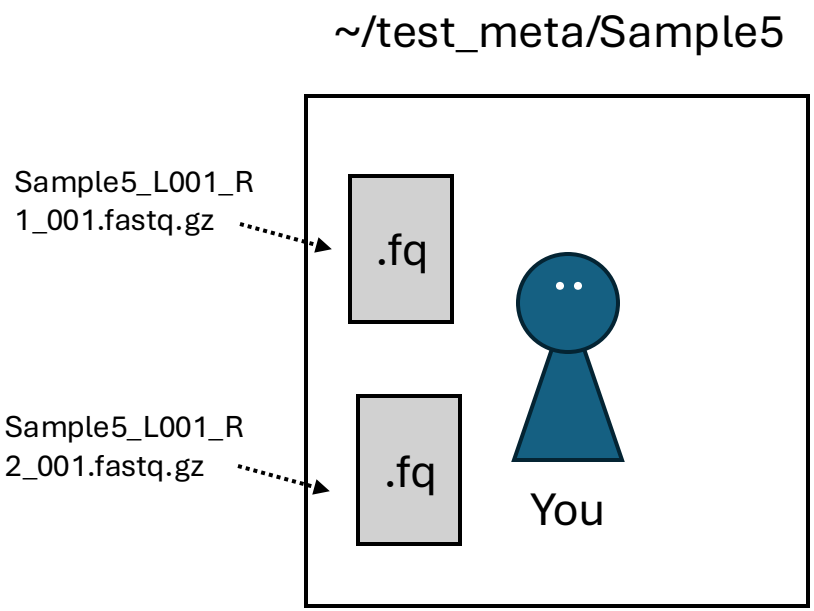
```
sudo apt update -qq
sudo apt install --no-install-recommends software-properties-common dirmngr
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc | sudo tee -a /etc/apt/trusted.gpg.d/cran_
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_release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

Quality check

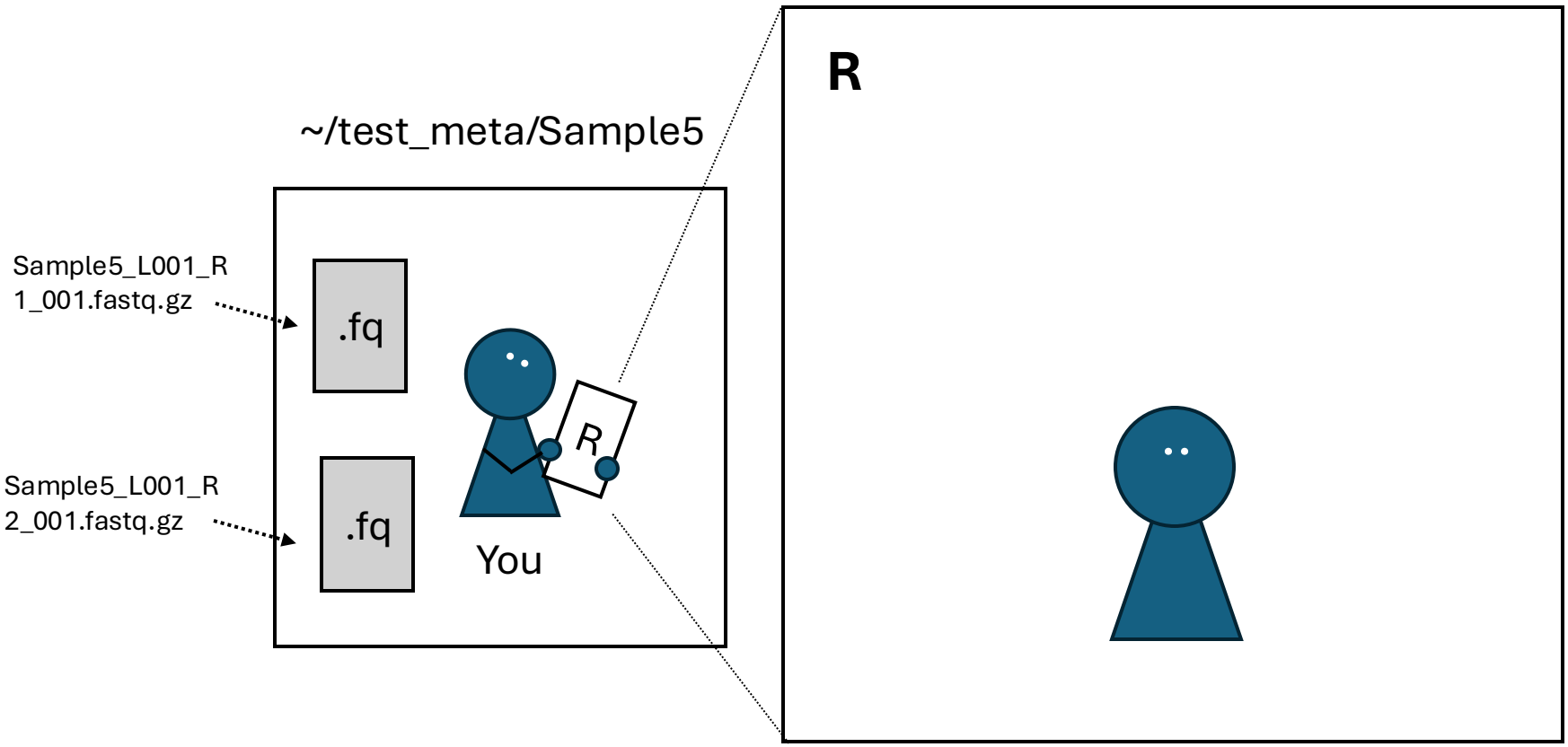
```
R
#open R console
```



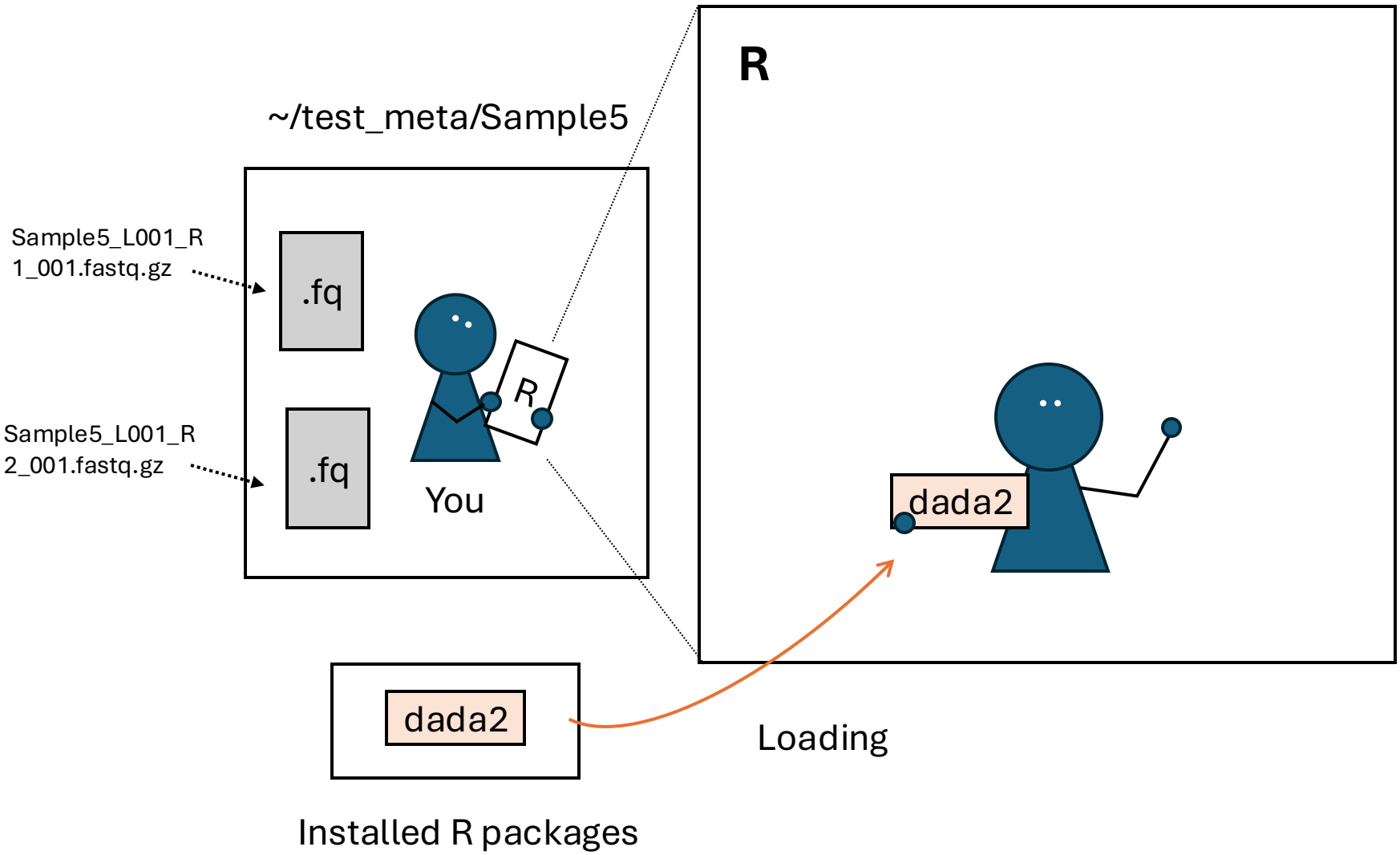
Dada2 set up and data installing



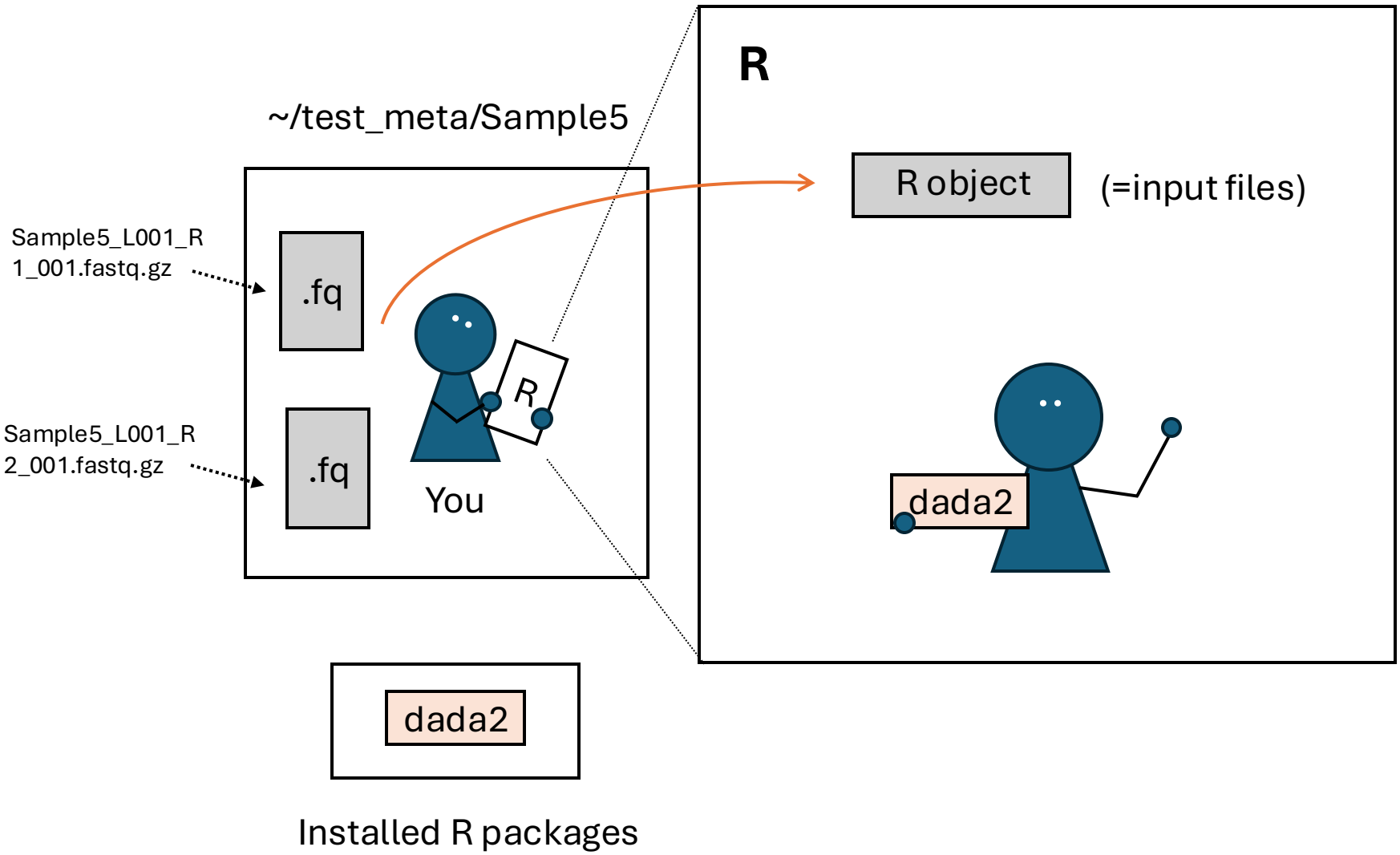
Dada2 set up and data installing



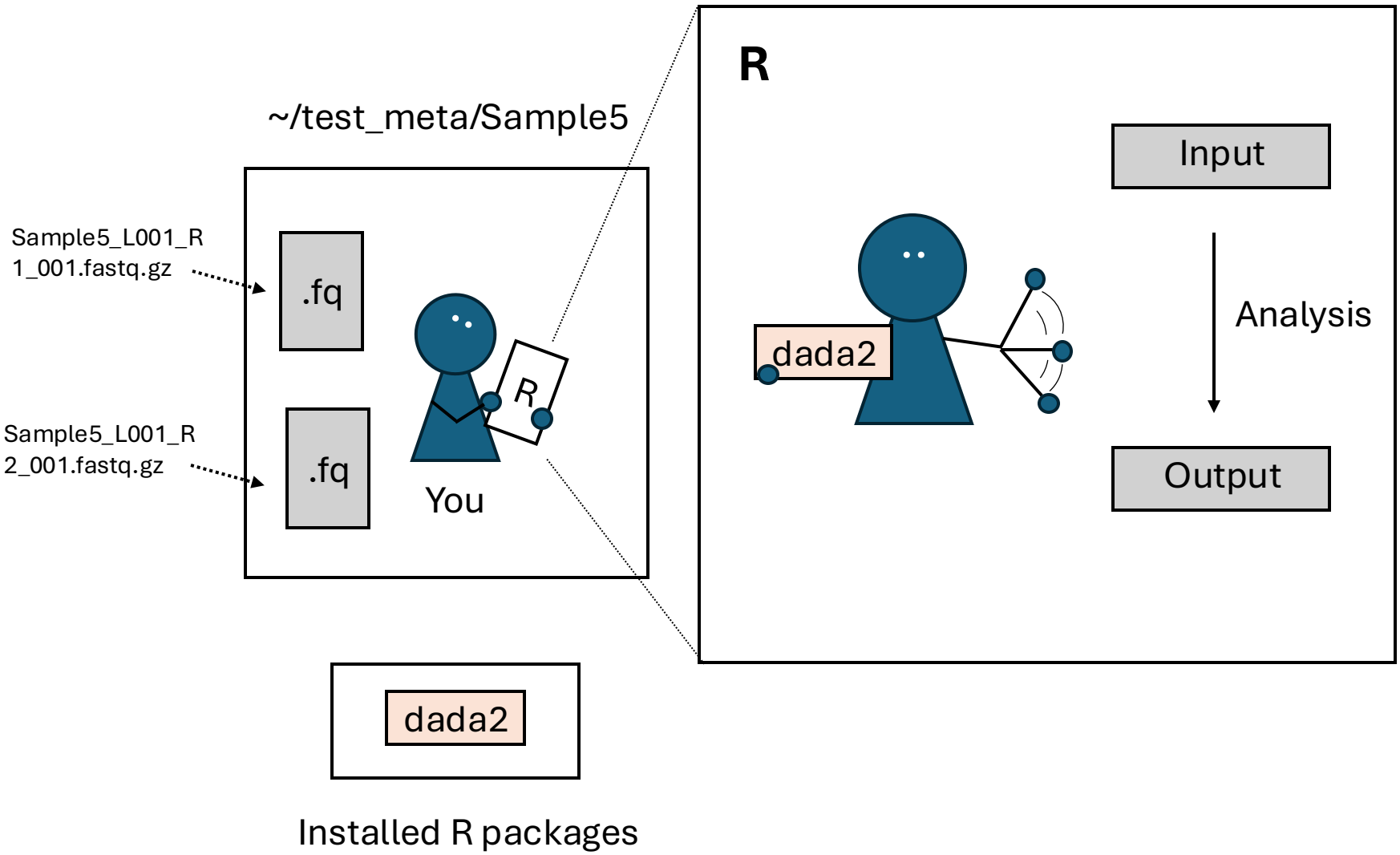
Dada2 set up and data installing



Dada2 set up and data installing



Dada2 set up and data installing



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")
```

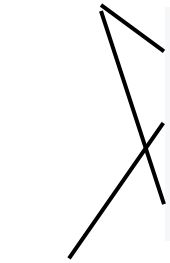
An **R object** is **any data structure** or **piece of data** that exists in R's memory — essentially, **everything you work with in R is an object**.

📦 Common Types of R Objects

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

Quality check

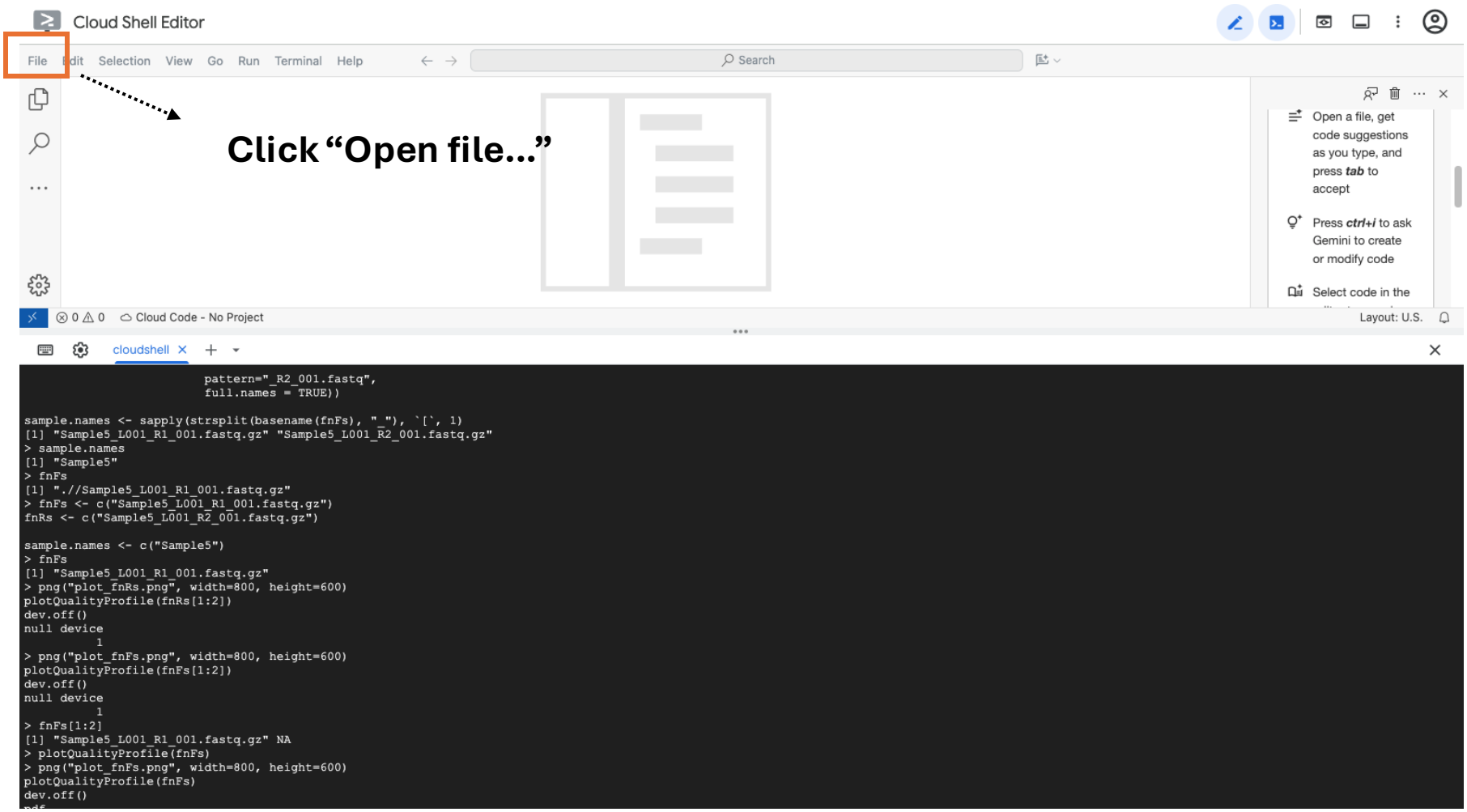
Functions for plotting**

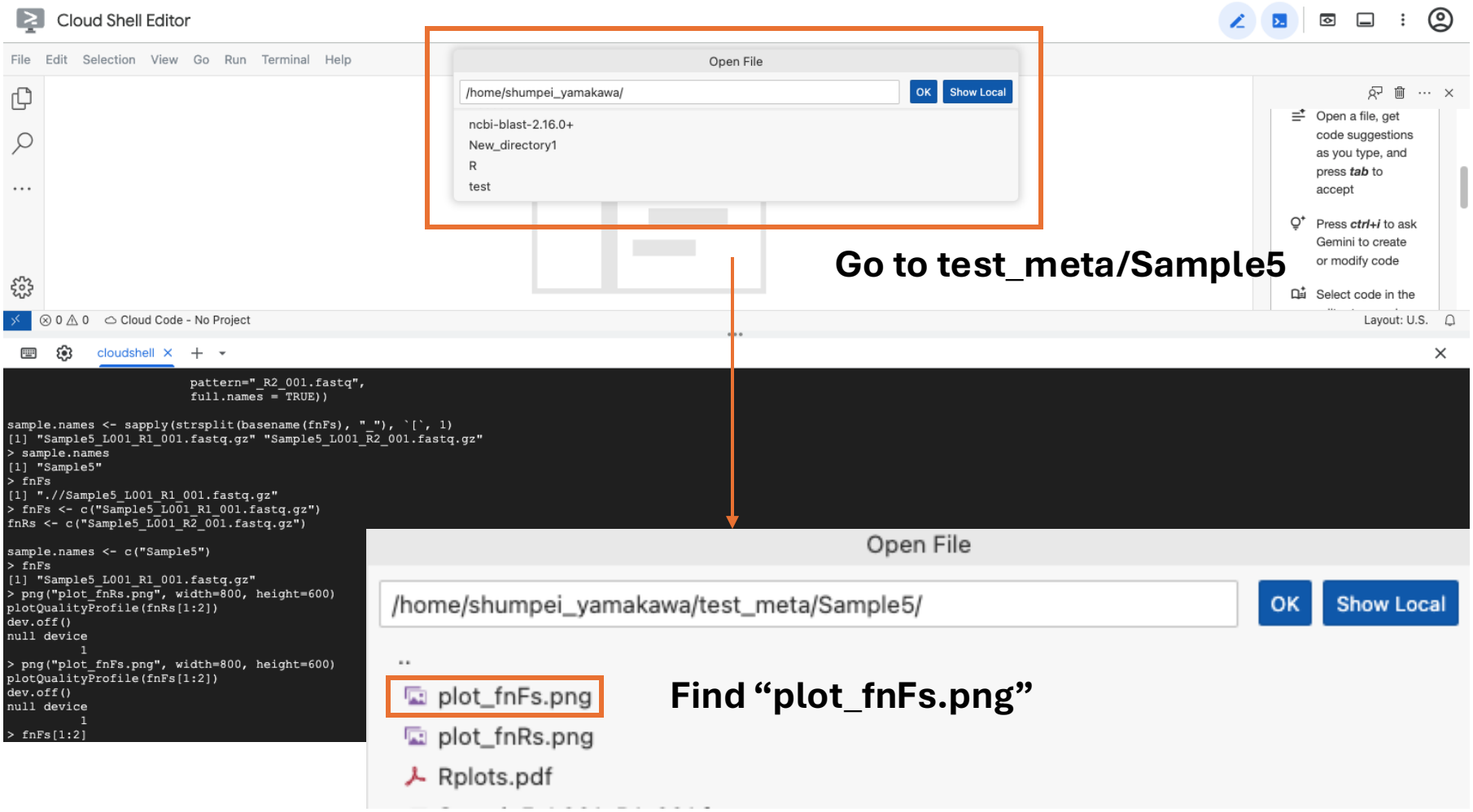


```
png("plot_fnFs.png", width=800, height=600)
plotQualityProfile(fnFs)
dev.off()
```

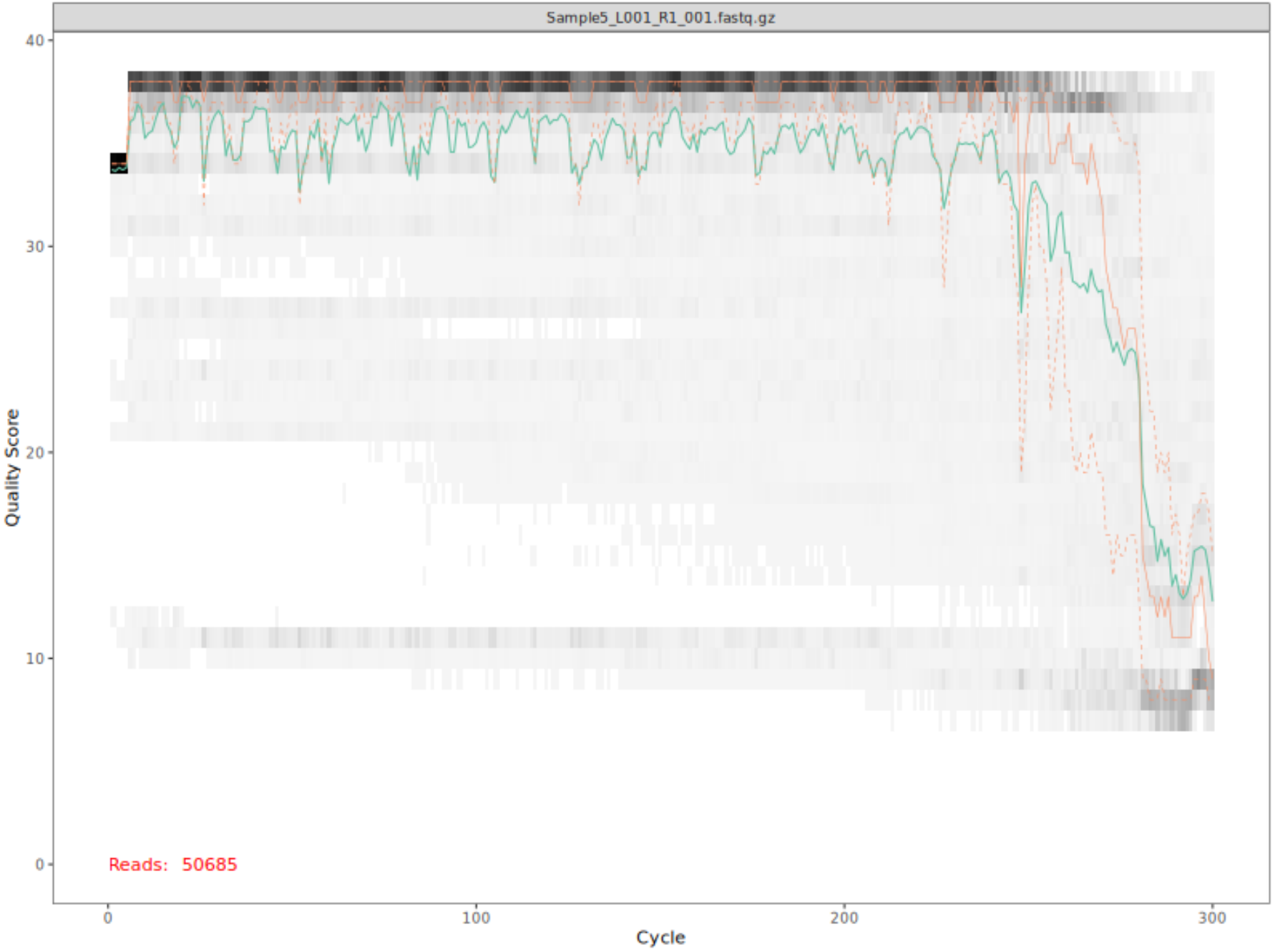
Function for quality check

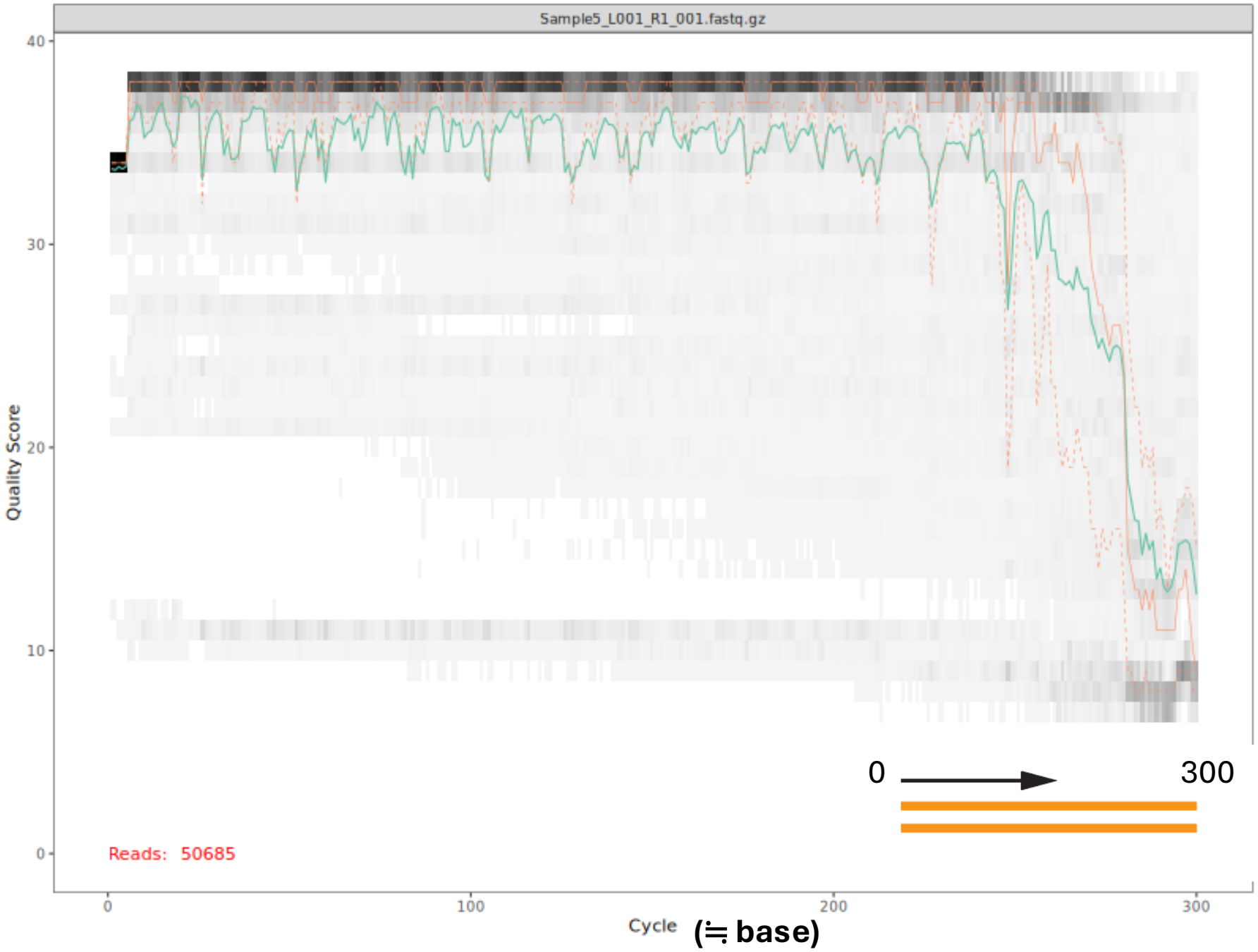
**If you are using your local PC, the png() functions are unnecessary. Enter plotQualityProfile(fnFs), and the plot will appear in a new window.


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
0 0
Cloud Code - No Project
Layout: U.S.
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_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()


Cloud Shell Editor
File Edit Selection View Go Run Terminal Help
Open File
/home/shumpei_yamakawa/
OK
Show Local
ncbi-blast-2.16.0+
New_directory1
R
test
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
Go to test_meta/Sample5
0 0
Cloud Code - No Project
Layout: U.S.
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_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]
Open File
/home/shumpei_yamakawa/test_meta/Sample5/
OK
Show Local
..
plot_fnFs.png
plot_fnRs.png
Rplots.pdf
Find “plot_fnFs.png”

“plot_fnFs.png” opens above the shell

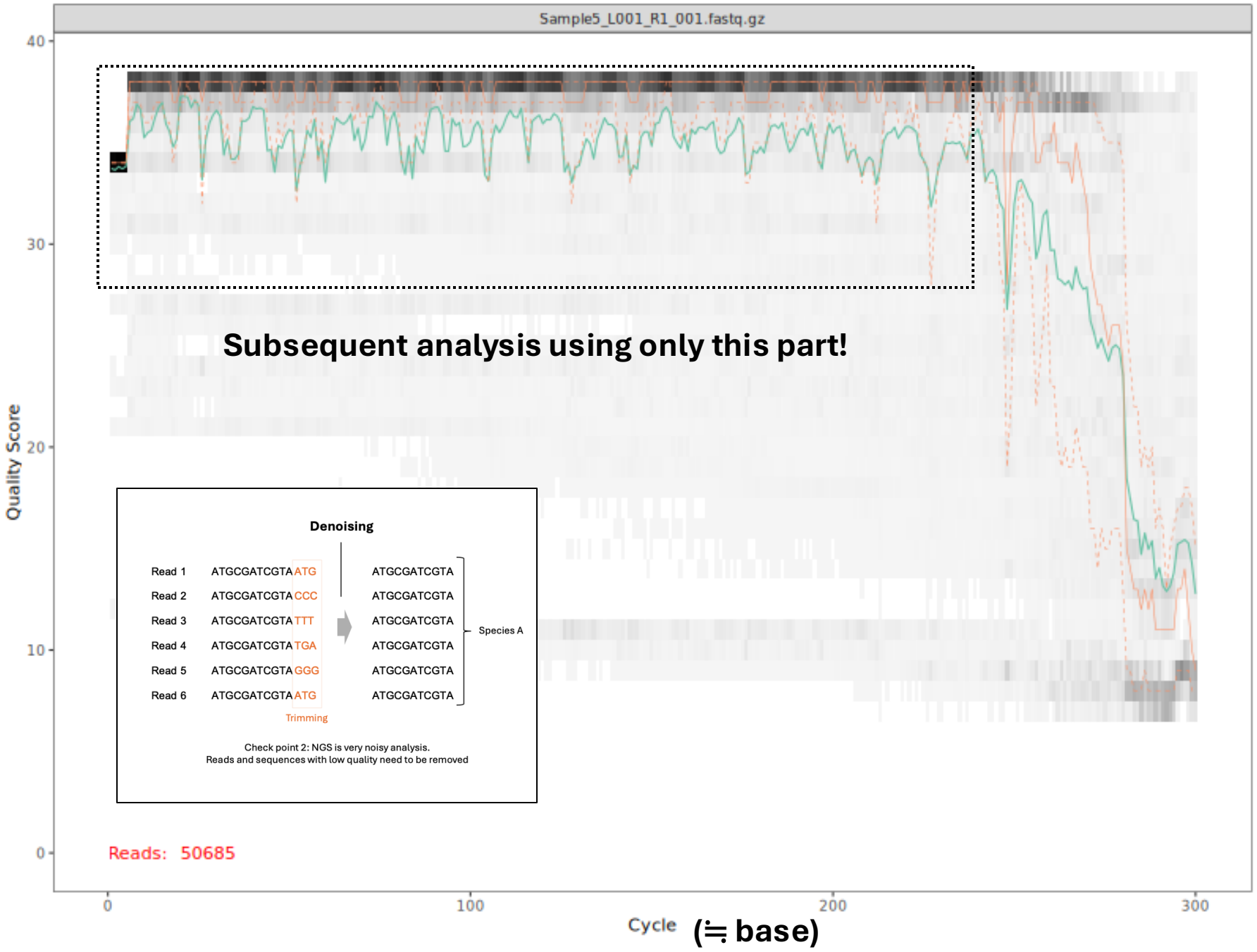

Sample5_L001_R1_001.fastq.gz
Quality Score
40
30
20
10
0
Reads: 50685
0
100
200
300
Cycle


Sample5_L001_R1_001.fastq.gz
Quality Score
40
30
20
10
0
0
100
200
300
Cycle (≒ base)
0
300
Reads: 50685


Sample5_L001_R1_001.fastq.gz
40
30
20
10
0
30
Error ratio = 1/1000
20
Error ratio = 1/100
Quality Score
Reads: 50685
0
100
200
300
Cycle (≒ base)


Sample5_L001_R1_001.fastq.gz
40
30
20
10
0
Quality Score
30
Error ratio = 1/1000
20
Error ratio = 1/100
Reads: 50685
0
100
200
300
Cycle (≒ base)
@read1
... A T C ...
+
...30 35 34 ...
@read2
...A T C G ...
+
...30 33 38...
@read3
...A T C G ...
+
...34 34 34 ...


Sample5_L001_R1_001.fastq.gz
40
30
20
10
0
Quality Score
30
Error ratio = 1/1000
20
Low quality...
Error ratio = 1/100
Reads: 50685
0
100
200
300
Cycle
(≒ base)


Sample5_L001_R1_001.fastq.gz
Quality Score
40
30
20
10
0
0
100
200
300
Cycle (≒ base)
Subsequent analysis using only this part!
Denoising
Read 1 ATGCGATCGTA ATG ATGCGATCGTA
Read 2 ATGCGATCGTA CCC ATGCGATCGTA
Read 3 ATGCGATCGTA TTT ATGCGATCGTA
Read 4 ATGCGATCGTA TGA ATGCGATCGTA
Read 5 ATGCGATCGTA GGG ATGCGATCGTA
Read 6 ATGCGATCGTA ATG ATGCGATCGTA
Species A
Trimming
Check point 2: NGS is very noisy analysis.
Reads and sequences with low quality need to be removed
Reads: 50685

Filtering

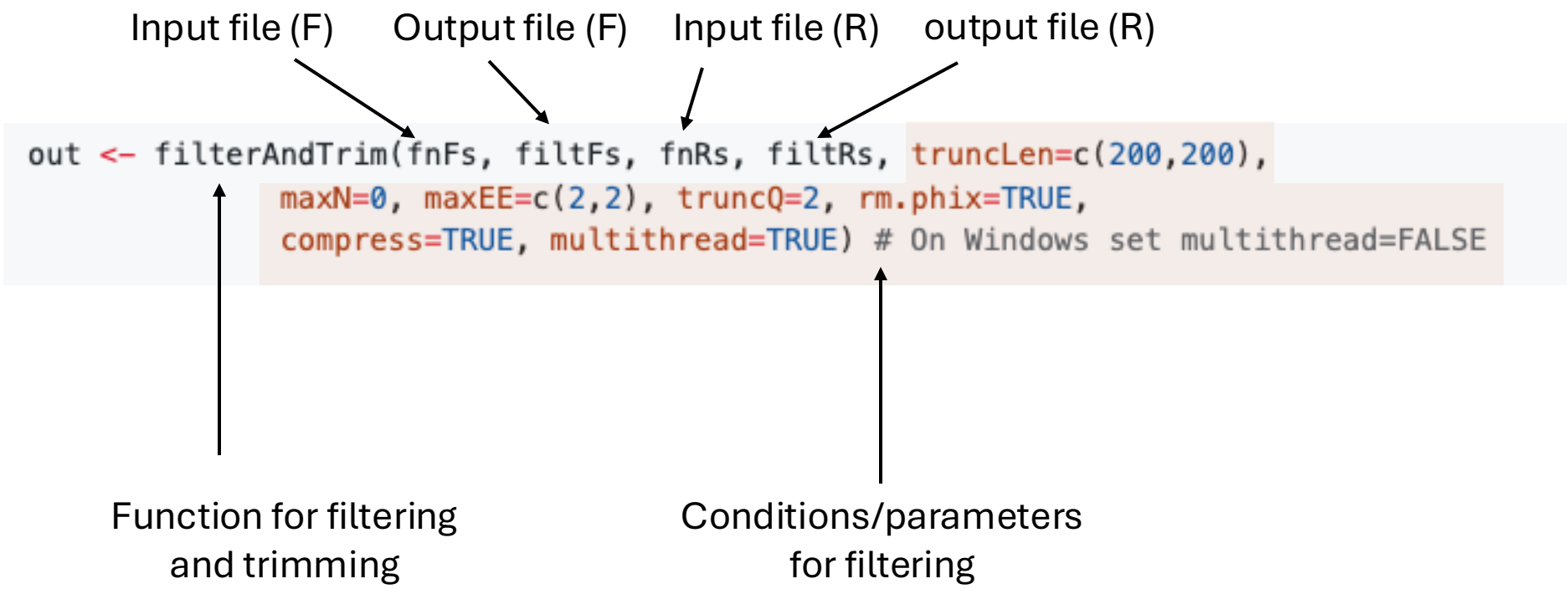
```
filtFs <- c(".//filtered/Sample5_L001_R1_filtered_001.fastq.gz")
filtRs <- c(".//filtered/Sample5_L001_R2_filtered_001.fastq.gz")
names(filtFs) <- sample.names
names(filtRs) <- sample.names

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

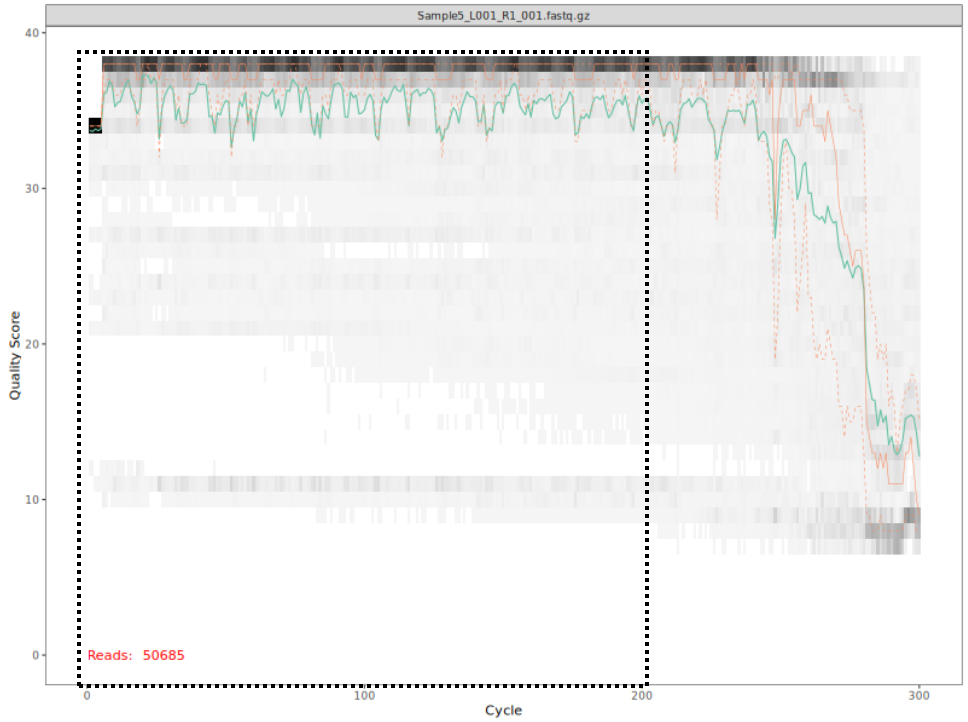
```
names(filtFs) <- sample.names
names(filtRs) <- sample.names
```

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)

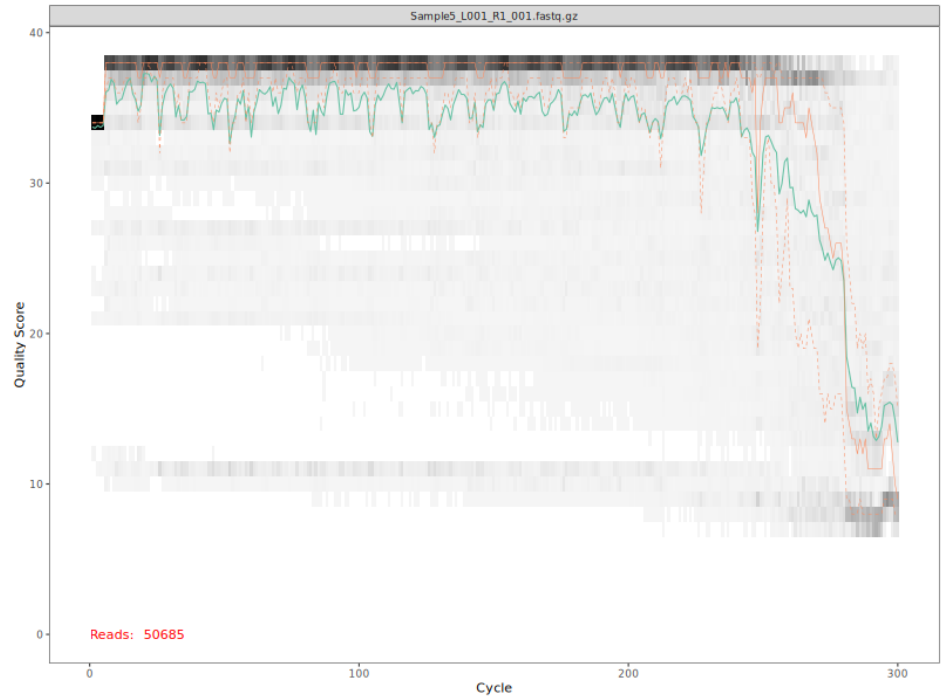
truncLen = c(200, 100)

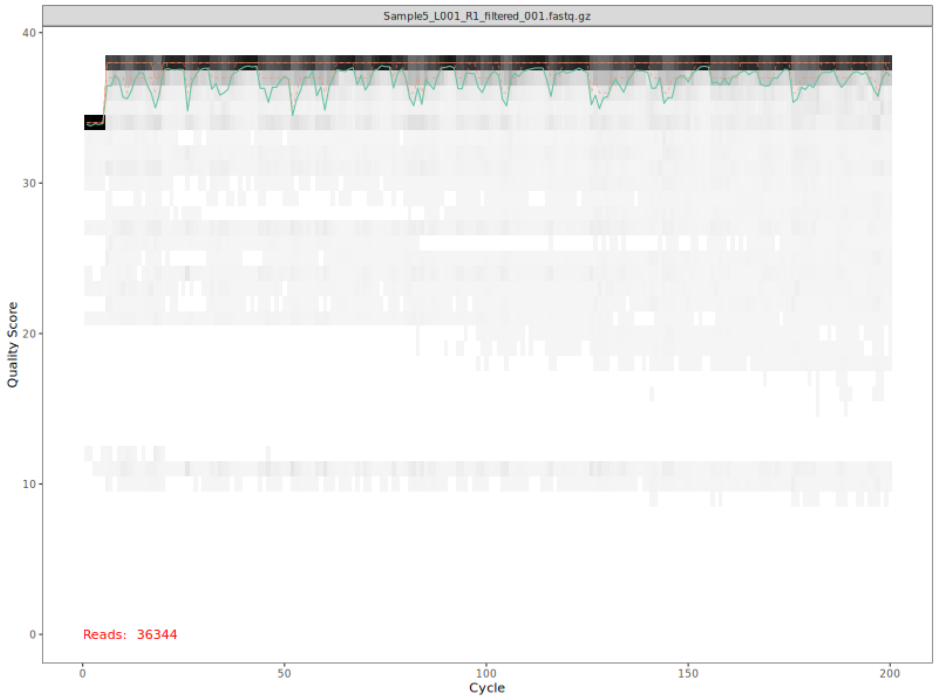
```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(200,200),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```

`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 ≤ 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).

Followed the tutorial (https://benjjneb.github.io/dada2/tutorial.html)

```
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
```





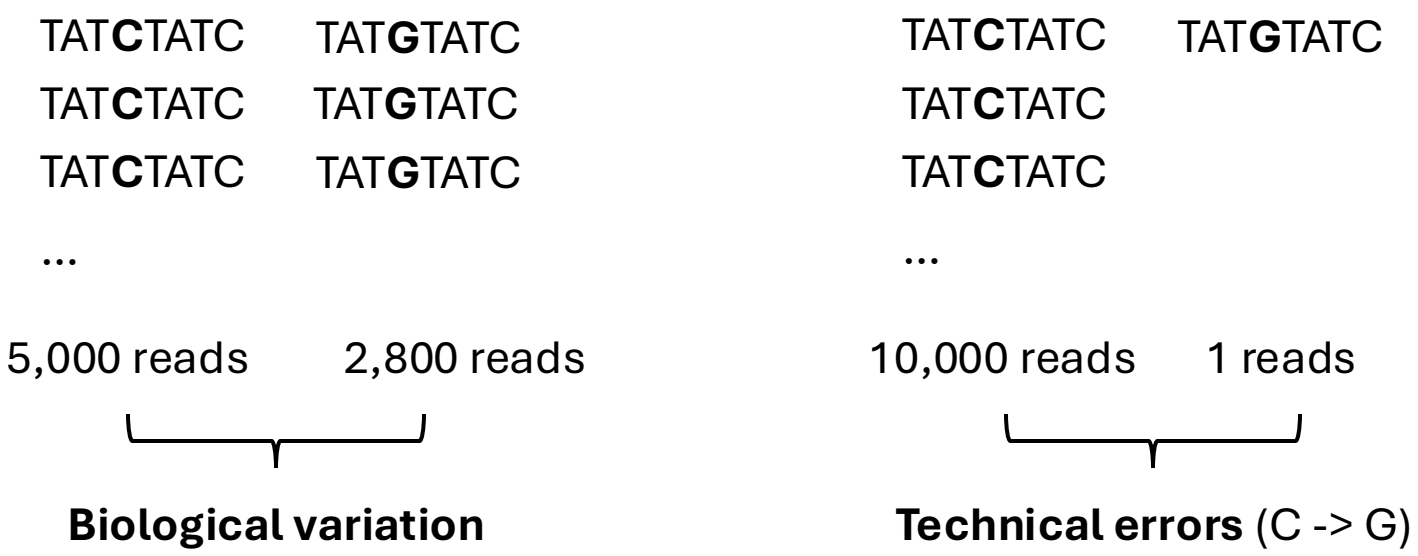
Before
(plot_fnFs.png)

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()
```

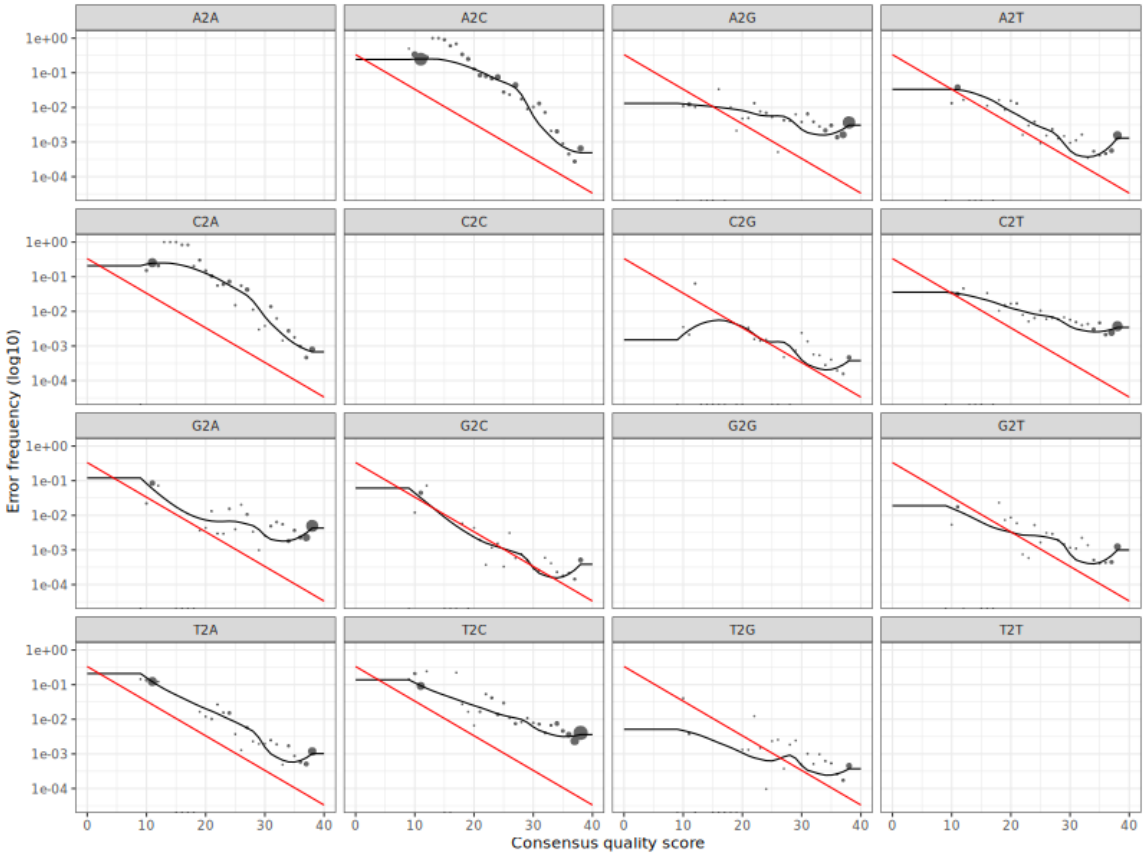
Ex) Estimate the errors



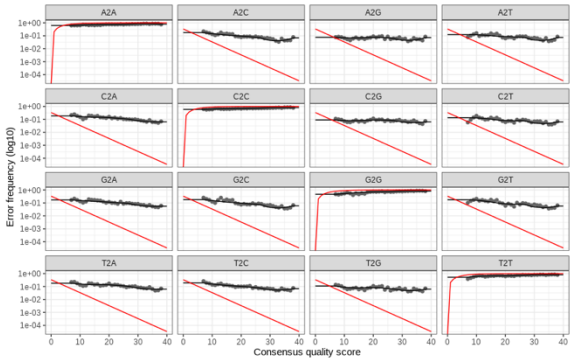
```
errF <- learnErrors(filtFs, multithread=TRUE)
errR <- learnErrors(filtRs, multithread=TRUE)
```

```
> 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()
```

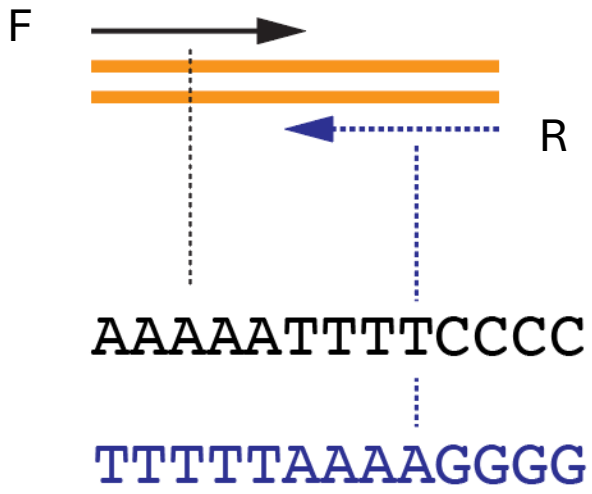


bad example



Marge

```
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

Denoising

Extraction of the accurate sequences

Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

ATGCGATCGCA

ATGCGAAAGTA

ATGCGATTTTA

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.
```

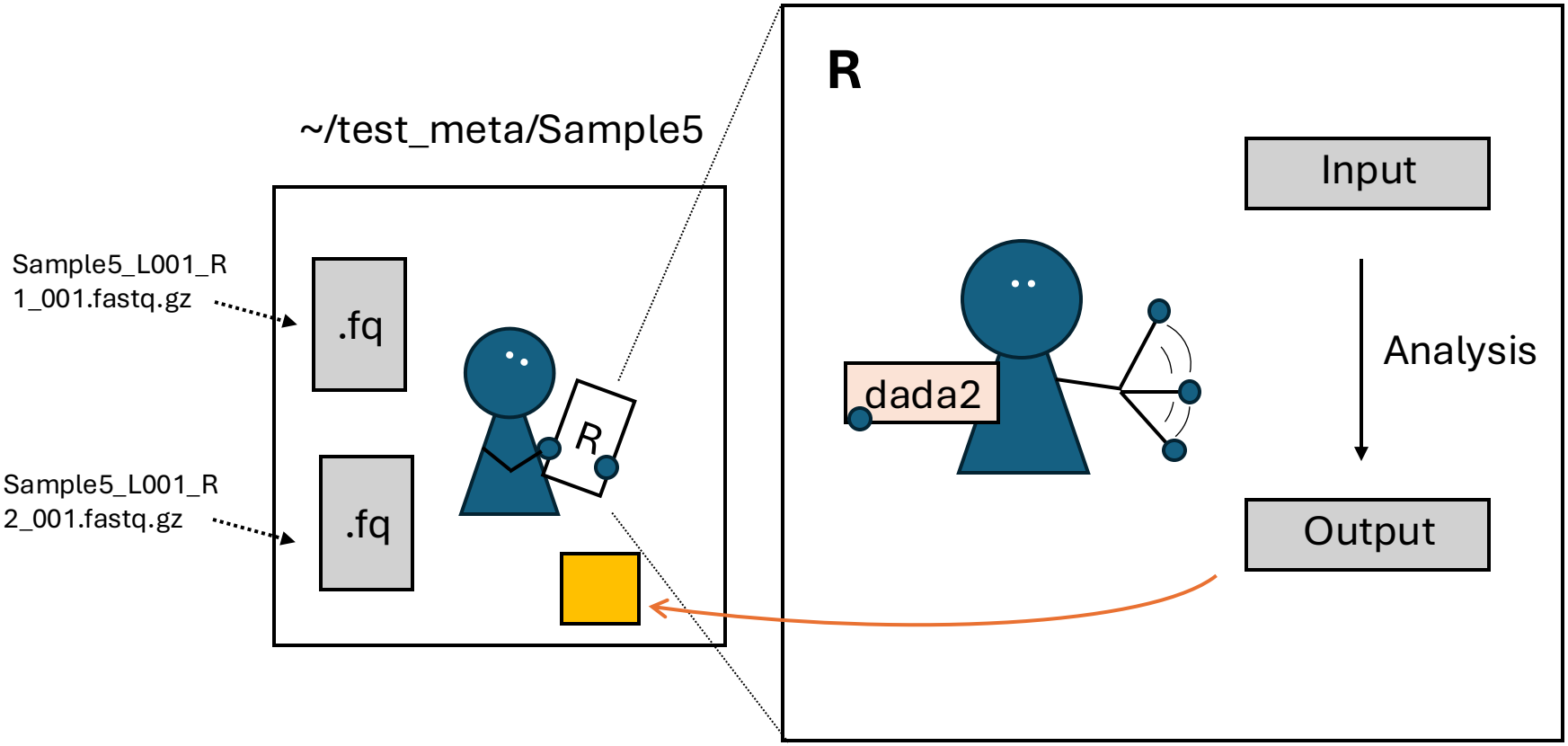
```
> head(mergers[[1]])
[1] "ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTT
ATCGATGAACACAATTACGACCTCGATGTTGGA"
[2] "ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTT
ATCGATGAACACAATTACGACCTCGATGTTGGA"
[3] "ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTT
ATCGATGAACACAATTACGACCTCGATGTTGGA"
[4] "ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTT
ATCGATGAACACAATTACGACCTCGATGTTGGA"
[5] "CGACGCTCTTCCGATCTACGAGAAGACCCTATGGAGCTTTACTTTAAGGTTAAGCCTACAAGTTTAAATGGGAACTTTTGAGGAACAAAATACCTCTGCCGACCTACTCTTCTATTCGTGGAGAAGCTACTTTGGGGATAACAGGGTAATACAGCTGGGGAGCACTTATCG
ATAGTTGTTCTTACGACCTCGATGTTGGATGATCGGAAGAGCACAC"
[6] "ACGAGAAGACCCTATGGAGCTTAATTTTACGCCTGTTTTATACCTATGCCCGGTATACAAATTTACATGGGGACTGTTGAGTGAGAAAATAATTTGAAACTCTTTACTATTGATCTAACTTCTTAATTATCTGAAAAGCTACTTTGGGGATAACAGGGTGAAGCGAGTGGA
GAGTTCTTATCGATACTTGCAATTACGACCTCGATGTTGGA"
```

```
> head(mergers[[1]])
[1] "ACGAGAAGACCCTGTGGAGCTTAATTTTT
ATCGATGAACACAATTACGACCTCGATGTTGGA"
```

ASV table also includes the “semi-quantitative” information about abundance

```
240    ACGAGAAGACCCCATGGAGCTTAAGGTACAAAGTTTAACCATGTTAAACAACTTATTAAAGAGC
CTCTAAGCCACAGAAAATCTGACCAAAAATGATCCGACACATAGTCGATCAACGAACCAAGTTACCCTAGG
262                                ACGAGAAGACCCTATGGAACTTTAATTCATTAATTCAA
GACACACAAGTCAAAGATACATCATAAATTGACCCAGAACTTATCTGATCAACGGACCAAGTTACCCTAGG
```

	abundance	forward	reverse	nmatch	nmismatch	nindel	prefer	accept
1	1095	2	2	196	0	0	1	TRUE
2	1094	1	4	196	0	0	1	TRUE
3	903	3	3	196	0	0	1	TRUE
4	836	5	16	196	0	0	1	TRUE
5	816	6	52	183	0	0	1	TRUE
6	804	4	8	188	0	0	1	TRUE
7	802	13	15	196	0	0	1	TRUE
8	790	10	9	188	0	0	1	TRUE
9	765	7	5	196	0	0	1	TRUE
10	761	8	12	183	0	0	1	TRUE
11	734	14	11	196	0	0	1	TRUE
12	733	9	6	196	0	0	1	TRUE
13	712	12	7	196	0	0	1	TRUE
14	686	16	10	196	0	0	1	TRUE
15	669	17	49	196	0	0	1	TRUE
16	649	11	18	183	0	0	1	TRUE
17	616	20	17	196	0	0	1	TRUE
18	581	19	53	188	0	0	1	TRUE
19	574	21	51	196	0	0	1	TRUE



Output to shell

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

q() #terminate R console
```

```
> q()
Save workspace image? [y/n/c]: n
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
```

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"
"ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",1095,2,2,196,0,0,1,TRUE
"ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",1094,1,4,196,0,0,1,TRUE
"ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",903,3,3,196,0,0,1,TRUE
"ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",836,5,16,196,0,0,1,TRUE
"CGACGCTCTTCCGATCTACGAGAAGACCCTATGGAGCTTTACTTTAAGGTTAAGCCTACAAGTTTAAATGGGAACTTTTGAGGAACAAAATACCTCTGCCGACCTACTCTTCTATTCGTGGAGAAGCTACTTTGGGGATAACAGGGTAATACAGCTGGGGAGCACTTATCGATAG
TTGTTCTTACGACCTCGATGTTGGATGATCGGAAGAGCACAC",816,6,52,183,0,0,1,TRUE
"ACGAGAAGACCCTATGGAGCTTAATTTTACGCCTGTTTTATACCTATGCCCGGTATACAAATTTACATGGGGACTGTTGAGTGAGAAAATAATTTGAAACTCTTTACTATTGATCTAACTTCTTAATTATCTGAAAAGCTACTTTGGGGATAACAGGGTGAAGCGAGTGGAGAGT
TCTTATCGATACTTGCAATTACGACCTCGATGTTGGA",804,4,8,188,0,0,1,TRUE
"ACGAGAAGACCCTGTGGAACTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",802,13,15,196,0,0,1,TRUE
"ACGAGAAGACCCTGTGGAGCTTAATTTTACGCCTGTTTTATACCTATGCCCGGTATACAAATTTACATGGGGACTGTTGAGTGAGAAAATAATTTGAAACTCTTTACTATTGATCTAACTTCTTAATTATCTGAAAAGCTACTTTGGGGATAACAGGGTGAAGCGAGTGGAGAGT
TCTTATCGATACTTGCAATTACGACCTCGATGTTGGA",790,10,9,188,0,0,1,TRUE
"ACGAGAAGACCCTGTGGAACTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTATGGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACTATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCG
ATGAACACAATTACGACCTCGATGTTGGA",765,7,5,196,0,0,1,TRUE
```

Sample5_L001_R1_001.fastaq

Denoising

Extraction of the accurate sequences

Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

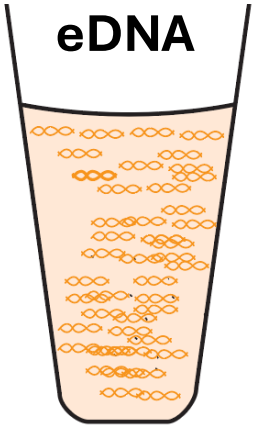
ATGCGATCGCA

ATGCGAAAGTA

ATGCGATTTTA

ATACGATCGTA

126 variants



Amplicon Sequence Variants (ASVs)

ATGCGATCGTT

ATGCGATCGCA

ATGCGAAAGTA

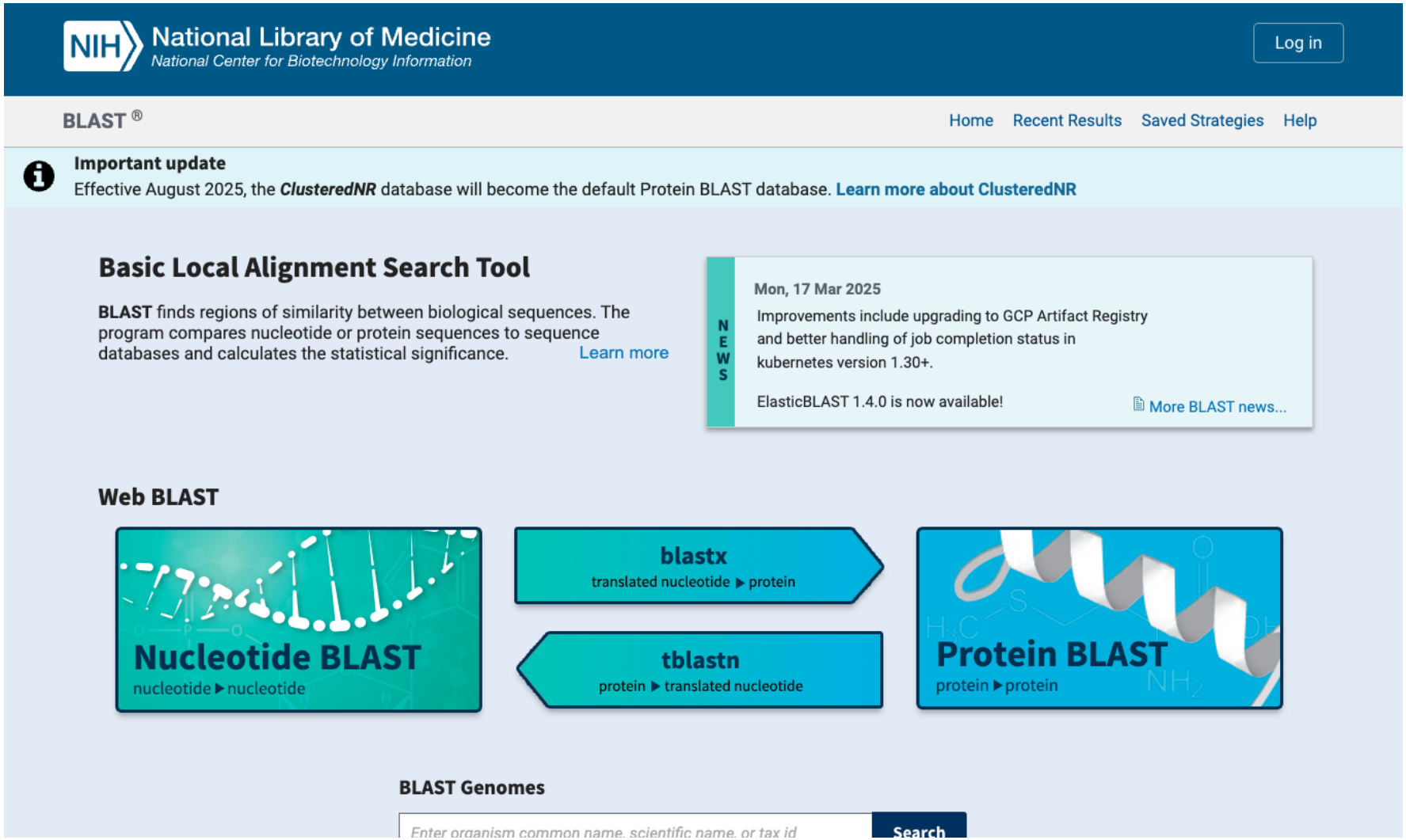
ATGCGATTTTA

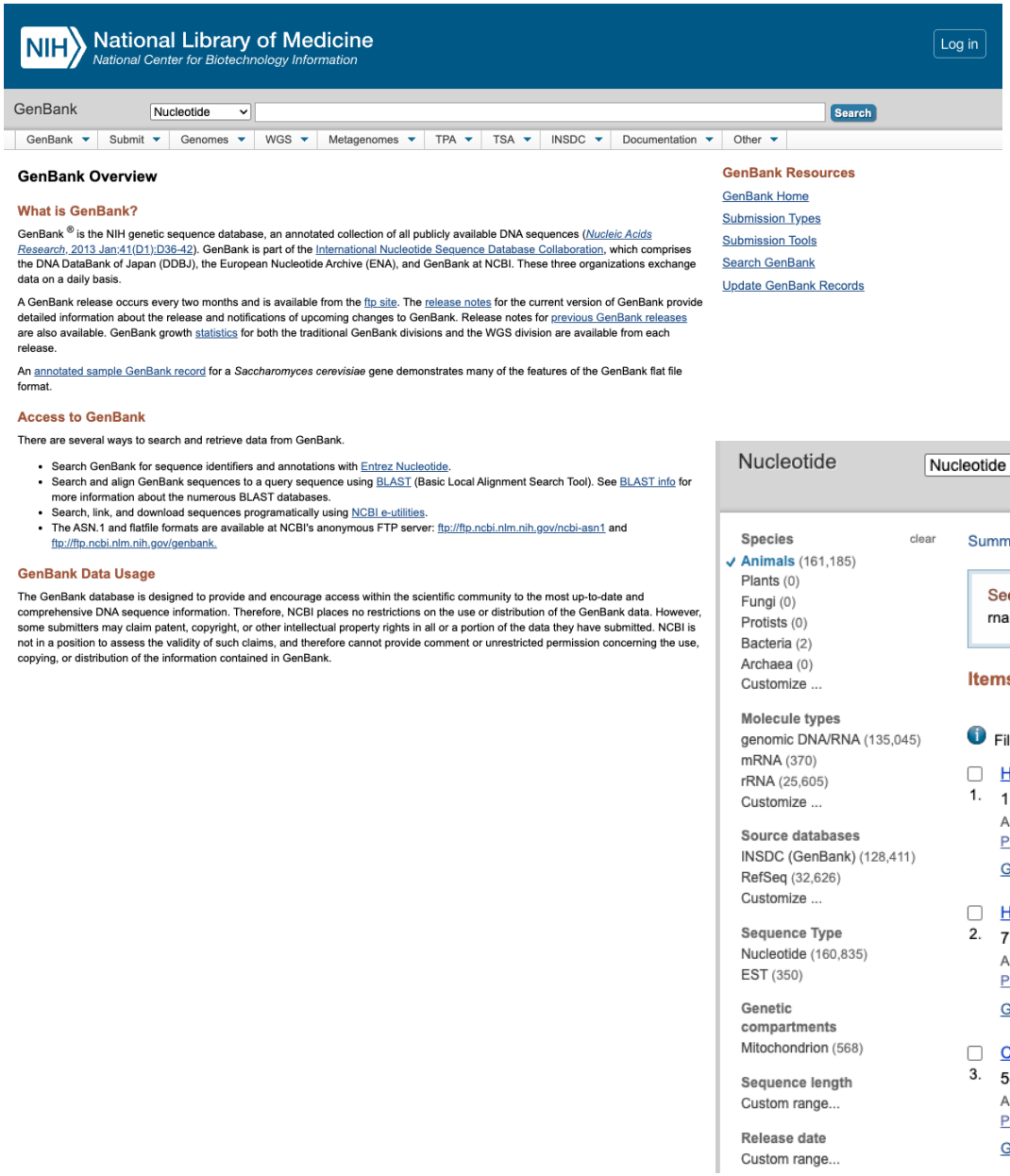
ATACGATCGTA

126 variants

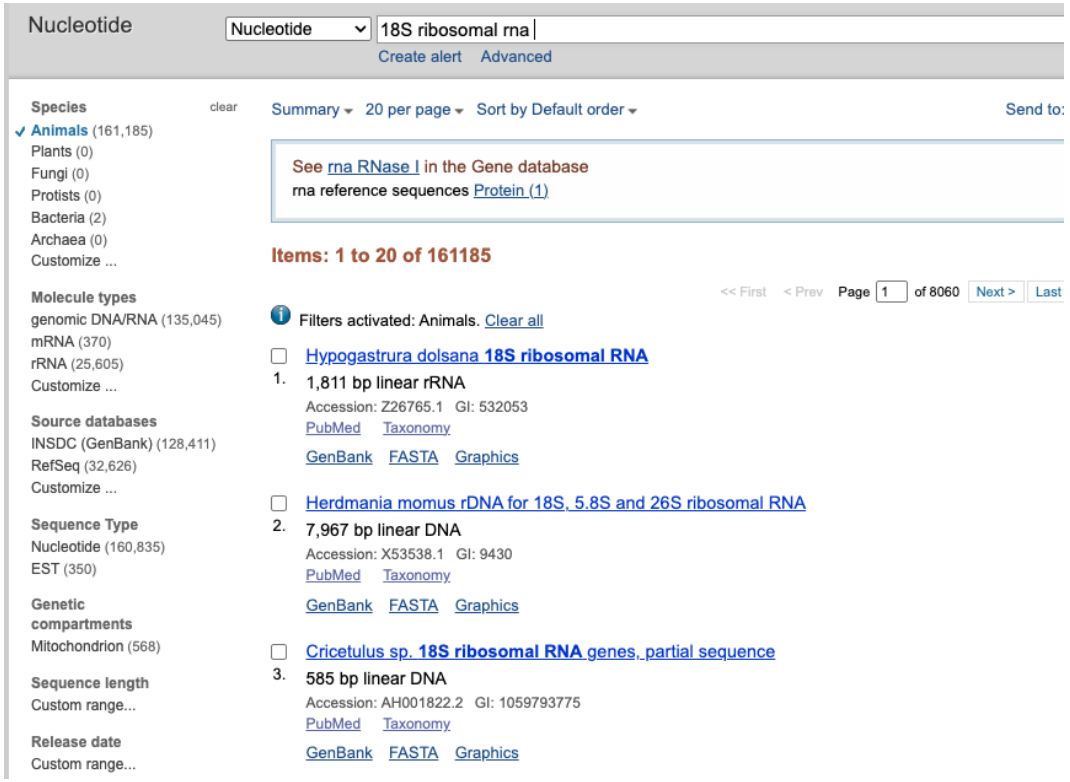
→Annotation

Practice 3: Annotation of Metabarcoding Data




National Library of Medicine
National Center for Biotechnology Information
Log in
GenBank
Nucleotide
Search
GenBank
Submit
Genomes
WGS
Metagenomes
TPA
TSA
INSDC
Documentation
Other
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 programatically 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

Genbank


Nucleotide
Nucleotide
18S ribosomal rna
Create alert
Advanced
Species
clear
Animals (161,185)
Plants (0)
Fungi (0)
Protists (0)
Bacteria (2)
Archaea (0)
Customize ...
Molecule types
genomic DNA/RNA (135,045)
mRNA (370)
rRNA (25,605)
Customize ...
Source databases
INSDC (GenBank) (128,411)
RefSeq (32,626)
Customize ...
Sequence Type
Nucleotide (160,835)
EST (350)
Genetic compartments
Mitochondrion (568)
Sequence length
Custom range...
Release date
Custom range...
Summary
20 per page
Sort by Default order
Send to:
See rna RNase I in the Gene database
rna reference sequences Protein (1)
Items: 1 to 20 of 161185
<< First
< Prev
Page 1 of 8060
Next >
Last
Filters activated: Animals. Clear all
Hypogastrura dolsana 18S ribosomal RNA
1. 1,811 bp linear rRNA
Accession: Z26765.1 GI: 532053
PubMed Taxonomy
GenBank FASTA Graphics
Herdmania momus rDNA for 18S, 5.8S and 26S ribosomal RNA
2. 7,967 bp linear DNA
Accession: X53538.1 GI: 9430
PubMed Taxonomy
GenBank FASTA Graphics
Cricetulus sp. 18S ribosomal RNA genes, partial sequence
3. 585 bp linear DNA
Accession: AH001822.2 GI: 1059793775
PubMed Taxonomy
GenBank FASTA Graphics

Database

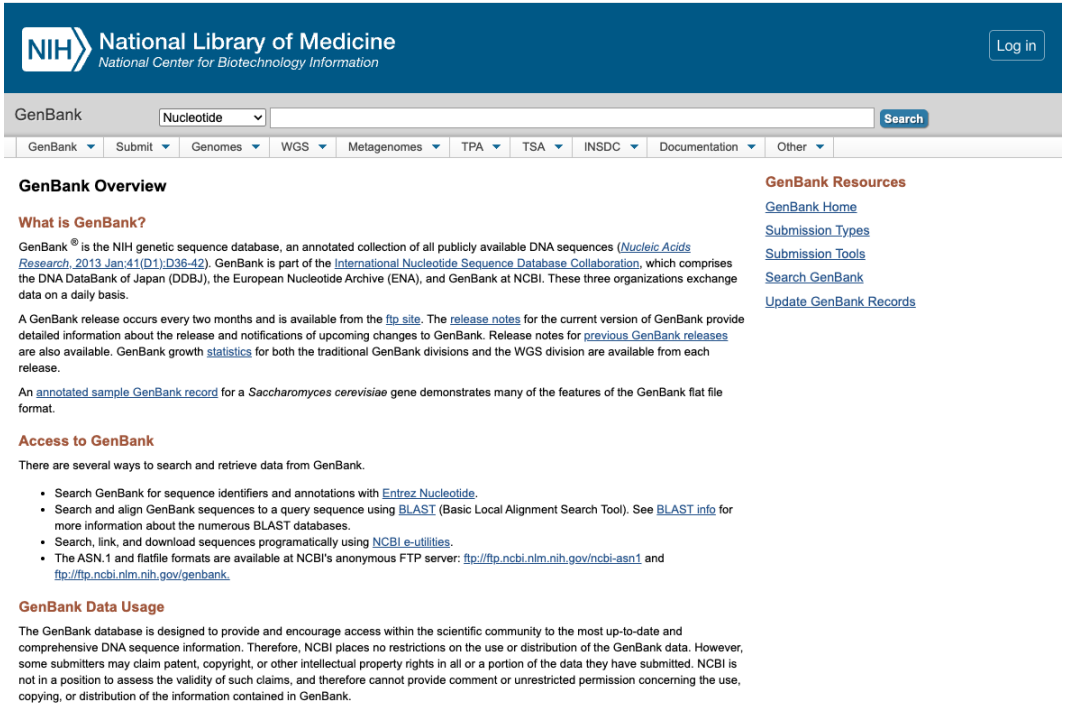
Query

ASVs

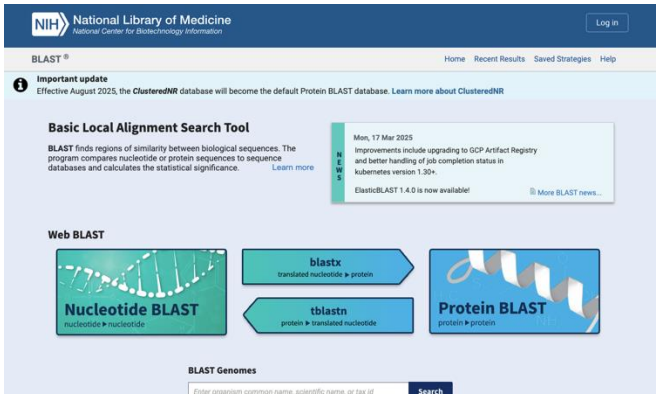
ATGCGATCGTT

ATGCGATCGCA

ATGCGAAAGTA



BLAST search



Making fasta file

```
cat Sample5_asv.csv | awk -F"," '{print $1}' | sed -e 1'd' | sed -e s/"\""//g | awk '{print "seq" NR "\t" $0}
```

.fastq file

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBR
```

Separated by “@” Sequence + Quality

.fasta file

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Separated by “>” Only sequence

```
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ cat Sample5_asv.csv | awk -F"," '{print $1}' | sed -e 1'd' | sed -e s/"\""//g | awk '{print "seq" NR "\t" $0}' | seqkit tab2fx
> Sample5_asv.fa
shumpei_yamakawa@cloudshell:~/test_meta/Sample5$ head Sample5_asv.fa
>seq1
ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq2
ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
```

126 ASVs

```
>seq1
ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq2
ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq3
ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq4
ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq5
CGACGCTCTTCCGATCTACGAGAAGACCCTATGGAGCTTTACTTTAAGGTTAAGCCTACA
AGTTTAAATGGGAACTTTTGAGGAACAAAATACCTCTGCCGACCTACTCTTCTATTCGTG
GAGAAGCTACTTTGGGGATAACAGGGTAATACAGCTGGGGAGCACTTATCGATAGTTGTT
CTTACGACCTCGATGTTGGATGATCGGAAGAGCACAC
>seq6
ACGAGAAGACCCTATGGAGCTTAATTTTACGCCTGTTTTATACCTATGCCCGGTATACAA
ATTTACATGGGGACTGTTGAGTGAGAAAATAATTTGAAACTCTTTACTATTGATCTAACT
TCTTAATTATCTGAAAAGCTACTTTGGGGATAACAGGGTGAAGCGAGTGGAGAGTTCTTA
TCGATACTTGCAATTACGACCTCGATGTTGGA
>seq7
ACGAGAAGACCCTGTGGAACTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATT
```

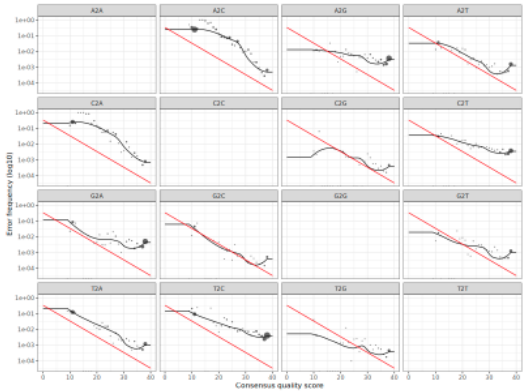
Cloud Shell Editor

File Edit Selection View Go Run ···

Search

plot_fnFs.png | plot_error_filtFs.png

home > shumpei_yamakawa > test_meta > Sample5 > plot_error_filtFs.png



restart

Upload

download

Default Mode

Safe Mode

Ephemeral Mode

theme

Usage statistics

About Cloud Shell

Privacy Policy

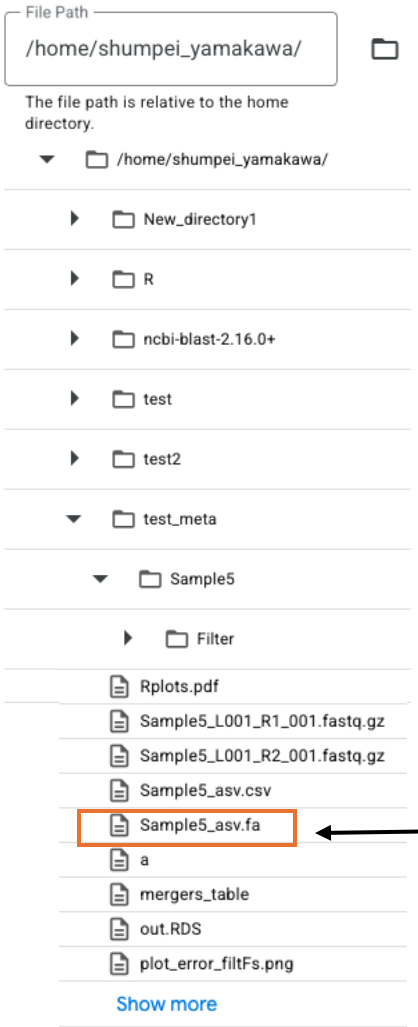
help

feedback

⊗ 0 ⚠ 0 Cloud Code - Sign in

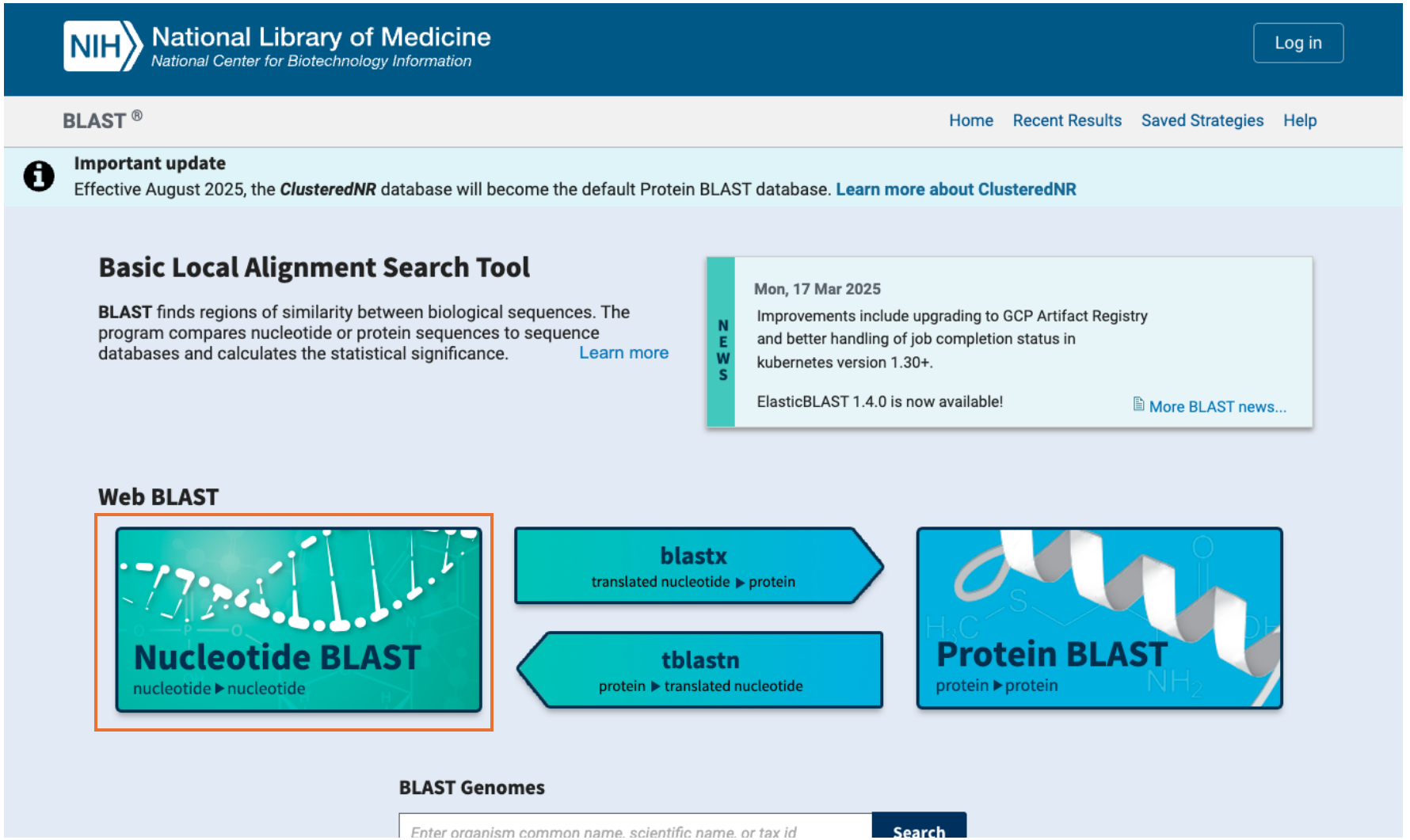
cloudshell | cloudshell

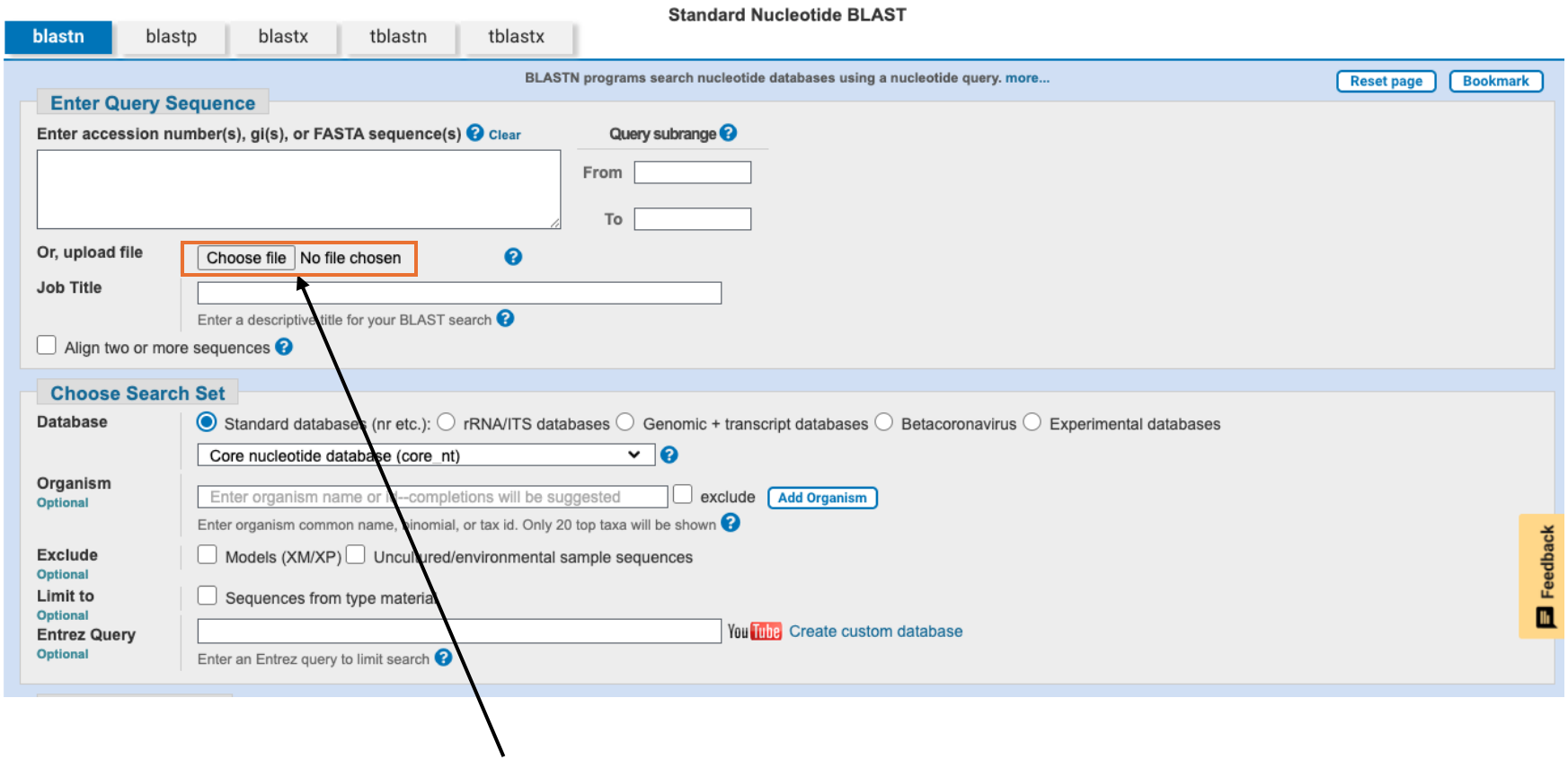
```
>seq1
ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq2
ACGAGAAGACCCTATGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACATATGACTTATATTACTTTTAGATTGATCTATTCCTTTAACT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
CACAATTACGACCTCGATGTTGGA
>seq3
ACGAGAAGACCCTGTGGAGCTTAATTTTATCGTAGCTAAACTCTGCCATTAAATTTGTAT
GGGGACTACTGAGTAAACAAATGACTTATATTACTTTTATATTGATCTATTCCTTTAATT
ATTTGAGAAGCTACTTTGGGGATAACAGGGTGTAGTGTTCGGGGAGTCCTTATCGATGAA
```



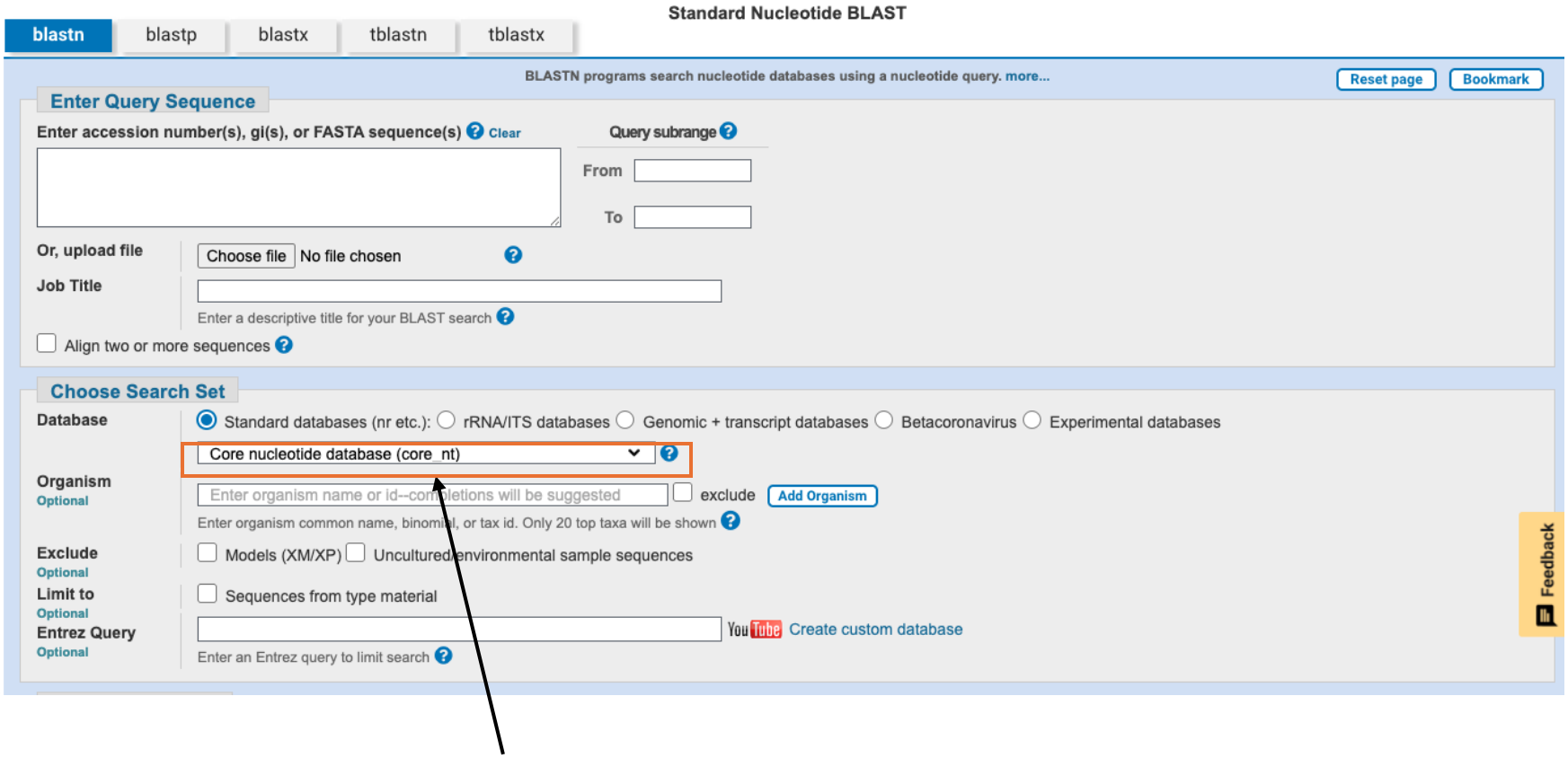
Download the fasta file to your local PC

https://blast.ncbi.nlm.nih.gov/Blast.cgi

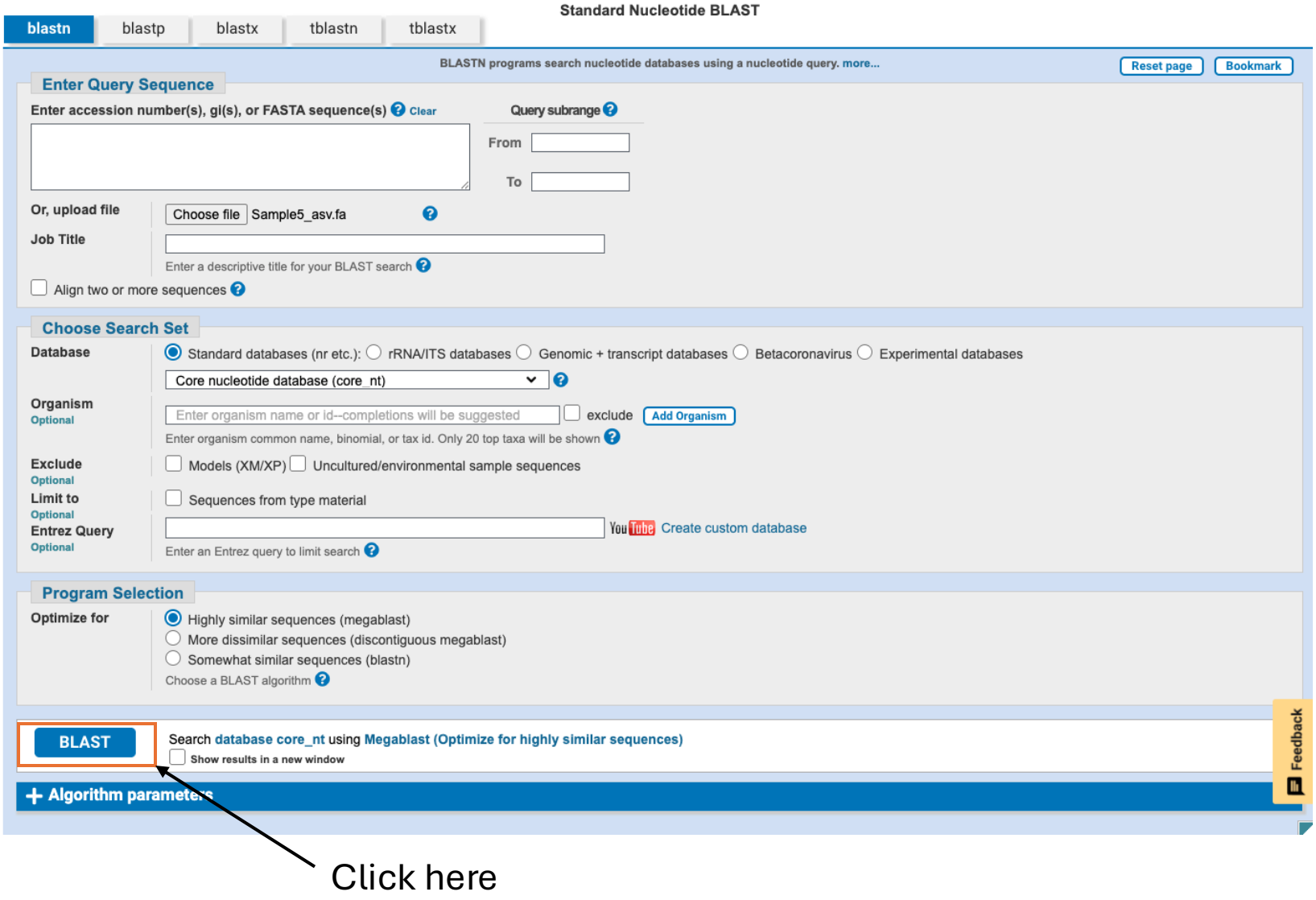




Choose “Sample_asv.fa” from you PC



Default is okay (Core nucleotide database)


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
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* Download All
Results for	1:lcl|Query_5674738 seq1(204bp)
Program	BLASTN Citation
Database	core_nt See details
Query ID	lcl|Query_5674738
Description	seq1
Molecule type	dna
Query Length	204
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage to

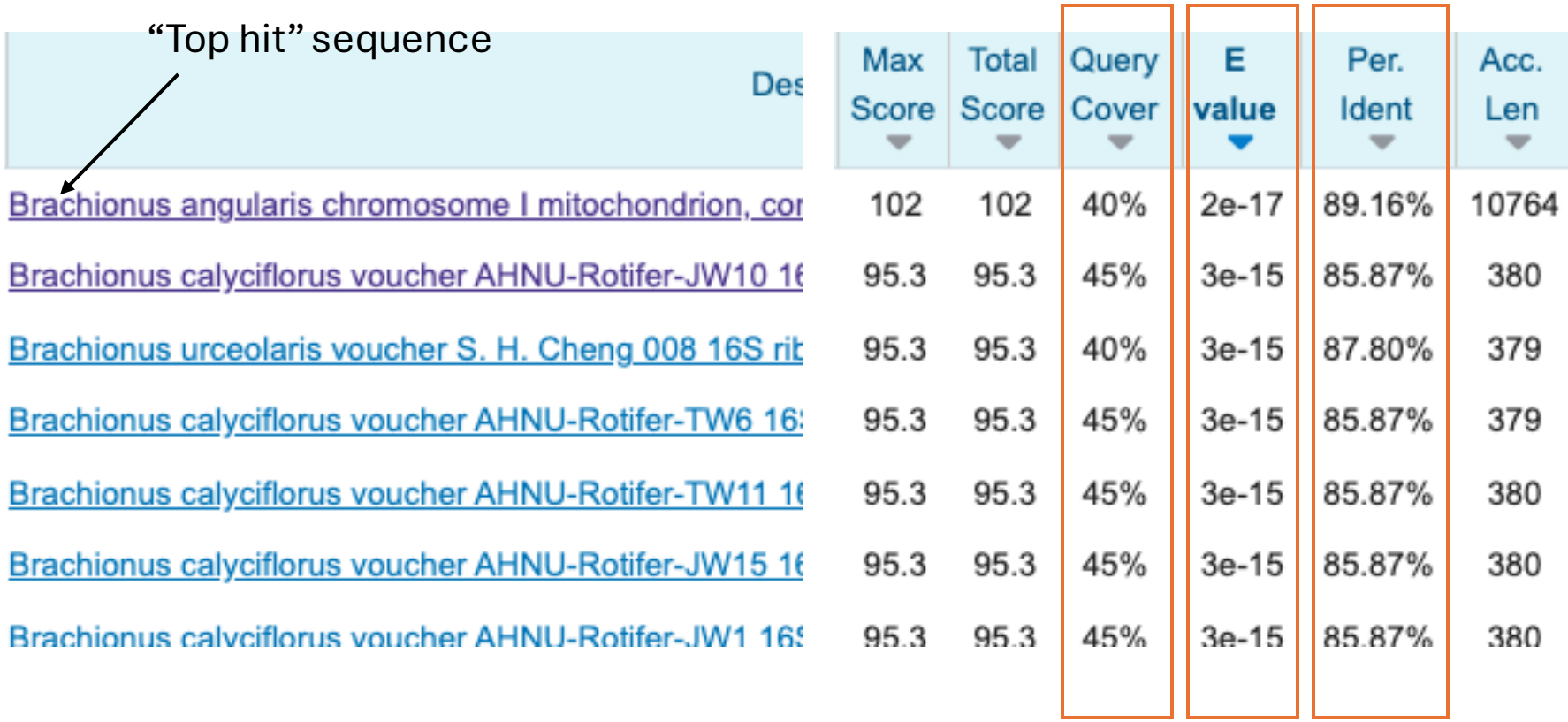
Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *72 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

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


“Top hit” sequence
Des
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Brachionus angularis chromosome I mitochondrion, cor 102 102 40% 2e-17 89.16% 10764
Brachionus calyciflorus voucher AHNU-Rotifer-JW10 16 95.3 95.3 45% 3e-15 85.87% 380
Brachionus urceolaris voucher S. H. Cheng 008 16S rib 95.3 95.3 40% 3e-15 87.80% 379
Brachionus calyciflorus voucher AHNU-Rotifer-TW6 16 95.3 95.3 45% 3e-15 85.87% 379
Brachionus calyciflorus voucher AHNU-Rotifer-TW11 16 95.3 95.3 45% 3e-15 85.87% 380
Brachionus calyciflorus voucher AHNU-Rotifer-JW15 16 95.3 95.3 45% 3e-15 85.87% 380
Brachionus calyciflorus voucher AHNU-Rotifer-JW1 16 95.3 95.3 45% 3e-15 85.87% 380

Des	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len
Brachionus angularis chromosome I mitochondrion, cor	102	102	40%	2e-17	89.16%	10764
Brachionus calyciflorus voucher AHNU-Rotifer-JW10 16	95.3	95.3	45%	3e-15	85.87%	380
Brachionus urceolaris voucher S. H. Cheng 008 16S rib	95.3	95.3	40%	3e-15	87.80%	379
Brachionus calyciflorus voucher AHNU-Rotifer-TW6 16S	95.3	95.3	45%	3e-15	85.87%	379
Brachionus calyciflorus voucher AHNU-Rotifer-TW11 16	95.3	95.3	45%	3e-15	85.87%	380
Brachionus calyciflorus voucher AHNU-Rotifer-JW15 16	95.3	95.3	45%	3e-15	85.87%	380
Brachionus calyciflorus voucher AHNU-Rotifer-JW1 16S	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    TTGAGAAGCTACTTTGGGGATAACAGGGTGTAG-TGTTCGGGGAGTCCTTATCGATGAAC  181
              ||||||||||||||||||||||||||||||||| || |||| || ||||||||||||| | |
Sbjct  10324  TTGAGAAGCTACTTTGGGGATAACAGGGTG-AGATGTTTGGAGAGTCCTTATCGATAAGC  10382

Query  182    ACAATTACGACCTCGATGTTGGA  204
              | | |||| ||||||||||||||
Sbjct  10383  ATAGTTACTACCTCGATGTTGGA  10405
```

Seq1

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Brachionus angularis chromosome I mitochondrion, complete sequence	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
Bufo bufo isolate Toad_3537 16S ribosomal RNA gene, partial sequence; mitochondrial	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
Myocastor coypus mitochondrion, complete genome	Myocastor coypus	508	508	100%	1e-139	99.64%	16874	MH182628.1

99.64%

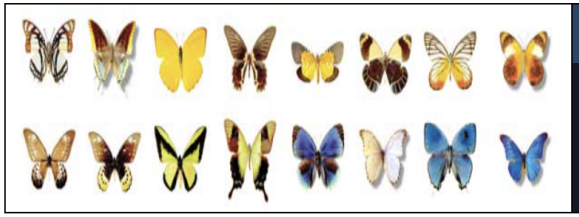
Nutria (*Mycastor coypus*)

https://www.ncbi.nlm.nih.gov/taxonomy

National Library of Medicine
National Center for Biotechnology Information

Log in

Taxonomy | Taxonomy | Search | Limits | Advanced | Hel



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
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

Other Resources

- GenBank
- LinkOut
- E-Utilities
- Batch Entrez
- INSDC

Myocastor coypus

Taxonomy ID: 10157 (for references in articles please use NCBI:txid10157)

current name

Myocastor coypus (Molina, 1782)

basionym: ***Mus coypus*** Molina, 1782

Genbank common name: **nutria**
NCBI BLAST name: **rodents**
Rank: **species**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:

common name(s)

coypu

Entrez records	
Database name	Direct links
Nucleotide	249
Protein	207
GEO Datasets	25
PubMed Central	249
Gene	37
SRA Experiments	19
Identical Protein Groups	165
BioProject	7
BioSample	25
Datasets Genome	1
Taxonomy	1

Lineage *(full)*

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Glires; Rodentia; Hystricomorpha; Myocastoridae; Myocastor

Comments and References:

image:Myocastor coypus
Image by Petar Miloaevia from Wikimedia Commons under a CC BY-SA license.
* Image may not have been verified for accuracy by NCBI Taxonomy.

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
Myocastor coypus	taxonomy/phylogenetic	Animal Diversity Web
2 records from this provider	taxonomy/phylogenetic	AnimalBase

Systematic annotation using a shell script (next week)

```
(bioinfo) shumpei@Shumpeis-MBP test % sh blast_annot.sh
Enter an ASV csv file: Sample5_asv.csv
Enter a threshold of species determination (identity of blast tophit): 95
###
Setting

blastn seqrch / database: 35267 seqs of 16S ribosormal 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
Start blast search
Complete blast search
Search taxnomony
Extract taxonomy
Summarize
done!

#####
Results

Abandance: total 31661 reads
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     Leucaspius 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,seq21,seq27,seq29,seq34,seq40,seq120
-------------------------------------------------------------------------------------------

Output files are in Sample5_asv.csv_95
```


NextSeq 550

Mycastor coypus

Gallinula chloropus

Cloeon dipterum

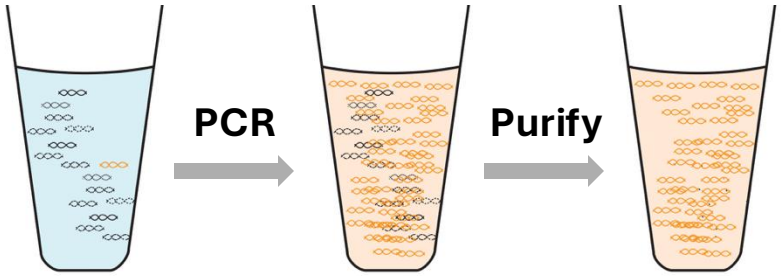
Bufo bufo

Daphnia longispina

Scapholeberis mucronata

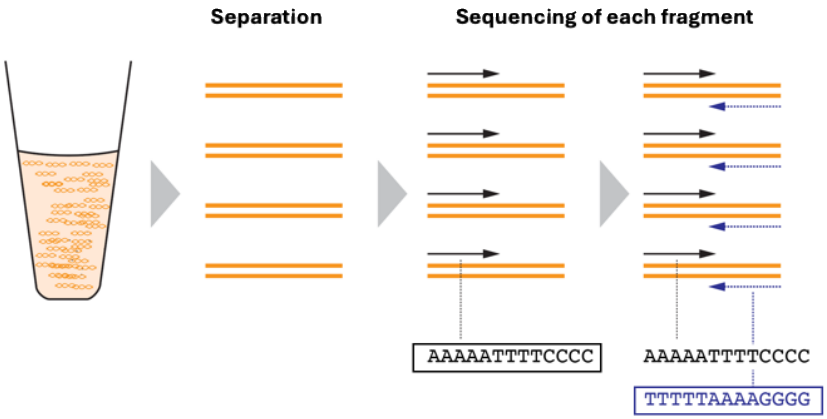
Plumatella repens

Gasterosteus aculeatus



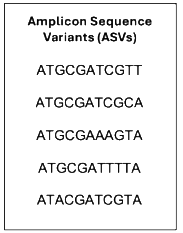
Send to the company or institute…

NGS analysis



Receive the sequencing data

Data processing



126 variants

Annotation