Long-read amplicon sequencing for microbiome analysis

Yan Hui
Postdoc
Department of Food Science
E-mail: huiyan@food.ku.dk

UNIVERSITY OF COPENHAGEN

Table of Contents

- **Reference-based and reference-free strategies for lengthy amplicon analysis**
- **De novo OTU picking from long amplicons with LACA**
- **Use NART for long amplicon profiling by read classification**
- **Exercises**

Reference-dependent **vs** Reference-free analysis

OUTPUT	RRF-dependent	REF-free
Representative sequences	No	Yes
Phylogenetic tree	No	Yes

EPI2ME

Per-read query against a known database:

- Limited by database
- No OTUs

illumina®

The consensus from built clusters:

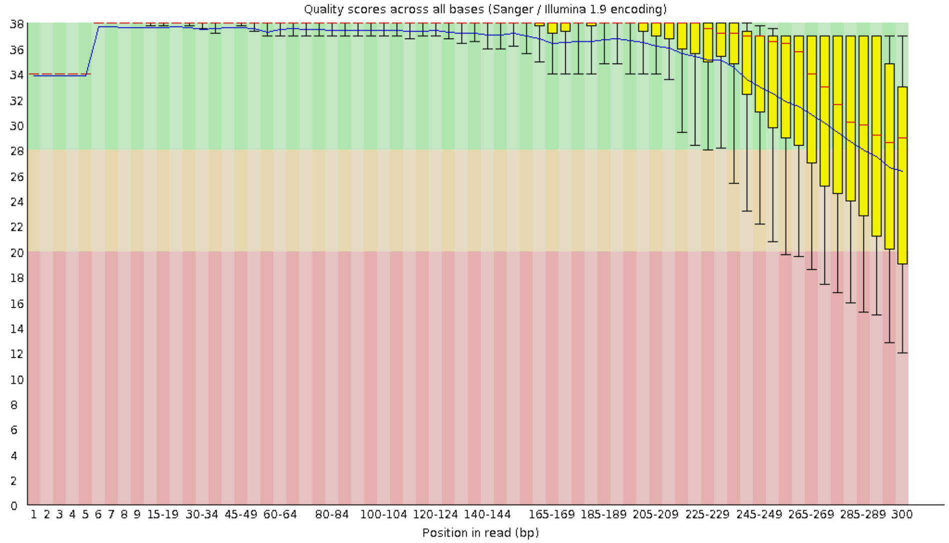
- Clustering by identity, etc.

Sequencing errors

Phred quality scores Q are logarithmically related to the base-calling error probabilities P and defined as

$$Q = -10\ \log_{10} P.$$

illumina

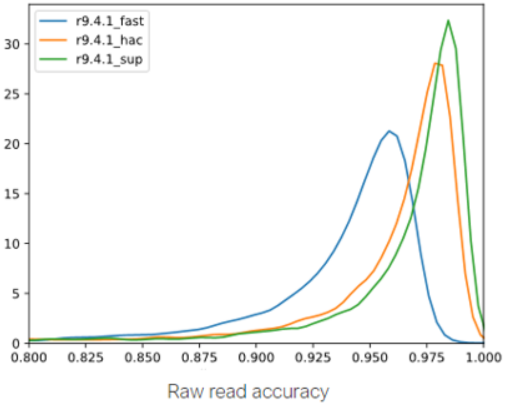


PCR >>>
More errors in the end

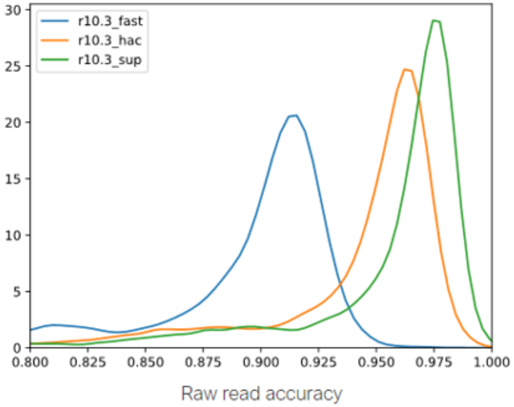
Oxford NANOPORE Technologies

High error rate

R9.4.1



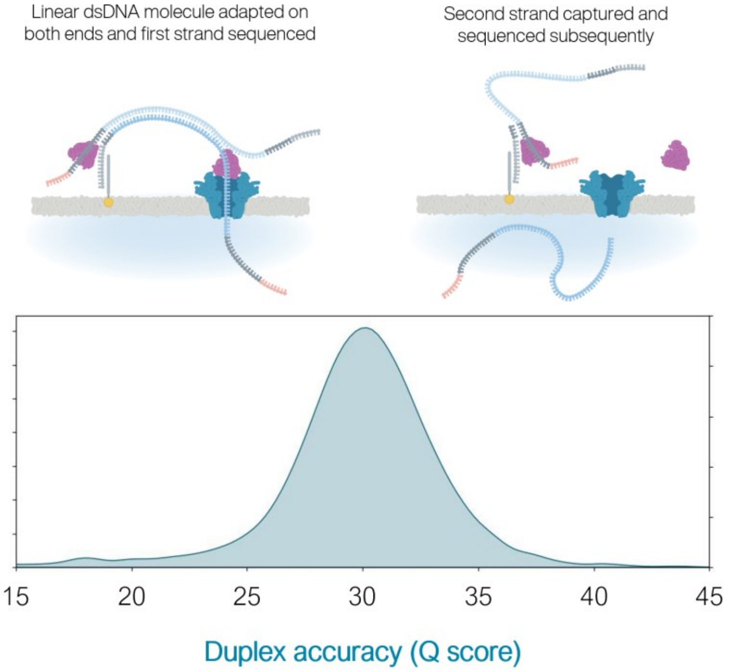
R10.3



Bio-pore >>>
Random error

Molecule-level correction

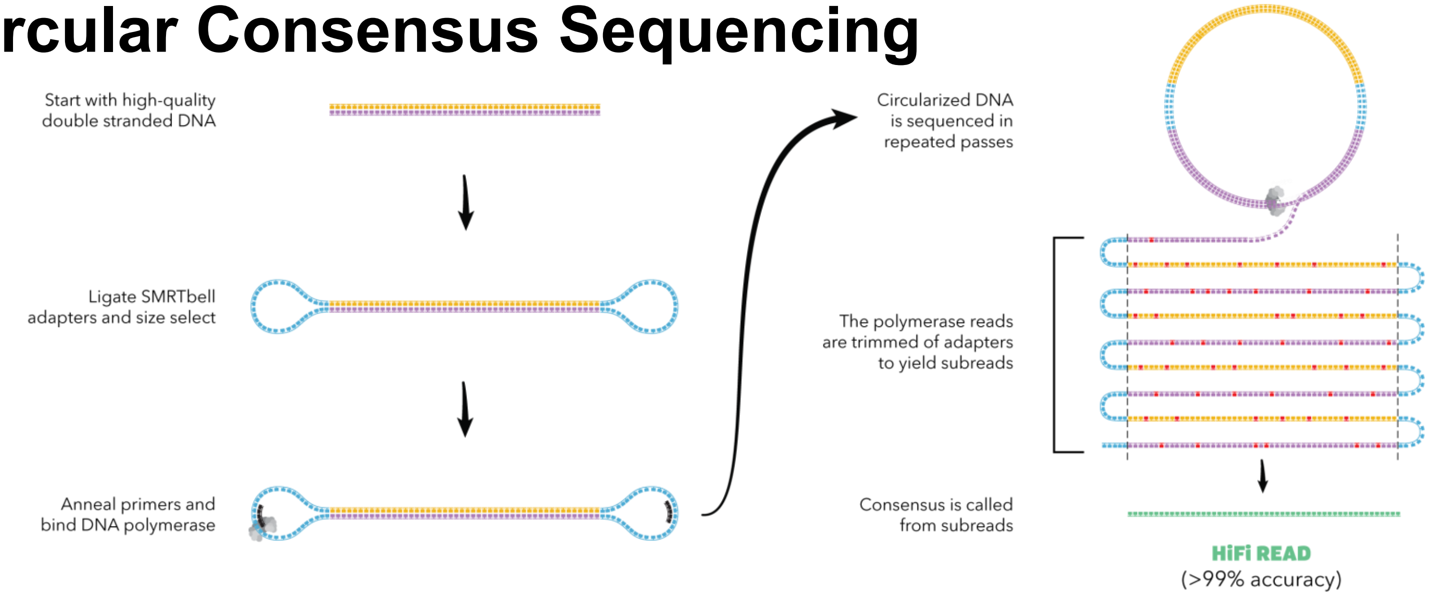
ONT Duplex



Linear dsDNA molecule adapted on both ends and first strand sequenced

Second strand captured and sequenced subsequently

PacBio Circular Consensus Sequencing (CCS)



Start with high-quality double stranded DNA

Ligate SMRTbell adapters and size select

Anneal primers and bind DNA polymerase

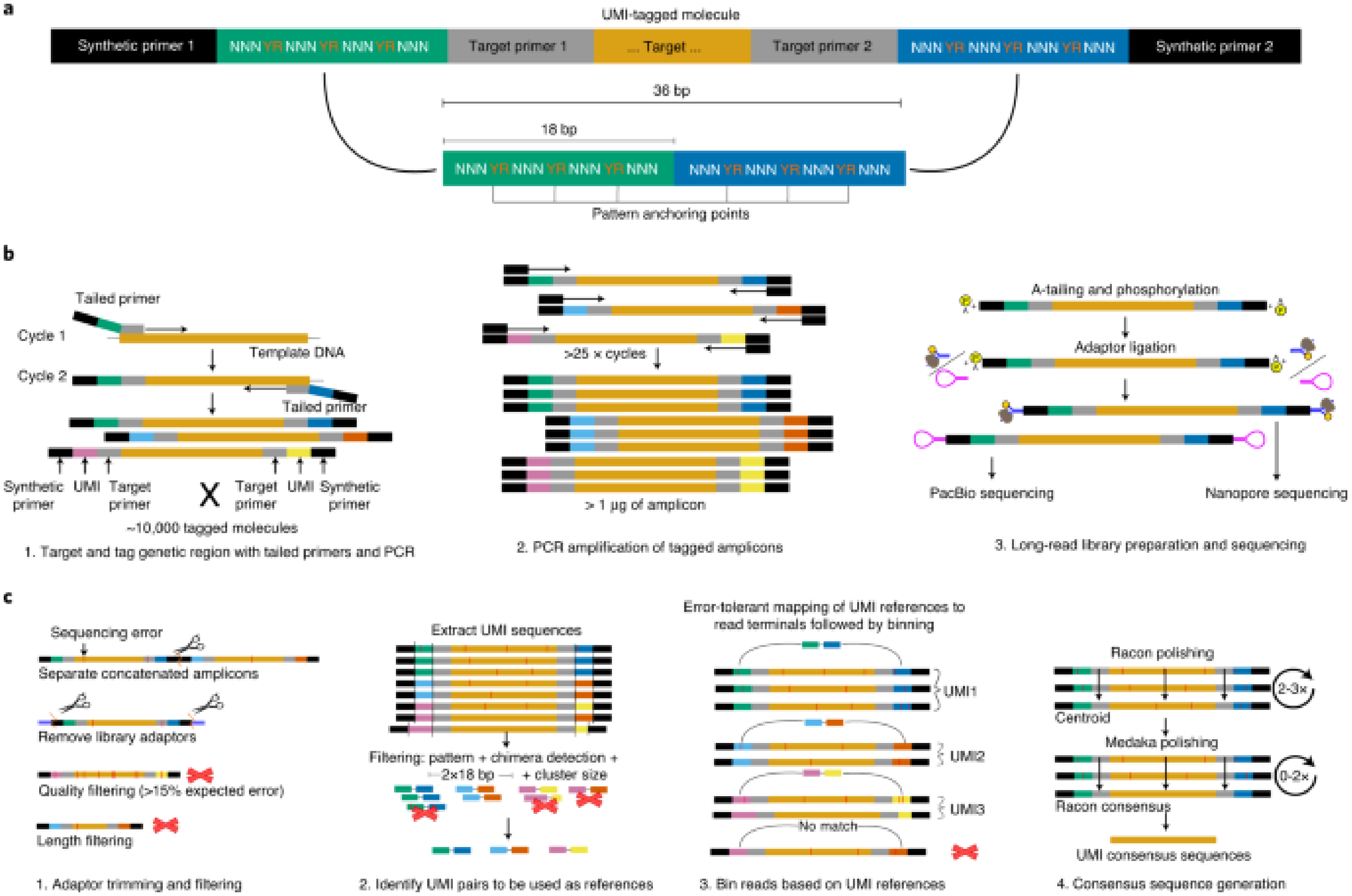
Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads

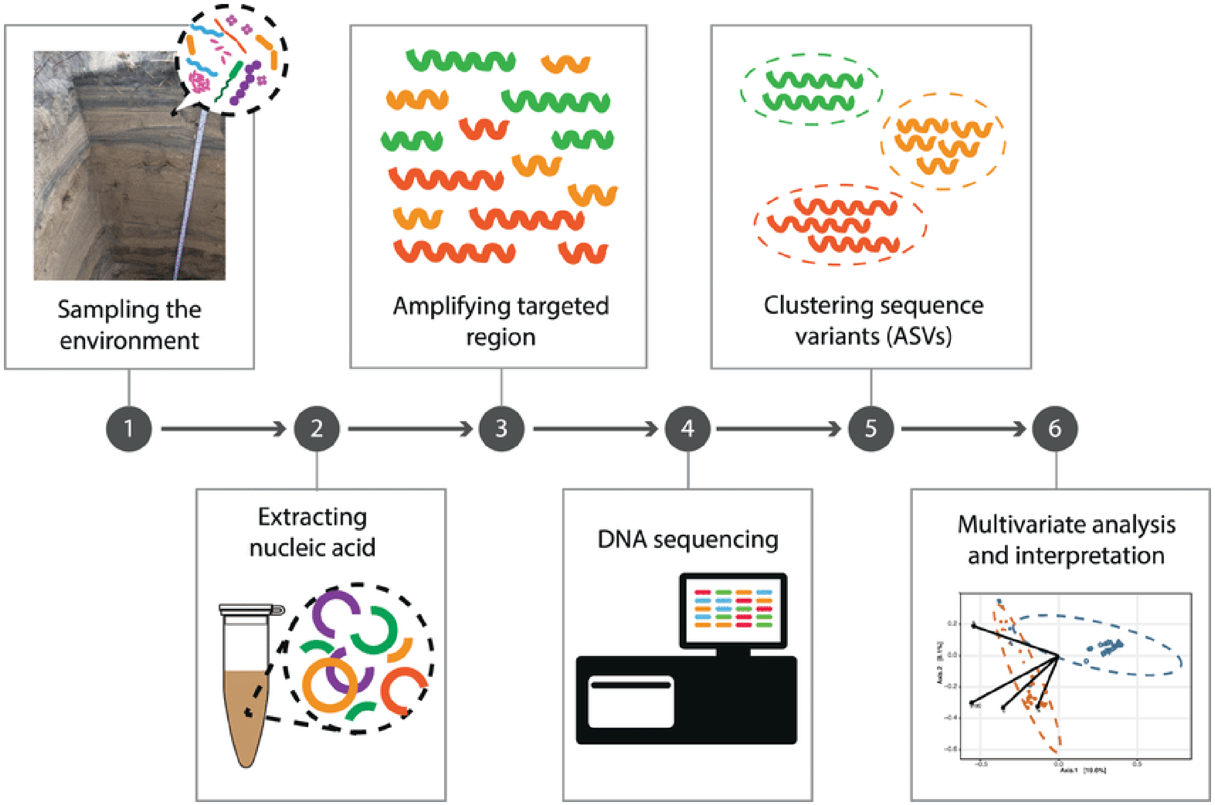
Molecule-level correction

Specialized library preparation with e.g., UMI



https://www.nature.com/articles/s41592-020-01041-y

Clustering-based correction


Sampling the environment
1
2
Extracting nucleic acid
Amplifying targeted region
3
4
DNA sequencing
Clustering sequence variants (ASVs)
5
6
Multivariate analysis and interpretation

Metagenomics: microbes in uneven abundance
UMI -> different template (including the phylogenetically same one)

https://www.nature.com/articles/s43247-021-00199-3

Clustering-based correction

Troubles with long-read alignment

- **Pairwise alignment**

Time complexity

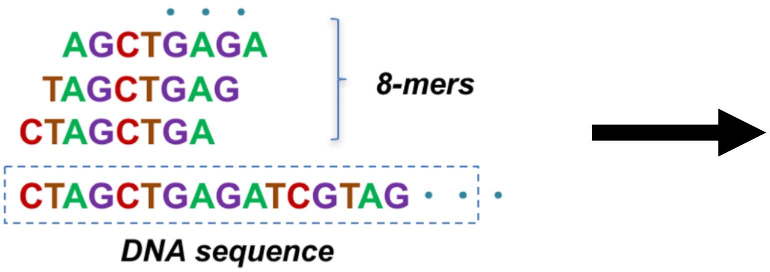
O(C(n,2)/2)=O(n!(n-1)!/2)

For amplicons, n can be millions if reads are pooled.

- **Noisy alignment with long reads**

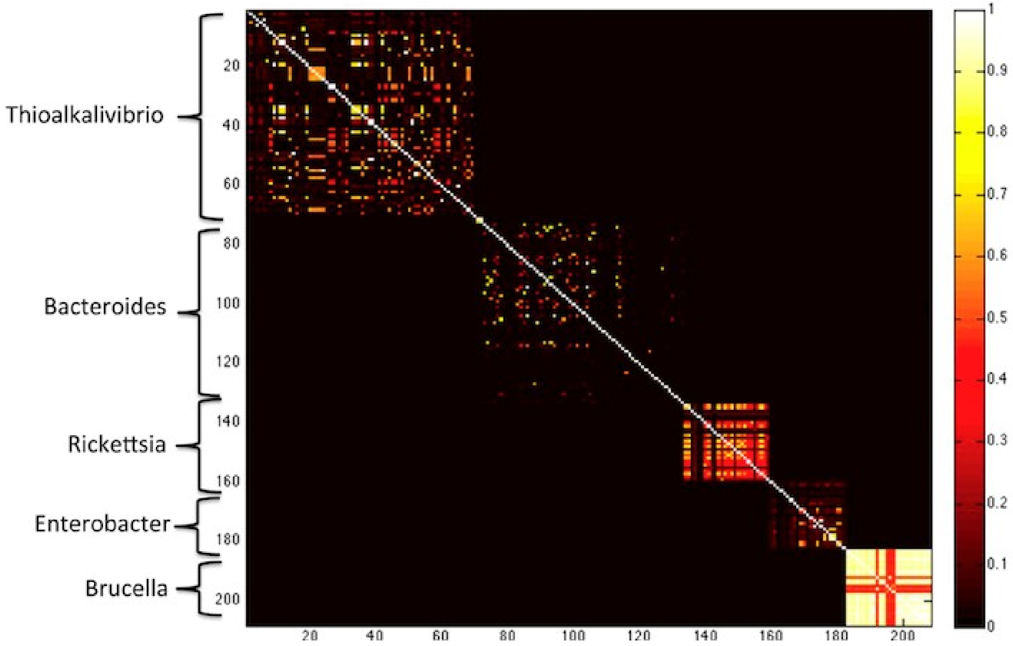
The relatively **high** error rate in relatively **long** reads

K-mers binning

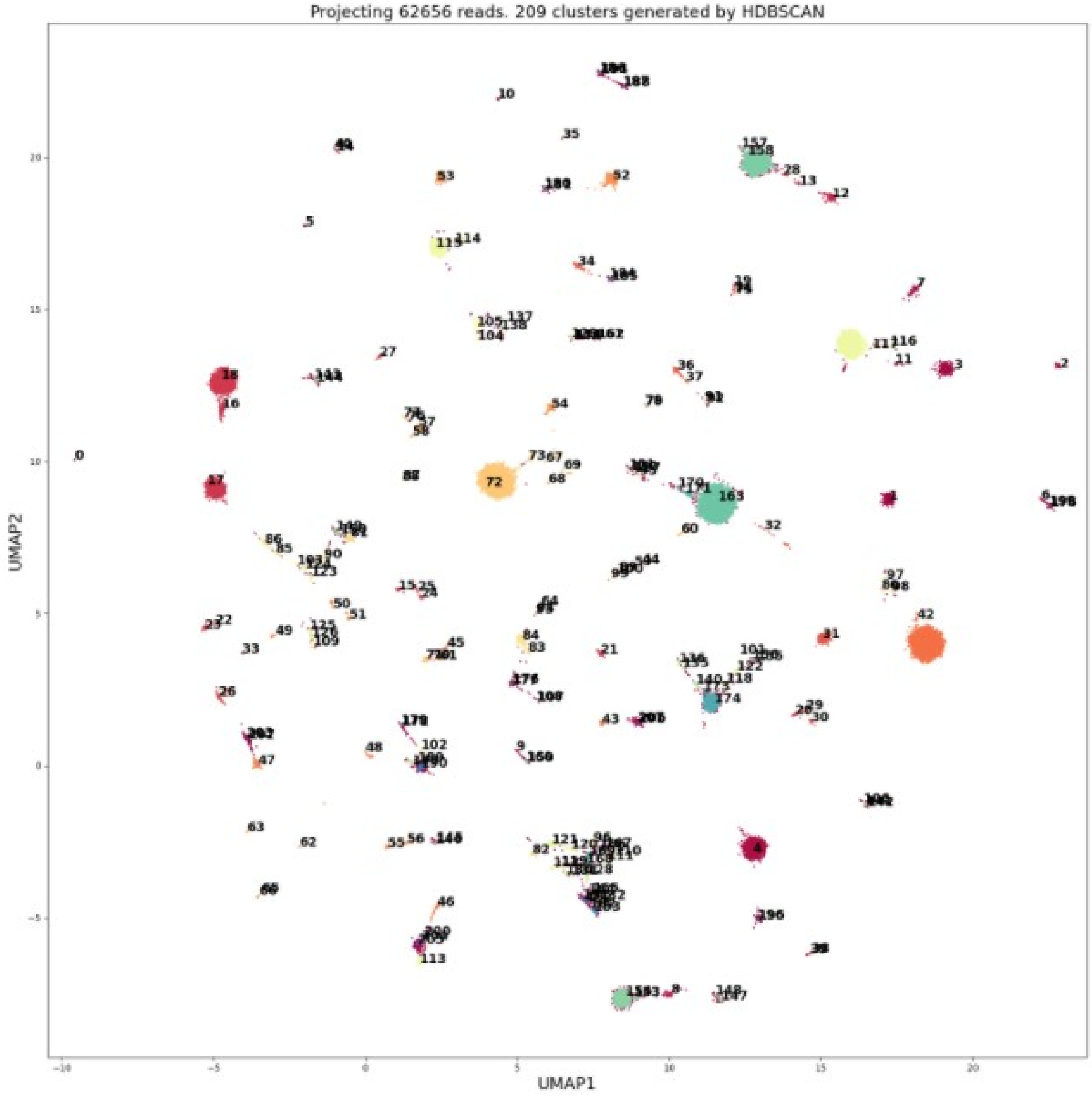

. . .
AGCTGAGA
TAGCTGAG
CTAGCTGA
8-mers
CTAGCTGAGATCGTAG . . .
DNA sequence

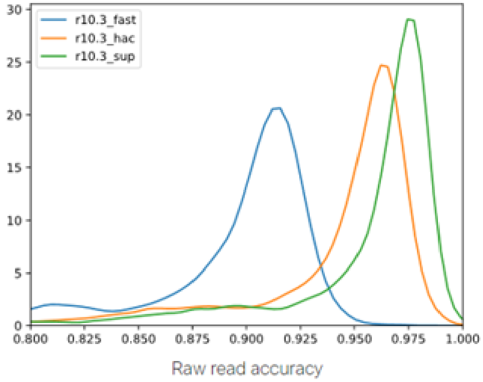
	atcga	tcgac	…	cgaaa
read1	1	6		7
read2	3	1		3
…				
readn	5	2		2

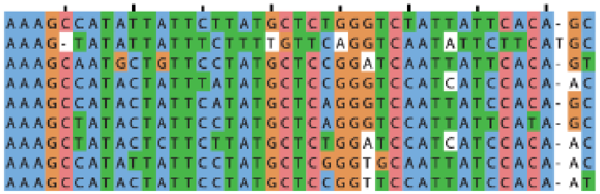
- Computers prefer k-mers than text: blast, binning
- Unique k-mer patterns between genomes



Pre-cluster: Use 5-kmer profiles to bin ONT reads







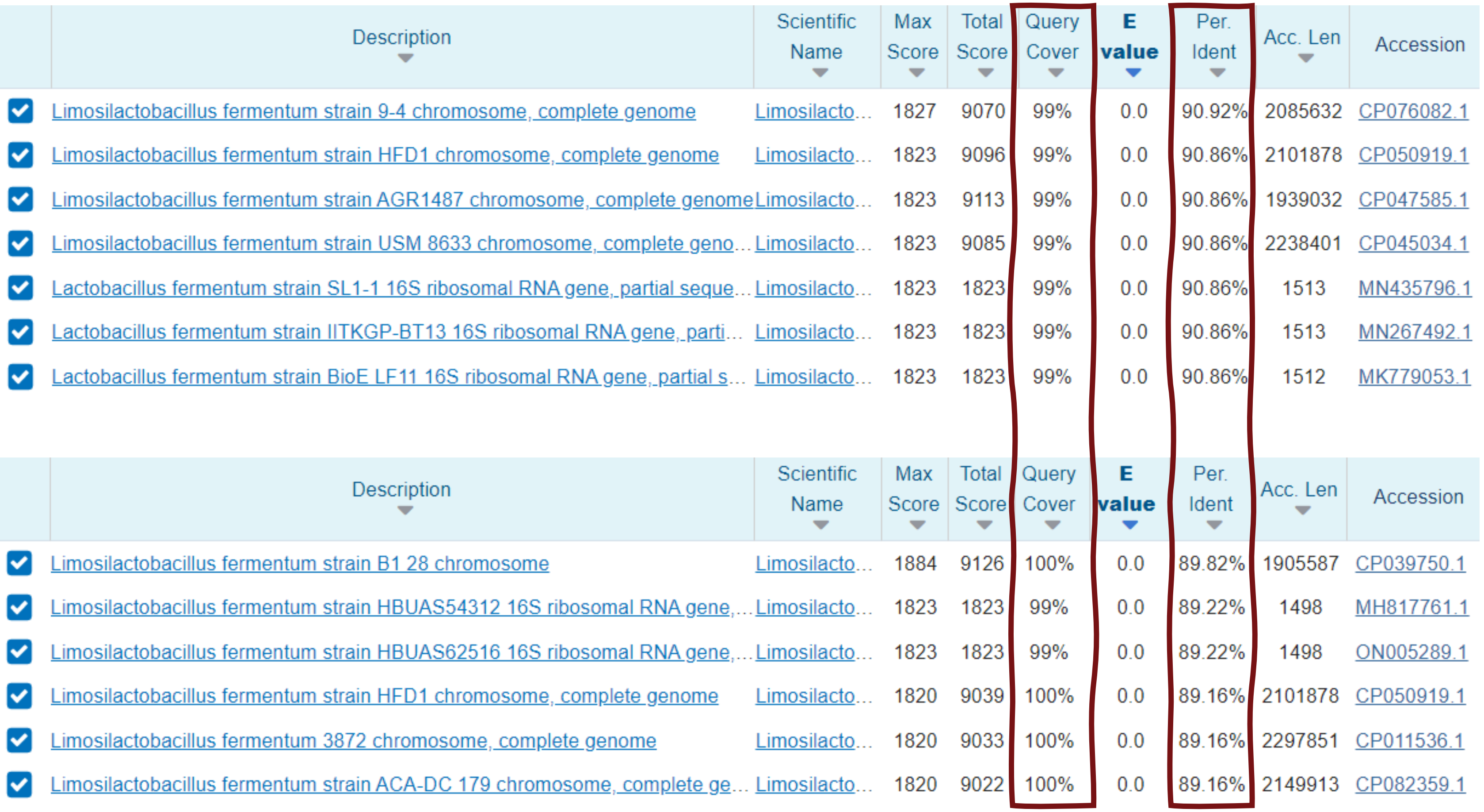
```
AAAGCCATATTATTCTTATGCTCTGGGTCTATTATTCACA-GC
AAAG-TATATTATTTCTTTTGTTCAGGTCAATATTCTTCATGC
AAAGCAATGCTGTTCCTATGCTCCGGATCAATTATTCACA-GT
AAAGCCATACTATTTATATGCTCCGGGTCCATCATCCACA-AC
AAAGCCATACTATTCATATGCTCAGGGTCAATTATCCACA-GC
AAAGCTATACTATTCCTATGCTCAGGGTCCATTATTCATA-GC
AAAGCTATACTCTTCTTATGCTCTGGATCCATCATCCACA-AC
AAAGCCATATTATTCCTATGCTCGGGTGCAATTATCCACA-AC
AAAGCCATACTATTCCTATGCTCCGGTTCCATTATCCACA-AT
```

	atcga	tcgac	…	cgaaa
read1	1	10		3
read2	1	10		4
…				
readn	1	10		3

*Take **blast** result as an example*

Raw reads within the cluster

Raw read



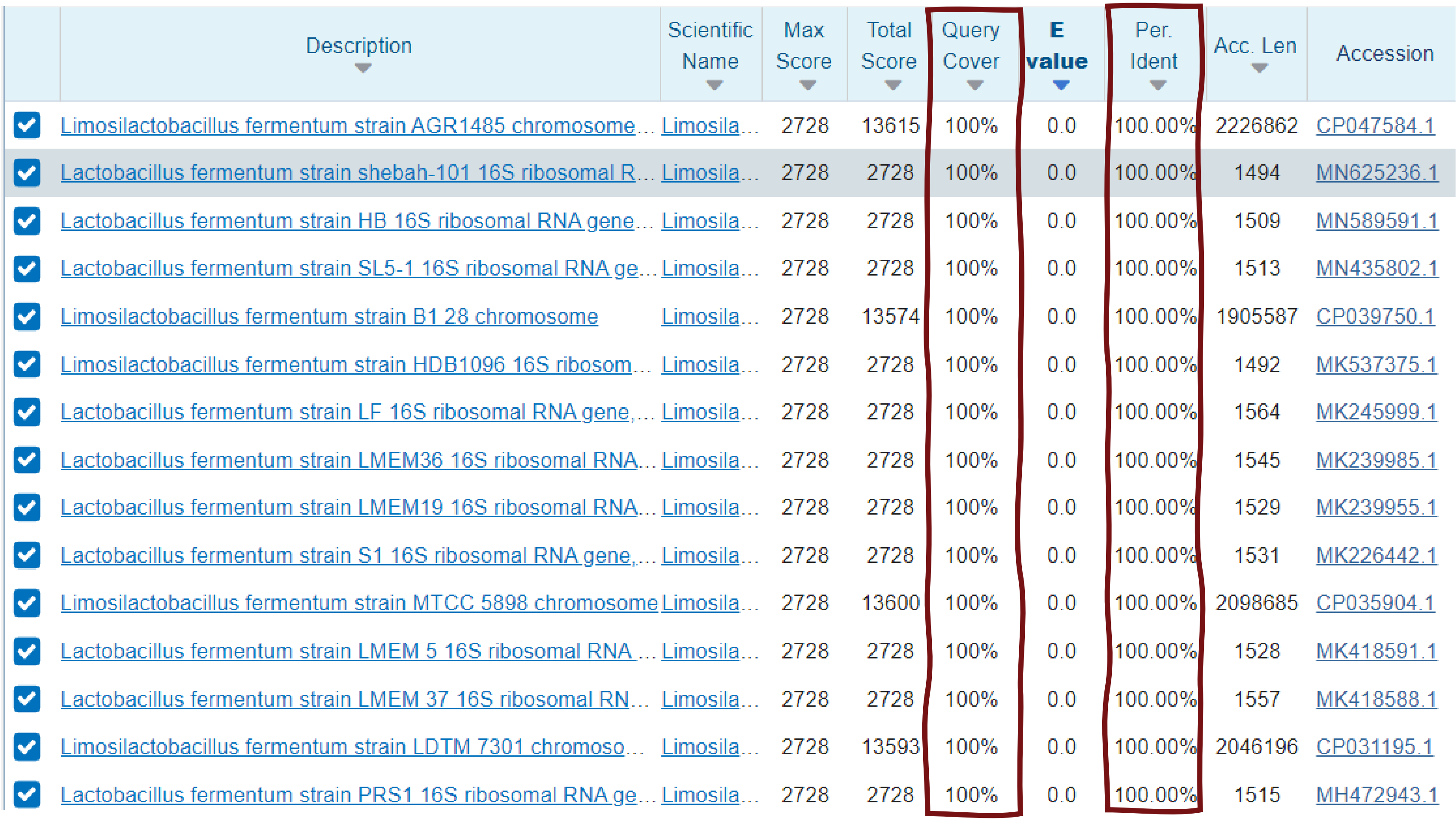
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Limosilactobacillus fermentum strain 9-4 chromosome, complete genome	Limosilacto...	1827	9070	99%	0.0	90.92%	2085632	CP076082.1
☑	Limosilactobacillus fermentum strain HFD1 chromosome, complete genome	Limosilacto...	1823	9096	99%	0.0	90.86%	2101878	CP050919.1
☑	Limosilactobacillus fermentum strain AGR1487 chromosome, complete genome	Limosilacto...	1823	9113	99%	0.0	90.86%	1939032	CP047585.1
☑	Limosilactobacillus fermentum strain USM 8633 chromosome, complete geno...	Limosilacto...	1823	9085	99%	0.0	90.86%	2238401	CP045034.1
☑	Lactobacillus fermentum strain SL1-1 16S ribosomal RNA gene, partial seque...	Limosilacto...	1823	1823	99%	0.0	90.86%	1513	MN435796.1
☑	Lactobacillus fermentum strain IITKGP-BT13 16S ribosomal RNA gene, parti...	Limosilacto...	1823	1823	99%	0.0	90.86%	1513	MN267492.1
☑	Lactobacillus fermentum strain BioE LF11 16S ribosomal RNA gene, partial s...	Limosilacto...	1823	1823	99%	0.0	90.86%	1512	MK779053.1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Limosilactobacillus fermentum strain B1 28 chromosome	Limosilacto...	1884	9126	100%	0.0	89.82%	1905587	CP039750.1
☑	Limosilactobacillus fermentum strain HBUAS54312 16S ribosomal RNA gene,...	Limosilacto...	1823	1823	99%	0.0	89.22%	1498	MH817761.1
☑	Limosilactobacillus fermentum strain HBUAS62516 16S ribosomal RNA gene,...	Limosilacto...	1823	1823	99%	0.0	89.22%	1498	ON005289.1
☑	Limosilactobacillus fermentum strain HFD1 chromosome, complete genome	Limosilacto...	1820	9039	100%	0.0	89.16%	2101878	CP050919.1
☑	Limosilactobacillus fermentum 3872 chromosome, complete genome	Limosilacto...	1820	9033	100%	0.0	89.16%	2297851	CP011536.1
☑	Limosilactobacillus fermentum strain ACA-DC 179 chromosome, complete ge...	Limosilacto...	1820	9022	100%	0.0	89.16%	2149913	CP082359.1

*Take **blast** result as an example*

Denoised consensus

Polished consensus



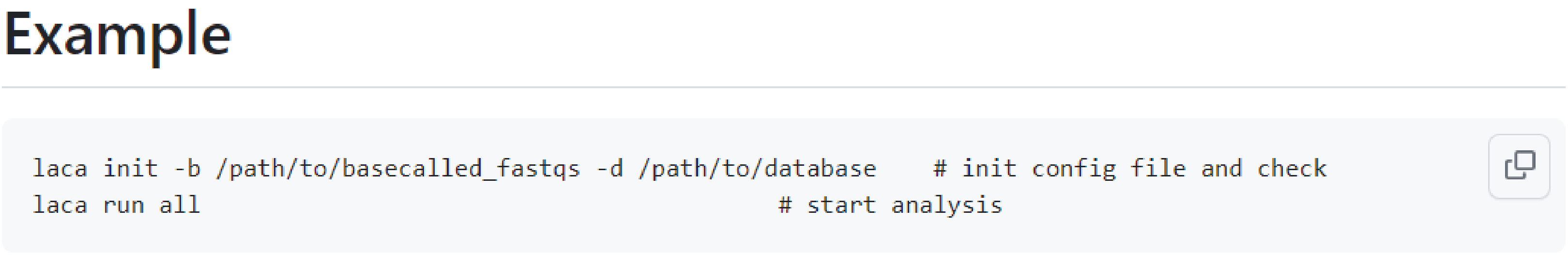
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Limosilactobacillus fermentum strain AGR1485 chromosome…	Limosila…	2728	13615	100%	0.0	100.00%	2226862	CP047584.1
☑	Lactobacillus fermentum strain shebah-101 16S ribosomal R…	Limosila…	2728	2728	100%	0.0	100.00%	1494	MN625236.1
☑	Lactobacillus fermentum strain HB 16S ribosomal RNA gene…	Limosila…	2728	2728	100%	0.0	100.00%	1509	MN589591.1
☑	Lactobacillus fermentum strain SL5-1 16S ribosomal RNA ge…	Limosila…	2728	2728	100%	0.0	100.00%	1513	MN435802.1
☑	Limosilactobacillus fermentum strain B1 28 chromosome	Limosila…	2728	13574	100%	0.0	100.00%	1905587	CP039750.1
☑	Limosilactobacillus fermentum strain HDB1096 16S ribosom…	Limosila…	2728	2728	100%	0.0	100.00%	1492	MK537375.1
☑	Lactobacillus fermentum strain LF 16S ribosomal RNA gene,…	Limosila…	2728	2728	100%	0.0	100.00%	1564	MK245999.1
☑	Lactobacillus fermentum strain LMEM36 16S ribosomal RNA…	Limosila…	2728	2728	100%	0.0	100.00%	1545	MK239985.1
☑	Lactobacillus fermentum strain LMEM19 16S ribosomal RNA…	Limosila…	2728	2728	100%	0.0	100.00%	1529	MK239955.1
☑	Lactobacillus fermentum strain S1 16S ribosomal RNA gene,…	Limosila…	2728	2728	100%	0.0	100.00%	1531	MK226442.1
☑	Limosilactobacillus fermentum strain MTCC 5898 chromosome	Limosila…	2728	13600	100%	0.0	100.00%	2098685	CP035904.1
☑	Lactobacillus fermentum strain LMEM 5 16S ribosomal RNA …	Limosila…	2728	2728	100%	0.0	100.00%	1528	MK418591.1
☑	Lactobacillus fermentum strain LMEM 37 16S ribosomal RN…	Limosila…	2728	2728	100%	0.0	100.00%	1557	MK418588.1
☑	Limosilactobacillus fermentum strain LDTM 7301 chromoso…	Limosila…	2728	13593	100%	0.0	100.00%	2046196	CP031195.1
☑	Lactobacillus fermentum strain PRS1 16S ribosomal RNA ge…	Limosila…	2728	2728	100%	0.0	100.00%	1515	MH472943.1

De novo OTU picking from long amplicons with **LACA**

UNIVERSITY OF COPENHAGEN

LACA: an automatic workflow for Long Amplicon Consensus Analysis

- GitHub: https://github.com/yanhui09/laca



Example

```
laca init -b /path/to/basecalled_fastqs -d /path/to/database    # init config file and check
laca run all                                         # start analysis
```

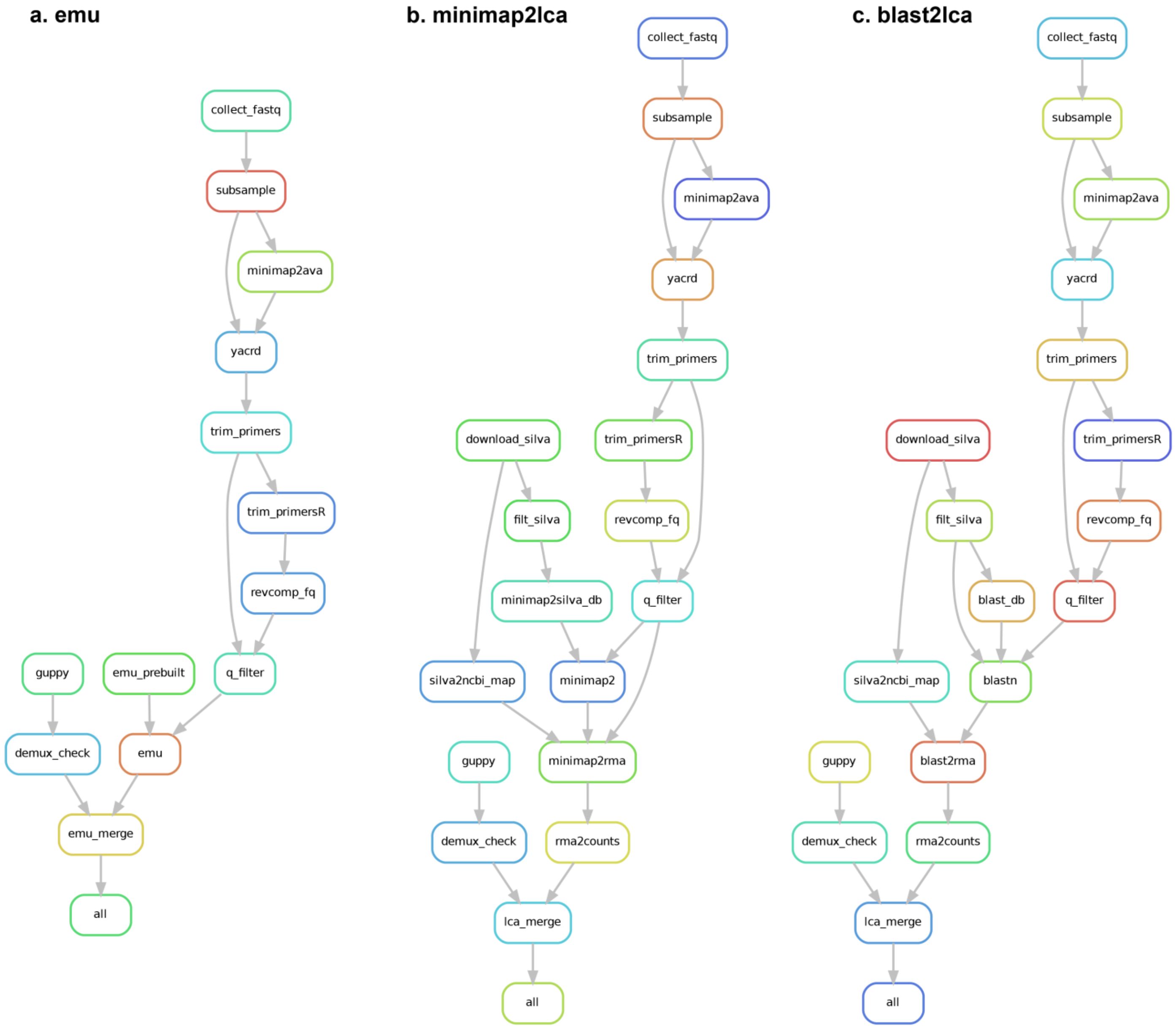
Use **NART** for long amplicon profiling by read classification

UNIVERSITY OF COPENHAGEN

NART: A tool for Nanopore Amplicon Real-Time analysis

- GitHub: https://github.com/yanhui09/nart
- Demo video: https://www.youtube.com/watch?v=TkdJGLOscPg

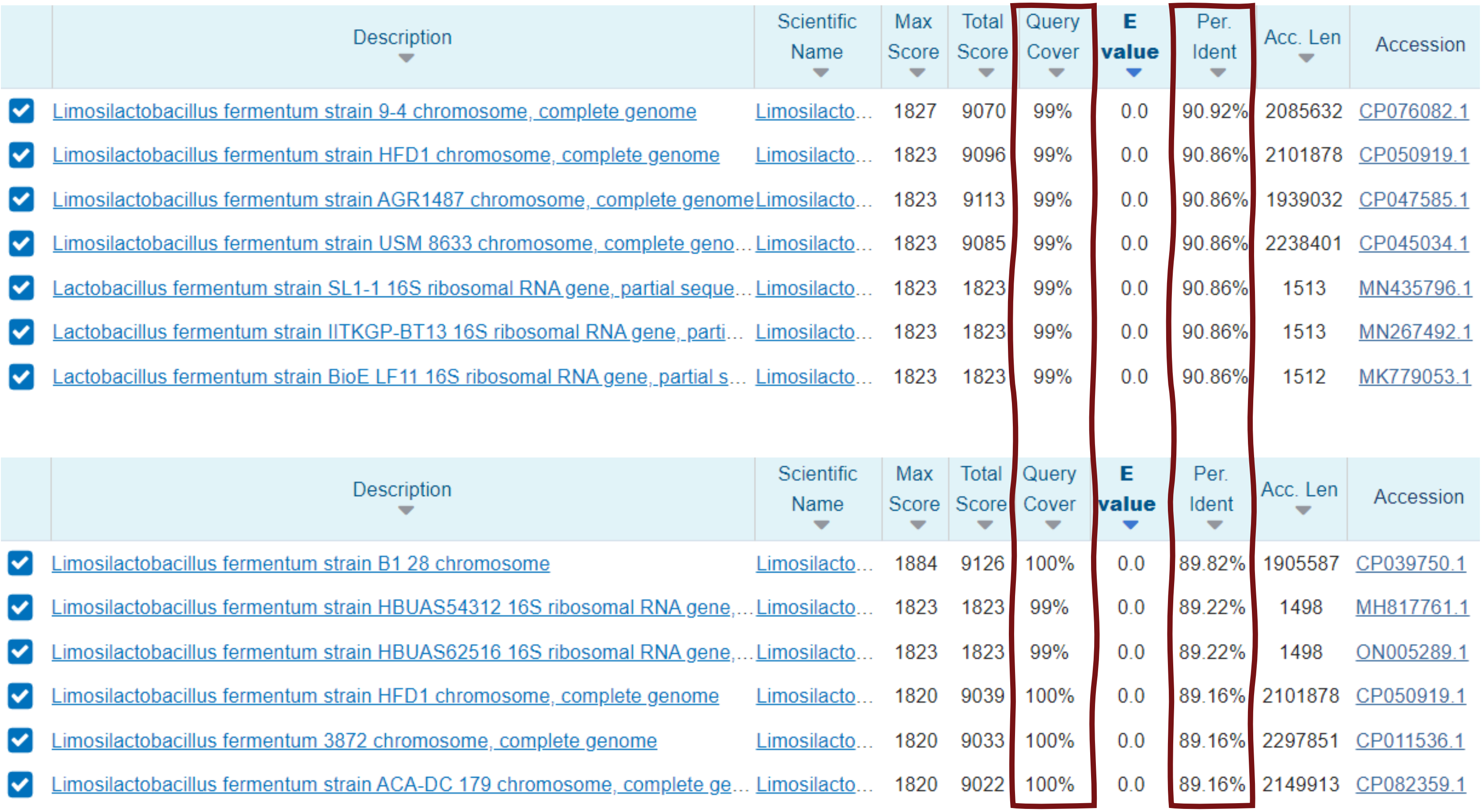
Directed Acyclic Graph (DAG)



Lowest Common Ancestor by read classification (minimap2lca, blast2lca)

Limosilactobacillus fermentum

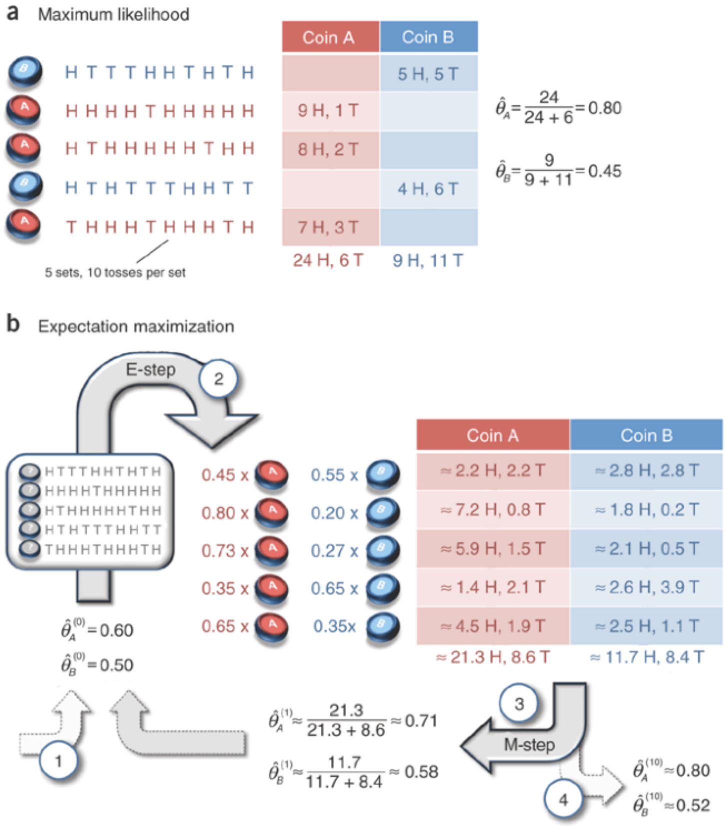
Raw read



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Limosilactobacillus fermentum strain 9-4 chromosome, complete genome	Limosilacto...	1827	9070	99%	0.0	90.92%	2085632	CP076082.1
☑	Limosilactobacillus fermentum strain HFD1 chromosome, complete genome	Limosilacto...	1823	9096	99%	0.0	90.86%	2101878	CP050919.1
☑	Limosilactobacillus fermentum strain AGR1487 chromosome, complete genome	Limosilacto...	1823	9113	99%	0.0	90.86%	1939032	CP047585.1
☑	Limosilactobacillus fermentum strain USM 8633 chromosome, complete geno...	Limosilacto...	1823	9085	99%	0.0	90.86%	2238401	CP045034.1
☑	Lactobacillus fermentum strain SL1-1 16S ribosomal RNA gene, partial seque...	Limosilacto...	1823	1823	99%	0.0	90.86%	1513	MN435796.1
☑	Lactobacillus fermentum strain IITKGP-BT13 16S ribosomal RNA gene, parti...	Limosilacto...	1823	1823	99%	0.0	90.86%	1513	MN267492.1
☑	Lactobacillus fermentum strain BioE LF11 16S ribosomal RNA gene, partial s...	Limosilacto...	1823	1823	99%	0.0	90.86%	1512	MK779053.1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Limosilactobacillus fermentum strain B1 28 chromosome	Limosilacto...	1884	9126	100%	0.0	89.82%	1905587	CP039750.1
☑	Limosilactobacillus fermentum strain HBUAS54312 16S ribosomal RNA gene,...	Limosilacto...	1823	1823	99%	0.0	89.22%	1498	MH817761.1
☑	Limosilactobacillus fermentum strain HBUAS62516 16S ribosomal RNA gene,...	Limosilacto...	1823	1823	99%	0.0	89.22%	1498	ON005289.1
☑	Limosilactobacillus fermentum strain HFD1 chromosome, complete genome	Limosilacto...	1820	9039	100%	0.0	89.16%	2101878	CP050919.1
☑	Limosilactobacillus fermentum 3872 chromosome, complete genome	Limosilacto...	1820	9033	100%	0.0	89.16%	2297851	CP011536.1
☑	Limosilactobacillus fermentum strain ACA-DC 179 chromosome, complete ge...	Limosilacto...	1820	9022	100%	0.0	89.16%	2149913	CP082359.1

Emu: Species-level abundance estimation through an expectation–maximization algorithm



	Coin A	Coin B
H T T T H H T H T H		5 H, 5 T
H H H H T H H H H H	9 H, 1 T	
H T H H H H H T H H	8 H, 2 T	
H T H T T T H H T T		4 H, 6 T
T H H H T H H H T H	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

$\hat{\theta}_A = \frac{24}{24+6} = 0.80$

$\hat{\theta}_B = \frac{9}{9+11} = 0.45$

HTTTHHTHTH
HHHHTHHHHH
HTHHHHHTHH
HTHTTTHHTT
THHHTHHHTH

$\hat{\theta}_A^{(0)} = 0.60$

$\hat{\theta}_B^{(0)} = 0.50$

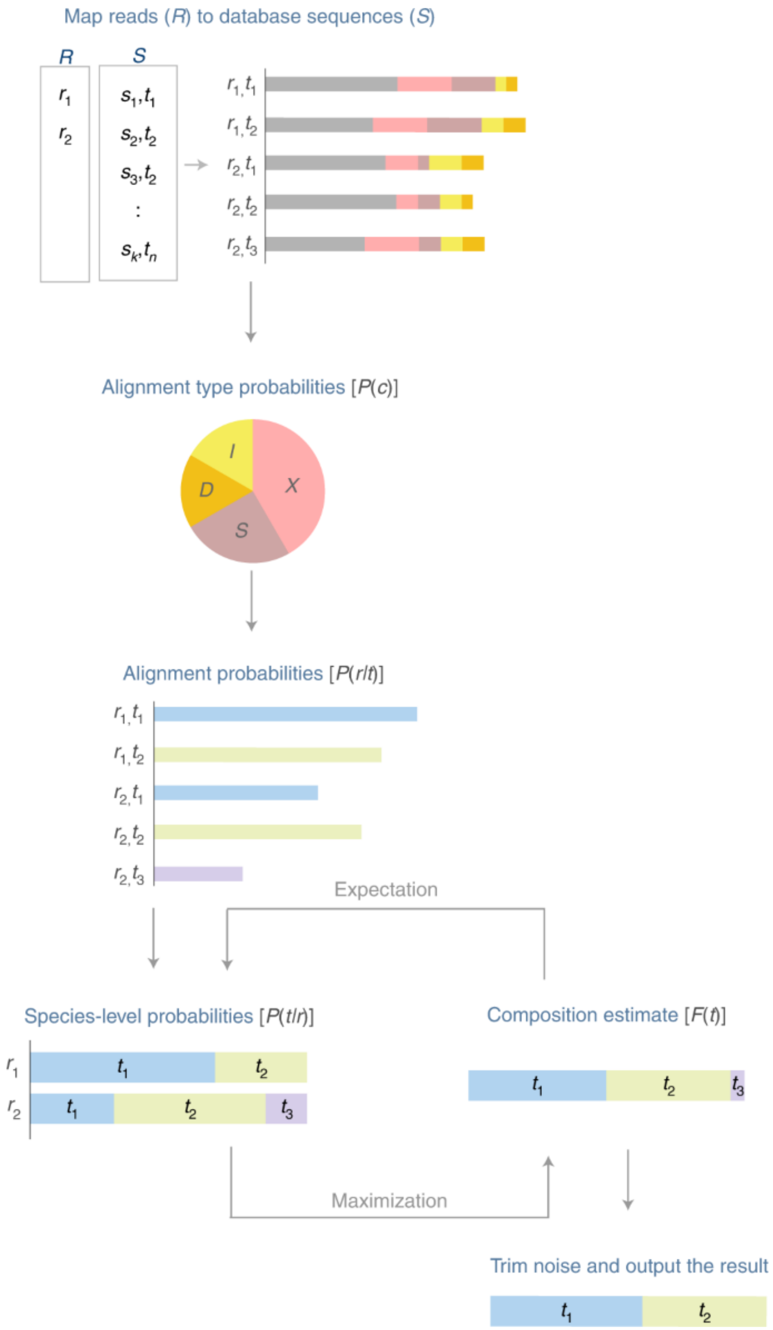
		Coin A	Coin B
0.45 x A	0.55 x B	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
0.80 x A	0.20 x B	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
0.73 x A	0.27 x B	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
0.35 x A	0.65x B	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
0.65 x A	0.35x B	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
		≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$

$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$

$\hat{\theta}_A^{(10)} \approx 0.80$

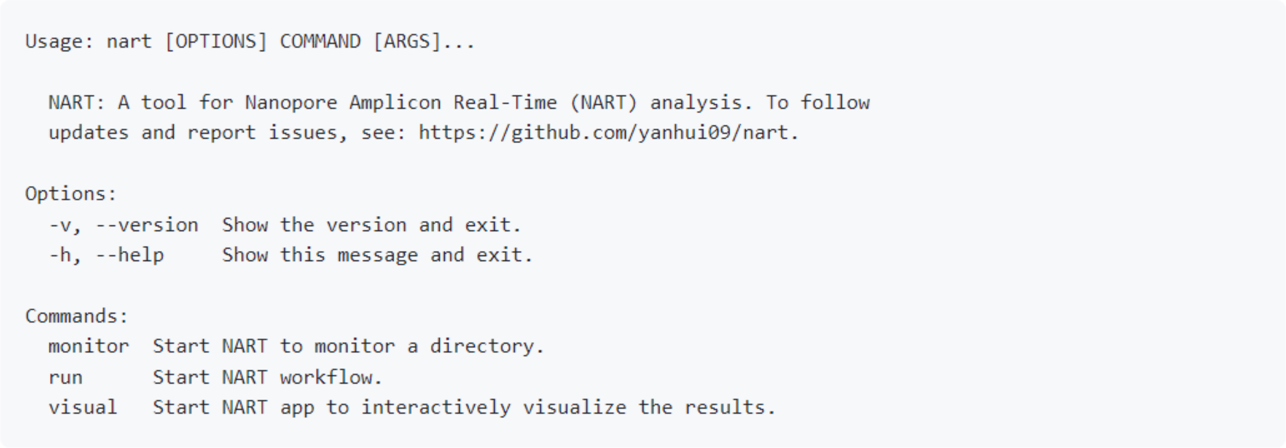
$\hat{\theta}_B^{(10)} \approx 0.52$


Map reads (R) to database sequences (S)
R
S
r1
r2
s1,t1
s2,t2
s3,t2
:
sk,tn
r1,t1
r1,t2
r2,t1
r2,t2
r2,t3
Alignment type probabilities [P(c)]
I
D
X
S
Alignment probabilities [P(r|t)]
r1,t1
r1,t2
r2,t1
r2,t2
r2,t3
Expectation
Species-level probabilities [P(t|r)]
r1
r2
t1
t2
t1
t2
t3
Composition estimate [F(t)]
t1
t2
t3
Maximization
Trim noise and output the result
t1
t2

https://www.nature.com/articles/nbt1406
https://www.nature.com/articles/s41592-022-01520-4

nart &nawf

`NART` is composed of two sets of scripts: `nart` and `nawf`, which controls real-time analysis and workflow performance, respectively.

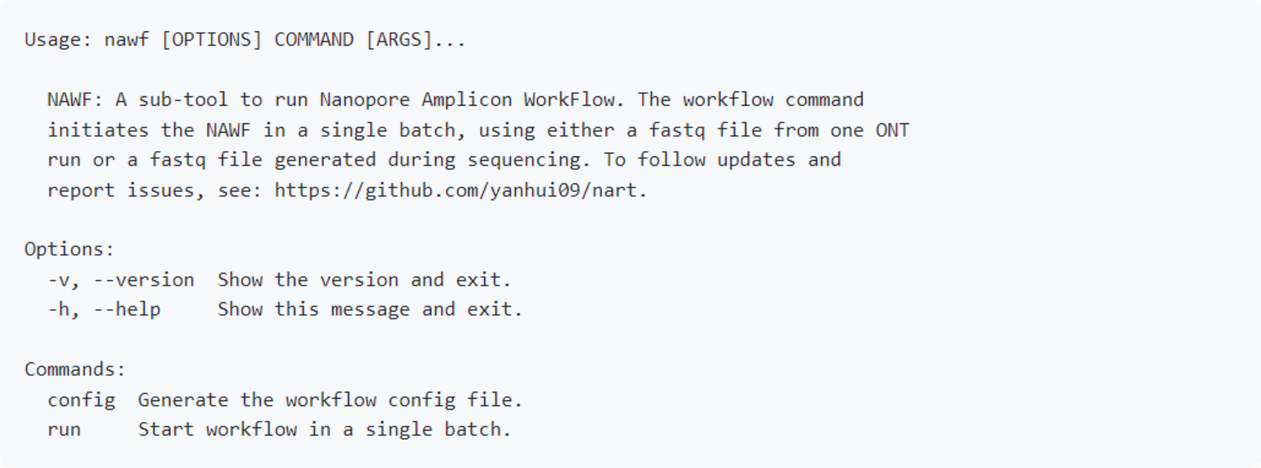


```
Usage: nart [OPTIONS] COMMAND [ARGS]...

  NART: A tool for Nanopore Amplicon Real-Time (NART) analysis. To follow
  updates and report issues, see: https://github.com/yanhui09/nart.

Options:
  -v, --version  Show the version and exit.
  -h, --help     Show this message and exit.

Commands:
  monitor  Start NART to monitor a directory.
  run      Start NART workflow.
  visual   Start NART app to interactively visualize the results.
```



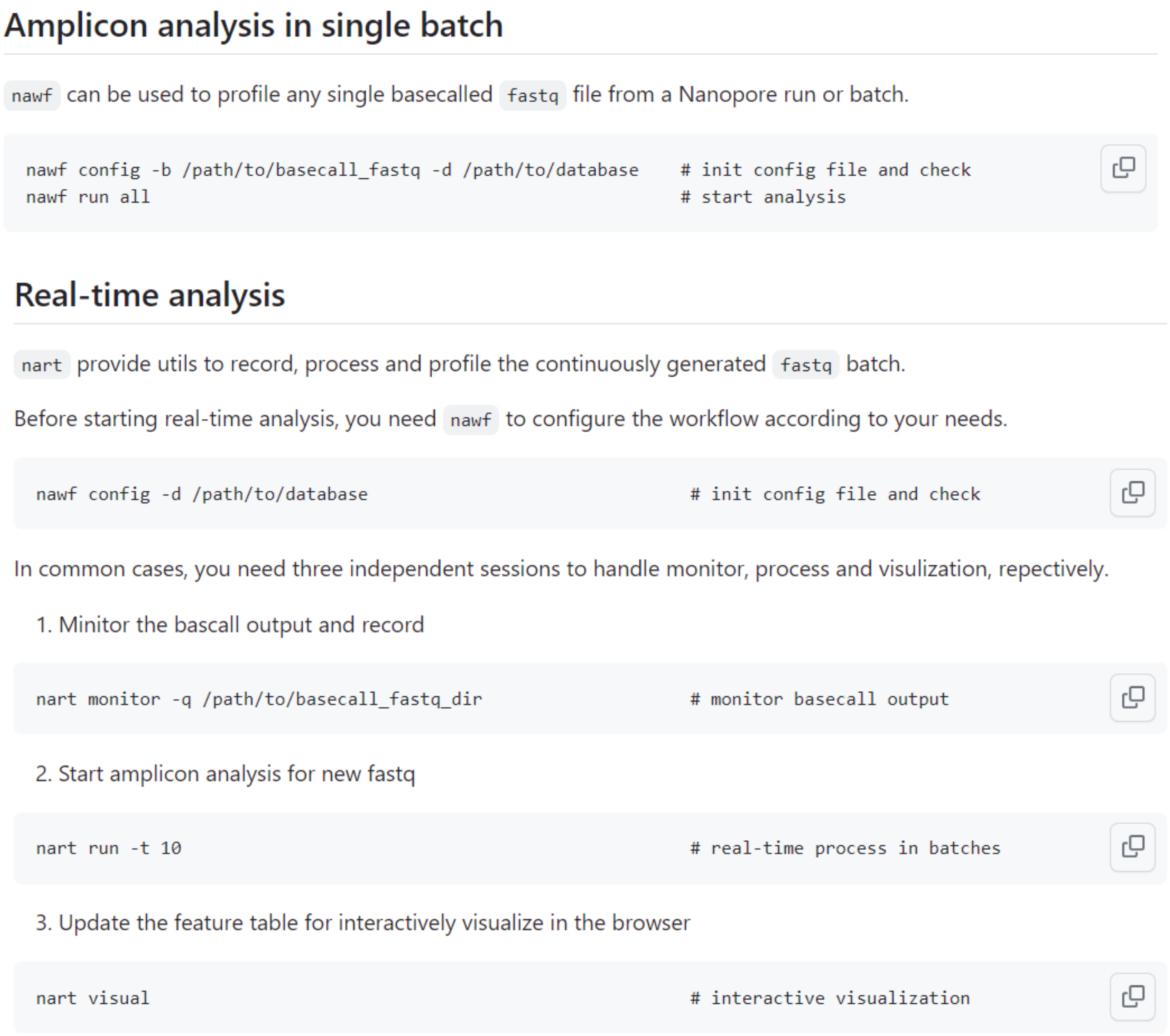
```
Usage: nawf [OPTIONS] COMMAND [ARGS]...

  NAWF: A sub-tool to run Nanopore Amplicon WorkFlow. The workflow command
  initiates the NAWF in a single batch, using either a fastq file from one ONT
  run or a fastq file generated during sequencing. To follow updates and
  report issues, see: https://github.com/yanhui09/nart.

Options:
  -v, --version  Show the version and exit.
  -h, --help     Show this message and exit.

Commands:
  config  Generate the workflow config file.
  run     Start workflow in a single batch.
```

Usage



Amplicon analysis in single batch

`nawf` can be used to profile any single basecalled `fastq` file from a Nanopore run or batch.

```
nawf config -b /path/to/basecall_fastq -d /path/to/database    # init config file and check
nawf run all                                                    # start analysis
```

Real-time analysis

`nart` provide utils to record, process and profile the continuously generated `fastq` batch.

Before starting real-time analysis, you need `nawf` to configure the workflow according to your needs.

```
nawf config -d /path/to/database                                # init config file and check
```

In common cases, you need three independent sessions to handle monitor, process and visulization, repectively.

1. Minitor the bascall output and record

```
nart monitor -q /path/to/basecall_fastq_dir                     # monitor basecall output
```

2. Start amplicon analysis for new fastq

```
nart run -t 10                                                  # real-time process in batches
```

3. Update the feature table for interactively visualize in the browser

```
nart visual                                                     # interactive visualization
```

RT-philosophy

ONT sequencing and basecalling in batches

- nart monitor => fqs.txt (record fastq files)
- nart run => nawf (start the workflow in batches & update the feature table)
- nart visual => interactively visualize profiles.

Exercises

Exercise

- MAC2023: https://yanhui09.github.io/MAC2023/

docker

Cross-platform support, incl. MacOS

CONDA® Mamba

Linux/amd64 platform

Thanks