From fastq to trim primers

Oliver Stringham

Contents

- Install softwares
- Concat fastqs into one fastq
- Quality filter
- Length Filter
- Primer match and trim

Let’s install them all

Software Used

Concat fastqs into one fastq	**command line**
Quality filter	**NanoFilt**
Length Filter	**seqkit**
Primer match and trim	**cutadapt**
Trim extra ONT adapters	**porechop**
Cluster sequences into MOTUs	**VSEARCH**
BLAST	**BLAST**

Other helper softwares: **bbmap, R, command line**

Set up

- Download Fastq folder from google drive
- Make folder in annotate home directory: data/florida_cf
- transer fastq folder to data/florida_cf
- transfer other files too

Concat fastqs into one fastq

```
cat data/florida_cf/fastq/**/*.fastq > data/florida_cf/barcodes-04-10.fastq
```

Quality Filter

Note: ONT uses Sanger fastq file notation

https://www.cs.cmu.edu/~pattis/15-1XX/common/handouts/ascii.html

Q = ascii value - 33

First, let's actually look at a fastq file

```
head data/florida_cf/barcodes-04-10.fastq
tail data/florida_cf/barcodes-04-10.fastq
```

Quality filter (rm reads with average q score < 10)

```
singularity exec images/nanofilt.sif \
NanoFilt --q 10 data/florida_cf/barcodes-04-10.fastq > \
data/florida_cf/barcodes-04-10.q10.fastq
```

View stats before and after

Now starts primer specific steps (needs to be repeated for each primer used)

Length Filter

- Let’s focus on MiFish (this run used Folmer, MiFish, and V5)
- Miya MiFish amplicon length = 163–185 bp
- Raw Length out of sequencer = ?
  - primers x2 + barcodes x2 + adapters x? + amplicon = ?

Can just look at empirical length distribution

```
singularity exec images/bbmap.sif \
readlength.sh \
in=data/florida_cf/barcodes-04-10.q10.fastq \
out=data/florida_cf/barcodes-04-10.q10.readlength.txt
```

(bonus to plot it)

Primer match and trim

>MiFish_Forward

GTCGGTAAAACTCGTGCCAGC

>MiFish_Reverse

CATAGTGGGGTATCTAATCCCAGTTTG

How many times to run cutadapt per primer set?

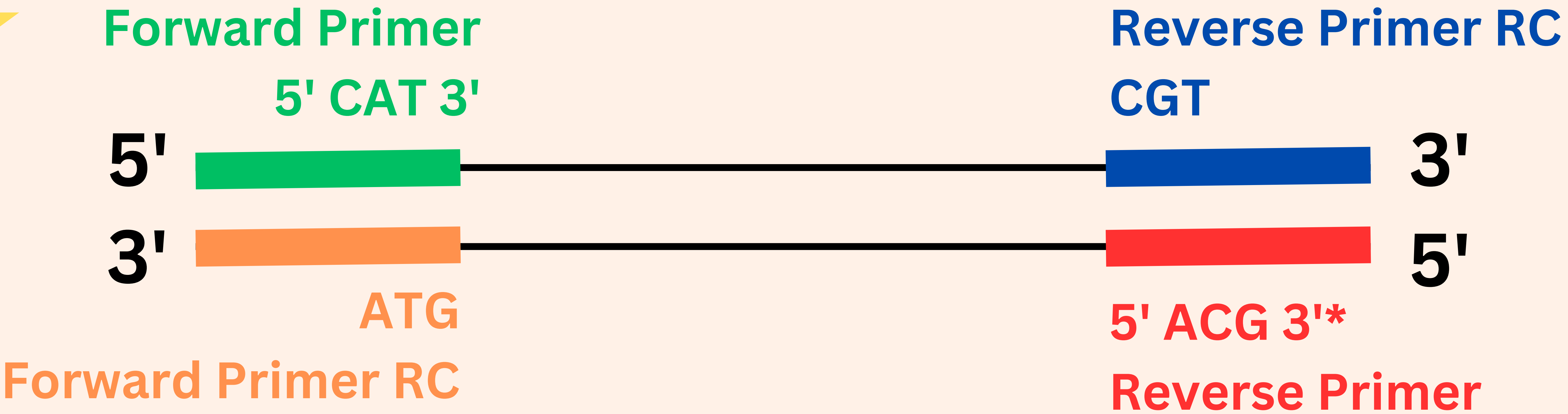
Need to consider

- Linked + unlinked primers
- Reverse complement (bc of pcr dna replication?)

Primer match and trim

con’t

Example:



Unlinked: CAT...	**Unlinked:** ...CGT	**Linked:** CAT...CGT
Unlinked: ACG...	**Unlinked:** ...ATG	**Linked:** ACG...ATG

Primer match and trim

con’t

Cutadapt parameters

https://cutadapt.readthedocs.io/en/stable/guide.html

-g	5' adapter
-a	3' adapter
-e	error tolerance (proportion different bwtn primer and read)
--no-indels	no insertions or deletions when matching primers
--discard-untrimmed	output will not contain reads that didn’t match the primer
--cores	number of cpu cores to use (don’t really need bc cutadapt is very fast)

Primer match and trim

con’t

>MiFish_Forward
GTCGGTAAAACTCGTGCCAGC
>MiFish_Reverse
CATAGTGGGGTATCTAATCCCAGTTTG

https://en.vectorbuilder.com/tool/dna-reverse-complement.html

Linked primer # 1 - cutadapt command

```
singularity exec images/cutadapt.sif \
cutadapt -g GTCGGTAAAACTCGTGCCAGC...CAAACTGGGATTAGATACCCCACTATG \
--cores 4 -e 0.2 --no-indels --discard-untrimmed \
data/florida_cf/barcodes-04-10.q10.fastq > \
data/florida_cf/barcodes-04-10.q10..mifish_linked_1.fastq
```

Next Steps

- Look at read length distribution
- Repeat for other linked primer
- Repeat for unlinked primers x4

- combine linked x2
- combine unlinked x4

Enough for today