

Practical - Quality control for short reads

In this section, we will be using some example data to assess the quality of short reads. We will use the tool, FASTQC. We will also use some tools to trim poor quality reads (or parts of reads).

You can also run [Kraken2 to detect contamination](#)

Where is the example data?

- https://zenodo.org/record/3977236/files/female_oral2.fastq-4143.gz?download=1
- https://zenodo.org/records/10018484/files/pKP1-NDM-1_R1.fastq.gz?download=1
- https://zenodo.org/records/10018484/files/pKP1-NDM-1_R2.fastq.gz?download=1

female_oral2.fastq.gz: This is a microbiome sample (16S) from a snake [Jacques et al. 2021](#).



Note

Remember, the pKP1-NDM-1 reads are simulated reads, with minimal error. These are effectively "perfect" and will not be representative of real data. We can use this to compare with problematic data (female_oral2.fastq.gz)

Required software

If you want to run this on the command-line, you may need to install some software.

- [FASTQC](#)
- [Cutadapt](#)

This is how to do it via conda:

```
conda install fastqc fastqc cutadapt -y
```


Downloading the reads via the command line

```
wget -O female_oral2.fastq.gz  
https://zenodo.org/record/3977236/files/female_oral2.fastq-4143.gz?download=1  
wget -O pKP1-NDM-1_R1.fastq.gz https://zenodo.org/records/10018484/files/pKP1-  
NDM-1_R1.fastq.gz?download=1  
wget -O pKP1-NDM-1_R2.fastq.gz https://zenodo.org/records/10018484/files/pKP1-  
NDM-1_R2.fastq.gz?download=1
```

Assess quality with FASTQC




One way we can check sequence quality is with [FastQC](#). It provides a modular set of analyses which you can use to check whether your data has any problems of which you should be aware before doing any further analysis. We can use it, for example, to assess whether there are known adapters present in the data. We'll run it on the FASTQ files.

FASTQC in Galaxy

 **FastQC** Read Quality reports (Galaxy Version 0.74+galaxy0)

Tool Parameters

Raw read data from your current history *






125: female_oral2.fastq-4143.gz

accepted formats ▼

Contaminant list - optional






Nothing selected

accepted formats ▼

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list - optional








Nothing selected

accepted formats ▼

List of adapters adapter sequences which will be explicitly searched against the library. It should be a tab-delimited file with 2 columns: r

Submodule and Limit specifying file - optional

Nothing selected

accepted formats ▼

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warnin

Disable grouping of bases for reads >50bp

☒ No

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You h

Lower limit on the length of the sequence to be shown in the report - optional

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your r

Length of Kmer to look for *

Note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

- Go to the Galaxy server's website. If you're using a public Galaxy server, you can usually access it through a web browser without needing to install anything locally.
- Before running FASTQC, you'll typically need to upload your data files to Galaxy. You can do this by clicking on the "Upload Data" button or using the "Get Data" menu to import data from various sources.

- Once your data is uploaded, find the FASTQC tool. Tools are organized into categories, and you can search for specific tools using the search bar.
- Click on the tool's name to open it. You'll see a form where you can configure the tool's inputs and parameters. Fill in the required fields and adjust any optional parameters as needed.
- After configuring the inputs, scroll down to the bottom of the form and click the "Execute" or "Run" button to start the tool.
- Galaxy will start running the tool, and you'll be redirected to the "History" panel where you can monitor the progress of your job. Depending on the tool and the size of your data, it may take some time to complete.
- Once the job is finished, you can view the results by clicking on the dataset in the history panel. You can download the results, visualize them, or use them as inputs for further analysis.

FASTQC on the command line

To run FastQC, open your terminal or command prompt and navigate to the directory where your data files are located. Then, use the `fastqc` command followed by the path to your data files. For example:

```
fastqc file1.fastq file2.fastq
```

You can also use wildcards to analyze multiple files at once, like this:

```
fastqc *.fastq
```

FastQC will process each file and generate an HTML report for each. *Are you able to open the report via the notebook file browser?* The reports contain various quality control metrics and visualizations. See the help via:

```
fastqc --help
```



Tip

FASTQC will also work for long reads.

Exercise 1: Run FASTQC

- Run FASTQC on `female_oral2.fastq.gz`.
- Run FASTQC on `pKP1-NDM-1_R1.fastq.gz` and `pKP1-NDM-1_R2.fastq.gz` together.

- Review and compare the HTML reports.

If you are unable to run FASTQC, here are some precalculated results; [female_oral2](#), [pKP1-NDM-1](#).

Which metrics are a major difference between the two reports?

What is the parts of the report are missing for pKP1-NDM? Can you explain why?

Review each metric for [female_oral2.fastq.gz](#), what part of each plot suggests there is a problem?



Tip

Remember, the pKP1-NDM-1 reads are simulated reads, with minimal error. These are effectively "perfect" and will not be representative of real data. We can use this to compare with problematic data ([female_oral2.fastq.gz](#))

[female_oral2.fastq.gz](#) data looks terrible, we should probably resequence it, but if we had to; how could we improve the quality?

[Answers to exercise 1](#)

Trim and filter - short reads


The quality drops in the middle of these sequences. This could cause bias in downstream analyses with these potentially incorrectly called nucleotides. Sequences must be treated to reduce bias in downstream analysis. Trimming can help to increase the number of reads the aligner or assembler are able to successfully use, reducing the number of reads that are unmapped or unassembled. In general, quality treatments include:

- Trimming/cutting/masking sequences
 - from low quality score regions
 - beginning/end of sequence
 - removing adapters
- Filtering of sequences
 - with low mean quality score
 - too short
 - with too many ambiguous (N) bases

To accomplish this task we will use [Cutadapt](#), a tool that enhances sequence quality by automating adapter trimming as well as quality control. We will:

- Trim low-quality bases from the ends. Quality trimming is done before any adapter trimming. We will set the quality threshold as 20, a commonly used threshold.
- Trim adapter with Cutadapt. For that we need to supply the sequence of the adapter. In this sample, Nextera is the adapter that was detected. We can find the sequence of the Nextera adapter on the Illumina website here [CTGTCTCTTATACACATCT](#) . We will trim that sequence from the 3' end of the reads.
- Filter out sequences with length < 20 after trimming

You can do this on galaxy:





 **Cutadapt** Remove adapter sequences from FASTQ/FASTA (Galaxy Version 4.8+galaxy0)

Tool Parameters

Single-end or Paired-end reads?

Single-end

FASTQ/A file *

125: female_oral2.fastq-4143.gz

accepted formats ▼

Should be of datatype "fastq.gz" or "fasta"

Read 1 Options

3' (End) Adapters

Sequence of an adapter ligated to the 3' end (paired data: of the first read). The adapter and subsequent bases see Help below.

1: 3' (End) Adapters

Source

Enter custom sequence

Custom 3' adapter name - optional

Optional if 'Multiple output' is selected in the Outputs selector

Custom 3' adapter sequence - optional

CTGTCTCTTATACACATCT|

(-a)

Filter Options

Discard Trimmed Reads

☐ No

Discard reads that contain the adapter instead of trimming them. Use the 'Minimum overlap length' option in order to avoid throwing away too many random

Discard Untrimmed Reads

☐ No

Discard reads that do not contain the adapter. (--discard_untrimmed)

Minimum length (R1) - optional

20

Discard trimmed reads that are shorter than LENGTH. Reads that are too short even before adapter removal are also discarded. (--minimum-length)

Maximum length (R1) - optional

Discard trimmed reads that are longer than LENGTH. Reads that are too long even before adapter removal are also discarded. (--maximum-length)

Max N - optional

Discard reads with more than this number of 'N' bases. A number between 0 and 1 is interpreted as a fraction of the read length. (--max-n)

Pair filter - optional

Read Modification Options

Cut bases from reads before adapter trimming - optional

0

Remove bases from each read (first read only if paired). If positive, remove bases from the beginning. If negative, remove bases from the end.

Quality cutoff *

20

Trim low-quality bases from 5' and/or 3' ends of each read before adapter removal. Applied to both reads for paired-end data. (--quality-cutoff)

NextSeq trimming *

0

Experimental option for quality trimming of NextSeq data. This is necessary because that machine cannot distinguish between 'A' and 'C' bases. (--nextseq-trim)

Trim Ns

☐ No

Trim N's on ends of reads. (--trim-n)

Trim poly-A tails

☐ No

Note, this trim poly-T 'heads' on R2 (--poly-a)

Strip suffix - optional

If you are unable to run this, here is the FASTQC output [pre trimming](#) and [post trimming](#) to compare.

Exercise 2: Trim and filter

Use cutadapt to trim the adapter sequence from the 3' end of the reads, and filter out sequences with a length less than 20 after trimming.

Run FASTQC on the trimmed data and compare to the original file.

Does the per base sequence quality look better?

Is the adapter gone?

What can you say about some of the other metrics?

If you are attempting this on the command-line, you can run cutadapt like:

```
cutadapt -q 20 -a CTGTCTCTTATACACATCT -m 20 female_oral2.fastq.gz | gzip -c > female_oral2.trimmed.fastq.gz
```

Can you explain what each of the options does?

[Answers to exercise 3](#)

Acknowledgements

Some of this material was adapted from:

- Bérénice Batut, Maria Doyle, Alexandre Cormier, Anthony Bretaudeau, Laura Leroi, Erwan Corre, Stéphanie Robin, Erasmus+ Programme, Cameron Hyde, Quality Control (Galaxy Training Materials). <https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html> Online; accessed Wed Oct 18 2023
- Hiltemann, Saskia, Rasche, Helena et al., 2023 Galaxy Training: A Powerful Framework for Teaching! PLOS Computational Biology 10.1371/journal.pcbi.1010752