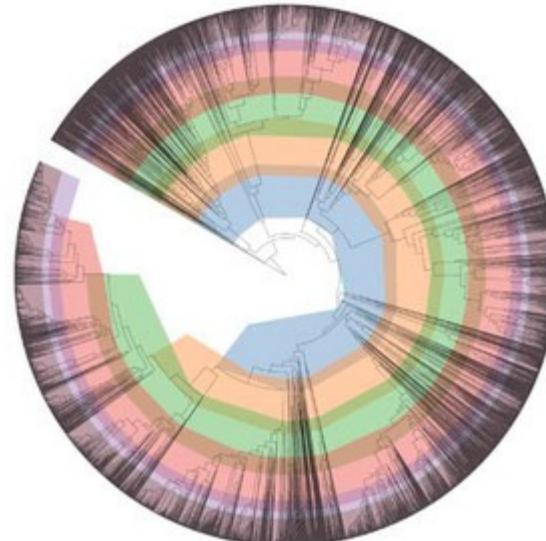


# ENVIRONMENTAL METAGENOMICS

Physalia course, online, 13-17 October 2025

## Data pre-processing: quality control and adapter removal

Nikolay Oskolkov, Group Leader of Metabolic Research Group at LIOS, Riga, Latvia  
Samuel Aroney, Postdoctoral Research Fellow, Queensland University of Technology



Physalia  
Courses

NB: original course material courtesy:  
Dr. Antti Karkman, University of Helsinki  
Dr. Igor Pessi, Finnish Environment Institute (SYKE)

# FASTQ File

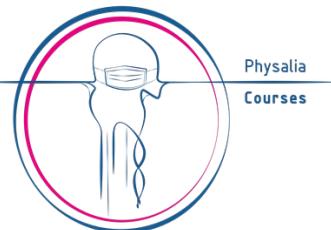
Example (files can be gigabytes in size!)

```
@K00233:37:HGHLYBBXX:3:1101:2646:1121 1:N:0:NACGCATC+NGCTGGTG
NCGCATGAGCCGCCTGTATCAGGCCTGATCGGCCGGCATTGCAGTTGGATAGATGGGGAGCACACGTCTG
+
#A7F<<GG<JFJFJJJJFFJJJJJAFFJFJJJJFJAFFFJAJFJJ<FJJJJFFF<FFA--FFFJJJJ
@K00233:37:HGHLYBBXX:3:1101:4655:1121 1:N:0:NACGCATC+NGCTGGTG
NATGCATGACAGGAGGTGAGGGCATTTCAGATTTCAGGCTGCGACCTTGAGCATTTGCCGTTCCAGCAC
+
#GG-<FFFF7JFF7JJJJFJJ<JJJJJA7FJJJJJJFF<JFF<J7-<FJJJJFJFFJJGGGGFFJJ--AJAJJ
```

**@ <read id, e.g. machine ID, location on flowcell> <extra metadata>**  
**<DNA sequence; Note: N = base couldn't be called!>**  
**+ <a separator>**  
**<base quality scores for each nucleotide in sequence>**

Quality score:

```
!"#$%&' ()*+,.-./0123456789:;<=>?@ABCDEFGHIJ
0.2.....26...31.....41
```



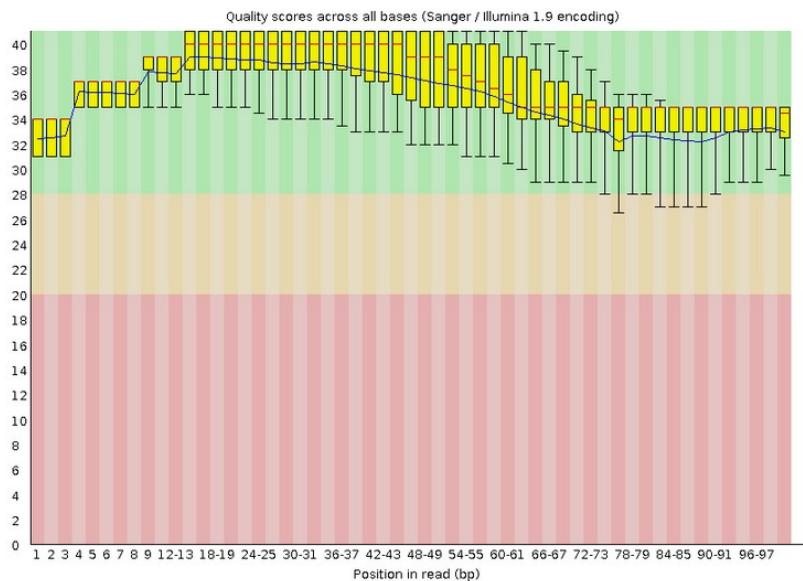
## Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Adapter Content

### Basic Statistics

Measure	Value
Filename	G69146_pe_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8997912
Sequences flagged as poor quality	0
Sequence length	77-101
%GC	49

### Per base sequence quality



## Quality Scores

A quality score (or Q-score) expresses an error probability. In particular, it serves as a convenient and compact way to communicate very small error probabilities.

Given an assertion, A, the quality score, Q(A), expresses the probability that A is not true, P(~A), according to the relationship:

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

where  $P(\sim A)$  is the estimated probability of an assertion A being wrong.

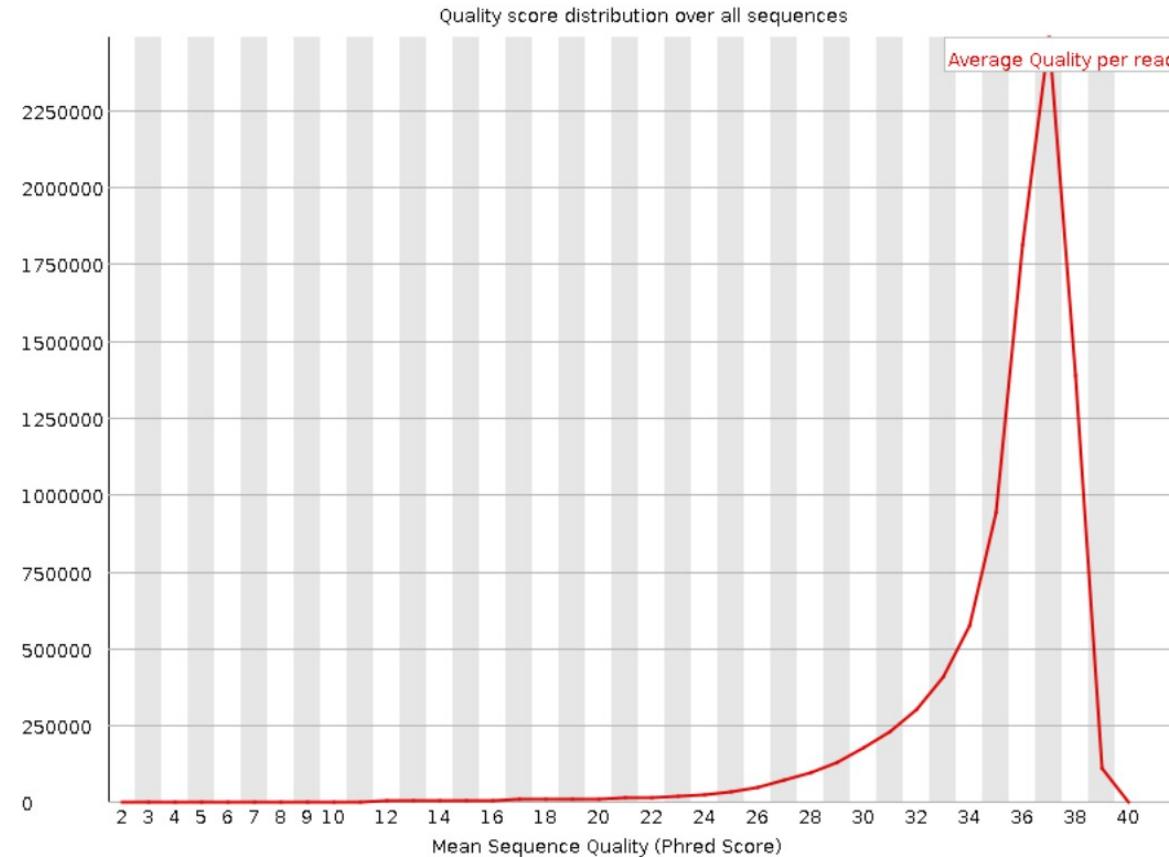
The relationship between the quality score and error probability is demonstrated with the following table:

Quality score, Q(A)	Error probability, P(~A)
10	0.1
20	0.01
30	0.001

## Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

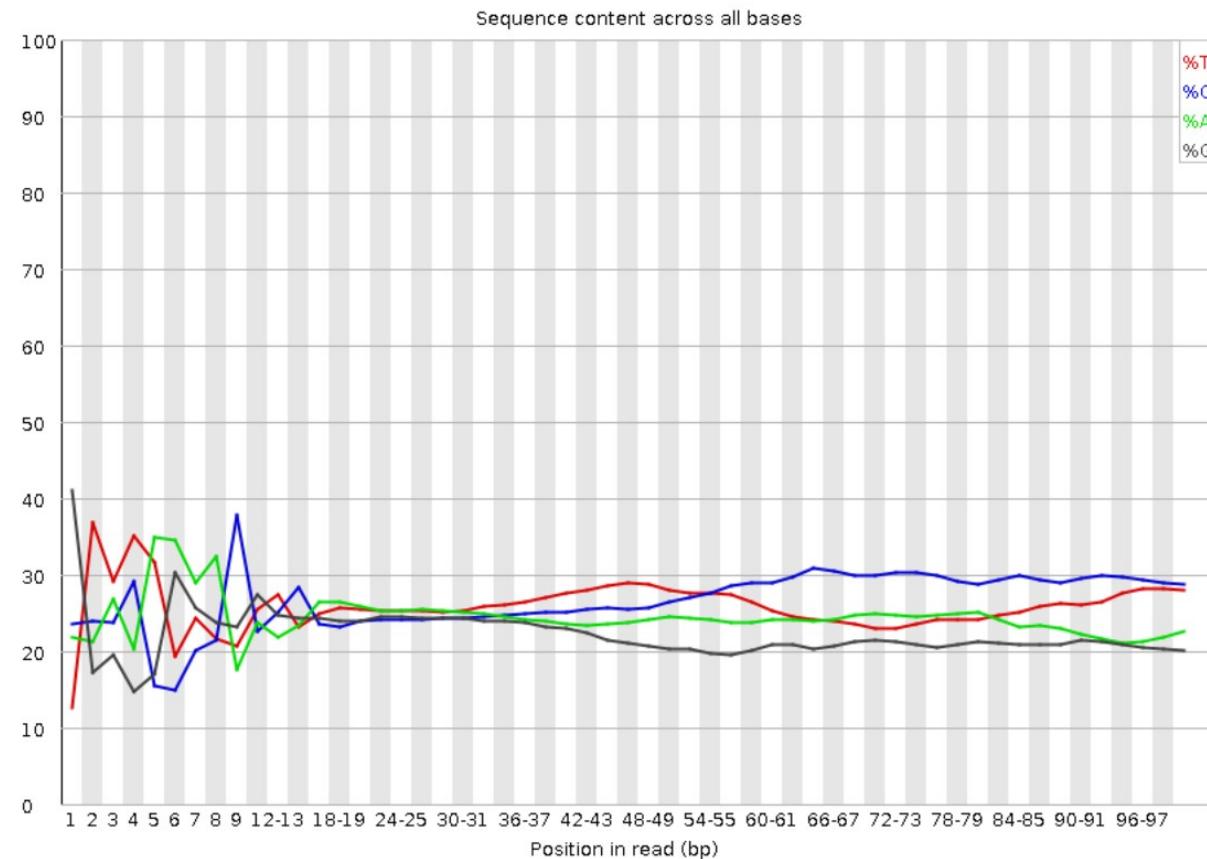
## Per sequence quality scores



## Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

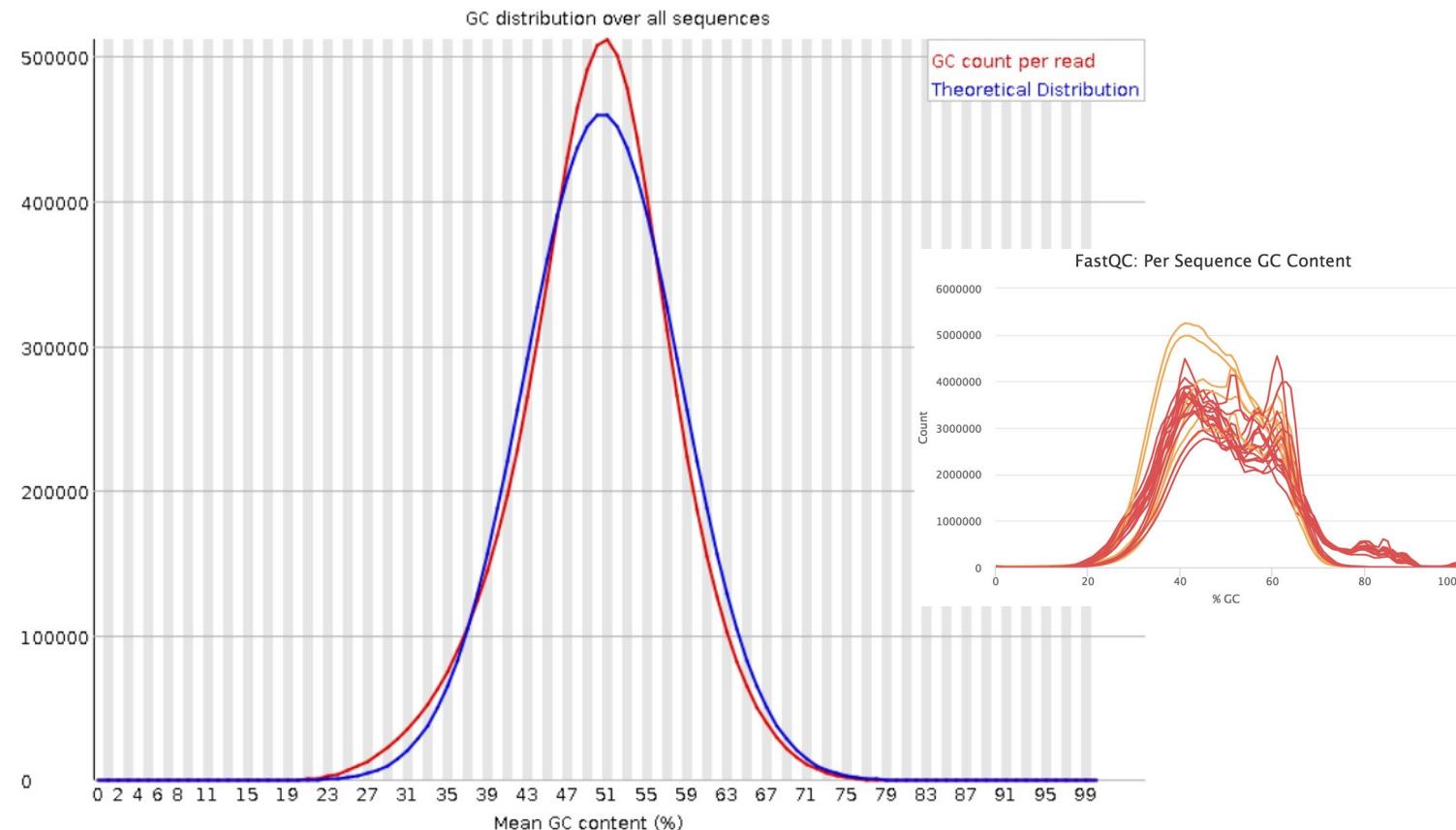
## ⚠ Per base sequence content



## Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)

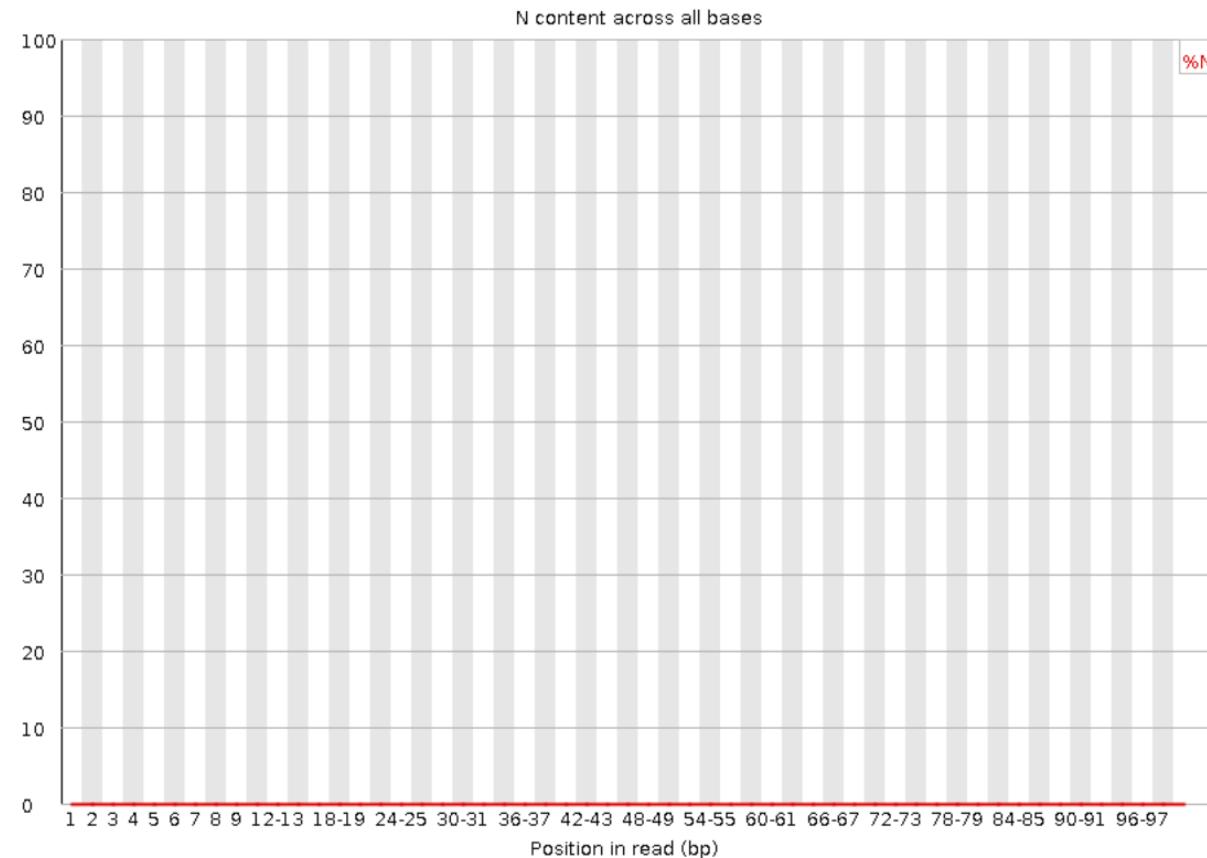
### ✓ Per sequence GC content



## Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)

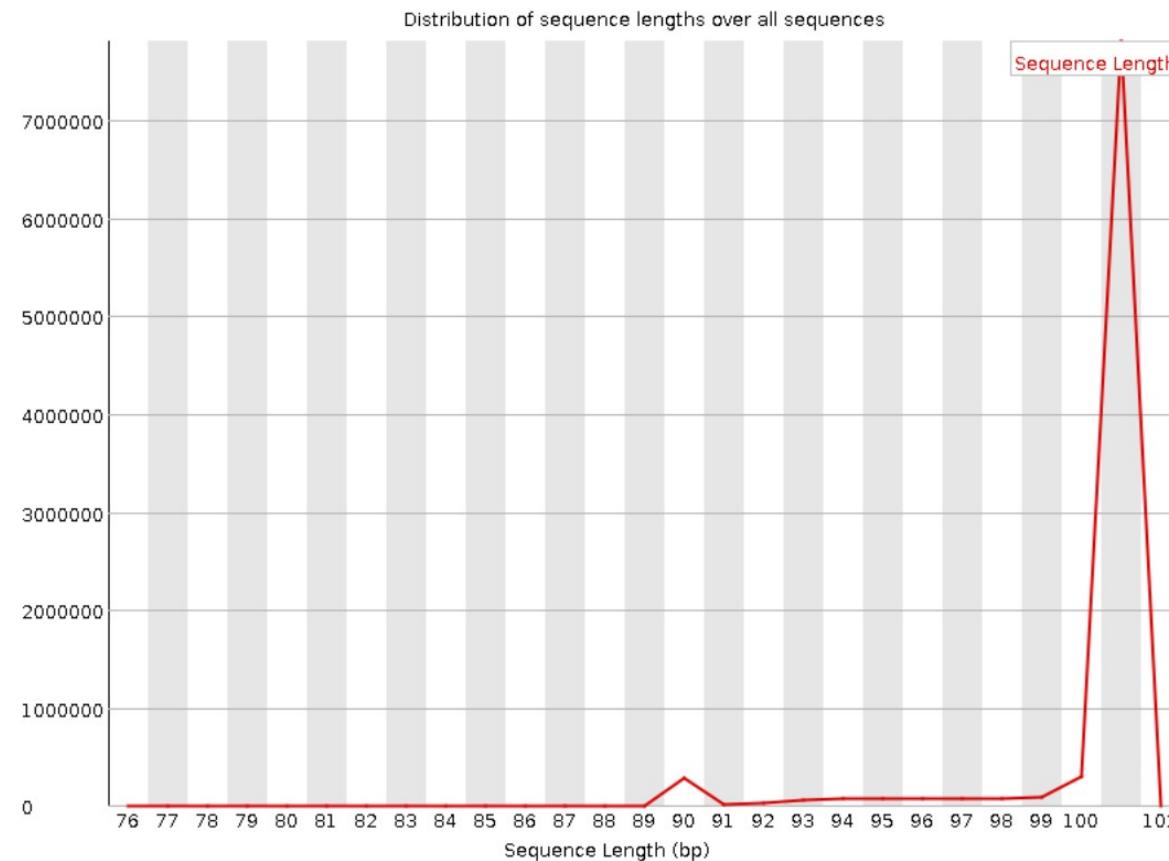
### ✓ Per base N content



## Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)

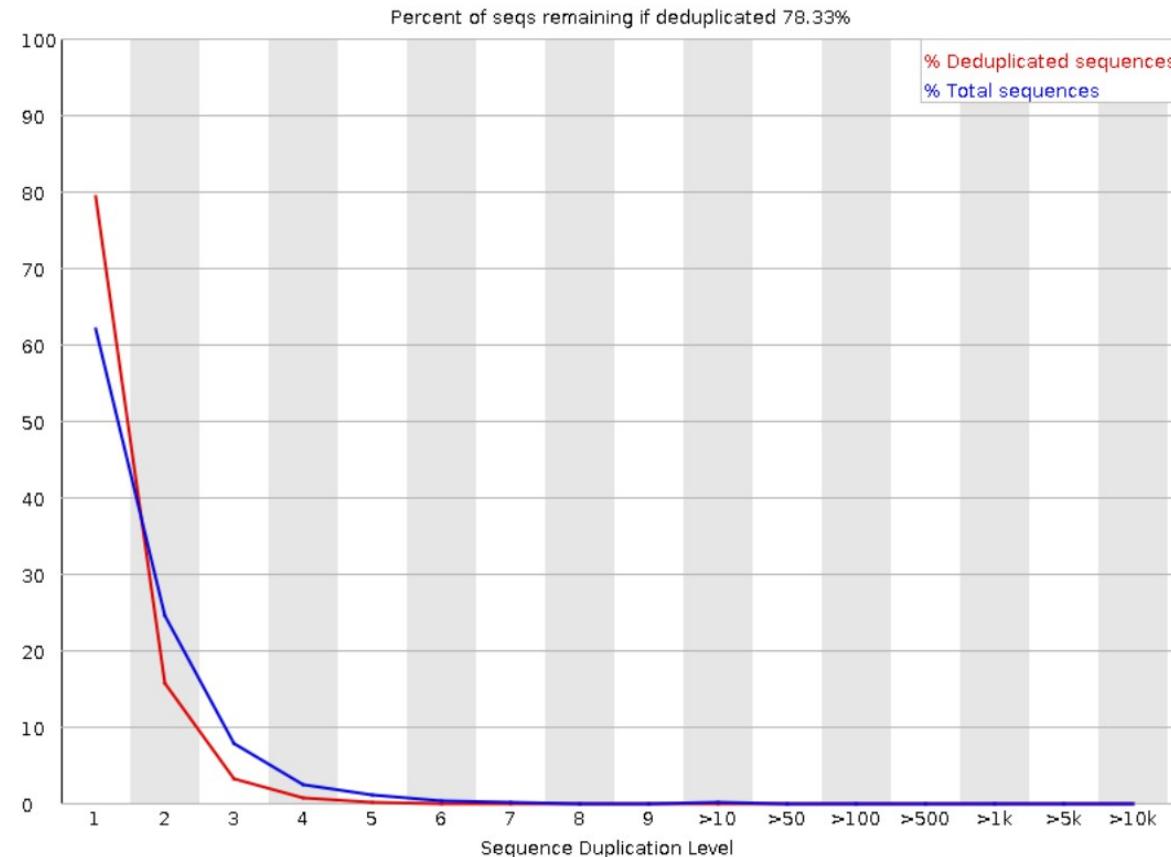
## ⚠ Sequence Length Distribution



## Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)

### ✓ Sequence Duplication Levels



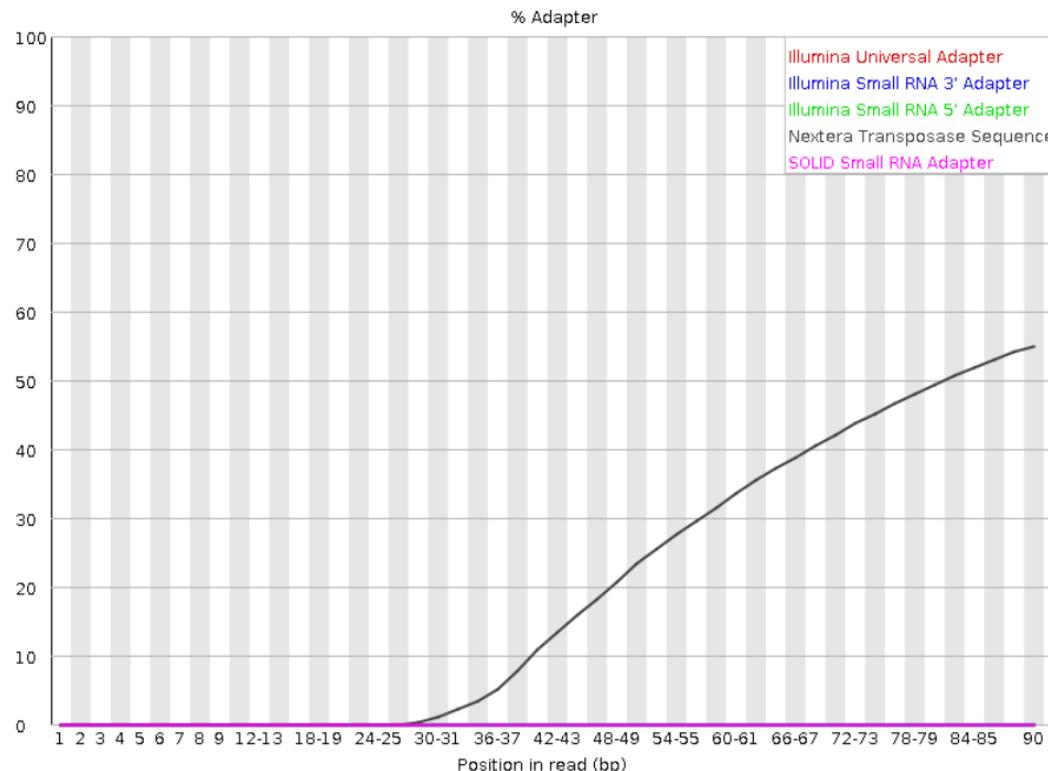
## Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)

## ✓ Overrepresented sequences

No overrepresented sequences

## ✗ Adapter Content



General Stats
FastQC
Sequence Quality Histograms
Per Sequence Quality Scores
Per Base Sequence Content
Per Sequence GC Content
Per Base N Content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2017-09-01, 09:26 based on data in: /pica/v11/b2017206\_nobackup/private/wabi/RNAseq/multiQC\_raw\_vs\_trimmed

ⓘ Welcome! Not sure where to start?

[Watch a tutorial video \(6:06\)](#)

don't show again

## General Statistics

Copy table

Configure Columns

Plot

Showing 18/18 rows and 4/5 columns.

Sample Name	% Dups	% GC	Length	M Seqs
P8205_101_S1_L001_R1_001	75.7%	46%	126 bp	81.2
P8205_101_S1_L001_R1_001.trimmomatic	75.7%	46%	126 bp	80.5
P8205_101_S1_L001_R1_001.trimmomatic_unpaired	38.7%	47%	106 bp	0.7
P8205_101_S1_L001_R2_001	71.9%	47%	126 bp	81.2
P8205_101_S1_L001_R2_001.trimmomatic	72.0%	46%	126 bp	80.5
P8205_101_S1_L001_R2_001.trimmomatic_unpaired	7.7%	47%	126 bp	0.0
P8205_102_S2_L001_R1_001	83.5%	46%	126 bp	86.2
P8205_102_S2_L001_R1_001.trimmomatic	83.5%	46%	126 bp	85.5
P8205_102_S2_L001_R1_001.trimmomatic_unpaired	50.3%	47%	108 bp	0.7
P8205_102_S2_L001_R2_001	80.0%	47%	126 bp	86.2
P8205_102_S2_L001_R2_001.trimmomatic	80.1%	47%	126 bp	85.5
P8205_102_S2_L001_R2_001.trimmomatic_unpaired	4.0%	41%	126 bp	0.0

## FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

### Sequence Quality Histograms

10 2 6

The mean quality value across each base position in the read. See the [FastQC help](#).

Y-Limits:  on

### Mean Quality Scores

Export Plot

# FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

General Stats
FastQC
Sequence Quality Histograms
Per Sequence Quality Scores
Per Base Sequence Content
Per Sequence GC Content
Per Base N Content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content

## Sequence Quality Histograms

10 2 6

The mean quality value across each base position in the read. See the [FastQC help](#).

Y-Limits:  on



A

H

D

M

?

### Mean Quality Scores



Export Plot

Created with MultiQC

## Per Sequence Quality Scores

15 3

The number of reads with average quality scores. Shows if a subset of reads has poor quality. See the [FastQC help](#).

Y-Limits:  on



### Per Sequence Quality Scores



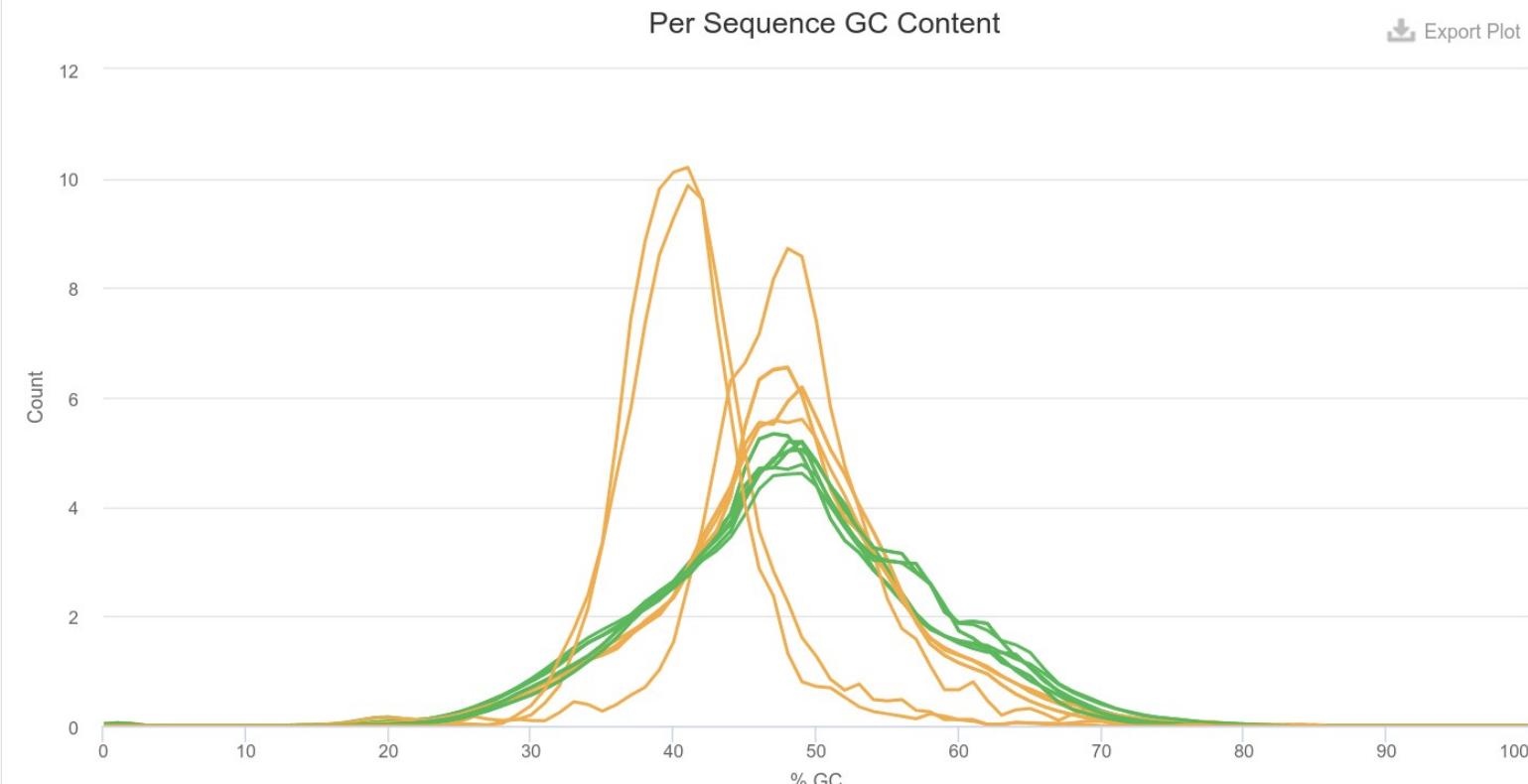
Export Plot

Created with MultiQC

## Per Sequence GC Content

10 8

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the [FastQC help](#).Y-Limits:  on

 Percentages  Counts

General Stats

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

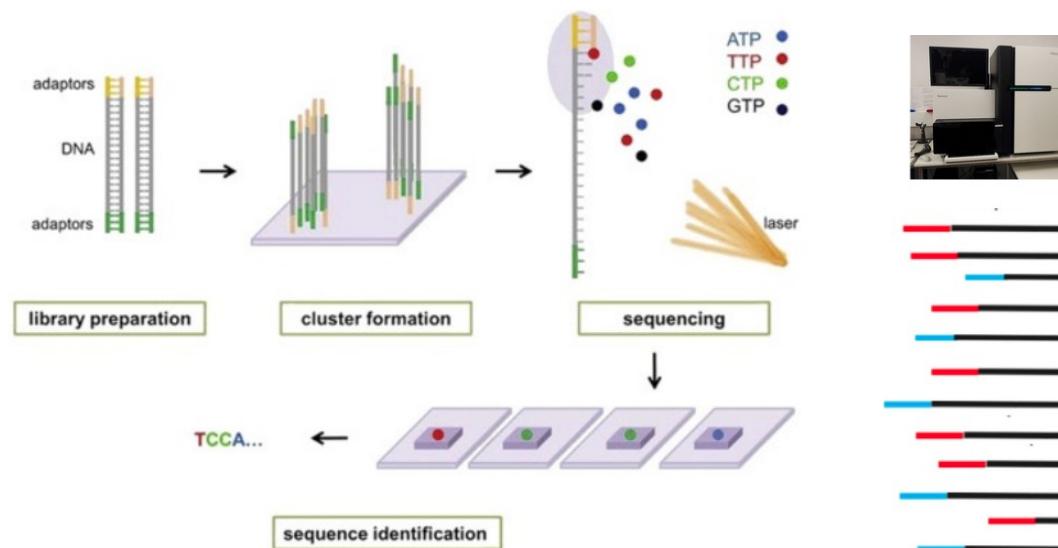
Overrepresented sequences

Adapter Content

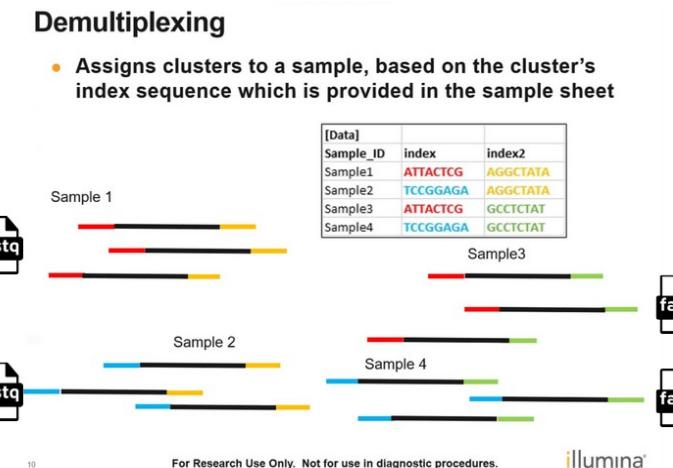
Toolbox



# Illumina DNA construct: indexes and adapters



Zhou et al. 2015. *Atlas Oral Microbiol.* <https://doi.org/10.1016/B978-0-12-802234-4.00002-1>



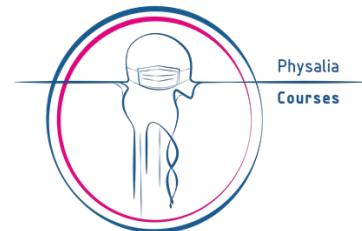
GGTGTACGGGACCAACaccgacGGCCCTACACGACGCTCTCCGATCTXXXXAGCACACGCTCTGGGCTCCAGTCACGacactaCCGTCTCTGCTTC

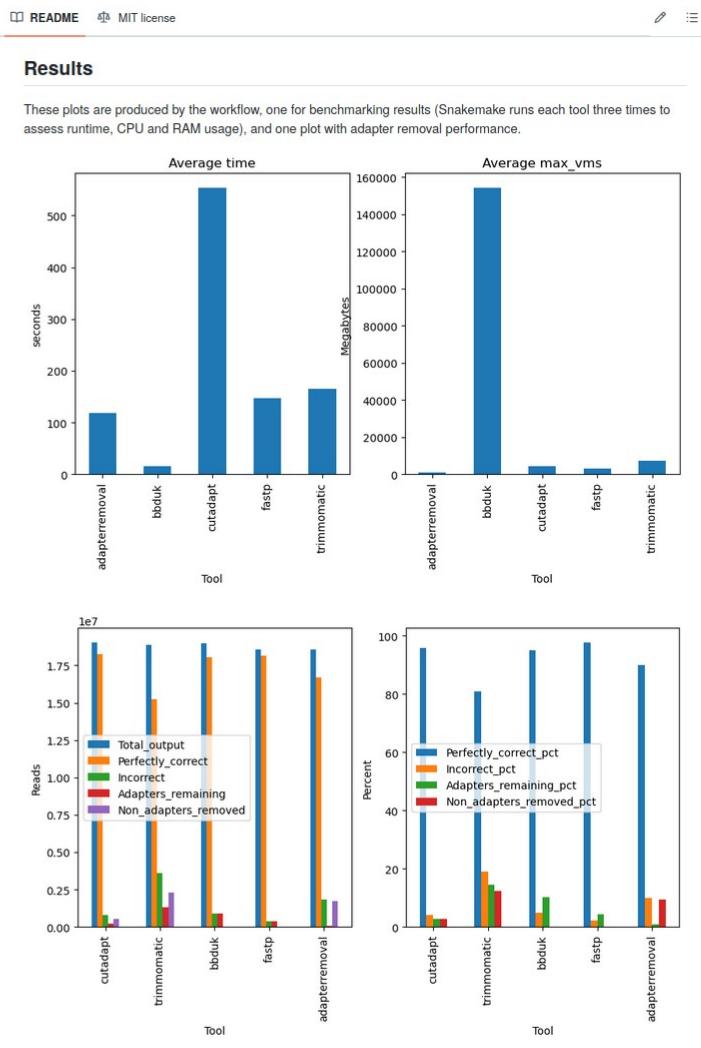
Digitized by srujanika@gmail.com

**T**TACTATGCCGCTGGTGGTG**tggctgttGGGATGTCGAGGGCTAGAXXXXXXXXCGTGTGCA**GACTTGAGGTAGT**G**ctgtga**tGGCAGGGGACGGGC**

[Adapter/Index Primer] [Index] [Target primer] [Target] [Target primer] [Index] [Adapter/Index Primer]

K00233:37:HGHLYBBXX:3:1101:2646:1121 1:N:0:NACGCATC+NGCTGGTG  
CGCATGAGCCGCTGTATCAGGCCTGATCGGGCCGGGATTGCAGTTGGATAGATCGGGGGAGCACACGTCTG  
A7F<<GG<JFJFJJJJFFJJJJJJJAFFJFJJJJJJFJAFFFJAJFJJ<FJJJJFFF<FFA--FFFJJJJ





Cutadapt

fastp

Trimmomatic

TrimGalore

AdapterRemoval

BBduk

## Benchmarking software tools for trimming adapters and merging next-generation sequencing data for ancient DNA

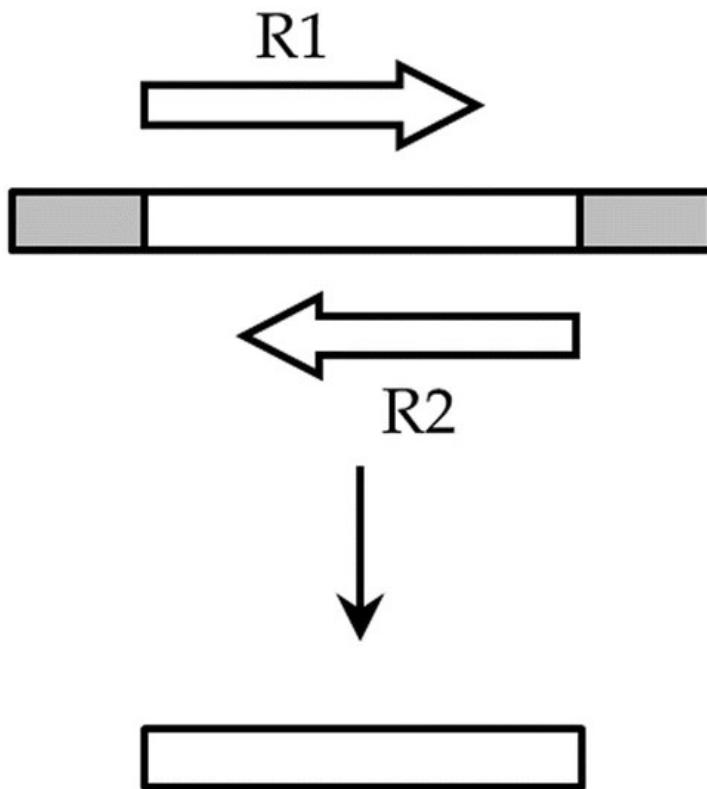
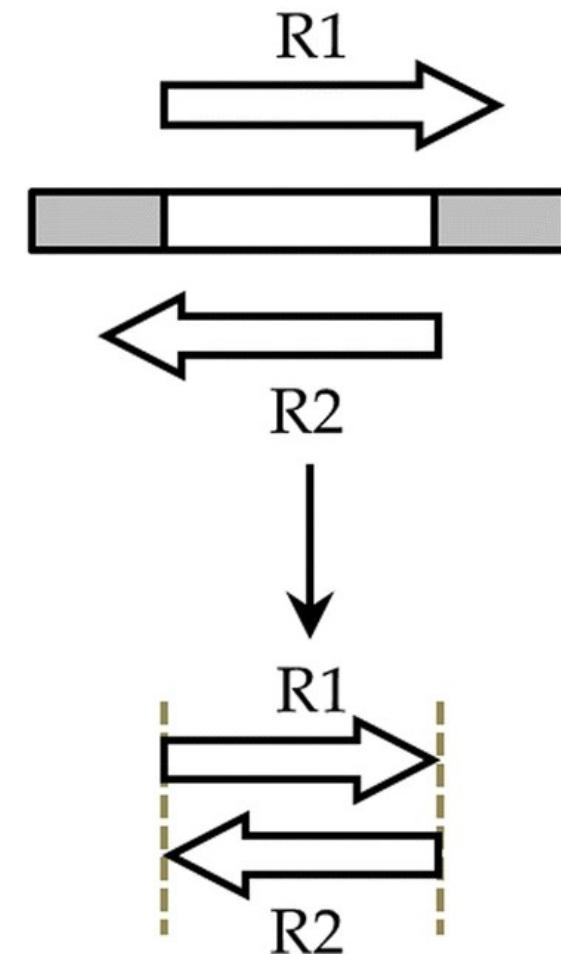
Annette Lien<sup>1</sup>, Leonardo Pestana Legori<sup>2</sup>, Louis Kraft<sup>1</sup>, Peter Wad Sackett<sup>1</sup> and Gabriel Renaud<sup>1\*</sup>

<sup>1</sup>Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark, <sup>2</sup>University of Debrecen, Debrecen, Hungary

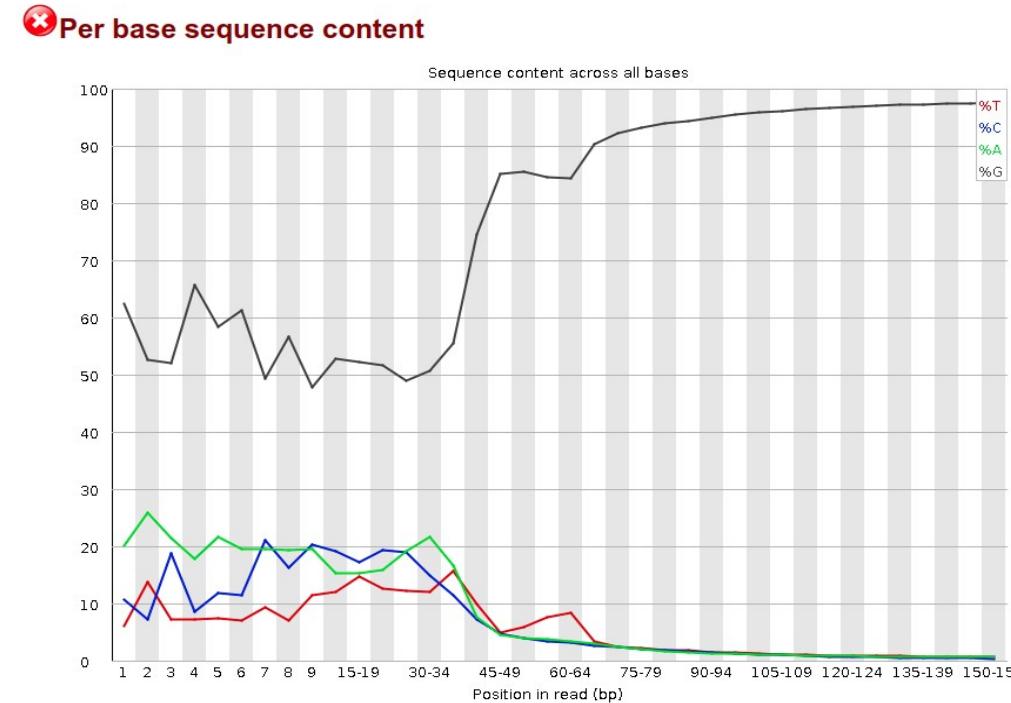
**CITATION**  
Lien A, Legori LP, Kraft L, Sackett PW and Renaud G (2023). Benchmarking software tools for trimming adapters and merging next-generation sequencing data for ancient DNA. *Front. Bioinform.* 3:1260486.  
doi: 10.3389/fbinf.2023.1260486

**COPYRIGHT**  
© 2023 Lien, Legori, Kraft, Sackett and Renaud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

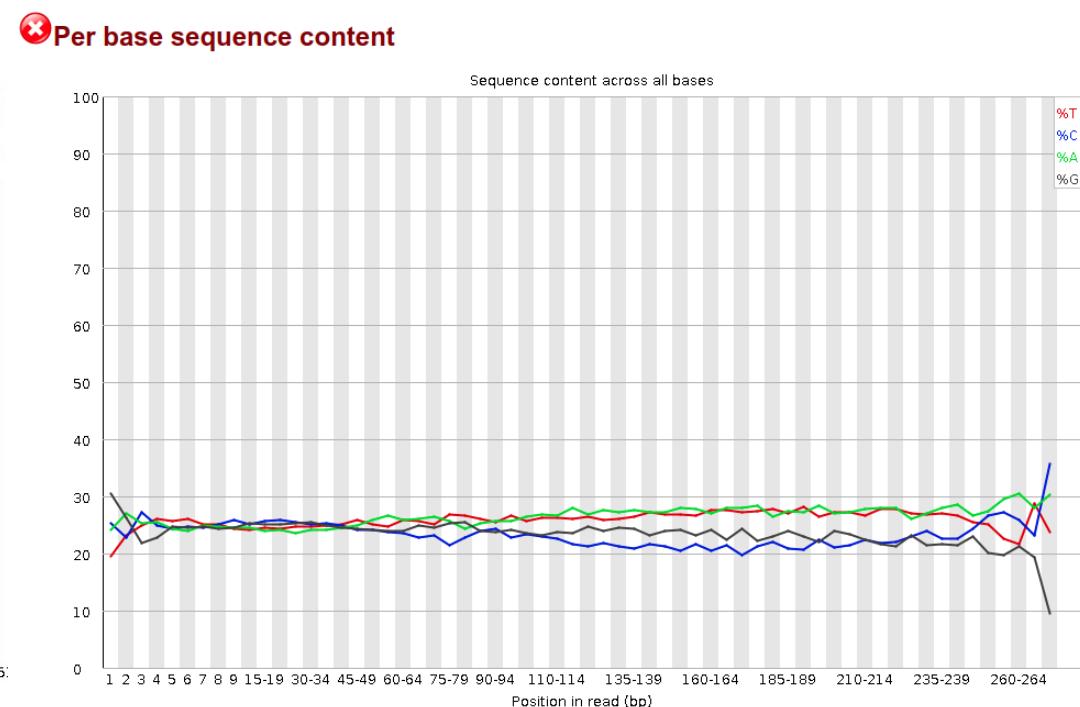
Ancient DNA is highly degraded, resulting in very short sequences. Reads generated with modern high-throughput sequencing machines are generally longer than ancient DNA molecules, therefore the reads often contain some portion of the sequencing adaptors. It is crucial to remove those adaptors, as they can interfere with downstream analysis. Furthermore, overlapping portions when DNA has been read forward and backward (paired-end) can be merged to correct sequencing errors and improve read quality. Several tools have been developed for adapter trimming and read merging, however, no one has attempted to evaluate their accuracy and evaluate their potential impact on downstream analyses. Through the simulation of sequencing data, seven commonly used tools were analyzed in their ability to reconstruct ancient DNA sequences through read merging. The analyzed tools exhibit notable differences in their abilities to correct sequence errors and identify the correct read overlap, but the most substantial difference is observed in their ability to calculate quality scores for merged bases. Selecting the most appropriate tool for a given project depends on several factors, although some tools such as fastp have some shortcomings, whereas others like BBduk outperform the other tools in most aspects. While the choice of tool did not result in a measurable difference when analyzing population genetics using principal component analysis, it is important to note that downstream analyses that are sensitive to wrongly merged reads or that rely on quality scores can be significantly impacted by the choice of tool.

**A****B**

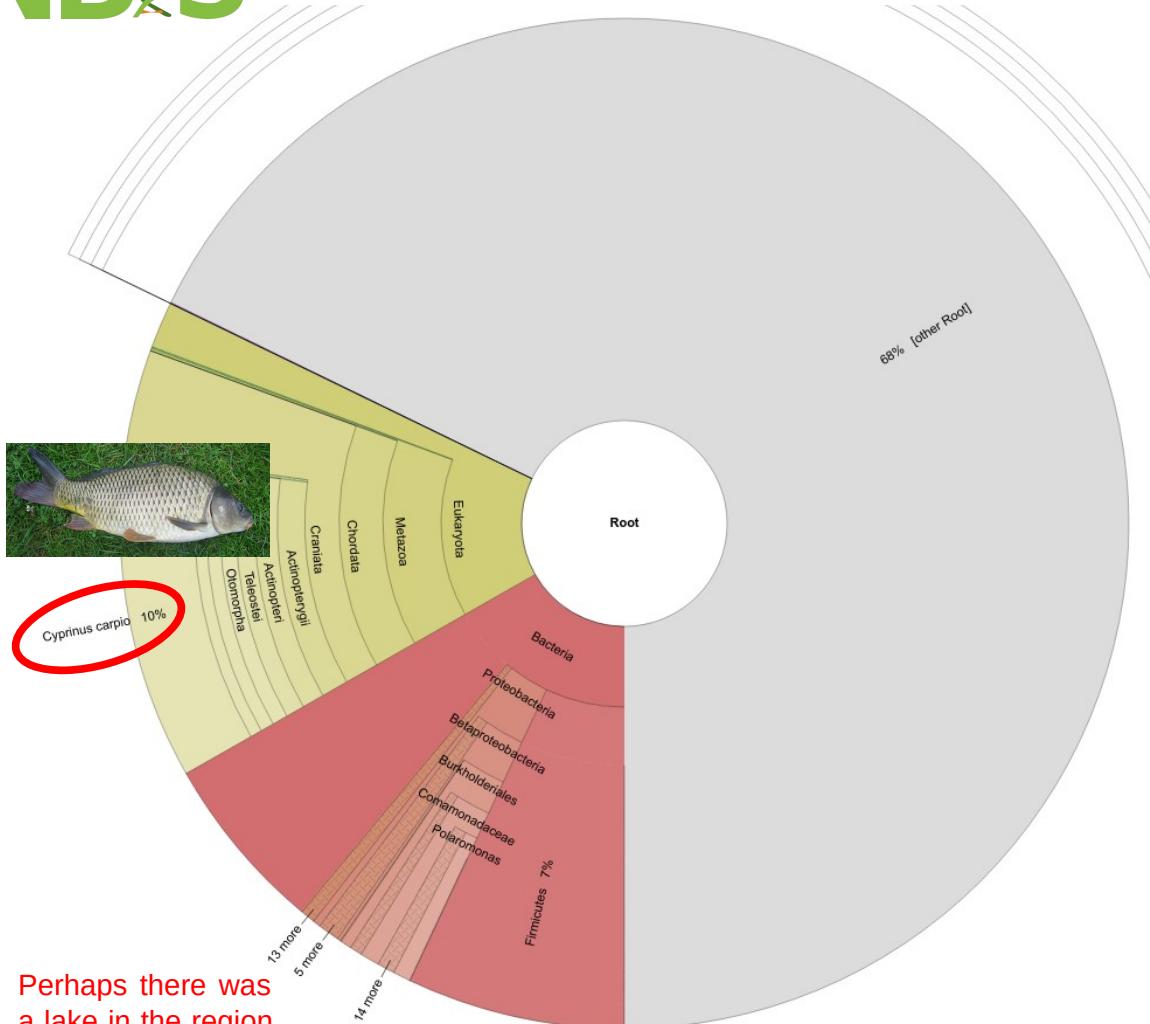
## Before fastp



## After fastp



Was NovaSeq or NextSeq used for sequencing?



It turns out [the Carp genome is full of Illumina adapters](#).

One of the first things we teach people in our [NGS courses](#) is how to remove adapters. It's not hard – we use [CutAdapt](#), but many other tools exist. It's simple, but really important – with De Bruijn graphs you will get paths through the graphs converging on kmers from adapters; and with OLC assemblers you will get spurious overlaps. With gap-filters, it's possible to fill the gaps with sequences ending in adapters, and this may be what happened in the Carp genome.

Why then are we finding such elementary mistakes in such important papers? Why aren't reviewers picking up on this? It's frustrating.

This is a separate, but related issue, to genomic contamination – [the Wheat genome has PhiX in it](#); [tons of bacterial genomes do too](#); and [lots of bacterial genes were problematically included in the Tardigrade genome](#) and declared as horizontal gene transfer.

## Bioinformatics Bits and Bobs

Rare and random blog posts about bioinformatics, genomics and evolution.

Monday, 29 September 2014

### Why you should QC your reads AND your assembly

The genome sequence of the Common Carp *Cyprinus carpio* was [published in Nature last week](#). By coincidence, I was doing some QC on some domesticated Ferret (*Mustela putorius furo*) reads, which had thrown some kmer warnings in the [FastQC tool](#). I blasted the kmers in NCBI and was quite perplexed by the number of hits that I found in the carp genome. Nearly all of the first 150 hits were all from the carp genome. Anyway, I looked a bit further into my odd kmers and it turns out that they were the ends of some Illumina adapter sequences that had presumably been incorporated into the paired-reads on the shorter ends of the insert size. This then took me back to the Carp Genome - what had crept into that?