

DTU





**DTU Health Technology
Bioinformatics**

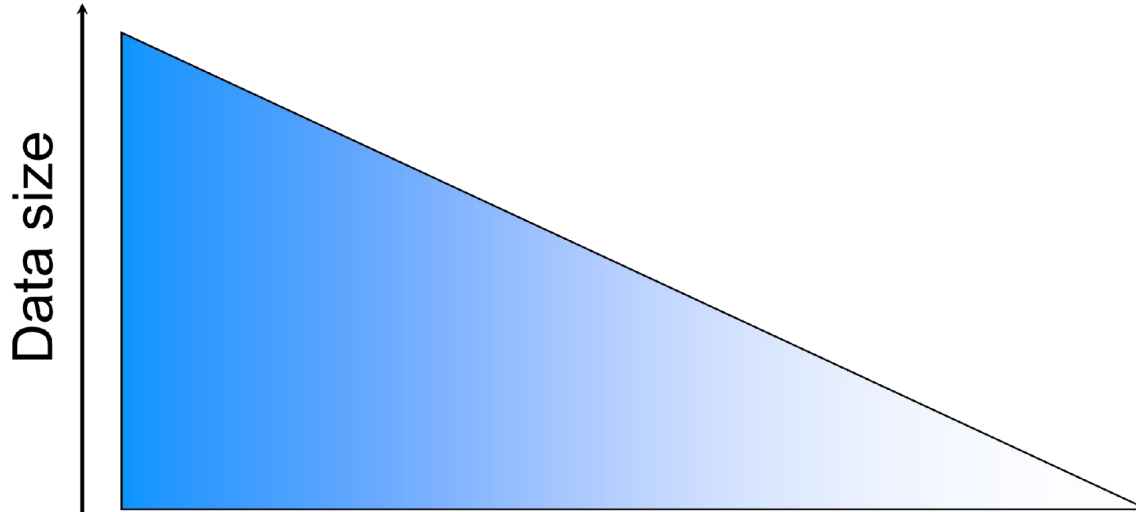
Data Preprocessing

*Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com*

Menu

- The main steps in NGS analysis
- Why is preprocessing important?
- Preprocessing
 - Fastqc reports for quality scores
 - Adapters
 - Depth of coverage vs Breadth of coverage

Generalized NGS analysis



Question

Raw reads

Pre-processing

Assembly:
Alignment /
de novo

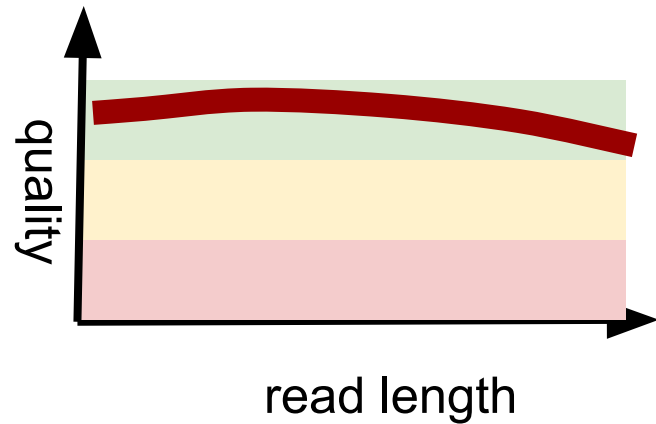
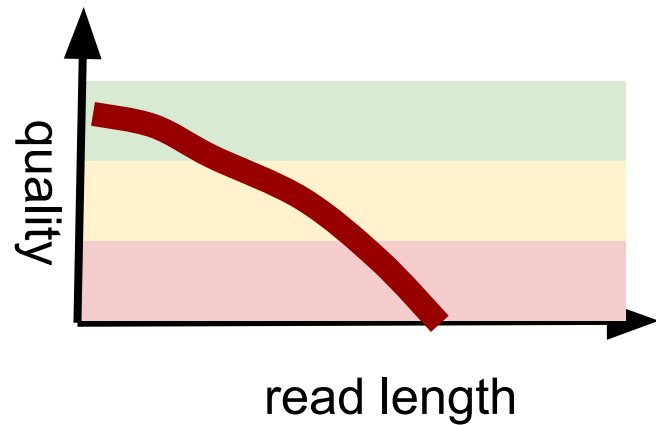
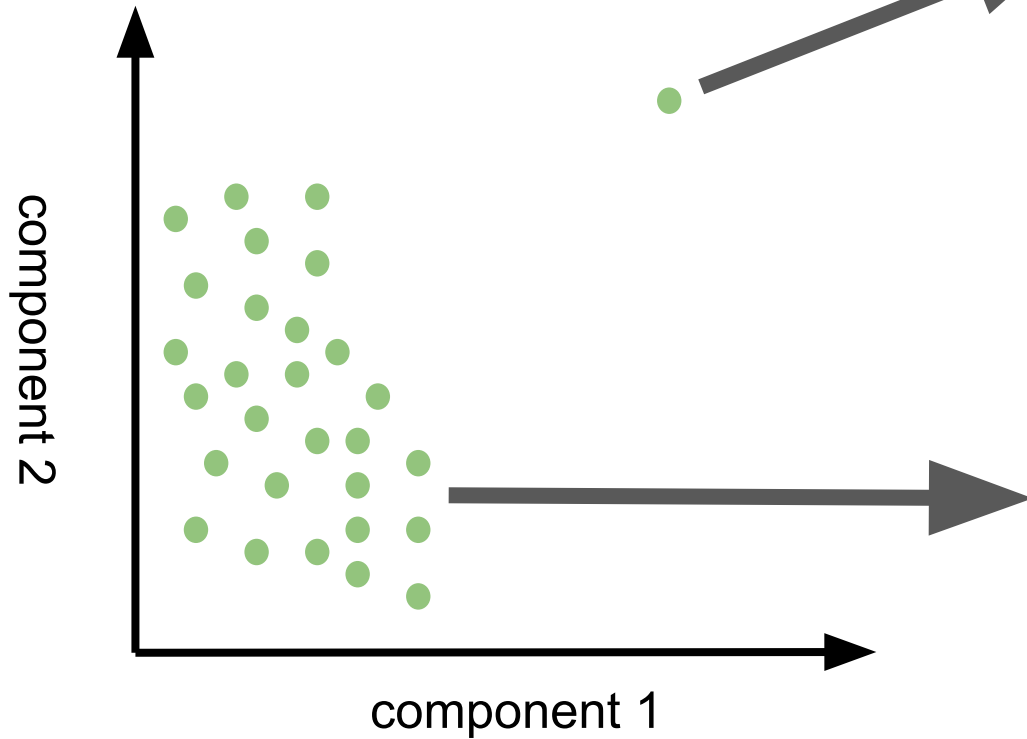
Application specific:
Variant calling,
count matrix, ...

Compare samples /
methods

Answer?

Why QC?

Principal component analysis













Quality scores

FastQC reports

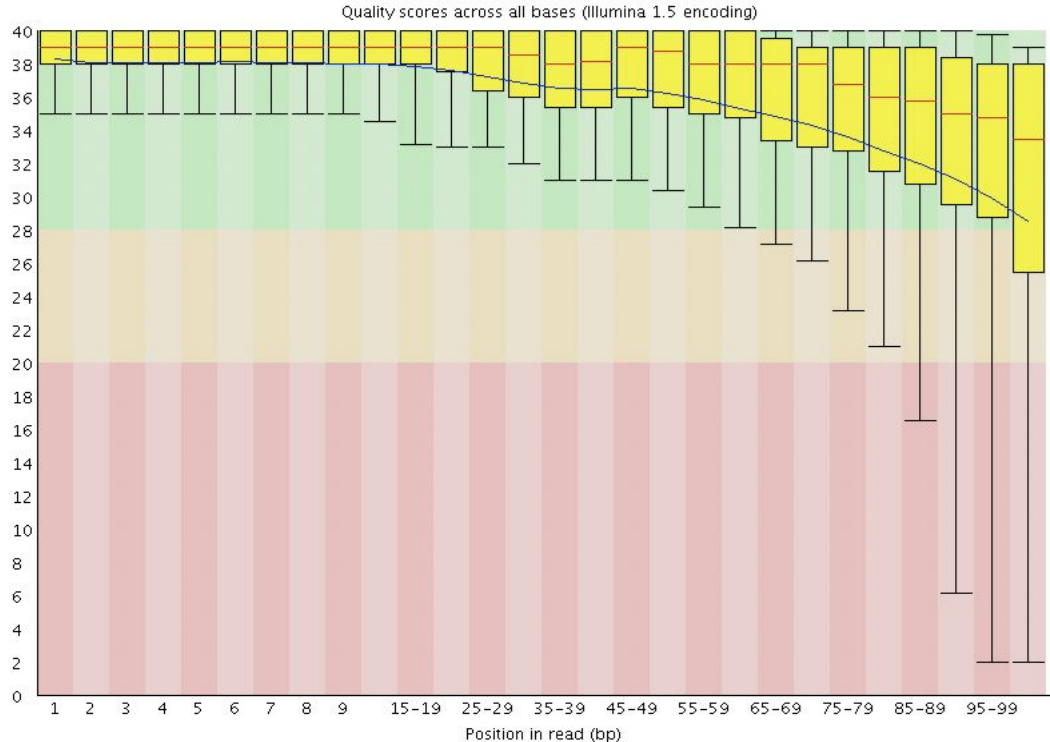
- Report basic statistics on your data
- Identify issues with your data
- **Use at each step of preprocessing to check progress**

FastQC Report

Summary

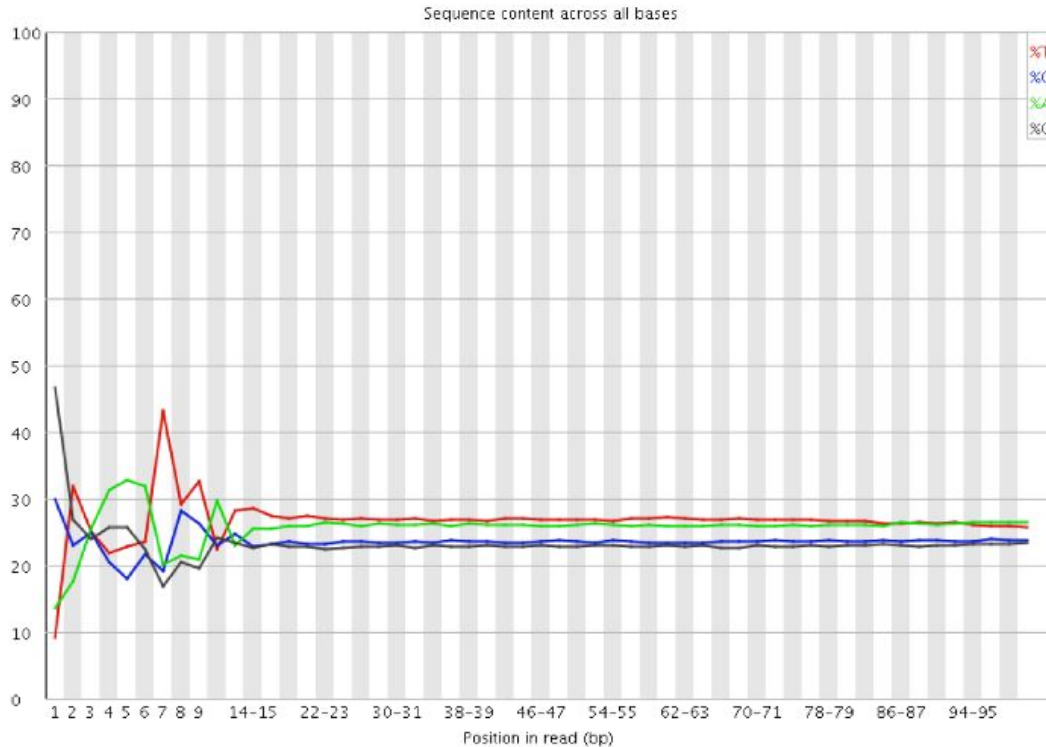
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented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

Per base sequence quality



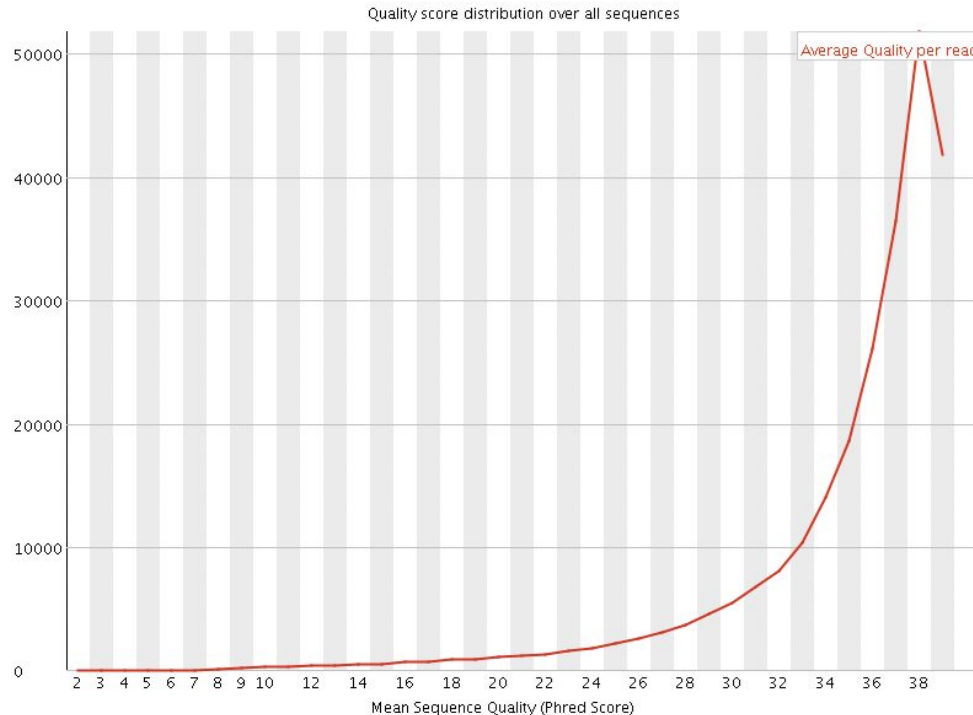
Do not panic
Quality often
decreases over the
read.

Trim from 5'



- Sometimes something is fishy in the beginning of the read.
- In case of bad quality, trim the first number of bases from the 5'.
- How many bases would you remove in this case?

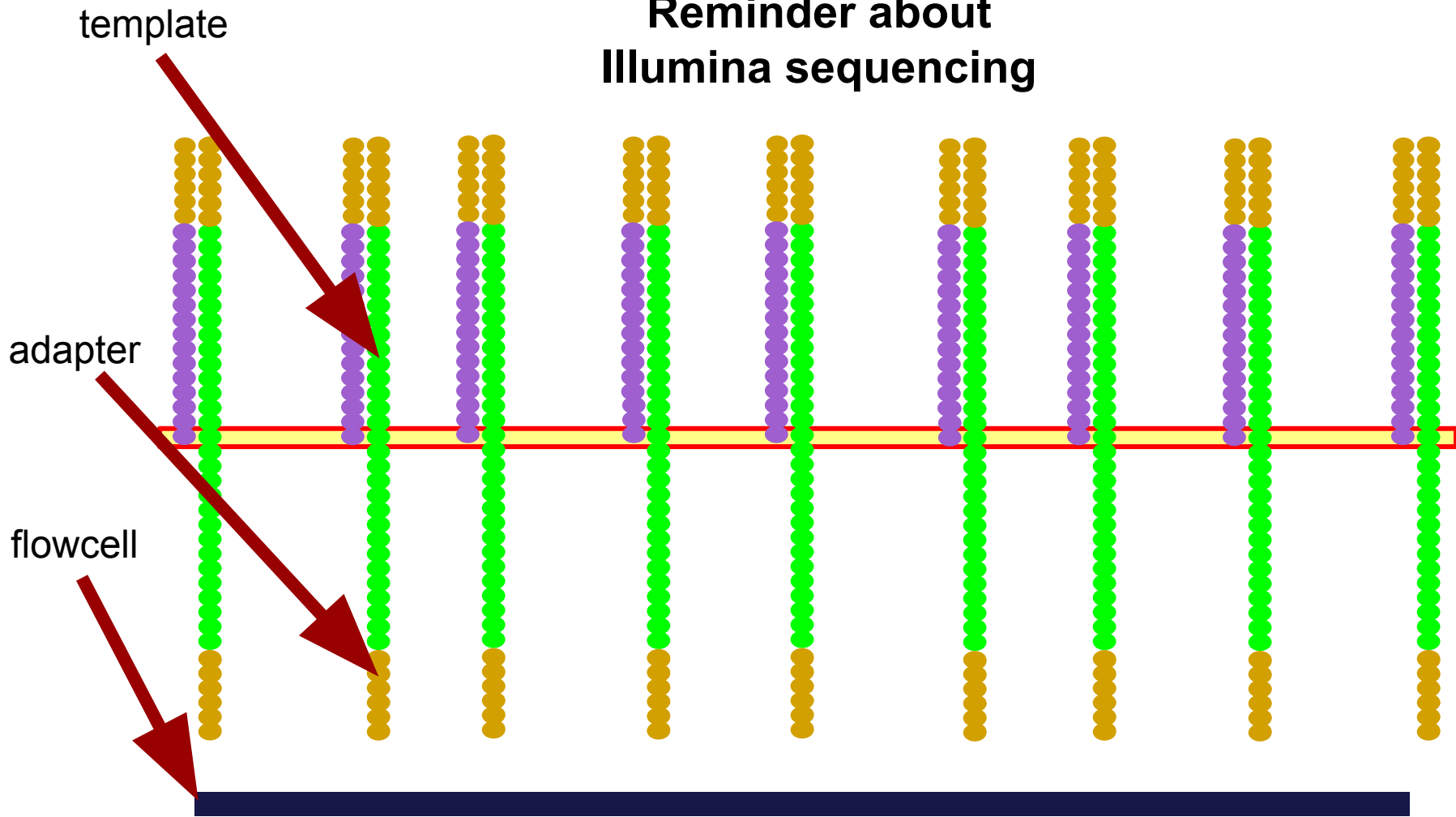
Short or low quality reads



- Sliding window to remove low quality regions
- Remove short reads

Adapters

Reminder about Illumina sequencing



adapter 1

adapter 2

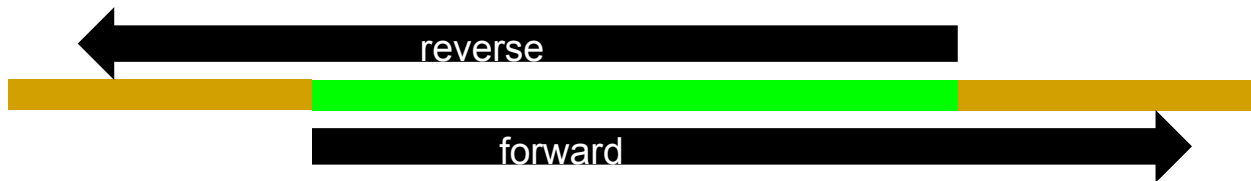
Single-end:



Single-end:

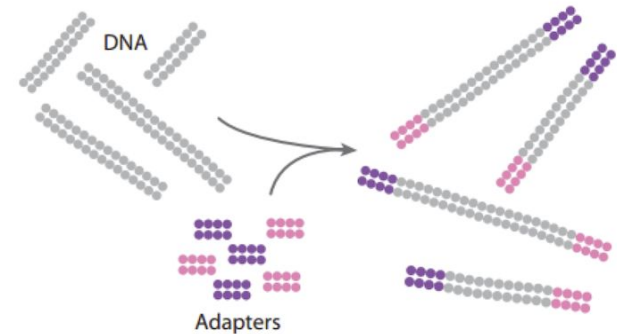


Paired-end:



Adapters

- Sometimes adapters / primers are also part of the read
- Short inserts are more prone to contain adapters
- Adapter / primers are non-biological sequences
- The artificial repeats will disturb alignments and *de novo* assembly
- The sequence is often known, if not, FastQC may still find them



Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

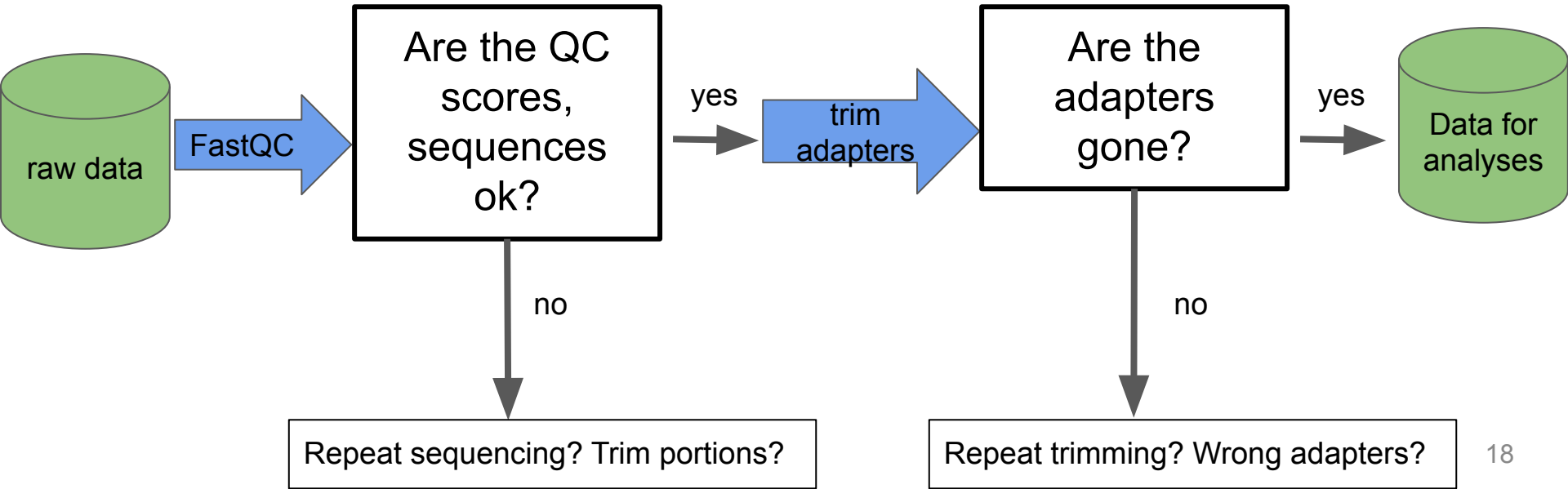
Adapters

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATATCGTATGC	1547768	38.192098035156306	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC	146635	3.61830603513262	TruSeq Adapter, Index 1 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCAAGATATCGTATGC	6639	0.16382128255358863	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATTCGTATGC	6462	0.15945370204267054	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACGATATCGTATGC	5433	0.1340625136486891	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAACGATATCGTATGC	5147	0.1270052931621209	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCAGATATCGTATGC	4703	0.11604932849066535	TruSeq Adapter, Index 1 (97% over 41bp)

Adapters

- Remove adapters before starting any analyses.



K-mer correction
(mostly for *de novo*)

K-mer correction

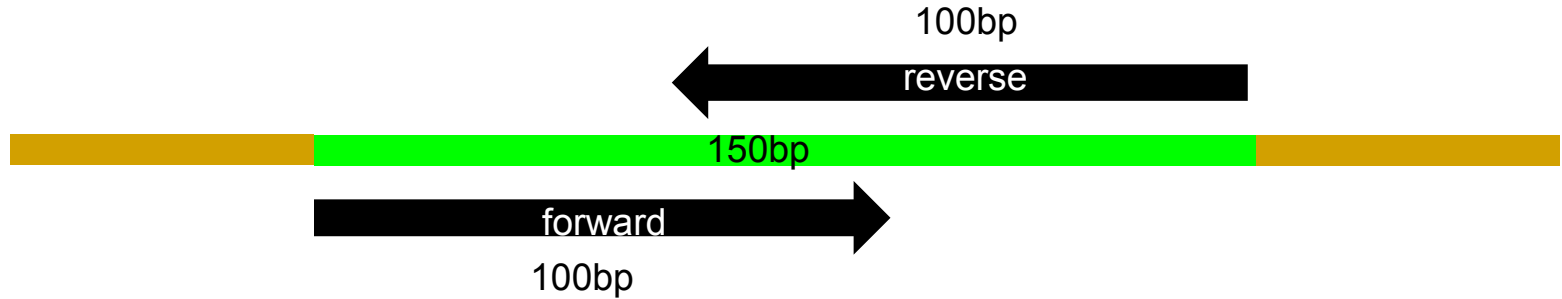
- Create a sliding window of size k , move it over all your reads and count occurrence of k -mers
- Example for 1 read, $k=4$:

read: ACGTGTAACGTGACGTTGGA

k-mers: ACGT
 CGTG
 GTGT
 TGTA

• • •

Merging paired-end reads



- 50 bp have been sequenced twice
- Merge both fwd+rev into a single sequence
- Error correction for free
- Useful for short inserts+*de novo* assembly

Quality control for long read technologies

We heard about other newer technologies yesterday

- PacBio, Nanopore etc.
- How can we do quality control on reads from these technologies?

Long reads quality control

Sequence analysis

NanoPack: visualizing and processing long-read sequencing data

Wouter De Coster^{1,*}, Sven D'Hert², Darrin T. Schultz³, Marc Cruts¹ and Christine Van Broeckhoven¹

¹Neurodegenerative Brain Diseases Group, ²Bioinformatics, Neuromics Support Facility, Center for Molecular Neurology, VIB & University of Antwerp, 2610 Antwerp, Belgium and ³Department of Biomolecular Engineering and Bioinformatics, University of California Santa Cruz, Santa Cruz, CA 95064, USA

JOURNAL ARTICLE

LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data

Yoshinori Fukasawa , Luca Ermini, Hai Wang, Karen Carty, Min-Sin Cheung 

[Author Notes](#)



Read & annotate PDF



Add to wizdom.ai

G3 Genes|Genomes|Genetics, Volume 10, Issue 4, 1 April 2020, Pages 1193–1196,
<https://doi.org/10.1534/g3.119.400864>

Published: 01 April 2020 **Article history** 

Final – but important note

- Lots of data - storage is expensive!
- Keep data compressed whenever possible (gzip, bzip, bam)
- Test workflows on subsets
- Remove intermediate files and files that can easily be re-created
- Learn Snakemake or Nextflow

