



LEIDEN UNIVERSITY MEDICAL CENTER

# Quality control

**Michiel van Galen**

**Leiden Genome Technology Center**

**Department of Human Genetics**

**Center for Human and Clinical Genetics**



## *Overview*

- Data and the flaws
- Quality control basics
- Tools and advanced methods

## *The data*

- FastQ: Expanded two line Fasta format
- Four lines per entry
- Sequence and per base phred quality combined
- Beware of different score offsets

```

1  @SEQ_ID
2  GATTTGGGGTTCAAAGCAGTA
3  +
4  ! ' * ((((***)%%)%%)++)(

```

Listing 1: FastQ format

*The flaws*

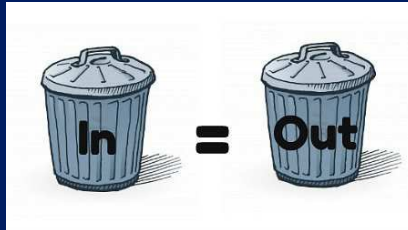
At any point from the start of the experiment until beginning analyses, quality can be jeopardized.

- Gathering material and sample prep
  - Contamination, degradation, sample swap
- Sequencing
  - Exhausted chemicals, technical issues
- Data integrity
  - File corruption
- Many other unexpected external factors

## *The consequence*

Low quality greatly influences the downstream analyses.

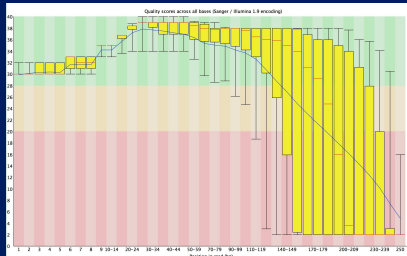
Figure: Garbage in garbage out



## *Quality assessment*

- FastQC: A quality control tool for high throughput sequence data.
- Assess the quality of your data in a fastq file

Figure: FastQC



## *Data properties*

Properties which can indicate possible biases in your data:

- Quality scores - Higher is better
- GC content - Expected vs observed
- Duplication rate - Lower is usually better
- N content - Less is more
- Adapter contaminants - More adapter, less sample
- kMer statistics - Expected vs observed

*Improving your data*

After identification of some issues, correction may be possible

- Low quality bases can be discarded
- Adapter sequences can be removed
- Downstream analyses can be tailored to identified problems



## *Quality trimming*

- Getting rid of low quality bases
- Only want to maintain the high-quality bases

```

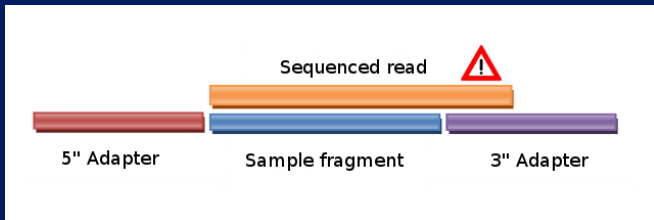
1  @Header
2  ACGTACGTACGT
3  +
4  !# II!JJJI##!
5
6  Will result in:
7  —GTACGTA—

```

## *Clipping adapters*

- FastQC can identify adapter contaminants which can hamper later analyses
- Specific tools can remove these specific sequences

Figure: Adapter Sequencing



## *Digital data quality*

Also digital data can be of low quality

- Hardware failure
  - Data corruption, insufficient disk space
- Human failure
  - Sample swaps, unclear file names, incomplete copies

## *kMer analysis*

- Analyzing the frequencies of words of length  $K$
- Proven to detect all sorts of factors which influence the data
  - Contamination, quality, duplication
- Also used to determine sample complexity

*Overview of tools*

- Quality assessment
  - FastQC, kMer, QCDB
- Trimming
  - Sickle: A windowed adaptive trimming tool
- Adapter clipping
  - Cutadapt
- File integrity
  - Md5checksums, GRP

## *QC process*

Good QC practice can be performed following the next steps:

- Assess the quality of raw data
- Identify possible factors that impact the data
- Apply the tools to improve the data
- Assess the quality again and evaluate the results

Preferably this can be done in a precompiled pipeline

## Acknowledgements:

Jeroen Laros

Martijn Vermaat

Jeroen Frank

LGTC