Assembly vs Alignment
and
Quality Assessment

Assembly vs. Alignment

- Alignment
  - Aligning reads to a reference sequence
- Assembly
  - You don’t have a reference, you need to build one
  - Genome or transcriptome

Alignment

- You have a reference genome

```
Known -> CTCCTAGAATGCTGGGAAGTGGAAGTCCAACTTCTTCCATGGGTTCACCT
```

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Alignment

- You have a reference genome

```
Known -> CTCCTAGAATGCTGGGAAGTGGAAGTCCAACTTCTTCCATGGGTTCACCT
```

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Alignment

- You have a reference genome

```
Known -> CTCCTAGAATGCTGGGAAGTGGAAGTCCAACTTCTTCCATGGGTTCACCT
                 TAATGCTGGGAAAGTG
```

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Alignment

- You have a reference genome

```
Known -> CTCCTAGAATGCTGGGAA-GTGGAAGTCCAACTTCTTCCATGGGTTCACCT
                TAATGCTGGGAAAGTG
```

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Alignment

- You have a reference genome

```
Known -> CTCCTAGAATGCTGGGAA-GTGGAAGTCCAACTTCTTCCATGGGTTCACCT
         CTCCTATAATGCTGGG
                TAATGCTGGGAAAGTG
                     CTGGGAAAGTGGAAG
                        GGAAAGTGGAAGTCC
                                    GTCCAACTTCTTG
                                         CTTCTTGGATGGG
                                              TGGATGGGTTAAC
                                                 ATGGGTTAACCT
```

Assembly

- No reference.
- What does this region look like?

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Assembly

- No reference.
- What does this region look like?

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
GGAAAGTGGAAGTCC
GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTGGGAAAGTGGAAG
CTTCTTGGATGGG
```

Assembly

- No reference.
- What does this region look like?

```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
         GGAAAGTGGAAGTCC
      CTGGGAAAGTGGAAG

GTCCAACTTCTTG
CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTTCTTGGATGGG
```

Assembly

- No reference.
- What does this region look like?

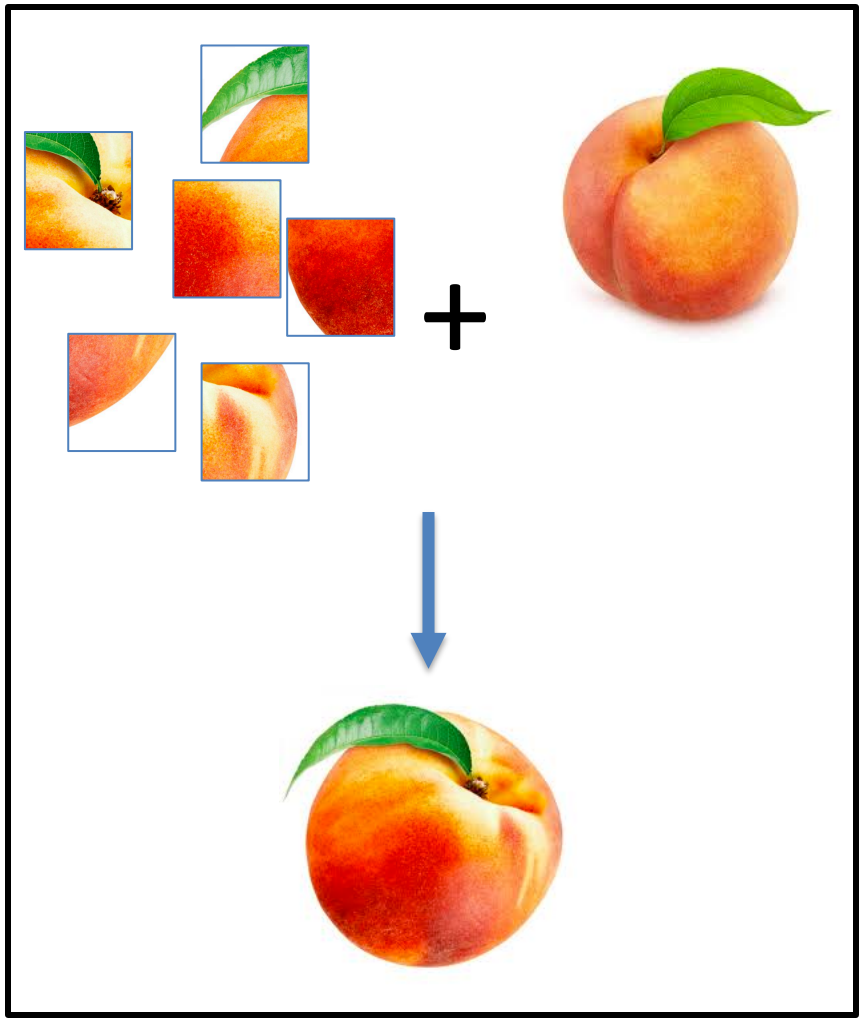
```
Sequences from Sequencing company:
TAATGCTGGGAAAGTG
         GGAAAGTGGAAGTCC
      CTGGGAAAGTGGAAG
                     GTCCAACTTCTTG

CTCCTATAATGCTGGG
TGGATGGGTTAAC
ATGGGTTAACCT
CTTCTTGGATGGG
```

Alignment

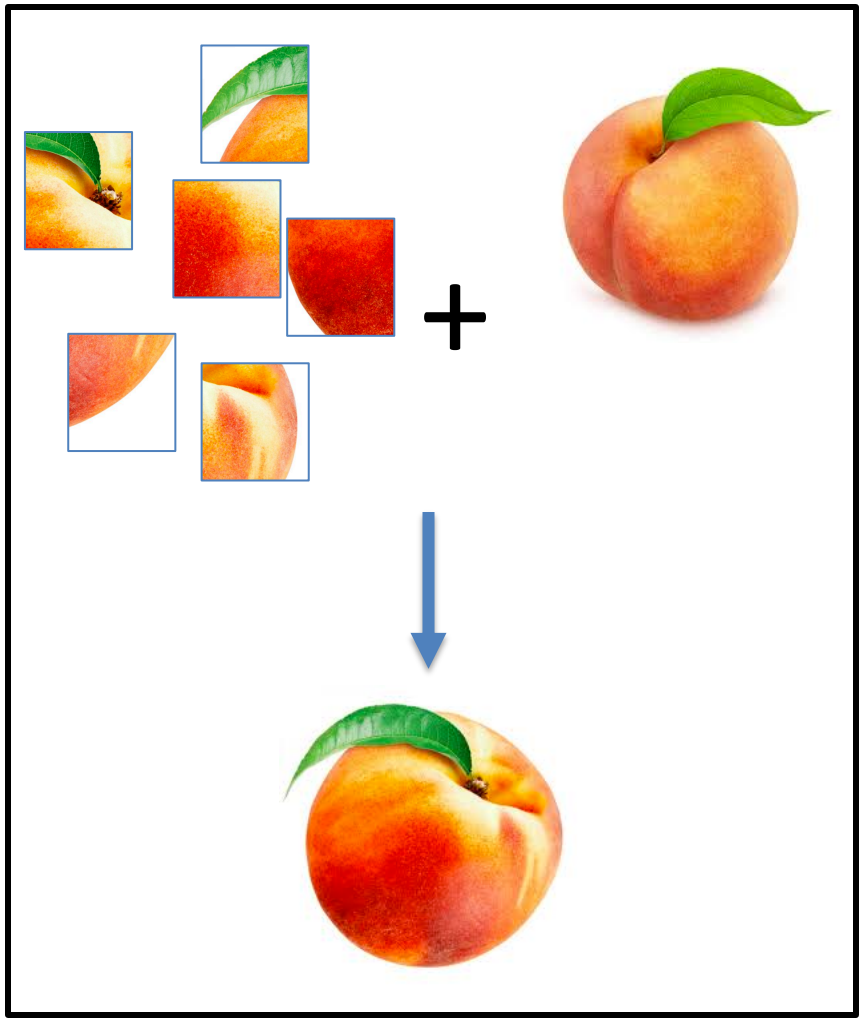
- You have a reference genome

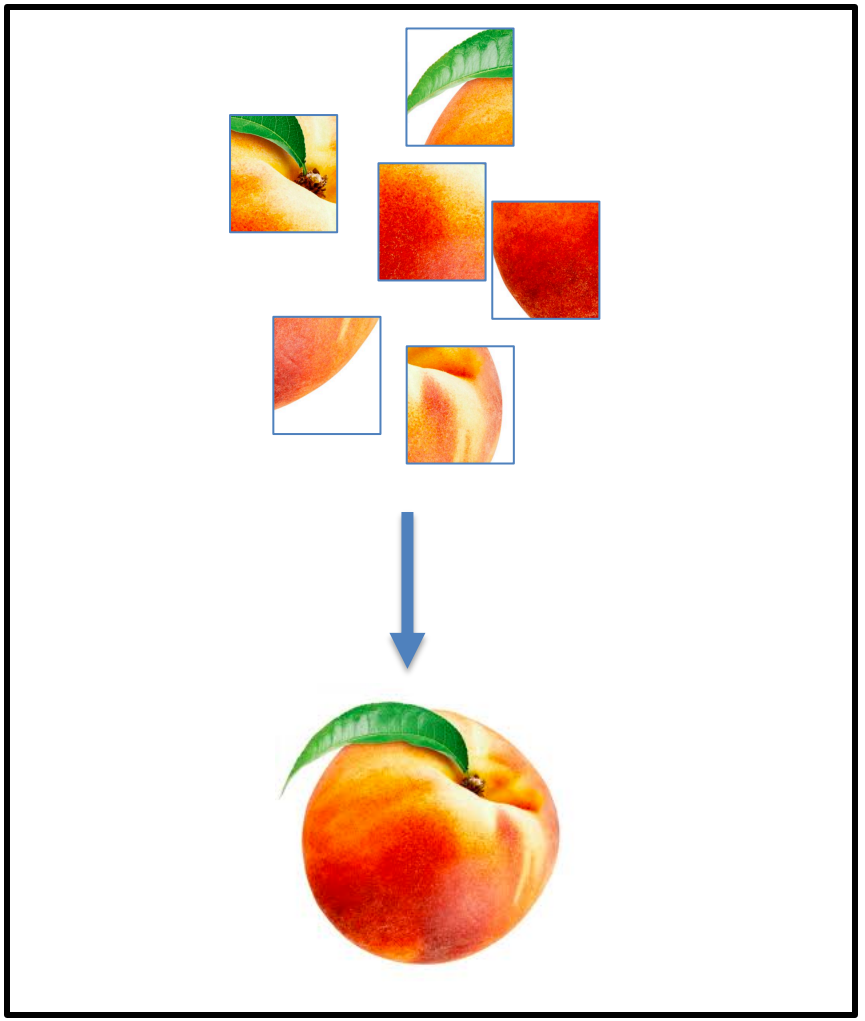
```
          CTCCTATAATGCTGGG
                TAATGCTGGGAAAGTG
                     CTGGGAAAGTGGAAG
                        GGAAAGTGGAAGTCC
                                    GTCCAACTTCTTG
                                          CTTCTTGGATGGG
                                               TGGATGGGTTAAC
                                                  ATGGGTTAACCT

Assembly -> CTCCTAGAATGCTGGGAAGTGGAAGTCCAACTTCTTGGATGGGTTAACCT
```

Alignment vs Assembly

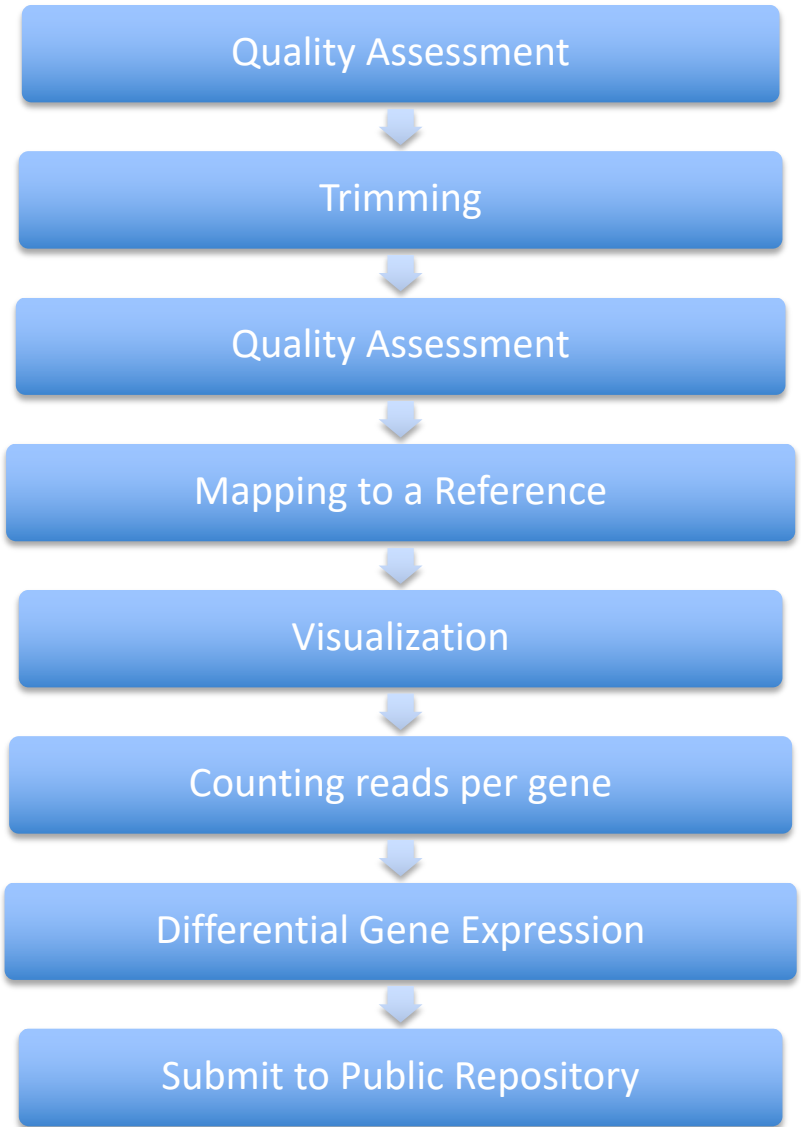


Alignment vs Assembly





mRNA Data Analysis Pipeline



FastQC — Babraham Bioinformatics

Skewer

FastQC — Babraham Bioinformatics

STAR

igv Integrative Genomics Viewer

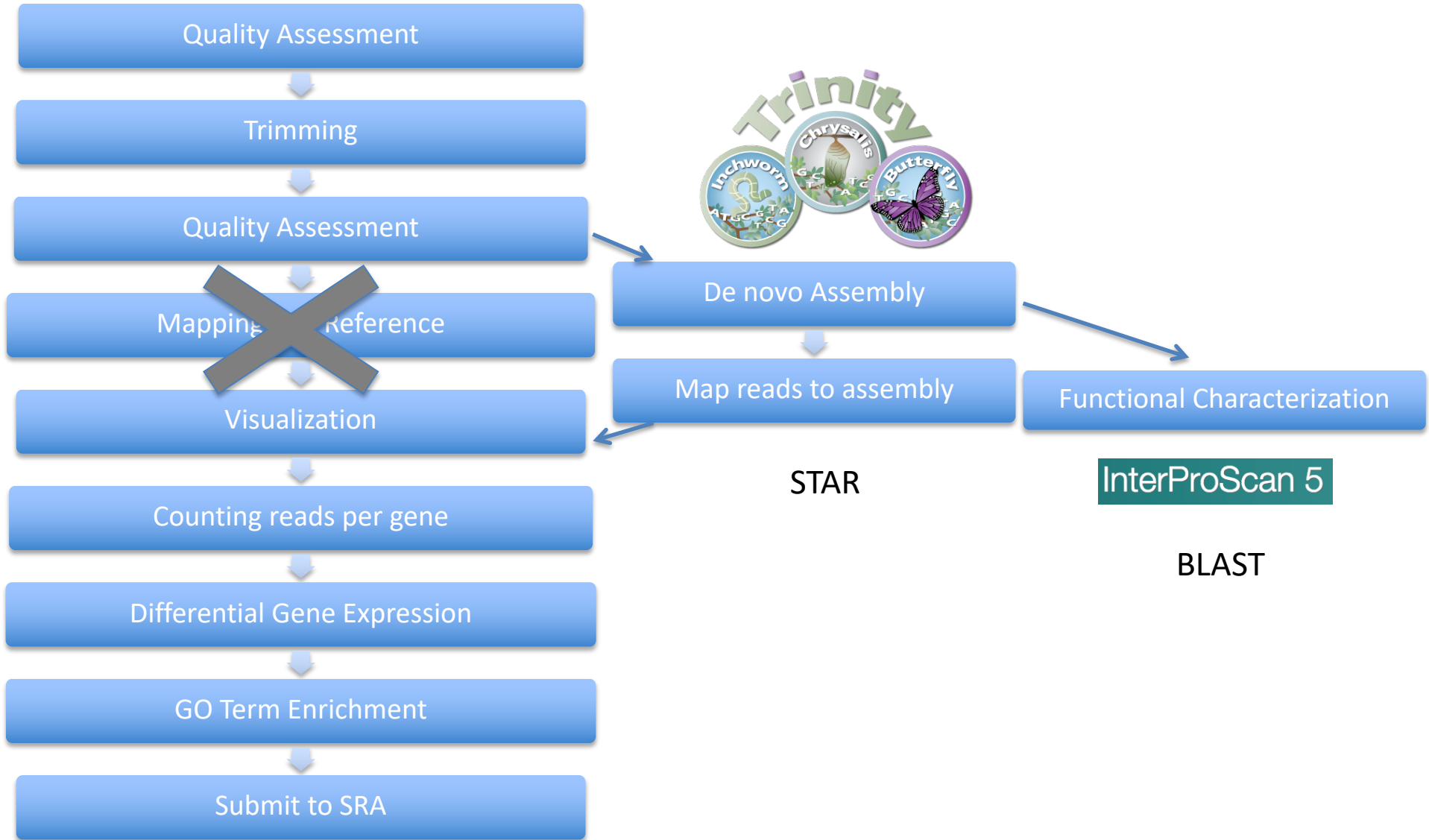
HTSeq

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

DESeq2

NCBI

What if you don't have a reference?



Quality Assessment

Quality Assessment

What the researcher cares about:

- Yield – did you get the number of reads you expected?
- Error - are the bases of reliable quality?
- Representative of sample – do the reads accurately represent the sample?

How do we get this information?

1. Think about sample quality and library quality

- Input sample quality is crucial for good sequencing
- High quality
- If quantity is low, use a kit designed fc
- Check
  - spectrophotometric (Nanodrop)
  - fluorimetric (Pico- and Ribo-Green)
  - gel electrophoretic methods (Bioanalyser)
    - RNA Integrity Number (RIN)
- If you have someone else prepare your libraries, they should be able to tell you minimum necessary quantity, concentration, and QC values.

1. Think about sample quality and library quality

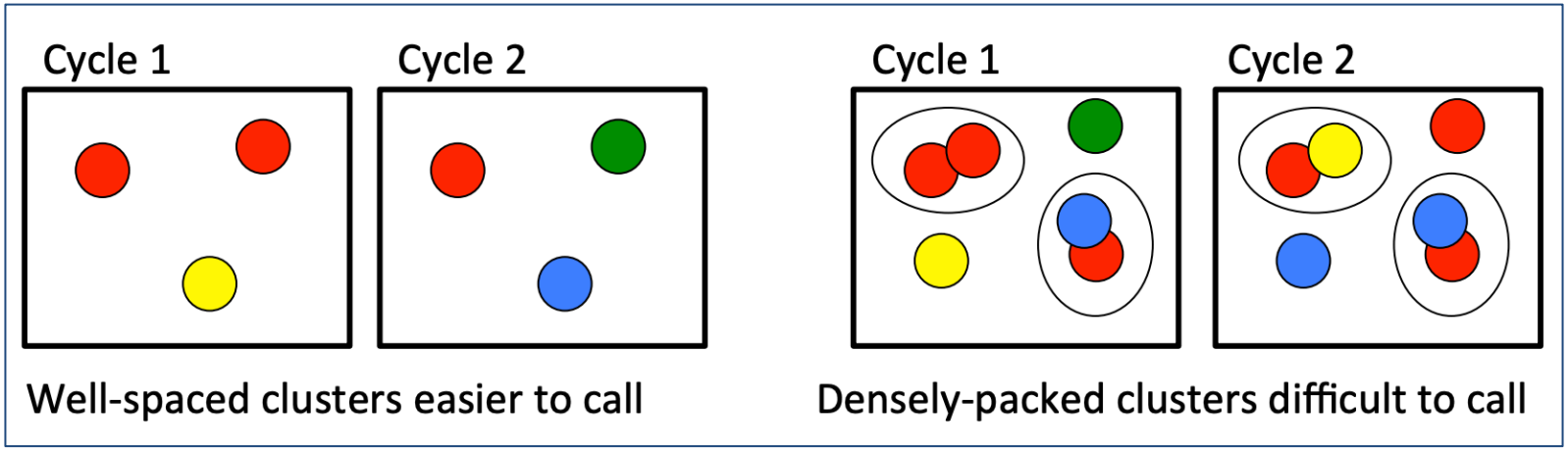
- Library quality
  - Bioanalyzer
  - Fragment Analyzer
  - Agarose Gel

https://support.illumina.com/bulletins/2016/05/library-quantification-and-quality-control-quick-reference-guide.html

- Library quantity
  - Flourometric = a dsDNA-specific fluorescent dye method, such as QuBit, PicoGreen, and AccuClear
  - qPCR
- Why?
  - confirm insert size
  - no primer dimers

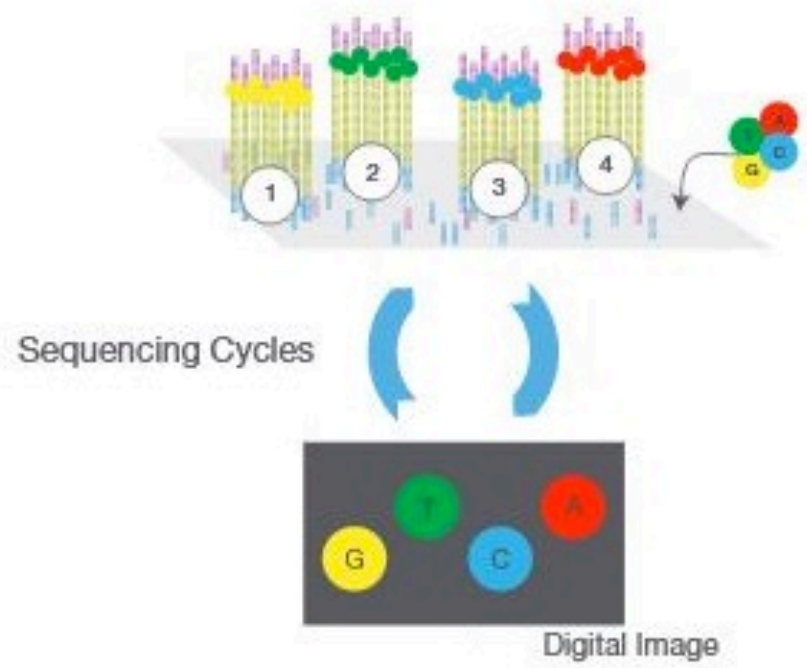
2. Instrument Metrics

- Cluster density
  - You need the right amount of DNA per run, which varies by instrument
  - Under clustering – quality is generally fine but you lost yield
  - Over clustering – quality problems!
  - Avoid by properly quantifying your library



2. Instrument Metrics

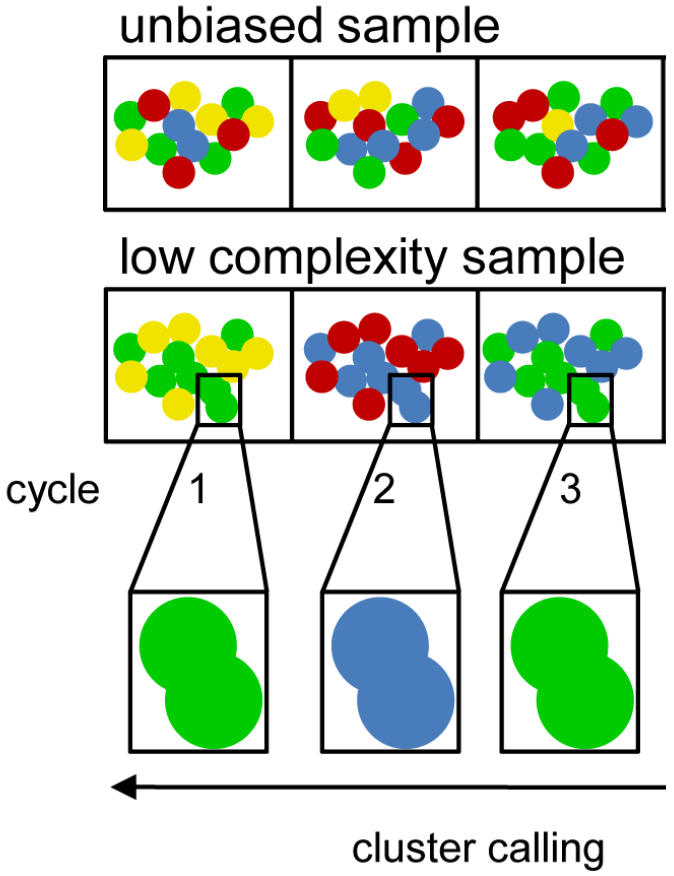
- PF% - percent passing filtering
  - single molecule = single cluster = clear signal
  - Anything else doesn't pass the filter
- Phasing/Prephasing
  - Sometimes individual molecules in a cluster become out of sync
- Q30 – bases over quality value 30



https://www.well.ox.ac.uk/ogc/sequencing-quality-monitoring-run/

PhiX Spike In

- Libraries must be diverse for proper sequencing
- Fix with a Phi-X spike in
  - A known quantity of DNA from a bacteriophage
  - 10-20% of sequences (or more)



Krueger et al 2011
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016607

https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html

3. Data Assessment

Software options

Illumina SAV – Sequence analysis viewer
- Part of BaseSpace
- Your sequencing facility probably uses this

FastQC
- Free
- Runs on linux

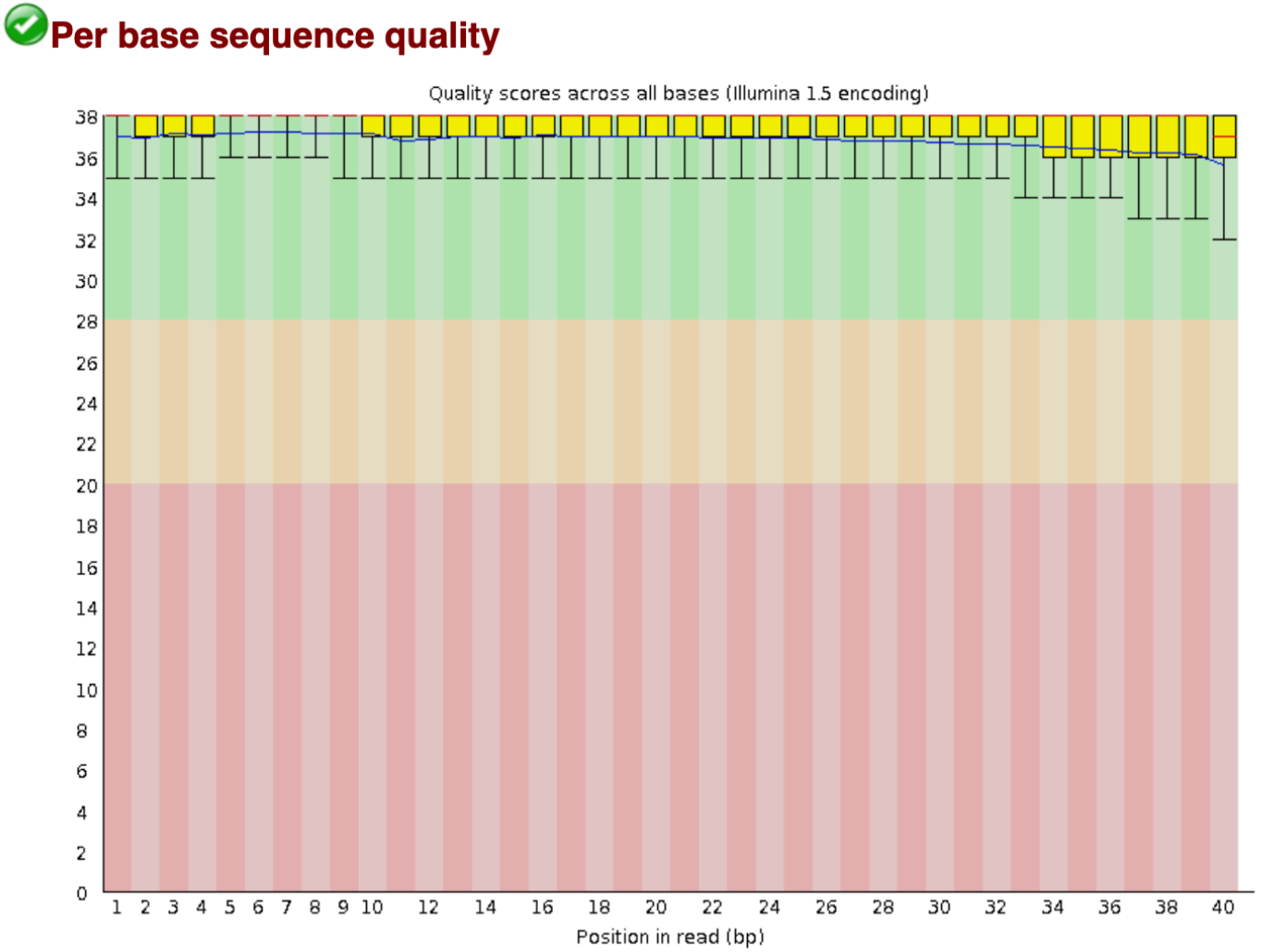
FastQC – Basic Statistics

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

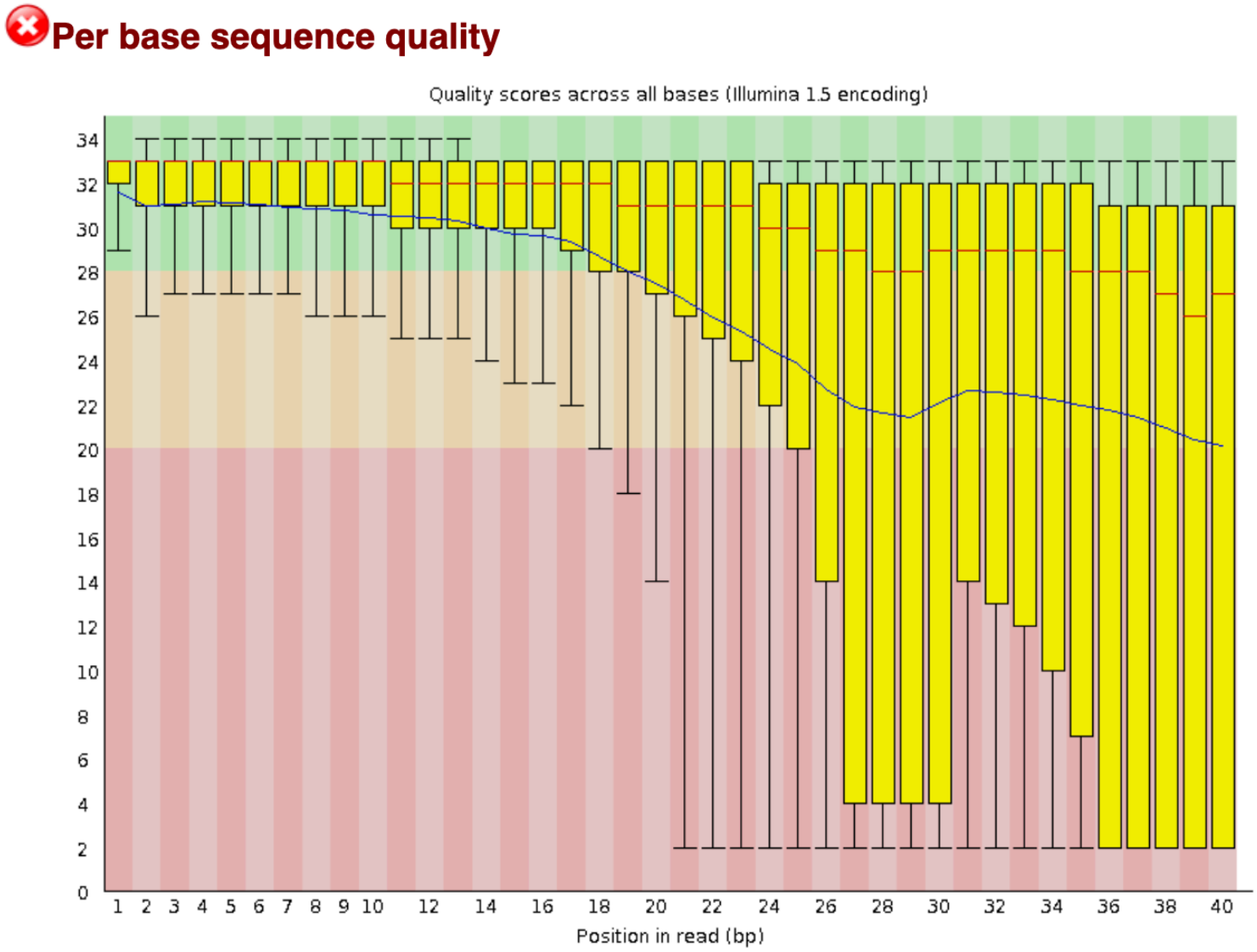
FastQC – Per Base Sequence Quality

GOOD



FastQC – Per Base Sequence Quality

BAD



FastQC – Example Reports

- Good report

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

- Bad report

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

- Adapter Dimer report (ligated adapters w/ no insert sequence)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/RNA-Seq_fastqc.html

- Small RNA with read through report:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/small_rna_fastqc.html