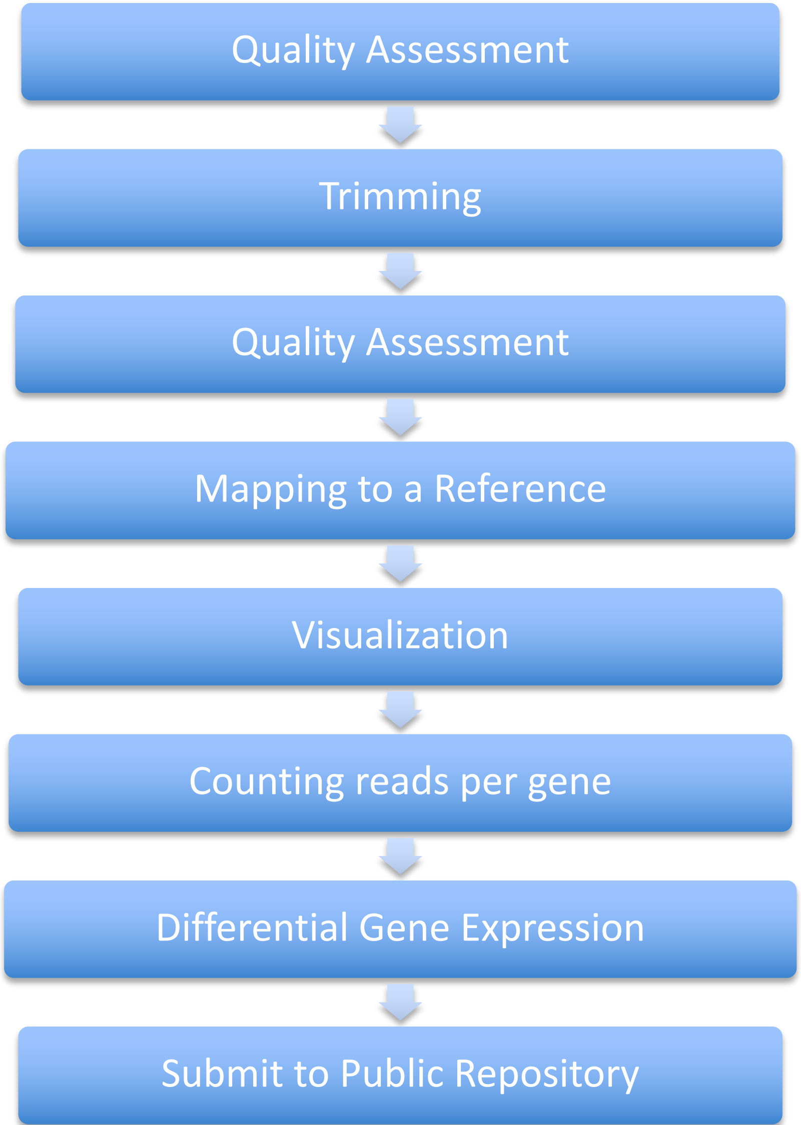
mRNA Data Analysis Pipeline


Quality Assessment

Trimming

Quality Assessment

Mapping to a Reference

Visualization

Counting reads per gene

Differential Gene Expression

Submit to Public Repository

FastQC Babraham Bioinformatics

Skewer

FastQC Babraham Bioinformatics

STAR

igv Integrative Genomics Viewer

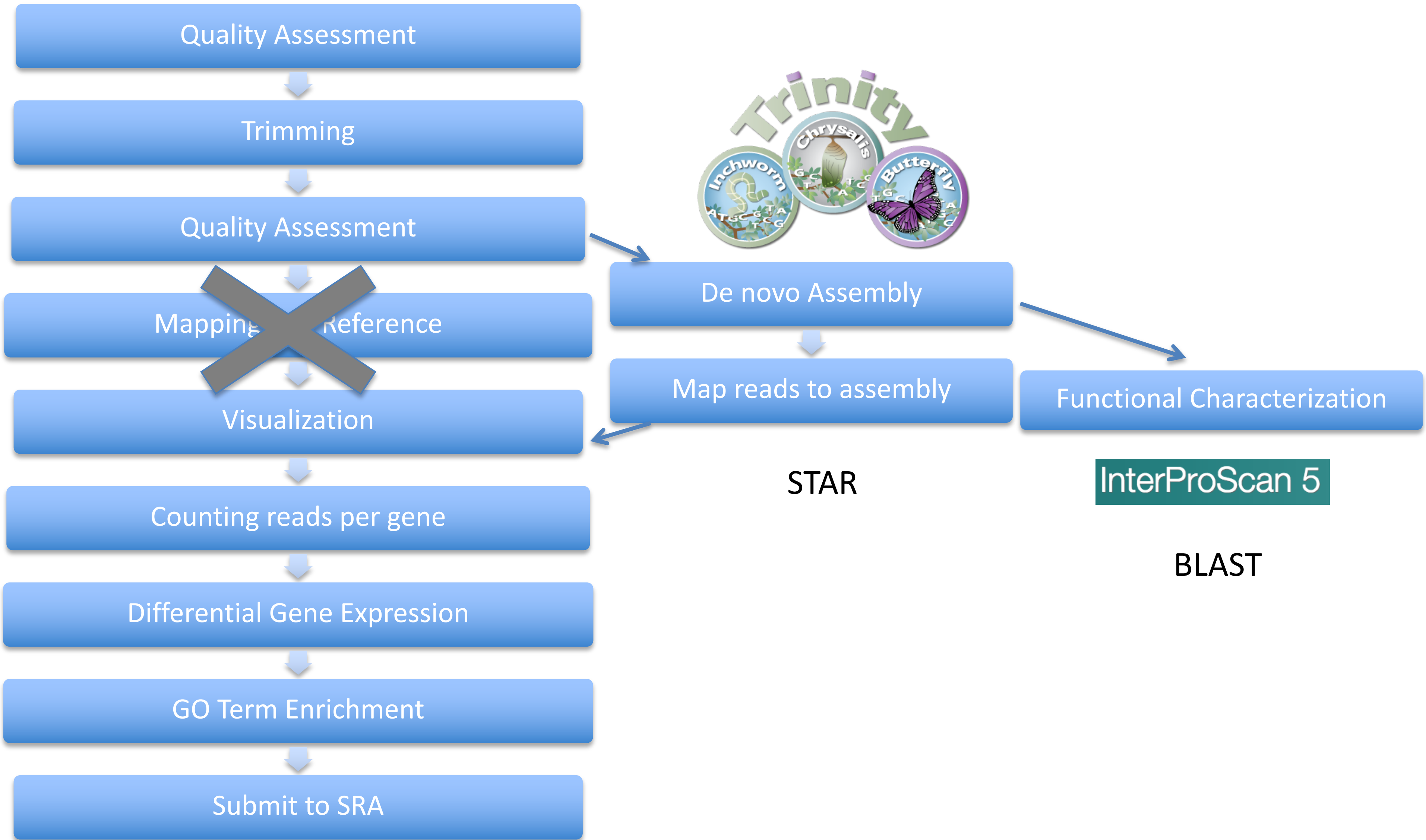
HTSeq

DESeq2

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

NCBI

What if you don't have a reference?



Quality Assessment

Trimming

Quality Assessment

Mapping to Reference

Visualization

Counting reads per gene

Differential Gene Expression

GO Term Enrichment

Submit to SRA

De novo Assembly

Map reads to assembly

Functional Characterization

STAR

InterProScan 5

BLAST

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.

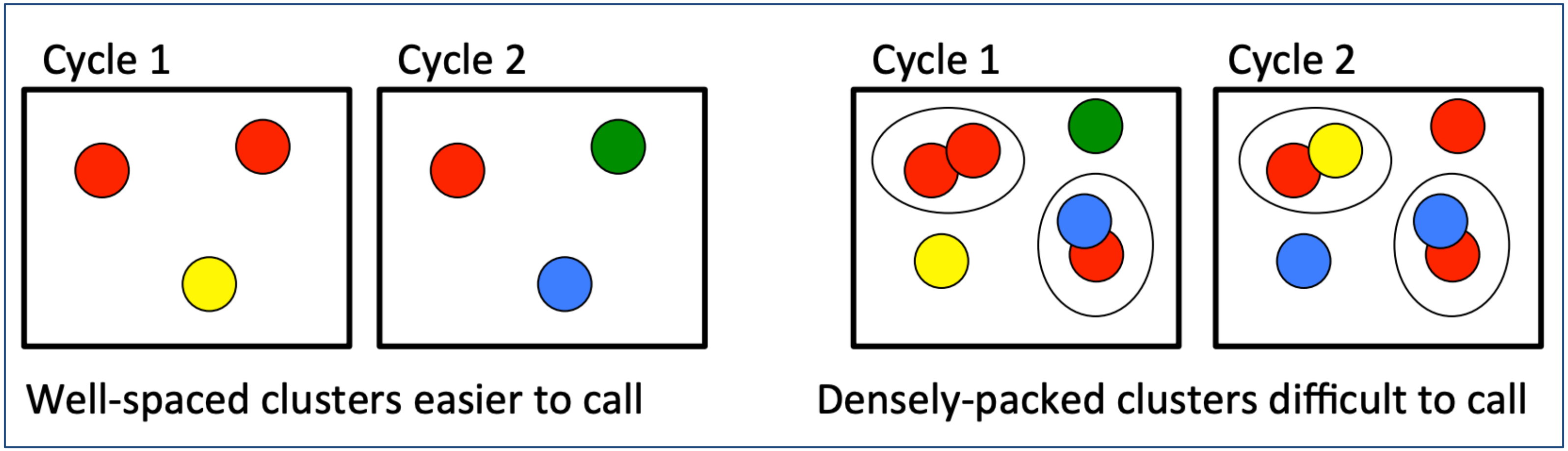
1. Think about sample quality and library quality

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

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

2. Instrument Metrics

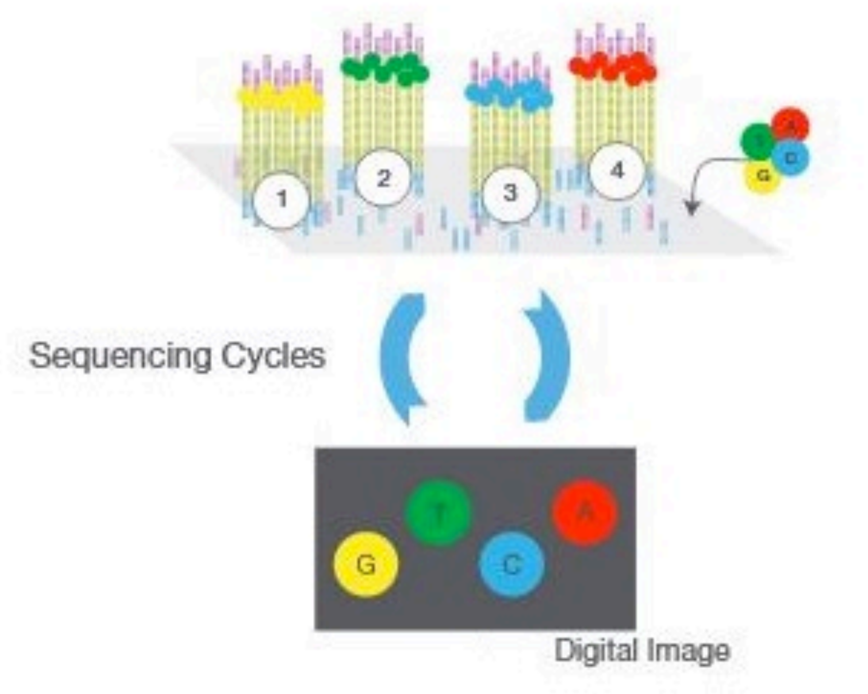
- Cluster density
  - You need the right amount of DNA per run, which varies by instrument
  - Under clustering – quality is generally fine but you lost yield
  - Over clustering – quality problems!
  - Avoid by properly quantifying your library


Cycle 1
Cycle 2
Cycle 1
Cycle 2
Well-spaced clusters easier to call
Densely-packed clusters difficult to call

https://bficores.colorado.edu/biofrontiers-core-facility-workshops/workshops-in-the-series/2017-illumina-sequencing-qc-troubleshooting-workshop/qc-workshop-files/part3-sav

2. Instrument Metrics

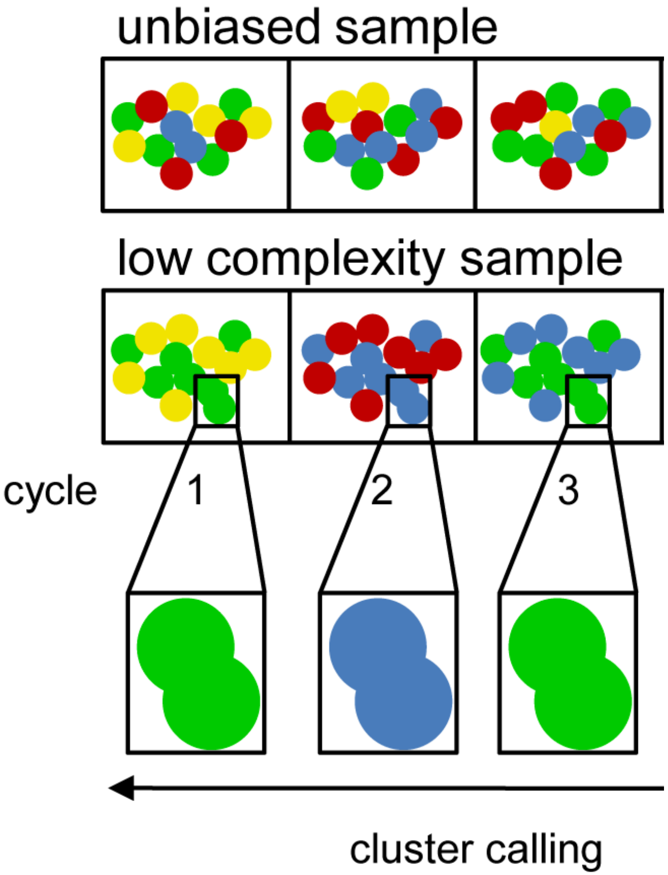
- PF% - percent passing filtering
  - single molecule = single cluster = clear signal
  - Anything else doesn’t pass the filter
- Phasing/Prephasing
  - Sometimes individual molecules in a cluster become out of sync
- Q30 – bases over quality value 30



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

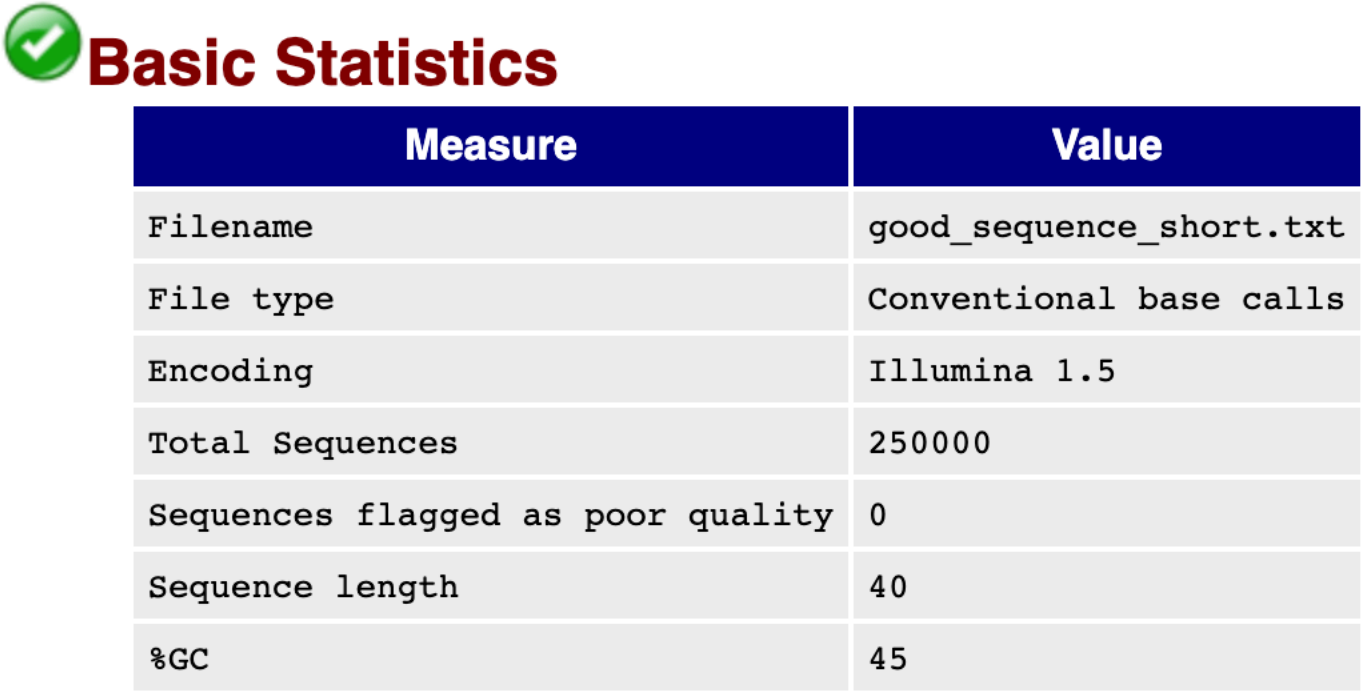
- Part of BaseSpace
- Your sequencing facility probably uses this

FastQC

- Free
- Runs on linux

Babraham Institute Babraham Bioinformatics

FastQC – Basic Statistics

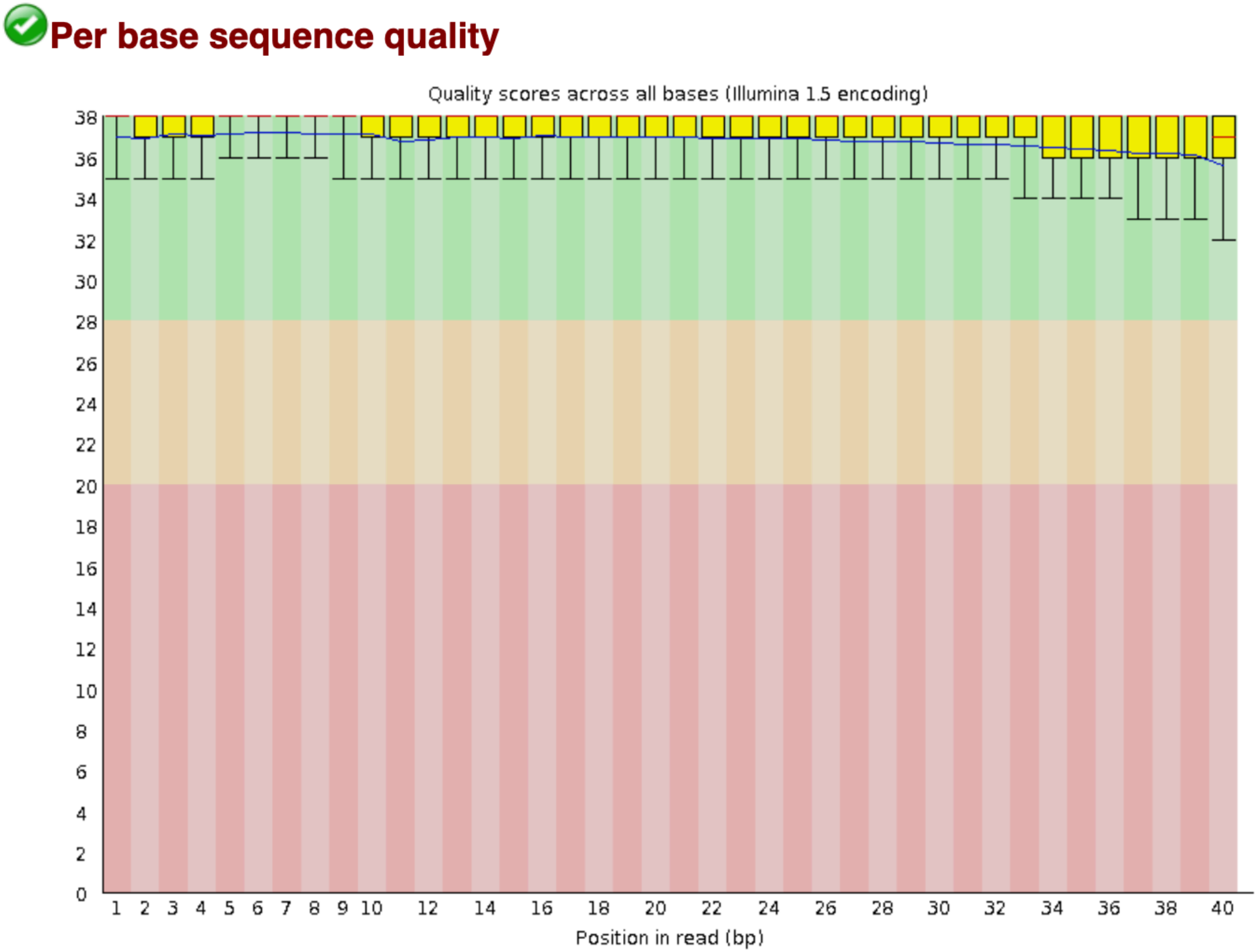


✓ Basic Statistics

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

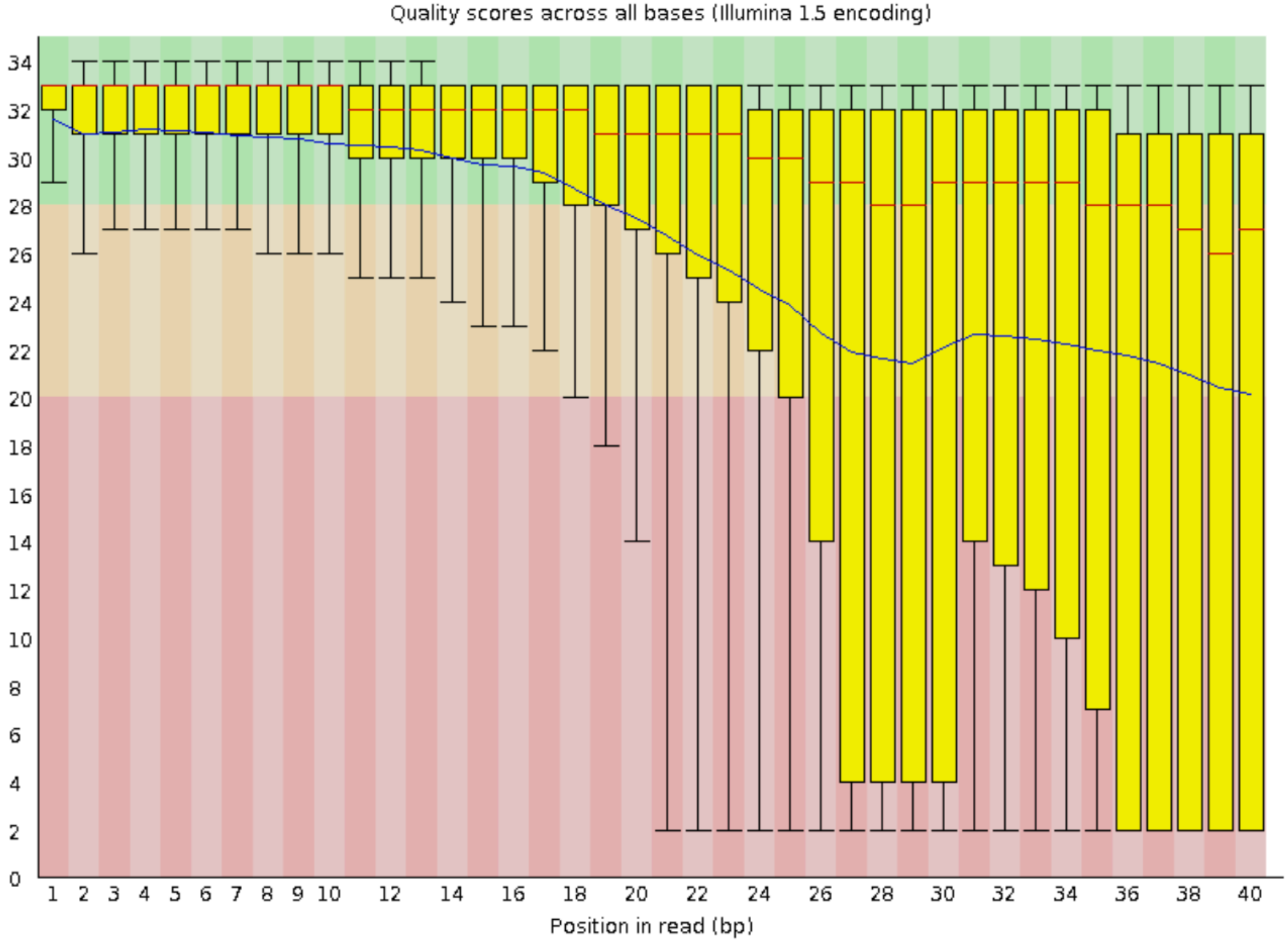
GOOD



FastQC – Per Base Sequence Quality

BAD

☒ **Per base sequence quality**



FastQC – Example Reports

- Good report

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

- Bad report

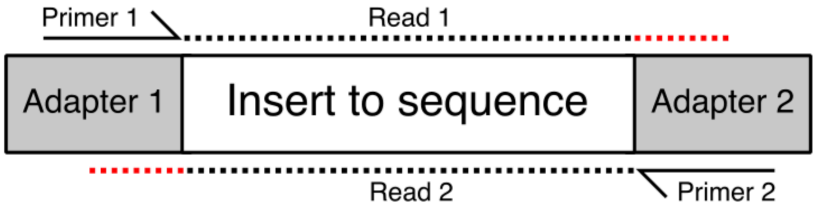
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

- 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



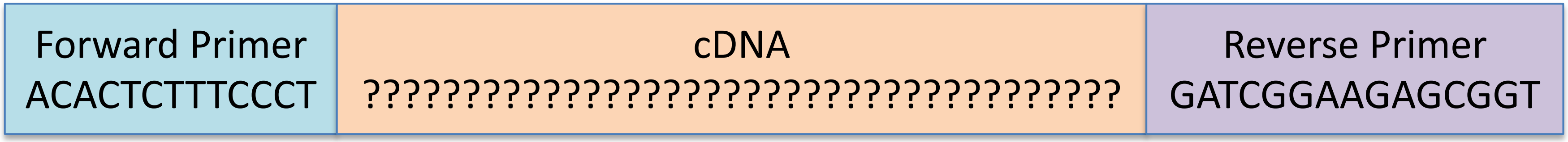
Trimming

Trimming

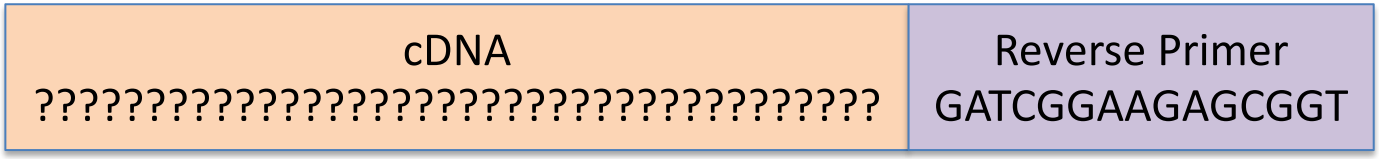
- From the quality control step, we know where the problems are
- All Illumina reads tend to have degrading quality at the end of the read
- Get rid of the bad data, keep the good data
  - Cut adapter sequences from the read.
  - Trim off low quality bases
  - Drop a read entirely if is too low quality or too short

Adapter Trimming

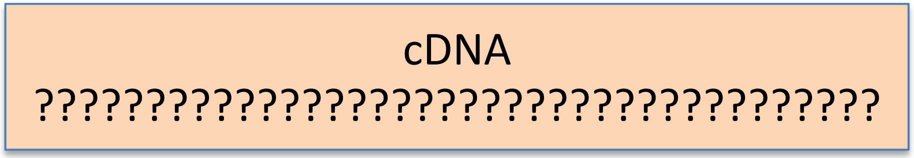
Library Fragment:


Forward Primer
ACACTCTTTCCCT
cDNA
??????????????????????????????????????
Reverse Primer
GATCGGAAGAGCGGT

Read returned from sequencing facility:

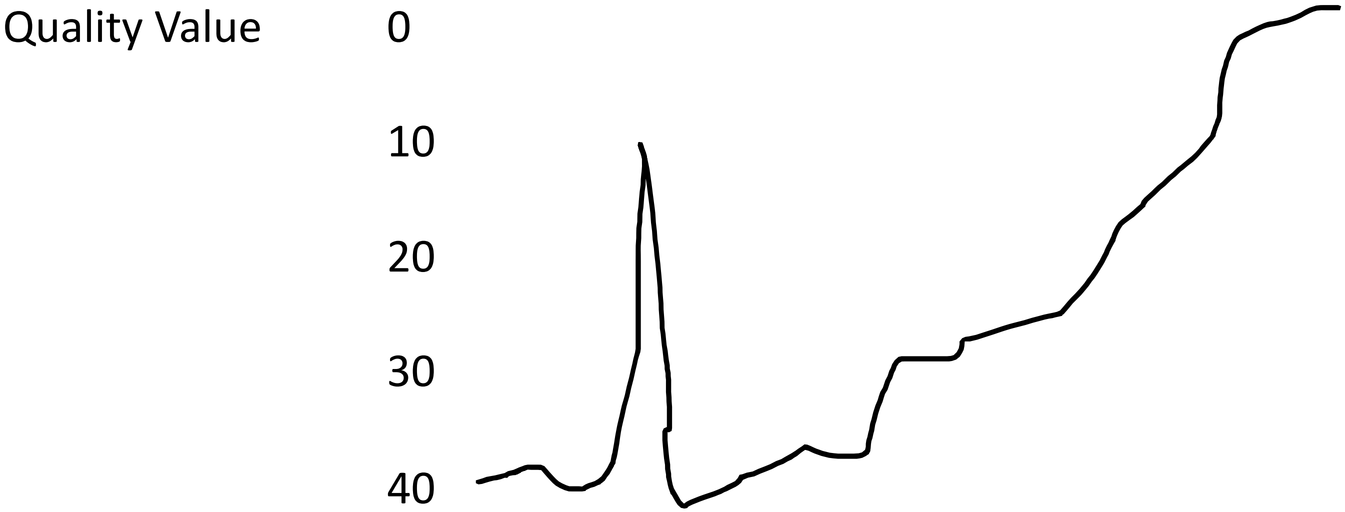

cDNA
??????????????????????????????????????
Reverse Primer
GATCGGAAGAGCGGT

After Trimming:


cDNA
??????????????????????????????????????

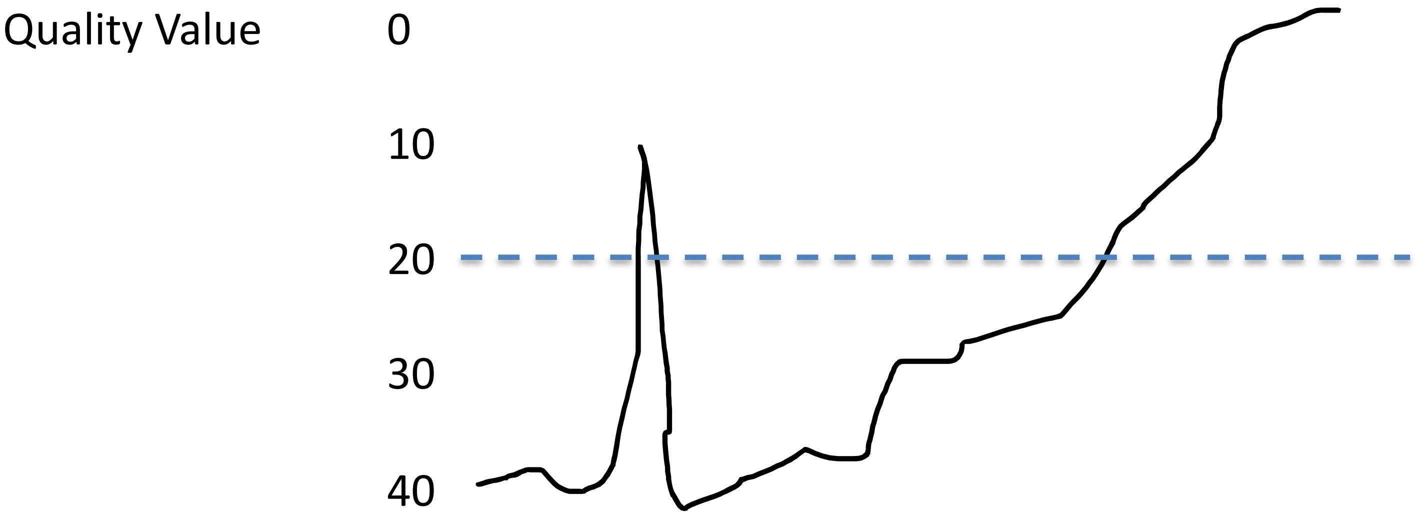
Quality Trimming

cDNA
??????????????????????????????????????



Quality Trimming

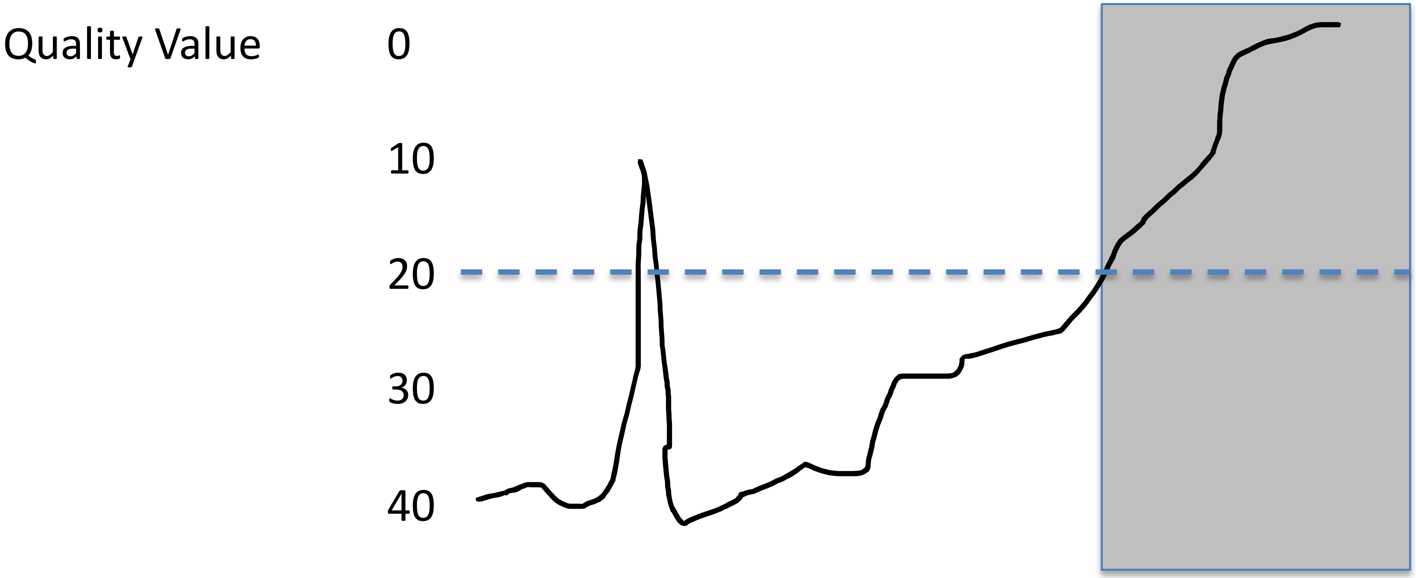
cDNA
??????????????????????????????????????



What is your cut off?

Quality Trimming

cDNA
???????????????????????????????????????



Generally its ok to keep a single base or a few low quality bases to preserve downstream quality. But once the quality has degraded to the point of the bases being largely useless, make the trim.

Is trimming necessary?

- Depends on what you are doing with the data
- Balance data loss with downstream accuracy

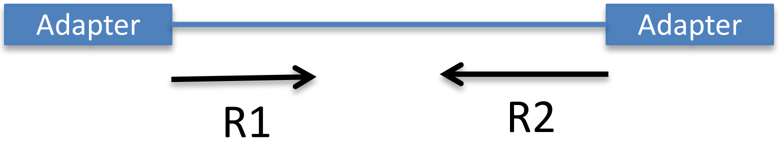
If you meet all these criteria, maybe not:

- The reads are of high quality and have minimal adapter contamination
- You are mapping the reads to a well annotated reference genome
- You are doing gene quantification (no assembly, no variant calling)

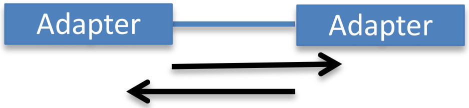
Trimmomatic

- Optimized for Illumina NGS
- Very flexible
- Handles paired end data well
- Threaded

- Detects adapter read through
- No read through:



- Read through:



Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

Trimmomatic

Defaults are very stringent, but you can adjust.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read. Must specify adapter sequence. Comes with basic Illumina adapter files, make sure yours are in there or add yours!
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length after trimming

Skewer

- Faster than trimmomatic
- “Gentle” quality trimming by default
- Utilizes quality scores in adapter identification and allows insertions/deletions

https://github.com/relipmoc/skewer

Jiang, H., Lei, R., Ding, S.W. and Zhu, S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics, 15, 182.

Trimming

- Current community wisdom:
  - Quality trimming reduces error
  - But also reduces content and contiguity
- Gentle trimming is preferred – many times the defaults are too stringent, you will lose lots of data!
- Application matters
  - For expression, gentle to no trimming (phred 3 to 5)
  - For assembly and variant calling, trimming is good (phred 10 to 15)
    - Also read correction can make a difference!

More reading on trimming

- Williams et al. 2016 **Trimming of sequence reads alters RNA-Seq gene expression estimates**
- MacManes 2014 **On the optimal trimming of high-throughput mRNA sequence data**

More on read correction for transcriptome assemblies:

- Song and Florea 2015 Rcorrector: efficient and accurate error **correction** for Illumina **RNA**-seq reads
- MacManes and Eisen 2013 Improving transcriptome assembly through error correction of high-throughput sequence reads
- Heydari et al 2017 Evaluation of the impact of Illumina error correction tools on de novo genome assembly