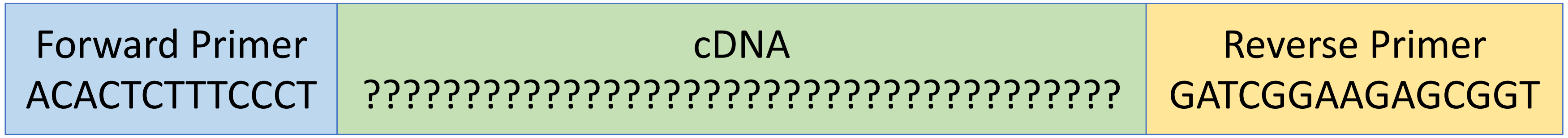
Trimming

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:

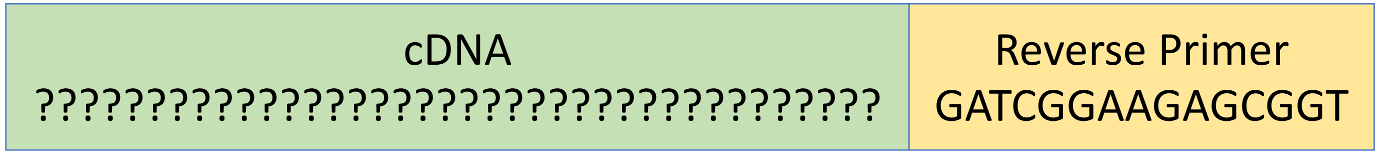


Forward Primer
ACACTCTTTCCCT

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

Reverse Primer
GATCGGAAGAGCGGT

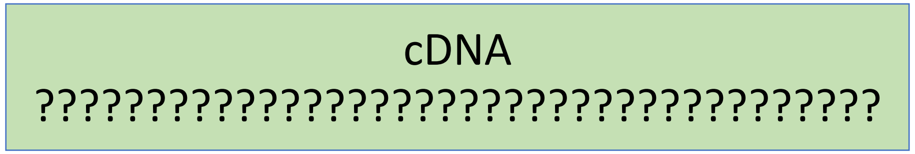
Read returned from sequencing facility:



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

Reverse Primer
GATCGGAAGAGCGGT

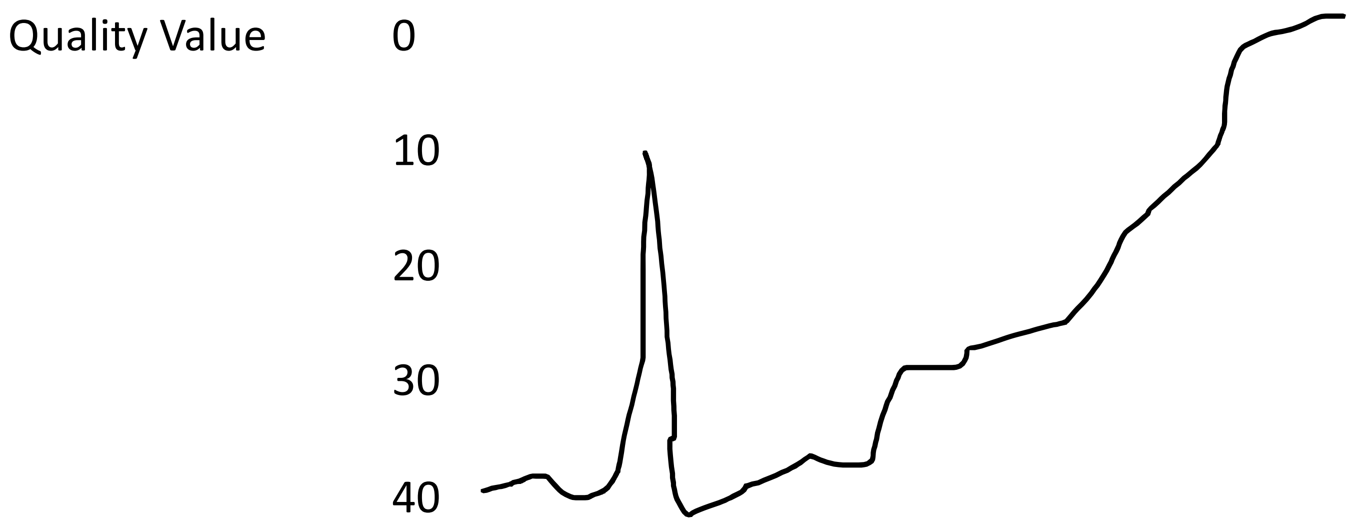
After Trimming:



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

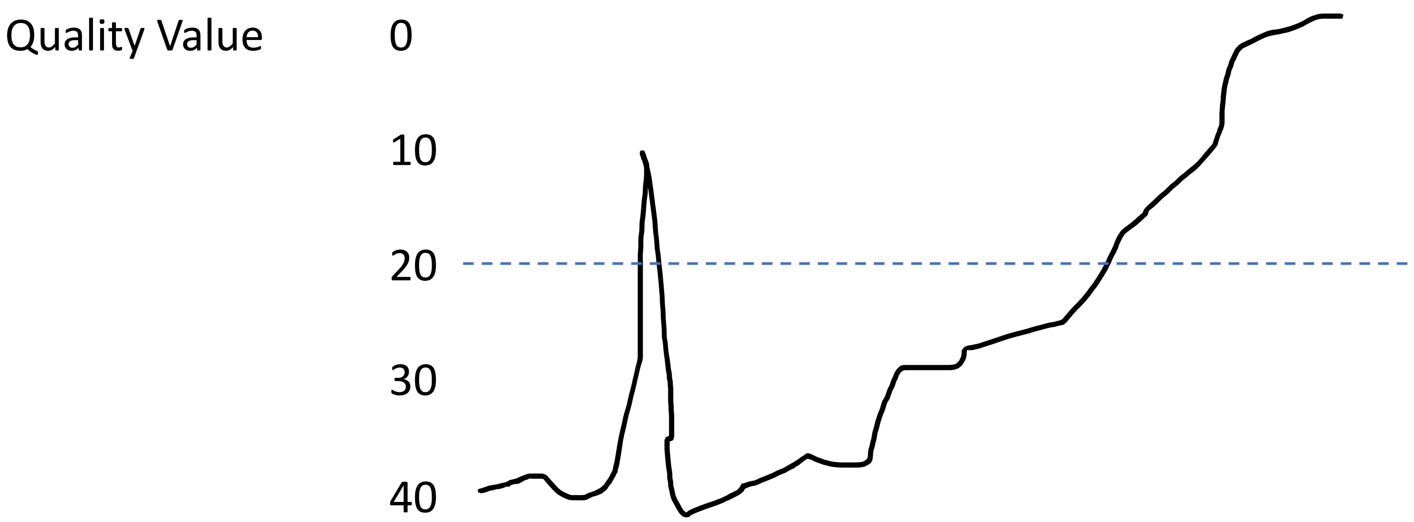
Quality Trimming

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



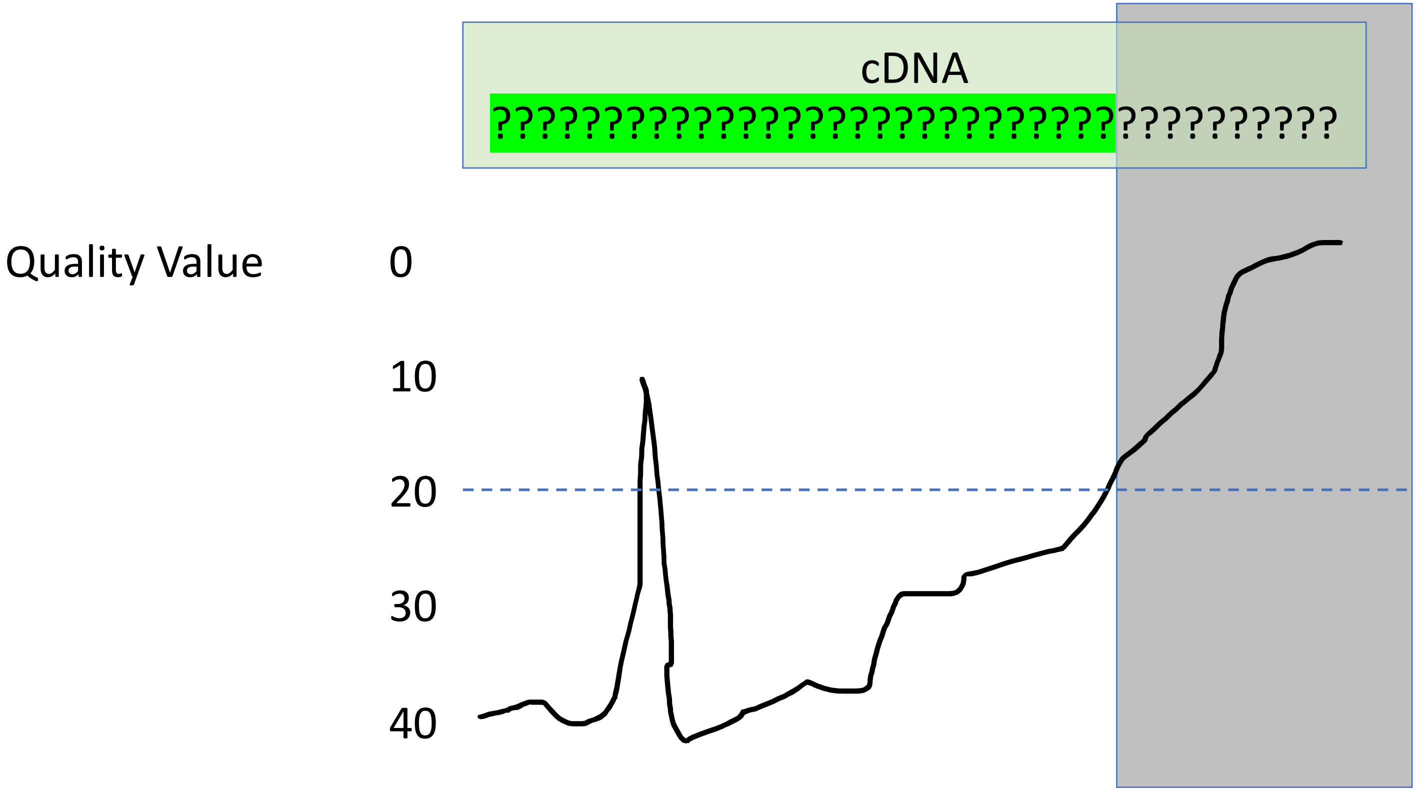
Quality Trimming

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



What is your cut off?

Quality Trimming



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

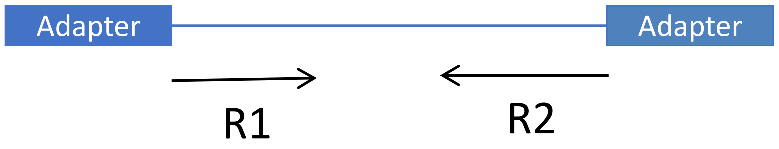
If you meet all these criteria, maybe not:

- 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)

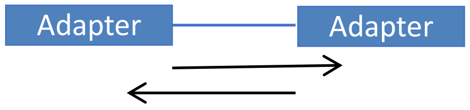
Trimmomatic

- 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 analysis, 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