Illumina Read Preprocessing

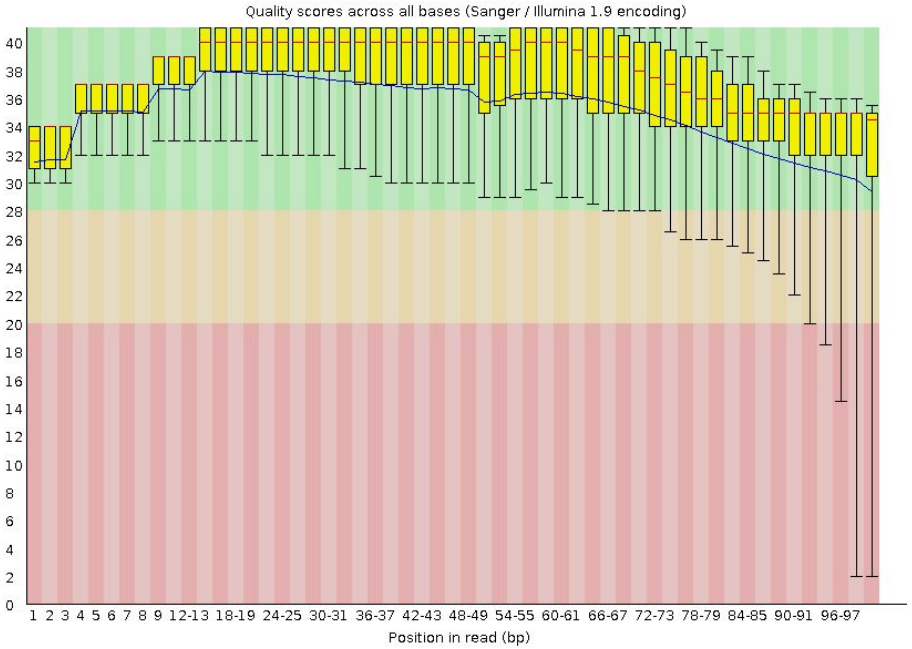
Monica Britton

Tuesday 22 August 2017

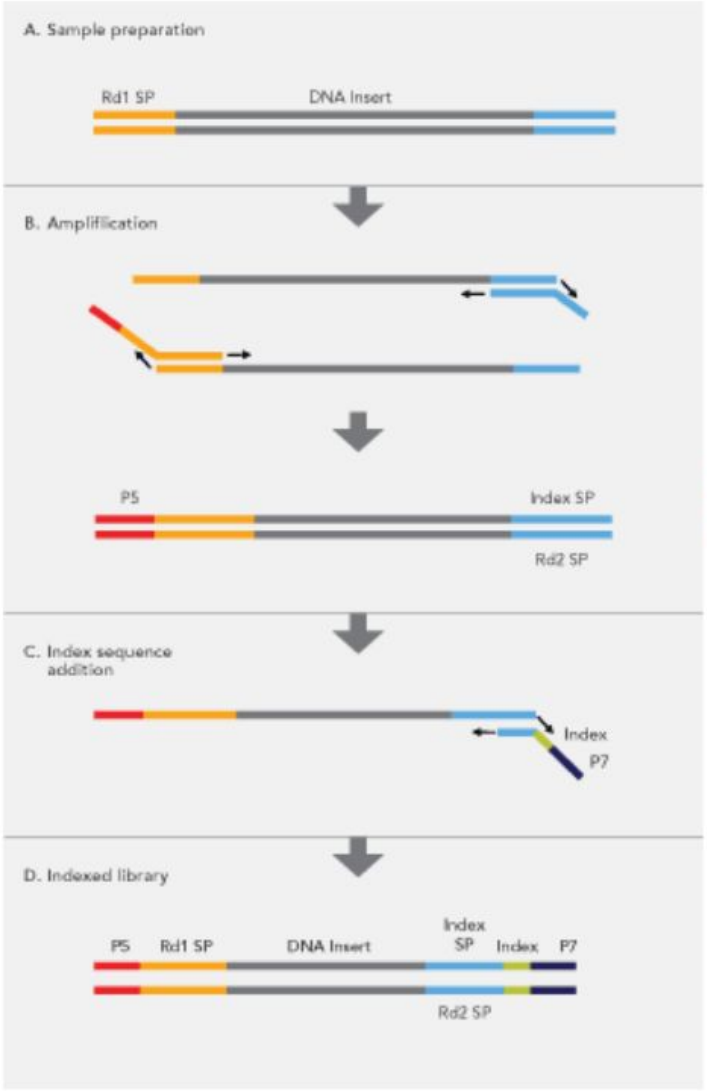
Read Preprocessing

Steps taken to improve the overall data quality.

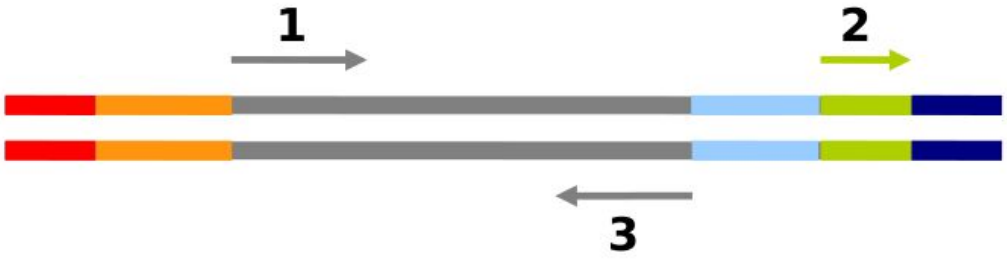
Typically includes removal of adapter sequences and low quality bases.



Adapter Contamination



Reads 1 and 3 are "forward" and "reverse" reads from your DNA-of-interest, and they are on opposite strands.

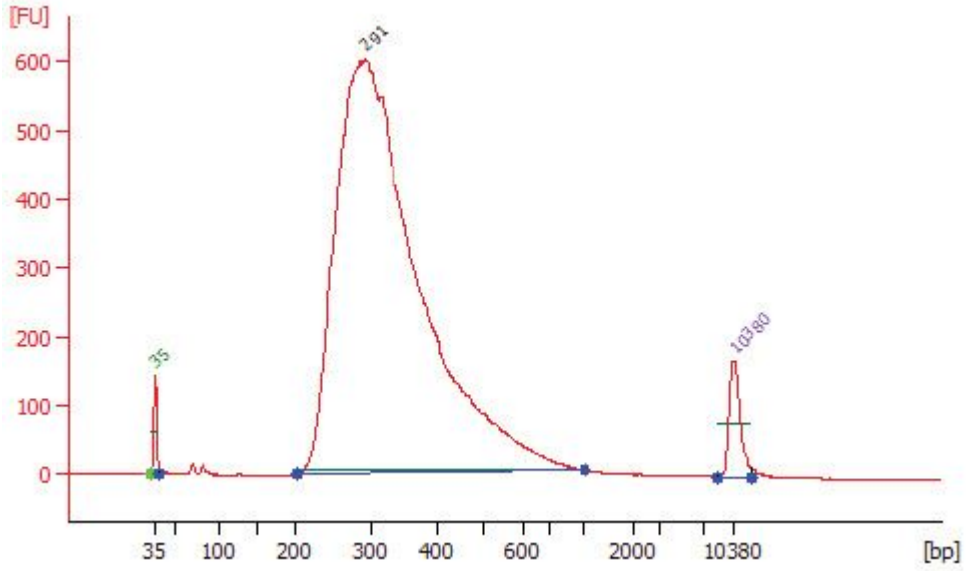


Read (2) is the "barcode," which identifies particular reads as belonging to a particular sample.

Adapter Contamination

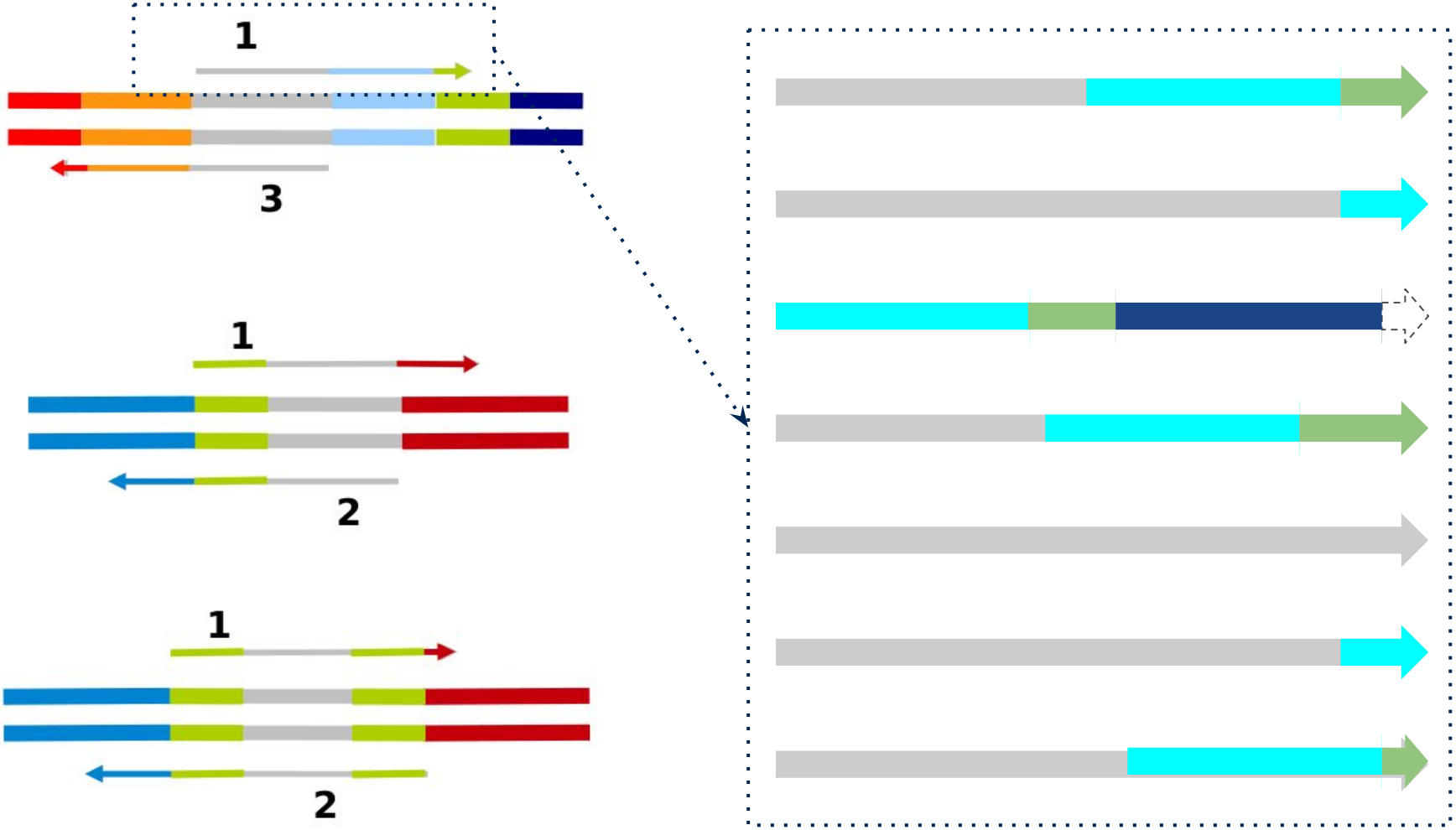
All sequencing libraries contain a distribution of molecule sizes. The smaller molecules (such as adapter dimers) will more easily find a place on the flowcell.

Any DNA insert that is shorter than the read length will generate some adapter contamination.



Adapter Contamination

Contamination is the result of the sequencer *reading through* a short read, into adapter sequence that *didn't come from your sample*!



Adapter Contamination

Where can you find adapter sequences?

- Illumina Adapter Sequences Document
- *Find them in your data*
- Contact the library kit manufacturer
- Google "github ucdavis-bioinformatics", look for Scythe, look for "*_adapters.fa"
- Check Seqanswers.com

Adapter Contamination

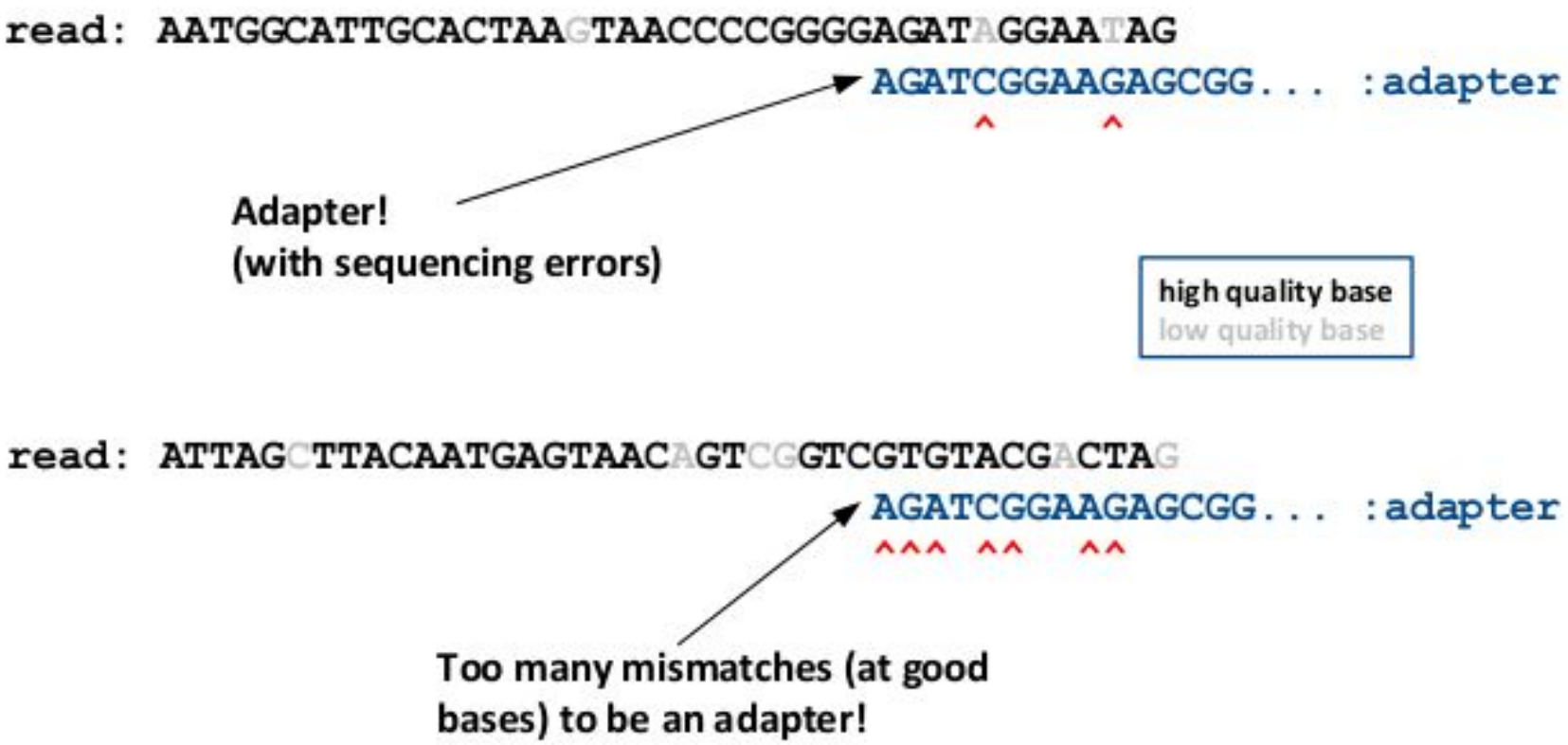
```
>TruSeq_forward_contam
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[8bp index]ATCTCGTATGCCGTCTTCTGCTTGAAAAA
>TruSeq_reverse_contam
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[8bp index]GTGGTCGCCGTATCATTAAAAA
```

```
>Nextera_forward_contam
CTGTCTCTTATACACATCTCCGAGCCCACGAGAC[8bp index]ATCTCGTATGCCGTCTTCTGCTTG
>Nextera_reverse_contam
CTGTCTCTTATACACATCTGACGCTGCCGACGA[8bp index]GTGTAGATCTCGGTGGTCGCCGTATCATT
```

```
>TruSeq_SmallRNA_forward_contam
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC[6bp adapter]ATCTCGTATGCCGTCTTCTGCTTG
>TruSeq_SmallRNA_reverse_contam
GATCGTCGGACTGTAGAACTCTGAACCTGTCG
```

Adapter Removal: Scythe

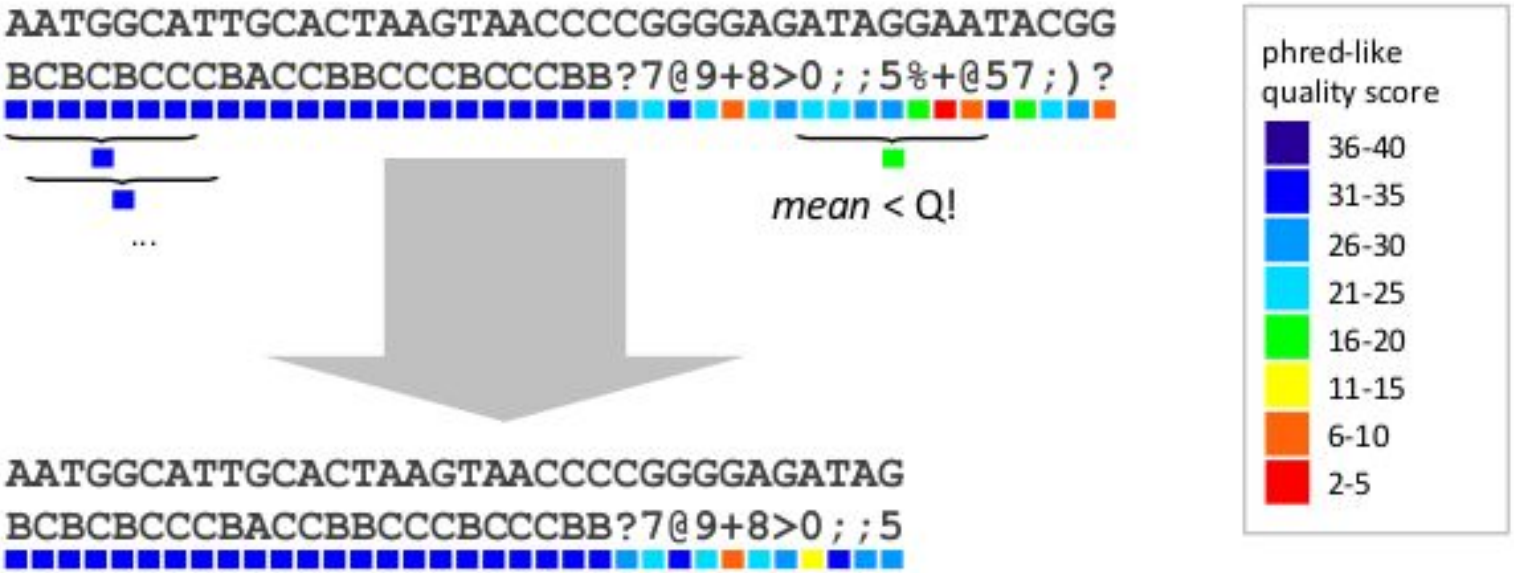
Scythe (*Vince Buffalo, Joe Fass*) is an adapter trimmer for Illumina reads that employs a Bayesian model that considers base qualities.



Low Quality Trimming: Sickle

Sickle (*Nik Joshi, Joe Fass*) is a sliding window trimmer for Illumina reads that tries to keep the longest high quality 5'-justified sequence from each read.

Moving from 5' to 3', windows of N bases are tested for average quality > Q. In the first window that fails, bases are trimmed starting with the first base with quality < Q.

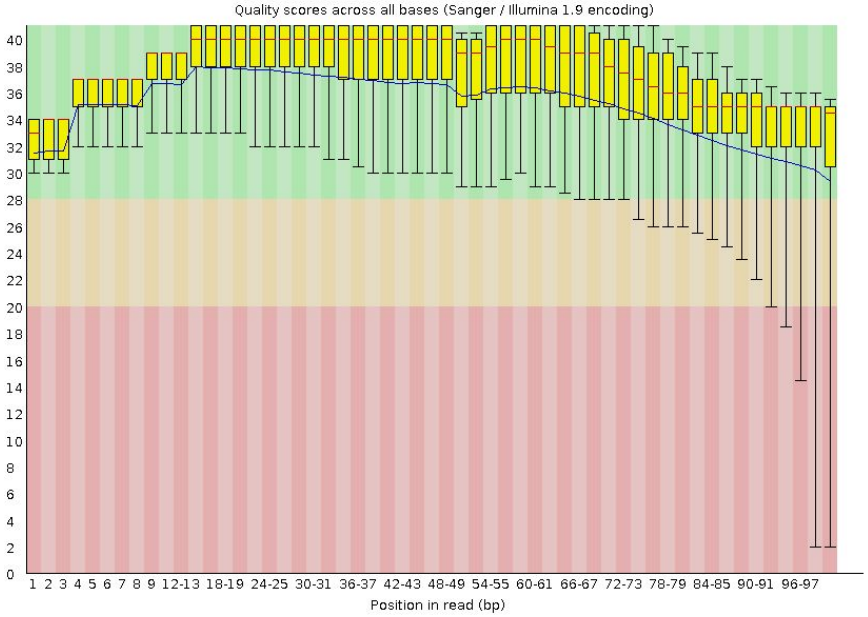


Read Preprocessing

Steps taken to improve the overall data quality.

Typically includes removal of adapter sequences and low quality bases.

Before



After