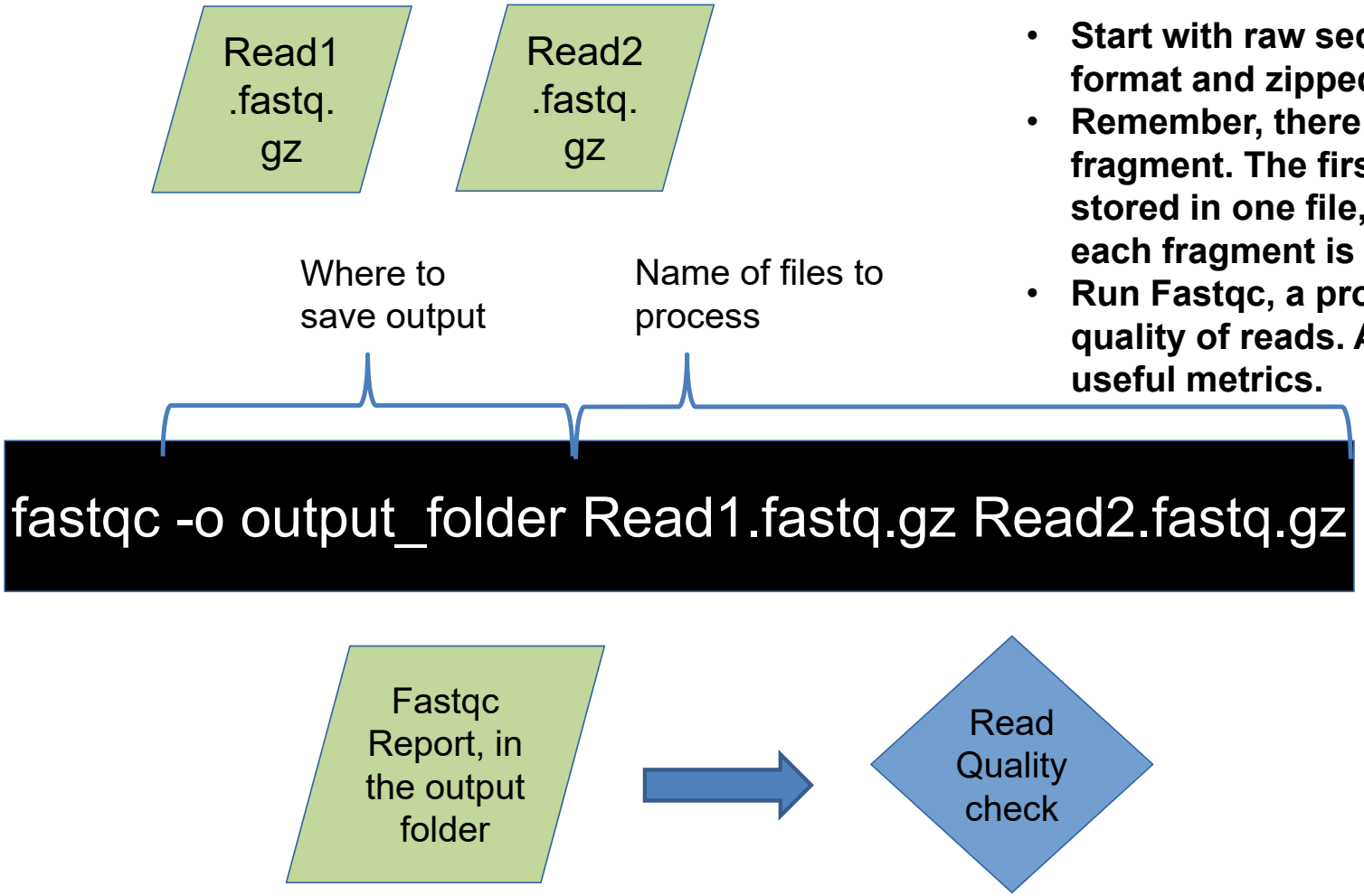


- **Start with raw sequencing data, in fastq format and zipped.**
- **Remember, there are two reads for each DNA fragment. The first read of each fragment is stored in one file, and the second read of each fragment is stored in another.**
- **Run Fastqc, a program that summarizes the quality of reads. Also outputs a number of useful metrics.**

FastQC Report

Summary

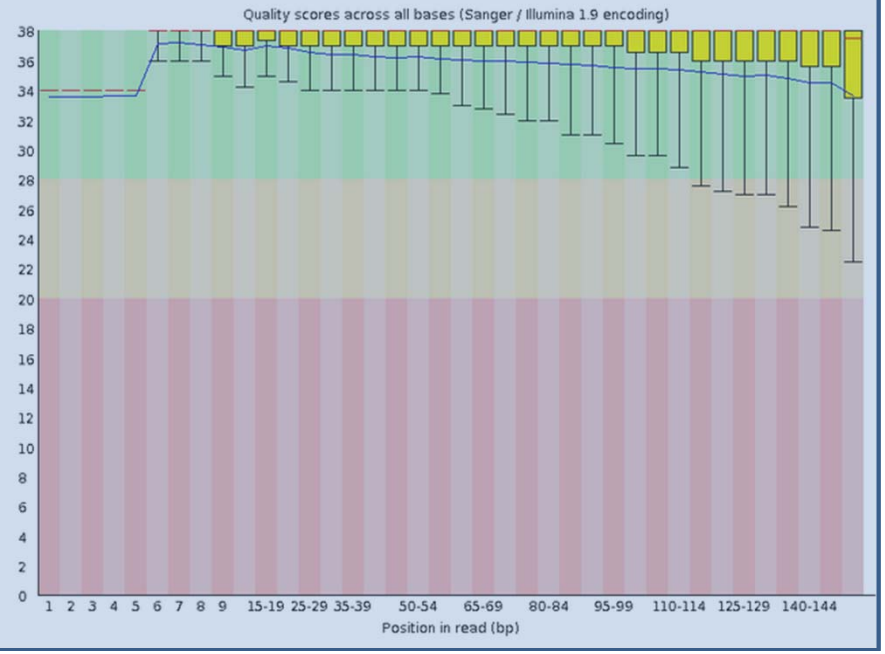
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Links to other reports

Basic Statistics

Measure	Value
Filename	293-412-5-12-16-16-B-R6_S20_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	849986
Sequences flagged as poor quality	0
Sequence length	151
%GC	46

Per base sequence quality



Fastqc Report

Preliminary information, number of sequences in file, average sequence length, etc.

Across all sequences, at each base position, what is the average quality score?

Quality>30 is good for most purposes.

The quality scores are high at each position of the read. We can proceed with the analysis.