NGS Data Analysis Course

Quality Control and Data Preprocessing

Marta Bleda, Javier Lopez, Ignacio Medina, David Montaner

2015-06-17

http://www.ngscourse.org

FastQ Format

- Standard Format for NGS data
- Conversion can be done from *sff*, *fasta* + *qual*, . . .
- Extension of the Fasta format
- Text-based formats (easy to use!)
- If not compressed, it can be huge

http://en.wikipedia.org/wiki/FASTQ_format

Quality measurements

Base-calling **error probabilities** are reported by sequencers.
Usually in **Phred** (quality) score.
Usually coded by ASCII characters

Phred score

$$Q = -10 log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

http://en.wikipedia.org/wiki/Phred_quality_score

NGS Data Preprocessing Steps

- File parsing: convert to **fastq** format form **sff**, **fasta** + **qual**
  . . .
- Split multiplex samples.
- Quality Control of the raw data.
- Filtering and trimming reads by quality.
- Adapter trimming
- Quality Control of the trimmed and filtered reads

Software

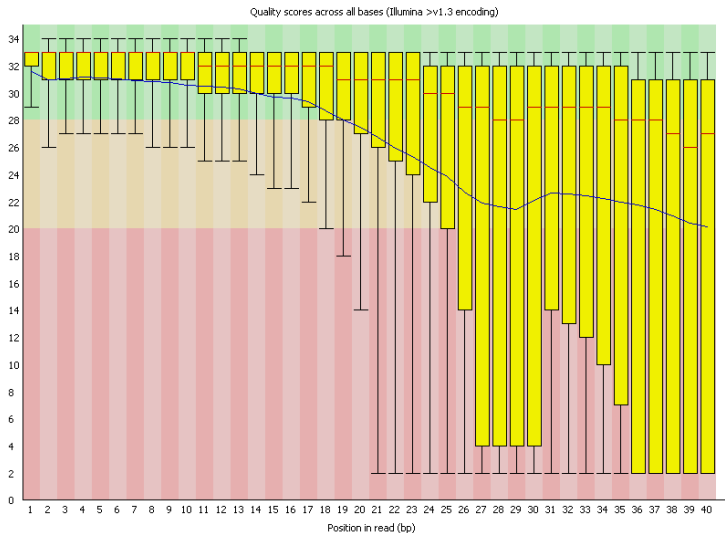- **FastQC**:
  - quality control
  - some filtering . . .

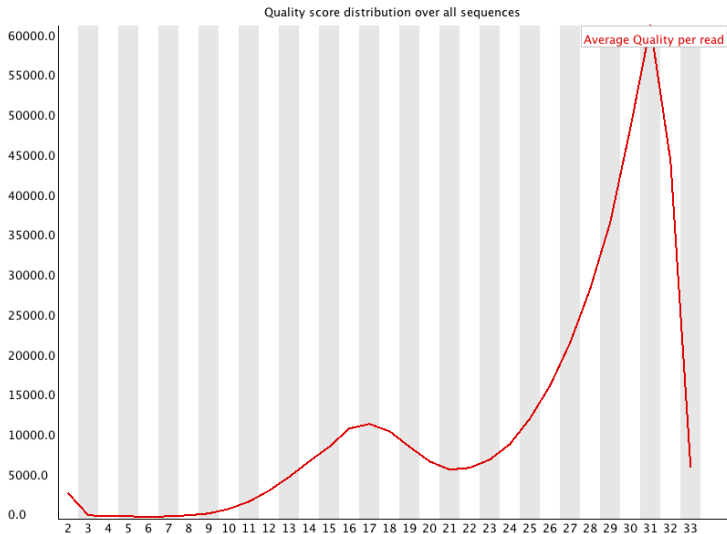www.bioinformatics.babraham.ac.uk/projects/fastqc

- **Cutadapt**:
  - adapter trimming
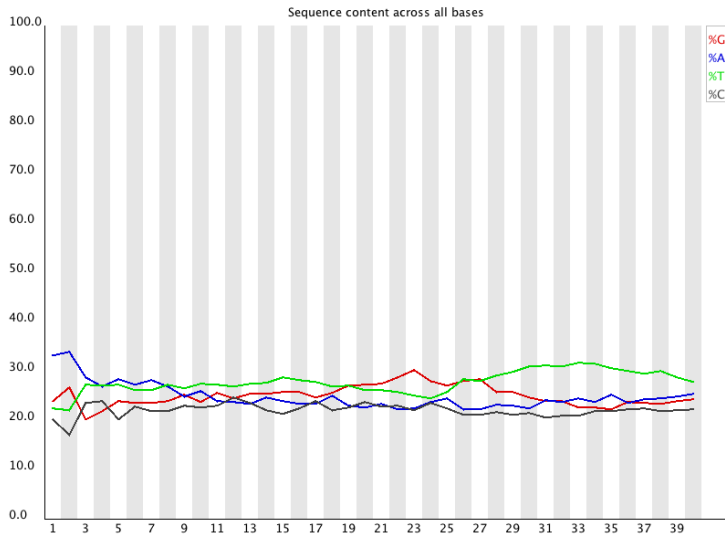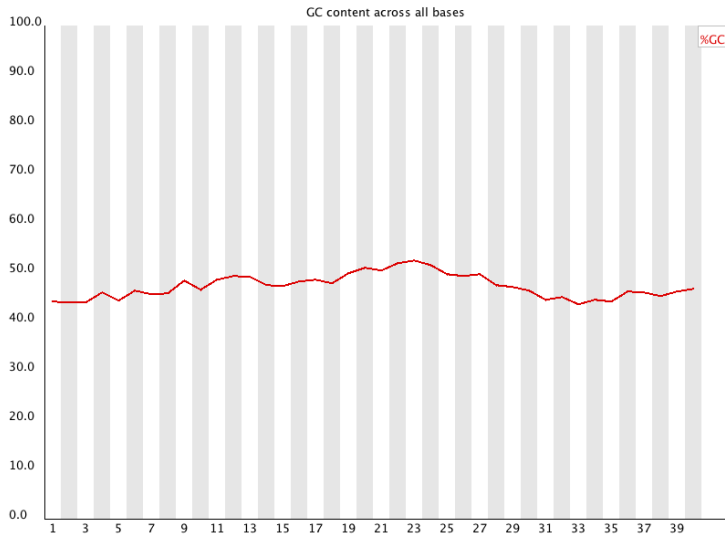 - filter reads by length (short, long)
  - filter reads by quality

http://code.google.com/p/cutadapt

Per Base Sequence Quality

Quality scores across all bases (Illumina >v1.3 encoding)

Per Sequence Quality

Quality score distribution over all sequences

Average Quality per read

Per Base Sequence Content

Sequence content across all bases

Per Base GC Content

GC content across all bases

Per Sequence Nucleotide Content

GC distribution over all sequences

Per Base N Content

N content across all bases
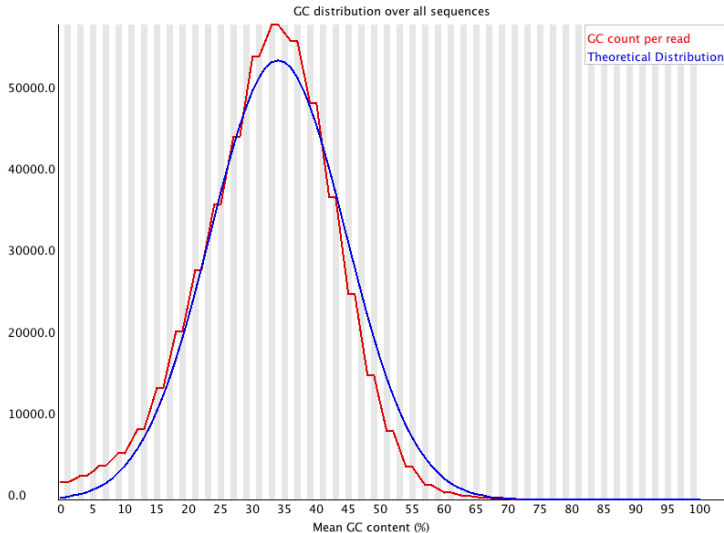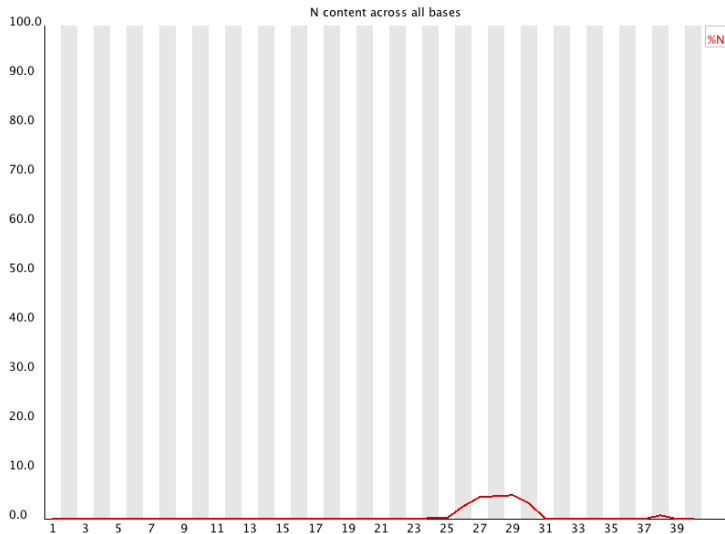
Sequence Length Distribution

Distribution of sequence lengths over all sequences
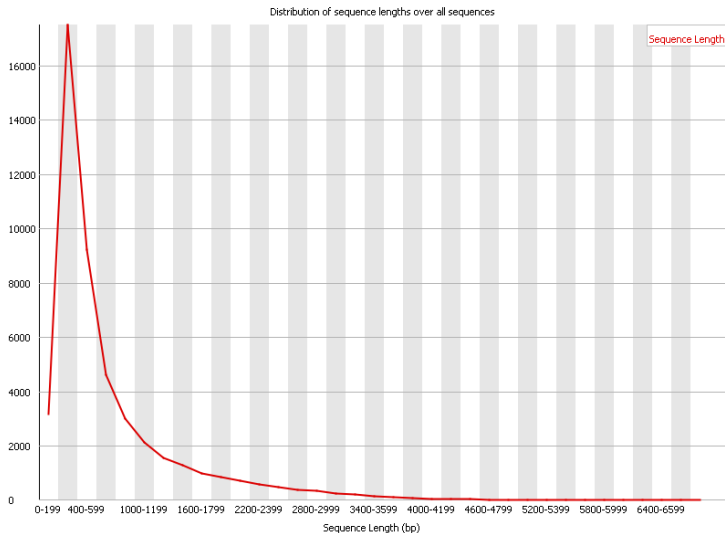
Duplicate Sequences Distribution

Over-represented Kmers
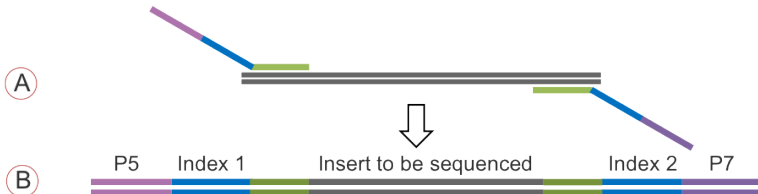

Relative enrichment over read length
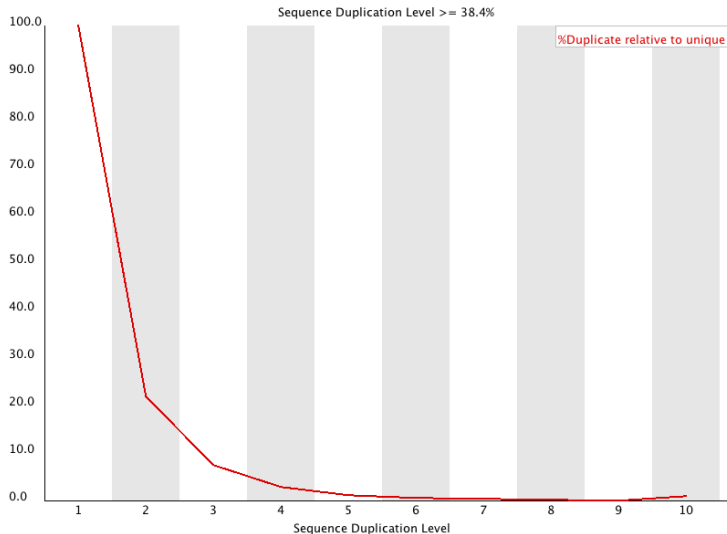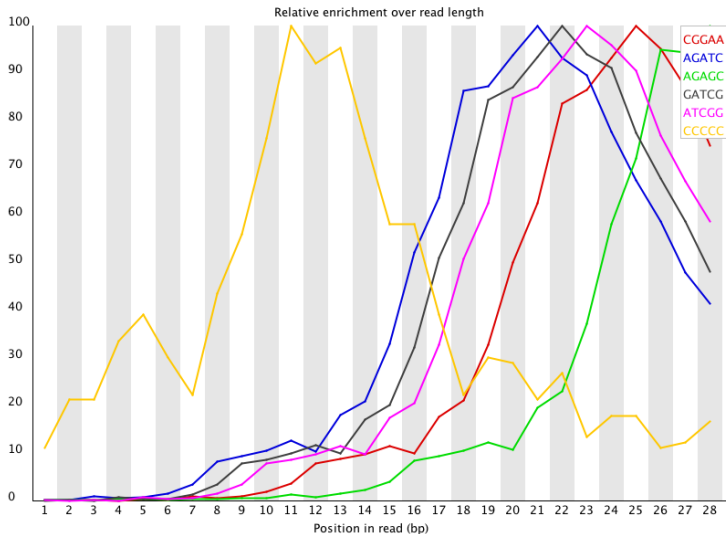
More FastQ examples and documentation

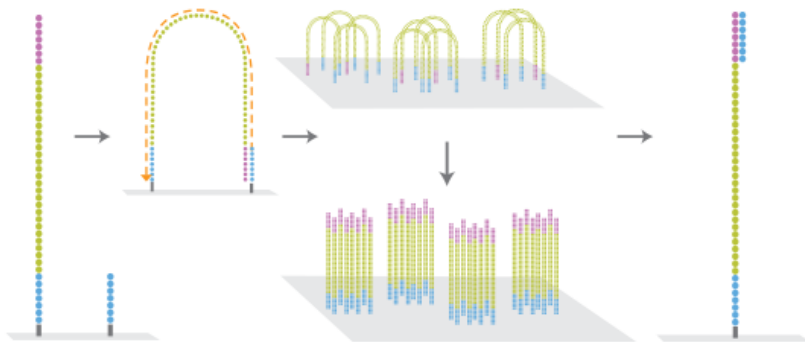. . . may be found at FastQ home page

- Example Reports

Sequencing process: PCR primers

One-step PCR Method

A — Target-specific PCR with indices and sequencing adaptors
B — Final amplicon ready to be sequenced

Sequencing process: PCR primers

NGS adaptors and Cutadapt