NGS data anlysis course

Quality control & Data Preprocessing

Ignacio Medina
David Montaner & Marta Bleda

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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#Definition

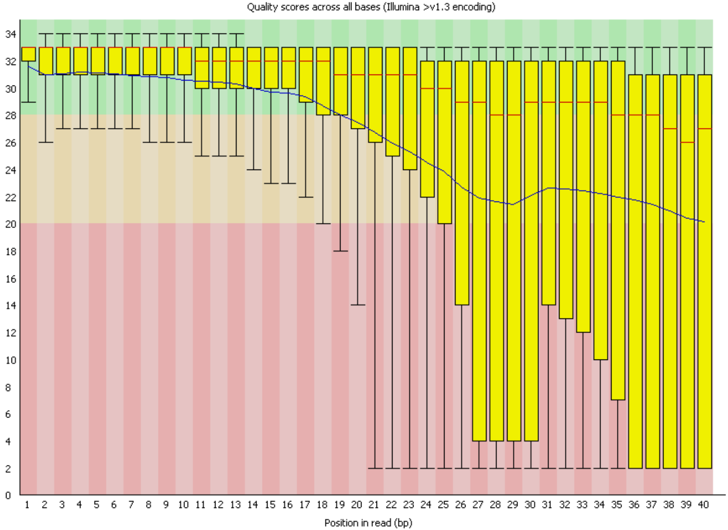
NGS Data Preprocessing Steps

- 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

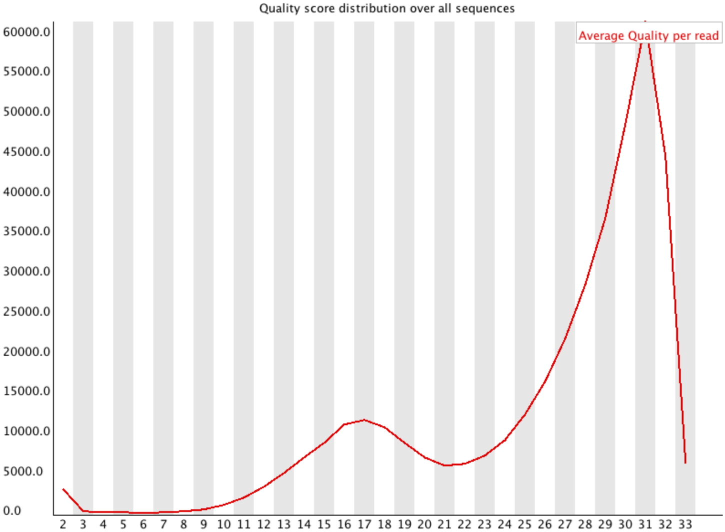
Software

- **FastQC**:
  - quality control
  - some filtering . . .
- **Cutadapt**:
  - adapter trimming
  - filter reads by length (short, long)
  - filter reads by quality

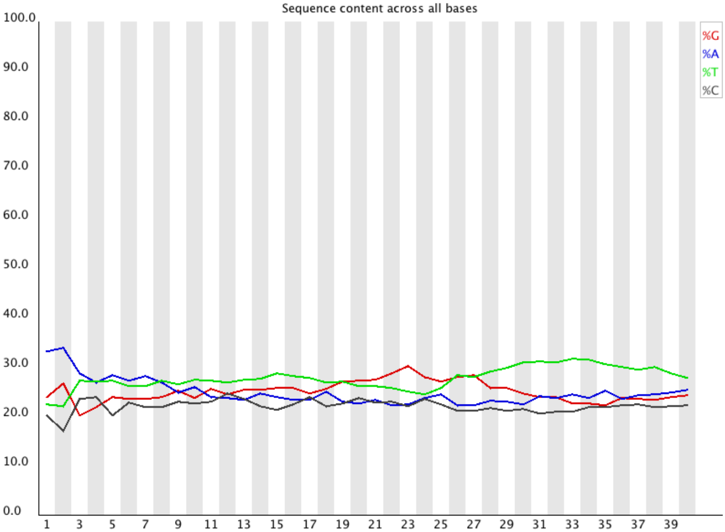
Per Base Sequence Quality



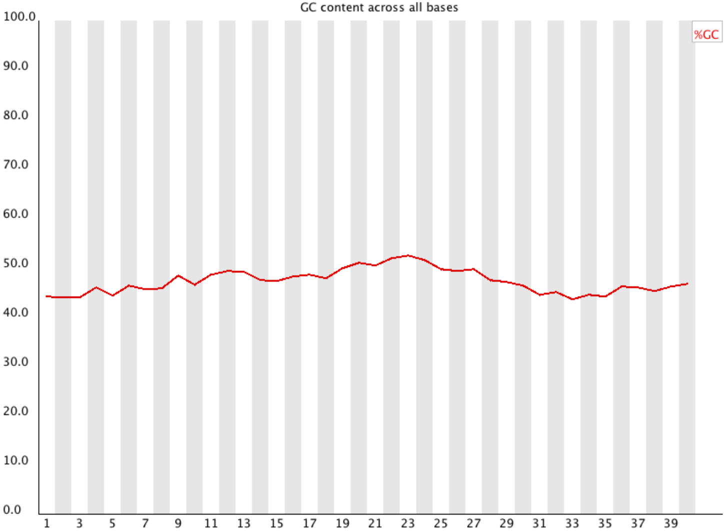
Per Sequence Quality



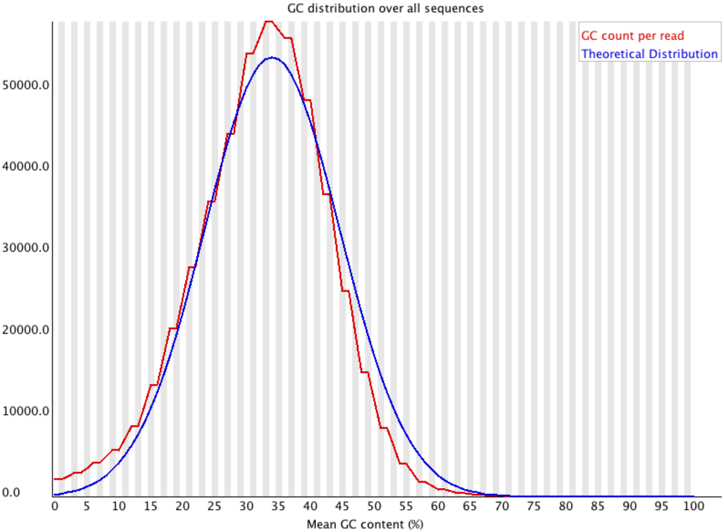
Per Base Sequence Content



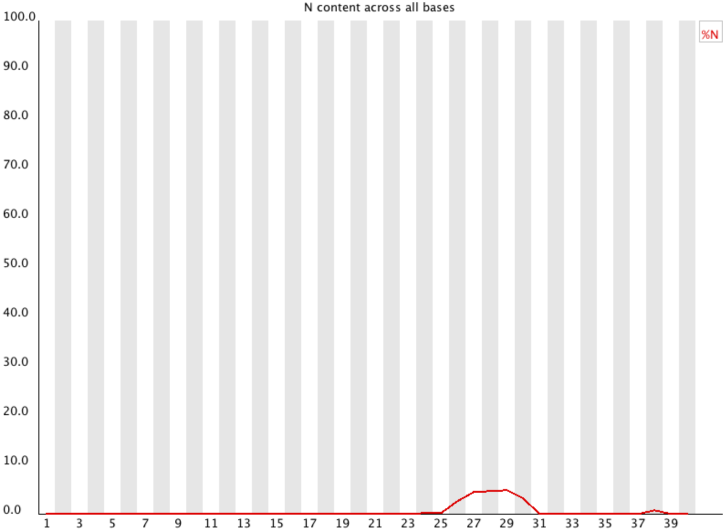
Per Base GC Content



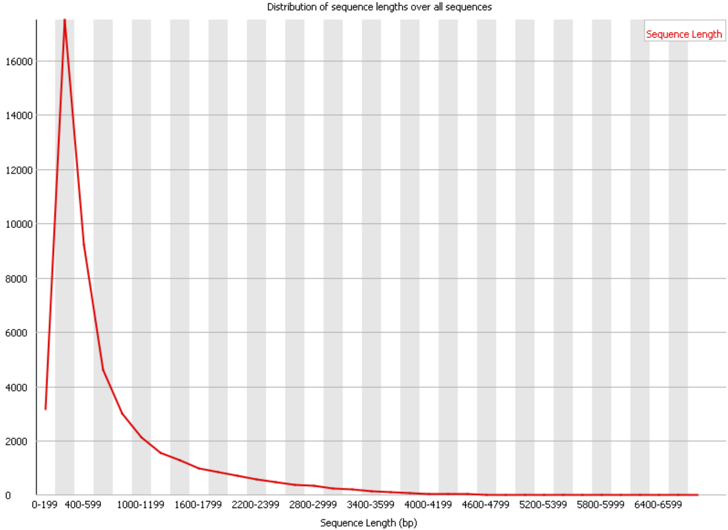
Per Sequence Nucleotide Content



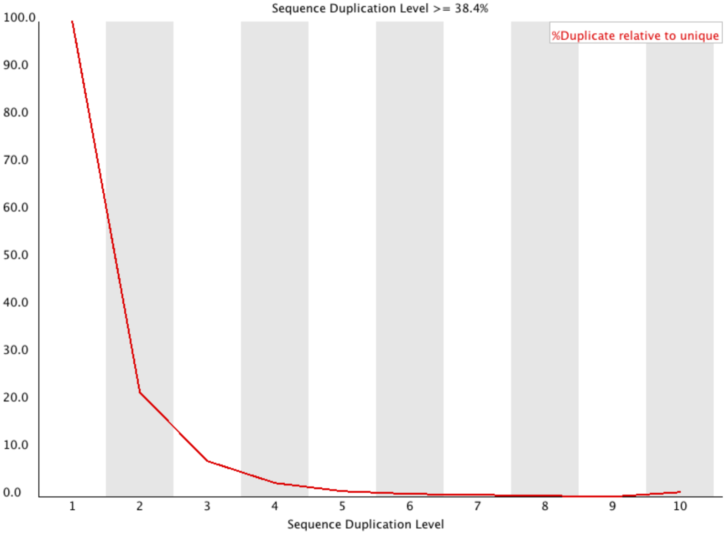
Per Base N Content



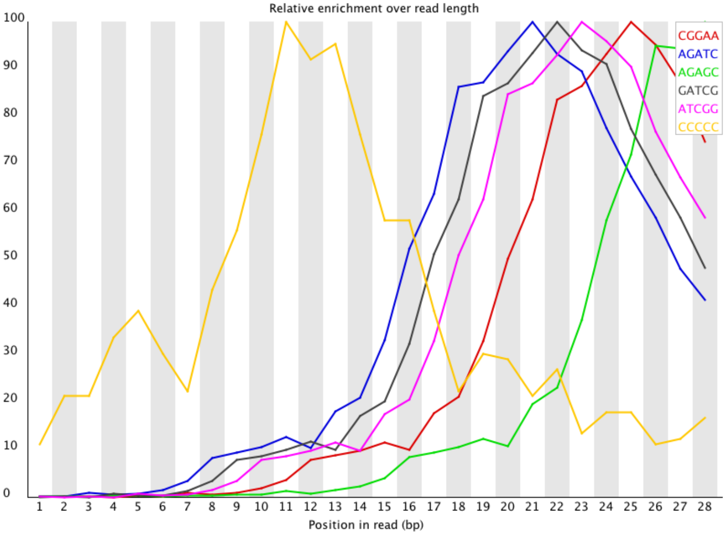
Sequence Length Distribution



Duplicate Sequences Distribution



Overrepresented Kmers


Relative enrichment over read length
Position in read (bp)
CGGAA
AGATC
AGAGC
GATCG
ATCGG
CCCCC