Data pre-processing

Arvind Sundaram

Oct 30, 2024

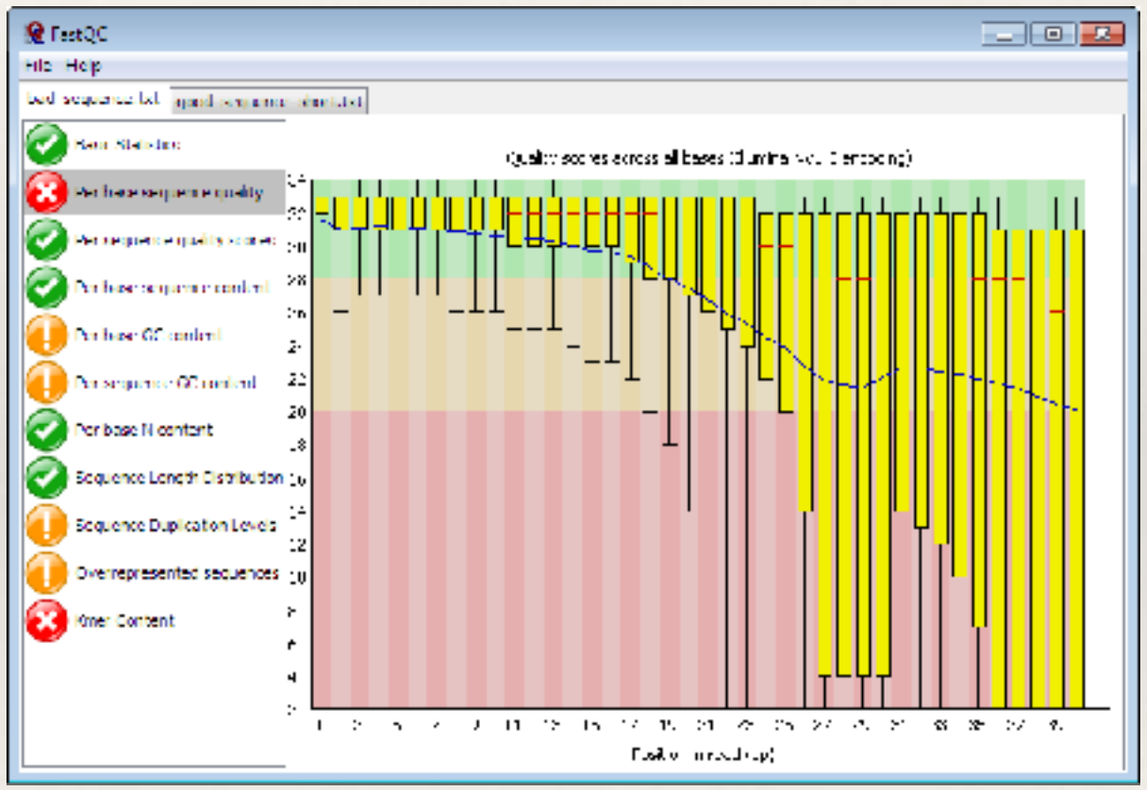
Norwegian Sequencing Centre
OUS, Ullevål, Oslo

Data pre-processing

- Quality control
- Why should we pre-process a sequence data
- Tools available
- Hands-on exercise

FastQC

- GUI, command line based
  - Import of data from BAM, SAM or FastQ files
  - Providing a quick overview to tell you in which areas there may be problems
  - Summary graphs and tables
  - HTML based permanent report
- requires Java



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

FastQC; MultiQC

- Video tutorial:
  - https://www.youtube.com/watch?v=bz93ReOv87Y
- Example reports:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- MultiQC
  - https://www.youtube.com/watch?v=BbScv9TcaMg
  - Not just for summarising FastQC reports but much more…..

http://multiqc.info/

FASTX-Toolkit

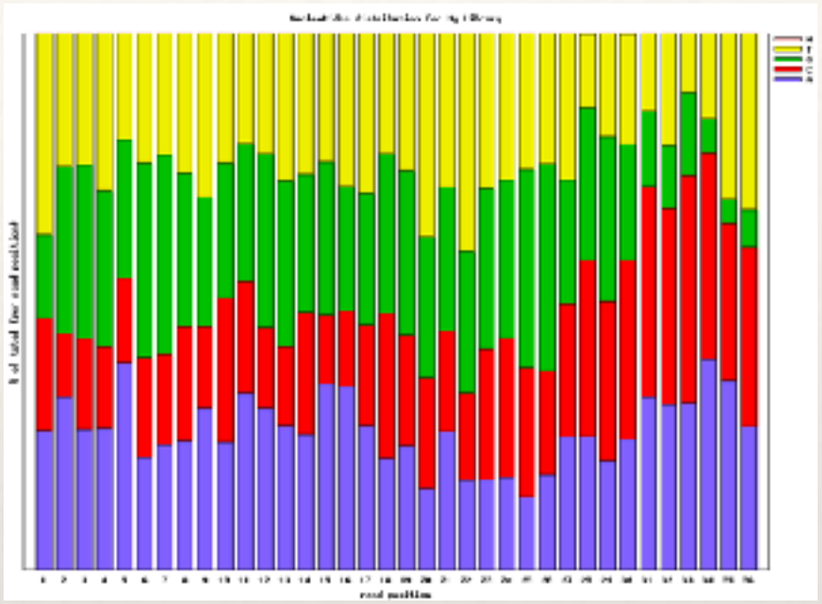
- Command line tool
  - Unix-based
- FastQ/A short-reads pre-processing tools

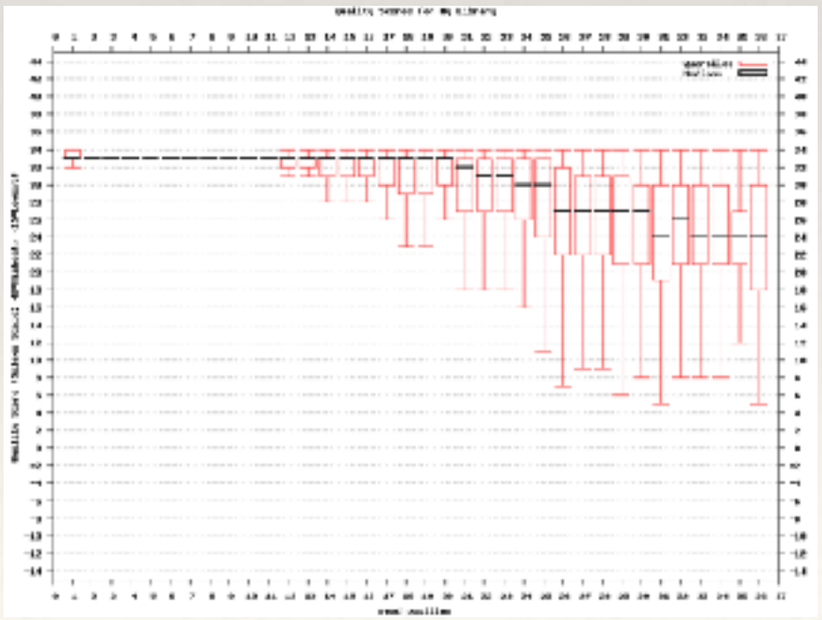
- FASTQ-to-FASTA
- FASTQ/A Quality Statistics
- FASTQ Quality chart
- FASTQ/A Nucleotide Distribution chart
- FASTQ/A Clipper
- FASTQ/A Renamer
- FASTQ/A Trimmer
- FASTQ/A Collapser
- FASTQ/A Artifacts Filter
- FASTQ Quality Filter
- FASTQ/A Reverse Complement
- FASTA Formatter
- FASTA nucleotides changer
- FASTA Clipping Histogram
- FASTX Barcode Splitter

http://www.usadellab.org/cms/index.php?page=trimmomatic

FASTX-Toolkit

- Command line usage:
  - http://hannonlab.cshl.edu/fastx_toolkit/commandline.html
- Remember to use '-Q 33' as a parameter





FastQ pre-processing

- Remove/Trim adapters
- Remove/Trim low quality reads
- Remove reads from spike-ins
  - PhiX for Illumina sequencing

- Trimmomatic*
- cutadapt
- PRINSEQ

- Make sure you understand what is going on under the hood

Do this if necessary

http://www.usadellab.org/cms/index.php?page=trimmomatic

Trimmomatic

- Quick start:
  - Paired End:
    - `trimmomatic PE -phred33 input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`
  - Single End:
    - `trimmomatic SE -phred33 input.fq.gz output.fq.gz ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`

Trimmomatic

- ILLUMINACLIP
  - Cut adapter and other Illumina-specific sequences from the read
  - Adapter file location
- 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

Hands-on

- Run FastQC on data
- Run your preferred FASTX toolkit tool
- Run trimmomatic on single end data
- Review the results
- Discuss

FastQC & FASTX toolkit Trimmomatic