UNAERP

Biotecnologia

Bioinformatics analysis tools

Wedson Gomes da S. Jr.

Prof. Dr. Mozart Marins

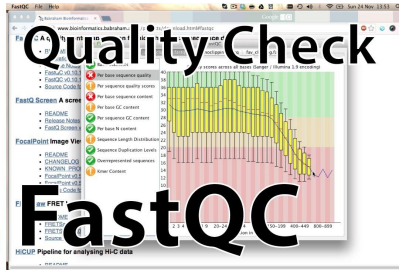
Pipeline

Analysis of data generated by the RNA-seq technique, that is, the sequencing of RNAs using performance.

- Evaluate the quality of sequences;
- Performing the sequence-based filtering step;
- Use mapping program;
- Use visualization program;

Pipeline

- FATSQC:http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- PRINSEQ: http://prinseq.sourceforge.net
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- IGV: http://www.broadinstitute.org/igv/download



FASTQC

Modern high throughput sequencers can generate tens of millions of sequences in a single run.

Before analysing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it.

Download data files

- SRR517961.fastq
  - wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR517/SRR517961/SRR517961.fastq.gz
  - gunzip SRR517961.fastq.gz
- Mus_musculus.GRCm38.dna.alt.fa.gz
  - wget ftp://ftp.ensembl.org/pub/release-94/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.alt.fa.gz
  - gunzip Mus_musculus.GRCm38.dna.alt.fa.gz

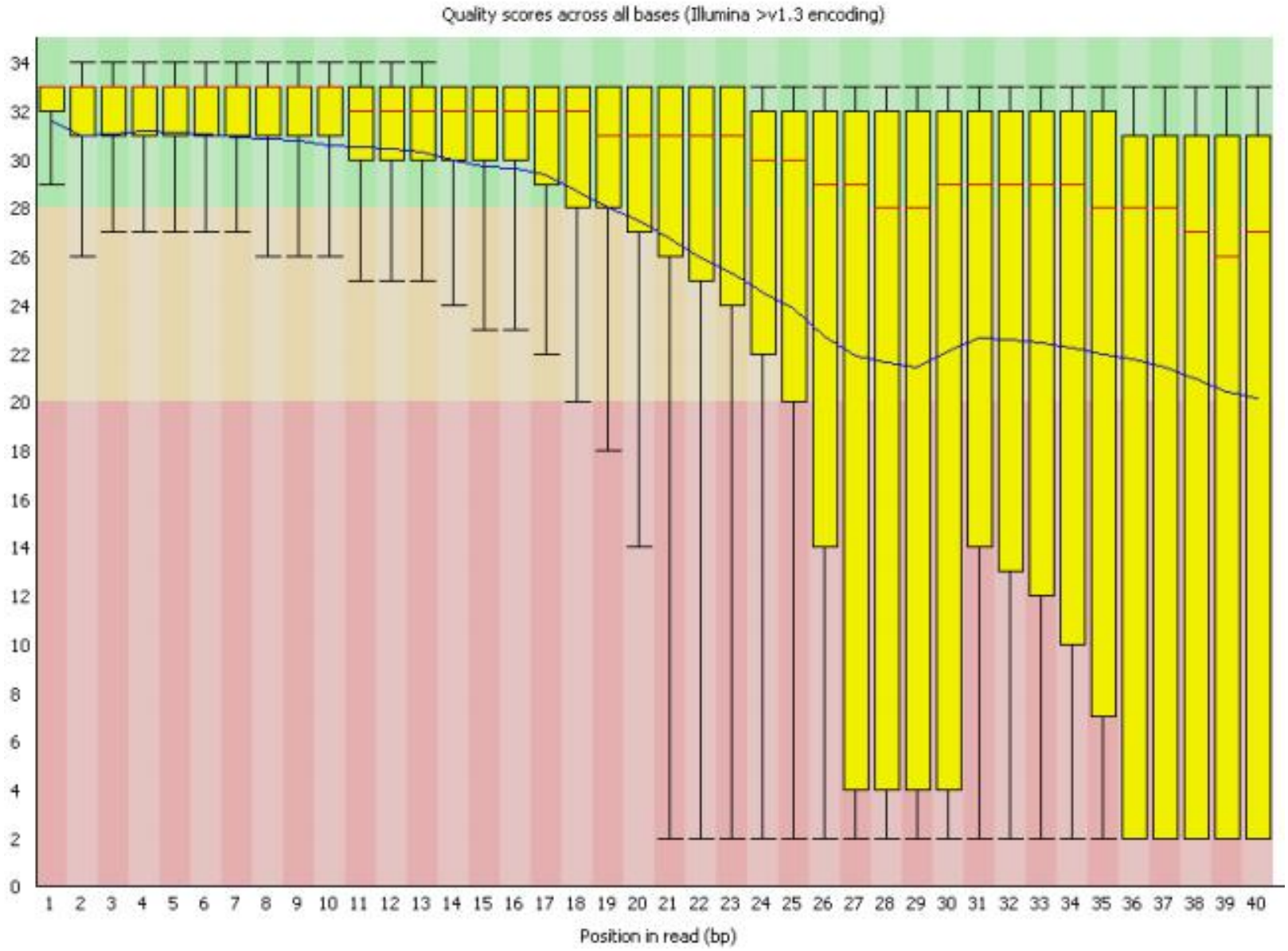
FASTQC

For each position a BoxWhisker type plot is drawn. The elements of the plot are as

follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

FASTQC


Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Position in read (bp)

Prinseq

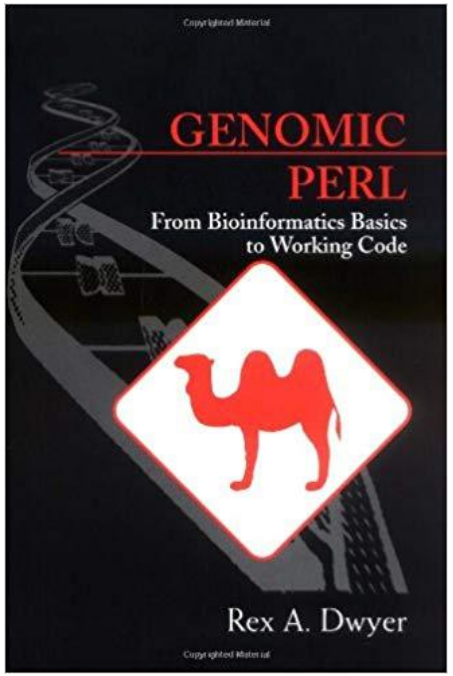
- Download http://prinseq.sourceforge.net/
- Rename to lower case and move to /etc

Prinseq - Trimming

In this step we will remove low quality sequences

perl prinseq-lite.pl -fastq SRR517961.fastq

-out_format 3 -trim_qual_right 20 -trim_qual_window 50 -trim_qual_step 25 -trim_qual_type mean -min_qual_mean 20 -min_len 50



Sequence mapping

- Bowtie is an ultrafast, memory-efficient short read aligner geared toward quickly aligning large sets of short DNA sequences (reads) to large genomes.
- It aligns 35-base-pair reads to the human genome at a rate of 25 million reads per hour on a typical workstation.

- command used to construct the index for further mapping of the reads in the mouse reference genome:
  - bowtie-build <nome do genoma>.fasta <nome do índice>
- To map the reads to the genome, run the command:
  - bowtie2 -x nome_do_arquivo.fasta.index -U nome_da_biblioteca.fastq -S nome_do_arquivo.sam -p 2 -a --no-unal --sensitive

Sequence mapping

- To view the mapping, you must sort and create an index of the .sam file generated in the previous step. To do this run:
  - samtools sort <nome da saida>.bam <nome da saida>.sorted
  - samtools index <nome da saida>.sorted.bam <nome da saida>.sorted.idx

Data visualization

- Use the IGV program to view the mapping.
  - IGV: http://www.broadinstitute.org/igv/download

- To open:
  - cd IGV_1.5.14; chmod 755 ./igv_linux.sh; ./igv_linux.sh