

Análise da Qualidade do Sequenciamento

Pablo Rodrigo Sanches

Departamento de Genética – FMRP/USP

psanches@usp.br

Roteiro de análise

1. FASTQC – 1ª rodada (análise de qualidade “pré-trimagem”)
2. Trimmomatic (“trimagem” de adaptadores e regiões de má qualidade)
3. FASTQC – 2ª rodada (análise de qualidade “pós-trimagem”)
4. Comparação dos resultados obtidos nas 1ª e 2ª rodadas
5. Desenvolvimento da tabela de resultados

Arquivos Fastq (Dados Brutos)

The screenshot shows the Galaxy web interface. The main panel displays a list of genomic data entries, including their IDs and coordinates. The right sidebar shows a history of datasets, including '48: FastQC on data 34: R awData' and '47: FastQC on data 34: Webpage'.

Tools

- Upload Data
- Get Data
- Collection Operations
- GENERAL TEXT TOOLS
- Text Manipulation
- Filter and Sort
- Join, Submatch and Group
- Datamash
- GENOMIC FILE MANIPULATION
- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over
- COMMON GENOMICS TOOLS
- Interactive tools
- Operate on Genomic Intervals
- Fetch Sequences/Alignments
- GENOMICS ANALYSIS
- Assembly
- Annotation
- Manning

History

- 48: FastQC on data 34: R awData
- 47: FastQC on data 34: Webpage
- 46: FastQC on data 33: R awData
- 45: FastQC on data 33: Webpage
- 44: 12h_III_R2.fastq.gz
- 43: 12h_III_R1.fastq.gz
- 42: 12h_II_R2.fastq.gz
- 41: 12h_II_R1.fastq.gz
- 40: 3h_III_R2.fastq.gz
- 39: 3h_III_R1.fastq.gz
- 38: 3h_II_R2.fastq.gz
- 37: 3h_II_R1.fastq.gz
- 36: 0h_III_R2.fastq.gz
- 35: 0h_III_R1.fastq.gz
- 34: 0h_II_R2.fastq.gz
- 33: 0h_II_R1.fastq.gz

FASTQC - Via Galaxy – 1ª rodada

Galaxy

usegalaxy.org

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

Using 8%

Tools

fastqc

Upload Data

Show Sections

fastp - fast all-in-one preprocessing for FASTQ files

FastQC Read Quality reports

Bio-TraDis reads to counts

Combine FASTA and QUAL into FASTQ

Manipulate FASTQ reads on various attributes

fastp - fast all-in-one preprocessing for FASTQ files

WORKFLOWS

All workflows

FASTQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Favorite Versions Options

Short read data from your current history

44: 12h_III_R2.fastq.gz
43: 12h_III_R1.fastq.gz
42: 12h_II_R2.fastq.gz
41: 12h_II_R1.fastq.gz
40: 3h_III_R2.fastq.gz
39: 3h_III_R1.fastq.gz

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

No tabular dataset available.

list of adapters adapter sequences which will be explicitly searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads > 50bp

No

Using this option will cause fastqc to crash and burn if you use it on really long reads. and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comparable statistics from datasets with somewhat variable read lengths. (--min_length)

length of Kmer to look for

7

note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

Email notification

No

History

search datasets

UDA

12 shown

20.9 GB

44: 12h_III_R2.fastq.gz

43: 12h_III_R1.fastq.gz

42: 12h_II_R2.fastq.gz

41: 12h_II_R1.fastq.gz

40: 3h_III_R2.fastq.gz

39: 3h_III_R1.fastq.gz

38: 3h_II_R2.fastq.gz

37: 3h_II_R1.fastq.gz

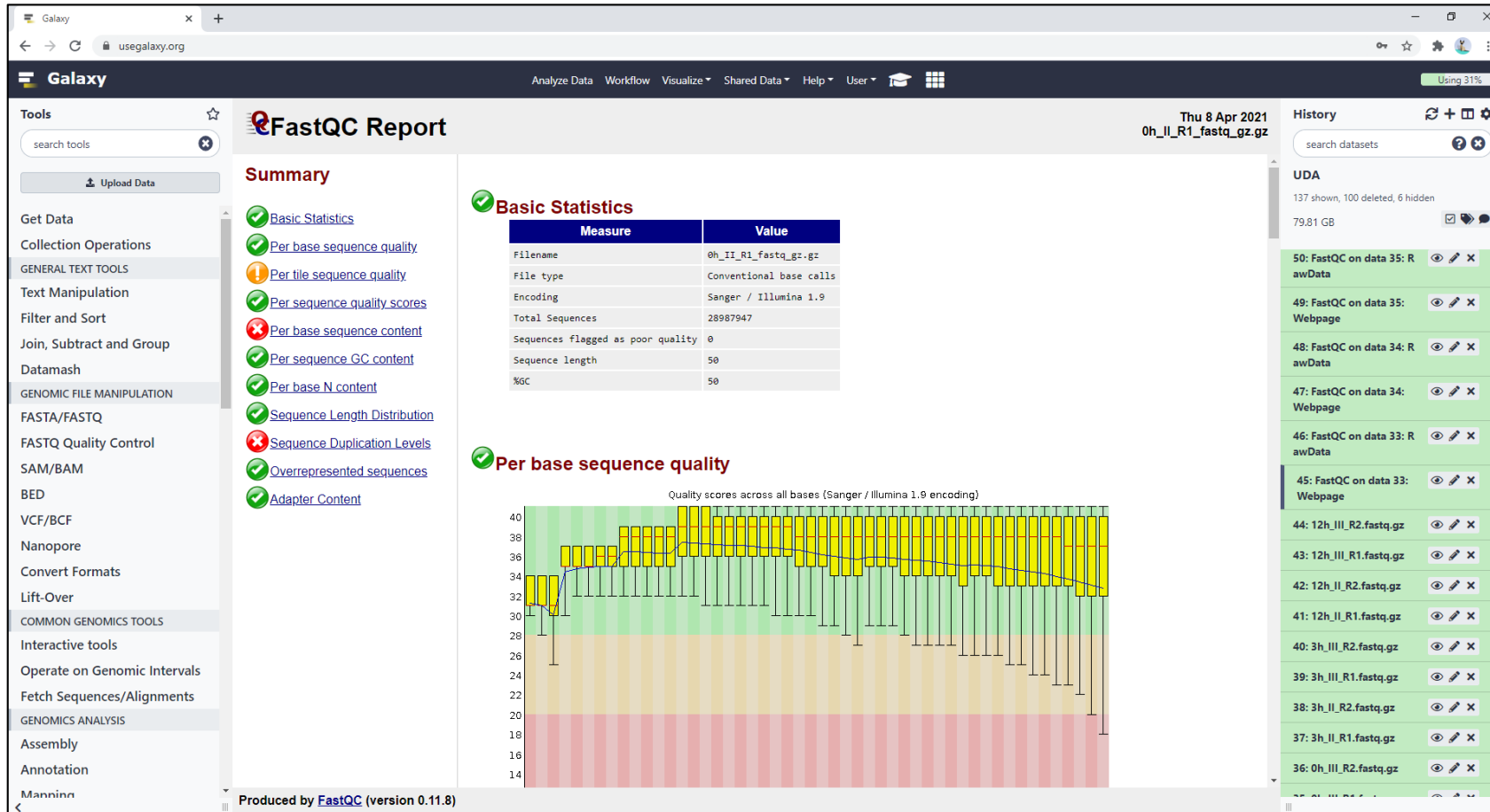
36: 0h_III_R2.fastq.gz

35: 0h_III_R1.fastq.gz

34: 0h_II_R2.fastq.gz

33: 0h_II_R1.fastq.gz

Exemplo de Resultado FASTQC



Trimmomatic – Via Galaxy

The screenshot displays the Galaxy web interface for configuring the Trimmomatic tool. The browser address bar shows 'usegalaxy.org'. The main header includes navigation links for 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', and 'User', along with a 'Using 31%' indicator.

Tools Panel (Left): Lists available tools, including 'trimmomatic' (selected), 'fastp', and 'Shovill'. It also includes 'Upload Data' and 'Show Sections' buttons.

Trimmomatic Configuration (Center):

- Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.38.0)**
- Single-end or paired-end reads?**: Paired-end (two separate input files)
- Input FASTQ file (R1/first of pair)**: 33: 0h_IL_R1.fastq.gz
- Input FASTQ file (R2/second of pair)**: 34: 0h_IL_R2.fastq.gz
- Perform initial ILLUMINA CLIP step?**: Yes
- Select standard adapter sequences or provide custom?**: Standard
- Adapter sequences to use**: TruSeq3 (paired-ended, for MiSeq and HiSeq)
- Maximum mismatch count which will still allow a full match to be performed**: 2
- How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment**: 30
- How accurate the match between any adapter etc. sequence must be against a read**: 10
- Minimum length of adapter that needs to be detected (PE specific/palindrome mode)**: 8

History Panel (Right): Shows a list of previous jobs, including several Trimmomatic runs on various datasets (e.g., '75: Trimmomatic on 0h_IL_R1.fastq.gz (R1 paired)').

Trimmomatic – Via Galaxy (cont.)

The screenshot displays the Galaxy web interface for configuring the Trimmomatic tool. The browser address bar shows `usegalaxy.org`. The top navigation bar includes links for `Analyze Data`, `Workflow`, `Visualize`, `Shared Data`, `Help`, `User`, and a user profile icon. A `Using 31%` indicator is visible in the top right.

Tools Sidebar:

- trimmomatic** (selected)
- Upload Data
- Show Sections
- fastp** - fast all-in-one preprocessing for FASTQ files
- Trimmomatic** flexible read trimming tool for Illumina NGS data
- Shovill** Faster SPAdes assembly of Illumina reads
- fastp** - fast all-in-one preprocessing for FASTQ files
- WORKFLOWS**
All workflows

Main Configuration Area:

- Always keep both reads (PE specific/palindrome mode)?**
 Yes
See help below
- Trimmomatic Operation**
1: Trimmomatic Operation
- Select Trimmomatic operation to perform**
Sliding window trimming (SLIDINGWINDOW)
- Number of bases to average across**
4
- Average quality required**
20
- + Insert Trimmomatic Operation**
- Output trimlog file?**
 Yes
(-trimlog)
- Output trimmomatic log messages?**
 Yes
these are the messages written to stderr (eg. for use in MultiQC)
- Job Resource Parameters**
Use default job resource parameters
- Email notification**
 No
Send an email notification when the job completes.
- Execute** button
- What it does**
Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.
This tool allows the following trimming steps to be performed:

History Sidebar:

- search datasets
- UDA**
137 shown, 100 deleted, 6 hidden
79.81 GB
- d)
- 75: Trimmomatic on 0h_III_R1.fastq.gz (R1 paired)
- 74: Trimmomatic on data 34 and data 33 (log file)
- 73: Trimmomatic on data 34 and data 33 (trimlog file)
- 72: Trimmomatic on 0h_II_R2.fastq.gz (R2 unpaired)
- 71: Trimmomatic on 0h_II_R1.fastq.gz (R1 unpaired)
- 70: Trimmomatic on 0h_II_R2.fastq.gz (R2 paired)
- 69: Trimmomatic on 0h_II_R1.fastq.gz (R1 paired)
- 68: FastQC on data 44: RawData
- 67: FastQC on data 44: Webpage
- 66: FastQC on data 43: R

Exemplo de Resultado Trimmomatic

```
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/jetstream/scratch0/main/jobs/34439361/_job_tmp
TrimmomaticPE: Started with arguments:
-threads 10 fastq_r1.fastqsanger.gz fastq_r2.fastqsanger.gz fastq_out_r1_paired.fastqsanger.gz fastq_out_r1_unpaired.fastqsanger.gz fastq_out_r2_paired.fastqsanger.gz
fastq_out_r2_unpaired.fastqsanger.gz ILLUMINACLIP:/cvmfs/main.galaxyproject.org/deps/_conda/envs/_trimmomatic@0.38/share/trimmomatic-0.38-1/adapters/TruSeq3-PE.fa:2:30:10:8:true
SLIDINGWINDOW:4:20 -trimlog trimlog
Using PrefixPair: 'TACACTCTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Read Pairs: 28987947 Both Surviving: 26800883 (92.46%) Forward Only Surviving: 1836438 (6.34%) Reverse Only Surviving: 274016 (0.95%) Dropped: 76610 (0.26%)
TrimmomaticPE: Completed successfully
```

Nro. Reads
totais

Nro. de pares de
Reads após trimagem

Reads apenas em R1
após trimagem

Reads apenas em R2
após trimagem

Reads retirados
após trimagem

FASTQC - Via Galaxy – 2ª rodada

Galaxy

usegalaxy.org

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

Using 31%

Tools

fastqc

Upload Data

Show Sections

fastp - fast all-in-one preprocessing for FASTQ files

FastQC Read Quality reports

fastp - fast all-in-one preprocessing for FASTQ files

Combine FASTA and QUAL into FASTQ

Bio-TraDis reads to counts

Manipulate FASTQ reads on various attributes

WORKFLOWS

All workflows

FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Favorite Versions Options

Short read data from your current history

77: Trimmomatic on 0h_III_R1.fastq.gz (R1 unpaired)
76: Trimmomatic on 0h_III_R2.fastq.gz (R2 paired)
75: Trimmomatic on 0h_III_R1.fastq.gz (R1 paired)
72: Trimmomatic on 0h_II_R2.fastq.gz (R2 unpaired)
71: Trimmomatic on 0h_II_R1.fastq.gz (R1 unpaired)
70: Trimmomatic on 0h_II_R2.fastq.gz (R2 paired)
69: Trimmomatic on 0h_II_R1.fastq.gz (R1 paired)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATA CGA

Adapter list

Nothing selected

list of adapters adapter sequences which will be explicitly searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads > 50bp

No

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comparable statistics from datasets with somewhat variable read lengths. (--min_length)

length of Kmer to look for

7

note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

Email notification

No

History

search datasets

UDA

137 shown, 100 deleted, 6 hidden

79.81 GB

d)

75: Trimmomatic on 0h_III_R1.fastq.gz (R1 paired)

74: Trimmomatic on data 34 and data 33 (log file)

73: Trimmomatic on data 34 and data 33 (trimlog file)

72: Trimmomatic on 0h_II_R2.fastq.gz (R2 unpaired)

71: Trimmomatic on 0h_II_R1.fastq.gz (R1 unpaired)

70: Trimmomatic on 0h_II_R2.fastq.gz (R2 paired)

69: Trimmomatic on 0h_II_R1.fastq.gz (R1 paired)

68: FastQC on data 44: RawData

67: FastQC on data 44: Webpage

66: FastQC on data 43: R

Exemplo de Resultado FASTQC

The screenshot displays the Galaxy web interface with a FastQC Report. The browser address bar shows 'usegalaxy.org'. The main header includes 'Galaxy', navigation links, and the date 'Fri 9 Apr 2021'. The report title is 'FastQC Report' for the file 'Trimmomatic on 0h_II_R1_fastq_gz_R1 paired_gz'. The left sidebar contains tool categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. The right sidebar shows a 'History' panel with a list of previous jobs.

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

Basic Statistics

Measure	Value
Filename	Trimmomatic on 0h_II_R1_fastq_gz_R1 paired_gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26800883
Sequences flagged as poor quality	0
Sequence length	1-50
%GC	50

Per base sequence quality

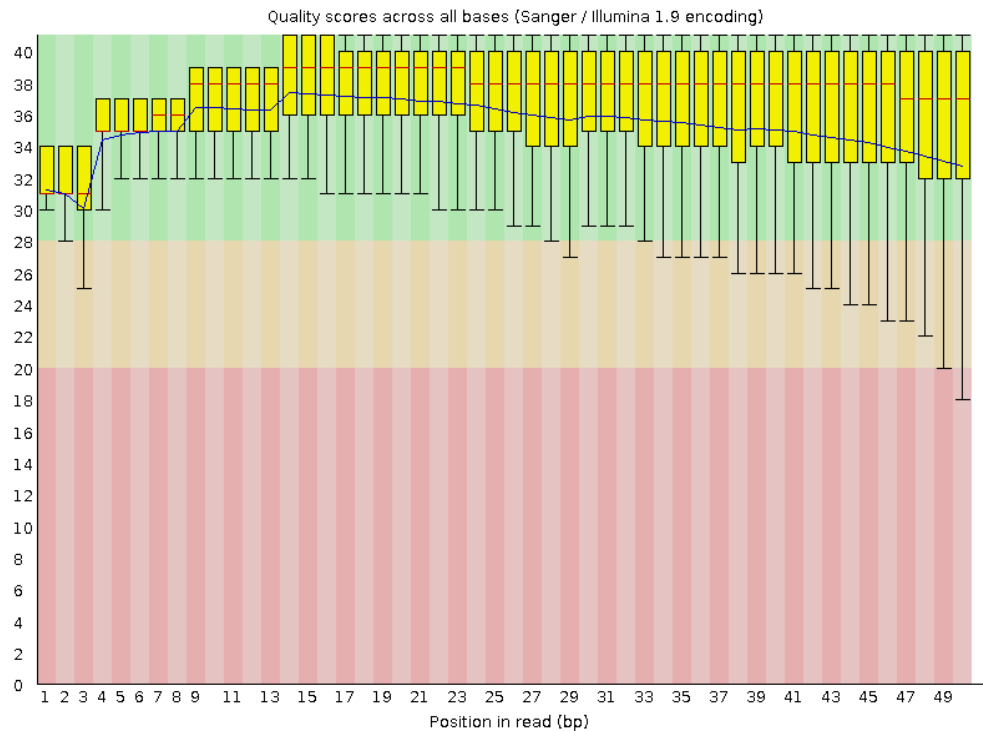
Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by [FastQC](#) (version 0.11.8)

Comparação dos resultados obtidos

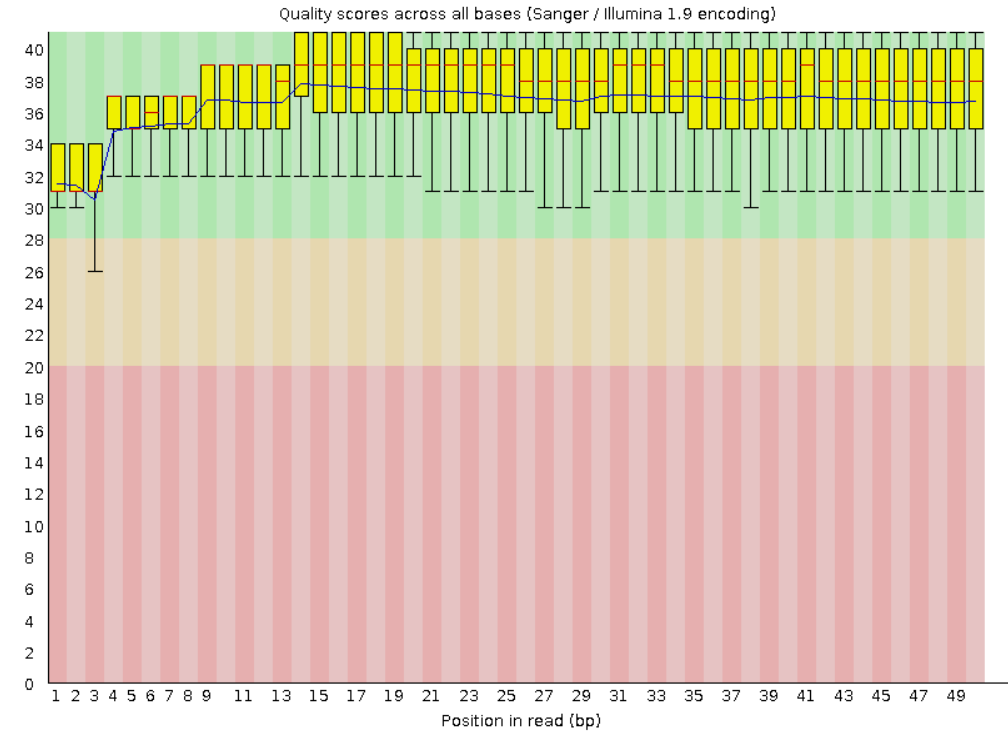
Antes (1ª rodada)

✔ Per base sequence quality



Depois (2ª rodada)

✔ Per base sequence quality



Desenvolver tabela de resultados

Table 1. General features of RNA-seq reads mapped to the *T. rubrum* reference genome

Sample	Raw reads	High-quality reads
0h II		
0h III		
3h II		
3h III		
12h II		
12h III		