Introduction

Cover letter:

Dr.Sara,
Thank you for your time on checking our App Development Proposal.
We are FASTQC2 app project team and we create a desktop software.
In this proposal, you'll see information about FASTQC2 project, and our services.

Thanks,

Gamela

Team overview

About team:

We are FASTQC2 project team, and we create a desktop software.

About vision:

Our vision is to produce desktop python program that executed by windows CMD and perform analysis on biological data in FASTQ format.

About mission:

Our mission is to make our users available to use this program to make analysis on biological data by executing program features independently using windows CMD.

Our team:

Jihad Shehata / UI designer and Back-end developer

Wessam Ahmed/ GitHub manager and Back-end developer
Sabah Gomaa/ Flutter developer

Michael Emad/ Graphic designer and video Editor
Gamela Hussien/ Back-end developer and GUI designer

Project details:

This project is a desktop application for FASTQC tool which is used to perform analysis on biological data in FASTQ files, CASAVA FASTQ files*, COLORSPASE FASTQ, GZIP compressed FASTQ, SAM, BAM and SAM/BAM Mapped only (normally used for COLORSPACE data).

FASTQC tool consists of 11 analysis modules:

1)Basic Statistics:

which gives file name, file type, encoding (ASCII encoding of quality values), total sequences, filtered sequences, sequence length and %GC (overall percentage of guanine and cytosine bases)

2)Per Base Sequence Quality

shows an overview of the range of quality values across all bases at each

position in the FASTQ file.

3)Per Sequence Quality Scores

allows to see if a subset of your sequences have universally low-quality values.

4)Per Base Sequence Content

plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

5)Per Base GC Content

Plots out the GC content of each base position in a file.

6)Per sequence GC Content

measures the GC content across the whole length of each sequence in a file.

7) Per Base N Content

If a sequencer is unable to make a base call with sufficient confidence, then it will normally substitute an N rather than a conventional base] call

This module plots out the percentage of base calls at each position for which an N was called.

8)Sequence Length Distribution

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.

This module generates a graph showing the distribution of fragment sizes in the file.

9)Duplicate Sequences

This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different

degrees of duplication.

10)Overrepresented Sequences

This module lists all of the sequence which make up more than0.1% of the total. To conserve memory only sequences which appear in the first 200,000 sequences are tracked to the end of the file. It is therefore possible that a

sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.

11)Overrepresented KMERS

This module counts the enrichment of every 5-mer within the sequence library.

The project goal is making desktop app provides same FASTQC analysis on biological data in FASTQ format created by ILLUMINA sanger/ 1.9 and is actually filtered. This program consists of the same 11 analysis modules. Source code is implemented by using python programming Language.

Sources:

click here