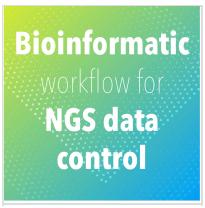

May 11, 2022

Bioinformatic workflow for NGS data control

Khalid El Moussaoui[1]

[1]Université de Liège

dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

Khalid El Moussaoui
Université de Liège

Workflow for data integrity and quality control of high throughput sequencing on Illumina NovaSeq6000. The analyses are performed on macOS Monterey 12.3.1 running on an ARM-architected Apple Silicon processor. This workflow considers that the user directory (~/) is structured as seen in the "work environment configuration" protocol. To avoid error messages, please follow this protocol and set up your computer before starting.

DOI

dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

Khalid El Moussaoui 2022. Bioinformatic workflow for NGS data control. **protocols.io** https://dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

protocol ,

May 11, 2022

May 11, 2022

62429

:

Activation of the environment

1 Open a terminal window.

Terminal 2.12.5
macOS Monterey 12.3.1
by Apple Inc.

2 Activate the previously created QC_env environment by typing the following command in the terminal :

```
conda activate QC_env
```

Data integrity check

3 Considering that the .gz archive downloaded from the GIGA servers has been unzipped under ~/fastq_files, that the original_md5.txt file has been stored under ~/md5 and that the python & R scripts previously created are stored under ~/KE_utilities, type the following command in the terminal to recompute the md5 hash and store it in a new file under ~/md5

Citation: Khalid El Moussaoui Bioinformatic workflow for NGS data control
https://dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

```
md5 ~/fastq_files/* > ~/md5/recomputed_md5.txt
```

4 After generating the ~/md5/recomputed_md5.txt file, type the following command in the terminal to launch the python script that allows the data integrity check :

```
python3 ~/KE_utilities/data_integrity_checker.py
```

This script can be downloaded on GitHub : https://github.com/elmoussaoui-k/drylab_workflow/blob/b41c0a03d7a3fdb023a94b331cab5460d706b6c8/data_integrity_checker.py

5 Specify the path to the original_md5.txt file and then to the recomputed_md5.txt file :

```
********** DATA INTEGRITY CHECKER **********

Please enter the path to original_md5.txt :
/users/khalid/md5/original_md5.txt
Please enter the path to recomputed_md5.txt :
/users/khalid/md5/recomputed_md5.txt

--------------------------------------------------
```

Run fastQC

6 Start the fastQC analysis on all existing files in the ~/fastq_files directory in recursive mode using "*". Moreover, the addition of the --outdir option allows to specify an output directory for the reports generated by fastQC. This generates an individual .html report for each file.

```
fastqc ~/fastq_files/* --outdir ~/fastqc_reports/
```

fastqc version : v. 0.11.9

Citation: Khalid El Moussaoui Bioinformatic workflow for NGS data control
https://dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

7 The generated reports can be opened by typing the following command in the terminal :

```
open ~/fastqc_reports/KE0xx_R1_fastqc.html
```

Run multiQC

8 To summarize the reports generated with fastQC into a single report, run multiQC. To do this, type the following command in the terminal :

```
multiqc ~/fastqc_reports --outdir ~/multiqc_report
```

multiqc version : v. 1.12

9 The generated report can be opened by typing the following command in the terminal :

```
open ~/multiqc_report/multiqc_report.html
```

Filter reads with fastp

10 The reads can be filtered automatically with fastp. Just launch the program, specify the 2 .fastq.gz files (R1 and R2) as input and specify the name and location of the 2 processed files. Adding the -h option allows to specify a folder for the HTML report. The option -j " " allows to cancel the creation of the JSON report. The -R option allows to give a name to the generated HTML report.

Citation: Khalid El Moussaoui Bioinformatic workflow for NGS data control
https://dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1

fastp -i ~/fastq_files/KE0xx_R1.fastq.gz
-I ~/fastq_files/KE0xx_R2.fastq.gz
-o ~/fastp/cleaned_fastq_files/KE0xx_R1_clean.fastq.gz
-O ~/fastp/cleaned_fastq_files/KE0xx_R2_clean.fastq.gz
-h ~/fastp/fastp_reports/KE0xx_fastp_report.html
-j ""
-R "Fastp report : KE0xx"

fastp version : v. 0.23.2

Citation: Khalid El Moussaoui Bioinformatic workflow for NGS data control
https://dx.doi.org/10.17504/protocols.io.8epv59bnjg1b/v1