RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Help4MultiQC

Gaston Rognon, Claire Hoede,
Sarah Maman

27 août 2024

Présentation de mon Cursus

Entré en M2 bio-informatique et génomique environnemental

Stage optionnel de 2 mois entre Juillet et août 2024

6 mois de projet tutoré dans le cadre du projet DEFIS (PALEOFISH)

Sommaire

Présentation du projet Help4MultiQC

Pourquoi et pour qui?

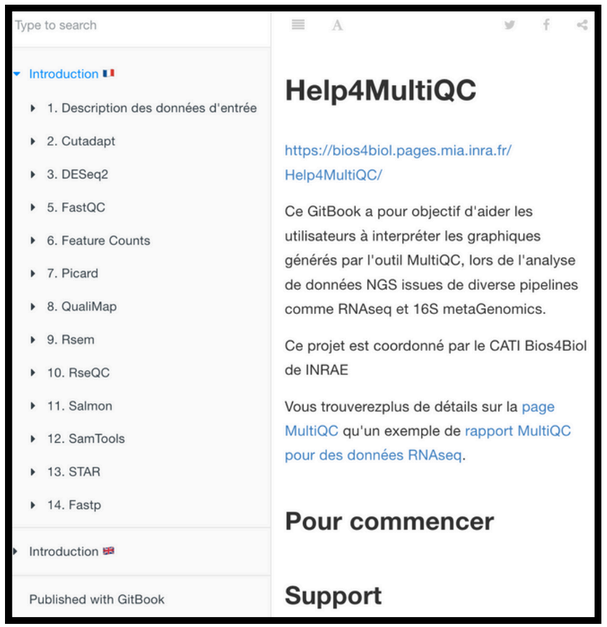
- Les sources des différents outils présentés sont très variés
- Help4MultiQC destiné aux biologistes pour affiner leur interprétation des sorties MultiQC.
- Début du projet en juin 2022 au sein du CATI BIOS4BIOL par Claire Hoede, Yannick Lippi, Cervin Guyomar et Sarah Maman.

Le contenu de ce Gitbook a été principalement inspiré par du contenu déjà publié dont voici les sources, d'autres sources sont listées tout au long du GitBook:

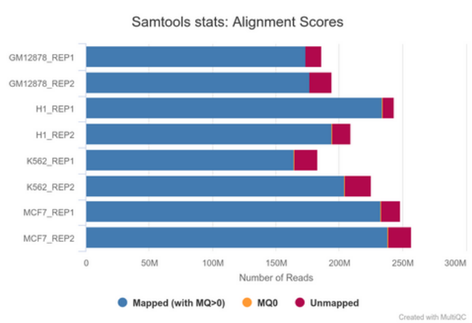
- https://elearning.formation-permanente.inrae.fr/course/view.php?id=196&section=1
- https://github.com/nf-core/rnaseq/blob/master/docs/output.md#quality-control
- https://github.com/hbctraining/Intro-to-rnaseq-hpc-salmon/blob/master/lessons/qc_fastqc_assessment.md
- https://youtu.be/qPbIlO_KWNo
- https://multiqc.info/docs/#using-multiqc
- https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx
- https://subread.sourceforge.net/featureCounts.html
- https://multiqc.info/
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1276-2
- https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard
- https://academic.oup.com/bioinformatics/article/28/20/2678/206551?login=true
- https://nf-co.re/rnaseq/3.14.0/results/rnaseq/results-b89fac32650aacc86fcda9ee77e00612a1d77066/aligner_star_rsem/multiqc/star_rsem/?file=multiqc_report.html
- http://qualimap.conesalab.org/
- https://rnnh.github.io/bioinfo-notebook/docs/featureCounts.html
- https://homolog.us/blogs/tech/2012/02/19/illumina-paired-end-libraries-inward-and-outward-looking-reads/
- https://genome.cshlp.org/content/suppl/2008/09/26/gr.078212.108.DC1/maq-supp.pdf
- http://www.htslib.org/doc/samtools-flagstat.html
- https://www.biostars.org/p/268550/
- https://support.bioconductor.org/p/91818/
- https://help.galaxyproject.org/t/unassigned-ambiguity-problem-in-featurecounts/5921/2
- https://cutadapt.readthedocs.io/en/stable/

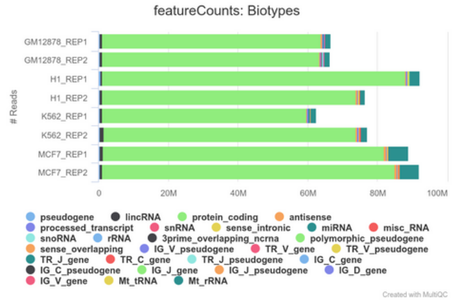
Présentation du projet Help4MultiQC

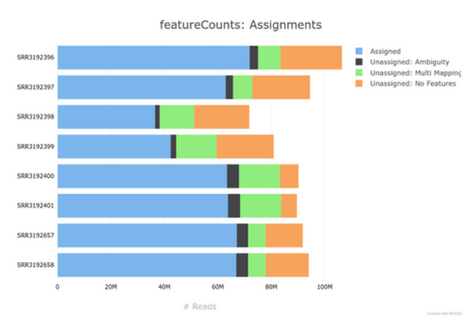
Objectif : réaliser un GitBook

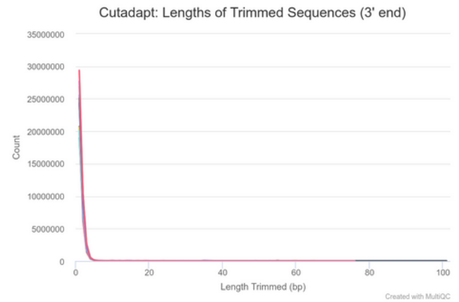


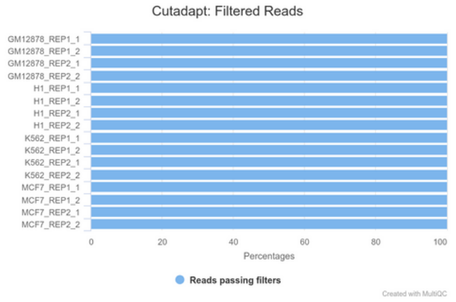
Pour les reports MultiQC :

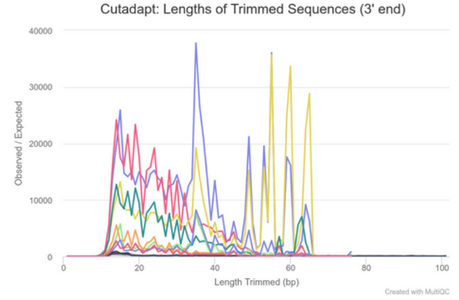












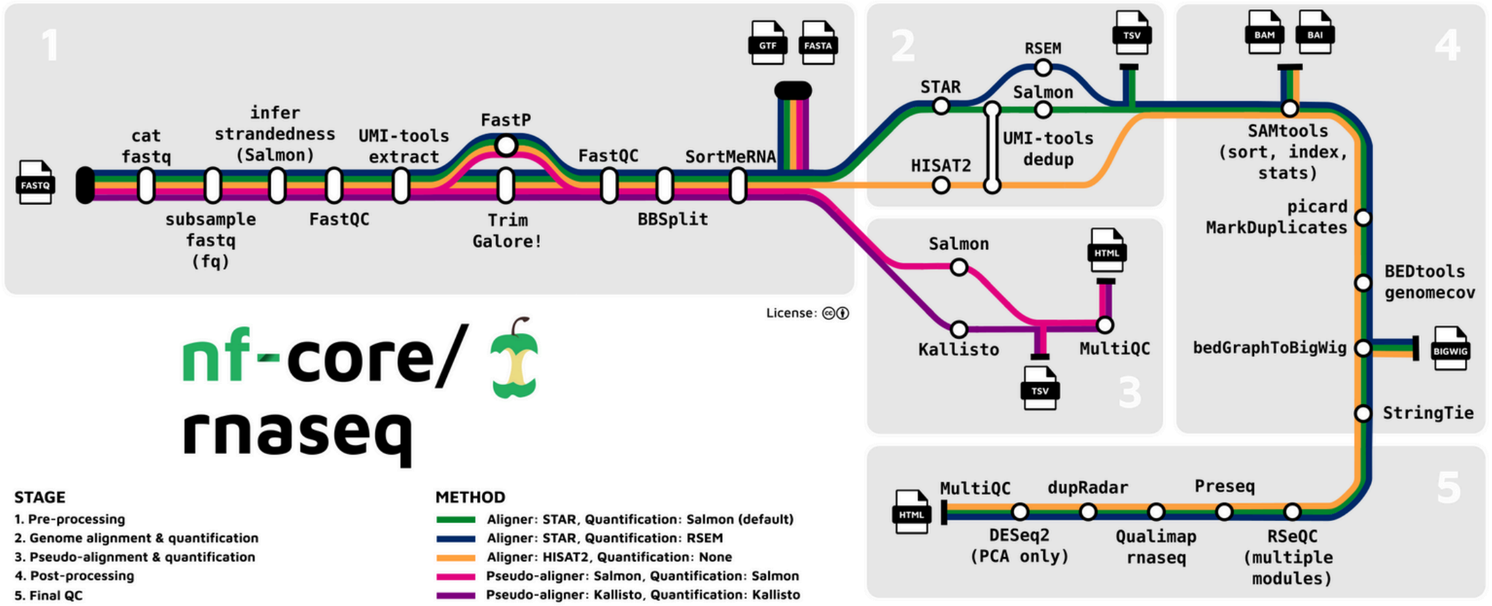
rnaseq: Results

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

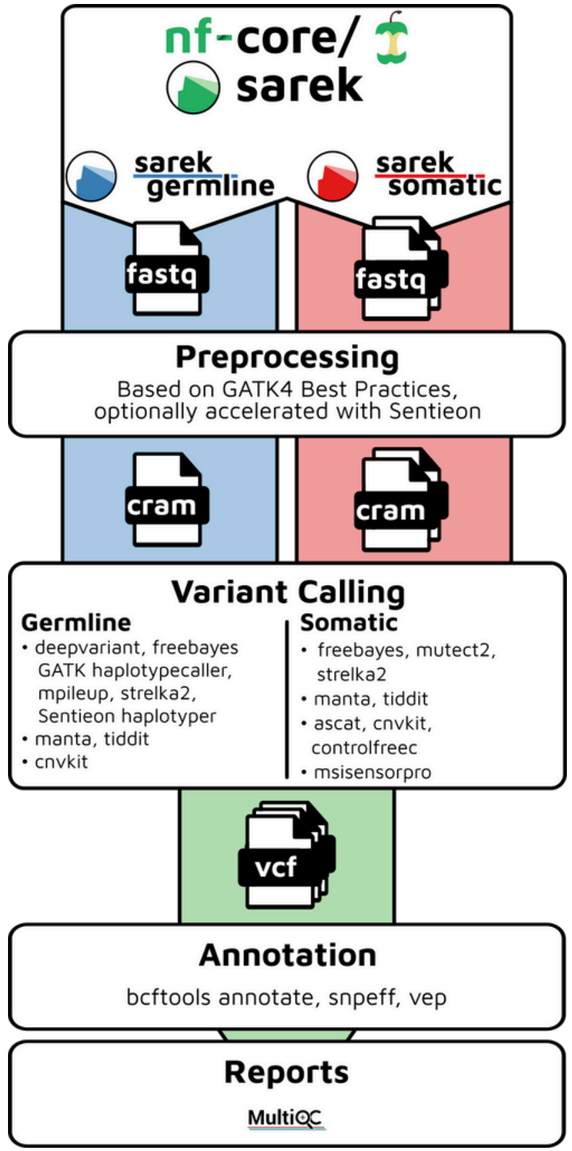
netlify.app

Récapitulation de Help4MultiQC

Focus sur deux pipelines Nextflow nf-core :



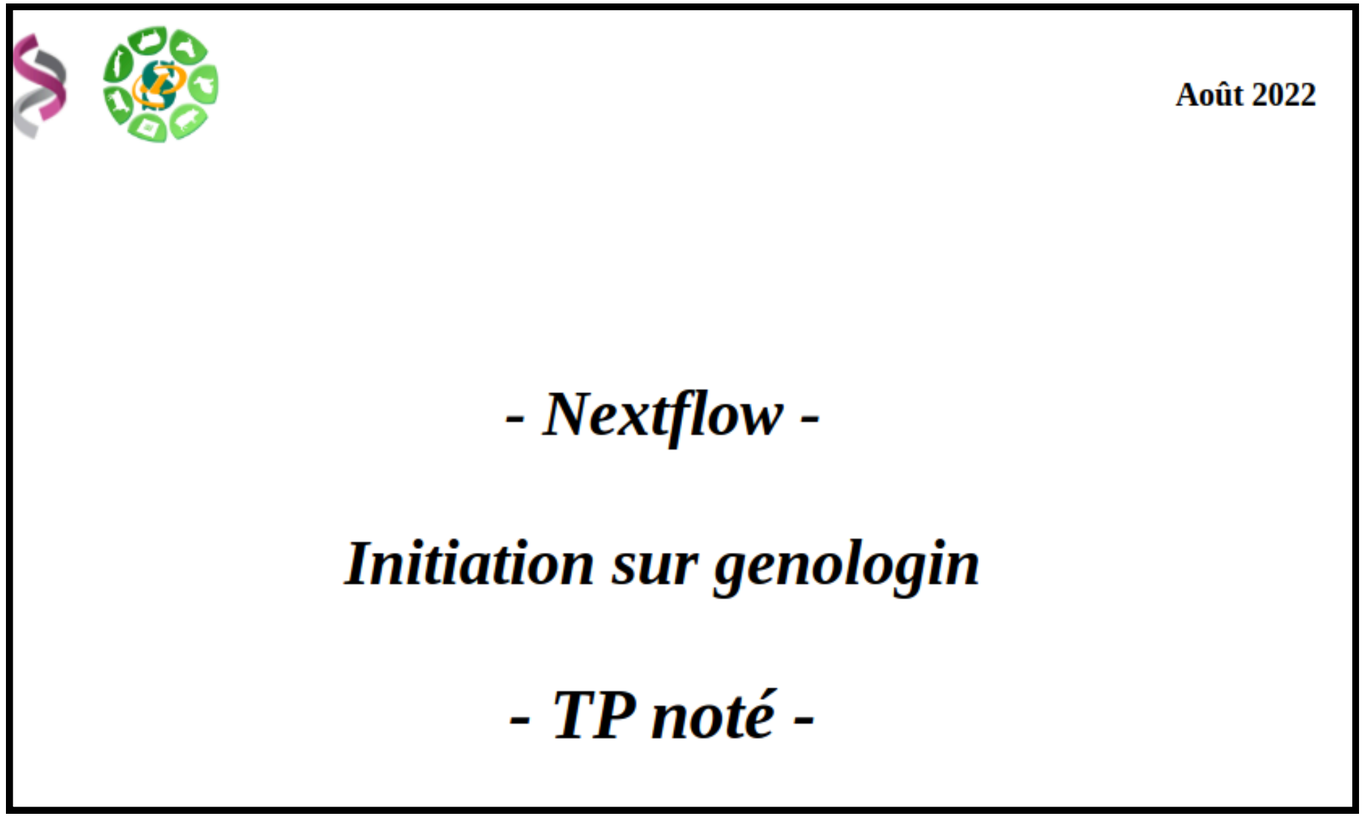
Source



Source

Matériels et Méthodes

Prise en main du pipeline nextflow nf-core/RNAseq



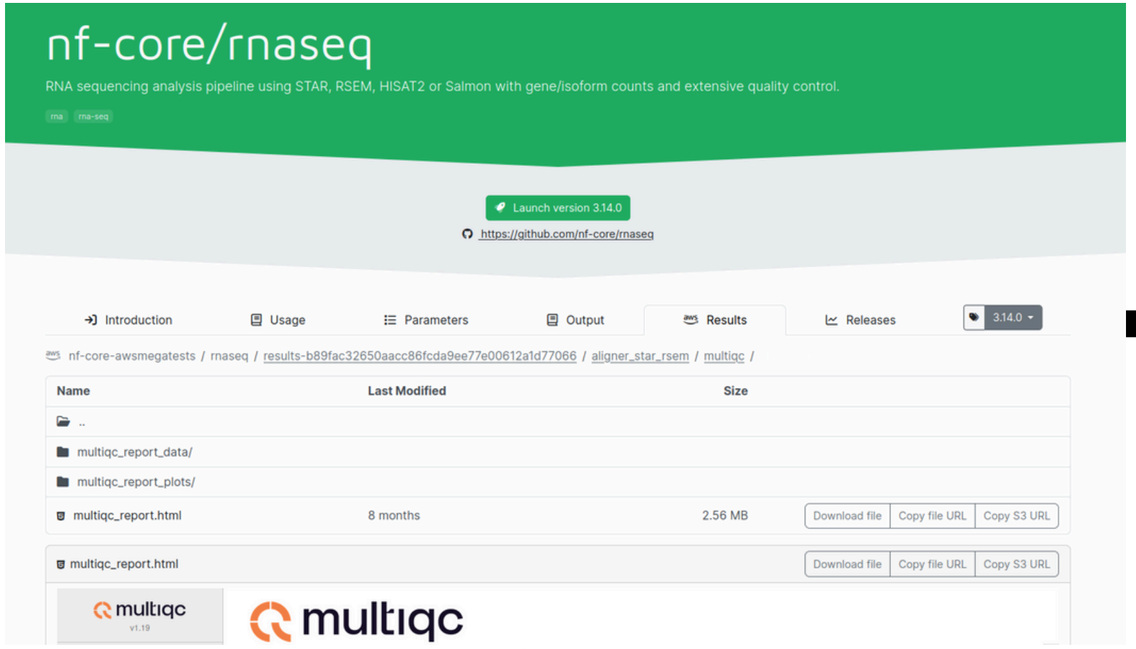
projet tutoré de mise en main d'un pipeline nextflow nf-core

Matériels et Méthodes

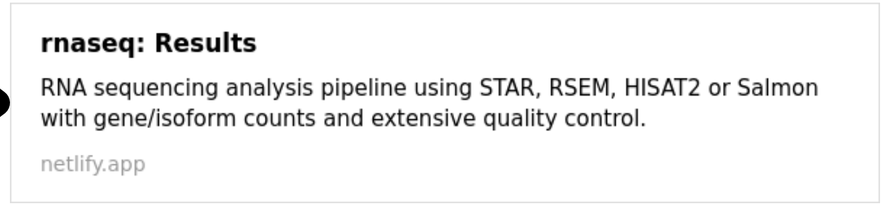
Recherche bibliographique

Rapports HTML MultiQC issus de la documentation Nextflow :

Source

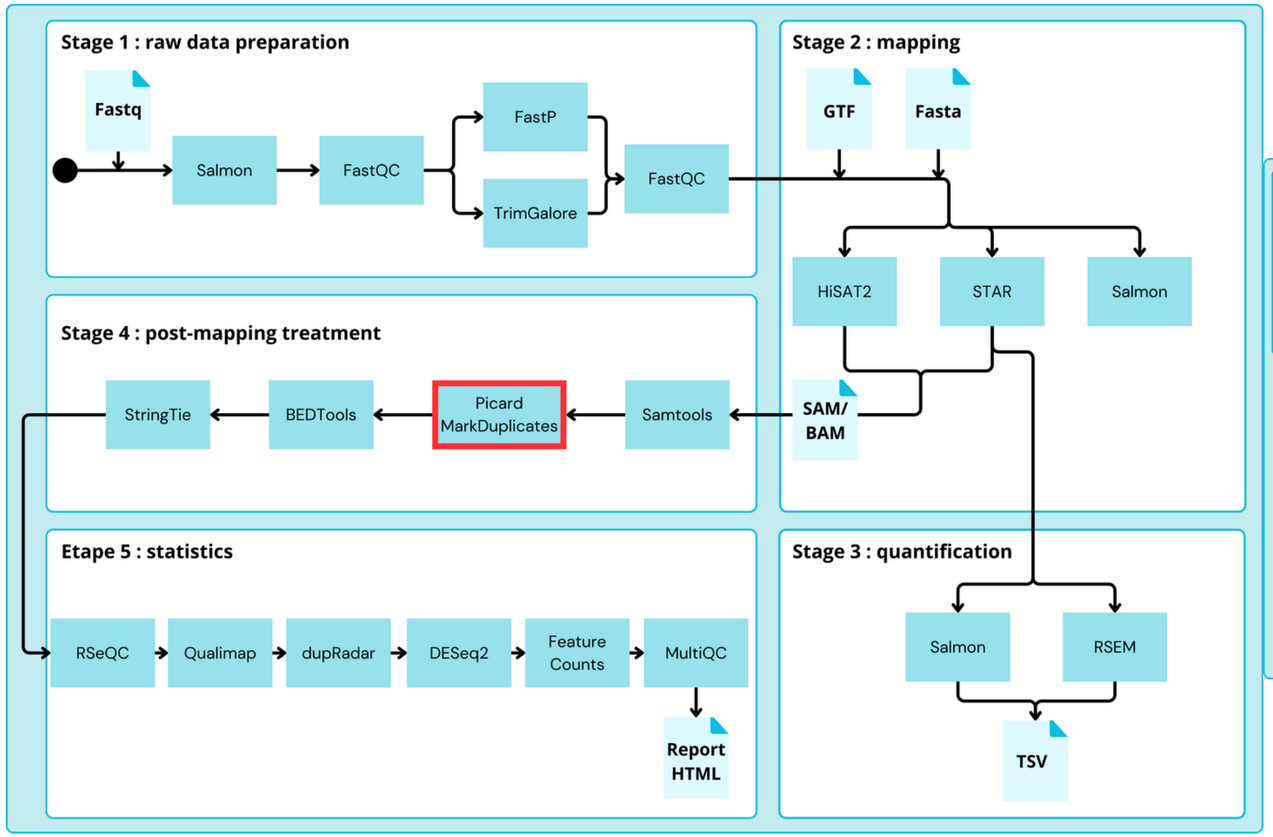


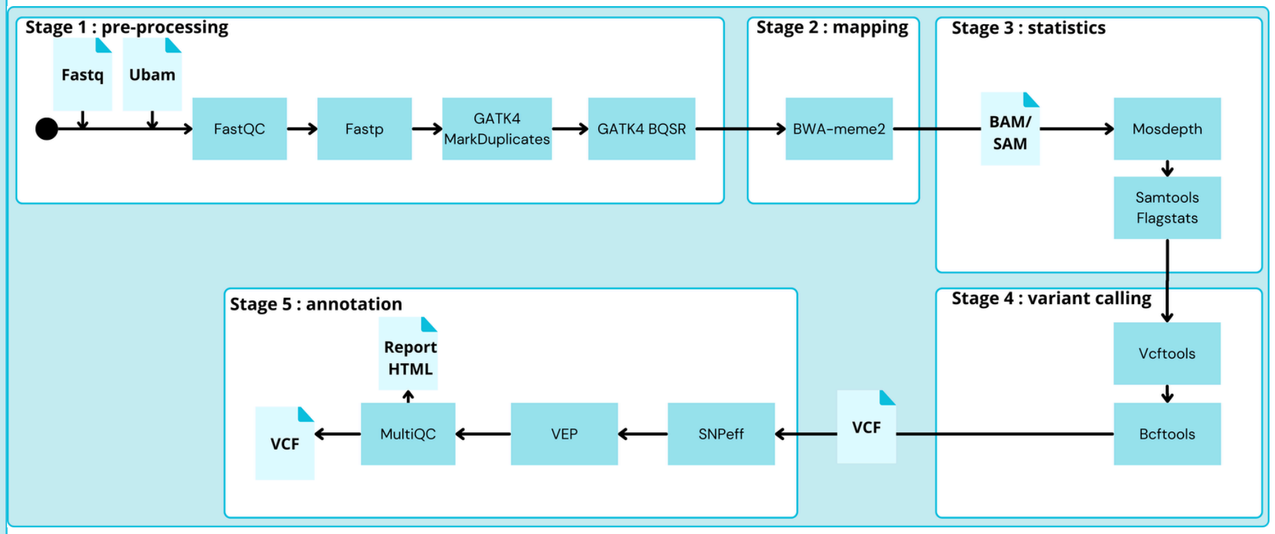
Source



Matériels et Méthodes

Prise en main des pipelines





Recréation des pipelines
RNAseq et Sarek

Matériels et Méthodes

Recherche de données utilisées dans Nextflow

Recherche de données utilisées dans les sorties manifest des outputs Nextflow nf-core ou sur le Gitlab.

manifest_2024-08-07_13-48-48.bco.json


```
        },
        {
            "step_number": 20,
            "name": "b5cc5e55bb7b9bcffc5e63adaa8071cf",
            "description": "NFCORE_SAREK:SAREK:FASTP (HCC1395N-1)",
            "input_list": [
                {
                    "uri": "s3://ngi-igenomes/test-data/sarek/SRR7890919_WES_HCC1395BL-EA_normal_2.fastq.gz"
                },
                {
                    "uri": "s3://ngi-igenomes/test-data/sarek/SRR7890919_WES_HCC1395BL-EA_normal_1.fastq.gz"
                }
```


Source

Description des données : méthode de séquençage, paired end ou single end, librairie, etc.

study_alias	run_accession	experiment_alias	encode_library_id	sample_description	instrument_model	library_
GSE78551	SRR3192657	GSM2072350	ENCLB038ZZZ	Homo sapiens GM12878 immortalized cell line	Illumina HiSeq 2000	PAIRED
GSE78551	SRR3192658	GSM2072351	ENCLB037ZZZ	Homo sapiens GM12878 immortalized cell line	Illumina HiSeq 2000	PAIRED
GSE78557	SRR3192408	GSM2072362	ENCLB055ZZZ	Homo sapiens K562 immortalized cell line	Illumina HiSeq 2000	PAIRED
GSE78557	SRR3192409	GSM2072363	ENCLB056ZZZ	Homo sapiens K562 immortalized cell line	Illumina HiSeq 2000	PAIRED

Source

Matériels et Méthodes

Prise de contact au sein du CATI Bios4Biol

Mail de présentation du projet

Bonjour,

Je me permets de vous relancer personnellement par conseil de Claire Hoede et Sarah Maman.
J'ai entendu que vous aviez déjà travaillé sur des sorties MultiQC RNAseq ou SAREK et si vous avez encore vos données MutliQC vous pourriez recouper mes informations et avoir des exemples de bons ou mauvais résultats.

Je suis Gaston Rognon, en stage d'été (juillet - août 24) au sein du CATI BIOS4BIOL pour poursuivre le projet Help4MultiQC.
J'aurais besoin de votre aide pour m'aider à décrire les graphiques générés par MultiQC pour les pipelines RNAseq et SAREK.
Si vous pouvez m'accorder 1h en visio https://inrae-fr.zoom.us/j/9810418696
Pour m'expliquer un ou plusieurs graphiques de votre choix, merci de compléter ce tableur :
https://docs.google.com/spreadsheets/d/1ZrG51OfYHsbJNhZvpKBJsImEwqffRz9th-wfmo1WTdw/edit?usp=sharing

Les graphiques à analyser sont listés en ligne et les plages horaires en colonne.
Dès que le plage horaire est grisé, cela signifie qu'elle n'est plus disponible.
Merci de préciser votre prénom et nom dans la cellule, ou les, cellule(s) choisies afin que je puisse vous envoyer une invitation Outlook en suivant.
N'hésitez pas à remplir plusieurs graphiques par sessions, votre participation est indispensable au bon déroulement de mon stage.

Les résultats du stage seront présentés en visio le mardi 27 août à 14 h.

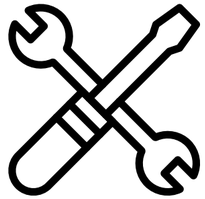
Je vous remercie de votre temps,
Gaston Rognon

Planning avec dates et outils

HORAIRE : →	08/07/2024 de 9h à 10h	08/07/2024 de 14h à 15h	09/07/2024 de 9h à 10h	09/07/2024 de 14h à 15h
Bam Stat				
Salmon				
Samtools :				
Percent Mapped				
Alignment metrics				
Samtools Flagstat				
XY counts				
Mapped reads per contig				
FastQC (raw) :				
Top overrepresented sequences				
Cutadapt :				
Filtered Reads				
Trimmed Sequence Lengths (3')				
SAREK		Mathieu		
FastP (Read preprocessing)				
Filtered Reads				
Insert Sizes				
Sequence Quality				
GC Content				
N content				
GATK4 MarkDuplicates				
Samtools Flagstat				
Percent mapped			Claire	
Alignment stats			Claire	
Mosdepth				
Cumulative coverage distribution				
Coverage distribution				

Matériels et Méthodes

Rédaction

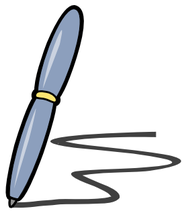


Outils d'aide à la rédaction :

DeepL

Copilot

scribens

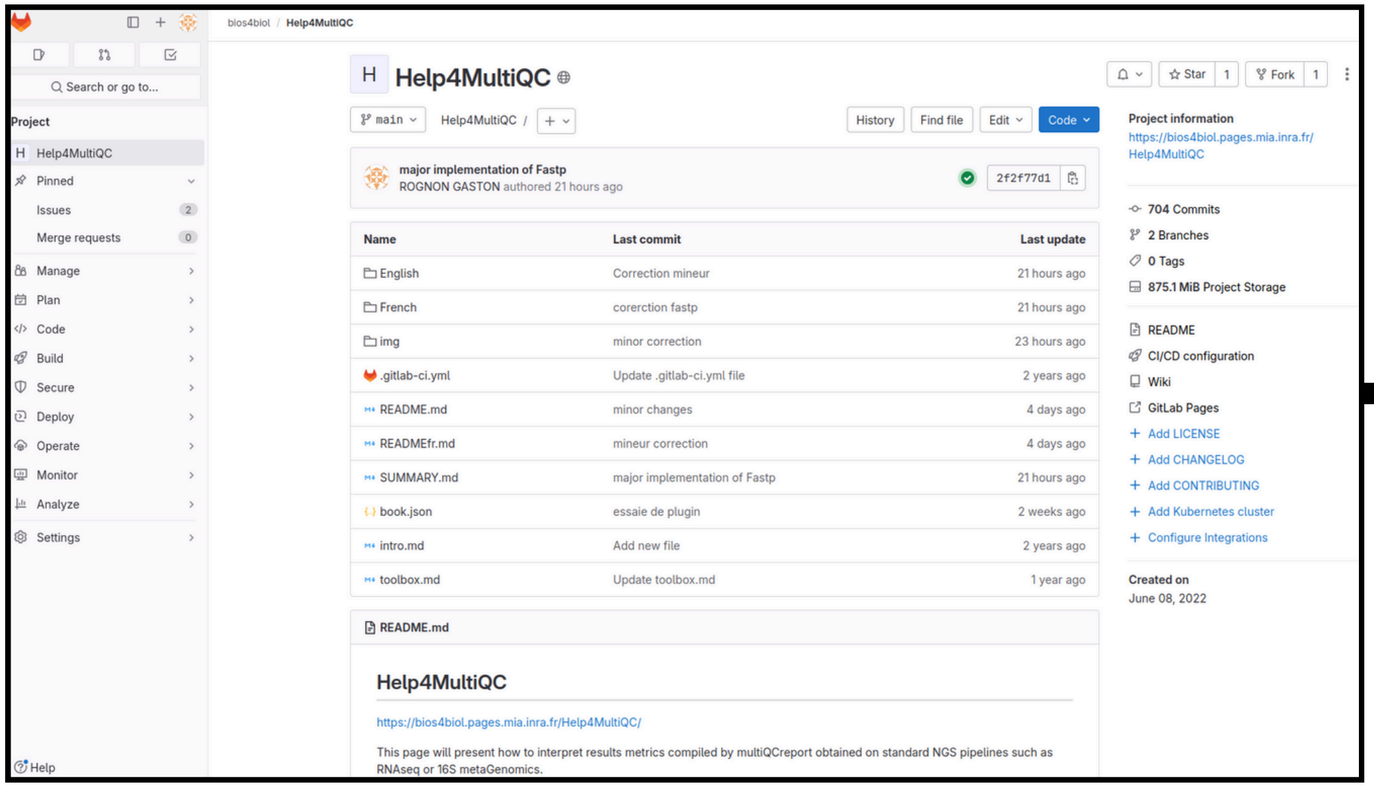


Rédaction sur la forge MIA au format Markdown :

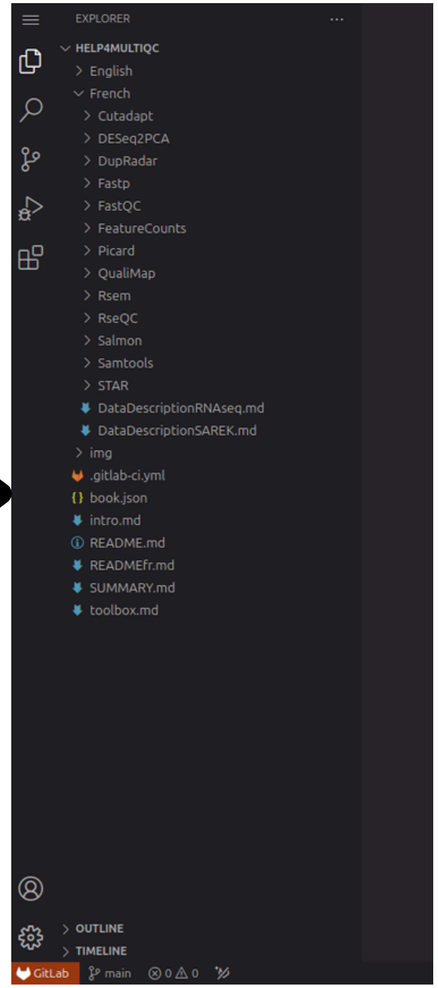
GitLab

Matériels et Méthodes

Travail sur le GitHub du CATI BIOS4BIOL



Les Edit sur le web IDE

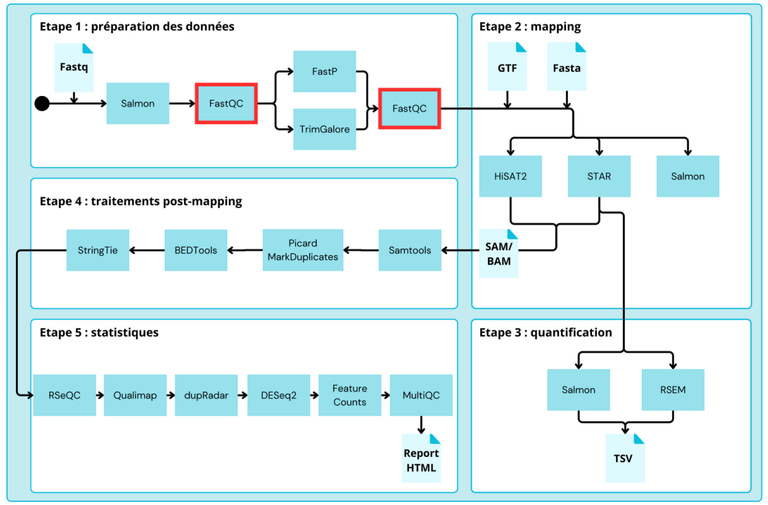


Source

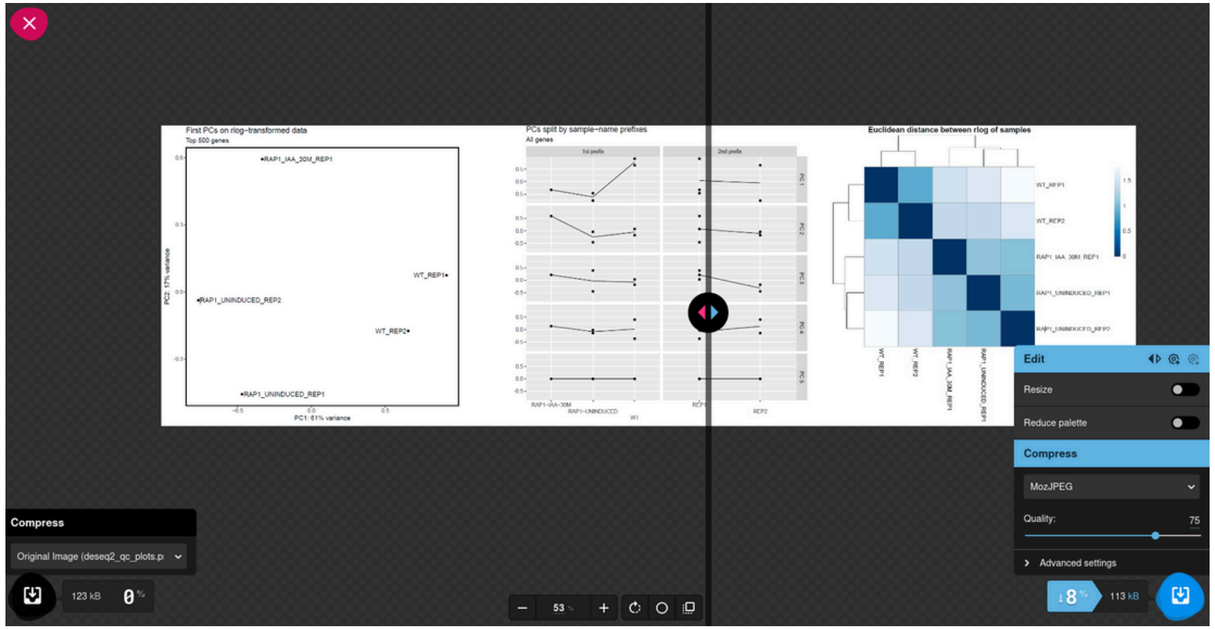
Matériels et Méthodes

Outils divers

Lien Canva



Squoosh



Réalisation

Recherches bibliographies et état de l’art

Synthèse de sources variées :

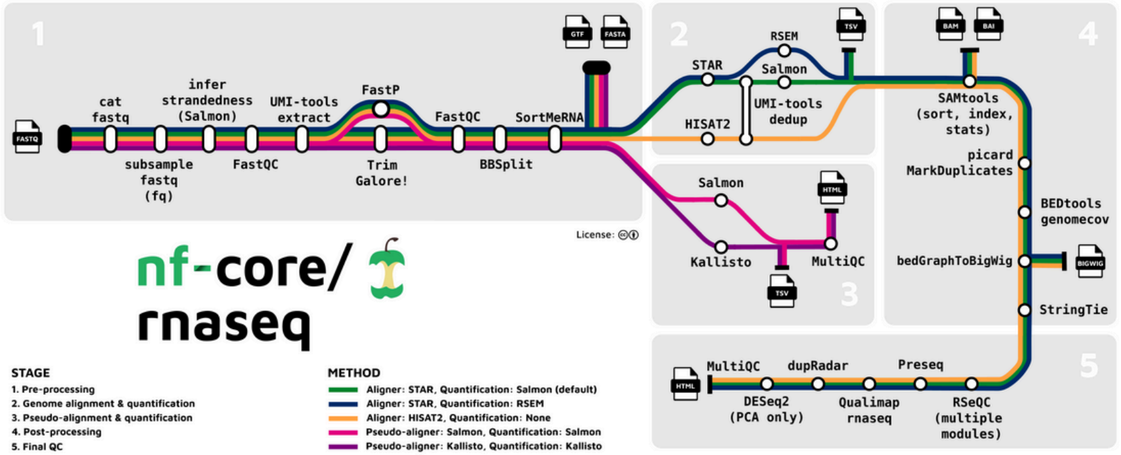
- Documentation des outils
- Manuel
- Entretien avec les membres du CATI BIOS4BIOL

Traduction anglais/français et reformulation

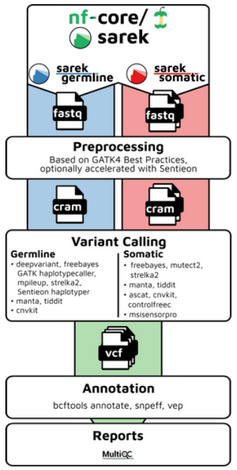
Mises en forme des informations :

- Prise en main du GitBook
- Mise en forme du texte et des images
- Pipelines sous canva

Présentation du GitBook



Source



Source

En cours

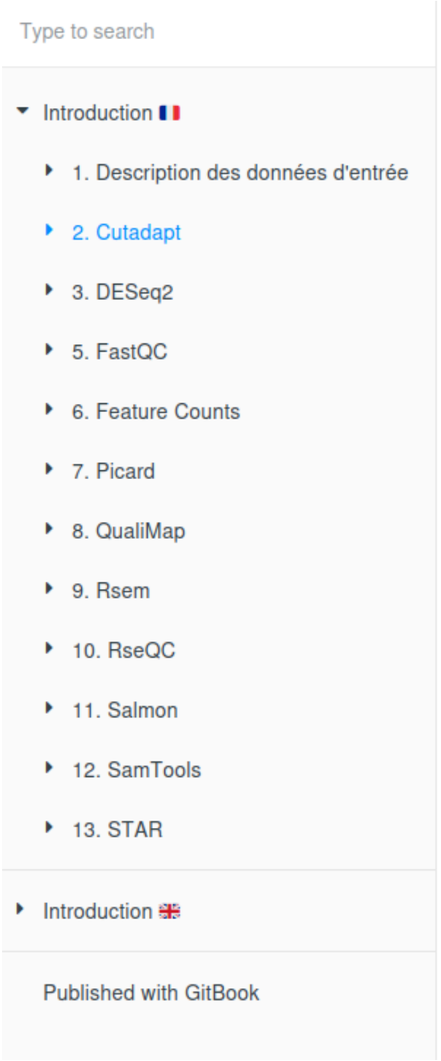
Présentation du GitBook

Organisation par langue et outils.

Chaque outil analysé par MultiQC fait l'objet d'un chapître.

L'ensemble du GitBook est disponible en français et en anglais.

https://bios4biol.pages.mia.inra.fr/Help4MultiQC/READMEfr.html



- 14 outils
- 44 grahiques

Présentation du GitBook

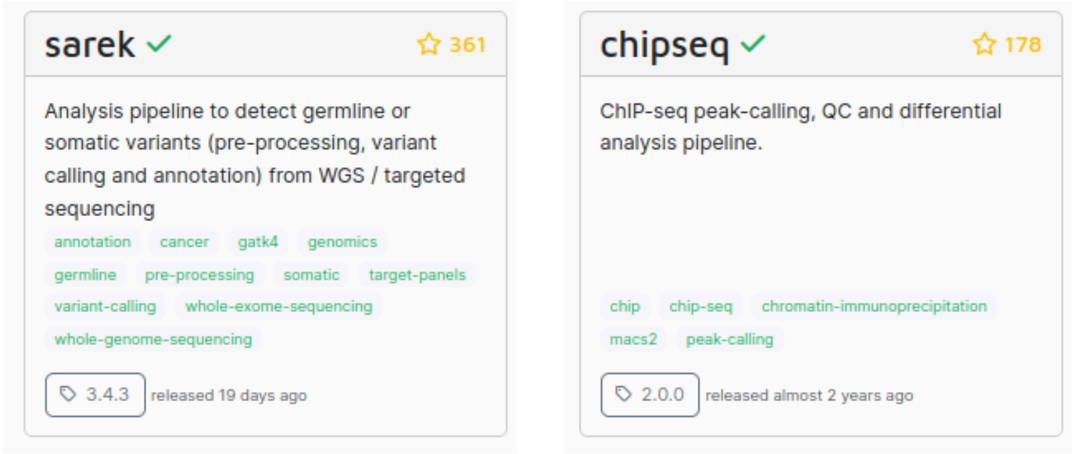
Accès au GitBook

Sur le GitHub du CATI BIOS4BIOL

Lien vers la formation

Perspectives

Pipeline à finir ou à commencer (SAREK Metagenomic, CHIPseq).



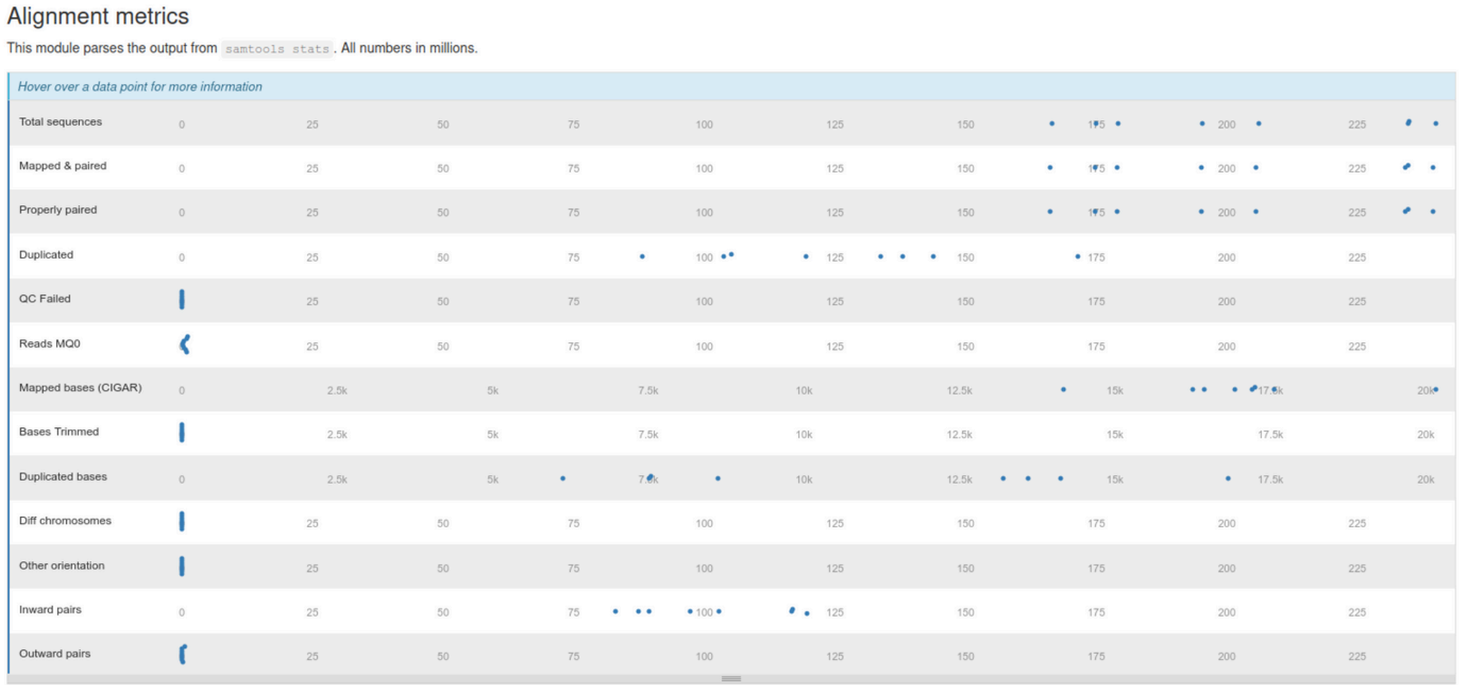
Source

Modifications mineurs à apporter sur le Gitbook:

- Supprimer les liens vers réseaux sociaux
- Optimiser les images du git (taille et format)

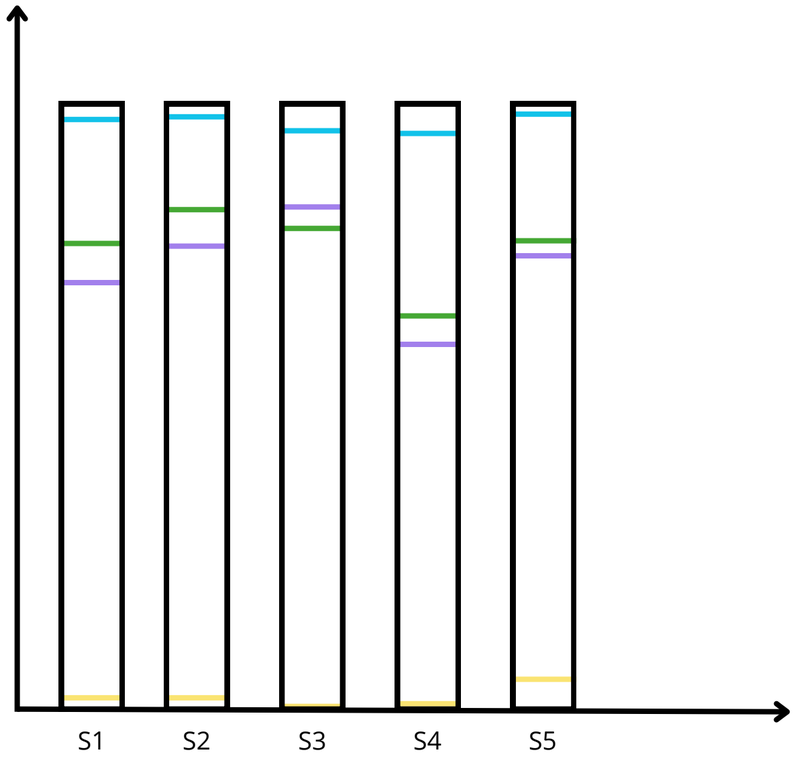
Perspectives

Proposer un nouveau graphique pour MultiQC :



Source

Originale



Nouveau

Problèmes :

- Mauvaise visualisation des valeurs d'un échantillon
- pas de violon dans ce diagramme en violon

Remerciement

Sarah Maman, Claire Hoede,
Matthias Zytnicki, Eric Casellas,
Han Phan et Mathieu Charles