FAIR Bioinfo: Retour d'expérience du MNHN

Marie Cariou

UAR 2700 Acquisition et Analyses de Données pour l'histoire naturelle

11 octobre 2023

- **Qui suis-je?**
  Présentation de l'UAR 2AD
  FAIR Bioinfo 2022

- **Mise en oeuvre au quotidien**

- **Perspectives: FAIR-bioinfo MNHN?**

La bioinfo au MNHN

UAR 2700 2AD

ACQUISITION ET ANALYSE DE DONNÉES POUR L'HISTOIRE NATURELLE

Pôle Acquisition

SSM
(Service de Systématique Moléculaire)

AST-RX
(Accès Scientifique à la Tomographie à Rayon X)

AIS
(Atelier d'Iconographie Scientifique)

Pôle Analyse

Service analyse de données

Patricia Wils	Analyse d'image
Jawad Abdelkrim	génomique
Marie Cariou	
Amandine Blin	statistique
Tristan Tchilinguirian	IA

PCIA
(Plateau Calcul Intensif et Algorithmique)

Fayçal Allouti

3 coordinateur.rices:
Mathilde Carpentier
Julien Mozziconacci
Alain Paris

3 départements de recherche

- Origine et Evolution (4 UMR, 1 USR)
- Homme et environnement (6 UMRs)
- Adaptation du vivant (5 UMRs)

Le Service Analyse de Données

- apporte un support aux unités de recherches du MNHN dans leurs analyses de données en génomique, statistique et imagerie.

Aide à la conception de projets

- Choix technologiques
- Plans expérimentaux
- Choix des outils d'analyse

Prise en charge d'analyses

- Conseils
- Encadrement
- Collaborations

Développement méthodologique

- Scripts et pipelines
- Modélisation

Animation et formation

- Séminaires scientifiques
- Ateliers/formations

Moi versus les *use case*:

Profile: first steps towards bioinformatics reproducibility

Aim: manage the traceability of the daily activities of a bioinformatician

- data analysis ⇒ notebook
- code development ⇒ code versioning

Schedule:

1. Data life cycle & project management
2. Notebook with jupyter
3. Code versioning with git

→ Analyse de données pour une équipe de recherche

Git + R markdown

Profile: about the code execution reproducibility

Aim: controlling the code execution and gain into its distribution to the communi

- control the execution environment of the code by its encapsulation
- how to distribute my code to the community?

Schedule:

1. environment management with CONDA
2. execution encapsulation with a docker container
3. code diffusion with GitHub

Profile: reproducibility on IFB resources

Aim: scaling up your analysis

- use the national IFB resources
- many data and many analysis steps: create a workflow!

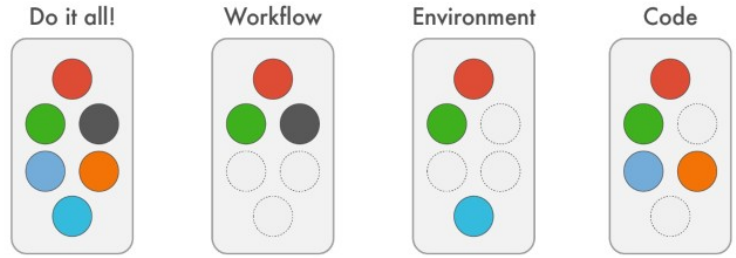
Schedule:

1. the IFB resources: jupyterlab + computational cluster with slurm
2. workflow with snakemake
3. application: run a workflow on the IFB computational cluster

→ Travail sur cluster MNHN

FAIR bioinfo 2022

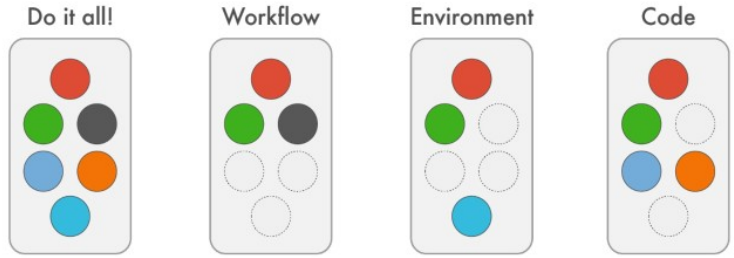
La présentation du programme de la formation:



https://nbis-reproducible-research.readthedocs.io/en/course_2104/introduction/

FAIR bioinfo 2022

Avant la formation:

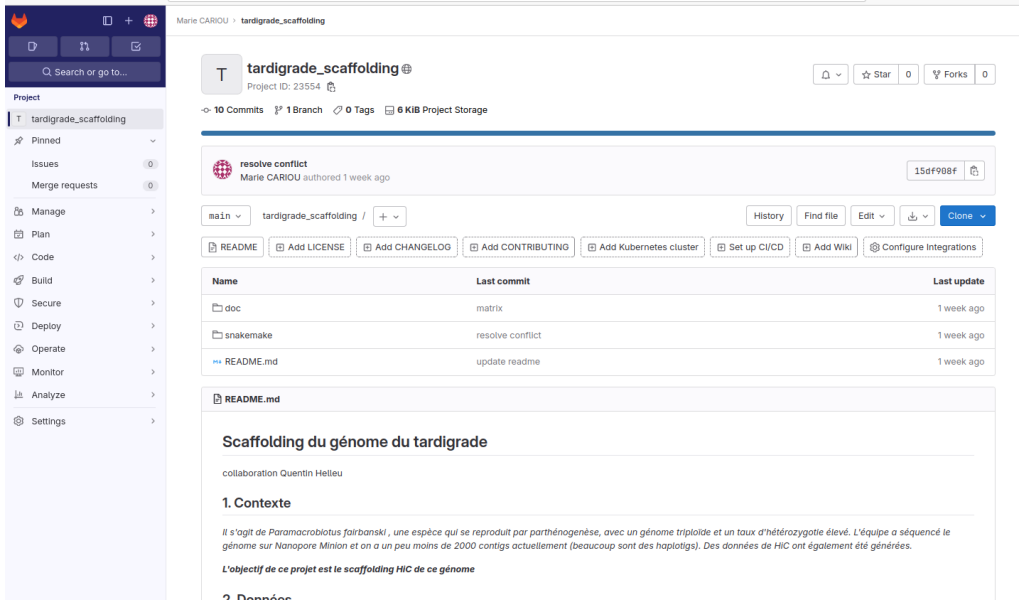


https://nbis-reproducible-research.readthedocs.io/en/course_2104/introduction/

Mise en oeuvre au quotidien

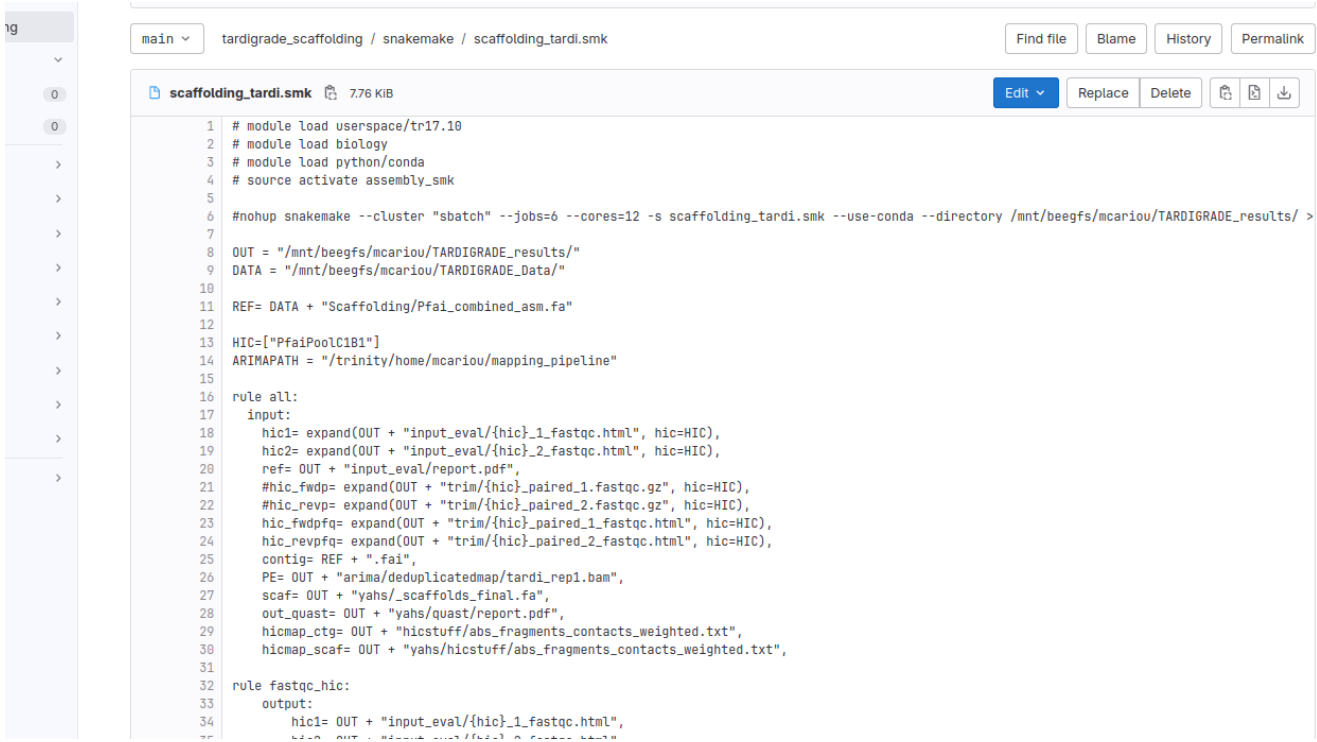
Projet exemple: scaffolding de génome avec des données HiC.

- Dépot git du projet (gitlab in2p3 utilisé dans l'UAR)
- Partage des scripts, notebook etc.
- Suivi du projet..



Mise en oeuvre au quotidien

Changement majeur dans ma pratique: snakemake



main / tardigrade_scaffolding / snakemake / scaffolding_tardi.smk

Find file | Blame | History | Permalink

scaffolding_tardi.smk 7.76 KiB

Edit | Replace | Delete

```
# module load userspace/tr17.10
# module load biology
# module load python/conda
# source activate assembly_smk

#nohup snakemake --cluster "sbatch" --jobs=6 --cores=12 -s scaffolding_tardi.smk --use-conda --directory /mnt/beegfs/mcariou/TARDIGRADE_results/ >

OUT = "/mnt/beegfs/mcariou/TARDIGRADE_results/"
DATA = "/mnt/beegfs/mcariou/TARDIGRADE_Data/"

REF= DATA + "Scaffolding/Pfai_combined_asm.fa"

HIC=["PfaiPoolC1B1"]
ARIMAPATH = "/trinity/home/mcariou/mapping_pipeline"

rule all:
  input:
    hic1= expand(OUT + "input_eval/{hic}_1_fastqc.html", hic=HIC),
    hic2= expand(OUT + "input_eval/{hic}_2_fastqc.html", hic=HIC),
    ref= OUT + "input_eval/report.pdf",
    #hic_fwdp= expand(OUT + "trim/{hic}_paired_1.fastqc.gz", hic=HIC),
    #hic_revp= expand(OUT + "trim/{hic}_paired_2.fastqc.gz", hic=HIC),
    hic_fwdpfq= expand(OUT + "trim/{hic}_paired_1_fastqc.html", hic=HIC),
    hic_revpfq= expand(OUT + "trim/{hic}_paired_2_fastqc.html", hic=HIC),
    contig= REF + ".fai",
    PE= OUT + "arima/deduplicatedmap/tardi_rep1.bam",
    scaf= OUT + "yahs/_scaffolds_final.fa",
    out_quast= OUT + "yahs/quast/report.pdf",
    hicmap_ctg= OUT + "hicstuff/abs_fragments_contacts_weighted.txt",
    hicmap_scaf= OUT + "yahs/hicstuff/abs_fragments_contacts_weighted.txt",

rule fastqc_hic:
    output:
        hic1= OUT + "input_eval/{hic}_1_fastqc.html",
```

Mise en oeuvre au quotidien

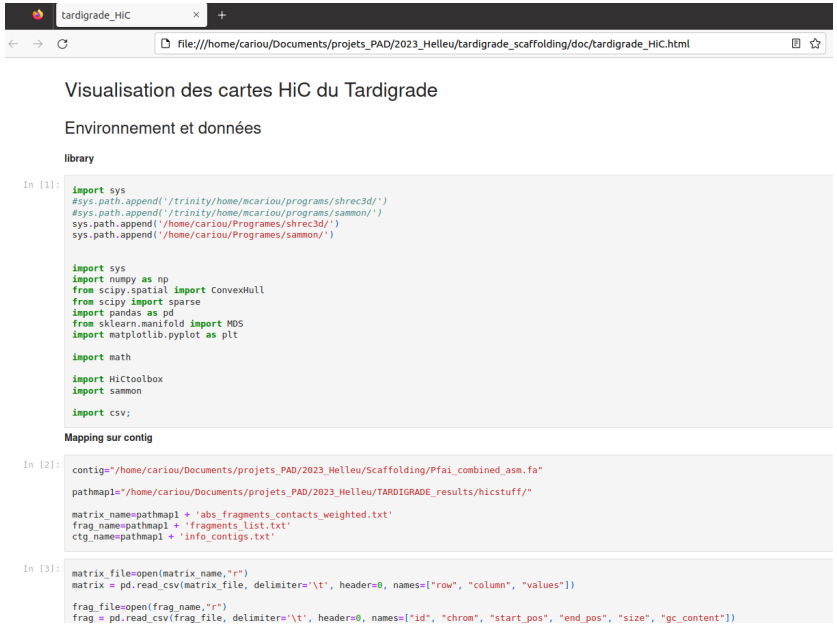
Changement majeur dans ma pratique: snakemake

- Sur cluster PCIA du MNHN
- Avec modules et conda
- Difficultés rencontrées:
  - Gestions des *rules* avec envmodules versus conda (et les 2?)
  - drmaa??

Mais aussi:

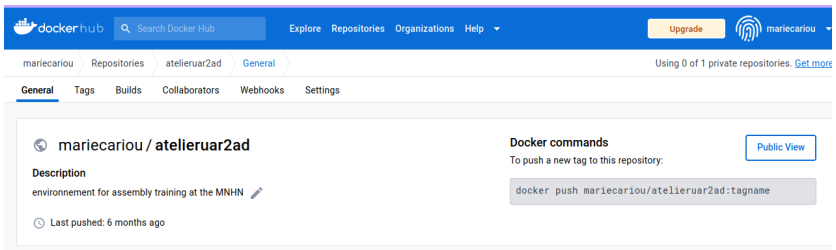
- des notebook python sans Jupyterlab
- des Docker

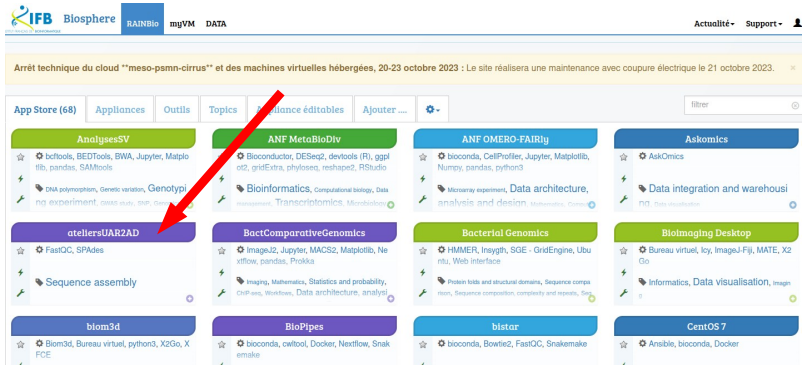
Des notebook



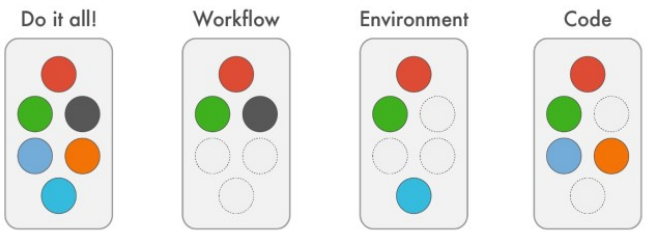
Un Docker

Pour la construction d'un environnement pour un TP sur l'assemblage des génomes.

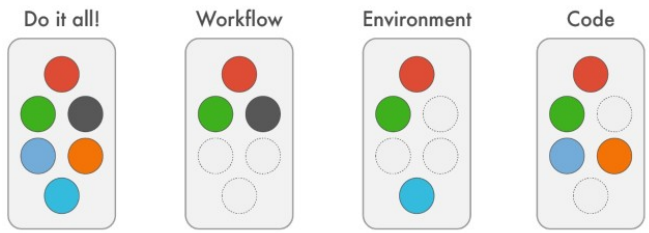




Après la formation:



https://nbis-reproducible-research.readthedocs.io/en/course_2104/introduction/



https://nbis-reproducible-research.readthedocs.io/en/course_2104/introduction/

La formation au FAIR au MNHN

Thématique déjà présente dans des formations existantes

- Module "Nettoyer et structurer les données" organisée par l'UAR BBEES
- Atelier "Introduction à la génomique" de l'UAR 2AD
- ...

Certains outils déjà présents dans des formations existantes

- Journée gitlab organisée par l'UAR BBEES
- Atelier "Rmarkdown" de l'UAR 2AD
- ...

Un atelier "snakemake" début 2024?

Conclusion

Merci :)