

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



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

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

PCIA
(Plateau Calcul Intensif et
Algorithmique)

Fayçal Alloui



3 départements de recherche

Origine et Evolution
(4 UMR, 1 USR)

Homme et
environnement
(6 UMRs)

Adaptation du vivant
(5 UMRs)

3 coordinateur.ices:
Mathilde Carpentier
Julien Mozzicopacci
Alain Paris



Moi versus les *use case*:

Profile: first steps towards bioinformatics reproducibility

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

- data analysis \Rightarrow notebook
- code development \Rightarrow 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 community

- 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!

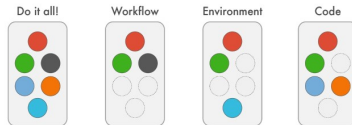
Schedule:

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

Travail sur cluster MNHN

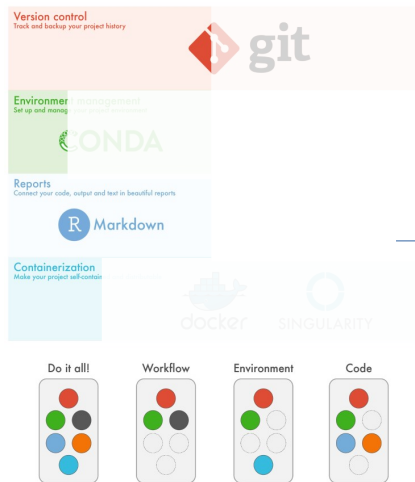
FAIR bioinfo 2022

La présentation du programme de la formation:



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

Avant la formation:

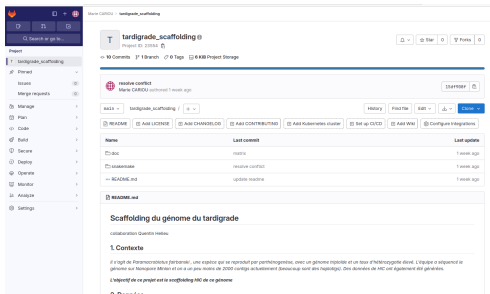


https://mbis-reproducible-research.readthedocs.io/en/course_2104/introduction/

Mise en oeuvre au quotidien

Projet exemple: scaffolding de génome avec des données HiC.

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



The screenshot shows a GitLab repository interface for a project named 'tardigrade_scaffolding'. The repository is owned by 'Marie CARROU' and is currently on the 'main' branch. The repository has 19 commits, 1 branch, 6 tags, and 6 KB of project storage. The repository is public and has a README file. The repository is currently in a state of 'resolved conflict'.

The repository contains a file named 'README.md'. The content of the README.md file is as follows:

```
Scfolding du génome du tardigrade

Collaboration Quentin Huetz

1. Contexte

Il s'agit de Paramacrobletus (parabarsi), une espèce qui se reproduit par parthénogenèse, avec un génome téploidé et un taux d'HiC/rozzette élevé. L'équipe a séquencé le génome sur Nanopore Minion et on a un jeu riche de 2000 cartés actuellement (Benoisjean sont très happy!). Des données de HiC ont également été générées.

L'objectif de ce projet est le scaffolding HiC de ce génome.
```


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 2.76 KiB Edit Replace Delete
1 # module load userspace/tr17.10
2 # module load bioinfo
3 # module load python/conda
4 # source activate ensemble_smk
5
6 #msub snakemake --cluster "dbatch" --jobs=6 --cores=12 -s scaffolding_tardi.smk --use-conda --directory /mnt/beegfs/ncariou/TARDIGRADE_results/ >
7
8 OUT = "/mnt/beegfs/ncariou/TARDIGRADE_results/"
9 DATA = "/mnt/beegfs/ncariou/TARDIGRADE_Data/"
10
11 REF= DATA + "Scaffolding/Pfa1_combined_asm.fa"
12
13 NIC=["Pfa1Pool1C181"]
14 ARIMAPATH = "/trinity/home/ncariou/mapping_pipeline"
15
16 rule all:
17     input:
18         hic1= expand(OUT + "input_eval/{hic}_1_fastqc.html", hic=NIC),
19         hic2= expand(OUT + "input_eval/{hic}_2_fastqc.html", hic=NIC),
20         ref= OUT + "input_eval/report.pdf",
21         #hic_fwdp= expand(OUT + "trim/{hic}_paired_1_fastqc.gz", hic=NIC),
22         #hic_revtp= expand(OUT + "trim/{hic}_paired_2_fastqc.gz", hic=NIC),
23         hic_fwdpfc= expand(OUT + "trim/{hic}_paired_1_fastqc.html", hic=NIC),
24         hic_revtpfc= expand(OUT + "trim/{hic}_paired_2_fastqc.html", hic=NIC),
25         contigs= REF + ".fa1",
26         PE= OUT + "arima/deduplicatedmap/tardi_rep1.bam",
27         scaf= OUT + "yaha/scaffolds_final.fa",
28         out_query= OUT + "yaha/query/report.pdf",
29         hicmap_ctpe= OUT + "hicstuff/abs_fragments_contacts_weighted.txt",
30         hicmap_scaf= OUT + "yaha/hicstuff/abs_fragments_contacts_weighted.txt",
31
32 rule fastqc_hic:
33     output:
34         hic1= OUT + "input_eval/{hic}_1_fastqc.html",
35         hic2= OUT + "input_eval/{hic}_2_fastqc.html",
36         ref= OUT + "input_eval/report.pdf",
37         #hic_fwdp= OUT + "trim/{hic}_paired_1_fastqc.gz",
38         #hic_revtp= OUT + "trim/{hic}_paired_2_fastqc.gz",
39         hic_fwdpfc= OUT + "trim/{hic}_paired_1_fastqc.html",
40         hic_revtpfc= OUT + "trim/{hic}_paired_2_fastqc.html",
41         contigs= REF + ".fa1",
42         PE= OUT + "arima/deduplicatedmap/tardi_rep1.bam",
43         scaf= OUT + "yaha/scaffolds_final.fa",
44         out_query= OUT + "yaha/query/report.pdf",
45         hicmap_ctpe= OUT + "hicstuff/abs_fragments_contacts_weighted.txt",
46         hicmap_scaf= OUT + "yaha/hicstuff/abs_fragments_contacts_weighted.txt"
```

Mise en oeuvre au quotidien

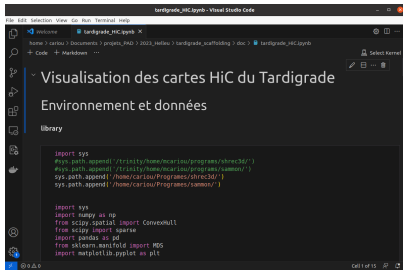
Changement majeur dans ma pratique: snakemake

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

Mais aussi:

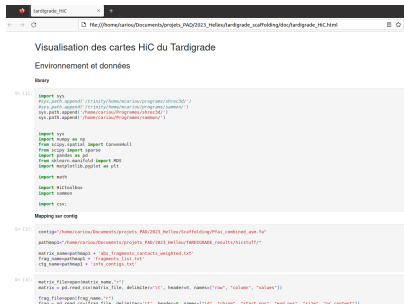
- des notebook python sans Jupyterlab
- des Docker

Des notebook



```
import sys
sys.path.append('/frinity/home/mcarisou/programes/shrec3d/')
sys.path.append('/frinity/home/mcarisou/programes/sammon/')
sys.path.append('/home/mcarisou/Programes/shrec3d/')
sys.path.append('/home/mcarisou/Programes/sammon/')

import sys
import numpy as np
from scipy.spatial import ConvexHull
from scipy import sparse
import pandas as pd
from sklearn.manifold import MDS
import matplotlib.pyplot as plt
```



```
import sys
sys.path.append('/frinity/home/mcarisou/programes/shrec3d/')
sys.path.append('/frinity/home/mcarisou/programes/sammon/')
sys.path.append('/home/mcarisou/Programes/shrec3d/')
sys.path.append('/home/mcarisou/Programes/sammon/')

import sys
import numpy as np
from scipy.spatial import ConvexHull
from scipy import sparse
import pandas as pd
from sklearn.manifold import MDS
import matplotlib.pyplot as plt

import math
import itertools
import os

import csv

Mapping sur config

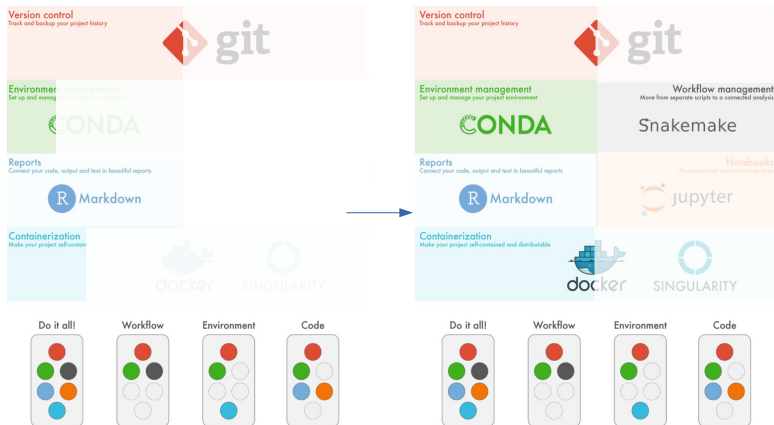
import sys
pathconfig = '/home/mcarisou/Documents/projects_PAG/2023_Hellou/tardigrade_scaffolding/HiC/tardigrade_HiC.html'
pathconfig = '/home/mcarisou/Documents/projects_PAG/2023_Hellou/tardigrade_scaffolding/HiC/tardigrade_HiC.html'

matrix_name=pathconfig + '_HiC_Fragments_contacts_weighted.txt'
Frag_name=pathconfig + '_Fragments_list.txt'
Frag_name=pathconfig + '_HiC_contacts.txt'

matrix = pd.read_csv(matrix_name, delimiter=';', header=0, names=['row', 'column', 'value'])
Frag_filenames=Frag_name.split('.')
Frag_filenames=[Frag_filenames[0].replace('.txt', '.bed') for Frag_filenames in Frag_filenames]
Frag_filenames=[Frag_filenames[0].replace('.txt', '.bed') for Frag_filenames in Frag_filenames]
```


FAIR bioinfo 2022

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 :)