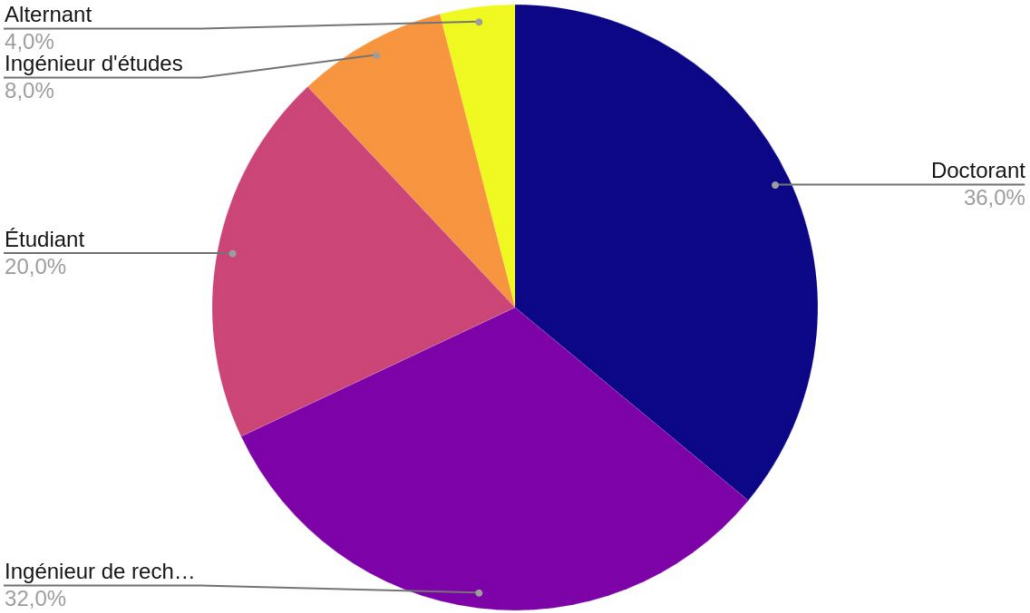
Good practices for reproducible bioinformatics data analysis

Sébastien Gradit, Elisabeth Hellec, Julien Fumey, Jérémy Rousseau, Benjamin Loire, Baptiste Imbert, Arthur Durante, Samuel Ortion, Ravy Leon Foun Lin, Maxime Ulysse Garcia, Savandara Besse

Participants



Program

1. Introduction: Reproducibility
2. Git - https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/01_git
3. Conda - https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/02_conda
4. Nextflow & Snakemake - https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/03_workflow
5. Docker - https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/04_docker
6. It's your turn to code - https://github.com/jebif/reprohackathon-jobim2024/tree/main/workflow

What does reproducibility means?

Reproducible research: Author provide all the necessary data and the computer codes to run the analysis again, re-creating the results.

Replicable research: A study that arrived at the same scientific finding as another study, collecting new data and completing new analyses.

Source: Introduction to reproducibility, Reproducibility 4 Everyone

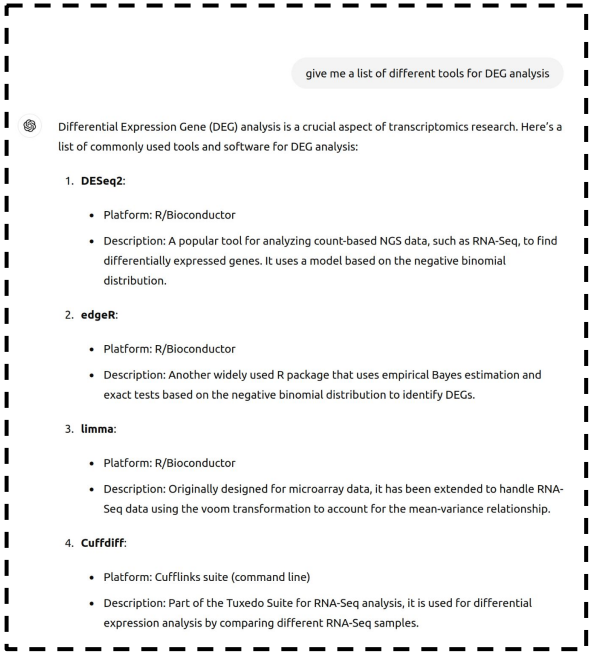
Differents modes of reproducibility

Methods	Same data	Different data
Same Methods	**Reproducibility**	**Replicability**
Different Methods	**Robustness**	**Generalizability**

Source: Introduction to reproducibility, Reproducibility 4 Everyone

3 good reasons to do reproducible methods (for a bioinformatician)

1. Why would you reinvent the wheel?



give me a list of different tools for DEG analysis

Differential Expression Gene (DEG) analysis is a crucial aspect of transcriptomics research. Here's a list of commonly used tools and software for DEG analysis:

1. **DESeq2**:
   - Platform: R/Bioconductor
   - Description: A popular tool for analyzing count-based NGS data, such as RNA-Seq, to find differentially expressed genes. It uses a model based on the negative binomial distribution.
2. **edgeR**:
   - Platform: R/Bioconductor
   - Description: Another widely used R package that uses empirical Bayes estimation and exact tests based on the negative binomial distribution to identify DEGs.
3. **limma**:
   - Platform: R/Bioconductor
   - Description: Originally designed for microarray data, it has been extended to handle RNA-Seq data using the voom transformation to account for the mean-variance relationship.
4. **Cuffdiff**:
   - Platform: Cufflinks suite (command line)
   - Description: Part of the Tuxedo Suite for RNA-Seq analysis, it is used for differential expression analysis by comparing different RNA-Seq samples.

- What type of data do you have?
- How was designed the experiments from where come from your data?
- Advantages / Limits

2. *We do love benchmarking*

METHOD	TP	TN	FP	FN	SENSIIVITY (%)	SPECIFICITY (%)	Q (%)	MCC
Aggrescan [33]	445	5210	1363	813	35.37	79.26	57.32	0.13
AmyloidMutants [41]	524	4924	1649	734	41.65	74.91	58.28	0.14
Amyloidogenic Pattern [23]	176	6208	365	1082	13.99	94.45	54.22	0.12
Average Packing Density [30]	361	5529	1044	897	28.70	84.12	56.41	0.12
Beta-strand contiguity [32]	417	5628	945	841	33.15	85.62	59.39	0.18
Hexapeptide Conf. Energy [27]	494	5172	1401	764	39.27	78.69	58.98	0.15
NetCSSP [16]	645	4287	2286	613	51.27	65.22	58.25	0.12
Pafig [37]	651	4695	1878	607	51.75	71.43	61.59	0.18
SecStr [17]	143	6205	368	1115	11.37	94.40	52.88	0.09
Tango [24]	172	6282	291	1086	13.67	95.57	54.62	0.14
Waltz [39]	710	4300	2273	548	56.44	65.42	60.93	0.16
AMYLPRED [42]	415	5668	905	843	32.99	86.23	59.61	0.19
AMYLPRED2	494	5553	1020	764	39.27	84.48	61.88	0.22

True/false positives (TP, FP) and true/false negatives (TN, FN) for each method were counted on a per residue basis. Sensitivity is measured as TP/(TP + FN), specificity as TN/(TN + FP), Q is calculated as (Sensitivity + Specificity)/2 and Matthews Correlation Coefficient (MCC) as (TP * TN – FP * FN)/$\sqrt{}$((TN + FN) * (TN + FP) * (TP + FN) * (TP + FP)).

doi:10.1371/journal.pone.0054175.t001

(Tsolis et al. PLoS One 2013)

3. We will be lost without code comments and documentation

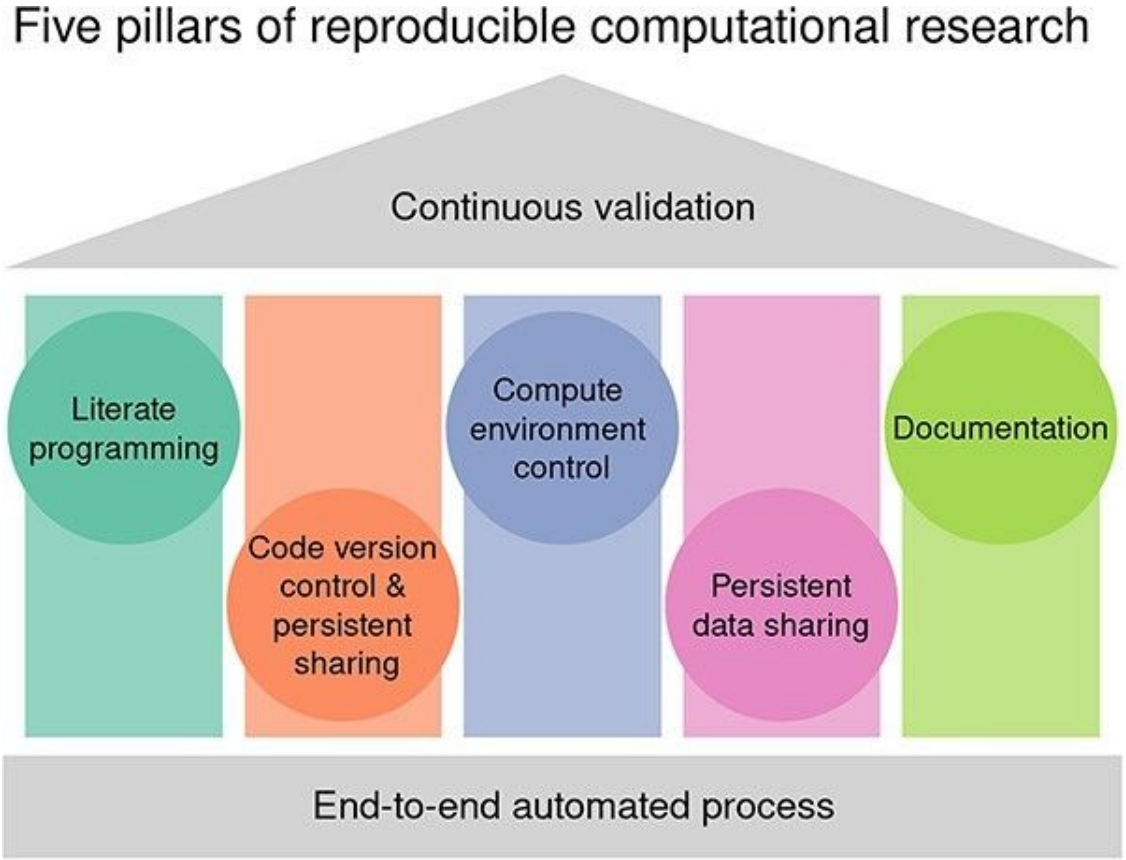
- *"It doesn't matter how good your software is, because **if the documentation is not good enough, people will not use it.**"* — Daniele Procida

- Or as **Yoda** says - credit to Chat GPT -

Documentation important it is. Share knowledge, it does. Without documentation, lost we become. Understand code, hard it will be. Communicate clearly, teams must. Future developers, thank you they will. Maintainable and scalable, your projects become. A guide, documentation is. Follow it, one must.

More about this:

- What nobody tells you about documentation, Daniele Procida (PyCon Australia, 2017)
- Documenting Python Code: A Complete Guide, James Mertz, Real Python



(From Ziemann et al. Briefings in Bioinformatics, 2023)

Data

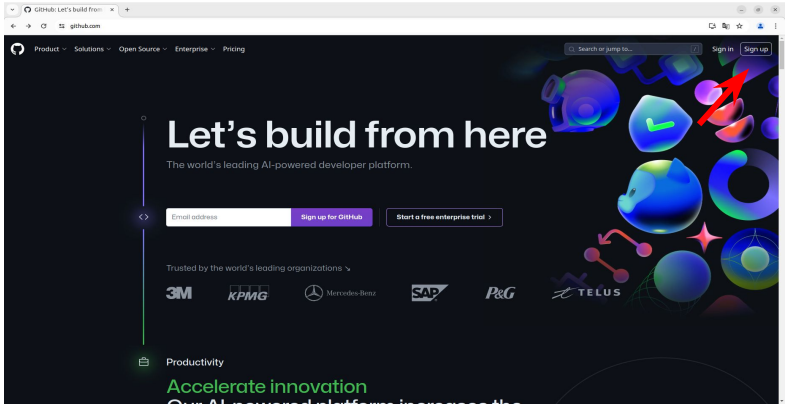
git clone git@github.com:jebif/reprohackathon-jobim2024.git

Git

Benjamin Loire

Git

- Git :
  - Created by Linus Torvalds in 2005
  - Decentralized versioning software (free and open source)
- GitHub :
  - Web hosting and software development management service
  - Owner: Microsoft
- Gitlab :
  - Same as GitHub
  - No owner



Git

IT'S UP TO YOU

Go to GitHub

https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/01_git

Conda

XXXX

Conda

- **Conda :**
  - **Characteristics : open-source, cross-platform, package manager and environment management system**
  - **Developed to solve package management challenges (Python and R)**
- **Anaconda :**
  - **Distribution of the Python and R programming languages for scientific computing**
  - **Simplify package management and deployment**
- **Miniconda :**
  - **Small version of Anaconda**
- **Mamba**
  - **Faster conda solver, now integrated by default**

https://en.wikipedia.org/wiki/Conda_(package_manager) & https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)

Conda

IT'S UP TO YOU

Go to GitHub

https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/02_conda

Nextflow

Maxime Ulysse Garcia

Snakemake

Sébastien Gradit

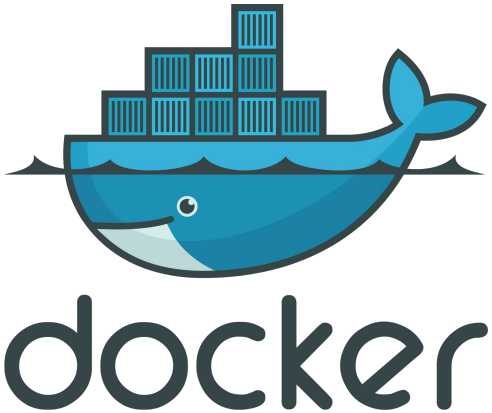
Nextflow & Snakemake

IT'S UP TO YOU

Go to GitHub

https://github.com/jebif/reprohackathon-jobim2024/tree/main/resources/03_workflow

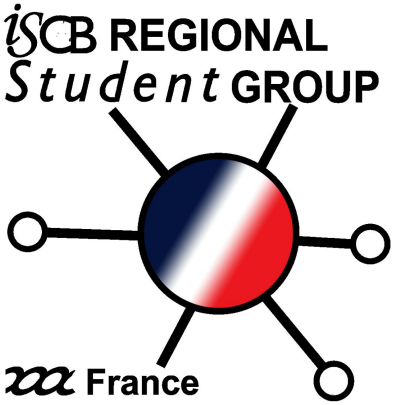
Docker



Benjamin Loire & Jérémy Rousseau

Project

What's JeBiF ?



XXXX

Société Française de Bioinformatique

XXXX

THE END !

Thanks to Marion Aguirrebengoa and Virginie Jouffret