

FAIR Bioinfo 2022

Best practice in your bioinformatic projects



P. Marin, M. Hiriart, P. Ruiz & N. Goué
aubi@uca.fr

Université Clermont Auvergne, AuBi, Mésocentre

15 mars 2023



UNIVERSITÉ
Clermont Auvergne



. This work is based on the IFB and I2BC formation offer

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful.

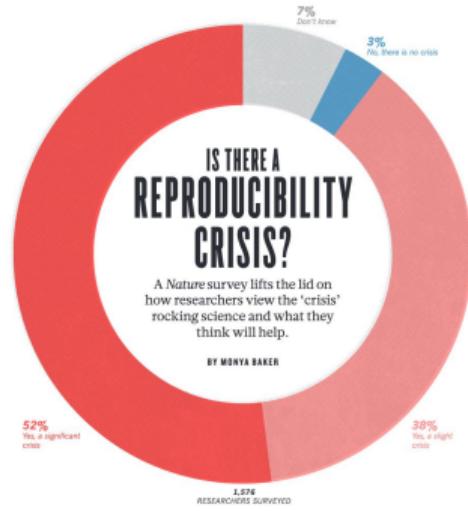
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that *c* relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta R)/(1 + \beta R)$.

Crisis elements

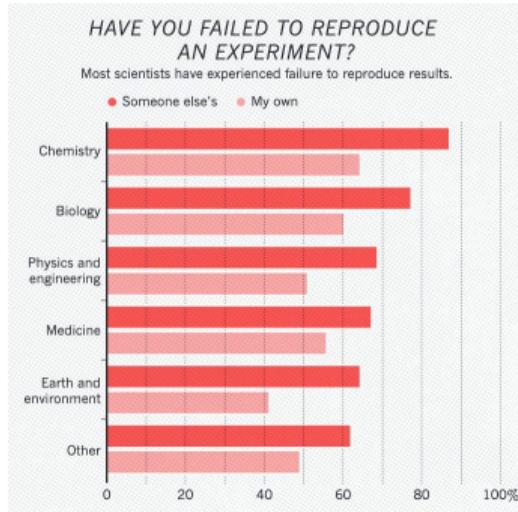
- Highlighted around 2005
- Since 2010 more articles related to the non reproducibility
- Medicine is one of the most impacted discipline

Reproducibility crisis

2016

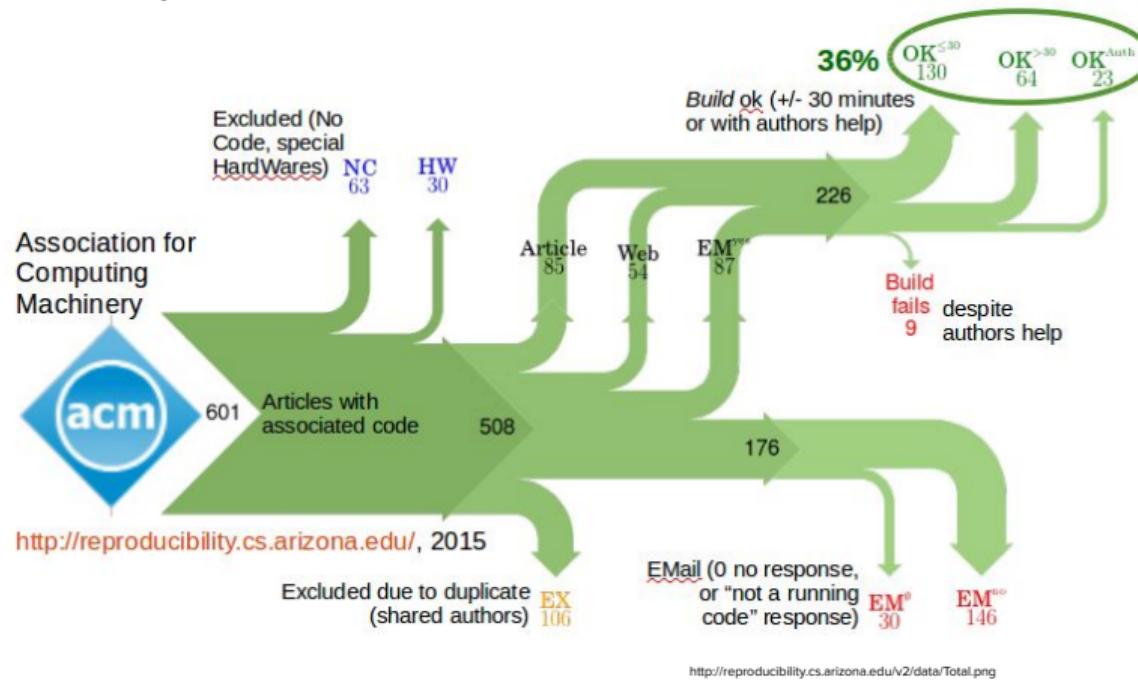


Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). <https://doi.org/10.1038/533452a>



4

Also in computer sciences



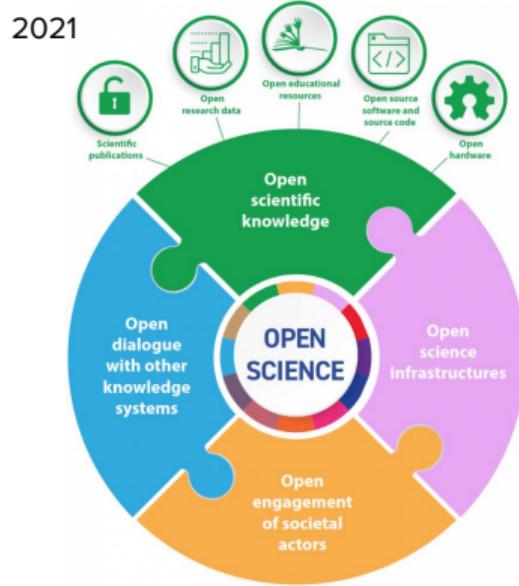
Long term negative impact of retracted papers

Article	Year of retraction	Citing Articles before retraction	Citing Articles after retraction	Total cites (journals indexed by Web of Science)
1. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N ENGL J MED; APR 2013 . Estruch R, et al.	2018	1919	816	2735
2. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. LANCET; FEB 28 1998 . Wakefield AJ, et al.	2010	642	867	1509
3. Visfatin: A protein secreted by visceral fat that mimics the effects of insulin. SCIENCE; JAN 2005 . Fukuhara A, et al.	2007	232	1192	1424
4. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. PLANT J; MAR 2003 . Voinnet O, et al.	2015	896	375	1271
5. Lysyl oxidase is essential for hypoxia-induced metastasis. NATURE; APR 2006 . Erler JT, et al.	2020	977	81	1058

Retraction Watch : Top 10 most highly cited retracted papers
<https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/>

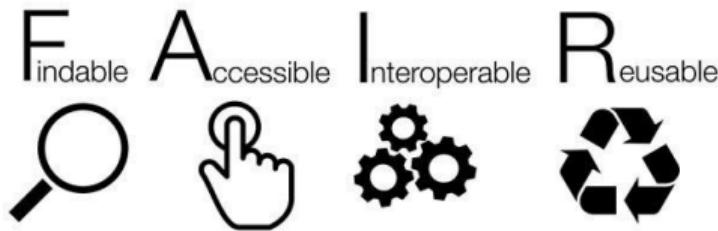
6

A way out: Open science and FAIR principles



Graphic on page 11. [UNESCO Recommendation on Open Science](#). CC BY IGO 3.0 C. Green

2016



Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
<https://doi.org/10.1038/sdata.2016.18>

FAIR history

- Born in 2016 with *The FAIR Guiding Principles for scientific data management and stewardship*
- How to build, stock, share, use and publish data
- Make criteria to better use our data

. <https://doi.org/10.1038/sdata.2016.18>

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

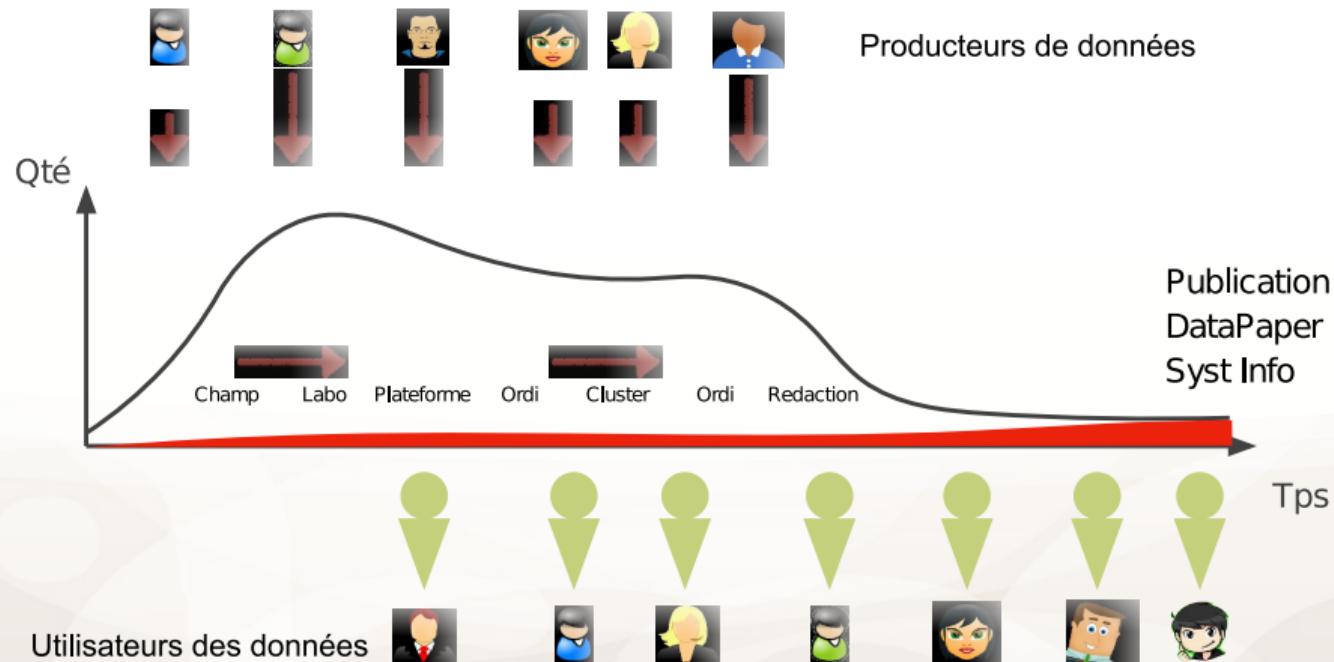
. <https://doi.org/10.1038/sdata.2016.18>

Apply FAIR TO

■ Your DATA

- Data lifecycle
- Data Management Plan (DMP)
- Metadata
- Data storage

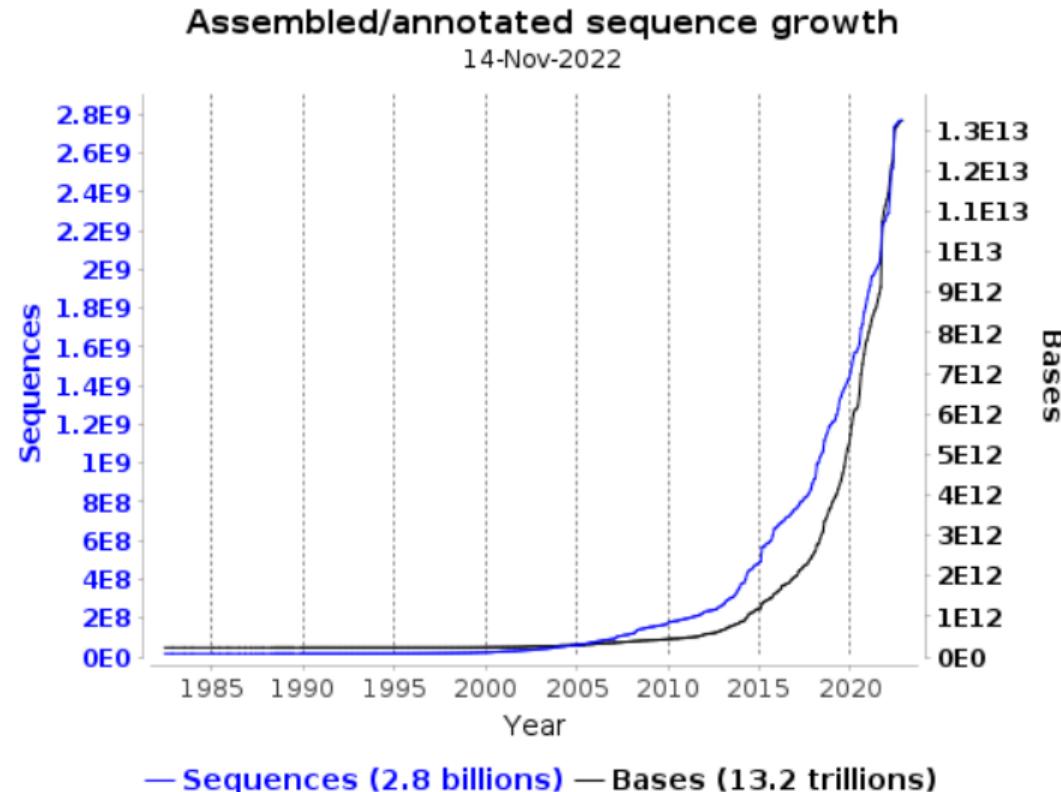
Focus on the data lifecycle



Focus on the data lifecycle



Focus on the data lifecycle



<https://www.ebi.ac.uk/ena/browser/about/statistics>

Focus on the data exchange

Transfert de vos données de recherche



Comment transmettre vos données ?





Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Apply FAIR TO

- Your DATA

- Data lifecycle
- Data Management Plan (DMP)
- Metadata
- Data storage

- Your scripts, environment...

- Objective of this training

FAIR principles

F
indable



By 維基小霸王 - Own work, CC BY-SA 4.0.
<https://commons.wikimedia.org/w/index.php?curid=88894774>

PID
Repository

8

<https://doi.org/10.1038/sdata.2016.18>

404

This is not the
web page you
are looking for.



To be Findable

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

FAIR principles

F
indable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

A
ccessible



<https://nitsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

PID
Repository

Protocols
(free, open, auth.)

9
<https://doi.org/10.1038/sdata.2016.18>

To be Accessible

- (meta)data are retrievable by their identifier using a standardized communication protocol
- the protocol is open, free, and universally implementable
- the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data are no longer available

FAIR principles

Findable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

Accessible



<https://nitsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

Interoperable



By Unknown author - Popular Science Monthly Volume 88, Public Domain
<https://commons.wikimedia.org/w/index.php?curid=22614407>

PID
Repository

Protocols
(free, open, auth.)

Standards
(format, vocabulary)

10
<https://doi.org/10.1038/sdata.2016.18>

To be Interoperable

- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data

FAIR principles

Findable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

Accessible



<https://nillsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

Interoperable



By Unknown author - Popular Science Monthly Volume 88, Public Domain
<https://commons.wikimedia.org/w/index.php?curid=22614407>

Reusable



By Sun Ladder - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=5746428>

PID
Repository

Protocols
(free, open, auth.)

Standards
(format, vocabulary)

Metadata
License
Origin

11
<https://doi.org/10.1038/sdata.2016.18>

To be Reusable

- meta(data) are richly described with a plurality of accurate and relevant attributes
- (meta)data are released with a clear and accessible data usage license
- (meta)data are associated with detailed provenance
- (meta)data meet domain-relevant community standard

FAIR tools

Findable



Accessible



Interoperable



Reusable



Data

Software
and
analyses

[FAIRsharing.org](https://fairsharing.org)
standards, databases, policies



CeCILL

A complete integrated FAIR environment

The Galaxy project



A complete integrated FAIR environment

The Galaxy project

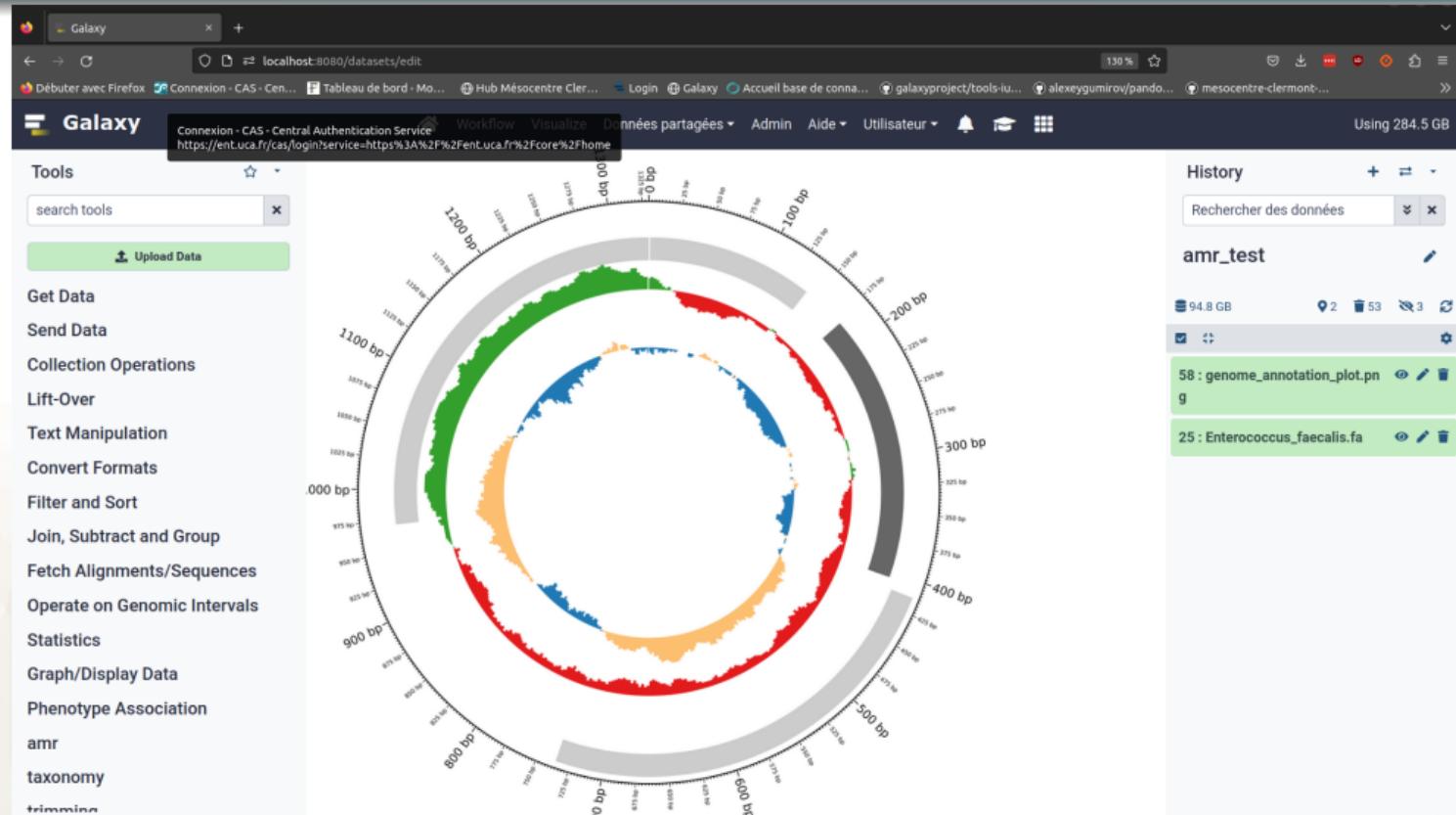
Galaxy

Galaxy is an open-source platform for FAIR data analysis that enables users to :

- Use tools from various domains (that can be plugged into workflows) through its graphical web interface.
- Run code in interactive environments (RStudio, Jupyter...) along with other tools or workflows.
- Manage data by sharing and publishing results, workflows, and visualizations.
- Ensure reproducibility by capturing the necessary information to repeat and understand data analyses.

A complete integrated FAIR environment

The Galaxy project



A complete integrated FAIR environment

The Galaxy project

The screenshot shows the Galaxy web interface running on localhost:8080. The main panel displays a workflow titled "staramr" which scans genome assemblies against three databases: ResFinder, PlasmidFinder, and PointFinder. The workflow has a single input file, "25 : Enterococcus_faecalis.fa", listed under the "genomes" section. Below the workflow, there are several configuration sliders for BLAST parameters:

- Percent identity threshold for BLAST: 98.0
- Percent length overlap of BLAST hit for ResFinder database: 60.0
- Percent length overlap of BLAST hit for PointFinder database: 95.0
- Percent length overlap of BLAST hit for PlasmidFinder database: 60.0

On the right side of the interface, there is a "History" panel titled "amr_test" containing a file named "58 : genome_annotation_plot.png". The bottom right corner of the slide features the text "Université Clermont Auvergne".

A complete integrated FAIR environment

The Galaxy project

The screenshot shows the Galaxy web interface running a workflow titled "Workflow: abromics_SR_PE_workflow". The workflow consists of nine steps:

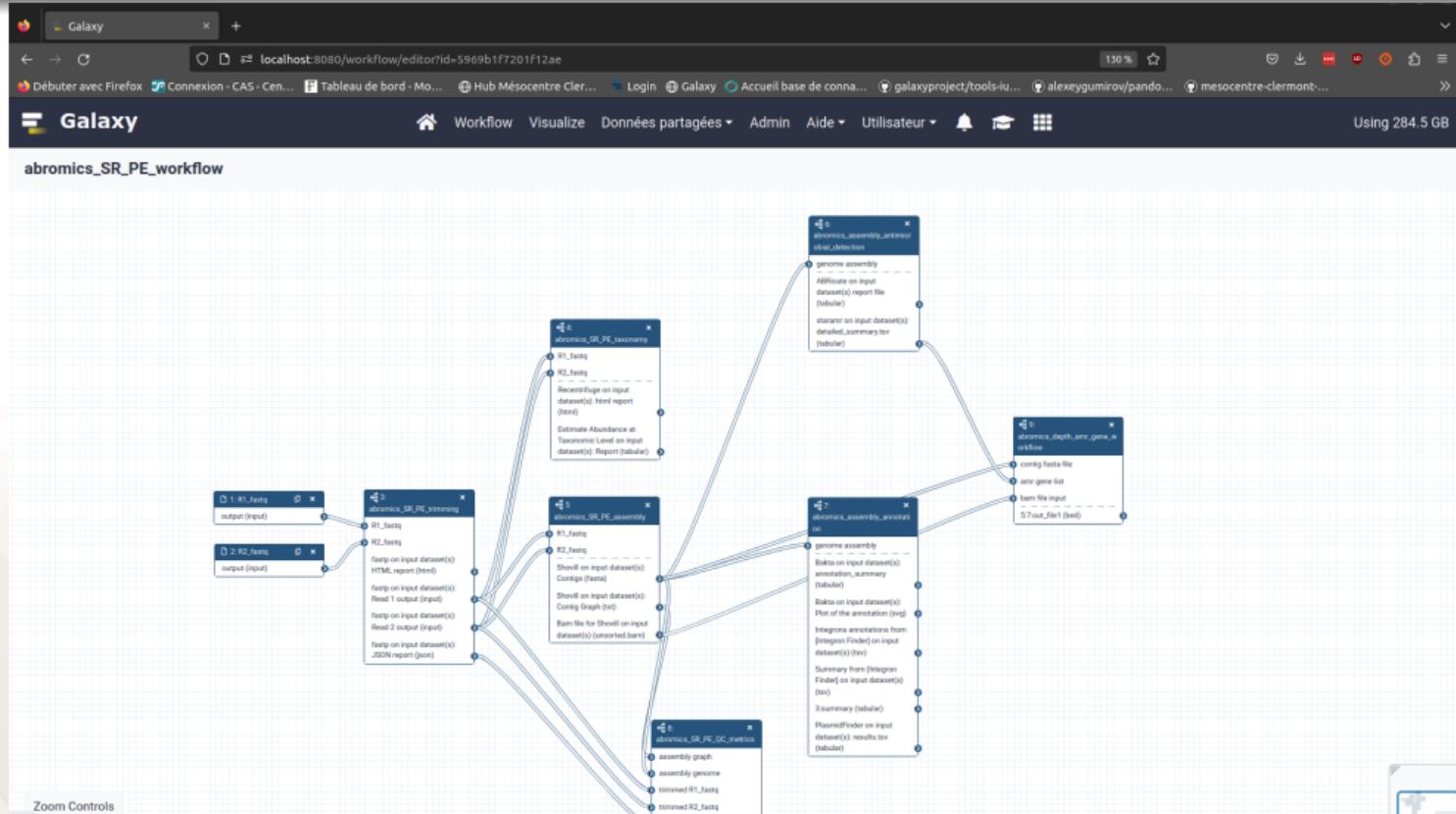
- 1: R1_fastq
- 2: R2_fastq
- 3: abromics_SR_PE_trimming
- 4: abromics_SR_PE_taxonomy
- 5: abromics_SR_PE_assembly
- 6: abromics_assembly_antimicrobial_detection
- 7: abromics_assembly_annotation
- 8: abromics_SR_PE_QC_metrics
- 9: abromics_depth_amr_gene_workflow

The "History" panel on the right shows a single history named "amr_test" containing three items:

- 94.8 GB (9 items)
- 58 : genome_annotation_plot.png (1 item)
- 25 : Enterococcus_faecalis.fa (1 item)

A complete integrated FAIR environment

The Galaxy project



How to integrate FAIR concepts in my work ?

Some tools for reproducible research



FAIR session with AuBi

Objectives

- Discover FAIR practices
- Discover tools for best practices
- Learn tools and best practices

FAIR session with AuBi

Objectives

- Discover FAIR practices
- Discover tools for best practices
- Learn tools and best practices
- 5 sessions for courses and practices
 - Day 1 : Introduction to FAIR and building training environment
 - Day 2 : Code versioning with Git
 - Day 3 : Environment managment using conda and docker/singularity
 - Day 4 : Workflow managment using snakemake
 - Day 5 : Documentation using Rmarkdown or Jupyter

Contents

- Introduction to FAIR practices

Contents

- Introduction to FAIR practices
- Code control using Git 

 - Git environment
 - Gitlab and Github  

Contents

- Introduction to FAIR practices
- Code control using Git 

 - Git environment
 - Gitlab and Github 

- Encapsulation process
 - Conda environment and packages use 
 - Containers as docker & singularity 
 - Reproducible workflow using snakemake 

Contents

- Introduction to FAIR practices
- Code control using Git 

 - Git environment
 - Gitlab and Github 

- Encapsulation process
 - Conda environment and packages use 
 - Containers as docker & singularity 
 - Reproducible workflow using snakemake 
- Literate programming and documentation
 - Markdown syntax 
 - Rmarkdown for R 
 - Jupyterlab for Python 

Good practice on...

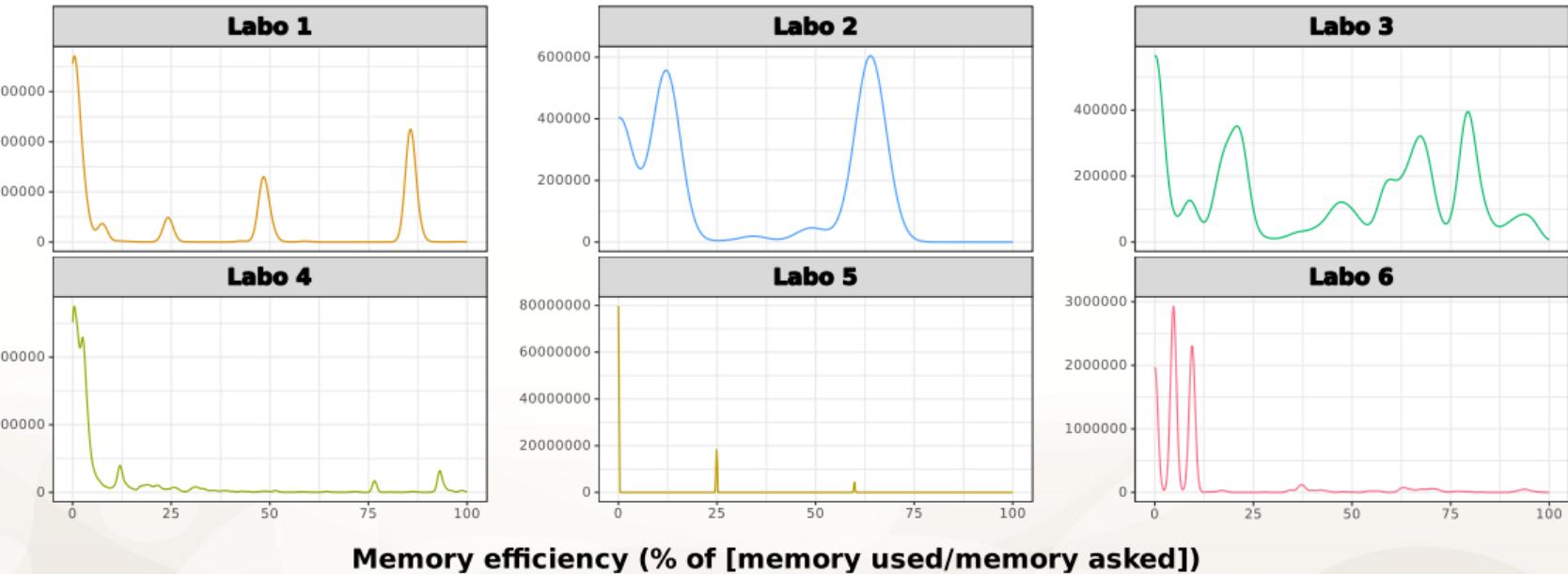
- Must have good practices on data usage
- Must have good practices on tool & code usage

Then...

Must have good practices on computing ressources usage ?

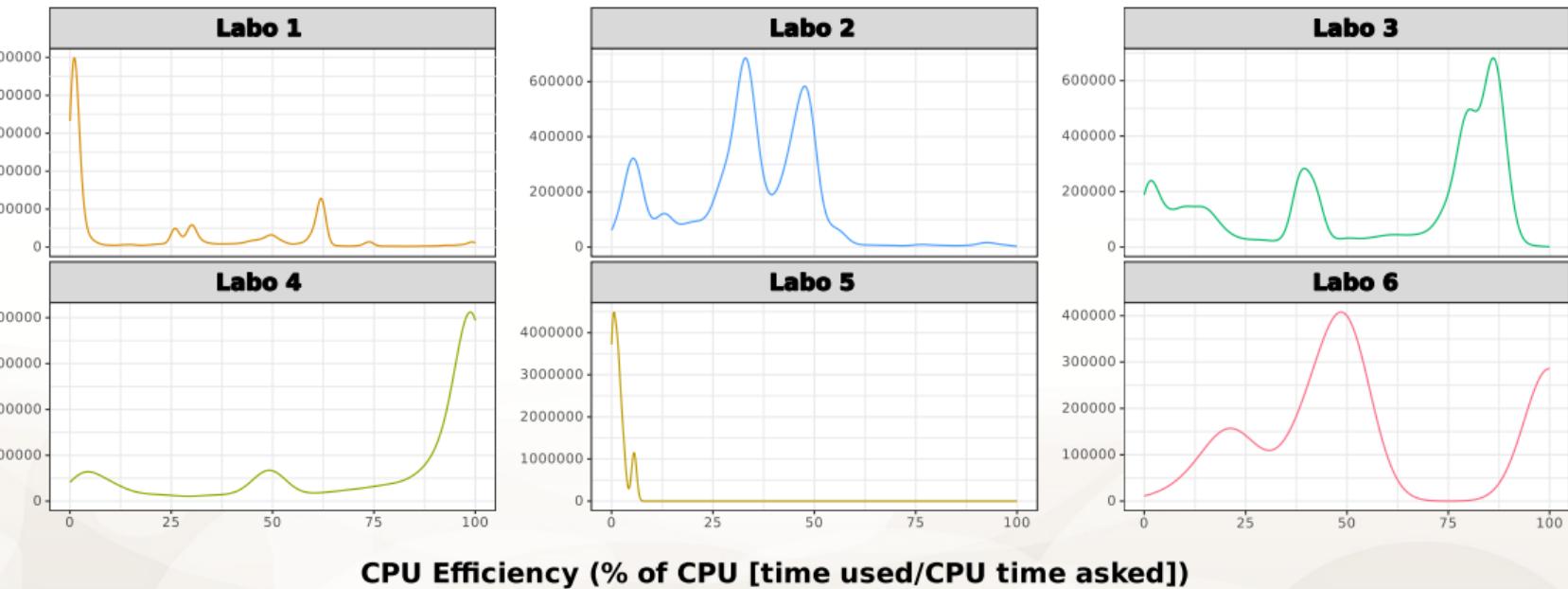
Efficiency of jobs on used RAM per account

Count of jobs



Efficiency of jobs on used CPU per account

Count of jobs



Biosphere a virtual environment for training

The screenshot shows the login interface for the Biosphere platform. At the top left is the IFB logo (Institut Français de Bioinformatique) with the text "Biosphere". To its right are links for "RAINBio", "myVM", and "DATA". On the far right are "Support", "[fr]", and "Se connecter". Below these, a "SE CONNECTER" button is visible. The main area features the IFB logo and the text "INSTITUT FRANÇAIS DE BIOINFORMATIQUE". A sub-instruction "Utilisez vos identifiants de votre organisme (CNRS, INRAE, Inserm, Universités...)" is followed by a "Se connecter" button. A note at the bottom states: "Nous utilisons la fédération d'identité européenne eduGAIN. Si votre organisme académique n'est pas dans cette fédération, vous pouvez utiliser un compte local avec votre adresse professionnelle.".



Biosphere a virtual environment for training

The screenshot shows the IFB Biosphere landing page. At the top, there's a navigation bar with the IFB logo, 'Biosphere', 'RAINBio', 'myVM', and 'DATA' buttons. On the right, there are 'Support' and a user profile icon. Below the header, a large blue banner says 'WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES'. Underneath, a text block explains what Biosphere offers: "French Institute of Bioinformatics (IFB) provides life scientists with a federation of clouds, Biosphere, and bioinformatics cloud services to analyze life science data. Biosphere is used for scientific production in the life sciences, developments, and to support events like scientific training sessions, university courses, hackathons or workshops." A section titled 'BIOSPHERE FEATURES' lists various capabilities, each with a bullet point and a list of sub-points.

BIOSPHERE FEATURES

With IFB-Biosphere, you get:

- A unified user portal ([Biosphere portal](#)) to deploy all bioinformatics environments on all clouds
 - Single sign-on, with your academic credentials ([Sign in](#)).
- Pre-defined bioinformatics environments, available with an one-click deployment from the [RAINBio catalogue](#)
- Infrastructure as Code: most configurations rely on public git repositories available in [Biosphere Commons](#).
- 8 cloud sites with more than 10,000 vCPU and 40 TB RAM ([System status](#))
 - High-availability thanks to the different sites usable equally.
- Modular cloud environments:
 - Single virtual machine (VM) to bunch of VMs.
 - Usual VM: up to 64 vCPUs-250 GB RAM.
 - BigMemory VM: up to 3 TB RAM.
 - HighFrequency VM : up to 3.8GHz vCPU.
 - ManyCores VM: up to 255 vCPUs in a VM
- Bring-Your-Own-Tools
 - Admin rights in the VM (all apps).
 - Deploy with your own container image (some apps).
 - Configure some apps with your own git repository (Infrastructure as Code).
- Useful public biological reference databases
- On-demand resources (CPU, RAM, storage, IFB experts) to support:
 - Training events, university courses, scientific schools, workshops, hackathons.
 - Scientific projects.

BIOSPHERE PORTAL

The Biosphere portal provides high-level cloud interfaces:

Biosphere a virtual environment for training

IFB Biosphère RAINBio myVM DATA Support pierre.marin@uca.fr (edu) GAIN

RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD

Catalogue des appliances bioinformatiques dans le cloud, filtrez-les en utilisant les termes présents dans l'ontologie EDAM, ou en langage naturel.

App Store (59) Appliances Outils Topics

Galaxy bioconda, Docker, Galaxy portal Informatics, Bioinformatics, Comparative genomics, Functional genomics	AnalysesSV bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, SAMtools DNA polymorphism, Genetic variation, Genotyping experiments, GWAS analysis	ANF MetaBioDiv Bioconductor, DESeq2, devtools (R), ggplot2, gridExtra, phyloseq, r Bioinformatics, Computational biology, Data management, Taxonomy	Askomics AskOmics Data integration and warehousing, Data visualisation	BactComparativeGenomics ImageJ2, Jupyter, MACS2, Matlab, Nextflow, pandas, Prokka Imaging, Mathematics, Statistics and probability, CHP-seq, Workflows, Data archiving	Bacterial Genomics HMMER, Insygh, SGE - GridEngine, Ubuntu, Web interface Protein folds and structural domains, Sequence comparison, Sequence composition, comparative
Bioimage Bureau virtuel, Icy, ImageJ-Fiji, X2Go, Xfce Informatics, Data visualisation, OF, Imaging	BioPipes bioconda, cwltool, Docker, Nextflow, Snakemake Informatics, Bioinformatics, Workflows	bistor bioconda, Bowtie2, FastQC, Snakemake Bioinformatics, Genomics, Mapping, Sequence alignment, Data quality management	CentOS 7 Ansible, bioconda, Docker Informatics, Bioinformatics	CentOS 7 Desktop Ansible, bioconda, Bureau virtuel, Docker Informatics, Bioinformatics	CoursAnalysesNanoporeSG bandage, Jupyter Data architecture, analysis and design, Mathematics, Statistics and probability
Cytoscape Bureau virtuel, Cytoscape, X2Go, Xfce Bioinformatics, Data visualisation, Molecular interactions, Pathway	Debian 10 Ansible, bioconda, Docker Bioinformatics, Informatics	Debian 11 Ansible, bioconda, Docker Bioinformatics, Informatics	DRomics Bioconductor, DESeq2, DRomics, R - base, RStudio, Shiny Bioinformatics, Computational biology, Data management, Taxonomy	DRomicsInterpreter Bioconductor, DESeq2, DRomics, R - base, RStudio, Shiny Bioinformatics, Computational biology, Data management, Taxonomy	EBAME-2022 oDNA BAMtools, BEDTools, bioconda, Bowtie2, metaDMG-cpp, python3, R Data management, Bioinformatics, Sequence alignment, Computer science, Sequencing
EBAME-2022 MetaTOR Bowtie2, CheckM, hicstrix, Pairix, SAMtools Sequence alignment, Phylogenetics, Population genomics	EBAME-Anvio Anvio	EBAME-Quince bam-readcount, BEDTools, BWA, CONCOCT, DESeq2, Diamond, FastQC Transcriptomics, Informatics	ETBII AnalyseMultivariée BioCStyle, Butcher, clusterProfiler, ComplexHeatmap, DESeq2, DT Literature and language, Transcriptomics	ETBII Réseaux BINGO (Cytoscape), Bureau virtuel, compositions (R), Cytoscape, igraph, Networkx Transcriptomics, Sequencing	formation_CIRI DESeq2, FastQC, HISAT2, RStudio, Trimomatic Transcriptomics, Whole genome sequencing

Biosphere a virtual environment for training



Biosphere

RAINBio myVM DATA

[Support](#)

VOS PARAMÈTRES

Informations personnelles

Adresse électronique	pierre.marin@uca.fr
Affiliation fournie par la fédération d'identité	
Prénom	Pierre
Nom	MARIN
Ville et code postal	63170
Formation initiale (optionnel)	<ul style="list-style-type: none"> • Biology • Computer Science
Pubkey	<ul style="list-style-type: none"> • ssh-rsa AAAAB3NzaC1yc2EAAAQABAAQDXgGQYRHqk4klU6XoeNqYVFqL14F5WkO3U2xSCDDQW8MpVQTCRqbzTq02GpIMF6014Q0gnC5-W-bTTJkZgnsigSziViug31MOdviZH8oWKXW6c0TZ1InMr0WjbgUBdB7RApTmHL4Laihvz/Ed/IASOTbKAQMLBtm4k3uL7l2RvFMdZcYVTKSbvcilEGShw8OHOHRicNEW3giiCIIUFSTZ5KEZ99/4KgarJRNcIXybzza+mtYGETfIQhJkyRD+R6DNrhUc8aJasgGpDpPwT/9ARVs9/avoF7e0JMUUdxWT9D139EPe44ony/s3k93HK4mpqGV piemari@hpclingo1 • ssh-rsa AAAAB3NzaC1yc2EAAAQABAAQDXgGQCFgYeYkm6YwmhKbf+mcckwcqzJVBGvRTgZwfhGWeWwfLyJsgDM9fWV+4ZucxtXLrxja69d9r2309E+IdUDoFYRvBtLQhX9X472DqbZ2duuDHvPKtgYaAtK4rjhjih32VuxSq7HeffrV5csTTbPJhunCTE6MIGLNVGH5rLjCwI-WtDVvaHYCdse2:YRAJeFsqAHtubRQucE2EKpRqixzS5cg8q2T3kdiYjaquOvCvQBlzswfkNyvQldg38LMf4JDqOW26ls7WEtavMD1X3EKUEmxm+TldNoc1hnNjgg8XEN/h

Connecté.e en tant que pierre.marin@uca.fr

Langues

[en] English

[fr] Français

Paramètres

Groups

Quota

Se déconnecter

<https://biosphere-france.bioinformatics.fr/cloudweb/account/settings/>

SSH key use

The principle

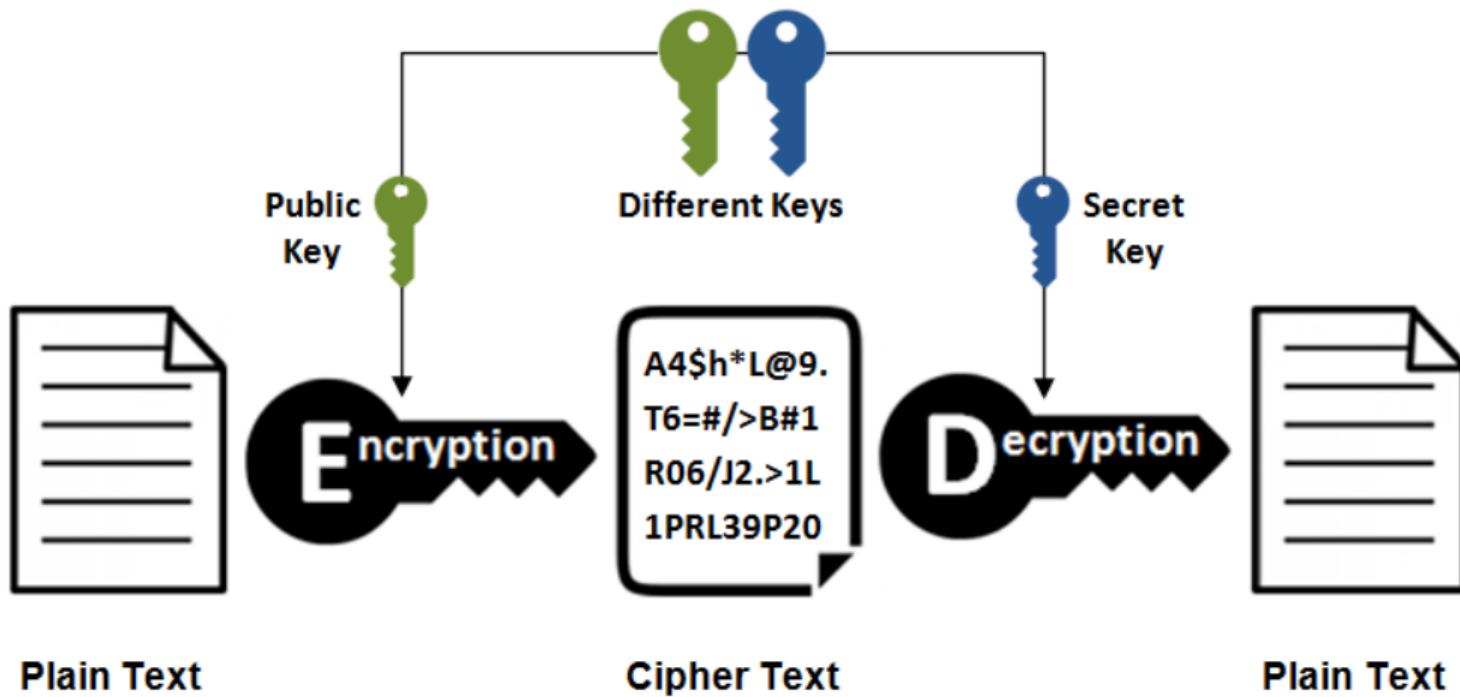
- Secure Shell (SSH)
- Securized protocol
- Encryption of informations between users
- Authentification without password
- Assymetric system based on keys

The keys

- Paired key based
- A public key for everyone to share
- A private key for myself (never share)

SSH key use

Asymmetric Encryption



SSH key use

```
$ ssh-keygen -t dsa
Generating public/private dsa key pair.
Enter file in which to save the key (/Users/tdd/.ssh/id_dsa): /Users/tdd/.ssh/id_dsa_
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /Users/tdd/.ssh/id_dsa_ga.
Your public key has been saved in /Users/tdd/.ssh/id_dsa_ga.pub.
The key fingerprint is:
65:31:7e:ee:49:3a:66:cd:92:7b:02:2b:bf:b3:1a:79 tdd@CodeMagic.local
The key's randomart image is:
+--[ DSA 1024]----+
|          o      |
|         . o     |
|          + .    |
|         o o     |
|        S o     |
|         .. B .   |
|        o EoB = |
```

SSH key use

```
$ cat ~/.ssh/id_dsa_ga
-----BEGIN DSA PRIVATE KEY-----
Proc-Type: 4,ENCRYPTED
DEK-Info: DES-EDE3-CBC,6ED59B013D8A361F

pB5eHHpvXxoz6i1jFzlKANv9W6SeHw664PV/1A90acR/Mw/ERQvTQKo3TaLaFhkb
...
NwhQFyxZZty2hn6xrv5UIAGTpjk+P2+waRmSno1Vg1x0epCp45kvFRv9AiXs0pt4
RgzPC5+a6kjPf8EtyozGoQ==
-----END DSA PRIVATE KEY-----
$ cat ~/.ssh/id_dsa_ga.pub
ssh-dss AAAAB3NzaC1kc3MAAACBAJhbQcZK81FMvpw7trbFj51Sqjd9nKBu2xkw/kvUAQ1PQPaiRL0iq92fx
```