



Université
de Paris



Unix : introduction et méthodes

DU-Bii 2020

Hélène Chiapello, Pierre Poulain

L'équipe



Hélène Chiapello



Benoist Laurent



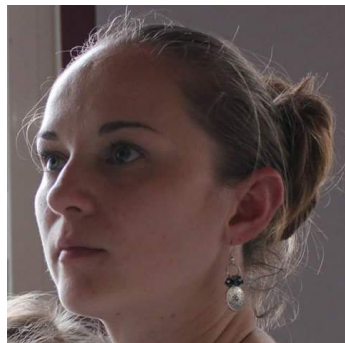
Jacques van Helden



Julien Seiler



Hubert Santuz



Sandra Dérozier



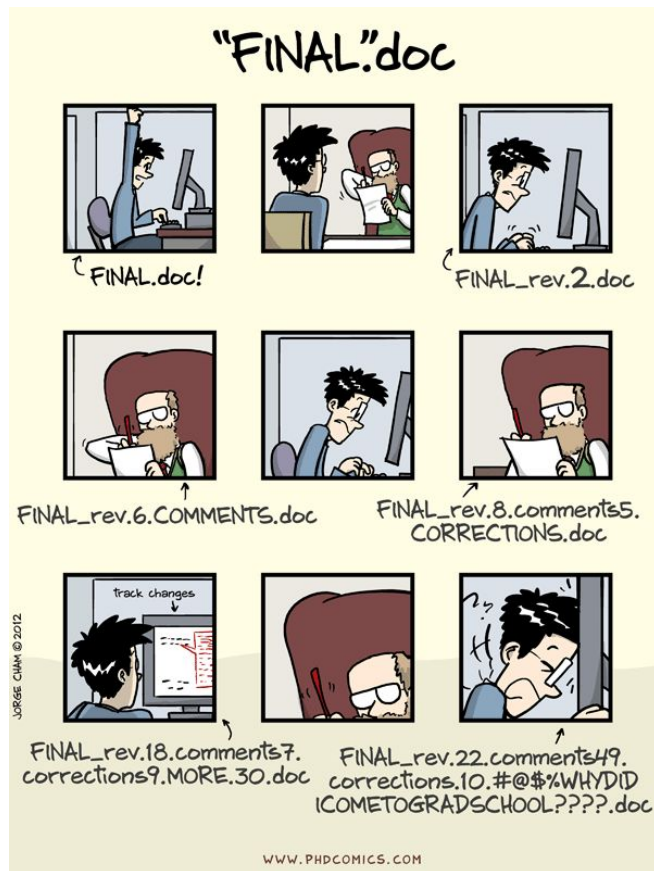
Pierre Poulain

Le planning

Jour	Horaire	Description
2 mars	14h30 - 17h30	Unix 1 : Premiers pas <i>Pierre Poulain, H��l��ne Chiapello, Benoist Laurent, Hubert Santuz</i> Manipuler des fichiers et des r��pertoires Manipuler les donn��es d��un fichier. ��diteur de texte. Bonnes pratiques en bioinfo.
3 mars	14h30 - 17h30	Unix 2 : Gestion de flux et extraction de donn��es <i>H��l��ne Chiapello, Benoist Laurent, Sandra D��rozier, Pierre Poulain</i> Encha��nement de commandes, compression et archivage
5 mars	9h00 - 12h00	Unix 3 : Utilisation des ressources de calcul IFB <i>Julien Seiler, Pierre Poulain, H��l��ne Chiapello, Benoist Laurent, Jacques van Helden</i> Pr��sentation et utilisation du cluster IFB. Utilisation de SLURM. Transfert de donn��es (ssh et scp)
10 mars	9h30 - 12h30	Unix 4 : Automatisation <i>Benoist Laurent, H��l��ne Chiapello, Pierre Poulain</i> Notion de variable, programmation Bash, calcul distribu��

Gestion des données et des logiciels

Gestion des données



Source : [PhD Comics](http://www.phdcomics.com)

Gestion des données : git / GitHub

- Garder une mémoire des modifications de fichiers
 - Travailler collaborativement
 - Partager des fichiers
-
- Git est un logiciel
 - GitHub est un site internet (une plateforme d'échange)

The screenshot shows a GitHub repository page for 'Diplôme Universitaire en Bioinformatique Intégrative' (DUBii). The repository is located in France and has 8 repositories, 24 people, 7 teams, and 1 project. The main content area lists four repositories:

- module-1-Environnement-Unix**: Environnement Unix. Languages: CSS, CC-BY-SA-4.0, 3 forks, 0 stars, 3 issues, 0 pull requests. Updated 1 hour ago.
- accueil**: Diplôme Universitaire en Bioinformatique Intégrative (DU-Bii). Languages: HTML, 2 forks, 0 stars, 0 issues, 1 pull request. Updated 2 hours ago.
- module-6-Integrative-Bioinformatics**: Integrative bioinformatics course of the Diplôme Universitaire en Bioinformatique Intégrative (DU-Bii). Languages: HTML, CC-BY-SA-4.0, 0 forks, 0 stars, 0 issues, 0 pull requests. Updated 8 days ago.
- module-3-Stat-R**: Analyse statistique avec R. Languages: HTML, CC-BY-SA-4.0, 4 forks, 0 stars, 0 issues, 1 pull request. Updated 16 days ago.

On the right side, there is a 'Top languages' section showing HTML, Shell, JavaScript, Python, and CSS. Below that is a 'People' section showing 24 contributors with their avatars. At the bottom right, there is an 'Invite someone' button.

Gestion des données : git / GitHub

OPEN ACCESS Freely available online

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve^{1,2*}, Anton Nekrutenko³, James Taylor⁴, Eivind Hovig^{1,5,6}

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, **2** Centre for Cancer Biomedicine, University of Oslo, Blindern, Oslo, Norway, **3** Department of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University Park, Pennsylvania, United States of America, **4** Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America, **5** Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway, **6** Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

Replication is the cornerstone of a cumulative science [1]. However, new tools and technologies, massive amounts of data, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts, as are increased pressures on scientists to advance their research [2]. As full replication of studies on independently collected data is often not feasible, there has recently been a call for reproducible research as an attainable minimum standard for assessing the value of scientific claims [3]. This requires that papers in experimental science describe the results and provide a sufficiently clear protocol to allow successful repetition and extension of analyses based on original data [4].

The importance of replication and reproducibility has recently been exemplified through studies showing that scientific papers commonly leave out experimental details essential for reproducibility [5], studies showing difficulties with replicating published experimental results [6], an increase in retracted papers [7], and through a high number of failing clinical trials [8]. This has led to the notion on how individual researchers, institutions, funding bodies, and journals can establish routines that increase transparency and reproducibility. In order to foster such aspects, it has been suggested that the scientific community needs to develop a "culture of reproducibility" for computational science, and to require it for published claims [3].

We want to emphasize that reproducibility is not only a moral responsibility with respect to the scientific field, but that a lack of reproducibility can also be a burden for you as an individual researcher. As an example, a good practice of reproducibility is necessary in order to allow previously developed methodology to be effectively applied on new data, or to allow reuse of code and results for new projects. In other words, good habits of reproducibility may actually turn out to be a time-saver in the longer run.

We further note that reproducibility is just as much about the habits that ensure reproducible research as the technologies that can make these processes efficient and realistic. Each of the following ten rules captures a specific aspect of reproducibility, and discusses what is needed in terms of information handling and tracking of procedures. If you are taking a bare-bones approach to bioinformatic analysis, i.e., running various custom scripts from the command line, you will probably need to handle each rule explicitly. If you are instead performing your analyses through an integrated framework (such as GenePattern [10], Galaxy [11], LON pipeline [12], or Taverna [13]), the system may already provide full or partial support for most of the rules. What is needed on your part is then merely the knowledge of how to exploit these existing possibilities.

In a pragmatic setting, with publication pressure and deadlines, one may face the need to make a trade-off between the ideals of reproducibility and the need to get the research out while it is still relevant. This trade-off becomes more important when considering that a large part of the analyses being tried out never end up yielding any results. However, frequently one will, with the wisdom of hindsight, recognize the missed opportunity to ensure reproducibility, as it may already be too late to take the necessary notes from memory (or at least much more difficult than to do it while underway). We believe that the rewards of reproducibility will compensate for the risk of having spent valuable time developing an annotated catalog of analyses that turned out as blind alleys.

As a minimal requirement, you should at least be able to reproduce the results yourself. This would satisfy the most basic requirements of sound research, allowing any substantial future questioning of the research to be met with a precise explanation. Although it may sound like a very weak requirement, even this level of reproducibility will often require a certain level of care in order to be met. There will for a given analysis be an exponential number of possible combinations of software versions, parameter values, pre-processing steps, and so on, meaning that a failure to take notes may make exact reproduction essentially impossible.

With this basic level of reproducibility in place, there is much more that can be wished for. An obvious extension is to go from a level where you can reproduce results in case of a critical situation to a level where you can practically and routinely reuse your previous work and increase your productivity. A second extension is to ensure that peers have a practical possibility of reproducing your results, which can lead to increased trust in, interest for, and citations of your work [6,14].

Citation: Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. *PLoS Comput Biol* 9(10): e1003285. doi:10.1371/journal.pcbi.1003285

Editor: Philip E. Bourne, University of California San Diego, United States of America

Published: October 24, 2013

Copyright: © 2013 Sandve et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors' laboratories are supported by US National Institutes of Health grants HG005113, HG005005, and HG005020 and US National Science Foundation grant DBI-0807013. Additional funding is provided in part by the Huck Institutes for the Life Sciences at Penn State, the Institute for Cyberscience at Penn State, and a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The funders had no role in the preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gks@iuh.no

PLOS Computational Biology | www.ploscompbiol.org 1 October 2013 | Volume 9 | Issue 10 | e1003285

When it comes to reproducible science, Git is code for success

And the key to its popularity is the online repository and social network, GitHub.

11 June 2018

Jeffrey Perkel



Lightcode/Getty

Sandve, PLOS Comput Biol, 2013
DOI 10.1371/journal.pcbi.1003285

J. Perkel, Nature Index, 2018

Gestion des données : git / GitHub

Débuter avec Git et Github en 30 min



<https://www.youtube.com/watch?v=hPfgekYUKgk>

La capsule, 2017

D'autres ressources :

- <https://cupnet.net/git-github/>
- <https://swcarpentry.github.io/git-novice/>


Gestion des logiciels

How to get a bioinformatics headache

1. See tweet about new published tool
2. Read abstract - sounds awesome!
3. Fail to find link to source code - eventually Google it
4. Attempt to compile and install it
5. Google for 30 min for fixes
6. Finally get it built
7. Run it on tiny data set
8. Get a vague error
9. Delete and never revisit it again



Gestion des logiciels : (bio)conda !

 MENU ▾

Correspondence | Published: 02 July 2018

Bioconda: sustainable and comprehensive software distribution for the life sciences

Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, Johannes Köster ✉ & The Bioconda Team

Nature Methods 15, 475–476(2018) | [Cite this article](#)

4061 Accesses | 82 Citations | 249 Altmetric | [Metrics](#)

Editorial | [Open Access](#) | Published: 27 February 2019

Improving the usability and archival stability of bioinformatics software

[Sergei Mangul](#) ✉, [Lana S. Martin](#), [Eleazar Eskin](#) & [Ran Blekhman](#)

Genome Biology 20, Article number: 47 (2019) | [Cite this article](#)

3099 Accesses | 9 Citations | 49 Altmetric | [Metrics](#)



PERSPECTIVE

Challenges and recommendations to improve the installability and archival stability of omics computational tools

Sergei Mangul^{1,2*}, Thiago Mosquero^{2*}, Richard J. Abdill¹, Dat Duong¹, Keith Mitchell¹, Varuni Sarwal¹, Brian Hill¹, Jaqueline Brito⁵, Russell Jared Littman¹, Benjamin Statz¹, Angela Ka-Mei Lam¹, Gargi Dayama³, Laura Grieneisen⁴, Lana S. Martin², Jonathan Flint⁶, Eleazar Eskin^{1,7}, Ran Blekhman^{3,8}

 Check for updates

 OPEN ACCESS

Citation: Mangul S, Mosquero T, Abdill RJ, Duong D, Mitchell K, Sarwal V, et al. (2019) Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS Biol* 17(6): e3000333. <https://doi.org/10.1371/journal.pbio.3000333>

These authors contributed equally to this work.
✉ Authorship confirmed by corresponding author.
* smangul@ucla.edu

Abstract

NATUREJOBS | NATUREJOBS BLOG

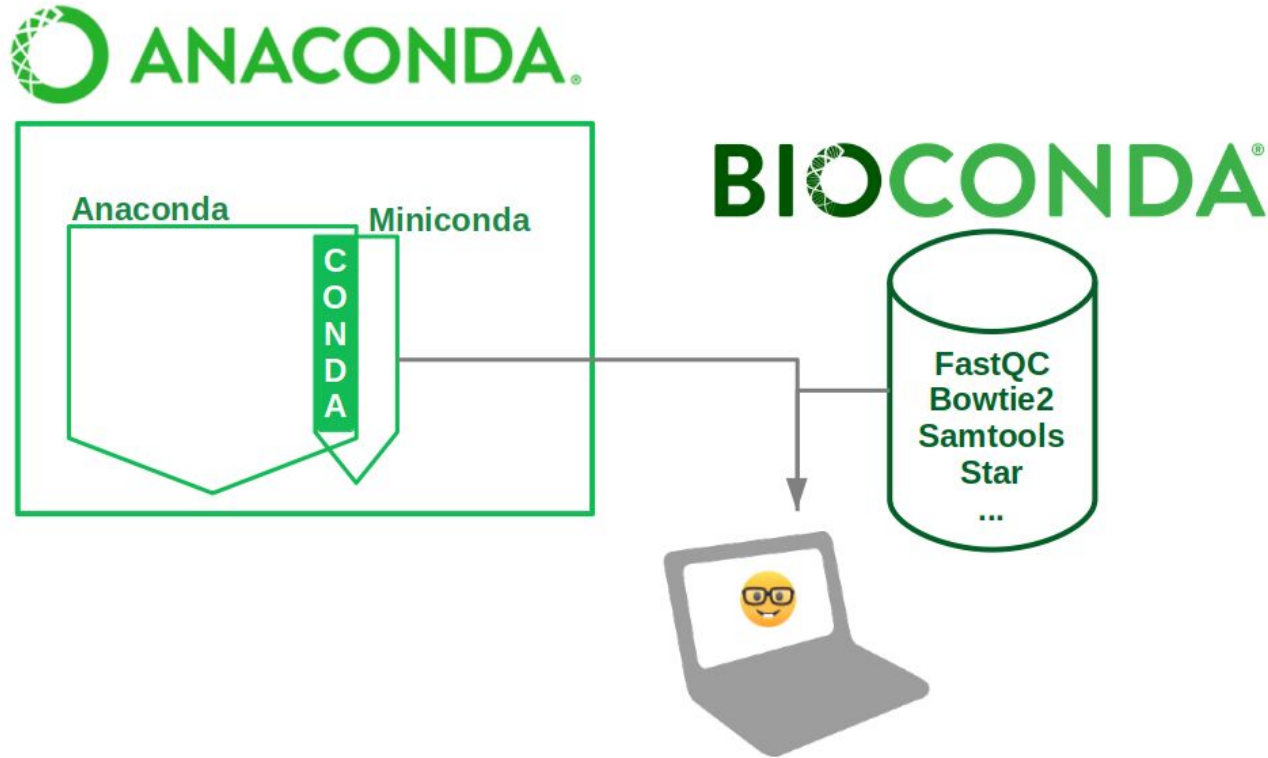
TechBlog: Bioconda promises to ease bioinformatics software installation woes

03 Nov 2017 | 12:00 GMT | Posted by Jeffrey Perkel | Category: Blog, Technology



JOHANNES KÖSTER/GITHUB

Gestion des logiciels : conda



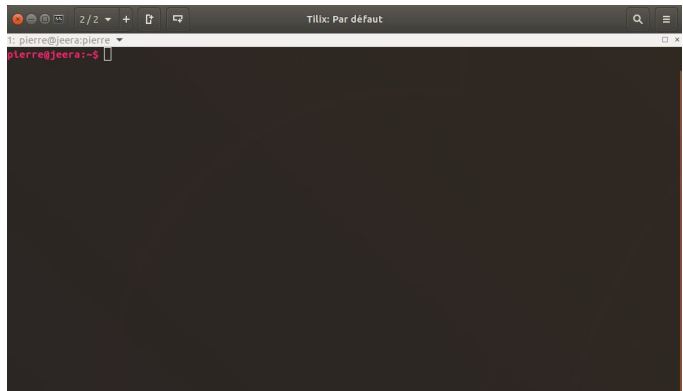
Unix !

Le *shell*

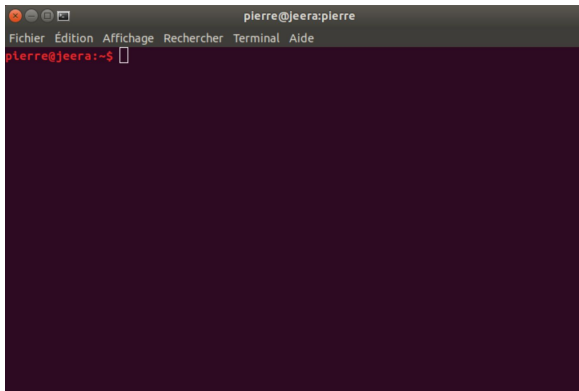
Une interface de commandes



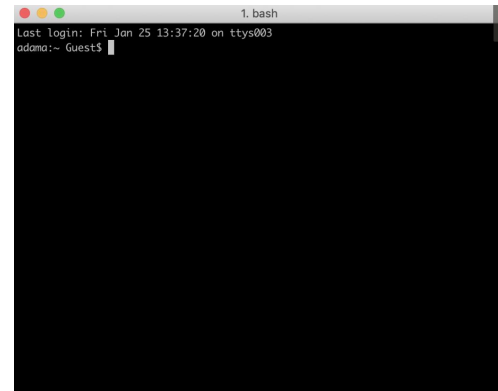
Une interface de commandes



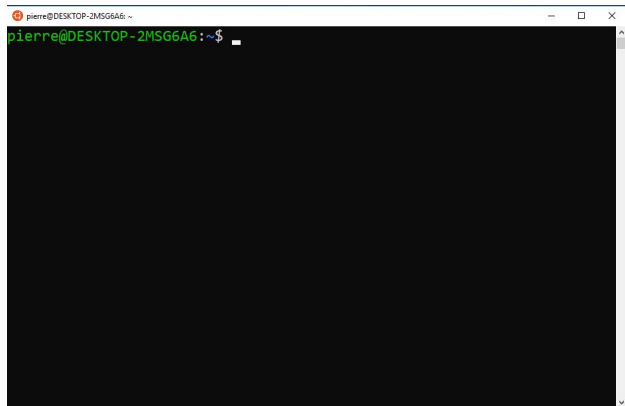
```
1: pierre@jeera:~$
```



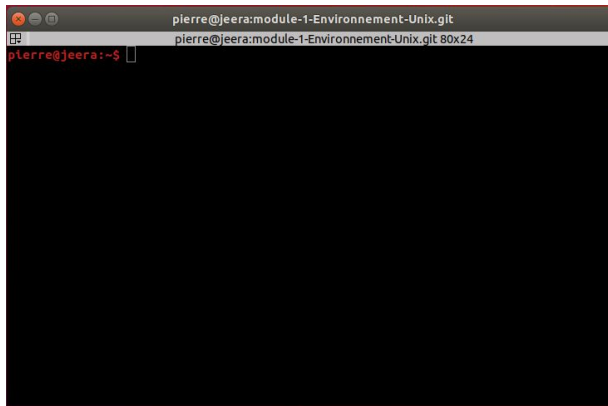
```
pierre@jeera:~$
```



```
Last login: Fri Jan 25 13:37:20 on ttys003
adama:~ Guests
```



```
pierre@DESKTOP-2MSG6A6:~$
```

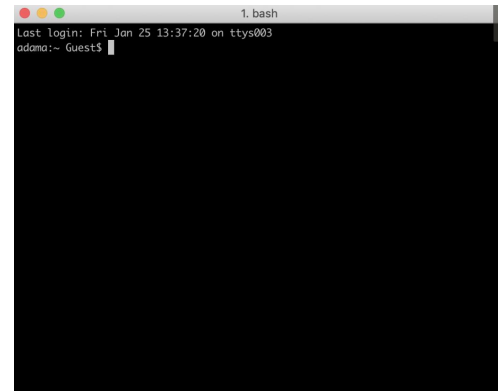
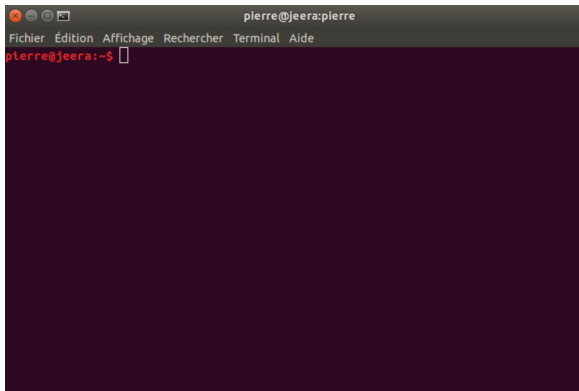
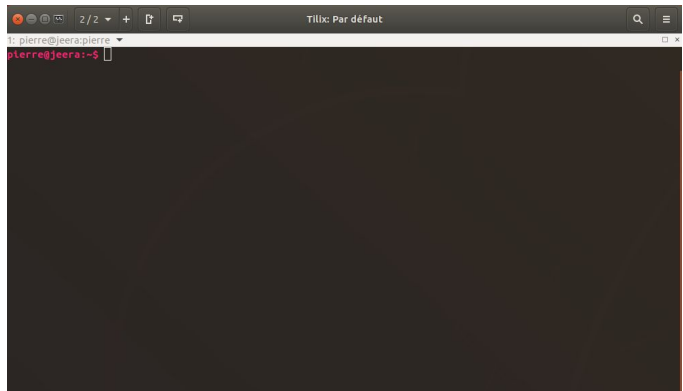


```
pierre@jeera:~$
```



```
pierre@jeera:~$
```

Une interface de commandes



À vous ! Prise de contact avec les machines de l'université :

1. Obtenez votre *login* et votre mot de passe.
2. Ouvrez votre session.
3. Lancez un *shell* via l'application terminal.
4. Changez votre mot de passe avec la commande **yppasswd** (avec 2 p et 2 s).
5. Fermez votre session puis reconnectez-vous.
6. Lancez un navigateur internet et ouvrez la page <https://huit.re/dubii-m1>

Une interface de **commandes**

Activités préparatoires sur DataCamp

INTERACTIVE COURSE

Introduction to Shell for Data Science

Continue Course

>_

INTRODUCTION TO
SHELL FOR DATA
SCIENCE

1 Manipulating files and directories

100%

This chapter is a brief introduction to the Unix shell. You'll learn why it is still in use after almost fifty years, how it compares to the graphical tools you may be more familiar with, how to move around in the shell, and how to create, modify, and delete files and folders.

[VIEW CHAPTER DETAILS](#) ▾

✓ Completed

wooclap

Activité [WooClap](#)

Tutoriels

<https://du-bii.github.io/module-1-Environnement-Unix/>