Université
de Paris

INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

Bonnes pratiques pour organiser vos projets en bioinfo

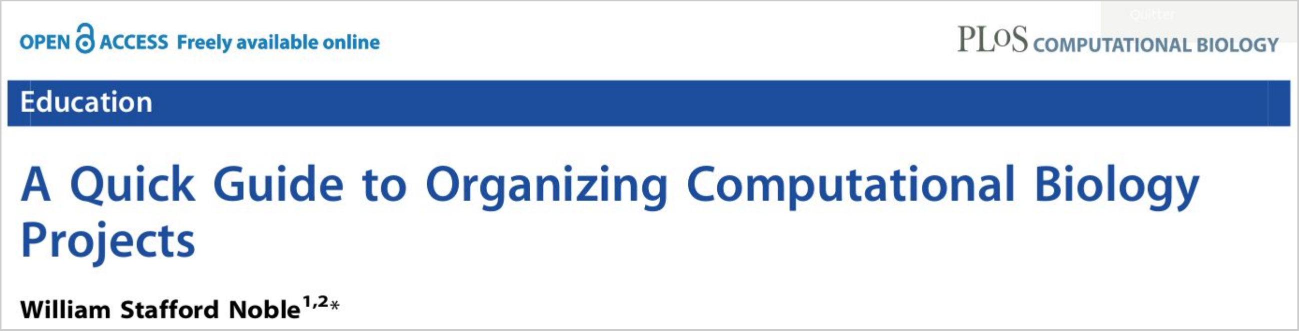
DUBii 2021

Hélène Chiapello & Pierre Poulain

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

Organisation des données

Deux références



OPEN ACCESS Freely available online

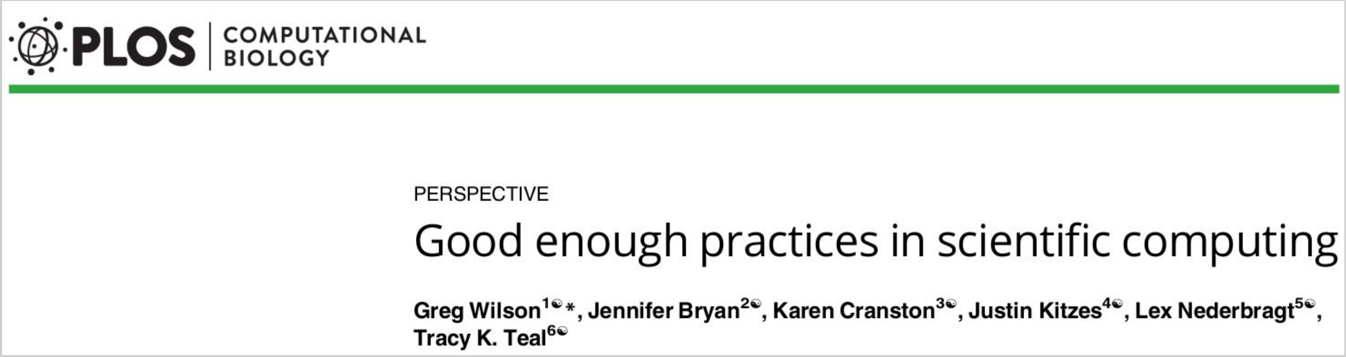
PLoS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424



PLOS | COMPUTATIONAL BIOLOGY

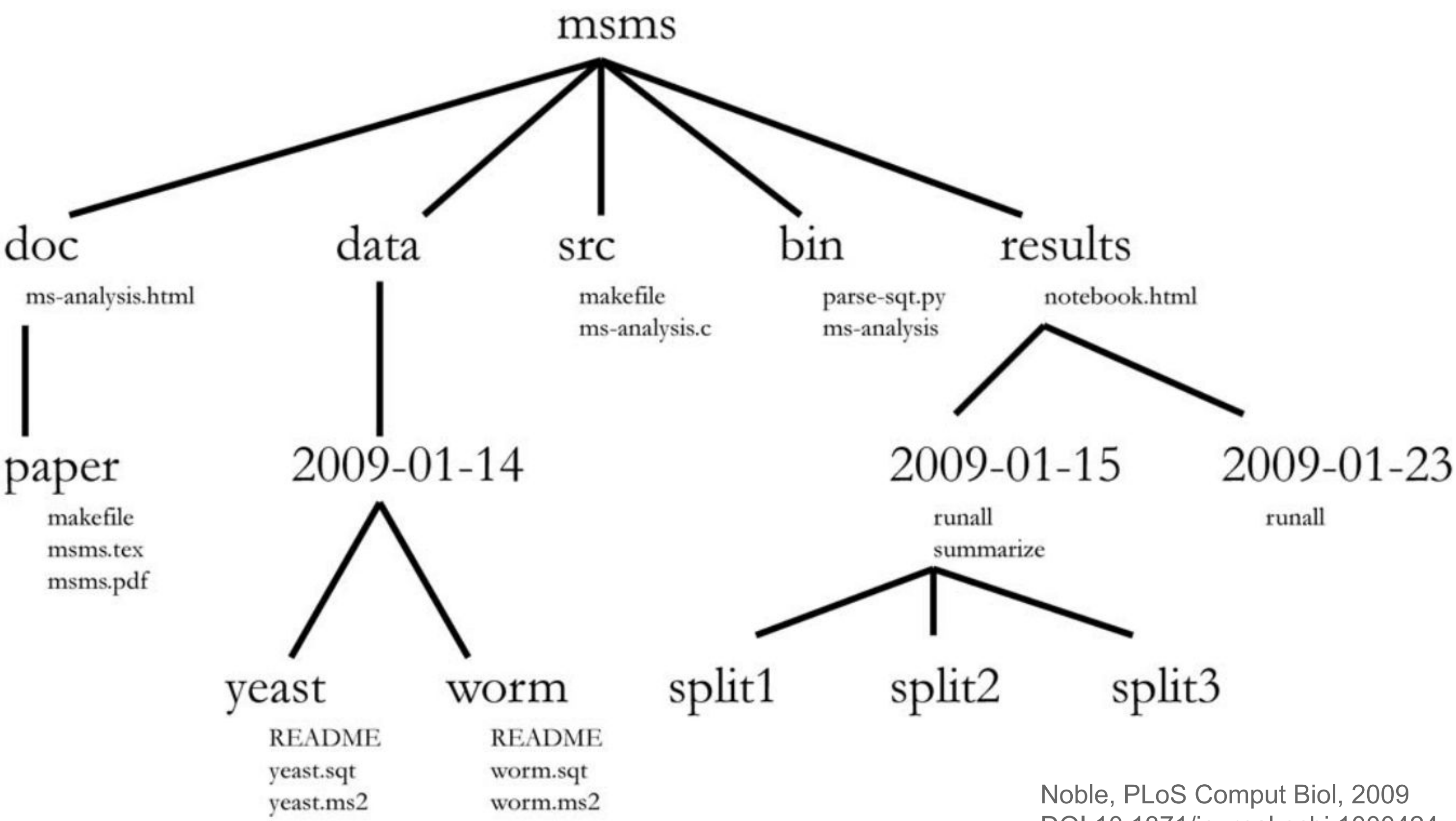
PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯*], **Jennifer Bryan**[2☯], **Karen Cranston**[3☯], **Justin Kitzes**[4☯], **Lex Nederbragt**[5☯], **Tracy K. Teal**[6☯]

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

Un exemple d'organisation



Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424

Noms de fichiers et répertoires

Pas d’espace

`_` ou `-` pour séparer les « mots »

Ex : `new_test`, `dubii-python`

Pas de caractères spéciaux

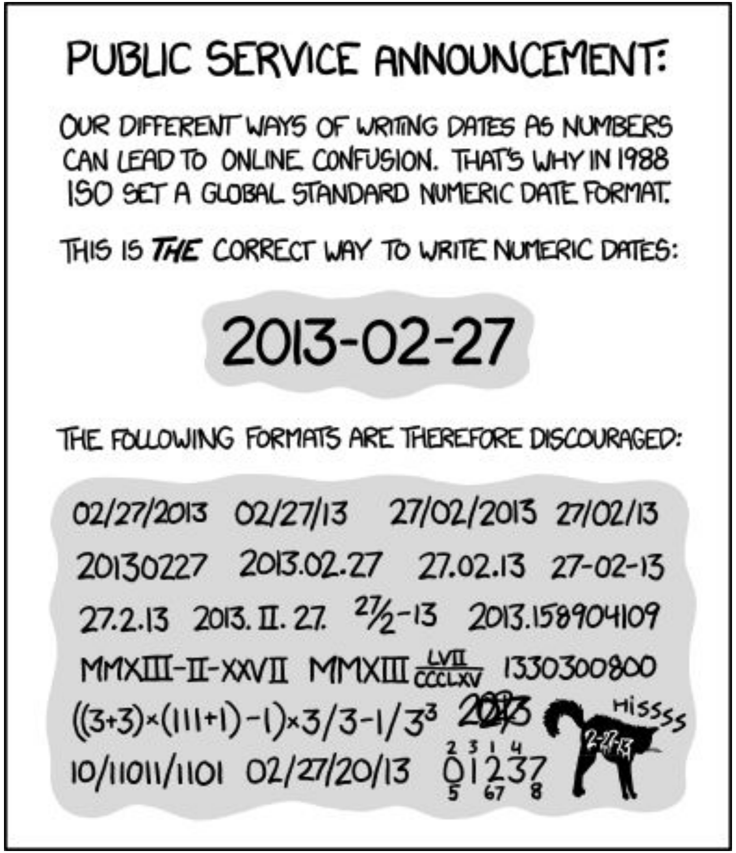
Format de date

ISO 8601 ?

Format de date

So what's your idea of a perfect date?
YYYY-MM-DD
I find other formats a bit confusing

Mahdi Yusuf / @myusuf3
https://twitter.com/myusuf3/status/865722106071453696


PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS ***THE*** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. $^{27}/_{2}$-13 2013.158904109

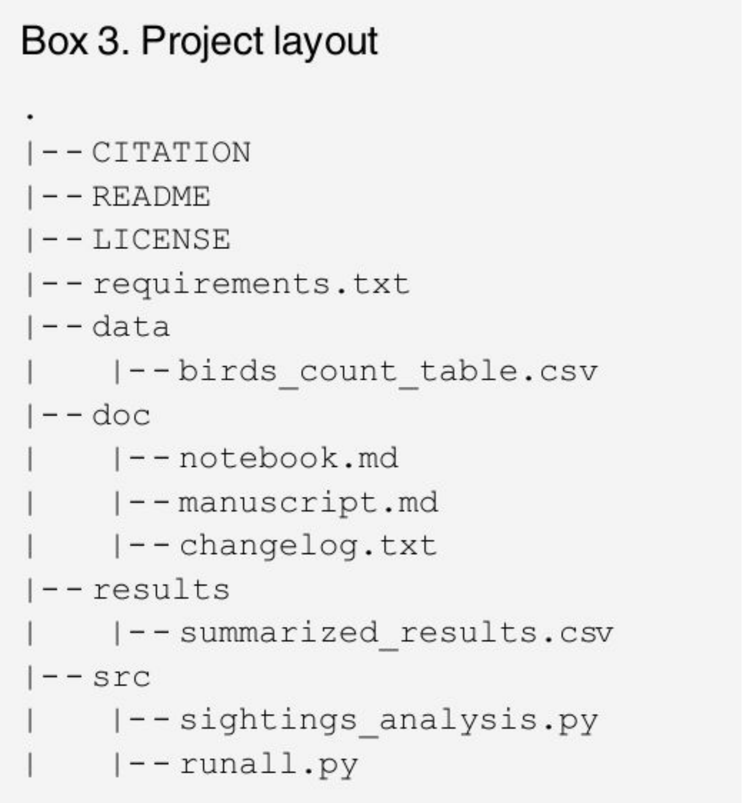
MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800

$((3+3)\times(111+1)-1)\times3/3-1/3^3$ 2013

10/11011/1101 02/27/20/13 $\overset{2}{\underset{5}{0}}\overset{3}{1}\overset{1}{\underset{67}{2}}\overset{4}{3}\underset{8}{7}$

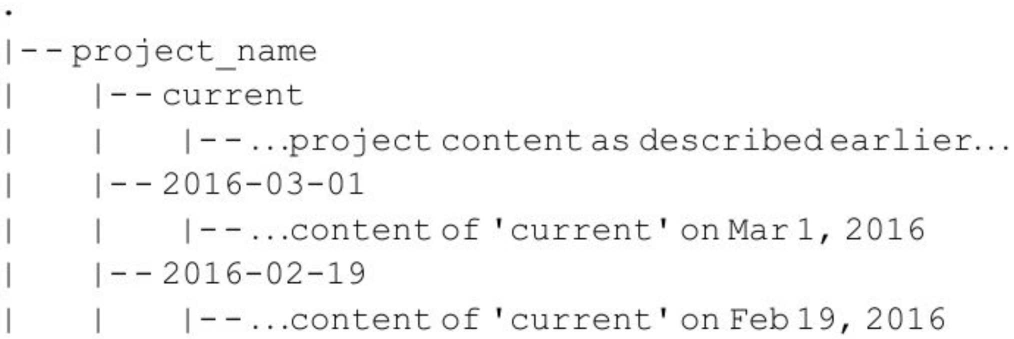
XKCD, ISO 8601
https://xkcd.com/1179/

Un autre exemple d’organisation



Box 3. Project layout

```
.
|--CITATION
|--README
|--LICENSE
|--requirements.txt
|--data
|    |--birds_count_table.csv
|--doc
|    |--notebook.md
|    |--manuscript.md
|    |--changelog.txt
|--results
|    |--summarized_results.csv
|--src
|    |--sightings_analysis.py
|    |--runall.py
```

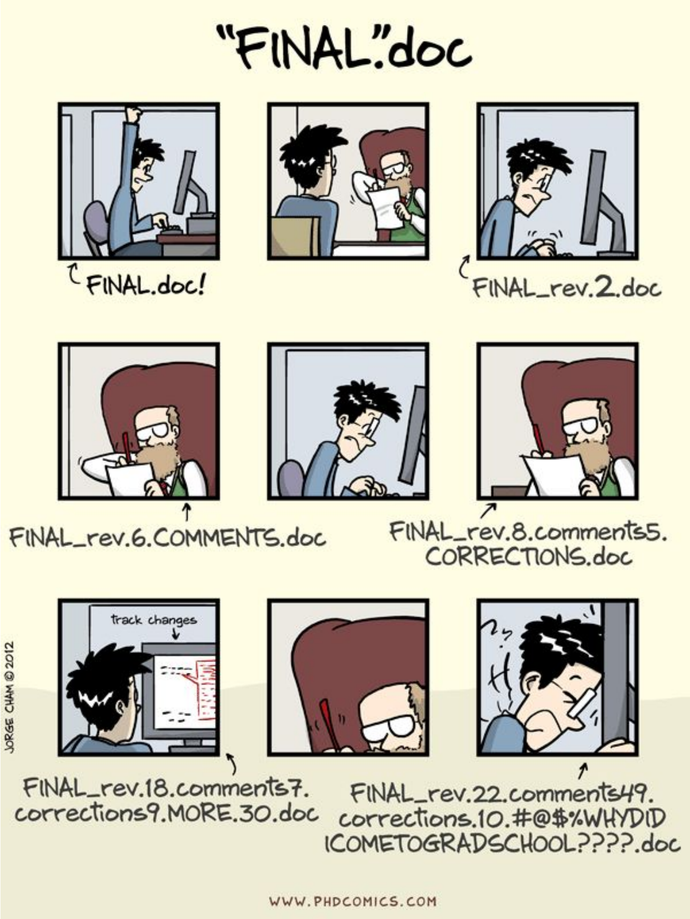


```
.
|--project_name
|    |--current
|    |    |--...project content as described earlier...
|    |--2016-03-01
|    |    |--...content of 'current' on Mar 1, 2016
|    |--2016-02-19
|    |    |--...content of 'current' on Feb 19, 2016
```

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

Gestion des données

Gestion des données

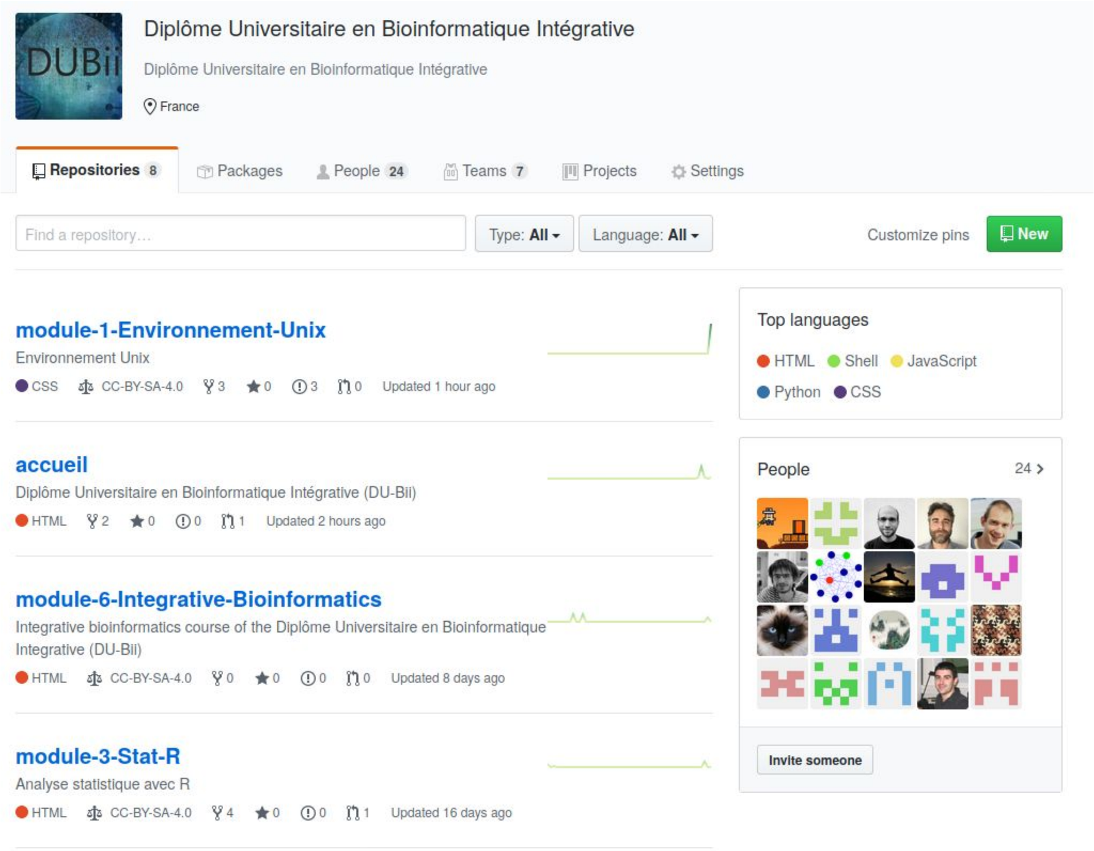


Source : PhD Comics

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)



Diplôme Universitaire en Bioinformatique Intégrative

Diplôme Universitaire en Bioinformatique Intégrative

France

Repositories 8 · Packages · People 24 · Teams 7 · Projects · Settings

Find a repository… · Type: All · Language: All · Customize pins · New

module-1-Environnement-Unix

Environnement Unix

CSS · CC-BY-SA-4.0 · 3 · 0 · 3 · 0 · Updated 1 hour ago

accueil

Diplôme Universitaire en Bioinformatique Intégrative (DU-Bii)

HTML · 2 · 0 · 0 · 1 · Updated 2 hours ago

module-6-Integrative-Bioinformatics

Integrative bioinformatics course of the Diplôme Universitaire en Bioinformatique Integrative (DU-Bii)

HTML · CC-BY-SA-4.0 · 0 · 0 · 0 · 0 · Updated 8 days ago

module-3-Stat-R

Analyse statistique avec R

HTML · CC-BY-SA-4.0 · 4 · 0 · 0 · 1 · Updated 16 days ago

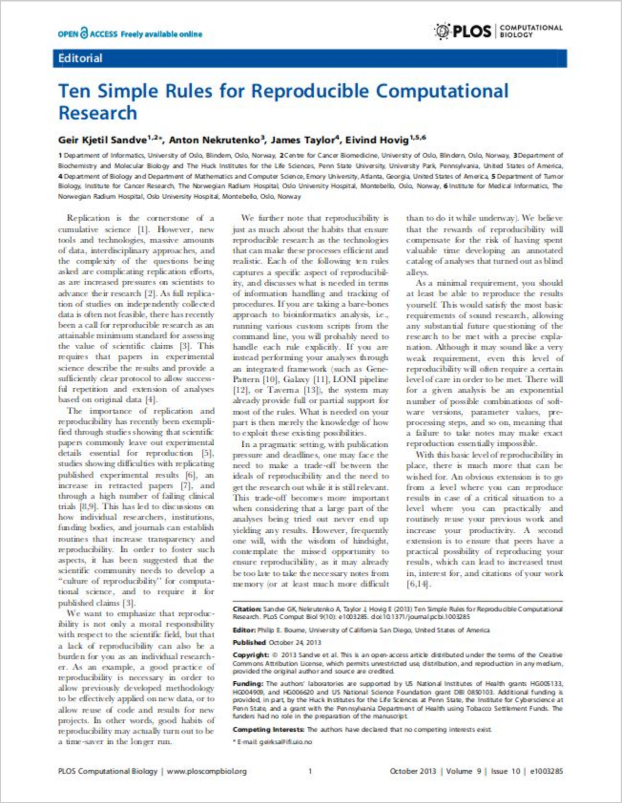
Top languages

HTML · Shell · JavaScript · Python · CSS

People 24

Invite someone

Gestion des données : git / GitHub



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2*], Anton Nekrutenko[3], James Taylor[4], 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 reproduction [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 [8,9]. This has led to discussions 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 bioinformatics 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], LONI 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, contemplate 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, preprocessing 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 HG005133, HG004909, and HG006620 and US National Science Foundation grant DBI 0850103. 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: geirksa@ifi.uio.no

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

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

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

Lightcome/Getty

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

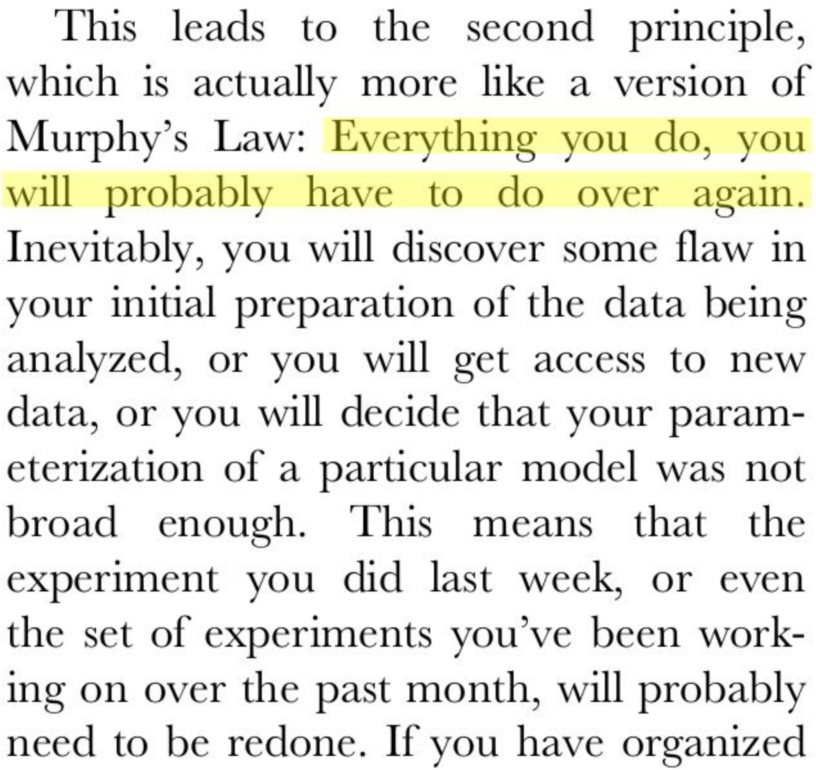
D’autres ressources :

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

intro : 10/03 après-midi 😆

Quelques conseils

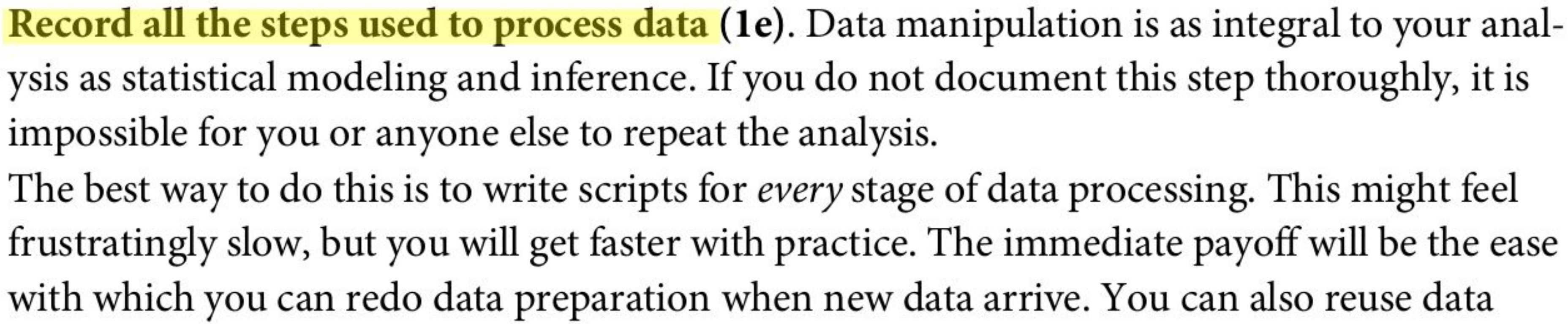
Quelques conseils



This leads to the second principle, which is actually more like a version of Murphy's Law: Everything you do, you will probably have to do over again. Inevitably, you will discover some flaw in your initial preparation of the data being analyzed, or you will get access to new data, or you will decide that your parameterization of a particular model was not broad enough. This means that the experiment you did last week, or even the set of experiments you've been working on over the past month, will probably need to be redone. If you have organized

Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424

Quelques conseils



Record all the steps used to process data (1e). Data manipulation is as integral to your analysis as statistical modeling and inference. If you do not document this step thoroughly, it is impossible for you or anyone else to repeat the analysis.

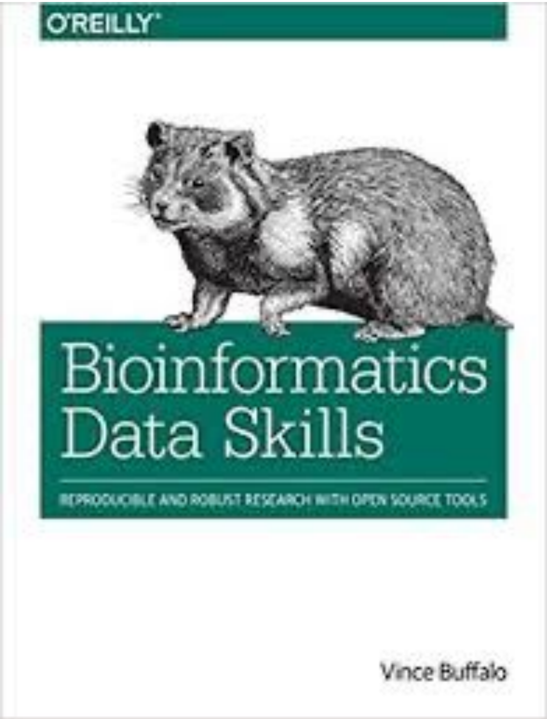
The best way to do this is to write scripts for *every* stage of data processing. This might feel frustratingly slow, but you will get faster with practice. The immediate payoff will be the ease with which you can redo data preparation when new data arrive. You can also reuse data

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

Des conseils, encore !

Adopter des pratiques **robustes** et **reproductibles**

- Code
  - Lisible
  - Documenté
  - Utiliser des librairies existantes dès que c'est possible
  - Versionné et partagé
- Données
  - Versioning
  - Plans de Gestion de Données (PGD)
- Code + données + résultats
  - Gestionnaires de workflows
  - Notebooks


O'REILLY
Bioinformatics Data Skills
REPRODUCIBLE AND ROBUST RESEARCH WITH OPEN SOURCE TOOLS
Vince Buffalo