Reproducibility for Bioinformatic tools and workflows

Authors: Peter Crisp, UMN, USA (0000-0002-3655-0130); Diep Ganguly, ANU, AUS (0000-0001-6746-0181)

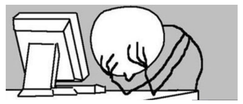
Topic →

Challenges →

Solutions →

Documentation

Why did I do this?



How can I record the steps of a bioinformatic analysis?

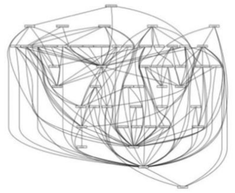
Version Control

What version of the program, data etc… did I use?

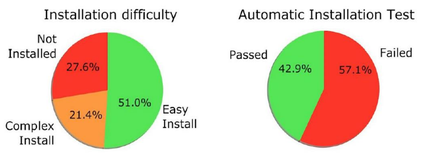
Version conflict

Software

Dependency hell



28% of the tools are impossible to install



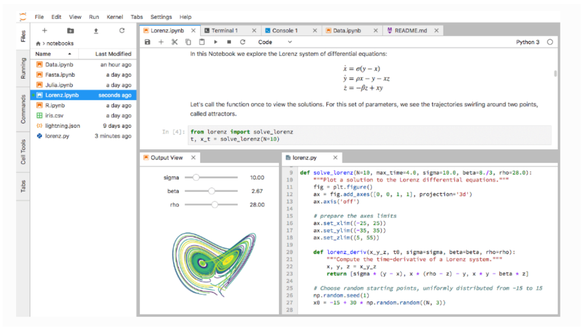
https://www.biorxiv.org/content/early/2018/10/25/452532

Notebooks

Notebooks are an essential tool for documenting analyses, enabling reproducibility and sharing.

- Keep track of analysis
- Interactive coding
- Interactive data exploration
- Imbedded visualization
- Easy access to docstrings
- Mix of code and documentation
- Over 40 programming languages
- Easily shared
- Widgets
- Interactive plots
- Run remotely on server

https://jupyter.org/documentation https://www.rstudio.com/

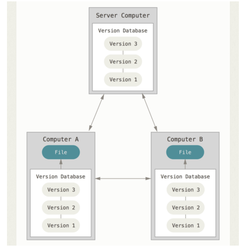


Git

Git tracks changes to files over time, enabling documentation of development and reversion to prior versions.

- Records and illustrates changes between versions
- Lets you share code and collaborate easily
- Active community with lots of tutorials and help pages
- Provides a backup for all your code to be accessed remotely as needed

https://git-scm.com/doc



Managers

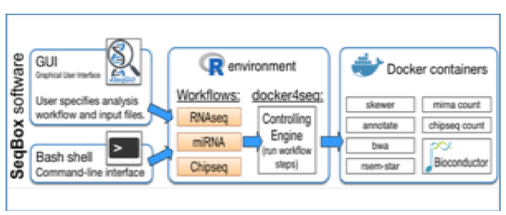
Managers and containers enable consistent use of software when emulating or reproducing an analysis.

- Handles installs and dependencies
- Allows for multiple independent environments
- Easily configurable
- Allows for manual installs as well
- Runs on all three major systems
- Open source
- You can package your own work and contribute

https://conda.io/docs/

https://bioconda.github.io/

https://snakemake.readthedocs.io/en/stable/



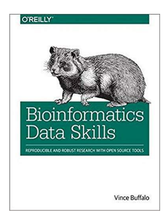
Containers

Docker runs images as containers that are

- self contained with all code, programs, libraries included. No subsequent installation required.
- Isolated
- Portable including dissemination
- Lightweight

Turns a GitHub repo with data and notebooks into a collection of interactive notebooks run in the cloud

Configuration, preservation, & reuse of executable code using containers for researchers



Resources →

Resources & further reading

➔ Reproducible computational research
Sandve et al. (2013), PloS Comp Bio,
https://doi.org/10.1371/journal.pcbi.1003285

➔ **Software**: reproducible installation
Mangul S et al. (2018), BioRxiv,
https://doi.org/10.1101/452532

➔ **Documentation**
- ◆ 'Bioinformatic data skills' Buffalo http://shop.oreilly.com/product/0636920030157.do
- ◆ Software carpentry https://software-carpentry.org/

➔ **Version control**:
http://smutch.github.io/VersionControlTutorial//

➔ **Containers:**
- ◆ Docker https://docs.docker.com/
- ◆ Biocontainers http://biocontainers.pro

➔ **Reviewing:**
- ◆ Computation checklist github.com/vivekbhr/reproChecklist