Data analysis at NBIS

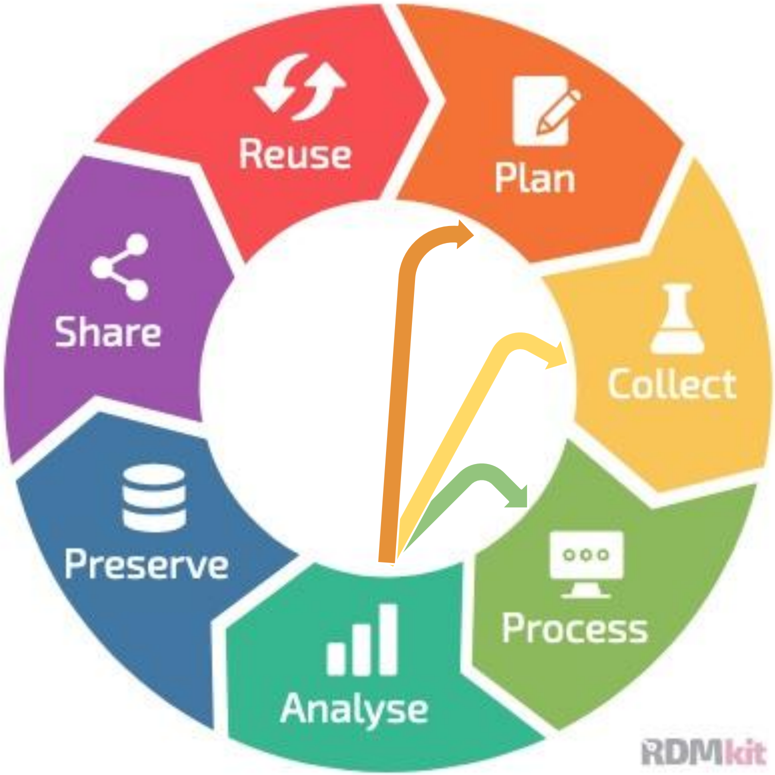
NBIS Data Management Team

data-management@scilifelab.se

RDM-kit

Data analysis is central to research

- Relies on Plan, Collect and Process
- Location of data
- Computing environment
- Tool selection
- Documentation
- Workflow publication (FAIR)



NBIS

- Inflow of data from multiple sources - Researchers, data platforms, repositories etc.
- Contract with NBIS specifies scope, goals and kind of analysis to be performed
- Pre-defined number of hours with per hour cost, per project
- Includes data access, scouting (pre-analysis), data exploration, analysis, and result report
- Possible to extend X hours per project
- Data enters the NBIS infrastructure (extended expertise)
- Cost-free access to Data Management at all stages (consultancy, advice, hands on)

NBIS resources

Course - Tools for Reproducible Research (1w)

https://uppsala.instructure.com/courses/73110

At the end of the course, students should be able to:

- Use good practices for data analysis and management
- Clearly organise their bioinformatic projects
- Use the version control system Git to track and collaborate on code
- Use the package and environment manager Conda
- Use and develop workflows with Snakemake and Nextflow
- Use R Markdown and Jupyter Notebooks to document and generate automated reports for their analyses
- Use Docker and Singularity to distribute containerized computational environments

NBIS resources

Support Drop-in Event (previously onsite, now online)

- Weekly, Tuesdays 14.00 CET
- Present your questions to Bioinformaticians for consultancy and support

Reproducibility group

Group at NBIS working with, and promoting, reproducible research internally among Bioinformaticians.

Training & Tools

Joint working group for Data Stewards at NBIS and Data Center

- Selection of tools
- Development of material, teaching and resources for usage

NBIS

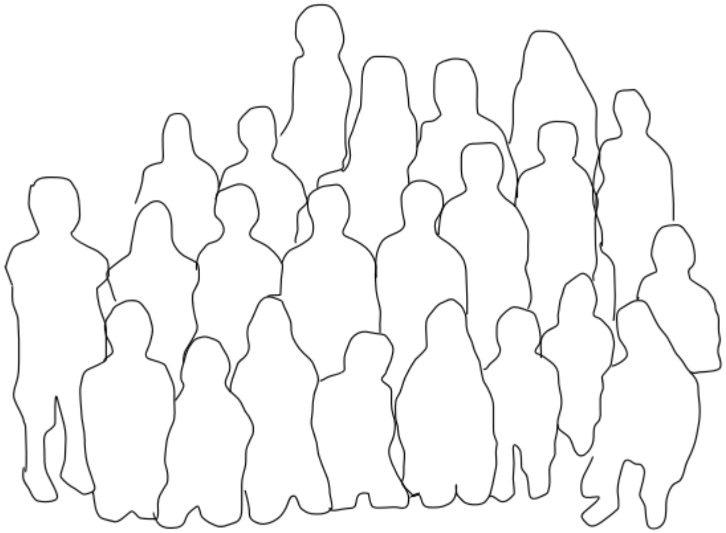
Resources for Bioinformatician at NBIS working with data analysis

- Access to research group data (agreements and contracts in place)
- Access to high performance computing resources (e.g UPPMAX)
- Can direct data management questions to DM Team
- Thematic teams for discussion on analysis issues, second opinions
- Reproducibility group for discussion
- Code review

Audiences

Who are we catering to?

- Researchers themselves
- Research group
- Data Steward
- Repositories
- Real and potential re-users of data
- Journals
- (International networks, e.g. ELIXIR, EOSC, RDA etc)



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

FAIR by design

SciLifeLab

ENA
European Nucleotide Archive

Study & data design | Sampling & specimen collection | Sample preparation | Sample analysis & data generation | Data processing to prepare inputs for analysis | Data analysis | Communicating results

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, …

Information system

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, …

Laboratory workspace

Biobank

Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data, …

Data delivery

Digital workspace

Outputs

publications,
data,
tools,
workflows,
reports,
dashboards, …

Local data archive etc

Research databases

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com

Reproducibility

Reproducibility - aka replicability, aka repeatability

What information and tools are required for results to be reproduced from data, over time, and across systems?

- Relies on level of documentation
- Not a step from nothing to everything
  Focus should be on Delta
  (Δ - gradual improvement over time)

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalizable

Reproducibility

Smallest unit in reproducibility for a Data Steward is - *Same code* + *Same data*

Reproducibility caters secondarily to Robustness and Replicability

- Reproducible environments
- Workflow managers
- Version control systems
- What to document?
    - Packaging and sharing executable code
    - Packaging and sharing source code
    - Referencing and sharing workflows

Reproducibility

Solutions for data processing - FAIR

- Persistent and explicit access to data
  - Exception for sensitive data
- Access to analysis tools and workflows
- Persistent access to code
- (Access to notes and documentation)
- Make all of the above available by PID's

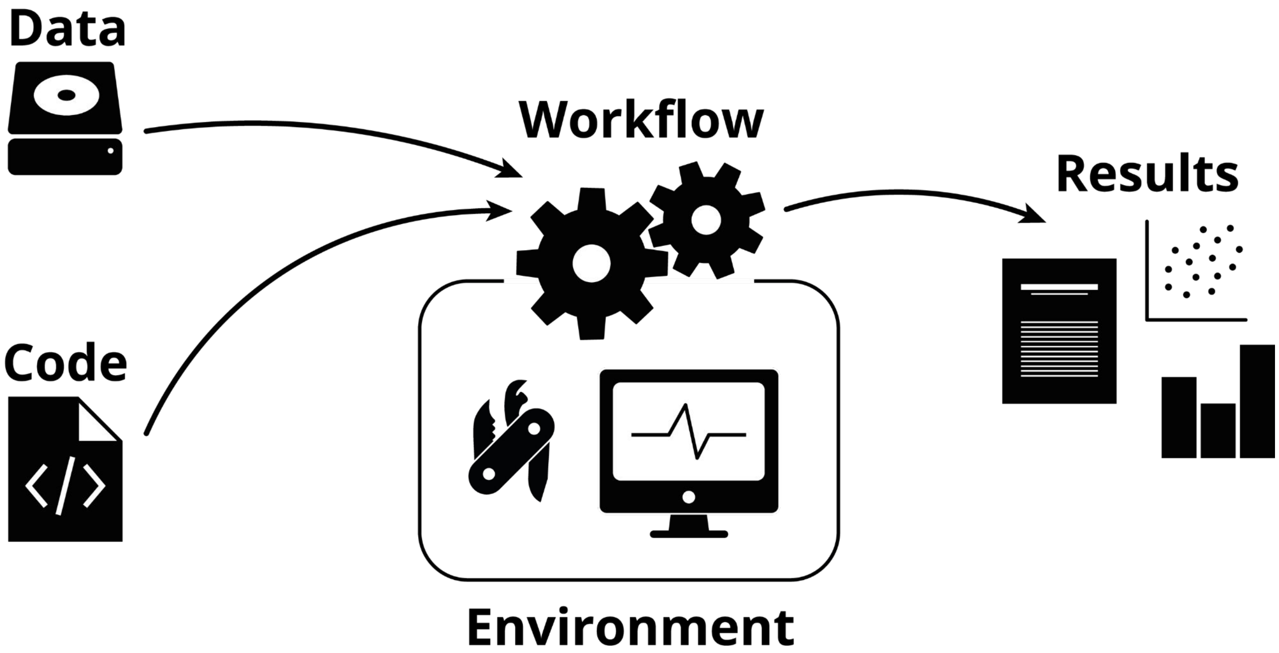
Restricted access to data increases the requirement for documentation quality. Important when analysis cannot easily be repeated.

Reproducibility

Desirables for data processing - FAIR

- Documentation of code execution
- Documentation of code failures and/or limitations
- Integrable documentation from data platforms
- Elaborative end report after project closure
- Independent reproducibility check
- Independent interoperability check
- Pre-defined folder structure
- Formulation of quality markers for follow-up

Analysis setting



Code reproducibility

Code reproducibility

- Git
  - Open source version control system
  - Snapshots, review

- Markdown
  - Explanation of code, provenance and usability

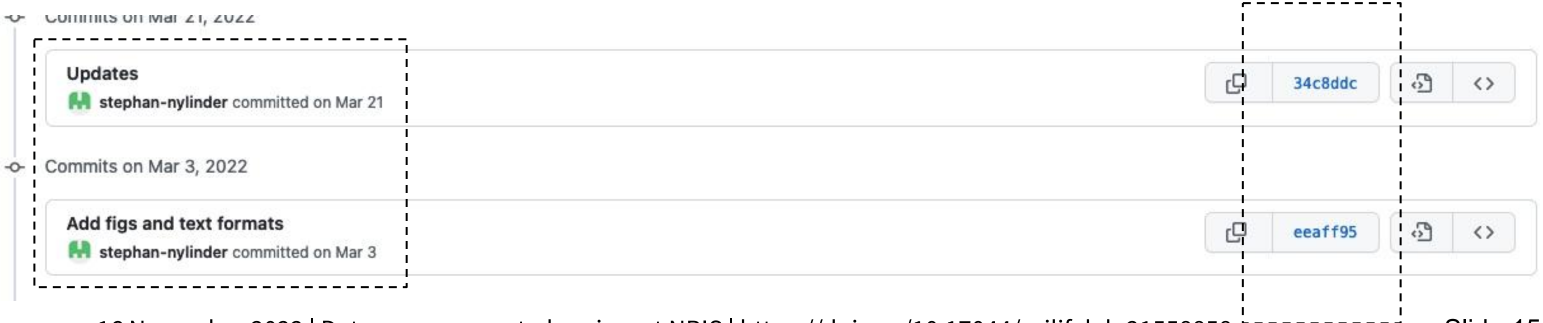
- Jupyter
  - Platform for e.g. documentation, description, publication of analysis
  - Exportable formats
  - Interoperable output?

Code reproducibility

Git (e.g. GitHub)

File histories with incremental changes (ID, commit snapshots)

- File reversion
- Compare files over time
- Who did what and when
- Always backed up
- Publishable repository



Commits on Mar 21, 2022

Updates
stephan-nylinder committed on Mar 21

34c8ddc

Commits on Mar 3, 2022

Add figs and text formats
stephan-nylinder committed on Mar 3

eaaff95

Code reproducibility

Git (e.g. GitHub)

- Git allows open review of e.g. code, easy publishing, link PID to repository, markdown documentation
- Always backed up
- Easy to share
- Suitable for small to medium size files (code, documents etc.)
    - Not large binary data files

Environments

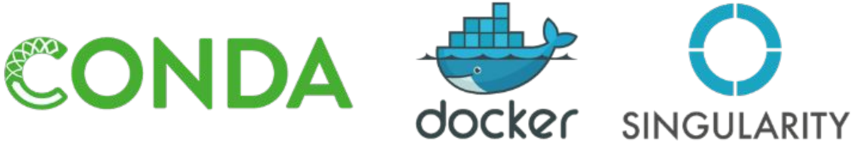
Analysis reproducibility issues over time…

- ➢ Tools can be specified
- ➢ Versions of tools can be specified
- ➢ Analysis can still fail to be reproducible due to wrong environment!
    - Tools missing for user OS
    - Tool dependencies lacking or non-functional
    - Tool versions may differ between different OS
    - Demanding work-arounds

Environments

Tools and versions of tools not always sufficient

Solution? Environments!

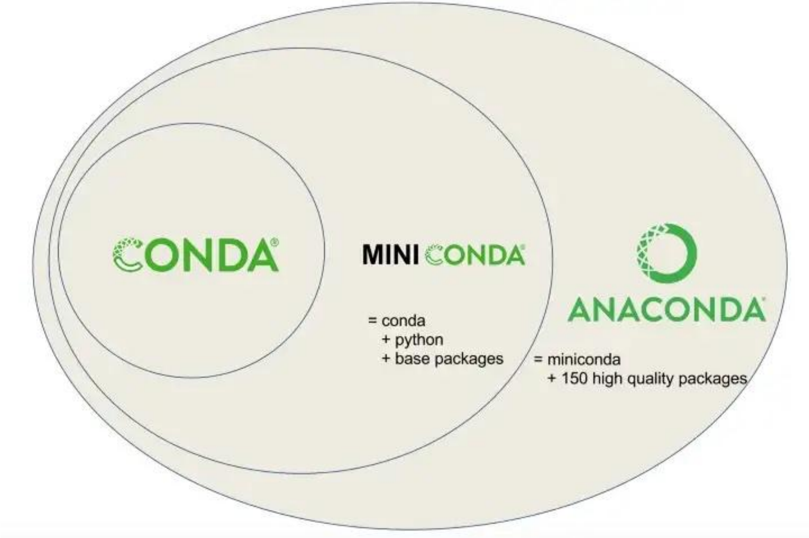


- Cross platform
  - Open source
    - Standalone
      - Light-weight
        - Language agnostic

Environments

Environment functionality (e.g Conda)

- Nested folder installation of packages
- Controlled dependencies
- Independent parallel environments
- Reproducible by env-file specification

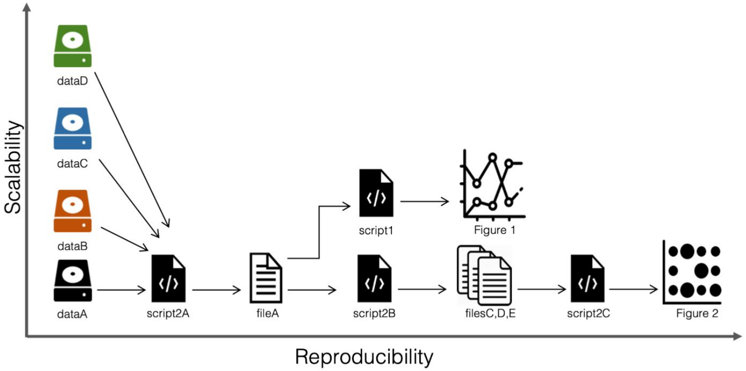


Bioconda

- Specific bioinformatic packages
- Limited to MacOS and 64-bit Linux
  https://bioconda.github.io/

Workflows

- Applicable for all (even ageing and growing) projects
- Keeping track of all parts of a (changing) analysis and how it fits together
- Branching analysis with many steps resulting in multiple output files
- Metadata documentation per step, and as a whole
- How to consider manual workflows (documentation issues?)

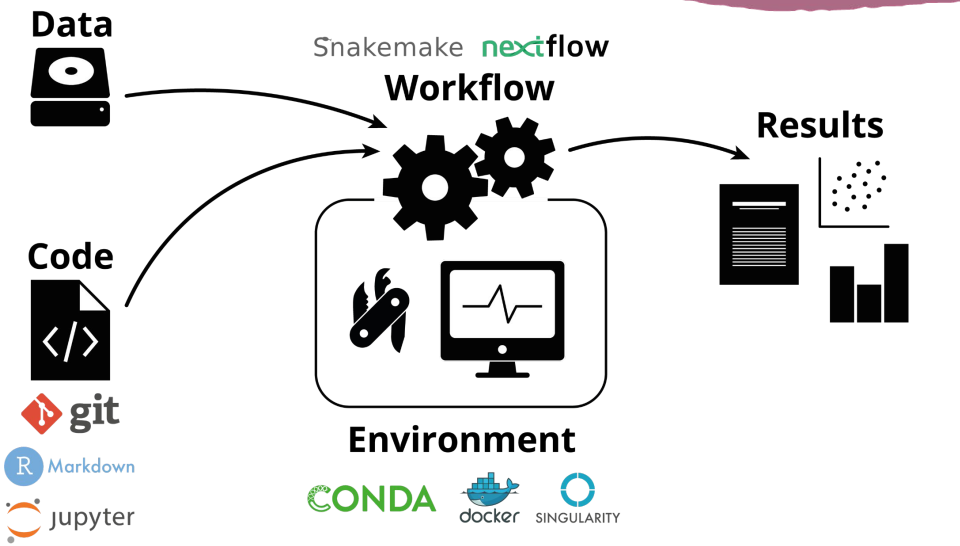


Workflow managers

Nextflow (nf-core), Snakemake, Galaxy

- GUI - Galaxy, Arvados
- CWL and likes - Nextflow, Snakemake

Integration with other reproducibility tools (Conda, Docker), cloud platforms, and GitHub



Workflow documentation

A workflow describes a pipeline for structured analysis of data in a specific context, where the result is derivable from explicit workflow specifications.

FAIR-ification requires analysis *citability*

No defined standards for workflow documentation

- RO-crates (https://www.researchobject.org/ro-crate/)
    - json-format?
- WorkflowHub (https://workflowhub.eu/)

Workflow documentation

10 Things for Curating Reproducible and FAIR Research

https://zenodo.org/record/6797657#.YsQB9OxBwlw / https://curating4reproducibility.org/10things/

1. Completeness The research compendium contains all of the objects needed to reproduce a predefined outcome.	6. Access It is clear who can use what, how, and under what conditions, with "open" being preferred.
2. Organisation It is easy to understand and keep track of the various objects in the research compendium.	7. Provenance The origin of the components of the compendium and how each has changed over time is evident.
3. Economy Fewer objects in the compendium mean fewer things that can break and less ongoing maintenance.	8. Metadata Information about the compendium and its components is embedded in a standardised schematic code.
4. Transparency The full context necessary to understand the research process is available.	9. Automation As much as possible, the computational workflow is script-based to allow re-execution with minimal actions.
5. Documentation The process and reasoning required to reproduce a scientific claim are readily available and understandable.	10. Review A series of managed activities are in place to ensure continued access to and functionality of the compendium.

Exercise

- How do researchers at your departments document analyes?
- Discuss what implementations you can strive towards at your department to increase general reproducibility (data and results).