University of Zurich[UZH]

URPP Evolution in Action

Bioinformatics session

Reproducibility

Dr. Heidi E.L. Tschanz-Lischer
University of Zurich
Switzerland

8 September 2017

Reproducibility

Reproducibility

- Minimal requirement: Be able to reproduce the results yourself!
  (Be aware that your future self is like another person)
  - Sometimes analysis needs to be repeated (e.g. reviews, new data set)
    → will increase your productivity
  - Allow previously developed methodology to be applied on new data set

- How individual researchers can increase transparency and reproducibility?

Reproducibility – 1. Rule

1. Rule: For every result, keep track of how it was produced

- Raw data → final result
  - involves many steps (single commands, scripts, programs)
  - automated or performed manually
  - Pre- and post-processing steps are often critical to reproduce results
- Critical details in steps using a computer program:
  - Name
  - Version
  - Exact parameters used
  - Inputs used
- Best solution:
  - simple pipeline/bash script: containing the full analysis workflow
    → additionally allows for automated reproduction
  - Include extensive comments or a README file: with all details of analysis

Reproducibility – 2. Rule

2. Rule: Avoid manual data manipulation steps

- Whenever possible: use programs instead of manual procedures to modify data
- Manual manipulations
  - Inefficient
  - Error-prone
  - Difficult to reproduce
- Examples
  - Modification with UNIX command line → better use small custom bash scripts
  - Analysis in R → use editors (like Rstudio) and execute commands from scripts
  - Modify data to fit specific input formats → use a data format converter (like PGDSpider Lischer and Excoffier 2012)
- Manual operations cannot be avoided:
  - note which data files were modified or moved and for what purpose

Reproducibility – 3. Rule

3. Rule: Archive the exact versions of all external programs used

- May be necessary to use programs in the exact versions used originally
  - To exactly reproduce a given result
  - Input/output format may change → not even possible to run it with a newer version
- Not always trivial to get old version of a program (e.g. Bioconductor)
  → archive exact versions of programs
- Minimum:
  Note the exact names and versions of the main programs

Reproducibility – 4. Rule

4. Rule: Version control all custom scripts

- Even the slightest changes in a computer program/script can have large consequences (intended or unintended)
- If code is not systematically archived → backtracking to certain results may not be possible

- Solution:
  - Use version control system (like git → see Tutorial on version control)
  - Makes possible to track evolution of code

- **Minimum:**
  Archive copies of your scripts from time to time

Reproducibility – 5. Rule

5. Rule: Record all intermediate results, when possible in standardized formats

- In case full process is tracked → all intermediate data can be regenerated
- In practice: having intermediate results may be of great value
- Intermediate files
  - like BAM, VCF or filtered VCF files
  - can reveal discrepancies toward what is assumed → uncover bugs or faulty interpretations (e.g. specific filtered SNPs)
  - directly reveals consequences of alternative programs and parameter choices
  - it allows parts of the process to be rerun
  - it allows any inconsistencies to be easier tracked to the step where it arose

- Minimum:
  Archive any (key) intermediate files that are produced running an analysis
  → as long as the required storage space is not limited

Reproducibility – 6. Rule

6. Rule: For analyses that include randomness, note underlying random seeds

- Some analysis may involve some element of randomness
  - Same program will typically give slightly different results every time it is executed
  - Example: simulations, Bowtie2 (multiple equally good alignments of reads)

- Using the same initial seed
  - all random numbers used in an analysis will be equal
  - random seed should be recorded

- Minimum:
  Note analysis steps involving randomness
  → certain level of discrepancy is expected when reproducing results

Reproducibility – 7. Rule

7. Rule: always store raw data behind plots

- Figures are often modified several times from first creation to publication
  - visual adjustments to improve readability
  - to ensure visual consistency between figures
- If raw data behind figures is stored
  - simple modification of plotting procedure → don't need to redo whole analysis
  - One can easily get detailed values of a figure by reading the raw numbers
- Useful to store both the underlying data and the processed values
  - Example: plotting of histograms → store values before binning (original data) and the counts per bin (heights of visualized bars)
- Also store the code used to make the figure (e.g. R code)
- Minimum:
  Note the data underlying a given plot and how it could be reconstructed

Reproducibility – 8. Rule

8. Rule: Generate hierarchical analysis output, allowing layers of increasing detail to be inspected

- Final results (e.g. plots or tables) of a paper often represent highly summarized data
  - Example: Each value along a curve may be an average from a distribution
- Often useful to inspect the detailed values underlying the summaries
  → to validate and fully understand the main result
- When storage allows: always include output with all underlying data
  - Use systematic naming: allows to easily relate full data to summarized value
- **Minimum:**
  At least once generate, inspect, and validate the detailed values underlying the summaries

Reproducibility – 9. Rule

9. Rule: Connect textual statements to underlying results

- Typical research project:
  - Different analyses are tried → stored on server or personal computer
  - Different interpretation are made → stored on personal notes or e-mails
- Textual interpretations:
  - carry extra information (connections to other results or theories)
  - supported in a given result
- Statements should be connected to underlying results already from the time the statements are initially formulated
  - Allow efficient retrieval of details behind textual statements
  - Include within text itself: file path or ID to results
  - Tools: Sweave, GenePattern Word add-in and Galaxy Pages
- Minimum:
  Provide enough details along with your textual interpretations → allows to track underlying results

Reproducibility – 10. Rule

10. Rule: Provide public access to scripts, runs, and results

- All input data, scripts, versions, parameters, and intermediate results should be made publicly and easily accessible
  - public databases
  - in online material of journals

- Minimum:
  - Submit the main data and source code as supplementary material
  - be prepared to respond to any requests for further data or methodology details

Conclusion

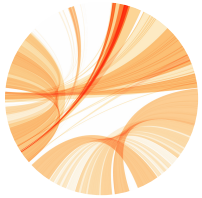
- Exponential number of possible combinations of software versions, parameter values, preprocessing steps, …
  → without notes the reproduction essentially impossible

- You should make your work reproducible
  - sends a strong signal of quality, trustworthiness, and transparency
  - could increase the
    - quality and speed of the reviewing process on your work
    - chances of your work getting published
    - chances of your work being taken further and cited by other researchers

Acknowledgment

- http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285

University of Zurich[UZH]



URPP Evolution in Action