Practices and Tools for Reproducible Analysis

Cliburn Chan

Reproducing analysis can be hard

> "Our hard disk crashed."

> "It only runs on Windows 95"

> "Our postdoc went back to Korea."

> "The software is no longer online."

> "The dog ate the code."

Overview

> Mostly common sense

Outline of talk

- Research is becoming more complex
- Reproducible research: big picture
- Practices and tools for reproducible analysis

Research is complex

Single discipline	â†’ Multiple disciplines
Single assay type	â†’ Multiple assay types
Single center	â†’ Multiple centers
GUI-driven analysis	â†’ Script-driven analysis

Multi-disciplinary

- Cancer immunotherapy
    - Surgeon
    - Oncologist
    - Cancer biologist
    - Immunologist
    - Platform experts
    - Bioinformatician
    - Biostatistician

Multiple complex assays

- Cancer immunotherapy
  - Whole exome or targeted sequencing for mutation load and neoantigens
  - RNA-seq for expression and pathway analysis
  - TCR sequencing for clonal diversity
  - Flow or mass cytometry for immunophenotyping and functional characterization
  - Multiplexed ELISA for cytokines, angiogenic factors, tumor growth factors
  - IHC for tumor architecture

Multi-center studies

- May be necessary to get adequate power
    - Sample processing
    - Sample shipping
    - Reconciliation
    - Data standards and annotation

Complex analysis pipelines

- Manual analysis is either already impossible or moving there rapidly
  - Whole exome sequencing: 180,000 exons or 3,000,000 base pairs
  - RNA-seq: 20,000 genes, millions of reads
  - TCR sequencing: possible TCRs > 10^{15}
  - Cytometry: CyTOF and BD Symphony up to 50 parameters
  - Multiplexed ELISA: 10s - 100s of soluble factors
  - IHC: Up to 10 distinct antibodies

Reproducible research

> Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

> BMJ 2003; 327

Reproducible Pieces

- Experimental design
- Data generation
- Data stewardship
- Data analysis

Experimental design

- Is there a statistical analysis plan?
- Is there a sample management plan?
- Is there a data management plan?

Statistical analysis

- What is the study objective?
- What is the outcome?
- What are the variables?
- What is the study design?
- Is there sufficient power?
- Will the samples be representative?

Sample management

- Shipping SOP
- Transfer and reconciliation SOP
- Sample labeling
- Sample tracking
- Use of LIMS

Data management

- Where will raw data be uploaded/stored?
- What variables will be recorded?
- What data standards will be used for annotation?
- Where will it be recorded?
- How are data entry errors handled?
    - Automated validation checks
    - Double-entry book-keeping
    - Review by supervisor
- Are data modifications tracked/logged?

Data generation

- Instrument calibration
- Assay QA/QC/reproducibility
- Operator training
- Written SOPs essential

Data stewardship

- FAIR principles to facilitate knowledge discovery
  - Findable
  - Accessible
  - Inter-operable
  - Reusable
- See guidelines

Reproducible analysis

> What I tell you three times is true.
>
> The Hunting of the Snark
>
> by Lewis Caroll

Data re-analysis scenarios

- A reviewer wants to reproduce your results
- A reader wants to reproduce your results
- Your study is chosen for a meta-analysis
- You need to update the data
- There is a bug in your script

Can you reanalyze your data?

- Can you find and reuse the data?
- Can you recreate the analysis environment?
- Can you find and use the analysis scripts?
- Can you replicate the report/poster/paper?

Can you find and reuse the data?

- Does the data even exist anymore?
- If exists, can you identify the exact data used?
- Can you link the laboratory and clinical data?

Practices and tools (Data)

- Keep raw data
- Create hash data signatures
- Use standard vocabularies for annotation
- Use standard exchangeable formats
- Make a copy of all data
- Deposit in public repository
    - Genomic Data Commons
    - ImmPort

Example: ReFlow

Can you recreate the environment?

- You used Windows XP
  - Your lab is now a Mac-only shop
- You used proprietary software A from Vendor X
  - Vendor X went bankrupt 2 years ago
- You ran the analysis with R 3.1.2 with packages X (version A), Y (version B) and Z (version C))
  - Versions have been updated and are not compatible with R 3.1.2

Practices and tools (Environment)

- Prefer open-source tools
- Use reproducible environments

Example: Using Docker (Jeremy's talk)

Can you use your analysis code?

- Using a GUI (e.g. Excel), you needs perfect memory
- Perfect memory does not exist
- Can you find the script(s) used for analysis?
- Are you 100% sure that is the script used?
- Dies your script still run? (see previous slide)

Practices and tools (Code)

- Don't do the final analysis using a GUI
- Use version control for scripts (e.g. `hg`, `git`)
- Use a version control service (e.g. `GitHub`, `Bitbucket`')
- Commit early and often
- Use informative log messages when commuting

Example: Using GitHub

Can you re-generate the report/poster/paper?

- Can you replicate the report/poster/paper when
  - you need to modify some data
  - you need to fix a bug in your script
  - you want to try an alternative algorithm
- Regeneration is hard with Excel/Prism/Word workflows

Practices and tools (Documents)

- Use a script for calculated values, tables and plots
- Practice literate programming
    - `knitr`
    - "Notebooks" - Jupyter, nteract, beaker, RStudio

Example: Literate programming with notebooks

Summary

- Upload data to public repository
- Prefer scripting to GUI applications
- Use version control
- Run scripts in reusable containers
- Program in a literate style
- Shameless plug: We offer [comprehensive free training in reproducible genomics analysis sponsored by NIH BD2K grant](https://biostat.duke.edu/education/high-throughput-sequencing-course)

Questions?

> Curiosity killed the cat.

> English folk saying