Computing Workflows for Biologists

Based on:

Shade & Teal, Computing Workflows for Biologists: A Roadmap, PLOS Biology

Data Carpentry data organization lessons

- How many people here plan to analyze data with a computer in their work?
- Are you working with other people on this analysis?
- Do other people need to understand your analysis?
- Do you need to remember and understand your analysis?

Elements of computing

- How data was generated (metadata)
- Data
- Data cleaning steps
- Data analysis steps
- Final plots and charts

Metadata

- The basis of any analysis
- Genomic data isn’t useful without metadata

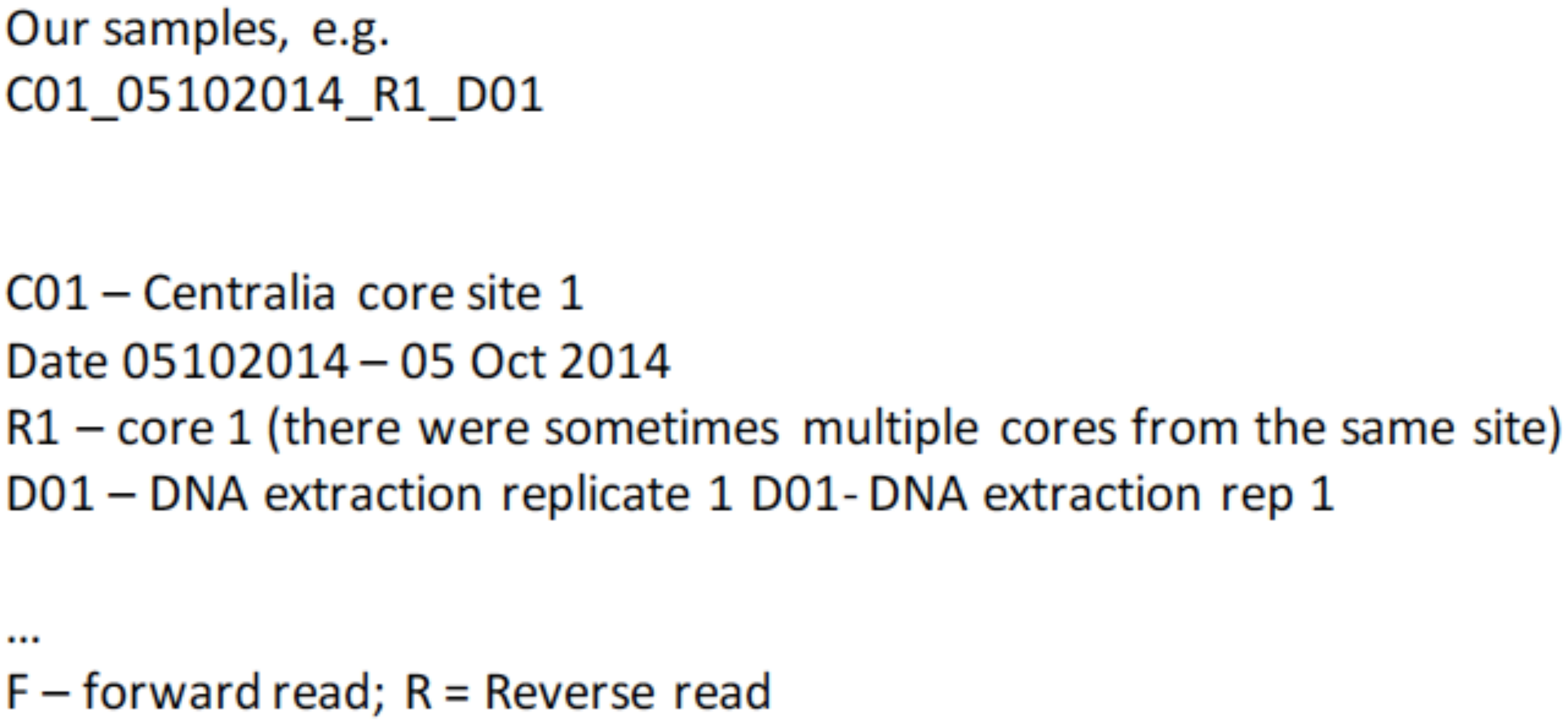
Data!

- Keep raw data raw
- Use meaningful names
- Organize your data so computers can read it

Keep raw data raw

- What is raw data?
- Why should I leave it alone?

Use meaningful names



Our samples, e.g.
C01_05102014_R1_D01

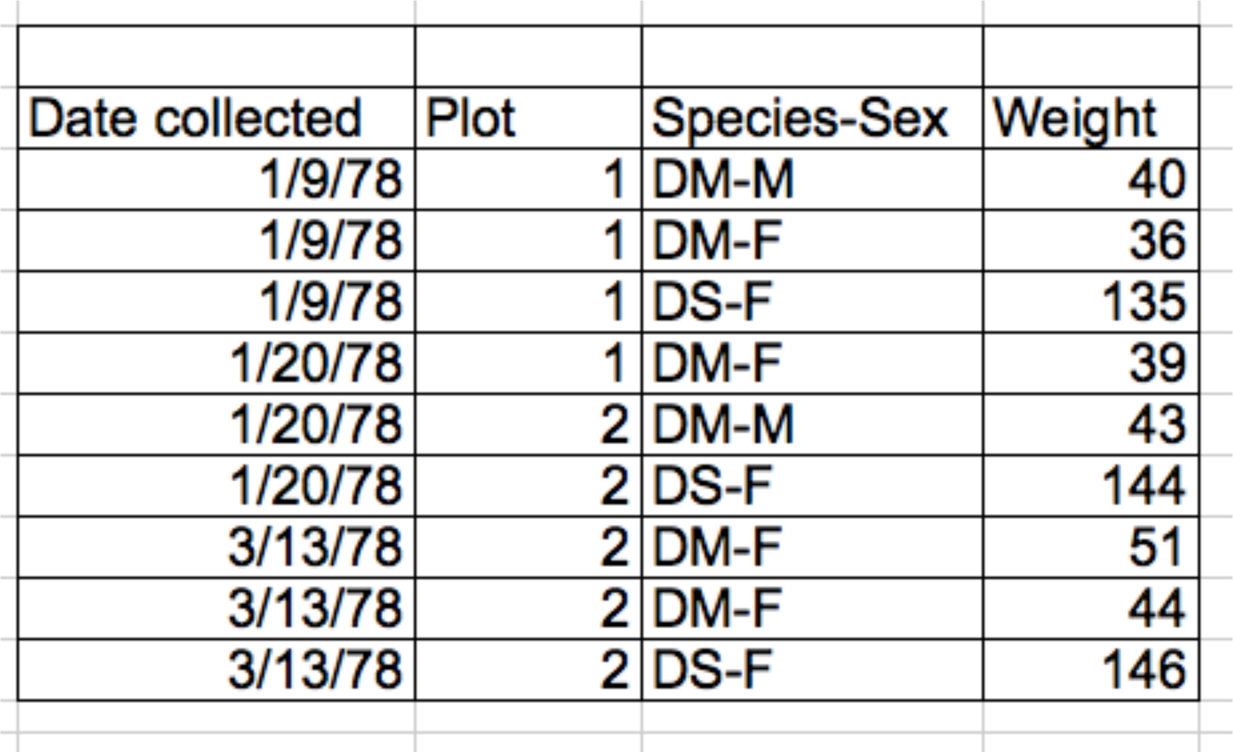
C01 – Centralia core site 1
Date 05102014 – 05 Oct 2014
R1 – core 1 (there were sometimes multiple cores from the same site)
D01 – DNA extraction replicate 1 D01- DNA extraction rep 1

...
F – forward read; R = Reverse read

Organize your data so computers can read it
(let’s talk about spreadsheets)

… also avoid formatting errors

http://www.datacarpentry.org/spreadsheet-ecology-lesson/00-intro.html

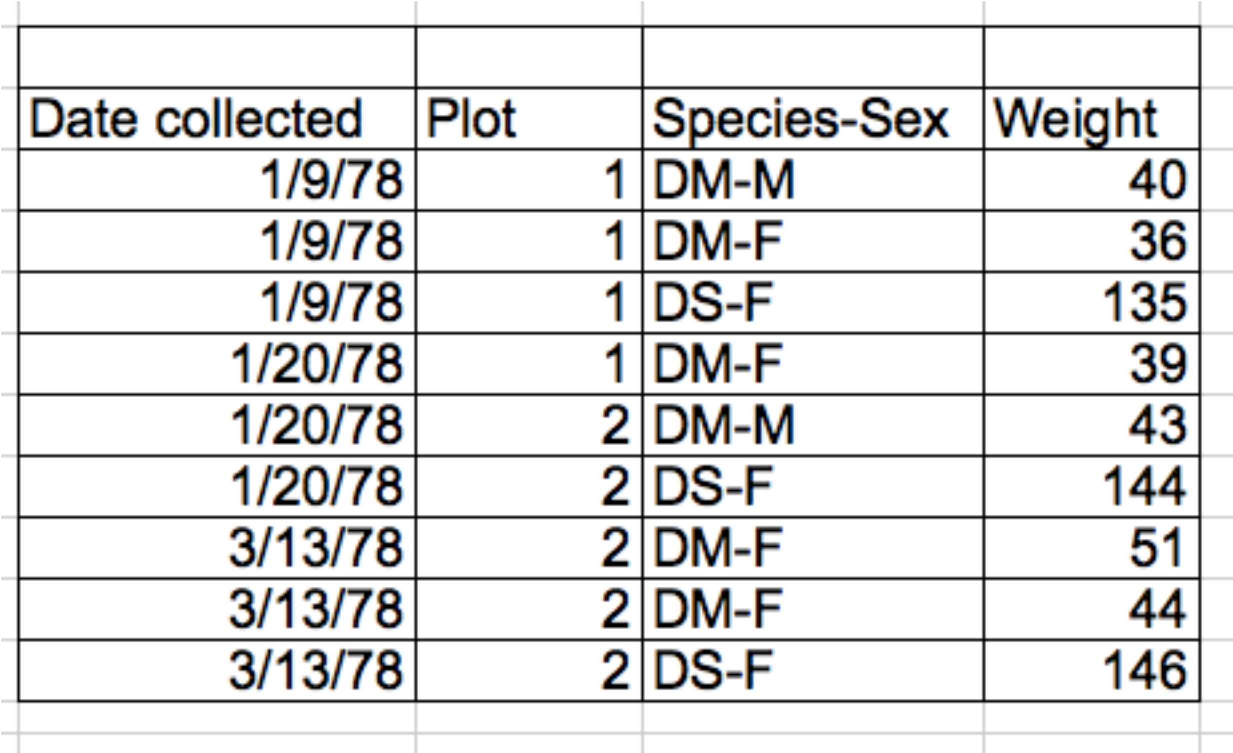


Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Organizing data in spreadsheets

The cardinal rules of using spreadsheet programs for data:

- Put all your **variables in columns** - the thing you're measuring, like 'weight' or 'temperature'.
- Put each **observation in its own row**.
- **Don't combine multiple pieces of information in one cell**. Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data.
- **Leave the raw data raw** - don't mess with it!
- Export the cleaned data to a **text based format** like CSV. This ensures that anyone can use the data, and is the format required by most data repositories.



Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Formatting problems

http://www.datacarpentry.org/spreadsheet-ecology-lesson/02-common-mistakes.html

A Roadmap for the Computing Biologist

- Consider the overarching goals of the analysis
- Adopt an Iterative, Branching Pattern to Systematically Explore Options
- Reproducibility Checkpoints
- Taking Notes for Computational Analysis
- Shared Responsibility: The Team Approach to Reproducibility and Data Management

Shade and Teal, Computing Workflows for Biologists: A Roadmap
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002303

Consider the Overarching Goals of the Analysis

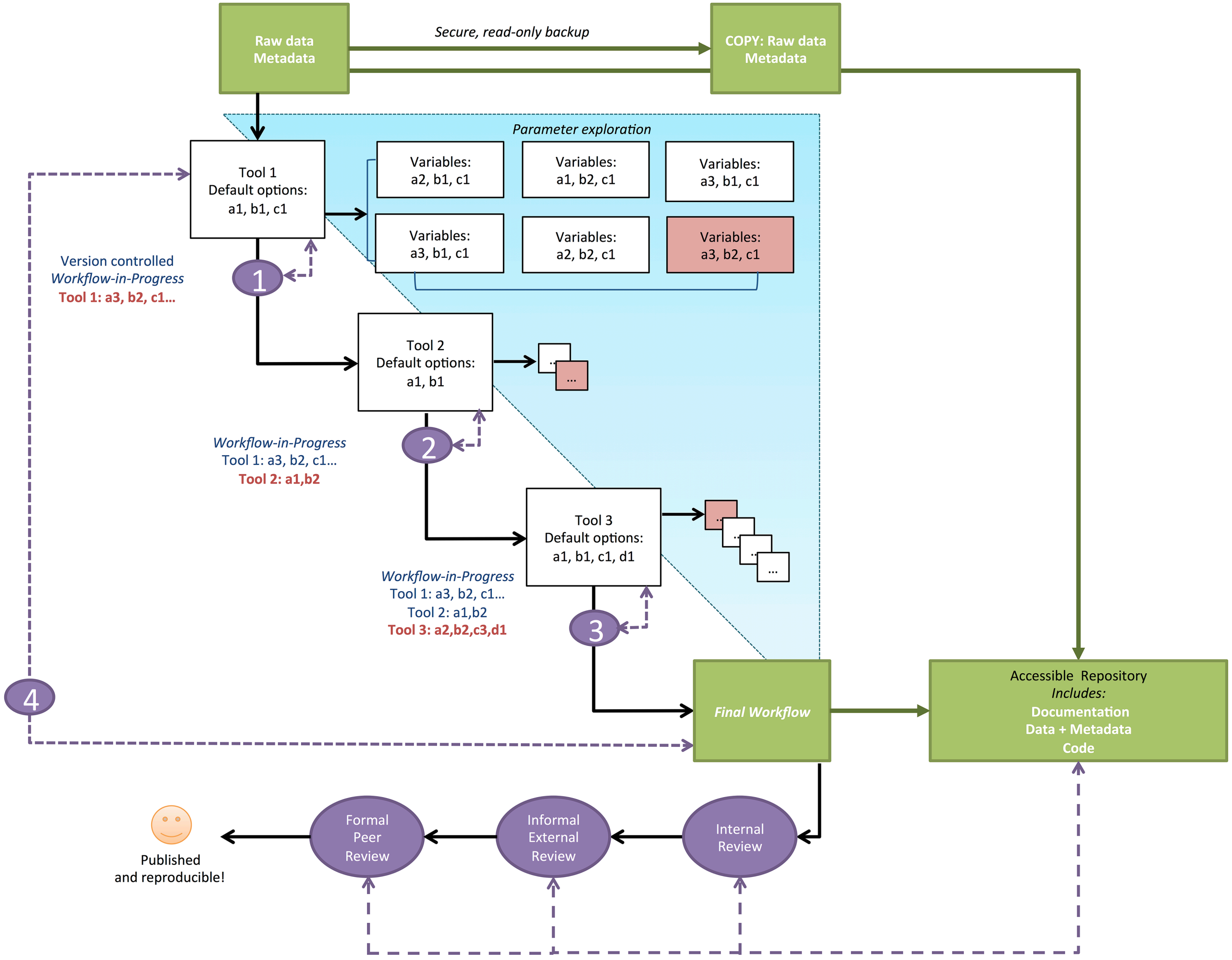
- Working to address a given hypothesis will motivate different analysis strategies than conducting data exploration

Reproducibility Checkpoints

Reproducibility checkpoints are places in a workflow devoted to scrutinizing its integrity

- the workflow (or step in the workflow) can be seamlessly used (it doesn't crash halfway or return error messages)
- the outcomes are consistent and validated across multiple, identical iterations
- results should make biological sense

Adopt an Iterative, Branching Pattern to Systematically Explore Options



Raw data
Metadata

Secure, read-only backup

COPY: Raw data
Metadata

Parameter exploration

Tool 1
Default options:
a1, b1, c1

Variables:
a2, b1, c1

Variables:
a1, b2, c1

Variables:
a3, b1, c1

Variables:
a3, b1, c1

Variables:
a2, b2, c1

Variables:
a3, b2, c1

1

Version controlled
Workflow-in-Progress
Tool 1: a3, b2, c1…

Tool 2
Default options:
a1, b1

…

…

2

Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2

Tool 3
Default options:
a1, b1, c1, d1

…

…

…

…

3

Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2
Tool 3: a2,b2,c3,d1

4

Final Workflow

Accessible Repository
Includes:
Documentation
Data + Metadata
Code

Internal
Review

Informal
External
Review

Formal
Peer
Review

Published
and reproducible!

Taking Notes for Computational Analysis

- Take notes like you would for experimental work
- Comment code
- Use version control (Github/Gitlab)

What needs to go in notes:

- Software versions used
- Description of what the software is doing/goal of that step
- Brief notes on deviations from default options
- Workflows can include different software (e.g., PANDAseq to QIIME to R), and should also include all "formatting steps" needed to move between tools hopefully you don't need to manually format too much; avoid if possible

Shared Responsibility: The Team Approach to Reproducibility and Data Management

We posit that integrity in computational analysis of biological data is enhanced if there is a sense of shared responsibility for ensuring reproducible workflows.

Research teams that work together to develop and debug code, perform internal reproducibility checkpoints for each other, and generally hold one another accountable for high-quality results likely will enjoy a low manuscript retraction rate, high level of confidence in their results, and strong sense of collaboration.

You, your lab mates and PI need to value the time it takes to do analyses reproducibly and correctly

Shared responsibility

- Shared storage and workspace can facilitate access to all group data
- Using version control repositories can provide access to code and documentation (Github, Dropbox)
- Setting expectations for ‘reproducibility checkpoints’ (team “hackathons”: open-computer group meetings dedicated to analysis)
- Paper reviews
- Looking for help/support outside the lab (bioinformatics or user groups, office hours, StackOverflow)

Exercise

http://tinyurl.com/mbl-workflows