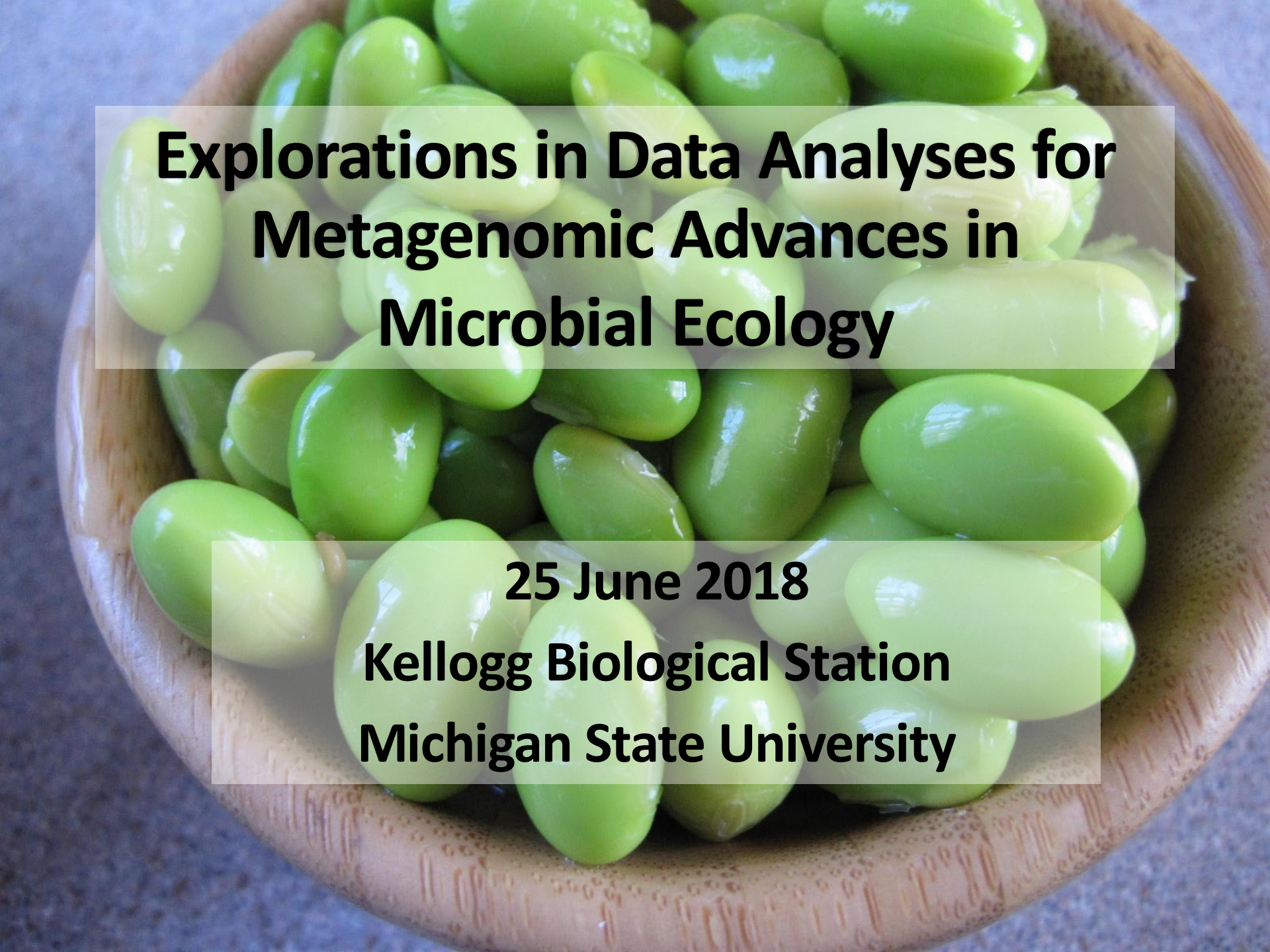


Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

25 June 2018

Kellogg Biological Station

Michigan State University

Case study: The Productive Post-doc Part #1

A very productive and confident post-doc in the lab, Jane, has performed some analyses on soil samples that were used for a 16S rRNA microbiome survey. She maintains the contextual data in an excel spreadsheet on her laptop, and backed up on the lab's server. To analyze these data, she routinely copies and pastes the values into a new "clean" tab to try different calculations, but she keeps the original data in the first tab. She also sometimes loads the excel tables into R for other statistical analyses. She keeps her R workflows on a private GitHub repository. In the end, she writes a paper describing her results, including an analysis of the explanatory value of the soil chemistry for microbiome community composition.

Discussion

1. Critique how Jane's stores and analyzes her data. What aspects of her strategy are reproducible? What aspects of her strategy can be improved?

Case study: The Productive Post-doc Part #2

A student in the lab, John, is a co-author on the paper that Jane is leading. He wants to check her workflow. He cannot access the private workflow on GitHub, and asks for the workflow to be made public. In the mean time, he checks the excel files that are backed up on the lab's servers and notices that a key sample, Sample1, is missing from all of the tabs in Jane's file. He also notices that some of the tabs were not copied correctly from the original data, and this has impacted some of the excel results.

Discussion

1. What should John do?
2. What aspects of this situation may make it difficult for John to decide what to do?

Case study: The Productive Post-doc Part #3

Jane makes the GitHub repo available to John and he reproduces the R workflow. He notices that the p-values that he generates are off from what Jane has reported in the results section of the paper. In fact, where Jane reports a p-value of 0.031, John generates a p-value of 0.31. John now feels very uncomfortable being an author on the paper.

Discussion

1. What would you advise John to do?
2. What steps should be taken before the paper is submitted?
3. Do you think that Jane's errors constitute misconduct or negligence? What is the difference?

Case study wrap-up & comparison to wet-bench

Consider the following scenarios:

1. A lab won't share their modifications to a DNA extraction protocol that they used to generate a fungal ITS leaf survey
2. A post-doc takes all of the freezer stocks of her genetic constructs when she moves to her new faculty position.
3. A graduate student never takes laboratory notes, and instead writes calculations on paper scraps and then discards.

What is a computing workflow?

- Exactly what you tell the computer to execute the analysis
- Each optimized step in a computing analysis
  - Verbatim scripts that were executed
  - Annotated:
    - 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., mothur 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

Workflows should be mindlessly complete – the computer is a literal beast

The Peanut Butter and Jelly Robot

https://www.youtube.com/watch?v=leBEFaVHllE

https://www.youtube.com/watch?v=Y-UEdr1wofM

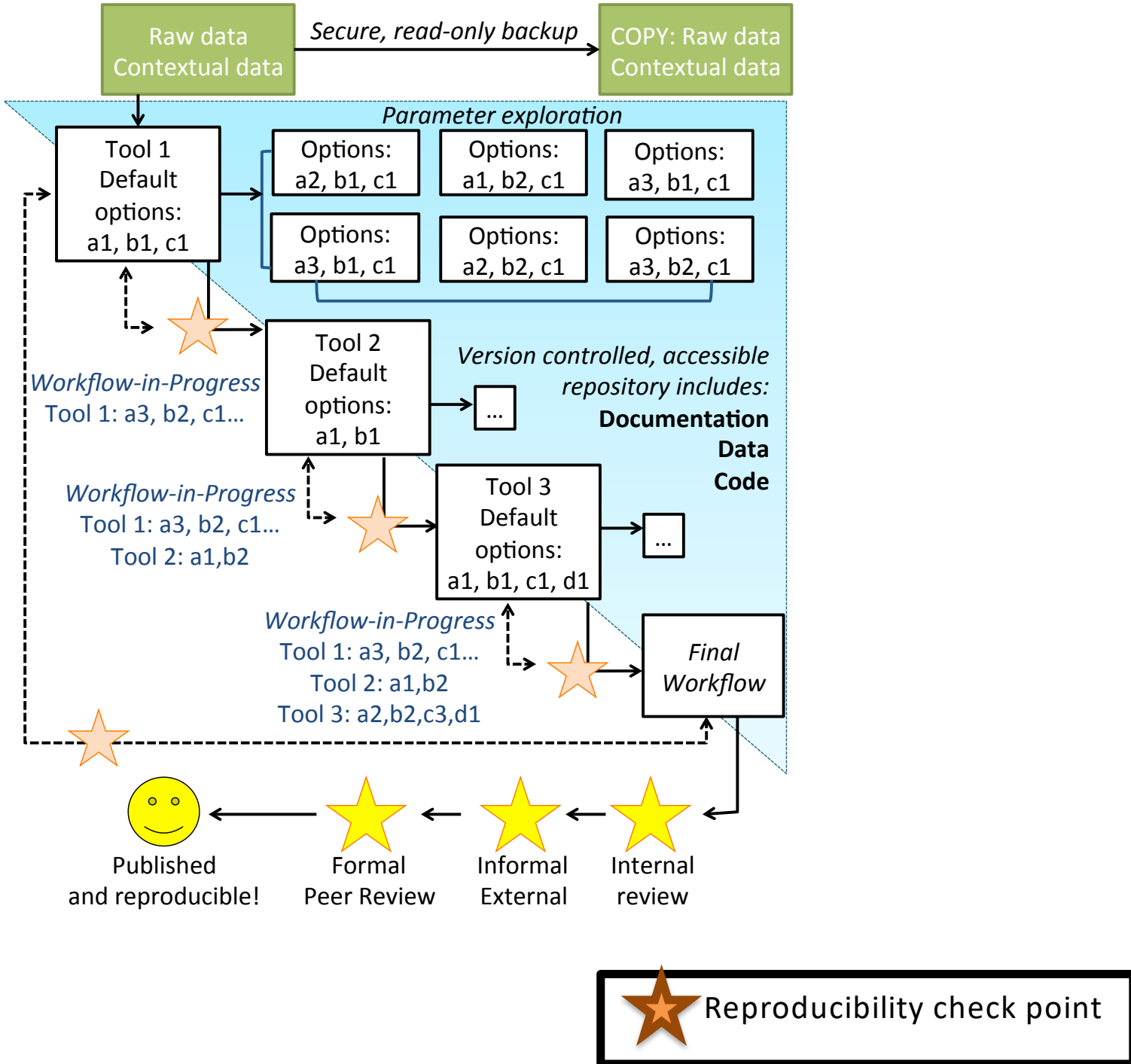
Computing Workflows for Biologists

- Papers of interest:
  - Wilson et al. 2014. Best practises for Computing. *PLoS Computational Biology*
  - Nobel 2009. Organizing Computational Biology Projects. *PLoS Computational Biology*
  - Sandve et al. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology.*

» *All of these references are posted in our Mendeley group*

Our suggestions for an analysis approach

1. Adopt a systematic, iterative exploration of parameter space.
   - Include “sanity checks”
   - Focus on exploring the parameters that *matter* for your objective/hypothesis
   - Organize your output and input for *someone who isn’t you*
2. Work towards an optimized, seamless workflow. Minimize manual steps.
3. Implement *reproducibility check-points.*
4. Maintain computing notes just as you would experimental notes
5. Do your part: cultivate a shared responsibility for reproducibility of results and data management


Raw data
Contextual data
Secure, read-only backup
COPY: Raw data
Contextual data
Parameter exploration
Tool 1
Default
options:
a1, b1, c1
Options:
a2, b1, c1
Options:
a1, b2, c1
Options:
a3, b1, c1
Options:
a3, b1, c1
Options:
a2, b2, c1
Options:
a3, b2, c1
Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2
Default
options:
a1, b1
…
Version controlled, accessible repository includes:
Documentation
Data
Code
Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2
Tool 3
Default
options:
a1, b1, c1, d1
…
Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2
Tool 3: a2,b2,c3,d1
Final
Workflow
Published
and reproducible!
Formal
Peer Review
Informal
External
Internal
review
Reproducibility check point

Heartaches: Naming Conventions

Exampl
20_A_T
rep1)

int 1,

Exampl
Ashley’s
A
Ashley

Exampl
ALS1, A

Improve
ALS01,

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

Subsampling – sometimes required for a large dataset to troubleshoot efficiently through an entire workflow

- Check out our tutorial about subsampling:

- https://github.com/edamame-course/2015-tutorials/blob/master/final/2015-06-23-QIIME1.md#ampliconsubsampling

Workflow Discussion - Etherpad

- Have you ever made a computing workflow? If so, what advice do you have?
- Has anyone ever checked your computing or statistical work? In that situation, what went well and what didn't?
- What do you think are the most challenging aspects of making computing workflows?