Healthy Habits for Data Science

Day 1: Open Science, Project organization
Open science concepts

- Open Access
- Open Data
- Open Educational resources
- Open Source (code)
- Open Protocols
FAIR

Findable

Accessible

Interoperable

Reusable
3-2-1 Rule for data backups

3 copies

2 formats

1 offsite
Avoidable problems

• Research methods dead end
• Poorly documented hand-me-down data
• Loss of data due to hardware failure/natural disaster
• Data not available for reanalysis
• Excel conversion problems
• Paper: https://doi.org/10.7554/eLife.81979
• GitHub: https://github.com/shahlab/LTEE_gene_expression_2/
"The landscape of transcriptional and translational changes over 22 years of bacterial adaptation" Favate et al 2022

E. coli exponential growth → Cell lysis → Footprint recovery → RNA recovery

12 evolved lines x 2 replicates

Are genes differentially expressed between the evolved lines and the ancestor? And how does that affect fitness?
Sequencing
• download from sequencing facility
• rename files

Clean-up
• remove primers
• de-multiplex
• de-duplicate
• trim ends
• deplete rRNA

Analysis
• quantify reads
• clean/combine data
• differential expression

Plotting
• ggplot in R

Fastq
Fastq
CSVs
Jpg/Tiff
Sequencing

Clean-up

Analysis

Plotting

data, raw

data, clean

numbers, analyzed data

images
Sequencing
Clean-up
Analysis
Plotting

data, raw
data, clean
numbers, analyzed data
images
1. Log in to FASRC cluster
2. Navigate to your lab’s folder in /holyscratch01/
3. Make a folder YOUR_USERNAME/healthy-habits
4. Copy today’s data into your folder
   - `cp /n/holylfs05/LABS/informatics/Everyone/workshop-data/healthy-habits-2024/day1/* .`
5. Make subfolders “software”, “data_raw”, and “scripts” and organize the files using `mv`
6. Make a project directory on your own computer
7. Use your preferred method (scp, rsync, Filezilla, etc) to copy the text files over to your computer from the cluster