

# 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

F

A

I

R



Findable



Accessible



Interoperable



Reusable



HMS Biomedical Data Lifecycle  
<https://datamanagement.hms.harvard.edu/>  
<https://hlrdm.library.harvard.edu/>

# 3-2-1 Rule for data backups



3 copies



2 formats



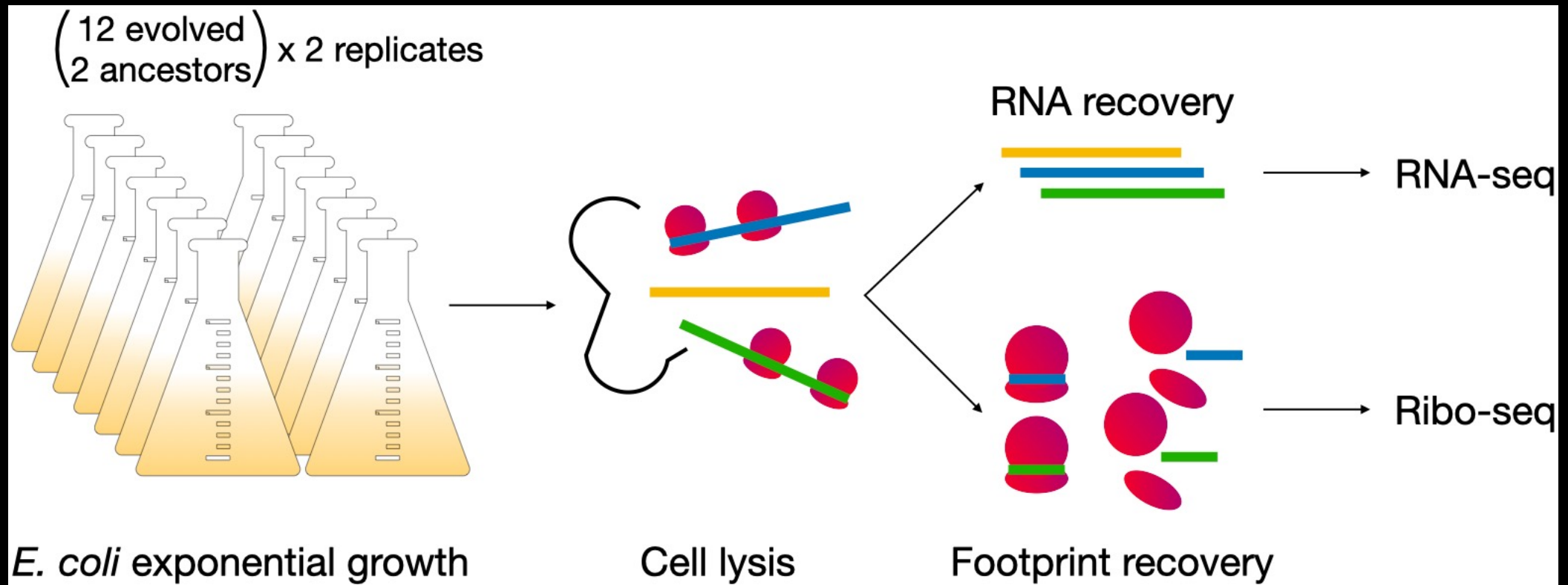
1 offsite

# 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

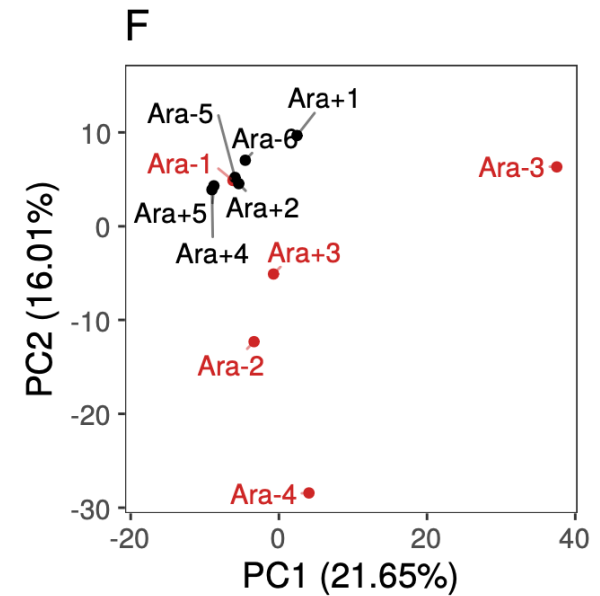
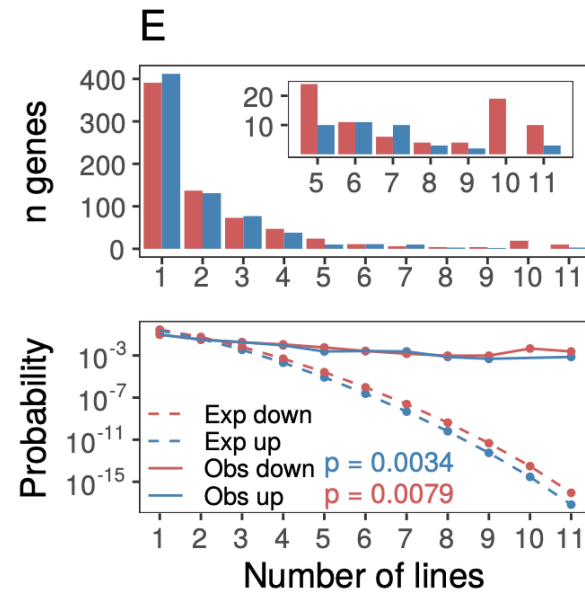
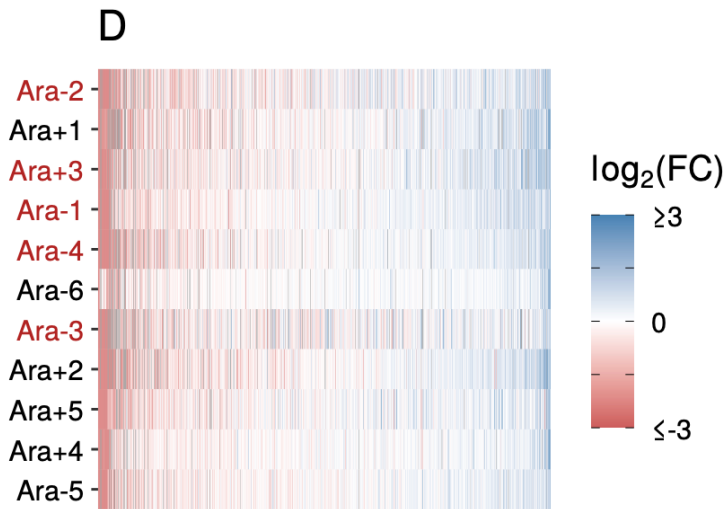
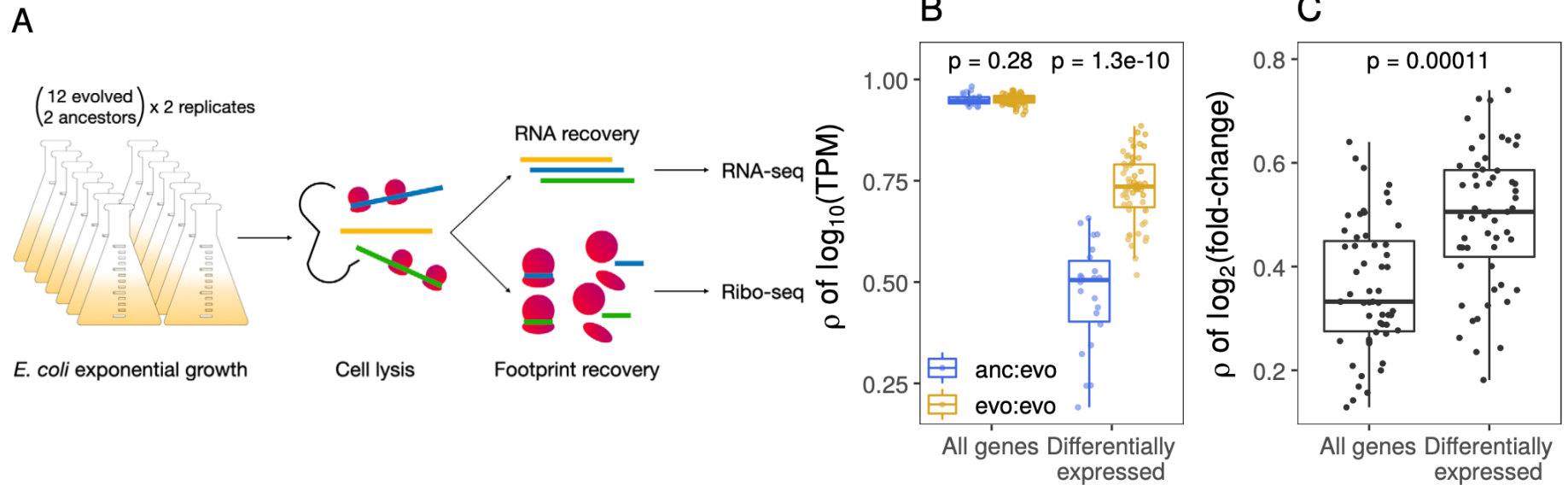
- Paper : <https://doi.org/10.7554/eLife.81979>
- GitHub: [https://github.com/shahlab/LTEE\\_gene\\_expression\\_2/](https://github.com/shahlab/LTEE_gene_expression_2/)

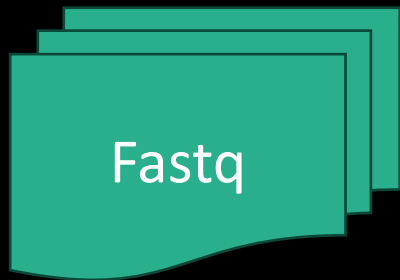
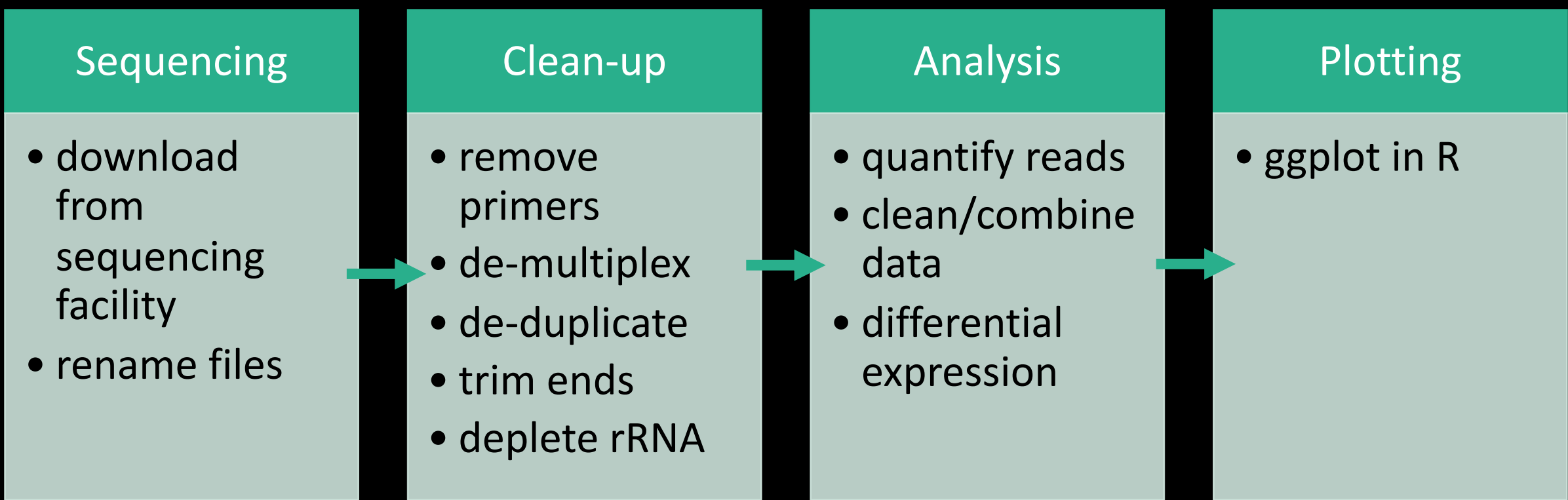
# "The landscape of transcriptional and translational changes over 22 years of bacterial adaptation" Favate et al 2022

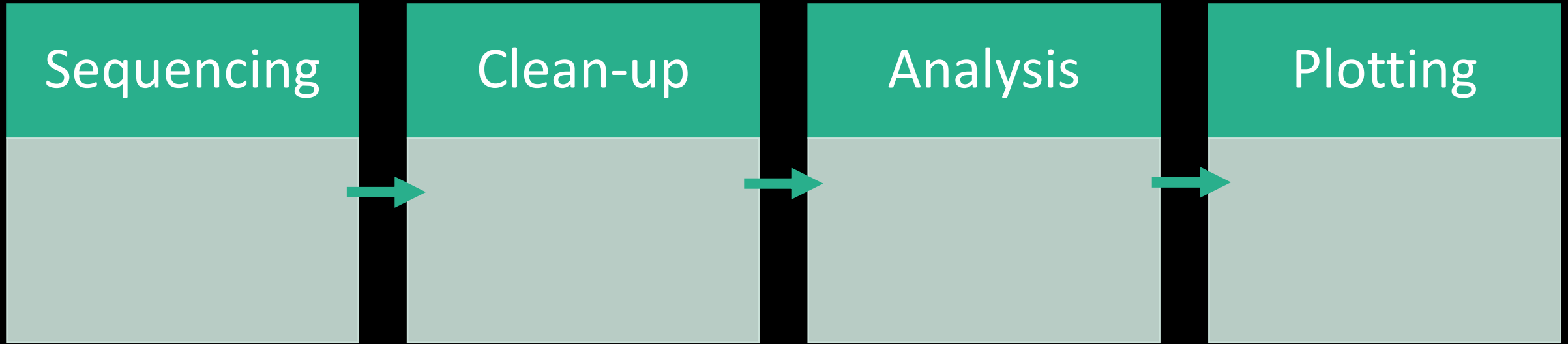


Are genes differentially expressed between the evolved lines and the ancestor? And how does that affect fitness?









data, raw

data, clean

numbers,  
analyzed  
data

images

## Name



figure\_1\_flipper\_bill\_11\_28\_2023.R



figure\_1\_flipper\_bill.png



figure\_1\_flipper\_bill.R



penguins\_raw.csv



penguins.csv



process\_data\_script\_from\_friend\_WORKING.R



process\_data\_script\_from\_friend.R

Sequencing

Clean-up

Analysis

Plotting


data, raw


data, clean

numbers,  
analyzed  
data

images


Name

▼  data\_processed

 penguins.csv

 README.md

▼  data\_raw

 penguins\_raw.csv


 README.md


▼  figures


 figure\_1\_flipper\_bill.png

▼  scripts

 figure\_1\_flipper\_bill.R

 process\_data.R

 penguins\_notebook.md

 environment.yml

/n/

|└ home/

| |└ lei/

| |└ jharvard/

|└ holylfs/

| |└ LABS/

| | |└ jharvard\_lab/

| | | |└ Everyone/

| | | |└ Lab/

| | | |└ Users/

|└ holyscratch01/

| |└ jharvard

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/holy1fs05/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



Sequencing

Clean-up

Analysis

Plotting

Fastq

Fastq

CSVs

Jpg/Tiff

Public Repo  
NAS  
RC storage  
HDD/SSD  
Harvard dataverse

NAS  
RC storage  
Not super important to store

Public Repo  
Paper supplement  
HDD/SSD  
Your computer  
Cluster Lab directory

Public Repo  
Paper  
HDD/SSD