Good ways to structure your data analysis projects

Gian Maria Niccolo’ Benucci (Nico)

March-29-2022

Where do we start when we begin a new project?

Goal: How to structure your working directory and your code to benefit you and your collaborators.

- What kind of project is it?
- Will I work in my computer or in a HPCC or both?
- Will it be mainly R or also Python, Shell etc. ?
- Does it need to be shared with others?
  - Do you need a Version Control system?

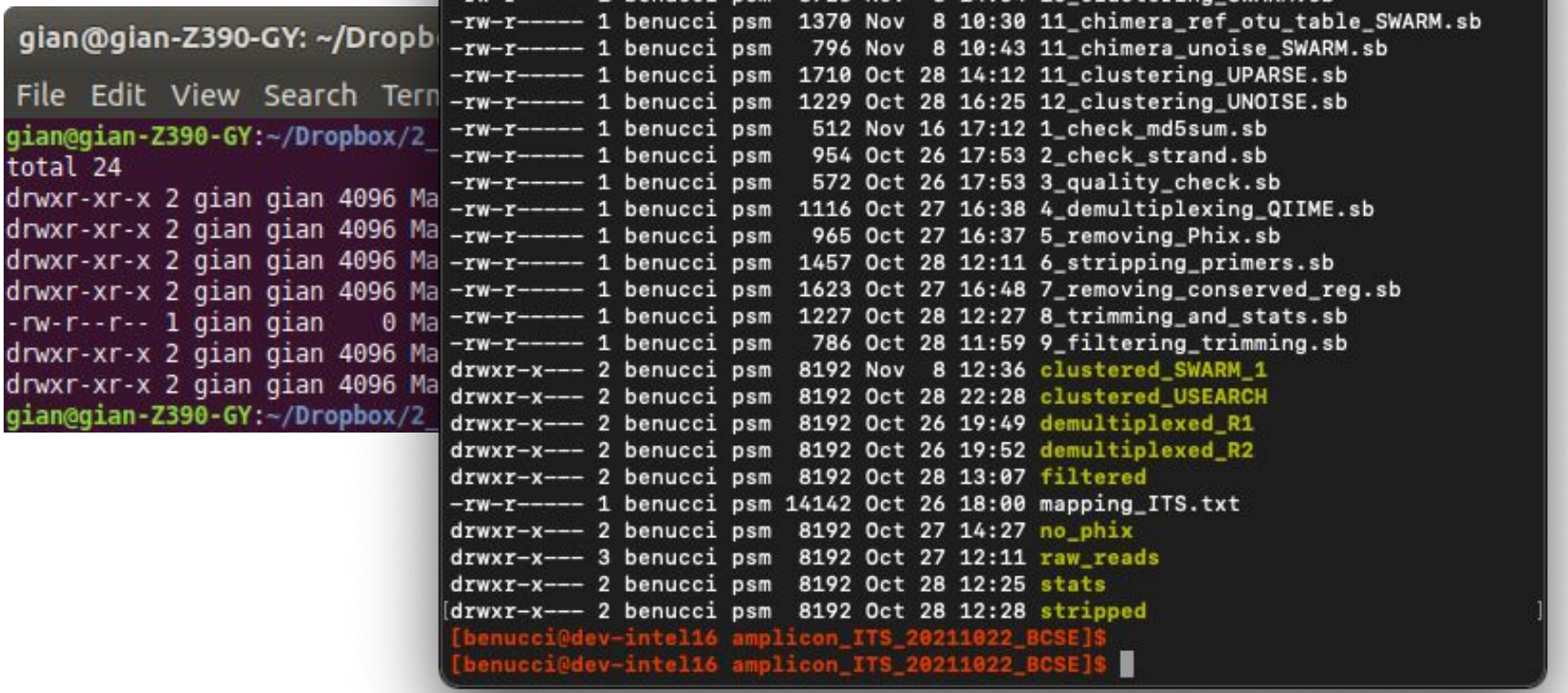
https://en.wikipedia.org/wiki/Coding_conventions

- Writing quality code benefits you and others;
- It saves time (e.g. No second edit and fix);
- It can be re-run more easily.

Main guidelines to follow…

- Use meaningful names (no spaces, ., -, etc.);
- Use DRY code (Don’t Repeat Yourself);
- Add documentation and comments (e.g. README.md).

An example


gian@gian-Z390-GY: ~/Dropb
File Edit View Search Tern
gian@gian-Z390-GY:~/Dropbox/2_
total 24
drwxr-xr-x 2 gian gian 4096 Ma
drwxr-xr-x 2 gian gian 4096 Ma
drwxr-xr-x 2 gian gian 4096 Ma
drwxr-xr-x 2 gian gian 4096 Ma
-rw-r--r-- 1 gian gian 0 Ma
drwxr-xr-x 2 gian gian 4096 Ma
drwxr-xr-x 2 gian gian 4096 Ma
gian@gian-Z390-GY:~/Dropbox/2_
gian-21 — benucci@dev-intel16:~/amplicon_ITS_20211022_BCSE — ssh b...
[benucci@dev-intel16 amplicon_ITS_20211022_BCSE]$ ll
total 143
-rw-r----- 1 benucci psm 3725 Nov 8 14:54 10_clustering_SWARM.sb
-rw-r----- 1 benucci psm 1370 Nov 8 10:30 11_chimera_ref_otu_table_SWARM.sb
-rw-r----- 1 benucci psm 796 Nov 8 10:43 11_chimera_unoise_SWARM.sb
-rw-r----- 1 benucci psm 1710 Oct 28 14:12 11_clustering_UPARSE.sb
-rw-r----- 1 benucci psm 1229 Oct 28 16:25 12_clustering_UNOISE.sb
-rw-r----- 1 benucci psm 512 Nov 16 17:12 1_check_md5sum.sb
-rw-r----- 1 benucci psm 954 Oct 26 17:53 2_check_strand.sb
-rw-r----- 1 benucci psm 572 Oct 26 17:53 3_quality_check.sb
-rw-r----- 1 benucci psm 1116 Oct 27 16:38 4_demultiplexing_QIIME.sb
-rw-r----- 1 benucci psm 965 Oct 27 16:37 5_removing_Phix.sb
-rw-r----- 1 benucci psm 1457 Oct 28 12:11 6_stripping_primers.sb
-rw-r----- 1 benucci psm 1623 Oct 27 16:48 7_removing_conserved_reg.sb
-rw-r----- 1 benucci psm 1227 Oct 28 12:27 8_trimming_and_stats.sb
-rw-r----- 1 benucci psm 786 Oct 28 11:59 9_filtering_trimming.sb
drwxr-x--- 2 benucci psm 8192 Nov 8 12:36 clustered_SWARM_1
drwxr-x--- 2 benucci psm 8192 Oct 28 22:28 clustered_USEARCH
drwxr-x--- 2 benucci psm 8192 Oct 26 19:49 demultiplexed_R1
drwxr-x--- 2 benucci psm 8192 Oct 26 19:52 demultiplexed_R2
drwxr-x--- 2 benucci psm 8192 Oct 28 13:07 filtered
-rw-r----- 1 benucci psm 14142 Oct 26 18:00 mapping_ITS.txt
drwxr-x--- 2 benucci psm 8192 Oct 27 14:27 no_phix
drwxr-x--- 3 benucci psm 8192 Oct 27 12:11 raw_reads
drwxr-x--- 2 benucci psm 8192 Oct 28 12:25 stats
drwxr-x--- 2 benucci psm 8192 Oct 28 12:28 stripped
[benucci@dev-intel16 amplicon_ITS_20211022_BCSE]$
[benucci@dev-intel16 amplicon_ITS_20211022_BCSE]$

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

OPEN ACCESS phytobiomesjournal.org APS
Phytobiomes Journal
A Transdisciplinary Journal of Sustainable Plant Productivity

PERSPECTIVE

Community-Driven Metadata Standards for Agricultural Microbiome Research

J. P. Dundore-Arias,[1,†] E. A. Eloe-Fadrosh,[2] L. M. Schriml,[3] G. A. Beattie,[4] F. P. Brennan,[5] P. E. Busby,[6] R. B. Calderon,[7] S. C. Castle,[8] J. B. Emerson,[9] S. E. Everhart,[10] K. Eversole,[11] K. E. Frost,[12] J. R. Herr,[13] A. I. Huerta,[14] A. S. Iyer-Pascuzzi,[15] A. K. Kalil,[16] J. E. Leach,[17] J. Leonard,[18] J. E. Maul,[19] B. Prithiviraj,[20] M. Potrykus,[21] N. R. Redekar,[22] J. A. Rojas,[23] K. A. T. Silverstein,[24] D. J. Tomso,[25] S. G. Tringe,[26] B. A. Vinatzer,[27] and L. L. Kinkel[28]

ABSTRACT

Accelerating the pace of microbiome science to enhance crop productivity and agroecosystem health will require transdisciplinary studies, comparisons among datasets, and synthetic analyses of research from diverse crop management contexts. However, despite the widespread availability of crop-associated microbiome data, variation in field sampling and laboratory processing methodologies, as well as metadata collection and reporting, significantly constrains the potential for integrative and comparative analyses. Here we discuss the need for agriculture-specific metadata standards for microbiome research, and propose a list of "required" and "desirable" metadata categories and ontologies essential to be included in a future minimum information metadata standards checklist for describing agricultural microbiome studies. We begin by briefly reviewing existing metadata standards relevant to agricultural microbiome research, and describe ongoing efforts to enhance the potential for integration of data across research studies. Our goal is not to delineate a fixed list of metadata requirements. Instead, we hope to advance the field by providing a starting point for discussion, and inspire researchers to adopt standardized procedures for collecting and reporting consistent and well-annotated metadata for agricultural microbiome research.

Keywords: genomics, meta-analysis, metagenomics, microbiome, omics, ontologies, phytobiome, synthetic

Terminal and HPCC Basics

Kristi Gdanetz MacCready

March-29-2022

Basic terminal navigation

- Text following `$` indicates what you should type
- Text in `Courier New` indicate commands
- Folder = “Directory”
- Important commands:
  - Change directory `$ cd`
  - Print working directory `$ pwd`
  - Make new directory `$ mkdir`
  - Remove directory `$ rmdir`
  - Move up one directory level `$ cd ..`
  - Home directory shortcut `$ ~/`

Basic terminal navigation

- Copy file:

```
$ cp file1.txt file1_copy.txt
```

- Rename file:

```
$ mv file1.txt file2.txt
```

- Move file:

```
$ mv ~/Downloads/file2.txt ~/Documents/file2.txt
```

- Delete file:

```
$ rm file1_copy.txt
```

Basic terminal navigation

- List files in directory:

```
$ ls
$ ls -lh
$ ls -a
```

- Read/view contents of file:

```
$ less
$ head
$ column
```

- Move to beginning of command prompt:

```
$ cntrl a
```

Useful shortcuts

- Save = `cmd s`
- Copy = `cmd c`
- Paste = `cmd v`
- Select all = `cmd a`
- Undo = `cmd z`

HPCC login

- We will do this together
- For more practice on your own time: https://wiki.hpcc.msu.edu/display/ITH/HPC+Tutorial+Series

To access a certain one, for example, dev-intel14, please log into HPCC and run: `ssh dev-intel14`

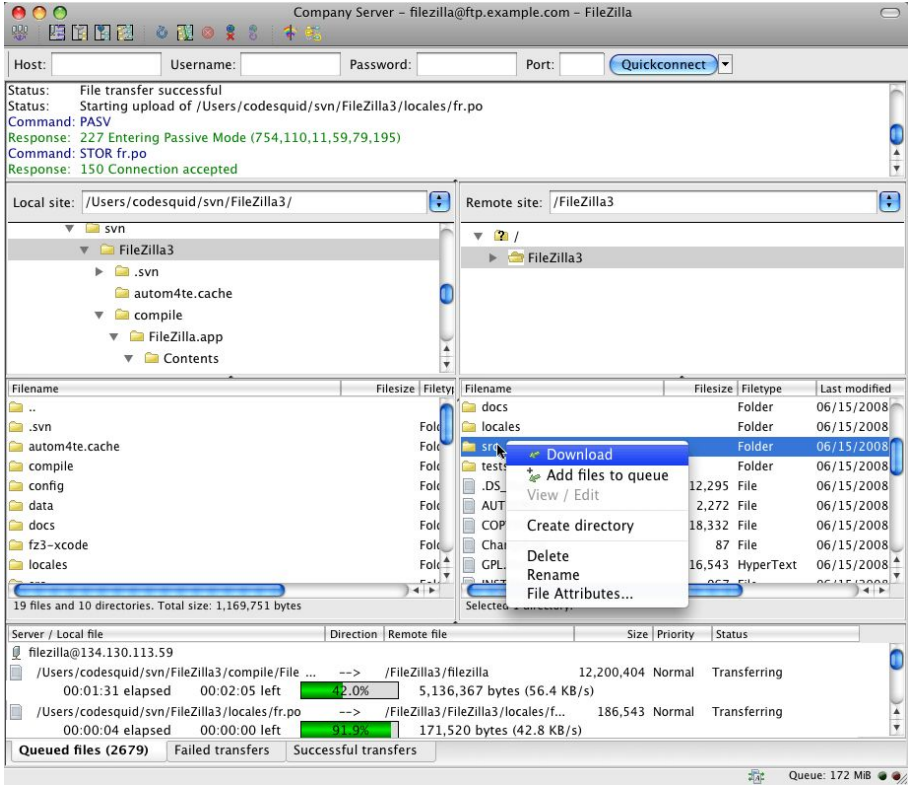
Node Hostname	Cores	Memory	Notes
dev-intel14	20	256GB	Large memory intel14 node
dev-intel14-k20	20	128GB	Two Nvidia K20 GPUs
~~dev-intel14-phi~~	~~20~~	~~128GB~~	~~Two Xeon Phi accelerators~~
dev-intel16	28	128GB	Two 2.4Ghz 14-core Intel Xeon E5-2680v4 (28 cores total)
dev-intel16-k80	28	256GB	Intel16 node with 4 Nvidia Tesla K80 GPUs
dev-intel18	40	377GB	Two 2.4Ghz 20-core Intel Xeon Gold 6148 CPU (40 cores total)
dev-amd20-v100	48	187GB	Intel(R) Xeon(R) Platinum 8260 CPU @ 2.40GHz and 4 Tesla V100S
dev-amd20	128	960GB	AMD EPYC 7H12 64-Core Processor @ 2.6GHz

Conda environment

- Initial anaconda install: https://wiki.hpcc.msu.edu/display/ITH/Using+conda
- Import commands:
  - `$ conda env list`
  - `$ conda create --name CONSTAX python=3.6`
  - `$ conda activate CONSTAX`
  - `$ conda deactivate CONSTAX`
  - `$ conda env remove --name CONSTAX`

FTP Client: FileZilla

- File Transfer Protocol (FTP)
- transfer files between a client and a server over the internet
- FTP Client is a program designed to transfer files between two computers



FTP Client: Globus

- Large data sets (do not need to maintain internet connection)
- File transfer between labs at MSU
- File transfer with external lab (other university)
- D2L course online: https://wiki.hpcc.msu.edu/display/ITH/Transferring+data+with+Globus

Long term storage

Research Space

Created by Chun-Min Chang, last modified by Xiaoge Wang on Dec 17, 2021

Research space is created upon request from a MSU principal investigator (PI) for his or her research group. The initial quota is 50GB and 1,000,000 on number of files. PI can increase to 1TB on space size for free. Additional space above 1TB may be purchased by completing Large Quota Increase Requet form, where you can find the annual fee. This space is associated with your group name and located at /mnt/research/[group name] by default. It is accessible to all users who have been added to the group and convenient for sharing files in the research directory. Please set your umask and the file permissions appropriately so the group members can access files and directories in the space. See the section **Instructions for using research space** below for more details.

Same as home directories, all research directories are also stored in the IBM GPFS under /mnt/ufs18. Files are backed up automatically (except saved in **`nodr`** space). To access any file backup, please submit a ticket and let us know the paths to the files or the directory with the time frame you would like them restored.

To learn about the space quota and usage in your research directory, please check **Space quota** section in Home Space page. If you would like to have more than 1 million files in your research space, please refer to **Limit on number of files** section in Home Space page. Please also read the following sections for how to use your research space.