Research Data Management Plan for meta-analysis (year 1):
This research relies on re-analyzing raw data and phylogeny results from multiple phylogenetic research papers. I consider it my responsibility to find accurate sampling and sequencing data, also, keep records of downloading and re-using them. In return, I will share organized reader-friendly analyzing procedures and results on publicly accessible online/offline platforms. All the data/records depositing and sharing plan is as follow:

Data Types	Where to store?	How to share?	Files for Documentation?	File Format	Nomenclature
Literature	Endnote	Endnote Library	-	pdf	-
Review of literature	Evernote (research data of PhD) & Folder (PhD Research)	-	Standard: Count those in the same genus. When subsect exist, only subsect.	mult-txt	-
Raw sequencing data	server (~500Gb is needed)	Public sequencing vault	*README_raw	fa/fq	RAD_genus_raw_20181010.fq
Intermediate files	server (~1T is needed, temporarily)	-	*README_mid	bam, csv, txt	RAD_genus_consensus_20181010.bam
Code scripts and procedures	server & GitHub & my computer	GitHub	*README_code	.py .sh .R .ipynb	01_draw_synteny.py (directory structure should be self-explanatory)
Public softwares and pipelines	server	-	*README_soft	-	-
Final results (figures, tables)	Folder (PhD Research)	DataSync	*NOTE_for_xxx	svg, ai, ps, pdf, xls, jpg	
Writing	Folder (writing up my PhD)	Publications	-	doc, pdf	-
Research diary	notebooks & Evernote (Record of PhD)	-	-	-	-
Discussion with supervisors and peers	Evernote (Record of PhD & Record of Supervisional Meetings)	-	-	-	2018.10.10

*This plan might change along the process of the research.
*the University of Edinburgh DataShare repository (http://datashare.is.ed.ac.uk/)
*data on my computer subject to weekly backup (Saturday) on hard drives.

***

*README_raw
Source of data (what, when, where, why, and how):
Data type (standards and methodologies):
Reference:
Downloading date:
Data size:
md5:
Back up dir:

*README_mid
Source of data:
Command and parameters:
Software Version/script dir:
Should be kept until:

*README_code
Description:
Usage:
date of first version:
Date of last update:
Back up dir:
Github branch:
Reference (If applicable):

*README_soft (usually included in files, or on webpage)
Downloading date:
version:
updating records:
parameters:
Reference (If applicable):

*NOTE_for_xxx
Figure legend should begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used. For contributions with methods sections, legends should not contain any details of methods, or exceed 100 words (fewer than 500 words in total for the whole paper).

Data -> information -> knowledge (what we understand based upon information) -> wisdom (use knowledge in decision making)

only 0.5% of the data are properly explored.

What data will you collect or create, how will it be created, and for what purpose?

▪ How will you manage any ethical issues? How will you manage copyright and Intellectual Property Rights issues?

▪ What file formats will be used? Are they non-proprietary, transparent and sustainable? What directory and file naming conventions will be used? Are there any formal standards that you will adopt? What documentation and metadata will accompany the data?

▪ How will the data be stored and backed up during the research? How will you manage access and security? Who will be responsible for data management?

▪ Are there existing procedures that you will base your approach on? For example, are there institutional data protection or security policies to follow, department or group data management guidelines, or Research Data Management policies defined by your institution or funder that must be considered?

▪ What is the long-term preservation plan for the dataset? For example, which data should be retained, shared, and/or preserved? How will you share the data, and are any restrictions on data sharing required?

▪ What resources will you require to deliver your plan? For example, are there tools or software needed to create, process, or visualise the data?

Read me may include:

- Admin:
- • Name and ID of the project
- • Project Description
- • Funding body/bodies
- • Principal Investigator name and possibly ID
- • Project Data Contact
- • Related Policies
- • Date of First Version
- • Date of Last Update
- Data collection
- • Data description, including anticipated type, format and volume
- • Existing datasets to be re-used
- • Methods by which data will be collected or created
- • Structures, naming and versioning system for folders and files
- • Quality assurance processes
- Metadata

Data management plans are a relatively new requirement for researchers; some say they require academic culture change. Here is a satirical response to what was then a new NSF requirement for

grantees from the blog post 'Daily Life in an Ivory Basement' by Dr. C Titus Brown (see MANTRA Acknowledgements to locate the original post).

Mon, 17 May 2010
My Data Management Plan - a satire

Dear NSF,

I am happy to respond to your request for a 2-page Data Management Plan.

First of all, let me say how enthusiastic I am that you have embraced this new field of "large scale data analysis". Ever since I started working with large Avida data sets in 1993, then with large meteorological data sets in 1995, and then again with large sequence data sets in 1999, I have seen the need for a systematic plan to manage the data. It is nice to see NSF stepping up to the plate in such a timely manner, and I am happy to comply.

Now, as to my actual data management plan, here is how I plan to deal with research data in the future.

I will store all data on at least one, and possibly up to 50, hard drives in my lab. The directory structure will be custom, not self-explanatory, and in no way documented or described. Students working with the data will be encouraged to make their own copies and modify them as they please, in order to ensure that no one can ever figure out what the actual real raw data is.

Backups will rarely, if ever, be done.

When required to make the data available by my program manager, my collaborators, and ultimately by law, I will grudgingly do so by placing the raw data on an FTP site, named with UUIDs like 4e283d36-61c4-11df-9a26-edddf420622d. I will under no circumstances make any attempt to provide analysis source code, documentation for formats, or any metadata with the raw data. When requested (and ONLY when requested), I will provide an Excel spreadsheet linking the names to data sets with published results. This spreadsheet will likely be wrong -- but since no one will be able to analyze the data, that won't matter.

Did I mention the click-through license? "You are provided this data for the sole purpose of reproducing our published results. Any attempt to publish your own analyses of this data will be rejected, if necessary during the anonymous review process, by pointing out all of the data cleanup steps you forgot to do correctly in your analysis. (We don't remember all of them ourselves, but there sure were a lot!) Give up now."

We will provide a short note -- in a Word document -- detailing the licensing restrictions, as above.

We will make sure that any CSV files we do eventually produce will have format errors, such as missing or extra commas. They will also be encoded in ISO 8859-16, "by accident".

On the off chance that we do choose to provide the source code, it will be in a file named 'source.tar.gz' that unpacks in to the current directory. There will be no explanation of contents, instructions on how to run it, or any enabling information -- it was hard to write, and it should be hard to run! Old, patched, or otherwise impossible-to-obtain versions of Redhat Linux, Perl 5, and associated CPAN libraries will be required before the code runs, even if it doesn't actually use any of them. No source code documentation

will be present, of course -- we don't need it ourselves, after all! Automated tests will also not be present (we don't have any of those, either). New versions of the code will be published under the identical file name, with no indication of what changes were made. (We'll be sure to use mixed DOS and Unix EOL editors for our files, so 'diff' won't work to figure out what has changed.)

Note, we didn't use a version control system, either. Or if we did, we made sure to use svn branching and merging profligately, with extremely obscure commit messages (our main programmer only speaks Chinese, so that's how she enters her commit notes. Wouldn't have it any other way). And our repository is not publicly available - you have to beg for permission. Note, I only answer e-mail on every other Tuesday.

Any design notes on the data analysis are in our private e-mail, and we will fight to the death -- up to and including ignoring FOIA requests -- to prevent you from obtaining them.

Meanwhile we will continue publishing exciting sounding (but irrerproducible) analyses, and submitting grants based on them, because that's the only thing that the reviewers care about.

sincerely yours,

--titus

(representing every computational scientist in the world.)