Managing Your Research Data

Data Structure and Formatting

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Reproducible Research

- At some point in the future someone might want to repeat your analysis for themselves or re-use your data

  This will most likely be you!

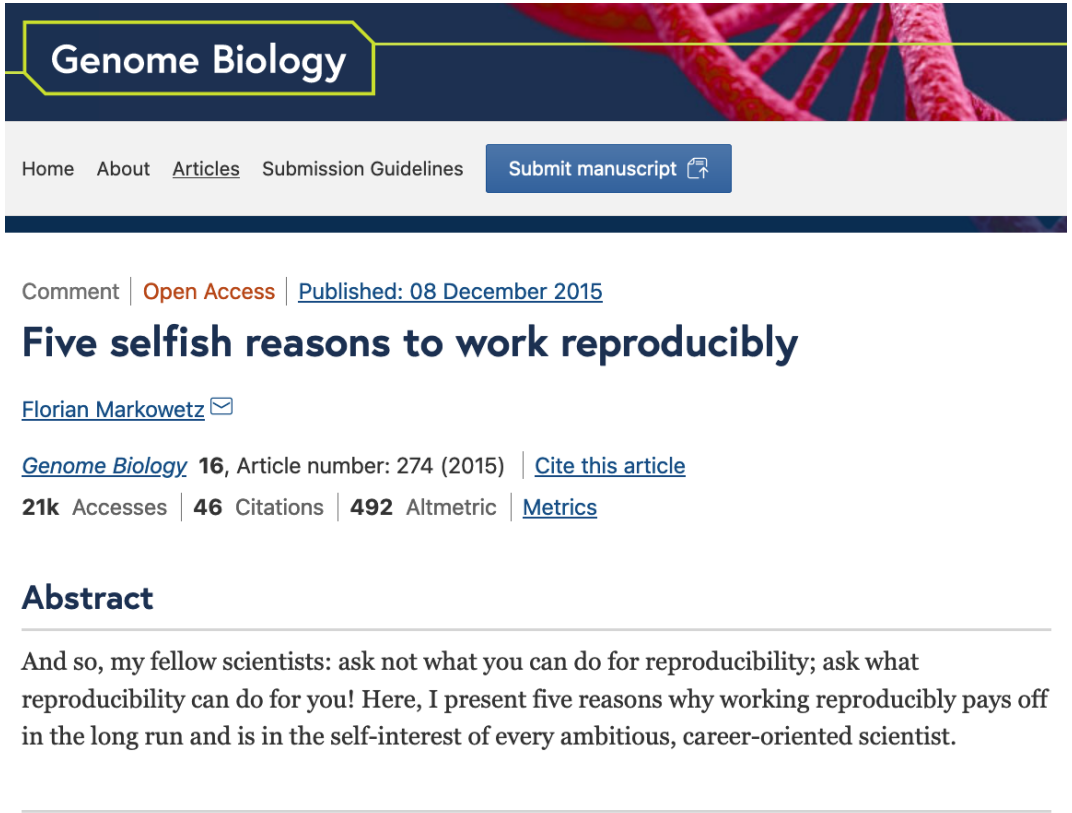
- Assuming that you'll be able to remember all the steps involved in generating the data is dangerous

  Making sure that everything is well documented is crucial

- Documentation should cover not only the methods used, but the files used as input and any transformations performed on them

Reproducible Research

- [Five selfish reasons to work reproducibly – Florian Markowetz (CRUK)]



YouTube video of Florian presenting this in a talk

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Reproducible Research

- Probably the most (in)famous example of failure to reproduce a study, which actually put people's lives at risk and rallied statisticians into action - Keith Baggerly's lecture on the scandal is a must-see.
- Retraction Watch

Are spreadsheets programs like Excel evil?

- … Not necessarily …
- Often much more convenient to eye-ball a spreadsheet and get an overall impression of your data
- But they have *limitations* making them not ideal for large-scale analyses
- Doing things by-hand increases the chances of mistakes, such as copy-and-paste errors
- Languages such as R or Python cannot read all files as if by magic

Data Validation in Excel for data entry

Excel and other spreadsheets have a data validation feature

e.g. In Excel

- Select a column
- In the menu bar, choose Data then Validation…
- You can then choose to limit the acceptable entries in the column to:
  - Integer or decimal number
  - A numeric range
  - List of possible text values
  - Limited length text
  - Date or time

Less helpful features in Excel

- When identifiers are long integers Excel converts them to exponents
  - 1000000 = 1e06
  - Issue with Illumina microarray chip IDs
- Excel can convert gene names to dates
  - SEPT2 (Septin 2) → '2-Sep'
  - MARCH1 (Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase) → '1-Mar'
- Conversion of ID codes!

Data Handling rules

- **Rule 1 - Never work directly on the raw data**
- **Rule 2 - Maintain consistency**
- **Rule 3 - Don't use 0 to mean missing**
- **Rule 4 - Fill in all the cells**
- **Rule 5 - Make it rectangular**
- **Rule 6 - Do your data entry in a timely fashion**

Data Handling rules

Rule 1 - Never work directly on the raw data

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Data Handling rules

Rule 1 - Never work directly on the raw data

- Hard to reverse all the manual steps performed
- Keep the original data somewhere **safe**
- Make a copy of the original and work on that
- Ideally write protect the original to avoid it being altered or overwritten

Data Handling rules

Rule 1 - Never work directly on the raw data

Write protection:

Mac

- Right click on the file in Finder
- Select “Get Info”
- Sharing and permission
- Under the “Privilege” column select “Read only”

Windows

- Right click on the file in Windows Explorer
- Properties
- General tab
- Attributes
- Select the box for “read only”

Data Handling rules

Rule 2 - Maintain consistency

How many inconsistencies can you spot?

Patient ID	Sex	Date of Diagnosis	Tumour Size
Patient_1	M	01-01-2013	3.1 cm
Patient_2	f	04-18-1998	1.5
Patient_3	Male	1st of April 2004	10.5
Patient_4	Female	NA	67
Patient_5	F	2010/03/12	4.2 cm
Patient_6	F		3.6 cm
Patient_7	M	1994-11-05	23.2 mm

Data Handling rules

Rule 2 - Maintain consistency

- Consistency: F, female, f, fem, 2, …
- Units - cm or mm; days, months or years
- You can introduce inconsistencies without realising it, e.g. blank spaces (whitespace) at the end of text - "Male " is not the same as "Male”
- Controlled vocabularies – standardize text entry option
- Use data validation where possible
- Document choices you make in a README file

Data Handling rules

Rule 2 - Maintain consistency

A note on dates:

- The formatting for dates varies widely e.g.:
  - 2nd May 2023
  - 02-05-2023
  - 2/5/2023
  - 5-2-2023
  - May 2nd 2023
- The clearest is to use YYYY-MM-DD
  - i.e. 2023-05-02
  - This is international standard ISO 8601

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Data Handling rules

Rule 2 - Maintain consistency

Corrected Data Table

Patient ID	Sex	Date of Diagnosis	Tumour Size (mm)
Patient_1	M	2013-01-01	31
Patient_2	F	1998-04-18	15
Patient_3	M	2004-04-01	10.5
Patient_4	F	NA	67
Patient_5	F	2010-03-12	42
Patient_6	F	NA	36
Patient_7	M	1994-11-05	23.2

Data Handling rules

Rule 3 – Don't use 0 to mean missing

- Zero values are data!
  - Sometimes extreme values such as 999 are sometimes used
- “NA” is okay, except if NA is a valid category in your data
  - Languages such as R or Python will recognise NA as a missing value and will ignore it in analyses
- Another option is to leave the cell *empty*
  - You need to be careful with blank spaces
  - Does it mean the data weren’t collected or was it a data entry error?

Data Handling rules

Rule 4 - Fill in all the cells

Sample ID	Date	Value
Sample_1	2015-06-14	213
Sample_2		76.5
Sample_3	2015-06-18	32
Sample_4		120.3
Sample_5		109
Sample_6	2015-06-20	95.6
Sample_7		143

Data Handling rules

Rule 4 - Fill in all the cells

- It is tempting to make the table look cleaner by not repeating some values
- Fill in all cells!
  - otherwise, problems when sorting
- Empty cell:
  - missing value?
  - value meant to be repeated multiple times?
- Make sure it’s clear that the data is missing and not unintentionally left blank

Data Handling rules

Rule 4 - Fill in all the cells

Sample ID	Date	Value
Sample_1	2015-06-14	213
Sample_2	2015-06-14	76.5
Sample_3	2015-06-18	32
Sample_4	2015-06-18	120.3
Sample_5	2015-06-18	109
Sample_6	2015-06-20	95.6
Sample_7	2015-06-20	143

Data Handling rules

Rule 5 - Make it rectangular

- Analysis software expects a very rigid shape of data with rows and columns
- Each column is a *variable* being examined
- Each row is an *observation*
- A concept commonly known as *tidy data*

Data Handling rules

Rule 5 - Make it rectangular

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

Data Handling rules

Rule 5 - Make it rectangular

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Data Handling rules

Rule 6 – Do your data entry in a timely fashion

- It is tempting to focus on data collection and leave data entry to later date
- This can lead to problems if your records are not perfect
- Lab/Field notebooks can be easily damaged or lost

Data Handling rules

- **Rule 1 - Never work directly on the raw data**
- **Rule 2 - Maintain consistency**
- **Rule 3 - Don't use 0 to mean missing**
- **Rule 4 - Fill in all the cells**
- **Rule 5 - Make it rectangular**
- **Rule 6 - Do your data entry in a timely fashion**

Additional best practices

- Don’t put too much information in one cell
  - 1 cell = 1 piece of information
- Don’t include units such as "30 g" → "g" in the column name
  - http://unitsofmeasure.org/ucum.html
- Write notes in a separate column or data dictionary or metadata
  - "0 (below threshold)”
- Design for machine readability
  - No calculations
  - Don’t use highlighting or font/fill colours to indicate data

Practical

- Download the file patient_data.txt to your computer
- Open it in Excel or any other spreadsheet software
- This is a simulated, but representative, example of ***bad data***
- Spend a few minutes identifying problems with the data
- In the next session we will look at the software OpenRefine, which we can use to clean up the problems with the data table.