Course Journal - Entry 1

17 September 2020

Yining Ding(1003898446)

Objective:

The objective of this assignment is to understand Plagiarism and academic integrityPlagiarism and academic integrity guidelines (Steipe, 2017) for BCB410. And develop some ideas about the package that I am going to make in this course.

Procedure:

1. Read through Plagiarism and academic integrity guidelines (Steipe, 2017) from 'How to reference' to 'Adjusting results'.

    - Link: http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Plagiarism

2. Read the Facets of R (Chambers, 2009)

    - Link: https://journal.r-project.org/archive/2009/RJ-2009-008/index.html

Results:

- How to reference

- We should **contextualize attributions** with statements (e.g inspired by, based on, etc.)
- **Reference in R**: citations should go directly into the comment
- **Image reference**: make sure to cite the author and context
- **Links**: must be original source

    - Code copied should reference the link to the post and the name of the author.

- **Citation style**: APA

- When and what to reference

- Code from R package vignette
- Code from this/other courses
- Code translated from other languages
- Code developed with/copied from other contributers
- Code that is open source/public domain

- Reusing previous material

- We can not get academic credit for the same material twice.
- Therefore, must cite these.

- "Adjusting" results

- Do not change anything in the printout to falsify yourself.

- What is the difference between NA and NaN in R?

- *NA* means that the data is missing for some reason

```
?NA             # NA is a logical constant of length 1 which contains a missing value indicator.
# For example, if we input a NA value (Howard, 2016)
c(1, 2, NA)   # [1]  1  2 NA
```

- *NaN* means that it is not a number and it can not be represented in the computer.

```
# For example, the square root of an unknown number (Howard, 2016)
sqrt(-1)      # /bin/sh: NaN: command not found
```

Conclusions

- I have learnt the difference between NA and NaN, their main difference is that NaN means that the data can not be represented in the computer, while NA means that the data is missing for some reason.

Next steps

- I have not developped a exact idea on what kind of R package/tool I may develop in the course. I noticed that on GEO many datasets are shared, but the publishers usually do not share the quality of the raw data. Therefore, when people try to analyze the data, they may question the quality of the data. I may want to develop a R package at the beginning of the NGS data processing workflow that can inform the user about the kind of instrument and the chemistry that the instrument used to produce the data. Therefore, when examining the bad data quality, bioinformaticians can understand the reason behind it. I understand there are many NGS platforms from different companies, therefore, I might just include instruments that are mostly used (such as Illumina).

- If I decide to make this tool, I would also need to learn the chemistry that Illumina uses for sequencing.

References:

1. Chambers, J. (2009). Facets of R. Retrieved September 17, 2020, from https://journal.r-project.org/archive/2009/RJ-2009-008/index.html

2. Howard, J. (2016, July 18). NaN versus NA in R. Retrieved September 17, 2020, from https://www.google.com/search?client=safari%2CWhat+is+the+difference+between+NA+and+NaN+in+R%3F

3. Steipe, B. (2017, August 05). ABC-Plagiarism. Retrieved September 17, 2020, from http://steipe.biochemistry.utoronto.ca/abc/index.php/ABC-Plagiarism