# **Manuscript Title**

This manuscript (permalink) was automatically generated from ferenckata/SQSeminarPaper@aff9359 on August 11, 2023.

### **Authors**

- John Doe
- Jane Roe <sup>⊠</sup>

Department of Something, University of Whatever; Department of Whatever, University of Something

☑ — Correspondence possible via <u>GitHub Issues</u> or email to Jane Roe <jane.roe@whatever.edu>.

## **Abstract**

### Introduction

This should probably be about bioinformatics and its software problems. Industry is the place with experience -> learn from them just because of that, not glorifying their ideas just acknowledging that they thought more

Bioinformatics and computational biology are continuously gaining importance in biological research. These fields are heavily relying on inventions of computer science and software engineering. As Goble pointed out in 2014, about 90% of researchers used or relied on results produced by scientific software [1]. This implies that incorrect software results in invalid scientific findings.

However, bioinformaticians lack formal education in computer science and related subjects [2]. Such lack of theoretical and practical foundations hinders the adoption of good practices (eg. continuous integration, unit tests, code reviews, planning of software architecture) established in other software-heavy endeavors. Many of these practices rely on redundancy of knowledge within team members, supported by practices such as pair programming, daily stand-up meetings and code reviews.

One main feature of academic research groups is the research projects conducted by a single PhD student or post-doc. In fact, academia prides itself on enabling individual achievements, and generally measures the success of individual effort while disregarding team effort. This often results in very few people maintaining the software product, which might be used by thousands of researchers worldwide [3,4]. Lack of funding contributes to the poor maintenance status of much of scientific software [1], which has been recognized and addressed by the Chan Zuckerberg Initiative Essential Open Source Software for Science fund [5].

The concept of a team is therefore different in a science project from a software project. While group members help each other with scientific suggestions, most often there is a single responsible person for the design and implementation of the code base. As official guidelines are rarely available on coding practices, the actual craft of software engineering is treated as secondary and up to individual judgment. We can observe (can we?) that these guidelines naturally emerge in larger groups (Seurat?), this requires a form of team organization not intrinsic to academic groups. In line with this, Hannay et al. [2] found that researchers tended to rank software engineering concepts higher if they worked in a team. Russel et al. [4] showed that high-profile code bases feature larger development teams than low-profile ones, and their activities are also associated with longer commit messages as a sign of better communication and documentation of changes. We asked ourselves whether a form of team structure organized around individual software products could improve the quality of our work.

In this work we review the findings of software engineer researchers on bioinformatics practitioners, and their suggestions for improving the overall quality of scientific software in bioinformatics and related fields (citations?). Inspired by the findings, especially related to process quality, our research group incorporated weekly discussions on code quality. We noticed that good practice requires investment in time and effort that may not pay off on an individual basis. Sharing the onboarding effort of new tools, and normalizing discussions on implementation details resulted in an overall better perceived process quality and code quality of the members.

We implemented a form of code review that fits our specific needs and context. We used our experience from these sessions, with special focus on team-based software development practices to ensure good code quality during in the update of the curation analysis software of the 2024 release of the JASPAR database (citation). In this project, as well as in some of our own individual research projects, we introduced user stories when documenting the assumptions, current features and new ideas, and we relied on development tools such as Jira and git. We note that the usage of these tools is not necessarily aligned with industry practices, due to the experimental nature of scientific software.

Nevertheless, as bioinformatics becomes a more and more software-heavy field, we believe a good direction is to collectively lower the barrier to adapting to new technologies.

We aim to provide guidelines on how to get started with improving the process quality of bioinformatics groups, even without members who have formal training in computer science or software engineering.

### References

#### 1. Better Software, Better Research

Carole Goble

IEEE Internet Computing (2014-09) https://doi.org/vjz

DOI: 10.1109/mic.2014.88

#### 2. How do scientists develop and use scientific software?

Jo Erskine Hannay, Carolyn MacLeod, Janice Singer, Hans Petter Langtangen, Dietmar Pfahl, Greg Wilson

2009 ICSE Workshop on Software Engineering for Computational Science and Engineering (2009-05) https://doi.org/bw966x

DOI: 10.1109/secse.2009.5069155

### 3. Empirical study on software and process quality in bioinformatics tools

Katalin Ferenc, Konrad Otto, Francisco Gomes de Oliveira Neto, Marcela Dávila López, Jennifer Horkoff, Alexander Schliep

bioRxiv (2022-03-13) https://doi.org/grx4jr

DOI: https://doi.org/10.1101/2022.03.10.483804

#### 4. A large-scale analysis of bioinformatics code on GitHub

Pamela H Russell, Rachel L Johnson, Shreyas Ananthan, Benjamin Harnke, Nichole E Carlson *PLOS ONE* (2018-10-31) <a href="https://doi.org/gskr8b">https://doi.org/gskr8b</a>

DOI: <u>10.1371/journal.pone.0205898</u> · PMID: <u>30379882</u> · PMCID: <u>PMC6209220</u>

#### 5. **CZI – Essential Open Source Software for Science**

Chan Zuckerberg Initiative

https://chanzuckerberg.com/eoss/

#### 6. Sci-Hub provides access to nearly all scholarly literature

Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene *eLife* (2018-03-01) <a href="https://doi.org/ckcj">https://doi.org/ckcj</a>

DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

### 7. Reproducibility of computational workflows is automated using continuous analysis

Brett K Beaulieu-Jones, Casey S Greene

Nature biotechnology (2017-04) <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/</a>
DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/">PMCID: 28288103 · PMCID: 282888103 · PMCID: 28288103 · PMCID: 282888103 · PMCID: 28288103 · PMCID: 282888103 · PMCID: 28288103 · PMCID: 2828810

#### 8. Bitcoin for the biological literature.

Douglas Heaven

Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888

DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

# 9. Plan S: Accelerating the transition to full and immediate Open Access to scientific publications

cOAlition S

(2018-09-04) https://www.wikidata.org/wiki/Q56458321

#### 10. **Open access**

Peter Suber *MIT Press* (2012)

ISBN: 9780262517638

#### 11. Open collaborative writing with Manubot

Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter

Manubot (2020-05-25) https://greenelab.github.io/meta-review/

#### 12. Opportunities and obstacles for deep learning in biology and medicine

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, ... Casey S Greene

Journal of The Royal Society Interface (2018-04) <a href="https://doi.org/gddkhn">https://doi.org/gddkhn</a> DOI: <a href="https://doi.org/gddkhn">10.1098/rsif.2017.0387</a> • PMID: <a href="pubmed">29618526</a> • PMCID: <a href="pubmed">PMC5938574</a>

#### 13. Open collaborative writing with Manubot

Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) https://doi.org/c7np

DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

This manuscript is a template (aka "rootstock") for <u>Manubot</u>, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (.md files in the /content directory) to the output you see below.

### **Basic formatting**

**Bold text** 

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined italics and bold

#### Strikethrough

- 1. Ordered list item
- 2. Ordered list item
  - a. Sub-item
  - b. Sub-item
    - i. Sub-sub-item
- 3. Ordered list item
  - a. Sub-item
- List item
- List item
- List item

subscript: H<sub>2</sub>O is a liquid

superscript: 2<sup>10</sup> is 1024.

unicode superscripts<sup>0123456789</sup>

unicode subscripts<sub>0123456789</sub>

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to <u>editing</u> and <u>version</u> control.

Line break without starting a new paragraph by putting two spaces at end of line.

### **Document organization**

Document section headings:

# **Heading 1**

### **Heading 2**

**Heading 3** 

**Heading 4** 

**Heading 5** 

**Heading 6** 



#### Horizontal rule:

Heading 1's are recommended to be reserved for the title of the manuscript.

Heading 2's are recommended for broad sections such as Abstract, Methods, Conclusion, etc.

Heading 3's and Heading 4's are recommended for sub-sections.

### Links

Bare URL link: <a href="https://manubot.org">https://manubot.org</a>

Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah

Link with text

Link with hover text

Link by reference

#### **Citations**

Citation by DOI [6].

Citation by PubMed Central ID [7].

Citation by PubMed ID [8].

Citation by Wikidata ID [9].

Citation by ISBN [10].

Citation by URL [11].

Citation by alias [12].

Multiple citations can be put inside the same set of brackets [6,10,12]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [7,8,12,13].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

## Referencing figures, tables, equations

Figure 1

Figure 2

```
Figure 3

Figure 4

Table 1

Equation 1

Equation 2
```

### **Quotes and code**

Quoted text

Quoted block of text

Two roads diverged in a wood, and I—I took the one less traveled by, And that has made all the difference.

Code in the middle of normal text, aka inline code.

Code block with Python syntax highlighting:

```
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

### **Figures**



**Figure 1:** A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.



**Figure 2:** An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.



Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.



Figure 4: A vector .svg image loaded from GitHub. The parameter sanitize=true is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image. Note that if you want to export to Word (.docx), you need to download the image and reference it locally (e.g. content/images/vector.svg) instead of using a URL.

### **Tables**

**Table 1:** A table with a top caption and specified relative column widths.

Bowling Scores	Jane	John	Alice	Bob
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

**Table 2:** A table too wide to fit within page.

		Digits 1-33	Digits 34-66	Digits 67-99	Ref.
k	oi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	piday.org
E	è	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	nasa.gov

Table 3: A table with merged cells using the attributes plugin.

	Colors	
Size	Text Color	Background Color
big	blue	orange
small	black	white

### **Equations**

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$$
(2)

### **Special**

▲ WARNING The following features are only supported and intended for .html and .pdf exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as .docx.

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot. Manubot Manubot. Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot attributes plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

Manubot Manubo

Available background colors for text, images, code, banners, etc:

white lightgrey grey darkgrey black lightred lightyellow lightgreen lightblue lightpurple red orange yellow green blue purple

Using the **Font Awesome** icon set:

Light Grey Banner
useful for general information - manubot.org

## **1** Blue Banner

useful for important information - manubot.org

**♦ Light Red Banner** useful for *warnings* - <u>manubot.org</u>