Response to reviewer comments, part II (The five pillars)

Mark Ziemann, Pierre Poulain, Anusuiya Bora

2023-09-22

Dear Editor,

Again, many thanks for your consideration of our manuscript. Below is our point-by-point response to Reviewer 3’s remaining comments.

### Point 2

In my view, the paper is not suitable for “anyone who analyzes biological data.” After all, the journal is Briefings in Bioinformatics. Much of the material is too technical for the stated audience. The authors have provided “Practical guides for implementing ‘Five Pillar’ principles.” This may be helpful to some, but it is not realistic to assume that people will consume all of those materials before reading the paper or that by doing so they will be able to understand all of the technical details in the manuscript. In my view, the intended audience should be clarified in the manuscript.

RESPONSE: Indeed much of the manuscript is oriented towards gold-standard practices and this could be out of the immediate reach for many biological data analysts, but this is not a futile exercise. Firstly, papers such as this one set out a road map of self-education which may take years to accomplish. Secondly, where self-education reaches its limits, they can seek workshops and other training to strengthen these priority areas.

Therefore, we have altered the phrase in the abstract and body to recognize the journey of bioinformaticians in training.

ABSTRACT:

This guide is designed for bioinformatics data analysts and bioinformaticians in training, but should be relevant to other domain of study.

RECOMMENDATIONS:

The intended audience is bioinformatics data analysts and bioinformaticians in training, however the principles described here could equally apply to other domains of study.

If the reviewer has a specific suggestion, then it would be helpful to provide this.

### Point 10

1. What about when your data files are too large to fit on a personal computer? Please address in the manuscript.

RESPONSE: In accordance with this suggestion we have added the following passage to the literate programming section.

While it is common to run literate programs on a PC with graphical interfaces such as RStudio, they can also be executed in “headless” mode on high performance computing clusters or cloud servers using command-line interfaces to take advantage of greater computational resources. Headless mode is also important for integrating literate scripts into larger workflows and for automated testing.

### Point 17

1. The Biocontainers project should be mentioned.

RESPONSE: We have added an additional mention of BioContainers which appears just after where DockerHub is mentioned.

In bioinformatics, containers are already extensively used. For example, BioContainers is a registry for sharing bioinformatics software containers [1]. Containers are equally useful to data analysts by encapsulating the environment in which analytical scripts are executed.

### Point 20

1. Figure 5: I don’t think it’s really necessary to resummarize the FAIR principles. People can go to the source article for that. I don’t mean to be snarky, but everyone has limited time. I still believe that spelling out these points only serves to make the manuscript longer than necessary and thus make it less likely for people with “limited time” to read it. FAIRness is not the main thrust of this manuscript.

RESPONSE: We appreciate the reviewer’s point of view. We have removed mention of FAIR and we refer readers to the literature on this topic. FAIR has also been removed from the pillars infographic and Figure 5 is also gone.

### R3 Point 21

1. Page 11, line 13: I disagree that repositories like GEO and SRA are FAIR. There are lots of problems with FAIRness in these repositories.

RESPONSE: Possibly the reviewer hasn’t seen the rebuttal letter had it’s own bibliography. Anyway, since “FAIR” is not described anymore, it simplifies the description of these databases, which satisfy most best practices. We cite the work of Bonaretti et al 2019. We also cite Hassenruck et al 2021, who report deficiencies in these databases, but spread the blame between database administrators, researchers and other stakeholders. The two important passages we have amended are provided below:

In the wake of the Human Genome Project, a number of repositories for specific biological data types were established including GEO, SRA, ENA and PRIDE [2–4]. Although these repositories suffer issues around interoperability and reusability, they do support data reuse through findability and accessibility [5,6].

…

Researchers should be aware, however, that these repositories are often not moderated, and that it is their responsibility to provide sufficient metadata.

## Bibliography

1. Veiga Leprevost F da, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: An open-source and community-driven framework for software standardization. Bioinformatics. 2017;33: 2580–2582.

2. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

3. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res. 2012;40: D54–6.

4. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, et al. PRIDE: The proteomics identifications database. Proteomics. 2005;5: 3537–3545.

5. Bonaretti S, Willighagen E. Two real use cases of FAIR maturity indicators in the life sciences. bioRxiv. 2019.

6. Hassenrück C, Poprick T, Helfer V, Molari M, Meyer R, Kostadinov I. FAIR enough? A perspective on the status of nucleotide sequence data and metadata on public archives. bioRxiv. 2021. p. 2021.09.23.461561.