Reproducible analysis and interactive reports using containers and R/Shiny in bioinformatics pipelines

PEC1

Luis Morís Fernández

# Title

Reproducible analysis and interactive reports using containers and R/Shiny in bioinformatics pipelines

# Keywords

reproducibility, interactive, pipeline, bioinformatics

# Context and justification of this work

Reproducibility is one of the cornerstones of the scientific method, nonetheless is one of the most commonly forgotten, sometimes shadowed by the novelty or alleged impact of the results (John P. A. Ioannidis 2005; Errington et al. 2021). With the advent of informatics more powerful and flexible analyses are available to researchers in all fields everyday (Gauthier et al. 2018). Unfortunately, this power and flexibility comes usually at the cost of complexity and length. Bioinformatics pipelines are paramount examples of the benefit of these powerful and flexible analyses and also of the cost of complexity (John P. A. Ioannidis et al. 2009; Mangul et al. 2019; Markowetz 2015; Brito et al. 2020). This issue is not solved exclusively by properly storing the scripts of analysis and the data, additional considerations must be taken, such as the pathing, the version of the language used, the version of the libraries, the operating system and its version, etc (Wratten, Wilm, and Göke 2021; Wallach, Boyack, and Ioannidis 2018). All these problem hinders reproducibility and utmost care must be taken when preparing and storing these pipelines, so they can be run not only at the moment of the initial analysis but at any moment in the future when another researcher may need to verify the analyses, extend them or run them on new data.

This work will explore two possible ways of improving reproducibility when developing bioinformatic pipelines. The first one is making the analysis steps traceable from the start to end. This regards the actual scripts that will be executed during the analysis. To address this issue I will take advantage of workflow creation tools that allows creating long reproducible analysis pipelines. The second is creating a reproducible a light-weight stable environment were the previous pipeline will through the use of containers.

The other objective of this work has to do with the interactivity of the the reports obtained during the analysis process. It is rarely the case, particularly in clinical environments, that the same person has the ability to program and run analyses (e.g.: a bioinfomatics expert) and also to make decisions based on those data (e.g.: a medical doctor) (Ludt et al. 2022). This usually creates a communication overhead between the data expert and the decision-maker until the exact data needed to make the decision is found. This work will also create an interactive report that will try to ameliorate this problem by providing the decision-maker with a small set of tools that would decrease this communication overhead.

# General description

The intention of this work is to create a proof of concept that would address two problems commonly found when developing bioinformatics pipelines. The first one is how to improve its reproducibility through the use of workflow tools and containers. And the second one is to explore interactive reports as a mean to reduce the overhead between bioninformatics and medical personnel improving the decision making loop.

# Objectives

## General Objectives

Create a proof of concept that includes a traceable and reproducible pipeline through the use of workflow tools. This pipeline must run inside a container and produce an interactive report as a result.

## Specific objectives

1. Describe the reproducibility problem in bioinformatics
2. Explore containers and workflow tools as a mean to improve the reproducibility of bioinformatics pipelines
3. Explore interactive reports as a mean to improve the decision making loop in clinical settings
4. Create a microarray analysis pipeline using containers that produce an interactive report as a result in a real setting

# Approach and method

Firstly, this work will address the state of the art in reproducibility with particular emphasis in the bioinformatics field. Secondly, this work will explore the creation of a proof of concept microarrays analysis pipeline created in a workflow tool that will run inside a container that will produce, among other products, an interactive report. In this second part, I will count with the support of the *Plataforma de Bioinformática de la Unidad de Estadística y Bioinformática (UEB)* for having access to data and creating a useful pipeline that could be used in a real clinical setting.

From a technical point of view, albeit other tools may be explored during this work:

* The R (R Core Team 2022) language will be used as the foundation for all the software produced in it.
* The maUEB package functions, currently under development by the *UEB*, will be used to implement the microarray analysis pipeline.
* The targets (Landau 2021) package will be used as the workflow tool to create the pipeline itself in combination with the maUEB functions.
* R/Shiny (Chang et al. 2022) framework will be used to create an interactive report.
* Docker (Merkel 2014) container technology will be used for containerizing the pipeline.

While selecting the tools for this work the following considerations were kept:

* Minimize friction between components when developing the proof of concept. That is the reason why all tools are R-based, with the exception of docker.
* Use components easily installed and run inside a container.
* Favor analysis scripts over graphical user interfaces.
* Favor simplicity and closeness to the R language (Ousterhout 2018).

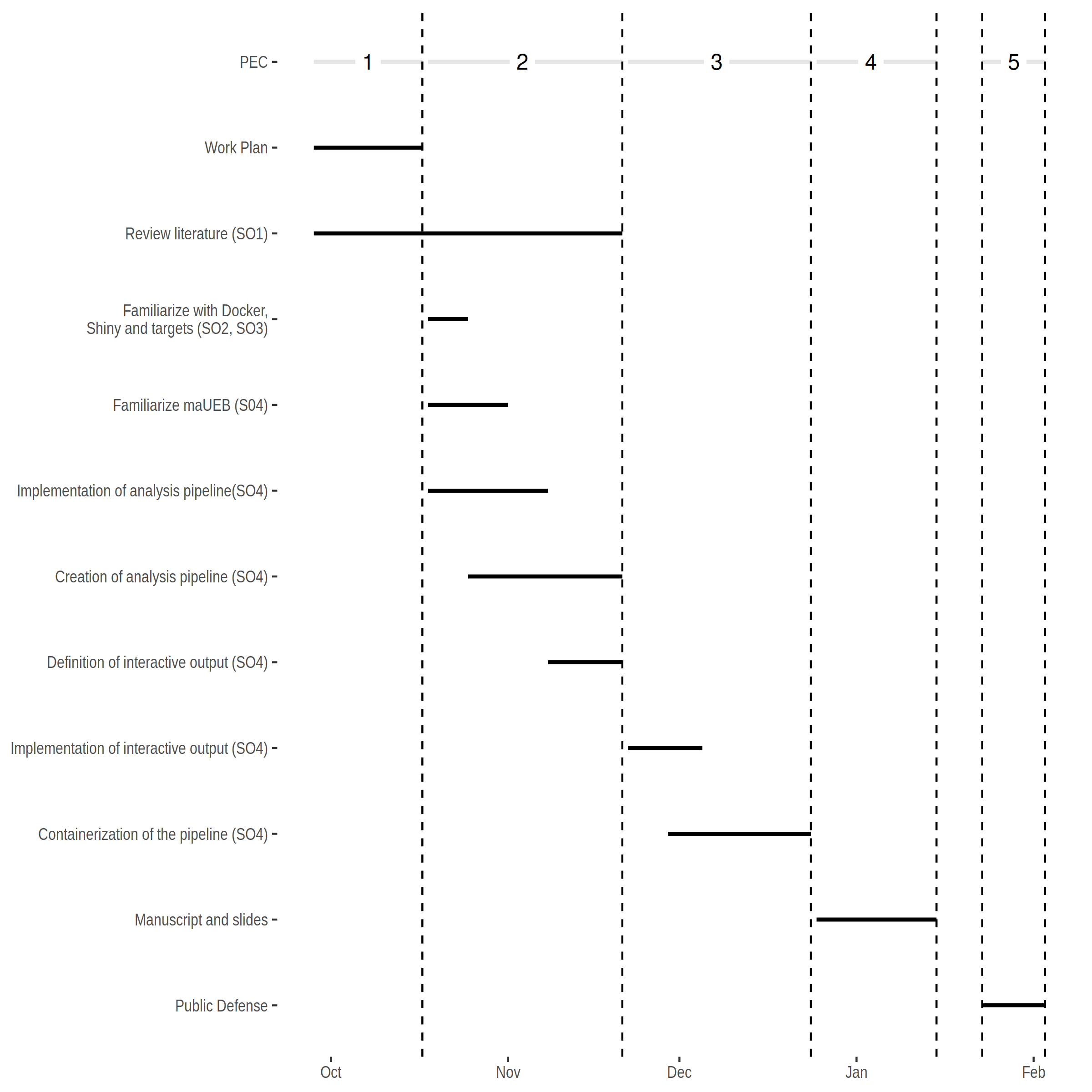
# Planning

## Tasks

Each task will be related to a specific objective using the acronym *SO*

1. Review literature for:
   1. **(SO1)** General problems of reproducibility in science and in bioinformatics
   2. **(SO1)** Current approaches to solve this problem
2. Software and technical knowledge and familiarization with:
   1. **(SO2, SO3)** Docker, R/Shiny and targets R package
   2. **(SO4)** maUEB package
3. **(SO2, SO4)** Definition and implementation of the analysis pipeline
4. **(SO3, SO4)** Definition and implementation of the interactive report
5. **(SO2, SO4)** Containerization of the pipeline

## Schedule



## Milestones

1. Workplan **(17/07/2022)**
2. Literature review and state of the art
3. Creating a pipeline using targets
4. Defining the interactive output
5. PEC2 **(21/11/2022)**
6. Implementing interactive output
7. Containerize the pipeline
8. PEC3 **(24/12/2022)**
9. Manuscript and slides **(27/12/2022 - 15/01/2023)**
10. Public defense **(23/01/2023 - 03/02/2023)**

## Risk analysis

Here I present possible risks and proposed solutions.

1. Oversized scope. This work tries to address two well-diferentiated problems, the reproducibility and the report interactivity. The interactivity scope can be reduced or removed from the work if the amount of time is not enough to cover both. The solution proposed is to reevaluate the scope when PEC2 is finalized and the scope is better defined.
2. Package availability. The initial intention is to use a recently developed package maUEB. If the package is not available other alternative packages covered during the master classes will be used.
3. Data availability. If data is unavailable, for example confidentiality for confidentiality reasons, open data sources will be used.
4. Pipeline definition. If the the pipeline is not possible to define in a real clinical setting, steps similar to those used in the other pipelines during the master classes will be used.

# Expected results

1. Work plan

* A detailed workplan for the thesis work.

1. Thesis document

* A document containing an state of the art, methodology, results and conclusions obtained during the master’s thesis.

1. A working proof of concept

* A containerized pipeline, delivered as an R package.

1. Slides and virtual presentation

* Slides and a video presentation for the defence.

# CCEG

Although the work has a strong technical nature, its nature can be related to at least one *Sustainable Development Goals* (SDG) directly and another one which is transversal to many research works.

In first place we are directly related to *SDG 9 Build resilient infrastructure, promote sustainable industrialization and foster innovation* in particular to point *9.5 Enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending*. One of the core aspects of this work is to improve the reproducibility of research in the field of Bioinformatics. This reproducibility is valuable in itself as it improves the quality of the evidence produced, but its impact goes beyond that. Low quality evidence may impact the allocation of research resources, as researchers may pursue lines that may have been discarded in the light of more solid and reproducible evidence. In fact research indicates that low quality evidence, although not restricted to reproducibility, wastes billions of dollars each in year in the biomedical field (Macleod et al. 2014).

In second place almost all research works can impact into *SDG 5: Achieve gender equality and empower all women and girls*. In particular, a common issue is that cites are not equally distributed between genders and men tends to cite other men more often (Chatterjee and Werner 2021). Therefore in this work particular caution will be placed on considering all works that may be referenced regardless of the gender of the author/s and citing using full names. Additionally, inclusive language will be used during the writing and the third person plural will be used to avoid making assumptions about gender, for example of possible users of the work produced (“Singular "They"” 2022). Although initially none is detected, this work will also consider possible biases or issues with this product beyond gender, regarding race, social position or level of technology access among others. One key consideration will be the open source nature of the work and the licensing of the software used to maximize its availability in institutions in developing countries (also related to *SDG 4 Quality education*).

# Bibliography

Brito, Jaqueline J, Jun Li, Jason H Moore, Casey S Greene, Nicole A Nogoy, Lana X Garmire, and Serghei Mangul. 2020. “Recommendations to Enhance Rigor and Reproducibility in Biomedical Research.” *GigaScience* 9 (6). <https://doi.org/10.1093/gigascience/giaa056>.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2022. *Shiny: Web Application Framework for r*. <https://shiny.rstudio.com/>.

Chatterjee, Paula, and Rachel M. Werner. 2021. “Gender Disparity in Citations in High-Impact Journal Articles.” *JAMA Network Open* 4 (7): e2114509–9. <https://doi.org/10.1001/jamanetworkopen.2021.14509>.

Errington, Timothy M, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021. “Challenges for Assessing Replicability in Preclinical Cancer Biology.” *eLife* 10 (December). <https://doi.org/10.7554/elife.67995>.

Gauthier, Jeff, Antony T Vincent, Steve J Charette, and Nicolas Derome. 2018. “A Brief History of Bioinformatics.” *Briefings in Bioinformatics* 20 (6): 1981–96. <https://doi.org/10.1093/bib/bby063>.

Ioannidis, John P A, David B Allison, Catherine A Ball, Issa Coulibaly, Xiangqin Cui, Aedín C Culhane, Mario Falchi, et al. 2009. “Repeatability of Published Microarray Gene Expression Analyses.” *Nature Genetics* 41 (2): 149–55. <https://doi.org/10.1038/ng.295>.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.

Landau, William Michael. 2021. “The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing.” *Journal of Open Source Software* 6 (57): 2959. <https://doi.org/10.21105/joss.02959>.

Ludt, Annekathrin, Arsenij Ustjanzew, Harald Binder, Konstantin Strauch, and Federico Marini. 2022. “Interactive and Reproducible Workflows for Exploring and Modeling RNA‐seq Data with pcaExplorer, Ideal, and GeneTonic.” *Current Protocols* 2 (4). <https://doi.org/10.1002/cpz1.411>.

Macleod, Malcolm R., Susan Michie, Ian Roberts, Ulrich Dirnagl, Iain Chalmers, John P. A. Ioannidis, Rustam Al-Shahi Salman, An-Wen Chan, and Paul Glasziou. 2014. “Biomedical Research: Increasing Value, Reducing Waste.” *The Lancet* 383 (9912): 101–4. <https://doi.org/10.1016/S0140-6736(13)62329-6>.

Mangul, Serghei, Thiago Mosqueiro, Richard J. Abdill, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, et al. 2019. “Challenges and Recommendations to Improve the Installability and Archival Stability of Omics Computational Tools.” *PLOS Biology* 17 (6): e3000333. <https://doi.org/10.1371/journal.pbio.3000333>.

Markowetz, Florian. 2015. “Five Selfish Reasons to Work Reproducibly.” *Genome Biology* 16 (1). <https://doi.org/10.1186/s13059-015-0850-7>.

Merkel, Dirk. 2014. “Docker: Lightweight Linux Containers for Consistent Development and Deployment.” *Linux Journal* 2014 (239): 2.

Ousterhout, John. 2018. *A Philosophy of Software Design*. 1st ed.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

“Singular "They".” 2022. *American Psychological Association*. American Psychological Association. <https://apastyle.apa.org/style-grammar-guidelines/grammar/singular-they>.

Wallach, Joshua D., Kevin W. Boyack, and John P. A. Ioannidis. 2018. “Reproducible Research Practices, Transparency, and Open Access Data in the Biomedical Literature, 2015–2017.” Edited by Ulrich Dirnagl. *PLOS Biology* 16 (11): e2006930. <https://doi.org/10.1371/journal.pbio.2006930>.

Wratten, Laura, Andreas Wilm, and Jonathan Göke. 2021. “Reproducible, Scalable, and Shareable Analysis Pipelines with Bioinformatics Workflow Managers.” *Nature Methods* 18 (10): 1161–68. <https://doi.org/10.1038/s41592-021-01254-9>.