Reproducible analysis pipelines using containers and data exploration using R/Shiny

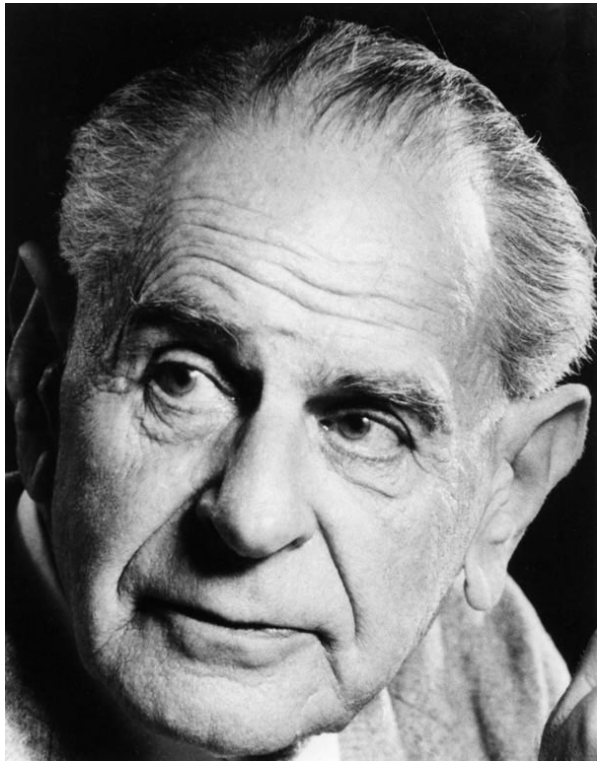
Máster en bioinformática y bioestadística

Luis Morís Fernández

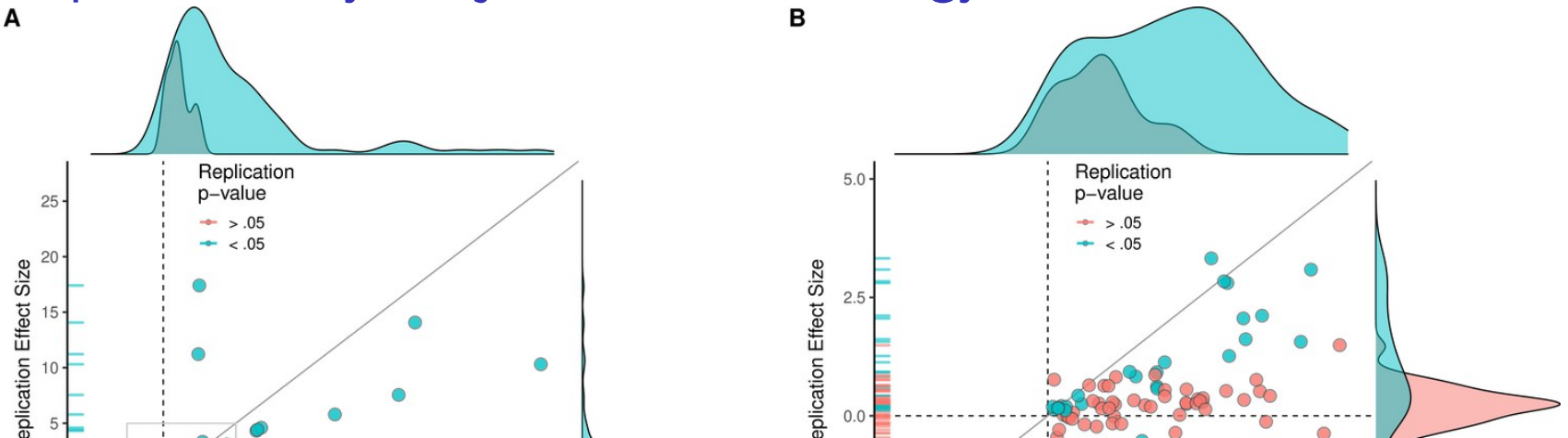
The reproducibility problem

"Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. [...] Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence'[...]"

— Karl R. Popper. The Logic of Scientific Discovery (1959)



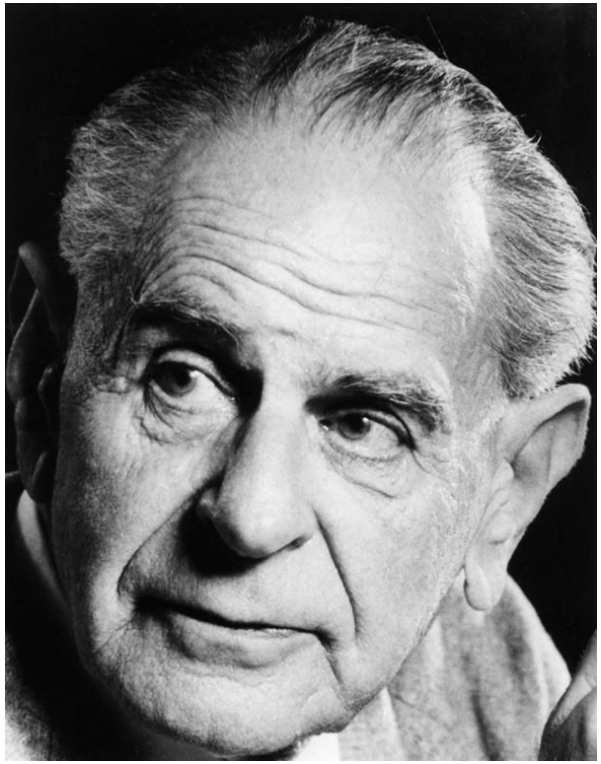
Reproducibility Project: Cancer Biology



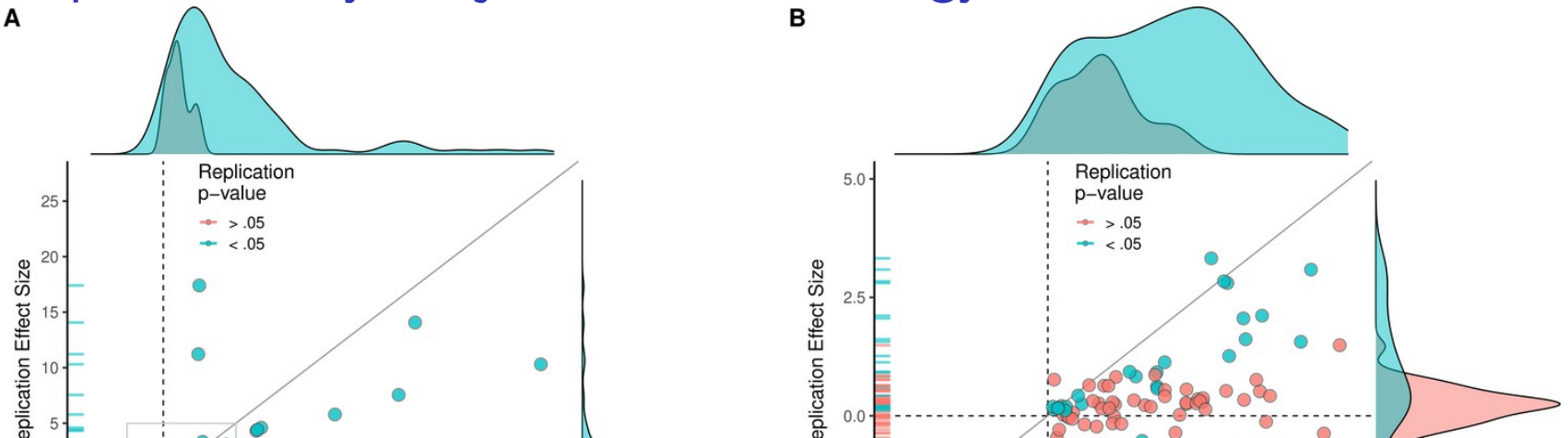
The reproducibility problem

"Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. [...] Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence'[...]"

— Karl R. Popper. The Logic of Scientific Discovery (1959)



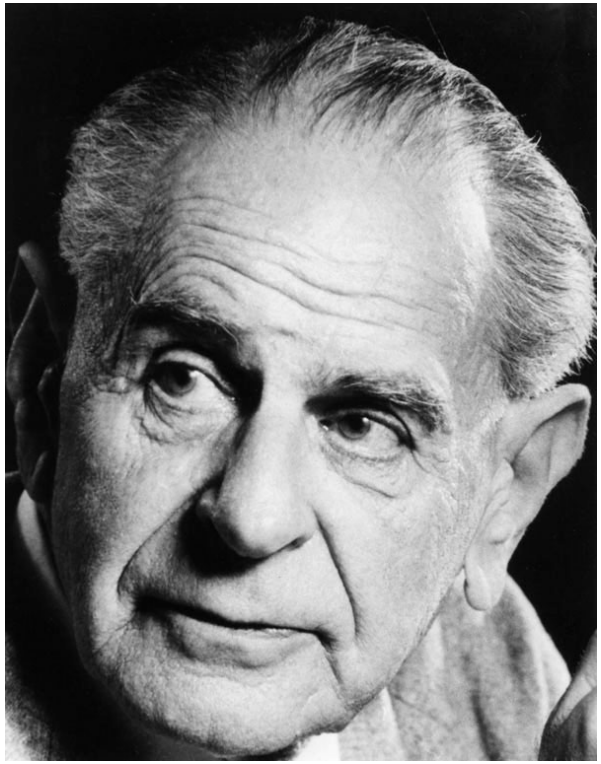
Reproducibility Project: Cancer Biology



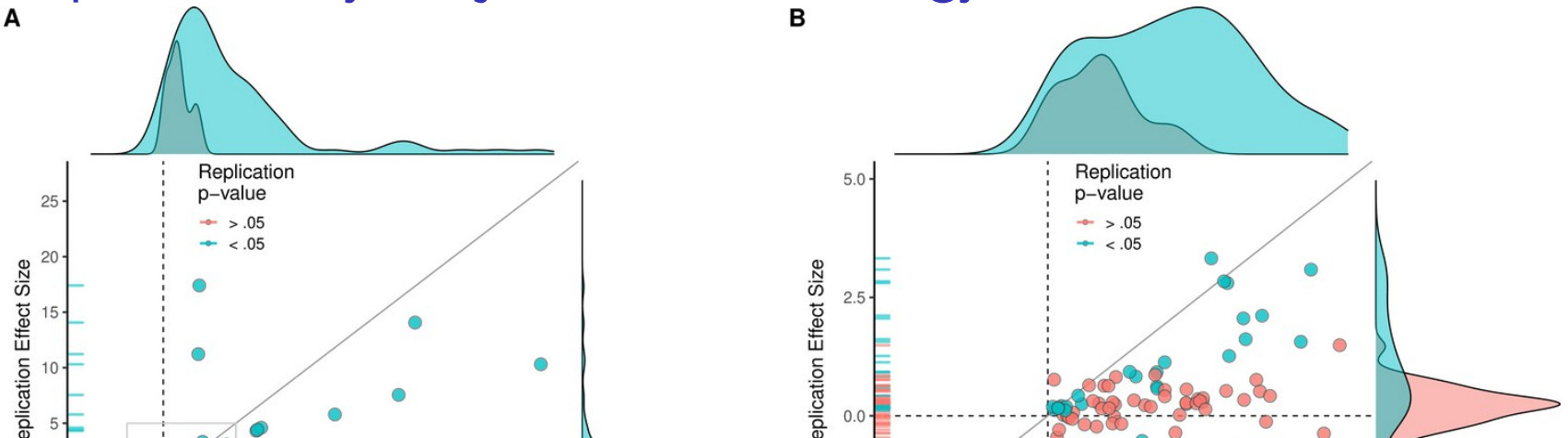
The reproducibility problem

"Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. [...] Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence'[...]"

— Karl R. Popper. The Logic of Scientific Discovery (1959)



Reproducibility Project: Cancer Biology



Results

Targets microarray analysis pipeline: Steps

1. Data Loading
2. Quality Control
3. Differential Expression Analysis
4. Gene Set enrichment Analysis

Targets microarray analysis pipeline: Advantages

- Step behavior is defined by a list of parameters
- Each step has an specific list of parameters
- All parameters are packed in a single section of the pipeline
- User can focus exclusively on the parameter lists instead of modyfing the pipeline itself

Targets microarray analysis pipeline: Advantages

- Code is based on small multiples

```
tar_target(
  name = qc_raw_boxplot_file,
command = do.call(
```

Discussion

A targets containerized microarray pipeline

- Users can concentrate on a smaller portion of the script for their changes
- Target declaration complexity was minimized by using small multiples
- Easily reproducible and automatic tracking of dependencies

Containerizing in a Docker

- Easily generalized to other pipelines
- Easy to archive and use in the future
- Helpful solving the reproducibility problem

An interactive application for data exploration using R/Shiny

- Simple but effective in reducing the data-analyst vs data-decision-makers loops
- Reusable can composable

Conclusions

List of achieved 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 tools as a mean to improve the decision making loop in clinical settings
4. Create a microarray analysis pipeline using containers that produce a report

Future lines of work

A targets containerized microarray pipeline

- Modifications to the maUEB package could be done to allow implementing an interactive report
- A more extensive manual on the pipeline could be written
- Additional versions of this pipeline could be written for other analyses
- Another strategies for creating the parameters list could be explored

Containerizing in a Docker

- Data analysis could be included in the Docker
- Docker image could be uploaded to a Docker image repository for direct use by researchers

An interactive application for data exploration using R/Shiny

- Run refinement sessions with users to prioritize new functionalities, for example:
  - Download only genes that are significant in several comparisons
  - Provide more links to external databases with information

The end

Thanks for your time and attention