CEA JupyterHub platform for multi-omics data analysis

Solène MAUGER[1], Florian JEANNERET, Pauline BAZELLE, Christophe BATTAIL and KATY consortium[2]

[1] Laboratoire Biologie et Biotechnologies pour la Santé, IRIG, UMR 1292 INSERM-CEA-UGA,
Univ Grenoble Alpes, 38 000 Grenoble, France

[2] https://katy-project.eu, European Union's Horizon 2020 research and innovation program,
Grant agreement No 101017453

contact: christophe.battail@cea.fr

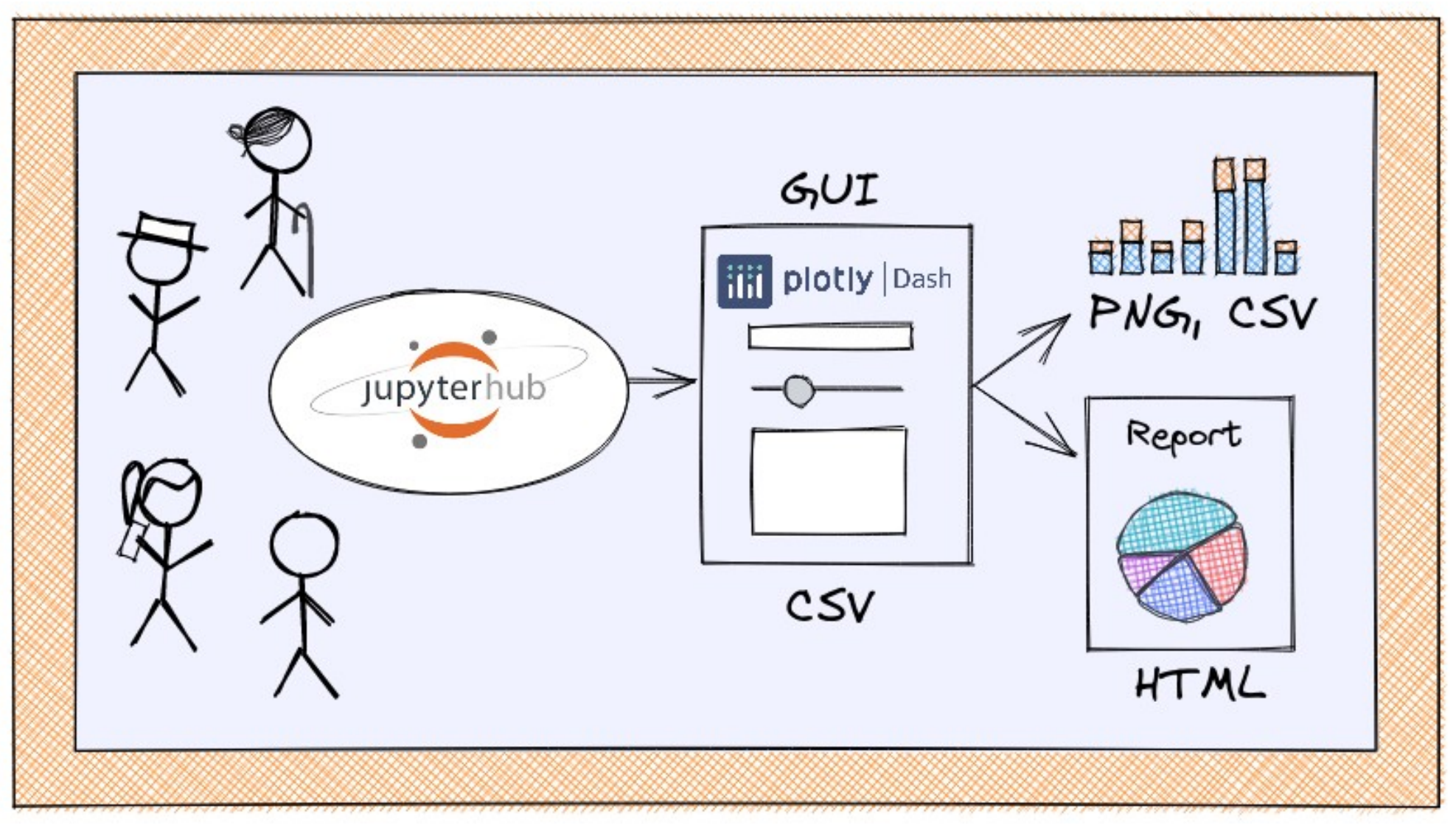
Abstract

CEA Grenoble has recently set up a **JupyterHub** infrastructure capable of putting the power of **Jupyter Notebooks** [1] in the hands of researchers and students. This shared work environment allows Data Scientists and Biologists to perform their exploration on shared computing resources. This paradigm shift allows the implementation of good practices associated with **Open Science**, **reproducibility** of analyses and the FAIR principles [2].

The Notebooks were developed by bioinformaticians, connected and deposited on the CEA JupyterHub platform to allow easy **access** to **various research members**, such as students, researchers, technicians and engineers.

Methods

A workflow was developed by connecting several **Jupyter Notebooks**. This choice was made because it is a powerful system for reproducibility, and its interface can be easily used by a user to run an analysis **without having to edit code**.



In order to automate the analyses and make them accessible to someone with no programming knowledge, a **GUI Dash-driven** was implemented. From the data generated and saved through the graphical interface, one can then obtain figures in **PNG**, tables in **CSV** format and an interactive report in **HTML** format.

Workflow of notebooks

1) In order to make the best use of Notebooks, here are the instructions for use : **Transcriptomics data** is input to perform differential gene expression analysis.

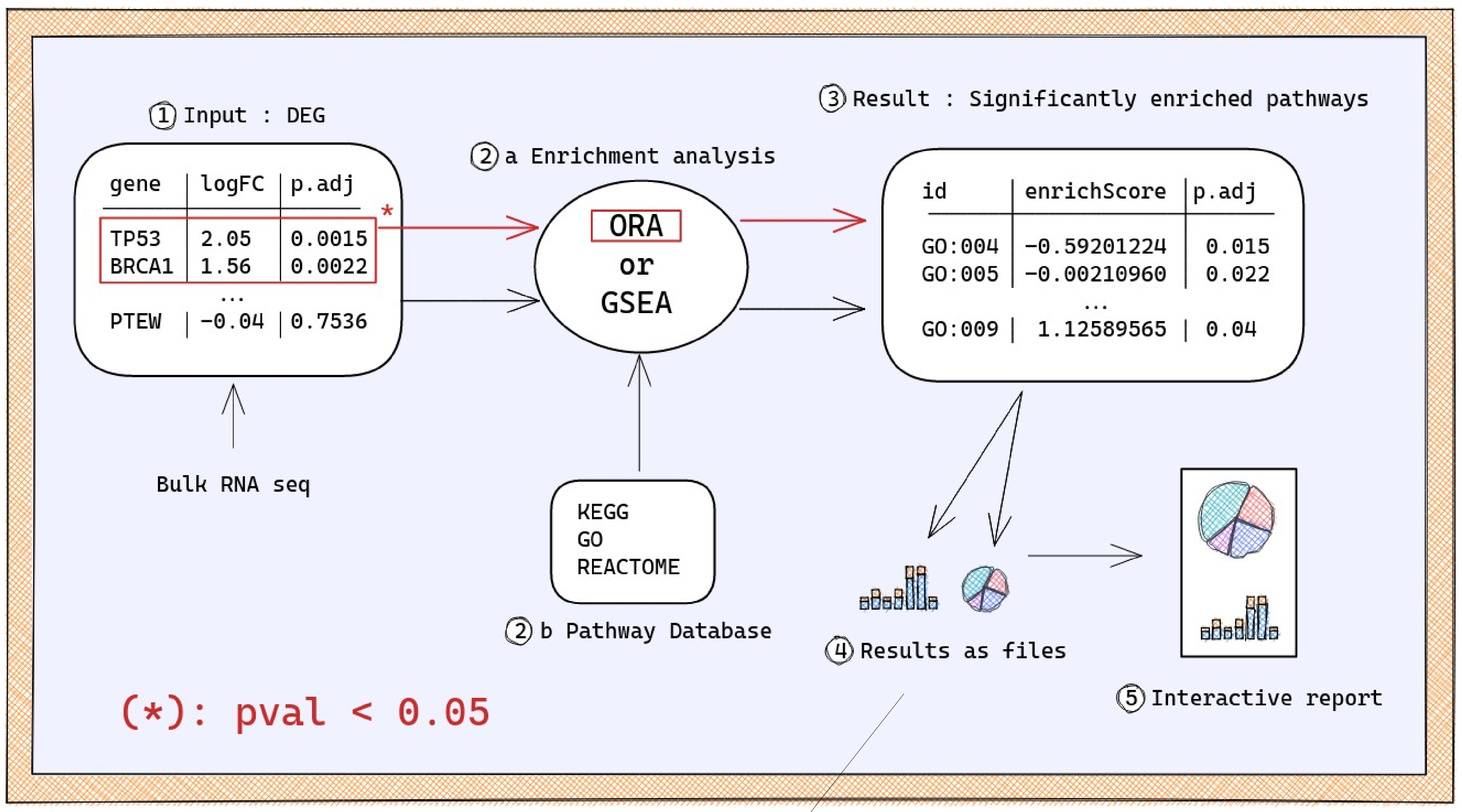
2 a) Next, a pathway enrichment analysis method must be chosen; either an **Over Representation Analysis** (ORA) or a **Gene Set Enrichment Analysis** (GSEA). If you want to perform an ORA, you must nevertheless keep only the significant genes. (e.g. p_value < 0.05)

2 b) Before starting the pathway enrichment analysis you also have to select a pathway database; either **KEGG**, **GO** or **REACTOME**.

3) The result is a **CSV** file with the list of enriched pathways associated with their enrichment scores.

4) A collection of **plots** is then generated and saved as **files**.

5) The final output of this workflow is an **interactive and documented report** to facilitate the understanding of the results by a non-bioinformatician.



① Input : DEG

gene	logFC	p.adj
TP53	2.05	0.0015
BRCA1	1.56	0.0022
...		
PTEW	-0.04	0.7536

Bulk RNA seq

② a Enrichment analysis

ORA or GSEA

② b Pathway Database

KEGG
GO
REACTOME

③ Result : Significantly enriched pathways

id	enrichScore	p.adj
GO:004	-0.59201224	0.015
GO:005	-0.00210960	0.022
...		
GO:009	1.12589565	0.04

④ Results as files

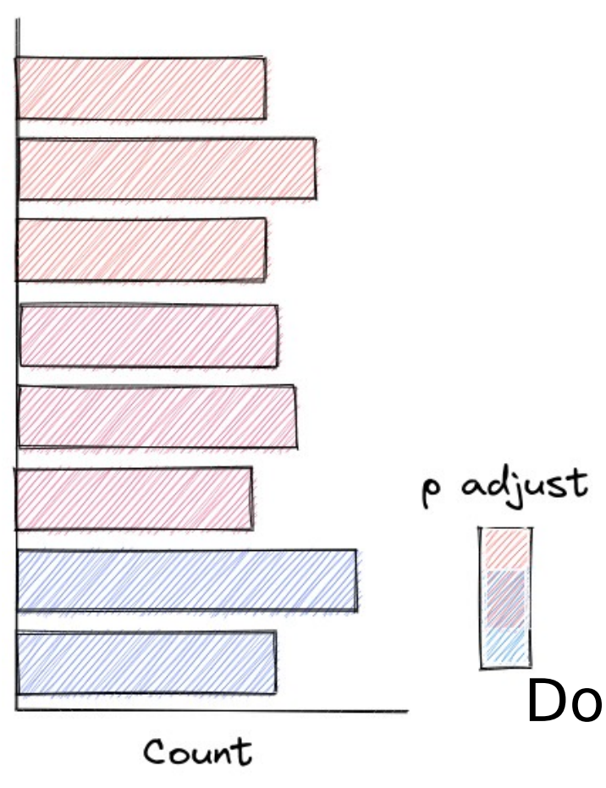
⑤ Interactive report

(*): pval < 0.05

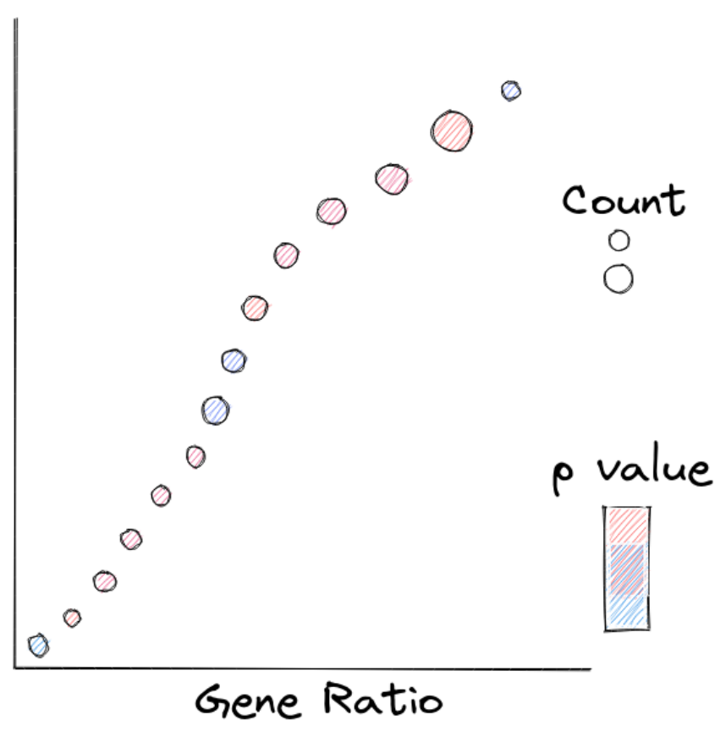
Plot types

Depending on whether we have an ORA or a GSEA we have chosen to obtain different types of plots using the ClusterProfiler R package [3]:

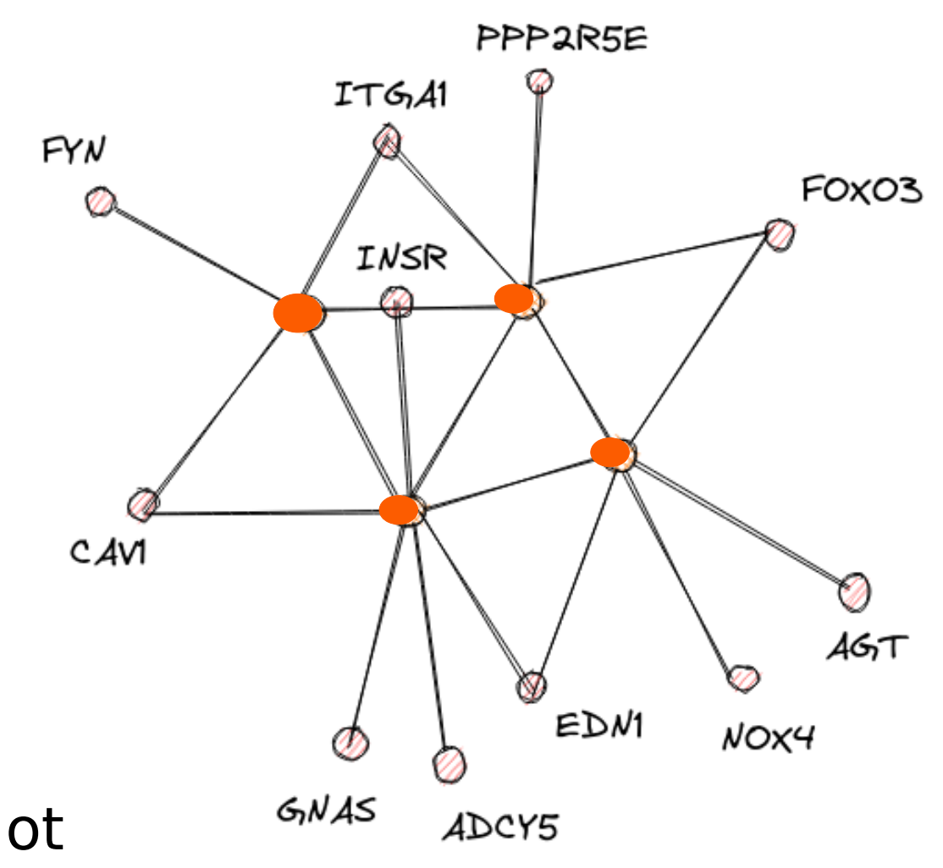
GSEA & ORA



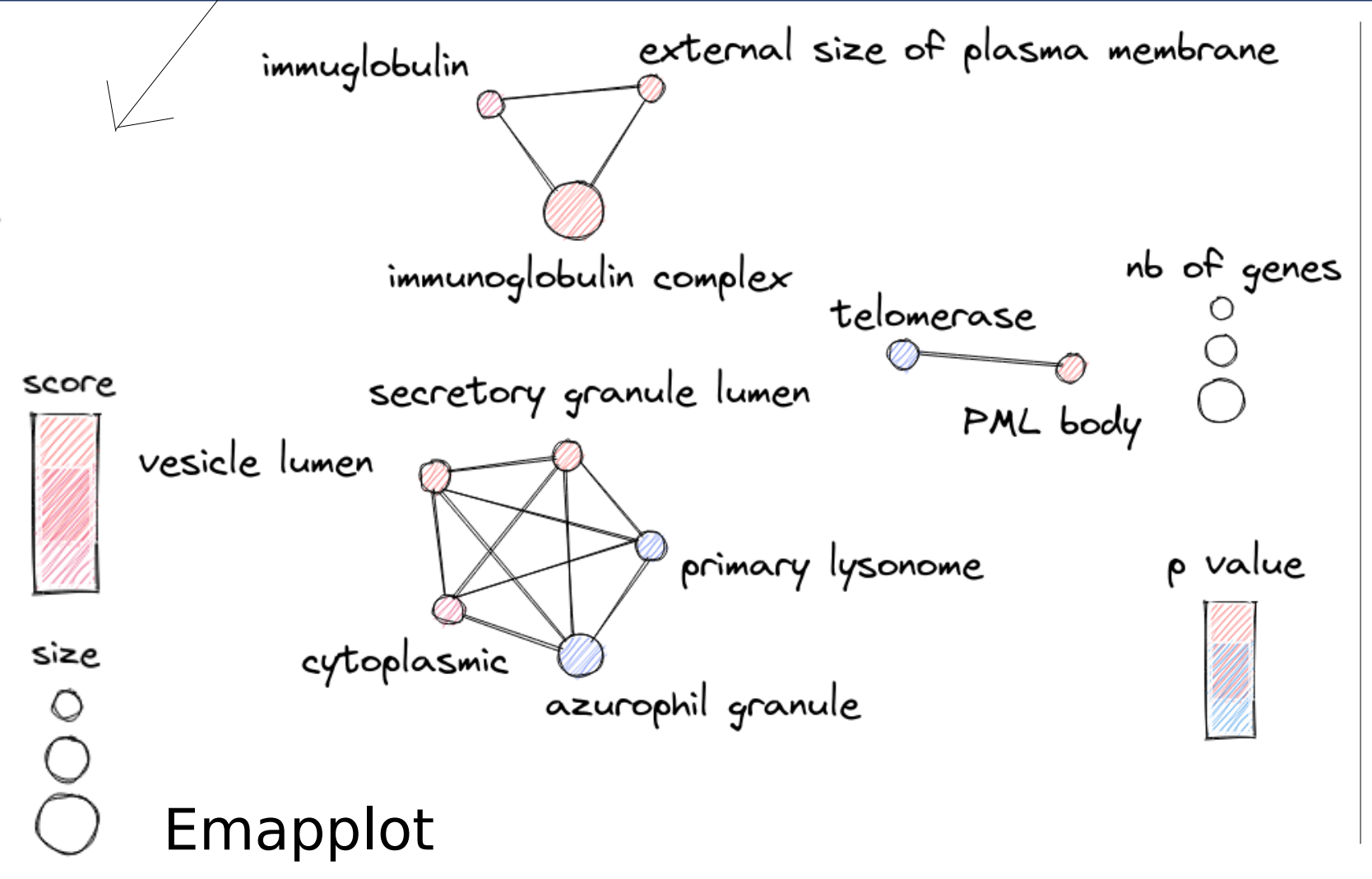
Barplot



Dotplot

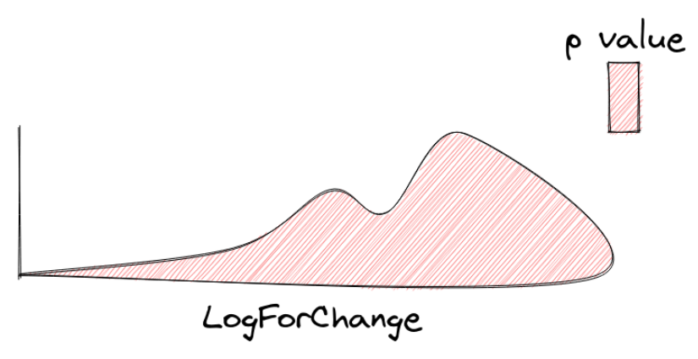


Cnetplot



Emapplot

Only GSEA



Ridgeplot

Conclusion

This **CEA JupyterHub environment** is currently made of Notebooks allowing to carry out differential gene expression analyses, ontological enrichment studies, using different statistical methods and several reference databases. To facilitate access to this data analysis environment by non-bioinformaticians, the user starts its analysis by a Notebook with a graphical user interface, implemented with the **Python Dash** software library, allowing to enter input data and to select the analyses to perform associated with the appropriate parameters, without having to modify the Notebook code.

It is a work in progress that will be **available** on **GitLab** when the project is mature enough. The next step will be to add new Jupyter Notebooks into our **Transcriptomics workflow**, such as the identification of transcription factors involved in DEG regulation, and to develop other workflows to analyse **proteomics and phosphoproteomics data**.

References

[1] Kluyver T et al. and Jupyter development team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press. pp. 87-90. 2016. doi: 10.3233/978-1-61499-649-1-87.

[2] Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18.

[3] Guangchuang Yu et al, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data, the Innovation, 2021 doi : 10.1016/j.xinn.2021.100141

The KATY project has received funding from the European Union's Horizon 2020 innovation program under grant agreement No 101017453.