Bioinformatics workflows: how to provide reproducible and scalable analyses

Marco Beccuti[1], Francesca Cordero[1], Raffaele Calogero[2]

Università degli Studi di Torino

[1]*Dip. di Informatica*

[2]*Dip. di Biotecnologie Molecolari e Scienze per la Salute*

marco.beccuti@unito.it

Ivrea, Italy - October 2020

Reproducible Bioinformatics Project

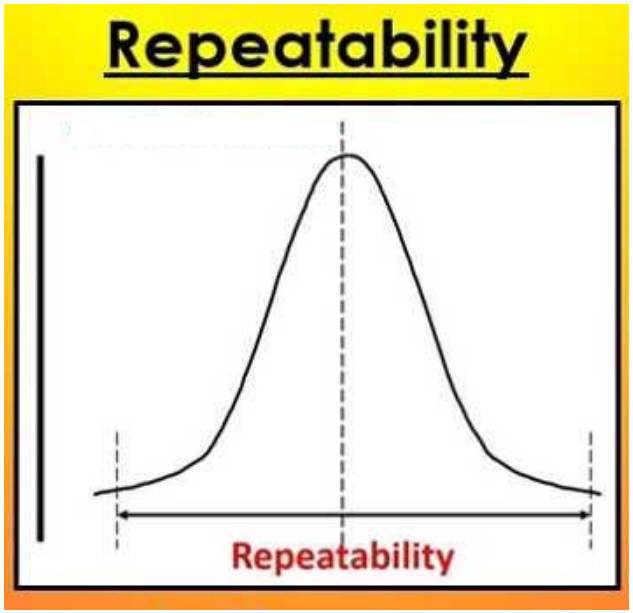
A project to provide reproducible results in Bioinformatics using Docker images

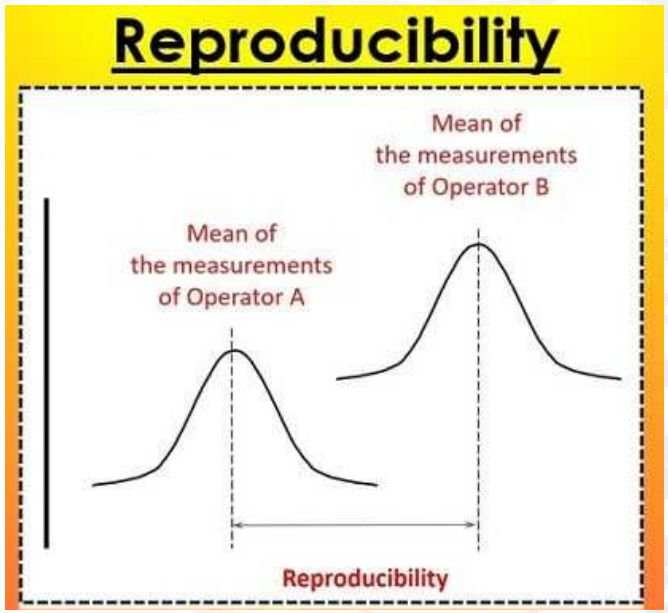
+

Computational Reproducibility in the Life Sciences

Repeatability VS Reproducibility

- ***Repeatability*** measures the variation in measurements taken by a single instrument or person under the same conditions $\Rightarrow$ *Statistical methods*;
- ***Reproducibility*** measures whether an entire study or experiment can be reproduced in its entirety $\Rightarrow$ *???*.





Most scientists 'can't replicate studies by their peers'

BBC NEWS

By Tom Feilden
Science correspondent, Today programme

22 February 2017

Share

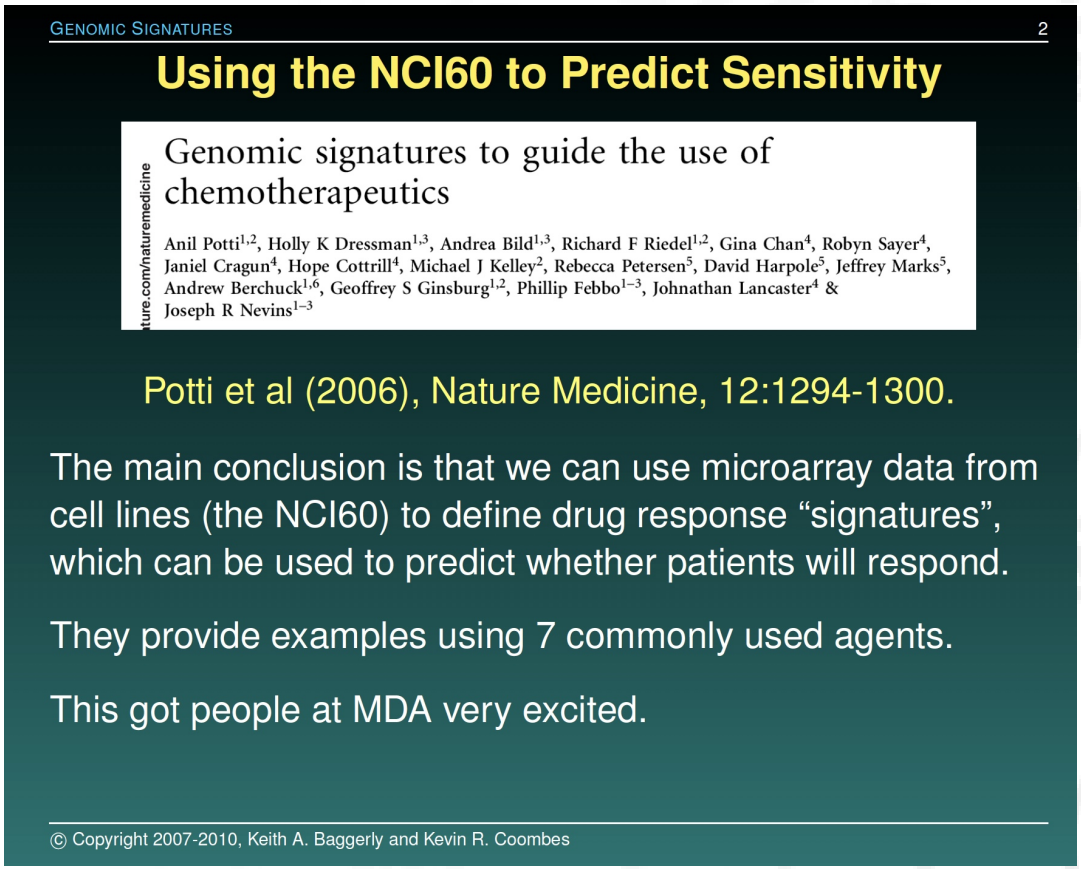
Scientists attempting to repeat findings reported in five landmark cancer studies confirmed only two

Reproducibility crisis in Life Science

- *reproducibility crisis* is a term used to point out that a large percentage of academic literature is not reproducible;
- Starting point was in 2011:
  - the attempt of Gleen Begley, head of hematology and oncology research at Amgen, to reproduce 53 foundational papers in oncology $\Rightarrow$ ***they were able to reproduce only 6 of them;***
  - the attempt of Bayer company to reproduce randomly selected bio-medical papers $\Rightarrow$ ***only the*** 65% ***of the experiments in the sample were reproduced.***
- Reproducibility has also found to be an issue in different scientific fields: computer science, psychology, economics, ...

Reproducibility crisis in Omics Research

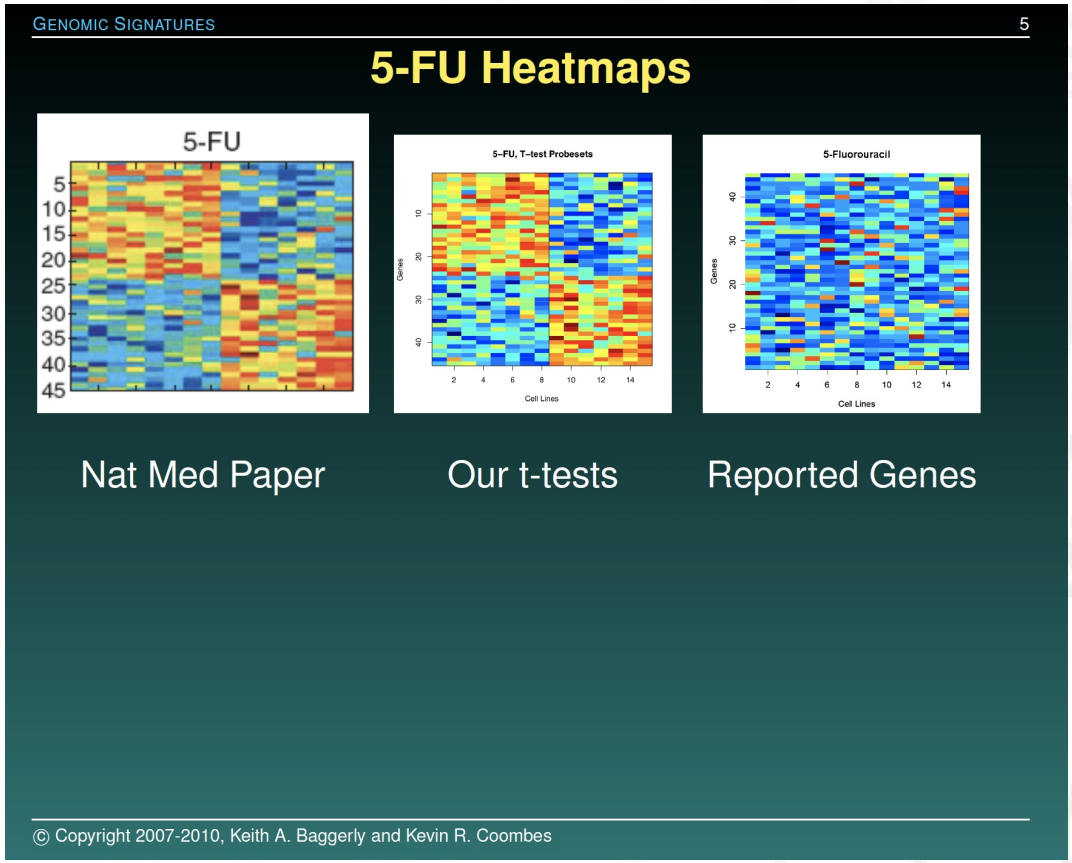
- Keith A. Baggerly, researcher at Anderson Cancer Center: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."



https://youtu.be/7gYIs7uYbMo

Reproducibility crisis in Omics Research

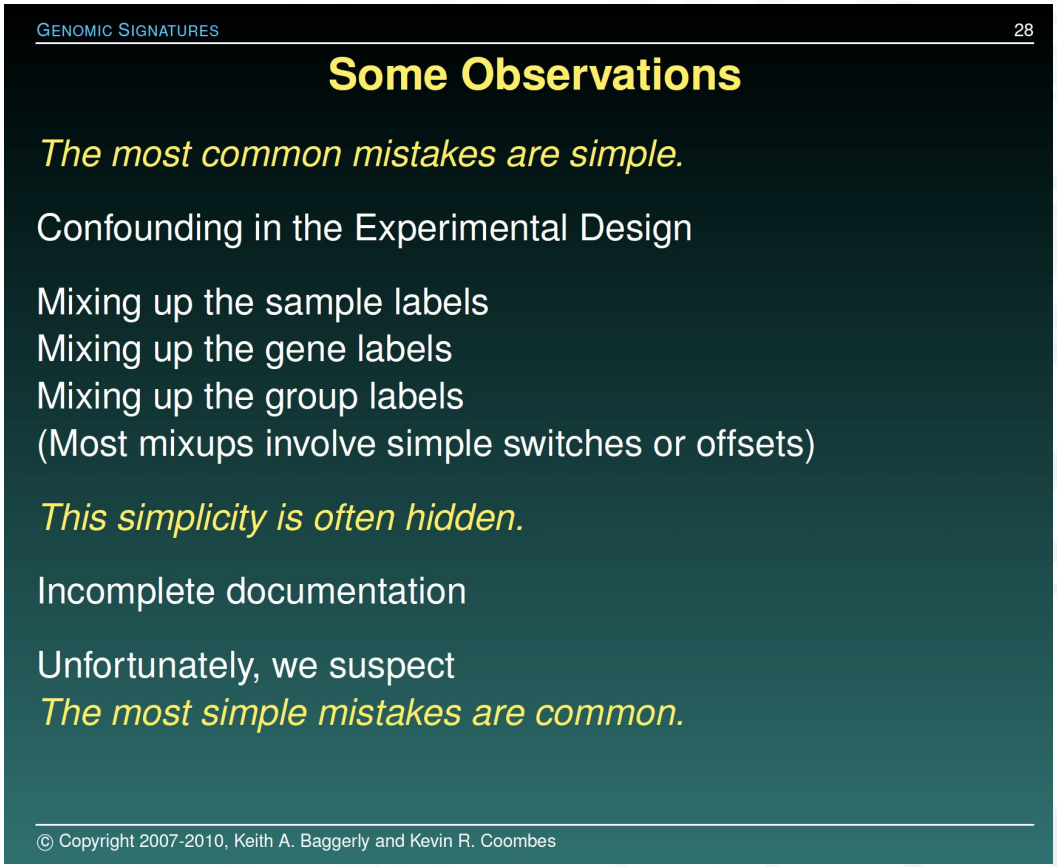
- Keith A. Baggerly, researcher at Anderson Cancer Center: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."



https://youtu.be/7gYIs7uYbMo

Reproducibility crisis in Omics Research

- Keith A. Baggerly, researcher at Anderson Cancer Center: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."



https://youtu.be/7gYIs7uYbMo

Ten Simple Rules for Reproducible Computational Research

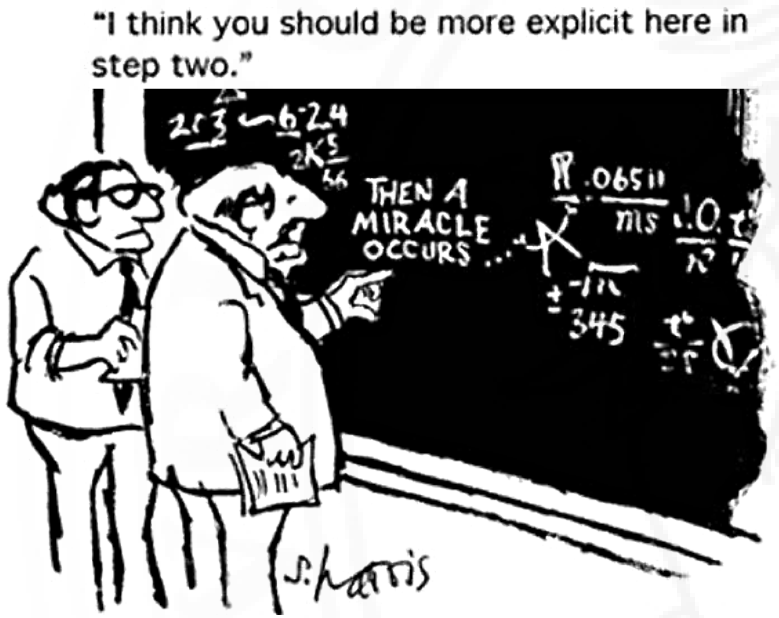
1. For Every Result, Keep Track of How It Was Produced
2. **Avoid Manual Data Manipulation Steps**
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Sandve et al. PLoS Comp Biol. 2013

Reproducibility crisis in Omics Research

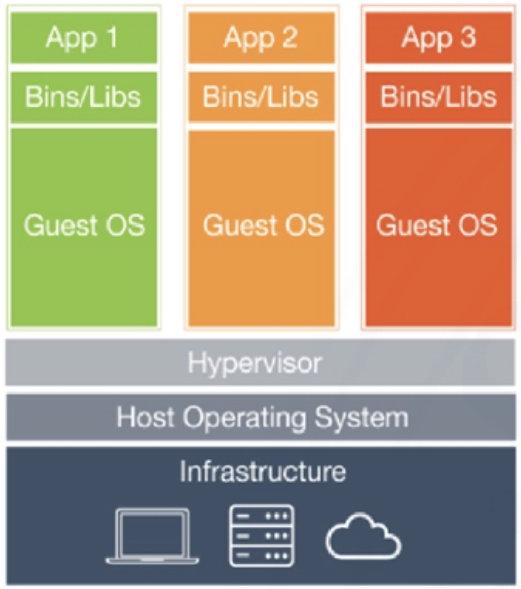
- ***Functional reproducibility***: the information about data and the utilized tools are saved as meta-data along with the generated data;

- ***Computational reproducibility***: extends the functional one storing the real image of the computation environment used to generate the data.

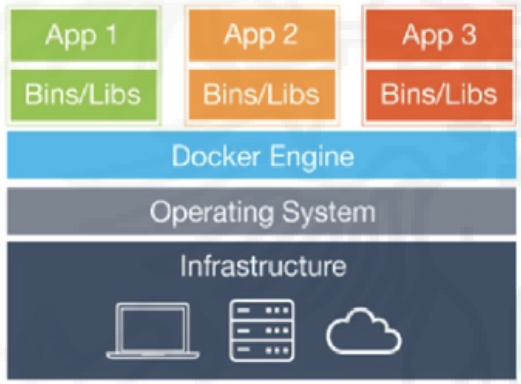


Reproducible research in Omics Research

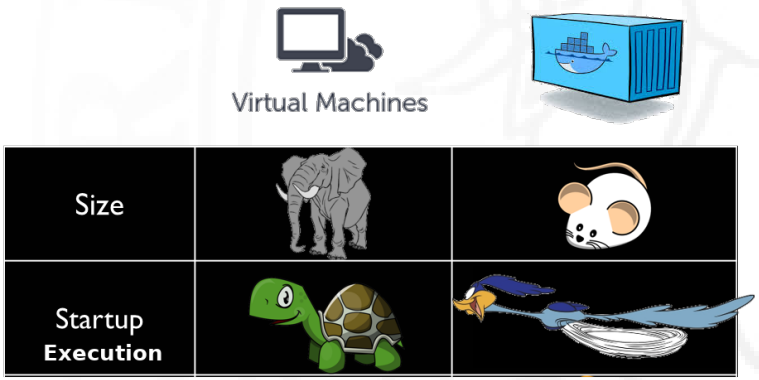
Container and Virtual machine are two virtualization techniques:



Virtual Machines



Containers



In Containers the sharing of same OS Kernel with the real machine reduces the portability, but ....

Reproducible research in Omics Research

- Reproducible Bioinformatics Project (RBP) is community opens to anyone interested to shared workflows under the umbrella of reproducibility;
- To enable an easy access NGS data analysis pipeline to users without advanced computer science skills;
- To provide robust and reproducible workflows fulfilling the Sandve's rules.

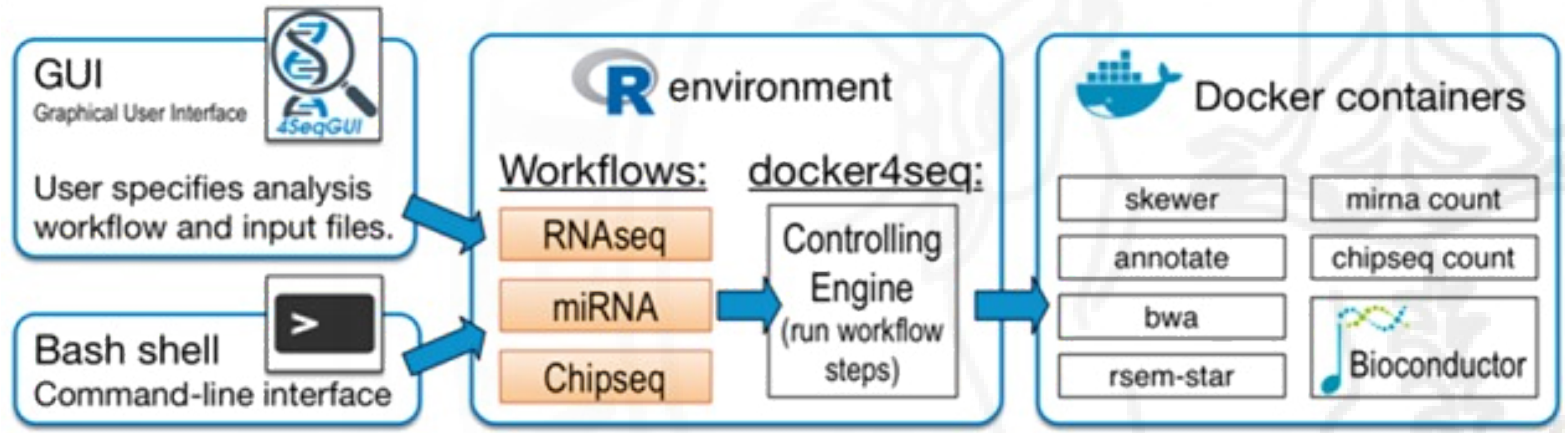
http://reproducible-bioinformatics.org/

- N. Kulkarni, L. Alessandrí, R. Panero, M. Arigoni, M. Olivero, F. Cordero, ***M. Beccuti*** and R. A. Calogero. **Reproducible Bioinformatics Project: A community for reproducible bioinformatics analysis pipelines.** BMC Bioinformatics, Volume 19, Issue 10, pages 211-219, October 2018.

Reproducible research in Omics Research

RBP general schema:

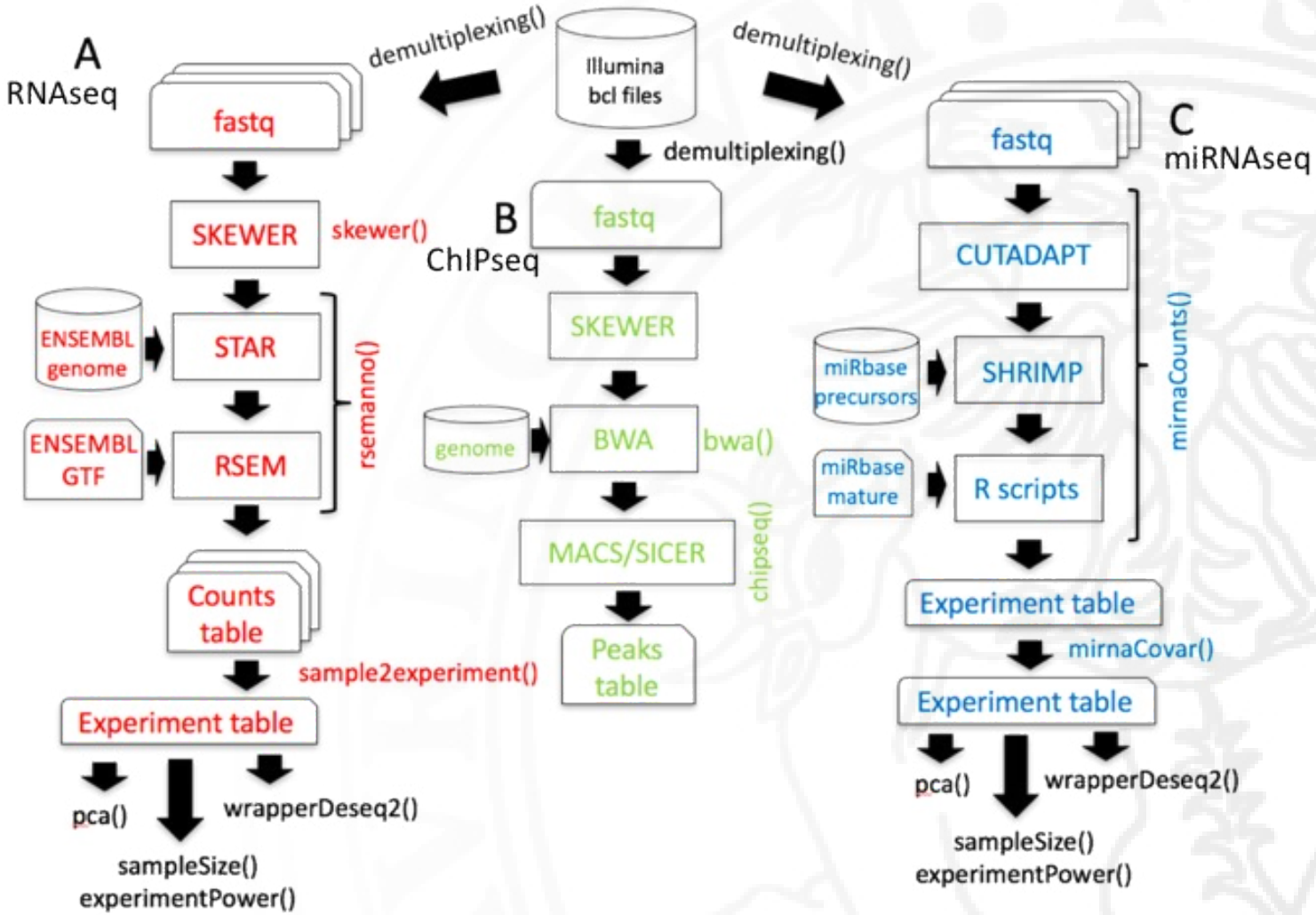
- Any workflow is specified as an R function that defines and controls the correct execution of all its tasks;
- any single task is encapsulated into docker images to guarantee the computation reproducibility.



Any workflow must be supported by an explanatory vignette

Reproducible research in Omics Research

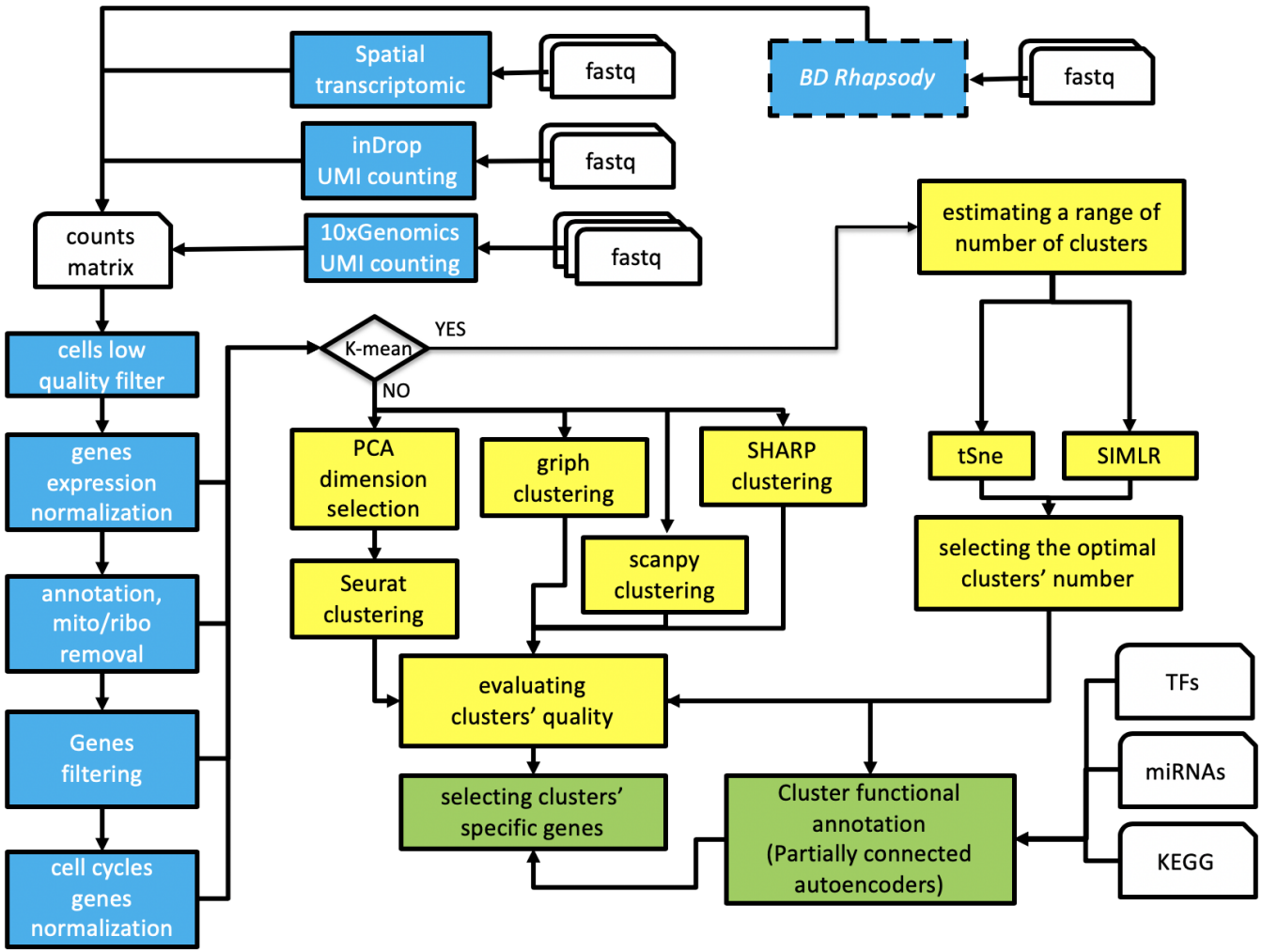
RNAseq, ChiPseq and miRNAseq analysis workflows



- ***M. Beccuti***, F. Cordero M. Arigoni, R. Panero, E. G. Amparore, S. Donatelli and R. A. Calogero. **SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer.** Bioinformatics, Volume 34, Issue 5, 1 March 2018, Pages 871-872. Oxford University Press.

Reproducible research in Omics Research

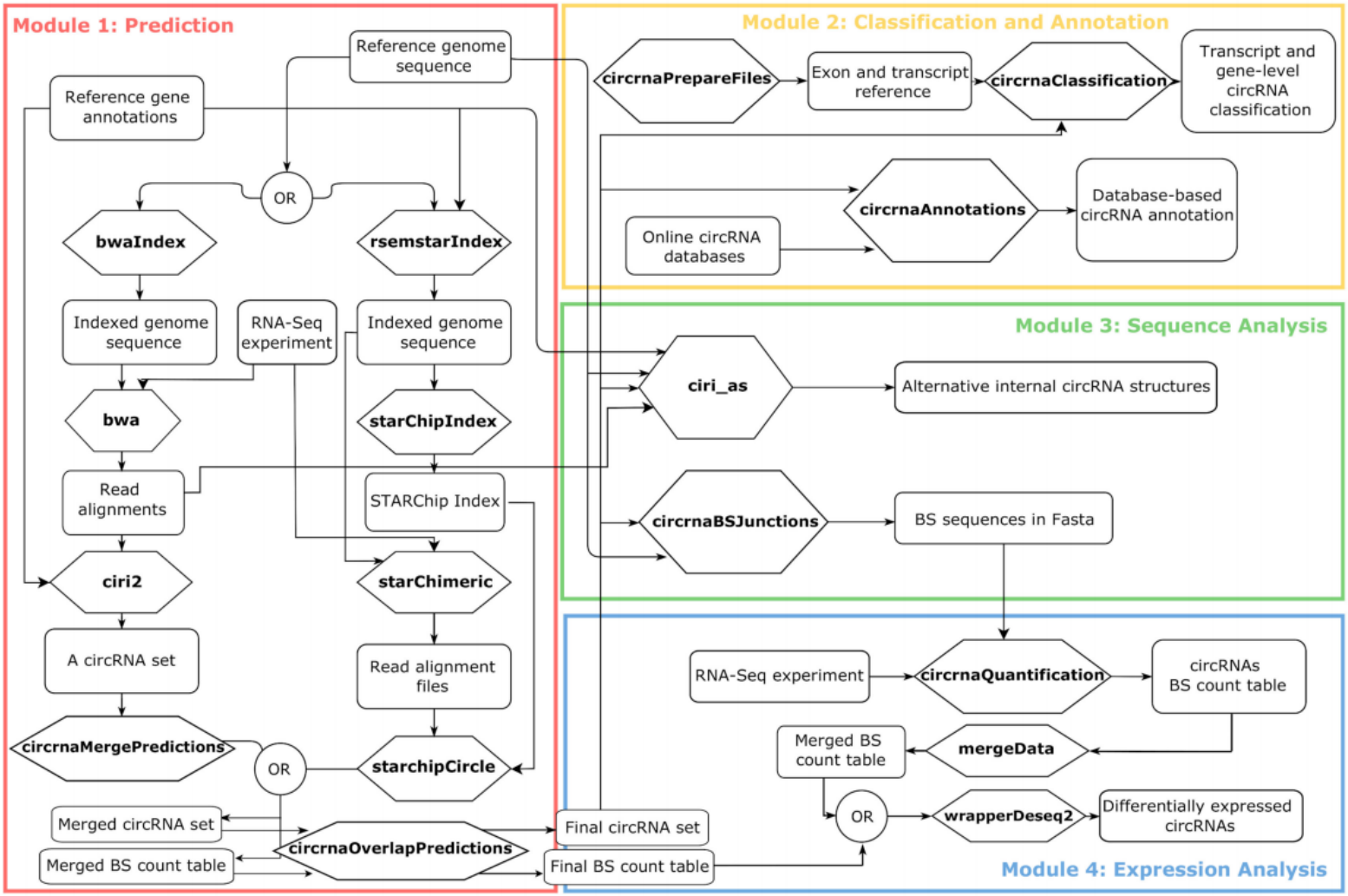
Single cell analysis workflow (rCASC)



- L. Alessandrí, F. Cordero, ***M. Beccuti***, M. Arigoni, M. Olivero, G. Romano, S. Rabellino, N. Licheri, G. De Libero, L. Pace, R.A. Calogero. **rCASC: reproducible classification analysis of single-cell sequencing data**. Gigascience, Volume 8, Issue 9, September 2019.

Reproducible research in Omics Research

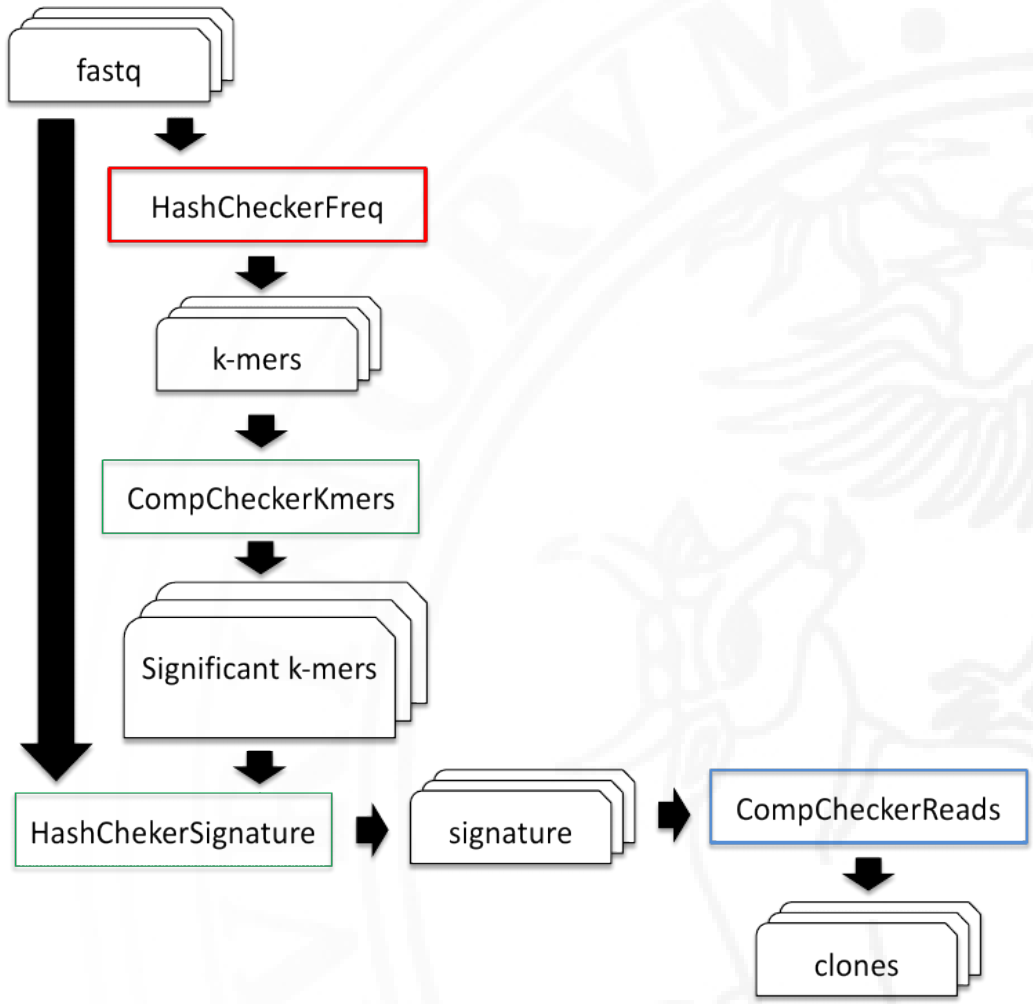
circular RNA workflow



- G. Ferrero, N. Licheri, L.C. Tarrero, C. De Intinis, V. Miano, R.A. Calogero, F. Cordero, M. De Bortoli, and ***M. Beccuti***. **Docker4circ: A framework for the reproducible characterization of circRNAs from RNA-seq data.** International Journal of Molecular Sciences, Volume 21, Issue 1, 1 January 2020, Article number 293.

Reproducible research in Omics Research

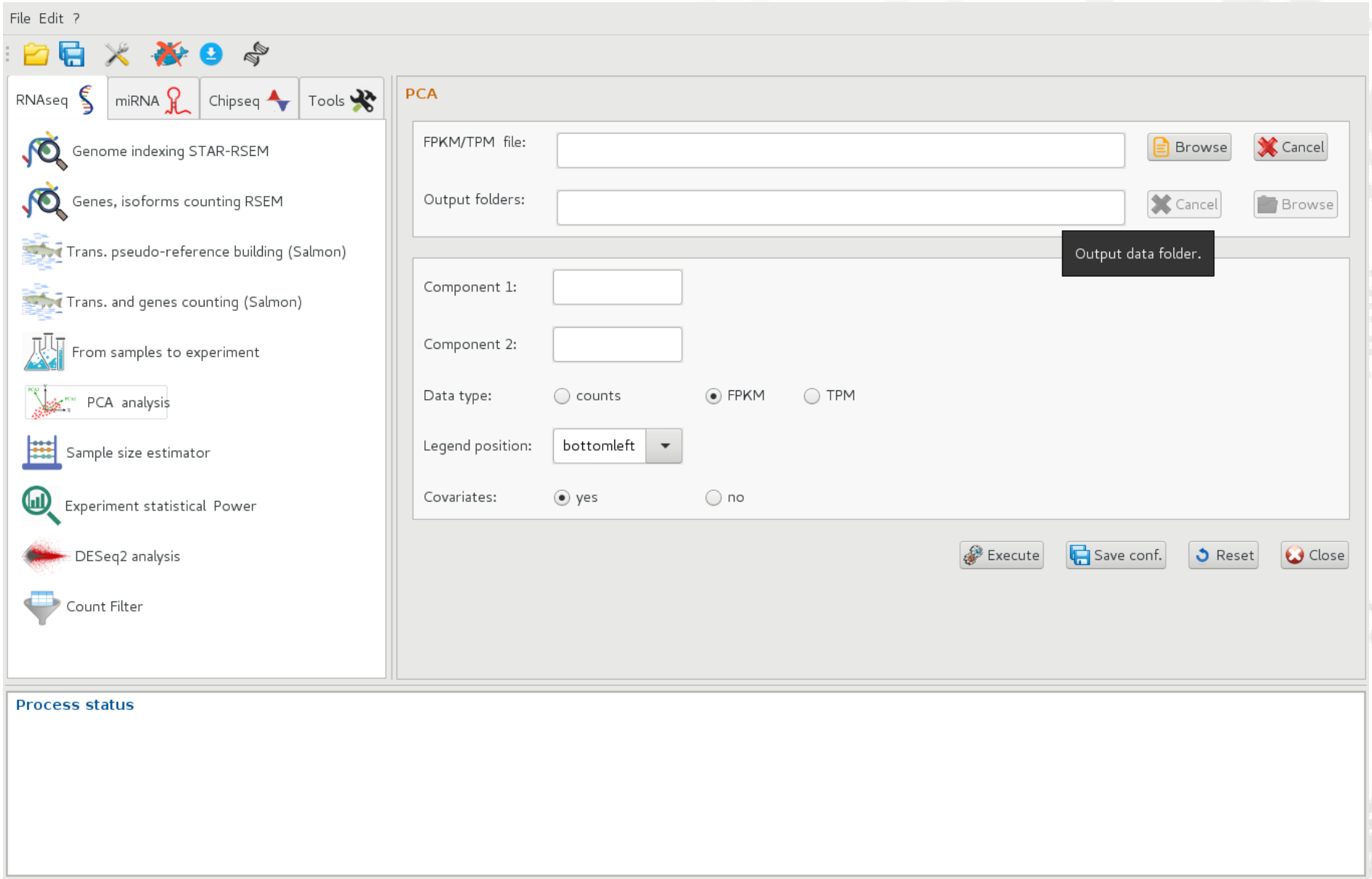
B-cells clonality assessment and minimal residual disease monitoring workflow



- ***M. Beccuti***, E. Genuardi, G. Romano, L. Monitillo, D. Barbero, M. Boccadoro, M. Ladetto, R. A. Calogero, S. Ferrero and F. Cordero. **HashClone: a new tool to quantify the minimal residual disease in B-cell lymphoma from deep sequencing data**. BMC Bioinformatics Volume 18, Issue 1, 23 November 2017.

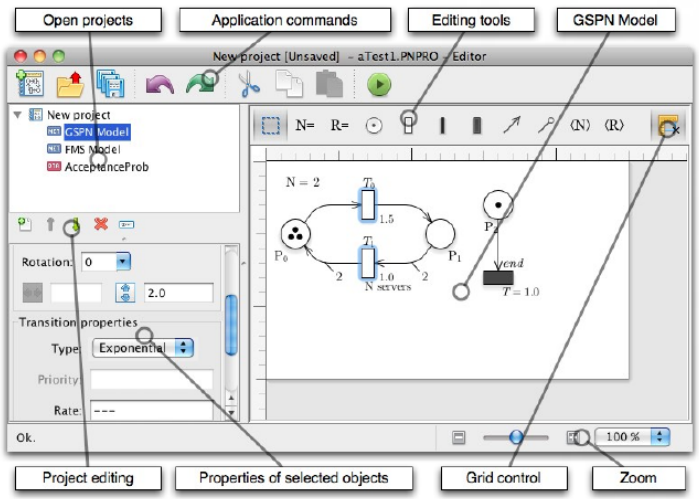
Reproducible research in Omics Research

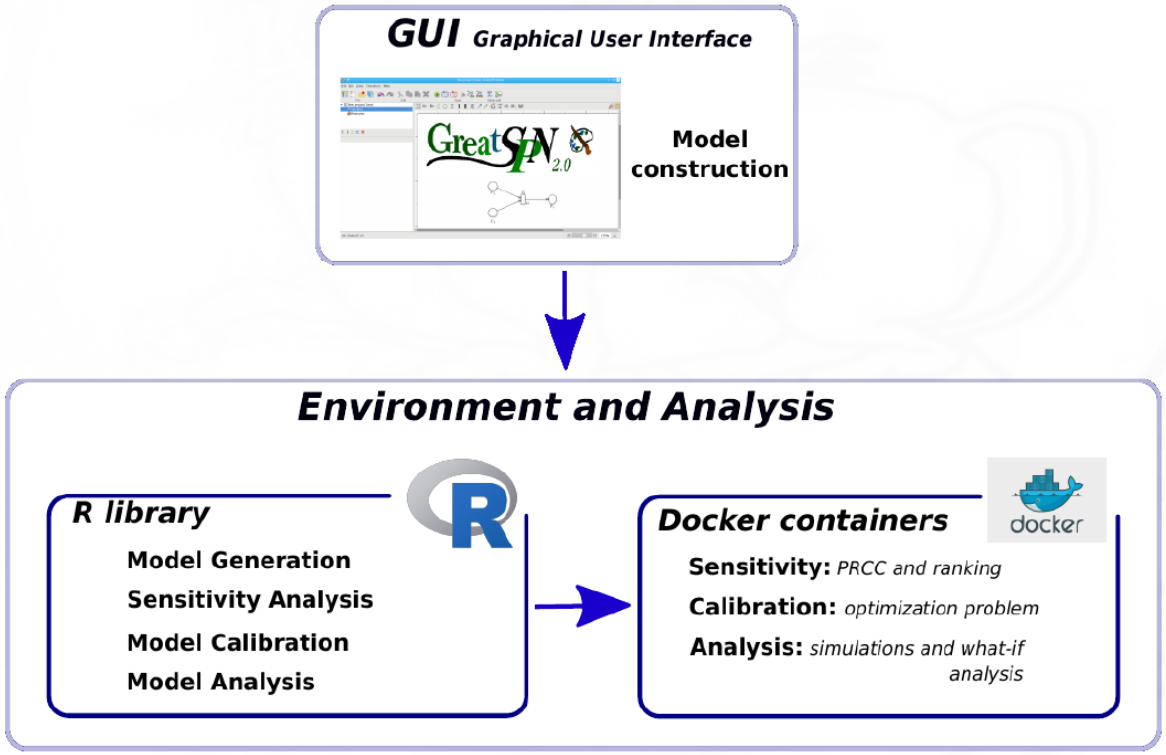
How to make easier the use of these workflows for beginner users:



Reproducible research in computational biology

Epimod is the reference RBP modeling framework based on a graphical formalism providing different analysis techniques for study biological and epidemiological systems.





Homepage:
https://github.com/qBioTurin/epimod

Reproducible research: a modelling framework

We apply **Epimod** framework on different case studies:

- to study Multiple Sclerosis under different strategies of drug administration;

  S. Pernice, et al. ***A computational approach based on the Colored Petri Net formalism for studying Multiple Sclerosis.*** BMC Bioinformatics, Vol. 20, 10 Dec. 2019, Article n.623.

- to study Pertussis Epidemiology and Vaccination;

  Castagno, et al. ***A computational framework for modeling and studying pertussis epidemiology and vaccination.*** BMC Bioinformatics, Vol. 21, 16 Sept. 2020, Page 344.

- to study Complex Metabolic Networks;

  S. Pernice, et al. ***Integrating Petri nets and Flux Balance methods in computationalbiology models: a methodological and computational practice.*** Fundamenta Informaticae, Vol. 171, Issue 1-4, 2019, Pages 367-392.

- to study the Impact of Reopening Strategies for COVID-19 Epidemic;

  S. Pernice, et al. ***Impacts of Reopening Strategies for COVID-19 Epidemic: A Modeling Study in Piedmont Region.***BMC Infectious Diseases. To be published.

- ...

Conclusion

- In this presentation we introduced the **Reproducible Bioinformcatics Project (RBP)**;
- We discussed how RBP workflows provide robust and reproducible NGS data analysis;
- We showed all the available workflows in RBP;
- We pointed out that all these worklfows are currently available on HPC4AI infrastructure.

Aknowledgements

di.unito.it

DIPARTIMENTO DI INFORMATICA
Università degli Studi di Torino

Q-bio group

Università di Torino
M B C
Molecular Biotechnology Center

Francesca Cordero, Ph.D.
Giulio Ferrero, Ph.D.
Laura Follia, Ph.D.
Beatrice Piaggeschi

Marco Beccuti, Ph.D.
Simone Pernice
Nicola Licheri
Vladimir Nosi

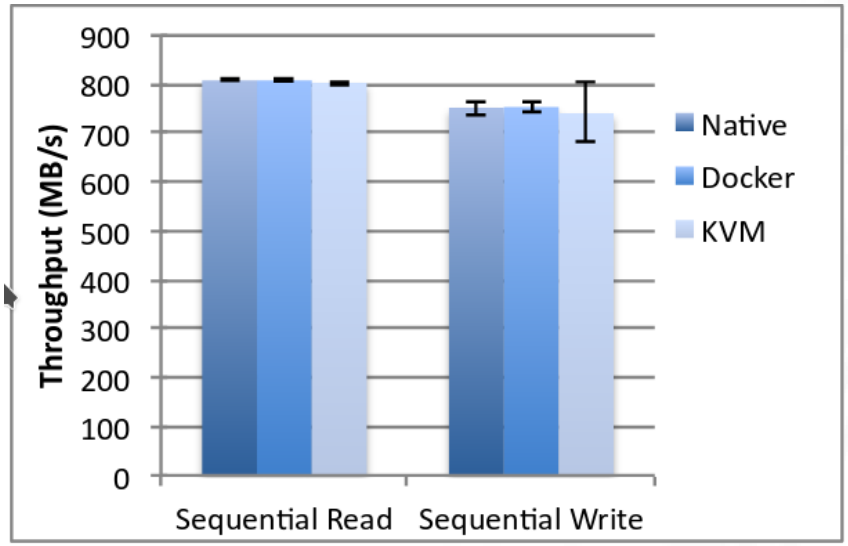
Prof. Raffale Calogero
Maddalena Arigoni, Ph.D.
Martina Olivero, Ph.D.
Luca Alessandri'

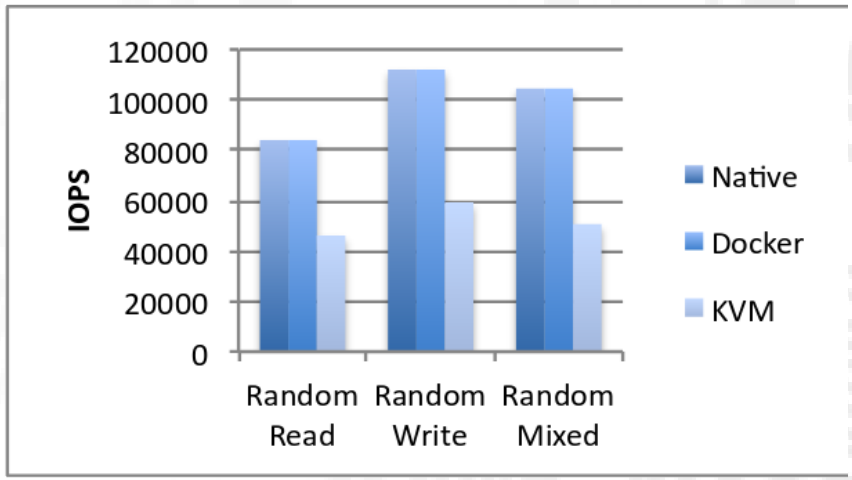

Thanks!


Any
Questions

Container, VM and real server: a comparison

In [1] a comparison among physical server, KVM, and Docker is reported.





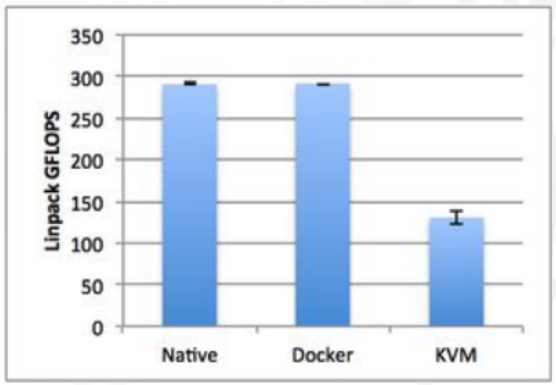
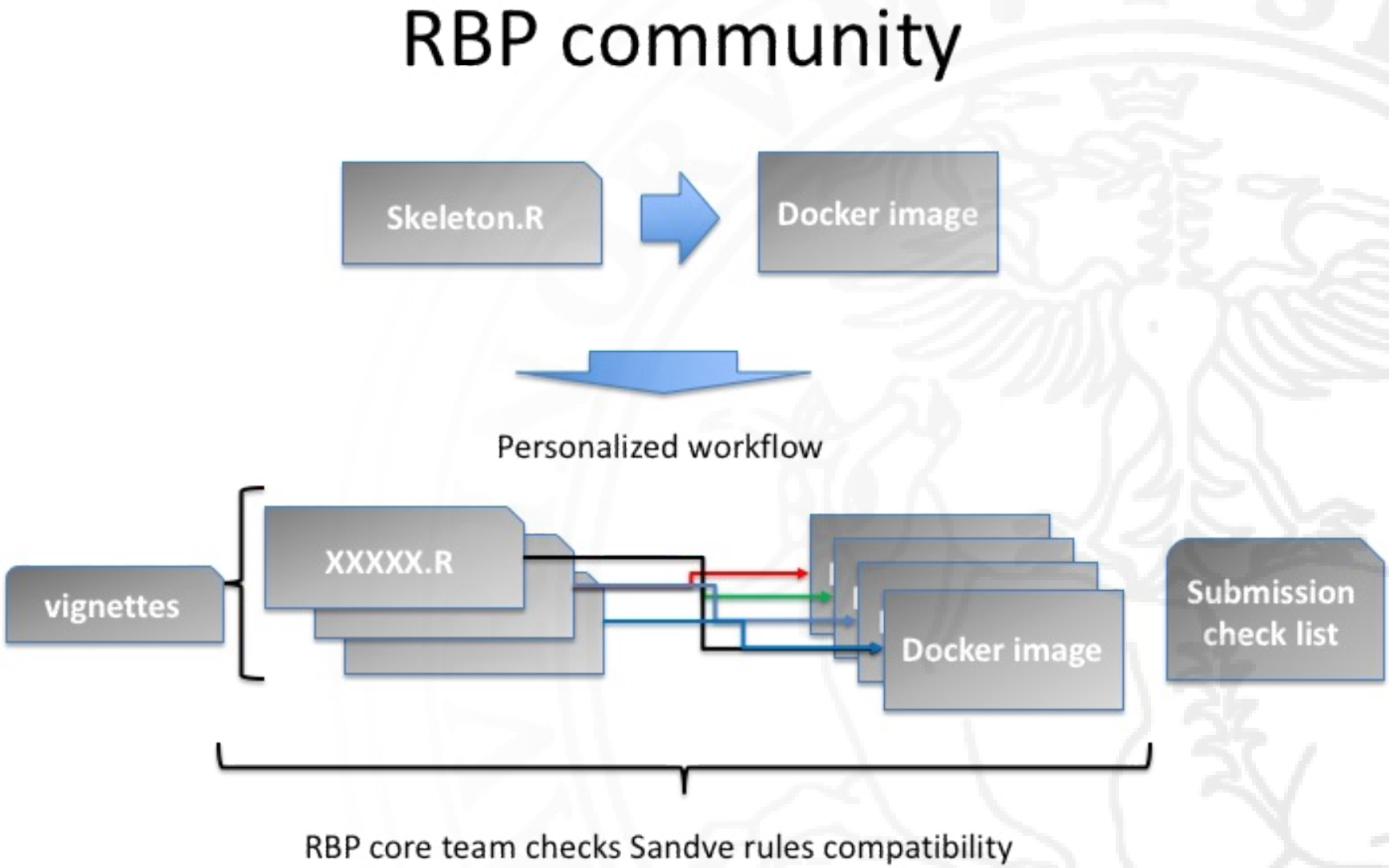


Figure 1. Linpack performance on two sockets (16 cores). Each data point is the arithmetic mean obtained from ten runs. Error bars indicate the standard deviation obtained over all runs.

[1] W. Felter, A. Ferreira, R. Rajamony and J. Rubio, ***An updated performance comparison of virtual machines and Linux containers***, 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Philadelphia, PA, 2015, pp. 171-172.

Reproducible research for NGS analysis

How to submit a new workflow:



Reproducible research in Omics Research

Some communities that fulfills many of the Sandve's rules:

- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data;
- It uses the R statistical programming language;
- it provides a version control for all its tools;

https://www.bioconductor.org/

Reproducible research in Omics Research

Some communities that fulfills many of the Sandve's rules:

- Galaxy is an open source, web-based platform for data intensive biomedical research;
- It provides substantial CPU and disk space, pre-installed tools making it possible to analyze large datasets;
- it guarantees (only) functional reproducibility;

https://galaxyproject.org/

Reproducible research in Omics Research

Some communities that fulfills many of the Sandve's rules:

BaseSpace SEQUENCE HUB

- it is a cloud-based genomics analysis and storage platform that directly integrates with all Illumina sequencers;
- Commercial tool implemented in Amazon cloud;
- Possible issue related to data privacy;
- it guarantees computation reproducibility using containerization and virtualization;

https://basespace.illumina.com