

rCASC implementation in Laniakea: porting containerization-based-reproducibility to a cloud Galaxy on-demand platform.



Integrating rCASC in Laniakea: rCASC, Cluster Analysis of Single Cells [Alessandri et al. BioRxiv], is part of the *reproducible-bioinformatics.org* project and provides single cell analysis functionalities within the reproducible rules described by Sandve et al. [PLoS Comp Biol. 2013].

Laniakea [Tangaro et al. BioRxiv Bioinformatics] provides

the possibility to automate the creation of Galaxy-based virtualized environments through an easy setup procedure, providing an on-demand workspace ready to be used by life scientists and bioinformaticians. The final goal is to offer rCASC as a Galaxy flavor in the Laniakea Galaxy on-demand environment.

Motivation

Reproducibility is an essential factor in establishing the reliability of results reported in the scientific literature and ultimately in the knowledge that is generated. This knowledge demands that the experiments and the analysis workflow used to generate their outcomes should be verifiable and potentially extendable through open and transparent processes.

In bioinformatics reproducibility issue is particularly magnified due to the short half-life of the bioinformatics software, the complexity of the pipelines, the uncontrolled effects induced by changes in the system libraries.

Methods

The Laniakea [Tangaro et al. BioRxiv Bioinformatics] Galaxy-on-demand framework (Fig. 1) was used as platform for the integration of rCASC [Alessandri et al. BioRxiv], which is a modular workflow (Fig. 2) providing an integrated analysis environment for the analysis of single cell sequencing data (from counts generation to cell subpopulation identification) exploiting docker containerization to achieve both computational (the ability to reproduce results independently by the underlying hardware) and functional (following the best practice rules for reproducible computational research, proposed in 2013 by Sandve [PLoS Comp Biol. 2013]) reproducibility in single cell sequencing data analysis.

Results

The rCASC workflow consists of 25 Docker images, each one implementing a specific analysis step. It was designed to use as orchestration software R (<https://github.com/kendomaniac/rCASC>), and it was not meant to be implemented on the Galaxy [4] platform. During this porting we identified a set of specific modifications in the docker structure, which are required to allow the integration of rCASC functionalities into Galaxy. Thus, we have outlined some guidelines for the integration of rCASC tools on Galaxy, e.g., how R scripts must be modified, how the management of inputs and outputs between the analysis steps should take place, etc. Six functions of rCASC workflow have already been successfully integrated and tested. Thanks to this pilot experiment, the implementation of the remaining functions should be straightforward since they are structurally similar to the six functions already integrated. Moreover, we have tested a specific Galaxy configuration which allows one to run both the common Galaxy tools, i.e., based on Conda packages, and the rCASC tools, which are based on Docker containers.

The final goal is to offer rCASC as a Galaxy flavor in the Laniakea Galaxy on-demand environment, thus giving the possibility to easily deploy a Galaxy instance containing rCASC tools which will allow a user-friendly and reproducible way to run this workflow thanks to the Galaxy platform graphical interface.

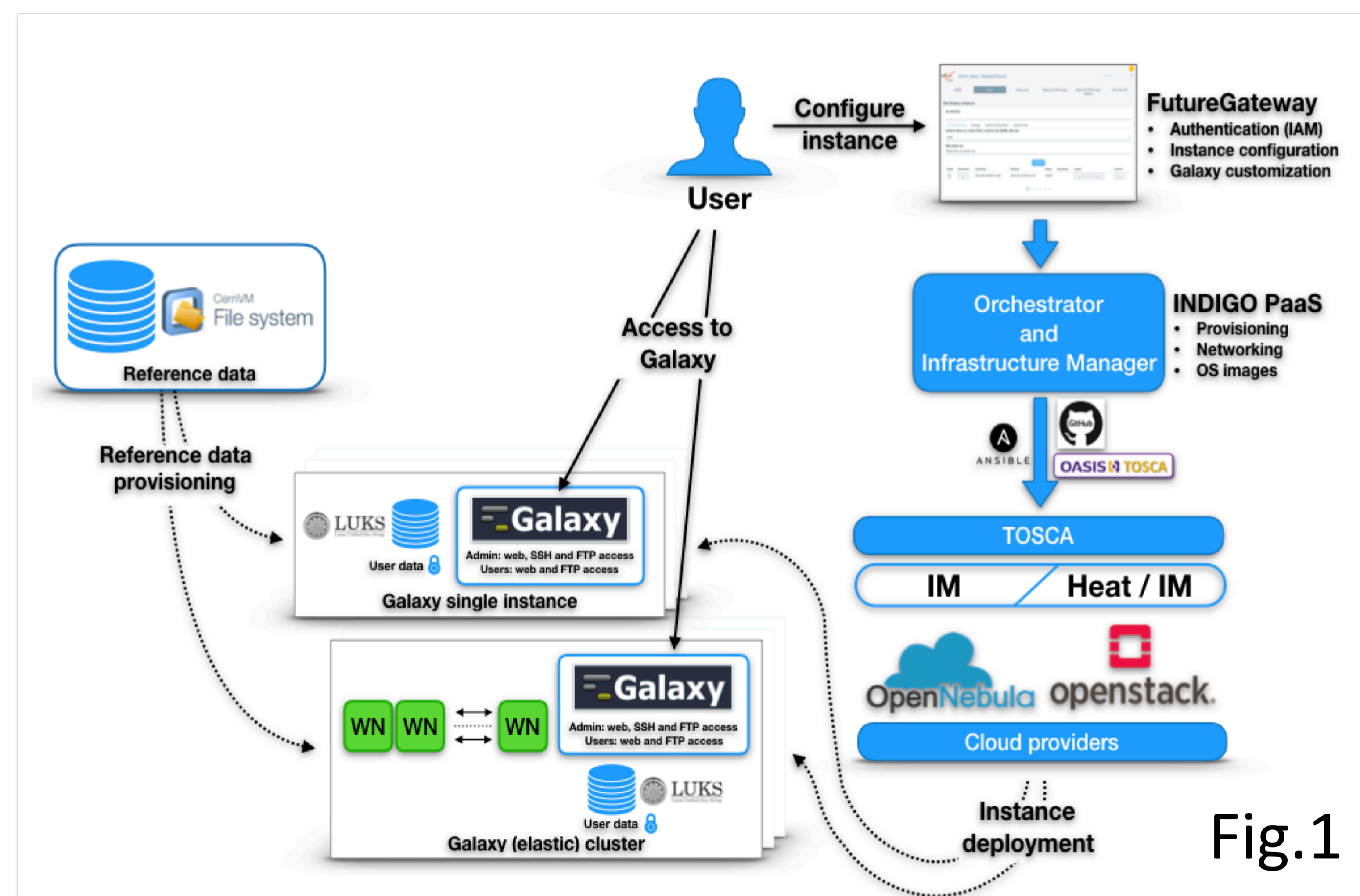


Fig.1

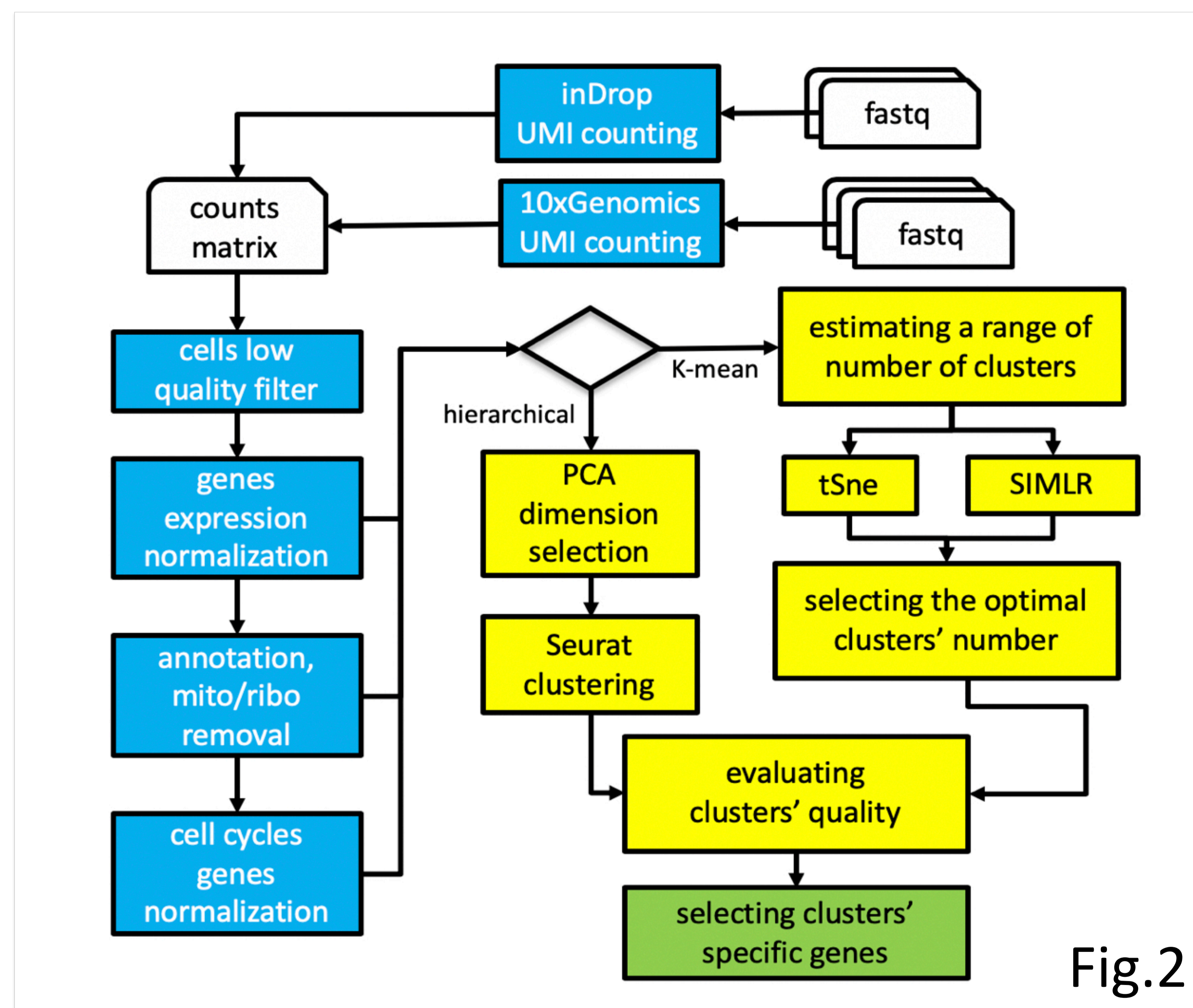


Fig.2

Contact

Elixir-Italy: Alessandri L (1), Mandreoli P (2), Tangaro MA, Beccuti M (4), Calogero RA (1), Zambelli F (2,3)

1) Bioinformatics and Genomics Unit, Dept. of Molecular Biotechnology and Health Sciences, University of Torino - Italy

2) Bioinformatics, Evolution and Comparative Genomics Lab Dept. of Biosciences, University of Milano - Italy

3) Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies - National Research Council - Italy

4) Dept of Computer Science, University of Torino - Italy

