

Engaging Galaxy in Microbiology

Gianmauro Cuccuru, Massimiliano Orsini, Andrea Pinna, Andrea Sbardellati, Nicola Soranzo, Antonella Travaglione, Paolo Uva, Gianluigi Zanetti, Giorgio Fotia



CRS4
Parco Scientifico e Tecnologico POLARIS, 09010 Pula (CA), Italy

Orione.crs4.it at a glance

Next Generation Sequencing is today widely applied in microbiology and metagenomics for research and diagnostic applications (Loman et al., 2012). The setup of the complete workflow to perform downstream analysis requires a significant effort to integrate software and data for each of the post sequencing steps.

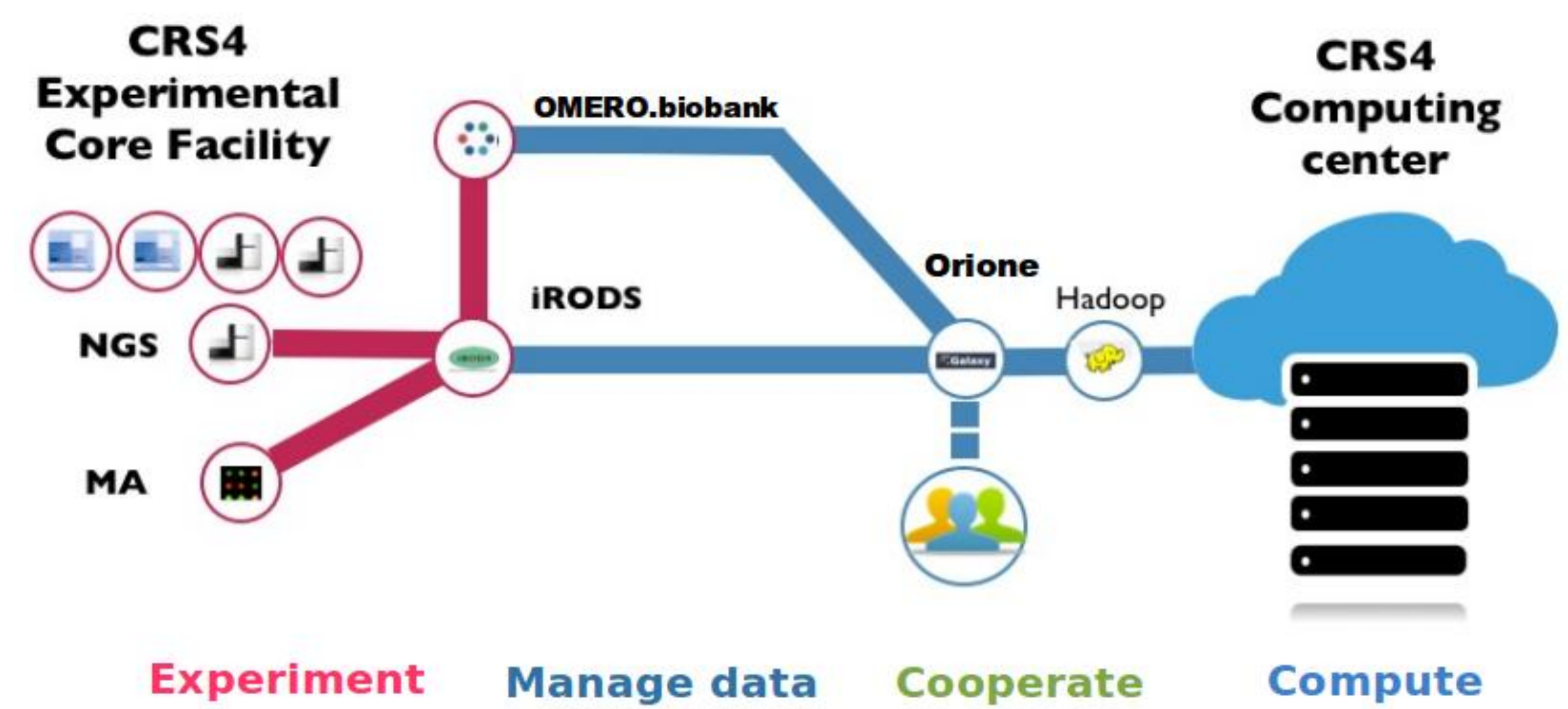
While many of the necessary tools are already available in Galaxy, there is currently a lack of a specialized framework in this area. To fill this gap we integrated into Galaxy the relevant software that cover the whole lifecycle of microbiology data analysis and created Orione, an environment offering many different tools, workflows, and options for microbiology ranging from bacterial genome assembling to differential transcriptional or microbiomes analysis.

The complete set of Orione tools and wrappers will be available soon through the Galaxy Community Toolshed <http://toolshed.g2.bx.psu.edu>.

Availability: Orione is available online at <http://orione.crs4.it>

A gateway to a large scale NGS facility

Orione is backed with a computational infrastructure specifically designed to support large scale data processing and storage in a scalable manner integrating state-of-the-art processing technologies, such as Hadoop, OMERO (Allan et al., 2012) and iRODS (Rajasekar et al., 2010), and fully integrated with a high throughput NGS facility that is the largest in Italy, capable to offer services as: Genotyping, mRNA and microRNA expression, Whole-genome sequencing, Exome sequencing, RNA-seq, ChIP-seq and much more (Cuccuru et al., 2013).



Preprocessing

Quality control
Filtering
Trimming
Manipulation

Assembly

Mapping:
Blast
Blat
Bwa/Seal, ...

DeNovo:
SSAKE
Velvet, Abyss
Edena, ...

Scaffolding

SSPACE
SSAKE
SOPRA
SEQuel

Post Assembly

SAM tools
Mugsy
MUSCLE

Annotation

Glimmer
Prokka
tRNA scan

RNA-seq

Get EDGE-pro
EDGE-pro

Metagenomics

Galaxy-Metagenomics
Metaphlan
READSCAN
Krona

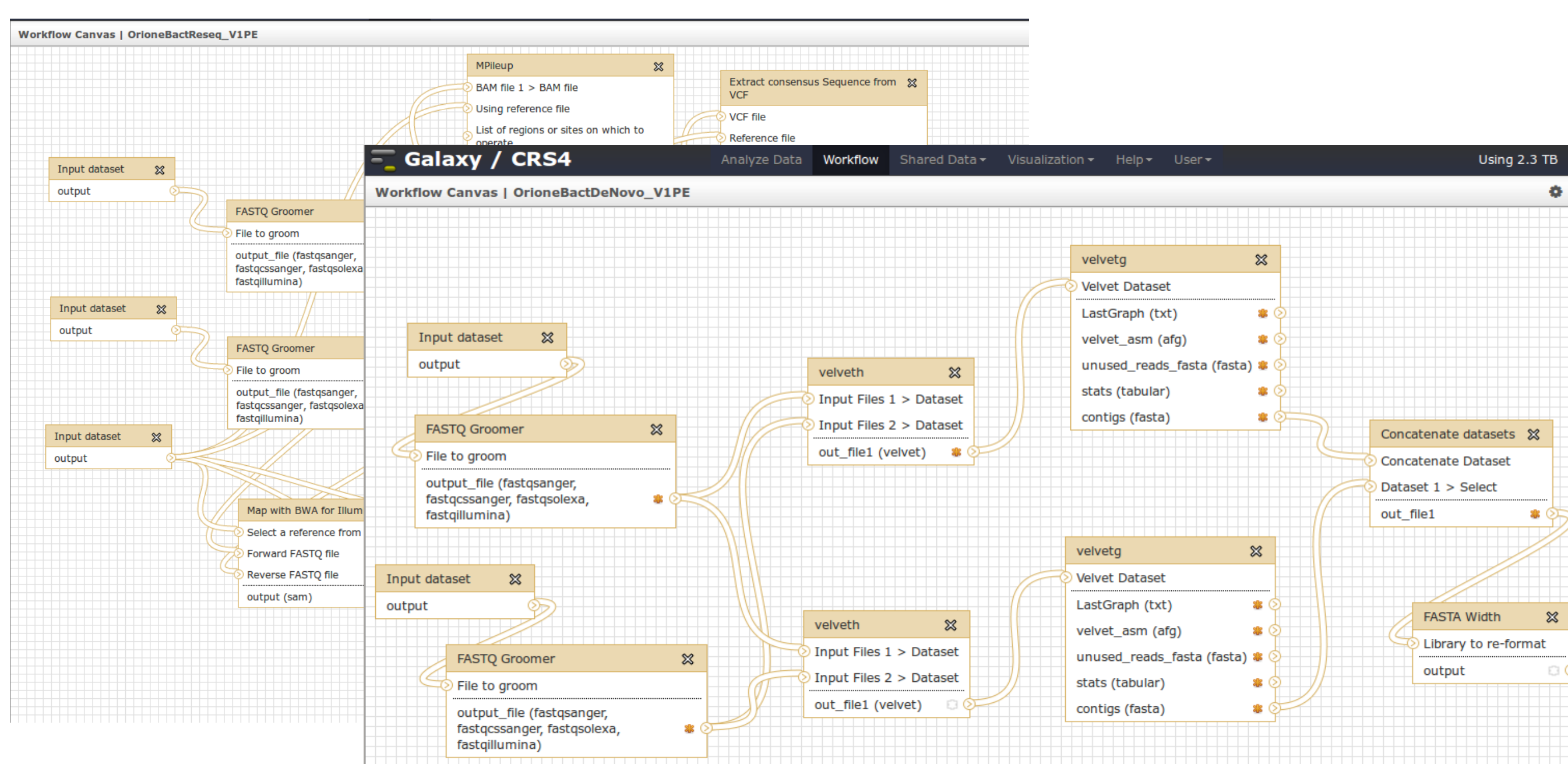
ORIONE.CRS4.IT

Overall schema of the Orione functionalities. Boxes represent collections of tools performing specific tasks. Italics indicates tools and wrappers developed by CRS4

Orione functionalities

Orione consists of 'best-of-breed' NGS downstream analysis tools covering end-to-end microbiology data analysis.

To provide the Orione users with some 'ready to go' applications we created a set of workflows covering bacterial resequencing, denovo assembly, draft annotation and metagenomics



Conclusions

We developed Orione, a Galaxy-based framework to build complex reproducible workflows for NGS microbiology data analysis.

Providing state-of-the-art software, pipelines and workflows specifically dedicated to applications in microbiology and metagenomics, Orione expands the constellation of specialized Galaxy-based web servers such as Nebula, Cistrome and several others offering additional capabilities to perform integrative, reproducible and transparent bioinformatic data analysis in microbiology.

References

- Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., et al. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature reviews. Microbiology*, 10(9), 599–606. doi:10.1038/nrmicro2850
- Goecks, J., Nekrutenko, A., Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86.
- Rajasekar A, Moore R, Hou CY, Lee CA, Marciano R, et al. (2010) iRODS Primer: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concept, Retrieval and Services* 2(1):1-143.
- Allan C., Burel J.-M., Moore J., Blackburn C., Linkert M., et al. (2012). OMERO: flexible, model-driven data management for experimental biology. *Nature Methods*, 9(3), 245-253. doi:10.1038/nmeth.1896
- Cuccuru, G., Fotia, G., Moore, J., Lianas, L., Pireddu, L., Swedlow, J., Zanetti G. Reproducible and automated processing in high-throughput NGS facilities. *Galaxy Community Conference 2013, Oslo, Norway.*