Data-intensive computational analysis of large scale microbiological data

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

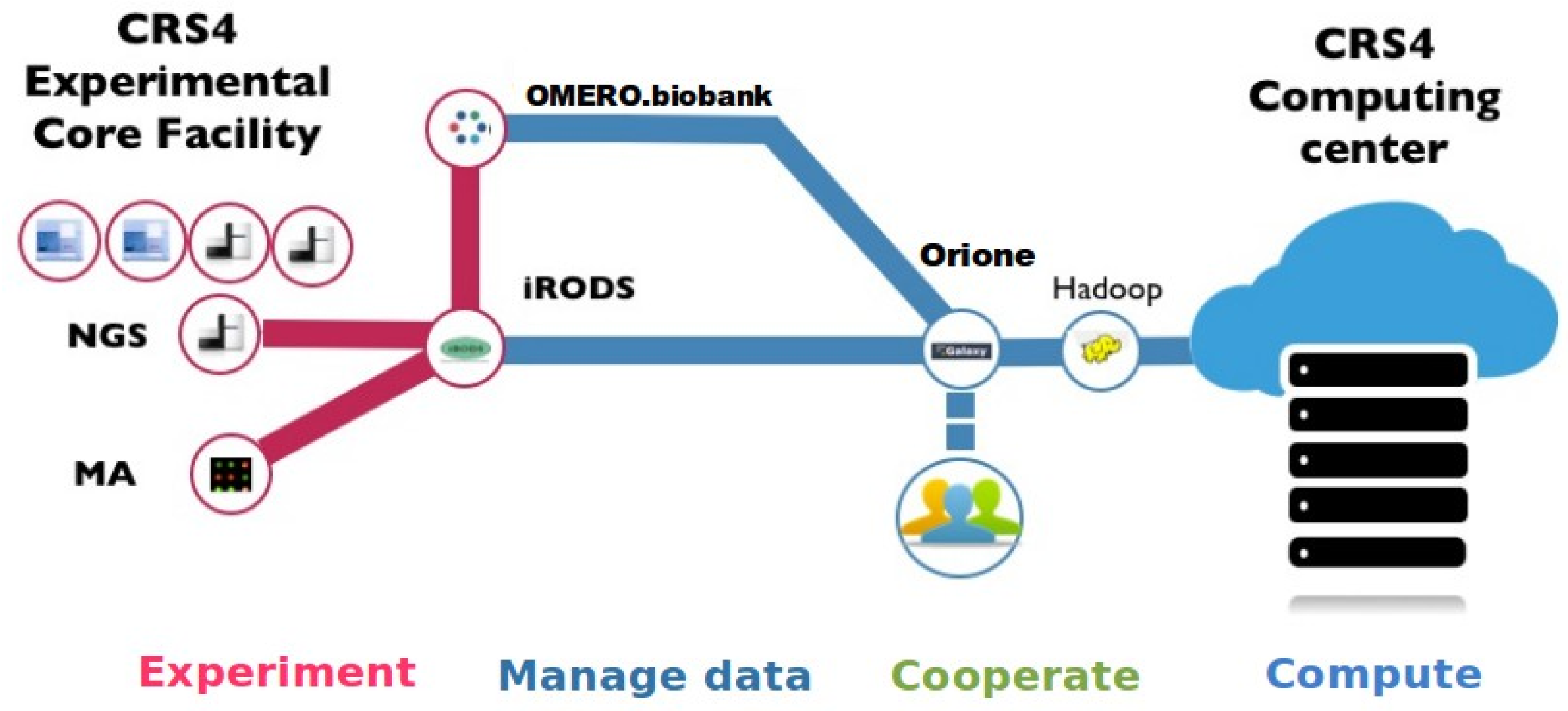
Application of Next Generation Sequencing (NGS) in microbiology is becoming a common practice with a profound impact on research, diagnostic and clinical microbiology (Loman et al., 2012).There exists serious technical challenges in the analysis of large scale NGS data sets that require adequate availability of resources (Chan et al., 2012). Furthermore, there is an increasing need to analyse and present data in a fashion that is transparent and reproducible and to provide analysis frameworks that are usable and cost-effective for biomedical researchers.

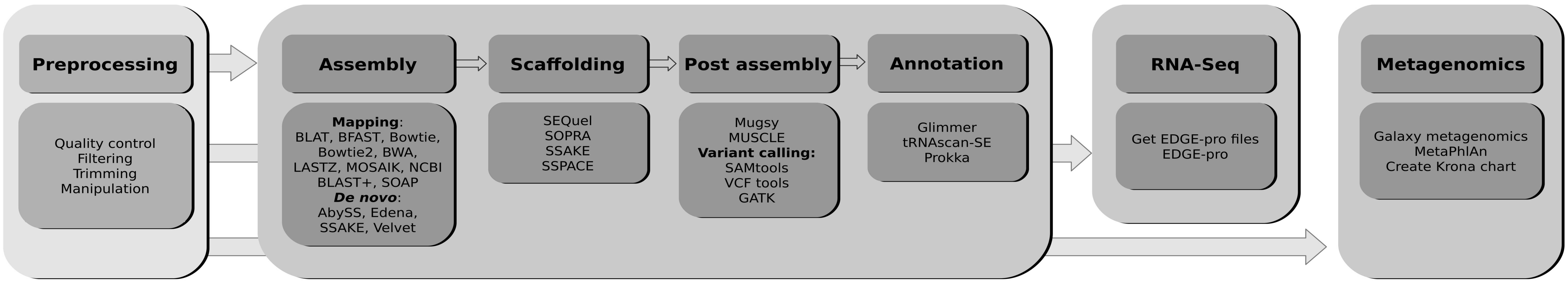
To address these challenges, we developed Orione, an online framework for downstream analysis of NGS microbiology data. Orione is based on Galaxy (Goecks et al., 2010), an open platform for data-intensive computational analysis and data integration utilized in many diverse biomedical research environments.

Orione is the first available platform that supports the whole life cycle of microbiology research data from production and annotation to publication and reuse. Orione is available online at http://orione.crs4.it

A gateway to a large scale NGS facility

Orione is part of an ongoing project to develop a scalable infrastructure for reproducible and traceable data-intensive biology which integrates Galaxy with state-of-the-art technologies such as Hadoop (Pireddu et al., 2011; Leo et al., 2012); OMERO (Allan et al., 2012) and iRODS (Rajasekar et al., 2010). This infrastructure is fully integrated with a high throughput NGS facility that is the largest in Italy and it is already used in production at CRS4 for the automated processing of sequencing data (Pireddu et al., 2013) and for quality control in gene therapy applications (Biffi et al., 2013).

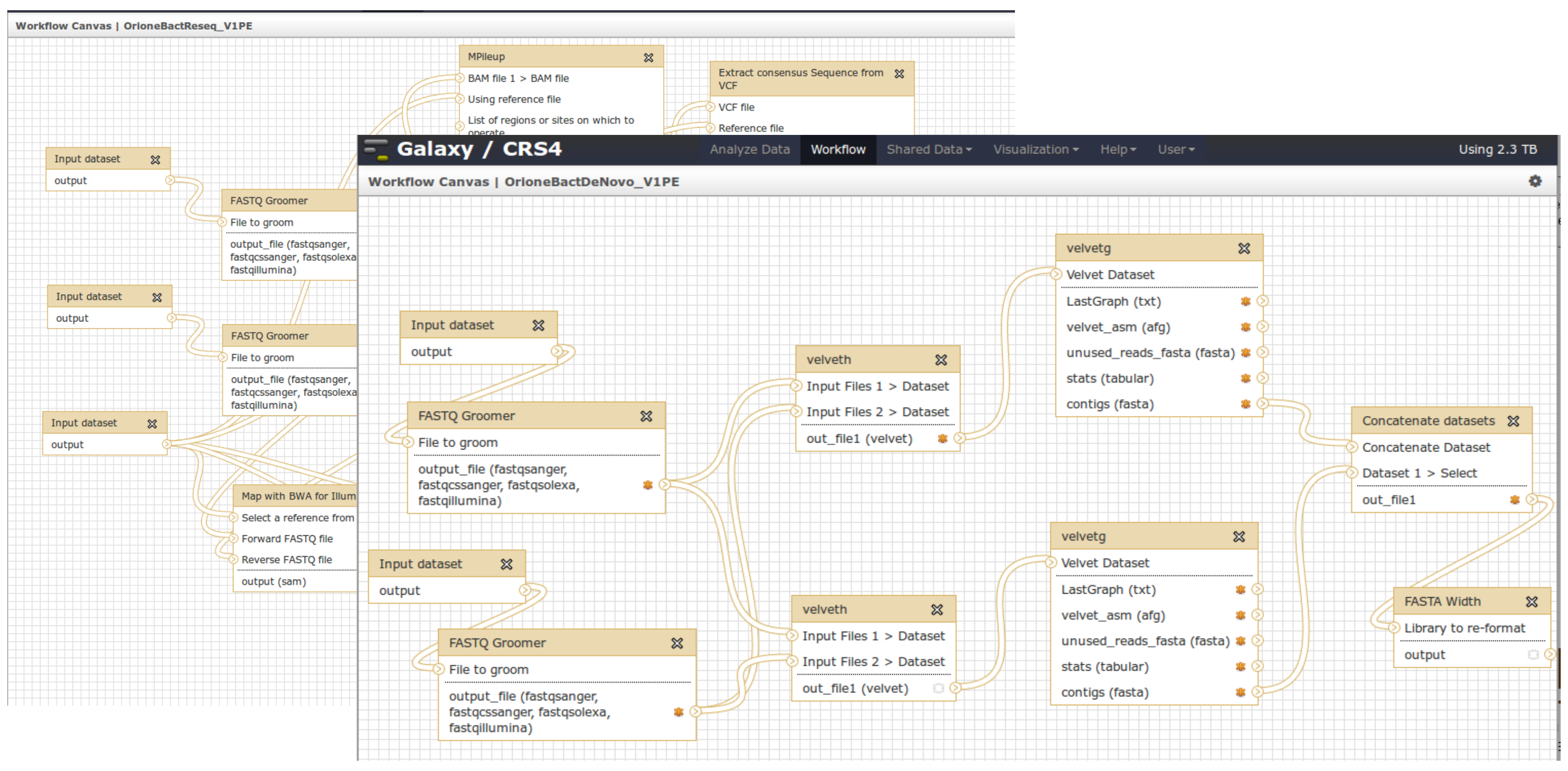




Overall schema of the Orione functionalities: boxes represent collections of tools performing specific tasks.

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. Orione is currently applied to a variety of microbiological projects, including bacteria resequencing, de novo assembling of both prokaryotic and eukaryotic organisms and microbiome investigations. Orione is also routinely used to analyse sequencing bacteria data at IZSAM (Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise), one of the Italian reference laboratories for foodborne pathogens.

Funding: this work was partially supported by the Sardinian Regional Authorities and by grant 095931 from the Wellcome Trust.

References

Allan, C. et al. (2012). OMERO: flexible, model-driven data management for experimental biology. Nat. Methods, 9(3), 245–253

Biffi, A. et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science, 341(6148), 1233158.

Chan, J.Z.M. et al. (2012). Genome sequencing in clinical microbiology. Nat.Biotechnol., 30(11), 1068–1071.

Goecks, J. et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol., 11(8), R86.

Leo, S. et al. (2012). SNP genotype calling with MapReduce. In Proceedings of The Third International Workshop on MapReduce and its Applications, MapReduce '12, pages 49–56, New York, NY, USA. ACM.

Loman, N.J. et al. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol., 10(9),599–606.

Pireddu, L. et al. (2013). Automated and traceable processing for large-scale high-throughput sequencing facilities. EMBnet.journal, 19(A), 23–24.

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.

Contact: gianmauro.cuccuru@crs4.it | http://orione.crs4.it

2013 International CAE conference, October 21-22, Pacengo del Garda (VR), Italy