Automated and traceable processing for large-scale high-throughput sequencing facilities

Luca Pireddu, Gianmauro Cuccuru, Luca Lianas, Matteo Vocale, Giorgio Fotia, Gianluigi Zanetti ✉
CRS4, Pula, Italy

Motivation and Objectives

Scaling up production in medium and large high-throughput sequencing facilities presents a number of challenges. As the rate of samples to process increases, manually performing and tracking the center's operations becomes increasingly difficult, costly and error prone, while processing the massive amounts of data poses significant computational challenges. We present our ongoing work to automate and track all data-related procedures at the CRS4 Sequencing and Genotyping Platform, while integrating state-of-the-art processing technologies such as Hadoop, OMERO (Allan, 2012), iRODS (Rajasekar, 2010), and Galaxy (Goecks, 2010) into our automated workflows.

Our main objective is to completely automate workflows from the moment the sequencer is started to the delivery of the processed sequencing data while keeping complete data traceability. The second, future, objective will be to reach full automation of selected pipelines for NGS downstream analysis such as variant calling. Through this effort, the Platform aims to cut the number of operator-hours required to process each sample while lowering its problem rate and costs. Simultaneously, the system will enable immediate access to detailed documentation of the processing undergone by each sample, even for special processing recipes.

Methods

To overcome the challenges related to the processing and tracking of large volumes of data and a large number of samples, we are building an automated and traceable processing platform based on four core services: Hadoop, OME Remote Objects, iRODS and Galaxy.

The computational heavy lifting with the sequencing platform is performed with the Hadoop framework. To enable this technology within the context of a sequencing center, we have adopted the Hadoop-based Seal toolkit (Pireddu, 2011), SeqPig (http://seqpig.sf.net), and Pydoop (Leo, 2010). Though CRS4 has its own computing facility with more then 3200 processors, it is not dedicated exclusively to the sequencing platform; therefore, an "elastic" Hadoop cluster allocation scheme has been developed in-house to allow effective use of shared storage and distributed computational cluster. In short, a Job Tracker (head node) is always up and ready to accept jobs. New worker nodes are allocated on demand through a standard Grid Engine queuing system. For this approach to work, we forego the HDFS file system and instead use a shared GPFS.

As the sequencing platform processes a relatively large number of samples, tracking their progress and tracing the specific processing steps applied to the resulting data is not a simple feat. To address this issue we have extended the OMERO system with models for high-throughput sequencing data; OMERO is a flexible, client-server, model-driven data management platform for experimental biology. Within the architecture presented in this work, it is the element that recalls how a given datum was produced – both in the source data used and the operations applied.

The sequencing operation generates a significant amount of data split over a large number of files and data sets. In addition, frequent collaborations with geographically dispersed entities introduce a requirement for fast and controlled remote access to data. To simplify and manage access to the data we have adopted iRODS, which provides a single point of access for all data sets, which may instead be distributed across a number of disjoint storage systems. In addition, the system allows one to associate meta-data and storage policies to data files and collections – e.g., compress all text files larger than 500 KB – and implements an optimized data transfer protocol.

To glue all these components together our architecture relies on the Galaxy web-based workflow engine. We have extended the Galaxy fork

created by Brad Chapman at the Harvard School of Public Health Bioinformatics to handle native Illumina flow cell descriptors and support the retrieval of all such information via a web service. Thus, the laboratory enters the complete flow cell composition through the modified Galaxy web interface; then, through the web service the sample tracking software can fetch the flow cell information and integrate it with the operations-related meta-data.

In addition, we have integrated with Galaxy the Hadoop-based tools used by our pipeline. Therefore, Galaxy is used to manage all workflow-based operations, while a custom "automator" daemon is used to execute and monitor workflow progress and to link workflows to each other – e.g., execute sample-based workflows after a flow cell-based workflow. One of the main advantages of this approach is that Galaxy natively tracks operations with its histories. In our system, successful completion of a workflow will trigger an action within the automator that will commit the data set to iRODS, extract the associated history from Galaxy and save it to OMERO. Additional integration work exposes these data sets registered within iRODS through the Galaxy libraries feature, from where users can perform their additional analyses on the data.

Results and Discussion

Currently, the core system is in its testing phase and it is on schedule to be in production use at CRS4 by May 2013. The results thus far obtained by combining Hadoop, OMERO, iRODS and Galaxy are encouraging and the authors are confident that the CRS4 Platform will increase its efficiency and capacity thanks to this system. In the near future, we plan to release the integration components as open source software. In addition, we are also working on the extending the framework to integrate different pipelines for downstream analysis of the sequencing data with a focus on microbiology and metagenomics.

Acknowledgements

This work has been funded by the Sardinian (Italy) Regional Grant L7-2010/COBIK.

References

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

Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**(8), R86. doi:10.1186/gb-2010-11-8-r86

Leo S, Zanetti G. (2010) Pydoop: a Python MapReduce and HDFS API for Hadoop. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 819–825. doi:10.1145/1851476.1851594

Pireddu L, Leo S, Zanetti G. (2011) SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27**(15), 2159-2160. doi:10.1093/bioinformatics/btr325

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.