Project Acronym: MuG

Project title: Multi-Scale Complex Genomics (MuG)

Call: H2020-EINFRA-2015-1

Topic: EINFRA-9-2015

Project Number: 676556

Project Coordinator: Institute for Research in Biomedicine (IRB Barcelona)

Project start date: 1/11/2015

Duration: 36 months

Milestone 15: Deployment of components in public and private clouds

Lead beneficiary: European Molecular Biology Laboratory

Dissemination level: PUBLIC

Due date: 31/10/2017

Actual submission date: 31/10/2017

Document History

Version	Contributor(s)	Partner	Date	Comments
0.1	Mark McDowall	EMBL-EBI	03/10/2017	First draft
0.2	Laia Codo	BSC	30/10/2017	Extended list of pipelines

Table of Contents

1 EXECUTIVE SUMMARY

There are 15 tools deployed with the VRE, currently 4 tools that are using the same architecture as used during the training workshop (Multi-scale study of 3D chromatin structure) that was run at EBI in April 2017. There are 6 tools that use the new version of the Tools API and are capable of running with the COMPSs environment and can be deployed by PMES. Of the 6 pipelines, 5 are in the process of integration with the VRE. The DM API and RESTful interfaces have been deployed and are currently (M24) being integrated into the VRE.

2 INTRODUCTION

The deployment of components has multiple stages. These split into identification, development and integration. As part of the development of the Tool API this formulated a guide for making a tool available within the VRE. Involved in this process has been the development/conversion of pipelines to allow them to function within the VRE.

3 INTEGRATING PIPELINES

As defined in MS13, all pipelines and tools that have been developed are made available on the Multiscale Genomics organisation portal on GitHub. The list of repositories and pipelines that they include are detailed in table 3.1. This is a continually evolving list with current pipelines evolving to match updates to the VRE and Tool API and new pipelines being identified. There are several tools that are already present within the VRE, but they do not use the Tool API. The Tool API relies on PMES and COMPSs as a way of controlling job management within the cluster of which there are several pipelines and tools. At the moment there are pipelines using 2 versions of the Tools API. The pipelines that use version 0.5 are capable of running within the COMPSs using PMES. Pipelines that are using version 0.6 of the Tools API also include the JSON config files so that they can be integrated into and deployed by the VRE.

Once the pipelines and tools have been created they are integrated into the VRE. This includes installation of the required software within a VM containing the latest version of COMPSs provided by the BSC and interaction with the VRE developers to register the VM. If there are any custom views of the generated data this also needs to be taken into account.

4 INTEGRATING THE DATA MANAGEMENT API

Along with the deployment of pipelines, there is also the DM module (mg-dm-api) along with the RESTful interface (mg-rest-dm) and file passing interface (mg-rest-file). These have been installed within the VRE and are in the process of integration within the VRE ready for release.

Pipeline/Tool	Description	Tool API Version	In VRE
3DConsensus	Identify 3D protein-DNA interactions	Pre API	Yes
Chromatin Dynamics	Visualise the chromotin filament base on nucleR results	Pre API	Yes
DNAShapeScan	Binding site prediction	Pre API	Yes
MC-DNA	Calculate B-DNA conformations	Pre API	Yes
MD Energy Refinement	Energy refinement	Pre API	Yes
NAFlex analysis	Fold modelling of DNA	Pre API	Yes
Nucleosome Dynamics	Nucleosome positioning and variability	Pre API	Yes
Process ChIP-seq	Processing of FASTQ sequence reads. Aligned using BWA, filtered by BioBamBam and peak calling with MACS2	0.6	Yes (dev)
Process RNA-seq	Alignment and analysis using Kallisto	0.6	Code is ready
Process WGBS	Alignment of FASTQ sequence reads with Bowtie2 and analysis by BS-Seeker2	0.6	Code is ready
Process MNase-seq	Nucleosome position calling	0.6	Code is ready
Process Genomes	Indexing of genomes (Bowtie2, BWA, GEM)	0.6	Yes (dev)
Process HiC	Alignments and analysis of HiC data using the GEM aligner and TADbit	0.6	Ongoing
pyDockDNA	Docking of DNA fragments	Pre API	Yes
TADbit bin		0.5	Yes (dev)
TADbit map, parse and filter	Alignment and filtering of fastq files	0.5	Yes (dev)
TADbit model		0.5	Yes (dev)
TADbit normalize	Normalisation of adjacency matrices	0.5	Yes (dev)
TADbit segmentation		0.5	Yes (dev)

TADbit Tool	Alignments and analysis of HiC data using the GEM aligner and TADbit	Pre API	Yes

Table 3.1: *List of the pipelines and current deployment. Tool API pipelines rely on pyCOMPs and PMES for deployment by the VRE. The pre API pipelines rely on Sun Grid Engine (SGE). All repositories are available from the MuG GitHub https://github.com/Multiscale-Genomics, where they are not present there are links to the relevant software or server locations.*

5 CONCLUSION

Development of the MuG VRE is progressing along with new pipelines being identified, integrated and deployed. As the APIs and service mature, the more pipelines and tools that are integrated into the system helps with the development of a standard integration procedure. A standardised procedure helps reduce the effort required by developers for integrating new tools and pipelines. As it becomes easier to integrate new applications it should enable a developer community to grow around the MuG VRE.