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 17: Early prototypes of the computational infrastructure

Lead beneficiary: Barcelona Supercomputing Center (BSC-CNS)

Dissemination level: PUBLIC

Due date: 01/09/2016

Actual submission date: 05/09/2016

Document history

Version	Contributor(s)	Partner	Date	Comments
0.1	Josep Ll. Gelpi	BSC	01/09/2016	First draft
0.2	Laia Codó	BSC	05/09/2016	Revised references
1.0	-		20/09/2016	Final version

Table of contents

1 INTRODUCTION

A Virtual Research Environment should provide their users with an adequate combination of relevant information, data, and computational tools. The combination should help the researcher to analyze data, either from repositories, or obtained from experiment or simulation; combine and compare such analysis results with related studies and reference data; and provide access in a friendly manner.

This document presents the installation of the initial prototype of MuG basic computational infrastructure. A more extended description will be available as project's deliverable D5.1. The initial installation provides a cloud-based environment able to perform a series of operations packed as virtual machines, and several interfaces of access, including specific web portals, and programmatic access.

2 COMPUTATIONAL INFRASTRUCTURE INITIAL DESIGN AND DEPLOYMENT

MuG computational infrastructure has been designed to fulfil the following principles:

1. Flexible environment, able to adapt to the specific needs of the analysis tools (from WP6), both in terms of software requirements, or computational resources.
2. Software scheduler, able to manage analysis workflows in a transparent and adaptable manner.
3. Multi-scale execution. Analysis workflows could be executed either at the cluster level, in HPC environments, or distributed infrastructures like EGI (https://www.egi.eu/).
4. Data repository, with a flexible infrastructure, and fully compatible with other repositories used by the project (see D4.1 and D4.2 from WP4). Data repository will integrate a personal user workspace.
5. Web-based access centered in the MuG multi-scale browser (designed in WP3). This will be complemented by programmatic access using well-known interfaces.

MuG infrastructure has been designed as an evolution of the cloud-infrastructure built for the project transPLANT (TransNational infrastructure for Plant Genomics. EC FP7 283496, http://www.transplantdb.eu/sites/transplantdb.eu/files/D5.2-transPLANT.pdf), located at http://transplantdb.bsc.es. The initial prototype of MuG infrastructure is a direct adaptation of such, with a series of MuG specific additions. Figure 1 shows a general schema of the projected infrastructure. Table 1 shows a list of the components that are part of the initial prototype, and those that are being currently developed. Finally, Table 2 shows a list of the tools already available in the infrastructure with specification of the type of access.

The infrastructure is available at http://www.multiscalegenomics.eu/MuGVRE/.

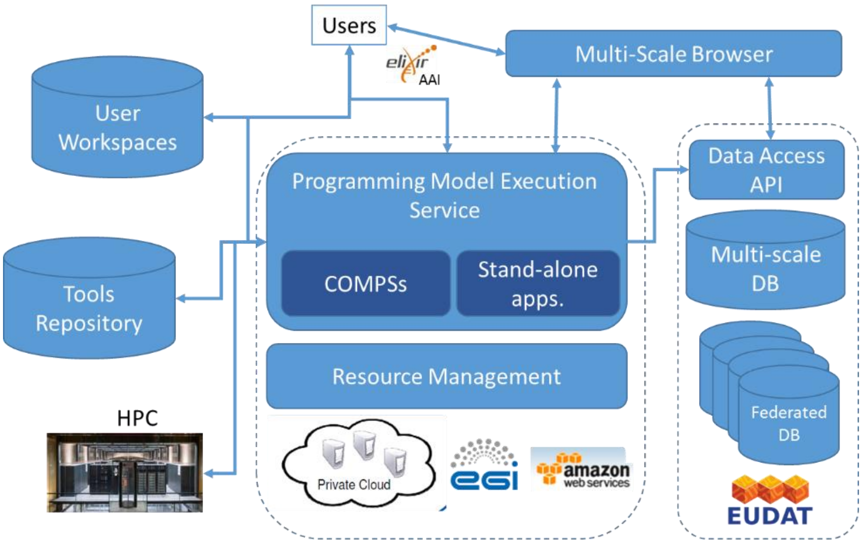

Figure 1. Projected layout of MuG's computational infrastructure

Table 1. Status of MuG infrastructure initial deployment

Component	Description	Status
Cloud manager	OpenNebula[1] cloud manager administrates hardware resources & Virtual Machine repository	Installed
COMP Superscalar	Multi-scale programming model	Installed
Programming Model Enactment system (PMES)[2]	Manages the execution of both stand-alone and COMPSs[3] based workflows. Controls OpenNebula resources through the rOCCI[4] protocol. Provides Programmatic Access to applications.	Installed
Sun Grid Engine	Queuing system to manage workflow execution. Complements PMES for specific applications.	Installed
Personal workspaces	User workspaces managed by traditional file systems, and a noSQL database. User authentication through traditional procedures.	Installed. The initial version maintains application-based workspaces where available. There is ongoing work to integrate them in a unique space. Additional authentication methods under study.
Data Repository	Data storage based in noSQL technology (MongoDB[5] & Cassandra[6])	Data available for 3D atomistic structure and simulation data, powered by the BiGNASim[14], and IRB's structural data repositories.
Multi-scale browser	Visual access to data analysis	Being developed at WP3. The prototype is linked to the TADkit tool for chromatin structure visualization,

		BIGNASim portal for atomistic simulation data, and to sequence data through JBrowse
Data Access APIs	All data will be accessible with the appropriate API based on REST protocol	To be developed

Table 2. Applications available at the prototype infrastructure

Application	Description	Type and Access
Serial-Maker +	General purpose genome annotation tool. Includes Maker[7], Exonerate[8], Augustus[9]	Virtual Machine, access through PMES Dashboard and WS
Bwapipeline	Sequence aligner for ngs data. Includes BWA[10], samtools[11], bcftools	Virtual Machine, access through PMES Dashboard
Bowtie+	Sequence aligner. Includes Bowtie[12], tophat[13], samtools[11]	Virtual Machine, access through PMES Dashboard
BIGNASim	Access to the MuG Section of BIGNASim[14], a general purpose database and analysis portal for Nucleic Acids simulations data	Web access.
Flexibility browser[15]	Access to flexibility data extracted from NA Simulations	Web access.
NucleosomeDynamics	Software suite for analysis of MNase-seq for Nucleosome positioning related data. Includes NucleR[16].	Web access.
TADKit 3D	Access to 3D Representation for chromatin conformation modeled from 3C-Data	Web access.
Genome Browser	Browser for reference genome data for Human, Drosophila and Yeast	Web access.

3 REFERENCES

(1) OpenNebula – http://www.opennebula.org

(2) Lordan F, Tejedor E, Ejarque J, Rafanell R, Álvarez J, Marozzo F, Lezzi D, Sirvent R, Talia D and Badia RM. *ServiceSs: An Interoperable Programming Framework for the Cloud*. Journal of Grid Computing. 2014 12(1) p. 1267-91.

(3) Tejedor E. and Badia RM. *COMP Superscalar: Bringing GRID superscalar and GCM Together*. 8th IEEE International Symposium on Cluster Computing and the Grid 2008.

(4) Open Cloud Computing Interface – http://occi-wg.org

(5) MongoDB –https://www.mongodb.org

(6) Cassandra – http://cassandra.apache.org/

(7) Cantarel B., Korf I., Robb SMC, Parra G., Ross E., Moore B., Holt C., Sanchez Alvarado A., Yandell M. *MAKER: An Easy-to-use Annotation Pipeline Designed for Emerging Model Organism Genomes*. Genome Research. 2008 18(1) p188-96.

(8) Slater G.S., Birney E. *Automated generation of heuristics for biological sequence comparison*. BMC Bioinformatics. 2005 6 p31.

(9) Stanke M, Waack S. *Gene Prediction with a Hidden-Markov Model and a new Intron Submodel*. Bioinformatics. 2003 19(2) p215-225.

(10) Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. 2013 arXiv:1303.3997v1.

(11) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. *The Sequence Alignment/Map format and SAMtools.* Bioinformatics. 2013 25(16) p2078–2079.

(12) Langmead B, Trapnell C, Pop M, Salzberg SL. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. *Genome Biol* 2009 10(3) R25.

(13) Trapnell C., Pachter L., Salzberg S. L. *TopHat: discovering splice junctions with RNA-Seq* Bioinformatics 2009 25(9) p1105-1111.

(14) Hospital A., Andrio P., Cugnasco C., Codo L., Becerra Y., Dans P.D., Battistini F., Torres J., Goñi R., Orozco M., Gelpí JL. BIGNASim: *A NoSQL database structure and analysis portal for nucleic acids simulation data.* Nucleic Acids Res. 2016 44(D1) p272-8.

(15) Hospital A., Faustino I., Collepardo-Guevara R., González C., Gelpí JL., Orozco M. *NAFlex: a web server for the study of nucleic acid flexibility.* Nucl. Acids Res. 2013 41(1) p47-55.

(16) Flores O, Orozco M. *nucleR: a package for non-parametric nucleosome positioning*. Bioinformatics. 2011 27 p2149–50.