Multiscale Complex Genomics

Project Acronym: MuG

Project title: Multi-Scale Complex Genomics (MuG)

Call: H2020-EINFRA-2015-1

Topic: EINFRA-9-2015: e-Infrastructures for virtual re.search environments (VRE)

Project Number: 676556

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

Project start date: 1st November 2015

Duration: 36 months

Deliverable 2.1: Project Fact Sheet

Lead beneficiary: Institute for Research in Biomedicine (IRB Barcelona)

Due date: 01/12/2015

Actual Submission date: 14/01/2016

Document history

Version	Contributor(s)	Partner	Date	Comments
0.1	Anna Montras	IRB Barcelona	30/12/2015	Circulated to supervisory board for approval
0.2	Rosa M. Badia	BSC	07/01/2016	Minor corrections to text
0.3	Anna Montras	IRB Barcelona	11/01/2016	Minor improvements to the brochure as suggested by BSC and IRB have been implemented.
0.4	Giacomo Cavalli	CNRS-IGH	12/01/2016	Name of PIs been added next to institutions in brochure and other minor changes.
1.0	Anna Montras	IRB Barcelona	14/01/2016	Revised version including remarks from supervisory board.

Table of contents

Executive summary

The present deliverable compiles the key facts about the project, including duration and costs, key expertise, objectives, challenges and impacts.

Following the development of the project visual image, a brochure has been created with the objective to support dissemination actions to be undertaken by all consortium partners.

The printable version of the brochure has been made available for download on the project website (http://www.multiscalegenomics.eu).

1 PROJECT FACTS AND FIGURES

As a first step towards the creation of the MuG dissemination materials (factsheet, website, etc.) the contents were defined. The main points over which the contents were developed are:

- **Project at a glance**: containing the main figures about the project, such as project duration, funding received and link to project website.
- **The need**: the motivation that instigated the project.
- **Consortium**: the partners, with very heterogeneous backgrounds, that have joined forces to give an answer to the target need.
- **Vision and mission**: the ambition of the project in the mid to long term.
- **Objectives and expected results**: the specific objectives that the consortium has defined for the three-year project, paving the way towards this long-term vision
- **Pilot projects**: the pilot cases that will be the main test bench for the developed innovations
- **Impacts**: the ultimate challenge of the project is to make a strong impact on end users at different levels, from academia to industry.

The contents of these points are developed below, and based on this information, the final printable project factsheet, in the format of a 3-page brochure and incorporating the project visual image, has been created (section 2).

MuG project at a glance

Project title: **Multi-scale Complex Genomics (MuG)**
Coordinator: Dr. Modesto Orozco (IRB Barcelona)
Project website: http://www.multiscalegenomics.eu
Contact: irbmultiscalegenomics@irbbarcelona.org
Start date: **1st November 2015**
Project duration: **36 months**
EC contribution: **2.961.163 €**

The need

> **3D and 4D genomics represent one of the greatest challenges for biology and biomedicine in the next decade. Understanding how genome is organized in the space and how this affects gene regulation would be instrumental to fully understand the time-dependent connection between genome and phenome.**

Unlike 1D genomics, which is a single-resolution problem that can be tackled using a common set of tools, 3D and 4D genomics represent a multi-resolution physics problem. From a computational point of view, 3D/4D genomics is in urgent need to develop **standards**, to complete **software and data infrastructure** and to design simple procedures to approach High Performance Computing **(HPC) to the community**.

Research efforts in 3D and 4D genomics are currently experiencing an exponential growth. However, the community is facing major problems as a result of the lack of efficient IT resources, which is leading to duplication of efforts, loss of information and severe misuse of existing resources.

Strongly fragmented, the community faces severe practical problems:

- How to integrate huge, noisy, and diverse data related to widely different size and time scales.
- A lack of standardized analysis and simulation tools, which threatens to compromise the reproducibility and robustness of the conclusions.
- The complete disconnection of associated databases.
- A lack of validated and flexible visualization engines.

MuG brings together different levels of expertise: (i) **biologists** with interest in chromatin structure, (ii) **methods developers** and (iii) **HPC facilities** with strong history of supporting Bio-computational problems, with the ambition of contributing to the **progress on this new and exciting field**.

MuG Consortium

Coordinating institution: Institute for Research in Biomedicine (IRB Barcelona)

Coordinator: Dr. Modesto Orozco: modesto.orozco@irbbarcelona.org

Joint research activities are the core of the MuG project, as shown by the strong **multidisciplinary** character of our consortium:

1. The Life sciences department (Josep Ll. Gelpí) and Computer Siencies department (Rosa M Badia) of a supercomputing centre (BSC-CNS) with large experience in supporting bio-HPC problems. Broad experience in modelling protein-protein and protein-nucleic acid complexes

2. Two groups (Charles Laughton of UNOT and Modesto Orozco of IRB Barcelona) with expertise in developing tools for high-resolution DNA simulation and with large experience in analysis of DNA physical properties.

3. Broad experience in modelling protein-protein and protein-nucleic acids complexes (BSC-CNS)

4. A group responsible for the development of the most used tools for analysis and simulation of nucleosome fibers (Modesto Orozco of IRB Barcelona)

5. A group with a world-leading activity in developing approach for reconstructing chromatin structure from Hi-C data (Marc Martí-Renom of CNAG-CRG)

6. One of the world-leading groups (at the European Bioinformatics Institute) in genomics data management and vii) one of the most active groups (Giacomo Cavalli of CNRS-IGH) in connecting chromatin structure with cell behaviour.

MuG Vision

The interface between 3D/4D genomics and HPC remains an unexplored area. Current heterogeneity of data, disconnection among databases and lack of standardization in analysis and simulation tools, are threatening to hold back advances in this field. MuG defines a new paradigm of relationship in 3D/4D genomics, providing new tools, favouring the transfer of codes, the generation of a scheme to share data, standardization of protocols and interaction with the HPC world.

The ultimate ambition of the Multi-Scale Complex Genomics (MuG) Virtual Research Environment (VRE) is to respond to the latest computational challenges of 3D/4D genomics by bringing this community closer to the HPC world and providing a suitable set of tools and infrastructure.

- MuG aims to **nucleate the 3D/4D community** and approach it to the HPC and Big Data world.

- MuG aims to **prevent a potential collapse** in the 3D/4D genomics field due to inability to manage the computational problems originated by massive sequencing experiments and simulations.

- MuG aims to provide an **integrated three-dimensional picture of the genome**.

MuG Mission

Targeting the abovementioned vision, the mission to be accomplished within the framework of the 3-year MuG project can be summarized in the following points:

- ✓ **To stimulate and engage the active and growing biology community** to interact with the HPC world. This is to be achieved by defining standards in software and data and creating a sustainable, HPC-optimized infrastructure to foster research in this field.

- ✓ To provide 3D/4D genomics users with **suitable tools to reduce current uncertainties** and inefficiencies.

- ✓ To develop the first **scalable infrastructure** for multi-scale (3D/4D) complex genomics

- ✓ To develop the first **multi-scale browser**.
- ✓ To ensure that the developed tools are embraced by the **Bio-community**.
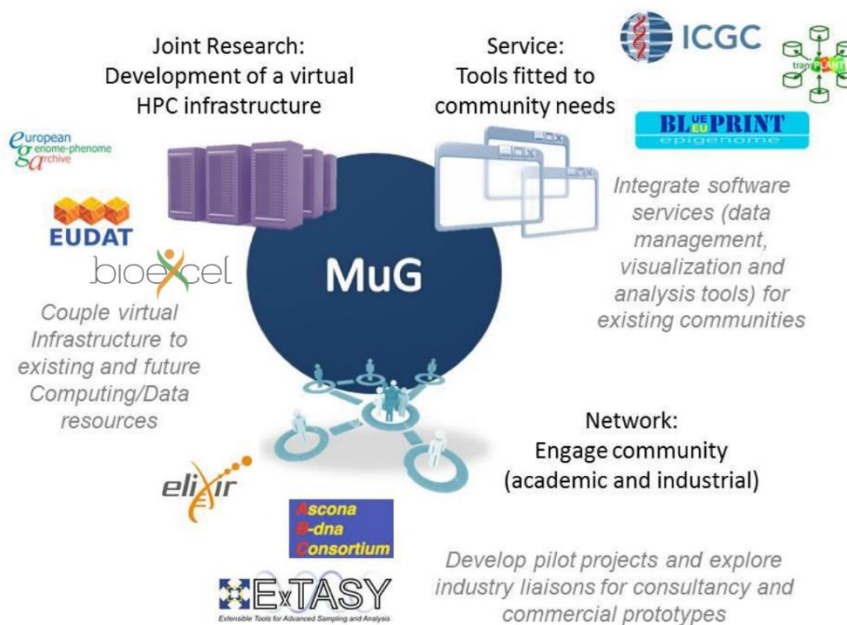
Objectives

The key **technological and scientific objectives** of the MuG project are:

1. To develop, benchmark, standardize, and integrate in common workflows multi-physics **simulation and analysis tools** used in 3D/4D genomics.

2. To create a common data infrastructure, generate a common database repository, generate the required data ontology, and integrate 3D/4D data with structural and 1D genomics databases.

3. To create a multi-resolution genome-browser with new visualization tools able to move from 1D genomics, to low- and even high-resolution levels.

4. As a result of the above developments and through pilot studies to prove the power of the developed methodology: improved understanding of chromatin structure and dynamics and other biological relevant topics.

5. To develop user oriented solutions and to engage both the academic and the industrial sectors.

Outputs and impacts

MuG aims to go beyond the state of the art in different fields to convince potential users on the potential of our vision and on the advantage of joining the initiative.

MuG Innovations
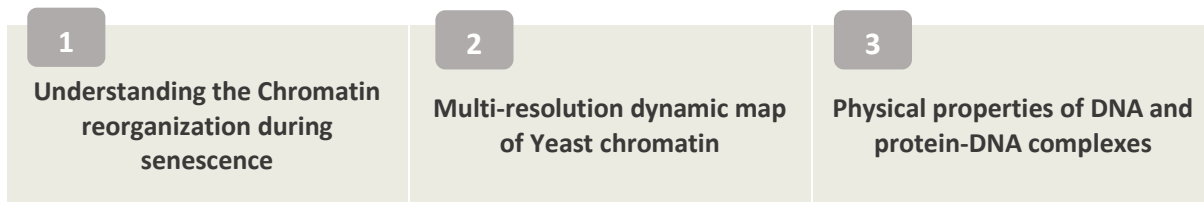
1. Connection of multi-scale data
2. Visualization complex genomic machinery with a 4D multiresolution picture
3. Modelling of chromatin
4. Definition of a new paradigm of relationship in the 3D/4D genomics community

Service activities developed and deployed:

- Infrastructure provisioning: user accounts, user certificates and assignment of a gateway for access to the European data and computing infrastructure
- Development of user interfaces
- Interoperability
- Deployment, maintenance and support
- Documentation and training

The pivotal point of all development tasks of the MuG project is the development and deployment of the computing infrastructure necessary to support the VRE services. These tasks are centralized in WP5, which integrates 3D/4D Visualization (WP3), Data Management (WP4) and Analysis and simulation tools (WP6).

The utility of MuG will be demonstrated through the development of **3 pilot projects**, which will have early access to VRE services, providing valuable feedback and thus helping to develop the data management, computational infrastructure and analysis modules.

1	2	3
Understanding the Chromatin reorganization during senescence	**Multi-resolution dynamic map of Yeast chromatin**	**Physical properties of DNA and protein-DNA complexes**

MuG provides a virtual research environment addressed to the growing worldwide 3D/4D genomics community: general genomics, structural genomics, molecular modelling and medical genomics.

MuG will make an impact on thousands of researchers worldwide, ultimately contributing to the fast emerging field of personalized medicine.

The MuG Virtual Research Environment targets the **3D/4D genomics community**, including both **academia** and **industry** end users:

Academia and research infrastructures	**Genomics and personalized medicine** research community: IPCG-PAWG, Blueprint, EGA, Genomics England, potential CoE on personalized medicine.
	Biosimulation and 3D Genomics community (direct beneficiaries include ABC, ExTASY, HEC-BioSim, potencial CoE on Biosimulation).
	HPC for Life Sciences initiatives (EESI2 and Bio HPC CoEs)
	Users of European Infrastructures for HPC/Big Data (EUDAT, PRACE, EGI)
	Genomics research community
	Computer Science research community
Life Sciences Industry	Sequencing instrument vendors
	Biotech/Pharma and Public Health

Significant effort will be devoted to **training and outreach activities** in order to reach the broadest possible number of stakeholders in both the public and the private sector.

2 PRINTABLE BROCHURE

The basic project information presented above has been compiled into a printable brochure. The brochure is available for download as a pdf from the MuG project website (http://www.multiscalegenomics.eu/docs/MuG_Factsheet.pdf)

EXPLORING GENOME
BEYOND SEQUENCE

3D and 4D genomics represent one of the greatest challenges for biology and biomedicine in the next decade.

Research efforts in 3D and 4D genomics are experiencing exponential growth. However, the current lack of standardization in analysis and simulation tools is threatening to become a bottleneck. MuG responds to the latest computational challenges of 3D/4D genomics by bringing this community closer to the HPC world and providing a suitable set of tools and infrastructure.

Understanding how genome is organized in the space and how this affects gene regulation would be instrumental to fully understand the time-dependent connection between genome and phenome.
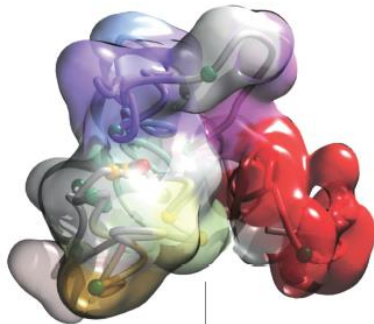
3D REPRESENTATION OF A HUMAN DNA FRAGMENT
(COPYRIGHT: MARC MARTÍ-RENOM, CNAG-CRG)

Multiscale Complex Genomics

Coordinator: Dr. Modesto Orozco (IRB Barcelona)
Start date: 1st November 2015
Project duration: 36 months
EC contribution: 2.961.163,00 €
Project website: www.multiscalegenomics.eu

Coordinator

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

Partners

Barcelona Supercomputing Center — Centro Nacional de Supercomputación · EMBL · cnag · cnrs · The University of Nottingham (UNITED KINGDOM · CHINA · MALAYSIA)

EXCELENCIA SEVERO OCHOA — IRB Barcelona, BSC and CNAG-CRG have been awarded the "Severo Ochoa center of excellence accreditation" by the Spanish government in recognition of their outstanding performance in the international research arena.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676556.

Multiscale Complex Genomics

BRINGING SCIENCE AND TECHNOLOGY CLOSER

The vision of the MuG project lays on the achievement of the following challenges:

- **To prevent a potential collapse** in 3D/4D genomics due to inability to manage the computational problems originated by massive sequencing experiments and simulations.
- **To nucleate the 3D/4D community** and bring it closer to the HPC and Big Data world.
- To provide an **integrated three-dimensional picture of the genome.**

JOINT RESEARCH: Development of a virtual HPC infrastructure
Couple virtual infraestructure to existing and future Computing/Data resources

SERVICE: Tools fitted to community needs.
Integrate software sevice (data management, visualization and analysis tools) for existing communities

NETWORK: Engage community (academic and industrial)
Develop pilot projects and explore industry liaisons for consultancy and commercial prototypes

KEY INNOVATIONS

A scalable computational infrastructure for multi-scale complex genomics, defining a new scheme for interaction between the 3D/4D genomics community and the HPC world

Modelling of chromatin re-organizations aimed at improving the knowledge on cell biology and predicting links to disease.

Connect 1D to 3D/4D genomics data, allowing for the first time to navigate across data from different sources and scales in a single "complex" model.

Visualization of complex genomic machinery with a 4D multiresolution picture.

HPC FACILITIES

BIOLOGISTS

METHODS DEVELOPERS

WORLD LEADING EXPERTS

A high-profile, multidisciplinary consortium:

IRB Barcelona (Modesto Orozco, Coordinator): Developer of the most used tools for analysis and simulation of nucleosome fibers. Expertise in developing tools for high-resolution DNA simulation; Analysis of DNA physical properties;

BSC-CNS-Life Sciences (Josep Ll. Gelpí) and Computer Sciences (Rosa M. Badia) departments: Supercomputing centre with large experience in supporting bio-HPC problems. Broad experience in modelling protein-protein and protein-nucleic acid complexes.

CNAG-CRG (Marc Martí-Renom): World-leading activity in developing approach for reconstructing chromatin structure from Hi-C data.

CNRS-IGH (Giacomo Cavalli): one of the most active groups in connecting chromatin structure with cell behaviour.

University of Nottingham (Charles Laughton): Expertise in developing tools for high-resolution DNA simulation and with large experience in analysis of DNA physical properties.

EMBL-EBI (Andrew Yates): World-leading group in genomics data management.

USER DRIVEN

The utility of MuG will be demonstrated through the development of **3 pilot projects** that will have early access to the MuG Virtual Research Environment services, providing valuable feedback and thus helping to develop the data management, computational infrastructure and analysis modules

1 Understanding the Chromatin reorganization during senescence

2 Multi-resolution dynamic map of Yeast chromatin

3 Physical properties of DNA and protein-DNA complexes

IMPACTS

MuG provides a virtual research environment addressed to the growing worldwide 3D/4D genomics community: general genomics, structural genomics, molecular modelling and medical genomics.

MuG will make an impact on **thousands of researchers** worldwide, ultimately contributing to the fast emerging field of personalized medicine.

The innovations developed by MuG will have a **strong impact on academia and industry alike.**

ACADEMIA AND RESEARCH INFRASTRUCTURES

Genomics and personalized medicine research community
Biosimulation and 3D Genomics community
HPC for Life Sciences initiatives
Users of European Infrastructures for HPC/Big Data
Genomics research community
Computer Science research community

LIFE SCIENCES INDUSTRY

Sequencing instrument vendors
Biotech/Pharma and Public Health