

MuGVRE. A virtual research environment for 3D/4D genomics

Laia Codó¹, Genís Bayarri², Javier Alvarez Cid-Fuentes¹, Javier Conejero¹, Adam Hospital², Romina Royo¹, Dmitry Repchevsky¹, Marco Pasi³, Athina Meletiou³, Mark D. McDowall⁴, Fatima Reham⁴, José A. Alcantara², Brian Jimenez-Garcia¹, Jürgen Walther², Ricard Illa², François Serra⁵, Michael Goodstadt⁵, David Castillo⁵, Satish Sati⁶, Diana Buitrago², Isabelle Brun-Heath², Juan Fernandez-Recio^{1,7}, Giacomo Cavalli⁶, Marc Marti-Renom^{5,8}, Andrew Yates⁴, Charles A. Laughton³, Rosa M. Badia¹, Modesto Orozco^{2,8}, Josep Ll. Gelpi^{1,8*}

1. Barcelona Supercomputing Center, Barcelona, Spain,
2. Institute for Research in Biomedicine, the Barcelona Institute of Science and Technology. Barcelona, Spain,
3. Sch. of Pharmacy and Centre for Biomolecular Sciences, Nottingham, UK,
4. European Molecular Biological Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
5. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain. Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
6. Institute of Human Genetics, CNRS, Univ. Montpellier, Montpellier, France.
7. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain,
8. Dept. of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain.

* To whom correspondence should be addressed. Tel: 34934034009; Fax: 34934021559;
Email: gelpi@ub.edu.

Marco Pasi, present address: Ecole Normale Supérieure de Cachan, Laboratory of Biology and Applied Pharmacology, Paris, Île-de-France, FR

ABSTRACT

Multiscale Genomics (MuG) Virtual Research Environment (MuGVRE) is a cloud-based computational infrastructure created to support the deployment of software tools addressing the various levels of analysis in 3D/4D genomics. Integrated tools tackle needs ranging from high computationally demanding applications (e.g. molecular dynamics simulations) to high-throughput data analysis applications (like the processing of next generation sequencing). The MuG Infrastructure is based on openNebula cloud systems implemented at the Institute for

research in Biomedicine, and the Barcelona Supercomputing Center, and has specific interfaces for users and developers. Interoperability of the tools included in MuGVRE is maintained through a rich set of metadata allowing the system to associate tools and data in a transparent manner. Execution scheduling is based in a traditional queueing system to handle demand peaks in applications of fixed needs, and an elastic and multi-scale programming model (pyCOMPSs, controlled by the PMES scheduler), for complex workflows requiring distributed or multi-scale executions schemes. MuGVRE is available at <https://vre.multiscalegenomics.eu> and documentation and general information at <https://www.multiscalegenomics.eu>. The infrastructure is open and freely accessible.

INTRODUCTION

While major milestones have been achieved in determining the sequence of DNA, understanding its 3D folding, the connection between chromatin structure and genome functionality and the links between changes in chromatin structure and pathology are still major challenges that are attracting a large research effort and have created a new area of knowledge: 3D/4D genomics. Opposite to traditional sequencing projects, 3D/4D genomics faces a major problem related to the diversity of data types and formats generated, the variety of analysis methods, and the multi-resolution nature of the navigation in a multi-scale data space. Data used in the field range from simple sequence (1D genomics), typically enriched by functional and structural annotations (2D Genomics) to contact maps, single or multiple structures (at a very wide range of resolutions) and images.

ENA (1), EGA (2) and Ensembl (3) are the most common sources of 1D genomic data in Europe, whose contents are visualized thanks to tools such as “Genome browsers” (4-8) that provide an integrated view of both sequence and annotations (2D genomics). Many research infrastructures deal with this level of data, Galaxy (9), being by far the most popular.

A second level of data includes nucleic acid and protein structures determined at the atomistic level by X-Ray crystallography or NMR spectroscopy and deposited in the Protein Data Bank (10). Analysis and visualization of data at this level have, in practice, little overlap with the 2D genomics level. This represents a major caveat for people interested in specific DNA-protein complexes. An additional type of data includes simulation trajectories, i.e. a set of structures defining the conformational ensemble of DNA (and associated proteins), which is typically obtained through atomistic or coarse-grained molecular dynamics (MD) simulations (11-17). Specific databases on DNA simulation data (18) or protein-DNA complexes (19, 20), and tools to perform flexibility analysis of nucleic acids like NAFlex (21) or 3DNA (22) are available. Few tools allow mapping 3D structures onto the genomic sequences (23-25).

A third level of data is represented by nucleosome and protein positioning studies (26-30). Raw data is obtained from sequencing analysis (MNase-seq, Chip-seq) and is usually limited to the annotation of specific binding sites. Data at this level is available from public repositories like ArrayExpress (31), where it is presented in a 1D manner without any connection to the 3D structure and flexibility of the chromatin fiber. At the upper end of the scale the main experimental strategies include FISH (32), and chromosome conformation capture (3C-like techniques (33)), which provide chimeric sequences representing interactions within the genome, from which structural insight can be derived after a complex manipulation of the data. Data visualization at this level is complex (34); several tools are already available (35-38), but they ignore any atomistic detail, or even the general description of the nucleosome string.

In summary, 3D (structure) and 4D (dynamics) genomics cover a large variety of data types and experimental and theoretical strategies. Tools and data repositories exist; but they are not integrated. As a result, research at the different levels is performed in isolation, and researchers almost ignore alternative views coming from different levels. The Multiscale Complex Genomics project (MuG, <http://www.multiscalegenomics.eu/MuG>) is one of the initiatives aiming to bridge the gap among the different levels of chromatin study and provide such an integrated view. Here, we present the MuG Virtual Research Environment (MuGVRE) infrastructure, designed to provide researchers with a single access point where data and tools covering the full spectrum, from sequence and atomistic data to chromosome capture results, can be used and combined to obtain a holistic picture of chromatin. MuGVRE is cloud based, allowing for an easy deployment and extension at the technical level. It is a base infrastructure where additional tools can be plugged-in to extend the functionality. The initial offer of tools already hosted, includes sequence, structure and Hi-C data analysis and visualizers; tools to handle nucleosome positioning, and tools for performing and analyzing simulation data from atomistic to coarse-grained levels. All the tools are accessible through an intuitive personal workbench, where most technical decisions are taken automatically by the system, allowing the user to concentrate on the scientific aspects of the analysis. We believe that MugVRE is the missing element to integrate the different views of physiological DNA.

INFRASTRUCTURE DESIGN AND COMPONENTS

The MuGVRE infrastructure has been designed to fulfil the following principles:

Technical:

1. Flexible structure, able to adapt to the specific needs of the analysis tools, both in terms of software requirements, and computational resources.

2. Software scheduler(s), able to manage analysis workflows, and computational resources in a transparent and adaptable manner. This should constitute an elastic infrastructure with automatic adaptation to user loads.
3. Multi-scale execution. Analysis workflows should be executed either at the cluster level, in HPC environments, or distributed infrastructures like EGI (<https://www.egi.eu>), and eventually in the forthcoming European Open Science Cloud (EOSC) ecosystem.

Usage:

4. Web-based access centered on the user workspace and complemented by full programmatic access using well-established interfaces.
5. Data should be kept private, through the appropriate Authentication and Authorization Infrastructure applied to all data transactions.

Supplementary Figure S1 shows a general schema of the computational infrastructure underlying MuGVRE.

MuGVRE Main components

Cloud deployments: MuGVRE infrastructure has been designed as a fully virtualized environment. At its present state, MuGVRE has been deployed in at the Starlife cloud infrastructure (<https://www.bsc.es/marenostrum/star-life>), at the Barcelona Supercomputing Center, using OpenNebula (<https://opennebula.org/>) and the KVM hypervisor (<https://www.linux-kvm.org>).

Process management: MuGVRE uses two complementary layouts for process management: i) Sun Grid Engine (SGE, <https://sourceforge.net/projects/gridscheduler/>), in combination with OneFlow (https://docs.opennebula.org/5.4/advanced_components/application_flow_and_auto-scaling/app_flow_use_cli.html), a component of the OpenNebula framework that allows managing Multi-VM applications and auto-scaling. SGE is used to manage applications where no complex workflows are necessary, requiring only to deploy additional workers on peaks of demand, ii) the COMPS Superscalar (COMPSs) programming model (and its python binding pyCOMPSs (39)), managed by the Programming Model Enactment Service (PMES) (40), which interacts with cloud infrastructures through Open Cloud Computing Interface (OCCL, <http://occi-wg.org/>) servers. PMES/pyCOMPSs are used to control complex workflows and distributed execution.

Database manager: Operational data and metadata regarding installed tools, public repositories, and user's files are maintained in a MongoDB database (<https://www.mongodb.com>). The MongoDB server also contains reference data as a copy of

Protein Data Bank (10), Uniprot (41), and BiGNASim (18) database. User's data is stored in a standard filesystem in its original format.

Authentication and authorization system: Data privacy is maintained using the authentication and authorization server Keycloak (<http://www.keycloak.org/>) to handle all internal communications and user access. Keycloak implements OpenID Connect which allows for Web access on the code authorization flow of OAuth2, and a token-based authentication for the REST services. For registered users, authentication schemes based on username/password, but also third-party identity providers (Google, LinkedIn, Elixir) are accepted.

See Supplementary Material Section 1 for additional information about MuGVRE software components.

USER WORKSPACE

User access

MuGVRE can be accessed without authentication. Users are granted a private workspace to hold data and analysis results. Data is maintained for one week after the last access and can be recovered during this period through a unique URL address. Users desiring a longer interaction with MuGVRE are advised to register to get a permanent workspace. In such cases the user space is maintained up to two months after the last access.

Personal workspace

The MuGVRE personal workspace is the central environment for user activity. It is based on a filesystem-based layout (see Figure 2), where both uploaded data and analysis results are available. Uploaded data should be annotated to specify data types and formats. This allows the MuGVRE workspace to offer an adapted toolkit for each file, including only compatible tools and visualizers (see Suppl. Material Section 2 for a description of the procedure, and Suppl. Table S3 for a description of accepted data types and formats). The user workspace has been laid out to provide an intuitive look-and-feel. The workspace itself is structured in projects, to keep data logically organized. Within each project the input data is distributed in folders: *Uploads* (uploaded data), and *Repository* (data obtained from public repositories). The remaining folders correspond to analysis operations (a new folder is generated for any new process started in the VRE). File lists can be filtered by any of the fields (name, format, data type, or project). Additionally, a tool-based filter is available to select only valid input files for the given tool.

Three interactive toolkits containing the following options are available:

- File toolkit: Download, rename, move, compress, delete files and folders, edit their metadata, rerun jobs.

- Visualization toolkit: Available visualizers for the specific file/s selection (based on their data type and format).
- Tools toolkit: Available tools for the specific file/s selection (based on their data type and format).

Tools based access

In addition to the data centric approach, more experienced user may prefer to access directly to the analysis operation. To this end, user may select the desired operation from an ontology, covering all the operations provided by the installed tools. The user should then select the desired tool and fill in the required parameters. At this point, a selection of input files filtered out from the user workspace are suggested according to the metadata accompanying both, user's data and tool input restrictions.

Applications and developer access

MuGVRE is designed as an infrastructure open to any application designed for the analysis of 3D/4D genomics data. Tools installed in MuGVRE should accept free, unrestricted usage. Guidelines are available for developers (see Suppl. Material, Section 2). Developers wishing to include their applications are granted access to a specific workspace to manage tool definitions and execution details, and to edit the tool's help pages, and perform execution tests. Also developers get access to usage statistics, execution logs, and job associated files and metadata. The current offer of tools is detailed in Suppl. Table S4. Several data visualizers are available (Suppl, Table S5). Figure 2 shows a representative set of the types of data that can be visualized in the infrastructure. JBrowse (5) is used for sequence data (see Figure 2A). In the figure, data from a Nucleosome Dynamics analysis of MNase-seq data is depicted (blue blocks for nucleosome calls, and coverage plot). The NGL visualizer (42) is used for 3D atomistic (see Figure 2B for a transcription factor-DNA complex) or coarse-grained structure data (output from Chromatin Dynamics, Figure 2D, depicting the structure of a fragment of chromatin fiber containing nucleosomes at positions shown in Fig. 2A). Finally, the TadKit visualizer (<http://3DGenomes.org/tadkit>, Figure 2C) allows the combination of 1D to 3D information taken from chromosome capture data.

USAGE

Users can populate the MuGVRE workspace in several ways:

- *Direct upload into the workspace*, using the HTTPS protocol
- *Create files using an embedded text editor*. This is intended for data or metadata of reduced size.
- *Upload from an External URL*: MuGVRE can access external sites to download. This is the recommended procedure to obtain data from public repositories, or import bulky

data, as the upload process becomes a batch job. HTTP and FTP protocols are accepted, also when user credentials are required.

- *From repository*: Data imported from the list of public repositories made browsable at the infrastructure (currently ArrayExpress (31) and BigNASim(18)).
- *From sample data*: selected input and output examples for the available tools are provided as help to start using the interface.

Files can be selected anywhere in the workspace, and added to the execution list, where a specific tool should be then selected. Alternatively, the user can select first an analysis tool from a list of available operations. Either selection mode opens a configuration screen where the user can assign data files (Suppl. Figure S7) to the appropriate input parameters, define additional settings, and launch the tool. Executions are performed in the background and do not require the user to keep the interactive session open. Results of the analysis are added to the workspace under a separate folder that contains the output files generated by the tool, log files, and a customized results page (Suppl. Figure S10). For a complete usage example, see Supplementary Material Section 3, where we show a series of screenshots of a session centered on the analysis of MNase-seq data from yeast Chr II on phases M and G2 of the cell cycle (data taken from (43)), analyzed with the Nucleosome Dynamics tool.

DISCUSSION

3D/4D Genomics is an emerging field originated from the unplanned aggregation of different disciplines which have developed their tools, and associated data types and formats, independently. This diversity is a major obstacle towards the generation of a complete picture of chromatin structure and dynamics. MuGVRE has been designed as an integration space. It follows the traditional concept of the personal workbench, already used in general genomic workbenches like Galaxy (9) or GenePattern (44), or spaces designed for simulation data analysis like NAFlex (21). In this kind of environment, data and tools are available and the user has the freedom to design his/her own analysis pipelines. MuGVRE has an initial offer of tools and visualizers covering all levels of resolution, from atomistic simulation to chromatin fiber simulation, or Hi-C data analysis. Tools and visualizers are offered in a single space where, for example, chromosome conformation capture or nucleosome positioning data can be visualized along with sequence annotations and the structures and binding modes of the transcription factors affecting the same DNA region. A strong commitment of MuGVRE design is to free the user from understanding the technical side of the infrastructure. With this aim, not only the computational layout is hidden, but also most technical decisions are taken automatically by the system. For example, we have designed a comprehensive ontology of data types and formats that are checked internally to configure the options offered to the user. The user can just select a tool, and the workspace selects those files that match its input requirements. Output from the

analyses are reusable following the same philosophy. As a result, the user can easily configure a pipeline taking only scientific decisions and not bothering about technicalities.

MuGVRE has been designed as a large and sustainable infrastructure, which relies now on the computational capabilities of the Barcelona Supercomputing Center (<http://www.bsc.es>), but technical decisions have been taken to assure the compatibility with other infrastructures like Elixir computational platforms (<http://elixir-europe.org>), EGI (<https://www.egi.eu>), or EUDAT (<https://eudat.eu>). The choice of a fully flexible cloud system, controlled with a multiscale software scheduler, and linked to HPC facilities, assures the usage of the optimal computational environment for each specific analysis.

MuGVRE is presented as an open platform with the aim of growing in functionality, since new tools can be easily incorporated by external developers. Hosting at MuGVRE can be an option for developers alternative to build a dedicated web site to run their tools.

MuGVRE was presented to the multiscale genomics community in November 2017, and has performed already over 6,500 analysis runs.

MuGVRE is a unique tool that aims to help researchers in the 3D/4D genomics field to gain an integrated view of discipline, sharing data among the diverse analysis levels and providing a complete and integrated view on DNA. We hope that MuGVRE will foster the development, deployment and use of new strategies for the analysis of the chromatin structure that were not envisioned simply because data was kept in separate silos.

AVAILABILITY

MuGVRE: <https://vre.multiscalegenomics.eu>

General information and documentation: <https://www.multiscalegenomics.eu>

ACKNOWLEDGEMENT

We are indebted to the entire MuG consortium and to the β -testers of the application for suggestions and comments.

FUNDING

This work has been supported by the Spanish MINECO [grants BIO2015-64802-R; BFU2015-61670-EXP, TIN2015-65316-P, TEC2015-67774-C2-2-R, BFU2013-47736-P and BFU2017-85926-P], the Catalan Government [grants 2014-SGR-134, 2014-SGR-1051]; the Instituto de Salud Carlos III-Instituto Nacional de Bioinformática [INB; grants PT13/0001/0019 and PT13/0001/0028]; France: The Fondation pour la Recherche Médicale, [grant

DEI20151234396], Laboratory of Excellence EpiGenMed; European Union, H2020 programme [grants Elixir-Excelerate: 676559; BioExcel: 674728 and MuG: 676566]. ERC Council [grants 291433, 609989]; IRB, CRG, and BSC are recipients of a Severo Ochoa Award of Excellence from MINECO (Government of Spain).

Funding for open access charge: European Union and Spanish Ministry of Science.

REFERENCES

1. Karsch-Mizrachi, I., Takagi, T., Cochrane, G. and Collaboration, I.N.S.D. (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 46, D48-D51.
2. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R. et al. (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*, 47, 692-695.
3. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, I.D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018) Ensembl 2018. *Nucleic Acids Res*, 46, D308-D314.
4. Barrios, D. and Prieto, C. (2017) D3GB: An Interactive Genome Browser for R, Python, and WordPress. *J Comput Biol*, 24, 447-449.
5. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. et al. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*, 17, 66.
6. Medina, I., Salavert, F., Sanchez, R., de Maria, A., Alonso, R., Escobar, P., Bleda, M. and Dopazo, J. (2013) Genome Maps, a new generation genome browser. *Nucleic Acids Res*, 41, W41-46.
7. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, 29, 24-26.
8. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., et al. (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res*. 46, D762-D769
9. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C. et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*, 44, W3-W10.
10. Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35, D301-D303.

11. Dans, P.D., Zeida, A., Machado, M.R., and Pantano, S. (2010) A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics. *J Chem Theory Comput.* 6, 1711-1725.
12. Dans, P.D., Perez, A., Faustino, I., Lavery, R. and Orozco, M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic acids res.* 40, 10668-10678.
13. Dans, P.D., Walther, J., Gómez, H., Orozco, M. (2016) Multiscale simulation of DNA. *Curr Opin Struct Biol.* 37, 29-45.
14. Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., III, Dans, P.D., Jayaram, B., Lankas, F., Laughton, C. et al. (2014) μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Research*, 42, 12272-12283.
15. Beveridge, D.L., Barreiro, G., Byun, K.S., Case, D.A., Cheatham, T.E., Dixit, S.B., et al. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(C(p)G) steps. *Biophys. J.* 87, 3799-7813.
16. Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T. III, et al. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* 38, 299-313.
17. Zakrzewska, K., Lavery, R. (2012) Towards a molecular view of transcriptional control. *Curr Opin Struct Biol.* 22, 160-167.
18. Hospital, A., Andrio, P., Cugnasco, C., Codo, L., Becerra, Y., Dans, P.D., Battistini, F., Torres, J., Goñi, R., Orozco, M. et al. (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res*, 44, D272-278.
19. Park, B., Kim, H. and Han, K. (2014) DBBP: database of binding pairs in protein-nucleic acid interactions. *BMC Bioinformatics*, 15 Suppl 15, S5.
20. Kirsanov, D.D., Zanegina, O.N., Aksianov, E.A., Spirin, S.A., Karyagina, A.S. and Alexeevski, A.V. (2013) NPIDB: Nucleic acid-Protein Interaction DataBase. *Nucleic Acids Res*, 41, D517-523.
21. Hospital, A., Faustino, I., Collepardo-Guevara, R., Gonzalez, C., Gelpi, J.L., Orozco, M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Research.* 41, W47-W55.
22. Lu, X.J., Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols.* 3, 1213-1227.
23. Duran, E., Djebali, S., Gonzalez, S., Flores, O., Maria Mercader, J., Guigo, R., Torrents, D., Soler-Lopez, M. and Orozco, M. (2013) Unravelling the hidden DNA

- structural/physical code provides novel insights on promoter location. *Nucleic Acids Research*, 41, 7220-7230.
24. Goni, J.R., Perez, A., Torrents, D. and Orozco, M. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol*, 8, R263.
 25. Goni, J.R., de la Cruz, X. and Orozco, M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res*, 32, 354-360.
 26. Chereji, R.V., Ramachandran, S., Bryson, T.D. and Henikoff, S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol*, 19, 19.
 27. Flores, O., Deniz, O., Soler-Lopez, M. and Orozco, M. (2014) Fuzziness and noise in nucleosomal architecture. *Nucleic Acids Research*, 42, 4934-4946.
 28. Sharma, S., Ding, F., Dokholyan, N.V. (2007) Multiscale modeling of nucleosome dynamics. *Biophys J*. 92, 1457-1470.
 29. Zuiddam, M., Everaers, R., Schiessel, H. (2017) Physics behind the mechanical nucleosome positioning code. *Phys Rev E*. 96, 052412.
 30. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316, 1497-1502.
 31. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T. et al. (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res*, 43, D1113-1116.
 32. Cremer, M., Grasser, F., Lanctôt, C., Müller, S., Neusser, M., Zinner, R., Solovei, I. and Cremer, T. (2008) Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol Biol*, 463, 205-239.
 33. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, 295, 1306-1311.
 34. Goodstadt, M. and Marti-Renom, M.A. (2017) Challenges for visualizing three-dimensional data in genomic browsers. *FEBS Lett*, 591, 2505-2519.
 35. Zhou, X., Lowdon, R.F., Li, D., Lawson, H.A., Madden, P.A., Costello, J.F. and Wang, T. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat Methods*, 10, 375-376.
 36. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S. and Aiden, E.L. (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*, 3, 99-101.
 37. Yang, D., Jang, I., Choi, J., Kim, M.S., Lee, A.J., Kim, H., Eom, J., Kim, D., Jung, I. and Lee, B. (2018) 3DIV: A 3D-genome Interaction Viewer and database. *Nucleic Acids Res*, 46, D52-D57.
 38. Tang, B., Li, F., Li, J., Zhao, W. and Zhang, Z. (2017) Delta: a new Web-based 3D genome visualization and analysis platform. *Bioinformatics*.

39. Tejedor, E., Becerra, Y., Alomar, G., Queralt, A., Badia, R.M., Torres, J., Cortes, T. and Labarta, J. (2017) PyCOMPSs: Parallel computational workflows in python. *Intl. J. High Perf. Comput. Appl.*, 31, 66-82.
40. Lordan, F., Tejedor, E., Ejarque, J., Rafanell, R., Alvarez, J., Marozzo, F., Lezzi, D., Sirvent, R., Talia, D. and Badia, R.M. (2013) ServiceSs: An Interoperable Programming Framework for the Cloud. *J. Grid. Comput.*, 12, 67-91.
41. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, 38, D142-148.
42. Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res*, 43, W576-579.
43. Deniz, Ö., Flores, O., Aldea, M., Soler-López, M. and Orozco, M. (2016) Nucleosome architecture throughout the cell cycle. *Sci Rep*, 6, 19729.
44. Reich, M., Tabor, T., Liefeld, T., Thorvaldsdóttir, H., Hill, B., Tamayo, P. and Mesirov, J.P. (2017) The GenePattern Notebook Environment. *Cell Syst*, 5, 149-151.e141.

LEGENDS TO FIGURES

Figure 1. Screenshot of MuGVRE personal workspace

Figure 2. Sample visualizations at MuGVRE.

A: Genome browser showing reference sequences, annotations, including a nucleosome positioning track (blue marks) (Jbrowse). B: Transcription Factor structure bound to cognate DNA fragment (NGL). C: HiC adjacent matrix, combined with genomics annotations, and a 3D model of chromatin structure (Tadkit), and D: Coarse-grained model of DNA fiber with several attached nucleosomes, from ChromatinDynamics (NGL).

SELECT FILE(S) Please select the file or files you want to use

Reload Workspace

Filter files by tool

Intuitive toolbox for data mngt.
analysis & visualization

File system layout

Rich set of
data types & formats

Search:

| File | File type | Data type | Project | Date | Size | Actions |
|------------------------|-----------|------------------------|------------------|------------------|----------|---------|
| File | All | All | All | | | |
| uploads | | | uploads | 2018/02/26 16:01 | 117.35 M | |
| cellcycleG2_chr11.bam | BAM | MNase-Seq | uploads | 2018/02/26 08:38 | 55.52 M | |
| cellcycleM_chr11.bam | BAM | MNase-Seq | uploads | 2018/02/26 08:38 | 27.62 M | |
| NR_G2_chr11.gff | GFF3 | Nucleosome positioning | uploads | 2018/02/26 16:01 | 615.64 K | |
| NR_M_chr11.gff | GFF3 | Nucleosome positioning | uploads | 2018/02/26 16:01 | 610.00 K | |
| SRR1232_1.fastq | FASTQ | HiC sequencing reads | uploads | 2018/02/26 09:32 | 16.51 M | |
| SRR1232_2.fastq | FASTQ | HiC sequencing reads | uploads | 2018/02/26 09:32 | 16.51 M | |
| chromDyn_40m_12N | | | chromDyn_40m_12N | 2018/02/26 16:01 | 2.66 M | |
| average_str.pdb | PDB | Chromatin 3D structure | chromDyn_40m_12N | 2018/02/26 16:01 | 177.66 K | |
| chromdyn_start_str.pdb | PDB | Chromatin 3D structure | chromDyn_40m_12N | 2018/02/26 16:01 | 177.66 K | |

