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

Deliverable 7.2: Report on the use of MuG VRE on integration of the whole yeast genome data

Lead beneficiary: Centre National de la Recherche Scientifique (CNRS)

Dissemination level: PUBLIC

Due date: 31/10/2018

Actual submission date: 16/11/2018

Copyright© 2015-2018 The partners of the MuG Consortium

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

Document history

Version	Contributor(s)	Partner	Date	Comments
0.1	Isabelle Brun Heath/Diana Buitrago/Jurgen Walther	IRB Barcelona	10/11/2018	First draft
1.0			15/11/2018	Final version. Approved by Supervisory Board.

Table of contents

Executive summary

This pilot project, led by IRB Barcelona, uses a yeast model to generate a full 3D picture of the genome, from nucleotide to chromosome. Indeed, visualization of the genome from single basepairs up to multiple chromosomes is a real challenge and yeast provides a perfect model thanks to its relatively small genome (12Mb).

1 INTRODUCTION

The main objectives of the pilot project 7.2 were (i) to generate multidimensional data from yeast to contribute to the development of new tools, both for the analysis and the visualisation of the data, and (ii) to play a key role in the testing of the tools and of the VRE in general.

Our lab is interested in studying the effect of DNA perturbation, either DNA damage or DNA methylation, on the 3D structure of the genome of the yeast *Saccharomyces cerevisiae*. In that context, we set up the Hi-C and Micro-C techniques in our lab and successfully generated interaction maps at different resolutions (Belton and Dekker 2015; Belton et al. 2012; Hsieh et al. 2016; Hsieh et al. 2015). Also, to correlate the effect observed at the structural level with the DNA perturbation, we also produced data at the 1D (CpG methylation and gene expression) and 2D (Nucleosome positioning and Protein binding) levels.

During this last part of the project, our efforts were mainly focused on using our data to contribute to the integration of the DNA and chromatin analysis tools into the VRE. However, we also generated new 3D interaction maps at higher resolution and we performed ChIP-seq experiments to characterize the effect we saw either at the expression or at the structural level.

2 DATA PRODUCTION

2.1 Effect of DNA methylation on nucleosome positioning and 3D chromatin structure

DNA methylation is one of the epigenetic marks most studied and its critical role in gene expression and cell differentiation has been clearly established (Suzuki and Bird 2008). However, there are many ways DNA methylation can regulate gene expression: (i) by preventing the binding of a specific transcription factor (TF), either directly, or indirectly by recruiting a Methylated DNA binding protein that will compete with the TF to access its target sequence or (ii) by acting on the chromatin conformation, for example through the displacement of nucleosome or the modification of DNA compaction. In order to study the effect of DNA methylation on the chromatin 3D structure, we artificially induced DNA methylation in the yeast genome, an organism normally unmethylated and in which there are no known homologues of the DNA Methyl Binding proteins. Using this system, we could test the effect of DNA methylation on gene expression, nucleosome positioning and 3D genome structure.

To map the methylated CpG at bp resolution, we performed Whole Genome Bisulfite Sequencing (WGBS). We also did Micrococcal Nuclease-sequencing (MNase-seq) to produce the nucleosome map and, finally, RNA-seq to follow the modification in gene expression and ChIP-seq for a couple of Histone marks (H3K4me1 and H3K4me3) as well as for a transcription factor that seemed to be responsible for the expression changes observed in our system. Finally, we generated Hi-C data to have a general view of the effect of DNA methylation on 3D genome organisation.

In total, during this project, we generated 2 datasets of WGBS, 4 of RNA-seq, 4 of MNase-seq and 12 of ChIP-seq, and we produced 4 Hi-C maps at 5kb resolution. Some of these data have already been submitted to Array express but the most recent ones are still being analysed and will be submitted by the end of the year.

2.2 Effect of DNA damage on 3D genome structure

The second project focuses on the effect of DNA damage on chromatin conformation and genome 3D structure. We used two different sources of DNA damage: Oxidative stress (OS) and UV, whose effects on DNA are very well documented. In addition, the group of N. Friedman studied 26 different histone marks in yeast under OS, and showed that OS induces changes in chromatin, suggesting that it could have repercussions on the 3D genome structure (Weiner et al. 2015).

To confirm that the conditions of treatment we applied were causing DNA damage, we checked if the DNA repair machinery genes were activated. Total RNA were extracted and sequenced using the stranded mRNA-seq protocol. We then performed MNase-seq to establish the nucleosome maps in each condition. Finally, we produced several types of Hi-C experiments. First, we generated Hi-C map at 5 kb resolution but the small size of the yeast genome and, more precisely, of the yeast genes (1-3kb in average) required a much higher resolution to allow us to correlate nucleosome positioning, gene expression and 3D structure. Therefore, we generated a new Hi-C map with a resolution of 500bp, and we also performed Micro-C experiments that allowed us to reach the nucleosome resolution.

To adopt its 3D structure, the genome requires structural proteins such as cohesin and condensin (Lazar-Stefanita et al. 2017). Therefore, in order to localize these proteins in our system, we performed ChIP-seq experiment in yeast treated or not by H_2O_2. In addition, we tested several histone marks known to be associated with DNA damage (i.e. H2Aph).

For this project, we generated 10 datasets of RNA-seq, 10 datasets of MNAse-seq, 16 datasets of MNase-seq and 11 datasets of Hi-C/Micro-C data. These data are still being analyzed but will be submitted to array express by the end of the year.

2.3 Imaging data / FISH

Amongst the different tasks of the pilot project 7.2 was the production of FISH (Fluorescence in Situ Hybridization) data. However, the small size of *S. cerevisiae* nucleus (2 to 3 μm^3) together with the progress of our projects made this task too difficult to complete. In addition, the pilot project 7.1 was in a far better position to fulfil this task, and therefore FISH data were instead produced in the context of the "Senescence" project.

3 DATA ANALYSIS

3.1 1D Genomic Analysis

3.1.1 Alignment

Mark McDowall – EMBL-EBI

The first step of any primary analysis of raw sequencing data analysis is read alignment, mapping reads contained in FASTQ files to a reference genome. Most of the sequencing analysis pipelines have an integrated aligner. However, some don't and it is therefore crucial to have mapping tools available in the VRE. So far, two of the most common aligners, Bowtie2 and BWA MEM are integrated in the VRE and a third one, Gem3, is planned to be integrated in the near future.

3.1.2 RNA-seq, and WGBS

Mark McDowall – EMBL-EBI pipelines

During this last year of the project, several pipelines have been integrated into the VRE to allow each biologist to perform RNA-seq and WGBS analysis. However, these pipelines are only starting to be fully functional now so we could not perform all the early analysis using the integrated version of the tools. On the other hand, our large set of data allowed us to contribute extensively to the integration of those tools for example in helping with the choice of output files that should be displayed and more importantly in testing the tools.

Our WGBS data were analysed using GemBS ((Merkel et al. 2018)) and the RNA-seq data were analysed using STAR for the alignment and RSEM for the quantification. The developers of GemBS are currently working on a way to integrate their tool in the VRE in the near future.

The tool currently available in the VRE to perform WGBS analysis uses Bowtie2 as the aligner and BS seeker 2 for the methylation calling while RNA-seq data are analysed with Kallisto. We've already performed several tests and they seem to perform correctly within the VRE.

3.2 2D Genomic Analysis

We consider here the second genomic dimension as the organisation of the DNA fibre in interaction with proteins such as histones to form the chromatin fibre, or transcription factors, structural proteins etc.

3.2.1 Nucleosome Dynamics

by Diana Buitrago and Ricard Illa (IRB Barcelona)

The first level of organisation of the chromatin fibre is achieved through the association of the DNA with a histone octamer to form the nucleosome. The position of the nucleosomes can be determined using Micrococcal Nuclease, an enzyme that specifically digests the linker DNA. The sequencing of the undigested fragment (MNase-seq) allows us to re-constitute the nucleosome map. Nucleosome Dynamics (ND) offers several tools dedicated to the analysis of MNase-seq data and providing nucleosome positions, Nucleosome Free Region (NFR) localisation, the classification of each Transcription Start Site (open or closed, -1 and +1 well positioned (W) or Fuzzy (F)), and the periodicity and the stiffness of the nucleosomes (Manuscript in preparation). ND also performs differential nucleosome positioning allowing comparisons between two different conditions. It is among the first set of tools that was integrated into the VRE and it was therefore used a lot in the analysis of our MNase-seq data.

❖ Example of ND analysis

Thanks to ND, we were able to detect changes in nucleosome positioning during the cell cycle, upon DNA methylation and after oxidative stress. Notably, we saw that the percentage of well-positioned nucleosomes decreased when the genome was methylated or when the cells were subjected to stress. Also, it appears that there are less promoters in the W-open-W configuration (2321 vs 2893) and more promoters in the F-closed-W (764 vs 537) configuration in thesamples after oxidative stress than in the control (fig. 1).

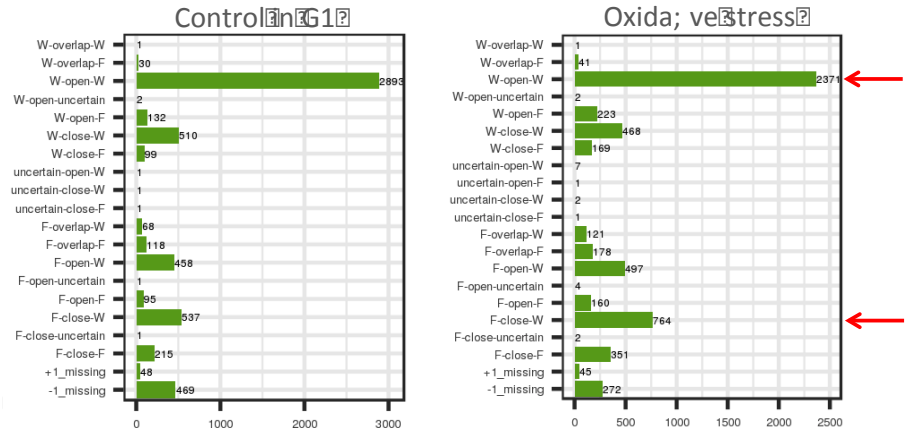

Figure 1 : Nucleosome position at promoters in cells before (A) and after (B) oxidative stress.

3.2.2 ChIP-seq

ChIP-seq is a technique largely used to map transcription factors, histone modifications and chromatin binding proteins. The most commonly used pipeline for ChiP-seq analysis uses the BWA aligner and MACS2 caller. This pipeline is now available in the VRE and we are now in the process of using it to analyse our latest ChIP-seq data.

❖ Example of ChiP-seq analysis in the case of the DNA methylation project

In the DNA methylation project, we observed that many genes specifically involved in meiosis were upregulated in the methylated cells. Looking more closely into the literature about those genes, we found out that they were all regulated by a transcription factor known as UME6, a subunit of the histone deacetylase complex rpd3. UME6 binds the sequence TCGGCCGT and normally acts as a repressor of meiotic gene expression. Therefore, we hypothesized that when its binding site gets methylated, UME6 is no longer able to bind DNA leading to the de-repression of the meiotic genes. To test this hypothesis, we performed UME6 ChIP-seq analysis. Preliminary results showed that depending on the target genes, UME6 binding can either be identical or enhanced in the methylated sample. These results still need to be confirmed but if this is the case, UME6 function as a repressor might need to be revisited.

We also tested two histone marks, H3K4me1 and H3K4me3 that are known to play a role in DNA methylation and confirmed that those marks are deposited equally in the methylated and unmethylated samples in our system.

3.3 3D genome structure and Modeling

3.3.1 Chromatin Dynamics

by Jürgen Walther (IRB Barcelona)

Chromatin Dynamics provides a user-friendly way to create individual 'beads-on-a-string' like representations of a chromatin fibre. The user decides the sequence of the linker DNA and where the nucleosomes are inserted which gives the user full control over designing his own individual chromatin fibre.

Chromatin fiber conformations strongly depend on the positions of its nucleosomes. It is common to mimic chromatin conformations of an organism according to average properties of linker DNA length and nucleosome free region length of the whole genome (Bascom, Kim, and Schlick 2017). This procedure however does not hold if the chromatin properties of a specific region within the genome have to be assessed. In a specific genomic region, nucleosome positioning is much more diverse than can be modelled just taking into account the average linker length and average size of a nucleosome free region. By using experimentally determined nucleosome positions of the region of interest as an input to our model, one can obtain realistic chromatin configurations. We used NucleR to determine nucleosome positions in a non-parametric manner (Flores and Orozco 2011). The nucleosome positions derived by NucleR are based on a population of cells, and simulating a fiber with the nucleosome positions derived by NucleR results in physical clashes and thus results in a non-realistic fiber configuration (Figure 5A). To overcome this issue, we developed an automatic way to obtain "*in-vivo*" like fiber configurations without any clashes by deconvoluting the MNase-seq signal. According to the parsimony principle we assume that the final signal in MNase-seq is a convolution of the signal of the minimum possible number of families of cells. A probability distribution consisting of a mix of Gaussian distributions where each single Gaussian represents a nucleosome detected by nucleR mimics the coverage. From this probability distribution, a certain number of physically realistic structures are sampled and clustered. The weight of each cluster is optimized so that the combination of all weighted clusters represents best the MNase-seq signal. The clusters with a weight above a given threshold are kept for further processing. The deconvolution process is done on regions of 3-5 Kb, depending on the number of nucleosomes in the segment. Larger fibers of up to 30 Kb can be reached by adding in a sequential way several 3-5 Kb segments.

3.3.2 TADbit and TADkit

by François Serra and David Castillo (CNAG-CRG) – WP3

Finally, we tested the effect of DNA methylation and DNA damage on chromatin and whole genome 3D structure. First, we used TADbit to produce the interaction maps which were then visualised with TADkit within the VRE (Serra et al. 2017). However, it was important for us to be able to navigate from the nucleosome map to the 3D genome structure and test the correlation between nucleosome positioning and chromatin folding at the level of the whole genome. Therefore, new functionalities were added into TADkit and are now available in the integrated version in the VRE.

- ❖ Interaction map from high resolution Hi-C experiment in the oxidative stress project

In order to test how oxidative stress affects the 3D genome organisation, we performed high-resolution Hi-C experiments (Hi-C 2.0) (Belaghzal, Dekker, and Gibcus 2017) using a 4-cutter restriction enzyme in order to get smaller fragments. As shown in figure 2, this method allowed us to reach a much higher resolution (up to 0.5kb) than the one we obtained with the original Hi-C protocol. Using this approach, we were able to see TAD-like structures, similar to the Chromosomal

Interaction Domain (CID) described using the Micro-C technique (Hsieh et al. 2015). Comparing the interaction map in cells before and after Oxidative stress (Fig 2C and D), the global structure of the genome does not seem to be changing upon stress. However, looking at a much higher resolution, some interactions appear to be changing upon oxidative stress (fig 2E and 2F).

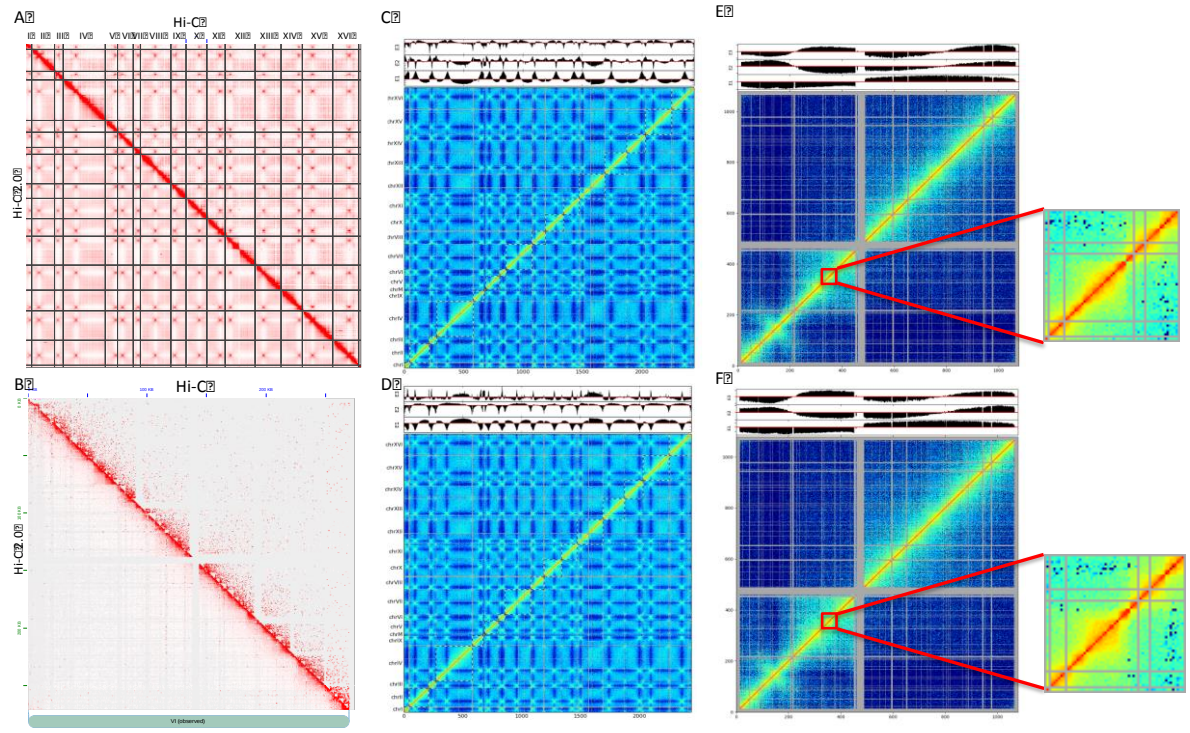

Figure 2: Comparison of Hi-C resolution obtained using either HindIII (6-cutter) or NdeI (4-cutter) restriction enzyme (A) Whole genome interaction map (B) interaction map in chromosome IV at 0.5kb resolution. High resolution Hi-C Interaction map from yeast cells synchronized in G1 phase before (C, E) and after (D, F) treatment with 10mM H_2O_2 during 30min. Whole genome interaction map (C and D) and chromosome XII at 1kb resolution (E and F).

In order to study the relation between interaction and Nucleosome positioning, we performed Micro-C experiments. However, due to the absence of defined restriction sites, the version of TADbit integrated in the VRE is not adapted to Micro-C analysis. However, this method is only very sparsely used and is only adapted to yeast, therefore, the relevance of adapting the VRE version of TADbit for Micro-C analysis is arguable.

3.3.3 Whole genome 3D model at bp resolution

by Diana Buitrago (IRB Barcelona)

We used Hi-C contact matrices (5Kb binning) to obtain an ensemble of 3D structures for yeast genome. The 3D models were obtained by transforming the Hi-C contact frequencies to spatial distance restraints using previously published relationship (Varoquaux et al. 2014) that connects the genomic distances in base-pairs with the contact frequencies from Hi-C experiments and the spatial distance between genomic loci. Afterwards, the computed spatial distances were applied as restraints (flat-well parabola potentials) during the simulations. Every chromosome was folded separately and then we applied inter-chromosomal restraints to join all the chromosomes and have a structure of yeast whole genome. We select an ensemble of the top structures that satisfy a given percentage of the restraints applied (typically 80%, Fig 3)

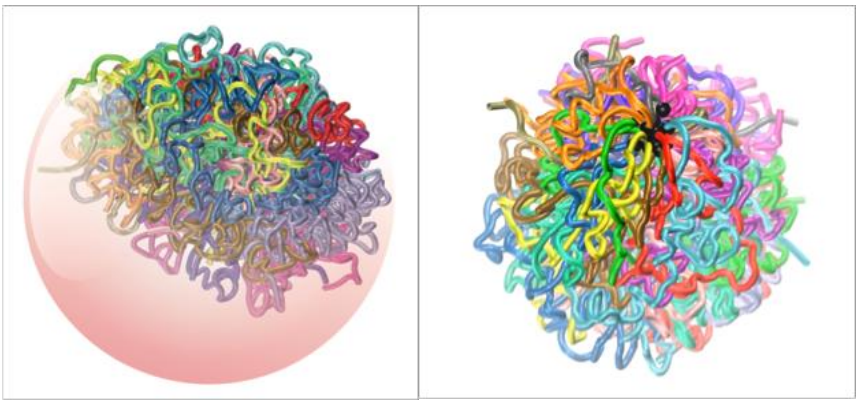

Figure 3 : S. cerevisiae whole genome 3D model. Each chromosome is represented in a different color. 40% of the nuclear volume (pink sphere) correspond to rDNA. The centromeres of all 16 chromosomes co-localise in one region of the nucleus.

We are now using this approach to model the whole genome structure in our different projects and see the effect of DNA methylation and oxidative stress on 3D genome organisation.

4 CONCLUSION

Considering the extensive variety and quantity of data produced, and the extensive VRE usage required for the data analysis, the pilot project 7.2 has shown its complete relevance for the setting up of the VRE. Also, it highlights the importance and the great advantage of having tools capable of analysing genomic data from the DNA sequence up to the 3D genome structure all in one environment. Even though we did not have time to complete the analysis of all our data and some tools have only been integrated recently, the pilot project 7.2 has benefited from the MuGVRE and we're convinced that this virtual environment will be extremely beneficial for the scientific community.

5 REFERENCES

Bascom, G. D., T. Kim, and T. Schlick. 2017. 'Kilobase Pair Chromatin Fiber Contacts Promoted by Living-System-Like DNA Linker Length Distributions and Nucleosome Depletion', *J Phys Chem B*, 121: 3882-94.

Belaghzal, H., J. Dekker, and J. H. Gibcus. 2017. 'Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation', *Methods*, 123: 56-65.

Belton, J. M., and J. Dekker. 2015. 'Hi-C in Budding Yeast', *Cold Spring Harb Protoc*, 2015: 649-61.

Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. 2012. 'Hi-C: a comprehensive technique to capture the conformation of genomes', *Methods*, 58: 268-76.

Flores, O., and M. Orozco. 2011. 'nucleR: a package for non-parametric nucleosome positioning', *Bioinformatics*, 27: 2149-50.

Hsieh, T. H., G. Fudenberg, A. Goloborodko, and O. J. Rando. 2016. 'Micro-C XL : Assaying chromosome conformation from the nucleosome to the entire genome', *Nature Methods*, 13: 1009-11.

Hsieh, T. H., A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando. 2015. 'Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C', *Cell*, 162: 108-19.

Lazar-Stefanita, L., V. F. Scolari, G. Mercy, H. Muller, T. M. Guerin, A. Thierry, J. Mozziconacci, and R. Koszul. 2017. 'Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle', *EMBO J*, 36: 2684-97.

Merkel, A., M. Fernandez-Callejo, E. Casals, S. Marco-Sola, R. Schuyler, I. G. Gut, and S. C. Heath. 2018. 'gemBS - high throughput processing for DNA methylation data from Bisulfite Sequencing', *Bioinformatics*.

Serra, F., D. Bau, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom. 2017. 'Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors', *PLoS Comput Biol*, 13: e1005665.

Suzuki, M. M., and A. Bird. 2008. 'DNA methylation landscapes: provocative insights from epigenomics', *Nat Rev Genet*, 9: 465-76.

Varoquaux, N., F. Ay, W. S. Noble, and J. P. Vert. 2014. 'A statistical approach for inferring the 3D structure of the genome', *Bioinformatics*, 30: i26-33.

Weiner, A., T. H. Hsieh, A. Appleboim, H. V. Chen, A. Rahat, I. Amit, O. J. Rando, and N. Friedman. 2015. 'High-resolution chromatin dynamics during a yeast stress response', *Mol Cell*, 58: 371-86.