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 4.2: Data Management Plan

Lead beneficiary: The European Bioinformatics Institute (EMBL-EBI)
Dissemination level: PUBLIC

Due date: 01/05/2016
Actual submission date: 10/05/2016

This project 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	Andrew Yates	EMBL-EBI	18/04/2016	First draft
0.2	Andrew Yates	EMBL-EBI	19/04/2016	Second draft
0.3	Josep Lluís Gelpí	BSC	26/04/2016	Third draft
0.4	Marc Marti-Renom	CNAG-CRG	06/05/2016	Edits integrated regarding 3C
1.0			09/05/2016	Approved by Technical and Supervisory Boards

Table of Contents

1 EXECUTIVE SUMMARY

The MuG consortium will release a number of data sets as part of its pilot projects to demonstrate the usefulness of the MuG platform. We will capture gene expression, the methylation of genomes, 3-dimensional structure of genomes, chromatin patterns and visualisation of genome structure. The consortium will submit these data to public archives to be made available after an up-to 12 month embargo period. Both raw and processed data will be submitted. Where no suitable archive for data currently exists (for example imaging data) the consortium elects to host the data itself until a point when a suitable archive can be identified. We will also use existing metadata standards and formats where available to maximise our compatibility with similar studies.

Following these steps will ensure that MuG consortium generated data will be publically accessible, discoverable through correct metadata handling, linkable by creating studies in these archives and ensure the usefulness of MuG data beyond our consortium.

2 INTRODUCTION

The MuG consortium is due to create a number of data sets as described in Work Package 7: Pilot Projects and data generated by external users as part of their use of our VRE. We are committed to not only maximising the usefulness of the data set to the consortium but also to other scientists. Ensuring the availability of comprehensive data sets is essential to both increasing the usefulness of the data and also to detail the usefulness of advances the MuG consortium will develop to integrate data ranging from gene expression analysis to the 3-dimensional folding of the genetic material.

Our data set management plan concerns itself with submitting both raw and processed data into public archives. Upon submission data sets will receive accessions making that data set identifiable within the bioinformatics domain. A large number of archives targeted are mirrored across the globe ensuring that this information is readily available to investigators. Where no suitable archive is available the consortium aims to provide alternative methods of making such data accessible and promote those links through appropriate channels.

We define a study as the creation of a work package data set from a consortium member.

All data generated by the consortium will be released using standard formats. Those that are domain specific are detailed in the annex section of this document.

3 DATA SET MANAGEMENT

3.1 Metadata Management and Standards

Study metadata can be coordinated by the BioStudies (https://www.ebi.ac.uk/biostudies) and BioSamples (https://www.ebi.ac.uk/biosamples/) resources. BioStudies allows the grouping of related data sets into a single accessioned record. Sample management will occur via BioSamples where sample information will be deposited and accessioned. These accessions will be reused on submission into the individual archives. We will follow recommendations made by MINSEQE when submitting sequencing data to archives (1).

3.2 Data Release Policy

After consulting materials generated by the Toronto International Data Release workshop we have elected to withhold data until publication since none of the data sets being generated here constitute whole genome reference samples (2). Data will be submitted to archive and then released after a maximum period of 12 months. Deviations from this will be indicated on a per data-set basis later in this document.

3.3 Sequencing Data

All DNA sequencing based analyses produce raw reads, which will be submitted to one of two archives based at EMBL-EBI. Reads will be submitted in standard formats e.g. FASTQ, BAM or CRAM. The choice of format depends on the type of data submitted by the consortium, which could be aligned or raw reads. ENA (http://www.ebi.ac.uk/ENA) provides archives for non-sensitive data and will be a primary archive for data generated by the MuG consortium. These submissions can be brokered by alternative archives including ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). Alignment will be performed against the most suitable version of the *Homo sapiens* reference genome. As of writing this will be either GRCh37 (hg19) or GRCh38 (hg38). Concerning *Saccharomyces cerevisiae*, alignment will be performed against the sequence of the reference strain S288C - Genome annotation Release R64-2-1 - with minor modifications to match the specific genotype of the studied strain.

Where data is identified as sensitive this will flow into EGA (http://www.ebi.ac.uk/EGA) where data access will be controlled by a DAC of suitable consortium members to control. Sensitive sequencing data is classified as identifiable human samples. No data at the time of writing to be produced by the consortium has been flagged as sensitive however this management plan will be updated accordingly as and when data is identified.

Processed data will flow into other archives as detailed below with BAM (3), CRAM (4) and the UCSC BigWig and BigBED formats (5) being the most prevalent form of data.

Experiment	Type	Format	Target Archive
RNA-seq	Expression analysis	BAM/CRAM and GTF	ArrayExpress
ChIP-seq	Predicted Peaks	BigWig and BigBED	ArrayExpress
Capture Sequencing (Hi-C)	Contact matrices	BED-like and raw text matrices	ArrayExpress
WGBS	Methylation values and coverage	BigWig and BigBED	ArrayExpress
MNase-seq	Predicted Peaks	BigWig and BED	ArrayExpress

3.4 FISH Imaging Data

No suitable public archive currently exists for the long-term storage of imaging data. Instead we suggest the use of a public FTP site hosted by the consortium and linking any data files back to a BioStudies record. Doing so means if a suitable archive is identified the images will be submitted to it and the BioStudies record updated accordingly.

3.5 Structural and simulation data

Atomic-resolution structural data and metadata will follow formats and recommendations by Protein Data Bank, hosted at the IRB Barcelona's structural data repository (http://mmb.irbbarcelona.org/pdb) and eventually mirrored in other suitable repositories for theoretical models. Simulation trajectories and the corresponding flexibility analysis results will be deposited under a MuG specific section in the BIGNASim nucleic acids simulation database (http://mmb.irbbarcelona.org/BIGNASim/ [6]). A specific policy for the data deposition is available at http://mmb.irbbarcelona.org/BIGNASim/help.php?id=submission. There is no specific standard to represent higher level (chromatin fibers, large DNA fragments, etc.) structure and simulation data. PDB-based formats are meant mainly for atomistic coordinates, but they can be easily extended to higher level structures as long as they can be represented as a set of spatial coordinates (In fact this is the general solution in coarse-grained simulations data). MuG will work with the 3D/4D genomics community to generate standards for representing higher-order structures of genomes. Indeed, currently we are adopting a JSON format adapted to be used with TADkit (http://3DGenomes.org/tadkit). In the meantime, data will be available from MuG repositories, in the most suitable format. All data will be linked to the appropriate BioStudies record.

3.6 Software and methods

All code and software will be developed on systems with regular backup strategies. Major and minor versions will be versioned controlled and carry an Apache 2.0 software license. Code will be documented, with appropriate help and tutorials and released publically.

3.7 Discoverability

In addition to the interfaces available from archives we host a web interface detailing (hosted alongside our project website at http://www.multiscalegenomics.eu/) to better organise data sets at a project level.

4 CONSORTIUM DATA SETS

4.1 Work Package 7.1: Senescence

Leads: IGH-CNRS & CNAG-CRG

Targets: These experiments will be run on non-identifiable human cell-lines WI38 human fetal lung fibroblasts. Samples will be registered with the BioSamples database.

4.1.1 RNA-seq

Description: Quantitative expression analysis of human cell-lines

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. GTF for expression.

Sharing: Data available from ArrayExpress.

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.1.2 Chromosome Conformation Capture (3C)-based Sequencing using Hi-C

Description: Using 3C-based sequencing (specifically Hi-C) to capture the 3D genetic material structure.

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Contact matrices submitted as a standard modified-BED format. These are detailed in Annex 7.2.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.1.3 WGBS

Description: Discovery of the methylation pattern of cell-lines.

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Methylation values and coverage submitted to ArrayExpress as BigWig and BigBED files.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.1.4 FISH

Description: Visualisation of the chromosome structure using fluorescent probes

Standards and metadata: Study submitted to BioStudies. Images held as TIFFs.

Sharing: Data available from FTP and linked from the BioStudies record.

Archiving: Held on FTP with backups onto tape available within the host institute. Will move to an archive once a suitable archive appears. Data will be held for publication.

4.2 Work Package 7.2: Whole Yeast Data

Leads: IRB Barcelona & CNAG-CRG

Targets: These experiments will be run on *Saccharomyces cerevisiae* strain using Illumina's standard sequencing protocols.

4.2.1 RNA-seq

Description: Quantitative expression analysis of *Saccharomyces cerevisiae* strains

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. GTF for expression.

Sharing: Data available from ArrayExpress.

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.2.2 Chromosome Conformation Capture (3C)-based using Hi-C

Description: Using 3C-based sequencing (specifically Hi-C) to capture the 3D genetic material structure.

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Contact matrices submitted as a standard modified-BED format. These are detailed in Annex 7.2.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.2.3 WGBS

Description: Discovery of the methylation pattern of cell-lines.

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Methylation values and coverage submitted to ArrayExpress as BigWig and BigBED files.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.2.4 ChIP-seq

Description: Prediction of transcription factors and chromatin-associated proteins on a genome

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Peaks submitted as BigWig and BigBED files.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.2.5 MNase-seq

Description: Protocol for finding the location nucleosomes; a unit of genomic organisation

Standards and metadata: MINSEQE guidelines for submission to ArrayExpress. Peaks submitted BigWig and BED files of possible locations.

Sharing: Data available from ArrayExpress

Archiving: Submitted to ArrayExpress. Data will be held for publication.

4.3 Work Package 7.3: Transcript Factor Binding and DNA Bending

Leads: University of Nottingham, IRB Barcelona, BSC

Targets: These experiments will be run to study several DNA-binding transcription factors of diverse origin, bound to a wide variety of different DNA fragments.

4.3.1 3D Structural models of Protein-DNA Interactions

Description: Theoretical models of Protein-DNA complexes obtained by docking procedures and eventually refined through simulation.

Standards and metadata: Data will follow Protein Data Bank (PDB) recommendations and formats

Sharing: Structures and analysis data will be available through MuG portal.

Archiving: Provided at Model Archive (http://www.modelarchive.org/); a part of the Protein Model Portal. Data will be held for publication.

4.3.2 MD Simulations

Description: Trajectories and metadata for molecular dynamics trajectories of nucleic acids and protein-nucleic acids complexes

Standards and metadata: Simulation metadata and nucleic acids description will follow data ontologies as described in [6]

Sharing: Data and analysis available at BigNASim database (http://mmb.irbbarcelona.org/BIGNASim/)

Archiving: Provided at IRB Barcelona-BSC long-term archive. Data will be held for publication.

4.3.3 DNA Flexibility

Description: Results of flexibility analysis done on nucleic acids structures and simulation trajectories

Standards and metadata: Flexibility results will follow data ontologies as described in [6]

Sharing: Data and analysis available at BigNASim database (http://mmb.irbbarcelona.org/BIGNASim/)

Archiving: Provided at IRB Barcelona-BSC long-term archive. Data will be held for publication.

5 CONCLUSIONS

Extensive use of archives hosted at EMBL-EBI and IRB Barcelona will ensure the long-term viability and availability of data sets generated by the consortium. When no suitable archive is present we believe making the data available is important and make every effort to ensure that data is promoted and linked correctly. Good use of existing standards will ensure that MuG generated data will be compatible with other similar studies. We also have a desire to make data available as soon as is feasible whilst still retaining the rights of the data generators to publish.

6 REFERENCES

1. MINSEQE:Minimum Information about a high-throughput Nucleotide SeQuencing Experiment - a proposal for standards in functional genomic data reporting (2012) http://fged.org/site_media/pdf/MINSEQE_1.0.pdf

2. Toronto International Data Release Workshop Authors. Prepublication data sharing (2009) Nature 461, 168-170

3. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

4. Cochrane G, Cook CE and Birney E (2012) The future of DNA sequence archiving. GigaScience 2012 1:2

5. W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs and D. Karolchik (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics (2010) 26 (17): 2204-2207

6. Hospital A, Andrio P, Cugnasco C, Codo L, Becerra Y, Dans PD, 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 Jan 4;44(D1):D272-8. doi: 10.1093/nar/gkv1301.

7 ANNEXES

7.1 Abbreviations

DAC: Data access committee

EGA: European Genome-phenome Archive

ENA: European Nucleotide Archive

MINSEQE: Minimum Information about a high-throughput Sequencing Experiment

SRA: Sequence read archive

VRE: Virtual Research Environment

WGBS: Whole Genome Bisulphite Sequencing