Multiscale Complex Genomics

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 4.1 – Data types, Processing and Data Model Specification

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

Dissemination level: PUBLIC

Due date: 01/05/2016

Actual submission date: 18/05/2016

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 Mark McDowall Daniel Zerbino	EMBL-EBI	28/04/2016	First draft
0.2	Juan Fernández Recio Josep Lluis Gelpí Isabelle Brun-Heath Andrew Yates Giacomo Cavalli Saitish Sati Marco Pasi	BSC IRB Barcelona EMBL-EBI IGH-CNRS University of Nottingham	05/05/2016	Second draft. Expansion of sequencing protocols and atomistic data.
1.0	-	-	18/05/2016	Final version approved by technical and supervisory boards.

Table of Contents

1 EXECUTIVE SUMMARY

We present a document describing the methods required to process and integrate the diverse data sets MuG will handle. Well characterised, standard processing methods and tools exist for RNA-seq, ChiP-seq and WGBS sequencing experiments. There also exists well defined rules for metadata and ontologies as promoted by the IHEC consortium including EFO and OBI. We also nominate reference data sets to ensure comparability between data sets. Hi-C, due to it being an emerging technology, lacks a single coherent method of analysis to create contact matrices. However we outline the current best available methods capable of processing this data. Challenges also exist for making sense of microscopy data sets such as FISH. Existing work in *Schizosaccharomyces pombe* has shown a possible method to reduce the representation of such data. Finally we show two possible levels of data integration at the traditional genome level and at the 3D level and how this will feed into the MuG standard.

2 INTRODUCTION

Our goal is to integrate multiple cutting edge data sets spanning multiple types of analysis and the creation of a usable virtual research environment for consortium members and external users. Providing a suite of standard analysis programs will allow MuG to quickly integrate new data sets into the platform and maintain comparable and consistent analysis. Here we present an evaluation of the best in class analysis techniques for each type of data MuG will process. Where no such technique exists we instead present the best available analyses.

2.1 Data Types Available in MuG

In M2 BSC and EMBL-EBI conducted a survey of data types and formats available in the MuG consortium. Its results are detailed in Table 1.

Data	Source	Format	Pilot WPs
Genomic sequences	INSDC	FASTA	7.1, 7.2
Sequence annotations	Ensembl, SGE	GFF, BED, Wig	7.1, 7.2
Nucleosome positioning	Experiment/Prediction	GFF, Wig, FASTQ, BAM	7. 2
Nucl Acids 3D Structures	PDB	PDB	7.3
Protein-Nucl Acids complex 3D Structures	PDB, simulations	PDB	7.3
Nucl Acids MD Atomistic Trajectories	Simulations	Various	7.3
Nucl Acids MD CG Trajectories	Simulations	Various	7.3
Nucl Acids flexibility properties	Simulations	Various	7.3
RNA-seq	Experiment	FASTQ	7.1, 7.2
ChiP-seq	Experiment	FASTQ	7.1, 7.2
WGBS	Experiment	FASTQ	7.1, 7.2
Hi-C	Experiment	FASTQ	7.1, 7.2
Hi-C contact matrices	Computational	Wash-U Pairwise Interactions, HDF5	7.1, 7.2
Hi-C contact coverage in close/far-cis, trans	Computational	Wig	7.1, 7.2
Hi-C contact coverage	Computational	Wig	7.1, 7.2
Hi-C differential contacts	Computational	TSV	7.1, 7.2
Hi-C scaling Factor	Computational	Wig	7.1, 7.2

Hi-C contact peak	Computational	TSV	7.1, 7.2
Hi-C TADs	Computational	BED	7.1, 7.2
HiC directionality index	Computational	Text	7.1, 7.2
TADs Data	Computational	Text	7.1, 7.2
Compartments data	Computational	Text	7.1, 7.2
Ensemble of chromatin 3D structures	Computational	JSON	7.1, 7.2
FISH	Experiment	TIFF	7.2

Table 1: The data types and data formats to be produced by the MuG consortium

3 PROCESSING WORKFLOWS

3.1 Describing Data Sets

In addition to the existing IHEC recommendations for samples and protocols they also make recommendations for the ontologies to use for describing epigenome metadata. These are detailed in Table 2. MuG will reuse these ontologies to define its own data sets and add them as and when terms do not exist. At the moment however MuG is working with well known epigenomic data types and so the need for new terms should be low. We also see ontologies such as the EFO extensible enough to handle the new terms required to describe our imaging challenges.

Ontology	Domain	Data Type
Experimental Factor Ontology	Sample Ontology	Cell Lines
Cell Ontology	Sample Ontology	Primary Cells
Uberon	Sample Ontology	Primary Tissue
NCI Metathesaurus	Disease Ontology	Disease
Ontology for Biomedical Investigations	Experimental Ontology	Assays and Platforms

Table 2: The recommended set of ontologies as defined by the IHEC consortium.

For describing Nucleic Acids structure and simulations data, including protein-DNA complexes, MuG will reuse and eventually extend the data ontology described in [1].

3.2 Sequencing Processing

Many of the data sets MuG will draw from are derived from DNA sequencing. Work package 7 has two projects where their primary source of annotation is DNA sequencing including RNA-seq, ChIP-seq, Hi-C and WGBS. A standard flow of analysis is detailed in Figure 1. Each protocol assigns its own criteria to all phases including the strictness/looseness of the genome mapping procedure. The current dominant sequence aligner is bwa [2] and forms the basis of many pipelines. Re-alignment and re-processing is a common procedure and normally carried out to normalise data across diverse producers and to enable comparisons between existing data including the ENCODE [3] and Roadmap Epigenomics projects [4]. In addition to the major project producers additional groups provide re-analysis of the epigenetic data sets including Ensembl's regulatory build [5].

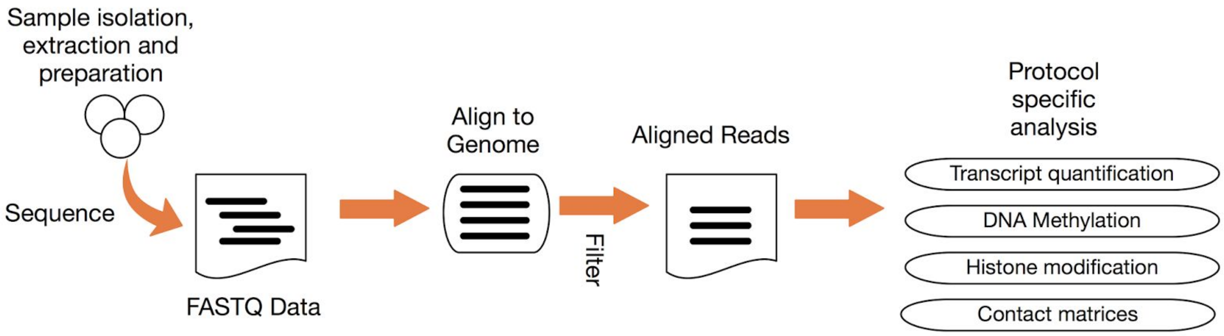

Figure 1: Diagram to show the common flow of DNA sequencing experiments from raw data to final product. All data must move through a process of alignment to a genome and filtering low quality reads. High quality data then moves into the analysis phase of the protocol.

Of the four protocols mentioned RNA-seq, ChIP-seq and WGBS are all mature methods and have well characterised methodologies. The International Human Epigenome Consortium (IHEC) [6] has outlined a set of best practices and filtering that should be applied to any experiment being conducted on human derived samples. These standards have been adopted into other consortia including the Functional Annotation of Animal Genomes (FAANG) [7] showing the usefulness and wider applicability of IHEC beyond the human genome. Our aim is to reuse those protocols and recommendations for analysis.

3.2.1 RNA-seq Processing

Whilst a mature area of analysis no clear consensus has been reached as to the best way to analysis RNA-seq data [8]. However we do highlight two common used strategies within the community to quantify gene expression. The first is based on traditional alignment based strategies, normally using splice aware aligners and transcript quantifiers such as Tophat [9] and Cufflinks [10]. The GENCODE gene set is regarded as a high quality source of human gene annotation and extensively used in the ENCODE project [11]. Differential gene expression can be calculated against reference data sets and then used to highlight genes expressed at different levels in the sequenced samples. Many different competing programs have been developed for alignment and quantification but the protocol is well characterised. The Gene Expression Atlas group at EMBL-EBI have developed a RNA-seq analysis pipeline, iRAP, that is capable of integrating multiple tools whilst retaining a consistent interface for analysis [12]. iRAP is available as a command line tool and as a Docker image enabling fast deployment in multiple environments.

More recently a new set of quantification tools have been developed and concentrate on reducing the computational load. Salmon [13], Sailfish [14] and Kallisto [15] all perform lightweight alignment processes avoiding whole genome alignment to reduce the overall computational load by an order of magnitude. Whilst not the dominant method of analysis these tools are being quickly adopted by the community at large.

In all cases the final data sets generated are expression levels of genes mapped to stable identifiers against a tissue/cell type/experimental condition.

3.2.2 ChiP-seq Processing

ChiP-seq allows us to find regions of the genome where DNA-associated proteins have bound or nucleosome proteins have been modified. Once sequencing reads have been mapped to the genome the focus switches to enriched regions, or peaks. Multiple algorithms exist to perform this function [16,17]. Peaks can be further filtered using the ENCODE Irreproducibility Discovery Rate (IDR) framework.

3.2.3 WGBS Processing

WGBS is the gold standard assay for measuring DNA methylation (modification of cytosine nucleotides by the addition of a methyl group, often but not exclusively in CpG context), on the whole genome at base level resolution. Genomic DNA is treated with sodium bisulfite that converts cytosine to uracil with high efficiency (with typically >98% of cytosines being converted), while 5-methylcytosine is only converted at low efficiency (<5% converted). This different behaviour of cytosine and 5-methylcytosine allows prediction of the methylation state of a given cytosine. The dominant analysis software is Bis-SNP [18] which allows the detection of sequence variants (SNVs) and prediction of DNA methylation from the same WGBS experiment. This is important because otherwise SNVs can give false evidence of differential methylation between samples.

3.2.4 Mnase-seq Processing

Mnase-seq attempts to assay the nucleosome population by digesting unprotected DNA. It indirectly assays chromatin accessibility by exposing areas of the genome already containing regulatory elements or nucleosomes. As with ChiP-seq processing Mnase-seq is a question of mapping reads to the genome and enriched areas. Again this is to find peaks of sequence reads and therefore the locations of the nucleosomes. Those areas devoid of reads will be the regions with accessible chromatin. Similar algorithms for ChiP-seq can be reused for Mnase-seq.

3.3 Chromosome Conformation Capture Analysis

Capture sequencing is a major focus of the MuG consortium in particular Hi-C [19]. Hi-C captures the interactions that occur between regions of chromatin by covalently linking them, digesting them using a restriction enzyme, followed by tagging, pull down of tagged fragments and high throughput sequencing [20]. Once sequenced it is then possible to map these sequences (fragments) back to the genome using standard alignment protocols followed by filtering and bias correction [21]. Representing a way to find the 3D folded conformation of chromosomes and the methods for generating the data sets are well understood. However interpretation of the data is still an open problem with many choosing to use end to end analysis pipelines such as Juicer and the bioconductor package diffHic. MuG is fortunate to have an analysis method for creating contact matrices and detecting TAD (topologically associating domains) developed by CNAG in TADbit. TADbit currently chooses to store the resulting matrices as plain text. The generated matrices are sparsely populated since bins will have no significant interactions with the vast majority of the genome. HDF5 provides a number of efficient sparse array storage solutions and also fast retrieval methods to reconstruct the interaction matrices. Called TADs can be stored in TADbit's native format or as called boundaries on the reference genome. The selected method will depend solely on the future use of the data.

3.4 Microscopy/Image Data

As noted in our previous report "Deliverable 3.1: A critical evaluation of the problems on data structure the browser has to solve" MuG must handle data across a wide range of scales from nm (unpacked DNA) to µm (whole cells) with 10^3 jumps between each scale. In many cases our post-processed data naturally fits into each individual scale. The report noted issues with FISH images and their ability to integrate into our platform. Our challenge is to bring representations of all data into one of two domains; mapping to the 3D structures to be generated or mapping back to the reference genome assembly coordinate space.

Recent work performed by Justin O'Sullivan's lab has concentrated on the discovery of the nuclear structure in *Schizosaccharomyces pombe*. It was discovered that there is a correlation between chromosomal structure and positioning in the G1, G2 and M phase of cell development and gene expression within *S. pombe [22]*. Further work has used a 2D representation of the *S.pombe* nucleus (as shown in Figure 2) to highlight the locations of replication origins in both late and early firing genes [23]. This has significantly reduced the data dimension and can define the chromosomes to have a probability of existing within a part of the nucleus.

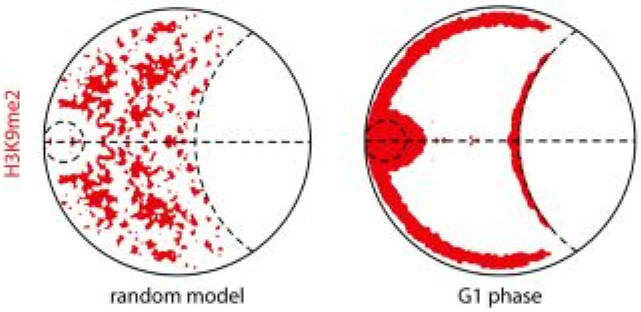

Figure 2: A modified figure of the 2D representation of the S. pombe nucleus taken from Pinchugina et al. 1000 genome structures were generated by polymer modelling with data sets enriched for histone modifications or proteins. The location of chromosome 1 (indicated in red) as predicted by a random model versus the actual position in G1 phase. Loci enriched for H3K9me2 were found at the nuclear periphery.

By defining nucleus as shells it is possible to attach a probabilistic location to a genomic loci enabling 3D spatial information to be integrated with traditional genomic annotation i.e. the 1D level. The attachment of genomic regions will mean the platform can be capable of answering complex 3D positional questions of multiple loci across timepoints in a manner not currently possible. In conjunction with this information we will annotate 3D models with the same information to ensure the data is represented at each logical level. Both levels of annotation require the decomposition of FISH data from raw images into values mapped against a predictable reference so as to allow comparisons between experiments.

3.5 Setup and processing of simulation data

3.5.1 Atomistic simulations

Atomic-detail 3D structural data will be used in MuG to study flexibility and dynamics of nucleic acids and protein-nucleic acids complexes. In particular, pilot project 7.3 will use this data to analyse the sequence-specificity of DNA-binding transcription factors (TF). Atomistic molecular dynamics simulations technique will be used to generate trajectories from 3D structures of protein-DNA complexes. A standard flow of MD setup and analysis is detailed in Figure 3. The initial 3D structures can be predicted from sequence (see next section), or retrieved from the Protein Data Bank. A setup process needs to be performed (using for example high-throughput simulation tools as NAFlex [24]) before running the MD simulation. Once the trajectory is generated, a large number of analyses can be performed to gain insight into the structural determinants of TF specificity and the dynamics of the complexes. In particular, the flexibility properties of DNA, and how they are affected by TF binding, will be studied by calculating stiffness constants as a function of the Curves+ helical parameters [25].

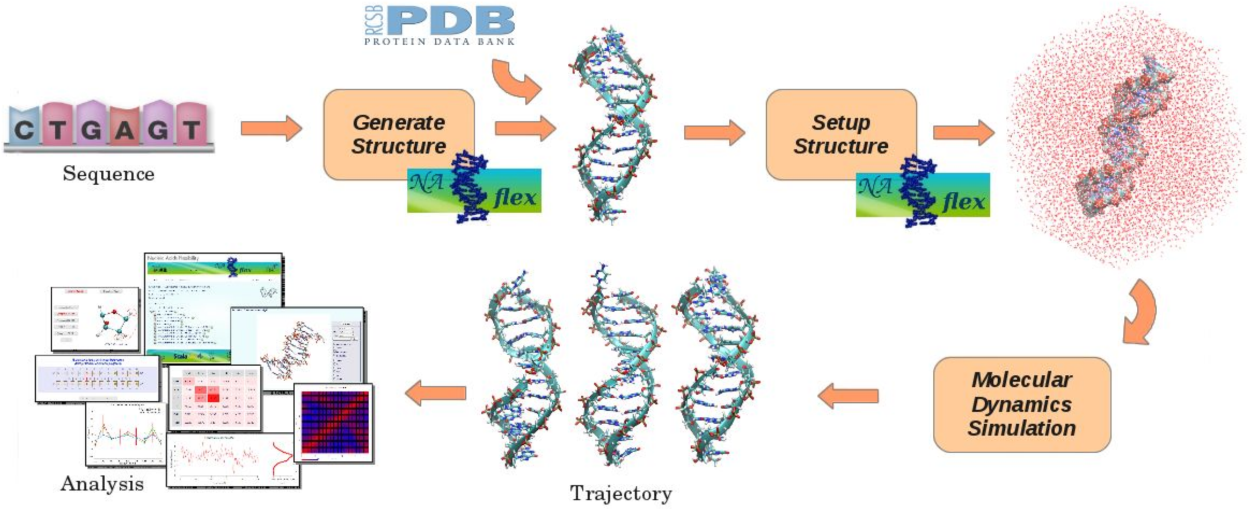

Figure 3: Diagram to show the common flow of DNA structure MD setup and analysis. Starting from a sequence or a structure, an initial setup is performed before the actual simulation run, and a final set of analyses are computed using the generated trajectory.

3.5.2 Atomistic protein-protein and protein-nucleic acids 3D structural models

Very often, the 3D structure of protein-protein and protein-nucleic acids complexes involved in the transcriptional factor complexes, as those studied in Pilot project 7.3, will not be available. In these cases, structural models will be generated using the appropriate tools that will be integrated in the

MuG VRE (see Task 6.1 of WP6), as follows. When needed, protein atomistic models will be built by using widely accepted homology modeling tools, like Modeller [26] and i-Tasser [27]. When possible, protein-protein complexes will be modelled based on available templates, or alternatively by ab initio docking [28]. The structure of DNA will be predicted from the sequence using information derived from long MD simulations of naked DNA [29], and refined, when required, using further specific MD simulations. The resulting individual protein and DNA structural models will be the input for the structural modeling of protein-DNA complexes. For a few cases, we will be able to inherit protein-DNA general orientation from homologous templates. However, for the majority of cases, we will build atomistic models of protein-DNA complexes by ab initio docking. Existing software for the identification of potential nucleic acids binding sites will be used to locate the DNA molecule on the surface of the protein. Finally, the assembly of multimolecular transcription factors will require specialized docking tools, with ad-hoc restraints from experimental data like SAXS [30], or from the models built at coarser-grained (mesoscopic) scale. The format of all the atomistic models that will be generated will follow PDB standards. Existing repositories of atomistic theoretical models can be used to deposit the generated data.

4 THE MULTI-SCALE DATA MODEL

4.1 Stable Reference Data Sets

Due to the nature of our data it is essential to define stable reference data sets. When considering the aforementioned analyses two types of reference data naturally select themselves; genomes and gene sets. The first is to provide a consistent reference coordinate system to map alignments to and the second provides a context to build value added resources upon e.g. gene expression levels. These references are detailed in Table 3.

Species	Assembly	Gene Set	Recommended for Analysis
Homo sapiens	GRCh37	GENCODE19	No (not the live reference)
Homo sapiens	GRCh38	GENCODE24	Yes
Saccharomyces cerevisiae	R64-2-1	2015 Annotation	Yes

Table 3: Available major assemblies

4.2 Defining the Multi-Scale Model and Database

Considering the levels of data involved in MuG no one single model can expose the domain specific complexities ranging from those changes at the base pair level to chromosome position within a nucleus. The first is to build efficient mappers capable of translating between the major coordinate spaces i.e. from the genomic level to the 3D space. Due to the nature of the data sets involved the genome provides a useful stable backbone to map a large proportion of data to. A data set can be represented at the genomic level so long as a data value can be mapped back to a genomic loci. This also has the advantage that the reference genome coordinates do not assume a 3D positioning and therefore elements mapped against it can be viewed as stable in their location. Some annotations require additional information and cannot be annotated alone e.g. Hi-C derived contact matrices require two genomic locations to be linked to each other resulting in a bidirectional relationship on the genome. Our second solution is to nominate distinct levels to annotate data against. The linear genome is one such level; as are the 3D models of chromatin. This has the effect of denormalising all data held by MuG for the sake of faster retrieval.

No single database can handle the multiple scales of data being described here. Each level brings with it new challenges and optimisations required for efficient storage, fast access and effective delivery to analysis tools. For example we previously described contact matrices are sparsely populated and therefore any solution that does not model this will suffer from an explosion of data storage costs. Instead we suggest a single database engine that can mediate to multiple types of back-end storage each optimised for a particular data type. This not only allows us to use the best format for storage but also to be flexible in future solutions for data storage. We can also prototype using faster NoSQL technologies including MongoDB until more efficient solutions become available. In this direction, atomistic structure and simulation data is being managed at IRB Barcelona's Data repository using

NoSQL technologies (Cassandra and MongoDB). MuG will base its development in extending the knowledge obtained in building such repository as well the experience of other partners, particularly EMBL-EBI, in the management of large scale data.

5 CONCLUSION

Whilst many data types to be processed by MuG have well defined analysis protocols some still lack sufficient maturity. In the case of the former providing tools via the VRE will enable external researchers to quickly integrate into the MuG environment and will also allow our partners to quickly push data through to the eventual database stores for further analysis and visualisation. Key to this second target is the ensure the data we store do not lose accuracy but also can provide data at a level where the interfaces can remain responsive over a modest internet connection. The MuG consortium and VRE must remain adaptable in the face of future analysis techniques and be open to integrating these new technologies as and when required. Our description of the current state of affairs in RNA-seq gene transcription analysis is a key example of a domain where there is a potential for competing analysis standards. MuG has begun the steps towards a multi-scale data model by embracing the benefits of each level and enabling the transfer of annotation between the different scales.

6 REFERENCES

1. Hospital A, Andrio P, Cugnasco C, Codo L, Becerra Y, Dans PD, et al. BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. Nucleic Acids Res. 2016;44: D272–8.

2. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv:1303.3997v1 [q-bio.GN]. 2013. Available: http://arxiv.org/abs/1303.3997

3. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, et al. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. 2011;39: D871–5.

4. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. 2015;518: 314–316.

5. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015;16: 56.

6. GmbH E. Welcome to IHEC · IHEC [Internet]. [cited 2 May 2016]. Available: http://ihec-epigenomes.org/

7. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16: 57.

8. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13.

9. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25: 1105–1111.

10. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28: 511–515.

11. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–1774.

12. Fonseca NA, Petryszak R, Marioni J, Brazma A. iRAP - an integrated RNA-seq Analysis Pipeline [Internet]. 2014 Jun. doi:10.1101/005991

13. Patro R, Duggal G, Kingsford C. Accurate, fast, and model-aware transcript expression quantification with Salmon [Internet]. 2015 Jun. doi:10.1101/021592

14. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32: 462–464.

15. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; doi:10.1038/nbt.3519

16. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9: e1003326.

17. Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei C-L, et al. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010;26: 1199–1204.

18. Liu Y, Yaping L, Siegmund KD, Laird PW, Berman BP. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012;13: R61.

19. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326: 289–293.

20. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58: 268–276.

21. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72: 65–75.

22. Grand RS, Pichugina T, Gehlen LR, Jones MB, Tsai P, Allison JR, et al. Chromosome conformation maps in fission yeast reveal cell cycle dependent sub nuclear structure. Nucleic Acids Res. 2014;42: 12585–12599.

23. Pichugina T, Sugawara T, Kaykov A, Schierding W, Masuda K, Uewaki J, et al. A diffusion model for the coordination of DNA replication in Schizosaccharomyces pombe. Sci Rep. 2016;6: 18757.

24. Hospital A, Faustino I, Collepardo-Guevara R, González C, Gelpí JL, Orozco M. NAFlex: a web server for the study of nucleic acid flexibility. Nucleic Acids Res. 2013;41: W47–55.

25. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K. Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res. 2009;37: 5917–5929.

26. Webb B, Benjamin W, Andrej S. Comparative Protein Structure Modeling Using MODELLER. Current Protocols in Bioinformatics. 2014. pp. 5.6.1–5.6.32.

27. Yang J, Jianyi Y, Renxiang Y, Ambrish R, Dong X, Jonathan P, et al. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2014;12: 7–8.

28. Cheng TM-K, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins. 2007;68: 503–515.

29. Pasi M, Maddocks JH, Beveridge D, Bishop TC, Case DA, Cheatham T 3rd, et al. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res. 2014;42: 12272–12283.

30. Jiménez-García B, Pons C, Svergun DI, Bernadó P, Fernández-Recio J. pyDockSAXS: protein-protein complex structure by SAXS and computational docking. Nucleic Acids Res.

2015;43: W356–61.

7 APPENDIX

7.1 Abbreviations

- FAANG - Functional Annotation of Animal Genomes
- IHEC - International Human Epigenome Consortium
- RRBS - Reduced Representation Bisulfite Sequencing
- TAD - Topologically associating domains
- VRE - Virtual Research Environment
- WGBS - Whole Genome Bisulfite Sequencing

7.2 Commonly Used Analysis Programs

Program	Data Type	Program Type	Analysis
BWA	Sequencing Reads	Aligner	Alignment to the genome
Bowtie	Sequencing Reads	Aligner	Alignment to the genome
GEM	Sequencing Reads	Aligner	Alignment to the genome
Tophat	RNA-seq reads	Mapper	Splice junction mapper
Cufflinks	Mapped RNA-seq reads	RNA-seq quantification	Recreates transcripts from RNA-seq reads and estimates expression levels
Kallisto	RNA-seq reads	RNA-seq quantification	Estimates transcript expression levels
iRAP	RNA-seq reads	Meta-pipeline	Pluggable pipeline for estimating transcript expression
MACSv2	ChiP-seq alignments	Peak caller	Calling peaks of significant alignment
SWEMBL	ChiP-seq alignments	Peak caller	Calling peaks of significant alignment
SPP	ChiP-seq alignments	Peak caller	Calling peaks of significant alignment
CCAT	ChiP-seq alignments	Peak caller	Calling peaks of significant alignment
ENCODE Irreproducibility Discovery Rate	Peaks	Peak quality caller	Measures consistency between replicates
Bis-SNP	WGBS reads and RRBS reads	Methylation detection	Finding methylated reads on the genome
TADbit	Hi-C reads	Hi-C analysis	Alignment, filtering, binning and TAD detection
diffHic	Aligned Hi-C reads	Hi-C analysis	Differential analysis of Hi-C data including binning, filtering and TAD detection

Juicer	Hi-C reads	Hi-C analysis	Hi-C contact matrices generation and feature annotation