Analysis of NGS Data on the Transcriptional Regulation[1]

Semyon K. Kolmykov[1,2,3], Ivan S. Evshin[1,2], Fedor A. Kolpakov[1,2]

[1] Institute of Computational Technologies SB RAS, Novosibirsk, Russian Federation
[2] BIOSOFT.RU, LLC, Novosibirsk, Russian Federation
[3] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russian Federation

Abstract. The GTRD database (http://gtrd.biouml.org) contains over 30,000 uniformly processed NGS experiments on transcriptional regulation (ChIP-seq, ChIP-exo, DNase-seq, ATAC-seq, MNase-seq and FAIRE-seq). To process these types of data, pipelines have been developed for the eGrid distributed computing management system and the BioUML platform.

Keywords: GTRD; BioUML; NGS; transcriptional regulation; ChIP-seq; ChIP-exo; DNase-seq; ATAC-seq; FAIRE-seq.

1 Introduction

The GTRD database (Gene Transcription Factors Database, http://gtrd.biouml.org) [1] contains information on the main components of transcriptional regulation: transcription factor binding sites and open chromatin regions. These data have been experimentally identified in various cell types and tissues using high-throughput methods: ChIP-seq and DNase-seq, respectively. Nowadays, GTRD database is the largest database in terms of size of uniformly annotated and processed ChIP-seq experiments on TFBSs identification (http://wiki.biouml.org/index.php/GTRD_comparison). The Statistics section on the GTRD start page provides detailed information on the number of analyzed data for each type of NGS experiment.

For the further development of the GTRD database, pipelines of processing NGS data of various types of experiments were built:

- ChIP-seq - mapping of transcription factor binding sites (TFBSs) and various types of modifications of histones (HM) on the reference genome;
- ChIP-exo - TFBSs mapping;
- DNase-seq, ATAC-seq, MNase-seq and FAIRE-seq - mapping of open chromatin regions on the reference genome, as well as localization of individual nucleosomes.

To process these types of NGS experiments, corresponding pipelines were developed for:

- eGrid distributed computing management systems. This system was designed to control the distributed processing of NGS data included in the GTRD on a cluster of 12 servers (each - 2 Intel® Xeon® CPU X5650 processors, 32-48GB RAM). The pipelines described below were implemented in the form of programs written in Java language with a set of software packages necessary for NGS data processing.
- The BioUML platform - web-platform for the analysis of biomedical data (https://ict.biouml.org). BioUML includes a wide range of capabilities, including access to databases with experimental data, tools for a formalized description of the structure and functioning of biological systems, as well as tools for their visualization, modeling, selection of parameters and analysis [2]. In this case, the developed scenarios are implemented as pipelines (workflows) and, unlike eGrid, are available to all users of this platform. Each user has the ability to change the parameters of the analyzes used in their own copy of the workflow. Also, due to the graphical representation of workflows and the available functionality of the workflow editor, users have the ability to both change the structure and build their own workflows for data processing.

2 Materials and Methods

In developing these scenarios, we followed the recommendations of the ENCODE project [3]. In the developed pipelines, the first steps of data processing are similar. At first, an initial quality analysis of raw data is performed using the FastQC software package (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) [4]. If necessary, adapter sequences are deleted using Trimmomatic package [5]. Alignment of the raw reads to the reference genome is performed using bowtie2 (bowtie2 --seed 0) [6]. Subsequent filtering of the obtained alignment (MAPQ ≥ 10) and sorting

by coordinate value are performed using the samtools software package: “samtools view -bq 10” and “samtools sort”, respectively. If data is represented by a library of double-end readings, then PCR duplicates are removed by Picard MarkDuplicates (https://broadinstitute.github.io/picard) [7]. The next step of processing is searching for genome regions enriched with aligned reads (peak calling). To this date, more than 30 peak calling algorithms have been published [8]. Since there is a difference in the distribution of aligned reads on the reference genome for different types of NGS data, the efficiency of peak calling depends on both the selected peakcaller and a set of its initial parameters. Therefore, this stage is specific for each type of NGS experiments under consideration.

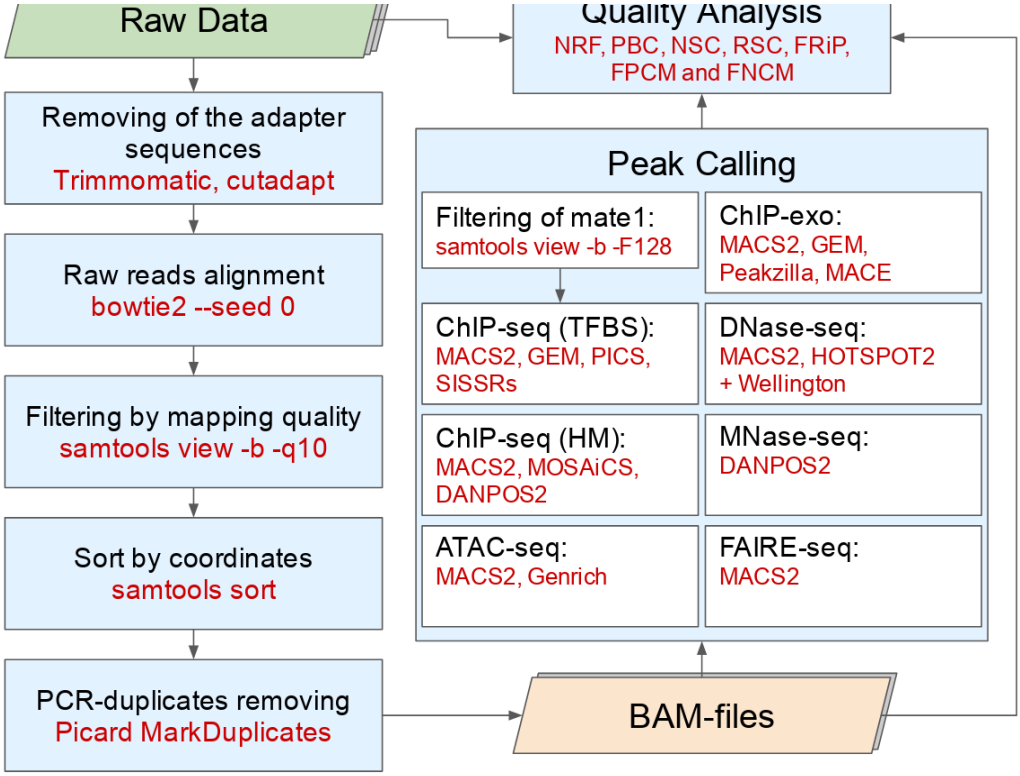

Figure 1. General pipeline of NGS data processing.

The ChIP-seq method is used for the whole genome searching for transcription factor binding sites (TFBSs) and various types of histone modifications (HM). When searching for TFBSs in ChIP-seq experiments represented by a library of double-end reads, only the first mate of aligned read pairs (mate-1; “samtools view -F128”) is selected for subsequent analysis. TFBSs are identified using the following set of peakcallers: MACS2 [9], GEM [10], SISSRS [11] and PICS [12]. It is worth noting that when using MACS2, the Phantompeakqualtools package [13] is used for modeling the shift size and calculation of the fragment length.

In the case of processing ChIP-seq experiments on various types of HMs identifying, a set of initial parameters of the software used depends on the type of HMs (see Table 1). For wide marks, MACS2 starts up with the following parameters: “macs2 callpeak --broad --broad-cutoff 0.1”, and the default parameters are used to search for narrow marks. MOSAiCS method for narrow marks and MOSAiCS-HMM method for broad peaks are also used for HMs identification [14].

Table 1. Types of histone modifications (https://www.encodeproject.org/chip-seq/histone/#histone)

To map nucleosomes on a reference genome based on processing data from MNase-seq experiments, the DANPOS2 software package (“danpos.py dpos”) is used [15].

A group of methods, consisting of DNase-seq, ATAC-seq, and FAIRE-seq, is used for mapping of open chromatin regions on a reference genome, thereby identifying potential regulatory regions. MACS2 and Hotspot2 (https://github.com/Altius/hotspot2) are used for identification of regions with high sensitivity to cleavage by endonuclease DNase I [16]. Due to the differences in preparation of the libraries, for single-cut DNase-seq experiments, MACS2 was used with the initial parameters: “–nomodel –shift -100 –extsize 200”; for other cases, the default parameters were used. To map open chromatin regions based on ATAC-seq or FAIRE-seq experiments, MACS2 is also used, but in the vast majority of cases (when the data is represented by a pair-reading library), this method starts with the parameters “--keep-dup all -f BAMPE”, and in the rest with “--keep-dup all –shift 100 –extsize 200”. Also, Genrich (https://github.com/jsh58/Genrich) [17] is used to process data from ATAC-seq experiments. The final step in the

processing of DNase-seq experiments is the identification of the DNase I footprints (putative areas of protein-DNA interaction; small regions of DNase-seq profiles with reduced sensitivity to cleavage by DNase I and location in regions with high activity of this enzyme). This step is performed using Wellington [18].

Analysis of the quality of NGS data begins with an assessment of the quality of the raw reads using the FastQC program. At the next stage, the alignment quality of the initial reads is estimated. In addition to calculating basic alignment statistics (samtools flagstat), the complexity of the library (NRF, PBC1 and PBC2) and the signal-to-noise ratio (based on cross-correlation (NSC and RSC) estimated using Phantompeakqualtools package [13]) are analysed. Upon completion of the peak calling stage, the percentage of reads that fall into the obtained TFBSs (FRiP) is calculated. In the case of processing ChIP-seq experiments, the data processing process was completed by evaluating FPCM and FNCM values [19]. In the future, it is planned to expand the set of methods for quality analysis of NGS data.

3 Results

As a part of the GTRD database development project, the scenarios for NGS data processing integrated into the eGrid distributed computing management system were used in preparation of the current release of the database. The information on the current release statistics is available on GTRD starting page (http://gtrd.biouml.org).

Furthermore, the software packages used for processing NGS data were installed on the BioUML platform using the Galaxy interface. To combine this set of programs into pipelines for processing various types of NGS experiments, we used workflow edition tools available on BioUML platform. Thus, 7 scenarios for processing NGS experiment data were built: “ChIP-seq TF pipeline”, “ChIP-seq HM pipeline”, “ChIP-exo pipeline”, “DNase-seq pipeline”, “ATAC-seq pipeline”, “MNase-seq pipeline” and “FAIRE-seq pipeline”. It is worth noting that it is possible for users to change both the initial parameters of the software packages used in workflow and the structure of the pipelines themselves.

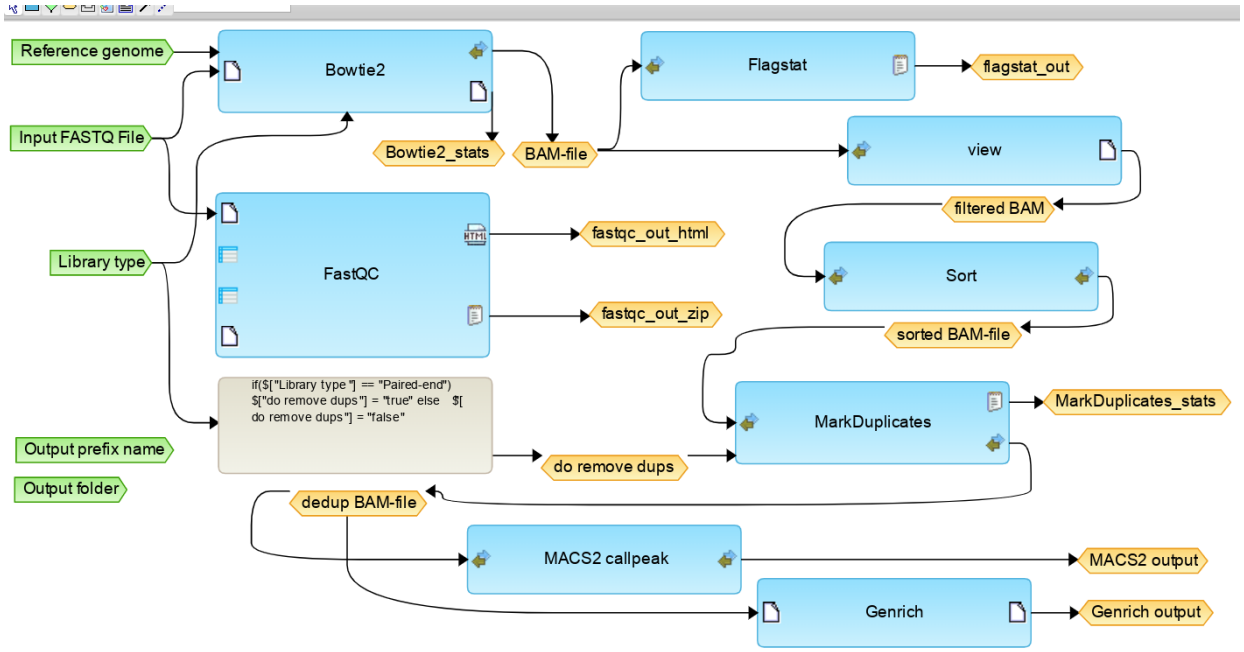

Figure 2. Diagram representation of “ATAC-seq pipeline” on BioUML platform. Green blocks - initial parameters of workflow; blue blocks - analyses in use; orange blocks - files generated by analyses in use.

4 Conclusion

Thus, we compiled 7 scenarios for processing different types of NGS data: ChIP-seq TF, ChIP-seq HM, ChIP-exo, DNase-seq, ATAC-seq, MNase-seq and FAIRE-seq. The developed data processing scenarios were used in preparation of the current release of database of regulatory elements (GTRD). Also, these scenarios were implemented as workflows on BioUML platform.

Acknowledgements. This work was supported by the Russian Science Foundation, grant number: 19-14-00295.

References

[1] Yevshin I. et al. GTRD: a database on gene transcription regulation—2019 update //Nucleic acids research. – 2018. – Т. 47. – №. D1. – С. D100-D105.

[2] Kolpakov F. et al. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data //Nucleic acids research. – 2019. – Т. 47. – №. W1. – С. W225-W233.

[3] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome //Nature. – 2012. – Т. 489. – №. 7414. – С. 57.

[4] Andrews S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[5] Bolger A. M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data // Bioinformatics. – 2014. – Т. 30. – №. 15. – С. 2114-2120.

[6] Langmead B., Salzberg S. L. Fast gapped-read alignment with Bowtie 2 // Nature methods. – 2012. – Т. 9. – №. 4. – С. 357.

[7] Broad Institute. Picard Toolkit: Java command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. https://github.com/broadinstitute/picard.

[8] Thomas R. et al. Features that define the best ChIP-seq peak calling algorithms //Briefings in bioinformatics. – 2017. – Т. 18. – №. 3. – С. 441-450

[9] Zhang Y. et al. Model-based analysis of ChIP-Seq (MACS) // Genome biology. – 2008. – Т. 9. – №. 9. – С. R137.

[10] Guo Y., Mahony S., Gifford D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints //PLoS computational biology. – 2012. – Т. 8. – №. 8. – С. e1002638.

[11] Narlikar L., Jothi R. ChIP-Seq data analysis: identification of Protein–DNA binding sites with SISSRs peak-finder //Next Generation Microarray Bioinformatics. – Humana Press, 2012. – С. 305-322.

[12] Zhang X. et al. PICS: probabilistic inference for ChIP-seq //Biometrics. – 2011. – Т. 67. – №. 1. – С. 151-163.

[13] Landt S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia // Genome research. – 2012. – Т. 22. – №. 9. – С. 1813-1831.

[14] Chung D., Zhang Q., Keles S. MOSAiCS-HMM: A model-based approach for detecting regions of histone modifications from ChIP-seq data //Statistical Analysis of Next Generation Sequencing Data. – Springer, Cham, 2014. – С. 277-295.

[15] Chen K. et al. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing //Genome research. – 2013. – Т. 23. – №. 2. – С. 341-351.

[16] Rynes E. et al. Hotspot2: Program for identifying genomic regions with statistically significant "hotspots," or enrichments, of cleavage activity in DNase-seq experiments. https://github.com/Altius/hotspot2

[17] Genrich: a peak-caller for genomic enrichment assays. https://github.com/jsh58/Genrich

[18] Piper J. et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data // Nucleic acids research. – 2013. – Т. 41. – №. 21. – С. e201-e201.

[19] Kolmykov S. K. et al. Population size estimation for quality control of ChIP-Seq datasets //PloS one. – 2019. – Т. 14. – №. 8.