NGS Data Analysis with Unipro UGENE

Olga Golosova, Yuriy Vaskin, German Grekhov, Yuliya Algaer

UNIPRO, Lavrentieva ave. 6/1, Novosibirsk, Russia, ugene@unipro.ru

Andrei Gabrielian, Vijayaraj Nagarajan, Andrew J. Oler, Mariam Quiñones, Alex Rosenthal

Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD, USA

Next-generation sequencing (NGS), also called high-throughput sequencing, technologies have revolutionised the study of genomics and molecular biology. They allow sequencing DNA and RNA more quickly than the previously used Sanger sequencing, and, with the lower cost, that decreases year after year [1]. Therefore, the technologies have become a preferable choice and de facto standard for many modern researchers. This implies that the researchers face the need to analyze data generated by the next-generation sequencing facilities, such as Illumina, Ion Torrent, and others. That is where Unipro UGENE NGS framework comes in play.

Unipro UGENE [2] is a desktop multiplatform open-source software package, integrating dozens of widely used bioinformatics tools and providing a rich graphical interface for various biological objects (DNA/RNA/protein sequences, multiple alignments, 3D protein structures, phylogenetic trees, sequencing reads assemblies, etc.) and associated tools. Moreover, the UGENE toolkit allows one to automate common tasks by creating workflows with the integrated tools using the UGENE Workflow Designer.

Analysis of NGS data with UGENE may start, for example, from quality control of raw NGS data, received from a sequencing facility, with the integrated FastQC [3] tool. However, alternatively a researcher may use one of the sample workflows for processing of raw NGS data, consisting of such steps as filtration of sequencing reads by CASAVA header (for Illumina platform), cutting of adapter sequences, trimming of the sequencing reads, mapping of the sequencing reads to a reference genome, additional filtration with SAMtools, removing of PCR duplicates, and others.

The selection of the raw NGS data processing sample workflow and, therefore, the type of processing steps and the output, produced by the workflow, depend on the type of NGS experiment the researcher performed.

For example, if this is a common DNA-Seq experiment, a popular tool BWA-MEM is used by default for the mapping step of the workflow. The output will be a sorted and indexed BAM file with sequencing reads aligned to the reference genome. The BAM file can be used for further analysis with other sample workflows. For example, if a diploid individual was sequenced, the researcher may use a sample "Call Variants with SAMtools" workflow [4], which will output SNPs and short INDELs for the data. The variants can be additionally annotated with the SnpEff [5] tool. This tool also reports the predicted effects of the variants (for example, amino acid changes).

When an RNA-Seq experiment is performed, only transcriptome parts of genome are covered by sequencing reads, so a dedicated tool for sequencing reads alignment must be used, that takes into account mapping of a single sequencing read to multiple genome areas. That is why the sample workflow for processing raw RNA-Seq data uses the TopHat tool for the mapping step of the workflow. To perform differential gene and transcript expression analysis of the RNA-Seq data in UGENE a researcher can follow a standard Tuxedo protocol [6, 7].

For a ChIP-Seq experiment, investigating DNA-protein interactions, the sequencing reads cover only such DNA regions that are linked by the protein, so the raw ChIP-Seq data processing workflow outputs a BED file with the list of the regions, although the BAM file with aligned sequencing reads is also available, if required. UGENE integrates a complex Cistrome workflow [6, 8] for analysis of the data.

Thus, the described sample workflows are ready-to-use right after the UGENE installation. More advanced users can use individual workflow blocks to create their own workflows. For example, the mapping to a reference step may be replaced by a *de novo* assembly step with the SPAdes tool [9].

UGENE has a rich graphical interface to simplify the use of the described features by biologists. For example, the Workflow Designer infrastructure allows configuring workflows parameters using wizards, results of workflows execution are stored in dashboards, providing details about the input parameters used and the output results obtained [6]. Using the UGENE Assembly Browser researches can navigate assembled sequencing reads, apply different highlighting modes, export consensus and coverage, and more.

However, words are but wind, but seeing is believing, so we conduct free introductory webinars about the NGS framework for researchers who prefer to get an online overview of the capabilities available in UGENE.

Development of the NGS framework in UGENE was supported by grants RUB1-31097-NO-12 and OISE-14-60510-1 from NIH/NIAID.

References

1. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 06/01/2015.
2. FastQC web page: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
3. Li H. and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform, *Bioinformatics*, **25**:1754-60. BWA-MEM algorithm: http://bio-bwa.sourceforge.net/.
4. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118, *Fly*, Apr-Jun; **6**(2):80-92. PMID: 22728672.
5. Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y (2014) Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses, *PeerJ*, **2**:e644. doi:10.7717/peerj.644.

6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols*, **7**:562-578.
7. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y et al. (2011) Cistrome: an integrative platform for transcriptional regulation studies, *Genome Biology*, **12**:R83.
8. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, *Journal of Computational Biology*, **19**(5), 455-477. doi:10.1089/cmb.2012.0021.