

Bio2M platform: find everything in your RNA-Seq data

Benoit GUIBERT^{1,2}, Florence RUFFLÉ^{1,2}, Chloé BESSIÈRE^{1,2}, Anthony BOUREUX^{1,2} and Thérèse COMMES^{1,2}

¹ Centre Hospitalier Universitaire de Montpellier, 191 av. du Doyen Giraud, 34295, Montpellier, France

² IRMB U1183 INSERM-Université Montpellier, 80 rue Augustin Fliche, 34295, Montpellier, France

Corresponding author: Anthony.Boureaux@inserm.fr

URL: <https://bio2m.montp.inserm.fr/platform/>

Abstract *The Bioinformatics and BioMarkers (Bio2M) platform is dedicated mainly to the analysis of RNA-Seq data obtained from next-generation sequencer (NGS). As other platforms, we can do gene expression analysis, but also, transcript assembly for protein-coding and long non-coding genes already annotated or new genes. We are able also to look for rare expression variant, splice and chimera transcripts. As the public data bank of RNA-Seq data increase rapidly, new methods are required to analyse these huge data files. We are developing, with collaborators, new methods which permits to count and detect all kind of transcriptional events, in less than five minutes, in thousands of samples.*

The platform is also involved in bioinformatic training for biologist, from R beginners, RNA-Seq data analysis to mutation research.

Keywords NGS, RNA-Seq, Chimera, lncRNA, kmers.

1 Introduction

The aim of this platform is to provide bioinformatic analysis to researcher in the NGS field. We have an extensive range of expertise in the RNA-Seq analysis data. The platform belongs to the Montpellier's CHU. The core facility is located at the Institute of Regenerative Medicine and Biotherapies (IRMB) in Montpellier. The platform was created from a bioinformatic research team at IRMB to meet the demands of biologists. It has exchange with different laboratories around Montpellier in the academic or private research.

2 Actitivity

Computational techniques are used to analyze high-throughput sequences data. In the last 10 years, the sequencing technology has grown faster than the computational biology, which makes more difficult and complex to analyze of all generated data. Therefore, the software to process sequencing data evolved rapidly and required a lot of expertise before using them. The platform members are also belonging to a research team which analyze, develop pipeline and software dedicated to RNA-Seq data.

Major services provided by the platform:

1. Help to design experiment for NGS
2. RNA-Seq analysis
 - Quality check of sequencing data: FastQC, kmerExploR
 - Alignment on reference sequence: CRAC[1], STAR, Hisat2
 - Differential gene expression based on k-mers: DE-kupl[2], CountTags
 - Differential gene expression based on transcript or gene sequences: Kallisto, Sleuth, DESeq2, EdgeR
 - Search for chimeric genes and/or RNA: ChimCT[3,4], STARFusion
 - Transcripts assembly for new unannotated genes, protein-coding or non-coding genes (lncRNA): Stringtie
3. DNA-Seq analysis
 - Variation and mutation search: CracTools, Hisat2/Freebayes, VEP

4. Specific analysis on demand

The research team associated with the platform also develop new software based on kmer search, which permits to look in large sample data libraries (more than 1000 samples) for gene or transcript expression, chimera transcript or other transcriptional events.

3 IT infrastructure

Computational biology, in the NGS field, require huge computer resources for a wide range of computationally intensive tasks. The platform shares IT infrastructure with other teams in our building. Different servers are available for services like web server as well as the IT management. The low and small intensive analysis are done on a local cluster (about 20 nodes and some with large memory). This cluster, as IT managed entirely by the team, give the freedom to the team to do what they want.

For large computer intensive tasks, the team use the high-throughput computing (HPC) cluster provided by the Meso@LR (<https://meso-lr.umontpellier.fr/>) located in Montpellier.

All the pipeline analysis are under control of a bioinformatic workflow management system (snake-make) and use singularity container that can be run directly in any HPC cluster available. All results and files are given to the scientist via a web site, with restricted access for each project. All generated files are under versioning control to track the provenance of the workflow execution results and to certify for quality control.

4 Training

The platform manages few training in the bioinformatic fields. Theses training are designed principally towards biologist that are interested to understand what is a behind a bioinformatic workflow.

The next one will be about how to find a mutation in a gene panel or exome experiment. The purpose of this training is to learn:

- how to do control quality check on fastq sequences
- how to do the mapping on reference sequences
- how to do variant calling
- how to do variant filtering and why

Theses training are organized once or twice a year, generally between March and June of each year, in collaboration with the Mobidic bioinformatic team of Montpellier CHU and Seq.one corporate. The expertise required for these courses are provided by associate professor and professor of Montpellier University, and researchers from INSERM.

Acknowledgements

This work is supported by CHRU of Montpellier, Montpellier University and INSERM.

References

- [1] Nicolas Philippe, Mikaël Salson, Thérèse Commes, and Eric Rivals. CRAC: an integrated approach to the analysis of RNA-seq reads. 14(3):R30.
- [2] Jérôme Audoux, Nicolas Philippe, Rayan Chikhi, Mikaël Salson, Mélina Gallopin, Marc Gabriel, Jérémy Le Coz, Emilie Drouineau, Thérèse Commes, and Daniel Gautheret. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. 18(1):243.
- [3] Florence Rufflé, Jerome Audoux, Anthony Boureux, Sacha Beaumeunier, Jean-Baptiste Gaillard, Elias Bou Samra, Andre Megarbane, Bruno Cassinat, Christine Chomienne, Ronnie Alves, Sebastien Riquier, Nicolas Gilbert, Jean-Marc Lemaitre, Delphine Bacq-Daian, Anne Laure Bougé, Nicolas Philippe, and Therese Commes. New chimeric RNAs in acute myeloid leukemia. [version 2; peer review: 2 approved]. F1000Research 2017, 6(ISCB Comm J):1302.
- [4] Anne-Laure Bougé, Florence Rufflé, Sébastien Riquier, Benoit Guibert, Jérôme Audoux, and Thérèse Commes. RNA-seq analysis to detect abnormal fusion transcripts linked to chromothripsis. 1769:133–156, 2018.