ORSON: a nextflow workflow for transcriptome and proteome annotation

Cyril Noël, Pierre Cuzin, Laura Leroi, Alexandre Cormier & Patrick Durand

IFREMER-IRSI-Service de Bioinformatique (SeBiMER), Centre Bretagne - 29280 PLOUZANE, FRANCE

INTRODUCTION

One of the key steps in transcriptomic and proteomic analyses is to link sequences to biology through annotation. Namely, it consist in adding relevant biological information to these sequences by inferring their putative function and other features. However, this process requires a complex combination of successive tools and reference databases, as well as significant computing resources due to the amount of data. Then, it is quite difficult to group together results in order to obtain a complete biological understanding of the sequences because of the many tools and data formats involved. The implementation of an automated, standardized and user-friendly tool to process transcriptomic and proteomic annotations is therefore essential

IMPLEMENTATION

We have developed ORSON to combine state-of-the-art tools for annotation processes within a Nextflow [1] pipeline. ORSON combines sequence similarity search, functional annotation retrieval and functional prediction. While ORSON results can be analyzed through the command-line, it also offers the possibility to be compatible with BlastViewer graphical tool

CONCLUSION

ORSON, combined with BlastViewer, offers a real alternative to the complex use of bioinformatic annotation tools by providing the best of both worlds: a scalable workflow running on the command-line to fit any computing infrastructures, and a GUI tool to analyze results.

Key features

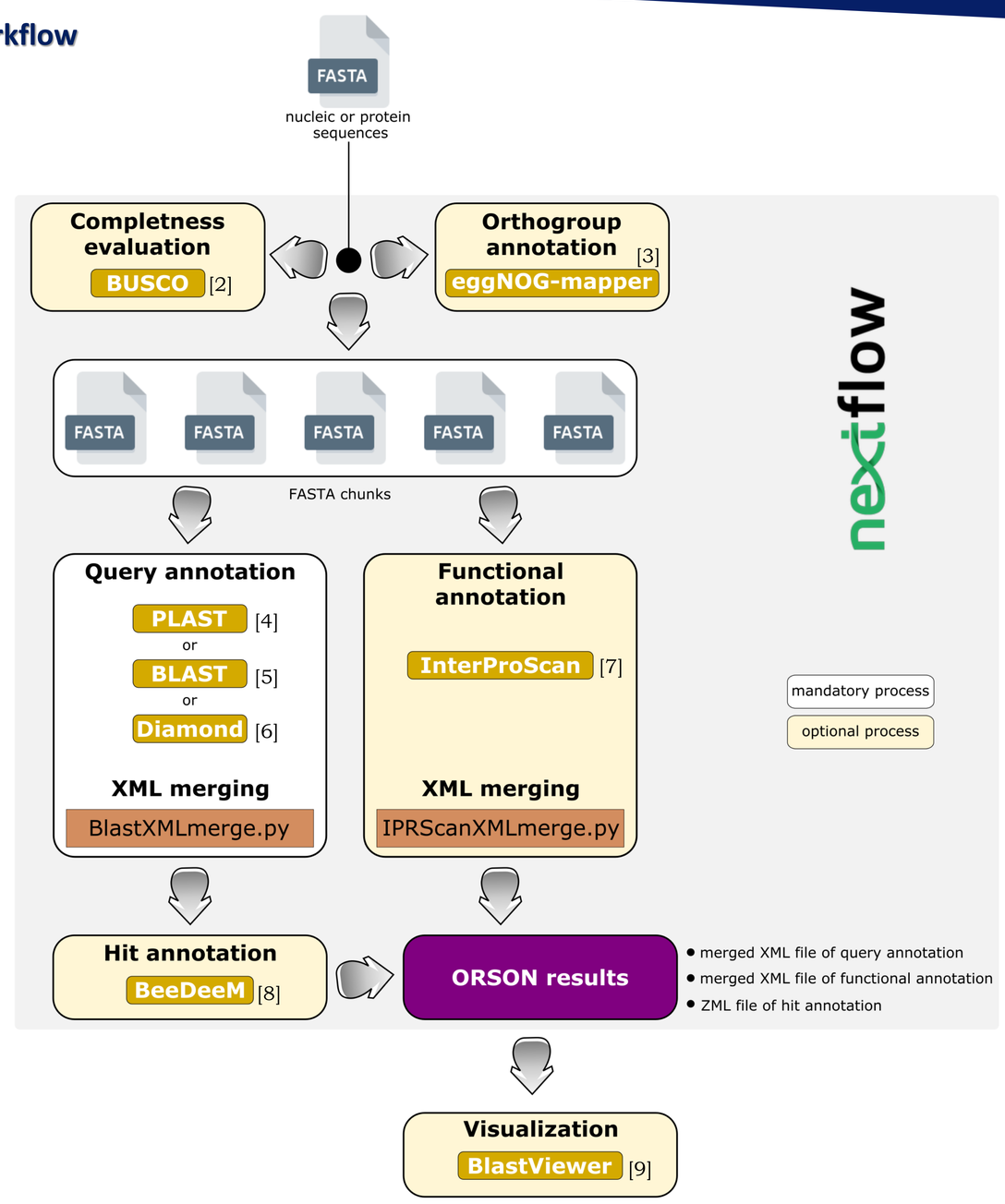
- FAIR workflow:
  - Reproducible results
  - Portable
- Automated installation of reference banks:
  - eggNOG-mapper
  - Uniprot/KB, RefSeq, *etc.*
    - Ready to be used with PLAST, BLAST, Diamond
  - InterProScan DB
- Standard XML output files compatible with Blast2GO and BlastViewer

Coming features

＋ lncRNA identification process

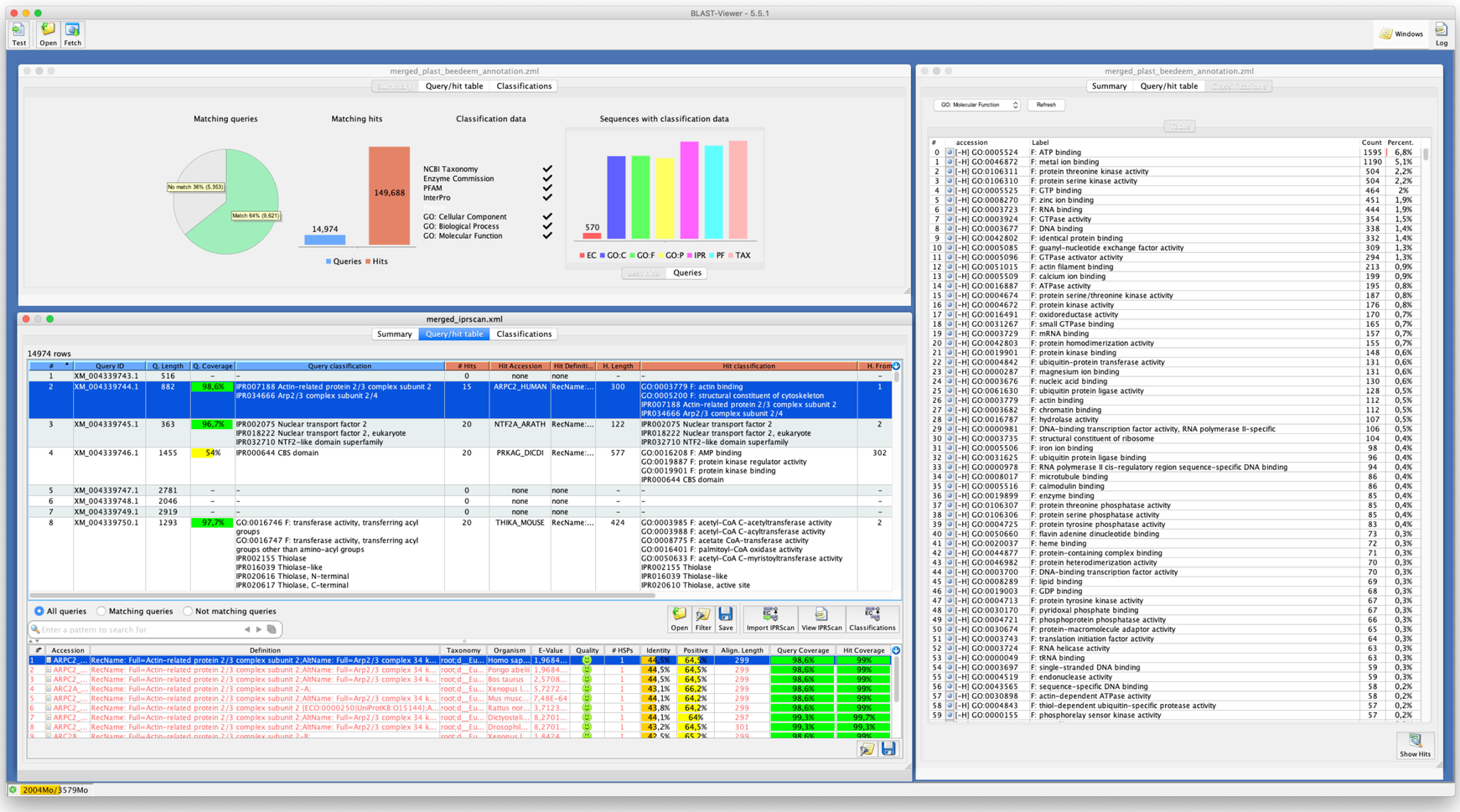
＋ rRNA inference process

Overview of ORSON workflow



Case study: annotation of a transcriptome

For evaluating ORSON workflow, we annotated the 14,974 sequences of *Acanthamoeba castellanii* str. Neff reference transcriptome (RefSeq:GCF_000313135.1) on Ifremer's DATARMOR supercomputer. Workflow was configured to create chunks of 1,000 sequences. With such a configuration, Nextflow has distributed 36 jobs on computing nodes. Overall processing was achieved in 133 minutes and included PLASTx against SwissProt, BUSCO completeness, InterProScan predictions and BeeDeeM annotations. Analysis of XML data files was achieved using BlastViewer, as illustrated below.



BlastViewer analysis session. Top left: summary of data set (matching queries against SwissProt; Classification data, i.e. InterProScan predictions on queries and SwissProt domains on hits). Bottom left: overview of all annotated queries; « Query classification » contains predicted domains from ORSON's InterProScan step; « Hit classification » contains domain features retrieved from SwissProt during ORSON's BeeDeeM annotation step. Right: overview of Gene Ontology Molecular Function IDs contained in this data set (other classifications can be selected by user: Gene Ontology, InterPro, Enzyme, PFAM, Taxonomy, *etc.*)

ORSON source code, installation instructions and user documentation are freely available at:
https://github.com/ifremer-bioinformatics/orson

1. Di Tommaso *et al.*, 2017. *Nature biotechnology*, **35**(4):316-319
2. Seppey *et al.*, 2019. *Methods in Molecular Biology*, 1962
3.Huerta-Cepas *et al.*, 2017. *Mol Biol Evol*, **35**(8):2115-2122
4. Van Nguyen *et al.*, 2009. *BMC bioinformatics*, **10**(1):1-13
5. Altschul *et al.*, 1990. *J. Mol. Biol.*, **215**:403-410
6. Buchfink *et al.*, 2015. *Nature Method*, **12**:59-60
7.Jones *et al.*, 2014. *Bioinformatics*, **30**(9):1236-40.
8. https://github.com/pgdurand/BeeDeeM
9.https://github.com/pgdurand/BlastViewer