The Journal of Open Source Software

GUESSmyLT: Software to guess the RNA-Seq library type of paired and single end read files

Erik Berner Wik[1], **Hampus Olin**[1], **Caitlin Vigetun Haughey**[1], **Lisa Klasson**[1], **and Jacques Dainat**[2, 3]

1 Molecular Evolution, Department of Cell and Molecular Biology, Uppsala University, 75124 Sweden. **2** IMBIM - Department of Medical Biochemistry and Microbiology, Box 582, S-751 23 Uppsala, SWEDEN. **3** National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, S-751 23 Uppsala, SWEDEN.

DOI: 10.21105/joss.01344

Software

- Review
- Repository
- Archive

Submitted: 19 March 2019
Published: 07 July 2019

Summary

Short-read RNA sequencing (RNA-seq) is a powerful approach allowing among others to investigate the expression of genes, to perform genome annotation, to detect the single nucleotide polymorphisms or look at the alternative gene spliced transcripts. Sequenced reads have characteristics that differ according to the RNA-seq library preparation protocols used. i) the reads can be either single end (only one side of a fragment is sequenced) or paired-end (both ends of a fragment are sequenced); ii) the reads can be stranded (the information about which strand was originally transcribed is preserved) or unstranded; iii) the right-most end of the fragment is the first sequenced (or only sequenced for single-end reads) or the left-most end of the fragment is the first sequenced (or only sequenced for single-end reads); iv) paired-end reads can be inward or outward looking; v) paired-end reads can both come from the original RNA strand/template or from the opposite strand or even one comes from the original RNA strand/template and the other from the opposite strand (Figure 1). The information regarding the library type is helpful for improving the reads mapping to a reference assembly/genome or to assemble them into a transcriptome. This is because the library type can help to discern location of ambiguous reads by using the read's relative orientation and from which strand it was sequenced. Unfortunately, this information regarding the library type used is not included in sequencing output files and can consequently be lost or miss-labelled before to be used the end user. Most of time it can be solved by contacting the parties involved in the generation of the RNA-seq data. But when it's not possible, this might yield in a waste of resources and time. Indeed one can launch the analysis with sub-optimum parameters, resulting in lower quality results or one can try to guess the library type with the available approaches: i) Using infer_experiment.py from the RSeQC package ("RSeQC: quality control of RNA-seq experiments," 2012), ii) Launching mappers with different parameters and compare the results; iii) mapping the reads and look at them within a genome browser; iv) using Salmon ("Salmon provides fast and bias-aware quantification of transcript expression," 2017). But none of them can guess the full information of the library type, they can require specific inputs (e.g. an annotation file), they can deal with only specific library types, or can be done only with substantial manual work. GUESSmyLT aims to automate the different steps needed for identifying the RNA-Seq library type as comprehensively as possible, and can deal with any type of input data: from mapped reads, from raw reads, with or without annotation, with or without reference genome. GUESSmyLT was developed as a snakemake pipeline consisting of three pre existing softwares (bowtie2 (Langmead B, 2012), trinity (Grabherr MG, 2011) and busco (Robert M. Waterhouse & Zdobnov, 2017)) and an inference step at the end that performs the library type prediction.

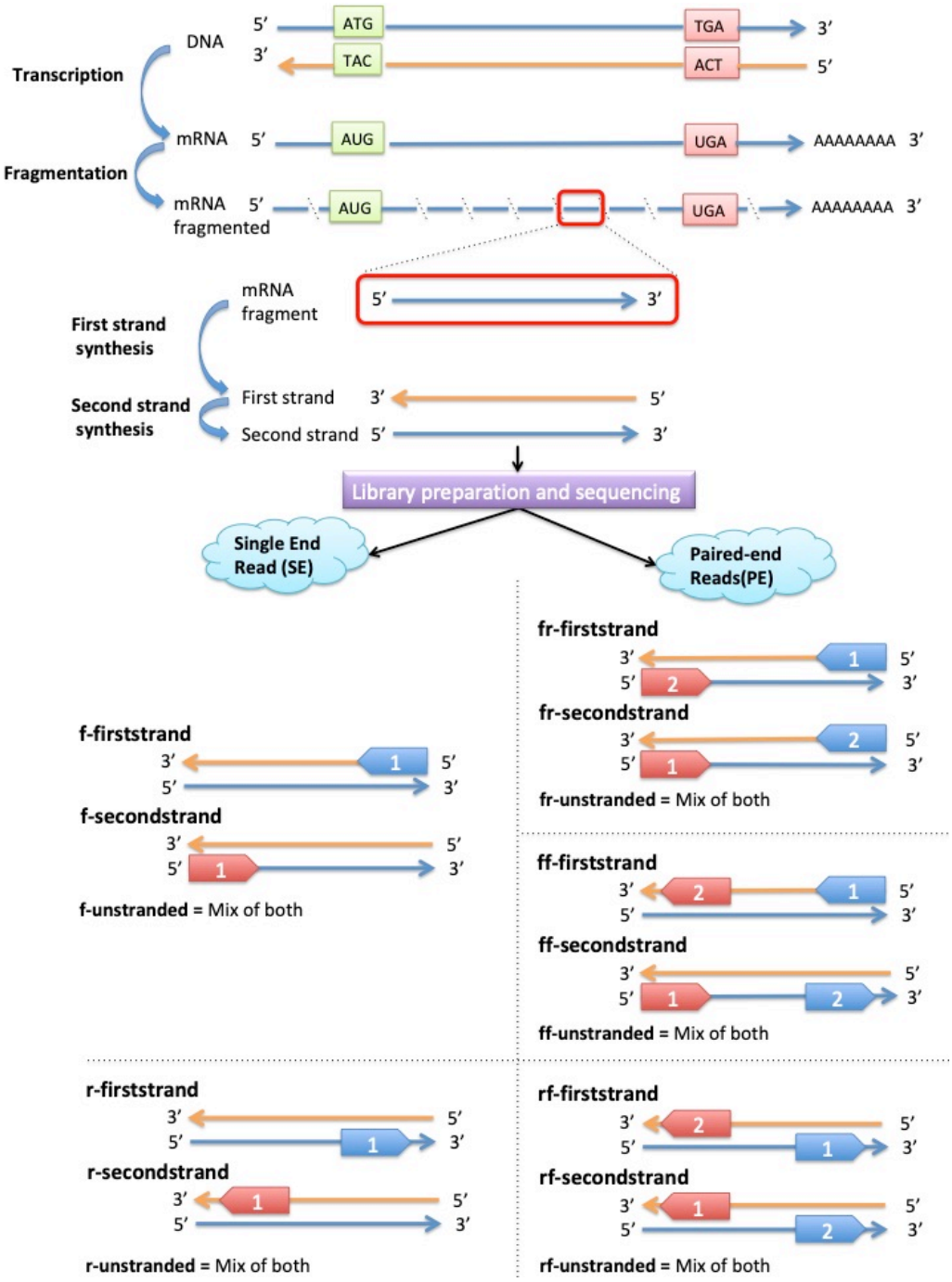

Figure 1: Overview of the different library types.

Authors Contributions

Berner Wik E., Olin H. and Vigetun Haughey C. contributed equally to this work.

Acknowledgements

We acknowledge Istvan Albert, Bengt Sennblad, Hadrien Gourlé, Jonas Söderberg and the NBIS development team for their inputs.

References

Grabherr MG, Y. M., Haas BJ. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol.* doi:10.1038/nbt.1883

Langmead B, S. S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. doi:10.1038/nmeth.1923

Robert M. Waterhouse, F. A. S., Mathieu Seppey, & Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. doi:doi:10.1093/molbev/msx319

RSeQC: quality control of RNA-seq experiments. (2012).*Bioinformatics*, *28*(16). doi:10.1093/bioinformatics/bts356

Salmon provides fast and bias-aware quantification of transcript expression. (2017).*Nature Methods*. doi:10.1038/nmeth.4197