MSc thesis project: An R package for inferring RNA half-lives from in-vivo SLAM-seq data

Background

Analysis of gene expression with methods of Next Generation Sequencing (https://en.wikipedia.org/wiki/RNA-Seq) is a key research tools in biomedical research and depends heavily on data science. Typical transcriptome sequencing data produces negative binominal-distributed gene counts [1]. However, more optimal models are required for newly developed methods, for example to account for data (multi)dimensionality or variables that cannot be empirically obtained from a biological system.

Required student background

MSc thesis project for a person keenly interested in applying statistical inference to big data in biomedical reasearch. Basic knowledge of biology and familiarity with transcriptome sequencing data (Next Generation Sequencing) will be an advantage.

Research objectives

- Implement an approach similar to [2] in `R` to infer RNA turnover kinetics from data obtained using SLAM-seq (thiol (**SH**)-**l**inked **a**lkylation for the **m**etabolic **seq**uencing of RNA) technology [3]
- Extend a Model described in [2] to account for variable label kinetics in *in vivo* experiments, such as those perfomed using Tagger transgenic mouse line [4]. Other relevant literature includes [5,6].

Short data description

- **Data type:** RNA (transcriptome) sequencing data and nucleotide (https://en.wikipedia.org/wiki/RNA) conversion rates (count data, see: https://kasperdanielhansen.github.io/genbioconductor/html/Count_Based_RNAseq.html)
- **Tentative data volume:** Approx. total volume of currently available data is 10^8 sequencing reads x 300 bases x 30 samples. For this project a subset of data may be used to reduce computation time. Also, if need be, it is possible to access an HPC cluster

Contact person

Lech Kaczmarczyk (lecka48@liu.se), Krzystof Bartoszek (krzysztof.bartoszek@liu.se). Project will be done in collaboration with the Department of Biomedical and Clinical Sciences of LiU.

References

[1] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology 2010;11. https://doi.org/10.1186/gb-2010-11-10-r106.

[2] Jürges C, Dölken L, Erhard F. Dissecting newly transcribed and old RNA using GRAND-SLAM. Bioinformatics 2018;34:i218–26. https://doi.org/10.1093/bioinformatics/bty256.

[3] Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. Nature Methods 2017;14:1198–204. https://doi.org/10.1038/nmeth.4435.

[4] Kaczmarczyk L, Bansal V, Rajput A, Rahman R, Krzyżak W, Degen J, et al. Taggera swiss army knife for multiomics to dissect cell typespecific mechanisms of gene expression in mice. PLOS Biology 2019;17:e3000374. https://doi.org/10.1371/journal.pbio.3000374.

[5] Uvarovskii A, Vries ISN, Dieterich C. On the optimal design of metabolic RNA labeling experiments. PLOS Computational Biology 2019;15:e1007252. https://doi.org/10.1371/journal.pcbi.1007252.

[6] Neumann T, Herzog VA, Muhar M, Haeseler A von, Zuber J, Ameres SL, et al. Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. BMC Bioinformatics 2019;20. https://doi.org/10.1186/s12859-019-2849-7.