Chair Digital Health & Machine Learning (Prof. Lippert)

Master Thesis

Differential expression (DE) analysis of TCGA RNA-seq dataset and integrating all it as a reproducible workflow

Aim:
The DE genes between specific conditions gives key information about the phenotypic variations in cancer samples **(Costa-Silva et al., 2017)**. The project aims to perform a DE analysis of TCGA 13000 samples from openly available resources such as The Cancer Genome Project (TCGA) dataset between tumor and adjacent healthy samples. In addition to that, the DE analysis (Love et al., 2014) between the known cancer subtypes. At the end, we are aiming to have multiple sets of long non-coding RNAs (lncRNAs), for all avalaible TCGA cancer types. Furthermore, instead of just analysising the DE lncRNAs which between tumor and normal samples for BRCA (breast cancer) as done in (Zhang et al., 2019), we will include all cancer samples, as well as will look into the DE lncRNAs between the various tumor subtypes.

Workplan:

1. Collect the TCGA RNA-seq counts dataset for 13000 cancer samples and their adjacent healthy samples
2. Collect the clinical data information and label the cancer subtypes
3. Use the sample information to perform the DE analysis between normal tumor and subtypes
4. Develop a reproducible pipeline which would perform all the above steps

Prerequisites:

- Good knowledge of R and python programming language
- Background in genomics or molecular biology
- Acquaintance with application of statistical models
- Interest in Bioinformatics pipeline/workflow development using Nextflow (DI Tommaso et al., 2017)

Reference:

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. In *PLoS ONE*. https://doi.org/10.1371/journal.pone.0190152

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. In *Nature Biotechnology*. https://doi.org/10.1038/nbt.3820

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. https://doi.org/10.1186/s13059-014-0550-8

Zhang, J., Sui, S., Wu, H., Zhang, J., Zhang, X., Xu, S., & Pang, D. (2019). The transcriptional landscape of lncRNAs reveals the oncogenic function of LINC00511 in ER-negative breast cancer. *Cell Death and Disease*. https://doi.org/10.1038/s41419-019-1835-3

Contact:
Alva Rani James, Phd