Humboldt Universität zu Berlin
Prof. Dr. Ulf Leser
Dr. Raik Otto

Freie Universität Berlin

October 6, 2021

MASTER THESIS EXPOSÉ

Working title: Clinical classification of common cancers by means of deconvolution-based machine learning

Melanie Fattohi

1 Introduction

Over the past decade, personalized medicine has become an essential aspect of oncology. Through personalized medicine, the treatment a patient receives is individually fitted to the molecular characteristics of the tumor from which the patient suffers. The premise of this approach is an accurate determination and classification of the molecular subtype of the tumor since different tumor subtypes require different treatments and hold different prognoses.

In the field of bioinformatics, tumor samples are often classified using a machine learning (ML) approach or clustering approach based on their similarity in gene expression measured by bulk RNA-sequencing (RNA-seq). In this way, further investigations of each tumor subtype can be made for the identification of novel gene signatures or independent biomarkers, which are unique to the tumor subtype [Zhao et al., 2018].

The limitation in using bulk RNA-seq for the molecular classification of tumors is that it measures aggregated gene expression of a heterogeneous mixture of cell types. Since cell types have varying levels of gene expression, the individual contribution of each cell type to the gene expression profile of a molecular subtype is obfuscated. By taking into account in which ratio certain cell types underlie bulk RNA-seq data of tumors may reveal more in-depth characteristics about each tumor subtype.

Transcriptional deconvolution is a computational method that infers the cell type composition of bulk RNA-seq data by using cell type-specific gene expression profiles as references. Recently, deconvolution approaches have been developed, which use single-cell RNA-sequencing (scRNA-seq) datasets as gene expression profile references for the prediction of cell type proportions [Baron et al., 2016] [Wang et al., 2019] [Moffitt et al., 2015]. The performance of the deconvolution is typically assessed by an empirical p-value.

SCDC [Dong et al., 2020] is a deconvolution method, which predicts cell type proportions of bulk RNA-seq data based on scRNA-seq data via a non-negative least squares regression framework.

Additionally, if multiple scRNA-seq reference datasets of the same tissue and cell types are available, SCDC offers an ENSEMBLE function for the integration of scRNA-seq datasets, which can improve deconvolution performance. In [Avila Cobos et al., 2020], SCDC was among the best performing deconvolution methods. However, SCDC does not provide an empirical p-value for the evaluation of the deconvolution result.

Transcriptional deconvolution gives rise to the idea that cell type proportions not only elucidate cellular heterogeneity in tumors but may also reveal statistically significant relations to clinical characteristics of tumor subtypes. [Otto, 2021] developed a framework, in which cell type proportions resulting from transcriptional deconvolution were further used as predictor variables in ML models for clinical classification of neuroendocrine neoplasms (NENs). The novelty of this framework was the utilization of scRNA-seq data of healthy tissue instead of cancerous tissue as references for deconvolution. Thus, [Otto, 2021] successfully addressed the issue of a lack of scRNA-seq data of NENs, which would otherwise make clinical classification more difficult.

2 Motivation and approach

In this master thesis we will investigate whether deconvolution of cancer types other than NENs is feasible and if resulting predictions are correlated with clinical characteristics. To this end, we intend to deconvolve tumor bulk RNA-seq data based on scRNA-seq data of healthy or, if available, cancerous tissue with the deconvolution method SCDC [Dong et al., 2020]. We chose to use SCDC for the deconvolution because it performed well compared to other methods [Avila Cobos et al., 2020] and because it allows integration of multiple scRNA-seq datasets. To evaluate the deconvolution performance of SCDC, we will implement the calculation of an empirical p-value as a function.

Based on the predicted cell type proportions and their reconstruction error, we will train a ML model, which will classify bulk RNA-seq samples in regards to a clinical characteristic relevant for the cancer type (e.g. tumor grading, expression of a certain receptor gene or other protein-coding genes).

We will integrate the deconvolution step including its evaluation, the ML step as well as other intermediate analyses and evaluations of results into a computational framework in form of an R package. Thus, assuming scRNA-seq data of the same tissue is provided, the R package will be applicable to bulk RNA-seq data of any tumor type, thereby addressing the limitation of bulk RNA-seq.

The feasibility of the deconvolution as well as the effectivity of the ML model in discerning between subtypes form the bioinformatic aspect of the work. Biologically, we may potentially detect molecular subtypes of cancers that have not been described before on the basis of cell type proportions predicted by deconvolution. Differing cell type compositions within tumor subtypes may reveal intermediate subtypes or differing cells-of-origin.

We will investigate bulk RNA-seq of pancreatic ductal adenocarcinoma (PDAC) as well as another cancer type, which uses the same tumor grading system (e.g. colorectal cancer). Moreover, we also intend to include a cancer type, which uses a different biomarker or clinical characteristic for subtyping the tumor (e.g. HER2 receptor in breast cancer).
Bulk RNA-seq datasets including meta information and scRNA-seq datasets of healthy and, if available, cancerous tissue will be obtained from The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and the Gene Expression Omnibus (GEO) database.

To guide the analysis and discussion of the results, the interpretation of biological findings will be supported by clinicians, who each have in-depth knowledge about a cancer type, that will be included into the master thesis.

The contributions of the master thesis could be biologically relevant for personalized medicine in oncology if the deconvolution-based ML model successfully predicts clinical and molecular characteristics of tumors, which supports the design of a treatment that is individually fitted to the specific tumor subtype. Otherwise, the master thesis would reveal that clinical classification of tumors based on cell type proportions predicted by deconvolution requires a more complex approach.

References

[Avila Cobos et al., 2020] Avila Cobos, F., Alquicira-Hernandez, J., Powell, J., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*.

[Baron et al., 2016] Baron, M., Veres, A., Wolock, S., Faust, A., Gaujoux, R., Vetere, A., Ryu, J., Wagner, B., Shen-Orr, S., Klein, A., Melton, D., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems*.

[Dong et al., 2020] Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C., Zou, F., and Jiang, Y. (2020). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*.

[Moffitt et al., 2015] Moffitt, R., Marayati, R., Flate, E., Volmar, K., Loeza, S., Hoadley, K., Rashid, N., Williams, L., Eaton, S., Chung, A., Smyla, J., Anderson, J., Kim, H., Bentrem, D., Talamonti, M., Iacobuzio-Donahue, C., Hollingsworth, M., and Yeh, J. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics*.

[Otto, 2021] Otto, R. (2021). *Distance-based methods for the analysis of Next-Generation sequencing data*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

[Wang et al., 2019] Wang, X., Park, J., Susztak, K. N.R. Zhang, N., and Li, M. (2019). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Nat Commun*.

[Zhao et al., 2018] Zhao, L., Lee, V., Ng, M., Yan, H., and Bijlsma, M. (2018). Molecular subtyping of cancer: current status and moving toward clinical applications. *Briefings in Bioinformatics*.