Discussion

Our findings from pan-cancer expression data show promising results. Via GSVA analysis we identified four clusters in the cancer types correlating strongly to the associated histological type. Glioblastoma seem to take a special role as they are predominantly characterized by the high activity of neural crest differentiation pathways and receptor tyrosine kinases. This is in line with previous studies showing that glioblastomas derive from neural crest cells [@neuralcrest].

This was also found for some melanoma like UVM, which explains the observed clustering of UVM with other glioblastoma. Also, the high receptor tyrosine kinase activity has been linked to the formation of UVM and glioblastoma and suggested as a possible target for therapy [@dis1; @dis2].

Further, especially liver and kidney adenocarcinoma seemed to form a strong subcluster within the other adenocarcinoma. They are characterized by exceptionally high activity of metabolic pathways such as carbohydrate metabolism, lipid, and amino acid synthesis. Again, this change in metabolism was previously found in hepatocellular carcinoma [@dis3].

The most significant classification we found was the clustering of tumor types by their differentiation stage. Poorly differentiated tumors like leukemia and squamous cell carcinoma show an upregulation of pathways associated with embryonic stem cell-like expression signatures. In contrast highly differentiated tumors like most adenocarcinoma as well as most glioblastoma underexpress these gene sets. Such a clustering by differentiation stage was previously described by Ben-Porath et al.. However, these findings cannot be verified directly as provided annotation data did not contain information regarding the differentiation stage [@dis4].

Taken together our results are in line with current research and allow for the following hypothesis: The expression profile of a given cancer type depends highly on its differentiation stage and its histological type but little on the actual tumor type itself. Understanding how these changes in expression link to mutational signatures might help in developing druggable targets for therapy.

From our GSEA and pan-cancer GSVA results, we identify two separate ways of carcinogenesis in THCA. The follicular subtype upregulates proliferative signaling through mTOR/PI3K and MAPK signaling pathways. This was previously shown by Furuya *et al* [@dis5].

A second way of carcinogenesis by signaling through alpha6beta4, RAS, JAK/STAT, and EWSR1/FLI1-fusion mediated pathways was observed in the data. This way of carcinogenesis was linked to non-follicular types of THCA [@result3; @dis6; @dis7].

Pan-cancer GSVA shows three distinct clusters in the expression data, upregulating either one or both ways of proliferative signaling. While the follicular subtype seemed to strongly correlate with one cluster, a similar process was not observed in tall-cell and classical phenotypes. With more detailed annotation data it might be possible to link anaplastic and papillary histological subtypes of THCA to the two yet unassigned clusters.

Despite differences in proliferative signaling, all clusters share an upregulated hedgehog signaling pathway which is consistent with the literature [@dis8]. Also, metabolic changes in line with the Warburg effect were observed in all clusters.

Our data suggest that our neuronal network is well suited to predict pathway activities from GSEA data. The model shows an excellent fit to data and produces only minor errors. However, both linear models struggle in predicting the data accurately. This might be since GSEA pathway activity data usually clusters into an up- and downregulated group with no values in between. Since the REACTOME_INTERLEUKIN_36_PATHWAY also shows this problem, the two clusters might produce larger correlation values that might impact the accuracy of the regression coefficients and the intercept. Secondly, the correlation of the residuals with the test data values did not approach zero, thus, our linearity assumption is not met. Therefore, it can be concluded that a linear regression model is not well suited to predict the REACTOME_INTERLEUKIN_36_PATHWAY activity accurately.

Outlook

Futher ways of analysis could be the prediction of the histological type of THCA as well as the way of carcinogenesis with a neuronal network. This might be possible with a larger training data set as well as more detailed and specified annotations. Furthermore, it might be possible to link whole genome sequencing and methylation data to pathway activity. In that way, one could suggest a suitable targeted therapy option for a THCA patient based only on sequencing data from a small biopsy sample.