WEEK 7 ASSIGNMENT

Research in Biotechnology – BIOL 51200 | Spring 1 2021

Christina Morgenstern

Full Project Proposal

1. Introduction

1.1 Omics technologies and multi-omics research

Omics broadly refers to data of a specific type of biomolecule that have been generated by next-generation sequencing (NGS) or other high-throughput technologies on a genome-wide scale. Genomics, transcriptomics, proteomics and metabolomics aim to quantify DNA, mRNA, proteins and metabolites, respectively. The reduction in cost and sample processing time has led to an explosion of omics data in different fields. However, the analysis of single omics data lacks the resolution of inferring causal relationships between molecular layers. Multi-omics aims at analyzing multiple omics datasets from the same cell/tissue at the same time. Thereby gaining a better understanding of the associations between biological entities, untangling regulatory networks or identifying biomarkers (Krassowski et al., 2020; Manzoni et al., 2016).

1.2 Approaches of multi-omics data analysis in cancer research

Cancer is a complex and heterogeneous disease with alterations at different levels of the information-flow from genome to proteome. The transition of a normal cell to a malignant cell requires the acquisition of cancer hallmarks. Typically, these hallmarks are based on molecular changes leading to uncontrolled and sustained proliferation, resisting cell death, evading growth suppressors, replicative immortality, faulty angiogenesis and metastasis. Complex phenotypic changes driving tumor progression further arise through an altered energy metabolism as well as the evasion of immune destruction (Hanahan & Weinberg, 2011). Acquiring these hallmarks requires alterations in the cellular machinery within tumor cells and tissues driven by molecular changes at the genome, epigenome, transcriptome, proteome and metabolome layers. While single-omics studies have contributed to our understanding of what drives the acquisition of individual hallmarks those studies often fall short in making connections between several hallmarks and in understanding the multifaceted etiology of this disease (Chakraborty et al., 2018). Thus, integrating multiple omics datasets is necessary to get a more holistic view on tumorigenesis, discover new therapeutic targets and identify novel cancer biomarkers (Nicora et al., 2020). Multi-omics approaches are capable of decoding links between a cancerous genotype and its phenotypic characteristics with the goal of driving efforts in personalized oncomedicine.

1.3 Objective of study

Knowing about cancer heterogeneity between different patients is of great importance for choosing the optimal treatment regime as well as for predicting clinical outcomes. Several studies have shown that the integration of multiple omics data sets and the application of systems biology or machine learning approaches can reveal cancer subtypes and disease mechanisms (reviewed in Biswas & Chakrabarti, 2020; Menyhárt & Győrffy, 2021; Nicora et al., 2020).

Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive malignancies with a 5-year survival rate of less than 10% (Ryan et al., 2014). The difficulty in early diagnosis and the limited response to treatment are attributed to a considerable heterogeneity among patients and within a primary tumor (Sarantis et al., 2020). Traditional classification is based on histopathology and has led to the identification of two to six subtypes of PDAC (Roy et al., 2021). While it is known that genetic mutations initiate tumorigenesis those alterations are unable to capture the vast heterogeneity observed between different PDAC patients. It has further been proposed that epigenetic modifications are implicated in driving this heterogeneity rather than genetic alterations alone (Juiz et al., 2019). A multi-omics study integrating data from mRNA, miRNA and DNA methylation of 150 samples using a Similarity Network Fusion (SNF) approach also revealed a complex molecular landscape of PDAC suggesting distinct subtypes (Raphael et al., 2017). Four subtypes of PDAC were identified in a study involving a multi-omics integration analysis based on genomics, epigenomics and transcriptomics data on datasets from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) (Kong et al., 2020). Recently, Roy et al., have employed an integrated unsupervised clustering approach to the PDAC dataset TCGA called TCGA-PAAD involving gene expression and methylation data and obtaining five relevant subtypes (Roy et al., 2021). These results show that there is an inconsistency on how many different subtypes of PDAC can be described and how the molecular heterogeneity is driving clinical outcomes.

So far, there hasn´t been a study that has comprehensively integrated more than three layers of information from the genome to the proteome in an integrated manner to assess PDAC heterogeneity.

In this research the TCGA-PAAD cohort is used to infer sources of variability within the samples through the integration of data from SNV (single nucleotide variants), CNV (copy number variation), methylation, mRNA-Seq, miR-Seq and RPPA (Reverse Phase Protein Assays) analyses. In performing such an integrated study and taking into account information from the genome to the proteome, a better molecular classification of this cancer type is anticipated which provides the basis for the discovery of new prognostic biomarkers.

2. Data for multi-omics cancer research

2.1 Repository for cancer omics data

The Cancer Genome Atlas TCGA is a rich resource for different types of omics data from mRNA-Seq, DNA-Seq, miRNA-Seq, SNV (single-nucleotide variant), CNV (copy number variation), DNA methylation and RPPA (reverse phase protein array) assays (Das et al., 2020). LinkedOmics has curated data from TCGA and corresponding proteome data from The Cancer Proteome Atlas (CPTA) and provides the possibility to download and visually explore the data (Vasaikar et al., 2018). The different datasets are available as matrix files which can serve as inputs for the downstream multi-omics analysis pipeline.

The dataset can be retrieved via the following link:
http://linkedomics.org/data_download/TCGA-PAAD/

The TCGA-PAAD cohort comprises 185 samples containing clinical annotation, methylation data, miRNA expression, somatic mutations, protein expression, RNAseq data and copy number changes (see Figure 1).

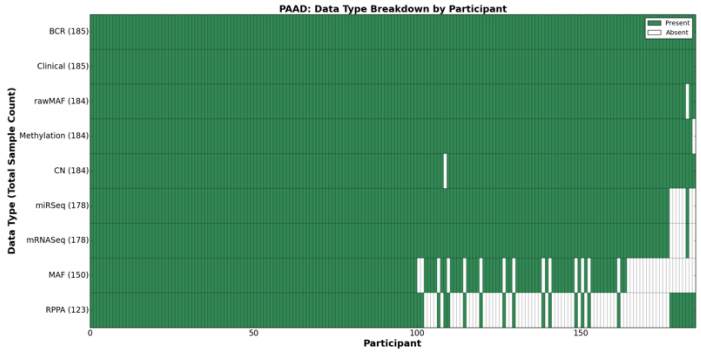

Figure 1. Overview of TCGA-PAAD cohort and the available data. Green indicates that the data is present for this patient sample whereas grey indicates the absence of this data point (*Broad GDAC Firehose*, n.d.)

Several multidimensional online data portals exist that allow further inspection and visualization of the TCGA data resources (Das et al., 2020). For example, Firebrowse is an interface that contains processed TCGA data and can be used to visualize and download data pertaining to the TCGA-PAAD cohort. In this research, Firebrowse will be used for initial data exploration in addition to LinkedOmics.

3. Analysis outline

3.1 Challenges

Several challenges arise when integrating multiple omics datasets which need to be addressed (Mirza et al., 2019). The curse of dimensionality is one of those challenges and implies a high number of features while the sample sizes are most of the time relatively small. The TCGA-PAAD dataset comprises 185 patient samples but the different assays contain a varying number of attributes from 14 (clinical data) to 334,357 (methylation). Furthermore, heterogeneity of the data which arises from the different high-throughput assays as well as the different study designs need to be taken into account. Missing data can arise for various reasons for example within the acquisition of measurements from the high-throughput platforms. The TCGA-PAAD dataset has some missing values, predominantly for the reverse phase protein assay but has a complete set of readings for about 100 patients (Figure 1). Samples with missing data will not be discarded or imputed in the first place but the chosen model that will be applied to the data is able to handle missing data and once established can impute missing entries afterwards.

3.2 Integration of different omics data types

Artificial intelligence and especially machine learning (ML) algorithms are becoming more and more important in the analysis of multi-omics data. In cancer research ML has been applied to

multi-omics data in order to identify disease subtypes, aide in cancer prognosis or help in identifying therapeutic targets (Biswas & Chakrabarti, 2020).

Next to the challenges that come with the individual omics datasets the integration of those different types of data is a further task. Several ways to integrate diverse omics data have been described such as horizontal and vertical integration. While the former approach uses the same data e.g., genomics data from different cancer types, the latter aims at integrating various omics types from one cancer type. A vertical omics data analysis can further be performed on an individual basis of the different omics datasets or it can be carried out in an integrated manner. Algorithms proposed for vertical omics data integration fall in five categories: network-based, Bayesian, fusion-based, similarity-based, correlation-based and further multivariate methods (Das et al., 2020). In this research, an integration-based vertical analysis of different omics data from a single cancer type – PDAC – is employed with the help of a factor analysis which falls in the last category of multivariate algorithms.

Argelaguet et al., developed a feature transformation algorithm called Multi-Omics Factor Analysis (MOFA) which is capable of identifying the sources of variability in multi-omics datasets (Argelaguet et al., 2018). The software can also be used for a variety of downstream analyses such as data imputation, detection of outliers and identification of subgroups from a common set of samples. The researchers have successfully used MOFA to identify the major dimensions of disease heterogeneity in chronic lymphocytic leukemia (Argelaguet et al., 2018). A recent update of the software resulted in the release of MOFA+ which next to analyzing bulk data is also capable of analyzing single-cell RNA-seq data (Argelaguet et al., 2020). The downturn of MOFA is that it can only capture linear relationships. This is a limitation because there might be a lot of non-linear relationships between the biomolecules of the different layers.

Modern data science techniques such as machine learning and especially deep learning can capture non-linear relationships between the different layers of information. For example, the deep learning algorithm developed by Lemsara et al, PathME, is able to integrate multi-omics data in an effective and interpretable manner and derives patient specific pathway information next to tumor subtyping (Lemsara et al., 2020).

In order to address the question of tumor heterogeneity sources in PDAC, the statistical framework MOFA+ tool suite is chosen because of its established pipeline and extensive documentation. The model will be implemented in Python (Rossum, 2007). All available data matrices of the TCGA-PAAD multi-omics data will be used in the analysis in order to infer an interpretable low-dimensional representation by the discovery of a few latent factors. These learned factors represent the sources of variation across the different data modalities. This can further help in identifying cellular states or disease subtypes in PDAC. Figure 2 outlines the MOFA computational framework of unsupervised discovery of latent factors demonstrating the principal axes of heterogeneity across all samples. The model can successfully handle missing values as well as different data types. This is important because the TCGA-PAAD data contains some missing values (see Figure 1) as well as is heterogenous due to different measurements and scales.

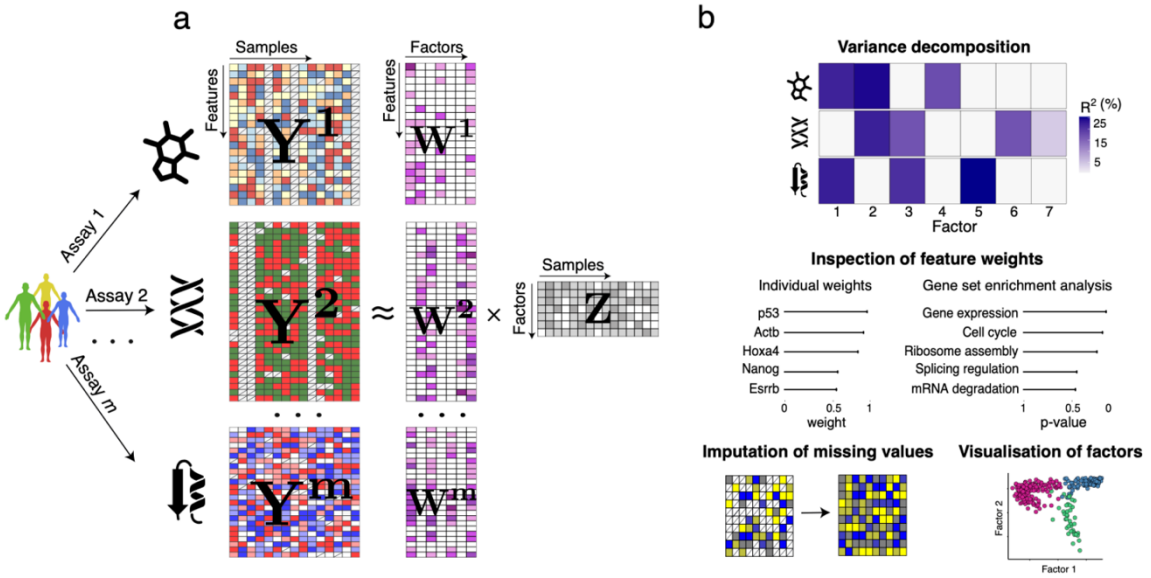

Figure 2. Overview of MOFA model (a) and downstream analyses (b) (Argelaguet et al., 2018).

3.3 Downstream analysis of integrated multi-omics TCGA-PAAD data

After the model has been established in MOFA+, the output will be used for further downstream analyses such as visualization of the factors, clustering and classification of samples in low-dimensional spaces and further factor annotation using gene set enrichment analysis. It is also possible to determine outliers as well as perform an imputation of missing values which might be addressed as well (Argelaguet et al., 2018). For the downstream analyses the programming language R and various R packages such as the accompanying *MOFAtools* will be used.

Initially, the identified factors and their contribution to the overall variation will be inspected and their etiology identified. A quantification of the amount of variance explained by each factor is performed to investigate the different factor loadings. Individual factors will be plotted against each other to assess the combined effect. Several types of visualizations like scatter plots and heatmaps can aide in visually inspecting the data from the model. Non-linear dimensionality reduction techniques, such as t-SNE (t-distributed stochastic neighbor embedding) or UMAP (Uniform Manifold Approximation and Projection), can be employed using the identified factors.

Performing Gene Set Enrichment Analysis (GSEA) can be a useful application in inspecting the information across genes within the context of biological pathways. This information is important in uncovering the sources of variability and thus establishing subtypes within the TCGA-PAAD samples.

4. Conclusion

This research performs an integrated study of different omics datasets from the TCGA-PAAD cohort using the MOFA+ model and with the goal of deriving the major sources of heterogeneity in this tumor type.

5. References

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21(1), 111. https://doi.org/10.1186/s13059-020-02015-1

This paper describes an update of the Multi-Omics Factor Analysis software (MOFA+) a statistical framework for comprehensive and scalable integration of multi-modal bulk and single-cell omics data.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124. https://doi.org/10.15252/msb.20178124

The authors describe the Multi-Omics Factor Analysis (MOFA) framework for unsupervised integration of multi-omics data which is capable of characterizing heterogeneity between samples.

Biswas, N., & Chakrabarti, S. (2020). Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer. Frontiers in Oncology, 10, 588221. https://doi.org/10.3389/fonc.2020.588221

In this review the authors discuss machine learning approaches for multi-omics data integration in cancer research. They discuss the major approaches of ML and list studies which have employed in ML in their multi-omics analysis.

Broad GDAC Firehose. (n.d.). Retrieved March 6, 2021, from https://gdac.broadinstitute.org/

Chakraborty, S., Hosen, M. I., Ahmed, M., & Shekhar, H. U. (2018, October 3). Onco-Multi-OMICS Approach: A New Frontier in Cancer Research [Review Article]. BioMed Research International; Hindawi. https://doi.org/10.1155/2018/9836256

In this review the authors describe the importance of multi-omics analyses in cancer research to drive the field of personalized oncomedicine. They give a description of the different high-throughput techniques and discuss applications and studies.

Das, T., Andrieux, G., Ahmed, M., & Chakraborty, S. (2020). Integration of Online Omics-Data Resources for Cancer Research. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.578345

The authors of this review discuss available databases for retrieving publicly available omics data. Furthermore, they state the different methodologies along with their strengths and weaknesses for integrating different omics modalities.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. Cell, 144(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hanahan and Weinberg have proposed six hallmarks of cancer in 2000. In this paper they discuss the addition of new enabling hallmarks and the importance of the tumor microenvironment on the cancer phenotype.

Juiz, N. A., Iovanna, J., & Dusetti, N. (2019). Pancreatic Cancer Heterogeneity Can Be Explained Beyond the Genome. Frontiers in Oncology, 9. https://doi.org/10.3389/fonc.2019.00246

In this study, Juiz et al., demonstrate that PDAC phenotypes arise through a combination of epigenetic and transcriptional changes rather than gene mutations alone.

Kong, L., Liu, P., Zheng, M., Xue, B., Liang, K., & Tan, X. (2020). Multi-omics analysis based on integrated genomics, epigenomics and transcriptomics in pancreatic cancer. Epigenomics, 12(6), 507–524. https://doi.org/10.2217/epi-2019-0374

The authors have performed a multi-omics analysis of genomics, epigenomics and transcriptomics data of TCGA-PAAD cohort identifying five cancer subtypes.

Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.610798

Krassowski et al., discuss in their review the challenges that come with multi-omics integration. They also list available tools and discuss their strengths and limitations.

Lemsara, A., Ouadfel, S., & Fröhlich, H. (2020). PathME: Pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinformatics, 21(1), 146. https://doi.org/10.1186/s12859-020-3465-2

The authors present a deep learning framework called PathME which is capable of both integrating multi-omics data and providing interpretable pathway information.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2016). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics, 19(2), 286–302. https://doi.org/10.1093/bib/bbw114

A review that states the rise of the omics technologies.

Menyhárt, O., & Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Computational and Structural Biotechnology Journal, 19, 949–960. https://doi.org/10.1016/j.csbj.2021.01.009

The authors review frameworks for multi-omics data integration.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., & Ping, P. (2019). Machine Learning and Integrative Analysis of Biomedical Big Data. Genes, 10(2), 87. https://doi.org/10.3390/genes10020087

In this review the authors discuss the challenges that arise with multi-omics integration studies. They focus on the curse of dimensionality, data heterogeneity, missing data, class imbalance and scalability issues.

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Frontiers in Oncology, 10, 1030. https://doi.org/10.3389/fonc.2020.01030

This paper reviews approaches to multi-omics analyses in cancer research. The authors describe the aims of ML in multi-omics studies in oncology as well as list research papers that have employed different ML algorithms specifically in cancer research.

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., Robertson, A. G., Cherniack, A. D., Gupta, M., Getz, G., Gabriel, S. B., Meyerson, M., Cibulskis, C., Fei, S. S., Hinoue, T., Shen, H., Laird, P. W., Ling, S., Lu, Y., … Zenklusen, J. C. (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell, 32(2), 185-203.e13. https://doi.org/10.1016/j.ccell.2017.07.007

An integrated genomic, transcriptomic and proteomic profiling of 150 PDAC specimens supports distinct molecular subtypes of pancreatic cancer.

Rossum, G. van. (2007). Python Programming Language. In J. Chase & S. Seshan (Eds.), Proceedings of the 2007 USENIX Annual Technical Conference, Santa Clara, CA, USA, June 17-22, 2007. USENIX.

Roy, S., Singh, A. P., & Gupta, D. (2021). Unsupervised subtyping and methylation landscape of pancreatic ductal adenocarcinoma. Heliyon, 7(1). https://doi.org/10.1016/j.heliyon.2021.e06000

The authors performed unsupervised clustering using the intNMF model on the TCGA-PAAD cohort and identified five distinct cancer subtypes.

Ryan, D. P., Hong, T. S., & Bardeesy, N. (2014, September 10). Pancreatic Adenocarcinoma (world) [Review-article]. Http://Dx.Doi.Org/10.1056/NEJMra1404198; Massachusetts Medical Society. https://doi.org/10.1056/NEJMra1404198

Review on Pancreatic Adenocarcinoma.

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. World Journal of Gastrointestinal Oncology, 12(2), 173–181. https://doi.org/10.4251/wjgo.v12.i2.173

This paper describes the tumor microenvironment of PDAC and addresses challenges in diagnosis and treatment.

The Cancer Genome Atlas Program—National Cancer Institute (nciglobal,ncienterprise). (2018, June 13). [CgvMiniLanding]. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Vasaikar, S. V., Straub, P., Wang, J., & Zhang, B. (2018). LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Research, 46(D1), D956–D963. https://doi.org/10.1093/nar/gkx1090

In this article the authors describe the LinkedOmics database, a platform that holds multi-omics data from TCGA and CPTAC.