LEWIS UNIVERSITY

UNSUPERVISED INTEGRATION OF MULTI-OMICS DATA REVEALS MOLECULAR ETIOLOGY OF INTERTUMOR HETEROGENEITY IN PANCREATIC DUCTAL ADENOCARCINOMA

RESEARCH PROJECT SUBMITTED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE, DATA SCIENCE

BY

CHRISTINA MORGENSTERN, PH.D.

DIRECTOR: PIOTR SZCZUREK, PH.D.

ROMEOVILLE, IL

JULY 2021

ABSTRACT

Cancer is a complex and heterogeneous disease with alterations at different levels of the information-flow from genome to proteome. While single *omics* studies have contributed to our understanding of what drives the acquisition of cancer hallmarks, such studies often fall short in making connections between several hallmarks, and in understanding the multifaceted etiology of cancer. Thus, integrating multiple high-throughput *omics* data sets is necessary to get a more holistic view on tumorigenesis, discover new therapeutic targets, and identify novel cancer biomarkers. Multi-*omics* approaches are capable of decoding links between a cancerous genotype and its phenotypic characteristics with the goal of driving efforts in personalized onco-medicine.

Pancreatic Ductal Adenocarcinoma (PDAC) is one of the most aggressive malignancies with a 5-year survival rate of less than 10%. The difficulty in early diagnosis and the limited response to treatment are attributed to a considerable heterogeneity among patients and within the tumor. While it is known that genetic mutations initiate tumorigenesis those alterations are unable to capture the vast heterogeneity observed between different PDAC patients.

In this study, the unsupervised Multi-Omics Factor Analysis (MOFA) framework is used to integrate data from a variety of *omics* assays to address the molecular sources of PDAC intertumor variation. The publicly available PDAC data set from The Cancer Genome Atlas (TCGA) comprising 185 patients was used to comprehend and interpret the interrelationships between the biomolecules at the genome, epigenome, transcriptome, and proteome layer. The observed major sources of heterogeneity were attributed to the top three factors characterized as roles in cell signaling, cilia movement and immune system regulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

AI	Artificial Intelligence
DNA	Deoxyribonucleic acid
ELBO	Evidence of Lower Bound
GSEA	Gene Set Enrichment Analysis
HGP	Human Genome Project
jDR	joint Dimensionality Reduction
MOFA	Multi-Omics Factor Analysis
mRNA	messenger RNA
miRNA	microRNA
ML	Machine Learning
MS	Mass Spectrometry
NGS	Next-Generation Sequencing
PDAC	Pancreatic Ductal Adenocarcinoma
PCA	Principal Component Analysis
RNA	Ribonucleic acid
RNA-seq	RNA sequencing

RPPA	Reverse Phase Protein Assay
SNV	Single Nucleotide Variant
SCNV	Somatic Copy Number Variation
TCGA	The Cancer Genome Atlas
TNM	Tumor, Node, Metastasis

CHAPTER 1 INTRODUCTION

To understand human health as well as diseases comprehensively, it is necessary to identify and interpret the interrelationships between different biomolecules at the various molecular layers such as the genome, epigenome, transcriptome, and proteome layer. With the rise of high-throughput technologies a revolution in the biomedical sciences and healthcare sector has been triggered which led to the generation of massive amounts of data representing the status of the biomolecules at these different layers. While the past decade has seen the analyses of this information in a layer-dependent manner, currently the integration of data obtained from measuring different biomolecules provides a promising avenue to assess and understand complex biological systems in a more systematic and holistic manner.

Within a cell, the flow of information starts with deoxyribonucleic acid (DNA) which stores the hereditary information. DNA is transcribed into messenger RNA (mRNA) and further processed by translation into amino acid chains, resulting in a genotype to phenotype transfer of information. Starting with the Human Genome Project (HGP), which revealed specific sequence of the alphabet of the human DNA, researchers set out to decipher the many different biomolecules in a comprehensive and global manner. The term “omics” was added to molecular biology terms to refer to an area of study which aims at deciphering the totality of the biomolecule within a system.

1.1 *Omics* – layers

The suffix -*ome* describes the study of molecules within a specific biological field such as the gene (genome), mRNA transcript (transcriptome), protein (proteome), and metabolite (metabolome). Similarly, the suffix -*omics* in biology refers to the analysis of a pool of biological molecules to untangle structure, function, and dynamics of an organism such as genomics, transcriptomics, proteomics, and metabolomics. These areas of scientific endeavor have in common that they study a *totality* of some sort of biological entity.

After the completion of the HGP in 2001, technologies capable of assessing cells, tissues, or organisms in a holistic way were initiated to detect all genes, mRNA transcripts, proteins, and metabolites within a complete cell, tissue, or organism. The term *Omics* was coined by Marc Wilkins in 1996 (Wilkins, Pasquali, et al., 1996) and genomics paved the way for the development followed by transcriptomics and proteomics (Arivaradarajan & Misra, 2018).

Spatial and/or temporal dynamics, as well as chemical modifications, drive variation and complexity at the different omics levels. Whereas the genome is largely static and the least complex level, the transcriptome is characterized by temporal dynamics and alternative splicing leading to an increase in complexity. The proteome increases further in complexity due to spatio-temporal dynamics and posttranslational modifications. Thus, the information flow from the genome to the proteome shows an exponential increase in complexity which is mirrored in the data obtained from the different levels (Chakraborty et al., 2018).

Table 1 is an overview of different omics technologies and how they are applied to understand biological systems and to uncover molecular signatures in health and disease.

Table 1. Omics techniques and their application (adapted from Chakraborty et al., 2018).

Omics	Type	Principle	Throughput	Application
Genomics	Whole exome sequencing	Next-generation sequencing	High	Genome-wide mutational analysis
	Whole genome sequencing	Next-generation sequencing	High	Exome-wide mutational analysis
	Targeted gene / exome sequencing	Sanger-sequencing	Low	Mutational analysis in targeted gene / exon
Epigenomics	Methylomics	Whole-genome bisulfite sequencing	High	Genome-wide mapping of DNA methylation pattern
	Chromatin Immunoprecipitation (ChIP)-sequencing	Chromatin Immunoprecipitation and Next-generation sequencing	High	Genome-wide mapping of epigenetic marks
miRNomics	Small RNA-sequencing	Next-generation sequencing	High	Genome-wide expression pattern analysis
Transcriptomics	RNA-sequencing	Next-generation sequencing	High	Genome-wide differential gene expression analysis
	Microarray	Hybridization	High	Differential gene expression analysis
Proteomics	Reverse Phase Protein Assay	Antibody-based	Low	Differential protein abundance analysis
	Deep-proteomics	Mass-spectrometry	High	Genome-wide differential protein expression analysis
Metabolomics	Deep-metabolomics	Mass-spectrometry	High	Differential metabolite expression analysis

1.1.1 Genomics

Genomics focuses upon studying genes, the entities of inheritance, and further structural and functional analyses of whole genomes of organisms. Hereditary information is stored in DNA in the form of four nucleotides, with the bases adenine (A), thymine (T), guanine (G) and cytosine (C). While genetics studies one gene at a time, genomics considers the full complement of hereditary material within a cell, tissue, or organism. Moreover, interactions between alleles and loci within a genome as well as effects of one gene affecting multiple traits (pleiotropy), the interaction between genes (epistasis) and the improved biological function as a mixture of genes (heterosis) are of interest within the study of genomics.

Analysis of genome data includes nucleotide sequences, expressed sequence tags (ESTs), complementary DNAs (cDNA) and chromosomal gene arrangements. The generation of sequences from whole genomes was driven by the advances in sequencing by pioneering work of Fred Sanger and recent next-generation sequencing (NGS) technologies as well as the development of *in silico* computational algorithms (Goodwin et al., 2016).

The creation of a human reference genome which determined the exact order of the base pairs in the human DNA by sequencing was a major endeavor which took 13 years and cost US$2.7 billion. The initial human genome draft sequence was the result of sequencing approximately 20 individuals whose DNA was extracted and sheared into ~150-200-kb. Thus, the human genome is a mosaic which has undergone further updates since its publication in 2001 and in the current version, GRch38.p13, contains 2.95 Gb of sequences and 349 gaps (Sherman & Salzberg, 2020). This effort has been the basis of many studies focusing on identifying gene variants implicated in disease, response to treatment or patient prognosis. Genome-wide association studies (GWAS) are

a study-design in which thousands of individuals are genotyped for genetic markers (single nucleotide variants, SNV) and the differences between disease and control states are assessed statistically with differences attributing to evidence of association (Hasin et al., 2017).

Somatic copy number variations (SCNV) refer to changes in chromosome structure that lead to the gain or loss of copies of sections of DNA and represent a hallmark of cancer. Through the changes in gene dosage and structure, SCNVs may activate oncogenes or inactivate tumor suppressors. One difficulty in the analysis of this kind of data is distinguishing between alterations driving cancer development and those that are randomly acquired.

1.1.2 Epigenomics

Epigenetics is defined as the *"stable heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence"* (Berger et al., 2009). Mechanisms that control epigenetic regulation are DNA methylation, histone modification, non-histone binding proteins, and non-coding RNA (ncRNAs) which jointly control chromatin architecture in different cellular processes such as DNA replication, transcription, and DNA repair (Han & He, 2016). Chemical modifications of the genetic information such as methylation and acetylation of the DNA and/or DNA-binding histone proteins can alter the expression of DNA stretches. DNA methylation occurs predominantly at the cytosine residue leading to 5-methylcytosine (5-mC) and mainly within CG dinucleotide sequence clusters such termed CpG islands which often span transcriptionally inactive promoter sequences. Modification of histones further regulates the accessibility of the transcription machinery to the chromatin. Euchromatin correlates with actively transcribed chromatin and is characterized by acetylation and trimethylated histone tails of H3K4, H3K36 and H3K79 (referring to histone 3 and different lysine residues). Transcriptionally inactive

chromatin, or heterochromatin, has low levels of acetylation and high levels of methylation on the histone tails of H3K9, H3K27, and H4K20 (Li et al., 2007). Thus, such modifications are inherent regulatory mechanisms controlling gene expression and cellular phenotypes.

Epigenetic modifications can be captured at different levels using a variety of approaches whereby epigenetic information is converted into genetic information by means of biomedical methods followed by standard DNA microarray technologies or high-throughput sequencing. Interpretation of the resulting data is performed using computational and statistical analyses (Han & He, 2016). To identify DNA methylation patterns, samples need to be pretreated using endonuclease digestion followed by bisulfite conversion (Laird, 2010). After the epigenetic information has been converted to genetic information, array-based or sequencing-based technologies are used to determine the nucleotide sequence. With the rise in NGS technologies, it is possible to determine the methylation status of every cytosine in the genome using bisulfite sequencing (BS-seq) which performs bisulfite conversion of unmethylated cytosines to thymines followed by high-throughput sequencing. Whole-Genome Bisulfite Sequencing (WGBS) (Lister et al., 2009) and Reduced Representation Bisulfite Sequencing (RRBS) (Meissner et al., 2008) are among the most widely used methods capable of producing an enormous volume of DNA methylation data.

Histone modifications and chromatin-binding factors are captured by chromatin immunoprecipitation (ChIP)-based techniques, such as ChIP-seq and ChIP-chip, which take an antibody-based approach to enrich for DNA fragments. DNAase-seq is a further technology that assesses the chromatin packing by using the DNA degradation enzyme DNase I which is able to

cut DNA that is not condensed and thus exposed to the transcription-machinery (Song & Crawford, 2010).

Lastly, non-coding RNAs (ncRNA) interact with DNA, mRNA, and proteins and act as epigenetic regulators to affect chromatin conformation. A variety of technologies exist to characterize the ncRNAs such as RNA-binding protein immunoprecipitation (RIP) followed by chip or NGS sequencing, as well as UV cross-linking and immunoprecipitation (CLIP) amongst others (Han & He, 2016).

Methylation of cytosine bases within CpG islands in DNA is one of the major epigenetic mechanisms. In cancer development, tumor suppressor genes become highly methylated across promoter regions which consequently leads to the silencing of those genes. To understand the gene regulation mediated by this epigenetic mark, high-throughput profiling of DNA methylation status is crucial.

1.1.3 Transcriptomics

Transcriptomics aims to detect and classify all RNA transcripts within an organism. Gene expression starts off with the transcription of DNA into mRNA which serves as an intermediary molecule in the information network from DNA to protein. Further noncoding RNAs perform diverse regulatory functions. Within cell types or as a response to developmental stimuli as well as to environmental conditions and extracellular cues, variations in the transcriptome can arise (Hager et al., 2009). Measuring the expression of genes within an organism contributes to the understanding of how genes are regulated in certain biological processes as well as disease states. While studies of individual transcripts have been performed for decades, transcriptomics started in the 1990s with the development of expressed sequence tags (ESTs) which allowed researchers

to determine the gene content of an entire organism without the need of sequencing (Lowe et al., 2017). The highly laborious techniques of Northern blotting, nylon membrane arrays, and reverse transcriptase quantitative PCR were also used but could only detect a fraction of the transcriptome.

Sanger sequencing was applied in the sequencing-based transcriptomic technology of serial analysis of gene expression (SAGE). This and other methods were eventually overtaken by high-throughput sequencing technologies of entire transcripts.

Microarrays are capable of measuring the abundance of defined set of transcripts through hybridization of complementary probes to a glass array surface. Thousands of transcripts are analyzed simultaneously by spotted oligonucleotide arrays or Affymetrix chips (Nelson, 2001).

Today RNA-sequencing (RNA-seq) is the method of choice to determine transcripts in a system-wide manner. Influenced by the technological revolution of high-throughput sequencing, RNA-seq is able to capture and quantify all transcripts in a cell (Ozsolak & Milos, 2011). RNA-seq makes use of deep sampling of the transcriptome whereby many short fragments from the transcriptome are aligned to a reference genome or in the case of de novo assembly, aligned to each other by means of computational tools. Advantages of RNA-seq over microarray analysis are the lower input RNA material, the ability to detect a more dynamic range of expression, the detection of all sequences as well as the low background. There are also advances in RNA-seq which allow for quantitation of RNA transcripts at the single cell level (single-cell RNA-sequencing).

1.1.4 miRNomics

microRNAs (miRNAs) are a family of small, endogenous, evolutionarily conserved RNA molecules with a length of 21-25 nucleotides. They have been shown to be implicated in the negative post-transcriptional regulation of gene expression. Initially, miRNAs were identified in 1993 and described in the worm *Caenorhabditis elegans* with the miRNA *lin-4* being the founding member (He & Hannon, 2004).

miRNAs are implicated in the regulation of development and disease by mechanisms such as RNA interference (RNAi) whereby the miRNA binds through base-pairing to its target mRNA and silences it. miRNAs are generated in process that involves the transcription by RNA polymerase II into a long primary miRNA of several hundreds of nucleotides which is capped and polyadenylated and subsequent processing into a precursor RNA (pre-miRNA) and after export to the cytoplasm processed to yield the mature miRNA (Bhaskaran & Mohan, 2014).

Aberrant expression of miRNA is a hallmark of cancer and miRNA expression profiling has shown that miRNAs are implicated in tumor development, progression, and therapy response. Also, miRNAs themselves can function as oncogenes or tumor suppressor genes (Iorio & Croce, 2012). Since miRNAs play pivotal roles at the post-transcriptional level in development and disease the abundances of miRNAs are assessed through high-throughput sequencing. The field of miRNomics refers to the genome-wide profiling of miRNA expression patterns (Yousef & Allmer, 2014).

1.1.5 Proteomics

Investigating the total number of proteins expressed by a cell, within a tissue, or in an organism is referred to as the study of proteomics. The proteome of a cell is described as the total protein content and characterized by localization, interactions, post-translational modifications, and turnover at a certain time point (Aslam et al., 2017). Marc Wilkins coined the word proteomics in 1996 and described it as *"PROTein complement of a genOME"* (Wilkins, Sanchez, et al., 1996). As with transcripts, the proteome is highly variable and varies over time, within species, or due to environmental factors. mRNA transcripts encode 20 amino acids, the sequence of which in the process of translation is transformed into an amino acid chain. Such chains are further folded into different conformations and are post-translationally modified by phosphorylation, acetylation, and/or glycosylation, which further adds to the complexity of the proteome. In contrast to DNA which resides in the nucleus and mRNA which is produced in the nucleus and later transported into the cytoplasm, the final destination of proteins can greatly vary. Proteins have different subcellular localizations beyond the nucleus and the cytoplasm, including extracellular and other organelles (Harper & Bennett, 2016). Proteomics is an important methodology in understanding gene function in health and disease i.e., for disease diagnosis or monitoring disease progression, as well as drug development.

The proteome can be studied using various techniques. Chromatography-based techniques such as affinity purification allow for enrichment and purification of proteins and their interaction partners. Analysis of specific proteins is done using enzyme-linked immunosorbent assay (ELISA) and Western blotting, the latter assay uses protein mixtures in a two-dimensional separation using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Aslam et al., 2017). In

quantitative proteomics, analyses such as Isotope-coded affinity tag (ICAT), stable isotope labeling with amino acids in cell culture (SILAC), and isobaric tag for relative and absolute quantitation (iTRAQ) experiments are used. The progression of high-throughput technologies has also accelerated data generation and curation of databases in this field. Protein microarrays or protein chips are such a high-throughput technique capable of detecting proteins in small samples. Three classes of protein microarrays can be distinguished: analytical protein microarray, functional protein microarray and reverse-phase protein microarray (RPPA). While analytical arrays detect specific proteins directed to antibody capture, functional assays allow for various interaction studies such as protein-DNA, protein-RNA, and protein-drug. The RPPA uses antibodies against the protein of interest on cell lysates which have been arrayed on nitrocellulose slides (Aslam et al., 2017). Although these protein chips are able to measure proteins in a semi-quantitative way, mass spectrometry approaches have taken over because they are high-throughput techniques capable of assessing the proteomic content of a cell.

1.1.6 Metabolomics

Metabolites are small molecules including lipids or hormones that are products and substrates of the host metabolism and responsible for mediating essential cellular functions such as energy production, epigenetic regulation, and signal transduction. These molecules are not only derived from the host but can also originate from microorganisms or diet. In total those molecules make up the metabolome. A variety of factors contribute to the variability in metabolic phenotypes such as the genetic make-up, environmental factors, or lifestyle habits. Since metabolites have far-reaching implications in the cell and within an organism, metabolomics aims at identifying and understanding metabolites and associated metabolic pathways to gain understanding in their

physiological roles (Johnson et al., 2016). Mass spectrometry is the analytical platform to assess metabolites in a high-throughput manner with high sensitivity and reproducibility.

1.2 *Omics* – technologies

Two major technologies are behind the driving force in the omics field: Next-generation sequencing (NGS) and mass-spectrometry (MS). The following sections give an overview of how these technologies work and what type of data they are generating.

1.2.1 Next-Generation Sequencing (NGS) based approaches

Since its first adoption in the mid-2000s, NGS technologies have evolved in terms of read lengths, capacity, and reduced cost. Today it is possible to sequence an entire human genome within a day and at a cost of around US $1000. This makes sequencing an interesting tool for clinical applications. In contrast to traditional Sanger sequencing, NGS approaches yield vast amounts of data. However, this data can be prone to a higher error rate (~0,1-15%) while sequence reads are shorter (35-700 bp) than the traditional reads generated using Sanger sequencing. NGS technologies capable of producing longer reads come at the expense of time, cost, and throughput (Goodwin et al., 2016).

In general, NGS refers to massive parallel sequencing of millions of DNA copies within a single reaction. To determine the sequence of DNA bases, single molecule templates are fixed to a surface followed by clonal amplification (Brown, 2015). There are two ways of short-sequencing technologies leading to clonal template generation: Sequencing by ligation (SBL) and sequencing by synthesis (SBS). In the SBL approach, the enzyme DNA ligase is used to determine the

underlying DNA sequence. DNA ligase can not only join the ends of DNA molecules but is also capable of detecting mismatches in base-pairing. Initially, DNA is fragmented, and an anchor sequence attached. This allows probes to hybridize with the pool of oligonucleotides that have been labeled with fluorescent tags according to the four nucleotides. DNA ligase catalyzes the joining of fluorescent-labelled probe, to the primer and the template DNA leading to a signal if sequences match. After the removal of the anchor primer, the cycle restarts using a different query primer. Two platforms that use SBL are Support oligonucleotide ligation detection (SOLiD) and Complete Genomics (Goodwin et al., 2016).

The SBS technology incorporates four fluorescently labelled nucleotides one at each cycle to the nucleic acid chain within the millions of clusters on the flow cell. The incorporation of theses labelled nucleotides results in the termination of the polymerization followed by imaging of the fluorescent signal and enzymatic cleavage of the terminator-bound deoxynucleoside triphosphate (dNTP). Illumina Sequencers exploit this technology of sequencing by reversible terminator and accounts for the most used technology today. SBS and SBL are part of the second-generation sequencing technologies and third-generation technologies have already arrived. Nanopore DNA sequencing is such a third-generation sequencing technique by which DNA is passed through a nanopore and the thereby altered electrical conductance is measured (Song & Crawford, 2010).

1.2.2 Mass-spectrometry based approaches

Mass spectrometry (MS) identifies, characterizes, and quantifies proteins within complex samples and at an increasing sensitivity. By measuring the mass-to-charge ratio *(m/z)* and the subsequent calculation of the exact molecular weight, MS can identify unknown compounds by molecular weight determination. Three components make up a mass spectrometer: ionization

source, mass analyzer and ion detection system. In the first step, molecules are converted to gas-phase ions to make them move and be manipulated by electric and magnetic fields. Routinely, two techniques are used in this step – electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). Next, the mass analyzer sorts and separates the ions based on the mass-to-charge *(m/z)* ratio. Lastly, the ion detection system measures the *m/z* ratios which can be visualized as mass spectrum, a diagram with *m/z* ratios plotted against their intensities (Domon & Aebersold, 2006). The *m/z* ratio is further used for molecule identification in dedicated databases.

A variety of instruments and methods have been developed that allow for accurate mass determination and protein characterization in many different applications. High Speed and accuracy in measuring fragmentation spectra and in the identification of proteins are achieved by using tandem mass spectrometry (MS/MS).

1.3 Multi-omics – integration of multiple omics data

High-throughput analysis techniques and the resulting high-dimensional data are ubiquitous in biology. While analysis of single omics data is standard when addressing biological questions, the joint analysis of multiple omics data has remained a challenge. To study biological processes holistically and to reduce the experimental and biological noise, integrating different complementary omics data is a necessity. The goal of a multi-omics approach should be a model that is able to capture signals shared by all omics data but also those that arise from the complementary analysis across all layers.

1.3.1 Online data resources

The curation and storage of omics data in public databases has grown exponentially in the past decade and accessing those has data been eased. National and international consortia have profiled various samples with a multitude of molecular arrays enabling researchers to find patterns in those data pertaining to their research interest.

The Cancer Genome Atlas (TCGA) a joint effort between National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) started in 2006 and aimed at molecularly characterizing different cancer types. Thus far, the TCGA produced over 2.5 petabytes of data describing 33 different tumor types including 10 rare cancers and contains cancerous as well as healthy tissue sets from more than 11,000 patients. The aims of this gigantic effort involving more than 20 collaborating institutions across the US and Canada are the improved understanding of the genomic basis of cancer, to contribute to the classification of tumor subtypes, as well as drive the discovery of therapeutic targets and drug development (*The Cancer Genome Atlas Program - National Cancer Institute*, 2018). The 33 cancer types selected for analysis by the TCGA program are listed on the homepage and have been chosen according to specific criteria such as poor prognosis, impact on health, and the availability of samples. In all cases patients have given consent for the use of their data. The data collected over the last decade comprises omics data representing genomic, epigenomic, transcriptomic and proteomic data (Das et al., 2020). To quantify single nucleotide variants (SCNV), mRNA and miRNA expression as well as DNA methylation, high-throughput NGS and RPPA assays as well as MS technologies were employed.

While TCGA stores data from the genome, epigenome and transcriptome, the proteome data from RPPA assays and generated from the same patient have been deposited in The Cancer

Proteome Atlas (TCPA) database. Since RPPA assays are limited in terms of their low throughput character and dependent on antibody specificity, some TCGA samples were subjected to LC-MS/MS for quantification and the data stored in the Clinical Proteomic Tumor Analysis Consortium (CPTAC) database. Currently, proteomic, phosphoproteomic, and glycoproteomic data from several TCGA cancer cohorts can be retrieved via the CPTAC data portal.

Although there are numerous other data portals containing single and multi-omics data for cancer research such as COSMIC (Catalogue Of Somatic Mutations In Cancer) or ICGC (International Cancer Genome Consortium), the TCGA database stands out because of the multiple omics profiles available from the same patients. Moreover, those data can be accessed, visualized and downloaded via a multitude of different portals such as Genomic Data Commons (GDC), Firebrowse and cBioPortal (Das et al., 2020).

LinkedOmics is another multi-omics database that contains omics data and corresponding clinical data for 32 of the currently 33 cancer types from a total of 11,158 patients of the TCGA project. (Vasaikar et al., 2018). It not only stores more than a billion data points but also allows for comprehensive analysis of these data using three web analysis applications. LinkFinder, LinkCompare, and LinkInterpreter modules explore the associations between clinical or molecular attributes across different cancer types, compare identified associations, and perform pathway and network analyses to understand biological significance, respectively. The Firehose of the Broad Institute (*Broad GDAC Firehose*, n.d.) served as data source for clinical, genomic, epigenomic, and transcriptomic data for the 32 cancer types from TCGA. Proteomic data was downloaded from the CPTAC data portal. Curated data was preprocessed and rows containing values of type “NA” of greater than 60% or zeros greater than 95% were removed, followed by normalization and

storage as attribute by sample in matrix files. The benefit of LinkedOmics over other data access portals is the straight-forward retrieval of the different single omics data in a convenient format which stores the samples and associated features in a compressed manner. Also, general preprocessing steps and normalization of the data have been performed making the data readily available for multi-omics integration.

1.3.2 Challenges of multi-omics research

Generation of biological insights from data is divided into steps of data generation, data processing, data integration and data analysis. Within the first step, biological assays are used to generate for example sequence reads, spectra from MS or imaging data. Those data are then preprocessed using manual curation and computational methods such as aligning sequences to the genome. In the third step, various data types are integrated in a meaningful way making use of specialized algorithms. Finally, the data need to be interpreted to gain novel insights which requires a deep understanding of the underlying biological field (Palsson & Zengler, 2010).

Several challenges arise when integrating multiple omics data sets such as the high dimensionality of the data, its heterogeneity, and missing values in biomedical data. The curse of dimensionality pertains to the *"large p, small n problem"* in Machine Learning (ML) and refers to the fact that the number of features within a data set are substantially higher than the number of samples (Altman & Krzywinski, 2018). This can be problematic and lead to overfitting, i.e., the model does well on the training data but has a poor generalization on new unseen data. Therefore, it is common to use dimensionality reduction techniques such as feature extraction or feature selection. The former reduces the high-dimensional space by projecting it to a lower dimensional space using linear methods like Principal Component Analysis (PCA) or non-linear methods like

t-distributed stochastic neighbor embedding (*t*-SNE). Feature selection aims at reducing the feature space by selecting only a subset of relevant features from the original data through filtering or embedding methods (Mirza et al., 2019).

The heterogeneity of biological data is another challenge in the joint analysis of multi-omics data. The difference in the number of variables, and the distributions thereof, as well as the data types ranging from binary data for e.g., mutation data to continuous data for gene expression values are inherent to the type of biomolecule analyzed. Moreover, each data is processed in different ways in terms of scaling and normalization adding up to the heterogenous nature of the data to be integrated. Gene expression, DNA methylation, miRNA expression, copy number alterations, protein quantities and clinical data are diverse on their own and provide one of the greatest challenges for multi-omics integrated analysis (Mirza et al., 2019).

Missing data from high-throughput platforms are also problematic for integrated multi-omics models. Such missingness can arise from low sequencing coverage of NGS, low sensitivity in protein detection of RPPA, and flawed protein and metabolite measurements by MS. This leads to sparse models that have integrated omics data for some but not all assay types leading to samples with missing values for certain data modalities. To deal with missing data, imputation using statistical methods or supervised ML algorithms can be employed. Other approaches comprise Deep Learning methods or integrative imputation strategies (Mirza et al., 2019).

1.3.3 Algorithms for the integration of different omics data types

Various integrative approaches have been proposed for the joint analysis of different omics data. Generally, such strategies can be separated into horizontal and vertical data integration methods whereby the former integrates single omics layer information across different studies and

the latter aims at integrating multiple different omics layers for the same sample (Das et al., 2020). The Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium analyzed genomic data across a variety of different tumor types and found that the majority of tumors (91%) harbored well-defined driver mutations (Campbell et al., 2020). While this horizontal integrative approach is important for investigating the origin and evolution of tumorigenesis, it lacks molecular associations arising from genotype to phenotype layer that may identify new features for biomarkers or stratification of patients. Vertical integration on the other hand can capture the multidimensionality of cancer across molecular layers and can be performed either in a post-analysis fashion or through an integrated approach. The former strategy analyzes each single omics data set individually and considers only overlapping features for interpretation. It is evident that this approach fails in identifying latent molecular signatures present across multiple omics layers (Das et al., 2020). Joint analysis of multiple omics data makes use of algorithms for an integrated vertical data integration which can be broadly classified into three categories: Bayesian methods, Network-based methods and joint Dimensionality Reduction (jDR) approaches (Cantini et al., 2021). Whereas Bayesian methods assume data distribution and dependency when building a model and the network-based approach infers relationships of features and combines those in a network, the jDR methods project the omics data into a low-dimensional latent space. Not only within the context of biological high-throughput data have jDR approaches emerged as well-performing in solving high-dimensional problems (Cantini et al., 2021).

Data from single omics analyses can provide a snapshot of the cells or tissues state in time reflecting active genes, methylation patterns and detectable proteins. In fact, the biological state of a cell or tissue is determined by generic processes such as proliferation to cell-specific processes

such as immunological activity. Thus, there is a convoluted mix of various active biological signals which jDR can deconvolute. Many different jDR techniques differ in their underlying mathematical formulation. PCA, Factor analysis, Gaussian latent model, matrix-tri-factorization and others represent the mathematical structures able to combine more than two omics modalities (Cantini et al., 2021). iCluster (Shen et al., 2009), Multi-Study Factor Analysis (MSFA) (De Vito et al., 2019) and Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018) are different varieties of Factor analysis while Multiple co-inertia analysis (MCIA) (Bady et al., 2004) and Joint and individual Variation Explained (JIVE) (Lock et al., 2013) algorithms represent extensions of PCA. Most of the algorithms make different assumptions of the underlying data but all do require a match between samples of the different omics data. In a jDR benchmarking study, Cantini et al., compared 9 different jDR approaches on simulated data as well as on TCGA data. Among the best-performing methods was MOFA which outperformed other methods in clustering of retrieved latent factors to group cancer patients (Cantini et al., 2021).

1.4 Multi-omics analyses in cancer research

The human organism contains about 20,000 – 22,000 protein-coding genes, ~30,000 mRNA transcripts, 2300 miRNAs, between ~20,000 – several million different proteins, and 114,000 metabolites (Biswas & Chakrabarti, 2020). The proper functioning of these biomolecules at the individual molecule layer as well as the interrelation thereof is the basis for a healthy life. Few diseases are attributed to the malfunctioning of a single biomolecule, most often it is the interplay between several factors as disease develops and progresses. Cystic fibrosis is a disease which is caused by a mutation in a single chloride channel and which can be researched upon focusing on the single gene (Welsh & Smith, 1993). Complex diseases such as cancer develop through the

combination of genetic and environmental factors with a non-distinctive molecular disease etiology involving many different biomolecules. Thus, the integration of data types generated from the different biological layers is key to further the understanding of complex diseases such as cancer.

1.4.1 Cancer: a complex disease

Cancer is a general term for a group of diseases that can impact almost every part of the body. Neoplasm or malignant tumor are synonymous terms for cancer. Albeit diverse in its location and molecular etiology, the common feature of cancer is the rapid and limitless growth of abnormal cells which can also invade neighboring tissues through a process called metastasis. According to the World Health Organization (WHO) cancer is the leading cause of death worldwide and attributed for about 10 million deaths in 2020 (Ferlay et al., 2021). The most common cancers in 2020 were breast cancer, lung cancer, and cancer of the colon and rectum. The highest death rates occurred in patients with lung cancer, cancer of the colon and rectum followed by liver cancer (Ferlay et al., 2021). The intricate interplay between a person's genetic factors and environmental agents are responsible for the transformation of normal cells into tumor cells. Such external factors are physical carcinogens such as radiation, chemical agents such as tobacco smoke and biological carcinogens such as viruses.

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 (Hanahan & Weinberg, 2000, 2011). 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. Proto-oncogenes and tumor-suppressor genes are two broad classes of genes that when mutated play a key role in the induction of cancer. Such genes encode proteins that are implicated in regulating growth and proliferation such as the RAS oncogenes and the tumor suppressor p53 (Kontomanolis et al., 2020).

The increased rate of change in the number of chromosomes as well as their structure leads to intratumor heterogeneity and is known as chancer chromosomal instability (CIN). This phenomenon is observed in many solid tumors and has been related to resistance to chemotherapy as well as poor outcomes (Burrell et al., 2013).

1.4.2 Pancreatic Ductal Adenocarcinoma

Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive malignancies with a 5-year survival rate of less than 10%. The difficulty in early diagnosis and the limited response to treatment are attributed to a considerable heterogeneity among patients and within primary tumors (Ryan et al., 2014). The rate of PDAC cases is increasing and it is projected that PDAC will become the second leading cause of cancer mortality before 2030 (Rahib et al., 2014). Although, the estimated time of development of PDAC is more than 20 years, it is mostly detected at a metastatic stage when it is too late to intervene. Tumorigenesis is driven by genetic mutations followed by increased age, chronic pancreatitis, smoking, obesity, and type 2 diabetes. Moreover,

PDAC has been shown to be resistant to drugs, contributing to its fatality (Storz & Crawford, 2020).

A number of mutations and SCNVs have been identified by whole-exome sequencing studies impacting the function of oncogenes such as *KRAS* and tumor suppressor genes including *TP53, SMAD4* and *CDKN2A* (Jones et al., 2008). Additional alterations in genes involved in axon guidance and DNA repair in a subset of PDACs has added to the mutational landscape in this tumor type. Furthermore, chromosomal rearrangements have been detected in PDAC tumor cells, driving cancer progression.

Molecular changes are accompanied by morphological changes at the tissue level with the formation of pancreatic intraepithelial neoplasia (PanIN) precursor lesions. As the cancer grows, the tissue surrounding the tumor cells also undergoes morphological changes leading to a favorable tumor microenvironment. Specifically, cells surrounding the tumor are activated to produce fibrosis and resulting in desmoplasia. This stromal effect subsequentially isolates the tumor cells like a barrier and is responsible for poor response to chemotherapy as well as immune cell infiltration (Storz & Crawford, 2020).

Although PDAC is a highly aggressive cancer, its neoplastic cellularity i.e. the amount of tumor cells within the primary tumor is 5% - 20%, a relatively low percentage compared to other tumor types (Raphael et al., 2017). Thus, historically, researchers focused upon tumor samples with higher tumor cellularity through dissection methods or enrichment procedures. This however led to an underrepresentation of the analysis of PDAC cancers with low tumor cellularity in studies. Moreover, the effect of low tumor cellularity complicated the analysis of less abundant

biomolecules such as lncRNAs, miRNAs, DNA methylation patterns, and protein abundances which have been confounded by tumor purity (Raphael et al., 2017).

PDAC tumors are classified based on morphologic characteristics as well as the progressive behavior of cancer cells and associated patient outcomes. Underlying genetic driver mutations are not sufficient to robustly group patients for treatment pointing to a more complex molecular and metabolic landscape as well as epigenetic mechanisms (Storz & Crawford, 2020). The four-stage classification system into squamous, immunogenic, pancreatic progenitor, and aberrantly differentiated exocrine types is based on gene expression and data clustering. Collisson et al., (2011) made use of combined transcriptional profiling including primary tumor samples and integrating information from human and mouse pancreatic cancer cell lines leading to three subtypes: classical, quasimesenchymal and exocrine-like. Two distinct tumor subtypes, basal-like or classical were proposed by Moffitt et al., (2015) in a gene expression study involving primary tumor as well as metastatic and healthy samples.

1.4.3 Multi-omics in cancer research

Unraveling the molecular etiology of tumorigenesis is of importance to identify the alterations across multiple molecular layers. Using NGS and MS technologies, alterations such as somatic mutations at the genome layer, altered methylation patterns at the epigenome layer, differential expression in mRNAs at the transcriptome layer and differential abundances of proteins at the proteome layer can be detected and quantified (Chakraborty et al., 2018). The goal of personalized onco-medicine is to unravel the combined action of such alterations from the genotype to the phenotype and subsequently develop personalized treatment plans based upon each patient's tumor composition.

Bioinformatics and artificial intelligence (AI) provide powerful computational and mathematical frameworks capable of reducing the fuzziness and randomness in biological data as well as build robust platforms for data mining and data analysis. Precision medicine refers to the use of diagnostic tools and treatments targeted to the needs of the individual patient and based on their genetics or other biomarkers (Ramaswami et al., 2018). With the wealth of information that is recorded from patients, computational methods play an important role in integrating multi-omics data for the identification of molecular disease mechanisms, diagnostic and prognostic markers as well as monitoring patient's response to drug treatments (Patel et al., 2020).

A comprehensive review by Nicora et al., (2020) describes various approaches of researchers combining different omics data and making use of a variety of algorithms to study cancer pathways, identify patient subgroups or contribute to drug discovery in cancer research.

1.4.4 Multi-omics studies addressing PDAC heterogeneity

Several attempts have been made in addressing PDAC heterogeneity to infer intertumor and intra patient heterogeneity. Traditional classification is based on histopathology and has led to the identification of two to six PDAC subtypes (Roy et al., 2021). While it is known that genetic mutations initiate tumorigenesis those alterations are unable to capture the vast heterogeneity observed between 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 PDAC subtypes were identified in a study involving a multi-omics integration analysis based on genomics, epigenomics and

transcriptomics data on data sets from TCGA and Gene Expression Omnibus (GEO) (Kong, et al., 2020). Recently, Roy et al., (2021) have employed an integrated unsupervised clustering approach to the PDAC data set from TCGA involving gene expression and methylation data and obtaining five relevant subtypes 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. Lack of evidence for generic number of clinical subtypes makes it a necessity to study PDAC heterogeneity with focus on post-genetic mechanisms.

1.5 Objective of study

In this research, data from TCGA is used to infer the major sources of PDAC heterogeneity within the samples of the PDAC cohort (also referred to as PAAD) through the integration of multiple different omics data. 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 address PDAC disease etiology. The objective of this study is to integrate six different data modalities from the PAAD cohort using the Multi-Omics Factor Analysis (MOFA) framework. Specifically, data from SCNV, methylation, mRNA sequencing, miRNA sequencing, and RPPA assays are used to build a model capable of investigating major sources of heterogeneity underlying PDAC.

Figure 1 shows the workflow that was undertaken in this research.

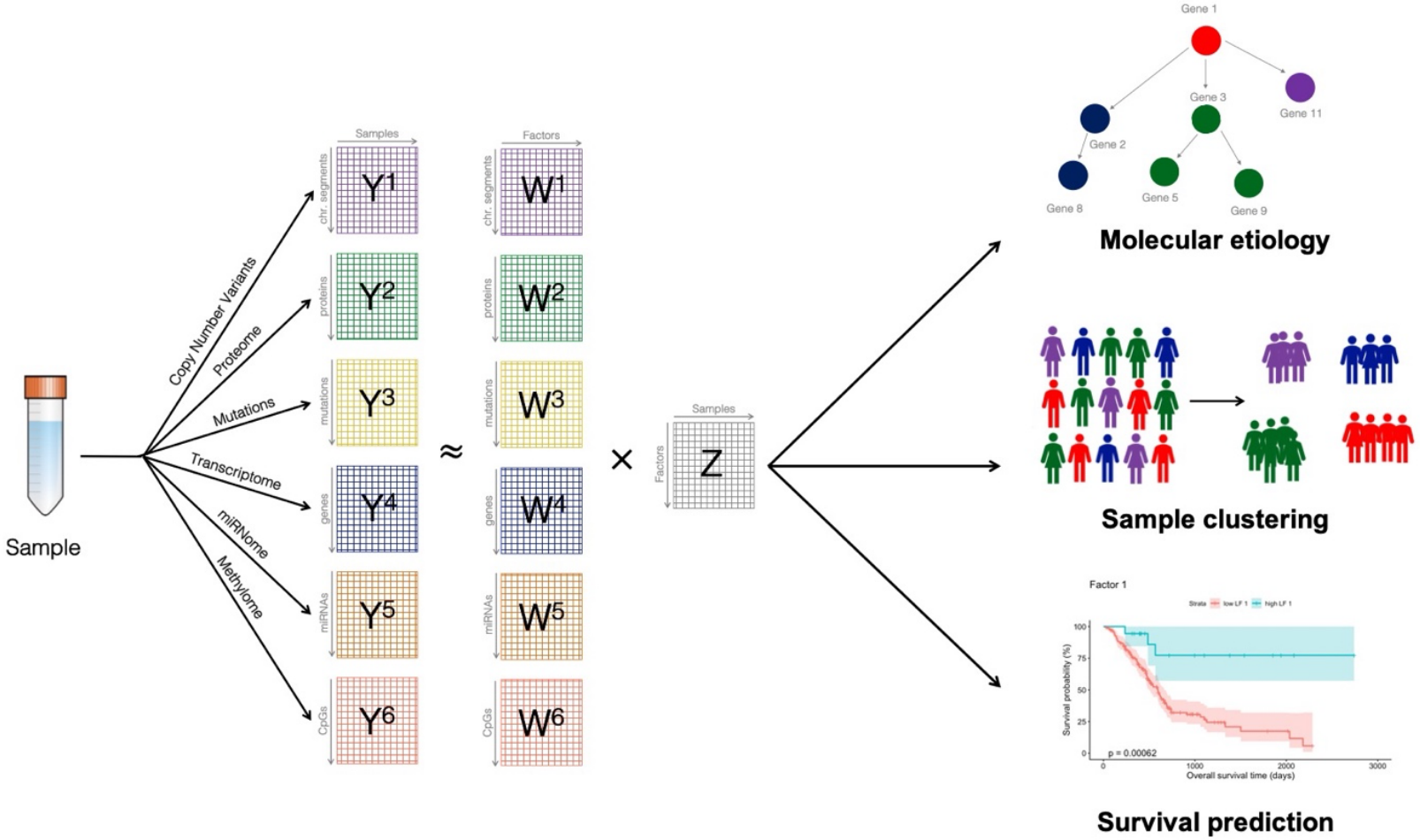


Figure 1. Research Outline. Multi-Omics Factor Analysis framework applied to PDAC data set to uncover molecular etiology, perform patient clustering, and predict overall survival (Own illustration modified from: Cantini et al., 2021).

Using the factor analysis framework MOFA, the omics data from the various assays and represented as matrices *Y* are factorized into the product of factor matrix *Z* and weight matrices *W*. The resulting matrices representing the lower dimensional latent space will be used for further analyses to uncover the molecular mechanisms present in the most dominant factors. Clustering of samples in latent space will be carried out to infer if a classification upon distinct subtypes is feasible. Employing gene set enrichment analysis will further identify metagenes and pathways adding to the bigger picture of PDAC etiology. Lastly, a clinical model is built, to anticipate any effects in survival based on the retrieved factors.

In performing such an integrated study and considering information from the genome to the proteome, a better molecular classification of this cancer type is anticipated which provides the basis for patient stratification, the discovery of new prognostic biomarkers and the response to treatment options.

CHAPTER 2 METHODS

2.1 Data availability

The data from the TCGA PDAC cohort, referred to as PAAD cohort, and comprising multiple omics assays was downloaded from LinkedOmics in the form of individual matrix data files (Vasaikar et al., 2018): http://linkedomics.org/data_download/TCGA PAAD/

While the original raw data were generated by TCGA, the pre-processing and normalization steps were carried out by the Linked Omics team. This section briefly explains the data generation and processing that have been applied to the omics data from the TCGA PDAC cohort. Exploratory data analysis and descriptive statistics on data matrices obtained from LinkedOmics was done in Python 3.9.0. (Rossum, 2007) using pandas (McKinney, 2010), numpy (van der Walt et al., 2011), matplotlib (Hunter, 2007) and, plotnine (Kibirige et al., 2021) libraries. The data matrices were further loaded into RStudio and combined to yield a long matrix in the format of sample, feature, value, and view (i.e., the type of omics data). Subsequent analyses were performed in R using RStudio version 1.4.1717 (Team, 2021).

2.1.1 Clinical data

Clinical data includes demographic patient attributes as well as survival and tumor staging information. The TNM pathologic staging system is the most widely used cancer staging system and describes the amount and spread of a tumor in a patient's body where T describes the size of

the original tumor and its nearby spread, N describes the spread to regional lymph nodes, and M is an indicator of metastasis i.e. the spread of cancer to other distant body parts (*Cancer Staging - National Cancer Institute*, 2015).

While the TNM staging provides great detail on the size and spread of cancer, another pathologic staging system has been recorded in the clinical data (*Cancer Staging - National Cancer Institute*, 2015). Table 2 explains the different staging systems as well as the residual tumor classification present in the PAAD data.

Table 2. Tumor Staging used in clinical patient data of PAAD cohort.

Tumor Staging	Stage	Description
TNM	T0	Primary tumor is not found
	T1, T2, T3, T4	Primary tumor size with 1 being the smallest size and 4 indicating large tumors and invasion into nearby tissues
	N0	No cancer in nearby lymph nodes
	N1, N2, N3	Cancer containing lymph nodes with higher numbers indicating more affected lymph nodes
	M0	Cancer has not spread to distant body parts
	M1	Metastasis. Cancer has spread to distant body parts
I-IV	0	Presence of abnormal cells without spread to nearby tissue
	I, II, III	Cancer is present and the higher the stage the larger the tumor and the more it has spread to nearby tissues
	IV	Metastasis to distant body parts
Residual tumor (R)	R0	No residual tumor
	R1	Microscopic residual tumor
	R2	Macroscopic residual tumor

2.1.2 Mutation data

Mutations discovered by DNA sequencing were analyzed using the MutSig2CV software tool. MutSig stands for "Mutation Significance" and CV for "covariates". This tool identifies genes that are mutated more often than expected just by chance or background mutation processes (Lawrence et al., 2013). Values were binarized using 0 and 1 for absence and presence of a mutation, respectively.

2.1.3 Somatic Copy Number Variants (SCNV) data

Affymetrix SNP 6.0 array was used to determine genomic regions with amplifications or deletions. To identify likely driver SCNVs, the GISTIC (Genomic Identification of Significant Targets in Cancer) algorithm was used to infer the frequency and amplitude of observed events (Mermel et al., 2011).

Four data matrices are available for this data modality: SCNVs were recorded in PAAD at the focal and gene level, respectively and analyzed using a thresholded or a log-ratio scale. Only protein-coding genes are kept, thresholded, and categorized using a noise cutoff of 0.3. Genes are categorized as a "loss" (-1) if focal CNV values are smaller than -0.3. Genes are categorized as a "gain" (+1) if focal CNV values are larger than 0.3. And genes with values between -0.3 and 0.3 are categorized as "neutral" (0).

2.1.4 Methylation data

The Illumina Infinium assay was used by TCGA consortium to generate the methylation data. Briefly, this technology makes use of two probes, one methylated and one unmethylated, to measure the intensities of individual methylated and unmethylated CpG sites respectively (Estécio et al., 2007). Based upon the intensities of both probes, the methylation level is inferred. The percentage of methylation at gene and loci level is calculated as Beta-values and ranges between 0 and 1 where 0 indicates a completely unmethylated site and 1 denotes that every copy of this site is methylated (Du et al., 2010). Specifically, it is calculated by taking the ratio of the methylated intensity and the overall intensity for a locus (sum of methylated and unmethylated intensities). Equation 1 denotes the calculation of Beta-values where $y_{i,methy}$ and $y_{i,unmethy}$ refer to the intensities measured by the i^{th} methylated and unmethylated probes, respectively. The α parameter is a constant value that regularizes Beta-values when intensities are low. Beta-values were recorded at CpG site and gene-level.

$$Beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}$$

Equation 1. Calculation of Beta-Value.

2.1.5 microRNA data

miRNA expression at the gene level for each tumor sample and its normalized expression was recorded using the RPM (Reads per million mapped reads) value metric (Equation 2).

$$RPM = \frac{\text{Number of reads mapped to gene } \times 10^6}{\text{Total number of mapped reads}}$$

Equation 2. Calculation of RPM value.

2.1.6 Transcriptome data

RNA sequencing reads were generated on an Illumina HiSeq platform and the abundance of genes or transcripts expressed as reads per kilo base per million mapped reads (RPKM). The values were further normalized and log2 transformed.

$$RPMK = \frac{\text{Number of reads mapped to gene } \times 10^3 \times 10^6}{\text{Total number of mapped reads } \times \text{ gene length in bp}}$$

Equation 3. Calculation of RPKM value.

2.1.7 Proteomic data

Proteomic data of the PAAD cohort was generated from RPPA assays which are a high-throughput antibody-based technique. The obtained data was further processed in four steps and protein expression quantified using a "Supercurve Fitting" approach (M. Sun et al., 2015). This analysis was carried out by the Department of Bioinformatics and Computational Biology at MD Anderson Cancer Center who reported the results to TCPA.

2.2 Software availability

Multi-Omics Factor Analysis V2 (MOFA+) is an open-source software that is implemented in Python but recommended to be run in R (Argelaguet et al., 2020). The stable release of MOFA+ was installed from Bioconductor in R.

2.3 Feature selection

The 5000 most highly variable features from RNA-seq and SCNV data were selected using the VST algorithm built in Seurat (Hao et al., 2020). The function `FindVariableFeatures` was applied to the generated Seurat objects and the `selection.method` "vst" chosen. VST fits a line to the log (variance) and log (mean) relationship and subsequently standardizes feature values based on the fitted line and using the observed mean and the expected variance.

For the methylation data, the top 1% of features with the highest scores was retrieved using the `TopFeatures` function in Seurat.

2.4 Multi-Omics Factor Analysis (MOFA)

Each of the M omics matrices $Y^1 \ldots Y^M$, is a $N \times D_m$ dimensional matrix with N the number of samples and D_m the number of features in data matrix m. MOFA performs a decomposition of these matrices as described in Equation 4 where Z denotes the factor matrix (common to all matrices) and W^m denotes the weight matrices for each data matrix m (referred to as "view"). ε^m denotes a noise term which specifies the structure of each data modality. For continuous data matrices, RNA-seq, methylation, SCNV, protein and miRNA data, a Gaussian noise model was used. For the binary mutation data, a Bernoulli model was applied. MOFA can combine different noise models and thereby considering a diversity of data types.

$$Y^m = ZW^{mT} + \varepsilon^m \qquad m = 1, \ldots, M$$

Equation 4. Matrix factorization by MOFA (Argelaguet et al., 2018).

Formulated as a probabilistic Bayesian framework the model further uses sparsity priors, which enable Automatic Relevance Determination (ARD) of the factors.

2.5 Model training and selection

Model training was done on the previously assembled MOFA object which contains the selected data modalities ("views") stacked with samples as rows and feature, value, and view as columns. MOFA options can be defined within the data options, model options, and training options parameters. Within the data options, the argument `scale_views` was set to `TRUE` in some models to scale the views to the same total variance. The number of factors and the likelihoods were specified in the model options as 10 or 15 factors and Gaussian and Bernoulli, respectively. Within the training options, maximum number of iterations was the default 1000, but the convergence mode was set to "slow". All other parameters were used as default.

To build a suitable model for addressing the research question, an incremental integration approach using MOFA was applied, as well as different model parameters. The architecture of the various PDAC MOFA models is described in Supplemental Table A (Appendix).

Selecting the best performing model, a comparison was done based on the Evidence of Lower Bound (ELBO) metric. The metric refers to the likelihood function evaluated at a fixed parameter θ (Equation 5) which should be maximized in model selection (Cherief-Abdellatif, 2019).

$$evidence := \log p(x;\ \theta)$$

$$ELBO := E_{Z\sim q}\left[log\frac{p(x,Z;\theta)}{q(Z)}\right]$$

Equation 5. ELBO Likelihood Function fixed at θ and ELBO Equation.

Based on this quantity, the model p and the parameter θ when the marginal probability of the observed data x is high was chosen. Thus, models with higher values of $\log p(x;\ \theta)$ indicate a better model fit. The function `compare_models` was used to compare different models by calculating and plotting the ELBO quantities.

2.6 Variance decomposition and inspection of feature weights

The R package, *MOFAtools,* and its comprehensive functions for semi-automated analysis was used for characterization and interpretation of latent factors. Explanation of variance by each factor and in each view was done by calculating the fraction of variance explained (R^2) using the `plot_variance_explained` function. Heatmap plots were used to visualize values of input features in a low-dimensional space.

Line plots (`plot_top_weights`) were used to inspect the individual feature matrices and the top 10 factors within each data modality plotted. Each feature weight is scaled by the absolute value between 0 and 1. It is of note, that feature values of different views are not directly comparable.

2.7 Gene set enrichment analysis (GSEA)

GSEA was performed using `run_enrichment` function of MOFA on gene sets with positive and negative weights of mRNA and methylation data. From the *msigdbr* R library (Dolgalev, 2021), the Gene Ontology gene set comprising 7481 gene sets from the GO Biological Process ontology was used (C5, GO:BP) (Liberzon et al., 2011). *p*-values from the parametric statistical test were adjusted for multiple testing for each factor using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) to control for the false discovery rate (FDR) at a threshold of 1%.

2.8 k-Means clustering

k-Means clustering on factors 1, 2 and, 3 was done with k=6 clusters and the `cluster_samples` function from *MOFAtools* package.

2.9 Association analysis

The correlation of clinical metadata with all 15 MOFA factors was done using the MOFA `correlate_factors_with_covariates` function with all available clinical covariates. log10 adjusted *p*-values were plotted in a heatmap diagram. A *p*-value threshold of 0.05 was used.

2.10 Cox proportional hazards model

The impact of MOFA factors on overall survival was modeled using the Cox proportional hazards model by employing the `coxph` function from the *survival* package (Therneau, 2021; Therneau & Grambsch, 2000). As response variable in the Cox model was overall survival chosen

2.11 Kaplan-Meier plots

The R survival and *maxstat* packages were used to estimate the survival function (Hothorn & Lausen, 2002). To determine the two groups, low factor and high factor, an optimal cut-point was determined using maximally selected rank statistics implemented in R package *survminer* and *p*-values based on log-rank test between the two groups (Kassambara et al., 2017).

2.12 Code availability

The software supporting the outlined methods is implemented in R scripts and is available with full documentation at https://github.com/morgen01/PDAC_MOFA

CHAPTER 3 RESULTS – A MULTI-OMICS INTEGRATION MODEL FOR PDAC

The TCGA database was chosen as data resource for this research because it contains multiple omics data from the same patient and provides open access. Specifically, data from the cohort representing pancreatic cancer and referred to as PAAD on TCGA was used for the multi-omics model. To circumvent the problematic of downloading the raw data from all samples individually, the LinkedOmics database was used to retrieve the data from all samples in a matrix format with samples as columns and features as rows (Vasaikar et al., 2018). It is important to mention, that the data provided by LinkedOmics already underwent some preprocessing and normalization steps. A description of the data, the type of values that were recorded and how it was preprocessed is detailed in the Methods section.

Normalization of data is important to remove technical biases such as sequencing depths and gene length and to make expression data comparable across samples. While the availability of preprocessed data has been beneficial for this multi-omics integration approach in terms of time, the multi-omics model will reflect the quality of the data that is fed into it. This refers to the *"Garbage in, garbage out"* concept in computer science in which low quality and flawed data will result in low quality and nonsense output. Upon initial inspection of the data this was accepted.

3.1 Description of PDAC data set

3.1.1 Clinical data

Data from 185 patients who have undergone surgical resection of their primary infiltrating pancreatic adenocarcinomas were included in this study. To get an overview of the patient population, descriptive statistical analysis of the clinical data was performed. The following features are part of the clinical data: age, gender, race, ethnicity, histological type, pathologic staging according to the 0-IV staging system, pathologic staging using the TNM staging system, radiation therapy, residual tumor, vital status, and overall survival.

The age of the patients ranges from 35 to 88 years with a median age of 65 years. Slightly, more male than female patients are part of the cohort with 102 and 83 patients, respectively. The majority of patients (n = 162) are of white origin, and only 11 patients are of Asian descent and 7 patients of Black or African American descent. This results in a bias towards white patients and must be considered when discussing clinical outcomes.

Most patients presented with PDAC (n = 154), but 25 patients had another unknown pancreatic adenocarcinoma subtype. Four patients harbor a different rare subtype of pancreatic cancer known as colloid carcinoma of the pancreas or mucinous non-cystic carcinoma which is characteristic of a better prognosis than PDAC (GAO et al., 2015).

The majority of patient samples are categorized as stage II (n = 152) using the cancer pathology staging 0-IV or pathology stage T3 (n = 148), referring to the TNM staging system. These TNM staging results suggest that in most patients the cancer has already spread to nearby tissues. More than 70% of patients (n = 130) have pathology stage N1 indicating that the cancer

has spread to regional lymph nodes. The number of affected lymph nodes ranges between 0 and 16 with a mean of 3. However, metastasis could only be confirmed in 5 patients with more than 50% of patients having no metastasis status (n = 95).

Radiation therapy refers to a cancer treatment that uses high-energy particles to destroy tumor cells. In PDAC radiation therapy is applied either after surgical removement of the primary tumor or in addition to chemotherapy if tumors have grown beyond the pancreas as well as for pain relief (Hall & Goodman, 2019). Almost 70% of patients in the PAAD cohort (n = 125) haven't received any form of radiation therapy while 24% have received this form of therapy (n = 80).

Following cancer treatment, the tumor status is evaluated using the residual tumor (R) classification (Hermanek & Wittekind, 1994). Interestingly, most of the patients (n = 111) are classified as R0 with no residual tumor present while almost 29% of patients (n = 53) are classified as R1 with microscopic residual tumor and only approx. 3% of patients (n = 5) are classified as R2 and harbor macroscopic tumor lesions after treatment.

At the time the data was recorded, 80 patients had deceased, and 99 patients were still alive. The overall survival time ranges between 31 and 2741 days with a mean survival period of 585.3 days. The full descriptive statistics on the TCGA PAAD cohort is represented in Table 3.

Table 3. Summary of patient characteristics of TCGA PAAD cohort.

Descriptor	Total patients (n=185)
Age, years *(median, range)*	65 (35 – 88)
Sex, *n (%)*	
Male	102 (55.1)
Female	83 (44.9)
Race*, n (%)*	
White	162 (87.6)
Asian	11 (5.9)
Black or African American	7 (3.8)
NaN	5 (2.7)
Ethnicity*, n (%)*	
Non-hispanic or latino	137 (74.1)
Hispanic or latino	5 (2.7)
NaN	43 (23.2)
Cancer type, *n (%)*	
Pancreas adenocarcinoma ductal type	154 (83.2)
Pancreas adenocarcinoma other subtype	25 (13.5)
Pancreas colloid (mucinous non-cystic) carcinoma	4 (2.2)
NaN	2 (1.1)
Pathologic stage, *n (%)*	
I	21 (11.4)
II	152 (82.2)
III	5 (2.7)
IV	5 (2.7)
NaN	2 (1.1)
Pathology T stage, *n (%)*	
T1	7 (3.8)
T2	24 (13)
T3	148 (80)
T4	4 (2.2)
NaN	2 (1.1)
Pathology N stage, *n (%)*	
N0	50 (27)
N1	130 (70.3)
NaN	5 (2.7)
Pathology M stage, *n (%)*	
M0	85 (45.9)
M1	5 (2.7)
NaN	95 (51.4)

Table 3. Summary of patient characteristics of TCGA PAAD cohort (continued).

Descriptor		**Total patients (n=185)**
Number of lymph nodes, *(mean, range)*		3 (0 – 16)
Radiation therapy, *n (%)*		
	Yes	45 (24.3)
	No	125 (67.6)
	NaN	15 (8.1)
Residual tumor classification		
	R0	111 (60)
	R1	53 (28.6)
	R2	5 (2.7)
	NaN	16 (8.6)
Overall survival, days *(mean, range)*		585.3 (31 – 2741)
Vital status, *n (%)*		
	0 (alive)	80 (43.2)
	1 (dead)	99 (53.5)
	NaN	6 (3.2)

3.1.2 Omics data

The retrieved omics data from LinkedOmics comprises mixed matrices of genomic (mutation, and copy number variation), transcriptomic (mRNA expression), epigenomic (methylation levels and miRNA expression) and proteomic (from RPPA analysis) data.

At the genomics layer, mutations have been recorded as single nucleotide variants (SNV) both at gene and chromosomal site level with 2317 and 504 features, respectively. The mutation status has been documented for 126 patients in a binary format with 0 and 1 denoting the absence and presence of a respective mutation, respectively.

Somatic copy number variations (SCNV) were recorded in four matrices, at focal i.e., chromosomal region and gene level site. For each level, the significantly amplified or deleted regions have been identified using the GISTIC2 algorithm and values in a continuous as well as discrete data type recorded referring to log-ratio transformed and thresholded values. Within the 184 samples, there were no missing values recorded.

mRNA transcripts were detected and quantified using RNA-seq and recorded for 178 patients across 19,774 features. The data is stored as continuous values with no missing values.

At the epigenetic layer, methylation data was recorded at CpG-site as well as at gene level and comprises 334,357 and 20,098 features of continuous data type, respectively. The 184 samples each harbor 8.4% and 17.1% missing entries for CpG-site and gene level, respectively. miRNA expression data adds information to the epigenetic layer for 178 samples and 734 features of continuous data type. No missing values were present in the miRNA data modality.

The expression of proteins was recorded at both gene and analyte (antigen) level for a total of 123 patients comprising 182 and 149 features of continuous data type, respectively. The proteomic layer is the layer with the highest degree of sparsity and has 22.8% and 28.3% of missing values at gene and analyte level, respectively in addition to the fewest samples.

Interestingly, for none of the 185 patient samples there is a complete set of omics data available. Table 4 shows a description of the different omics data in terms of number of samples, features, data type and ranges as well as the number of missing values within each data modality.

Table 4. Summary of available omics data for TCGA PAAD cohort from LinkedOmics.

Omics data set	**Numbers of samples**	**Numbers of features**	**Data type, range *(min-max)***	**Total missing values (%)**
Mutation (SNV)				
Gene level	126	2317	Binary	0 (0)
Site level	126	504	Binary	0 (0)
SCNV				
Focal level (log-ratio)	184	56	Continuous (-1.2929 – 3.6569)	0 (0)
Focal level (thresholded)	184	23	Discrete (0, 1, 2)	0 (0)
Gene level (log-ratio)	184	24776	Continuous (-1.293 – 3.657)	0 (0)
Gene level (thresholded)	184	24776	Discrete (-2, -1, 0, 1, 2)	0 (0)
RNAseq	178	19774	Continuous (0 – 22.6756)	0 (0)
Methylation data				
CpG-site level	184	334357	Continuous (-0.4956 – 0.4956)	34264 (8.4)
Gene level	184	20098	Continuous (-0.4944 – 0.4949)	3779 (17.1)
miRNA				
Gene level	178	734	Continuous (0 – 19.345)	0 (0)
RPPA				
Analyte level	123	182	Continuous (-4.1159 – 6.076)	53 (28.3)
Gene level	123	149	Continuous (-4.1159 – 6.076)	35 (22.8)

3.2 Training the MOFA model with PDAC omics data

To build a suitable model for addressing the research question, an incremental integration approach using MOFA was applied, as well as different model parameters. Since there are more data matrices for all omics data except for the RNA sequencing data for transcripts and miRNA, a rational decision was made on which matrices to use for model building. Because of subsequent interpretation of the model, it was decided that choosing all data matrices at gene level is a reasonable choice as gene names can be inferred from the literature and made sense of more easily than regions. Thus, the mutation, the thresholded SCNV, and the RPPA data were used in gene level format. An exception was the methylation data, which in some models was included as methylation status at CpG-site level. The architecture of the different PDAC-MOFA models is described in (Supplemental Table A, Appendix). To select the best performing model, a comparison was done using the Evidence of Lower Bound (ELBO) algorithm built into MOFA with higher resulting ELBO values indicating a better model fit.

The first model built was incorporating mutation and RNA-sequencing data. As suggested by the developer of the tool (personal communication), starting with mutation data and the 5000 most variable genes from the RNA-seq data is suggested for a start as mutation data is straight-forward in its explanation because of the binary nature with 0 and 1 denoting absence or presence of a certain gene mutation. Subsequently, methylation (using top 1% of variable features, or 3199 features), SCNV (5000 most variable features), miRNA and protein data were added one at a time. These models were further tweaked using either 10 or 15 MOFA factors. Generally, the higher the number of MOFA factors the more explanation of variance is possible. However, there will never

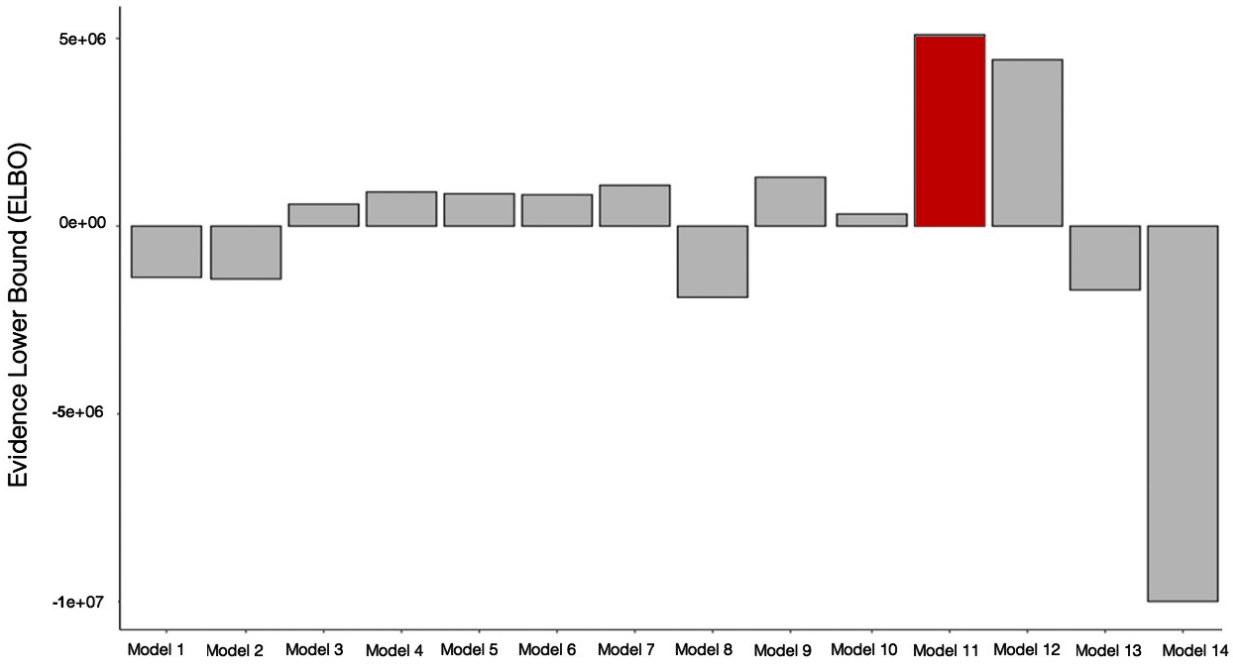
be a number of factors capable of explaining 100% of the underlying variability (Argelaguet et al., 2018).

The combination of different likelihood models such as Gaussian for continuous data (mRNA, methylation, protein, SCNV and miRNA) and Bernoulli for binary data (mutation) was set in the model parameters. Since the data values of the different omics modalities had different ranges, the data scaling parameter within the data options was used in some models to improve model performance. This parameter scales the views to have the same unit variance. However, scaling the features didn't improve model performance (Figure 2, models 8, 13 and 14) and made it even worse as reflected by the negative ELBO scores.

For subsequent downstream analysis, MOFA model number 11 was used as it had the highest ELBO score (Figure 2). This was striking because model 11 had all six data modalities integrated without any feature selection and retrieving 15 factors. This suggests that MOFA is robust to incorporate data from different types and with different dimensionalities.

Of concern was that the best performing model 11 could have overfit because of including the high number of features in RNA-seq, methylation and SCNV data modalities. But linear dimensionality reduction models are not prone to overfitting and in general should be able to deal with data of varying dimensions. However, if one data modality outcompetes other data in terms of features this can lead to over-representation of these features in the final model (Argelaguet et al., 2018).

A



B

Model	ELBO value
1	-1366636.0
2	-1407734.7
3	584224.4
4	908100.1
5	861675.9
6	836036.1
7	1087498.1
8	-1896148.6
9	1301855.2
10	323034.2
11	5096198.9
12	4431253.4
13	-1700954.6
14	-9995418.7

Figure 2. Comparison of PDAC MOFA models.

(A) Diagram showing ELBO values (from table in B) of 14 different PDAC MOFA models.

A comparison of model 11 and the best performing model with reduced input feature space, model 9, shows that some of the retrieved MOFA factors correlate suggesting that those factors are similar in nature (Figure 3). Importantly, factor 1 and factor 3 are strongly correlated across the two models indicating that these two factors are almost identical. Other factors that are strongly correlated are factor 8 of model 9 with factor 9 of model 11 as well as factor 2 of model 9 and factor 4 of model 11. Factors 4 and 3 of models 9 and 11, respectively, are also correlated albeit at a lower scale. In both models, factor 5 shows correlation and the same is true for factors 4 and 3 of models 9 and 11.

This comparison across models suggests that MOFA can integrate a variety of different PDAC data types and of different dimensions but still able to grasp the most relevant sources of heterogeneity present.

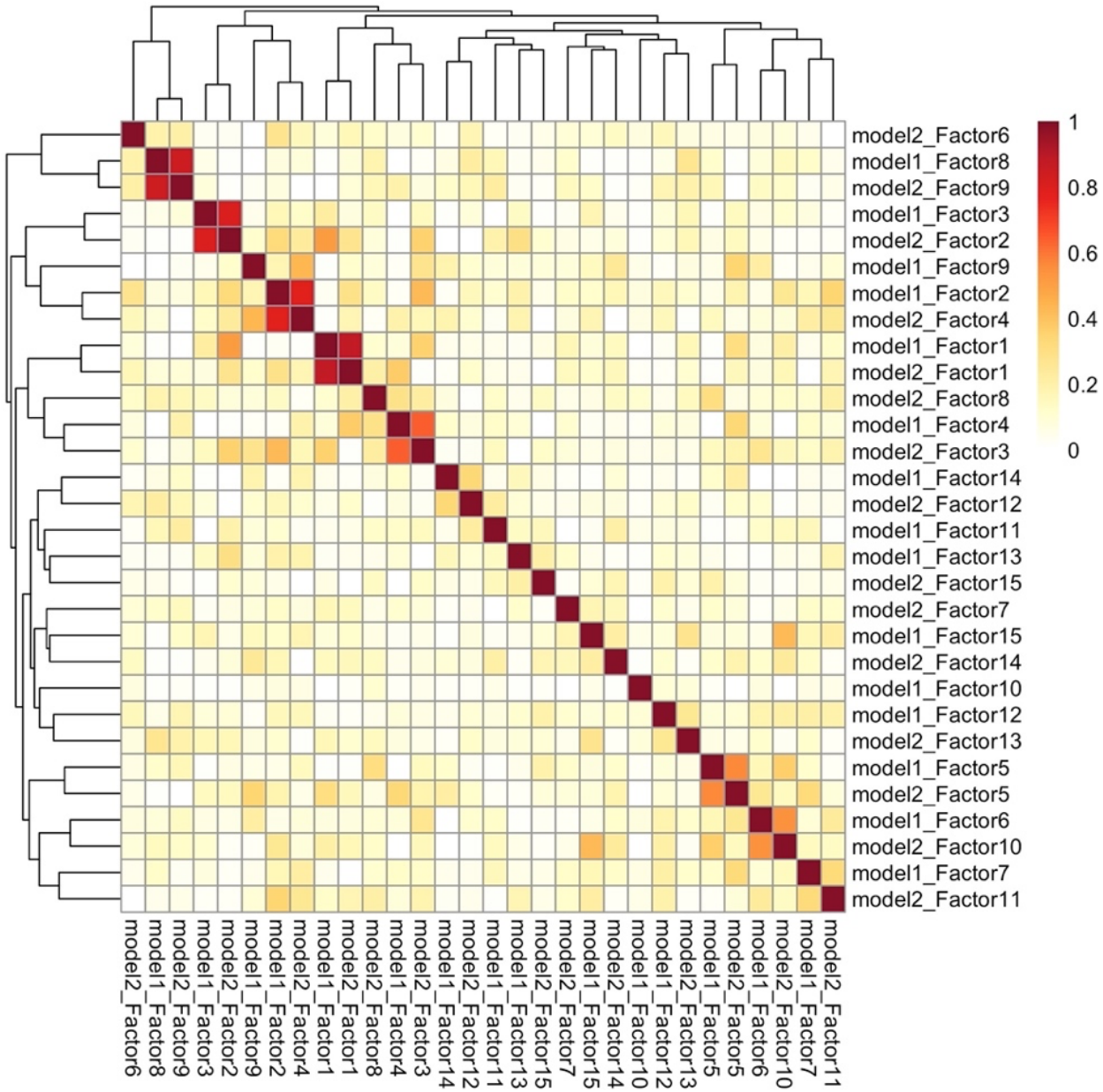
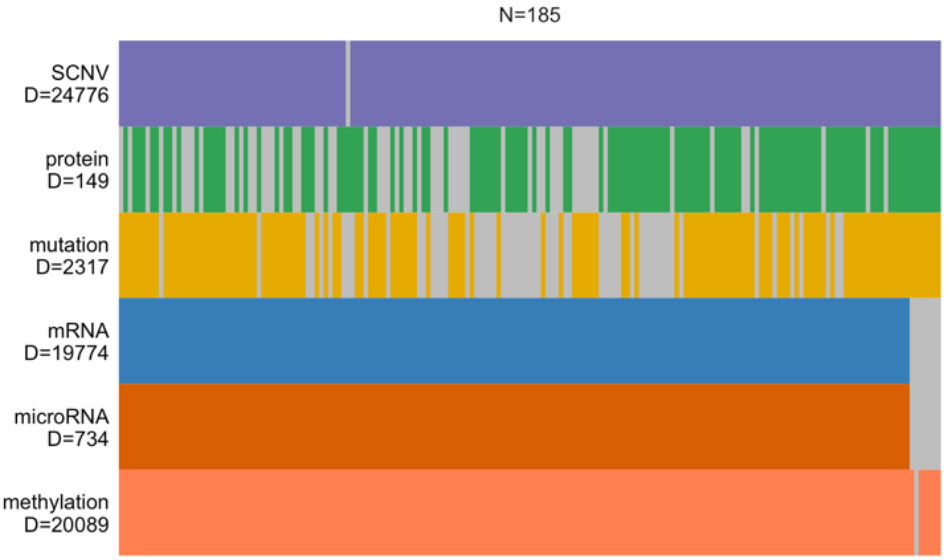


Figure 3. Comparison of MOFA factors from different PDAC MOFA models. Comparison of latent factors from model 9 (model 1) and model 11 (model 2).

The architecture of the MOFA model 11 used for variance decomposition and further downstream analysis has the following number of features: 2317, 24776, 20089, 734, 19774 and 149 for mutation, SCNV, methylation, miRNA, mRNA, and protein data modalities, respectively. Figure 4A shows the model architecture and represents which data is available across 185 patient samples. Of note is the sparsity of the model indicated by the grey areas in the graphic. As MOFA uses sparsity priors, it is capable of dealing with missing values, but the important aspect is that there needs be an overlap between samples i.e., the multi-omics assays have to originate from the same set of samples. This is the case for the PDAC data set and has been a reason why it was chosen.

The trained PDAC MOFA model has learned 15 factors which capture the major sources of variation within the PDAC data set and across data modalities from the genome, transcriptome, epigenome, and proteome. The retrieved factors were largely orthogonal and didn't show any correlation with each other, indicating that they can capture different sources of variation (Figure 4B). The factor correlation plot is also a quality check on model performance and shows that the chosen model without a reduced feature set is a reasonable choice.

A



B

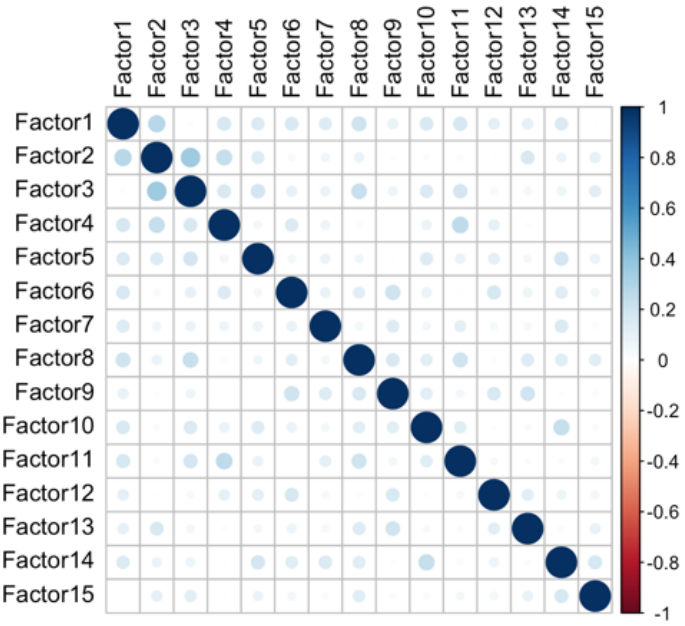


Figure 4. Application of MOFA to PAAD cohort.

(A) MOFA model for PAAD representing integrated omics data modalities with *D* features as rows and *N* samples as columns. Grey areas indicate missing data for the respective patient sample. (B) Factor correlation plot showing that none of the factors strongly correlate with each other.

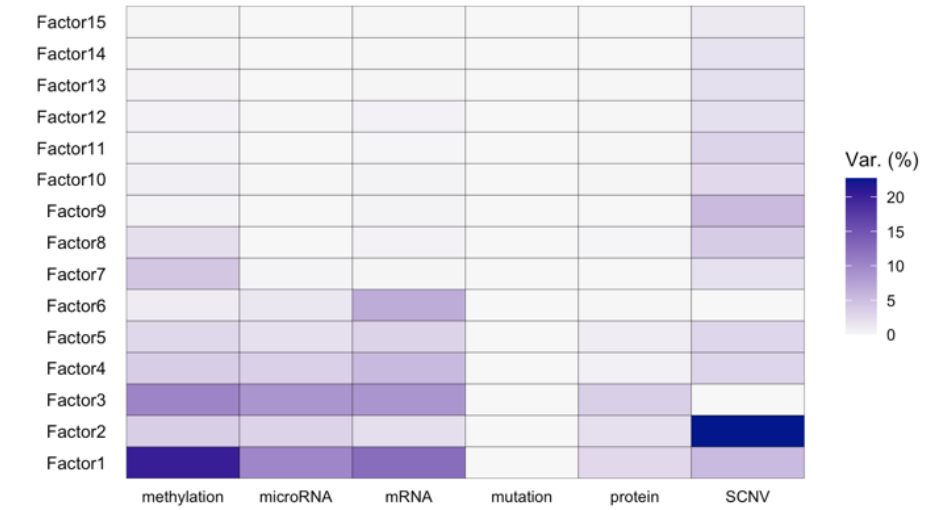
3.3 Variance decomposition

To explore the variance captured by each latent factor, the percentage of explained variance was plotted in a heatmap diagram for all 15 factors and separated by all data modalities (Figure 5A).

Factors 1 and 2 were active in all but the mutation data indicating shared variance across several molecular layers (Figure 5A). While the methylation data modality explains the most variability within factor 1, the SCNV data modality has the highest variance explained in factor 2. The first six factors contribute the most to the heterogeneity whereas the last five factors don't contribute a lot. An exception is the SCNV data which except from factor 3 and 6 adds to the variability within each factor.

Cumulatively, the 15 factors explain 55% in the SCNV data, 49% in the methylation data, 40% in the mRNA data, 28% in the microRNA data, 11% in the protein data, and only 0.01% in the mutation data (Figure 5B). Interestingly, the model hardly captures any variation in the mutation data suggesting that gene mutations don't contribute to the observed tumor heterogeneity in PDAC. This low contribution to variation by the mutation data was also seen in other models which had fewer data modalities incorporated, such as models 1, 2, 3, and 4 (data not shown). Generally, the higher the number of factors, the more variance can be explained. Conversely, smaller total explained variance is a result of noisy data sets with strong non-linearities. Whereas the nature of MOFA as a linear and sparse model reduces effects from overfitting, it will never be able to explain 100% of the variance in a data set.

A



B

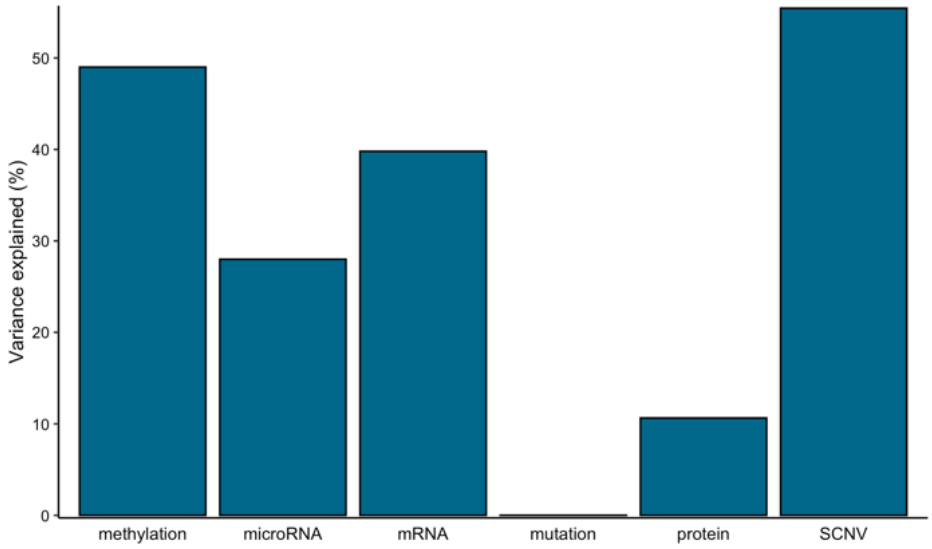


Figure 5. Explained variance by MOFA model applied to PAAD cohort.

(A) Variance plot showing the proportion of variance explained (%) by individual factors and for each assay. (B) Cumulative explained variance (%) for each data modality.

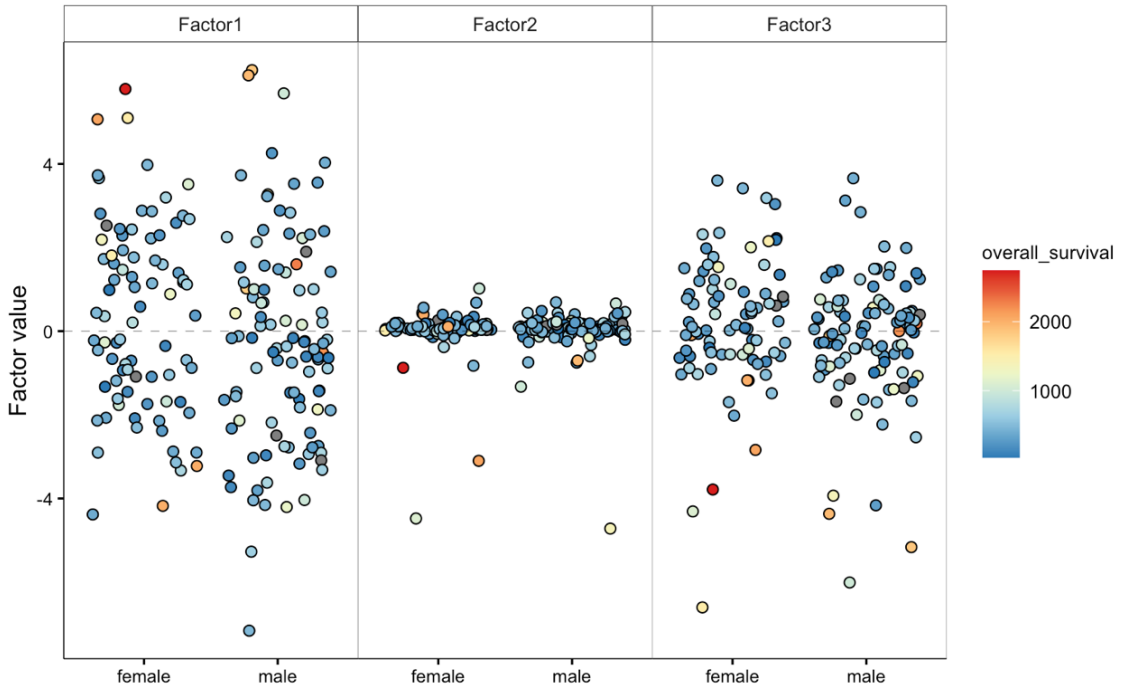
3.4 Analysis of MOFA factors

A MOFA factor is defined as a linear combination of input features. As such, each factor is capable of capturing a different source of variability within the PDAC data. The resulting factor values arrange the patient samples along a one-dimensional axis that is centered at zero. Samples with different factor values are arranged on opposite site of this axis as indicated by positive and negative signs and higher values indicating stronger effects.

Initially, the top three factors were characterized and visualized in latent space using beeswarm plots and scatterplots and annotated with gender and overall survival (Figure 6). While within factors 1 and 3, the samples are spread on either side of the zero-centered axis, almost all samples arrange at the zero-value axis for factor 2. Thus, most samples do not exhibit great variation in factor 2.

No difference was made out in the separation of the first three factors with respect to gender. The grouping of the 83 female and 102 male patients according to the three factors didn't show any differences in these groups visually (Figure 6). Concerning the overall survival, most samples within factors 1, 2, and 3 have poor survival denoted by the blue color which is indicative of a survival period of less than 100 days from initial diagnosis. No immediate effect based on gender or impacting survival was detected at this stage.

A



B

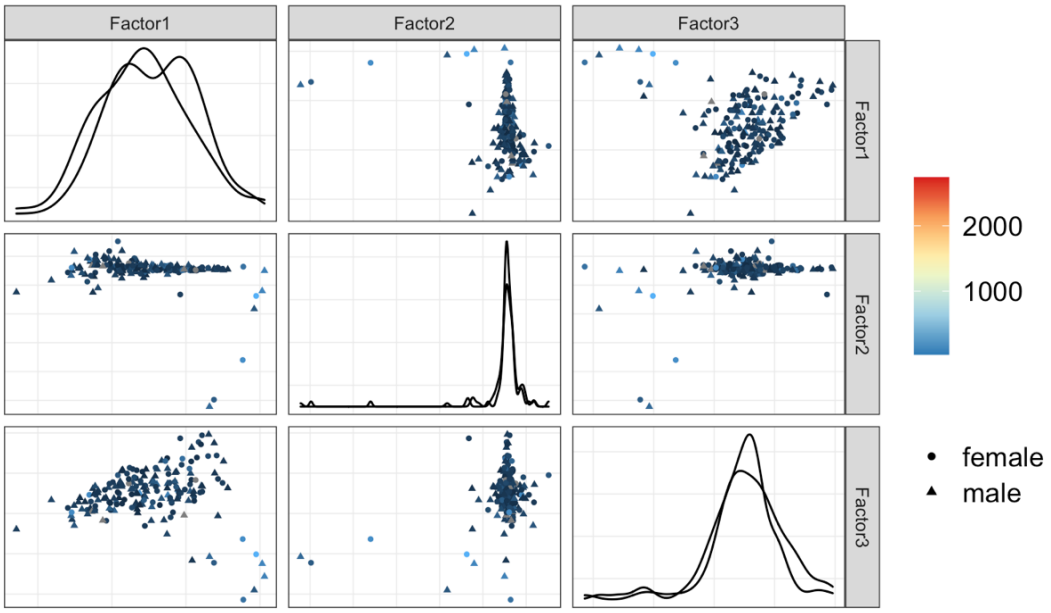


Figure 6. Visualization of MOFA factors 1, 2, and 3 in latent space.

(A) Beeswarm plot of latent factors 1, 2, and 3 showing the spread of samples grouped by gender and colored by overall survival. (B) Scatter plots of factor combinations between factors 1, 2, and 3 showing the combination of two factors, colored by overall survival, and split by gender.

3.5 Molecular characterization of latent factors

Each factor contains weighted features of each data modality. Feature weights of zero characterizes features with no association to the corresponding factor whereas features with large absolute values confer strong associations with the factor. Positive and negative signs of the feature weights refer the direction of the effect with positive weights indicating higher levels of the feature in sample/cells and vice versa.

Since factors 1, 2, and 3 captured most of the heterogeneity, those factors were further characterized by comprehensive literature search on MEDLINE®/PubMed®. Particularly, it was investigated whether the top ten features within the different data modalities of the first three factors have published connections to PDAC. A previous association with PDAC was marked in red on the feature line plots.

3.5.1 Characterization of MOFA factor 1

Latent factor 1 was able to capture 20% of variability within the methylation, 13% of variability within the mRNA, 10% of variability within the miRNA, 2.6% of variability within the protein, 5% of variability within the SCNV, and no variation related to the mutation data.

3.5.1.1 Mutation data modality and factor 1

The contribution of the mutation data to the variability within factor 1 is with 0.0007% negligible. Despite the clinical importance of mutations like in *KRAS, TP53, CDKN2A*, and *RNF43*, amongst others, those genes do not appear to play a role in determining the heterogeneity of PDAC. The mentioned genes can be found among the top features of the mutation view and are strong negative signs (Figure 7A).

Mutations in *RYR2* have not been linked to PDAC previously but according to a recent bioinformatics study, are implicated in esophageal adenocarcinoma tumorigenesis and prognosis. Similarly, the neuroblastoma breakpoint family member 3 (*NBPF3*) has not been a documented mutation in PDAC but has been implicated in oncogenesis (Shi et al., 2018). Additionally, no published evidence for an involvement of mutation of the putative tyrosine-protein phosphatase *TPTE* (or *PTEN2*), the RAN binding protein 2 (*RANBP2*), and the protein arginine methyltransferase 8 (*PRMT8*) were found.

While it was described that *FMN2* expression is linked to overall survival in PDAC patients, it was not evident whether this overexpression was mediated by a gain-of-function mutation or through other regulatory mechanisms (Raja et al., 2021).

The results of the line plots were further confirmed by heatmap visualization showing the absence and presence of the top gene mutations in the input data. Almost all patient tumors harbor *KRAS* mutations and a vast majority of samples further contain mutations of *TP53* (Figure 7B). Taken together, these results suggest other sources than gene mutation are responsible for PDAC heterogeneity.

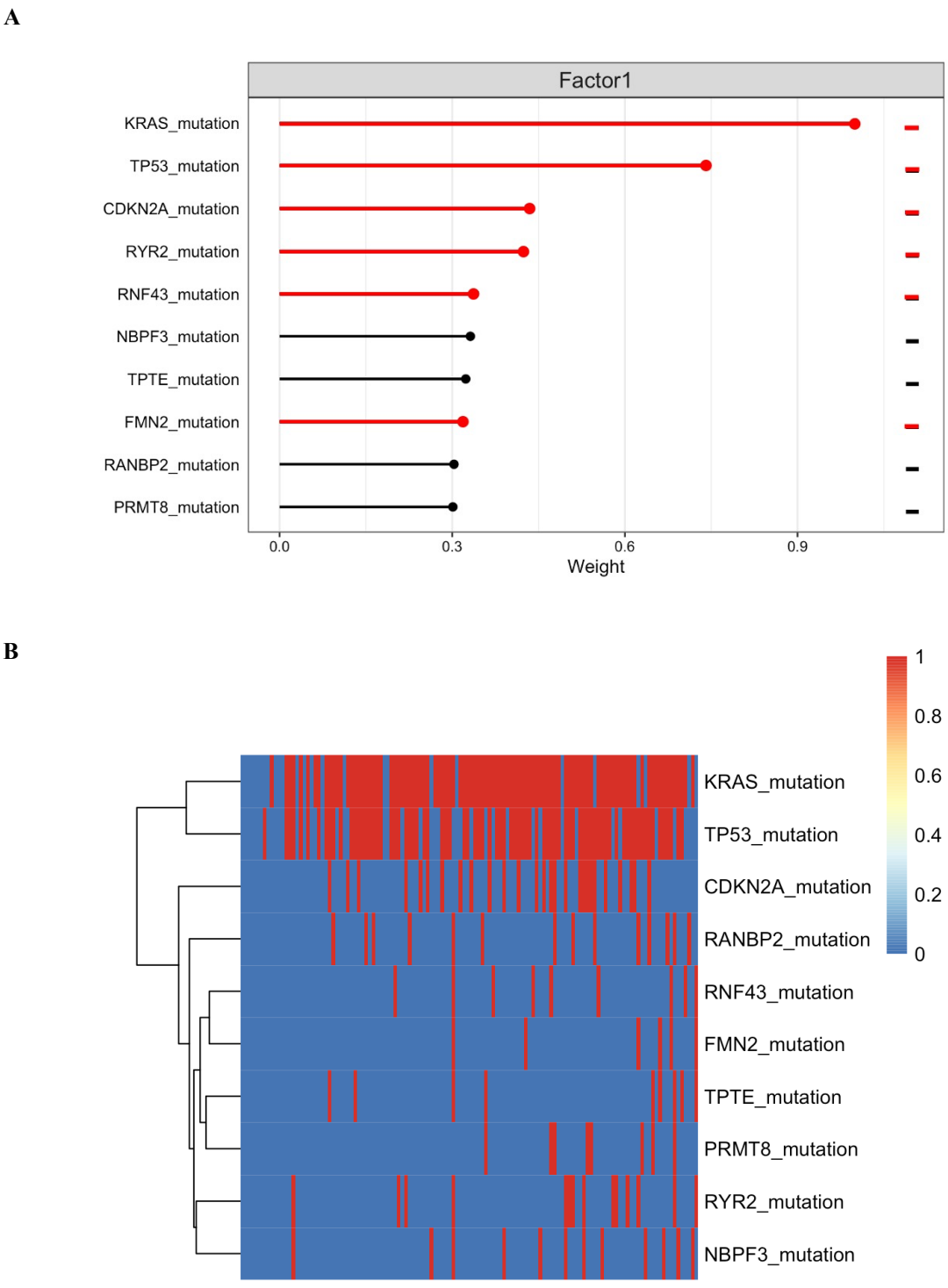


Figure 7. Top 10 features within mutation data modality of factor 1.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows absence (blue) and presence (red) of gene mutation in patient samples (N = 185).

3.5.1.2 SCNV data modality and factor 1

Comprising 5% of all explained variance, the SCNV layer further adds to the explanation of factor 1. The top three features with positive weights are *C9orf53*, *CDKN2A* and *CDKN2B* (Figure 8A). While the protein-coding gene chromosome 9 open reading frame 53 (*C9orf53*) hasn't been linked to pancreatic cancer before, deletions in *CDKN2A* and *CDKN2B* are known oncogenic drivers (Ho et al., 2010). However, *C9orf53* has been shown to impact the growth of head and neck squamous cell carcinoma and has been linked to poor prognosis (Wang et al., 2019). Similarly, somatic mutations in *DMRTA1* together with *CDKN2A/CDKN2B* and *C9orf53* (and *MTAP*) have been linked to alterations in immune regulation in glioma (Wu et al., 2020).

Among the top 10 features in the SCNV data modality is also *SMAD4* which has previously been reported to contribute to the oncogenic process in PDAC (Figure 8A).

Co-deletions of the mitochondrial enzyme malic enzyme 2 (*ME2*) in addition to its neighboring gene *SMAD4* has been described as "collateral lethality" and adds to the cancerogenic progress (Dey et al., 2017).

The ribonuclease ELAC1 and the RNA binding protein MEX3C have been linked to other cancer types and up to now haven't been associated with PDAC (Burrell et al., 2013; Takaku et al., 2003). No published connections have been found of RN7SL69P and the U13 small nucleolar RNA to cancer.

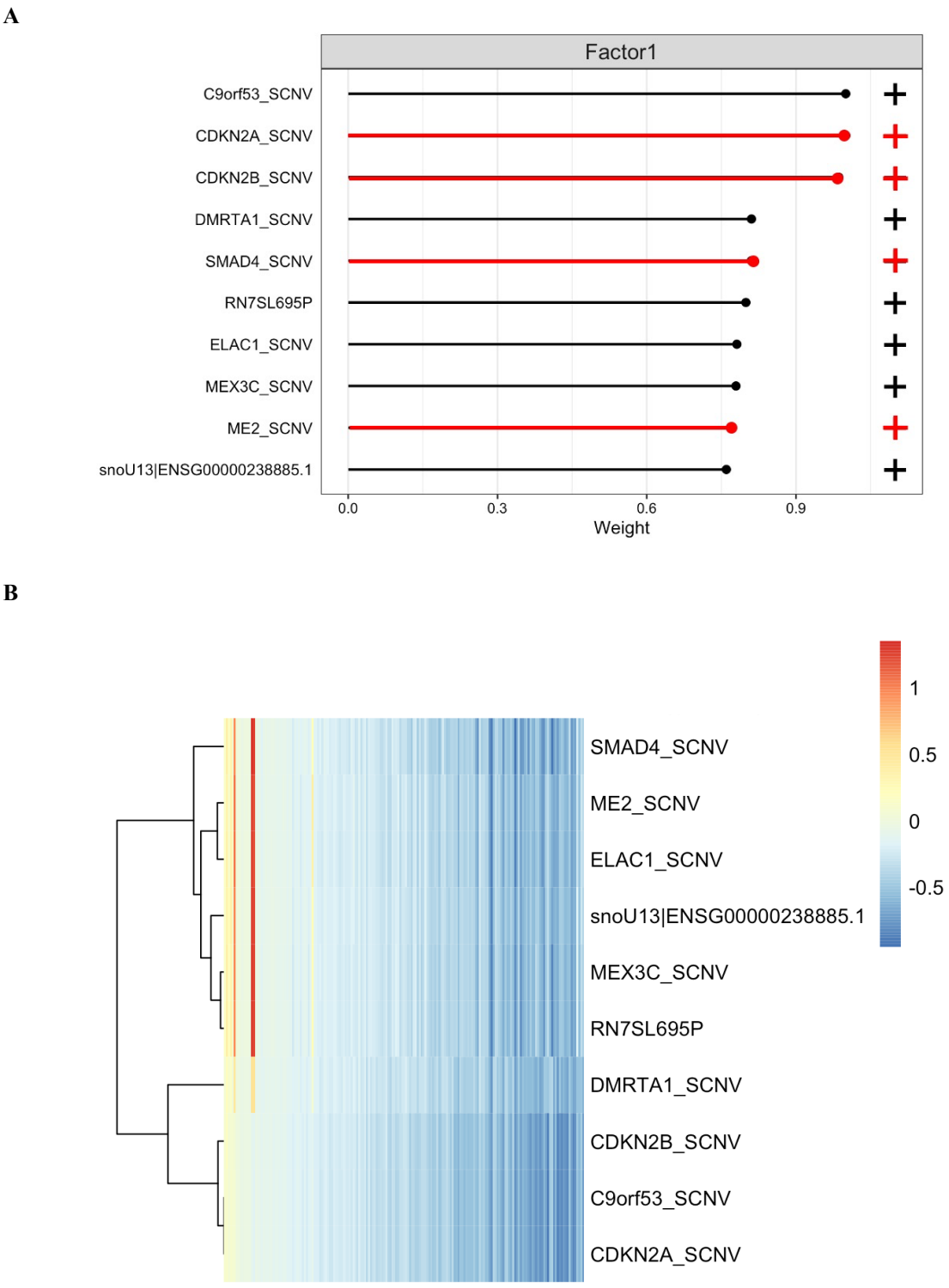


Figure 8. Top 10 features within SCNV data modality of factor 1.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows pattern of SCNV in patient samples (N = 185).

3.5.1.3 mRNA data modality and factor 1

The mRNA modality is with 12% of explained variation the second important component within factor 1 with the ten most important features having negative signs (Figure 9A). The top feature *CEACAM5* (Carcinoembryonic antigen-related cell adhesion molecule) has a negative impact on survival and has been shown to have a role in cell adhesion, invasion and metastasis in pancreatic and other cancers (Blumenthal et al., 2007). Also, the second most important feature, *S100P* has a history in PDAC as playing a role in the aggressiveness of the pancreatic cancer and with high expression levels contributing to higher proliferation, migration and invasion rates (Arumugam et al., 2005). Another promoter of PDAC metastasis is *TFF1* whose expression has been linked to the aggressive potential of pancreatic cancer (Arumugam et al., 2011). Similarly, *CDN18* is highly expressed in PDAC although normally its expression is restricted to the stomach epithelium (Zhu et al., 2019). A recent machine learning study by Ye et al., revealed *TMPRSS4*, *CTSE* and *SDR16C5* as network hub genes which can be exploited as diagnostic markers for pancreatic cancer. Interestingly, the PDACMOFA model found these three genes in the mRNA modality, the fourth biomarker discovered by the group was *TSPAN1* which is not found in this research. Generally, the mentioned genes have been associated with proliferation and metastasis in PDAC.

Anterior gradient 2 (AGR2) is a proto-oncogenic protein that is involved in the maturation process of proteins and has been shown to be expressed in many cancers. Within PDAC *AGR2* expression is linked to the initiation of pancreatic cancer (Dumartin et al., 2017).

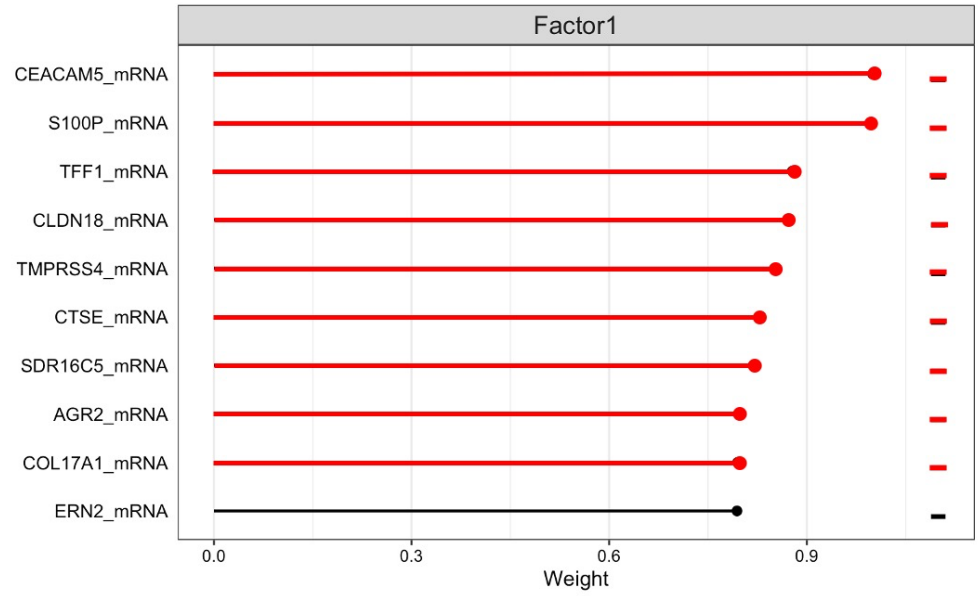
COL17A1 was among nine genes which have been identified by differential gene expression analysis of an integrated GEO data set as prognostic gene signature for PDAC (Wu et al., 2019).

The other eight genes proposed by this study were not among the top ten features for the mRNA view. Although not directly related to pancreatic cancer, *ERN2* expression has been linked to lung adenocarcinoma (Xia et al., 2020).

In summary, all but the *ERN2* expression has been associated with PDAC before in proliferative and metastasis promoting mechanisms.

The mRNA expression pattern of the top 10 genes within factor 1 were further examined using heatmaps. Figure 9B shows that these molecules are differentially expressed within factor 1 and that a specific pattern correlates with the pathologic stage of the tumor and subsequently with patient outcome. Lower mRNA expression levels of the top mRNA features within factor 1 were associated with the pathologic stage I, indicating smaller tumor sizes. This further impacts overall survival which is increased in these samples designated by the dark, purple-colored bars on top. Furthermore, this has an impact on the vital status of patients which were still alive. To conclude, the expression levels of the top ten mRNA features within factor 1 are likely to have an impact on driving tumor progression and thus controlling tumor aggressiveness and survival.

A



B

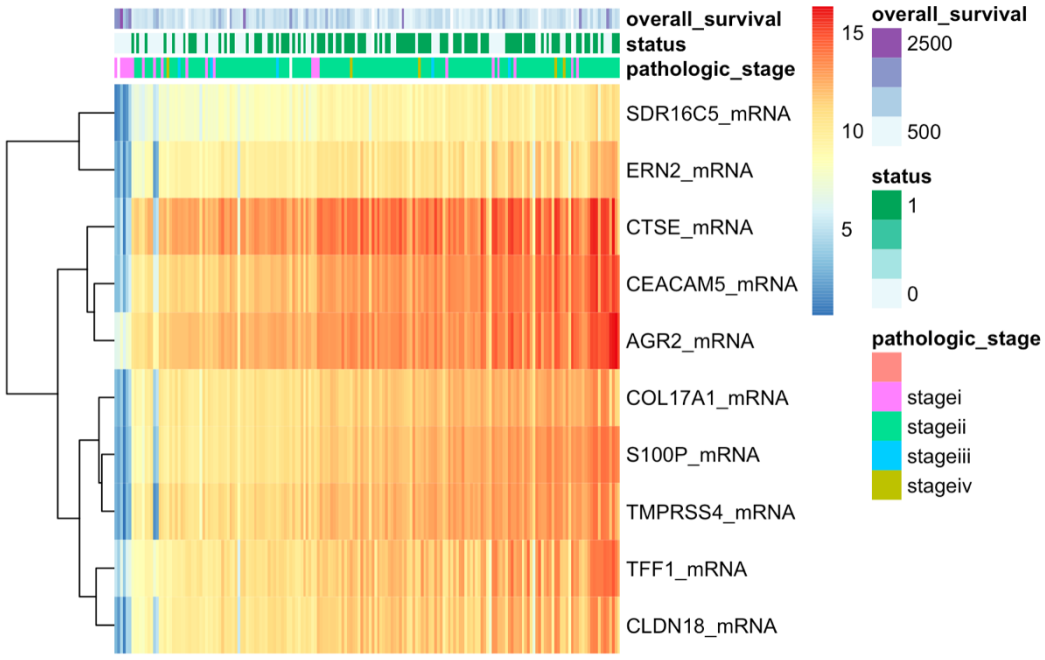


Figure 9. Top 10 features within mRNA sequencing data modality of factor 1.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows mRNA expression pattern, annotated with overall survival, vital status, and pathologic stage in patient samples (N = 185)

3.5.1.4 miRNA data modality and factor 1

Within the miRNA data which explains 10% of the variability of factor 1, the highest ranked feature with a negative sign is miR-203 which has been shown to regulate proliferation, migration and invasion in pancreatic cancer (Ren et al., 2016).

Several studies have identified the second top feature in the list miR-135b to be associated with PDAC and especially conferring to a poor prognosis (Zhou et al., 2019). Specifically, high expression of miR-135b is implicated in the proliferative and invasive behavior of pancreatic cancer stem cells (PSC) by activating the AKT/mTOR signaling pathway via its target JADE-1 (Zhou et al., 2020).

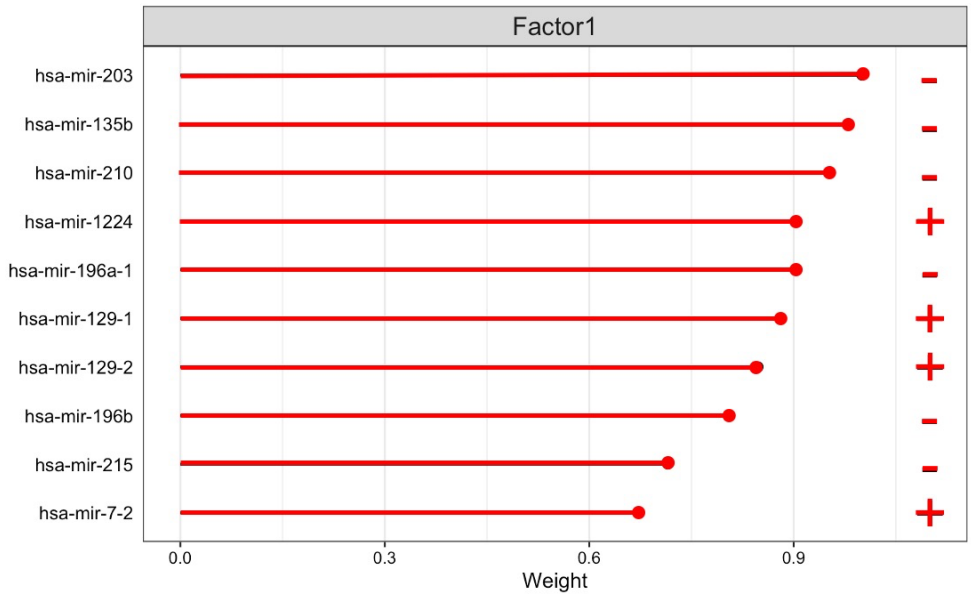
Similarly, miR-210 has been reported to be involved in pancreatic cancer in particular in the epithelial-mesenchymal transition (EMT) in an in vitro system under hypoxia, a hallmark of PDAC and responsible for adverse outcomes (Ho et al., 2010).

miR-1224 has been proposed as a tumor-suppressor within PDAC by regulating the transcription factor ELF3 which has also been linked to other cancers and promotes progression *in vivo*, *in vitro* and *in silico* experiments via the PI3K/AKT/Notch/EMT signaling pathways (Kong, Liu, Zheng, Wang, et al., 2020). Together with miR-129-1 and miR-129-2, miR-1224 was found to be part of a set of 12 prognostic markers for PDAC and using an integrated multi-omics study of mRNA, miRNA, methylation and SNP data from the TCGA PAAD cohort (Jia et al., 2020). Another potential prognostic or therapeutic target could be miR-196b which has been shown to decrease apoptosis in a cell line experimental study through targeting CADM1, a cell adhesion molecule (Wang et al., 2017).

miRNAs act as negative regulators at the post-transcriptional level. Oncogenic roles in PDAC have been ascribed to miR-21, miR-221 and mir-155 while miR-34, miR-200, miR-let7, miR-15a, miR-506, miR-k96, miR-17-92 and miR-145 have been suggested to function as tumor suppressors in PDAC (Guo et al., 2018). A further putative tumor suppressor was found within the top ten features of the miRNA layer of factor 1, miR-215 (Li et al., 2015). Lastly miR-7-2 completes the list of miRNAs that have been shown to be implicated in the regulatory mechanisms of pancreatic tumorigenesis (Figure 10A).

Similarly, to the expression of mRNA, the miRNA expression levels show a characteristic pattern for samples of different pathologic stages. The expression levels of the top miRNAs are differentially regulated between pathologic stage I and II. While the expression of miR-203, miR-210, miR-7-2 and miR-1224 is low in stage I tumors, in stage II tumors, the expression of the former two increases whereas the expression of the latter two decreases. Of note is also that mRNA and miRNA data are explaining similar amounts of variability within factor, indicating that miRNA and mRNA co-regulate processes. Thus, transcriptional, and post-transcriptional alterations might jointly contribute to PDAC cancer progression.

A



B

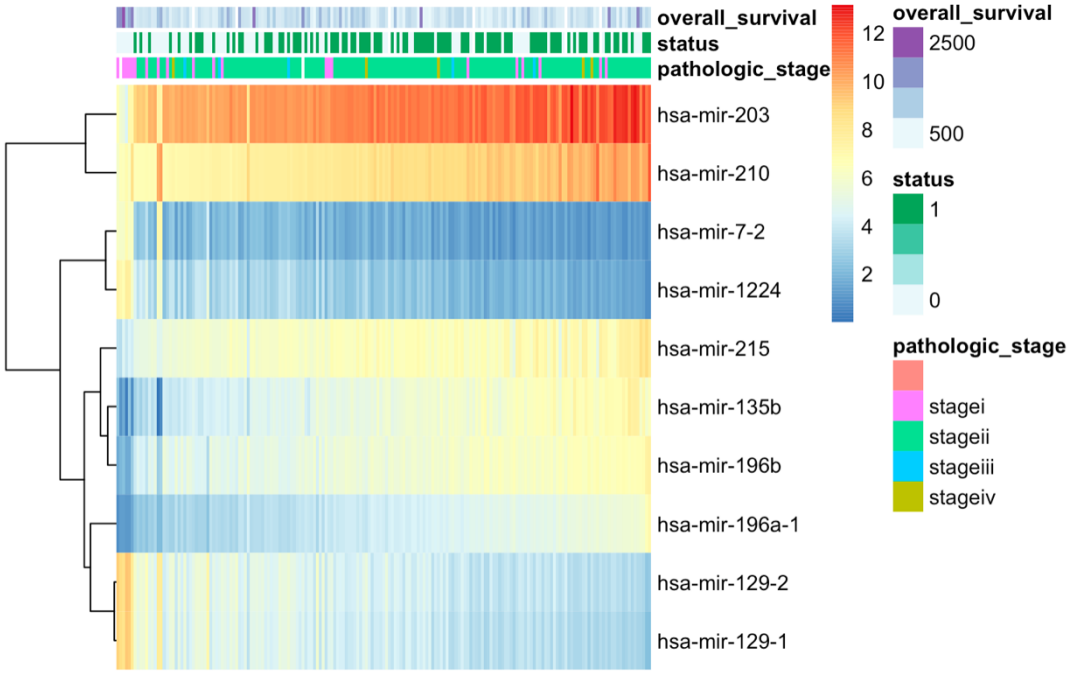


Figure 10. Top 10 features within miRNA data modality of factor 1.

(A) Line plot shows scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows miRNA expression pattern and annotated with overall survival, vital status, and pathologic stage.

3.5.1.5 Methylation data modality and factor 1

The methylation data with about 20% could explain the most variety contained within factor 1. The top loadings of the methylation data modality with positive sign is methylation of *CELSR1* (Cadherin EGF LAG seven-pass G-type receptor), a G protein-coupled receptor involved as regulators in many biological processes such as neuronal/endocrine cell differentiation, vessel valve formation and the control of planar cell polarity during embryonic development (Wang et al., 2014). Moreover, aberrant expression of *CELSR1* has been documented in several different cancers such as breast cancer, lung cancer and glioma in which it acts as an oncogene (Wang et al., 2020). So far, it hasn't been linked to PDAC.

While there is not much known about *LOC399959*, the third top hit, *MSMB* encodes the tumor suppressor PSP94 which is repressed in androgen-refractory prostate tumors through hypermethylation of the CpG island within the promoter region (Beke et al., 2007).

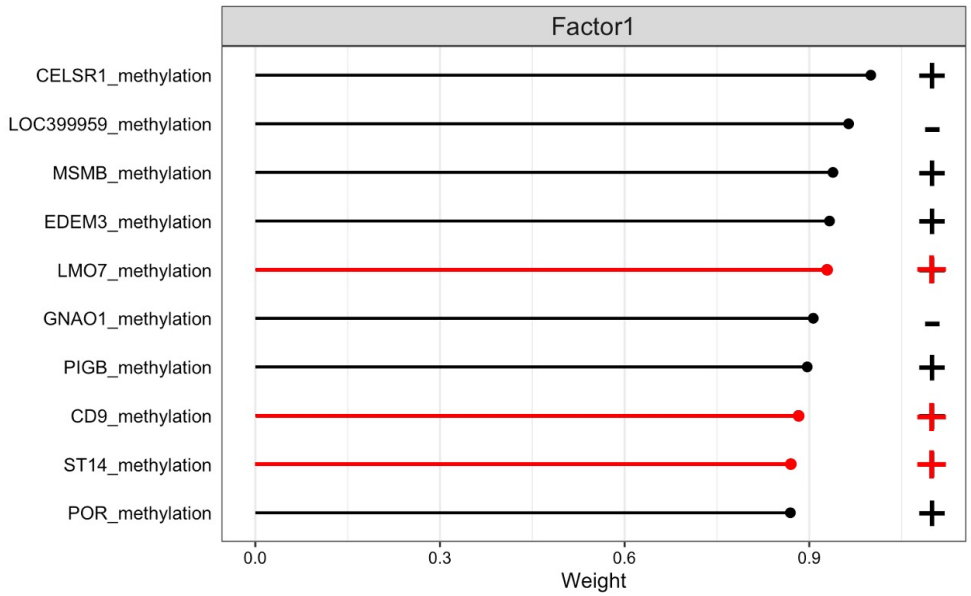
EDEM3 a protein that trims cellular glycoproteins has been linked to cancer progression in a study of hepatocellular carcinoma (HCC) (Zhang et al., 2021). Abnormal expression of another gene, *GNAO1*, through hypermethylation has also been implicated in HCC (Xu et al., 2018).

LMO7 is highly expressed in several murine and human tumors and has been shown to promote pancreatic cancer progression and metastasis in a mouse model (Liu et al., 2021). The tetraspanin CD9 is another marker that has been linked to pancreatic cancer. Interestingly, CD9 has been characterized as a driver of PDAC and initiator of PDAC heterogeneity (Wang et al., 2019). The matripase ST14 is a protease that performs activating proteolytic cleavage of Macrophage stimulating protein (MSP) and has been shown to be upregulated in PDAC (Li et al., 2019).

Lastly, the Cytochrome P450 reductase (POR) has not been linked to PDAC but has been identified as a prognostic marker for triple-negative breast cancer (Pedersen et al., 2019). In summary, from the top ten features of the methylation modality four (*ST14*, *CD9*, *LMO7* and *CELSR1*) have been linked to PDAC previously. All other genes, except *LOC399959*, have been associated with other cancer types and implicated in cancer progression and metastasis processes (Figure 11A).

The methylation pattern of the top 10 features within the methylation data modality of factor 1 further confirms the polarity of the features. While *LOC399959* and *GNAO1* are predominantly hypomethylated in patient samples, all other features are hypermethylated (Figure 11B). There appears to be a difference in the strength of the methylation for different tumor stages with stage I being more strongly hypo- and hypermethylated for the respective gene promoters.

A



B

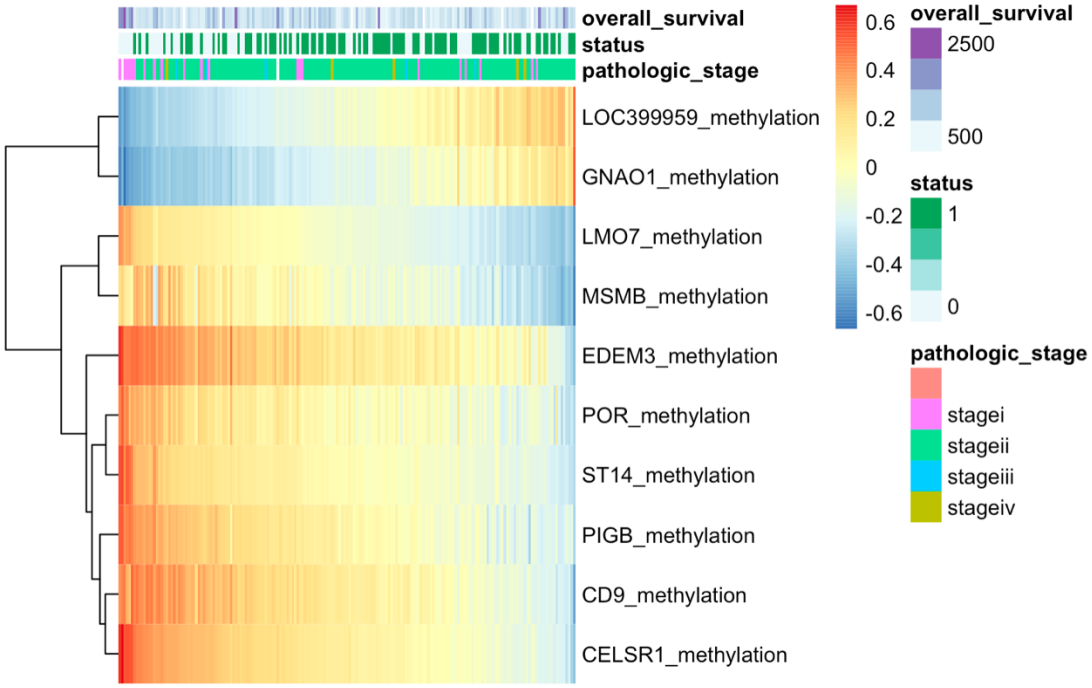


Figure 11. Top 10 features within methylation data modality of factor 1.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows methylation pattern, annotated with overall survival, vital status, and pathologic stage for patient samples (N = 185).

3.5.1.6 Protein data modality and factor 1

Lastly, the protein view can explain 2.6% of the variability within factor 1. Since ten protein features characterizing factor 1 have the same weight (0.9), more than ten have been listed in the line plot (Figure 12). The feature with the top weight is the protein Cdkn1a which has a negative weight indicating that this protein is absent. This makes sense as deletions of *CDKN1A* do not yield a functional protein. The second top feature with strong positive weight is A-Raf a member of the RAF kinase family and implicated in regulating the MAPK/ERK signaling pathway and promoting pancreatic cancer metastasis (Meng et al., 2020).

The adhesion molecule E-cadherin or CDH1 has demonstrated interactions with β-catenin which is involved in intracellular signaling, transcription and cell adhesion with aberrant function promoting chronic inflammation and carcinogenesis (Katoh, 2018).

The membrane-bound transferrin receptor TFRC has been reported as highly expressed in proliferating cells and was shown to be expressed in 80% of the tested tumor cell lines (Ryschich et al., 2004).

The protein encoded by the YWHAZ gene 14-3-3 protein zeta/delta (14-3-3ζ) has a central role in cell signaling especially as a regulator of apoptosis. Within PDAC 14-3-3ζ was linked to chemoresistance and as a consequence proposed as therapeutic target (D'Errico et al., 2019).

PRKAA1 is a serine/threonine protein kinase and catalytic subunit of the 5'-prime-AMP-activated protein kinase (AMPK) a sensor of the cellular energy implicated in cancer protective mechanisms such as inhibition of cell proliferation and induction of apoptosis. On the other hand, AMPK also contributes to EMT which is a hallmark of cancer metastasis (Ye et al., 2016).

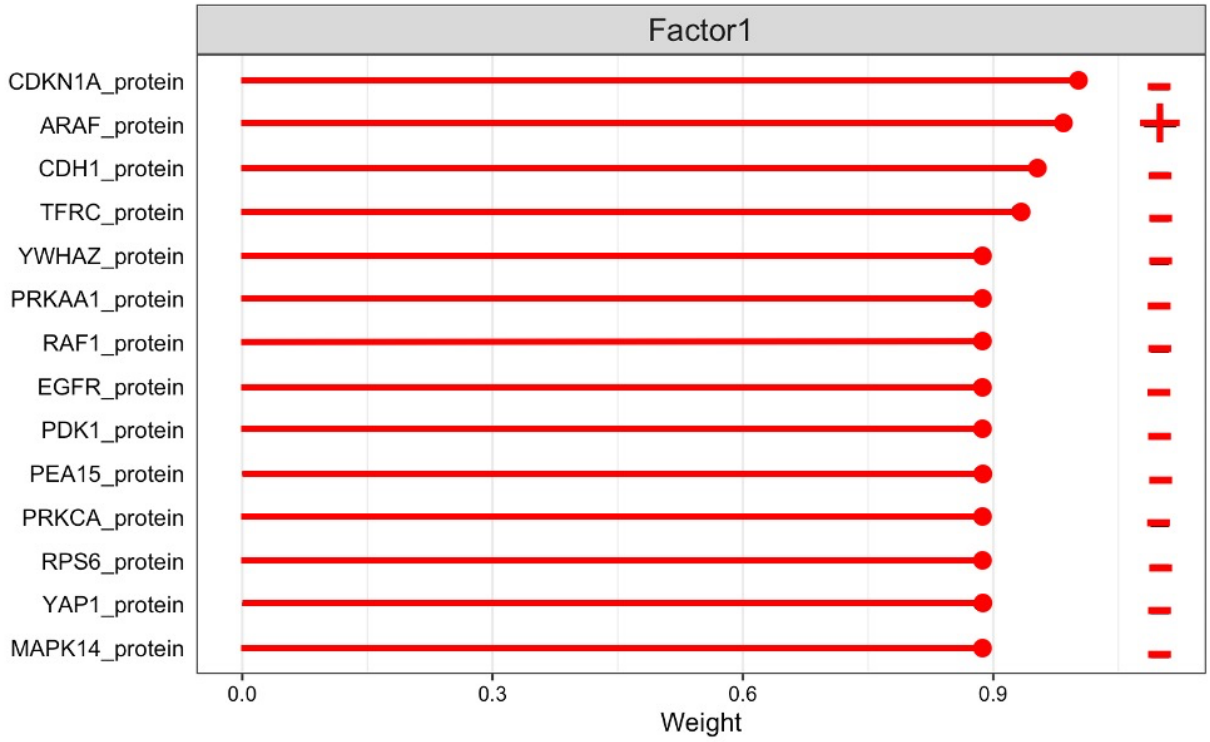


Figure 12. Top 10 features for protein data modality within factor 1.

Line plot of scaled feature weights with red lines indicating previous known associations with PDAC.

The deregulation of the RAS/mitogen-activated protein kinase (MAPK) signaling pathway plays a central role in human cancers. Mutations in KRAS are most frequently observed within this pathway and also constitute up to 96% of mutations in PDAC. Next to ARAF, RAF1 is also downstream of KRAS and affected by *KRAS* mutation such as MAPK (Drosten & Barbacid, 2020). The cell surface receptor EGFR functions as an activator of the MAPK/ERK (also Ras-Raf-MEK-ERK) pathway and is also found among the features with the highest negative weights in the protein modality of factor 1. Since this pathway is affected in PDAC through mutations in *KRAS*, it makes sense that proteins further downstream the signaling cascade are impaired. The oncogene KRAS has further consequences on other signaling pathways as well such as the phosphoinositide 3-kinase (PI3K) and 3-phophoinositide-dependent protein kinase 1 (PDK1) are effectors of KRAS in PDAC. The latter one is also a key feature in the protein data of factor 1 (Eser et al., 2013).

The small scaffold protein PED/PEA-15 has been linked to regulating apoptosis and cell proliferation in the pancreas and dysregulation leading to diabetes and cancer (Fiory et al., 2014).

With the peroxisome proliferator-activated receptor (PPAR) signaling pathway, another signaling pathway is known to be deregulated in PDAC and effector proteins such a PRKCA have been discovered in the protein features of factor 1 (Liu et al., 2020). The PTEN/PI3K/AKT/mTOR pathway is also implicated in the development of PDAC. An effector of mTOR, the ribosomal protein S6 is another verified protein implicated in PDAC that is retrieved by MOFA. Its role has been in initiating pancreatic intraepithelial neoplasia (PanIN) a pre-stage of PDAC through its phosphorylation status (Khalaileh et al., 2013). Also linked to pancreatic cancer initiation is YAP1 which together with TAZ regulates JAK-STAT3 signaling promoting tumorigenesis in a murine model (Gruber et al., 2016).

In conclusion, all retrieved protein features of MOFA factor 1 have been linked to PDAC previously and are components of several cell signaling pathways (Figure 12). Common mechanisms in driving tumorigenesis are promoting cell proliferation and suppressed apoptosis.

Altogether, the analysis of the top features from each data view within factor 1 suggests that factor 1 is implicated in the invasive and proliferative potential of pancreatic cancer mediated by cell signaling pathways.

3.5.2 Characterization of MOFA factor 2

MOFA factor 2 was able to capture 23% within the SCNV data, 3.6% within the methylation, 3.1% within the miRNA, 2.1% within the mRNA, and 2% within the protein data. Again, the mutation data could hardly explain any variation (0.00064%).

3.5.2.1 Mutation data modality and factor 2

The two top features within the mutation data of factor 2 have also been the top features of factor 1, *KRAS* and *TP53*. However, this time the two genes come with large positive weights. Mutations in *MAMLD1* have not been associated with pancreatic cancer before but gene fusions of *YAP1* (which has been reported as a key feature of the protein modality of factor 1) with *MAMLD1* has shown oncogenic potential in mice (Szulzewsky et al., 2020).

The gene encoding the heat shock protein *HSP90AA1* has been shown to be one of six major hub genes that are targets of MALAT1 a lncRNA and impacting several downstream signaling pathways in pancreatic cancer based on a bioinformatics study (Xie et al., 2017).

While not directly related to PDAC, *LPIN3* has been identified as a prognostic signature gene for tumor mutation burden in epithelial ovarian cancer using a multi-omics analysis study from TCGA data (Liu et al., 2020). Similarly, mutations in *MLLT3* have not been related to PDAC but occur in other tumors like acute myeloid leukemia albeit as fusions or deletions (Matsuo et al., 2018).There is published evidence that the p21-activated kinase PAK3 serves as a key effector in pancreatic cancer via Akt-GSK3β-β-catenin signaling in cancer cell lines (Wu et al., 2019).

Since the bottom three genes had very low weights, they were not further investigated in detail and didn't show a connection to PDAC.

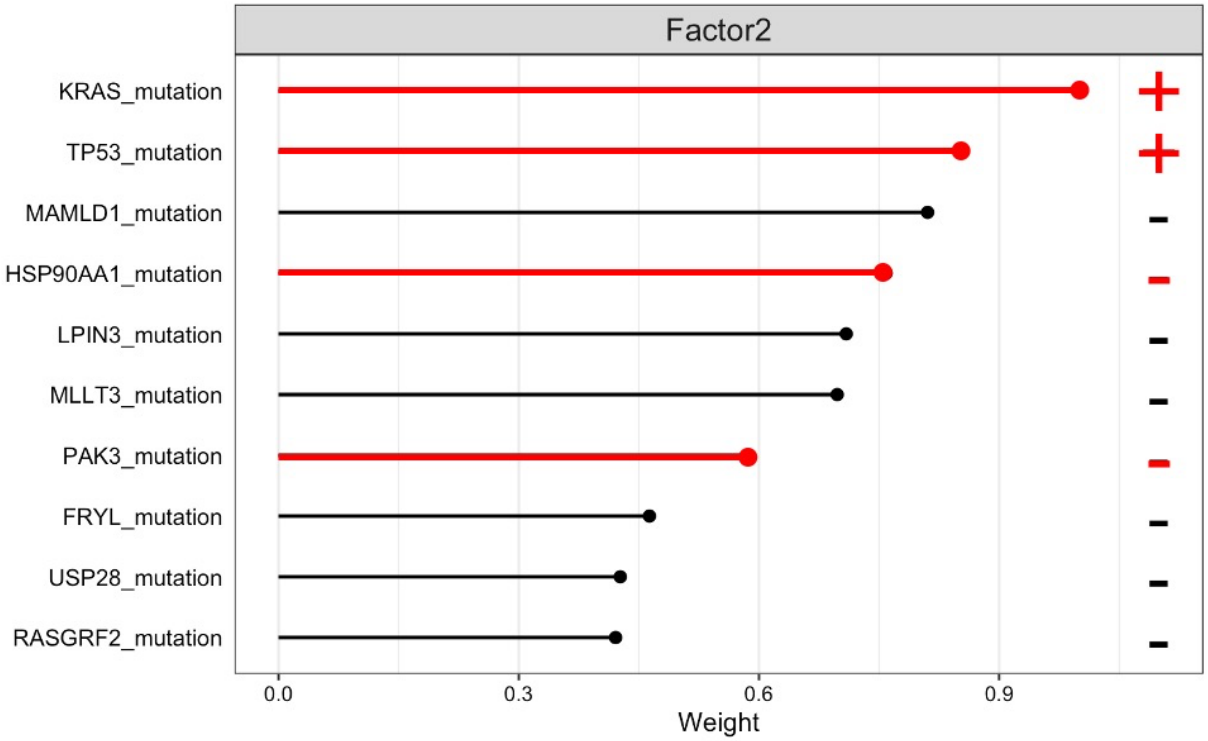


Figure 13. Top 10 features for mutation data modality within factor 2.

Line plot shows scaled feature weights with red lines indicating previous known associations with PDAC.

3.5.2.2 SCNV data modality and factor 2

The top 10 features of the SCNV view of factor 2 all show strong negative weights (Figure 14). The ETS-transcription factor ETV1 is the top feature and has been reported to promote PDAC metastasis and stromal expansions by mechanisms of epithelial-mesenchymal transition (Heeg et al., 2016).

Death receptor adaptor molecule TRAF2 has been identified as a key player in PDAC through an over-expression study and leading to apoptosis resistance (Trauzold et al., 2005).

The role of intracellular ion channels in PDAC has been investigated and one of the genes encoding a chloride channel *CLIC1* is implicated in oncogenic mechanisms when deregulated (Patel et al., 2019).

Nothing is known about the complement C8 gamma chain of C8G and a link to cancer in general. Also, no previously reported connection of *FBXW5* with PDAC but this gene seems to promote tumorigenesis and metastasis in gastric cancer (Yeo et al., 2019). The same is true for *LCN12* which has linked this and other immune-related genes to the tumor immune microenvironment in gastric cancer (Xu et al., 2021). Not much is known about the two long intergenic non-protein coding RNAs *C9orf141* and *C9orf142*. Albeit not directly linked to PDAC, LCNL1 has been implicated in shaping the tumor immune environment in bladder cancer (Lyu et al., 2021). The same is true for the prostaglandin D2 synthase PTGDS which deregulated expression can be seen in a variety of cancers.

Although the SCNV view had the strongest contribution to factor 2, and top ten features with high feature weights, only three genes could be confirmed in being related in PDAC tumorigenesis.

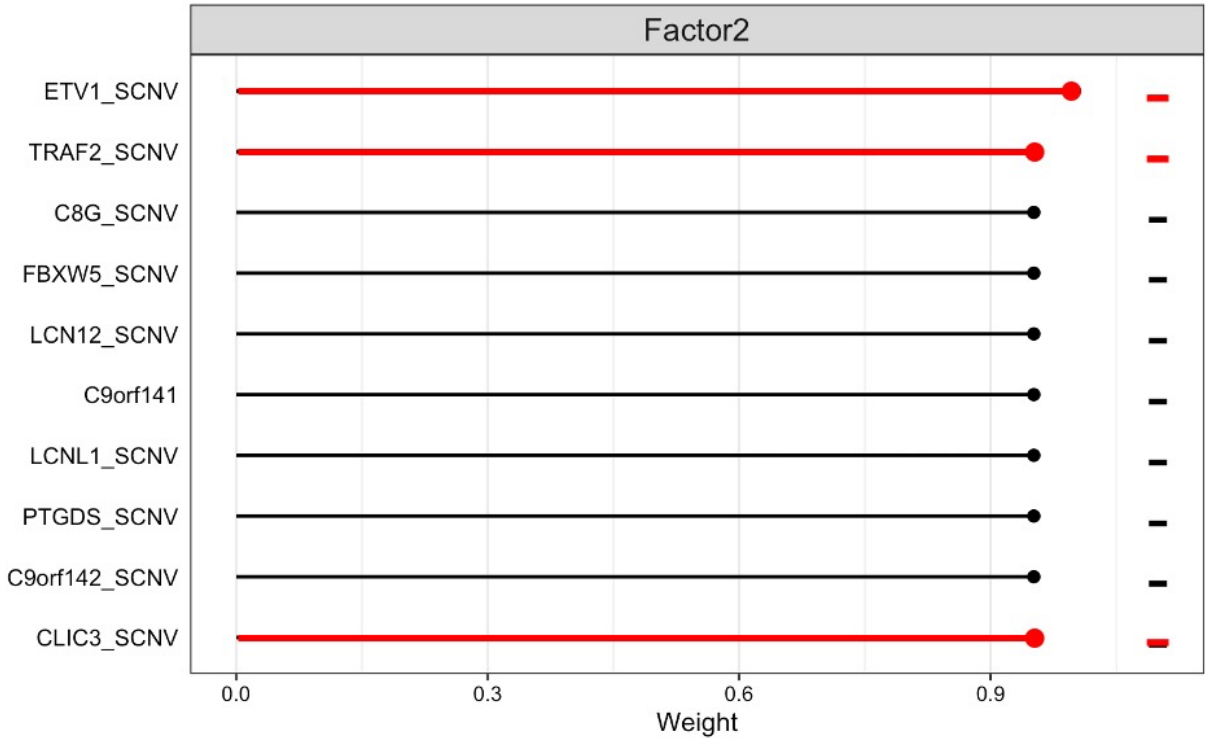


Figure 14. Top 10 features for SCNV data modality within factor 2.

Line plot shows scaled feature weights with red lines indicating previous known associations with PDAC.

3.5.2.3 mRNA data modality and factor 2

The top feature of the mRNA view with a strong negative sign is the glycine-N-acyltransferase like 3 *GLYATL3* for which there is only limited published information. Likewise, *DNAI2* hasn't been linked to pancreatic cancer in the literature before but has been identified as a differentially expressed gene in nasopharyngeal carcinoma with mechanisms involved in cilia movement (Ye et al., 2019). BPIFA2 (formerly C20orf70) is a salivary protein which has not been investigated in cancer but interestingly, *BPIFA2*-depleted mice show metabolic changes such as increased insulin production (Nandula et al., 2020). Little is also known for the solute carrier family 22 member 10 (*SLC22A10*) and its impact on tumorigenesis. It could be shown that *SLC22A10* is part of set of genes that are correlated with the pathological staging of hepatocellular carcinoma (Zhang et al., 2019). *TTC29*, a gene expressed in the testis and involved in cilia- and flagella movements shows up as fifth top feature with a strong negative weight. Its role in cancer has only been restricted to one computational study in gastric cancer where *TTC29* has been identified as a major driver of tumorigenesis as well as linked to increased mortality rates (Wang et al., 2020).

While the top five genes within the mRNA view were associated with a negative sign, the following genes with exception of the last *WDR16* have a positive weight. The lipase H encoding gene *LIPH* has no history in pancreatic cancer as well. However, a recent study in breast cancer patients found that over expression of *LIPH* was associated with increased metastasis and poor prognosis (Zhang et al., 2020).

Strong expression of the poliovirus receptor-related protein 4 (PVRL4) or nectin-4 which provides a host entry point for certain viruses has been detected in several cancers involving

pancreatic cancer using an immunohistochemistry approach. A possible pathological function of nectin-4 could be a role in cell polarity and the establishment and maintenance of adherens junctions which are affected when cancer progresses (Challita-Eid et al., 2016). The nectin-4 receptor is further exploited as entry point for potential oncolytic viruses in therapeutic approaches to pancreatic cancer (Awano et al., 2016). *VWA2* encodes the extracellular-matrix protein AMACO and has been reported as a putative biomarker for colorectal cancer. While *VWA2* is differentially expressed in tumor as compared healthy tissue in the colon, no differences in expression have been observed for PDAC (González et al., 2018). The tight junction protein 3 encoded by TJP3 has no direct association with pancreatic cancer but has been linked to suppressing EMT in colon cancer via a WNT-mediated signaling cascade (Nfonsam et al., 2019). WDR16 is a marker for ciliated cells and has been investigated as a potential cell type marker for patient stratification in endometrial carcinoma in a single-cell study (Cochrane et al., 2020).

Next to researching the connections to PDAC, the mRNA expression of the top 10 genes in the input data was visualized by a heatmap and confirming the status for upregulated and downregulated mRNAs as denoted by the positive and negative signs (Figure 15A).

Analyzing the features from the mRNA view shows some commonality such as the implication of the genes in cilia development or movement. The loss of primary cilia, organelles important in signal transduction, is a hallmark of tumorigenesis and has also been reported for PDAC (Kobayashi & Itoh, 2017). This adds to evidence that the tumor microenvironment which includes endothelial cells, immune cells, and cancer-associated fibroblasts among others, is an important determinant in tumor development and metastasis.

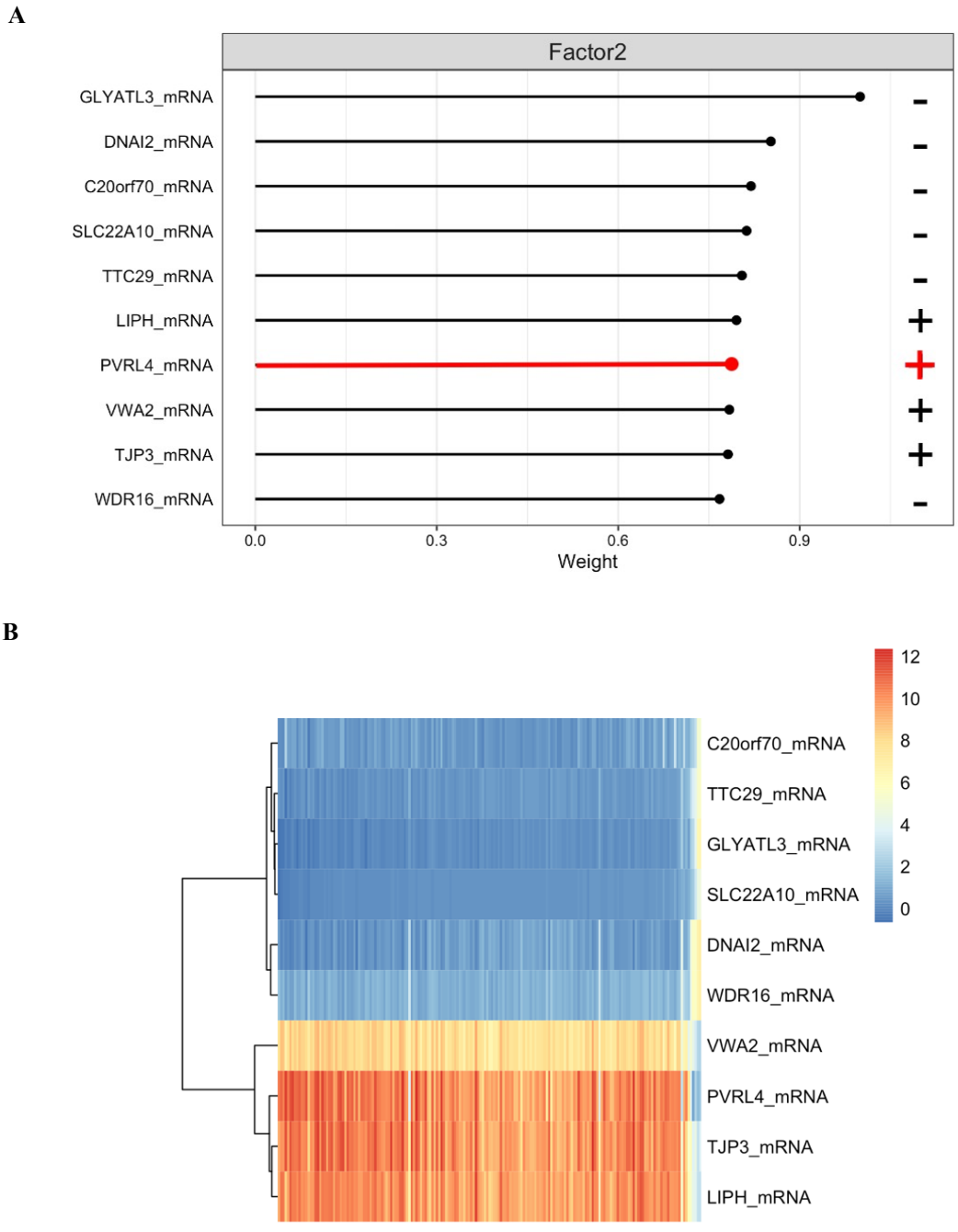


Figure 15. Top 10 features for mRNA data modality within factor 2.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows mRNA expression levels.

3.5.2.4 Methylation data modality and factor 2

The top two features of the methylation data are the chemokines CCL14 and CCL15 which are signaling proteins or cytokines secreted by cells and able to induce chemotaxis in nearby cells i.e., the movement of a cell in response to a chemical stimulus (Figure 16).

CCL14 has been shown to serve as a chemotactic agent for immune cells of the type monocytes and neutrophils. Its function in cancer can be either pro- or anti-cancer promoting. Although CCL14 can recruit monocytes to a tumor and promote the conversion to tumor-associated macrophages, it has also been shown that CCL14 can induce cancer migration in breast cancer (Korbecki et al., 2020). CCL15 acts as a chemotactic agent not only for monocytes but also eosinophils, and neutrophils and has been shown to be important for keeping immune balance in the gut and the liver. Like CCL14, CCL15 can have positive as well as negative effects within carcinogenesis by recruiting immune cells to the tumor or mediating tumor metastasis (Korbecki et al., 2020).

SLC26A3 and other plasma membrane transporters have been shown to act as tumor suppressors albeit not in PDAC (Bhutia et al., 2016). Similarly, COL15A1 has not been related to PDAC before but shows up in a variety of cancer studies. Also, *REP15* was linked to prognostic outcomes in colorectal cancer using a epigenome study but not in PDAC (Xu et al., 2020). No association with PDAC and the following genes *CASQ2, NAA11, GNG2, LOC340017* and *SLC10A6*.

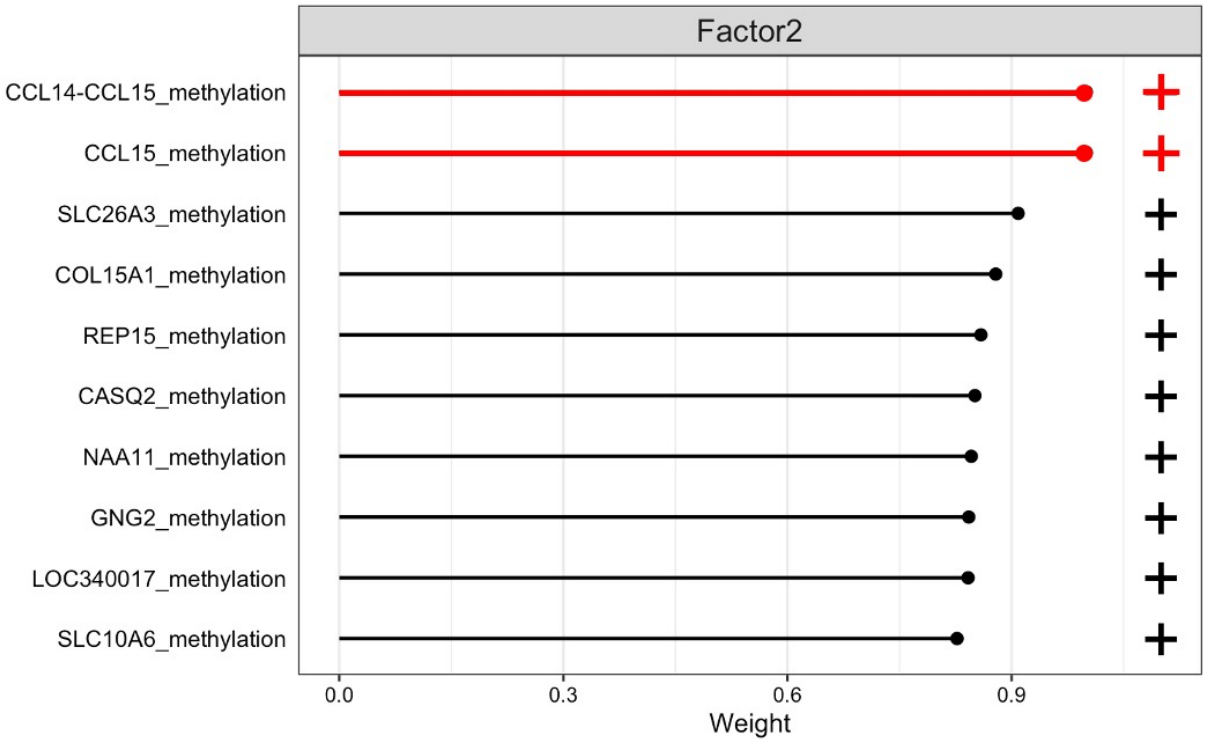


Figure 16. Top 10 features for methylation data modality within factor 2.

Line plot of scaled feature weights with red lines indicating previous known associations with PDAC.

3.5.2.5 miRNA data modality and factor 2

miR-1180 the top feature of the miRNA view with negative sign has been suggested to exert anti-cancer effects on pancreatic cancer when suppressed (Gu et al., 2017). It has also been noted as one of eight miRNAs serving as survival predictors for PDAC (Dou et al., 2018). miR-135, part of the feature signature and with a strong positive weight, is also among the eight miRNAs.

Differential expressed miR-376a has been reported to affect pancreatic tumorigenesis and has already been reported for factor 1 (Lee et al., 2007). Similarly, miR-210 has been identified in latent factor 1 of the miRNA view. And miR-376b joins as a further prognostic miRNA for pancreatic cancer (Liang et al., 2018).

While miR-7-2 has been retrieved as a major feature from factor 1 miRNA view, miR-7-3 is a top 10 feature component of factor 2. Aberrant over-expression of miR-7-2 and others by the MAPK pathway led to the inhibition of pancreatic cancer cells in culture (Ikeda et al., 2012).

Lastly, miR-485 has been linked as a prognostic marker in PDAC through a T cell immune-related miRNA regulatory network in a bioinformatics study (Gu et al., 2020). Seven out of ten listed miRNAs have been associated with PDAC before (Figure 17).

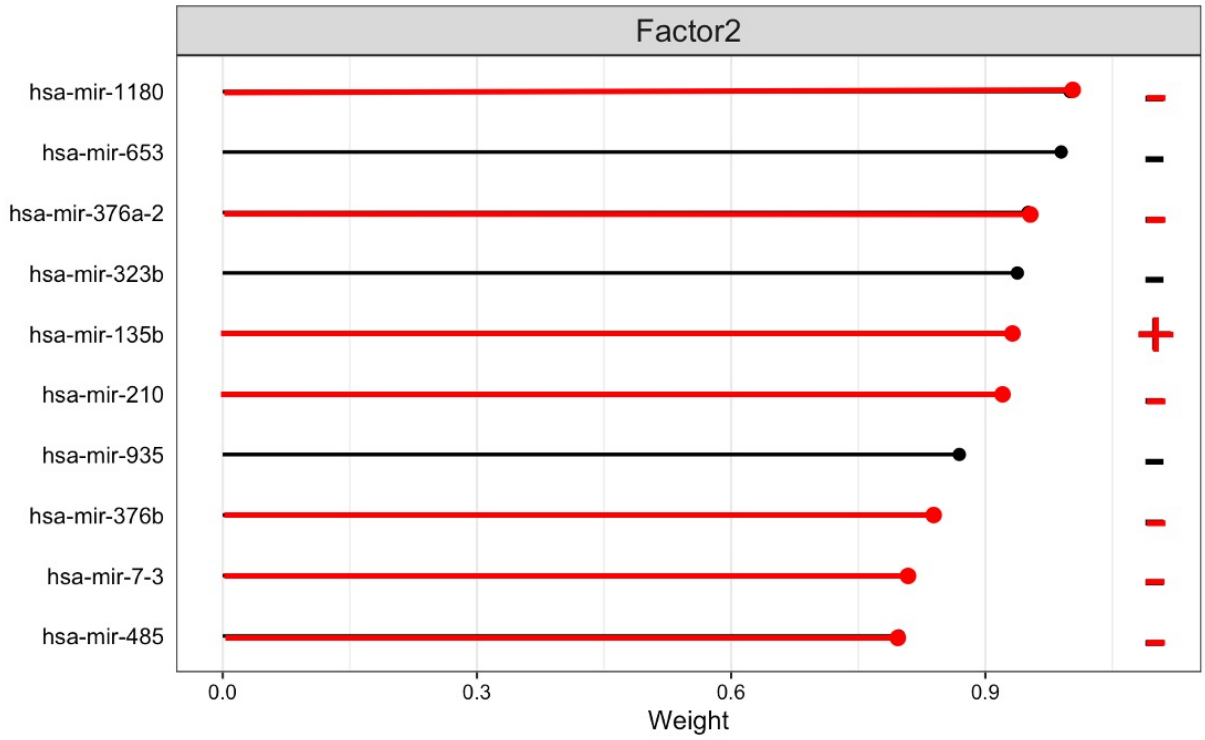


Figure 17. Top 10 features for miRNA data modality within factor 2.

Line plot of scaled feature weights with red lines indicating previous known associations with PDAC.

3.5.2.6 Protein data modality and factor 2

The protein data modality of factor 2 returns four features with negative signs and weights greater than 0.6 (Figure 18). The top feature found is the protein X-ray repair cross complementing 5, XRCC5, which has been linked in a single study to Rapamycin mediated radiosensitivity of pancreatic cancer cells (Dai et al., 2013).

Next, mutations in the serine/threonine kinase 11 STK11 have not only be linked as a component of familial cancer in patients with Peutz-Jeghers Syndrome (Hruban et al., 1999) but it functions also as a tumor suppressor in acinar cell carcinoma a form of pancreatic cancer through loss of mTOR inhibition (Heestand & Kurzrock, 2015).

The protein ALK1 encoded by ACVRL1 gene and implicated in EMT has been proposed as a pharmacologic target in several cancers and PDAC specifically to block tumor angiogenesis which is a further cancer hallmark (Eleftheriou et al., 2016). Similarly, the E-cadherin CDH1 is implicated in regulating EMT in PDAC and thus promoting metastasis (Sato et al., 2020). The Src homolog and collagen homolog 1 SHC1 has been shown to be involved in the collagen induced N-cadherin up-regulation in a mouse cell line a hallmark of EMT in pancreatic cancer (Huang et al., 2016). Further evidence to the biological mechanism of the factor 2 protein view adds the fact that the retrieved feature WWTR1 (also TAZ) and Hippo pathway transducer functions as an oncogene in pancreatic cancer through promotion of EMT (Xie et al., 2015).

The asparagine synthetase ASNS has previously been shown to promote metastasis and cancer cell survival with the majority of PDAC cells expressing no or low ASNS rendering them unable to produce the amino acid asparagine (Cohen et al., 2015). While the feature ASNS points

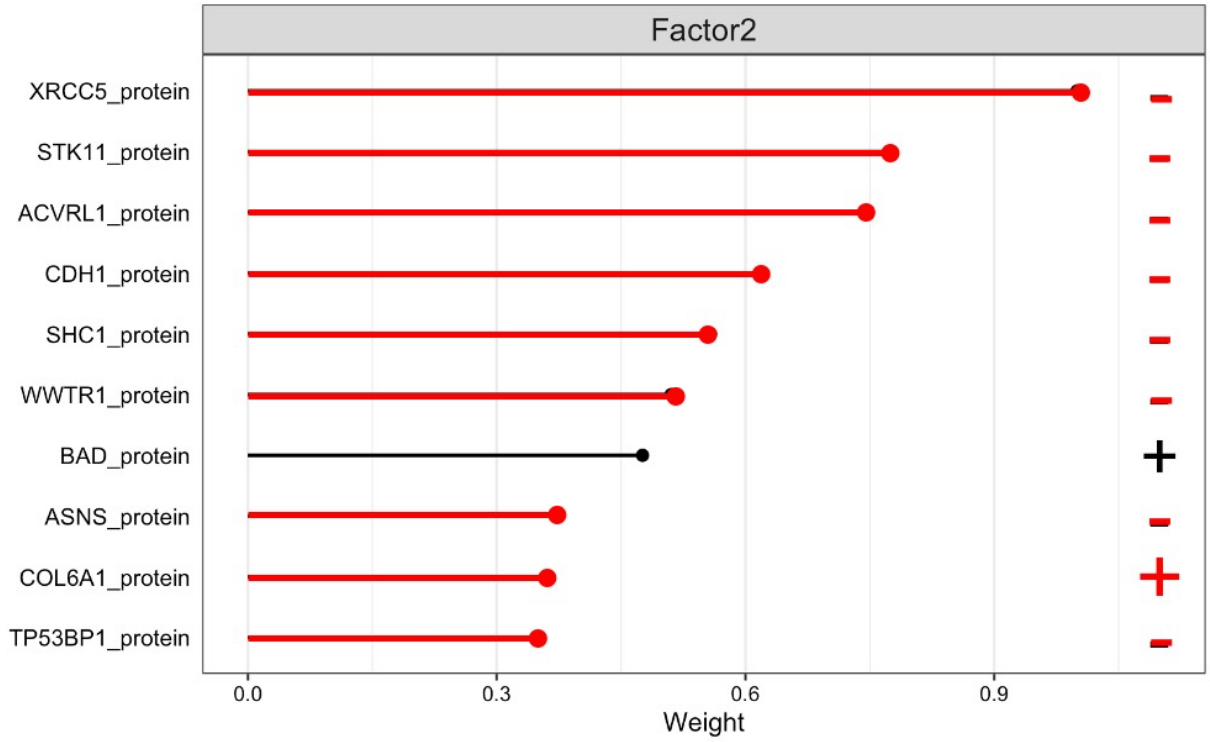


Figure 18. Top 10 features for protein data modality within factor 2.

Line plot of scaled feature weights with red lines indicating previous known associations with PDAC.

to a deregulated cancer cell metabolism, COL6A1 has been described as a promoter of metastasis and poor prognosis and shows up with a positive feature weight in the protein data modality of factor 2 (Owusu-Ansah et al., 2019).

Lastly tumor protein p53 binding protein 1 TP53BP1 is no stranger in PDAC as it binds to the tumor suppressor p53 a key tumor suppressor and regulates DNA damage and repair. Within PDAC its role hasn't been fully characterized but there is reasoning that TP53BP1 alters tumor behavior (Ausborn et al., 2013).

3.5.3 Characterization of MOFA factor 3

MOFA factor 3 is interesting, because it captures more variability than factor 2. In summary, it was able to capture, 10% within the methylation, 8.7% within the miRNA, 8.7% within the mRNA, 3.5% within the protein data, and no variation within the SCNV data. Again, the mutation data could hardly explain any variation (0.00052%). Interestingly, within this factor the SCNV data modality doesn't contribute to the variation as well.

3.5.3.1 Mutation data modality and factor 3

The two top features within the mutation data of factor 3 are TP53 and KRAS which also showed up in the previous two factors. However, this time, the weights of both features are associated with a strong positive sign.

PAK1 has been shown to play a role in tumor immunity by promoting the infiltration of CD4+ and Cd8+ T cells and further killing pancreatic cancer cells (Wang et al., 2020).

None of the other listed features were associated previously with PDAC and due to the lower weights were not further investigated (data not shown).

3.5.3.2 SCNV data modality and factor 3

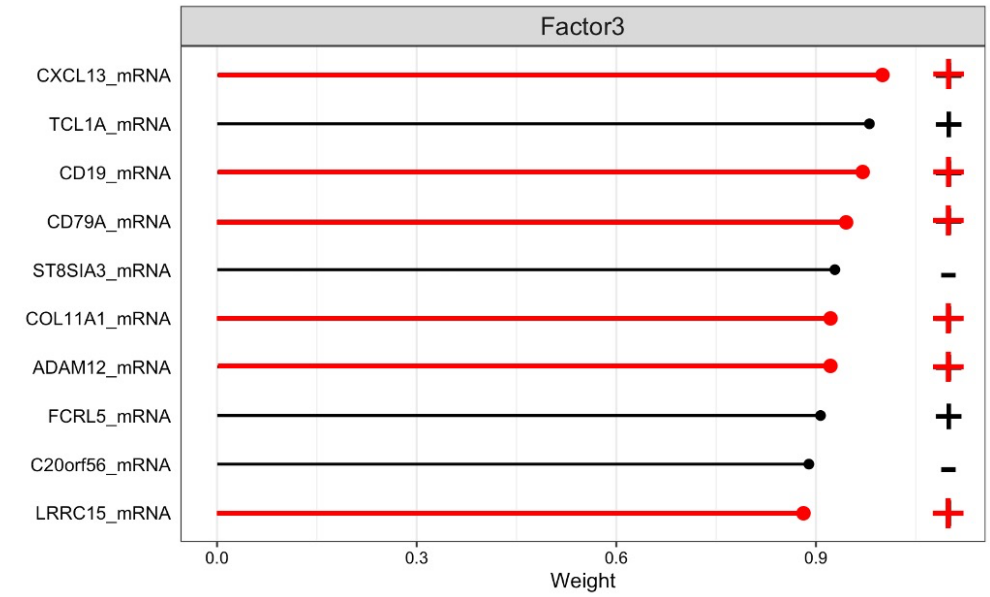
None of the retrieved ten features showed a connection to PDAC using a literature search. No further investigation was done in understanding the features in different biological contexts or in other cancers because of the negligible contribution of the SCNV data to variability in this factor (data not shown).

3.5.3.3 mRNA data modality and factor 3

Apart from *ST8SIA3* and *C20orf56* all other features harbor strong positive weights (Figure 19). *CXCL13* is part of set of targets of non-canonical NF-kappaB signaling which is constitutively active in pancreatic tumor (Wharry et al., 2009). *CD19* and *CD79A* point to molecules on immune cells and thus, a role in immune regulation. The differentially regulated *COL11A1* gene has been identified as a prognostic marker for PDAC with relation to immune filtration (Sun et al., 2018).

ADAM12 like other A Disintegrin and Metalloproteases is overexpressed in PDAC patients and associated with poor outcomes and linked to immune infiltration (Qi et al., 2020; Veenstra et al., 2018). Further evidence for a role in immune cell infiltration comes from the leucine-rich-repeat containing 15 (*LRR15*) gene which is expressed in stromal fibroblasts of pancreatic cancer (Purcell et al., 2018).

A



B

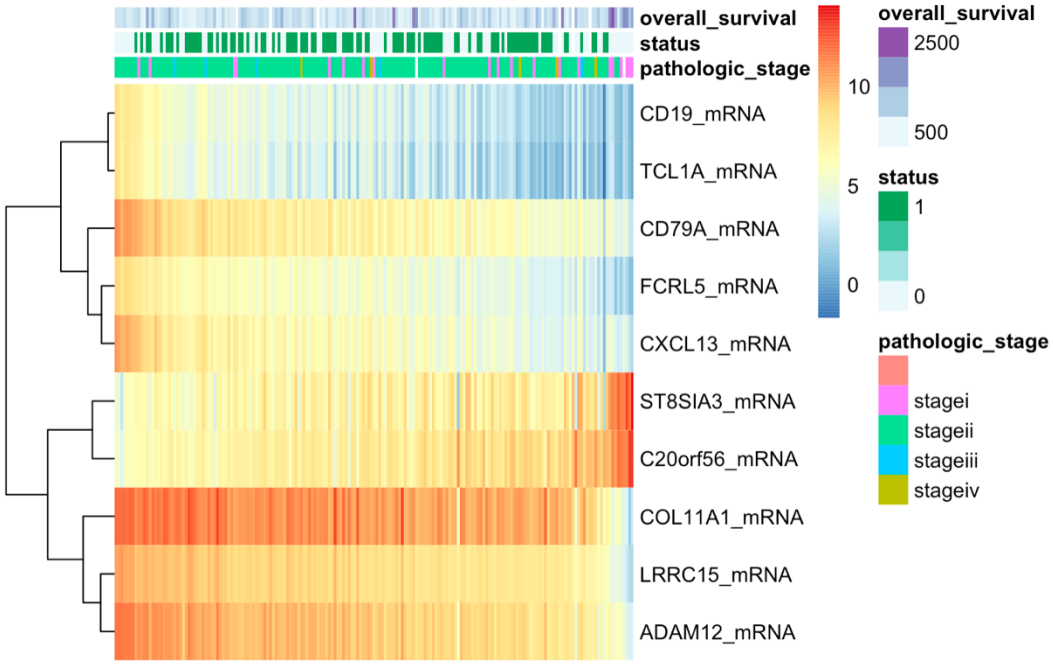


Figure 19. Top 10 features for mRNA data modality within factor 2.

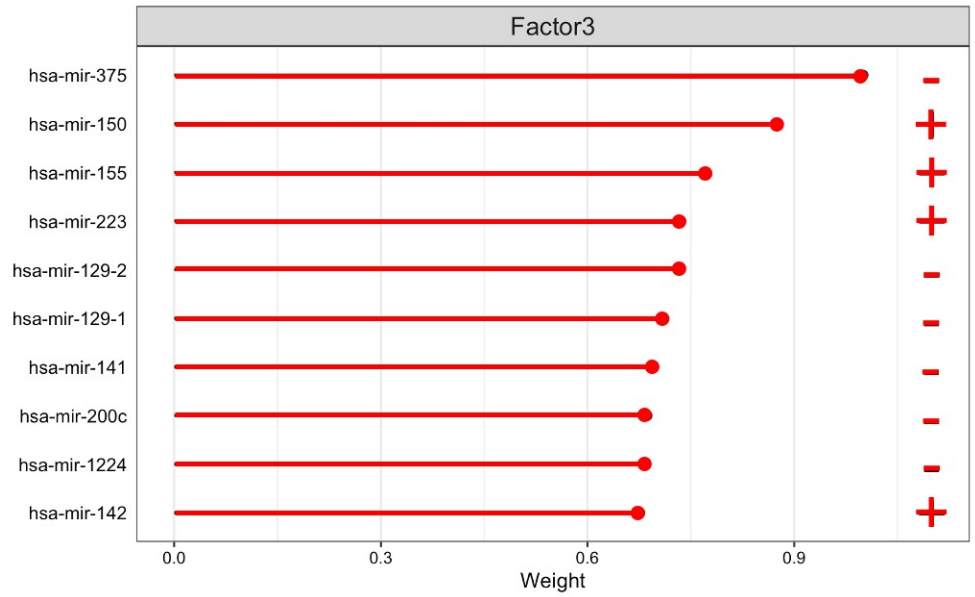
(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows mRNA expression pattern, annotated with overall survival, vital status, and pathologic stage for patient samples (N = 185).

3.5.3.4 microRNA data modality of factor 3

miR-375 represents the top feature of the microRNA data within factor 3 and has been shown to exert a negative regulation on the expression of PDK1. miR-375 can be differentially expressed with downregulation promoting cell growth and inhibiting cell apoptosis (Zhou et al., 2014). miR-150 has also been shown to be implicated in the malignancy of pancreatic cancer as a tumor suppressor (Arora et al., 2014). A similar role in regulating the proliferation and apoptosis of pancreatic cancer cells exhibits miR-115 (Wang et al., 2020). No stranger in the pancreatic microRNA regulation landscape is miR-223 which has been shown to promote cell proliferation and invasion in pancreatic cancer (Ma et al., 2019). The two miRNAs miR-129-1 and miR-129-2 as well as miR-1224 have already been described in the factor 1 features of the miRNA modality with both having been associated with PDAC before. Downregulated miR-141 has been observed in pancreatic tumorigenesis with an impact on tumor size and TNM staging as well as metastasis and association with lymph nodes (Zhao et al., 2013). A lot is also known about miR-200s which together with miR-141 is part of the miR-200 family and its implication in cancer (Huang et al., 2019). Lastly, miR-142 has also been described as a regulator of pancreatic cancer cell proliferation and apoptosis (Yao et al., 2019).

In summary, all retrieved features from the miRNA view have previous associations with pancreatic cancer pointing to complex post-transcriptional regulatory mechanisms of gene regulation (Figure 20).

A



B

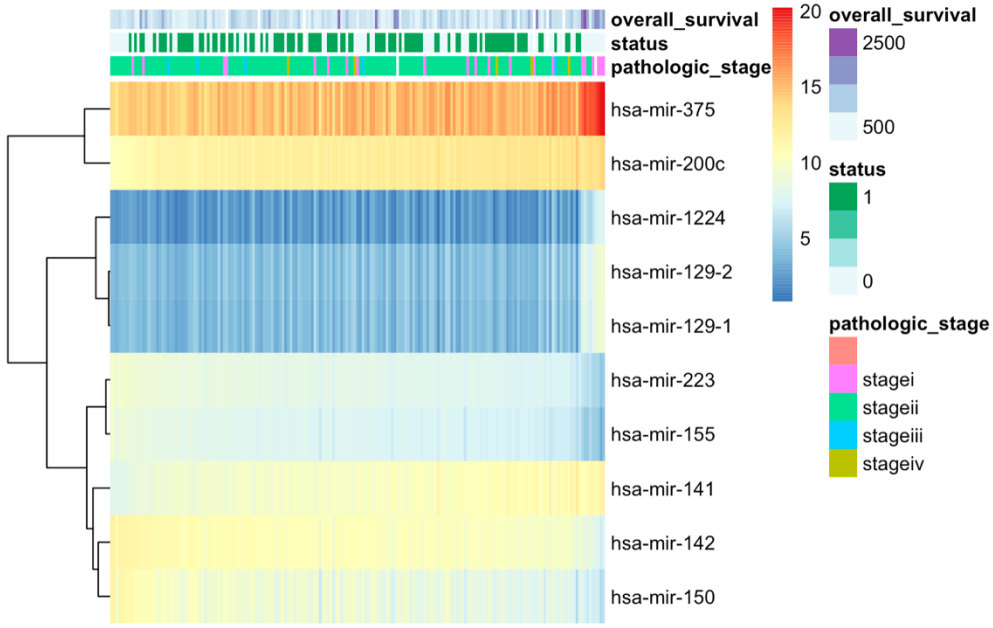


Figure 20. Top 10 features for miRNA data modality within factor 3.

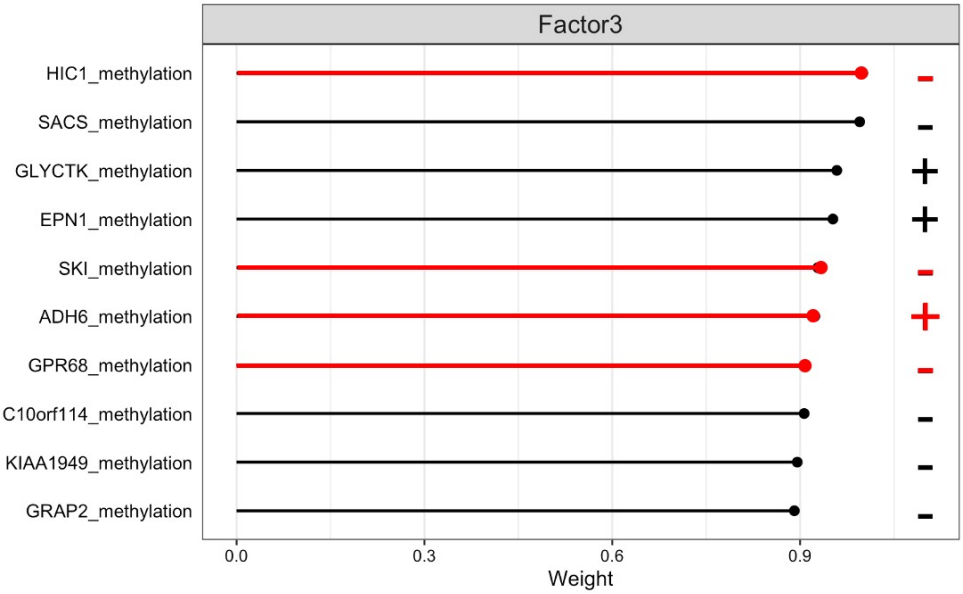
(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows miRNA expression pattern, annotated with overall survival, vital status, and pathologic stage for patient samples (N = 185).

3.5.3.5 Methylation data modality of factor 3

The top ten features retrieved from the methylation data modality of factor 3 all have very high weights which are mostly negative (Figure 21). DNA promoter hypermethylation is a hallmark of cancer and has been observed in PDAC for the *HIC1* promoter region (Zhao et al., 2013). Conversely, the promoter region of the oncogene *SKI* has been shown to be hypomethylated (Kinugawa et al., 2018) and a role in NK cell-mediated immunosurveillance has been proposed (Ponath et al., 2020). The alcohol dehydrogenase ADH6 as well as other members of this family have been shown to be potential prognostic markers for PDAC capable of predicting better prognosis (Liao et al., 2017). The expression of G-protein-coupled receptor *GPR68* is upregulated in pancreatic cancer where it role has been associated with defining the tumor microenvironment (Wiley et al., 2019).

An integrated multi-omics analysis of genomic, epigenomic and transcriptomic data from the TCGA cohort identified four molecular subtypes of PDAC and a set of genes as prognostic markers. GRAP2 showing up in tenth position of the features from the methylation data was one of the described markers (Kong, et al., 2020).

A



B

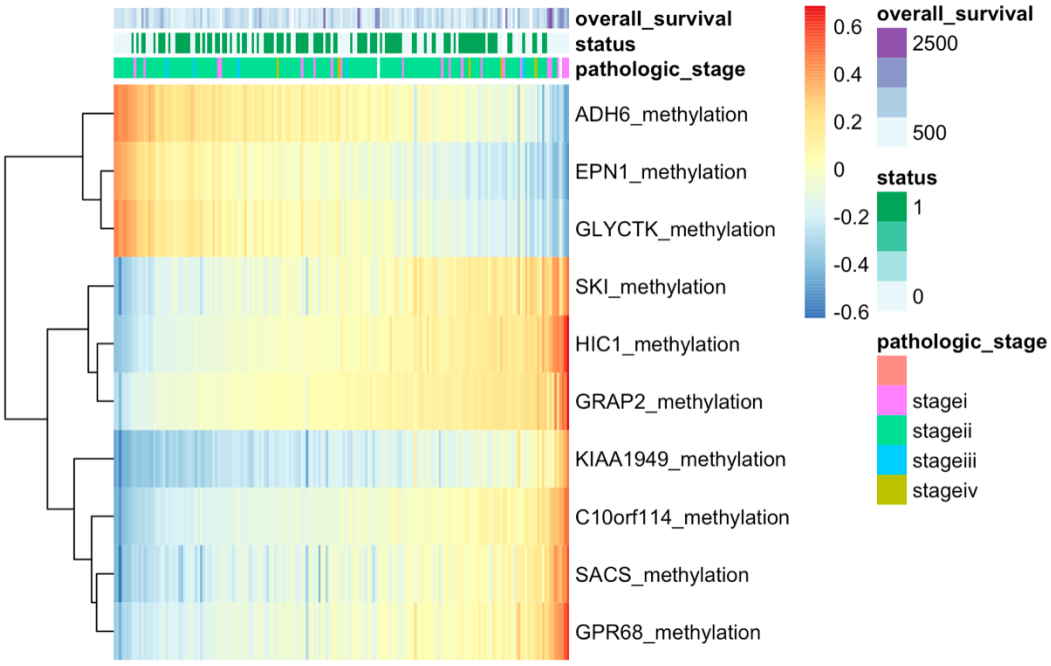


Figure 21. Top 10 features for methylation data modality within factor 3.

(A) Line plot of scaled feature weights with red lines indicating previous known associations with PDAC. (B) Heatmap shows methylation pattern, annotated with overall survival, vital status, and pathologic stage for patient samples (N = 185).

3.5.3.6 Protein data modality of factor 3

The ten most significant features within the protein modality of factor 3 come with both negative and positive signs (Figure 22). Some proteins already showed up in previous analyses such as the ARAF protein in factor 1 and the BAD protein in factor 2. The DNA repair protein RAD50 showed up in the top position and elevated levels have been observed in PDAC previously where it has also been suggested as an early biomarker (Jia et al., 2020). A wealth of information has been published on the implication of STAT3 in PDAC and its importance in pancreatic cancer pathogenesis mediated by STAT and MAPK signaling pathways (Ligorio et al., 2019). Next, TSC1 reduction has been linked to activation of KRAS/MEK/ERK-mediated mTOR signaling and further promoting PDAC metastasis in a transgenic mouse model (Kong et al., 2016). Fibronectin 1 (FN1) has been shown to be expressed in the stroma and indicative of the existence of macrophages and associated with poor PDAC prognosis (Hiroshima et al., 2020). Furthermore the insulin receptor substrate 1 has been shown to be implicated in the proliferation, invasion and metastasis in pancreatic cancer (Huang et al., 2018). In factor 2, CDH1 has been described as a top feature of the protein modality, within factor 3 it is CDH3. Both cadherins have been determinants of aggressiveness within PDAC with Cadherin-1 driving type I collagen organization within the tumor as well as invadopodia activity and Cadherin-3 regulating tumor growth by cell migration (Siret et al., 2018). Lastly, the translation elongation factor eEF2 has been described to be over expressed in many tumors such as pancreatic cancer resulting in an oncogenic role (Oji et al., 2014). The research on the genes retrieved by the different data modalities of factor 3 hasn't been consistent in a possible underlying mechanism. There could be an immune system relevant component or more general molecular mechanisms like cell proliferation and metastasis.

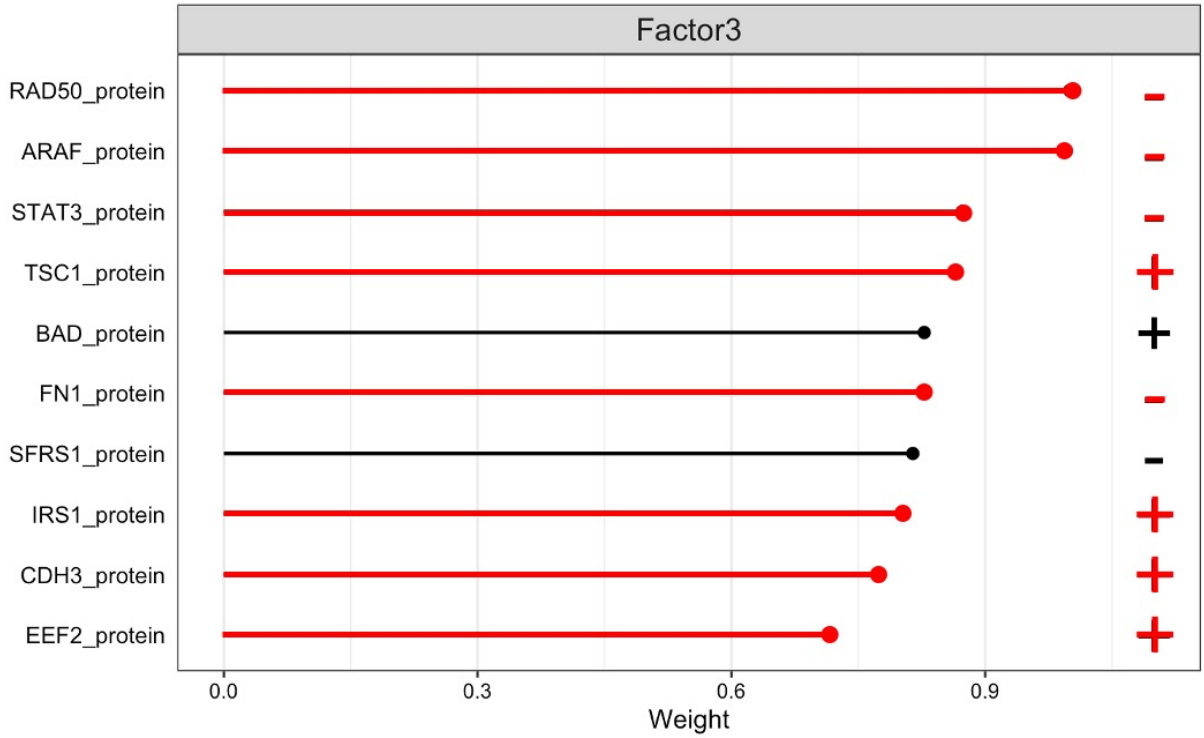


Figure 22. Top 10 feature weights for protein data modality within factor 3.

CHAPTER 4 RESULTS – DOWNSTREAM ANALYSIS OF DERIVED FACTORS

4.1 Gene set enrichment analysis reveals biological signature of retrieved factors

To get more mechanistic insights into the identified factors and their weights, Gene Set Enrichment Analysis (GSEA) was performed on factors 1, 2 and, 3. GSEA performs functional annotation of the factors by inferring associated biological functions and molecular processes based on the statistical assessment of enrichment. Using the gene sets from mRNA and methylation data modalities separated by plus and minus signs, a comparison to the Gene Ontology Biological Pathways (BP) reference set was performed. The reason for a separate analysis of features with plus and minus weights is that they might relate to different pathways that cannot be untangled when analysis is performed jointly.

The GSEA on the positive feature set from the mRNA modality of factor 1 showed a statistically significant enrichment of cell-cell signaling pathways (Supplemental Figure A, Appendix). In particular synaptic signaling relating to neurogenesis processes were retrieved as major biological process (Figure 23A). This data was confirmed from GSEA on the methylation feature set from the methylation modality of factor 1 with negative weights (Figure 23C). This mutual relationship points to an interconnected regulation between epigenetic modifications and transcriptional activity.

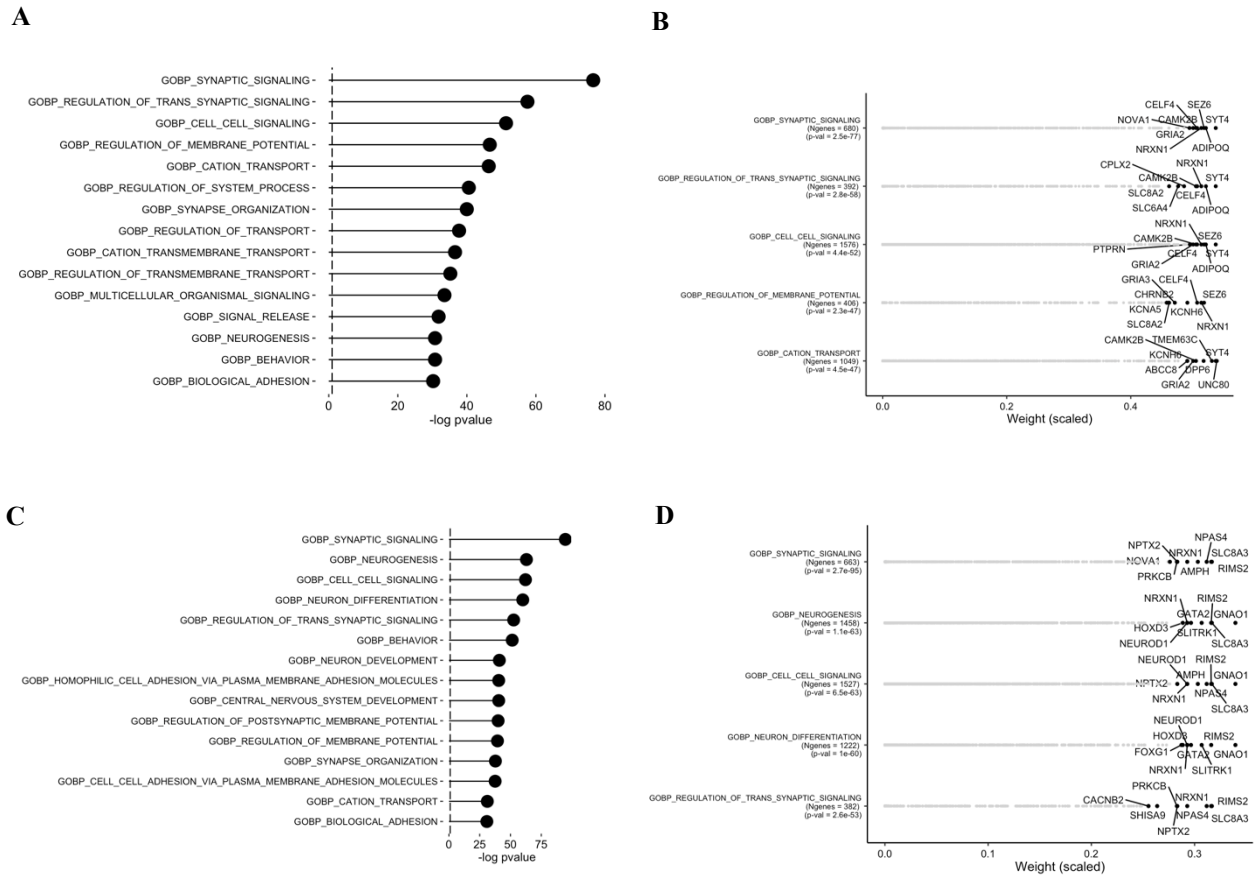


Figure 23. Gene set enrichment analysis for factor 1 on mRNA feature weights with positive sign (A) and methylation feature weights with negative sign (C). The top 15 biological pathways for each analysis are listed as line graphs. The signature genes driving the top five pathways are displayed in the right column with associated scaled feature weights.

The results for the GSEA on factor 2 didn't lead to statistically enriched pathways. But results from GSEA on factor 3 using mRNA and methylation data modalities were consistent in being related to immune system processes (Figure 24). The top 15 retrieved biological pathways within the mRNA data modality of factor 3 with positive weights were strongly enriched for immune system processes and defense response. This result was reflected in the GSEA analysis of factor 3 using the methylation features with negative feature weights. Similarly, as in factor 1, those two enrichment analyses are complementary in that strong positive feature weights of the mRNA gene set and negative feature weights of the methylation data are implicated in the same processes. Regulation of immune processes are a hallmark of cancer as avoiding immune destruction is important for tumor cells to strive (Hanahan & Weinberg, 2000).

GSEA was also performed on the other data modalities but didn't show any similar significant enrichment (data not shown). Interestingly, the methylation features with negative weights, further showed significant enrichment within factors 4 and 5. While factor 4 seems to capture variation in the sensory perception as well as the detection of chemical stimuli, factor 5 most likely is implicated in structural organization and adhesion (Figure 25).

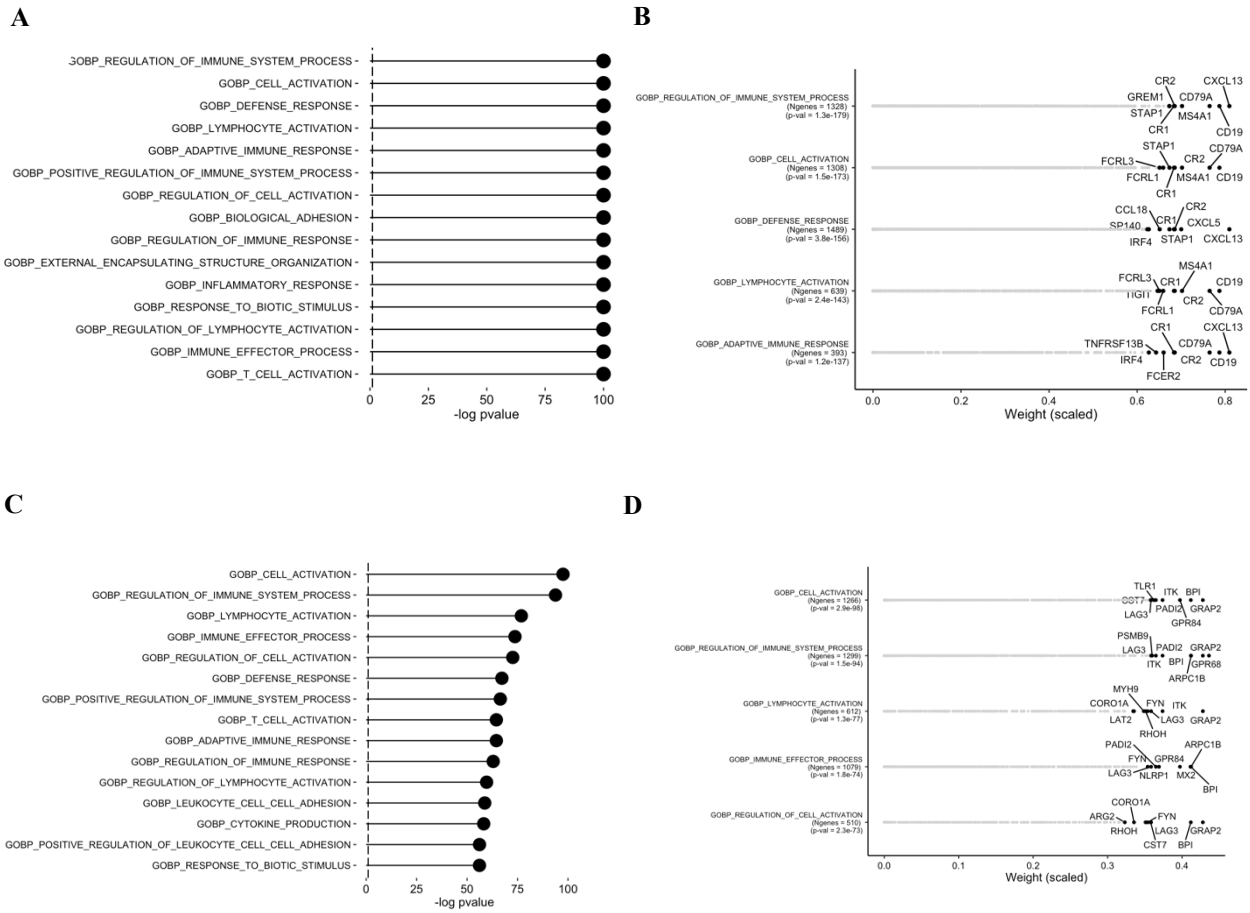
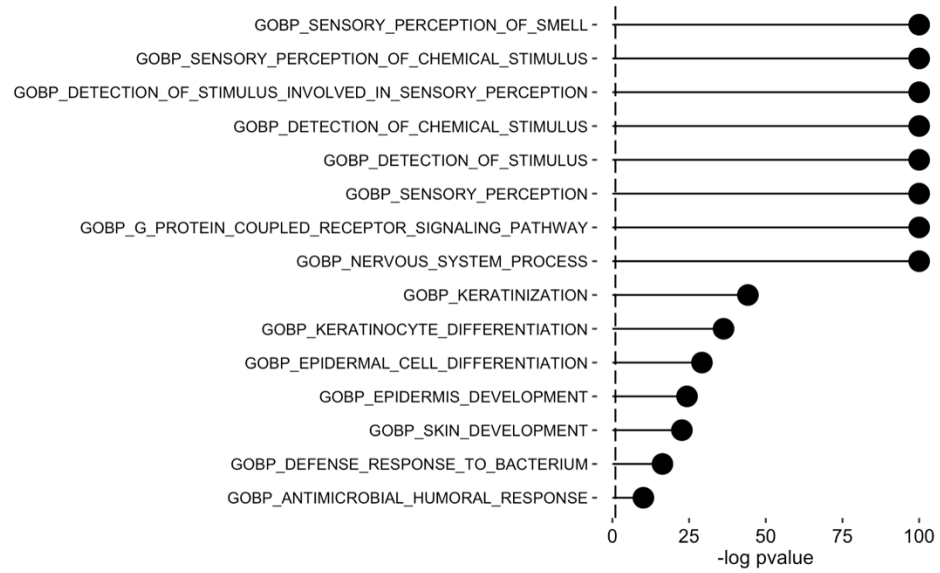


Figure 24. Gene set enrichment analysis for factor 3 on mRNA feature weights with positive sign (A) and methylation feature weights with negative sign (C). The top 15 biological pathways for each analysis are listed as line graphs. The signature genes driving the top five pathways are displayed in the right column with associated scaled feature weights.

A



B

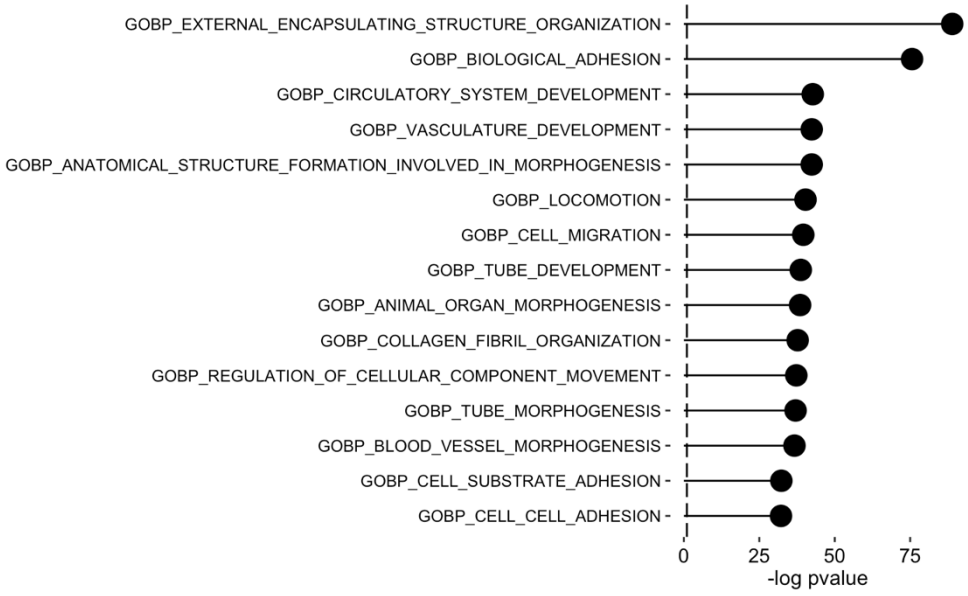


Figure 25. Gene set enrichment analysis for factor 4 and 5 on methylation feature weights with negative sign. The top 15 biological pathways for each analysis are listed as line graphs for factor 4 (A) and factor 5 (B).

4.2 Clustering of PDAC samples in low-dimensional space

Although continuous, MOFA factors can be used to perform clustering of samples. Since the first three factors seem to confer most of the heterogeneity, clustering of patient samples was performed using factors 1, 2 and, 3. The *MOFAtools* package contains the k-Means clustering algorithm which was applied using six clusters. Although a good separation of samples in the six clusters could be achieved (cluster sizes: 51, 41, 22, 43, 9 and 19) as well as the measure of goodness was at 83.6%, the cluster annotation was not helpful for visualization of samples. No aggregated pattern of patients was observed when plotting the samples in two-dimensional space of factor 3 and factor 1. Similarly, applying t-SNE (t-distributed stochastic neighbor embedding) or UMAP (Uniform Manifold Approximation and Projection) projections didn't help in visualizing the clustering result in lower-dimensional space (data not shown).

The reason for the promising clustering result but the failure to visualize its outcome might be the underlying heterogeneity which has been traced to various biological processes but still remains too variable to be captured. Of note is also the choice of the number of factors for the k-Means algorithm as well as the number of classes employed. Results didn't improve with the number of factors included in the analysis suggesting that the top factors are most relevant. Variation of the number of factors did have an impact on the result with k equals ten clusters leading to best fit. The caveat with k-Means clustering is that the number of clusters needs to be determined in advance. For further research, an elbow plot can be used to empirically infer a suitable choice of clusters. Moreover, k-Means might not be the optimal choice for a clustering algorithm on this data set but other algorithms like density-based clustering algorithms might perform better.

However, the factors can be used to perform clustering in patient samples possibly leading to better patient stratification.

4.3 Survival analysis

4.3.1 Association with clinical covariates

To relate the retrieved factors to clinical outcome, an association analysis was performed which correlates clinical metadata to the MOFA factors. The association between all factors and overall survival, pathologic staging (TNM and I-IV), gender, ethnicity, race, residual tumor, and radiation therapy was plotted as log10 adjusted p-values (Figure 26A). While most of the factors don't relate to any of the clinical variables, factors 1, 2 and, 3 correlates with overall survival with factor 3 having the strongest correlation. Factor 1 further correlates with the pathologic staging, in particular with the T staging and further indicating a correlation with vital status. Factors 2 and 4 strongly correlate with the histological type of the pancreatic tumor. Next to overall survival, factor 3 correlates with pathological staging of tumor size (T) and nearby node invasion (N). Investigating the nature of the correlation in more detail by plotting the correlation strength and direction per factor and covariate shows that factor 1 is negatively correlated to the pathological staging while impacting overall survival in a positive correlation. The reverse is true for factors 3 and to a lesser extent factor 2 where these factors are positively correlated with pathologic staging and negatively correlated with overall survival (Figure 26B).

In summary, the retrieved latent factors capture variability within the PDAC cohort impacting tumor staging and overall survival.

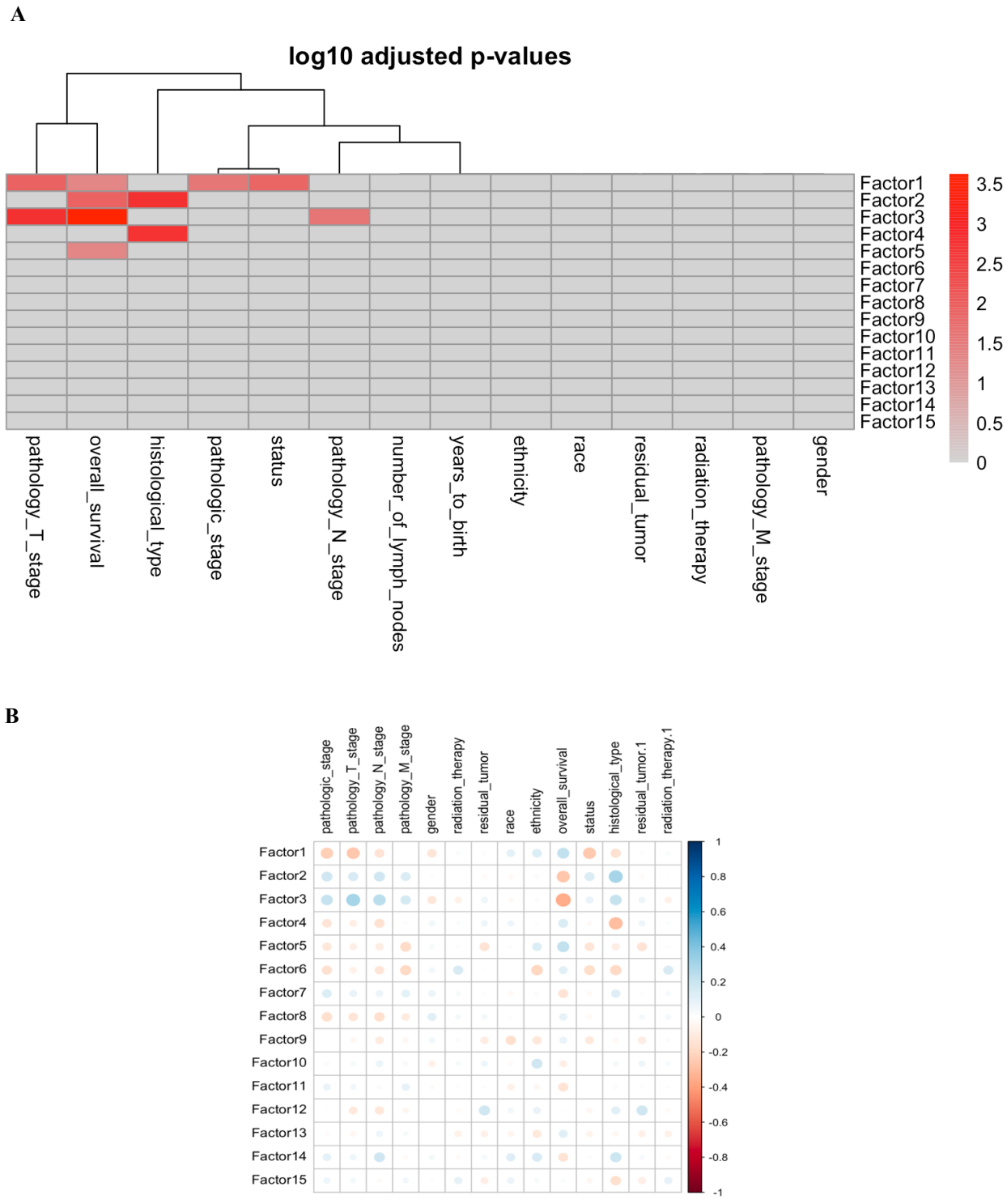


Figure 26. Association Analysis of retrieved latent MOFA factors with clinical covariates. (A) Heatmap of correlation with covariates as log10 adjusted p-values. (B) Correlation plot of factors with clinical covariates indicating the strength and the direction of correlation.

4.3.2 Survival prediction model

Since some of the latent factors were related to overall survival, it was investigated if the factors can be used in a predictive model for clinical outcome. The Cox proportional hazards model was chosen to explore the influence of the 15 factors on patient survival. The model explores the influence of the specified factors and the occurrence of the hazard death. The returned hazard ratios were visualized as line graphs and denote no effect if the hazard ratio is equal to 1, an increased hazard if hazard ratio is greater than 1 (bad prognostic factor) and a reduction in the hazard if the hazard ratio is smaller than 1 (good prognostic factor). The results indicate that factors 1, 5 and, 6 are good prognostic factors with significant smaller hazard ratios. Factor 3 seems to be a bad prognostic factor as indicated by the greater than one hazard ratio but in in this analysis it was not significant (Figure 27).

Further visualization of the cox model using Kaplan-Meier curves shows that the levels of the factors have an impact on the length of survival. High levels of factor 1, factor 5 and factor 6 are related to prolonged survival with factor 1 having the strongest impact. Conversely, low levels of factor 2 and factor 3 are indicative of increased survival length. Factor 1 and factor 3 exhibit the strongest effects on survival duration and stand for good and bad prognostic outcomes, respectively (Figure 28).

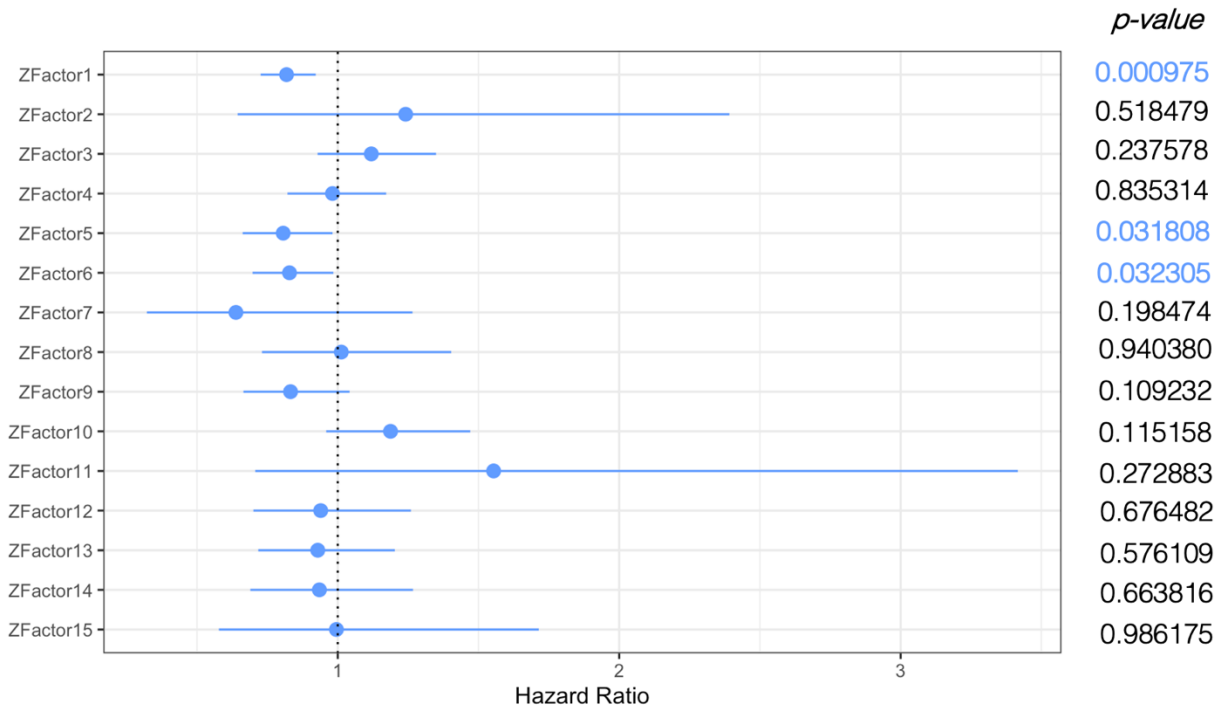


Figure 27. Cox proportional hazards model applied to PDAC MOFA analysis.

Association of MOFA factors with overall survival time using a univariate Cox regression with N = 185 samples. Error bars denote 95% confidence intervals. Numbers on the right show the *p*-values for each predictor.

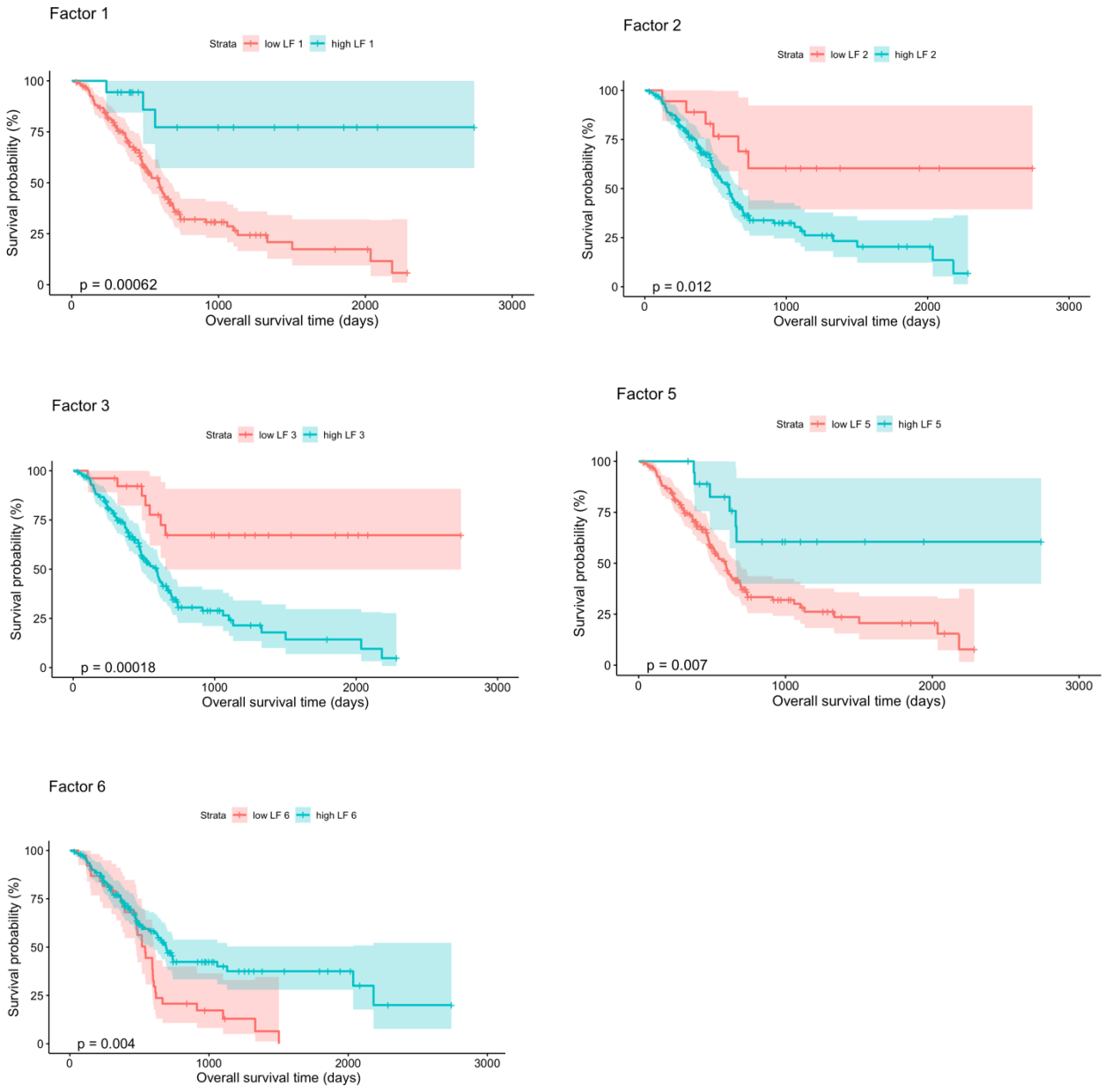


Figure 28. Kaplan-Meier plots measuring the overall survival for the individual MOFA factors.

The cut-points on each factor were chosen using maximally selected rank statistics, and *p*-values were calculated using a log-rank test on resulting groups.

CHAPTER 5 CONCLUSIONS AND PERSPECTIVES

In this research the successful integration of multiple omics data from a pancreatic cancer cohort has been demonstrated. Using publicly available data from TCGA and comprising high-throughput data from the genome to the proteome, the Multi-Omics Factor Analysis (MOFA) framework was employed to infer the major sources of heterogeneity driving inter-tumor variability in PDAC. The MOFA model has proved as a valuable tool for this endeavor because it is capable of integrating heterogenous data from various molecular assays that differ in the type of data as well as the scale it was recorded and the way it has been normalized. Furthermore, none of the samples had a complete analysis at each molecular layer but showed missingness to a varying degree. As opposed to other integration models, MOFA is capable of dealing with the sparsity. In this study, MOFA has also proved to work well with data of various dimensionality.

The use of preprocessed, curated omics data is a limitation in this research as it requires trust in the source in terms of accuracy and completeness. Little was known about how the data was processed in detail and which quality measures were in place. The methods and algorithms used for processing the raw data was only vaguely described on the LinkedOmics platform or the publication (Vasaikar et al., 2018), and the data curators didn't respond to inquiries. However, initial inspection of the data and descriptive statistical analyses indicated a good quality dataset.

Next to being able to incorporate data from different molecular assays, MOFA has also contributed to understanding the between patient heterogeneity within PDAC considering

information from the genome to the proteome. So far, no other study has investigated the sources of heterogeneity in such a comprehensive manner. The top retrieved latent factors were able to capture the global sources of variability within the data. Factors 1 and 2 were shared across all but the mutation data modality implying that they capture heterogeneity from the genome to the proteome layer. The mutation data didn't contribute to explaining the observed variability at all suggesting that those mutations are not the underlying source of heterogeneity but merely implicated in initiating the cancerous trajectory. This is consistent with previous studies that have associated mutations in KRAS with triggering the progression of PDAC (Storz & Crawford, 2020).

The major sources of inter-tumor variation were identified as neurological signaling within factor 1 and immune system evasion within factor 3. Tumor-induced neurogenesis hasn't been a cancer hallmark in the past but the role of nerves in tumorigenesis in some cancers is becoming more and more important in determining cancer progression. The tumor microenvironment i.e., the tissue surrounding the tumor are coopted by tumors to favor growth and metastasis. Nervous and immune cells, amongst other various types of cells, are harnessed by the cancer cells in complex communication circuits to fuel the expansion of the tumor (Cervantes-Villagrana et al., 2020). While this study has inferred the neuro-immune axis as the determining component of heterogeneity within PDAC further investigation in the underlying genes and molecular pathways is necessary. This requires deeper literature searches and connecting the results from the individual factor analyses with the enriched biological pathways from the GSEA.

Furthermore, a comparison of the PDAC MOFA model with a PDAC model comprising the same data but using a different algorithm would be interesting and necessary for reproducing the results. Of interest would be similar joint Dimensionality Reduction (jDR) approaches such as

iCluster which like MOFA is an extension of factor analysis and has been successfully applied to subtyping breast and lung cancer (Shen et al., 2009). Next to jDR techniques, it would also be interesting to employ Deep Learning to a set of multi omics data to uncover the patterns between tumors of the same kind and within tumors.

MOFA has been beneficial in determining the heterogeneity within the PDAC data set, but the clustering of patients was not satisfactory due to the limitations in visualizing distinct samples in latent space. It has been reported before that MOFA is not particularly well suited for clustering as it has not been intrinsically designed for this purpose (Cantini et al., 2021). Despite this fact, the retrieved latent factors do lend themselves for sample clustering but probably needs a different kind of algorithm or more investigation into the nature of the retrieved clusters.

MOFA has been acknowledged as one of the best algorithms to associate factor-level information with clinical outcomes (Cantini et al., 2021). This has been confirmed in the survival analysis where the latent factors were associated with increased or decreased hazard of death. Furthermore, levels of latent factors were attributed to increased or decreased survival indicating that the factors are capable of conferring clinical outcome. Future extensions to the model should involve the possibility to exploit the built MOFA model in a predictive setting where having a read-out from multi-omics assays of future patients are used to make prognostic decisions or treatment choices.

Further directions also include the incorporation of other data modalities such as drug response data to relate drug treatments with retrieved factors. This would facilitate the way towards precision medicine were based on the molecular nature of the patient treatment options can be individualized and adverse effects minimized. However, this would entail data from the sample

patient cohort which at this point is not available. Increasingly TCGA data is supplemented with mass spectrometry data for the proteomics data in addition to RPPA assays. This type of data can be incorporated instead of the RPPA data and thus add to the information on the proteomic layer. To be even more comprehensive in the analysis, the metabolome layer can also be incorporated but this information is not yet available for the TCGA cohort. And there are further sources of data that might be of interest such as microbiome data. Intuitively, it seems that incorporating ever more data must result in a more comprehensive and precise model.

Eric Topol has proposed that every human individual is unique in its biological content and that the wealth of information that is present within our cells and is constantly produced by our digital devices are key in understanding and preserving our health (Topol, 2014). However, how much data is too much and at which point do we introduce biases and redundancies. These questions are future questions that need to be addressed to achieve the vision of personalized medicine.

CHAPTER 6 LITERATURE CITED

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

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

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

Arivaradarajan, P., & Misra, G. (Eds.). (2018). *Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data*. Springer Singapore. https://doi.org/10.1007/978-981-13-2925-8

Arora, S., Swaminathan, S. K., Kirtane, A., Srivastava, S. K., Bhardwaj, A., Singh, S., Panyam, J., & Singh, A. P. (2014). Synthesis, characterization, and evaluation of poly (D,L-lactide-co-glycolide)-based nanoformulation of miRNA-150: Potential implications for pancreatic cancer therapy. *International Journal of Nanomedicine*, *9*, 2933–2942. https://doi.org/10.2147/IJN.S61949

Arumugam, T., Brandt, W., Ramachandran, V., Moore, T. T., Wang, H., May, F. E., Westley, B. R., Hwang, R. F., & Logsdon, C. D. (2011). Trefoil Factor 1 Stimulates Both Pancreatic Cancer and Stellate Cells and Increases Metastasis. *Pancreas*, *40*(6), 815–822. https://doi.org/10.1097/MPA.0b013e31821f6927

Arumugam, T., Simeone, D. M., Golen, K. V., & Logsdon, C. D. (2005). S100P Promotes Pancreatic Cancer Growth, Survival, and Invasion. *Clinical Cancer Research*, *11*(15), 5356–5364. https://doi.org/10.1158/1078-0432.CCR-05-0092

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and Their Applications. *Journal of Chromatographic Science*, *55*(2), 182–196. https://doi.org/10.1093/chromsci/bmw167

Ausborn, N. L., Wang, T., Wentz, S. C., Washington, M. K., Merchant, N. B., Zhao, Z., Shyr, Y., Chakravarthy, A. B., & Xia, F. (2013). 53BP1 expression is a modifier of the prognostic value of lymph node ratio and CA 19–9 in pancreatic adenocarcinoma. *BMC Cancer*, *13*, 155. https://doi.org/10.1186/1471-2407-13-155

Awano, M., Fujiyuki, T., Shoji, K., Amagai, Y., Murakami, Y., Furukawa, Y., Sato, H., Yoneda, M., & Kai, C. (2016). Measles virus selectively blind to signaling lymphocyte activity molecule has oncolytic efficacy against nectin-4-expressing pancreatic cancer cells. *Cancer Science*, *107*(11), 1647–1652. https://doi.org/10.1111/cas.13064

Bady, P., Dolédec, S., Dumont, B., & Fruget, J.-F. (2004). Multiple co-inertia analysis: A tool for assessing synchrony in the temporal variability of aquatic communities. *Comptes Rendus Biologies*, *327*(1), 29–36. https://doi.org/10.1016/j.crvi.2003.10.007

Beke, L., Nuytten, M., Van Eynde, A., Beullens, M., & Bollen, M. (2007). The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the Polycomb group protein EZH2. *Oncogene*, *26*(31), 4590–4595. https://doi.org/10.1038/sj.onc.1210248

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://www.jstor.org/stable/2346101

Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, *23*(7), 781–783. https://doi.org/10.1101/gad.1787609

Bhaskaran, M., & Mohan, M. (2014). MicroRNAs: History, Biogenesis, and Their Evolving Role in Animal Development and Disease. *Veterinary Pathology*, *51*(4), 759–774. https://doi.org/10.1177/0300985813502820

Bhutia, Y. D., Babu, E., Ramachandran, S., Yang, S., Thangaraju, M., & Ganapathy, V. (2016). SLC transporters as a novel class of tumour suppressors: Identity, function and molecular mechanisms. *The Biochemical Journal*, *473*(9), 1113–1124. https://doi.org/10.1042/BJ20150751

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

Blumenthal, R. D., Leon, E., Hansen, H. J., & Goldenberg, D. M. (2007). Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers. *BMC Cancer*, *7*, 2. https://doi.org/10.1186/1471-2407-7-2

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

Brown, S. M. (2015). *Next-Generation DNA Sequencing Informatics* (Second Edition). Cold Spring Harbor Laboratory Press.

Burrell, R. A., McClelland, S. E., Endesfelder, D., Groth, P., Weller, M.-C., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S. M., Gronroos, E., Chew, S. K., Rowan, A. J., Schenk, A., Sheffer, M., Howell, M., Kschischo, M., Behrens, A., Helleday, T., Bartek, J., … Swanton, C. (2013). Replication stress links structural and numerical cancer chromosomal instability. *Nature*, *494*(7438), 492–496. https://doi.org/10.1038/nature11935

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C.-L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Cancer Staging—National Cancer Institute (nciglobal,ncienterprise). (2015, March 9). [CgvArticle]. https://www.cancer.gov/about-cancer/diagnosis-staging/staging

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, *12*(1), 124. https://doi.org/10.1038/s41467-020-20430-7

Cervantes-Villagrana, R. D., Albores-García, D., Cervantes-Villagrana, A. R., & García-Acevez, S. J. (2020). Tumor-induced neurogenesis and immune evasion as targets of innovative anti-cancer therapies. *Signal Transduction and Targeted Therapy*, *5*(1), 99. https://doi.org/10.1038/s41392-020-0205-z

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

Challita-Eid, P. M., Satpayev, D., Yang, P., An, Z., Morrison, K., Shostak, Y., Raitano, A., Nadell, R., Liu, W., Lortie, D. R., Capo, L., Verlinsky, A., Leavitt, M., Malik, F., Aviña, H., Guevara, C. I., Dinh, N., Karki, S., Anand, B. S., … Stover, D. R. (2016). Enfortumab Vedotin Antibody–Drug Conjugate Targeting Nectin-4 Is a Highly Potent Therapeutic Agent in Multiple Preclinical Cancer Models. *Cancer Research*, *76*(10), 3003–3013.

Cherief-Abdellatif, B.-E. (2019). Consistency of ELBO maximization for model selection. *Symposium on Advances in Approximate Bayesian Inference*, 11–31. http://proceedings.mlr.press/v96/cherief-abdellatif19a.html

Cochrane, D. R., Campbell, K. R., Greening, K., Ho, G. C., Hopkins, J., Bui, M., Douglas, J. M., Sharlandjieva, V., Munzur, A. D., Lai, D., DeGrood, M., Gibbard, E. W., Leung, S., Boyd, N., Cheng, A. S., Chow, C., Lim, J. L., Farnell, D. A., Kommoss, S., … Huntsman, D. G. (2020). Single cell transcriptomes of normal endometrial derived organoids uncover novel cell type markers and cryptic differentiation of primary tumours. *The Journal of Pathology*, *252*(2), 201–214. https://doi.org/10.1002/path.5511

Cohen, R., Neuzillet, C., Tijeras-Raballand, A., Faivre, S., de Gramont, A., & Raymond, E. (2015). Targeting cancer cell metabolism in pancreatic adenocarcinoma. *Oncotarget*, *6*(19), 16832–16847. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4627277/

Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G. E., Jakkula, L., Feiler, H. S., Ko, A. H., Olshen, A. B., Danenberg, K. L., Tempero, M. A., Spellman, P. T., Hanahan, D., & Gray, J. W. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Medicine*, *17*(4), 500–503. https://doi.org/10.1038/nm.2344

Dai, Z.-J., Gao, J., Kang, H.-F., Ma, Y.-G., Ma, X.-B., Lu, W.-F., Lin, S., Ma, H.-B., Wang, X.-J., & Wu, W.-Y. (2013). Targeted inhibition of mammalian target of rapamycin (mTOR) enhances radiosensitivity in pancreatic carcinoma cells. *Drug Design, Development and Therapy*, *7*, 149–159. https://doi.org/10.2147/DDDT.S42390

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

De Vito, R., Bellio, R., Trippa, L., & Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, *75*(1), 337–346. https://doi.org/10.1111/biom.12974

D'Errico, G., Alonso-Nocelo, M., Vallespinos, M., Hermann, P. C., Alcalá, S., García, C. P., Martin-Hijano, L., Valle, S., Earl, J., Cassiano, C., Lombardia, L., Feliu, J., Monti, M. C., Seufferlein, T., García-Bermejo, L., Martinelli, P., Carrato, A., & Sainz, B. (2019). Tumor-associated macrophage-secreted 14-3-3ζ signals via AXL to promote pancreatic cancer chemoresistance. *Oncogene*, *38*(27), 5469–5485. https://doi.org/10.1038/s41388-019-0803-9

Dey, P., Baddour, J., Muller, F., Wu, C. C., Wang, H., Liao, W.-T., Lan, Z., Chen, A., Gutschner, T., Kang, Y., Fleming, J., Satani, N., Zhao, D., Achreja, A., Yang, L., Lee, J., Chang, E., Genovese, G., Viale, A., … DePinho, R. A. (2017). Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature*, *542*(7639), 119–123. https://doi.org/10.1038/nature21052

Dolgalev, I. (2021, May 5). *MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format [R package msigdbr version 7.4.1]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=msigdbr

Domon, B., & Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. *Science*, *312*(5771), 212–217. https://doi.org/10.1126/science.1124619

Dou, D., Yang, S., Lin, Y., & Zhang, J. (2018). An eight-miRNA signature expression-based risk scoring system for prediction of survival in pancreatic adenocarcinoma. *Cancer Biomarkers: Section A of Disease Markers*, *23*(1), 79–93. https://doi.org/10.3233/CBM-181420

Drosten, M., & Barbacid, M. (2020). Targeting the MAPK Pathway in KRAS-Driven Tumors. *Cancer Cell*, *37*(4), 543–550. https://doi.org/10.1016/j.ccell.2020.03.013

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1), 587. https://doi.org/10.1186/1471-2105-11-587

Dumartin, L., Alrawashdeh, W., Trabulo, S. M., Radon, T. P., Steiger, K., Feakins, R. M., di Magliano, M. P., Heeschen, C., Esposito, I., Lemoine, N. R., & Crnogorac-Jurcevic, T. (2017). ER stress protein AGR2 precedes and is involved in the regulation of pancreatic cancer initiation. *Oncogene*, *36*(22), 3094–3103. https://doi.org/10.1038/onc.2016.459

Eleftheriou, N. M., Sjölund, J., Bocci, M., Cortez, E., Lee, S.-J., Cunha, S. I., & Pietras, K. (2016). Compound genetically engineered mouse models of cancer reveal dual targeting of ALK1 and endoglin as a synergistic opportunity to impinge on angiogenic TGF-β signaling. *Oncotarget*, *7*(51), 84314–84325. https://doi.org/10.18632/oncotarget.12604

Eser, S., Reiff, N., Messer, M., Seidler, B., Gottschalk, K., Dobler, M., Hieber, M., Arbeiter, A., Klein, S., Kong, B., Michalski, C. W., Schlitter, A. M., Esposito, I., Kind, A. J., Rad, L., Schnieke, A. E., Baccarini, M., Alessi, D. R., Rad, R., … Saur, D. (2013). Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer. *Cancer Cell*, *23*(3), 406–420. https://doi.org/10.1016/j.ccr.2013.01.023

Estécio, M. R. H., Yan, P. S., Ibrahim, A. E. K., Tellez, C. S., Shen, L., Huang, T. H.-M., & Issa, J.-P. J. (2007). High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome Research*, *17*(10), 1529–1536. https://doi.org/10.1101/gr.6417007

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*. https://doi.org/10.1002/ijc.33588

Fiory, F., Parrillo, L., Raciti, G. A., Zatterale, F., Nigro, C., Mirra, P., Falco, R., Ulianich, L., Di Jeso, B., Formisano, P., Miele, C., & Beguinot, F. (2014). PED/PEA-15 inhibits hydrogen peroxide-induced apoptosis in Ins-1E pancreatic beta-cells via PLD-1. *PloS One*, *9*(12), e113655. https://doi.org/10.1371/journal.pone.0113655

GAO, Y., ZHU, Y.-Y., & YUAN, Z. (2015). Colloid (mucinous non-cystic) carcinoma of the pancreas: A case report. *Oncology Letters*, *10*(5), 3195–3198. https://doi.org/10.3892/ol.2015.3733

González, B., Fece de la Cruz, F., Samuelsson, J. K., Alibés, A., & Alonso, S. (2018). Epigenetic and transcriptional dysregulation of VWA2 associated with a MYC -driven oncogenic program in colorectal cancer. *Scientific Reports*, *8*(1), 11097. https://doi.org/10.1038/s41598-018-29378-7

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gruber, R., Panayiotou, R., Nye, E., Spencer-Dene, B., Stamp, G., & Behrens, A. (2016). YAP1 and TAZ Control Pancreatic Cancer Initiation in Mice by Direct Up-regulation of JAK–STAT3 Signaling. *Gastroenterology*, *151*(3), 526–539. https://doi.org/10.1053/j.gastro.2016.05.006

Gu, J., Zhang, J., Huang, W., Tao, T., Huang, Y., Yang, L., Yang, J., Fan, Y., & Wang, H. (2020). Activating miRNA-mRNA network in gemcitabine-resistant pancreatic cancer cell associates with alteration of memory CD4+ T cells. *Annals of Translational Medicine*, *8*(6), 279. https://doi.org/10.21037/atm.2020.03.53

Gu, L., Zhang, J., Shi, M., & Peng, C. (2017). The effects of miRNA-1180 on suppression of pancreatic cancer. *American Journal of Translational Research*, *9*(6), 2798–2806. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5489882/

Guo, S., Fesler, A., Wang, H., & Ju, J. (2018). MicroRNA based prognostic biomarkers in pancreatic Cancer. *Biomarker Research*, *6*(1), 18. https://doi.org/10.1186/s40364-018-0131-1

Hager, G. L., McNally, J. G., & Misteli, T. (2009). Transcription Dynamics. *Molecular Cell*, *35*(6), 741–753. https://doi.org/10.1016/j.molcel.2009.09.005

Hall, W. A., & Goodman, K. A. (2019). Radiation therapy for pancreatic adenocarcinoma, a treatment option that must be considered in the management of a devastating malignancy. *Radiation Oncology*, *14*(1), 114. https://doi.org/10.1186/s13014-019-1277-1

Han, Y., & He, X. (2016). Integrating Epigenomics into the Understanding of Biomedical Insight. *Bioinformatics and Biology Insights*, *10*, 267–289. https://doi.org/10.4137/BBI.S38427

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

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

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., … Satija, R. (2020). Integrated analysis of multimodal single-cell data. *BioRxiv*, 2020.10.12.335331. https://doi.org/10.1101/2020.10.12.335331

Harper, J. W., & Bennett, E. J. (2016). Proteome complexity and the forces that drive proteome imbalance. *Nature*, *537*(7620), 328–338. https://doi.org/10.1038/nature19947

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

He, L., & Hannon, G. J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, *5*(7), 522–531. https://doi.org/10.1038/nrg1379

Heeg, S., Das, K. K., Reichert, M., Bakir, B., Takano, S., Caspers, J., Aiello, N. M., Wu, K., Neesse, A., Maitra, A., Iacobuzio-Donahue, C. A., Hicks, P., & Rustgi, A. K. (2016). The ETS-Transcription Factor ETV1 Regulates Stromal Expansion and Metastasis in Pancreatic Cancer. *Gastroenterology*, *151*(3), 540-553.e14. https://doi.org/10.1053/j.gastro.2016.06.005

Heestand, G. M., & Kurzrock, R. (2015). Molecular landscape of pancreatic cancer: Implications for current clinical trials. *Oncotarget*, *6*(7), 4553–4561. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4467098/

Hermanek, P., & Wittekind, C. (1994). Residual tumor (R) classification and prognosis. *Seminars in Surgical Oncology*, *10*(1), 12–20. https://doi.org/10.1002/ssu.2980100105

Hiroshima, Y., Kasajima, R., Kimura, Y., Komura, D., Ishikawa, S., Ichikawa, Y., Bouvet, M., Yamamoto, N., Oshima, T., Morinaga, S., Singh, S. R., Hoffman, R. M., Endo, I., & Miyagi, Y. (2020). Novel targets identified by integrated cancer-stromal interactome analysis of pancreatic adenocarcinoma. *Cancer Letters*, *469*, 217–227. https://doi.org/10.1016/j.canlet.2019.10.031

Ho, A. S., Huang, X., Cao, H., Christman-Skieller, C., Bennewith, K., Le, Q.-T., & Koong, A. C. (2010). Circulating miR-210 as a Novel Hypoxia Marker in Pancreatic Cancer. *Translational Oncology*, *3*(2), 109–113. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847318/

Hothorn, T., & Lausen, B. (2002). Maximally Selected Rank Statistics in R. *R News*, *2*, 3–5.

Hruban, R. H., Petersen, G. M., Goggins, M., Tersmette, A. C., Offerhaus, G. J., Falatko, F., Yeo, C. J., & Kern, S. E. (1999). Familial pancreatic cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *10 Suppl 4*, 69–73.

Huang, G.-L., Sun, J., Lu, Y., Liu, Y., Cao, H., Zhang, H., & Calin, G. A. (2019). MiR-200 family and cancer: From a meta-analysis view. *Molecular Aspects of Medicine*, *70*, 57–71. https://doi.org/10.1016/j.mam.2019.09.005

Huang, H., Svoboda, R. A., Lazenby, A. J., Saowapa, J., Chaika, N., Ding, K., Wheelock, M. J., & Johnson, K. R. (2016). Up-regulation of N-cadherin by Collagen I-activated Discoidin Domain Receptor 1 in Pancreatic Cancer Requires the Adaptor Molecule Shc1. *The Journal of Biological Chemistry*, *291*(44), 23208–23223. https://doi.org/10.1074/jbc.M116.740605

Huang, Y., Zhou, L., Meng, X., Yu, B., Wang, H., Yang, Y., Wu, Y., & Tan, X. (2018). IRS-1 regulates proliferation, invasion and metastasis of pancreatic cancer cells through MAPK and PI3K signaling pathways. *International Journal of Clinical and Experimental Pathology*, *11*(11), 5185–5193.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ikeda, Y., Tanji, E., Makino, N., Kawata, S., & Furukawa, T. (2012). MicroRNAs associated with mitogen-activated protein kinase in human pancreatic cancer. *Molecular Cancer Research: MCR*, *10*(2), 259–269. https://doi.org/10.1158/1541-7786.MCR-11-0035

Iorio, M. V., & Croce, C. M. (2012). MicroRNA dysregulation in cancer: Diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine*, *4*(3), 143–159. https://doi.org/10.1002/emmm.201100209

Jia, K., Zhao, X., & Dang, X. (2020). Mass spectrometry-based iTRAQ analysis of serum markers in patients with pancreatic cancer. *Oncology Letters*, *19*(6), 4106–4114. https://doi.org/10.3892/ol.2020.11491

Jia, Y., Shen, M., Zhou, Y., & Liu, H. (2020). Development of a 12-biomarkers-based prognostic model for pancreatic cancer using multi-omics integrated analysis. *Acta Biochimica Polonica*, *67*(4), 501–508. https://doi.org/10.18388/abp.2020_5225

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews. Molecular Cell Biology*, *17*(7), 451–459. https://doi.org/10.1038/nrm.2016.25

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., … Kinzler, K. W. (2008). Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science*, *321*(5897), 1801–1806. https://doi.org/10.1126/science.1164368

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

Kassambara, A., Kosinski, M., Biecek, P., & others. (2017). survminer: Drawing Survival Curves using'ggplot2'. *R Package Version 0.3*, *1*.

Katoh, M. (2018). Multi-layered prevention and treatment of chronic inflammation, organ fibrosis and cancer associated with canonical WNT/β-catenin signaling activation (Review). *International Journal of Molecular Medicine*, *42*(2), 713–725. https://doi.org/10.3892/ijmm.2018.3689

Khalaileh, A., Dreazen, A., Khatib, A., Apel, R., Swisa, A., Kidess-Bassir, N., Maitra, A., Meyuhas, O., Dor, Y., & Zamir, G. (2013). Phosphorylation of Ribosomal Protein S6 Attenuates DNA Damage and Tumor Suppression during Development of Pancreatic Cancer. *Cancer Research*, *73*(6), 1811–1820.

Kibirige, H., Lamp, G., Katins, J., gdowding, austin, matthias-k, Funnell, T., Finkernagel, F., Arnfred, J., Blanchard, D., Astanin, S., Chiang, E., Kishimoto, P. N., Sheehan, E., stonebig, Willers, B., Gibboni, R., smutch, Halchenko, Y., … Saiz, D. (2021). *has2k1/plotnine: V0.8.0*. Zenodo. https://doi.org/10.5281/zenodo.4636791

Kinugawa, Y., Uehara, T., Matsuda, K., Kobayashi, Y., Nakajima, T., Hamano, H., Kawa, S., Higuchi, K., Hosaka, N., Shiozawa, S., Ishigame, H., Nakamura, T., Maruyama, Y., Nakazawa, K., Nakaguro, M., Sano, K., & Ota, H. (2018). Promoter hypomethylation of SKI in autoimmune pancreatitis. *Pathology, Research and Practice*, *214*(4), 492–497. https://doi.org/10.1016/j.prp.2018.03.005

Kobayashi, T., & Itoh, H. (2017). Loss of a primary cilium in PDAC. *Cell Cycle*, *16*(9), 817–818. https://doi.org/10.1080/15384101.2017.1304738

Kong, B., Wu, W., Cheng, T., Schlitter, A. M., Qian, C., Bruns, P., Jian, Z., Jäger, C., Regel, I., Raulefs, S., Behler, N., Irmler, M., Beckers, J., Friess, H., Erkan, M., Siveke, J. T., Tannapfel, A., Hahn, S. A., Theis, F. J., … Michalski, C. W. (2016). A subset of metastatic pancreatic ductal adenocarcinomas depends quantitatively on oncogenic Kras/Mek/Erk-induced hyperactive mTOR signalling. *Gut*, *65*(4), 647–657. https://doi.org/10.1136/gutjnl-2014-307616

Kong, L., Liu, P., Zheng, M., Wang, Z., Gao, Y., Liang, K., Wang, H., & Tan, X. (2020). The miR-1224-5p/ELF3 Axis Regulates Malignant Behaviors of Pancreatic Cancer via PI3K/AKT/Notch Signaling Pathways</p>. *OncoTargets and Therapy*, *13*, 3449–3466. https://doi.org/10.2147/OTT.S248507

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

Kontomanolis, E. N., Koutras, A., Syllaios, A., Schizas, D., Mastoraki, A., Garmpis, N., Diakosavvas, M., Angelou, K., Tsatsaris, G., Pagkalos, A., Ntounis, T., & Fasoulakis, Z. (2020). Role of Oncogenes and Tumor-suppressor Genes in Carcinogenesis: A Review. *Anticancer Research*, *40*(11), 6009–6015. https://doi.org/10.21873/anticanres.14622

Korbecki, J., Kojder, K., Simińska, D., Bohatyrewicz, R., Gutowska, I., Chlubek, D., & Baranowska-Bosiacka, I. (2020). CC Chemokines in a Tumor: A Review of Pro-Cancer and Anti-Cancer Properties of the Ligands of Receptors CCR1, CCR2, CCR3, and CCR4. *International Journal of Molecular Sciences*, *21*(21). https://doi.org/10.3390/ijms21218412

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, *11*(3), 191–203. https://doi.org/10.1038/nrg2732

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

Lee, E. J., Gusev, Y., Jiang, J., Nuovo, G. J., Lerner, M. R., Frankel, W. L., Morgan, D. L., Postier, R. G., Brackett, D. J., & Schmittgen, T. D. (2007). Expression profiling identifies microRNA signature in pancreatic cancer. *International Journal of Cancer. Journal International Du Cancer*, *120*(5), 1046–1054. https://doi.org/10.1002/ijc.22394

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–719. https://doi.org/10.1016/j.cell.2007.01.015

Li, C., Morvaridi, S., Lam, G., Chheda, C., Kamata, Y., Katsumata, M., Edderkaoui, M., Yuan, X., Nissen, N., Pandol, S. J., & Wang, Q. (2019). MSP-RON Signaling Is Activated in the Transition From Pancreatic Intraepithelial Neoplasia (PanIN) to Pancreatic Ductal Adenocarcinoma (PDAC). *Frontiers in Physiology*, *10*. https://doi.org/10.3389/fphys.2019.00147

Li, Q. W., Zhou, T., Wang, F., Jiang, M., Liu, C. B., Zhang, K., Zhou, Q., Tian, Z., & Hu, K. W. (2015). MicroRNA-215 functions as a tumor suppressor and directly targets ZEB2 in human pancreatic cancer. *Genetics and Molecular Research*, *14*(4), 16133–16145. https://doi.org/10.4238/2015.December.8.2

Liang, L., Wei, D.-M., Li, J.-J., Luo, D.-Z., Chen, G., Dang, Y.-W., & Cai, X.-Y. (2018). Prognostic microRNAs and their potential molecular mechanism in pancreatic cancer: A study based on The Cancer Genome Atlas and bioinformatics investigation. *Molecular Medicine Reports*, *17*(1), 939–951. https://doi.org/10.3892/mmr.2017.7945

Liao, X., Huang, R., Liu, X., Han, C., Yu, L., Wang, S., Sun, N., Li, B., Ning, X., & Peng, T. (2017). Distinct prognostic values of alcohol dehydrogenase mRNA expression in pancreatic adenocarcinoma. *OncoTargets and Therapy*, *10*, 3719–3732. https://doi.org/10.2147/OTT.S140221

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Ligorio, M., Sil, S., Malagon-Lopez, J., Nieman, L. T., Misale, S., Pilato, M. D., Ebright, R. Y., Karabacak, M. N., Kulkarni, A. S., Liu, A., Jordan, N. V., Franses, J. W., Philipp, J., Kreuzer, J., Desai, N., Arora, K. S., Rajurkar, M., Horwitz, E., Neyaz, A., … Ting, D. T. (2019). Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer. *Cell*, *178*(1), 160-175.e27. https://doi.org/10.1016/j.cell.2019.05.012

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322. https://doi.org/10.1038/nature08514

Liu, J., Xu, W., Li, S., Sun, R., & Cheng, W. (2020). Multi-omics analysis of tumor mutational burden combined with prognostic assessment in epithelial ovarian cancer based on TCGA database. *International Journal of Medical Sciences*, *17*(18), 3200–3213. https://doi.org/10.7150/ijms.50491

Liu, X., Qian, D., Liu, H., Abbruzzese, J. L., Luo, S., Walsh, K. M., & Wei, Q. (2020). Genetic variants of the peroxisome proliferator-activated receptor (PPAR) signaling pathway genes and risk of pancreatic cancer. *Molecular Carcinogenesis*, *59*(8), 930–939. https://doi.org/10.1002/mc.23208

Liu, X., Yuan, H., Zhou, J., Wang, Q., Qi, X., Bernal, C., Avella, D., Kaifi, J. T., Kimchi, E. T., Timothy, P., Cheng, K., Miao, Y., Jiang, K., & Li, G. (2021). LMO7 as an Unrecognized Factor Promoting Pancreatic Cancer Progression and Metastasis. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/fcell.2021.647387

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. *The Annals of Applied Statistics*, *7*(1), 523–542. https://doi.org/10.1214/12-AOAS597

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5). https://doi.org/10.1371/journal.pcbi.1005457

Lyu, X., Wang, P., Qiao, Q., & Jiang, Y. (2021). Genomic stratification based on microenvironment immune types and PD-L1 for tailoring therapeutic strategies in bladder cancer. *BMC Cancer*, *21*(1), 646. https://doi.org/10.1186/s12885-021-08350-1

Ma, J., Cao, T., Cui, Y., Zhang, F., Shi, Y., Xia, J., & Wang, Z. P. (2019). MiR-223 Regulates Cell Proliferation and Invasion via Targeting PDS5B in Pancreatic Cancer Cells. *Molecular Therapy. Nucleic Acids*, *14*, 583–592. https://doi.org/10.1016/j.omtn.2019.01.009

Matsuo, H., Yoshida, K., Fukumura, K., Nakatani, K., Noguchi, Y., Takasaki, S., Noura, M., Shiozawa, Y., Shiraishi, Y., Chiba, K., Tanaka, H., Okada, A., Nannya, Y., Takeda, J., Ueno, H., Shiba, N., Yamato, G., Handa, H., Ono, Y., … Adachi, S. (2018). Recurrent CCND3 mutations in MLL-rearranged acute myeloid leukemia. *Blood Advances*, *2*(21), 2879–2889. https://doi.org/10.1182/bloodadvances.2018019398

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., & Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, *454*(7205), 766–770. https://doi.org/10.1038/nature07107

Meng, L.-D., Shi, G.-D., Ge, W.-L., Huang, X.-M., Chen, Q., Yuan, H., Wu, P.-F., Lu, Y.-C., Shen, P., Zhang, Y.-H., Cao, S.-J., Miao, Y., Tu, M., & Jiang, K.-R. (2020). Linc01232 promotes the metastasis of pancreatic cancer by suppressing the ubiquitin-mediated degradation of HNRNPA2B1 and activating the A-Raf-induced MAPK/ERK signaling pathway. *Cancer Letters*, *494*, 107–120. https://doi.org/10.1016/j.canlet.2020.08.001

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

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

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., Rashid, N. U., Williams, L. A., Eaton, S. C., Chung, A. H., Smyla, J. K., Anderson, J. M., Kim, H. J., Bentrem, D. J., Talamonti, M. S., Iacobuzio-Donahue, C. A., Hollingsworth, M. A., & Yeh, J. J. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature Genetics*, *47*(10), 1168–1178. https://doi.org/10.1038/ng.3398

Nandula, S. R., Huxford, I., Wheeler, T. T., Aparicio, C., & Gorr, S.-U. (2020). The parotid secretory protein BPIFA2 is a salivary surfactant that affects lipopolysaccharide action. *Experimental Physiology*, *105*(8), 1280–1292. https://doi.org/10.1113/EP088567

Nelson, N. J. (2001). Microarrays have arrived: Gene expression tool matures. *Journal of the National Cancer Institute*, *93*(7), 492–494. https://doi.org/10.1093/jnci/93.7.492

Nfonsam, L. E., Jandova, J., Jecius, H. C., Omesiete, P. N., & Nfonsam, V. N. (2019). SFRP4 expression correlates with epithelial mesenchymal transition-linked genes and poor overall survival in colon cancer patients. *World Journal of Gastrointestinal Oncology*, *11*(8), 589–598. https://doi.org/10.4251/wjgo.v11.i8.589

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

Oji, Y., Tatsumi, N., Fukuda, M., Nakatsuka, S.-I., Aoyagi, S., Hirata, E., Nanchi, I., Fujiki, F., Nakajima, H., Yamamoto, Y., Shibata, S., Nakamura, M., Hasegawa, K., Takagi, S., Fukuda, I., Hoshikawa, T., Murakami, Y., Mori, M., Inoue, M., … Sugiyama, H. (2014). The translation elongation factor eEF2 is a novel tumor-associated antigen overexpressed in various types of cancers. *International Journal of Oncology*, *44*(5), 1461–1469. https://doi.org/10.3892/ijo.2014.2318

Owusu-Ansah, K. G., Song, G., Chen, R., Edoo, M. I. A., Li, J., Chen, B., Wu, J., Zhou, L., Xie, H., Jiang, D., & Zheng, S. (2019). COL6A1 promotes metastasis and predicts poor prognosis in patients with pancreatic cancer. *International Journal of Oncology*, *55*(2), 391–404. https://doi.org/10.3892/ijo.2019.4825

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Palsson, B., & Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, *6*(11), 787–789. https://doi.org/10.1038/nchembio.462

Patel, S. H., Edwards, M. J., & Ahmad, S. A. (2019). Intracellular Ion Channels in Pancreas Cancer | Cell Physiol Biochem. *Cellular Physiology & Biochemistry*, *53*(S1), 44–51. https://www.cellphysiolbiochem.com/Articles/000193/

Patel, S. K., George, B., & Rai, V. (2020). Artificial Intelligence to Decode Cancer Mechanism: Beyond Patient Stratification for Precision Oncology. *Frontiers in Pharmacology*, *11*. https://doi.org/10.3389/fphar.2020.01177

Pedersen, M. H., Hood, B. L., Ehmsen, S., Beck, H. C., Conrads, T. P., Bak, M., Ditzel, H. J., & Leth-Larsen, R. (2019). CYPOR is a novel and independent prognostic biomarker of recurrence-free survival in triple-negative breast cancer patients. *International Journal of Cancer*, *144*(3), 631–640. https://doi.org/10.1002/ijc.31798

Ponath, V., Frech, M., Bittermann, M., Al Khayer, R., Neubauer, A., Brendel, C., & Pogge von Strandmann, E. (2020). The Oncoprotein SKI Acts as A Suppressor of NK Cell-Mediated Immunosurveillance in PDAC. *Cancers*, *12*(10), 2857. https://doi.org/10.3390/cancers12102857

Purcell, J. W., Tanlimco, S. G., Hickson, J., Fox, M., Sho, M., Durkin, L., Uziel, T., Powers, R., Foster, K., McGonigal, T., Kumar, S., Samayoa, J., Zhang, D., Palma, J. P., Mishra, S., Hollenbaugh, D., Gish, K., Morgan-Lappe, S. E., Hsi, E. D., & Chao, D. T. (2018). LRRC15 Is a Novel Mesenchymal Protein and Stromal Target for Antibody-Drug Conjugates. *Cancer Research*, *78*(14), 4059–4072. https://doi.org/10.1158/0008-5472.CAN-18-0327

Qi, B., Liu, H., Dong, Y., Shi, X., Zhou, Q., Zeng, F., Bao, N., Li, Q., Yuan, Y., Yao, L., & Xia, S. (2020). The nine ADAMs family members serve as potential biomarkers for immune infiltration in pancreatic adenocarcinoma. *PeerJ*, *8*. https://doi.org/10.7717/peerj.9736

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014). Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. *Cancer Research*, *74*(11), 2913–2921. https://doi.org/10.1158/0008-5472.CAN-14-0155

Raja, A., Malik, M. F. A., & Haq, F. (2021). Genomic relevance of FGF14 and associated genes on the prognosis of pancreatic cancer. *PLOS ONE*, *16*(6), e0252344. https://doi.org/10.1371/journal.pone.0252344

Ramaswami, R., Bayer, R., & Galea, S. (2018). Precision Medicine from a Public Health Perspective. *Annual Review of Public Health*, *39*(1), 153–168. https://doi.org/10.1146/annurev-publhealth-040617-014158

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

Ren, Z.-G., Dong, S.-X., Han, P., & Qi, J. (2016). MiR-203 promotes proliferation, migration and invasion by degrading SIK1 in pancreatic cancer. *Oncology Reports*, *35*(3), 1365–1374. https://doi.org/10.3892/or.2015.4534

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

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

Ryschich, E., Huszty, G., Knaebel, H. P., Hartel, M., Büchler, M. W., & Schmidt, J. (2004). Transferrin receptor is a marker of malignant phenotype in human pancreatic cancer and in neuroendocrine carcinoma of the pancreas. *European Journal of Cancer (Oxford, England: 1990)*, *40*(9), 1418–1422. https://doi.org/10.1016/j.ejca.2004.01.036

Sato, M., Matsumoto, M., Saiki, Y., Alam, M., Nishizawa, H., Rokugo, M., Brydun, A., Yamada, S., Kaneko, M. K., Funayama, R., Ito, M., Kato, Y., Nakayama, K., Unno, M., & Igarashi, K. (2020). BACH1 Promotes Pancreatic Cancer Metastasis by Repressing Epithelial Genes and Enhancing Epithelial–Mesenchymal Transition. *Cancer Research*, *80*(6), 1279–1292.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912. https://doi.org/10.1093/bioinformatics/btp543

Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews. Genetics*, *21*(4), 243–254. https://doi.org/10.1038/s41576-020-0210-7

Shi, X., Cheng, L., Jiao, X., Chen, B., Li, Z., Liang, Y., Liu, W., Wang, J., Liu, G., Xu, Y., Sun, J., Fu, Q., Lu, Y., & Chen, S. (2018). Rare Copy Number Variants Identify Novel Genes in Sporadic Total Anomalous Pulmonary Vein Connection. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00559

Siret, C., Dobric, A., Martirosyan, A., Terciolo, C., Germain, S., Bonier, R., Dirami, T., Dusetti, N., Tomasini, R., Rubis, M., Garcia, S., Iovanna, J., Lombardo, D., Rigot, V., & André, F. (2018). Cadherin-1 and cadherin-3 cooperation determines the aggressiveness of pancreatic ductal adenocarcinoma. *British Journal of Cancer*, *118*(4), 546–557. https://doi.org/10.1038/bjc.2017.411

Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384. https://doi.org/10.1101/pdb.prot5384

Storz, P., & Crawford, H. C. (2020). Carcinogenesis of Pancreatic Ductal Adenocarcinoma. *Gastroenterology*, *158*(8), 2072–2081. https://doi.org/10.1053/j.gastro.2020.02.059

Sun, D., Jin, H., Zhang, J., & Tan, X. (2018). Integrated whole genome microarray analysis and immunohistochemical assay identifies COL11A1, GJB2 and CTRL as predictive biomarkers for pancreatic cancer. *Cancer Cell International*, *18*, 174. https://doi.org/10.1186/s12935-018-0669-x

Sun, M., Lai, D., Zhang, L., & Huang, X. (2015). Modified SuperCurve Method for Analysis of Reverse-Phase Protein Array Data. *Journal of Computational Biology*, *22*(8), 765–769. https://doi.org/10.1089/cmb.2015.0007

Szulzewsky, F., Arora, S., Hoellerbauer, P., King, C., Nathan, E., Chan, M., Cimino, P. J., Ozawa, T., Kawauchi, D., Pajtler, K. W., Gilbertson, R. J., Paddison, P. J., Vasioukhin, V., Gujral, T. S., & Holland, E. C. (2020). Comparison of tumor-associated YAP1 fusions identifies a recurrent set of functions critical for oncogenesis. *Genes & Development*, *34*(15–16), 1051–1064. https://doi.org/10.1101/gad.338681.120

Takaku, H., Minagawa, A., Takagi, M., & Nashimoto, M. (2003). A candidate prostate cancer susceptibility gene encodes tRNA 3′ processing endoribonuclease. *Nucleic Acids Research*, *31*(9), 2272–2278. https://doi.org/10.1093/nar/gkg337

Team, Rs. (2021). *RStudio: Integrated Development Environment for R. RStudio*. http://www.rstudio.com/

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

Therneau, T. M. (2021, April 26). *Survival Analysis [R package survival version 3.2-11]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-3294-8

Topol, E. J. (2014). Individualized Medicine from Prewomb to Tomb. *Cell*, *157*(1), 241–253. https://doi.org/10.1016/j.cell.2014.02.012

Trauzold, A., Röder, C., Sipos, B., Karsten, K., Arlt, A., Jiang, P., Martin-Subero, J. I., Siegmund, D., Müerköster, S., Pagerols-Raluy, L., Siebert, R., Wajant, H., & Kalthoff, H. (2005). CD95 and TRAF2 promote invasiveness of pancreatic cancer cells. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *19*(6), 620–622. https://doi.org/10.1096/fj.04-2984fje

Tu, Q., Hao, J., Zhou, X., Yan, L., Dai, H., Sun, B., Yang, D., An, S., Lv, L., Jiao, B., Chen, C., Lai, R., Shi, P., & Zhao, X. (2018). CDKN2B deletion is essential for pancreatic cancer development instead of unmeaningful co-deletion due to juxtaposition to CDKN2A. *Oncogene*, *37*(1), 128–138. https://doi.org/10.1038/onc.2017.316

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

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

Veenstra, V. L., Damhofer, H., Waasdorp, C., van Rijssen, L. B., van de Vijver, M. J., Dijk, F., Wilmink, H. W., Besselink, M. G., Busch, O. R., Chang, D. K., Bailey, P. J., Biankin, A. V., Kocher, H. M., Medema, J. P., Li, J. S., Jiang, R., Pierce, D. W., van Laarhoven, H. W. M., & Bijlsma, M. F. (2018). ADAM12 is a circulating marker for stromal activation in pancreatic cancer and predicts response to chemotherapy. *Oncogenesis*, *7*(11), 87. https://doi.org/10.1038/s41389-018-0096-9

Wang, G. D., Huang, J. T., Liu, Y., Wu, Y., Chen, X. Q., Zhao, Y. F., & Wang, D. L. (2019). Deletion of C9orf53 promotes the growth of head and neck squamous cell carcinoma and is associated with poor prognosis of patients with head and neck squamous cell carcinoma. *Oncology Letters*, *17*(1), 1223–1228. https://doi.org/10.3892/ol.2018.9675

Wang, G., Li, Y., Zhang, D., Zhao, S., Zhang, Q., Luo, C., Sun, X., & Zhang, B. (2020). CELSR1 Acts as an Oncogene Regulated by miR-199a-5p in Glioma. *Cancer Management and Research*, *12*, 8857–8865. https://doi.org/10.2147/CMAR.S258835

Wang, H., Shen, L., Li, Y., & Lv, J. (2020). Integrated characterisation of cancer genes identifies key molecular biomarkers in stomach adenocarcinoma. *Journal of Clinical Pathology*, *73*(9), 579–586. https://doi.org/10.1136/jclinpath-2019-206400

Wang, H.-L., Zhou, R., Liu, J., Chang, Y., Liu, S., Wang, X.-B., Huang, M.-F., & Zhao, Q. (2017). MicroRNA-196b inhibits late apoptosis of pancreatic cancer cells by targeting CADM1. *Scientific Reports*, *7*(1), 11467. https://doi.org/10.1038/s41598-017-11248-3

Wang, J., Guo, J., & Fan, H. (2020). MiR-155 regulates the proliferation and apoptosis of pancreatic cancer cells through targeting SOCS3. *European Review for Medical and Pharmacological Sciences*, *24*(24), 12625. https://doi.org/10.26355/eurrev_202012_24143

Wang, K., Zhan, Y., Huynh, N., Dumesny, C., Wang, X., Asadi, K., Herrmann, D., Timpson, P., Yang, Y., Walsh, K., Baldwin, G. S., Nikfarjam, M., & He, H. (2020). Inhibition of PAK1 suppresses pancreatic cancer by stimulation of anti-tumour immunity through down-regulation of PD-L1. *Cancer Letters*, *472*, 8–18. https://doi.org/10.1016/j.canlet.2019.12.020

Wang, V. M.-Y., Ferreira, R. M. M., Almagro, J., Evan, T., Legrave, N., Zaw Thin, M., Frith, D., Carvalho, J., Barry, D. J., Snijders, A. P., Herbert, E., Nye, E. L., MacRae, J. I., & Behrens, A. (2019). CD9 identifies pancreatic cancer stem cells and modulates glutamine metabolism to fuel tumour growth. *Nature Cell Biology*, *21*(11), 1425–1435. https://doi.org/10.1038/s41556-019-0407-1

Wang, X.-J., Zhang, D.-L., Xu, Z.-G., Ma, M.-L., Wang, W.-B., Li, L.-L., Han, X.-L., Huo, Y., Yu, X., & Sun, J.-P. (2014). Understanding CELSRs—Cadherin EGF LAG seven-pass G-type receptors. *Journal of Neurochemistry*, *131*(6), 699–711. https://doi.org/10.1111/jnc.12955

Welsh, M. J., & Smith, A. E. (1993). Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell*, *73*(7), 1251–1254. https://doi.org/10.1016/0092-8674(93)90353-R

Wharry, C. E., Haines, K. M., Carroll, R. G., & May, M. J. (2009). Constitutive noncanonical NFκB signaling in pancreatic cancer cells. *Cancer Biology & Therapy*, *8*(16), 1567–1576. https://doi.org/10.4161/cbt.8.16.8961

Wiley, S. Z., Sriram, K., Salmerón, C., & Insel, P. A. (2019). GPR68: An Emerging Drug Target in Cancer. *International Journal of Molecular Sciences*, *20*(3), 559. https://doi.org/10.3390/ijms20030559

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J. C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., & Hochstrasser, D. F. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology (Nature Publishing Company)*, *14*(1), 61–65. https://doi.org/10.1038/nbt0196-61

Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., & Williams, K. L. (1996). Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnology and Genetic Engineering Reviews*, *13*(1), 19–50. https://doi.org/10.1080/02648725.1996.10647923

Wu, F., Li, G.-Z., Liu, H.-J., Zhao, Z., Chai, R.-C., Liu, Y.-Q., Jiang, H.-Y., Zhai, Y., Feng, Y.-M., Li, R.-P., & Zhang, W. (2020). Molecular subtyping reveals immune alterations in IDH wild-type lower-grade diffuse glioma. *The Journal of Pathology*, *251*(3), 272–283. https://doi.org/10.1002/path.5468

Wu, H.-Y., Yang, M.-C., Ding, L.-Y., Chen, C. S., & Chu, P.-C. (2019). P21-Activated kinase 3 promotes cancer stem cell phenotypes through activating the Akt-GSK3β-β-catenin signaling pathway in pancreatic cancer cells. *Cancer Letters*, *456*, 13–22. https://doi.org/10.1016/j.canlet.2019.04.026

Wu, M., Li, X., Zhang, T., Liu, Z., & Zhao, Y. (2019). Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. *Frontiers in Oncology*, *9*, 996. https://doi.org/10.3389/fonc.2019.00996

Xia, L., Wang, H., Xiao, H., Lan, B., Liu, J., & Yang, Z. (2020). EEF1A2 and ERN2 could potentially discriminate metastatic status of mediastinal lymph node in lung adenocarcinomas harboring EGFR 19Del/L858R mutations. *Thoracic Cancer*, *11*(10), 2755–2766. https://doi.org/10.1111/1759-7714.13554

Xie, D., Cui, J., Xia, T., Jia, Z., Wang, L., Wei, W., Zhu, A., Gao, Y., Xie, K., & Quan, M. (2015). Hippo transducer TAZ promotes epithelial mesenchymal transition and supports pancreatic cancer progression. *Oncotarget*, *6*(34), 35949–35963. https://doi.org/10.18632/oncotarget.5772

Xie, Z., Dang, Y., Wei, D., Chen, P., Tang, R., Huang, Q., Liu, J., & Luo, D. (2017). Clinical significance and prospective molecular mechanism of MALAT1 in pancreatic cancer exploration: A comprehensive study based on the GeneChip, GEO, Oncomine, and TCGA databases. *OncoTargets and Therapy*, *10*, 3991–4005. https://doi.org/10.2147/OTT.S136878

Xu, D., Du, M., Zhang, J., Xiong, P., Li, W., Zhang, H., Xiong, W., Liu, F., & Liu, J. (2018). DNMT1 mediated promoter methylation of GNAO1 in hepatoma carcinoma cells. *Gene*, *665*, 67–73. https://doi.org/10.1016/j.gene.2018.04.080

Xu, X., Gong, C., Wang, Y., Hu, Y., Liu, H., & Fang, Z. (2020). Multi-omics analysis to identify driving factors in colorectal cancer. *Epigenomics*, *12*(18), 1633–1650. https://doi.org/10.2217/epi-2020-0073

Xu, X., Lu, Y., Wu, Y., Wang, M., Wang, X., Wang, H., Chen, B., & Li, Y. (2021). A signature of seven immune-related genes predicts overall survival in male gastric cancer patients. *Cancer Cell International*, *21*(1), 117. https://doi.org/10.1186/s12935-021-01823-0

Yao, R., Xu, L., Wei, B., Qian, Z., Wang, J., Hui, H., & Sun, Y. (2019). MiR-142-5p regulates pancreatic cancer cell proliferation and apoptosis by regulation of RAP1A. *Pathology, Research and Practice*, *215*(6), 152416. https://doi.org/10.1016/j.prp.2019.04.008

Ye, H., Li, T., Wang, H., Wu, J., Yi, C., Shi, J., Wang, P., Song, C., Dai, L., Jiang, G., Huang, Y., Yu, Y., & Li, J. (2021). TSPAN1, TMPRSS4, SDR16C5, and CTSE as Novel Panel for Pancreatic Cancer: A Bioinformatics Analysis and Experiments Validation. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.649551

Ye, T., Su, J., Huang, C., Yu, D., Dai, S., Huang, X., Chen, B., & Zhou, M. (2016). Isoorientin induces apoptosis, decreases invasiveness, and downregulates VEGF secretion by activating AMPK signaling in pancreatic cancer cells. *OncoTargets and Therapy*, *9*, 7481–7492. https://doi.org/10.2147/OTT.S122653

Ye, Z., Wang, F., Yan, F., Wang, L., Li, B., Liu, T., Hu, F., Jiang, M., Li, W., & Fu, Z. (2019). Bioinformatic identification of candidate biomarkers and related transcription factors in nasopharyngeal carcinoma. *World Journal of Surgical Oncology*, *17*(1), 60. https://doi.org/10.1186/s12957-019-1605-9

Yeo, M. S., Subhash, V. V., Suda, K., Balcıoğlu, H. E., Zhou, S., Thuya, W. L., Loh, X. Y., Jammula, S., Peethala, P. C., Tan, S. H., Xie, C., Wong, F. Y., Ladoux, B., Ito, Y., Yang, H., Goh, B. C., Wang, L., & Yong, W. P. (2019). FBXW5 Promotes Tumorigenesis and Metastasis in Gastric Cancer via Activation of the FAK-Src Signaling Pathway. *Cancers*, *11*(6). https://doi.org/10.3390/cancers11060836

Yousef, M., & Allmer, J. (Eds.). (2014). *miRNomics: MicroRNA biology and computational analysis*. Humana Press.

Zhang, J., Zheng, B., Zhou, X., Zheng, T., Wang, H., Wang, Y., & Zhang, W. (2021). Increased BST-2 expression by HBV infection promotes HBV-associated HCC tumorigenesis. *Journal of Gastrointestinal Oncology*, *12*(2). https://doi.org/10.21037/jgo-20-356

Zhang, X., Kang, C., Li, N., Liu, X., Zhang, J., Gao, F., & Dai, L. (2019). Identification of special key genes for alcohol-related hepatocellular carcinoma through bioinformatic analysis. *PeerJ*, *7*. https://doi.org/10.7717/peerj.6375

Zhang, Y., Zhu, X., Qiao, X., Gu, X., Xue, J., Han, Y., Sun, L., Cui, M., & Liu, C. (2020). LIPH promotes metastasis by enriching stem-like cells in triple-negative breast cancer. *Journal of Cellular and Molecular Medicine*, *24*(16), 9125–9134. https://doi.org/10.1111/jcmm.15549

Zhao, G., Qin, Q., Zhang, J., Liu, Y., Deng, S., Liu, L., Wang, B., Tian, K., & Wang, C. (2013). Hypermethylation of HIC1 promoter and aberrant expression of HIC1/SIRT1 might contribute to the carcinogenesis of pancreatic cancer. *Annals of Surgical Oncology*, *20 Suppl 3*, S301-311. https://doi.org/10.1245/s10434-012-2364-9

Zhao, G., Wang, B., Liu, Y., Zhang, J.-G., Deng, S.-C., Qin, Q., Tian, K., Li, X., Zhu, S., Niu, Y., Gong, Q., & Wang, C.-Y. (2013). MiRNA-141, downregulated in pancreatic cancer, inhibits cell proliferation and invasion by directly targeting MAP4K4. *Molecular Cancer Therapeutics*, *12*(11), 2569–2580. https://doi.org/10.1158/1535-7163.MCT-13-0296

Zhou, J., Hui, X., Mao, Y., & Fan, L. (2019). Identification of novel genes associated with a poor prognosis in pancreatic ductal adenocarcinoma via a bioinformatics analysis. *Bioscience Reports*, *39*(8). https://doi.org/10.1042/BSR20190625

Zhou, J., Song, S., He, S., Zhu, X., Zhang, Y., Yi, B., Zhang, B., Qin, G., & Li, D. (2014). MicroRNA-375 targets PDK1 in pancreatic carcinoma and suppresses cell growth through the Akt signaling pathway. *International Journal of Molecular Medicine*, *33*(4), 950–956. https://doi.org/10.3892/ijmm.2014.1638

Zhou, J., Wang, H., Che, J., Xu, L., Yang, W., Li, Y., & Zhou, W. (2020). Silencing of microRNA-135b inhibits invasion, migration, and stemness of CD24+CD44+ pancreatic cancer stem cells through JADE-1-dependent AKT/mTOR pathway. *Cancer Cell International*, *20*(1), 134. https://doi.org/10.1186/s12935-020-01210-1

Zhu, G., Foletti, D., Liu, X., Ding, S., Melton Witt, J., Hasa-Moreno, A., Rickert, M., Holz, C., Aschenbrenner, L., Yang, A. H., Kraynov, E., Evering, W., Obert, L., Lee, C., Sai, T., Mistry, T., Lindquist, K. C., Van Blarcom, T., Strop, P., … Liu, S.-H. (2019). Targeting CLDN18.2 by CD3 Bispecific and ADC Modalities for the Treatments of Gastric and Pancreatic Cancer. *Scientific Reports*, *9*(1), 8420. https://doi.org/10.1038/s41598-019-44874-0

CHAPTER 7 APPENDIX

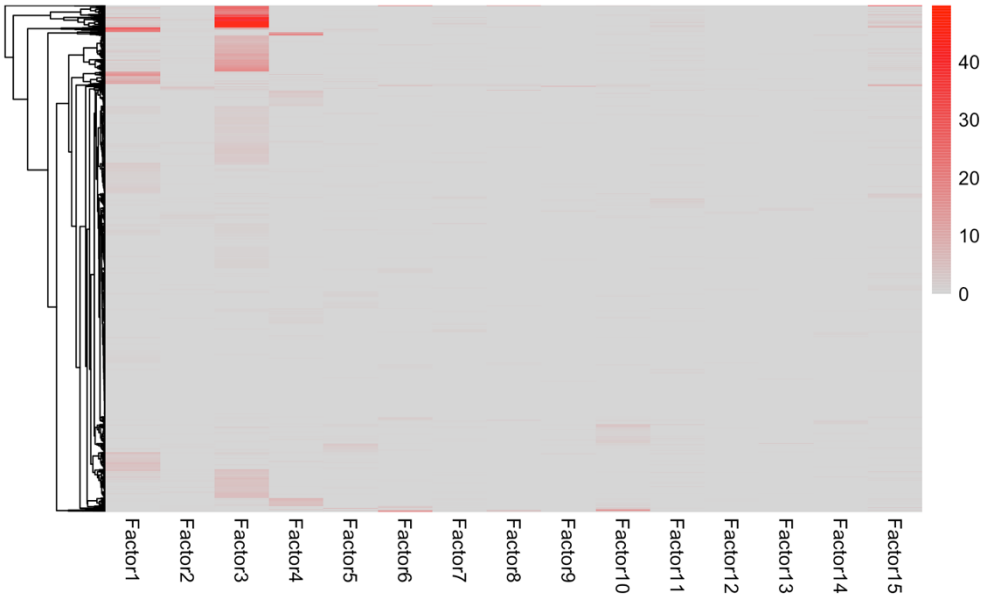
Supplemental Table A. Properties of MOFA models for PDAC.

Model	Data modality	Number of features	Pre-processing	Number of retrieved factors	Model scaling
1	Mutation	2317	No	15	No
	RNA sequencing	5000	5000 most variable genes		
2	Mutation	2317	No	10	No
	RNA sequencing	5000	5000 most variable genes		
3	Mutation	2317	No	10	No
	RNA sequencing	5000	5000 most variable genes		
	Methylation CpG	3199	Top 1% variable features		
4	Mutation	2317	No	10	No
	RNA sequencing	5000	5000 most variable genes		
	Methylation CpG	3199	Top 1% variable features		
	SCNV	5000	5000 most variable features		
5	Mutation	2317	No	10	No
	RNA sequencing	5000	5000 most variable genes		
	Methylation CpG	3199	Top 1% variable features		
	SCNV	5000	5000 most variable features		
	miRNA	734	No		
6	Mutation	2317	No	10	No
	RNA sequencing	5000	5000 most variable genes		
	Methylation CpG	3199	Top 1% variable features		
	SCNV	5000	5000 most variable features		
	miRNA	734	No		
	Protein	149	No		
7	same as Model 6			15	No

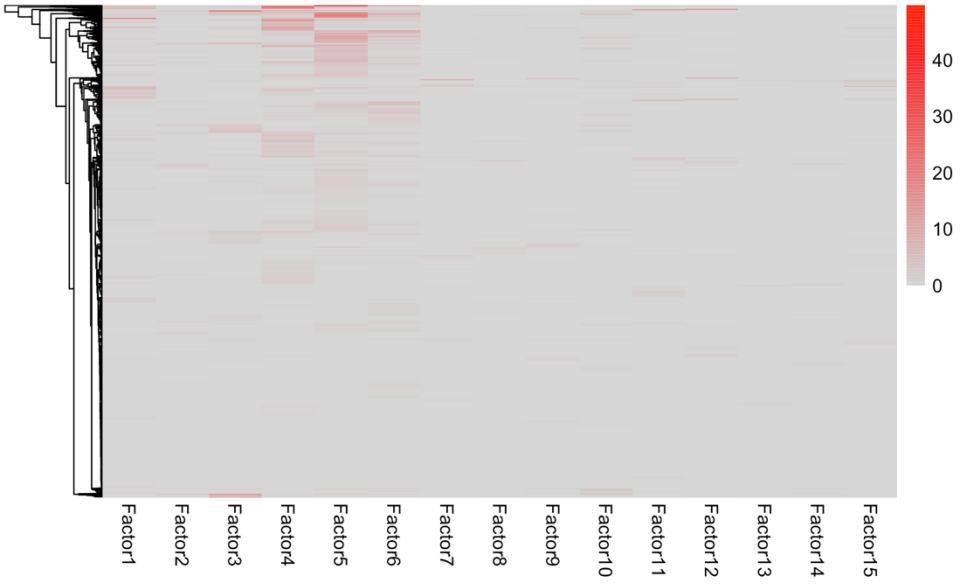
Supplemental Table A. Properties of MOFA models for PDAC (continued)

Model	Data modality	Number of features	Pre-processing	Number of retrieved factors	Model scaling
8	same as Model 6			15	Yes
9	RNA sequencing	5000	5000 most variable genes	15	No
	Methylation CpG	3199	Top 1% variable features		
	SCNV	5000	5000 most variable features		
	miRNA	734	No		
	Protein	149	No		
10	RNA sequencing	5000	5000 most variable genes	15	No
	Mutation	2317	No		
	Methylation at gene level	4571	Top 25% of features		
	SCNV	5000	5000 most variable features		
	miRNA	734	No		
	Protein	149	No		
11	Mutation	2317	No	15	No
	RNA sequencing	19774	No		
	Methylation at gene level	20089	No		
	SCNV	24776	No		
	miRNA	734	No		
	Protein	149	No		
12	same as model 11			10	No
13	same as model 9			15	Yes
14	same as model 11			15	Yes

A



B



Supplemental Figure A. Gene Set Enrichment Analysis of factor 1 mRNA modality with positive (A) and negative feature (B) weights.