Detection of somatic mutations and mutational signatures in pancreatic cancer using RNA-sequencing data

Pantlik, Jan

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:857654

University of Zagreb
Faculty of Science
Department of Biology

Jan Pantlik

Detection of somatic mutations and mutational signatures in pancreatic cancer using RNA-sequencing data

Master thesis

Zagreb, 2023.

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Jan Pantlik

Detekcija somatskih mutacija i mutacijskih potpisa tumora gušterače korištenjem podataka RNA-sekvenciranja

Diplomski rad

Zagreb, 2023.

This thesis was done in the Bioinformatics group at the Division of Molecular Biology, under the supervision of Assistant Professor Rosa Karlić and co-supervision of Assistant Paula Štancl. The thesis is submitted for grading to the Department of Biology at the Faculty of Science, University of Zagreb, with the aim of obtaining the Master's degree in molecular biology.

I would like to take this chance and express my gratitude to my mentor, Assoc. Prof. Dr. Rosa Karlić, and my co-mentor, Mag. Biol. Mol. Paula Štancl, for their patience, advice, and assistance during the preparation of this paper over the past year.

Furthermore, I would like to thank my friends and colleagues for all the gatherings, coffees, and “study sessions” without which these past few years of study wouldn't have been nearly as enjoyable and interesting.

My immense and endless gratitude goes to my parents, mom Nataša and dad Miro, my brother Vilim, uncle Boris, as well as grandmothers Nena and Višnja, and grandfathers Mladen and Živko. They have always been an unmeasurable source of support in everything I do, enabling me to be where I am today.

Lastly, I would like to give a special thanks to Lucija, who went through all the nerve-wracking moments with me during the creation of this master thesis and was an inexhaustible source of laughter and inspiration when it was most needed.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek Diplomski rad

Detekcija somatskih mutacija i mutacijskih potpisa tumora gušterače korištenjem podataka RNA-sekvenciranja

Jan Pantlik

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Duktalni adenokarcinom gušterače (engl. *pancreatic ductal adenocarcinoma*, PDAC) izrazito je smrtonosan karcinom, čije kasno otkrivanje i agresivnost predstavljaju velike prepreke koje dovode do neuspjeha u liječenju. Napreci u polju onkologije, prvenstveno unaprjeđenje tehnika ranog otkrivanja karcinoma i personalizirane terapije, znatno su poboljšali stope preživljavanja pacijenata oboljelih od karcinoma. Zahvaljujući napretku tehnologija sekvenciranja nove generacije značajno su proširene mogućnosti genetskog testiranja te je sad moguća preciznija kategorizacija genetskih varijanti detektiranih iz uzoraka unutar populacije te određivanje onkogena. Sekvenciranje RNA (RNA-Seq) ima veliki potencijal za identifikaciju somatskih mutacija i mutacijskih potpisa povezanih s različitim mutacijskim procesima. Cilj ovog istraživanja je korištenje RNA-seq podataka za otkrivanje varijanti i identificiranje mutacijskih potpisa prisutnih u uzorcima PDAC tumora. Mutacije su određene pomoću alata za pozivanje varijanti Strelka2, anotirane i uspoređene sa somatskim varijantama identificiranim iz podataka sekvenciranja eksoma. Detektirane mutacije su zatim korištene za izradu mutacijskih kataloga u svrhu identifikacije mutacijskog potpisa pomoću signature.tools.lib alata. Rezultati su pokazali da su SBS1, SBS5, SBS18 i SBS123 najčešći mutacijski potpisi u analiziranim uzorcima PDAC. Analiza mutacijskih potpisa korištenjem podataka sekvenci RNA pokazala je uzbudljiv potencijal kao nova i precizna metoda za karakterizaciju tumora, koja bi mogla pomoći u dijagnozi i liječenju tumora.

Ključne riječi: računalna genomika, mutacijski potpisi, RNA sekvenciranje
(73 stranice, 20 slika, 1 tablica, 141 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici

Mentor: doc. dr. sc. Rosa Karlić
Komentor: Paula Štancl, mag. biol. mol.

Ocjenitelji:

Doc. dr. sc. Rosa Karlić
Izv. prof. dr. sc. Tomislav Ivanković
Izv. prof. dr. sc. Maja Matulić

Rad prihvaćen: 07.09.2023.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Master thesis

Detection of somatic mutations and analysis of mutational signatures in pancreatic cancer cells using RNA-sequencing data

Jan Pantlik

Rooseveltov trg 6, 10000 Zagreb, Croatia

Pancreatic ductal adenocarcinoma (PDAC) is an extremely deadly cancer whose late detection and aggressive nature pose significant challenges, leading to treatment failures. Recent oncological advancements have substantially enhanced survival rates for various cancers by utilizing novel techniques for early detection and personalized therapies. The scope of clinical genetic testing has been significantly broadened by recent advancements in next-generation sequencing technologies, which have enabled the cataloging of genetic variation in population samples and the determination of cancer driver genes. RNA sequencing (RNA-Seq) has great potential for the identification of somatic mutations and mutational signatures associated with different mutational processes. The goal of this research is to use RNA-seq data to detect variants and identify mutational signatures present in PDAC tumor samples. Mutations were called with the variant-calling tool Strelka2, annotated, and compared to somatic variants identified from exome sequencing data. Detected mutations were then utilized to build mutational catalogs for mutational signature fitting with signature.tools.lib. The results showed that SBS1, SBS5, SBS18, and SBS123 are the most frequent mutational signatures in analyzed PDAC samples. Mutational signature analysis using RNA sequence data showed exciting potential as a novel and precise method for tumor characterization, which could help tumor diagnosis and treatment.

Keywords: computational genomics, mutational signatures, RNA sequencing

(73 pages, 20 figures, 1 table, 141 references, original in: English)
Thesis is deposited in Central Biological Library.

Mentor: Asst. Prof. Rosa Karlić, PhD
Co-mentor: Paula Štancl, MSc

Reviewers:
Asst. Prof. Rosa Karlić, PhD
Assoc. Prof. Tomislav Ivanković, PhD
Assoc. Prof. Maja Matulić, PhD
Thesis accepted: 07.09.2023.

ABBREVATIONS:

TCGA	The cancer genome atlas
PAAD	Pancreatic adenocarcinoma
RNA-seq	RNA sequencing
PDAC	Pancreatic ductal adenocarcinoma
miRNA	micro RNA
GATK	Genome Analysis Toolkit
WXS	Whole exome sequencing
WGS	Whole genome sequencing
SKCM	Skin cutaneous melanoma
NGS	Next Generation Sequencing

Table of contents

1 Introduction

1.1 Next generation sequencing technologies

A remarkable advancement in sequencing technologies can be witnessed in the past decade, which led to a revolution in the field of genomics, the expansion of research capabilities, and the discovery of new clinical applications. The greatest impact of these next-generation sequencing technologies (NGS) is that they enabled rapid, cost-effective, and high-throughput sequencing of DNA and RNA molecules. There are a number of different NGS platforms (Illumina, Ion Torrent, Pacific Biosciences) that employ different sequencing strategies but still share common principles. The key concept is to break the target genome into millions of small fragments and then sequence them at the same time (Behjati and Tarpey, 2013). This allows for generation of vast amounts of sequencing data in a single run, thereby accelerating genomic research (Behjati and Tarpey, 2013). Recent advancements in NGS technology have revolutionized large-scale resequencing of human samples for medical and population genetics purposes. There are numerous prominent initiatives such as the 1000 Genomes (Fairley et al., 2020), The Cancer Genome Atlas (Weinstein et al., 2013), and various other expansive exome sequencing projects that have been launched with the goal of comprehensively understanding the entirety of human genetic diversity (Depristo et al., 2011). This ability to objectively inspect the whole genome enables thorough exploration of genetic variations associated with diseases, identification of underlying mutations in Mendelian diseases, and investigation of spontaneously arising mutations with no existing genetic mapping (like the ones occurring in cancer genomes) (Depristo et al., 2011). These advancements have provided cost-effective avenues for uncovering invaluable insights into the genetic landscape, paving the way for improved medical diagnoses and clinical treatments for patients.

1.2 RNA sequencing

RNA sequencing, also referred to as RNA-seq, emerged as a transformative technique in the field of molecular biology and genomics more than a decade ago, and since then it has become an omnipresent instrument in the field of molecular biology, significantly influencing our comprehension of genomic function across various domains (Stark et al., 2019). RNA-seq is a powerful method of deep sequencing that allows the examination of all expressed genes in an organism, referred to as the transcriptome, including noncoding RNAs like micro-RNAs

(Whitley et al., 2016). The RNA-seq technique leverages next-generation sequencing technologies to generate millions of short sequence reads from RNA molecules present in a biological sample. The process involves several essential steps, the first of which is RNA extraction in the laboratory, where high-quality RNA is isolated from the sample of interest. This is followed by mRNA enrichment or ribosomal RNA depletion if needed (Stark et al., 2019). After isolation of the whole RNA and selective enrichment, RNA (or cDNA) must be fragmented to create short sequences of 200–500 base pairs that are adequate for sequencing (Whitley et al., 2016). Subsequently, the extracted RNA is converted into a complementary DNA (cDNA) library, often using reverse transcription, and then sequenced to a read depth of 10–30 million reads per sample on a high-throughput platform (Stark et al., 2019). The process of sequencing fragmented cDNAs generally generates concise reads with lengths ranging from 250 to 400 nucleotides (Whitley et al., 2016) and to make sense of this vast amount of raw sequencing data, a series of computational analyses are performed. These analyses include read alignment to a reference genome or *de novo* assembly, quantification of transcript abundances, identification of differentially expressed genes, functional annotation, and alternative splicing detection (Kukurba and Montgomery, 2015). These analyses provide valuable insights into gene expression patterns, differential gene regulation across various conditions or tissues, and the underlying molecular mechanisms involved in biological processes.

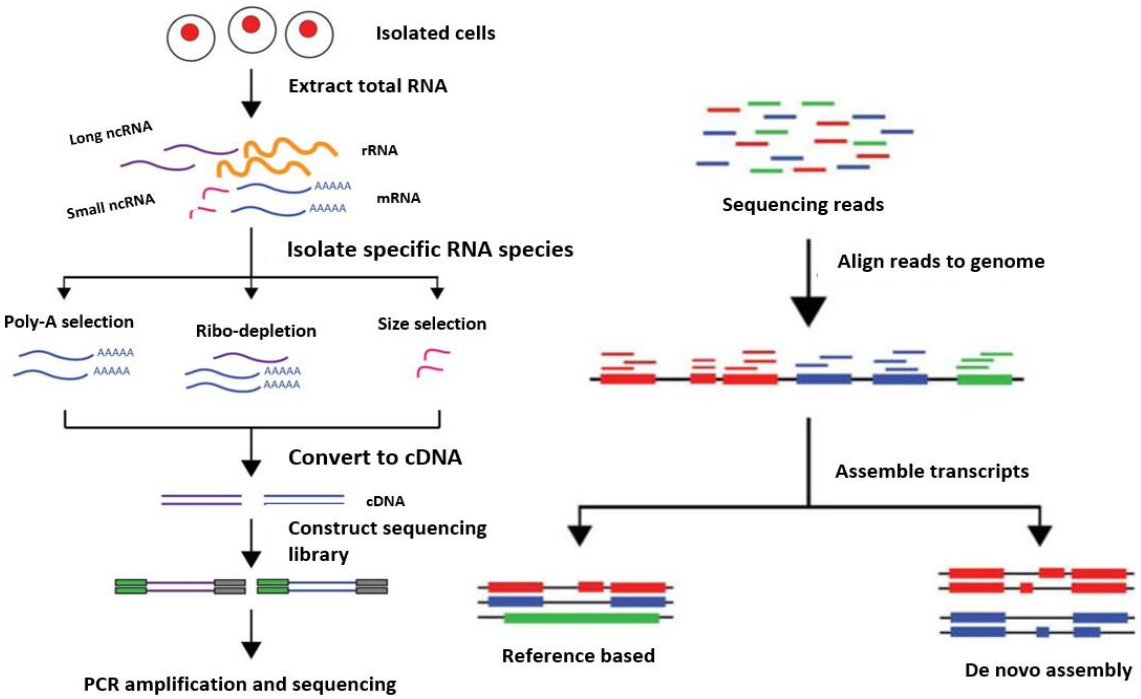

Figure 1. The illustration of the typical RNA-seq workflow and data analysis. Adapted from Kukurba & Montgomery, 2015

Although the term RNA-seq is frequently employed to encompass diverse methodological approaches and biological applications, differential gene expression (DGE) analysis continues to be the principal application of RNA-seq and is regarded as a standard research technique (Stark et al., 2019). The progress and advancement of RNA-seq have been fueled by advancements in technology, encompassing both wet-lab procedures and computational approaches and these developments have paved the way for a more comprehensive and unbiased understanding of RNA biology and the transcriptome compared to previous microarray-based methods (Stark et al., 2019). Constant improvements in sequence technology along with new scientific findings led to the development of numerous distinct methodologies derived from conventional RNA-seq protocols. Although RNA-seq is generally used to measure gene expression, it can potentially also be used for the analysis of somatic mutations (Coudray et al., 2018). Currently, most studies focusing on the identification of somatic mutations have primarily relied on analyzing DNA at the molecular level. With the development of somatic mutation detection within whole-genome or whole-exome sequencing data, significant advancements have been made in the field of the precision oncology (Xiao et al., 2021). One would think that all mutations within exons are transcribed into RNA and can be detected, but that is not always the case. Nevertheless, RNA serves as a dependable resource for distinguishing mutations that have actively influenced cellular functions (Long et al., 2022). Previous research efforts demonstrated that it is not only feasible and cost-effective to identify genomic variations in expressed exons through RNA-seq data analysis (Chepelev et al., 2009; Long et al., 2022; Radenbaugh et al., 2014), but also that developed protocols for detecting novel, tumor related somatic mutations can complement whole-exome sequencing in identifying somatic mutations specific to tumor genomes (Coudray et al., 2018).

1.3 Databases for genomic research

The National Center for Biotechnology Information (NCBI) database is an invaluable resource used for various research projects in the fields of biotechnology and genetics (NCBI, 2023). The NCBI database serves as a centralized repository of genetic and molecular biology information, encompassing a vast array of data derived from various sources, such as scientific literature, research projects, and publicly available sequence databases (NCBI, 2023). The NCBI database operates as a comprehensive suite of databases and tools that cover a wide range of biological data. These include databases like GenBank (DNA and protein

sequences); PubMed (repository of biomedical literature); GEO (gene expression data); and more. It also combines a user-friendly interface with powerful search functionalities, which enables conducting complex bioinformatics analyses and exploring different molecular relationships (NCBI, 2023).

The GEO (Gene Expression Omnibus) database is a public repository maintained by the NCBI (Edgar et al., 2002). The primary purpose of GEO is to provide a platform for researchers to deposit, access, and analyze gene expression data from a wide range of organisms and experimental conditions. GEO contains various types of gene expression data, including raw data files, processed data, and metadata associated with each experiment. The availability of vast amounts of gene expression data in GEO makes it a valuable resource for a wide range of research areas, including genomics, functional genomics, transcriptomics, and systems biology (Edgar et al., 2002).

The University of California, Santa Cruz (UCSC) database is a comprehensive and widely used resource in genomic research (UCSC Genome Browser, 2022). It offers a plentiful collection of genomic data and annotations for various organisms. The UCSC database is comprised of diverse genomic data types, including DNA sequences, gene annotations, regulatory elements, epigenetic marks, genetic variations, and more (Kent et al., 2002). All this information is easily accessible through the web interface, which enables researchers to explore and analyze genomic data within the context of a genome browser.

SRA (Sequence Read Archive) is a publicly available repository that stores raw sequencing data generated from technologies such as next-generation sequencing (NGS) platforms (Leinonen et al., 2011). The SRA accepts data from various sequencing methods, including whole-genome sequencing, transcriptome sequencing (RNA-seq), metagenomic sequencing, and others. The SRA provides resources for data exploration, analysis, and visualization, but raw sequences can also be retrieved from the SRA by using specific keywords, metadata, or accession numbers associated with experiments and further analyzed using downstream bioinformatics pipelines.

1.4 Cancer genomics

According to the NCI Dictionary of Cancer Terms, a tumor is defined as an abnormal cluster of tissue that arises when cells exhibit excessive growth and division or fail to undergo programmed cell death as expected (NCI Dictionary of Cancer Terms, 2023). Tumors can be

categorized as either benign, meaning they are not cancerous and do not invade nearby tissues, or malignant, meaning they are cancerous and can spread into nearby tissues or other parts of the body (NCI Dictionary of Cancer Terms, 2023).

Cancer is known to be one of the leading diseases with the greatest death expectancy worldwide. Based on the estimates provided by GLOBOCAN 2020 regarding the occurrence of cancer and related fatalities, approximately 19.3 million individuals were diagnosed with new cases of cancer, while the number of cancer-related deaths reached 10.0 million in the year 2020 (Sung et al., 2021). The hallmarks of cancer consist of eight biological capabilities that emerge throughout the progressive stages of tumor development in humans. They include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion, metastasis, reprogramming of energy metabolism and evading immune destruction (Hanahan and Weinberg, 2011). These fundamental characteristics of cancer serve as a framework for understanding and making sense of the intricate nature of tumorous diseases (Hanahan and Weinberg, 2011). In the past, the diagnosis and treatment of cancer were primarily determined by examining the physical characteristics of the tumor, its clinical symptoms, and its location within the body. Comprehensive investigations of cancer genomes over the last decade caused a paradigm shift, and now the concept of "cancer" encompasses a wide range of diseases, each of which is distinguished by unique combinations of gene mutations (Hudson et al., 2010; Stratton et al., 2009; Van Hoeck et al., 2019).

1.4.1 Cancer somatic mutations

Cancer, being a genetic condition, is partly fueled by the buildup of somatic mutations (Coudray et al., 2018; Weinstein et al., 2013). Cancer cells commonly accumulate somatic variations typically induced by environmental factors, whose accumulation affects specific pathways linked to cell growth, survival, angiogenesis, motility, and other fundamental characteristics of cancer, resulting in malignant transformation and ultimately cancer (Hanahan and Weinberg, 2011; Watson et al., 2013). Therefore, the comprehensive identification of somatic mutations in cancer, such as through the utilization of the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (COSMIC 2023, Tate et al., 2019), serves to characterize the intricate genomic complexities of the disease (Long et al., 2022; Watson et al., 2013), and also aids in the discovery of oncogenic mutations and driver genes that play a significant role in cancer development (Bailey et al., 2018; Long et al., 2022).

Interestingly, these mutations also present opportunities for targeted precision therapies aimed at combating the mutations responsible for tumor growth (Yu et al., 2015). Apart from that, individual-level somatic mutations possess their own oncogenic and therapeutic implications in various types of cancer, including lung cancer (Skoulidis and Heymach, 2019), bladder cancer (Wen et al., 2021), and glioblastoma (Lin et al., 2021; McDonald et al., 2015). It is expected that targeting these specific somatic mutations holds potential for personalized treatment strategies (Coudray et al., 2018).

Recent advancements in next-generation sequencing technologies have facilitated rapid, accurate, and cost-effective analysis of DNA and RNA samples, leading to the discovery of crucial mutations that drive cancer development (Coudray et al., 2018; Raphael et al., 2014). To date, the advancement of cancer diagnostic methods has primarily concentrated on the identification of these driver mutations, which confer growth advantages to cancer cells and promote the development of tumors (Stratton et al., 2009; Van Hoeck et al., 2019). Genetic testing targeting these driver genes enabled the identification of potential targets for treatment, the development of accurate mutation detection assays for cancer diagnosis, and the design of drugs that directly target proteins encoded by mutated driver genes (Bernards, 2010). While this knowledge has significantly contributed to drug development and improved cancer care, a considerable number of patients do not benefit from this approach due to low response rates to targeted drugs and a lack of reliable biomarkers (Van Hoeck et al., 2019). Next-generation sequencing (NGS) technologies have emerged as a valuable tool to address the need for enhanced molecular profiling of tumors and the identification of dependable biomarkers for patient stratification in cancer diagnostics (Van Hoeck et al., 2019). While these findings are laying the groundwork for novel targeted treatments across various types of cancer, despite advancements, some obstacles still need to be addressed for treatments to work properly (Coudray et al., 2018; E. Taylor et al., 2012; Paez et al., 2004).

1.4.2 Mutational signatures

Somatic mutations found in cancer genomes are the result of mutational processes occurring during the lineage of cells between the fertilized egg and the formation of the cancer cell (Alexandrov et al., 2020; Stratton et al., 2009). These mutational processes can stem from both exogenous factors, including exposure to environmental carcinogens or UV radiation, as well as from endogenous processes occurring within the body. Endogenous

mechanisms that contribute to somatic mutations involve normal mutational decay caused by the spontaneous deamination of methylated nucleotides, errors in DNA replication by error-prone polymerases leading to base misincorporation, and impaired function of DNA damage response (DDR) genes resulting in unrepaired or improperly repaired DNA damage (Helleday et al., 2014; Van Hoeck et al., 2019). Notably, each of these processes results in a distinct pattern of mutations, which are referred to as mutational signatures. (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Van Hoeck et al., 2019). Although mutational signatures are a relatively recent concept in cancer biology, the idea of linking mutational processes with mutational patterns is not new (Van Hoeck et al., 2019). The abundance of mutations present in each patient's cancer has provided us with a remarkable capability to identify these common patterns of mutations known as mutational signatures that emerge during the process of tumor formation (tumorigenesis) (Koh et al., 2021). Each mutational signature is characterized by base substitutions, small insertions and deletions (indels), genome rearrangements, and chromosome copy number changes (Alexandrov et al., 2020). The initial mutational signatures that were introduced are base substitutions (Van Hoeck et al., 2019). In these types of mutations, a signature is characterized by a specific change in the DNA base, along with the 5' and 3' flanking bases (Alexandrov et al., 2020). The main classification of SBS consists of 96 distinct trinucleotide classes since there are six categories of base substitution and a total of 16 possible sequence contexts. It is possible to extract mutational signatures from large groups of cancer patients whose DNA has been sequenced using a computational framework to identify recurring patterns within the cohort's 96-mutation matrix. Ultimately, each pattern represents the relative proportion of trinucleotide mutations and serves as a reflection of a mutational signature (Van Hoeck et al., 2019).

It is important to note that it is possible to detect mutations in an individual cancer genome that may result from multiple different mutational processes due to the simultaneous or sequential action of several internal or external factors causing mutations in a cell's genetic material over its lifetime, and because of that this, genome will then incorporate multiple overlapping mutational signatures (Van Hoeck et al., 2019). Certain mutational processes remain consistently active throughout the entire lifespan of the cancer cell (referred to as 'clock signatures'), while others operate periodically, some of which are influenced by the patient's lifestyle (Alexandrov et al., 2015). Consequently, numerous mathematical approaches have been employed to systematically interpret and characterize the mutational processes involved in cancer. These methods aim to extract mutational signatures from

collections of somatic mutations, estimate the proportion of mutations associated with each signature in individual samples, and annotate specific mutation classes within each tumor (Alexandrov et al., 2015, 2020; Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013; Blokzijl et al., 2018; Fischer et al., 2013; Gehring et al., 2015). Therefore, mutational signatures can serve as indicators of the occurrence (or non-occurrence) of various cellular activities in cancer cells.

Recent research has demonstrated that mutational signatures can serve as biomarkers for specific characteristics of cancer (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Nik-Zainal et al., 2012; Van Hoeck et al., 2019). Consequently, they can be hold promising clinical value as diagnostic tools and as predictors of cancer treatment response (Davies et al., 2017). However, a crucial requirement for conducting mutational signature analysis is the availability of comprehensive mutational data covering the entire genome from numerous distinct cancer cases (Van Hoeck et al., 2019). The cost of whole-exome sequencing (WXS) has decreased in the last decade, and with that came the successful completion of extensive pan-cancer genomic datasets, including The Cancer Genome Atlas (TCGA) (McLendon et al., 2008), Welcome Trust Sanger Institute's Cancer Genome Project (Pleasance et al., 2010), and the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010). These datasets provided essential resources and data necessary for conducting cancer research and helped to establish mutational signature analysis as a novel way for discovering biomarkers, diagnosing tumors, and guiding treatment decisions (Van Hoeck et al., 2019). In the present day, mutational signature analyses have emerged as a fundamental aspect of genomic research due to their ability to offer fresh perspectives on the underlying causes of specific cancers (Koh et al., 2021). They can uncover both environmental and endogenous causes of mutations within each tumor. As a result, the field of mutational signatures analysis is gaining increasing recognition and holds great promise for enhancing our understanding of individual cancers and their potential implications for clinical practice (Koh et al., 2021). Until recently, mutational signatures analyses were made using whole genome sequencing data (WGS), but newly developed protocols allowed using RNA-seq and WXS data as well in an attempt to gain a better understanding of complex tumor biology (Coudray et al., 2018).

1.4.3. Large-scale cancer genomics projects

The ICGC (International Cancer Genome Consortium) is an international collaborative effort focused on deciphering the genomic alterations associated with various

types of cancer (Hudson et al., 2010). The primary goal of the ICGC project is to thoroughly characterize the genomic changes occurring in different cancer types. This involves analyzing DNA sequencing data to identify somatic mutations, structural variations, copy number alterations, and epigenetic modifications in tumor genomes. The ICGC Data Portal serves as a central repository for accessing and exploring the genomic data generated by member projects. It also provides different tools and resources for data analysis and visualization.

The Cancer Genome Atlas (TCGA) (NCI The Cancer Genome Atlas Program, 2023) is a large-scale collaborative initiative aimed at comprehensively analyzing various types of cancer at the molecular level, including genomic, transcriptomic, and epigenomic alterations (Weinstein et al., 2013). This project integrates high-throughput sequencing technologies and innovative bioinformatics approaches to provide a comprehensive molecular characterization of cancer, enabling researchers to identify key genetic mutations, gene expression patterns, and epigenetic modifications associated with different cancer types. TCGA maintains an open-access policy for its database, which allows scientists worldwide to access and utilize this invaluable resource to accelerate cancer research and ultimately improve patient treatment outcomes.

1.5 Pancreatic ductal adenocarcinoma (PDAC)

Pancreatic ductal adenocarcinoma (PDAC) is the predominant type of pancreatic cancer, comprising over 80% of pancreatic cancer cases (NCI Pancreatic Ductal Adenocarcinoma Study, 2023). PDAC is an extremely destructive disease, one of the most aggressive and lethal malignancies, characterized by a bleak prognosis and increasing occurrence (Orth et al., 2019). The development of pancreatic ductal adenocarcinoma is characterized by a poor prognosis, primarily due to its complex and multifactorial nature (Sarantis et al., 2020). As of now, pancreatic ductal adenocarcinoma ranks as the fifth leading cause of cancer-related deaths in the United States, with a 5-year overall survival rate of 12% (Siegel Mph et al., 2023). The incidence of PDAC is predicted to continue increasing in the future, and projections suggest a more than two-fold rise in the number of new diagnoses and PDAC-related deaths within the next decade in the United States and European countries (Orth et al., 2019; Cancer Research UK, 2023; Quante et al., 2016; Rahib et al., 2014). This disease originates in the ducts of the pancreas, which are responsible for transporting digestive enzyme-containing fluids to the small intestine (NCI Pancreatic Ductal Adenocarcinoma Study, 2023). The most commonly observed symptoms in patients with

PDAC include weight loss, abdominal pain, and jaundice (Porta et al., 2005). However, there are also less common symptoms, such as new-onset type 2 diabetes and thromboembolic disease, that can be associated with PDAC (De Souza et al., 2016; Khorana, 2012; Sarantis et al., 2020).

According to the Pancreatic Ductal Adenocarcinoma Study by the National Cancer Institute (NCI), several risk factors are associated with the development of PDAC. These include having a family history of the disease, a previous history of chronic inflammation of the pancreas (pancreatitis), Lynch syndrome, diabetes, being overweight or obese, and smoking (NCI Pancreatic Ductal Adenocarcinoma Study). In a subgroup comprising around 5-6% of all PDAC patients, there are additional risk factors in the form of genetic predispositions. These include germline mutations in genes such as *BRCA1/2*, *ATM*, *MLH1*, *TP53*, or *CDKN2A* (Hu et al., 2018; Orth et al., 2019; Petersen et al., 2010; Pihlak et al., 2017).

Challenges associated with early detection and the limited effectiveness of available treatments led to PDAC being an extremely aggressive and fatal malignancy. The significant obstacles that contribute to therapeutic failure are the late identification of the disease and its notably aggressive nature (Orth et al., 2019). The effectiveness of treatments for pancreatic ductal adenocarcinoma such as surgery, radiation, and chemotherapy are limited due to various factors, including the extensive heterogeneity of genetic mutations and the dense stromal environment (Sarantis et al., 2020). The outcome of pancreatic ductal adenocarcinoma treatment primarily depends on the stage of the disease at the time of diagnosis. Typically, PDAC is diagnosed at an advanced stage since symptoms tend to manifest only when the disease has already progressed and metastasized to different sites (Sarantis et al., 2020). Treating pancreatic cancer is a challenging task that involves addressing issues at both the genetic and cellular levels. The high degree of genetic mutations in pancreatic tumors contributes to gene instability, which plays a crucial role in the growth of PDAC and its resistance to treatments (Sarantis et al., 2020). PDAC is characterized by significant genetic heterogeneity, not only among different patients but even within a single primary tumor (Sarantis et al., 2020). Currently, the only potentially curative treatment option available is surgical resection followed by adjuvant chemotherapy (Orth et al., 2019). However, in recent years, combined treatments with immunotherapy have shown success in treating different cancer types (Sarantis et al., 2020). Despite the revolutionary impact of immunotherapy in cancer treatment, it presents significant challenges when applied to

pancreatic ductal adenocarcinoma. This is primarily due to the nonimmunogenic nature of PDAC as well as its immune-suppressive and therapy-resistant microenvironment (Sarantis et al., 2020). In contrast to cancers like lung cancer with EGFR mutations and melanoma with BRAF mutations, where targeted treatments have shown efficacy due to a relatively high percentage of patients sharing the same cancer-causing mutation (Bethune et al., 2010; Cheng et al., 2018), pancreatic cancer presents a different challenge because it exhibits a wide variety of mutations that contribute to its development, with each mutation being present in a small percentage of patients (Grant et al., 2016; Sarantis et al., 2020). This genetic heterogeneity makes it difficult to develop targeted therapies that can effectively address the diverse mutational landscape of pancreatic cancer (Sarantis et al., 2020). These factors make it difficult to achieve successful outcomes with targeted immunotherapy in the context of PDAC.

1.5.1 PDAC mutational signatures

The most common mutational signature found in pancreatic cancer is Signature 3 (COSMIC 2023; Forbes et al., 2017, Tate et al., 2019). This specific mutational pattern has been found to be strongly associated with alterations in genes involved with homologous recombination repair machinery, showing both germline and somatic mutations in the *BRCA1* and *BRCA2* genes (Forbes et al., 2017; Polak et al., 2017). Research done by Polak et al. in 2017 demonstrated that Signature 3 can also be found in samples without *BRA1* or *BRCA2* mutations but with a mutational landscape similar to that of samples with homolog repair deficiency. Signature 3 is found particularly in pancreatic, but is also common in breast and ovarian cancers (Forbes et al., 2017). In the case of pancreatic cancer, patients who respond positively to platinum-based therapy often exhibit mutations associated with Signature 3 (Forbes et al., 2017).

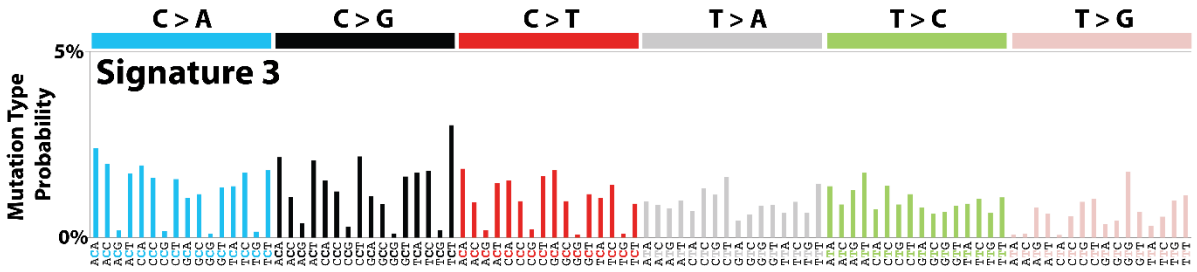

Figure 2. The pattern of Mutational Signature 3 displayed using the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G. Information on the different combinations of bases immediately 5' and 3' to the mutated base is taken into account to allow for the generation of 96 possible mutation types (6 types of substitution ∗ 4 types of 5' base ∗ 4 types of 3' base). Adapted from Forbes et al., 2017.

When researching signatures of mutational processes in human cancer samples in 2013, Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. detected three more signatures in pancreatic cancer sample sequences derived from WGS and WXS. These were Signature 1B attributed to age, Signature 2 attributed to activity of the AID/APOBEC family of cytidine deaminases, and Signature 6 associated with the presence of defective DNA mismatch repair commonly found in microsatellite unstable tumors (Alexandrov et al., 2020; Forbes et al., 2017). In research from 2017. by Connor et al., mutational signatures that correlated with detected mutations in PDAC sequencing data from International Cancer Genome Consortium (ICGC) database were also Signature 8 and Signature 17 (Connor et al., 2017). Considering useful results from recent cancer genomics studies, it is clear that mutational signature analysis, along with mutation profile analysis, will have an important role in optimizing the diagnosis and treatment of cancer patients, including those suffering from PDAC.

1.6 Bioinformatic tools and software environments used for variant discovery in cancer genome research

1.6.1 Programming language R

R is a programming language and open-source software environment widely used for statistical analysis and data visualization. It can be run on the majority of platforms (Windows, UNIX), and it provides a wide variety of statistical and graphical techniques that are highly extensible (R: The R Project for Statistical Computing, n.d.). One of the biggest strengths of R is its broad selection of user contributed packages, which greatly extends its functionality. Additionally, the graphics capabilities of R are robust and adaptable, making it possible to generate high-quality visual representations of analyzed data.

1.6.2 FastQC

FastQC is a widely used bioinformatics tool designed for quality control of high-throughput sequencing data. It provides the user with a comprehensive assessment of the quality and composition of raw sequencing reads and generates detailed reports that point out potential issues with generated read sequences (Andrews, 2010). These insights help researchers evaluate the overall quality of their sequencing data, identify potential sources of bias or errors, and make decisions about data processing and downstream analysis (Andrews, 2010).

1.6.3 Samtools

Samtools is a commonly used software package for the manipulation and analysis of data stored in the widely adopted SAM (Sequence Alignment/Map) and BAM (Binary Alignment/Map) file formats (Danecek et al., 2021). These formats are prevalent in the field of bioinformatics and serve as standards for storing and exchanging sequence alignment data. Samtools provides a broad set of functionalities to effectively work with SAM and BAM files. It offers to complete multiple operations on sequence alignment data, such as format conversion, sorting, indexing, filtering, and variant calling (Danecek et al., 2021).

1.6.4 Read Alignment

One of the essential stages in most genomic analysis workflows is matching sequenced reads to a reference genome. This process is called read alignment (also known as read mapping), and its purpose is to determine the potential location of each read using the sequence of the reference genome as a template (Schilbert et al., 2020). This is done by using computational algorithms that have progressed through the years along with technological advances, leaving us with a wide variety of alignment methods. Some of the most popular bioinformatic tools used for aligning reads are: Bowtie2, BWA-MEM, STAR, CLC Genomics Workbench (Qiagen), GEM3, Novoalign, Segemehl, and BBMap (Alser et al., 2021).

1.6.4.1 STAR aligner

Spliced transcripts alignment to a reference (STAR) is a tool that gives out information about where on the human genome given reads originated from (Dobin et al., 2013). It is specifically designed to address many of the challenges with RNA sequenced reads mapping. The STAR algorithm uses an approach that considers the existence of spliced alignments

(Dobin et al., 2013). Splice alignment can be found in RNA-seq data, considering isolated RNA sequences are formed by the splicing of transcribed intron regions from pre-messenger RNA molecules (Regan et al., 2021). These alignments are the result of the existence of spliced junctions, defined as borders separating the introns from the exons in non-mature messenger RNA molecules. (Regan et al., 2021). Mapping with STAR has two phases: 1) Indexing of a reference genome using the information from FASTA file reference sequences and GTF file reference genome annotations; 2) Mapping read sequences on the indexed reference genome (Dobin et al., 2013).The highly efficient STAR mapping algorithm consists of two steps. The first step is seed searching, where STAR searches for the longest read sequences that exactly match one or more locations in the reference genome (Dobin et al., 2013). These are marked as seed1. After that, STAR searches again for only the unmapped part of the read to find the next longest sequence that exactly matches the reference genome and marks it as seed2 (Dobin et al., 2013). If it is impossible to find exact matches for each part of the read (because of mismatches or indels, etc.), the previous seed will be extended, and if the extension has poor quality alignment, it will be soft clipped (Dobin et al., 2013). The second step is clustering, stitching, and scoring, where STAR stitches together the separate seeds to get full alignment (Dobin et al., 2013). The seeds are first clustered by the proximity of the “anchor seeds”, the alignment of the read is scored based on mismatches, indels, gaps, etc., and then the seeds are stitched based on the best alignment (Dobin et al., 2013). An illustration of this process can be seen in Figure 3.

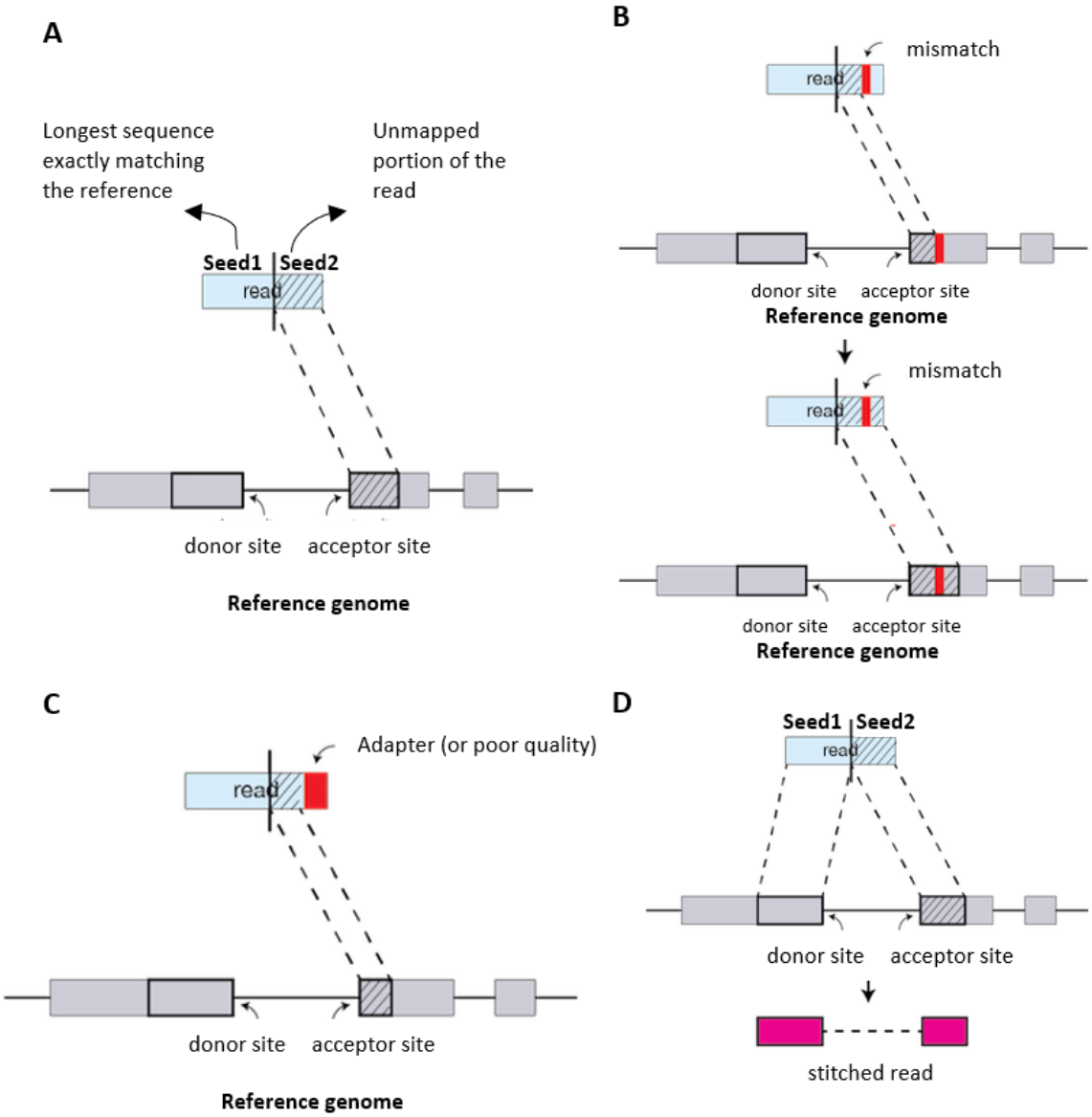

Figure 3.Illustration of the STAR aligner algorithm strategy; A) Seed searching, searching for the longest sequence that exactly matches one or more locations on the reference genome;. B) Extension of previous MMPs if STAR didn't find an exact matching sequence on the reference; C) Soft clipping of the poor quality or adapter sequence if extension didn't produce good alignment; D) The separate seeds are clustered together and stitched based on proximity. Adapted from Introduction to RNA-seq using high-performance computing (HPC), 2021).

1.6.5 GATK

GATK stands for Genome Analysis Toolkit (GATK, 2023). It is a widely used software package developed by the Broad Institute for the purpose of analyzing high-throughput sequencing data (GATK, 2023). GATK software offers a broad suite of tools and algorithms for processing next-generation sequencing data. Its notable capabilities include variant calling, including single nucleotide polymorphisms (SNPs) and structural variations, as well

as sequence data preprocessing steps like base quality score recalibration and indel realignment, which are designed to enhance variant calling accuracy (Auwera and O'Connor, 2020; Depristo et al., 2011). GATK abilities extend further and involve variant annotation, haplotype phasing, and variant filtering, making it a valuable resource in genomics research and clinical applications (Auwera and O'Connor, 2020).

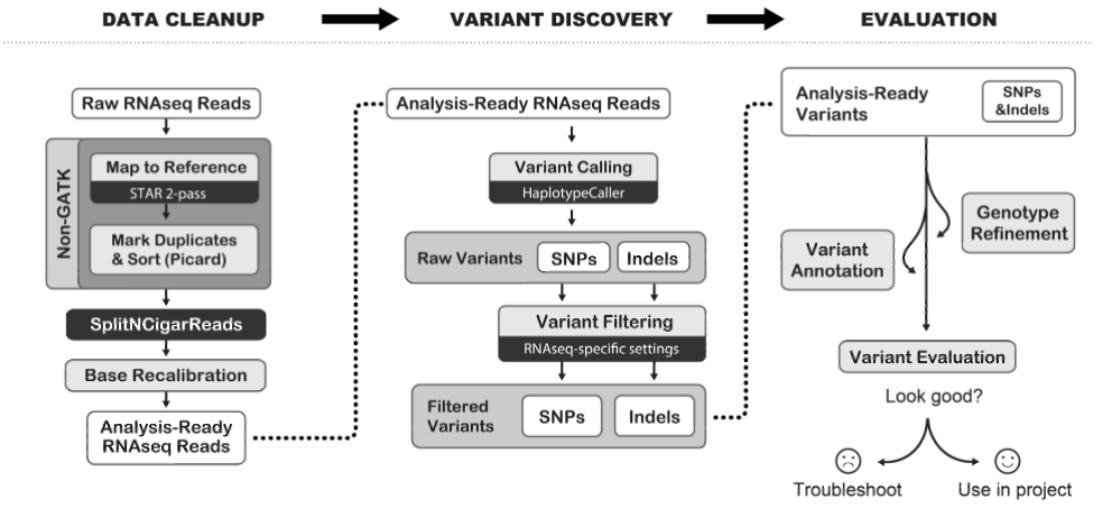

Figure 4. Illustration of GATK best practices workflow for RNAseq short variant discovery (SNPs + Indels). Adapted from (GATK, 2023)

1.6.6 Variant calling and annotation

One of the primary applications of next-generation sequencing is to identify genetic variations within large populations of closely related samples (Danecek et al., 2011). The name of this process is variant calling. Variant calling is essentially the procedure of distinguishing genuine genetic variations from anomalies arising during library preparation, sample enrichment, sequencing, and read alignment (Xu, 2018). This has been a highly dynamic area of research for numerous years, leading to the development of numerous variant calling tools, many of which are freely accessible. Widely used variant calling tools are: DeepVariant, Strelka2, Octopus, FreeBayes, Platypus, Samtools/mpileup, SNVer, VarScan, VarDict.

The process of analyzing and interpreting genetic variants identified in an individual's genome is called variant annotation (McCarthy et al., 2014). Variant annotation involves adding functional and clinical information to each variant (affected gene, type of variant, potential implications for health, etc.). This is a crucial step in the analysis of sequencing data,

considering its results have the greatest impact on the conclusions about diseases that were researched in the study (McCarthy et al., 2014). Furthermore, it is essential for understanding the functional and clinical significance of discovered genetic variants, enabling accurate diagnosis and personalized treatment, and advancing knowledge of genetic diseases. Frequently used tools for annotating detected mutations are SnpEff, AnnoVar and VEP.

1.6.6.1 Strelka2

Strelka2 is a useful open-source tool for detecting small variants in sequencing data (Kim et al., 2018). Compared to its predecessor Strelka, Strelka2 exhibits enhanced accuracy, speed, and efficiency in identifying genetic variations, including single nucleotide variants (SNVs), insertions, and deletions (Kim et al., 2018). The Strelka2 variant calling algorithm can be simply described as a two-step process. In the first step, the algorithm finds potential variant sites and implements local assembly around those sites (Kim et al., 2018). The second step consists of utilizing a refined statistical model that considers local sequence context, mapping qualities, and other characteristics to estimate the probability of occurrence of each identified variant at that specific place in the genome (Kim et al., 2018). The result of these refined steps is highly accurate variant calling with reduced rates of false positives and false negatives.

The somatic variant calling algorithm begins with parameter estimation from sample data (Kim et al., 2018). Parameter estimation involves promptly estimating the sequencing depth for each chromosome using read alignments in the BAM file, but estimation is specifically performed solely on the normal samples (Kim et al., 2018). This step is crucial for accurately assessing the likelihoods and making reliable somatic variant calls because it estimates the parameters of the statistical models used to evaluate the evidence for somatic variants.

By comparing the genotype likelihoods in the tumor sample with the expected distribution based on the normal sample model, Strelka2 estimates the probability for each candidate variant position in the tumor sample to be a true somatic mutation (Kim et al., 2018). The final set of somatic variants is determined by applying an empirical threshold to calculated somatic variant probabilities, and only variants with somatic variant probabilities higher than this threshold are confident somatic variant calls (Kim et al., 2018).

The final phase of variant calling is empirical scoring and filtering. In this phase, additional information is extracted in the form of predictive features and used together with

calculated probabilities to enhance the precision of variant calls. This integration is executed with the Empirical Variant Scoring (EVS) model, which acts as a supervised random forest classifier (Kim et al., 2018). This model is trained on labeled data from sequencing runs conducted under diverse conditions (different sequencers, different sample preparation techniques, and different coverage levels) (Kim et al., 2018). The EVS model assigns an aggregate quality score to each variant, enabling the exploration of the precision-recall curve in a convenient manner (Kim et al., 2018).

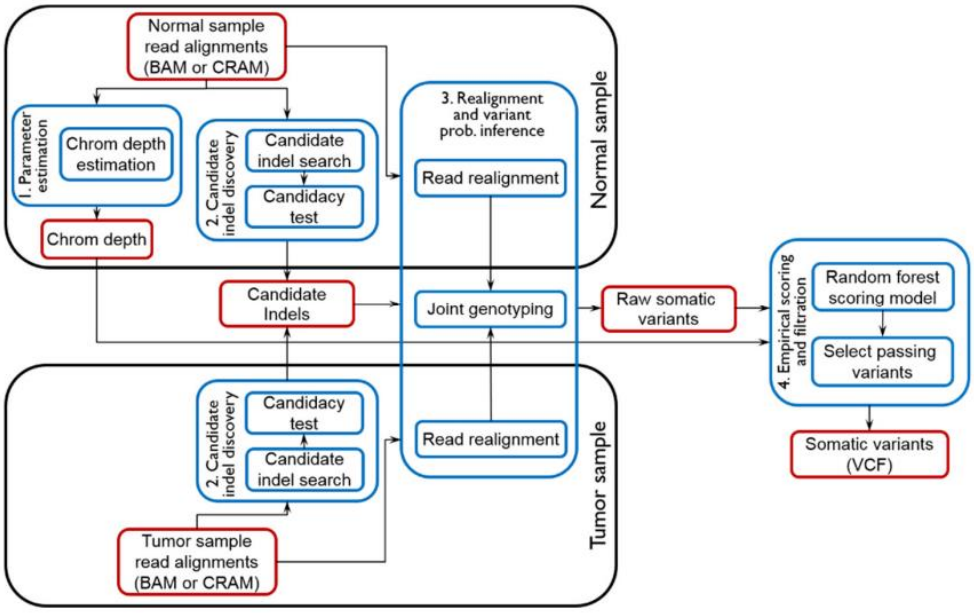

Figure 5. The Strelka2 algorithm workflow. Adapted from (Kim et al., 2018)

1.6.6.2 SnpEff

SnpEff is a practical, multiplatform, open-source computer program designed to annotate variants and predict the coding effects of genetic variations (Cingolani et al., 2012). This tool is capable of categorizing effects of various genome aberrations resulting from mutational processes such as single nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs), insertions, and deletions (indels) in whole genome sequencing results (Cingolani et al., 2012). SnpEff is extensively used in major academic institutions for research, pharmaceutical companies and clinical sequencing projects because of its main assets that are: (1) speed – it enables generation of thousands of predictions per second; (2) flexibility – it allows the inclusion of custom made genomes and annotations; (3) it integrates with Galaxy, an open-access web-based platform for computational bioinformatic research (Afgan et al., 2022); (4) it is compatible with multiple species and multiple codon usage tables (such as mitochondrial genomes); (5) it seamlessly integrates with the GATK which makes

work much easier; and (6) it demonstrates the ability to perform non-coding annotations, enabling researchers to explore and interpret variants located outside of coding regions (Cingolani et al., 2012).

2 Goals

Next-generation sequencing technologies have revolutionized cancer research by providing tools for fast and accurate sequencing of large numbers of samples, leading to a better understanding of the complex genomic landscape of tumors. Tools for identifying and annotating variations across the genome, like Strelka2, are well developed and widely accessible, making it possible to detect genetic variations and explore the mutational landscape of tumor samples. In addition to the currently used whole-genome sequencing and whole-exome sequencing technologies, RNA sequencing (RNA-seq) is also thought to have great potential in cancer genomics by providing information that could be used in clinical diagnosis and treatment. Since the use of RNA-seq is not as ubiquitous in clinical practice, in this research I will use RNA-seq data to detect variants and identify mutational signatures in PDAC tumor samples and compare my results to those obtained by more established techniques. The goals of this research are:

1. To identify and annotate somatic mutations using RNA-seq data from PDAC tumor tissue samples;
2. To characterize identified mutations and developed mutational landscape;
3. To analyze the mutational signatures of identified somatic variants;
4. To compare results with publicly available whole exome sequencing (WXS) pancreatic cancer data and assess the concordance of the results of RNA sequencing and whole exome analyses.

3 Materials and methods

3.1 Downloading and preprocessing of raw sequencing data

In this work, I analyzed publicly available sequencing data from the NCBI database. The analyzed data can be found in the GEO database under the GEO series designation GSE130688, while the raw sequences are saved in the SRA database. Samples were collected from patients with PDAC by the Biochemistry Department of the University of São Paulo. Surgically removed fragments of PDAC and peritumoral nonmalignant tissues were collected and then frozen in liquid nitrogen. The RNA isolation was done using Trizol reagent (ThermoFisher), and purification was done according to the manufacturer's protocol. After isolation, RNA-seq libraries were generated using the Illumina TruSeq Stranded Total RNA LT sample preparation kit with Ribo-Zero Gold according to the standard manufacturer's protocol (da Paixão et al., 2022). Total strand-specific RNA-seq libraries from 15 paired samples were sequenced using the Illumina HiSeq 1500/2500 platform, and the resulting sequences were then uploaded to the SRA database. I downloaded raw sequenced reads from the publicly available SRA database under the SRA Study designation mark SRP194936 to conduct further analysis. Analysis was not possible for three of the downloaded samples because of the occurrence of corrupted files during pipeline implementation and because of that analysis was made on the remaining twelve samples.

3.2 Quality control

Raw sequences can be unreliable because of errors during sequencing, such as using the wrong primers or old polymerases, sequencing machine failure, etc. This causes a loss of quality, which is why one of the most important steps in the whole study is quality control. I used publicly available FastQC software to generate a detailed report of the characteristics of raw sequenced reads and access their quality. Afterwards, I used Trimmomatic to filter and trim raw sequenced reads using the parameters SLIDINGWINDOW:5:10 MINLEN:60 HEADCROP:5. I did one more quality control over trimmed and filtered reads.

3.3 Read mapping and data cleanup

Before calling variants, it must be known exactly where sequenced reads belong on the reference genome for the variant caller to be able to compare them with the reference. For that

purpose, I used the STAR aligner to map our sequenced reads of interest to GRCh38 (hg38) downloaded from the NCBI database. For both phases of mapping, I used only the default options when running STAR. The result of mapping with STAR are alignments in BAM format, which hold the information about where on the reference genome the extracted read sequences come from.

After mapping and before variant calling, mapped read data needs to be cleaned for the variant caller to make good calls on the existence of possible variants. I followed GATK's best practices workflow for RNAseq short variant discovery for the steps of data cleanup. First, I used the MarkDuplicates tool with standard options for effectively identifying and marking duplicated reads. Duplicated reads are the result of errors occurring in PCR steps during sequencing, and their occurrence varies depending on the phase of the sequencing process. The result is a BAM file with marked duplicates and a metrics file with the number of duplicates.

After cleaning the duplicated reads, I used the GATK tool AddOrReplaceReadGroups to add information about the read group, which other tools used in the pipeline need to function properly. Added information included read-group ID, read-group library, read-group platform, read-group platform unit (e.g., run barcode), and read-group sample name. Next, I needed to index the BAM file with read alignments, and for that, I used the BuildBamIndex tool from the GATK best practices workflow. The result is a BAM file with indexed and sorted reads. Considering that in RNA-seq data, reads can span exon-exon junctions or harbor splice junctions, it's necessary to split the reads at these junctions for variants to be called correctly. The tool I used for reconstruction of alignments that span intron regions is called SplitNCigarStrings. The output is a BAM file with reads split at N CIGAR elements and CIGAR strings updated.

One of the most important steps during data preprocessing is base quality score recalibration. The primary goal of this step is to identify and flag systematic errors that occurred during the estimation of the accuracy of individual base calls performed by a sequencing machine. Since the quality scores play an important role in the decision of the algorithm whether something will be considered a variant or not, it is important to exclude any kind of systematic bias in my data. In the process of Base Quality Score Recalibration (BQSR), machine learning methods need to be utilized to empirically model these errors and make appropriate adjustments to the quality scores. For that, I used two tools from the GATK best practices workflow called BaseRecalibrator and ApplyBQSR. BaseRecalibrator is

responsible for building a recalibration model by using all reads from an indexed BAM file and a collection of known variants (that I downloaded from the UCSC genome browser), while ApplyBQSR creates a new BAM file by adjusting quality scores based on the previously created recalibration model. The result of this step is analysis-ready RNA-seq reads saved in a BAM file.

3.4 Variant discovery and filtering

After the read data is prepared, variant calling can be done. For variant discovery, I used Strelka2, an open-source tool specifically designed for small variant calling in research and clinical germline and somatic sequencing applications (Kim et al., 2018). I used Strelka2 in the mode for calling somatic variants from paired tumor—normal samples originating from the same patient.

3.5 Variant annotation

For annotating identified variants and predicting their effects on genes and proteins, I used the bioinformatic tool SnpEff. For variant annotation, certain files are required: 1) a reference genome sequence, which serves as a baseline against which the variants are compared; 2) an organism-specific annotation database, which provides information about genes, transcripts, and functional elements; 3) VCF (Variant Call Format) file with genetic variants obtained through sequencing and variant identification. SnpEff determines the impact and potential consequences of a variant by considering its genomic location, the affected gene(s), and the type of variant (missense, nonsense, synonymous, or frameshift). The functional consequences of each variant are predicted based on the annotation, and variants are classified into categories (high impact, moderate impact, low impact, and modifier). The result of variant annotation with SnpEff is a detailed output file in VCF format containing information about each variant's impact, gene annotations, and predicted effect, which can be further processed in downstream analyses.

3.6 Analysis of detected variants

I loaded the annotated VCF files with annotated somatic variants from 12 analyzed samples into R Studio using the read.vcfR function from the vcfR package, saved them into a list, and assigned them patient names (Patient_1 to Patient_15).

Starting the analysis, first I filtered all the variants that didn't pass variant filtration, and these are all the variants that didn't have a "PASS" annotation in the FILTER column of the data frame. Since I am interested only in detected mutations positioned on the chromosomes, I also filtered all the scaffolds in the file. I filtered out all non-protein-coding genes by keeping all protein-coding genes, which I downloaded using the BiomaRt package. For analyzing and visualizing detected somatic mutations in patient cancer genomes, I used the R package maftools. I modified the imported VCFs to appropriate the MAF object format using a custom R script.

3.6.1 Mutational landscape analysis

To see how detected mutations are distributed across the genome, I counted the number of all detected mutations on each chromosome of all analyzed patients. I normalized the counted mutations per chromosome by dividing the number of detected mutations by the calculated number of exon nucleotides in each chromosome. The results were plotted using the ggplot function from the ggplot2 package in R Studio.

Using the maftools package and its titv function, I classified the detected mutations into transitions and transversions. Furthermore, I used the plot function to visually represent the overall distribution of six different nucleotide conversions, the fraction of conversions in each patient, and the percentage of transitions and transversions across all detected SNPs in all samples (patients).

Oncoplots, alternatively known as waterfall plots, serve as an improved means of displaying the information contained in a MAF file. To generate oncoplots depicting the mutational landscape of my samples, I used the oncoplot function from the maftools package. To show only the top 10 most mutated genes, I used the argument top = 10, and to show the top 3 pathways with the most mutated genes, I used the argument pathways = TRUE.

I searched for the co-occurring and mutually exclusive set of genes affected by the mutations with the somaticInteractions function from the maftools package. This function performs a pair-wise Fisher's Exact test to detect such a pair of genes and shows the results graphically.

Using literature, I found the commonly mutated genes in PDAC. According to my research, the genes *KRAS, CDKN2A, TP53, SMAD4, BRCA1, BRCA2, ATM, PALB2,* and *BRAF* are highly connected to PDAC (Hu et al., 2021; Kamisawa et al., 2016; Maitra and

Hruban, 2008; Orth et al., 2019; Sarantis et al., 2020). I used the oncoplot function with the argument genes = c("KRAS", "CDKN2A", "TP53", "SMAD4", "BRCA1", "BRCA2", "ATM", "PALB2" and "BRAF") to draw oncoplot only for these genes. I inspected the generated mutational summary data to see how many mutations affected this group of genes and to compare them with the most mutated genes detected in patient samples. I performed an analysis of each gene found to be highly connected to PDAC. First, I extracted the information considering mutations for individual genes from a data frame containing all the mutations across all samples by using their Hugo symbols. I counted the classes of variants affecting each gene and plotted them on individual bar plots. I used the lollipopPlot function to visually represent the locations of detected variants on each gene and cause amino acid changes.

I wanted to compare mutation load in analyzed patient data against 33 TCGA cohorts from the MC3 project, and for this I used the tcgaCompare function included in the maftools package. Apart from that, I explored the differences and similarities between mutational data detected in my samples and mutational data in the pancreatic adenocarcinoma (PAAD) cohort from TCGA. To do this comparative analysis, I downloaded PAAD somatic mutation whole exome sequencing data from the ICGC data portal. I read the data into R Studio and filtered it by mutation type to keep only single-base substitutions. I counted the number of mutations on each chromosome (across all samples) and generated a bar plot to explore which chromosomes carried the most mutations.

Furthermore, I generated a maf object from PAAD cohort data and used the titv and plot functions to show percentages of transitions and transversions, the distribution of nucleotide conversions, and the portion of each nucleotide conversion in each sample. I generated the oncoplot by using the oncoplot function to showcase the most mutated genes and their mutation types across all samples. I also used the mafCompare function to detect differentially mutated genes between PAAD and my cohort. I wanted to include only genes that are mutated in at least nine samples in one of the cohorts to avoid bias due to genes mutated in a single sample. The results of the comparison were plotted using the coBarplot function. To see if there is a significant difference between the most mutated genes in my sample data and the PAAD cohort sample data, I carried out Chi-square test.

3.7 Mutational signature analysis

To expose and explore mutational signatures in patient data, I used the mutational.signatures.lib package. First, I created mutational catalogs from VCF files with identified somatic variants in patient sample sequencing data. These mutational catalogs refer to collections of distinct patterns of DNA mutations observed in cancer genomes.

I fitted known mutational signatures to mutation count data (contained in mutational catalogs) using the Fit and FitMS functions, which apply non-negative matrix factorization to decompose the mutation count matrix into a set of mutational signatures and their corresponding contributions to the mutational pattern.

The result of the mutational signature fitting process using the signature.tools.lib algorithm and its Fit and FitMS functions is the decomposition matrix $C \approx SE$, where C is the catalog matrix, with mutation types as rows and samples as columns; S is the signature matrix, with mutation types as rows and signatures as columns; and E is the exposure matrix, with signatures as rows and samples as columns (Degasperi et al., 2020, 2022). Vector e indicates how many mutations in C are associated with each of the k mutational signatures.

To compare original and reconstructed mutational catalogs, cosine similarity is calculated and displayed. This similarity score has a value between 0 and 1, and it serves as measure of reconstruction error. A high cosine similarity score between the original and reconstructed mutational catalog indicates that the signatures used for the fitting process can explain the original mutational pattern well.

I performed signature fit using organ-specific signatures with the Fit function. This function uses a one-step bootstrap (resampling with replacement) approach to calculate the empirical probability of an exposure being larger or equal to a given threshold (Degasperi et al., 2020, 2022). I extracted signatures specific for the pancreas using the getOrganSignatures function with the arguments "Pancreas", and typemut = "subs". I did the multi-step signature fit using the FitMS function. According to the specified organ, this function automatically selects the common signatures that will be used in the first step of the fitting process and the rare signatures whose presence will try to be determined (Degasperi et al., 2020, 2022). In both fitting procedures, I used the Gini-based exposure filter to compute the unique thresholds for every fitted mutational signature. For visual representation of fitting results, I used the plotFit and plotFitMS functions, respectively.

4 Results

4.1 The distribution of mutations across the chromosomes

The normalized number of mutations per chromosome detected in each patient's cancer RNA-seq data is depicted in the plot in Figure 6. Each boxplot provides a summary of the distribution of the normalized mutation count for each analyzed patient on a specific chromosome. The graph shows that all the chromosomes harbor mutations, although some of them are affected more than others. The median values of normalized mutation count per chromosome revolve between 0.00066 and 0.00025, while interquartile ranges are relatively wide, spanning from 0.0013 (chromosome 7) to 0.00016 (chromosome 17). The highest median value is detected for chromosome Y (0.0006558), followed by chromosomes 8, 4, and 20, while the lowest median value is detected for chromosome 9 (0.0002511), followed by chromosomes 22, 17, and 16. The outliers are not common, being detected in data for only four chromosomes.

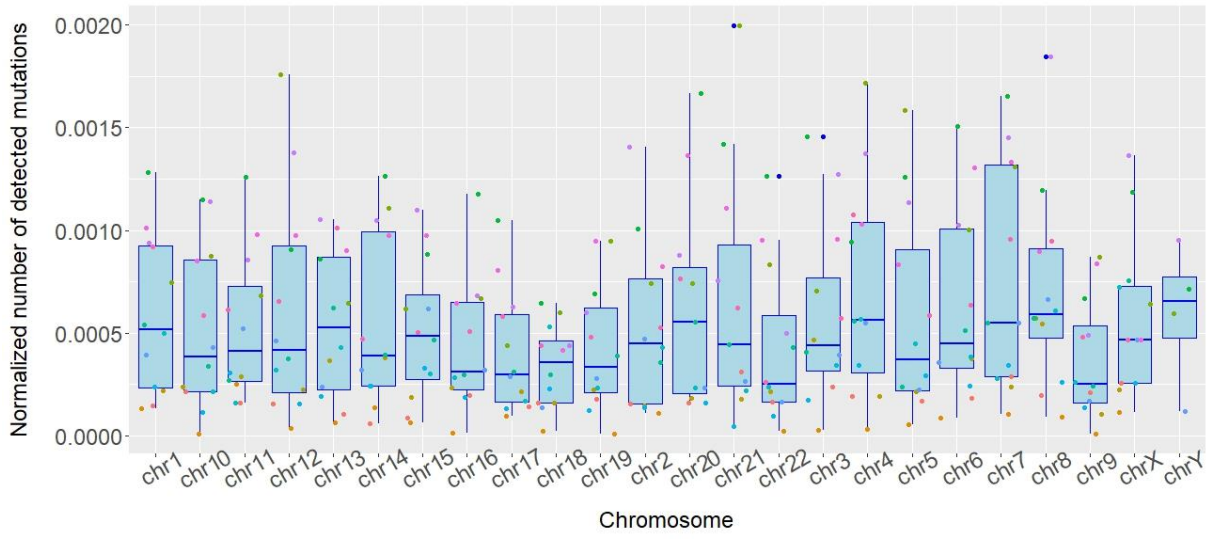

Figure 6. A box plot of the normalized mutation count per chromosome in the analyzed patient cohort data. Each boxplot represents a summary of the normalized number of mutations detected in each analyzed patient for a specified chromosome. The names of chromosomes are listed along the x-axis, while the y-axis represents the normalized number of detected mutations. The median value is represented by a thick blue line in each boxplot. The dots represent the normalized number of mutations detected in specific patient sequence data, while their color corresponds to the specific sample from the PDAC cohort.

4.2 Mutational landscape

The overall distribution of six different nucleotide conversions is shown with a box and whiskers plot in Figure 7. A. The most common nucleotide conversion detected is the T>C transition, with a median occurrence rate of 54.72%. The second most common nucleotide conversion is reverse mutation, that is, C>T transition, with a median value of 24.34%, while the rest of the four types of nucleotide conversions, transversions T>A, C>A, C>G, and T>G, are less common, with median values of occurrence of 5.65%, 5.92%, 4.92%, and 2.54%, respectively.

The ratio of mutations classified into transitions and transversions is shown with a box and whiskers plot on Figure 7. B. The percentage of transitions is significantly higher than that of transversion mutations. The median value for the percentage of transition mutations is 79.38%, while for the percentage of transversion mutations, it adds up to 20.62%.

The contribution of detected nucleotide conversions in each patient's sequencing data is shown with a stacked bar chart in Figure 7. C. Percentages of nucleotide conversions follow the overall distribution of six different nucleotide conversions depicted in Figure 7 for the most part, whereas in most patients sequencing data, the most numerous nucleotide conversion is a T>C transition, apart from Patient_2, Patient_8, and Patient_9, where the most nucleotide conversions come from C>T transitions. The percentages of transversions in all patients are small, while their contribution to each patient's mutational landscape does not follow any pattern but rather is different from one patient to another.

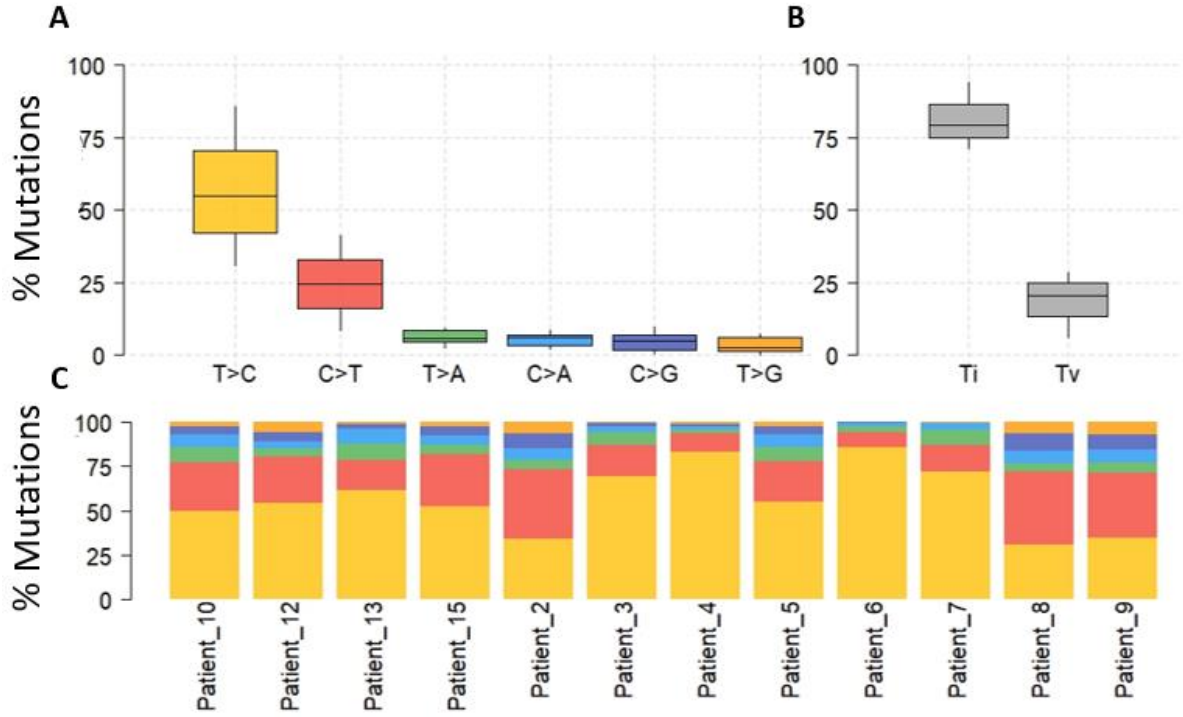

Figure 7. Transitions and transversions classification of mutations and nucleotide conversions contribution to the mutational profile. The percentage of contribution is shown on the y-axis of all three graphs, while six different nucleotide conversions are shown on the x-axis of A), transition and transversion classification on the x-axis of B), and analyzed patient samples on the x-axis of C). A) Box and whiskers plot depicting the overall distribution of six different nucleotide conversions detected in patient tumor genome sequence data. B) Box and whiskers plot showing the calculated contribution of transitions and transversions detected in the analyzed patient tumor sequence data. C) Stacked bar charts showing the contribution of detected nucleotide conversions in each patient's sequencing data.

An overview of the genomic alterations across patient samples is shown in Figure 8. with an oncoplot, also called a waterfall plot. The top 10 mutated genes in patient samples are listed along the y-axis of the plot. The gene with the most detected mutations is *XIAP*, which is altered in all the samples included in the analysis. It is hit by multiple mutations in most of the samples, apart from Patient_5, Patient_8, and Patient_12 samples, where only a 3' prime UTR mutation is detected. The next three genes, *CTNND1*, *CTSB*, and *MIR612* are altered in 92% of the samples, while the next six genes, *ACOX1*, *EIF2AK2*, *H2AZ2*, *METTL7A*, *SLC4A4*, and *SOD2*, are altered in 82% of the samples.

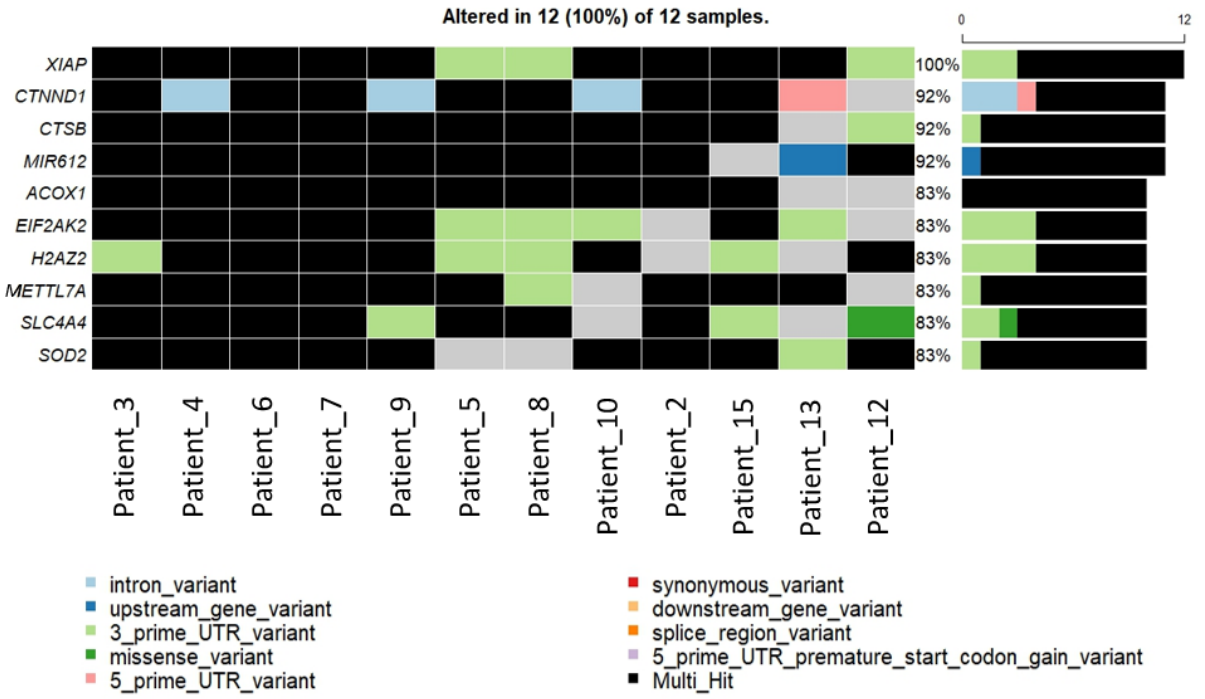

Figure 8. Oncoplot depicting the top 10 most mutated genes in the analyzed tumor genome sequence data. Mutated genes are listed along the y-axis, while sample names are listed along the x-axis. The mutation frequency and prevalence of variants in the corresponding gene across patient cohorts are shown on the right side of the matrix with a stacked bar chart.

The top 3 most mutated pathways, that is, the pathways whose genes accumulated the most mutations, are shown in Figure 9. with oncoplot. The biological pathways with the most mutated genes in the analyzed tumor genome sequence data are *Wnt/B-catenin signaling* and *Transcription factor* with alteration in 92% of patient samples. Mutated genes in the *Wnt/B-catenin signaling*, pathway are *CTNND1, CTNNB1, TCF7L2,* and *APC,* with mostly detected multiple mutations, alongside intron variants, a 5' prime UTR variant, and a 3' prime UTR variant. There are 17 detected mutated genes in the analyzed data that are involved in the pathway *Transcription factor*, which is altered in 92% of samples, and these are: *MECOM, ZBTB20, EPAS1, MAX, MYC, RUNX1, CBFB, ELF3, FUBP1, KLF5, MGA, NFE2L2, RXRA, TBX13, TCF12, ZFHX3,* and *ZMYM2*. The most common class of variants with these genes is multiple hit, followed by intron variant and 3' prime UTR variant, but other classes such as missense variant, 5' prime UTR variant, upstream gene variant, and downstream gene variant also appear. The third most mutated biological pathway is *Other signaling*, with mutations detected in 19 genes included in it. These genes are *FAT1, GNAQ, CDH1, LATS2, MAP3K1, PRKAR1A, RAC1, RHOA, ARHGAP35, DIAPH2, GNA11, GNA13, GNAS, MAP2K4, PLXXNB2, PTPDC1, PTPN11, RHOB,* and *SOS1*. Most detected variants are multiple hits and

3’ prime UTR variant, followed by missense variant, upstream gene variant, downstream gene variant, intron variant and synonymous variant.

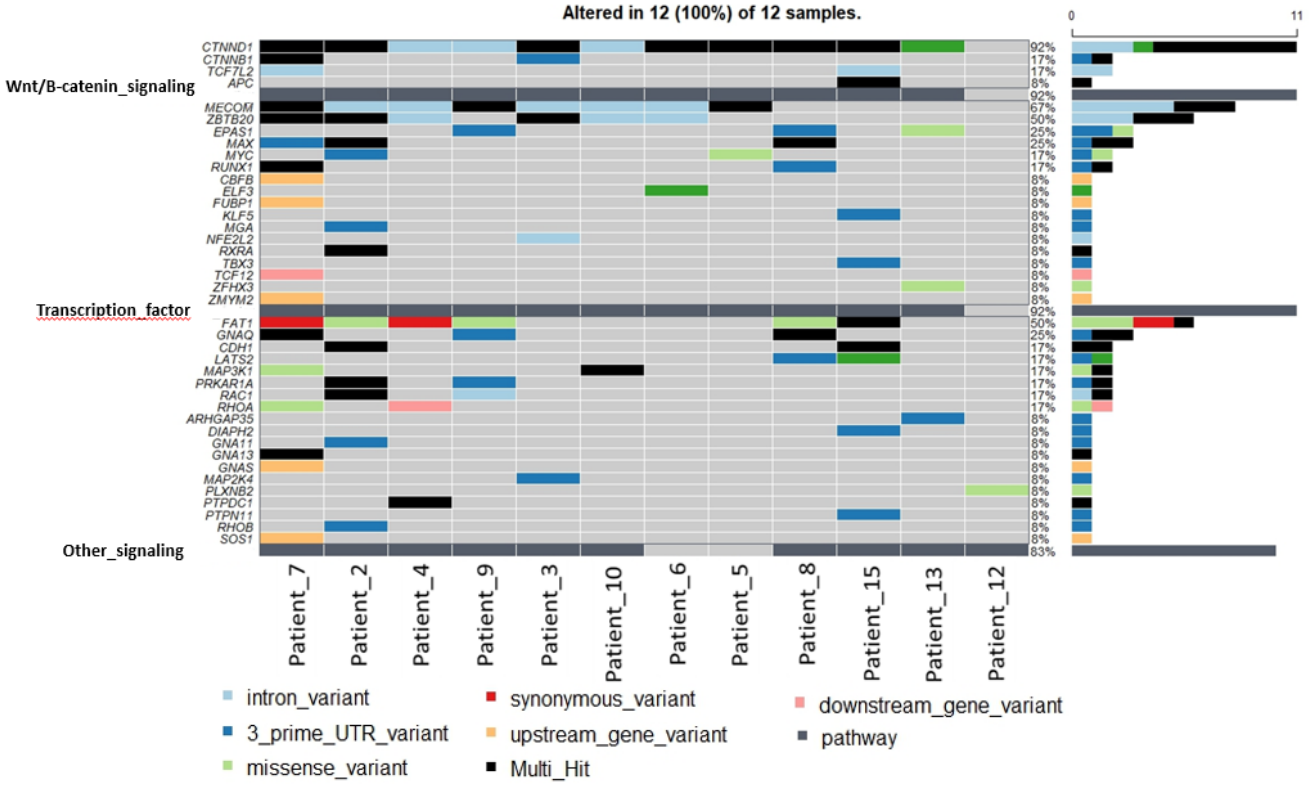

Figure 9. Oncoplot with the top 3 most mutated biological pathways with belonging genes in analyzed tumor genome RNA-seq data. The pathways are listed along the y-axis, while sample names are listed along the x-axis. The genes are grouped based on their biological processes above the pathways along the y-axis. Pathway mutation frequency in analyzed samples is shown on the right side of the matrix with a corresponding bar chart, while the mutation frequency and prevalence of variants in the corresponding gene across patient cohorts are also shown on the right side of the matrix, but with a stacked bar chart.

Somatic interactions, that is, co-occurrences of genomic alterations, in tumor genome sequencing data are visualized with a plot in Figure 10. Here, somatic interactions are shown for the top 20 genes. Co-occurrence refers to the simultaneous presence of two or more alterations in the same sample, while mutual exclusivity indicates that the alterations or events are rarely or never observed together in the same sample. Positive associations among the top 20 mutated genes in the analyzed samples can be seen for fifteen pairs of genes: *ACOX1* and *SLC35F5*, *ACOX1* and *KAT8*, *ACOX1* and *FNDC3B*, *ACOX1* and *AKAP13*, *EIF2AK2* and *SMIM14*, *EF2AK2* and *ELL2*, *H2AZ2* and *NIBAN1*, *METTL7A* and *SLC35F5*, *METTL7A* and *KAT8*, *METTL7A* and *HNRNPNC*, *METTL7A* and *ELL2*, *SLC4A4* and *SLC35F5*, *SLC4A4* and *KAT8*, *SLC4A4* and *AHR*, *KAT8* and *SLC35F5*. This means that Fisher’s exact test revealed

that the mutations of these genes tend to occur together in the same sample more frequently than expected by chance ($p < 0.05$). There are no two genes that demonstrate mutual exclusivity with statistical significance in any of the analyzed samples.

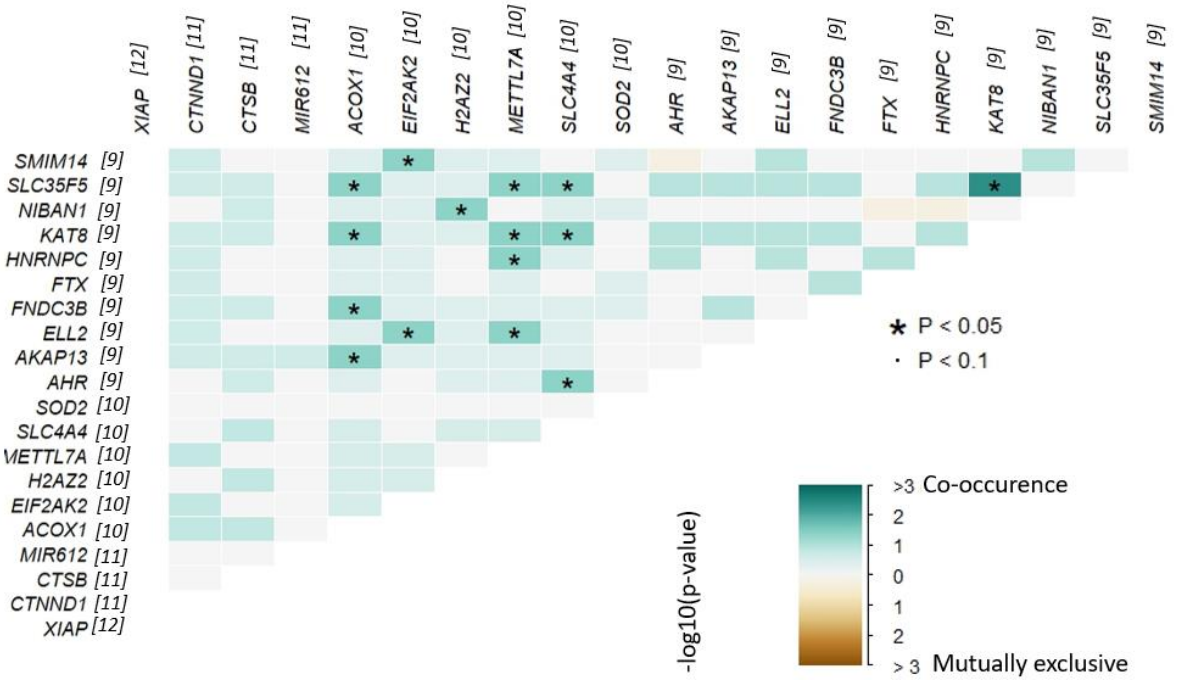

Figure 10. Plot depicting the somatic interactions of the top 20 mutated genes in the analyzed samples in an oncoplot-like grid; each row and column represents a specific mutated gene. Gene names are listed above the columns and on the left side of the rows. The number in square brackets represents the number of samples in which a particular genomic alteration was observed for that gene. The dot (".") and asterisk ("*") symbols represent the statistical significance of the co-occurrence or mutual exclusivity of a pair of mutated genes. The color of the square represents the calculated p-value with Fisher's exact test; green indicates a tendency toward co-occurrence, whereas brown indicates a tendency toward exclusiveness.

4.3 Commonly mutated genes in PDAC

The oncoplot for genes found to be highly connected to PDAC is shown in Figure 11., along with a stacked bar plot showing the proportion of six different nucleotide conversions in each patient's sequencing data. Only in eight patient samples was a mutation of one of these genes detected. The most mutated gene is *BRAF*, whose mutation classified as a 3' prime UTR variant is detected in Patient_7, Patient_8, and Patient_10 samples, while a mutation classified as a downstream gene variant is detected in Patient_15. Genes *KRAS* and *ATM* share second place with detected mutations in two samples. *KRAS* variants are detected in Patient_8 and Patient_9 samples, and both are classified as multiple hit mutations, while ATM

has detected variants in Patient_2 and Patient_6 samples classified as 3' prime UTR variant and missense variant respectively. *TP53* has detected mutation only in the Patient_4 classified as missense variant, while *SMAD4* has detected mutation only in the Patient_2 classified as synonymous variant. The other four genes, *BRCA1*, *BRCA2*, *CDKN2A*, and *PALB2*, have no detected mutations in any of the samples. In all samples, the most numerous nucleotide conversion is the T>C transition followed by the C>T transition, except in the Patient_8 sample, where the most numerous nucleotide conversion is the C>T transition followed by the T>C transition.

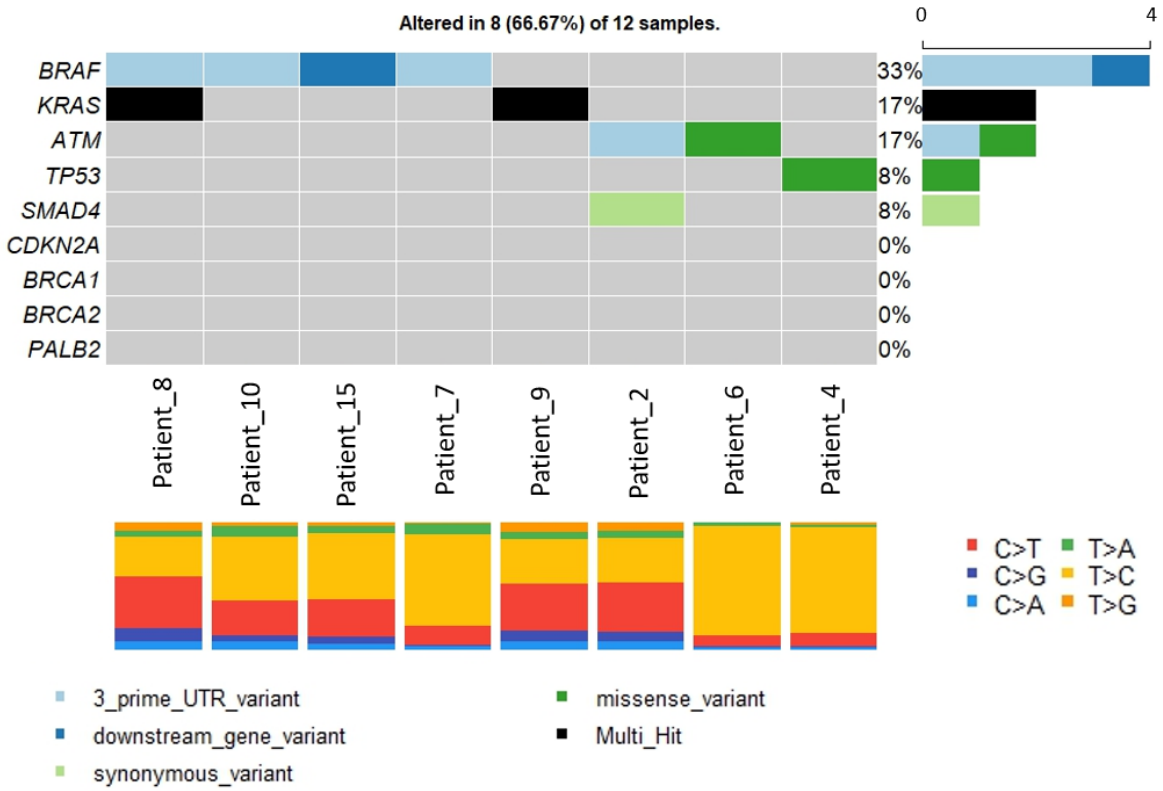

Figure 11. Oncoplot of genes highly connected to PDAC, along with stacked bar plots with contributions of detected nucleotide conversions in each sample of sequencing data. Genes are listed along the y-axis, while sample names are listed along the x-axis. The mutation frequency and prevalence of variants in the corresponding gene across patient cohorts are shown on the right side of the matrix with a stacked bar chart.

All the genes with detected mutations in the analyzed data were ranked by the number of samples in which alterations were found and ordered from the most mutated ones to the least mutated ones. Table 1. contains the ranks of five genes highly connected to PDAC with detected mutations in the analyzed data, together with the total number of mutations and the number of samples in which mutations of the specific gene were found. The highest-ranking

gene that is commonly mutated in PDAC is *BRAF*, with a total of 4 mutations in 4 mutated samples and a rank of 387. Gene *KRAS* received a rank of 886 with six detected mutations in two samples. The ranks of the remaining three genes are considerably low; *ATM* has a rank of 1271 with two detected mutations in two samples, while the ranks of *SMAD4* and *TP53* are 4126 and 4335, respectively, with only 1 detected mutation in 1 sample each.

Table 1. Mutation ranks of genes highly connected to PDAC retrieved from detected mutations in analyzed sample data.

Hugo symbol	Rank	Total mutations	Number of mutated samples
BRAF	387	4	4
KRAS	886	6	2
ATM	1271	2	2
SMAD4	4126	1	1
TP53	4335	1	1

Bar charts illustrating the count of classes of variants detected for genes *BRAF, KRAS, TP53, ATM,* and *SMAD4* are represented in Figure 12. The genes *CDKN2A, BRCA1, BRCA2,* and *PALB2* are excluded from this analysis, considering neither of them has any detected mutations in the analyzed patient tumor sequence data. The *BRAF* gene variant classification is depicted in Figure 12. A. Two mutations are classified as intron variants, four as downstream gene variants, and three as 3' prime UTR variants. There are four different classes of *KRAS* gene variants depicted in Figure 12. B. Eight mutations belong to the 3-prime UTR variant class, six to the non-coding transcript exon variant class, six to the 3' prime UTR variant class, and four to the downstream gene variant class, while the least populated class is the intron variant with one detected mutation. Detected variants of the *ATM* gene, divided by classes, are depicted in Figure 12 C. Classes 3' prime UTR variant and non-coding transcript exon variant have two assigned mutations, while classes upstream gene variant, missense variant, and downstream gene variant have one assigned mutation each. Variant classifications of the *TP53* gene are shown in Figure 12 D. The most numerous class of variants is the missense variant, with four assigned mutations, while downstream gene variants, intron variants, noncoding transcript exon variants, and upstream gene variants each have one assigned mutation. Gene *SMAD4* variant classes are shown in Figure 12 E. Mutations in these genes are assigned to four different classes: downstream gene variant, noncoding transcript exon variant, synonymous variant, and upstream gene variant, and each of them has only one assigned mutation.

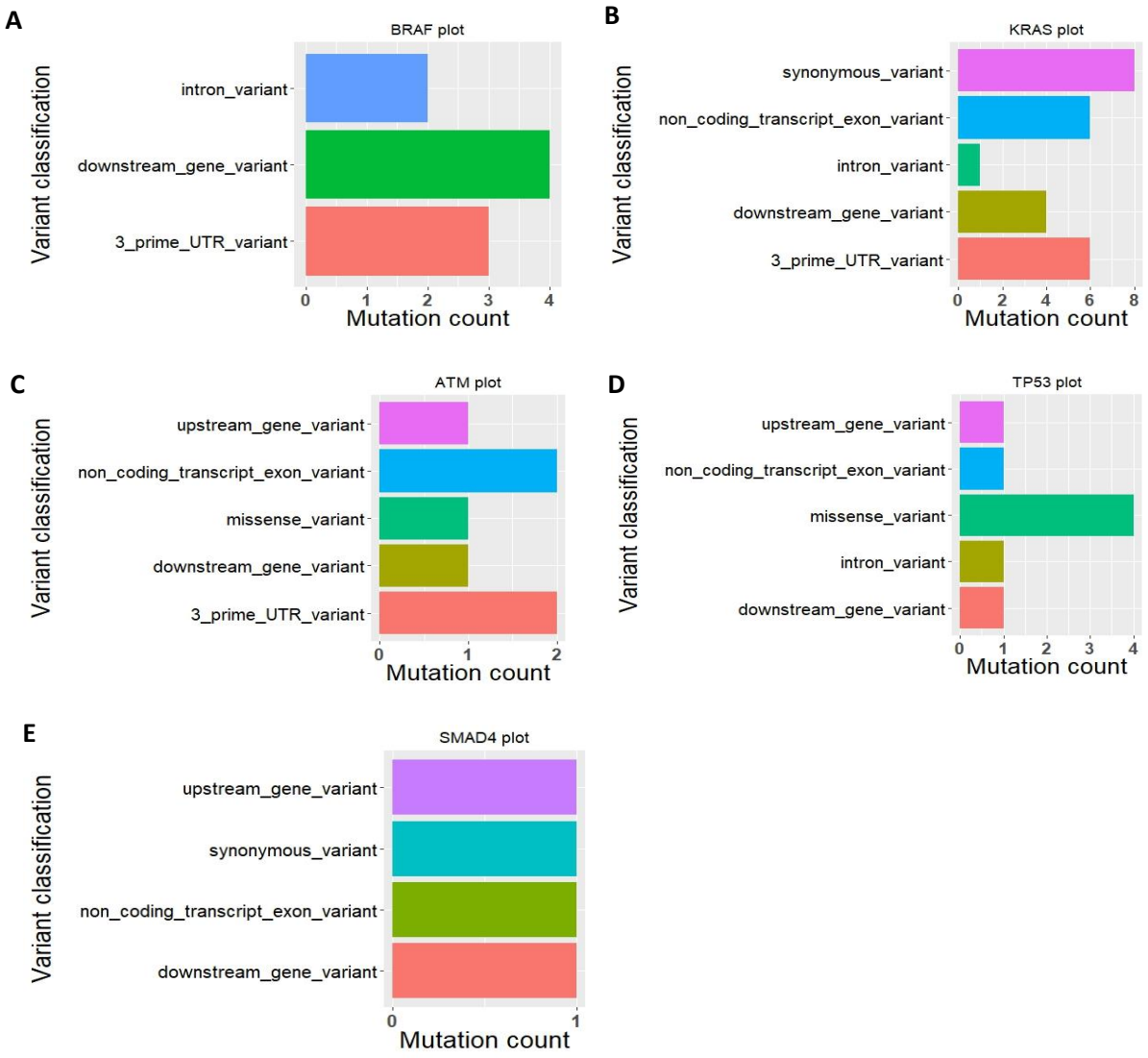

Figure 12. Barplots of counts of detected variant classes in commonly mutated genes in PDAC Variant classes are listed along the y-axis, while count numbers are listed along the x-axis. A) BRAF gene variant class count; B) KRAS gene variant class count; C) ATM gene variant class count; D) TP53 gene variant class count; E) SMAD4 gene variant class count.

I used the lollipopPlot function to visually represent the locations of detected mutations on each gene that caused amino acid changes. This plot displays the different types of mutations found within a gene and their corresponding positions along the gene sequence. In this graph, each gene is represented by a horizontal line consisting of gene domains drawn as rectangles in different colors, while each mutation along the gene is depicted as a tick mark with labeled amino acid changes. In Figure 13., only the lollipop plots *of KRAS, TP53, ATM,* and *SMAD4* are shown because only in these genes are there detected mutations in the analyzed samples. A lollipop plot for the *BRAF* gene couldn't be drawn because there are no annotated amino acid changes in the data set. Some mutations in drawn lollipop plots have no labeled amino acid change, and for that fact, these are not marked on the lollipop plot

corresponding to that gene. Figure 13. A represents the *KRAS* gene lollipop plot. The detected mutation rate is 16.67%, and a synonymous mutation in the COG1100 domain is marked on the 173rd nucleotide, where aspartic acid remains unchanged. On Figure 13. B, the *TP53* gene lollipop plot is shown. There is an 8.33% somatic mutation rate observed. A marked mutation in the P53 domain on the 270th nucleotide causes the conversion of phenylalanine into serine. The *ATM* gene lollipop plot is depicted in Figure 13. C. The detected mutation rate is 16.67%. The detected mutation lies in the FAT domain on the 2314th nucleotide and causes glutamine to change to leucine. The lollipop plot of the *SMAD4* gene is shown in Figure 13. D. The somatic mutation rate is 8.33%, and the detected mutation is located on the 202nd nucleotide, marked as synonymous, which means alanine remains unchanged.

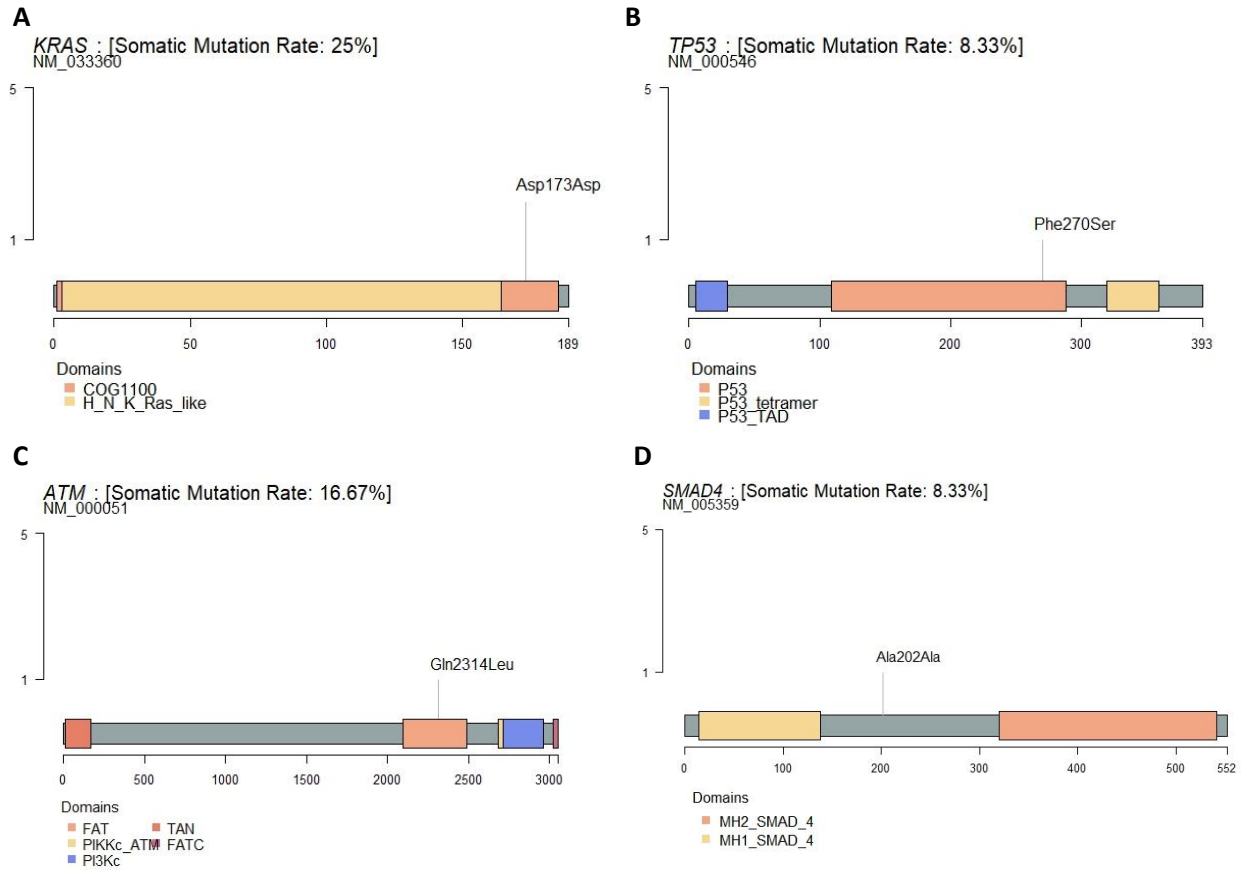

Figure 13. Lollipop plots of genes found to be commonly mutated in PDAC. Each graph consists of a gene depicted as a horizontal line, with colored rectangles representing gene domains. Variants are marked as a vertical line with labeled amino acid changes and nucleotide positions. The x-axis consists of numbers representing nucleotide positions in gene sequence, while the y-axis shows the number of detected variants. Next to a gene name in the title, there is a percentage of the somatic mutation rate written for that gene. A) KRAS lollipop plot; B) TP53 lollipop plot; C) ATM lollipop plot; D) SMAD4 lollipop plot

4.4 Comparison of the mutational landscape of RNA-seq called mutations in PDAC samples with mutational landscape of TCGA cohorts obtained by WXS

The comparison of detected mutational load between analyzed patient data and 33 TCGA cohorts from the MC3 project is depicted in Figure 14. The tumor mutational burden (TMB) is measured in the number of detected mutations per megabase of DNA. The mutational burden of all 33 TCGA cohorts seems to follow the student-t distribution, while the same cannot certainly be said for the samples analyzed in this paper considering there are only 12 samples. The median tumor mutational burden for analyzed patient data has a value of 15.9 mutations per megabase, which is rather higher than for any of the TCGA cohorts involved in the comparison. When looking at the distribution and median value for all 33 TCGA cohorts, the Skin Cutaneous Melanoma (SKCM) cohort shows the most similarity to the cohort of samples analyzed in this research.

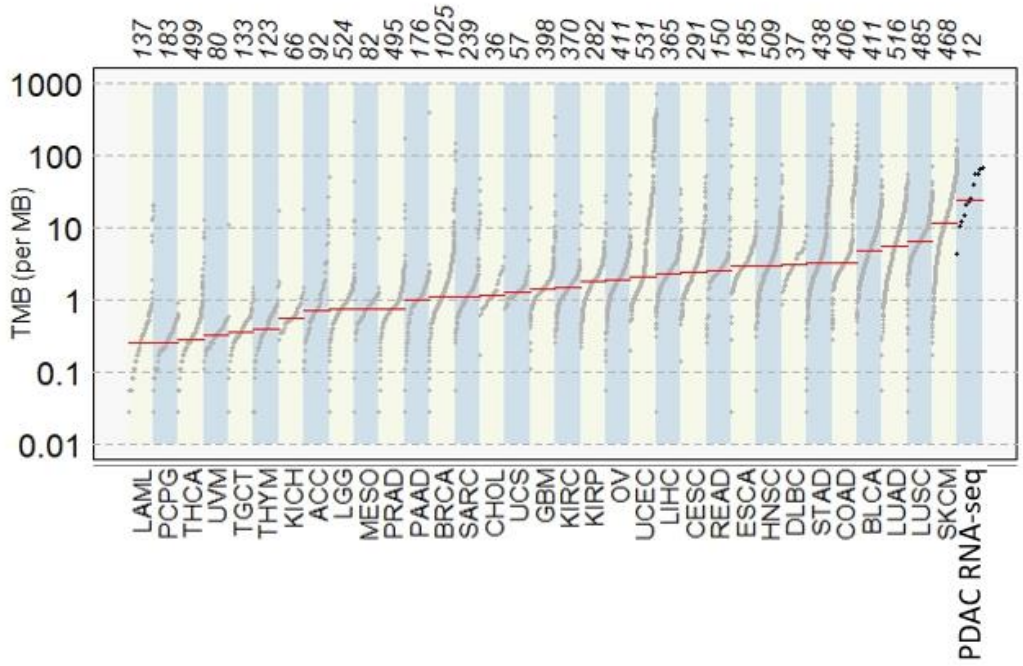

Figure 14. Scatter plot for comparison of mutational load between cohorts of analyzed patient data and 33 TCGA cohorts from the MC3 project. Mutational load is expressed as TMB (tumor mutational burden) in the form of the number of detected mutations per megabase in the analyzed DNA. TMB values are listed along the y-axis. The names of cohorts are listed along the x-axis from the downside, while the number of analyzed samples in each cohort is shown from the upper side of the graph. Each dot represents the TMB value for one sample in a cohort, while the red horizontal line represents the calculated median value of TMB in each cohort.

4.5 Comparison with the TCGA Pancreatic adenocarcinoma cohort

Since the analyzed data in this work comes from PDAC tumor samples, I used the PAAD cohort from the TCGA M3 project for comparison. Figure 15. depicts the normalized number of mutations per chromosome detected for each sample in the PAAD cohort. Each boxplot provides a summary of the distribution of the normalized mutation count for each analyzed patient on a specific chromosome. The mutations are detected on all chromosomes, but their accumulation is not equally distributed. Chromosome Y stands out among all others with a median normalized mutation count of 0.0005365, which is double the recorded median values of the other 23 chromosomes. The median values of normalized mutation count are appearing in the range between 0.00035 and 0.0001 when chromosome Y is excluded. Two samples from the PAAD cohort presented strong outliers for every chromosome with values of normalized mutation count from 0.03 to more than 0.15, which is why they were excluded from further analysis.

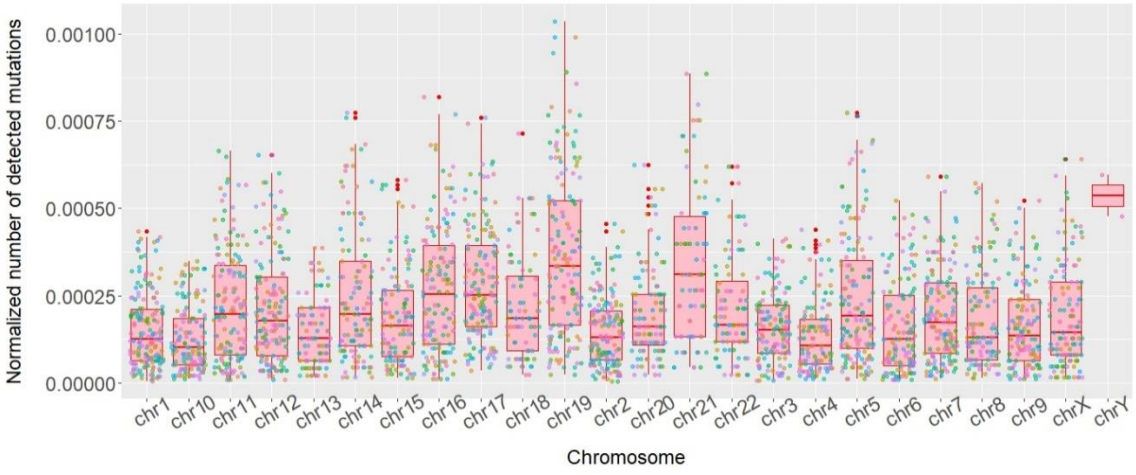

Figure 15. Box plot of the normalized mutation count per chromosome in PAAD cohort samples. Each boxplot represents a summary of the normalized number of mutations detected in each analyzed PAAD sample for a specified chromosome. The names of chromosomes are listed along the x-axis, while the y-axis represents the normalized number of detected mutations. The median value is represented by a thick blue line in each boxplot. The dots represent the normalized number of mutations detected in specific patient sequence data, while their color corresponds to the specific sample from the PAAD cohort.

The PAAD cohort transitions and transversions (TiTv) graph is depicted in Figure 15. A shows the overall distribution of six different nucleotide conversions with a box and whiskers plot. The most common nucleotide conversion detected is the C>T transition, with a

median value of 63.12% in the data set. This mutation is by far the most numerous, considering the other five mutations do not add up to even 35% of total mutations. The second most common nucleotide conversion is C>A transversion, with a median value of 11.51%, while the other four types of nucleotide conversions, T>C, C>G, T>A, and T>G, have less than 10% occurrence, with median values of 9.38%, 6.72%, 4.44%, and 2.6%, respectively.

The ratio of mutations classified into transitions and transversions is shown with a box and whiskers plot on Figure 15. B. The percentage of transitions is significantly higher than the percentage of transversion mutations. The median value for the percentage of transition mutations is 72.65%, while for the percentage of transversion mutations, it adds up to 27.35%.

The mutation contribution of detected nucleotide conversions in each patient's sequencing data is depicted in Figure 15. C with a stacked bar plot. The mutation with the most contribution to the mutation profile of almost all analyzed patient samples is the C>T transition. Only a few patient samples have the C>T transition or C>G transversion as the most common nucleotide conversion.

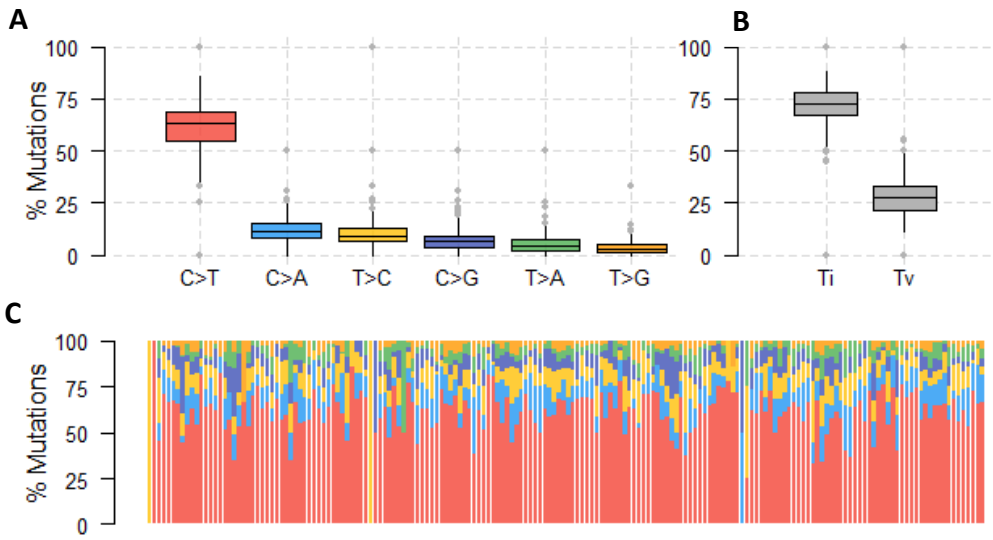

Figure 16. Transitions and transversions classification of mutations and nucleotide conversions contribution to the mutational profile. The percentage of contribution is shown on the y-axis of all three graphs, while six different nucleotide conversions are shown on the x-axis of A), transition and transversion classification on the x-axis of B), and analyzed patient samples on the x-axis of C). A) Box and whiskers plot showing the calculated contribution of each of the six different nucleotide conversions detected in the analyzed sample data. B) Box and whiskers plot showing the calculated contribution of transitions and transversions detected in the analyzed sample data. C) A stacked bar chart showing the contribution of detected nucleotide conversions in each patient's sequencing data.

The oncoplot for the top 10 mutated genes in the PAAD cohort, along with a stacked bar plot showing the proportion of six different nucleotide conversions in each patient's sequencing data, is shown in Figure 16. The genes that accumulated the most mutations are: *KRAS*, *TP53*, *TTN*, *MUC16*, *FLG*, *OBSCN*, *RYR1*, *DNAH11*, *MYO18B*, and *SCN5A*. The most mutated gene is *KRAS*, with a 62% mutation rate in analyzed samples, and most of them are classified as missense mutations, while a small amount is classified as a multiple hit mutation. In second place is the *TP53* gene, with alteration in 19% of analyzed samples, with mutations being classified as missense mutations for the most part, five as nonsense mutations, and one as frame insertion. 10% of the samples have detected mutations in the *TTN* gene, and the majority of them are classified as missense mutations, but there are also three detected nonsense mutations and one multi-hit mutation. Seven of the top ten genes have a mutation rate below 10% in the analyzed PAAD cohort. Most of the detected variants are classified as missense mutations, but there are also detected multi-hit mutations, splice site mutations, frame shift insertions, and nonsense mutations.

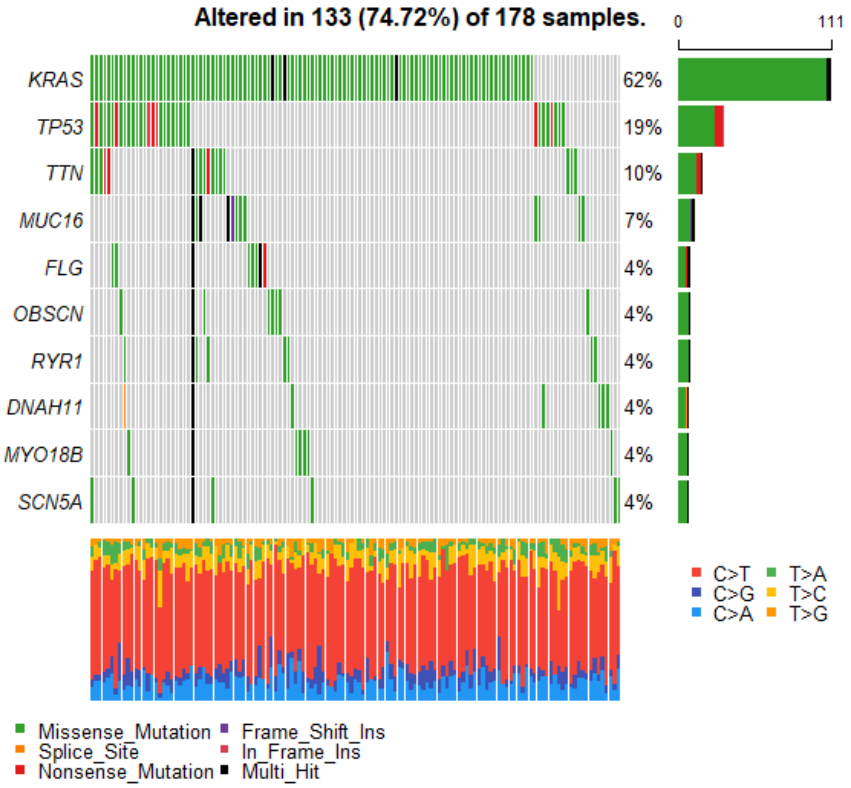

Figure 17. Oncoplot of the top 10 mutated genes in the PAAD cohort, along with a stacked bar plot with the contributions of detected nucleotide conversions in each sample of sequencing data Genes are listed along the y-axis, while sample names are listed along the x-axis. The mutation frequency and prevalence of variants in the corresponding gene across patient cohorts are shown on the right side of the matrix with a stacked bar chart.

The oncoplot made for the top 3 most mutated pathways is shown in Figure 17. The biological pathway with the most mutated genes in PAAD cohort data is *MAPK signaling*, with alterations in 62% of patient samples. Mutated genes involved in this pathway are *KRAS* and *BRAF*. All the samples contain mutations in the *KRAS* gene, with most of them classified as missense variants, while three are classified as multi-hit variants. One of the samples has detected a mutation in *BRAF* categorized as a missense variant, along with a missense variant of *KRAS*. The second most mutated pathway is *Genome integrity*, with 11 mutated genes: *TP53, ATM, ATR, BRCA1, BRCA2, CHEK2, ERCC2, PDS5BPOLQ, RFC1,* and *STAG2*. The most mutated gene in this pathway is *TP53*, with alterations in 19% of the samples classified as missense mutations or nonsense mutations. The other ten genes contribute much less to the mutation rate of the pathway, with four (*ATM*), two (*ATR* and *BRCA*), or just one (all others) detecting mutations in analyzed samples. Pathway *Other* takes third place among the most mutated pathways. Mutated genes involved in this pathway are: *APOB, SPATA1, CACNA1A, COL5A1, FLNA, KIF1A, CNBD1, DMD, GABRA6, GRIN2D, KEL, MUC6, MYH9,* and *SPTAN1*. Most of these mutations are classified as missense or multi-hit mutations, except for two nonsense mutations and one frame shift insertion. Recorded alterations in these genes contribute 3% or less to the overall mutation rate for the analyzed samples.

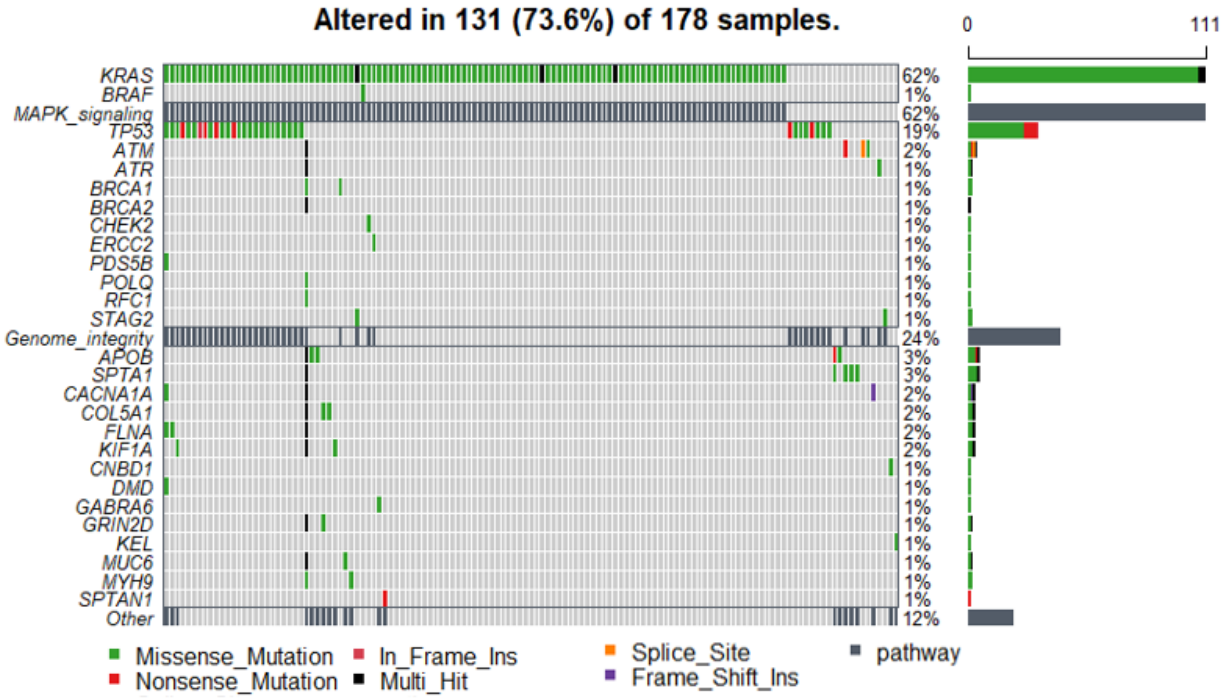

Figure 18. Oncoplot with the top 3 most mutated biological pathways with belonging genes in analyzed tumor genome sequence data. The pathways are listed along the y-axis, while sample names are listed along the x-axis. The genes are grouped based on their biological processes above the pathways along the y-axis. Pathway mutation frequency in analyzed samples is shown on the right side of the matrix with a corresponding bar chart, while the mutation frequency and prevalence of variants in the corresponding gene across patient cohorts are also shown on the right side of the matrix, but with a stacked bar chart.

The visualization of differences between mutational patterns in the PAAD cohort and the analyzed data cohort is represented by a co-bar plot in Figure 18. This graph shows the mutational frequencies of the most mutated genes in the PAAD cohort and in the cohort of analyzed PDAC patient samples researched in this study. Five out of ten displayed genes appear to be mutated in both cohorts: *CTNND1*, *KRAS*, *TPP3*, *TTN*, and *MUC16*, but their mutational frequencies are different for these two cohorts. The mutational frequency of *CTNND1* for PAAD is 1% compared to 92% for PDAC; the mutational frequency of *KRAS* is 62% for PAAD compared to 17% for PDAC cohort; the mutational frequency of *TP53* for PAAD is 19% compared to 8% for PDAC cohort; the mutational frequency of *TTN* is 10% for PAAD compared to 25% for PDAC cohort; and the mutational frequency of *MUC16* is 7% for PAAD compared to 8% for PDAC cohort. We found a statistically significant difference in proportions of mutated samples between PDAC and PAAD samples for the following genes:

XIAP, CTSB, MIR612, CTNND1, ACOX1, and *KRAS* (chi-square test, p-value < 0.05, p-values were adjusted for multiple comparisons using Bonferroni's method).

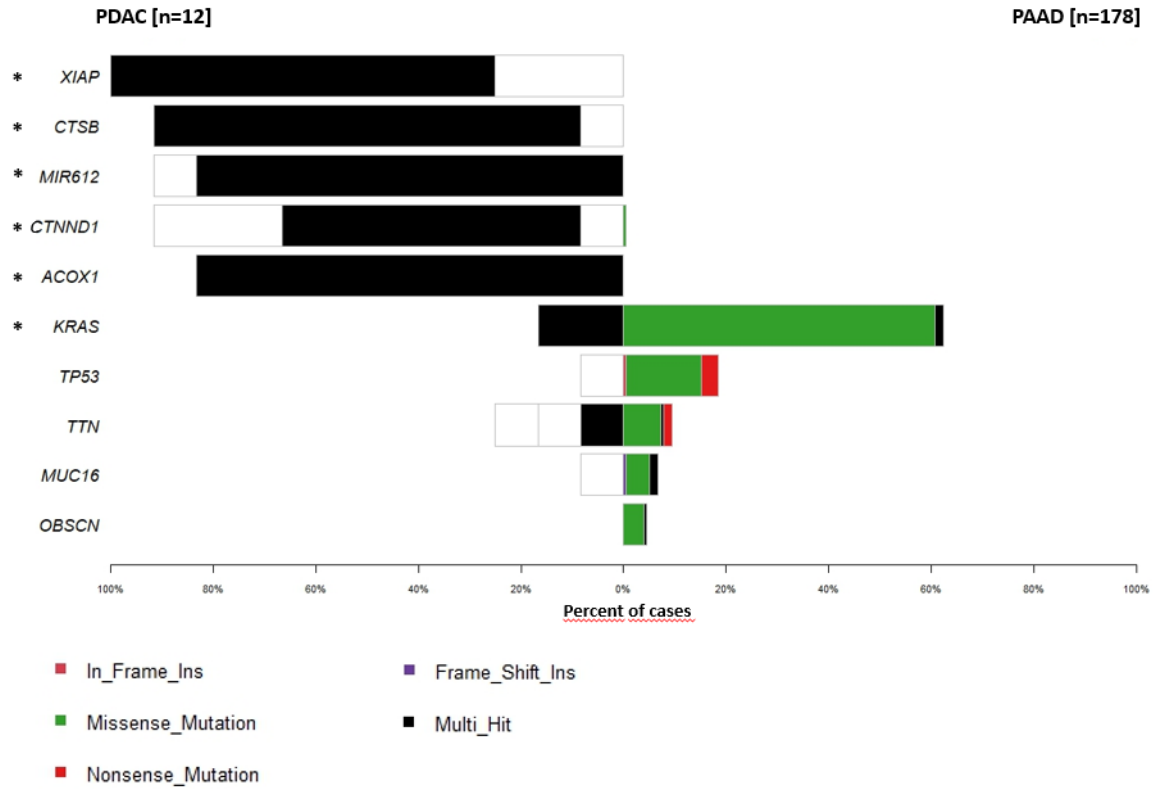

Figure 19. Co-barplot displaying mutational frequencies of the top 5 mutated genes in the PAAD cohort and cohort of analyzed patient samples in this research. Gene names are listed along the y-axis, while percentages describing mutational frequency are displayed along the y-axis for each cohort. The name of the cohort is displayed on the upper side, with cohort size in brackets. The size of the bar represents the detected mutational frequency for the specified gene in each of the two cohorts. Rectangles of different colors represent different classes of detected gene variants. Genes that showed a statistically significant difference in proportions of mutated genes between PDAC and PAAD cohorts are marked with an asterisk (*).

4.6 Mutational signatures fitting

The point estimate exposures calculated with the Fit function are visualized as proportions of total mutations and can be seen in Figure 19. The calculated cosine similarity between a given catalog and the corresponding catalog reconstructed using the signatures indicates that none of the sample catalogs can be well explained with the signatures used. The highest cosine similarity can be seen in the Patient_8 (0.89) and Patient_2 (0.88) sample catalogs, while the Patient_9 sample catalog is not so far behind with 0.87 calculated cosine similarity

but still below the threshold of 0.95, which would be considered a meaningful cosine similarity. All other sample catalogs show a cosine similarity less than 0.8, which cannot be considered when explaining mutational catalogs with this group of signatures. Every sample catalog except Patient_6 has at least twenty six unassigned mutations. Most of the mutations are assigned to Signature GEL-Pancreas common SBS1+15+18.

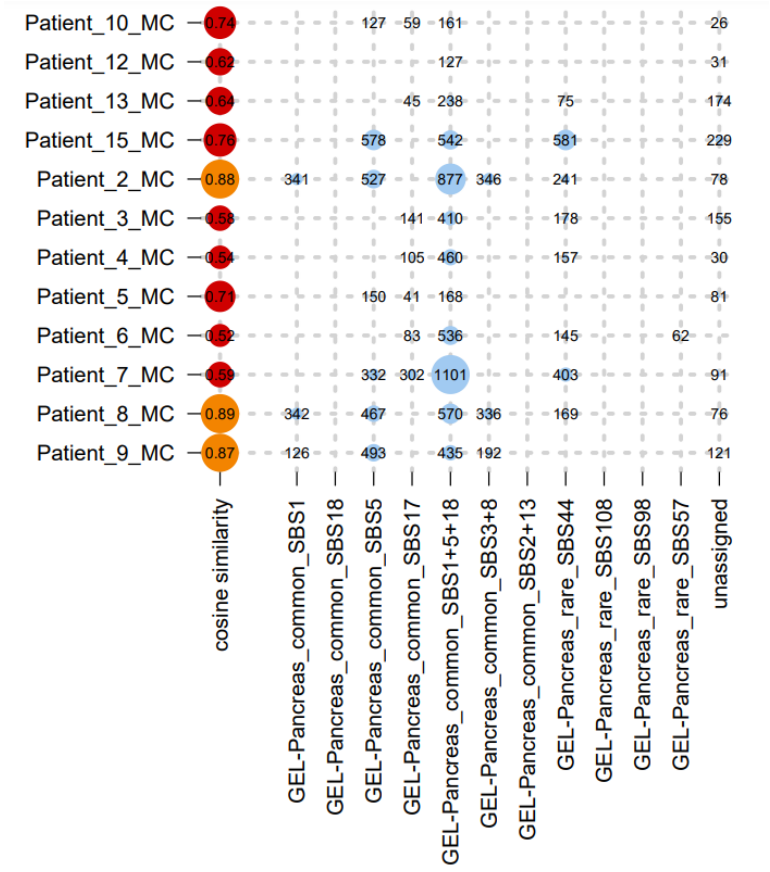

Figure 20. Proportions of total mutations from mutational catalogs assigned to mutational signatures used for the fitting process and calculated point estimate exposures. In this matrix-like visualization, sample names are listed vertically, while signatures used for fitting are listed horizontally. The cosine similarity scores between the original patient mutational catalog and the one reconstructed based on chosen signatures are listed in the first column. The size of the circle around the number of assigned mutations or calculated cosine similarity is proportional to the absolute value of the number it corresponds to. The color of the circle around the calculated cosine similarity represents how the point estimate of exposure corresponds to the thresholds computed with the Gini-based exposure filter.

The point estimates from the multi-step fitting process using the FitMS function are visually represented as proportions of total mutations in Figure 20. The calculated cosine similarity scores between the original mutational catalog and the reconstructed catalog indicate that seven out of twelve sample catalogs can be well explained using the best linear

combination of signatures chosen in this multi-step fitting process. The highest cosine similarity of 0.95 can be seen in Patient_2, followed by Patient_8 with a calculated cosine similarity of 0.94. Patient_5 and Patient_9 sample catalogs showed a 0.93 cosine similarity with their respective reconstructed catalogs, while the last sample catalogs whose calculated cosine similarity is considered meaningful are Patient_4, Patient_7, and Patient_15 catalogs with a recorded similarity of 0.90. The cosine similarity of the Patient_3, Patient_6, Patient_13, and Patient_10 sample catalogs is high (0.89, 0.88, 0.88, and 0.86, respectively), but not high enough to reach the threshold to be considered acceptable. The worst cosine similarity was detected for Patient_12's mutational catalog, which scored 0.62. This means that the difference between the reconstructed catalog and the mutational catalog is so big that this mutational catalog cannot be explained with chosen mutational signatures. Apart from the GEL-pancreas common signatures SBS1, SBS5, and SBS1+5+18, the rare signature SBS123 is also detected in analyzed samples, and it has most mutations assigned to it. Some of the mutations from the patient catalogs remained unassigned, most of them belonging to Patient_2, Patient_8 and Patient_13, while Patient_3, Patient_6, and Patient_7 have no unassigned mutations.

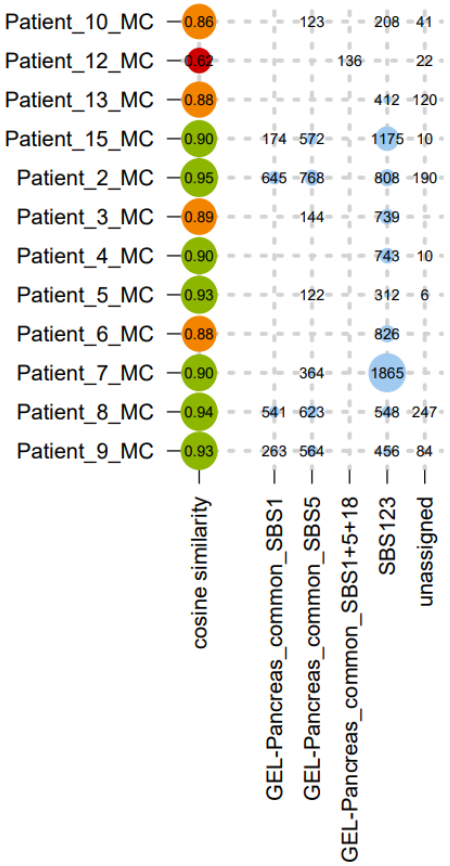

Figure 3. Proportions of total mutations from mutational catalogs assigned to mutational signatures used for multiple fitting process and calculated point estimate exposures. In this matrix-like visualization, sample names are listed vertically, while signatures used for fitting are listed horizontally. The size of the circle around the number of assigned mutations or calculated cosine similarity is proportional to the absolute value of the number it corresponds to. The color of the circle around the calculated cosine similarity represents how the point estimate of exposure corresponds to the thresholds computed with the Gini-based exposure filter.

5 Discussion

Analyzed cohort samples are firstly characterized by the number of detected mutations per chromosome to explore the distribution of mutations across the analyzed genomes and possibly detect chromosomes with a high accumulation of mutations. When looking at the normalized count of mutations per chromosome, it can be seen that mutations are well distributed across all 24 chromosomes of the analyzed patients. The highest mutation count median value is recorded for chromosome Y, which is the smallest and contains the least number of nucleotides in its exons, but only two samples in the whole PAAD cohort carry mutations on this chromosome. This kind of mutation count across the genome suggests the advanced stage of pancreatic cancer in patients whose samples were analyzed in this study. The overall pattern or profile of genetic mutations present in tumor cells is called the mutational landscape, and it provides a comprehensive view of the genetic alterations that have occurred in the tumor cells (Moore et al., 2021). The landscape can provide important insights into the underlying molecular mechanisms driving tumor growth, metastasis, and response to therapy. The mutational landscape of the analyzed tumor sample data cohort is consistent across all 12 patient samples, with mutations mainly being classified as transitions, most of them being T>C nucleotide conversions, with reverse conversion (C>T) taking second place. This kind of profile could be the result of DNA replication or damage and repair processes such as oxidative damage or exposure to reactive chemicals formed during tumorigenesis, as well as exposure to certain environmental factors either before or during tumorigenesis (Harris, 2013).

The analysis of detected mutations in analyzed patient tumor samples determined *XIAP, CTNND1, CTSB, MIR612, ACOX1, EIF2AK2, H2AZ2, METTL7A, SLC4A4,* and *SOD2* as the top 10 mutated genes. Neither of them was found to be listed among commonly mutated genes in previous PDAC tumor studies. The most mutated gene in analyzed patient tumor samples is *XIAP,* with alterations in all the analyzed samples, and it was mostly hit by multiple mutations, except in three samples where the 3' prime UTR variant was the only detected mutation. *The XIAP* gene encodes a multi-functional protein belonging to a family of apoptotic suppressors that regulates not only caspases and apoptosis but also modulates inflammatory signaling and immunity, copper homeostasis, mitogenic kinase signaling, cell proliferation, as well as cell invasion and metastasis (Tu and Costa, 2020). This protein also acts as an important regulator of innate immune signaling via regulation of Nod-like receptors

(NLRs) (GeneCards - Human Genes, 2023; Tu & Costa, 2020). The overexpression of *XIAP* is known to cause resistance to apoptosis and is well documented in pancreatic cancer studies. (Li et al., 2013; Vogler et al., 2009). It was also shown that elevated levels of *XIAP* expression posed a risk factor for the development of pancreatic cancer and served as an indicator for predicting the prognosis of post-operative pancreatic cancer patients (Li et al., 2013). *MIR612,* along with *CTNND1* and *CTSB,* holds second place for the most mutated gene, with alteration in 92% of the samples. *MIR612* (MicroRNA 612) is an RNA gene and is affiliated with the miRNA class. It is involved in a post-transcriptional gene silencing pathway where regulatory microRNAs (miRNAs) are responsible for silencing specific target genes (GeneCards - Human Genes, 2023). Due to their regulatory role in tumorigenesis, miRNAs have been used as therapeutic methods and diagnostic biomarkers, and miR-612 has demonstrated tumor-suppressive activity in various cancer types (Javadrashid et al., 2021). It is important to note that Javadrashid et al. demonstrated in the study from 2021 on pancreatic cells *in vitro* that miR-612 has the potential to be a focal point for therapeutic intervention in pancreatic cancer when combined treatment strategies are used. The *CTNND1* and *CTSB* genes are both protein-coding genes with different roles. *CTSB* encodes a member of the C1 family of peptidases that possesses both endopeptidase and exopeptidase activities and is classified as a lysosomal cysteine protease involved in protein turnover processes within the cell (GeneCards - Human Genes, 2023; Mort, 2013). Its main physiological function is to maintain the stability of the intracellular proteome by degrading different proteins in the lysosomes, but it also has a significant role in different intracellular signaling pathways, including cell proliferation, migration, autophagy, antigen presentation, and apoptosis (Ma et al., 2022; Mort, 2013). The research from 2021 done by Fujimoto et al. demonstrated high expression of *CTSB* in pancreatic cancer stem-like cells and also suggested that its expression in surgically removed tumor samples correlated with unfavorable outcomes after surgery (Fujimoto et al., 2021). The *CTNND1* gene, which is also found to be mutated in 92% of analyzed patient tumor samples, encodes a member of the Armadillo protein family with the main function of regulating cell-cell adhesion through the surface stability of C-, E-, and N-cadherins (Alharatani et al., 2020). Furthermore, it also regulates gene transcription through several transcription factors (GeneCards - Human Genes, 2023). It has been shown that pancreatic cancer patients with high expression of *CTNND1* have a poor prognosis because its encoded protein is one of the key molecules involved in pancreatic cancer metastasis (Huang et al., 2023). On top of that, it was also shown that downregulation of *CTNND1* expression in PDAC leads to a poor prognosis (Huang et al., 2023). The remaining six genes are altered in

82% of the analyzed samples. *ACOX1* and *EIF2AK2* both encode enzymes. The first enzyme of the fatty acid beta-oxidation pathway is encoded by *ACOX1,* and in some studies, its overexpression is linked to pancreatic cancer development (GeneCards - Human Genes, 2023; Nowara & Huszno, 2016). The protein encoded by *EIF2AK2* (Eukaryotic Translation Initiation Factor 2 Alpha Kinase 2) is a serine/threonine protein kinase that becomes activated by autophosphorylation upon binding to double-stranded RNA (dsRNA) and plays an important role in the innate immune response against multiple DNA and RNA viruses (Kuipers et al., 2021). This enzyme is also involved in the regulation of signal transduction, apoptosis, cell proliferation, and differentiation by phosphorylating other substrates, so it is not a surprise that its overexpression is thought to contribute to tumor development (GeneCards - Human Genes, 2023; Kuipers et al., 2021; Wang et al., 2022). *H2AZ2* (H2A.Z Variant Histone 2) is a protein-coding gene for replication-independent histone protein variants found in different organisms. It belongs to the H2A family of histones involved in various pathways, including RNA polymerase, promoter opening, and packaging of telomere ends (Ávila-López et al., 2021; GeneCards - Human Genes, 2023). It has been shown that this histone isoform is highly expressed in PDAC patients, leading to tumor growth and chemoresistance (Ávila-López et al., 2021; Salmerón-Bárcenas et al., 2021). Protein encoded by *METTL7A* (also known as *TMT1A*, or Thiol Methyltransferase 1A) belongs to the methyltransferase-like protein family and can be found in the endoplasmic reticulum, where it plays a role in lipid metabolism and recruits cellular proteins for the assembly of functional organelles (Liu et al., 2023; Zehmer et al., 2008, 2009). The exact function of *METTL7A* is not fully understood, but it is believed to play a role in cellular processes such as protein methylation and post-translational modifications, as well as cell development, migration, and drug resistance (Liu et al., 2023). Its altered expression has been observed in several types of cancer, including breast cancer, ovarian cancer, colorectal cancer, and hepatocellular carcinoma, implying its potential as a molecular marker for the diagnosis of tumors (GeneCards - Human Genes, 2023; Liu et al., 2023). *SLC4A4* is also a protein-coding gene; it encodes for electrogenic sodium bicarbonate cotransporter 1 (NBCe1), whose role is to control intracellular pH and regulate bicarbonate absorption and secretion (GeneCards - Human Genes, 2023). Research revealed that *SLC4A4* is the most expressed bicarbonate transporter in PDAC, and it was demonstrated that it plays an important role in regulating extracellular pH levels throughout the progression of PDAC (Cappellesso et al., 2022). *SOD2* is a protein-coding gene and a member of the iron/manganese superoxide dismutase family. It encodes a mitochondrial protein working as a homotetramer that binds one manganese ion per

subunit (Alateyah et al., 2022). Its main role lies in the detoxification of mitochondrial reactive oxygen species (mtROS) (Alateyah et al., 2022; Infantino et al., 2023). ROS cause DNA damage, leading to genetic mutations and genomic instability and promoting carcinogenesis (Alateyah et al., 2022). Consequently, any changes in the expression or activity of *SOD2* inevitably impact mitochondrial function, potentially contributing to the emergence of numerous diseases (Infantino et al., 2023; Steven et al., 2019). Diverse research studies have provided evidence that *SOD2* has the capacity to function as a tumor suppressor gene but also as an oncogene (Alateyah et al., 2022; Kim et al., 2017). The expression of *SOD2* is found to be elevated in several tumors, including PDAC (Nie et al., 2022). The results of the 2022 study by Nie et al. suggest that overexpression of SOD1 inhibits mitochondrial ROS-dependent cell apoptosis and endorses the proliferation of pancreatic cancer cells. On the other hand, several studies also documented the downregulation of the *SOD2* gene in pancreatic cancer as a consequence of epigenetic silencing and microRNA activation (mi-R301a) (Alateyah et al., 2022; Pandit et al., 2015).

When analyzing the somatic interactions of mutated genes, eleven pairs of genes are found to be mutated in the same sample more frequently than expected by chance ($p < 0.05$). That means positive somatic interactions are recorded for these pairs of mutated genes in the analyzed sample data. These pairs of genes are: *ACOX1* and *SLC35F5, ACOX1* and *FNDC3B, ACOX1* and *AKAP13, EIF2AK2* and *SMIM14, EF2AK2* and *ELL2, H2AZ2* and *NIBAN1, METTL7A* and *SLC35F5, METTL7A* and *HNRNPNC, METTL7A* and *ELL2, SLC4A4* and *SLC35F5, SLC4A4* and *AHR*. Considering the size of the analyzed cohort, the co-occurrence of these genes needs to be assessed carefully. The cooccurrence of two genes does not imply causation or involvement in the same biological pathways. None of these pairs of genes seems to have connected functions in a cell, but that does not necessarily mean that their respective pathways are not intertwined. The sheer complexity and interconnection of cellular pathways need to be taken seriously when analyzing the cooccurrence of mutated genes because it is possible that a cooccurring pair of genes is not connected in any sort of way, and that connection exists but has not yet been documented in any research study. It would be useful to do a bioinformatic analysis of the biological pathways in which these genes are involved and explore their possible connections.

The analysis of the mutational landscape of acquired patient tumor sample data also determined which biological pathways accumulated the most mutations. Three pathways that were altered the most by variations are *Wnt/β-catenin signaling*, *Transcription factor* and

Other signaling. The *Wnt/β-catenin signaling* pathway has essential functions in the maintenance of somatic stem cells development in embryos and maintaining the homeostasis of adult tissues (Liu et al., 2022; Makena et al., 2019). This pathway is involved in pancreatic cancer development because it has a role in overseeing processes such as cell cycle progression, apoptosis, epithelial-mesenchymal transition (EMT), angiogenesis, and many more, but it was also demonstrated that when disrupted, it can contribute to drug resistance (Makena et al., 2019). Emerging evidence suggests a potential connection between the *Wnt/β-catenin signaling* pathway and PDAC pathogenesis by exerting diverse influences on cellular proliferation, survival, differentiation, stemness, and the tumor immune microenvironment (Aguilera and Dawson, 2021). Aberrant activation of the pathway, often resulting from mutations in key components such as *APC*, *CTNNB1* (β-catenin), or other upstream regulators, leads to the accumulation and nuclear translocation of β-catenin, promoting the transcription of target genes involved in cell proliferation, invasion, and metastasis (Liu et al., 2022). Mutations in *CTNNB1* as well as in *CTNND1* were detected in the analyzed samples; *CTNND1* was altered in 92% of the samples, while *CTNNB1* was altered in 19% of the analyzed samples. It is important to note that microRNAs greatly participate in the regulation of *Wnt/β-catenin signaling* (Liu et al., 2022) and the *MIR612* gene was found to be altered in eleven out of twelve analyzed samples. Moreover, crosstalk between the *Wnt/β-catenin signaling* pathway and some other signaling cascades, such as Hedgehog and Notch, further contributes to PDAC development (Liu et al., 2022; Xia et al., 2022) The biological pathway that is altered in the same number of samples as the *Wnt/β-catenin signaling* pathway (92%) is the *Transcription factor* pathway which encapsulates all the proteins acting as key regulators of intrinsic cellular processes, such as differentiation and development, as well as orchestrators of cellular responses to external disturbances via signaling pathways (Weidemüller et al., 2021). Since transcription factors influence the creation of proteins (which themselves can act as signaling molecules), there are a wide variety of pathways and proteins that fall under this area (Weidemüller et al., 2021). Considering the nature of tumors as a pathogenic disease, it is expected that with their progression, more transcriptional factors will accumulate mutations that alter their activity and enable the progression of tumorigenesis. Apart from transcriptional factors, it is not surprising that the third most mutated pathway is the *Other signaling* containing genes that encode proteins that receive signals from the cell environment and carry them further down the pathway chain.

Although the most mutated genes found in the patient tumor cohort may not be directly linked to the formation of PDAC, they play important roles in the regulation of the cell cycle, cell communication and signaling, metabolism, and overall homeostasis. These roles are greatly devastated by multiple mutations causing the break of the cell cycle, inhibition of signaling pathways, and further progression of tumorigenesis. Even though these genes do not belong to a group of genes commonly found to be mutated in PDAC samples, this study of PDAC implicates that continuous mutational pressure caused by the formation of PDAC causes additional mutations in this set of genes. The documented destruction of important biological pathways used for signaling and regulating the cell cycle is expected in these stages of tumorigenesis. Further analysis should also include results of variant calling with WXS data and WGS data from the same samples to find the cause of the accumulation of mutations in this set of genes and try to connect them with the driver mutations causing the PDAC.

Analysis conducted on the group of genes found to be commonly mutated in previous studies of PDAC showed that these genes accumulated a small number of mutations in the analyzed PDAC sequence data. When all protein-coding genes with detected mutations are ranked and ordered by the number of samples in which they carry mutations, the group of commonly mutated genes in PDAC is far from being the most mutated, which was not expected. The highest ranked of them is the *BRAF* gene, placed in 310th place with four detected mutations in four samples, all of them being 3' prime UTR variants, followed by *KRAS* in 736th place with six detected mutations in two samples, all of them being classified as multi-hit mutations. Considering *KRAS* encodes a small GTPase oncogenic protein that plays an important role in the regulation of cell proliferation and, when mutated, promotes oncogenic events by inducing transcriptional silencing of tumor suppressor genes (GeneCards – Human Genes, 2023; Hajdúch et al., 2010), it is expected to be ranked much higher on this list. The same thing can be said for the *BRAF* gene, which encodes a protein of the same name belonging to the RAF family of serine/threonine protein kinases (GeneCards - Human Genes, 2023; Hussain et al., 2015). This protein plays a role in regulating the *MAP kinase/ERK signaling* pathway, which affects cell division, differentiation, secretion, and apoptosis (Hussain et al., 2015). *BRAF* is a known oncogene that, when mutated, is continuously active and transmits messages to the nucleus that encourage cell growth and division, even in the absence of these chemical signals (Hussain et al., 2015). The lollipop plot for *BRAF* couldn't be drawn because there were no detected amino acid changes in the dataset. Lollipop plot representing *KRAS* and its detected variants showed only synonymous mutation that wasn't

present in any previous research because it should not be relevant for protein function. *ATM*, *SMAD4*, and *TP53* genes are not even in the top 1000 mutated genes, which is particularly curious when *TP53* is considered to be the most common mutated gene across all cancer types (Guimaraes and Hainaut, 2002). *TP53* shows alteration only in one sample, and it is classified as a missense variant. This gene has been nicknamed the "guardian of the genome" because it encodes a tumor suppressor protein p53 essential for regulating DNA repair and cell division in response to a wide range of cellular stresses and regulates the expression of target genes (GeneCards - Human Genes, 2023; Guimaraes & Hainaut, 2002). This regulation leads to diverse cellular outcomes such as cell cycle arrest, apoptosis, senescence, DNA repair, or alterations in metabolism (Guimaraes and Hainaut, 2002). Even though it is not among the most mutated genes, the lollipop plot of *TP53* shows that the detected missense mutation causing the conversion of phenylalanine into serine is located in the DNA-binding domain (NM000546). This missense mutation is classified as pathogenic in the ClinVar database, and it is a commonly observed single nucleotide variation in TP53 that has lost tumor suppressor function (ClinVar, 2023; Landrum et al., 2018). Alterations in the *ATM* gene have been detected in two samples, one being classified as a missense variant and the other as a 3' prime UTR variant. The *ATM* gene encodes for a protein belonging to the PI3/PI4-kinase family that is thought to be the master controller of cell cycle checkpoint signaling pathways that are required for cell response to DNA damage and for genome stability (GeneCards - Human Genes, 2023; Khanna, 2000). The lollipop plot of *ATM* shows a missense variant in the FAT domain necessary for *ATM* dimerization, causing glutamine to change to leucine, which could affect its function and subsequently cause disturbance in the cell cycle. The *SMAD4* protein encoded by the *SMAD4* gene is part of the TGF-β signaling pathway and functions in dual roles as both a transcription factor and a suppressor of tumor growth (GeneCards - Human Genes, 2023; Zhao et al., 2018). The detected synonymous mutation is shown on the lollipop plot, and considering it is located in the non-active domain and there is no amino acid change, its function cannot be considered disturbed.

When the analyzed cohort is compared to the PAAD cohort from the TCGA database, there are much more dissimilarities than expected. Looking at the mutational landscape, transitions are the predominant type of mutation in both cohorts, but in the PAAD cohort, C>T nucleotide conversion is by far the most common one, followed by C>A transversion. This was expected considering the C>T transition occurs often, not only during tumorigenesis but also as a spontaneous one (Rünger, 2008). The most common detected nucleotide

conversion in the analyzed PDAC cohort (T>C) takes third place in the PAAD cohort, with more than 50% fewer detected mutations than the C>T transition. The normalized number of detected mutations per chromosome in TCGA cohort samples shows again that mutations are widespread across all chromosomes, as is expected in advanced stages of cancer. When looking at the median values of normalized mutation count as well as interquartile ranges, it should be noted that these values are significantly smaller than the same values calculated for the PDAC cohort. That could mean PDAC samples accumulated more mutations during tumorigenesis than samples in the PAAD cohort, or that more mutations were detected with RNA-seq data in this pipeline than WXS data from the PAAD cohort, but it could also just be the result of the fewer samples in the PDAC cohort. These assumptions should be further investigated by calling mutations from the WXS sequencing data of PDAC samples and comparing the results. One other thing that should be mentioned is that only chromosome Y shows similar values in both cohorts, with median values of 0.0006558 for the PDAC cohort and 0.0005365 for the PAAD cohort, bearing in mind that only two samples in the whole PAAD cohort even accumulated mutations on chromosome Y.

Among the top 10 most mutated genes in the PAAD cohort, only five genes were also found to be mutated in the analyzed patient sample cohort. The two genes that were expected to be altered in both cohorts are *KRAS* and *TP53*. The *KRAS* gene came on top of the list with alterations in 62% of the samples in the PAAD cohort and recorded a 17% mutation rate in the PDAC cohort, while *TP53* had detected alterations in 19% of the samples from PAAD and 8% in the PDAC cohort. In the PAAD cohort, both genes harbor missense mutations with a small percent of multi-hit mutations (*KRAS*) and nonsense mutations (*TP53*), while in the PDAC cohort, the detected mutations of *KRAS* are classified as multi-hit mutations and *TP53* mutations are classified as nonsense mutations. There are also two more genes found among the top ten mutated genes in the PAAD cohort that are also marked as mutated in the analyzed PDAC RNA-seq samples. These two genes are *TTN* and *MUC16*. In the PAAD cohort, their variants are mostly classified as missense mutations, but there are a few nonsense mutations and multi-hit mutations. On the other hand, the PDAC cohort had variants classified as multi-hit mutations for the *TTN* gene and frame shift insertion for the *MUC16* gene. *TTN* encodes for the large, abundant protein titin, which is a key structural and mechanical component for the assembly and normal functioning of striated muscles (Chauveau et al., 2014; GeneCards - Human Genes, 2023). Even though its mutations have a strong correlation to muscular diseases, mutated *TTN* is often observed in solid tumors and shows a correlation with the

raising tumor mutational burden (TMB) (Xue et al., 2021). Observed *TTN* variants can also be connected with chemotherapy, which is known to induce *TTN* mutations, causing truncating of the titin and subsequently cardiomyopathy (chemotherapy induced cardiomyopathy, CCMP) (Xue et al., 2021). The last gene that accumulated mutations in both cohorts is *MUC16*. *MUC16* encoded protein is membrane-bound mucin, which is thought to provide a protective barrier against different infectious agents at mucosal surfaces (GeneCards - Human Genes, 2023; Haridas et al., 2014). It is proven to be overexpressed in multiple human malignancies and has a significant impact on tumorigenicity as well as acquired resistance to therapy (Aithal et al., 2018). Furthermore, recent research shows its oncogenic role in PDAC; however, the pathological roles of *MUC16* in PDAC progression, tumor microenvironment, and metastasis are yet to be discovered (Lakshmanan et al., 2022).

Analysis of the most mutated pathways in PAAD cohort samples showed no similarities with the most mutated pathways in the analyzed PDAC cohort. The most mutated biological pathway is *MAPK signaling* with alteration in 62% of the samples, and this result was anticipated since the most mutated gene is *KRAS* (with alteration in 62% of the samples), which is one of the crucial parts of this pathway. In second place is the *Genome integrity* pathway, with genes such as *TP53, ATM, ATR, BRCA1, BRCA2*, and others. Even though it is not among the top 3 most mutated pathways in the analyzed PDAC patient samples, the alterations in genes included in this pathway are very well expected in any kind of pathogenic disease, considering the nature of tumors as disturbers of genome integrity and cell cycle.

The comparison of mutational load between analyzed patient tumor samples and 33 TCGA cohorts from the MC3 project by the number of detected mutations per magabase (TMB) implies that the mutational pattern of the analyzed cohort has the most similarity to the SKCM cohort. This kind of result is unexpected considering these are two completely unrelated carcinomas affecting two different organs. One of the explanations for such results could be that the number of mutations called using RNA-seq data is very large and highly prone to false positives, in part due to sequencing and mapping errors (Coudray et al., 2018; Piskol et al., 2013). On the other hand, previous research showed that SKCM is known to accumulate a large number of mutations during tumorigenesis, which can be explained by the nature of mutations occurring in this tumor (Li et al., 2021). That is why both datasets have recorded similarities in high TMB numbers and appear to be similar when they are not.

Mutations detected in patient tumor sequences generated with the RNA-seq method have proven to be sufficient for calling variants and portraying the mutational landscapes of

patients. The differences in the derived mutational landscape between PDAC and PAAD cohorts are clearly observed in this research. This disparity can be justified by the fact that the data from the TCGA MC3 project comes from exome sequences, while the patient cohort data analyzed in this study comes from RNA sequences. Some of the highly expressed genes in cancer cannot be detected because they fall outside the boundaries of exon capture kits, but also because RNA-seq is prone to false positives, in part due to errors during the RNA to cDNA conversion, mapping mismatches, RNA editing processes, and the existence of splice sites and fusion genes (Coudray et al., 2018). This result could also be attributed to the unrepresentative size of the analyzed cohort (12 samples) compared to the other 33 TCGA cohorts, which are much larger (the minimum size of the cohort is 36). It is possible that the mutational pattern of PDAC could not be detected with such a small amount of data, leading to unreliable results. The cause could not be attributed to sequence quality, considering it was checked and assessed to meet the requirements. Other possible causes of divergency are that there was no filtration for RNA-specific mutations, which can greatly impact the result of variant calling (Long et al., 2022). The work from 2022 by Long et al. demonstrated that the implementation of multiple filtering along with machine learning models can greatly improve the accuracy of RNA somatic variants identification. This approach should be used in future studies when RNA-seq data is used for variant calling. Also, it should be noted that variant callers using RNA-seq data are not perfected yet and can produce additional errors (Goode et al., 2013). It was shown that if you want to enhance validation rates while preserving optimal sensitivity, the consensus approach with multiple variant callers provides much better results, which is why it should be used in future research (Goode et al., 2013).

The comparison couldn’t be made with RNA-seq data from TCGA because this data is not publicly available yet. To get a better understanding of the advantages and disadvantages of variant calling using RNA-seq data and the nature of PDAC in general, this data should be included in comparative analysis in the future.

The analysis was made on extracted mutational catalogues to try to observe possible existing mutational signatures among the mutational patterns of PDAC RNA-seq sequences. The goal was to find already defined mutational signatures that would most precisely describe the given pattern of detected mutations in each patient's sequence data. The analysis was conducted with two different algorithms to produce comparable results. Signatures used for matching mutational catalogs with the Fit function algorithm were the signatures that are most commonly found in pancreatic cancer. This group consists of eleven signatures found to

represent characteristic patterns of mutations detected in previous research on pancreatic cancer and reflect its underlying mutational processes. The results of the signature fit analysis with this function couldn't be taken into account considering the low calculated cosine similarity between the provided mutational catalog and the one reconstructed using the chosen mutational signatures. This kind of bad fit could be the result of the large number of signatures trying to be fitted in a one-step manner. Also, signatures used for fitting are defined using whole-genome variant calls produced by the ICGC/TCGA Pan Cancer Analysis of Whole Genomes (PCAWG) Network, while sequences of patient tumor samples researched in this study are generated using RNA sequencing. These two methods of sequence generation differ in parts of the genome which they capture what implicates different mutational patterns, and because of that, they are unable to accurately describe the acquired mutational catalog with the desired mutational signatures (O'Brien et al., 2015).

The algorithm of the FitMS function that conducts fitting in a multi-step manner achieved far better results regarding cosine similarity. This kind of outcome could be attributed to a multi-step method in which the first step of fitting is done on common signatures selected based on a specified organ, and then the presence of rare signatures is detected (Degasperi et al., 2020). A recent study revealed that different organs possess distinct collections of organ-specific signatures, but importantly, the study also demonstrated that the majority of cancers in each organ exhibit a common set of signatures, while only a small subgroup displays a few rare signatures (Degasperi et al., 2022). When considering the tissue-specificity and tumor-specificity of mutational signatures and the distinction between rare and common signatures, there is a lesser number of common signatures trying to be fitted to the mutational catalog, producing a more robust vector of exposures (Degasperi et al., 2022). Additionally, once the common signatures have been accounted for, it is possible to focus on identifying just one rare signature that could enhance the overall fit rather than attempting to incorporate multiple rare signatures simultaneously.

Analysis using the FitMS algorithm appointed two common signatures, SBS1 and SBS5; one rare signature, SBS123; and one combination of common signatures, SBS1+5+18, to fit on mutational catalogues derived from PDAC samples RNA sequences. Signature SBS1 is reported to highly correlate with the age of individuals, which is why it is regarded as a mitotic clock (Alexandrov et al., 2020; Connor et al., 2017; Jianlong et al., 2022). The rates of acquisition of SBS1 mutations differ between cancer tissues and normal tissues, corresponding to rates of stem cell division in different cell types (Jianlong et al., 2022; Koh

et al., 2021). The SBS1 mutational burden is dominated by C>T mutations at CpG sites (Koh et al., 2021). This is the result of an endogenous mutational process led by spontaneous or enzymatic deamination of 5-methylcytosine generating G:T mismatches, which, if failed to be recognized and removed before DNA replication, result in the fixation of the T substitution for C (Koh et al., 2021; Nik-Zainal et al., 2012). Some research shows a correlation between the activity of SBS1 and the activity of SBS5 within different cancer types, although their mutational burdens do not clearly correlate with their activity because of different fundamental processes (COSMIC 2023). SBS5 signature mutations are also connected to the age of the individual (Alexandrov et al., 2020). While the rates of acquisition differ between cancer types and normal cell types, they do not show a correlation with the estimated rates of stem cell division in these tissues (COSMIC 2023). The mutational burden of the SBS5 signature is characterized by T>C and A>G mutations at ATA, ATG, and ATT sequence sites, and it is known to be increased in many cancer types due to tobacco smoking, but its aetiology is yet unknown (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). Research from 2022 by Jianlong et al. showed that in most types of cancer, there was a positive connection between a decreased mutation load and a heightened contribution of SBS1 and SBS5 mutational signatures. The SBS18 mutational signature has been closely connected with damage by reactive oxygen species and shows association with defective base excision repair due to the MUTYH mutation (Jianlong et al., 2022; Jin et al., 2022; Kucab et al., 2019). Its mutational burden consists mostly of C>A substitutions with peaks on TCT, CCA, and GCA sequence sites (COSMIC 2023). An interesting thing to note is that when analyzing the changes in mutational signature activity during human cancer evolution from prostate adenocarcinoma patient samples in a study from 2022, Jianlong et al. found that SBS1 and SBS5 contributed a greater share of mutations during the initial stages of disease progression, whereas SBS18 demonstrated an escalation during later stages.

It is important to mention that the obtained mutational catalogues did not exhibit the mutational pattern of the SBS3 mutational signature, which was not expected taking into account that SBS3 is considered a common one in pancreatic cancer (COSMIC 2023; Hayashi et al., 2021). This flat signature is strongly associated with BRCA1 and BRCA2 mutations and is proposed as a predictor of defective homologous recombination-based repair (Abbasi and Alexandrov, 2021). It would be reasonable to attribute this failed detection of SBS3 to the small size of the analyzed sample cohort, which may not be representative of the overall population of PDAC. The low prevalence of SBS3 in this population, along with various

mutagenic processes, and the fact that SBS3 is known as a 'broad' signature whose motifs overlap with other signatures (signature bleeding) make detection of SBS3 difficult (Abbasi and Alexandrov, 2021). On top of that, mutational signature analysis on RNA-seq data has a bias towards transcribed regions, causing the absence of some genomic mutations and an incomplete representation of the mutational landscape across the entire genome.

Mutational signatures for the PAAD cohort from the TCGA M3 project were not extracted in this study, considering it was already done in the work from 2020 by Alexandrov et al. The study demonstrated the presence of four different mutational signatures in pancreatic cancer data. These are Signature 1B, Signature 2, Signature 3, and Signature 6 (Alexandrov et al., 2020). Neither one of these was detected in the PDAC RNA-seq samples analyzed in this research.

The mutational signature fitting method successfully extracted three common and one rare mutational signature, which was able to explain the mutational pattern. Future research should aim to discover mutational signatures *de novo*, using larger groups of patient sequences for analysis. Also, it would be useful to do differential expression analysis between normal samples and tumor samples to see how tumorigenesis influenced gene expression and connect that information with mutational signature analysis. RNA sequencing once again proved itself as a valuable addition to WXS and WGS, providing valuable insights into the PDAC mutational landscape by identifying its mutational signatures and potentially enhancing diagnostics and therapeutic approaches in PDAC treatment.

The next generation sequencing technologies opened new horizons in the field of genome research and provided invaluable tools for understanding mutational processes in tumors and discovering new clinical applications. While whole-exome sequencing (WXS) has traditionally been the primary method employed for identifying somatic mutations in cancer genomes, new research along with this one indicates that using RNA-seq data from tumor samples to call variants can provide a valuable addition to our understanding of the mutational landscape in tumors (Coudray et al., 2018). In this study, RNA-seq analysis uncovered variants that are linked to the mutational landscape of PDAC, and they served as the primary ingredient in the search for mutational signatures that might be present in the analyzed tumor samples. RNA-seq data can provide valuable information for mutational signature discovery, although it has some limitations in accurately capturing the full mutational landscape of a tumor (compared to WGS or WXS) because it does not obtain all types of mutations, especially noncoding mutations or mutations occurring in regions that are not transcribed.

Therefore, integrating RNA-seq data with other sequencing approaches, such as DNA sequencing, can provide a more comprehensive understanding of the mutational landscape in tumors (Coudray et al., 2018).

6 Conclusions

The detection of somatic mutations and mutational signatures in pancreatic cancer using RNA-sequencing data resulted in following findings:

1. Somatic mutations were detected and annotated using RNA sequencing data from paired tumor and normal tissue patient samples.
2. The mutational landscape of the analyzed samples was constructed based on the characterization of mutations.
3. Genes *XIAP, CTNND1, CTSB, MIR612, ACOX1, EIF2AK2, H2AZ2, METTL7A, SLC4A4 and SOD2* were identified as the most mutated in the analyzed samples, and their impact on tumor development is estimated.
4. The most mutated biological pathways in the analyzed samples are *Wnt/B-catenin signaling*, *Transcription factor*, and *Other signaling*.
5. The mutational impact of commonly mutated genes found in PDAC was assessed and visually represented with lollipop plots.
6. The mutational pattern of the exome-sequenced PAAD cohort from the TCGA database showed little similarity with the one observed in RNA-sequenced patient data.
7. Mutational signatures SBS1, SBS5, SBS18, and SBS123 proved to be the best choice for fitting on mutational catalogues derived from patient sample sequences.
8. RNA sequencing is able to complement exome sequencing in variant detection and mutational signature analysis for a better understanding of tumorigenesis and the development of personalized therapies.

7 References

Abbasi, A. and Alexandrov, L. B. (2021). Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures. *DNA Repair*, *107*, 103200. p. https://doi.org/10.1016/J.DNAREP.2021.103200

Afgan, E., Nekrutenko, A., Grüning, B. A., Blankenberg, D., Goecks, J., Schatz, M. C., Ostrovsky, A. E., Mahmoud, A., Lonie, A. J., Syme, A., Fouilloux, A., Bretaudeau, A., Nekrutenko, A., Kumar, A., Eschenlauer, A. C., Desanto, A. D., Guerler, A., Serrano-Solano, B., Batut, B., … Briggs, P. J. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. pp. https://doi.org/10.1093/NAR/GKAC247

Aguilera, K. Y. and Dawson, D. W. (2021). WNT Ligand Dependencies in Pancreatic Cancer. *Frontiers in Cell and Developmental Biology*, *9*, 671022. p. https://doi.org/10.3389/FCELL.2021.671022

Aithal, A., Rauth, S., Kshirsagar, P., Shah, A., Lakshmanan, I., Junker, W. M., Jain, M., Ponnusamy, M. P. and Batra, S. K. (2018). MUC16 as a novel target for cancer therapy. *Expert Opinion on Therapeutic Targets*, *22*(8), 675–686. pp. https://doi.org/10.1080/14728222.2018.1498845

Alateyah, N., Gupta, I., Rusyniak, R. S. and Ouhtit, A. (2022). SOD2, a Potential Transcriptional Target Underpinning CD44-Promoted Breast Cancer Progression. *Molecules*, *27*(3). https://doi.org/10.3390/MOLECULES27030811

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S. and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, *47*(12), 1402–1407. pp. https://doi.org/10.1038/NG.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Boutros, P., … Yu, W. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. pp. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature 2013 500:7463*, *500*(7463), 415–421. pp. https://doi.org/10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. and Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, *3*(1), 246–259. pp. https://doi.org/10.1016/J.CELREP.2012.12.008

Alharatani, R., Ververi, A., Beleza-Meireles, A., Ji, W., Mis, E., Patterson, Q. T., Griffin, J. N., Bhujel, N., Chang, C. A., Dixit, A., Konstantino, M., Healy, C., Hannan, S., Neo, N., Cash, A., Li, D., Bhoj, E., Zackai, E. H., Cleaver, R., … Liu, K. J. (2020). Novel truncating mutations in CTNND1 cause a dominant craniofacial and cardiac syndrome. *Human Molecular Genetics*, *29*(11), 1900–1921. pp. https://doi.org/10.1093/HMG/DDAA050

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., Yang, H. T., Xue, V., Knyazev, S., Singer, B. D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O., & Mangul, S. (2021). Technology dictates algorithms: recent developments in read

alignment. *Genome Biology 2021 22:1*, *22*(1), 1–34. https://doi.org/10.1186/S13059-021-02443-7

Auwera, G. A. V. der. and O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra - Geraldine A. Van der Auwera, Brian D. O'Connor - Google Books*. Sebastopol : O'Reilly. https://books.google.hr/books?hl=hr&lr=&id=vsXaDwAAQBAJ&oi=fnd&pg=PP1&dq=Genomics+in+the+Cloud:+Using+Docker,+GATK,+and+WDL+in+Terra&ots=5kjk8GTts3&sig=85y1Rp443cnQBu_KpsTduXuAprQ&redir_esc=y#v=onepage&q=Genomics%20in%20the%20Cloud%3A%20Using%20Docker%2C%20GATK%2C%20and%20WDL%20in%20Terra&f=false

Ávila-López, P. A., Guerrero, G., Nuñez-Martínez, H. N., Peralta-Alvarez, C. A., Hernández-Montes, G., Álvarez-Hilario, L. G., Herrera-Goepfert, R., Albores-Saavedra, J., Villegas-Sepúlveda, N., Cedillo-Barrón, L., Montes-Gómez, A. E., Vargas, M., Schnoor, M., Recillas-Targa, F. and Hernández-Rivas, R. (2021). H2A.Z overexpression suppresses senescence and chemosensitivity in pancreatic ductal adenocarcinoma. *Oncogene 2021 40:11*, *40*(11), 2065–2080. pp. https://doi.org/10.1038/s41388-021-01664-1

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K. S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371-385.e18. pp. https://doi.org/10.1016/J.CELL.2018.02.060

Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, *98*(6), 236. p. https://doi.org/10.1136/ARCHDISCHILD-2013-304340

Bernards, R. (2010). It's diagnostics, stupid. *Cell*, *141*(1), 13–17. pp. https://doi.org/10.1016/J.CELL.2010.03.018

Bethune, G., Bethune, D., Ridgway, N. and Xu, Z. (2010). Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *Journal of Thoracic Disease*, *2*(1), 48. p. /pmc/articles/PMC3256436/

Blokzijl, F., Janssen, R., van Boxtel, R. and Cuppen, E. (2018). MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Medicine*, *10*(1), 1–11. pp. https://doi.org/10.1186/S13073-018-0539-0/TABLES/1

Cancer Research UK, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer#heading-Zero, Accessed 7 2023

Cappellesso, F., Orban, M. P., Shirgaonkar, N., Berardi, E., Serneels, J., Neveu, M. A., Di Molfetta, D., Piccapane, F., Caroppo, R., Debellis, L., Ostyn, T., Joudiou, N., Mignion, L., Richiardone, E., Jordan, B. F., Gallez, B., Corbet, C., Roskams, T., DasGupta, R., … Mazzone, M. (2022). Targeting the bicarbonate transporter SLC4A4 overcomes immunosuppression and immunotherapy resistance in pancreatic cancer. *Nature Cancer 2022 3:12*, *3*(12), 1464–1483. pp. https://doi.org/10.1038/s43018-022-00470-2

Chauveau, C., Rowell, J. and Ferreiro, A. (2014). A Rising Titan: TTN Review and Mutation Update. *Human Mutation*, *35*(9), 1046–1059. pp. https://doi.org/10.1002/HUMU.22611

Cheng, L., Lopez-Beltran, A., Massari, F., Maclennan, G. T. and Montironi, R. (2018). Molecular testing for BRAF mutations to inform melanoma treatment decisions: a move toward precision

medicine. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *31*(1), 24–38. pp. https://doi.org/10.1038/MODPATHOL.2017.104

Chepelev, I., Wei, G., Tang, Q. and Zhao, K. (2009). Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Research*, *37*(16). https://doi.org/10.1093/NAR/GKP507

CIGAR Strings Explained – Replicon Genetics (2023) https://replicongenetics.com/cigar-strings-explained/ (accessed 01.06.2023.).

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. pp. https://doi.org/10.4161/fly.19695

ClinVar (2023) https://www.ncbi.nlm.nih.gov/clinvar/ (accessed 14.08.2023.).

Connor, A. A., Denroche, R. E., Jang, G. H., Timms, L., Kalimuthu, S. N., Selander, I., McPherson, T., Wilson, G. W., Chan-Seng-Yue, M. A., Borozan, I., Ferretti, V., Grant, R. C., Lungu, I. M., Costello, E., Greenhalf, W., Palmer, D., Ghaneh, P., Neoptolemos, J. P., Buchler, M., … Gallinger, S. (2017). Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncology*, *3*(6), 774–783. pp. https://doi.org/10.1001/JAMAONCOL.2016.3916

COSMIC (2023) https://cancer.sanger.ac.uk/cosmic (accessed 07.08.2023.).

Coudray, A., Battenhouse, A. M., Bucher, P. and Iyer, V. R. (2018). Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ*, *2018*(7). https://doi.org/10.7717/peerj.5362

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/BIOINFORMATICS/BTR330

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A. M., Simpson, P. T., King, T. A., Raine, K., Eyfjord, J. E., Kong, G., Borg, Å., Birney, E., Stunnenberg, H. G., Van De Vijver, M. J., … Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, *23*(4), 517–525. pp. https://doi.org/10.1038/NM.4292

De Souza, A., Khawaja, K. I., Masud, F. and Saif, M. W. (2016). Metformin and pancreatic cancer: Is there a role? *Cancer Chemotherapy and Pharmacology*, *77*(2), 235–242. pp. https://doi.org/10.1007/S00280-015-2948-8/FIGURES/1

Degasperi, A., Amarante, T. D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., Morganella, S., Nanda, A. S., Badja, C., Koh, G., Momen, S. E., Georgakopoulos-Soares, I., Dias, J. M. L., Young, J., Memari, Y., Davies, H. and Nik-Zainal, S. (2020). A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature Cancer*, *1*(2), 249–263. pp. https://doi.org/10.1038/s43018-020-0027-5

Degasperi, A., Zou, X., Amarante, T. D., Martinez-Martinez, A., Koh, G. C. C., Dias, J. M. L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V. Y. W., Nanda, A. S., Bernstein, A., Momen, S. E., Young, J., Perez-Gil, D., Memari, Y., Badja, C., Shooter, S., Czarnecki, J., … Nik-Zainal, S. (2022). Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science (New York, N.Y.)*, *376*(6591). https://doi.org/10.1126/SCIENCE.ABL9283

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics 2011 43:5*, *43*(5), 491–498. pp. https://doi.org/10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. pp. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. pp. https://doi.org/10.1093/NAR/30.1.207

Fairley, S., Lowy-Gallego, E., Perry, E. and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, *48*(D1), D941–D947. pp. https://doi.org/10.1093/NAR/GKZ836

Fischer, A., Illingworth, C. J. R., Campbell, P. J. and Mustonen, V. (2013). EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, *14*(4), 1–10. pp. https://doi.org/10.1186/GB-2013-14-4-R39/COMMENTS

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., YinKok, C., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T. and Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. pp. https://doi.org/10.1093/NAR/GKW1121

Fujimoto, T., Tsunedomi, R., Matsukuma, S., Yoshimura, K., Oga, A., Fujiwara, N., Fujiwara, Y., Matsui, H., Shindo, Y., Tokumitsu, Y., Suzuki, N., Kobayashi, S., Hazama, S., Eguchi, H. and Nagano, H. (2021). Cathepsin B is highly expressed in pancreatic cancer stem-like cells and is associated with patients' surgical outcomes. *Oncology Letters*, *21*(1), 1–9. pp. https://doi.org/10.3892/OL.2020.12291

GATK (2023) https://gatk.broadinstitute.org/hc/en-us (accessed 25.05.2023.).

Gehring, J. S., Fischer, B., Lawrence, M. and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics (Oxford, England)*, *31*(22), 3673–3675. pp. https://doi.org/10.1093/BIOINFORMATICS/BTV408

GeneCards – Human Genes (2023) https://www.genecards.org/ (accessed 16.07.2023.).

Goode, D. L., Hunter, S. M., Doyle, M. A., Ma, T., Rowley, S. M., Choong, D., Ryland, G. L. and Campbell, I. G. (2013). A simple consensus approach improves somatic mutation prediction accuracy. *Genome Medicine*, *5*(9), 1–14. pp. https://doi.org/10.1186/GM494/TABLES/6

Grant, T. J., Hua, K. and Singh, A. (2016). Molecular Pathogenesis of Pancreatic Cancer. *Progress in Molecular Biology and Translational Science*, *144*, 241–275. pp. https://doi.org/10.1016/BS.PMBTS.2016.09.008

Guimaraes, D. P. and Hainaut, P. (2002). TP53: a key gene in human cancer. *Biochimie*, *84*(1), 83–93. pp. https://doi.org/10.1016/S0300-9084(01)01356-6

Hajdúch, M., Jančík, S., Drábek, J. and Radzioch, D. (2010). Clinical relevance of KRAS in human cancers. *Journal of Biomedicine and Biotechnology*, *2010*. https://doi.org/10.1155/2010/150960

Hanahan, D. and Weinberg, R. A. (2011). *Leading Edge Review Hallmarks of Cancer: The Next Generation*. https://doi.org/10.1016/j.cell.2011.02.013

Haridas, D., Ponnusamy, M. P., Chugh, S., Lakshmanan, I., Seshacharyulu, P. and Batra, S. K. (2014). MUC16: Molecular analysis and its functional implications in benign and malignant conditions. *FASEB Journal*, *28*(10), 4183–4199. pp. https://doi.org/10.1096/FJ.14-257352

Harris, R. S. (2013). Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications. *Genome Medicine*, *5*(9), 87. p. https://doi.org/10.1186/GM490

Hayashi, A., Hong, J. and Iacobuzio-Donahue, C. A. (2021). The pancreatic cancer genome revisited. *Nature Reviews Gastroenterology & Hepatology 2021 18:7*, *18*(7), 469–481. pp. https://doi.org/10.1038/s41575-021-00463-z

Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews. Genetics*, *15*(9), 585–598. pp. https://doi.org/10.1038/NRG3729

Hu, C., Hart, S. N., Polley, E. C., Gnanaolivu, R., Shimelis, H., Lee, K. Y., Lilyquist, J., Na, J., Moore, R., Antwi, S. O., Bamlet, W. R., Chaffee, K. G., DiCarlo, J., Wu, Z., Samara, R., Kasi, P. M., McWilliams, R. R., Petersen, G. M. and Couch, F. J. (2018). Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer. *JAMA*, *319*(23), 2401–2409. pp. https://doi.org/10.1001/JAMA.2018.6228

Hu, H. Feng, Ye, Z., Qin, Y., Xu, X. Wu, Yu, X. Jun, Zhuo, Q. Feng and Ji, S. Rong. (2021). Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications. In *Acta Pharmacologica Sinica* (Vol. 42, Number 11, 1725–1741. pp.). Springer Nature. https://doi.org/10.1038/s41401-020-00584-2

Huang, J., Wang, H., Xu, Y., Li, C., Lv, X., Han, X., Chen, X., Chen, Y. and Yu, Z. (2023). The Role of CTNNA1 in Malignancies: An Updated Review. *Journal of Cancer*, *14*(2), 219. p. https://doi.org/10.7150/JCA.79236

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusuda, J., Lane, D. P., … Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998. pp. https://doi.org/10.1038/NATURE08987

Hussain, M. R. M., Baig, M., Mohamoud, H. S. A., Ulhaq, Z., Hoessli, D. C., Khogeer, G. S., Al-Sayed, R. R. and Al-Aama, J. Y. (2015). BRAF gene: From human cancers to developmental syndromes. *Saudi Journal of Biological Sciences*, *22*(4), 359–373. pp. https://doi.org/10.1016/J.SJBS.2014.10.002

Infantino, V., Todisco, S. and Convertini, P. (2023). Mitochondrial physiology: An overview. *Mitochondrial Intoxication*, 1–27. pp. https://doi.org/10.1016/B978-0-323-88462-4.00001-8

Introduction to RNA-seq using high-performance computing (HPC) (2021) https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html (accessed 20.06.2023.).

Javadrashid, D., Mohammadzadeh, R., Baghbanzadeh, A., Safaee, S., Amini, M., Lotfi, Z., Baghbani, E., Shahgoli, V. K. and Baradaran, B. (2021). Simultaneous microRNA-612 restoration and 5-FU treatment inhibit the growth and migration of human PANC-1 pancreatic cancer cells. *EXCLI Journal*, *20*, 160. p. https://doi.org/10.17179/EXCLI2020-2900

Jianlong, L., Xu, J., Li, Y. K. and Li, X. (2022). Clinical and genomic characterization of mutational signatures across human cancers. *Article in International Journal of Cancer*. https://doi.org/10.1002/ijc.34402

Jin, S. G., Meng, Y., Johnson, J., Szabó, P. E. and Pfeifer, G. P. (2022). Concordance of hydrogen peroxide–induced 8-oxo-guanine patterns with two cancer mutation signatures of upper GI tract tumors. *Science Advances*, *8*(22), 3815. p. https://doi.org/10.1126/SCIADV.ABN3815/SUPPL_FILE/SCIADV.ABN3815_SM.PDF

Kamisawa, T., Wood, L. D., Itoi, T. and Takaori, K. (2016). Pancreatic cancer. In *The Lancet* (Vol. 388, Number 10039, 73–85. pp.). Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(16)00141-0

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, and D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. pp. https://doi.org/10.1101/GR.229102

Khanna, K. K. (2000). Cancer Risk and the ATM Gene: a Continuing Debate. *JNCI: Journal of the National Cancer Institute*, *92*(10), 795–802. pp. https://doi.org/10.1093/JNCI/92.10.795

Khorana, A. A. (2012). Cancer and coagulation. *American Journal of Hematology*, *87 Suppl 1*(Suppl 1). https://doi.org/10.1002/AJH.23143

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P. and Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. pp. https://doi.org/10.1038/S41592-018-0051-X

Kim, Y. S., Vallur, P. G., Phaëton, R., Mythreye, K. and Hempel, N. (2017). Insights into the Dichotomous Regulation of SOD2 in Cancer. *Antioxidants*, *6*(4). https://doi.org/10.3390/ANTIOX6040086

Koh, G., Degasperi, A., Zou, X., Momen, S. and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. In *Nature Reviews Cancer* (Vol. 21, Number 10, 619–637. pp.). Nature Research. https://doi.org/10.1038/s41568-021-00377-7

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., Arlt, V. M., Phillips, D. H. and Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell*, *177*(4), 821-836.e16. pp. https://doi.org/10.1016/j.cell.2019.03.001

Kuipers, D. J. S., Mandemakers, W., Lu, C. S., Olgiati, S., Breedveld, G. J., Fevga, C., Tadic, V., Carecchio, M., Osterman, B., Sagi-Dain, L., Wu-Chou, Y. H., Chen, C. C., Chang, H. C., Wu, S. L., Yeh, T. H., Weng, Y. H., Elia, A. E., Panteghini, C., Marotta, N., … Bonifati, V. (2021). EIF2AK2 Missense Variants Associated with Early Onset Generalized Dystonia. *Annals of Neurology*, *89*(3), 485–497. pp. https://doi.org/10.1002/ANA.25973

Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951. p. https://doi.org/10.1101/PDB.TOP084970

Lakshmanan, I., Marimuthu, S., Chaudhary, S., Seshacharyulu, P., Rachagani, S., Muniyan, S., Chirravuri-Venkata, R., Atri, P., Rauth, S., Nimmakayala, R. K., Siddiqui, J. A., Gautam, S. K., Shah, A., Natarajan, G., Parte, S., Bhyravbhatla, N., Mallya, K., Haridas, D., Talmon, G. A., … Batra, S. K. (2022). Muc16 depletion diminishes KRAS-induced tumorigenesis and metastasis by altering tumor microenvironment factors in pancreatic ductal adenocarcinoma. *Oncogene*, *41*(48), 5147–5159. pp. https://doi.org/10.1038/S41388-022-02493-6

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. pp. https://doi.org/10.1093/NAR/GKX1153

Leinonen, R., Sugawara, H. and Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database issue), D19. p. https://doi.org/10.1093/NAR/GKQ1019

Li, L., Bai, L., Lin, H., Dong, L., Zhang, R., Cheng, X., Liu, Z., Ouyang, Y. and Ding, K. (2021). Multiomics analysis of tumor mutational burden across cancer types. *Computational and Structural Biotechnology Journal*, *19*, 5637. p. https://doi.org/10.1016/J.CSBJ.2021.10.013

Li, S., Sun, J., Yang, J., Zhang, L., Wang, L., Wang, X. and Guo, Z. (2013). XIAP expression is associated with pancreatic carcinoma outcome. *Molecular and Clinical Oncology*, *1*(2), 305–308. pp. https://doi.org/10.3892/MCO.2013.58

Lin, W., Qiu, X., Sun, P., Ye, Y., Huang, Q., Kong, L. and Lu, J. J. (2021). Association of IDH mutation and 1p19q co-deletion with tumor immune microenvironment in lower-grade glioma. *Molecular Therapy - Oncolytics*, *21*, 288–302. pp. https://doi.org/10.1016/J.OMTO.2021.04.010

Liu, J., Xiao, Q., Xiao, J., Niu, C., Li, Y., Zhang, X., Zhou, Z., Shu, G. and Yin, G. (2022). Wnt/β-catenin signalling: function, biological mechanisms, and therapeutic opportunities. *Signal Transduction and Targeted Therapy 2021 7:1*, *7*(1), 1–23. pp. https://doi.org/10.1038/s41392-021-00762-6

Liu, Z., Chen, Y. and Shen, T. (2023). Evidence Based on an Integrative Analysis of Multi-Omics Data on METTL7A as a Molecular Marker in Pan-Cancer. *Biomolecules*, *13*(2), 195. p. https://doi.org/10.3390/BIOM13020195/S1

Long, Q., Yuan, Y. and Li, M. (2022). RNA-SSNV: A Reliable Somatic Single Nucleotide Variant Identification Framework for Bulk RNA-Seq Data. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.865313

Ma, K., Chen, X., Liu, W., Chen, S., Yang, C. and Yang, J. (2022). CTSB is a negative prognostic biomarker and therapeutic target associated with immune cells infiltration and immunosuppression in gliomas. *Scientific Reports 2022 12:1*, *12*(1), 1–15. pp. https://doi.org/10.1038/s41598-022-08346-2

Maitra, A. and Hruban, R. H. (2008). Pancreatic cancer. In *Annual Review of Pathology: Mechanisms of Disease* (Vol. 3, 157–188. pp.). https://doi.org/10.1146/annurev.pathmechdis.3.121806.154305

Makena, M. R., Gatla, H., Verlekar, D., Sukhavasi, S., Pandey, M. K. and Pramanik, K. C. (2019). Wnt/β-Catenin Signaling: The Culprit in Pancreatic Carcinogenesis and Therapeutic Resistance. *International Journal of Molecular Sciences*, *20*(17). https://doi.org/10.3390/IJMS20174242

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J. B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 1–16. https://doi.org/10.1186/GM543/TABLES/2

McDonald, K. L., Tabone, T., Nowak, A. K. and Erber, W. N. (2015). Somatic mutations in glioblastoma are associated with methylguanine-DNA methyltransferase methylation. *Oncology Letters*, *9*(5), 2063–2067. pp. https://doi.org/10.3892/OL.2015.2980

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Yung, W. K. A., Bogler, O., Weinstein, J. N.,

VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. pp. https://doi.org/10.1038/NATURE07385

Mort, J. S. (2013). Cathepsin B. *Handbook of Proteolytic Enzymes*, *2*, 1784–1791. pp. https://doi.org/10.1016/B978-0-12-382219-2.00406-3

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [2023 Aug 11]. Available from: https://www.ncbi.nlm.nih.gov/

National Cancer Institute NCI Dictionary of Cancer Terms 21.09d September 27, 2021 Bethesda, MD (accessed 30.06.2023.).

NCI Pancreatic Ductal Adenocarcinoma Study (2023) https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/pancreatic-ductal-adenocarcinoma-study (accessed 03.06.2023.).

NCI The Cancer Genome Atlas Program (TCGA) (2023) https://www.cancer.gov/ccg/research/genome-sequencing/tcga/ (accessed 06.07.2023.).

Nie, S., Shi, Z., Shi, M., Li, H., Qian, X., Peng, C., Ding, X., Zhang, S., Lv, Y., Wang, L., Kong, B., Zou, X. and Shen, S. (2022). PPARγ/SOD2 Protects Against Mitochondrial ROS-Dependent Apoptosis via Inhibiting ATG4D-Mediated Mitophagy to Promote Pancreatic Cancer Proliferation. *Frontiers in Cell and Developmental Biology*, *9*, 745554. p. https://doi.org/10.3389/FCELL.2021.745554/BIBTEX

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–993. pp. https://doi.org/10.1016/J.CELL.2012.04.024

Nowara, E. and Huszno, J. (2016). Masitinib plus gemcitabine for personalized treatment of PDAC patients with overexpression of ACOX1. *Https://Doi.Org/10.1080/23808993.2016.1257911*, *1*(6), 479–485. pp. https://doi.org/10.1080/23808993.2016.1257911

O'Brien, T. D., Jia, P., Xia, J., Saxena, U., Jin, H., Vuong, H., Kim, P., Wang, Q., Aryee, M. J., Mino-Kenudson, M., Engelman, J. A., Le, L. P., Iafrate, A. J., Heist, R. S., Pao, W. and Zhao, Z. (2015). Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods*, *83*, 118–127. pp. https://doi.org/10.1016/J.YMETH.2015.04.016

Orth, M., Metzger, P., Gerum, S., Mayerle, J., Schneider, G., Belka, C., Schnurr, M. and Lauber, K. (2019). Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiation Oncology 2019 14:1*, *14*(1), 1–20. pp. https://doi.org/10.1186/S13014-019-1345-6

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, M. J., Sellers, W. R., Johnson, B. E. and Meyerson, M. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science (New York, N.Y.)*, *304*(5676), 1497–1500. pp. https://doi.org/10.1126/SCIENCE.1099314

Pandit, H., Zhang, W., Li, Y., Agle, S., Li, X., Li, S. P., Cui, G., Li, Y. and Martin, R. C. G. (2015). Manganese superoxide dismutase expression is negatively associated with microRNA-301a in

human pancreatic ductal adenocarcinoma. *Cancer Gene Therapy*, *22*(10), 481–486. pp. https://doi.org/10.1038/CGT.2015.46

Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., Arslan, A. A., Bueno-De-Mesquita, H. B., Gallinger, S., Gross, M., Helzlsouer, K., Holly, E. A., Jacobs, E. J., Klein, A. P., Lacroix, A., Li, D., Mandelson, M. T., Olson, S. H., Risch, H. A., … Chanock, S. J. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, *42*(3), 224–228. pp. https://doi.org/10.1038/NG.522

Pihlak, R., Valle, J. W. and McNamara, M. G. (2017). Germline mutations in pancreatic cancer and potential new therapeutic options. *Oncotarget*, *8*(42), 73240–73257. pp. https://doi.org/10.18632/ONCOTARGET.17291

Piskol, R., Ramaswami, G. and Li, J. B. (2013). Reliable Identification of Genomic Variants from RNA-Seq Data. *American Journal of Human Genetics*, *93*(4), 641. p. https://doi.org/10.1016/J.AJHG.2013.08.008

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordõez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), 191–196. pp. https://doi.org/10.1038/NATURE08658

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K. W., Kamburov, A., Maruvka, Y. E., Leshchiner, I., Lander, E. S., Golub, T. R., Zick, A., Orthwein, A., Lawrence, M. S., Batra, R. N., … Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, *49*(10), 1476. p. https://doi.org/10.1038/NG.3934

Porta, M., Fabregat, X., Malats, N., Guarner, L., Carrato, A., De Miguel, A., Ruiz, L., Jariod, M., Costafreda, S., Coll, S., Alguacil, J., Corominas, J. M., Solà, R., Salas, A. and Real, F. X. (2005). Exocrine pancreatic cancer: symptoms at presentation and their relation to tumour site and stage. *Clinical & Translational Oncology : Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, *7*(5), 189–197. pp. https://doi.org/10.1007/BF02712816

Quante, A. S., Ming, C., Rottmann, M., Engel, J., Boeck, S., Heinemann, V., Westphalen, C. B. and Strauch, K. (2016). Projections of cancer incidence and cancer-related deaths in Germany by 2020 and 2030. *Cancer Medicine*, *5*(9), 2649–2656. pp. https://doi.org/10.1002/CAM4.767

Radenbaugh, A. J., Ma, S., Ewing, A., Stuart, J. M., Collisson, E. A., Zhu, J. and Haussler, D. (2014). RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection. *PLOS ONE*, *9*(11), e111516. p. https://doi.org/10.1371/JOURNAL.PONE.0111516

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M. and 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. pp. https://doi.org/10.1158/0008-5472.CAN-14-0155

Raphael, B. J., Dobson, J. R., Oesper, L. and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. *Genome Medicine*, *6*(1), 1–17. pp. https://doi.org/10.1186/GM524/FIGURES/4

Regan, K., Saghafi, A. and Li, Z. (2021). Splice Junction Identification using Long Short-Term Memory Neural Networks. *Current Genomics*, *22*(5), 384. p. https://doi.org/10.2174/1389202922666211011143008

Rünger, T. M. (2008). C→T Transition Mutations Are Not Solely UVB-Signature Mutations, Because They Are Also Generated by UVA. *Journal of Investigative Dermatology*, *128*(9), 2138–2140. pp. https://doi.org/10.1038/JID.2008.165

Salmerón-Bárcenas, E. G., Zacapala-Gómez, A. E., Lozano-Amado, D., Castro-Muñoz, L. J., Leyva-Vázquez, M. A., Manzo-Merino, J. and Ávila-López, P. A. (2021). Comprehensive bioinformatic analysis reveals oncogenic role of H2A.Z isoforms in cervical cancer progression. *Iranian Journal of Basic Medical Sciences*, *24*(11), 1470–1481. pp. https://doi.org/10.22038/IJBMS.2021.58287.

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G. and Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World Journal of Gastrointestinal Oncology*, *12*(2), 173. p. https://doi.org/10.4251/WJGO.V12.I2.173

Schilbert, H. M., Rempel, A., & Pucker, B. (2020). Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. *Plants*, *9*(4). https://doi.org/10.3390/PLANTS9040439

Siegel Mph, R. L., Miller, K. D., Sandeep, N., Mbbs, W., Ahmedin, |, Dvm, J. and Siegel, R. L. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(1), 17–48. pp. https://doi.org/10.3322/CAAC.21763

Skoulidis, F. and Heymach, J. V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nature Reviews. Cancer*, *19*(9), 495–509. pp. https://doi.org/10.1038/S41568-019-0179-8

Stark, R., Grzelak, M. and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, *20*(11), 631–656. pp. https://doi.org/10.1038/s41576-019-0150-2

Steven, S., Frenis, K., Oelze, M., Kalinovic, S., Kuntic, M., Jimenez, M. T. B., Vujacic-Mirski, K., Helmstädter, J., Kröller-Schön, S., Münzel, T. and Daiber, A. (2019). Vascular inflammation and oxidative stress: Major triggers for cardiovascular disease. *Oxidative Medicine and Cellular Longevity*, *2019*. https://doi.org/10.1155/2019/7092151

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. pp. https://doi.org/10.1038/NATURE07943

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. pp. https://doi.org/10.3322/CAAC.21660

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. pp. https://doi.org/10.1093/NAR/GKY1015

Taylor, T. E., Furnari, F. B. and K. Cavenee, W. K. (2012). Targeting EGFR for treatment of glioblastoma: molecular basis to overcome resistance. *Current Cancer Drug Targets*, *12*(3), 197–209. pp. https://doi.org/10.2174/156800912799277557

Tu, H. and Costa, M. (2020). XIAP's Profile in Human Cancer. *Biomolecules*, *10*(11), 1–15. pp. https://doi.org/10.3390/BIOM10111493

UCSC Genome Browser (2023) https://genome.ucsc.edu/index.html (accessed 07.08.2023.).

Van Hoeck, A., Tjoonk, N. H., Van Boxtel, R. and Cuppen, E. (2019). Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer 2019 19:1*, *19*(1), 1–14. pp. https://doi.org/10.1186/S12885-019-5677-2

Vogler, M., Walczak, H., Stadel, D., Haas, T. L., Genze, F., Jovanovic, M., Bhanot, U., Hasel, C., Möller, P., Gschwend, J. E., Simmet, T., Debatin, K. M. and Fulda, S. (2009). Small molecule XIAP inhibitors enhance TRAIL-induced apoptosis and antitumor activity in preclinical models of pancreatic carcinoma. *Cancer Research*, *69*(6), 2425–2434. pp. https://doi.org/10.1158/0008-5472.CAN-08-2436

Wang, Z., Lu, Y., Fornage, M., Jiao, L., Shen, J., Li, D. and Wei, P. (2022). Identification of novel susceptibility methylation loci for pancreatic cancer in a two-phase epigenome-wide association study. *Epigenetics*, *17*(11), 1357–1372. pp. https://doi.org/10.1080/15592294.2022.2026591/SUPPL_FILE/KEPI_A_2026591_SM6769.ZIP

Watson, I. R., Takahashi, K., Futreal, P. A. and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature Reviews. Genetics*, *14*(10), 703–718. pp. https://doi.org/10.1038/NRG3539

Weidemüller, P., Kholmatov, M., Petsalaki, E. and Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics*, *21*(23–24). https://doi.org/10.1002/PMIC.202000034

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics 2013 45:10*, *45*(10), 1113–1120. pp. https://doi.org/10.1038/ng.2764

Wen, L., Britton, C. J., Garje, R., Darbro, B. W. and Packiam, V. T. (2021). The emerging role of somatic tumor sequencing in the treatment of urothelial cancer. *Asian Journal of Urology*, *8*(4), 391–399. pp. https://doi.org/10.1016/J.AJUR.2021.06.005

Whitley, S. K., Horne, W. T. and Kolls, J. K. (2016). Research Techniques Made Simple: Methodology and Clinical Applications of RNA Sequencing. *Journal of Investigative Dermatology*, *136*(8), e77–e82. pp. https://doi.org/10.1016/J.JID.2016.06.003

Xia, R., Xu, M., Yang, J. and Ma, X. (2022). The role of Hedgehog and Notch signaling pathway in cancer. *Molecular Biomedicine*, *3*(1). https://doi.org/10.1186/S43556-022-00099-8

Xiao, W., Ren, L., Chen, Z., Fang, L. T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., Blomquist, T. M., Hung, T., Sultan, M., Idler, K., Lu, C., Scherer, A., Kusko, R., Moos, M., Xiao, C., … Shi, L. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nature Biotechnology 2021 39:9*, *39*(9), 1141–1150. pp. https://doi.org/10.1038/s41587-021-00994-5

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, *16*, 15–24. https://doi.org/10.1016/J.CSBJ.2018.01.003

Xue, D., Lin, H., Lin, L., Wei, Q., Yang, S. and Chen, X. (2021). TTN/ TP53 mutation might act as the predictor for chemotherapy response in lung adenocarcinoma and lung squamous carcinoma patients. *Translational Cancer Research*, *10*(3), 1284–1294. pp. https://doi.org/10.21037/TCR-20-2568

Yu, B., O'Toole, S. A. and Trent, R. J. (2015). Somatic DNA mutation analysis in targeted therapy of solid tumours. *Translational Pediatrics*, *4*(2), 125. p. https://doi.org/10.3978/J.ISSN.2224-4336.2015.04.04

Zehmer, J. K., Bartz, R., Bisel, B., Liu, P., Seemann, J. and Anderson, R. G. W. (2009). Targeting sequences of UBXD8 and AAM-B reveal that the ER has a direct role in the emergence and regression of lipid droplets. *Journal of Cell Science*, *122*(20), 3694–3702. pp. https://doi.org/10.1242/JCS.054700

Zehmer, J. K., Bartz, R., Liu, P. and Anderson, R. G. W. (2008). Identification of a novel N-terminal hydrophobic sequence that targets proteins to lipid droplets. *Journal of Cell Science*, *121*(11), 1852–1860. pp. https://doi.org/10.1242/JCS.012013

Zhao, M., Mishra, L. and Deng, C. X. (2018). The role of TGF-β/SMAD4 signaling in cancer. *International Journal of Biological Sciences*, *14*(2), 111. p. https://doi.org/10.7150/IJBS.23230

CURRICULUM VITAE

Jan Pantlik

jan.pantlik@gmail.com

Born: 23.02.1999., Zagreb, Croatia

EDUCATION

2020.-2023.	University Graduate Programme in Molecular Biology Faculty of Science, University of Zagreb (Croatia)
2017.-2020.	Undergraduate Programme of Science in Biology Faculty of Science, Division of Biology, Zagreb (Croatia)

WORK EXPERIENCE

June 2021. – Aug 2023.	Lifeguard at *Red Cross*, Zagreb
Sep 2022. – Aug 2023.	Barman at night club *Katran*, Zagreb
Aug 2020. – May 2021.	Waiter at caffe bar *Bubby*; Megabox d.o.o., Zagreb
May– Sep 2019.	Waiter at caffe bar *Pejo*, Betina
Jun – Sep 2018.	Waiter at restaurant *Zameo ih vjetar*; Gangaro d.o.o., Murter
Apr – Jun 2018.	Warehouse worker; *HNK transport*, Zagreb

INTERNSHIPS

Apr 2022. – Aug 2022.	Erasmus+ internship at Biotechnology and Synthetic Biology laboratory at the Institute for Integrative Systems Biology I2SysBio; University of Valencia

CONFERENCES

2019.	15. European Youth Parliament for Water, Nizhniy Novgorod - participant

SKILLS

Languages	Croatian - native proficiency, English - B2
Software	R

EXTRA-CURRICULAR ACTIVITIES

2017. – 2021.	Water polo player at VK Zagreb, competing in Second division of Croatian Water polo League