UMEÅ UNIVERSITY

In search of early biomarkers in pancreatic ductal adenocarcinoma using multi-omics and bioinformatics

Emmy Borgmästars

Department of Surgical and Perioperative Sciences
Umeå 2022

Research activities

- The Swedish National Graduate School in Medical Bioinformatics (medbioinfo.se)

- Research stay at International Agency for Research of Cancer (IARC, World Health Organization, Lyon, France)

Dissertation for PhD
ISBN: 978-91-7855-928-2 (print)
ISBN: 978-91-7855-929-9 (digital)
ISSN: 0346–6612
New Series Number 2211
Cover design by Anja Sundberg
Electronic version available at: http://umu.diva-portal.org/
Printed by: Cityprint i Norr AB
Umeå, Sweden 2022

Till Karla och Anton

Start where you are. Use what you have. Do what you can.

- Arthur Ashe

Table of Contents

Abstract

Background: Pancreatic ductal adenocarcinoma (PDAC) is a very aggressive malignancy with a 5-year survival of 10 %. Surgery is the only curative treatment. Unfortunately, few patients are eligible for surgery due to late detection. Thus, we need ways to detect the disease at an earlier stage and for that, good screening biomarkers could be used. Previous studies have analyzed circulating analytes in prospective studies to identify early PDAC signals. One such class is microRNAs (miRNAs). MicroRNAs are non-coding RNAs of around 22 nucleotides that act as post-transcriptional regulators by interaction with messenger RNAs (mRNAs). The function of a miRNA can be elucidated by target prediction, to identify its potential targets, followed by enrichment analysis of the predicted targets. Challenges with this approach includes a lot of false positives being generated and that miRNAs can perform their role in a tissue- or disease-specific manner. Other classes of analytes that have previously been studied in prospective PDAC cohorts are metabolites and proteins.

Aims: This thesis has three aims. First, to build a miRNA functional analysis pipeline with correlation support between miRNA and its predicted target genes. Second, to identify potential circulating biomarkers for early detection of PDAC using multi-omics. Third, to identify potential prognostic metabolites in a prospective PDAC cohort.

Methods: We used publicly available data from the cancer genome atlas-pancreatic adenocarcinoma (TCGA-PAAD) and pre-diagnostic plasma samples from the Northern Sweden Health and Disease Study. We built a pipeline in R including miRNA, mRNA, and protein expression data from TCGA-PAAD for *in silico* miRNA functional analysis. Pre-diagnostic plasma samples from future PDAC patients as well as matched healthy controls were analyzed using multi-omics. Tissue polypeptide specific antigen (TPS) was analyzed by enzyme linked immunosorbent assay in 267 future PDAC samples and 320 healthy controls. Metabolomics and clinical biomarkers (carbohydrate antigen (CA) 19-9, carcinoembryonic antigen (CEA), and CA 15-3) were profiled in 100 future PDAC samples and 100 healthy controls using liquid chromatography-mass spectrometry (MS), gas chromatography-MS, and multi-plex technology. Of these, a subset of 39 future PDAC patients and 39 healthy controls were

profiled for 2083 microRNAs using targeted sequencing and 644 proteins using proximity extension assays. Circulating levels of multi-omics analytes were analyzed using conditional or unconditional logistic regression. Least absolute shrinkage and selection operator (LASSO) in combination with 500 bootstrap iterations identified the most informative variables. The prognostic value of metabolites was assessed using cox regression. Multi-omics factor analysis (MOFA) and data integration analysis for biomarker discovery using latent components (DIABLO) were used for multi-omics integration analyses.

Results: An automated pipeline was built consisting of 1) miRNA target prediction, 2) correlation analyses between miRNA and its targets on mRNA and protein expression levels, and 3) functional enrichment of correlated targets to identify enriched Kyoto encyclopedia of genes and genomes (KEGG) pathways and gene ontology (GO) terms for a specific miRNA. The pipeline was run for all microRNAs (~700) detected in the TCGA-PAAD cohort. These results can be downloaded from a shiny app (https://emmbor.shinyapps.io/mirfa/). TPS was not altered in pre-diagnostic PDAC patients up to 24 years prior to diagnosis, but increased at diagnosis (OR = 1.03, 95 % CI: 1.01-1.05). Internal area under curves of 0.74, 0.80, and 0.88 were achieved for five metabolites, two proteins, and two miRNAs, that were selected by LASSO and bootstrap iterations, in combination with CA 19-9. Neither MOFA nor DIABLO separated well between future PDAC cases and healthy controls.

Conclusions: Our bioinformatics pipeline for *in silico* functional analysis of microRNAs successfully identifies enriched KEGG pathways and GO terms for miRNA isoforms. The investigated plasma samples are heterogeneous, but among the analyzed variables, we identified five metabolites, two proteins, and two microRNAs with highest potential for early PDAC detection. CA 19-9 levels increased closer to diagnosis. We identified five fatty acids that could be studied in a diagnostic PDAC cohort as prognostic biomarkers.

Enkel sammanfattning på svenska

Bukspottkörtelcancer är mycket aggressiv med en dyster prognos. Kirurgi är den enda botande behandlingen, men på grund av sen upptäckt kan endast en liten andel patienter opereras med botande syfte. Vi behöver därför sätt att upptäcka sjukdomen i ett tidigare skede. För detta ändamål skulle bra biomarkörer, mätbara nivåer av biologiska substanser, som påverkas av en sjukdom vara av värde. I tidigare studier har biomarkörer studerats i pre-diagnostiska blodprover från individer som senare utvecklar bukspottkörtelcancer med syfte att identifiera tidiga bukspottkörtelcancer-signaler för att kunna behandla sjukdomen i ett tidigare skede. En sådan klass av biomarkörer är mikroRNA (miRNA). MiRNA är korta RNA-molekyler på cirka 22 nukleotider och reglerar genuttryck genom att binda till budbärar-RNA (mRNA). Funktionen av ett miRNA kan studeras genom att identifiera potentiella målgener som ett specifikt miRNA kan binda till, följt av funktionell analys av de identifierade målgenerna. Utmaningar med detta tillvägagångssätt inkluderar att många falsk-positiva målgener genereras och att ett miRNA kan utföra sin roll på ett vävnads- och sjukdomsspecifikt sätt. Andra klasser av biomarkörer som tidigare har studerats i bukspottkörtelcancer är metaboliter och proteiner.

Denna avhandling har tre syften. För det första, att bygga ett bioinformatiskt verktyg med korrelationsstöd mellan miRNA och dess potentiella målgener på mRNA- och proteinuttrycksnivåer. För det andra, att identifiera potentiella cirkulerande biomarkörer för tidig upptäckt av bukspottkörtelcancer med hjälp av multi-omik, dvs mikroRNA, metaboliter och proteiner. För det tredje, att definiera metaboliter som kan förutspå överlevnadstid för bukspottkörtelcancer-patienter.

Metoder: Fyra olika patientgrupper analyserades i denna avhandling. I den första studien inkluderade vi offentligt data från bukspott-körtelcancer inom the Cancer Genome Atlas (TCGA-PAAD). Vi byggde ett bioinformatiskt verktyg med miRNA, mRNA och protein-uttrycksdata från TCGA-PAAD för funktionell analys av miRNA *in silico* med stöd från miRNA-målgenskorrelationer. I de andra tre studierna använde vi plasmaprover från Northern Sweden Health and Disease Study (NSHDS) biobanken. Vi inkluderade plasmaprover från individer som senare utvecklade bukspottkörtelcancer samt matchade friska kontroller. Tissue

polypeptide-specific antigen (TPS) analyserades i 267 plasmaprover från framtida bukspottkörtelcancer-patienter och 320 friska kontroller. Metaboliter och tre kliniska biomarkörer mättes i 100 plasmaprover från patienter som senare utvecklar bukspottkörtelcancer och 100 friska kontroller. Av dessa 100 definierade vi en undergrupp bestående av 39 framtida bukspott-körtelcancer-patienter och 39 friska kontroller för ytterligare analys av 2083 mikroRNA och 644 proteiner. Uppmätta nivåer av TPS, metaboliter, mikroRNA, proteiner och kliniska biomarkörer i plasma jämfördes mellan framtida bukspottkörtelcancer-patienter och friska kontroller.

Ett bioinformatiskt verktyg byggdes i mjukvaran R bestående av 1) miRNA-målgenprediktion, 2) korrelationsanalyser mellan miRNA och dess målgener på mRNA- och proteinuttrycksnivåer, samt 3) funktionell anrikning av korrelerade målgener för att identifiera över-representerade signalvägar och funktionstermer för ett specifikt miRNA. Metodflödet kördes för alla mikroRNA (~700) som mätts i TCGA-PAAD-kohorten. Resultat för dessa finns tillgängligt för nerladdning (https://emmbor.shinyapps.io/mirfa/). TPS skiljde sig inte mellan framtida bukspottkörtelcancer-patienter och friska kontroller upp till 24 år innan bukspottkörtelcancer-diagnos. En skillnad observerades däremot vid tidpunkten för diagnos. Några potentiella metaboliter för att förutspå överlevnadstid identifierades men dessa behövas undersökas ytterligare vid tidpunkten för bukspottkörtelcancer-diagnos. Genom kombinerad multi-omik i plasma lyckades vi inte tydligt separera framtida bukspottkörtelcancer-patienter från friska kontroller.

Sammanfattningsvis så identifierar vårt bioinformatiska verktyg för funktionell analys av miRNA framgångsrikt över-representerade signalvägar och funktionstermer för miRNA. De undersökta plasmaproverna uppvisar stor variation. Bland de analyser vi har gjort, så har vi identifierat fem metaboliter, två proteiner och två microRNA med mest potential att fungera som biomarkörer för tidig detektion av bukspottkörtelcancer. Vi identifierade fem metaboliter som kan studeras vidare i bukspottkörtelcancer-patienter som biomarkörer för att förutspå överlevnadstid.

Abbreviations

ACC	Accuracy
AI	Artificial intelligence
AU	Arbitrary units
AJCC	American joint committee on cancer
BCAA	Branched chain amino acids
BMI	Body mass index
BRCA2	BRCA2 DNA repair associated
CA 19-9	Carbohydrate antigen 19-9
CAPS	International cancer of the pancreas screening
CCL15	C-C motif chemokine ligand 15
CDKN2A	Cyclin dependent kinase inhibitor 2A
CEA	Carcinoembryonic antigen
CoD	Curse of dimensionality
CP	Chronic pancreatitis
CT	Computer tomography
CV	Coefficient of variation
DIABLO	Data integration analysis for biomarker discovery using latent components
EPIC	European prospective investigation into cancer and nutrition
EUS	Endoscopic ultrasound
FDR	False discovery rate
FGF	Fibroblast growth factor
GCMS	Gas chromatography mass spectrometry
GDPR	General data protection regulation
GNAS	GNAS complex locus
GO	Gene ontology
HR	Hazard ratio
IFG	Impaired fasting glucose
IPMN	Intraductal papillary mucinous neoplasm
ITPN	Intraductal tubulopapillary neoplasms
kAU	Kiloarbitrary units
KEGG	Kyoto encyclopedia of genes and genomes
KIF2C	Kinesin family member 2C
KRAS	Kirsten rat sarcoma virus
LASSO	Least absolute shrinkage and selection operator

LCMS	Liquid chromatography mass spectrometry
MCN	Mucinous cystic neoplasm
Mice	Multiple imputation chained equations
MiRFA	MicroRNA functional analysis
MiRNA	MicroRNA
MODY	Maturity-onset diabetes of the young
MOFA	Multi-omics factor analysis
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
NFATC3	Nuclear factor of activated T cells 3
NFG	Normal fasting glucose
NGS	Next generation sequencing
NOS	Not otherwise specified
NPV	Negative predictive value
OPLS-DA	Orthogonal projections to latent structures discriminant analysis
OPLS-EP	Orthogonal projections to latent structures effect projections
OR	Odds ratio
OS	Overall survival
PAAD	Pancreatic adenocarcinoma
PanIN	Pancreatic intraepithelial neoplasia
PCA	Principal component analysis
PDAC	Pancreatic ductal adenocarcinoma
PDX1	Pancreatic and duodenal homeobox 1
PEA	Proximity extension assay
PP	Pancreatic polypeptide
PPV	Positive predictive value
PSC	Pancreatic stellate cells
PTF1A	Pancreas associated transcription factor 1a
qPCR	Real-time polymerase chain reaction
rpm	Reads per million
ROC	Receiver operating characteristic
RT-qPCR	Reverse transcriptase real-time polymerase chain reaction
SEER	Surveillance, Epidemiology, and End Results
SKP1	S-Phase Kinase Associated Protein 1
SMA	Superior mesenteric artery
SMAD4	SMAD Family Member 4
SN	Sensitivity

SP	Specificity
STK11	Serine/threonine kinase 11
TCGA	The cancer genome atlas
TCPA	The cancer proteome atlas
TNM	Tumor-node-metastasis
TP53	Tumor protein p53
TPS	Tissue polypeptide specific antigen
US	United States

List of original papers

Paper I

Borgmästars E, de Weerd HA, Lubovac-Pilav Z, Sund M. miRFA: an automated pipeline for microRNA functional analysis with correlation support from TCGA and TCPA expression data in pancreatic cancer. BMC Bioinformatics, BioMed Central 2019, Vol. 20. (Open Access)

Paper II

Borgmästars E, Lundberg E, Öhlund D, Nyström H, Franklin O, Lundin C, Jonsson P, Sund M. Circulating tissue polypeptide-specific antigen in pre-diagnostic pancreatic cancer samples Cancers, MDPI 2021, Vol. 13, (21). (Open Access)

Paper III

Borgmästars E, Jacobson S, Simm M, Johansson M, Billing O, Lundin C, Nyström H, Öhlund D, Lubovac-Pilav Z, Jonsson P, Franklin O, Sund M. Metabolomics for early pancreatic cancer detection in plasma samples from a Swedish prospective population-based biobank. (Submitted manuscript)

Paper IV

Borgmästars E, Ulfenborg B, Johansson M, Jonsson P, Billing O, Franklin O, Lubovac-Pilav Z, Sund M. Plasma multi-omics in pre-diagnostic pancreatic ductal adenocarcinoma samples from a Swedish prospective biobank. (Manuscript)

Additional papers

Not included in this thesis.

Hagglund M, Backman S, Macellaro A, Lindgren P, **Borgmastars E**, Jacobsson K, Dryselius R, Stenberg P, Sjodin A, Forsman M, Ahlinder J. Accounting for Bacterial Overlap Between Raw Water Communities and Contaminating Sources Improves the Accuracy of Signature-Based Microbial Source Tracking. Frontiers in Microbiology, Frontiers Media S.A. 2018, Vol. 9: 2364.

Borgmästars E, Persson S, Hellmér M, Simonsson M, Eriksson R. Comparison of Skimmed Milk and Lanthanum Flocculation for Concentration of Pathogenic Viruses in Water. Food and Environmental Virology, Springer 2021, Vol. 13, (3) : 380-389.

Jurcevic S, Keane S, **Borgmästars E,** Lubovac-Pilav Z, Ejeskär K. Bioinformatics analysis of miRNAs in the neuroblastoma 11q-deleted region reveals a role of miR-548l in both 11q-deleted and MYCN amplified tumour cells. (In Press Scientific Reports)

Background

In this thesis, I focus on *in silico* analysis of microRNA functions in a pancreatic cancer context as well as metabolites, proteins, and microRNAs in plasma samples from pre-diagnostic pancreatic cancer patients.

Chapter 1 – The pancreas

The pancreas, meaning 'all flesh' in Greek, is an organ involved in food digestion and glucose homeostasis. It was first described in 300 BC by the 'Father of anatomy' Herophilus of Chalcedon (Busnardo et al. 1983). However, it was not until the 19th century when Claude Bernard clarified the role of pancreas in food digestion. Paul Langerhans described the islets of Langerhans in 1869 and in 1893 its role in diabetes was suggested by M.E. Laguesse. In 1922, Frederic Banting and Charles Best treated a diabetic dog with insulin, where the dog recovered from its coma after injection of insulin.

The pancreas is located in the upper abdomen, it is 14-25 cm long and weighs around 100 g. It can be divided into three parts; head (caput), body (corpus) and tail (cauda), and consists of the exocrine and endocrine compartments (**Figure 1**). The bile duct runs through pancreas and fuses with the main pancreatic duct. The fused part is between a few mm to 1 cm long and connects to the duodenum through the major papilla.

1.1. Pancreatic function and cell types

Acinar and ductal cells are the two exocrine cell types, which make up > 95 % of the pancreas. Acinar cells are the most abundant cell type and secrete enzymes involved in food digestion. Acinar cells form clusters of cells, acini, at the end of ducts. The enzymes are stored in acini in so called zymogen granules. Some enzymes are secreted as inactive enzymes, such as trypsin, chymotrypsin, carboxypeptidase, and elastase. Ribonuclease, deoxyribonuclease, amylase, and lipase are released as active enzymes. Ductal cells are located along the pancreatic ducts. The ducts secrete and transport pancreatic juice into the duodenum, where the inactive enzymes are subsequently activated.

The islets of Langerhans constitute the endocrine pancreas (1-2 % of the pancreatic mass) and are located scattered within the pancreas. Langerhans islets contain glucagon-producing alpha cells, insulin-producing beta cells, somatostatin-producing delta cells, ghrelin-producing epsilon cells and pancreatic polypeptide-producing (PP) cells. Insulin lowers blood glucose levels and glucagon increases them. Somatostatin regulates insulin and glucagon secretion. Ghrelin is the 'hunger hormone' that increases appetite, and pancreatic polypeptide regulates pancreatic secretion.

In addition, pancreatic stellate cells (PSC), capillaries, arteries, lymphatics, nerve fibers, fat cells, and veins are found in the pancreas. During inflammation or tissue injury, PSC are activated to form fibroblasts that contribute to fibrosis. This process is central in pancreatic cancer, which is characterized by a rich stroma.

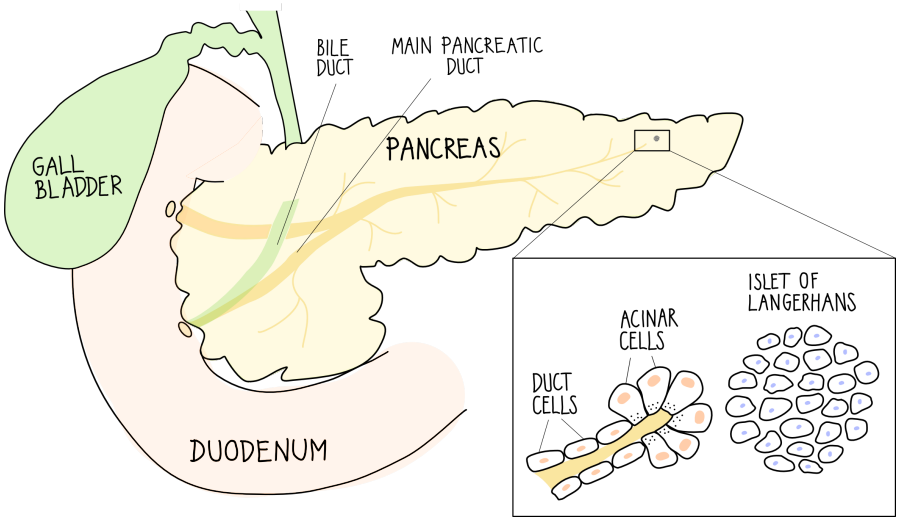

Figure 1. Anatomy of the pancreas.

1.2. Development of the pancreas

Ectoderm, endoderm, and mesoderm are the three germ layers in human embryonic development. Pancreas originates from the endoderm (Lewis and Mao 2018). Initially two buds are independently differentiated from the foregut that later fuse into one pancreatic organ. The ventral bud forms into part of the pancreas head, and the dorsal part into head, body,

and tail of pancreas. The central early transcription factors are pancreatic and duodenal homeobox 1 (PDX1) and pancreas associated transcription factor 1a (PTF1A), as well as signaling pathways, such as the fibroblast growth factor (FGF) signaling and sonic hedgehog signaling in pancreas development.

An undifferentiated tubular network is formed by epithelial cells growing into the surrounding mesenchyme (Lewis and Mao 2018). These tubular networks are premature duct systems with acinar cells located at the end of the branches. Isolated clusters of endocrine cells within the parenchyma are formed from delaminated cells. These clusters proliferate and differentiate into islets of Langerhans. Most endocrine cells identified at 12 weeks of gestation express glucagon and are believed to be alpha cells. Zymogen granules start to appear at week 16, which indicates acinar cell differentiation. The transition from an immature network into a mature ductal system is not fully understood, although WNT signaling has been suggested to have a role (Heiser et al. 2006).

Many of the pathways important in pancreas development do also play a role in pancreas disease or cancer, such as pancreatic ductal adenocarcinoma (PDAC) and maturity-onset diabetes of the young (MODY) (Lewis and Mao 2018). The adult pancreas shows outstanding plasticity in pancreatitis, pancreatic injury, and tumorigenesis.

Chapter 2 – Pancreatic cancer

Cancer is defined as abnormal cell growth and invasion of the basal membrane. Cancer progresses through genetic mutations, classified as driver or passenger mutations. Genetics as well as environmental factors affect the risk of developing cancer. For some cancer types, there is one clear cause of cancer, for instance cervical cancer, which is mainly caused by human papillomavirus infection. Another example is lung cancer, where 80-90 % of patients are smokers or have previously smoked. Fourteen hallmarks or enabling characteristics of cancer have been defined, e.g. deregulating cellular metabolism, sustaining proliferative signaling, evading growth suppressors, and resisting cell death (Hanahan 2022; Hanahan and Weinberg 2000, 2011).

PDAC constitutes the most common (around 90 %) type of pancreatic cancer. In this thesis, the term 'pancreatic cancer' will be used interchangeably with PDAC. Pancreatic cancer is one of the most aggressive malignancies with a 5-year overall survival (OS) of around 9-11 % (Siegel et al. 2022; Wild et al. 2020). In Sweden, the 5-year survival of pancreatic cancer is only about 6 % (Regionala cancercentrum i samverkan 2022a). The relative 5-year survival rate has improved slightly between 1996-2017 from 5 to 11 % in the US (Jemal et al. 2007; Siegel et al. 2022). The reason for the poor prognosis is that PDAC is detected at a late stage with spread disease in most patients. In addition, among the patients with localized disease where curative surgery is an option, the survival rate still remains low at around 40 % (Siegel et al. 2022). Pancreatic ductal adenocarcinoma is the 3rd most common cause of cancer death in the United States (US) but has been estimated to become the second leading cause of cancer deaths in 2026 (Rahib et al. 2021).

2.1. Epidemiology and risk factors

A total of 1387 new cases with a malignant tumor in pancreas were reported in Sweden, in 2020 (Socialstyrelsen 2022). Incidence rates ranged between 1-8 per 100 000, being highest in Western Europe (8.3 per 100 000) and North America (7.6 per 100 000) in 2018 (Wild et al. 2020). Non-modifiable risk factors include higher age, African-American race, non-O blood group, and increased adult height. Most patients are between 60-80 years at diagnosis (Wild et al. 2020). Family history of PDAC and hereditary pancreatitis increase the risk. Modifiable risk factors include smoking, excessive alcohol use, obesity, germ-line mutations, in for instance BRCA2 DNA repair associated (*BRCA2*) or serine/threonine kinase 11 (*STK11*) (Peutz-Jeghers syndrome). Medical conditions such as pancreatitis and diabetes mellitus are also associated with a higher risk of developing pancreatic cancer.

2.2. Symptoms and metabolic changes

Symptoms associated with PDAC are non-specific and emerge late in disease progression. Weight loss, pain, and jaundice are the most common symptoms (**Table 1**). Cachexia is the loss of skeletal muscle mass, which is one explanation for weight loss being a common pancreatic cancer symptom (Kordes et al. 2021). High prevalence of cachexia among PDAC patients can be attributed to systemic metabolic changes, pathogenic

signals produced by the tumor, disrupted pancreatic function as well as the physical proximity between pancreas and the gut. In addition to being a risk factor, diabetes is also a PDAC symptom and only around 14 % of PDAC patients have normal fasting glucose levels (Pannala et al. 2008). Thus, PDAC patients display a severe metabolic shift at diagnosis. In addition, metabolic alterations can emerge before pancreatic cancer diagnosis (Jacobson et al. 2021; Sah et al. 2019; Sharma et al. 2018).

Table 1. Pancreatic cancer symptoms. Modified from the Swedish treatment guidelines for pancreatic cancer (Regionala cancercentrum i samverkan 2021).

Symptom	Frequency
Involuntary weight loss	60-90 %
Pain	50-80 %
Jaundice	50-75 %
Nausea	30 %
Non-hereditary diabetes	5 %
Acute pancreatitis	3 %

2.3. Imaging

Computer tomography (CT) is usually performed on patients with a PDAC suspicion (Regionala cancercentrum i samverkan 2021). If a spread malignancy is not visible on a transabdominal CT, a pancreas-specific CT is performed that consists of two phases; one pancreas- and one venous liver phase. A common feature for tumors in the pancreas head (caput) is a dilated bile and pancreatic duct, known as the 'double duct sign'. This can also be caused by bile stones obstructing the bile duct as well as intraductal papillary mucinous neoplasms (IPMN). Magnetic resonance imaging (MRI) can be used if a clear conclusion cannot be made based on the CT. Use of endoscopic ultrasound (EUS) is encouraged as it has the advantage of finding small pancreatic tumors. Positron emission tomography-CT (PET-CT) is not routinely used in PDAC investigations but can sometimes be used postoperatively to monitor patients with a high risk of developing metastatic disease. Ultrasound can be useful for excluding bile stones as the reason for jaundice. Many asymptomatic PDAC patients are discovered through imaging when other diseases are being investigated.

2.4. PDAC staging

Tumors are classified preoperatively through pancreas-specific CT according to AJCC 8th edition tumor-node-metastasis (TNM) system (**Table 2**) (Brierley 2017). The tumor diameter cutoffs proposed in the 8th AJCC TNM version were shown statistically valid in R0 resected patients and more reproducible compared to the 7th edition (Allen et al. 2017).

Table 2. TNM classification of PDAC (8th edition). Modified from (Brierley 2017).

TNM Stage	Tumor (T)	Lymph node metastasis (N)	Distant Metastasis (M)
IA	< 2 cm (T1)	N0	M0
IB	2-4 cm (T2)	N0	M0
IIA	≥ 4 cm (T3)	N0	M0
IIB	≥ 2 cm (T1-3)	N1	M0
III	≥ 2 cm (T1-3)	N2	M0
III	Tumor involves celiac axis, SMA, and/or hepatic artery (T4)	Any	M0
IV	Any	Any	M1

SMA = superior mesenteric artery, N0 = no lymph node metastasis, N1 = metastasis in one-three lymph nodes, N2 = metastasis in ≥ four lymph nodes, M0 = no distant metastasis, M1 = distant metastasis

2.5. Cancer progression

Pancreatic intraepithelial neoplasia (PanIN), mucinous cystic neoplasm (MCN), intraductal tubulopapillary neoplasms (ITPN), and IPMN are precursor lesions for PDAC. Precancerous lesions can be divided into low- or high-grade lesions (Regionala cancercentrum i samverkan 2021).

The most common precursor pancreatic cancer lesions are PanIN. Telomere shortening and Kirsten rat sarcoma virus (*KRAS*) mutations are early events found in low-grade PanIN (PanIN-1A, -1B or 2) but are not by themselves sufficient for PDAC development (**Figure 2**) (Kanda et al. 2012). Additional mutations are required for progression to PDAC, typically involving cyclin dependent kinase inhibitor 2A (*CDKN2A*), SMAD Family Member 4 (*SMAD4*), and tumor protein p53 (*TP53*)

(Maitra et al. 2003; Waters and Der 2018). Overall, at least 90 % of PDAC tumors contain activating *KRAS* mutations.

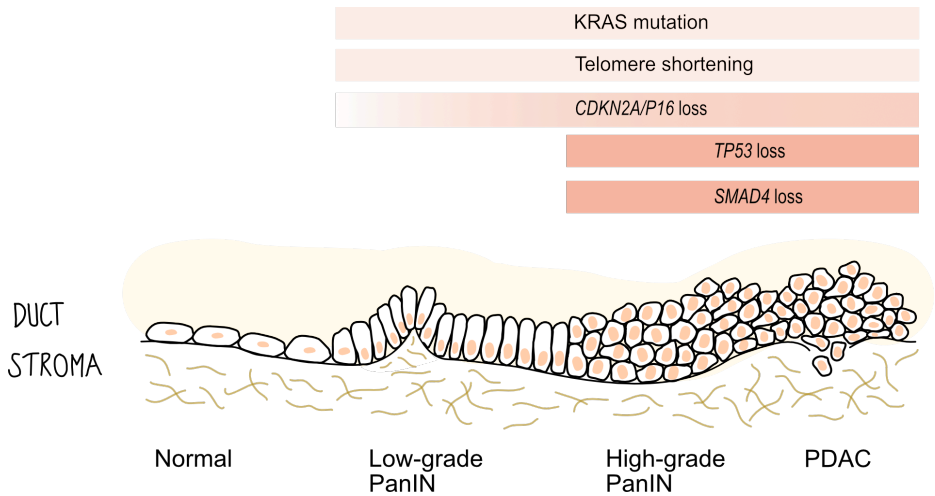

Figure 2. Molecular hallmarks from pancreatic intraepithelial neoplasia (PanIN) to pancreatic ductal adeno-carcinoma (PDAC) progression. Modified from (Maitra et al. 2003; Noë et al. 2018).

The IPMN are divided into main duct-IPMN and branch-duct IPMN, where IPMN found in the main duct increase the risk of developing PDAC. Many IPMN harbor early *KRAS* or GNAS complex locus (*GNAS*) mutations. MCN develop mostly in the pancreas body or tail in women and have ovarian-like stroma. ITPN is a type of rare intraductal neoplasm.

Some efforts have been made to understand progression time to PDAC by using mutation rates or patients' age and stage at diagnosis (**Figure 3**). A tumor evolution model based on sequencing data from seven primary pancreatic cancers and paired metastases suggested a time frame of at least ten years from initiating tumor to a founder cancer cell, and at least another five years for the establishment of metastatic lesions (Yachida et al. 2010). From the establishment of metastasis, Yachida et al. estimated another two years on average before the patients die. Genomic characterization of matched IPMN and PDAC suggested an average progression rate of more than three years from high-grade precursor lesions to pancreatic cancer (Noë et al. 2020). A simulation model, using information from the National Cancer Institute's Surveillance,

Epidemiology and End Results (SEER) database and reported PanIN prevalence by age, predicted 9.5 years progression time on average between high grade-PanIN and PDAC (Peters et al. 2018). Altogether, these studies indicate that precancerous lesions that later develop into PDAC can be detected early.

Yu et al. compared the mean age at diagnosis in the SEER database stratified by tumor stage (Yu et al. 2015). They found stage IV patients to be on average 1.3 years (adjusted by sex, ethnicity, tumor location, and neoplastic grade) older than stage I patients. This suggests a rapid disease course from low to high PDAC tumor stages. The two models by Yachida et al. and Yu et al. are slightly different but not necessarily contradictive since the rapid PDAC progression predicted by Yu et al. can fit into the cancer-metastatic phase predicted by Yachida et al. (Yachida et al. 2010; Yu et al. 2015). As mentioned, early-stage patients can have a short post-operative survival, which suggests that micro-metastatic disease is already established and thus the progression from stage I to stage IV occurs rapidly and is consistent with the PDAC progression model by Yu et al. In addition, the time until stage I is not taken into account in the model by Yu et al. as this is a cross-sectional study (Gallmeier et al. 2015). In addition to a gradual progression, chromothripsis has been shown to be a common event in PDAC, which leads to a fast catastrophic event by shattered chromosomes (Cortes-Ciriano et al. 2020; Notta et al. 2016). These studies estimated chromothripsis events occurring in 56-65 % of PDAC tumors.

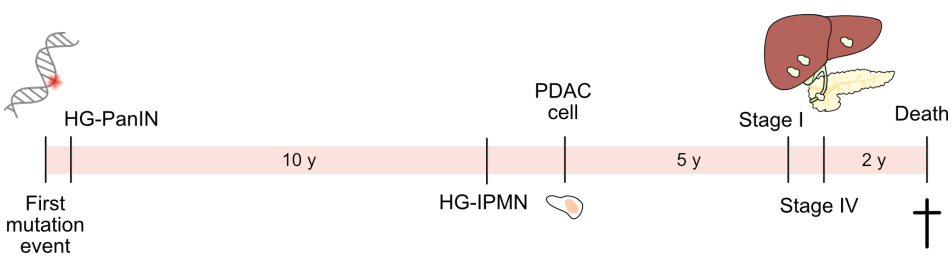

Figure 3. Models for PDAC progression time. Estimated progression time between high grade-PanIN or -IPMN to PDAC, between the first mutation event and death, as well as between PDAC stages I and IV (Noë et al. 2020; Peters et al. 2018; Yachida et al. 2010; Yu et al. 2015).

2.6. Surgical Treatment

Radical surgery in combination with chemotherapy is needed to achieve long-term survival in PDAC patients. The first recorded one-stage pancreaticoduodenectomy, also known as Whipple's procedure, dates back to 1940 and was performed by Whipple and Nelson (Busnardo et al. 1983). Pancreatic surgery is a major surgery with a long recovery time and high morbidity (Regionala cancercentrum i samverkan 2021). Unfortunately, due to late disease stage discovery, only 20-30 % of patients have resectable disease at diagnosis. In addition, curative surgery is cancelled in around 11 % of resectable patients due to discovery of metastatic disease or locally advanced PDAC during surgery.

There are three main surgical strategies, Whipple's procedure, distal pancreatectomy and total pancreatectomy, and the choice between these depends on the size and location of the PDAC tumor. The Whipple's procedure (pancreatoduodenectomy) is the most common PDAC surgery performed on pancreas head tumors, where the pancreas head, gall bladder, duodenum and part of the bile ducts are removed (**Figure 4A**). The small intestine is anastomosed to the stomach and the remnant pancreas so that digestive enzymes can still be secreted into the small intestine after surgery (**Figure 4B**). A Whipple's procedure can also be performed for patients with main-duct IPMN.

Distal pancreatectomy is the removal of the body and tail of the pancreas, and often together with the spleen (and sometimes left kidney, left diaphragm and left adrenal gland). Benign or low-malignant lesions can be surgically removed by laparoscopy and has been shown to benefit postoperative recovery.

Total pancreatectomy is the removal of the whole pancreas, and sometimes the bile duct, spleen, gallbladder, part of small intestine and surrounding lymph nodes are removed as well, depending on the tumor spread. The patients will be severely diabetic after a total pancreatectomy and needs to take insulin and digestive enzymes for the rest of their lives. Total pancreatectomy is recommended for patients with multi-focal PDAC, widespread tumors, widespread IPMN and those with a high risk of pancreatic leakage.

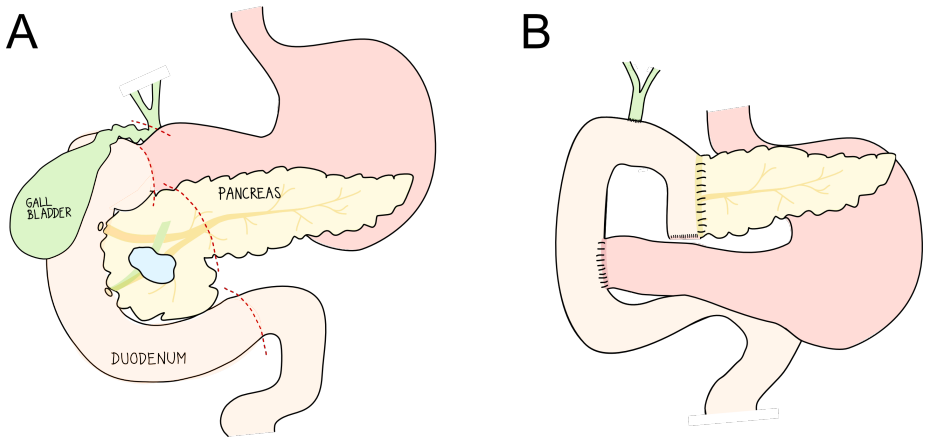

Figure 4. Illustration of Whipple's procedure. A) The pancreas head is removed as well as the gall bladder and part of the duodenum. B) The remaining pancreas and stomach are anastomosed with the small intestine.

Postoperative complications in PDAC surgery are common, especially after Whipple's procedure, which is associated with an in-hospital mortality incidence of 3 % (Merath et al. 2020) and complications in 30-40 % of resected patients (Regionala cancercentrum i samverkan 2021). The patient often needs to take digestive enzymes and hormonal supplements to control food digestion and blood glucose levels. Leakage in the pancreas-jejunum anastomosis occurs in 2-30 % of operated patients and is the leading cause of postoperative death (Regionala cancercentrum i samverkan 2021). Some patients show long-term survival after radical surgery, whereas others undergo extensive operations with questionable benefit. Thus, methods for the prediction of surgical benefit and better selection of patients for surgery are needed.

2.7. Oncological treatment

Adjuvant chemotherapy is routinely administered after PDAC surgery (Regionala cancercentrum i samverkan 2017). Palliative chemotherapy or best supportive care is given to unresectable patients.

First line adjuvant treatment is FOLFIRINOX or gemcitabine-capecitabine. Randomized controlled trials have concluded that this treatment is superior to single-agent chemotherapy or observation only (Conroy et al. 2018; Neoptolemos et al. 2017). Standard palliative

treatment is FOLFIRINOX or gemcitabine in combination with nab-paclitaxel (Regionala cancercentrum i samverkan 2021).

Administering neoadjuvant therapy will delay surgery, and due to the aggressive nature of PDAC, it can be argued to perform the surgery as soon as possible to prevent further tumor progression. At the same time, neoadjuvant chemotherapy could be a mean to select truly resectable patients for surgery while treating micro-metastatic disease. Neoadjuvant chemotherapy is currently recommended to borderline resectable and locally advanced PDAC to achieve resectability. However, evidence and use of neoadjuvant chemotherapy in PDAC is increasing, especially in the US (Aquina et al. 2021). Long-term results from the PREOPANC trial concluded that neoadjuvant single-agent chemotherapy in combination with radiotherapy improved overall survival (OS) compared to upfront surgery, with 5-year OS rates of 20.5 % and 6.5 %, respectively (Versteijne et al. 2022). However, this difference was mainly driven by included borderline resectable patients. Studies evaluating multiagent neoadjuvant chemotherapy in upfront resectable patients are ongoing (NorPACT-1, PREOPANC-3, and NEOPAC) with mature data expected soon.

The oncological treatments have a high toxicity and can give severe side effects for the patients. Thus, biomarkers for predicting treatment response would be very valuable to avoid unnecessary side-effects and restrict administration to those patients that will benefit from oncological treatment.

2.8. Radiotherapy

Radiotherapy combined with chemotherapy, radiochemotherapy, is not routinely used in PDAC treatment in Sweden (Regionala cancercentrum i samverkan 2021). It is only applied within clinical trials focused on converting borderline resectable tumors to resectable (as mentioned in section '2.7. Oncological treatment').

2.9. Histopathology

Pancreatic cancer is characterized by duct-like structures grown in a haphazard pattern surrounded by a dense stroma (**Figure 5**) (Pittman and Hruban 2018). PDAC cells often contain intracytoplasmic mucin and are shaped as columnar or cuboidal cells. Perineural and intravascular

invasion are common features in PDAC. Tumors that resemble normal ducts are considered 'well-differentiated', whereas poorly formed glands are categorized as 'poorly differentiated'.

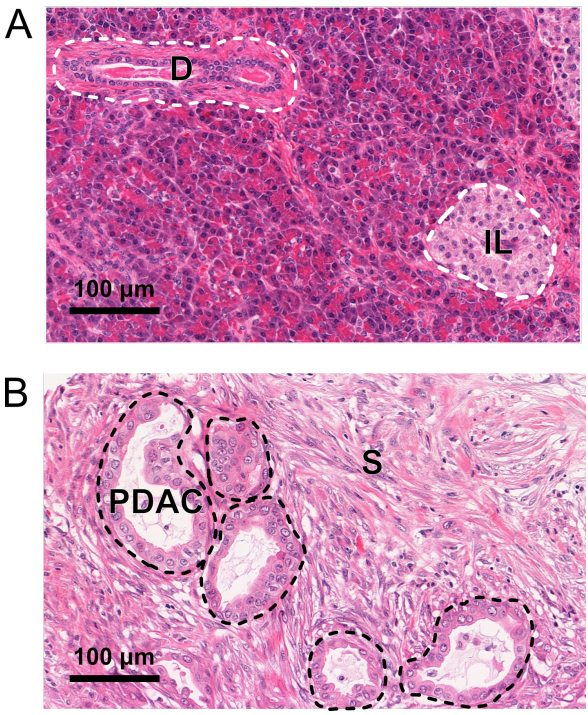

Figure 5. Histology of the pancreas. A) Normal pancreas tissue showing a duct (D) and an islet of Langerhans (IL). B) Pancreatic ductal adenocarcinoma (PDAC) with surrounding stroma (S). Scale bar = 100 µm. Image courtesy of Anette Berglund.

Chapter 3 – Early detection of cancer

Identifying a tumor early is essential for a good patient survival. According to World Health Organization (WHO), early detection can be divided into two strategies: early diagnosis and screening.

Early diagnosis concerns setting a correct diagnosis in individuals that seek medical care after onset of symptoms. A challenge in PDAC management is that many individuals are asymptomatic and have an advanced disease by the time they develop symptoms. However, some individuals develop symptoms and can still be offered curative surgery.

Screening programs are used to detect a cancer in asymptomatic, high-risk individuals. National screening programs for cervical cancer and breast cancer are currently performed in Sweden. Screening for colorectal cancer is established in some Swedish regions with ongoing implementation in additional regions. Cervical cancer patients have decreased from 25 to 8.4 per 100 000 in Sweden since 1965 when the screening program started (Regionala cancercentrum i samverkan 2022b). PDAC-screening is limited today, mainly because of the low incidence, but also because it is challenging to define a high-risk group to include in a screening setting. Patients with an increased PDAC risk, due to having a risk-increasing syndrome (e.g. Peutz-Jeghers) or with familial PDAC, are monitored from the age of 50 or 10 years prior to when the index case developed PDAC. Yearly surveillance is also recommended for individuals with high-risk cysts. One risk of screening is over-diagnosis, especially for PDAC where curative surgery is a major intervention.

Chapter 4 – Biomarkers in pancreatic cancer

4.1. What is a biomarker?

A biomarker is defined by WHO as "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" (WHO 2001). It can be measured in blood, tissue, urine, stool, or other body fluids. There are different types of biomarkers and these can be divided into the following five categories; diagnostic, prognostic, predictive, monitoring, and screening biomarkers. Depending on the type of biomarker studied, different criteria are needed. For a diagnostic biomarker, there must be a good discriminative ability between individuals with a specific disease and healthy controls. A good prognostic biomarker holds information regarding survival of the patients and thus the biomarker is mainly assessed among the patients. To assess the response to a certain treatment, good predictive biomarkers are needed. A monitoring biomarker can be used in the follow-up of patients to study the response to treatment or potential disease relapse. Screening biomarkers can be used with the purpose of finding a disease at an early, curable stage.

How do we decide if a biomarker is good or not? Biomarker performance can be assessed by various measures such as sensitivity (also called recall),

specificity, positive predictive value (PPV, also called precision), negative predictive value (NPV), and accuracy (i.e. fraction of correctly classified samples) (**Figure 6**). An area under the receiver operating characteristic (ROC) curve (AUC) is a combination of two performance parameters, e.g. sensitivity and specificity where an AUC of 0.5 has similar performance as a random classifier and AUC of 1.0 represents a perfect classifier (100 % of all samples are correctly classified). These performance parameters can then be compared for the newly identified biomarker(s) and compared to the currently used biomarker or other suitable reference biomarkers.

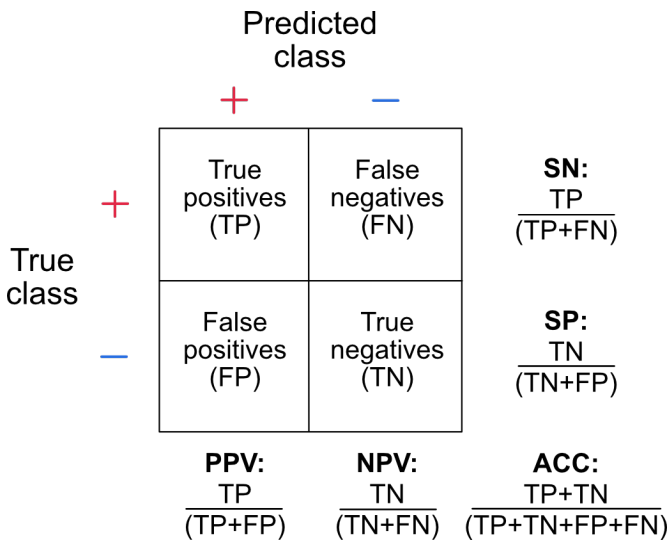

Figure 6. Performance assessment of a biomarker. Red plus sign and blue minus sign represent two different conditions, e.g. disease and control. SN = sensitivity, SP = specificity, PPV = positive predictive value, NPV = negative predictive value, ACC = accuracy.

4.2. Clinically used biomarkers in PDAC

4.2.1. CA 19-9

Carbohydrate antigen 19-9 (CA 19-9), a sialyl-Lewis A antigen expressed on cell surfaces, is the gold standard biomarker in PDAC and new potential biomarkers are often compared to CA 19-9 in terms of performance. CA 19-9 is utilized for PDAC diagnostics in symptomatic patients but has some disadvantages. It does not have an optimal sensitivity since 10 % of the Caucasian population lacks expression of the

Lewis antigen. Furthermore, the specificity is suboptimal since it can be elevated in other gastrointestinal diseases (Poruk et al. 2013).

In some individuals, CA 19-9 starts to increase between 2-3 years before PDAC diagnosis (Honda et al. 2019; Mason et al. 2022; O'Brien et al. 2015). However, considerable CA 19-9 elevations are noted very close to diagnosis, limiting its use as biomarker for early PDAC detection (Mason et al. 2022). CA 19-9 does not hold promise as a screening biomarker in asymptomatic individuals due to low positive predictive values (0.5-0.9 %) using a circulating CA 19-9 cutoff > 37 U/mL (Chang et al. 2006; Kim et al. 2004). However, among symptomatic patients presenting with abdominal complaints, screening was concluded as effective in finding patients eligible for pancreatic cancer resection (Satake et al. 1994). CA 19-9 has also been indicated as a useful pre- and post-operative follow-up marker and as a prognostic biomarker in several studies as reviewed in (Ballehaninna and Chamberlain 2012).

4.2.2. CEA

The glycoprotein carcinoembryonic antigen (CEA) is involved in cell adhesion and mainly used as a colorectal cancer biomarker. A meta-analysis found CEA to have lower sensitivity but slightly higher specificity in separating benign and malignant pancreatic disease as compared to CA 19-9 (Poruk et al. 2013). A more recent study found CEA to be a more robust predictor of advanced PDAC than CA 19-9 in 214 patients with suspected PDAC (van Manen et al. 2020). Furthermore, combining CEA and CA 19-9 had a higher PPV than either biomarker alone of predicting advanced PDAC (91.4%).

Chapter 5 – MicroRNAs

MicroRNAs (miRNAs) are small non-coding RNAs of around 22 nucleotides that act as post-transcriptional regulators by binding to messenger RNA (mRNA) (Bhaskaran and Mohan 2014). MicroRNAs originate from a miRNA gene that, through different hairpin precursors, are formed into two mature miRNA isoforms, termed -3p and -5p arms (**Figure 7**). Usually, one of the mature miRNAs plays a role in post-transcriptional regulation, termed the guide strand, whereas the other strand referred to as passenger strand is degraded. However, sometimes both miRNA strands can be involved in miRNA-mediated regulation.

MicroRNAs are mostly known for degrading or repressing their target mRNAs, but a few studies suggest an up-regulating role of miRNAs in some cell conditions (Rusk 2008; Vasudevan et al. 2007). MiRNAs can be detected and quantified by reverse transcription-real time polymerase chain reaction (RT-qPCR), microarrays or next-generation sequencing (NGS) (Git et al. 2010).

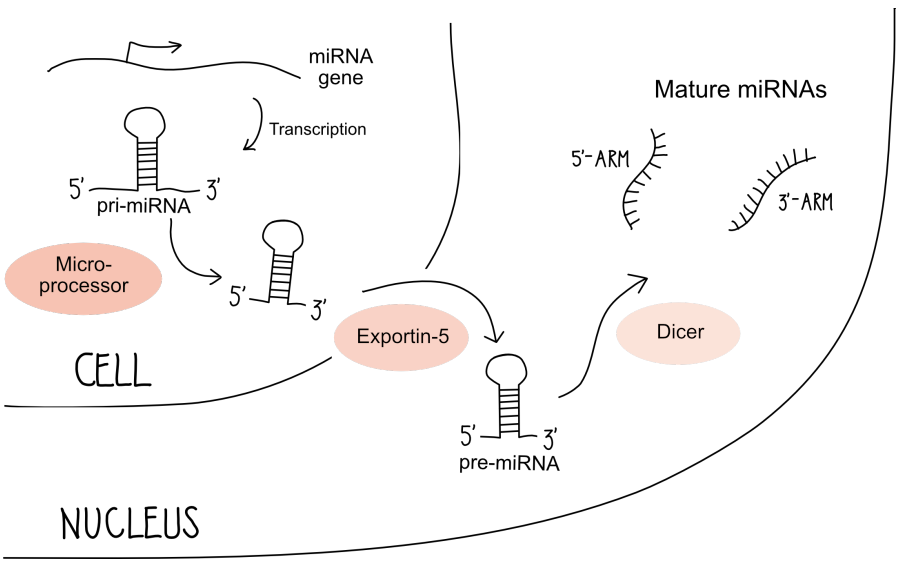

Figure 7. Biogenesis of microRNAs. A miRNA gene is transcribed and a pri-miRNA is formed. Microprocessor complex is involved in forming pre-miRNA, which is transported into the cytoplasm by Exportin-5. Finally, two mature miRNA strands, 3'- and 5'-arms, are formed by the Dicer enzyme (Bhaskaran and Mohan 2014).

5.1. Functional analysis of miRNAs

In silico-based functional analysis of miRNAs typically consists of an indirect annotation comprising 1) miRNA target prediction followed by 2) functional enrichment of predicted targets.

5.1.1. MiRNA target prediction

There are numerous miRNA target prediction resources available today, including miRNA target prediction tools and databases collecting experimentally validated miRNA targets. The prediction tools utilize different algorithms and parameters, such as seed region match, interaction site accessibility, free energy and conservation, in their miRNA target prediction (Peterson et al. 2014). There are over 75

databases available and a lot of these use combinations of existing tools (Tokar et al. 2018). Most tools search the 3'-untranslated region for miRNA targets, for instance miRanda (Enright et al. 2003; John et al. 2004), microRNA.org (Betel et al. 2008), miRDB (Wong and Wang 2015), DIANA-microT (Paraskevopoulou et al. 2013) and TargetScan (Agarwal et al. 2015). DIANA-microT-CDS algorithm also searches the coding sequence for potential miRNA binding regions (Reczko et al. 2012).

5.1.2. Functional enrichment

Functional enrichment tools have been developed to understand long gene lists. Predicted miRNA target genes can be used as input in these tools to find over-represented functions connected to a specific miRNA. There are a few functional enrichment tools available such as DAVID (Huang da et al. 2009a, 2009b), GOMA (Huang et al. 2013), g:Profiler (Raudvere et al. 2019) and FunRich (Pathan et al. 2015). In addition, the R package edgeR provides functions for analyzing over-representation of gene ontology (GO) terms and Kyoto Encyclopedia of genes and genomes (KEGG) pathways (McCarthy et al. 2012; Robinson et al. 2010). The online tool DIANA-miRPath combines miRNA target prediction, using two experimentally validated databases and one experimentally validated database, and functional enrichment for miRNA functional analysis (Vlachos, Zagganas, et al. 2015). The user can identify predicted miRNA targets and functionally enriched GO terms and KEGG pathways from the same resource, instead of using another tool for functional enrichment of a list of miRNA targets. MiRNet is another tool for microRNA functional analysis (Chang et al. 2020).

5.1.3. Validation

One problem with *in silico*-based functional analysis of miRNAs is the large number of false positive targets generated by prediction algorithms and that the expression or miRNA-mRNA interactions might differ between tissues or disease states (Singh 2017; Wu et al. 2017). The most optimal method to validate miRNA targets is experimental validation, however this approach is not possible for a large number of predicted miRNA targets. Functional analysis of miRNAs can be performed *in vitro* by knockdown/knockin to increase or decrease levels of a certain miRNA. Another method is using databases that collects experimentally validated miRNA-mRNA interactions from the literature, such as DIANA-Tarbase

(Karagkouni et al. 2018), miRecords (Xiao et al. 2009), miRTarBase (Chou et al. 2018) or miRWalk (Sticht et al. 2018).

5.2. MicroRNAs in pancreatic cancer

MicroRNA levels can be altered in cancerous tissues. Both tumor suppressors, e.g. hsa-miR-141 and hsa-miR-200, and oncogenic miRNAs (onco-miRs) such as hsa-miR-21 and hsa-miR-31, have been identified in pancreatic cancer (Baradaran et al. 2019). MicroRNAs are suitable as circulating biomarkers due to their remarkable stability in blood, and previous studies have identified differentially altered circulating microRNAs in PDAC (Franklin et al. 2018; Hussein et al. 2017). The stability of miRNAs could be due to protection from degradation by exosomes, associations to proteins or miRNA modifications making them more resistant to degradation (Mitchell et al. 2008).

Franklin and co-workers previously identified a 15-miRNA signature with an AUC of 0.96 at PDAC diagnosis, which outperformed CA 19-9 (AUC = 0.92) (Franklin et al. 2018). However, this 15-miRNA signal as well as the CA 19-9 signal were lost in a pre-diagnostic cohort including blood samples from future PDAC patients up to ten years prior to diagnosis (AUC ≤ 0.65). A pre-diagnostic PDAC cohort derived from the European Prospective Investigation into Cancer and Nutrition (EPIC) study was screened for eight selected miRNAs (Duell et al. 2017). This 8-miRNA panel was selected based on previous studies investigating overexpressed miRNAs in PDAC. No clear biomarker with potential for early PDAC diagnosis emerged, but the three most promising miRNAs were hsa-miR-21-5p, hsa-miR-30c, and hsa-miR-10b. The best performing AUCs were observed within 5 years prior to a PDAC diagnosis and these were 0.79 (hsa-miR-21-5p) and 0.77 (hsa-miR-30c). Recently, a panel of two miRNAs, hsa-miR-33a-3p + hsa-miR-320a, in combination with CA 19-9 has been suggested to have a potential role in early detection of PDAC as this signature was found to discriminate PDAC and patients with IPMN from healthy controls with an AUC of 0.95 (Vila-Navarro et al. 2019).

Chapter 6 – Metabolomics in pancreatic cancer

Metabolomics involves studying small molecules resulting from metabolic processes. A meta-analysis of 25 previous metabolomics studies in PDAC shows that most studies have found altered amino acid pathways (Long et

al. 2018). The most commonly altered metabolites found are glutamic acid and histidine (found in seven studies), as well as glutamine and isoleucine (found in five studies). Due to different sample sizes and lack of external validations, there were a lot of inconsistencies between the studies. Of 25 included studies, only nine included any kind of external validations of their findings.

In 2018, a 9-metabolite profile + CA 19-9 was identified excellent in distinguishing between PDAC and chronic pancreatitis (CP) with an AUC of 0.94 (Mayerle et al. 2018). To make it more translatable into clinical use, an improved signature was derived, resulting in a minimalistic (m)-metabolic signature of CA 19-9 and four metabolites (ceramide [d18:1, C24:0], lysophosphatidylethanolamine [C18:9], phosphatidylethanolamine [C18:0,C22:6], and sphingomyelin [d17:1,C16:0]) (Mahajan et al. 2022). The m-metabolic signature was excellent at distinguishing between individuals with PDAC, chronic pancreatitis (CP) and non-pancreatic disease controls (NPC) in Lewis-positive patients with an AUC of 0.924. A 12-metabolite signature was also applicable to Lewis-antigen negative subjects that do not express CA 19-9 and displayed an AUC of 0.805. The 9-metabolic signature, termed MxP® PancreasScore, will be validated in a prospective study of individuals with undefined pancreatic mass on imaging (Deutsche Register Klinischer Studien registration ID: DRKS00010866). Furthermore, validation of the m-metabolic signature in pre-diagnostic individuals with a pancreatic mass lesion is underway (Mahajan et al. 2022).

Analyses in pre-diagnostic pancreatic cancer cohorts with regards to altered metabolites have previously been performed. Branched-chain amino acids (BCAA; isoleucine, leucine and valine) were found up-regulated in PDAC patients, with the strongest association found at 2-5 years prior to PDAC diagnosis (Mayers et al. 2014). The findings were supported by elevated BCAA in early PDAC progression in mouse models with *KRAS*-driven tumors, and muscle catabolism was suggested as the source of the elevated BCAA. Circulating BCAAs were validated in a separate Japanese cohort, where the strongest association was observed ≥ 10 years prior to PDAC diagnosis (Katagiri et al. 2018). These findings were not replicated by Fest et al., who instead found downregulated histidine and glutamine associated with an increased PDAC risk (Fest et

al. 2019). However these were not statistically significant after multiple hypothesis correction.

Chapter 7 – Proteins in pancreatic cancer

7.1. Circulating TPS

Tissue polypeptide specific antigen (TPS) is a fragment of cytokeratin 18. Using cut-offs of 37 kiloarbitrary units (kAU)/L for CA 19-9 and 40 AU/L for TPS resulted in higher sensitivity of TPS, but lower specificity compared to CA 19-9 with respect to benign diseases (22 % for TPS; 60 % for CA 19-9) and a control group of blood donors (88 % for TPS; 100 % for CA 19-9) (Banfi et al. 1993). It has been suggested to have a better discriminative ability between PDAC and chronic pancreatitis compared to CA 19-9 (Slesak et al. 2000). Furthermore, TPS was differentially altered when comparing PDAC and benign hepatopancreabiliary diseases but had low discriminative ability (Pasanen et al. 1994). TPS has been suggested to be more suitable for monitoring the clinical status in post-operative PDAC patients compared to CA 19-9 (Slesak et al. 2004). However, a later study found no diagnostic value of TPS in discriminating between pancreatobiliary diseases (Ozkan et al. 2011).

7.2. Circulating protein biomarkers

IMMray® (Immunovia, Inc.) is a platform of 349 human recombinant antibodies targeting 156 antigens that correspond to systemic disease response and tumor secretome (Mellby et al. 2018). This platform was utilized to develop a 29-biomarker signature that displayed an AUC of 0.963 for separating PDAC stages I & II from normal controls in a validation cohort. In 2022, the signature was refined to include 8-biomarkers + CA 19-9, which was named the PanCan-d test (Brand et al. 2022). PanCan-d outperformed CA 19-9 alone (specificity 97.6 %, sensitivity 75.8 %) for PDAC stages I & II compared to a high-risk cohort (PanFAM, PanCan-d: specificity 98 %, sensitivity 85 %) or healthy controls (PanCan-d: specificity 99 %, sensitivity 85 %). Analyzing the 8-biomarkers without CA 19-9 in individuals with CA 19-9 levels < 2 U/mL generated an AUC of 0.874.

Chapter 8 – Multi-omics profiling

Systems biology aims to understand complex biological systems by studying interactions between biological components. With high-throughput analyses dropping in price, it is nowadays feasible to generate omics data at multiple levels. Multi-omics consists of combining data from at least two different omics levels. There are various methods to analyze multiple omics levels together. The most straightforward way is looking at data modalities separately or by concatenating all variables into the same matrix (Singh et al. 2019). Unsupervised and supervised methods for multi-omics data also exist to include the covariance between different omics levels (Argelaguet et al. 2020; Rohart et al. 2017; Singh et al. 2019).

Unsupervised analyses do not label the outcome and can be performed to look at the overall variation in the data. It can be used to look at whether confounders are main contributors to certain variation or if outliers exist. Principal component analysis (PCA) is a linear method widely used to get an overview of data by reducing dimensions. Multi-omics factor analysis (MOFA) is an extension of PCA for multi-omics data (Argelaguet et al. 2020). It allows us to understand how much each omics layer contributes to the latent factors (equivalent to principal components) and to identify the features with highest weights in the latent factors.

Supervised analysis is performed to find the best separation on the outcome or conditions of interest, such as disease versus healthy. Data integration analysis for biomarker discovery using latent components (DIABLO) is a supervised method that allows disentangling the different omics layers, compare them and see how they correlate indirectly through the components (Rohart et al. 2017; Singh et al. 2019). One limitation of DIABLO is that the user has to specify the number of components and variables to include in the model. An algorithm is available to find the most appropriate parameter tunings by splitting the cohort into train and test data. However, this can be challenging for small datasets.

8.1. Multi-omics biomarkers

CancerSEEK consists of a panel of eight circulating proteins and mutations in 1933 genomic positions that detects eight different cancer types, including pancreatic cancer (Cohen et al. 2018). The mean

sensitivity was 62 % at > 99 % specificity for all cancer types combined. The highest sensitivities, in all eight cancer types combined, were achieved in cancer stages II & III (75-78 %), whereas sensitivity dropped to 40 % in stage I. Larger prospective cohorts will be needed to further evaluate the clinical utility of CancerSEEK.

There is an ongoing observational clinical trial named the DAYBREAK study with the aim to separate between pancreatic cancer and benign pancreatic disease by a multi-omics profile in blood (ClinicalTrials.gov Identifier: NCT05495685). The omics modalities include, but are not restricted to, cell-free DNA (cfDNA) methylation, blood miRNAs, serum proteins in 450 participants. The study is currently (October 2022) recruiting.

In translational research and for discovery of biomarkers that would eventually be implemented in clinics, the cost-efficiency question is central. If the best biomarker signature spans a multi-omics panel, then it has to be sufficiently good to be worth implementing. However, it might also be that multi-omics can be used as a first screening to gain more biological insights and generate hypotheses. Eventually, a biomarker signature using only one omics layer consisting of a few biomarkers might be derived with clinical value.

Aims

Overall aim

The overall aim of this thesis is to understand the biological functions of circulating microRNAs in pancreatic cancer. Moreover, to identify circulating biomarkers with a potential use in early pancreatic cancer detection to be able to offer curative surgery to more patients and subsequently improve patient survival.

Specific aims

- To build a tool for *in silico* functional analysis of microRNAs in pancreatic cancer with correlation support of predicted microRNA targets

  (Paper I)

- To determine the use of TPS as a biomarker for early pancreatic cancer detection

  (Paper II)

- To identify biomarkers for early pancreatic cancer detection by multi-omics, including microRNAs, metabolites, and proteins

  (Papers III & IV)

- To assess the prognostic value of circulating metabolites in pre-diagnostic plasma samples of individuals that later develop pancreatic cancer

  (Paper III)

Materials and Methods

In this thesis, publicly available expression data derived from pancreatic cancer tissue and plasma samples from pre-diagnostic or diagnostic pancreatic cancer was used (**Table 3**).

Table 3. Overview of materials and methods in this thesis.

	Paper			
	I	**II**	**III**	**IV**
Study design	Software development	Nested case-control study	Nested case-control study	Nested case-control study
Data	MiRNA, mRNA, protein expression (public data)	TPS	Metabolomics, clinical biomarkers	Metabolomics, miRNomics, proteomics, clinical biomarkers
Sample size	183 PDAC patients	267 future PDAC, 26 PDAC, 328 controls	100 future PDAC, 100 controls	39 future PDAC, 39 controls

TPS = tissue polypeptide specific antigen, miRNA = microRNA, mRNA = messenger RNA

Chapter 9 – Patient cohorts and characteristics

9.1. Ethics statement

Public data was used in **Paper I**. Remaining studies in this thesis were approved by the ethical committee at Umeå University according to the Helsinki Declaration of 1975. Written informed consent was given by participants at inclusion into the Northern Sweden Health and Disease Study (NSHDS). Pre-diagnostic PDAC cohorts were subsequently obtained from NSHDS and used for Papers II-IV. Participants in the diagnostic PDAC cohort gave their written informed consent before inclusion into the biobank at the Department of Surgery.

9.2. TCGA and TCPA

We used miRNA, mRNA, and protein expression from the cancer genome atlas (TCGA) pancreatic adenocarcinoma (PAAD) generated by the TCGA Research Network (http://cancergenome.nih.gov/) and the cancer proteome atlas (TCPA)-PAAD projects (Li et al. 2013). MiRNA isoform quantification data from the GDC portal (https://portal.gdc.cancer.gov/) and mRNA expression data from the Xena browser (https://xenabrowser.net/datapages/) were downloaded (Goldman et al. 2020). Mature miRNA expression was annotated from the miRNA isoform quantification by summarizing values ≥ 1 reads per million (rpm) for each miRNA using plyr R package (Wickham 2011).

Protein expression was derived from the TCPA-PAAD project (http://tcpaportal.org/tcpa/download.html) (Li et al. 2013). The list of 15 differentially expressed miRNAs at PDAC diagnosis by Franklin et al. was used (Franklin et al. 2018) to demonstrate the functionality of the *in silico* miRNA functional analysis (miRFA) pipeline developed in **Paper I** and to compare it to miRCancerdb (Ahmed et al. 2018).

9.3. Pre-diagnostic cohorts

Pre-diagnostic PDAC plasma samples and matched healthy controls were withdrawn from NSHDS. Two healthy controls without malignancy were matched to the first sampling occasion of the future PDAC patient by sex, age (± 6 months), and sample date (± 6 months) in the first withdrawal (**Paper II**). A second withdrawal was performed in the same way to include additional diagnosed PDAC patients. A matching procedure within the withdrawn cohort was performed by matching one healthy control to each PDAC sample by sex, age (± 6 years), and sampling date (± 6 years) up to 5.2 years lag-time to diagnosis (**Papers III & IV**). There was two exceptions, for one case the sample date differed 9 years and for another case age differed 15 years.

9.4. Pre-diagnostic symptoms

Pancreatic cancer-related symptoms were reviewed for PDAC patients in the cohort used in **Paper III** up to six years prior to diagnosis from medical records in orthopedic, surgical, medical, and health care centers. Symptoms reviewed included back pain, abdominal pain, jaundice, new-onset diabetes, diarrhea, weight loss, gallstone, pancreatitis, and fatigue.

Inclusion and exclusion criteria of the reviewed symptoms are listed in **Paper III, Supplementary Table 1**.

9.5. Diagnostic cohorts

Plasma samples at PDAC diagnosis before surgical or oncological treatment were collected at the Department of Surgery, Umeå University Hospital (**Paper II**).

Chapter 10 – Bioinformatics

10.1. MiRNA functional analysis pipeline

The miRNA functional analysis (miRFA) pipeline was built in R project for statistical computing (R Core Team 2021), where one miRNA is queried separately and consists of the following steps:

1) MiRNA target prediction
2) Correlation analyses of the predicted miRNA target between miRNA-mRNA expression and miRNA-protein expression
3) Functional enrichment analysis of correlated predicted miRNA targets

MicroRNA target prediction was done using two prediction databases; DIANA-microT-CDS (Reczko et al. 2012) and TargetScan version 7.1 (Agarwal et al. 2015), as well as the experimentally validated database DIANA-TarBase version 7 (Vlachos, Paraskevopoulou, et al. 2015). DIANA-microT-CDS was downloaded from http://diana.imis.athena-innovation.gr/DianaTools/index.php and a prediction score threshold of 0.7 was used. DIANA-TarBase v7 was downloaded from http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index. We downloaded predicted targets for conserved sites for miRNAs, conserved miRNA families, and predicted non-conserved sites miRNAs for TargetScan (http://www.targetscan.org/vert_71/). The downloaded miRNA prediction databases were combined with miRNA, mRNA, and protein expression data into an sqlite database, which was queried from R using the RSQLite package (Müller et al. 2022). Over-representation analysis of KEGG pathways and GO terms was performed using the edgeR package (McCarthy et al. 2012; Robinson et al. 2010).

10.2. Network analysis

Networks are displayed by nodes interconnected by edges (**Figure 8**). The nodes with a high number of direct neighbors are referred to as hubs (Barabasi and Oltvai 2004). Protein-protein interaction networks were generated in the STRING database (https://string-db.org/) (Szklarczyk et al. 2021). The network was analyzed in Cytoscape software (Shannon et al. 2003) and hub genes identified by the cytohubba plugin (Chin et al. 2014). CluePedia plugin was used to visualize the overlap between different KEGG pathways (Bindea et al. 2013).

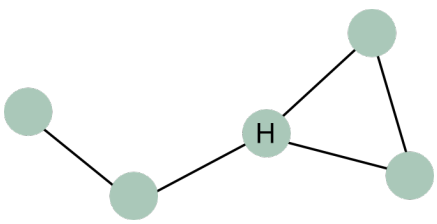

Figure 8. Simple network. Five nodes (green) are connected by edges (black) showing the most highly connected node or hub (H).

10.3. Shiny web app

To enable non-bioinformaticians to run the miRFA pipeline, we created a shiny app for all 775 miRNAs detected in the TCGA-PAAD data (https://emmbor.shinyapps.io/mirfa/). The shiny app includes predicted miRNA targets, as well as correlation results for miRNA-miRNA target mRNA expression and miRNA-miRNA target protein expression. A Venn diagram is generated showing the overlap of the three miRNA target prediction databases. Over-represented KEGG pathways and GO terms of miRNA targets that were significantly correlated to the miRNA expression level can be downloaded.

Chapter 11 – Metabolite profiling

Metabolites were detected by untargeted liquid chromatography mass spectrometry (LCMS) and gas chromatography-MS (GCMS). LCMS and GCMS detect different metabolites depending on volatility and polarity (**Figure 9**). The volatility also depends on how the samples are pre-processed and whether metabolites are derivatized or not. There is some overlap between the two platforms, such as amino acids and fatty acids, whereas other metabolites are specific for LCMS or GCMS. Metabolite

profiling was performed at the Swedish Metabolomics Centre (Umeå, Sweden).

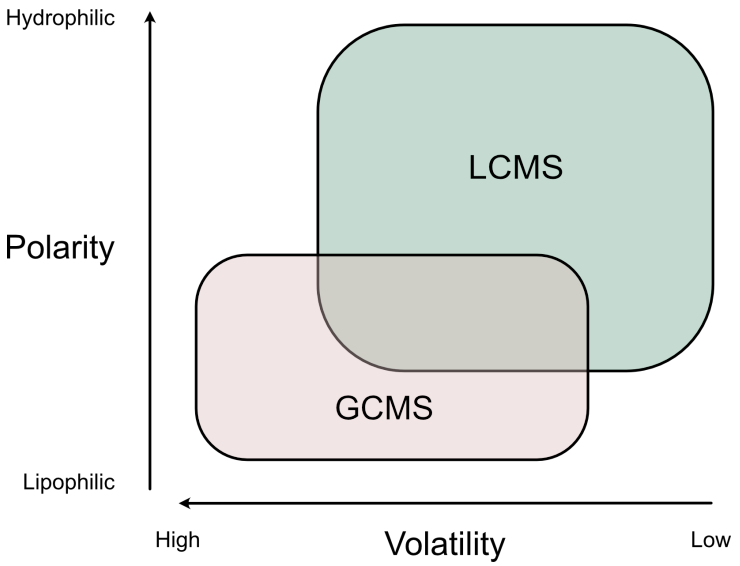

Figure 9. Types of metabolites quantified by LCMS and GCMS. Detected metabolites differ by the two methodologies in terms of polarity and volatility (Brack et al. 2016). LCMS = liquid chromatography mass spectrometry, GCMS = gas chromatography mass spectrometry.

11.1. Metabolite extraction

Plasma samples were prepared as previously described (A et al. 2005). Briefly, 900 µL extraction buffer (90/10 v/v HPLC grade methanol (Fisher Scientific, Waltham, MA, USA)/milliQ water) including internal standards was added to 100 µL plasma followed by shaking for 2 minutes at 30 Hz in a mixer mill. Proteins were precipitated at 4 °C on ice followed by centrifugation at 14,000 rpm, 4 °C for 10 minutes. A volume of 200 µL supernatant for LC-MS analysis and 100 µL for GC-MS were evaporated to dryness in a speed-vac concentrator. Furthermore, solvents were evaporated followed by storage at -80 °C until analysis. Quality control samples were created by pooling a small aliquot of remaining supernatants. The QC samples were analyzed by MSMS (LCMS) for metabolite identification purposes.

11.2. LCMS

Samples were resuspended in 10 μL methanol and 10 μL water followed by LCMS analysis in positive and negative mode. Chromatographic separation was performed on Agilent 1290 Infinity UHPLC-system (Agilent Technologies, Waldbronn, Germany) followed by detection with Agilent 6550 Q-TOF mass spectrometer connected to a jet stream electrospray ion source. Data was processed using Agilent Masshunter Profinder version B.08.00 (Agilent Technologies Inc., Santa Clara, CA, USA). Batch Targeted feature extraction in Masshunter Profinder was used to search a pre-defined list of commonly detected metabolites in plasma and serum. An inhouse LCMS library of authentic standards analyzed on the same system with similar settings for targeted processing. Metabolites were identified by information of MS, MSMS, and retention time.

11.3. GCMS

Derivatization and GC-MS profiling was performed on a Pegasus HT time-of-flight mass spectrometer, GC/TOFMS (Leco Corp., St Joseph, MI) as previously described (A et al. 2005). Non-processed MS-files were exported from ChromaTOF software to MATLAB 2018a (Mathworks, Natick, MA, USA), where following pre-treatment procedures were performed; base-line correction, chromatogram alignment, data compression, and multivariate curve resolution (Jonsson et al. 2005). Mass spectra were identified by comparing retention index and mass spectra with those found in libraries using NIST MS 2.2. software (Schauer et al. 2005). Reverse and forward searches were done, with extra caution taken on masses and ratio between masses indicative of a derivatized metabolite. A peak was identified by the mass spectrum with the highest probability and a maximum difference of five between library and sample for the suggested metabolite.

Chapter 12 – Protein profiling

Plasma proteins were analyzed using enzyme-linked immunosorbent assay (ELISA, **Paper II**), Luminex (**Papers III & IV**) or proximity extension assays (PEA, Olink®, **Paper IV**) (**Figure 10**). All three protein methods require antibody recognition of two different epitopes of the protein, which reduces false positive signals and increases specificity. The choice of protein assay depends on the purpose of the study. Milliplex and

PEA are more sensitive than ELISA and require smaller sample volumes. However, ELISA is cheap and fast to run. PEA can be multiplexed up to around 3000 proteins using NGS.

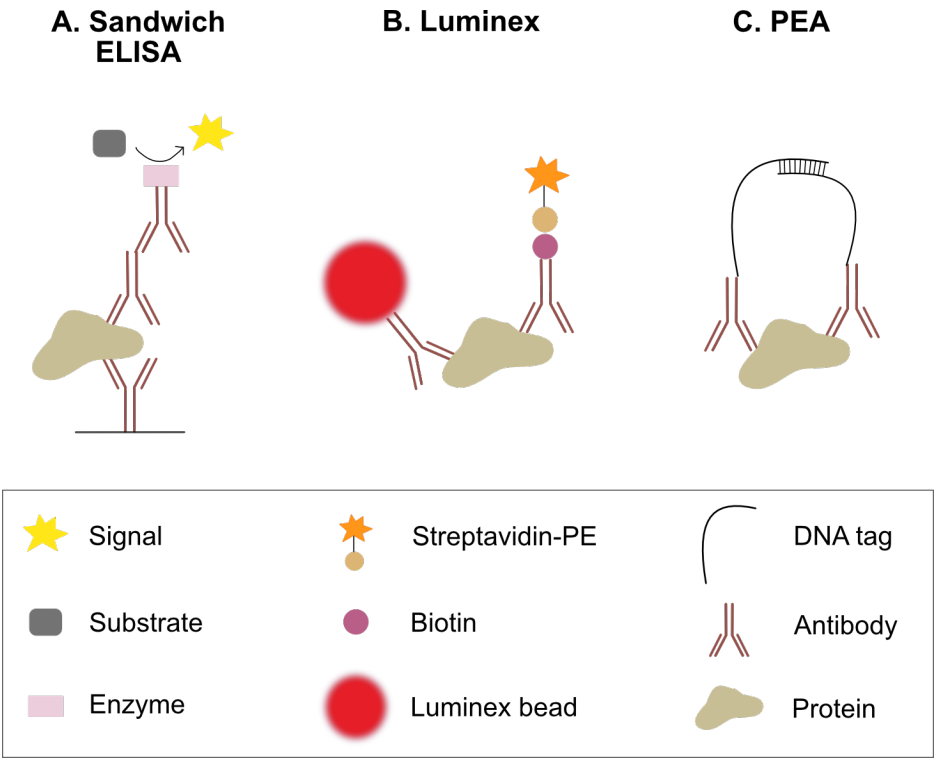

Figure 10. Methods employed for quantification of plasma proteins. A) In sandwich enzyme-linked immunosorbent assay (ELISA), a protein binds to a capturing antibody that is attached to the bottom of a well in a microplate. A primary antibody is added followed by a secondary antibody ligated to an enzyme. Substrate is added and consumed by the enzyme generating a signal. B) In Luminex assays, a protein binds to antibodies coupled to fluorescent beads. Another antibody is added conjugated to biotin. A streptavidin ligated to phycoerythrin (PE) generates a fluorescent signal. C) In proximity extension assays, antibodies conjugated to DNA tags bind to the target protein. When the two antibodies bind to two proximal target epitopes of a protein, the DNA tags are hybridized and amplified by qPCR or NGS.

12.1. ELISA

Circulating TPS was assessed in 50 µL plasma using TPS ELISA kit (IDL Biotech, Bromma, Sweden) following the manufacturer's instructions. Samples were analyzed in duplicates and a coefficient of variation (CV)

below 15 % was accepted. For samples with TPS < 80 U/L (lowest reference value), a higher CV was accepted. The average of duplicates was used in statistical analyses.

12.2. Luminex analyses

The clinically used biomarkers CA 19-9, CEA, and CA 15-3 were assessed in 6 µL plasma using Milliplex Multiplex assays for Luminex kit Human circulating biomarker panel 1 (Merck) according to manufacturer's instructions. Samples were analyzed in duplicates with an accepted CV below 15 %. The average of the two replicates was used in statistical analyses. Values for samples above the dynamic range of the standard curve were determined from an extrapolated standard curve. Values below limit of detection were imputed by lowest detectable value divided by two.

12.3. PEA

Proximity extension assays (PEA) were performed by Olink® (Uppsala, Sweden). Plasma samples were analyzed using seven Olink® panels; metabolism, immune response, inflammation, oncology II & III, cardiometabolic, and cardiovascular III. A volume of 1 µL plasma was analyzed in each Olink® panel.

Chapter 13 – MicroRNA profiling

A total of 2083 miRNAs were analyzed in 15 µL plasma using whole miRNA transcriptome assay (HTG Molecular Diagnostics, Inc) by TATAA Biocenter AB (Göteborg, Sweden). Quantification and sequencing are performed on hybridized miRNA-specific probes (**Figure 11**). The protocol for Plasma, Serum and PAXgene samples (HTG EdgeSeq System User Manual [ROU], version 10254600 Revision F) was followed. Briefly, 15 µL plasma was lysed, followed by hybridization between microRNAs and probes. Indexing, library amplification and cleanup followed by quality check and qPCR quantification by TATAA NGS Library Quantification kit (Part no: TA20-NGSQ, TATAA Biocenter AB) were subsequently performed. Libraries were normalized, pooled and sequenced using NextSeq500 (Illumina) with the parameter single end 50 base pairs mid output in Illumina's cloud-based service BaseSpace. Demultiplexed fastq-files were aligned, parsed and raw miRNA counts were generated in the HTG Edgeseq system software.

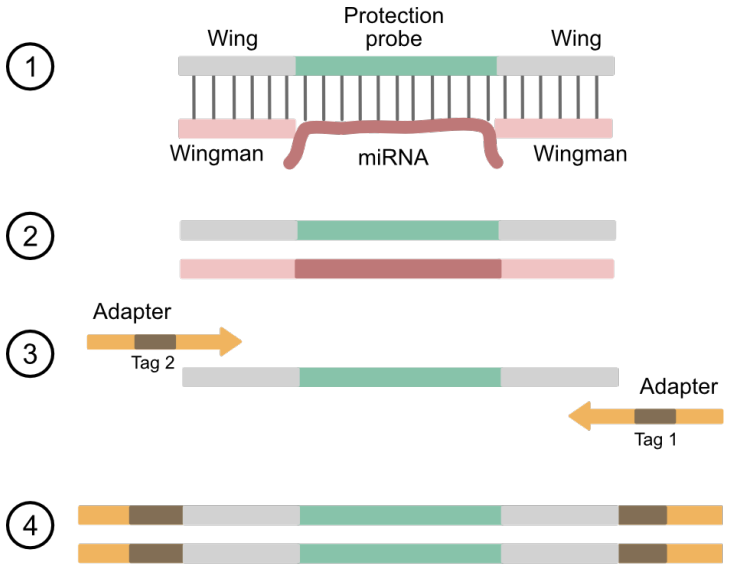

Figure 11. Whole miRNA transcriptome assay by HTG Edgeseq. 1) MicroRNA (miRNA) is hybridized to a protection probe ligated to 'wings' that will bind to a 'wingman'. 2) Non-hybridized probes and RNA are degraded by S1 nuclease, which results in a 1:1 ratio of probes and miRNAs. 3) Primers carrying sequencing adaptors and molecular barcodes amplify the remaining probes. 4) Resulting PCR amplicons are purified, quantified, and combined into a sequencing library.

Chapter 14 – Statistical analyses

All statistical analyses were performed in R (R Core Team 2021).

14.1. Univariate analysis

Conditional or unconditional logistic regression was performed using survival R package (Therneau 2022; Therneau and Grambsch 2000) or glm function from stats package (R Core Team 2021), respectively. For comparing continuous clinical characteristics, the Student's t-test or Mann-Whitney U rank test was performed. For categorical clinical characteristics, Fisher's exact test was calculated.

14.2. Multivariate analysis

Orthogonal projections to latent structures (OPLS) models were generated using ropls R package (Thevenot et al. 2015). OPLS models separate the latent variables by predictive components associated to the outcome and orthogonal components associated to other sources of

variation. OPLS effect projections (OPLS-EP) was used for paired analyses (Jonsson et al. 2015) and OPLS discriminant analysis (OPLS-DA) for unpaired analyses. Subset OPLS-EP models of circulating metabolomics were built stratified by time to diagnosis versus overall survival (time between sample date and death) or TNM stage (Jonsson et al. 2020). Each subset consisted of the 15 cases closest to a specific time to diagnosis-value and overall survival or TNM coordinate, as well as their matched healthy controls.

Multi-omics integration was performed using data integration analysis for biomarker discovery using latent components (DIABLO) (Rohart et al. 2017; Singh et al. 2019) and multi-omics factor analysis (MOFA) (Argelaguet et al. 2020).

14.3. LASSO regression

Least absolute shrinkage and selection operator (LASSO) regression was performed using glmnet R package in combination with bootstrapping with replacement (Friedman et al. 2010). LASSO is a variable selection method that sets some coefficients to zero for variables that are not informative for the outcome, it thus simplifies the regression model. Bootstrapping with replacement will generate new cohorts consisting of some of the samples in the original cohort. This can be viewed as changing the cohort multiple times. We performed LASSO on 500 bootstrapping iterations and finally we included variables with the highest proportion in the bootstrap-cohorts into final logistic regression models.

14.4. Survival analysis

Survival analyses were performed using survival (Therneau 2022; Therneau and Grambsch 2000), survminer (Kassambara 2018), and RTCGA.clinical R packages (Kosinski 2018). Cox regression as well as Kaplan Meier estimates using median expression as cut-off were performed to determine the prognostic value. Multivariable cox regression of miR-885-5p adjusted for age at diagnosis, tumor stage, sex, and histological grade was performed in the TCGA-PAAD cohort.

14.5. Imputation

Mice package was used o impute missing clinical information on BMI, smoking and type of surgery (van Buuren and Groothuis-Oudshoorn

2011). Mean values were imputed for BMI and smoking in the final logistic regression models of LASSO-selected variables.

Chapter 15 – Visualization

Figures were constructed in R using ggplot2 (Wickham 2016), ggrepel (Slowikowski 2021), the network analysis and visualization software Cytoscape (Shannon et al. 2003), and the graphic design software Affinity designer (Serif Europe Ltd).

Results

Chapter 16 – Patient cohorts

16.1. TCGA-PAAD cohort

We utilized the TCGA-PAAD and TCPA-PAAD data in **Paper I**. Clinical characteristics for TCGA-PAAD patients are described in **Table 4**.

Table 4. Clinical characteristics of TCGA-PAAD patients.

Variable	TCGA-PAAD patients (n = 177)
Mean age (years, range)	65 (36-89)
Sex	
Women, n (%)	80 (45 %)
Men, n (%)	97 (55 %)
Median survival (months)	7.9
Histological grade	
1	31 (18 %)
2	94 (53 %)
3	48 (27 %)
4	2 (1 %)
NA	2 (1 %)
Tumor stage, n (%)	
I	21 (12 %)
II	146 (82 %)
III	3 (2 %)
IV	4 (2 %)
Information missing	3 (2 %)

NOS = not otherwise specified

16.2. Included patients in pre-diagnostic and diagnostic cohorts

Included pre-diagnostic PDAC samples are presented in **Figure 12**. The cohort-specific clinical characteristics are described in **Papers II-IV**. A diagnostic PDAC cohort was also used in **Paper II** (**Paper II, Figure 1B**).

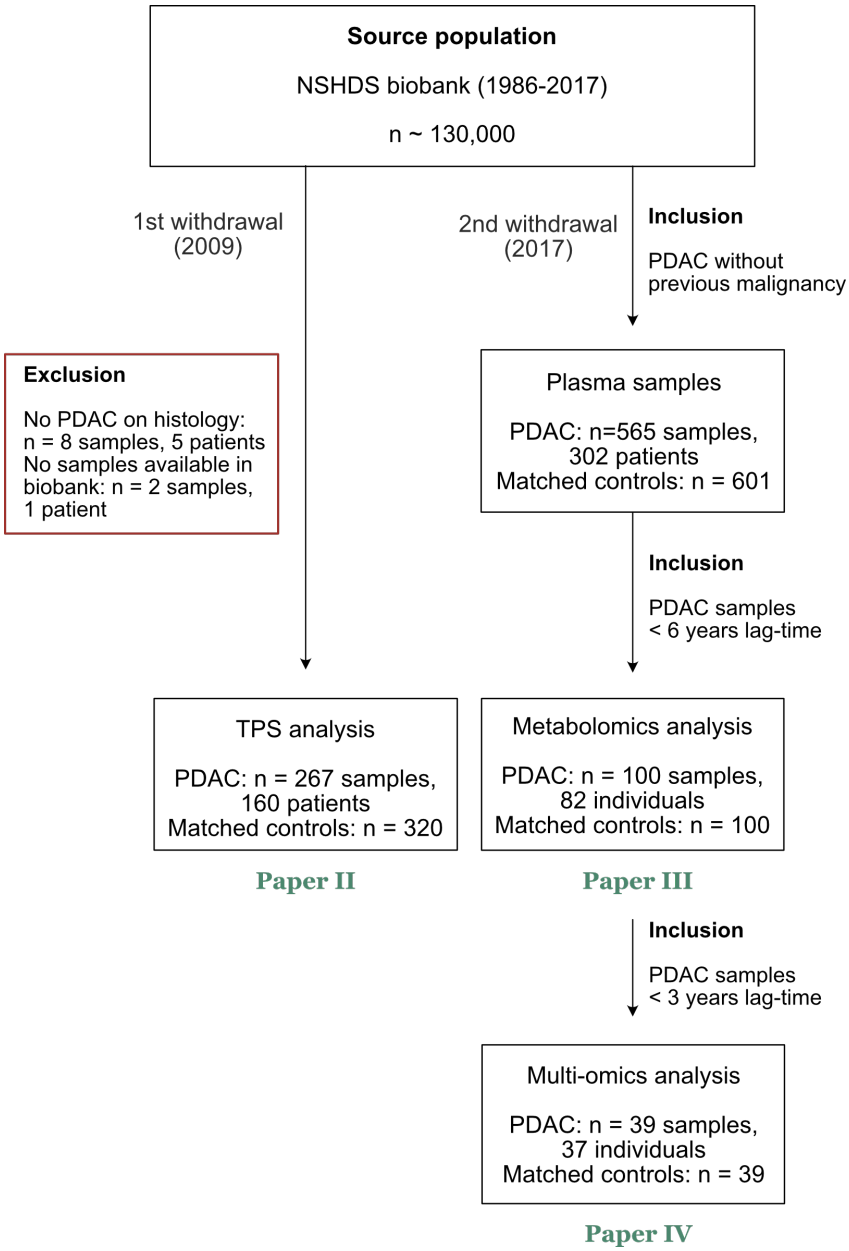

Figure 12. Flowchart of prospective cohorts. Included pre-diagnostic PDAC plasma samples in papers II-IV.

Chapter 17 – miRFA: microRNA functional analysis in pancreatic cancer

We developed a bioinformatics pipeline consisting of the following steps; miRNA target prediction, miRNA-target correlations (including mRNA and protein expression levels), and over-representation analysis of correlated miRNA targets (**Paper I, Figure 1**). We separated between the mature miRNA isoforms to be able to provide a more precise functional analysis of miRNAs. A list of 15 differentially expressed circulating miRNAs at PDAC diagnosis was used to show the functionality of our pipeline (Franklin et al. 2018). An example of mir-144 shows that there will be a difference in expression levels depending on whether the miR-144-3p and miR-144-5p isoforms are studied together or separately (**Paper I, Figure 2**). The number of predicted miRNA targets was reduced by including only those with a correlation support on mRNA or protein expression levels (**Paper I, Figure 4**).

We included correlation analyses between miRNAs and mRNAs or protein expression levels of predicted targets to provide further support for identified miRNA targets before proceeding to functional enrichment. Since there are studies of up-regulation by microRNAs on their targets, we included both positive correlations as well as negative ones in functional enrichment analysis (Rusk 2008; Vasudevan et al. 2007). We extended the miRFA tool by running each miRNA isoform detected in TCGA-PAAD data and made the results easy to download from a shiny web app (https://emmbor.shinyapps.io/mirfa/, **Figure 13**). The correlated miRNA targets can also be analyzed by other downstream analyses, such as network analyses (**Paper I, Figures 6 & 7**). We analyzed miR-885-5p as an example and identified the top 10 hub genes (**Paper I, Figure 6**). For the downstream analyses, positively and negatively correlated targets were analyzed separately. Different top 10 hub genes were identified for negatively and positively correlated targets with the top hits kinesin family member 2C (KIF2C) and S-Phase Kinase Associated Protein 1 (SKP1), respectively.

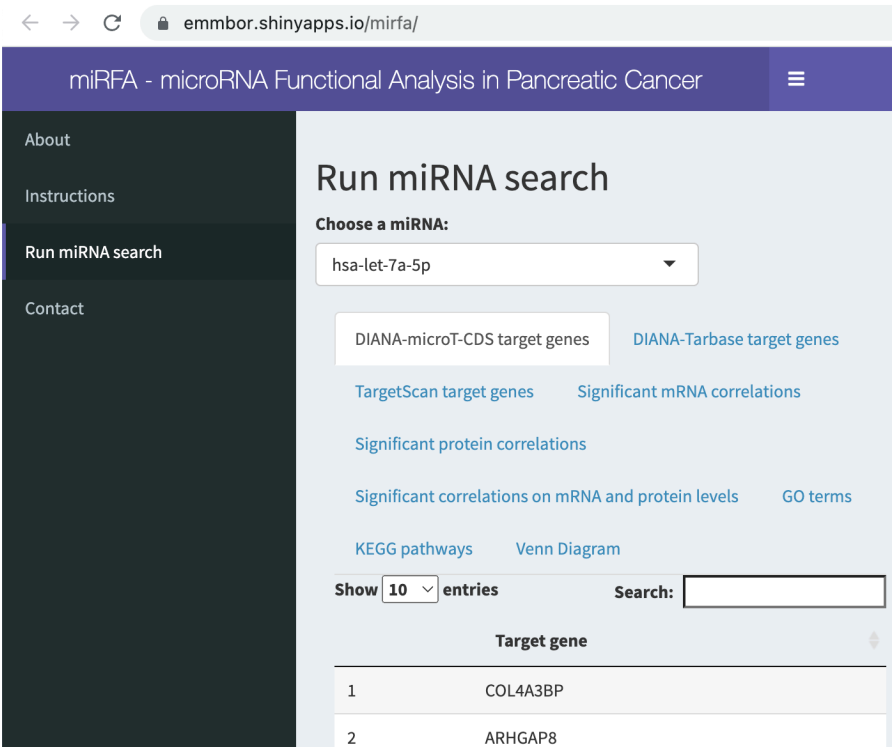

Figure 13. Screenshot of miRFA shiny web app.

We compared miRFA to miRCancerdb (Ahmed et al. 2018). The list of genes in KEGG pathway hsa05212 'Pancreatic cancer' was chosen as benchmark dataset to compare the 15 miRNAs (Franklin et al. 2018) used throughout **Paper I**. Since input miRNA in miRCancerdb are not in the form of mature isoforms, we modified the 15-miRNA list before running miRCancerdb. In general, miRFA picked up more correlations on mRNA expression level compared to miRCancerdb (**Paper I, Table 9**). On protein expression levels, miRCancerdb and miRFA generated different correlations for different miRNAs (**Paper I, Table 10**).

Chapter 18 – Potential pre-diagnostic PDAC biomarkers

A multi-omics approach was applied in this thesis to search widely for novel biomarkers with a potential for early PDAC detection. Around 3000 variables were studied in total. Multi-omics allows us to combine biomarkers, not only within a certain omics-type but also across different omics levels. This approach also offers the ability to validate findings at

other omics levels, compare performance of different omics candidates, and to facilitate biological interpretation.

18.1. Plasma TPS was not altered in pre-diagnostic PDAC

Circulating TPS was assessed in plasma samples from future PDAC cases up to 18.8 years lag-time to diagnosis. Low TPS levels were found in future PDAC cases (n=267 plasma samples) as well as in matched healthy controls (n = 320) (**Paper II, Figures 2B-D, Table 3**).

18.2. Circulating metabolites

Univariate logistic regression models adjusted for matching factors age, sex, and sample date returned 12 circulating metabolites with a nominal P-value < 0.05 (**Paper III, Figure 2A, Supplementary Table 2**). These were not significant after adjusting for multiple hypothesis testing. Since circulating histidine was previously found downregulated in pre-diagnostic PDAC (Fest et al. 2019), we examined the levels in time to diagnosis intervals of < 2 years (y), 2-5 y, and > 5 y. A slight drop in histidine level is visual < 2 y lag-time to PDAC diagnosis in concordance with Fest et al. (**Paper III, Supplementary Figure 2A**). This pattern seems to be specific for females in our data (**Paper III, Supplementary Figure 2B-C**). Longitudinal samples were available for 15 future PDAC cases. Among the 12 metabolites with a nominal P-value < 0.05, homoarginine (P-value = 0.03) and an unidentified metabolite with retention index (RI):2745.4 (P-value = 0.03) differed between first and last blood collection within the individuals (**Paper III, Supplementary Figure 3**). However, these differences were not significant after adjusting for multiple hypothesis testing.

We performed LASSO regression of the 12 metabolites with a nominal P-value < 0.05 using bootstrapping with replacement 500 times. Five informative metabolites (occurred in ≥ 80 % of the bootstrap sub-cohorts) were identified that were included into a final logistic regression including a baseline model (BMI, fasting status, smoking status, sex, sample date, and age) and CA 19-9. An improved AUC of the final model of 0.738 (95 % CI: 0.669-0.807) was achieved compared to 0.641 (95 % CI: 0.563-0.719) for the baseline model + CA 19-9 (P-value = 0.01, **Paper III, Figure 2B**).

18.3. Metabolites related to pre-diagnostic PDAC symptoms

Medical records were investigated for symptoms in PDAC patients up to six years before diagnosis in the cohorts in **Papers III & IV** (**Paper III, Supplementary Table 1**). The three most common pre-diagnostically reported symptoms were abdominal pain, weight loss, and back pain (**Paper III, Supplementary Table 4**). An OPLS-DA analysis revealed no symptom-specific metabolite signature for these three pre-diagnostic PDAC symptoms. We also stratified the pre-diagnostic cohort into patients with reported symptoms and a sample taken within the same time interval or after, and an asymptomatic cohort. We hypothesized that this would give cleaner cohorts and increase the chances of findings in those patients with reported symptoms that could possibly be due to a PDAC. Different metabolite profiles were obtained in the symptomatic compared to the asymptomatic cohort, however none of these metabolites were statistically significant after adjusting for multiple hypothesis testing (**Figure 3, Supplementary Tables 5 & 7**). A disadvantage of this approach is that the cohorts became smaller after splitting based on symptoms profile.

18.4. Circulating metabolites in relation to fasting glucose levels

Newly onset diabetes is a risk factor as well as a consequence of PDAC (Wild et al. 2020). Diabetic status can be reflected by fasting glucose levels. We thus split the cohort into individuals with normal fasting glucose (NFG < 6.1 mmol/L) and impaired fasting glucose (IFG ≥ 6.1 mmol/L). Different metabolite profiles were obtained with nominal P-value < 0.05 (**Figure 4, Supplementary Tables 10 & 11**). However, these were not significant after correcting for multiple hypothesis testing.

18.5. Subset OPLS-EP models of metabolites

OPLS-EP is a multivariate method used for paired study designs (Jonsson et al. 2015). We did not find any significant multivariate paired model (OPLS-EP) for the whole pre-diagnostic cohort in **Paper III**. Therefore, we sought to identify OPLS-EP models in subsets of the pre-diagnostic cohort since the case samples differ in terms of TNM stage and lag-time to PDAC diagnosis (Jonsson et al. 2020). The subsets were determined by coordinates in either time to diagnosis vs TNM stage (**Figure 14A & -C**), or time to diagnosis vs overall survival (**Figure 14B & -D**). A grid search

was performed of the whole coordinate system extracting the 15 cases closest to the coordinate, along with their matched controls. This creates different subsets from which an OPLS-EP model was created. Statistically significant OPLS-EP models of 15 case-control subsets could be identified using both LCMS and GCMS data (**Table 5**). However, no clear pattern was seen in relation to lag-time, TNM stage or overall survival.

Table 5. Summary statistics for OPLS-EP models of different case-control sub-cohorts.

Model	Max R2Y	Max Q2	Unique models	Significant models	Significant models/total unique models
LCMS – tnm	0.97	0.64	144	75	52 %
LCMS – surv	0.98	0.72	203	128	63 %
GCMS – tnm	0.98	0.53	144	58	40 %
GCMS – surv	0.96	0.45	203	84	41 %

LCMS = liquid chromatography mass spectrometry, GCMS = gas chromatography mass spectrometry, tnm = tumor node metastasis stage, surv = survival time between time of blood collection and death

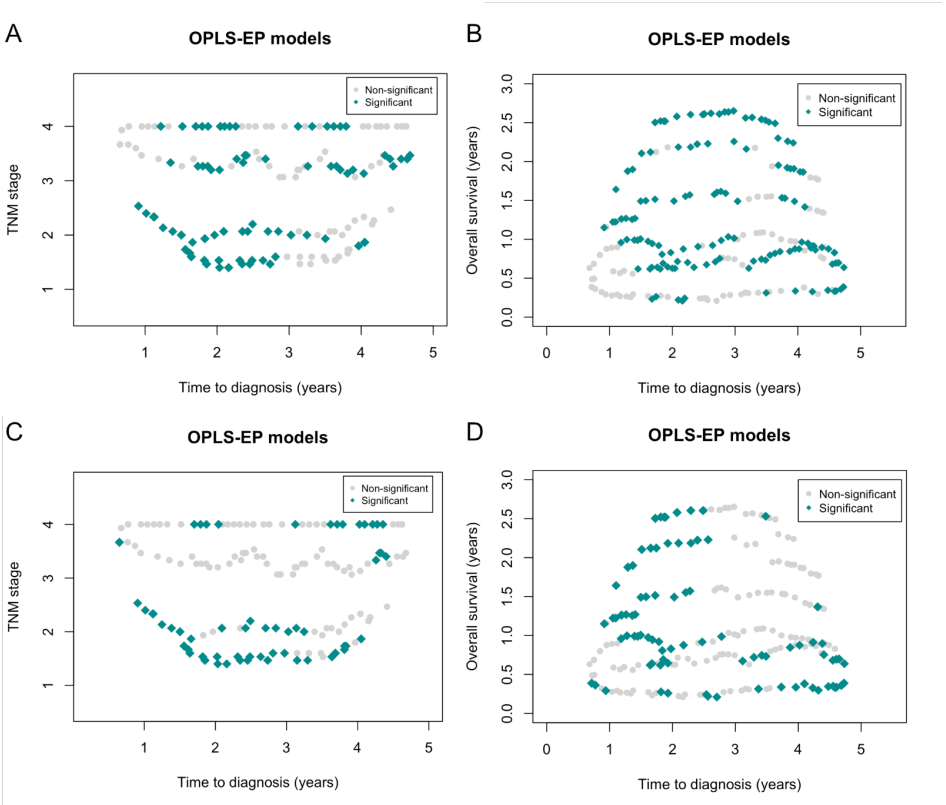

Figure 14. Subset analyses using OPLS-EP. Subsets were stratified based on time between sample date and death (overall survival) or tumor-node-metastasis (TNM) stage versus time to diagnosis. Each dot represents the mean value of each subset. Orthogonal projections to latent structures-effect projections (OPLS-EP) models were generated using LCMS data (A-B) or GCMS data (C-D). Each subset contains 15 cases and their matched controls. Since one matched case-control pair was excluded from the GCMS data, some subsets contain only 14 matched case-control pairs.

18.6. Multi-omics of pre-diagnostic PDAC

We profiled plasma proteins, metabolites, and microRNAs in 39 future PDAC samples and 39 matched healthy controls. Univariate logistic regression analysis revealed no significant (Q-value < 0.01) metabolite, protein, or microRNA after adjusting for multiple hypothesis testing, although we identified 96 variables with a nominal P-value <0.05 (**Paper IV, Figure 2**). Among these variables, none were selected in ≥ 70 % of LASSO-iterations when all omics levels were analyzed together. When performing LASSO on the omics types separately, two proteins (C-C motif chemokine ligand 15 [CCL15] and nuclear factor of activated T cells 3

[NFATC3]) and two miRNAs (miR-3646 and miR-132-5p) were selected in ≥ 70 % of the models. Final logistic regression models were built with miRNAs or proteins and were adjusted for sex, age, sample date, BMI, smoking status, and fasting status. The internal AUC for miRNAs in combination with CA 19-9 (0.884 [95 % CI: 0.810-0.959]) performed better than without miRNAs (0.769 [95 % CI: 0.661-0.878], P-value = 0.01, **Paper IV, Figure 3**). No statistically significant difference was found between a model including proteins in combination with CA 19-9 (0.802 [95 % CI: 0.703-0.902]) and a model without the proteins.

To find the variables that separated best between future PDAC cases and healthy controls, we performed DIABLO, a supervised multi-omics method (Rohart et al. 2017; Singh et al. 2019). The best visual separation was seen between future PDAC cases and healthy controls for miRNA and protein blocks (**Paper IV, Figure 4**). However, the performance, assessed by leave-one-out cross-validation, of these was poor (**Paper IV, Table 4**). Unsupervised MOFA factors correlated more to clinical parameters than to the outcome of interest, namely case-control status (**Paper IV, Figure 5B**).

We also created a DIABLO model with male/female as outcome. This was regarded as a positive control for biological signals in our data. Perfect discrimination by LC-Metabolites was found with the male hormone testosterone giving the highest loading of the component (**Paper IV, Supplementary Figure 2, Supplementary Table 1**).

Chapter 19 – Pre-diagnostic CA 19-9 levels

Circulating CA 19-9 levels start to increase < 2 years before PDAC diagnosis (**Figure 15**). The odds ratio increased from 1.38 in the cohort in **Paper III** (up to six years lag-time to PDAC diagnosis) to 2.18 in the cohort in **Paper IV** (up to three years lag-time to PDAC diagnosis) (**Table 6**).

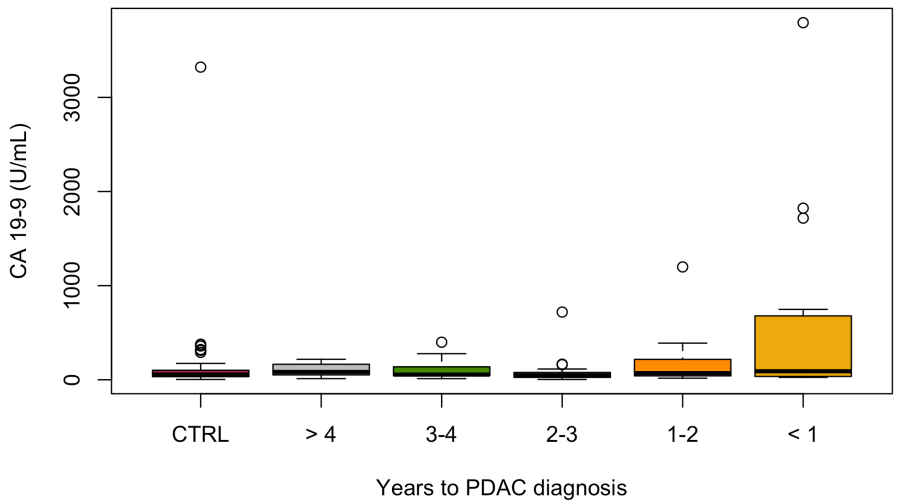

Figure 15. CA 19-9 in pre-diagnostic pancreatic cancer. Plasma CA 19-9 levels stratified by different lag-time to pancreatic cancer diagnosis intervals and matched healthy controls (CTRL).

Table 6. Conditional logistic regression models of plasma CA 19-9 in pre-diagnostic pancreatic cancer. Odds ratios are shown for unadjusted ('Crude') models and models adjusted for smoking status and BMI ('Adjusted').

Lag-time to PDAC diagnosis	Crude OR (95 % CI)	Crude P-value	Adjusted OR (95 % CI)	Adjusted P-value
< 6 y (Paper III)	1.38 (1.02-1.86)	0.04	1.46 (1.06-2)	0.02
< 3 y (Paper IV)	2.18 (1.15-4.13)	0.02	2.45 (1.21-4.95)	0.02

PDAC = pancreatic ductal adenocarcinoma, OR = odds ratio

Chapter 20 – Circulating TPS at diagnosis

Circulating TPS was found higher in PDAC patients compared to benign controls by logistic regression adjusted for age (**Paper II, Figure 2A, Table 3**, P-value < 0.001). The mean TPS level in PDAC patients and controls were 208 ± 196 U/L and 48 ± 28 U/L, respectively. The levels did not differ between different PDAC stages (**Paper II, Figure 3**, P-value = 0.3, Kruskal Wallis rank sum test).

Chapter 21 – Prognostic biomarkers

21.1. The prognostic value of miR-885-5p in TCGA-PAAD

We found miR-885-5p to be a possible prognostic tissue miRNA biomarker in PDAC using information from the TCGA-PAAD project (**Paper I, Figure 5**). However, it was not significant after adjusting for multiple testing and we did not adjust for any potential confounders. MiR-885-5p was moreover not significant after adjusting for potential confounders (**Table 7**).

Table 7. Crude and adjusted cox regression models of miR-885-5p in the TCGA-PAAD cohort. P-values for the coefficient are shown. Models were adjusted for age at diagnosis, sex, tumor stage, and histological grade.

MiRNA	Crude HR (95 % CI)	Crude P-value	Adjusted HR (95 % CI)	Adjusted P-value
miR-885-5p	0.61 (0.38-0.96)	0.032*	0.67 (0.40-1.12)	0.13

HR = hazard ratio

21.2. Prognostic circulating metabolites

Five circulating fatty acids (3-hydroxypalmitate, 13-HODE + 9-HODE or alpha-Dimorphecolic acid, hydroxystearate, 3-hydroxydecanoate, and hydroxymyristate) with prognostic value were identified in pre-diagnostic PDAC samples (Q-value < 0.1, **Paper III, Figure 5**). Higher levels of these fatty acids indicated a worse prognosis.

Discussion

Pancreatic ductal adenocarcinoma presents at a late stage with diffuse symptoms. In this thesis, a bioinformatics tool was developed for investigating miRNA functions in pancreatic cancer. Furthermore, we have investigated a total of 2083 miRNAs, 645 proteins, and 408 metabolites in pre-diagnostic plasma with the main aim to identify potential biomarkers for early PDAC detection. The miRFA tool was developed into a shiny app, where the user can download results for the miRNAs detected in the TCGA-PAAD data. TPS was downregulated in pre-diagnostic PDAC but increased at PDAC diagnosis. Among analytes with a nominal P-value < 0.05, LASSO selected five metabolites, two proteins, and two microRNAs to be most informative. Logistic regression models were built in combination with baseline variables, such as BMI, smoking status, matching factors, and fasting status, as well as CA 19-9. Internal AUCs for these models were 0.74, 0.80, and 0.88 for five metabolites, two proteins, and two miRNAs, respectively.

Chapter 22 – miRFA

The miRNA functional analysis approach implemented in miRFA is an indirect annotation consisting of miRNA target prediction and functional enrichment of predicted target genes. We compared miRFA to another tool called miRCancerdb (Ahmed et al. 2018), which we considered most similar to our tool. Functional enrichment analysis was not implemented in miRCancerdb, and we thus only compared the obtained miRNA-target correlations in the two tools. We defined the gene list in KEGG pathway hsa05212 'pancreatic cancer' as our benchmark data set. MiRFA and miRCancerdb generated somewhat different correlations on miRNA-target mRNA (**Paper I, Table 9**), as well as miRNA-target protein expression levels (**Paper I, Table 10**). There are at least two differences that could explain this. First, miRNA expression differ between the two tools. In miRFA, we implemented mature miRNA isoform expression, whereas miRCancerdb use the hairpin miRNA expression level. Second, different resources of miRNA target predictions were used in the two tools. DIANA-microT-CDS, Tarbase, and TargetScan were implemented for miRNA target prediction in miRFA. MiRCancerdb implements targetscan.Hs.eg.db R package for miRNA target prediction of conserved

miRNA targetes using TargetScan (Csardi 2013). Thus, different miRNA targets will be identified in miRFA and miRCancerdb.

Chapter 23 – Pre-diagnostic plasma analyses

The search for new early detection biomarkers was done directly in pre-diagnostic cohorts in this thesis. We have previously identified 15 miRNAs deregulated at PDAC diagnosis, which were then studied pre-diagnostically (Franklin et al. 2018). However, the 15-miRNA signature was not discriminative between PDAC and healthy controls in the pre-diagnostic cohort. Hence, there could be other important circulating signals pre-diagnostically that disappears at diagnosis or later in the disease course.

23.1. Metabolomics in pre-diagnostic PDAC

We identified a few metabolites with a nominal P-value < 0.05 that did not remain significant after multiple hypothesis correction. This could be due to low power and a heterogeneous cohort, with different lag-times to PDAC diagnosis and different clinical characteristics at diagnosis. Interestingly, we found histidine as one of the downregulated metabolites identified in our pre-diagnostic pancreatic cancer cohort. This is in line with Fest et al., who also found histidine to be downregulated with a nominal P-value < 0.05 in future PDAC patients using plasma samples from five European population-based biobanks (Fest et al. 2019). Not only has histidine been found downregulated pre-diagnostically, but a systematic review also identified histidine as one of the most frequently reported downregulated metabolites in pancreatic cancer at diagnosis (Long et al. 2018). Histidine has also been found to have an inverse association with colorectal cancer risk in the EPIC cohort (Breeur et al. 2022).

We further stratified the pre-diagnostic cohort based on fasting glucose levels and pre-diagnostically reported PDAC symptoms. We identified different deregulated metabolites between future PDAC patients and controls within these sub-cohorts, yet no metabolite reached statistical significance after adjusting for multiple hypothesis testing. By this stratification approach, we expect the sub-cohorts to be slightly less heterogeneous, however we lose power as the cohorts get smaller, which

could be why no metabolites remained significant after adjusting for multiple hypothesis testing.

A prognostic biomarker would in a clinical setting be measured at the time of disease onset to predict survival of patients. In **Paper III**, we identified five potential prognostic fatty acids (3-hydroxypalmitate, 13-HODE + 9-HODE or alpha-Dimorphecolic acid, hydroxystearate, 3-hydroxydecanoate, and hydroxymyristate) that were measured pre-diagnostically. Cox regression models adjusted for lag-time and the interaction term lag-time*metabolite were constructed. In this way we adjusted for the fact that the measurements were performed pre-diagnostically. However, translating lag-time to a clinical setting is difficult and thus further validation of these fatty acids in PDAC diagnostic cohorts will be necessary to evaluate the clinical utility.

23.2. Multi-omics analyses in pre-diagnostic PDAC

We performed multi-omics on pre-diagnostic PDAC plasma samples with the aim to identify a biomarker signature that could be useful in early PDAC detection. None of the metabolites, proteins, or miRNAs studied were significant after adjusting for multiple testing in univariate analyses. This could – similar to the case in **Paper III** – be due to heterogeneous cohorts and low power. Among the nominal variables with a P-value < 0.05, LASSO combined with 500 bootstrap iterations identified CCL-15, NFATC3, miR-3646, and miR-132-5p as the most potential candidates for early PDAC detection.

A model constructed by supervised multi-omics method DIABLO successfully identified different omics profiles that separated between males and females. This provided a positive control for biological signals in our data. Furthermore, this might offer important biological insights that could benefit other studies of plasma samples using any of the omics methods we have included. As expected, testosterone was most important for separating males and females, which indicates that the data is reliable. Another expected finding was Chorionic Gonadotropin Subunit Beta 3 (CGB3), a subunit of chorionic gonadotropin (CG), which is important in maintaining pregnancy. However, there was another protein with much higher loading on the proteomics DIABLO-component; persephin (PSPN). Persephin promotes survival of neuronal populations and is

enhanced in early and late spermatids by RNA single cell data in the human protein atlas (proteinatlas.org).

Chapter 24 – Strengths

In our developed miRFA tool for *in silico* functional analysis of miRNAs, we separated between miRNA isoforms to enable a more specific functional analysis of mature miRNAs. MiRFA was extended to a freely available shiny app containing results for all miRNA isoforms identified in TCGA-PAAD data. This makes the tool more widely accessible to users not familiar with R in which the pipeline was built.

Our pre-diagnostic plasma samples were obtained from the NSHDS biobank, which contains good quality plasma samples that have been frozen within 1 hour after sampling and stored at -80 °C. In **Paper III**, we combined metabolomics analyses with symptoms reported up to six years prior to PDAC diagnosis, which made it possible for us to separate the search for screening biomarkers in an asymptomatic population and early diagnosis biomarkers in symptomatic individuals. We performed wide screens of potential biomarkers by including validated high throughput methods. By combining multi-omics levels we could gain more biological insights by correlating different omics layers. This was done by correlating the latent MOFA factors with other clinical variables, such as BMI, smoking status, and sex. Moreover, we could directly compare the performance of different types of biomarkers. We could see in our DIABLO model that the protein block separated future PDAC patients from healthy controls better than the other omics modalities studied.

Chapter 25 – Limitations

There are several limitations in the current miRFA version. First, it is only possible to query one miRNA at a time. Adding the option to query a whole list would be of value. Second, it is limited to pancreatic cancer, but it would be possible to extend to other cancer types in TCGA. Third, the tool is limited to perform enrichment analysis of both positively and negatively correlated miRNA targets together. An option would be to allow the user to perform over-representation analysis of positively and negatively correlated miRNA targets separately. Even though we included both correlation directions with the aim of being unbiased, we might miss

interesting GO terms or KEGG pathways when combining positively and negatively correlated miRNA targets. Fourth, a relatively narrow protein expression dataset (~200 proteins) was available in TCPA-PAAD. Thus, many miRNA targets were not available for correlation analyses between a miRNA and its predicted targets on protein expression levels. Fifth, miRFA is currently restricted to pancreatic cancer.

Another limitation in this thesis is that we obtained rather small sub-cohorts by splitting the metabolomics cohort (**Paper III**) into individuals with IFG, NFG, pre-diagnostic symptoms and no pre-diagnostic symptoms. We imputed smoking status and BMI for individuals with missing information and that introduced some uncertainty. The potential prognostic circulating metabolites identified were measured pre-diagnostically and hence these results should be interpreted with caution. We restricted our search to include seven Olink® panels leaving out other potential biomarkers as more panels were available. We only focused on pre-diagnostic sample from individuals who later developed pancreatic cancer. This made it impossible to interpret whether deregulated variables related to a cancer in general or specifically to PDAC.

Our studied cohorts were challenging to work with for several reasons. First, we did not know how PDAC progresses in each case, i.e. we cannot be sure that a tumor actually exists at the sample occasion. In addition, even if there was a tumor present at blood sample collection, when does it become detectable in blood? Second, larger cohorts are needed. We analyzed a larger cohort in paper II, however the lag-time to diagnosis for PDAC patients was very long – up to 19 years. Third, the cohorts were extremely heterogeneous. The heterogeneity was mainly due to different lag-times to diagnosis as well as a variation in clinical characteristics at the time of PDAC diagnosis. Adding to that, there was also variation in life style factors and other clinical parameters in the cohort participants. Fourth, although we had a lot of information on lifestyle factors, such as smoking and BMI, there could also be other confounders that we have not measured.

Chapter 26 – Challenges in biomarker discovery

The curse of dimensionality (CoD) is a major challenge in large scale data (Altman and Krzywinski 2018). The concept includes for instance data

sparsity, multicollinearity, overfitting, and multiple testing. As the number of variables grow, data will become more sparse. Multicollinearity describes how we can predict one variable by a linear combination of the rest when the number of variables outgrows the number of samples. Overfitting happens when a prediction model gets too adapted to the samples, partly due to random associations in the samples. This will create a very good model for the cohort studied but the external validity will be low. We aimed to mitigate overfitting by applying bootstrapping iterations in our data and perform variable selection by LASSO. This reduced the number of variables in logistic regression models, which decreases the chances of overfitting. Adjusting for multiple testing is performed with the goal to reduce the number of false positive results but it can also lead to false negatives due to reduced power. None of our studied multi-omics variables in association to early PDAC detection remained significant after we applied multiple hypothesis correction. Statistical improvements, larger sample sizes, and machine learning can to some extent mitigate CoD. However, the most important aspect as suggested by Altman & Kzywinski, is to separate between exploratory studies and confirmatory studies (Altman and Krzywinski 2018). The multi-omics analysis of pre-diagnostic PDAC samples performed in this thesis was explorative. Thus, our identified metabolites, proteins, and microRNAs might be worth exploring further despite being non-significantly altered after multiple hypothesis correction. Another challenge is that the choice of bioinformatics or statistical tools as well as parameter definitions can have a great impact on the results. This makes comparing and reproducing studies challenging.

Many promising biomarker candidates fail clinical implementation. CA 19-9 is still the only clinically used biomarker for PDAC. One explanation for that is lack of validation and the presence of confounders. Many studies are published today with good biomarker performance that were not validated in an external cohort (Long et al. 2018). This relates to the previously mentioned problem of overfitting a model. A confounder is a variable correlated to the exposure as well as the outcome, causing a false association between the exposure and the outcome. The potential confounders highly depend on the research question, but smoking, socioeconomic status, and risk factors for the disease studied are examples of confounders. In this thesis, we adjusted the models by known confounders that we have measured, such as sex, BMI, and smoking

status. However, the unmeasured, but known potential confounders could affect the results in a way that we cannot control for. Heavy alcohol consumption is a risk factor in PDAC (Wild et al. 2020). However, we lacked information on alcohol habits for around 80 % of individuals in our cohort. We did therefore not adjust our statistical models for alcohol consumption. In addition, there might exist confounders that we do not know of and that we have not measured.

Chapter 27 – Opportunities in biomarker discovery

In 1990, the human genome project (HUGO) started with the aim to sequence the whole human genome (https://www.hugo-international.org/). In 2003, more than 90 % of the human genome was mapped. In January 2022, the full human genome was complete. The cost for producing the whole genome sequence was around the figures of several billion dollars at the time, whereas today companies offer whole genome sequencing for costs of around a few hundred dollars. This decline in sequencing costs has led to a lot of data being generated throughout the years, much of them publicly accessible. Sharing or publishing data, and decreasing costs will make research faster and more effective. This makes it easier to reuse data. Mendelian Randomization is a statistical method that uses nature's random allocation of alleles to explain causality in observational studies. It can also be applied in biomarker discovery for risk prediction and can facilitate biological understanding by for instance associating circulating biomarkers with known risk factors and the disease of interest. Public data can also be used for creating accessible bioinformatics tools. TCGA-PAAD and TCPA-PAAD are examples of publicly available data resources that we used for building the miRFA tool.

Chapter 28 – Challenges in PDAC screening

The biomarker discovery conducted in this thesis is just the first step in identifying a potential screening biomarker or diagnostic biomarker. There are many challenges to tackle before reaching clinical implementation of candidate PDAC biomarkers. Screening the whole population for PDAC will probably never become a reality since a biomarker with an excellent specificity of 99% will still result in 1000 false-positives per 100,000 screened individuals due to low life-time risk of developing pancreatic cancer (Lucas and Kastrinos 2019). One risk with screening programs is overdiagnosis and overtreatment. For pancreatic

cancer screening, this is very challenging as pancreatic surgery is a major procedure with high morbidity. Another challenge is that risk factors for PDAC are still poorly characterized (Wild et al. 2020). Some risk factors are defined such as chronic pancreatitis, or a history of familial PDAC, however these still explain a small portion of patients. This makes it difficult to define a high risk population to include in a screening setting and to enable prevention.

The International cancer of the pancreas screening (CAPS) consortium has suggested annual imaging screening of individuals carrying a mutation with higher risk for developing PDAC or familial pancreatic cancer kindreds (Goggins et al. 2020). The main goal of the pancreatic surveillance program is to prevent PDAC death and progression to PDAC by treating precursor lesions. This is achieved by early PDAC detection at either stage I or precursors with high-grade dysplasia. Some individuals enrolled in PDAC surveillance programs with annual imaging still present with advanced disease without an option for curative surgery once a PDAC is formed. A recent study followed 366 high-risk individuals (165 mutation carriers and 201 mutation-negative familial PDAC kindreds) annually by imaging using EUS and MRI/MRCP (Overbeek, Levink, et al. 2022). Ten PDAC patients were identified among mutation carriers, with an over-representation of individuals with *CDKN2A* mutation or Peutz-Jeghers syndrome. Of these PDAC patients, only six were resected with curative intent and only three (30 %) met the formal goal of surveillance defined by CAPS consortium. A multi-center study in 16 centers and seven countries surveilling 2552 high-risk individuals detected 28 individuals that developed pancreatic cancer or high-grade dysplasia (Overbeek, Goggins, et al. 2022). Thirteen had a new lesion since prior examination and ten had a lesion with progression beyond pancreas at diagnosis. Fifteen had a previously detected lesion out of which eleven had a lesion with progression beyond pancreas. This highlights the timing challenge of pancreatic cancer surveillance, which would have to be further investigated before implementing novel screening biomarkers.

Future perspectives

Chapter 29 – Early detection of PDAC

Despite years of research, the overall survival of pancreatic cancer patients remains poor. Early detection is crucial to be able to offer curative surgery to a greater proportion of pancreatic cancer patients and subsequently enhance patient survival. Further analyses of larger pre-diagnostic cohorts could pave the way towards discovery of novel biomarkers. However, pre-diagnostic pancreatic cancer samples are rare. Thus, international collaborations are valuable to be able to increase the sample size and simultaneously limit the lag-time to up to a few years before pancreatic cancer diagnosis. New potential biomarkers might be discovered by other platforms, such as lipidomics or additional Olink® panels. There are also other sample matrices available in NSHDS biobank, such as red blood cells, and buffy coat, which also might be interesting to analyze for potential biomarkers.

One aspect of early detection is to be able to set the correct diagnosis fast in patients presenting with symptoms. In paper III, we identified future pancreatic cancer patients that experienced symptoms before disease onset. It would be interesting to collect more samples of individuals with pancreatic cancer-associated symptoms and identify altered analytes associated to pancreatic cancer among symptomatic individuals. This would enable us to compare molecular patterns between individuals that experience symptoms due to pancreatic cancer or other reasons.

The other aspect of early detection is screening. Since PDAC is a rare disease, this is not feasible in the whole population and a high-risk group needs to be identified. The United Kingdom Early Detection Initiative has started constructing a biobank of individuals with new-onset diabetes for early detection of pancreatic cancer (Oldfield et al. 2022). Individuals with familial pancreatic cancer or carriers of mutations that entail an increased pancreatic cancer risk are monitored today. As some individuals display a rapid disease progression in current surveillance programs, it would be of value to find biomarkers that could detect malignant changes before these appear on imaging. Performing large scale explorative studies on blood samples collected in high-risk individuals would provide an opportunity to identify potential screening biomarkers for pancreatic cancer.

Chapter 30 – Future bioinformatics studies

Expression data for 32 cancer types is available in the TCGA database. The miRFA pipeline could be extended to include other cancer types than pancreatic cancer, or a pan-cancer dataset since many features are shared across different cancer types (Chen et al. 2018). Hence, this would give the opportunity to study miRNA functional analysis in a wide cancer-context. It would also be interesting to compare mechanisms that are specific to pancreatic cancer and mechanisms shared across cancer types.

In this thesis we defined four omics modalities; microRNAs, proteins, LC-metabolites, and GC-metabolites. However, different types of omics can be redefined. It would be of value to further zoom into the proteins block and disentangle variation specific to the seven Olink® panels. In this way we could determine which one(s) of the protein panels is most interesting to choose in further analyses of larger cohorts. We have performed extensive multi-omics analyses on plasma samples, which we hope to make publicly available in some way. Since the samples fall under general data protection regulation (GDPR) and are thus considered personal data, one possibility would be to make results instead of raw data freely available.

In addition to the studied pre-diagnostic analytes in this thesis, imaging data could also be explored for early PDAC detection. Artificial intelligence (AI) could be applied on imaging data to determine whether subtle pancreatic changes visible through imaging can be detected. A good AI tool could also remove the differences in inter-individual performance by radiologists. Recently, an AI tool was published for early pancreatic cancer detection on CT scans (Chen et al. 2022). This tool could be implemented for CT scans in Umeå retrospectively. It would also be interesting to study plasma analytes in relation to findings by imaging to see whether biomarkers could be identified that detects the tumor before visible on imaging.

Conclusions

In this thesis, we developed a novel bioinformatics pipeline for microRNA functional analysis and performed multi-omics analysis of high-quality, pre-diagnostically collected blood samples. Early pancreatic cancer detection is difficult to achieve by analyzing microRNAs, Olink® protein panels and metabolomics in pre-diagnostic plasma samples. Thus, identifying biomarkers for early detection of pancreatic cancer still remains challenging.

- Our miRFA tool successfully identifies and correlates predicted miRNA targets, as well as generates enriched pathways, in a pancreatic cancer-specific context

- Circulating TPS is increased in pancreatic cancer at diagnosis

- Among the studied multi-omics analytes, we identified five metabolites, two proteins, and two miRNAs as most informative for separating between future pancreatic cancer patients and healthy controls

- Circulating CA 19-9 levels increase closer to pancreatic cancer diagnosis

- Five circulating fatty acids (3-hydroxypalmitate, 13-HODE + 9-HODE or alpha-Dimorphecolic acid, hydroxystearate, 3-hydroxydecanoate, and hydroxymyristate) had prognostic value in pre-diagnostic pancreatic cancer and could be validated in pancreatic cancer patients at diagnosis

- Liquid chromatography-derived metabolites and proteins successfully separate males from females with highest weights in testosterone and persephin

Acknowledgements

I would like to thank Biobanken Norr for guidance and the Swedish Metabolomics Centre, Umeå, Sweden, for performing untargeted metabolomics (www.swedishmetabolomicscentre.se). I would also like to thank Hans Stenlund and Annika Johansson at the Swedish Metabolomics Centre for valuable discussions. Information on pre-diagnostic symptoms was collected by **Maja Simm** and **Sara Jacobson**. Clinical information on the pancreatic cancer patient cohorts was collected by **Erik Lundberg, Hanna Nyström, Daniel Öhlund, and Malin Sund**. **Christina Lundin** performed protein analyses by ELISA and milliplex. **Anette Berglund** performed hematoxylin/eosin staining in pancreatic tissue sections. I would also like to thank **Tajaswi Badam** and **Thomas Hillerton** for programming assistance, **Xiaoshuang Feng** for assistance in statistical analyses, and **James Mason** for English proofreading of Paper I.

Jag är så tacksam över alla som gjort denna resa möjlig och trevligare; mitt dreamteam-handledargäng, mina kollegor, min familj och mina vänner.

Till **alla mina handledare**: tack för ert stöd och för att ni ville ställa upp som mina handledare! Ni kompletterar verkligen varandra med olika bakgrund och kunskaper. Detta har gjort att jag fått så otroligt bra och bred input i mina projekt. Jag är också tacksam över att ni har inkluderat mig i andra projekt och introducerat mig för nya kontakter och samarbeten.

Till min huvudhandledare professor **Malin Sund**; jag ser verkligen upp till dig och din otroliga kompetens och engagemang som du besitter. Tack för att du har låtit mig jobba självständigt men också stöttat och väglett när det behövts. Tack för ett gott samarbete och jag hoppas det kommer fortsätta!

Till min bi-handledare **Zelmina Lubovac-Pilav**; tack för att du gett mig en god bioinformatisk grund att bygga vidare på och dina uppmuntrande ord. Till min bi-handledare **Ola Billing**; tack för din värdefulla handledning kring praktiskt labbarbete och trevliga luncher med många skratt. Till min bi-handledare **Pär Jonsson**; tack för din statistiska

handledning och hjälp, samt för att du uppmärksammat framgångar. Till min bi-handledare **Oskar Franklin**; tack för att du gett mig goda insikter i det kliniska arbetet och varit så positiv.

Jag vill tacka stjärnorna på labb **Anette Berglund** och **Christina Lundin**. Ni gör ett fantastiskt arbete. Tack också till andra medlemmar i kirlab **Hanna Nyström, Oskar Hemmingsson, Fredrik Nilsson, Gunilla Rask, Niklas Löfgren, Adrian Molnár** och **Malin Jansson** för ett gott samarbete och trevliga kickoffs. Känns alltid tryggt när man har ett gäng kirurger runt sig. Thank you to the former kirlab member **Szymon Gorgon** for all the challenging brain puzzles you have given me.

Till mina vänner/doktorandkollegor **Moa Lindgren, Josefin Jonsson, Sara Karlsson** och **Sara Jacobson**, tack för alla skratt och samarbeten genom doktorandtiden. Kämpa på med resten av era doktorandprojekt, jag hejar på er!

Till **Henrik Antti**; tack för att jag fått följa med på era gruppmöten. Det har varit otroligt givande.

I would like to thank **Mattias Johansson** and **Hillary Robbins**, as well as their team members for my research stay at IARC, Lyon. It was very nice meeting you and working with you for a short period of time. I learned a lot!

Jag vill tacka min tidigare chef **Martin Sandberg** och alla fantastiska före detta kollegor på Livsmedelsverket för att jag fick jobba i så intressanta projekt samt ägna tid åt värdefull kompetensutveckling och nationella samarbeten.

I would like to thank professor **Laura Machesky**, **Benjamin Thyrell**, and **Kevin Myant** who were my supervisors during my thesis degree project in molecular medicine at Beatson Institute of cancer research in Glasgow. It was thanks to this well-designed project I grew an interest to continue in research.

To **Hendrik Arnold de Weerd**, it has been nice to work with you and thank you for always helping me with R issues.

Till min vän/tidigare kollega **Sofia Persson**, jag har alltid roligt med dig och du har alltid kloka visdomar att komma med som exempelvis gjort att jag blivit bekväm med att känna mig dum.

Till min barndomsvän **Nina Liikanen**, tack för att jag kan berätta allt för dig och tack för alla jamsessioner. Tack till **Jonna Illanvuori, Linda Ekman, My Haga** och **Linda Pitkänen** för all kvalitetstid jag får med er. Tack **Anna Viklund** för att du gjorde gymnasietiden så mycket roligare och för att du är så otroligt peppande. Tack **Irina Borgmästars** för alla roliga barndoms-sommarminnen. Jag är också otroligt glad över att jag fick lära känna **Helen Kahsay-Seiron, Frida Wennerholm, Emelie Wallén, Lilian Kempe, Sebastian Kapell** och **Feria Hikmet** under studietiden i Uppsala. Det är fantastiskt att få har er som vänner!

Stort tack till **Viktor Boman** och **Anton Brännvall** för att ni ställde upp som toastmasters!

Till mamma **Yvonne**, tack för att du alltid är så stöttande, lugn och tålmodig. Du hittar lösningar på varje problem. Till pappa **Tage**, när jag flyttade till Uppsala sa du: "du kan ju alltjämt koma hejm". Det fick mig att inse att 'hejm' alltid finns kvar oavsett var jag befinner mig och gjort att jag vågat testa på saker utanför min bekvämlighetszon. Tack till er båda för att ni alltid stöttar mig och mina livsval, det betyder så otroligt mycket.

Till mina fantastiska syskon **Erica**, **Fredrik** och **Melvin**, jag har alltid så otroligt roligt med er och ni betyder så mycket för mig. Tack till dom som förgyller mina syskons vardag; **Kristoffer, Alma, Troy, Elvi, Elin, Signe, Agnes** och **Tilda**.

Till fammo **Mona**, faffa **Per**, mommo **Maggie** och moffa **Kalli**. Tack för alla kära barndomsminnen och äventyr.

Tack till **Britt-Inger** och **Jan-Erik**. Jag kunde inte fått bättre svärföräldrar. Tack också till min svägerska **Joanna**, **Andreas, Lova** och **Edvard**.

Till min klippa **Anton**, tack för att du alltid stöttar mig i mina livsval, för att du alltid finns där för mig och pushar mig till att våga testa nya saker trots att det ofta innebär att du också tvingas ur din bekvämlighetszon. Du gör min vardag mycket roligare. Till min dotter **Karla**, du är så otroligt fin och jag är så tacksam över att jag får vara din mamma. Ni båda är mitt allt, älskar er.

Funding

- The Swedish Research Council (Vetenskapsrådet)
- The Swedish Cancer Society (Cancerfonden)
- The Sjöberg Foundation (Sjöbergsstiftelsen)
- Västerbotten County Council (Region Västerbotten)
- The JC Kempe Memorial Foundation Scholarship Fund
- Umeå University
- Lion's Cancer Research Foundation
- Knut and Alice Wallenberg Foundation

References

A, J., J. Trygg, J. Gullberg, et al. 2005. 'Extraction and GC/MS analysis of the human blood plasma metabolome', *Anal Chem*, 77: 8086-94.

Agarwal, V., G. W. Bell, J. W. Nam, et al. 2015. 'Predicting effective microRNA target sites in mammalian mRNAs', *Elife*, 4.

Ahmed, M., H. Nguyen, T. Lai, et al. 2018. 'miRCancerdb: a database for correlation analysis between microRNA and gene expression in cancer', *BMC Res Notes*, 11: 103.

Allen, P. J., D. Kuk, C. F. Castillo, et al. 2017. 'Multi-institutional Validation Study of the American Joint Commission on Cancer (8th Edition) Changes for T and N Staging in Patients With Pancreatic Adenocarcinoma', *Ann Surg*, 265: 185-91.

Altman, N., and M. Krzywinski. 2018. 'The curse(s) of dimensionality', *Nature Methods*, 15: 399-400.

Aquina, C. T., A. Ejaz, A. Tsung, et al. 2021. 'National Trends in the Use of Neoadjuvant Therapy Before Cancer Surgery in the US From 2004 to 2016', *JAMA Netw Open*, 4: e211031.

Argelaguet, R., D. Arnol, D. Bredikhin, et al. 2020. 'MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data', *Genome Biol*, 21: 111.

Ballehaninna, U. K., and R. S. Chamberlain. 2012. 'The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal', *J Gastrointest Oncol*, 3: 105-19.

Banfi, G., A. Zerbi, S. Pastori, et al. 1993. 'Behavior of Tumor-Markers Ca19.9, Ca195, Cam43, Ca242, and Tps in the Diagnosis and Follow-up of Pancreatic-Cancer', *Clinical Chemistry*, 39: 420-23.

Barabasi, A. L., and Z. N. Oltvai. 2004. 'Network biology: understanding the cell's functional organization', *Nat Rev Genet*, 5: 101-13.

Baradaran, B., R. Shahbazi, and M. Khordadmehr. 2019. 'Dysregulation of key microRNAs in pancreatic cancer development', *Biomed Pharmacother*, 109: 1008-15.

Betel, D., M. Wilson, A. Gabow, et al. 2008. 'The microRNA.org resource: targets and expression', *Nucleic Acids Res*, 36: D149-53.

Bhaskaran, M., and M. Mohan. 2014. 'MicroRNAs: history, biogenesis, and their evolving role in animal development and disease', *Vet Pathol*, 51: 759-74.

Bindea, G., J. Galon, and B. Mlecnik. 2013. 'CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data', *Bioinformatics*, 29: 661-3.

Brack, W., S. Ait-Aissa, R. M. Burgess, et al. 2016. 'Effect-directed analysis supporting monitoring of aquatic environments--An in-depth overview', *Sci Total Environ*, 544: 1073-118.

Brand, R. E., J. Persson, S. O. Bratlie, et al. 2022. 'Detection of Early-Stage Pancreatic Ductal Adenocarcinoma From Blood Samples: Results of a Multiplex Biomarker Signature Validation Study', *Clin Transl Gastroenterol*, 13: e00468.

Breeur, M., P. Ferrari, L. Dossus, et al. 2022. 'Pan-cancer analysis of pre-diagnostic blood metabolite concentrations in the European Prospective Investigation into Cancer and Nutrition', *BMC Med*, 20: 351.
Brierley, J.D. 2017. *TNM Classification of Malignant Tumours* (John Wiley & Sons: Chisester).
Busnardo, A. C., L. J. DiDio, R. T. Tidrick, et al. 1983. 'History of the pancreas', *Am J Surg*, 146: 539-50.
Chang, C. Y., S. P. Huang, H. M. Chiu, et al. 2006. 'Low efficacy of serum levels of CA 19-9 in prediction of malignant diseases in asymptomatic population in Taiwan', *Hepato-Gastroenterology*, 53: 1-4.
Chang, L., G. Zhou, O. Soufan, et al. 2020. 'miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology', *Nucleic Acids Res*, 48: W244-W51.
Chen, F., Y. Zhang, D. L. Gibbons, et al. 2018. 'Pan-Cancer Molecular Classes Transcending Tumor Lineage Across 32 Cancer Types, Multiple Data Platforms, and over 10,000 Cases', *Clin Cancer Res*, 24: 2182-93.
Chen, P. T., T. Wu, P. Wang, et al. 2022. 'Pancreatic Cancer Detection on CT Scans with Deep Learning: A Nationwide Population-based Study', *Radiology*: 220152.
Chin, C. H., S. H. Chen, H. H. Wu, et al. 2014. 'cytoHubba: identifying hub objects and sub-networks from complex interactome', *BMC Syst Biol*, 8 Suppl 4: S11.
Chou, C. H., S. Shrestha, C. D. Yang, et al. 2018. 'miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions', *Nucleic Acids Res*, 46: D296-D302.
Cohen, J. D., L. Li, Y. Wang, et al. 2018. 'Detection and localization of surgically resectable cancers with a multi-analyte blood test', *Science*, 359: 926-30.
Conroy, T., P. Hammel, M. Hebbar, et al. 2018. 'FOLFIRINOX or Gemcitabine as Adjuvant Therapy for Pancreatic Cancer', *N Engl J Med*, 379: 2395-406.
Cortes-Ciriano, I., J. J. Lee, R. Xi, et al. 2020. 'Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing', *Nat Genet*, 52: 331-41.
Csardi, Gabor. 2013. 'targetscan.Hs.eg.db: TargetScan miRNA target predictions for human'.
Duell, E. J., L. Lujan-Barroso, N. Sala, et al. 2017. 'Plasma microRNAs as biomarkers of pancreatic cancer risk in a prospective cohort study', *Int J Cancer*, 141: 905-15.
Enright, A. J., B. John, U. Gaul, et al. 2003. 'MicroRNA targets in Drosophila', *Genome Biol*, 5: R1.
Fest, J., L. S. Vijfhuizen, J. J. Goeman, et al. 2019. 'Search for Early Pancreatic Cancer Blood Biomarkers in Five European Prospective Population Biobanks Using Metabolomics', *Endocrinology*, 160: 1731-42.
Franklin, O., P. Jonsson, O. Billing, et al. 2018. 'Plasma Micro-RNA Alterations Appear Late in Pancreatic Cancer', *Ann Surg*, 267: 775-81.
Friedman, J., T. Hastie, and R. Tibshirani. 2010. 'Regularization Paths for Generalized Linear Models via Coordinate Descent', *Journal of Statistical Software*, 33: 1-22.
Gallmeier, E., R. Hernaez, and T. M. Gress. 2015. 'Controversy on the time to progression of pancreatic ductal adenocarcinoma', *Gut*, 64: 1676-7.

Git, A., H. Dvinge, M. Salmon-Divon, et al. 2010. 'Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression', *RNA*, 16: 991-1006.
Goggins, M., K. A. Overbeek, R. Brand, et al. 2020. 'Management of patients with increased risk for familial pancreatic cancer: updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium', *Gut*, 69: 7-17.
Goldman, M. J., B. Craft, M. Hastie, et al. 2020. 'Visualizing and interpreting cancer genomics data via the Xena platform', *Nat Biotechnol*, 38: 675-78.
Hanahan, D. 2022. 'Hallmarks of Cancer: New Dimensions', *Cancer Discov*, 12: 31-46.
Hanahan, D., and R. A. Weinberg. 2000. 'The hallmarks of cancer', *Cell*, 100: 57-70.
Hanahan, D., and R. A. Weinberg. 2011. 'Hallmarks of cancer: the next generation', *Cell*, 144: 646-74.
Heiser, P. W., J. Lau, M. M. Taketo, et al. 2006. 'Stabilization of beta-catenin impacts pancreas growth', *Development*, 133: 2023-32.
Honda, K., V. A. Katzke, A. Husing, et al. 2019. 'CA19-9 and apolipoprotein-A2 isoforms as detection markers for pancreatic cancer: a prospective evaluation', *Int J Cancer*, 144: 1877-87.
Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009a. 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res*, 37: 1-13.
Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009b. 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat Protoc*, 4: 44-57.
Huang, Q., L. Y. Wu, Y. Wang, et al. 2013. 'GOMA: functional enrichment analysis tool based on GO modules', *Chin J Cancer*, 32: 195-204.
Hussein, N. A., Z. A. Kholy, M. M. Anwar, et al. 2017. 'Plasma miR-22-3p, miR-642b-3p and miR-885-5p as diagnostic biomarkers for pancreatic cancer', *J Cancer Res Clin Oncol*, 143: 83-93.
Jacobson, S., P. Dahlqvist, M. Johansson, et al. 2021. 'Hyperglycemia as a risk factor in pancreatic cancer: A nested case-control study using prediagnostic blood glucose levels', *Pancreatology*, 21: 1112-18.
Jemal, A., R. Siegel, E. Ward, et al. 2007. 'Cancer statistics, 2007', *CA Cancer J Clin*, 57: 43-66.
John, B., A. J. Enright, A. Aravin, et al. 2004. 'Human MicroRNA targets', *PLoS Biol*, 2: e363.
Jonsson, P., H. Antti, F. Spath, et al. 2020. 'Identification of Pre-Diagnostic Metabolic Patterns for Glioma Using Subset Analysis of Matched Repeated Time Points', *Cancers (Basel)*, 12.
Jonsson, P., A. I. Johansson, J. Gullberg, et al. 2005. 'High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses', *Anal Chem*, 77: 5635-42.
Jonsson, P., A. Wuolikainen, E. Thysell, et al. 2015. 'Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples', *Metabolomics*, 11: 1667-78.

Kanda, M., H. Matthaei, J. Wu, et al. 2012. 'Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia', *Gastroenterology*, 142: 730-33 e9.
Karagkouni, D., M. D. Paraskevopoulou, S. Chatzopoulos, et al. 2018. 'DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions', *Nucleic Acids Res*, 46: D239-D45.
Kassambara, A.; Kosinski, M. 2018. 'survminer: Drawing Survival Curves using 'ggplot2''.
Katagiri, R., A. Goto, T. Nakagawa, et al. 2018. 'Increased Levels of Branched-Chain Amino Acid Associated With Increased Risk of Pancreatic Cancer in a Prospective Case-Control Study of a Large Cohort', *Gastroenterology*, 155: 1474-82 e1.
Kim, J. E., K. T. Lee, J. K. Lee, et al. 2004. 'Clinical usefulness of carbohydrate antigen 19-9 as a screening test for pancreatic cancer in an asymptomatic population', *Journal of Gastroenterology and Hepatology*, 19: 182-86.
Kordes, M., L. Larsson, L. Engstrand, et al. 2021. 'Pancreatic cancer cachexia: three dimensions of a complex syndrome', *Br J Cancer*, 124: 1623-36.
Kosinski, M. 2018. 'RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project'.
Lewis, B. , and J. Mao. 2018. 'Development of the Pancreas and Related Structures.' in H.G. Beger, A.L. Warshaw, R. H. Hruban, M.W. Büchler, M.M. Lerch, J.P. Neoptolemos, T. Shimosegawa and D.C. Whitcomb (eds.), *The Pancreas : An Integrated Textbook of Basic Science, Medicine, and Surgery* (John Wiley & Sons Ltd: Hoboken NJ, USA).
Li, J., Y. Lu, R. Akbani, et al. 2013. 'TCPA: a resource for cancer functional proteomics data', *Nature Methods*, 10: 1046-7.
Long, N. P., S. J. Yoon, N. H. Anh, et al. 2018. 'A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer', *Metabolomics*, 14: 109.
Lucas, A. L., and F. Kastrinos. 2019. 'Screening for Pancreatic Cancer', *JAMA*, 322: 407-08.
Mahajan, U. M., B. Oehrle, S. Sirtl, et al. 2022. 'Independent Validation and Assay Standardization of Improved Metabolic Biomarker Signature to Differentiate Pancreatic Ductal Adenocarcinoma From Chronic Pancreatitis', *Gastroenterology*.
Maitra, A., N. V. Adsay, P. Argani, et al. 2003. 'Multicomponent analysis of the pancreatic adenocarcinoma progression model using a pancreatic intraepithelial neoplasia tissue microarray', *Mod Pathol*, 16: 902-12.
Mason, J., E. Lundberg, P. Jonsson, et al. 2022. 'A Cross-Sectional and Longitudinal Analysis of Pre-Diagnostic Blood Plasma Biomarkers for Early Detection of Pancreatic Cancer', *International Journal of Molecular Sciences*, 23: 12969.
Mayerle, J., H. Kalthoff, R. Reszka, et al. 2018. 'Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis', *Gut*, 67: 128-37.
Mayers, J. R., C. Wu, C. B. Clish, et al. 2014. 'Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development', *Nat Med*, 20: 1193-98.

McCarthy, D. J., Y. Chen, and G. K. Smyth. 2012. 'Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation', *Nucleic Acids Res*, 40: 4288-97.
Mellby, L. D., A. P. Nyberg, J. S. Johansen, et al. 2018. 'Serum Biomarker Signature-Based Liquid Biopsy for Diagnosis of Early-Stage Pancreatic Cancer', *J Clin Oncol*, 36: 2887-94.
Merath, K., R. Mehta, D. I. Tsilimigras, et al. 2020. 'In-hospital Mortality Following Pancreatoduodenectomy: a Comprehensive Analysis', *J Gastrointest Surg*, 24: 1119-26.
Mitchell, P. S., R. K. Parkin, E. M. Kroh, et al. 2008. 'Circulating microRNAs as stable blood-based markers for cancer detection', *Proc Natl Acad Sci U S A*, 105: 10513-8.
Müller, K., H. Wickham, DA. James, et al. 2022. 'RSQLite: SQLite Interface for R'.
Neoptolemos, J. P., D. H. Palmer, P. Ghaneh, et al. 2017. 'Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial', *Lancet*, 389: 1011-24.
Noë, M., A.A.B. Lodewijk, and J Offerhaus. 2018. 'Pancreatic Cancer: Precancerous Lesions.' in H.G. Beger, A.L. Warshaw, R. H. Hruban, M.W. Büchler, M.M. Lerch, J.P. Neoptolemos, T. Shimosegawa and D.C. Whitcomb (eds.), *The Pancreas: An Integrated Textbook of Basic Science, Medicine, and Surgery* (John Wiley & Sons Ltd: Hoboken NJ, USA).
Noë, M., N. Niknafs, C. G. Fischer, et al. 2020. 'Genomic characterization of malignant progression in neoplastic pancreatic cysts', *Nat Commun*, 11: 4085.
Notta, F., M. Chan-Seng-Yue, M. Lemire, et al. 2016. 'A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns', *Nature*, 538: 378-82.
O'Brien, D. P., N. S. Sandanayake, C. Jenkinson, et al. 2015. 'Serum CA19-9 is significantly upregulated up to 2 years before diagnosis with pancreatic cancer: implications for early disease detection', *Clin Cancer Res*, 21: 622-31.
Oldfield, L., M. Stott, R. Hanson, et al. 2022. 'United Kingdom Early Detection Initiative (UK-EDI): protocol for establishing a national multicentre cohort of individuals with new-onset diabetes for early detection of pancreatic cancer', *BMJ Open*, 12: e068010.
Overbeek, K. A., M. G. Goggins, M. Dbouk, et al. 2022. 'Timeline of Development of Pancreatic Cancer and Implications for Successful Early Detection in High-Risk Individuals', *Gastroenterology*, 162: 772-85 e4.
Overbeek, K. A., I. J. M. Levink, B. D. M. Koopmann, et al. 2022. 'Long-term yield of pancreatic cancer surveillance in high-risk individuals', *Gut*, 71: 1152-60.
Ozkan, H., S. Demirbas, M. Ibis, et al. 2011. 'Diagnostic validity of serum macrophage inhibitor cytokine and tissue polypeptide-specific antigen in pancreatobiliary diseases', *Pancreatology*, 11: 295-300.
Pannala, R., J. B. Leirness, W. R. Bamlet, et al. 2008. 'Prevalence and clinical profile of pancreatic cancer-associated diabetes mellitus', *Gastroenterology*, 134: 981-7.
Paraskevopoulou, M. D., G. Georgakilas, N. Kostoulas, et al. 2013. 'DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows', *Nucleic Acids Res*, 41: W169-73.

Pasanen, P. A., M. Eskelinen, K. Partanen, et al. 1994. 'A Prospective-Study of Serum Tumor-Markers Carcinoembryonic Antigen, Carbohydrate Antigen-50 and Antigen-242, Tissue Polypeptide Antigen and Tissue Polypeptide Specific Antigen in the Diagnosis of Pancreatic-Cancer with Special Reference to Multivariate Diagnostic Score', *British Journal of Cancer*, 69: 562-65.

Pathan, M., S. Keerthikumar, C. S. Ang, et al. 2015. 'FunRich: An open access standalone functional enrichment and interaction network analysis tool', *Proteomics*, 15: 2597-601.

Peters, M. L. B., A. Eckel, P. P. Mueller, et al. 2018. 'Progression to pancreatic ductal adenocarcinoma from pancreatic intraepithelial neoplasia: Results of a simulation model', *Pancreatology*, 18: 928-34.

Peterson, S. M., J. A. Thompson, M. L. Ufkin, et al. 2014. 'Common features of microRNA target prediction tools', *Front Genet*, 5: 23.

Pittman, M.E., and R.H. Hruban. 2018. 'Pathology of Exocrine Pancreatic Tumors.' in H.G. Beger, A.L. Warshaw, R. H. Hruban, M.W. Büchler, M.M. Lerch, J.P. Neoptolemos, T. Shimosegawa and D.C. Whitcomb (eds.), *The Pancreas: An Integrated Textbook of Basic Science, Medicine, and Surgery* (John Wiley & Sons Ltd: Hoboken NJ, USA).

Poruk, K. E., D. Z. Gay, K. Brown, et al. 2013. 'The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates', *Curr Mol Med*, 13: 340-51.

R Core Team. 2021. 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing*.

Rahib, L., M. R. Wehner, L. M. Matrisian, et al. 2021. 'Estimated Projection of US Cancer Incidence and Death to 2040', *JAMA Netw Open*, 4: e214708.

Raudvere, U., L. Kolberg, I. Kuzmin, et al. 2019. 'g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)', *Nucleic Acids Res*, 47: W191-W98.

Reczko, M., M. Maragkakis, P. Alexiou, et al. 2012. 'Functional microRNA targets in protein coding sequences', *Bioinformatics*, 28: 771-6.

Regionala cancercentrum i samverkan. 2017. "Nationellt vårdprogram Bukspottskörtelcancer." In.

Regionala cancercentrum i samverkan. 2021. "Nationellt vårdprogram för bukspottkörtelcancer; version 3.1." In. Stockholm, Sweden: Regionala cancercentrum i samverkan.

Regionala cancercentrum i samverkan. 2022a. "Kvalitetsregister för tumörer i pankreas och periampullärt - Årsrapport nationellt kvalitetsregister, diagnosår: 2021." In.

Regionala cancercentrum i samverkan. 2022b. "Livmoderhalscancer och vaginalcancer, Nationellt vårdprogram." In.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Rohart, F., B. Gautier, A. Singh, et al. 2017. 'mixOmics: An R package for 'omics feature selection and multiple data integration', *PLoS Comput Biol*, 13: e1005752.

Rusk, N. 2008. 'When microRNAs activate translation', *Nature Methods*, 5: 122-23.

Sah, R. P., A. Sharma, S. Nagpal, et al. 2019. 'Phases of Metabolic and Soft Tissue Changes in Months Preceding a Diagnosis of Pancreatic Ductal Adenocarcinoma', *Gastroenterology*, 156: 1742-52.
Satake, K., T. Takeuchi, T. Homma, et al. 1994. 'Ca19-9 as a Screening and Diagnostic-Tool in Symptomatic Patients - the Japanese Experience', *Pancreas*, 9: 703-06.
Schauer, N., D. Steinhauser, S. Strelkov, et al. 2005. 'GC-MS libraries for the rapid identification of metabolites in complex biological samples', *FEBS Lett*, 579: 1332-7.
Shannon, P., A. Markiel, O. Ozier, et al. 2003. 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res*, 13: 2498-504.
Sharma, A., T. C. Smyrk, M. J. Levy, et al. 2018. 'Fasting Blood Glucose Levels Provide Estimate of Duration and Progression of Pancreatic Cancer Before Diagnosis', *Gastroenterology*, 155: 490-500 e2.
Siegel, R. L., K. D. Miller, H. E. Fuchs, et al. 2022. 'Cancer statistics, 2022', *CA Cancer J Clin*, 72: 7-33.
Singh, A., C. P. Shannon, B. Gautier, et al. 2019. 'DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays', *Bioinformatics*, 35: 3055-62.
Singh, N. K. 2017. 'miRNAs target databases: developmental methods and target identification techniques with functional annotations', *Cell Mol Life Sci*, 74: 2239-61.
Slesak, B., A. Harlozinska-Szmyrka, W. Knast, et al. 2000. 'Tissue polypeptide specific antigen (TPS), a marker for differentiation between pancreatic carcinoma and chronic pancreatitis - A comparative study with CA 19-9', *Cancer*, 89: 83-88.
Slesak, R., A. Harlozinska-Szmyrka, W. Knast, et al. 2004. 'TPS and CA 19-9 measurements in the follow-up of patients with pancreatic cancer and chronic pancreatitis', *International Journal of Biological Markers*, 19: 115-19.
Slowikowski, K. 2021. 'ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. '.
Socialstyrelsen. 2022. 'Statistikområden, Cancer', Socialstyrelsen, Accessed [cited 2022-08-25. https://www.socialstyrelsen.se/statistik-och-data/statistik/statistikdatabasen.
Sticht, C., C. De La Torre, A. Parveen, et al. 2018. 'miRWalk: An online resource for prediction of microRNA binding sites', *PLoS One*, 13: e0206239.
Szklarczyk, D., A. L. Gable, K. C. Nastou, et al. 2021. 'The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets', *Nucleic Acids Res*, 49: D605-D12.
Therneau, T. 2022. 'A Package for Survival Analysis in R'.
Therneau, TM., and PM Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model* (Springer, New York).
Thevenot, E.A., A. Roux, Y. Xu, et al. 2015. 'Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses', *Journal of Proteome Research*, 14: 3322-35.

Tokar, T., C. Pastrello, A. E. M. Rossos, et al. 2018. 'mirDIP 4.1-integrative database of human microRNA target predictions', *Nucleic Acids Res*, 46: D360-D70.
van Buuren, S., and K. Groothuis-Oudshoorn. 2011. 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software*, 45: 1-67.
van Manen, L., J. V. Groen, H. Putter, et al. 2020. 'Elevated CEA and CA19-9 serum levels independently predict advanced pancreatic cancer at diagnosis', *Biomarkers*, 25: 186-93.
Vasudevan, S., Y. Tong, and J. A. Steitz. 2007. 'Switching from repression to activation: microRNAs can up-regulate translation', *Science*, 318: 1931-4.
Versteijne, E., J. L. van Dam, M. Suker, et al. 2022. 'Neoadjuvant Chemoradiotherapy Versus Upfront Surgery for Resectable and Borderline Resectable Pancreatic Cancer: Long-Term Results of the Dutch Randomized PREOPANC Trial', *J Clin Oncol*, 40: 1220-30.
Vila-Navarro, E., S. Duran-Sanchon, M. Vila-Casadesus, et al. 2019. 'Novel Circulating miRNA Signatures for Early Detection of Pancreatic Neoplasia', *Clin Transl Gastroenterol*, 10: e00029.
Vlachos, I. S., M. D. Paraskevopoulou, D. Karagkouni, et al. 2015. 'DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions', *Nucleic Acids Res*, 43: D153-9.
Vlachos, I. S., K. Zagganas, M. D. Paraskevopoulou, et al. 2015. 'DIANA-miRPath v3.0: deciphering microRNA function with experimental support', *Nucleic Acids Res*, 43: W460-6.
Waters, A. M., and C. J. Der. 2018. 'KRAS: The Critical Driver and Therapeutic Target for Pancreatic Cancer', *Cold Spring Harb Perspect Med*, 8.
WHO. 2001. 'World Health Organisation International Programme on Chemical Safety Biomarkers in Risk Assessment: Validity and Validation'. http://www.inchem.org/documents/ehc/ehc/ehc222.htm.
Wickham, H. 2011. 'The Split-Apply-Combine Strategy for Data Analysis', *Journal of Statistical Software*, 40: 1-29.
Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York).
Wild, C.P., E. Weiderpass, B.W. Stewart, et al. 2020. "World Cancer Report: Cancer Research for Cancer Prevention." In.: Lyon, France: International Agency for Research on Cancer.
Wong, N., and X. Wang. 2015. 'miRDB: an online resource for microRNA target prediction and functional annotations', *Nucleic Acids Res*, 43: D146-52.
Wu, W. S., B. W. Tu, T. T. Chen, et al. 2017. 'CSmiRTar: Condition-Specific microRNA targets database', *PLoS One*, 12: e0181231.
Xiao, F., Z. Zuo, G. Cai, et al. 2009. 'miRecords: an integrated resource for microRNA-target interactions', *Nucleic Acids Res*, 37: D105-10.
Yachida, S., S. Jones, I. Bozic, et al. 2010. 'Distant metastasis occurs late during the genetic evolution of pancreatic cancer', *Nature*, 467: 1114-7.
Yu, J., A. L. Blackford, M. Dal Molin, et al. 2015. 'Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages', *Gut*, 64: 1783-9.