Spring 5-8-2021

Development of In-silico Pipelines for Identification and Characterization of Biomarker Panels and Therapeutic Interventions in Gastro-Intestinal (GI) Cancers

Pranita Atri
University of Nebraska Medical Center

Development of *In-silico* Pipelines for Identification and Characterization of Biomarker Panels and Therapeutic Interventions in Gastro-Intestinal (GI) Cancers

By

Pranita Atri

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Biochemistry and Molecular Biology Graduate Program

Under the Supervision of Professor Dr. Surinder K. Batra

University of Nebraska Medical Center

Omaha, Nebraska

April 2021

Development of *In-silico* Pipelines for Identification and Characterization of Biomarker Panels and Therapeutic Interventions in Gastro-Intestinal (GI) Cancers

Pranita Atri, Ph.D.

University of Nebraska Medical Center, 2021

Advisor: Surinder K. Batra, Ph.D.

Gastro-intestinal (GI) malignancies, including gastric, colorectal, and pancreatic cancers, have maintained their high overall mortality due to a lack of prognostic and diagnostic biomarkers and potential therapeutic modalities. While efforts have been made to improve both early detection and therapeutic interventions in these cancers, failure of conventional approaches have proven to be a big challenge, and alternate approaches are needed. Computational biology approaches owing to lesser time and more per target success rate offer a unique solution here. The current study explored the use of computational biology techniques to study the various aspects relating to GI malignancies. First, we sought to understand the role of mucins in colorectal cancer, which helped establish the role of MUC16 and its associated signaling in a subset of patients in colorectal cancer (CRC) as a potential therapeutic target. Interestingly, the role of MUC16 in CRC had remained unexplored up until this point. Further, we carried out a comprehensive study of all mucins in gastric cancer (GC). This study helped us identify and establish a 5-mucin prognostic panel for GC, proving to be highly beneficial in this high-mortality malignancy. Further, our study, for the first time, explored the presence of

intrinsically disordered regions (IDRs) in mucins. Interestingly, IDRs have known to have significant functional relevance and hence the high percentage IDRs found within various mucins have the potential to be extremely relevant therapeutic targets. Furthermore, our *in-silico* identification and pre-clinical assessment of the novel therapeutic ISOX showed extremely high efficacy of ISOX in pancreatic cancer, which can help improve the overall survival of this highly lethal cancer. Overall, this dissertation successfully applies computational tools in highly lethal GI cancers establishing various novel biomarker panels and therapeutic interventions.

Table of Contents

List of Figures

Abbreviations

5’FU – 5-flurouracil

AMOP-adhesion-associated domain

CMAP- Connectivity map

COAD- Colorectal Adenocarcinoma

CRC- Colorectal Cancer

D^2P^2-Database of Disordered Protein Predictions; CH, charge hydropathy;

ECM-extracellular matrix

EGF-epidermal growth factor-like domain

GC- Gastric Cancer

GEM- Gemcitabine

GI- Gastrointestinal

HP- Hyperplastic polyps

IDP- Intrinsically disordered proteins

IDR- Intrinsically disordered regions

MoRF- molecular recognition feature

MSI- Microsatellite instability

MSS- Microsatellite stability

NIDO-nidogen-like domain

PDAC- Pancreatic Ductal Adenocarcinoma

PPI- protein-protein interaction

PTS sequence- proline, threonine and serine sequence

SEA-sea-urchin sperm protein enterokinase and agrin module

SSA/P- Sessile serrate adeonoma/polyps

STAD- Stomach Adenocarcinoma

TA- Tubular Adenomas

TM-transmembrane

VNTR-variable number of tandem repeat domain

vWD-von Willebrand factor D domain

Acknowledgments

I want to start by thanking my mentor, Dr. Surinder K. Batra, without whom this journey would not have been possible. From the very beginning of my Ph.D., he has been the most supportive and encouraging mentor that a student can hope to get. He took a great deal of chance with me since my project and interests were entirely different from all of his current and previous students. While there were many days when I questioned my ability to carry this forward, his faith in me never wavered and that is his greatest gift to me. He cared for me and protected me like his own child, and in my weakest moments, his faith is what kept me going. I am eternally grateful to Dr. Batra for all that he has done for me. While I will never be able to repay him for all that he has given me, I will keep trying for the rest of my life.

I would also like to thank Dr. Sukhwinder Kaur for her constant mentorship and support. She has given her all to guide me through this process, and for that, I am extremely grateful. From teaching me the fundamentals of experiments, scientific writing, and presentations to supporting me through failures and encouraging me never to give me, she has done it all. She has always been available for me no matter what day or time and I am tremendously thankful to her for all her support.

I would also like to extend my gratitude to the other members of my supervisory committee. More specifically, Dr. Moorthy P. Ponnusamy for his guidance, support, and encouragement throughout my Ph.D. journey and for always being readily available to help. Dr. Dario Ghersi for helping me navigate through some extremely

challenging computational problems. Dr. Maneesh Jain for his suggestions and encouragement, and Dr. Punita Dhawan, Dr. Lynette Smith, and Dr. Apar Ganti for their constant guidance and for shaping me into the person I am today.

I am also extremely thankful to Batra Lab membersfor creating an extremely positive work environment for me and teaching me various aspects of science. Specially Dr. Parthasarathy Seshacharyulu for teaching and helping me with all the animal experiments. His patience and guidance mean a great deal to me. I would also like to extend my thanks to Dr. Gopalkrishnan Natrajan, Dr. Satyanaryana Rachgani, Dr. Ashu Shah, Dr. Seema Chugh, Dr. Raghupathy Vengoji, Dr. Sarvanakumar Marimuthu, Dr. Imayavaramban Lakshmanan, Dr. Andrew Cannon, Dr. Joseph Carmicheal, Koelina Ganguly, Sanchita Rauth, Frank Leon, Christopher Thompson and Sophia Kisling for all their invaluable suggestions, help and support. I also want to extend my gratitude to Kim from the IGBPS office, everyone from graduate studies and the biochemistry office staff- Karen, Amy, Coleen, April and Jeanette to navigate through the various requirements of graduate school.

I would never have been here if it was not for the support and love that I get every day from my amazing friends and family here and all over the world. No amount of gratitude is enough to thank each and every one of these wonderful individuals for understanding me, protecting me, and loving me beyond words. But above all, I am forever in the debt of my best friends and parents, Vandana Sharma and Devender Singh, whom I can never thank for their sacrifices, love, and support.

Chapter 1: Introduction

Chapter 1A: Gastrointestinal Malignancies

1A.1 Overview of gastrointestinal malignancies with a focus on colorectal, gastric, and pancreatic cancers.

The term gastrointestinal (GI) malignancies encompass all the malignancies related to the gastro-intestinal tract and further in the related organs, mainly esophageal, gastric, colorectal, pancreatic, and liver cancers. Considering the pivotal role of these organs in normal human functioning, these cancers have an unfortunate level of mortality. Amongst the various reasons for this high mortality are late diagnosis, a lack of potent therapeutic options, and an overall lack of understanding of the complex biology of the disease(s). Considering this, the goals of this dissertation are to apply computational tools for biomarker prediction, understanding the tumor biology, and therapeutic interventions in GI malignancies. The subsequent sections discuss the specific aspects of colorectal, gastric, and pancreatic cancers that the dissertation covers.

1A.1.1 Colorectal Cancer

Colorectal cancer (CRC), often also referred to as either colon or rectal cancers, is a disease group encompassing the malignancies relating to the large intestine (*1*). Over the past many years, CRC has maintained its high mortality, wherein currently it is the third leading cause of cancer-related deaths in the United States (Cancer Statistics, 2021). Further, in addition to this high mortality, CRC is also the third most common cancer hence increasing the public health relevance of studying the disease (*2*). Amongst the various risk factors, genetic predisposition

mixed with lifestyle factors like diet, alcohol use, and smoking has been established to be of most significance. While lifestyle and diet-based changes in conjunction with progress in diagnostic and screening methods have brought about some improvement, the overall survival for CRC patients remains dismal. Furthermore, personalized therapeutic approaches based on gene expression or other biological parameters have further helped in increasing overall survival (*3*). Although improvements have been made in survival, the heterogeneity of the disease makes it extremely hard to effectively target a wide population, and hence the overall survival continues to be appalling, signifying the need for a better understanding of the disease.

1A.1.1.1 Colorectal Cancer Development.

Pivotal to this understanding of the disease is the processes and pathways involved in the early development and progression. In general terms, CRC can develop through two major pathways- the convention pathway consisting of a transition from polyps to adenomas to carcinoma or the sessile serrate pathway, which develops from hyperplastic polyps to sessile serrated adenomas/polyps (SSA/Ps) to carcinomas (*4*). Interestingly, each of these pathways has unique clinicopathological and molecular features and hence leads to phenotypically unique disease types (*5*). Some of these key differences are highlighted in each of the sections below.

1A.1.1.1.1 Precursor Lesions of Colorectal Cancer

A. Tubular Adenomas/Adenomatous Polyps.

Tubular Adenomas (TAs) or adenomatous polyps are irregular growths or protrusions formed on the inside of the intestinal canal. While many of these polyps are non-cancerous, a combination of chromosomal instability (CIN) and accumulation of APC, KRAS, TP53, SMAD4 PIK3CA, and other mutations lead to the progression of these polyps to adenomas and eventually to invasive carcinoma (*6*). Various studies have established TAs to be the most common colonic polyp, with these tube-like polyps comprising 80% of all polyps detected. Furthermore, while these are not most likely to become cancers, ironically, 80-90% of CRC tumors arise from these precursor lesions. Considering the historical importance of these lesions, the most widely used CRC tumor models are designed to mimic these pathways. Additionally, adenocarcinomas arising from these conventional CIN pathways are more likely to have microsatellite stability (MSS), whereas those from the alternative serrated pathways are more likely to have microsatellite instability (MSI).

B. Hyperplastic polyps

Another type of colonic polyps known as the hyperplastic polyps (HPs) is often regarded as harmless since these are essentially non-malignant. However, morphologic similarities with other lesion types make it almost impossible for these to be classified through colonoscopy. Furthermore, recent advances have identified an extremely small subset of HPs, which are believed to be neo-plastic

and give rise to tumors. This, however, warrants further studies considering the morphological similarities between HPs and other neoplastic precursor lesions and the lack of lesion-specific biomarker panels (*7*).

C. Sessile Serrated Adenomas/Polyps

Furthermore, a subset of CRC (~10-20%) develops through a non-traditional serrated pathway which is characterized by the presence of sessile serrate adenomas/polyps (SSA/Ps). Unlike the flattened tube-like TAs, SSA/Ps are characterized by a jagged or serrated morphology. Further, the carcinomas that arise through this pathway are characterized by genetic alterations in BRAF and a CpG island methylator phenotype (CIMP). Additionally, this subset of carcinomas is more likely to have high-level microsatellite instability (MSI-H).

The distinct morphological and molecular differences between these precursor lesions and the carcinomas that develop through them bring about pertinent questions concerning the gaps in our understanding of CRC, and further studies are warranted.

1A.2 Gastric Cancer

Gastric Cancer (GC), a cancer of the gastric track or mainly stomach is a major global health concern considering its high morality and the high total number of cases wherein currently it is the fifth most common cancer worldwide and the third leading cause of cancer-related deaths (*8*). Additionally, risk factors like

Helicobactor pylori and Epstein-Barr virus infections in conjunction with diet and alcohol consumption make controlling the disease extremely difficult. While improvements have been made in the past few years, survival remains abysmal mainly due to late diagnosis, early metastasis, and lack of potent therapeutic options for advanced GC. Furthermore, the clinical presentation of GC and the similarities between early symptoms of the tumors and other gastric problems complicated diagnosis even further. Various predictive and diagnostic biomarkers like CEA, CA19-9, MUC2, CD10, CD31, etc., have been explored and have also been successful in certain cohorts; a consensus in both diagnostic and predictive biomarkers is missing (*9*). A better understanding of the disease biology would lead to the identification of better biomarker panels and therapeutic interventions.

1A.3 Pancreatic Cancer

1A.3.1 Survival and therapy

Over the past many years, Pancreatic Cancer (PC) has remained one of the major causes of cancer-related deaths, mainly due to a lack of therapeutic options. The Pancreatic Cancer Action Network (PanCAN) blames the lack of a promising therapeutic modality as the major cause for our inability to reach a significant improvement in median overall survival (MOS) in PC patients. Over the past decade, gemcitabine (GEM) based chemotherapy has been established as one of the most promising therapeutic modalities for PC patients. However, major limitations to GEM mediated chemo-regimen are toxicity, lack of specificity towards molecules specifically altered in PC, ineffectiveness in a subgroup of patients, and

poor penetration over hypo-vascularized dense PC stroma (*10*). Through the years, hundreds of clinical trials have been conducted using GEM and non-GEM (Paclitaxel or 5-FU) based chemo-radiotherapy, but a minor improvement has been observed in the median overall survival (MOS) of PC patients (4.2-11.1 months). Over the past two decades, attempts had been made by combining the cytotoxic agent Gemcitabine (GEM) with molecular targeted agents as an alternative strategy in PC, but all these attempts have been in vain, though the initial response of FOLFIRINOX (oxaliplatin, irinotecan, leucovorin, and 5FU) compared to GEM alone was dramatic it showed a significant rate of grade 3/4 toxicity in PC patients (*11*). The initial response of US-FDA approved combination therapy of GEM with erlotinib (EGFR specific) is appreciable, the majority of PC patients are not responsive due to the existence of a KRAS mutation and compensatory pathway activation of other HER family proteins (*11*). Of note, the majority of USFDA approved therapeutic modalities fail to render their required clinical outcomes due to toxicity, off-target effects, and therapy resistance; hence there is a compelling need for the identification of new potent therapeutics.

Chapter 1B: Mucins in Gastrointestinal Malignancies

1B.1 Overview of Mucins

The protective mucus layers that cover various body cavities, including the gastrointestinal cavity, are made up of epithelial cells, which then home many members of the mucin family of glycoproteins. These glycoproteins are affected the pathological transformations and are known to play important roles in cancer initiation and progression. Further, this 22-member family is mainly divided into two major groups, namely transmembrane and secretory mucins. These two groups are distinct in their molecular functions as well structural variations wherein each of the groups is classified by its own set of domains. These domains are known to have specific functions in various malignancies. Specifically, in gastro-intestinal malignancies, mucins have been studied for therapeutic potential and as biomarkers and play significant roles in tumor progression.

1B.2 Overview of Mucins in Colorectal Cancer

Specifically, in colorectal malignancies, overexpression of MUC1, MUC5AC, and MUC17 has been reported. Loss of MUC4 and MUC2 has been reported as we move from the polyp to adenoma to carcinoma stages. Furthermore, the serrated pathway has been reported to have a different expression pattern wherein MUC5AC is lost as we move from sessile serrated polyps to adenomas. While these studies relating to specific mucins provide some insight, this varied evidence warrants the need for more in-depth studies. A comprehensive assessment of mucins in colorectal cancer has not been carried out to date and can be extremely

helpful in a better understanding of the disease biology and can have application in therapeutic interventions and biomarker activity.

1B.3 Overview of Mucins in Gastric Cancer

Further, in GC- MUC1 has been associated with *Helicobacter pylori* infection which is a known risk factor for GC. Additionally, a study assessing the prognostic relevance of MUC1, MUC2, and MUC5AC found downregulation of MUC1, aberrant expression of MUC5AC, and de-novo expression of MUC2 in GC patients. Additionally, various studies have found overexpression of MUC13 in various subtypes of GC. Furthermore, a recent study has correlated the mutational load of MUC16 with tumor mutation load and survival of GC patients. Considering this varied evidence and the lack of in-depth studies, a comprehensive study of mucins in GC will prove to be extremely beneficial.

Chapter 1C: Intrinsically disordered proteins

1C.1 Overview of intrinsically disordered regions/proteins

While human proteins are believed to be highly structured and ordered, within the human proteome is a subset of proteins that function without the need to require a unique structure. This subset of proteins that functions in a structure-independent way has been classified as intrinsically disordered proteins (IDPs), and the regions within these proteins that cause this disorder are referred to as intrinsically disordered regions (IDRs). This disorder, in turn, affects various downstream biological functions, including the protein-protein interactions, often turning these IDPs into protein hubs. Further, various studies have suggested that the prevalence of IDRs increases with an increase in complexity of an organism suggesting that that the human proteome is most susceptible to the presence of these IDRs.

1C.2 Intrinsically disorders proteins in disease biology with a focus on cancer

This said prevalence then translates into specific functional differences such as translation, alternate splicing, signaling, etc., making IDPs specifically important in biological diseases. Various studies have implicated the role of IDPs in disease biology, more specifically in diseases like cancer (*12, 13*), diabetes (*14*), cardiovascular defects (*15*), and neurogenerative disorders (*16, 17*). Specifically, in cancer, IDPs are of great importance considering the importance of these proteins in cellular processes like translation, transcription (*18*), and cell cycle

regulation (*19*). Further, IDPs play a very important role in regulating cellular machinery like ribosomes, chromatin organization, and various processes related to microfilaments and microtubules. These processes, in turn, have been established to have extremely important roles in disease pathology, specifically in cancer biology. IDPs roles in various cellular processes, in conjunction with specific tumor suppressors and other tumor-relevant proteins, have a high propensity of being intrinsically disordered, which warrants the need to study IDPs in cancer biology.

1C.3 Methods to assess intrinsically disordered regions

Various in-vivo and in-vitro methods can be applied to identify IDRs in any given protein, specifically NMR spectrometry and X-Ray crystallography-based techniques. These however, are limited due to high cost and lack of 3-D structures. Considering this, recently, computational methods have gained popularity and are now being more readily utilized for IDR prediction. There are currently over 25 prediction tools that are available as stand-alone tools, web-based tools, or python scripts. In this regard, the database for disordered proteins (D^2P^2) is a compilation of 9 such prediction tools and hence offers a deeper understanding and a higher significance level.

Chapter 1D: Connectivity mapping in therapy

Understanding drug-disease interaction remains one of the most challenging aspects of effective cancer therapeutics. Over the past few decades' large-scale genomic sequencing and global gene expression profiling, which in turn has led to the development of big data-driven resources like the connectivity map (CMAP), has become common play. CMAP a source perturbation database by the Broad institute, can potentially help biologists understand and build the said drug-disease relationship. The review's focus is to assess what has already been done concerning CMAP and what can be done further to understand these complex interactions better to make effective and well-informed clinical decisions.

1D.1 Introduction

Conventional methods of drug development involve the assessment of one drug target at a time, making it extremely difficult to target a complex disease like cancer. In recent times, computational assessment to identify potential drug targets has become common (*20-22*) However, up until this point, there has not been a clear understanding of what resources are available and what can be done to make these assessments. One of the biggest of such resources is the Connectivity Map (CMAP), a large-scale dataset of gene expression profiling from over 2800 drugs on various cancer cell lines. This humongous data collected by the broad institute is now available for us to query and connect with various user-defined gene signatures. This resource can be beneficial for chemists, and

biologists alike wherein chemists can leverage from this in the process of drug designing and biologists can identify mechanisms by existing and new drugs act.

The advent of large-scale genomic sequencing and global gene expression profiling has led to the discovery of large-scale perturbation databases, like the Connectivity Map (CMAP). CMAP was a portal developed by the broad institute with the sole aim of bridging the gap between physicians, chemists, and biologists to make better-informed decisions about patient care. The major question that Justin Lamb et. al. answered if the use of big data analytics can help the scientific community make better decisions to aid in pre-clinical studies that eventually can lead to better clinical decisions and effective therapy. A major issue that they identified was a lack of knowledge of the connection between diseases and drugs. Considering this, the aim of their study was two-way, first to find this information and second to make it easily accessible without the need for specialized computational skills. Thus, CMAP was put together containing gene-expression data obtained from genetic variations of genes and treating human cell lines with chemicals and reagents, which they refer to as perturbagens.

The next question at hand was, even if the connections were made, how it would lead to the identification of new drugs. This called for a proper scoring scheme to quantify these said connections. CMAP works on a simple scoring method wherein a user-defined gene signature (up-regulated and down-regulated genes from a disease condition) is connected to the drugs' gene signature (upregulated and downregulated genes observed when the human cell lines are treated with the said perturbagens) and a positive, negative and neutral score with values ranging from

–1 to +1 which directly reflect the strength of this connection are obtained. The initial set of perturbagens contained 164 drugs which were selected over a broad spectrum of activities. There were multiple drugs with the same target (the most common example being histone deacetylase inhibitors) and efforts were made to determine the similarities and differences between the mode of action of these inhibitors. This information will be extremely helpful to biologists since the subtle differences in the mode of action of inhibitors targeting the same family of proteins could answer that we have been searching for so long.

The initial set of 164 was just the beginning, and as of the last update, CMAP contains data from over 27000 perturbagens ranging from various small molecule compounds to gene knockout signatures. The method and the subsequent updates have all followed the simple underlying principle that gene expression is the “universal language” of drug perturbations. Over the years, CMAP has continued to make great heaps and jumps over these many years. The most important came up just last year wherein a new age CMAP was launched with almost a 1000-fold scale up from the existing data. This collaborative effort of the Broad Institute with the National Institutes of Health’s (NIH) library of integrated network-based cellular signature (LINCS) was developed to help make clinical decisions mainly on side effects of exciting drugs to aid in clinical trials.

In complex diseases including cancer, the understanding between the relationship of a drug and the disease could very well be the key to effective targeting. This review aims to compile all the applications of CMAP in various disease models as

well other modifications to assess the applicability of this great resource for biologists and pharmacists (*19, 20*)

1D.2 How Was the Connectivity Map Developed?

As J. Lamb and colleagues rightly mention, for a very long-time, gene expression profiling has been applied specifically to the pathway-driven analysis of disease conditions or, more recently, the determination of “gene signatures” for various subtypes of cancers. The vision that the authors had was to move beyond this historical role and make the most use of these profiles that are becoming more and more readily available. The idea was to find a better way to harvest and assess this big data so that we can learn from the past in moving towards the future. While considering all this, they came across a landmark study (*23*) wherein, in a yeast model, it was established that gene expression data can be directly correlated to functional responses to small molecules and genes. However, if this would translate to mammalian cultures was something they still had to test and hence came about their hypothesis that gene expression data can be used to make assessments of drug responses in various disease settings (*23*).

1D.3 Modifications of CMAP

Over the years, CMAP has been modified outside its initial framework to be applied in various other forms, many of which have extremely important roles.

1.D.3.1 Evaluation of RNA interference and CRISPR through CMAP.

While CMAP started as a perturbation dataset, the addition of genetic perturbations opened avenues for other applications. One such application is the

evaluation of RNA interference and CRISPR using this data. Both RNAi and CRISPR have been extremely helpful in assessing the loss of function of a wide range of targets. However, the question that Ian Smith and colleagues (*24*) wondered about was the off-target effects of these technologies. Even though it is known that both these techniques have off-target effects, a comprehensive study had not been conducted to compile these effects.

As a part of the "new" CMAP, the gene consequences of over 13,000 shRNAs can be studied in the same 9 cell lines as used for initial perturbagens based studies. Smith et al utilized this information to develop a consensus gene signature (CGS) which is a compilation of all off-target and targeted effects of the said shRNAs. Furthermore, to better understand the targeting of each shRNA, the authors established a consensus seed signature (CSS) which was also analyzed across cell lines to identify cell line-specific activity of particular shRNAs. This resource can now be utilized for designing experiments pertaining to the use of these technologies. Further on, they performed a projection analysis to assess a gene expression analysis with a combination of on-target activity, off-target effects, and assay noise. This data is now available for further assessment.

1.D.3.2 Drug-Induced Apoptosis Subnetwork From CMAP Data.

Another such ingenious application of this data came about in 2015. Considering the importance of tumor-selective cell death for effective cancer therapy Jiyang Yu and colleagues (*25*) used the breast cancer-specific MCF7 data from CMAP to model a subnetwork specifically for apoptosis. They applied Gaussian Bayesian network approach to this meta-data from all the drugs and identified apoptosis as

a major drug-induced cellular pathway. They then identified the major genes causing these effects and came up with 13 apoptotic genes that showed differential expression across all drug-perturbed samples. These 13 genes were used to make what they decided to call the apoptosis subnetwork. The identification has potential applicability in the drug designing process wherein an effective apoptosis therapy should target the newly identified 13-gene signature. This analysis can also be carried out for other cancer-specific pathways to make discoveries.

1. D.3.3 Functional Module Connectivity Map.

Another application of the CMAP data which falls into the realm of drug repurposing and drug development is the functional module connectivity map (FMCM). As the authors rightly discuss, drug discovery is an expensive and long process often due to a lack of understanding of the underlying mechanism in a disease model and hence potential. Gaining this understanding becomes even more difficult in a complex disease like cancer which has various complex biological processes associated with its initiation and progression. Based on these issues and the availability of the large CMAP dataset, Chung et al in 2014 set out to assess if they could establish a computational drug screening procedure that addresses these complex issues and hence came out the FMCM. The authors taking colorectal adenocarcinoma as an example, used the CMAP data and combined it with a functional assessment to come up with a functionality-based application of CMAP. The study assesses gene-gene interaction-based functional networks developed in adenoma and normal cases and the use thereof to identify drugs specific to these varied functional modalities. The authors went on to

compare the individual gene query to this functional module-based assessment and propose that the functional module renders a better assessment of potential therapeutics (*26*).

1. D.3.4 Reversal of CMAP-Identification of Genes.

The original idea behind CMAP was simply to have a better understanding between drugs, genes, and diseases however, it has been applied to direct identification of small molecules rather than identification of the underlying genes associated with the disease condition and this is what Liu et al set out to explore in this 2017 study. The overall premise of this was to identify specific gene targets which are perturbed by drugs.

To answer this, they assessed genes significantly affected in the CMAP project and denoted it with differential expression number (DEN), and compared the genes with high DENs with the others. They further carried out a network topology-based analysis and explored the subcellular localization of these genes to decipher the potential connection with disease conditions (*27*).

1. D.3.5 Compound Carcinogenicity

The toxicity of compounds has been an age-old question that researchers have struggled with in drug discovery. In an interesting application, Caiment et. al. use the CMAP and hepatocellular carcinoma (HCC) data to develop a model to use the CMAP data in toxicity analysis. The first step of this process was to develop a gene signature for liver toxicity which was done with the help of gene expression data from various datasets containing normal liver samples (*28*). This gene

signature was then used in comparison to the gene signature from toxico-genomics datasets from various tumor samples to identify carcinogenicity of various samples.

1. D.3.6 Proteomic Connectivity Map

The original study was further extended to include “a library of phospho-proteomic and chromatin signatures for characterizing cellular responses to drug perturbations” or a proteomic connectivity map. The proteomics connectivity hub is an effort that the authors define as being complementary to the original CMAP study in adding the liquid chromatography-mass spectrometry (LCMS) data to measure the phospho-signaling and chromatin state both of which can add to the information surrounding the mechanism of action (MoA) of any drug perturbant. The question that the authors were faced with was while gene expression profiling had been becoming more and more relevant to drug perturbation studies, are these profiles the “universal language” of the readout of these drugs or other perturbagens. Based on this information and the basic “connectivity” framework, they profiled 90 drugs in 6 cell lines to study the global chromatin profile (GCP) and a reduced representation proteomic assay (P100). This data now can be used to study the overall state of the cell with respect to any said perturbations. The resource is open access and can be used through the CMAP portal to assess the “proteomic signatures” (*29*).

1. D.3.7 Epigenetics Connectivity Map.

Within the same realm as the proteomics study, CMAP is another such application, the epigenetics connectivity map. Epigenetic marks or chemical modifications to DNA and histone proteins play a central role in gene regulation, but many gene expression compilation studies have been presented similar studies are missing for these epigenetic marks. The study uses a SILAC mix which is an equal combination of Hela, 293T, and K562 for histone modification studies (*30*).

1. D.4 CMAP Based Studies in Gastro-Intestinal Cancers

1. D.4.1 Pancreatic Cancer

Pancreatic cancer (PC) and its most deadly subtype, pancreatic ductal adenocarcinoma (PDAC), is currently one of the deadliest cancers. In 2020 alone, over 47,050 people were projected to die from this disease. Unfortunately, the disease is extremely hard to target by therapeutics mainly due to the inherent drug resistance of these tumors (*31*). Over the years, clinical trials and laboratory-based studies have explored potential therapeutics; however, significant improvements have not been seen in the overall survival of these patients. Considering this, a tool like CMAP can be extremely useful in identifying potential new therapeutics. CMAP has been applied indirectly to various PC studies as summarized here.

1. D.4.1.1 Identification of New Therapeutics for Dasatinib Resistant Cell Lines

Dasatinib, an FDA-approved tyrosine kinase inhibitor, has shown great potential in PC (*32*). However, as common for many other drugs PC cells have the potential

to gain resistance to this drug. To identify drugs with a potential synergistic effect with dasatinib, Chien et. al., used RNA expression from 3 dasatinib-resistant and 3 dasatinib-sensitive and the CMAP data to identify FDA-approved thioridazine. Thioridazine was found to cause apoptosis, affect the cell cycle and multiple kinase activities. Additionally, it was identified as a protein phosphatase 2 (PP2A) inhibitor, which led to the identification of PP2A and its subunit as being potential targets for effective PC targeting (*33*).

1. D.4.1.2 Identification of Inhibitors Synergizing With Gemcitabine

On a similar note, Jian Lin Er et. al. in 2018 (*34*) presented a similar study for gemcitabine-resistant cells. The study aimed to specifically look at the squamous subtype and identify potential therapeutic agents using CMAP in concordance with the International Cancer Genome Consortium (ICGC) and the Cancer Cell Line Encyclopedia (CCLE). The aim of this study was mainly two-fold: one to identify drugs specific to the squamous subtype of PDAC and the second for this identified therapeutic to be synergistic with gemcitabine. Differential gene expression from the squamous specific samples was subjected to a CMAP analysis to identify 26 candidate drugs that were tested in squamous type cell lines. These cell lines were further subjected to combination treatment with gemcitabine to identify the most synergistic drugs leading to the identification of SRC or MEK inhibitors as potential synergistic drugs.

1. D.4.1.3 Metabolic Networks In Response To Perturbations In Pancreatic Cancer.

To add further to this quest of using CMAP to treat this horrific malignancy, Biancur et al (*35*) assessed the CMAP data to identify drugs for the synergistic effect with glutamine inhibitor CB839. Metabolic reprogramming has been shown to be an integral part of pancreatic cancer initiation and progression. The premise of the study is that specifically in PDAC the conversion of glutamine to glutamate catalyzed by the enzyme glutaminase (GLS) leads to an increased reducing potential in the form of NADPH and glutathione (GSH). The question that leads to is whether GLSi is a therapeutic strategy worth pursuing and if it is then can a CMAP analysis help identify drugs with a synergy with GLSi therapy. They utilize proteomics data to identify significantly synergistic target drugs which can further be utilized for combinatorial therapeutic approaches.

1. D.4.2 Colorectal Cancer

Colorectal adenocarcinoma is currently the third leading cause of cancer-related deaths in the United States. The cancer genome atlas (TCGA) contains within itself has the RNA-seq data from 437 colorectal cancer tumors and 39 control samples (COAD) which can be utilized for big data studies. W.D. XI et al (*36*), used the COAD data to assess first the differential gene expression and then further to assess the CMAP to identify small molecule drugs, which are potential therapeutics for colorectal cancer. Out of the drugs identified, the histone deacetylase (HDAC) inhibitor; scriptaid is observed to have chemo-sensitization on human colorectal cancer cells. In a similar study, Qing Wen et al used five

microarray datasets from colon cancer and used them to identify a query gene signature for the identification of potential therapeutics. Interestingly, the CMAP identification from this gene signature helped the authors identify 10 potential drugs, including current chemotherapies validating the method. The combination of both these studies is suggestive of the potential use of HDAC inhibitors in colorectal adenocarcinoma, but further studies are needed to establish these as potential therapeutics. A further application of CMAP in colorectal cancer came about more recently in 2014 by Wen Q and co-authors. The article is one of the quintessential applications of CMAP to identify wherein they combine 5 colorectal cancer (CRC) microarray datasets with normal and tumor samples. They further go on to rank the various differentially expressed genes in each of these datasets to establish a combined gene signature which they go on to test for potential drug matches through CMAP. Interestingly, many of the negatively connected drugs were current chemotherapies in use (*37*).

1. D.4.3 Gastric Cancer

To add to these wonderful studies, CMAP found another of its application in gastric cancer (GC) wherein Li Zhang et al used a weighted gene co-expression network analysis to identify HDAC2 inhibition using lovastatin as a potential therapeutic modality for GC. This comes as a breakthrough for GC since the long-standing treatment option has been surgery. They used gene expression data from GC to first construct a gene expression network of differentially associated genes and then subsequently use a combination of the network and functional analysis to

identify valproic acid and lovastatin as potentially therapeutic for GC. Interestingly, HDAC2 was identified as a target for both drugs suggesting it as a potential target ineffective GC targeting. (*38*) A further study was carried out (*39*) in GC by Zu-Xuan Chen and colleagues wherein TCGA tumor data was compared to normal data from genotype-tissue expression (GTEx) to identify differentially expressed genes in between tumor and normal samples. The DEGs were then used for a CMAP analysis to identify potential therapeutics for GC, establishing another successful use of the CMAP data.

1. D.4.3 Esophageal Carcinoma

An application of the CMAP data was also seen in the esophageal carcinoma (ESCA) wherein Yu-Ting Chen and colleagues first used the TCGA and GTEx data to identify differentially expressed genes followed by a functional analysis of the identified genes then the excavation of hub genes from this data followed by a prediction of drugs associated with these identified genes. They went on to do a detailed analysis of the identified drugs and the corresponding gene networks, a functional analysis as well as a docking analysis to study the potential interactions of the identified drugs with key molecules relating to ESCA hence establishing a comprehensive method that will find its application beyond the current study (*40*).

1. D.4.4 Liver Cancer

Another application of the CMAP data was carried out by Li-Min Liu and colleagues for the horrid disease of liver cancer. While exploring the use of nitidine chloride (NC) for this horrid disease, the authors set out to independently elucidate the

potential regulatory mechanism of this alkaloid. They used differentially expressed genes from a microarray study for NC treatment to connect to the various drug signatures within CMAP to identify pathways that can be affected by NC treatment, hence establishing a unique application of the CMAP data. They found that this analysis could successfully identify pathways like cell growth and hence supplement the gene set enrichment and in-vitro data (*41*).

1. D.4.5 Hepatoblastoma

Hepatoblastoma (HB) is the most common hepatic tumor in infants and children, claims for half of all liver tumors in children. The best treatment option is surgery followed by liver transplantation. However, the increase in liver-based illness and hence the subsequent increase in patients needing a transplant warrant for other treatment options. Specifically, for rare diseases like hepatoblastoma, drug discovery through CMAP can be extremely helpful since its extremely affordable (*42*) . In this backdrop, Beck et al in 2016 (*43*) applied the CMAP data to identify 13 potential drugs for high-risk HB, including 2 PI3K/AKT inhibitors which already have a known application in HB. Furthermore, the data helped them identify 2 HDAC inhibitors that have a potential application in HB since HDACs are known to be overexpressed in HB and hence the potential application of HDACi in HB. They further studied the potential of the HDACi SAHA to sensitize HB to cisplatin and doxorubicin, which showed a synergistic effect of SAHA and cisplatin. Through a combination of in-vivo and in-vitro studies, the authors successfully identified a

combination of HDACi and cisplatin as a potential therapeutic strategy for high-risk HB.

1. D.4.6 Kidney and Renal Cancer

Primary renal cancer is one of the few cancers which is curable at early stages, but metastatic renal cell cancer shows a poor overall survival. The identification for a potential therapeutic using CMAP is a promising application in this cancer partly due to well-established gene expression profiling, which supports the fact that it remains one of the only cancers wherein more than one application of CMAP has been seen. The first of which came in 2013 from Zhong Y et. al. (*44*) wherein they queried the 1300 drugs in CMAP for gene expression profile of mice with HIV-associated nephropathy a condition that often leads to renal cancer. What they were looking for was a potent combination of drugs that could help in reducing this condition. They identified that a combination of angiotensin-converting enzyme (ACE) inhibitor and a histone deacetylase inhibitor potentially the best combination for this condition. They went on to test this combination in the same mice and observed that the combination could help in reducing the condition significantly. The second application was specifically with respect to renal cell carcinoma (RCC), wherein Zerbini LF *et al (44, 45)* used their previously established signature together with a gene set enrichment analysis to identify potential FDA-approved drugs for clear cell renal cell carcinoma. They established a workflow, Individualized Bioinformatics Analysis (IBA), which first establishes a differential gene expression in healthy *vs.* the disease tissues and further uses these for a

CMAP analysis. They successfully identified various current chemotherapies as well as potential new therapeutics. Interestingly, these high-scoring drugs induced a high level of apoptosis in RCC cell lines. The further assessment led to the identification of pentamidine as a potential therapeutic for RCC with high efficacy in RCC mice models and affecting major cancer-specific pathways (*46*).

1. D.5 CMAP Based Studies in Other Cancers

1. D.5.1 Medulloblastoma

A cerebellar primitive neuroectodermal tumor (PNET) or medulloblastoma is the tumor that starts in the base of the skull, or the posterior fossa is cancer commonly found amongst children. Medulloblastoma has been divided into clinical subtypes wherein Group 3 has the worst prognosis of all (*47*). One of the first applications of CMAP based drug identification was within this subgroup. Fara CC et. al. in 2015, applying a subtype-specific CMAP analysis, successfully identified alsterapullone as a novel small molecule inhibitor to target group 3 medulloblastoma. They went on to assess the efficacy of alsterapullone *in-vitro* and *in-vivo* showing a very high efficacy, specifically targeting cell cycle genes in group 3 medulloblastoma showing potential applicability (*48*).

1. D.5.2 Acute Myeloid Leukemia

Acute myeloid leukemia (AML) is the cancer of the bone marrow that makes abnormal myeloblasts, red blood cells, or platelets. While a lot has been studied about AML therapy, the disease still takes the lives of over 11,000 people within a

single year with a 23-25% overall survival. Previous studies in the field have shown an inactivation of C/EBPα in AML cells, a transcription factor that is known to be mutated in over 10% of AML patients. However, the therapeutic targeting of the gene has not been successful. With this background, Manzotti et al (*49*) decided to use the CMAP data to specifically look for small molecule inhibitors, which would specifically target C/EBPα. They first identified the C/EBPα signature, which consisted of genes regulated by the transcription factor followed by the identification of small molecule inhibitors that mimic the effect of the biological effect of the transcription factor C/EBPα. They identified eight drugs as potential therapeutics in AML. They found ATRA a drug commonly used in AML as one of the potential drugs, and tried various combinations of the other (diperodon and amantadine) identified drugs with ATRA. GSEA assessment revealed similarities between the C/EBPα affected genes and the downstream effects of treatment with the small molecule drug amantadine. Diperodon however, failed to achieve a profound difference in the effector gene expressions supporting the use of amantadine over it. The identification of both these inhibitors in combination with ATRA is a promising strategy that can be used in AML.

1. D.5.3 Acute Lymphoblastic Leukemia (All)

Acute Lymphoblastic Leukemia (ALL), like its counter AML, is a blood and bone marrow cancer but, unlike AML, affects the white blood cells. ALL is the most commonly occurring childhood cancer which was slated to claim 24% of the lives in 2020 (*50*). A subset of ALL, mixed-lineage leukemia (MLL)-rearranged infant ALL is specifically even more aggressive than the other subtypes. It is

characterized by its unique gene-expression profile which can then be targeted using specific inhibitors. Considering this Stumpel et. al. applied the CMAP identification in this specific subset of patients led to the identification of the HDAC inhibitors as a potential treatment for this horrid disease. Before conducting this analysis, the authors had previously recognized characteristic promoter hypermethylation in these MLL rearranged infant ALL through this study; however, they establish a specific subset of genes which were hypomethylated and overexpressed in these patients. This gene set was then used for CMAP analysis, which led to the identification of various HDAC inhibitors. The authors went on to validate these assessments establishing the potential of epigenetic therapies in this specific ALL. Off note, this study, for the first time, successfully established that CMAP analysis can also be applied to DNA methylation patterns and hence establishing another potential application of DNA methylation data in drug discovery studies (*51*).

1. D.5.4 Breast Cancer

Over the years, Breast cancer (BC) has remained one of the most common causes of death in women in the United States. While a lot has been studied the authors, Fang E et al (*52*), observed gaps in the knowledgebase for the molecular mechanism of BC. They hypothesized that identification of hub genes and the corresponding pathways would, in turn, lead to the identification of the molecular mechanism of the disease, followed by identification of potential therapeutics using CMAP. They obtained the differential gene expression and the subsequent functional implications in BC. Furthermore, they identified genes with a prognostic

significance in BC since those would be the targets to be assessed as potential therapeutic targets. This helped them identify the small molecule agent “emetine” which might have potential application in BC. Similar to this study was a meta-analysis carried out by Thillaiyampalam et al in 2017 wherein public data repositories were used to analyze over 7000 BC samples led to the identification of 26 potential therapeutics for BC. Interestingly, 14 out of these were known, anti-cancer agents. The methodology established in this study offered a framework of combining various datasets while accounting for batch effects and which in turn can be applied to various other disease states utilizing the abundant gene expression data. In another study in BC, Busby et al combined epidemiological evidence with the CMAP approach to assessing existing medications that alter BC risk. The CMAP correlations led them to identify a total of 10 drugs with 6 cancer-causing and 4 cancer-preventing drugs. Unfortunately, further analysis of the overall survival associated with these drugs did not lead to any conclusive results. However, this study opens the door to another application that can be explored much more in the future. A further application in BC was carried out by Tong Liu and colleagues in 2017 wherein the CMAP data was connected to a signature associated with the knockdown signature of FSIP1 specifically in a HER2+ setting. Interestingly, the identified compounds showed very high efficacy in BC and hence establishing another branched application of CMAP (*53-55*).

1. D.5.5 Prostate Cancer

Similar to their BC study, Fang et al carried out a hub gene and CMAP study in Prostate cancer (PCa), considering that it is the second leading cause of cancer-

related deaths among men in the US. Like BC the molecular mechanism of PCa is not yet fully established. To do this, they identified the differentially expressed genes between prostate cancer and normal cells, used this for a functional assessment using gene ontology (GO) and KEGG, and additionally to identify a potential drug for PCa (*56*). This analysis helped them establish the importance of cell adhesion molecules (CAMs) and TGF-beta signaling in PCa and identify the small molecule therapeutic: phenoxybenzamine as a potential drug for PCa. A similar study was carried out by Jian Li et al in 2013 to identify networks of differentially expressed genes and the associated small molecules which led to the identification of various pathways similar to the aforementioned study (*57*). Furthermore, another application of the CMAP was seen in the study by McArt et al wherein they use a combination of RNA-seq and microarray data from various PCa patients to identify the nicotine derivative cotinine as a potential therapeutic. Interestingly, through independent experiments, they further established its effect in reducing proliferation through an androgen-dependent mechanism (*57, 58*).

1. D.5.6 Lung Cancer

Another such different application was carried out by Youchum et al where they looked specifically at the gene knockdown signature of TWIST1 to identify potential TWIST1 targeting compounds. This led to identifying 6100 potential TWIST1 inhibitors that the authors scaled down to 8 using a combination of score assessment and a literature review. Further assessment with various non-small

cell lung models successfully established harmine as a TWIST1 inhibitor as well as a potential therapeutic for lung cancer (*59*).

Further, another study in lung cancer was carried out using gene expression data tyrosine kinase inhibitor (TKI) sensitive lung adenocarcinoma and TKI-sensitive clones from the same cell line and used the differentially expressed genes to identify HDAC inhibitor-valproic acid (VPA) as a potential therapeutic for lung adenocarcinoma. The study of VPA in lung adenocarcinoma could establish its effects in various cell line models hence establishing it as a potential therapeutic (*60*).

1. D.5.7 Ovarian Cancer

Epithelial ovarian cancer or EOC is one of the most lethal cancers accounting for over 5% of the cancer-related details. While therapeutic options are available, they are unfortunately not extremely effective, with most patients coming back with a relapse soon after treatment. Considering this, Rama Raghavan et al, used the CMAP data to connect gene expression data from tumor vs normal samples to identify 11 drugs, five out of which were independently validated and were shown to have a cytotoxic effect on ovarian cancer cell lines (*61*).

1. D.5.8 Bladder Cancer

Further, in bladder cancer, a bioinformatics-driven study was carried out to assess the role of DAPK1 as a prognostic marker and used the CMAP data to identify specific drugs against the DAPK1 signature. Interestingly, out of the top 10 small molecules identified, many were Braf/MEK/ERK pathway inhibitors with previously known potency in other cancers like melanoma. Further, they could successfully

validate the effect of one of the identified drugs, vemurafenib, both on DAPK1 and subsequently on bladder cancer (*62*).

1. D.6 CMAP Based Studies in Other Diseases

While cancer is the most common application, CMAP data has been applied to other diseases as well. Some of these studies are summarized in the subsequent sections.

1. D.6.1 Candidate Agents for Diabetes

Diabetes, a disease essentially characterized by a high blood sugar level, is mainly divided into type 1 and type 2 diabetes. Specifically, type 2 diabetes (T2D) results from an impaired insulin secretion and the effect it has on various target tissues. Recent studies have shown the underlying importance of metabolic syndrome (*63*) as a predictor of T2D, but the metabolic signature or the genes associated with causing these effects have not been very well studied. The authors, Wang et al, (*64*) decided to use a computational approach to gather more information about the underlying genetic mechanism behind the disease and then eventually assess the therapeutic targets for it. In order to do this, they extracted microarray gene expression data of human pancreatic islets with or without T2D, carried out differential gene expression, followed by a pathway analysis to identify dysfunctional pathways, regulatory gene networks, and finally, identification of candidate small molecule inhibitors. This systematic process led to the identification of potential therapies: sanguinarine and DL-thiorphan, which can act as successful therapies for T2D. In a similar study, Zhang et al (*65*) combined

gene-wide association studies (GWAS) with proteomics and metabolomics studies first to reveal specific proteins with underlying importance in diabetes and further to identify druggable targets and the corresponding drugs. The identified targets combined with a study of the associated pathogenesis led to the identification of 9 potential drugs which can be repurposed for diabetes treatment.

1. D.6.2 Sex Linkage Of Therapies In Cancers

In 2016, Ma J *et al (66)* identified a unique application of CMAP. Cancer researchers and epidemiologists have often alluded to a sex linkage of various cancer forms wherein the progression of the disease varies greatly with gender, but the question at hand at this point was if therapies should be gender-specific. A systematic analysis was then carried out utilizing CMAP to differentiate between genetic make-up of these diseases in men vs women. They utilized the RNA-seq from 17 cancers reported within the cancer genome atlas (TCGA) in order to identify genes, which had gender specific impact. They also assessed the differences in these tumors with respect to the corresponding normal samples. Interestingly, they found the sex-linkage in these tumors goes up to as high as 76.47% in melanoma. They were further went on to carry out a functional assessment of tumor vs normal samples which had been pre-stratified by sex. They found various pathways across different types of cancers which were significant in one or the other genders supporting the initial hypothesis of functional differences across these cancers. Furthermore, this was followed up by a CMAP

analysis in the sex-stratified patients which led to alarmingly different results in both the genders establishing a sex-linkage in cancer therapies.

1. D.6.3 Mutations and CMAP

Researchers over the years have established the importance of mutations in therapy response and therein lies the premise behind targeting significantly mutated genes (SMGs) found in the cancer genomes and this is what Cheng F. et al (*67*) set out to establish with this 2016 study. They utilized the TCGA data to identify these SMGs and established how these or their neighbors could be potential druggable targets. They successfully identified 693 SMGs across these various cancer types. Interestingly, many of these SMGs are currently targeted using various inhibitors. Further on, they curated a list of drug-gene signatures and compared it to the SMGs and the corresponding networks to establish a global drug-target interaction network. This led to the systematic identification of 121 druggable proteins encoded by the SMGs and FDA-approved drugs which can successfully target these hence establishing an effective drug repurposing strategy for mutations in cancer.

1. D.7 Limitations and Future Prospects

While the connectivity mapping approach has its merits, this ingenious method is not without its limitations. One of the major limitations is the points of differences between cell lines and tissues in general. Gene expression data from cell lines grown on tissue culture flasks and treated with the perturbagens are used to connect to tissue samples from patients to assess and identify potential therapeutics for various diseases; however, there are way too many differences

between these cells that grow in artificial conditions and live tissues, hence adding higher variability. There might be changes that the original expression techniques might have missed out on and hence remain masked. Further, handling differences and batch-to-batch variation in drugs being used can add to the variability, which can be mistaken for differences in responses. While these are notable issues, the CMAP team works continuously to ensure that these issues are met, and hence the data is getting updated time and again.

Considering this background information, the overarching aims of this dissertation were to make use of big data resources to address these many problems surrounding GI malignancies. These overall objectives study aims are elaborated further in the next chapter.

Chapter 2: Hypothesis and Overall Objectives

2.1 HYPOTHESIS AND OVERALL OBJECTIVES

Gastro-intestinal (GI) malignancies or more specifically colorectal and gastric and pancreatic cancers have maintained an extremely high molarity mainly due to the gaps in our understanding which in turn leads to poor therapeutic targeting and lack of potent biomarkers. Over the years, many attempts have been made to gain a better understanding of these malignancies in order to identify better biomarker panels and therapeutic targets but most of these attempts have, unfortunately, had limited success. Furthermore, considering the importance of mucosal barriers in these malignancies, the mucin family of proteins has remained central to the understanding of these cancers. While mucin based biomarkers and therapeutic targeting has shown promise, a better understanding of these proteins can prove to be beneficial. In this regard, recent advances in computational techniques leveraging the abundant patient data can prove helpful considering the reduction in time taken and better per target success rate. Considering this, this dissertation proposes to utilize big-data resources to assess mucin and non-mucin based biomarkers and therapeutic targets. The central hypothesis surrounding this study is *"a big data driven approaches taking advantage of the large-scale datasets will lead to a better understanding, better biomarker prediction and better therapeutic targeting in high molarity GI cancers.*" More specifically, the following studies were considered:

Aim 1: Global *in-silico* analysis of mucins in Colorectal Cancer.

The mucin family of glycoproteins is known to play a significant role in colorectal cancer (CRC). Over the years, studies have elucidated the role of MUC2, MU4,

MUC5AC, MUC16 and MUC17 but an overall consensus for the role of all 22-family members has not been carried out till date. Further, various datasets have become available both from early precursor lesions and CRC tumor samples. These datasets can now prove to be beneficial in order to study the role of mucins in CRC initiation and progression. Considering this, the central question to be answered from this part of the study is to establish the overall role of mucins in CRC using various computational resources.

Aim 2: Global *in-silico* analysis of mucins in Gastric Cancer.

Similar to CRC, mucins have been established to play significant roles in the initiation, progression, and metastasis of gastric adenocarcinoma. Specifically, MUC1, MUC2, MUC5AC and MUC13 have known to be of great importance with potential as both prognostic and diagnostic biomarkers and as therapeutic targets. However, a comprehensive analysis of the understanding of mucins in gastric cancer (GC) is missing. Considering this, this sub-aim of the study proposes to carry out a comprehensive expression, mutation and survival analysis of mucins in GC.

Aim 3: Presence and structure-activity relationship of intrinsically disordered regions across mucins.

Further, to explore the therapeutic potential of mucins and gain a deeper understanding, this sub-aim proposes to study the presence of intrinsically disordered regions (IDRs) across all mucins. IDRs have been established to play a significant role in downstream signaling of various proteins, however, no such study has been carried out in mucins. Considering the importance of mucins in

various malignancies it is extremely relevant to study IDRs in mucins. Considering this, this sub-aim proposes to carry out a detailed study of IDRs in mucins.

Aim 4: Connectivity Mapping-based identification and evaluation of ISOX: A novel therapeutic strategy for Pancreatic Cancer.

The final aim of this dissertation is to carry out an *in-silico* identification and validation of novel therapeutic for Pancreatic Cancer (PC). Considering the high molarity of PC, the lack of potential therapeutic options and the failure of the one target at a time approach, a systems biology approach targeting the global gene expression of PC can prove to be highly beneficial. Considering this, the sub-aim proposed to use the big data repository connectivity map to identify a specific and potent therapeutic for PC.

Chapter 3: Global *in-silico* analysis of Mucins in Colorectal Cancer identifies MUC16 signaling

3.1 SYNOPSIS

The mucin family of glycoproteins has been established to play a highly significant role in gastrointestinal (GI) malignancies. Specifically, in colorectal cancer (CRC) multiple studies have elucidated the role of MUC2, MUC4, MUC5AC, MUC16, and MUC17, however, a comprehensive in-depth study has not been carried out yet. In this regard, the availability of microarray and RNA-seq datasets and the improvement in our ability to harness this data can prove to be extremely useful. Considering this, the current study uses microarray data from early precursor lesions of CRC and TCGA-RNA-seq dataset to study the expression, mutation, and survival differences rendered by the 22-member mucin family. Our *in-silico* analysis reiterated the importance of MUC2, MUC4, MUC5AC, and MUC17 in CRC and additionally helped in the identification of the important role of MUC16 in CRC. Further, our analysis helped in the identification of higher expression of MUC16 in a subset of MSI-H patients and a very high overall mutation in MUC16 which established the imperative role of MUC16 in CRC.

3.2 BACKGROUND AND RATIONALE

Colorectal cancer (CRC) has remained one of the deadliest cancers in the United States for the past few years, with over a hundred thousand new cases expected in 2021 alone (Cancer Statistics, 2021). Furthermore, the 5-year survival in patients with late-stage CRC stands at a meager 14%. Amongst major reasons for this poor survival are the gaps in our understanding of the initiation and

progression of this disease. What makes it even harder is the inherent heterogeneity of this disease and the subsequent varied response to therapy. The majority (~70%-80%) of the CRC patient follow a convention pathway of initiation from polyp-adenoma to carcinoma (*68, 69*). This pathway is characterized by mutations in common tumor-associated genes (KRAS, TP53) and a very high percentage mutation in the adenomatous polyposis coli (APC) gene. Furthermore, a smaller subset of CRC patients develops tumors through what is known as the serrated pathway considering the development from hyperplastic polyps to serrated polyps to carcinoma which is characterized by a mutation in BRAF and a CpG island methylator phenotype (*70, 71*). Additionally, tumors developing from these two pathways also show differences in propensity to show microsatellite instability (MSI-H) wherein tumors emerging from the serrated pathway are more likely to be MSI-H as opposed to those from the traditional pathway. These major differences warrant more in-depth studies into the understanding of molecular dissimilarities in patients with CRC.

Colonic mucus which is made up predominately of highly glycosylated mucin proteins plays a central role in normal colonic function. However, mutations in conjunction with various environmental stimuli lead to changes in the composition of the mucus which eventually play a significant role in CRC development (*72*). Considering this, an integral part of CRC initiation and progression is the aberrant glycosylation and expression of various members of the 22-member mucin family. Over the past many years, various studies have elucidated the role of specific mucins in CRC (*73-75*). More specifically, studies have established a differential

expression of MUC2, secretory mucin inherently expressed in normal colon which is found to be downregulation or lost in CRC. Subsequently, this loss leads to faster tumor progression, resistance to apoptosis, etc. Further, mice lacking in Muc2 show increased proliferation and migration of cells, and a significantly lower level of apoptosis, and further a higher propensity to develop CRC (*76, 77*). Additionally, various studies have associated MUC2 silencing with CRC metastasis through various important signaling pathways like IL6 signaling (*78, 79*). In addition to MUC2, various studies have also elucidated the role of the MUC1, (*80-82*), MUC4 (*83, 84*), MUC5AC (*75, 85*), and MUC16 (*86, 87*). While these studies provide insight into the role of mucins in CRC initiation and progression, a comprehensive, patient cohort driven study has not been carried out to date.

Considering this, the current study uses a bioinformatics-driven approach harnessing the Cancer Genome Atlas (TCGA) and microarray datasets to carry out an in-depth analysis of mucin alteration in early precursor lesions and tumor samples. To investigate how mucins are expressed in the various precursor lesions for CRC, we first assessed microarray data from tubular adenomas (TA), sessile serrated adenomas/polyps (SSA/Ps), and normal samples to delineate expressional alteration across mucin family. Expression differences in mucins in CRC tumor samples were evaluated using the expression data from the adenocarcinoma dataset (COAD) from The Cancer Genome Atlas (TCGA). Additionally, to further study the relevance of mucins in CRC, we explored the TCGA data to study the survival differences in CRC patients with respect to the expression of various mucins. A deeper understanding of the role of mucins was

investigated using the web-tool from CBioPortal (https://www.cbioportal.org/) by assessing the mutational profile of mucins in the TCGA-COAD dataset. To fully understand the role of MUC16-which was found to be upregulated and highly mutated in a subset of patients, we combined the clinical information from within the TCGA dataset to the expression profile to identify specific subsets of patients with expression of specific mucins. Overall, we carried out an in-depth *in-silico* analysis to study and establish the role of mucins in CRC.

3.3 MATERIALS AND METHODS

3.3.1 Microarray data processing and analysis.

Raw microarray data files (.CEL) files were downloaded for both the GEO datasets-GSE43841 and GSE45720 and processed using RMA normalization using the “affy” package in R-Bioconductor. Furthermore, all mucins were extracted using gene symbols and boxplots plotted using MedCalc.

3.3.2 TCGA RNA-seq data processing.

Processed RNA-seq data (FPKM) was downloaded from the UCSC-Xena webserver (https://xena.ucsc.edu/). Clinical information was also downloaded from UCSC-Xena and matched in a sample-specific way using R-Bioconductor.

3.3.3 Survival analysis of mucins in CRC patient cohort.

Survival data were extracted from the clinical files from UCSC-Xena and matched to the expression data in a sample-specific manner using R-Bioconductor. Survival analysis was then carried out in JMP (v.14) defining median as the differentiating variable.

3.3.4 Mutational analysis of CRC in TCGA-COAD patient cohort.

Mutational analysis of all mucins was carried out using the web tool cBioPortal (https://www.cbioportal.org/). The point mutation data for MUC16 mutations was also downloaded using the webtool.

3.3.5 Domain mapping for MUC16

MUC16 sequence (ENST00000397910.8) was downloaded from Ensembl and domain analysis was using various online tools and servers including NCBI-CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), Pfam (http://pfam.xfam.org/) and prosite (https://prosite.expasy.org/). The dissection of motifs in the sequence was done using prosite and HMMER (http://hmmer.org/) tools. The regions for each motif and domain are represented in the figure.

3.4 RESULTS

3.4.1 Study cohort demographics.

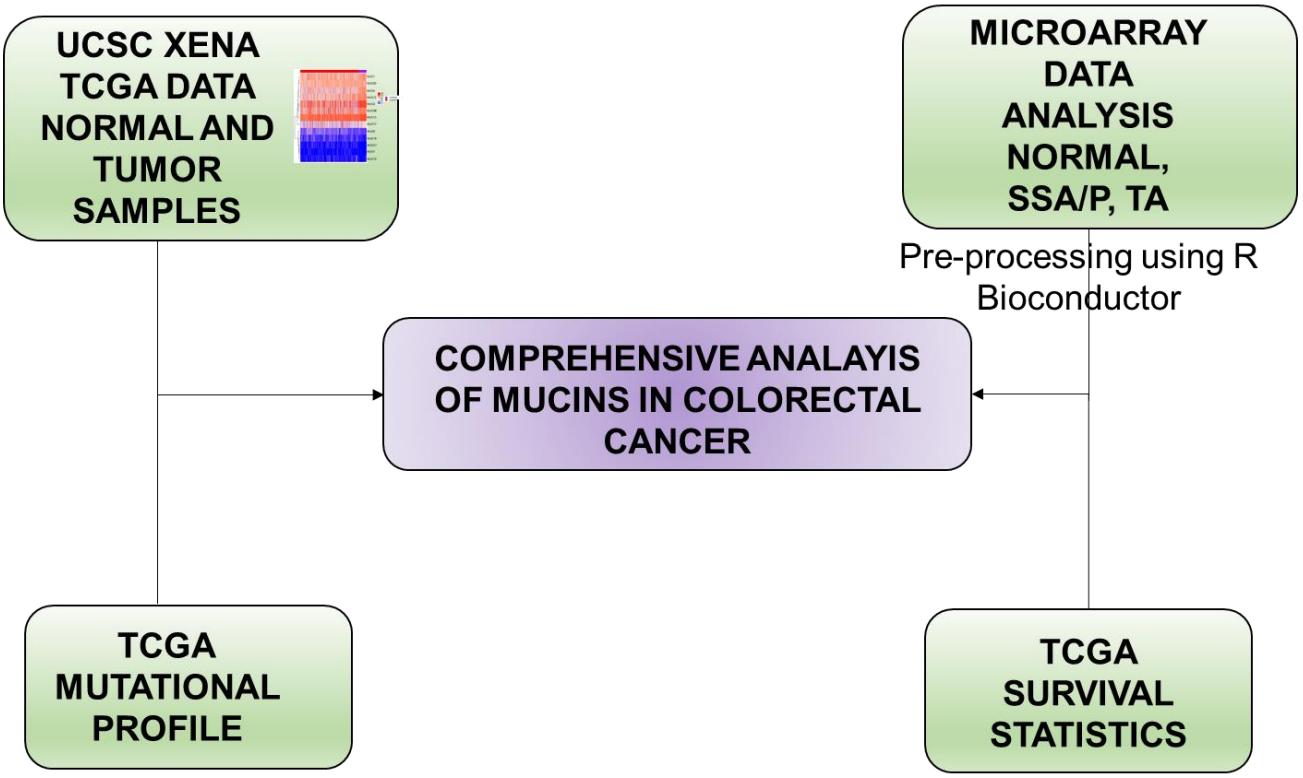
To study the expression differences in early precursor lesions from CRC, GEO datasets (https://www.ncbi.nlm.nih.gov/gds) was evaluated for datasets containing data from tubular adenomas (TAs), hyperplastic polyps (HPs), and sessile serrated adenomas/polyps (SSA/Ps) which led to the identification of two datasets- GSE43841 and GSE45720 (**Fig. 3.1B**). The Cancer Genome Atlas (TCGA) repository was studied for analyzing the alteration in the expression, mutation, and survival data from the colorectal adenocarcinoma (COAD) dataset. A total of 380 tumor samples were studied with representation from various subtypes (**Fig. 3.2A**, different microsatellite stability status (**Fig. 3.2B**), across race (**Fig. 3.2C**), and gender (**Fig. 3.2D**).

Figure 3.1 Overall study design.

Microarray data from GEO datasets was used in conjunction with the TCGA RNA-seq data to carry out comprehensive expression, mutation, and survival analyses in normal, early precursor lesion and tumor samples from colorectal cancer (CRC). **A.** Schematic representation of the overall methodology for assessing the expression, mutation, and survival status of mucins in CRC. TCGA RNA-Seq expression data from 380 tumors and 51 normal samples was downloaded using UCSC Xena and the expression of mucins was compared across tumor and normal samples. Survival information was also downloaded from UCSC Xena. TCGA mutation profile was assessed using the online tool; CBioPortal. **B.** Schematic representation of early precursor lesion datasets. Early precursor lesions data was assessed using microarray datasets from GEO datasets. Two separate microarray datasets (GSE43841 and GSE45720) were used to compare normal samples with tubular adenomas (TAs) and sessile serrated adenomas/polyps (SSA/Ps).

Figure 3.1

A.


UCSC XENA TCGA DATA NORMAL AND TUMOR SAMPLES
MICROARRAY DATA ANALYSIS NORMAL, SSA/P, TA
Pre-processing using R Bioconductor
COMPREHENSIVE ANALAYIS OF MUCINS IN COLORECTAL CANCER
TCGA MUTATIONAL PROFILE
TCGA SURVIVAL STATISTICS

B.

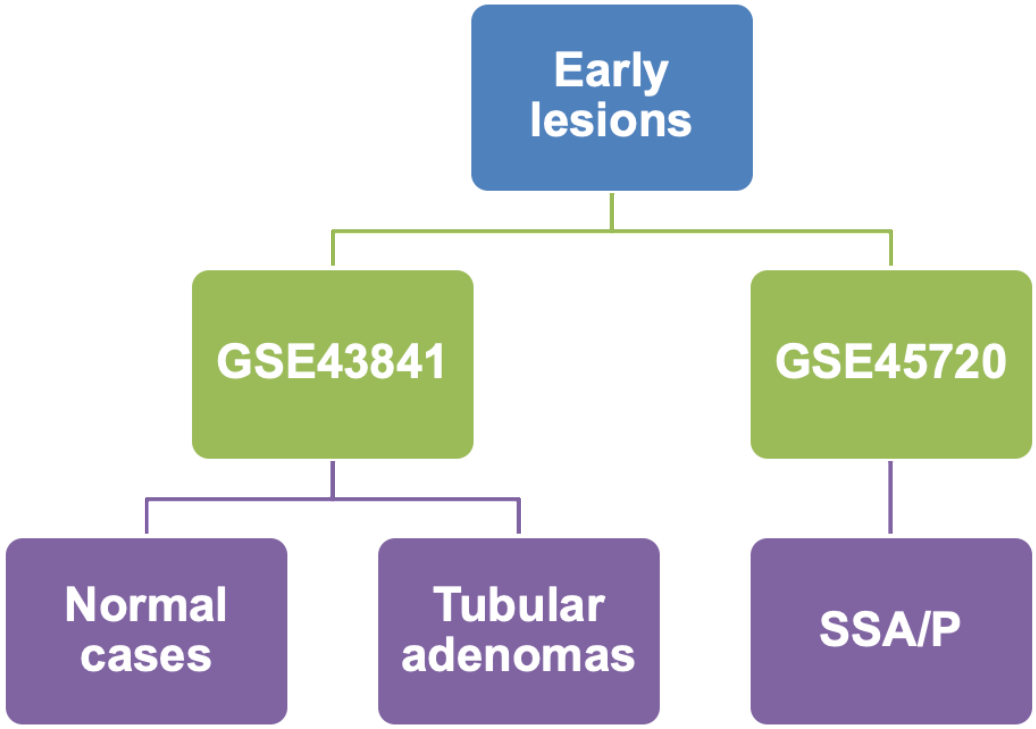
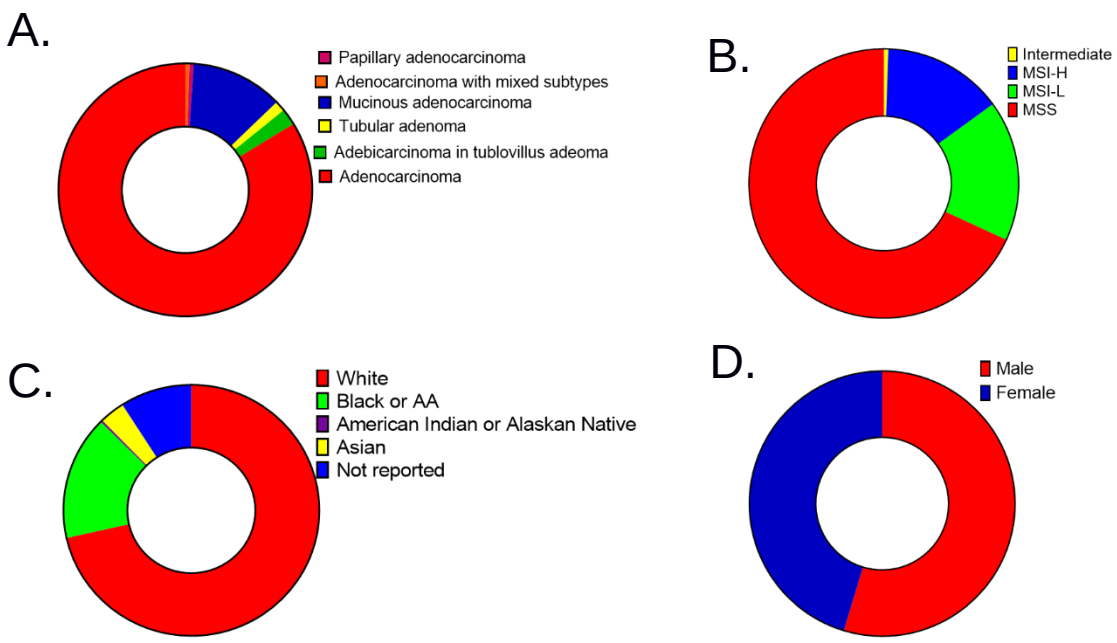

Early lesions
GSE43841
GSE45720
Normal cases
Tubular adenomas
SSA/P

Figure 3.2 Overall study population

A. Pie chart representation of various subtypes of CRC represented within the TCGA-RNA seq data. **B.** Pie chart representation of the microsatellite instability status of the TCGA samples represented as high microsatellite instability (MSI-H), low microsatellite instability (MSI-L), and microsatellite stable (MSS). The same codes have been used all throughout the study. **C.** Pie chart representation of race distribution across the TCGA RNA seq samples used for the current study. **D.** Pie chart representation of gender distribution across the TCGA RNA seq samples used for the current study.

Figure 3.2


A.
Papillary adenocarcinoma
Adenocarcinoma with mixed subtypes
Mucinous adenocarcinoma
Tubular adenoma
Adebicarcinoma in tublovillus adeoma
Adenocarcinoma
B.
Intermediate
MSI-H
MSI-L
MSS
C.
White
Black or AA
American Indian or Alaskan Native
Asian
Not reported
D.
Male
Female

3.4.2 Lesion-specific analysis of mucin expression reveals aberrant expression of various mucins across specific lesion types.

Comparison of normalized expression data from normal, TA, and SSA/P samples highlighted a statistically significant difference in expression of various mucins. Specifically, MUC1, MUC4, MUC6, MUC15, MUC16, and MUC19 were found to be upregulated in SSA/Ps when compared to normal samples and further in TAs when compared to SSA/Ps (**Fig.3.3A**). Further, interestingly MUC5AC, MUC13, and MUC17 were found to be significantly higher specifically in the SSA/P sample set when compared to both normal and the TAs samples. (**Fig. 3.3B**)

3.4.3 Expression analysis of TCGA CRC tumor samples reveals aberrant expression of mucins.

To assess the role of mucins in CRC tumor samples, normal and tumor data from the TCGA COAD cohort was assessed for expression alterations of mucins in CRC. This analysis led us to identify the upregulation of specific mucins like MUC6 and MUC15 (**Fig. 3.4A and Fig. 3.4B**). Furthermore, MUC2 and MUC4 were found to be significantly downregulated (**Fig 3.4C**). Interestingly, a specific subset of patients showed an upregulation of MUC16 (**Fig. 3.4D**).

3.4.5 MUC16 found to be upregulated in MSI-H patients.

Considering the specific elevation of MUC16 in a specific subset of patients, we correlated each of these MUC16 high (~34% of the total population, **Fig. 3.4E**) samples to the corresponding clinical variables. Comparison of MUC16 expression with respect to MSI status led to the identification that a higher percentage of MSI-H patients having an elevated expression of MUC16. Furthermore, no such

differences were observed in the MSS and MSI-L subpopulations (**Fig. 3.4F**). An independent immuno-histochemistry based validation showed a high-level MUC16 expression in tissue samples from human CRC patients (**Fig. 3.4G**). Further, gene set enrichment analysis of MUC16-HIGH and MUC16-LOW patient population showed important cancer-related pathways like KRAS-signaling, JAK-STAT signaling, and epithelial-mesenchymal transition highly enriched in the MUC16-HIGH population (**Fig. 3.4H**).

Figure 3.3 Differential expression analysis of mucins across early precursor lesions of CRC.

Microarray data from two microarray datasets (GSE43841 and GSE45720) was downloaded and processed using R-Bioconductor. The figure is a boxplot representation of mucins compared across normal, sessile serrated adenomas/polyps (SSA/Ps), and tubular adenomas (TAs). Statistical significance was assessed by carrying out a pairwise assessment comparing normal samples with sessile serrated adenomas and tubular adenomas.

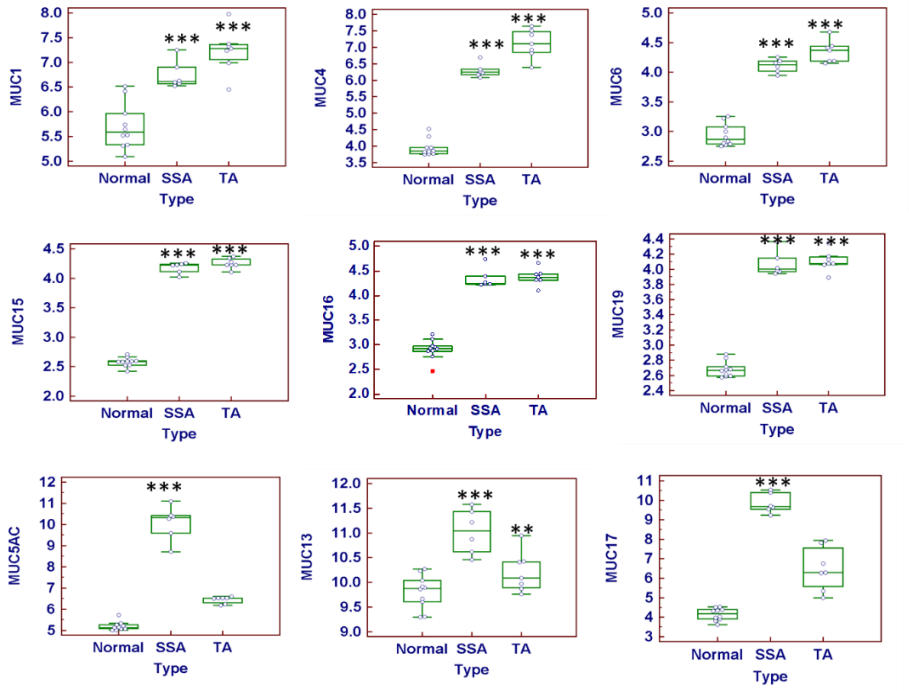
*** p.value < 0.01

** p. value < 0.05

Figure 3.3

A.

B.



*** p.value < 0.01
** p. value < 0.05

Figure 3.4 Differential gene expression of 22-member mucin family in tumor and normal samples.

A. Heatmap representation of expression differences between tumor (red) and normal (purple) samples. Overall assessment of expression differences in between normal and tumor samples showed overexpression of various mucins like MUC1, MU5AC, MUC15, and MUC16 with a loss of others like MUC2 and MUC4. **B**. Representative boxplots showing overexpression of mucins MUC6 and MUC15. **C.** Representative boxplots showing downregulation of MUC2 and MUC4.

Figure 3.4

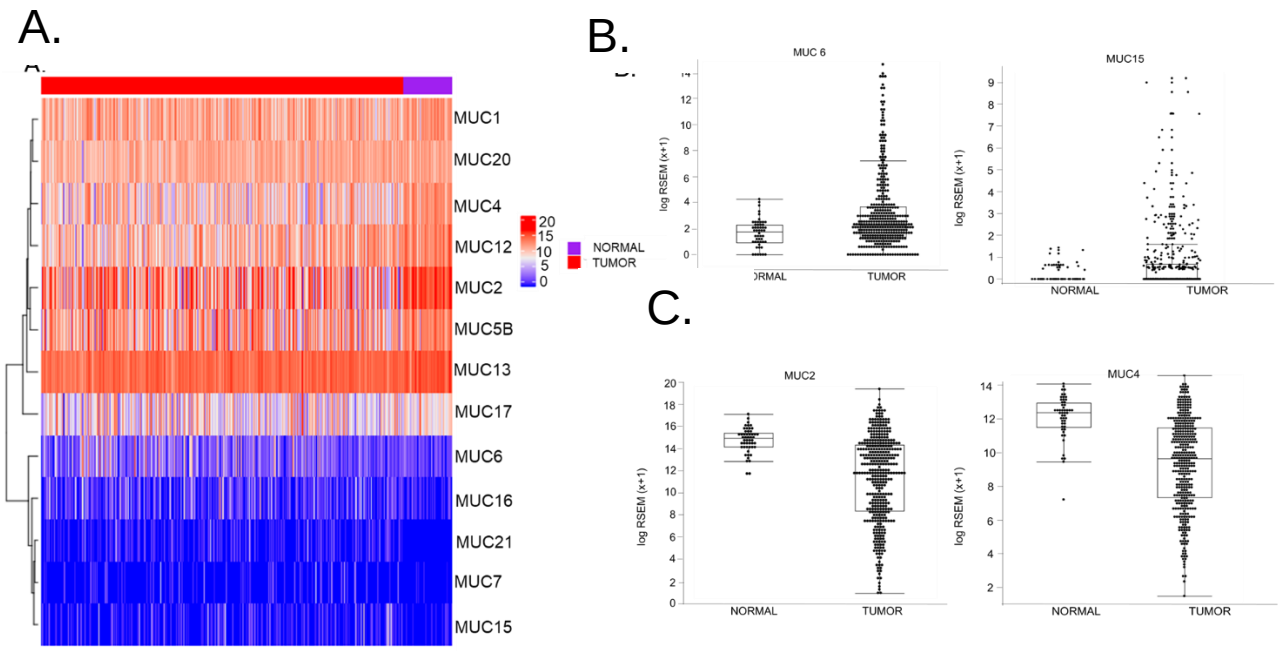
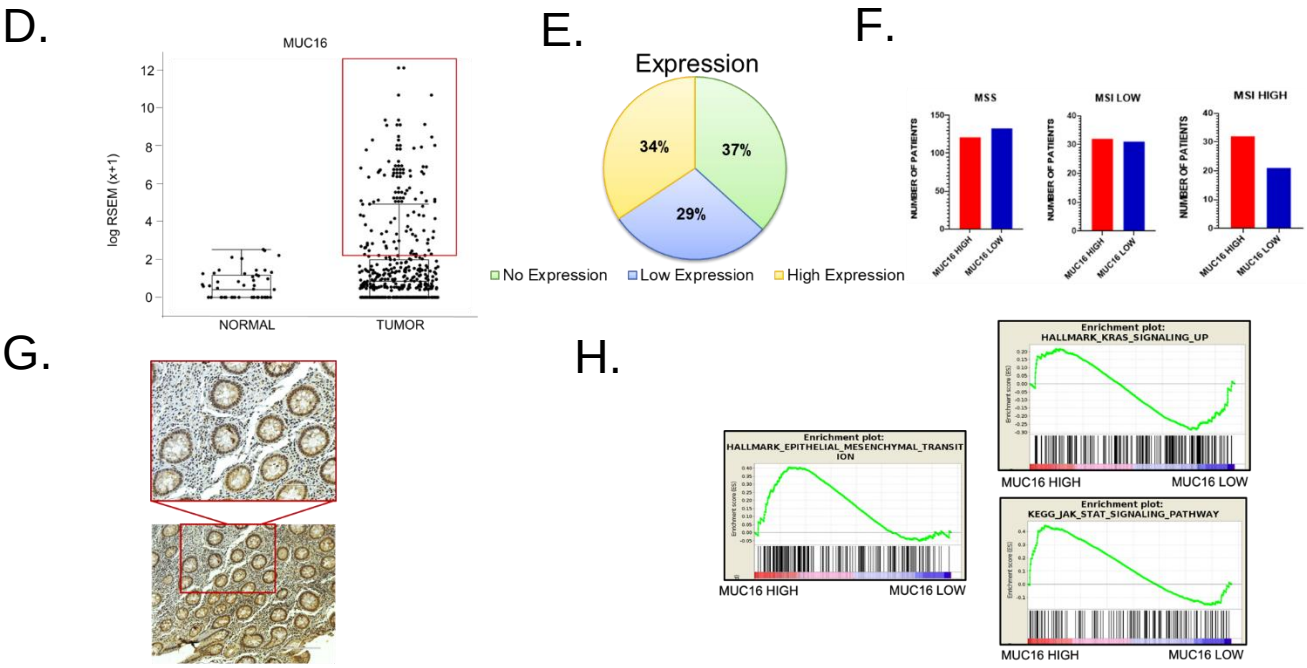

A.
MUC1
MUC20
MUC4
MUC12
MUC2
MUC5B
MUC13
MUC17
MUC6
MUC16
MUC21
MUC7
MUC15
20
15
10
5
0
NORMAL
TUMOR
B.
MUC 6
log RSEM (x+1)
TUMOR
MUC15
log RSEM (x+1)
NORMAL
TUMOR
C.
MUC2
log RSEM (x+1)
NORMAL
TUMOR
MUC4
log RSEM (x+1)
NORMAL
TUMOR

Figure 3.4 Differential gene expression of 22-member mucin family in tumor and normal samples.

D. A subset of patients were found to have an upregulation of MUC16 (marked by the red box). **E.** Pie chart representation of the distribution of MUC16 expression shows 34% of total samples show an extremely high expression of MUC16. **F.** Closer assessment of high MUC16 expressing patients showed an association of MSI status with MUC16 expression with a high percentage of MSI-H patients showing an overexpression of MUC16. **G.** Representative figure of IHC expression studies of MUC16 expression in CRC patient tissue samples. Independent IHC validation of MUC16 expression in CRC samples showed high MUC16 expression in a subset of all tested samples, hence validating the initial finding. **H.** Gene set enrichment analysis graphs comparison MUC16 high and MUC16 low samples.

Figure 3.4


D.
MUC16
log RSEM (x+1)
NORMAL
TUMOR
E.
Expression
34%
37%
29%
No Expression
Low Expression
High Expression
F.
MSS
MSI LOW
MSI HIGH
NUMBER OF PATIENTS
MUC16 HIGH
MUC16 LOW
G.
H.
Enrichment plot:
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
Enrichment plot:
HALLMARK_KRAS_SIGNALING_UP
Enrichment plot:
KEGG_JAK_STAT_SIGNALING_PATHWAY

3.4.5 Mucin expression differences found to be associated with survival.

Further, to study the prognostic relevance of mucins in CRC, survival statistic from TCGA was studied for each of the mucins after dividing the whole patient cohort into high and low expressing patient cohorts based on the median expression of each of the mucin. Interestingly, various mucins were found to affect the survival of CRC patients. Most interestingly, MUC6 was found to be statistically significantly correlated with survival wherein patients showing high expression were found to have a worse prognosis than patients showing low expression (**Fig. 3.5**).

3.4.6 High percentage mutations observed in various mucins.

Further, in the quest to study the role of mucins in CRC, we analyzed the percentage mutation of all mucins in the TCGA-COAD cohort. Interestingly, many mucins like MUC16, MUC5B, and MUC17 were found to be highly mutated. Moreover, MUC16 was found to be mutated in 28% of the TCGA-COAD cohort which when taken in conjunction with the overexpression of MUC16 in a subset of the population was an extremely interesting observation (**Fig. 3.6A**). Of note, various previous studies across different types have correlated MUC16 mutations with overall survival, tumor mutation load, and immune response but such a comprehensive study is missing for CRC. This finding was also independently validated in other CRC datasets, which consistently showed a high percentage mutation in CRC (**Fig. 3.6B**). Next, we evaluated the correlations between MUC16 mutations and the other commonly mutated genes in CRC (TP53, APC, KRAS, PIK3CA, SMAD4, BRAF, etc.). Interestingly, MUC16 mutations were found to be

highly correlated to these various other types of mutations with a tendency of co-occurrence amongst these mutations (**Fig. 3.6C**).

3.4.7 MUC16 mutations.

Considering this, we further carried out an in-depth analysis of mutations of MUC16 in CRC. We first assessed the type of mutations present in MUC16 in terms of variant classification, single nucleotide variation, and variant type. Interestingly, we observed that most of the mutations in MUC16 were missense mutations (**Fig. 3.6D**), with predominately T/G-G/T and T/C-C/T single nucleotide variations (**Fig. 3.6E**), and the majority of the patients showing a single nucleotide polymorphism (**Fig. 3.6F**). Furthermore, we overlapped independently mapped domains in MUC16 and the location of mutations and found that most MUC16 mutations were found in the SEA domain of MUC16 (**Fig. 3.6G**). Interestingly, various studies have elucidated the role of the SEA domain present within mucins in therapeutic interventions, and the overall progression of tumors.

Figure 3.5 Survival analysis of CRC patients

Survival analysis was carried out using SAS-JMP (v14) using the median expression of each mucin as the differentiating variable.

Figure 3.5

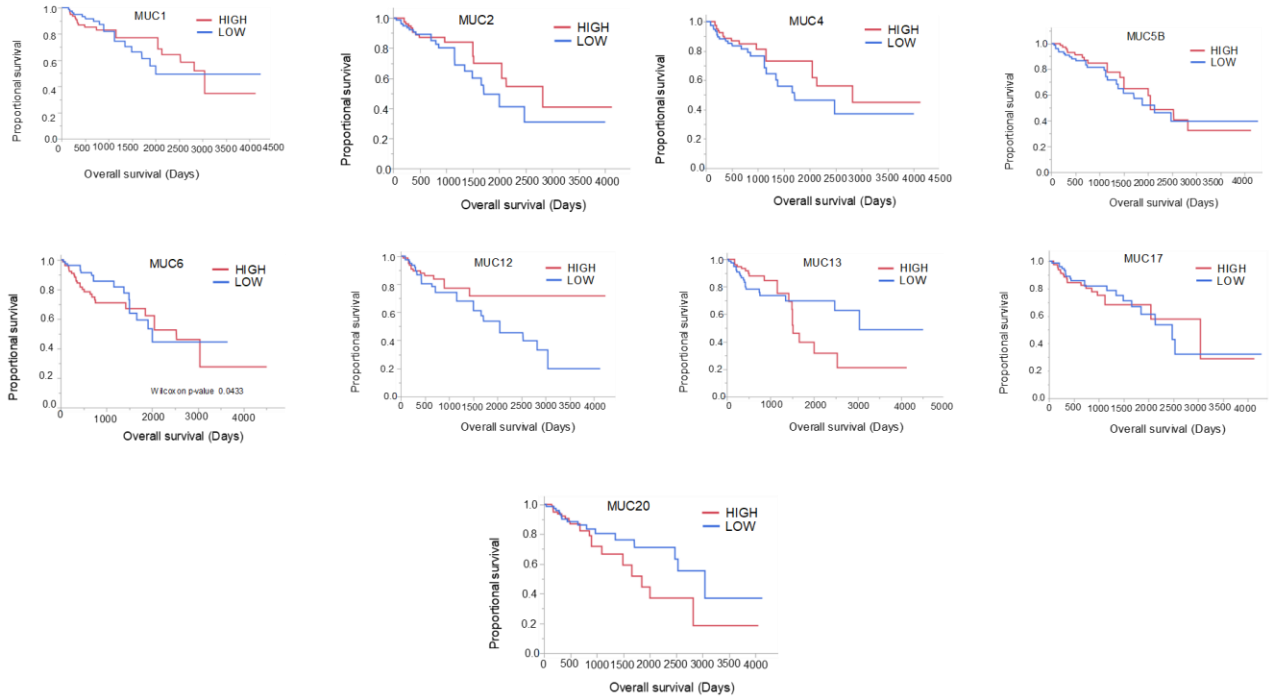

MUC1
HIGH
LOW
Proportional survival
Overall survival (Days)
MUC2
MUC4
MUC5B
MUC6
Wilcoxon p-value 0.0433
MUC12
MUC13
MUC17
MUC20

Figure 3.6 Mutational analysis of mucins in TCGA CRC samples identifies high mutation percentage of MUC16 mutated patients

A. CBioPortal web portal was used to assess the mutations of all members of the mucin family in TCGA CRC patients. Interestingly, MUC16 was mutated in 28% of the CRC patients.

Figure 3.6

A.

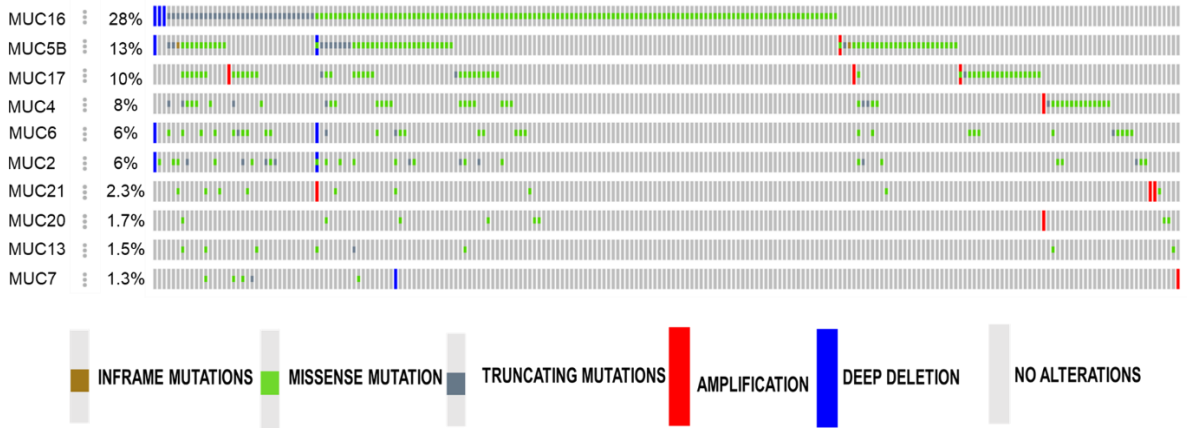
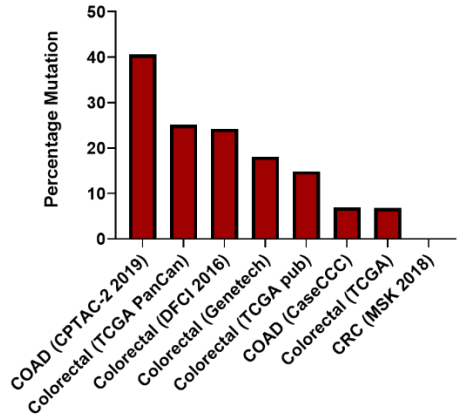

MUC16 28%
MUC5B 13%
MUC17 10%
MUC4 8%
MUC6 6%
MUC2 6%
MUC21 2.3%
MUC20 1.7%
MUC13 1.5%
MUC7 1.3%
INFRAME MUTATIONS
MISSENSE MUTATION
TRUNCATING MUTATIONS
AMPLIFICATION
DEEP DELETION
NO ALTERATIONS

Figure 3.6. B. Assessment of MUC16 mutations in other datasets. **C**. Comparison of MUC16 mutations to other commonly mutated genes in CRC showed high tendency of co-occurrence of MUC16 and other genes like BRAF, MSH6, PIK3CA amongst others

Figure 3.6

B.


Percentage Mutation
50
40
30
20
10
0
COAD (CPTAC-2 2019)
Colorectal (TCGA PanCan)
Colorectal (DFCI 2016)
Colorectal (Genetech)
Colorectal (TCGA pub)
COAD (CaseCCC)
Colorectal (TCGA)
CRC (MSK 2018)

C.

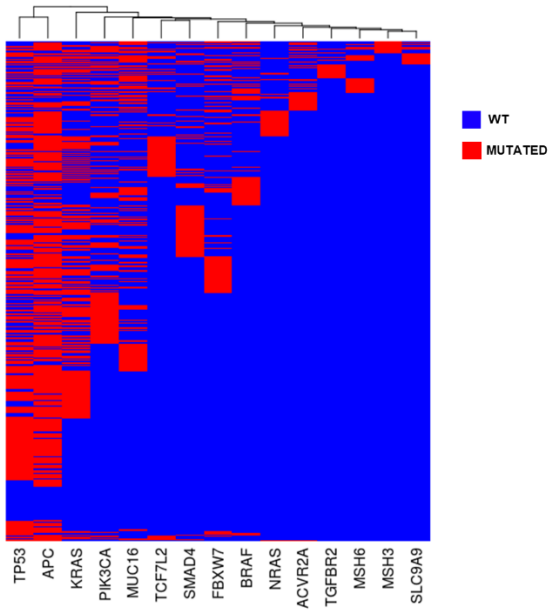
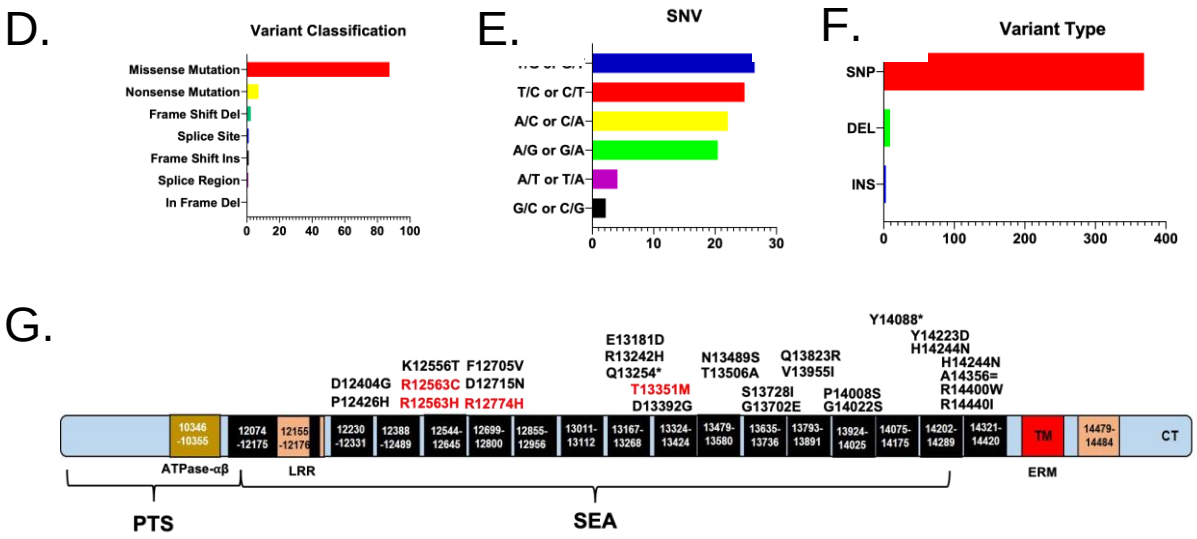

WT
MUTATED
TP53
APC
KRAS
PIK3CA
MUC16
TCF7L2
SMAD4
FBXW7
BRAF
NRAS
ACVR2A
TGFBR2
MSH6
MSH3
SLC9A9

Figure 3.6. D. Closer assessment of MUC16 mutations assessing variant classification. **E.** Assessment of MUC16 mutations for single nucleotide variation classes. **F.** Assessment of MUC16 mutations for variant types. **G.** Further, we assessed the specific mutations across MUC16 domains and compared mutations across cancers. Interestingly, various mutations were found to be common to other cancers (red). Most of these mutations were found within the SEA domain of MUC16.

Figure 3.6


D.
Variant Classification
Missense Mutation
Nonsense Mutation
Frame Shift Del
Splice Site
Frame Shift Ins
Splice Region
In Frame Del
0 20 40 60 80 100
E.
SNV
T/C or C/T
A/C or C/A
A/G or G/A
A/T or T/A
G/C or C/G
0 10 20 30
F.
Variant Type
SNP
DEL
INS
0 100 200 300 400
G.
D12404G
P12426H
K12556T
R12563C
R12563H
F12705V
D12715N
R12774H
E13181D
R13242H
Q13254*
T13351M
D13392G
N13489S
T13506A
Q13823R
V13955I
S13728I
G13702E
P14008S
G14022S
Y14088*
Y14223D
H14244N
H14244N
A14356=
R14400W
R14440I
10346 -10355
12074 -12175
12155 -12176
12230 -12331
12388 -12489
12544- 12645
12699- 12800
12855- 12956
13011- 13112
13167- 13268
13324- 13424
13479- 13680
13635- 13736
13793- 13891
13924- 14025
14075- 14175
14202- 14289
14321- 14420
TM
14479- 14484
CT
ATPase-αβ
LRR
ERM
PTS
SEA

3.5 DISCUSSION

While various studies have elucidated the role of specific mucins in the initiation and progression of CRC, a detailed analysis of all members of the mucin family has not been carried out to date. Considering this, the current study uses various bioinformatics tools to assess the role of mucins in CRC. First, mucin expression patterns were studied in microarray data from early precursor lesions of CRC. Further, to gain insight into differences between tumor and normal samples, data from the TCGA-COAD cohort were studied. Furthermore, mutation, survival, and clinical information from the TCGA cohort were used to assess the role of mucins in CRC.

The analysis of the microarray sample set from precursor lesions of CRC led to the identification of MUC5AC, MUC13, and MUC17 to be specifically overexpressed in SSA/Ps. Interestingly, multiple previous studies have explored the biomarker potential of MUC5AC and MUC17 in SSA/P (*88-90*). MUC13, however, remains unexplored. Considering this, further studies combining MUC5AC-MUC13-MUC17 can prove to be beneficial in CRC early detection. Furthermore, analysis of tumor samples from the TCGA-COAD patient cohort showed significant downregulation of MUC2 and MUC4 in the tumor samples in comparison to the normal samples. Multiple studies have established the functional significance of the loss of MUC2 and MUC4 in CRC. We observed an upregulation in the expression of MUC6 and MUC15. Upregulation observed in the MUC16 was restricted to a subset of patients which were predominately MSI-H. Interestingly, previous studies have correlated MUC2 and MUC5AC expression

with MSI status suggesting that the expression of these mucins can be a predictor of MSI status (*91*) but no such study has been carried out in MUC16. Considering this, MUC16 expression can have direct clinical implications which should be further explored through more studies. Additionally, MUC16 was also found to be highly mutated in the TCGA-COAD cohort further supporting the important role of MUC16 in CRC. Interestingly, studies across cancer types have elucidated the importance of MUC16 mutations in cancer progression. MUC16 mutations were found to be correlated with prognosis in endometrial cancers (*92*), gastric cancer (*93*) cutaneous melanoma (*94*), and with response to immune checkpoint inhibitors overall in solid tumors (*95*). These observations and the high percentage of mutations seen in MUC16 in our patient cohort suggest a potential clinical relevance of MUC16 and warrant further studies regarding these mutations.

Chapter 4: Computational Analysis of Mucins in Gastric Cancer Identifies Prognostically Relevant Clusters

4.1 SYNOPSIS

Gastric adenocarcinoma is an immensely deadly disease with a 5-year survival rate of 6% in advanced tumors and over 1 million annual cases worldwide. The mucin family of glycoproteins has been established to be of great importance in the initiation, progression, and metastasis of various gastrointestinal (GI) malignancies. This role of mucins is in turn responsible for significant applications as biomarkers and therapeutic interventions. In GC, multiple studies have reported the importance of these mucins, specifically MUC1, MUC2, MUC5AC, and MUC13. However, to date, global studies for assessing the combined expression and functional implications of mucin family are lacking. In this regard, large-scale datasets analyses such as The Cancer Genome Atlas (TCGA), along with comprehensive information on the expression, survival, and mutational information, can provide a holistic picture of the mucin family in gastric cancer pathobiology. Considering this, the present study uses the stomach adenocarcinoma (STAD) TCGA. It assesses mucin expression patterns, mutational patterns, survival statistics and pathobiological significance of mucin-associated signature. Furthermore, the study also uses the large-scale normal dataset (genotype-tissue expression-GTEx) to measure the differences between the normal and tumor samples. Overall, our study reiterates the importance of MUC1, MUC5AC, MUC6, MUC13, and MUC16, and establishes the significance of MUC2 and MUC9. Further, it provides the MUC2-MUC12-MUC13-OVGP1/MUC9-MUC4-MUC20 as a prognostically relevant cluster of mucins in GC.

4.2 BACKGROUND AND RATIONALE

Over the years, gastric cancer (GC) has maintained its high mortality rate, wherein currently it is the third most common cause of cancer deaths globally (*8*). Additionally, what makes it even worse is the high number of cases, with over one million cases detected annually (*96*). One of the major reasons for this high mortality is a lack of understanding of the complexity and the heterogeneity within the disease. Researchers believe that a better understanding of the disease biology would lead to efficient biomarker discovery and better therapeutic interventions resulting in improved survival rates (*97*). In this regard, a better understanding of the role of the 22-member mucin family of glycoproteins might prove beneficial considering their importance in cancer initiation and progression. Specifically, in gastro-intestinal malignancies, mucins have been studied for therapeutic potential (*98-100*) as biomarkers (*73, 101*) and to play significant roles in tumor progression (*75, 102*). Further, MUC1 has been associated with *Helicobacter pylori* infection which is a known risk factor for GC (*103*). Additionally, a study assessing the prognostic relevance of MUC1, MUC2, and MUC5AC found downregulation of MUC1, aberrant expression of MUC5AC, and de-novo expression of MUC2 in GC patients (*104*). Additionally, various studies have found overexpression of MUC13 in various GC subtypes (*105, 106*). Furthermore, a recent study has correlated the mutational load of MUC16 with tumor mutation load and survival of GC patients (*107*). While all these studies report about one or a combination of a few mucins, in-depth knowledge of mucin expression and its correlation with GC is urgently needed. In this regard, the recent boom in

sequencing technology and the patient data that has become available since then offers unique insights into expression, mutation, and survival patterns. Hence, it can prove extremely useful in answering questions pertaining to understanding the complexity of various proteins and protein families across various malignancies.

Considering this, the present study uses open-source data and various web tools to carry out an in-depth analysis of mucins in gastric cancer. We first assessed the stomach adenocarcinoma (STAD) dataset from the cancer genome atlas (TCGA) for expression differences between tumor and normal samples. Further, to carry out a more specific analysis, we compared the STAD-TCGA dataset to the normal samples from the genotype-tissue expression (https://www.gtexportal.org/home/; GTEx, (*108*) portal using the web-based tool GEPIA (http://gepia.cancer-pku.cn/) (*109*)**.** Furthermore, an independent validation study was carried out using the immuno-histochemistry data from the pathology atlas within the human protein atlas (https://www.proteinatlas.org/, (*110*)). Further, to identify mucins closely associated with each other and forming functionally relevant clusters, we first carried out a protein network analysis using the STRING (https://string-db.org/;(*111*)**)** database. Additionally, for a GC-specific co-expression analysis, Spearman correlation values for all mucins within the TCGA-STAD dataset were assessed using GEPIA, which led to the identification of the highly correlated MUC2-MUC3A-MUC12-MUC13-MUC17-OVGP1/MUC9-MUC4-MUC20 cluster. Further, to understand this cluster's prognostic relevance, a cox-proportional survival analysis was carried out in the TCGA-STAD dataset using the webtool SurvExpress (*112*) using the maximization algorithm on risk groups. Interestingly,

GC's prognostic stratification based on the MUC2-MUC12-MUC13-OVGP1/MUC9-MUC4-MUC20 cluster was extremely significant with a log-rank p-value of 0.00017. Furthermore, to study these mucins more closely, we went on to assess the mutations present within these glycoproteins. Interestingly, many of the mucins showed a high percentage of mutations, including MUC16 (35%), MUC17 (17%), MUC5B (14%), MUC4 (13%), and MUC6 (13%), and further studies exploring these mucins in GC would be extremely interesting.

4.3 MATERIALS AND METHODS

4.3.1 Expression data extraction and processing. Expression data from the TCGA-STAD dataset was assessed in two ways. USSC-Xena was used to download the log normalized FPKM expression data which was then processed using "Complex Heatmaps" from R-Bioconductor. Furthermore, the webtool GEPIA (http://gepia.cancer-pku.cn/) was assessed for comparison analysis between TCGA-STAD tumor expression and normal organ expression from GTEx.

4.3.2 Correlation analysis. Spearman correlation values were calculated for each of the mucins with each other using the web-interface of the tool GEPIA (http://gepia.cancer-pku.cn/). Further, the "corplots" package within R-Bioconductor was used to plot the correlation plot using hierarchical clustering as the order variable.

4.3.3 Network analysis. Functional protein network analysis was carried out using the webtool string (https://string-db.org/). Mucins were clustered into distinct groups using the k-means algorithm.

4.3.4 Survival-analysis. The webtool SurvExpress (http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp) was analyzed for the correlation cluster (MUC2, MUC3A, MUC12, MUC13, OVGP1/MUC9, MUC4, and MUC20) and each of the genes within the cluster individually. “Survival days” was used as the censoring variable with maximization of the risk groups.

4.3.5 Mutational analysis. Mutational analysis was carried out using the webtool cBioPortal (https://www.cbioportal.org/). CBioPortal was assessed for all mucins in the TCGA-stomach adenocarcinoma pan-cancer atlas (N=440) cohort. Further, independent validation studies were carried out in other stomach adenocarcinoma datasets (Pfizer; N=100, TCGA-Firehose Legacy; N=478, Nature 2014; N=295, Nature Genetics 2014; N=30, and Nature Genetics 2011; N=22).

4.3.6. IHC analysis using protein atlas. Protein atlas was analyzed for immunohistochemistry (IHC) assessment of all mucins, and quantitative results from specific antibodies (MUC1-CAB000036, MUC2-CAB016275, MUC3A-HPA010871, MUC4-HPA005895, MUC5AC-CAB002774, MUC6-CAB002165, MUC7-HPA006411, MUC9/OVGP1-HPA062205, MUC12-HPA023835, MUC13-HPA045163, MUC14-HPA005928, MUC15-HPA026110, MUC16-CAB055172, MUC17-HPA031634, MUC18-CAB002147) analyzed for quantitative staining Representative figures from each of these specific mucins were downloaded.

4.4 RESULTS

4.4.1 Study population details and characteristics. The national cancer institute (NCI) repository, the Cancer Genome Atlas (TCGA), contains gene expression, mutation, and survival data from various malignancies. To explore the role of mucins in GC, the stomach adenocarcinoma (STAD) dataset from within TCGA was assessed for expression, survival, and mutation data. The key characteristic distribution, including gender (**Fig.4.1A**), race (**Fig. 4.1B**), ethnicity (**Fig. 4.1C**), vital status (**Fig. 4.1D**), stage (**Fig. 4.1E**), and treatment received (**Fig. 1F**), has been compiled shows a diverse patient population.

4.4.2 Aberrant expression of multiple mucins observed in GC. Expression differences between normal and tumor samples were explored in two ways. First, processed expression (fragments per kilobase per million; FPKM) data from GC tumor and adjacent normal samples were downloaded from the UCSC Xena webserver (http://xena.ucsc.edu/). A heatmap-based visualization (**Fig. 4.2A**) of the data helped in identifying expression differences in certain mucins wherein mucins like MUC13 and MUC1 were found to be upregulated in tumor cases when compared to normal and were also found be clustering together. While this gave some insight, to further understand the complexities of the expression differences, the tumor expression data (N=408) was compared to the normal samples (N=211) from the GTEx portal using the web-based tool GEPIA. Interestingly, a statistically significant (p-value < 0.05) increase was observed in MUC2, MUC3A, MUC4, MUC5B, MUC12, MUC13, MUC17, and a significant decrease was seen in the

case of MUC5AC and MUC6 when tumor samples were compared to normal samples (**Fig. 4.2B**). Stage-specific expression differences identified an increase in MUC13 expression with stages (not-significant) and a uniform distribution across other mucins (**Fig. 4.2C).**

Figure 4. 1 Demographics of the patient population.

The various demographics of the TCGA-STAD cohort were assessed to reveal the patient population's distribution in terms of **A.** gender, **B.** race, **C.** ethnicity, **D.** vital status, **E.** stage, and **F.** treatment type.

Figure 4.1

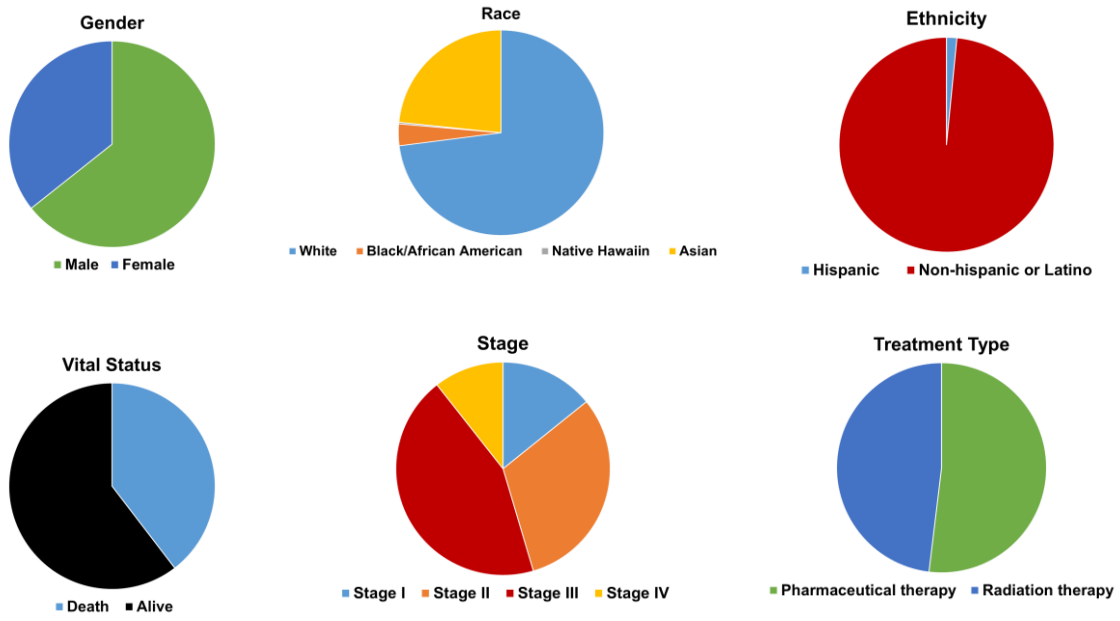

Gender
Male
Female
Race
White
Black/African American
Native Hawaiin
Asian
Ethnicity
Hispanic
Non-hispanic or Latino
Vital Status
Death
Alive
Stage
Stage I
Stage II
Stage III
Stage IV
Treatment Type
Pharmaceutical therapy
Radiation therapy

Figure 4. 2 Expression patterns for Mucins in TCGA-STAD.

The TCGA-STAD dataset was studied for expression patterns of all 22-members of the mucin family of glycoproteins. **A**. A heatmap representation of processed (FPKM) RNA-sequencing data downloaded from UCSC-XENA webserver. Mucins have been clustered using the hierarchical clustering within “Complex-Heatmaps”.

Figure 4.2

A.

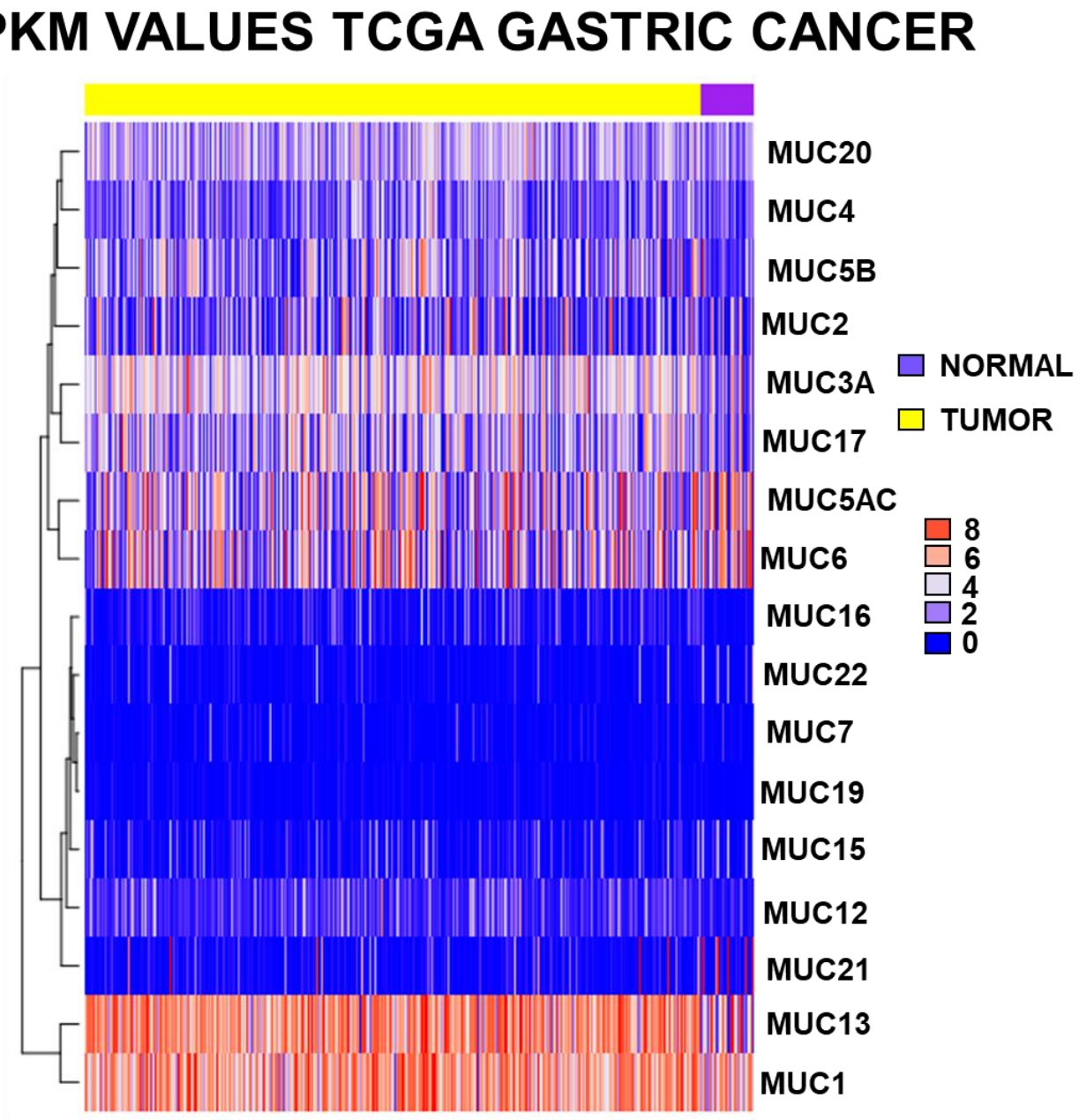

FPKM VALUES TCGA GASTRIC CANCER
MUC20
MUC4
MUC5B
MUC2
MUC3A
MUC17
MUC5AC
MUC6
MUC16
MUC22
MUC7
MUC19
MUC15
MUC12
MUC21
MUC13
MUC1
NORMAL
TUMOR
8
6
4
2
0

Figure 4.2. B. Boxplot representations of significantly differentially expressed mucins in between normal and tumor samples.

Figure 4.2

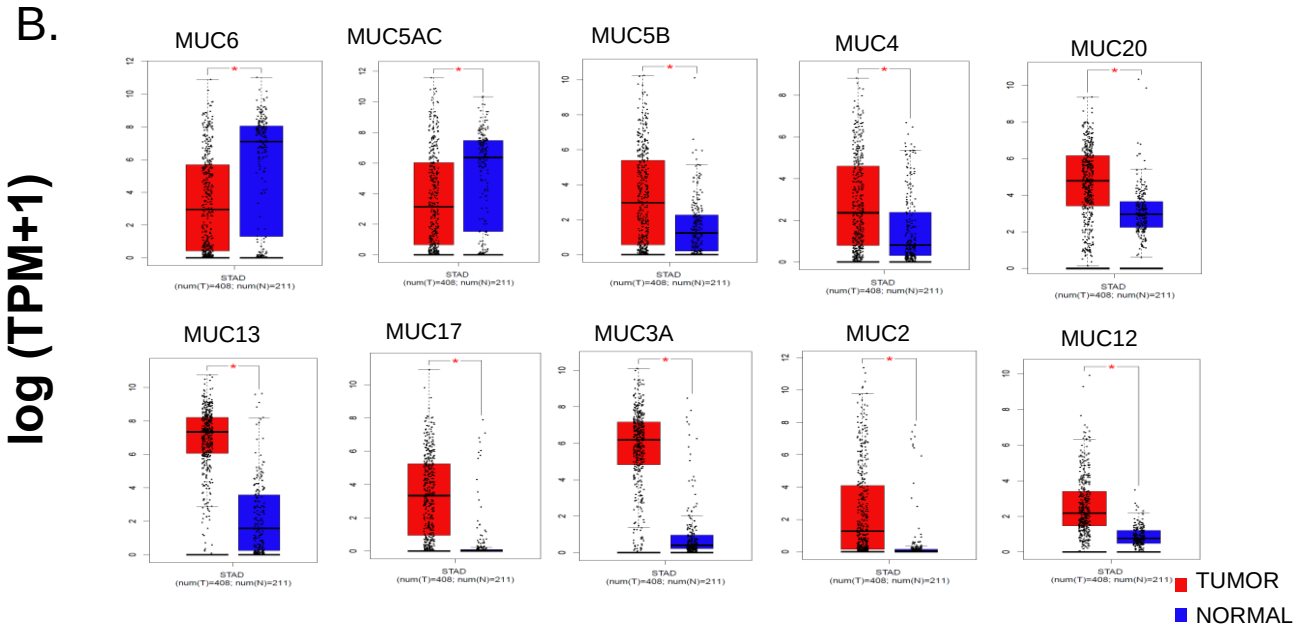

B.
log (TPM+1)
MUC6
MUC5AC
MUC5B
MUC4
MUC20
MUC13
MUC17
MUC3A
MUC2
MUC12
STAD
(num(T)=408; num(N)=211)
TUMOR
NORMAL

Figure 4.2. C. Boxplot representations of all other mucins in between normal and tumor samples.

Figure 4.2

C.

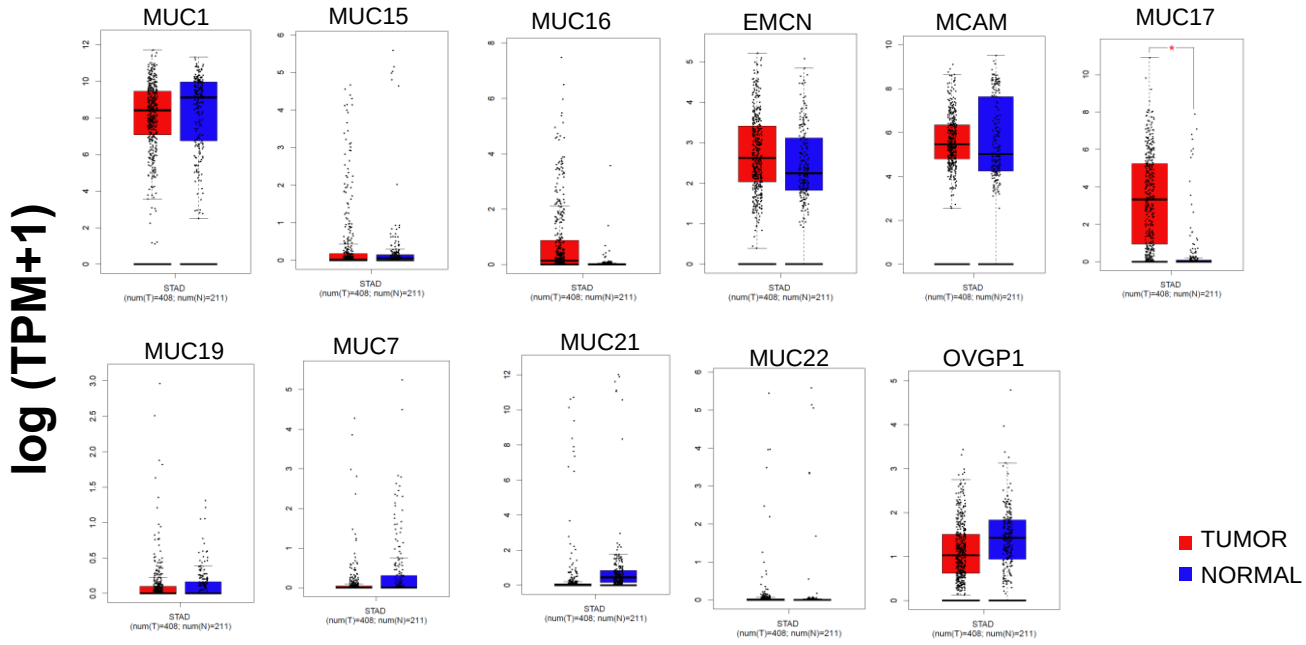
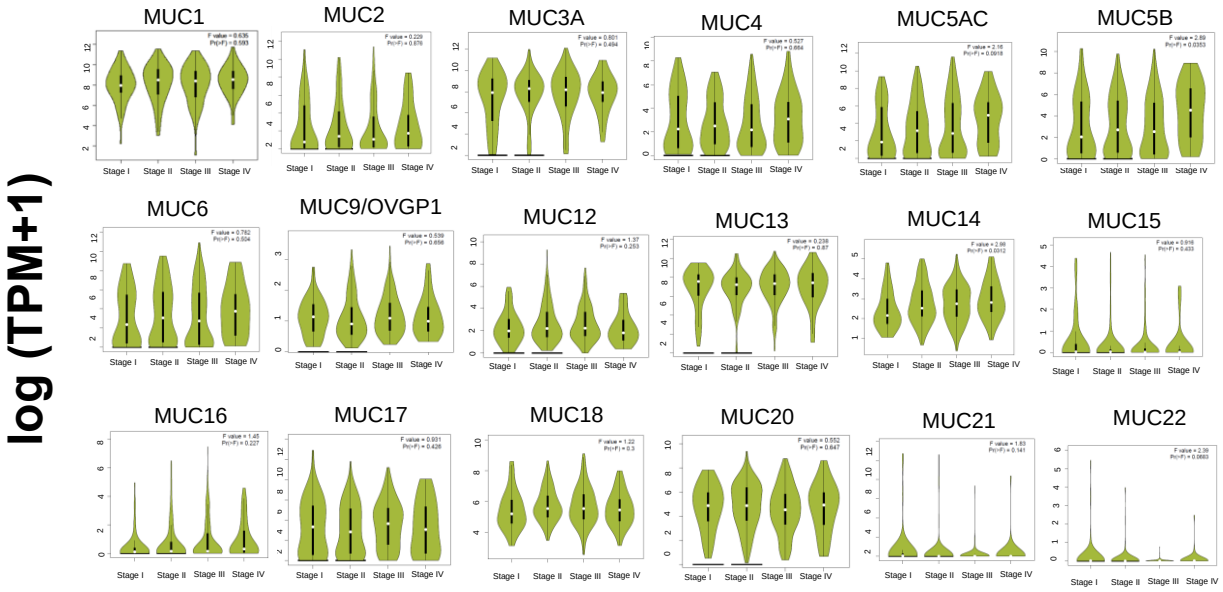

log (TPM+1)
MUC1
MUC15
MUC16
EMCN
MCAM
MUC17
MUC19
MUC7
MUC21
MUC22
OVGP1
STAD
(num(T)=408; num(N)=211)
TUMOR
NORMAL

Figure 4.3 Expression patterns for Mucins in TCGA-STAD across stage.

Boxplot representations of differentially expressed mucins across stages.

Figure 4.3


log (TPM+1)
MUC1
MUC2
MUC3A
MUC4
MUC5AC
MUC5B
MUC6
MUC9/OVGP1
MUC12
MUC13
MUC14
MUC15
MUC16
MUC17
MUC18
MUC20
MUC21
MUC22
Stage I
Stage II
Stage III
Stage IV

4.4.3 Independent validation of expression mucins. Further, to independently validate the expression of these mucins in GC, immunohistochemistry (IHC) images from the Human Protein Atlas were assessed for expression. Quantification (**Fig. 4.4A**) of expression in GC patient tissue samples revealed an extremely high expression of MUC1. Furthermore, MUC2 MUC3, MUC4, MUC5AC, MUC13, MUC14, MUC17, and MUC18 were found to be upregulated in a high number of cases suggesting a potential role of these mucins in GC, which needed to be explored further. Representative IHC images from high or medium expression cases were downloaded from the Human Protein Atlas **(Fig. 4.4B)**

4.4.4 Correlation analysis identifies specific clusters of mucins. Next, to study the relationship of mucins with each other, protein network analysis was first carried out to compare and identify closely associated mucins. Interestingly, k-means clustering of the network showed a close association between MUC3A-MUC12-MUC13-MUC15-MUC20-MUC21-MUC22 and also in between MUC4-MUC5AC-MUC5B-MUC6-MUC7-EMCN-MUC16 (**Fig 4.5A**)**.** Additionally, to assess the co-expression between mucins specifically in GC, a spearman correlation analysis was carried out using GEPIA and plotted using the “corrplot” library from R-Bioconductor with hierarchical clustering as the ordering parameter. Interestingly, MUC2-MUC3A-MUC4-MUC12-MUC17-OVGP1-MUC20 was found to be a highly correlated cluster. (**Fig. 4.5B**)

4.4.5 Survival analysis identifies prognostic biomarker signature. Next, to assess the prognostic relevance of these mucins’ survival analysis was carried out in the TCGA-STAD dataset using the risk maximization algorithm within

SurvExpress. Interestingly, from the aforementioned cluster, MUC4, MUC12, MUC13, and MUC20 were found to be independently prognostically relevant (**Fig. 4.6A**). Furthermore, interesting was the fact that the aforementioned correlation cluster of MUC2-MUC12-MUC13-MUC4-OVGP1-MUC20 was found to be highly prognostically relevant (Log Rank-p-value 0.0001778), establishing the prognostic relevance of these mucins in GC (**Fig. 4.6B**).

4.4.6 Mutational patterns identify a high percentage of mutations in various mucins. Next, to supplement the understanding of mucins in GC, a mutational analysis was carried in the TCGA-STAD dataset using the webtool CBioPortal. Intriguingly, MUC16 was found to be one of the most highly mutated genes in the TCGA-STAD cohort, with over 35% showing at least some mutation in MUC16 (**Fig. 4.7A**). Additionally, MUC17, MUC5B, MUC4, and MUC6 were also found to be highly mutated in the studied cohort. MUC16 was also found to be highly mutated in other GC datasets (**Fig. 4.7B**), further establishing the important role of MUC16 in GC.

Figure 4.4 Immunohistochemistry analysis of Mucins in GC.

The Human Protein Atlas was studied for immunohistochemistry expression of all mucins in GC. **A**. Bar graph representation of percentage high (red), medium (green), low (yellow), and no-expression (black) across all mucins. **B.** Representative IHC images of all mucins revealed high expression of MUC1.

Figure 4.4

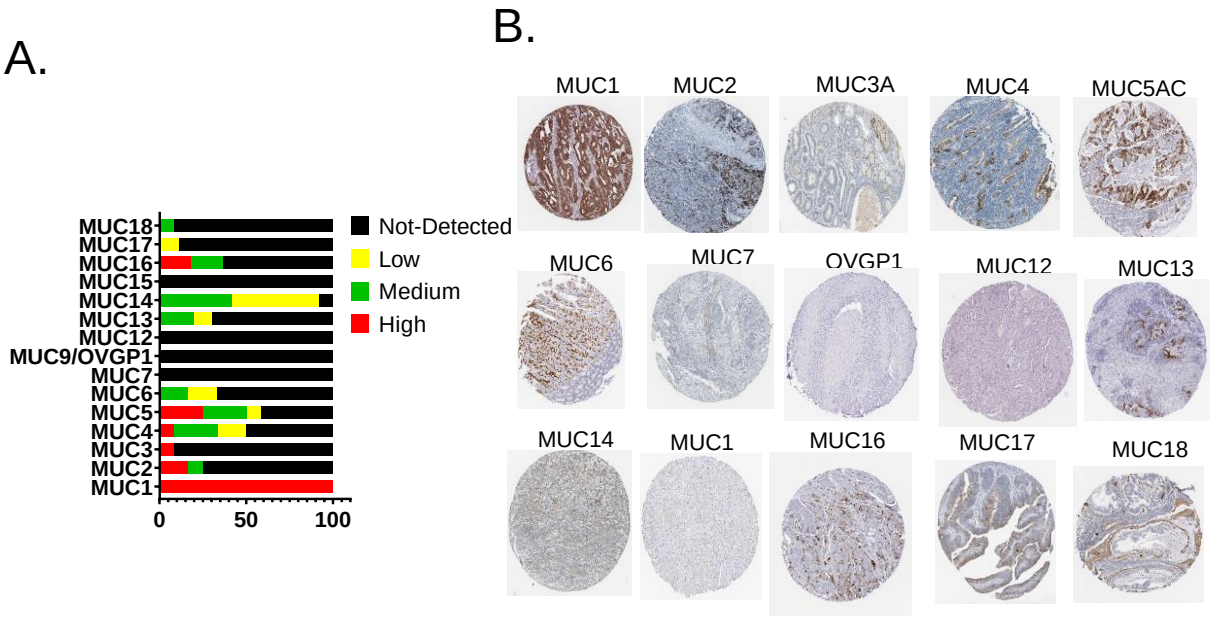
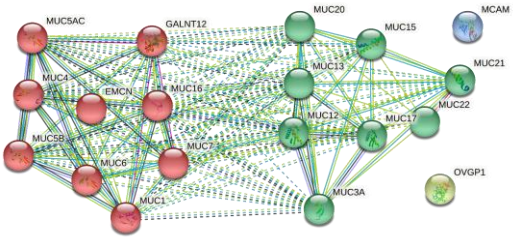

A.
B.
MUC18
MUC17
MUC16
MUC15
MUC14
MUC13
MUC12
MUC9/OVGP1
MUC7
MUC6
MUC5
MUC4
MUC3
MUC2
MUC1
0
50
100
Not-Detected
Low
Medium
High
MUC1
MUC2
MUC3A
MUC4
MUC5AC
MUC6
MUC7
OVGP1
MUC12
MUC13
MUC14
MUC1
MUC16
MUC17
MUC18

Figure 4.5 Correlation of Mucins in GC.

Correlation and co-expression study mucins was studied using the STRING database and “corrplot” package in R-Bioconductor **A.** Protein-protein interaction network of all mucins with k-means (N=4) based clustering. Each color represents a specific cluster. **B.** Correlation plot of mucins from TCGA-STAD expression with highly positively correlated cluster marked.

Figure 4.5

A.


MUC5AC
GALNT12
MUC20
MUC15
MCAM
MUC21
MUC4
EMCN
MUC2
MUC13
MUC22
MUC12
MUC17
MUC5B
MUC7
MUC6
OVGP1
MUC1
MUC3A

B.

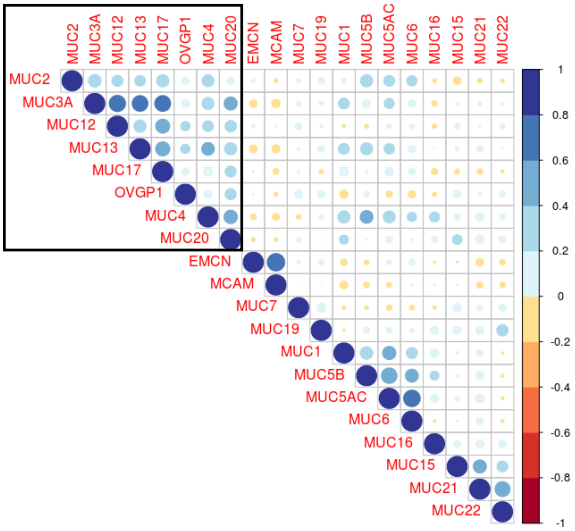
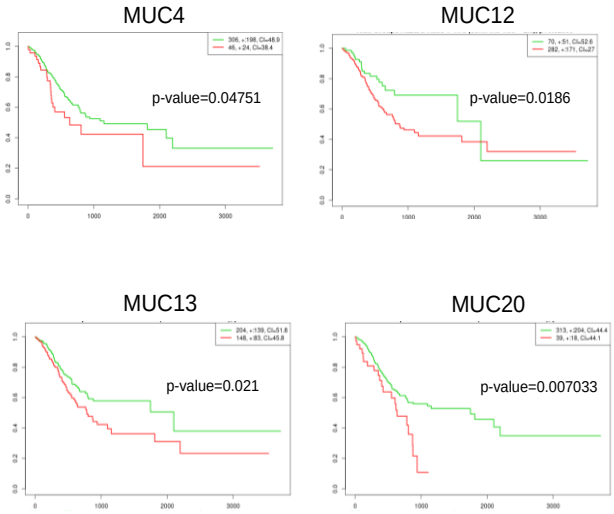

MUC2
MUC3A
MUC12
MUC13
MUC17
OVGP1
MUC4
MUC20
EMCN
MCAM
MUC7
MUC19
MUC1
MUC5B
MUC5AC
MUC6
MUC16
MUC15
MUC21
MUC22
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 4. 6 Survival analysis of Mucins.

Survival differences were studied for individual and previously identified clusters of Mucins together. **A.** Individual survival plots for MU4, MUC12, MUC13, and MUC20. **B.** Survival differences of highly correlated cluster MUC2-MUC12-MUC13-OVGP1-MUC4-MUC20.

Figure 4.6

A.


MUC4
p-value=0.04751
MUC12
p-value=0.0186
MUC13
p-value=0.021
MUC20
p-value=0.007033

B.

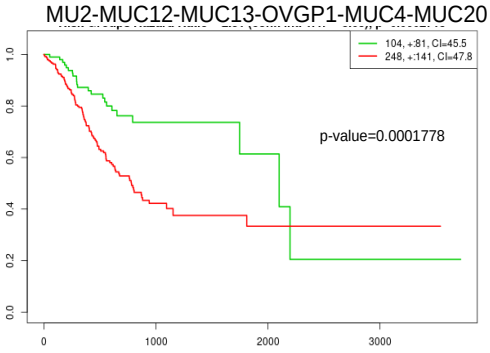
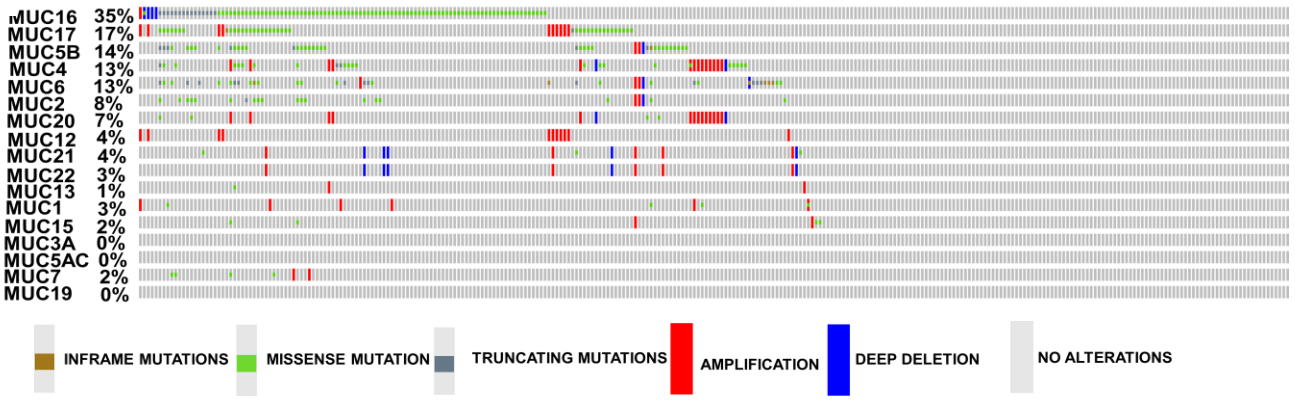

MU2-MUC12-MUC13-OVGP1-MUC4-MUC20
104, +:81, CI=45.5
248, +:141, CI=47.8
p-value=0.0001778
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000

Figure 4. 7 Mutational analysis of Mucins.

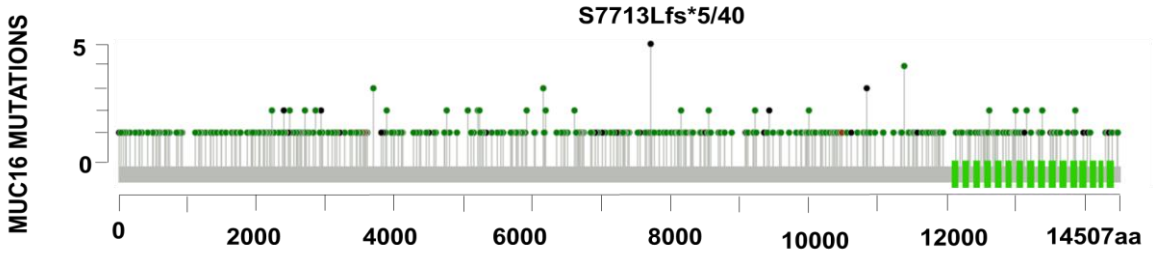
Mutational analysis was carried out in the TCGA-STAD cohort using the web-tool cBioPortal. **A.** Oncoprint is depicting the percentage mutations in all mucins in GC. **B.** Positional mutations in MUC16 are based on increasing length.

Figure 4.7

A.


MUC16 35%
MUC17 17%
MUC5B 14%
MUC4 13%
MUC6 13%
MUC2 8%
MUC20 7%
MUC12 4%
MUC21 4%
MUC22 3%
MUC13 1%
MUC1 3%
MUC15 2%
MUC3A 0%
MUC5AC 0%
MUC7 2%
MUC19 0%
INFRAME MUTATIONS
MISSENSE MUTATION
TRUNCATING MUTATIONS
AMPLIFICATION
DEEP DELETION
NO ALTERATIONS

B.


S7713Lfs*5/40
MUC16 MUTATIONS
5
0
0
2000
4000
6000
8000
10000
12000
14507aa

4.5 DISCUSSION

Gastrointestinal (GI) cancers, including pancreatic, colorectal, and gastric cancers, have an aberrant expression and an extremely significant role of the mucin family in their initiation and progression. Considering this important role, mucins have in turn been established as the pertinent biomarkers (*90, 101*) and in therapeutics (*75, 113, 114*). Specifically, in GC, various mucins like MUC1, MUC2, MUC5AC MUC6, and MUC13 have been studied for their important roles as the therapeutic targets or prognostic and predictive biomarkers. (*115-120*). While studies have helped us make some progress in understanding the role of mucins in GC, various gaps remain. The current study, with the help of various bioinformatics tools, sought to assess the current status of various members of the mucin family in GC.

The high-throughput expression analysis of the FPKM data from the TCGA-STAD dataset led to the identification of specific clusters of mucins within the GC TCGA dataset. Specifically, MUC1 and MUC13 were found to be strongly associated with each other, which is intriguing since studies have been conducted to study the collective role of these two transmembrane mucins (*121*)**.** However, this analysis was limited by the imbalance in the normal and tumor samples. Considering this, an in-depth analysis was carried out using the web tool GEPIA by comparing the TCGA (only tumor samples, N=408) dataset to the normal data from the genotype-tissue expression (GTEx) portal. This comparison led to the identification of various significantly differentially expressed genes. Significant downregulation of MUC6 and MUC5AC was observed from normal to cancer cases. Interestingly, MUC6

and MUC5AC are resident mucins of the gastric canal and various studies have previously reported a downregulation or loss of these mucins in patient cohorts (*120, 122-125*). Further, the expression analysis led to the identification of various upregulated mucins like MUC2, MUC3A, MUC4, MUC5B, MUC12, MUC13, and MUC17. Interestingly, both MUC13, and MUC17 are known to be upregulated and have specific roles in GC (*105, 126*) and various efforts are currently underway to target and study these two mucins further in GC. Furthermore, an independent assessment of expression differences in immunohistochemistry data from the Human Protein Atlas reiterated the high expression of MUC3A, MUC4, MUC13, and MUC17. Furthermore, a very high expression of MUC1 has observed in these tissue sections in contrast to the TCGA dataset since the expression differences in TCGA vs GTEx normal cases were not found to be significant. This observation further goes back to the initial observation of MUC1 and MUC13 forming an extremely strong cluster with each other. Furthermore, to identify the relationship of these mucins with each other and their interdependence in functionality, a two-way correlation and co-expression analysis were carried out. Interestingly, through a k-means clustering of the STRING-based protein-protein interaction analysis, two major clusters namely MUC3A-MUC12-MUC13-MUC15-MUC17-MUC20-MUC22 and MUC1-MUC4-MUC5AC-MUC5B-MUC6-MUC7-MUC16 were observed. Further, a spearman correlation analysis in the TCGA-STAD cohort identified a strongly positively correlated cluster containing MUC2-MUC3A-MUC12-MUC13-MUC17-OVGP1/MUC9-MUC4-MUC20. This observation opened the doors for further questions, one of which was if the hence identified cluster

possesses a prognostic role in GC. To answer this, we went on to assess the survival differences in the TCGA-STAD concerning the expression of these mucins. Interestingly, MUC4, MUC12, MUC13, and MUC20 were found to have an independent prognostic relevance in the TCGA-STAD cohort when analyzed with a risk-maximization algorithm and survival days as the censoring variable. Further, more intriguing was the fact that when analyzed together, the cluster of MUC2-MUC12-MUC13-MUC9/OVGP1-MUC20 was found to significant (Log-Rank-p-value=0.0001778) associated with survival hence identifying a prognostic biomarker panel. Further, mutational analysis of all mucins in GC revealed a very high percentage mutation in MUC16, MUC17, MUC5B, MUC4, and MUC6. Interestingly, MUC16 mutations have been known to have prognostic and functional implications in GC and various other cancers, opening up various avenues for further research and potential applications in GC.

Chapter 5 Presence and structure-activity relationship of intrinsically disordered regions across mucins

Portions of the content covered in this chapter are the subject of a published article in FASEB (Carmicheal J, Atri P, Sharma S, Kumar S, Chirravuri Venkata R, Kulkarni P, Salgia R, Ghersi D, Kaur S, Batra SK. Presence and structure-activity relationship of intrinsically disordered regions across mucins FASEB J. 2020 02; 34(2):1939-1957).

Presence and structure-activity relationship of intrinsically disordered regions across mucins

Joseph Carmicheal[#1], Pranita Atri[#1], Sunandini Sharma[#1], Sushil Kumar[1,2] Ph.D., Ramakanth Chirravuri Venkata[1] Ph.D., Prakash Kulkarni[3] Ph.D., Ravi Salgia[3] M.D., Ph.D., Dario Ghersi[4,*] Ph.D., Sukhwinder Kaur[1,2*] Ph.D., Surinder K. Batra[1,2*] Ph.D

[1]Department of Biochemistry and Molecular Biology, University of Nebraska Medical center, Omaha, USA

[2]Buffett Cancer Center, University of Nebraska Medical Center, Omaha, NE, USA

[3]Department of Medical Oncology and Therapeutics Research, City of Hope, Duarte, California, USA

[4]School of Interdisciplinary Informatics, University of Nebraska at Omaha, Omaha, NE, USA

[#]Equal contribution

***To whom correspondence should be addressed**:

Dario Ghersi, M.D., Ph.D. Sukhwinder Kaur, Ph.D., Surinder K. Batra, Ph.D.

Department of Biochemistry and Molecular Biology,

Eppley Institute for Research in Cancer and Allied Diseases,

University of Nebraska Medical Center, 985870 Nebraska Medical Center, Omaha, NE, 68198-5870, U.S.A., Phone: 402-559-5455; Fax: 402-559-6650,
E-mails: dghersi@unomaha.edu; skaur@unmc.edu; sbatra@unmc.edu

Running Title: Intrinsic disorder across mucins

NON-STANDARD ABBREVIATIONS LIST

IDR, intrinsically disordered region; IDP, intrinsically disordered protein; PTM, post-translational modification; MoRF, molecular recognition feature; PPI, protein-protein interaction; TM, transmembrane; VNTR, variable number of tandem repeat domain; PTS sequence, proline, threonine and serine sequence; EGF, epidermal growth factor-like domain; SEA, sea-urchin sperm protein enterokinase and agrin module; vWD, von Willebrand factor D domain; NIDO, nidogen-like domain; AMOP, adhesion-associated domain in MUC4 and other proteins domain; ECM, extracellular matrix; D^2P^2, Database of Disordered Protein Predictions; CH, charge hydropathy;

5.1 SYNPOSIS

Many of the 20-member mucin family are evolutionarily conserved proteins that are often aberrantly expressed and glycosylated in various benign and malignant pathologies including oncogenic signaling leading to tumor invasion, metastasis, and immune evasion. Large size and extensive glycosylation present challenges to study mucin structure using traditional methods, including crystallography. We offer the hypothesis that the functional versatility of mucins may be attributed to the presence of intrinsically disordered regions (IDRs), which provide dynamism and flexibility; further, that these sites offer potential therapeutic targets. Herein, we examined the links between mucin structure and function based on IDRs, post-translational modifications (PTMs), and potential impact on their interactome. Using sequence-based bioinformatics tools, we observed that mucins are predicted to be moderately (20-40%) to highly (>40%) disordered and many conserved mucin domains could be disordered. Phosphorylation sites overlap with IDRs throughout the mucin sequences. Additionally, the majority of predicted *O*- and *N*- glycosylation sites in the tandem repeat regions occur within IDRs, and these IDRs contain a large number of functional motifs, i.e. molecular recognition features (MoRFs), which directly influence PPIs. This investigation provides a novel perspective and offers an insight into the complexity and dynamic nature of mucins.

5.2 BACKGROUND AND RATIONALE

5.2.1 Mucin Protein Family

Mucins (MUCs) are heavyweight (over 10^6 Dalton) glycoproteins that are expressed by epithelial cells in many organs throughout the body (*127*). In humans, the mucin protein family contains over twenty members and is subdivided, based on structural differences, into transmembrane and secretory mucins. The primary distinction between these two groups is the presence or absence of a transmembrane domain (TM), which anchors them to the cell membrane. Mucins contain a characteristic large polymorphic variable number of tandem repeat domain (VNTR) that is rich in proline, threonine and serine residues (PTS). The VNTR is susceptible to enzymatic modification by *O*-linked and *N*-linked oligosaccharides (*128*). All mucins harbor one or more domains with high sequence similarity to a known functional domain present in other proteins. These include the EGF-like domain (EGF), sea-urchin sperm protein, enterokinase and agrin (SEA) domain, von Willebrand factor D domain (vWD), nidogen-like domain (NIDO), the adhesion-associated domain in MUC4 and other proteins (AMOP), and the D-domain (*129*). These domains have been implicated in several biological processes such as cell-to-cell interaction (*130*), cell-to-ECM interaction (*131*), apoptosis inhibition (*132*), and cell signaling complexes (*133*).

Aberrant expression, splicing, and glycosylation in various members of the mucin family is a characteristic feature of several malignancies including pancreatic ductal adenocarcinoma (*101, 127, 134)*, colorectal (*73, 135)*, lung (*136, 137*) and ovarian cancer (*138*). Further, tumor cells exploit mucin differential localization,

alternative splicing, cellular adhesive/anti-adhesive properties, and alterations in glycosylation profile, to metastasize to distant locales and survive in hostile environments (*127, 139*).

5.2.2 Intrinsically Disordered Proteins

Until recently, it was generally held that the three-dimensional structure of a protein defined its function (*140*). However, it is now well established that intrinsically disordered proteins (IDPs) and regions (IDRs), are complete proteins (or segments of proteins) that lack a traditional globular secondary or tertiary structure yet are fully functional (*141-146*). Disordered regions generally are sequences of low complexity with a low proportion of hydrophobic residues and a high number of repeating residues with a preponderance of polar and charged residues (*147, 148*). This lack of bulky hydrophobic amino acids prevents the formation of an ordered core that comprises a traditional structured domain (*149*). Disorder is ubiquitous throughout the human proteome. A study estimated that 30% of all proteins harbor some degree of disorder with a majority of these proteins containing disorder ranging between 20-40% of their total sequence (*149-151*)

IDPs/IDRs have wide-ranging implications in various physiological and biological processes such as transcription, splicing, translation, and signaling (*144, 152-155*), scaffolding (*156, 157*), cell cycle regulation (*19, 158, 159*), protein-protein interactions (PPIs) (*143, 160-164*), chaperoning (*141*), and phenotypic plasticity (*165, 166*) (that is the ability to switch phenotypes). Further, IDP/IDR-mediated modulations are implicated in the pathogenesis of various diseases such

as cancer, diabetes, cardiovascular defects, amyloidosis, and neurodegeneration (*167*). More specifically, many cancer-associated proteins have been shown to have a higher percentage of IDRs relative to the rest of the human proteome (*13, 167-169*).

This functional versatility of IDPs/IDRs encourages us to investigate their presence in known cancer-associated proteins like mucins. Molecular events such as mutations that increase protein hydrophilicity, or alter protein splicing, can lead to changes in IDR length and affect protein-protein interactions, leading to pathological properties. This often affects protein solubility and aggregation, leading to nonproductive or over-productive complexes that disturb regulatory (*149, 170*).

Considering the significant role of mucins in normal physiology as well as pathological conditions, hub protein characteristics, and their simple abundance and aberrant expression tendency in a variety of cancers, IDR/IDP presence within mucins could have important clinical implications. Mucins are also prime targets for IDP/IDR analyses because conventional methods of structural delineation are limited by large size, the high number of PTMs, and the presence of multiple splice variants.

For many proteins implicated in cancer, structural biology information has proven invaluable for understanding their functional implications as well as discovering novel therapeutic modalities (*171-174*). Unfortunately, it is difficult to study mucins structurally by traditional methods. Crystallographic methods falter because of the

sheer size of mucins (up to 14,000 kDa), extensive variation in the number of tandem repeats within the VNTR (up to 120), sequence variation, and inability to clone, express, and purify fully folded and glycosylated forms. While specific domains have been cloned, purified, and studied (i.e. SEA (*175*)), domain homology between mucins and other proteins varies. What structural analyses have been accomplished (via x-ray crystallography) were conducted domain-by-domain and not as a part of the complete protein (*176*). In addition, improper refolding of solitary domains is a constraint on these experiments. This dearth of advanced structural knowledge constrains the investigation of mucins as possible therapeutic targets.

Based on the earlier studies, we hypothesized that the functional versatility of mucins may be attributed to the presence of intrinsically disordered regions (IDRs); further, that these sites offer potential therapeutic targets.

To support this hypothesis, we analyzed the protein sequences of mucins using the Database of Disordered Protein Predictions (D^2P^2) to predict disorder based on a ≥75% consensus between the nine disorder prediction models incorporated within the tool itself (*177*). The presence of IDRs was determined within conserved mucin domains, inter-domain sequences, C-terminal, and transmembrane domains. Next, we assessed the relationships of IDRs with curated phosphorylation sites and predicted *N*- and *O*-glycosylation sites, to discern whether posttranslational modifications occurred preferentially in IDRs within the mucin sequences. Finally, we also assessed the effect of conformational disorder on the mucin family interactome.

5.3 MATERIALS AND METHODS

5.3.1 Mucin disorder prediction with D^2P^2

Mucin sequences were searched for predicted disorder using the text search option provided on the web-based D^2P^2 (version v0.3-689) (*177*) portal (http://d2p2.pro/). D^2P^2 comprises nine different disorder prediction tools involving a variety of prediction methods (Espritz-D, Espritz-X, Espritz-N, IUPred-L, IUPred-S, PV2, PrDOS, VSL2b, VLXT). Due to variable length and high sequence ambiguity for tandem repeat domains of mucins, the longest available human mucin transcripts that were present in D^2P^2 were used for our analysis. Disorder was predicted based upon 75% consensus of all nine predictors, and high confidence disorder regions were then obtained. The percentage disorder was computed by dividing the total length of disordered regions by the protein sequence length. For our study, mucins with < 20% disorder will be considered to have low levels of disorder, those with >20% and <40% will be defined as moderately disordered, and those with >40% will be considered as highly disordered.

5.3.2 Mucin disorder prediction with FoldIndex

An algorithm originally developed by Uversky and colleagues (*178*) was implemented using an in-house python script that predicts if a region in a protein sequence would assume a folded or intrinsically unfolded state. This algorithm works on two properties of an amino acid: net charge and hydrophobicity of amino acids. The net charge represents the difference between the positive and

negative amino groups at a physiological pH = 7.0, and the mean hydrophobicity is the sum of the individual hydrophobicity of each residue divided by the total number of residues. The Kyte-Doolittle scale was used to determine the hydrophobic propensities of amino acids in the protein sequence, and the following equation was used to calculate the disorder score (I):

$$\boldsymbol{I = 2.785 \times \langle H \rangle - |\langle R \rangle| - 1.151}$$

In the above equation, <H> represents the mean hydrophobicity, i.e. the sum of hydrophobicity of all residues, and |<R>| is the absolute difference between positively and negatively charged residues. The protein sequences were inputted to the python script, which calculates the score for each residue in the sequence. It is noteworthy to mention that this algorithm assumes that different regions of a protein vary in their folding properties, so a sliding window scores specific regions of proteins rather than the whole protein. Note, the length of the sliding window was 51 aa, as used in the original study.

5.3.3 Domain-wise disorder prediction of mucins

Protein domains were predicted on mucin sequences using the open-access Pfam (version 32.0 produced at the European Bioinformatics Institute, September 2018) (https://pfam.xfam.org/) and CD-Search (*179*) (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) interfaces. Only those with significant E-values (a parameter denoting significance of the actual number of sequences aligned compared to the number expected by chance) were utilized. Domain coordinates for each mucin were compared with predicted disordered

regions (D^2P^2). **Figure 1c** represents the presence or absence of these domains across mucins (x-axis) and their corresponding disorder (represented by *) across mucins (y-axis).

5.3.4 Disorder prediction in the cytoplasmic tail and transmembrane domains of mucins

Literature searches provided the starting point for the cytoplasmic tail (CT) sequences of mucins (*129*). Furthermore, the transmembrane (TM) sequences were obtained from Mucin database 2.0, 2015 (*180*) (http://www.medkem.gu.se/mucinbiology/databases/). Disorder predicted by D^2P^2 was compared within the transmembrane and cytoplasmic region to determine the specific residues disordered in these regions.

5.3.5 MoRFs prediction

The D^2P^2 database also identifies molecular recognition features (MoRFs) across proteins by using ANCHOR (*181*). This web server predicts disordered binding regions using protein sequences. The total number of MoRFs in each mucin was divided by its length. For instance, MUC12 has 145 predicted MoRFs and a length of 5478 aa, yielding a representative value of ~0.026. This assessment helped us identify the relative MoRFs per base pair in each of the mucins, enabling a relative assessment across mucins.

5.3.6 PhosphoSitePlus® curated phosphorylation site

PhosphoSitePlus® is a database of mammalian post-translational modification sites curated from the scientific literature. Over 95% of the presented PTM sites

have been elucidated by tandem mass spectrometry experimentation requiring a P <.05 for each site assignment (http://www.phosphosite.org/) (*182*). Phosphorylation sites curated by PhosphoSitePlus® were presented as a part of D^2P^2 analyses. These were subsequently aligned with the predicted IDRs and MoRFs for each mucin individually, and the proportion found within IDRs.

5.3.7 Predicted O- and N-linked Glycosylation sites

N- and *O*-linked glycosylation sites were predicted for all mucin sequences using NetNGlyc 1.0 (*183*) server and NetOGlyc 4.0 (*184*) server, respectively. The NetNGlyc 1.0 server is an artificial neural network-based program that examines the Asn-Xaa-Ser/Thr sequons, AA sequence at which an *N* glycosylation can occur and predicts if a residue can act as a potential *N* glycosylation site with an accuracy of 77%. The NetOGlyc 4.0 server was derived by a 'bottom-up' ETD-based mass spectrometric analysis of 12 human cancer cell lines to develop a training set of the human *O*-glycoproteome. Mucin sequences from D^2P^2 were queried into these tools and potential *N*- and *O*-glycosylation sites were predicted. These predicted sites were then compared with the pre-computed disordered regions across mucins. We performed these predictions for two transmembrane mucins: MUC1 and MUC4 and two secretory mucins: MUC2 and MUC6. A threshold of 0.5 was used to predict a potential *N*- and *O*-glycosylation site within the tandem repeat domains of these mucins and individually analyzed the ability of these domains to serve as potential sites for *N*- and *O*-glycosylation post-translational modifications.

5.3.7 PONDR-VSL2

To further analyze the disorder regions in mucins and assess their inter-species pattern, the PONDR (*185*) based tool VSL2 was used. Mucin sequences from D^2P^2 were loaded into the online tool and the graphical representation was analyzed. With this method, an amino acid residue with a score close to 0 is considered to be ordered whereas the score approaching 1 is considered as disordered with a defined cutoff of ordered vs. disordered at a value of 0.5.

5.3.8 Mucin interactome and functional annotation of mucin interaction partners

All mucins that were assessed for intrinsic disorder (MUC1, MUC2, MUC3, MUC4, MUC5B, MUC6, MUC7, MUC9, MUC12, MUC13, MUC14, MUC15, MUC16, MUC17, MUC18, MUC19, MUC20, MUC21, MUC22) were included for the Reactome functional analysis. We obtained 25 major mucin pathways and retained the pathway names along with the adjusted p-value. The protein-protein interaction information was extracted using the Biological General Repository for Interaction Datasets (BioGRID) *Homo sapiens* database (*186*). Duplicate edges emerging due to different validation methods were removed from the network to prevent redundancy. A total of 144 unique interactions between mucins and other proteins were obtained. BioGRID interactions are either peer-reviewed or experimentally validated by empirical protein-protein interaction methods. Further, we performed GO (*187*) analysis for mucins to illustrate the functional versatility and the implications in cancer-associated pathways. To further elucidate the pathways to which mucins contribute, we next used the FunSet webserver (*188*) to cluster and

visualize the enriched GO pathway terms. This technique detects semantic similarity between terms and spectrally clusters them into neighborhoods in a bubble chart format.

5.4 RESULTS

5.4.1 Intrinsic disorder analysis across Mucins

A substantial portion of mucin sequences lacks meaningful structural annotation. In order to analyze the degree of disorder and their location across mucins, the percentage of disorder was calculated with 75% prediction consensus by nine disorder predictors present in D^2P^2 database (i.e., 7 out of 9 tools in agreement). Therefore, these meta-prediction tools provide more reliable predictions than a single disorder prediction tool (*177*).

The longest available protein sequence of mucins present within D^2P^2 was used in this analysis. The mucins analyzed included MUC1, MUC2, MUC4, MUC5B, MUC6, MUC7, MUC9, MUC12, MUC13, MUC14, MUC15, MUC16, MUC17, MUC20, and MUC21. Due to the unavailability of the sequences within D^2P^2, some mucins such as MUC3, MUC18, MUC19, and MUC22, were not included in the analysis. Also, MUC5AC (with an extremely short transcript available within the database), MUC8 (which is not a mucin protein although called MUC8 (*180*), MUC10 (which is found only in mice) and MUC11 (which is part of MUC12 (*180*)) were all excluded from the analysis.

The disorder prediction analysis of mucin sequences showed that all mucins were moderately (20% - 40%) to highly (40%-90%) disordered. MUC12 (90%), MUC17 (87%), and MUC21 (83%) were the most disordered transmembrane mucins (**Fig. 5.1A**). Of the other transmembrane members, MUC20 (79%) and MUC4 (77%) were more disordered compared to MUC1 (60%) and MUC16 (63%). All transmembrane mucins were considered highly disordered with the exception of MUC13 (34%) and MUC15 (25%) which were moderately disordered. For secreted mucins, MUC7 (80%), MUC5B (48%), and MUC6 (44%) were highly disordered, whereas MUC9 (25%) and MUC2 (23%) were moderately disordered (**Fig. 5.1B**). We observed that disorder occurs more in transmembrane mucins compared to secreted mucins with the exception of MUC7 (**Fig. 5.1A**).

Next, we determined the domains for all mucins using Pfam (*189*) and the Conserved Domain Database (CDD) (*190*). These sequences were subsequently analyzed for the presence of IDRs. Domains determined using two databases allowed higher confidence predictions with a significant E-value (E). We observed that the SEA domain was correctly predicted across transmembrane mucins including MUC1, MUC12, MUC13, MUC16 and MUC17 as confirmed by the Mucin Biology Database (**Fig. 5.1C**). Similarly, vWD was predicted to be present in MUC2, MUC4, MUC5B, and MUC6 (**Fig. 5.1C**)

Combined analyses of the mucin domain sequence and disorder prediction identified the vWD domain of MUC4 to be disordered. In addition to the vWD domain, MUC4 also contains AMOP and NIDO domains (**Fig. 1c**), but the disorder predicted by D^2P^2 did not reach the 75% consensus pre-set cut-off. Three of the

nine tools predicted sections of these domains to be disordered. The Mucin2_WxxW (a.k.a. CysD) domain known to contain a conserved repeat sequence motif (WxxW) of at least six conserved cysteine residues, also displayed a high level (>40%) of disorder; CysD was also predicted to be present in MUC2 and MUC5B. The Endomucin domain in MUC14, a highly *O*-glycosylated region that affects cell adhesion, was predicted to be disordered as well (**Fig. 1c**). MUC21 contained repeating motifs, represented by epiglycan TR and epiglycan C, which were also found to contain disorder (**Fig. 5.1C**).

5.4.2 Intrinsic disorder in trans-membrane and intracellular c-terminal domains of Mucins

Next, we used D^2P^2 to predict disorder in the transmembrane and C-terminal domains of mucins. The cytoplasmic tail protein sequences of MUC1, MUC4, MUC12, MUC13, MUC15, MUC16, MUC17 and MUC20 (**Fig. 5.2A**) were obtained from earlier published findings (*129*). Though the majority of the cytoplasmic tail sequences of MUC4 and MUC16 did not reach the pre-set 75% disorder consensus cutoff, a high level of agreement between tools, 6 out of 9, found them to be disordered (66%) (**Fig. 5.2B**).

Similarly, we obtained transmembrane domain sequences from the Mucin Biology Database (Human) and assessed if those transmembrane sequences were disordered in the global disorder prediction of mucins by D^2P^2. No disorder was observed within the transmembrane domains of mucins (**Fig. 5.2C**). It is established that disorder prediction consensus approaches are generally more

accurate (*191*), however, to verify our observation, we analyzed the transmembrane sequences with DisEMBL for further confirmation of our predictions. Similar to D^2P^2, no disorder was observed for the transmembrane domain of mucins with DisEMBL. Representative figures for MUC4 and MUC12 show low disorder probability within the transmembrane domains (**Fig. 5.2D**). This is in line with the fact that transmembrane regions are largely hydrophobic static regions and are not involved in dynamic protein-protein interactions, thus decreasing the probability of disorder.

Figure 5. 1 Intrinsic disorder across mucins determined using the Database of Disordered Protein Predictions (D^2P^2).

D^2P^2 is a database of pre-computed disorder predictions on a large library of proteins from completely sequenced genomes. **A.** Bar graph displaying the percentage of intrinsic disorder across transmembrane (grey) and secreted (black) mucins. All mucins analyzed were either highly disordered (defined as >40% disorder) or moderately disordered (>20% and <40% disorder). The disorder was calculated by dividing the total number of 75% consensus disordered residues, by the total mucin length to obtain a percentage disorder for each mucin. **B.** Pictorial representation of intrinsic disorder observed in MUC1 (length=550 amino acids) with a truncated tandem repeat region, as available within D^2P^2. The portion of the protein sequence with a high degree of consensus between tools (at least 6 of 9) is demarcated by green coloring. The regions with a lower degree of consensus (3-5 of 9 tools) but still predicted as disordered, are demarcated by shades of blue (darker blue denotes higher consensus). Both the N-terminus and C-terminus contain disordered sequences, as does much of the extracellular domain. **c.** Intrinsic disorder observed within different mucin domains as predicted by Pfam and CD-search databases. The presence of intrinsic disorder is denoted by (*). D^2P^2 data analyses suggested that disorder is present within the vWD domain of MUC4, WxxW domain of MUC2 and MUC5B, endomucin domain of MUC14, and the Epiglycan TR and C domains of MUC21

Figure 5.1

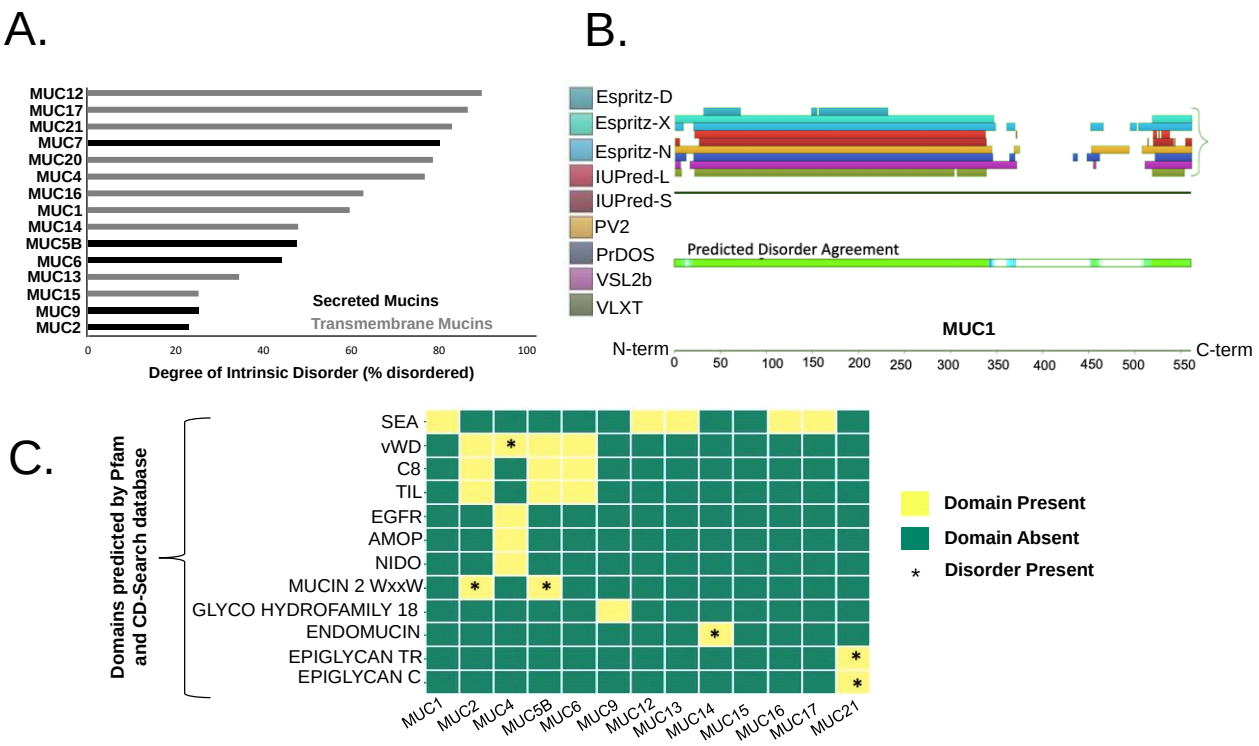

A.
MUC12
MUC17
MUC21
MUC7
MUC20
MUC4
MUC16
MUC1
MUC14
MUC5B
MUC6
MUC13
MUC15
MUC9
MUC2
Secreted Mucins
Transmembrane Mucins
0
20
40
60
80
100
Degree of Intrinsic Disorder (% disordered)
B.
Espritz-D
Espritz-X
Espritz-N
IUPred-L
IUPred-S
PV2
PrDOS
VSL2b
VLXT
Predicted Disorder Agreement
MUC1
N-term
C-term
0
50
100
150
200
250
300
350
400
450
500
550
C.
Domains predicted by Pfam and CD-Search database
SEA
vWD
C8
TIL
EGFR
AMOP
NIDO
MUCIN 2 WxxW
GLYCO HYDROFAMILY 18
ENDOMUCIN
EPIGLYCAN TR
EPIGLYCAN C
MUC1
MUC2
MUC4
MUC5B
MUC6
MUC9
MUC12
MUC13
MUC14
MUC15
MUC16
MUC17
MUC21
Domain Present
Domain Absent
* Disorder Present

Figure 5. 2 Assessment of intrinsic disorder in cytoplasmic and transmembrane domains of membrane-tethered mucins.

Considering the differential sequence attributes for transmembrane (hydrophobic and lacking protein-protein interacting sites) and cytoplasmic domains (sites for purported signaling functions of mucins), we assessed intrinsic disorder regions across these domains. **A.** Intrinsically disordered residues (red color) observed within cytoplasmic tails of MUC1, MUC12, MUC13, MUC15, and MUC20 with a 75% consensus of D^2P^2. **B.** Of note, the entirety of the MUC16 and MUC4 CT domains are predicted to be disordered by 6 of 9 tools reaching 66% consensus. Pictorial representation from the D^2P^2 disorder predicted in the cytoplasmic tails of MUC4 and MUC16. **C.** Assessment of IDR in transmembrane regions of mucins. No intrinsic disorder was observed within the transmembrane domains. **D.** Representative figure of MUC4 and MUC12 transmembrane domain displaying extremely low disorder probability as determined by DisEMBL. Disorder probability increases as values approach 1 and decreases as it approaches 0. The greenish-yellow curve is the disorder prediction for missing residues (those lacking crystal structure), the red curve indicates residues predicted as hot loops, and the blue curve indicates those predicted as coils. Of note, loops and coils are considered necessary but not sufficient for disorder and the lack of these features as predicted by DisEMBL is indicative of a low level of disorder

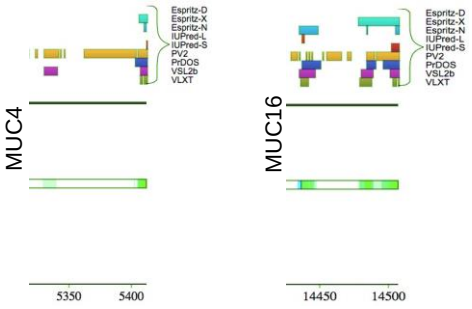
Figure 5.2

A.

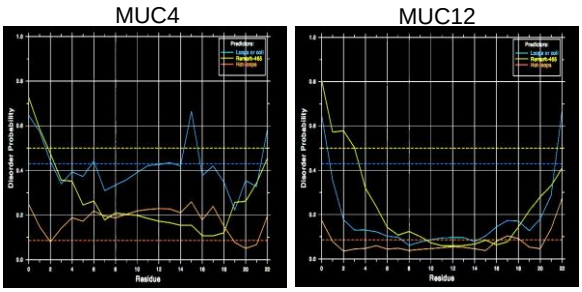
Disorder in cytoplasmic tails of transmembrane Mucins	
MUC1	RRKNYGQLDIFPARDTYHPMSEYPTYHTHGRYVPPSSTDRSPYEKVSAGNGGSSLSYTNPAV AATSANL
MUC4	GCSGARFSYFLNSAEALP
MUC12	SQRKRHREQYDVPQEWRKEGTPGIFQKTAIWEDQNLESRFGLENAYNNFRPTLETVDSGTEK HIQRPEMVASTV
MUC13	TARSNNKTKHIEEENLIDEDFQNLKLRSTGFTNLGAEGSVFPKVRITASRDSQMQNPYSRHSSM PRPDYDIPPLRTSV
MUC15	CGKAKTDSFSHRRLYDDRNEPVLRLDNAPEPYDVSFGNSSYYNPTLNDSAMPESEENARDGI PMDDIPPLRTSV
MUC16	VTTRRRKKEGEYNVQQQCPGYYQSHLDLEDLQ
MUC17	RSKREVKRQKYRLSQLYKWQEEDSGPAPGTFQNIGFDICQDDSIHLESIYSNFQPSLRHIDPET KIRIQRPQVMTTSF
MUC20	RNSLSLRNTFNTAVYHPHGLNHGLGPGGNHGAPHRPRWSPNWFWRRPVSSIAMEMSGRNS GP

Transmembrane domains of Mucins	
MUC1	WGIALLVLVCVLVALAIVYLIAL
MUC4	IFFGALGGLLLLGVGTFVVLRFW
MUC12	GIVGAVMAVLLLALIILIILTMFSL
MUC13	LILTIVGTIAGIVILSMIIALIV
MUC14	LPVVIALIVITLSVFVLVTMGLY
MUC15	IVFGAILGAILGVSLLTLVGYLL
MUC16	VILIGLAGLLGLITCLICGVLTMVT
MUC17	YGLVGAGVVLMLIILVALLMLVF
MUC22	WAIILISLAAVVAAVGLSVGTML

B.



C.



5.4.3 Assessment of molecular recognition features (MoRFs) in mucin IDRs

Disorder-to-order transition of IDRs is facilitated by stretches of amino acids known as MoRFs, which facilitate molecular recognition and signal transduction (*192*). MoRFs undergo a conformational change to a lower energy state when interacting with an appropriate binding partner and are thus one of the keys to the executioner function of IDRs (*192*).

The presence of MoRFs was determined via ANCHOR, as a component of D^2P^2 that determines sequence motifs within an IDR that have a decrease in free energy upon binding with another protein (*193*). Interestingly, mucins contain a large number of predicted MoRFs within their IDRs (**Fig. 5.3**). Transmembrane mucins, particularly MUC12 and MUC16, were predicted to contain greater numbers of MoRFs within their IDRs compared to other mucins (145 and 413, respectively, **Fig. 5.3**). Further analysis showed that mucins implicated in multiple human cancers particularly MUC4, MUC17, and MUC16 contain a large number of MoRFs >30 residues (38, 46, and 46, respectively, **Fig. 5.3**). When the number of MoRFs are normalized to mucin protein length by dividing with the total number of residues in each mucin, MUC12 and MUC4 have the greatest number of MoRFs, at a ratio of 0.026 and 0.022, respectively. Within secreted mucin members, MUC5B has the highest total number of MoRFs with 72, as well as the highest normalized quantity at 0.013.

5.4.4 Delineating the association of mucin IDRs & PTMs

Many post-translational modifications (PTMs) that modulate protein actions and interactions are predicted within disordered regions of mucins and, more

specifically, some reside within IDRs that harbor MoRFs (**Fig. 5.4A**, yellow and black bars are predicted as MoRFs). The predicted presence of smaller MoRFs within IDRs provides structural insight into the function of each mucin. Further investigation in this regard would help to associate the presence of disorder and MoRFs with domain and inter-domain functionality.

IDRs within structured protein domains are preferred and accessible sites for a variety of PTMs, including glycosylation and phosphorylation (*194*). With this in mind, we compared the phosphorylation sites to the regions of disorder for each mucin. For this, we utilized curated phosphorylation sites provided by PhosphoSitePlus® as a part of D^2P^2 as well as direct inquiry via PhosphoSitePlus® and subsequent sequence alignment. We found that the majority of phosphorylation sites were found within the predicted disordered regions when assessing entire mucin sequences (**Fig. 5.4B**). The proportion of phosphorylation sites found within IDRs for each mucin are as follows: MUC15 – 1.0, MUC4 – 1.0, MUC14 – 1.0, MUC20 – 1.0, MUC12 – 0.97, MUC17 – 0.91, MUC6 – 0.89, MUC16 – 0.86, MUC5B - 0.76, MUC13 - 0.75, MUC21 - 0.71, MUC9 - 0.5, MUC9 – 0.5, MUC2 - 0.44, and MUC1 - 0.36 (**Fig. 5.4B**). MUC7 did not have any curated phosphorylation sites presented within D^2P^2 nor with direct inquiry via PhosphoSitePlus®. The proportions for phosphorylation sites found to reside in MoRFs within IDRs are as follows: MUC15 – 0.0, MUC4 – 1.0, MUC14 – 0.2, MUC20 - 0.57, MUC12 – 0.42, MUC17 – 0.59, MUC6 – 0.38, MUC16 – 0.49, MUC5B - 0.14, MUC13 – 0.0, MUC21 – 0.0, MUC9 – 0.0, MUC2 – 0.0, and MUC1 – 0.0 (**Fig. 5.4B**).

We then specifically analyzed the tandem repeat regions of mucins to correlate *N*- and *O*- glycosylation, and high level of disorder predicted by D^2P^2. The predicted *N*- and *O*- glycosylation sites reside almost exclusively in IDRs compared to the ordered regions within tandem repeat regions (representative transmembrane mucins, MUC1 & MUC4, representative secreted mucin, MUC6, **Fig. 5.4C and D**). No prediction was made for MUC2 tandem repeat sequence by the NetNGlyc tool due to the absence of asparagine residues in the input sequence, as indicated by (*) in **Fig. 5.4C**.

Figure 5. 3 Prediction of Molecular Recognition Features (MoRFs) within Intrinsically Disordered Regions (IDRs) in mucins.

MoRFs are disorder-to-order (order upon binding) recognition motifs that influence and participate in protein-protein interactions (PPIs). Considering this determining the prevalence and location of such motifs would improve our understanding of mucin function and interaction. We found mucins contain many large-sized MoRFs that have a greater propensity to affect PPIs. Bar graph of the number of MoRFs normalized to mucin length by divided MoRFs with the number of residues for each mucin (available lengths in D^2P^2). Interestingly, MUC4 and MUC16, two transmembrane mucins that are differentially expressed in multiple malignancies, have a high MoRF/length ratio

Figure 5.3

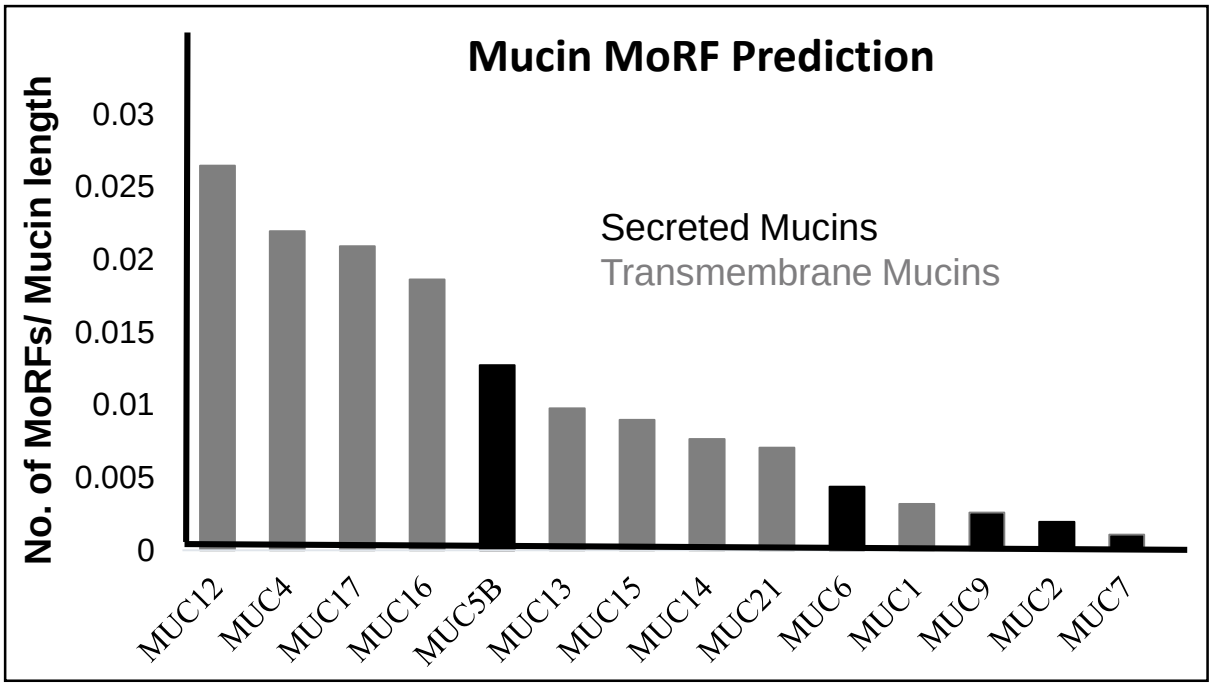
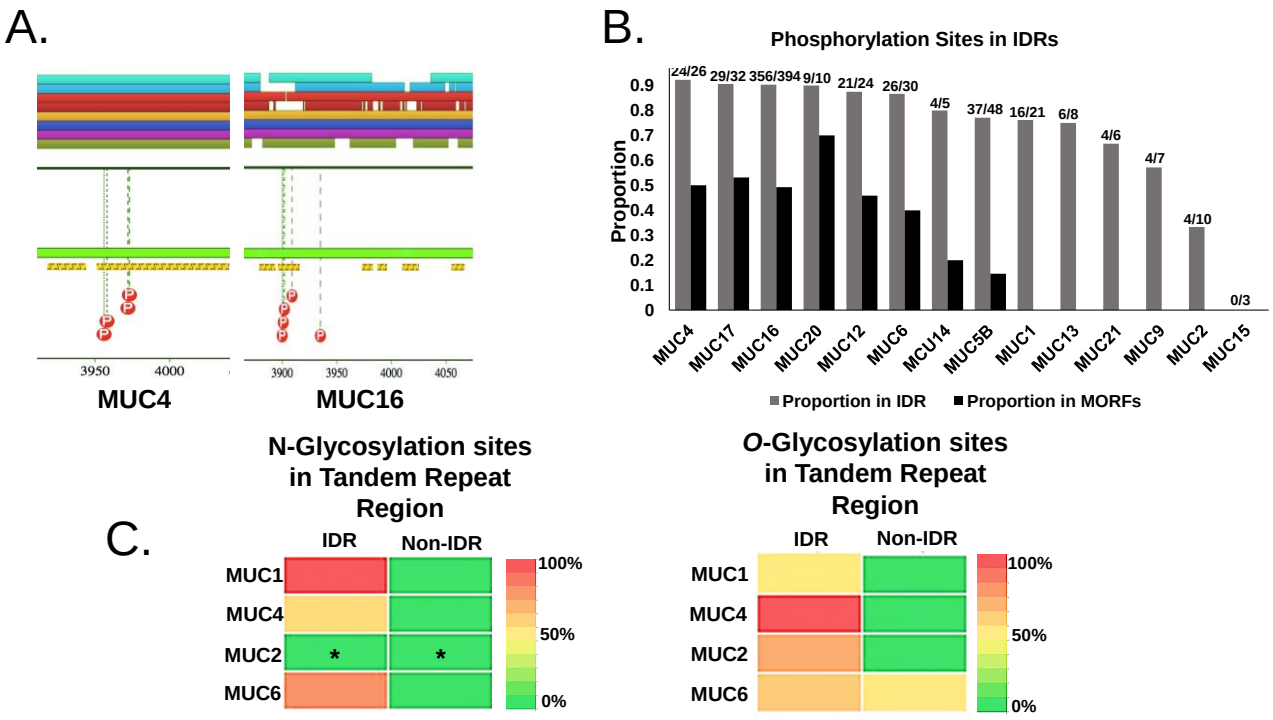

Mucin MoRF Prediction
No. of MoRFs/ Mucin length
0.03
0.025
0.02
0.015
0.01
0.005
0
Secreted Mucins
Transmembrane Mucins
MUC12
MUC4
MUC17
MUC16
MUC5B
MUC13
MUC15
MUC14
MUC21
MUC6
MUC1
MUC9
MUC2
MUC7

Figure 5. 4 Association of Post Translational Modifications, namely phosphorylation, and glycosylation, with IDRs and MoRFs in mucins.

Disordered regions are shown to be amenable to various forms of post-translational modification such as phosphorylation and glycosylation. **A.** Pictorial representation of phosphorylation sites found in D^2P^2 along with the IDR and MoRF predictions. MoRFs observed in transmembrane mucin MUC4 and MUC16 determined using ANCHOR as a part of D^2P^2. The yellow and black bar represents IDR-associated MoRFs predicted by ANCHOR (a tool that determines sequence motifs within an IDR that have a decrease in free energy upon binding with another protein). Curated phosphorylation sites (PhosphoSitePlus®) are displayed by red dots with "P" inside. **B.** Bar graph showing the proportion of curated phosphorylation sites found in D^2P^2 that are inside regions of predicted disorder (grey bars) as well as in regions predicted as MoRFs (black bars). **C.** Heatmap representing percentage occurrence of predicted *N*-glycosylation sites within IDR and Non-IDR across the VNTR domain of representative transmembrane mucins, MUC1 and MUC4, and representative secreted mucins, MUC2 and MUC6. The analysis was conducted by NetNGlyc 1.0 server. N-glycosylation occurs almost exclusively in IDRs as compared to non-IDR regions within the tandem repeat domain for MUC1, MUC4, and MUC6. **D.** Heatmap representing percentage occurrence of *O*-glycosylation sites within IDR and Non-IDR across tandem repeat domain. Representative transmembrane mucins, MUC1 and MUC4, representative secreted mucin, MUC2 and MUC6 were analyzed using NetOGlyc 4.0 server. *O*-glycosylation occurs almost exclusively in IDRs compared to Non-IDR regions within the tandem repeat domains of mucins.

Figure 5.4


A.
MUC4
MUC16
B.
Phosphorylation Sites in IDRs
Proportion
24/26
29/32
356/394
9/10
21/24
26/30
4/5
37/48
16/21
6/8
4/6
4/7
4/10
0/3
MUC4
MUC17
MUC16
MUC20
MUC12
MUC6
MUC14
MUC5B
MUC1
MUC13
MUC21
MUC9
MUC2
MUC15
Proportion in IDR
Proportion in MORFs
C.
N-Glycosylation sites in Tandem Repeat Region
IDR
Non-IDR
MUC1
MUC4
MUC2
MUC6
100%
50%
0%
O-Glycosylation sites in Tandem Repeat Region
IDR
Non-IDR
MUC1
MUC4
MUC2
MUC6
100%
50%
0%

5.4.5 Assessment of IDR Conservation across Mucins

To evaluate the potential functional significance of IDRs in mucins, we examined the evolutionarily conserved regions between mouse and human. Interestingly, MUC4 and MUC16, each with important implications in oncogenic development, were found to have similar patterns of disorder across human and mouse (**Fig. 5.5A and B**). *N*-terminal residues in the protein sequences of MUC4 and MUC16, in both human and mouse, have a consistently high degree of disorder, while the C-terminal residues fluctuated between order and disorder.

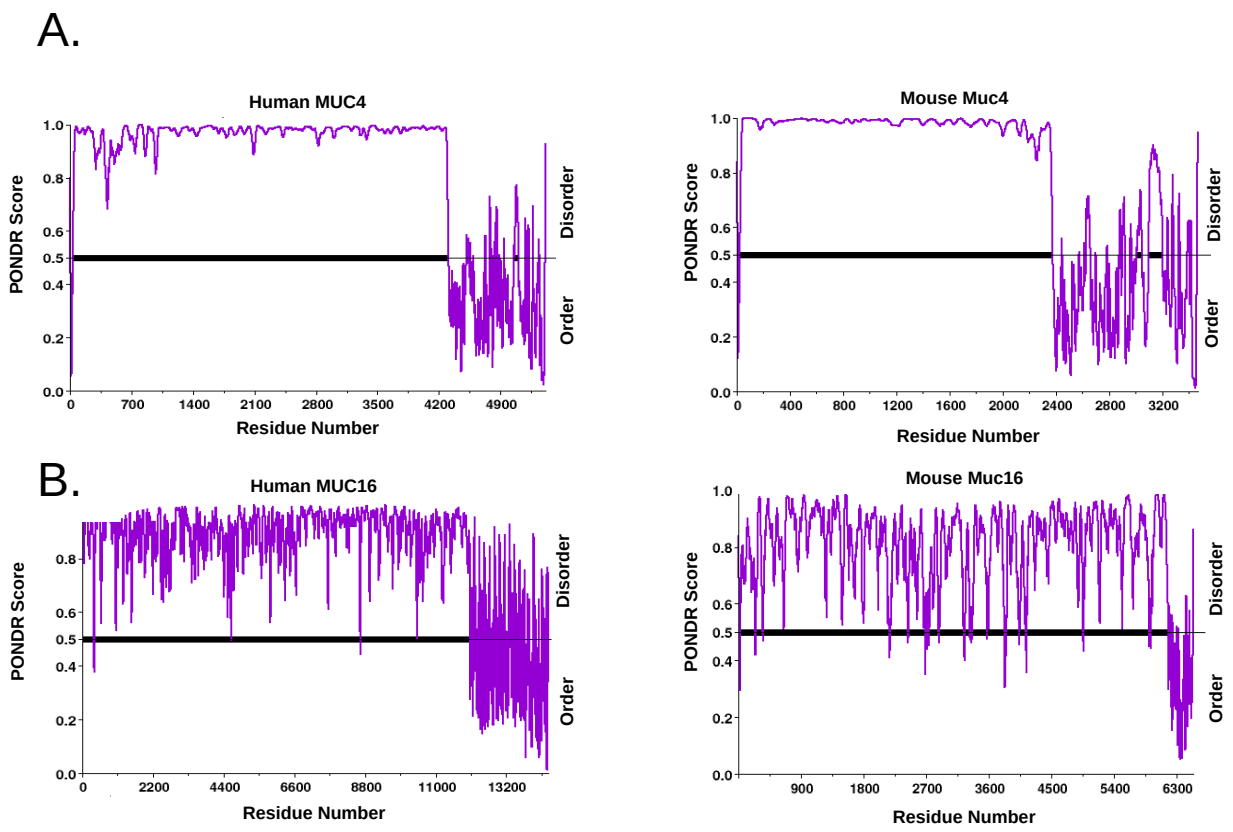
5.4.6 Mucin interactomes

Disorder allows for rapid on/off binding kinetics with other proteins because of high specificity yet low affinity for their partners, frequently observed in hub and signaling proteins (*144, 152-155*). Considering this attribute of IDRs, we asked if mucins with high predicted disorder can interact with multiple partners or occupy hub positions. Using BioGRID, interacting partners of MUC1, MUC2, MUC3A, MUC5B, MUC7, MUC9 (OVGP1), MUC12, MUC13, MUC14, MUC15, MUC16, and MUC20 were retrieved. No interaction partners were found for MUC3B, MUC4, MUC6, MUC8, MUC10, MUC17, MUC19, MUC21 and MUC22. Interaction partners for MUC4 and MUC17 were identified from a literature search.

Figure 5. 5 Intrinsic disorder patterns in human and mouse MUC4 and MUC16.

Evolutionary conservation of protein structure/sequence highlights the preservation of necessary biological functions. Though the actual interspecies sequence homology for both of these proteins is minimal, we speculated that the pattern of disorder may be conserved. We analyzed the IDR homology between mouse and human MUC4 and MUC16. Predictors of Natural Disordered Regions (PONDR) is an online compilation of five artificial intelligence tools that utilize previously defined structures for predicting intrinsic disorder. Due to the lack of mice full-length sequences in D^2P^2, PONDR was chosen for an inter-species comparison of disorder. Mice and human sequences (longest transcripts) were assessed with VSL2, a tool within PONDR, which makes a length-dependent prediction of protein intrinsic disorder to facilitate inter-species comparisons. **A.** Disorder prediction in human MUC4 and mouse MUC4 by PONDR VSL2 (a tool that predicts disorder and addresses protein length bias). **B.** PONDR VSL2 disorder prediction across human MUC16 and mouse MUC16. Residue values above 0.5 are predicted to be disordered. Both mucins displayed a significantly high degree of interspecies IDR pattern conservation

Figure 5.5


A.
Human MUC4
PONDR Score
Disorder
Order
Residue Number
Mouse Muc4
PONDR Score
Disorder
Order
Residue Number
B.
Human MUC16
PONDR Score
Disorder
Order
Residue Number
Mouse Muc16
PONDR Score
Disorder
Order
Residue Number

We next associated the number of interacting partners of mucins with the percentage of disorder present within the mucins. We observed that most transmembrane mucins including MUC1, MUC16, MUC13, and MUC20, which tend to have more IDRs and the longest MoRFs, had a higher number of interacting partners (**Fig. 5.6A and Fig.5.7**). MUC20, which is highly disordered at 79%, has interactions with 24 other proteins involved in various pathways. Similarly, MUC16 (63% disorder) has 10 interacting partners (**Fig. 5.6A and Fig.5.7**) and 46 MoRFs. However, transmembrane mucin MUC12, the most disordered protein in our analysis at 89%, has only two interacting partners. This is likely due to few studies on MUC12, and a dearth of knowledge regarding its interactome.

For secreted mucins, the most disordered family members, MUC7 (80%) and MUC5B (48%), had a greater number of interactions when compared to the other secreted mucins with lower levels of predicted disorder, MUC9 (25%) and MUC2 (23%). Unfortunately, in the case of other membrane-bound mucins, due to lack of information on their interactome, it was difficult to discern an accurate overall representation.

5.4.7 Functional Diversity of Mucins and Their Interactome

We next explored the functional significance of the mucin interactome. Reactome pathway analysis of the entire mucin family revealed significant involvement in a variety of important mechanisms including immune function, protein metabolism and signal transduction (**Fig. 5.9A**). Additionally, the mucin interacting partners generated from the BioGRID database were subsequently analyzed for their

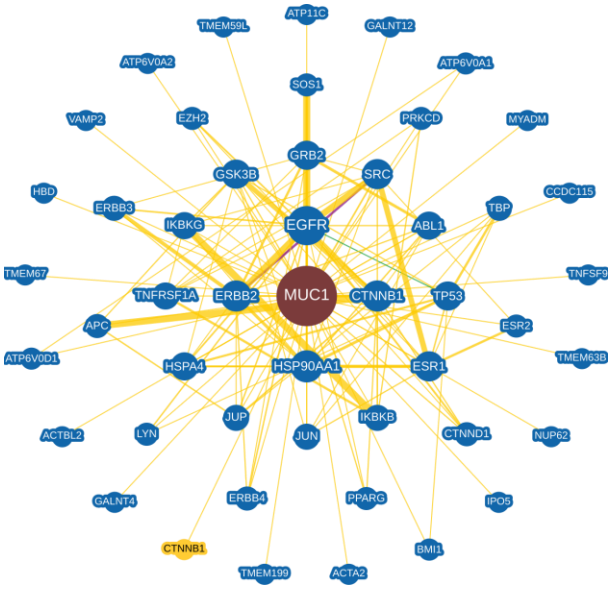
contribution to functional pathways. These interactome members are involved in a variety of functions associated with cancer including response to antineoplastic agents, cell migration, ERBB2 signaling, cell adhesion, and protein glycosylation (Fig. 7b). These pathways are directly involved in many aspects of oncogenesis, invasion, metastasis, and response to treatment. As mentioned, mucins have been shown to impact these cancer-associated pathways, further corroborating our findings.

Figure 5. 6 Transmembrane mucin MUC1 and secretory mucin MUC7 interacting partners determined using the BioGRID database

To assess if the quantity and prevalence of intrinsic disorder affect mucin interactomes, we utilized BioGrid, an online repository of protein chemical and genetic interactions. Physical interactions of all mucins were assessed using BioGrid to identify the interactions and the functional implications thereof. **A.** The network of the physical interactions of representative transmembrane mucin MUC1. **B.** The network of the physical interactions of representative secreted mucin, MUC7. Each Mucin is in the center of each interactome. The edge thickness connecting the mucin with its partners is linked to the number of times the interaction has been experimentally verified. Interactions with proteins not observed in humans, but in other organisms, are indicated by a yellow nodes.

Figure 5.6

A.


ATP11C
TMEM59L
GALNT12
ATP6V0A2
ATP6V0A1
SOS1
VAMP2
EZH2
PRKCD
MYADM
GRB2
GSK3B
SRC
HBD
ERBB3
CCDC115
IKBKG
EGFR
TBP
ABL1
TMEM67
TNFSF9
TNFRSF1A
ERBB2
MUC1
CTNNB1
TP53
APC
ESR2
ATP6V0D1
TMEM63B
HSPA4
HSP90AA1
ESR1
JUP
IKBKB
ACTBL2
LYN
JUN
CTNND1
NUP62
GALNT4
ERBB4
PPARG
IPO5
CTNNB1
BMI1
TMEM199
ACTA2

B.

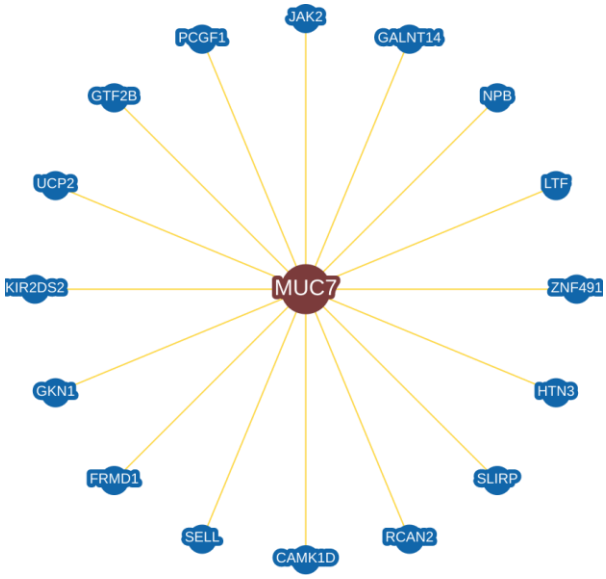

JAK2
PCGF1
GALNT14
GTF2B
NPB
UCP2
LTF
KIR2DS2
MUC7
ZNF491
GKN1
HTN3
FRMD1
SLIRP
SELL
CAMK1D
RCAN2

Figure 5. 7 Interacting partners of other mucins.

Mucins and their interacting partners determined using the BioGRID database. BioGrid an online repository of protein, chemical and genetic interactions. Physical interactions of all mucins were assessed using BioGrid to identify the interactions and the functional implications thereof. **.** The network of the physical interactions of transmembrane mucins, MUC12, MUC13, MUC14, MUC15, MUC16, and MUC20.

Figure 5.7

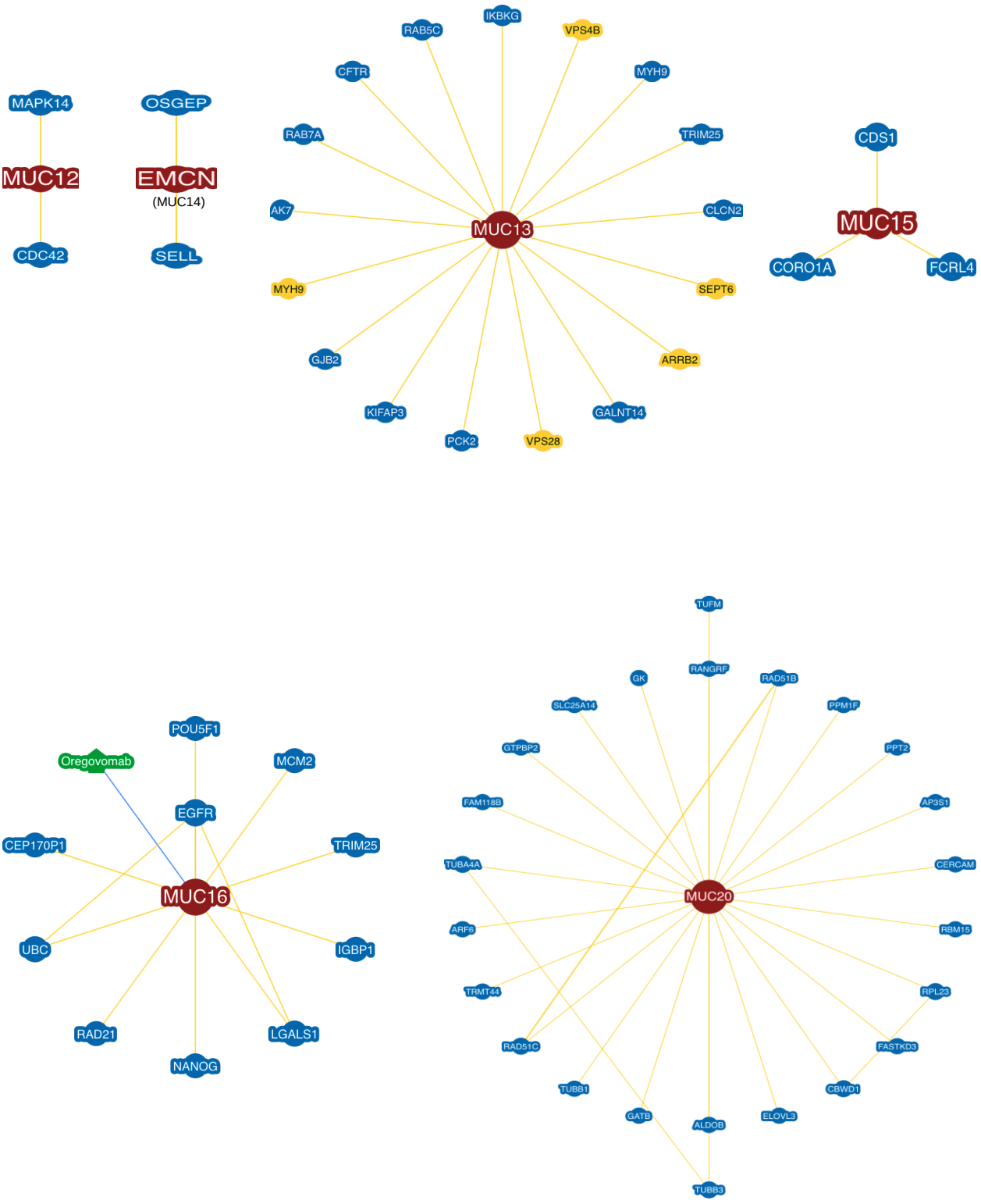
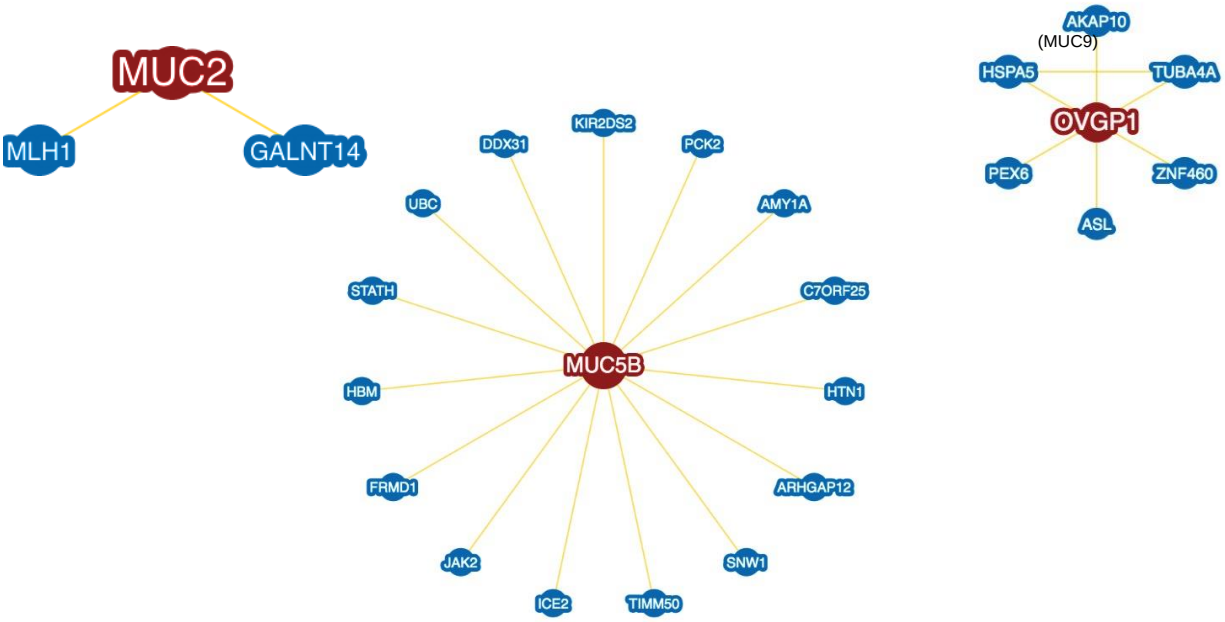

MAPK14
MUC12
CDC42
OSGEP
EMCN
(MUC14)
SELL
IKBKG
RAB5C
VPS4B
CFTR
MYH9
RAB7A
TRIM25
AK7
CLCN2
MUC13
MYH9
SEPT6
GJB2
ARRB2
KIFAP3
GALNT14
PCK2
VPS28
CDS1
MUC15
CORO1A
FCRL4
POU5F1
Oregovomab
MCM2
EGFR
CEP170P1
TRIM25
MUC16
UBC
IGBP1
RAD21
LGALS1
NANOG
TUFM
RANGRF
GK
RAD51B
SLC25A14
PPM1F
GTPBP2
PPT2
FAM118B
AP3S1
TUBA4A
CERCAM
MUC20
ARF6
RBM15
TRMT44
RPL23
RAD51C
FASTKD3
TUBB1
CBWD1
GATB
ALDOB
ELOVL3
TUBB3

Figure 5.8 Interacting partners for MUC2, MUC5B, and MUC9

The network of the physical interactions of secreted mucins, MUC2, MUC5B, and MUC9. Mucins are in the center of each interactome. The edge thickness connecting the mucin with its partners is linked to the number of times the interaction has been experimentally verified. Interactions with proteins not observed in humans, but in other organisms, are indicated by a yellow node

Figure 5.8


MUC2
MLH1
GALNT14
KIR2DS2
DDX31
PCK2
UBC
AMY1A
STATH
C7ORF25
MUC5B
HBM
HTN1
FRMD1
ARHGAP12
JAK2
SNW1
ICE2
TIMM50
AKAP10
(MUC9)
HSPA5
TUBA4A
OVGP1
PEX6
ZNF460
ASL

5.5 DISCUSSION

The hypothesis was supported by the predicted prevalence of intrinsic disorder across the entire mucin family. The presence of IDRs within the functional domains, between domains, extracellular, and cytoplasmic tail regions were analyzed. We observed that all mucins were predicted to have high (>40%) to moderate levels (>20% and <40%) of disorder. Indeed, 11 out of 15 of the mucins assessed were >40% disordered. Transmembrane mucins were more disordered compared to secreted mucins with the exception of MUC7. The average predicted disorder across all assessed mucins (58%) far exceeds the average of what is present throughout the human (30%) and eukaryotic (32%) proteomes (149-151, 195). In fact, this would place the mucin family in the top 10-15% most disordered proteins found in the human Ensemble database analyzed by D2P2 (149) with the majority of individual mucins harboring a far greater amount of disorder.

Apart from the nine predictors included in D2P2, we assessed mucin disorder with FoldIndex and PONDR CH plots, which are tools based on the dual assumption that IDPs/IDRs are generally enriched in polar and charged residues and depleted in the hydrophobic regions of proteins (141, 178). The mucin sequences used for the disorder analysis in D2P2 were used for FoldIndex disorder. This method predicted mucins to be far more ordered as compared to the D2P2 consensus results (**Fig.5.10A**). Confirmation with PONDR CH plots corroborates the FoldIndex findings and predicts native folding considering sequence charge and net hydrophobicity (**Fig. 5.10B**).

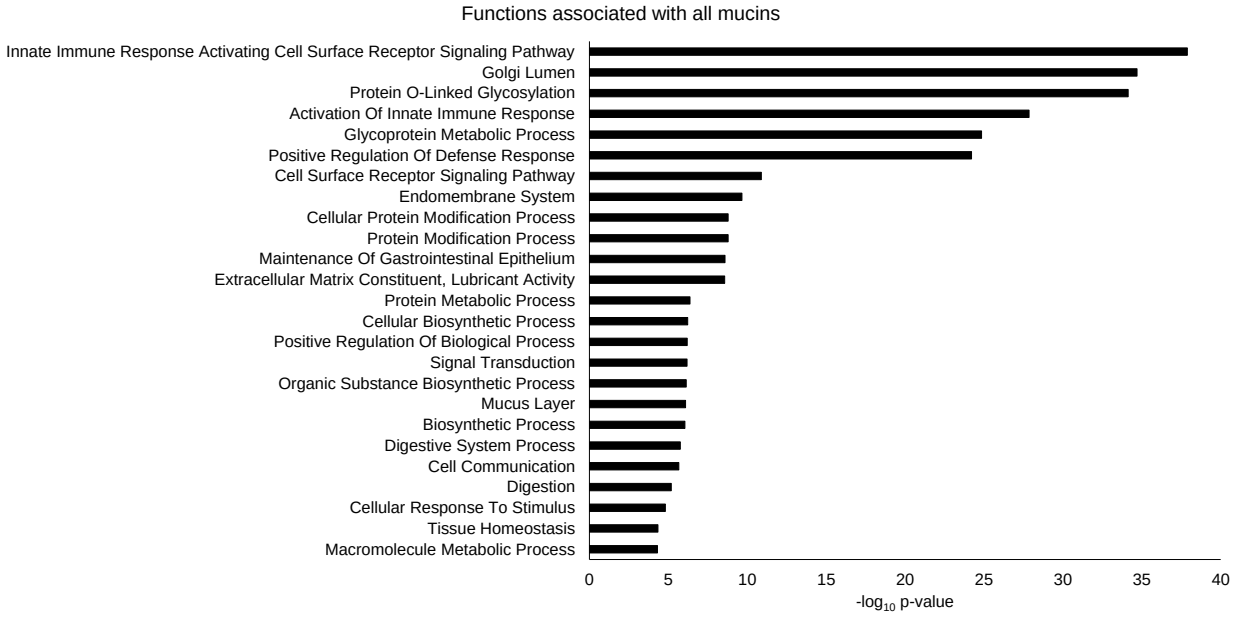
Though these findings seem incongruous with the other prediction methods, we postulate that charge hydropathy is not the best method to predict disorder in mucins for two reasons. First, FoldIndex has been shown to be the least accurate predictor of transmembrane protein disorder (196), and secondly, high number, variability, and degeneration of the tandem repeat sequences present in mucins are not accurately characterized leading to falsely low disorder prediction.

Correlations between the organization and evolution of chromosomes and chromosomal gene congregation with the extent of disorder have previously been evaluated. A study by Rajagopalan et al. found that cancer/testis antigens, a family of proteins often aberrantly expressed in cancer, are highly disordered (168). Further, they found that CTAs located on the X-chromosome (CT-Xs) displayed the largest extent of disorder compared to the family members located on other chromosomes (168). To help bolster the disorder prediction for mucins, we assessed if any correlation existed between their degree of disorder and chromosomal location. Interestingly, MUC4 and MUC20 with a similar percentage of disorder are located at the same 3q29 locus. Also, MUC12 and MUC17, which are predicted as the most disordered transmembrane mucins, cluster at 7q22 locus. Among secreted mucins, MUC5B and MUC6 predicted as almost equally disordered, are located at 11p15.5 locus.

Figure 5. 9 Functional annotation of mucins and their interacting partners.
To further assess the functional implications of the previously identified interacting partners a gene ontology (GO) based functional enrichment analysis was carried out. GO enrichment analyses are used to assess the functional implications of a gene or set of genes. The mucins and the interacting partners were compared to the GO database to assess the enriched terms and the enriched pathways assessed more closely. **A.** Graphical representation of GO Reactome pathway analysis of the mucin family (false discovery rate (FDR) < 0.05). The bar length is indicative of $-\log_{10}$ p-value. **B.** A bubble plot depicting the functions of the mucin interacting partners grouped into 15 neighborhoods. Each circle represents a GO pathway term to which mucin interacting partners contribute. The size of each circle is representative of the enrichment score of each pathway. Each numbered circle cluster (0-14) is demarcated into neighborhoods by color. Within each neighborhood is a circle highlighted in red, labeled with a pathway term. This shows mucin partners are involved in a variety of functions associated with cancer including response to antineoplastic agents, cell migration, ERBB2 signaling, cell adhesion, and protein glycosylation

Figure 5.9

A.


Functions associated with all mucins
Innate Immune Response Activating Cell Surface Receptor Signaling Pathway
Golgi Lumen
Protein O-Linked Glycosylation
Activation Of Innate Immune Response
Glycoprotein Metabolic Process
Positive Regulation Of Defense Response
Cell Surface Receptor Signaling Pathway
Endomembrane System
Cellular Protein Modification Process
Protein Modification Process
Maintenance Of Gastrointestinal Epithelium
Extracellular Matrix Constituent, Lubricant Activity
Protein Metabolic Process
Cellular Biosynthetic Process
Positive Regulation Of Biological Process
Signal Transduction
Organic Substance Biosynthetic Process
Mucus Layer
Biosynthetic Process
Digestive System Process
Cell Communication
Digestion
Cellular Response To Stimulus
Tissue Homeostasis
Macromolecule Metabolic Process
0
5
10
15
20
25
30
35
40
-log10 p-value

B.

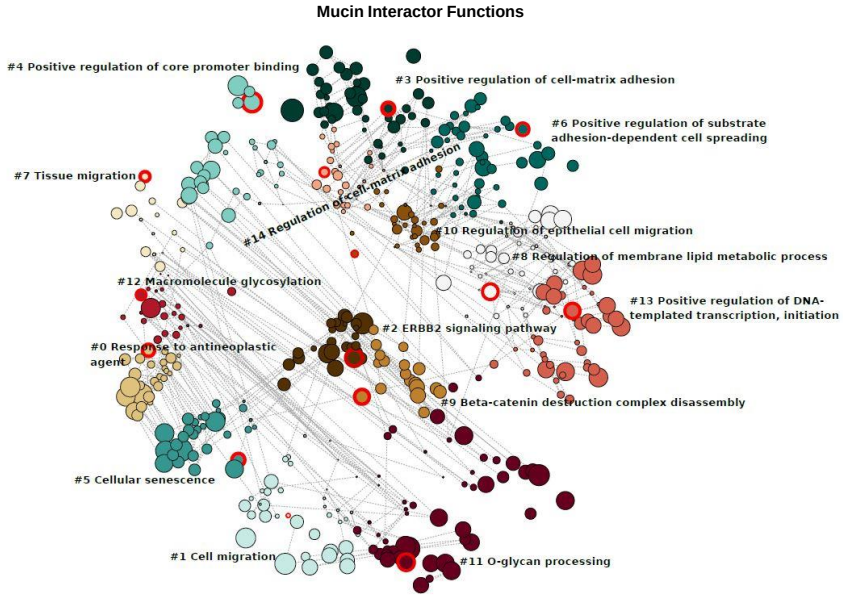
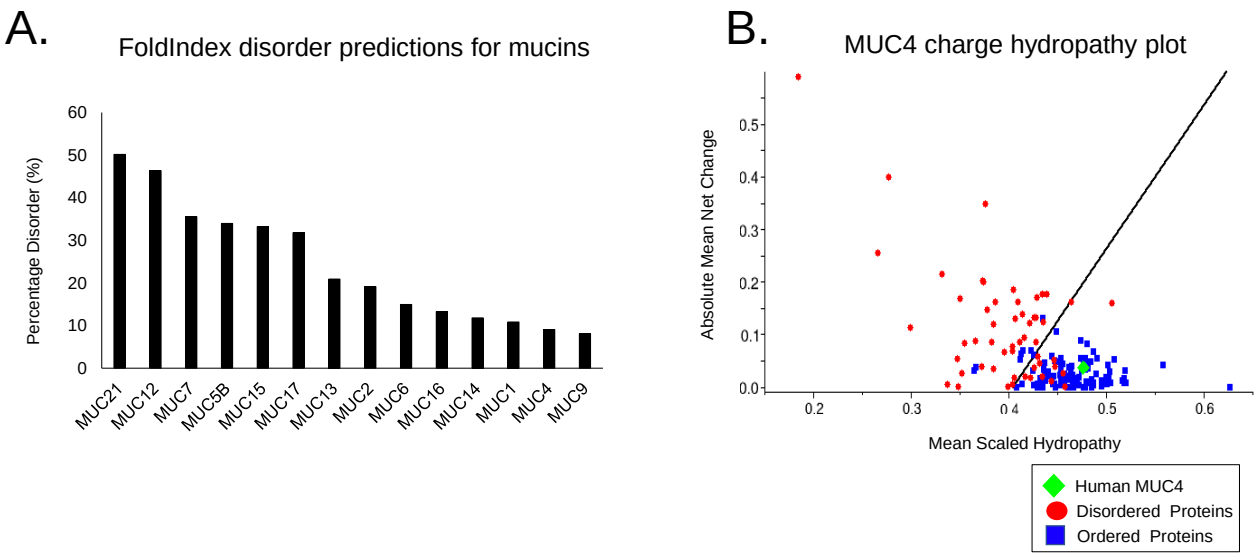

Mucin Interactor Functions
#4 Positive regulation of core promoter binding
#3 Positive regulation of cell-matrix adhesion
#6 Positive regulation of substrate adhesion-dependent cell spreading
#7 Tissue migration
#14 Regulation of cell-matrix adhesion
#10 Regulation of epithelial cell migration
#8 Regulation of membrane lipid metabolic process
#12 Macromolecule glycosylation
#13 Positive regulation of DNA-templated transcription, initiation
#2 ERBB2 signaling pathway
#0 Response to antineoplastic agent
#9 Beta-catenin destruction complex disassembly
#5 Cellular senescence
#1 Cell migration
#11 O-glycan processing

Figure 5. 10 FoldIndex based assessment of mucin disorder

Percentage of intrinsic disorder across all mucins determined using FoldIndex. The percentage of intrinsic disorder is determined relative to the total length of the protein. Among transmembrane and secreted mucins, the highest percentage of disorder was observed in MUC21 and MUC12, respectively. **A.** Bar graph displaying the percentage of intrinsic disorder across transmembrane (green) and secreted (red) mucins. **B.** Charge hydropathy chart showing verification of the in-house FoldIndex tool for MUC4. FoldIndex scores are based on sequence charge and net hydrophobicity, which predicts MUC4 to harbor little disorder

Figure 5.10


A.
FoldIndex disorder predictions for mucins
Percentage Disorder (%)
60
50
40
30
20
10
0
MUC21
MUC12
MUC7
MUC5B
MUC15
MUC17
MUC13
MUC2
MUC6
MUC16
MUC14
MUC1
MUC4
MUC9
B.
MUC4 charge hydropathy plot
Absolute Mean Net Charge
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.3
0.4
0.5
0.6
Mean Scaled Hydropathy
Human MUC4
Disordered Proteins
Ordered Proteins

Conserved regions of disorder contribute to myriad biological activities (197). These activities have been categorized into six functional classes by Tompa et al (198) and highlighted in comprehensive review articles on disordered proteins by the Uversky and Babu groups (141, 149). Entropic chain classifiers (does not acquire ordered confirmation for their functioning) is the first class where IDRs can act as flexible inter-domain linkers and spacers necessary for appropriate functional-domain activity effectors, where IDR act to modulate (inhibit or activate) interaction partner activity. A third class is assembler functioning when IDRs facilitate and provide scaffolding for large multi-protein complexes including signaling complexes. They can also have scavenger functions, where IDPs/IDRs interact with small ligands and capture, neutralize, or store them for later release. Chaperone class disordered proteins facilitate the folding of various molecules into their functional conformation. Finally, IDRs also harbor display site functions, where they provide conformational flexibility allowing PTM enzymes access to the protein backbone, thus facilitating their action including phosphorylation and glycosylation. Based on our data, we speculate that many of these functional classifications of IDRs will hold true for IDR present within mucins.

Many of these attributed IDR functions overlap with mucin activities. For example, MUC15 is moderately disordered (25%). MUC15/EGFR interaction is shown to diminish the aggressiveness of hepatocellular carcinoma by preventing EGFR dimerization (*199*). EGFR dimerization promotes the loss of intrinsic disorder in its kinase domain (disorder-to-order transition) leading to an increase in kinase signaling activity and the presence of an L834R mutation facilitates this

dimerization by suppression of disorder within the kinase domain (*170*). MUC15/EGFR interaction and subsequent inhibition of EGFR dimerization may be facilitated by an effector function of these disordered sequences present in both proteins or even MUC15 prevention of EGFR kinase domain loss of disorder. Another mucin with effector type functioning is MUC4. It has been shown that MUC4 interacts and stabilizes the receptor tyrosine kinase HER2 in the setting of pancreatic cancer, thus promoting cell proliferation (*200*). Interestingly, our analysis showed that MUC4 is over 76% disordered. Given the impact EGFR and HER2 signaling have in many cancers, the precise mechanism of these interactions, their effect on disorder, and the respective inhibition or activation of kinase capability warrants further study.

Other functional contributions of mucin IDRs could exist. For example, salivary proline-rich glycoproteins (much like mucins) contain high levels of IDRs and have been shown to have scavenger functions and bind small ligands such as ions and organic compounds either for disposal, sequestration, or later release (*201*) Congruently, our results show salivary MUC7 contains the highest percentage of disorder among all secreted mucins, at 80% predicted disorder. With this, many IDPs, or ordered proteins rich in IDRs, can form proteinaceous membrane-less organelles (PMLOs) *via* a liquid-liquid phase separation (*202*). Given the prevalence of intrinsic disorder in mucins, it is conceivable that mucins could contribute to PMLO formation and scavenger functioning within the local tumor milieu trapping nutrients, growth factors, and various other cytokines, thus forming a synergistic oncogenic environment. Another possible function of mucin IDRs

arises in light of the observation that viruses with a greater number of IDRs in their coat proteins are able to evade memory T-cells (*203-205*). Mucin polymerization and bulky glycosylation could also form an immunomodulating “glycoblanket” that shields cancer cells from detection and killing by leukocytes, thereby facilitating the unchecked growth of the disease. Along with this, our results show that IDRs are present in the cytoplasmic tail of transmembrane mucins as well as the extracellular region. Many oncogenic molecules involved in signaling are enriched in IDRs in their cytoplasmic tail (*206, 207*), thus, their existence in mucin CTs could impact mucin activity.

Due to the inherent ability to engage in promiscuous interactions, and the ability of rapid on/off binding, IDP/IDR ensembles are associated with dosage sensitivity. The higher the protein concentration, the larger the interaction pool. This, in turn, can lead to a dose-dependent non-specific response (*208, 209*). In conjunction, proteins with the most disorder are associated with hub positions in cancer-associated protein-protein interaction networks (*143, 168*). The combination of these two IDR aspects may explain why mucins can bind with and activate a great number of surface receptors (*210*), signaling molecules (*129*), and transcription factors (*211, 212*).

It is known that protein-protein interactions involving IDPs are influenced by molecular recognition features (MoRFs)(*189*). We found mucins contain higher numbers and many large-sized MoRFs, suggesting that they may participate in a variety of mucin interactions including the aforementioned effector activities of MUC15 and MUC4. The presence of MoRFs and their ability to undergo rapid

binding events further insinuates mucins in a myriad of cellular functions including cell signaling as well as rheostat (on/off) functioning. For example, MUC16 has the highest number of MoRFs that are >30 residues in length (413 and 46, respectively). Consistently, MUC16 has multiple interaction partners (*213-216*). MUC1 has the highest number of interaction partners as it is the best characterized of all mucins and the most studied. While MUC1 does not contain a large number of MoRFs, it contains the largest found in all mucins, at 214 residues in length.

MUC12 was predicted to have a large number of MoRFs with 145 but has only two interacting partners. This could be due to the fact that few studies have characterized MUC12 interactions. However, both of the MUC12 interaction partners (MAPK14 and CDC42) are involved in MAPK signaling and cell division, indicating that MUC12 harbors certain motifs/IDRs that contribute to oncogenic signaling and may impact cancer progression.

Though IDRs have been implicated in PTMs for years, a concept of an IDR-PTM-AS (alternative splicing) toolkit has recently been proposed (217). This toolkit allows for a single protein-coding gene to produce multiple disparate functional units, predicated in tissue or cell-specific manner. This also correlates with observed site-specific and context-dependent signaling of mucins under normal physiological as well as pathological conditions (e.g. MUC5AC deleterious effects in pancreatic cancer (218) and its protective role in the lung epithelium (219). Specific tissues or cell types are able to rewire/remodel protein pathways and gene expression patterns (via transcription factors) through changes in PTMs and alternative splicing, which are impacted by the presence, prevalence, and size of

IDRs (217). Mucins exhibit a high level of alternative splicing thus warranting further investigation into the effect of disorder on splicing events and impact on PTMs.

In addition to splice events, numerous structural modifications of mucins drive protein-protein interactions which serve as a key to their oncogenic role in cancer progression (220). The flexibility of IDRs facilitates access to enzymes involved in PTM (143) and the ability of an IDR to interact with target proteins is dramatically altered by the presence of these PTMs (217). Our findings show a high degree of overlap between the PhosphoSitePlus® curated phosphorylation sites found in D2P2 and IDRs, throughout the mucin protein family. The proportions of phosphorylation sites residing in IDRs are extremely high (ranging from 1.0 to .86) for each mucin with the lone exception of MUC9 (.5). This finding corroborates the amenability of mucin IDRs to PTM and further underscores the importance of these regions in signaling and interaction dynamics due to phosphorylation events.

Another PTM assessed in conjunction with disorder was glycosylation, specifically within the VNTR region. We found that predicted mucin glycosylation sites within the VNTR, overlap markedly with IDRs. A specific signaling motif (e.g., Proline-Threonine-Serine PTS sequence) could be working in combination with disorder, to facilitate glycosylation of this region. Disordered regions lying outside of the tandem repeats may not harbor this signaling motif allowing a variety of other PPIs, thus conferring the aforementioned hub protein characteristics of mucins. Alteration in the amount of disorder present within the VNTR including expression, mutations, and repeat expansion, could augment susceptibility to enzymatic

modification and dramatically affect the glycome, thus altering mucin interactomes. As aberrant glycosylation has long been a hallmark of cancer (134), understanding the amount and variety of disorder present within the system is incredibly important.

There is also overwhelming evidence that glycosylation is associated with protein stability (221). Contrarily, the presence and length of IDRs are negatively correlated with protein half-life, due to facile interaction with ubiquitin ligases (222). Thus, the balance between the presence of IDRs and their glycosylation status may act as a homeostatic mechanism to modulate mucin turnover and associated signaling pathways. Aberrant glycosylation of mucins, like what is observed in cancer, could alter their half-life and thus facilitate dose-dependent promiscuous binding and subsequent increases in pro-growth cellular signaling. Altogether, these observations indicate that IDRs may influence mucin protein-protein interactions as well as half-life. Future studies characterizing the complete mucin interactome and the mucin functional life-spans could enhance the associations reported here and further elucidate the relationship between IDRs, MoRFs, and PTMs.

Our assessments also show that IDR patterns are conserved between human and mouse, even though the number and order of tandem repeats within the PTS domain of mucins vary between species. We speculate that this IDR pattern conservation between human and mouse mucins is evidence that these regions preserve an important biological function despite underlying genetic variations and limited sequence homology. Additionally, IDRs may also explain the expansion of

the repeat sequence in human mucins compared to mice. Their lack of structural rigidity and the conferred decrease in evolutionary constraint can allow for replication slippage. If mucin function is predicated not only on their structure but also their ability to undergo plastic and dynamic structural changes, understanding of mucin IDRs becomes incredibly valuable for the assessment of their biological functioning and oncogenic implications.

IDRs are excellent cancer therapeutic targets for a variety of unique reasons. Disordered proteins can be sensitive to modulation through various methods and mechanisms of action. IDRs can be targeted directly via small molecules which can affect the affinity of the parent protein for binding partners, thus altering specific protein-protein interactions. Along with this, small molecule binding to IDRs can act through a variety of mechanisms including steric and/or allosteric hindrance, induced order upon binding, dimerization prevention, and conformation “locking” which decreases the dynamism of the protein. A specific unique advantage associated with IDRs is that since dynamism is key to their interactions, a small molecule that is able to diminish this dynamism, could have dramatic effects on the function throughout the entire region, regardless of where the binding occurs (i.e. not necessarily at the site of parent-partner interaction). Converse to this but equally as effective, small molecules or peptides can be used to target the IDR interactor proteins in (referred to as a “clamp”) and prevent the undesired PPI. Along with this, binding regions within IDRs can be predicted by utilizing MoRFs and computer-aided drug design to identify binding partners and

subsequently substituted for small molecules allowing for a facile and high-throughput method of discovery (223).

IDR therapeutic relevance is corroborated by the fact that many oncogenic molecules involved in signaling are enriched in IDRs. Importantly, IDRs have been confirmed in many cancer-associated proteins including p53, BRACA1, PAGE4, and PTEN (*204, 224-227*) and successful attempts have been made to target these regions. For example, a small molecule inhibitor was used to lock the normally dynamic IDR in the MYC protein in a static conformation that was unable to bind MAX, thereby preventing its oncogenic signaling (*228*). In another study, an alpha-helix-stapled peptide was engineered to interact with an IDR in P53, preventing its activation and subsequent anti-apoptotic effects (*229*).

Mucins, like the aforementioned proteins, are cancer-associated molecules that have eluded traditional therapeutic modalities. The development of compounds to target mucins is still in its infancy partly because detailed structures of this family are unavailable. We hypothesize that IDRs could serve as novel drug target sites for mucins, but this requires a detailed elucidation of their location and functional contributions. For instance, MUC16 cleavage and shedding of the EC domain is a major barrier to efficient MUC16 targeting in cancers (*230*). Where antibody therapy has failed, the IDRs present within the remaining membrane-bound MUC16 could be utilized as a means of targeting cancer cells. Alternatively, MUC16 has a cleavage site within the cytoplasmic tail region (*99*), and the sequence distal to the cleavage is predicted as completely disordered (6 out of 9 tools (66% consensus) in D^2P^2). When the intracellular cleaved portion of MUC16

is released, it increases cell proliferation, prevents apoptosis and influences the transcription of oncogenic genes (*99*). A cleaved MUC16 IDP would present a valuable therapeutic target for disruption of these functions and further investigation is warranted.

Another mucin that holds therapeutic relevance is MUC1. Monoclonal antibody (Mab) intervention attempts have been of limited success for MUC1 (*206*). In one study, Raina et al. were unsuccessful in attempts to crystalized the MUC1 CT and subsequent structural analysis with ROBETTA (*231*) and IGB-SSPro (*232*) revealed no identifiable secondary structure. Given these results, they determined that the MUC1 CT has features characteristic of an IDP. Notably, despite a lack of structure, IDPs are emerging as attractive drug targets (*233-238*). Further investigation into these MUC1 disordered regions is warranted which could provide insight into their relevance to its oncogenic signaling. In turn, this opens up a new possibility for therapeutic intervention *by* providing new targets for small molecule inhibitors or stapled peptides that can bind to MUC1 IDRs and inhibit its oncogenic function.

As mentioned prior, many studies are warranted to validate these *in-silico* findings as well as accurate attribution of IDR functions in mucins. Experimental characterization methods including (but not limited to) nuclear magnetic resonance (NMR) spectroscopy (including in-cell NMR), small-angle X-ray scattering (SAXS), 20S proteasomal degradation, and single-molecule fluorescence resonance energy transfer (smFRET) are necessary to determine the accuracy of the D^2P^2 disorder prediction. These techniques will help to elucidate the overall degree of disorder,

mucin structure and dynamics, and conformation changes upon partner interaction. In addition, experimental strategies to investigate the role of IDRs on mucin function are required as well. For example, selective mutations to residues that alter overall order in CT regions of MUC1 and MUC16 could be utilized to assess the effects of disorder on dimerization, proliferation, and/or oncogene transcription. Phage displays could also be utilized for the CT region of these mucins to determine what peptides bind and could be used as a therapeutic strategy. Another strategy to determine disorder effect on PPIs would be to use various isoforms of MUC4 (i.e. MUC4X, MiniMUC4, MUC4β, and WT MUC4) with different lengths of the tandem repeat regions (found to be highly disordered in our analysis) in pulldown assays or SPR based studies. Furthermore, disordered links between mucin domains could be deleted to determine if these have entropic chain characteristics and are required for adequate domain functioning.

The prevalence of IDRs within mucins could have vast clinical potential. Though we have utilized multiple prediction tools to determine the level of disorder, these computational findings must be validated experimentally. These studies would provide validation of the predictions and hypotheses presented herein, and justify a new and alternative perspective when assessing mucin structure.

Chapter 6: Connectivity Mapping-based Identification and Preclinical Evaluation of ISOX: A Novel Therapeutic for Pancreatic Cancer

6.1 SYNOPSIS

Pancreatic Cancer (PC) remains one of the most lethal malignancies due to high resistance to present-day chemotherapies. The expensive and burdensome process of drug discovery makes targeting this horrid disease highly challenging and novel ways for therapeutic targeting are urgently needed. Considering this, the present study utilizes a genomic-data-driven drug repurposing strategy based on Connectivity Mapping (CMAP). CMAP data was first mapped to gene expression from 106 PC patients identifying nine negatively connected drugs which were further narrowed down using a similar analysis for PC patient-derived-xenografts, human cell lines, and human tumoroids identifying ISOX as the most potential agent to target PC. Further, validation studies showed that ISOX exhibited strong anti-proliferative activity as well as anti-apoptotic activity (~48%) with a G0/G1 arrest in PC cell-lines. Notably, ISOX synergistically improved the efficacy of 5FU and Gemcitabine. At 500 nM ISOX, 60% loss in the viability of pancreatic tumoroids was observed. In orthotopic mouse studies, a 10-fold reduction in tumor weight was observed with ISOX alone (p-value=0.014) and ISOX-5FU combination (p-value=0.02) with significant survival difference (log-rank-p-value<0.05) and no metastasis in the combination drug treatment group. Further, RNA-sequencing and pathway analysis of the ISOX-LINCS signature, led to the identification of the MYC-MAX-MAD complex as an ISOX target which could also be corroborated in two independent transcription-factor analyses from the RNA-sequencing data. Finally, mechanistic studies established the acetylation-dependent regulation of

MYC action, suggesting HDAC dependent mechanism of action. Taken together, our data establish a novel approach of using CMAP for cancer therapy discovery and the potency of ISOX as a potential therapeutic.

6.2 BACKGROUND AND RATIONALE

Pancreatic Cancer (PC) maintains its unglorified third rank amongst all malignancy-related deaths in the United States and is projected to escalate to the second position by 2030 (*50*). Late diagnosis, early metastasis, and resistance to first-line therapies contribute to the poor prognosis of this horrid disease, with a dismal five-year survival rate of 10%. Furthermore, the Pancreatic Cancer Action Network (PanCAN) explicitly declares the lack of promising therapeutic modalities as a major confounder for this appalling survival. Sadly, therapeutic interventions in PC have been challenging for researchers and physicians alike. The current options, which include surgery, radiation therapy, chemotherapy, targeted therapy, and immunotherapy, have major limitations. While surgery is one curative option, only 20% of patients are rendered suitable for surgical interventions (*239*). Further, Gemcitabine (GEM) based chemotherapy regimens have been established as one of the most promising therapeutic modalities for PC patients' major limitations like toxicity, lack of specificity towards molecules specifically altered in PC, ineffectiveness in a subgroup of patients, and poor penetration due to hypo-vascularized dense PC stroma limit its use (*240, 241*)

While various combination chemotherapies have been tried, only a minor improvement has been observed in the median overall survival (MOS) of 4 to 11 months in PC patients (*242*). Though the initial response of FOLFIRINOX

(oxaliplatin, irinotecan, leucovorin, and 5FU) compared to GEM alone was dramatic with a 21.6 month MOS it was limited by the high toxicity in these patients (*243*). Across studies, FOLFIRINOX and Abraxane (ABX) in combination with GEM have only marginally improved the survival rates, 11 months and 9 months, respectively, (*244-246*) keeping lethality of PC at its top. Further, considering only 25% of patients respond to GEM and the high degree of toxicities associated with FOLFIRINOX, there is a compelling need for the identification of new potent therapeutics (*11*). While currently, there are 18 monotherapies and 4 combination therapies that the USFDA has approved for use in PC these approved therapeutic modalities fail to render their required clinical outcomes due to toxicity, off-target effects, and therapy resistance. Furthermore, this aggressive malignancy has an intrinsic resistance to chemotherapy (*247*), and better chemotherapeutic agents are highly needed.

The failure of the conventional one target at a time approach has created a need to assess and develop robust methods for the identification of novel drug targets. Following the sequencing of the human genome, genomic-driven approaches have gained importance in recent years (*248-250*) and can provide a great advantage in dissecting new targets. Various computational methods have been established to identify new therapeutics, including the BROAD institute tool, CMAP (https://www.broadinstitute.org/ cmap/).CMAP is a big repository of gene expression data compiled from the effects rendered or the transcription readout of treatment by various FDA-approved as well as preclinical small molecule inhibitors. Altered gene expression data is utilized to connect to a user-defined gene

signature to render positive, neutral, and negative connections. Effectively, a positive connection between the disease signature and the gene expression from a drug would mean that the similar signature got affected, i.e., upregulated and downregulated across drug and disease, a neutral connection simply means no connection, but most importantly, a negative connection would mean a reversal of the user defined gene expression by the drug. These negative connections would then be useful to researchers seeking to identify promising therapeutics which will reverse the gene expression from the biological disease of interest (*251, 252*).

The present study establishes the use of CMAP for drug identification in PC and validates its utility through *in-vitro* and *in-vivo* assessment of the identified small molecule, ISOX (N-[4-[3-[[[7-(hydroxyamino)-7-oxoheptyl] amino] carbonyl]-5-isoxazolyl] phenyl]-1, 1-dimethylethyl ester, carbamic acid) (**Fig.1A**). To do this, we first used four PC microarray datasets to establish “gene-signatures” for PC and, in turn, used these signatures to identify drugs that were commonly negatively connected across (score -30 or higher) all four datasets. The therapeutic effect of ISOX was tested across a spectrum of PC cell lines, including MiaPaCa2, CD18/HPAF, AsPC1, and BxPC3 using various functional assays such as MTT-based proliferation assay, cell cycle and apoptosis, invasion using a matrigel assisted Boyden-chamber, and migration analyses. Various doses of ISOX were further assessed in combination with established PC therapeutics, Gemcitabine and 5’FU. Furthermore, in order to mimic the tumor biology more closely, ISOX was tested in 3D tumoroids developed from KPC ($KRAS^{G12D}$; $p53^{R172H}$; Pdx-1-Cre) mice and human PDAC patient samples followed by a drug-efficacy assessment

in pancreatic orthotopic mice model. In order to decipher the drug mechanism of action, RNA-sequencing analysis was carried out using ISOX treated and untreated PC cells, followed by a comprehensive pathway analysis using ingenuity pathway analysis (IPA), a two-way transcription factor enrichment analysis using TFactS (http://www.tfacts.org/) and ENCODE. A combination of literature review and RNA-seq and western blot studies led to the identification of the HDAC inhibition-dependent inhibition of cMYC and its related pathways.

6.3 MATERIALS AND METHODS

6.3.1 Identification of datasets. Gene expression omnibus (GEO) datasets was queried for datasets containing PC and normal samples. The first step filter used were the keywords “tissue” and “homo sapiens” which led to the identification of over 245 datasets. Further, these datasets were filtered on a multifold criterion-the normal and tumor sample within the same dataset, no pre-treatment, and no inherent bias in sample selection which helped us identify four datasets GSE32676 (25 tumor 7 normal), GSE15471 (6 tumor 16 normal), GSE16515 (36 tumor 16 normal) and GSE18670 (6 tumor 6 normal). Further, GEO was queried to identify datasets with PC cell lines (GSE45757), tumor xenografts (GSE46385), and human patient-derived tumoroids (GSE107610) (*253-259*).

6.3.2 Identification of differentially regulated genes. The CEL files (Affymetrix raw data files) from each of the identified datasets was downloaded and processed using the “affy” (*46)* package from R Bioconductor (version 3.6). The expression

was assessed using a robust multi-array average (RMA) from the “affy” package. The array probes were converted to gene names using the hgu133plus2.db library. A linear model was fit (using limma) across individual datasets to identify the most differentially expressed genes in tumors in comparison to normal ones. The genes were then arranged according to $\log_2$ fold changes and the top 150 upregulated and downregulated genes were assessed for the CMAP analysis.

6.3.3 Determination of perturbagens targeting PDAC tissues. Top 150 upregulated and top 150 downregulated genes from individual datasets were queried into the connectivity map tool (https://clue.io/). Negatively connected drugs (connectivity score -30 and higher) were compared across datasets to identify the most common drugs across datasets. To evaluate the specificity of PC’s identified drug spectrum of PC, differentially expressed gene signature from GEO of human PC cell lines, PC patient derived xenografts, and PC tumoroids was queried for CMAP. The 9 commonly negatively connected drugs from the previous analysis were then compared to negatively connected drugs to each of these studies, and ISOX was identified as the single common drug and hence was chosen for further analysis.

6.3.4 Cell culture and reagents. Human pancreatic cell lines (AsPC1, MiaPaCa2, CD18/HPAF, and BxPC3) were obtained from ATCC and confirmed using short tandem repeat (STR) profiling routinely during the experiments. The cells were cultured in 10% FBS supplemented DMEM or RPMI medium supplemented with

glutamine and penicillin as suggested and maintained in a cell culture incubator at 5% CO_2 and 37°C. For drug studies, ISOX (CAY10603) was obtained from Cayman Chemicals (CAS number 1045792-66-2), while tubastatin A (CAS number 1252003-15-8) and riclinostat (ACY1215) were purchased from MedChemExpress.

6.3.5 Cell viability studies. Cell viability studies were carried out using MTT-assay. Previously cultured cells (AsPC1, MiaPaCa 2, CD18/HPAF, BxPC3) were plated at 5000 cells per well in 96 well plates. After overnight incubation, PC cells were cultured with varied doses of ISOX (10 nM, 100 nM, 1 mM, 10 mM, and 100 mM) to assess its impact on cellular viability. Treatments were carried out for 24, 48, and 72 hours following which the 100mL of 5 mg/mL MTT reagent was added to each of the wells and cells were incubated for 3 hours. At the end of the incubation period the cells were lysed using 100 mL/well DMSO (10 minutes) and read at 570 and 640 nm. A similar analysis was also carried out for other HDAC6 inhibitors; Riclinostat and Tubastatin A. Similarly, drug combination studies were carried out with fixed doses of Gemcitabine (2 µM) and 5-flurouracil (5 µM) in addition to varied doses of ISOX (0.156 µM, 0.3125 µM, 0.625 µM, 1.25 µM, 2.5 µM, 5 µM, and 10 µM) and the proliferation rates were calculated with respect to the untreated samples. **Apoptosis analysis.** Cells were plated in triplicates for both untreated and treated groups. Followed by a 12-hour incubation, treatment group cells were treated with 1 µM of ISOX. Following a 48-hour treatment cells were trypsinzed, collected and centrifuged and the pellet re-suspended in the

annexin V binding buffer. The cells were then stained with annexin V and PI and assessed using flow cytometry.

6.3.6 Cell cycle analysis. Cell cycle analysis was carried out using propidium iodide staining. Various PC cell lines (AsPC1, MiaPaca2, CD18/HPAF, and BxPC3) were first seeded in triplicates. To ensure a uniform assessment cell cycles, cell were synchronized using a 12-hour thymidine block followed by a 9-hour deoxycytidine treatment and a final 12-hour thymidine block. Following the synchronization, cells in the treatment group were treated with 1 mM ISOX for 48 hours. At the end of the treatment the cells were fixed in 70% ethanol, followed by a propidium iodide (with telford reagent) staining and flow cytometry assessment.

6.3.7 Migration analysis. Cells were seeded in triplicates for treated and untreated groups at $1x10^6$ per well in the top chamber of matrigel coated plates (Corning BioCoat Matrigel Invasion chambers) for 12 hours. Following the initial incubation, cells were treated with 1mM ISOX for 48 hours, and the bottom half of the chamber filled with serum-containing media for chemoattractant purposes. At the end of the 48-hour treatment, the migrated cells were stained using a quick-diff staining kit and compared in between groups.

6.3.8 Efficacy assessment in KPC and human tumoroids. Tumor tumoroids were established using tumors from KPC autochthonous mouse and human donor tissues by firstly enzymatically digesting with 0.012% (w/v) collagenase XI (Sigma)

and 0.012% (w/v) dispase (GIBCO) in DMEM media containing 1% FBS (GIBCO) and then embedding in growth factor reduced Matrigel (BD Biosciences). These tumoroids were maintained and cultured in complete DMEM/F12 medium supplemented with HEPES [Invitrogen], Glutamax [Invitrogen], penicillin/streptomycin [Invitrogen], B27 , Primocin [1 mg/ml, InvivoGen], N-acetyl-L-cysteine [1 mM, Sigma Aldrich], mouse recombinant Wnt3a [100ng/ml, EMD Milipore], human recombinant RSpondin1 [1µg/ml, PeproTech], Noggin[0.1 mg/ml, PeproTech], epidermal growth factor [EGF, 50 ng/ml, PeproTech], Gastrin [10 nM, Sigma], fibroblast growth factor 10 [FGF10, 100 ng/ml, PreproTech], Nicotinamide [10 mM, Sigma], and A83- 01 (0.5 mM, Tocris Biosciences).

6.3.9 Orthotopic mice model studies. All animal experiments were approved by the UNMC Institutional Animal Care and Use Committee. Luciferase labelled CD18/HPAF (viability >95%) were orthotopically implanted into the pancreas of athymic nude mice (male and female) at $2.5X10^5$ cells in 50 µL tissue culture grade PBS. The mice were then imaged using the small imaging IVIS system to monitor tumor formation. At the end of 2 weeks following implantation and confirmation of tumor formation, the mice were randomly distributed into four groups- control (PBS), ISOX (50 mg/mL), 5FU (50 mg/mL), and combination (ISOX and 5FU together). The treatment was carried out for 15 days (3 cycles of 5 days continuous followed by 2 days break) with imaging at day 10 and day 15. Half of the mice from every group were sacrificed, and the other half followed for survival. The tumor weight was then assessed for each mouse and compared using a Mann-Whitney

U test across groups. For the survival analysis, day of death was measured either as the natural death or the veterinarian suggested euthanasia. Survival across groups was compared using a log-rank test. The experiment was repeated twice.

6.3.10 RNA-sequencing analysis. RNA from treated (1 µM ISOX, 48 hours) and untreated CD18/HPAF were assessed using Illimuna TrueSeq (mid-output 75 paired-end) RNA sequencing (RNA seq). The raw reads were mapped to Hg38 human genome from iGenomes using TopHat (version v2.1.0), quantified using CuffLinks (version v2.2.1), and differential expression across the untreated and treated samples calculated using CuffDiff. The differentially expressed genes were subjected to a pathway assessment using ingenuity pathway analysis (IPA v01-12) and gene set enrichment analysis (GSEA). Furthermore, the RNA-seq data was used for a transcription factor analysis using the online tool TFactS (http://www.tfacts.org/) and a more detailed analysis using the ENCODE transcription factor tool within the web tool iLincs (http://www.ilincs.org/ilincs/)

6.3.11 Immunoblotting. Protein isolation and western blot analysis was carried out to compare untreated and ISOX treated (24, 42, 72 hours) cells to assess the difference in protein expression of targets identified through the *in-silico* screen. Proteins were isolated using

radio-immunoprecipitation assay (RIPA) buffer for lysing the cells, followed by removing the cell debris by centrifuging at 13,000 rpm at 4^{o}C. Protein

concentrations were measured DC Bio-Rad protein assay kit. Equal concentrations of protein were loaded for the untreated and treated samples.

6.3.12 Immunohistochemistry analysis. Slides from untreated and treated animal tissue sections were baked overnight at 56°C followed by deparaffinization using 2 xylene washes, followed by rehydration using varied concentrations of ethanol. Endogenous peroxidases were blocked using 3% H_2O_2 for 1 hour, followed by antigen retrieval in citrate buffer (pH 6) for 15 mins. The slides were then blocked using normal horse serum (Vector Laboratories) and incubated overnight with the various primary antibodies. Universal secondary antibodies (Vector Laboratories) were used for 1 hour, and the slides were developed using DAB substrate kit (Vector Laboratories). Hematoxylin was used for nuclear counterstain. Similarly, human patient tissues were stained for HDAC3, HDAC6, and HDAC10. Tissues were dehydrated and mounted using permount.

6.4 RESULTS

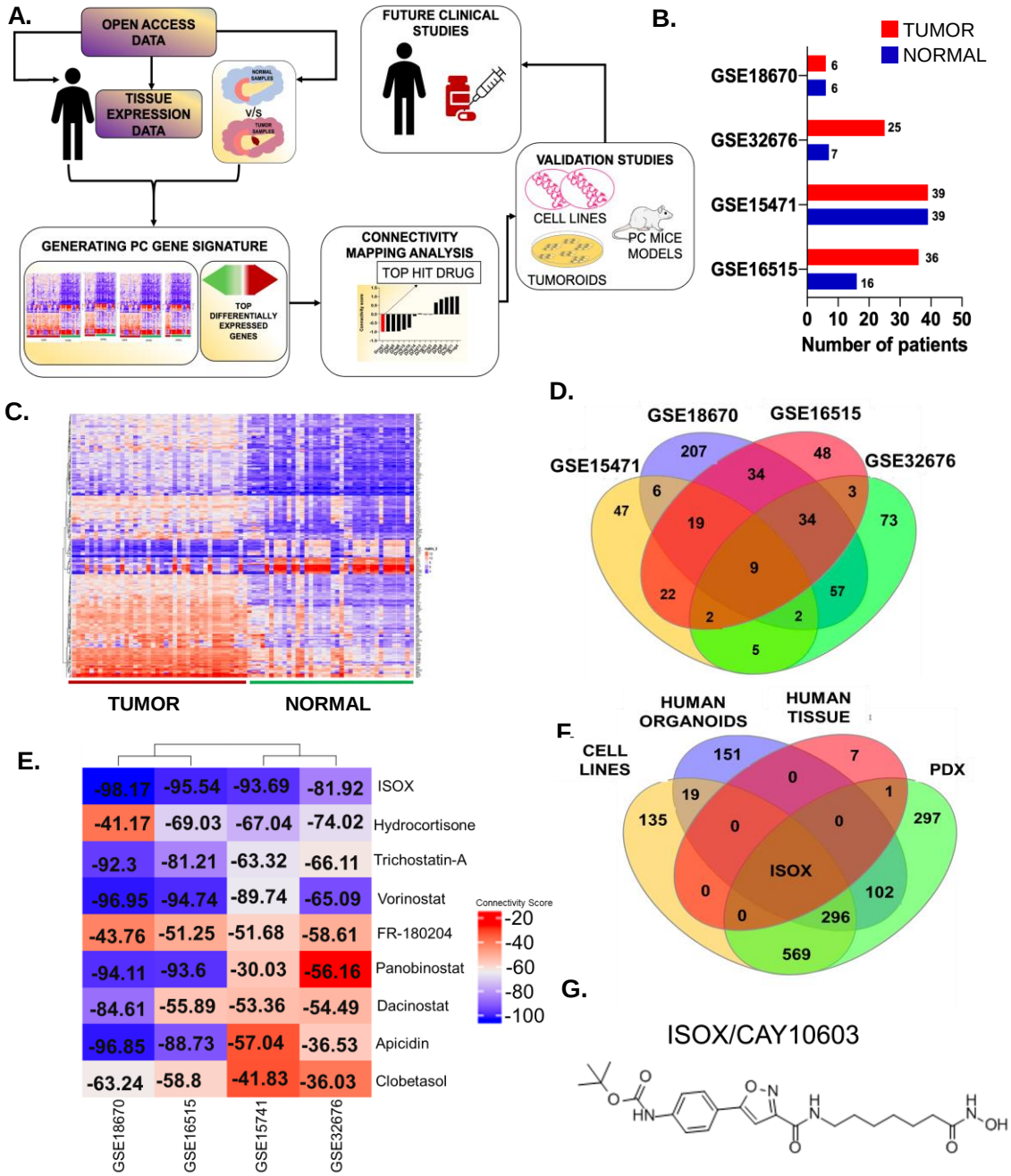
6.4.1 Connectivity mapping analysis identifies ISOX as a potential therapeutic for PC.

CMAP, a big data repository from the BROAD Institute, is a collection of gene expression data from cell lines treated with a huge variety of inhibitors. We used the CMAP database and queried it for the differential gene signatures from four (GSE16515, GSE15471, GSE18670, and GSE32676) PC datasets (**Fig. 6.1B, Fig. 6.1C**). We then assessed the drugs negatively connected to each of these PC signatures due to their propensity to reverse the disease state's gene signature. Nine common drugs with a significant negative score (-30 or above), including ISOX, Trichostatin A, Vorinostat, Apicidin, Panobinostat, Hydrocortiosone, Dacinostat, FR-180204, and Clobetasol were identified (**Fig. 6.1D**). Interestingly, ISOX was the top negative scoring drug across all datasets (**Fig. 6.1E**). However, the scoring showed a slight variability across the datasets, with drugs like panobinostat showing a high score in two of the datasets (-94.11 and -93.6) but lower in the other two. Considering this and to identify a highly specific for PC, we next carried out a CMAP analysis in datasets from PC cell lines (GSE45757), human tumoroids (GSE107610), and patient-derived xenografts (GSE46385, PDX). The negatively connected drugs across these datasets were compared to the nine commonly identified drugs (**Fig. 6.1F**). Among various drugs, ISOX (CAY10603, tert-butyl N-[4-[3-[[7-(hydroxylamino)-7-oxoheptyl] carbamoyl]-1, 2-oxazol-5-yl] phenyl] carbamate) was identified as the most potential therapeutic for PC (**Fig. 6.1G)**

Figure 6. 1 *In-silico* identification of highly specific therapeutic for pancreatic cancer.

Connectivity mapping was used to identify negatively connected drugs specific to four PC datasets. **A.** Schematic representation of overall study design followed to identify and validate specific therapeutic for PC. **B.** Sample cohort- Bar graph representation of a number of tumors and normal samples within the four microarray datasets GSE18670 (24 samples), GSE32676 (32 samples), GSE15471 (78 samples), GSE16515 (52 samples). A differential gene expression was carried out using limma package from R bioconductor to identify top differentially expressed genes between normal and tumor samples. **C.** Representative heatmap from differential gene expression in GSE15471. The top 150 up-regulated and down-regulated genes from the differential gene expression was put into connectivity map to identify negatively connected drugs for each of the datasets separately. **D.** Venn diagram representing negatively connected drugs across the four datasets. Nine common drugs were identified as being common between all the datasets. **E.** Heat map representing the connectivity scores of all the 9 commonly negatively connected drugs**. F.** Highly specific drug for PC across cell lines, human tumoroids, human tissue, and patient-derived xenografts (PDX). In order to delineate a highly specific drug connectivity mapping was run again for human tumoroids, cell lines, and PDX models. The negatively connected drugs were compared to the nine drugs identified from the human tissue samples. ISOX was identified as the only drug common between all these four models. **G.** ISOX structure as obtained for CMAP showing a clear HDACi moiety.

Figure 6.1


A.
OPEN ACCESS DATA
TISSUE EXPRESSION DATA
v/s
FUTURE CLINICAL STUDIES
VALIDATION STUDIES
CELL LINES
PC MICE MODELS
TUMOROIDS
GENERATING PC GENE SIGNATURE
TOP DIFFERENTIALLY EXPRESSED GENES
CONNECTIVITY MAPPING ANALYSIS
TOP HIT DRUG
B.
TUMOR
NORMAL
GSE18670
GSE32676
GSE15471
GSE16515
Number of patients
C.
TUMOR
NORMAL
D.
GSE18670
GSE16515
GSE15471
GSE32676
E.
ISOX
Hydrocortisone
Trichostatin-A
Vorinostat
FR-180204
Panobinostat
Dacinostat
Apicidin
Clobetasol
Connectivity Score
GSE18670
GSE16515
GSE15741
GSE32676
F.
HUMAN ORGANOIDS
HUMAN TISSUE
CELL LINES
PDX
ISOX
G.
ISOX/CAY10603

6.4.2 ISOX inhibits the proliferation of PC cell lines at low concentrations.

To validate our *in-silico* findings, ISOX's impact on tumor cell proliferation, migration, and apoptosis was evaluated in a panel of PC cell lines, including AsPC1, MiaPaCa, BxPC3, and CD18/HPAF. The cell lines were chosen based on the varied genetic background, differentiation status (covering well to poor) and tumor source (primary tumor, metastatic site and ascitic fluid). The impact of ISOX as a monotherapy was first evaluated on cellular proliferation via MTT method as described in our earlier publications (*260*). ISOX was found to be extremely potent in reducing the growth of all the cell lines with an IC_{50} value ranging 2.4 nM-1.4 µM (**Fig. 6.2A**). Interestingly, the varied response of ISOX was observed across the PC cell line panel, further suggesting the inherent heterogeneity and varied therapy response of the disease. Further, to determine the toxicity level of ISOX on normal pancreatic cells, MTT was carried out on normal human pancreatic immortalized cell line HPNE (*261*). ISOX had minimal effect on HPNE cells, wherein even at higher concentrations, the proportional viability was at 80% as opposed to a much higher toxicity of 5FU (**Fig. 6.2B**).

6.4.3 ISOX affects the cell cycle by inducing G0/S arrest of PC cells.

Next, to gain a better understanding of the mechanism of action for ISOX on cell cycle abrogation, we treated PC cells with ISOX and examined its effect on the various phases of the cell cycle. An initial synchronization followed by propidium iodide (PI) based FACS analysis was performed to assess the effect of ISOX on cell cycle arrest in PC cells. We observed that across our cell line panel and at

various time-points, ISOX induced a G0/S phase arrest. This change was most prominently observed at 48 hours after the treatment with a significant reduction in cells in S phase and a statistically (Mann-Whitney U p-value< 0.05) significant increase in the GO/S phase (**Fig. 6.2C**). Interestingly, MiaPaCa2 showed a G2/M arrest suggesting heterogeneity within the cell lines.

6.4.4 ISOX induces apoptosis in PC cells.

To further evaluate whether ISOX impacts tumor cell resistance to apoptosis, we performed apoptosis studies using Annexin V and PI *via* flow cytometry analysis. ISOX treated group showed a high percentage of both early apoptotic cells (annexin V positive and PI negative) and late apoptosis (annexin V and PI positive) across the cell line panel. Significant apoptosis induction was observed in all the PC cells examined. Notably, up to 40% apoptosis was observed in both BxPC3 and MiaPaCa2 cells (**Fig. 6.2D**).

6.4.5 ISOX reduces the invasion and migration abilities of PC cell.

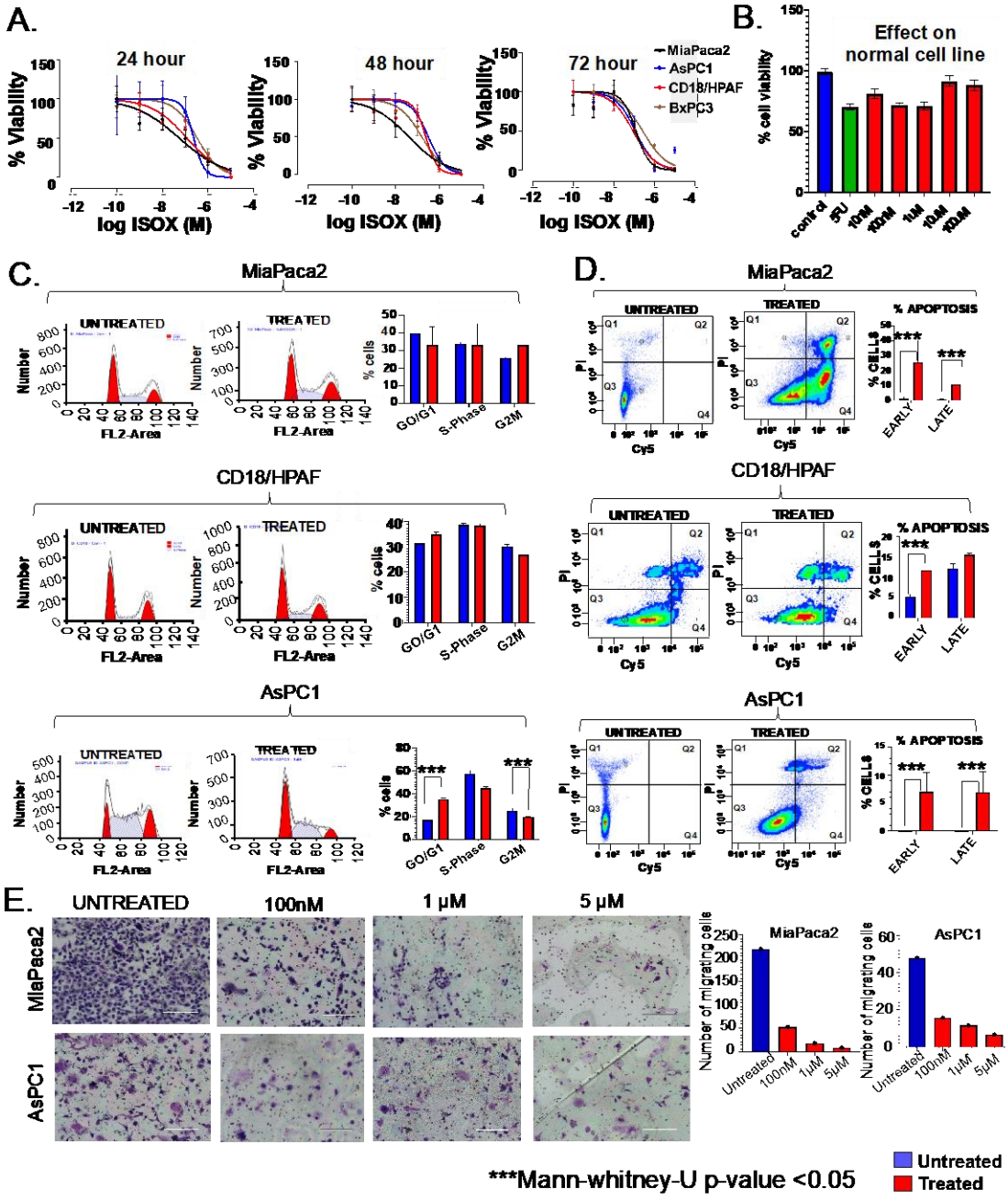
Considering that invasion and migration are established hallmark features of cancer cell aggressiveness and that chemotherapeutic agents often fail to inhibit these processes effectively, we next carried out a matrigel assisted invasion assay using various concentrations of ISOX (100 nM, 1 µM, and 5 µM) with AsPC-1 and MiaPaCa2 cell lines. ISOX successfully reduced the invasion of both the cell lines in a dose-dependent manner (**Fig. 6.2E**). Excitingly, even at 100 nM, ISOX could successfully reduce the percentage of migratory cells by over 83% compared with untreated control. Furthermore, a wound healing assay was carried out at the same concentrations compared to the untreated cells, ISOX treatment at various

concentrations significantly reduced the wound closure examined at multiple time-points (**Fig. 6.3**).

Figure 6. 2 Evaluation of ISOX therapeutic efficacy across pancreatic cancer cell lines.

To validate our *in-silico* identification of ISOX as a potential therapeutic, PC cell-line-based assessment of ISOX was carried out across various PC cell lines. **A. MTT assay**. The effect of ISOX on proliferation was tested using across four commonly used PC cell lines (MiaPaCa2, CD18/HPAF, AsPC1, and BxPC3). An IC50 of as low as 7nM-2.3uM was observed across various cell lines. **B. Effect on normal (HPNE) cells.** To identify the toxicity on normal PC cell lines, an MTT based analysis of varied ISOX doses was carried out on HPNE cell lines using 5FU as the control. The comparison showed minimal toxicity of ISOX even at higher concentrations (80% viability at 100µM). **C. Cell cycle analysis.** In order to identify where in the cell cycle ISOX acting was, a FACS based cell cycle analysis was carried out through PI-based staining of synchronized MiaPaca2, CD18/HPAF, and AsPC1 cells. The representative FACS figures with the associated bar graph representation establish a GO/S arrest with a significant reduction in S phase and a significant increase in the GO/S phase with ISOX treatment. **D. Apoptosis analysis.** The cell cycle analysis was followed up by an apoptosis analysis to compare apoptosis between treated and untreated samples using a FACS based comparison of annexin 5 and PI staining in the same cell lines. **E. Invasion assay.** Invasion is one of the most important properties of cancer cells. To assess how ISOX affects the invasion, a dose-dependent analysis matrigel assisted invasion was carried out. A dose-dependent significant decrease was observed in both MiaPaca2 and AsPC1 cell lines.

Figure 6.2


A.
24 hour
48 hour
72 hour
% Viability
log ISOX (M)
MiaPaca2
AsPC1
CD18/HPAF
BxPC3
B.
Effect on normal cell line
% cell viability
control
5FU
10nM
100nM
1uM
10uM
100uM
C.
MiaPaca2
CD18/HPAF
AsPC1
UNTREATED
TREATED
Number
FL2-Area
% cells
G0/G1
S-Phase
G2M
D.
MiaPaca2
CD18/HPAF
AsPC1
UNTREATED
TREATED
% APOPTOSIS
% CELLS
EARLY
LATE
PI
Cy5
Q1
Q2
Q3
Q4
***
E.
UNTREATED
100nM
1 µM
5 µM
MiaPaca2
AsPC1
Number of migrating cells
Untreated
100nM
1µM
5µM
***Mann-whitney-U p-value <0.05
Untreated
Treated

6.4.6 ISOX inhibits the proliferation of PC cells in combination with 5FU and Gemcitabine.

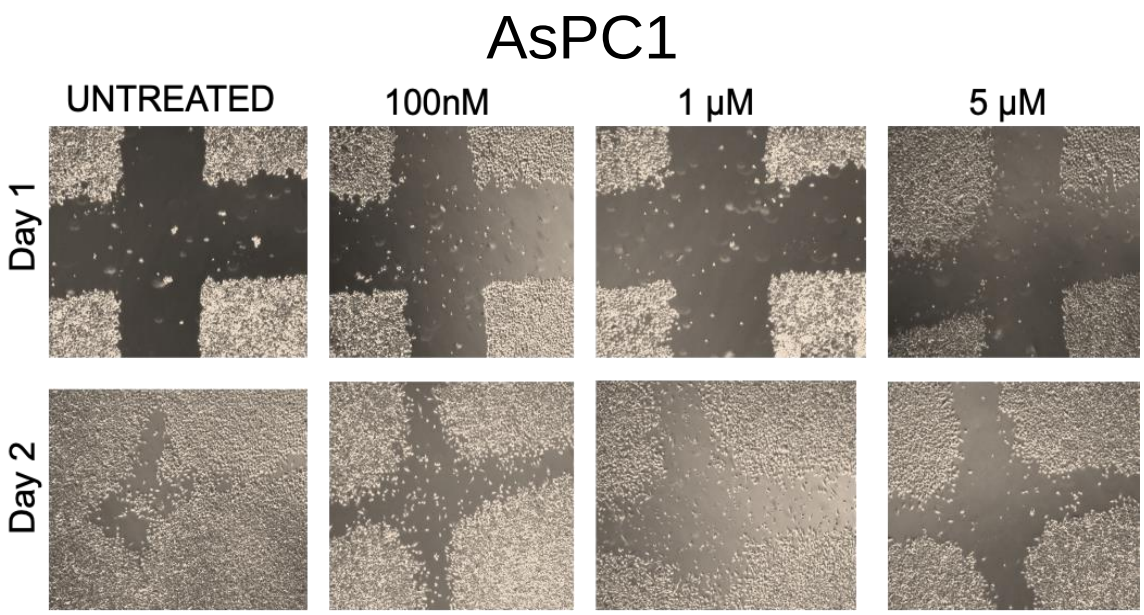
FOLFIRINOX has recently gained popularity as a first-line therapy for patients with advanced pancreatic cancer (*262, 263*). However, toxicity associated with FOLFIRINOX limits its applicability. Notably, 5FU serves as the key component of FOLFIRINOX. Resistance to 5FU due to deficient drug uptake, alterations of targets, activation of DNA repair pathways, apoptosis resistance is considered as a serious challenge for both PC and other malignancies (colon, lung, and stomach) (*264*). Additionally, Gemcitabine serves as a candidate first-line therapy for PC. Considering this and to further our effort into studying ISOX as a potential therapeutic for PC, we next evaluated the impact of ISOX in combination with 5FU and Gemcitabine. MTT analysis was carried out with 5FU (5µM, **Fig. 6.AA**) and Gemcitabine (2 µM, **Fig. 6.4B**) alone and in combination with increasing doses of ISOX (0.156 µM, 0.3125 µM, 0.625 µM, 1.25 µM, 2.5 µM, 5 µM, and 10 µM). The combination of ISOX with 5FU and Gemcitabine significantly reduced the percentage viability across PC cell lines (**Fig. 6.4A and B**). Of note, the combinations reduced the viability of MiaPaCa2 and CD18/HPAF in a dose-dependent manner both at 48 and 72 hours. AsPC1 showed slightly more variability wherein certain lower (1.25 µM ISOX with both drugs) combinations were better than the subsequent higher concentrations (**Fig. 6.4A and B**).

Figure 6.3 Wound healing assay in response to ISOX treatment.

Wound healing assay was carried out comparing the wound healing capabilities of untreated and ISOX treated (100nM, 1μM, and 5μM) AsPC1 and MIaPaca2 cell lines. In conjunction with our earlier studies, as compared to the untreated samples, wounds closure was significantly slower in ISOX treated samples when compared to untreated.

Figure 6.3

A.


AsPC1
UNTREATED
100nM
1 μM
5 μM
Day 1
Day 2

B.

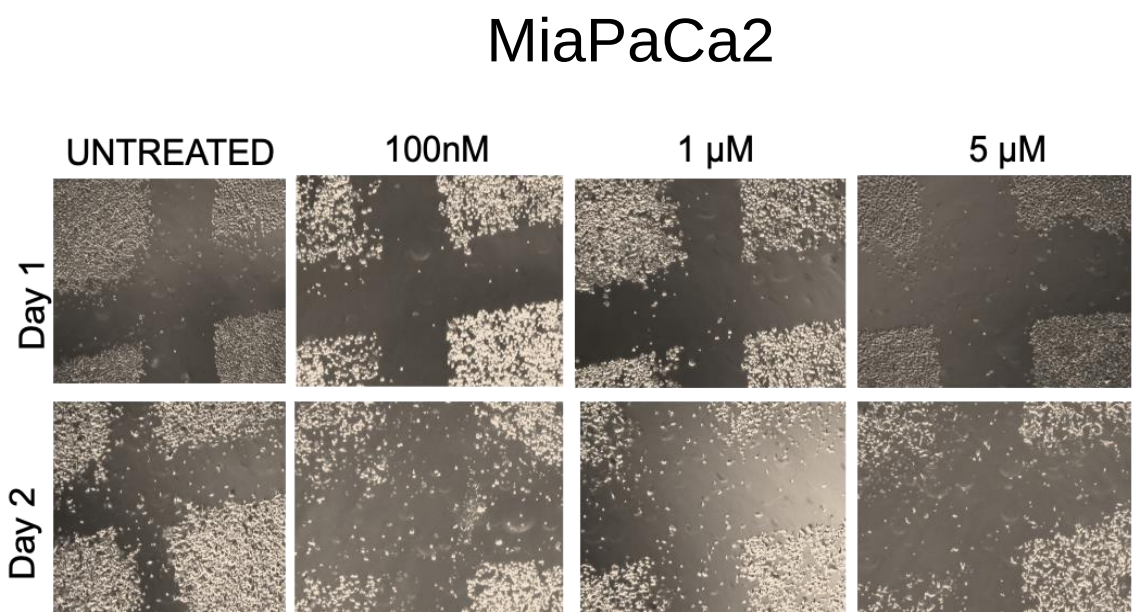
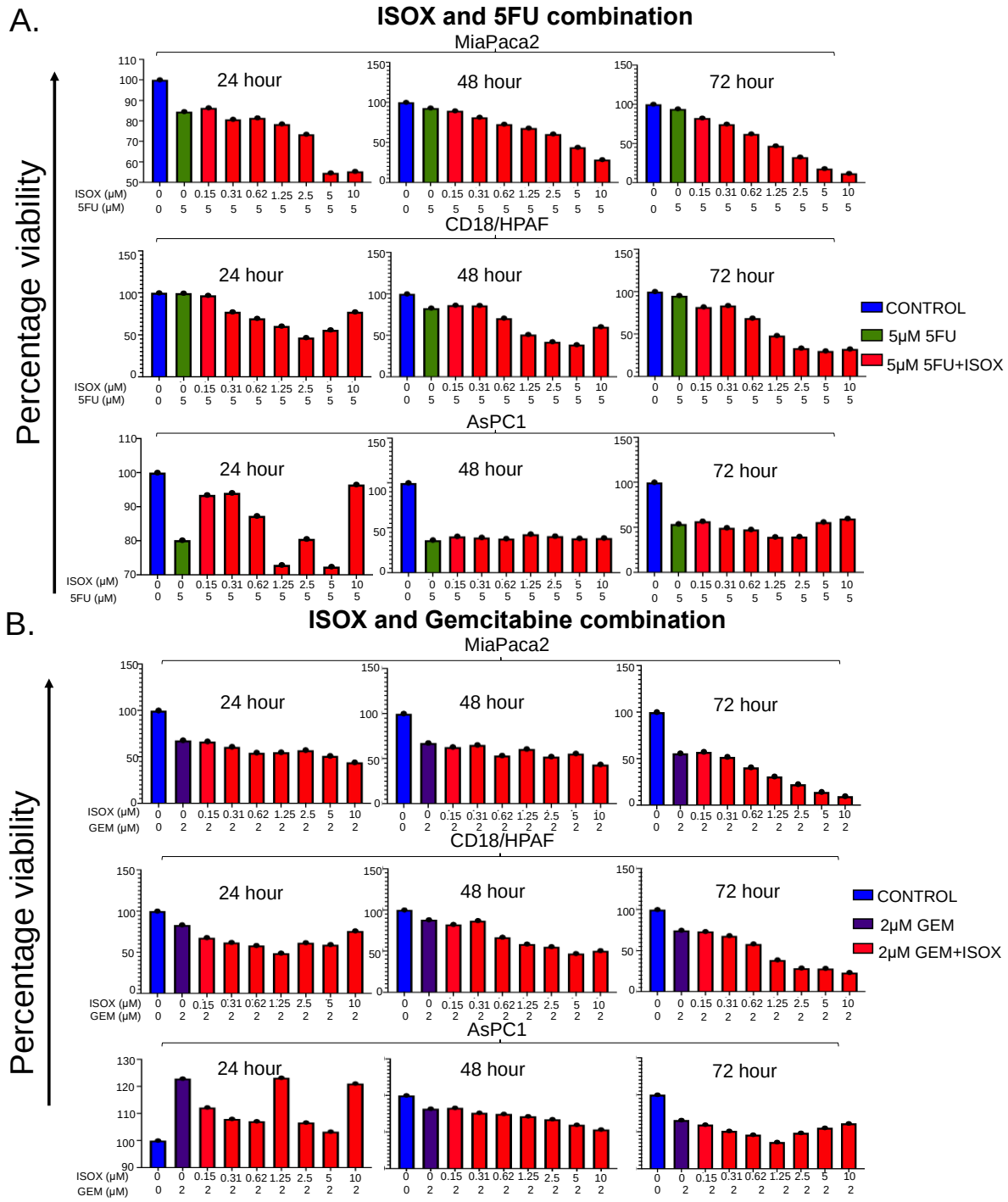

MiaPaCa2
UNTREATED
100nM
1 μM
5 μM
Day 1
Day 2

Figure 6.4 ISOX therapeutic efficacy in combination with two of the most commonly used PC therapeutics; 5-flurouracil (5FU) and Gemcitabine (GEM).

A. Combination with 5FU. To further assess the potential of ISOX in PC, a combination study with 5 µM of 5FU was carried out in the same PC cell lines (MiaPaCa, CD18/HPAF, AsPC1)). In support of our earlier data, ISOX showed very high potential in this combination. The IC50 for the combination was further lowered than the initial values to as low as 0.6 nM. **B. Combination with GEM**. The same PC cell lines (MiaPaCa, CD18/HPAF, AsPC1) were assessed for combination studies for GEM and ISOX. Based on previous studies, 2µM of GEM was used in combination with increasing doses of ISOX.

Figure 6.4


A.
ISOX and 5FU combination
MiaPaca2
CD18/HPAF
AsPC1
24 hour
48 hour
72 hour
Percentage viability
ISOX (μM)
5FU (μM)
CONTROL
5μM 5FU
5μM 5FU+ISOX
B.
ISOX and Gemcitabine combination
GEM (μM)
2μM GEM
2μM GEM+ISOX

ISOX is highly efficacious in inducing growth inhibition in mice and human-derived tumoroids.

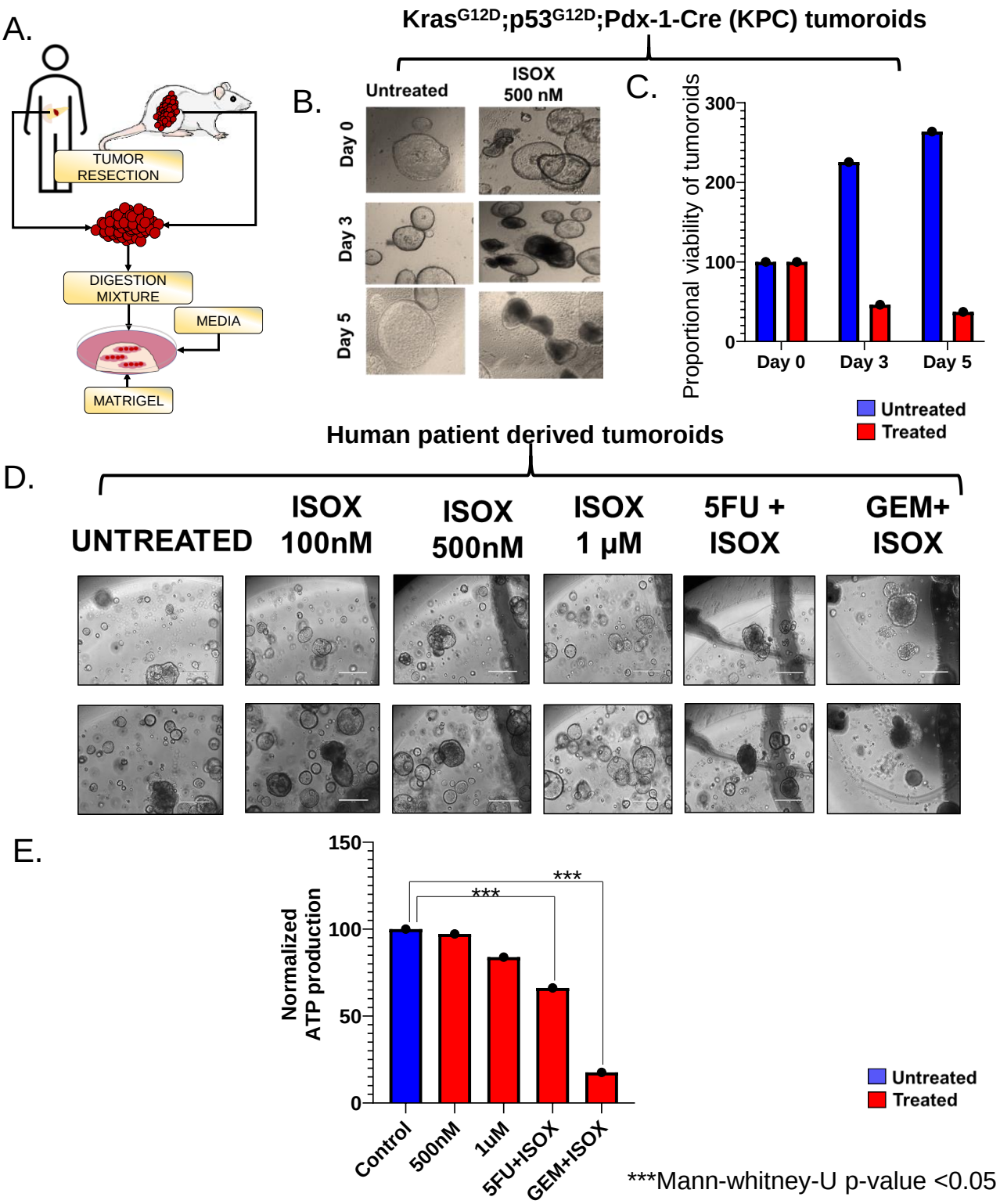
The anti-cancer effects of cancer therapeutics observed in preclinical models often cannot be mirrored in clinical models mainly due to the discordance between the biology of these monolayer cultures and the complex tumor microenvironment of most cancers and specifically complex diseases like PC. In this regard, the 3D cancer tumoroids offer a near-native structure and provide great promise and applicability in drug efficacy studies. Considering this, the efficacy of ISOX was next evaluated in tumoroids-derived from the most common mouse model of PC; KrasG12D; $p53^{R172H}$; Pdx- Cre (KPC) mouse and human PC patient-derived tumoroids (**Fig. 6.5A**). The KPC tumoroids were treated with 500 nM of ISOX and followed for 7 days through imaging. A significant reduction in the viability with morphological changes in the tumoroids with a significant reduction in size and visible darkening of the structures (**Fig. 6.5B**) were observed across drug-treated groups. Proportional viability of tumoroids was assessed where the total number of live tumoroids at day 0 was considered to be the baseline. Interestingly, the proportional viability reduced significantly within 3 days of 500nM of ISOX treatment (**Fig. 6.5C**). A similar assessment was carried out in a set of human patient-derived tumoroids using a range of ISOX concentrations (100nM, 1 µM, and 5µM) and a combination with 5µM of 5FU and 2µM of Gemcitabine (**Fig. 6.5D**). Next, a 3D cyber-glow assay was performed to assess the ATP production in these tumoroids. Dose-dependent reduction of ATP production was observed upon ISOX treatment, thus, suggesting a higher efficacy of ISOX in killing human

tumoroids and inhibiting the viability PC (**Fig. 6.5E**). Interestingly, tumoroids exposed with gemcitabine and ISOX alone or in combination up to 48 hours showed minimal ATP production. Similar effects were observed in an additional group of tumoroids derived from a second PC patient. These second sets of treated human tumoroids were sectioned at the end of the 48-hour treatment. H&E staining of these sections supported the initial loss of viability, wherein a loss of structure could be observed with an increase in concentrations. A tunnel assay and caspase 3 staining also supported this observation with the maximum expression observed in the 5 µM ISOX and 5FU drug combination groups (Fig.6.

Figure 6.5 ISOX is highly effective in inhibiting the growth of mice and human-derived tumoroids.

To assess if this high potency observed in cell lines can also be translated to 3D models, the efficacy of ISOX was observed in KPC mice and human tumor-derived tumoroids. **A. Schematic representation of protocol used for establishing tumoroids from mice and human tumors.** Tumors resected from $KRAS^{G12D}$; $p53^{R172H}$ (KPC) mice and human patient donors were digested using a specific digestive mixture and embedded in matrigel. These tumoroids were then used to assess the efficacy of ISOX. **B. Representative figure of untreated and ISOX treated KPC mouse derived tumoroids.** KPC tumoroids were treated with 500nM of ISOX and followed for 7 days. Untreated tumoroids went on to increase in size and remained healthy the ISOX treated tumoroids showed a significant decrease in size and loss of viability observed through blackening of the tumoroids. **C. Bar graph representation of proportional viability of tumoroids.** The proportional viability of tumoroids in each of the group were observed across the 5 days. The bar graph shows a significant reduction in the viability of tumoroids in the treated group when compared to the untreated group, which remained unaffected. **D. Human tumoroids**. To further establish ISOX as a potential therapeutic, a dose-dependent effect of ISOX alone and in combination with GEM and 5FU was assessed in patient tumor-derived tumoroids. **E. Bar graph representation of diameter of normalized ATP production.** The efficacy of the treated samples was measured using a CellTiter-Glo 3D cell viability assay. The normalized ATP production reduced significantly with an increase in the dose of ISOX and was found to be the least in the combination groups.

Figure 6.5



***Mann-whitney-U p-value <0.05

6.4.7 ISOX alone and in combination reduces the growth of PC orthotropic tumors.

Further, after observing the significant potential of ISOX in cell line and tumoroid models, studies were focused on assessing the effect of ISOX under in-vivo conditions. Luciferase labeled CD18/HPAF cells were orthotopically implanted into the pancreas of athymic nude mice. The tumors were allowed to grow for two weeks, and animals were randomized into four treatment groups by comparing the size of tumors and animals were equally distributed into four groups based on the positive luciferase images. The control group was given intraperitoneal (i/p) injections of PBS, while the treatment groups included 50 mg/kg 5FU, 50 mg/kg ISOX alone, and a combination of both drugs. All treatments were carried out for 5 days followed by a 2-day interval, and this routine was followed for 2 cycles. The mice were subjected to IVIS imaging on day 0, day 10, and day 15. Interestingly, treatment with ISOX alone or in combination with 5FU led to a significant reduction in the size of tumors observed at the end of day 10, with minimal IVIS signal across multiple mice in the treatment group (**Fig. 6.7A**). Comparison of tumor weight between untreated and treated mice showed a statistically significant reduction in both the ISOX alone and combination groups in comparison to the untreated group (**Fig. 6.7B**). Furthermore, these mice were assessed for specific metastatic spots in various organs (peritoneal, mesenteric lymph nodes, intestinal, and kidney). Intriguingly, no distant metastasis was observed in the drug combination group (**Fig. 6.7C**). Similar to the sectioned tumoroids, the sectioned tissues were

subjected to tunnel staining, which showed high level of apoptosis in the ISOX alone and the 5FU groups. This could be corroborated with a caspase 3 staining. Of note is a similar caspase 3 levels in cell lines post-treatment with ISOX alone and in combination with 5FU. While imaging these mice across the treatment cycle (day 10), a few of the mice were imaged using the 3D-BLIT settings within the IVIS system. A representative figure (**Fig. 6.7D**) shows that while the control and 5FU treated mice showed distant metastasis, no metastasis was observed in the ISOX alone and its combination with 5-FU groups. The rest of the mice, when assessed for survival (day of death defined as a natural death or if suggested by the veterinarian) showed that the combination mice survived significantly longer. **Fig. 6.7E**).

6.4.8 Mechanism of ISOX action on PC cells.

While the aforementioned pre-clinical studies helped us to determine the efficacy of ISOX in PC cells, the question of the mechanism of action was entirely unexplored up until this point. A literature review identified ISOX as a potent HDAC inhibitor with the highest inhibitory potential towards HDAC6 (IC_{50} 0.002nM), and significant efficacy towards HDACs 3 (IC_{50} 0.42nM), and 10 (IC_{50} 90.7 nM) (**Fig 6.8A**) (*265*)**.** Based on this, we next explored the status of these HDACs in PC cell lines and tissues. First, we analyzed this expression in PC cell lines (MiaPaca2, CD18/HPAF, AsPC1, CFPAC1, SW1990, Colo357 and T3M4) by immunoblotting with HDAC antibodies and compared their expression relative to normal immortalized pancreas (HPNE and HPDE) cell lines. All these HDACs were found

to be upregulated in the PC cell lines when compared to the normal cell lines (**Fig. 6.8B**). Further, an independent assessment of TCGA pancreatic (tumor) tissue samples with GTex pancreatic (normal) samples showed a higher expression for HDAC3, HDAC6 and HDAC10 in the tumor cells. An IHC analysis for the aforementioned HDACs in human PC tissue sections showed an upregulation of all the three HDACs in both early (PanINs) and PC tumor lesions when compared to normal specimens (**Fig. 6.8C**). All these results together suggested the important role of these specific HDACs in PC initiation and progression. Still, it remained unclear how ISOX mediates its impact in pancreatic tumor cases

6.4.9 ISOX shows an acetylation-dependent effect on c-MYC.

Furthermore, western blot analyses of ISOX treated cells showed a significant increase in the acetylated tubulin level suggesting effective HDAC inhibition. Moreover, ISOX treatment increased the acetylation of cMYC and hence reduced cMYC by itself and the key downstream target p21 and CDK6. (**Fig. 6.8D**)

.

Figure 6. 6 H&E Tunnel and caspase 3 staining in tumoroids treated with ISOX.

Tumoroids treated with ISOX were sectioned and stained using H&E, caspase 3 and subjected to a tunnel staining. In conjunction with our earlier results, a high level of apoptosis was observed in ISOX and combination treated tumoroids.

Figure 6.6

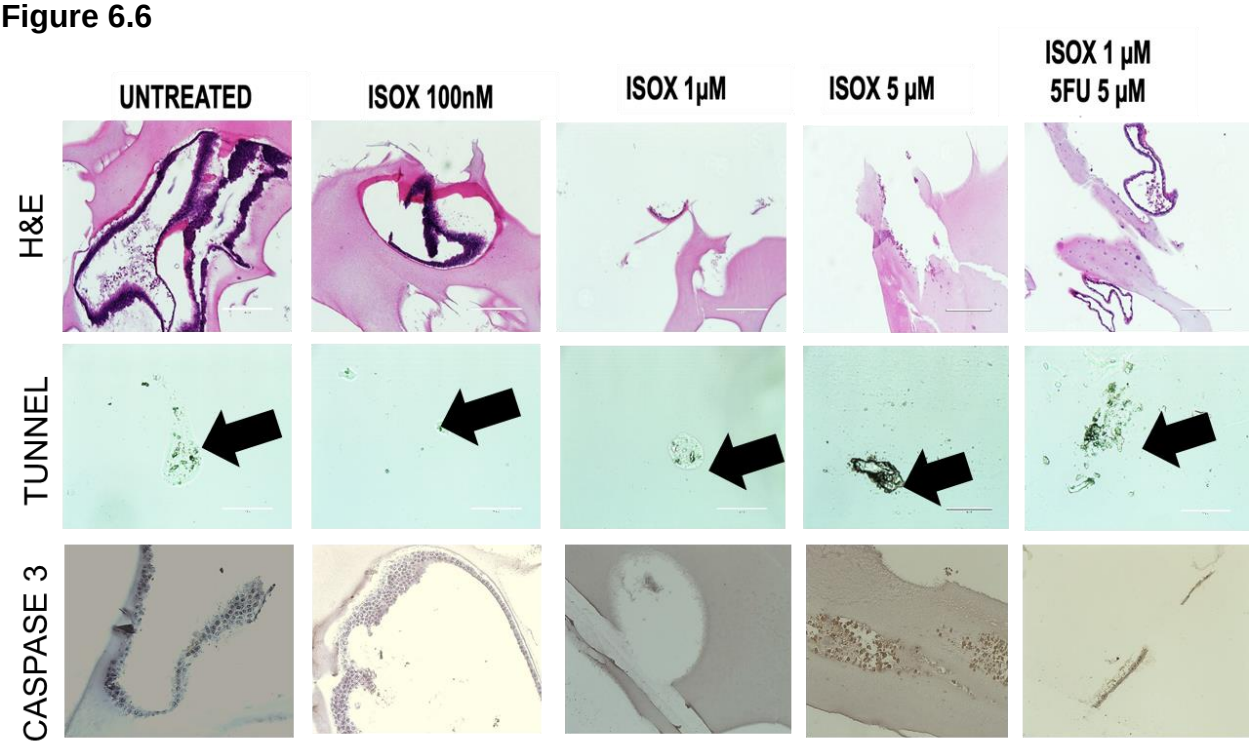
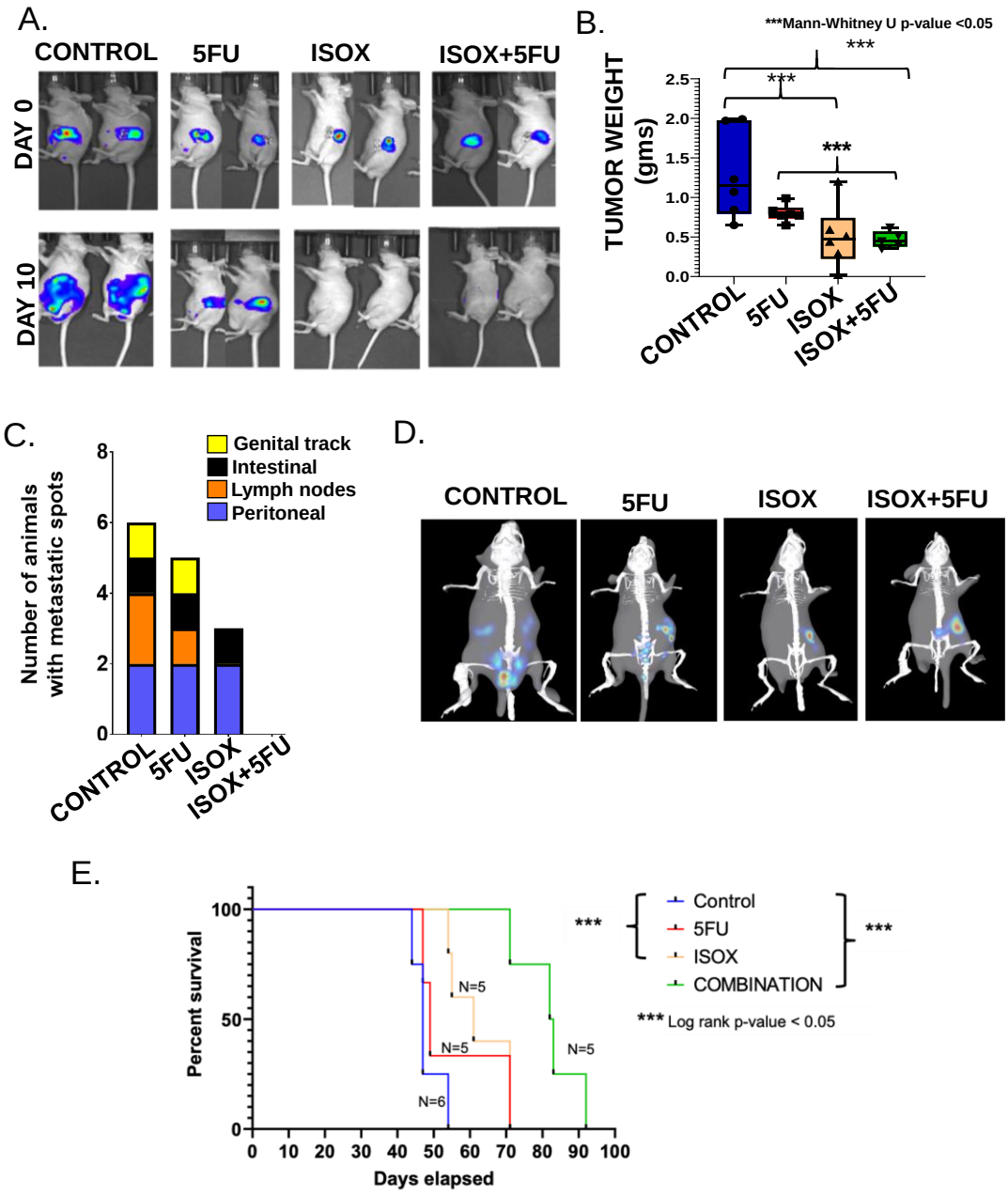

UNTREATED
ISOX 100nM
ISOX 1μM
ISOX 5 μM
ISOX 1 μM
5FU 5 μM
H&E
TUNNEL
CASPASE 3

Figure 6. 7 ISOX is highly effective in reducing the tumor load and increasing the survival of PC orthotopic mice models.

Luciferase labeled CD18/HPAF cells were orthotopically implanted into the pancreas of athymic nude mice. The mice were randomly divided into four groups: control mice (received PBS), 5FU alone (50 mg/kg 5FU), ISOX alone (50 mg/kg ISOX), and combination Mice were treated for 15 days (3 cycles of 5 days treatment with 2 days break) and imaged at the beginning, middle (day 10), and end of the treatment (day 15). At end of day 15, half the mice were sacrificed for tumor weight calculations, and the other half were followed for survival analysis. **A.** IVIS images of two representative mice from each of the groups at the start and middle of the treatment. ISOX alone and combination mice showed a huge reduction in tumor size with the gross tumors gone within 10 days. **B.** Box plot representation of tumor weight from each of the groups show a statistically significant reduction in tumor weight in both ISOX alone and combination groups. **C.** Bar graph representation of a number of metastatic spots. Metastasis was significantly reduced in ISOX treated mice, and there was no metastasis in the combination group. **D.** Representative BLIT images of untreated mice. BLIT images with 3-D reconstructions showed a reduction a high amount of metastasis in the control and 5FU mice but no metastasis in the ISOX treated and combination group mice. **E.** Survival plot for the untreated and treated mice. Survival assessment using a JMP pro (version 14) shows a statistically significant increase in survival in the combination group. The days to death were days post orthotopic implantation where either the mouse died on its own or sacrificing the mouse was suggested by the veterinarian.

Figure 6.7


A.
CONTROL
5FU
ISOX
ISOX+5FU
DAY 0
DAY 10
B.
***Mann-Whitney U p-value <0.05
TUMOR WEIGHT (gms)
CONTROL
5FU
ISOX
ISOX+5FU
C.
Genital track
Intestinal
Lymph nodes
Peritoneal
Number of animals with metastatic spots
D.
CONTROL
5FU
ISOX
ISOX+5FU
E.
Percent survival
Days elapsed
N=5
N=5
N=5
N=6
Control
5FU
ISOX
COMBINATION
*** Log rank p-value < 0.05

6.4.10 Global analyses of ISOX affected pathways.

Further, to get a more in-depth idea of the global effects of ISOX, library of integrated network-based cellular signatures (LINCS) was queried for the ISOX signature using the iLINCS (http://www.ilincs.org/ilincs/) webtool. Specifically, the cancer therapeutics response signature from ISOX/CAY10603 was studied for downstream effectors and similarity with other drug signatures (**Fig. 6.8E**). Interestingly, while exploring this signature, we observed that all members of the MYC-MAX-MAD as key transcription factors affecting the various genes in the ISOX signature (**Fig. 6.8E**). It is also noteworthy to mention that various studies have elucidated the regulation of cMyc through an acetylation-dependent mechanism.

Moreover, to study the global impact of ISOX treatment on PC cells, RNA-sequencing analysis of untreated (CD18/HPAF) and treated cells (CD18/HPAF; 1µM; 48 hours) was carried out, followed by a differential expression analysis (**Fig. 6.8F**). The RNA-sequencing data was further analyzed using TFacts and ENCODE transcription factor analysis tool to assess the transcription factors, whose functions were modulated by ISOX treatment. Intriguingly, both transcription factor analyses showed MYC as a regulator of the differentially expressed genes (**Fig. 6.8F & 6.8G**). Furthermore, pathway analysis of the RNA-seq data showed a high enrichment of sirtuin signaling, epithelial-mesenchymal transition, ERK/MAPK, PI3K-mTOR-AKT, and sonic hedgehog pathway in the treatment group in comparison to the untreated samples (**Fig. 6.8H**). Interestingly, all the aforementioned pathways have been shown to be associated with MYC.

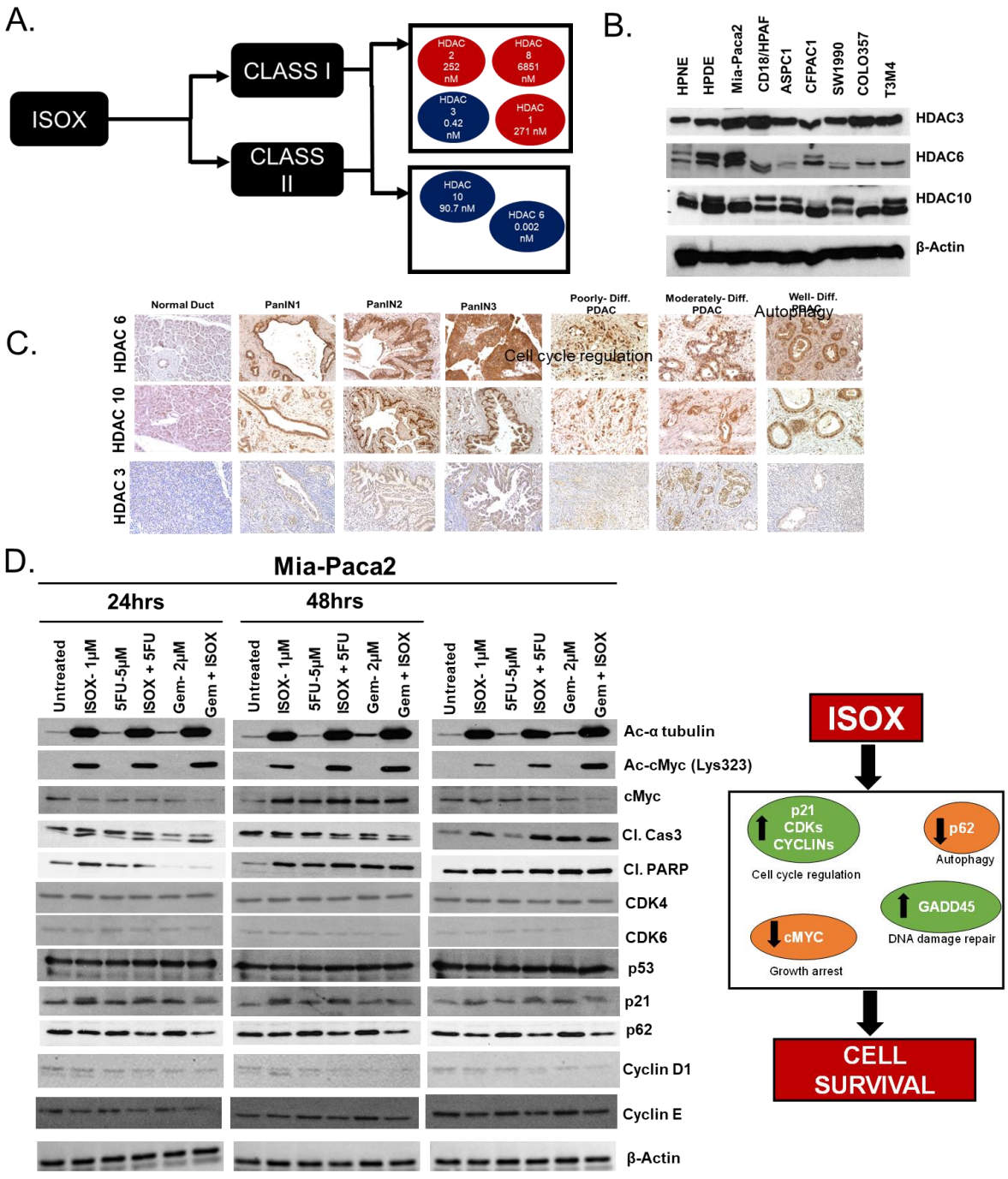
6.4.11 ISOX fares better than other HDAC inhibitors Tubastatin A and Ricolinostat.

To identify the HDAC dependency of ISOX in its mechanism of action, a head-to-head comparison was carried out between ISOX, tubastatin A, and riclinostat. Reassuringly, ISOX performed better in reducing the proliferation of all PC cell lines while tubastatin A and ricolinostat were unable to induce reduction even at 1-10 µM concentrations supporting the efficacy of the method and the unique potential of ISOX as a PC therapeutic. (**Fig. 6.8J**)

Figure 6. 8 Mechanistic studies of ISOX action.

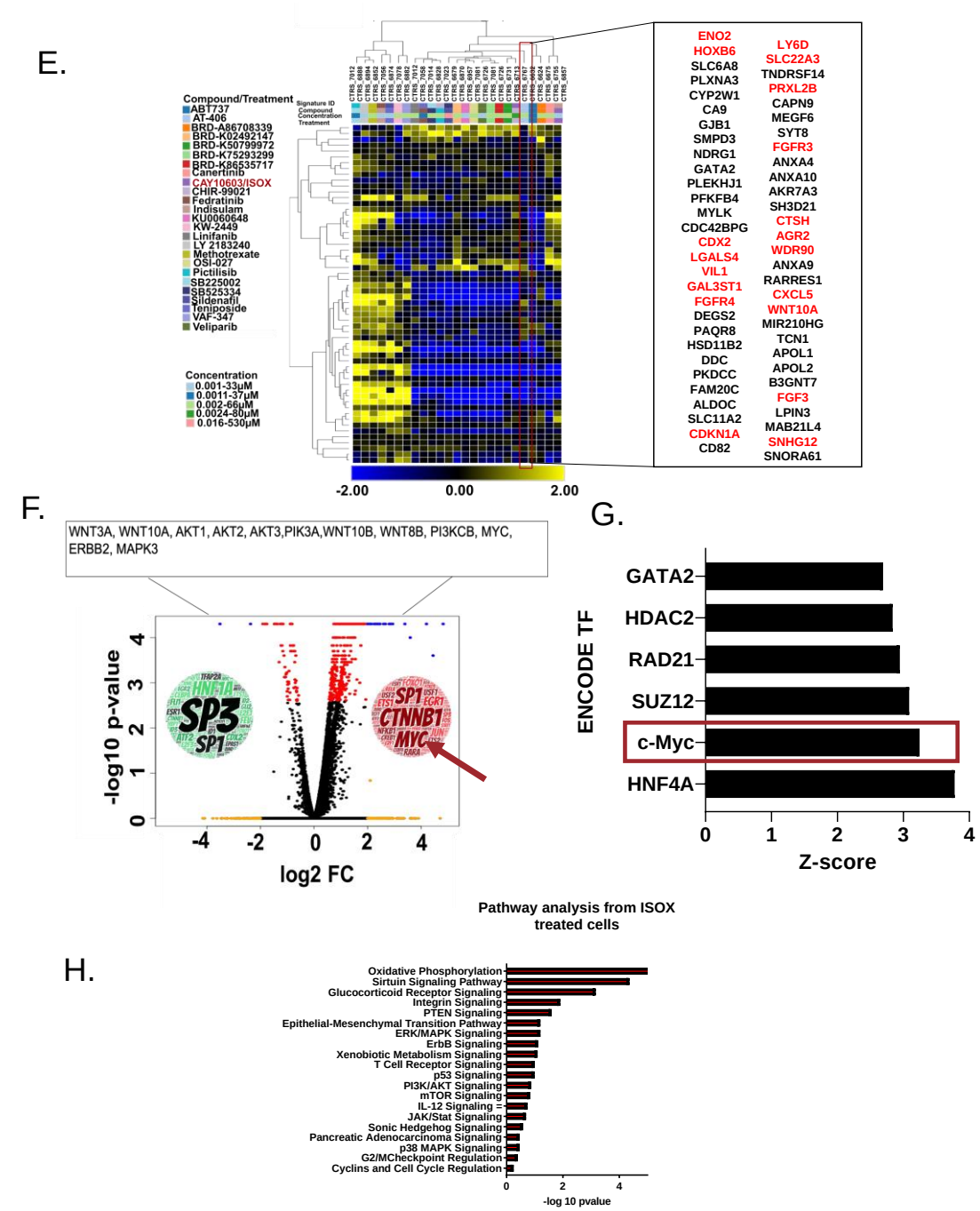
Mechanistic studies were carried out using a combination of in-vitro and *in-silico* assessments, including RNA-seq, western blotting assessment and reverse engineering study. **A.** Schematic representation of IC_{50} values of HDAC inhibition by ISOX. Literature review helped us identify the IC_{50} values of ISOX for inhibiting the activity of various HDACs (*265*). Noteworthy to mention is the fact that ISOX is highly efficacious towards HDACs 3, 6, and 10 (blue color). **B.** Independent assessment of the aforementioned HDACs in PC cell lines by immunoblotting. PC cell lines (MiaPaca2, CD18/HPAF, AsPC1, CFPAC1, SW1990, Colo357 and T3M4) with normal pancreas (HPNE and HPDE) cell lines were assessed for the expression levels of HDACs 3, 6, and 10. Interestingly, there was an upregulation of all these 3 HDACs in PC cell lines when compared to normal cell lines. **C.** Immunohistochemistry (IHC) analysis of normal and tumor tissue from patient samples. IHC successfully corroborated the earlier assessment of overexpression of HDACs 3, 6, and 10 in the tumor samples when compared to the normal samples. **D.** Western blot analysis of ISOX treated samples

Figure 6.8


A.
ISOX
CLASS I
CLASS II
HDAC 2 252 nM
HDAC 8 6851 nM
HDAC 3 0.42 nM
HDAC 1 271 nM
HDAC 10 90.7 nM
HDAC 6 0.002 nM
B.
HPNE
HPDE
Mia-Paca2
CD18/HPAF
ASPC1
CFPAC1
SW1990
COLO357
T3M4
HDAC3
HDAC6
HDAC10
β-Actin
C.
Normal Duct
PanIN1
PanIN2
PanIN3
Poorly- Diff. PDAC
Moderately- Diff. PDAC
Well- Diff. PDAC
HDAC 6
HDAC 10
HDAC 3
Autophagy
Cell cycle regulation
D.
Mia-Paca2
24hrs
48hrs
Untreated
ISOX- 1μM
5FU-5μM
ISOX + 5FU
Gem- 2μM
Gem + ISOX
Ac-α tubulin
Ac-cMyc (Lys323)
cMyc
Cl. Cas3
Cl. PARP
CDK4
CDK6
p53
p21
p62
Cyclin D1
Cyclin E
β-Actin
ISOX
p21 CDKs CYCLINs
Cell cycle regulation
p62
Autophagy
GADD45
DNA damage repair
cMYC
Growth arrest
CELL SURVIVAL

E. Heat map representation of ISOX gene signature in conjunction with the other related therapeutics. Marked in red are the genes related directly or indirectly to MYC. **F.** Volcano plot representing RNA-seq analysis of ISOX treated CD18/HPAF cells. CD18/HPAF cells were treated with 1 μM of ISOX (for 48 hours) and subjected to RNA-seq analysis. The word cloud within the volcano plot represents transcription factors regulating the genes differentially expressed between untreated and treated samples. **G.** Bar graph representation of z-scores from encoding transcription factor analysis of the genes downregulated between untreated and treated samples. Genes downregulated by the treatment were subjected a transcription factor analysis using the iLincs server. Noteworthy to mention is that various important transcription factors, including HDAC2, GATA2, and cMYC, were significantly affected by this treatment. **H.** Bar graph representation of pathways affected by ISOX as observed through an ingenuity pathway analysis (IPA)

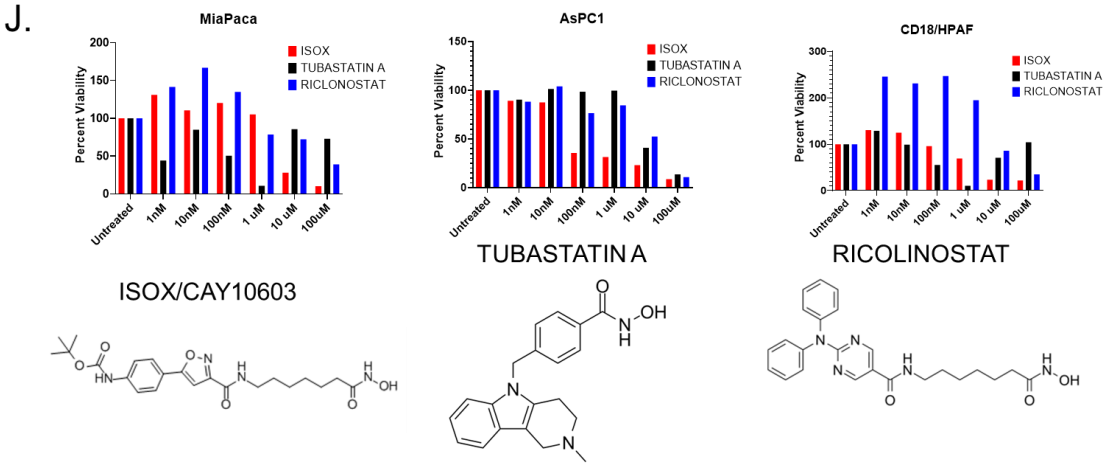
Figure 6.8


E.
Compound/Treatment
ABT737
AT-406
BRD-A86708339
BRD-K02492147
BRD-K50799972
BRD-K75293299
BRD-K86535717
Canertinib
CAY10603/ISOX
CHIR-99021
Fedratinib
Indisulam
KU0060648
KW-2449
Linifanib
LY 2183240
Methotrexate
OSI-027
Pictilisib
SB225002
SB525334
Sildenafil
Teniposide
VAF-347
Veliparib
Concentration
0.001-33μM
0.0011-37μM
0.002-66μM
0.0024-80μM
0.016-530μM
-2.00
0.00
2.00
ENO2
HOXB6
SLC6A8
PLXNA3
CYP2W1
CA9
GJB1
SMPD3
NDRG1
GATA2
PLEKHJ1
PFKFB4
MYLK
CDC42BPG
CDX2
LGALS4
VIL1
GAL3ST1
FGFR4
DEGS2
PAQR8
HSD11B2
DDC
PKDCC
FAM20C
ALDOC
SLC11A2
CDKN1A
CD82
LY6D
SLC22A3
TNDRSF14
PRXL2B
CAPN9
MEGF6
SYT8
FGFR3
ANXA4
ANXA10
AKR7A3
SH3D21
CTSH
AGR2
WDR90
ANXA9
RARRES1
CXCL5
WNT10A
MIR210HG
TCN1
APOL1
APOL2
B3GNT7
FGF3
LPIN3
MAB21L4
SNHG12
SNORA61
F.
WNT3A, WNT10A, AKT1, AKT2, AKT3,PIK3A,WNT10B, WNT8B, PI3KCB, MYC, ERBB2, MAPK3
-log10 p-value
log2 FC
G.
ENCODE TF
GATA2
HDAC2
RAD21
SUZ12
c-Myc
HNF4A
Z-score
H.
Pathway analysis from ISOX treated cells
Oxidative Phosphorylation
Sirtuin Signaling Pathway
Glucocorticoid Receptor Signaling
Integrin Signaling
PTEN Signaling
Epithelial-Mesenchymal Transition Pathway
ERK/MAPK Signaling
ErbB Signaling
Xenobiotic Metabolism Signaling
T Cell Receptor Signaling
p53 Signaling
PI3K/AKT Signaling
mTOR Signaling
IL-12 Signaling =
JAK/Stat Signaling
Sonic Hedgehog Signaling
Pancreatic Adenocarcinoma Signaling
p38 MAPK Signaling
G2/MCheckpoint Regulation
Cyclins and Cell Cycle Regulation
-log 10 pvalue

J. Head-to-head comparison of ISOX with other HDAC inhibitors. Bar graph representation of percentage viability of various PC cell lines with increasing doses of ISOX, two other HDAC inhibitors tubastatin A and riclinostat. Interestingly, ISOX (red) was far better in reducing the viability of PC cell lines when compared to its counterparts.

Figure 6.8

J.


MiaPaca
AsPC1
CD18/HPAF
ISOX
TUBASTATIN A
RICLONOSTAT
Percent Viability
Untreated
1nM
10nM
100nM
1 uM
10 uM
100uM
ISOX/CAY10603
TUBASTATIN A
RICOLINOSTAT

6.5 DISCUSSION

Targeting PC has been a challenge to both researchers and clinicians. While various drugs and combinations thereof have been assessed in clinical trials, we have not been able to either achieve promising survival or improve the quality of life of these patients. The failure of these therapies can be attributed to the failure of the one target at a time approach, and better methods with a multiple gene targeting approaches can prove beneficial. Recent advances in molecular profiling of tumors led to the identification of distinct molecular signatures amongst patients, which has led to efforts directed towards establishing targeted therapies or devising precision medicine strategies. Efforts have been made to develop tailored treatments for a subset of the patients with one or multiple actionable mutations the response, however, remains minuscule. Considering this, the current study uses an *in-silico* approach to target the whole gene expression profile of PC.

Our global *in-silico* assessment led to identifying a highly specific and novel therapeutic for PC; ISOX. Through our in-depth analysis using PC cell lines, tumoroid and orthotopic mice models, we could successfully establish the potential of ISOX as a potential therapeutic for PC. The thorough literature review helped us to identify ISOX as a histone deacetylase (HDAC) inhibitor with the highest efficacy towards HDAC6, with a potent effect on HDAC3 and HDAC10**.** (*265, 266*) Interestingly, an independent assessment of comparison of normal pancreas and PC cell lines and the cancer genome atlas (TCGA) based comparison of normal and tumor tissue samples showed an upregulation of HDACs 3, 6, and 10 in both

PC cell line and the tissue samples. Interestingly, the HDAC family of proteins has been established to play essential roles in the initiation, promotion, and cancer progression due to their direct effect on various histone and non-histone substrates (*267, 268*). Notably, prominent genes like tubulin, PI3K-AKT, p53, HSP90, NF-KB, ERK have been established to be direct targets of the HDAC family (*269*). Interestingly, these substrates are known to play a crucial function in cancer biology by playing a direct role in cell cycle regulation, apoptosis, DNA-damage response, autophagy, and other such vital pathways (*270*). Furthermore, epigenetic deregulation has been established as a hallmark of cancer, and effectively targeting these deregulations can prove to be extremely effective. Various HDAC inhibitors such as vorinostat (SAHA), panobinostat, valopric acid, abexinostat, etc. have been studied through multiple *in-vitro* and *in-vivo* models in the laboratory, and many more have fared well in clinical trials suggesting the potential of HDACi as a cancer therapeutic strategy.

However, studies have reported that pan-HDAC or class independent HDAC inhibitors (HDACi) often have a variety of adverse effects. To solve this problem, medicinal chemist Hyun-Jung Kim and colleagues (*271*) suggest the development of isoform-specific inhibitors either targeting one of the specific classes or, more specifically, one of the HDACs. As mentioned, ISOX shows the highest efficacy towards HDAC6; a structurally unique HDAC owing to the presence of two catalase domains has been established as a promising target (*272*). Recent studies have demonstrated the role of HDAC6 as a central target in cancer therapy, because of

its role in oncogenic cell transformation. But owing to the structural complexity, it is extremely challenging to target HDAC6 (*273*).

Additionally, ARID1A (a SWI/SNF -complex chromatin-component gene), is commonly mutated in several cancers, has been established to have an HDAC6 dependence in their action (*274*). Interestingly, ARID1A is one of the most mutated genes in the TCGA PC dataset and highly mutated in various other PC datasets, suggesting the importance of targeting HDAC6 for effective therapeutic intervention in PC. Interestingly, ISOX has been shown to have a very high specificity towards HDAC6. Further, our comparison studies between ISOX and other HDAC6 inhibitors, trichostatin A and riclinostat showed that ISOX fairs better than both the other inhibitors in inhibiting the proliferation of PC cell lines supporting using ISOX in a PC setting. This can possibly be explained by that besides HDAC6, ISOX also shows an excellent efficacy towards HDAC3 and HDAC10, both of which are also critical targets in cancer biology and in PC. HDAC3 has been known to have roles in lung, ovarian, and colorectal cancers due to its central role in mitosis regulation (*275, 276*). HDAC10, on the other hand, has been established to play an important role in the regulation of stem-like cell properties of KRAS-driven lung adenocarcinoma, which supports its use in other KRAS-driven cancers like PC (*277*). Interestingly, the other HDACi(s) do not have the same kind of specificity (to HDACs 3, 6, and 10) as ISOX does, further supporting it as a potential therapeutic.

While this background was strong enough to support the exploration of ISOX as a potential therapeutic, our RNA-sequencing analysis helped us to gain insight into the differences that make ISOX more efficacious than the other HDAC inhibitors. In conjunction with a two-way enrichment study for the regulating transcription factors and pathway analysis, the RNA-sequencing analysis helped establish various important pathways like PTEN signaling, ERK signaling, PI3K/AKT/mTOR signaling, etc., as targets for ISOX. Furthermore, a combination of studying the ISOX signature from LINCS and our RNA-seq data paved the path to the identification of MYC and related pathways as the direct downstream effectors of ISOX action. MYC and its related signaling pathways like EGFR, PI3K-AKT-MTOR signaling have been established to have extremely important roles in PC. Interestingly, studies have established MYC to be regulated by HDAC dependent acetylation suggesting a direct downstream effect of the HDAC family proteins (*278, 279*); however, further in-depth studies need to be carried out to understand that these cMYC and related pathways are getting affected through HDAC inhibition or independently.

While the study has been extremely successful in establishing ISOX as a potential therapeutic, there are various limitations that need to be addressed as we move forward. Firstly, the study is limited by CMAP since it only queries against the drugs tested within the database. There could be a potential drug targeting the exact signature for PC, which is yet to be included in the CMAP dataset. Furthermore,

the variability between the responses observed in various cell lines needs to be studied for specific mechanism differences.

Overall, this study successfully establishes ISOX as a new therapeutic for PC. ISOX proved to be highly efficacious in PC cell lines, tumoroids, and mice models. More interestingly, through an HDAC driven mechanism, ISOX effectively targets multiple pathways, which are known to be important in PC. While years of research have identified drugs affecting single targets, this approach has not been successful in PC due to its late diagnosis, high complexity, early metastasis, and dense stroma. We have successfully demonstrated the impact of ISOX on master regulators through our study targeting multiple pathways, including HDAC6, 3, and 10 and MYC, and hence multiple pertinent downstream pathways. Owing to the low dose of ISOX action, low toxicity in normal cell lines, and multiple pathways targeting establishes it as a unique and highly potent therapy for PC. This comprehensive preclinical assessment has led to the identification of ISOX as a potential therapeutic for PC. We believe that this pipeline and ISOX in itself will prove beneficial for PC patients. Further assessment in PC progression models and clinical trials will help us establish the use of ISOX in the clinics and eventually have direct implications on better patient care.

Chapter 7: Overall Conclusions and Future Directions

7.1 OVERALL CONCLUSIONS

Advances in computational power, along with our ability to collect and investigate large datasets, have had overwhelming implications in disease diagnosis, detection, and therapeutic interventions. The advanced present-day algorithms combined with Moore's law increase in computational power has given us the capability of uncovering otherwise undetectable patterns. This, combined with our ability to validate these identified patterns in the laboratory, has thus led to a shift in the way we approach biological questions and seek solutions. Considering this, the overarching aim of this Ph.D. dissertation was the development of *in-silico* pipelines for the identification and characterization of biomarker panels and therapeutic interventions in high mortality gastrointestinal cancers. This chapter provides a summary and future directions of the different aspects covered in this dissertation.

7.1.1 Global *in-silico* analysis of mucins in Colorectal Cancer identifies specific MUC16 signaling.

The 22-member mucin family consists of high molecular weight glycoproteins with imperative roles in the initiation and progression of various cancers. These roles, in turn, lead to specific applications both in the early discovery of malignancies (as biomarkers) and using targeting mucins for therapeutic interventions. Specifically, in colorectal malignancies, differential expression and glycosylation profile of these mucins have been associated with benign and malignant pathologies. While studies relating to specific mucins provide some insight, a comprehensive

assessment has not been carried out to date. Furthermore, there is varied evidence of aberrant expression and mutation profiles which makes the applicability extremely challenging. Considering this lack of consensus, this study carried out a bioinformatics-driven assessment of mucins in colorectal cancer to identify expression and mutation patterns considering early precursor lesions and tumor samples.

The expression, survival, and mutational data from the cancer genome atlas (TCGA) for tumor (N=380) and normal samples (N= 51) were used in conjunction with microarray data for precursor lesions, namely tubular adenomas (TA) and sessile serrated adenomas/polyps (SSA/Ps) from GEO datasets. For TCGA, the pre-processed expression data was downloaded from the UCSC-Xena web server and further processed using R-Bioconductor and SAS-JMP (v15). The expression data were then matched in a sample-specific way to the survival information from the TCGA server and processed further using SAS-JMP. Furthermore, mutational data from the TCGA sample set was assessed using the cBioPortal webtool and further in R-Bioconductor. The microarray data were pre-processed and assessed for expression using the “affy” and “limma” packages within R-Bioconductor.

Overall assessment of expression data comparing the tumor and normal samples from TCGA showed an upregulation of MUC1, MUC5AC, and MUC15 within the tumors. Furthermore, we saw a loss of MUC2 and MUC4. Interestingly, MUC16 was seen to be highly expressed in a specific subset of patients. Closer assessment of these patients led to the identification of a strong correlation between microsatellite instability (MSI) status and MUC16 expression, wherein

most of the MSI-high patients also showed a higher expression of MUC16. Furthermore, the study of the expression differences between normal colon, tubular adenomas (TA), and sessile serrated adenomas/polyps (SSA/Ps) led to the identification of MUC5AC and MUC17 as being uniquely overexpressed in SSA/Ps and not in the TAs. Furthermore, mutational analysis of tumor samples from TCGA led to an interesting finding wherein 28% of the TCGA patients had a mutation in MUC16. This assessment could be independently validated in RNA-seq data from various other sources with a mutation range of 10-40% all across. Intriguingly, these mutations were highly correlated with other mutations common in colorectal cancer like BRAF, MSH6, TP53, KRAS, and others. Additional assessment of these mutations based on MUC16 domain structure helped in identifying most of these as part of the SEA domain of this large mucin. Interestingly, the SEA domain has been widely studied for its role in cell migration and cancer metastasis. Considering both the overexpression and high mutation of MUC16 in colorectal cancer, a functional assessment of MUC16 and its associated gene signature was then carried out. Interestingly, similar to other cancers the MUC16 associated signature was significantly associated with immune-based signatures suggesting its potential importance in colorectal cancer setup (data not presented).

The study, for the first time, identifies this specific subset of colorectal cancer patients with a MUC16 overexpression. Considering the importance of MUC16 in cancer biology and studies in other cancers, this can prove a potential therapeutic avenue for future studies.

7.1.2. Computational analysis of Mucins in Gastric Cancer identifies prognostically relevant clusters.

The aforementioned mucin family of proteins is known to have important roles within gastric adenocarcinoma. However, similar to colorectal cancer, a comprehensive analysis of these mucins has not yet been carried out in the gastric cancer setting. Considering this and taking leverage of the big data resources, this study aimed to assess differential mucin signature using bioinformatics techniques.

Like the colorectal cancer study, the expression, mutation, and survival data and statistics from the TCGA gastric adenocarcinoma dataset were downloaded from UCSC Xena and assessed using R-Bioconductor and SAS-JMP. Furthermore, the mutational analysis was verified using cBioPortal a web-based tool.

This *in-silico* assessment helped us identify a loss of MUC5AC in gastric adenocarcinoma cases. Furthermore, a significant upregulation of MUC13, MUC20, and MUC15 was observed. Interestingly, MUC13 and MUC15 were also observed to have a significant prognostic significance. Additionally, similar to the earlier study, a very high percentage of MUC16 high cases showed a mutational significance, with over 30% of patients showing a MUC16 mutation. Interestingly, this could be corroborated in various other datasets from gastric cancers. Furthermore, an unsupervised clustering analysis identified specific mucin signatures as potential biomarkers. Most interestingly, MUC1 and MUC13 formed a very strong cluster which could also be corroborated in independent validation datasets.

Overall, this study has led to the identification of potential mucin-based biomarker panels, which can find their application in prognostication and further the study of mucins in the gastric cancer setting.

7.1.3 Presence and structure-activity relationship of intrinsically disordered regions across mucins.

Intrinsically disordered regions (IDRs) are sequences of low complexity with a low proportion of hydrophobic residues and a high number of repeating residues. Furthermore, these regions show a preponderance of polar and charged residues and a general lack of an ordered core that comprises a traditional structured domain. Additionally, these IDRs, in turn, show a specific biological implication, such as in cell cycle regulation, transcription, splicing, translation, and signaling. Further, IDRs or intrinsically disordered proteins (IDPs) are known to have specific roles in various diseases such as many cancers, cardiovascular defects, diabetes, and others. Considering this and the aforementioned role of mucins in various cancers, this study focused on identifying IDRs within the various members of the mucin family of proteins.

The study uses a sequence-based approach making use of the web-based prediction tool D^2P^2 to identify regions of intrinsic disorder with mucins. Further, to identify the functional role of these IDRs, Pfam domain prediction was applied to identify IDRs present within predicted domains. This was then followed by an assessment of phosphorylation and glycosylation sites. Furthermore, mucin

interactome studies and a functional assessment were carried out to correlate the presence of IDRs with their functions.

Interestingly, the study identified a high prevalence of IDRs across the entire mucin family. All mucins were predicted to have a high (>40%) to moderate (>20% and <40%) levels of disorder. Specifically, transmembrane mucins were disordered compared to secreted mucins except for MUC7. Interestingly, the average predicted disorder (58%) within the mucin family exceeded the average disorder (30%) within the human genome. Additionally, many of the mucins were found to have a high number of large-sized molecular recognition features (MoRFs) known to have implications for protein-protein interactions. Specifically, the largest known mucin, MUC16, was found to have the highest number of MoRFs and, in turn, the highest number of interacting partners. The study further successfully identified cancer-related pathways to be affected by mucins and the interacting partners, corroborating earlier known studies and establishing the importance of IDRs.

This study, for the first time, studied the presence of IDRs within the mucin family. Considering the important role of mucins in cancer biology, this identification can lead to specific applications of targeting these IDRs for therapeutic interventions. Furthermore, the correlation of the post-translation modifications with these IDRs can help in studying the biological effects of mucins in greater detail and will be of great importance to the field.

7.1.4 Connectivity Mapping-based identification and evaluation of ISOX: A novel therapeutic strategy for Pancreatic Cancer.

Pancreatic Cancer (PC) maintains its unglorified third rank amongst all malignancy-related deaths in the United States and is projected to escalate to the second position by 2030. The Pancreatic Cancer Action Network (PanCAN) explicitly declares the lack of promising therapeutic modalities as a major confounder for this appalling survival. Therapeutics interventions in PC have been challenging for researchers and physicians alike. While various efforts have been made to identify new potent therapeutics, the one target at a time approach has failed to identify robust therapeutic interventions. Better methods are urgently needed. In this regard, genomics-driven computational methods have started gaining popularity in recent times. This study uses one such approach called the Connectivity Map (CMAP), a big data repository of gene expression profiles of various small molecule inhibitors.

The data from 106 tumors from PC patients and 68 normal samples was processed using R-Bioconductor, and a limma-based analysis used to generate a global gene signature for PC. This signature was mapped to the over 2300 drug profiles within the big-data CMAP to identify nine drugs with the propensity to reverse the global gene signature from PC. Further, to identify drugs with higher specificity, a similar CMAP analysis was carried out using patient-derived xenografts (PDX), human cell lines, and human tumoroids. This comprehensive CMAP analysis led to the identification of ISOX as a potential therapeutic drug for PC. This *in-silico* identification was then followed by an *in vitro* and *in vivo* assessment of ISOX

efficacy in various PC models. Furthermore, RNA-seq assessment followed by a broad *in-silico* analysis was carried out to identify the mechanism of action for the small molecule ISOX in PC.

Computationally identified small molecule ISOX was found to be highly efficacious (IC_{50} 2.4 nM-1.5 µM) in myriad PC cell lines. Furthermore, combination studies with commonly used chemotherapies gemcitabine and 5-FU showed that ISOX could synergistically increase the efficacy of both drugs. Additionally, ISOX induced over 50% apoptosis in PC cell lines and caused a G0/G1 arrest of PC cells. Interestingly, this effect of ISOX could also be observed in 3D tumoroid models of PC derived from KPC ($KRAS^{G12D}$, $p53^{R172H}$; Pdx-1-Cre) mice and human tumor tissues. In pancreatic orthotopic mouse studies, a 10-fold reduction in tumor weight was observed at 50 mg/kg of ISOX alone (p-value = 0.014) and in combination with 5-FU (p-value=0.02). Furthermore, the combination-treated mice showed no local or distant metastasis. Interestingly, the ISOX alone and combination-treated mice survived significantly (log-rank p-value < 0.05) better than the control animals, further supporting the therapeutic potential of ISOX. Further, RNA-seq analysis of ISOX treated mice led to the identification of HDAC dependent cMYC inhibition as a mechanism of action of ISOX. This was further supplemented by a pathway analysis of ISOX-LINCS signature, which helped to identify specific enrichment of MYC-MAX-MAD complex and hence establish the mechanism of action.

While CMAP has been used in other malignancies, this study provides one of the first comprehensive assessments of the CMAP data in PC setting. Furthermore,

this study establishes a pipeline starting from an *in-silico* identification followed by an in-vitro and in-vivo validation that can be applied to other cancers as well as other disease settings. Additionally, the study successfully established ISOX as a potential therapeutic strategy for PC.

7.2 FUTURE DIRECTIONS

Overall, through the use of bioinformatics and sophisticated laboratory-based validation techniques, we have been able to answer pertinent problems relating to GI cancers. The comprehensive analysis of mucins in disease initiation and progression in gastric and colorectal cancers helped us establish a consensus for the role of these mucins in both these pathologies. Noteworthy to mention is that this study, for the first time, established the clinical relevance of MUC16 in colorectal cancer, which can have direct implications in therapeutic interventions. Additionally, our published study of the intrinsically disordered regions in mucins has not only led to a better understanding of these proteins but also holds relevance in the therapeutic targeting of these mucins. Further, our pre-clinical assessment of ISOX helped identify and establish ISOX as a highly effective therapy for pancreatic cancer, and we are working on translating this into a clinical setting. While these studies have helped us understand various aspects of GI cancers, a lot remains to be explored.

7.2.1 Exploratory studies for MUC16 based therapeutic interventions in CRC.

MUC16 expression and mutation have been associated with therapy in various cancers (*95*), (*107*), (*280*). Our study found MUC16 to be upregulated in a subset

of CRC patients paving the path for such interventions in CRC. While a lot of research has been carried out to identify a potential therapy for CRC patients, most effort has limited success, with survival showing very slight improvement. In this regard, if we can find therapeutics which can increase the survival of even a smaller subset of patients, it would still be worth pursuing. Considering this, exploration of MUC16 based therapeutics in MSI high CRC patients will prove beneficial.

7.2.2 Validation of IDR through NMR and X-Ray crystallography studies.

As previously mentioned, our study of IDR in mucins identified a large percentage of intrinsically disordered regions within various mucins. IDRs have been implicated to have roles in protein-protein interaction, diseases, and hence therapeutic interventions. Further, considering the important role of mucins in various GI cancers, information about IDRs can prove beneficial for therapeutic interventions in these high mortality cancers. However, computational methods applied in this study warrant a wet lab validation before these intrinsically disordered regions can be explored further. Considering that, NMR or X-Ray crystallography-based validation of mucin IDRs will prove to be beneficial and will have implications in targeting mucins for therapeutic interventions.

7.2.3 ISOX mechanistic studies.

Our studies in PC helped us identify and establish ISOX as a potential therapeutic for PC. Mechanistic studies of ISOX action have led to the identification of both HDAC inhibition as well as a MYC inhibition driven action. However, further studies

into the mechanism of ISOX's action would prove extremely valuable to our understanding of this novel therapeutic. Our RNA-sequencing analysis has given us a novel insight into the mechanism of action of ISOX and has also opened avenues for further research. The pathway analysis also identified specific stem cell related pathways which are known to be extremely relevant in PC progression. Current studies are underway to identify the more specific effects of ISOX on the stem cell population and the downstream effectors. Further, various other pathways like PI3K-AKT-mTOR signaling are worth pursuing. Furthermore, previous studies have elucidated that ARID1A mutations are dependent on HDAC6 in tumor progression (*274*). Furthermore, ARID1A is one of the top mutated genes in various PC datasets. Considering this, it would be extremely interesting to study the interdependence of ARID1A and ISOX action.

7.2.4 ISOX efficacy studies in patient-derived xenograft models.

To gain more insight into the efficacy of ISOX in pancreatic cancer, patient-derived-xenograft models (PDX) of pancreatic cancer can be used in addition to the cell-line orthotopic models already used in the study. Furthermore, the most common mouse models for PC namely, KC and KPC, can be used to study the efficacy of ISOX.

7.2.5 Toxicity studies on ISOX.

Additionally, considering that the final goal of these therapy-based studies are human clinical trials, it would be imperative to carry out toxicity studies in mouse

and bigger animal models. While no initial toxicity was observed in the orthotopic mouse models, these results might vary in immune-efficient animal models. These studies would add to the translation potential of ISOX and hence make it easier to translate into a clinical setting.

7.2.6 Clinical trials for ISOX efficacy.

The eventual aim of this study is to be able is to test the efficacy of ISOX in a clinical setting in the form of a clinical trial.

Chapter 8: References

1. J. Weitz, M. Koch, J. Debus, T. Hohler, P. R. Galle, M. W. Buchler, Colorectal cancer. *Lancet* **365**, 153-165 (2005).
2. A. Recio-Boiles, B. Cagir, in StatPearls. (Treasure Island (FL), 2021).
3. G. Binefa, F. Rodriguez-Moranta, A. Teule, M. Medina-Hayas, Colorectal cancer: from prevention to personalized medicine. *World J Gastroenterol* **20**, 6786-6808 (2014).
4. L. Yamane, C. Scapulatempo-Neto, R. M. Reis, D. P. Guimaraes, Serrated pathway in colorectal carcinogenesis. *World J Gastroenterol* **20**, 2634-2640 (2014).
5. T. Murakami, N. Sakamoto, A. Nagahara, Clinicopathological features, diagnosis, and treatment of sessile serrated adenoma/polyp with dysplasia/carcinoma. *J Gastroenterol Hepatol* **34**, 1685-1695 (2019).
6. M. S. Pino, D. C. Chung, The chromosomal instability pathway in colon cancer. *Gastroenterology* **138**, 2059-2072 (2010).
7. J. R. Jass, Hyperplastic polyps and colorectal cancer: is there a link? *Clin Gastroenterol Hepatol* **2**, 1-8 (2004).
8. E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. van Grieken, F. Lordick, Gastric cancer. *Lancet* **396**, 635-648 (2020).
9. D. M. Ye, G. Xu, W. Ma, Y. Li, W. Luo, Y. Xiao, Y. Liu, Z. Zhang, Significant function and research progress of biomarkers in gastric cancer. *Oncol Lett* **19**, 17-29 (2020).
10. Y. Binenbaum, S. Na'ara, Z. Gil, Gemcitabine resistance in pancreatic ductal adenocarcinoma. *Drug Resist Updat* **23**, 55-68 (2015).
11. M. J. Moore, D. Goldstein, J. Hamm, A. Figer, J. R. Hecht, S. Gallinger, H. J. Au, P. Murawa, D. Walde, R. A. Wolff, D. Campos, R. Lim, K. Ding, G. Clark, T. Voskoglou-Nomikos, M. Ptasynski, W. Parulekar, G. National Cancer Institute of Canada Clinical Trials, Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol* **25**, 1960-1966 (2007).
12. D. Ophir, T. Hahn, A. Schattner, D. Wallach, A. Aviel, Tumor necrosis factor in middle ear effusions. *Arch Otolaryngol Head Neck Surg* **114**, 1256-1258 (1988).
13. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, A. K. Dunker, Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**, 573-584 (2002).
14. Z. Du, V. N. Uversky, A Comprehensive Survey of the Roles of Highly Disordered Proteins in Type 2 Diabetes. *Int J Mol Sci* **18**, (2017).
15. Y. Cheng, T. LeGall, C. J. Oldfield, A. K. Dunker, V. N. Uversky, Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* **45**, 10448-10460 (2006).
16. P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, P. T. Lansbury, Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* **35**, 13709-13715 (1996).
17. A. H. S. Martinelli, F. C. Lopes, E. B. O. John, C. R. Carlini, R. Ligabue-Braun, Modulation of Disordered Proteins with a Focus on

Neurodegenerative Diseases and Other Pathologies. *Int J Mol Sci* **20**, (2019).
18. J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, A. K. Dunker, Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873-6888 (2006).
19. C. A. Galea, Y. Wang, S. G. Sivakolundu, R. W. Kriwacki, Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* **47**, 7598-7609 (2008).
20. M. Schenone, V. Dancik, B. K. Wagner, P. A. Clemons, Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* **9**, 232-240 (2013).
21. H. Wang, Q. Gu, J. Wei, Z. Cao, Q. Liu, Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. *Clin Pharmacol Ther* **97**, 451-454 (2015).
22. N. U. Sahu, P. S. Kharkar, Computational Drug Repositioning: A Lateral Approach to Traditional Drug Discovery? *Curr Top Med Chem* **16**, 2069-2077 (2016).
23. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S. H. Friend, Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).
24. I. Smith, P. G. Greenside, T. Natoli, D. L. Lahr, D. Wadden, I. Tirosh, R. Narayan, D. E. Root, T. R. Golub, A. Subramanian, J. G. Doench, Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol* **15**, e2003213 (2017).
25. J. Yu, P. Putcha, J. M. Silva, Recovering drug-induced apoptosis subnetwork from Connectivity Map data. *Biomed Res Int* **2015**, 708563 (2015).
26. F. H. Chung, Y. R. Chiang, A. L. Tseng, Y. C. Sung, J. Lu, M. C. Huang, N. Ma, H. C. Lee, Functional Module Connectivity Map (FMCM): a framework for searching repurposed drug compounds for systems treatment of cancer and an application to colorectal adenocarcinoma. *PLoS One* **9**, e86299 (2014).
27. X. Liu, P. Zeng, Q. Cui, Y. Zhou, Comparative analysis of genes frequently regulated by drugs based on connectivity map transcriptome data. *PLoS One* **12**, e0179037 (2017).
28. F. Caiment, M. Tsamou, D. Jennen, J. Kleinjans, Assessing compound carcinogenicity in vitro using connectivity mapping. *Carcinogenesis* **35**, 201-207 (2014).
29. L. Litichevskiy, R. Peckner, J. G. Abelin, J. K. Asiedu, A. L. Creech, J. F. Davis, D. Davison, C. M. Dunning, J. D. Egertson, S. Egri, J. Gould, T. Ko, S. A. Johnson, D. L. Lahr, D. Lam, Z. Liu, N. J. Lyons, X. Lu, B. X. MacLean,

A. E. Mungenast, A. Officer, T. E. Natoli, M. Papanastasiou, J. Patel, V. Sharma, C. Toder, A. A. Tubelli, J. Z. Young, S. A. Carr, T. R. Golub, A. Subramanian, M. J. MacCoss, L. H. Tsai, J. D. Jaffe, A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. *Cell Syst* **6**, 424-443 e427 (2018).

30. A. L. Creech, J. E. Taylor, V. K. Maier, X. Wu, C. M. Feeney, N. D. Udeshi, S. E. Peach, J. S. Boehm, J. T. Lee, S. A. Carr, J. D. Jaffe, Building the Connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry. *Methods* **72**, 57-64 (2015).
31. P. E. Oberstein, K. P. Olive, Pancreatic cancer: why is it so hard to treat? *Therap Adv Gastroenterol* **6**, 321-337 (2013).
32. L. Ma, J. Wei, G. H. Su, J. Lin, Dasatinib can enhance paclitaxel and gemcitabine inhibitory activity in human pancreatic cancer cells. *Cancer Biol Ther* **20**, 855-865 (2019).
33. W. Chien, Q. Y. Sun, K. L. Lee, L. W. Ding, P. Wuensche, L. A. Torres-Fernandez, S. Z. Tan, I. Tokatly, N. Zaiden, L. Poellinger, S. Mori, H. Yang, J. W. Tyner, H. P. Koeffler, Activation of protein phosphatase 2A tumor suppressor as potential treatment of pancreatic cancer. *Mol Oncol* **9**, 889-905 (2015).
34. J. L. Er, P. N. Goh, C. Y. Lee, Y. J. Tan, L. W. Hii, C. W. Mai, F. F. Chung, C. O. Leong, Identification of inhibitors synergizing gemcitabine sensitivity in the squamous subtype of pancreatic ductal adenocarcinoma (PDAC). *Apoptosis* **23**, 343-355 (2018).
35. D. E. Biancur, J. A. Paulo, B. Malachowska, M. Quiles Del Rey, C. M. Sousa, X. Wang, A. S. W. Sohn, G. C. Chu, S. P. Gygi, J. W. Harper, W. Fendler, J. D. Mancias, A. C. Kimmelman, Compensatory metabolic networks in pancreatic cancers upon perturbation of glutamine metabolism. *Nat Commun* **8**, 15965 (2017).
36. W. D. Xi, Y. J. Liu, X. B. Sun, J. Shan, L. Yi, T. T. Zhang, Bioinformatics analysis of RNA-seq data revealed critical genes in colon adenocarcinoma. *Eur Rev Med Pharmacol Sci* **21**, 3012-3020 (2017).
37. Q. Wen, P. O'Reilly, P. D. Dunne, M. Lawler, S. Van Schaeybroeck, M. Salto-Tellez, P. Hamilton, S. D. Zhang, Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC Syst Biol* **9 Suppl 5**, S4 (2015).
38. L. Zhang, W. Kang, X. Lu, S. Ma, L. Dong, B. Zou, Weighted gene co-expression network analysis and connectivity map identifies lovastatin as a treatment option of gastric cancer by inhibiting HDAC2. *Gene* **681**, 15-25 (2019).
39. Z. X. Chen, X. P. Zou, H. Q. Yan, R. Zhang, J. S. Pang, X. G. Qin, R. Q. He, J. Ma, Z. B. Feng, G. Chen, T. Q. Gan, Identification of putative drugs for gastric adenocarcinoma utilizing differentially expressed genes and connectivity map. Mol Med Rep **19**, 1004-1015 (2019).
40. Y. T. Chen, J. Y. Xie, Q. Sun, W. J. Mo, Novel drug candidates for treating esophageal carcinoma: A study on differentially expressed genes, using

connectivity mapping and molecular docking. Int J Oncol **54**, 152-166 (2019).
41. L. M. Liu, P. Lin, H. Yang, Y. W. Dang, G. Chen, Gene profiling of HepG2 cells following nitidine chloride treatment: An investigation with microarray and Connectivity Mapping. *Oncol Rep* **41**, 3244-3256 (2019).
42. D. Sharma, G. Subbarao, R. Saxena, Hepatoblastoma. *Semin Diagn Pathol* **34**, 192-200 (2017).
43. A. Beck, C. Eberherr, M. Hagemann, S. Cairo, B. Haberle, C. Vokuhl, D. von Schweinitz, R. Kappler, Connectivity map identifies HDAC inhibition as a treatment option of high-risk hepatoblastoma. *Cancer Biol Ther* **17**, 1168-1176 (2016).
44. Y. Zhong, E. Y. Chen, R. Liu, P. Y. Chuang, S. K. Mallipattu, C. M. Tan, N. R. Clark, Y. Deng, P. E. Klotman, A. Ma'ayan, J. C. He, Renoprotective effect of combined inhibition of angiotensin-converting enzyme and histone deacetylase. *J Am Soc Nephrol* **24**, 801-811 (2013).
45. L. F. Zerbini, M. K. Bhasin, J. F. de Vasconcellos, J. D. Paccez, X. Gu, A. L. Kung, T. A. Libermann, Computational repositioning and preclinical validation of pentamidine for renal cell cancer. *Mol Cancer Ther* **13**, 1929-1941 (2014).
46. J. S. Pang, Z. K. Li, P. Lin, X. D. Wang, G. Chen, H. B. Yan, S. H. Li, The underlying molecular mechanism and potential drugs for treatment in papillary renal cell carcinoma: A study based on TCGA and Cmap datasets. *Oncol Rep* **41**, 2089-2102 (2019).
47. O. Menyhart, F. Giangaspero, B. Gyorffy, Molecular markers and potential therapeutic targets in non-WNT/non-SHH (group 3 and group 4) medulloblastomas. *J Hematol Oncol* **12**, 29 (2019).
48. C. C. Faria, S. Agnihotri, S. C. Mack, B. J. Golbourn, R. J. Diaz, S. Olsen, M. Bryant, M. Bebenek, X. Wang, K. C. Bertrand, M. Kushida, R. Head, I. Clark, P. Dirks, C. A. Smith, M. D. Taylor, J. T. Rutka, Identification of alsterpaullone as a novel small molecule inhibitor to target group 3 medulloblastoma. *Oncotarget* **6**, 21718-21729 (2015).
49. G. Manzotti, S. Parenti, G. Ferrari-Amorotti, A. R. Soliera, S. Cattelani, M. Montanari, D. Cavalli, A. Ertel, A. Grande, B. Calabretta, Monocyte-macrophage differentiation of acute myeloid leukemia cell lines by small molecules identified through interrogation of the Connectivity Map database. *Cell Cycle* **14**, 2578-2589 (2015).
50. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2020. *CA Cancer J Clin* **70**, 7-30 (2020).
51. D. J. Stumpel, P. Schneider, L. Seslija, H. Osaki, O. Williams, R. Pieters, R. W. Stam, Connectivity mapping identifies HDAC inhibitors for the treatment of t(4;11)-positive infant acute lymphoblastic leukemia. *Leukemia* **26**, 682-692 (2012).
52. E. Fang, X. Zhang, Identification of breast cancer hub genes and analysis of prognostic values using integrated bioinformatics analysis. Cancer Biomark **21**, 373-381 (2017).

53. J. Busby, L. Murray, K. Mills, S. D. Zhang, F. Liberante, C. R. Cardwell, A combined connectivity mapping and pharmacoepidemiology approach to identify existing medications with breast cancer causing or preventing properties. Pharmacoepidemiol Drug Saf **27**, 78-86 (2018).
54. T. Liu, H. Zhang, L. Sun, D. Zhao, P. Liu, M. Yan, N. Zaidi, S. Izadmehr, A. Gupta, W. Abu-Amer, M. Luo, J. Yang, X. Ou, Y. Wang, X. Bai, Y. Wang, M. I. New, M. Zaidi, T. Yuen, C. Liu, FSIP1 binds HER2 directly to regulate breast cancer growth and invasiveness. *Proc Natl Acad Sci U S A* **114**, 7683-7688 (2017).
55. G. Thillaiyampalam, F. Liberante, L. Murray, C. Cardwell, K. Mills, S. D. Zhang, An integrated meta-analysis approach to identifying medications with potential to alter breast cancer risk through connectivity mapping. *BMC Bioinformatics* **18**, 581 (2017).
56. E. Fang, X. Zhang, Q. Wang, D. Wang, Identification of prostate cancer hub genes and therapeutic agents using bioinformatics approach. *Cancer Biomark* **20**, 553-561 (2017).
57. J. Li, Y. H. Xu, Y. Lu, X. P. Ma, P. Chen, S. W. Luo, Z. G. Jia, Y. Liu, Y. Guo, Identifying differentially expressed genes and small molecule drugs for prostate cancer by a bioinformatics strategy. *Asian Pac J Cancer Prev* **14**, 5281-5286 (2013).
58. D. G. McArt, P. D. Dunne, J. K. Blayney, M. Salto-Tellez, S. Van Schaeybroeck, P. W. Hamilton, S. D. Zhang, Connectivity Mapping for Candidate Therapeutics Identification Using Next Generation Sequencing RNA-Seq Data. *PLoS One* **8**, e66902 (2013).
59. Z. A. Yochum, J. Cades, L. Mazzacurati, N. M. Neumann, S. K. Khetarpal, S. Chatterjee, H. Wang, M. A. Attar, E. H. Huang, S. N. Chatley, K. Nugent, A. Somasundaram, J. A. Engh, A. J. Ewald, Y. J. Cho, C. M. Rudin, P. T. Tran, T. F. Burns, A First-in-Class TWIST1 Inhibitor with Activity in Oncogene-Driven Lung Cancer. *Mol Cancer Res* **15**, 1764-1776 (2017).
60. W. Zhuo, L. Zhang, Y. Zhu, Q. Xie, B. Zhu, Z. Chen, Valproic acid, an inhibitor of class I histone deacetylases, reverses acquired Erlotinib-resistance of lung adenocarcinoma cells: a Connectivity Mapping analysis and an experimental study. *Am J Cancer Res* **5**, 2202-2211 (2015).
61. R. Raghavan, S. Hyter, H. B. Pathak, A. K. Godwin, G. Konecny, C. Wang, E. L. Goode, B. L. Fridley, Drug discovery using clinical outcome-based Connectivity Mapping: application to ovarian cancer. *BMC Genomics* **17**, 811 (2016).
62. J. Y. Xie, P. C. Chen, J. L. Zhang, Z. S. Gao, H. Neves, S. D. Zhang, Q. Wen, W. D. Chen, H. F. Kwok, Y. Lin, The prognostic significance of DAPK1 in bladder cancer. *PLoS One* **12**, e0175290 (2017).
63. J. A. Shin, J. H. Lee, S. Y. Lim, H. S. Ha, H. S. Kwon, Y. M. Park, W. C. Lee, M. I. Kang, H. W. Yim, K. H. Yoon, H. Y. Son, Metabolic syndrome as a predictor of type 2 diabetes, and its clinical interpretations and usefulness. *J Diabetes Investig* **4**, 334-343 (2013).

64. Q. Wang, Z. Zhao, J. Shang, W. Xia, Targets and candidate agents for type 2 diabetes treatment with computational bioinformatics approach. *J Diabetes Res* **2014**, 763936 (2014).
65. M. Zhang, H. Luo, Z. Xi, E. Rogaeva, Drug repositioning for diabetes based on 'omics' data mining. *PLoS One* **10**, e0126082 (2015).
66. J. Ma, S. Malladi, A. H. Beck, Systematic Analysis of Sex-Linked Molecular Alterations and Therapies in Cancer. *Sci Rep* **6**, 19119 (2016).
67. F. Cheng, J. Zhao, M. Fooksa, Z. Zhao, A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. *J Am Med Inform Assoc* **23**, 681-691 (2016).
68. D. W. Day, The adenoma-carcinoma sequence. *Scand J Gastroenterol Suppl* **104**, 99-107 (1984).
69. B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. Smits, J. L. Bos, Genetic alterations during colorectal-tumor development. *N Engl J Med* **319**, 525-532 (1988).
70. J. R. Jass, Serrated route to colorectal cancer: back street or super highway? *J Pathol* **193**, 283-285 (2001).
71. M. J. Makinen, S. M. George, P. Jernvall, J. Makela, P. Vihko, T. J. Karttunen, Colorectal carcinoma associated with serrated adenoma--prevalence, histological features, and prognosis. *J Pathol* **193**, 286-294 (2001).
72. A. Biondi, R. Fisichella, F. Fiorica, M. Malaguarnera, F. Basile, Food mutagen and gastrointestinal cancer. *Eur Rev Med Pharmacol Sci* **16**, 1280-1282 (2012).
73. S. R. Krishn, S. Kaur, L. M. Smith, S. L. Johansson, M. Jain, A. Patel, S. K. Gautam, M. A. Hollingsworth, U. Mandel, H. Clausen, W. C. Lo, W. T. Fan, U. Manne, S. K. Batra, Mucins and associated glycan signatures in colon adenoma-carcinoma sequence: Prospective pathological implication(s) for early diagnosis of colon cancer. *Cancer Lett* **374**, 304-314 (2016).
74. M. D. Walsh, M. Clendenning, E. Williamson, S. A. Pearson, R. J. Walters, B. Nagler, D. Packenas, A. K. Win, J. L. Hopper, M. A. Jenkins, A. M. Haydon, C. Rosty, D. R. English, G. G. Giles, M. A. McGuckin, J. P. Young, D. D. Buchanan, Expression of MUC2, MUC5AC, MUC5B, and MUC6 mucins in colorectal cancers and their association with the CpG island methylator phenotype. *Mod Pathol* **26**, 1642-1656 (2013).
75. R. Pothuraju, S. Rachagani, S. R. Krishn, S. Chaudhary, R. K. Nimmakayala, J. A. Siddiqui, K. Ganguly, I. Lakshmanan, J. L. Cox, K. Mallya, S. Kaur, S. K. Batra, Molecular implications of MUC5AC-CD44 axis in colorectal cancer progression and chemoresistance. *Mol Cancer* **19**, 37 (2020).
76. A. Velcich, W. Yang, J. Heyer, A. Fragale, C. Nicholas, S. Viani, R. Kucherlapati, M. Lipkin, K. Yang, L. Augenlicht, Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science* **295**, 1726-1729 (2002).
77. M. Wu, Y. Wu, J. Li, Y. Bao, Y. Guo, W. Yang, The Dynamic Changes of Gut Microbiota in Muc2 Deficient Mice. *Int J Mol Sci* **19**, (2018).

78. Y. S. Shan, H. P. Hsu, M. D. Lai, M. C. Yen, J. H. Fang, T. Y. Weng, Y. L. Chen, Suppression of mucin 2 promotes interleukin-6 secretion and tumor growth in an orthotopic immune-competent colon cancer animal model. *Oncol Rep* **32**, 2335-2342 (2014).
79. H. P. Hsu, M. D. Lai, J. C. Lee, M. C. Yen, T. Y. Weng, W. C. Chen, J. H. Fang, Y. L. Chen, Mucin 2 silencing promotes colon cancer metastasis through interleukin-6 signaling. *Sci Rep* **7**, 5823 (2017).
80. W. Li, N. Zhang, C. Jin, M. D. Long, H. Rajabi, Y. Yasumizu, A. Fushimi, N. Yamashita, M. Hagiwara, R. Zheng, J. Wang, L. Kui, H. Singh, S. Kharbanda, Q. Hu, S. Liu, D. Kufe, MUC1-C drives stemness in progression of colitis to colorectal cancer. *JCI Insight* **5**, (2020).
81. C. Li, T. Liu, L. Yin, D. Zuo, Y. Lin, L. Wang, Prognostic and clinicopathological value of MUC1 expression in colorectal cancer: A meta-analysis. *Medicine (Baltimore)* **98**, e14659 (2019).
82. M. Guo, B. Luo, M. Pan, M. Li, F. Zhao, J. Dou, MUC1 plays an essential role in tumor immunity of colorectal cancer stem cell vaccine. *Int Immunopharmacol* **85**, 106631 (2020).
83. S. Das, S. Rachagani, Y. Sheinin, L. M. Smith, C. B. Gurumurthy, H. K. Roy, S. K. Batra, Mice deficient in Muc4 are resistant to experimental colitis and colitis-associated colorectal cancer. *Oncogene* **35**, 2645-2654 (2016).
84. S. Lu, C. Catalano, S. Huhn, B. Pardini, L. Partu, V. Vymetalkova, L. Vodickova, M. Levy, T. Buchler, K. Hemminki, P. Vodicka, A. Forsti, Single nucleotide polymorphisms within MUC4 are associated with colorectal cancer survival. *PLoS One* **14**, e0216666 (2019).
85. S. D. Rico, D. Hoflmayer, F. Buscheck, D. Dum, A. M. Luebke, M. Kluth, C. Hube-Magg, A. Hinsch, C. Moller-Koop, D. Perez, J. R. Izbicki, M. Neipp, H. Mofid, H. Larusson, T. Daniels, C. Isbert, S. Coerper, D. Ditterich, H. Rupprecht, A. Goetz, C. Fraune, K. Moller, A. Menz, C. Bernreuther, T. S. Clauditz, G. Sauter, R. Uhlig, W. Wilczak, R. Simon, S. Steurer, P. Lebok, E. Burandt, T. Krech, A. H. Marx, Elevated MUC5AC expression is associated with mismatch repair deficiency and proximal tumor location but not with cancer progression in colon cancer. *Med Mol Morphol*, (2020).
86. K. Bjorkman, H. Mustonen, T. Kaprio, C. Haglund, C. Bockelman, Mucin 16 and kallikrein 13 as potential prognostic factors in colon cancer: Results of an oncological 92-multiplex immunoassay. *Tumour Biol* **41**, 1010428319860728 (2019).
87. M. M. Streppel, A. Vincent, R. Mukherjee, N. R. Campbell, S. H. Chen, K. Konstantopoulos, M. G. Goggins, I. Van Seuningen, A. Maitra, E. A. Montgomery, Mucin 16 (cancer antigen 125) expression in human tissues and cell lines and correlation with clinical outcome in adenocarcinomas of the pancreas, esophagus, stomach, and colon. *Hum Pathol* **43**, 1755-1763 (2012).
88. D. A. Delker, B. M. McGettigan, P. Kanth, S. Pop, D. W. Neklason, M. P. Bronner, R. W. Burt, C. H. Hagedorn, RNA sequencing of sessile serrated colon polyps identifies differentially expressed genes and immunohistochemical markers. *PLoS One* **9**, e88367 (2014).

89. M. Khaidakov, K. K. Lai, D. Roudachevski, J. Sargsyan, H. E. Goyne, R. K. Pai, L. W. Lamps, C. H. Hagedorn, Gastric Proteins MUC5AC and TFF1 as Potential Diagnostic Markers of Colonic Sessile Serrated Adenomas/Polyps. *Am J Clin Pathol* **146**, 530-537 (2016).
90. S. R. Krishn, S. Kaur, Y. M. Sheinin, L. M. Smith, S. K. Gautam, A. Patel, M. Jain, V. Juvvigunta, P. Pai, A. J. Lazenby, H. K. Roy, S. K. Batra, Mucins and associated O-glycans based immunoprofile for stratification of colorectal polyps: clinical implication for improved colon surveillance. *Oncotarget* **8**, 7025-7038 (2017).
91. F. Renaud, A. Vincent, C. Mariette, M. Crepin, L. Stechly, S. Truant, M. C. Copin, N. Porchet, E. Leteurtre, I. Van Seuningen, M. P. Buisine, MUC5AC hypomethylation is a predictor of microsatellite instability independently of clinical factors associated with colorectal cancer. *Int J Cancer* **136**, 2811-2821 (2015).
92. J. Hu, J. Sun, MUC16 mutations improve patients' prognosis by enhancing the infiltration and antitumor immunity of cytotoxic T lymphocytes in the endometrial cancer microenvironment. *Oncoimmunology* **7**, e1487914 (2018).
93. E. C. Smyth, R. C. Fitzgerald, MUC16 Mutations and Prognosis in Gastric Cancer: A Little Goes a Long Way. *JAMA Oncol* **4**, 1698-1699 (2018).
94. X. Wang, X. Yu, M. Krauthammer, W. Hugo, C. Duan, P. A. Kanetsky, J. K. Teer, Z. J. Thompson, D. Kalos, K. Y. Tsai, K. S. M. Smalley, V. K. Sondak, Y. A. Chen, J. R. Conejo-Garcia, The Association of MUC16 Mutation with Tumor Mutation Burden and Its Prognostic Implications in Cutaneous Melanoma. *Cancer Epidemiol Biomarkers Prev* **29**, 1792-1799 (2020).
95. L. Zhang, X. Han, Y. Shi, Association of MUC16 Mutation With Response to Immune Checkpoint Inhibitors in Solid Tumors. *JAMA Netw Open* **3**, e2013201 (2020).
96. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424 (2018).
97. T. Matsuoka, M. Yashiro, Biomarkers of gastric cancer: Current topics and future perspective. *World J Gastroenterol* **24**, 2818-2832 (2018).
98. M. Bose, P. Mukherjee, Potential of Anti-MUC1 Antibodies as a Targeted Therapy for Gastrointestinal Cancers. *Vaccines (Basel)* **8**, (2020).
99. A. Aithal, S. Rauth, P. Kshirsagar, A. Shah, I. Lakshmanan, W. M. Junker, M. Jain, M. P. Ponnusamy, S. K. Batra, MUC16 as a novel target for cancer therapy. *Expert Opin Ther Targets* **22**, 675-686 (2018).
100. S. K. Gautam, S. Kumar, V. Dam, D. Ghersi, M. Jain, S. K. Batra, MUCIN-4 (MUC4) is a novel tumor antigen in pancreatic cancer immunotherapy. *Semin Immunol* **47**, 101391 (2020).
101. S. Kaur, L. M. Smith, A. Patel, M. Menning, D. C. Watley, S. S. Malik, S. R. Krishn, K. Mallya, A. Aithal, A. R. Sasson, S. L. Johansson, M. Jain, S. Singh, S. Guha, C. Are, M. Raimondo, M. A. Hollingsworth, R. E. Brand, S. K. Batra, A Combination of MUC5AC and CA19-9 Improves the Diagnosis

of Pancreatic Cancer: A Multicenter Study. *Am J Gastroenterol* **112**, 172-183 (2017).
102. K. Ganguly, S. R. Krishn, S. Rachagani, R. Jahan, A. Shah, P. Nallasamy, S. Rauth, P. Atri, J. L. Cox, R. Pothuraju, L. M. Smith, S. Ayala, C. Evans, M. P. Ponnusamy, S. Kumar, S. Kaur, S. K. Batra, Secretory Mucin 5AC Promotes Neoplastic Progression by Augmenting KLF4-Mediated Pancreatic Cancer Cell Stemness. *Cancer Res* **81**, 91-102 (2021).
103. J. L. McAuley, S. K. Linden, C. W. Png, R. M. King, H. L. Pennington, S. J. Gendler, T. H. Florin, G. R. Hill, V. Korolik, M. A. McGuckin, MUC1 cell surface mucin is a critical element of the mucosal barrier to infection. *J Clin Invest* **117**, 2313-2324 (2007).
104. O. Ilhan, U. Han, B. Onal, S. Y. Celik, Prognostic significance of MUC1, MUC2 and MUC5AC expressions in gastric carcinoma. *Turk J Gastroenterol* **21**, 345-352 (2010).
105. T. Shimamura, H. Ito, J. Shibahara, A. Watanabe, Y. Hippo, H. Taniguchi, Y. Chen, T. Kashima, T. Ohtomo, F. Tanioka, H. Iwanari, T. Kodama, T. Kazui, H. Sugimura, M. Fukayama, H. Aburatani, Overexpression of MUC13 is associated with intestinal-type gastric cancer. *Cancer Sci* **96**, 265-273 (2005).
106. L. He, L. Qu, L. Wei, Y. Chen, J. Suo, Reduction of miR1323p contributes to gastric cancer proliferation by targeting MUC13. *Mol Med Rep* **15**, 3055-3061 (2017).
107. X. Li, B. Pasche, W. Zhang, K. Chen, Association of MUC16 Mutation With Tumor Mutation Load and Outcomes in Patients With Gastric Cancer. *JAMA Oncol* **4**, 1691-1698 (2018).
108. G. T. Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
109. Z. Tang, C. Li, B. Kang, G. Gao, C. Li, Z. Zhang, GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* **45**, W98-W102 (2017).
110. M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
111. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. V. Mering, STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).
112. R. Aguirre-Gamboa, H. Gomez-Rueda, E. Martinez-Ledesma, A. Martinez-Torteya, R. Chacolla-Huaringa, A. Rodriguez-Barrientos, J. G. Tamez-Pena, V. Trevino, SurvExpress: an online biomarker validation tool and

database for cancer gene expression data using survival analysis. *PLoS One* **8**, e74250 (2013).
113. D. W. Kufe, Mucins in cancer: function, prognosis and therapy. *Nat Rev Cancer* **9**, 874-885 (2009).
114. D. B. Rubinstein, M. Karmely, E. Pichinuk, R. Ziv, I. Benhar, N. Feng, N. I. Smorodinsky, D. H. Wreschner, The MUC1 oncoprotein as a functional target: immunotoxin binding to alpha/beta junction mediates cell killing. *Int J Cancer* **124**, 46-54 (2009).
115. Y. Ge, G. Ma, H. Liu, Y. Lin, G. Zhang, M. Du, M. Wang, H. Chu, H. Zhang, Z. Zhang, MUC1 is associated with TFF2 methylation in gastric cancer. *Clin Epigenetics* **12**, 37 (2020).
116. J. Yang, Identification of novel biomarkers, MUC5AC, MUC1, KRT7, GAPDH, CD44 for gastric cancer. *Med Oncol* **37**, 34 (2020).
117. X. T. Wang, F. B. Kong, W. Mai, L. Li, L. M. Pang, MUC1 Immunohistochemical Expression as a Prognostic Factor in Gastric Cancer: Meta-Analysis. *Dis Markers* **2016**, 9421571 (2016).
118. K. Jiang, H. Liu, D. Xie, Q. Xiao, Differentially expressed genes ASPN, COL1A1, FN1, VCAN and MUC5AC are potential prognostic biomarkers for gastric cancer. *Oncol Lett* **17**, 3191-3202 (2019).
119. L. Giraldi, M. B. Michelazzo, D. Arzani, R. Persiani, R. Pastorino, S. Boccia, MUC1, MUC5AC, and MUC6 polymorphisms, Helicobacter pylori infection, and gastric cancer: a systematic review and meta-analysis. *Eur J Cancer Prev* **27**, 323-330 (2018).
120. A. Z. Khattab, W. A. Nasif, M. Lotfy, MUC2 and MUC6 apomucins expression in human gastric neoplasm: an immunohistochemical analysis. *Med Oncol* **28 Suppl 1**, S207-213 (2011).
121. Y. H. Sheng, S. Triyana, R. Wang, I. Das, K. Gerloff, T. H. Florin, P. Sutton, M. A. McGuckin, MUC1 and MUC13 differentially regulate epithelial inflammation in response to inflammatory and infectious stimuli. *Mucosal Immunol* **6**, 557-568 (2013).
122. S. D. Babu, V. Jayanthi, N. Devaraj, C. A. Reis, H. Devaraj, Expression profile of mucins (MUC2, MUC5AC and MUC6) in Helicobacter pylori infected pre-neoplastic and neoplastic human gastric epithelium. *Mol Cancer* **5**, 10 (2006).
123. D. Boltin, Y. Niv, Mucins in Gastric Cancer - An Update. *J Gastrointest Dig Syst* **3**, 15519 (2013).
124. C. A. Reis, L. David, P. A. Nielsen, H. Clausen, K. Mirgorodskaya, P. Roepstorff, M. Sobrinho-Simoes, Immunohistochemical study of MUC5AC expression in human gastric carcinomas using a novel monoclonal antibody. *Int J Cancer* **74**, 112-121 (1997).
125. Y. Jia, C. Persson, L. Hou, Z. Zheng, M. Yeager, J. Lissowska, S. J. Chanock, W. H. Chow, W. Ye, A comprehensive analysis of common genetic variation in MUC1, MUC5AC, MUC6 genes and risk of stomach cancer. *Cancer Causes Control* **21**, 313-321 (2010).
126. B. Yang, A. Wu, Y. Hu, C. Tao, J. M. Wang, Y. Lu, R. Xing, Mucin 17 inhibits the progression of human gastric cancer by limiting inflammatory responses

through a MYH9-p53-RhoA regulatory feedback loop. *J Exp Clin Cancer Res* **38**, 283 (2019).
127. S. Kaur, S. Kumar, N. Momi, A. R. Sasson, S. K. Batra, Mucins in pancreatic cancer and its microenvironment. *Nat Rev Gastroenterol Hepatol* **10**, 607-620 (2013).
128. H. S. Silverman, S. Parry, M. Sutton-Smith, M. D. Burdick, K. McDermott, C. J. Reid, S. K. Batra, H. R. Morris, M. A. Hollingsworth, A. Dell, A. Harris, In vivo glycosylation of mucin tandem repeats. *Glycobiology* **11**, 459-471 (2001).
129. J. P. M. van Putten, K. Strijbis, Transmembrane Mucins: Signaling Receptors at the Intersection of Inflammation and Cancer. *J Innate Immun* **9**, 281-299 (2017).
130. R. J. Quin, M. A. McGuckin, Phosphorylation of the cytoplasmic domain of the MUC1 mucin correlates with changes in cell-cell adhesion. *Int J Cancer* **87**, 499-506 (2000).
131. J. Wesseling, S. W. van der Valk, H. L. Vos, A. Sonnenberg, J. Hilkens, Episialin (MUC1) overexpression inhibits integrin-mediated cell adhesion to extracellular matrix components. *J Cell Biol* **129**, 255-265 (1995).
132. S. Bafna, S. Kaur, S. K. Batra, Membrane-bound mucins: the mechanistic basis for alterations in the growth and survival of cancer cells. *Oncogene* **29**, 2893-2904 (2010).
133. P. Pai, S. Rachagani, P. Dhawan, S. K. Batra, Mucins and Wnt/beta-catenin signaling in gastrointestinal cancers: an unholy nexus. *Carcinogenesis* **37**, 223-232 (2016).
134. S. Pan, R. Chen, Y. Tamura, D. A. Crispin, L. A. Lai, D. H. May, M. W. McIntosh, D. R. Goodlett, T. A. Brentnall, Quantitative glycoproteomics analysis reveals changes in N-glycosylation level associated with pancreatic ductal adenocarcinoma. *J Proteome Res* **13**, 1293-1306 (2014).
135. Y. Niv, T. Rokkas, Mucin Expression in Colorectal Cancer (CRC): Systematic Review and Meta-Analysis. *J Clin Gastroenterol*, (2018).
136. I. Lakshmanan, S. Salfity, P. Seshacharyulu, S. Rachagani, A. Thomas, S. Das, P. D. Majhi, R. K. Nimmakayala, R. Vengoji, S. M. Lele, M. P. Ponnusamy, S. K. Batra, A. K. Ganti, MUC16 Regulates TSPYL5 for Lung Cancer Cell Growth and Chemoresistance by Suppressing p53. *Clin Cancer Res* **23**, 3906-3917 (2017).
137. A. K. Bauer, M. Umer, V. L. Richardson, A. M. Cumpian, A. Q. Harder, N. Khosravi, Z. Azzegagh, N. M. Hara, C. Ehre, M. Mohebnasab, M. S. Caetano, D. T. Merrick, A. van Bokhoven, Wistuba, II, H. Kadara, B. F. Dickey, K. Velmurugan, P. R. Mann, X. Lu, A. E. Baron, C. M. Evans, S. J. Moghaddam, Requirement for MUC5AC in KRAS-dependent lung carcinogenesis. *JCI Insight* **3**, (2018).
138. T. Baert, J. Van Camp, L. Vanbrabant, P. Busschaert, A. Laenen, S. Han, E. Van Nieuwenhuysen, I. Vergote, A. Coosemans, Influence of CA125, platelet count and neutrophil to lymphocyte ratio on the immune system of ovarian cancer patients. *Gynecol Oncol* **150**, 31-37 (2018).

139. A. P. Singh, S. Senapati, M. P. Ponnusamy, M. Jain, S. M. Lele, J. S. Davis, S. Remmenga, S. K. Batra, Clinical potential of mucins in diagnosis, prognosis, and therapy of ovarian cancer. *Lancet Oncol* **9**, 1076-1085 (2008).
140. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).
141. J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, Introducing protein intrinsic disorder. *Chem Rev* **114**, 6561-6588 (2014).
142. M. Kjaergaard, B. B. Kragelund, Functions of intrinsic disorder in transmembrane proteins. *Cell Mol Life Sci* **74**, 3205-3224 (2017).
143. G. Hu, Z. Wu, V. N. Uversky, L. Kurgan, Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. *Int J Mol Sci* **18**, (2017).
144. J. Burgi, B. Xue, V. N. Uversky, F. G. van der Goot, Intrinsic Disorder in Transmembrane Proteins: Roles in Signaling and Topology Prediction. *PLoS One* **11**, e0158594 (2016).
145. W. Boomsma, S. V. Nielsen, K. Lindorff-Larsen, R. Hartmann-Petersen, L. Ellgaard, Bioinformatics analysis identifies several intrinsically disordered human E3 ubiquitin-protein ligases. *PeerJ* **4**, e1725 (2016).
146. M. Larion, B. Miller, R. Bruschweiler, Conformational heterogeneity and intrinsic disorder in enzyme regulation: Glucokinase as a case study. *Intrinsically Disord Proteins* **3**, e1011008 (2015).
147. V. N. Uversky, What does it mean to be natively unfolded? *Eur J Biochem* **269**, 2-12 (2002).
148. S. Lise, D. T. Jones, Sequence patterns associated with disordered regions in proteins. *Proteins* **58**, 144-150 (2005).
149. R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631 (2014).
150. M. M. Pentony, D. T. Jones, Modularity of intrinsic disorder in the human proteome. *Proteins* **78**, 212-221 (2010).
151. Y. J. Edwards, A. E. Lobley, M. M. Pentony, D. T. Jones, Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome Biol* **10**, R50 (2009).
152. A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, Z. Obradovic, Intrinsic disorder and protein function. *Biochemistry* **41**, 6573-6582 (2002).
153. S. L. Shammas, Mechanistic roles of protein disorder within transcription. *Curr Opin Struct Biol* **42**, 155-161 (2017).
154. V. N. Uversky, C. J. Oldfield, A. K. Dunker, Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* **18**, 343-384 (2005).
155. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**, 18-29 (2015).

156. V. Astro, I. de Curtis, Plasma membrane-associated platforms: dynamic scaffolds that organize membrane-associated events. *Sci Signal* **8**, re1 (2015).
157. M. Guharoy, B. Szabo, S. Contreras Martos, S. Kosol, P. Tompa, Intrinsic structural disorder in cytoskeletal proteins. *Cytoskeleton (Hoboken)* **70**, 550-571 (2013).
158. D. M. Mitrea, M. K. Yoon, L. Ou, R. W. Kriwacki, Disorder-function relationships for the cell cycle regulatory proteins p21 and p27. *Biol Chem* **393**, 259-274 (2012).
159. M. K. Yoon, D. M. Mitrea, L. Ou, R. W. Kriwacki, Cell cycle regulation by the intrinsically disordered proteins p21 and p27. *Biochem Soc Trans* **40**, 981-988 (2012).
160. A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, V. N. Uversky, Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* **272**, 5129-5148 (2005).
161. J. Gsponer, M. M. Babu, WITHDRAWN: The rules of disorder or why disorder rules. *Prog Biophys Mol Biol*, (2009).
162. C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal, L. M. Iakoucheva, Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* **2**, e100 (2006).
163. C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky, A. K. Dunker, Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**, S1 (2008).
164. A. Patil, K. Kinoshita, H. Nakamura, Hub promiscuity in protein-protein interaction networks. *Int J Mol Sci* **11**, 1930-1943 (2010).
165. B. Xue, C. J. Oldfield, Y. Y. Van, A. K. Dunker, V. N. Uversky, Protein intrinsic disorder and induced pluripotent stem cells. *Mol Biosyst* **8**, 134-150 (2012).
166. B. Xue, A. K. Dunker, V. N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**, 137-149 (2012).
167. V. N. Uversky, C. J. Oldfield, A. K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**, 215-246 (2008).
168. K. Rajagopalan, S. M. Mooney, N. Parekh, R. H. Getzenberg, P. Kulkarni, A majority of the cancer/testis antigens are intrinsically disordered proteins. *J Cell Biochem* **112**, 3256-3267 (2011).
169. V. N. Uversky, V. Dave, L. M. Iakoucheva, P. Malaney, S. J. Metallo, R. R. Pathak, A. C. Joerger, Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem Rev* **114**, 6844-6879 (2014).
170. Y. Shan, M. P. Eastwood, X. Zhang, E. T. Kim, A. Arkhipov, R. O. Dror, J. Jumper, J. Kuriyan, D. E. Shaw, Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* **149**, 860-870 (2012).

171. G. Dong, W. Chen, X. Wang, X. Yang, T. Xu, P. Wang, W. Zhang, Y. Rao, C. Miao, C. Sheng, Small Molecule Inhibitors Simultaneously Targeting Cancer Metabolism and Epigenetics: Discovery of Novel Nicotinamide Phosphoribosyltransferase (NAMPT) and Histone Deacetylase (HDAC) Dual Inhibitors. *J Med Chem* **60**, 7965-7983 (2017).
172. F. Merino, S. Raunser, Electron Cryo-microscopy as a Tool for Structure-Based Drug Development. *Angew Chem Int Ed Engl* **56**, 2846-2860 (2017).
173. S. Goyal, S. Jamal, A. Shanker, A. Grover, Structural investigations of T854A mutation in EGFR and identification of novel inhibitors using structure activity relationships. *BMC Genomics* **16 Suppl 5**, S8 (2015).
174. G. M. Verkhivker, Computational Modeling of the Hsp90 Interactions with Cochaperones and Small-Molecule Inhibitors. *Methods Mol Biol* **1709**, 253-273 (2018).
175. T. Pelaseyed, M. Zach, A. C. Petersson, F. Svensson, D. G. Johansson, G. C. Hansson, Unfolding dynamics of the mucin SEA domain probed by force spectroscopy suggest that it acts as a cell-protective device. *FEBS J* **280**, 1491-1501 (2013).
176. T. Maeda, M. Inoue, S. Koshiba, T. Yabuki, M. Aoki, E. Nunokawa, E. Seki, T. Matsuda, Y. Motoda, A. Kobayashi, F. Hiroyasu, M. Shirouzu, T. Terada, N. Hayami, Y. Ishizuka, N. Shinya, A. Tatsuguchi, M. Yoshida, H. Hirota, Y. Matsuo, K. Tani, T. Arakawa, P. Carninci, J. Kawai, Y. Hayashizaki, T. Kigawa, S. Yokoyama, Solution structure of the SEA domain from the murine homologue of ovarian cancer antigen CA125 (MUC16). *J Biol Chem* **279**, 13174-13182 (2004).
177. M. E. Oates, P. Romero, T. Ishida, M. Ghalwash, M. J. Mizianty, B. Xue, Z. Dosztanyi, V. N. Uversky, Z. Obradovic, L. Kurgan, A. K. Dunker, J. Gough, D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* **41**, D508-516 (2013).
178. J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, J. L. Sussman, FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435-3438 (2005).
179. A. Marchler-Bauer, S. H. Bryant, CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327-331 (2004).
180. T. Lang, S. Klasson, E. Larsson, M. E. Johansson, G. C. Hansson, T. Samuelsson, Searching the Evolutionary Origin of Epithelial Mucus Protein Components-Mucins and FCGBP. *Mol Biol Evol* **33**, 1921-1936 (2016).
181. R. Sharma, S. Kumar, T. Tsunoda, A. Patil, A. Sharma, Predicting MoRFs in protein sequences using HMM profiles. *BMC Bioinformatics* **17**, 504 (2016).
182. P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520 (2015).
183. G. Y. Chuang, J. C. Boyington, M. G. Joyce, J. Zhu, G. J. Nabel, P. D. Kwong, I. Georgiev, Computational prediction of N-linked glycosylation

incorporating structural properties and patterns. *Bioinformatics* **28**, 2249-2255 (2012).
184. C. Steentoft, S. Y. Vakhrushev, H. J. Joshi, Y. Kong, M. B. Vester-Christensen, K. T. Schjoldager, K. Lavrsen, S. Dabelsteen, N. B. Pedersen, L. Marcos-Silva, R. Gupta, E. P. Bennett, U. Mandel, S. Brunak, H. H. Wandall, S. B. Levery, H. Clausen, Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J* **32**, 1478-1488 (2013).
185. B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, V. N. Uversky, PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**, 996-1010 (2010).
186. A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B. J. Breitkreutz, K. Dolinski, M. Tyers, The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
187. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
188. M. L. Hale, I. Thapa, D. Ghersi, FunSet: an open-source software and web server for performing and displaying Gene Ontology enrichment analysis. *BMC Bioinformatics* **20**, 359 (2019).
189. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141 (2004).
190. A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, S. H. Bryant, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**, D200-D203 (2017).
191. I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann, S. C. Tosatto, Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* **31**, 201-208 (2015).
192. V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, A. K. Dunker, Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* **6**, 2351-2366 (2007).
193. N. Malhis, E. T. Wong, R. Nassar, J. Gsponer, Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. *PLoS One* **10**, e0141603 (2015).
194. A. Kurotani, T. Sakurai, In Silico Analysis of Correlations between Protein Disorder and Post-Translational Modifications in Algae. *Int J Mol Sci* **16**, 19812-19835 (2015).

195. W. Basile, M. Salvatore, C. Bassot, A. Elofsson, Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol* **15**, e1007186 (2019).
196. E. E. Pryor, Jr., M. C. Wiener, A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder. *Biophys J* **106**, 1638-1649 (2014).
197. J. W. Chen, P. Romero, V. N. Uversky, A. K. Dunker, Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J Proteome Res* **5**, 888-898 (2006).
198. P. Tompa, Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527-533 (2002).
199. R. Y. Wang, L. Chen, H. Y. Chen, L. Hu, L. Li, H. Y. Sun, F. Jiang, J. Zhao, G. M. Liu, J. Tang, C. Y. Chen, Y. C. Yang, Y. X. Chang, H. Liu, J. Zhang, Y. Yang, G. Huang, F. Shen, M. C. Wu, W. P. Zhou, H. Y. Wang, MUC15 inhibits dimerization of EGFR and PI3K-AKT signaling and is associated with aggressive hepatocellular carcinomas in patients. *Gastroenterology* **145**, 1436-1448 e1431-1412 (2013).
200. P. Chaturvedi, A. P. Singh, S. Chakraborty, S. C. Chauhan, S. Bafna, J. L. Meza, P. K. Singh, M. A. Hollingsworth, P. P. Mehta, S. K. Batra, MUC4 mucin interacts with and stabilizes the HER2 oncoprotein in human pancreatic cancer cells. *Cancer Res* **68**, 2065-2070 (2008).
201. H. Boze, T. Marlin, D. Durand, J. Perez, A. Vernhet, F. Canon, P. Sarni-Manchado, V. Cheynier, B. Cabane, Proline-rich salivary proteins have extended conformations. *Biophys J* **99**, 656-665 (2010).
202. Y. Lin, S. L. Currie, M. K. Rosen, Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J Biol Chem* **292**, 19110-19120 (2017).
203. A. Tarakhovsky, R. K. Prinjha, Drawing on disorder: How viruses use histone mimicry to their advantage. *J Exp Med* **215**, 1777-1787 (2018).
204. E. R. Tamarozzi, S. Giuliatti, Understanding the Role of Intrinsic Disorder of Viral Proteins in the Oncogenicity of Different Types of HPV. *Int J Mol Sci* **19**, (2018).
205. J. Charon, A. Barra, J. Walter, P. Millot, E. Hebrard, B. Moury, T. Michon, First Experimental Assessment of Protein Intrinsic Disorder Involvement in an RNA Virus Natural Adaptive Process. *Mol Biol Evol* **35**, 38-49 (2018).
206. D. Raina, P. Agarwal, J. Lee, A. Bharti, C. J. McKnight, P. Sharma, S. Kharbanda, D. Kufe, Characterization of the MUC1-C Cytoplasmic Domain as a Cancer Target. *PLoS One* **10**, e0135156 (2015).
207. G. Machkalyan, P. Trieu, D. Petrin, T. E. Hebert, G. J. Miller, PPIP5K1 interacts with the exocyst complex through a C-terminal intrinsically disordered domain and regulates cell motility. *Cell Signal* **28**, 401-411 (2016).
208. E. M. Marcotte, M. Tsechansky, Disorder, promiscuity, and toxic partnerships. *Cell* **138**, 16-18 (2009).

209. T. Vavouri, J. I. Semple, R. Garcia-Verdugo, B. Lehner, Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198-208 (2009).
210. I. Lakshmanan, P. Seshacharyulu, D. Haridas, S. Rachagani, S. Gupta, S. Joshi, C. Guda, Y. Yan, M. Jain, A. K. Ganti, M. P. Ponnusamy, S. K. Batra, Novel HER3/MUC4 oncogenic signaling aggravates the tumorigenic phenotypes of pancreatic cancer cells. *Oncotarget* **6**, 21085-21099 (2015).
211. M. S. A. Yamashita, E. O. Melo, Mucin 2 (MUC2) promoter characterization: an overview. *Cell Tissue Res* **374**, 455-463 (2018).
212. P. K. Singh, M. A. Hollingsworth, Cell surface-associated mucins in signal transduction. *Trends Cell Biol* **16**, 467-476 (2006).
213. S. Muniyan, D. Haridas, S. Chugh, S. Rachagani, I. Lakshmanan, S. Gupta, P. Seshacharyulu, L. M. Smith, M. P. Ponnusamy, S. K. Batra, MUC16 contributes to the metastasis of pancreatic ductal adenocarcinoma through focal adhesion mediated signaling mechanism. *Genes Cancer* **7**, 110-124 (2016).
214. S. Das, S. Rachagani, M. P. Torres-Gonzalez, I. Lakshmanan, P. D. Majhi, L. M. Smith, K. U. Wagner, S. K. Batra, Carboxyl-terminal domain of MUC16 imparts tumorigenic and metastatic functions through nuclear translocation of JAK2 to pancreatic cancer cells. *Oncotarget* **6**, 5772-5787 (2015).
215. A. Rump, Y. Morikawa, M. Tanaka, S. Minami, N. Umesaki, M. Takeuchi, A. Miyajima, Binding of ovarian cancer antigen CA125/MUC16 to mesothelin mediates cell adhesion. *J Biol Chem* **279**, 9190-9198 (2004).
216. S. H. Chen, M. R. Dallas, E. M. Balzer, K. Konstantopoulos, Mucin 16 is a functional selectin ligand on pancreatic cancer cells. *FASEB J* **26**, 1349-1359 (2012).
217. J. Zhou, S. Zhao, A. K. Dunker, Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. *J Mol Biol* **430**, 2342-2359 (2018).
218. H. Hoshi, T. Sawada, M. Uchida, H. Iijima, K. Kimura, K. Hirakawa, H. Wanibuchi, MUC5AC protects pancreatic cancer cells from TRAIL-induced death pathways. *Int J Oncol* **42**, 887-893 (2013).
219. C. Ridley, D. J. Thornton, Mucins: the frontline defence of the lung. *Biochem Soc Trans* **46**, 1099-1106 (2018).
220. S. Senapati, S. Das, S. K. Batra, Mucin-interacting proteins: from function to therapeutics. *Trends Biochem Sci* **35**, 236-245 (2010).
221. P. W. Lo, J. J. Shie, C. H. Chen, C. Y. Wu, T. L. Hsu, C. H. Wong, O-GlcNAcylation regulates the stability and enzymatic activity of the histone methyltransferase EZH2. *Proc Natl Acad Sci U S A* **115**, 7302-7307 (2018).
222. R. van der Lee, B. Lang, K. Kruse, J. Gsponer, N. Sanchez de Groot, M. A. Huynen, A. Matouschek, M. Fuxreiter, M. M. Babu, Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep* **8**, 1832-1844 (2014).
223. Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky, A. K. Dunker, Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* **46**, 13468-13477 (2007).

224. V. N. Uversky, p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. *Int J Mol Sci* **17**, (2016).
225. L. R. Yadav, S. Rai, M. V. Hosur, A. K. Varma, Functional assessment of intrinsic disorder central domains of BRCA1. *J Biomol Struct Dyn* **33**, 2469-2478 (2015).
226. P. Malaney, V. N. Uversky, V. Dave, Identification of intrinsically disordered regions in PTEN and delineation of its function via a network approach. *Methods* **77-78**, 69-74 (2015).
227. R. Salgia, M. K. Jolly, T. Dorff, C. Lau, K. Weninger, J. Orban, P. Kulkarni, Prostate-Associated Gene 4 (PAGE4): Leveraging the Conformational Dynamics of a Dancing Protein Cloud as a Therapeutic Target. *J Clin Med* **7**, (2018).
228. S. J. Metallo, Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**, 481-488 (2010).
229. Y. S. Chang, B. Graves, V. Guerlavais, C. Tovar, K. Packman, K. H. To, K. A. Olson, K. Kesavan, P. Gangurde, A. Mukherjee, T. Baker, K. Darlak, C. Elkin, Z. Filipovic, F. Z. Qureshi, H. Cai, P. Berry, E. Feyfant, X. E. Shi, J. Horstick, D. A. Annis, A. M. Manning, N. Fotouhi, H. Nash, L. T. Vassilev, T. K. Sawyer, Stapled alpha-helical peptide drug development: a potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc Natl Acad Sci U S A* **110**, E3445-3454 (2013).
230. V. P. Balachandran, M. Luksza, J. N. Zhao, V. Makarov, J. A. Moral, R. Remark, B. Herbst, G. Askan, U. Bhanot, Y. Senbabaoglu, D. K. Wells, C. I. O. Cary, O. Grbovic-Huezo, M. Attiyeh, B. Medina, J. Zhang, J. Loo, J. Saglimbeni, M. Abu-Akeel, R. Zappasodi, N. Riaz, M. Smoragiewicz, Z. L. Kelley, O. Basturk, I. Australian Pancreatic Cancer Genome, R. Garvan Institute of Medical, H. Prince of Wales, H. Royal North Shore, G. University of, H. St Vincent's, Q. B. M. R. Institute, C. f. C. R. University of Melbourne, I. f. M. B. University of Queensland, H. Bankstown, H. Liverpool, C. O. B. L. Royal Prince Alfred Hospital, H. Westmead, H. Fremantle, H. St John of God, H. Royal Adelaide, C. Flinders Medical, P. Envoi, H. Princess Alexandria, H. Austin, I. Johns Hopkins Medical, A. R.-N. C. f. A. R. o. Cancer, M. Gonen, A. J. Levine, P. J. Allen, D. T. Fearon, M. Merad, S. Gnjatic, C. A. Iacobuzio-Donahue, J. D. Wolchok, R. P. DeMatteo, T. A. Chan, B. D. Greenbaum, T. Merghoub, S. D. Leach, Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551**, 512-516 (2017).
231. D. Chivian, D. E. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. E. Strauss, R. Bonneau, C. A. Rohl, D. Baker, Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53 Suppl 6**, 524-533 (2003).
232. G. Pollastri, D. Przybylski, B. Rost, P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-235 (2002).

233. S. Wojcik, M. Birol, E. Rhoades, A. D. Miranker, Z. A. Levine, Targeting the Intrinsically Disordered Proteome Using Small-Molecule Ligands. *Methods Enzymol* **611**, 703-734 (2018).
234. J. L. Neira, J. Bintz, M. Arruebo, B. Rizzuti, T. Bonacci, S. Vega, A. Lanas, A. Velazquez-Campoy, J. L. Iovanna, O. Abian, Identification of a Drug Targeting an Intrinsically Disordered Protein Involved in Pancreatic Adenocarcinoma. *Sci Rep* **7**, 39732 (2017).
235. C. Yu, X. Niu, F. Jin, Z. Liu, C. Jin, L. Lai, Structure-based Inhibitor Design for the Intrinsically Disordered Protein c-Myc. *Sci Rep* **6**, 22298 (2016).
236. K. Y. Jung, H. Wang, P. Teriete, J. L. Yap, L. Chen, M. E. Lanning, A. Hu, L. J. Lambert, T. Holien, A. Sundan, N. D. Cosford, E. V. Prochownik, S. Fletcher, Perturbation of the c-Myc-Max protein-protein interaction via synthetic alpha-helix mimetics. *J Med Chem* **58**, 3002-3024 (2015).
237. S. Ambadipudi, M. Zweckstetter, Targeting intrinsically disordered proteins in rational drug discovery. *Expert Opin Drug Discov* **11**, 65-77 (2016).
238. K. Tsafou, P. B. Tiwari, J. D. Forman-Kay, S. J. Metallo, J. A. Toretsky, Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm. *J Mol Biol* **430**, 2321-2341 (2018).
239. L. Moletta, S. Serafini, M. Valmasoni, E. S. Pierobon, A. Ponzoni, C. Sperti, Surgery for Recurrent Pancreatic Cancer: Is It Effective? *Cancers (Basel)* **11**, (2019).
240. A. Cannon, C. Thompson, B. R. Hall, M. Jain, S. Kumar, S. K. Batra, Desmoplasia in pancreatic ductal adenocarcinoma: insight into pathological function and therapeutic potential. *Genes Cancer* **9**, 78-86 (2018).
241. A. N. Hosein, R. A. Brekken, A. Maitra, Pancreatic cancer stroma: an update on therapeutic targeting strategies. *Nat Rev Gastroenterol Hepatol* **17**, 487-505 (2020).
242. B. R. Hall, A. Cannon, P. Atri, C. S. Wichman, L. M. Smith, A. K. Ganti, C. Are, A. R. Sasson, S. Kumar, S. K. Batra, Advanced pancreatic cancer: a meta-analysis of clinical trials over thirty years. *Oncotarget* **9**, 19396-19405 (2018).
243. T. Conroy, P. Hammel, M. Hebbar, M. Ben Abdelghani, A. C. Wei, J. L. Raoul, L. Chone, E. Francois, P. Artru, J. J. Biagi, T. Lecomte, E. Assenat, R. Faroux, M. Ychou, J. Volet, A. Sauvanet, G. Breysacher, F. Di Fiore, C. Cripps, P. Kavan, P. Texereau, K. Bouhier-Leporrier, F. Khemissa-Akouz, J. L. Legoux, B. Juzyna, S. Gourgou, C. J. O'Callaghan, C. Jouffroy-Zeller, P. Rat, D. Malka, F. Castan, J. B. Bachet, G. Canadian Cancer Trials, G. I. P. G. the Unicancer, FOLFIRINOX or Gemcitabine as Adjuvant Therapy for Pancreatic Cancer. *N Engl J Med* **379**, 2395-2406 (2018).
244. D. D. Von Hoff, T. Ervin, F. P. Arena, E. G. Chiorean, J. Infante, M. Moore, T. Seay, S. A. Tjulandin, W. W. Ma, M. N. Saleh, M. Harris, M. Reni, S. Dowden, D. Laheru, N. Bahary, R. K. Ramanathan, J. Tabernero, M. Hidalgo, D. Goldstein, C. E. Van, X. Wei, J. Iglesias, M. F. Renschler, Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N. Engl. J. Med* **369**, 1691-1703 (2013).

245. R. K. Sahoo, L. Kumar, Albumin-bound paclitaxel plus gemcitabine in pancreatic cancer. *N. Engl. J. Med* **370**, 478-479 (2014).
246. M. W. Saif, U.S. Food and Drug Administration approves paclitaxel protein-bound particles (Abraxane(R)) in combination with gemcitabine as first-line treatment of patients with metastatic pancreatic cancer. *JOP* **14**, 686-688 (2013).
247. S. Zeng, M. Pottler, B. Lan, R. Grutzmann, C. Pilarsky, H. Yang, Chemoresistance in Pancreatic Cancer. *Int J Mol Sci* **20**, (2019).
248. M. G. Rees, B. Seashore-Ludlow, J. H. Cheah, D. J. Adams, E. V. Price, S. Gill, S. Javaid, M. E. Coletti, V. L. Jones, N. E. Bodycombe, C. K. Soule, B. Alexander, A. Li, P. Montgomery, J. D. Kotz, C. S. Hon, B. Munoz, T. Liefeld, V. Dancik, D. A. Haber, C. B. Clish, J. A. Bittker, M. Palmer, B. K. Wagner, P. A. Clemons, A. F. Shamji, S. L. Schreiber, Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* **12**, 109-116 (2016).
249. B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, B. Alexander, A. Li, P. Montgomery, M. J. Wawer, N. Kuru, J. D. Kotz, C. S. Hon, B. Munoz, T. Liefeld, V. Dancik, J. A. Bittker, M. Palmer, J. E. Bradner, A. F. Shamji, P. A. Clemons, S. L. Schreiber, Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov* **5**, 1210-1223 (2015).
250. A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, A. L. Bracha, T. Liefeld, M. Wawer, J. C. Gilbert, A. J. Wilson, N. Stransky, G. V. Kryukov, V. Dancik, J. Barretina, L. A. Garraway, C. S. Hon, B. Munoz, J. A. Bittker, B. R. Stockwell, D. Khabele, A. M. Stern, P. A. Clemons, A. F. Shamji, S. L. Schreiber, An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151-1161 (2013).
251. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, T. R. Golub, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935 (2006).
252. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W. N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, T. R. Golub, A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452 e1417 (2017).

253. T. Seino, S. Kawasaki, M. Shimokawa, H. Tamagawa, K. Toshimitsu, M. Fujii, Y. Ohta, M. Matano, K. Nanki, K. Kawasaki, S. Takahashi, S. Sugimoto, E. Iwasaki, J. Takagi, T. Itoi, M. Kitago, Y. Kitagawa, T. Kanai, T. Sato, Human Pancreatic Tumor Organoids Reveal Loss of Stem Cell Niche Factor Dependence during Disease Progression. *Cell Stem Cell* **22**, 454-467 e456 (2018).
254. P. A. Toste, L. Li, B. E. Kadera, A. H. Nguyen, L. M. Tran, N. Wu, D. L. Madnick, S. G. Patel, D. W. Dawson, T. R. Donahue, p85alpha is a microRNA target and affects chemosensitivity in pancreatic cancer. *J Surg Res* **196**, 285-293 (2015).
255. K. A. Ellsworth, B. W. Eckloff, L. Li, I. Moon, B. L. Fridley, G. D. Jenkins, E. Carlson, A. Brisbin, R. Abo, W. Bamlet, G. Petersen, E. D. Wieben, L. Wang, Contribution of FKBP5 genetic variation to gemcitabine treatment and survival in pancreatic adenocarcinoma. *PLoS One* **8**, e70216 (2013).
256. T. Idichi, N. Seki, H. Kurahara, K. Yonemori, Y. Osako, T. Arai, A. Okato, Y. Kita, T. Arigami, Y. Mataki, Y. Kijima, K. Maemura, S. Natsugoe, Regulation of actin-binding protein ANLN by antitumor miR-217 inhibits cancer cell aggressiveness in pancreatic ductal adenocarcinoma. *Oncotarget* **8**, 53180-53193 (2017).
257. G. Sergeant, R. van Eijsden, T. Roskams, V. Van Duppen, B. Topal, Pancreatic cancer circulating tumour cells express a cell motility gene signature that predicts survival after surgery. *BMC Cancer* **12**, 527 (2012).
258. S. Gysin, J. Paquette, M. McMahon, Analysis of mRNA profiles after MEK1/2 inhibition in human pancreatic cancer cell lines reveals pathways involved in drug sensitivity. *Mol Cancer Res* **10**, 1607-1619 (2012).
259. T. E. Newhook, E. M. Blais, J. M. Lindberg, S. J. Adair, W. Xin, J. K. Lee, J. A. Papin, J. T. Parsons, T. W. Bauer, A thirteen-gene expression signature predicts survival of patients with pancreatic cancer and identifies new genes of interest. *PLoS One* **9**, e105631 (2014).
260. M. P. Torres, S. Rachagani, J. J. Souchek, K. Mallya, S. L. Johansson, S. K. Batra, Novel pancreatic cancer cell lines derived from genetically engineered mouse models of spontaneous pancreatic adenocarcinoma: applications in diagnosis and therapy. *PLoS. One* **8**, e80580 (2013).
261. K. M. Lee, C. Nguyen, A. B. Ulrich, P. M. Pour, M. M. Ouellette, Immortalization with telomerase of the Nestin-positive cells of the human pancreas. *Biochem Biophys Res Commun* **301**, 1038-1044 (2003).
262. T. Conroy, F. Desseigne, M. Ychou, O. Bouche, R. Guimbaud, Y. Becouarn, A. Adenis, J. L. Raoul, S. Gourgou-Bourgade, C. de la Fouchardiere, J. Bennouna, J. B. Bachet, F. Khemissa-Akouz, D. Pere-Verge, C. Delbaldo, E. Assenat, B. Chauffert, P. Michel, C. Montoto-Grillot, M. Ducreux, U. Groupe Tumeurs Digestives of, P. Intergroup, FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med* **364**, 1817-1825 (2011).
263. R. Kim, FOLFIRINOX: a new standard treatment for advanced pancreatic cancer? *Lancet Oncol* **12**, 8-9 (2011).

264. W. B. Wang, Y. Yang, Y. P. Zhao, T. P. Zhang, Q. Liao, H. Shu, Recent studies of 5-fluorouracil resistance in pancreatic cancer. *World J Gastroenterol* **20**, 15682-15690 (2014).
265. A. P. Kozikowski, S. Tapadar, D. N. Luchini, K. H. Kim, D. D. Billadeau, Use of the nitrile oxide cycloaddition (NOC) reaction for molecular probe generation: a new class of enzyme selective histone deacetylase inhibitors (HDACIs) showing picomolar activity at HDAC6. *J Med Chem* **51**, 4370-4373 (2008).
266. K. V. Butler, J. Kalin, C. Brochier, G. Vistoli, B. Langley, A. P. Kozikowski, Rational design and simple chemistry yield a superior, neuroprotective HDAC6 inhibitor, tubastatin A. *J Am Chem Soc* **132**, 10842-10846 (2010).
267. Y. Li, E. Seto, HDACs and HDAC Inhibitors in Cancer Development and Therapy. *Cold Spring Harb Perspect Med* **6**, (2016).
268. L. Hontecillas-Prieto, R. Flores-Campos, A. Silver, E. de Álava, N. Hajji, D. J. García-Domínguez, Synergistic Enhancement of Cancer Therapy Using HDAC Inhibitors: Opportunity for Clinical Trials. *Frontiers in Genetics* **11**, (2020).
269. M. Mrakovcic, L. F. Frohlich, Molecular Determinants of Cancer Therapy Resistance to HDAC Inhibitor-Induced Autophagy. *Cancers (Basel)* **12**, (2019).
270. S. Ropero, M. Esteller, The role of histone deacetylases (HDACs) in human cancer. *Mol Oncol* **1**, 19-25 (2007).
271. H. J. Kim, S. C. Bae, Histone deacetylase inhibitors: molecular mechanisms of action and clinical trials as anti-cancer drugs. *Am J Transl Res* **3**, 166-179 (2011).
272. M. Haberland, R. L. Montgomery, E. N. Olson, The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat Rev Genet* **10**, 32-42 (2009).
273. G. I. Aldana-Masangkay, K. M. Sakamoto, The role of HDAC6 in cancer. *J Biomed Biotechnol* **2011**, 875824 (2011).
274. B. G. Bitler, S. Wu, P. H. Park, Y. Hai, K. M. Aird, Y. Wang, Y. Zhai, A. V. Kossenkov, A. Vara-Ailor, F. J. Rauscher, III, W. Zou, D. W. Speicher, D. G. Huntsman, J. R. Conejo-Garcia, K. R. Cho, D. W. Christianson, R. Zhang, ARID1A-mutated ovarian cancers depend on HDAC6 activity. *Nat Cell Biol* **19**, 962-973 (2017).
275. X. F. Lu, X. Y. Cao, Y. J. Zhu, Z. R. Wu, X. Zhuang, M. Y. Shao, Q. Xu, Y. J. Zhou, H. J. Ji, Q. R. Lu, Y. J. Shi, Y. Zeng, H. Bu, Histone deacetylase 3 promotes liver regeneration and liver cancer cells proliferation through signal transducer and activator of transcription 3 signaling pathway. *Cell Death Dis* **9**, 398 (2018).
276. Y. Jiang, J. Hsieh, HDAC3 controls gap 2/mitosis progression in adult neural stem/progenitor cells by regulating CDK1 levels. *Proc Natl Acad Sci U S A* **111**, 13541-13546 (2014).
277. Y. Li, X. Zhang, S. Zhu, E. A. Dejene, W. Peng, A. Sepulveda, E. Seto, HDAC10 Regulates Cancer Stem-Like Cell Properties in KRAS-Driven Lung Adenocarcinoma. *Cancer Res* **80**, 3265-3278 (2020).

278. A. Nebbioso, V. Carafa, M. Conte, F. P. Tambaro, C. Abbondanza, J. Martens, M. Nees, R. Benedetti, I. Pallavicini, S. Minucci, G. Garcia-Manero, F. Iovino, G. Lania, C. Ingenito, V. Belsito Petrizzi, H. G. Stunnenberg, L. Altucci, c-Myc Modulation and Acetylation Is a Key HDAC Inhibitor Target in Cancer. *Clin Cancer Res* **23**, 2542-2555 (2017).
279. N. Yu, P. Chen, Q. Wang, M. Liang, J. Qiu, P. Zhou, M. Yang, P. Yang, Y. Wu, X. Han, J. Ge, J. Zhuang, K. Yu, Histone deacetylase inhibitors differentially regulate c-Myc expression in retinoblastoma cells. *Oncol Lett* **19**, 460-468 (2020).
280. R. Coelho, L. Marcos-Silva, S. Ricardo, F. Ponte, A. Costa, J. M. Lopes, L. David, Peritoneal dissemination of ovarian cancer: role of MUC16-mesothelin interaction and implications for treatment. *Expert Rev Anticancer Ther* **18**, 177-186 (2018).