Leveraging Big Data Resources and Data Integration in Biology: Applying Computational Systems Analyses and Machine Learning to Gain Insights into the Biology of Cancers

by

Musalula Sinkala

SNKMUS003

A thesis submitted to the graduate faculty

in partial fulfilment of the requirements for the degree of Doctor of Philosophy

PhD Bioinformatics

Faculty of Health Sciences

UNIVERSITY OF CAPE TOWN

October 2019

Supervisor: **Darren Martin**, Department of Integrative Biomedical Sciences, University of Cape Town

Co-Supervisors: **Nicola Mulder**, Department of Integrative Biomedical Sciences, University of Cape Town

Stefan Barth, Department of Integrative Biomedical Sciences, University of Cape Town

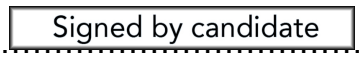
Declaration

I declare that this thesis is entirely my own and only aided by the guidance of my Supervisors. We have published one paper during the preparation of this thesis, a second is under peer-review, and a third has been provisionally accepted for publication. This thesis is a reformatting of the published and submitted manuscripts. I am the lead author on each of these manuscripts and took responsibility for abstracting the methodology, data analysis, writing of the original draft manuscript, reviewing and editing, and visualization.

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publication(s) in my thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publication(s):

1. Sinkala M, Mulder N, Martin DP. Integrative landscape of dysregulated signaling pathways of clinically distinct pancreatic cancer subtypes. Oncotarget 2018;9:29123–39. doi:10.18632/oncotarget.25632.
2. Sinkala, Musalula, Nicola Mulder, and Darren Martin. "Machine Learning and network Analyses Reveal Disease Subtypes of pancreatic cancer and their Molecular characteristics." *Scientific Reports* 10.1 (2020): 1-14. https://doi.org/10.1038/s41598-020-58290-2 .
3. Sinkala, Musalula, Nicola Mulder, and Darren Patrick Martin. "Metabolic gene alterations impact the clinical aggressiveness and drug responses of 32 human cancers." *Communications biology* 2.1 (2019): 1-14. https://doi.org/10.1038/s42003-019-0666-1 .

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: …  …

Date: 31st August 2019

Dedication

I would like to dedicate this work to my mother, Beatrice Nambela. She has provided me with a lifetime of love, which has brought me to this day. Also, I would like to dedicate this work to my siblings, Cornelius, Hellen, Paul, Margaret, and Kelvin, who have been dear to my heart. They have shared every victory and sacrifice with me throughout my life. I also thank my many friends, grandma Nakaundi and family for their love and support over the years. Finally, I dedicate this work to Kapambwe, who was both a brother and friend who loved me the most. Grandma Nakaundi and Kapambwe left us to be with the Lord, but their love will be with us always.

Acknowledgements

First and foremost, I thank my supervisors, Darren Martin and Nicola Mulder – for offering me this PhD Position and their willingness to provide me advice and support whenever I needed it, and for setting a great example in both a personal and academic capacity. Darren and Nicola, I am indebted to you for the confidence you gave me to do what was difficult and for the numerous opportunities that you opened for me my during research. I would also like to thank my co-supervisor, Professor Stefan Bath, for your guidance, assistance and drive for excellence, which is contagious. I appreciate all my fellow post-graduate students in the Computational Biology laboratory for creating a friendly and conducive environment even when things were not going well. I would also like to thank my collaborators for the Barth Lab, Krupa Naran and Suresh Madheswaran for their valuable advice concerning my project.

Abstract

Recently, many “molecular profiling” projects have yielded vast amounts of genetic, epigenetic, transcription, protein expression, metabolic and drug response data for cancerous tumours, healthy tissues, and cell lines.

We aim to facilitate a multi-scale understanding of these high-dimensional biological data and the complexity of the relationships between the different data types taken from human tumours. Further, we intend to identify molecular disease subtypes of various cancers, uncover the subtype-specific drug targets and identify sets of therapeutic molecules that could potentially be used to inhibit these targets.

We collected data from over 20 publicly available resources. We then leverage integrative computational systems analyses, network analyses and machine learning, to gain insights into the pathophysiology of pancreatic cancer and 32 other human cancer types.

Here, we uncover aberrations in multiple cell signalling and metabolic pathways that implicate regulatory kinases and the Warburg effect as the likely drivers of the distinct molecular signatures of three established pancreatic cancer subtypes. Then, we apply an integrative clustering method to four different types of molecular data to reveal that pancreatic tumours can be segregated into two distinct subtypes. We define sets of proteins, mRNAs, miRNAs and DNA methylation patterns that could serve as biomarkers to accurately differentiate between the two pancreatic cancer subtypes. Then we confirm the biological relevance of the identified biomarkers by showing that these can be used together with pattern-recognition algorithms to infer the drug sensitivity of pancreatic cancer cell lines accurately.

Further, we evaluate the alterations of metabolic pathway genes across 32 human cancers. We find that while alterations of metabolic genes are pervasive across all human cancers, the extent of these gene alterations varies between them. Based on these gene alterations, we define two distinct cancer supertypes that tend to be associated with different clinical outcomes and show that these supertypes are likely to respond differently to anticancer drugs.

Overall, we show that the time has already arrived where we can mine available data resources to potentially elicit more precise and personalised cancer therapies that would yield better clinical outcomes at a much lower cost than is currently being achieved.

Table of Contents

Declaration I

Dedication II

Acknowledgements III

Abstract IV

Table of Contents V

List of Tables VIII

List of Figures IX

List of Abbreviations XI

Chapter 1 : General Introduction *1*

1.1 Introduction 1

1.2 Thesis Organization 10

Chapter 1 10
Chapter 2 10
Chapter 3 11
Chapter 4 11
Chapter 5 12

Chapter 2 : Integrative Landscape of Dysregulated Signalling Pathways of Clinically Distinct Pancreatic Cancer Subtypes *13*

2.1 Introduction 13

2.2 Results 15

2.2.1 Pancreatic cancer subtypes display distinctive clinical outcomes 15
2.2.2 Gene expression and pathway characteristics of different PDAC subtypes 16
2.2.3 The gene alterations landscape of PDAC subtypes 24
2.2.4 Integrative pathway analysis 27
2.2.5 Connectivity of genomic alterations to transcription factors and their pathway activities in pancreatic cancer 32

2.3 Discussion 36

2.4 Methods 38

2.4.1 Transcriptome-based classification 38
2.4.2 Treatment outcomes 39
2.4.3 Differential gene expression, functional and pathways analyses 39
2.4.4 Prediction of regulator kinases 40
2.4.5 Mutation and copy number alteration analyses 41
2.4.6 Integrative analysis of expression and genomic alterations 41
2.4.7 Statistical analyses 44

2.6 Supplementary Information 45

2.7 Funding 45

2.8 Ethics Approval **45**

Chapter 3 : Machine Learning and Network Analyses Reveals Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics ***46***

3.1 Introduction **46**

3.2 Results **48**

3.2.1 Subtypes of pancreatic cancer and their clinical characteristics 48
3.2.2 Proteomics-based signalling pathway analyses distinguish disease subtypes 51
3.2.3 Pancreatic cancer subtypes exhibit functional differences in mRNA levels and DNA methylation patterns 55
3.2.4 Biomarker genes, proteins and miRNA sets that define the pancreatic cancer subtypes 56
3.2.5 Subtyping pancreatic cancer cell lines 62
3.2.6 Predicting drug responses using machine learning 62
3.2.7 Validation of our machine learning drug prediction method using GDSC data 64

3.3 Discussion **65**

3.4 Methods **69**

3.4.1 RPPA-based classification of pancreatic cancer 69
3.4.2 Integrative subtyping of pancreatic cancer 70
3.4.3 Patient's clinical characteristics of the pancreatic cancer subtypes 70
3.4.4 Pathways and kinase enrichment analyses 70
3.4.5 Identification and evaluation of biomarker sets 71
3.4.6 Validating biomarker molecular datasets 74
3.4.7 Subtype classification of cell lines 74
3.4.8 Machine learning method to predict a cell line's drug response 75
3.4.9 Validation of our machine learning drug prediction method using GDSC data 75
3.4.9 Statistical analyses 76

3.5 Acknowledgements **76**

3.6 Ethics approval **76**

3.7 Supplementary Information **77**

Chapter 4 : Metabolic Genes Alterations Impacts the Clinical Aggressiveness and Drug Response of 32 Human Cancers ***1***

4.1 Introduction **1**

4.2 Results **3**

4.2.1 Alterations to genes involved in metabolism distinguish human cancers 4
4.2.2 Alterations of genes involved in carbohydrate, amino acid and lipid metabolic pathways across all cancers 7
4.2.3 Alterations of genes involved in metabolism are associated with alterations of mRNA transcript levels 10
4.3.4 The drug responses of cancer cell lines are associated with metabolic gene alterations 14
4.3.5 The subtypes within each cancer exhibit diverse responses to anticancer drugs 17

4.3 Discussion **20**

4.4 Methods **23**

4.4.1 Metabolic gene alterations in the TCGA cancers 23
4.4.2 Analysis of mRNA expression profiles of metabolic pathway genes across cancers 27
4.4.3 Alterations of metabolic genes in cancer cell lines 28

4.4.4 Dose-response characteristics of the LM and HM cancer cell lines 29
4.4.5 Identification of metabolic disease subtypes for each cancer type 29
4.4.6 Comparison of dose-response profiles within each cancer type for tumours with or without specific pathway alterations 29
4.4.7 Survival analysis 30
4.4.8 Statistical analyses 31
4.5 Ethics Approval 31
4.6 Supplementary Information 31
Chapter 5 : General Conclusion 33
5.1 General Discussion 33
5.2 Future Work 39
References 42
Appendices 78
Appendix A: Supplementary File Description 78
Supplementary files for Chapter 3 78
Supplementary files for Chapter 4 78
Appendix B: 80

List of Tables

Table 1-1: Big data and biological knowledge resources for bioinformatics and systems biology 3
Table 2-1: Some dysregulated pathway drug targets in clinical trial..... 36
Table 2-2: Co-occurring gene sets in pancreatic cancer 42
Table 3-1: Comparison of gene mutations between subtypes..... 62
Table 3-2: Trained supervised learning models 72
Table B-1: Key resources used in this thesis 80

List of Figures

Figure 1-1: Overall representation of the data integration challenge in computational systems biology 7
Figure 2-1: Pancreatic cancer subtypes and their clinical characteristics 17
Figure 2-2: Enrichment Map of QM-PDAC vs C-PDAC and EL-PDAC tumours 18
Figure 2-3: Enrichment Map of Classical vs. Exocrine-like tumours 19
Figure 2-4: Metabolic characteristic of PDAC subtypes..... 21
Figure 2-5: Subtype-specific EGFR signalling pathway activity 22
Figure 2-6: Subtype-specific TGF-β signalling pathway activity..... 23
Figure 2-7: Identification of regulatory kinases 24
Figure 2-8: Mutation and gene copy number analyses..... 26
Figure 2-9: Enriched Reactome pathways 27
Figure 2-10: Integrative pathway analysis 29
Figure 2-11: Genetic alterations in three critical signalling pathways..... 31
Figure 2-12: TieDIE subnetworks..... 34
Figure 2-13: MAPK heat diffusion sub-network 35
Figure 2-14: Expression2Kinases pipeline 41
Figure 2-15: Genomic alteration bias and TieDIE subnetwork..... 44
Figure 3-1: Classification of pancreatic cancer..... 50
Figure 3-2: Treatment outcomes 51
Figure 3-3: Pathway enrichment analyses..... 52
Figure 3-4: Protein-level differences between pancreatic cancer subtypes 54
Figure 3-5: Functional analyses and mutational landscape of pancreatic tumours 58
Figure 3-6: Classification of pancreatic tumours using biomarker sets 59
Figure 3-7: Biomarker set validation and mutational landscape of pancreatic tumours..... 61
Figure 3-8: Subtyping pancreatic cancer cell lines and predicting drug responses..... 63
Figure 3-9: Kappa scores of all GDSC pancreatic cancer cell lines with drug data..... 65
Figure 3-10: K-mean clustering plots..... 70
Figure 3-11: Feature selection and machine learning classification of pancreatic cancer... 73
Figure 4-1: Mutational landscape of TCGA tumours..... 5
Figure 4-2: Disease outcomes of the HM and LM tumours..... 7
Figure 4-3: Frequency of tumours of different cancer types with altered genes that are involved in second-tier metabolic pathways of carbohydrate, lipid and amino acid metabolism 8
Figure 4-4: Major catabolic and anabolic pathways of glucose and lipid metabolism in human cells 11
Figure 4-5: TCGA tumour grouping based on metabolic gene transcript level..... 12
Figure 4-6: DBSCAN tumour classification..... 14
Figure 4-7: GDSC cell lines and the mutational and dose-response characteristics..... 15
Figure 4-8: Metabolic pathway gene alteration-dependent dose-responses and disease outcomes..... 19
Figure 4-9: Unsupervised hierarchical clustergram of tumours assigned to the two metabolic supertypes of human cancers..... 24
Figure 4-10: Highlight table showing the fractions of tumours with alterations in glycolytic pathway genes across all human cancers..... 25

Figure 4-11: Highlight table showing the fractions of tumours with alterations in gluconeogenic pathway genes across all human cancers 26
Figure 4-12: Highlight table showing the fractions of tumours with alterations in mitochondrial fatty acid oxidation pathway genes across all human cancers 26
Figure 4-13: Highlight table showing the fractions of tumours with alterations in lipid biosynthesis pathway genes across all human cancers 27
Figure 4-14: Heatmap of GDSC cancer lines showing the percentages of cell lines with alterations to genes involved in each of the 16 first-tier metabolic pathways 28
Figure 4-15: Metabolic pathway gene alterations within cancer types 30
Figure 5-1: Our current knowledge of human proteins 34

List of Abbreviations

ABC	ATP-Binding Cassette
ACC	Adrenocortical Carcinoma
BioGrid	Biological General Repository for Interaction Datasets
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
C-PDAC	Classical PDAC
cBioPortal	Computational Biology Cancer Data Portal
CCLE	Cancer Cell Line Encyclopaedia
CDK	Cyclin Dependent Kinase
CESC	Cervical Squamous Cell Carcinoma
ChEA	Chromatin Immunoprecipitation Enrichment Analysis
CHOL	Cholangiocarcinoma
CNA	Copy Number Alteration
COADREAD	Colorectal Adenocarcinoma
CoMDP	Co-Occurring Mutated Driver Pathway
COSMIC	Catalogue of Somatic Mutations in Cancer
CTD2	Cancer Target Discovery and Development
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DFS	Disease-Free Survival
DGID	Drug Gene Interaction Database
DLBC	Diffuse Large B-Cell Lymphoma
DUSP4	Dual-Specificity Phosphatase
EGFR	Epidermal Growth Factor Receptor
EL-PDAC	Exocrine-Like PDAC
ENCODE	Encyclopaedia of DNA Elements
ERK	Extra-Cellular Regulatory Kinase
ESCA	Oesophageal Adenocarcinoma
FDA	Food and Drug Administration
GBM	Glioblastoma Multiforme
GDSC	Genomics of Drug Sensitivity in Cancer

GEO	Gene Expression Omnibus
GLUT	Glucose Transporter
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GSKB	Glycogen Synthase Kinase Beta
GTEX	Genotype-Tissue Expression Portal
HIF-1	Hypoxia-Inducible Factor 1
HM	Higher Frequency of Metabolic Gene Alterations
HNSC	Head and Neck Squamous Cell Carcinoma
ICGC	International Cancer Genomic Consortium
KEA	Kinase Enrichment Analysis
KEGG	Kyoto Encyclopaedia of Genes of Genomics
KICH	Kidney Chromophobe
KIRP	Kidney Renal Papillary Cell Carcinoma
KNN	K-Nearest Neighbour
LAML	Acute Myeloid Leukaemia
LDHA	Lactate Dehydrogenase
LGG	Brain Lower Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
LINCS	Library of Integrated Cellular-based Signature
LM	Lower frequency of metabolic gene alterations
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MAPK	Mitogen Activated Protein Kinase
MESO	Mesothelioma
miRNA	micro Ribose Nucleic Acid
mRNA	Messenger Ribose Nucleic Acid
mTOR	Mammalian Target of Rapamycin
NCA	Neighbourhood Component Analysis
OGs	Oncogenes
OMIM	Online Mendelian Inheritance in Man
OS	Overall Survival

OV	Ovarian Serous Cystadenocarcinoma
PAAD	Pancreatic Adenocarcinoma
PCPG	Pheochromocytoma And Paraganglioma
PDAC	Pancreatic Ductal Adenocarcinoma
PDK-1	Pyruvate Dehydrogenase Complex Kinase-1
PharmaGKB	Pharmacogenomics Knowledgebase
PI3K	Phosphoinositide 3-Kinase
PKM	Pyruvate Kinase Muscle Isoform
PRAD	Prostate Adenocarcinoma
QM-PDAC	Quasi-Mesenchymal PDAC
RNA-Seq	RNA-sequencing
RPPA	Reverse Phase Protein Array
SARC	Sarcoma
SIDER	Side Effect Resource
SKCM	Skin Cutaneous Melanoma
SNF	Similarity Network Fusion
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machines
t-SNE	t-Distributed Stochastic Neighbourhood Embedding
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TGCT	Testicular Germ Cell Tumours
TGF-β	Transforming Growth Factor–Beta
THCA	Thyroid Carcinoma
THYM	Thymoma
TieDIE	Tied Diffusion Through Interacting Event
TSGs	Tumour Suppressor Genes
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UCSC	University of California, Santa Cruz
UniProt	Universal Protein Resource
UVM	Uveal Melanoma

X2K	Expression2Kinase

Chapter 1 : General Introduction

If I had to nominate one key to success, it's a focus on, well, everything…
— Andrew Fitzgibbon

1.1 Introduction

Recent advances in various high-throughput experimental technologies and associated computational analysis techniques, have provided modern medicine with powerful tools for devising novel strategies to treat some of the world's most burdensome diseases. These advances have enabled the emergence of systems biology: a holistic, global, interdisciplinary and integrative approach to understanding the molecular processes that manifest as living organisms. Systems biology has, in turn, prompted the rise to systems medicine: a science focusing on the accurate modelling of complex diseases [1]. Systems biology approaches are illuminating both how the molecular components of cells interact with one another to form functional units within healthy cells, and the specific perturbances within and between these units that yield diseased cellular states [2].

We now live in an era where a constellation of amazing technologies is allowing us to construct, disturb, and observe biological model systems in the laboratory with unprecedented fidelity and throughput [3–7]. Over the last few years large-scale "molecular profiling" projects applying these technologies have yielded vast amounts of genetic, epigenetic, transcription, protein expression, metabolic and drug response data for cancerous tumours, diseased tissues, and cell lines [8,9]. Concurrently, comprehensive information has been gathered on the properties of thousands of cellular proteins, their functions, their interaction partners and the signalling or metabolic pathways within which they function [10,11]. While the accessibility of these "big data" has been enabled by the development and population of various large databases and data repositories, various statistical tools implemented in an array of different data analysis software have also been devised to extract actionable insights from the data. Although these developments are promising to transform the treatment

of all human diseases – from communicable infections to genetic disorders and cancers – the unprecedented complexity and scale of the accumulating data is presenting formidable analytical challenges [12].

The extraction of meaningful information on cellular processes from large scale genetic (or genomic), epigenetic (or epigenomic), transcription (or transcriptomic), protein expression (or proteomic), metabolism (or metabolomic), requires the application of innovative data mining and data integration approaches [13]. Further, cellular components in general form complex biological systems containing huge numbers of interconnected agents [14–16]. Such systems are dynamic, and frequently have "emergent" properties that cannot be entirely understood even given a comprehensive understanding of each of their individual parts.

Concretely, complex systems can be perceived as scale-free networks wherein the molecular components within the systems (for example proteins) are represented as nodes and the possibility of interactions between the components (for example protein-protein binding) are represented by edges between nodes [17,18]. These networks are arranged in a hierarchical manner, whereby sub-complex systems form modules of the higher level complex systems generating what is known as a hierarchical network [14,19]. Network representations of complex systems has revolutionised many physics, chemistry, and mathematics research fields. Besides also having useful applications in ecology and evolutionary biology research, network representations of complex systems are also crucial in systems biology and cell signalling research. Casting the complexity of molecular cellular systems as networks can be instrumental in understanding how these systems are organised and regulated. When, for example, the networks in question represent systems with regulatory and/or signalling functions, the nodes with the highest numbers of edges may represent "central" proteins, or inflexions that could be targeted to disrupt/modify cell signalling/regulation [20,21].

To facilitate a multiscale understanding of high-dimensional biological data and the complexity of the relationships between different data types (such as between

genomic, transcriptomic and proteomic data types), there is presently a concerted drive towards the development of novel data integration techniques [2,22–24]. In most cases, these computational techniques aim to leverage an understanding of the functional cellular components in physiology and disease using prior-knowledge network representations of biological systems (such as protein-protein interaction or gene regulatory networks) in conjunction with statistical and machine learning methods to analyse different data types derived from the same sets of biological specimens. This has led to a paradigm shift; from testing a small number of hypotheses that are defined well before data is even collected, to using the data to develop testable hypothesis based on observed connections between potential causes and effects that we never thought existed.

To facilitate the generation of hypothesis, data from most large-scale molecular profiling projects are freely accessible, searchable, and downloadable. Many of these datasets (see table 1-1 for some examples) contain multiple sources of information on the same sets of biological samples with, for example, individual specimens having combinations of associated epidemiological and/or clinical data, genome sequence data, methylation data, RNA transcription and/or protein expression data and drug response data.

Table 1-1: *Big data and biological knowledge resources for bioinformatics and systems biology*

#	Resource	Available Data	Web Portal	Statistics
1	The Cancer Genome Atlas [8]	Multi-molecular data profiling from patients with clinical information	https://portal.gdc.cancer.gov/	33,605 donors and 374,699 data files
2	International Cancer Genomic Consortium [9]		https://dcc.icgc.org/repositories	15,513 donors and 231,751 data files
3	Gene Expression Omnibus [25]	Public functional genomics data repository	https://www.ncbi.nlm.nih.gov/geo/	30,97,229 samples and 4,348 datasets
4	Expression Atlas [26]	Gene expression across species and biological conditions	https://www.ebi.ac.uk/gxa/home	1,284 experiments
5	Genotype-Tissue Expression Portal [27]	Resource of tissue-specific gene expression and regulation	https://www.gtexportal.org/home/	11,688 samples
6	Cancer Cell Line Encyclopaedia [6]	Detailed genetic and pharmacologic characterisation of	https://portals.broadinstitute.org/ccle	136,488 dose-response profiles

7	Genomics of Drug Sensitivity in Cancer [5]	a large panel of human cancer models	https://www.cancerrxgene.org/	224,202 dose-response profiles
8	Cancer Target Discovery and Development [28]	Data generated from different types of approaches, e.g. chemical genetics, genome-wide gain-of-function and loss-of-function studies	https://ctd2-dashboard.nci.nih.gov/dashboard/#	1,004 drugs and 35 cancers
9	Library of Integrated Cellular-based Signature [3,29]	Gene expression and functional changes after drug and molecular perturbations	http://lincsproject.org	1,127 cells treated with 41847 small molecules
10	BioGrid [30]	Genetic and protein interaction data	https://thebiogrid.org/	Over 1,670,000 interactions
11	KEGG pathways [11]	Molecular interactions, reaction and pathway database	https://www.kegg.jp/	533 pathways and 29,545,122 genes
12	Reactome pathways [10,31]		https://reactome.org/	2272 pathways, 12505 reactions and 10,883 human proteins
13	Sider 4.1 Side Effect Resources [32]	Information on marketed medicines and their recorded adverse drug reactions	http://sideeffects.embl.de/	1430 drug and 5868 Side effects
14	DrugBank [33]	Detailed drug data with comprehensive drug target information	https://www.drugbank.ca/	13,338 drugs, and 5,177 proteins
15	PharmaGKB [34]	Curated knowledge about the impact of genetic variations on drug response	https://www.pharmgkb.org/	656 drugs, 136 pathways and 132 clinical guidelines
16	ClinicalTrail.gov [35]	Database of privately and publicly funded clinical studies	https://clinicaltrials.gov/	308,234 clinical research studies
17	Universal Protein Resource [36]	A comprehensive resource of protein sequence and functional information	https://www.uniprot.org/	560,292 proteins
18	Encyclopedia of DNA Elements [37,38]	A comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.	https://www.encodeproject.org/	15,074 various experimental data

19	The Gene Ontology Resource [39]	Comprehensive information on the functions of genes	http://geneontology.org/	44,990 ontology terms, 6,394,055 annotations and 1,152,603 gene products
20	Mammalian Phenotype Ontology [40]	Data on spontaneous, induced, and genetically-engineered mutations and their strain-specific phenotypes.	http://www.informatics.jax.org/phenotypes.shtml	318,738 Mapped genes/markers
21	The Human Phenotype Ontology [41]	Data on phenotypic abnormalities encountered in human disease	https://hpo.jax.org/app/	Over 13,000 terms and over 156,000 annotations to hereditary disease
22	Online Mendelian Inheritance in Man [42,43]	Catalogue of Human Genes and Genetic Disorders	https://www.omim.org/	6,441phenotypes and 4,106 disease-causing genes
23	Proteomic Database [44]	Database for exploration and analysis of human proteome data	https://www.proteomicsdb.org/#overview	15,721 proteins and 43,237,800 spectra
24	Allen Brain Atlas [45]	Data and web application to explore the biological complexity of the mammalian brain	http://portal.brain-map.org/	Data from ~1,000 experiments and image analysis datasets ~63,000 cells
25	Drug Gene Interaction Database [46]	User-friendly browsing, searching, and filtering of information on drug-gene interactions and the druggable genome	http://www.dgidb.org/	40,000 genes and 10,000 drugs involved in over 15,000 drug-gene interactions
26	Catalogue of Somatic Mutations in Cancer [47,48]	World's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer	https://cancer.sanger.ac.uk/cosmic	1,420,135 samples and over 44,451,862 genetic alterations
27	Project Achilles [4,49]	Genome-scale RNAi and CRISPR-Cas9 genetic perturbation to silence or knockout individual genes and identify those genes that affect cell survival	https://depmap.org/portal/achilles/	17,309 genes, 1,627 cell lines across 38 primary diseases
28	Human Proteome Map [49]	Integrated massive peptide sequencing result from the draft	http://www.humanproteomemap.org/	17,000 human genes and ~290,000 non-

		map of the human proteome project		redundant peptide identifications
29	cBioPortal [50]	Large-scale cancer genomics data sets	https://www.cbioportal.org/	158 cancer studies and 40,199 samples
30	The National Center for Biotechnology Information [13]	Various biomedical and genomic information	https://www.ncbi.nlm.nih.gov/	N/A
31	Human Protein Reference Database [51]	Integrated information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein	http://www.hprd.org/	30,047 proteins, 41,327 protein-protein interactions and 93,710 post-translational modifications
32	Human Protein Atlas [52,53]	Expression and localization of human proteins across tissues and organs	https://www.proteinatlas.org	Tissue-specific expression of 17,000 individual proteins
33	Strings [54]	Protein-protein interaction networks, integrated over the tree of life	https://string-db.org/	19,257 human proteins with network connections
34	The Molecular Interaction Database [55]	Experimentally verified protein-protein interactions mined from the scientific literature	https://mint.bio.uniroma2.it/	8,931 human proteins and their verified interactions
35	CORUM [56]	The comprehensive resource of mammalian protein complexes	https://mips.helmholtz-muenchen.de/corum/	4274 mammalian protein complexes
36	Integrated interactions database [57]	Tissue-specific protein–protein interactions	http://iid.ophid.utoronto.ca/	1,566,043 PPIs among 68,831 proteins
37	Human Integrated Protein–Protein Interaction Reference [58]	large-scale protein–protein interactions database	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/	more than 270 000 confidence scored PPIs
38	High-quality Interactomes [59]	Curated compilation of high-quality protein-protein interactions from 8 interactome resources	http://hint.yulab.org/	62,435 interactions and 116,897 co-complex
39	biophysical interactions of ORFeome-based complexes [60]	Protein interaction networks and co-complexes	http://bioplex.hms.harvard.edu/	Over 50,000 interactions among ~11,000 proteins

The data from each of these resources can often be integrated or augmented with data either from other resources or from additional laboratory experiments. By enabling a more holistic multi-dimensional view of human diseases, integrative analysis of multiple data types from these various freely accessible resources can be used to fast-track the development of new therapies by focusing more traditional biomedical research and development activities on the *in vitro* validation of computationally predicted drug targets and therapeutic molecules (Figure 1-1).

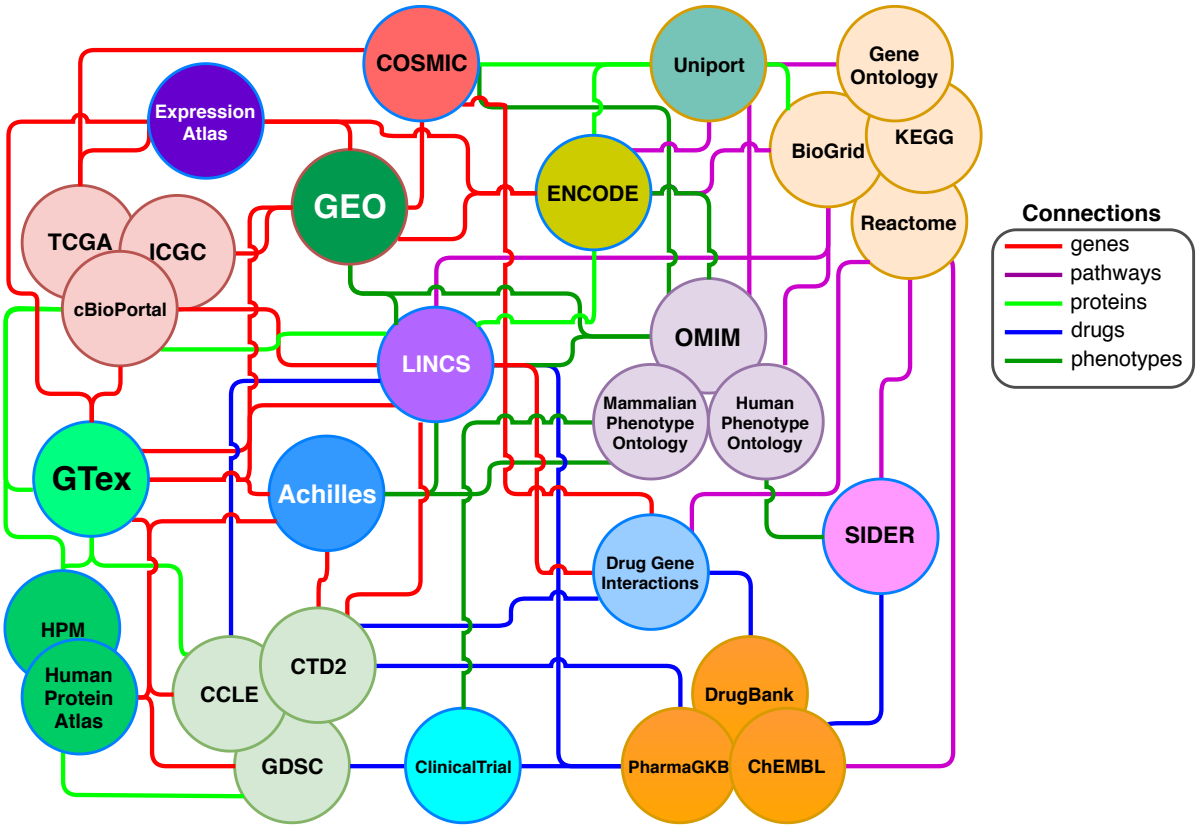


Figure 1-1: *Overall representation of the data integration challenge in computational systems biology: The resources that contain similar datasets are shown in groups. The large-scale datasets from each resource can be individually analysed or integrated with data from other resources that provide complimentary information on genes, proteins, drugs, pathways or protein-protein interactions, and phenotypes. Note that I have omitted some of the possible connections to aid visualisation of the global picture.*

This integrative "data-first" approach has already yielded solutions to numerous problems than would have taken decades longer to solve using more traditional "hypothesis first" approaches. Successes include the design of optimized therapeutic

molecules [61–63], correctly predicting drug targets [64–67], and correctly predicting the responses of tumours to anticancer drugs [68–74].

Given that many of the currently available datasets have a strong cancer focus, much of the activity in this burgeoning research area is on curing cancer. Although, integrative data-first approaches have yielded small molecule inhibitors for targeted-therapies of tumours that have improved the survival of patients afflicted with many forms of cancers [75–78], there remain many other forms of cancers for which such approaches have not yet yielded effective therapies [79–81]. For example, pancreatic cancers, triple negative breast cancers and lung cancers remain very difficult to treat because they have variable responses to available anticancer drugs. However, the underlying molecular causes of these variable responses should themselves be discernible using integrative data-first approaches and, once the correct subtypes of the cancer in question have been determined it should be possible to both identify prospective subtype-specific drug targets for each of these, and identify sets of therapeutic molecules that could potentially be used to inhibit these targets.

Given the perpetually increasing cost and difficulty of developing new drugs using traditional reductionist approaches [82–84], it would be completely untenable to consider the development of therapeutics that would only be useable in, for example, small subsets of cancer patients. The strength of computational systems approaches is that, when they are applied to molecular data (i.e. genomic sequence, RNA transcription or protein expression data) for even just a single tumour, it is plausible that they could both identify the metabolic aberrations within the cells of that tumour, and indicate the chemical compound(s) that should be used to most effectively and specifically treat that tumour. In such cases the cost of drug development would simply be the cost of obtaining the pertinent metabolic data: a cost that, due to economy of scales and continuing technological innovation, is anticipated to progressively decrease.

At the moment, however, publicly available data for patients with tumours displaying similar molecular profiles is enough for us to begin using computational approaches to find the best treatment options for some of the less common and/or more variable

cancers. Despite the promise of integrative data-first approaches in the development of therapeutics, using publicly available data to accurately identify appropriate drug targets that could enable the treatment of specific cancer variants remains non-trivial. Besides the computational complexity of both characterizing the different subtypes of cancer that can arise within a given tissue, and then identifying suitable therapeutics for these cancer subtypes, the effective deployment of these therapeutics will require (potentially costly) molecular profiling and categorization of patient tumours.

Fortunately, once we have accurately identified the different subtypes of any particular cancer, the molecular profiling and classification of new tumours need not involve the extensive determination of the tumours' genomes, transcriptomes and proteomes. Rather, the identification of genetic polymorphisms in a few "marker" genes or the determination of a small number of marker mRNA transcription levels and/or marker protein expression levels may suffice to accurately determine which drugs a tumour will be most sensitive to. In this regard machine learning strategies (both supervised and unsupervised) have proven to be highly effective for the identification of small numbers of marker genes, transcripts and proteins that can be used to accurately classify tumours. These methods have additionally been useful for pinpointing both the molecular dependencies responsible for the varied responses of different cancer subtype to different drugs, and, in so doing, identifying the specific therapeutic compounds that could best be deployed to treat these cancer subtypes [22,24,85].

In this thesis, we collect and mine data from over 20 publicly available resources. We then leverage integrative computational systems analyses, network analyses and machine learning, to gain insights into the pathology of pancreatic cancer subtypes. Using these contemporary analytical approaches, we also evaluate the alterations of metabolic pathway genes across 32 human cancers and the impact that these genetic alterations have on both treatment outcomes and the responses of tumours to different anticancer drugs. We show that using these big data resources we can: (1) accelerate the discovery of drug targets by identifying cellular signalling network nodes that can be used to disrupt the disease process, and (2) predict either patient-specific or cancer subtype specific drug candidates. The overarching message of this thesis is that the time has already arrived where we are able to leverage available data resources to

potentially elicit more precise and personalised cancer therapies that would yield better clinical outcomes at a much lower cost than is currently being achieved.

1.2 Thesis Organization

This thesis is organized into five chapters. Chapters 2-4 are manuscripts that have either been published in a peer reviewed journal or have been submitted to one. Chapter 5 is a general discussion on the significance and impact of the studies presented in Chapters 2-4.

Chapter 1

We provide an introduction to big data and data integration in the context of computational systems biology.

Chapter 2

We present results published in the paper entitled “Integrative landscape of dysregulated signalling pathways of clinically distinct pancreatic cancer subtypes”. Here, we used publicly available pancreatic cancer datasets from The Cancer Genome Atlas (including transcriptome profiles, gene copy number alterations, mutation profiles, proteome profiles and data on disease outcomes) to provide, for the first time, a comprehensive landscape of pathway alterations that are associated with pancreatic cancer. Using transcriptome data, we identified three pancreatic cancer subtypes that displayed distinctive survival outcomes. Further, using various pathway and network analysis approaches of these data, we uncovered aberrations in multiple cell signalling and metabolic pathways that implicate regulatory kinases and the Warburg effect as the likely drivers of the distinct molecular signatures of the different pancreatic cancer subtypes. Through integrative analysis of mRNA expression profiles, gene mutations, gene copy number alterations, proteome datasets, and prior knowledge, we found that, to varying extents in the three pancreatic cancer subtypes, hyperactivation of the PI3K-mTOR and MAPK pathways, and dysregulation of p53 and cell cycle control processes, are apparently the drivers of pancreatic cancer progression.

Chapter 3

We addressed the challenge of inconsistent classifications of pancreatic cancer patient tumours when the tumours are subtyped using different types of molecular data using an integrative clustering method called Similarity Network Fusion. Using targeted proteomics and other molecular data compiled by The Cancer Genome Atlas we revealed that pancreatic tumours can be broadly segregated into two distinct subtypes. Besides being associated with substantially different clinical outcomes, tumours belonging to each of these subtypes also display notable differences in diverse signalling pathways and biological processes. At the proteome level, we show that tumours belonging to the less severe subtype are characterised by aberrant mTOR signalling, whereas those belonging to the more severe subtype are characterised by disruptions in SMAD and cell cycle-related processes. We use machine learning algorithms to define sets of proteins, mRNAs, miRNAs and DNA methylation patterns that could serve as biomarkers to accurately differentiate between the two pancreatic cancer subtypes. Lastly, we confirm the biological relevance of the identified biomarkers by showing that these can be used together with pattern-recognition algorithms to accurately infer the drug sensitivity of pancreatic cancer cell lines. Our study shows that integrative profiling of multiple data types enables a biological and clinical representation of pancreatic cancer that is comprehensive enough to provide a foundation for future therapeutic strategies.

Chapter 4

We evaluate the “The Frequency of Metabolic Genes Alterations Impacts the Clinical Aggressiveness and Drug Response of 32 Human Cancers”. Here, using genomics data from the Cancer Genome Atlas for 10,528 tumours of 32 different cancer types, we characterise the alterations of genes involved in various metabolic pathways across all human cancers. We find that while mutations and copy number variations of metabolic genes are pervasive across all human cancers, the extent of these gene alterations varies between them. We identify that the most common alterations are to genes involved in lipid, carbohydrate and amino acid metabolism. Based on the frequencies of metabolic pathway gene alterations, we further find that there are two distinct cancer supertypes that tend to be associated with different clinical outcomes.

By utilising the known dose-response profiles of 825 human cancer cell lines from the Genomics of Drug Sensitivity in Cancer dataset, we infer that cancers belonging to these supertypes are likely to respond differently to various anticancer drugs. Further, we show that for each of the 32 human cancer types, the altered genes involved in particular metabolic pathways are associated with observed variations in the drug responses of tumours to many different anticancer drugs. Collectively our analyses define the foundational metabolic features of different cancer supertypes and subtypes upon which discriminatory strategies for treating particular tumours could be constructed.

Chapter 5

We provide a conclusion to this thesis and recommendations for future work in big data and big data integration in computational systems biology.

Chapter 2 : Integrative Landscape of Dysregulated Signalling Pathways of Clinically Distinct Pancreatic Cancer Subtypes

This section is a reformatting of a paper published in Oncotarget [86]:
https://doi.org/10.18632/oncotarget.25632

Musalula Sinkala, Nicola Mulder, Darren Martin

2.1 Introduction

Pancreatic cancer is the most lethal form of cancer. It has an extremely poor prognosis with less than 20% of patients surviving for more than one year following diagnosis [87,88]. Factors contributing to reduced survival rates are the difficulty of diagnosing the disease during its early stages, the rapid progression of tumours with few specific associated symptoms, and the diversity of responses that different forms of pancreatic cancer have to anticancer drugs [89,90]. Despite progress having been made towards understanding the histological phenotypes and molecular mechanisms at play, positive responses to conventional chemotherapy regimens have remained infrequent, and the overall survival rates of patients have not substantially improved [91].

A significant challenge to achieving better treatment outcomes has been the heterogeneity of pancreatic cancers. Underlying this heterogeneity is the vast array of somatic mutations that are acquired during oncogenesis, and the varied effects that these mutations have on cell signalling pathways [92,93]. Recent analyses of genomic sequence datasets from patients with advanced disease have identified potential activating mutations, many of which occur in genes encoding proteins that might be suitable drug targets [94,95]. In this regard the discovery of mutation hot-spots in various signalling kinases has already prompted the development of highly selective kinase inhibitors that are capable of specifically killing pancreatic cancer cells. Although the antitumor activities of some of these kinase inhibitors have been strong, they have rarely been long-lasting, with the targeted cancers frequently developing

resistance [93]. There is, therefore, a pressing need to identify additional potential drug targets amongst the dysregulated signalling and metabolic pathway components that differentiate pancreatic cancer subtypes. Used in conjunction with kinase-inhibitors, novel drugs targeting these pathway components could yield pancreatic cancer therapies with longer lasting effectiveness.

Aiding in the discovery of novel drug targets has been the use of next-generation sequencing based analytical methods that simultaneously identify mutations in sequences and quantify the expression of all the cellular genes that might have an impact on cancer progression. In this regard, the Cancer Genome Atlas (TCGA) project, has performed a systematic genomic, transcriptomic and proteomic characterisation of matched healthy and cancerous tissue samples from thousands of individuals afflicted with a variety of cancers [96]. This data, together with matched clinical information is publicly available and includes data for 185 pancreatic cancer patients. Combining data on transcription levels, gene mutations, copy number alteration, protein expression levels and clinical information, the intention of such datasets is to uncover causal relationships between specific genetic and/or cellular aberrations and the onset of disease [97].

Recent genomic studies using such datasets have both provided insights into the biological heterogeneity of pancreatic cancer and identified genomic aberrations that may be of therapeutic and prognostic value [94,95,98]. These studies have identified somatic mutations in the proto-oncogene *KRAS* as a hallmark of pancreatic cancers in that more than 90% of pancreatic cancer cases have mutations in this gene [95,98–100]. Several other mutations are also strongly associated with the onset of pancreatic cancers, including homozygous deletions in the *TP53*, *SMAD4*, and *CDKN2A* tumour suppressor genes [98,100]. Alteration in the *KRAS*, *TP53*, *SMAD4* and *CDKN2A* are considered as the critical drivers of pancreatic tumorigenesis and altered signalling through KRAS and p53 is associated with varied treatment response to therapy and disease outcomes [100–102] . Nonetheless, as in other cancers, genetic alterations and variations in gene expression also occur in many other genes. Specifically, genes involved in the RB, beta-catenin, PI3K-Akt, and NOTCH pathways, commonly exhibit alterations that likely contributed to tumour development and progression[87,95,100].

Further, while tumours displaying KRAS-pathway alterations either alone or in combination with TP53 pathway alterations have a poor prognosis, it has been shown that tumours with more complex pathway disruptions tend to have even poorer outcomes [101,103].

Notably, these studies have highlighted the alterations in key genes that function in these various signalling pathways. However, these studies have relied on smaller and/or less detailed datasets than those which are now available and have therefore failed to comprehensively define the specific signalling network perturbations that arise during different forms of pancreatic cancer. Here we explore the molecular characteristics of the three major pancreatic cancer subtypes and define the altered signalling pathways and subcellular process that, besides differentiating these subtypes and potentially being the underlying drivers of oncogenesis, also present a variety of potential prognostic biomarkers and drug targets.

2.2 Results

We assembled a TCGA pancreatic ductal adenocarcinoma (PDAC) dataset comprising clinical information for 185 patients together with their associated cellular transcription data (based on RNA sequencing (RNA-Seq)), protein expression data (based on reverse phase protein array (RPPA)), and information on genomic mutations and copy number alterations (CNA). We performed, survival, clustering, and integrative pathway and network analyses of these diverse data types, both to classify the pancreatic cancers into different subtypes, and to reveal their clinical characteristics and the potential underlying causes of oncogenesis in of these different subtypes.

2.2.1 Pancreatic cancer subtypes display distinctive clinical outcomes

Based on the mRNA transcription data we identified three major mRNA expression profiles using unsupervised hierarchical clustering (Figure 2-1A). Upon returning only exemplars for each profile, we identify the three PDAC subtypes as: (1) quasi-mesenchymal PDAC (QM-PDAC; 35 samples), (2) classical PDAC (C-PDAC; 87 samples), and exocrine-like PDAC (EL-PDAC; 48 samples), based on transcriptomic

classification framework established by Collision *et al*. [104]. These three subtypes were associated with distinct overall survival, and duration of disease-free survival (Figure 2-1B and Figure 2-1C). Specifically, both overall survival and the duration of disease-free survival were shorter for patients with the QM-PDAC subtype, intermediate for patients with the C-PDAC subtype and longer for patients with the EL-PDAC subtype. We observed a similar trend for treatment outcomes, with patients having QM-PDAC and EL-PDAC respectively displaying the worst and best outcomes (Figure 2-1D and Figure 2-1E). Further, we did not observe significant associations between age, gender, or diabetes with the distribution of the PDAC subtypes (Figure 2-1H). Finally, we further validated the consistency of tumours within each cluster using a support vector machine classifier which yielded an average 10-fold cross validation classification accuracy of 95.5% over ten models (Figure 2-1F). We have summarised the distribution of tumour grades across these PDAC subtypes in Figure 2-1G.

2.2.2 Gene expression and pathway characteristics of different PDAC subtypes

We compared gene expression profiles between all pairs of PDAC subtypes and identified genes that were differentially expressed within the tumours of each PDAC subtype (see Supplementary file 1). Using gene set enrichment analysis (GSEA), we established that the genes which were differentially expressed between the subtypes were involved in a variety of different signalling pathways [105]. Compared with tumours of the other subtypes, those of the QM-PDAC subtype displayed elevated transcription levels for genes involved in the epidermal growth factor receptor (EGFR) signalling pathway, the transforming growth factor–beta (TGF-β) signalling pathway, the phosphoinositide 3-kinase-mechanistic target of rapamycin (PI3K-mTOR) oncogenic pathway, the mitogen-activated protein kinase (MAPK) oncogenic pathway and among others (Figure 2-2). Dysregulation of the EGFR signalling pathway has previously been linked to tumour aggressiveness and reduced patient survival in various cancers including those of the breast and lung [106,107]. The PI3K-mTOR pathway was inactive in EL-PDAC tumours but was activated in C-PDAC and QM-PDAC tumours. In other cancers, including those of the breast, gastrointestinal tract

and prostate, activation of this pathway has been previously associated with significantly decreased 5-year survival rates [108].

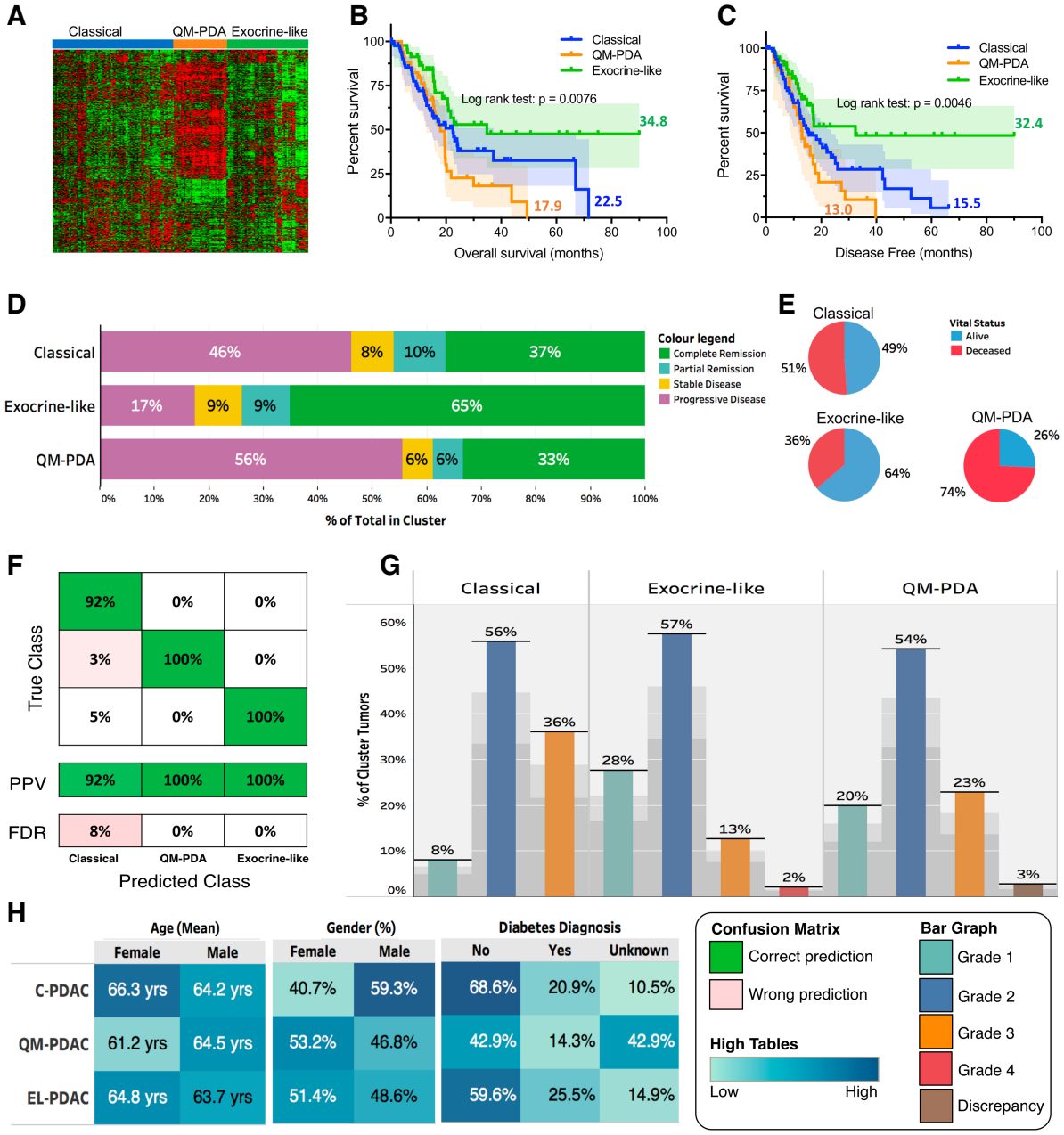


Figure 2-1: *Pancreatic cancer subtypes and their clinical characteristics:* ***(A)*** *Clustering of mRNA expression data identified three major pancreatic cancer subtypes, each with distinct expression patterns.* ***(B)*** *Kaplan-Meier Curves: overall patient survival periods were lower for patients with QM-PDAC and highest for those with EL-PDAC. Pairwise comparisons showed statistically significant differences between: C-PDAC vs. EL-PDAC (χ^2 = 4.4, p = 0.036) and QM-PDAC vs. EL-PDAC (χ^2 = 9.7, p = 0.002).* ***(C)*** *Kaplan-Meier Curves: disease-free survival months were lower for patients with QM-PDAC and highest for those with EL-PDAC: C-PDAC vs. EL-PDAC (χ^2 = 5.3, p = 0.02) and QM-PDAC vs. EL PDAC (χ^2 = 9.2, p = 0.002).* ***(D)*** *Vital statistics after the first course of treatment; only a*

quarter of QM-PDAC patients were alive compared with nearly half of C-PDAC patients and two-thirds of EL-PDAC patients. Odds ratios (95% CI): C-PDAC vs QM-PDAC = 2.7(1.158 – 6.568), C-PDAC vs EL-PDAC = 0.541(0.261 – 1.122), EL-PDAC vs QM-PDAC = 5.51(1.95 – 13.33). ***(E)*** *Treatment outcomes after the first course of therapy were most favourable for EL-PDAC patients and least favourable for QM-PDAC patients.* ***(F)*** *Support Vector Machine confusion matrix: a representative confusion matrix for the SVM classifier used to validated the purity of each molecular subtype of pancreatic cancer. We obtain on average a classification accuracy of 95.5% with an F1-score of 0.96.* ***(G)*** *Distribution of tumour grades across molecular subtypes: showing percentage of total count of the number of tumours for each grade broken down by molecular subtype.* ***(H)*** *Highlight tables of participant's age, gender, and diabetes diagnosis from left, centre and right respectively. The overall mean age of participants that were afflicted by C-PDAC, QM-PDAC and EL-PDAC were 65yrs., 64yrs. and 63yrs, respectively: one-way ANOVA test: F = 0.499, p = 0.608. No significant association was found between gender and diabetes diagnosis, and the presence of a particular PDAC subtypes, X^2 = 2.351, p = 0.308, and X^2 = 0.611, p = 0.737, respectively.*

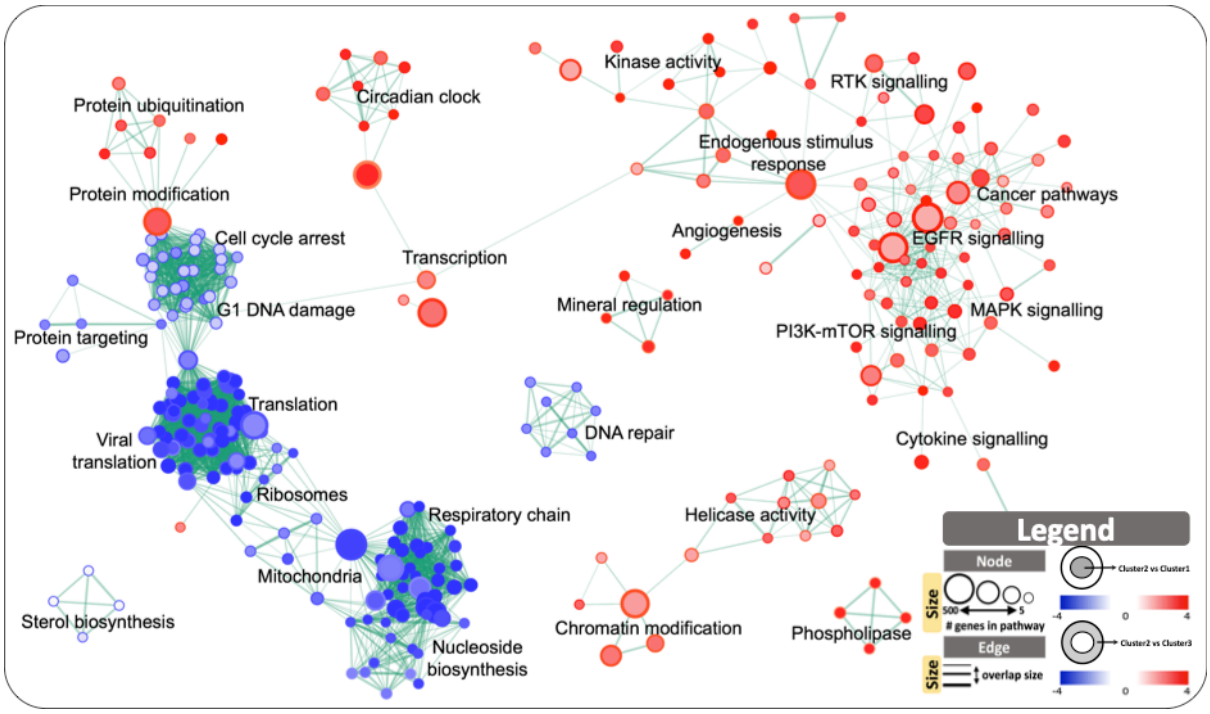


Figure 2-2: *Enrichment Map of QM-PDAC vs C-PDAC and EL-PDAC tumours: GSEA was used to obtain enriched gene ontology (GO)-terms that were visualised using the Enrichment Map plug-in for Cytoscape. Each node represents a GO-term with similar nodes clustered together and connected by edges with the number of known interactors between the nodes being represented by the thickness of edges. The size of each node denotes the gene set size for each specific node GO-term. A map comparing C-PDAC and EL-PDAC tumours is shown in Fig S2.*

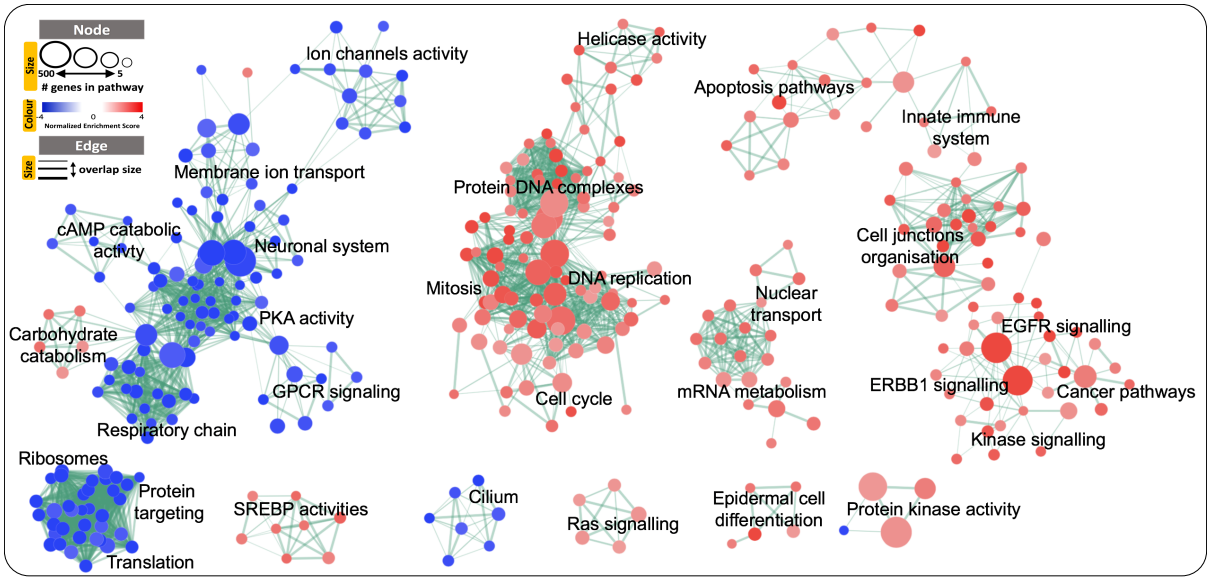


Figure 2-3: *Enrichment Map of Classical vs. Exocrine-like tumours: visualisation of enriched GO-terms obtained from GSEA results of classical vs. exocrine-like tumours. Each node represents a GO-term with similar nodes clustered together and connected by edges whereby the number of known interactors between the nodes specifies the edge thickness. The size of each node denotes the gene set size for each specific node GO-term. The map was created the Enrichment Map plug-in for Cytoscape.*

Further, we observed reduced expression of genes involved in electron transport chain and oxidative phosphorylation in the QM-PDAC tumours and, to a lesser degree, in the C-PDAC tumours (Figure 2-2 and Figure 2-4A). Since patients with QM-PDAC and C-PDAC tumours exhibited the worst clinical outcomes, these findings are consistent with the hypothesised link between the Warburg effect (characterized by decreased mitochondrial respiration and increased glycolytic activity) and tumour aggressiveness [109,110]. It is well established that hypoxia-inducible factor 1 (HIF-1), which regulates glucose homeostasis by controlling the expression of multiple glycolytic genes and glucose transporters [111,112], drives the Warburg phenotypes of various cancers, including those of the lungs and clear renal cells [111,113].

In support of our findings that QM-PDAC tumours have a Warburg phenotype, we found elevated levels of *HIF1A* and concomitantly lower levels of its corepressor, *SIRT6* (Figure 2-4B) [114]. Also, we found higher transcript levels of *SLC2A1* and *HK2* in C-PDAC tumours relative to EL-PDAC tumours and the highest *SLC2A3* transcript levels in QM-PDAC tumours. The transcription of *SLC2A1*, *SLC2A3* and *HK2* is up-regulated by HIF1A [114]. Whereas *SLC2A1* and *SLC2A3* respectively encode the

glucose transporters, GLUT1 and GLUT3, and *HK2* encodes hexokinase II. Interestingly, among the class 1 GLUT transporters GLUT3 has the highest affinity for glucose and among the hexokinase isoforms, hexokinase II has the highest catalytic efficiency. Further, GLUT 3 and hexokinase II are both reported elevated in various cancers [115–117]. This is particularly significant as the combined action of GLUT3 and hexokinase II should afford tumour cells preferential access to available glucose for energy production via glycolysis.

We also found that the transcript levels of certain key glycolytic pathway enzymes varied between PDAC subtypes: these included the transcript levels of pyruvate kinase (*PKM*), lactate dehydrogenase (*LDHA*) and pyruvate dehydrogenase complex kinase-1 (*PDK1*) which were highest in C-PDAC tumours and lowest in EL-PDAC tumours (Figure 2-4B). Recently, studies have shown that the HIF-1A induced expression of PDK1 limits the oxidation of pyruvate to acetyl-CoA by inhibiting the pyruvate dehydrogenase complex in cancers of the breast and kidney [118,119]. Accordingly, we suggest that in QM-PDAC and C-PDAC tumours, the upregulation of *PDK1* and *LDHA* would likely favour the conversion of glucose-derived pyruvate to lactate, thereby promoting the Warburg effect in these PDAC subtypes (Figure 2-4C).

To further investigate the degree of EGFR and TGF-β signalling pathway activation in the different PDAC subtypes, we mapped mRNA expression levels onto these pathways. We observed that there were higher mRNA levels for genes involved in these pathways in the QM-PDAC tumours than in the C- and EL-PDAC tumours (Figure 2-5 and Figure 2-6). Here and subsequently, we opted to compare QM-PDAC to C-PDAC and EL-PDAC tumours (referred to collectively as the “other” subtypes) because QM-PDAC has the most distinct molecular signature of the three PDAC subtypes.

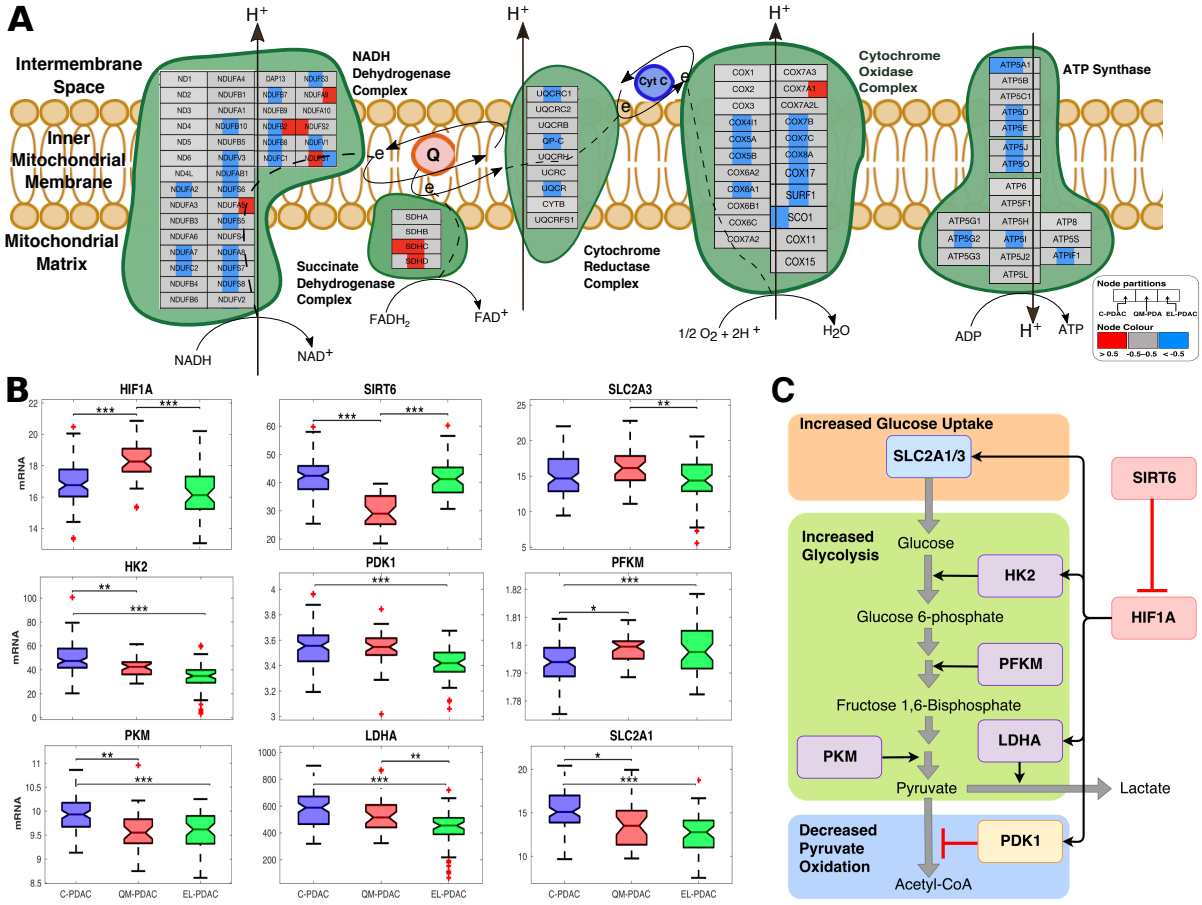


Figure 2-4: *Metabolic characteristic of PDAC subtypes:* ***(A)*** *PDAC subtype-specific electron transport chain activity: a comparison of electron transport chain activity between pancreatic cancer subtype based on mRNA expression data. Node denote genes-- left section = C-PDAC, middle section = QM-PDA and right section = EL-PDAC tumours. Node are coloured based on overall subtype mRNA-expression z-score (blue = low, grey = no change, and red = high). Edges represent various types of protein interaction (refer to legend to full notations for all edges).* ***(B)*** *PDAC subtype-specific transcript levels of Warburg effect associated mRNA: levels were compared between the tumour subtypes using one-way analysis of variance. The data are where transformed using the Box-Cox transformation. ***, **, and *, denote pairwise student t-test statistical significance for p values of 0.001, 0.01 and 0.05, respectively. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.* ***(C)*** *Control of glucose metabolism by HIF1A: SLC2A1, SCL2A3, HK2, LDHA, PDK1 are positively regulated (black arrows) by HIF1A. PDK1 negatively regulates (red blunt arrows) the conversion of pyruvate to acetyl-CoA.*

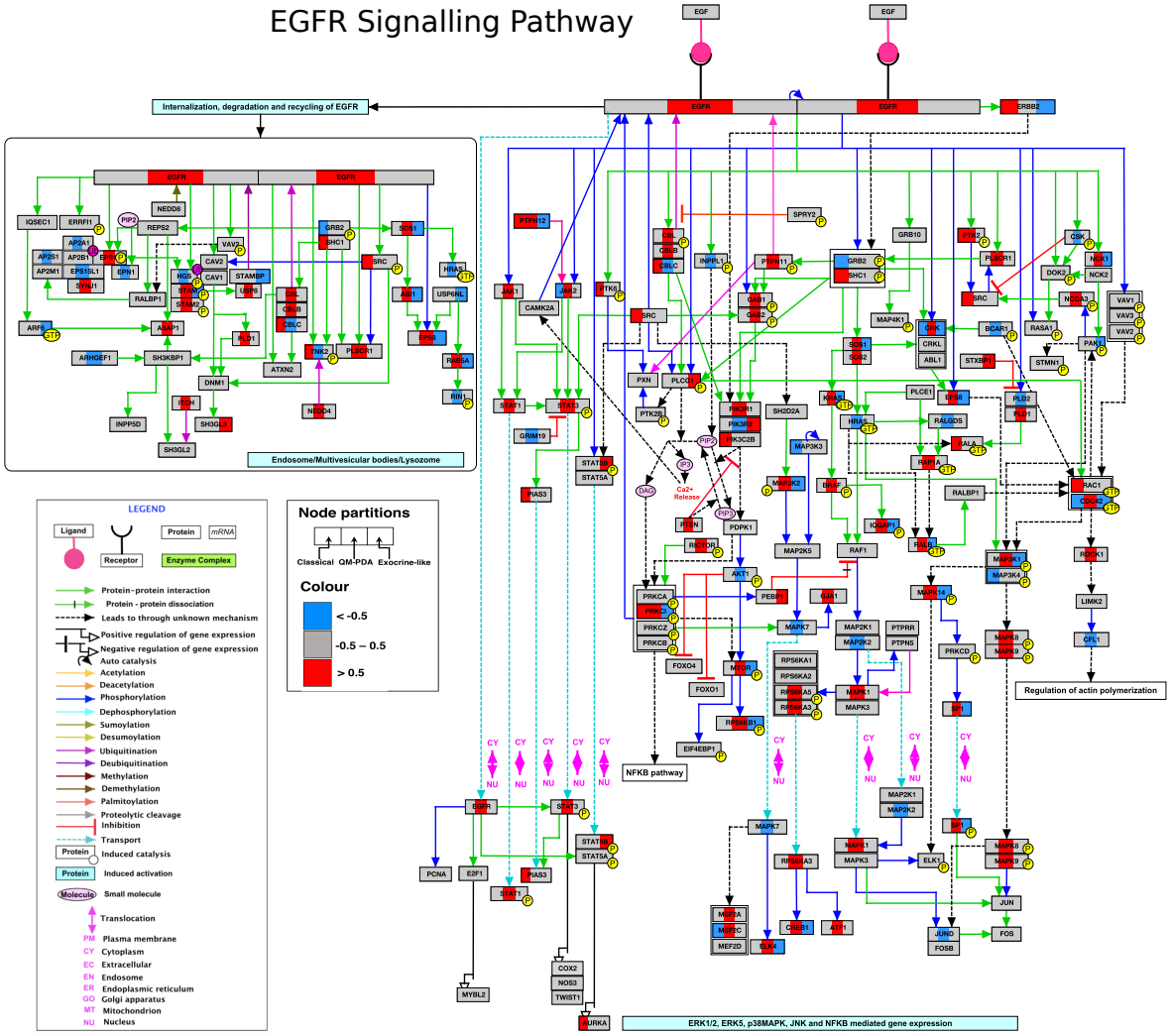


Figure 2-5: *Subtype-specific EGFR signalling pathway activity: a comparison of EGFR pathway activity between pancreatic cancer subtype based on mRNA expression data. Node denote genes-- left section = classical, middle section = QM-PDA and right section = exocrine-like tumours. Node are coloured based on overall subtype mRNA-expression z-score (blue = low, grey = no change, and red = high).*

Edges represent various types of protein interaction (refer to legend to full notations for all edges).

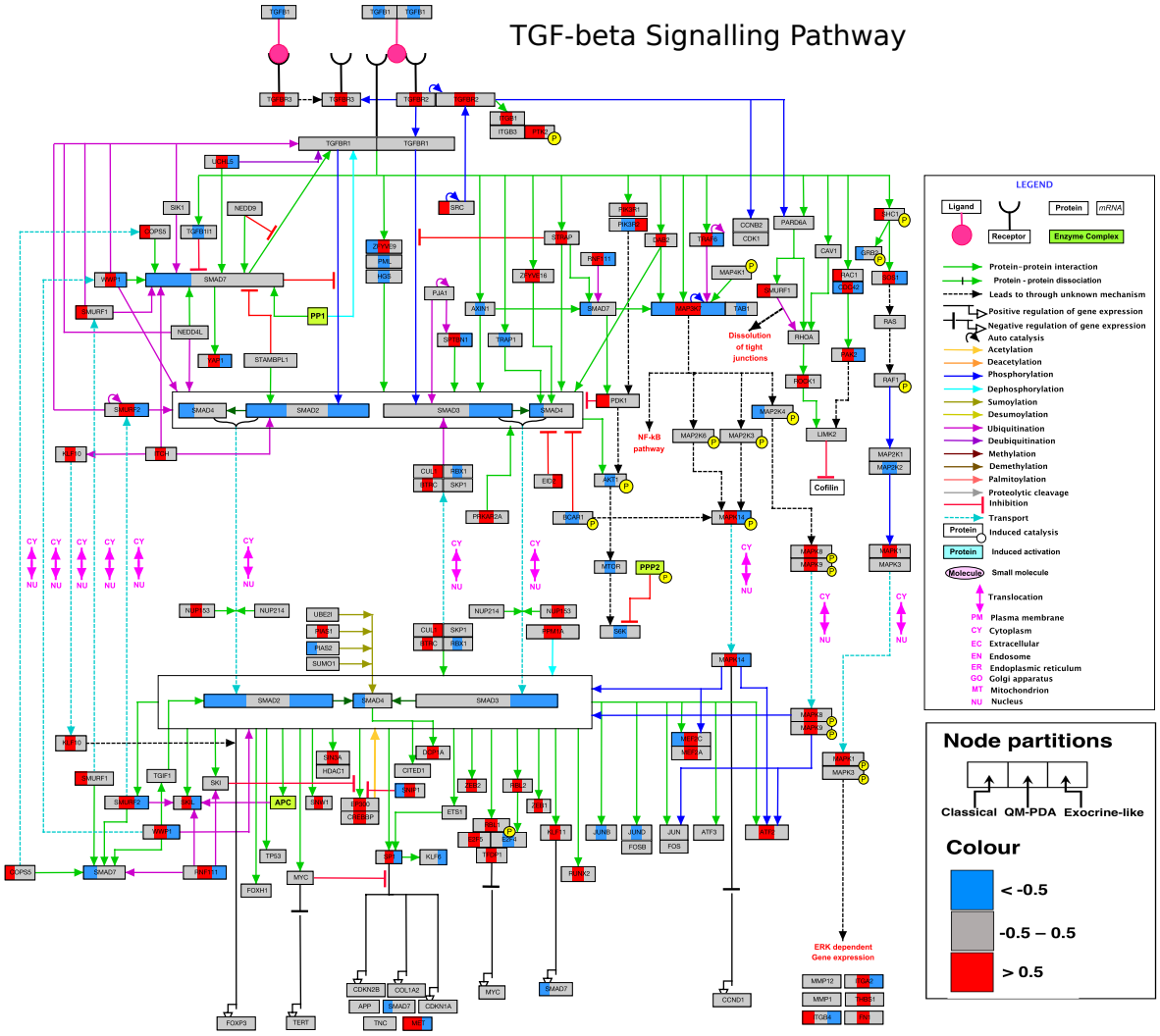


Figure 2-6: *Subtype-specific TGF-β signalling pathway activity: a comparison of TGF-β pathway activity between pancreatic cancer subtype based on mRNA expression data. Node denote genes; left section = classical, middle section = QM-PDA and right section = exocrine-like tumours. Node are coloured based on overall subtype mRNA-expression z-score (blue = low, grey = no change, and red = high). Edges represent various types of protein interaction (refer to legend to full notations for all edges).*

Many kinases have been previously implicated in carcinogenesis and, accordingly, we identified variations between the PDAC subtypes in the mRNA levels of the various kinases that are involved in the EGFR and TGF-β pathways [120]. Further, we aimed to pinpoint kinases that might be driving the QM-PDAC as opposed to the other subtypes. For this, we used an unbiased computational approach called

Expression2Kinases to identify the kinases that might be driving the hyperactivation of the EGFR and TGF-β pathways in QM-PDAC tumours. Among the top ten kinases identified were six well-documented oncogenes (*AKT1, GSKB, MTOR, MAPK1, MAPK14*, and *MAPK7*), all of which are involved in the EGFR and TGF-β pathways (Figure 2-7A and 2C) [47]. Also present in the top ten list were CDK1, CDK2 and ATM: kinases which are involved in cell cycle control. Interestingly, among the top ten kinase that were predicted by the Expression2Kinases methods, all showed higher expression levels in the QM-PDAC, except AKT1 (Figure 2-7C).

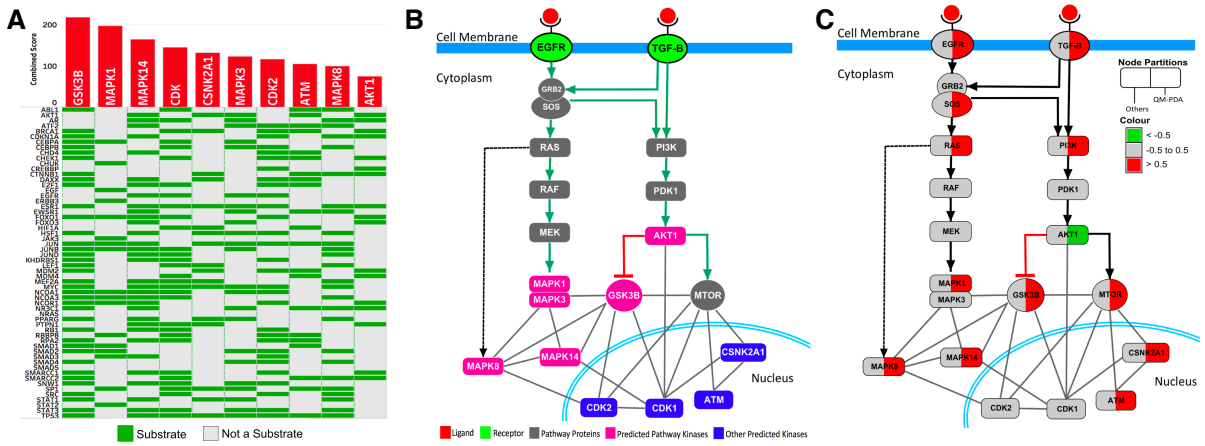


Figure 2-7: *Identification of regulatory kinases:* ***(A)*** *Expression2Kinases solution: Heat map showing the top ten predicted kinases ranked according to their combined statistical score based on the number of substrates they phosphorylate within a protein-protein interaction subnetwork. Along the rows of the heatmap are proteins which are the substrates for kinases given along the columns of the heatmap.* ***(B)*** *Mapping of the top ten predicted kinases onto simplified models of the EGFR and TGF-B signalling pathways. Six of the top ten ranked kinases (pink nodes) fall within these two pathways whereas the other predicted proteins are involved either directly in the cell cycle, or in the regulation of the cell cycle (blue nodes).* ***(C)*** *Simplified EGFR and TFG-β signalling pathway for X2K: mapping of mRNA expression data onto a model pathway showing the top ten X2K predicated kinase; AKT1, GSKB, MTOR, MAPK1, MAPK14, MAPK7, CDK1, CDK2 and ATM. Each node denotes a pathway protein, the left section denotes QM-PDA z-scored mRNA expression, and the right section denotes Other subtypes z-scored mRNA expression. Green = low expression, grey = no change, and red = high expression.*

2.2.3 The gene alterations landscape of PDAC subtypes

We evaluated the scope of genomic alterations in PDAC subtypes by focusing on the types of genetic changes that are known to promote oncogenesis. Specifically, these encompassed gain-of-function mutations in oncogenes (OGs), amplification of OGs,

loss-of-function mutations in tumour suppressor genes (TSGs), and deletions in TSGs. Across all the PDAC subtypes we found that, as has been reported elsewhere, *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, and *CDKN2B* were the most commonly altered genes (Figure 2-8A) [87,94,95].

SMAD4 signals through the canonical TGF-β pathway and therefore deletions in *SMAD4* should limit signalling through this pathway [121]: a factor that may seem inconsistent with our earlier finding that genes in this pathway display elevated levels of transcription (Figure 2-6 and Figure 2-7C). However, SMAD4 loss does not initiate tumorigenesis in human pancreatic cancers [121–123]. Further, in pancreatic tumours displaying either *SMAD4* deletions or *SMAD4* under-expression, ligand stimulation of the TGF-β pathways activates non-canonical pathways including the MAPK, p38/JNK, and PI3K-mTOR pathways which can function independently of SMAD signalling [121,124,125]. Moreover, we found that among the PDAC subtypes, QM-PDAC had the lowest transcript levels of genes in the SMAD gene family (SMAD2/3/4/7; Figure 2-8a). Therefore, consistent with other studies, these results indicated an association between reduced SMAD expression and poor survival [124,126].

Whereas we observed higher gene deletion frequencies in EL-PDAC tumours than in QM-PDAC and C-PDAC tumours, QM-PDAC and C-PDAC tumours displayed higher gene amplification frequencies (Figure 2-8C). We observed 118 that were common across all the PDAC subtypes (Figure 2-8B), mostly impacting genes involved in diverse cell signalling pathways (Figure 2-8B and 2-8D); a finding consistent with the hypothesis that, like most other cancers, PDAC is primarily a consequence of disrupted signal transduction pathways (Figure 2-8D and 2-8E) [99,127].

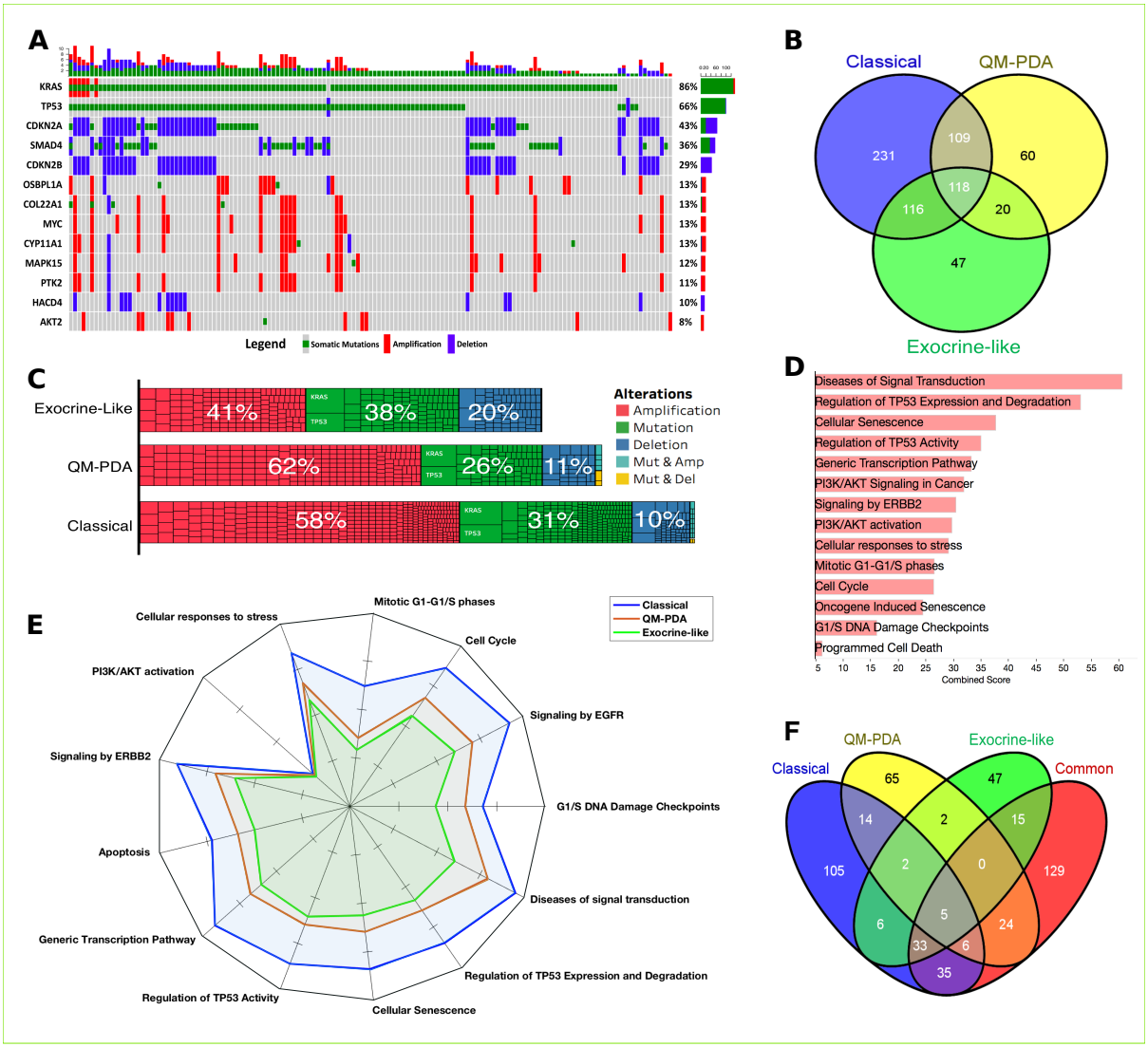


Figure 2-8: *Mutation and gene copy number analyses:* ***(A)*** *Genes with the most alterations in PDAC tumours. The only genetic alterations considered are mutations in, and amplifications of known oncogenes, and mutations in, and deletions of, known tumour suppressor genes.* ***(B)*** *The distribution of alterations among the three PDAC subtypes. Refer to supplementary file 5 for details concerning alterations in each set.* ***(C)*** *The extent of genomic alterations expressed as a percentage of total numbers of alterations found within the tumours of each PDAC subtype. Each cell in the bar-grid represents a mutant gene.* ***(D)*** *Reactome pathway enrichment results of the 118 genes that are commonly altered in tumour cells of all three PDAC subtypes. Refer to supplementary file 6 for the complete list of Reactome pathways that represent significantly enriched genetic alterations in the different PDAC subtypes.* ***(E)*** *The predicted extent of mutation-induced pathway dysregulation for the different PDAC subtypes.* ***(F)*** *The distribution of mutation-induced pathway dysregulations for mutations specifically associated with particular PDAC subtypes and common pathway enrichment. The non-overlapping pathways are those disrupted only in single PDAC subtypes (see supplementary file 6).*

We uncovered little overlap in the signalling pathways that were impacted by genetic alterations that were observed in only one of the PDAC subtypes (Figure 2-8F and Figure 2-9A; 2-9B; 2-9C). Further, we found that 62% of the pathways affected by mutations were altered in tumours belonging to at least two PDAC subtypes, whereas only 18%, 11%, and 8% of altered pathways were unique to C-PDAC, QM-PDAC, and E-PDAC tumours, respectively. This suggests that only a small proportion of mutations contribute to differences in the signalling pathway perturbations that are seen between the subtypes. Conversely, we therefore also suggest that most of the oncogenesis promoting mutations perturb pathways that are active in all three of the PDAC subtypes.

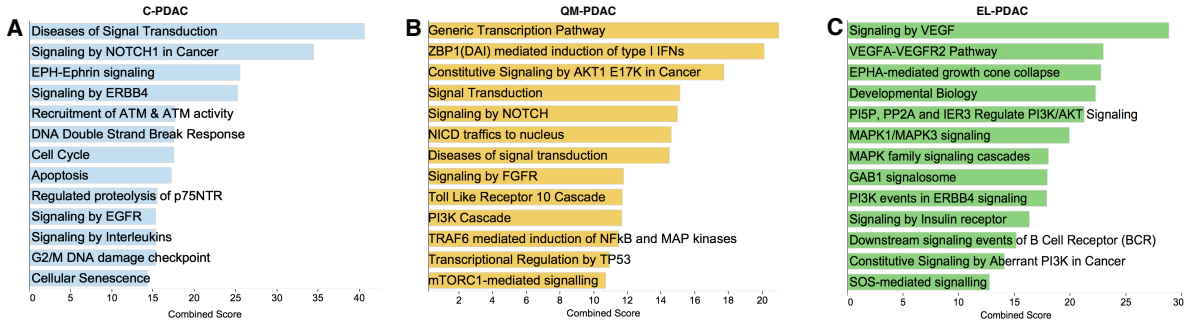


Figure 2-9: *Enriched Reactome pathways: Showing the top ranked dysregulated Reactome pathways based on mutations that are specific to each pancreatic cancer subtype. The size of each bar represents the combined score for each pathway (also see supplementary file 5).* ***(A)*** *Top 13 pathways over presented by mutations unique to C-PDAC tumours.* ***(B)*** *Top 13 pathways over presented by mutations unique to QM-PDAC.* ***(C)*** *Top 13 pathways over presented by mutations unique to EL-PDAC tumours.*

2.2.4 Integrative pathway analysis

We used the co-occurring mutated driver pathway (CoMDP) mathematical algorithm to discover *de novo,* two co-occurring pathways that may be driving the progression of PDAC; the first pathway involved the genes *KRAS, COL4L4 and FBWX7*, and the second involved the genes *SMAD4, TP53, PHF24, PRG4, PI3KCA, RPTOR and EP300* (Figure 2-10A) [128]. We expanded the CoMDP solution pathways using known protein-protein interactions to generate a network enriched with *MAPK, PI3K, and TP53* pathway members. Here, we found that two of the genes in the CoMDP solution pathways, *PRG4* and *PHF24*, did not map to any signal transduction pathway, emphasising the fact that the roles of some potentially key proteins in oncogenesis still

need de be defined. In particular, virtually nothing is known about *PHF24* [129].

To further investigate the degrees of alteration that are evident in the expanded CoMDP solution networks for the different PDAC subtypes, we mapped a combined dataset of mRNA transcript levels, protein expression levels, mutations and CNA onto these networks. We found that alterations in p53 and cell cycle checkpoint pathway genes were most apparent in C-PDAC tumours (Figure 2-10B and 2-10C), whereas alterations in specific MAPK and PI3K-mTOR pathway genes were more apparent in QM-PDAC tumours (Figure 2-10D and 2-10E). We found that the PI3K-mTOR and MAPK pathways were altered in 93% of all PDAC tumours. Alterations, included the activation of, among others, the PI3KCA (in 13% of tumours), AKT (in 27%) and BRAF (in 9%) genes, and the inactivation of the PTEN (in 4% of tumours), TSC1/2 (in 7%) and FOXO3 (in 5%) genes. Alterations in the *PI3KCA* oncogene and its negative regulator, *PTEN,* occur in cancers of the colon, breast, and prostate: cancers where the co-occurrence of *PI3KCA* and *PTEN* mutations appears to both drive oncogenesis, and reduce anticancer drug sensitivity [98,130]. Furthermore, we observed that PIK3R1 (the regulatory subunit of PI3K) was inactivated in 6% of PDAC tumours; inactivated PIK3R1 promotes the phosphorylation of AKT, which itself promotes oncogenesis as it activates numerous OGs and inhibits TSGs within the cell [131,132].

Alterations in the p53 and cell cycle pathways are frequent in cancer, and here we found that such changes were apparent in 85% of the tumours examined [133]. In addition to activated MDM2 and MDM4 (which both inhibit p53 activity), MYC, and CDK2/4/6, we found that p53 was inactivated in 58% of all tumours. Similarly, CDKN2A, CDKN2B and ATM (a kinase that activates p53) were inactivated in 35%, 42% and 3% of all tumours, respectively [133,134]. This suggests that, within the p53 and cell cycle checkpoint pathways, hyperactivated OGs such as MYC, MDM2, CDKs and inactivated TSGs such as TP53, CDKN2A and CDKN2B, may act together to promote oncogenesis both by limiting the repair of damaged DNA, and by permitting affected cells to proliferate uncontrollably through the inhibition of apoptosis [133,134].

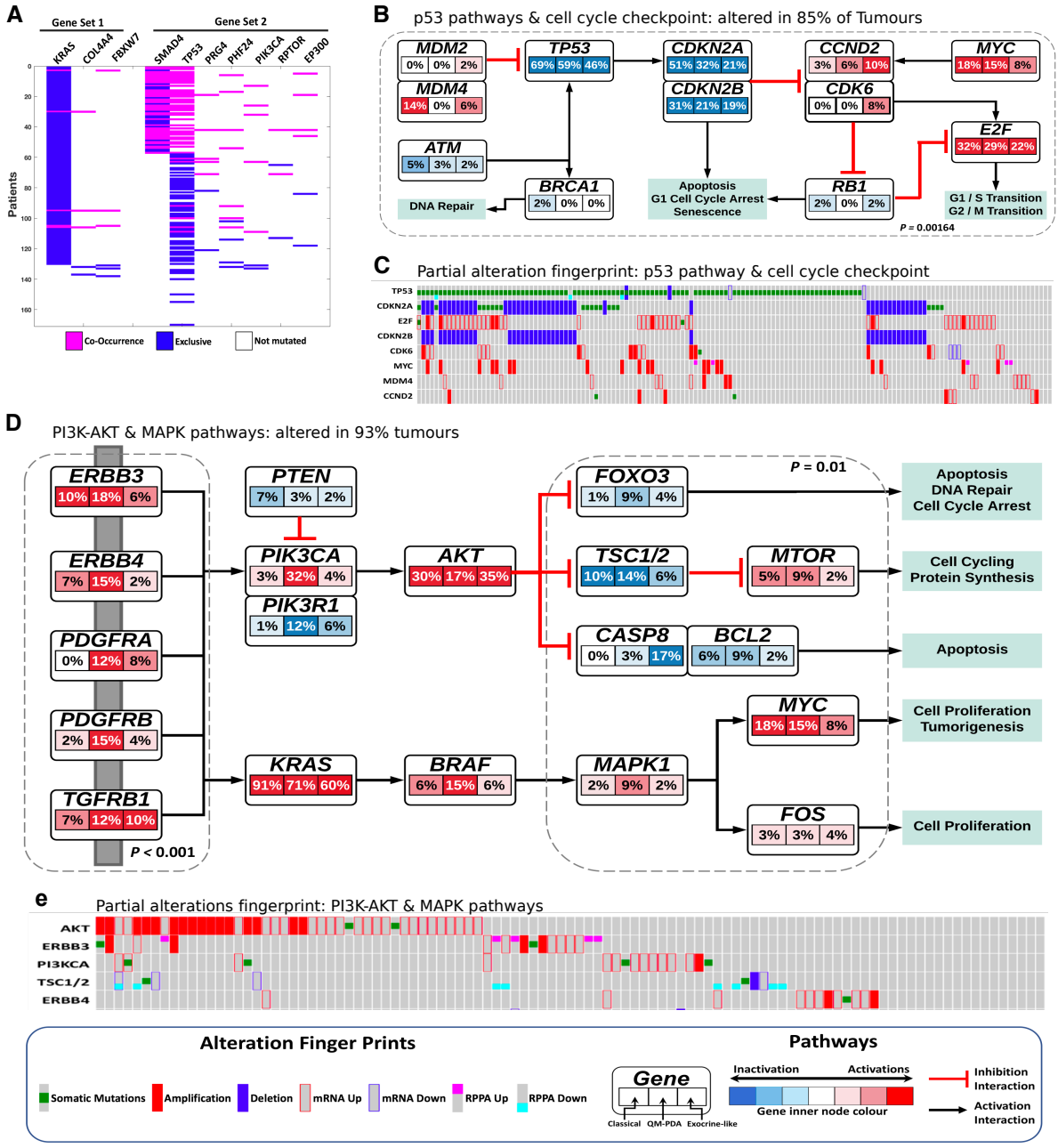


Figure 2-10: *Integrative pathway analysis:* ***(A)*** *The two co-occurring pathways in PDAC that are predicted to drive oncogenesis based on the mutational landscape representing two common driver pathways.* ***(B)*** *Alterations in p53 and cell cycle checkpoint pathways. p-value = C-PDAC vs Other subtypes, calculated using Fisher's exact test. Red indicates activating genetic alterations whereas blue indicates inactivating alterations. Darker shades correspond to higher alteration frequencies. Each node within the pathway represent a gene and the highlighted segments within each node and the percentage representing the alteration in the three PDAC subtype: C-PDAC, QM-PDAC and EL-PDAC from left, centre, and right, respectively.* ***(C)*** *The pattern of genetic alteration in selected genes that encode proteins involved in the p53 and cell cycle checkpoint pathways.* ***(D)*** *Alterations in the MAPK, RTKs and PI3K signalling pathways. p-values = QM-PDAC vs Other subtypes, calculated using Fisher's exact test. The connectivity of network components was extracted from Reactome Pathways, BioGrid and*

the literature. ***(e)*** *The pattern of genetic alterations in selected genes that encode components of the MAPK and PI3K-mTOR pathways.*

Consistent with previous observations, we found mutations, CNAs and changes in mRNA transcription and protein expression levels for proteins that participate in various signalling pathways that have previously been associated with pancreatic cancer [87,95,98]. Specifically, we observed alterations in the Notch (61%), apoptosis (32%) and NF-kβ (19%) pathways (Figure 2-11). Also, we found a variety of alterations in 41 genes that encode the ATP-binding cassette (ABC) transporter proteins in 72%, 82% and 79% of all tumours that belong to the C-PDAC, QM-PDAC and EL-PDAC subtypes, respectively (Figure 2-11D). The most altered ABC transporter gene in any PDAC subtype was *ABCC9*; found altered in 21% of all in QM-PDAC tumours. Overall, we observed that 78% of all PDAC tumours harboured a genetic alteration in at least one ABC transporter gene. ABC transporter-mediated energy-dependent efflux of a multitude of unrelated classes of anticancer drugs across membranes is a major cause of multidrug resistance and chemotherapeutic failures during cancer therapy [135,136]. Therefore, future efforts to determine tumour cell ABC transporter gene mutations that accentuate the activities of their encoded transporters are expected to guide precision medicine [137].

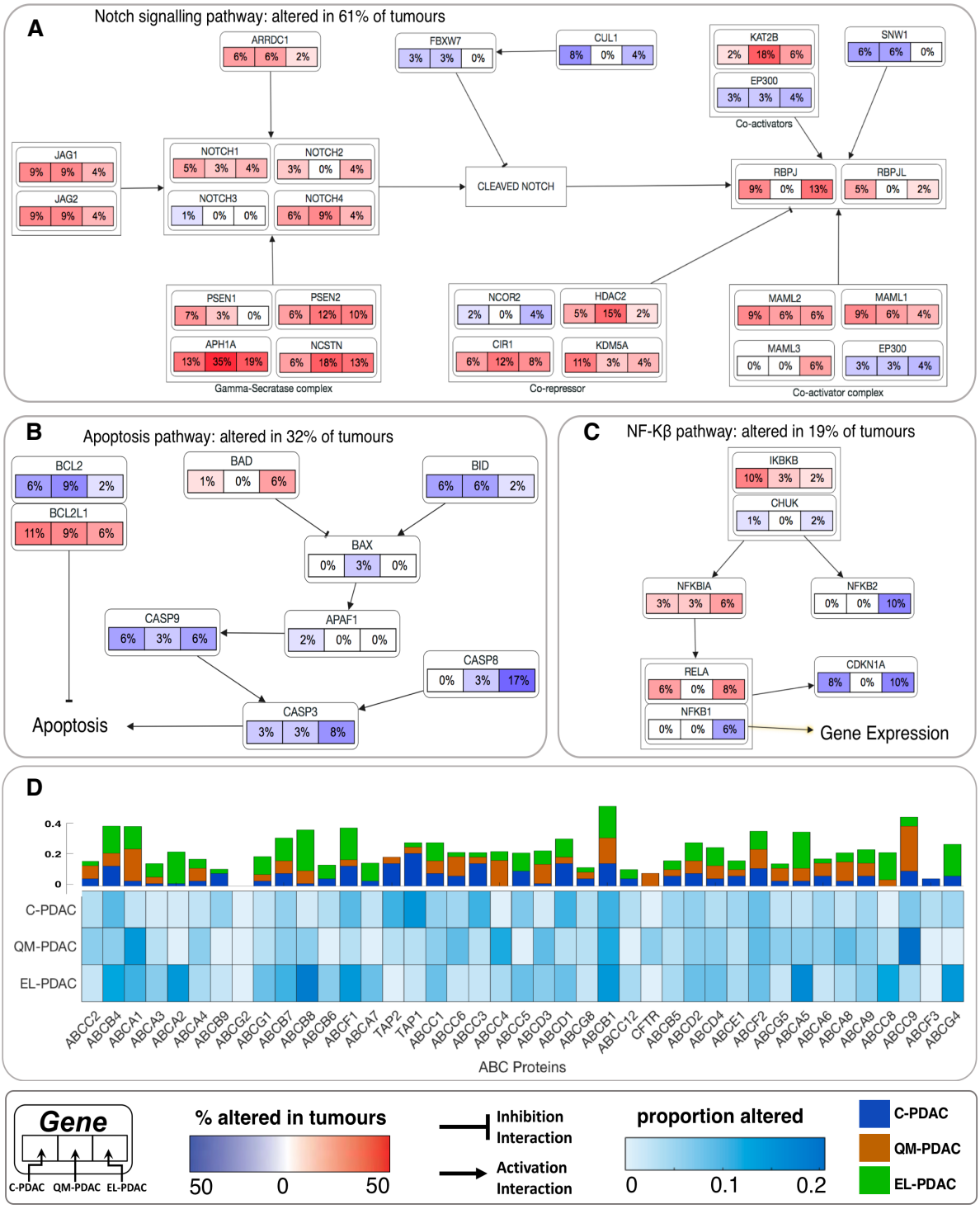


Figure 2-11: *Genetic alterations in three critical signalling pathways: Integrated alterations of mutations, copy number alteration, mRNA expression and protein levels are shown for* ***(A)*** *Notch signalling pathway* ***(B)*** *apoptosis pathway, and* ***(C)*** *NF-kβ pathway. Red indicates activating genetic alterations whereas blue indicates inactivating alterations. Darker shades correspond to higher alteration frequencies. Each node within the pathway represent a gene and the highlighted segments within each node and the percentage representing the alteration in the three PDAC subtype: C-PDAC, QM-PDAC and EL-PDAC from left, centre, and right, respectively.* ***(D)*** *The heatmap of integrated alterations in*

ATP-binding cassette encoding genes. The intensity denotes the frequency of alterations, with darker shades of blue representing a higher proportion of alterations.

2.2.5 Connectivity of genomic alterations to transcription factors and their pathway activities in pancreatic cancer

To link genomic changes to transcriptional events, we applied the Tied Diffusion Through Interacting Event (TieDIE) approach to reveal a protein interaction subnetwork that connects altered genes to transcription factors and their putative targets [138]; a network referred to below as the TieDIE subnetwork. Additionally, we used the PARADIGM-shift algorithm to infer pathway activity levels of all proteins that are known to participate in various signalling pathways in each of the three PDAC subtypes [139]. Furthermore, using heat diffusion analysis from the MAPK1 and TP53 nodes of the TieDIE subnetwork, we extracted two pathways that recapitulated signalling via the MAPK and p53 pathways.

The MAPK1 network was enriched with proteins whose associated mRNA transcription levels were significantly higher in QM-PDAC tumours compared to other PDAC subtypes. These proteins included EGFR (a receptor of the EGFR pathway that we found activated in QM-PDAC tumours), SOS1 and GRB2 (both of which are upstream signalling proteins in the canonical MAPK signalling pathway; Figure 2-12A) [140]. Also, the MAPK network connected TFs that are induced upon activation of the MAPK pathway, to proteins which are known to promote oncogenesis (such as FOS, JUN, ATF2 and ESR1). This inferred connectivity was further supported by the PARADIGM analysis which predicted that ESR1, JUN, GRB2 and CBL would have high degrees of activity [140,141].

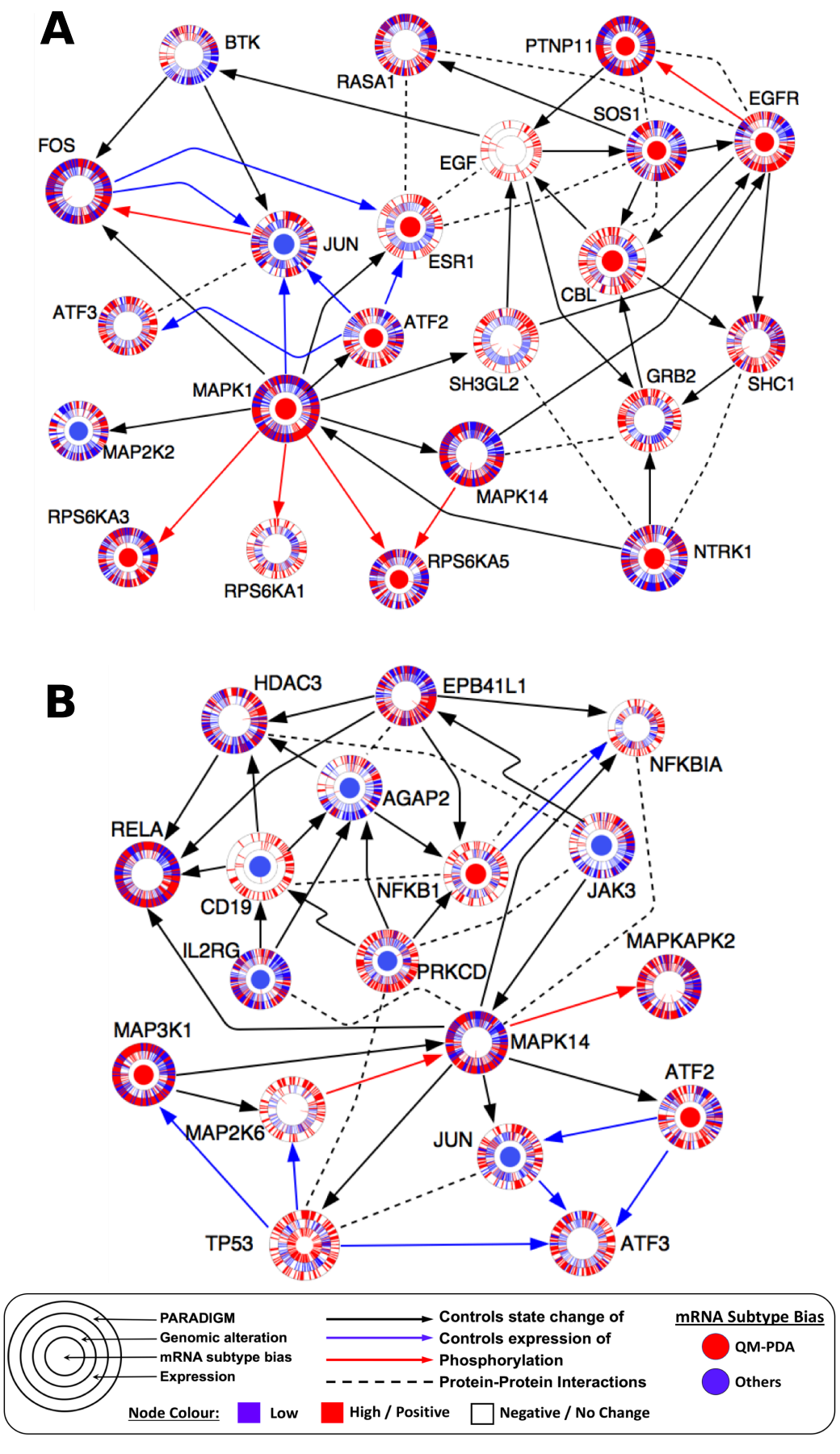

A
BTK
RASA1
PTNP11
EGFR
SOS1
FOS
EGF
JUN
ESR1
CBL
ATF3
ATF2
SH3GL2
SHC1
GRB2
MAPK1
MAP2K2
MAPK14
RPS6KA3
RPS6KA1
RPS6KA5
NTRK1
B
HDAC3
EPB41L1
NFKBIA
AGAP2
RELA
CD19
NFKB1
JAK3
IL2RG
PRKCD
MAPKAPK2
MAP3K1
MAPK14
ATF2
MAP2K6
JUN
TP53
ATF3
PARADIGM
Genomic alteration
mRNA subtype bias
Expression
Controls state change of
Controls expression of
Phosphorylation
Protein-Protein Interactions
mRNA Subtype Bias
QM-PDA
Others
Node Colour:
Low
High / Positive
Negative / No Change

Figure 2-12: *TieDIE subnetworks:* ***(A)*** *MAPK heat diffusion sub-network: pathway extracted from the TieDIE subnetwork using heat diffusion analysis from the MAPK1 network node.* ***(B)*** *p53 heat diffusion sub-network: pathway extracted from the TieDIE solution network using heat diffusion analysis from the TP53 network node. Each node indicates a pathway protein shown as concentric rings. The inner node denotes differential mRNA expression (Benjamini-Hochberg adjusted $p < 0.05$) bias for genes when comparing the QM-PDAC subtype tumours to those of the other PDAC subtypes (red = QM-PDAC bias, blue = bias towards other subtypes). The second ring indicates the presence of genomic alterations for that gene in each patient's tumour, with each patient's tumour being denoted by a spoke within the ring. The third ring shows mRNA expression levels for each tumour sample (red = high, blue = low). The outer ring indicates the PARADIGM inferred pathway activity for that protein in each tumour sample (red = high, blue = low). Arrows indicate known protein-protein interactions extracted from UCSC Super pathway, KEA, ChEA or inferred from the literature. We have attempted to make the visualisation clearer by omitting some interactions between some network nodes.*

The p53 network, on the other hand, was more prominent in C-PDAC and EL-PDAC tumours, and it connected signalling proteins to various TFs that are known to promote carcinogenesis, including ATF3, PRKCD, NFKB1 and NFKBIA [142–144]. These TFs were also predicted by PARADIGM to have a high degree of activity (Figure 2-12B).

Collectively these analyses emphasise that certain pathways may be more prominent than others in the different PDAC subtypes. Differences between the PDAC subtypes in the activity of specific signal transduction proteins suggests that some of these proteins could be targets of PDAC subtype-specific anti-cancer drugs (Figure 2-13 and Table 2-1).

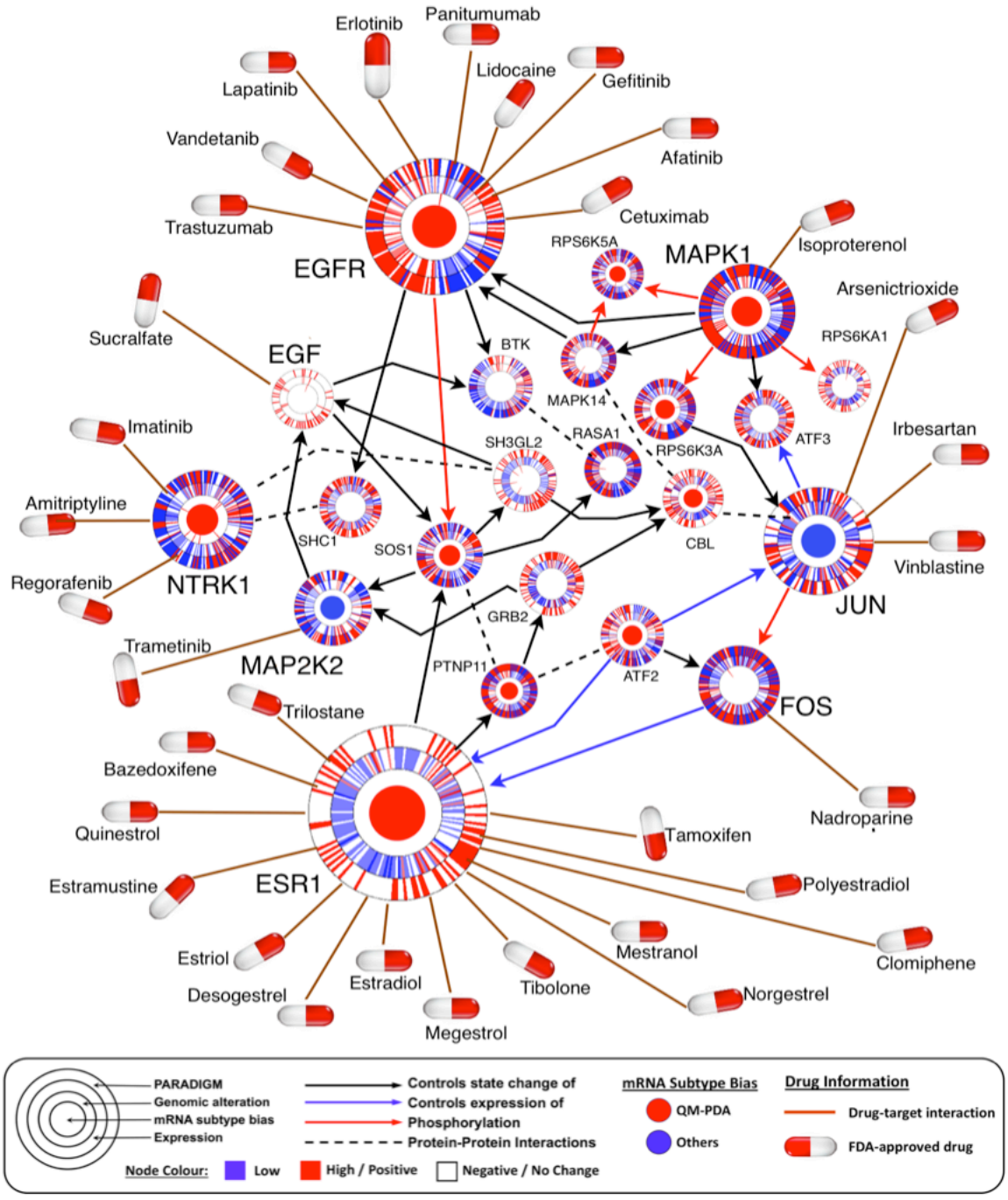


Figure 2-13: *MAPK heat diffusion sub-network: pathway extracted from the TieDIE solution network by heat diffusion for the MAPK1 network node and annotated with FDA approved anticancer drugs that are known to target specific pathway proteins. Each node indicates a pathway protein shown as concentric rings. The size of the node is proportional to the number of FDA approved drug for that particular pathway protein. The inner node detonates mRNA differential expression (Benjamini-Hochberg adjusted p < 0.05) biased for the statistically significant expressed gene between QM-PDA and other subtypes (red = QM-PDA bias, blue = bias towards other subtypes). The second ring shows the presence of genomic alterations for that gene in each sample that is denoted by a spoke within the*

ring. The third ring shows mRNA expression levels per samples (red = high, blue = low). The outer ring shows PARADIGM inferred activity for that gene in each sample (red = high, blue = low). Arrows indicate known protein-protein interaction extracted from UCSC Super pathway, KEA, ChEA or curated from the literature. We have made the visualisation easy by omitting some interactions between network nodes.

Table 2-1: *Some dysregulated pathway drug targets in clinical trial*

Drug	Target	Phases	Status	NCT Number
Palbociclib, Gedatolisib	CDK4/CDK6	Phase 1	Recruiting	NCT03065062
MK0752, gemcitabine, +	NOTCH	Phase 1	Completed	NCT01098344
NIS793, PDR001	TGF	Phase 1	Recruiting	NCT02947165
Nimotuzumab, Gemcitabine	EGFR	Phase 3	Recruiting	NCT02395016
MK-0646, Gemcitabine, Erlotinib	mTOR	Phase 1/2	Active, not recruiting	NCT00769483
Ixabepilone, Cetuximab	EGFR	Phase 2	Completed	NCT00383149
Irinotecan, Hydrochloride, Veliparib	ESR1	Phase 1	Recruiting	NCT00576654
Gemcitabine, Capecitabine, Erlotinib	EGFR/mTOR/PI3K	Phase 1	Completed	NCT00480584
Galunisertib, Durvalumab	TGF	Phase 1	Recruiting	NCT02734160
Everolimus, Octreotide, Acetate	mTOR	Phase 1	Completed	NCT01204476
Erlotinib, Gemcitabine, Oxaliplatin	EGFR	Phase 2	Completed	NCT01505413
Docetaxel, Irinotecan Hydrochloride	KRAS	Phase 2	Completed	NCT00042939
Cetuximab, Erlotinib, Hydrochloride	EGFR	Phase 1	Completed	NCT00397384
Cediranib, Maleate, Olaparib	ESR1	Phase 2	Recruiting	NCT02498613
Capecitabine, Cetuximab, Everolimus	mTOR/KRAS/EGFR	Phase 1/2	Completed	NCT01077986
Afatinib, Selumetinib, Docetaxel	PI3K	Phase 1/2	Recruiting	NCT02450656
Afatinib, Binimetinib, Capivasertib, +	CDK4/CDK6	Phase 2	Recruiting	NCT02465060

Targeted-therapy drug that are currently being evaluated in clinical trial for treatment of pancreatic cancer. + denotes more drug being used in combination. NCT denotes the national clinical trial. Information concerning clinical trials was obtained from www.clinicaltrials.gov.

2.3 Discussion

Through comprehensive transcriptomic and integrative profiling of pancreatic cancer, we have uncovered various functional alterations and signalling pathway perturbations and revealed how these alterations and perturbations might be associated with clinically relevant differences between patients with different PDAC subtypes. In particular, the discovery that QM-PDAC tumours are characterised by what is likely to be ESR1 and NTRK1 transcription factor-mediated over-activation of genes associated with the EGFR and TGF-β pathways, provides a rationale to target these tumours with drugs that either downregulate ESR1 and NTRK1, or inhibit EGFR and TGFBR2 (Figure 2-13) [145,146].

Furthermore, we find that, in general, PDAC is characterised by pervasive RTK, MAPK and PI3K-mTOR alterations [94,95,147] and that over 90% of the potentially oncogenic alterations occur in genes that are directly involved in the RTK, MAPK and PI3K-mTOR signalling pathways. It is well established that PDAC tumours frequently respond to MAPK and/or PI3K-mTOR pathway inhibitors [148,149]. While cases where these inhibitors have failed to provide therapeutic benefits have highlighted the heterogeneity of PDAC, they have also emphasised the importance of finding additional drugs that are either more generally applicable to PDAC treatment, or which can be used to target the signalling pathways that are most relevant for specific PDAC subtypes [149–151]. We identified that the most prominent cell signalling changes in EL-PDAC and C-PDAC tumours were within the p53 and cell cycle checkpoint pathways; hinting that these tumours might respond to cell cycle inhibitors. Consistent with our findings, other PDAC studies have also reported co-occurring mutations in genes involved in the p53 and cell cycle pathways. Collectively these studies provide a rationale for potentially treating PDAC using an approach that synchronously targets all of these pathways [152–154]. Furthermore, we have uncovered other receptors, intermediary signal transduction proteins and TF targets that may drive oncogenesis: all of which could be targeted by drugs designed to specifically treat QM- C- or EL-PDAC tumours.

Alterations in metabolism and cellular bioenergetics are hallmarks of cancer cells and represent an active area of research that is anticipated to yield novel anti-cancer drugs that could be used in combination with targeted-therapies or chemotherapy [155–157]. Here, we found that QM-PDAC and, albeit to a lesser extent, C-PDAC tumours exhibit a Warburg metabolic phenotype (Figure 2-4) [153]. Associations between the Warburg phenotype and both increased disease aggressiveness and poorer clinical outcomes have been previously reported [153,156]. As expected, we observed decreased overall survival and a shorter duration of disease-free survival in patients with QM- and C-PDAC tumours (i.e. tumours with the Warburg phenotype) relative to patients with EL-PDAC tumours (i.e. those without the Warburg phenotype). In this regard,

scrutinising the metabolic differences between PDAC tumour subtypes is likely to yield further leads for the development of novel therapeutic approaches.

We observed that most of the genomic alterations which are found within PDAC tumour cells are found in tumours belonging to all three of the defined PDAC subtypes. This finding suggests that improved responses to targeted-therapies may be achievable by systematic targeting of hub kinases within the multiple alternative signalling pathways that enable cancer cells to frequently acquire resistance [158,159].

By integrative analyses of genomic, transcriptomic, and proteomic data, we have uncovered novel signalling pathway aberrations that exist in PDAC tumour cells at the DNA, mRNA and protein levels. Altogether, our analyses have revealed widespread signalling network perturbations in PDAC subtypes, many of which could likely impact treatment outcomes and which are therefore also potential targets for novel anticancer drugs.

2.4 Methods

We obtained data for 185 PDAC patients involved in the TCGA project. Besides treatment outcomes these data include: whole exome sequence (WES; n = 185), transcriptome data (determined using RNAseq; n = 179); DNA copy number and mutation data (n = 179), and targeted proteome data (determined using RPPA; n = 123). Not all types of data were available for all patients because of assay failures, incomplete specimen availability and quality of issues with certain samples. All data used in our analyses are available from the TCGA website; https://portal.gdc.cancer.gov/repository.

2.4.1 Transcriptome-based classification

We performed unsupervised hierarchical clustering on RNA-seq data to identify three distinct PDAC subtypes. Before clustering, we removed data for unexpressed genes and genes that exhibited little variation between patients. Then, using the

transcriptomic classification framework established by Collision *et al.* [97], we classified the pancreatic cancer clusters as C-PDAC, QM-PDAC and EL-PDAC; respectively corresponding to clusters 1, 2 and 3 [97]. To return only exemplars for each cluster, we applied an anomaly detection algorithm based on an approximate Gaussian distribution [160]. Finally, we further validated the consistency of tumours within each cluster using a support vector machine classifier which yielded an average 10-fold cross validation classification accuracy of 95.5% over ten models (Figure 2-1F). We have summarised the distribution of tumour grades across these PDAC subtypes in Figure 2-1G.

2.4.2 Treatment outcomes

We integrated mRNA expression-based classification of PDAC subtypes with clinical information to review tumour characteristics specific to each of the PDAC subtypes. The Kaplan-Meier method was used to estimate overall survival and the duration of disease-free survival in a pairwise manner between subtypes [160]. Furthermore, the Fisher exact test was used evaluate associations between tumour subtypes and various clinical variables including treatment outcomes at the first, and later courses of treatment.

2.4.3 Differential gene expression, functional and pathways analyses

The identification of differentially expressed genes was performed in MATLAB using an implementation based on the negative binomial model (see supplementary file 1) [161,162]. Gene set enrichment analysis (GSEA) was employed to extract knowledge of overrepresented Gene Ontology (GO) terms for various functional processes and signalling pathways between molecular subtypes [105,163]. Complete GSEA results are provided in supplementary files 2A, 2B and 2C, for C-PDAC vs QM-PDAC, C-PDAC vs EL-PDAC and QM-PDAC vs EL-PDAC, respectively. Visualisation of significantly enriched GO terms of functional process and signalling pathways between subtypes was done in the Cytoscape plugin, Enrichment Map [164,165]. Furthermore, the mapping of gene expression levels onto template WikiPathways of the EGFR and TGF-β signalling pathways and the electron transport chain was done using the software PathVisio 3 (See Figure 2-5, 2-6, 2-7A, and 2-7B) [166,167]. For

this, we used z-score normalised expression data categorised into three levels: 1) Low for z-scores below -0.5; 2) no change for z-scores between -0.5 to 0.5; and 3) high for z-score above 0.5. The highlighted scale was chosen to consistently capture variations in gene expression across the entire pathways.

2.4.4 Prediction of regulator kinases

We computationally predicted upstream regulatory kinases that likely effect the observed differences in the gene expression signatures between QM-PDAC and the other PDAC subtypes using Expression2Kinases (X2K) [168]. X2K employs a reverse engineering network-based approach to predict upstream regulatory kinases based on prior knowledge. We obtained a list of differentially expressed genes between QM-PDAC and the others PDAC subtypes: 242 up-regulated genes and 1011 down-regulated genes based empirical Bayes statistics. Using this gene list, we predicted upstream regulatory TFs that are likely to be responsible for the observed changes in gene expression using the Chromatin Immunoprecipitation (ChIP) Enrichment Analysis (ChEA; 2016) [169]. In the next analysis, we linked the top 10 predicted TFs to upstream regulatory mechanisms by generating a TF-intermediate protein-protein interaction sub-network based on prior knowledge (Figure 2-14). Intermediate protein-protein interaction sub-network had 180 nodes with 1816 edges and is enriched in co-regulators, kinases and TFs that are experimentally verified to physically interact. Finally, we analysed the sub-network for enriched targets of known protein kinases that are likely to phosphorylate proteins within the sub-network using the Kinase Enrichment Analysis (KEA; 2015) [170]. See supplementary file 3 for a full list of computationally predicted kinases and their rankings.

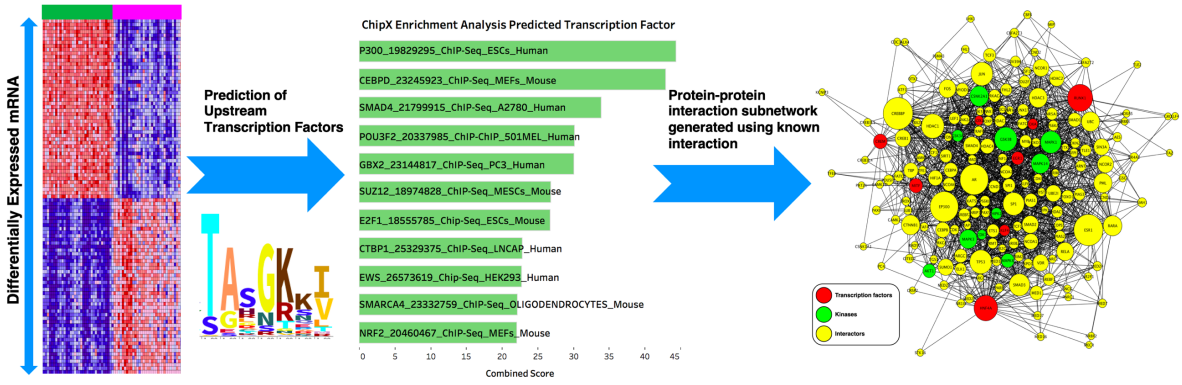


Figure 2-14: *Expression2Kinases pipeline. The computationally predicted kinases are extracted from the protein-protein subnetwork and ranked according to the number of targets that each kinase phosphorylated with in the subnetwork.*

2.4.5 Mutation and copy number alteration analyses

We evaluated the scope of genomic alterations in PDAC subtypes using significantly mutated genes, and copy number alteration identifications obtained from MutSigCV and GISTIC2.0 outputs, respectively [171,172]. Data for the genomic alteration analysis was processed as follows: oncogenes (OGs) and tumour suppressor genes (TSGs) in the samples were annotated using information from multiple sources. These include the Sanger Consensus Cancer Gene Database (699 OGs and TSGs), the UniProt Knowledgebase (304 OGs and 741 TSGs), the TSGene database (1,219 TSGs) and the ONGene database (725 OGs) [36,47,173,174]. Collation of data from these sources yielded a list of 3,688 OGs and TSGs, representing 2,773 unique genes (969 OGs and 1,804 TSGs). We utilised this list of OGs and TSGs to extract genetic changes anticipated to have a potential impact on the oncogenesis of pancreatic cancer. Explicitly, we returned only gain-of-function mutations and gene amplifications for known OGs. Also, for known TSGs, we returned loss-of-function genetic changes that involve mutations and deletions. Using these processed data, we identified frequently altered genes that likely have detrimental impacts concerning pancreatic carcinogenesis. We compared gene mutations between the PDAC subtypes to generate lists of mutations that are common among subtypes or unique to particular subtypes (see supplementary file 4). Using these lists, we performed a Reactome pathway enrichment analysis by querying Enrichr either with genes that were consistently altered in tumours of all three PDAC subtypes, or with genes that were altered in only one of the PDAC subtypes (see supplement file 5) [175].

2.4.6 Integrative analysis of expression and genomic alterations

2.4.6.1 Identification of co-occurring driver pathways

To discover driver pathways based on the patterns of mutations associated with PDAC, we applied the CoMDP algorithm which employs a mathematical programming method to identify *de novo* driver pathways in cancer from mutation profiles [128]. Briefly, this method identifies pathways that have a set of mutated genes with both

high coverage (i.e. present in the tumours of multiple individuals) and high exclusivity, and the pathways exhibit a statistically significant co-occurrence pattern. Using mutation data, we ran the CoMDP test with $K = 5$ to 11 (K equals the total gene set size) to return mutated driver pathways for all K values (Table 2-1). Genes in the CoMDP solution for $K = 10$ were connected using experimentally verified protein-protein interactions to generate an intermediary network. The solution network was enriched with members of the PI3K, MAPK, p53 and cell cycle regulation pathways. To visualise the extent of pathway aberration at DNA, mRNA and protein levels, we mapped genomic alteration data, mRNA transcript abundance data and protein expression data onto the network. For the genomic alteration data, we only considered gain-of-function mutations and gene amplifications for the OGs, and loss-of-function mutations and deletions for the TSGs. For the transcription and protein expression data we only considered OGs that had a degree of upregulation indicated by a > 2 Z-score and for the TSGs a degree of downregulation indicated by a < -2 X-score. The generated combined dataset was mapped on signalling pathways over-represented in the CoMDP solution network expanded using a prior-knowledge network. Additionally, plots of alteration patterns in genes among tumours were generated using the R package complex heatmaps [128]. Mapping of alterations onto genes in pathways shown in figure S8a, S8b and S8c were done using the software PathwayMapper [176].

Table 2-2: *Co-occurring gene sets in pancreatic cancer*

K	Gene set 1	Gene set 2	n1	n2	r1,2	Co-occurrence
5	*KRAS*	*SMAD4 TP53 PHF24 CDKN2A*	130	134	0.84	< 0.001
6	*KRAS CDKN2B*	*SMAD4 TP53 PHF24 CDKN2A*	136	134	0.89	< 0.001
7	*KRAS CDKN2B COL4A4*	*SMAD4 TP53 PHF24 CDKN2A*	138	134	0.9	< 0.001
8	*KRAS COL4A4 FBXW7*	*SMAD4 TP53 PHF24 RPTOR EP300*	135	128	0.88	< 0.001
9	*KRAS COL4A4 FBXW7*	*SMAD4 TP53 PHF24 RPTOR EP300 PRG4*	135	130	0.89	< 0.001
10	*KRAS COL4A4 FBXW7*	*SMAD4 TP53 PHF24 RPTOR EP300 PRG4*	135	132	0.91	< 0.001

11	*KRAS COL4A4 FBXW7 SALL2*	*SMAD4 TP53 PHF24 RPTOR EP300 PRG4 PRDM11*	137	132	0.91	< 0.001

In the Table, K is the combined gene set size (gene set 1 and gene set 2). Gene sets 1 and 2, represent co-occurring driver pathway in pancreatic cancer. n1 and n2 denote the coverage of genomic alterations in pancreatic cancer and $r_{1,2}$ is the ratio of the common coverage to their union coverage (i.e., co-occurrence ratio). Co-occurrence P represents the p-value of the co-occurrence significance of both pathways

2.4.6.2 Inferring gene activity from pathway analysis of copy number and expression data

We used PARADIGM-shift, a probabilistic graphical model approach that infers the activity of signalling pathway proteins by detecting differences in the expected activity of a protein on its downstream target relative to what is expected given it upstream modulator [128]. We ran PARADIGM with default settings using three datasets as inputs: (i) a dataset including only statistically significant CNA as determined by GISTIC2, (ii) a normalised gene expression dataset matching the CNA input file, and (iii) a custom UCSC Pathway formatted file. Pathway information of known gene interactions was created from various sources including Reactome pathways, KEGG Pathways, the KEA database, the ChEA database and the UCSC Super pathway [10,139,170]. PARADIGM predicted integrated pathway levels results are provided in supplementary file 6.

2.4.6.3 Identification of genomic perturbation associated with transcriptional changes

Genomic perturbations in PDAC subtypes were connected to associated transcriptional changes using TieDIE [138]. This method uses a heat diffusion process to identify relevant pathways that might be altered in tumours. To reveal sub-networks that distinguish QM-PDAC from the other PDAC subtypes, using genes that we found altered in at least 5% of all tumours, we generated a ranked list of genes that were differentially mutated between QM-PDAC tumours and those of the other PDAC subtypes using the Fisher's exact test (Figure 2-15A). The resulting genes are assumed to be responsible for the distinctive molecular signatures between subtypes—these were used as upstream inputs in TieDIE. A downstream input file was generated by computationally identifying the TFs that are most likely to be responsible for the difference in the transcriptome signatures between the QM-PDAC tumours and those of the other PDAC subtypes. The upstream and downstream input

files, together with a custom super pathway, were used in TieDIE to compute a sub-network connecting genomic alterations to transcriptional events (Figure 2-15B). The resulting sub-network had 158 nodes with 1127 edges and is enriched in mutated proteins and their TF targets likely response for the molecular differences between QM-PDA and other pancreatic cancer subtypes. We performed a secondary heat diffusion query on the TieDIE solution sub-network from the MAPK1 and TP53 nodes. Both MAPK1 And TP53 were flagged as being of likely importance based on both the numerous alterations of these genes within PDAC tumours and the pathway analysis that we had previously performed. These analyses produced two subnetworks that recapitulate signalling through the MAPK1 and TP53 pathways to their downstream TFs.

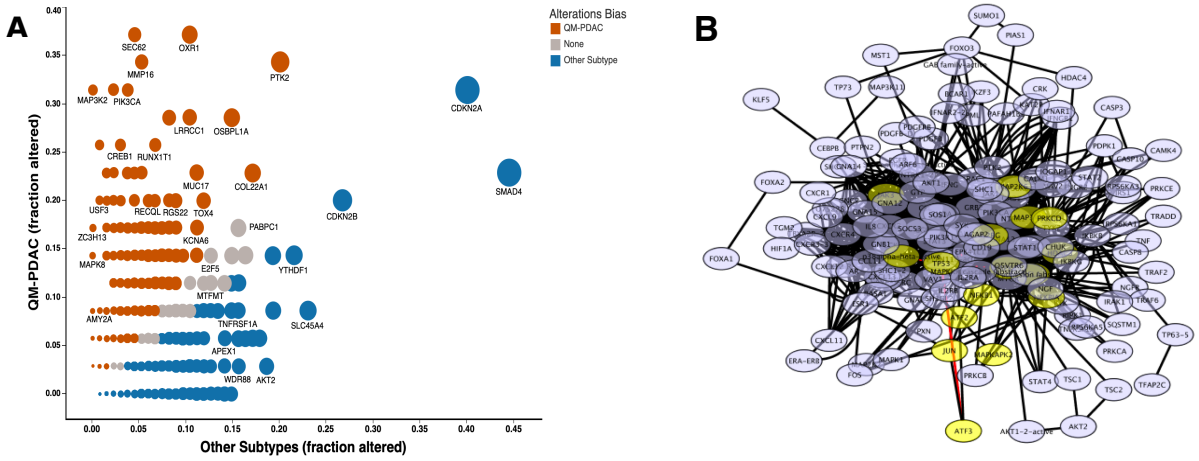


Figure 2-15: *Genomic alteration bias and TieDIE subnetwork:* ***(A)*** *Genomic alterations in QM-PDA vs. Other Subtypes. Each node represents a gene, who's colour shows details about the bias alterations and the node size is proportional to the overall alteration frequency in in pancreatic cancer. TP53 and KRAS are excluded from the plot and they don't show any bias.* ***(B)*** *TieDIE solution network: Visualisation of the TieDIE solution gene connectivity subnetwork. The nodes in highlighted in yellow represent the TP53 heat-diffusion network extracted the TieDIE solution network.*

2.4.7 Statistical analyses

Except were stated otherwise all statistical analyses were performed in MATLAB 2017b. The Fisher's exact test was used assess associations between categorical variables. The Wilcoxon rank sum test and Kruskal-Wallis test or independent sample Student *t*-test and one-way analysis of variance were used for continuous variables

were appropriate. Statistical tests were considered significant at $p < 0.05$ for single comparisons, and Benjamini-Hochberg adjusted *p*-values for multiple comparisons.

2.6 Supplementary Information

Supplemental Information can be found with this article online at https://doi.org/10.18632/oncotarget.25632

2.7 Funding

The H3ABioNet NIH Common Fund funded this study, grant number: U41HG006941

2.8 Ethics Approval

The University of Cape Town; Health Sciences Research Ethics Committee (HREC) IRB00001938 approved the protocol of this study.

Chapter 3 : Machine Learning and Network Analyses Reveals Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics

This section is a reformatting of a paper published in Scientific Reports [177]: https://www.nature.com/articles/s41598-020-58290-2

Musalula Sinkala, Nicola Mulder, Darren Martin

3.1 Introduction

Pancreatic cancer is a heterogeneous disease that is characterised by poor clinical outcomes and few effective treatment options. Attempts to define a standard classification for tumours of the pancreas have been ongoing for decades [178–180]. In general, the approaches that are currently used for making both outcome predictions and treatment decisions are based on histological subtyping and clinical parameters such as the disease stage, metastasis, and the resectability of tumours [181,182]. Recently, however, the advent of molecular profiling has laid the foundation for quantitatively profiling tumours based on their genome-wide gene transcription profiles, protein expression profiles and/or mutational landscapes [183–186]. These profiling methods promise a more accurate and precise definition of tumour subtypes and better predictions of how particular tumour types will respond to different treatments.

Further, molecular data that is used to construct the molecular profiles of particular cancers have been used to identify the perturbances in the cellular regulatory networks that characterize these cancers: often revealing numerous potential drug targets within various signalling pathways. These molecular data together with the known molecular profiles of numerous well characterized cancer cell lines can even be leveraged using

machine learning methods to predict the responses of particular patient tumour subtypes to different anticancer drugs [7,187].

A crucial resource for the discovery of useful diagnostic biomarkers and potential anticancer drug targets are large-scale datasets comprising, among other data types, extensive genomic, transcriptomic and proteomic profiles of matched healthy and tumorous tissues. These datasets, some of which are compiled and maintained by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are helping us uncover the molecular characteristics and signalling pathway perturbations that define specific cancer subtypes [9,96].

Among the cancers that are well represented in these data collections is pancreatic cancer. Molecular profiling analyses of the pancreatic tumour datasets have identified both distinct pancreatic cancer subtypes, and mutations of the genes, KRAS, TP53, SMAD4 and CDKN2A as potential drivers of pancreatic cancer [86,188–191]. Although the biomarkers that differentiate between different pancreatic cancer subtypes could eventually inform treatment decisions, there are as yet no available subtype-specific treatment options for this type of cancer. There is, therefore, a pressing need to, firstly, find a set of biomarkers that can be used to accurately and sensitively identify the molecular subtypes of pancreatic cancer and, secondly, to identify suitable targets for drug development among these biomarkers.

Definitions of disease subtypes is a perpetual process, with classifiers and cut-offs that differentiate between the subtypes, essentially needing to be continually re-defined and refined as more molecular data and better molecular profiling tools become available. As classification schemes for pancreatic cancers improve, it is expected that additional specific molecular correlates of patient survival, responses to anticancer drugs, and tumour aggressiveness will be uncovered. Armed with such knowledge, we could develop better prognostic and diagnostic methods, and select the best drugs to treat specific pancreatic cancer subtypes. Further, more subtype-specific molecular features could potentially enhance the accuracy with which machine learning methods could predict the drug response profiles of specific pancreatic tumours, thus leading to improved disease outcomes.

However, it remains technically difficult to effectively leverage the diverse and ever-increasing data relating to pancreatic tumours [192–194]. These difficulties include, but are not limited to, inconsistent classifications of patient tumours when the tumours are subtyped using different types of molecular data, and the efficient integration and analysis of different data types to yield consistent identifications of the causal disruptors of the molecular processes that underlie the observed differences between pancreatic cancer subtypes [192]. Ultimately, these difficulties undermine efforts to predict the responses of tumours to drugs: an endeavour involving comparisons between the relevant molecular features of a novel tumour with those of well-characterized tumour subtypes or tumour cell lines.

With these issues in mind, we attempted to identify clinically relevant subtypes of pancreatic cancer accounting for the full spectrum of molecular available for pancreatic cancer tumours in the TCGA dataset. We address the problem of inconsistent tumour classifications that are obtained using different types of molecular data, by applying an integrative classification approach that considered all the available molecular data types. As expected, our analyses identified discrepancies between various classification schemes but ultimately supported the existence of two major pancreatic cancer subtypes. Besides uncovering the likely molecular causes of altered biological processes within the tumours of these two subtypes, we identified biomarker sets that can be used to accurately and sensitively classify novel pancreatic tumours. Further, in the face of multiple high-dimensional data types, we show that statistical models that capture the complexity of disease can aid in the identification of relevant drugs and drug targets that might offer substantial benefits for patients afflicted with tumours belonging to either of the pancreatic cancer subtypes.

3.2 Results

3.2.1 Subtypes of pancreatic cancer and their clinical characteristics

We applied K-means clustering to the reverse phase protein array (RPPA) determined proteomics data of the 45 high-purity pancreatic cancer samples that are available in the TCGA to identify two coherent clusters of patient tumours [195]. Then, we compared this clustering of pancreatic cancer samples to other subtypes that are

reported in the literature for various other molecular data types (DNA methylation status, protein expression levels and mRNA/ miRNA transcription levels) and established that the samples clustered differently depending on the specific molecular data type used (Figure 3-1A).

To mitigate this problem, we applied a multi-platform integrative clustering method called similarity network fusion (SNF). SNF solves the disparate clustering problem by constructing similarity networks of samples for each available molecular data type and then efficiently fuses these into one network that represents clustering based on all the underlying data types (Figure 3-1B) [192]. Using DNA methylation status, protein expression, mRNA transcription and miRNA data of the 45 high purity cancer tumour samples available in TCGA, we applied the SNF clustering method to identify two-cluster and three-cluster clustering solutions (Figure 3-1C).

The pancreatic cancer subtypes in the two-cluster solution comprised 25 and 20 tumours, which we provisionally named as subtype-1 and subtype-2, respectively. Interestingly, the SNF clustering solutions were highly concordant with each of the clustering solutions obtained using individual molecular data types but were most similar to that obtained using the proteomics data (refer to Figure 3-1C).

Next, we sought to understand whether the identified pancreatic cancer subtypes were associated with different clinical outcomes. Indeed, we found that the two groups of patients differed with respect to the overall percentages of individuals with progressive disease and the percentages of individuals who eventually died. Here we found that the patients with subtype-1 tumours were more likely to survive than those with subtype-2 tumours (65% vs 35% survival, respectively; Figure 3-2A). We further observed a nearly 50% lower median disease-free survival (DFS) period for patients with subtype-2 tumours (DFS = 12.42 months) than for patients with subtype-2 tumours (DFS = 25.07 months; Figure 3-2B). Likewise, the overall survival (OS) periods for the patients with subtype-2 tumours (OS = 16.05 months) were shorter than those with subtype-1 tumours (OS = 23.06 months; Figure 3-3A). However, our analysis of OS and DFS periods using the Kaplan-Meier methods revealed no

statistically significant difference between the pancreatic cancer subtypes; possibly due to the small sample size (Figure 3-2B and 3-2B) [160].

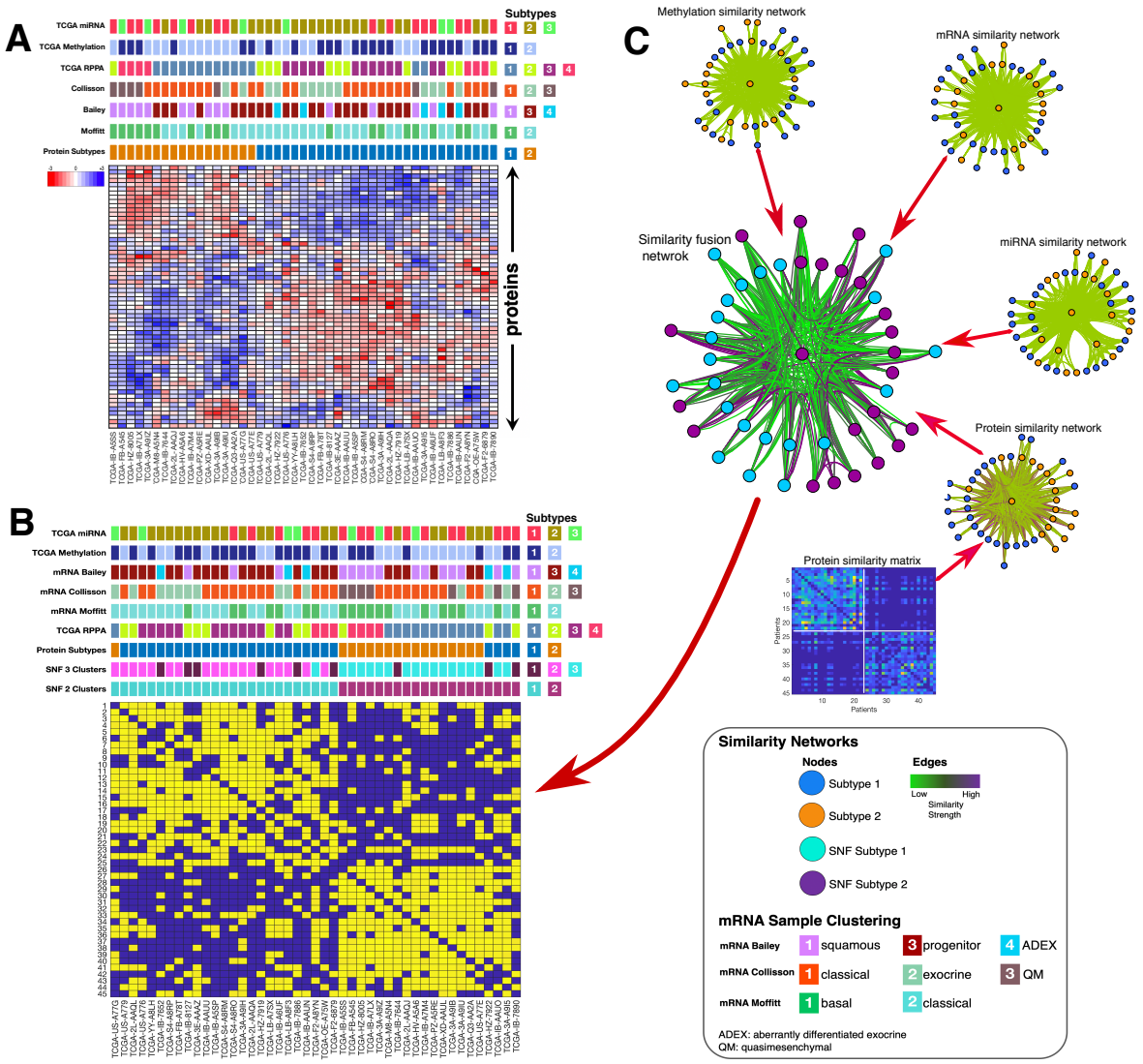


Figure 3-1: *Classification of pancreatic cancer:* ***(A)*** *Comparison between the proteomics-based subtyping of pancreatic cancers using unsupervised hierarchical clustering, to other classification schemes from top to bottom: TCGA's (Raphael et al, 2017) miRNA, RPPA, and DNA methylation; mRNA-based classification schemes using the gene biomarkers established by Collosson et al; Bailey et al; and Moffitt et al.* ***(B)*** *illustrative example of SNF steps: similarity matrices are used to create patient networks from protein, mRNA, miRNA and DNA methylation data showing patient-to-patient similarities for the 45 pancreatic cancer patients. The network nodes represent patients. The colours of edges joining nodes indicate the degree of similarity between pairs of patients. The nodes of the fused network are coloured according to the subtypes to which the patient tumours were assigned using spectral clustering of the combined patient network.* ***(C)*** *Comparison between the SNF subtyping using*

spectral clustering to other classification schemes from top to bottom: TCGA's [188] *miRNA, and DNA methylation classifications; mRNA-based classification schemes* [94,104]*; TCGA's RPPA classification, our K-means clustering classification; our 3-cluster SNF classification; and our 2-clusters SNF classification.*

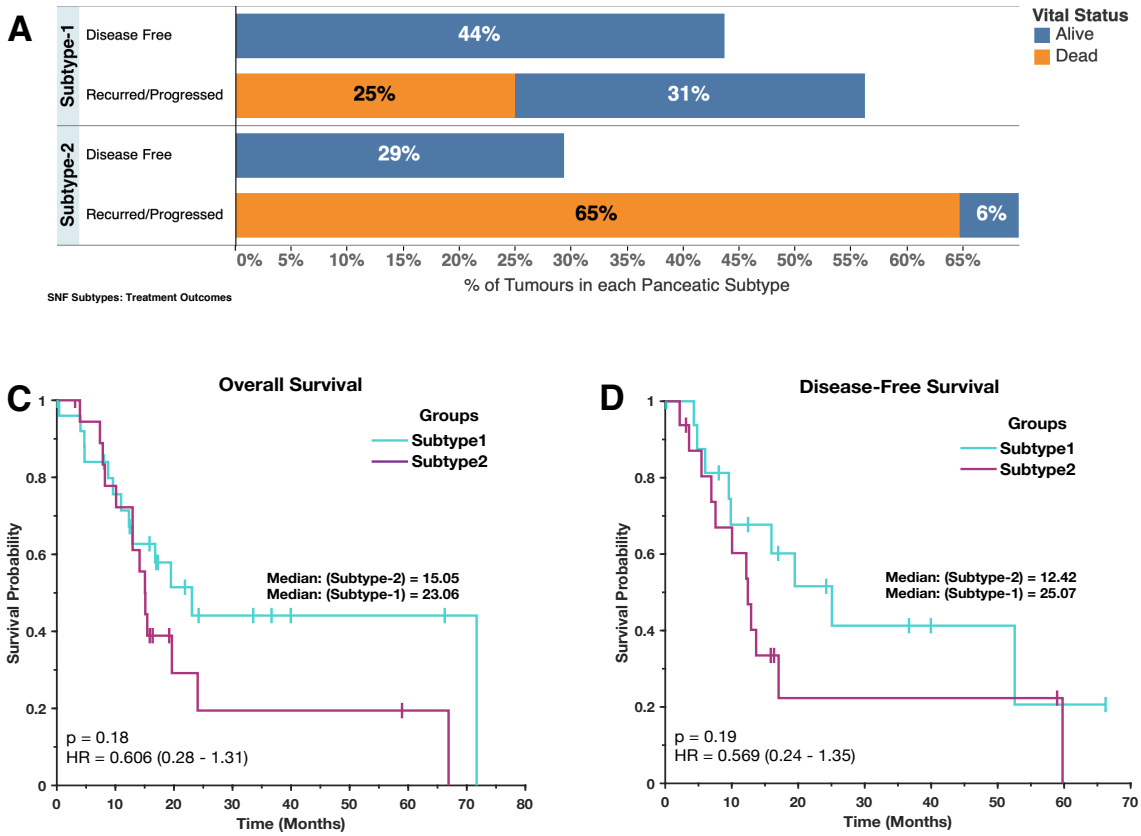


Figure 3-2: *Treatment outcomes:* ***(A)*** *Percentage of total number of pancreatic cancer patients for each DFS status after the first course of treatment broken down by the patient's vital status (dead or alive).* ***(B)*** *Kaplan-Meier curve of the disease-free survival months of patients afflicted by each of the pancreatic cancer subtypes* ***(C)*** *Kaplan-Meier curve of the overall survival months of patients afflicted by each pancreatic cancer subtypes.*

3.2.2 Proteomics-based signalling pathway analyses distinguish disease subtypes

For each disease subtype, we compared the enrichment of KEGG pathways and Gene Ontology (GO) biological process classifications of proteins that were upregulated within tumour belonging to each of the subtypes using Enrichr [175]. We found that whereas certain pathways were differentially altered between tumours belonging to different subtypes, other pathways were consistently altered (albeit to different extents

in some cases) in the tumours of both subtypes (Figure 3-3A, also see Supplementary File 1).

The mTOR signalling pathway was altered in subtype-1 tumours but not in subtype-2 tumours (combined score = 85, hypergeometric test; $p = 2.1 \times 10^{-19}$). Within the mTOR pathway of subtype-1 tumours, we found increased expression of well-documented oncogenes including MTOR and BRAF: both of which have previously been linked to pancreatic carcinogenesis (Figure 3-3B) [196–198].

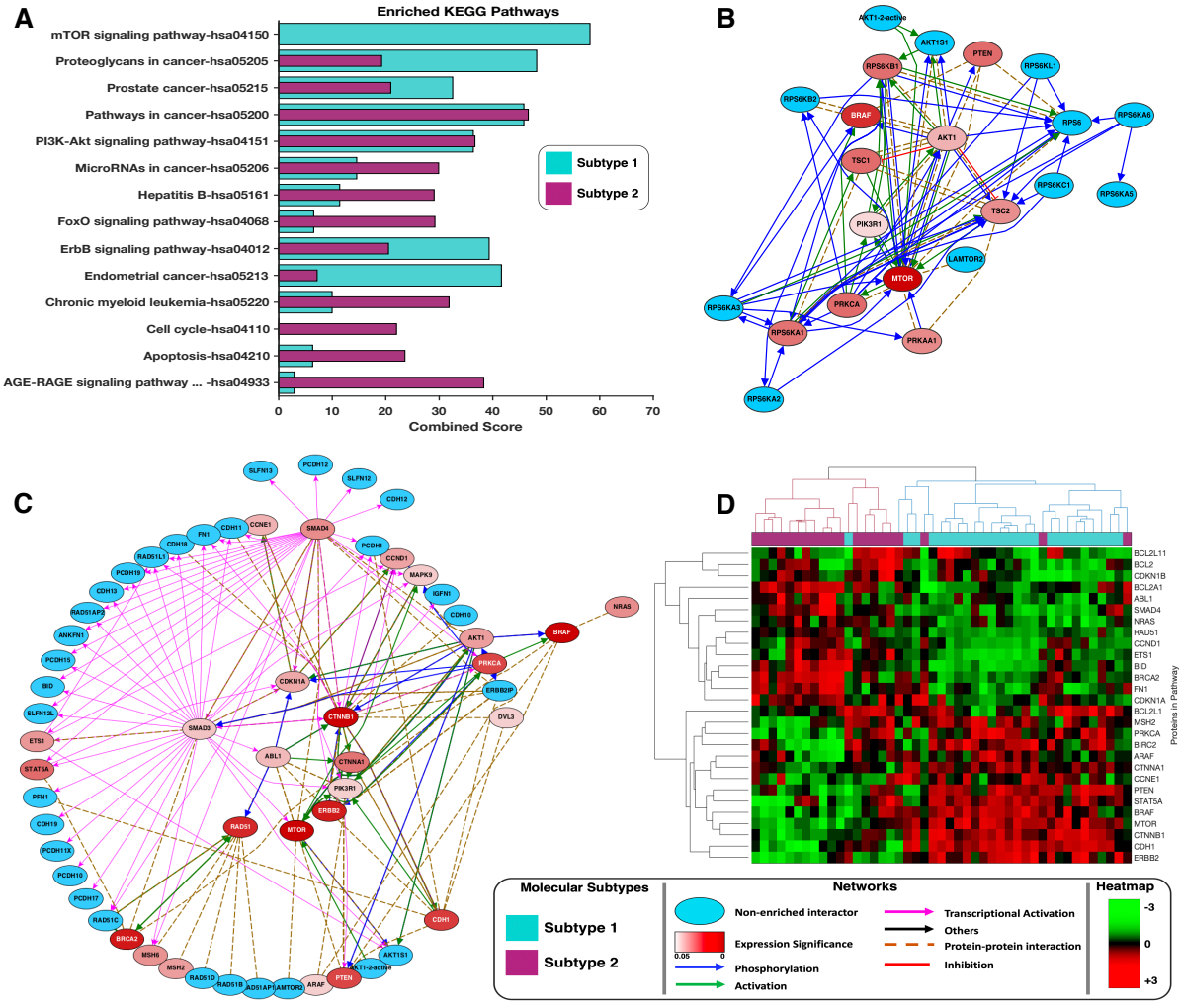


***Figure 3-3**: Pathway enrichment analyses: **(A)** KEGG pathways enrichment results of the most significantly altered pathways in tumours belonging to each of the inferred pancreatic cancer subtypes. Refer to supplementary file 1 for the complete list of KEGG pathways enriched based on the proteomics data. **(B)** mTOR signalling pathways found to be uniquely altered in subtype-1 tumours. Blue nodes indicate proteins with expression levels that were either not significantly altered between the subtypes or were not measured by the TCGA. Red coloured nodes represent proteins with significantly altered*

expression levels with the degree of statistical significance being expressed as the negative logarithm of Benjamin-Hochberg adjusted p-values. The connectivity of network components was extracted from the KEA, ChEA, and UCSC super pathway databases. ***(C)*** *KEGG cancer pathways found to be consistently altered in tumours belonging to both pancreatic cancer subtypes.* ***(D)*** *Clustergram of tumours using only the proteins that are members of the KEGG cancer pathways ontology.*

Further, we found that proteins that are involved in the KEGG cancer pathways were dysregulated in both the subtype-1 and subtype-2 tumours; these pathways encompass several known oncogenes (such as RAD51, BRAC1, and ERBB2) and tumour suppressor genes (such as PTEN and CDK2A1)[199–201] (Figure 3-3C). Despite the upregulation of the KEGG cancer pathways in tumours belonging to both subtypes, we found that the clustering of patients using only proteins within these cancer pathways was concordant with our subtype classification (Figure 3-3D). Such a clustering pattern indicates that even when the same pathways are altered both subtype-1 and subtype-2 tumours, the exact nature of the alterations within these pathways still differs between the two tumour subtypes. For example, whereas subtype-1 tumours exhibit hyperactivation of mTOR-associated signalling, subtype-2 tumours display increased activation of SMAD4-associated signalling. Also, we found that other proteins involved in mTOR signalling were both more strongly correlated and more highly expressed in subtype-1 tumours than they were in subtype-2 tumours, indicating the hyperactivation of this pathway requires the increased expression of most of the mTOR signalling proteins (Figure 3-4A). Likewise, SMAD4 signalling pathway protein expression levels also differed significantly ($p = 2 \times 10^{-4}$) between these subtypes (Figure 3-4B).

We further attempted to identify the kinases that likely phosphorylate substrates within the various signalling pathways of pancreatic tumour cells. Using kinase enrichment analysis (KEA), we found a subset of kinases that might drive pancreatic carcinogenesis, including, among others (Supplementary File 1), AKT1 ($p = 8.2 \times 10^{-03}$), MTOR ($p = 0.011$), and RPS6KA1 ($p = 0.0499$) (Figure 3-4C) [170]. We observed a moderate positive correlation between proteins involved in mTOR signalling and their phosphorylated forms (Figure 3-4D). Further, our results show that the protein phosphorylation pattern among the two pancreatic cancer subtypes is distinctive.

Here, we found that in subtype-1 tumours various phosphoproteins that participant in mTOR signalling – such as MTOR-pS2448, GSKB-pS21-S9, PDK-pS241 and growth factor receptors EGFR-pY1068 and ERBB-pY1248 – all exhibited increased phosphorylation (Figure 3-4E) [202,203].

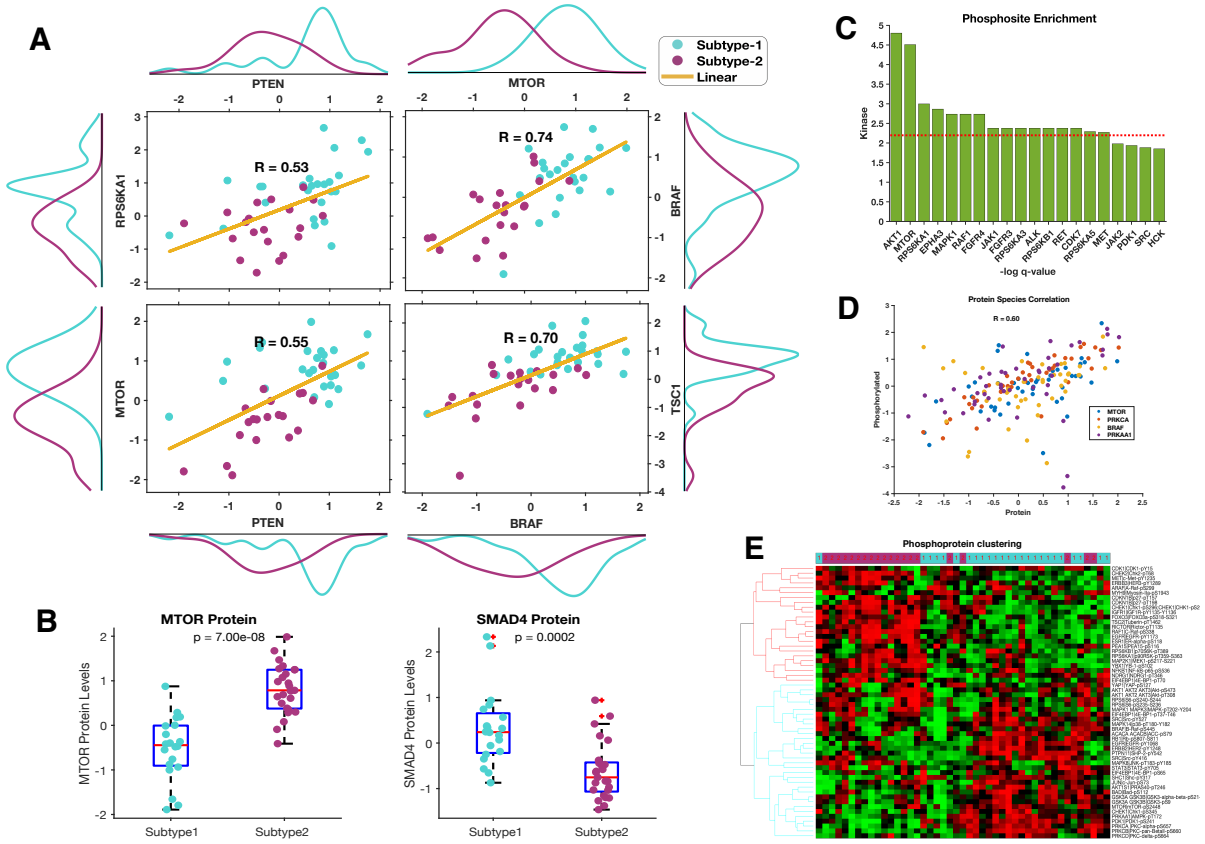


Figure 3-4: *Protein-level differences between pancreatic cancer subtypes:* ***(A)*** *Pearson's correlation values of some proteins involved in mTOR signalling. The plot shows relatively higher expression levels of these proteins in subtype-1 tumours compared to subtype-2 tumours. (B) Boxplots shows mTOR and SMAD4 protein expression biomarker of the subtypes.* ***(C)*** *Enriched phosphosites identified by kinase enrichment analysis: the negative logarithm values of the Benjamin-Hochberg adjusted p-value are plotted on the y-axis while kinases are plotted along the x-axis. The red line represents the cut-off values at the 10% false discovery rate.* ***(D)*** *Correlation between the phosphorylated and de-phosphorylated proteins species for proteins involved in the mTOR signalling pathway.* ***(E)*** *Unsupervised hierarchical clustergram of tumours phosphoproteins showing high concordance with the clustering obtained from all the proteins (de-phosphorylated and non-phosphorylated protein) profiled by the TCGA. The clustergram was produced using the Spearman correlation distance metric and the complete linkage.*

These phosphoproteomics analyses support our initial findings (using dephosphorylated proteins) that subtype-1 tumours display increased mTOR signalling. Conversely, for subtype-2 tumours, we found elevated phosphorylation

levels of proteins such as CDK1-pY15, p27-pT158 and p27-pT198 (Figure 3-4E) which are involved in cell-cycle-associated processes [204].

Overall, our findings suggest that for tumours of the two major pancreatic cancer subtypes, oncogenesis may be primarily driven by perturbation in either SMAD4 or mTOR signalling.

3.2.3 Pancreatic cancer subtypes exhibit functional differences in mRNA levels and DNA methylation patterns

We attempted to determine whether any GO molecular functions were enriched for among the overexpressed genes that differentiated the two pancreatic cancer subtypes. Specifically, we queried Enrichr using mRNA transcripts that were significantly upregulated across the tumours belonging to each particular cancer subtype (see Supplementary File 2) [175]. We found that the over-transcribed genes in subtype-2 tumours were enriched for, among others, molecular functions associated with transmembrane transporter and G-protein coupled receptor activities (Figure 3-5A, see Supplementary File 1). Alternatively, the over-transcribed genes in subtype-1 tumours were enriched for, among others, molecular functions that are associated with phosphoinositide 3-kinase signalling, peptidase enzyme activity and growth factor receptors (Figure 3-5A, see Supplementary File 1).

We explored the enriched KEGG pathways that were differentially expressed between the two pancreatic cancer subtypes using lists of genes with methylation profiles and mRNA transcription levels that differed between the subtypes (see Supplementary File 1). Interestingly, we found that only subtype-1 tumours displayed enrichment for pancreatic secretions (Figure 3-5B). These results corroborate both our previously noted enrichment in subtype-1 tumours of mRNAs involved in transmembrane transport, and published observations that the secretion of compounds from the pancreas and other organs is associated with increased transmembrane transporter activity [205].

Similarly, for both enrichment analyses using differentially expressed mRNA and proteins, we found enrichment for components of the AGE-RAGE signalling pathway in subtype-2 tumours (Figure 3-3A and 3-5B). The AGE-RAGE system promotes the development of various types of cancers, including those of the pancreas and prostate, through diminished apoptosis and increased cell viability [206,207]. Therefore, targeted inhibition of RAGE may serve as an effective treatment strategy against subtype-2 tumours.

In addition to these findings, the DNA methylation data revealed that while the methylation landscapes of subtype-1 and subtype-2 tumours were generally similar, the subtype-1 tumours had some additional genes displaying significantly increased DNA methylation (Supplementary File 2). We noted that these hypermethylated genes participate in various cellular pathways including focal adhesion, RAP1-signalling, and actin cytoskeleton regulation (Figure 3-5C). Since these DNA methylation alterations are unique to subtype-1 tumours, they could be associated with reduced pancreatic tumour aggressiveness.

Unexpectedly, we observed no significant differences in mutation distributions and gene copy number alterations for the genes with transcription and translation profiles that differed between the two subtypes (Figure 3-5D).

3.2.4 Biomarker genes, proteins and miRNA sets that define the pancreatic cancer subtypes

Given that different types of molecular data yield different patterns of tumour clustering, we attempted to identify biomarker genes, proteins or miRNA sets that best differentiated between the two pancreatic cancer subtypes. It was anticipated that these sets of biomarker genes might allow for consistent classification of pancreatic cancer patients using machine learning methods applied to only one category of molecular data.

To extract relevant features for each category of molecular data, we applied the diagonal adaptation of neighbourhood component analysis (NCA) for classification

with regularisation[208]. NCA learns feature weights for minimisation of an objective function that measures the average leave-one-out classification loss over the training data (Figure 3-9A and 3-9B in the methods section) [208].

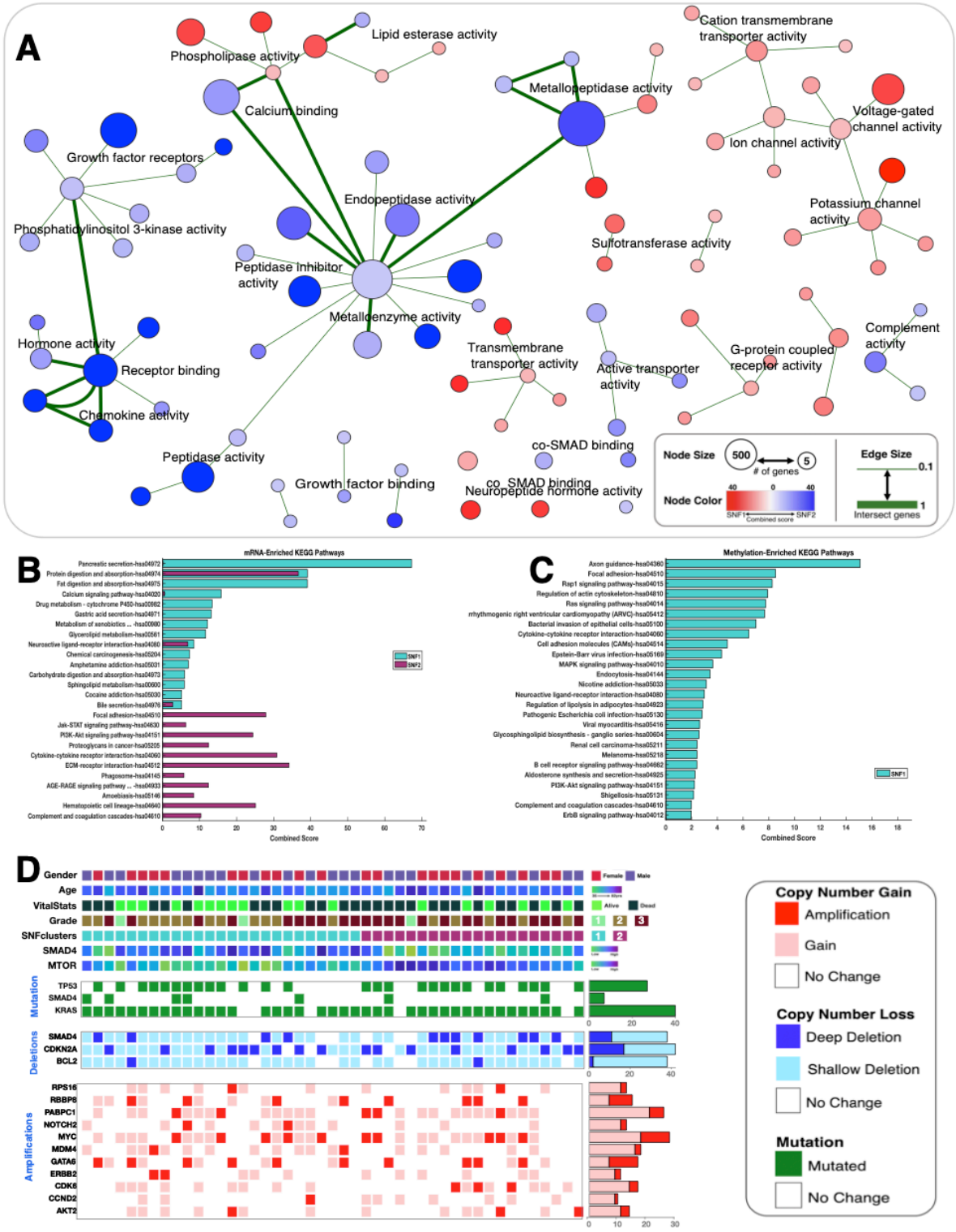


Figure 3-5: *Functional analyses and mutational landscape of pancreatic tumours:* ***(A)*** *Network of Gene Ontology (GO) molecular functions found enriched between the two pancreatic cancer subtypes. Enrichr was used to obtain enriched GO-terms that were visualised in Cytoscape (refer to the methods section). Each node represents a GO-term with similar nodes clustered together and connected by edges with the number of shared genes between the nodes being represented by the thickness of the edges. The size of each node denotes the gene set size of the represented GO-term. The colour of each node represents the magnitude of the combined enrichment score: red represent enrichment in subtype-1 tumours and blue represents enrichment in subtype-2 tumours. KEGG pathways: showing the top-ranked dysregulated KEGG pathways for each disease subtype based on the* ***(B)*** *mRNA transcript levels,* ***(C)*** *and DNA methylation levels.* ***(D)*** *The integrated plot of clinical and molecular features of 45 tumour samples ordered by their SNF clustering positions. From top to bottom panels indicate: patient gender; Age at which a condition or disease was first diagnosed; neoplasm histological grade; SNF subtype of tumour; SMAD4 protein expression level; mTOR protein expression level; significantly mutated genes: TP53, SMAD4 and KRAS gene mutations; SMAD4, CDKN2A and BCL2 gene deep deletion (dark blue) and shallow deletion (pale blue); gene amplification (red) and copy number gain (pink) of multiple genes.*

Using NCA, we identified biomarker sets comprising 50 mRNAs, 49 methylated genes, 14 proteins, and 20 miRNAs. For these biomarker sets, we separately applied hierarchical clustering to each of the different molecular data categories to consistently and accurately reproduce the pancreatic cancer subtype classifications (Figure 3-7A, 3-7B, 3-7C and 3-7D). Also, we individually applied supervised machine learning methods to the 50 mRNA, and the 49 methylated gene sets to classify tumours into subtype-1 and subtype-2 categories. For this, we used the K-nearest neighbour (KNN) algorithm for the mRNA expression data and the support vector machines (SVM) classifier (see methods section) for the DNA methylation data to achieve very accurate subtype classifications of the tumours (Figure 3-7E and 3-7F)[209,210]. Specifically, we observed five-fold cross-validation classification accuracies of 99% for the mRNA-based KNN classifier and 98% for the DNA methylation-based SVM classifier, with an agreement of 97% (see methods sections).

Decreasing the number of biomarker genes needed to accurately classify tumours from new pancreatic cancer patients would improve the utility of these sets in a clinical diagnostic setting. To identify smaller biomarker gene sets, we used supervised machine learning methods (see methods section) to define a biomarker set of fewer

than ten genes, miRNA or proteins that would minimise incorrect classifications. Also, we used these biomarker sets to consistently re-classify TCGA pancreatic cancer patients using hierarchical clustering (Figure 3-9 in the methods section). These results imply that smaller gene sets could potentially be useful in a clinical diagnostic setting.

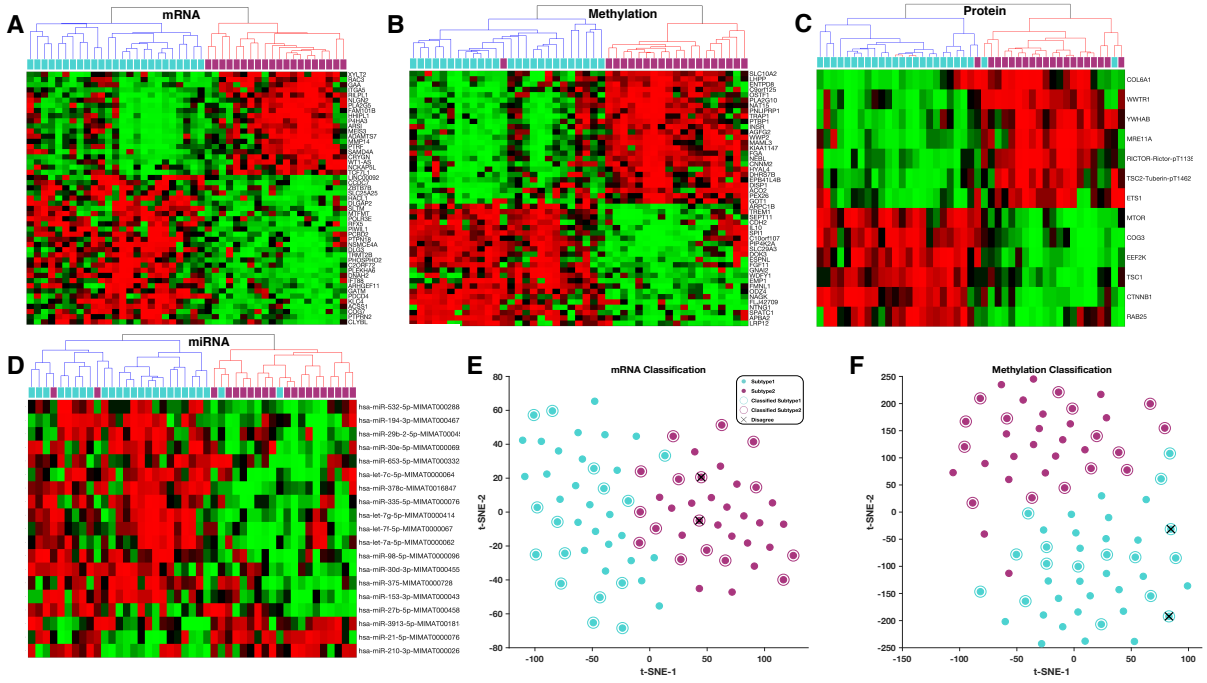


Figure 3-6: *Classification of pancreatic tumours using biomarker sets: Clustered heatmap of tumours using the* ***(A)*** *mRNA biomarker gene set,* ***(B)*** *DNA methylation biomarker gene set,* ***(C)*** *protein biomarker set, and* ***(D)*** *miRNA biomarker set. All the heatmaps (In A, B, C and D) were produced using unsupervised hierarchical clustering with the cosine distance metric and complete linkage. The coloured bars on each clustergram shows the original subtype classification of each patient's tumour found by applying SNF and spectral clustering to all molecular data sets.* ***(E)*** *Supervised classification of cancer patients using the mRNA biomarker set trained on a KNN-model,* ***(F)*** *the DNA methylation biomarker set trained on an SVM-model. For both plots (E and F), t-SNE was used to visualise the tumour classes using the exact algorithm and squared Euclidean distance metric. Circled points represent newly classified TCGA pancreatic cancer patients, whereas un-circled points represent the original 45 tumour samples that were used to train the models. Crossed points represent disagreement between the mRNA-based model and the DNA methylation-based model.*

To validate the performance of our 50-mRNA biomarker set, we downloaded pancreatic cancer data from the ICGC data portal [211]. Using the mRNA-based KNN classifier that was trained on TCGA data, we tested the reproducibility of the two-

subtype classification scheme by classifying 96 ICGC pancreatic cancer patients into subtypes-1 and subtypes-2 (Figure 3-8A). We also applied unsupervised hierarchical clustering to the mRNA biomarker set extracted from the ICGC RNAseq data to reproduce a two-subtype classification analogous to that obtained fusing the TCGA datasets (Figure 3-8B). The grouping of ICGC patients yielded by the supervised "TCGA classifier" and the unsupervised "ICGC classifier" agreed on the classifications of 94% of the patients. We observed that 5% of patients with posterior subtype membership probabilities that were less than 0.7, were more likely to be among the discordant cases, accounting for five out of the seven discordant patients (Figure 3-8B) [212].

We examined mutational data for the genes that are frequently altered in pancreatic cancer together with the clinical features of subtype-1 and subtype-2 tumours from all of the patients represented in the TCGA and ICGC datasets (Figure 3-8C). Here, we found no significant differences in the gene mutations between the tumour subtypes (see Table 3-1). Also, we observed that no genes were consistently altered in all of the tumours belonging to either of the subtypes. Similar to other studies, we discovered that some tumours lack mutations in any of the frequently mutated genes [213,214]. This diversity in the mutational landscape of pancreatic cancer tumours is likely to complicate the discovery of broadly applicable treatment regimens that target driver mutations [214].

Concerning histological features of tumours that might be useful for differentiating between the subtypes, we observed that only subtype-1 tumours displayed evidence of intraductal papillary mucinous neoplasm, whereas only subtype-2 tumours were categorised by histological inspection as being pancreatic adenocarcinomas (Figure 3-8C). Further, we found that subtype-1 tumours tended to be assigned a lower grade than subtype-2 tumours ($\chi^2 = 10.3$, $p < 0.01$).

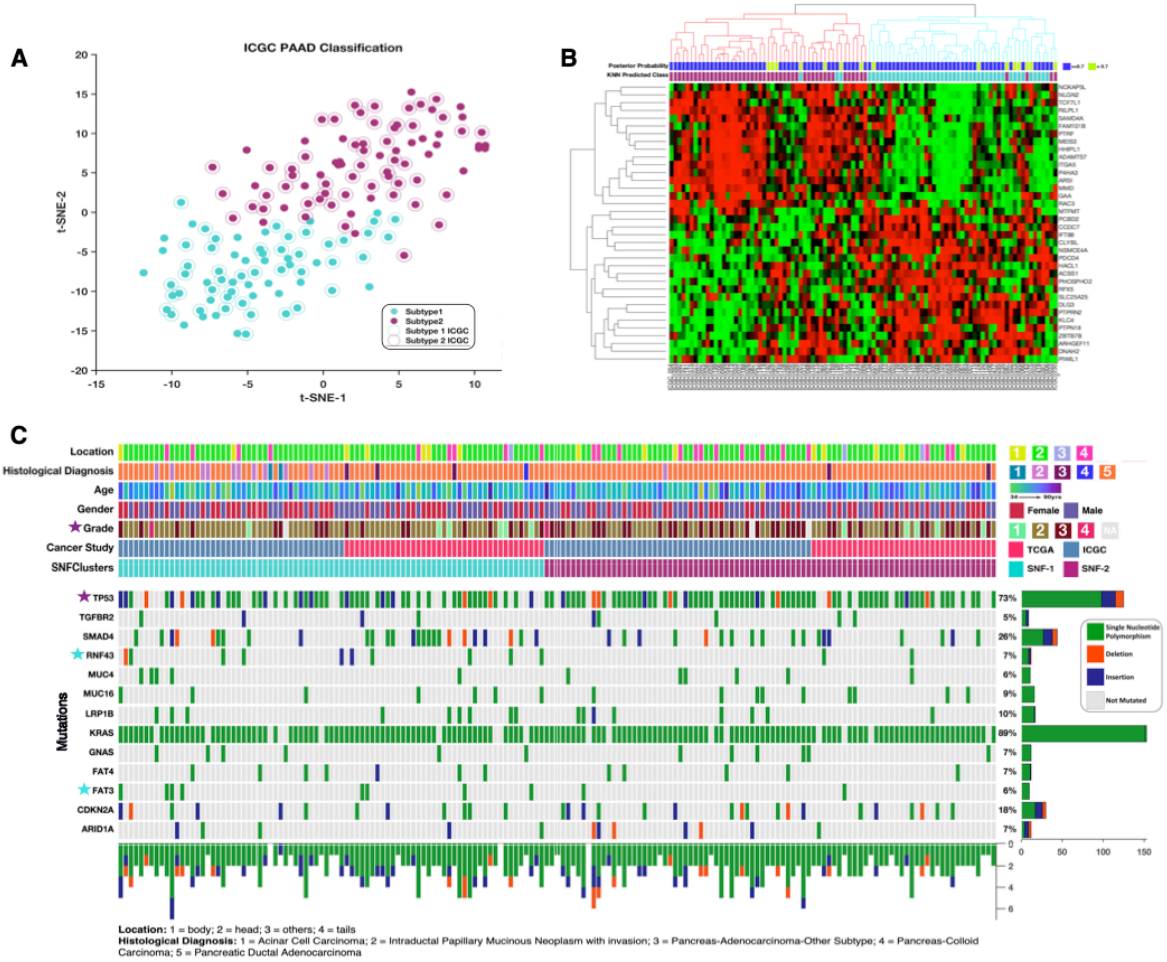


Figure 3-7: *Biomarker set validation and mutational landscape of pancreatic tumours:* ***(A)*** *Supervised classification of ICGC cancer patients using the mRNA-based KNN model trained on TCGA data. Circled points represent newly classified ICGC pancreatic cancer patients, whereas un-circled points represent the original 45 TCGA tumour samples that were used to train the model.* ***(B)*** *Unsupervised hierarchical clustering of the ICGC patients using the mRNA biomarker gene. The coloured bar on the clustergram shows the KNN model predicted class.* ***(C)*** *The integrated plot of clinical and molecular features for the TCGA and ICGC patient's data, ordered by their integrative (SNF) clustering. From top to bottom panels indicate primary tumour location; neoplasm histological type; patient gender; age at diagnosis; neoplasm histological grade; cancer study; integrative tumour subtypes; non-silent gene mutations. The key to the number coding of tumour location and histological diagnosis is at the bottom.*

Table 3-1: *Comparison of gene mutations between subtypes*

Gene	Chi Square	P-Value	Adjusted P-Value
ARID1A	0.333786103	0.563438152	0.843615793
CDKN2A	1.371503915	0.241553658	0.52336626
FAT3	3.8986332	0.048325408	0.209410101
FAT4	1.484768861	0.223029878	0.52336626
GNAS	0.00044137	0.983238629	0.983238629
KRAS	1.046475416	0.306320229	0.568880425
LRP1B	0.078952302	0.778722271	0.843615793
MUC16	0.25061223	0.616644317	0.843615793
MUC4	0.085064047	0.770548573	0.843615793
RNF43	5.837118963	0.0156915	0.101994749
SMAD4	3.219745336	0.072754948	0.236453581
TGFBR2	0.105266758	0.745598589	0.843615793
TP53	6.055228692	0.01386518	0.101994749

Chi-square test results for the frequently altered genes between the two pancreatic cancer subtypes.

3.2.5 Subtyping pancreatic cancer cell lines

We obtained mRNA expression and drug response data for 45 pancreatic cancer cell lines from the Cancer Cell Line Encyclopaedia (CCLE) [7]. We attempted to subtype these cell lines using the KNN classifier that we trained on the TCGA mRNA biomarker gene set identified using NCA (Figure 3-9A). It is known that cell lines with similar transcription profiles are likely to exhibit similar responses to drug perturbations [7,215,216]. It follows, therefore, that the drug response profiles of cell lines should be predictable based on their gene expression profiles [7,216,217].

3.2.6 Predicting drug responses using machine learning

Therefore, we predicted the anti-cancer drug responses of the cell lines from the drug response profiles of the cell lines that are most similar (i.e., the nearest neighbours) to each “query” cell line as determined using an exhaustive KNN searcher model (see methods section). The Searcher model quantified and stored information concerning similarities between the transcription profiles of all the cell lines. A “query” cell line’s nearest neighbours based on squared Euclidean distances were retrieved from the Searcher model. To infer the drug response of the query cell line, we calculated the median drug response of the retrieved nearest neighbour cell lines to each of the 24

anticancer drugs that were profiled by the CCLE (Figure 3-9B). For example, in Figure 3-9B, the cell lines SU8686 and PANC1005 both have available drug response profiles in the CCLE database, and both are the nearest neighbours of the cell line, PANC0203. Therefore, we used the mean drug responses of SU8686 and PANC1005 to predict the drug responses of PANC0203 (see methods section) (Figure 3-9C).

After predicting the drug responses of all the pancreatic cancer cell lines that also had measured drug response data in the CCLE database, we compared the predictions to the observed drug responses. Our drug response predictions displayed substantial agreement with the actual drug responses in that they yielded an average Kappa statistic of 0.67 (Figure 3-9D).

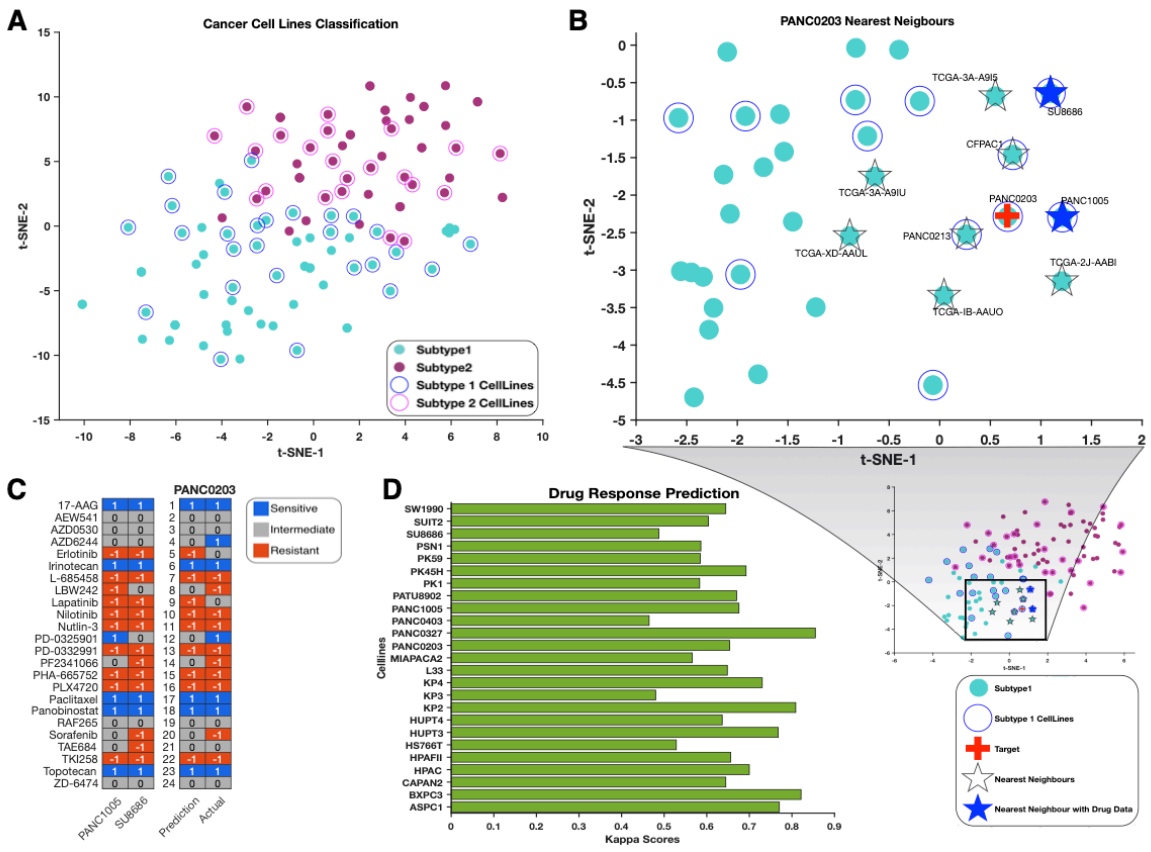


Figure 3-8: *Subtyping pancreatic cancer cell lines and predicting drug responses:* ***(A)*** *Supervised classification of CCLE pancreatic cancer cell lines using the mRNA-based KNN-model trained on TCGA data. t-SNE was used to visualise the tumour classes using the exact algorithm and squared Euclidean distance metric. Circled points represent classified CCLE cell lines, whereas un-circled points represent the TCGA samples used to train the models.* ***(B)*** *The t-SNE plot represents the KNN search for the nearest neighbours of PANC0203 in the exhaustive searcher model. Refer to the legend at the right*

bottom of the figure for interpretation. ***(C)*** *Drug response prediction: first two lanes represent the ranked drug responses to the 24 anticancer drugs of the PANC0203 nearest neighbours (PANC1005 and SU8686) for which such data is available. The last two lanes represent PANC0203's predicted drug responses and its actual drug responses.* ***(D)*** *Kappa scores of all CCLE pancreatic cancer cell lines with drug data. The kappa score was calculated using the quadratic method by comparing the actual and predicted drug responses of cell lines to the 24 CCLE anticancer.*

3.2.7 Validation of our machine learning drug prediction method using GDSC data

To validate the performance of the our Exhaustive search algorithm to predict of drug response to pancreatic cell lines profiles by other studies, mRNA expression data from 26 pancreatic cancer cell lines together with their response profiles to 309 small molecule inhibitors were downloaded from the GDSC database [5]. Then we used the 50-mRNA biomarker sets of both the GDSC cell lines and TCGA tumours to create an KNN exhaustive searcher model and inferred the response to each cell line to the small molecule inhibitors as previously described.

After predicting the drug responses of all the GDSC cancer cell lines that also had observed drug response data in the GDSC, we compared the predictions to the observed drug responses. Here, our drug response predictions displayed substantial agreement with the actual drug responses in that they yielded an average Kappa statistic of 0.65, a prediction accuracy that is comparable to the average Kappa statistics of 0.67 that we found for the prediction made using CCLE datasets (Figure 3-9).

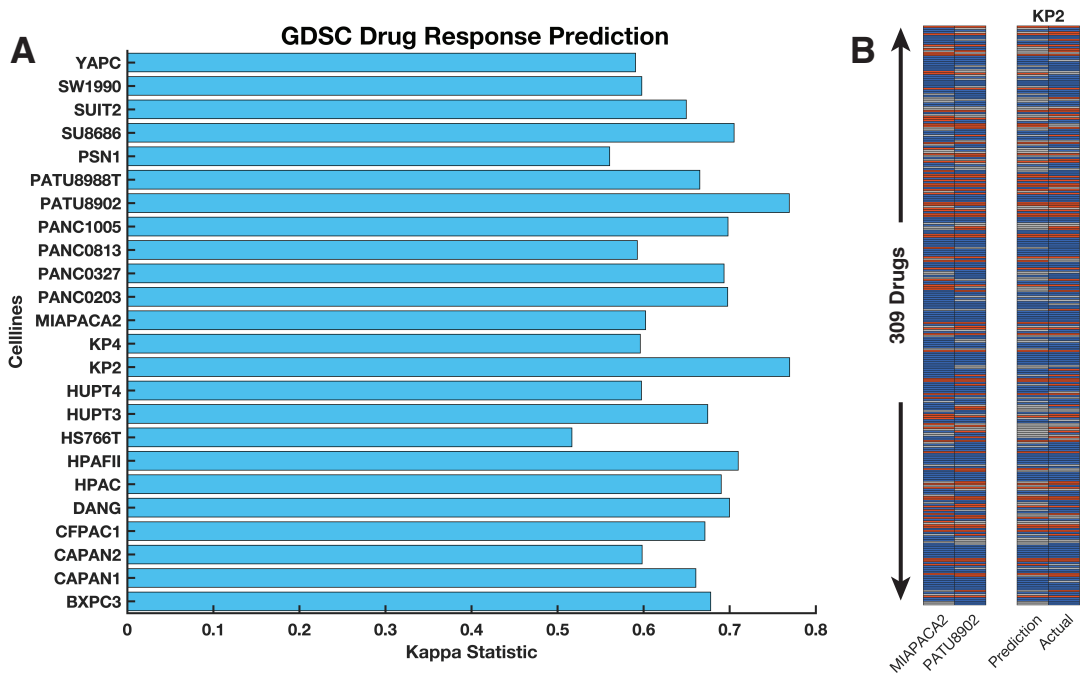


Figure 3-9: *Kappa scores of all GDSC pancreatic cancer cell lines with drug data.* ***(A)*** *The kappa score was calculated using the quadratic method by comparing the actual and predicted drug responses of cell lines to the 309 anticancer drugs.* ***(B)*** *Example drug response prediction plots for the cell lines KP2: first two lane(s) represent the ranked drug responses to the 309 anticancer drugs of the nearest neighbours (MIAPACA1 and PATU8902) for which drug response data are available. The last two lanes represent the predicted drug responses and the actual drug responses.*

3.3 Discussion

We conducted a comprehensive analysis of clinically relevant patterns of mutation, gene methylation, transcription, protein expression, and miRNA synthesis within pancreatic tumours. Several studies have previously highlighted the limitations of utilising a single molecular data type to accurately classify pancreatic cancers (Figure 3-1A) [185,186,188,218]. Here, we attempted to resolve this issue by employing a multidimensional clustering method capable of simultaneously utilising protein expression, mRNA transcription, DNA methylation and miRNA synthesis data. We found that by integrating across all these molecular data types, pancreatic cancer tumours could be classified into two clinically distinct subtypes: which we have simply named subtype-1 and subtype-2.

We observed that subtype-1 tumours were characterised by alterations of the mTOR signalling pathway, and the expression levels of different mTOR pathway proteins

were positively correlated to each other (Figure 3-4A, 3-4B and 3-4D). This finding is consistent with previous studies based on analyses of mRNA transcription and mutation data which also observed alterations of the mTOR pathway in pancreatic cancers [219–221]. Further, it is well established that some pancreatic cancer subtypes respond well to drugs which inhibit the mTOR pathway [222,223]. Accordingly, we anticipate that subtype-1 tumours will likely be more responsive to such therapies than will subtype-2 tumours.

Interestingly, subtype-2 tumours display unique alterations to cell cycle pathways (Figure 3-3A). This is consistent with the observation that subtype-2 tumours are clinically more aggressive than subtype-1 tumours in that an element of aggressiveness is the hyperactivation of the cell cycle processes that accelerate tumour growth [224–227].

We noted that, in addition to differences in patterns of protein expression, the two pancreatic cancer subtypes differ with respect to patterns of protein phosphorylation, implying that the kinases that are involved in oncogenic transformation differ between the subtypes. Specifically, whereas subtype-1 tumours show upregulation of mTOR signalling associated kinases (among others, MTOR-pS2448, GSKB-pS21-S9, and PDK-pS241), subtype-2 tumours display upregulation of cell cycle associated kinases (among others, CDK1-pY15, p27-pT158, and p27-pT198; Figure 3-4). Most of these kinases represent credible targets for small molecule inhibitors that might prove useful for subtype-specific anticancer therapies. Such small molecule kinase inhibitors are currently either being tested in clinical trials or are already in use as cancer therapies [227–230].

In addition to displaying alterations in the mTOR signalling pathway, subtype-1 tumours also display evidence of elevated ion channel (Figure 3-5A) and secretion pathway activities: a phenotype that is likely associated with increased trans-membrane transport of cell products (Figure 3-5B). Changes in the expression patterns of ion channel proteins are also found in breast and prostate cancers [231,232]. In pancreatic cancers, ion channel proteins likely play crucial roles in

cellular processes that are integral to oncogeneses such as cellular proliferation, motility, tissue invasion, and the excretion of lactic acid produced as a consequence of anaerobic respiration [233,234]. It is plausible therefore that subtype-1 tumours may be responsive to anti-cancer treatments that target ion channels and membrane pump proteins [233].

Subtype-2 tumours on the other hand display elevated peptidase activities (Figure 3-5A). Peptidases regulate various proteins that play essential roles in regulatory signalling networks. As is presently the case for tumours of the kidney, peptidases may be useful as diagnostic and/or prognostic biomarkers of subtype-2 pancreatic cancers [235,236].

We found no significant differences in the mutational landscape between the two pancreatic cancer subtypes, indicating that the accumulation of similar genetic mutations drive the formation of tumours belonging to both subtypes. Recently, the paradigm of oncogenesis has been expanded beyond the classical view that oncogenesis is entirely driven by the accumulation of genetic mutations [155,237]. This paradigm now includes the disruption of epigenetic regulatory mechanisms and variations in miRNA expression [237–242]. Unlike with mutations, we currently lack adequate conceptual knowledge and the analytical framework needed to identifying putative driver and passenger changes in epigenetic and miRNA based regulatory processes [74–76].

Nevertheless, we observed several differences between subtype-1 and subtype-2 tumours with respect to epigenetic (DNA methylation profile) and miRNA signatures. These suggest that epigenetic and/or miRNA variations may be primarily drivers of the differences in the transcriptome and proteome profiles of subtype-1 and substype-2 tumours.

In line with other studies that have identified biomarkers to classify tumour subtypes, some of which have important treatment and prognostic implications, we identified biomarker mRNA, DNA methylation, protein or miRNA sets that could be used to

accurately subtype pancreatic tumours [185,186,218,246]. We are optimistic that any of these four biomarker sets could be individually used to obtain accurate subtype classifications for new pancreatic tumours. Nevertheless, the utility of these four biomarkers sets for predicting clinical outcomes and guiding treatment strategies will need to be evaluated in future studies.

Encouragingly, we were able to demonstrate that, by focusing on just the transcription levels of the mRNA molecules that are represented in our mRNA biomarker set, we could accurately predict the drug responses of cell lines based on the drug responses of other cell lines with similar mRNA expression profiles. Although others have also been able to predict the drug responses of cell lines using similar machine learning approaches [7,187,216,217,247], our approach is novel in that it utilizes tumour subtyping based on all available molecular data to mine for biomarkers that differentiate disease subtypes: biomarkers which are then used to inform our KNN exhaustive search model with respect to quantifying the similarity of cell lines. What this means is that our approach is capable of utilizing matched molecular data and drug responses from either cancer patients or cell lines to predict, with reasonable accuracy, the drug responses of tumours for which we have only information on the concentrations of the mRNAs, proteins or miRNAs that are included within the biomarker sets which we have identified. As with other machine learning based inference schemes, the accuracy of the predictions that are made should improve given additional matched molecular and drug response data [248].

Altogether, our analyses have revealed the molecular underpinnings of, and potential treatment strategies for, two clinically distinct forms of pancreatic cancer. We are optimistic that an approach such as we have used, where multiple different molecular data types are leveraged to subtype and characterise particular tumour variants, could yield valuable insights into the management of other difficult to treat cancers such as those of the lungs and triple negative breast cancer.

3.4 Methods

We analysed data from 185 of the pancreatic cancer patients who had contributed samples to the TCGA project [96]. Data on these patient samples within the TCGA included: reverse phase protein array-based proteomics data (RPPA; n = 45), whole exome sequencing data (n = 76), transcriptome data determined using RNAseq (n = 76); DNA copy number and mutation data (n = 76), miRNA data (n= 56), and comprehensive clinical data. For our analyses, we only considered the 76 "high purity" samples for which transcriptome and whole exome sequencing data was available. Out of these 76 samples only 45 have RPPA data. All data used in our analyses were obtained from cBioPortal (http://www.cbioportal.org)[50].

3.4.1 RPPA-based classification of pancreatic cancer

K-means clustering of proteomic data was performed to identify subtypes of the 45 high purity TCGA pancreatic tumour datasets with available RPPA data [209]. To find the most informative number of clusters, K-means clustering was run over 500 iterations for cluster sizes (K values) of two, three, four, and five (i.e., K = 2 to 5). The average silhouette values for each value of K were compared, revealing that the two-cluster solution had the highest mean silhouette value and was therefore deemed to be the most coherent (Figure 3-9A). To aid in visualizing the most informative features that differentiated between the two inferred tumour subtypes, the 112 proteins with the highest entropy values across samples were used to reproduce the two-cluster K-mean classification using semi-supervised hierarchical clustering (Figure 3-1A) [249]. The clustering pattern thus obtained was visualised using a principal component analysis plot (Figure 3-9B) [250]. The clustering of these 45 pancreatic cancer tumours based on protein, miRNA and DNA methylation data has been previously published by Raphael *et al* [188], and the results of these clustering analyses were extracted from the supplementary file of that publication.

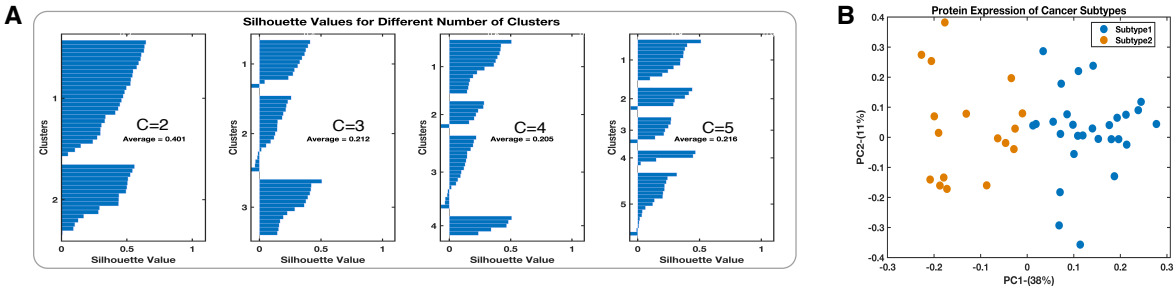


Figure 3-10: *K-mean clustering plots:* ***(A)*** *Choosing the number of clusters: plots of Silhouette values for each value of K (number of clusters).* ***(B)*** *Visualisation of the K-means clustering of the 45 high-purity pancreatic tumours. Principle component analysis was used to reduce the proteomic data dimensions, and the first two principal components plotted on the x-axis and y-axis, respectively.*

3.4.2 Integrative subtyping of pancreatic cancer

Similarity Network Fusion (SNF) is a clustering method that considers information from multiple molecular profiles. It has previously been used to segregate tumours of various cancer types based on multiple different sources of molecular data [192]. Briefly, standard normalised protein, mRNA, miRNA, and DNA methylation data derived from the 45 high-purity samples were used to create patient similarity networks (Figure 3-1B). Next, we ran SNF to fuse the similarity networks over 25 iterations, with hyperparameter settings of 24 and 0.7 for the number of neighbours and alpha value, respectively. Finally, spectral clustering with two specified as the best number of clusters (identified according to the eigengap) was applied to the unified similarity network to obtain the final tumour classification (Figure 3-1C) [192].

3.4.3 Patient's clinical characteristics of the pancreatic cancer subtypes

The Kaplan-Meier method was used to compare overall survival and the duration of progression-free survival of patients with tumours belonging to the different pancreatic cancer subtypes [160].

3.4.4 Pathways and kinase enrichment analyses

The differentially expressed proteins between the pancreatic cancer subtypes were identified using the Student *t*-test with unequal variance and with the Benjamin-Hochberg correction applied to p-values [251,252]. Further, we queried Enrichr with two lists of 60 and 30 proteins found to be upregulated in subtype-1 and subtype-2

tumours, respectively, to return enriched KEGG pathways for each subtype (see Supplementary File 1) [39,175]. The enriched KEGG pathways were compared to identify pathways that are unique to each of the disease subtypes [11]. The proteins that participate in pathways that are uniquely altered in sybtype-1 or subtype-2 tumours were used to construct protein-protein interaction networks using known interactions from each of the following databases: the University of California Santa Cruz Super pathway (101,525 protein-protein interactions), the Kinase Enrichment Analysis (428 kinases and their 10,792 targets), and Chromatin Immunoprecipitation Enrichment Analysis 2016 (667 transcription factors and their 464,967 targets) [169,170,253]. We visualised the resulting networks in yEd (Figure 3-3B and Figure 3-3C). Lastly, Kinase Enrichment Analysis was used to computationally identify the kinases that are responsible for the observed phosphorylation patterns in pancreatic cancer [170].

The moderated student *t*-test based on the negative binomial model was used to identify differentially expressed mRNAs and variations in DNA methylation patterns (see Supplementary File 2) [254,255]. Additionally, functional enrichment analyses were performed using lists of differentially expressed mRNA transcripts or altered DNA methylation patterns associated with each disease subtype. These were used to query Enrichr to return Gene Ontology (GO) molecular functions and KEGG pathways enriched for each disease subtype (Figure 3-1B, Figure 3-1C, and see Supplementary File 2). A custom MATLAB script was used to create an enrichment network based on the enriched GO-molecular function designations. This enrichment network was visualised in Cytoscape (Figure 3-5A) [256].

3.4.5 Identification and evaluation of biomarker sets

We used various data mining and machine learning methods to identify biomarker sets of mRNAs, DNA methylation, miRNAs or proteins that individually and consistently best stratified the two pancreatic cancer subtypes. The diagonal adaption of neighbourhood component analysis (NCA) with regularisation method was used to select the most useful features for each molecular data type (Figure 3-10A and 3-10B) [208]. Briefly, NCA attached feature weights to each attribute where the feature

weights are used to select the most important attributes for classification. For each molecular biomarker dataset identified using NCA, unsupervised hierarchical clustering was applied to the TCGA datasets to reproduce the two-subtype pancreatic cancer classification (Figure 3-6A, 3-6B, 3-6C and 3-6D). To apply supervised machine learning methods that accurately predict the tumour subtypes while utilising only one molecular data type, 23 different machine learning classifiers were trained ranging from linear discriminate analysis, support vector machines, decision trees, logistic regression, ensemble trees, and K-nearest neighbour algorithms. Then, the best performing classifier for each molecular biomarker dataset was selected based on their 5-fold cross-validation accuracy and area under the receiver operating characteristic curve (Table 3-2). The selected models were the cubic K-nearest neighbour for the mRNA biomarker set (98.7% accuracy), quadratic SVM for the DNA methylation biomarker set (97.8% accuracy), Ensemble bagged trees for the protein biomarker set (95.6%), and the course Gaussian SVM for the miRNA biomarker set (93.3% accuracy) [257].

Table 3-2: *Trained supervised learning models*

Genetic Data	Biomarker Size	Best Algorithm	Accuracy	AUC
mRNA	50	Cubic KNN	98.4%	0.99
Protein	14	Ensemble Bagged Trees	95.1%	0.97
Methylation	49	Quadratic SVM	97.6%	0.99
microRNA	20	Course Gaussian SVM	93.3%	0.95

Machine learning models that were trained for each biomarker genes, proteins and miRNA. All models were training using 5-fold cross-validation, and the best performing models were selected based on the classification accuracy, and the curve area under the curve.

To improve the accuracy of these models, the optimal hyperparameters that minimise the five-fold cross-validation loss were obtained using Bayesian hyperparameter optimisation (Figure 3-10C and 3-10D) [258–260]. This improved the overall classification accuracy of the models on the cross-validation set to 100% for the mRNA-based KNN model and 99% for the DNA methylation-based SVM model. The

trained models were then used to classify 31 other high-purity pancreatic tumours from the TCGA (Figure 3-6E and 3-6F). Supervised learning models based on the proteomic or miRNA biomarkers datasets were not trained because there were too few other high purity samples profiled by TCGA for these data types. Further, for each molecular data biomarker set, between three and ten features were selected based on the lowest cross-validation loss of the best performing algorithm (Figure 3-10E). These features were then used to classify TCGA pancreatic cancer samples using unsupervised hierarchical clustering (Figure 3-10F, 3-10G, 3-10H and 3-10I).

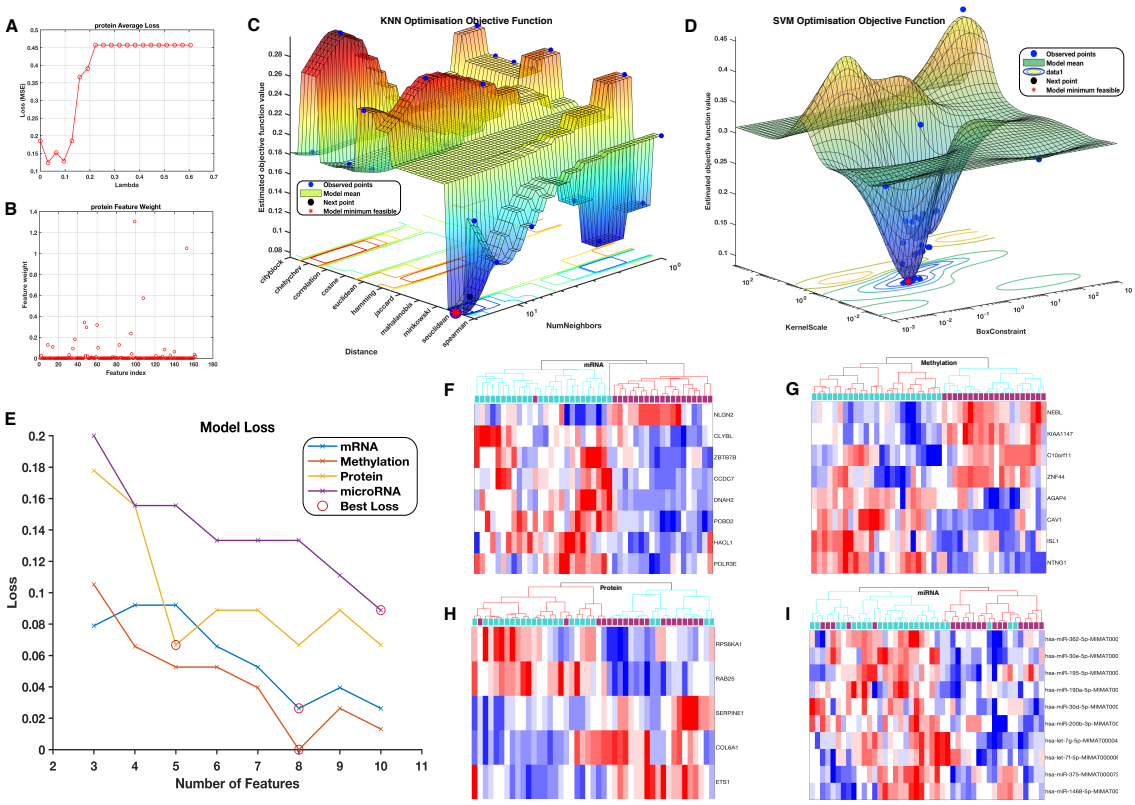


Figure 3-11: *Feature selection and machine learning classification of pancreatic cancer:* ***(A)*** *Identifying the best regularisation value (lambda) for NCA: plot of average loss values vs the lambda values of classification. In this example, the best lambda value that corresponds to the minimum average loss for the protein features was 0.1203****. (B)*** *A plot of learned feature (protein) weights using NCA of the proteins in the RPPA data that would be used to classify patients into the two integrative subtypes. The weights of the irrelevant features are close to zero.* ***(C)*** *Plot showing the selected optimal machine learning hyperparameters using Bayesian optimisation for the mRNA-based KNN model. On the x-axis are the number of neighbours, y-axis the distance metric and the z-axis average model loss. The optimal value is shown by the red star, i.e., one neighbour and the squared Euclidean distance.* ***(D)*** *Plot showing the selected optimal machine learning hyperparameters using Bayesian optimisation for the DNA*

methylation-based SVM model. On the x-axis is the box constraint, the y-axis is the SVM kernel scale and estimate average model loss on the z-axis. The optimal values are shown by the red star; box constraint of 0.46 and kernel scale of approximately 0.02. ***(E)*** *The number of features plotted against the average loss value of each supervised learning model for the Molecular data. The best in the top ten features that may be used to reproduce the integrative classification were selected based on the best loss (circled point).* ***(F)*** *Clustered heatmaps of the selected five mRNA transcripts, (G) four DNA methylation gene, (I) six proteins and (J) eight miRNAs: plots were produced using unsupervised hierarchical clustering with the correlation distance metric and the complete linkage.*

3.4.6 Validating biomarker molecular datasets

To evaluate the performance of the biomarker mRNA on a different pancreatic cancer dataset, we downloaded pancreatic cancer data from the ICGC data portal [9]. From the initial 50 mRNA biomarker set identified using the TCGA dataset, only 45 had corresponding data in the ICGC mRNA dataset. Therefore, we extracted the 45 gene biomarker set from both the TCGA and ICGC data. The mRNA-based KNN model was then re-trained on the TCGA 45 mRNA biomarker set. Here, standard normalisation was applied as a pre-processing step both to avoid platform associated biases, and because it was previously performed on the data before SNF clustering. Thereafter, the TCGA mRNA-based KNN model was used to predict the subtype of tumours in the ICGC dataset using a standard normalised mRNA biomarker set that we extracted from the ICGC RNAseq data (Figure 3-7A). Also, unsupervised hierarchical clustering was applied to the ICGC biomarker gene set (Figure 3-7B). Finally, the mutational landscape and clinical characteristics of the two pancreatic cancer subtypes of both the ICGC and TCGA datasets were compared (Figure 3-7C).

3.4.7 Subtype classification of cell lines

mRNA expression data from 45 pancreatic cancer cell lines together with their response profiles to 24 anticancer drugs were downloaded from the Cancer Cell Line Encyclopaedia [7]. The 50-mRNA biomarker set was extracted from the mRNA expression dataset and standard normalised. Then, the normalised CCLE mRNA biomarker genes were to subtype the cell lines by running the mRNA transcript levels for these genes through the mRNA-based KNN-model trained on TCGA data. The

predicted subtypes of the CCLE cell lines were visualised using t-distributed stochastic neighbour embedding (t-SNE; Figure 3-8A).

3.4.8 Machine learning method to predict a cell line's drug response

An exhaustive nearest neighbour searcher model was created using standard normalised mRNA biomarker sets of both the CCLE cell lines and TCGA tumours [261]. The exhaustive searcher model takes as input the training data (in this case the mRNA biomarker set), distance metrics, and parameter values of the distance metrics for an exhaustive nearest neighbour search and can then be used to identify the nearest neighbours to a particular patient tumour or cell line within a specified radius of the distance matric. Here, the nearest neighbours to a particular cell line suggest similarity at the molecular level based on mRNA, DNA methylation, protein and miRNA data encoded in the SNF subtyping. The ten nearest neighbouring cell lines or tumours were determined using a nearest neighbour search algorithm based on a Euclidean distance metric (see Figure 3-8B for intuition). After that, the drug response activity areas of the nearest neighbour cell lines were z-normalized and categorised as sensitive (for z-scored activity areas >0.8), intermediate (for z-scored activity areas between 0.8 and -0.8), or resistant (for z-scored activity areas <-0.8). A simple prediction model was employed where the median responses to a particular drug of the nearest neighbouring cell lines was used to infer a target cell line's drug response (Figure 3-8B and 3-8C). Following this the quadratic Cohen's Kappa score was used to evaluate the goodness of fit between the predicted and the actual drug response profiles of the cell lines (Figure 3-8D) [262].

3.4.9 Validation of our machine learning drug prediction method using GDSC data

mRNA expression data from 26 pancreatic cancer cell lines together with their response profiles to 309 small molecule inhibitors were downloaded from the GDSC database [5]. The 50-mRNA biomarker set was extracted from the mRNA expression dataset and standard normalised.

Next, described for predictions made using the CCLE dataset, an exhaustive nearest neighbour searcher model was created using standard normalised mRNA biomarker sets of both the GDSC cell lines and TCGA tumours [261]. The 8 nearest neighbouring cell lines or tumours were determined using a nearest neighbour search algorithm based on a squared Euclidean distance metric (refer to Figure 3-8A and 3-8B for intuition). Then we used the predicted the response of each GDSC cell line to each of the 309 small molecule inhibitors using the methods previously described in section 3.2.5.

3.4.9 Statistical analyses

All statistical analyses were performed in MATLAB 2018a except where stated otherwise. Fisher's exact tests were used to assess associations between categorical variables. Wilcoxon rank sum tests or independent sample Student *t*-tests were used for continuous variables where appropriate. Statistical tests were considered significant at $p < 0.05$ for single comparisons, and for Benjamini-Hochberg adjusted p-values < 0.05 for multiple comparisons.

3.5 Acknowledgements

Student bursary funding for this project was provided by H3ABioNet, supported by the National Institutes of Health Common Fund under grant number U24HG006941. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

3.6 Ethics approval

The University of Cape Town; Health Sciences Research Ethics Committee (HREC) IRB00001938 approved the protocol of this study.

3.7 Supplementary Information

Supplemental Information can be found online at https://zenodo.org/record/3253831#.XRCiFi2B2u4. The supplementary file descriptions are provided in appendix A.

Chapter 4 : Metabolic Genes Alterations Impacts the Clinical Aggressiveness and Drug Response of 32 Human Cancers

This section is a reformatting of a paper published in Communications Biology [263]: https://www.nature.com/articles/s42003-019-0666-1

Musalula Sinkala, Nicola Mulder, Darren Martin

4.1 Introduction

The transformation of normal cells into cancer cells requires the adaptation of multiple metabolic processes to satisfy the high energy demands of malignant cellular growth, proliferation and survival [264,265]. Accordingly, metabolic dysregulation is recognised as a hallmark of malignant cellular phenotypes [155,266]. Although many of the metabolic processes occurring in cancer cells are similar to those occurring in healthy proliferating cells, a series of genetic and epigenetic modifications in cancer cells can result in the aberrant regulation of these processes [267,268]. Among these genetic alterations are those occurring in a range of genes that are involved in metabolism. These alterations include diverse “driver” mutations and gene copy number alterations, which can impart a substantial degree of metabolic heterogeneity to different tumours of the same cancer type [269]. There is, therefore, keen interest in determining how genetic alterations within various types of malignant cells relate to specific aspects of the metabolic dysregulation occurring within these cells.

Transcriptomic and metabolomic analyses of various human tumours have revealed the numerous metabolic peculiarities of cancer cells that likely play essential roles in oncogenesis and cancer progression [269–274]. In general, these peculiarities can be traced to abnormal variations in the expression levels of either particular metabolic enzymes or the proteins that regulate these enzymes [114,275]. These and other studies [264,266,268,276–280] have also yielded a growing appreciation of how the aberrant metabolic changes in cancer cells influence the anticancer drug responses of different tumours.

Besides enabling the selection of the most appropriate available drugs, a better understanding of the metabolic differences between different cancer cell types will also likely yield better disease outcome predictions. This is because some of the metabolic features of cancer cells are likely to be directly associated with disease aggressiveness and clinical outcomes [281–283].

Until recently, a major impediment to linking specific metabolic dysregulations in cancer cells to particular disease outcomes or drug responses has been that the relevant metabolic pathways and their participating proteins were only partially known. The effective leveraging of detectable metabolic dysregulation to achieve either accurate prognosis or actionable treatments was further hampered by the unavailability of both large numbers of genomic/transcriptome datasets for the tumours of cancer patients with known clinical features and outcomes, and consistent data on the drug response profiles of large numbers of different human cancer cell types.

Today, however, comprehensive pathway curation projects (such as, for example, the Reactome and KEGG pathway projects) have successfully gathered high-quality information on human metabolic proteins and has accurately mapped these to metabolic pathways [10,11]. Cancer profiling projects such as that carried out by The Cancer Genome Atlas (TCGA) have yielded detailed genetic, transcriptomic, proteomic, and epigenetic data for thousands of human tumours each of which is annotated with clinical information for the patient from which it was taken [8]. Analysis of the TCGA data in the context of our present understanding of human metabolism should both illuminate the metabolic differences between different cancer types, and identify which of these differences has the most meaningful prognostic value. If this information is then coupled with the known drug responses of different cancer cell types, it should also be possible to identify the most suitable drugs to treat any particular cancer.

In this regard, large-scale drug response screening projects are extremely valuable. The immortalised human cell lines that have been widely applied as models of human disease can also be used for both drug discovery and the evaluation of drug dependencies [85,225,231]. For example, the Genomics of Drug Sensitivity in Cancer (GDSC) project, has provided genetic, transcriptome and epigenetic profiles for over one thousand human cancer cell lines together with their dose-response profiles to hundreds of anticancer drugs [5]. The genetic, transcriptomic and epigenetic profiles of tumour samples from the TCGA and those of cancer cell lines from the GDSC can be directly compared to systematically test for metabolic similarities and differences that might have a bearing on drug responses. More specifically, the subset of the cancer cell lines that have genomic and transcriptomic features that are most similar to those of tumour cells from a patient could be used to interrogate how metabolic perturbations in the patient's tumour cells are likely to influence the effectiveness of particular anticancer drugs.

Here, we used data on mutations and copy number variations from the TCGA in conjunction with Reactome Pathways data to identify the heterogeneous metabolic features of 32 human cancers. We then used these features together with drug response data from the GDSC to identify specific metabolic perturbations in tumour cells that are likely to impact their responses to different anticancer drugs.

4.2 Results

We analysed a TCGA dataset comprising lists of gene alterations (mutations and copy number variations) together with clinical information collected from 10,528 patients afflicted by 32 different human cancers (Figure 4-1A). Also, we analysed lists of gene alterations found within the genomes of 812 human cancer cell lines together with the drug-response profiles of these cell lines to 251 anticancer drugs to reveal associations between gene alterations and drug responses.

4.2.1 Alterations to genes involved in metabolism distinguish human cancers

We obtained curated human metabolic pathway data and the names of genes involved in these pathways from the Reactome Pathways database using the annotation search term "metabolism" (see https://reactome.org/PathwayBrowser/#/R-HSA-1430728) [10]. In this database, the term "metabolism" encompasses 68 different metabolic pathways involving 2,325 genes. Within the TCGA dataset, we found that out of these 2,325 genes, 2,095 contained an alteration in at least one of the 10,225 analysed patients.

Among the 2,095 metabolic genes displaying some alteration (a copy number variation or a mutation) in at least one patient, the most frequently altered were *PIK3CA* in 1,384 individual tumour samples, *APOB* in 976 and *LRP2* in 961 (Figure 4-1B). Most of the genes displaying some alterations in tumours of different cancer types have well-defined roles in carcinogenesis. For instance, mutations of *PIK3CA* reprogram metabolism and are associated with poorer survival outcomes in several cancers, including those of the colon, rectum, breast and lungs [284–287]. *APOB* is a lipid metabolism regulator that is linked to carcinogenesis and tumour progression in the liver, lungs and other tissues [288–290]. *LRP2* encodes a low-density lipoprotein receptor-related protein-2 which mediates endocytic uptake of various lipids, is linked to the enhanced metabolism of lipids and vitamin D, and promotes the transformation, proliferation and survival of various types of cancer cells [291–293].

Next, we calculated the frequency of alterations among the 16 first-tier metabolic pathways across all 32 of the cancer types. Here we found that genes involved in lipid metabolism were the most commonly altered, followed by those involved in carbohydrate metabolism and then those involved in amino acid metabolism (Figure 4-1C). These findings echo the well-established tenet of molecular oncogenesis, that meeting the cellular energy and biosynthetic demands of malignancy require alterations to the lipid, carbohydrate and amino acid metabolic pathways [266,280].

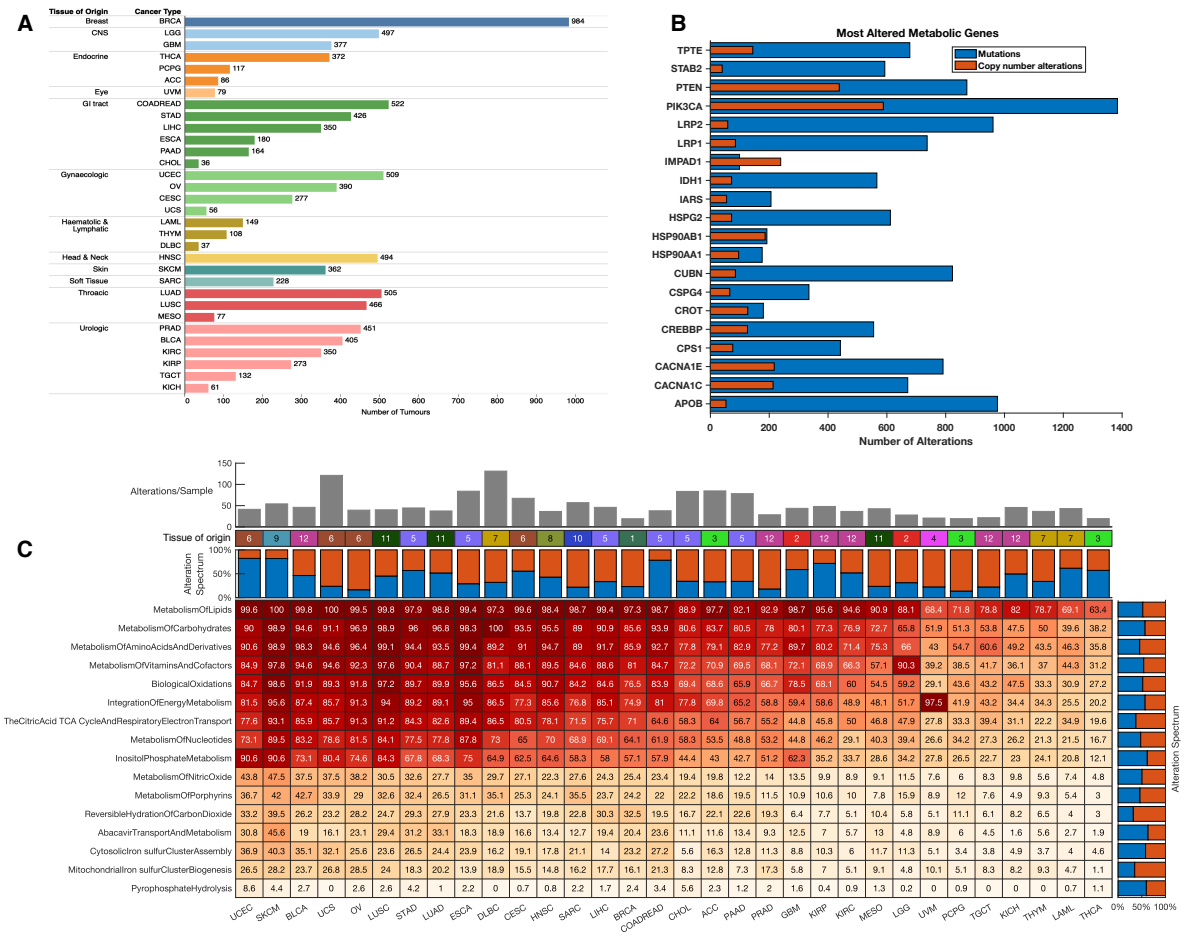


Figure 4-1: *Mutational landscape of TCGA tumours:* ***(A)*** *Distribution of 10,528 TCGA tumours across 32 human cancer types broken down by tissue of origin. TCGA disease codes and abbreviations: UCEC, uterine corpus endometrial carcinoma; SKCM, skin cutaneous melanoma; BLCA, bladder urothelial carcinoma; UCS, uterine carcinosarcoma; OV, ovarian serous cystadenocarcinoma; LUSC, lung squamous cell carcinoma; STAD, stomach adenocarcinoma; LUAD, lung adenocarcinoma; ESCA, oesophageal adenocarcinoma; DLBC, diffuse large b-cell lymphoma; CESC, cervical squamous cell carcinoma; HNSC, head and neck squamous cell carcinoma; SARC, sarcoma; LIHC, liver hepatocellular carcinoma; BRCA, breast invasive carcinoma; COADREAD, colorectal adenocarcinoma; CHOL, cholangiocarcinoma; ACC, adrenocortical carcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; GBM, glioblastoma multiforme; KIRP, kidney renal papillary cell carcinoma; KIRC, kidney renal clear cell carcinoma; MESO, mesothelioma; LGG, brain lower grade glioma; UVM, uveal melanoma; PCPG, pheochromocytoma and paraganglioma; TGCT, testicular germ cell tumours; KICH, kidney chromophobe; THYM, thymoma; LAML, acute myeloid leukaemia; THCA, thyroid carcinoma.* ***(B)*** *Genes involved in metabolism found to be most altered across all human cancers.* ***(C)*** *Clustered heatmap of cancer types using the percentage of tumours with first-tier metabolic pathway genes displaying alterations. Pathways are ordered by decreasing frequencies of alterations. Increasing colour intensities denote higher percentages. The heat map was produced using unsupervised hierarchical clustering with the Euclidean distance metric and complete linkage (see*

Figure 4-10). The bar graph represents the fraction of tumours with either mutations or copy number variations in each cancer type and metabolic pathway. The coloured bars on the heatmap show the tissue of origin for each cancer: 1 = Breast; 2 = CNS, 3 = Endocrine; 4 = Eye; 5 = GI tract; 6 = Gynecologic; 7 = Haematologic & Lymphatic; 8 = Head & Neck; 9 = Skin; 10 = Soft Tissue; 11 = Thoracic; 12 = Urologic. The bar graph represents the overall frequency of genomic alterations in each human cancer.

We clustered the 32 human cancers based on the frequencies of genomic alterations of metabolic pathways. Our clustering revealed two major groups of cancers (Figure 4-9 in the methods section): those cancers with a higher frequency of metabolic gene alterations (which we named as HM; n = 6,191) and those with a lower frequency of metabolic gene alterations (named as LM; n = 3,329). Interestingly, we observed that the landscape of metabolic gene alterations varied across the 32 cancer types. The median alteration frequencies for genes involved in each of the 16 first-tier metabolic pathways was higher in skin cutaneous melanoma (occurring in 90% of patients with this type of cancer) and lung squamous cell carcinoma (occurring in 84% of patients), whereas only 21% of patients with acute myeloid leukaemias and 14% of patients with thyroid carcinomas exhibited metabolic gene alterations (Figure 4-1C).

We examined whether the HM and LM cancer supertypes were associated with different clinical outcomes. Remarkedly, we observed that the median disease-free survival (DFS) periods was significantly lower ($p = 1.3 \times 10^{-7}$; log-rank test [160]) for the HM cancer patients (median = 58.3 months) than it was for the LM cancer patients (median = 116.2 months; Figure 4-2A). Similarly, the duration of overall survival (OS) periods for the HM cancer patients (OS = 68.9 months) were significantly shorter ($p = 6.8 \times 10^{-10}$) relative to those of the LM cancer patients (OS = 116.2 months; log-rank test; Figure 4-2B). We validated these findings with an independent dataset of patients afflicted with various cancers from the ICGC databases [9]. As with the patients recorded in TCGA the median OS period for patients recorded in the ICGC databases who had cancers belonging to the HM supertype (OS = 1,759 days) was significantly lower ($p = 6.3 \times 10^{-17}$) than that of patients with cancers belonging to the LM supertype (OS = 3,681 days; Figure 4-2C). Our results, therefore, demonstrate an association between the extent to which metabolic genes in cancer cells are altered (and therefore

probably the degree of metabolic dysregulation within these cells), and the aggressiveness of cancers.

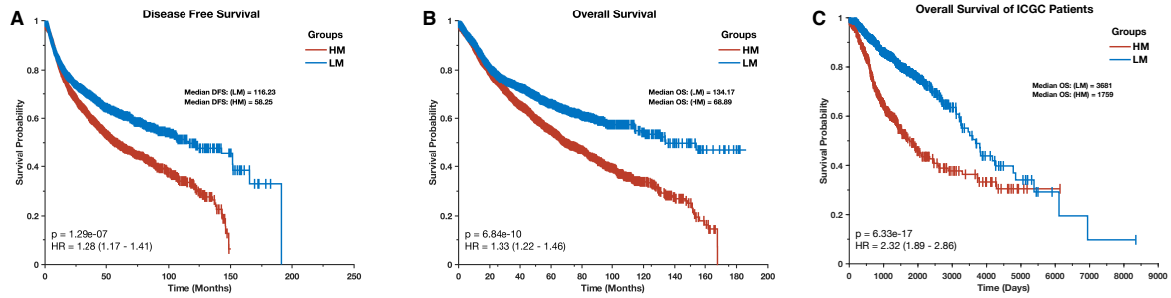


Figure 4-2: *Disease outcomes of the HM and LM tumours: Kaplan-Meier curve of the disease-free survival periods* ***(A)*** *and overall survival periods* ***(B)*** *of TCGA patients afflicted by the HM and LM cancer supertypes.* ***(C)*** *Kaplan-Meier curve of the overall survival periods of ICGC patients afflicted by the HM and LM cancer supertypes.*

There is noteworthy, however, that when we separately examined each of the 32 cancer types in isolation and compared the clinical outcomes of patients with tumours displaying higher and lower numbers of metabolic gene alterations, we detected no statistically significant difference in the duration of the DFS and OS periods for any of the 32 cancer types other than adrenocortical carcinomas (supplementary file 2).

4.2.2 Alterations of genes involved in carbohydrate, amino acid and lipid metabolic pathways across all cancers

We evaluated the extent of alterations to genes involved in second-tier lipid, carbohydrate, and amino acid metabolic pathways as these pathways had the highest gene alteration frequencies across all 32 of the cancer types (Figure 4-1C). We found that, as with the genes involved in the first-tier pathways, alterations to genes involved in second-tier pathways were more frequent in the HM cancers than in the LM cancers (Figure 4-3). Among the genes involved in second-tier carbohydrate metabolism pathways, those involved in the glycosaminoglycan metabolism (in 67% of all patients' tumours) and glucose metabolism (in 58% of tumours) pathways were the most commonly altered across all cancers. In recent years, cellular glycosaminoglycan profiles have been shown to be markedly altered during tumour pathogenesis and progression. Glycosaminoglycans influence cell signalling, angiogenesis, tumour

invasiveness, and metastasis, and have therefore emerged as essential pharmacological targets for the treatment of cancer [294–296].

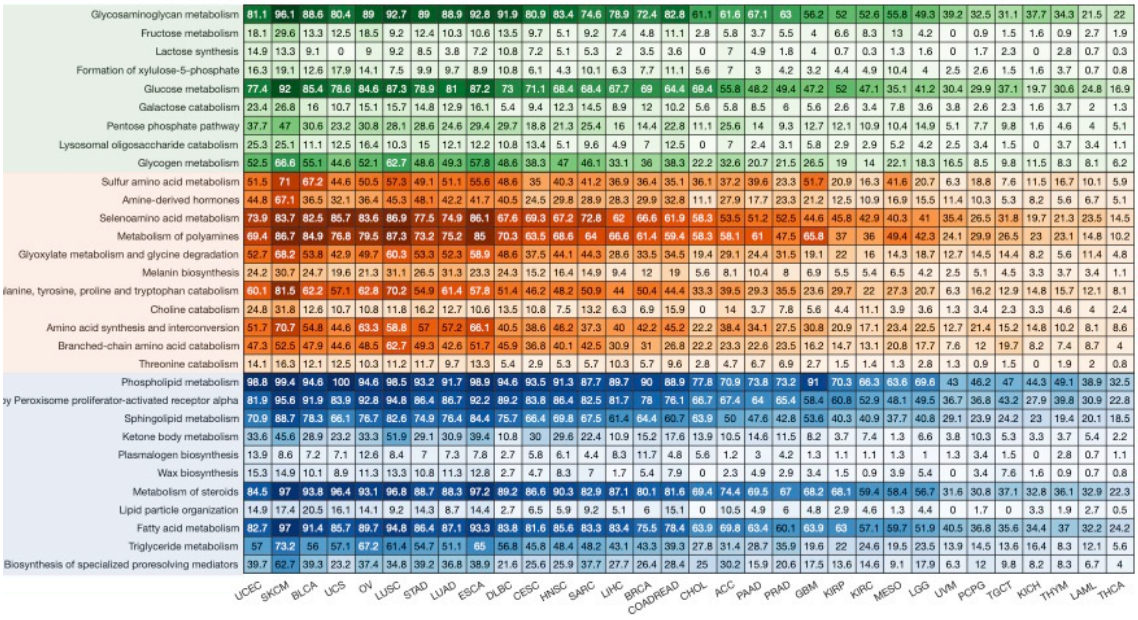


Figure 4-3: *Frequency of tumours of different cancer types with altered genes that are involved in second-tier metabolic pathways of carbohydrate, lipid and amino acid metabolism. The cancers are arranged according to how they clustered based on similarities between their first-tier metabolic pathway gene alterations (as in Figure 4-1C). Increasing colour intensities denote higher percentages of tumours with gene alterations).*

Among the genes involved in second-tier amino acid metabolism pathways, those involved in selenoamino acid metabolism (in 56% of all patients' tumours) and polyamine metabolism (in 56% of tumours) were the most altered across all the cancer types. Increased polyamine metabolism is associated with neoplasia: an important risk factor for the development of cancer in humans [297–301]. Drugs that target polyamine metabolism, several of which are in clinical trials, have been considered for the treatment of many cancers, including those of the colon, prostate, and skin [297,298,302]. Unlike with polyamines, the roles of selenoamino acids in cancer remain poorly explored; although an enrichment of selenoamino acids has been noted in breast cancer cells [303]. We anticipate that studying alterations of selenoamino acid metabolism could yield targets for the development of new

therapeutics and predictive biomarkers that would aid the treatment of various cancers.

Abnormal lipid metabolism has emerged as a metabolic hallmark of oncogenesis and tumour progression [304]. Here, we found that across all cancers, the most frequently altered of the lipid metabolism genes were those involved in the phospholipid metabolism (in 79% of all patients' tumours) and fatty acid metabolism (in 68%). Changes in the transcripts of genes that encode membrane phospholipids and the actual levels of phospholipids have been shown in various cancers, including those of the breast and lung [305–307]. Since the changes in phospholipid metabolism can affect the proliferation of cancer cells and their responses to drugs, it is plausible that at least some of the observed alterations in genes involved in phospholipid metabolism may have biological and clinical relevance [307,308].

Some of the most studied metabolic pathways in cancer are the glycolytic and fatty acid oxidation and biosynthesis pathways [264,276–279,309]. Here, we also explored the degree to which genes that are involved in these pathways were altered in each of the 32 cancers. In all cancers, we found alterations to some of the genes involved in the glycolytic and fatty acid oxidation and biosynthesis pathways (Figure 4-10, 4-10, 4-11, and 4-12 in the methods section). We found that these gene alterations were most frequent in uterine corpus endometrial carcinomas and skin cutaneous melanomas.

Finally, we used the literature to identify a subset of genes that encode proteins which are either key metabolic enzymes of the central metabolic pathways or are regulators of these enzymes. We discovered that 78% of all tumours harbour alteration in these genes (Figure 4-4). Among the most frequently altered metabolic regulators were *PTEN* (in 14% of all tumours), *KRAS* (in 11%) and *MYC* (in 11%). These gene alterations were most frequent in uterine carcinosarcoma (98.2% of patients' tumours) and least frequent in thyroid carcinomas (in 12.4% of tumours; Figure 4-4B).

Collectively these results reiterate that alterations within genes involved in particular aspects of lipid, carbohydrate and amino acid metabolism are found in many different cancers.

4.2.3 Alterations of genes involved in metabolism are associated with alterations of mRNA transcript levels

We next determined whether alterations in genes that are involved in metabolism are associated with alterations to the encoded mRNA transcript levels of these genes. We first examined whether the HM and LM cancer supertypes displayed distinct mRNA signatures for genes involved in metabolic pathways. Among the 2,325 genes involved in metabolism, we found that only 1,977 genes had transcript measurements in the TCGA. Therefore, focusing only on these 1,977 mRNA transcripts in all patients afflicted with the 32 different cancers, we applied t-distributed stochastic neighbourhood embedding (t-SNE) to reduce the dimensions of these data and visualised the relationships between cancers using scatter plots. We found that whereas the HM cancers displayed similar patterns of mRNA expression (i.e. they clustered closer to one another in the scatter plots; Figure 4-5A), the LM cancers tended to display more diverse mRNA expression patterns (i.e. they did not cluster as much in the scatter plots; Figure 4-5B). Specifically, whereas a three-dimensional t-SNE plot indicated that the HM cancers tended to group in the centre of the gene expression space, the LM cancers were scattered around the periphery of this space (Figure 4-5C).

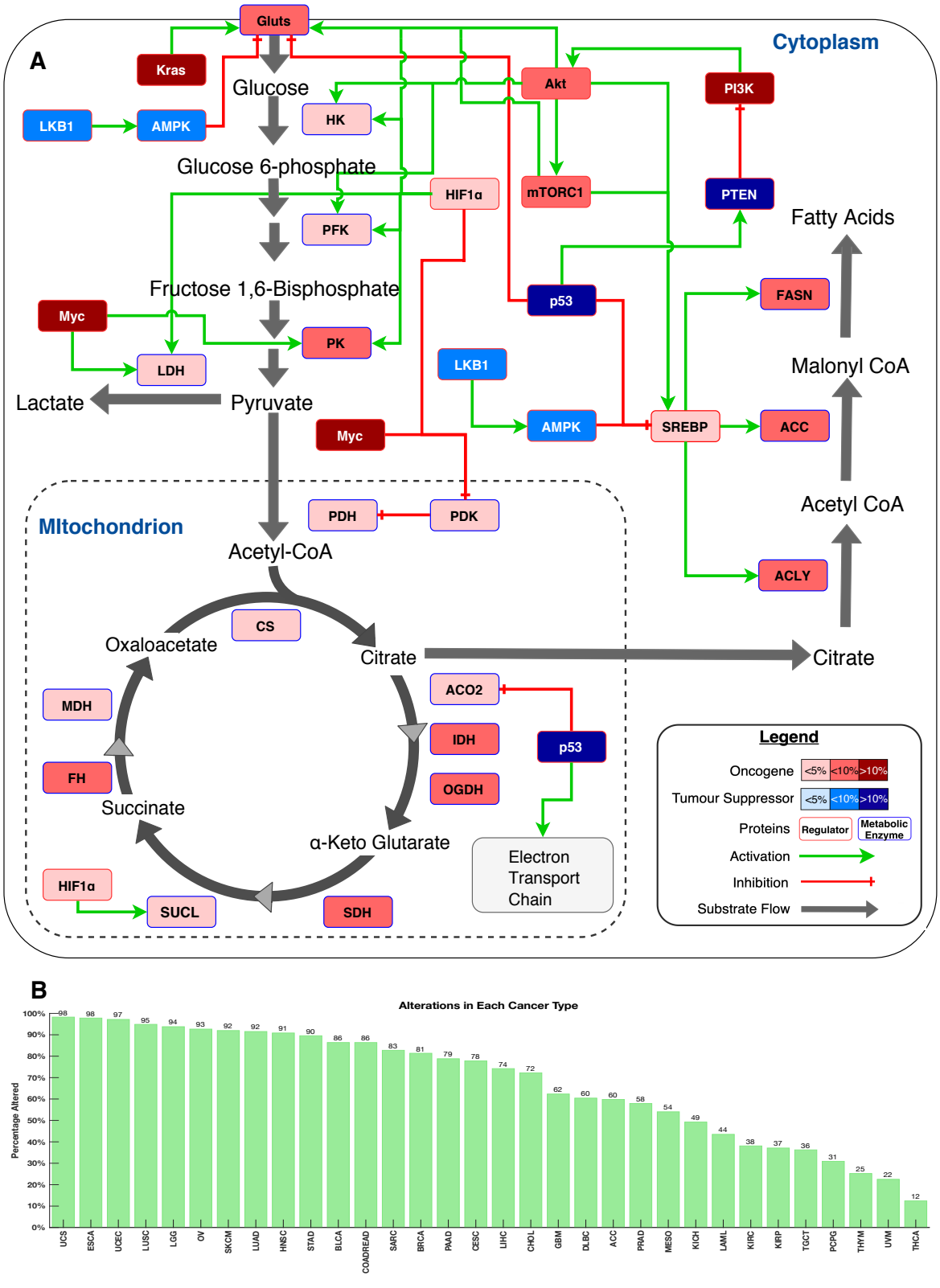


Figure 4-4: *Major catabolic and anabolic pathways of glucose and lipid metabolism in human cells. Nodes represent either enzymes (blue outline colour) or metabolic regulators (red outline colour). Node colours represent tumour suppressors (blue) and oncogenes (red) and their increasing colour intensities denote higher percentages of tumours with alterations in the genes encoding these enzymes or regulatory proteins. Edges indicate known types of interaction: red for inhibition and green arrows for*

activation. Abbreviations: GLUTs, all glucose transporters; HK, hexokinase; PFK, phosphofructokinase; PK, pyruvate kinase; LDH, lactate dehydrogenase; PDH, pyruvate dehydrogenase complex, PDK; pyruvate dehydrogenase kinase; CS, citrate synthase; ACO2, cis-aconitase; IDH, isocitrate dehydrogenase; OGDH, α-ketoglutarate; SDH, succinate dehydrogenase; SUCL, succinyl-CoA lyase; FH, fumarate hydratase; MDH, malate dehydrogenase; ACLY, ATP-dependent citrate lyase; ACC, acetyl-CoA carboxylase; FASN, fatty acid synthase; PTEN, phosphatase and tensin homolog; AMPK, 5'-AMP-activated protein kinase; mTORC1, mechanistic target of rapamycin complex-1; PI3K, phosphoinositide-3 kinase; SREBP, Sterol regulatory element-binding protein; Akt, RAC-alpha serine/threonine-protein kinase; Kras, Kirsten rat sarcoma viral oncogene homolog; Myc, MYC proto-oncogene; HIF1α, hypoxia-inducible factor 1-alpha; LKB1, Liver Kinase B1; p53, p53 tumour suppressor. ***(B)*** *overall fraction of samples with the central metabolic pathways gene alterations across 32 human cancers.*

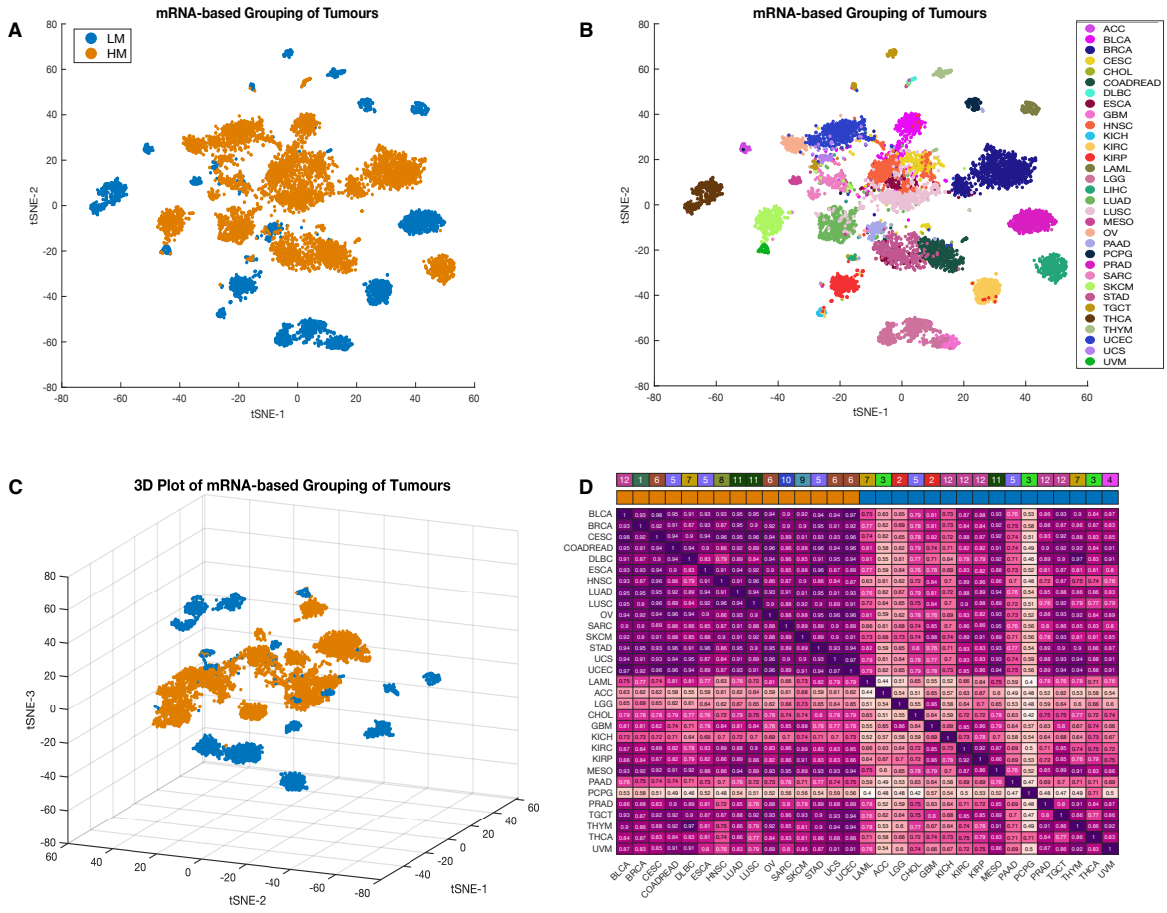


Figure 4-5: *TCGA tumour grouping based on metabolic gene transcript level****: (A)*** *Clustering of HM (orange points) and LM (blue points) tumours based on mRNA transcript levels.* ***(B)*** *Clustering of 32 different cancer types based on mRNA transcript levels. Points are coloured according to the type of cancer they represent. For both plots (A and B), t-SNE was used to visualise the tumour classes using*

the exact algorithm and standardised Euclidean distance metric. ***(C)*** *Three-dimensional plot of the HM/LM tumour supertype grouping based on mRNA transcript levels.* ***(D)*** *The integrated plot of mRNA expression correlations ordered by whether cancers belong to the HM or LM supertypes. From top to bottom, panels indicate: The tissue of origin; whether tumours belong to the HM or LM supertype; heatmap of inter-tumour linear Pearson's correlation scores with increasing colour intensities denoting higher degrees of correlation.*

Since the HM cancers tended to cluster together, we hypothesized that their metabolic gene expression profiles were highly correlated. To test this hypothesis, we measured the Pearson's linear correlation coefficients between transcript abundances across each pair of the 32 human cancers (see methods section). We establish that whereas the mRNA transcript levels of the 1,977 metabolic genes of each pair of HM cancers tended to be strongly positively correlated (mean Pearson's correlation = 0.9; range: 0.79 to 0.98), there tended to be weaker positive correlations between the mRNA transcript levels seen between the LM cancers (mean Pearson's correlation = 0.68; range: 0.40 to 0.92; Figure 4-5D).

Overall, these results indicate that while gene expression profiles are relatively conserved among the HM cancers, they are more diverse in the LM cancers.

Since the relative uniformity of the HM group was intriguing, we decided to further evaluated tumours in this supertype using data on all 20,502 of the mRNA transcripts that are available in the TCGA database (i.e., not only the transcript of metabolic genes). Here, we applied t-SNE to visualise the grouping of HM tumours (Figure 4-6A) and also applied Density-based spatial clustering of applications with noise (DBSCAN; [310]) approach to classify the tumour into various subgroups(Figure 4-6B). We found that patients afflicted with the different subgroups of tumours identified using DBSCAN exhibited different durations of DFS (Figure 4-6C) and OS (Figure 4-6D).

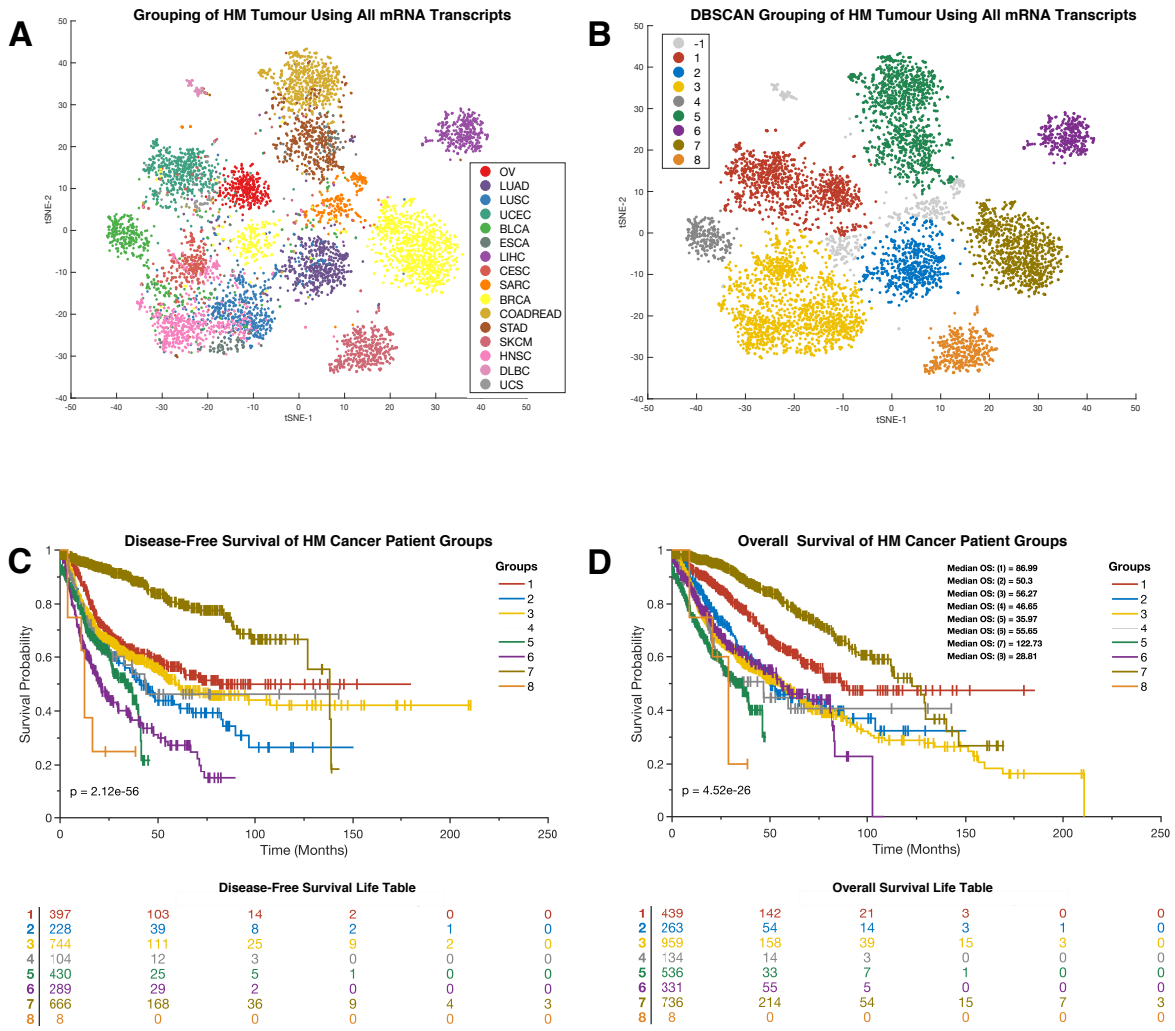


Figure 4-6: *DBSCAN tumour classification* ***(A)*** *Clustering of HM tumours based on all 20,502 mRNA transcript levels that were measured by the TCGA project. The colour legend represents different cancer types.* ***(B)*** *Clustering of HM tumours based on all 20,502 mRNA transcript levels that were measured by the TCGA. Points are coloured according to the clustering of the tumour using DBSCAN. -1 indicates the outlier points. For both plots (A and B), t-SNE was used to visualise the tumour classes using the exact algorithm and standardised Euclidean distance metric.* ***(C)*** *Kaplan-Meier curve of the disease-free survival periods and the life table of patients afflicted with each DBSCAN disease subtype.* ***(D)*** *Kaplan-Meier curve of the overall survival periods and life table of patients afflicted with each DBSCAN disease subtype. For both survival curve plots (C and D), the colours represent the tumour groupings yielded by DBSCAN in panel B.*

4.3.4 The drug responses of cancer cell lines are associated with metabolic gene alterations

From the GDSC database, we collected gene alteration data (including single nucleotide mutations, indels and copy number alterations) for 812 cancer cell lines of

30 different human cancer types. Each of these cell lines also has dose-response profiles to 251 anticancer drugs (Figure 4-6A) [5]. We assessed the patterns of metabolic gene alterations within these cancer cell lines and discovered that these were similar to those of the primary tumours (Figure 4-6B).

Given that previous studies have underlined differences in molecular characteristics between cancer cell lines and their primary tumour tissues [311,312], we directly compared metabolic gene alterations between cell lines and tumours of the same type. This revealed that, with only two exceptions, there were no significant differences in the frequencies of metabolic gene alterations between the cell lines and primary tumours of a given cancer type. The two exceptional cases were acute myeloid leukaemia (chi-square = 22.7, $p = 1.9 \times 10^{-6}$) and thyroid carcinoma (chi-square = 16.7, $p = 5 \times 10^{-4}$) for which the cell lines have significantly higher frequencies of metabolic gene alterations than did primary tumours (Figure 4-6B; supplementary file 1).

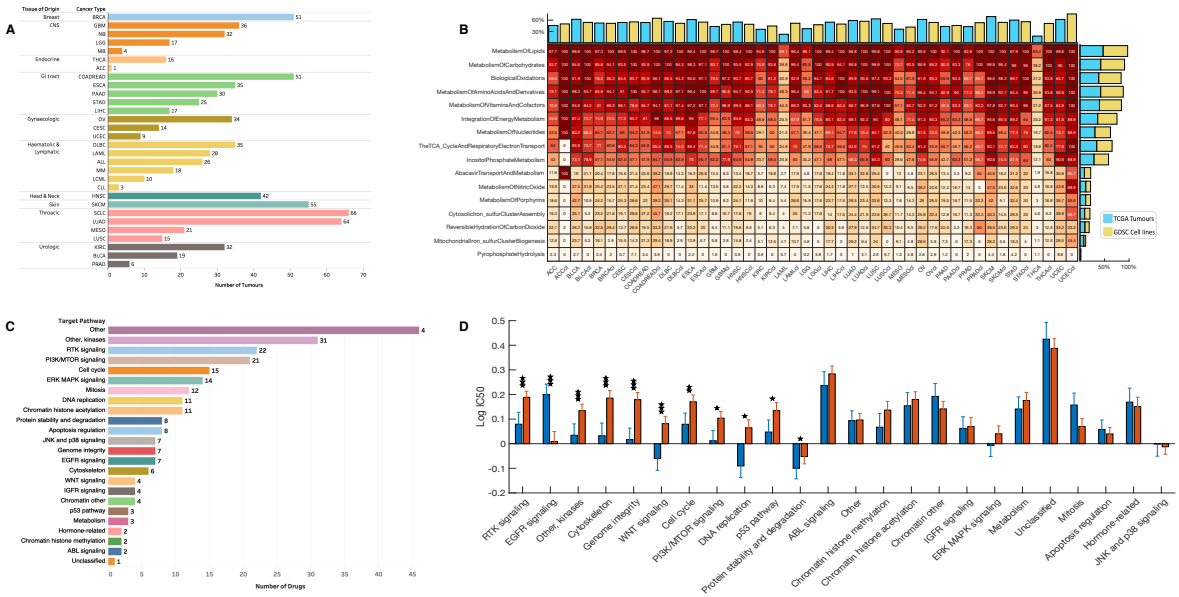


Figure 4-7: *GDSC cell lines and the mutational and dose-response characteristics:* ***(A)*** *Distribution of 1,001 cancer cell lines derived from 32 human cancer types broken down by tissue of origin.* ***(B)*** *Heatmap of the fraction of altered GDSC cancer cell line genes that are involved in each first-tier metabolic pathway in relation to corresponding patient tumour data from TCGA. Pathways are ordered according to numbers of observed alterations within genes that are involved in the pathways. Increasing colour intensities denote higher percentages of tumours containing alterations in the genes involved in the represented pathways. Bar graphs above the heatmap indicate overall percentages of gene*

alterations within GDSC cell lines (blue bars) or TCGA tumours (tan bars) of a particular cancer type. Bar graphs on the right of the heatmap indicate the overall percentage of alterations within each first-tier metabolic pathway for the GDSC cell lines (blue bars and TCGA tumours (tan bars). ***(C)*** *The number of anticancer drugs that target 24 signalling pathways and/or biological process that were used by the GDSC to treat cancer cell lines. Colours indicate the targeted pathways.* ***(D)*** *Comparison of the dose-response profiles between the LM and HM supertypes of the GDSC cancer cell lines. Bar graphs indicate logarithm transformed mean IC50 values of the cancer cell lines that correspond to the HM (orange bars) and LM (blue bar) cancer supertypes. The error bars indicate the standard error of the mean logarithm transformed IC50 value for each class of anticancer drug. The start indicates the levels of statistical significance; three stars for p values less than 0.001, two stars for p values less than 0.01 and one star for p values less than 0.05.*

Next, we classified the cancer cell lines into either the HM or LM supertypes using the TCGA cancer type labels of each cell line that are provided within the GDSC database. We then compared drug IC50 values between HM and LM cell lines for 24 classes of drugs that target 24 signalling pathways and/or biological processes (Figure 4-7C). Remarkedly, we uncovered differences between the HM and LM cancer cell lines in their observed dose-responses to various classes of anticancer drugs. Compared to the HM cell lines, the LM cell lines were significantly more sensitive to seven out of the 24 classes of anticancer drugs (Figure 4-7D; supplementary file 2). The drugs to which the LM cell lines were significantly more sensitive than the HM cell lines were those targeting pathways and biological processes such as receptor tyrosine kinase signalling, cytoskeleton structure, genome integrity processes, Wnt signalling, the cell cycle, PI3K/mTOR signalling, DNA replication, and kinases that are involved in multiple signalling pathways. Surprisingly, the HM cell lines were only significantly more sensitive than the LM cell lines to drugs that target the EGFR signalling pathway (Figure 4-7D; supplementary file 2).

We next compared the IC50 values of all 251 individual drugs with which the LM and HM cell lines were treated, regardless of the drugs' modes of action. Here, we found that, after correcting for multiple comparisons, the IC50 values of 41 anticancer drugs differed significantly between the LM and HM cell lines (supplementary file 2). Interestingly, the HM cell lines were more sensitive to only five of these 41 drugs. These included afatinib ($p = 2.1 \times 10\text{-}9$), CP724714 ($p = 5.5. \times 10\text{-}4$), gefitinib ($p = 6.3$

x 10-4), TAK-715 (p = 0.02), and vinorelbine (p = 0.049). Among these, afatinib, CP724714 and gefitinib target the EGFR signalling pathway, whereas TAK-715 targets JNK and p38 signalling, and vinorelbine inhibits mitosis by destabilising microtubules (Figure 4-7D and supplementary file 2). Conversely, we observed that the LM cell lines were significantly more sensitive than the HM cell lines to 36 of the anticancer drugs including CHIR-99021(p = 4.6 x 10-7), QL-XI-92 (p = 4.6 x 10-7) and SN-38 (p = 9.2 x 10-5; see supplementary file 2).

Overall this indicates that frequencies of metabolic gene alterations (our exclusive criterion for placing cell lines into the LM and HM supertypes) is a highly relevant variable when attempting to predict the drug responsiveness of cell lines and, therefore, that it may also be a clinically relevant variable when predicting the drug responsiveness of primary tumours.

4.3.5 The subtypes within each cancer exhibit diverse responses to anticancer drugs

For each of the 32 TCGA cancer types, we applied unsupervised hierarchical clustering to counts of alterations within genes involved in the 16 first-tier metabolic pathways to identify disease subtypes within each cancer type (see examples in Figure 4-15). Here, we found that (aside from genes involved in lipid, carbohydrate, and amino acid metabolism) subgroups of patients are likely to harbour additional alterations in other metabolic pathways. For example, in the case of glioblastoma multiforme (Figure 4-15A) and lung adenocarcinoma (Figure 4-15B), we found that while almost all the tumours represented in TCGA have alterations to genes involved in lipid, carbohydrate and amino acid metabolism pathways, small groups of tumours usually show even higher numbers of alterations to genes involved in, amongst others, the abacavir and nitric oxide metabolism pathways.

Since the frequencies of alterations to genes involved in metabolic pathways are likely to influence the responses of patients to anticancer drugs, we identified GDSC cancer cell lines displaying similar gene alterations to those found in individual primary tumours to test whether this might be the case (see methods section). Here, we

applied an approach were, for all cell lines of a particular human cancer, we compared their IC50 values for each of the 251 anticancer drugs between the cell lines with or without alterations to genes involved in each of the 16 first-tier metabolic pathways. Interestingly, we found that for cell lines of a particular cancer type, there are gene-alteration-dependent differences in their dose-responses to various anticancer drugs (supplementary file 2). For example, 51 anticancer drugs demonstrated significantly higher efficacies on oesophageal adenocarcinoma cell lines that have alterations in genes involved in abacavir metabolism pathways than on oesophageal adenocarcinoma cell lines without alterations to these genes (Figure 4-8A). Also, cell lines of lung adenocarcinoma with alterations in genes involved in the biological oxidation pathways are significantly more resistant to 52 anticancer drugs than are those without alterations to these genes (Figure 4-8A).

Altogether, we found 2,186 instances where alterations to genes involved in a specific metabolic pathway are associated with the efficacy in the cancer cell lines (supplementary file 2). Among the metabolic pathways, we found that those of cytoplasmic iron-sulphur clusters (447 instances), nucleotide metabolism (292 instances), and amino acid and derivatives metabolism (293 instances; Figure 4-8A) were associated with varied efficacies of the highest numbers of anticancer drugs for all the cancer cell lines across all tumour types.

Given that we had found that tumours displaying different numbers of alterations to metabolic genes exhibit different clinical and survival outcomes, we decided to examine this in more detail for particular cancer types. Using data of primary cancers from the TCGA, for patients' tumours with or without alterations in genes involved in abacavir metabolism, we found that the durations of the disease-free progression periods were significantly lower for oesophageal adenocarcinoma patients with alterations to these genes (log rank $p = 0.004$; Figure 4-8C). Conversely, disease-free progression periods were higher for uterine corpus endometrial carcinoma patients with alterations to genes involved in abacavir metabolism (log rank $p = 0.041$; Figure 4-8D). This then indicates that, even within each cancer type, the numbers of

alterations found in metabolic genes involved in particular pathways can, in addition to influencing anticancer drug responses, detectably impact patient survival.

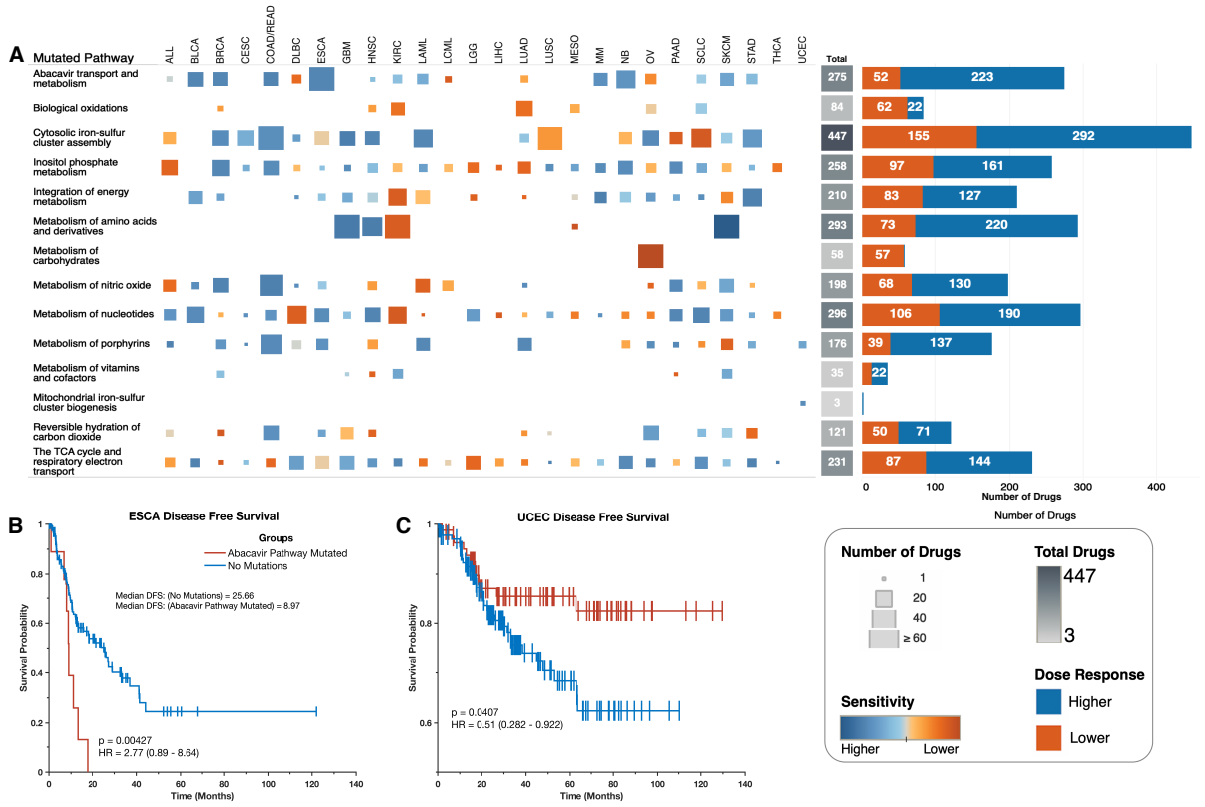


Figure 4-8: *Metabolic pathway gene alteration-dependent dose-responses and disease outcomes:* ***(A)*** *Dose-response profiles for drugs that have a degree of efficacy that is influenced by alterations in genes involved in specific metabolic pathways. From left to right: the columns represent GDSC cancer cell lines of various cancer types. The sizes of squares represent the number of drugs with efficacies that differ significantly between cell lines with and without gene alterations in the pathways indicated along the rows. The marks are coloured based on the overall influence of the metabolic gene alterations on drug efficacy: with increasing blue intensities denoting increasing sensitivity and increasing orange intensity denoting increasing resistance. The heatmap represents the overall numbers of drugs whose efficacy is influenced by the altered metabolic genes that are involved in the represented pathways. The bar graphs represent the total numbers of drugs whose dose-responses are increased (blue) or decreased (orange) by alterations of genes that are involved in the respective pathways.* ***(B)*** *Kaplan-Meier curve of the disease-free survival periods of patients afflicted with oesophageal adenocarcinoma with or without alterations to genes involved in the abacavir metabolism pathway.* ***(C)*** *Kaplan-Meier curve of the disease-free survival periods of patients afflicted with uterine corpus endometrial carcinoma, with or without alterations to genes involved in the abacavir metabolism pathway.*

4.3 Discussion

Metabolism is an essential process within all living cells. Thus, we examined the relationships between the numbers of alterations within the metabolic genes of primary tumours and cell lines of 32 different human cancer types and both clinical outcomes and likely drug responses. Others have used mRNA transcript data to show that alterations in metabolic pathways likely differ substantially between human cancer types [269,270]. To the best of our knowledge, ours is the first study to characterise metabolic gene alterations across such a large number of primary tumours (10,528) for so many distinct cancer types (32).

While we found at least one altered metabolic gene in every one of the 10,528 analysed tumours, the numbers of altered metabolic genes varied between the 32 cancer types that these tumours belonged to. We demonstrated that a clinically relevant clustering of patient tumours, irrespective of the type of cancer they represented, could be achieved by simply dividing the tumours into two supertypes based entirely on the numbers of alterations they displayed in metabolic genes: an LM supertype for low numbers of metabolic gene alterations and an HM supertype for high numbers of metabolic gene alterations (Figure 4-1B). Just as others have shown that alterations of genes involved in signalling pathways can have clinical implications [85,313], we show here that individuals with HM tumours tend to have significantly worse clinical outcomes than those afflicted with LM tumours. As such, our results suggest that simple counts of metabolic gene alterations in a tumour can provide a quantitative approximation of the extent of metabolic dysregulation within the tumour and, hence, an indirect approximation of the aggressiveness of the tumour.

Our analyses indicate that alterations of genes involved in the central metabolic pathways and the regulators of these pathways are pervasive across all human cancers (Figure 4-4). Among the most commonly altered of the regulatory genes that are involved in cellular metabolism were *PIK3CA* (in 32% of tumours), *MYC* (in 14%) and *HIF1A* (in 11%). In various cancers, *MYC* and *HIF1A* alterations dysregulate multiple metabolic enzymes including, hexokinase, isocitrate dehydrogenase, pyruvate dehydrogenase kinase and lactate dehydrogenase [314,315]. Further,

PIK3CA, *MYC*, *HIF1A* and other genes with frequent alterations in primary tumours are known to dysregulate cellular metabolism by increasing the rate of glycolysis while reducing the rate of aerobic respiration; a phenomenon referred to as the Warburg effect [265,314,316]. Tumours that exhibit a Warburg phenotype are known to be more aggressive and respond more poorly to most anticancer drugs [317]. Accordingly, compared to the LM cancers, we found higher alteration rates of the Warburg phenotype-associated genes in the HM cancers, which could explain why patients afflicted with HM cancers tend to have worse survival outcomes.

Changes in various signalling pathways are associated with variations in the response of cancer cells to drug perturbations, and these changes can, therefore, impact disease treatment outcomes [222,223]. Prior to the provision of anticancer drugs, it is desirable to know the drugs to which a particular tumour is most likely to be responsive. Since it is practically impossible to test hundreds of individual drugs on a specific tumour, cell lines that have phenotypic features resembling that of the tumour may be useful in predicting the drug responses of that tumour [5,7,73,227]. Accordingly, using drug response data for cancer cell lines, we inferred that HM and LM cancers are likely to respond differently to various anticancer drugs. Specifically, HM cancers tended to be less responsive to most anticancer drugs than LM cancers. This suggests that in addition to HM tumours potentially being more aggressive than LM tumours, patients afflicted with HM cancers may also exhibit worse clinical outcomes simply because HM cancers are more refractory to most anticancer drugs (supplementary file 2). Also, since our results indicate that HM tumours are likely to only respond to higher doses of anticancer drugs, it would follow that patients with such tumours would tend to experience more adverse drug effects and treatment-associated complications, both of which could unfavourably impact their survival [318–321].

Drugs such as afatinib and gefitinib, which target the EGFR signalling pathway were, however, found to have significantly higher efficacies in HM cell lines than in LM cell lines. Currently, afatinib is the first-line treatment for patients with metastatic non-small cell lung cancer, and it has also been evaluated for the treatment of head and neck

squamous cell carcinoma [322,323]. In our analyses, both non-small cell lung cancer and head and neck squamous cell carcinoma are HM cancers, and we predict, therefore, that there is a strong likelihood that many other HM cancers such as skin cutaneous melanoma, bladder urothelial carcinoma and lung adenocarcinoma may also respond to drugs that target the EGFR signalling pathway.

It is important to emphasise that our LM/HM classification is very simplistic. Taking a step back, we are reminded that among tumours that are derived from any particular tissue, there exist distinct tumour subtypes that differ from one another both in the gene alterations they display, and in the actual metabolic perturbations that these gene alterations cause [86,89,96,188,269]. In many respects, these distinct tumour subtypes are different diseases requiring different treatments [7,85,227].

We noted differences in the efficacy of various anticancer drugs between cell lines of the same primary cancer type. In some cases, these differences were associated with the presence or absence of alterations to genes involved in a particular metabolic pathway. This is in concordance with several recent studies that have established links between gene alterations and drug action [5,73,85,187,324,325]. This then supports the assertion that for any given cancer patient, the overall landscape of metabolic gene alterations could be used to identify generally applicable anticancer drug classes, following which alterations to specific metabolic genes could be used to eliminate the remaining drug choices that have the highest chances of failure.

Our results have shown that within each of the 32 cancer types, there exist subtypes that have alterations in genes that are involved in metabolic pathways that are less commonly associated with cancers (Figure 4-15). Interestingly, we found that for different cancer types, alterations of genes involved in a particular metabolic pathway may not produce similar clinical outcomes. For example, we found that for patients with alterations to genes involved in abacavir metabolism, those afflicted with oesophageal adenocarcinoma present with worse outcomes whereas those afflicted with uterine corpora endometrial carcinoma present with better outcomes (Figure 4-8B and Figure 4-8C). Such a scenario has been shown in other cancers. For instance,

activation of the mitogen-activated kinase pathway is associated with worse clinical outcomes in ovarian and colorectal cancer [326,327], but with better clinical outcomes in hormone receptor-negative breast cancer and astrocytoma [328,329].

Altogether, we have shown both that metabolic gene alterations which potentially dysregulate metabolic pathways are a pervasive phenomenon across all 32 of the investigated human cancer types, and that numbers of metabolic gene alterations are linked to treatment outcomes. Further, our analysis of the drug response profiles of well-characterised cancer cell lines suggests that alterations of genes of various metabolic pathways may also be predictive of drug responses. While we cannot guarantee that simply scoring gene alterations of particular metabolic pathways in patient tumours will reveal the best available treatment choices for these patients, it is apparent that such scores could nevertheless be leveraged to increase the probability of making a good treatment choice.

4.4 Methods

We analysed a TCGA project dataset of 10,528 patient-derived tumours representing 32 distinct human cancers (see Figure 4-1A) [96], obtained from cBioPortal [50] version 2.20 (http://www.cbioportal.org). The elements of the data that we used to identify gene alterations were gene copy number counts and somatic mutations (point mutations and small insertions/deletions). We also used mRNA expression data and comprehensive deidentified clinical data for all patients.

4.4.1 Metabolic gene alterations in the TCGA cancers

We accessed information of all human metabolic pathways from the Reactome pathways database version 68 [10]. Reactome pathways are arranged into several tiers with the Reactome term “metabolism” (Reactome ID: R-HSA-1430728), encompassing 68 different metabolic pathways (see https://reactome.org/PathwayBrowser/#/R-HSA-1430728). The first-tier pathways include sixteen curated metabolic pathways which involve 2,325 genes.

For each of the 32 human cancers, we calculated the overall percentage of samples with mutations and/or copy number alterations in these 2,325 genes. This provided us with alteration frequencies for each metabolic pathway in each human cancer (Figure 4-1C). We applied unsupervised hierarchical clustering with the squared Euclidean distance metric to these data to identify “altered metabolic gene” supertypes of human cancers (Figure 4-9). Based on the clustering dendrogram that that this yielded, we identified two cancer supertypes, which for simplicity, we named as either HM or LM, for those that respectively displayed higher or lower numbers of first-tier metabolic pathway associated gene alterations. The clustering of tumours into the two supertypes was highly coherent, with a cophenetic correlation coefficient of 0.89 and a Spearman's rank correlation between the dissimilarities and the cophenetic distances of 0.9 [310].

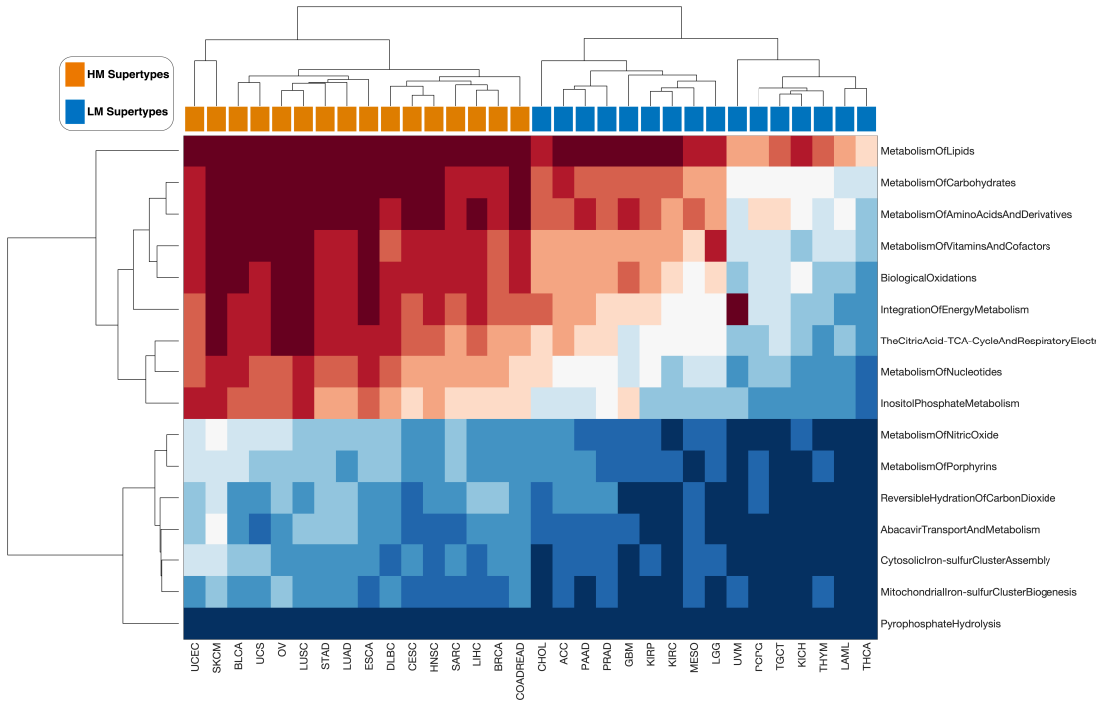


Figure 4-9: *Unsupervised hierarchical clustergram of tumours assigned to the two metabolic supertypes of human cancers. The fractions of tumours with altered genes that are involved in each of the 16 first-tier metabolic pathways were used for clustering. The clustergram was produced using the Spearman correlation distance metric with complete linkage.*

We extracted information relating to the genes that encode proteins of the second-tier metabolic pathways for only three of the first-tier pathways: those of carbohydrate, lipid and amino acid metabolism. Again, we used the approach for determining the extent of gene alterations (as described above) to calculate the fraction of tumours with alterations to genes involved in second-tier metabolic pathways across each cancer type (Figure 4-3). Also, using the same approach, we calculated the fraction of tumours with alterations in the genes that encode enzymes of the central metabolic pathway and their regulators (Figure 4-4). Further, we calculate the fraction of tumour with alterations in genes involved in the glycolytic, gluconeogenesis, fatty acid oxidation and biosynthesis pathways (Figure 4-10, 4-10, 4-11, and 4-12).

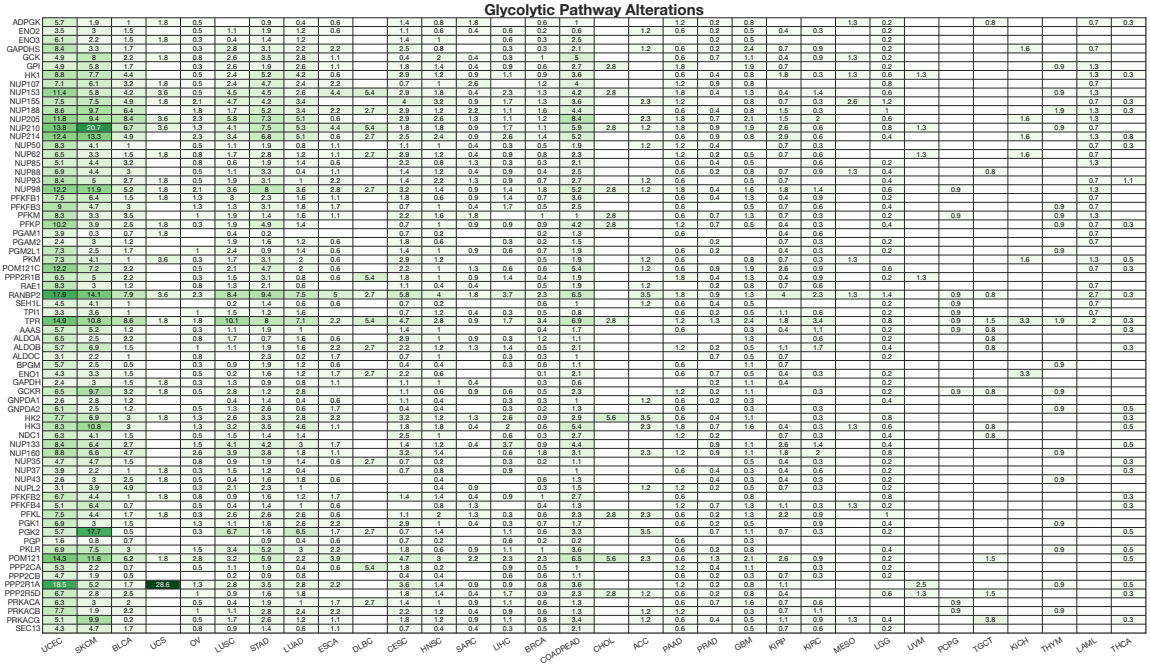


Figure 4-10: *Highlight table showing the fractions of tumours with alterations in glycolytic pathway genes across all human cancers. The increasing colour intensities denote higher percentages of altered genes.*

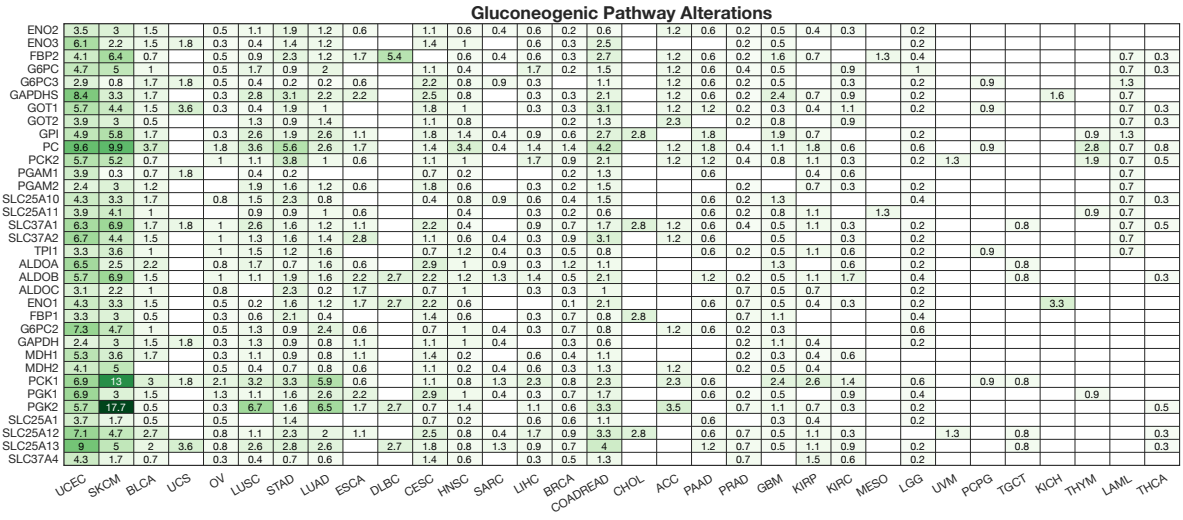


Figure 4-11: *Highlight table showing the fractions of tumours with alterations in gluconeogenic pathway genes across all human cancers. Increasing colour intensities denote higher percentages of altered genes.*

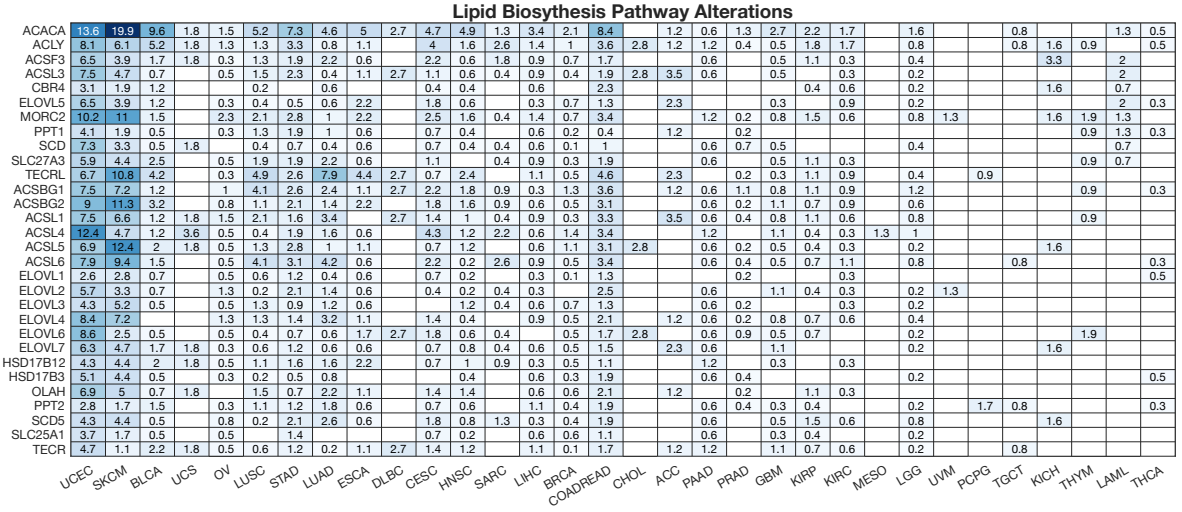


Figure 4-12: *Highlight table showing the fractions of tumours with alterations in mitochondrial fatty acid oxidation pathway genes across all human cancers. Increasing colour intensities denote higher percentages of altered genes.*

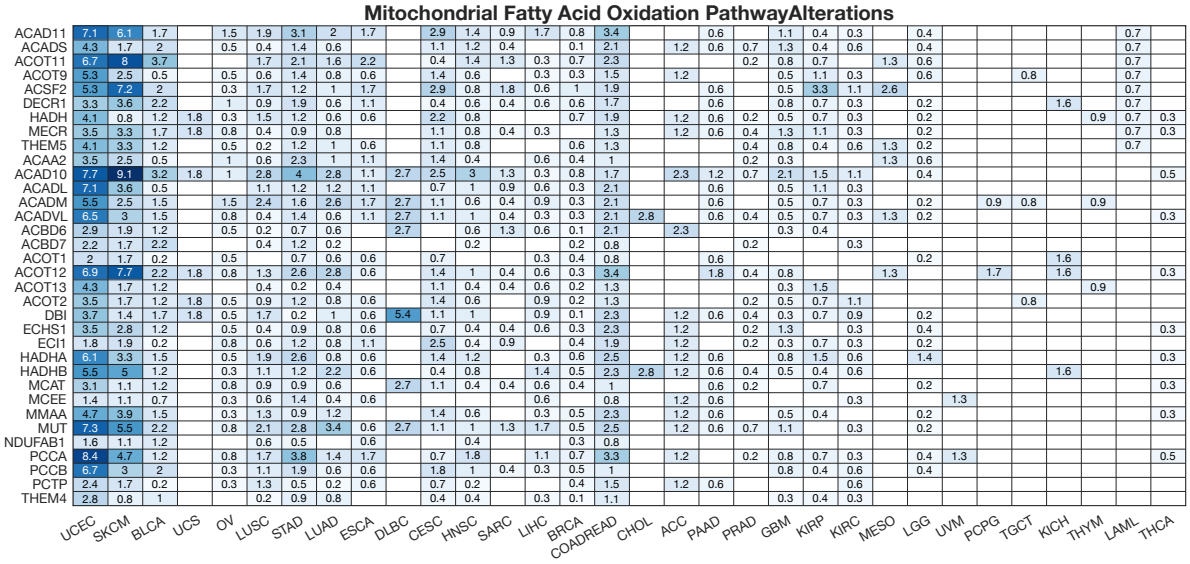


Figure 4-13: *Highlight table showing the fractions of tumours with alterations in lipid biosynthesis pathway genes across all human cancers. Increasing colour intensities denote higher percentages of altered genes.*

4.4.2 Analysis of mRNA expression profiles of metabolic pathway genes across cancers

We collected mRNA expression data of the genes that were profiled by the TCGA. Among this mRNA transcript data, we found information on only 1,977 genes out of the 2,325 genes that are involved in metabolism. We used the t-Distributed stochastic neighbourhood embedding algorithm to minimise the divergence between the 1,977-mRNA transcripts across cancers to return a two- and three-dimensional embedding of the 32 human cancers [330]. The overall structure of these transcript embeddings was visualised using scatter plots, first based on the cancers' metabolic supertypes (HM and LM) and second based on the cancer types (Figure 4-5A, Figure 4-5B, Figure 4-5C). To test for correlations between the mRNA transcripts of human cancers, we first calculated the mean transcript levels of the 1,977 metabolic pathway genes for each cancer and then used these mean values to calculate pairwise Pearson's linear correlation coefficients between each pair of the 32 human cancers (Figure 4-5D) [331].

We retrieved all 20,502 of the mRNA transcripts that were measured by the TCGA project across all cancer studies and applied t-SNE to visualise the clustering of HM tumours across a dimensional space (Figure 4-8A). Further, we applied DBSCAN [310] to cluster tumours belonging to the HM cancer supertype into various subgroups (Figure 4-8A).

4.4.3 Alterations of metabolic genes in cancer cell lines

We obtained mutation and copy number alteration data for 1,002 cancer cell lines and 224,202 dose-response profiles of these cell lines to 267 anticancer drugs from the Genomics of Drug Sensitivity in Cancer (GDSC) database version 7.0 (www.cancerRxgene.org) [5]. For downstream analyses, we focused on only the 812 cancer cell lines for which a complete set of gene alterations and drug response data was available.

Next, we calculated the frequencies of alterations in genes involved in the sixteen first-tier metabolic pathways in the cancer cell lines using the approach previously described for the 32 human cancers (Figure 4-6B, Figure 4-14). Finally, we used *chi*-square tests to identify possible differences between the TCGA cancers and the GDSC cell lines concerning the alteration counts of genes involved in the first-tier metabolic pathways (see results in supplementary file 1).

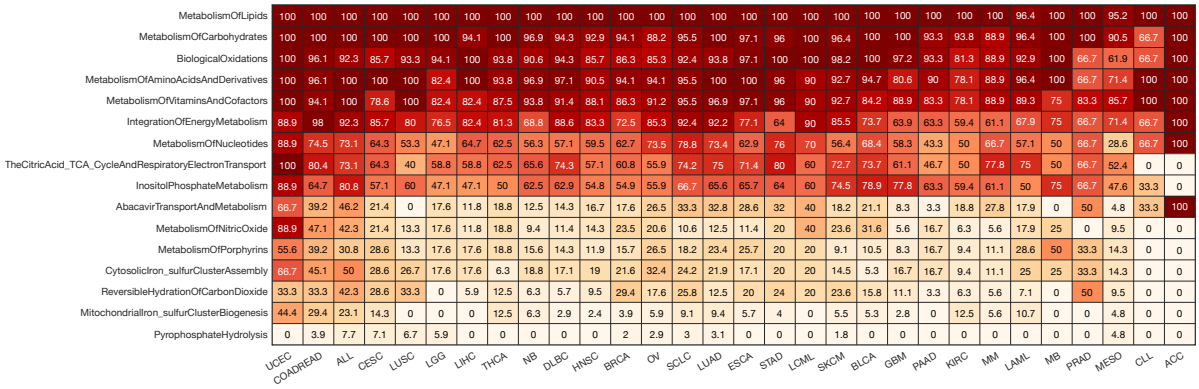


Figure 4-14: *Heatmap of GDSC cancer lines showing the percentages of cell lines with alterations to genes involved in each of the 16 first-tier metabolic pathways. Pathways are ordered by decreasing frequencies of gene alterations. Increasing colour intensities denote higher percentages of cell lines with gene alterations.*

4.4.4 Dose-response characteristics of the LM and HM cancer cell lines

From the list of 812 GDSC cancer cell lines, we returned only 653 cancer cell lines for which the GDSC have assigned a TCGA classification to the cell lines' primary cancer. Altogether, these 653 cell lines corresponded to only 23 of the 32 different human cancers profiled by the TCGA (Figure 4-6A). The GDSC treated these 653 cancer cell lines with 251 distinct anticancer drugs that target 24 different signalling pathways and biological processes (Figure 4-6C).

We used Student t-tests to compare the mean differences in the logarithm transformed IC50 values between the HM and LM cell lines for each class of anticancer drugs that we segregated based on the target signalling pathway and/or biological process (Figure 6D, also see supplementary file 2). Additionally, we compared the mean differences in the logarithm transformed IC50 values between HM and LM cell lines for each anticancer drug separately (supplementary file 2).

4.4.5 Identification of metabolic disease subtypes for each cancer type

For each of the 32 human cancer types, we calculated the frequency of alterations to genes involved in the 16 first-tier metabolic pathways. We then applied unsupervised hierarchical clustering to these data to identify subtypes of disease for each cancer (see examples in Figure 4-15).

4.4.6 Comparison of dose-response profiles within each cancer type for tumours with or without specific pathway alterations

For each particular human cancer, we collected all corresponding cell lines from the GDSC database. Then, for each of the 16 first-tier metabolic pathways, we segregated these cancer cell lines into two groups: those with and those without alterations in genes involved in a particular metabolic pathway. Finally, we compared the logarithm transformed IC50 values for each of the 251 anticancer drugs between the two groups of cell lines using the Wilcoxon rank sum test. Subsequently, we only returned drugs that had associated IC50 values which differed significantly between cell lines of human cancers with and without alterations of genes involved in a particular metabolic

pathway (Figure S8 and supplementary file 2). Note that these comparisons were only made in the cases were at least four cell lines had alterations of genes involved in a particular metabolic pathway and at least four other cell lines did not have such alterations.

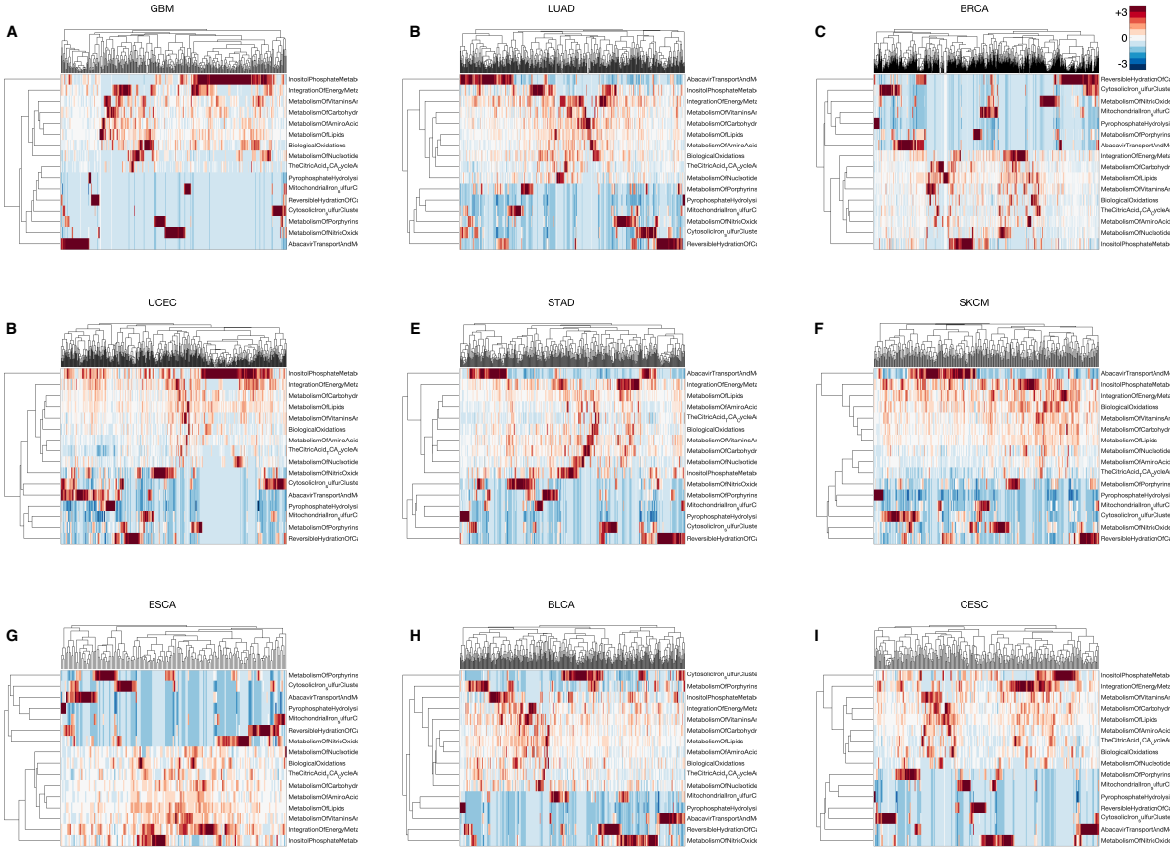


Figure 4-15: *Metabolic pathway gene alterations within cancer types. The clustergrams of gene alterations in various human cancer types. Only nine examples of cancer types are shown. The clustergrams show the percentage of tumours with alterations to genes involved in each of the 16 first-tier metabolic pathways.*

4.4.7 Survival analysis

The Kaplan-Meier method was used to estimate overall survival and the duration of progression-free survival between the HM and the LM supertypes of human cancer [160]. To validate our findings concerning the overall survival of the TCGA HM and LM supertypes, we downloaded an independent dataset of overall survival outcomes from the ICGC data portal [9] for individuals afflicted with tumours of types corresponding to those in the TCGA database. Since the ICGC data portal also contains some cancer

datasets from the TCGA, we removed these to return a dataset of 3,146 patient tumours that are unique to the ICGC. Next, we classified these the ICGC patient tumours into the HM or LM supertype categories based on the TCGA classification label provided within the ICGC database. We then compared the overall survival of these HM and LM patients. Also, the Kaplan-Meier method was applied to assess the survival outcomes of oesophageal adenocarcinoma and uterine corpus endometrial carcinoma patients who had tumours with or without alterations to genes involved in the abacavir metabolism pathway. The Kaplan-Meier method was also used to estimate OS and DFS between the subgroups of HM tumours that we identified using DBSCAN.

4.4.8 Statistical analyses

All statistical analyses were performed in MATLAB 2019a. Fisher's exact test was used to assess associations between categorical variables. The independent sample Student *t*-test or the Wilcoxon rank sum test and the one-way Analysis of Variance were used to compare continuous variables where appropriate. Statistical tests were considered significant at $p < 0.05$ for single comparisons, whereas the p-values of multiple comparisons were adjusted using the Benjamini-Hochberg method.

4.5 Ethics Approval

The University of Cape Town; Health Sciences Research Ethics Committee IRB00001938 approved the protocol of this study. This study involved the analysis of publicly available datasets that were collected by the TCGA, ICGC, GDSC and other databases from consenting participants. All methods were performed following the relevant policies, regulations and guidelines provided by the TCGA, ICGC, GDSC and other databases for analysing their datasets and reporting of the findings.

4.6 Supplementary Information

Supplemental Information can be found online at https://zenodo.org/record/3253843#.XRCj7C2B2u4. The supplementary file descriptions are provided in appendix A.

Chapter 5 : General Conclusion

5.1 General Discussion

I have highlighted how, in the post-genomics era, we can integrate big data from various resources to decipher the molecular characteristics of human diseases. Using large-scale datasets produced by high-throughput technologies such as next-generation nucleotide sequencing and protein mass spectrometry, my PhD project has exemplified how it is possible to move away from only studying the role(s) in a given disease of one or a few genes and/or proteins, to simultaneously studying the roles of, and relationships between, almost all the genes and proteins that might impact that disease [332,333]. In the current study, by applying an integrative "data-first" approach, I have shown that we can establish a multiscale appreciation of the cellular regulatory mechanisms that are at play in different forms of pancreatic cancer. For example, I showed that MAPK-PI3K-mTOR signalling pathways primarily drive the QM-PDAC disease subtype, whereas the p53 and cell cycle checkpoint pathways primarily drive the EL-PDAC subtypes (Figure 2-7 and 2-10). Also, I revealed that these subtypes of pancreatic cancer could likely be treated by targeting hub kinases within the dysregulated signalling pathways that are unique to each disease subtype (Figure 2-13).

To arrive at these findings, I integrated genome-wide expression data with prior biological knowledge extracted from the KEGG pathways, Reactome Pathway, ChEA, KEA, GO, and UCSC super-pathways databases. The use of such prior knowledge from multiple sources to construct metabolic and signal transduction networks of protein-protein interactions has drastically improved our predictive understanding of cellular systems [1,332,334,335]. However, just as prior knowledge has helped inform my network-based analyses, the remaining gaps in our current knowledge of the human interactome [332,333,336–340] have also constrained the precision of the inferences that I could draw from my analyses. To achieve maximum power with the types of analyses carried out in this thesis, there is, therefore, a pressing need to

improve the coverage of information on the various pathway databases to include all of the interactions between all human genes and all human proteins. Only with a complete comprehension of how cellular regulatory networks coordinate the underlying behaviours of complex biological systems can we fully comprehend how perturbations of these systems result in disease.

To put our current knowledge of cellular proteins into context; despite advances in bioinformatics and systems biology, we only understand the functions of ~65% of all the known human proteins. Further, only 3% of all the cellular proteins are presently targeted by drugs in clinical settings (Figure 6-1). We urgently need to identify the functions and interaction partners of the 35% of human proteins that we know virtually nothing about; only then can we develop more detailed network models that could then be used to more effectively mine transcriptomics, proteomics and other publicly available large-scale datasets for applicable disease biomarkers and novel therapeutic targets.

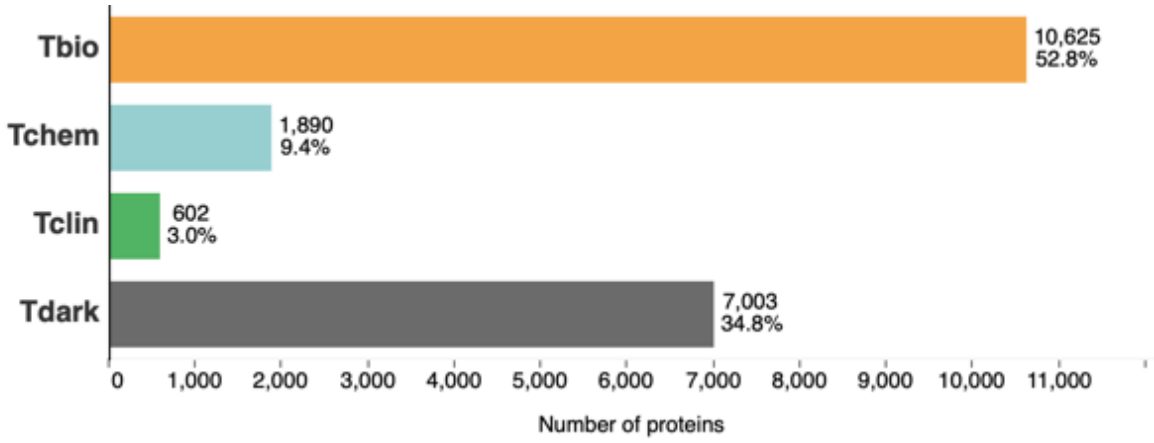


Figure 5-1: Our current knowledge of human proteins. *Tdark; these are human proteins about which virtually nothing is known. They do not have known interactions with either any known drugs or other small molecules. Tbio; these proteins also do not have interactions with either any known drugs or other small molecules, but are both annotated with a gene ontology molecular function or biological process term(s) derived from experimental evidence, and have known phenotype annotations. Tchem; these are proteins with known functions that interact strongly with at least one ChEMBL compound (with an activity cutoff of < 30 nM) that does not have an approved application in clinical settings. Tclin; these are proteins with known functions that interact with at least one approved drug (i.e. a ChEMBL compound with approved application(s) in a clinical setting). We retrieved information of these targets*

from Pharos (https://pharos.nih.gov/targets/)*; an integrated knowledgebase for illuminating the druggable genome* [129].

Despite these challenges, I have shown that numerous opportunities still exist to leverage multiple high-throughput data sources to define the clinical relevance of disease subtypes. For instance, I showed that the mutational profiles of most cancer cell lines are similar to those of the primary tumours from which the cell lines originated. Also, I showed that patients' tumours of various cancers could be compared to their corresponding cell lines to understand how genetic changes within cancer cells that impact metabolic pathways affect the responses of tumour cells to hundreds of anticancer drugs.

The backdrop to these results are studies which, to the contrary, have suggested that such comparisons are misleading because cancer cell lines are generally genetically not similar enough to the primary tumours from which they were derived [311,312]. Further, others have suggested that the comparison of genomic features and cell line drug response profiles from databases such as the GDSC may yield inconsistent associations [341]. Some of the observed inconsistencies are likely attributable to the diverse molecular profiling platforms and different computational analysis pipelines that have been employed in different studies [342–345]. Therefore, in this big-data era, it is essential to continually benchmark and standardise molecular datasets between and across experiments, so that various datasets obtained using different platforms are easily comparable [346–348]. In all such instances, knowledge of the underlying biology should guide the implementation of the data mining, machine learning and network modelling techniques that are applied to extract meaningful insights from such data. As such, our ability to integratively analyse large-scale datasets that have been obtained from multiple publicly available resources, and interpret the computational outputs of these analyses to identify the molecular predictors of drug dependencies, remains a pivotal challenge of precision cancer medicine.

In chapter three, I established that pancreatic cancer tumours clustered differently depending on the specific molecular data types that are examined (DNA methylation status, protein expression levels and mRNA/ miRNA transcription levels; Figure 3-1A). To mitigate this problem, I applied SNF, a multi-platform integrative clustering method that, in this case, took into consideration the full spectrum of available molecular data to define the two main subtypes of pancreatic cancer. Further, I integrated known protein-protein interactions with either mRNA transcription data or proteomics data to identify signalling pathways that are altered in these two pancreatic cancer subtypes. Here, I uncovered little overlap in the enriched signalling pathways identified when focusing either on upregulated mRNA transcripts or on upregulated proteins that I observed in only one of the two pancreatic cancer subtypes (Figure 3-5B and 3-3A).

Importantly, these results showed that besides issues related to the heterogeneous nature of the available molecular datasets, we are also still challenged by the fundamental complexity of biological regulatory networks. One of the consequences of this is the disparate clustering of the same sets of patients based on different sources of molecular data. This problem has been observed in studies of many other human cancers, including acute myeloid leukaemia, adrenocortical carcinoma, muscle-invasive bladder carcinoma, breast cancer, cervical cancer, cholangiocarcinoma, colorectal cancer, oesophageal carcinoma, glioblastoma, hepatocellular carcinoma, papillary renal cell carcinoma, malignant pleural mesothelioma, and prostate cancer [349–359].

Given the growing availability and size of big datasets, we advocate an approach where different types of molecular datasets that are all gathered for the same sets of patients are accurately integrated to identify unified disease subtypes. Ideally, we need to reach a consensus regarding the unified molecular disease subtypes of each human cancer; this would allow for the consistent identification of the causal molecular process disruptors that underlie each disease subtype. Once such subtypes are defined for a particular cancer, the clinical utility of such a classification scheme will depend on both the practical viability of assaying tumours for subtype-specific biomarkers, and the reliability with which these biomarkers can be used to assign

tumours to these subtypes. Altogether, my results demonstrate that, when one takes a multidimensional view of cellular regulatory systems using different molecular datasets, it is possible to identify small numbers of prospective biomarkers that can be used to reliably classify complex diseases such as cancer.

I have shown that alterations to metabolic genes which potentially dysregulate metabolic pathways are a pervasive phenomenon across 32 forms of human cancer. Interestingly, alterations to genes involved in specific metabolic pathways may be associated with contrasting disease outcomes. For example, using data from the TCGA for patients' tumours with or without alterations in genes involved in abacavir metabolism, we found that the duration of DFS periods was significantly lower for oesophageal adenocarcinoma patients with alterations to these genes (log-rank $p = 0.004$; Figure 4-8C). Conversely, DFS periods were higher for uterine corpus endometrial carcinoma patients with alterations to genes involved in abacavir metabolism (log-rank $p = 0.041$; Figure 4-8D).

Further, we found that alterations to genes involved in specific metabolic pathways are associated with the dose-response of cancer cell lines to anticancer drugs. Paradoxically, we find that in some cases, these associations differ between different cancers. For example, 126 anticancer drugs demonstrated significantly higher efficacies on glioblastoma multiforme cell lines that have alterations in genes involved in amino acid metabolism pathways than on glioblastoma multiforme cell lines without alterations to these genes (Figure 4-8A). Conversely, 67 anticancer drugs demonstrated significantly lower efficacies on kidney renal clear cell carcinoma cell lines that have alterations in genes involved in amino acid metabolism pathways than on kidney renal clear cell carcinoma cell lines without alterations to these genes (Figure 4-8A). These results illustrate the complexity of biological systems at the tissue level, i.e., genetic alterations have a tissue-specific context [360] which determines how they affect certain tumour characteristics: including the response of tumours to anticancer drugs and the aggressiveness of the disease.

Alterations to a particular gene may have different impacts on the physiology of different tissues because the gene may carry out different functions in different cell types and tissues [360–362]. For example, whereas in some cell lines, p53 levels oscillate with fixed periodicity in response to cellular perturbations, in other cell lines p53 levels change dynamically by either a single broad pulse or by continuous induction [363]. Likewise, DUSP4, which is a well-known negative regulator of ERK activity in various tissues, unexpectedly promotes ERK activity in various cancers, including that of the lung [364,365]. DUSP4 and other DUSP family proteins are thus considered a double-edged sword in cancer progression as they may act as oncogenes or tumour suppressor genes depending on the tissue or cancer type [364,366]. Likewise, we have recently showed that the mutations in MAPK pathway genes may be associated with mixed responses of the cell lines to the MAPK pathway inhibitors, i.e., some cells lines with these mutations tend to display more robust responses to MAPK inhibitors than others [367].

The fact that many other signalling pathways are likely to exhibit similarly variable biological impacts to those I observed during my analysis of metabolic pathways highlights the limitations of the current signalling pathway models and other prior-knowledge network models that are widely applied in systems biology research. These models are generally not cell-type or tissue-specific but are instead curated based on our understanding of whole-body cellular physiology. Making universal analytical assumptions about cell signalling based on these models may yield difficult to interpret results and/or biologically implausible conclusions [363]. Currently, this is a challenging unsolved problem that must be directly confronted when deciphering the causes and biological consequences of signalling pathway or network alterations. Altogether, I have concluded that there is a critical need to define regulatory network connections individually for all of the different cell types and tissues so as to enable the formulation of accurate cell-type and tissue-specific network models.

In this regard, the library of integrated cellular-based signatures (LINCS, http://www.lincsproject.org) has recently embarked on a multi-centre project to address some of the challenges of using the data integration techniques that I have

presented in this thesis to more fully understand complex cellular systems [3,29]. The overall goal of this project is to determine the molecular and functional changes that occur in thousands of different human cell types following drug exposure and/or genetic perturbations. For example, the Harvard Medical School LINCS Center is "addressing a range of technical and conceptual challenges related to collecting and analysing feature-rich data on cellular pathways and mechanisms of drug action" (https://lincs.hms.harvard.edu). The LINCS datasets offer fresh prospects to the broader scientific community to link biochemical and imaging data on cell signalling pathways (measured using protein levels, protein activities, receptor activation, kinase modification and transcription factor activation) to gene expression and other perturbations caused by the exposure of cells to small-molecules, drugs and cytokines. By integrating the LINCS datasets, and linking these data with other big data resources, we will be able to construct complex curated regulatory networks that are cell-type and context-specific. Besides providing completely novel biological insights, such a multiscale systems-level view of different cellular states in various contexts will both help us gain a more detailed understanding of known signalling pathways, and support efforts to develop therapies aimed at restoring perturbed pathways and networks to their normal states.

Altogether, in this thesis, I have focused on showing the practical utility of leveraging data-first approaches, machine learning, and network analyses to achieve novel, potentially actionable insights into the pathophysiology of cancers. Besides showing the ways in which big data resources can be used to accelerate the discovery of proteins within signalling pathways that could be targeted by novel anticancer drugs, I have also been particularly cognisant of showing how mining big-data resources can be used to predict the drugs that would be most effective for the treatment of specific categories of patient tumours. My results firmly support the view that the time is ripe for integrative analyses of big data in computational system biology to yield revolutionary new treatments for diseases like cancer.

5.2 Future Work

There is an urgent need to bridge the gap between advances in generating big biological datasets and our ability to integrate, analyse, and interpret these datasets. We could address many of these challenges by developing novel bioinformatics tools that would enable scientists that have little or no programming expertise to fully explore this data as effectively as our best-trained coder bioinformaticians. Soon, we aim to create a freely available web-based tool that will allow users to efficiently perform the various types of systems-levels analyses which I have showcased in this thesis. These will include: 1) performing functional enrichment analyses and visualising the results of these analyses. 2) the identification of regulatory kinases that are most likely the best drug targets within the signalling pathways/networks that are dysregulated in specific diseases or disease subtypes, 3) the definition of clinically relevant disease subtype classifications based on one or more different types of large-scale molecular dataset(s), and 4) the machine learning-based prediction of candidate drug compounds based on the similarity between primary patient tumours and cancer cell lines. We are optimistic that a user-friendly tool for experimental biologists will at least partially bridge the gap between our ability to generate large-scale datasets and extract biologically relevant insights from these.

Our current regulatory network models unrealistically imply that the activation (or inhibition) of a regulatory kinase will, across different cell types, propagate a signal through the same cellular circuitry that will lead to the consistent activation of downstream kinases and transcription factors. Further, these models also unrealistically imply that the activation of a specific transcription factor will lead to the same changes in gene expression regardless of the cells in which they occur. In reality, however, we know that the dynamics of complex regulatory networks vary substantially between tissues and cell types [360–366], and there is, therefore, a vital need for the development of a tissue-specific interactome knowledgebase. In the near future, I intend to focus my efforts on addressing this pressing need in bioinformatics and system biology. Specifically, using healthy tissue and cell line-specific transcriptome profiles from the Genotype-Tissue Expression and the Human Protein Atlas projects and datasets from various other resources (such as those from the LINCS and Achilles projects), I aim to redefine kinase and transcription factor

interactions in a tissue-specific context. Concretely, for a given tissue type, I will leverage information from these projects to remove connections between genes and/or proteins from the global human interactome that are unsupported (i.e. interactions which do not seem to occur) in that specific tissue type. This will allow me to extract cell-type or tissue-specific subnetworks from the entire human interactome that will better represent the biology of different tissues and/or cell types. These subnetworks could then be applied to perform better informed regulatory network analyses for higher precision mining of big datasets for drug targets.

References

[1] Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. N Biotechnol 2012;29:613–24. doi:10.1016/J.NBT.2012.03.004.

[2] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015;8:33. doi:10.1186/s12920-015-0108-y.

[3] Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst 2018;6:13–24. doi:10.1016/J.CELS.2017.11.001.

[4] Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, Scott JA, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A 2011;108:12372–7. doi:10.1073/pnas.1109363108.

[5] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 2012;41:D955–61. doi:10.1093/nar/gks1111.

[6] Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 2017;17:513. doi:10.1186/s12885-017-3500-5.

[7] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7. doi:10.1038/nature11003.

[8] Weinstein JN, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott

K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Publ Gr 2013;45. doi:10.1038/ng.2764.

[9] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026. doi:10.1093/database/bar026.

[10] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res 2016;44:D481–7. doi:10.1093/nar/gkv1351.

[11] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61. doi:10.1093/nar/gkw1092.

[12] Ioannidis JPA. Expectations, validity, and reality in omics. J Clin Epidemiol 2010;63:945–9. doi:10.1016/J.JCLINEPI.2010.04.002.

[13] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10. doi:10.1093/nar/30.1.207.

[14] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13. doi:10.1038/nrg1272.

[15] Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47–97. doi:10.1103/RevModPhys.74.47.

[16] Ferrell JE. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. Curr Opin Cell Biol 2002;14:140–8. doi:10.1016/S0955-0674(02)00314-9.

[17] Albert R. Scale-free networks in cell biology. J Cell Sci 2005;118:4947–57. doi:10.1242/JCS.02714.

[18] Barabási A-L. Scale-free networks: a decade and beyond. Science 2009;325:412–3. doi:10.1126/science.1173299.

[19] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–5. doi:10.1126/science.1073374.

[20] Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex

networks. Nature 2000;406:378–82. doi:10.1038/35019019.

[21] Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001;411:41–2. doi:10.1038/35075138.

[22] Brown N, Cambruzzi J, Cox PJ, Davies M, Dunbar J, Plumbley D, et al. Big Data in Drug Discovery. Prog Med Chem 2018;57:277–356. doi:10.1016/BS.PMCH.2017.12.003.

[23] Kedaigle A, Fraenkel E. Turning omics data into therapeutic insights. Curr Opin Pharmacol 2018;42:95–101. doi:10.1016/J.COPH.2018.08.006.

[24] Gagneur J, Friedel C, Heun V, Zimmer R, Rost B. Bioinformatics Advances Biology and Medicine by Turning Big Data Troves into Knowledge. 50 Jahre Univ. München, Berlin, Heidelberg: Springer Berlin Heidelberg; 2017, p. 33–45. doi:10.1007/978-3-662-54712-0_3.

[25] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2012;41:D991–5. doi:10.1093/nar/gks1193.

[26] Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 2016;44:D746–52. doi:10.1093/nar/gkv1045.

[27] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5. doi:10.1038/ng.2653.

[28] Aksoy BA, Dancík V, Smith K, Mazerik JN, Ji Z, Gross B, et al. CTD2 Dashboard: a searchable web interface to connect validated results from the Cancer Target Discovery and Development Network. Database (Oxford) 2017;2017. doi:10.1093/database/bax054.

[29] Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res 2018;46:D558–66. doi:10.1093/nar/gkx1063.

[30] Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The

BioGRID interaction database: 2019 update. Nucleic Acids Res 2019;47:D529–41. doi:10.1093/nar/gky1079.

[31] Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics 2017;18:142. doi:10.1186/s12859-017-1559-2.

[32] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44:D1075–9. doi:10.1093/nar/gkv1075.

[33] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. doi:10.1093/nar/gkx1037.

[34] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther 2012;92:414–7. doi:10.1038/clpt.2012.96.

[35] Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov — The Final Rule. N Engl J Med 2016;375:1998–2004. doi:10.1056/NEJMsr1611785.

[36] UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45:D158–69. doi:10.1093/nar/gkw1099.

[37] Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74. doi:10.1038/nature11247.

[38] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 2018;46:D794–801. doi:10.1093/nar/gkx1081.

[39] Gene Ontology Consortium: going forward. Nucleic Acids Res 2015;43:D1049–56. doi:10.1093/nar/gku1179.

[40] Brown SDM, Moore MW. The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm Genome 2012;23:632–40. doi:10.1007/s00335-012-9427-x.

[41] Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res 2017;45:D865–76. doi:10.1093/nar/gkw1039.

[42] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging

knowledge across phenotype–gene relationships. Nucleic Acids Res 2019;47:D1038–43. doi:10.1093/nar/gky1151.

[43] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015;43:D789–98. doi:10.1093/nar/gku1205.

[44] Schmidt T, Samaras P, Frejno M, Gessulat S, Barnert M, Kienegger H, et al. ProteomicsDB. Nucleic Acids Res 2018;46:D1271–81. doi:10.1093/nar/gkx1029.

[45] Shen EH, Overly CC, Jones AR. The Allen Human Brain Atlas: Comprehensive gene expression mapping of the human brain. Trends Neurosci 2012;35:711–4. doi:10.1016/J.TINS.2012.09.005.

[46] Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Res 2018;46:D1068–73. doi:10.1093/nar/gkx1143.

[47] Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–11. doi:10.1093/nar/gku1075.

[48] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 2011;39:D945–50. doi:10.1093/nar/gkq929.

[49] Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci Data 2014;1:140035. doi:10.1038/sdata.2014.35.

[50] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4. doi:10.1158/2159-8290.CD-12-0095.

[51] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res 2009;37:D767–72. doi:10.1093/nar/gkn892.

[52] Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science 2017;356:eaal3321. doi:10.1126/science.aal3321.

[53] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80-) 2015;347:1260419–1260419. doi:10.1126/science.1260419.

[54] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43:D447–52. doi:10.1093/nar/gku1003.

[55] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 2012;40:D857–61. doi:10.1093/nar/gkr930.

[56] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res 2010;38:D497–501. doi:10.1093/nar/gkp914.

[57] Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 2016;44:D536–41. doi:10.1093/nar/gkv1115.

[58] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res 2017;45:D408–14. doi:10.1093/nar/gkw985.

[59] Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 2012;6:92. doi:10.1186/1752-0509-6-92.

[60] Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature 2017;545:505–9. doi:10.1038/nature22366.

[61] Liu T, W. Tang G, Capriotti E. Comparative Modeling: The State of the Art and Protein Drug Target Structure Prediction. Comb Chem High Throughput Screen 2011;14:532–47. doi:10.2174/138620711795767811.

[62] Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr Comput Aided-Drug Des

2011;7:146–57. doi:10.2174/157340911795677602.

[63] Jorgensen WL, Duffy EM. Prediction of drug solubility from structure. Adv Drug Deliv Rev 2002;54:355–66. doi:10.1016/S0169-409X(02)00008-X.

[64] Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data. PLoS One 2012;7:e37608. doi:10.1371/journal.pone.0037608.

[65] Ishida S, Umeyama H, Iwadate M, Taguchi Y. Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery n.d.

[66] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23:1241–50. doi:10.1016/J.DRUDIS.2018.01.039.

[67] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Informatics Assoc 2014;21:e278–86. doi:10.1136/amiajnl-2013-002512.

[68] Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug Repositioning: A Machine-Learning Approach through Data Integration. J Cheminform 2013;5:30. doi:10.1186/1758-2946-5-30.

[69] Gönen M, Margolin AA. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics 2014;30:i556–63. doi:10.1093/bioinformatics/btu464.

[70] Jang IS, Neto EC, Guinney, Justin Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Biocomput. 2014, WORLD SCIENTIFIC; 2013, p. 63–74. doi:10.1142/9789814583220_0007.

[71] Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, et al. Predicting in vitro drug sensitivity using Random Forests. Bioinformatics 2011;27:220–4. doi:10.1093/bioinformatics/btq628.

[72] Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and

Drug Repurposing Using Transcriptomic Data. Mol Pharm 2016;13:2524–30. doi:10.1021/acs.molpharmaceut.6b00248.

[73] Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One 2013;8:e61318. doi:10.1371/journal.pone.0061318.

[74] Majumder B, Baraneedharan U, Thiyagarajan S, Radhakrishnan P, Narasimhan H, Dhandapani M, et al. Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. Nat Commun 2015;6:6169. doi:10.1038/ncomms7169.

[75] Chan BA, Hughes BGM. Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. Transl Lung Cancer Res 2015;4:36–54. doi:10.3978/j.issn.2218-6751.2014.05.01.

[76] Tewari KS, Sill MW, Long HJ, Penson RT, Huang H, Ramondetta LM, et al. Improved Survival with Bevacizumab in Advanced Cervical Cancer. N Engl J Med 2014;370:734–43. doi:10.1056/NEJMoa1309748.

[77] Flaherty KT, Robert C, Hersey P, Nathan P, Garbe C, Milhem M, et al. Improved Survival with MEK Inhibition in BRAF-Mutated Melanoma. N Engl J Med 2012;367:107–14. doi:10.1056/NEJMoa1203421.

[78] Gross S, Rahal R, Stransky N, Lengauer C, Hoeflich KP. Targeting cancer with kinase inhibitors. J Clin Invest 2015;125:1780–9. doi:10.1172/JCI76094.

[79] Koinis F, Kotsakis A, Georgoulias V. Small cell lung cancer (SCLC): no treatment advances in recent years. Transl Lung Cancer Res 2016;5:39–50. doi:10.3978/j.issn.2218-6751.2016.01.03.

[80] Enewold L, Harlan LC, Tucker T, McKenzie S. Pancreatic Cancer in the USA: Persistence of Undertreatment and Poor Outcome. J Gastrointest Cancer 2015;46:9–20. doi:10.1007/s12029-014-9668-x.

[81] Bae SY, Kim S, Lee JH, Lee H, Lee SK, Kil WH, et al. Poor prognosis of single hormone receptor- positive breast cancer: similar outcome as triple-negative breast cancer. BMC Cancer 2015;15:138. doi:10.1186/s12885-015-1121-4.

[82] Howard DH, Chernew ME, Abdelgawad T, Smith GL, Sollano J, Grabowski DC. New Anticancer Drugs Associated With Large Increases In Costs And Life

Expectancy. Health Aff 2016;35:1581–7. doi:10.1377/hlthaff.2016.0286.

[83] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ 2016;47:20–33. doi:10.1016/J.JHEALECO.2016.01.012.

[84] DiMasi JA, Grabowski HG, Hansen RW. The Cost of Drug Development. N Engl J Med 2015;372:1972–1972. doi:10.1056/NEJMc1504317.

[85] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740–54. doi:10.1016/J.CELL.2016.06.017.

[86] Sinkala M, Mulder N, Martin DP. Integrative landscape of dysregulated signaling pathways of clinically distinct pancreatic cancer subtypes. Oncotarget 2018;9:29123–39. doi:10.18632/oncotarget.25632.

[87] Waddell N, Pajic M, Patch A-M, Chang DK, Kassahn KS, Bailey P, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 2015;518:495–501. doi:10.1038/nature14169.

[88] Chang DK, Grimmond SM, Biankin A V. Pancreatic cancer genomics. Curr Opin Genet Dev 2014;24:74–81. doi:10.1016/j.gde.2013.12.001.

[89] Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108:479–85. doi:10.1038/bjc.2012.581.

[90] Martin JD, Fukumura D, Duda DG, Boucher Y, Jain RK. Reengineering the Tumor Microenvironment to Alleviate Hypoxia and Overcome Cancer Heterogeneity. Cold Spring Harb Perspect Med 2016;6:a027094. doi:10.1101/cshperspect.a027094.

[91] Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin 2013;63:11–30. doi:10.3322/caac.21166.

[92] De Sousa E Melo F, Vermeulen L, Fessler E, Medema JP. Cancer heterogeneity--a multifaceted view. EMBO Rep 2013;14:686–95. doi:10.1038/embor.2013.92.

[93] Hidalgo M, Cascinu S, Kleeff J, Labianca R, Löhr J-M, Neoptolemos J, et al. Addressing the challenges of pancreatic cancer: future directions for improving outcomes. Pancreatology 2015;15:8–18. doi:10.1016/j.pan.2014.10.001.

[94] Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016;531:47–52. doi:10.1038/nature16965.

[95] Witkiewicz AK, McMillan EA, Balaji U, Baek G, Lin W-C, Mansour J, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun 2015;6:6744. doi:10.1038/ncomms7744.

[96] Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20. doi:10.1038/ng.2764.

[97] Werner HMJ, Mills GB, Ram PT. Cancer Systems Biology: a peek into the future of patient care? Nat Rev Clin Oncol 2014;11:167–76. doi:10.1038/nrclinonc.2014.6.

[98] Biankin A V, Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012;491:399–405. doi:10.1038/nature11547.

[99] Jones S, Zhang X, Williams Parsons D, Cheng-Ho Lin J, Leary RJ, Angenendt P, et al. Cancers Revealed by Global Genomic Analyses Core Signaling Pathways in Human Pancreatic Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses 2008. doi:10.1126/science.1164368.

[100] Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008;321:1801–6. doi:10.1126/science.1164368.

[101] Muzumdar MD, Chen P-Y, Dorans KJ, Chung KM, Bhutkar A, Hong E, et al. Survival of pancreatic cancer cells lacking KRAS function. Nat Commun 2017;8:1090. doi:10.1038/s41467-017-00942-5.

[102] Morton JP, Timpson P, Karim SA, Ridgway RA, Athineos D, Doyle B, et al. Mutant p53 drives metastasis and overcomes growth arrest/senescence in pancreatic cancer. Proc Natl Acad Sci U S A 2010;107:246–51. doi:10.1073/pnas.0908428107.

[103] Collins MA, Bednar F, Zhang Y, Brisset J-C, Galbán S, Galbán CJ, et al. Oncogenic Kras is required for both the initiation and maintenance of

pancreatic cancer in mice. J Clin Invest 2012;122:639–53. doi:10.1172/JCI59227.

[104] Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med 2011;17:500–3. doi:10.1038/nm.2344.

[105] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. doi:10.1073/pnas.0506580102.

[106] Okita R, Maeda A, Shimizu K, Nojima Y, Saisho S, Nakata M. PD-L1 overexpression is partially regulated by EGFR/HER2 signaling and associated with poor prognosis in patients with non-small-cell lung cancer. Cancer Immunol Immunother 2017;66:865–76. doi:10.1007/s00262-017-1986-y.

[107] De Robertis M, Loiacono L, Fusilli C, Poeta ML, Mazza T, Sanchez M, et al. Dysregulation of EGFR Pathway in EphA2 Cell Subpopulation Significantly Associates with Poor Prognosis in Colorectal Cancer. Clin Cancer Res 2017;23:159–70. doi:10.1158/1078-0432.CCR-16-0709.

[108] Ocana A, Vera-Badillo F, Al-Mubarak M, Templeton AJ, Corrales-Sanchez V, Diez-Gonzalez L, et al. Activation of the PI3K/mTOR/AKT Pathway and Survival in Solid Tumors: Systematic Review and Meta-Analysis. PLoS One 2014;9:e95219. doi:10.1371/journal.pone.0095219.

[109] Gatenby RA, Gillies RJ. Why do cancers have high aerobic glycolysis? Nat Rev Cancer 2004;4:891–9. doi:10.1038/nrc1478.

[110] Gogvadze V, Zhivotovsky B, Orrenius S. The Warburg effect and mitochondrial stability in cancer cells. Mol Aspects Med 2010;31:60–74. doi:10.1016/J.MAM.2009.12.004.

[111] Zhao T, Zhu Y, Morinibu A, Kobayashi M, Shinomiya K, Itasaka S, et al. HIF-1-mediated metabolic reprogramming reduces ROS levels and facilitates the metastatic colonization of cancers in lungs. Sci Rep 2015;4:3793. doi:10.1038/srep03793.

[112] Zhong L, D'Urso A, Toiber D, Sebastian C, Henry RE, Vadysirisack DD, et al. The histone deacetylase Sirt6 regulates glucose homeostasis via Hif1alpha.

Cell 2010;140:280–93. doi:10.1016/j.cell.2009.12.041.

[113] Semenza GL. HIF-1 mediates the Warburg effect in clear cell renal carcinoma. J Bioenerg Biomembr 2007;39:231–4. doi:10.1007/s10863-007-9081-2.

[114] Zwaans BMM, Lombard DB. Interplay between sirtuins, MYC and hypoxia-inducible factor in cancer-associated metabolic reprogramming. Dis Model Mech 2014;7:1023–32. doi:10.1242/dmm.016287.

[115] Muzi M, Freeman SD, Burrows RC, Wiseman RW, Link JM, Krohn KA, et al. Kinetic characterization of hexokinase isoenzymes from glioma cells: implications for FDG imaging of human brain tumors. Nucl Med Biol 2001;28:107–16. doi:10.1016/S0969-8051(00)00201-8.

[116] Mathupala SP, Ko YH, Pedersen PL. Hexokinase II: Cancer's double-edged sword acting as both facilitator and gatekeeper of malignancy when bound to mitochondria. Oncogene 2006;25:4777–86. doi:10.1038/sj.onc.1209603.

[117] Simpson IA, Dwyer D, Malide D, Moley KH, Travis A, Vannucci SJ. The facilitative glucose transporter GLUT3: 20 years of distinction. Am J Physiol Endocrinol Metab 2008;295:E242-53. doi:10.1152/ajpendo.90388.2008.

[118] Dupuy F, Tabariès S, Andrzejewski S, Dong Z, Blagih J, Annis MG, et al. PDK1-Dependent Metabolic Reprogramming Dictates Metastatic Potential in Breast Cancer. Cell Metab 2015;22:577–89. doi:10.1016/J.CMET.2015.08.007.

[119] Ma X, Li C, Sun L, Huang D, Li T, He X, et al. Lin28/let-7 axis regulates aerobic glycolysis and cancer progression via PDK1. Nat Commun 2014;5:5212. doi:10.1038/ncomms6212.

[120] Gross S, Rahal R, Stransky N, Lengauer C, Hoeflich KP. Targeting cancer with kinase inhibitors. J Clin Invest 2015;125:1780–9. doi:10.1172/JCI76094.

[121] Yang G, Yang X. Smad4-mediated TGF-beta signaling in tumorigenesis. Int J Biol Sci 2010;6:1–8.

[122] Blobe GC, Schiemann WP, Lodish HF. Role of Transforming Growth Factor β in Human Disease. N Engl J Med 2000;342:1350–8. doi:10.1056/NEJM200005043421807.

[123] Malkoski SP, Wang X-J. Two sides of the story? Smad4 loss in pancreatic cancer versus head-and-neck cancer. FEBS Lett 2012;586:1984–92.

doi:10.1016/j.febslet.2012.01.054.

[124] Zhang YE. Non-Smad pathways in TGF-beta signaling. Cell Res 2009;19:128–39. doi:10.1038/cr.2008.328.

[125] Lee MK, Pardoux C, Hall MC, Lee PS, Warburton D, Qing J, et al. TGF-beta activates Erk MAP kinase signalling through direct phosphorylation of ShcA. EMBO J 2007;26:3957–67. doi:10.1038/sj.emboj.7601818.

[126] Shugang X, Hongfa Y, Jianpeng L, Xu Z, Jingqi F, Xiangxiang L, et al. Prognostic Value of SMAD4 in Pancreatic Cancer: A Meta-Analysis. Transl Oncol 2016;9:1–7. doi:10.1016/j.tranon.2015.11.007.

[127] Dreesen O, Brivanlou AH. Signaling Pathways in Cancer and Embryonic Stem Cells. Stem Cell Rev 2007;3:7–17. doi:10.1007/s12015-007-0004-8.

[128] Zhang J, Wu L-Y, Zhang X-S, Zhang S. Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics 2014;15:271. doi:10.1186/1471-2105-15-271.

[129] Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, et al. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res 2017;45:D995–1002. doi:10.1093/nar/gkw1072.

[130] Thorpe LM, Yuzugullu H, Zhao JJ. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. Nat Rev Cancer 2015;15:7–24. doi:10.1038/nrc3860.

[131] Lin Y, Yang Z, Xu A, Dong P, Huang Y, Liu H, et al. PIK3R1 negatively regulates the epithelial-mesenchymal transition and stem-like phenotype of renal cancer cells through the AKT/GSK3β/CTNNB1 signaling pathway. Sci Rep 2015;5:8997. doi:10.1038/srep08997.

[132] Cheung LW, Mills GB. Targeting therapeutic liabilities engendered by PIK3R1 mutations for cancer treatment. Pharmacogenomics 2016;17:297–307. doi:10.2217/pgs.15.174.

[133] Joerger AC, Fersht AR. The p53 Pathway: Origins, Inactivation in Cancer, and Emerging Therapeutic Approaches. Annu Rev Biochem 2016;85:375–404. doi:10.1146/annurev-biochem-060815-014710.

[134] Visconti R, Della Monica R, Grieco D. Cell cycle checkpoint in cancer: a therapeutically targetable double-edged sword. J Exp Clin Cancer Res

2016;35:153. doi:10.1186/s13046-016-0433-9.

[135] Dean M. ABC Transporters, Drug Resistance, and Cancer Stem Cells. J Mammary Gland Biol Neoplasia 2009;14:3–9. doi:10.1007/s10911-009-9109-9.

[136] Li W, Zhang H, Assaraf YG, Zhao K, Xu X, Xie J, et al. Overcoming ABC transporter-mediated multidrug resistance: Molecular mechanisms and novel therapeutic drug strategies. Drug Resist Updat 2016;27:14–29. doi:10.1016/J.DRUP.2016.05.001.

[137] Kathawala RJ, Gupta P, Ashby CR, Chen Z-S. The modulation of ABC transporter-mediated multidrug resistance in cancer: A review of the past decade. Drug Resist Updat 2015;18:1–17. doi:10.1016/J.DRUP.2014.11.002.

[138] Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics 2013;29:2757–64. doi:10.1093/bioinformatics/btt471.

[139] Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 2012;28:i640–6. doi:10.1093/bioinformatics/bts402.

[140] ZHANG W, LIU HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res 2002;12:9–18. doi:10.1038/sj.cr.7290105.

[141] Fang JY, Richardson BC. The MAPK signalling pathways and colorectal cancer. Lancet Oncol 2005;6:322–7. doi:10.1016/S1470-2045(05)70168-6.

[142] Cheng C-W, Su J-L, Lin C-W, Su C-W, Shih C-H, Yang S-F, et al. Effects of NFKB1 and NFKBIA Gene Polymorphisms on Hepatocellular Carcinoma Susceptibility and Clinicopathological Features. PLoS One 2013;8:e56130. doi:10.1371/journal.pone.0056130.

[143] Royds JA, Dower SK, Qwarnstrom EE, Lewis CE. Response of tumour cells to hypoxia: role of p53 and NFkB. Mol Pathol 1998;51:55–61.

[144] Yao L, Wang L, Li F, Gao X, Wei X, Liu Z. MiR181c inhibits ovarian cancer metastasis and progression by targeting PRKCD expression. Int J Clin Exp

Med 2015;8:15198–205.

[145] Sun M, Estrov Z, Ji Y, Coombes KR, Harris DH, Kurzrock R. Curcumin (diferuloylmethane) alters the expression profiles of microRNAs in human pancreatic cancer cells. Mol Cancer Ther 2008;7:464–73. doi:10.1158/1535-7163.MCT-07-2272.

[146] Stan SD, Singh S V, Brand RE. Chemoprevention strategies for pancreatic cancer. Nat Rev Gastroenterol Hepatol 2010;7:347–56. doi:10.1038/nrgastro.2010.61.

[147] Eser S, Schnieke A, Schneider G, Saur D. Oncogenic KRAS signalling in pancreatic cancer. Br J Cancer 2014;111:817–22. doi:10.1038/bjc.2014.215.

[148] Soares HP, Ming M, Mellon M, Young SH, Han L, Sinnet-Smith J, et al. Dual PI3K/mTOR Inhibitors Induce Rapid Overactivation of the MEK/ERK Pathway in Human Pancreatic Cancer Cells through Suppression of mTORC2. Mol Cancer Ther 2015;14:1014–23. doi:10.1158/1535-7163.MCT-14-0669.

[149] Ning C, Liang M, Liu S, Wang G, Edwards H, Xia Y, et al. Targeting ERK enhances the cytotoxic effect of the novel PI3K and mTOR dual inhibitor VS-5584 in preclinical models of pancreatic cancer. Oncotarget 2017;8:44295–311. doi:10.18632/oncotarget.17869.

[150] Ciuffreda L, Del Curatolo A, Falcone I, Conciatori F, Bazzichetto C, Cognetti F, et al. Lack of growth inhibitory synergism with combined MAPK/PI3K inhibition in preclinical models of pancreatic cancer. Ann Oncol 2017;28:2896–8. doi:10.1093/annonc/mdx335.

[151] Pettazzoni P, Viale A, Shah P, Carugo A, Ying H, Wang H, et al. Genetic events that limit the efficacy of MEK and RTK inhibitor therapies in a mouse model of KRAS-driven pancreatic cancer. Cancer Res 2015;75:1091–101. doi:10.1158/0008-5472.CAN-14-1854.

[152] Stojanovic N, Hassan Z, Wirth M, Wenzel P, Beyer M, Schäfer C, et al. HDAC1 and HDAC2 integrate the expression of p53 mutants in pancreatic cancer. Oncogene 2017;36:1804–15. doi:10.1038/onc.2016.344.

[153] Gurpinar E, Vousden KH. Hitting cancers' weak spots: vulnerabilities imposed by p53 mutation. Trends Cell Biol 2015;25:486–95. doi:10.1016/J.TCB.2015.04.001.

[154] Weissmueller S, Manchado E, Saborowski M, Morris JP, Wagenblast E, Davis CA, et al. Mutant p53 Drives Pancreatic Cancer Metastasis through Cell-Autonomous PDGF Receptor β Signaling. Cell 2014;157:382–94. doi:10.1016/J.CELL.2014.01.066.

[155] Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. vol. 144. Elsevier; 2011. doi:10.1016/j.cell.2011.02.013.

[156] Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab 2016. doi:10.1016/j.cmet.2015.12.006.

[157] Wilde L, Roche M, Domingo-Vidal M, Tanson K, Philp N, Curry J, et al. Metabolic coupling and the Reverse Warburg Effect in cancer: Implications for novel biomarker and anticancer agent development. Semin Oncol 2017;44:198–203. doi:10.1053/j.seminoncol.2017.10.004.

[158] Ramos P, Bentires-Alj M. Mechanism-based cancer therapy: resistance to therapy, therapy for resistance. Oncogene 2015;34:3617–26. doi:10.1038/onc.2014.314.

[159] Obenauf AC, Zou Y, Ji AL, Vanharanta S, Shu W, Shi H, et al. Therapy-induced tumour secretomes promote resistance and tumour progression. Nature 2015;520:368–72. doi:10.1038/nature14336.

[160] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res 2010;1:274–8. doi:10.4103/0974-7788.76794.

[161] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106. doi:10.1186/gb-2010-11-10-r106.

[162] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881–7. doi:10.1093/bioinformatics/btm453.

[163] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst 2015;1:417–25. doi:10.1016/J.CELS.2015.12.004.

[164] Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PLoS One 2010;5:e13984. doi:10.1371/journal.pone.0013984.

[165] Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new

features for data integration and network visualization. Bioinformatics 2011;27:431–2. doi:10.1093/bioinformatics/btq675.

[166] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 2018;46:D661–7. doi:10.1093/nar/gkx1064.

[167] Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. PLOS Comput Biol 2015;11:e1004085. doi:10.1371/journal.pcbi.1004085.

[168] Chen EY, Xu H, Gordonov S, Lim MP, Perkins MH, Ma'ayan A. Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. Bioinformatics 2012;28:105–11. doi:10.1093/bioinformatics/btr625.

[169] Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 2010;26:2438–44. doi:10.1093/bioinformatics/btq466.

[170] Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. Bioinformatics 2009;25:684–6. doi:10.1093/bioinformatics/btp026.

[171] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12:R41. doi:10.1186/gb-2011-12-4-r41.

[172] Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8. doi:10.1038/nature12213.

[173] Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res 2016;44:D1023–31. doi:10.1093/nar/gkv1268.

[174] Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. J Genet Genomics 2017;44:119–21. doi:10.1016/J.JGG.2016.12.004.

[175] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G, et al. Enrichr:

interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013;14:128. doi:10.1186/1471-2105-14-128.

[176] Bahceci I, Dogrusoz U, La KC, Babur Ö, Gao J, Schultz N. PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. Bioinformatics 2017;33:2238–40. doi:10.1093/bioinformatics/btx149.

[177] Sinkala M, Mulder N, Martin D. Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. Sci Rep 2020;10. doi:10.1038/s41598-020-58290-2.

[178] Isaji S, Kawarada Y, Uemoto S. Classification of pancreatic cancer: comparison of Japanese and UICC classifications. Pancreas 2004;28:231–4.

[179] Baylor SM, Berg JW. Cross-classification and survival characteristics of 5,000 cases of cancer of the pancreas. J Surg Oncol 1973;5:335–58. doi:10.1002/jso.2930050410.

[180] Cubilla AL, Fitzgerald PJ. Classification of pancreatic cancer (nonendocrine). Mayo Clin Proc 1979;54:449–58.

[181] Varadhachary GR, Tamm EP, Abbruzzese JL, Xiong HQ, Crane CH, Wang H, et al. Borderline Resectable Pancreatic Cancer: Definitions, Management, and Role of Preoperative Therapy. Ann Surg Oncol 2006;13:1035–46. doi:10.1245/ASO.2006.08.011.

[182] Hidalgo M. Pancreatic Cancer. N Engl J Med 2010;362:1605–17. doi:10.1056/NEJMra0901557.

[183] Biankin A V., Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012;491:399–405. doi:10.1038/nature11547.

[184] Waddell N, Pajic M, Patch A-M, Chang DK, Kassahn KS, Bailey P, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 2015;518:495–501. doi:10.1038/nature14169.

[185] Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016;531:47–52. doi:10.1038/nature16965.

[186] Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes

of pancreatic ductal adenocarcinoma. Nat Genet 2015;47:1168–78. doi:10.1038/ng.3398.

[187] Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One 2013;8:e61318. doi:10.1371/journal.pone.0061318.

[188] Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu TCGAR, Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell 2017;32:185-203.e13. doi:10.1016/j.ccell.2017.07.007.

[189] Dreyer SB, Chang DK, Bailey P, Biankin A V. Pancreatic Cancer Genomes: Implications for Clinical Management and Therapeutic Development. Clin Cancer Res 2017;23:1638–46. doi:10.1158/1078-0432.CCR-16-2411.

[190] Costello E, Greenhalf W, Neoptolemos JP. New biomarkers and targets in pancreatic cancer and their application to treatment. Nat Rev Gastroenterol Hepatol 2012;9:435–44. doi:10.1038/nrgastro.2012.119.

[191] Bournet B, Muscari F, Buscail C, Assenat E, Barthet M, Hammel P, et al. KRAS G12D Mutation Subtype Is A Prognostic Factor for Advanced Pancreatic Adenocarcinoma. Clin Transl Gastroenterol 2016;7:e157. doi:10.1038/ctg.2016.18.

[192] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11:333–7. doi:10.1038/nmeth.2810.

[193] Bauer DC, Gaff C, Dinger ME, Caramins M, Buske FA, Fenech M, et al. Genomics and personalised whole-of-life healthcare. Trends Mol Med 2014;20:479–86. doi:10.1016/J.MOLMED.2014.04.001.

[194] Keogh E, Mueen A. Curse of Dimensionality. Encycl. Mach. Learn. Data Min., Boston, MA: Springer US; 2017, p. 314–5. doi:10.1007/978-1-4899-7687-1_192.

[195] ACM Special Interest Group for Algorithms and Computation Theory. D, SIAM Activity Group on Discrete Mathematics. S, Association for Computing

Machinery., Society for Industrial and Applied Mathematics. Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms. Association for Computing Machinery; 2007.

[196] Ishimura N, Yamasawa K, Rumi MAK, Kadowaki Y, Ishihara S, Amano Y, et al. BRAF and K-ras gene mutations in human pancreatic cancers. Cancer Lett 2003;199:169–73. doi:10.1016/S0304-3835(03)00384-7.

[197] Heidorn SJ, Milagre C, Whittaker S, Nourry A, Niculescu-Duvas I, Dhomen N, et al. Kinase-Dead BRAF and Oncogenic RAS Cooperate to Drive Tumor Progression through CRAF. Cell 2010;140:209–21. doi:10.1016/J.CELL.2009.12.040.

[198] Testa JR, Bellacosa A. AKT plays a central role in tumorigenesis. Proc Natl Acad Sci U S A 2001;98:10983–5. doi:10.1073/pnas.211430998.

[199] Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. J Genet Genomics 2017;44:119–21. doi:10.1016/J.JGG.2016.12.004.

[200] Lee EYHP, Muller WJ. Oncogenes and tumor suppressor genes. Cold Spring Harb Perspect Biol 2010;2:a003236. doi:10.1101/cshperspect.a003236.

[201] de Leon MP. Oncogenes and Tumor Suppressor Genes, Springer, Berlin, Heidelberg; 1994, p. 35–47. doi:10.1007/978-3-642-85076-9_4.

[202] Schmid K, Bago-Horvath Z, Berger W, Haitel A, Cejka D, Werzowa J, et al. Dual inhibition of EGFR and mTOR pathways in small cell lung cancer. Br J Cancer 2010;103:622–8. doi:10.1038/sj.bjc.6605761.

[203] ZAROGOULIDIS P, LAMPAKI S, TURNER JF, HUANG H, KAKOLYRIS S, SYRIGOS K, et al. mTOR pathway: A current, up-to-date mini-review (Review). Oncol Lett 2014;8:2367–70. doi:10.3892/ol.2014.2608.

[204] Harashima H, Dissmeyer N, Schnittger A. Cell cycle control across the eukaryotic kingdom. Trends Cell Biol 2013;23:345–56. doi:10.1016/J.TCB.2013.03.002.

[205] Frizzell RA, Hanrahan JW. Physiology of epithelial chloride and fluid secretion. Cold Spring Harb Perspect Med 2012;2:a009563. doi:10.1101/cshperspect.a009563.

[206] Kang R, Tang D, Schapiro NE, Livesey KM, Farkas A, Loughran P, et al. The

receptor for advanced glycation end products (RAGE) sustains autophagy and limits apoptosis, promoting pancreatic tumor cell survival. Cell Death Differ 2010;17:666–76. doi:10.1038/cdd.2009.149.

[207] Abe R, Yamagishi S. AGE-RAGE System and Carcinogenesis. Curr Pharm Des 2008;14:940–5. doi:10.2174/138161208784139765.

[208] Yang W, Wang K, Zuo W. Neighborhood Component Feature Selection for High-Dimensional Data 2012. doi:10.4304/jcp.7.1.161-168.

[209] Wu Y, Ianakiev K, Govindaraju V. Improved k-nearest neighbor classification. Pattern Recognit 2002;35:2311–8. doi:10.1016/S0031-3203(01)00132-7.

[210] Kecman V, Huang T-M, Vogt M. Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance, Springer, Berlin, Heidelberg; 2005, p. 255–74. doi:10.1007/10984697_12.

[211] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database 2011;2011:bar026–bar026. doi:10.1093/database/bar026.

[212] Platt JC, Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv LARGE MARGIN Classif 1999:61--74.

[213] Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer 2012;12:323–34. doi:10.1038/nrc3261.

[214] Witkiewicz AK, McMillan EA, Balaji U, Baek G, Lin W-C, Mansour J, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun 2015;6:6744. doi:10.1038/ncomms7744.

[215] Dhar S, Nygren P, Csoka K, Botling J, Nilsson K, Larsson R. Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance. Br J Cancer 1996;74:888–96. doi:10.1038/bjc.1996.453.

[216] Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, et al. A community computational challenge to predict the activity of pairs of compounds. Nat Biotechnol 2014;32:1213–22. doi:10.1038/nbt.3052.

[217] Geeleher P, Cox NJ, Huang R. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines.

Genome Biol 2014;15:R47. doi:10.1186/gb-2014-15-3-r47.

[218] Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med 2011;17:500–3. doi:10.1038/nm.2344.

[219] Mohammed A, Janakiram NB, Brewer M, Ritchie RL, Marya A, Lightfoot S, et al. Antidiabetic Drug Metformin Prevents Progression of Pancreatic Cancer by Targeting in Part Cancer Stem Cells and mTOR Signaling. Transl Oncol 2013;6:649-IN7. doi:10.1593/TLO.13556.

[220] Jiao Y, Shi C, Edil BH, Wilde RF de, Klimstra DS, Maitra A, et al. DAXX/ATRX, MEN1, and mTOR Pathway Genes Are Frequently Altered in Pancreatic Neuroendocrine Tumors. Science (80-) 2011;331:1199–203. doi:10.1126/SCIENCE.1200609.

[221] Morran DC, Wu J, Jamieson NB, Mrowinska A, Kalna G, Karim SA, et al. Targeting mTOR dependency in pancreatic cancer. Gut 2014;63:1481–9. doi:10.1136/gutjnl-2013-306202.

[222] Soares HP, Ming M, Mellon M, Young SH, Han L, Sinnet-Smith J, et al. Dual PI3K/mTOR Inhibitors Induce Rapid Overactivation of the MEK/ERK Pathway in Human Pancreatic Cancer Cells through Suppression of mTORC2. Mol Cancer Ther 2015;14:1014–23. doi:10.1158/1535-7163.MCT-14-0669.

[223] Ning C, Liang M, Liu S, Wang G, Edwards H, Xia Y, et al. Targeting ERK enhances the cytotoxic effect of the novel PI3K and mTOR dual inhibitor VS-5584 in preclinical models of pancreatic cancer. Oncotarget 2017;8:44295–311. doi:10.18632/oncotarget.17869.

[224] Loddo M, Kingsbury SR, Rashid M, Proctor I, Holt C, Young J, et al. Cell-cycle-phase progression analysis identifies unique phenotypes of major prognostic and predictive significance in breast cancer. Br J Cancer 2009;100:959–70. doi:10.1038/sj.bjc.6604924.

[225] Teodoro A, Oliveira F, Martins N, Maia G de, Martucci R, Borojevic R. Effect of lycopene on cell viability and cell cycle progression in human cancer cell lines. Cancer Cell Int 2012;12:36. doi:10.1186/1475-2867-12-36.

[226] Williams GH, Stoeber K. The cell cycle and cancer. J Pathol 2012;226:352–64. doi:10.1002/path.3022.

[227] Diaz-Moralli S, Tarrado-Castellarnau M, Miranda A, Cascante M. Targeting cell cycle regulation in cancer therapy. Pharmacol Ther 2013;138:255–71. doi:10.1016/J.PHARMTHERA.2013.01.011.

[228] Dickson MA. Molecular pathways: CDK4 inhibitors for cancer therapy. Clin Cancer Res 2014;20:3379–83. doi:10.1158/1078-0432.CCR-13-1551.

[229] McCubrey JA, Steelman LS, Bertrand FE, Davis NM, Sokolosky M, Abrams SL, et al. GSK-3 as potential target for therapeutic intervention in cancer. Oncotarget 2014;5:2881–911. doi:10.18632/oncotarget.2037.

[230] Madhok BM, Yeluri S, Perry SL, Hughes TA, Jayne DG. Dichloroacetate induces apoptosis and cell-cycle arrest in colorectal cancer cells. Br J Cancer 2010;102:1746–52. doi:10.1038/sj.bjc.6605701.

[231] Fraser SP, Diss JKJ, Chioni A-M, Mycielska ME, Pan H, Yamaci RF, et al. Voltage-Gated Sodium Channel Expression and Potentiation of Human Breast Cancer Metastasis. Clin Cancer Res 2005;11:5381–9. doi:10.1158/1078-0432.CCR-05-0327.

[232] Furuya Y, Lundmo P, Short AD, Gill DL, Isaacs JT. The role of calcium, pH, and cell proliferation in the programmed (apoptotic) death of androgen-independent prostatic cancer cells induced by thapsigargin. Cancer Res 1994;54:6167–75. doi:10.1158/0008-5472.can-04-2146.

[233] Pedersen SF, Stock C. Ion Channels and Transporters in Cancer: Pathophysiology, Regulation, and Clinical Potential. Cancer Res 2013;73:1658–61. doi:10.1158/0008-5472.CAN-12-4188.

[234] Monteith GR, Davis FM, Roberts-Thomson SJ. Calcium channels and pumps in cancer: changes and consequences. J Biol Chem 2012;287:31666–73. doi:10.1074/jbc.R112.343061.

[235] Varona A, Blanco L, López JI, Gil J, Agirregoitia E, Irazusta J, et al. Altered levels of acid, basic, and neutral peptidase activity and expression in human clear cell renal cell carcinoma. Am J Physiol Physiol 2007;292:F780–8. doi:10.1152/ajprenal.00148.2006.

[236] Larrinaga G, Blanco L, Sanz B, Perez I, Gil J, Unda M, et al. The impact of peptidase activity on clear cell renal cell carcinoma survival. Am J Physiol Physiol 2012;303:F1584–91. doi:10.1152/ajprenal.00477.2012.

[237] Duesberg P, Rausch C, Rasnick D, Hehlmann R, Woodgate R, Goodman MF, et al. Genetic instability of cancer cells is proportional to their degree of aneuploidy. Proc Natl Acad Sci 1998;95:13692–7. doi:10.1073/pnas.95.23.13692.

[238] Coyle KM, Boudreau JE, Marcato P. Genetic Mutations and Epigenetic Modifications: Driving Cancer and Informing Precision Medicine. Biomed Res Int 2017;2017:9620870. doi:10.1155/2017/9620870.

[239] Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis 2010;31:27–36. doi:10.1093/carcin/bgp220.

[240] Reddy KB. MicroRNA (miRNA) in cancer. Cancer Cell Int 2015;15:38. doi:10.1186/s12935-015-0185-1.

[241] Mishra NK, Guda C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. Oncotarget 2017;8:28990–9012. doi:10.18632/oncotarget.15993.

[242] Khatri I, Ganguly K, Sharma S, Carmicheal J, Kaur S, Batra SK, et al. Systems Biology Approach to Identify Novel Genomic Determinants for Pancreatic Cancer Pathogenesis. Sci Rep 2019;9:123. doi:10.1038/s41598-018-36328-w.

[243] Kazanets A, Shorstova T, Hilmi K, Marques M, Witcher M. Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. Biochim Biophys Acta - Rev Cancer 2016;1865:275–88. doi:10.1016/J.BBCAN.2016.04.001.

[244] Chatterjee A, Rodger EJ, Eccles MR. Epigenetic drivers of tumourigenesis and cancer metastasis. Semin Cancer Biol 2018;51:149–59. doi:10.1016/J.SEMCANCER.2017.08.004.

[245] Shen H, Laird PW. Interplay between the Cancer Genome and Epigenome. Cell 2013;153:38–55. doi:10.1016/J.CELL.2013.03.008.

[246] Prat A, Parker JS, Fan C, Perou CM. PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. Breast Cancer Res Treat 2012;135:301–6. doi:10.1007/s10549-012-2143-0.

[247] Volm M, Efferth T. Prediction of Cancer Drug Resistance and Implications for Personalized Medicine. Front Oncol 2015;5:282. doi:10.3389/fonc.2015.00282.

[248] Chu C, Hsu A-L, Chou K-H, Bandettini P, Lin C. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 2012;60:59–70. doi:10.1016/J.NEUROIMAGE.2011.11.066.

[249] Hastie T, Tibshirani R, Friedman J. Unsupervised Learning, Springer, New York, NY; 2009, p. 485–585. doi:10.1007/978-0-387-84858-7_14.

[250] Jolliffe I. Principal Component Analysis. Int. Encycl. Stat. Sci., Berlin, Heidelberg: Springer Berlin Heidelberg; 2011, p. 1094–6. doi:10.1007/978-3-642-04898-2_455.

[251] Benjamini Y. Discovering the false discovery rate. J R Stat Soc Ser B (Statistical Methodol 2010;72:405–16. doi:10.1111/j.1467-9868.2010.00746.x.

[252] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863. doi:10.3389/fpsyg.2013.00863.

[253] Wong CK, Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, et al. The UCSC Interaction Browser: multidimensional data views in pathway context. Nucleic Acids Res 2013;41:W218–24. doi:10.1093/nar/gkt473.

[254] Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, et al. Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res 2011;21:193–202. doi:10.1101/gr.108662.110.

[255] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. doi:10.1038/nmeth.1226.

[256] MathWorks T. MATLAB (R2017b). MathWorks Inc 2017. doi:10.1007/s10766-008-0082-5.

[257] Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990;36:265–70.

[258] Research AB-J of ML, 2011 undefined. Convergence rates of efficient global optimization algorithms. JmlrOrg n.d.

[259] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms 2012:2951–9.

[260] Gelbart MA, Snoek J, Adams RP. Bayesian Optimization with Unknown

Constraints 2014.

[261] Friedman JH, Bentley JL, Finkel RA. AN ALGORITHM FOR FINDING BEST MATCHES IN LOGARITHMIC EXPECTED TIME. 1975.

[262] Ben-David A. Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Syst Appl 2008;34:825–32. doi:10.1016/J.ESWA.2006.10.022.

[263] Sinkala M, Mulder N, Patrick Martin D. Metabolic gene alterations impact the clinical aggressiveness and drug responses of 32 human cancers. Commun Biol 2019;2. doi:10.1038/s42003-019-0666-1.

[264] DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. Cell Metab 2008;7:11–20. doi:10.1016/J.CMET.2007.10.002.

[265] Courtnay R, Ngo DC, Malik N, Ververis K, Tortorella SM, Karagiannis TC. Cancer metabolism and the Warburg effect: the role of HIF-1 and PI3K. Mol Biol Rep 2015;42:841–51. doi:10.1007/s11033-015-3858-x.

[266] Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab 2016;23:27–47. doi:10.1016/J.CMET.2015.12.006.

[267] Johnson C, Warmoes MO, Shen X, Locasale JW. Epigenetics and cancer metabolism. Cancer Lett 2015;356:309–14. doi:10.1016/J.CANLET.2013.09.043.

[268] Koppenol WH, Bounds PL, Dang C V. Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer 2011;11:325–37. doi:10.1038/nrc3038.

[269] Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. Cell Rep 2018;23:255-269.e4. doi:10.1016/J.CELREP.2018.03.077.

[270] Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. Nat Commun 2018;9:5330. doi:10.1038/s41467-018-07232-8.

[271] Cantor JR, Sabatini DM. Cancer Cell Metabolism: One Hallmark, Many Faces. Cancer Discov 2012;2:881–98. doi:10.1158/2159-8290.CD-12-0345.

[272] Patel S, Ahmed S. Emerging field of metabolomics: Big promise for cancer

biomarker identification and drug discovery. J Pharm Biomed Anal 2015;107:63–74. doi:10.1016/J.JPBA.2014.12.020.

[273] Vander Heiden MG, Locasale JW, Swanson KD, Sharfi H, Heffron GJ, Amador-Noguez D, et al. Evidence for an alternative glycolytic pathway in rapidly proliferating cells. Science 2010;329:1492–9. doi:10.1126/science.1188015.

[274] Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. J Pharm Biomed Anal 2014;87:1–11. doi:10.1016/J.JPBA.2013.08.041.

[275] Stine ZE, Walton ZE, Altman BJ, Hsieh AL, Dang C V. MYC, Metabolism, and Cancer. Cancer Discov 2015;5:1024–39. doi:10.1158/2159-8290.CD-15-0507.

[276] Beloribi-Djefaflia S, Vasseur S, Guillaumond F. Lipid metabolic reprogramming in cancer cells. Oncogenesis 2016;5:e189–e189. doi:10.1038/oncsis.2015.49.

[277] Zong W-X, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell 2016;61:667–76. doi:10.1016/J.MOLCEL.2016.02.011.

[278] Liberti M V., Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? Trends Biochem Sci 2016;41:211–8. doi:10.1016/J.TIBS.2015.12.001.

[279] Hsu PP, Sabatini DM. Cancer Cell Metabolism: Warburg and Beyond. Cell 2008;134:703–7. doi:10.1016/J.CELL.2008.08.021.

[280] DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv 2016;2:e1600200. doi:10.1126/sciadv.1600200.

[281] Serna E, Morales JM, Mata M, Gonzalez-Darder J, San Miguel T, Gil-Benso R, et al. Gene Expression Profiles of Metabolic Aggressiveness and Tumor Recurrence in Benign Meningioma. PLoS One 2013;8:e67291. doi:10.1371/journal.pone.0067291.

[282] Martinez-Outschoorn UE, Peiris-Pagés M, Pestell RG, Sotgia F, Lisanti MP. Cancer metabolism: a therapeutic perspective. Nat Rev Clin Oncol 2017;14:11–31. doi:10.1038/nrclinonc.2016.60.

[283] Pertega-Gomes N, Felisbino S, Massie CE, Vizcaino JR, Coelho R, Sandi C, et al. A glycolytic phenotype is associated with prostate cancer progression and aggressiveness: a role for monocarboxylate transporters as metabolic targets for therapy. J Pathol 2015;236:517–30. doi:10.1002/path.4547.

[284] Wang L, Hu H, Pan Y, Wang R, Li Y, Shen L, et al. PIK3CA Mutations Frequently Coexist with EGFR/KRAS Mutations in Non-Small Cell Lung Cancer and Suggest Poor Prognosis in EGFR/KRAS Wildtype Subgroup. PLoS One 2014;9:e88291. doi:10.1371/journal.pone.0088291.

[285] Hao Y, Samuels Y, Li Q, Krokowski D, Guan B-J, Wang C, et al. Oncogenic PIK3CA mutations reprogram glutamine metabolism in colorectal cancer. Nat Commun 2016;7:11971. doi:10.1038/ncomms11971.

[286] Vivanco I, Sawyers CL. The phosphatidylinositol 3-Kinase–AKT pathway in human cancer. Nat Rev Cancer 2002;2:489–501. doi:10.1038/nrc839.

[287] Majewski IJ, Nuciforo P, Mittempergher L, Bosma AJ, Eidtmann H, Holmes E, et al. PIK3CA Mutations Are Associated With Decreased Benefit to Neoadjuvant Human Epidermal Growth Factor Receptor 2–Targeted Therapies in Breast Cancer. J Clin Oncol 2015;33:1334. doi:10.1200/JCO.2014.55.2158.

[288] Whitfield AJ, Barrett PHR, van Bockxmeer FM, Burnett JR. Lipid disorders and mutations in the APOB gene. Clin Chem 2004;50:1725–32. doi:10.1373/clinchem.2004.038026.

[289] Ashur-Fabian O, Har-Zahav A, Shaish A, Wiener Amram H, Margalit O, Weizer-Stern O, et al. apoB and apobec1, two genes key to lipid metabolism, are transcriptionally regulated by p53. Cell Cycle 2010;9:3785–94. doi:10.4161/cc.9.18.12993.

[290] Borgquist S, Butt T, Almgren P, Shiffman D, Stocks T, Orho-Melander M, et al. Apolipoproteins, lipids and risk of cancer. Int J Cancer 2016;138:2648–56. doi:10.1002/ijc.30013.

[291] Marzolo M-P, Farfán P. New Insights into the Roles of Megalin/LRP2 and the Regulation of its Functional Expression. Biol Res 2011;44:89–105. doi:10.4067/S0716-97602011000100012.

[292] Anderson LN, Cotterchio M, Cole DEC, Knight JA. Vitamin D-Related Genetic Variants, Interactions with Vitamin D Exposure, and Breast Cancer Risk among Caucasian Women in Ontario. Cancer Epidemiol Biomarkers Prev 2011;20:1708–17. doi:10.1158/1055-9965.EPI-11-0300.

[293] Andersen RK, Hammer K, Hager H, Christensen JN, Ludvigsen M, Honoré B,

et al. Melanoma tumors frequently acquire *LRP2* /megalin expression, which modulates melanoma cell proliferation and survival rates. Pigment Cell Melanoma Res 2015;28:267–80. doi:10.1111/pcmr.12352.

[294] Belting M. Glycosaminoglycans in cancer treatment. Thromb Res 2014;133:S95–101. doi:10.1016/S0049-3848(14)50016-3.

[295] Afratis N, Gialeli C, Nikitovic D, Tsegenidis T, Karousou E, Theocharis AD, et al. Glycosaminoglycans: key players in cancer cell biology and treatment. FEBS J 2012;279:1177–97. doi:10.1111/J.1742-4658.2012.08529.X@10.1002/(ISSN)1742-4658(CAT)FREEREVIEWCONTENT(VI)REVIEWS1213.

[296] Nikitovic D, Kouvidi K, Voudouri K, Berdiaki A, Karousou E, Passi A, et al. The motile breast cancer phenotype roles of proteoglycans/glycosaminoglycans. Biomed Res Int 2014;2014:124321. doi:10.1155/2014/124321.

[297] Gerner EW. Cancer Chemoprevention Locks onto a New Polyamine Metabolic Target. Cancer Prev Res 2010;3:125–7. doi:10.1158/1940-6207.CAPR-09-0252.

[298] Paz EA, Garcia-Huidobro J, Ignatenko NA. Polyamines in cancer. Adv Clin Chem 2011;54:45–70. doi:10.1016/B978-0-12-387025-4.00002-9.

[299] Murray-Stewart TR, Woster PM, Casero RA, Jr. Targeting polyamine metabolism for cancer therapy and prevention. Biochem J 2016;473:2937. doi:10.1042/BCJ20160383.

[300] Nowotarski SL, Woster PM, Casero RA, Jr. Polyamines and cancer: implications for chemotherapy and chemoprevention. Expert Rev Mol Med 2013;15:e3. doi:10.1017/erm.2013.3.

[301] Casero RA, Murray Stewart T, Pegg AE. Polyamine metabolism and cancer: treatments, challenges and opportunities. Nat Rev Cancer 2018;18:681–95. doi:10.1038/s41568-018-0050-3.

[302] Babbar N, Gerner EW. Targeting Polyamines and Inflammation for Cancer Prevention, Springer, Berlin, Heidelberg; 2010, p. 49–64. doi:10.1007/978-3-642-10858-7_4.

[303] Herbert B-S, Chanoux RA, Liu Y, Baenziger PH, Goswami CP, McClintick JN, et al. A molecular signature of normal breast epithelial and stromal cells from

Li-Fraumeni syndrome mutation carriers. Oncotarget 2010;1:405–22. doi:10.18632/oncotarget.101004.

[304] Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. Nat Rev Cancer 2013;13:227–32. doi:10.1038/nrc3483.

[305] Dória ML, Cotrim Z, Macedo B, Simões C, Domingues P, Helguero L, et al. Lipidomic approach to identify patterns in phospholipid profiles and define class differences in mammary epithelial and breast cancer cells. Breast Cancer Res Treat 2012;133:635–48. doi:10.1007/s10549-011-1823-5.

[306] Li J, Ren S, Piao H, Wang F, Yin P, Xu C, et al. Integration of lipidomics and transcriptomics unravels aberrant lipid metabolism and defines cholesteryl oleate as potential biomarker of prostate cancer. Sci Rep 2016;6:20984. doi:10.1038/srep20984.

[307] Marien E, Meister M, Muley T, Fieuws S, Bordel S, Derua R, et al. Non-small cell lung cancer is characterized by dramatic changes in phospholipid profiles. Int J Cancer 2015;137:1539–48. doi:10.1002/ijc.29517.

[308] Zalba S, ten Hagen TLM. Cell membrane modulation as adjuvant in cancer therapy. Cancer Treat Rev 2017;52:48–57. doi:10.1016/J.CTRV.2016.10.008.

[309] Gogvadze V, Orrenius S, Zhivotovsky B. Mitochondria in cancer cells: what is so special about them? Trends Cell Biol 2008;18:165–73. doi:10.1016/J.TCB.2008.01.006.

[310] Rohlf FJ, Fisher DR. Tests for Hierarchical Structure in Random Data Sets. Syst Biol 1968;17:407–12. doi:10.1093/sysbio/17.4.407.

[311] Stein WD, Litman T, Fojo T, Bates SE. Cancer Research. Cancer Res 2004;62:2281–6. doi:10.1158/0008-5472.can-03-3383.

[312] Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. Nat Commun 2013;4:2126. doi:10.1038/ncomms3126.

[313] Ge Z, Leighton JS, Wang Y, Peng X, Chen Z, Chen H, et al. Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types. Cell Rep 2018;23:213-226.e3. doi:10.1016/J.CELREP.2018.03.047.

[314] Zhao S, Lin Y, Xu W, Jiang W, Zha Z, Wang P, et al. Glioma-Derived Mutations in IDH1 Dominantly Inhibit IDH1 Catalytic Activity and Induce HIF-

1α. Science (80-) 2009;324:261–5. doi:10.1126/SCIENCE.1170944.

[315] Semenza GL. HIF-1 mediates metabolic responses to intratumoral hypoxia and oncogenic mutations. J Clin Invest 2013;123:3664–71. doi:10.1172/JCI67230.

[316] Dang C V. The Interplay Between MYC and HIF in the Warburg Effect, Springer, Berlin, Heidelberg; 2008, p. 35–53. doi:10.1007/2789_2008_088.

[317] Icard P, Shulman S, Farhat D, Steyaert J-M, Alifano M, Lincet H. How the Warburg effect supports aggressiveness and drug resistance of cancer cells? Drug Resist Updat 2018;38:1–11. doi:10.1016/J.DRUP.2018.03.001.

[318] Silver JK, Baima J. Cancer Prehabilitation. Am J Phys Med Rehabil 2013;92:715–27. doi:10.1097/PHM.0b013e31829b4afe.

[319] Postow MA, Sidlow R, Hellmann MD. Immune-Related Adverse Events Associated with Immune Checkpoint Blockade. N Engl J Med 2018;378:158–68. doi:10.1056/NEJMra1703481.

[320] Epstein JB, Thariat J, Bensadoun R-J, Barasch A, Murphy BA, Kolnick L, et al. Oral complications of cancer and cancer therapy. CA Cancer J Clin 2012;62:400–22. doi:10.3322/caac.21157.

[321] Ranpura V, Hapani S, Wu S. Treatment-Related Mortality With Bevacizumab in Cancer Patients. JAMA 2011;305:487. doi:10.1001/jama.2011.51.

[322] Longton E, Schmit K, Fransolet M, Clement F, Michiels C. Appropriate Sequence for Afatinib and Cisplatin Combination Improves Anticancer Activity in Head and Neck Squamous Cell Carcinoma. Front Oncol 2018;8:432. doi:10.3389/fonc.2018.00432.

[323] Park K, Tan E-H, O'Byrne K, Zhang L, Boyer M, Mok T, et al. Afatinib versus gefitinib as first-line treatment of patients with EGFR mutation-positive non-small-cell lung cancer (LUX-Lung 7): a phase 2B, open-label, randomised controlled trial. Lancet Oncol 2016;17:577–89. doi:10.1016/S1470-2045(16)30033-X.

[324] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. No Title. Nature 2012;483. doi:10.1038/nature11003.

[325] Ammad-Ud-Din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, et al. Drug response prediction by inferring pathway-response associations

with kernelized Bayesian matrix factorization. Bioinformatics, vol. 32, 2016. doi:10.1093/bioinformatics/btw433.

[326] Hew KE, Miller PC, El-Ashry D, Sun J, Besser AH, Ince TA, et al. MAPK Activation Predicts Poor Outcome and the MEK Inhibitor, Selumetinib, Reverses Antiestrogen Resistance in ER-Positive High-Grade Serous Ovarian Cancer. Clin Cancer Res 2016;22:935–47. doi:10.1158/1078-0432.CCR-15-0534.

[327] Kalady MF, Dejulius KL, Sanchez JA, Jarrar A, Liu X, Manilich E, et al. BRAF mutations in colorectal cancer are associated with distinct clinical characteristics and worse prognosis. Dis Colon Rectum 2012;55:128–33. doi:10.1097/DCR.0b013e31823c08b3.

[328] Derin D, Eralp Y, Ozluk Y, Yavuz E, Guney N, Saip P, et al. Lower Level of MAPK Expression Is Associated with Anthracycline Resistance and Decreased Survival in Patients with Hormone Receptor Negative Breast Cancer. Cancer Invest 2008;26:671–9. doi:10.1080/07357900801891628.

[329] Hawkins C, Walker E, Mohamed N, Zhang C, Jacob K, Shirinian M, et al. BRAF-KIAA1549 Fusion Predicts Better Clinical Outcome in Pediatric Low-Grade Astrocytoma. Clin Cancer Res 2011;17:4790–8. doi:10.1158/1078-0432.CCR-11-0034.

[330] Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[331] Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient, Springer, Berlin, Heidelberg; 2009, p. 1–4. doi:10.1007/978-3-642-00296-0_5.

[332] Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004;306:640–3. doi:10.1126/science.1104635.

[333] Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics 2013;29:1060–7. doi:10.1093/bioinformatics/btt099.

[334] Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med 2009;1:2. doi:10.1186/gm2.

[335] Bousquet J, Anto JM, Sterk PJ, Adcock IM, Chung K, Roca J, et al. Systems

medicine and integrated care to combat chronic noncommunicable diseases. Genome Med 2011;3:43. doi:10.1186/gm259.

[336] Eduati F, De Las Rivas J, Di Camillo B, Toffolo G, Saez-Rodriguez J. Integrating literature-constrained and data-driven inference of signalling networks. Bioinformatics 2012;28:2311–7. doi:10.1093/bioinformatics/bts363.

[337] Tan M, Alshalalfa M, Alhajj R, Polat F. Influence of Prior Knowledge in Constraint-Based Learning of Gene Regulatory Networks. IEEE/ACM Trans Comput Biol Bioinforma 2011;8:130–42. doi:10.1109/TCBB.2009.58.

[338] Ritchie MD, Davis JR, Aschard H, Battle A, Conti D, Du M, et al. Incorporation of Biological Knowledge Into the Study of Gene-Environment Interactions. Am J Epidemiol 2017;186:771–7. doi:10.1093/aje/kwx229.

[339] de Souza MC, Higa CHA. Reverse Engineering of Gene Regulatory Networks Combining Dynamic Bayesian Networks and Prior Biological Knowledge, Springer, Cham; 2018, p. 323–36. doi:10.1007/978-3-319-95162-1_22.

[340] Squillario M, Barbieri M, Verri A, Barla A, Squillario M, Barbieri M, et al. Enhancing Interpretability of Gene Signatures with Prior Biological Knowledge. Microarrays 2016;5:15. doi:10.3390/microarrays5020015.

[341] Haibe-Kains B, El-Hachem N, Birkbak N, Jin A, Beck A, Aerts H, et al. Inconsistency in large pharmacogenomic studies. Nature 2013;504:389–93. doi:10.1038/NATURE12831.

[342] Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst 2018;6:271-281.e7. doi:10.1016/J.CELS.2018.03.002.

[343] Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature 2016;533:333–7. doi:10.1038/nature17987.

[344] Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature 2015;528:84–7. doi:10.1038/nature15736.

[345] Geeleher P, Gamazon ER, Seoighe C, Cox NJ, Huang RS. Consistency in large pharmacogenomic studies. Nature 2016;540:E1–2.

doi:10.1038/nature19838.

[346] Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol 2014;15:484. doi:10.1186/s13059-014-0484-1.

[347] Jagodnik KM, Koplev S, Jenkins SL, Ohno-Machado L, Paten B, Schurer SC, et al. Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. J Biomed Inform 2017;71:49–57. doi:10.1016/J.JBI.2017.05.006.

[348] Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol Sci 2014;35:450–60. doi:10.1016/J.TIPS.2014.07.001.

[349] Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, et al. Integrative Molecular Characterization of Malignant Pleural Mesothelioma. Cancer Discov 2018;8:1548–65. doi:10.1158/2159-8290.CD-18-0804.

[350] Network TCGAR. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. Https://DoiOrg/101056/NEJMoa1505917 2016. doi:10.1056/NEJMOA1505917.

[351] Cancer Genome Atlas Research Network A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015;163:1011–25. doi:10.1016/j.cell.2015.10.025.

[352] Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell 2013;155:462–77. doi:10.1016/j.cell.2013.09.034.

[353] Network TCGAR. Integrated genomic characterization of oesophageal carcinoma. Nature 2017;541:169–75. doi:10.1038/nature20805.

[354] Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7. doi:10.1038/nature11252.

[355] Farshidfar F, Zheng S, Gingras M-C, Newton Y, Shih J, Robertson AG, et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. Cell Rep 2017;18:2780–94. doi:10.1016/j.celrep.2017.02.033.

[356] Network TCGAR. Integrated genomic and molecular characterization of cervical cancer. Nature 2017;543:378–84. doi:10.1038/nature21386.

[357] Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70. doi:10.1038/nature11412.

[358] Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell 2016;29:723–36. doi:10.1016/j.ccell.2016.04.002.

[359] Network TCGAR. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med 2013;368:2059–74. doi:10.1056/NEJMoa1301689.

[360] Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-) 2015;348:648–60. doi:10.1126/science.1262110.

[361] Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 2014;13:397–406. doi:10.1074/mcp.M113.035600.

[362] Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science 2015;348:660–5. doi:10.1126/science.aaa0355.

[363] Stewart-Ornstein J, Lahav G. p53 dynamics in response to DNA damage vary across cell lines and are shaped by efficiency of DNA repair and activity of the kinase ATM. Sci Signal 2017;10:eaah6671. doi:10.1126/scisignal.aah6671.

[364] Caunt CJ, Keyse SM. Dual-specificity MAP kinase phosphatases (MKPs). FEBS J 2013;280:489–504. doi:10.1111/j.1742-4658.2012.08716.x.

[365] Britson JS, Barton F, Balko JM, Black EP. Deregulation of DUSP activity in EGFR-mutant lung cancer cell lines contributes to sustained ERK1/2 signaling. Biochem Biophys Res Commun 2009;390:849–54. doi:10.1016/J.BBRC.2009.10.061.

[366] Prabhakar S, Asuthkar S, Lee W, Chigurupati S, Zakharian E, Tsung AJ, et al. Targeting DUSPs in glioblastomas - wielding a double-edged sword? Cell Biol

Int 2014;38:145–53. doi:10.1002/cbin.10201.

[367] Sinkala M, Nkhoma P, Mulder N, Martin DP. Integrated Molecular Characterisation of the MAPK Pathways in Human Cancers. BioRxiv 2020:2020.03.14.989350. doi:10.1101/2020.03.14.989350.

[368] Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res 2015;43:D470-8. doi:10.1093/nar/gku1204.

Appendices

Appendix A: Supplementary File Description

Supplementary files for Chapter 3

Supplementary file 1: Enrichment analyses of the KEGG pathways, Kinase Enrichment Analysis, and GO biological process results for the subtypes of pancreatic cancer obtained using the differentially expresses proteins, mRNA transcripts, and DNA methylation.

Supplementary file 2: Differentially expressed mRNA transcripts, DNA methylation, proteins and miRNAs for the subtypes of pancreatic cancer.

Supplementary file 3: Biomarker sets of mRNA transcripts, DNA methylation, proteins and miRNAs that each may be used to classify pancreatic cancer patients into the integrative (SNF) subtypes.

Supplementary files for Chapter 4

Supplementary file 1: Data of cancer studies and genomic alterations: The spreadsheet contains the following results according to the sheet name. TCGA Cancer Types; list and description of TCGA studies. GDSC Cancer Types; list and description of GDSC studies. Metabolic Pathways - First Tier; list of and description of first-tier Reactome metabolic pathways. GDSC vs TCGA Alterations; comparison between genomic alterations of GDSC cancer cell lines and TCGA cancers.

Supplementary file 2: Dose-Response Differences. The spreadsheet contains the following results according to the sheet name. Pathway Drug Response Results; comparison of dose-responses of the HM and LM cell lines for all anticancer drugs that target the 24 signalling pathways and/or biological processes. 251 Drug comparisons; comparison of dose-responses of the HM and LM cell lines to each of the 251 anticancer drugs. 41 Sig Drug Results; 41 drugs that showed statistically significant differences in their dose-response comparisons between the HM and LM

cancer cell lines. Within Cancer Efficacy Variation; results that show differences in log IC50 values between cancer cell lines of each cancer type with or without alterations to genes involved in each of the 16 first-tier metabolic pathways.

Supplementary File 3: Differential Expression Results; list of statistically significantly differentially expressed genes between the HM and LM cancer supertypes. HM Supertype Upregulated Genes: list of upregulated genes in the HM tumours compared to the LM tumours. LM Supertype Upregulated Genes: list of upregulated genes in the LM tumours compared to the HM tumours. LM Supertype GO Mol Function; enriched gene ontology molecular function in the LM cancer supertypes. HM Supertype GO Mol Function; enriched gene ontology molecular function in the HM cancer supertypes.

Appendix B:

Table B-1: *Key resources used in this thesis*

Resource	Paper	Source
Software and Algorithms		
MATLAB	N/A	https://www.mathworks.com/products/matlab.html
Python	N/A	https://www.python.org
R	N/A	https://www.r-project.org
Gene Set Enrichment Analysis	[105]	http://software.broadinstitute.org/gsea/index.jsp
Cytoscape	[165]	https://cytoscape.org
Enrichment Map	[164]	https://nrnb.org/tools/enrichmentmap.html
yEd	N/A	https://www.yworks.com/products/yed
PathVasio3	[167]	https://www.pathvisio.org/
Expression2Kinases	[168]	http://www.maayanlab.net/X2K/
Kinase Enrichment Analysis	[170]	https://www.maayanlab.net/KEA2/
ChIP-x Enrichment Analysis	[169]	http://amp.pharm.mssm.edu/Enrichr/
Enrichr	[175]	http://amp.pharm.mssm.edu/Enrichr/
Co-occurring Mutated Driver Pathways	[128]	http://page.amss.ac.cn/shihua.zhang/software.html
PARADIGM-Shift	[139]	https://github.com/ucscCancer/paradigm-scripts
Tied Diffusion Through Interacting Event	[138]	https://github.com/epaull/TieDIE
PathwayMapper	[176]	https://www.pathwaymapper.org/
Similarity Fusion Network	[192]	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
Big Data and Biological Knowledge Resources		
The Cancer Genome Atlas Data Portal	[8]	https://portal.gdc.cancer.gov/
International Cancer Genomics Consortium Data Portal	[9]	https://dcc.icgc.org/
cBio Cancer Data Portal	[50]	https://www.cbioportal.org/
Genomics of Drug Sensitivity in Cancer Database	[5]	https://www.cancerrxgene.org/
Uniprot Knowledgebase	[36]	https://www.uniprot.org/
KEGG Pathways Database	[11]	https://www.genome.jp/kegg/pathway.html
Reactome Pathways Database	[10,31]	https://reactome.org/
WikiPathways Database	[166]	https://www.wikipathways.org/index.php/WikiPathways
Kinase Enrichment Database	[170]	https://www.maayanlab.net/KEA2/
Molecular Signatures Database	[163]	http://software.broadinstitute.org/gsea/msigdb
Sanger Consensus Cancer Gene Database	[47]	https://cancer.sanger.ac.uk/census
Tumour Suppressor Gene Database	[173]	https://bioinfo.uth.edu/TSGene/
Oncogene Database	[199]	http://ongene.bioinfo-minzhao.org/
UCSC Super Pathway Database	[253]	https://github.com/ucscCancer/superpathway_db
Cancer Cell Line Encyclopaedia	[6]	https://portals.broadinstitute.org/ccle
Gene Ontology Consortium	[39]	http://geneontology.org/
Pharos Knowledgebase	[129]	https://pharos.nih.gov/
BioGrid Database	[368]	https://thebiogrid.org/
ChEA Database	[169]	http://amp.pharm.mssm.edu/Enrichr/
ClinicalTrail.gov	[35]	https://clinicaltrials.gov/
DrugBank	[33]	https://www.drugbank.ca/