TECHNISCHE UNIVERSITÄT MÜNCHEN

Klinikum rechts der Isar
Chirurgische Klinik und Poliklinik

In vivo and *in silico* models of early pancreatic carcinogenesis

Philipp Edgar Bruns

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Roland M. Schmid
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Helmut Friess
	2. Univ.-Prof. Dr. Dr. Fabian J. Theis

Die Dissertation wurde am 04.09.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 26.05.2016 angenommen.

Abstract

Oncogenic Kras mutations are a critical genetic event in early pancreatic carcinogenesis. However, in the adult organ, the expression of mutated Kras alone is not sufficient to induce tissue transformation. Instead, inflammation is required as an additional trigger. Here, we perform experiments with two well-established models of acute pancreatitis and early pancreatic carcinogenesis, initialized with caerulein injections in wild-type and $Kras^{G12D}$-mutated mice. Gene expression profiling and histological analysis with a high temporal resolution are applied in order to allow for a comparative analysis of both mouse strains.

Based on our findings, we define pancreatic regeneration as a process characterized by three distinct stages each showing a specific histological and molecular profile: inflammation, regeneration and refinement. Comparative analysis shows that regeneration in $Kras^{G12D}$ mice is impaired and characterized by a sustained inflammatory status.

In addition, we observed that sustained inflammation in $Kras^{G12D}$ mice is accompanied by altered proliferation cascades with respect to acinar, progenitor-like and mesenchymal cells. In particular, a massive mesenchymal expansion is observed. We conclude that these characteristics, sustained inflammation and altered prolifertion cascades, are hallmarks of early pancreatic carcinogenesis mediating precursor lesion formation.

We extracted a signature characterizing the inflammatory status and analyzed a dataset of 45 human pancreatic ductal adenocarcinoma (PDAC) with respect to homologous genes of this signature. We found that patients could be classified into two distinct clusters, with one exhibiting an increased average survival time (high signature expression) compared to the other one (low signature expression). According to this finding, we conclude that the inflammatory status could be possibly regarded as an indicator of early pancreatic carcinogenesis.

Based on our results, we set up a "dual model" of pancreatic regeneration and early pancreatic carcinogenesis, comprising four states (healthy, inflammation, regeneration and refinement). As divergence of wild-type and $Kras^{G12D}$ profiles starts during the *inflammation-regeneration transition* in wild-type mice, we assemble correlation-based networks for the corresponding time frame from 3h to 84h in both mouse strains.

We found that among the identified interactions of acinar homeostasis-related genes and intercellular signaling-related genes there is a significant enrichment for interaction pairs involving members of the Fgf/Fgfr family, specifically in wild-type mice. Particularly, we found Fgfr3 to be highly connected in the wild-type network, suggesting a potential role in the mediation of the inflammation-regeneration-transition. This finding is in line with recent studies which showed that Fgfr3, which is generally regarded as oncogene, can limit cell growth and promote differentiation if expressed in epithelial cells.

Zusammenfassung

Onkogene Kras-Mutationen sind ein kritisches Ereignis im Frühstadium der Pankreaskarzinogenese. Allerdings ist die Expression von mutiertem Kras allein nicht ausreichend, um die Gewebetransformation im adulten Organ zu induzieren. Stattdessen wird Inflammation als zusätzlicher Auslöser benötigt. Für unser Experiment verwenden wir etablierte Mausmodelle der akuten Pankreatitis und der frühen Pankreaskarzinogenese, basierend auf konsekutiven Caerulein-Injektionen in Wildtyp- und $Kras^{G12D}$-mutierten Mäusen. Wir führen Genexpressionsanalysen und histologische Analysen mit hoher zeitlicher Auflösung durch, um die ablaufenden Prozesse in beiden Mausmodellen zu vergleichen.

Basierend auf unseren ersten Ergebnissen beschreiben wir die pankreatische Regeneration als Prozess der aus drei klar abgrenzbaren Phasen hinsichtlich der Genexpression und der Morphologie besteht: Inflammation, Regeneration und Refinement. Unsere komparative Analyse zeigt, dass die unterbrochene Regeneration in $Kras^{G12D}$-mutierten Mäusen durch eine anhaltende Inflammation gekennzeichnet ist.

Zudem beobachten wir veränderte Proliferationskaskaden im Hinblick auf die Azinus-, Progenitor-, und mesenchymalen Zellen. Insbesondere wird eine massive Expansion des Mesenchyms beobachtet. Wir folgern daraus, dass diese veränderten Proliferationskaskaden, zusammen mit der anhaltenden Inflammation, die unverwechselbaren Kennzeichen des Frühstadiums der Pankreaskarzinogenese sind.

Wir extrahieren eine Signatur, welche die anhaltende Inflammation in $Kras^{G12D}$-mutierten Mäusen beschreibt und untersuchen einen Patienten-Datensatz, basierend auf 45 Proben von duktalen pankreatischen Adenokarzinomen (PDAC), hinsichtlich der Expression der homologen Gene der Signatur. Hier zeigt sich, dass die Patienten in zwei klar abgrenzbare Cluster unterteilbar sind, mit einer erhöhten durchschnittlichen Überlebenszeit im ersten Cluster (hohe Expression der Signatur), verglichen mit dem zweiten Cluster (niedrige Expression der Signatur).

Im Einklang mit unseren vorherigen Ergebnissen, definieren wir ein "duales Modell" der pankreatischen Regeneration und der frühen Pankreaskarzinogenese, welches aus vier Zuständen besteht: gesund, Inflammation, Regeneration und Refinement. Da die Divergenz der Wildtyp- und $Kras^{G12D}$-Profile mit der *Inflammation-Regeneration Transition* in Wildtyp-Mäusen beginnt, berechnen Korrelationsnetzwerke für das Zeitfenster 3h-84h auf beiden Datensätzen um potentielle Vermittler dieser Transition zu entdecken.

Wir sehen, dass an den für Wildtyp-Mäuse identifizierten Interaktionen zwischen Genen die zur Aufrechterhaltung des azinären Phänotyps benötigt werden und Genen die zu interzellulären Signalwegen gehören, besonders häufig Mitglieder der Fgf/Fgfr Familie beteiligt sind. Insbesondere ist Fgfr3 ein hochvernetzter Knoten im resultierenden Netzwerk, was nahelegt, dass dieses Gen eine potentiell wichtige Rolle bei der Vermittlung der Inflammation-Regeneration Transition in Wildtyp-Mäusen einnimmt. Dieses Ergebnis steht im Einklang mit neuesten Studien die gezeigt haben, dass Fgfr3, welches allgemein als Onkogen bekannt ist, Zellwachstum hemmen und Differenzierung fördern kann, wenn es in Epithelzellen exprimiert wird.

Acknowledgements

Firstly, I would like to gratefully thank Nikola Müller and Bo Kong for their support regarding research and organizational questions. Thank you for all your scientific advice and encouragement which helped me a lot for the completion of this thesis.

Furthermore, my greatest thanks go to Christoph Michalski and Fabian Theis for providing me with the opportunity to perform my doctoral work under their supervision. Thank you very much for your support, expertise, inspiration and patience.

I would further like to thank Jörg Kleeff for the regular scientific feedback. Thank you also for the constructive discussions regarding new interdisciplinary scientific approaches.

My great thanks go to Nora Behler, Ziying Jian, Sina Fritzsche, Natalya Valkovskaya, Weiwei Wu, for your excellent and intensive work in the lab and your expertise. Furthermore, I thank Manja Thorwirth and Nadja Maeritz for their excellent work in the lab.

I would like to thank Ivonne Regel, Susanne Raulefs, Dominik Lutter and Carsten Jäger for their advice and constructive support. Thanks also go to Daniela Stangl, Simone Hausmann, Martin Preusse and Andrea Ocone for the great working atmosphere.

Many thanks also goes to Anna Melissa Schlitter, Irene Esposito, Martin Irmler, Johannes Beckers, Peter Winter, Sören Müller, Günter Kahl, Rolf Thermann, Kurt Roth, Björn Rotter and Klaus Hoffmeier for our successful collaborations in the last years. It has been a pleasure to work with you.

For the help with organizational issues, I gratefully thank Katrin Offe, Elisabeth Noheimer, Marianne Antunes and Monika Hane.

Finally, I would like to thank Sebastian, my parents and my brother Rich for their permanent support and encouragement on the road of life.

Contents

Table of contents **IX**

List of figures **XIII**

List of tables **XV**

List of abbreviations **XVII**

1 Introduction **1**

1.1 Biology of the pancreas 1

1.1.1 Anatomy and physiology 1

1.1.2 Cell types of the mature pancreas 1

1.1.2.1 Exocrine cells 1

1.1.2.2 Endocrine cells 3

1.1.2.3 Mesenchymal cells 4

1.1.3 Pancreatic development in mice 4

1.1.3.1 Morphogenesis and signaling 4

1.1.3.2 Pancreatic progenitor cells 5

1.2 Pancreatitis 8

1.2.1 Acute and chronic pancreatitis: differences and common features 8

1.2.1.1 Causes and mechanisms of acute pancreatitis 8

1.2.1.2 Causes and mechanisms of chronic pancreatitis 9

1.2.2 Cellular processes and signaling 10

1.2.2.1 Immune response 10

1.2.2.2 Acinar-to-ductal transdifferentiation 11

1.2.3 Induction of pancreatitis in mouse models 11

1.3 Pancreatic ductal adenocarcinoma (PDAC) 12

1.3.1 Epidemiology and pathology of PDAC 12

1.3.1.1 Pathology . . . 12
1.3.1.2 Epidemiology . . . 13
1.4 Early pancreatic carcinogenesis and progression to PDAC . . . 14
1.4.1 The PanIN model for progression of precursor lesions to carcinoma . . . 16
1.4.1.1 Stepwise progression from PanIN-1A to PanIN-3 and PDAC . . . 16
1.4.1.2 Direct transition from normal pancreas to PanIN-2 or PanIN-3 . . . 16
1.4.2 Pathways activated in PDAC . . . 18
1.4.3 Cross-talk of epithelial and mesenchymal cells . . . 19
1.4.4 Epithelial-mesenchymal transition . . . 20
1.5 Other types of pancreatic tumors . . . 22

2 Motivation **25**
2.1 The $Kras^{G12D}$ model . . . 25
2.2 KRAS mutations in human PDAC . . . 26
2.3 Previous PDAC studies on Kras-mutated mice . . . 26

3 Material and methods **29**
3.1 Experimental design . . . 29
3.1.1 Treatment protocol . . . 29
3.1.2 RNA isolation . . . 30
3.1.3 Immunohistochemistry and immunofluorescence analysis . . . 30
3.1.3.1 Tissue extraction, H&E, IHC and IF protocol . . . 30
3.1.3.2 Antibodies . . . 31
3.2 Microarrays . . . 32
3.2.1 Transcriptional profiling with microarrays . . . 32
3.2.2 Affymetrix arrays . . . 32
3.2.3 Applied protocol . . . 33
3.3 Data preprocessing . . . 33
3.3.1 Quality control . . . 33
3.3.2 Normalization . . . 34
3.3.3 Probe filtering . . . 34
3.3.4 Application of preprocessing steps . . . 34
3.4 Data modeling and analysis . . . 35
3.4.1 Hierarchical clustering . . . 35
3.4.2 PCA and PARAFAC . . . 36

3.4.2.1 Singular value decomposition . . . 36
3.4.2.2 PCA . . . 36
3.4.2.3 Extension to tensors: PARAFAC . . . 37
3.4.2.4 Confidence ellipsoids . . . 37
3.4.3 Differential expression analysis . . . 38
3.4.3.1 t-test for detection of differential expression between two groups 39
3.4.3.2 F-test for more than two groups (ANOVA) . . . 39
3.4.3.3 Linear models for microarrays and RNA-Seq (Limma) . . . 40
3.4.4 Multiple testing correction . . . 41
3.4.4.1 False discovery rates . . . 41
3.4.4.2 Bonferroni correction . . . 41
3.4.5 Detection of differential expression of a gene set without predefined groups 41
3.4.5.1 Kolmogorov-Smirnov test . . . 42
3.4.5.2 Shapiro-Wilk test . . . 42
3.4.6 Enrichment analysis . . . 43
3.4.6.1 Hypergeometric test . . . 43
3.4.6.2 GO and KEGG enrichment . . . 44
3.4.6.3 Model-based gene set analysis for gene ontologies . . . 44
3.4.7 Cross-validation for gene signature extraction . . . 45
3.4.8 Identification of active subcircuits in regulatory networks . . . 46
3.4.9 k-means clustering . . . 48
3.4.10 Survival analysis . . . 49
3.4.10.1 Log-rank test . . . 50
3.4.10.2 Peto-Peto test . . . 50
3.4.11 Support vector machines . . . 50
3.4.12 Correlation networks . . . 52

4 Results **55**

4.1 Temporal course of acute pancreatitis . . . 55
4.1.1 Regeneration in wild-type mice is characterized by three stages: inflammation, regeneration and refinement . . . 55
4.1.1.1 Histological observations suggest three predefined stages . . . 55
4.1.1.2 Transcriptional profiling confirms three stages . . . 56
4.1.2 Temporal gradient unveils regeneration as a stepwise self-resolution process 58

4.2 Inflammatory response in Kras^{G12D} mice . . . 58
4.2.1 Histological observations suggest sustained inflammation . . . 58
4.2.2 Transcriptional profiling indicates disordered process . . . 60
4.2.3 PCA indicates end-to-open pattern . . . 60
4.3 Persistent inflammation in Kras^{G12D} mice . . . 60
4.4 Comparison of the identified time frames 64
4.4.1 Comparison of inflammatory phase samples . . . 65
4.4.2 Comparison of regeneration phase samples . . . 65
4.4.3 Comparison of refinement phase samples . . . 66
4.5 Alterations of the cellular programs . . . 69
4.5.1 Loss of exocrine maintenance characterizes inflammatory phase and early carcinogenesis . . . 69
4.5.2 Expansion of progenitor-like cells characterizes the inflammatory phase and early carcinogenesis . . . 72
4.5.3 Expansion of mesenchymal cells characterizes the inflammatory phase and early carcinogenesis . . . 76
4.6 Transfer to human PDAC dataset . . . 79
4.6.1 Extraction of an inflammation-specific signature . . . 79
4.6.2 Differential expression and survival analysis . . . 80
4.7 Identification of potential mediators . . . 84

5 Discussion **89**
5.1 Loss of coordination in cell proliferation 89
5.2 Transferability to pancreatic carcinogenesis in humans . . . 92
5.3 Mediators of the inflammation-regeneration transition . . . 93

Conclusion **97**

Bibliography **99**

List of figures

1.1 Anatomy of the pancreas, adapted from MediVisuals [6]. 2

1.2 Cellular organization of the pancreas, adapted from Yale University [7]. 3

1.3 Formation of pancreatic progenitors during embryonic development and specification of pancreatic cell lineages. 7

1.4 PanIN model for stepwise progression from normal pancreatic tissue to PDAC. Left: HE stainings (coloring all nuclei) of different progression stages; staining between PanIN-1A and PanIN-1B is an intermediate stage. Stainings adapted from [109]. Right: Illustration of acinar structures undergoing stepwise disruption. Sketch adapted from [110]. 17

1.5 The progressive stages of EMT. 21

2.1 Ras signaling. Adapted from [157]. 25

3.1 Time course of the experiment. Wild-type mice were sacrificed at 13, Kras^{G12D} mice were at nine different time points. 30

3.2 Two dendrograms visualizing the same clustering result. In the right part, leaves are visualized with a uniform distance. 36

3.3 Illustration of the cross-validation procedure. 46

4.1 Representative H&E-stained sections with histological features of the inflammation, regeneration and refinement phases in WT pancreata (scale bars: 50 μm). 56

4.2 Unsupervised clustering of transcriptional profiles in WT mice: homogeneous clusters correspond to the histologically predefined phases. 57

4.3 PCA biplot of transcriptional profiles with 95% confidence ellipsoids: the temporal course of inflammation in WT pancreata is a self-resolving process forming an end-to-end pattern (as underscored by arrows pointing to the centers of the groups' samples, right panel). 57

4.4 Representative H&E-stained sections with histological features of inflammation (3h - 36h), of the anticipated regeneration phase (48h - day 7) and of the anticipated refinement phase (day 7 to day 14; scale bars: 50 μm). 61

4.5 Clustering of Kras^{G12D} profiles confirms disordered regeneration as indicated by heterogeneous clusters. 62

4.6 PCA biplot confirms that regeneration is blocked and forms and end-to-open pattern in Kras^{G12D} mice. 62

4.7 Proliferation indices in WT and Kras^{G12D} pancreata in different phases (or anticipated phases) of inflammation, as obtained by quantifying the percentage of BrdU-positive cells and hematoxylin-counterstained nuclei. 63

4.8 Representative IHC pictures show BrdU-positive cells in the different phases (or anticipated phases) of inflammation in WT (left panel) and Kras^{G12D} pancreata (right panel; scale bars: 50μm). 63

4.9 Quantitative data shows the time-dependent changes in the number of Cd45-positive cells in the different phases (or anticipated phases) in WT and Kras^{G12D} pancreata. 64

4.10 Representative IHC pictures show Cd45-positive cells in the different phases (or anticipated phases) of inflammation in WT (left panel) and Kras^{G12D} pancreata (right panel; scale bars: 50 μm). 64

4.11 PARAFAC analysis: direct comparison of transcriptional profiles of WT and Kras^{G12D} pancreata on a genome-wide scale; 95% confidence ellipsoids reflect the previously defined phases; all samples from Kras^{G12D} mice (except for the controls) are highly similar to the inflammation phase samples from WT mice. . 65

4.12 Number and percentage of proliferating α-amylase-positive cells in WT and Kras^{G12D} pancreata. 70

4.13 Representative IF of proliferating α-amylase-positive cells in WT (upper panel) and Kras^{G12D} (lower panel) pancreata in different phases (or anticipated phases) of inflammation; scale bars: 50 μm. 70

4.14 Heatmap of 11 genes linked to acinar cell homeostasis as indicated on the right; voxel color: transcriptional up- (red) and down-regulation (blue) as compared to the mean; each voxel indicates gene activity in one mouse sample arranged in chronological order; colored bar indicates control (green) and regenerative phases (yellow, red and blue). . . 71

4.15 Absolute and relative amount of proliferating acinar cells in wild-type inflammation, regeneration and Kras^{G12D} samples from 3h to 36h. 71

4.16 Number and percentage of proliferating Sox9-positive cells in WT and Kras^{G12D} pancreata. 72

4.17 Representative IF of proliferating Sox9-positive cells in WT (upper panel) and Kras^{G12D} (lower panel) pancreata in different phases of inflammation (upper panel) and anticipated phases of Kras^{G12D}-mediated early carcinogenesis (lower panel); scale bars: 50 μm. 73

4.18 Bioinformatic analysis of the activity of the entire GRN of pancreatic organogenesis in different phases of inflammation (upper panel) and anticipated phases of Kras^{G12D}-mediated early carcinogenesis (lower panel). 74

4.19 Percentage of proliferating Sox9^{+} cells in wild-type inflammation, regeneration and Kras^{G12D} samples between 3h and 36h. . . . 75

4.20 Number and percentage of proliferating α-SMA-positive cells in WT and Kras^{G12D} pancreata. 76

4.21 Staining for α-SMA-positive cells in WT and Kras^{G12D} pancreata. 77

4.22 Heatmap of 8 genes linked to activation of PSC in vivo; voxel color: transcriptional up- (red) and down-regulation (blue) as compared to the mean; each voxel indicates gene activity in one mouse sample arranged in chronological order; colored bar: different phases or anticipated phases in WT and Kras^{G12D} samples, respectively. 77

4.23 Percentage of proliferating α-SMA^{+} cells in wild-type inflammation, regeneration and Kras^{G12D} samples between 3h and 36h. . . . 77

4.24 Workflow for selection of the "PSC activation signature", overlap of a literature-based gene set of PSC-specific genes and genes specifically up-regulated during the wild-type inflammation phase (compared to wild-type controls and regeneration). 78

4.25 Workflow for selection of inflammation-specific genes in wild-type mice after one-way ANOVA. 79

4.26 Clustered heatmap of the gene signature of the inflammation phase allows differentiation of early carcinogenesis from control samples. . 81

4.27 k-means clustering allows separation of human PDAC samples into two clusters with high (black cluster) and low (red cluster) expression of the inflammation signature (homologous genes). 82

4.28 Left: Distribution of signature expression scores in the two clusters (aggregated z-scores). Center: Sampling with random gene sets indicates significant differential expression of the identified signature (red bar, p=0.0012). Right: Survival analysis of the two clusters: longer median survival of patients with a high expression of the signature (24 months vs. 12 months). 83

4.29 SVM based on z_A score and clustering information allows classification of long-term and short-term survival patients with an accuracy of 0.86 (leave-out-one cross-validation). 83

4.30 The petri net describes our model of pancreatic regeneration and early pancreatic carcinogenesis, derived from the results presented in sections 4.1 and 4.2. 84

4.31 Interactions of epithelial genes and cell-cell signaling related genes in wild-type mice between 3 hours and 84 hours. Occurrence of 16 interactions involving Fgf/Fgfr-related genes suggests that corresponding pathways contribute to the transition from inflammation phase to regeneration phase. 85

4.32 Neighbourhood of the Fgf/Fgfr nodes in the wild-type network. 86

4.33 In Kras^{G12D} mice, 14 interaction pairs were found. 87

List of tables

1.1 Transcription factors contributing to the pancreatic cell fate during primary transition. . . 6

1.2 Transcription factors contributing to the pancreatic cell fate during secondary transition. . 7

3.1 Number of collected tissues samples and number of microarrays used at different time points. 31

4.1 Histological observations in wild-type and Kras^{G12D} mice. 59

4.2 Genes differentially expressed between wild-type and Kras^{G12D} inflammation. Due to high similarity of both stages, the number of genes found to be differentially expressed is small, and no GO terms or KEGG pathways were found to be enriched. 66

4.3 GO terms that were found to be enriched for genes differentially expressed between wild-type regeneration and Kras^{G12D} anticipated regeneration (mgsa). 67

4.4 KEGG pathways that were found to be enriched for genes differentially expressed between wild-type regeneration and Kras^{G12D} anticipated regeneration. 67

4.5 GO terms that were found to be enriched for genes differentially expressed between wild-type refinement and Kras^{G12D} anticipated refinement (mgsa). 67

4.6 KEGG pathways that were found to be enriched for genes differentially expressed between wild-type refinement and Kras^{G12D} anticipated refinement. 68

List of abbreviations

PDAC pancreatic ductal adenocarcinoma

ECM extracellular matrix

EMT epithelial-to-mesenchymal transition

HSC hepatic stellate cell

PSC pancreatic stellate cell

RA retinoic acid

ROS reactive oxygen species

MMP matrix metalloproteinases

ADM acinar-to-ductal metaplasia

DNA deoxyribonucleic acid

GTP guanosine-5'-triphosphate

GEF guanine nucleotide exhange factor

PanIN pancreatic intraepithelial neoplasia

LOH loss of heterozygosity

IPMN intraductal papillary mucinous neoplasms

MCN mucinous cystic neoplasms

ACC acinar cell carcinomas

PEN pancreatic endocrine neoplasms

SPN solid pseudopapillary neoplasms

MEF mouse embryonic fibroblasts

mRNA messenger ribonucleic acid

TF transcription factor

miRNA micro ribonucleic acid

GEMM genetically engineered mouse model

i.p. intraperitoneal

HE hematoxylin and eosin

IHC immunohistochemistry

IF immunouorescence

PFA paraformaldehyde

BSA bovine serum albumin

PM/MM perfect match/mismatch oligo

WT wild-type

RMA robust multi array normalization

SVD singular value decomposition

PCA principal component analysis

PARAFAC parallel factor analysis

ANOVA analysis of variance

BGV between-group variability

WGV within-group variability

FDR false discovery rate, q-value

GO Gene Ontology

KEGG Kyoto Encyclopedia of Genes and Genomes

MSigDB molecular signatures database

MCMC markov chain monte carlo

SVM support vector machine

GGM gaussian graphical model

TC tubular complex

AFL atypical flat lesion

GRN gene regulatory network

FC fold change

Chapter 1

Introduction

1.1 Biology of the pancreas

1.1.1 Anatomy and physiology

The pancreas is a glandular organ composed of an exocrine and an endocrine compartment. The exocrine compartment is a key part of the digestive system regulating the secretion of digestive enzymes. The endocrine compartment is essential for glucose homeostasis controlled by glucagon and insulin release.

The human pancreas measures 16-20 cm in length, 3-4 cm in width and 1-2 cm in height, weighing approximately 40-120 grams [1].

It is located in the retroperitoneal space between the stomach, the duodenum, the spleen and the liver [2]. The pancreatic duct, which has a width of approximately 2 mm transports the produced enzymes towards the small intestine and joins the common bile duct [3]. The union of both ducts forms the ampulla of Vater which is a connection to the duodenum.

On the macroscopic level, the pancreas is separated into three parts, the head (*Caput pancreatis*), the body (*Corpus pancreatis*) and the tail (*Cauda pancreatis*). The head is encompassed by the duodenum and forms the biggest part of the pancreas. The tail is located close to the spleen. An overview of the pancreatic anatomy is given in Fig. 1.1.

1.1.2 Cell types of the mature pancreas

1.1.2.1 Exocrine cells

The exocrine compartment of the pancreas is mainly structured in components called acini which are small clusters of acinar cells. Together with the ductal cells, also belonging to the exocrine compartment, this functional unit accounts for more than 90% of the whole pancreas [4]. Organization of the different cells within the pancreas is illustrated in Fig. 1.2.

- *Acinar cells* are responsible for synthesization, storage and secretion of inactive forms of digestive enzymes. They are characterized by a pyramidal shape, a basal surface, filamentous mitochondria, a prominent Golgi complex and regular arrays of rough endoplasmatic

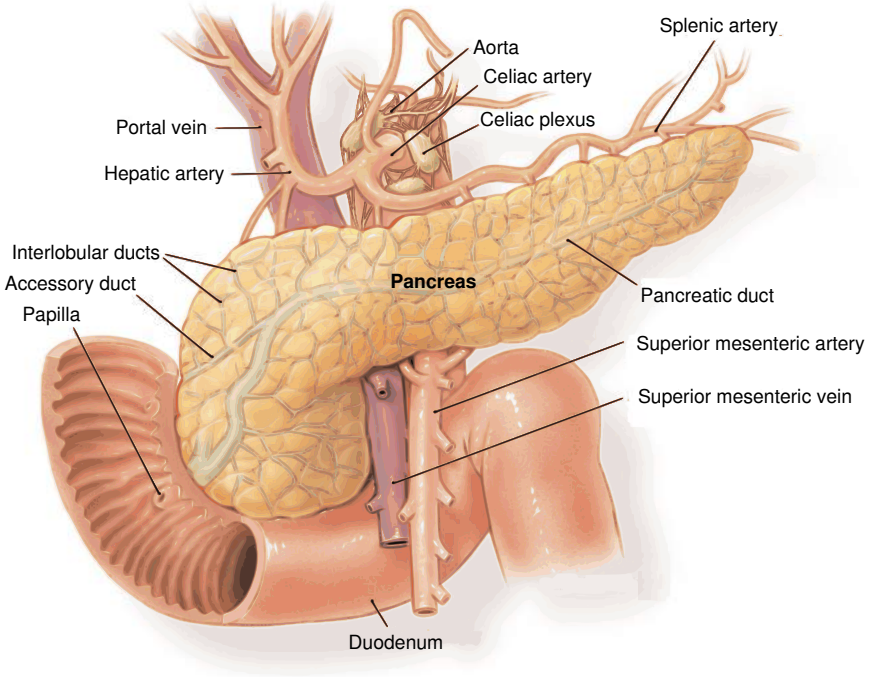

Figure 1.1: Anatomy of the pancreas, adapted from MediVisuals [6].

reticulum [5]. Additionally, they have zymogen granules, storing the different digestive enzymes. Secretion of pancreatic juice is regulated by gastrin, secretin and other hormones. One acinus is composed of around 40-50 acinar cells [5].

Both, development and maintenance of the acinar phenotype, are controlled by distinct molecular programs, in which *Pdx1* and *Mist1* are two key players. Alterations of these programs have impact on the cell fate or can lead to dedifferentiation or transdifferentiation of mature acinar cells to ductal cells (*acinar-to-ductal metaplasia*) [8, 9].

- *Ductal cells* are epithelial cells transporting the secreted enzymes to the gastrointestinal tract. Together they form a network of branched tubes. They arise from the same progenitor cells as pancreatic acinar cells but undergo a slightly different differentiation program as will be described in section 1.1.3.2. Previous studies have suggested that adult ductal cells are multipotent cells or even pancreatic stem cells being able to produce endocrine, insulin-producing β-cells [10]. This hypothesis was driven by the observation that endocrine cells arise from the Sox9$^+$ ductal domain. Nonetheless, recent *in vivo* studies did not succeed to produce endocrine cells from adult ductal cells. On the contrary, they suggest that it is impossible [11]. Ductal cells have been shown to share morphological similarities with tumor cells from pancreatic ductal adenocarcinoma (PDAC).

  They are characterized by a cuboidal to columnar shape, a perinuclear Golgi complex,

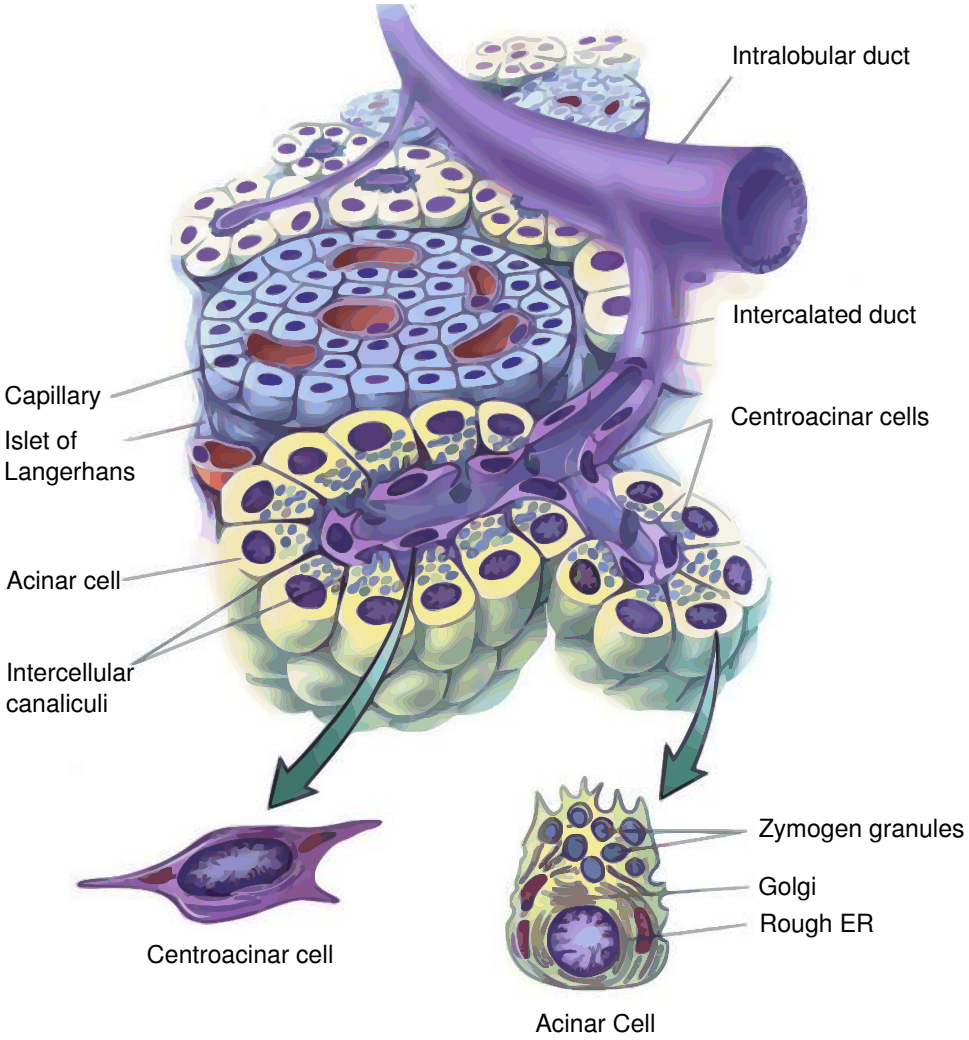

Figure 1.2: Cellular organization of the pancreas, adapted from Yale University [7].

scattered or apically localized cytoplasmic vesicles, and a large number of apical plasma membrane mircovilli [12].

The junction between the terminal ductal epithelium and the acini is formed by so-called centroacinar cells; yet it has not been proven whether these cells form another distinct cell type or are functionally equivalent to ductal cells.

1.1.2.2 Endocrine cells

Functions of the endocrine pancreas include glucose homeostasis by controlled hormone release as well as the regulation of metabolism. The endocrine compartment is organized in so-called islets of Langerhans which are cell clusters that are integrated in the exocrine pancreas tissue [13]. Different cell types with distinct functions are involved in hormone level maintenance.

- *β-cells* are responsible for insulin and amylin release, amylin being an insulin-antagonist. They account for more than 60% of all cells in the Langerhans islets.

- *α-cells* accounting for around 20% of the islet cells are responsible for synthesization and secretion of glucagon.
- *δ-cells* are responsible for somatostatin production and release. They account for 3-10% of the islet cells.
- *ϵ-cells* are producing and secreting ghrelin. They account for less than 1% of the islet cells.
- *PP-cells* sometimes also referred as γ-cells, are responsible for pancreatic polypeptide production and account for 3-5% of the islet cells. They are characterized by a polygonal shape.

1.1.2.3 Mesenchymal cells

Apart from the above cells which are directly linked to a functional task of the pancreas, mesenchymal cells are part of the pancreas located in the extracellular matrix (ECM) of the organ. In contrast to the epithelial cells or endocrine cells which are connected via junctions or organized in islets, respectively, the non-polar mesenchymal cells (stroma cells) are not directly connected to each other but loosely organized in the ECM instead. Mesenchymal cells are known to impact β-cell production during embryonic development [14]. Additionally, they play an important role during pancreatic carcinogenesis as epithelial cells may undergo a transition to mesenchymal cells during early tissue transformation, leading to a massive expansion of the pancreatic mesenchyme. This process is called *epithelial-mesenchymal transition* (EMT) and will be discussed detailed in section 1.4.4. It has been shown that EMT and corresponding mesenchymal expansion are contributing to pancreatic cancer metastasis and treatment resistance [15]. In this context, identification of putative EMT-targeting agents is of great interest [15].

The major mesenchymal cell in the pancreas is the *pancreatic stellate cell* (PSC) accounting for approximately 4% of all cells in the pancreas [16, 17]. Morphologically, PSCs have high similarities with hepatic stellate cells (HSCs). Like HSCs, PSCs are able to switch between a quiescent and an activated phenotype. After inflammatory injury, PSCs are activated and become myofibroblast-like cells which are commonly identified by α-SMA expression. In this activated state PSCs begin to proliferate, migrate and produce ECM components. In addition, they start to secrete pro-inflammatory cytokines, chemokines as matrix metalloproteinases and their inhibitors [16]. In case of a prolonged inflammation (i.e. chronic pancreatitis), prolonged PSC activation causes pancreatic fibrosis. It was shown that the release of the above-mentioned molecules promotes acinar cells to undergo EMT [18]. Accordingly, PSCs play a key role for mesenchymal expansion and tissue transformation during pancreatic carcinogenesis.

1.1.3 Pancreatic development in mice

1.1.3.1 Morphogenesis and signaling

As pancreatic morphogenesis was best studied in mice, all descriptions in this section refer to mice. The pancreas arises from the primitive gut tube, derived from the definitive endoderm. At

E9.5, a thickening on the dorsal side of the foregut epithelium indicates morphological formation of the pancreas. Nearly at the same time, two ventral buds arise in the epithelium [19]. During this early period of development, as a result of gut rotation as well as elongation of the dorsal and the ventral stalk, dorsal and ventral buds contact each other, forming a C-loop of the duodenal anlage. This fusion occurs between E12 and E13. Additionally the epithelial compartment branches into the surrounding mesenchyme where multipotent progenitor cells are able to form all lineages discussed before [19]. The commitment which type of cell a given progenitor cell will become between E11 and E14 depends on the location; cells at the tips of the branching epithelium will become pre-acinar $Ngn3^{-}$ cells, other cells will become bipotent duct/endocrine $Ngn3^{+}$ cells and populate the branches. At E15.5, all cells are specified and undergo growth and maturation until birth [19].

Derivation of the pancreas from the definitive endoderm is accompanied by inductive signals of the surrounding mesenchyme; current cancer-related research studies are focusing on these signals, as similar processes might be promoted during pancreatic carcinogenesis [20, 21].

Cell fate of pancreatic and liver progenitor cells are induced by Wnt and Fgf4 suppression in the foregut. Positioning of liver and pancreas is mediated by retinoic acid (RA) signaling. Fgf and Bmp signals from the mesenchyme induce liver formation and suppress pancreas formation [22]. Afterwards, during foregut closure, lateral ventral endoderm cells initiate pancreatic development by inactivation of these signals [23].

1.1.3.2 Pancreatic progenitor cells

Several genes are involved in specification of multipotent pancreatic progenitor cells which then give rise to differentiation to the above described cell types. Pancreatic endoderm formation is characterized by expression of two transcription factors: pancreas and duodenal homeobox gene-1 (Pdx1) and pancreatic transcription factor-1 (Ptf1a/p48) [24].

Previous studies have shown that $Pdx1^{+}$ cells contribute to all pancreatic cell fates whereas inactivation of Pdx1 completely suppresses pancreas organogenesis [25]. In the adult organ however, Pdx1 expression is low in exocrine cells but high in islet cells, in particular β- and δ-cells. $Pdx1^{+}$ expression starts at E8.5 during embryonic development in mice and is an initiating force of pancreas formation.

Ptf1a (p48) expression is required for evagination of the ventral bud as well as growth of the dorsal bud and first expressed at E9.5. Pancreas formation entirely fails when Ptf1a expression is suppressed [26]. In the adult state, Ptf1a expression is exocrine-specific and part of the acinar homeostatic program [27].

Apart from the two key factors, other transcription factors contribute to the pancreatic cell fate, as described in table 1.1.

Table 1.1: Transcription factors contributing to the pancreatic cell fate during primary transition.

Factor	Function
Pdx1	generally required for pancreatic cell fate; all pancreatic lineages are Pdx1$^+$ [25]
Ptf1a (p48)	evagination of ventral bud; growth of dorsal bud [28]
Sox17	first up-regulated, promotes pancreatic and biliary primordia; then required to be down-regulated for pancreas formation [29]
Hhex	required for correct positioning of pancreatic progenitor cells; required for Pdx1 and Ptf1a induction [20]
Hlxb9	dorsal bud formation; promotes Pdx1 expression [30]
Hnf6 (Onecut1)	required for pancreas specification; endocrine progenitor formation; duct cell development [31, 32]
Foxa1 (Hnf3a)	specification of pre-pancreatic endoderm [33]
Foxa2 (Hnf3b)	specification of pre-pancreatic endoderm; maturation of endocrine cells [33, 34]
Hnf1b (vHNF, TCF2)	required for dorsal and ventral bud formation; expressed in a subpopulation of pancreatic multipotent cells [35, 36]
Gata4	required for epithelial cell formation and ventral pancreas formation [37]
Gata6	required for epithelial cell formation and ventral pancreas formation [37]
Sox9	indicator for multipotent progenitor cells; required for pancreatic development; control of endocrine progenitor cell formation [38]
Hes1	maintenance of undifferentiated state in progenitor cells [39]
Myc	expressed in multipotent progenitor cells; inactivation leads to switch of cell production during development [40]
Rbpj	required for progenitor expansion; suppression of acinar maturation [41]

The above described formation of pancreatic progenitors is termed *primary transition*. This process terminates around E14 and is followed by the *secondary transition* which includes depletion of previously formed progenitors by specification to the distinct pancreatic lineages, i.e. mature exocrine and endocrine cells. Accordingly, the secondary transition is accompa-

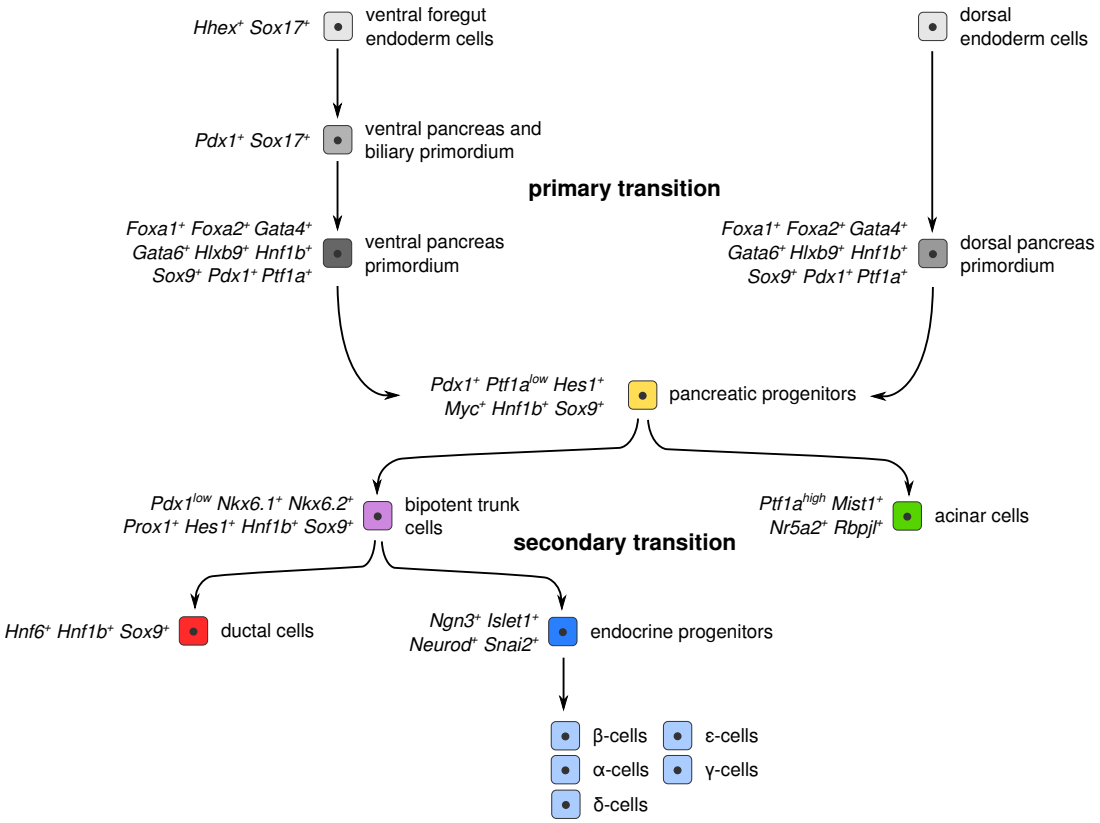

Figure 1.3: Formation of pancreatic progenitors during embryonic development and specification of pancreatic cell lineages.

nied by increase of digestive enzymes and endocrine hormones. During this process, pancreatic progenitors first undergo differentiation to pre-acinar cells and bipotent trunk cells which can differentiate into ductal or endocrine cells depending on Ngn3 expression.

Table 1.2 provides an overview of factors involved in lineage specification. Additionally, Fig. 1.3 gives a schematic overview of both primary and secondary transition with relevant intermediate steps.

Table 1.2: Transcription factors contributing to the pancreatic cell fate during secondary transition.

Factor	Function
Ptf1a	specification of the acinar phenotype together with Nr5a2; promotes induction of acinar program by activation of Rbpjl and Mist1 (Bhlha15) [42]
Nkx6.1, Nkx6.2	antagonists of Ptf1a activity mediating formation of bipotent trunk cells [44]
Prox1	suppressor of acinar differentiation [45]
Hes1	Notch target leading to Ptf1a inhibiton and thus blocks acinar differentiation [39]

Ngn3 (Neurog3)	commits cells to endocrine fate [46]
Hnf6 (Onecut1)	subsequently active in exocrine cells, low expression in endocrine cells; seems to be required for duct cell formation [32]
Hnf1b (vHNF, TCF2)	expressed in centroacinar and ducatal cells [47]
MafA, MafB	required for α- and β-cell maturation [48]
Neurod, Isl1, Pax genes Nkx, Arx, Irx genes	maintenance of balance of different islet cell types [49]

1.2 Pancreatitis

Pancreatitis is a disease characterized by inflammation of the pancreas. An *acute pancreatitis* occurs suddenly and lasts for several days when properly treated, a *chronic pancreatitis*, on the contrary, is a long-lasting disease accompanied by years of severe complications. Nonetheless, a clear distinction of both forms is not entirely possible as there are intermediate forms suggesting that there is a continuous range of pancreatitis intensity and that the established clinical terms are two extrema on the scale of this range. Also, it is known that the repeated occurrence of acute pancreatitis can lead to chronic pancreatitis.

1.2.1 Acute and chronic pancreatitis: differences and common features

1.2.1.1 Causes and mechanisms of acute pancreatitis

Gallstones and alcohol abuse are the most common causes of acute pancreatitis [50]. Other possible causes include infections (e.g. hepatitis), metabolic diseases, antibiotics, stomach or duodenal cancer. Additionally, smoking is asummed to be a risk-increasing factor [50].

The inflammatory process involves peri-pancreatic tissues (sometimes also surrounding tissues) and leads to acinar cell necrosis accompanied by edema and (pseudo-)cysts formation. Mild forms of the acute pancreatitis affect the pancreas itself only, whereas more severe forms can affect other organs and lead to death of the patient. The overall morality rate of all forms of acute pancreatitis is approximately 5% [51]. However, an acute pancreatitis is a disease which is usually reversible.

The early stage of acute pancreatitis is characterized by protease activation, secretion alterations and accumulation of inflammatory molecules in the acinar cell compartments. Under physiologic conditions, enzyme secretion is limited to the duct lumen by passing the apical membrane. During acute pancreatitis, however, noxious stimuli prevent exocytosis of zymogen granules (containing enzyme precursors) at the apical membrane, leading to an exocytosis at the basolateral membrane instead. Combined with an increased zymogen granule production, this can induce auto-digestion of the pancreas, explaining the pathogenesis of the disease [52]. Several molecular mechanisms, including trypsinogen activation, are thought to mediate these

inflammation-driven changes [53]. Up to a certain level, these abnormalities can be compensated with antagonizing effects including trypsin autoproteolysis, SPINK1 production, trypsin-degrading enzyme production and bicarbonate secretion. However, after the critical threshold has been passed, auto-digestion cannot be avoided any more.

1.2.1.2 Causes and mechanisms of chronic pancreatitis

In contrast to acute pancreatitis, chronic pancreatitis implies a persistent damage to the exocrine and the endocine part of the pancreas. In many cases, impaired digestion ability and diabetes include two of the severe causalities. Further complications include (pseudo-) cyst formation, pancreatic stones, exocrine insufficiency and bilary/duodenal stenosis [54]. Patients with chronic pancreatitis have an increased risk of pancreatic cancer development. Unspecific abdominal pain is the most common symptom of chronic pancreatitis. Incidence of chronic pancreatitis ranges up to 10 per 100.000 population in industrialized countries [55].

Alcohol abuse is the most common cause (70%) of chronic pancreatitis. Ethanol oxidation results in accumulation of acetaldehyde and production of ROS, the first one being toxic for pancreatic cells and the latter one being generally harmful for most types of cells. Non-oxidative metabolization of ethanol leads to fatty acid ethyl ester production which also leads to damage of pancreatic cells. It is commonly assumed that the risk of developing an alcohol-induced chronic pancreatitis increases logarithmically with the dose of daily consumed alcohol, with 40 grams/day (approx. 1 liter of beer, 5% vol.) being the critical dose for women and 80 grams/day (approx. 2 liters of beer, 5% vol.) being the critical dose for men, given an exposure time of 5 to 15 years [56].

However, chronic pancreatitis is also a hereditary disease. 10% to 30% of all cases of chronic pancreatitis are assumed to be hereditary. One common mutation associated with hereditary chronic pancreatitis is PRSS1^{R122H} which enhances the auto-activation of cationic trypsinogen (PRSS1) [57]. This mutation has been studied in transgenic mouse models which show an enhanced response after inflammatory injury. Another well-studied mutation is PRSS2^{G191R} affecting the anionic trypsinogen gene. In contrast to the PRSS1 mutation, this mutation shows a protective function [57]. Furthermore, mutations of the chymotrypsin C gene have been shown to impair pancretic function due to impaired activity or reduced levels of secretion [58]. Mutations of SPINK1 (serine protease inhibitor Kazal type 1) which is an inhibitor of trypsin activity in the pancreas, were also documented in cases of hereditary pancreatitis. Mouse models confirmed the role of SPINK1 in chronic pancreatitis: While overexpression led to a protection against pancreatitis, deletion of the gene caused a complete degeneration of acinar structurs within 14 days [59]. These outcomes suggest that also mutations of other trypsin- or trypsinogen-regulating genes impact the occurrence and course of chronic pancreatitis. Apart from these mutations, CFTR, a regulator of bicarbonate and chloride secretion has been reported to be mutated in chronic pancreatitis patients.

Tropical pancreatitis is another form of chronic pancreatitis which is, in accordance with its name, reported in regions with tropical climate (e.g. India). Viral and parasitic infections

as well as autoimmune mechanisms, malnutrition, oxidative stress and other factors are known causes of tropical pancreatitis [62]. The disease was also reported to coincide with mutations of SPINK1, Cathepsin B and Chymotrypsin C [61].

Finally, autoimmune pancreatitis forms a rare subtype of the disease, often coinciding with other autoimmune diseases. It is mainly characterized by increased serum γ-globulin levels as well as IgG4 autoantibodies and lymphocytic infiltration [60].

Regardless of the subtype, the clinical picture of chronic pancreatitis is homogeneous: acinar atrophy, fibrosis, necro-inflammation. Clinical observations suggest that the described transformation of the pancreas results from multiple, silently occuring cascades of pancreatic inflammation ("invisible forms of acute pancreatitis") [57]. This hypothesis was termed "necrosis-fibrosis model" and is currently accepted as the most accurate model of chronic pancreatitis. It has been discovered that pancreatic stellate cells (PSCs) play an important role during the observed fibrogenic process.

1.2.2 Cellular processes and signaling

1.2.2.1 Immune response

After inflammatory injury, damaged acinar cells themselves produce inflammatory mediators such as cytokines and chemokines which then recruit inflammatory cells, including neutrophils and macrophages. Activation of these cells in turn causes further acinar cell damage, leading to an increase of further pro-inflammatory mediators including TNFα (tumor necrosis factor alpha), interleukines 1, and 6, ICAM-1 (intercellular adhesion molecule 1), CD40L (cluster of differentiation 40, ligand), MCP-1 (monocyte chemotactic protein 1) and others [63]. Endothelial cell activation enables leucocyte migration leading to release of further harmful enzymes. Decreased oxygen supply in the pancreas causes additional injury.

Inflammatory response is assumed to be mediated by NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells), MAPK (mitogen-activated protein kinases), STAT3 (signal transducer and activator of transcription 3), AP-1 (activator protein 1) and PI3K (phosphatidylinositol-4,5-bisphosphate 3-kinase) signaling as these pathways were found to show an increased activity after inflammatory injury in the pancreas [51].

Apart from inflammatory cell recruitment, cytokine and chemokine production in the damaged acinar cells also induces activation of stellate cells (PSCs). Being in a quiescent state under normal conditions, PSCs transform into myofibroblast-like, α-SMA (alpha smooth muscle actin) expressing cells under the influence of chemokines and cytokines. Activated PSCs then release additional cytokines and chemokines, recruit further immune cells and produce extracellular matrix (ECM) components as well as matrix metalloproteinases (MMPs) and their inhibitors [16]. Activated PSCs can finally revert to their quiescent state or undergo apoptosis. Prolonged activation of these cells causes pancreatic fibrosis by ECM transformation [57]. Accordingly, there is special interest in understanding the role of PSCs during pancreatic carcinogenesis with respect to mesenchyme expansion and tumor-stroma interactions. Yet, it has not been clarified whether

these interactions may have a protective influence for the tissue in the tumor environment or whether they worsen the survival prognosis of PDAC patients [64].

1.2.2.2 Acinar-to-ductal transdifferentiation

Previously described acinar cell fate is not only determined but also maintained by a distinct molecular program. The homeostatic program incorporates Mist1 (Bhlha15) and Pdx1 expression. Studies with mouse models unveiled Pdx1 and Notch signaling being reactivated after inflammatory injury [11], recapitulating elements of embryonic development. Additionally, EGFR (epidermal growth factor receptor), Numb (protein numb homolog) and hedgehog signaling was shown to be active.

Activation of these signals is considered the most likely explanation for the observation of $Sox9^{+}$-cell expansion after inflammatory injury. It is assumed that due to disturbance of the homeostatic program, acinar cells dedifferentiate into a premature, progenitor-like state which then gives rise to ductal cell formation, a process termed *acinar-to-ducatal transdifferentiation*. Subsequently, transdifferentiation leads to formation of *acinar-to-ductal metaplasia* (ADMs) which in turn are able to progress to pancreatic tumors.

1.2.3 Induction of pancreatitis in mouse models

In order to study diversity of molecular signals mediating the inflammatory process and pancreatic regeneration, animal models were established for both acute and chronic pancreatitis. Non-surgical animal models induce pancreatitis by nutrition, e.g. ethanol or by injection of toxic substances or caerulein. Surgical models, also termed invasive models, are based on direct manipulation of the pancreatic duct [66].

Yet, none of the established animal models is sufficient for full representation of human pancreatitis. Accordingly, results from animal model-based experiments should be treated with caution.

One well-established animal model for acute pancreatitis is based on subcutaneous injections of caerulein in mice or rats [69]. Advantages of this model include easy application, low costs and high reproducibility.

Caerulein, an oligopeptide with a composition similar to cholecystokinin, increases digestive activity and stimulates smooth muscle [69]. After subcutaneous injection, it causes intracellular upregulation of NF-κB in the pancreas which in turn leads to ICAM-1 (acinar cell intercellular adhesion molecule-1) upregulation. Acinar ICAM-1 expression then increases ability of other cells to adhere; subsequent adhesion of neutrophils then promotes inflammation induction as described in section 1.2.2.1 [67]. As caerulein causes also dysregulation of digestive activity, acinar cells are destroyed and edema are formed which further enhances the inflammatory process. In addition, Jak signaling and NADPH oxidase (member of ROS family) are activated [68].

Most common protocols are based on consecutive caerulein injections with distinct time intervals.

1.3 Pancreatic ductal adenocarcinoma (PDAC)

Accounting for more than 85% of all cases, *pancreatic ductal adenocarcinoma* (PDAC) is the most frequently occurring type of tumor in the pancreas [70]. Ranking fifth (Europe) and fourth (USA) in the statistics of cancer-related death cases, PDAC is a major disease problem, though it accounts for less then 3% of all reported cancer-related cases. Accordingly, PDAC is one of the tumors with the worst survival prognosis with a median survival of less than six months and a five-year survival rate of less than 5%.

Though comprehensive, cost-intensive research studies have helped gaining detailed insights into the mechanisms of pancreatic carcinogenesis, essential processes and particularly targetable molecules remain still unknown. Mostly regardless of previous scientific and medical efforts, survival rates did not improve over the past five decades [71]. One reason for the poor prognosis is that PDAC is typically diagnosed at a late stage, also due to its unspecific symptoms.

1.3.1 Epidemiology and pathology of PDAC

1.3.1.1 Pathology

Typical symptoms of PDAC incorporate deep upper abdominal pain, dull, anorexia, asthenia, skin yellowing, depression and weight loss [43, 77]. Most of the tumors arise from head or neck of the pancreas (∼78%) [43]. PDAC causes obstructive cholestasis of the main duct which in turn can cause chronic pancreatitis and dysglycemia [72, 73]. In many cases, PDAC is also manifested by occurrence of diabetes as a causality of the tumor. Distant metastases as well as local invasion are common features of PDAC, rendering it one of the most aggressive tumor types in general. Most common locations of PDAC metastases include liver, lung, peritoneum and lymph nodes [74, 75]. Resected tumor tissue samples often show perineural, vascular and lymphatic invasion. Apart from PDAC as the most common pancreatic tumor type, other tumor types are documented, including colloid, adenosquamous and sarcomatoid tumors. Microscopically, PDAC is characterized by its infiltrating gland-forming neoplastic epithelium [76]. Additionally, it is constituted of dense stroma formed by a mixture of fibroblastic, endothelial and inflammatory cells. The expansion of the stroma is an important characteristic of PDAC compared to other tumor entities [77]; yet it has neither been proven nor confuted whether it contributes to the tumor's aggressiveness or not [78, 79]. Studies indicate that the so-called *desmoplasia* around the tumor, formed by PSCs, may blunt effective intratumoral drug delivery on the one hand [80]. On the other hand it was shown that targeting the stroma can result in undifferentiated aggressive pancreatic tumors, suggesting a protective role of the stroma [81]. According to the stromal character of PDAC, cytokeratines and mucins are commonly upregulated in the tumor tissue [82].

Depending on the stage of progression of PDAC, the disease can be treated surgically (tumor removal) or by appliance of chemotherapy or radiotherapy; also, a consecutive combination of these methods is possible. If progressed too far, palliative care is applied.

In most cases (around 80%), surgical treatment is not possible due to metastasis and tumor invasiveness [83]. If possible however, surgery is generally followed by adjuvant treatment. In Europe, chemotherapeutic treatment with gemcitabin or 5-fluorouracil is the current standard of care, while in the USA radiotherapy is also considered [84]. For this group of patients, which can be regarded as "early stage group", average survival time is 18-20 months with a five-year survival rate of 10%; in cases of complete tumor resection, five-year survival rate can raise up to 25% if the tumor has not yet spread to the lymph nodes [75]. For patients with an advanced stage of PDAC progression however, the average expected survival time is as low as 6 months [75].

Recent progresses in chemotherapy have been made with Folfirinox (folinic acid, fluorouracil, irinotecan and oxaliplatin) treatment, achieving an average survival benefit of 4 months which is the best improval documented yet [85]. It thus exceeds established treatment methods using gemcitabine or fluorouracil only. Folinic acid (Fol) is used in order to reduce fluorouracil side effects, F (fluorouracil) itself blocks DNA synthesis, Irin (irinotecan) prevents DNA from uncoiling and Ox (oxaliplatin) inhibits DNA repair/synthesis. Folfirinox has become one of the active chemotherapeutic regimens [85]. Clinical studies for identification of new regimens are focussing on TGFβ, hedgehog, Notch and VEGF/VEGFR inhibition [86].

1.3.1.2 Epidemiology

The median age of occurrence of PDAC is 70 years [87]. The risk of developing pancreatic cancer once in a life is around 1% for both men and women. Risk factors for pancreatic carcinogenesis are associated with genetic and environmental factors including personal life-style. Smoking has been documented to cause a 2-fold risk increase for developing PDAC and is assumed to be responsible for 20%-30% of all cases [88]. Alcohol abuse is an indirect factor as it increases the risk for developing chronic pancreatitis which in turn increases PDAC risk. Obesity is an additional risk-increasing factor. Furthermore, previous diseases including gastric ulcers, Chron's disease and diabetes increase the risk of developing PDAC. Correlations with nutrition style or drug abuse have not been identified [87].

Apart from these individual life-style preferences and personal disease history it has been shown that PDAC incidence correlates with latitude being higher in northern countries and lower in southern countries. An unproven assumption is that decreased vitamin D production due to decreased sunlight exposure has negative impact on the risk of developing PDAC [87]. Studies have shown that there is also a race-dependent factor with an increased risk for the black population and lowest risk for the Asian population [87].

However, genetic susceptibility appears to have the greatest impact on the personal PDAC risk. Persons with first-degree relatives that had PDAC have an up to 4-fold increased risk which can further increase up to 57-fold if more than 3 relatives are affected with PDAC [89]. This can be explained by the fact that several mutations have great impact on PDAC development when coinciding with a triggering factor like inflammation.

The affected genes mainly comprise functional players required for cell homeostasis or growth

signal transduction in the pancreas, respectively. Previously discussed PRSS1 mutation increases PDAC risk as well as alterations of the cystic fibrosis gene (CFTR) [90, 91]. Also, several germline mutations were reported to increase the risk, including BRCA1/2, FAMMM, p16, p53 or DNA repair gene mutations including ATM and PALB2 [92]. These hereditary factors account for ∼20% of familiar pancreatic cancer cases. Also NR5A2 (liver receptor homolog-1), required for acinar differentiation, was reported as a relevant factor [93].

1.4 Early pancreatic carcinogenesis and progression to PDAC

Apart from hereditary mutations, spontaneous genetic alterations can accompany or give rise to PDAC: Generally, these specific mutations can act as a "switch" dramatically increasing the risk of developing PDAC, though tissue transformation requires an additional trigger; or vice versa, they can occur as a consequence of already initiated tissue transformation and then promote further tumor growth, metastasation and invasion. Major genes whose abnormalities can act as this kind of "switch" include KRAS, p16, p53 and SMAD4 [76].

KRAS (Kirsten rat sarcoma viral oncogene homolog) is an oncogene that is found to be mutated in more than 90% of all PDAC patients. It is commonly assumed to be an initiating force for PDAC development as it is also mutated in many precursor lesions [76]. As an important growth-mediating factor it is not only associated with initiation but also progression.

Being a GTPase, KRAS normally works as a switch (not to be confused with the above mentioned genetic "switch") between an on- and an off-state. When bound to GTP, KRAS transduces incoming growth signals causing activation of the RAF/MEK/ERK and PI3K/PDK1/AKT cascade. Due to its intrinsic enzymatic activity (GTPase characteristic) it cleaves the terminal phosphate and thereby automatically changes into the off-state. It can then be reactivated by SOS1 or other guanine nucleotide exchange factors (GEFs) [94].

When mutated, e.g. frequently occuring $KRAS^{G12D}$, the ability to hydolyze GTP is lost and KRAS remains in the on-state; accordingly, transient growth signal changes into a persistent growth signal.

Both major downstream cascades of KRAS, RAF and PI3K play an important role during tumorigenesis. PI3K/AKT signaling is required for cell cycle maintenance and proliferation by differentiation of adult stem cells and thus promotes uncontrolled proliferation while at the same time preventing apoptosis. Amplified AKT2 was found in different tumor entities without KRAS mutations [95]. Loss of Pten, a natural inhibitor of PI3K signaling, also promotes uncontrolled proliferation in mouse models [96]. The RAF/MEK/ERK signaling cascade is required for further growth signal transduction and, accordingly, can also cause uncontrolled proliferation mutated. Mutations of B-RAF were found in several pancreatic tumors that did not show KRAS mutations [97].

p16 is a tumor suppressor which is commonly deactivated in PDAC cells [76]. As it is also found to be deactivated in precursor lesions it is regarded as an early genetic event contributing to further malignant transformation and uncontrolled growth. However, p16 is generally assumed to

be mutated after the KRAS alteration occurred [98, 99]. As a cyclin-dependent kinase inhibitor (CDK), p16 inhibits cell cycle progression between G1 and S phase, leading to a decreased proliferation. p16 prevents proliferation by binding to CDK4/6 which then cannot bind to cyclin D and phosphorylate pRB anymore [100]. This phosphorylation normally allows pRB to dissociate from transcription factor E2F1, which then can migrate from the cytoplasm to the nucleus where it initiates transcription of genes relevant for the transition from G1 to S phase [100].

Loss of p53 tumor suppressor genes is observed in most PDAC patients (around 85%). However, this mutation is usually observed after p16 and KRAS mutations, i.e. in progressed lesions or in the tumor stage. As p16, p53 is able to arrest growth by targeting the G1/S regulation point of the cell cycle by binding to CDK4/6, CDK2 and CDK1 and inducing cell cycle inhibitors. Accordingly, it can also induce apoptosis. Usually, p53 is in an inactive state, being bound to MDM2. After cell damage, dissociation from MDM2 activates p53 to allow for cell repair. Depending on the extent of damage, the p53 then either arrests the cell cycle as described in order to induce DNA repair or directly induces apoptosis. The exact decision mechanism remains still unknown. When p53 is mutated, DNA-damaged cells are able to reproduce their genetic abnormalities, leading to an accumulation of the deficiencies.

SMAD4 mutations are another critical genetic event in early pancreatic carcinogenesis. The gene is found to be mutated in 50% of PDAC patients [101]. Comparable to p53, it is mostly found to be mutated in progressed precursor lesions and tumors, indicating that SMAD4 alterations are likely to be the consequence of prior genetic events as KRAS mutations. SMAD4 is a member of the SMAD family. It dimerizes, trimerizes or hexamerizes together with other SMAD members (e.g. SMAD1-SMAD2-SMAD4 or SMAD3-SMAD4), forming different transcription factors. Those relatively unspecific factors then contribute to regulation of many genes contributing to functions as differentiation, apoptosis, cell cycle and, importantly, embryonic development. As a mediator of the TGFβ pathway, SMAD4 mutations are a double-edged sword: During early stages of pancreatic carcinogenesis, TGFβ signaling malfunction can contribute to uncontrolled growth, as TGFβ signals mediate epithelial growth and survival inhibition and apoptotic signals. However, also proliferation and migration *epithelial-mesenchymal* (EMT) signals are transduced by TGFβ activity; accordingly, SMAD4 mutations can prevent further tumor growth in advanced stages of PDAC formation [102, 103].

Apart from the discussed mutations, dysregulation of mucins (heavily glycosylated high molecular weight glycoproteins), is another characteristic of early pancreatic carcinogenesis. Mucins are gel-forming proteins functioning as chemical barrier and lubricators. In early pancreatic precursor lesions, MUC1 and MUC6 were shown to be strongly upregulated. In advanced PDAC, MUC1 and MUC4 showed a strong upregulation [104]. Recent studies are considering mucins to be potential drug targets for PDAC treatment [105].

An appropriate quantifier of PDAC progression is S100P, a calcium-binding protein. Expression of this protein has been reported to increase stepwise with progression of precursor lesions to tumors [106]. It has also been shown to promote tumor growth [104]. However, though it is

an accurate indicator of PDAC, it appears to be drug resistant.

A non-genetic characteristic of early pancreatic carcinogenesis is telomere shortening. Telomeres are composed of repetetive sequences and bind with shelterins in order to avoid truncation of functionally relevant genes during cell division. Usually, in case of critically short telomeres, DNA damage response is activated by p53. As most cells in PDAC are p53-deficient, corresponding damage response will be absent here. However, in those cells which are not p53-deficient, damage response would have to be prevented artificially. Accordingly, short-telomere cells could be potential targets of therapy [70].

1.4.1 The PanIN model for progression of precursor lesions to carcinoma

1.4.1.1 Stepwise progression from PanIN-1A to PanIN-3 and PDAC

The most common type of pancreatic preneoplastic lesion is called pancreatic intraepithelial neoplasia (PanIN). PanINs are small lesions (< 5 mm) located in the pancreatic ducts. The *PanIN progression model* with different subtypes of PanINs was designed in order to establish a descriptive model reflecting different stages of progression to PDAC with regard to tissue transformation. Yet it is not clear whether all precursor lesions undergo the same stepwise process of tissue transition or whether there are special subtypes of transition in which certain steps are "skipped".

The model assumes five transitional steps to take place for the transformation of normal pancreatic tissue to invasive carcinoma: Normal tissue $\longrightarrow$ PanIN-1A $\longrightarrow$ PanIN-1B $\longrightarrow$ PanIN-2 $\longrightarrow$ PanIN-3 $\longrightarrow$ PDAC. Visualization and histological examples are given in Fig. 1.4.

PanIN-1A and -1B lesions are characterized by different extent of mucin expression and papillary growth. The above described KRAS-mutation is found in most of these lesions [107]. First nuclear abnormalities are observed in PanIN-2 lesions, usually coinciding with p16 mutations [98]. PanIN-3 is then characterized by pronounced nuclear atypia, mitosis and luminal budding and generally accompanied by p53 and SMAD4 mutations [108]. PDAC tissue shows invasion and desmoplasia.

1.4.1.2 Direct transition from normal pancreas to PanIN-2 or PanIN-3

However, the described linear progression of PanIN has not been proven to be correct but rather describes an attempt to embed commonly observed intermediate steps of tissue transformation into a well-defined model. In fact, the model contains several implications which are not necessarily correct. Firstly, it implies that once tissue transformation is initiated, PanIN-1A/B is formed and KRAS is mutated, the lesion will progress towards advanced PanIN stages and finally PDAC in all cases. However, it could also be possible that KRAS-mutated acinar cells in PanIN-1 undergo cell cycle arrest leading to destruction of the lesion. An alternative model by Real et al. [111] assumes that only additional loss of heterozygosity (LOH) in a tumor suppressor gene (e.g. p16, p53, SMAD4) will induce further progression. Accordingly, PanINs will usually be destroyed if not being promoted by additional genetic events. In addition, when

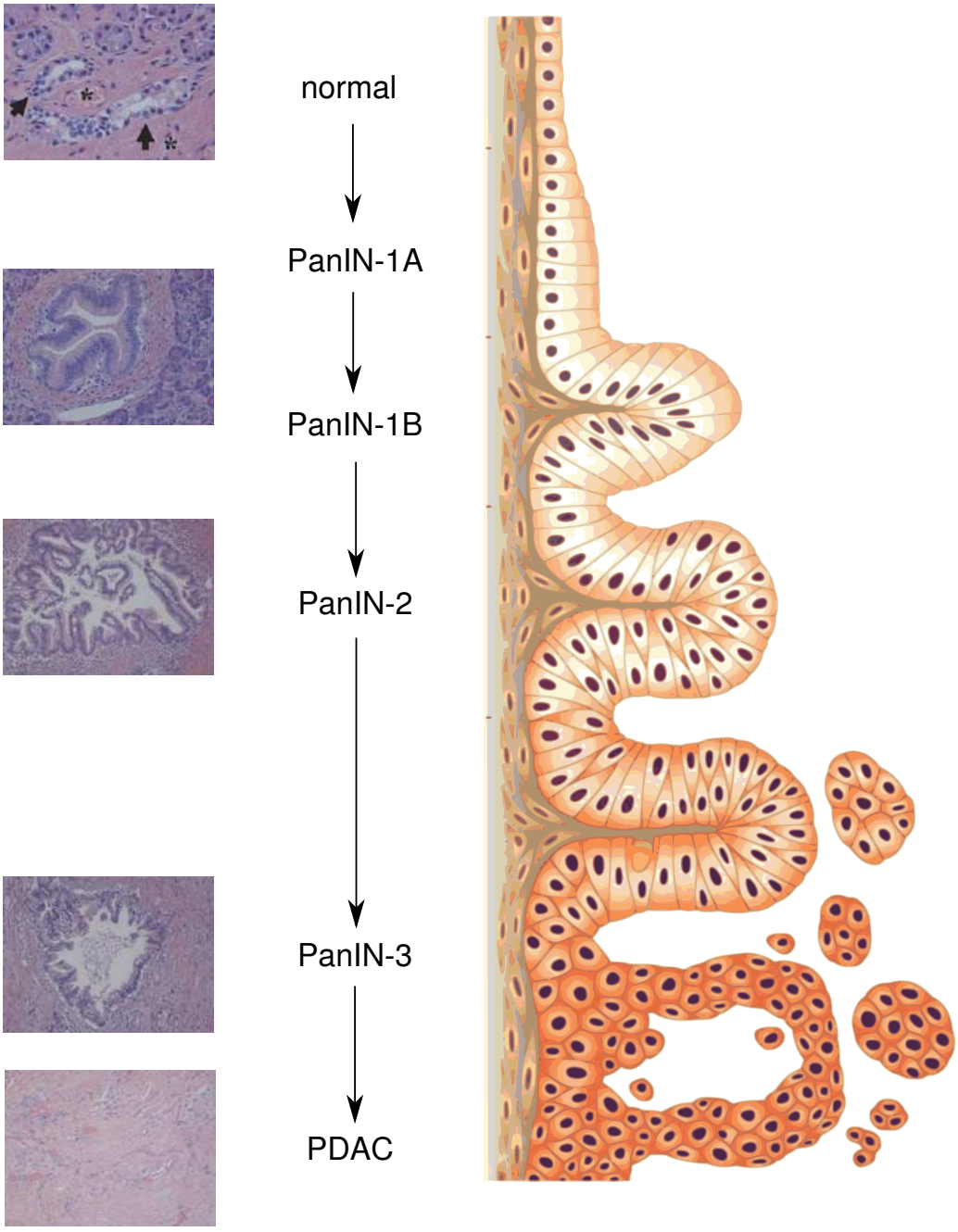

Figure 1.4: PanIN model for stepwise progression from normal pancreatic tissue to PDAC. Left: HE stainings (coloring all nuclei) of different progression stages; staining between PanIN-1A and PanIN-1B is an intermediate stage. Stainings adapted from [109]. Right: Illustration of acinar structures undergoing stepwise disruption. Sketch adapted from [110].

several mutation events occur at the same time (e.g. KRAS, p16) it could also be possible that normal tissue undergoes direct transition to PanIN-2 or even PanIN-3. In accordance with these objections it was recently found that PanIN-2 and PanIN-3 share many common features with PDAC on the transcriptional level, in contrast to PanIN-1 which shows higher similarity

to normal pancreatic tissue [111].

1.4.2 Pathways activated in PDAC

In order to understand molecular mechanisms in different stages of PDAC progression, studies are not only focusing on the mutational status but moreover on the activity of intercellular signaling pathways with special regard to growth-relevant embryonic pathways, i.e. Hedgehog Notch and Wnt/β-catenin signaling [112, 113, 117].

Hedgehog signaling is proven to mediate PDAC progression with sonic hedgehog being hyperactivated in all intermediate PanIN stages [112]. Pdx-Shh mouse models with Shh being dysregulated in the pancreatic endoderm have been shown to develop PanIN lesions and Kras mutations as well as HER-2/neu mutations [112]. In addition, cyclopamine-mediated inhibition of Hh signaling induces apoptosis and blocks proliferation in a subset of PDAC cell lines. These findings indicate that Hedgehog signaling is playing a key role in initiation of PDAC, in particular as it is predating Kras mutation in mouse models.

Elevated expression of Notch members and the Notch target Hes1 was reported in human PDAC and PanIN lesions but also in genetic mouse models [113]. In mice, Notch has also been shown to contribute not only to progression but also to initiation of tissue transformation [114]. Additionally, inhibition of γ-secretase, a protease required for Notch processing, has been shown to reduce the number of PanIN lesions and suppress tumor formation in oncogenic mouse models [115]. This inhibition has also been shown to promote hypoxic necrosis in advanced mouse PDAC and can thus be regarded as future rationale for clinical translation [116].

β-catenin, member of the Wnt signaling pathway, is commonly found to be enriched in human PanIN and PDAC in the nucleus as well as in the cytoplasm [117]. However, activity of the corresponding Wnt pathway has been only demonstrated in a subset of cell lines [118]. Mouse model experiments have proven the growth-mediating function of β-catenin activity by inducing β-catenin overexpression using an early Pdx1-driven promoter: When overexpressed in all pancreatic progenitors, mice develop exocrine agenesis, cysts and postnatal lethality [119]. Usage of a late Pdx1-driven promoter however restricted β-catenin activity to the acinar compartments and islet cells [119]. However, no tumors were found to be developing in these models up to 1 year of age [119]. Ptf1a-Cre mediated activation, on the contrary, has been shown to induce cystic tumor formation in mice [120].

mTOR (mechanistic target of rapamycin) is a non-embryonic pathway regulating cell growth, survival, motility, proliferation, protein synthesis and transcription. It has also been shown to be dysregulated in pancreatic tumor entities and its role in PDAC progression is currently being studied. Mouse models with Kras mutations and $\text{Pten}^{-/-}$ were shown to exhibit a remarkable dependance on mTOR signaling where mTOR inhibition leads to proliferation arrest and tumor regression. In mice with Kras and p53 point mutations however, these effects could not be observed [121]. This finding is particularly relevant because p53 mutations are frequently occurring in human PDAC patients. It has been shown that single inhibition of mTOR (as also of Mek) does not lead to proliferation arrest due to feedback activation of Erk/Akt [122]. Instead, dual

inhibition of Mek and PI3K signaling is required for an efficient inhibition of mTOR, leading to induction of apoptosis [122].

1.4.3 Cross-talk of epithelial and mesenchymal cells

In section 1.2.2.1 it was introduced that under the influence of cytokines and chemokines, pancreatic stellate cells (PSCs) can switch from a quiescent state to an active state in which they become α-SMA expressing, myofibroblast-like cells. In the active state, PSCs then produce ECM components as well as their inhibitors [57]. Consequently, prolonged production of these components causes stromal expansion and can further induce fibrosis [57, 16].

PDAC are usually surrounded by an extensive stroma originating from the described effect. This stroma formed by activated PSCs is also called *desmoplasia*. Tumor environment studies raised the question whether there is a cross-talk of tumor cells and surrounding mesenchymal cells, and if so, which signaling pathways are mediating this cross-talk [136]. After a variety of pathways has been shown to contribute to the communication of epithelial and mesenchymal cells, more and more studies are focusing on this subject called *tumor-stroma interactions*.

Once regarded as a host response for protection of the environmental tissue, desmoplasia formation is now being investigated as a *"dynamic series of autocrine and paracine signaling interactions between host cells and tumor cells that both enhances the desmoplastic response and accelerates pancreatic cancer initiation, progression and metastasis"* [136]. Indeed, a variety of experiments has shown that PSCs contribute to progression of PDAC: conditioned PSC media induces proliferation, invasion and migration when added to pancreatic cancer cells *in vitro* [137]. *In vivo*, injection of a mixture of pancreatic cancer cells and PSCs causes increased tumorigenesis as compared to injection of pancreatic cancer cells alone [138]. MMP production of stroma cells suggests that migration of tumor cells is supported through degradation of the extracellular matrix, explaining the promoting effect. Other tumor-promoting factors released by stroma cells include PDGF, FGF, SPARC, EGF, HGF [136].

Though the exact mechanisms of the tumor-stroma cross-talk remain still unknown, recent studies have shown that elements of embryonic development are recapitulated, including in particular activity of the Hedgehog, Notch and TGFβ signaling pathway [136].

1. An experiment for assessing the role of Hedgehog (Hh) is based on the implantation of human pancreatic cancer xenografts in mice: implantation of the generally Hh-ligand producing human PDAC xenografts in sub-optimal numbers alone did not cause tumor formation, whereas additional injection of mouse embryonic fibroblasts (MEFs) subsequently induced tumor formation [139]. In contrast, when the experiment was repeated with wild-type non-Smo-expressing MEFs (Smo is a Hh transducing membrane protein), tumors were found to be reduced in size significantly [139], suggesting that pronounced Hh signaling activity contributes to the tumor-stroma cross-talk, promoting tumor invasion. Further studies confirm the important role of Hh in this process [140, 141].

2. Notch is not only an important regulator of pancreatic development but also expressed

in adult PSCs [142]. Co-culture experiments with pancreatic cancer cells and pancreatic myofibroblast cells showed that mRNA levels of Hes1 (a downstream target of Notch) are increased in both of these cell types, suggesting a cross-talk mediated by Notch which can either be just stromal-to-epithelial or also epithelial-to-epithelial [138].

3. TGFβ is known to be secreted by pancreatic cancer cells; accordingly, corresponding signaling pathway was regarded a potential candidate for the tumor-stroma cross-talk. TGFβ receptors II and III were frequently found to be upregulated in PDAC patients, TGFβ-RII being upregulated in cancer cells and TGFβ-RIII being upregulated in stroma cells [143]. This finding could indicate a potential role of TGFβ-mediated interactions between tumor and stroma cells. Still, this hypothesis has not been proven yet.

Further investigation of the tumor-stroma cross-talk will be the focus of future research studies.

1.4.4 Epithelial-mesenchymal transition

Epithelial-mesenchymal transition (EMT) is a process in which epithelial cells undergo specific changes of the molecular program which finally leads to acquirement of the mesenchymal phenotype. Depending on the context of occurrence, EMT is classified into three subtypes: developmental (type I), fibrotic and regenerative (type II) and cancer-specific (type III). Studies focusing on pancreatic cancer have shown that type III EMT plays an important role for PDAC progression, maintenance and generation of stem cells, metastasis and treatment resistance [15, 144, 145]. Consequently, considerations about molecular mechanisms initiating EMT are required for studies focusing on early pancreatic carcinogenesis.

EMT can be regarded as a four-step process:

1. *Loss of tight junctions, adherens junctions and desmosomes.* Epithelial cells are arranged in clearly shaped compartments, forming a characteristic apico-basal axis of polarity. This arrangement is supported by tight junctions, adherens junctions and desmosomes. In contrast, mesenchymal cells are organized loosely in the extracellular matrix (ECM). Initiation of EMT is characterized by loss of the described adhesion structures. This is usually driven by loss of E-cadherin which is required for their maintenance [147]. In addition, cell polarity is lost [147]. As E-cadherin loss is the initiating step of EMT, any transcription factor (TF) leading to E-cadherin inhibiton can be regarded as potential EMT inducer; these TFs include Snail1/2, ZEB1/2, Slug, E47 and KLF8 [147, 148].

2. *Cytoskeletal changes.* As a consequence of loss of E-cadherin, β-catenin can now translocate to the cell nucleus, leading to activation of its target genes [147]. The actin cytoskeleton starts to form stress fibers anchoring to the focal adhesion complexes, allowing cell migration.

3. *Transcriptional shift.* Activation of Snail, ZEB and bHLH transcription factors leads to inhibiton of epithelial marker genes, while at the same time vimentin is upregulated and production of ECM components (e.g. collagens) is induced [147].

4. *Migration.* Expression of ECM components promote formation of focal adhesion complexes, and activation of N-cadherin allows cell motility. Finally, MMP production leads to destruction of the extracellular matrix, which in turn allows the cell to migrate to the mesenchyme [147].

The four steps are illustrated in Fig. 1.5.

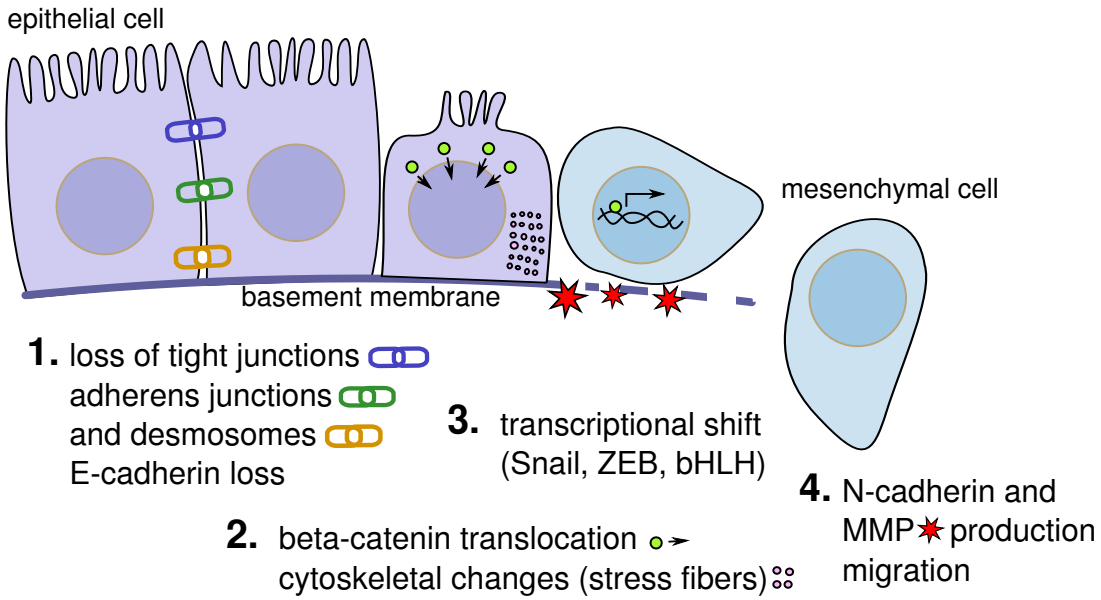

Figure 1.5: The progressive stages of EMT.

Once EMT is initiated by loss of E-cadherin, remaining steps are following immediately. Correspondingly, in order to identify the sources of origin of EMT type III-mediated mesenchymal expansion during pancreatic carcinogenesis, signaling cascades causing inhibiton of E-cadherin must be considered relevant. Recent studies have shown that several developmental pathways, including Wnt, Notch and TGFβ, but also FGF, EGF and HGF signaling can induce loss of E-cadherin [149]. As several elements of embryonic development have been shown to be recapitulated during pancreatic carcinogenesis, we here want to consider the mechanisms of EMT induction through developmental pathways.

1. Increased β-catenin levels activate TFs of the TCF family which in turn activate several EMT-inducing factors. Accordingly, Wnt signaling as an inhibitor of β-catenin degradation is an inducer of EMT. Wnt-mediated EMT has been observed during carcinogenesis, gastrulation and cardiac valve [147].

2. TGFβ signaling mediates EMT induction by SMAD phosphorylation or regulation of cell polarity or tight junction proteins like Par6; additionally it actives the PI3K/Akt/mTORC1 cascade which leads to activation of EMT inducers through NF-κB activation. Furthermore, TGFβ induces RhoA activity, leading to cytoskeletal changes which the EMT process is accompanied by [151]. Due to these manifold mechanisms,

TGFβ signaling can be regarded as the most relevant EMT-inducing pathway. Accordingly, it was reported to mediate all three types of EMT.

3. Notch signaling can induce EMT indirectly through activation of TGFβ signaling or NF-κB activation [147].

4. Additionally, microRNA-mediated EMT induction was documented [152, 153].

EMT is thought to play a striking role in PDAC, as reduced E-cadherin expression has been documented in 42%-53% of PDAC patients [154]. In addition reduction or loss of E-cadherin was correlated with metastasis formation [155]. Furthermore, it was shown that resistance to chemotherapy of PDAC patients is correlated to acquisition of EMT traits including ZEB expression [156]. In this regard, there is great interest in further investigating EMT mechanisms.

1.5 Other types of pancreatic tumors

The above explanations were focusing on PDAC as the major pancreatic tumor subtype. However, a variety of other tumor subtypes is well-documented though less extensively studied.

Intraductal papillary mucinous neoplasms (IPMNs) are cystic tumors arising from the ductal system which are characterized by mucin-producing cells. Though there is a broad range of IPMN subtypes with respect to extent of dysplasia and invasiveness, IPMNs are often regarded as another subtype of precursor lesion showing similar genetic variations as compared to PDAC. Patients with IPMNs are older on average, additionally the percentage of female patients is greater [123]. Additionally, a significant relation between IPMN-related symptoms and malignant cystic neoplasia was reported [123]. Due to rising incidences, clinical interest in IPMNs is growing [124]. Generally, it is distinguished between three subtypes of IPMNs, main-duct, branch-duct an combined IPMNs, with main-duct IPMNs showing an increased malignant potential and a higher percentage of invasiveness, worsening survival prognosis [125]. Patients with non-invasive IPMNs have a much better survival prognosis (five-year survival rate 77-100%) than patients with invasive IPMNs (five-year survival reate 43-65%) [126]. Subtypes of IPMNs are characterized by specific expression of different mucins [127].

Mucinous cystic neoplasms (MCNs) are another type of cystic pancreatic tumor characterized by columnar muncin producing cells [126]. They are regarded as distinct subtype as they are not communicating with the surrounding ductal system, in contrast to IPMNs [128]. MCNs are almost exclusively found in women and have a good survival prognosis compared to PDAC (five-year survival rate approaching 100%) [128].

Acinar cell carcinomas (ACCs) are tumors exhibiting less stroma than PDAC that are characterized by their production of digestive enzymes. ACCs are highly aggressive with a survival prognosis comparable to PDAC [129]. They account for 1%-2% of all pancreatic tumors [130]. They are large in size and patients with ACC have very unspecific symptoms (e.g. weight loss) [130].

Pancreatoblastoma form another pancreatic tumor formed by cells with acinar phenotype. Presence of squamoid nests and extensive lobulation are characteristics of pancreatoblastoma that are not shared with ACCs. Common symptoms indicating pancreatoblastoma include abdominal pain, emesis and jaundice [132]. Survival prognosis for pancreatoblastoma is generally good. The tumor occurs mainly in the childhood.

Pancreatic neuroendocrine tumors (PNETs) are uncommon endocrine tumors with an annual incidence of less than 1 per 100,000 persons [133]. It is generally distinguished between functional PNETs which produce hormones and non-functional PENs that do not produce hormones and form the majority of PNETs. Clinical prognosis depends on the extent of the disease and treatment; patients with resectable tumors have a better prognosis [133]. Also, prognosis of functional PENs is better than of non-functional PNETs. Functional PNETs are named after the hormone they produce, e.g. gastrinoma, glucagonoma with insulinoma being the most frequent functional PNET type. PNETs do not share typical genetic variations observed in PDAC, including Kras, p53, p16 or SMAD4 mutations [134].

Solid pseudopapillary neoplasms (SPNs) are cystic tumors. The cell of origin which this type of tumor is composed of, is unknown [130]. SPNs account for only 2%-5% of all pancreatic tumors. They are typically round in shape and have a diameter of 2-17 cm [135]. Also they were shown to be positive for vimentin and alpha 1-antitrypsin expression [135]. SPNs are typically observed in young women [130]. Mutations of β-catenin are found in around 90% of all SPNs; also, nuclear accumulation of β-catenin is observed very frequently [131].

Chapter 2

Motivation

2.1 The $Kras^{G12D}$ model

Oncogenic KRAS mutations are found in approximately 90% of all pancreatic cancer patients [76]. As KRAS mutations have also been documented as a critical genetic event in early pancreatic carcinogenesis, occurring in most PanIN-1A/B lesions, there is great interest in understanding causal effects of KRAS mediated tissue transformation.

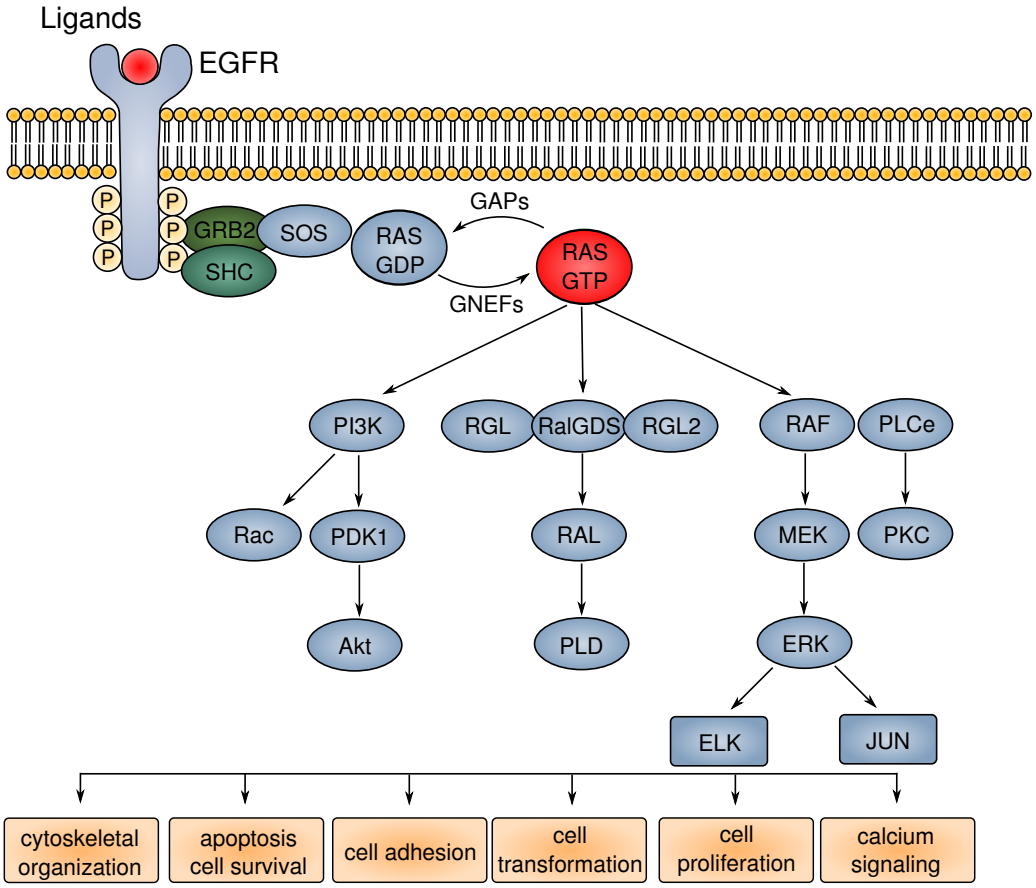

Figure 2.1: Ras signaling. Adapted from [157].

Pancreas-specific expression of $Kras^{G12D}$ in mice is a well-established method allowing to study early pancreatic carcinogenesis [158]. In section 1.4 it was described how the expression of oncogenic Kras impairs the switching function of the protein, resulting in a persistent activation of downstream cascades like PI3K/Akt and RAF/Mek/Erk (illustrated in Fig. 2.1). Correspondingly, once Kras signaling is induced, uncontrolled proliferation and tissue transformation is induced [159].

2.2 KRAS mutations in human PDAC

Due to the highly frequent occurrence of KRAS mutations in PDAC patients, KRAS has been considered as a potential drug target, respectively. However, all clinical attempts trying to interfere with KRAS-induced signals have failed and it is now commonly assumed that KRAS is undruggable [107]. Consequently, current studies are focusing on targeting effector pathways of KRAS, including PI3K/Akt or RAF/Mek/Erk. Notably, it was shown that tumor microenvironment effects like tumor-stroma interactions also depend on KRAS activity and that the observed stromal expansion regresses after KRAS inactivation [169]. In this regard, persistent KRAS activity can be regarded as required maintenance factor for pancreatic tumor cells and inhibition of KRAS-induced effects is a promising perspective for treatment of PDAC.

2.3 Previous PDAC studies on Kras-mutated mice

Previous studies have shown that during development, $Kras^{G12D}$-mutated mice invariably produce preneoplastic lesions which in one third of the cases further progress into invasive tumors [159]. Additional mutations of tumor suppressor genes further decrease survival prognosis. Mice cohorts with a combination of p53 and Kras mutations exhibit 100% mortality within 12 months, simple Kras-mutated mice have a survival time increased about 5 months on average [159].

In mature epithelial cells however, tissue transformation cannot be observed in most samples, even in case of additional deletion of tumor suppresor genes [160]. Also, healthy individuals are carriers of oncogenic Kras mutations without developing pancreatic cancer [161].

These outcomes suggest that oncogenic Kras alone is not an initiating force for the development of PDAC. However, it has been suggested that malignant transformation in Kras-mutated mice can be triggered with an additional inflammatory stimulus, causing acute or chronic inflammation of the pancreas [162, 163]. This can be possibly explained by a crosstalk of Kras signals and signals originating from inflammatory response including, for instance, Notch or Hedgehog signals discussed in section 1.2.2.2. Causal effects of pro-inflammatory chemokine and cytokine release as discussed in section 1.2.2.1 further contribute to the complex pathway interference in early pancreatic carcinogenesis. Studies focusing on these crosstalks have unveiled that a canonical component of the NF-κB pathway, Ikk2, synergizes with basal Notch signaling in order to activate transcriptional Notch targets [164]. Genetic deletion of Ikk2 in Kras-mutated mice led to a substancially delayed carcinogenesis and also caused downregulation of Notch targets Hes1

and Hey1 [164]. Hes1 is a suppressor of Pparγ which causes induction of antiinflammatory signals. The described effect requires NF-κB signaling to be activated. Studies have shown that this activation is directly tracing back to Kras. Kras activates transcription factor AP-1 which promotes Il-1α production [165]. Il-1α than induces a feed-forward loop by activating NF-κB signals which in turn promote Il-1α and p62 expression again. It was further shown that these feed-forward loops are required for PDAC development and that Il-1α upregulation coincides with KRAS mutations and NF-κB activation in PDAC patients.

Another study has shown that blocking a single isoform of PI3K, p110α is sufficient to block the transition from acinar cells to preneoplasic lesions in Kras-mutated mice [168], motivating further studies on downstream targets of Kras.

Whatever the distinct molecular mechanisms are, it is generally assumed that the combination of the inflammatory stimulus and oncogenic Kras mutation creates a "transformation-permissive microenvironment" impairing coordinated proliferation of epithelial and mesenchymal cell lineages. In order to analyze the mechanisms of normal pancreatic regeneration in wild-type mice and impaired regeneration in Kras-mutated mice after inflammatory injury, we, for the first time analyze histological and transcriptional changes in both mouse strains with a high temporal resolution. In particular, the aims of our study include:

- A detailed description of the temporal course of natural pancreatic regeneration after inflammatory injury in wild-type mice, with respect to histological and molecular changes.
- Comparative analysis of natural pancreatic regeneration and impaired regeneration in Kras^{G12D}-mutated mice.
- Characterization of proliferation cascades of acinar, progenitor-like and mesenchymal cells after inflammatory injury in both mouse strains.
- Detection of marker genes / gene sets indicating an *early stage* of pancreatic carcinogenesis; transferability to human PDAC data.
- Identification of intercellular signaling pathways required for successful mediation of regeneration in wild-type mice or tissue transformation in Kras^{G12D}-mutated mice.

Chapter 3

Material and methods

In the previous chapter it was discussed that the use of Kras^{G12D} GEMMs is a well-established method allowing to model early pancreatic carcinogenesis. However, Kras mutations alone are not sufficient to cause a malignant transformation in adult, but only in developmental mouse models. In adult mouse models, an additional inflammatory stimulus is required. Caerulein-induced pancreatitis is a matching trigger inducing tissue transformation.

Our study focuses on the exploration of the molecular and histological landscape of early pancreatic carcinogenesis. To do this, we perform a comparative analysis between pancreatic regeneration in wild-type mice and malignant transformation in Kras-mutated mice after caerulein-induced inflammatory injury. Particularly, the divergence of the natural course of inflammatory response leading to complete regeneration of the organ in wild-type mice and tissue transformation in Kras-mutated mice will be addressed.

3.1 Experimental design

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258].

In order to achieve pancreas-specific expression of Kras^{G12D}, strain crossing of Kras-mutated mice and mice with pancreas-specific Cre recombinase was performed: Mice carrying the *Loxp-STOP-Loxp-Kras*^{G12D} gene, obtained from Jackson Laboratory (Bar Harbor, USA) were crossed with mice carrying p48$^{Cre/+}$, also known as Ptf1a$^{Cre/+}$, courtesy of Prof. Roland M. Schmidt and Jens T. Siveke (Dept. of Gastroenterology, TU Munich). These mice will be referred as "Kras^{G12D} mice" in the following. Wild-type mice (C57BL/6J) were obtained from Charles River Laboratory (Sulzfeld, Germany).

3.1.1 Treatment protocol

Acute pancreatitis in wild-type (WT) and Kras^{G12D} mice was induced at eight to nine weeks of age by administering caerulein according to the "consecutive protocol" [171], see also section

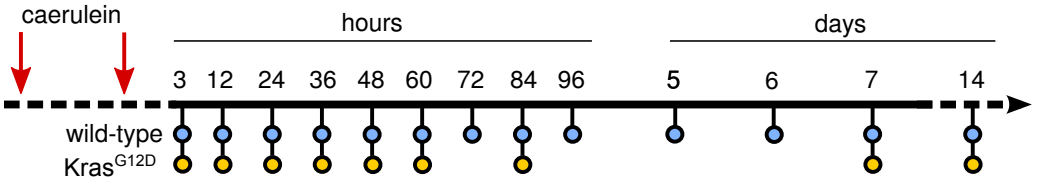

Figure 3.1: Time course of the experiment. Wild-type mice were sacrificed at 13, Kras^{G12D} mice were at nine different time points.

1.2.3. Caerulein treatment was performed by eight hourly intraperitoneal (i.p.) injections (2 mg per injection) on two consecutive days. Control animals were treated with 0.9% sodium chloride. On each day of injection, animals were treated with a subcutaneous bolus of Buprenorphine (Temgesic; 1 mg/kg bodyweight). The last injection was considered "hour 0" or "day 0". Control animals were sacrificed one hour after injections. Mice were injected with 2.5 mg BrdU i.p. two hours before sacrifice. To follow histological and molecular changes dynamically after the onset of acute pancreatitis, organs were collected at 13 time points in wild-type and 9 time points in Kras^{G12D} mice, from 3 hours to day 14 after the last injection. Additionally, organs from control samples were collected, respectively. Fig. 3.1 provides an overview of the experimental time course.

3.1.2 RNA isolation

In order to allow analysis of the transcriptional changes, RNA isolation was performed for the extracted tissue samples. Bulk tissues were processed according to instructions provided in the RNeasy mini kit (Qiagen, Venlo, Netherlands) using 50μl of β-mercaptoethanol in 1ml RLT buffer. The extracted RNA was eluted in purified water and shock-frozen in liquid nitrogen. Samples were stored at -80°C. Cell sorting was not performed.

3.1.3 Immunohistochemistry and immunofluorescence analysis

Hematoxylin and eosin stainings (H&E) as well as immunohistochemistry (IHC) and immunofluorescence (IF) are established methods allowing specific coloring for histological analysis of the tissue samples. H&E stainings color all nuclei, IHC and IF are antibody-based techniques, coloring specifically for a protein. IHC staining for Sox9$^+$ cells was performed in order to detect pancreatic progenitor cells, or, particularly dedifferentiated acinar cells acquiring a progenitor-like state as discussed in section 1.2.2.2. Stainings for α-SMA$^+$ were performed for identification of stroma cells, α-Amy$^+$-specific stainings were used to color acinar cells. Proliferating cells were colored with stainings for BrdU and immune cells were identified by coloring Cd45$^+$ cells. The following sections will provide information about the applied protocols and the antibodies used.

3.1.3.1 Tissue extraction, H&E, IHC and IF protocol

Harvested tissues were conserved overnight in 4% paraformaldehyde (PFA), then embedded in paraffin and cut into 3.5 μm thick slices. Sections were subjected to H&E, IHC and IF, as

previously described [162]. Briefly, citric buffer was used for antigen retrieval. IHC staining was performed using 3% BSA (bovine serum albumin, Sigma-Aldrich, Munich, Germany) in PBS as blocking and antibody dilution. IF staining was performed using Triton x100 diluted to 0.6% in PBS plus 3% BSA for blocking, except for Ck19 staining where 10% goat serum in PBS plus 0.6% Triton x100 was used for blocking, primary antibody dilution and dilution of the secondary antibody.

For quantification IHC pictures, five non-overlapping bright-field images were acquired per animal, resulting in an image size of 355 x 263 μm. For quantification of IF pictures, five non-overlapping images were acquired, resulting in an image size of 697 x 522 μm (for the quantification of proliferating α-SMA-positive cells) or an image size of 355 x 263 μm (for the quantification of proliferating α-amylase- and Sox9-positive cells).

3.1.3.2 Antibodies

The following antibodies were used: rat anti-mouse CD45 (IHC, BD Bioscience, Heidelberg, Germany), rat anti-BrdU (IF, Santa Cruz Biotechnology, Heidelberg, Germany); mouse anti-BrdU (IHC&IF, Cell Signaling Technology, NEB, Frankfurt/Main, Germany); mouse anti-Amylase (G-10, IF, Santa Cruz Biotechnology); mouse anti-α-SMA (IF, Dako, Hamburg, Germany); rabbit anti-Sox9 (IF, Merck Millipore, Billerica, MA, USA). For IHC, HRP labelled polymer anti-mouse or anti-rat secondary antibodies (Dako) was applied. For IF, fluorescence-labelled secondary antibodies: goat anti-mouse IgG Alexa Fluor 488 (Dako), goat anti-rat 596 and chicken anti-rabbit 488 (Alexa fluor, Invitrogen, Carlsbad, CA, USA) were used at a dilution of 1:200 and nuclei were counterstained with DAPi. Five non-overlapping images at a 200-fold magnification were acquired per animal.

Table 3.1: Number of collected tissues samples and number of microarrays used at different time points.

time point	WT tissues	WT arrays	$Kras^{G12D}$ tissues	$Kras^{G12D}$ arrays
control	5	5	4	4
3h	13	3	6	4
12h	6	3	8	3
24h	13	3	5	4
36h	6	3	9	3
48h	15	3	1	1
60h	2	3	9	3
72h	14	4		
84h	2	3	6	2
96h	13	3		
d5	16	4		
d6	2			
d7	13	3	7	2
d14	8	4	4	3
$\sum$	128	44	59	29

3.2 Microarrays

DNA microarrays are a well-established technology allowing to perform *transcriptome analysis*, based on given RNA samples. The following sections will give an overview about the concept of microarrays.

3.2.1 Transcriptional profiling with microarrays

Generally, there are different types of microarrays including DNA, protein, tissue arrays and others. Following explanations are referring to DNA microarrays as those chips are used for transcriptome analysis.

Today, microarrays are the most frequently applied technology for transcriptional profiling, as they are relatively inexpensive and easy to handle with respect to data volume. Recent advances in next-generation sequencing (NGS) approaches suggest that sooner or later microarray technology will be outdated largely due to constantly decreasing costs of NGS technologies [172]. In contrast to microarrays which allow only for quantification of predefined transcripts, NGS methods cover the entire transcriptome, making them particularly useful for detection of unknown transcripts. Still, at the moment, microarrays provide an easy-to-use method for quantification of most transcripts with standardized procedures for data normalization and further analysis.

The principle of DNA microarrays is based on strand hybridizations. A preselected set of sequences is placed on the chip during its production. These so-called *probes* are specifically designed to measure the expression of a corresponding transcript with the complementary sequence. Isolated RNA from the sample of interest is transcribed into complementary DNA (cDNA) and labeled with a fluorescent dye. Hybridization is then performed under controlled conditions. Finally, fluorescence intensity of the different probe spots is measured with a laser and transformed into a quantitative value.

Microarrays can be divided into two classes in terms of production: spotted arrays and *in situ* arrays. For spotted arrays, probes are produced from mRNA and then spotted on a glass slide. For *in situ* arrays on the contrary, oligonucleotides are produced in place. Spotted arrays are usually produced using two different dyes, e.g. Cy3 and Cy5 (two-channel-arrays) which allow for a comparative analysis of two samples with one array. For *in situ* arrays, only a single dye is used (one-channel-arrays). In place production of oligonucleotides for *in situ* arrays is usually based on a photolithographic process; predefined sequences of nucleotides are extended base by base using a photholithographic mask [173]. Today, *in situ* arrays are gereally preferred due to greater precision [174, 175].

3.2.2 Affymetrix arrays

GeneChip by *Affymetrix* is today the most popular platform of *in situ* microarrays. Gene-specific oligonucleotides (*oligos, probes*), have a length of approx. 50 base pairs (bp). One

chip contains around 1 million distinct oligonucleotides which are grouped in up to 50,000 *probe sets*, each probe set reflecting one predefined gene. In contrast to other manufacturers, Affymetrix includes two different oligonucleotides in each probe: perfect match oligos (PM) and mismatch oligos (MM). PM oligos are perfectly complementary to the cDNA of the gene they are supposed to hybridize with. MM oligos contain a single base alteration compared to the corresponding PM oligos. The PM/MM concept was designed to allow noise reduction, as there are always non-matching sequences binding to the probes. As the non-matching probes are binding unspecifically, they will bind to both PM and MM oligos whereas the actual cDNA will only bind to the PM oligos. Accordingly, noise can be removed by subtraction of the MM signal, theoretically.

3.2.3 Applied protocol

Transcriptional profiling was performed using Affymetrix DNA microarrays of type GeneChip Mouse Gene 1.0 ST. Total RNA (100 ng) was amplified using the WT Expression kit (Ambion) and the WT Terminal labeling and Fragmentation Kit (Affymetrix). 2 μg of amplified cDNA were hybridized on the arrays containing about 29,000 probe sets. The Agilent 2100 Bioanalyzer was used to assess RNA quality and only high quality RNA (RIN $>$ 7) was used for microarray analysis. Staining (Fluidics script FS450_0007) and scanning was done according to the Affymetrix expression protocol.

Profiling was performed for control samples and samples from all time points except day 6. Number of collected tissues and number of chips used at different stages is provided with table 3.1.

3.3 Data preprocessing

In order to obtain well-interpretable data, transcriptional profiles are preprocessed in several steps: quality control, normalization and filtering.

3.3.1 Quality control

Quality control is performed in order to discard microarrays that show large deviations, relative to other samples, which cannot be explained by biological effects and, accordingly, indicate experimental failure. These outliers can be detected with different methods including *between array comparison* (L_1-distance of profiles), *comparison of signal intensity distributions* and *variance-mean dependance*. Additionally, individual array quality can be assessed with *MA plots*.

MA plots were originally established for dual-channel arrays with two signal intensities per probe [176]. Given two intensities I_1, I_2, the MA plot visualizes log ratios $M = log_2(I_1) - log_2(I_2)$ versus mean averages $A = log_2(I_1 I_2)/2$. Both are generally assumed to be correlated. Thus, a correlation or dependence coefficient and its corresponding test can be used for quality assessment [177]. For single-channel arrays the reference chip is usually imitated using an averaged

combination of other chips from the same experiment. For Affymetrix chips, further quality control can be performed using the PM/MM ratios [178].

The above described steps are readily applicable using the arrayQualityMetrics package for Bioconductor/R [179].

3.3.2 Normalization

Normalization of the expression data is performed in order to correct samples for technical variation. Here, RMA (robust multi array) normalization is an established standard [180].

In a first step, signals are background-corrected and log_2-transformed [180]. In contrast to previous approaches, RMA background correction does not take advantage the PM vs. MM ratios of Affymetrix chips which are supposed to improve measurement precision (see section 3.2.2); it has been shown that total exclusion of MM probes yields better results [180]. Instead, RMA uses so-called *convolution background correction* which is based on a statistical model for the PM probes [180].

Afterwards, quantile normalization is performed according to Bolstad [181, 180]. Finally, each *probe set* (multiple probes on the chip referring to the same gene) is summarized to a single gene expression level using *Tukey's median polish* [180]. Resulting gene-specific expression profiles are then used for further analysis.

3.3.3 Probe filtering

After normalization, gene filtering is performed. As generally less than 30-40% of all genes are expressed in a given tissue, most of the normalized expression profiles can be discarded [182]. Selection is performed variance-based: It was shown, that discarding genes with a small variance is a viable method for improving the results of further analyses [172]. Particularly, due to the decreased number of genes, differential expression analysis will yield more true positives after p-value correction [172]. By default, 50% of the profiles with the smallest variance are discarded using the genefilter package for Bioconductor/R [183].

3.3.4 Application of preprocessing steps

Our experimental data comprises 73 microarrays (44 for wild-type mice, 29 for $Kras^{G12D}$ mice). For most time points, three replicates were made; a precise listing of the sample counts is provided with table 3.1. In a first step, quality control was performed for all 73 arrays. None of the chips was withdrawn, as all arrays passed at least 4 out of the 5 quality control methods provided with R package arrayQualityMetrics. Subsequently, RMA normalization was performed as provided with R package affy. As a final step of data preprocessing, probe filtering was performed; here, all probes that could not be assigned to an Entrez ID were removed. In case of multiple probes referring to the same Entrez ID, only the probe with the greatest variance was retained. Afterwards, 50% of the remaining probes with the smallest variance were discarded.

After filtering, the normalized expression dataset comprised 10,360 gene expression profiles and 73 samples and was used for all further analysis.

3.4 Data modeling and analysis

3.4.1 Hierarchical clustering

Hierarchical clustering is a distance-based method allowing to identify tree-like structures of feature or sample similarities in a given dataset. *Complete linkage clustering* is an agglomerative hierarchical clustering procedure. Given m samples with n dimensions, x_1, ..., $x_m \in \mathbb{R}^n$. Initially, each sample forms an own cluster.

$$\mathcal{X}_1 = \{x_1\}\,,\ \ldots\,, \mathcal{X}_m = \{x_m\}$$

Based on the initial set $\mathcal{X}^{(1)} = \{\{\mathcal{X}_1\}\,,\ \ldots,\ \{\mathcal{X}_m\}\}$, clusters are merged iteratively according to the shortest between-cluster distance computed with the complete linkage function. For two clusters $\mathcal{X}_i$, $\mathcal{X}_j$ this is $D(\mathcal{X}_i, \mathcal{X}_j) = \max_{x\in\mathcal{X}_i,\, y\in\mathcal{X}_j} d(x,y)$ where $d(x,y)$ is the distance function. Commonly, the L_2-norm (Euclidean distance) is used as distance function:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The step from iteration k to iteration $(k+1)$ is then

$$\mathcal{X}_1^*, \mathcal{X}_2^* = \operatorname{argmin}_{\mathcal{X}_i\in\mathcal{X}^{(k)},\, \mathcal{X}_j\in\mathcal{X}^{(k)}} D(\mathcal{X}_i, \mathcal{X}_j)$$
$$\mathcal{X}^{(k+1)} = \left\{\mathcal{X}^{(k)} \cup \{\mathcal{X}_1^*, \mathcal{X}_2^*\}\right\} \setminus \{\mathcal{X}_1^*\} \setminus \{\mathcal{X}_2^*\}$$

The clustering procedure terminates when all subclusters are merged to one big cluster; distances can then be visualized in a *dendrogram*.

A *dendrogram* is a visualization method for hierarchical clustering results. According to the clustering procedure, a tree-like structure arises. Original samples can then be arranged according to this structure, i.e. each sample or subcluster is arranged together with the sample or subcluster it has been merged with. Then, the dendrogram complements data visualization with a tree reflecting the computed cluster distances. An example is given in Fig. 3.2.

Within the dendrogram, the sum of distances between a node and its two child nodes reflects the distance between the clusters and its subclusters.

However, paths cannot be interpreted as distances, except for a special case on the leaf level where the distances of two leaves connected via a common father node can be summed up. In order to avoid misinterpretation, the leaves are often visualized with a uniform distance. Then, distance to the father cluster can still be anticipated by distance to the x-axis.

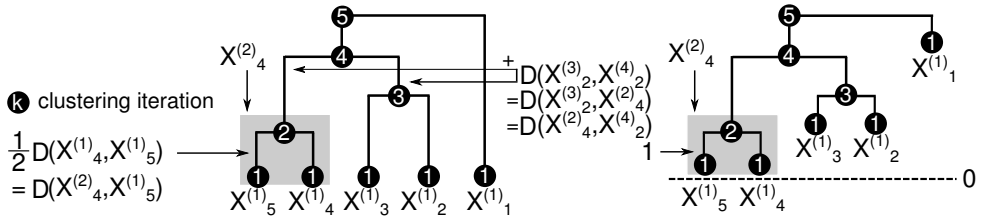

Figure 3.2: Two dendrograms visualizing the same clustering result. In the right part, leaves are visualized with a uniform distance.

3.4.2 PCA and PARAFAC

Several factorization-based methods allow general investigation of a dataset with respect to variance. In this context, they help identifying influential variables and visualizing data using principal components.

Let $\mathbf{X}_{m\times n} = (x_{ij})$ be the *expression matrix* containing the expression levels of m different genes in n different samples after preprocessing has been performed. x_{ij} refers to the expression of gene i in sample j. Correspondingly, the i^{th} row of $\mathbf{X}$ reflects the *gene expression profile* of gene i across samples and the j^{th} column reflects the *sample expression profile* of sample j for all genes.

3.4.2.1 Singular value decomposition

Singular value decomposition (SVD) is the factorization of $\mathbf{X}$ into

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U}$ is a unitary $m \times m$ matrix, $\mathbf{V}^\top$ is the conjugate transpose of a unitary $n \times n$ matrix and $\mathbf{\Sigma}$ is a diagonal $m \times n$ matrix. As $\mathbf{U}$ and $\mathbf{V}$ are defined to be unitary, $\mathbf{U}^\top\mathbf{U} = \mathbf{I_m}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I_n}$. Geometrically, if $\mathbf{X}$ is regarded as linear function $f : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ with $f(\mathbf{w}) = \mathbf{X}\mathbf{w}$, application of $\mathbf{X}$ can be interpreted as a sequence of three operations: a transformation, followed by a rescaling, followed by another transformation. $\mathbf{U}$ corresponds to the first transformation, $\mathbf{V}^\top$ to the second transformation and $\mathbf{\Sigma}$ to the rescaling operation. Thus, the gain of information achieved by application of SVD is the identification of these operations. Particularly, columns of $\mathbf{U}$ yield an orthonormal basis of the *variables* reflected by the rows of $\mathbf{X}$ and rows of $\mathbf{V}^\top$ yield an orthonormal basis of the *samples* reflected by the columns of $\mathbf{X}$. The elements of the diagonal matrix $\mathbf{\Sigma}$ are termed *singular values*.

3.4.2.2 PCA

Principal component analysis (PCA) is a method which is conceptually closely related to SVD. Here, the covariance matrix $\mathbf{X}^\top\mathbf{X}$ is decomposed in order to transform data into a new vector space with the basis being formed by the eigenvectors of $\mathbf{X}^\top\mathbf{X}$, the principal components, which

allow for a variance-maximizing transformation. In a first step, variables in $\mathbf{X}$ are standardized row-wise. Then, factorization into two matrices is performed in a way that

$$\mathbf{X}^\top\mathbf{X} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top$$

with $\mathbf{W}$ containing the eigenvectors of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{\Lambda}$ containing the eigenvalues as diagonal entries. Columns of $\mathbf{W}$ are then reordered with respect to singular values (decreasingly). The first L eigenvectors of $\mathbf{W}$ are finally used as *loading matrix* $\mathbf{W}_L$. Then, $\mathbf{X}\mathbf{W}_L$ allows the desired transformation. This "truncated transformation" is used for *dimensionality reduction*; the underlying assumption is that those components with the greatest variance also contain the most information.

The following consideration unveils the relationship between PCA and SVD: Since $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma^2}\mathbf{V}^\top$, $\mathbf{V}$ contains the eigenvectors of the covariance matrix $\mathbf{X}^\top\mathbf{X}$ and thus corresponds to $\mathbf{W}$, and $\mathbf{\Sigma^2}$ contains the eigenvalues and thus corresponds to $\mathbf{\Lambda}$.

After PCA has been applied, a *biplot* can be generated in order to visualize the data in the new vector space, using two or three of the computed principal components.

3.4.2.3 Extension to tensors: PARAFAC

Given a matrix, PCA is a very useful method for dimensionality reduction and variance maximization. However, an extension of the concept is required when factorization should be performed for *multi-way arrays*. An approximation algorithm called *parallel factor analysis (PARAFAC)* which was described by Richard A. Harshman [186] can be applied in order to minimize the error in multi-way decomposition. Consider p gene expression matrices $\mathbf{X}_1$, ..., $\mathbf{X}_p$ resulting from p different *experimental protocols*, with the same set of m genes and n samples prepared under equal or *comparable* conditions, with $(x_{ij})_k$ reflecting expression of gene i in sample j for protocol k. As dimensions are equal $\dim(\mathbf{X}_1) = \ ... \ = \dim(\mathbf{X}_p) = (m \times n)$ and indexes are matching, this data can be summarized in a *three-way array* (tensor) $\mathbf{Y}_{m\times n\times p}$. Then, PARAFAC performs a three-way array decomposition and reduction to L dimensions by minimizing the error term in

$$\mathbf{Y} = \sum_{\ell=1}^{L} \mathbf{a}_\ell \otimes \mathbf{b}_\ell \otimes \mathbf{c}_\ell + \boldsymbol{\epsilon}$$

with $\mathbf{a}_\ell$, $\mathbf{b}_\ell$, $\mathbf{c}_\ell$ being the ℓ^{th} columns of the resulting loading matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Depending on the context of interest, each of the three truncated loading matrices can then be used in order to visualize the data, i.e. visualization of genes, samples or protocols. Implementations of PARAFAC are available with the PTAk package for R [187].

3.4.2.4 Confidence ellipsoids

Consider a biplot that was generated after PCA (or another decomposition method) has been applied to $\mathbf{X}^\top$, with transformed values $\mathbf{Y} := \mathbf{X}^\top\mathbf{W}_L$. Let $\mathcal{A} \subseteq \{1, ..., n\}$ be a *subset of*

samples from expression matrix $\mathbf{X}$, e.g. samples prepared under a certain condition. Then, a *confidence ellipsoid* for a significance level α can be drawn in order to visualize the part of the space in the plot in which a random sample prepared under the certain condition will be located with a probability of $1-\alpha$ (called $(1-\alpha)$ confidence ellipsoid)). This part of the space is estimated using the sample means $\hat{\boldsymbol{\mu}}_{\mathbf{Y}[\mathcal{A},]}$ and the covariance matrix $\mathbf{C} := (\mathbf{Y}[\mathcal{A},])^{\top}(\mathbf{Y}[\mathcal{A},])$. Let $\mathbf{v}_1^{\mathbf{C}}$, ..., $\mathbf{v}_L^{\mathbf{C}}$ be the eigenvectors of $\mathbf{C}$. The ellipsoid is now created in two steps. First, the size is set up with the standard ellipsoid function for a coordinate $\mathbf{z} = (z_1, \ldots, z_L)$

$$\frac{(z_1 - (\hat{\mu}_{\mathbf{Y}[\mathcal{A},]})_1)^2}{(\hat{\sigma}^2_{\mathbf{Y}[\mathcal{A},]})_1} + \ldots + \frac{(z_L - (\hat{\mu}_{\mathbf{Y}[\mathcal{A},]})_L)^2}{(\hat{\sigma}^2_{\mathbf{Y}[\mathcal{A},]})_L} = s$$

where $(\hat{\sigma}^2_{\mathbf{Y}[\mathcal{A},]})_i$ is the sample variance of row i in $\mathbf{Y}[\mathcal{A},]$ and s is a scaling factor. This scaling factor s is used to adjust the ellipsoid size according to the significance level α and since we assume sample values to be distributed normally, and sum of L squared normally distributed data points is following a χ^2-distribution with L degrees of freedom, s can be found using the inverse cumulative distribution function so that $P(\chi^2_L \leq s) = 1-\alpha$. Finally, the ellipsoid must be rotated. The rotation angle for axis i is calculated by $\theta_i = \arctan\left(\left(\mathbf{v}_i^{\mathbf{C}}\right)_L / \left(\mathbf{v}_i^{\mathbf{C}}\right)_i\right)$, for $1 \leq i \leq L-1$. The distance between any coordinate (e.g. other sample) to the center of the ellipsoid can be simply calculated with $d\left(\mathbf{z}, \hat{\boldsymbol{\mu}}_{\mathbf{Y}[\mathcal{A},]}\right)$, e.g. using the L_2-norm.

3.4.3 Differential expression analysis

Given an *expression matrix* $\mathbf{X}_{m\times n}$ (m genes, n samples) and u disjoint *sets of samples* $\mathcal{A}_1$, ..., $\mathcal{A}_u$, defined with column indexes of $\mathbf{X}$, i.e. the following conditions are fulfilled: $\forall k, 1 \leq k \leq u : (\forall a_{k\ell} \in \mathcal{A}_k : 1 \leq a_{k\ell} \leq n \wedge \forall \mathcal{A}_{k'}, k' \neq k : \mathcal{A}_k \cap \mathcal{A}_{k'} = \emptyset)$. These sets will also be called *sample groups* in the following.

Then, *differential expression analysis* can be performed in order to determine genes which are expressed at a different level in at least one of the groups. The analysis is based on comparison of the mean expression levels. For each single gene i, the null hypothesis is

$$H_0^{(i)} : \forall \left\{k, k'\right\}, k \neq k', 1 \leq k \leq u,\ 1 \leq k' \leq u : \mu_{ik} = \mu_{ik'}$$

and the alternative hypothesis is

$$H_1^{(i)} : \exists \left\{k, k'\right\}, k \neq k', 1 \leq k \leq u,\ 1 \leq k' \leq u : \mu_{ik} \neq \mu_{ik'}$$

where μ_{ik} is the mean expression level of the gene in group k (which can be estimated using samples $\mathcal{A}_k$). For the following considerations, let $\mathbf{y} = \mathbf{X}[i,]$ be the *gene expression profile* of a gene i, i.e. $\mathbf{y}$ is a vector of n elements.

3.4.3.1 t-test for detection of differential expression between two groups

For $u = 2$, the testing procedure reduces to a single t-test. Given samples of two random variables reflecting expression of a gene i in two groups k, k', i.e. gene expression levels $\mathbf{y}_{\mathcal{A}_k}, \mathbf{y}_{\mathcal{A}_{k'}}$, following a normal distribution. Then, the null hypothesis of the two-sample t-test is

$$H_0^{(i)} : \mu_{ik} = \mu_{ik'}$$

where $\mu_{ik}, \mu_{ik'}$ are the mean expression levels of gene i in groups k, k'. Welch's t-statistic for two random variables with different variance is then calculated by

$$t = \frac{\hat{\mu}(\mathbf{y}_{\mathcal{A}_k}) - \hat{\mu}(\mathbf{y}_{\mathcal{A}_{k'}})}{\sqrt{\frac{\hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_k})}{N_{\mathcal{A}_k}} + \frac{\hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_{k'}})}{N_{\mathcal{A}_{k'}}}}}$$

where $\hat{\mu}(\mathbf{y}_{\mathcal{A}_k}), \hat{\mu}(\mathbf{y}_{\mathcal{A}_{k'}})$ are the sample means, $N_{\mathcal{A}_k} = |\mathcal{A}_k|$, $N_{\mathcal{A}_{k'}} = |\mathcal{A}_{k'}|$ and $\hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_k}), \hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_{k'}})$ are the sample variances. For the case the null hypothesis is correct, t-statistic is now assumed to follow a t-distribution with ν degrees of freedom where ν is being estimated with the Welch-Satterthwaite equation

$$\nu \approx \frac{\left(\frac{\hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_k})}{N_{\mathcal{A}_k}} + \frac{\hat{\sigma}^2(\mathbf{y}_{\mathcal{A}_{k'}})}{N_{\mathcal{A}_{k'}}}\right)^2}{\frac{\hat{\sigma}^4(\mathbf{y}_{\mathcal{A}_k})}{N_{\mathcal{A}_k}^2 + \nu_1} + \frac{\hat{\sigma}^4(\mathbf{y}_{\mathcal{A}_{k'}})}{N_{\mathcal{A}_{k'}}^2 + \nu_2}}$$

where $\nu_1 = N_{\mathcal{A}_k} - 1$ and $\nu_2 = N_{\mathcal{A}_{k'}} - 1$. Null hypothesis is rejected at significance level α if $|t| > t(1 - \frac{1}{2}\alpha, \nu)$ which is calculated with the probability density function of the t-distribution, or using a reference table. Unless stated differently, this *two-sided* variant of Welch's t-test is applied in the following. The *one-sided* variant could be applied in order to test $H_0^{(i)} : \mu_{ik} \leq \mu_{ik'}$; in this case, null hypothesis is rejected at significance level α if $t > t(1 - \alpha, \nu)$.

3.4.3.2 F-test for more than two groups (ANOVA)

For $u > 2$, however, application of multiple t-tests would result in α-error accumulation (i.e. false positive discoveries, null hypotheses that are rejected erroneously). Here, one-way analysis of variance (ANOVA) can be applied, using the F-test.

The idea of F-test application is based on the comparison of *between-group sample variability* $\hat{\sigma}^2_{BGV}$ estimating the variance *between* groups and *within-group sample variability* $\hat{\sigma}^2_{WGV}$ estimating the variance *within* the groups. $\hat{\sigma}^2_{BGV}$ is calculated by

$$\hat{\sigma}^2_{BGV} = \sum_{k=1}^{u} \frac{N_{\mathcal{A}_k}(\hat{\mu}(\mathbf{y}_{\mathcal{A}_k}) - \hat{\mu}(\mathbf{y}))^2}{u - 1}$$

and $\hat{\sigma}^2_{WGV}$ is calculated by

$$\hat{\sigma}^2_{WGV} = \sum_{k=1}^{u} \sum_{a_{k\ell} \in \mathcal{A}_k} \frac{(\mathbf{y}_{a_{k\ell}} - \hat{\mu}(\mathbf{y}_{\mathcal{A}_k}))^2}{N_{\mathcal{A}_k} - 1}$$

From these definitions, it directly follows that $\sigma^2_{BGV} \neq \sigma^2_{WGV} \Leftrightarrow \exists k \neq k' : \mu_{ik} \neq \mu_{ik'}$, because, if variances would be equal, then for any two groups k, k', members would be *exchangeable* without $\sigma^2_k, \sigma^2_{k'}$ and $\mu_k, \mu_{k'}$ being changed. Correspondingly, the null hypothesis of the F-test is

$$H_0^{(i)} : \sigma^2_{BGV} = \sigma^2_{WGV}$$

and the F-statistic is calculated by

$$F = \frac{\hat{\sigma}^2_{BGV}}{\hat{\sigma}^2_{WGV}}$$

and assumed to follow an F-distribution with $u - 1$, $u(n - 1)$ degrees of freedom.

As the above procedure is repeated for each gene separately, the information about multiple testing must be included when significance is assessed. To this end, multiple testing correction is applied, as described in section 3.4.4. Then, if the multiple testing correction yields a significant positive result for differential expression of a gene, the *decision* between which pairs of groups the gene is differentially expressed can be performed based on *two-sample t-tests* (after F-test application, differential expression is detected, but the affected groups are not yet identified). Then, after t-tests were performed, corresponding p-values are corrected again with one of the multiple testing correction methods.

3.4.3.3 Linear models for microarrays and RNA-Seq (Limma)

An advanced method for differential expression analysis has been developed by Smyth et al. and is provided in R package limma (linear models for microarrays and RNA-Seq) [191]. Here, linear models are set up for each gene. Let $\mathbf{y} = \mathbf{X}[i,]$ be the *expression profile of gene i across all samples*. Then, a binary *design matrix* $\mathbf{\Gamma} = (\gamma_{\ell k}) \in \{0,1\}^{n \times u}$ is set up, based on *sample grouping*:

$$\gamma_{\ell k} = \begin{cases} 1, & \text{if } \ell \in \mathcal{A}_k \\ 0, & \text{otherwise.} \end{cases}$$

As groups were defined to be disjoint, $\mathbf{\Gamma}$ can contain only up to one positive entry (1) per row. The linear model is then

$$\mathbf{y} \sim \mathbf{\Gamma}\boldsymbol{\beta} + \beta_0 + \boldsymbol{\epsilon}$$

After $\boldsymbol{\beta}$ has been estimated, statistics are computed according to the concept in the above sections. As sample values are replaced by the linear model here, *moderated* F-statistics and t-statistics are computed instead of the standard statistics described above. These statistics take advantage of the variance distribution across different genes which can be estimated. Estima-

tion of the moderated statistics is performed with an empirical Bayes approach [191]. Finally, multiple testing correction is performed.

3.4.4 Multiple testing correction

With each null hypothesis being tested on the same dataset, the probability for a null hypothesis being rejected erroneously increases. Multiple testing correction can be applied in order to counterpoise this α-error accumulation effect.

3.4.4.1 False discovery rates

P-values can be transformed into *false discovery rates* (FDR) in order to correct for multiple testing. Concept of FDRs was first introduced by Benjamini and Hochberg [188]. Let p_1, ..., p_m be the p-values of independent testing procedures (null hypotheses $H_0^{(1)}$, ..., $H_0^{(m)}$) and let α be the desired significance level. Then, the Benjamini-Hochberg procedure searches for the largest i with

$$\widetilde{p_i} \leq \frac{i}{m}\alpha$$

where $\widetilde{p}$ is the reordered list of p-values with $\widetilde{p_1} < ... < \widetilde{p_m}$, i.e. $\widetilde{p_i}$ is the i^{th} smallest p-value. Then, all null hypotheses $\widetilde{H_0}^{(1)}$, ..., $\widetilde{H_0}^{(i)}$ are rejected, i.e. corresponding variables are regarded as positive discoveries, where $\widetilde{H_0}$ is the reordered list of null hypotheses, according to the reordering procedure from p to $\widetilde{p}$.

3.4.4.2 Bonferroni correction

Bonferroni correction is a familywise error rate (FWER) method which is more strict than Benjamini and Hochberg's FDR concept [189, 190]. Here, any null hypothesis $\widetilde{H_0}^{(i)}$ is rejected only if

$$\widetilde{p_i} \leq \frac{1}{m}\alpha$$

3.4.5 Detection of differential expression of a gene set without predefined groups

In section 3.4.3 it was discussed how differential expression analysis is performed, given two or more groups. Another question arising is how differential expression can be detected for a given set of genes if the groups are unknown. Here, the concept of aggregated z-scores can be used.

Assume, expression levels are distributed normally across samples in an expression matrix $\mathbf{X}_{m\times n} = (x_{ij})$ (m genes, n samples). The z-score of a single gene i with expression profile $\mathbf{y} = \mathbf{X}[i,]$ for a given sample j, is calculated by $z_{ij} = (x_{ij} - \hat{\mu}(\mathbf{y}))/\hat{\sigma}(\mathbf{y})$, with $\hat{\mu}(\mathbf{y})$ and $\hat{\sigma}^2(\mathbf{y})$ referring to sample mean and sample variance. For a set of gene indexes $\mathcal{G} = \{g_1, ..., g_k \mid \forall g_\ell : 1 \leq g_\ell \leq m\}$, we can aggregate this z-score by simply calculating $Z_{\mathcal{G}j} = 1/\sqrt{k}\sum_{g_\ell \in \mathcal{G}} z_{g_\ell j}$; correspondingly $Z_{\mathcal{G}j}$ summarizes the expression of genes from the set $\mathcal{G}$ in sample j. These values can be summarized in the n-element vector $\mathbf{Z}_{\mathcal{G}}$.

According to our assumption, for a given gene set $\mathcal{G}$, $\mathbf{Z}_\mathcal{G}$ should follow a standard normal distribution across the samples, if there is no *differential expression* which we want to detect. It can now simply be tested whether this is the case or not, e.g. using a *Kolmogorov-Smirnov test* or *Shapiro-Wilk test*.

If the null hypothesis $H_0 : \mathbf{Z}_\mathcal{G} \sim \mathcal{N}(0,1)$ can then be rejected for a certain significance level α, differential expression can be assumed with a p-value $p_\mathcal{G}$. However, this result does not consider whether the given gene set has a *specific* characteristic that allows to separate samples into two or more groups. Assume all genes in the dataset to be differentially expressed between two distinct groups of samples; then, the null hypothesis would be rejected for any random set of genes. Accordingly, we set up a null model based on the p-value distribution for randomly assembled gene sets. Let $\mathbf{E}(p_\mathcal{S})$ be the expected p-value for differential expression of a random gene set $\mathcal{S}$ with $|\mathcal{S}| = |\mathcal{G}|$ (equal size). Then, we design a null model in order to test for the null hypothesis

$$H_0 : p_\mathcal{G} \geq \mathbf{E}(p_\mathcal{S})$$

To this end, N (e.g. $N = 100,000$) random sets are generated and p-values are computed and collected in a list. Finally the list of p-values is sorted increasingly. Let now $r_\mathcal{G}$ be the list index of the first p-value which is greater than $p_\mathcal{G}$. Then, significance can be assessed by calculating

$$p^*_\mathcal{G} = \frac{r_\mathcal{G}}{N}$$

3.4.5.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test can be applied in order to compare two distributions (two-sample test) or to compare a given sample distribution to a reference distribution (one-sample test). Let $\widetilde{\mathbf{Z}_\mathcal{G}}$ be the vector of aggregated z-scores, with values reordered increasingly. Then, the one-sample test for normality is based on the empirical distribution function $E_{\widetilde{\mathbf{Z}_\mathcal{G}}}(i) = \xi(i)/n$. Here, n refers to the number data points $\widetilde{Z_{\mathcal{G}_1}}$, ..., $\widetilde{Z_{\mathcal{G}_n}}$ and $\xi(i)$ is the number of values smaller than i. Test statistic is then computed by

$$D = \sup_{1 \leq i \leq n} \left| \Phi(\widetilde{\mathbf{Z}_{\mathcal{G}_i}}) - E_{\widetilde{\mathbf{Z}_\mathcal{G}}}(i) \right|$$

where Φ represents the cumulative standard normal distribution function. Significance can then be assessed with tables, published by Nikolai V. Smirnov [203].

3.4.5.2 Shapiro-Wilk test

The Shapiro-Wilk test is specifically designed for testing whether the random variable of a given sample follows a normal distribution. Test statistic is

$$W = \frac{b^2}{(n-1)\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ is the sample variance. b considers the structure of the sample in order to estimate variance. Based on the reordered vector of aggregated z-scores $\widetilde{\mathbf{Z}_\mathcal{G}}$ with $\widetilde{Z_{\mathcal{G}_1}} < \ ... \ < \widetilde{Z_{\mathcal{G}_n}}$, it is calculated by

$$b = a_1(\widetilde{Z_{\mathcal{G}_n}} - \widetilde{Z_{\mathcal{G}_1}}) + a_2(\widetilde{Z_{\mathcal{G}_{(n-1)}}} - \widetilde{Z_{\mathcal{G}_2}}) + ...$$

where a is a vector calculated by

$$a = \left((\boldsymbol{\psi}^\top \mathbf{V}^{-1}\mathbf{V}^{-1}\boldsymbol{\psi})^{-\frac{1}{2}}\right)\boldsymbol{\psi}^\top \mathbf{V}^{-1}$$

where $\mathbf{V}^{-1}$ is the inverse of the covariance matrix $\mathbf{Cov}(\widetilde{\mathbf{Z}_\mathcal{G}})$ and $\boldsymbol{\psi}$ is an n-element vector with $\psi_i = \Phi^{-1}(\widetilde{Z_{\mathcal{G}_i}})$. Here, $\Phi^{-1}(\widetilde{Z_{\mathcal{G}_i}})$ is the inverse function of the cumulative normal distribution function $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ with $\hat{\mu}$ being the mean estimated from the sample and $\hat{\sigma}^2$ being the variance estimated from the sample. The closer W is to 1 (equality of both variance estimators), the more likely it is that $\mathbf{Z}_\mathcal{G}$ follows a normal distribution. Critical values of W for a given significance level α and a given sample size n are provided with tables.

3.4.6 Enrichment analysis

Given a selection of features (e.g. genes) from a set, it can be interesting to test whether in this selection, there is an overrepresentation of features that share a certain characteristic (e.g. genes with a specific function). To this end, *enrichment analysis* is applied, based on hypergeometric tests.

3.4.6.1 Hypergeometric test

In order to find out whether there are significant *enrichments* of set members (e.g. groups of functionally related genes) in a given subset (e.g. set of differentially expressed genes), a hypergeometric test can be applied.

Let $\mathcal{G} = \{g_1, \ ..., \ g_m\}$ be a *gene universe*, given with the names of the genes of an expression matrix $\mathbf{X}_{m\times n}$. Furthermore, let $\mathcal{J} \subset \mathcal{G}$ be a subset of the universe, e.g. a set of genes that were found to be differentially expressed with one of the above described methods, and $J = |\mathcal{J}|$. Finally, let $\mathcal{K} \subset \mathcal{G}$ be another subset, covering all genes in $\mathcal{G}$ that share a certain functional relationship in terms of biology (e.g. genes involved in a certain process), and $K = |\mathcal{K}|$. It can now be interesting to find out whether there are more of those genes included in the set of differentially expressed genes than it would be expected randomly. The number of these cases of interest is called the number of *hits* $q = |\mathcal{J} \cap \mathcal{K}|$. q is assumed to follow a hypergeometric distribution. Hypergeometric test can now be applied, in order to find whether q is larger than expected; if so, this can be regarded as indicator that the process or function which is covered with the set of genes $\mathcal{K}$ can play an important role in one or multiple of those groups compared with differential expression analysis which yielded the set $\mathcal{J}$. Hypergeometric test is then based

on the probability mass function for a random variable Y

$$P(Y = q) = \frac{\binom{K}{q}\binom{J-K}{m-q}}{\binom{J}{m}}$$

which describes the probability to observe exactly q hits. P-value is simply calculated by $P(Y \geq q)$.

3.4.6.2 GO and KEGG enrichment

Several ontologies have been created in order to document the particular functions of genes, the most popular ones being MSigDB (molecular signatures database, often combined with GSEA), GO (gene ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) [192, 193, 194]. Genes are annotated with so-called *terms* which refer to a certain function or subgroup of genes.

Corresponding datasets can be easily imported into Bioconductor/R [195, 196, 197]. Given a set of differentially expressed genes, hypergeometric test is then applied to all sets of an ontology, summarizing genes which are linked to the same function. The analysis then unveils groups of functionally related genes that are overrepresented in the set of differentially expressed genes.

3.4.6.3 Model-based gene set analysis for gene ontologies

After enrichment analysis was applied for a multitude of sets, multiple testing correction should be applied in order to derive significant results, respectively. However, many ontologies, in particular GO, exhibit tree-like structures. Accordingly, an ontology set might be entirely included in another, larger ontology set which is located on a higher level of the GO tree. In practice, this often leads to redundant results, because hypergeometric test yields small p-values for a specific term and also corresponding less specific terms. This is called the *inheritance problem* [198]. In addition, due to the tree-like structure, ordinary p-value correction is not possible; accordingly, established R packages implementing GO enrichment analysis do not perform p-value correction.

Several approaches are addressing the inheritance problem, including the elim algorithm [199] which eliminates all genes from a set that are included in a subset which has already been added to the "list of significant sets". Still, elim and similar algorithms are based on sequential testing procedures. An alternative concept was introduced with an approach called GenGO [200]. Here, a model is set up for simultaneous fitting of all ontology terms; optimization of the model is performed in order to determine the most likely *combination of active terms* [198].

An even more advanced method is provided with *model-based gene set analysis* (mgsa), a Bayesian approach using a Markov Chain Monte Carlo (MCMC) algorithm [198]. Its Bayesian network model consists of three layers, a *term layer*, in which each node represents one term, a *hidden layer* in which each node represents one gene, "capturing" the output of the term layer and an *observation layer* in which each node represents one gene again, which is connected to the

hidden layer in a one-to-one-fashion. All three layers are boolean. The term layer and the hidden layer are connected according to the annotation in the respective database; the observation layer reflects whether the respective gene is differentially expressed or not. Nodes of the term layer are then modelled to a Bernoulli distribution.

According to the concept of mgsa, posterior distributions are provided instead of p-values here. In practice, mgsa has been shown to outperform GenGO [198] and yields less but more accurate results than normal GO term enrichment, respectively.

3.4.7 Cross-validation for gene signature extraction

Commonly, differential expression analysis is performed in order to identify a set of genes that are dysregulated in the respective group. However, if the focus is not put on the individual genes but more on the general purpose of *characterizing* or *classifying* a group of samples, there is great interest in the extraction of a *robust gene signature*.

Robustness of an estimation or inference method refers to its *sensitivity to outliers*. For instance, consider the sample $\mathbf{q} = (q_1, ..., q_u)^\top$ of a random variable Q, with values ordered increasingly. Then, two different estimators for the average of Q are given with the sample mean $\hat{\mu}_Q^{(1)}$ and the sample median $\hat{\mu}_Q^{(2)}$

$$\hat{\mu}_Q^{(1)} = \frac{1}{u}\sum_{i=1}^{u} q_i, \quad \hat{\mu}_Q^{(2)} = \begin{cases} q_{\frac{u+1}{2}}, & \text{if } 2 \nmid u \\ \frac{1}{2}\left(q_{\frac{u}{2}} + q_{\frac{u}{2}+1}\right), & \text{otherwise.} \end{cases}$$

The sample median is more *robust* against outliers than the sample mean, because more incorrect observations or outliers are required in order to cause a large deviation of the estimator from the real mean of the random variable. More precisely, its *breakdown point* is greater. The breakdown point ϵ_u^* of an estimator is the proportion of samples that must be outliers in order to make the estimation arbitrary. Since one outlier is enough to achieve this for the sample mean, $\epsilon_u^*\left(\hat{\mu}_Q^{(1)}\right) = 1/u$, whereas for the sample median $\epsilon_u^*\left(\hat{\mu}_Q^{(2)}\right) = 1/2$ which is also the greatest possible breakdown point value. The idea of robustness can also be transferred to gene signatures. In order to reduce a set of genes (e.g. that were found to be differentially expressed) to a robust signature allowing for a reliable classification of groups, genes with outliers must be discarded from the set. This *training procedure* can be performed with a k-fold cross-validation. Here, the dataset is split up into a training set and a test set, and validation is performed in k steps. For instance, if $k = 5$ cross-validation is a 5-step process and in each step, the classifier is trained on 80% of the dataset and then tested on 20% of the set. Fig. 3.3 illustrates the process. Algorithm 1 provides an implementation of a k-fold cross-validation for gene signature extraction, given a set of differentially expressed genes. In this specific case, training is based on t-tests assessing differential expression and testing is based on fold change computation.

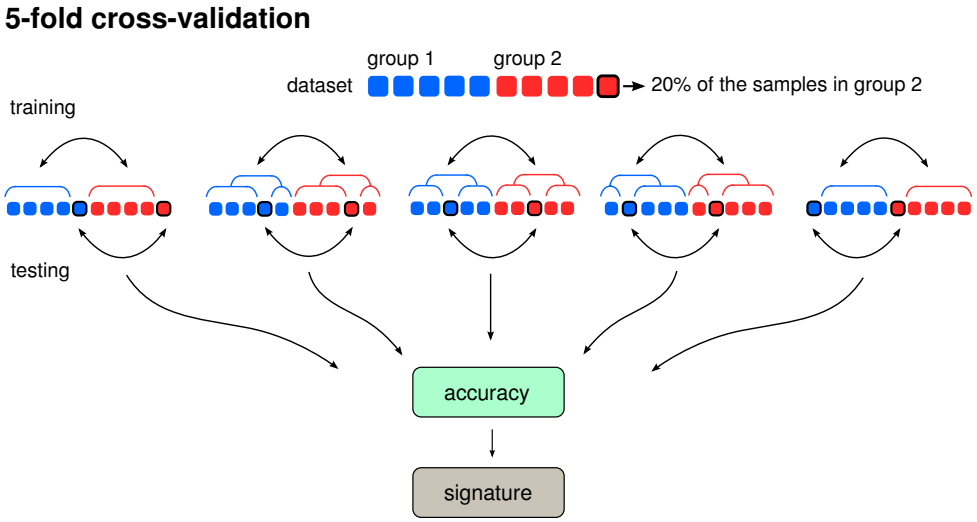

Figure 3.3: Illustration of the cross-validation procedure.

3.4.8 Identification of active subcircuits in regulatory networks

Given a network with gene regulatory interactions describing a distinct biological process, one might be interested in finding out whether the process described by this network is mediating the switch of gene expression which is observed between two different states recorded with gene expression profiling. Moreover, it can be interesting to find out whether there is a subcomponent in the network which is specifically active.

Ideker et al. have developed a method which can be applied to address this issue [202]. The underlying assumption is that if the players of a subcomponent or even of a whole network are differentially expressed in gene expression data with two distinct states, then it is likely that the dynamics reflected by the considered network or subcomponent are driving the expressional change which is observed in the data. For instance, consider a gene expression matrix $\mathbf{X}_{m \times n}$ (m genes, n samples) and two disjoint *groups of samples*, $\mathcal{A}_1 \subseteq \{1, \ldots, n\}, \mathcal{A}_2 \subseteq \{1, \ldots, n\}, \mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. Assume that sample groups $\mathcal{A}_1, \mathcal{A}_2$ here refer to two different steady states and that there is a transcriptional switch driving the change from the state reflected by samples in $\mathcal{A}_1$ to the state reflected by samples in $\mathcal{A}_2$; this could be the case for two different time points where gene expression was recorded. Then, if there are several interconnected players in the provided prior network which are also differentially expressed between the two states, it can be likely that this network component is the cause of the transcriptional switch which is observed.

Given a fixed prior network with u nodes (genes) and a set of corrected p-values $\mathcal{P} = \{p_1, \ldots, p_u\}$ referring to the significance of differential expression of the corresponding genes between samples $\mathcal{A}_1$ and $\mathcal{A}_2$. p-values can be converted into z-scores $z_i = \Phi^{-1}(1 - p_i)$ with Φ^{-1} being the inverse function of the standard normal cumulative distribution function. z-scores then follow a standard normal distribution. The so-called *aggregated z-score* for a set of p-values

Algorithm 1 k-fold cross-validation for an expression matrix (of preselected features)

procedure GeneSignatureCrossValidation
  $X_{m\times n} \leftarrow$ expression matrix of group 1 samples
  $Y_{m\times n} \leftarrow$ expression matrix of group 2 samples
  $A_{m\times k} \leftarrow$ accuracy matrix initialized with zeros
  $\theta_{FDR} \leftarrow$ FDR-value threshold
  $\theta_{fc} \leftarrow$ fold change threshold
  $\theta_{acc} \leftarrow$ accuracy threshold
  $S \leftarrow \emptyset$
  for $i := 1$ .. k **do**
    $s \leftarrow (i-1)(n/k)+1$ .. $i(n/k)$
    $X_{test}, X_{train} \leftarrow X[,s], X[,-s]$
    $Y_{test}, Y_{train} \leftarrow Y[,s], Y[,-s]$
    $excl \leftarrow \emptyset$
    for j := 1 .. rows(X) **do**
      $p \leftarrow$ t-test$(X_{train}[j,], Y_{train}[j,])$
      $fc \leftarrow \mu(X_{train}[j,]) - \mu(Y_{train}[j,])$
      if $p \geq \theta_{FDR}$ or $|fc| < \theta_{fc}$ **then**
        $excl \leftarrow \{excl \cup j\}$
    for j := 1 .. rows(X) **do**
      for h := 1 .. cols(X_{test}) **do**
        $acc \leftarrow 0$
        $fc \leftarrow X_{test}[j,h] - Y_{test}[j,h]$
        if $|fc| \geq \theta_{fc}$ **then**
          $acc \leftarrow a+1$
      $A[j,i] \leftarrow \frac{acc}{k}$
    $X \leftarrow X[-excl,],\ Y \leftarrow Y[-excl,],\ A \leftarrow A[-excl,]$
  for j := 1 .. rows(A) **do**
    if $\frac{\sum_{i=1}^{k} A[j,]}{k} \geq \theta_{acc}$ **then**
      $S \leftarrow \{S \cup j\}$
  return S

$\mathcal{P}' \subseteq \mathcal{P}$ is also following a standard normal distribution

$$Z_{\mathcal{P}'} = \frac{1}{\sqrt{|\mathcal{P}'|}} \sum_{p_i \in \mathcal{P}'} \Phi^{-1}(1-p_i) \sim \mathcal{N}(0,1)$$

In principle, random sampling could now be applied in order to find the largest aggregated z-score for the given set of p-values, regardless of the *size* of the set. To this, random sets of genes p-values are drawn from the network repeatedly, and the aggregated z-score is calculated. The set $\mathcal{P}^*$ with the best aggregated Z score is regarded as active subcomponent if the score does not exceed a predefined threshold.

However, several problems have to be considered.

1. The number of possible candidate gene sets is 2^u for an u-node network. Accordingly, optimization must be performed in an organized, algorithmic fashion.

2. Not each gene set is actually a subcomponent. Only if each gene can be reached from all other genes in the set, the set can be regarded as subcomponent (*conntected component* in terms of graph theory). To consider subcomponents only, the *topology* of the network must be used.

3. The number of possible subcomponents depends on the network's topology but can also be 2^u in the worst case.

4. For a given network and corresponding p-values, the expected average of the aggregated z-score differs for different set sizes, i.e. for $\mathcal{P}_1, \mathcal{P}_2$ with $|\mathcal{P}_1| \neq |\mathcal{P}_2|$, $\mathbf{E}(Z_{\mathcal{P}_1}) \neq \mathbf{E}(Z_{\mathcal{P}_2})$.

The last issue can be addressed easily: In order to correct calculated aggregated z-scores for randomly expected $Z_{\mathcal{P}'}$, Monte Carlo sampling is applied for each possible set size $k, 1 \leq k \leq u$. During optimization, calculated $Z_{\mathcal{P}'}$ scores can then always be corrected against these values. The corrected $Z_{\mathcal{P}'}$ is then computed by

$$S_{\mathcal{P}'} = \frac{Z_{\mathcal{P}'} - \hat{\mu}_k}{\hat{\sigma}_k}$$

for a given subcomponent of size k. Here, $\hat{\mu}_k$ and $\hat{\sigma}_k$ are the values derived from Monte Carlo sampling of random subcomponents of the same size.

The other issues can be addressed by optimizing the $S_{\mathcal{P}'}$ scores, using a *Simulated Annealing* algorithm. Here, subcomponents are optimized by toggling nodes with a temperature-function dependent probability stepwise until a predefined number of iterations has been performed.

3.4.9 k-means clustering

k-means is a method allowing to partition a set of samples into k clusters in a way that the sum of distances of cluster members to their corresponding cluster centers is minimal.

Given an expression matrix $\mathbf{X}_{m \times n}$ (m genes, n samples). Assume that samples should partitioned into k clusters. For k given clusters $\mathcal{S}_1$, ..., $\mathcal{S}_k$, $\forall i,\ 1 \leq i \leq k : \mathcal{S}_i \subseteq \{1, ..., n\}$ with $\forall i, j,\ i \neq j : \mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, sum of distances between *sample expression profiles* $\mathbf{X}[,1]$, ..., $\mathbf{X}[,n]$ and their corresponding centers is calculated by

$$J = \sum_{i=1}^{k} \sum_{s_{ij} \in \mathcal{S}_i} d(\mathbf{X}[, s_{ij}], \boldsymbol{\mu}_i)$$

where d is the L_2-norm and $\boldsymbol{\mu}_i = \overline{\mathbf{X}[, \mathcal{S}_i]}$ are the vectors of cluster means calculated using the element-wise sample mean for all cluster members. Optimization of J (and corresponding assingment of samples to clusters) is usually performed with *Lloyd's algorithm* [204].

Initially, random clusters $\mathcal{S}_1^{(1)}$, ..., $\mathcal{S}_k^{(1)}$ are formed and corresponding means $\boldsymbol{\mu}_1^{(1)}$, ..., $\boldsymbol{\mu}_k^{(1)}$ are calculated. The iteration step from t to $(t+1)$ now includes sample reassignment and mean

recalculation. Reassignment:

$$\mathcal{S}_i^{(t)} = \left\{ \ell \in \{1,\ ...,\ n\} : d(\mathbf{X}[,\ell], \boldsymbol{\mu}_i^{(t)}) \leq d(\mathbf{X}[,\ell], \boldsymbol{\mu}_{i'}^{(t)})\ \forall\ 1 \leq i' \leq k, i \neq i' \right\}$$

Mean recalculation:

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|\mathcal{S}_i^{(t)}|} \sum_{s_{ij} \in \mathcal{S}_i^{(t)}} \mathbf{X}[, s_{ij}]$$

Termination condition is $t > 1 \wedge \mathcal{S}_i^{(t)} = \mathcal{S}_i^{(t-1)}\ \forall i, 1 \leq i \leq k$.

k is the crucial parameter in k-means clustering. Established methods for determination of the best k are based mainly on functions $f(k)$ evaluating the corresponding clustering results. Here, the so-called *gap statistic* which analyzes change of the within-cluster dispersion $\hat{\sigma}_i^2/\hat{\mu}_i$ (for some cluster i) with increasing k, is commonly used [205]. The resulting distribution is corrected against the within-cluster dispersion expected for normally distributed random data. Then, k is selected according to the minimal gap statistic [205]. An implementation of k-means clustering with Lloyds's algorithm is available in R package stats [206], gap statistic calculation is available in R package cluster [207].

3.4.10 Survival analysis

Given a population of N subjects with known death events t_i at $t_1 \leq\ ...\ \leq t_N$ after one common initial event t_0, e.g. diagnosis. Then, the *Kaplan-Meier estimator* is applied to predict the probability for a random subject from the population to survive longer then a given time t. It is defined as

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

where d_i is the number of *death events at time* t_i and n_i is the number of *subjects under risk* until then (including death cases at t_i). The great advantage of the estimator is that it can be applied to censored data, i.e. studies in which survival information about certain subjects is lost (e.g. patients leaving the study). These cases, called *lost to follow-ups*, are subtracted from the number of *subjects under risk* n_i, but not included in the death cases d_i, respectively. Correspondingly, n_i is the number of subjects which are still alive with certainty. Example: Consider two time points t_i, t_{i+1}. If there are no lost to follow-ups between t_i and t_{i+1}, then $n_{t_{i+1}} = n_{t_i} - d_{t_i}$, and accordingly, $\hat{S}(t_{i+1}) = \hat{S}(t_i)\frac{n_{t_i} - d_{t_i} - d_{t_{i+1}}}{n_{t_i} - d_{t_i}}$. If, however, there is an intermediate time point t', i.e. $t_i < t' < t_{i+1}$ with 1 lost to follow-up, then, $n_{t_{i+1}} = n_{t_i} - d_{t_i} - 1$, and correspondingly, $\hat{S}(t_{i+1}) = \hat{S}(t_i)\frac{n_{t_i} - d_{t_i} - 1 - d_{t_{i+1}}}{n_{t_i} - d_{t_i} - 1}$. Still, t' itself is not part of the "list of events", because recalculation of the estimator is not required at t', since $d_{t'} = 0$ and thus $\hat{S}(t') = \hat{S}(t_i)$.

Survival statistics can be easily visualized in so-called *Kaplan-Meier curves* visualizing $\hat{S}(t)$ which allow direct comparison of multiple populations. In order to assess whether the expected survival is different for two populations, significance tests can be applied. Here, most commonly, log-rank test is used.

3.4.10.1 Log-rank test

Given two populations of subjects (e.g. patients), $X = \{x_1, ..., x_\ell\}$, $Y = \{y_1, ..., y_m\}$ and time points of events (in either group) $T = \{t_1, ..., t_N\}, t_1 \leq ... \leq t_N$. Let $\mathbf{n}_X, \mathbf{n}_Y$ be the N-element vectors of *subjects under risk* i.e. $n_{Xi}, n_{Yi}, 1 \leq i \leq N$ be the number of subjects in X, Y that have not had an event (death) or been censored at time point t_i, according to the definition in section 3.4.10; and define $\mathbf{n} := \mathbf{n}_X + \mathbf{n}_Y$. Additionally, let $\mathbf{d}_X, \mathbf{d}_Y$ be the N-element vectors of events (deaths), and correspondingly d_{Xi}, d_{Yi} be the number of events in X, Y taking place exactly at time t_i, and define $\mathbf{d} := \mathbf{d}_X + \mathbf{d}_Y$.

Now, if $\mathbf{d}_X, \mathbf{d}_Y, \mathbf{n}_X, \mathbf{n}_Y$ are given, the null hypothesis is that survival functions of X and Y are equal, i.e. proportions of events that have taken place and events that have not taken place are equal, $d_i/n_i = d_{Xi}/n_{Xi} = d_{Yi}/n_{Yi}$. Correspondingly, if the null hypothesis is correct, d_{Xi} follows a hypergeometric distribution with parameters n_i, n_{Xi}, d_i. In this regard also variance of d_X must match the hypergeometric distribution. Accordingly, the null hypothesis is

$$H_0 : \forall\, 1 \leq i \leq N : \ \mathbf{E}(d_{Xi}) = n_{Xi}\frac{d_i}{n_i} \ \wedge \ \mathbf{Var}(d_{Xi}) = \frac{d_i(n_{Xi}/n_i)(1 - n_{Xi}/n_i)(n_i - d_i)}{n_i - 1}$$

If H_0 is correct, i.e. d_{Xi} follows the described hypergeometric distribution then the *log-rank statistic* Z follows a standard normal distribution function

$$Z = \frac{\sum_{i=1}^{N} d_{Xi} - \hat{\mu}(d_{Xi})}{\sqrt{\sum_{i=1}^{N} \hat{\sigma}^2(d_{Xi})}}$$

Here, $\hat{\mu}(d_{Xi})$, $\hat{\sigma}^2(d_{Xi})$ are the estimated mean and the estimated variance, being calculated analogously to the terms in the null hypothesis. Null hypothesis can now be rejected at signficance level α if $Z > z_\alpha$ where z_α is the upper α-quantile of the standard normal distribution function.

3.4.10.2 Peto-Peto test

An alternative to the log-rank test was introduced by Peto and Peto [208]. Generally, the above described test can be regarded as a special case in the class of weighted log-rank tests. In this class of tests, a *weight function* $w(i)$ is applied to both the nominator and the denominator of the above described long rank statistic Z in a way that $Z = \frac{\sum_{i=1}^{N} w(i)(d_{Xi} - \hat{\mu}(d_{Xi}))}{\sqrt{\sum_{i=1}^{N} w^2(i)(\hat{\sigma}^2(d_{Xi}))}}$. For the Peto-Peto test, $w(i) := \hat{S}(t_i)$, accordingly, early events have more impact here [208].

3.4.11 Support vector machines

Support vector machines (SVM) are supervised learning models that can be applied for binary classification problems. In many cases, data is not separable linearly. Using the so-called *kernel*

trick, SVMs transform the original data into a higher dimensional vector space where the data is separable. SVMs are *maximum margin classifiers.*

Given a set of training data $\{\mathbf{x_i}, y_i\}$ with $i = 1, \ ..., \ n$, $\mathbf{x_i} \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Generally, any hyperplane in the d-dimensional input space can now be described by

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

where b is a constant and $\mathbf{w}$ is the vector which is orthogonal to the hyperplane. Then, a linear classification function$f$ is described by

$$f(x) = \text{sign}\,(\mathbf{w} \cdot \mathbf{x} + b)$$

ideally yielding the original group where $\mathbf{x}$ belongs to. Requiring a minimal distance of 1 of each sample to the hyperplane, we can define the canonical hyperplane requiring

$$y_i(\mathbf{w} \cdot x + b) \geq 1 \qquad \forall\ 1 \leq i \leq n$$

Since the distance of the two margin hyperplanes $\mathbf{w} \cdot \mathbf{x} + b = 1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$ is $\frac{2}{||\mathbf{w}||}$, the optimization problem is now to minimize the norm $||\mathbf{w}||$ subject to the above conditions. This minimization problem corresponds to the minimization of $\frac{1}{2}||\mathbf{w}||^2$ which eliminates the square root from the norm. In order to include the constraints in the optimization problem, Lagrange multipliers must be introduced. For a two-dimensional function $f(x, y)$ that should be optimized with the constraint $g(x, y) = c$ where g is another two-dimensional function (that should be constant here), the conditions $\nabla_{x,y} f = -\lambda \nabla_{x,y} g$ (tangency) and $\nabla_{x,y} g \neq 0$ must be satisfied, where λ is the so-called *Lagrange multiplier* and

$$\nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

The problem can be formulated using the Lagrange function which is here defined as

$$\Lambda(x, y, \lambda) := f(x, y) + \lambda \cdot (g(x, y) - c)$$

The optimization problem is now to solve the equations

$$\nabla_{x,y,\lambda} \Lambda(x, y, \lambda) = 0$$

The third equation guarantees the original constraint $g(x, y) = c$ to be fulfilled and the two first equations guarantee $\nabla_{x,y} f = -\lambda \nabla_{x,y} g$. If the Lagrange multipliers have been derived for the n original constraints from above, the original optimization problem can be formulated as

$$\text{argmin}_{\mathbf{w},b} \text{max}_{\lambda \geq 0} \left\{ \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \lambda_i \left(g_i(y_i, \mathbf{x_i}) - 1\right) \right\}$$

where $g_i(\mathbf{x_i}, y_i) = y_i(\mathbf{w} \cdot x + b)$. According to the *Karush-Kuhn-Tucker condition* for stationarity, the solution is a linear combination of training vectors $\mathbf{x_i}$

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x_i}$$

which also corresponds to the geometric definition of the optimal hyperplane which is based on the vectors. It can be shown that

$$\lambda_i(y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1) = 0 \qquad \forall 1 \leq i \leq n$$

which means that $\lambda_i \neq 0$ only for vectors $\mathbf{x_i}$ whose functional distance is smaller than 1 to the hyperplane. These vectors are termed *support vectors*.

After optimization, b is identified by calculating

$$b = -\frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}^+ + \mathbf{w} \cdot \mathbf{x}^-\right)$$

where $\mathbf{x}^+$ are those support vectors for which $\mathbf{w} \cdot \mathbf{x}^+ + b = 1$ and $\mathbf{x}^-$ are those for which $\mathbf{w} \cdot \mathbf{x}^+ + b = -1$.

As many datasets are not separable linearly, the *kernel trick* is applied. Here, input vectors $\mathbf{x_i}$ are first transformed using a predefined *kernel function* and optimization described above is performed with the transformed values. A function $K : X \times X \to \mathbb{R}$ is a *kernel* if there is an *inner product space* $(F, \langle \cdot, \cdot \rangle)$ and a function $\phi : X \to F$ with $K(x, y) = \langle \phi(x), \phi(y) \rangle \, \forall x, y$. F is also called *feature space* and ϕ *feature mapping*. According to *Mercer's theorem*, the feature space is not required to be known, because all kernel functions are characterized by clear properties. Several kernel functions are established as standards in SVM application, the most popular ones being the linear kernel

$$K(x, y) = \langle x, y \rangle$$

the polynomial kernel

$$K(x, y) = \langle x, y \rangle^d$$

and the radial basis function kernel

$$K(x, y) = \exp\left\{-\frac{||x - y||^2}{2\sigma^2}\right\}$$

A powerful SVM implementation is available in R package e1071 [210].

3.4.12 Correlation networks

Correlation networks provide a simple method for identification of potential interactions between a set of preselected features. Given two random variables X, Y, Pearson correlation coefficient

is defined as

$$\varrho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where $cov(X,Y)$ is the covariance of X and Y and σ_X, σ_Y are the standard deviations of X and Y. Given samples $\mathbf{x} = \{x_1, ..., x_n\}^\top$, $\mathbf{y} = \{y_1, ..., y_n\}^\top$ of two random variables X, Y, the estimator for the Pearson correlation coefficient is calculated by

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

According to this definition, r_{XY} estimates the linear relationship of two random variables and, as ϱ_{XY}, can vary between -1 and 1. An interaction network can be assembled based on pairwise correlation coefficients between any two genes; a significance test for the coefficient can then be applied in order to derive the set of edges which are considered as "significant interactions", using a hard cut-off threshold:

For two independent random variables X, Y which are following normal distributions, ϱ_{XY} is following a t-distribution with $(n-2)$ degrees of freedom with n being the sample size of X and Y [213]. To test for significant correlation of X and Y, the null hypothesis is

$$H_0 : \ \varrho_{XY} = 0$$

and the alternative hypothesis is

$$H_1 : \ \varrho_{XY} \neq 0$$

Corresponding test statistic is calculated by

$$t^* = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$$

H_0 is rejected at a significance level α if $|t^*| > t_{\alpha/2}$ with t_α being the α^{th} upper percentile of the corresponding t-distribution with $(n-2)$ degrees of freedom.

However, several limitations of this approach must be considered when Pearson correlation based networks are calculated:

1. If X and Y are regarded as *significantly correlated* based on the test result, this information does not include the influence of any third variable which can impact the correlation coefficient, so-called *confounders*. Consider a transcription factor represented by variable Z that controls expression of both genes (corresponding to X and Y), then $|r_{XY}|$ can be large though there is no interaction between these genes. Here, Z would be the confounder. Generally, similar scenarios must be expected to occur in a large number in our dataset.

The problem of removing confounding effects has been addressed with partial correlations. A method allowing computation of partial correlations even for cases with a large number of variables and a small number of samples has been developed by Strimmer et al. [212], based on the inverse variance lemma by Whittaker [201]. Though this method has been successfully applied for metabolomics data in the past [209], application on genome-wide scale is not feasible, due to the extremely large variable number.

2. The complexity of the resulting correlation network is regulated according to the selection of the significance cut-off, i.e. the number of resulting edges is chosen arbitrarily.

3. Generally, if two genes "interact", e.g. a transcription factor controlling a target gene, corresponding reactions are following complex kinetics, e.g. Michaelis-Menten kinetics. Still, for large time intervals between different points of measurement, the exact underlying reaction kinetics can be disregarded.

4. Non-directionality: The direction of regulation cannot be identified with Pearson's correlation coefficient.

5. Hard cut-off threshold: Selection of a hard cut-off threshold is arbitrary. Interactions which are slightly under this threshold remain undetected which can lead to distortion of the interaction network's topology. A popular approach addressing this issue is *weighted correlation network analysis* (WGCNA) which has often been applied to transcriptional data and is particularly useful for identification of highly connected nodes [215].

Even though there are several limitations for correlation-based networks and whilst the identified interactions can also be indirect, this approach allows us to find interaction *candidates*, and the list of candidates can then be evaluated, e.g. for enrichments. This allows us to identify pathways whose members are overrepresented among the interaction list.

After the above described steps have been applied, the so-called *relevance network* is extracted. This relevance network can be analyzed like an *undirected graph*. A *graph* is a 2-tuple $G = (V, E)$ where $V = \{v_1, ..., v_n\}$ is a set of *vertices* or *nodes* and $E = \{\{e_{11}, e_{12}\}, ..., \{e_{m1}, e_{m2}\}\}$ is a set of *edges* or *links*. G is called *undirected* if the set E is unordered. n is the *order* of the graph (number of nodes) and m is the size of the graph (number of edges). The *degree* $deg(v_j)$ of a node v_j is the sum of the number of edges starting at v_j and ending at v_j: $deg(v_j) = |\{e_i \in E : e_{i1} = v_j\}| + |\{e_i \in E : e_{i2} = v_j\}|$ (loops are counted twice). The *degree distribution* in biological networks has been shown to be *scale-free*, i.e. the degrees follow a power-law distribution [216]. Based on the observed degrees in the relevance network, probability that an additional randomly inserted edge will connect nodes v_i and v_j is calculated by $\frac{deg(v_i)deg(v_j)}{\sum_{k=1}^{m} deg(v_k)}$, assuming $\max_i w_i^2 < \sum_{k=1}^{m} deg(v_k)$ [217]. Accordingly, an edge between v_i and v_j is $deg(v_i)deg(v_j)$ times more likely to occur than an edge between a pair of two nodes with degree 1. This aspect has to be considered with care when the relevance network is analyzed. Highly connected nodes are also called *hubs*.

Chapter 4

Results

4.1 Temporal course of acute pancreatitis

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258]. Stainings were performed by Nora Behler and Sina Fritzsche.

In order to investigate the natural inflammatory response of the pancreas, acute pancreatitis was induced with consecutive caerulein injections in wild-type mice, as previously reported [171, 219]. As previous studies did not describe pancreatic regeneration with a high temporal resolution, histological changes and transcriptional profiles were analyzed for a multitude of time points (12 time points between 3 hours to 14 days after the first injection and control samples), as described in section 3.1. The purpose of this high-resolution design of the experiment is to uncover the dynamic changes of transcriptional activity and morphology after inflammatory injury.

4.1.1 Regeneration in wild-type mice is characterized by three stages: inflammation, regeneration and refinement

4.1.1.1 Histological observations suggest three predefined stages

Histological analysis revealed that acute pancreatitis is a self-dissolving process which can be divided into three phases which were termed “inflammation” (from 3 to 36 hours), “regeneration” (from day 2 to day 6) and “refinement” (from day 7 to day 14). Further examined samples were able to be assigned post-hoc to the identified phases. At 3 hours, pancreata were edematous and infiltrated by immune cells (Fig. 4.1). Vacuolization of acinar cells with scattered apoptotic cells was observed. Formation of acinar-to-ductal metaplasia (ADM) or tubular complexes (TCs) was initiated. As the inflammatory process progressed, immune cell infiltration and ADM formation became more pronounced (Fig. 4.1). This process was completed at 36 hours, where atrophic acini and ADM lesions with apoptosis were observed in the pancreatic parenchyma. Hereafter, histological changes compatible with organ regeneration started to appear. This phase lasted

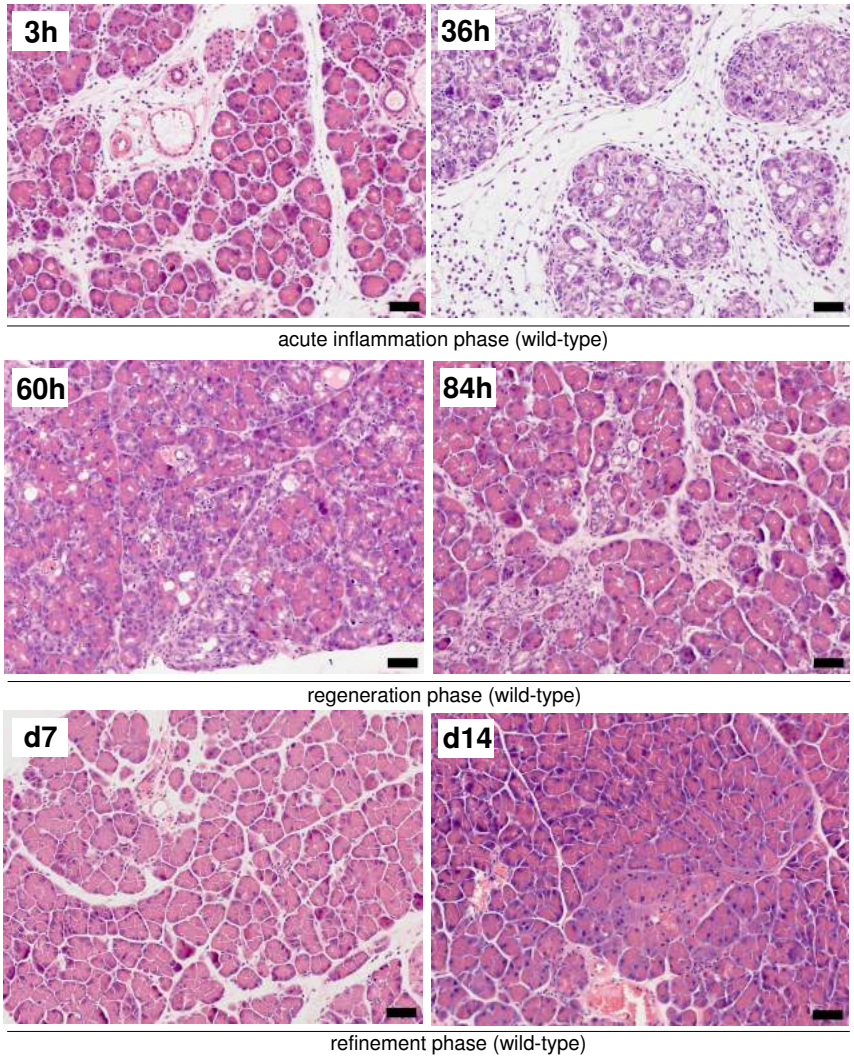

Figure 4.1: Representative H&E-stained sections with histological features of the inflammation, regeneration and refinement phases in WT pancreata (scale bars: 50 μm).

from day 2 to day 6 (Fig. 4.1).

From day 2 on, immune cell infiltration began to resolve; ADM lesions slowly disappeared and regular pancreatic parenchyma was gradually re-established. On day 6, only minimal inflammation, focal atrophic acini and ADM were observed. From day 7 to day 14, organ regeneration was almost complete and residual inflammation resolved (Fig. 4.1), accordingly the phase was termed refinement.

4.1.1.2 Transcriptional profiling confirms three stages

To test whether histologically predefined phases from above were also reflected by microarray data, hierarchical clustering with complete linkage distance function and L_2-norm was performed for wild-type mice. Samples were clustering according to histological observations showing high transcriptional similarity between samples from the same predefined phase as compared to samples from different phases. Particularly, four distinct clusters were observed, each representing

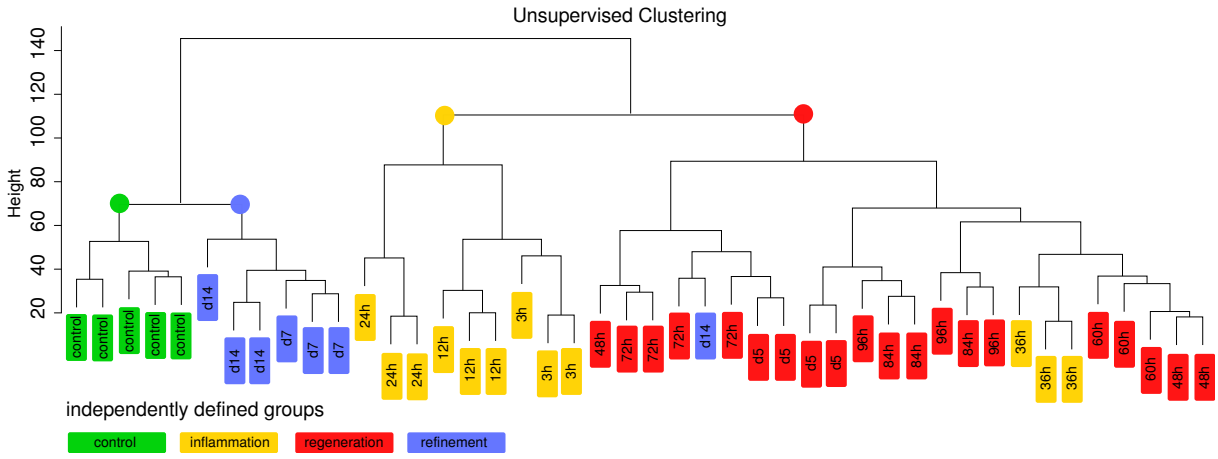

Figure 4.2: Unsupervised clustering of transcriptional profiles in WT mice: homogeneous clusters correspond to the histologically predefined phases.

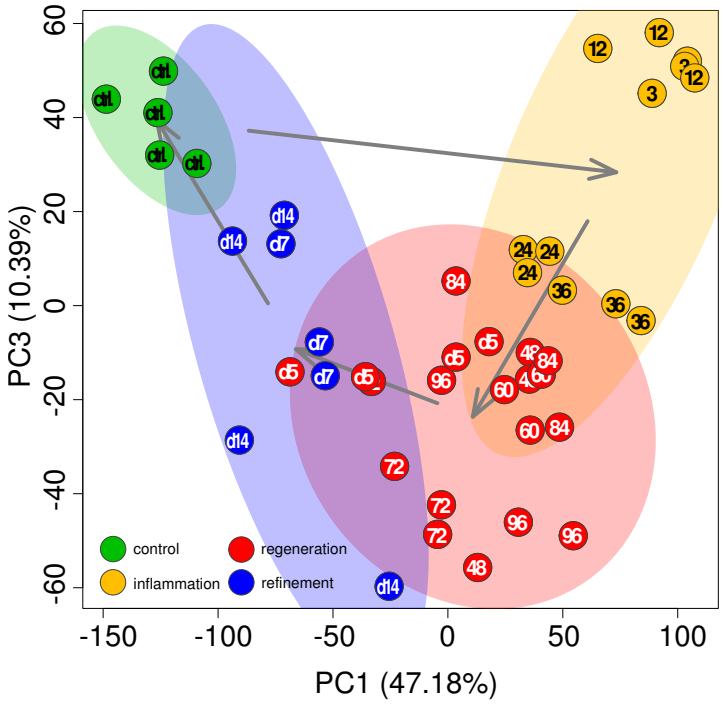

Figure 4.3: PCA biplot of transcriptional profiles with 95% confidence ellipsoids: the temporal course of inflammation in WT pancreata is a self-resolving process forming an end-to-end pattern (as underscored by arrows pointing to the centers of the groups' samples, right panel).

one group (three predefined phases and control samples). Except for the "regeneration cluster" containing all three 36h samples from the inflammatory phase and a single day 14 sample from the refinement phase, all remaining three clusters were homogeneous. Transcriptional profiles of samples from day 7 and day 14 were similar to control samples but formed a separate "refinement cluster". Clustering results are visualized with the dendrogram given in Fig. 4.2.

4.1.2 Temporal gradient unveils regeneration as a stepwise self-resolution process

Principal component analysis (PCA) was applied in order to visualize the temporal gradient of inflammatory response in wild-type mice. Using PC1 and PC3 which together explain 57.57% of the variance (PC1: 47.18%, PC3: 10.39%), biplot unveils that the course of normal pancreatic regeneration follows a cyclic "end-to-end" pattern as shown in Fig. 4.3. 95% confidence ellipsoids and arrows, pointing to the centers of the ellipsoids of the consecutive stage, illustrate that predefined phases form distinct sub-steps of the process. PC2, which explains 11.27%, was not chosen, because the temporal gradient is not visible here, though samples of the same stage are clustering together. In sum, we conclude that histological characterization of pancreatic regeneration as a three-phase process is confirmed on the molecular level.

4.2 Inflammatory response in Kras^{G12D} mice

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258]. Stainings were performed by Nora Behler and Sina Fritzsche.

Similarly, pancreatitis was induced in 8-week old P48$^{cre/+}$; LSL-Kras^{G12D}/+ mice (hereafter referred to as "Kras^{G12D} mice") and pancreata were collected at 9 time points for histological and transcriptional analysis (from 3 hours to day 14; as well as pancreata from NaCl-treated control mice; see also section 3.1). We did not use an inducible Kras^{G12D} system (e.g. Elastase-CreERT) because we aimed at collecting pancreatic tissues from mice with Cre recombination rather than not being able to strictly control time and efficiency of recombination in each animal. As previously recorded, only low grade pancreatic intraepithelial neoplasia (PanIN) lesions were detected in control mice.

4.2.1 Histological observations suggest sustained inflammation

In contrast to WT mice, no clear histological phases after caerulein injection could be identified. Within the first 36 hours (inflammation phase of WT mice), histological patterns of Kras^{G12D} mice were similar to those of WT mice. Focal vacuolization of acinar cells and incomplete ADM lesions were found; immune cells were scattered in the edematous pancreas. ADM formation was complete after 36 hours, accompanied by incipient lobular fibrosis (Fig. 4.4). During further progression, no clear organ regeneration was observed - which is in line with the published data [219, 171]. Particularly, ADM lesions remained in their ductal phenotypes. Immune cell infiltration persisted and progression of lobular fibrosis was seen (Fig. 4.4). On day 7 and 14, putative precursor lesions of PDAC - atypical flat lesions (AFLs [220]) and PanINs - were observed in atrophic areas (Fig. 4.4). Thus, activation of Kras^{G12D} together with injection of caerulein induces a phenotype without clear histological phases, as compared to the sequence of events in WT mice. A comparison of histological changes in both mouse strains is provided in

table 4.1.

Table 4.1: Histological observations in wild-type and Kras^{G12D} mice.

time	histological changes in WT mice	histological changes in Kras^{G12D} mice
0h	normal pattern	Focal vacuolisation of acinar cells, Acinar-to-ductal metaplasia (ADM)
3h	Vacuolisation of acinar cells Single apoptotic acinus cells Initial transformation to tubular complexes (TC)	Focal vacuolisation of acinar cells Dilated/cystic ducts
12h	Acini with larger lumina, more inflammation, initial TC Less apoptotic cells compared to 3 h Acini with larger lumina, More inflammation, Initial TC	Acini with larger lumina, More inflammation, Initial TC Edema, Pancreatic intraepithelial lesions (PanIN I; rare)
24h	No complete TC, single apoptotic cells, single mitosis More neutrophile granulocytes	Acini with larger lumina, No complete TC, More inflammation + Edema
36h	Edema, atrophy of acini, several complete TC Plenty of apoptosis	Complete TC, Lobular fibrosis Edema, Less inflammation
48h	Apoptosis and complete TC	Complete TC, Lobular fibrosis, Small areas with intact (regenerated) acinar cells
60h	Regeneration, Mild changes, only few TC, less inflammation, no apoptosis	Complete TC, Lobular fibrosis, Small areas with intact (regenerated) acinar cells
72h	Small foci with TC (10-15%) and atrophy, regeneration, minimal inflammation	
84h	Single lobuli with TC and atrophy Regeneration, minimal inflammation	Complete TC, Lobular fibrosis, Small areas with intact (regenerated) acinar cells Initial (atypical flat lesions) AFLs
96h	Less atrophie and TC (only small foci) compared to 72 and 84 h	
d5	Minimal inflammation, only small foci with TC and atrophy	
d6	Pattern comparable to day 5	

d7	pattern comparable to day 5 and 6; Small foci with TC and atrophy (less than day 5 and 6) Sparse inflammation	Lobuli with TC, Initial AFL Focal lobuli with acinar cell regeneration can be observed
d14	Almost normal pattern, almost complete regeneration	Lobuli with TC + Fibrosis, AFL Focal areas with solid AFL high-grade suspicious of initial carcinoma, Lobular architecture preserved, dilated ducts

4.2.2 Transcriptional profiling indicates disordered process

Unsupervised clustering of transcriptional profiles from $Kras^{G12D}$/caerulein pancreata was performed analogously to the above description (section 4.1.1.2). Consistent with the histological observations, this analysis failed to identify distinctive sub-clusters reflecting the clearly distinct phases in WT mice. All transcriptome profiles of $Kras^{G12D}$ pancreata after caerulein treatment formed two heterogeneous clusters, whereas non caerulein-treated control samples formed one cluster of similar expression profiles (Fig. 4.5).

4.2.3 PCA indicates end-to-open pattern

PCA was performed analogously to above. Here we did not find any combination of PC1 (covering 51.36% of the variance) and another principal component (PC2: 11.28%, PC3: 8.24%, ...) demonstrating an "end-to-end" pattern as described above. Instead, pronounced dissimilarity between control and late stage samples (day 7 and day 14) is observed and samples are disordered, indicating great heterogeneity. According to the similarity of all samples except controls, we identify and "end-to-open" pattern as visualized in Fig. 4.6.

4.3 Persistent inflammation in $Kras^{G12D}$ mice corresponds to the WT mouse inflammatory phase

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258]. Stainings were performed by Nora Behler and Sina Fritzsche.

We then went on to test the biological relevance of these histological and molecular findings and to assess proliferative contribution of different cellular compartments of the pancreas to inflammation, regeneration and refinement (and its absence in $Kras^{G12D}$ mice). First, the index of proliferating cells (BrdU-positive) and the number of immune cells (Cd45-positive) over time was determined in WT and $Kras^{G12D}$ mice. Two distinctive waves of proliferation were found in WT mice (Fig. 4.7 and 4.8, left panel). The first wave corresponded exactly to the time frame of the inflammation phase, whereas the second wave covered the regeneration phase. In $Kras^{G12D}$ mice, however, no clear segmentation of the proliferation wave was found (Fig. 4.7

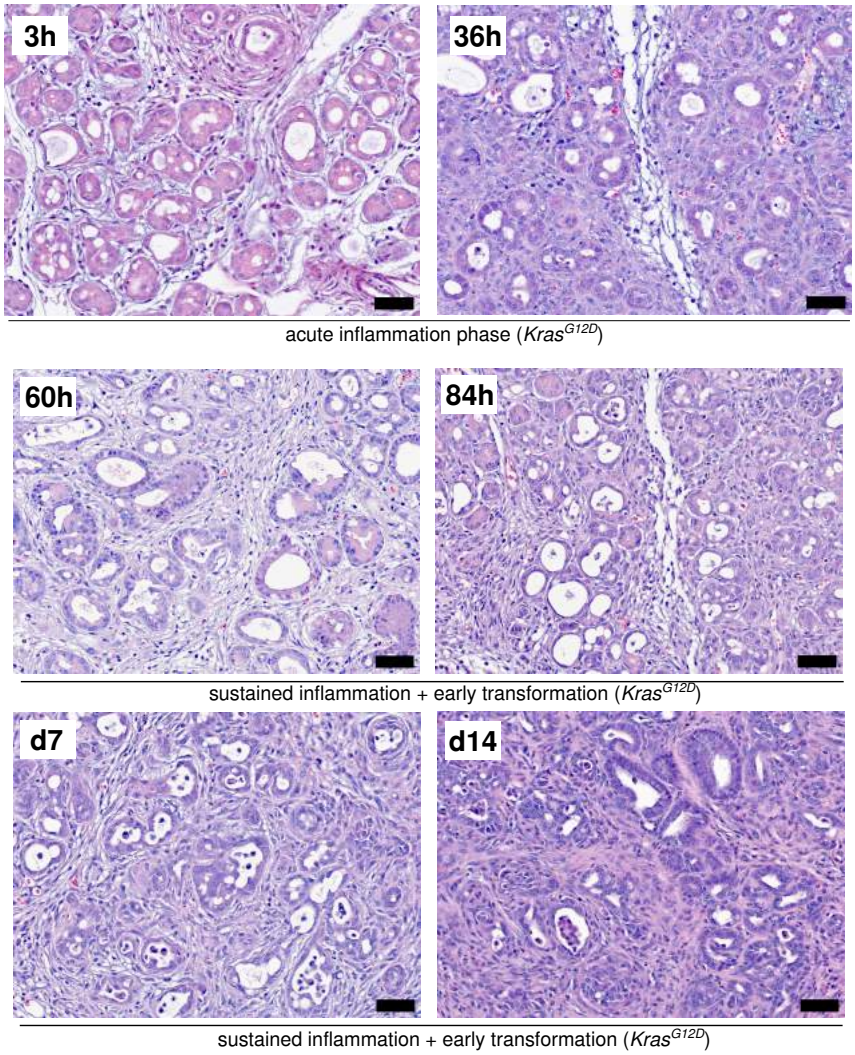

Figure 4.4: Representative H&E-stained sections with histological features of inflammation (3h - 36h), of the anticipated regeneration phase (48h - day 7) and of the anticipated refinement phase (day 7 to day 14; scale bars: 50 μm).

and 4.8, right panel). Instead, the organ proliferation index, describing the count of proliferative cells in one high-power field, continuously increased from 3 hours to day 7 until its decrease on day 14. Immune cell infiltration in WT mice increased dramatically in the inflammatory phase and then slowly decreased during regeneration and refinement (Fig. 4.9 and 4.10, left/upper panel). In Kras^{G12D} animals, the number of Cd45-positive cells steadily increased, indicating a non-resolving immune response within the observation period (Fig. 4.9 and 4.10, right/lower panel).

To allow for direct analysis of all transcriptional profiles taking account of two conditions, parallel factor analysis (PARAFAC) was applied [184]. In line with our previous findings, the analysis reproduced three predefined phases and the "end-to-end" pattern of the natural course of inflammation in WT samples (Fig. 4.11, left panel). The "end-to-open" pattern seen in Kras^{G12D} mice was also reproduced (Fig. 4.11, right panel). Areas of different phases were

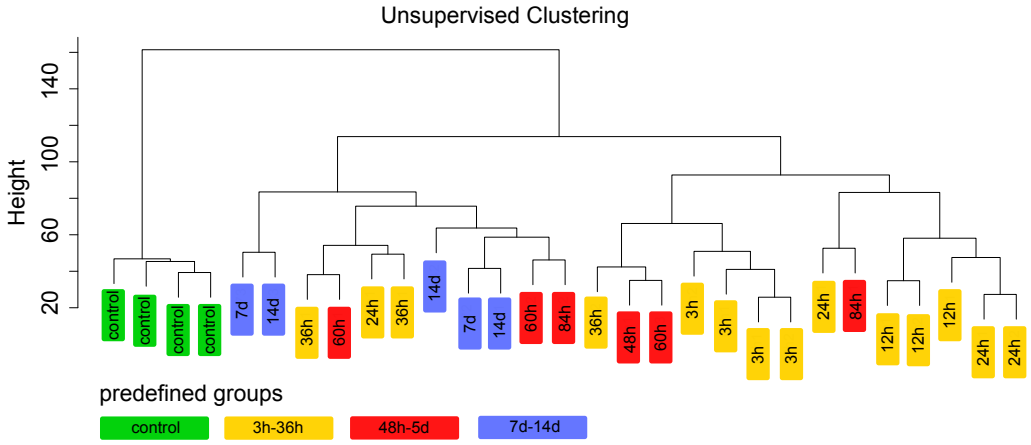

Figure 4.5: Clustering of Kras^{G12D} profiles confirms disordered regeneration as indicated by heterogeneous clusters.

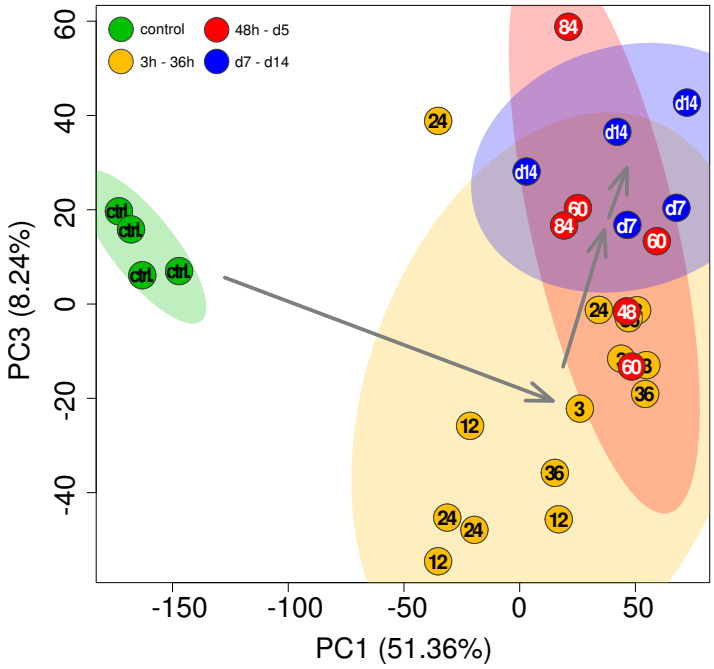

Figure 4.6: PCA biplot confirms that regeneration is blocked and forms and end-to-open pattern in Kras^{G12D} mice.

distinguished using confidence ellipsoids indicating a probability of 95% for respective samples of this phase to be located in this region. Interestingly, 74% (17/23) of Kras^{G12D} samples (except for controls) overlap with the confidence ellipsoid of the WT inflammation phase. Additionally, four of the remaining six outlying Kras^{G12D} samples were located closer to the center of the WT inflammation phase ellipsoid rather than to any other WT ellipsoid (based on Euclidean distance), indicating that 91% (21/23) of the Kras^{G12D} samples demonstrate highest similarity in their transcriptome profile to the WT inflammation phase. Remarkably, all ellipsoids from Kras^{G12D} samples (except controls) spatially overlap with the WT inflammation phase, which

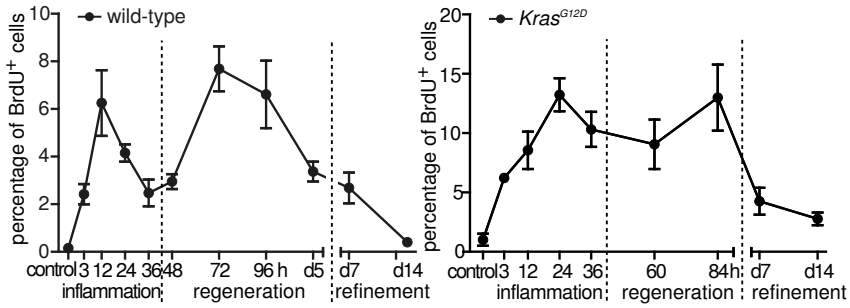

Figure 4.7: Proliferation indices in WT and Kras^{G12D} pancreata in different phases (or anticipated phases) of inflammation, as obtained by quantifying the percentage of BrdU-positive cells and hematoxylin-counterstained nuclei.

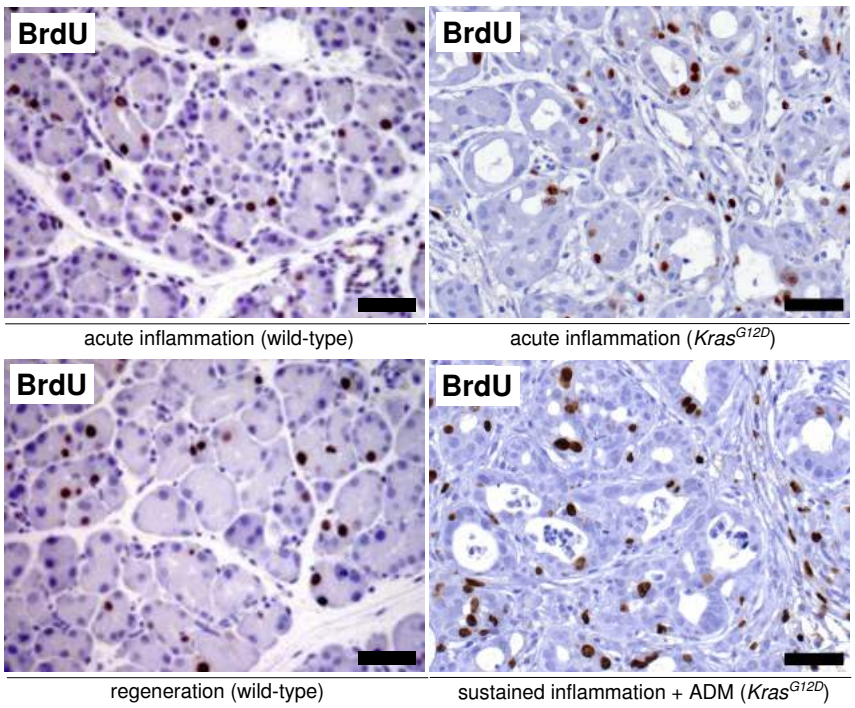

Figure 4.8: Representative IHC pictures show BrdU-positive cells in the different phases (or anticipated phases) of inflammation in WT (left panel) and Kras^{G12D} pancreata (right panel; scale bars: 50μm).

indicates that their transcriptional profiles are not clearly distinguishable.

These data demonstrate that the transcriptional signatures of Kras^{G12D} mice during regeneration and refinement correspond to the WT mouse inflammatory phase, showing a persistent inflammatory status.

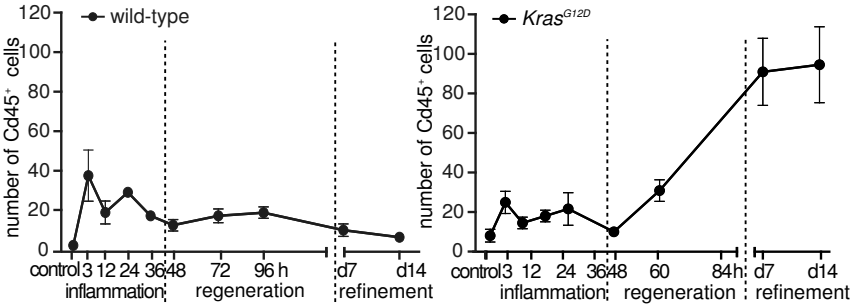

Figure 4.9: Quantitative data shows the time-dependent changes in the number of Cd45-positive cells in the different phases (or anticipated phases) in WT and Kras^{G12D} pancreata.

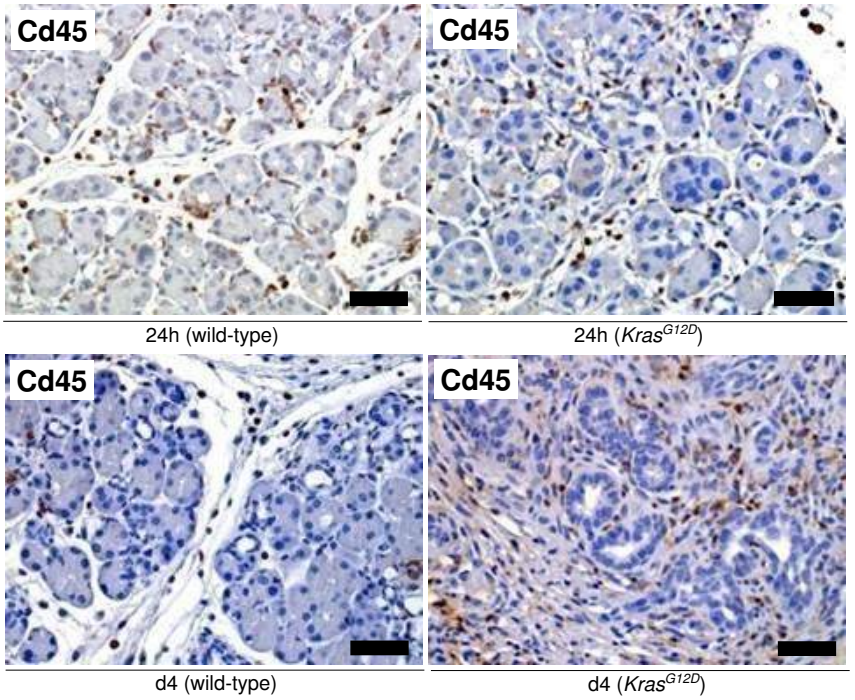

Figure 4.10: Representative IHC pictures show Cd45-positive cells in the different phases (or anticipated phases) of inflammation in WT (left panel) and Kras^{G12D} pancreata (right panel; scale bars: 50 μm).

4.4 Comparison of the identified time frames in wild-type mice and Kras^{G12D} mice

In order to compare the identified stages of pancreatic regeneration in wild-type mice to the anticipated stages in Kras^{G12D} mice, we performed a differential expression analysis (absolute FC > 2, FDR < 0.05, Benjamini-Hochberg correction) between each pair (wild-type inflammation vs. Kras^{G12D} 3-36h frame, wild-type regeneration vs. Kras^{G12D} 48h-d5 frame, wild-type refinement vs. Kras^{G12D} d7-d14 frame), according to section 3.4.3.3.

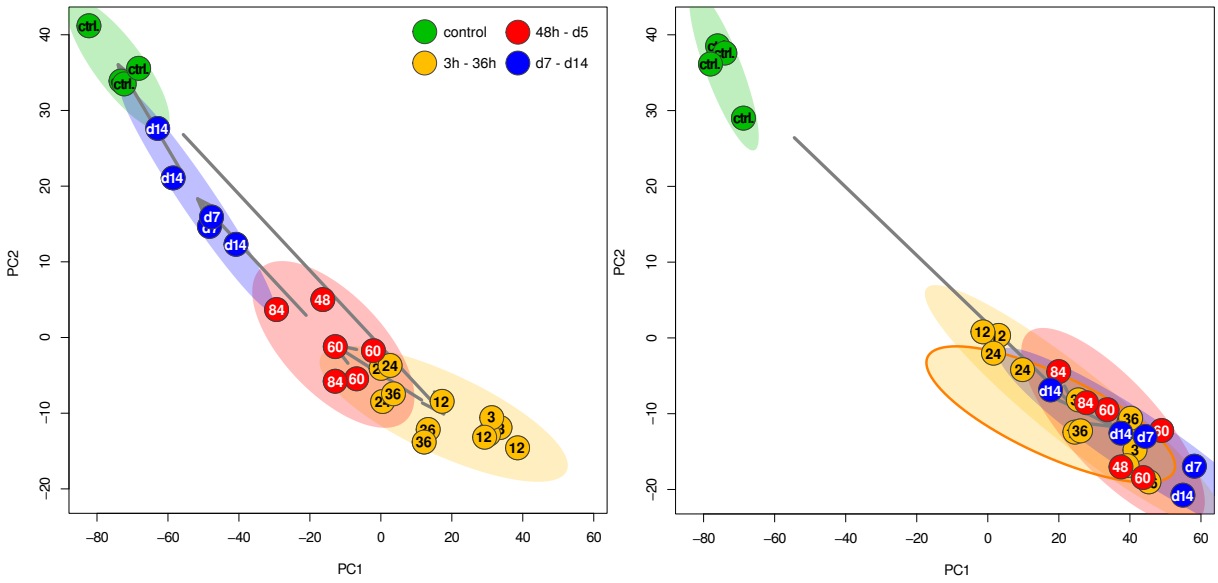

Figure 4.11: PARAFAC analysis: direct comparison of transcriptional profiles of WT and Kras^{G12D} pancreata on a genome-wide scale; 95% confidence ellipsoids reflect the previously defined phases; all samples from Kras^{G12D} mice (except for the controls) are highly similar to the inflammation phase samples from WT mice.

4.4.1 Comparison of inflammatory phase samples

According to our expectations, only few genes were found to be differentially expressed comparing samples from the inflammation phase (n=20). Those included Agr2 which was found upregulated in Kras^{G12D} and is known to be expressed in preneoplastic lesions and correlate with downregulation of p53 response [251] [252]. Also, Tspan8 which has been reported to be activated in various cancer types, was found upregulated [253]. Vsig1 which is known to be expressed in gastric, pancreatic and ovarian cancer types, was also found upregulated in Kras^{G12D} [254]. A list of all 20 genes is given in table 4.2.

4.4.2 Comparison of regeneration phase samples

Comparison of the regeneration phase showed that more than 1,000 genes were differentially expressed; consequently, GO term enrichment was performed using mgsa and KEGG term enrichment was performed, based on hypergeometric tests. The analysis shows that genes annotated by the GO term “inflammatory response” were overrepresented among differentially expressed genes, confirming our previous observations. In addition, “cytokine activity” related genes were found to be enriched, indicating that this subtype of inflammatory response-related genes is contributing significantly to the tissue transformation process. We also found genes annotated by the term “extracellular matrix” being enriched, indicating mesenchymal expansion. KEGG enrichment yielded similar results, but additionally identified “pancreatic secretion” related genes to be enriched, indicating an impaired function of the organ. Also “p53-signaling” related genes were found to be enriched. A detailed list of the identified terms and pathways is given in tables

4.3 and 4.4.

4.4.3 Comparison of refinement phase samples

For the refinement phase, more than 1,500 genes were found to be differentially expressed. Here, GO term enrichment and KEGG enrichment yielded similar results as for the regeneration phase, but also additional terms, including in particular the GO terms "cell adhesion" and "actin cytoskeleton reorganization" which could indicate epithelial-to-mesenchymal transitions. Among the identified KEGG pathways, we found "pancreatic cancer", which indicates that tissue transformation has already progressed in Kras-mutated mice at this time; also, we found "chemokine activity" to be enriched, suggesting a potential role of these molecules for further progression of tissue transformation. Furthermore, additional indicators for EMT could be found including terms "focal adhesion" and "tight junctions". Also, "TGFβ signaling"-related genes were found to be enriched. Lists of terms and pathways are given in tables 4.5 and 4.6.

Table 4.2: Genes differentially expressed between wild-type and Kras^{G12D} inflammation. Due to high similarity of both stages, the number of genes found to be differentially expressed is small, and no GO terms or KEGG pathways were found to be enriched.

Gene	log2 Fold Change	adj.P.Val
Tff1	-3.017362076	3.33E-07
Tspan8	-2.105441086	8.87E-07
Mcpt2	-3.840491666	1.57E-06
Gkn2	-4.031858542	1.64E-06
Slamf7	1.494899801	4.60E-06
Sprr2a1	-3.642284448	4.60E-06
2210407C18Rik	-3.70143059	4.60E-06
Gpnmb	2.471844476	4.60E-06
Gal	1.621528891	5.43E-06
Lgals4	-2.39862183	7.10E-06
Ly9	1.026111156	2.41E-05
Mpeg1	1.007947001	6.70E-05
Agr2	-3.278496002	0.000114556
Rgs17	-1.161133977	0.000199657
Ifit3	1.679235749	0.000209965
Gsta1	-3.231977051	0.000304763
Pllp	-1.153871986	0.000367459
Mgat4a	-1.022773585	0.000402205
Gcnt3	-2.527271481	0.000500391
Vsig1	-2.79386996	0.000624111

Table 4.3: GO terms that were found to be enriched for genes differentially expressed between wild-type regeneration and Kras^{G12D} anticipated regeneration (mgsa).

GOID	Count	Size	Posterior	Term
GO:0016323	31	146	0.9692666	Basolateral plasma membrane
GO:0008236	28	167	0.9533530	Serine-type peptidase activity
GO:0006954	38	214	0.9372224	Inflammatory response
GO:0001725	14	51	0.9246292	Stress fiber
GO:0031225	20	121	0.8599868	Anchored component of membrane
GO:0004364	8	24	0.8522842	Glutathione transferase activity
GO:0005125	30	171	0.8186694	Cytokine activity
GO:0008652	8	25	0.6077514	Cellular amino acid biosynthetic process
GO:0031012	35	201	0.5996836	Extracellular matrix
GO:0004806	6	15	0.5484838	Triglyceride lipase activity

Table 4.4: KEGG pathways that were found to be enriched for genes differentially expressed between wild-type regeneration and Kras^{G12D} anticipated regeneration.

KEGGID	Pvalue	Count	Size	Term
4060	1.43E-09	39	228	Cytokine-cytokine receptor interaction
4972	4.65E-08	22	98	Pancreatic secretion
4640	2.51E-06	17	78	Hematopoietic cell lineage
480	4.93E-06	13	50	Glutathione metabolism
4610	1.76E-05	15	72	Complement and coagulation cascades
4512	0.000116488	15	84	ECM-receptor interaction
5146	0.000262696	17	110	Amoebiasis
4670	0.000405606	17	114	Leukocyte transendothelial migration
260	0.002908055	7	33	Glycine, serine and threonine metabolism
5140	0.003492912	10	62	Leishmaniasis
4975	0.004112055	8	44	Fat digestion and absorption
4115	0.004955945	10	65	p53 signaling pathway
1100	0.005230295	84	1081	Metabolic pathways
4974	0.005283946	11	76	Protein digestion and absorption
561	0.00805778	8	49	Glycerolipid metabolism
982	0.008418386	10	70	Drug metabolism - cytochrome P450
4145	0.009847888	17	153	Phagosome

Table 4.5: GO terms that were found to be enriched for genes differentially expressed between wild-type refinement and Kras^{G12D} anticipated refinement (mgsa).

GOID	Count	Size	Posterior	Term
GO:0070062	381	1408	0.9508358	Extracellular vesicular exosome

GO:0008289	52	213	0.9290848	Lipid binding
GO:0003779	91	305	0.8960582	Actin binding
GO:0043065	72	255	0.8806580	Positive regulation of apoptotic process
GO:0045121	60	191	0.8541354	Membrane raft
GO:0002376	89	291	0.8511042	Immune system process
GO:0006935	39	106	0.8423098	Chemotaxis
GO:0006629	105	414	0.8096078	Lipid metabolic process
GO:0007155	130	470	0.7966920	Cell adhesion
GO:0031012	62	201	0.7906754	Extracellular matrix
GO:0031532	16	42	0.5164660	Actin cytoskeleton reorganization
GO:0004252	39	159	0.5146582	Serine-type endopeptidase activity

Table 4.6: KEGG pathways that were found to be enriched for genes differentially expressed between wild-type refinement and Kras^{G12D} anticipated refinement.

KEGGID	Pvalue	Count	Size	Term
5140	7.47E-11	33	62	Leishmaniasis
4145	1.15E-09	57	153	Phagosome
4060	5.37E-09	74	228	Cytokine-cytokine receptor interaction
4810	8.48E-09	68	205	Regulation of actin cytoskeleton
5150	8.78E-09	25	46	Staphylococcus aureus infection
4380	1.06E-08	44	111	Osteoclast differentiation
4670	2.74E-08	44	114	Leukocyte transendothelial migration
4062	9.26E-08	57	170	Chemokine signaling pathway
4666	3.91E-07	34	85	Fc gamma R-mediated phagocytosis
5323	4.35E-07	32	78	Rheumatoid arthritis
5144	4.25E-06	20	42	Malaria
480	7.19E-06	22	50	Glutathione metabolism
4210	2.08E-05	30	83	Apoptosis
5142	2.67E-05	34	100	Chagas disease (American trypanosomiasis)
4662	3.19E-05	26	69	B cell receptor signaling pathway
4142	5.57E-05	38	120	Lysosome
4620	5.64E-05	32	95	Toll-like receptor signaling pathway
4540	7.57E-05	28	80	Gap junction
5146	9.75E-05	35	110	Amoebiasis
5100	0.000161504	24	67	Bacterial invasion of epithelial cells
4650	0.000162279	33	104	Natural killer cell mediated cytotoxicity
4610	0.000209581	25	72	Complement and coagulation cascades
4972	0.00027028	31	98	Pancreatic secretion
5145	0.000348944	36	121	Toxoplasmosis
4510	0.000411873	52	195	Focal adhesion

4514	0.000460875	39	136	Cell adhesion molecules (CAMs)
600	0.000669752	16	41	Sphingolipid metabolism
4010	0.000984848	63	255	MAPK signaling pathway
1100	0.001058046	220	1081	Metabolic pathways
4530	0.001409668	35	125	Tight junction
5214	0.001571412	20	60	Glioma
4360	0.001640115	35	126	Axon guidance
4971	0.0018037	22	69	Gastric acid secretion
4115	0.001890891	21	65	p53 signaling pathway
4640	0.001968879	24	78	Hematopoietic cell lineage
5212	0.002216216	22	70	Pancreatic cancer
4350	0.00222498	25	83	TGF-beta signaling pathway
4664	0.002542365	23	75	Fc epsilon RI signaling pathway
5020	0.002621346	13	34	Prion diseases
5210	0.003031006	20	63	Colorectal cancer
4970	0.003283823	22	72	Salivary secretion
4512	0.005710118	24	84	ECM-receptor interaction
5219	0.00602403	14	41	Bladder cancer
5220	0.007290356	21	72	Chronic myeloid leukemia
5412	0.008634601	21	73	Arrhythmogenic right ventricular cardiomyopathy (ARVC)

4.5 Alterations of the cellular programs

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258]. Stainings were performed by Nora Behler and Sina Fritzsche.

4.5.1 Loss of exocrine maintenance characterizes inflammatory phase and early carcinogenesis

To address whether the similarity in transcriptional profiles of acute inflammation and of early carcinogenesis was reflected in the patterns of the proliferating cells, proliferation was analyzed in different cellular compartments. Co-stainings of the proliferation marker BrdU with cell-specific markers for pancreatic acinar (α-Amylase), progenitor-like (Sox9) and mesenchymal cells (α-SMA) were performed. This allowed for generating time- and cell type-specific quantitative data including the exact number of proliferating cells and its proportion to overall proliferation.

Acinar cells form the largest cell population of the exocrine compartment, which is specifically injured by caerulein injection. We thus expected this population of cells to compose the largest number of proliferating cells during repopulation. In WT pancreata, we observed two waves of cell replication, the first one peaking at around 24h, the second one on day 4 (Fig. 4.12, 4.13). The overall number of proliferating acinar cells during the regeneration phase was higher

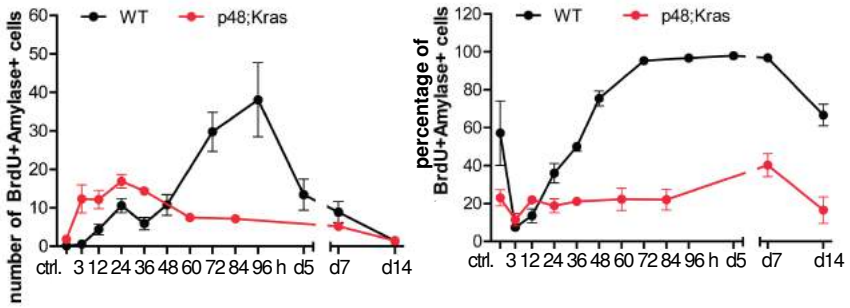

Figure 4.12: Number and percentage of proliferating α-amylase-positive cells in WT and Kras^{G12D} pancreata.

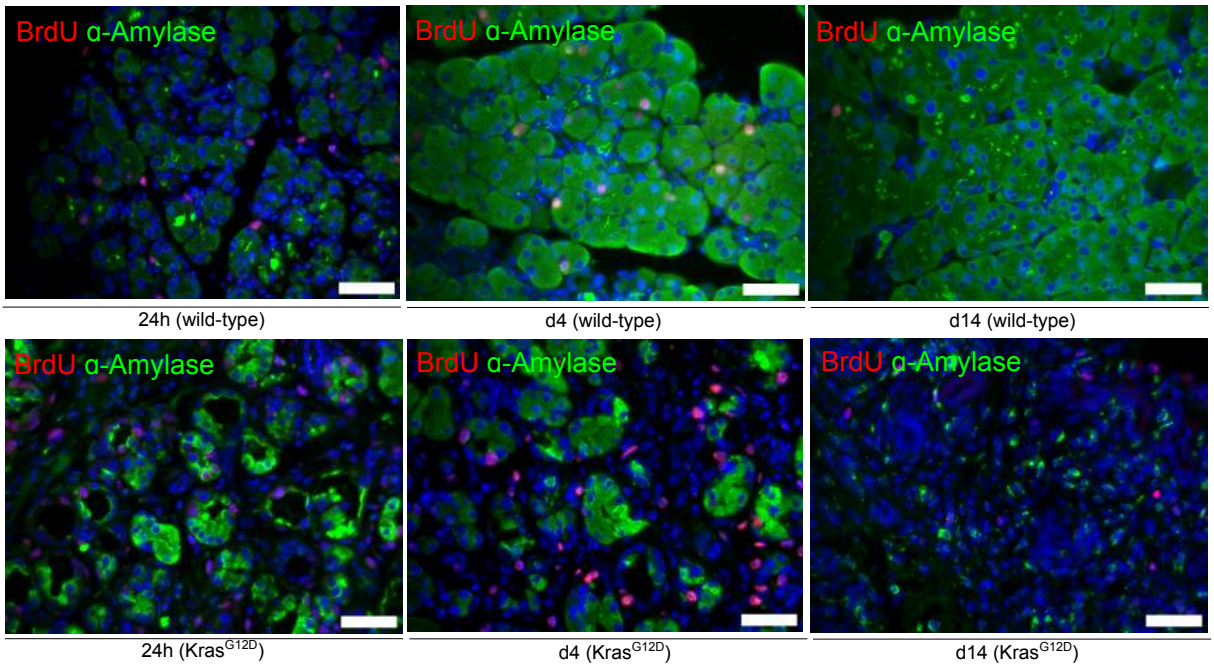

Figure 4.13: Representative IF of proliferating α-amylase-positive cells in WT (upper panel) and Kras^{G12D} (lower panel) pancreata in different phases (or anticipated phases) of inflammation; scale bars: 50 μm.

than that in the inflammatory phase ($p < 0.0001$; t-test, Fig. 4.15, left panel). Furthermore, only 25% of proliferating cells were acinar cells during the acute inflammation phase while this number increased to 92% during regeneration ($p < 0.0001$, t-test, Fig. 4.15, right panel). These data demonstrate that acinar cell proliferation is transiently compromised during the acute inflammation phase in WT pancreata and that, as previously described, many cells negative for α-amylase were proliferative [218]. Interestingly, the number of proliferating acinar cells in Kras^{G12D} pancreata during the first 36 hours (i.e. corresponding to the WT inflammation phase) was comparable to that in WT pancreata (Fig. 4.12, 4.13). However, during anticipated regeneration and refinement, the total number of proliferating acinar cells was significantly lower than in WT mice ($p = 0.0001$; t-test, Fig. 4.15, left panel). Overall, acinar cells accounted only

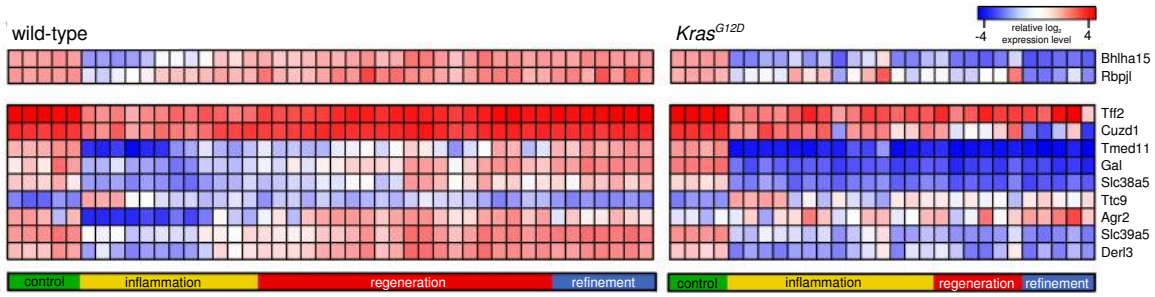

Figure 4.14: Heatmap of 11 genes linked to acinar cell homeostasis as indicated on the right; voxel color: transcriptional up- (red) and down-regulation (blue) as compared to the mean; each voxel indicates gene activity in one mouse sample arranged in chronological order; colored bar indicates control (green) and regenerative phases (yellow, red and blue).

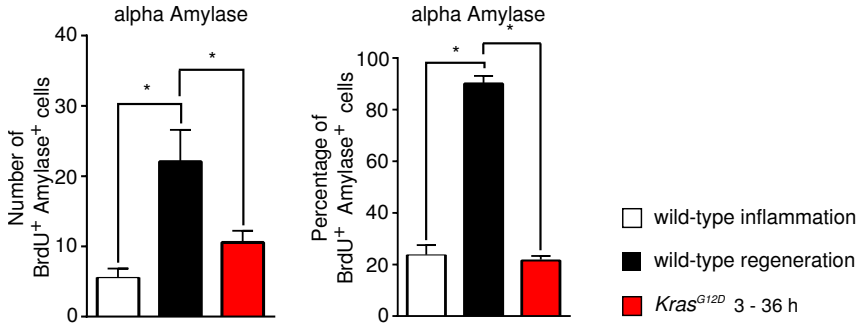

Figure 4.15: Absolute and relative amount of proliferating acinar cells in wild-type inflammation, regeneration and Kras^{G12D} samples from 3h to 36h.

for an average of 21% of all proliferating cells in Kras^{G12D} pancreata over the entire time course, which is similar to that of inflammatory phase (25%) but significantly lower than that of the regeneration (92%) phases in WT pancreata ($p < 0.0001$; t-test, Fig. 4.15, right panel).

From the quantitative data of the dynamic proliferation of acinar cells, we hypothesized that the molecular machinery governing the exocrine program may exhibit a similar pattern. To test this hypothesis, we extracted a list of 61 "acinar cell homeostasis" genes from the literature [218, 221, 222, 223, 224] and investigated the temporal courses of expression of these genes during different phases of inflammation and Kras^{G12D}-driven early carcinogenesis. This list included Bhlha15 (also known as Mist1), Xbp1, Nr5a2, Hnf1a and Gata6 and a set of 53 genes related to the transcriptional complex switch between Ptf1a-Rbpj and Ptf1a-Rbpjl. Based on the WT dataset, differential expression analysis was performed, with three groups being defined by the control samples, inflammation phase samples and regeneration phase samples. Specifically, we extracted genes that were differentially expressed during inflammation as compared to control samples and regeneration phase samples. Of 61, eleven genes were differentially expressed between inflammatory phase and regeneration phase or control samples in WT mice ($FDR < 0.05$, absolute $FC > 2$, two independent t-tests, Benjamini-Hochberg correction, Fig. 4.14, left panel), differential expression analysis with limma. This proportion (11/61) was significantly higher than expected for a random gene set (hypergeometric test: $p = 5.98 \times 10^{-10}$). We

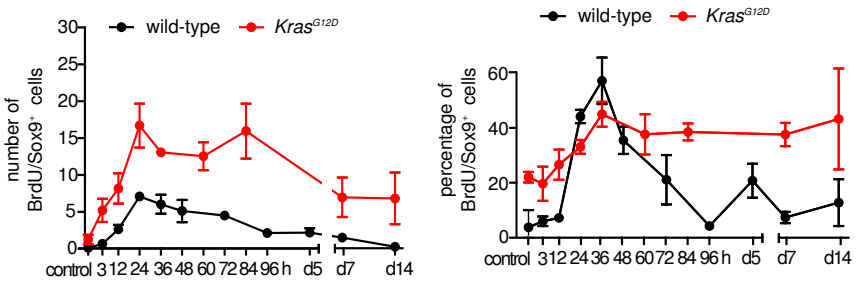

Figure 4.16: Number and percentage of proliferating Sox9-positive cells in WT and Kras^{G12D} pancreata.

thus concluded that the transcriptional program responsible for maintenance of the acinar phenotype is altered during inflammation in WT mice. Except for Ttc9 (tetratricopeptide repeat domain 9), the expression of 10 other genes displayed a similar kinetic pattern with transient down-regulation during the inflammatory phase. In particular, transient down-regulation of the transcription factors Bhlha15 (Mist1) and Rbpjl terminated earlier than down-regulation of the remaining 8 genes. This suggests that these transcription factors are major players in the complex dynamics governing loss of acinar cell homeostasis. We then compared these results with those from the Kras^{G12D} samples; here, constantly altered gene expression patterns were observed with an exception for Ttc9 and Agr2 (Anterior gradient 2), comparable to the above mentioned changes during inflammation in WT animals. Importantly, two transcription factors (Bhlha15 and Rbpjl) and target genes of the Ptf1a-Rbpjl complex such as Tmed11, Cuzd1, Derl3, Gal and Slc39a5 (important for protein transport and thus exocrine cell function) remained persistently down-regulated in Kras^{G12D} pancreata (Fig. 4.14, right panel).

Thus, we provide quantitative histological and molecular evidence that disturbance of the exocrine program characterizes both the inflammatory phase in WT mice and Kras^{G12D}-driven early carcinogenesis. As acinar cells constitute only 25% and 21% of proliferating cells during the inflammatory phase of the inflammatory response WT mice and early carcinogenesis, respectively, these data argue for a significant contribution of other non-acinar proliferative cells to inflammation and early carcinogenesis.

4.5.2 Expansion of progenitor-like cells characterizes the inflammatory phase and early carcinogenesis

It has been described that mature acinar cells transiently assume a “progenitor state” or “duct-like morphology” that allows for rapid repopulation of the exocrine compartment after inflammatory injury [219, 225]. We thus set out to quantify the proliferative activity of pancreatic progenitor-like cells in different phases of the inflammatory response (comparing WT to the respective Kras^{G12D} results). Pancreatic tissues were co-stained with BrdU and Sox9 - which labels cells with progenitor properties during embryonic development and in the adult organ [226, 227]. This analysis revealed that in WT mice, Sox9-positive proliferating cells were mainly expanded during the inflammatory phase and that the number of proliferating cells decreased during the

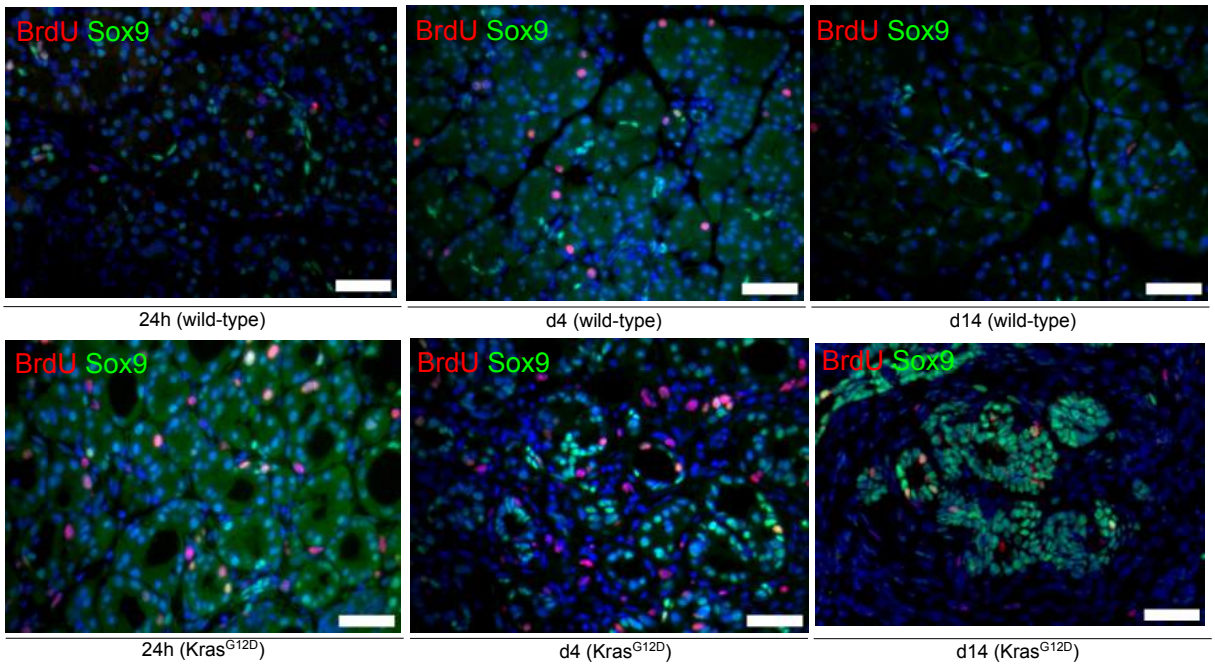

Figure 4.17: Representative IF of proliferating Sox9-positive cells in WT (upper panel) and $Kras^{G12D}$ (lower panel) pancreata in different phases of inflammation (upper panel) and anticipated phases of $Kras^{G12D}$-mediated early carcinogenesis (lower panel); scale bars: 50 μm.

regeneration phase (Fig. 4.16, 4.17). In contrast, the number of proliferative Sox9-positive cells was consistently high in $Kras^{G12D}$ samples compared to WT pancreata. Sox9-positive cells constituted up to 50% of proliferating cells in WT pancreata at 36h (the end of the inflammation phase) but their contribution decreased rapidly afterwards (Fig. 4.16). In $Kras^{G12D}$ pancreata, Sox9-positive cells accounted for an average of 31% of the proliferating cells and showed no decrease over time. Thus, the proportion of Sox9-positive cells in $Kras^{G12D}$ pancreata throughout the experiment was 50% higher than in WT animals during the regeneration phase ($p = 0.011$; t-test) but not during the inflammation phase (Fig. 4.16, Fig. 4.19). These data demonstrate that putative pancreatic progenitors or cells with progenitor features are transiently expanded within the inflammatory phase in WT mice but that their number is continuously increased in $Kras^{G12D}$-mediated early carcinogenesis.

Next we analyzed a key gene regulatory network (GRN) from the gene expression data that may support progenitor-like cell expansion in the inflammation phases of WT and of caerulein-treated $Kras^{G12D}$ pancreata. Because it has been demonstrated previously [228] that elements of embryonic development are recapitulated during organ regeneration in adult animals, we hypothesized that GRNs governing organ development might be partially active in our model. To test this, we extracted a previously compiled transcription factor (TF) GRN governing the embryonic development of the exocrine pancreas [229].

We reimplemented and applied a method developed by Ideker et al. [202] which allows to identify active subcomponents in a network, based on p-values for differential expression of the

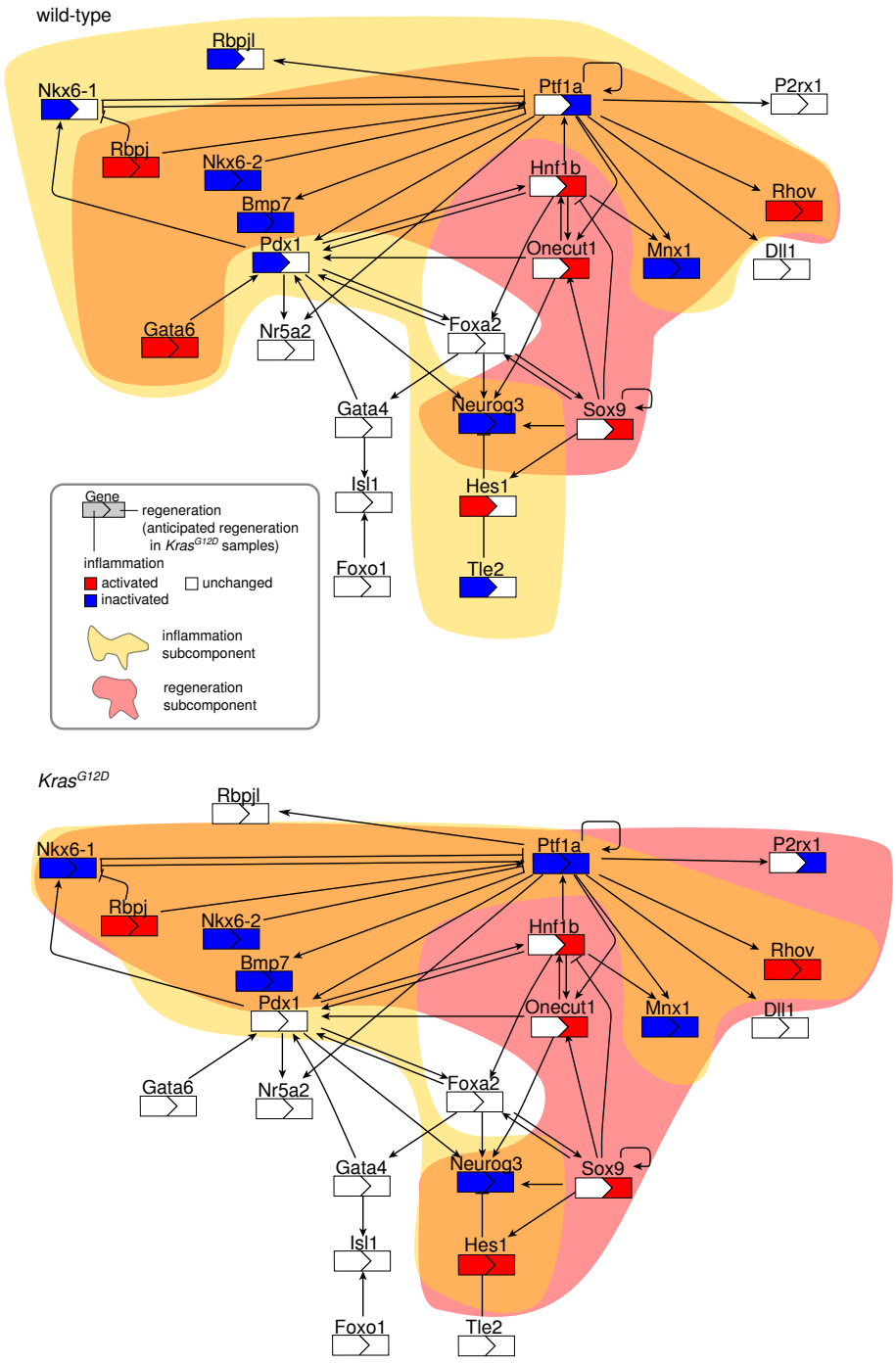

Figure 4.18: Bioinformatic analysis of the activity of the entire GRN of pancreatic organogenesis in different phases of inflammation (upper panel) and anticipated phases of Kras^{G12D}-mediated early carcinogenesis (lower panel).

single genes between two or more groups of samples, i.e. phases in our case (C++ implementation of the method is available at ICB, Helmholtz Zentrum München). These subcomponents which could also comprise the entire network are assembled based on corrected aggregated z-scores s_A (for details see section 3.4.8). The method was applied four times, in order to identify active subcomponents mediating transcriptional changes during 1) wild-type inflammation, 2) wild-

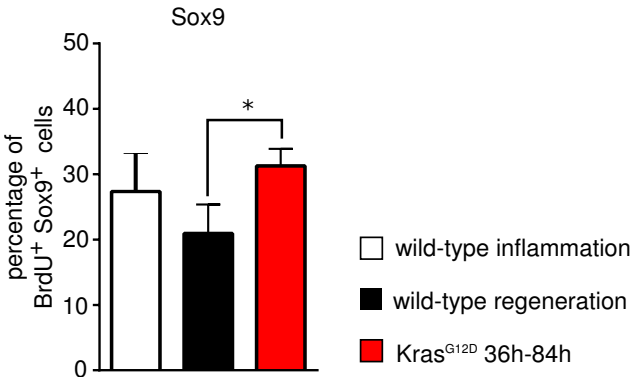

Figure 4.19: Percentage of proliferating Sox9$^+$ cells in wild-type inflammation, regeneration and Kras^{G12D} samples between 3h and 36h.

type regeneration and 3) Kras^{G12D} inflammation 4) anticipated regeneration in Kras^{G12D} mice. Hereto, differential expression analysis was performed individually for each of the four cases, according to section 3.4.3.3. In contrast to above where we performed pairwise comparison of groups, we are here interested in genes which are dysregulated in the respective group compared to *both surrounding groups*, e.g. wild-type inflammation samples are compared to wild-type control samples and also wild-type regeneration samples. Correspondingly, corrected p-values (FDR, Benjamini-Hochberg correction) were derived from one-way ANOVA incorporating the respective phase and the two surrounding phases. Afterwards, we applied a simulated annealing algorithm as part of Ideker's method, in order to optimize the active subcomponent in the network extracted from literature, scored with s_A. 100,000 iterations were performed in the simulated annealing procedure for each of the four cases separately. For wild-type mice, we obtained an active subcomponent with a corrected aggregated z-score $s_A = 3.21$ ($p = 0.0007$) for inflammation and another one with $s_A = 3.17$ ($p = 0.0008$) for regeneration. For Kras^{G12D} mice, we identified a subcomponent with $s_A = 2.92$ ($p = 0.0018$) for inflammation and another one with $s_A = 3.24$ ($p = 0.0006$) for regeneration. Subcomponents are visualized in Fig. 4.18; each of the network visualizations represents one of the two mouse strains, and each node is split up into two parts, the left part representing transcriptional activity during inflammation and the right part representing transcriptional activity during regeneration. Color refers to the expression change compared to the previous phase with red indicating up-regulation and blue indicating down-regulation.

Our analysis unveiled that, though not all individual players show differential expression when regarded separately, regulatory subcircuits comprising extensive parts of the network show significant activity during inflammation and regeneration, indicating that the progenitor differentiation network mediates cellular reorganization in both pancreatic regeneration and tissue transformation in Kras-muated mice. Two "regulatory elements" within the identified subcomponents are of specific biological interest, according to the dynamics of pancreatic development described in section 1.1.3.2. Ptf1a-Rbpj/Rbpjl-Nkx6.1 and Hes1-Sox9-Onecut1-Hnf1b-Neurog3; both of these have previously been linked with proliferation of pancreatic progenitors. Within the negative feedback loop consisting of Ptf1a, Rbpj/Rbpjl and Nkx6.1, the Ptf1a-Rbpj complex

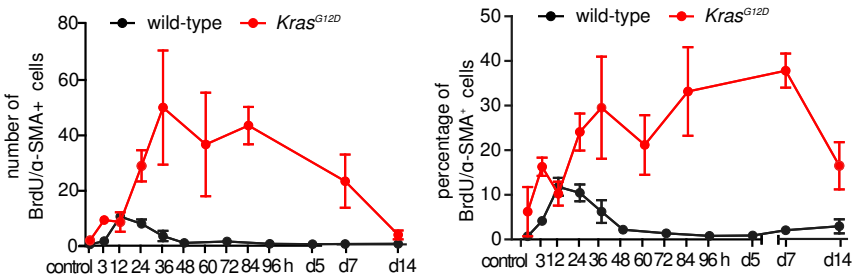

Figure 4.20: Number and percentage of proliferating α-SMA-positive cells in WT and Kras^{G12D} pancreata.

is presumably activated while the activity of Rbpjl and Nkx6.1 is suppressed in the inflammatory phase, which would hypothetically promote the formation of "immature" acinar cells. At the same time, only the negative feedback component of the Hes1-Sox9-Onecut1-Hnf1b-Neurog3 subcircuit is active: Hes1 integrates the intercellular signaling (particularly from the Notch pathway) and suppresses the activity of Neurog3 [230], stimulating proliferation of progenitor-like cells. These data suggest a coordination of the Ptf1a-Rbpj complex and the Notch-Hes1 pathway in expanding progenitor-like cells in the inflammatory phase of WT pancreata. During progression into the regeneration phase, the Ptf1a-Rbpjl complex seems to be "re-activated" and suppression of the activity of Nkx6.1 is released within the Ptf1a-Rbpj/Rbpjl-Nkx6.1 subcircuit, which facilitates the expansion of mature acinar cells. In the Hes1-Sox9-Onecut1-Hnf1b-Neurog3 subcircuit, Hes1 activity decreases whereas the signal loop between Sox9, Onecut1 (also known as Hnf6) and Hnf1b is activated. Together with the function of Ptf1a-Rbpjl, the activation of this signal loop (in the absence of Hes1) may provide the proliferative momentum for more differentiated acinar cells under physiological circumstances.

In Kras^{G12D} and exogenous inflammation-induced early carcinogenesis, however, we did not observe the dynamic changes in the activity of the Ptf1a-Rbpjl complex, Nkx6.1 and the Notch-Hes1 pathway in the course of time. Here, no "re-activation" of the Ptf1a-Rbpjl complex was observed; Nkx6.1 was persistently suppressed while the Notch-Hes1 pathway was constantly activated. These changes seem to provide continuous proliferating signals for (expansion of) progenitor-like cells during early carcinogenesis. This analysis also revealed a consistent activation of Gata6 in WT samples and its entire absence in Kras^{G12D} samples. This observation is largely in line with the recently uncovered role of Gata6, where its presence is required for maintaining the function of the exocrine pancreas while its absence promotes carcinogenesis by dedifferentiating acinar cells [224, 231].

4.5.3 Expansion of mesenchymal cells characterizes the inflammatory phase and early carcinogenesis

Because the pancreatic mesenchyme plays an important role in instructing expansion of pancreatic progenitors during embryonic development and because pancreatic injury is linked with expansion of resident fibroblasts, known as pancreatic stellate cells (PSCs) [232, 233], we quanti-

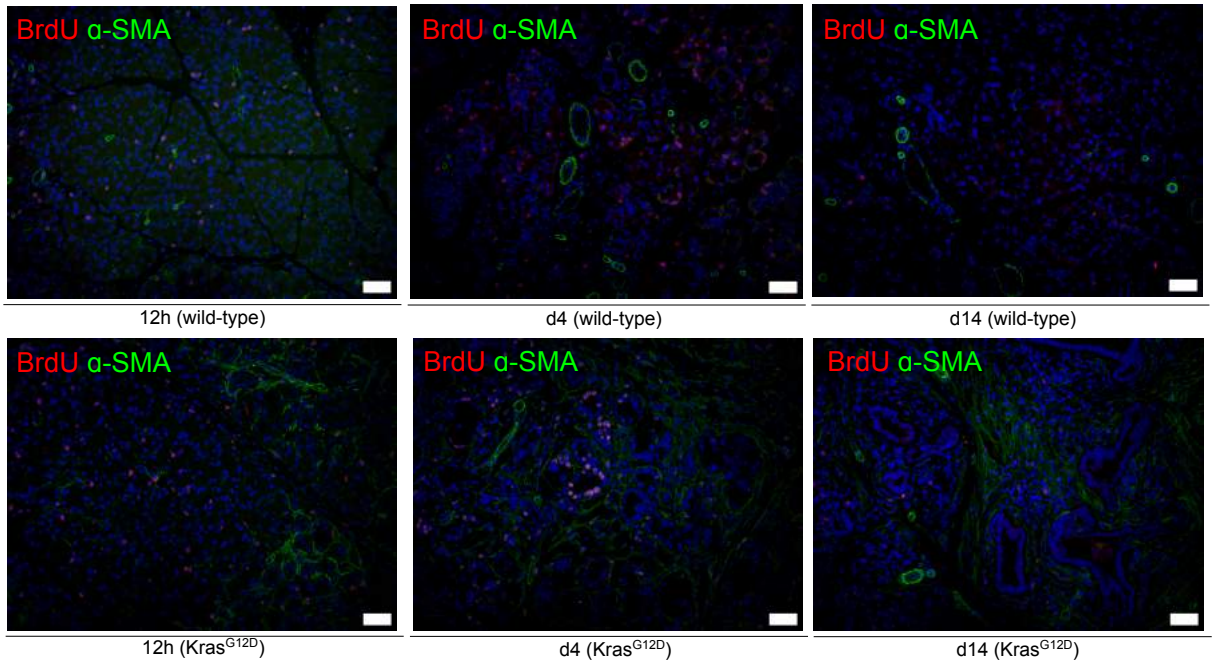

Figure 4.21: Staining for α-SMA-positive cells in WT and Kras^{G12D} pancreata.

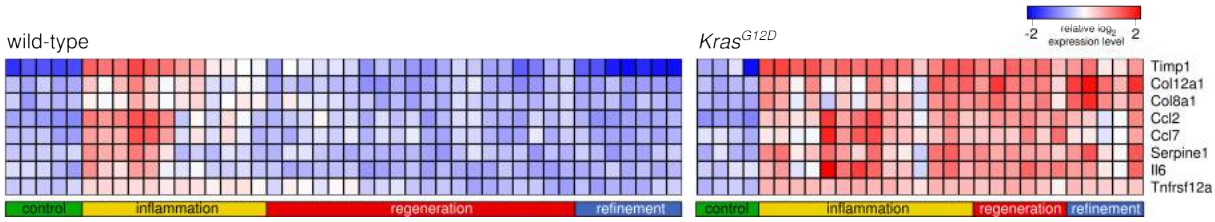

Figure 4.22: Heatmap of 8 genes linked to activation of PSC in vivo; voxel color: transcriptional up- (red) and down-regulation (blue) as compared to the mean; each voxel indicates gene activity in one mouse sample arranged in chronological order; colored bar: different phases or anticipated phases in WT and Kras^{G12D} samples, respectively.

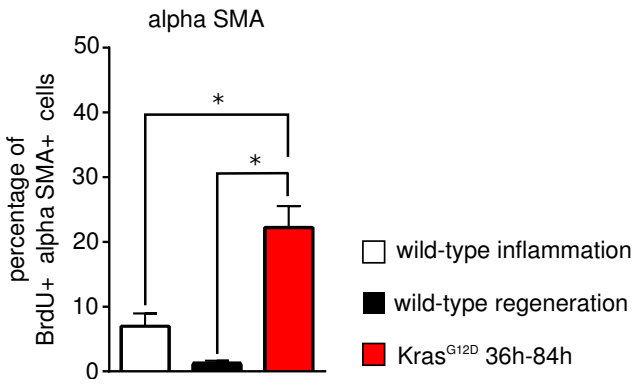

Figure 4.23: Percentage of proliferating α-SMA^{+} cells in wild-type inflammation, regeneration and Kras^{G12D} samples between 3h and 36h.

fied the number of such proliferating cells. In response to injury, PSC proliferated and expressed

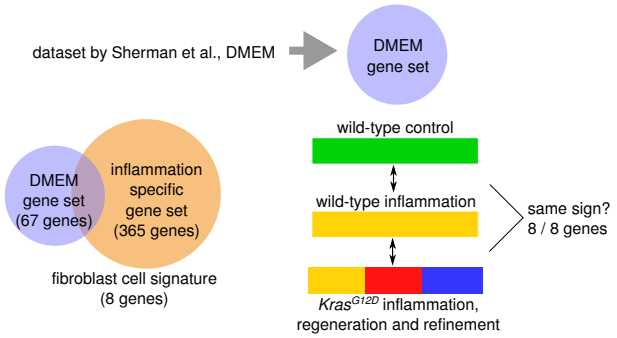

Figure 4.24: Workflow for selection of the "PSC activation signature", overlap of a literature-based gene set of PSC-specific genes and genes specifically up-regulated during the wild-type inflammation phase (compared to wild-type controls and regeneration).

α-SMA. Compellingly, we observed that proliferation of α-SMA-positive cells was strictly limited to the inflammatory phase in WT pancreata and that the magnitude of proliferation was relatively low. In contrast, proliferation of α-SMA-positive cells in $Kras^{G12D}$ samples was more persistent (from 3 hours to day 7) and more excessive (Fig. 4.20, 4.21). Accordingly, mesenchymal cells composed on average 22% of proliferating cells in $Kras^{G12D}$ tissues throughout the experiment whereas in WT pancreata, these constituted only 8% of the proliferating cells during the inflammation phase and less than 1% during the regeneration phase, respectively (Fig. 4.20, Fig. 4.23).

To validate this histological finding on the gene expression level, we extracted a previously characterized in vitro PSC "activation signature" from published data which was defined by comparing the transcriptomes of isolated pre-activated (3-day culture) to culture-activated PSCs (7-day culture) [234]. This in vitro PSC activation signature, which contained 67 genes including extracellular matrix (ECM) proteins, cell adhesion molecules, inflammatory cytokines/chemokines and modulators of signal transduction, was used as a positive reference to identify a potential *in vivo* PSC activation signature (Fig. 4.24). Eight of the 67 genes were differentially expressed in the inflammatory phase in WT samples compared to controls and regeneration phase samples ($FDR < 0.05$, absolute $FC > 2$, one-way ANOVA with three groups) Fig. 4.22, left panel). This ratio (8/67) was significantly higher than expected for a random gene set (hypergeometric test: $p = 2.57 \times 10^{-6}$). Furthermore, these eight genes exhibited exactly the same kinetic expression pattern: transient up-regulation in the inflammatory phase and normalization in the regeneration phase. Notably, all of these showed sustained up-regulation in $Kras^{G12D}$ samples treated with caerulein (Fig. 4.22, right panel), which is consistent with the persistent expansion of PSC on histological levels. Selection procedure is illustrated in Fig. 4.24.

Among these eight genes, the proteinase inhibitors responsible for ECM turnover, tissue inhibitor of metalloproteinase 1 (Timp1) and serine (or cysteine) peptidase inhibitor, clade E, member 1 (Serpine1, also known as PAI-1) have been reported to be associated with PSC activation [235]. Col12a1 (collagen, type XII, α1) and Col8a1 (collagen, type VIII, α1) are ECM molecules whose function in pancreatic carcinogenesis requires further investigation. Ccl2 (chemokine (C-C motif) ligand 2), Ccl7 (chemokine (C-C motif) ligand 7) and Il6 (interleukin

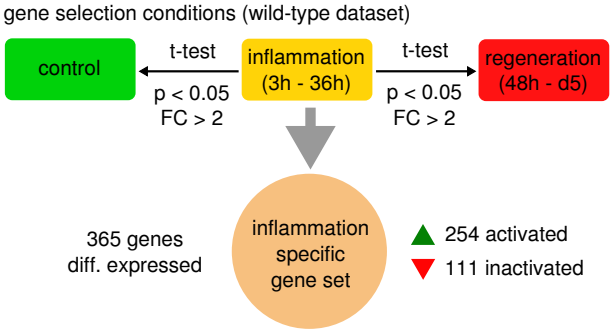

Figure 4.25: Workflow for selection of inflammation-specific genes in wild-type mice after one-way ANOVA.

6) are cytokines and chemokines, which promote pancreatic carcinogenesis by amplifying the inflammatory response [236, 237]. Our data thus suggest a pro-inflammatory role of PSC in early carcinogenesis. In particular, Tnfrsf12a (also known as Fn14 (fibroblast growth factor inducible-14 receptor)), which is the receptor for Tnfsf12 (also known as TWEAK (tumor necrosis factor-like weak inducer of apoptosis)), was identified as a novel PSC activation marker *in vivo*, which is in line with the recently uncovered function of the TWEAK/Fn14 pathway in promoting inflammation, cell proliferation and the fibrotic response in other biological contexts [238, 239].

4.6 The signature of early carcinogenesis is recapitulated in human PDAC and allows for survival stratification

This section, including all subsections, figures and any other material, is adapted from *Kong, Bruns, Behler et al.* [258].

4.6.1 Extraction of an inflammation-specific signature

To identify genes characterizing the inflammatory phase, differential expression analysis followed by Benjamini-Hochberg p-value correction was performed on the transcriptional profiles of WT samples, based on three groups, control samples, inflammation samples and regeneration samples (one-way ANOVA, followed by independent t-tests for the groups). The analysis yielded a set of 365 genes (254 up-regulated; 111 down-regulated) with significant differential regulation (absolute $FC > 2$, $FDR < 0.05$) in the inflammation phase of WT mice compared to control samples and to regeneration samples (Fig. 4.25). A 5-fold cross-validation was applied to generate an "inflammation-specific" signature consisting of 143 genes, allowing for significant differentiation between inflammation phase samples and other samples. As shown above, early carcinogenesis in $Kras^{G12D}$ mice and the inflammatory phase in WT mice are similar in both the composition of proliferating cells and in the corresponding selected gene set. We thus went on to test whether this also holds true without using the "histology-guided" gene selection (analyses from the above sections were motivated from histological observations, gene selection was performed according the function of specific cell type; here, in contrast no "prior information" except the identified

phases are used, in order to find a signature best characterizing the inflammatory state). Thus, unsupervised clustering was performed using all WT and Kras^{G12D} samples, based on the 143 genes (the "inflammatory signature"). As visualized in Fig. 4.26, all Kras^{G12D} samples treated with caerulein were clustered together with the samples of the inflammatory phase of WT mice. In contrast, this gene set did not distinguish the regeneration phase from control or refinement samples in WT pancreata. As such, we identified a highly specific signature for early carcinogenesis that may be present even before the acquisition of oncogenic Kras. We thus asked whether this signature would be preserved in (advanced) human pancreatic adenocarcinomas.

4.6.2 Differential expression and survival analysis

An independent dataset of human PDAC tissue samples and samples from the surrounding tissue was obtained from GEO (Gene Expression Omnibus), dataset ID GSE28735, and used for further analysis (previously published by Zhang et al. [240]). Homologous genes of the "inflammation-specific" signature (143 genes) were identified using Biomart. Gene expression in PDAC samples was corrected against expression in the surrounding tissue by subtraction of the matched sample signals. The aggregated z-score z_A of the signature's genes was used to describe the level of expression of the signature as a whole in a given sample (Fig. 4.27). We found that this score was not normally distributed in the samples, as it would be expected for a random set of genes (Shapiro-Wilk-Test, $p = 0.001$, Fig. 4.28).

A null model with random gene sets was applied to correct this distribution for chance, showing that the cross-validated, inflammation-specific signature was not equally expressed in the PDAC samples with p=0.002 (Fig. 4.28). K-means clustering was performed separating the patients into two clusters ($k = 2$ was chosen according to the best gap statistic), one reflecting a high score (high expression of the signature's genes, $n = 16$) and the other one reflecting a low score ($n = 29$) (Fig. 4.28). Survival analysis was performed to determine whether expression of the signature related to survival time (median survival 12 months; Fig. 4.28, right panel). Even though the observed clustering was robust, significance for differences in survival was not reached, most likely because of the relatively small sample size (log-rank test: $p = 0.106$ and Peto-Peto test: $p = 0.056$). Kaplan-Meier survival curves showed an increased survival time for patients with a high score (median survival: 24 months) compared to low-scoring patients (median survival 12 months; 4.28, right panel).

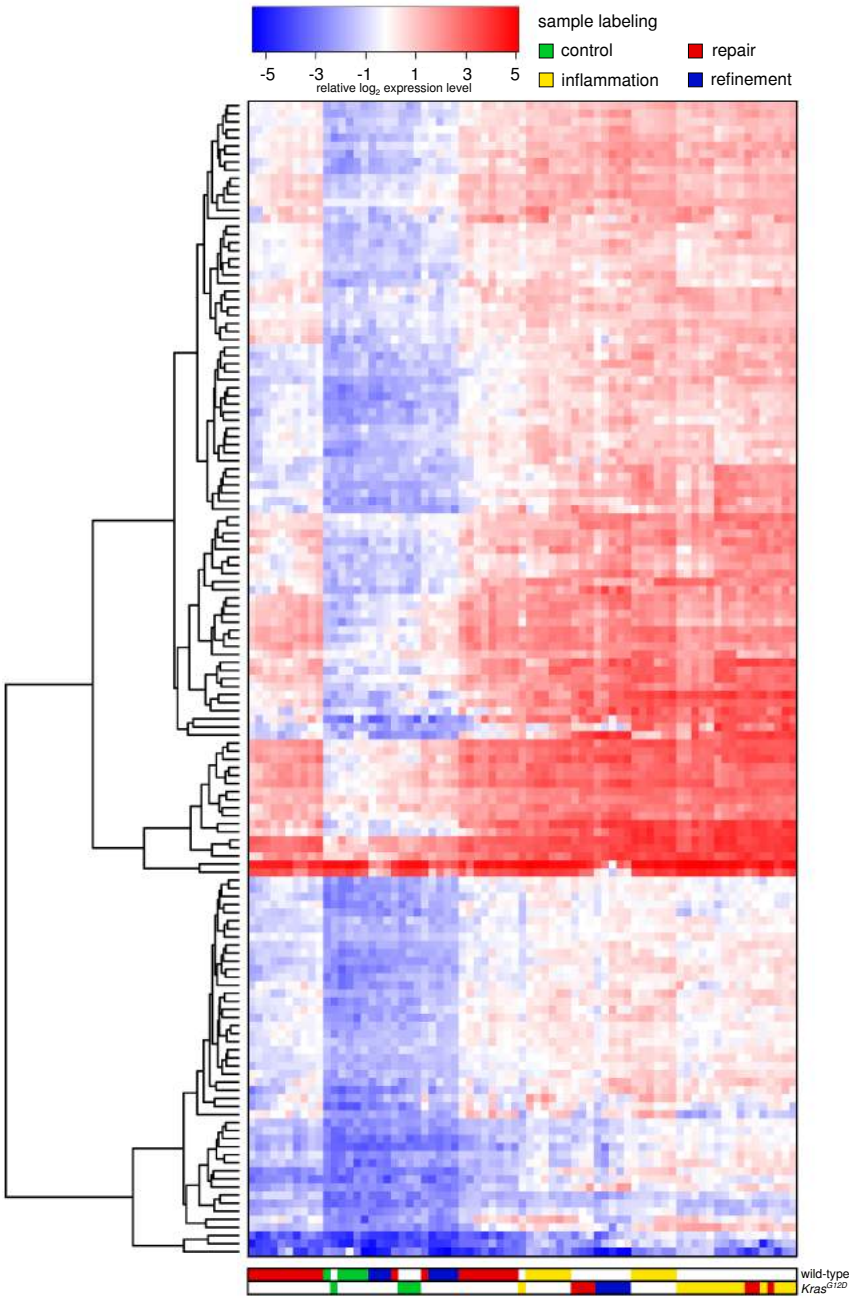

Figure 4.26: Clustered heatmap of the gene signature of the inflammation phase allows differentiation of early carcinogenesis from control samples.

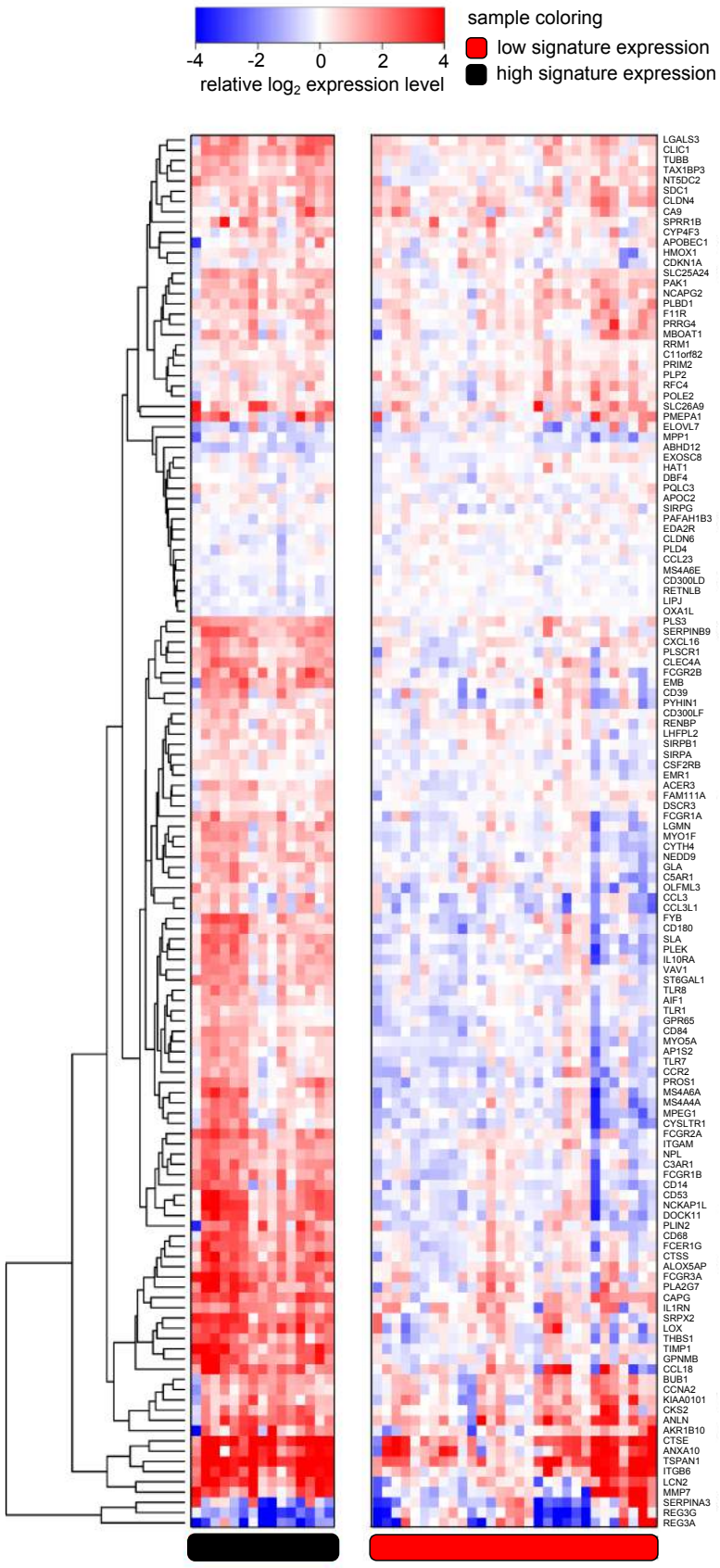

Figure 4.27: k-means clustering allows separation of human PDAC samples into two clusters with high (black cluster) and low (red cluster) expression of the inflammation signature (homologous genes).

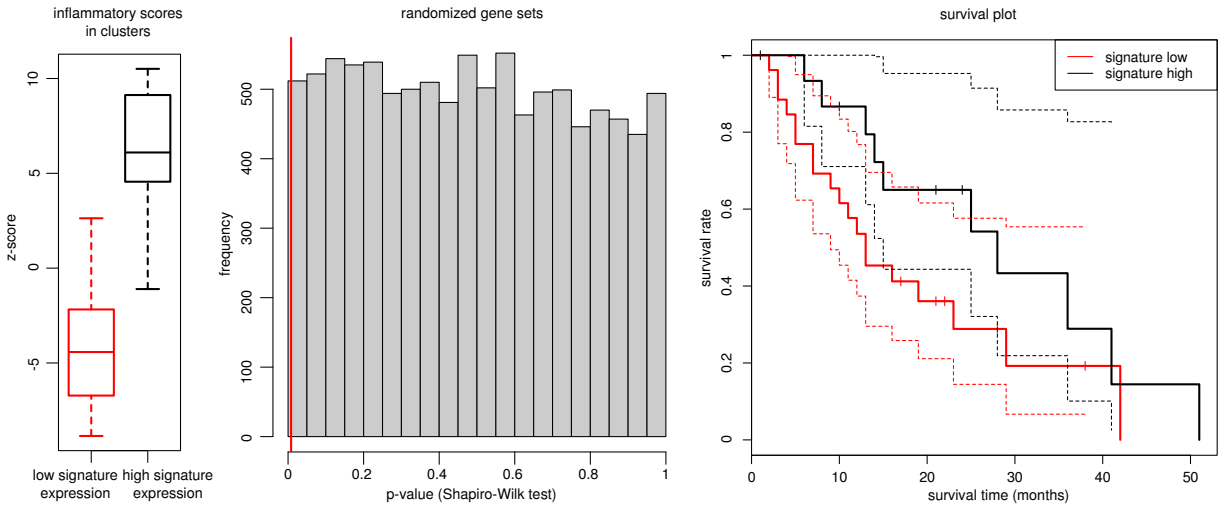

Figure 4.28: Left: Distribution of signature expression scores in the two clusters (aggregated z-scores). Center: Sampling with random gene sets indicates significant differential expression of the identified signature (red bar, p=0.0012). Right: Survival analysis of the two clusters: longer median survival of patients with a high expression of the signature (24 months vs. 12 months).

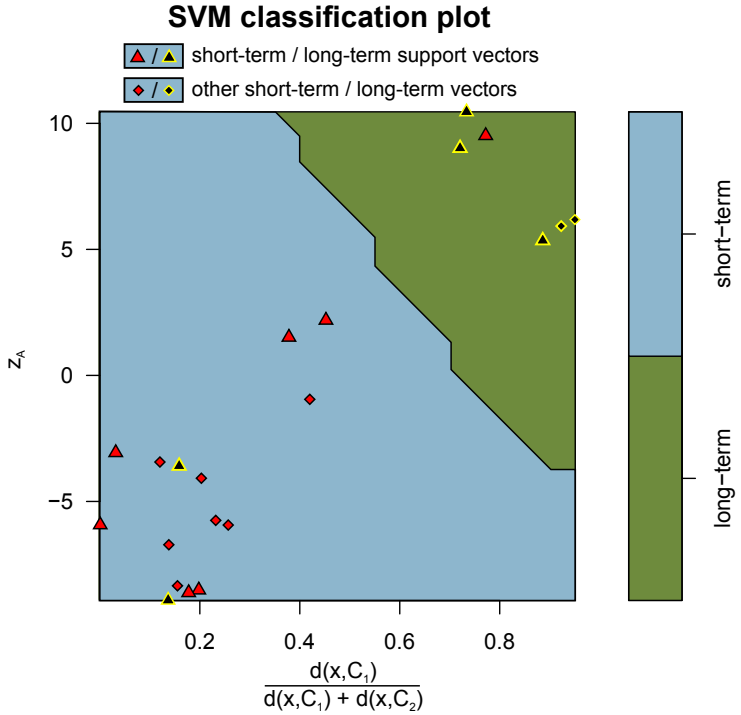

Figure 4.29: SVM based on z_A score and clustering information allows classification of long-term and short-term survival patients with an accuracy of 0.86 (leave-out-one cross-validation).

In order to find whether the signature expression score z_A can be used for classification of long-term and short-term survival patients, linear regression was performed. According to clinical definitions, patients with a survival time less or equal to 12 months were regarded as *short-term* samples ($n = 14$) and patients with a survival time greater or equal to 24 months

($n = 7$) were regarded as *long-term* samples, other samples were discarded.

However, regression did not yield significant results. In a next step, a support vector machine (SVM) with radial kernel function was trained on the samples; still samples could not be separated with a satisfying accuracy. We then used the relative distance (based on L_2-norm) of a sample to the centers of the clusters as an additional variable: $\frac{d(x,C_1)}{d(x,C_1)+d(x,C_2)}$ where x represents the signature expression vector (143 elements) of the respective sample and C_1, C_2 represent the two cluster centers. According to the definition, the variable ranges from 0 to 1 describing distance of the sample to cluster 1 and at the same time closeness to cluster 2. We trained another SVM with radial kernel function, based on the two variables, and found that samples could be classified correctly with an accuracy of 0.86 now. Leave-out-one cross-validation was performed in order to test for robustness of the classifier.

4.7 Identification of potential mediators of the inflammation-regeneration transition

Our study unveiled that pancreatic regeneration in wild-type mice is a *three-stage* process; in Kras-mutated mice, however, regeneration is blocked and persistent inflammation is observed. Here, we want to address the question which regulatory mechanisms are mediating the regeneration process in wild-type mice or blocking regeneration in Kras-mutated mice.

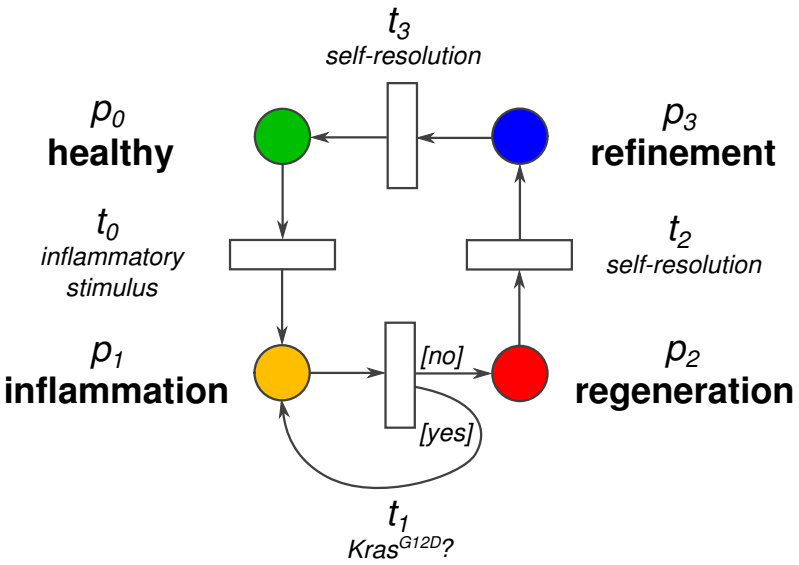

Figure 4.30: The petri net describes our model of pancreatic regeneration and early pancreatic carcinogenesis, derived from the results presented in sections 4.1 and 4.2.

In order to concretize our previous observations formally, we define the "dual model" of pancreatic regeneration and early pancreatic carcinogenesis as our *working model*, depicted in the petri net in Fig. 4.30. The model is composed of four steady states each representing the identified stages including the control state, termed "healthy state". Based on our findings from above, we know that the temporal course of inflammatory response in Kras^{G12D} mice diverges

from natural pancreatic regeneration after 36 hours, as inflammatory response in Kras^{G12D} mice fails to enter the "regeneration phase". Accordingly, we are particularly interested in identification of gene interactions mediating the *inflammation-regeneration transition* in wild-type mice and the prolonged inflammation in Kras^{G12D} mice, between 3h and 84h after caerulein injections. As the regeneration phase is characterized by repair of the acinar cell structures

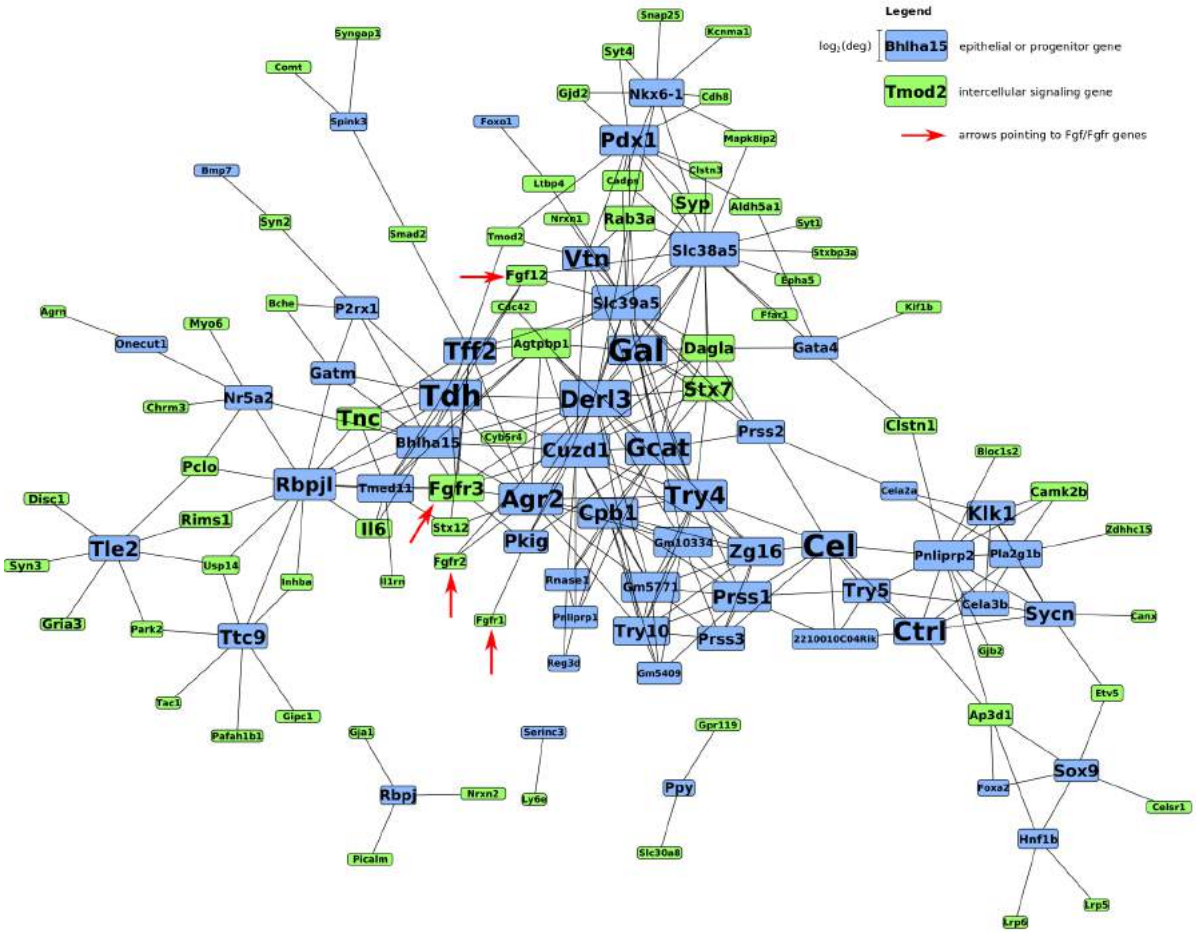

Figure 4.31: Interactions of epithelial genes and cell-cell signaling related genes in wild-type mice between 3 hours and 84 hours. Occurrence of 16 interactions involving Fgf/Fgfr-related genes suggests that corresponding pathways contribute to the transition from inflammation phase to regeneration phase.

and redifferentiation of progenitor-like cells into acinar cells, we particularly focused on genes involved in maintenance of the acinar cell fate as well as differentiation of Sox9$^+$ genes. Here, we used the same genes extracted from literature in sections 4.5.1 (acinar homeostasis) and 4.5.2 (progenitor network) which together form a set of 77 genes; these genes will be called *epithelial genes* in the following. As we are interested in identification of intercellular signals impacting the expression of those genes, we further selected a set of signaling genes, based on the GO term "cell-cell signaling" (GO:0007267) which yielded a set of 420 genes; these genes will be called *intercellular signaling genes* in the following. Now, correlation networks were generated in four steps:

1. *Sample selection.* As we aimed to compare interactions in wild-type mice and Kras^{G12D} mice for the time frame 3h-84h, we selected samples from corresponding time points; however, if correlation networks are assembled based on a hard significance cut-off, then there will be detected more interactions for a dataset with more samples, as the confidence interval is larger here (see methods). Accordingly, sample number was reduced for wild-type mice, leaving out the time point 72h which was not available for $Kras^{G12D}$ mice. In addition, for several time points, the number of samples available was differing (e.g. for 12h, four Kras^{G12D} but only three wild-type samples were available). Here, numbers of samples were reduced for the dataset with more samples; to this end, samples with the greatest sample ID were removed, i.e. selection was performed arbitrarily in these cases. Finally, two datasets, each comprising 18 samples, were used. For each sample of a given time point in one dataset, there was one corresponding sample from the same time point in the other dataset.

2. *Correlation tests and p-value correction.* In a next step, for each of the two datasets independently, *all-to-all* Pearson correlation networks were generated by calculating Pearson's r pairwise for all of the available genes; correlation tests were performed based on Fisher's z transformation. Afterwards, p-value correction was performed using Bonferroni's method.

3. *Extraction of relevant interactions.* Then, from the assembled global networks, *interactions of interest* were extracted. These include interactions between *two epithelial genes* or between an *epithelial gene* and an *intercellular signaling gene*. All of these interactions with a corrected p-value smaller than 0.05 were extracted.

4. *Analysis and visualization.* Finally, networks were analyzed and visualized. For wild-type mice, 296 interactions were observed; 162 of those were interactions between an *epithelial gene* and an *intercellular signaling gene*. Corresponding wild-type network is visualized in Fig. 4.31. For Kras^{G12D} mice, on the contrary, we obtained 14 interaction pairs only as visualized in Fig. 4.33. Among these 14 interactions there was not a single one between an *epithelial gene* and an *intercellular signaling gene*.

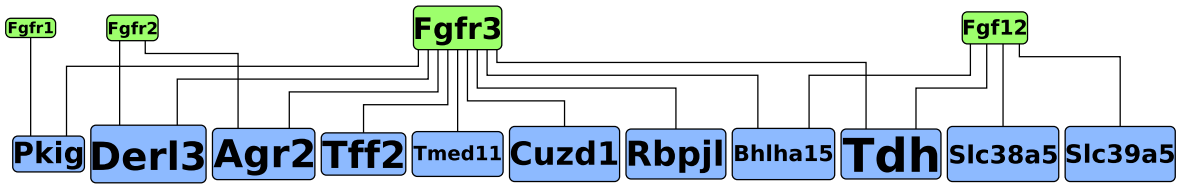

Figure 4.32: Neighbourhood of the Fgf/Fgfr nodes in the wild-type network.

Notably, with Fgfr1, Fgfr2, Fgfr3 and Fgf12, four representatives of the fibroblast growth factor family and corresponding receptors were included in the wild-type network, suggesting that Fgf signaling plays a crucial role in the mediation of the inflammation-regeneration transition. Particularly, Fgfr3 is a highly connected player in the network, interacting with nine of the

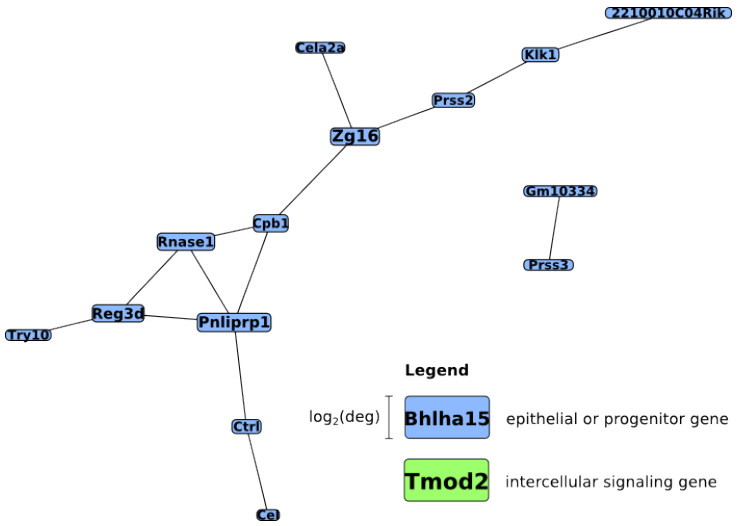

Figure 4.33: In Kras^{G12D} mice, 14 interaction pairs were found.

preselected acinar/progenitor genes, including Mist1 (Bhlha15) which is a key factor of acinar homeostasis, and also other important acinar-specific genes. Neighbourhood of Fgf/Fgfr genes is in the wild-type network is visualized in Fig. 4.32.

Hypergeometric test was applied in order to find whether there is a significant enrichment of edges involving Fgf/Fgfr family members among the identified interactions between *epithelial genes* and *intercellular signaling genes*. Here, the number of all possible interactons is 77×420 (each epithelial gene interacts with each intercellular signaling gene) and the number of positive cases is 77×13, as there are 13 Fgf/Fgfr family members included in the set of 420 intercellular signaling genes. The number of identified interactions is 162 and among these, there were 16 interactions involving Fgf/Fgfr family members. However, this test is based on the assumption that for any pair of random genes A and B, it is equally likely that there will be an interaction between both. As most biological networks are *scale-free* (i.e. degree distribution is following a power law function), this assumption is not correct; it could be the case, that Fgf/Fgfr genes are hubs in the global network and in this case it would not be surprising that interactions involving these genes are also "enriched" in the identified subnetwork. However, the number of positive cases (16) can be corrected: Let $\deg(A)$ and $\deg(B)$ be the degrees of genes A and B in the global *all-to-all* network, then, the "expected number" of interactions between A and B is $\deg(A)\deg(B)$ times greater as compared to two nodes with a degree of 1. Correspondingly, we rescale the identified interactions. While initially each interaction was counted with weight 1, now each interaction is counted with weight $(\deg(A)\deg(B))^{-1}$. Afterwards, each weight is divided by the sum of weights for the identified interactions (i.e. now the sum of all weighted interactions is 1) and finally multiplied with 162, so that the total number of identified interactions remains constant; now, weights for members of the Fgf/Fgfr family are summed up, yielding a total number of 10.85 (weighted) interactions. As the hypergeometric distribution is discrete, we test for the ratio $10/162$ against $(13 \times 77)/(420 \times 77)$, which yields a p-value of 0.012. This indicates, that there is a significant enrichment for Fgf/Fgfr-involving

interactions among the identified set of interactions between epithelial and intercellular signaling genes.

We conclude, that the *inflammation-regeneration transition* in wild-type mice, described with t_1 in our working model in Fig. 4.30, which is required for successful pancreatic regeneration, is potentially mediated by members of the Fgf/Fgfr family.

Previous studies strengthen the potential relevance of this finding: It was reported that Fgfr3 has a dual role in pancreatic carcinogenesis [255]. Being generally regarded as oncogene which is often found to be mutated in PDAC patients, it has been shown that Fgfr3 limits cell growth and promote differentiation in epithelial cells [255]. In mesenchymal cells however, it possesses oncogenic properties [255]. In this regard, Fgfr3 could play an important role, mediating the redifferentiation into epithelial cells after inflammatory injury in wild-type mice. Accordingly, missing Fgfr3 activity in Kras^{G12D} mice could be the explanation for missing redifferentiation and finally mesenchymal expansion.

Generally, comparison of wild-type mice and Kras^{G12D} mice shows that a complex interaction network is required for successful initiation of regeneration; this complex network requires both interactions between acinar- or progenitor-specific genes and exterior signals.

Chapter 5

Discussion

5.1 Loss of coordination in cell proliferation initiates early carcinogenesis

For the first time, we here analyzed the natural course of pancreatic regeneration after inflammatory injury with a high temporal resolution on both histological and transcriptional level. Tissue and RNA samples were extracted at 13 time points after injections of supramaximal concentrations of caerulein. Our analysis unveiled that pancreatic regeneration is a process which can be separated into three distinct phases which we termed inflammation, regeneration and refinement. Corresponding definitions were based on histological observations and could be entirely confirmed on the transcriptional level by application of hierarchical clustering. Comparative analysis unveiled that the inflammatory phase is perpetuated in mice, conditionally expressing oncogenic $Kras^{G12D}$. Correspondingly, tissue damage cannot be resolved; instead altered proliferation cascades and persistent cell expansions are observed. Here, the key question is which molecular mechanisms are required for mediation of regular cellular reorganization in wild-type mice and which mechanisms are driving the divergence of the response in $Kras^{G12D}$ mice.

Quantification of Amy^+ cells and histological analysis indicated that acinar complexes are destroyed during the inflammatory phase in both mouse strains. Transcriptional profiles confirmed this observation, suggesting that the acinar homeostasis program is disturbed by caerulein-induced signaling cascades. As we can observe a simultaneous expansion of $Sox9^+$ cells, we agree with previous studies suggesting a dedifferentiation of acinar cells into progenitor-like cells expressing Sox9 [256]. It has been previously proposed that lacking expression of Mist1 (Bhlha15) is initiating the loss of acinar homeostasis which is also in line with our findings. Particularly, we observe, that shortly after reactivation of Mist1 and Rbpjl in wild-type mice at t=24h, other acinar homeostasis related genes are reactivated as well and that the repair of acinar structures is initiated as indicated by the up-following peak of proliferation of $BrdU^+$ Amy^+ cells. In $Kras^{G12D}$ however, reactivation of the acinar program fails; instead, massive expansion of $Sox9^+$ cells is observed, finally leading to formation of acinar-to-ductal metaplasia. Correspondingly, we addressed the question how the redifferentiation of $Sox9^+$ cells is impaired

in Kras^{G12D} mice as we assume this to be a crucial step preventing these mice to enter the regeneration phase after inflammatory injury. To this end, we considered a previously published gene regulatory network describing differentiation programs of progenitor cells during pancreatic development, assuming that the same mechanisms are also influencing the cell fate of the observed progenitor-like cells. We were searching for subcomponents in this network that are particularly activated during the inflammation phase and the regeneration phase for both mouse strains separately. Here, we regarded subcomponents to be "active" if corresponding genes taken together showed differential expression (corrected aggregated z-score). Though subcomponents identified for wild-type mice and Kras^{G12D} mice showed a large overlap, several key factors showed a different expression pattern and potentially influence the cellular fate and thus altered proliferation cascades. In particular we found Hes1 to be persistently activated in Kras^{G12D} (in wild-type mice it was inactivated after 36 hours), potentially indicating a sustained Notch signaling. Previous studies have shown that after loss of the acinar cell fate (e.g. induced by loss of Mist1 expression), elements of embryonic development are recapitulated, including in particular Notch signaling but also other pathways [11]. Accordingly, once the acinar fate is lost, different scenarios give possible explanations for the induction of uncontrolled expansion of Sox9$^+$ cells in Kras^{G12D} mice. Either, the persistent inactivation of Mist1 is perpetuating the proliferation of Sox9$^+$ cells, with Notch signals being a side effect accompanying this process. Or, once the acinar cell fate is lost, persistent Notch signaling governs expansion of Sox9$^+$ cells and together with Ptf1a suppression prevents redifferentiation to acinar cells at the same time. Alternatively, it is also possible that a combination of these two forces is promoting uncontrolled expansion. Mist1 inactivation causes loss of the acinar cell fate, and Notch-mediated disturbance of the Hes1-Sox9-Onecut1-Hnf1b-Neurog3 regulatory circuit disallows redifferentiation, promotes further proliferation (due to Neurog3 inhibition [226]) and also intensifies further Mist1 inhibition. Recent studies based on Mist1 knock-out mice with Kras^{G12D} mutations have shown that Kras signaling cooperates with Notch and EGFR signaling promoting the formation of PanIN lesions [257]. Enhanced activity of these pathways and accelerated PanIN formation could indicate that effects caused by Mist1 inactivation and Notch signaling strengthen each other which finally initiates tissue transformation. According to these findings, elements of the Hes1-Sox9-Onecut1-Hnf1b-Neurog3 regulatory circuit could be considered as potential drug targets for efforts aiming at inhibition of uncontrolled expansion of Sox9$^+$ cells, particularly, because the other driving force, Mist1 inactivation is directly tracing back to Kras signaling, and Kras is an undruggable molecule. Additionally, activity of the Rbpjl-Ptf1a complex is potentially impaired in Kras^{G12D} mice, as we observe persistent down-regulation of Ptf1a here. As this complex is required for differentiation of progenitor cells to acinar cells, we assume that missing Ptf1a activity contributes to Sox9$^+$ cell expansion.

Apart from the distinct molecular mechanisms, temporal coordination of proliferation of the different cell types has been shown to be required for faultless development of organs. Recent studies uncovered that optimal control theory is governing intestinal development, with an expansion of stem cells and progenitor cells followed by proliferation of epithelial cells; these

well-defined cascades of cell proliferation have been shown to be the most efficient way for (re-)generation of the intestinal system [249]. In line with these findings, we observe similar cascades in the natural course of pancreatic regeneration. Here, proliferation of progenitor-like cells peaks prior to the strong elevation of the proliferation of acinar cells. In $Kras^{G12D}$ mice however, these clear cascades of proliferation could not be observed. Although it remains unknown whether acquisition of proliferative capacity and progenitor-like features is a default program after failure of the exocrine program in acinar cells, it has been shown that inflammation inevitably promotes the inactivation of the exocrine program and accelerates $Kras^{G12D}$-driven pancreatic carcinogenesis [250]. Taken together, these data suggest that interaction between inflammatory pathways and $Kras^{G12D}$ destabilizes the exocrine program and induces acquisition of progenitor-like features and proliferative capacities.

It is commonly known that cytokine and chemokine release as a causality of inflammatory response leads to recruitment of quiescent PSCs which in turn become α-SMA expressing, *active* PSCs [16]. Activated PSCs in turn release further cytokines and chemokines, but also produce MMPs and other ECM components as well as their inhibitors, leading to mesenchymal expansion [16]. We quantified the number of α-SMA^+ cells in order to compare the extent of mesenchymal expansion between wild-type and $Kras^{G12D}$ mice. We observe that mesenchymal expansion during the inflammatory phase is strongly increased in $Kras^{G12D}$ mice, compared to wild-type mice. Particularly, we found that TWEAK activation allows quantification of the extent of inflammatory response and PSC expansion in both mouse strains; TWEAK/Fn14 has also previously been reported to promote cell proliferation and fibrotic response [238, 239]. Increased PSC expansion in $Kras^{G12D}$ during the inflammatory phase can be possibly regarded as a self-induction effect; once pronounced inflammatory response is initiated in $Kras^{G12D}$ mice, accompanied by the above discussed cellular expansion, activated PSCs release additional cytokines and chemokines promoting further mesenchymal expansion. Possibly, once the population of active PSCs is sufficiently grown, signals induced by PSCs then also influence the population of remaining epithelial cells. Cross-talk of epithelial and mesenchymal cells has been studied and documented extensively for patients in advanced stages of PDAC as discussed (tumor-stroma interactions, see section 1.4.3). As impaired regeneration and loss of epithelial cells in $Kras^{G12D}$ mice is accompanied by massive mesenchymal expansion during the anticipated regeneration phase (48h - d5), such a cross-talk could further accelerate disruption of acinar structures. Additionally, we assume that acinar cells undergo EMT during inflammation phase and anticipated regeneration phase, because we observed a persistent Notch signaling which can cause activation of EMT-inducing transcription factors [149] (see also section 1.4.4). This effect would explain further enhancement the observed mesenchymal expansion and disruption of acinar cells.

Our findings indicate that an extended and enhanced inflammatory response in $Kras^{G12D}$ mice generates a transformation-permissive microenvironment which, once initiated, causes a loss of coordination in cell proliferation, characterized by acinar dedifferentiation, missing redifferentiation, extensive PSC recruitment and possibly EMT and epithelial-mesenchymal cross-talk. The resulting premalignant cellular mileu of stromal cells, $Sox9^+$ cells and inflammatory

cells then potentially leads to a self-induction effect promoting further tissue transformation. We indicate that the extended inflammatory phase in $Kras^{G12D}$ mice is the "incipient phase of early pancreatic carcinogenesis" and provide histological and molecular evidence that pancreatic carcinogenesis adheres to the old concept proposed by Dvorak: "cancer is a wound that never heals" [247]. One major characteristic of impaired regeneration in $Kras^{G12D}$ mice is uncontrolled mesenchymal proliferation. Here, the question raises in how far the steps of pancreatic organogenesis are recapitulated during normal regeneration. Studies focusing on pancreatic growth and branching during organogenesis have shown the emergence of a strict temporal order regarding proliferation of precursor cells, mesenchymal cells and differentiated cells [248]. Assuming that similar coordinated proliferation cascades are required for organ regeneration, altered temporal course of mesenchymal expansion explains induction of tissue transformation. Still, biological relevance of the early mesenchymal expansion of in the long-term carcinogenic process requires further investigation with special focus on the epithelial-mesenchymal cross-talk. Particularly it is unknown whether early mesenchymal expansion finally causes desmoplastic reaction leading to PDAC formation.

We conclude that inflammation-accelerated early pancreatic carcinogenesis in $Kras^{G12D}$-mutated mice is characterized by two properties,

- persistent inflammation, corresponding to the inflammatory phase in wild-type mice on both histological and transcriptional level
- loss of coordinated cell proliferation the lineages of epithelial, progenitor-like and mesenchymal cells.

In order to further investigate the impact of Notch signaling, Ptf1a-Rbpjl inactivation and other molecular factors on the expansion of progenitor-like cells, we suggest to perform cell-sorting based experiments. Additionally, in order to identify molecular interactions of epithelial and mesenchymal cells which potentially induce massive mesenchymal expansion and tissue transformation, further *in vitro* studies are proposed.

5.2 Transferability to pancreatic carcinogenesis in humans

Generally, genetically engineered mouse models (GEMM) provide an excellent opportunity to investigate early carcinogenesis and recent successes have shown that GEMM-based studies provide indispensable contributions for cancer research. In particular, invention of Kras models is regarded as milestone in cancer research [170]. However, attention must be paid when conclusions drawn from mouse model-based experiments are generalized.

Human PDAC show a high intrinsic variation across different patients due to differences in genetic, epigenetic and microenvironmenal factors. This implies variation in response to surgery and chemotherapy: Even for patients with the same mutational status, e.g. mutation of EGFR, therapies targeting the corresponding molecule can result in different outcomes [170]. In order to cover intrinsic variation in human PDAC, more differentiated GEMMs are required, covering

additional mutations or epigenetic factors. One further limitation of GEMMs is that given mutations are implicitly assumed to be the initiating event of early carcinogenesis. However, preceding events might have initiated mutagenesis; these events are not covered by GEMM-based experiments.

Nonetheless, our findings from above raise the question whether specific molecular characteristics of early pancreatic carcinogenesis in mice are also found in human PDAC patients. Based on differential expression analysis followed by a 5-fold cross-validation, we have extracted a signature of 143 genes, characterizing the inflammatory phase of pancreatic regeneration and, correspondingly, also sustained inflammation in Kras^{G12D}-mutated mice. Since we assume that expression of this signature corresponds to the *early stage* of carcinogenesis, it could be potentially applied for survival prediction of PDAC patients. Provided with a dataset of transcriptional profiles of 45 pancreatic cancer patients, published by Zhang et al. [240], we found our signature to be differentially expressed between two groups of patients which were identified using k-means clustering. Differential expression was highly significant ($p = 0.002$). Subsequent survival analysis showed that the median survival time of patients with a high signature expression was 24 months, whereas median survival time of patients with a low signature expression was 12 months; however, this result was not significant (log-rank test: $p = 0.106$, Peto-Peto test: $p = 0.056$). Based on the signature expression score (aggregated z-score), and the sample distance to the cluster centers, we trained a support vector machine, followed by leave-out-one cross-validation, which allowed to classify long-term ($\geq$ 24 months) and short term survival patients ($\leq$ 12 months) with an overall accuracy of 0.86. Our findings indicate, that expression of the previously identified signature might be applied as indicator of progression of PDAC. We suggest to evaluate the signature's expression on a larger dataset of human PDAC samples with a similar experimental design, comprising tumor tissue and surrounding tissue. If our hypothesis should turn out to be correct, a next step would be to find generalizations of this concept that possibly unveil a correlation of PDAC progression and degree of inflammation.

5.3 Mediators of the inflammation-regeneration transition

Based on our findings, we set up a "dual model" of pancreatic regeneration and early pancreatic carcinogenesis. Here, previously defined phases and control samples are regarded as four different steady states. In both mouse strains, transition between the "control state" and the "inflammation state" is initiated by caerulein injections and subsequent inflammatory response. In wild-type mice, all further transitions are self-resolution steps. In Kras^{G12D} mice, the "inflammation state" is perpetuated. According to these observations, we hypothesize that impaired regeneration in Kras^{G12D} mice mainly depends on the inability of entering the *inflammation-regeneration transition* which causes divergence of the temporal courses of both mouse strains. Correspondingly, investigation of the molecular mechanisms during the time segment between 3h (earliest time point regarded as "inflammation phase") and 84h (time point regarded as peak of the "regeneration phase") could help understanding this divergence. Particularly, as we assume

that loss of the acinar cell fate and missing redifferentiation of progenitor-like cells into acinar cells is driving the sustained inflammation in $Kras^{G12D}$ mice, we are interested in identification of intercellular singals impacting acinar homeostasis genes which are required for maintenance of the acinar phenotype but also redifferentiation.

Here, we applied a correlation-based approach in order to assemble networks of potential interactions. We assembled networks on the time segment from 3h to 84h for each mouse strain separately, incorporating previously introduced acinar-specific genes (literature-based list of 77 genes) and genes annotated by the GO term "cell-cell signaling". Finally, we extracted interactions which were identified exclusively for one of both mouse strains (using two different q-value cut-offs after Bonferroni correction). Analysis of the resulting networks showed that four members of the Fgf/Fgfr family were impacting acinar homeostasis genes with several of these genes being highly connected. For $Kras^{G12D}$ mice, on the contrary, only a single member of this family, Fgfr1, was included with a single interaction. As interactions involving Fgf/Fgfr members are significantly enriched among the identified interactions with acinar homeostasis genes in wild-type mice (p¡0.012), we conclude that Fgf signaling and expression of Fgf receptors are potential driving forces of the initiation of the inflammation-regeneration transition. Particularly, we found Fgfr3 being highly connected in the wild-type network; also Fgfr3 was predicted to interact with Mist1 directly. This result was first surprising as Fgfr3 is commonly regarded as oncogene also mediating mesenchymal expansion [255]. But notably, recent studies have shown that Fgfr3 has tumor-suppressing properties when expressed in epithelial cells [255]. Our findings confirm this result and indicate that Fgfr3 activity is potentially required for successful regeneration of the acinar components. In addition to this specific result, structural comparison of the assembled correlation networks shows that the number of intercellular signaling genes interacting with epithelial cell-linked genes is greater in wild-type mice. For $Kras^{G12D}$ mice, only a small number (14) of isolated interactions was predicted. Furthermore, among these interactions, we found several development-related genes to be involved (Gata4, Gata6, Hnf1b) but none of the most important acinar homeostasis-related genes (Ptf1a, Mist1). This further underscores that potentially, several exterior signaling effects required for induction of the acinar repair mechanism are absent in the anticipated regeneration phase of $Kras^{G12D}$ mice.

We chose pearson correlation coefficients with predefined q-value cutoffs for prediction of interactions. Accordingly, resulting networks must be treated with care, as it cannot be assumed that highly correlated genes interact with each other in general. Particularly, confounding effects must be expected: Two variables can be highly correlated due to a third variable (confounder) controlling both of them. This problem has been addressed in methods which use partial correlations instead of pearson correlations (e.g. GeneNet by Strimmer et al. [212]). However, partial correlation approaches are unsuitable for our dataset due to the extremely large number of variables compared to the number of samples. Also, our approach does not incorporate the temporal information from our dataset, as sample reordering does not affect calculation of the pearson correlation coefficient. However, as the temporal gradient for a small time frame which we analyzed (3h - 84h) is unclear with high variance across samples from surrounding

time points, we decided to exclude this information on purpose. Still, though we cannot assume that each individual predicted interaction is correct, the significant frequency of Fgf/Fgfr related genes and the large number of interactions with relevant genes (particularly Mist1) indicates that the corresponding signaling pathway has an important role for mediation of the inflammation-regeneration transition in our model.

Accordingly, we suggest to perform *in vitro* studies with populations of acinar, progenitor-like and mesenchymal cells in order to investigate the effect Fgf/Fgfr during pancreatic regeneration. Here, transcriptional profiling for the distinct cell types will further help to unveil other intercellular mechanisms mediating the described regeneration process. Furthermore, studies on Kras^{G12D} mice with cell-sorting based experiments could help unveil the impact of distinct cytokines and chemokines on tissue transformation, as we observe sustained up-regulation of these genes after inflammatory injury.

Conclusion

Our analysis provides new insights into the mechanisms of early pancreatic carcinogenesis. We found that

- Normal pancreatic regeneration is a default process characterized by three distinct stages: inflammation, regeneration and refinement.
- Regeneration is blocked in $Kras^{G12D}$-mutated mice and characterized by a sustained inflammatory status which we are able to describe with a signature.
- Altered proliferation cascades of acinar, progenitor-like and mesenchymal cells are characterizing early pancreatic carcinogenesis. Massive mesenchymal expansion is observed.

Application of the homologous genes of the extracted gene signature to a dataset of 45 human PDAC samples showed that

- PDAC patients are separable with regard to expression of the inflammation-specific signature.
- According to survival analysis, inflammation-specific signature can be possibly regarded as an indicator of an early stage of pancreatic carcinogenesis.

We studied our "dual model" of pancreatic regeneration and early pancreatic carcinogenesis, with special respect to the *inflammation-regeneration transition*, using correlation networks. We found that

- Interactions involving members of the Fgf/Fgfr family are overrepresented in the wild-type network, suggesting that corresponding genes could play a major role for successful regeneration of the pancreas.
- Fgfr3 is a highly connected player in the network. This is in line with recent studies repor ting that Fgfr3 expression in epithelial cells limits cell growth and promotes differentiation.

We suggest to continue with follow-up studies focussing on the distinct impact of Fgf/Fgfr members and selected cytokines and chemokines on the regeneration in wild-type mice or tissue transformation in $Kras^{G12D}$ mice.

Bibliography

[1] Drenckhahn D, Anatomie, vol. 1, edition 17. Urban&Fischer, München 2008, 723.

[2] Jaspers M, Sonographie, organ- und leitsymptomorientiert, Springer, 2011, 89.

[3] Graumann W, Sasse D, Compact Lehrbuch Anatomie, vol. 3, Stuttgart 2004, 120-122.

[4] Swift GH et al., An Endocrine-Exocrine Switch in the Activity of the Pancreatic Homeodomain Protein PDX1 through Formation of a Trimeric Complex with PBX1b and MRG1 (MEIS2), Mol. Cell. Biol. September 1998, 18(9):5109-5120.

[5] Neoptolemos JP et al., Pancreatic Cancer, Springer, 2010.

[6] `http://www.chasepto.org/wp-content/uploads/2015/01/Inflammation-of-the-pancreas.jpg`

[7] `http://medcell.med.yale.edu/systems_cell_biology_old/liver_and_pancreas/images/pancreas_cartoon.jpg`

[8] Collins MA et al., MAPK signaling is required for dedifferentiation of acinar cells and development of pancreatic intraepithelial neoplasia in mice, Gastroenterology 2014, 146(3):822-834.

[9] Houbracken I et al., Lineage tracing evidence for transdifferentiation of acinar to duct cells and plasticity of human pancreas, Gastroenterology 2011, 141(2):731-41.

[10] Grapin-Botton A, Ductal cells of the pancreas, Int J Biochem Cell Biol. 2005, 37(3):504-10.

[11] Kopp JL et al., Sox9+ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas, Development. 2011, 138(4):653-65.

[12] Madden ME, Sarras MP Jr., Morphological and biochemical characterization of a human pancreatic ductal cell line (PANC-1), Pancreas 1988, 3(5):512-28.

[13] Elayat AA, el-Naggar MM, Tahir M, An immunocytochemical and morphometric study of the rat pancreatic islets, Journal of Anatomy 1995, 186(3):629-37.

[14] Attali M et al., Control of beta-cell differentiation by the pancreatic mesenchyme, Diabetes 2007, 56(5):1248-58.

[15] Dangi-Garimella S et al., Pancreatic Cancer and Tumor Microenvironment, Chapter 5, Trivandrum (India) Transworld Research Network, 2012.

[16] Masamune A et al., Roles of pancreatic stellate cells in pancreatic inflammation and fibrosis, Clin Gastroenterol Hepatol. 2009 7(11 Suppl):S48-54.

[17] Means AL, Pancreatic stellate cells: small cells with a big role in tissue homeostasis, Laboratory Investigation (2013) 93, 4-7.

[18] Kikuta K et al., Pancreatic stellate cells promote epithelial-mesenchymal transition in pancreatic cancer cells, Biochem Biophys Res Commun. 2010, 403(3-4):380-4.

[19] Gittes G et al., Developmental biology of the pancreas: A comprehensive review, Developmental Biology 2009, 326:4-35.

[20] Bort R et al., Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. Development 2004; 131:797-806.

[21] Kabashima-Niibe A et al., Mesenchymal stem cells regulate epithelial-mesenchymal transition and tumor progression of pancreatic cancer cells, Cancer Sci. 2013 Feb;104(2):157-64.

[22] McLin VA, Rankin SA, Zorn AM, Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development. Development 2007; 134:2207-17.

[23] Wells JM, Melton DA, Early mouse endoderm is patterned by soluble factors from adjacent germ layers. Development 2000; 127:1563-72.

[24] Burlison JS et al., Pdx-1 and Ptf1a concurrently determine fate specification of pancreatic multipotent progenitor cells, Dev Biol. 2008 Apr 1; 316(1): 74-86.

[25] Gu G, Dubauskaite J, Melton DA. Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. Development 2002; 129:2447-57.

[26] Jonsson J, Carlsson L, Edlund T, Edlund H. Insulin-promoter-factor 1 is required for pancreas development in mice. Nature 1994; 371:606-9.

[27] Hesselson D, Anderson RM, Stainier DY, Suppression of Ptf1a activity induces acinar-to-endocrine conversion., Curr Biol. 2011 Apr 26;21(8):712-7.

[28] Yoshitomi H, Zaret KS, Endothelial cell interactions initiate dorsal pancreas development by selectively inducing the transcription factor Ptf1a, Development. 2004 Feb;131(4):807-17.

[29] Spence JR, Lange AW, Lin SC, Kaestner KH, Lowy AM, Kim I, et al., Sox17 regulates organ lineage segregation of ventral foregut progenitor cells. Dev Cell 2009; 17:62-74.

[30] Harrison KA, Thaler J, Pfaff SL, Gu H, Kehrl JH, Pancreas dorsal lobe agenesis and abnormal islets of Langerhans in Hlxb9-deficient mice, Nat Genet 1999; 23:71-5.

[31] Jacquemin P, Durviaux SM, Jensen J, Godfraind C, Gradwohl G, Guillemot F, et al., Transcription factor hepatocyte nuclear factor 6 regulates pancreatic endocrine cell differentiation and controls expression of the proendocrine gene ngn3, Mol Cell Biol 2000; 20:4445-54.

[32] Gannon M, Ray MK, Van Zee K, Rausa F, Costa RH, Wright CV, Persistent expression of HNF6 in islet endocrine cells causes disrupted islet architecture and loss of beta cell function. Development 2000; 127:2883-95.

[33] Gao N, LeLay J, Vatamaniuk MZ, Rieck S, Friedman JR, Kaestner KH, Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. Genes Dev 2008; 22:3435-48.

[34] Lee CS, Sund NJ, Behr R, Herrera PL, Kaestner KH. Foxa2 is required for the differentiation of pancreatic alpha-cells. Dev Biol 2005; 278:484-95.

[35] Haumaitre C, Barbacci E, Jenny M, Ott MO, Gradwohl G, Cereghini S, Lack of TCF2/vHNF1 in mice leads to pancreas agenesis. Proc Natl Acad Sci USA 2005; 102:1490-5.

[36] Solar M, Cardalda C, Houbracken I, Martin M, Maestro MA, De Medts N, et al, Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. Dev Cell 2009; 17:849-60.

[37] Xuan S, Borok MJ, Decker KJ, et al., Pancreas-specific deletion of mouse Gata4 and Gata6 causes pancreatic agenesis, J Clin Invest. 2012 Oct 1; 122(10): 3516-3528.

[38] Akiyama H, Kim JE, Nakashima K, Balmes G, Iwai N, Deng JM, et al. Osteochondroprogenitor cells are derived from Sox9 expressing precursors. Proc Natl Acad Sci USA 2005; 102:14665-70.

[39] Esni F, Ghosh B, Biankin AV, Lin JW, Albert MA, Yu X, et al., Notch inhibits Ptf1 function and acinar cell differentiation in developing mouse and zebrafish pancreas. Development 2004; 131:4213-24.

[40] Zhou Q, Law AC, Rajagopal J, Anderson WJ, Gray PA, Melton DA, A multipotent progenitor domain guides pancreatic organogenesis. Dev Cell 2007; 13:103-14.

[41] Fujikura J, Hosoda K, Iwakura H, Tomita T, Noguchi M, Masuzaki H, et al. Notch/Rbp-j signaling prevents premature endocrine and ductal cell differentiation in the pancreas. Cell Metab 2006; 3:59-65.

[42] Macdonald RJ, Swift GH, Real FX, Transcriptional control of acinar development and homeostasis. Prog Mol Biol Transl Sci 2010; 97:1-40.

[43] Smith GB, Banga N, Tobi MH, Imber CJ, Pancreatic adenocarcinoma, BMJ, 2012 May 16;344:e2476.

[44] Afelik S, Qu X, Hasrouni E, Bukys MA, Deering T, Nieuwoudt S, Rogers W, Macdonald RJ, Jensen J, Notch-mediated patterning and cell fate allocation of pancreatic progenitor cells, Developmentm 2012 139(10):1744-53.

[45] Schaffer AE, Freude KK, Nelson SB, Sander M, Nkx6 transcription factors and Ptf1a function as antagonistic lineage determinants in multipotent pancreatic progenitors. Dev Cell 2010; 18:1022-9.

[46] Puri S, Hebrok M. Cellular plasticity within the pancreas–lessons learned from development. Dev Cell 2010; 18:342-56.

[47] Kopp JL, Dubois CL, Hao E, Progenitor cell domains in the developing and adult pancreas, Cell Cycle 10:12, 1921-1927; June 15, 2011.

[48] Nishimura W, Kondo T, Salameh T, El Khattabi I, Dodge R, Bonner-Weir S, et al., A switch from MafB to MafA expression accompanies differentiation to pancreatic beta-cells. Dev Biol 2006; 293:526-39.

[49] Wilson ME, Scheel D, German MS, Gene expression cascades in pancreatic development. Mech Dev 2003; 120:65-80.

[50] Lankisch PG, Apte M, Banks PA, Acute pancreatitis, Lancet. 2015 Jan 20.

[51] Frossard JL, Steer ML, Pastor CM. Acute pancreatitis. Lancet 2008; 371:143- 52.

[52] Gaisano HY, Gorelick FS, New insights into the mechanisms of pancreatitis. Gastroenterology 2009; 136:2040-4.

[53] Lerch MM, Halangk W, Human pancreatitis and the role of cathepsin B. Gut 2006; 55:1228-30.

[54] Sebastiano PD, di Mola FF, Bockman DE, Friess H, Büchler MW, Gut. 2003 Jun; 52(6): 907-911.

[55] Mitchell RM, Byrne MF, Baillie J. Pancreatitis, Lancet 2003; 361:1447-55.

[56] Bülcher MW et al., Diseases of the pancreas, Karger 1999, p. 90.

[57] Witt H, Apte MV, Keim V, Wilson JS, Chronic pancreatitis: challenges and advances in pathogenesis, genetics, diagnosis, and therapy. Gastroenterology 2007; 132:1557-73.

[58] Rosendahl J, Witt H, Szmola R, Bhatia E, Ozsvari B, Landt O, et al. Chymotrypsin C (CTRC) variants that diminish activity or secretion are associated with chronic pancreatitis. Nat Genet 2008; 40:78-82.

[59] Pfutzer RH, Barmada MM, Brunskill AP, Finch R, Hart PS, Neoptolemos J, et al. SPINK1/PSTI polymorphisms act as disease modifiers in familial and idiopathic chronic pancreatitis. Gastroenterology 2000; 119:615-23.

[60] Kamisawa T, Okamoto A, Autoimmune pancreatitis: proposal of IgG4-related sclerosing disease, J Gastroenterol. 2006 Jul;41(7):613-25.

[61] Vonlaufen A, Wilson JS, Apte MV, Molecular mechanisms of pancreatitis: current opinion. J Gastroenterol Hepatol 2008; 23:1339-48.

[62] Paliwal S, Bhaskar S, Chandak GR, Genetic and phenotypic heterogeneity in tropical calcific pancreatitis Sumit Paliwal, Seema Bhaskar, and Giriraj R Chandak, World J Gastroenterol. 2014 Dec 14; 20(46): 17314-17323.

[63] Bhatia M, Inflammatory response on the pancreatic acinar cell injury, Scandinavian Journal of Surgery 94:97-102, 2005.

[64] Vonlaufen A, Joshi S, Qu C, Phillips PA, Xu Z, Parker NR, et al. Pancreatic stellate cells: partners in crime with pancreatic cancer cells. Cancer Res 2008; 68:2085-93.

[65] Zhu L, Tran T, Rukstalis JM et al., Inhibition of Mist1 Homodimer Formation Induces Pancreatic Acinar-to-Ductal Metaplasia, Mol Cell Biol. 2004 Apr; 24(7): 2673-2681.

[66] Otsuki M, Yamamoto M, Yamaguchi T. Animal models of chronic pancreatitis. Gastroenterol Res Pract 2010; 2010:403295.

[67] Zaninovic V, Gukovskaya AS, Gukovsky I, Mouria M, Pandol SJ, Cerulein upregulates ICAM-1 in pancreatic acinar cells, which mediates neutrophil adhesion to these cells, Am J Physiol Gastrointest Liver Physiol. 2000 Oct;279(4):G666-76.

[68] Hyeyoung K, Cerulein Pancreatitis: Oxidative Stress, Inflammation, and Apoptosis, Gut Liver. 2008 Sep; 2(2): 74-80.

[69] Niederau C, Ferrell LD, Grendell JH, Caerulein-induced acute necrotizing pancreatitis in mice: protective effects of proglumide, benzotript, and secretin, Gastroenterology. 1985 May;88(5 Pt 1):1192-204.

[70] Hezel AF, Kimmelman AC, Stanger BZ, Bardeesy N, Depinho RA, Genetics and biology of pancreatic ductal adenocarcinoma. Genes Dev 2006; 20:1218-49.

[71] Ruben P, Hansen JT, TNM Staging Atlas, Ch. 28, Wolters Kluwer, 2012.

[72] Porta M, Fabregat X, Malats N, et al., Exocrine pancreatic cancer: symptoms at presentation and their relation to tumour site and stage., Clin Transl Oncol. 2005 Jun;7(5):189-97.

[73] Krech RL, Walsh D, Symptoms of pancreatic cancer, J Pain Symptom Manage. 1991 Aug;6(6):360-7.

[74] Iacobuzio-Donahue CA, Fu B, Yachida S, Luo M, Abe H, Henderson CM, et al., DPC4 gene status of the primary carcinoma correlates with patterns of failure in patients with pancreatic cancer. J Clin Oncol 2009; 27:1806-13.

[75] Stathis A, Moore MJ, Advanced pancreatic carcinoma: current treatment and future challenges. Nat Rev Clin Oncol 2010; 7:163-72.

[76] Maitra A, Hruban RH, Pancreatic cancer, Annu Rev Pathol 2008; 3:157-88.

[77] Hidalgo M. Pancreatic cancer. N Engl J Med 2010; 362:1605-17.

[78] Gore J, Korc M, Pancreatic cancer stroma: friend or foe?, Cancer Cell, 2014.

[79] Rhim AD, Oberstein PE, Thomas DH, Stromal Elements Act to Restrain, Rather Than Support, Pancreatic Ductal Adenocarcinoma, Cell, 2014, 25(6):735-747.

[80] Provenzano PP, Cuevas C, Chang AE, Goel VK, Von Hoff DD, Hingorani SR, Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma, Cancer Cell, 2012, 21:418-429.

[81] Özdemir BC, Pentcheva-Hoang T, Castens JL, Zheng X, Wu CC, Simpson T, Laklai H, Sugimoto H, Kahlert C, Novitskiy SV, et al., Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival, Cancer Cell, 2014, 25:719-734.

[82] Lee MJ et al., Expression of Mucins and Cytokeratins in Primary Carcinomas of the Digestive System, Mod Pathol 2003;16(5):403-410

[83] Ishiwata T, Pancreatic Ductal Adenocarcinoma: Basic and Clinical Challenges for Better Prognosis. J Carcinogene Mutagene S9:005, 2013.

[84] Hoff DD, Evans DB, Hruban RH, Pancreatic Cancer, Jones and Bartlett Pub., 2004.

[85] Grippo PJ, Pancreatic cancer and the tumor microenvironment: Mesenchymes role in pancreatic carcinogenesis, Pancreatic Cancer Tumor Microenvironment, Transworld Research Network, 2012.

[86] Chiu J, Yau T, Metastatic Pancreatic Cancer: Are We Making Progress in Treatment?, Gastroenterology Research and Practice Volume 2012, Article ID 898931.

[87] Raimondi S, Maisonneuve P, Lowenfels AB, Epidemiology of pancreatic cancer: an overview, Nat Rev Gastroenterol Hepatol 2009, 6:699-708.

[88] Iodice S, Gandini S, Maisonneuve P, Lowenfels AB, Tobacco and the risk of pancreatic cancer: a review and meta-analysis, Langenbecks Arch Surg 2008, 393:535-45.

[89] Milne RL, Greenhalf W, Murta-Nascimento C, Real FX, Malats N, The inherited genetic component of sporadic pancreatic adenocarcinoma. Pancreatology 2009, 9:206-14.

[90] Landi S, Genetic predisposition and environmental risk factors to pancreatic cancer: A review of the literature. Mutat Res 2009; 681:299-307.

[91] McWilliams RR, Petersen GM, Rabe KG, Holtegaard LM, Lynch PJ, Bishop MD, et al. Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations and risk for pancreatic adenocarcinoma. Cancer 2010; 116:203-9.

[92] Rustgi AK, Familial pancreatic cancer: genetic advances, Genes Dev. 2014 Jan 1;28(1):1-7.

[93] Murtaugh LC, Putting GWAS to the functional test: NR5A2 and pancreatic cancer risk, Gut. 2014 Apr;63(4):535-6.

[94] Malumbres M, Barbacid M, RAS oncogenes: the first 30 years. Nat Rev Cancer 2003; 3:459-65.

[95] Cheng JQ, Rugerri B, Klein MW et al., Amplification of AKT2 in human pancreatic cells and inhibition of AKT2 expression and tumorigenicity by antisense RNA, Proc Natl Acad Sci U S A 1996; 93:3636-41.

[96] Hill R, Calvopina JH, Kim C, et al., PTEN loss accelerates KrasG12D-induced pancreatic cancer development, Cancer Res. 2010 Sep 15;70(18):7114-24.

[97] Calhoun ES, Jones JB, Ashfaq R, Adsay V, Baker SJ, Valentine V, et al., BRAF and FBXW7 (CDC4, FBW7, AGO, SEL10) mutations in distinct subsets of pancreatic cancer: potential therapeutic targets, Am J Pathol 2003; 163:1255-60.

[98] Moskaluk CA, Hruban RH, Kern SE. p16 and K-ras gene mutations in the intraductal precursors of human pancreatic adenocarcinoma. Cancer Res 1997; 57:2140-3.

[99] Wilentz RE, Geradts J, Maynard R, Offerhaus GJ, Kang M, Goggins M, et al. Inactivation of the p16 (INK4A) tumor-suppressor gene in pancreatic duct lesions: loss of intranuclear expression. Cancer Res 1998; 58:4740-4.

[100] Rayess H, Wang MB, Srivatsan ES, Cellular senescence and tumor suppressor gene p16, 2012, Int. J. Cancer 130(8):1715-25.

[101] Bradshaw RA, Dennis EA, Handbook of cell signaling, 2010, p. 2614.

[102] Bierie B, Moses HL, TGF-beta and cancer, Cytokine Growth Factor Rev 2006; 17:29-40.

[103] Zavadil J, Bottinger EP. TGF-beta and epithelial-to-mesenchymal transitions. Oncogene 2005; 24:5764-74.

[104] Kato K, Kamada H, Fujimori T et al., Molecular Biologic Approach to the Diagnosis of Pancreatic Carcinoma Using Specimens Obtained by EUS-Guided Fine Needle Aspiration, Gastroenterology Research and Practice Volume 2012 (2012), Article ID 243524.

[105] Jonckheere N, Skrypek N, van Seuningen I, Mucins and Pancreatic Cancer, Cancers 2010, 2, 1794-1812.

[106] Cuschieri A, Hanna GB, Essential Surgical Practice, CRC press, 2015, p. 821.

[107] Eser S, Schnieke A, Schneider G, Saur D, Oncogenic KRAS signalling in pancreatic cancer, British Journal of Cancer (2014) 111, 817-822.

[108] Hruban RH, Goggins M, Parsons J, Kern SE. Progression model for pancreatic cancer. Clin Cancer Res 2000; 6:2969-72.

[109] Cruz-Monserrate Z, Qiu S, Evers BM, O'connor KL, Upregulation and redistribution of integrin alpha6bold italic beta4 expression occurs at an early stage in pancreatic adenocarcinoma progression, Modern Pathology (2007) 20, 656-667.

[110] `http://what-when-how.com/wp-content/uploads/2012/04/tmpDC131.jpg`

[111] Real FX, Cibriun-Uhalte E, Martinelli P, Pancreatic cancer development and progression: remodeling the model, Gastroenterology. 2008 Sep;135(3):724-8.

[112] Thayer SP, di Magliano MP, Heiser PW, Nielsen CM, Roberts DJ, Lauwers GY, et al., Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis, Nature 2003; 425:851-6.

[113] Miyamoto Y, Maitra A, Ghosh B, Zechner U, Argani P, Iacobuzio-Donahue CA, et al., Notch mediates TGF alpha-induced changes in epithelial differentiation during pancreatic tumorigenesis. Cancer Cell 2003; 3:565-76.

[114] De La OJ, Emerson LL, Goodman JL, Froebe SC, Illum BE, Curtis AB, et al., Notch and Kras reprogram pancreatic acinar cells to ductal intraepithelial neoplasia. Proc Natl Acad Sci U S A 2008; 105:18907-12.

[115] De La OJ, Murtaugh LC, Notch signaling: where pancreatic cancer and differentiation meet? Gastroenterology 2009; 136:1499-502.

[116] Cook N, Frese KK, Bapiro TE et al., Gamma secretase inhibition promotes hypoxic necrosis in mouse pancreatic ductal adenocarcinoma, J Exp Med. 2012 Mar 12; 209(3): 437-444.

[117] Pasca di Magliano M, Biankin AV, Heiser PW, Cano DA, Gutierrez PJ, Deramaudt T, et al., Common activation of canonical Wnt signaling in pancreatic adenocarcinoma. PLoS One 2007; 2:e1155.

[118] Pujal J, Capella G, Real FX. The Wnt pathway is active in a small subset of pancreas cancer cell lines. Biochim Biophys Acta 2006; 1762:73-9.

[119] Heiser PW, Lau J, Taketo MM, Herrera PL, Hebrok M, Stabilization of beta-catenin impacts pancreas growth. Development 2006; 133:2023-32.

[120] Heiser PW, Cano DA, Landsman L, Kim GE, Kench JG, Klimstra DS et al., Stabilization of beta-catenin induces pancreas tumor formation, Gastroenerology, 2008, 135:1288-300.

[121] Morran DC, Wu J, Jamieson NB, et al., Targeting mTOR dependency in pancreatic cancer, Gut. 2014 Sep;63(9):1481-9.

[122] Kong B, Wu W, Cheng T, Schlitter AM, Qian C, Bruns P, Jian Z, Jäger C, Regel I, Raulefs S, Behler N, Irmler M, Beckers J, Friess H, Erkan M, Siveke JT, Tannapfel A, Hahn SA, Theis FJ, Esposito I, Kleeff J, Michalski CW, A subset of metastatic pancreatic ductal adenocarcinomas depends quantitatively on oncogenic Kras/Mek/Erk-induced hyperactive mTOR signalling, Gut. 2015 Jan 19 [Epub ahead of print].

[123] Kosmahl M, Pauser U, Peters K, Sipos B, Lüttges J, Kremer B, Klöppel G, Cystic neoplasms of the pancreas and tumor-like lesions with cystic features: a review of 418 cases and a classification proposal, Virchows Arch. 2004 Aug;445(2):168-78.

[124] Gress T, Neoptolemos J, Lemoine NR, Real FX, Exocrine pancreas cancer. Hannover: Felsenstein CCCP, 2004.

[125] Crippa S, Fernandez-Del Castillo C, Salvia R, Finkelstein D, Bassi C, Dominguez I, et al., Mucin-producing neoplasms of the pancreas: an analysis of distinguishing clinical and epidemiologic characteristics. Clin Gastroenterol Hepatol 2010; 8:213-9.

[126] Mulkeen AL, Yoo PS, Cha C, Less common neoplasms of the pancreas. World J Gastroenterol 2006; 12:3180-5.

[127] Yonezawa S, Nakamura A, Horinouchi M, Sato E, The expression of several types of mucin is related to the biological behavior of pancreatic neoplasms, J Hepatobiliary Pancreat Surg 2002; 9:328-341.

[128] Fernandez-del Castillo C, Adsay NV. Intraductal papillary mucinous neoplasms of the pancreas. Gastroenterology 2010; 139:708-13, 13 e1-2.

[129] Hoorens A, Lemoine NR, McLellan E, Morohoshi T, Kamisawa T, Heitz PU, et al., Pancreatic acinar cell carcinoma. An analysis of cell lineage markers, p53 expression, and Ki-ras mutation. Am J Pathol 1993; 143:685-98.

[130] Klimstra DS, Nonductal neoplasms of the pancreas. Mod Pathol 2007; 20, Suppl 1:S94-112.

[131] Abraham SC, Klimstra DS, Wilentz RE, Yeo CJ, Conlon K, Brennan M, Cameron JL, Wu TT, Hruban RH, Solid-pseudopapillary tumors of the pancreas are genetically distinct from pancreatic ductal adenocarcinomas and almost always harbor beta-catenin mutations, Am J Pathol, 2002, 160(4):1361-9.

[132] Saif MW, Pancreatoblastoma, JOP. J Pancreas (Online) 2007; 8(1):55-63.

[133] Halfdanarson TR, Rubin J, Farnell MB et al., Pancreatic endocrine neoplasms: Epidemiology and prognosis of pancreatic endocrine tumors, Endocr Relat Cancer. 2008 Jun; 15(2): 409-427.

[134] House MG, Herman JG, Guo MZ et al., PMCID: PMC1422710 Aberrant Hypermethylation of Tumor Suppressor Genes in Pancreatic Endocrine Neoplasms, Ann Surg. 2003 Sep; 238(3): 423-432.

[135] Notohara K, Hamazaki S, Tsukayama C, Solid-pseudopapillary tumor of the pancreas: immunohistochemical localization of neuroendocrine markers and CD10, Am J Surg Pathol. 2000 Oct;24(10):1361-71.

[136] Bailey JM, Leach SD, Signaling pathways mediating epithelial- mesenchymal crosstalk in pancreatic cancer: Hedgehog, Notch and TGFβ, Pancreatic Cancer Tumor Microenvironment, Transworld Research Network, 2012.

[137] Hwang RF, Moore T, Arumugam T, et al, Cancer-associated stromal fibroblasts promote pancreatic tumor progression, Cancer Res. 2008;68:918-926.

[138] Fujita H, Ohuchida K, Mizumoto K, et al., Tumor-stromal interactions with direct cell contacts enhance proliferation of human pancreatic carcinoma cells. Cancer Sci. 2009;100:2309-2317.

[139] Yauch RL, Gould SE, Scales SJ, et al., A paracrine requirement for hedgehog signalling in cancer, Nature. 2008;455:406-410.

[140] Tian H, Callahan CA, DuPree KJ, et al., Hedgehog signaling is restricted to the stromal compartment during pancreatic carcinogenesis, Proc Natl Acad Sci U S A. 2009;106:4254-4259.

[141] Olive KP, Jacobetz MA, Davidson CJ, et al., Inhibition of Hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer, Science. 2009;324:1457-1461.

[142] Kopinke D, Brailsford M, Shea JE, Leavitt R, Scaife CL, Murtaugh LC, Lineage tracing reveals the dynamic contribution of Hes1+ cells to the developing and adult pancreas, Development. 2011;138:431-441.

[143] Friess H, Yamanaka Y, Buchler M, et al., Enhanced expression of the type II transforming growth factor beta receptor in human pancreatic cancer cells without alteration of type III receptor expression, Cancer Res. 1993;53:2704-2707.

[144] Kalluri R, Weinberg RA, The basics of epithelial-mesenchymal transition, J ClinInvest 2009; 119:1420-8.

[145] Polyak K, Weinberg RA, Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits, Nat Rev Cancer 2009; 9:265-73.

[146] `http://www.rndsystems.com/Resources/Images/6868.pdf`

[147] Thiery JP, Acloque H, Huang RY, Nieto MA, Epithelial-mesenchymal transitions in development and disease, Cell 2009; 139:871-90.

[148] Gort EH, Groot AJ, van der Wall E, van Diest PJ, Vooijs MA, Hypoxic regulation of metastasis via hypoxia-inducible factors, Curr Mol Med 2008; 8:60-7.

[149] Lee KA, Nelson CM, New Insights into the Regulation of Epithelial-Mesenchymal Transition and Tissue Fibrosis, International Review of Cell and Molecular Biology, 29:171-221.

[150] Rhim AD, Stanger BZ, Molecular Biology of Pancreatic Ductal Adenocarcinoma Progression: Aberrant Activation of Developmental Pathways, Prog Mol Biol Transl Sci, 2011;97:4178.

[151] Lamouille S, Xu J, Derynck R, Molecular mechanisms of epithelial-mesenchymal transition Nature Reviews Molecular Cell Biology 2014, 15:178-196.

[152] Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al., The miR- 200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1, Nat Cell Biol 2008; 10:593-601.

[153] Park SM, Gaur AB, Lengyel E, Peter ME, The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2, Genes Dev 2008; 22:894-907.

[154] Karayiannakis AJ, Syrigos KN, Chatzigianni E, Papanikolaou S, Alexiou D, Kalahanis N, et al., Aberrant E-cadherin expression associated with loss of differentiation and advanced stage in human pancreatic cancer, Anticancer Res 1998; 18:4177-80.

[155] Pignatelli M, Ansari TW, Gunter P, Liu D, Hirano S, Takeichi M, et al., Loss of membranous E-cadherin expression in pancreatic cancer: correlation with lymph node metastasis, high grade, and advanced stage, J Pathol 1994; 174:243-8.

[156] Wang Z, Li Y, Kong D, Banerjee S, Ahmad A, Azmi AS, et al., Acquisition of epithelial-mesenchymal transition phenotype of gemcitabine-resistant pancreatic cancer cells is linked with activation of the notch signaling pathway, Cancer Res 2009; 69:2400-7.

[157] Normanno N, Tejpar S, Morgillo F, Implications for KRAS status and EGFR-targeted therapies in metastatic CRCl, Nature Reviews Clinical Oncology, 2009, 6:519-527.

[158] Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013, CA Cancer J Clin. Jan 2013;63(1):11-30.

[159] Hingorani SR, Wang L, Multani AS, et al., Trp53R172H and KrasG12D cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice., Cancer Cell. May 2005;7(5):469-483.

[160] Guerra C, Collado M, Navas C, et al., Pancreatitis-induced inflammation contributes to pancreatic cancer by inhibiting oncogene-induced senescence, Cancer Cell. Jun 14 2011;19(6):728-739.

[161] Eshleman JR, Norris AL, Sadakari Y, et al., KRAS and GNAS Mutations in Pancreatic Juice Collected From the Duodenum of Patients at High Risk for Neoplasia Undergoing Endoscopic Ultrasound, Clin Gastroenterol Hepatol. Dec 3 2014.

[162] Guerra C, Schuhmacher AJ, Canamero M, et al., Chronic pancreatitis is essential for induction of pancreatic ductal adenocarcinoma by K-Ras oncogenes in adult mice, Cancer Cell. Mar 2007;11(3):291-302.

[163] Carriere C, Young AL, Gunn JR, Longnecker DS, Korc M, Acute pancreatitis markedly accelerates pancreatic cancer progression in mice expressing oncogenic Kras, Biochem Biophys Res Commun. May 8 2009;382(3):561-565.

[164] Maniati E, Bossard M, Cook N, et al., Crosstalk between the canonical NF-kappaB and Notch signaling pathways inhibits Ppargamma expression and promotes pancreatic cancer progression in mice, J Clin Invest. Dec 2011;121(12):4685-4699.

[165] Ling J, Kang Y, Zhao R, et al., KrasG12D-induced IKK2/beta/NF-kappaB activation by IL-1alpha and p62 feedforward loops is required for development of pancreatic ductal adenocarcinoma., Cancer Cell. Jan 17 2012;21(1):105-120.

[166] Schlomann U, Koller G, Conrad C et al., ADAM8 as a drug target in pancreatic cancer, Nat Commun. 2015 Jan 28;6:6175.

[167] Deng X, Friedman E, et al., Mirk kinase inhibition blocks the in vivo growth of pancreatic cancer cells, Genes Cancer. 2014 Sep;5(9-10):337-47.

[168] Baer R, Cintas C, Dufresne M, Pancreatic cell plasticity and cancer initiation induced by oncogenic Kras is completely dependent on wild-type PI 3-kinase p110α, Genes Dev. 2014 Dec 1;28(23):2621-35.

[169] Collins MA, Bednar F, Zhang Y, et al., Oncogenic Kras is required for both the initiation and maintenance of pancreatic cancer in mice, J Clin Invest. 2012 Feb;122(2):639-53.

[170] O'Hagan RC, Heyer J, KRAS Mouse Models, Genes Cancer. 2011 Mar; 2(3): 335-343.

[171] Greer RL, Staley BK, Liou A, Hebrok M, Numb regulates acinar cell dedifferentiation and survival during pancreatic damage and acinar-to-ductal metaplasia, Gastroenterology. Nov 2013;145(5):1088-1097 e1088.

[172] Malone JH, Oliver B, Microarrays, deep sequencing and the true measure of the transcriptome, BMC Biology, 9(1):34, 2011.

[173] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP, Light-generated oligonucleotide arrays for rapid dna sequence analysis, Proc Natl Acad Sci U S A, 91(11):5022-5026, May 1994.

[174] Bammler T, Beyer RP, Bhattcharya S et al. (Members of the Toxicogenomics Research Consortium), Standardizing global gene expression analysis between laboratories and across platforms, Nat Methods, 2(5):351-356, May 2005.

[175] Spruill SE, Lu J, Hardy S, Weir B, Assessing sources of variability in microarray gene expression data. Biotechniques, 33(4):916-20, 922-3, Oct 2002.

[176] Yang YH, Dudoit S, Luu P et al., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, 2002, Nucleic Acids Research vol. 30(4):e15.

[177] Hoeffding W, A Non-Parametric Test of Independence, Ann. Math. Statist. Volume 19, Number 4 (1948), 546-557.

[178] Wilson CL, Miller CJ, Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis, Bioinformatics. 2005, 15;21(18):3683-5. Epub 2005 Aug 2.

[179] Kauffmann A, Gentleman R, Huber W, Bioinformatics. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data, 2009 Feb 1;25(3):415-6.

[180] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP, Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Research, 2003, 31(4):e15.

[181] Bolstad BM, Irizarry RA, Astrand M, Speed TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics. 2003 Jan 22;19(2):185-93.

[182] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB, Large-scale analysis of the human and mouse transcriptomes, Proc Natl Acad Sci USA. 2002 Apr 2;99(7):4465-70.

[183] Gentleman R, Carey V, Huber W and Hahne F, genefilter: genefilter: methods for filtering genes from high-throughput experiments, R package version 1.50.0.

[184] Kruskal JB, Rank, decomposition, and uniqueness for 3-way and n-way arrays. North-Holland Publishing Co. Amsterdam, The Netherlands: Elsevier Science Publishers B.V.; 1989

[185] Golub G, Kahan V, Calculating the singular values and pseudoinverse of a matrix, SIAM J. Numer. Anal. 2, pp 205-224, 1965.

[186] Harshman RA, Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis, UCLA Working Papers in Phonetics, 1970, 16:1-84

[187] Leibovici D, Sabatier R, A Singular Value Decomposition of k-Way Array for a Principal Component Analysis of Multiway Data, Linear Algebra and Its Applications, 269:307-329, 1998.

[188] Benjamini Y, Hochberg Y, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society, 1995, Series B 57(1):289-300.

[189] Bonferroni CE, Teoria statistica delle classi e calcolo delle probabilit, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[190] Dunn OJ, Estimation of the Medians for Dependent Variables, Annals of Mathematical Statistics 30(1):192-197.

[191] Smyth GK, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Statistical Applications in Genetics and Molecular Biology, 2004, 3, No. 1, Article 3.

[192] Subramanian, Tamayo et al., GSEA software, and Molecular Signature Database (MSigDB) 2005, PNAS 102, 15545-15550

[193] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, Nat Genet. May 2000;25(1):25-9.

[194] Kanehisa M, Goto S, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28, 27-30 (2000).

[195] Gentleman R, Falcon S, Using GOstats to test gene lists for GO term association, Bioinformatics, 2007, 23(2):257-8.

[196] Zhang J, Gentleman R, KEGGSOAP: A package that provides a client interface to the KEGG SOAP server, 2015, R package version 2.11.

[197] Morgan M, Falcon S, Gentleman R, GSEABase: Gene set enrichment data structures and methods, 2015, R package version 1.30.2.

[198] Bauer S, Gagneur J, Robinson PN, GOing Bayesian: model-based gene set analysis of genome-scale data, Nucleic Acids Research, 2010, Vol. 38, No. 11 3523-3532.

[199] Alexa A, Rahnenführer J, Lengauer T, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, Bioinformatics, 2006, 22, 1600-1607.

[200] Lu Y, Rosenfeld R, Simon I, Nau GJ, Bar-Joseph Z, A probabilistic generative model for GO enrichment analysis. Nucleic Acids Res, 2008, 36, e109.

[201] Whittaker J, Graphical Models in Applied Multivariate Statistics, New York: Wiley; 1990.

[202] Ideker T et al., Discovering regulatory signaling circuits in molecular interaction networks, 2002, 18:S223-40.

[203] Smirnov N, Table for estimating the goodness of fit of empirical distributions, Annals of Mathematical Statistics 19:279-281.

[204] Lloyd Stuart P, Least squares quantization in PCM, IEEE Transactions on Information Theory, 1982, 28(2): 129-137

[205] Tibshirani R, Walther G, Hastie T, Estimating the number of clusters in a data set via the gap statistic, J R Statist Soc 2001, 63(2):411-23.

[206] R Core Team, A language and environment for statistical computing, 2015, R Foundation for Statistical Computing, Vienna, Austria.

[207] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik, cluster: Cluster Analysis Basics and Extensions, 2015, R package version 2.0.2.

[208] Peto R, Peto J, Asymptotically Efficient Rank Invariant Test Procedures, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 2. (1972), pp. 185-207.

[209] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ, Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data, BMC Syst Biol. 2011 Jan 31;5:21.

[210] Dimitriadou E, Hornik K, Leisch, et al., e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2010.

[211] Lawson CL, Hanson RJ, Solving Least Squares Problems, 1995, SIAM.

[212] Schäfer J, Strimmer K, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, 2005, Statist. Appl. Genet. Mol. Biol. 4:32.

[213] Beversdorf LM, Tests for correlation on bivariate non-normal distributions, 2008, UNF Theses and Dissertations, 284.

[214] Fisher RA, Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, Biometrika, 1915, 10(4):507-521.

[215] Langfelder P, Mischel PS, Horvath S, When Is Hub Gene Selection Better than Standard Meta-Analysis? PLoS One. 2013 Apr 17;8(4).

[216] Barabási AL, Oltvai ZN, Network biology: understanding the cell's functional organization, Nature Reviews Genetics, 2004, 5:101-113.

[217] Chung F, Linyuan L, Connected Components in Random Graphs with Given Expected Degree Sequences, Annals of Combinatorics, 2002, 6(2):125-145.

[218] Molero X, Vaquero EC, Flandez M, et al., Gene expression dynamics after murine pancreatitis unveils novel roles for Hnf1alpha in acinar cell homeostasis, Gut 2012, 61(8):1187-1196.

[219] Morris JPt, Cano DA, Sekine S, Wang SC, Hebrok M, Beta-catenin blocks Kras-dependent reprogramming of acini into pancreatic cancer precursor lesions in mice, J Clin Invest. Feb 2010;120(2):508-520.

[220] Aichler M, Seiler C, Tost M, et al., Origin of pancreatic ductal adenocarcinoma from atypical flat lesions: a comparative study in transgenic mice and human tissues, J Pathol, Apr 2012;226(5):723-734.

[221] Direnzo D, Hess DA, Damsz B, et al., Induced Mist1 expression promotes remodeling of mouse pancreatic acinar cells, Gastroenterology, Aug 2012;143(2):469-480.

[222] Lee AH, Chu GC, Iwakoshi NN, Glimcher LH, XBP-1 is required for biogenesis of cellular secretory machinery of exocrine glands, Embo J. Dec 21 2005;24(24):4368-4380.

[223] Figura G, Morris JPt, Wright CV, Hebrok M. Nr5a2 maintains acinar cell differentiation and constrains oncogenic Kras-mediated pancreatic neoplastic initiation, Gut, Apr 2014;63(4):656-664.

[224] Martinelli P, Canamero M, del Pozo N, Madriles F, Zapata A, Real FX, Gata6 is required for complete acinar differentiation and maintenance of the exocrine pancreas in adult mice, Gut, Oct 2013;62(10):1481-1488.

[225] Kong B, Michalski CW, Erkan M, Friess H, Kleeff J, From tissue turnover to the cell of origin for pancreatic cancer, Nat Rev Gastroenterol Hepatol, Aug 2011;8(8):467-472.

[226] Seymour PA, Freude KK, Tran MN, et al., SOX9 is required for maintenance of the pancreatic progenitor cell pool, Proc Natl Acad Sci U S A, Feb 6 2007;104(6):1865-1870.

[227] Furuyama K, Kawaguchi Y, Akiyama H, et al, Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine, Nat Genet. Jan 2011;43(1):34-41.

[228] Jensen JN, Cameron E, Garay MV, Starkey TW, Gianani R, Jensen J, Recapitulation of elements of embryonic development in adult mouse pancreatic regeneration, Gastroenterology, Mar 2005;128(3):728-741.

[229] Arda HE, Benitez CM, Kim SK, Gene regulatory networks governing pancreas development, Dev Cell, Apr 15 2013;25(1):5-13.

[230] Jensen J, Pedersen EE, Galante P, et al., Control of endodermal endocrine development by Hes-1, Nat Genet, Jan 2000;24(1):36-44.

[231] Hermann PC, Sancho P, Canamero M, et al., Nicotine Promotes Initiation and Progression of KRAS-Induced Pancreatic Cancer via Gata6-Dependent Dedifferentiation of Acinar Cells in Mice, Gastroenterology, Nov 2014;147(5):1119-1133 e1114.

[232] Omary MB, Lugea A, Lowe AW, Pandol SJ, The pancreatic stellate cell: a star on the rise in pancreatic diseases, J Clin Invest, Jan 2007;117(1):50-59.

[233] Landsman L, Nijagal A, Whitchurch TJ, et al., Pancreatic mesenchyme regulates epithelial organogenesis throughout development, PLoS Biol, Sep 2011;9(9):e1001143.

[234] Sherman MH, Yu RT, Engle DD, et al., Vitamin d receptor-mediated stromal reprogramming suppresses pancreatitis and enhances pancreatic cancer therapy, Cell, Sep 25 2014;159(1):80-93.

[235] Erkan M, Adler G, Apte MV, et al., StellaTUM: current consensus and discussion on pancreatic stellate cell research, Gut, Feb 2012;61(2):172-178.

[236] Sanford DE, Belt BA, Panni RZ, et al., Inflammatory monocyte mobilization decreases patient survival in pancreatic cancer: a role for targeting the CCL2/CCR2 axis, Clin Cancer Res, Jul 1 2013;19(13):3404-3415.

[237] Lesina M, Kurkowski MU, Ludes K, et al., Stat3/Socs3 activation by IL-6 transsignaling promotes progression of pancreatic intraepithelial neoplasia and development of pancreatic cancer, Cancer Cell, Apr 12 2011;19(4):456-469.

[238] Burkly LC, Michaelson JS, Hahm K, Jakubowski A, Zheng TS, TWEAKing tissue remodeling by a multifunctional cytokine: role of TWEAK/Fn14 pathway in health and disease, Cytokine, Oct 2007;40(1):1-16.

[239] Maecker H, Varfolomeev E, Kischkel F, et al., TWEAK attenuates the transition from innate to adaptive immunity, Cell, Dec 2 2005;123(5):931-944.

[240] Zhang G, Schetter A, He P, et al., DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma, PLoS One, 2012;7(2):e31507.

[241] Kuo TC, Tan CT, Chang YW, Hong CC, Lee WJ, Chen MW, Jeng YM, Chiou J, Yu P, Chen PS, Wang MY, Hsiao M, Su JL, Kuo ML, Angiopoietin-like protein 1 suppresses SLUG to inhibit cancer cell motility, J Clin Invest. 2013 Mar;123(3):1082-95.

[242] Clausen KA, Blish KR, Birse CE, Triplette MA, Kute TE, Russell GB, D'Agostino RB Jr, Miller LD, Torti FM, Torti SV, SOSTDC1 differentially modulates Smad and beta-catenin activation and is downregulated in breast cancer, Breast Cancer Res Treat. 2011 Oct;129(3):737-46.

[243] Thawani JP, Wang AC, Than KD, Lin CY, La Marca F, Park P, Bone morphogenetic proteins and cancer: review of the literature, Neurosurgery. 2010 Feb;66(2):233-46.

[244] Wente MN, Mayer C, Gaida MM, Michalski CW, Giese T, Bergmann F, Giese NA, Büchler MW, Friess H, CXCL14 expression and potential function in pancreatic cancer, Cancer Lett. 2008 Feb 8;259(2):209-17.

[245] Hempen PM, Kurpad H, Calhoun ES, Abraham S, Kern SE, A double missense variation of the BUB1 gene and a defective mitotic spindle checkpoint in the pancreatic cancer cell line Hs766T, Hum Mutat. 2003 Apr;21(4):445.

[246] Chou RH et al., Suppression of the invasion and migration of cancer cells by SERPINB family genes and their derived peptides, Oncology Reports, Spandidos Publications, 2012, 27(1):238-245.

[247] Dvorak HF, Tumors: wounds that do not heal, Similarities between tumor stroma generation and wound healing, N Engl J Med. Dec 25 1986;315(26):1650-1659.

[248] Gittes GK, Developmental biology of the pancreas: a comprehensive review, Dev Biol, Feb 1 2009;326(1):4-35.

[249] Itzkovitz S, Blat IC, Jacks T, Clevers H, van Oudenaarden A, Optimality in the development of intestinal crypts, Cell, Feb 3 2012;148(3):608-619.

[250] Murtaugh LC, Keefe MD, Regeneration and Repair of the Exocrine Pancreas, Annu Rev Physiol. Oct 24 2014.

[251] Ramachandran V, Arumugam T, Wang H, Logsdon CD, Anterior gradient 2 is expressed and secreted during the development of pancreatic cancer and promotes cancer cell survival, Cancer Res. 2008 Oct 1;68(19):7811-8.

[252] Pohler E, Craig AL, Cotton J, Lawrie L, Dillon JF, Ross P, Kernohan N, Hupp TR, The Barrett's antigen anterior gradient-2 silences the p53 transcriptional response to DNA damage, Mol Cell Proteomics. 2004 Jun;3(6):534-47.

[253] Yue S, Mu W, Zöller M, Tspan8 and CD151 promote metastasis by distinct mechanisms, Eur J Cancer. 2013 Sep;49(13):2934-48.

[254] Uhlen M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F, Human Protein Atlas, Tissue-based map of the human proteome, Science 2015.

[255] Lafitte M, Moranvillier I, Garcia S, Peuchant E, Iovanna J, Rousseau B, Dubus P, Guyonnet-Duprat V, Belleanne G, Ramos J, Bedel A, de Verneuil H, Moreau-Gaudry F, Dabernat S, FGFR3 has tumor suppressor properties in cells with epithelial phenotype, Mol Cancer 2013, Jul 31;12:83.

[256] Pinho AV, Rooman I, Reichert M, De Medts N, Bouwens L, Rustgi AK, Real FX, Adult pancreatic acinar cells dedifferentiate to an embryonic progenitor phenotype with concomitant activation of a senescence programme that is present in chronic pancreatitis, Gut. 2011 Jul;60(7):958-66.

[257] Shi G, Zhu L, Sun Y, Bettencourt R, Damsz B, Hruban RH, Konieczny SF, Loss of the acinar-restricted transcription factor Mist1 accelerates Kras-induced pancreatic intraepithelial neoplasia, Gastroenterology, 2009 Apr;136(4):1368-78.

[258] **Kong B*, Bruns P*, Behler N*, Schlitter AM, Fritzsche S, Valkovska N, Jian Z, Regel I, Raulefs S, Irmler M, Beckers J, Friess H, Erkan M, Mueller NS, Esposito I, Theis FJ, Kleeff J, Michalski CW, Histological and molecular landscape of early pancreatic carcinogenesis, *submitted to Nat Commun, manuscript ID NCOMMS-15-07978*, * shared first authorship.**

Index

α-Amy, 30, 69
α-SMA, 4, 19, 31, 69, 78, 91
β-catenin, 18, 21
γ-globulin, 10
γ-secretase, 18

ACC, 22
acinar cell carcinoma, *see* ACC
acinar-to-ductal metaplasia, *see* ADM
acini, 1
actin, 20
ADM, 2, **11**, 11, 55, 58, 89
AFL, 58
Agr2, 72
amylin, 3
analysis of variance, *see* ANOVA
ANOVA, 39
antibodies, 31
AP-1, 10, 27
Arx, 8
ATM, 14
atypical flat lesions, *see* AFL

Bioconductor, 34
Bmp, 5
BRCA, 14
BrdU, 30, 31, 60, 72

caerulein, **11**, 11, 29
CANDECOMP, 37
Cathepsin B, 10
Ccl2, 79
Ccl7, 79
CD40L, 10
Cd45, 30, 31, 60
CDK1/2, 15
CDK4/6, 15
cells
    α-, 4
    β-, 3–5
    δ-, 4, 5
    ϵ-, 4
    γ-, *see* PP-cells
    acinar, **1**, 1, 8, 10, 11, 69
    centroacinar, 3
    ductal, **2**, 6, 11
    endocrine, **3**
    endothelial, 10, 12
    exocrine, **1**
    fibroblast, 12
    immune, 30, 58, 61
    inflammatory, 12
    mesenchymal, 4
    pancreatic stellate, *see* PSC
    PP-, 4
    progenitor, 5, 6, 11, 30, 72, 76, 90
    proliferating, 30
    stroma, 4
    trunk, 7
CFTR, 9, 14
chemokines, 4, 10, 26, 78, 91
chymotrypsin C, 9, 10
clustering
    hierarchical, 35, 56, 57, 60
    k-means, *see* k-means
Col12a1, 78
Col8a1, 78
collagens, 20
correlation
    pearson, 52

correlation networks, 52
cross-validation, 45, 80
Cuzd1, 72
cytokeratines, 12
cytokines, 4, 10, 26, 78, 91

DAPi, 31
decomposition, 36
dendrogram, 35
Derl3, 72
desmoplasia, 19
desmosomes, 20
development, **4**
differential expression analysis, 38, 79
    with unknown groups, 41

E-cadherin, 20, 22
E2F1, 15
E47, 20
early pancreatic carcinogenesis, **14**
ECM, 4, 10, 19, 78, 91
edema, 55
EGF, 19, 21
EGFR, 11, 92
EMT, 4, 15, **20**, 91
enrichment, 43, 44
enzymes
    digestive, 1, 22
epidemiology, 13
epithelial-mesenchymal transition, *see* EMT
ERK, 14, 26
experimental design, 29
extracellular matrix, *see* ECM

factorization, 36
false discovery rate, *see* FDR
FAMMM, 14
FDR, 41
FGF, 5, 19, 21, 94
FGFR, *see* FGF
fibrosis, 10
Folforinox, 13
Foxa1, 6
Foxa2, 6

Gal, 72
Gata4, 6
Gata6, 6, 76
GEMM, 29, 92
gene ontology, *see* GO
gene signature, 45
GeneChip, 32
ghrelin, 4
glucagon, 1, 4
glucose, 1
GO, 44

H&E, 30, 56
Hedgehog, *see* Hh
hematoxylin and eosin, *see* H&E
Hes1, 6, 7, 18, 20, 27, 76, 90
Hey1, 27
HGF, 19, 21
Hh, 11, 13, 18, 19, 26
Hhex, 6
Hlxb9, 6
Hnf1b, 6, 8, 76, 90
Hnf3a, *see* Foxa1
Hnf3b, *see* Foxa2
Hnf6, 6, 8
homeostasis, 5, 11, 13

ICAM-1, 10, 11
IF, 30
IgG, 31
IgG4, 10
IHC, 30
Ikk2, 26
Il-1α, 27
Il6, 79
immune response, **10**
immunofluorescence, *see* IF
immunohisochemistry, *see* IHC
inflammation, 4, 8, 13, 26, 55

inflammatory response, 55
insulin, 1, 3
interleukines, 10
intraductal papillary mucinous neoplasm, *see* IPMN
IPMN, 22
Irx, 8
Isl1, 8
islets of Langerhans, 3

Jak, 11
junctions
    adherens, 20
    tight, 20

k-means, 48
Kaplan-Meier curves, 49, 80
KEGG, 44
KLF8, 20
KRAS, 14, 16, **25**
Kras, 18, **25**, 29

Langerhans islets, *see* islets of Langerhans
lesions
    precursor, 14
linear model, 38

MafA, 8
MafB, 8
MAPK, 10
matrix metalloproteinases, *see* MMP
MCN, 22
MCP-1, 10
MDM2, 15
MEFs, 19
Mek, 19
metastases, 12
mgsa, 44
microarray, **32**
microRNA, 22
Mist1, 2, 11, 72, 89, 94
MMP, 4, 10, 91
model, 52
Monte Carlo, 48
morphogenesis, **4**
mouse embryonic fibroblasts, *see* MEFs
mouse models, 11, 92
mTOR, 18
mTORC1, 21
mucinous cystic neoplasm, *see* MCN
mucins, 12, 15, 16, 22
Myc, 6

N-cadherin, 21
necro-inflammation, 10
necrosis, 8, 18
Neurod, 8
Neurog3, *see* Ngn3, 76, 90
NF-κB, 10, 11, 21, 26
Ngn3, 5, 7, 8
Nkx, 8
Nkx6.1, 7, 76
Nkx6.2, 7
normalization, 34
Notch, 11, 13, 18, 19, 21, 26, 76, 90, 91
NR5A2, 14
Numb, 11

oligo, 33
Onecut1, *see* Hnf6, 76, 90
overrepresentation, 43

p110α, 27
p16, 14, 16
p48, *see* Ptf1a
p53, 14–16, 18, 26
p62, 27
PALB2, 14
pancreas, **1**
pancreatic ductal adenocarcinoma, *see* PDAC
pancreatic neuroendocrine tumor, *see* PNET
pancreatic polypeptide, 4
pancreatitis, **8**, 29, 55
    acute, **8**, 11

autoimmune, 10
chronic, **9**, 12
hereditary, 9
tropical, 9
pancreatoblastoma, 23
PanIN, **16**, 25, 58
PARAFAC, 37, 61
pathology, **12**
Pax, 8
PCA, 36, 57, 58
PDAC, 2, 11, **12**, 12, 13, 18
PDGF, 19
Pdx1, 2, 5, 6, 11, 18
petri net, 52
PI3K, 10, 14, 19, 21, 26, 27
PNET, 23
preprocessing, 33
primary transition, 6
Principal component analysis, *see* PCA
probe, 33
protocol
RNA isolation, 30
treatment, 29
Prox1, 7
PRSS1, 9, 14
PRSS2, 9
PSC, **4**, 10, 76, 91
Pten, 18
Ptf1a, 5–7, 18, 76, 90

q-value, 41
quality control, 33

RA, 5
RAF, 14, 26
Rbpj, 6, 76
Rbpjl, 72, 76, 89, 90
reactive oxygen species, *see* ROS
refinement, 55
regeneration, 55
regulatory networks, 46
retinoic acid, *see* RA
RhoA, 21
RMA, 34
robustness, 45
ROS, 9, 11

S100P, 15
secondary transition, 6
Serpine1, 78
signature, 80
Simulated Annealing, 48
Singular Value Decomposition, *see* SVD
Slc39a5, 72
Slug, 20
SMAD, 21
SMAD4, 14–16
Smo, 19
Snail, 20
solid pseudopapillary neoplasm, *see* SPN
somatostatin, 4
Sox17, 6
Sox9, 6, 30, 31, 69, 72, 76, 89
SPARC, 19
SPINK1, 9, 10
SPN, 23
STAT3, 10
stroma, 12, 20
subcircuits, 46
support vector machine, *see* SVM
survival analysis, 49, 80
survival prognosis, 13
SVD, 36
SVM, 50, 84

TCF, 21
TCF2, *see* Hnf1b
telomere, 16
test
correlation, 53
F-, 39
hypergeometric, 43
Kolmogorov-Smirnov-, 42
log-rank, 49

Peto-Peto-, 49
Shapiro-Wilk-, 42
t-, 39
TGFβ, 13, 15, 19–21
Timp1, 78
Tmed11, 72
TNFα, 10
Tnfrsf12a, *see* TWEAK
trypsin, 9
trypsinogen, 8
Ttc9, 72
tubular complexes, 55
tumor, 12, 22
tumor environment, 19
tumor-stroma interactions, 10, **19**
TWEAK, 79

VEGF/VEGFR, 13
vHNF, *see* Hnf1b
vimentin, 20
vitamin D, 13

Wnt, 5, 18, 21

z-score, 41, 80
aggregated, 41
ZEB, 20, 22
zymogen granules, 2, 8