Computational Approaches to Delineate Transcriptional and Functional Heterogeneity in Pancreatic Cancer

Sanjana Srinivasan

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations

Part of the Bioinformatics Commons, Computational Biology Commons, and the Genomics Commons

Recommended Citation
Srinivasan, Sanjana, "Computational Approaches to Delineate Transcriptional and Functional Heterogeneity in Pancreatic Cancer" (2021). The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access). 1070. https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1070

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.
COMPUTATIONAL APPROACHES TO DELINEATE TRANSCRIPTIONAL AND
FUNCTIONAL HETEROGENEITY IN PANCREATIC CANCER

by

Sanjana Srinivasan, B.Sc., M.P.H.

APPROVED:

__________________________________
Giulio F. Draetta, M.D., Ph.D.
Advisory Professor

__________________________________
Swathi Arur, Ph.D.

__________________________________
Traver Hart, Ph.D.

__________________________________
Anirban Maitra, M.B.B.S.

__________________________________
Paul Scheet, Ph.D.

__________________________________
Andrea Viale, M.D.

APPROVED:

__________________________________
Dean, The University of Texas
MD Anderson Cancer Center UTHHealth Graduate School of Biomedical Sciences
COMPUTATIONAL APPROACHES TO DELINEATE TRANSCRIPTIONAL AND FUNCTIONAL HETEROGENEITY IN PANCREATIC CANCER

A DISSERTATION

Presented to the Faculty of
The University of Texas
MD Anderson Cancer Center UTHealth
Graduate School of Biomedical Sciences
In Partial Fulfillment
For the Degree of
DOCTOR OF PHILOSOPHY

by

Sanjana Srinivasan, B.Sc., M.P.H.
Houston, Texas
May, 2021
Dedication

This dissertation is dedicated to my parents, Sujatha and Ravi Srinivasan, and my beloved grandparents, Komala and S. Challappa. Everything that I am, and everything that I ever will be, I owe to you.
Acknowledgements

Getting to this stage in my life – earning a doctoral degree, has been a lifelong goal. It is a culmination of several years of work, none of which would have been possible without a long list of mentors, colleagues, friends and family that I am eternally grateful for.

First, I would like to thank my advisor, Dr. Giulio Draetta for taking a chance on me, supporting me and investing in my success. You have been an inspiration, both in and out of lab. Thank you for creating an environment of innovation, learning and collaboration, I could not have asked for a better environment to grow as a scientist. To my co-advisor, Dr. Traver Hart, thank you for support, generosity and openness, and for your continued effort in exposing your students to the most cutting-edge technologies. I would like to thank Alessandro Carugo and Chris Bristow for teaching me everything I know. Thank you for your patience, your mentorship, your guidance and your willingness to share your time and knowledge with me.

To Dr. Johnathon Rose, I don’t know where to begin. I would not be here without your generosity and wisdom, constantly helping me navigate my way through science, grad school and life. Thank you for being a mentor, role model, cheerleader, drill sergeant, Sherpa, cat transporter, bodyguard, and everything in between. Thank you for every late night, every early morning, the countless cups of coffee and burgers. Most of all, thank you for being my best friend.

I would like to offer my sincere gratitude to my advisory committee – Dr. Swathi Arur, Dr. Anirban Maitra, Dr. Paul Scheet, Dr. Andrea Viale, and Dr. Scott Kopetz. Your guidance and inputs through the years has taught me to focus on the big picture and
direct my perspective on clinical implications and patient care. Thank you to the Dr. Kari Brewer-Savannah, Dr. Khandan Keyomarsi and Dr. Stephanie Watowich, and CPRIT training program for your support and mentorship over these past two years. It has been an honor to learn from you.

Thank you to my labmates and friends, Jintan Liu, Alex Li, Melinda, I Lin Ho, Nick Yen, Daniel Zamler, and the rest of the Draetta, Viale and Genovese Labs for the shared laughs and tears alike, and for setting the bar so high with your brilliance, skill and hard work. Thank you to the many friends and colleagues in the Hart lab – Megan McLaughlin, Eiru Kim, Merve Dede and Medina Colic for the many lunches, and for your support through my time in grad school.

I am forever grateful for the wonderful friendships and bonds that I have formed through graduate school. A special thank you to two thirds of the Three Musketeers - Frank Lenoir and Darlan Minussi. In addition to your friendship through the trials and tribulations of coursework and research, I am forever grateful to both of you for patiently taking the time to guide me through the scary world of quantitative sciences. Every cup of tea and fro-yo has meant more to me than I can ever express. Thank you to my dear friends, Vinay Nair, Anupallavi Srinivasamani, and Hannah Savage for putting up with me. You are brilliant scientists, and even better human beings. You make bad days seem better and good days feel great. Your friendship has been a highlight of my graduate experience. The laughs, memes and meals we have shared are memories I will treasure forever. Don’t forget me when you’re famous!

To my brilliant, beautiful and industrious cousin, Kritika Murli, thank you. You might be younger, but you are infinitely wiser. I owe my sanity to you and the rest of my Thunder
Indians – Pavithra Chidambaram, Maya Iyer and Shivani Mukhi. Thank you for always being there for me, and for laughing at my very lame jokes, and for never letting me feel alone. The four of you have been a constant source of joy and awe. Being able to grow with you during my time in Houston has been an honor and a privilege.

I would like to thank the GSBS and everyone at the grad school for creating such a pleasant and friendly environment for students. A sincere thank you to Ms. Amy Carter and Ms. Elisabeth Lindheim for going the extra mile to support us. Thank you to Dr. Bill Mattox and Brenda Gaughan for steering me in the right direction and for your advice whenever I lost my way. I would also specifically like to thank Dr. Prahlad Ram and the QS program for welcoming students from all backgrounds and for dispelling any fears I had joining this program as a novice.

Thank you to my family, without whom, none of this would be possible. To my uncle and aunt, Murli Challappa and Sujatha Murli, for being my second parents and making sure both my heart and belly are always full. To my grandfather, S. Challappa, for being the beacon of hard work and integrity throughout my life, and my grandmother Komala Challappa – your strength and resilience have been an inspiration to me, and I am forever grateful for your love. To my parents, thank you for your unconditional love and support. To my father, Ravi Srinivasan, your brilliance is exceeded only by your kindness. Thank you for your sage advice, and for moving heaven and earth to make sure that we, as a family, are where we are now. Finally, thank you to my mother, Sujatha Srinivasan, the smartest and bravest woman I know, for being my role model and best friend. I love you, and cherish every moment I have shared with you. If I grow into even half the woman that you are, I will consider myself a success.
Abstract:

Pancreatic ductal adenocarcinoma (PDAC) is an incurable disease characterized by poor survival, dense desmoplastic stroma and activating mutations in KRAS (>90%). These tumors are highly complex ecosystems composed of molecularly distinct sub-populations that exhibit a spectrum of genetic features and associated phenotypes. Despite recent advances in the transcriptomic characterization of PDAC into at least two tumor subtypes, this alone has been insufficient to define more specific patterns of oncogenic dependency. To fully leverage advancements in next generation sequencing and functional genomics, we have sought to establish computational methodologies to aid in refined target discovery, and to develop a novel platform to comprehensively characterize the transcriptional heterogeneity of PDAC. Specifically, focusing on a large PDAC PDX cohort, we focused on a) establishing a PDAC co-expression network to serve as a foundation for quantifying disease diversity within the cohort, while in parallel b) optimizing an analytical approach to allow for in vivo CRISPR-Cas9 functional genomics using select models from the cohort. Applying and integrating this novel computational methodology, we integrated CRISPR-based co-dependency annotations with a disease-specific co-expression network developed from patient-derived models to establish a framework to quantitatively associate gene-cluster patterns with genetic
vulnerabilities. We defined multiple prominent anti-correlating gene-cluster signatures and pathway-specific dependencies, both across genetically distinct PDAC models and intratumorally at the single-cell level. This characterization of intratumoral cluster representation was accomplished through a novel adaptation of network signatures for single-cell analysis. Of these network-defined cluster trends, one differential signature recapitulated the characteristics of classical and basal-like PDAC molecular subtypes on a continuous scale, which we validated using direct capture Perturb-seq. Our results demonstrate the utility of this integrated platform as a quantitative approach for characterizing specific genetic dependencies within defined molecular contexts represented in PDAC, with the potential to guide future clinical positioning for targeted therapeutics while also considering a constantly evolving intratumoral heterogeneity.
# Table of Contents

**Approval Page**........................................................................................................i

**Title Page**..................................................................................................................ii

**Dedication** .................................................................................................................. iii

**Acknowledgements** .................................................................................................. iv

**Abstract** ..................................................................................................................... vii

**Table of Contents** ...................................................................................................... ix

**List of Figures** ........................................................................................................ xii

**List of Tables** ........................................................................................................... xvi

**Chapter 1: Introduction** ......................................................................................... 1

1.1: PDAC Epidemiology............................................................................................... 1

1.2: PDAC Genetics and Biology ............................................................................... 2

1.3: PDAC Management and Treatment .................................................................... 4

1.4: PDAC Subtypes .................................................................................................... 7

  1.4.1: Collisson Classification ................................................................................... 7

  1.4.2: Bailey Classification ....................................................................................... 8

  1.4.3: Moffitt Classification ...................................................................................... 8

  1.4.4: Integrated Analysis of Classifications by the TCGA ....................................... 9

  1.4.5: Puleo Classification ....................................................................................... 10

  1.4.6: COMPASS Trial: Toronto Classification ....................................................... 10

  1.4.7: Current Consensus in PDAC Subtyping ....................................................... 11

1.5: Dissertation Objective .......................................................................................... 11

**Chapter 2: Materials and Methods** ...................................................................... 14

**Chapter 3: Novel Methodology for Target Library CRISPR Screening** ............... 23
3.1: CRISPR and its Uses in Functional Genomics ................................................................. 23

3.2: Analytical Tools in Genome-Wide CRISPR Screening .................................................... 26
   3.2.1: Redundant siRNA Activity (RSA) ........................................................................... 26
   3.2.2: MAGeCK .................................................................................................................. 27
   3.2.3: CERES ....................................................................................................................... 28
   3.2.4: JACKS ...................................................................................................................... 29
   3.2.5: BAGEL ..................................................................................................................... 29

3.3: Introduction of Surfaceome Library and Logic ................................................................. 30

3.4: Analysis of Custom Library Screens .............................................................................. 32

3.5: Surfaceome Library Analysis using BAGEL v1 .............................................................. 32

3.6: Low-Fat BAGEL in Surfaceome Screens ....................................................................... 42

3.7: Discussion ....................................................................................................................... 54

Chapter 4: Quantifying PDAC Diversity with Biological Networks ........................................ 55

4.1: Patient Derived Xenograft Models ................................................................................ 55

4.2: Utility of Co-expression Analysis in Large Datasets ..................................................... 62

4.3: Construction of the PDAC Co-Expression Network .................................................... 63

4.4: Clustering and GO Annotation of the PDAC Co-Expression Network ......................... 65

4.5: Associated Mutations with Cluster Enrichment ............................................................. 72

4.6: Prominent Anticorrelated Cluster Signatures ............................................................... 74

4.7: Cluster 1 and 23 Anticorrelation Recapitulates Moffitt Signatures ............................... 77

4.8: Anticorrelated Signatures Recapitulated in TCGA Samples ........................................ 84

4.9 Discussion ....................................................................................................................... 86

Chapter 5: Integration of the Co-Expression Network with Single Cell Transcriptomic
Signatures ............................................................................................................................... 87

5.1: Interrogating the Quasi-Basal Signature Intratumorally ............................................... 87
5.2: Single Cell Samples Utilized ........................................................................................................ 88
5.3: Cluster Signatures for Single Cells ............................................................................................... 89
5.4: Single-cell Transcriptomic Profiles Recapitulate Granularity of PDAC Subtypes ..... 89
5.5: Application of the Co-Expression Network in Single-cell Transcriptomic Profiles from Clinical Samples ......................................................................................................................... 100
5.6: Discussion ....................................................................................................................................... 111

Chapter 6: Intratumoral Validation of Functional Vulnerabilities with Co-Expression Network
.......................................................................................................................................................... 112
6.1: Integration of CRISPR Identified Dependencies with the Co-Expression Network 112
6.2: Analysis of the Feature Barcoded scRNAseq Validates Subtype Specific Functional Vulnerabilities .......................................................................................................................................... 120
6.3: Discussion ....................................................................................................................................... 132

Chapter 7: Discussion and Future Directions ....................................................................................... 134
Low-Fat BAGEL and in vivo Genetic Screens ....................................................................................... 135
PDAC Co-Expression Network Development and Applications ......................................................... 136
Linking Genetic Dependency with Network-based Transcriptomic Signatures ................................. 139
Conclusions .......................................................................................................................................... 141

References ........................................................................................................................................ 143

Vita ..................................................................................................................................................... 167
List of Figures

Figure 1: Distribution of the essential and non-essential gene set of the surfaceome library in PATC69, Day 12. ..................................................................................................................37

Figure 2: Bayes Factors of the core essential and non-essential genes in the PATC69 Day 12 screen derived from BAGEL v1. ..................................................................................................................38

Figure 3: Bayes Factors of the core essential and non-essential genes in the PATC69 Day 12 screen derived from Low-Fat BAGEL. ..................................................................................................................39

Figure 4: Precision-Recall curves comparing the performance of BAGEL v1 and Low-Fat BAGEL. ..................................................................................................................40

Figure 5: Comparison of F measures of known genome-wide CRISPR screen analytical methods to Low-Fat BAGEL across PATC69 and PATC124 in vitro and in vivo screens. ..................................................................................................................41

Figure 6: Paired density plots and Bayes Factor estimates for PATC69. ..................45

Figure 7: Paired density plots and Bayes Factor estimates for PATC124. ............47

Figure 8: Paired density plots and Bayes Factor estimates for PATC53. .............49

Figure 9: Comparing essential genes across models in vitro (top) and in vivo (bottom). ..................................................................................................................50

Figure 10: CRISPR screening results from PATC69. ........................................51

Figure 11: CRISPR screening results from PATC124. .......................................52

Figure 12: CRISPR screening results from PATC53. .......................................53

Figure 13: Mutational diversity of the PDAC PDX models included in this study. ....60

Figure 14: Moffitt Classification of PDAC PDX models. ....................................61
Figure 15: Network hairball of the PDAC Co-expression network. ........................................68
Figure 16: Treemaps of Cluster by Cluster GO Annotations within the PDAC Co-expression Network. ..........................................................................................................................69
Figure 17: Cluster centroids derived from the PDAC co-expression network and the associated correlations between clusters. ........................................................................................................71
Figure 18: Anticorrelations of clusters in the PDAC co-expression network with the corresponding GO annotations. ..........................................................................................................................76
Figure 19: Module centroid scores of Cluster 1 between Moffitt defined classical and basal models. .................................................................................................................................79
Figure 20: K means clustering reveals optimal k=3 using the Cluster 1 and Cluster 23 centroid scores across the PDX models. ..................................................................................................................80
Figure 21: Prominent anti-correlated clusters recapitulate a granular spectrum of “classical”, “quasi-basal”, and “basal-like” models that are associated with clinical features........................................................................................................................................82
Figure 22: GSEA conducted on PDAC co-expression network-defined classical and basal-like models identifies EMT as a top pathway enriched in basal-like models FDR = 0.000.................................................................................................................................83
Figure 23: Prominent anti-correlated clusters identifying “classical”, “quasi-basal”, and “basal-like” recapitulated in TCGA................................................................................................................85
Figure 24: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC69 single cells. ..................................................................................................................93
Figure 25: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC124 single cells. ..................................................................................................................94
Figure 26: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC53 single cells. .................................................................95
Figure 27: Cluster 1 and 23 differential in PATC69 reveals quasi-basal single cells. ....96
Figure 28: Cluster 1 and 23 differential in PATC124 reveals quasi-basal single cells. ..97
Figure 29: Cluster 1 and 23 differential in PATC53 reveals quasi-basal single cells. ....98
Figure 30: Difference in the distribution of cluster 1 and cluster 23 enrichment across three PDAC PDX models........................................................................99
Figure 31: Diagram of the locations and numbers of isolated patient core needle biopsy (CNB) samples used for single-cell RNA sequencing analysis.................................102
Figure 32: UMAP outlining the tumor microenvironment components of multiple single-cell sequenced primary (n = 4) and metastatic (n = 3) CNB samples from PDAC patients..............................................................................................................................104
Figure 33: Expression of known marker genes was used to identify cell type of each UMAP-identified cluster .................................................................................................................................107
Figure 34: Representation of the PDAC co-expression network genes across cell types in PDAC patient core needle biopsy samples (n=7). .................................................................109
Figure 35: Combined CNB UMAP with an overlay of a PDAC co-expression differential signature of cluster 1 and cluster 23 expression..............................................................110
Figure 36: Co-expression network anchoring of in vivo functional dependencies on a model by model basis..............................................................................................................................117
Figure 37: Overview of the feature barcoding strategy for intratumoral tracking of PDAC co-expression network signatures following sgRNA knockouts............................................119
Figure 38: Comparison of ILK, SMAD4, and ZEB1 knock-outs to negative control ABCG8 conditions in PATC69. ................................................................. 123

Figure 39: Comparison of ILK, SMAD4, and ZEB1 knock-outs to negative control ABCG8 conditions in PATC53. ................................................................. 125

Figure 40: UMAPs detailing the cluster 1 and cluster 23 signature in each single-cell RNA sequenced for PATC69 (top) and PATC53 (bottom). .............................................. 127

Figure 41: Comparative PATC69 density plots of cluster 1 and cluster 23 signature following knockout with sgILK, sgSMAD4, sgZEB1 and sgABCG8. ......................... 129

Figure 42: Comparative PATC53 density plots of cluster 1 and cluster 23 signature following knockout with sgILK, sgSMAD4, sgZEB1 and sgABCG8. ......................... 131
List of Tables

Table 1: Clinical features of patients corresponding to the PDAC PDX models used in this study ...........................................................................................................................................58

Table 2: Results from the UNCOVER analysis reveal complementary gene mutations associated with cluster centroid-based enrichment ..............................................................................73
Chapter 1: Introduction

Pancreatic Ductal Adenocarcinoma (PDAC), the most common form of pancreatic cancer, is an aggressive malignancy typically diagnosed at an advanced stage and resistant to most forms of treatment\(^1\). PDAC is the fourth most common cause of cancer related death in the United States\(^2\) and has an extremely poor prognosis - frequently detected as advanced non-localized tumors with a five-year survival rate of \(\leq 10\%\)\(^1\). The main factor influencing the survival outcome is the stage of tumor at diagnosis. Only 10-20\% of patients present with surgically resectable tumors at the time of diagnosis, while most patients present with locally advanced disease, unresectable tumors or metastases\(^3\).

1.1: PDAC Epidemiology

Epidemiological studies have identified risk factors for pancreatic cancer including cigarette smoking, type 2 diabetes, high alcohol consumption, being overweight or obese, and a family history of pancreatic cancer\(^2\). A history of chronic pancreatitis is also found to increase the risk of developing pancreatic cancer later in life\(^4\). Additionally, 80-90\% of patients diagnosed with pancreatic cancer are over the age of 55, with the majority of patients being diagnosed in their 70s\(^5\), and a slightly higher incidence in men compared to women\(^5\). Within the United States, African Americans have a 50%-90\% higher risk of pancreatic cancer compared to Caucasian Americans, with Pacific Islanders and Asian Americans having the lowest risk\(^5\). The higher rate in African Americans is possibly due to higher prevalence of lifestyle factors listed above that increase the risk for pancreatic cancer, but gene by environment interactions have also been proposed to play a role\(^1\).
Inherited genetic syndromes can also increase risk for PDAC, including hereditary breast and ovarian cancer syndrome involving mutations in the *BRCA1*, *BRCA2* and/or *PALB2*, familial atypical multiple mole syndrome caused by mutations in *p16/CDKN2A*, and Lynch syndrome, which is also implicated in familial colorectal cancer incidence (as reviewed by McGuigan, et al.2018)².

**1:2: PDAC Genetics and Biology**

The pancreas consists of exocrine (acinar), epithelial (ductal) and acinar (α, β, δ, ε) cells. The acinar cells present with a high degree of plasticity and go through a process called acinar-to-ductal metaplasia (ADM) - whereby acinar cells transdifferentiate into a more ductal morphology and phenotype under conditions like inflammation or stress⁶,⁷. Acinar cells undergoing ADM become more prone to oncogenic events, defining the genetic progression model of PDAC carcinogenesis⁸. These events include activating mutations in *KRAS* on chromosome 12p, which transforms mutated tissue into pancreatic intraepithelial neoplasia (PanIN). These lesions tend to be largely asymptomatic in nature, with minimal disruption of pancreatic function in early stages. PanINs were previously classified into three grades: PanIN1A (flat lesion) and PanIN1B (micropapillary type) which display low grade dysplasia, PanIN2 which exhibits loss of polarity, nuclear crowding, cell enlargement and papillary formation, and finally PanIN3, which present as advanced lesions with nuclear atypia, luminal necrosis, and epithelial cell budding into the ductal lumen⁹. Currently, the PanIN classification is divided into low grade (PanIN1 and PanIN2) and high grade (PanIN3)¹⁰.

During PDAC pathogenesis, the progression into higher grade PanINs are accompanied with sequential loss of tumor suppressor genes. Following activating
mutations in KRAS, loss of function mutations in tumor suppressor TP53, followed by CDKN2A, and SMAD4, which make up the most frequent alterations/mutations in PDAC\textsuperscript{11}. In addition to these four prominent mutations, advancement in sequencing and genetic techniques have provided further insight into the genetic complexity of PDAC, with an average of 63 genetic aberrations in an individual tumor\textsuperscript{12}. Large scale genomic studies have also characterized a long tail of mutations, each with prevalence in less than 10\% of patients, including RNF43, ARID1A, TGF\textbeta{}R2, GNAS, RREB1, and PBRM. In KRAS wild type tumors, genes such as GNAS, BRAF, CTNNB1, which were found to be mutated\textsuperscript{12}. These mutations lead to dysregulations of key pathways and processes including activation of RAS-ERK signaling, loss of the G1/S checkpoint, NOTCH signaling, Hedgehog signaling, Wnt/\beta{}-catenin, axon guidance, and chromatin remodeling (as reviewed by Ying, et al)\textsuperscript{13}. While over 90\% of patients have KRAS mutations, additional amplifications accelerate the tumor promoting the potential for lesions even more. Apart from mutations, AKT2 overexpression and PI3K activity that is elevated in PDAC leads to increased tumor cell proliferation and survival\textsuperscript{14,15}. In addition to this, impaired DNA damaged repair (DDR) genes BRCA1/2, ATM etc. lead to an increase in microsatellite instabilities\textsuperscript{13}. Epigenetic regulatory circuits including DNA methylation and histone post translational modifications are also dysregulated in PDAC, which leads to further repression of tumor suppressor genes and upregulation of oncogenes\textsuperscript{16}.

Another crucial hallmark of PDAC is the dense desmoplastic stroma, which comprises up to 90\% of the tumor volume. This stroma is made up of extracellular matrix (ECM), vasculature and cancer associated fibroblasts (CAFs)\textsuperscript{17}. CAFs have distinctive subtypes with myofibroblastic or inflammatory phenotypes. CAFs are hypothesized to
originate from pancreatic stellate cells that are activated upon injury or chronic inflammation, at which point they deposit large amounts of ECM including laminins, fibronectins, collagens and hyaluronan into the space surrounding tissues$^{18,19}$. The stromal component is purported to be responsible for challenges in drug delivery and alterations of metabolic features of the tumor$^{20}$. However, studies focused on targeting the stroma have either resulted in non-significant results on decreasing the tumor mass, or in more aggressive forms of PDAC$^{21}$. In addition to stroma, the PDAC microenvironment is extremely hypoxic due to desmoplasia induced hypovascularization, which further activates pancreatic stellate cells$^{22}$. This hypoxic environment, coupled with the dense desmoplasia acts as a barrier to both proliferation of treatment, and also prevents T cell infiltration$^{23}$. Macrophages within the microenvironment are also hypothesized to create an immunosuppressive and pro-angiogenic environment, blocking T cell entry into the microenvironment, and supporting PDAC progression$^{24}$.

PDAC is also characterized by early progression to metastasis$^{25}$. Common sites of metastases include the peritoneum, liver, other gastrointestinal organs, lungs, and the nervous system. Metastases appear to be a clonal process – with the primary tumor being composed of subclones exhibiting variable metastatic potential$^{25}$. A crucial component of metastases in PDAC is epithelial to mesenchymal (EMT) transition, especially mesenchymal driver gene $ZEB^9$.

1.3: PDAC Management and Treatment

At this time, surgical resection is the only treatment for pancreatic cancer that is considered potentially curable$^{26}$. Pancreatoco-duodenectomy, also known as the Whipple’s procedure, distal or total pancreatectomy are the main surgical options for
pancreatic cancer. However, only 10-20% of patients present with surgically resectable disease at the time of diagnosis\textsuperscript{3,26}. Following curative surgery, the median survival of patients 15-20 months and a five-year survival rate of 8-15\%\textsuperscript{26}. However, 69-75\% of patients who undergo surgical develop recurrent tumor disease within 2 years, and 80-90\% relapse within 5 years\textsuperscript{27}.

Adjuvant chemotherapy with fluorouracil and folinic acid showed significantly improved overall survival and five-year survival rate compared to patients who did not receive chemotherapy after surgery\textsuperscript{28}. Similarly, adjuvant chemotherapy with gemcitabine, compared to surgery alone, has been found to increase disease -survival time, resulting in a statistically significant, albeit limited increase in survival from 20 months to 23 months\textsuperscript{29}. Studies evaluating the effectiveness of fluorouracil plus folinic acid, compared to gemcitabine, demonstrated no significant differences in survival, but significantly more adverse events in the fluorouracil plus folinic acid condition\textsuperscript{30}. Combination therapy of capecitabine and gemcitabine showed a slight increase in adverse events but significant increase in survival compared to gemcitabine alone\textsuperscript{28}. Other chemotherapies such as a combination treatment – mFOLFIRONOX (fluorouracil, folinic acid, irinotecan and oxaliplatin) compared to gemcitabine have demonstrated increased disease-free survival as well as overall survival. mFOLFIRINOX was associated with increased adverse events, but toxicities were manageable depending on the age and physical fitness of the patient\textsuperscript{31}. The current recommended standard of care post-surgery is mFOLFIRINOX in patients physically fit enough to tolerate adverse events, or combination gemcitabine and capecitabine treatment in less fit patients\textsuperscript{26,32,33}. Medical management in metastatic PDAC patients involves symptom control, pain
management, and palliative chemotherapy with FOLFIRINOX (mFOLFIRINOX with 5-fluorouracil) or gemcitabine\textsuperscript{34}. 

Targeted therapies and precision medicine in PDAC are still in its infancy. Erlotinib, an EGFR inhibitor, developed and approved to treat non-small cell lung cancer was one of the first attempts at introducing targeted therapies within PDAC\textsuperscript{35}. Erlotinib was incorporated in combination with gemcitabine as part of a first line chemotherapy regimen. While this resulted in a statistically significant increase in overall survival compared to gemcitabine alone, this was only an increase of two weeks\textsuperscript{36}. While this was intended to be a targeted therapy, the trial did not stratify patients based on EGFR overexpression or mutation\textsuperscript{36}. In contrast, an ongoing clinical trial, POLO trial, is a targeted therapy based on stratification of patients with germline BRCA1 or BRCA2 mutations (approximately 2\% of patients)\textsuperscript{37}. In this trial, patients with metastatic PDAC patients, were treated with the PARP inhibitor Olaparib as maintenance therapy as opposed to gemcitabine. Initial reports showed a significant improvement in progression-free survival in the Olaparib treatment group compared to the gemcitabine placebo group (7.4 months versus. 3.8 months, \( p = 0.004 \)) with mild indications of toxicity. The latest update on the trial was discussed at the ASCO Gastrointestinal Cancers Symposium in January 2021. Unfortunately, these results indicated that there was no statistically significant difference in survival in patients in the Olaparib arm compared to the placebo\textsuperscript{38}. 

In order to better understand the heterogeneity in PDAC that can be clinically actionable for targeted therapy, the field has sought to expand into categorizing the disease into clinically prognostic subtypes.
1.4: PDAC Subtypes

Recent efforts in PDAC molecular subtyping have better characterized disease heterogeneity using transcriptomic signatures associated with clinical features, which can be used to define multiple PDAC subtypes. Here I will describe the most widely used and reproduced subtyping systems in PDAC, which primarily rely on quantifying transcriptional signatures.

1.4.1: Collisson Classification

The Collisson classification was the first major attempt at PDAC molecular subtyping method to be widely used\textsuperscript{20}. Using transcriptional profiles from primary PDAC samples, along with mouse and human PDAC cell lines, variable gene signatures using non-negative matrix factorization (NMF) on microarray samples to identify gene signature clusters that show differential expression. With these gene signatures, they performed consensus clustering to identify three subtypes – classical, quasi-mesenchymal and exocrine-like. The classical subtype demonstrated high expression of adhesion-associated genes and epithelial genes, while the quasi-mesenchymal subtype highly expressed genes associated to a mesenchymal profile, and the exocrine-like subtype exhibited a high expression of tumor-cell derived digestive enzyme genes\textsuperscript{20}. Within this system, patients whose tumors were classified as classical fared better in terms of survival compared to patients classified as quasi-mesenchymal. In cell lines, classical cells have shown relatively higher dependence on KRAS over quasi-mesenchymal cell lines\textsuperscript{20}. Additionally, classical cell lines show greater sensitivity to Erlotinib, an EGFR inhibitor, compared to quasi-mesenchymal cell lines, which showed greater sensitivity to gemcitabine\textsuperscript{20}. 
1.4.2: Bailey Classification

Bailey, et al, also defined a subtype classification for PDAC based on the analysis of a cohort of PDAC tumors with high epithelial content (> 40%) and identified four unique subtypes – squamous, pancreatic progenitor, immunogenic and endocrine exocrine or ADEX. This was based on variable expression of transcription factors and downstream targets. Squamous tumors harbored mutations in TP53, KDM6A and upregulated hypermethylated pancreatic endodermal cell fate determining genes. Progenitor tumors overexpressed genes involved in early pancreatic development. Immunogenic tumors upregulated genes involved in immune pathways, and the ADEX tumors highly expressed KRAS activation genes, and both exocrine and endocrine differentiation genes. These subtypes corresponded with specific histologies – squamous tumors with adenosquamous carcinoma, progenitor and immunogenic with mucinous non-cystic adenocarcinoma, and ADEX with rare acinar cell carcinoma. The squamous subtype was shown to have the poorest survival overall.

1.4.3: Moffitt Classification

In 2015, Moffitt, et al developed another classification scheme for PDAC by focusing on deconvoluting tumor cell intrinsic gene expression signatures. Using a cohort of patient and PDX PDAC tumors samples, they applied a virtual microdissection approach to computationally separate gene expression signals arising from the stroma and tumor microenvironment. Using consensus clustering, they identified two distinct stromal signatures – “activated” and “normal”. The normal stromal signature was characterized by expression of key marker genes of pancreatic stellate cells, whereas the activated stromal signature was characterized by genes associated with macrophages,
and genes involved in tumor promotion, including members of the Wnt family of genes. They also described two distinct cell intrinsic subtypes – classical and basal-like. The basal-like subtype was so called due to the similarity to the basal subtypes in bladder and breast cancers. Patients with basal-like tumors have decreased survival compared to classical tumors\textsuperscript{40} Basal-like tumors have responded more favorably to adjuvant chemotherapy than classical tumors, but have recently been associated with treatment resistance to FOLFIRINOX\textsuperscript{40}.

1.4.4: Integrated Analysis of Classifications by the TCGA

The Cancer Genome Atlas (TCGA) consortium applied the clustering techniques from the Collisson, Moffitt, and Bailey studies on their large cohort of patient PDAC tumors\textsuperscript{12}. While they were able to recapitulate Collisson’s three subtypes, Bailey’s four subtypes and Moffitt’s basal-like and classical subtypes, they found that only Moffitt’s subtypes were consistently independent of sample purity. Considering only high purity samples (>30% cellularity), the pancreatic progenitor subtype from Bailey’s classification and the Collisson classical subtype overlapped with Moffitt’s classical subtype, while the Bailey squamous subtype and Collison quasi-mesenchymal subtypes overlapped with Moffitt basal-like subtypes. The immunogenic and ADEX subtypes were strongly associated with low purity tumors, suggesting that these subtypes are likely a result of signals from non-tumor cells. It is hypothesized that the variation in the samples and analyses used explain the differences in the subtypes, but overlap across classification methods\textsuperscript{15}. 
1.4.5: *Puleo Classification*

Puleo et al. released a classification based on independent component analysis of PDAC tumors with deconvolution of normal, microenvironment and tumor compartments\(^{41}\). They identified two tumor subtypes, which recapitulated the Moffitt classical and basal-like subtypes. They also identified stromal components, activated stroma component – with overexpression of extracellular matrix organization, collagen formation, and focal adhesion. The second stromal component was the inflammatory stroma component – with high expression of IL-6 and other markers of inflammation. This classification also reported that the ADEX subtype previously described emerged from normal cell contamination. They ultimately identified five subtypes from unsupervised clustering of their entire cohort – “pure” classical and basal-like with signal emerging primarily from the tumor, and three microenvironment subtypes – immune classical, stroma activated and desmoplastic. The authors argue that this classification system, which is recapitulative of the Moffitt subtypes, provides more clinical value by incorporating microenvironmental heterogeneity\(^{41}\).

1.4.6: *COMPASS Trial: Toronto Classification*

Most recently, researchers leading the COMPASS trial for PDAC patients with advanced disease addressed two pressing issues facing PDAC subtyping – first, lack of accurate characterization of unresectable, advanced tumors, which make up the majority of patients in clinic, and second, low cellularity of samples\(^{42}\). Here, over 200 PDAC tumor samples were collected using laser capture microdissection – resulting in a minimum of 80% cellularity. Using NMF on RNAseq performed on these samples, they defined the Toronto classification, and identified five subtypes labeled “basal-like A”, “basal-like B”,...
“classical-A”, “classical-B” and a novel subtype that was previously inconsistently classified due to multiple gene expression signatures, termed “hybrid”\textsuperscript{42}. They found that classical A/B tumors were more likely to be early stage tumors, with basal A tumors making up a large portion of stage IV tumors. Furthermore, Moffitt basal-like tumors showed high association with the basal-like B and hybrid subtypes. In addition to identifying the hybrid subtype in bulk tumors – tumors that show expression of both classical and basal-like signatures, they also identified that classical and basal-like cells exist intratumorally within the same tumor – resulting in hybrid tumors at the bulk scale\textsuperscript{42}.

1.4.7: Current Consensus in PDAC Subtyping

Integration of these subtypes has given rise to a general consensus in the field of PDAC subtyping of two overarching cell intrinsic subtypes – the classical/progenitor subtype and the basal-like/squamous subtype\textsuperscript{39,12,43}. However, despite these advances, and associations being drawn in difference in response to treatment, transcriptomic subtyping within PDAC is still in its early stages. It has not yet been applied to identify potential avenues for targeted therapy. Subtypes are not currently widely assessed in clinical care, and are not used to inform on choice of treatment. Moreover, clonal and subclonal evolution as well as therapeutic intervention can result in molecular signatures that change the original subtype classification, highlighting the relevance of intratumoral molecular and functional heterogeneity that underlies high-level subtype groupings\textsuperscript{41,42}.

1.5: Dissertation Objective

Evaluation of the mutational and transcriptional landscape of PDAC has revealed a wealth of knowledge that has served as the foundation for further biological investigation about the genes and signaling pathways relevant in tumor maintenance and progression.
In addition to this, several groups have developed computational tools to identify subtypes of PDAC using mutational signatures and transcriptomic analyses in both patient and PDX models. Despite these biological and computational characterization efforts, **targeted therapy positioning in these tumors remains suboptimal and subsequently the clinical prognosis for these patients remains dismal.** So far, large-scale subtyping and characterization efforts based on statistical and bioinformatics approaches centered on transcriptomic and mutational landscapes. These characterization strategies serve as an indirect measure of tumor dependency, and have been insufficient to stratify and define pancreatic cancer on a patient-by-patient or tumor-by-tumor basis.

My long-term goal is to develop translational computational platforms for PDAC that enable disease characterization and patient stratification that will inform on potential avenues for therapeutic intervention. The objective of this dissertation is to establish computational methodologies that will lead to better target discovery and identification using functional genomics approaches and to develop a novel platform that will characterize the transcriptional heterogeneity of PDAC.

To this end, the focus of my research and this research is centered around (1) developing appropriate analytical tools for *in vivo* and targeted library screens conducted on PDAC models (2) quantifying PDAC transcriptional diversity using a novel network-based approach to capture continuous patterns of heterogeneity (3) adapt our network-based approach to single cell RNAseq samples across PDAC patient derived xenograft models and clinical core needle biopsy samples to recapitulate continuous intratumoral
heterogeneity and (4) anchor genetic dependencies within our transcriptional network to inform on subtype specific vulnerabilities intratumorally.

Taken together, these efforts: a) provide an accessible framework to serve as a discovery and hypothesis generation platform, b) allow comprehensive model characterization of functional genomics inferred oncogenic signaling, c) stratify tumors and models in terms of survival based on their enrichment, and d) serve as an avenue to support better clinical positioning of targeted therapies for patients with PDAC.
Chapter 2: Materials and Methods

PDX models and Sequencing (RNA and Whole Exome)

A total of 48 models were utilized in this paper. PDAC PDX models were obtained from the labs of Dr. Michael Kim (Department of Surgical Oncology, MD Anderson Cancer Center) and Dr. Scott Lowe (Memorial Sloan Kettering Cancer Center)\textsuperscript{44,45}. PDXs were propagated and maintained in NOD scid gamma (NSG) mice carrying NOD.Cg-Prkdc\textsuperscript{scid} Il2rg\textsuperscript{tm1Wjl}/SzJ (Jackson Labs).

PDX Sequencing

Whole exome library preparation and sequencing

Whole exome sequencing (WES) libraries were prepared using the Agilent SureSelect XT library preparation kit in accordance with the manufacturer’s instructions. Briefly, DNA was sheared using a Covaris LE220. DNA fragments were end-repaired, adenylated, ligated to Illumina sequencing adapters, and amplified by PCR. Exome capture was performed using the Agilent SureSelect XT v4 51Mb capture probe set and captured exome libraries were enriched by PCR. Final libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems), Qubit Fluorometer (Life Technologies) and Agilent 2100 BioAnalyzer and were sequenced on an Illumina HiSeq2500 sequencer using 2 x 125bp cycles. Base calling and filtering were performed using current Illumina software and adapters were trimmed using Trim Galore [55]. Sequences were aligned to both NCBI genome human build 37 and mouse build 38 using Burrows-Wheeler Aligner\textsuperscript{16}; identified mouse reads were removed from the original FASTQs and then the files were realigned again to NCBI build 37 using BWA. Picard was used to remove duplicate reads (http://picard.sourceforge.net); base quality scores were
recalibrated using GATK\textsuperscript{17}. Assessment of reads that do not align fully to the reference genome was performed, locally realigning around indels to identify putative insertions or deletions in the region. Variants were called using GATK HaplotypeCaller, which generates a single-sample GVCF. To improve variant call accuracy, multiple single-sample GVCF files were jointly genotyped using GATK GenotypeGVCFs, which generates a multi-sample VCF. Variant Quality Score Recalibration (VQSR) was performed on the multi-sample VCF, which adds quality metrics to each variant that can be used in downstream variant filtering.

\textit{RNA library preparation and sequencing}

RNA sequencing libraries were prepared using the Illumina TruSeq Stranded mRNA sample preparation kit in accordance with the manufacturer's instructions. Briefly, 100ng of total RNA was used for purification and fragmentation of mRNA. Following conversion of mRNA to cDNA, DNA was adenylated, ligated to Illumina sequencing adapters, and amplified by PCR (using 10 cycles). Final libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems), Qubit Fluorometer (Life Technologies) and Agilent 2100 BioAnalyzer and were sequenced on an Illumina HiSeq2500 sequencer (v4 chemistry) using 2 x 50bp cycles. Base calling and filtering were performed using current Illumina software. Reads were aligned to a joint index of NCBI genome human build 37 and mouse build 38 with STAR aligner\textsuperscript{18}. Reads that map uniquely and unambiguously to the mouse genome were removed from the FASTQ files and then the files (containing unmapped reads and reads mapped at least once to the human reference) were remapped to GRCh37 using STAR aligner and Gencode 19 annotation. Gene expression quantification was performed with featureCounts.
Genes with raw read counts present in less than 20% of samples were removed from further analysis, and log normalized counts were generated on the 14175 filtered genes using DESeq\textsuperscript{49}.

**Cell Culture**

PDX cell lines were seeded in treated tissue-culture plates (Corning) in DMEM/F12 medium (Gibco) supplemented with 10% Fetal Bovine Serum (Gibco), Penicillin (50 units/mL) and Streptomycin (50 µg/mL) (ThermoFischer Scientific). Phosphate Buffered Saline (PBS) was utilized prior to trypsinization and for general cell washing purposes (Thermo Fisher Scientific). Cells were regularly trypsonized (0.25%, Trypsin-EDTA, Gibco) prior to reaching 70% - 80% confluence, and maintained on 10 cm and 15 cm treated tissue-culture dishes (Corning). Viable cells were counted using a Cellometer mini and 0.2% Trypan Blue staining (Nexcelom).

**Design and construction of custom CRISPR Library**

The custom CRISPR-Cas9 sgRNA library was constituted by 3,367 sgRNAs and designed to both incorporate and expand upon the original RNAi-defined targets. The library was constructed using chip-based oligonucleotide synthesis and cloned into a pRSG16-U6-sg-HTS6C-UbiC-TagRFP-2A-Puro lentiviral vector (Cellecta) as a pool. RSA was used to rank essential genes from shRNA screens. Genes were curated based on RSA < 0.05 and FC < -2 in at least one model (N=100). Curated gene targets were further annotated by incorporating neighbors with a PPI score ≥ 0.80 (STRING, version 10), and TPM (transcripts per million) > 2 in that model\textsuperscript{48}. In addition, 50 non-essential genes and 50 essential genes were added to have a final set of 654 genes\textsuperscript{50}. 
CRISPR Cas9 Screening *in vivo and in vitro*

Using the custom barcodes lentiviral sgRNA library, PDX lines (PATC69, PATC124, PATC53 and PATC153) containing the lentiCas9-blast vector (addgene, plasmid #52962) were transduced *in vitro* using 8 µg/mL Polybrene (Sigma-Aldrich). PDX lines transduced with Cas9 were constantly kept at 10 µg/mL blasticidin (Thermo Fischer Scientific). Libraries were transduced at 1000X coverage and a multiplicity of infection (MOI) of 0.3. Media was replaced after 14 hours, and at 48 hours, and MOI was confirmed by checking RFP percentage with flow-cytometry 48 hours post-transduction. Immediately following flow cytometry, 2 ug/mL puromycin (Thermo Fisher) was added to each transduced cell population for 72 hours. Immediately following puromycin selection, a Reference population was collected (1000X) and stored at -80 °C. Cells were allowed to culture for an additional 12 days in order to allow for some initial sgRNA-directed cutting prior to starting the *in vivo* screen, to reduce potential noise derived from disproportional cell doubling following orthotopic implantation. A secondary Reference (1000X), at the time of injection, was then collected. Remaining cells were split and moved into *in vitro* and *in vivo* settings. Three independent *in vitro* screens were seeded (1000X) into 15 cm treated plates (Corning). For *in vivo* implantation, cells were combined into 1:1 PBS/Growth-Factor Reduced Matrigel (Corning), and injected orthotopically at 1000X coverage per mouse. Immunodeficient NSG mice were leveraged for *in vivo* screening. For *in vitro* screening populations, cell populations were collected at 10, 20, and 30 days post injection and secondary Reference collection. For *in vivo* screening, the entire pancreas and tumor of each mouse was collected at day 30 post injection. DNA extraction and barcode library preparation were conducted in similar fashion to the RNAi screens.
For each condition, cells were lysed using SDS and DNA was sheared using sterile 23-gauge 1-inch needles (Becton Dickinson). DNA was isolated using Phenol:Chloroform extraction and Ethanol precipitation. Nested PCR was utilized to amplify and prepare barcode populations for NGS (Cellecta).

**Network Display**

Cytoscape (version 3.7.2) was leveraged for visualization of both the PDAC co-expression network and STRING-anchored co-dependency networks\(^{51,52}\). The PDAC co-expression network was visualized using a Prefuse Force-Directed Layout, with node color displayed based on Random Walk defined clusters and node size representative of Betweenness Centrality. For visualization purposes, only nodes with 12 or more edges are represented, and edges are not displayed in representative co-expression network figures. For STRING-anchored co-dependency networks, a Perfuse Force-Directed Layout is also applied, with quantile-normalized Bayes Factors represented as a blue-to-red color distribution. For each *in vivo* and *in vitro* condition, networks were constructed for each independent PDX line and then merged for comparison based on overlapping nodes. All STRING-defined edges were maintained for *in vivo* and *in vitro* network merging.

**Clinical correlations with signature**

Clinical data from patients corresponding to the PDAC PDX cohort was provided by the lab of Dr. Eugene Koay\(^ {53} \). Differentiation status on 45 models was categorized as “Poor” and “Moderate” and histology status on 35 models was categorized as “Locoregional” if the records indicated as “regional” or “locoregional”, and “Distant” if site was outside the pancreas. Chi square tests were used to evaluate differences between
the cluster-defined Classical, Quasi-basal and Basal groups as well as the Moffitt binary classification.

**Single Cell Data for PDX-derived Cell Lines**

Seurat version 3.1\textsuperscript{54} was used to analyze all single cell analysis. Each of the lines, PATC124, PATC53 and PATC69 were analyzed separately. PATC53 contained two replicates, which were merged for analysis. For all PDX cell lines, single cells with a minimum of 350 expressed genes and less than 10% mitochondrial reads were retained. Genes expressed in less than 3 cells and mitochondrial genes were removed from further analysis. The data was log normalized, transformed using the “vst” function with top 2000 variant genes. The total RNA count, cell cycle score and mitochondrial reads were regressed out. For PATC124, principal-component analysis and uniform manifold approximation and projection (UMAP) with the first 15 dimensions was performed, followed by identifying clusters using a resolution of 0.15. For PATC69, principal-component analysis and uniform manifold approximation and projection (UMAP) with the first 15 dimensions was performed, followed by identifying clusters using a resolution of 0.10. PATC53 contained two replicates, which were integrated as one dataset following normalization and variant stabilization. Total RNA count, cell cycle score and mitochondrial reads were regressed out of the integrated dataset, and PCA and UMAP on first 15 dimensions was performed, further identifying clusters using a resolution of 0.10.

**Core Needle Biopsy Single Cell Analysis**

Seven CNB samples were used – four primary tumors, one liver, lung and vaginal metastases sample each\textsuperscript{55}. All seven samples were filtered to have a minimum of 350
expressed genes and less than 25% mitochondrial reads were retained. Similar to the PDX analysis, genes expressed in less than 3 cells and mitochondrial genes were removed from further analysis. Each sample was log normalized, transformed using the “vst” function with top 2000 variant genes. The seven samples were then integrated, following which, the total RNA count, cell cycle score and mitochondrial reads were regressed out. Principal-component analysis and UMAP with the first 20 dimensions was performed, followed by identifying clusters using a resolution of 0.20. Cell types associated with clusters were identified using established stromal markers and epithelial cell markers specific to liver and pancreas.

**Feature Barcoding Vector**

The feature barcode vector (LentiCRISPR-E-10xcs1) was built using the pLentiCRISPR-v2 (addgene: 52961) as the base vector. All of the molecular modifications were performed by Epoch Life Sciences (Missouri City, Tx). The pLentiCRISPR-v2 was modified to an optimized sgRNA scaffold that included the 3’ 10x capture sequence 1:

(cgtttCagagctaTCGTGgaaaCAGCAtagcaagttaaaataaggctagtccgtttatcaactgaaaaagtggca
ccgagtcggtgcGCTTTAAGGCCGGTCCTAGCAATtttt); a ccdb bacterial expression cassette between the BsmBI restriction sites was introduced to reduce background during sgRNA cloning and library generation; and an N-terminal Flag-sv40 NLS was added to Sp. Cas9-nucleoplasmin NLS-P2A-Puro.

**Preparation of Feature Barcoded sgRNA Knockout Populations**

Lentiviral transductions of four separate feature barcode sgRNA vectors (targeting ABCG8, ILK, SMAD4 and ZEB1), were conducted on separate cell populations for both
the PATC69 and PATC53 PDX lines. Lentivirus was concentrated through ultracentrifugation, resuspended in 200 µL of PBS, and stored at -80 °C until use. For each condition, 1x10^6 cells were transduced in 10 cm treated plates (Corning) using 8 µg/mL Polybrene (Sigma-Aldrich). Media was replaced after a 16-hour incubation, and each cell population was washed with PBS and then placed under puromycin selection for 72 hours. Following selection, all conditions were cultured for 22 days (or 10 days post “CRISPR Screen Injection point”) to match the in vitro CRISPR screen control-separation profile at day 10 (Extended Data Figure 2A). Prior to library preparation for scRNAseq, knockout populations were combined in equal proportion for each PDX line. Library preparation was conducted on a total of 10,000 cells per PDX line, resulting in an approximate coverage of 2500 cells per condition.

**sgRNA Phenotype Confirmation and Confirmation of Site-Specific Cutting**

Utilizing the same feature barcoded populations prepared for scRNAseq, 1500 cells/well were seeded in triplicate in 12 well tissue culture plates (Corning) immediately following 72 hours of 2 µg/mL Puromycin selection. Cells were then cultured for a minimum of 10 doublings. Individual plates were then stained with 0.5% crystal violet (in 25% methanol) for 2 hours. Plates were washed in water, dried overnight and then digitally scanned. After digitally scanning the plates, crystal violet was dissolved in equal volumes of 1% SDS, and 200 µL of each sample was moved into 96-well plates to measure absorbance at 570 nm. Relative growth was quantified based on the internal sgABCG8 negative control. All data was graphed using GraphPad Prism v 8.0.

Puromycin selected cells were collected for Sanger sequencing to confirm sgRNA induced indel formation relative to non-infected populations. Cell pellets for each sgRNA
across both PATC69 and PATC53, 1x10^6 cells each, were isolated at Day 12 and Day 40 for each PDX line (sgRPS27A indels representative at Day 12, all other sgRNAs at Day 40). Cell pellets were centrifuged, washed once with PBS, and frozen at -80 °C. All Sanger sequenced regions were normalized against respective non-transduced PDX line populations, 1x10^6 cells/pellet. Primers for each cut site were developed to allow for 400 - 800 bp products, and primer sites were run on 2% agarose gels and extracted following amplification. Site-specific sequencing primers were utilized for Sanger sequencing, and indels distributions were calculated using the Synthego ICE Analysis Tool\textsuperscript{61}. 
Chapter 3: Novel Methodology for Target Library CRISPR Screening

This work is based on efforts outlined in our recent BioRxiv submission “Diversity Across the Pancreatic Ductal Adenocarcinoma Disease Spectrum Revealed by Network-Anchored Functional Genomics” by Rose and Srinivasan, et al. (2020), DOI: https://doi.org/10.1101/2020.09.17.302034. As co-first author, Dr. Johnathon Rose provided substantial experimental support for this effort, enabling our findings. The manuscript is currently under review.

Existing methodologies can accurately identify essential genes within CRISPR screens of large library complexity, however they cannot be applied to small in vivo libraries due to overfitting these small datasets with limited controls, resulting in a loss of accuracy. Here, we have modified the currently used Bayesian classifier of gene essentiality, BAGEL, used to detect context-specific essential genes. Using this adapted method, Low-Fat BAGEL, we have compared essentiality profiles among PDX models and within the models by comparing the differentially essential genes in vitro and in vivo.

3.1: CRISPR and its Uses in Functional Genomics

Systematic perturbations of genes, and the subsequent tracking of these genetic perturbations to observe cellular level dropout, provides a phenotypic readout of causal genes underlying dependence in a cell and tissue specific manner\textsuperscript{62,63}. Furthermore, the comprehensive analysis of these genetic dependencies within the backdrop of known molecular interactions and pathways provides a wider view of the major contributors to proliferative signaling. The recent discovery and subsequent adaptation of the clustered regularly interspaced short palindromic repeats associated Cas endonuclease system
(CRISPR/Cas) technology has enabled high-precision knock-out screens for both arrayed and pooled settings. As a result, this recent advancement has vastly propelling the field of functional genomics forward by providing the capacity for precise systematic gene-level perturbations, using both targeted and whole-genome libraries, to uncover genetic dependencies in disease models.

The Nobel Prize winning discovery of the CRISPR/Cas system was first described as a form of adaptive immunity in prokaryotes. Identified in most species of bacteria and archaea, it is involved in developing resistance to bacteriophages. Of the various systems identified, one of focus and early investment for application in eukaryotic model systems was the *S. pyogenes* derived bacterial type II CRISPR system using the Cas9 endonuclease. Adaptation of this system for mammalian cells allows this engineered Cas9 system to introduce a frameshift mutation into early exon coding regions of specific genes of interest, resulting in targeted loss of function. Specifically, the Cas9 proteins generate a double stranded break at a target locus, located three bases upstream of an “NGG” protospacer adjacent motif (PAM) recognition site at the 3’ end of a 20-nucleotide base pair sequence complementary to the short guide RNA (sgRNA). The induced double stranded break is repaired by non-homologous end joining (NHEJ), which is error prone, leading to indels. Targeting coding regions thus allows for gene knockout, enabling the study of gene function in a high-throughput, pooled genome-wide scale to systematically interrogate function in different biological systems. The CRISPR/Cas screening system is a rapidly adapted and evolving mechanism to interrogate genes essential in specific genetic and disease contexts. Given this, several genome-wide CRISPR/Cas9 screening libraries have been developed for genetic screening in human cell cultures, mouse and
other model systems\textsuperscript{62-65}. This system offers an advantage over previously used systems such as RNA interference by avoiding the off-target effects and incomplete gene knockdown of the latter system. A typical genome wide CRISPR screening library consists of either mouse or human coding genes targeting multiple loci, with four to ten loci us sgRNAs per gene\textsuperscript{65,68}. They are available as either a single plasmid system with the sgRNA and Cas9 protein on a single vector\textsuperscript{69}, or a two-plasmid system with the sgRNA and Cas9 protein on separate lentiviral vectors.

High throughput, pooled CRISPR screening consists of designed sgRNAs synthesized and cloned into a lentiviral library (with or without Cas9) and transduced into living cells at a low multiplicity of infection (MOI) of 0.3, to prevent the likelihood of multiple guides integrating into a single cell. Successfully transduced cells are then cultured, and the integrated sgRNA are replicated along with the rest of the host genome during cell division. An initial time point (Day 0, immediately following selection) is collected to serve as a reference population\textsuperscript{67}. To identify essential genes, at least one other subsequent time point is collected at later doublings for comparison. These two, or more populations are then compared. Within these screens, the outcome of interest is gene essentiality within the context of cellular fitness\textsuperscript{64}. If a gene is essential, it is expected that a knock out of this gene, and depletion of the encoded protein within the cell, would result in decreased viability of that cell. In a negative selection screen, the later time-points are compared to the Day 0 timepoint to identify genes that have “dropped out” of the population. In the case of a certain genetic background or disease, “context essential” genes are ones whose normal function is required for the continued survival and proliferation of that cell. Knockout of these genes would result in the phenotype of cell
death, resulting in a fewer surviving cells with the corresponding sgRNA in the later time-points compared to Day 0. Thus, the typical statistical readout of CRISPR screens is foldchange, calculating the relative abundance of guides in the later time-points compared to the reference population. A major assumption in taking the normalized log foldchange of sgRNA abundance is that the majority of genes will have no effect or phenotype upon knockout – resulting in a foldchange centered around $0^{18}$.

3.2: Analytical Tools in Genome-Wide CRISPR Screening

With the advent of accessible CRISPR/Cas9 screening, there has been a vast increase in publicly available high-quality screens across various genetic and disease backgrounds. In turn, this has enabled the development of several new analysis pipelines. Methods for these large-scale screens involve the typical steps of quantifying sgRNA abundance by mapping sequencing reads back to the sgRNA library, comparing the later time-point screens to the reference timepoint to compute relative abundance of each sgRNA and aggregating the effects of all sgRNAs on a gene-wise scale to understand the effect of each gene in the library. Here, I will briefly describe a few analytical tools developed to address the bioinformatic complexity of screening data.

3.2.1: Redundant siRNA Activity (RSA)

This method was initially developed for the analysis of short hairpin RNA (shRNA) in RNAi screening$^{70}$. RSA relies on the core principle of negative selection CRISPR screening – an “hit” or an essential gene would show greater differences in overall foldchange in the at the endpoint of a screen compared to the reference population. All targeting guides are ranked by decreasing foldchange between the endpoint to the reference point. A p value is calculated using a hypergeometric distribution to calculate
the probability for each gene under the null distribution assuming all guides behave uniformly. P-values for all guides for a single gene are then used to determine whether the guides for that gene are significantly higher than the rest of the distribution. While RSA analysis is robust in shRNA screen analysis, it consistently does not perform as well with CRISPR screens when compared to methodology designed specifically for these screens.

3.2.2: MAGeCK

The Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) method was first developed in 2014 by Li et al, and is a widely used analytical tool for CRISPR screens\(^{71}\). MAGeCK employs a negative binomial distribution similar to RNAseq analysis platforms such as DESeq2\(^{49}\). The variance of median normalized raw sgRNA counts between the later time-points and control populations is used to calculate a guide level p value. The Robust Ranking algorithm (RRA) implemented within MAGeCK to combine the guide level p values to compute a gene level False Discovery Rate (FDR) using the Benjamini-Hochberg correction for multiple testing. The MAGeCK-RRA method is designed for both loss-of-function and positive selection gain-of-function screens. The MAGeCK toolset also contains a version for identifying gene hits across treatment and drug conditions. This Maximum Likelihood Estimate (MLE) method uses a design matrix of samples from each condition and extends the negative binomial model to include a generalized linear model to fit a coefficient at the gene level and compute a p value and estimated effect size. Like MAGeCK RRA, MLE also computes an FDR and gene hits for both loss-of-function and gain-of-function screens.
3.2.3: CERES

Measurement of gene effect sizes has been found to be confounded by copy number variations (CNVs) present at the target site of individual guides. By introducing Cas protein induced cuts, regions with high copy number might have induced an antiproliferative effect on transduced cells due to DNA damage, rather than context specific gene essentiality. To address this, the Broad Institute developed a method termed CERES that computationally corrects for gene variation bias\textsuperscript{72}. CERES was developed as part of the Achilles/DepMap effort that initially released genome-wide CRISPR screen data from 342 cancer cell lines, known as the Avana project. CERES compares the effect of the same guide across all cancer cell lines to normalize per cell line. While the CERES method successfully reduces the number of false positive hits in CRISPR screens, it is only applicable in experiments involving multiple cell lines to compare the effects of genomic amplification. In addition, CERES also requires copy number profiles for cell lines/models being screened, which can be prohibitive in cases where this data is not publicly available. A more recent method by investigators at the Sanger Institute developed a method termed CRISPRCleanR\textsuperscript{73}, aimed at addressing the issue of gene independent effects due to copy number amplification. CRISPRCleanR uses an unsupervised approach to segment single sgRNA foldchange across the genome, without requiring prior knowledge of the copy number of the targeted loci. This method outputs sgRNA fold changes and normalized read counts, making it compatible with other analysis tools for downstream hit discovery.
3.2.4: JACKS

To address the issue of alternate sgRNAs targeting the same gene confounding the essentiality estimates obtained from CRISPR screening, researchers at the Sanger institute developed JACKS (Joint Analysis of CRISPR/Cas9 Knock-out Screens)\textsuperscript{74}. This is a Bayesian method that models efficacies of all sgRNAs across multiple screens using the same library. JACKS models the log2 foldchange of sgRNAs comparing treatment to control (or an endpoint screen compared to a reference). The authors report that the JACKS method provides improvements on gene essentiality estimations, and ultimately allow for the scaling down of library sizes in terms of the number of sgRNAs per gene.

3.2.5: BAGEL

BAGEL (Bayesian Analysis of Gene EssentiaLity)\textsuperscript{75} is, as the name suggests, a Bayesian method that relies on a set of gold standard genes – “core essential” genes that are considered essential in all cell lines and contexts, and a set of “non-essential” genes that do not induce drop-out when knocked out in any cell lines or contexts. These gold standard genes were initially curated from previously published data describing essential and non-essential genes. This list was further expanded with increasing availability of CRISPR screens, with the final list containing 684 core essential genes and 927 non-essential genes\textsuperscript{69}. The fold-change in the guide RNA (sgRNA) abundance after knockout of each gene is measured, and compared against the fold change distribution of the core essential and the non-essential reference sets. BAGEL then computes a ratio, the Bayes Factor (BF), which is the log2 likelihood of each gene belonging to either the core essential gene distribution or non-essential gene distribution. An updated version of the BAGEL, platform, BAGEL version 2\textsuperscript{76}, was recently released which provides better quality
control and improved dynamic range of the BF, allowing for better inferences of the magnitude of essentiality.

3.3: Introduction of Surfaceome Library and Logic

Application of a perturbation-based functional genomics approaches, while incredibly informative in comprehensively characterizing the network of deregulated cellular signaling, is currently not feasible directly in tumor samples. Current genome-wide approaches in traditional cell lines, while optimized, may not be a faithful recapitulation of the diversity of signaling of the disease they represent. Especially in the case of cancer, cell lines may not maintain the level of heterogeneity of the original patient tumor, and tend to become more homogenous with further passaging. Patient-Derived Xenografts (PDXs), which are direct implants of patient tumors into murine models, along with associated PDX-derived early passage cell lines (PDX lines) address some of the shortcomings of using traditional cancer cell lines by better maintaining morphological, genetic and molecular characteristics of the original tumor. Leveraging this unique strength of PDX’s in more closely recapitulating the dynamics of patient tumors, we aim to establish a functional genomics approach to more comprehensively characterize the specific dependencies within the context of each PDAC model.

Genome-wide CRISPR library screens are demonstrated to be a feasible and easily accessible method to quantify model specific vulnerabilities in traditional cells lines and in vitro settings of pre-selected screen-able models. However, this remains challenging in alternatives such as patient derived xenografts (PDX’s), in vivo models or clinical samples. The large number of transduced cells required for adequate coverage may not be feasible in models with slow doubling time or in vivo implantation. CRISPR
screens have been conducted in human cancer cell lines on a whole-genome basis at a coverage of 200-300x, which translates to 14 to 16 million transfected cells. While this is feasible when conducting *in vitro* screens, it is not feasible to perform screen *-in vivo* while maintaining the complexity of the library. For this reason, custom, small library CRISPR screens asking a targeted question centering around a hypothesis are becoming more common\textsuperscript{79,80}.

To more closely recapitulate the signaling diversity of PDAC patient tumors, the Draetta lab, led by Dr. Wantong Yao, Dr. Johnathon Rose and Dr. Alessandro Carugo have optimized a screening protocol for parallel *in vitro* and *in vivo* CRISPR screening in PDX models\textsuperscript{81,82}. Owing to the limitations of conducting genome wide screens in this setting, we have developed a custom PDAC prioritized screening library, termed the “surfaceome” library. This library was initially developed as an RNAi library with an emphasis on targeting proteins localized and integrated into the extracellular face of the plasma membrane\textsuperscript{82}. Surface protein targets were included in the library based on evidence of differential expression in pancreatic tumors compared to matched-normal tissue, copy number versus RNA expression correlation trends from the TCGA PDAC dataset, and SILAC screening for mutant *Kras* dependency. Upon completion of the parallel *in vitro* and *in vivo* RNAi screens, we subsequently built a custom sgRNA library that expanded upon the RNAi-identified protein surface dependencies across all tested models. Using a CRISPR-based approach, the library incorporated sgRNAs targeting highly connected proteins defined by a stringent STRING protein-protein interaction (PPI) network used to evaluate oncogenic signaling redundancies and interpret co-dependencies. The final iteration of the surfaceome library is a PDAC prioritized surface
protein library including the downstream targets totaling 657 genes, with 5 sgRNA each, of which 50 “core” essential and 50 non-essential genes were selected from the Hart reference set. This library was jointly designed by Dr. Wantong Yao, Dr. Johnathon Rose, Dr. Alessandro Carugo and Sahil Seth.

3.4: Analysis of Custom Library Screens

Parallel *in vitro* and *in vivo* CRISPR screens were performed by Dr. Johnathon Rose on three PDX lines (PATC69, PATC124 and PATC53) (See methods). We conducted each *in vitro* CRISPR screen as a time course with a reference time-point and three follow up time-points, with a matched *in vivo* endpoint. Current analytical frameworks described above, such as MaGeCK, CERES, JACKS, and BAGEL are designed for genome-wide screens, and are not tailored to account for the far smaller training sets of these screens.

Thus, in addition to developing a sophisticated experimental framework to conduct custom library CRISPR screening, we have also developed an analytical framework to better model smaller datasets. Here, we introduce an adapted algorithm of the BAGEL method implemented as an option within the software, **Low-Fat BAGEL** for the analysis of small, targeted library CRISPR screens. We demonstrate the utility and improved accuracy of Low-Fat BAGEL compared to other genome-wide analysis tools benchmarked in our custom library CRISPR screens of two pancreatic cancer PDX derived cell lines.

3.5: Surfaceome Library Analysis using BAGEL v1

The surfaceome library contains PDAC prioritized surface proteins and downstream targets totaling 657 genes, with 5 sgRNA each, of which 50 core essential
and 50 non-essential genes were selected from the Hart reference set. With the inclusion of the Hart reference set, we first analyzed our screens using the BAGEL algorithm.

In the original BAGEL algorithm, the Bayes Factor of a given guide is computed as a log2 likelihood of being essential based on comparison to the foldchange distributions of the core essential and non-essential genes. Gene level Bayes Factors are obtained by summing across all guides corresponding to that gene. On a genome-wide scale, relying on our gold standard set of five guide 684 core and 927 nonessential genes, respectively, 1611 genes are bootstrapped 1000 times, and a mean BF per guide is calculated. As previously mentioned, version 1 of BAGEL also implemented a limited minimum and maximum kernel density estimate of log2 foldchange for calculating BF to decrease the impact of outliers.

Initial analysis of our custom library screens with 50 core essential and 50 nonessential genes, using version 1 of BAGEL yielded irregular results despite good separation of essential and non-essential genes. As an example, the results from Day 12 of PATC69 are shown below. In this screen, there was good separation of the core essential and non-essential gene distribution (Figure 1). Despite the early time-point, Day 12, the core essential genes demonstrate a clear negative foldchange, while the non-essential genes show a distribution centered at or around 0. Of note in the screen is an outlier in the non-essential distribution, indicated by the arrow in Figure 1.

The impact of this single outlier gene present in the nonessential gene set adversely affects the precision-recall of the BF in a PATC69 screen at Day 12 (Figure 1). The Bayes Factor distributions of these genes is demonstrated in Figure 2, where the same outlier gene in the nonessential set outperforms every other gene, despite most of
the essential genes having a lower log2 foldchange. This is an instance of the impact of the upper and lower bounds on the Bayes Factor distributions of another PATC69 screen after on Day 12 despite clear separation of core-essentials and non-essentials is demonstrated in Figure 1. This highlights the need for analysis strategies more suitable for correctly modeling and fitting custom library datasets.

To address this, we developed **Low-Fat BAGEL**, an adapted framework of the updated BAGEL algorithm to more accurately analyze small screens with limited training sets. In Low-Fat BAGEL, all of the analysis is focused on each sgRNA as an independent data point rather than on gene level. Only 100 permutations are performed on the individual guides of the reference set, rather than on the level of each gene. In the case of our PATC69 and PATC124 screens, 500 core-essential and nonessential guides are bootstrapped across 100 permutations, and the resulting BF for each guide computed is an average of all permutations. An aggregate BF for each gene is then obtained by summing the BFs of individual guides. Log2 fold-change was calculated on a guide level by comparing each time point to the reference time point for each model. Screen quality and efficient drop out was assessed by comparing the log density ratio log2 foldchange of core essential versus non-essential guides.

The BAGEL and BAGELv2 calculate a Bayes Factor (BF) as a log likelihood of gene essentiality trained on gold-standard core essential and non-essential genes\textsuperscript{75,76}.

\[ BF = \frac{\Pr(D \mid \text{essential})}{\Pr(D \mid \text{nonessential})} = \frac{\int \Pr(D \mid k, \text{essential}) \Pr(k \mid \text{essential}) \, dk}{\int \Pr(D \mid k, \text{nonessential}) \Pr(k \mid \text{nonessential}) \, dk} \]
In Low-Fat BAGEL, a random bootstrap of core essential and non-essential guides is used as the training set to generate log density curves of the essential and non-essential fold-change distribution. This distribution is used to generate a linear regression model (k) of the fold-change to control for the influence of outliers. The linear interpolation method replaces the maximum and minimum kde for the essential and non-essential distribution implemented in BAGEL v1. This linear interpolation model is also implemented in BAGEL v2.

Over 100 permutations of the training set, Low-Fat BAGEL calculates the BF of all guides in the library, by comparing the log2 fold-change of each guide (D) to the distribution of fold-changes of the essential and non-essential genes. The final BF of each individual guide is computed as the mean across 100 iterations, and a gene level BF was calculated as the sum of all guides for that gene. The impact of using Low-Fat BAGEL on the same screen (PATC69, Day 12) is demonstrated in Figure 3. Precision-Recall curves are a good measure of the success of prediction of classes. Precision, also called positive predictive value, is the fraction of the number of true positives over all the positive predictions. In this case, precision is the proportion of total core essential genes over the sum of both core and non-essential genes predicted to be essential. Recall, also known as sensitivity, is the proportion of positives that were correctly identified over the total population of true positives (true positives + false negatives). In the case of our screens, recall would be the total number of core essential genes correctly predicted to be essential over the total number of core essential genes in our library. The precision-recall curve provides a dynamic view of how these values change with varying cut offs of the BF to deem essentiality. The R package ‘ROCR’ was used to generate Precision-Recall curves.
of the essential and non-essential gene level BF	extsuperscript{83}. When we compare the Precision-Recall curves for PATC69, Day 12 between BAGELv1 and Low-Fat BAGEL, we are able to visually appreciate the improvement in performance and accuracy in modeling the smaller dataset and controlling for outliers (Figure 4).

With the final optimized Low-Fat BAGEL, we also compared its performance against JACKS, CERES and MaGeCK. This comparison was made with data from two of our models – PATC69 and PATC124, across all our time points. Each time-point contained three replicates, which were regarded individually to compare methods. To reduce the Precision-Recall curve to a single metric, we computed the F measure of the screen. The F measure is a harmonic mean of precision and recall. The nature of the precision-recall curve is a trade-off – as the precision of the screen nears a value of 1, recall will near a value of 0, and vice versa. Given this, the F measure, which identifies an optimal balance between precision and recall provides a value between 0 and 1, with higher values representing higher screen quality. Comparison to other genome-wide analytical methods, Low-Fat BAGEL analyses demonstrate better performance and a more accurate classification of essential and nonessential genes in our screens (Figure 5).
Figure 1: Distribution of the essential and non-essential gene set of the surfaceome library in PATC69, Day 12.
Kernel density estimate plots comparing the log2 fold-change distribution of all essential and non-essential guides (n=285, n=280 respectively) for PATC69 in vitro Day 12 screen demonstrating the differences in dropout between the two distributions. The arrow indicates the outlier guides in the non-essential distribution that demonstrate dropout in this context.
Figure 2: Bayes Factors of the core essential and non-essential genes in the PATC69 Day 12 screen derived from BAGEL v1.
Gene-wise BF of essential and non-essential genes only from PATC69 in vitro Day 12 screen calculated from BAGEL v1 designed for genome-wide CRISPR screens. Gene-wise BF is calculated as a sum of guide level BFs. The dotted line indicates a BF cutoff of 0. BFs above this value indicate that a gene is more likely to be essential than non-essential. The box highlights an outlier gene in the non-essential distribution (marked by the arrow in Figure 1) that scores higher than the true essential genes.
Figure 3: Bayes Factors of the core essential and non-essential genes in the PATC69 Day 12 screen derived from Low-Fat BAGEL.
Gene-wise Bayes Factors of essential and non-essential genes PATC69 in vitro Day 12 screen calculated from Low-fat BAGEL designed for custom library CRISPR screens. Gene-wise BF is calculated as a sum of guide-level BFs. The dotted line indicates a BF cutoff of 0. BFs above this value indicate that a gene is more likely to be essential than non-essential.
Figure 4: Precision-Recall curves comparing the performance of BAGEL v1 and Low-Fat BAGEL.

Precision-Recall curves of BFs of essential and non-essential genes from the PATC69 in vitro Day 12 screen comparing the performance of BAGEL v1 and Low-fat BAGEL. These curves were derived from the BFs computed in Figure 2 and Figure 3.
Figure 5: Comparison of F measures of known genome-wide CRISPR screen analytical methods to Low-Fat BAGEL across PATC69 and PATC124 in vitro and in vivo screens.
3.6: Low-Fat BAGEL in Surfaceome Screens

With the optimized method in Low-Fat BAGEL, we established an analytical protocol for our surfaceome screens. As mentioned, we used three available PDX lines (PATC69, PATC124 and PATC53) that were compatible with the CRISPR-based screening technology in vivo. We conducted each in vitro CRISPR screen as a time course, with three in vitro time-points and with a matched in vivo time-point matching the last in vitro time-point. Each time-point had three replicates each. The Day 0 Reference time-point had two replicates each. BFs were calculated for each replicate compared to the Reference time-point for each respective model. Results from the latest in vitro time-point, Day 30 and the in vivo time-point were considered to identify essential genes. Figures 6-8 illustrate the screen quality with the separation of core essential and non-essential genes across all replicates and time-points for PATC69, PATC124 and PATC53, along with the corresponding BFs of each screen per time-point calculated as the mean of all replicates.

In order to be able to compare the BFs across all our screens, we quantile normalized our BFs across the entire cohort. Quantile-normalized BFs (BF > 1) were utilized as our threshold for determining essential genes represented in the three models in vitro and in vivo. Dependencies among the three PDX lines, in both in vitro and in vivo conditions, were highly diverse, with little overlap. This is presented in Figure 9. Ribosome-associated RACK1, RPL30, and the MYC proto-oncogene were the only shared vulnerabilities identified in all in vivo and in vitro screening contexts. Figures 10-12 illustrate the diversity of the surfaceome genes that are essential in vivo in each context. For visualization, we are presenting a STRING PPI of the surfaceome genes that
showed essentiality in at least one model, with genes in red being essential in each model, and genes in blue not showing essentiality in that context. We integrated quantile-normalized BFs with the STRING PPI network used to build the sgRNA library and generated dependency networks for each model. Dependency networks were merged across all three PDX lines based on overlapping essentiality thresholds (BF >1) for in vivo and in vitro contexts.
Injection Point

In vitro

Non Essential

Essential

0.00
0.05
0.10

Density
Figure 6: Paired density plots and Bayes Factor estimates for PATC69.

(Right) Density plots of each time-point and replicate of PATC69. Log2 foldchange is calculated from the Day 0 reference time-point. The injection point time-point represents the population of cells that was implanted *in vivo*, at a point before we see separation of essential and non-essential guides. Density plots in blue represent the non-essential genes in the surfaceome library that have a distribution centered at or around 0. Density plots in red represent the essential genes in the surfaceome library, which have an increasingly negative log2 foldchange distribution in later time-points.

(Left) BF plots of each time-point averaged across replicates of PATC69. BF plots in blue represent the non-essential genes in the surfaceome library that have a tightly distributed and highly negative BF. BF plots in red represent the essential genes in the surfaceome library, which have a distribution with highly positive BFs, indicating increasing confidence in essentiality.
Injection Point

in-vitro1

in-vitro2

in-vivo

Non Essential
Essential

density

0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05
1.10
1.15
1.20
1.25
1.30
1.35
1.40
1.45
1.50
1.55
1.60
1.65
1.70
1.75
1.80
1.85
1.90
1.95
2.00
2.05
2.10
2.15
2.20
2.25
2.30
2.35
2.40
2.45
2.50
2.55
2.60
2.65
2.70
2.75
2.80
2.85
2.90
2.95
3.00
3.05
3.10
3.15
3.20
3.25
3.30
3.35
3.40
3.45
3.50
3.55
3.60
3.65
3.70
3.75
3.80
3.85
3.90
3.95
4.00
4.05
4.10
4.15
4.20
4.25
4.30
4.35
4.40
4.45
4.50
4.55
4.60
4.65
4.70
4.75
4.80
4.85
4.90
4.95
5.00
Figure 7: Paired density plots and Bayes Factor estimates for PATC124.
(Right) Density plots of each time-point and replicate of PATC124. Log2 foldchange is calculated from the Day 0 reference time-point. The injection point time-point represents the population of cells that was implanted in vivo, at a point before we see separation of essential and non-essential guides. Density plots in blue represent the non-essential genes in the surfaceome library that have a distribution centered at or around 0. Density plots in red represent the essential genes in the surfaceome library, which show a slight negative log2 foldchange distribution, but overlap with the non-essential distribution.
(Left) BF plots of each time-point averaged across replicates of PATC124. BF plots in blue represent the non-essential genes in the surfaceome library that have a moderately negative BF. BF plots in red represent the essential genes in the surfaceome library, which have a distribution with positive BFs, and clear separation from the non-essential distribution, indicating accuracy in identifying essentiality.
Injection Point

in-vitro1

in-vitro2

in-vitro3

in-vivo

Non Essential

Essential
Figure 8: Paired density plots and Bayes Factor estimates for PATC53.
(Right) Density plots of each time-point and replicate of PATC53. Log2 foldchange is calculated from the Day 0 reference time-point. The injection point time-point represents the population of cells that was implanted in vivo, at a point before we see separation of essential and non-essential guides. Density plots in blue represent the non-essential genes in the surfaceome library that have a distribution centered at or around 0. Density plots in red represent the essential genes in the surfaceome library, which a show a slight negative log2 foldchange distribution, but overlap with the non-essential distribution.
(Left) BF plots of each time-point averaged across replicates of PATC53. BF plots in blue represent the non-essential genes in the surfaceome library that have a moderately negative BF. BF plots in red represent the essential genes in the surfaceome library, which have a distribution with positive BFs, and clear separation from the non-essential distribution, indicating accuracy in identifying essentiality.
Figure 9: Comparing essential genes across models in vitro (top) and in vivo (bottom).
(top) Venn diagram of functional targets derived from *in vitro*, orthotopically implanted, CRISPR screening of PDX lines using a quantile-normalized BF > 1, with six total genes showing *in vitro* essentiality across all three models.
(bottom) Venn diagram of functional targets derived from *in vivo*, orthotopically implanted, CRISPR screening of PDX lines using a quantile-normalized BF > 1, with eight total genes showing *in vivo* essentiality across all three models.
Figure 10: CRISPR screening results from PATC69.

*In vivo* CRISPR screening results from PATC69 overlaid onto a merged STRING PPI force-directed diagram. The BF of each gene indicates the degree of vulnerability of the gene, with BF > 1 indicating an essential gene.
Figure 11: CRISPR screening results from PATC124. 
*In vivo* CRISPR screening results from PATC124 overlaid onto a merged STRING PPI force-directed diagram. The BF of each gene indicates the degree of vulnerability of the gene, with BF > 1 indicating an essential gene.
Figure 12: CRISPR screening results from PATC53. 
In vivo CRISPR screening results from PATC53 overlaid onto a merged STRING PPI force-directed diagram. The BF of each gene indicates the degree of vulnerability of the gene, with BF > 1 indicating an essential gene.
3.7: Discussion

Here, we applied a custom functional genomics approach for screening PDAC PDX lines in parallel *in vivo* and *in vitro* screens. We designed small, customized CRISPR libraries were used to ensure maintenance of library complexity *in vivo*, Data from these sgRNA screens was not amenable to analysis using current analytical frameworks, which are designed for genome-wide screens and are not tailored for the small training sets of our custom library. To address this, we adapted the BAGEL framework to create Low-Fat BAGEL. We demonstrated that Low-Fat vastly improves the ability to identify essential genes in the context of our PDX models.

With the advent of CRISPR screening technology to answer more and more targeted and nuanced biological questions, the popularity of targeted CRISPR libraries continues to grow. As experimental methodologies and protocols enable more in-depth *in vivo* modeling, as well as other platforms such as T-cell screening, organoids, etc., having an equally robust analytical platform is equally important. Low-Fat BAGEL is able to address this need. Currently, we have released Low-Fat BAGEL as a stand-alone Python based tool, and is also available as an option for “small library screens” within the BAGEL v2 platform.

Using Low-Fat BAGEL, we can the minimum threshold for the number of core essential and non-essential genes required to accurately calculate the likelihood of a given gene being essential. For instance, using the large number of publicly available genome-wide screens of varying quality, we can down-sample the number of control genes needed in a library to recapitulate genome-wide results, allowing for further optimization of targeted libraries.
Chapter 4: Quantifying PDAC Diversity with Biological Networks

This work is based on efforts outlined in our recent BioRxiv submission “Diversity Across the Pancreatic Ductal Adenocarcinoma Disease Spectrum Revealed by Network-Anchored Functional Genomics” by Rose and Srinivasan, et al. (2020), DOI: https://doi.org/10.1101/2020.09.17.302034. As co-first author, Dr. Johnathon Rose provided substantial experimental support for this effort, enabling our findings. The manuscript is currently under review.

4.1: Patient Derived Xenograft Models

PDAC is a disease characterized by its dense desmoplastic stroma and genetic alterations that lead to disease development and progression. Owing to this, researchers are focusing on ways to model the tumor beyond using traditional cell lines to better capture and recreate the complex factors involved in tumor development and progression. Transgenic mouse models, notably KRAS mutant genetically engineered mouse models, serve as good model systems to study PDAC disease development and treatment response. However, studying the complex cascade of genetic alterations beyond that of KRAS, remains challenging in this setting. One strategy to study late stage and metastatic PDAC tumors is to directly implant a patient tumor sample into an immune-deficient mouse, or Patient Derived Xenografts (PDXs). Both directly and indirectly implanted PDX’s utilize clinical patient samples retain signaling resulting from tumor/microenvironmental interaction, are capable of developing metastases, and at least partially recapitulate the heterogeneity evidenced in patients. An additional benefit of PDXs is the ability to delineate molecular signaling innately derived from human tumor cells, vs. the murine stromal component (as reviewed by Garcia, et al).
In collaboration with Dr. Michael Kim and Dr. Scott Lowe, the IACS/TRACTION group has established a growing list of PDAC PDX tumors that have been extensively characterized for use in preclinical drug testing. To assess the diversity of transcriptomic signatures among PDAC tumors, we utilized 48 of these early passage patient-derived PDAC xenografts, which maintain the cellular heterogeneity of tumor lesions while reducing the contribution of the stromal components prevalent in these tumors. Clinical data corresponding to the patients these PDXs were derived from are presented in Table 1.

We performed RNA sequencing and whole exome sequencing (WES) on the 48 PDX models. First, we sought to quantify our models along previously reported heterogeneity within PDAC. The “oncoplots” feature within the R package maftools was utilized to visualize the mutational spectrum across the genes identified as relevantly mutated by the PDAC TCGA paper. We found that the general mutational spectrum of our tumors matches the frequency of mutations previously described (Figure 13). In our cohort, 93% of models harbored a KRAS mutation, followed by 84% of patients with TP53 mutations, and 27% of patients with CDKN2A and SMAD4 mutations each. Interestingly, we found a higher than previously reported mutation rate in ARID1A (31%), ATM (29%) and GNAS (29%).

As previously described, there are currently several subtyping methodologies described within PDAC, with the general consensus of the field agreeing on two broader subtypes “Classical” and “Basal-like”. This two-subtype classification was described by Moffitt et al using computationally micro-dissected patient and PDX models. The Moffitt classification method is based on a set of 21 classical subtype genes and 25 basal-like
genes. We assigned the Moffitt classification of classical or basal to our 48 PDX models using log normalized and scaled counts of 21 classical and 25 basal genes for PDX PDAC model classification described by consensus clustering into two groups using the R package ConsensusClusterPlus\(^{88}\). The clusters were manually assigned as classical and basal based on high expression of each group of genes. The breakdown of classical and basal-like models within our cohort is presented in Figure 14. We captured a level of transcriptional diversity consistent with the previously defined Moffitt classification status and distribution of classical and basal-like pancreatic adenocarcinoma. Within the tumors categorized as classical and basal-like, only a small proportion of tumors exclusively express markers that are either classical or basal-like. The heatmap in Figure 14 shows that a sizable portion of tumors express some of the marker genes from the other subtype, or low expression of markers from both subtypes. Consensus clustering requires the user to specify the number of clusters, and in this case, assigning them as classical or basal-like binarizes tumors even though these subtypes might biologically present as a spectrum rather than a category. This “spectrum” is partially addressed with the more recent Toronto classification, that identifies a “hybrid” group that expresses both classical and basal-like genes, but is still a categorical classification\(^{42}\).
Table 1: Clinical features of patients corresponding to the PDAC PDX models used in this study
<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Date of Birth</th>
<th>Address</th>
<th>Occupation</th>
<th>Education</th>
<th>Marital Status</th>
<th>Relationship</th>
<th>City</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>Male</td>
<td>01/01/1980</td>
<td>123 Main St, Anytown, USA</td>
<td>Teacher</td>
<td>B.S. Education</td>
<td>Married</td>
<td>Father</td>
<td>New York</td>
<td>NY</td>
</tr>
<tr>
<td>Jane Smith</td>
<td>Female</td>
<td>02/02/1985</td>
<td>456 Oak Ave, Anytown, USA</td>
<td>Nurse</td>
<td>B.S. Nursing</td>
<td>Single</td>
<td>Daughter</td>
<td>Chicago</td>
<td>IL</td>
</tr>
<tr>
<td>Michael Johnson</td>
<td>Male</td>
<td>03/03/1990</td>
<td>789 Pine Rd, Anytown, USA</td>
<td>Engineer</td>
<td>M.S. Engineering</td>
<td>Married</td>
<td>Husband</td>
<td>Los Angeles</td>
<td>CA</td>
</tr>
<tr>
<td>Sarah Lee</td>
<td>Female</td>
<td>04/04/1987</td>
<td>101 Elm St, Anytown, USA</td>
<td>Lawyer</td>
<td>J.D. Law</td>
<td>Divorced</td>
<td>Daughter</td>
<td>San Francisco</td>
<td>CA</td>
</tr>
<tr>
<td>David Kim</td>
<td>Male</td>
<td>05/05/1982</td>
<td>222 Cedar Ln, Anytown, USA</td>
<td>Doctor</td>
<td>M.D. Medicine</td>
<td>Single</td>
<td>Son</td>
<td>Houston</td>
<td>TX</td>
</tr>
</tbody>
</table>
Figure 13: Mutational diversity of the PDAC PDX models included in this study. Oncoplot of common PDAC-associated mutations in the 48-model PDX cohort displays mutation frequencies similar to previous publications.
Figure 14: Moffitt Classification of PDAC PDX models.
Moffitt classification of PDAC PDX models (n=48) using the PDX-specific gene set. Consensus clustering of normalized gene expression was performed to classify models into classical (blue top annotation) and basal-like (yellow top annotation).
4.2: Utility of Co-expression Analysis in Large Datasets

Our aim in this section was to develop a comprehensive approach to quantify and transcriptional diversity arising from PDAC tumor cells. Methods such as differential gene expression face the disadvantage of requiring a comparison group, typically normal tissue, which is not easily obtained in large numbers. In addition, differential expression analysis identifies genes with concordantly dysregulated signaling between two groups, making interrogation of heterogeneity between models challenging. Comparing tumors vs normal samples would capture the general transcriptional differences defining PDAC, while comparing tumor models or conditions to each other would highlight nuanced changes between the two. However, neither of these methods would capture the intrinsic diversity across a large cohort of PDAC tumors.

An alternative to these approaches, is co-expression analysis to capture the transcriptional diversity of a cohort of models. Co-expression analysis can be a powerful tool to explore candidate genes within the context of the entire genome, and also characterize and interrogate specific biological processes and functions. Specifically, constructing a co-expression network is a powerful integration of large transcriptomic datasets\textsuperscript{89-92}. Characterization of modules, or groups of genes that are co-expressed across a large cohort. This can serve as a platform for evaluating specific biological functions, mine for future hypothesis, contextualize “your favorite gene” within the behavior of a large network of genes, and predict the function of genes with previously unknown functions, to name a few.

Global pairwise co-expression networks that are constructed as undirected networks, while computationally intensive, provide a comprehensive view of
transcriptional heterogeneity파터, 한바. Undirected global networks are based on pairwise relationships between all genes in a cohort, pruned based on a determined threshold for inclusion. Identification of highly connected genes into clusters that are further annotated can aid in the characterization of biological functions. Interconnectivity within each cluster can inform on the core module of genes that might be implicated in regulating a specific biological function, whereas high connectivity outside a cluster can inform on higher order relationships between different biological functions. The basic organization of a network consists of a node – individual units of information, in this case, a gene. These nodes are connected to each other with edges or links, which in this case, would be a co-efficient of co-expression. These nodes and edges together form a network, or a graph.

To associate genetic drivers with clinically predictive transcriptomic signatures, we developed a dedicated PDAC co-expression network that we derived from our curated cohort of 48 PDXs. Co-expression of a set of genes is indicative of genes controlled by the same regulatory program, belong to the same signaling pathway or have similar functions파터, 한바. By developing a comprehensive genome-wide co-expression network, we are able to identify large sets of genes that are co-regulated across a large set of PDAC models.

4.3: Construction of the PDAC Co-Expression Network

To identify patterns of co-expression between genes across our cohort, we first selected genes with dynamic expression. Genes with stable high or low expression across all models would not be informative to study trends of co-expression. With our cohort of 48 models, we retained 17168 genes with raw read counts in at least 80% of all samples. Following this, we computed gene-wise median absolute deviation across
samples from the normalized read counts. Median absolute deviation provides a measure of dispersion of the data that is more robust to outliers than measures such as standard deviation. With median absolute deviation, 8505 genes exceeding the 40th percentile were filtered to consider highly variable genes. Spearman correlations were calculated as a metric of pairwise gene co-expression across the 8505 highly variable genes. Spearman's rank correlation provides a nonparametric measure to assess the linear relationship between two variables. Spearman’s rho provides a more conservative estimate without assuming the normality of the distribution of every one of the 8000+ genes being considered. This gave over 72 million pairwise correlations. In order to prune our network, and do so while maximizing the likelihood of retaining interpretable gene-pairs, we limited our dataset to positively correlated gene pairs. Using an adjusted p value of 0.05, 366,836 correlations passed this threshold.

In order to prune the network to prioritize biologically relevant gene pairs, a Bayesian framework developed by Yang, et al, Log Likelihood Score (LLS), was applied. This paper curated a “positive gold standard” list of biologically relevant gene pairs and “negative gold standard” gene pairs with no known functional annotations. As with a standard Bayesian formula, the LLS is a likelihood ratio of how likely any given Spearman rho value between two genes is likely to represent biological relevance over the rho value representing no biological relevance. This is calculated using the distribution of spearman correlations of gene pairs from the positive gold standard set and the correlations of gene pairs from the negative gold standard as posterior probabilities. The resulting score is the likelihood of any given gene pair belonging to the biologically relevant group, with positive score indicating higher likelihood of biological relevance, with higher scores indicating
higher confidence. COEXPEDIA, a co-expression database employing the LLS method to prune edges, uses an LLS score of 1 to identify gene pairs of high predicted biological relevance. Here, we used an extremely stringent cut-off of a log likelihood score of 2.5 to identify gene pairs to be included in the network. This resulted in the PDAC co-expression network with a total 103,000 correlations of 7,828 genes, of which all but 75 genes had an adjusted \( p \) value less than 0.05.

### 4.4: Clustering and GO Annotation of the PDAC Co-Expression Network

Our next step was to identify clusters of co-expressed genes within the PDAC Co-Expression Network. For this, we utilized a tool called InfoMap\(^96\). It incorporates MapEquation\(^96\), a community detection algorithm for large networks, which was used to cluster the network. All gene pairs were inputted into Infomap, and three hierarchical tiers of clusters were produced. In order to assign clusters, if the third tier contained more than 50 genes, it was assigned to an individual cluster. In the event that the third tier contained less than 50 genes, it was folded into the second tier, all of which formed a cluster. This was the case with cluster 29-31, which are larger encompassing clusters. This resulted in a total of 31 clusters, which were further genome ontology (GO) annotated\(^97\). For the resulting 31 clusters, the R package GOseq\(^98\) was used to conduct hyper-enrichment analysis of Gene Ontology Biological Processes pathways on each cluster. The R package “revigo”\(^99\) was used to prioritize and visualize GO pathways to represent their hierarchical class. The resulting PDAC co-expression network is presented in Figure 15 along with its GO annotated pathways. Revigo annotated pathway labels are presented in Figure 16.
We measured PDAC diversity by applying a dimensional reduction approach on our 31 defined clusters. Specifically, we quantified the mean expression of each cluster, or centroid score, on a tumor-by-tumor basis, for all 48 PDX models in the PDAC cohort. To generate a cluster level enrichment score, we calculated a centroid score per cluster by taking the mean log-normalized expression of all genes in each cluster for each sample. A comprehensive overview of the 31 centroid scores across the PDX models was generated using the R package ComplexHeatmap\textsuperscript{100}. 
Figure 15: Network hairball of the PDAC Co-expression network.
Force-directed layout of the PCEN. Gene clusters, defined by Gene Ontology (GO), visually highlight nodes with a minimum of 12 edges.
Figure 16: Treemaps of Cluster by Cluster GO Annotations within the PDAC Co-expression Network.
Figure 17: Cluster centroids derived from the PDAC co-expression network and the associated correlations between clusters.
(Top) Heatmap depicting the cluster centroid scores of each of the 31 clusters across the 48 PDX models. The centroid is computed as the mean normalized expression of all genes per cluster. The column annotation above presents the Moffitt classification of the PDX models, with the blue annotation corresponding to classical models and orange annotation corresponding to basal models.
(Bottom) Heat map depicting Pearson positive (red) and negative (blue) correlations between cluster centroid scores across 48 PDX models. Starred correlations represent an adjusted P value < 0.05.
4.5: Associated Mutations with Cluster Enrichment

To determine if the mutational background of a model was significantly associated with cluster enrichment across the PDX cohort, we applied UNCOVER\textsuperscript{101}, a method to identify complementary patterns of mutation enrichment across groups using a filtered set of “high impact” mutations and the cluster enrichment centroid scores per model. The filtered set of mutations was generated by identifying canonical mutations with gnomAD\textsuperscript{102} allele frequency less than 1%, “moderate” and “high” impact, and limited to non-intronic/non-coding and synonymous mutations. Interestingly, we identified a limited cluster-mutation association between \textit{ACVR2A}, \textit{RREB1}, and \textit{MARK2} mutations and tumors with significant enrichment in Cluster 21 (Table 2). PDAC-associated loss-of-function mutations in \textit{RREB1}, which encodes a zinc finger transcription factor that binds to RAS-responsive elements, have been reported in the TCGA PDAC cohort. \textit{RREB1} is a positive regulator of the zinc transporter, ZIP3, with loss-of-function playing a potential role in limiting zinc uptake and shielding developing tumors from the cytotoxic effects of high cellular zinc concentrations\textsuperscript{103}. \textit{RREB1} has also been described as a KRAS-regulated SMAD co-factor involved in driving the expression of epithelial-to-mesenchymal (EMT) transcription factors\textsuperscript{104}. While the significance of mutations in the serine-threonine kinases, \textit{MARK2} and \textit{ACVR2A}, is relatively poorly understood in PDAC, dysregulation in these genes could implicate the regulation of epithelial polarity and downstream SMAD-associated signaling, respectively\textsuperscript{105-107}. Associating co-expressed and annotated gene clusters with these less frequent mutations in PDAC could aid in illuminating the molecular signaling, and potential therapeutic avenues, underlying these genomic alterations.
Table 2: Results from the UNCOVER analysis reveal complementary gene mutations associated with cluster centroid-based enrichment

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mutation 1</th>
<th>Mutation 2</th>
<th>Mutation 3</th>
<th>OFC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster_1</td>
<td>SETD2</td>
<td>PLXNB2</td>
<td>SIN3B</td>
<td>3.526925</td>
<td>0.752475</td>
</tr>
<tr>
<td>cluster_2</td>
<td>KDM6A</td>
<td>BCORL1</td>
<td>BRCA2</td>
<td>5.498805</td>
<td>0.277228</td>
</tr>
<tr>
<td>cluster_3</td>
<td>KDM6A</td>
<td>BCORL1</td>
<td>SDK1</td>
<td>4.765432</td>
<td>0.485149</td>
</tr>
<tr>
<td>cluster_4</td>
<td>NF1</td>
<td>BRCA2</td>
<td>RREB1</td>
<td>5.238452</td>
<td>0.326733</td>
</tr>
<tr>
<td>cluster_5</td>
<td>RREB1</td>
<td>MARK2</td>
<td>ACVR2A</td>
<td>6.091605</td>
<td>0.217822</td>
</tr>
<tr>
<td>cluster_6</td>
<td>SDK1</td>
<td>NF1</td>
<td>PALB2</td>
<td>5.14112</td>
<td>0.267327</td>
</tr>
<tr>
<td>cluster_7</td>
<td>RREB1</td>
<td>RNF43</td>
<td>MARK2</td>
<td>5.484781</td>
<td>0.287129</td>
</tr>
<tr>
<td>cluster_8</td>
<td>SDK1</td>
<td>ACVR2A</td>
<td>MARK2</td>
<td>5.674092</td>
<td>0.168317</td>
</tr>
<tr>
<td>cluster_9</td>
<td>RREB1</td>
<td>BRCA2</td>
<td>RNF43</td>
<td>5.141292</td>
<td>0.445545</td>
</tr>
<tr>
<td>cluster_10</td>
<td>PLXNB2</td>
<td>RNF43</td>
<td>RREB1</td>
<td>5.237245</td>
<td>0.376238</td>
</tr>
<tr>
<td>cluster_11</td>
<td>SPTB</td>
<td>PALB2</td>
<td>SDK1</td>
<td>5.575874</td>
<td>0.089109</td>
</tr>
<tr>
<td>cluster_12</td>
<td>SPTB</td>
<td>ITPR3</td>
<td>BRCA2</td>
<td>4.064193</td>
<td>0.782178</td>
</tr>
<tr>
<td>cluster_13</td>
<td>PALB2</td>
<td>ITPR3</td>
<td>SIN3B</td>
<td>4.180341</td>
<td>0.653465</td>
</tr>
<tr>
<td>cluster_14</td>
<td>PLXNB2</td>
<td>PALB2</td>
<td>ITPR3</td>
<td>4.296772</td>
<td>0.712871</td>
</tr>
<tr>
<td>cluster_15</td>
<td>SDK1</td>
<td>PLXNB2</td>
<td>RREB1</td>
<td>4.013054</td>
<td>0.792079</td>
</tr>
<tr>
<td>cluster_16</td>
<td>ITPR3</td>
<td>SETD2</td>
<td>SMARCA4</td>
<td>3.520863</td>
<td>0.980198</td>
</tr>
<tr>
<td>cluster_17</td>
<td>PALB2</td>
<td>TLE4</td>
<td>ITPR3</td>
<td>4.769161</td>
<td>0.574257</td>
</tr>
<tr>
<td>cluster_18</td>
<td>ACVR2A</td>
<td>ITPR3</td>
<td>BRCA1</td>
<td>5.513736</td>
<td>0.415842</td>
</tr>
<tr>
<td>cluster_19</td>
<td>PALB2</td>
<td>CALD1</td>
<td>SMARCA4</td>
<td>6.269093</td>
<td>0.287129</td>
</tr>
<tr>
<td>cluster_20</td>
<td>SDK1</td>
<td>CALD1</td>
<td>AKT2</td>
<td>4.720414</td>
<td>0.653465</td>
</tr>
<tr>
<td>cluster_21</td>
<td>ACVR2A</td>
<td>RREB1</td>
<td>MARK2</td>
<td>7.206207</td>
<td>0.029703</td>
</tr>
<tr>
<td>cluster_22</td>
<td>SDK1</td>
<td>PALB2</td>
<td>CALD1</td>
<td>4.122304</td>
<td>0.871287</td>
</tr>
<tr>
<td>cluster_23</td>
<td>CALD1</td>
<td>SMARCA4</td>
<td>PALB2</td>
<td>4.131839</td>
<td>0.920792</td>
</tr>
<tr>
<td>cluster_24</td>
<td>ACVR2A</td>
<td>TP53BP2</td>
<td>CALD1</td>
<td>4.385642</td>
<td>0.831683</td>
</tr>
<tr>
<td>cluster_25</td>
<td>PALB2</td>
<td>SETD2</td>
<td>CALD1</td>
<td>4.767625</td>
<td>0.524752</td>
</tr>
<tr>
<td>cluster_26</td>
<td>SMARCA4</td>
<td>RREB1</td>
<td>PALB2</td>
<td>4.372286</td>
<td>0.80198</td>
</tr>
<tr>
<td>cluster_27</td>
<td>CALD1</td>
<td>AKT2</td>
<td>BCORL1</td>
<td>4.055904</td>
<td>0.792079</td>
</tr>
<tr>
<td>cluster_28</td>
<td>AKT2</td>
<td>PLXNB2</td>
<td>CALD1</td>
<td>4.196129</td>
<td>0.772277</td>
</tr>
<tr>
<td>cluster_29</td>
<td>RNF43</td>
<td>PLXNB2</td>
<td>PALB2</td>
<td>4.640496</td>
<td>0.366337</td>
</tr>
<tr>
<td>cluster_30</td>
<td>PALB2</td>
<td>CALD1</td>
<td>PLXNB2</td>
<td>5.071861</td>
<td>0.60396</td>
</tr>
<tr>
<td>cluster_31</td>
<td>PALB2</td>
<td>SMARCA4</td>
<td>CALD1</td>
<td>4.305349</td>
<td>0.881188</td>
</tr>
</tbody>
</table>
4.6: Prominent Anticorrelated Cluster Signatures

We observed general patterns of anti-correlative clusters across the PDAC PDX cohort that were reflected in cluster positioning and subsequent cross-cluster connectivity. The most significant anticorrelated signatures were identified in clusters predominantly localized to adjacent ends of the force-directed layout (Figure 15), and the non-overlap of these adjacent anti-correlated cluster trends implicates multiple distinct molecular signaling contexts that represent the immense diversity across the PDAC disease spectrum. The top anti-correlative signatures were quantified between two opposing clusters: Cluster 1 vs. Cluster 23, respectively enriched for lipid metabolism vs. cell development; and Cluster 2 vs. Cluster 13, respectively enriched for Golgi-vesicle transport vs. nuclear-transcribed mRNA catabolism (Figure 18). Additionally, the PDAC co-expression network also highlighted GO annotated clusters critical for proliferating tumor cells; specifically, Cluster 15, cell cycle, and Cluster 16, mitochondrial transport and organization (Figure 18). These findings reveal that, along with 31 unique gene clusters, two distinct anti-correlated cluster signatures contribute to PDAC tumor diversity and have the potential to provide context for tumor cell-intrinsic vulnerabilities.
Figure 18: Anticorrelations of clusters in the PDAC co-expression network with the corresponding GO annotations.

(Top) Force-directed layout of the PCEN (top) and matching GO hierarchical treemaps (bottom) of prominent anti-correlative cluster centroid trends across the PDAC co-expression network highlighting Cluster 1 (Lipid metabolism) vs. Cluster 23 (Cell development).

(Bottom) Force-directed layout of the PCEN (top) and matching GO hierarchical treemaps (bottom) of prominent anti-correlative cluster centroid trends across PDAC co-expression network highlighting Cluster 2 (Golgi vesicle transport) vs. Cluster 13 (mRNA catabolism).
4.7: Cluster 1 and 23 Anticorrelation Recapitulates Moffitt Signatures

Upon observing the strong anti-correlative trend between Cluster 23 and Cluster 1, we evaluated the Cluster 1 centroid scores among the entire PDAC PDX cohort (Figure 19). Interestingly, Cluster 1 unbiasedly localized the entire set of 21 classical signature genes from the Moffitt classification. The Cluster 1 centroid scores across our PDX cohort showed low variance among Moffitt-defined classical models but a wider range and significant depletion of Cluster 1 gene expression \( p = 3.67 \times 10^{-6} \) was observed in basal-like models (Figure 19). Visualized another way, we paired the centroid scores of Cluster 1 and Cluster 23 (C1vC23) and identified a near perfect anticorrelation between the two clusters across the PDX cohort. Integrating the Moffitt classifications with the C1vC23 enrichment anticorrelation demonstrates that models enriched in C1 and depleted in C23 are all classical, and conversely, models enriched in C23 and depleted in C1 are all basal-like. There is a sizable group of models with partial enrichment of both C1 and C23, that are equally assigned to either classical or basal-like groups. This is a similar trend previously observed with the Moffitt classification genes itself, with a proportion of samples showing partial enrichment of both classical and basal-like marker genes. Taken together, these trends suggest that the classical and basal-like subtypes of PDAC are two ends of a continuum, rather than two discrete groups. This suggests the existence of a transitionary state of tumors that harbor both classical and basal-like phenotypes.

Indeed, by applying K means clustering on only the Cluster 1 and Cluster 23 centroid scores, we separated the PDX cohort into three groups (Figure 20): 1) enrichment in Cluster 1 represented Moffitt-defined classical models, 2) enrichment in Cluster 23 represented Moffitt-defined basal-like models, and 3) models with partial
enrichment of both Clusters 1 and 23 (Figures 21), which we termed “quasi-basal”. Thus, our C1vC23 signature provides a continuous transcriptomic signature that expands on the original binary Moffitt classification and uncovers a transitionary quasi-basal phenotype of PDAC with molecular signatures falling along a continuum.
Figure 19: Module centroid scores of Cluster 1 between Moffitt defined classical and basal models. Violin plot of average Cluster 1 centroid score across Moffitt-defined classical and basal-like models. Box-whisker plots show median ± first and third quartiles. P values are derived from t-test (n = 48 PDX tumors).
Figure 20: K means clustering reveals optimal $k=3$ using the Cluster 1 and Cluster 23 centroid scores across the PDX models.
Figure 21: Prominent anti-correlated clusters recapitulate a granular spectrum of “classical”, “quasi-basal”, and “basal-like” models that are associated with clinical features.

(Right) K means clustering reveals optimal k=3 using the cluster 1 and cluster 23 centroid scores across the PDX models into co-expression network (CEN) derived classical, quasi-basal and basal models paired with Consensus Clustering (k=2) into Moffitt’s classical and basal subtypes. The heatmap demonstrates the strong anticorrelation in the cluster enrichment between cluster 1 and 23.

(Left) Pie chart comparison of clinical histology and recurrence of patients associated with PDX models across CEN subtypes. P values are derived from Chi Square test. The three pie charts correspond with the breakdown of models into the CEN derived classical, quasi-basal and basal subtypes.
To gain further insight into the molecular signatures driving PDAC subtypes, we conducted gene-set enrichment analysis\textsuperscript{110} to compare C1vC23 defined classical vs. basal-like models within our PDX cohort (excluding models described as quasi-basal). This highlighted that EMT signaling was significantly enriched among PDX models that fell into the C1vC23 definition of basal-like, whereas this gene set was depleted in classical models (Figure 3E). Enrichment of an EMT signature in basal-like PDX models supports the hypothesis that the basal-like subtype is strongly associated with tumors where a majority of tumor cells has at least partly undergone trans-differentiation towards a mesenchymal phenotype\textsuperscript{56,111}.

![Enrichment plot: HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION](image)

**Figure 22:** GSEA conducted on PDAC co-expression network-defined classical and basal-like models identifies EMT as a top pathway enriched in basal-like models FDR = 0.000.
4.8: Anticorrelated Signatures Recapitulated in TCGA Samples

Consistently, analysis of clinical and histological data for the PDX cohort demonstrated that basal-like tumors were uniformly associated with poor differentiation status and distant metastatic recurrence (Figure 3B). Next, using high-epithelial-content PDAC patient data from TCGA (30% - 80% epithelial cells), we confirmed the presence of the network-derived subtyping, again identifying a quasi-basal continuum. Further, we recapitulated the association between tumor histology and the network-derived subtype in the TCGA dataset, wherein more poorly differentiated tumors were classified as basal-like, with strong enrichment in Cluster 23 gene expression (Figure 3F). By quantitatively characterizing a quasi-basal population, the C1vC23 differential enables a more precise definition of the clinically relevant basal-like tumor cohort while also expanding on associated molecular dependencies that may be considered for therapeutic intervention. Moreover, C1vC23 differential defines a broader, cluster-level characterization of the range of molecular signaling contributing to diversity in PDAC, which can be used to granularly ascertain pathways that may be therapeutically targeted to exert anti-tumor effects across this classical, quasi-basal and basal-like tumor spectrum.
Figure 23: Prominent anti-correlated clusters identifying “classical”, “quasi-basal”, and “basal-like” recapitulated in TCGA.

[left panel] Pie chart comparison of tumor differentiation status in high-epithelial-content TCGA comparing the two group Moffitt subtype and the co-expression network anticorrelating cluster 1 and 23 signatures.

[right panel] Heat map of C1vC23 anti-correlated cluster signature differential, matching Moffitt classification and tumor grade on TCGA tumor samples (samples with ≥ 30% epithelial content) with models comparing PDAC co-expression network-derived classifications and Moffitt classifications (n=76 tumors).
4.9 Discussion

Using our large cohort of PDAC PDX models, we constructed a cell intrinsic PDAC co-expression network. We applied a Bayesian method, the Log Likelihood Score to prune our network to not just statistically significant gene-pairs, but also maximize the likelihood that a given gene-pair is biologically relevant. Using a large network community detection tool, we also identified 31 unique clusters, which we further GO annotated for biological function. Interestingly, we identified anticorrelated signatures across two clusters – cluster 1 and cluster 23. These anticorrelated signatures recapitulated the Moffitt classical/basal-like subtypes on a continuous spectrum. Further, we were able to identify a transitionary group of tumors that partially express both cluster 1 and cluster 23. Using the three-tier classification, we were able to provide improved resolution in clinical associations with subtypes and consistently identify basal tumors with poor differentiation status and distant metastases.

Here we are able to demonstrate that the Moffitt classical and basal subtypes are likely two ends of a subtype continuum, rather than discrete groups. Current computational methods aiming to classify PDAC tumors into subtypes, regardless of what the subtypes are, fall short of capturing the granular heterogeneity in tumors due to their categorical nature. Using the co-expression network in this context provides an avenue to capture this heterogeneity on a continuous scale, but also categorize these tumors for ease of further analyses, as we have shown above.
Chapter 5: Integration of the Co-Expression Network with Single Cell Transcriptomic Signatures

This work is based on efforts outlined in our recent BioRxiv submission “Diversity Across the Pancreatic Ductal Adenocarcinoma Disease Spectrum Revealed by Network-Anchored Functional Genomics” by Rose and Srinivasan, et al. (2020), DOI: https://doi.org/10.1101/2020.09.17.302034. As co-first author, Dr. Johnathon Rose provided substantial experimental support for this effort, enabling our findings. The manuscript is currently under review.

5.1: Interrogating the Quasi-Basal Signature Intratumorally

We have validated the presence of the transitionary, “quasi-basal” subtype in the bulk PDAC PDX and TCGA samples. This raises the question of the origin of this signature. Two possibilities exist:

1. The quasi-basal signature identified in bulk tumors is a computational artifact – it is an average transcriptional signature of competing classical and basal-like subclones within the tumor.

2. The quasi-basal signature is an additional transitionary state that exists and is quantifiable intratumorally.

The Toronto classification, part of the COMPASS Trial, also recently provided clarity on the presence of a hybrid subtype in their bulk sequencing of tumor samples. These tumors were classified as hybrid due to expressing both the classical and basal-like signatures they derived using NMF. Interrogation of the signature intratumorally confirmed that tumors express both classical and basal-like programs within the same tumor. This was segregated into different tumor cell subsets within the same tumor42.
The findings of the Toronto classification suggest that the quasi-basal signature is an average signature of classical and basal-like subclones. However, given that this methodology results in categorical subtypes, it is likely that it insufficiently captures the spectrum of enrichment from both subtype signatures\(^{42}\).

Here we hypothesized that the quasi-basal tumors we characterized in the bulk samples are a quantifiable state, that can be captured using a more continuous classification method, and not an artifact of two competing subclonal populations. We investigated this by applying the PDAC co-expression network anticorrelating C1 and C23 signatures towards single cell analysis, as described below.

### 5.2: Single Cell Samples Utilized

To investigate whether the quasi-basal signature identified in bulk tumor populations represents either a quantifiable cell state or a mean signature derived from competing subcellular populations, we conducted single-cell RNAseq (scRNAseq) on early-passage cell lines derived from PDX models PATC69 (quasi-basal), PATC124 (quasi-basal), and PATC53 (basal-like). Single-cell samples of PATC124 and PATC53 were courtesy of Dr. Andrea Viale’s lab and cultured, collected and sequenced by Dr. Chieh-Yuan Li for a prior publication\(^{112}\). Single-cell samples for PATC69 were cultured, collected and sequenced by Dr. Johnathon Rose. Additionally, in order to evaluate the utility of the network-defined tumor subtypes in a clinical setting, in which small numbers of cells and stromal components may influence molecular analyses, we evaluated seven patient core needle biopsies (CNBs; four primary tumors, as well as one each of liver, lung, and vaginal metastases) using the network-aided scRNAseq analysis\(^{55}\).
Specifically, network-aided analysis applied clusters identified in the PDAC co-expression network to address limitations in scRNAseq. This is due to the fact that genome-wide expression in individual cells is currently not feasible. On average, a single cell expresses only up to approximately 4000 genes. Given this, in order to apply our PDAC co-expression network to single data to capture cluster level centroids, we sought to adapt the method to apply to a single cell setting.

5.3: Cluster Signatures for Single Cells

Network-based normalization of cluster signatures amongst single cells was achieved by first identifying a subset of genes whose expression was strongly correlated with its own cluster assignment (r > 0.4). For the PDX models, using the subset of highly correlated genes per cluster, if more than 30% of the cluster was captured per single cell, we calculated a centroid score by taking the mean of normalized UMI count. For patient CNB samples, centroid was calculated for all single cells if more than 20% of the cluster was captured. The percentage of cluster genes expressed within each cell type was calculated by taking the mean number of genes with UMI count above 0 per cluster across all the single cells within each cell type. The co-expression network was used to identify cells expressing more than 30% (20% for CNBs) of any cluster, and a centroid score for that cluster was calculated using genes with expression highly correlated to the cluster enrichment (r > 0.4).

5.4: Single-cell Transcriptomic Profiles Recapitulate Granularity of PDAC Subtypes

We applied this method to the PDAC PDX derived cell lines from PATC69, PATC124 and PATC53. We assigned centroid scores across all 31 clusters from the
PDAC co-expression network on each single cell on a model by model basis (Figures 24-26). A small proportion of single cells across all three models were coded as “NA” if they did not express enough of the cluster genes as described above. Specifically, we considered cluster 1 and cluster 23 scores to evaluate if a) a dynamic range in the enrichment of each cluster exists, b) if the spectrum of enrichment of each cluster is continuous or discrete (high expression vs. no expression) and c) whether the anticorrelation of the two clusters is present on a single cell scale.

This approach successfully confirmed that similar to the bulk tumors, cluster 1 and cluster 23 showed dynamic, continuous expression across single cells, and displays the same anticorrelation. We computed a cluster differential score (cluster 1 – cluster 23) to quantify the spectrum of enrichment across cluster 1 and 23. Single cells were grouped into classical, quasi-basal and basal based on the cluster 1 and cluster 23 differential signature, classified as basal if the differential signature was < -0.1, quasi-basal if they are -0.1 to 0.1 and classical if they were > 0.1. The results for each of the three samples is presented in Figures 27-29.

As is evident, even on a single cell scale, we are able to capture single cells that: a) are enriched in either cluster 1 or 23 and depleted in the other, and b) show equal enrichment of cluster 1 and cluster 23. As a result, we demonstrate that the quasi-basal state identified in bulk tumors, also exists as a quantifiable state within single cells. Using the cluster 1 and 23 differential, the quasi-basal signature is quantifiable state in individual cells within in each PDX model, confirming that the bulk readout represented an average of the intratumoral spectrum of classical, quasi-basal, and basal-like sub-clonal populations, rather than competing classical and basal-like signatures (Figures 27-29).
Interestingly, quasi-basal single cell analysis also composed the largest proportion of cells across all three subtypes (Figures 27-29). While this is not surprising in PATC69 and PATC124, quasi-basal tumors in bulk PDX models, it is surprisingly still the case in PATC53, a basal-like tumor in bulk state (Figure 14, Figure 21). Indeed, PATC53 displayed a large proportion of cells with much higher enrichment of cluster 23 than the other two models, explaining its classification as a basal-like tumor (Figure 14, Figure 21). Similarly, PATC69, on a bulk scale showed slightly higher enrichment in cluster 1 than 23, despite being classified as quasi-basal (Figure 21). Accordingly, PATC69 displayed the largest proportion of single cells that were classified as classical along with higher expression of cluster 1 compared to the other two models (Figure 27). This is also demonstrated in the density plot with the distribution of single cells across the highly negative (enriched in cluster 23) to near 0 (partial enrichment in both clusters) and highly positive (enriched in cluster 1) cluster differential scores (Figure 30). This displays a statistically significant difference in this distribution across all three models (Kolmogorov Smirnov test, p < 2.2e-16) (Figure 30). Thus, in similar fashion to the COMPASS trial\textsuperscript{42}, we also validated that all three of the models encompass both classical and basal subtypes, while also demonstrating that quasi-basal cells co-expressing both programs exist intratumorally (Figure 27-30).

We also applied the Moffitt classical and basal-like classification towards the single cell analysis of PDX cell lines, using consensus clustering to all three models. Here, we assigned Moffitt subtypes to single cells using a method called Nearest Template Prediction (NTP)\textsuperscript{113}. Using a set of template genes, the previously defined classical and basal-like gene sets released by Moffitt, et al, the NTP algorithm makes a class prediction
using a sample’s gene expression based with an accompanying $p$ value as a confidence measure of the prediction. In the case of PATC69 and PATC53, we find that no single cell is classified into a subtype with $p < 0.05$. Within PATC124, the majority of single cells were classified under NTP – all as basal, including single cells that showed high enrichment of cluster 1 and low enrichment of cluster 23 (Figures 24-26).
Figure 24: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC69 single cells.

Centroid expression of 31 clusters computed from single cells that meet the quality control thresholds. Single cells in grey represent cells that did not meet this threshold. The top annotation bar represents the Moffitt subtype, calculated using NTP. Annotation bar in grey represents cells that were not classified at p < 0.05. The bottom annotation bar represents the subtype determined by the co-expression network (CEN), with blue representing the classical subtype, beige representing quasi-basal subtype, and orange representing the basal subtype.
Figure 25: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC124 single cells.
Centroid expression of 31 clusters computed from single cells that meet the quality control thresholds. Single cells in grey represent cells that did not meet this threshold. The top annotation bar represents the Moffitt subtype, calculated using NTP. Annotation bar in grey represents cells that were not classified at p < 0.05, and orange represents the basal subtype. The bottom annotation bar represents the subtype determined by the co-expression network (CEN), with blue representing the classical subtype, beige representing quasi-basal subtype, and orange representing the basal subtype.
Figure 26: Heatmap of centroid expression of all 31 clusters of the PDAC co-expression network in PATC53 single cells.
Centroid expression of 31 clusters computed from single cells that meet the quality control thresholds. Single cells in grey represent cells that did not meet this threshold. The top annotation bar represents the Moffitt subtype, calculated using NTP. Annotation bar in grey represents cells that were not classified at $p < 0.05$. The bottom annotation bar represents the subtype determined by the co-expression network (CEN), with blue representing the classical subtype, beige representing quasi-basal subtype, and orange representing the basal subtype.
Figure 27: Cluster 1 and 23 differential in PATC69 reveals quasi-basal single cells. (Top) UMAP of single-cell PATC69 (7,857 cells), cell line with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential (Bottom) Respective percentages of cells corresponding to each subtype where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells.
Figure 28: Cluster 1 and 23 differential in PATC124 reveals quasi-basal single cells.
(Top) UMAP of single-cell PATC124 (9,482 cells), cell line with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential
(Bottom) Respective percentages of cells corresponding to each subtype where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells.
Figure 29: Cluster 1 and 23 differential in PATC53 reveals quasi-basal single cells. (Top) UMAP of single-cell PATC124 (14,791 cells), cell line with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential. (Bottom) Respective percentages of cells corresponding to each subtype where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells.
Figure 30: Difference in the distribution of cluster 1 and cluster 23 enrichment across three PDAC PDX models.
Density histogram of the C1v23 signature differential distributions of PATC124, PATC53, and PATC69 PDX lines, with more positive cluster differential indicating enrichment in Cluster 1 and more negative cluster differential indicating Cluster 23 enrichment.
5.5: Application of the Co-Expression Network in Single-cell Transcriptomic Profiles from Clinical Samples

To evaluate the utility of the network-defined tumor subtypes in a clinical setting, in which small numbers of cells and stromal components may influence molecular analyses, we evaluated seven patient core needle biopsies (CNBs; four primary tumors, as well as one each of liver, lung, and vaginal metastases) using the network-aided scRNAseq analysis\(^{55}\) (Figure 31). We first defined the distribution of cell types amongst our samples (Figure 32) using previously annotated marker genes (Figure 33) (see methods).

Up until this point, owing to the use of our PDX models, we have not characterized the expression of the PDAC co-expression network where microenvironment is heavily present. Our aim with the clinical single cell expression from CNB samples is two-fold. By evaluating a clinical sample with single cell sequencing, we are able to a) Deconvolute the signature derived from the epithelial tumor cell compartment from the microenvironment to assess if the signal is truly tumor cell intrinsic and b) Evaluate the feasibility of capturing network cluster level expression and the granular distribution of the Moffitt subtypes.

To deconvolute the signature derived from the epithelial tumor from the microenvironment, we first divided our first annotated and grouped our cells by cell type, as described above. Then, on for all the cells from a single cell type, we computed the proportion of genes from each cluster of the co-expression network that showed any expression (UMI > 0). We characterized cluster representation across the tumor microenvironment by quantifying the mean proportion of genes of each cluster for each
cell type. This identified the epithelial component of the tumor as the primary contributor to the co-expression network cluster signatures (Figure 34).

Next, applying the same cluster centroid normalization method described above for CNBs, the majority of cells that met our quality control cutoff were epithelial cells, with very little representation from two primary tumor samples (Primary 1 and Primary 2) that contained little to no epithelial content (Figure 32). Finally, to determine where each cell type was represented on the classical to basal-like continuum, we assessed the C1vC23 cluster differential. We found that the C1vC23 classification is largely present within epithelial cells, with Uniform Manifold Approximation and Projection (UMAP) clusters of fibroblasts and endothelial cells only representing a minority of misclassified C1vC23 signatures within the representative multiregional tumor microenvironment (Figure 35). This finding confirms a lack of sample purity bias in the characterization of C1vC23 signatures in the TCGA PDAC samples (Figure 23) and supports the feasibility of applying network-based cluster characterization to bulk clinical samples, including CNBs.
Figure 31: Diagram of the locations and numbers of isolated patient core needle biopsy (CNB) samples used for single-cell RNA sequencing analysis. A total of 4 primary samples were obtained (3 from female patients, and 1 from a male patient), along with three metastatic samples (one liver and lung sample each from male patients, and one vaginal metastasis from a female patient).
Figure 32: UMAP outlining the tumor microenvironment components of multiple single-cell sequenced primary (n = 4) and metastatic (n = 3) CNB samples from PDAC patients.
UMAPs of 25,954 single cells separated by the seven patient core needle biopsies. P1-p4 represents primary tumor samples. LiM represents the liver metastasis sample, LuM represents the lung metastasis sample and VM represents the vaginal metastasis sample. Analysis was performed using Seurat 3.1 and clusters were identified using a resolution of 0.2.
Figure 33: Expression of known marker genes was used to identify cell type of each UMAP-identified cluster.
Figure 34: Representation of the PDAC co-expression network genes across cell types in PDAC patient core needle biopsy samples (n=7).
The mean percentage of cluster representation for each cluster in the PDAC co-expression network from the cellular constituents of the tumor microenvironment. Each dot per cell type represents individual clusters. Box plot center: mean; box: quartiles 1–3; whiskers: quartiles 1–3 ± 1.5 × IQR.
Figure 35: Combined CNB UMAP with an overlay of a PDAC co-expression differential signature of cluster 1 and cluster 23 expression. UMAP of seven single-cell CNB samples (25,954 cells), with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells, and grey represents cells that did not meet the minimum threshold of PDAC co-expression genes expressed to have a cluster centroid score.
5.6: Discussion

Previous reports of a hybrid subtype on the bulk scale have been classified as an average signature of competing classical and basal-like subtypes existing intratumorally\textsuperscript{42}. Here, we validated the quasi-basal subtype observed in bulk samples as a quantifiable state in single cells as well. The partial enrichment of cluster 1 and cluster 23 in the bulk tumors is not as a result of single cells separately expressing cluster 1 or 23 intratumorally. However, on a single cell scale, across three models, we demonstrate that single cells also show equal enrichment of cluster 1 and 23. Again, we demonstrate that the cluster 1 and 23 anticorrelation is a continuous spectrum in our PDAC PDX derived cell lines.

We also investigated the expression of the co-expression network clusters in clinical grade samples from primary and metastatic patients. Using clinical patient allows us to characterize how different components of the PDAC microenvironment influences the co-expression network signatures we observe. Using our network normalized cluster enrichment method described above, we were able to confirm that the majority of the signature arises from the tumor epithelial component, confirming that the co-expression network is highly cell intrinsic.

By analyzing scRNAseq data in the context of the co-expression network, we circumvented the technical issue of signal dropout by prioritizing large gene clusters to represent transcriptional diversity rather than single gene expression. Thus, the co-expression network serves as an additional resource for disease-specific single cell analysis.
Chapter 6: Intratumoral Validation of Functional Vulnerabilities with Co-Expression Network

This work is based on efforts outlined in our recent BioRxiv submission “Diversity Across the Pancreatic Ductal Adenocarcinoma Disease Spectrum Revealed by Network-Anchored Functional Genomics” by Rose and Srinivasan, et al. (2020), DOI: https://doi.org/10.1101/2020.09.17.302034. As co-first author, Dr. Johnathon Rose provided substantial experimental support for this effort, enabling our findings. The manuscript is currently under review.

6.1: Integration of CRISPR Identified Dependencies with the Co-Expression Network

The distinct gene clusters we identified in the PDAC co-expression network provide a means to deeply characterize the diversity of any PDAC model by considering correlations in disease-specific cluster enrichment patterns, a parallel strategy to current approaches that use consensus clustering to assign a tumor subtype based on refined PDAC-specific gene sets. Thus, this refined co-expression network can comprehensively quantify global transcriptomic signaling trends on a tumor-by-tumor basis.

As previously reported, we interrogated the dependencies identified through our in vivo CRISPR screens in three models, PDX derived cell lines PATC69 and PATC124 (quasi-basal) and PATC53 (basal). Next, we integrated quantile-normalized BFs obtained from each screen with the PDAC co-expression network to generate dependency networks for each model (Figure 36). The in vivo PATC53 dependency network highlighted multiple unique groups of interconnected vulnerabilities when compared to
PATC69 and PATC124 dependency networks, despite PATC53 and PATC124 both being classified as basal-like, based on the Moffitt signature (Figure 14). Notably, interconnected vulnerabilities in PATC53 were associated with epithelial-to-mesenchymal transition (EMT) (e.g. SNAI1, ZEB1, SMAD4, MAPK11 and MAPK14), integrin signaling (e.g. ILK, ITGB1, ITGB2 and NPNT), heparan sulfate proteoglycan regulation (e.g. SDC3 and EXT1), cell junction regulation (e.g. CTTN, TJP1 and TJP2) (Figure 36).

The integration with the co-expression network also displayed a clear shift in the PATC53-associated dependency spectrum along the cluster 1 -to- cluster 23 axes, with many of the dependencies identified in the CRISPR screen localized within, or adjacent to, cluster 23 (Figure 36). Specifically, MAPK11 and FAM171A2 were localized within cluster 23 itself; NKAIN4 in Cluster 25; and SMAD4, ZEB1 and SDC3 in cluster 31, which exhibits high correlation with cluster 23 expression (Figure 17). The network localization of functionally annotated and PATC53-specific ZEB1, an EMT-associated transcription factor, and SMAD4, an EMT facilitator and oncogenic driver in advanced PDAC, implicated a connection between the cluster 1 to 23 axis, and its associated dependencies, with PDAC epithelial-to-mesenchymal transdifferentiation\textsuperscript{114,115}.

By integrating CRISPR screen-defined co-dependencies with the co-expression network, we were able to identify both common and unique dependencies within the larger context of PDAC diversity (Figure 9). As demonstrated in the previous chapter, single cell expression of the same three PDX derived cell lines, PATC69, PATC124 and PAT53 demonstrates a vast degree of intratumoral heterogeneity.

To address this, we developed an experimental platform to conduct a small scale, direct-capture PerturbSeq\textsuperscript{116}. Using two PDX derived cell lines, PATC69 and PATC53,
we knocked out four genes in an arrayed fashion – one negative control and three hypothesized basal cell dependencies. The objective of this approach was to integrate our approaches (the CRISPR knock-out screening, PDAC co-expression network derived signatures, and network normalized cluster signatures applied to single cell sequencing) to demonstrate that the unique dependencies we identified in our CRISPR screening in PATC53 do represent functional dependencies of the basal subtype. Conversely, this approach also provides an opportunity to study these transcriptional signatures in the context of targeted gene perturbation, as opposed to mainly focusing on correlation with clinical features. Given the intratumoral subtype heterogeneity we have described in our models, we sought to validate if these basal vulnerabilities preferentially impacted only basal single cells. Since the quasi-basal cells show partial expression of both cluster 1 and cluster 23, we sought to evaluate the impact that these gene knock-outs has on this specific population of cells.

We evaluated the effect of perturbing gene targets associated with the basal-like subtype that were prioritized through the integration of the CRISPR screening hits with the co-expression network in varied tumor contexts (schematic presenting the general experimental overview is presented in Figure 36). Based on our in vitro CRISPR screening results, we selected sgRNA sequences targeting SMAD4, ZEB1, and ILK, as well as the non-essential gene, ABCG8, as a negative control. The CRISPR-defined dependencies in PATC53 SMAD4 and ZEB1, both localized within the network, were targeted for cluster 1 to cluster 23 signature validation. ILK, a CRISPR-defined dependency in PATC53 not present in the co-expression network, was selected to
determine whether knockout of this potential EMT regulator would also have the capacity to influence this signature differential\textsuperscript{117-119}.

We applied a feature barcoding strategy where a complement sequence was incorporated into the 3’ end of the sgRNA sequences, enabling scRNAseq sample multiplexing and quantification of the C1vC23 signature shift relative to the sgABCG8 negative control distribution (see methods for details). Individual sgRNAs derived from the CRISPR library were transduced into quasi-basal PATC69 and basal-like PATC53 cells. Cells were cultured cells \textit{in vitro} and collected at the earliest point when separation of essential versus non-essential genes was observed in the original CRISPR screen (Day 20). Sanger sequencing was used to analyze the indel frequency of each sgRNA and colony growth was tracked for each sgRNA to confirm selective growth inhibition in the basal-like PATC53. Multiplexed scRNAseq was conducted on 10,000 cells total (2,500 cells per sgRNA) for each PDX line model. This experiment was conceptualized, designed and conducted by Dr. Johnathon Rose, co-first author on this manuscript.
Figure 36: Co-expression network anchoring of in vivo functional dependencies on a model by model basis. Overlay of context essential genes is represented as quantile-normalized Bayes Factors (BFs), to inform on cluster-associated vulnerability context in PATC69, PATC124, and PATC53 PDX lines. The size of the gene represents the magnitude of the BFs in each screen.
Figure 37: Overview of the feature barcoding strategy for intratumoral tracking of PDAC co-expression network signatures following sgRNA knockouts.

Targeting sgRNAs for ABCG8 (as a negative control), ILK, SMAD4, and ZEB1 were selected and based on CRIPSR screening results. PATC69 and PATC53 cell populations were individually transduced in vitro, selected with puromycin, and parallel assays were also prepared to confirm knockout phenotypes through colony growth, and sgRNA cutting through Sanger sequencing. Each knockout cell population was cultured in vitro for 22 days (CRISPR screening time-point day 10), and then combined scRNAseq library preparation. A total of 10,000 cells per PDX line were sequenced, 2,500 per condition.
6.2: Analysis of the Feature Barcoded scRNAseq Validates Subtype Specific Functional Vulnerabilities

Feature barcoded scRNAseq data were analyzed using Seurat3.1\textsuperscript{54}. Each cell line was individually evaluated. Cells expressing more than 350 genes and less than 25\% mitochondrial reads were retained and subsequently log normalized, variant stabilized, and total RNA count, mitochondrial reads and cell cycle were regressed out. All guide-level samples for each PDX line were merged using the Seurat Anchor Cell feature to provide a direct point of comparison for PATC69 and PATC53 perturbations. A total of 10,113 PATC69 (3449 \textit{ABCG8}, 1962 \textit{ILK}, 2662 \textit{SMAD4}, and 2040 \textit{ZEB1} knockout cells) and 11,439 PATC53 (3414 \textit{ABCG8}, 2952 \textit{ILK}, 2819 \textit{SMAD4}, and 2254 \textit{ZEB1} knockout cells) cells were retained and analyzed for processing. Principal-component analysis and UMAP with the first 20 dimensions was performed, with clustering performed at 0.15 resolution for. Cluster centroids were calculated using the method described above for the patient CNB samples.

PATC69 UMAPs revealed that the knockout conditions did not perform any differently than the \textit{ABCG8} negative control (Figure 38). Single gene knock-outs also displayed no differences in cell viability between \textit{ABCG8} and any of the knockout conditions. In contrast, PATC53 UMAPs revealed a clear separation between the \textit{ABCG8} negative control knockout cells and populations with perturbations in combined cluster 23 (\textit{SMAD4} and \textit{ZEB1}) and basal-like (\textit{ILK})-associated genes (Figure 39). Also, as expected, the PATC69 population transduced with negative control sg\textit{ABCG8} contained co-expression defined classical and quasi-basal cells, with very few basal cells (Figure 40). The basal-like PATC53 population contained quasi-basal and basal-like cells in the
population transduced with the negative control ABCG8, with little to no classical cells in this population (Figure 40).

To test the cluster 1 to cluster 23 signature distribution shift relative to the reference distribution defined by the ABCG8-null negative control, cluster 1 to cluster 23 density plots were generated for each sgRNA knockout cell line (Figure 41), and Kolmogorov’s D statistic was used. Deletion of ILK, SMAD4, and ZEB1 each resulted in a significant cluster 1 to cluster 23 shift towards the cluster 1-enriched classical signature in both PATC69 (Figure 41) and PATC53 (Figure 42). In parallel with scRNAseq, transduced populations were seeded in vitro immediately following selection to compare relative growth phenotypes for both PDX lines (Figure 37). Coinciding with the signature shift towards cluster 1 at the single-cell level, perturbation of each of the three basal-like-associated genes inhibited bulk population growth in PATC53 relative to sgABCG8-knockout controls (Figure 38), whereas no growth phenotype was observed in the classical PATC69 (Figure 38).
Figure 38: Comparison of ILK, SMAD4, and ZEB1 knock-outs to negative control ABCG8 conditions in PATC69.
(Top) UMAP of the PATC69 PDX line with defined ABCG8, ILK, SMAD4, and ZEB1 knockout populations (10113 total cells).
(Bottom) Normalized viability of PATC69 cells following knockout of ABCG8, SMAD4, ZEB1, ILK, or RPS27A with sgRNA. **p < 0.05.
Figure 39: Comparison of ILK, SMAD4, and ZEB1 knock-outs to negative control ABCG8 conditions in PATC53.
(Top) UMAP of the PATC53 PDX line with defined ABCG8, ILK, SMAD4, and ZEB1 knockout populations (11,439 total cells).
(Bottom) Normalized viability of PATC53 cells following knockout of ABCG8, SMAD4, ZEB1, ILK, or RPS27A with sgRNA. **p < 0.05.
Figure 40: UMAPs detailing the cluster 1 and cluster 23 signature in each single-cell RNA sequenced for PATC69 (top) and PATC53 (bottom).

(Top) UMAP of PATC69 with all four knock-out conditions, with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells, and grey represents cells that did not meet the minimum threshold of PDAC co-expression genes expressed to have a cluster centroid score.

(Bottom) UMAP of PATC53 with all four knock-out conditions, with an overlay of a PDAC co-expression network-normalized cluster 1 and cluster 23 signature differential where orange represents the percentage of basal cells, blue represents the percentage of classical cells and beige represents the percentage of quasi-basal cells, and grey represents cells that did not meet the minimum threshold of PDAC co-expression genes expressed to have a cluster centroid score. The red box represents the population of ABCG8 knock-out cells, which cluster separately from the other knock-out conditions.
Figure 41: Comparative PATC69 density plots of cluster 1 and cluster 23 signature following knockout with sgILK, sgSMAD4, sgZEB1 and sgABCG8. Differential is calculated as the differential between the Cluster 1 and 23 centroid score, with more positive differential indicating cluster 1 enrichment and more negative differential indicating cluster 23 enrichment. P values and D statistic derived from Kolmogorov Smirnov test.
Figure 42: Comparative PATC53 density plots of cluster 1 and cluster 23 signature following knockout with sgILK, sgSMAD4, sgZEB1 and sgABCG8. Differential is calculated as the differential between the Cluster 1 and 23 centroid score, with more positive differential indicating cluster 1 enrichment and more negative differential indicating cluster 23 enrichment. P values and D statistic derived from Kolmogorov Smirnov test.
6.3: Discussion

In summary, we provide functional validation of the cluster 1 and 23 signature based on concordant signature shifting, in both quasi-basal PATC69 and basal-like PATC53 cell populations, following genetic perturbation of basal-like associated targets. Importantly, these findings provide evidence for our platform to uncover the intratumoral context of vulnerabilities localized within and adjacent to cluster 23, by quantifying the targeted depletion of cluster 1 and cluster 23 defined basal-like cell populations. This represents a novel approach to leverage network-informed signaling perturbation to influence heterogeneity in pancreatic tumors.

In PATC69, given that there is little to no basal population of cells within the negative control, we cannot speak to the impact of the gene knock-outs on this population. However, it is of note that there is a statistically significant shift towards a positive differential in all three conditions compared to the negative control. This suggests that these identified basal dependencies, predictably do not have an impact on the classical cell population. Surprisingly, knock-out of these basal markers resulted in decreased density of quasi-basal cells, suggesting that in addition to partially expressing both cluster 1 and 23 markers, the quasi-basal population may show partial essentiality to basal dependencies.

This is further supported when we consider PATC53. As expected, the negative control condition in PATC53 displays a proportion of single cells with enrichment in cluster 23. In each of the three knock-out conditions, we see that there is large shift towards the right in these distributions, possibility due to preferential death of basal cells in PATC53. Similar to PATC69, there is an overall shift towards the right of the quasi-basal population
as well. In the negative control condition, there are no cells enriched for cluster 1. However, in all three knock-out conditions, there is an outgrowth of classical cell populations. Given there is no population of cells with classical enrichment in the negative control, it is unlikely that this outgrowth in the knock-out conditions is a result of competition.

This collectively suggests that the quasi-basal population demonstrates some degree of plasticity and is capable of reverting to a classical state in the absence of essential basal dependencies.
Chapter 7: Discussion and Future Directions

A disease characterized by a relatively flat mutational landscape, a dense desmoplastic stroma, adaptive chemoresistance and poor patient prognosis, Pancreatic Ductal Adenocarcinoma (PDAC) persists as a major clinical challenge with limited therapeutic opportunity\textsuperscript{1,3,12}. Despite major clinical efforts through efforts like the phase III POLO trial, establishing clear survival advantage through patient stratification and targeted therapy remains elusive\textsuperscript{37,38}. Recent efforts have sought to expand PDAC stratification efforts into transcriptomic landscape, starting with Moffitt, Collison and Bailey identifying concordant and tumor-intrinsic classical and basal-like subtypes\textsuperscript{43}. These subtypes were then subsequently expanded upon, identifying transitory hybrid groups\textsuperscript{42}. However, while these subtypes are reflective of elements such as patient prognosis and tumor differentiation status, they have yet to inform on underlying genetic dependencies of these tumors.

Advancements in CRISPR-Cas9 functional genomics, tumor modeling, and next-generation sequencing (NGS)\textsuperscript{66,67,69} are providing opportunities to better understand disease diversity and genetic dependency for PDAC. In order to take full advantage of these advancements, \textit{computational methodology to properly utilize these resources need to be innovated}. Here, I developed and implemented several methodologies critical for: a) deciphering pooled customized sgRNA libraries for pancreatic tumors modeled \textit{in vivo}, b) leveraging RNA sequencing data derived from a PDAC PDX cohort, and c) analyzing single-cell sequencing data in pancreatic tumor models and patient biopsies. The overarching purpose of these computation efforts being to anchor genetic
dependencies within the context of signatures underlying PDAC transcriptomic diversity, with a focus on combining tumor modeling and functional genomics.

**Low-Fat BAGEL and *in vivo* Genetic Screens**

As previously described, Low-Fat BAGEL was adapted to maximize guide level data from smaller training sets while employing a linear regression model to decrease the impact of outliers. In doing so, this approach allowed for a more accurate classification of controls, which in turn provided more confidence in the accuracy of essentiality determined for the remainder of the expanded surfaceome library. Applying Low-Fat BAGEL to small custom libraries allows for optimized screening *in vivo*. Leveraging this analysis allowed for a refined comparison of dependencies across multiple PDAC PDX cell lines *in vivo*, highlighting unique and common vulnerabilities and informing on the oncogenic signaling driving these tumors.

Future utilization for Low-Fat BAGEL also has potential for screening custom libraries in organoid settings, or even in direct PDX screening. This Bayesian approach for library analysis also has the potential to be expanded beyond the scope of targets that influence gene proliferation alone. For example, establishing a set of essential and non-essential controls within the context of transcriptional reporter activity could have the potential for a refined analysis, and expansion, of prominent biological pathways. For example, utilizing gene targets whose knockout enhances signaling/reporter activity vs. gene targets whose knockout represses signaling/reporter activity as a training set. Moreover, expanding Low-Fat BAGEL analysis for screens conducted in a syngeneic setting may provide further insight into tumor dependencies within the context of an active
immune systems, assuming random dropout of the implanted tumor population is largely negligible in this case.

**PDAC Co-Expression Network Development and Applications**

Efforts in transcriptomic subtyping efforts have largely been applied through NMF using patient samples with high epithelial content or using a virtually microdissected dataset to limit influence from stroma. While these gene sets are refined using patient data, and have prognostic relevance, this subtyping approach provides an inherently categorical stratification of tumors while primarily only focusing on the most prominent transcriptomic differences between essentially two groups. Development and refinement of a PDAC co-expression network utilizing LLS, and subsequent annotation of highly interconnected gene clusters using InfoMap and GO, served as an orthogonal approach to identify clusters of genes that contribute to transcriptomic diversity within the PDAC PDX cohort. Observation of cluster patterns individually, while providing dimensional reduction to the PDAC transcriptome, initially only highlighted a high level of heterogeneity across the cohort of PDXs. However, when observing cluster patterns within the context of each other, clear anticorrelating signatures pointed towards multiple clear transcriptomic shifts contributing to PDAC diversity. These patterns prominently focused around two anticorrelating gene cluster pairs, cluster 1 vs. cluster 23, and cluster 2 vs. cluster 13. Leveraging these anticorrelating clusters to establish a differential signature to characterize the disease spectrum, capturing transitioning transcriptomic states as a spectrum as oppose to leveraging NMF to categorically bin the disease into discrete groups. This is relevant as the cluster 1 vs. cluster 23 differential signature, along with enriching clinical features such as site of tumor reoccurrence and tumor
differentiation status, is highly enriched for epithelial-to-mesenchymal transition (EMT), based on GSEA. Thus, there is an argument to leveraging a differential signature to characterize disease-specific EMT patterns as a spectrum, as this is more representative of the progressive alterations in transitioning cells. In addition, these larger gene clusters provide a broader interpretation of transcriptional patterns underlying both sides of the transitory spectrum. Additionally, leveraging clusters co-expressed with cluster 23 can serve to provide additional interpretation in models already enriched for cluster 23.

This network model also served as a tool to characterize heterogeneity at the single cell scale. scRNAseq analysis of PDAC PDX models recapitulated findings from bulk tumors, revealing vast intratumoral heterogeneity and a cluster-defined classical to basal-like clonal spectrum with a discrete quasi-basal signature in individual cells. We are able to demonstrate that the quasi-basal subtype that we have identified is a quantifiable state in single cells, and form the majority of cells in all of the models we tested. Furthermore, scRNAseq of patient samples recapitulated that the cluster 1 and 23 anticorrelated signatures present as an intratumoral spectrum, revealing that the majority of the co-expression network signatures are intrinsic to tumor cells, with limited confounding noise from the microenvironment. The main exception here being a over-representation of cluster 13 across the majority of cell types within the PDAC tumor microenvironment.

There is potential for further refinement of the co-expression network. Incorporation of more models can provide greater granularity of co-expressed gene pairs. Additionally, with ongoing trials like COMPASS and POLO, access to omics data directly from patients rather than tumor models is becoming more feasible. As always, cell lines
can also provide an additional source of data. Using the current co-expression network as a foundation, we can construct parallel networks to begin comparing similarities and differences in patterns of co-expression between cell lines, PDXs and patients. By doing this, we can prioritize gene-pairs and clusters that are reinforced in patient datasets, and identify if cell lines and PDXs can serve as adequate representations of the disease.

PDAC is a disease characterized by vast epigenetic alterations and DNA damage. While we did not find strong or overarching associations between mutational signatures and cluster enrichment, there can be utility in exploring other sources of data such as protein-protein interactions, copy number alterations, and epigenomics. Incorporating data from RPPA or mass spectrometry, whole genome sequencing, ChIP-seq, ATAC-seq, DNA methylation, etc. can serve as additional layers to the co-expression network – essentially “activating” or “strengthening” certain gene pairs. Another interesting avenue to consider is applying the co-expression network to spatial transcriptomics. We can incorporate data from emerging technologies, such as slide-seq\textsuperscript{120}, to evaluate if co-expressed genes and clusters that are also colocalized. We can begin to elucidate if the transcriptional heterogeneity that we describe intratumorally is spatially localized, or if we identify a wider pattern of subclonal distribution across a tumor sample.

The PDAC co-expression network, and this analytical approach have provided a mechanism for us to capture the diversity in transcriptional signaling in a comprehensive manner. This platform is also malleable with potential scope for further refinement across many fronts based on the specific hypotheses at hand.
Linking Genetic Dependency with Network-based Transcriptomic Signatures

Recently, the Viale lab at MD Anderson have evaluated tumor evolution and heterogeneity in PDAC within some of the same models we have utilized in this our work (PATC124 and PATC53, quasi-basal and basal like, respectively). They have developed a clonal barcoding technology (CRT) wherein cohorts of PDX tumor derived clones are barcoded, and followed through time and with different treatment modalities, both in vivo and in vivo\textsuperscript{112}. They concluded that tumors display a high degree of intratumoral heterogeneity, which reach an equilibrium over time if left unchallenged by selection pressures\textsuperscript{112}. Of note, when they treated their barcoded tumors with gemcitabine, they identified subsets of clones sensitive and resistant to treatment. Moreover, while tumor size and viability were impacted upon treatment to gemcitabine, attenuation of the treatment lead to regrowth and relapse\textsuperscript{90}. Together this suggests that tumors consist of heterogenous subclones that are plastic in nature and are able to reach a new equilibrium upon selection pressures such as chemotherapy – explaining the general chemoresistant nature of PDAC\textsuperscript{112}. Similarly, the COMPASS trial found that early stage tumors were more likely to belong to the classical A/B subtypes, while stage III and IV tumors were more likely to be basal A/B, with previously reported resectable tumors classified as basal-like, typically being classified by COMPASS as hybrid. They also found that resectable tumors classified as basal A/B or hybrid were highly aggressive, and advanced stage basal A/B were resistant to both gemcitabine and FOLFIRINOX treatment\textsuperscript{42}.

Our work with the feature barcoding system supports the conclusion that these tumors are highly heterogenous. Specifically, the quasi-basal subclonal population that we characterize displays a degree of cellular plasticity. As a result of selection pressures,
here, depleting identified basal dependencies, these subclones are possibly reaching a new equilibrium by reverting to a classical state in order to survive. We can infer from this that patients that present in clinic, even at early or resectable stages, likely present with highly heterogenous tumors that harbor subclones along the entire spectrum of the cluster 1 to 23 subtype continuum. Initial successful treatment response observed in the clinic, especially with adjuvant chemotherapy, likely impacts the sensitive classical subclones within the tumor, leading to initial regression of tumor, or at least reduction in tumor size. However, based on our observations, it is possible that while the classical subclones are effectively depleted with gemcitabine or FOLFIRINOX, again, the quasi-basal subclones are able to adapt to a more basal state to survive. As the surviving tumor subclones reach a new equilibrium, there is increased opportunity for the smaller surviving populations of basal and quasi-basal to outgrow and vastly outnumber the classical subclones, explaining why relapsed tumors are often extremely chemorefractory. This highlights the importance of future studies and clinical assessment focusing on tumor heterogeneity. Focusing equally on how clonal dynamics are impacted as a result of treatment, rather than just tumor size and shrinkage might yield more successful treatments long term, and improve overall survival.

Future directions and expansion of our work can provide further insight into this possibility by specifically testing for it. Adapting the CRT platform developed by the Viale lab, we can track individual subclonal populations over time and treatment. With combined knock out of basal dependencies, we can invariably confirm if the shift to a more classical state is due to cell death of the basal and quasi-basal populations or cellular plasticity. Substituting gene knock-out with gemcitabine treatment, combined with
our co-expression network-based cluster enrichment, we can evaluate if the surviving populations confirm our hypothesis – that classical cells are preferentially depleted, while quasi-basal cells adapt to a more basal state, leading to an outgrowth of this population. Additionally, we can also evaluate the impact of multi-gene knockout, both sequentially and simultaneously. By knocking out multiple genes that represent classical and basal dependencies, we can utilize the barcoding approach to evaluate the impact on the quasi-basal population to provide more insight into the nature of its adaptability and plasticity.

Taken together, our work suggests that both interpatient and intra-patient heterogeneity need to play a bigger role in pancreatic cancer research and treatment. Especially in the development of targeted therapy, patient and model stratification is essential to ensure that we identify ideal treatment regimens based on transcriptional and mutational signatures. In addition to this, future research into pancreatic cancer should focus on the intratumoral impact of perturbation or treatment and evaluate the impact of sequential or simultaneous combinatorial treatment to specifically target different populations of subclones, rather than temporarily perturb the equilibrium of specific subclones.

Conclusions

This dissertation presents computational efforts developed to quantify transcriptional and functional heterogeneity within PDAC. While all of these approaches were custom developed to address specific avenues within our interest in pancreatic cancer research, they are applicable outside of this context as well.

1. Low-Fat BAGEL can be and is utilized for targeted library CRISPR screens in any context.
2. The PDAC co-expression network can serve as a hypothesis generating tool, as a platform to interrogate model by model transcriptional heterogeneity, and provide context for other vulnerabilities or single genes, as we have done. The network methodology in general can also be applied to other model systems, cancers, and diseases and can be easily tailored to addresses the unique constraints of each context.

3. Applying the network-based cluster enrichment methodology to single cell samples allows for interrogation of higher-level gene sets within a system where noise and signal drop out are still a technical concern.

4. Applying a combined gene knock-out and scRNAseq methodology with accompanying computational methods to assess the transcriptional impact of gene knock-out opens the doors for testing various hypotheses that can provide additional insight into gene dependencies beyond that of cell death.

In sum, the work presented in this dissertation has provided further insight into the budding field of pancreatic cancer subtyping and characterization, while also contributing to the field of computational biology and bioinformatics through development and further refinement of novel methodologies.
References


D.P., Iacobuzio-Donahue, C.A., Klimstra, D.S., Saltz, L.B., Sahin, I.H. & O’Reilly,
E.M. Evaluating Mismatch Repair Deficiency in Pancreatic Adenocarcinoma:
17. Iguchi, E., Safgren, S.L., Marks, D.L., Olson, R.L. & Fernandez-Zapico, M.E.
Pancreatic Cancer, A Mis-interpreter of the Epigenetic Language. Yale J Biol
therapeutic targeting strategies. Nature Reviews Gastroenterology & Hepatology
17, 487-505 (2020).
19. Melstrom, L.G., Salazar, M.D. & Diamond, D.J. The pancreatic cancer
20. Collisson, E.A., Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc,
J., Weinkle, J., Kim, G.E., Jakkula, L., Feiler, H.S., Ko, A.H., Olshen, A.B.,
Subtypes of pancreatic ductal adenocarcinoma and their differing responses to
stellate cells in pancreatic inflammation and fibrosis. Clin Gastroenterol Hepatol
7, S48-54 (2009).


64. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F.P., Rissland, O.S., Durocher, D., Angers, S. & Moffat, J. High-


75. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential

76. Kim, E. & Hart, T. Improved analysis of CRISPR fitness screens and reduced off-
target effects with the BAGEL2 gene essentiality classifier. *bioRxiv*,
2020.05.30.125526 (2020).

77. Bhimani, J., Ball, K. & Stebbing, J. Patient-derived xenograft models—the future

xenograft (PDX) models: the implications for translational research and

79. Michels, B.E., Mosa, M.H., Streibl, B.I., Zhan, T., Menche, C., Abou-El-Ardat, K.,
Darvishi, T., Czlonka, E., Wagner, S., Winter, J., Medyouth, H., Boutros, M. &
Farin, H.F. Pooled In Vitro and In Vivo CRISPR-Cas9 Screening Identifies Tumor

80. Szlachta, K., Kuscu, C., Tufan, T., Adair, S.J., Shang, S., Michaels, A.D., Mullen,
M.G., Fischer, N.L., Yang, J., Liu, L., Trivedi, P., Stelow, E.B., Stukenberg, P.T.,
Parsons, J.T., Bauer, T.W. & Adli, M. CRISPR knockout screening identifies
combinatorial drug targets in pancreatic cancer and models cellular drug

81. Carugo, A., Genovese, G., Seth, S., Nezi, L., Rose, J.L., Bossi, D., Cicales, A.,
Shah, P.K., Viale, A., Pettazzoni, P.F., Akdemir, K.C., Bristow, C.A., Robinson,
F.S., Tepper, J., Sanchez, N., Gupta, S., Estecio, M.R., Giuliani, V., Dellino, G.I.,
Riva, L., Yao, W., Di Francesco, M.E., Green, T., D'Allesio, C., Corti, D., Kang, Y.,


Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E.,
Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-
Hoffman, E., Zappala, Z., O’Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek,
M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly,
K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta,
N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N.,
Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K.,
Tolonen, C., Wade, G., Talkowski, M.E., Aguilar Salinas, C.A., Ahmad, T., Albert,
C.M., Ardissino, D., Atzmon, G., Barnard, J., Beaugerie, L., Benjamin, E.J.,
Boehnke, M., Bonnycastle, L.L., Bottinger, E.P., Bowden, D.W., Bown, M.J.,
Chambers, J.C., Chan, J.C., Chasman, D., Cho, J., Chung, M.K., Cohen, B.,
Correa, A., Dabelea, D., Daly, M.J., Darbar, D., Duggirala, R., Dupuis, J., Ellinor,
P.T., Elosua, R., Erdmann, J., Esko, T., Färkkilä, M., Florez, J., Franke, A., Getz,
G., Glaser, B., Glatt, S.J., Goldstein, D., Gonzalez, C., Groop, L., Haiman, C.,
Hanis, C., Harms, M., Hiltunen, M., Holi, M.M., Hultman, C.M., Kallela, M.,
Kaprio, J., Kathiresan, S., Kim, B.-J., Kim, Y.J., Kirov, G., Kooner, J., Koskinen,
S., Krumholz, H.M., Kugathasan, S., Kwak, S.H., Laakso, M., Lehtimäki, T.,
Loos, R.J.F., Lubitz, S.A., Ma, R.C.W., MacArthur, D.G., Marrugat, J., Mattila,
K.M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Meigs, J.B.,
Melander, O., Metspalu, A., Neale, B.M., Nilsson, P.M., O’Donovan, M.C., Ongur,
Pulver, A.E., Rahman, N., Remes, A.M., Rioux, J.D., Ripatti, S., Roden, D.M.,


Vita

Sanjana Srinivasan was born to Sujatha Srinivasan in Mumbai, India and graduated from Sir Winston Churchill High School, in Calgary in 2009. She received a Bachelor of Sciences in Psychology, Exceptionality in Human Learning, with a minor in Biology from the University of Toronto Mississuaga in June 2013. During her undergraduate studies, Sanjana conducted research in the Dr. Ashely Monnks’ lab studying the physiological and behavioral effects of overexpression of androgen receptors in the central and peripheral nervous system. Following this, Sanjana completed her Masters of Public Health in Epidemiology from the University of Texas School of Public Health in August 2016. During this time, she worked as a graduate research assistant in the Behavioral Science Department at MD Anderson Cancer Center with Dr. Karen Basen-Engquist. Sanjana joined The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences in August of 2016.