Somatic Copy Number Aberrations In Familial Pancreatic Cancer: Integrative Genomics And Gene Discovery

by

Zaheer Shamshudin Kanji, MD

A thesis submitted in conformity with the requirements for the degree of Masters of Science

Institute of Medical Science
University of Toronto

© Copyright by Zaheer S. Kanji 2013
Somatic Copy Number Aberrations In Familial Pancreatic Cancer: Integrative Genomics and Gene Discovery

Zaheer S. Kanji, MD
Masters of Science
Institute of Medical Science
University of Toronto
2013

Abstract

Familial Pancreatic Cancer (FPC) is an autosomal dominant condition with greater than 80% of genetic causes unknown. We hypothesize that an integrative approach employing germline exome sequencing and somatic microarray analysis of FFPE DNA will identify novel tumour suppressor genes (TSGs). 55 FPC and 21 sporadic PDAC tumours were analyzed on the Affymetrix Oncoscan FFPE Express 2.0 SNP microarray and compared to data from 33 germline FPC cases analyzed on the Illumina GAIIx Analyzer. We have demonstrated that FPC is genetically heterogeneous with recurrent loss at CDKN2A/p16, TP53 and SMAD4. Analysis of 2 sisters has shown a shared loss region involving DCLK3 and SERPINF1. By an integrative approach, we have identified ATPAF1-AS1 and MAP3K6 as potential TSGs. Germline variants of these genes have been confirmed by Sanger sequencing and somatic fluorescence in-situ hybridization. Future functional studies will better characterize the importance of these regions and novel putative TSGs in FPC.
Acknowledgments

The success of this project was not an individualistic effort and it is to those who have provided me countless hours of support that I would like to recognize.

Firstly, I would like to thank the patients with pancreatic cancer and their families who have demonstrated an unwavering determination and drive to conquer this horrible disease. Their will has provided me inspiration to find a cure for this disease and I can only hope that the work contained in this thesis will provide a future benefit for patients afflicted with this condition. I am grateful for their participation in the Ontario Pancreas Cancer Study, which has made this project possible.

To my lab colleagues, Kazy, Treasa, Hyeja, Colleen, Thomas, Iris, Nicolas, Ayelet and Spring. You have seen me go through both good times and bad and I want to commend you on your patience and in your unwavering belief that I would overcome and persevere. Through your technical expertise, prowess and willingness to teach, I have not only been able to complete this challenging project but also acquire the rudimentary skills in becoming a researcher. Your help will most definitely not be forgotten.

I am forever indebted to the support and expertise of my program advisory committee and informal mentors, specifically Drs. Thomas Hudson, David Malkin, Michael Taylor, John D. McPherson and Stefano Serra. I can attest to the challenges of finding time in the clinical/research world and I truly appreciate your sacrifice with very little remuneration in improving my research study and developing me as a clinical scientist.

To my supervisor, Dr. Steven Gallinger, who has made all this possible, thank you for believing in me after a random meeting in Toronto in October 2010. I only hope I can make you as proud as I feel privileged to call you mentor and hopefully one day colleague. I truly wish that this is only the beginning of our personal and professional friendship.

To my colleagues and mentors in the General Surgery Residency Program at the University of British Columbia and Clinician Investigator Program (Drs. Morad Hameed, Sonia Butterworth, Stephen Chung, Sian Spacey and Mrs. Tessa Feuchuk and Eleni Tsakumis), for which none of
this would exist without you, thank you for pushing me past my limits and believing that I would still stand afterwards.

Finally, to my parents, Shamshudin and Fatma Kanji and my life partner Dr. Siham Zerhouni, words cannot even express how blessed I feel to have such generous human beings in my life. Thank you for always being there whether I deserved it or not. To my mom, your ongoing battle with a chronic illness was the source of my motivation in becoming a doctor. I only hope that I can emulate the poise and dignity that you have displayed through this very tough fight and fulfill all the expectations you ever had for me. I love you all in its fullest sense.

“Success is not final, failure is not fatal: It is the courage to continue that counts.”

-Winston Churchill
# Table of Contents

## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>V</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>VIII</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>IX</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>XI</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>XII</td>
</tr>
<tr>
<td>SUMMARY OF EXPERIMENTS PERFORMED</td>
<td>XVI</td>
</tr>
</tbody>
</table>

## CHAPTER 1 BACKGROUND AND LITERATURE REVIEW

### 1 PANCREATIC CANCER

1.1 Anatomy, Histology and Epidemiology ........................................ 1
1.2 Clinical Presentation and Diagnosis ........................................... 3
1.3 Risk Factors ................................................................................... 5
1.4 Molecular Biology and Genetics of PDAC ......................................... 7
1.5 Familial Basis of PDAC .................................................................... 15
1.6 Familial Pancreatic Cancer (FPC) ................................................ 20

## CHAPTER 2 COPY NUMBER VARIATION

2.1 CNV Detection Methods .................................................................... 27
2.2 CNVs and Human Disease ................................................................... 30
2.3 CNVs and Human Cancers ................................................................... 32
2.4 CNVs and FPC .................................................................................. 33

## CHAPTER 3 NEXT GENERATION SEQUENCING

3.1 Sequencing and Human Genome Project ........................................... 35
3.2 Second Generation Sequencing ..................................................... 36

## CHAPTER 2 STUDY OUTLINE

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
</tr>
</tbody>
</table>
2 RATIONALE, AIMS AND HYPOTHESIS ................................................................. 38
2.1 RATIONALE ........................................................................................................ 38
2.2 AIMS .................................................................................................................... 38
2.3 HYPOTHESIS ....................................................................................................... 39

CHAPTER 3 STUDY DESIGN ...................................................................................... 41

3 MATERIAL AND METHODS ................................................................................... 41
3.1 ONTARIO PANCREAS CANCER STUDY AND TISSUE SPECIMENS .................. 41
3.2 HISTO-PATHOLOGICAL CONFIRMATION .......................................................... 41
3.3 LASER CAPTURE MICRODISSECTION (LCM) ....................................................... 43
3.4 DNA EXTRACTION ................................................................................................. 43
3.5 NANODROP SPECTROPHOTOMETRY .................................................................. 45
3.6 WHOLE GENOME AMPLIFICATION .................................................................. 45
3.7 QUBIT DOUBLE STRANDED DNA QUANTIFICATION ............................................ 46
3.8 AFFYMETRIX ONCOSCAN FFPE EXPRESS VERSION 2.0 ............................... 47
3.9 BIOINFORMATICS ................................................................................................. 51
3.10 GERMLINE NEXT GENERATION SEQUENCING AND GENE FILTERING ........ 53
3.11 SANGER SEQUENCING OF CANDIDATE VARIANTS ......................................... 55
3.12 FLUORESCENCE IN-SITU HYBRIDIZATION ...................................................... 59

CHAPTER 4 DATA COLLECTION .................................................................................. 60

4 RESULTS .................................................................................................................. 60
4.1 SAMPLE COHORT AND MICROARRAY PROFILE INTERPRETATION ............... 60
4.2 FPC ASCAT COPY NUMBER PROFILE .................................................................. 62
4.3 FPC SNP-FASST2 SEGMENTATION COPY NUMBER PROFILE ........................... 71
4.4 SPORADIC VS. FPC CNV PROFILE COMPARISON ............................................ 77
4.5 COMPARISON WITH MICROSATELLITE MARKER CNV PROFILE .................... 79
4.6 SHARED LOH REGIONS ......................................................................................... 81
4.7 FISH CONFIRMATION OF SHARED LOH REGIONS............................................. 85
4.8 2-HIT HYPOTHESIS GENES ........................................................................... 90
4.9 SANGER SEQUENCING VALIDATION OF 2-HIT GENES................................. 91
4.10 FISH CONFIRMATION OF SOMATIC LOH OF PUTATIVE 2-HIT TSGS........... 94
4.11 SOMATIC GISTIC ANALYSIS AND GENE FILTERING ............................... 97
4.12 GISTIC GENES GERMLINE AND SOMATIC CONFIRMATION .................... 100

CHAPTER 5 DATA ANALYSIS ............................................................................. 106

5 DISCUSSION ........................................................................................................ 106
5.1 FFPE DNA AND MICROARRAY ANALYSIS IN PDAC .................................. 106
5.2 SAMPLE CHARACTERISTICS AND MICROARRAY ANALYSIS IN PDAC ....... 107
5.3 COMPARISONS OF FPC WITH SPORADIC PDAC AND CN-LOH ................. 111
5.4 INTEGRATIVE GENOMICS ............................................................................. 114

6 CONCLUSION ....................................................................................................... 118

7 FUTURE DIRECTIONS ......................................................................................... 119

BIBLIOGRAPHY ...................................................................................................... 121

APPENDICES ........................................................................................................ 146
# List of Tables

Table 1: Staging and Prognosis in Pancreatic Cancer

Table 2: Age-Adjusted Incidence of Pancreatic Cancer by Race and Gender 2005-2009

Table 3: Candidate FPC genes PCR primer specifications

Table 4: Individual sample ASCAT Ploidy and Normal Cell Contamination Predictions

Table 5: Fisher Exact Test demonstrating significant regions of LOH difference between sporadic PDAC and FPC CNV profiles

Table 6: Gene Annotations for the Shared LOH regions in Family 385P

Table 7: 2-Hit Hypothesis Genes Summary

Table 8: Summary of GISTIC Loss Regions on Nexus 6.1

Table 9: Summary of GISTIC Genes with Rare Germline Inactivating Variants
List of Figures

Figure 1: Trends in Death Rates by Sex in the United States from 1930-2009 .......................... 2

Figure 2: Diagnostic Flow Algorithm in PDAC ................................................................. 5

Figure 3: Adenoma to Carcinoma Sequence of Sporadic PDAC .............................................. 13

Figure 4: Summary of Hereditary Pancreatic Cancer Syndromes ........................................... 20

Figure 5: A pedigree depicting a typical FPC Family ............................................................ 21

Figure 6: Illustration of the Knudson 2-Hit Hypothesis of TSGs ............................................ 39

Figure 7: MIP technology and array hybridization ............................................................... 50

Figure 8: Total FPC cases available for a paired matched analysis ........................................ 60

Figure 9: Individual genome-wide Oncoscan microarray tracing for an FPC sample ............... 61

Figure 10: ASCAT CNV Output Tracing for FPC cases ....................................................... 65

Figure 11: ASCAT Allelic Ratios CNV Output Tracing for FPC Cases ............................... 67

Figure 12: ASCAT Sporadic CNV Output Tracing ............................................................... 69

Figure 13: SNP FASST2 Segmentation FPC CNV Output Tracing ........................................ 73

Figure 14: SNP FASST2 Segmentation FPC Allelic Ratios CNV Plot ................................. 74

Figure 15: SNP FASST2 Segmentation Sporadic Allelic Ratios CNV Plot ............................ 76

Figure 16: MS marker vs. Oncoscan SNP array FPC CNV profile ...................................... 80

Figure 17: Pedigrees of families demonstrating shared LOH regions on the SNP-FASST2 Segmentation Algorithm ................................................................. 81

Figure 18: Shared Loss Regions identified on the SNP-FASST2 Segmentation Algorithm ...... 83
Figure 19: FISH assay for the shared LOH region on chromosome 3p22.2-23.3 using a DCKL3 specific BAC clone…………………………………………………………………………………………...85

Figure 20: FISH assay for the shared LOH region on chromosome 17p13.3 using a SERPINF1 specific BAC clone…………………………………………………………………………………………...88

Figure 21: Sanger Sequencing Electropherograms for Candidate 2-Hit Hypothesis Genes………92

Figure 22: FISH Results for candidate 2-Hit genes demonstrating somatic LOH………………94

Figure 23: Sanger Sequencing of Germline Tissue corresponding to GISTIC Gene variants...100

Figure 24: FISH Results for Candidate GISTIC Genes with Rare Germline Variants………...102
List of Appendices

APPENDIX TABLES:

Table S1: Summary of BAC clones used for FISH assays.................................146
Table S2: Summary of FPC Copy Gain and Loss Regions Called on ASCAT...........146
Table S3: Summary of FPC Allelic Loss Called on ASCAT................................146
Table S4: Summary of Sporadic PDAC Copy Gain and Loss Regions Called on ASCAT......146
Table S5: Summary of Sporadic PDAC Allelic Loss Called on ASCAT....................147
Table S6: Summary of FPC Copy Gain and Loss Regions Called on SNP-FASST2 Segmentation.................................................................147
Table S7: Summary of Sporadic Copy Gain and Loss Regions Called on SNP-FASST2 Segmentation.................................................................147
Table S8: Summary of CN Loss Genes identified by Fisher Test on SNP-FASST2 Segmentation.................................................................147
Table S9: Summary of CN Loss Genes identified by Fisher Test on ASCAT.............147
Table S10: CNV Regions for Cases with Germline Exome Sequencing and Oncoscan Data called on SNP-FASST2 Segmentation.................................................................147
Table S11: Genes identified on GISTIC on the ASCAT Algorithm........................149
Table S12: Genes identified on GISTIC on the SNP-FASST2 Segmentation Algorithm.......149

APPENDIX FIGURES:

Figure S1: BAC probe localization on peripheral blood lymphocyte chromosomes..........147
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCGH</td>
<td>Array Comparative Genomic Hybridization</td>
</tr>
<tr>
<td>ASCAT</td>
<td>Allele Specific Copy Number Analysis of Tumours</td>
</tr>
<tr>
<td>ASD</td>
<td>Autism Spectrum Disorder</td>
</tr>
<tr>
<td>AOMF</td>
<td>Advanced Optical Microscope Facility</td>
</tr>
<tr>
<td>APC</td>
<td>Adenomatous Polyposis Coli</td>
</tr>
<tr>
<td>AT</td>
<td>Ataxia Telangiectasia</td>
</tr>
<tr>
<td>aUPD</td>
<td>Acquired Uniparental Disomy</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>BAF</td>
<td>B-Allele Frequency</td>
</tr>
<tr>
<td>BD-IPMN</td>
<td>Branch Duct Intra-ductal Papillary Mucinous Neoplasm</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>bp</td>
<td>Base Pair</td>
</tr>
<tr>
<td>BR</td>
<td>Broad Range</td>
</tr>
<tr>
<td>BSR</td>
<td>Biospecimen Repository</td>
</tr>
<tr>
<td>CBS</td>
<td>Circular Binary Segmentation</td>
</tr>
<tr>
<td>CDKi</td>
<td>Cyclin Dependant Kinase Inhibitor</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CN</td>
<td>Copy Number</td>
</tr>
<tr>
<td>CN-LOH</td>
<td>Copy Neutral LOH</td>
</tr>
<tr>
<td>CNP</td>
<td>Copy Number Polymorphism</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number Variation</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal Cancer</td>
</tr>
<tr>
<td>CT</td>
<td>Computer Axial Tomography</td>
</tr>
<tr>
<td>ddH₂O</td>
<td>Double distilled water</td>
</tr>
<tr>
<td>ddNTP</td>
<td>dideoxynucleoside triphosphate</td>
</tr>
<tr>
<td>DGV</td>
<td>Database of Genomic Variants</td>
</tr>
<tr>
<td>DNA</td>
<td>De-oxy ribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleoside triphosphate</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double Stranded DNA</td>
</tr>
<tr>
<td>ES</td>
<td>Exome Sequencing</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>FAMMM</td>
<td>Familial Atypical Malignant Mole Melanoma</td>
</tr>
<tr>
<td>FAP</td>
<td>Familial Adenomatous Polyposis</td>
</tr>
<tr>
<td>FDR</td>
<td>First Degree Relative</td>
</tr>
<tr>
<td>FFPE</td>
<td>Formalin Fixed Paraffin Embedded</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescence In-Situ Hybridization</td>
</tr>
<tr>
<td>FPC</td>
<td>Familial Pancreatic Cancer</td>
</tr>
<tr>
<td>GATK</td>
<td>Genome Analysis Toolkit</td>
</tr>
<tr>
<td>Gb</td>
<td>Gigabase pair</td>
</tr>
<tr>
<td>GI</td>
<td>Gastro-Intestinal</td>
</tr>
<tr>
<td>GISTIC</td>
<td>Genomic Identification of Significant Targets in Cancer</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome Wide Association Studies</td>
</tr>
<tr>
<td>H+E</td>
<td>Hematoxylin and Eosin</td>
</tr>
<tr>
<td>HBOC</td>
<td>Hereditary Breast Ovarian Cancer</td>
</tr>
<tr>
<td>HGP</td>
<td>Human Genome Project</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HP</td>
<td>Hereditary Pancreatitis</td>
</tr>
<tr>
<td>HPC</td>
<td>Hereditary Pancreatic Cancer</td>
</tr>
<tr>
<td>ICGC</td>
<td>International Cancer Genome Consortium</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion Deletion Polymorphism</td>
</tr>
<tr>
<td>IPMN</td>
<td>Intra-ductal Papillary Mucinous Neoplasm</td>
</tr>
<tr>
<td>JPS</td>
<td>Juvenile Polyposis Syndrome</td>
</tr>
<tr>
<td>kb</td>
<td>Kilo base pair</td>
</tr>
<tr>
<td>LCM</td>
<td>Laser Capture Microdissection</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of Heterozygosity</td>
</tr>
<tr>
<td>LRR</td>
<td>Log R Ratio</td>
</tr>
<tr>
<td>LS</td>
<td>Lynch Syndrome</td>
</tr>
<tr>
<td>MAPD</td>
<td>Mean Absolute Percent Deviation</td>
</tr>
<tr>
<td>Mb</td>
<td>Mega base pair</td>
</tr>
<tr>
<td>MCN</td>
<td>Mucinous Cystic Neoplasm</td>
</tr>
<tr>
<td>MD-IPMN</td>
<td>Main Duct Intra-ductal Papillary Mucinous Neoplasm</td>
</tr>
<tr>
<td>MIP</td>
<td>Molecular Inversion Probe</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MMR</td>
<td>Mismatch Repair</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MS</td>
<td>Microsatellite</td>
</tr>
<tr>
<td>MSI</td>
<td>Microsatellite Instability</td>
</tr>
<tr>
<td>MSI-H</td>
<td>Microsatellite Instability-High</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>OD</td>
<td>Optical Density</td>
</tr>
<tr>
<td>OICR</td>
<td>Ontario Institute for Cancer Research</td>
</tr>
<tr>
<td>OPCS</td>
<td>Ontario Pancreas Cancer Study</td>
</tr>
<tr>
<td>OR</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>OS</td>
<td>Overall Survival</td>
</tr>
<tr>
<td>OSCC</td>
<td>Oral Squamous Cell Carcinoma</td>
</tr>
<tr>
<td>PACGENE</td>
<td>Pancreatic Cancer Genetic Epidemiology Consortium</td>
</tr>
<tr>
<td>PanIN</td>
<td>Pancreatic Intra-epithelial Neoplasia</td>
</tr>
<tr>
<td>PC</td>
<td>Pancreatic Cancer</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PDAC</td>
<td>Pancreatic Ductal Adenocarcinoma</td>
</tr>
<tr>
<td>PJS</td>
<td>Peutz-Jeghers Syndrome</td>
</tr>
<tr>
<td>PNET</td>
<td>Pancreatic Neuroendocrine Tumour</td>
</tr>
<tr>
<td>PRP</td>
<td>Pathology Research Processing</td>
</tr>
<tr>
<td>Qc</td>
<td>Quality Control</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RPM</td>
<td>Rotations per Minute</td>
</tr>
<tr>
<td>RR</td>
<td>Relative Risk</td>
</tr>
<tr>
<td>SCA</td>
<td>Serous Cystadenoma</td>
</tr>
<tr>
<td>SCNA</td>
<td>Somatic Copy Number Aberration</td>
</tr>
<tr>
<td>SIR</td>
<td>Standardized Incidence Ratio</td>
</tr>
<tr>
<td>SLRI</td>
<td>Samuel Lunenfeld Research Institute</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single Nucleotide Variant</td>
</tr>
<tr>
<td>SpG</td>
<td>Spectrum Green</td>
</tr>
<tr>
<td>SpO</td>
<td>Spectrum Orange</td>
</tr>
</tbody>
</table>

xv
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPN</td>
<td>Solid Pseudopapillary Neoplasm</td>
</tr>
<tr>
<td>SSR</td>
<td>Simple Sequence Repeat</td>
</tr>
<tr>
<td>TNC</td>
<td>Tenascin C</td>
</tr>
<tr>
<td>TSG</td>
<td>Tumour Suppressor Gene</td>
</tr>
<tr>
<td>UPD</td>
<td>Uniparental Disomy</td>
</tr>
<tr>
<td>US</td>
<td>Ultrasound</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular Endothelial Growth Factor</td>
</tr>
<tr>
<td>VNTR</td>
<td>Variable Number Tandem Repeat</td>
</tr>
<tr>
<td>VUS</td>
<td>Variant of Undetermined Significance</td>
</tr>
<tr>
<td>WES</td>
<td>Whole Exome Sequencing</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole Genome Amplification</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Sequencing</td>
</tr>
</tbody>
</table>
Summary of Experiments Performed

Research ethics approval for this study was obtained by the University Health Network and Mount Sinai Hospital Research Ethics Boards, Toronto, Ontario. Pathological review of cases was performed at the Toronto General Hospital Department of Pathology, Toronto, Ontario with the gracious assistance of Dr. Stefano Serra. Microdissection, DNA extraction, whole genome amplification and DNA quantification was conducted at the Samuel Lunenfeld Research Institute of Mount Sinai Hospital (SLRI), Toronto, Ontario. Zaheer S. Kanji performed the above mentioned procedures and the analysis involved with this. Microarray processing of somatic FFPE DNA and initial data normalization was performed by Affymetrix Inc. (Santa Clara, California) however subsequent data analysis was performed by Zaheer S. Kanji. Ms. Olga Ludkovski at Princess Margaret Hospital, Toronto, Ontario performed FISH at the Applied Molecular Pathology Laboratory. Finally, DNA extraction for germline ES was conducted by the Biospecimen Repository (BSR) at the SLRI and exome sequencing and variant calling was performed by Robert Grant and Dr. John D. McPherson at the Ontario Institute for Cancer Research (OICR), Toronto, Ontario.
Chapter 1
Background and Literature Review

Selected sections of this chapter have been published in the Canadian Medical Association Journal (Kanji, Gallinger 2013) with myself as first author and in Human Genomics (Grant, Al-Sukhni et al. 2013) with myself as a contributing author.

1 Pancreatic Cancer

1.1 Anatomy, Histology and Epidemiology

1.1.1 PANCREATIC ANATOMY AND HISTOPATHOLOGY

The pancreas, comprising the pancreatic head, uncinate process, neck, body and tail, is a major organ of digestion and metabolism. It is located in the retroperitoneal cavity, which along with the kidneys and great vessels, are concealed from the remaining abdominal viscera. The pancreas has both exocrine (digestion) and endocrine (metabolism) functions, which are central to the maintenance of physiological homeostasis. Histologically, exocrine functions are executed by the acini and ductal systems whereas metabolic functions, and specifically glucose monitoring, are controlled by the Islets of Langerhans. Digestive enzymes are secreted by the acini into terminal ductules as inactivated zymogen granules. These granules are transported via the main pancreatic duct and emptied into the small intestine (duodenum) where they execute the breakdown of carbohydrates, fats and proteins. Islet cells are divided into alpha, beta and delta cells, which secrete glucagon, insulin and somatostatin respectively. Alpha and beta cells are the main glucose regulators and have a rich blood supply facilitating secretion and systemic release of hormones.

Cancers of the pancreas arise from both exocrine and endocrine components of the organ. Cancer of the exocrine pancreas includes various histologies with ductal adenocarcinoma (PDAC) accounting for greater than 90% of cases (Samuel, Hudson 2011). The remaining subtypes (acinar cell carcinoma, adeno-squamous carcinoma, cystadenocarcinoma) are rare and generally excluded in the discussion of pancreatic cancer (PC). Cancers of the endocrine pancreas are termed pancreatic neuroendocrine tumours (PNETs) and include a variety of subtypes such as insulinoma, glucagonoma, somatostatinoma and non-functional tumours. PNETs account for 1-
3% of pancreatic tumours and display a distinct biology, incidence and survival compared to the exocrine type (de Wilde, Edil et al. 2012). As PDAC forms the overwhelming majority and clinically most relevant subtype, the term PC is used, in both research and clinical settings, to refer to this exclusively.

### 1.1.2 EPIDEMIOLOGY OF PDAC

Advances in molecular biology, diagnostic imaging and cancer therapeutics have deepened our understanding of PDAC over the past 50 years, yet PDAC continues to remain one of the most fatal cancers in the Western world. Approximately 45,000 new cases and 38,000 deaths from PDAC are estimated for 2013 in the United States ranking it as the 10\(^{th}\) most common malignancy and 4\(^{th}\) leading cause of cancer death (Siegel, Naishadham et al. 2013). In Canada, approximately 4,600 incident cases of PDAC will be diagnosed and 4,300 patients will die from this disease in 2013, ranking it as the 12\(^{th}\) most common malignancy and 4\(^{th}\) leading cause of cancer death (Canadian Cancer Society's Steering Committee on Cancer Statistics 2012). The age-adjusted incidence for PDAC is 10 cases per 100,000 and the baseline risk of developing PDAC in the general population is 1.45% (95% CI: 1.45-1.48) (Howlader, Noone et al. 2012). Whereas mortality rates for most solid malignancies appear to be declining, death rates from PDAC have been steadily increasing since the 1930s (Figure 1). Furthermore, with greater adoption and popularity of screening strategies for breast, prostate and colorectal cancer (CRC), it is anticipated that PC will eventually become the leading cause of cancer death in North America.

Figure 1: Trends in Death Rates by Sex in the United States from 1930-2009 (Siegel, Naishadham et al. 2013) (Reproduced with Permission)
1.2 Clinical Presentation and Diagnosis

The rapidly fatal nature of PDAC is partially attributable to the anatomic location of the pancreas, which permits cancer growth in the absence of early symptoms. Approximately 20% of patients at the time of incidental presentation to a physician will be candidates for surgical resection (Stage I and II). Unfortunately, only 20% of patients having undergone curative therapy will remain alive at 5 years (Al-Sukhni, Borgida et al. 2012). In comparison, these survival statistics for resectable PDAC are similar to those for late stage advanced breast cancer, indicating that the clinical definition of early disease does not reflect the underlying biology. The vast majority of patients with PDAC present with locally advanced or metastatic disease (Stage III and IV) and management for these individuals is palliative (chemo and/or radiation therapy). The overall 5-year survival for patients with PDAC at all stages is 6% (Table 1).

Table 1: Staging and Prognosis in Pancreatic Cancer (Reprinted from Kanji, Gallinger 2013 Diagnosis and Management of Pancreatic Cancer; Table 1 staging and prognosis in pancreatic cancer; Canadian Medical Association. This work is protected by copyright and the making of this article was with the permission of the Canadian Medical Association Journal.)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Tumour Grade</th>
<th>Node Status</th>
<th>Distant Metastases</th>
<th>5-Year Survival (%)*</th>
<th>Median Survival (mo)**</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>T1</td>
<td>N0</td>
<td>M0</td>
<td>14</td>
<td>24</td>
<td>&lt;2cm tumour in pancreas only</td>
</tr>
<tr>
<td>IB</td>
<td>T2</td>
<td>N0</td>
<td>M0</td>
<td>12</td>
<td>21</td>
<td>&gt;2cm tumour in pancreas only</td>
</tr>
<tr>
<td>IIA</td>
<td>T3</td>
<td>N0</td>
<td>M0</td>
<td>7</td>
<td>15</td>
<td>Tumour extends beyond the pancreas but no involvement of the celiac/SMA</td>
</tr>
<tr>
<td>IIB</td>
<td>T 1-3</td>
<td>N1</td>
<td>M0</td>
<td>5</td>
<td>13</td>
<td>Regional lymph node metastasis</td>
</tr>
<tr>
<td>III</td>
<td>T4</td>
<td>N 0-1</td>
<td>M0</td>
<td>3</td>
<td>11</td>
<td>Tumour involves the celiac/SMA</td>
</tr>
<tr>
<td>IV</td>
<td>T 1-4</td>
<td>N 0-1</td>
<td>M1</td>
<td>1</td>
<td>5</td>
<td>Distant metastases</td>
</tr>
</tbody>
</table>

* T- Primary Tumour, N= Regional Lymph Nodes, M= Distant Metastases
** Statistics based on patients receiving stage recommended treatment

1.2.1 CLINICAL SIGNS AND SYMPTOMS

The onset of symptoms in PDAC is often a sign of advanced disease and most early cases are incidentally detected by imaging requested for complaints unrelated to the cancer. Clinical symptoms in PDAC are dictated by stage and location of the tumour (Artinyan, Soriano et al. 2008): the distribution of tumours includes pancreatic head/neck/uncinate process (70%), body or tail (20%) and multifocal disease (10%). The majority of tumours develop in the pancreatic head, common symptoms include jaundice, right upper quadrant and epigastric pain, nausea and
vomiting due to mass effect on the stomach, new or worsening diabetes and diarrhea/steatorrhea secondary to pancreatic insufficiency and fat malabsorption (Kanji, Gallinger 2013). The presence of new onset or worsening back pain without a definable cause is also suggestive of PDAC localized to the body or tail, with secondary neural infiltration of the celiac plexus. Systemic manifestations of pancreatic cancer caused by paraneoplastic phenomena include weight loss, anorexia, depression and thromboembolic disease (Porta, Fabregat et al. 2005).

1.2.2 DIAGNOSIS OF PDAC

The detection of PDAC relies on radiological techniques (Figure 2). For patients who present on an emergent basis with non-specific abdominal pain and/or jaundice, trans-abdominal ultrasound (US) serves as the initial imaging modality of choice. Although US displays a high sensitivity (90%) and specificity (95%) for the detection of large masses (>3cm), its use is limited largely due to its operator dependency and its inability to distinguish cancer from chronic pancreatitis (Miura, Takada et al. 2006). The mainstay of PDAC diagnosis is with triple-phase contrast enhanced computer axial tomography (CT) scan. The resolution of modern scanners allows accuracy in diagnosis (sensitivity: 81-99%; specificity: 70-93%) and spatial relationships to surrounding organs and vascular structures. Other imaging modalities such as magnetic resonance imaging (MRI), although diagnostically equivalent to CT, is usually unnecessary for the diagnosis of PDAC (Birchard, Semelka et al. 2005), and is used in patients with contrast allergies or when there are diagnostic dilemmas. Routine pathological diagnoses with either endoscopic ultrasound (sensitivity: 92%; specificity: 96%), CT guided (sensitivity: 80-90%; specificity: 98-100%) biopsy or endoscopic retrograde pancreatography with cytology (sensitivity: 40-60%; specificity: 91-100%) are not indicated for patients scheduled for surgery as a negative biopsy will not alter management (Hidalgo 2010, Wight, Zaitoun et al. 2004) and potential risks are not insignificant (tumor seeding, bleeding, pancreatitis). Biopsies are generally required for patients with advanced disease awaiting palliative treatment or when diagnosis is uncertain (e.g. PDAC vs. autoimmune pancreatitis).

Figure 2: Diagnostic Flow Algorithm in PDAC (Reprinted from Kanji, Gallinger 2013 Diagnosis and Management of Pancreatic Cancer; Figure 1 diagnostic algorithm for pancreatic cancer; Canadian Medical Association. This work is protected by copyright and the making of this article was with the permission of the Canadian Medical Association)
1.3 Risk Factors

1.3.1 EXTRINSIC RISK FACTORS

Risk factors for PDAC are broadly divided into environmental and genetic causes. Although numerous environmental factors have been investigated for causative roles in PDAC, currently active smoking (Klein 2013) remains the most established risk factor (odds ratio [OR]: 1.74, 95% confidence interval [CI]: 1.61-1.87). Risk of PDAC for previous smokers reduces by 1.2 fold and resets to the general population baseline risk after 20 years cessation (Klein 2012). Other putative associations include an elevated body mass index (BMI) > 35 (OR: 1.55, 95% CI: 1.16-2.07) compared to individuals with a BMI<25 and heavy alcohol consumption >6 beverages per day (OR: 1.46, 95% CI: 1.16-1.83) compared to consumption of <1 beverage per day. The impact of chronic pancreatitis (usually secondary to chronic alcohol abuse) on PDAC risk remains controversial. A recent meta-analysis (Raimondi, Lowenfels et al. 2010) of 22
studies (12 case-control and 10 cohort) demonstrated that individuals with chronic pancreatitis in fact had a higher risk of developing PDAC (Relative Risk [RR]: 13.3, 95% CI: 6.1-28.9).

Interestingly, the presence of non-type O blood group (Wolpin, Kraft et al. 2010) increases risk of PDAC (OR: 1.45, 95% CI: 1.21-1.66) possibly due to aberrant expression of AB antigens on pancreatic tumor surfaces, which alter signal transduction and cellular adhesion mechanisms. Recent epidemiological studies have shown that allergies (Olson, Hsu et al. 2013, Eppel, Cotterchio et al. 2007) confer a protective role against PDAC (OR: 0.40, 95% CI: 0.26-0.87) although the mechanisms underlying this are unknown.

The impact of longstanding diabetes mellitus remains controversial. Although worsening or new onset diabetes without obvious cause increases the probability of diagnosing PDAC (OR: 7.9, 95% CI: 4.7-12.5), diabetes is more likely a result of PDAC rather than a direct cause. A recent meta-analysis (Ben, Xu et al. 2011) demonstrated an increase risk of PDAC in patients with type II diabetes (relative risk [RR]: 1.94, 95% CI: 1.66-2.27) however when stratified for duration of diabetes (< 1 year vs. 1-4 years vs. 5-9 years vs. >10 years), increased risk of PDAC was seen most profoundly in patients with short-onset disease (RR: 5.38, 95% CI: 3.49-8.30 vs. 1.95, 95% CI: 1.65-2.31 vs. 1.49, 95% CI: 1.05-2.12 vs. 1.47, 95% CI: 0.94-2.31). Although an association has been demonstrated between diabetes and PDAC, its exact role in tumorigenesis remains ill defined.

### 1.3.2 INTRINSIC RISK FACTORS

Intrinsic, patient specific risk factors are also important in cancer development. Advanced age, although not specifically genetic, increases risk of PDAC. The age adjusted rates for PDAC for <65 and >65 years of age are 4.0 and 68.3 cases per 100,000 respectively (Howlader, Noone et al. 2012). The median age of diagnosis is 71 years and patients older then 65 years account for 41% of cases. In contrast, individuals younger the 45 years of age represent 12% of patients with PDAC. Furthermore, race and ethnicity also impact PDAC risk. In the United States, Caucasians have an age adjusted incidence of 12.0 cases per 100,000 population in contrast to African and Asian Americans where the incidence is 15.8 and 9.5 cases per 100,000 respectively. Moreover, trends of increased risk with advanced age are preserved in these ethnic groups (Table 2).
Table 2: Age-Adjusted Incidence of Pancreatic Cancer by Race and Gender 2005-2009
(Howlader, Noone et al. 2012) (Reproduced with Permission

<table>
<thead>
<tr>
<th>Age Group</th>
<th>All Races</th>
<th>Caucasian</th>
<th>Blacks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>15-19</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20-24</td>
<td>0.1</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>25-29</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>30-34</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>35-39</td>
<td>1.0</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>40-44</td>
<td>2.5</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td>45-49</td>
<td>5.6</td>
<td>4.8</td>
<td>5.5</td>
</tr>
<tr>
<td>50-54</td>
<td>10.8</td>
<td>8.6</td>
<td>10.4</td>
</tr>
<tr>
<td>55-59</td>
<td>18.7</td>
<td>15.3</td>
<td>18.3</td>
</tr>
<tr>
<td>60-64</td>
<td>30.2</td>
<td>25.0</td>
<td>29.6</td>
</tr>
<tr>
<td>65-69</td>
<td>43.7</td>
<td>36.3</td>
<td>42.9</td>
</tr>
<tr>
<td>70-74</td>
<td>60.3</td>
<td>53.8</td>
<td>59.8</td>
</tr>
<tr>
<td>75-79</td>
<td>76.0</td>
<td>71.4</td>
<td>76.2</td>
</tr>
<tr>
<td>80-84</td>
<td>91.6</td>
<td>85.4</td>
<td>91.1</td>
</tr>
<tr>
<td>&gt;85</td>
<td>99.3</td>
<td>94.6</td>
<td>98.0</td>
</tr>
</tbody>
</table>

1.4 Molecular Biology and Genetics of PDAC

1.4.1 ADENOMA TO CARCINOMA SEQUENCE OF CANCER

In its simplest form, cancer represents the successive accumulation of key genetics events, which confer a selective advantage to a population of cells. The hallmarks or biological capabilities of cancer have been described as a cell’s ability to sustain a proliferate signal, evade growth suppression/cell death, confer replicative immortality, neo-angiogenesis and the ability of invade and metastasize (Hanahan, Weinberg 2011). A stepwise genetic model of carcinogenesis was first described in the colon and has since been referred to as the Vogelstein adenoma to carcinoma sequence of CRC (Fearon, Vogelstein 1990). In this model, invasive cancer (carcinoma) arises from a benign precursor (adenoma) that undergoes both sequential mutational activation of oncogenes, which promotes growth, and bi-allelic inactivation of tumor suppressor genes (TSGs) that prevents inhibition of the cell cycle.
Development of PC also demonstrates a stepwise progression involving the accumulation of key genetic events. Interestingly, recent studies have shown that within a primary PC, a collection of distinct subclones exist, each with a specific genetic signature (Yachida, Jones et al. 2010) originating from a source parental clone. Although, certain genetic events remain in common with the parental clone and between subclones, the distinct molecular signatures serve as “address tags” dictating a subclone’s future organ of metastasis. Mathematical modeling of these events demonstrate that >11 years are required from the first genetic event to the development of the parental clone, with another 6 years for the development of organ specific subclones. Similar genomic profiling of metastases demonstrates genetic events in common between tumours at all distant sites but also organ specific differences (Campbell, Yachida et al. 2010). Like primary disease, the tumour at a particular metastatic site also harbors subclones with distinct molecular signatures. The practical consequences of this genetic heterogeneity remain unknown. Nevertheless, these findings challenge the simplistic linear model proposed by Vogelstein and suggest that PDAC may in fact be a collection of diseases with common histology, progressing and continuously diverging in parallel.

1.4.2 PANCREATIC INTRA-EPITHELIAL NEOPLASM (PanIN)

Further evidence in support of a neoplastic precursor arises from studies demonstrating a cystic or intra-epithelial origin to PC, both of which display a distinct molecular pathway of development, clinical presentation, treatment strategy and prognosis. Pancreatic Intra-epithelial Neoplasms (PanINs) represent a varying structure of morphological and cytological atypia arising from the ductal cells and are classified on a scale from I-III (Matthaei, Molin et al. 2013). The earliest lesions, PanIN-I, generally consist of tall columnar cells, which may display a flat or micro-papillary structure with round basally located nuclei. PanIN-II lesions are mostly papillary with moderate cytological atypia involving enlarged nuclei, minimal loss of polarity and occasional mitoses. PanIN-III, also termed carcinoma in-situ, is the most advanced non-invasive precursor lesions prior to PDAC and contains a papillary structure with epithelial “budding” into the ductal lumen. These lesions have prominent nuclear irregularity, loss of polarity and abnormal mitoses however they remain confined to the basement membrane.

Though PanINs have been well characterized morphologically and histologically, and accepted as precursors to PDAC, their underlying genetic make-up in relation to the natural history of PC
remains an area of active investigation. PanIN-I lesions have been commonly reported in pancreatic specimens resected for non-oncologic reasons and, furthermore, the frequency of PanIN-I lesions increases with age without subsequent PDAC development (Hruban, Maitra et al. 2008). PanIN-II or III lesions are seen in up to 59% of specimens removed for PDAC (PanIN-III frequency of 40%) in comparison to 17% in non-cancer cases (Andea, Sarkar et al. 2003). Genetic events occur in a particular temporal sequence and the most common findings include: activating K-Ras mutations and telomere shortening in up to 92% of PanIN-I and in 96% of PanIN-II/III lesions whereas these findings are absent in normal ductal epithelium (Kanda, Matthaei et al. 2012, Hruban, Maitra et al. 2008). Activating mutations in BRAF are uncommon (<2% in PanIN-I and <5% in PanIN-III) and may cluster in K-Ras wild type cases. CDKNA2A/p16 loss is detected in up to 11.5% of PanIN-Ils but its frequency increases in high grade PanINs (56.4%). Loss of DPC4/SMAD4 is restricted to PanIN-III lesions and is reported at frequency of 31% (Wilentz, Iacobuzio-Donahue et al. 2000). Similarly, p53 expression is normal in low grade PanINs but loss at this locus is observed in approximately 40% of PanIN-IIIls. In-vivo models similarly corroborate these genetic findings as evidenced by the increased frequency of PanIN lesions (and subsequent PDAC) detected in the K-Ras mutant mouse, KRas$^{G12D}$ (Hruban, Goggins et al. 2000).

1.4.3 CYSTIC NEOPLASMS OF THE PANCREAS

Cystic disease represents <10% of pancreatic lesions (Brugge, Lauwers et al. 2004) and these lesions are separated histologically into inflammatory and neoplastic subtypes (Lennon, Wolfgang 2013). Inflammatory cysts, arising following complications from pancreatitis, lack a true epithelial lining and are referred to as “pseudocysts.” These lesions do not possess a malignant potential and are commonly excluded in the discussion of pancreatic cystic neoplasms.

NON-MUCINOUS:

Neoplastic cysts are further categorized into non-mucinous and mucinous. Non-mucinous cystic lesions include serous cystadenomas (SCAs) and solid pseudopapillary neoplasms (SPNs). SCAs are commonly diagnosed in females during the sixth decade of life and account for greater than one-third of pancreatic cysts. They originate mostly in the pancreatic body/tail displaying a microcystic honeycombing appearance and comprise a cuboidal, glycogen staining, epithelium.
Malignant degeneration to serous cyst-adenocarcinoma is uncommon (0-1.2%) and most individuals with SCAs are observed with treatment reserved for patients with symptomatic disease or where the diagnosis remains uncertain (Kimura, Moriya et al. 2012). As most cases are not resected, little is known genetically about SCAs other than genetic loss of chromosome 3p25 corresponding to the VHL locus and the absence of K-Ras mutations (Ishikawa, Nakao et al. 1998). In contrast, SPNs represent <10% of cystic neoplasms and occur mostly in females (>90%) before 50 years of age. These lesions originate commonly in the pancreatic body/tail, present with symptomatic disease (abdominal pain, jaundice etc.) and comprise both a solid and cystic component. SPNs demonstrate indolent growth, carry a borderline risk of malignant degeneration and demonstrate an excellent prognosis following surgery (Lennon, Wolfgang 2013). Few studies exist examining the genetics of SPNs. In a series of 20 cases (Abraham, Klimstra et al. 2002), frequent alteration of the APC/beta-catenin pathway and overexpression of cyclin D1 was observed. No mutations in K-Ras were found and approximately 80% of cases did not demonstrate a mutation in DPC4/SMAD4.

MUCINOUS:

Mucin producing lesions of the pancreas are the most well studied cystic tumours and include mucinous cystic neoplasms (MCNs) and intra-ductal papillary mucinous neoplasms (IPMNs). MCNs almost exclusively occur in females (90-98%) with a mean diagnostic age of 48.1 years (Yamao, Yanagisawa et al. 2011) and localize predominantly to the pancreatic tail (67-99%). Histologically, MCNs demonstrate a unilocular appearance with a characteristic ovarian like stroma. Malignant transformation to invasive mucinous cyst-adenocarcinoma has been observed in 12% of cases (Crippa, Salvia et al. 2008) with 57% 5-year survival following curative surgery. Molecular studies have been challenging in MCN largely due to its low incidence. Malignant MCN displays an increased expression of MUC1 and MUC2 on immunohistochemistry and genetic analysis reveals a varying degree of K-Ras and p53 mutations (Fukushima, Fukayama 2007). Hypermethylation of CDKN2A/p16 and genetic loss of the VHL gene has been reported in up to 17% of cases whereas no reports demonstrate alterations of the DPC4/SMAD4 gene.

IPMNs represent the most clinically significant cystic neoplasms of the pancreas affecting both sexes equally and predominantly in the 50s and 60s (Brugge, Lauwers et al. 2004, Lennon, Wolfgang 2013). IPMNs are most commonly diagnosed in the pancreatic head or uncinate
process and are often detected incidentally on imaging performed for reasons unrelated to the pancreatic mass (Sachs, Pratt et al. 2009). Risk of malignancy is ultimately contingent on the type of ductal involvement. Main duct IPMN (MD-IPMN) is associated with cancer in >40% of cases whereas branch duct IPMN (BD-IPMN) carries an approximate 16% risk of malignant degeneration. Therefore, resection is routinely recommended for all MD-IPMN whereas surgery is reserved for BD-IPMN with certain high-risk features as outlined by the Sendai Criteria (Tanaka, Chari et al. 2006). In a retrospective study spanning a 12-year period of 140 patients with malignant MD-IPMN who underwent curative-intent surgery, average overall 5-year survival was reported as 60% (Salvia, Fernandez-del Castillo et al. 2004). This is in sharp contrast to conventional PDAC where the minority of patients post curative intent operations are alive at 5 years. At a genetic level, recurrent mutations in K-Ras, and GNAS, overexpression of Her2/Neu and loss of chromosome 6q, 8p 17p and 18q are most commonly observed whereas DPC4/SMAD4 mutations are non-existent in malignancies arising from IPMNs (Wu, Matthaei et al. 2011, Hansel, Kern et al. 2003). Furthermore, the early detection of GNAS mutations in IPMNs may also predict future malignant degeneration. In contrast, cancers arising in the absence of an underlying IPMN do not harbor a GNAS mutation and this coupled with differences in outcomes, provides further evidence of a distinct biology between malignancies arising from cystic or intra-epithelial origin.

1.4.4 GENETICS OF SPORADIC PDAC

PDAC in the absence of a family history, termed sporadic PDAC, represents the majority of patients and molecular analyses of these tumors display a genetic progression from findings in PanIN lesions (Figure 3). Genetic instability is a hallmark of PDAC and includes several genetic, epigenetic and genomic structural events. High-resolution analysis of pancreatic tumours has illustrated a varied picture of genomic instability and inter-tumoral heterogeneity (Campbell, Yachida et al. 2010). This includes translocations involving multiple loci and a unique duplication with inversion at a chromosomal region. This phenomenon known as “fold back inversions” occurs in up to one-sixth of PDAC and likely arises from a repetitive cycle of breakage and fusion mediated by the ductal cell DNA repair machinery. Other structural changes include genomic gains and losses (termed Copy Number Variation-CNV) and a detailed discussion of this can be found in section 2 of this chapter.
Conventional sequencing studies have demonstrated that a single PDAC tumor contains on average 63 exonic genetic alterations, the majority of which are point mutations. Furthermore, genes involved in PDAC development can be classified into 12 core signaling pathways affecting 67-100% of tumors (Jones, Zhang et al. 2008, Yachida, Iacobuzio-Donahue 2013). Once again, activating mutations of K-Ras are the most frequent genetic abnormality occurring in >95% of cases at the codon 12 position. A single point mutation in one allele is sufficient to promote increased proliferation and apoptotic resistance confirming the oncogenic role of K-Ras (Lisiansky, Naumov et al. 2012). Other mutated oncogenes discovered in PDAC, however at much lower frequency (<15%), include BRAF, PI3/AKT, c-MyC and Her2/Neu (Abbas 2013).

The most common (95%) TSG affected in PDAC is CDKN2A/p16 located on chromosome 9p21.3 which functions in cell cycle control. CDKN2A is a cyclin dependent kinase inhibitor (CDKi) that prevents the phosphorylation of Rb-1 by CDK4 and 6, thus halting progression through the G1/S cycle in interphase (Hansel, Kern et al. 2003). Loss of this gene occurs by three mechanisms: homozygous deletion (i.e. simultaneous loss of homologous chromosomal arms in the tumor) in 40-45% of tumours, intragenic mutation with subsequent loss of the allele carrying the wild-type copy of the gene (loss of heterozygosity-LOH) in 40-45% or alternatively promoter hypermethylation and transcriptional silencing in 15%. The second most common TSG, p53, is involved in 50-75% of PDAC largely by single intragenic point mutations followed by LOH. The function of this gene is mostly as a “gate-keeper” of the cell cycle, promoting arrest in G1 and G2 following DNA damage. Loss, therefore, removes this checkpoint mechanism allowing the cell to continue mitosis despite the presence of altered DNA. The final “major player” of pancreatic tumorigenesis is DPC4/SMAD4. This TSG is located on chromosome 18q and similar to the aforementioned TSGs, SMAD4 functions in regulation of cell cycle control and specifically G1/S transition. SMAD4 is also involved in the TGF-beta signaling pathway and thus a regulator of cell growth. Alteration of this gene occurs in approximately 50-55% of cases commonly through homozygous deletion (30-35%) or a single intragenic point mutation with subsequent LOH. Interestingly, engineered KRas$^{G12D}$ transgenic mice develop PanINs and eventually PDAC faster as more of the above TSGs are conditionally knocked out providing further evidence of the causative and synergistic role of these genes in disease (Hruban, Maitra et al. 2008, Bardeesy, Cheng et al. 2006).
Despite the well-established role of these genes and the consistency of their findings among studies, all four of these genes are not always altered in all PDAC specimens. In a study of 70 autopsies, matched primary and metastatic disease underwent genetic analysis for K-Ras, CDKN2A, p53 and SMAD4 (Yachida, White et al. 2012). Only 37% of cases had an alteration of all 4 driver genes and furthermore, the number of genes altered correlated with patient survival. This would suggest that although these 4 genes are important in PDAC development, there are still yet to be defined genetic events important for tumorigenesis.

BRCA 1 and 2 are a set of TSGs that have an important role in double stranded DNA damage repair via homologous recombination and have been associated not only with the development of PC but also breast, prostate and ovarian cancer (Abbas 2013). Alterations at the BRCA locus occur usually in high-grade PanIN lesions or early invasive cancer most commonly by an intragenic point mutation with subsequent LOH (Greer, Whitcomb 2007). Other important however less common TSGs include: MAP2K4, STK11, EP300, ALK4, TGFBR1, TGFBR2, SMARCA2/4, ARID1A, ARID1B and PBRM1 (Abbas 2013, Shain, Giacomini et al. 2012).

Figure 3: Adenoma to Carcinoma Sequence of Sporadic PDAC (Iacobuzio-Donahue, Velculescu et al. 2012) (Reprinted by permission from the American Association for Cancer Research)

Epigenetic changes scattered throughout the genome have also demonstrated a role in PDAC development and may arise de-novo or inherited through successive generations. DNA promoter hypermethylation of CpG islands is the most well described epigenetic change and results in transcriptional silencing of the target gene. As previously discussed, DNA hypermethylation
accounts for 40% of alterations at the CDKN2A/p16 locus. Furthermore, NPTX2, SARP2 and CLDN5 have demonstrated hypermethylation in PDAC cell line models with an improvement in cell kinetics following the subsequent demethylation at these loci (Sato, Fukushima et al. 2003). These epigenetic changes appear in the earliest PanIN-I lesions and increase in frequency through PDAC development. Other identified but uncommon TSGs affected by DNA hypermethylation include: RARB (20%), CACNA1G (16%), TIMP3 (11%), CDH1 (7%) and MLH1 (4%) (Ueki, Toyota et al. 2000). In contrast, hypomethylation and the subsequent de-regulation of putative oncogenes have also been described. For example, loss of promoter silencing by demethylation at VAV1 leads to an increased activity of K-Ras and consequently increased cellular proliferation (Abbas 2013). More recently, non-coding RNAs (or micro-RNAs) have shown considerable importance in cancer biology and indeed PDAC is not an exception. Their role in maintenance of TSGs and oncogenes serves as another layer of control preventing cellular invasion, neo-vascularization, proliferation and epithelial to mesenchymal transition. One such well documented non-coding RNA would be miR-200 whereby overexpression leads to genomic hypomethylation, suppression of E-cadherin and further genetic instability.

One of the challenges encountered in novel gene discovery in PDAC is the overwhelming volume of genetic data generated; which complicates identifying genes involved with tumorigenesis (driver genes) from genes whose alterations are of no consequence (passenger genes). Conceptually, it is believed that a gene involved with disease causation occurs recurrently and non-randomly among tumors from different individuals; whereas passenger genes are scattered randomly throughout the genome and seldomly recur with a consistent frequency. One strategy has focused on grouping affected genes based on the major biological pathway rather than targeting individual genes, occurring at a varying mutational frequency. As previously discussed, Jones et al demonstrated that mutations in genes involved with PDAC can be lumped into 12 core major signaling pathways. Building off of these preliminary findings, a recent study employed next generation sequencing (NGS) and copy number profiling on 99 early sporadic PDAC specimens (Biankin, Waddell et al. 2012). Results from this study demonstrated tremendous genetic heterogeneity between tumors and once again re-affirmed the importance of the aforementioned major genes and core signaling pathways in PDAC. Interestingly, recurrent mutations and copy number events were detected in genes involved with axon guidance.
pathways including class 3 semaphorins (SEM3A), SLIT2 and ROBO 1-3. Furthermore, examination of expression changes of these genes in a murine model demonstrated increased metaplastic changes, ductal injury and neoplasia. The potential benefits of these and other pathway focused investigations is that future treatment strategies can be targeted globally to one or a collection of pathways as opposed to relying on gene specific therapies which may only affect a small subset of cancers.

1.5 Familial Basis of PDAC

Whereas the majority of patients (80%) develop sporadic disease, 20% of patients have an underlying family history of PDAC (Abe, Fukushima et al. 2007) with an autosomal dominant mode of inheritance. Familial aggregation of PDAC has been highlighted as a risk factor for disease in case-control studies (RR: 2.49, 95% CI: 1.32-4.69, (Schenk, Schwartz et al. 2001) and is further compounded by a smoking history (RR: 6.02, 95% CI: 1.98-18.29). More recent studies have demonstrated a 2.41 fold risk (95% CI: 1.04-4.74) for relatives of patients with sporadic PDAC and interestingly a 2.14 fold risk (95% CI: 0.58-5.49) in spouses of patients with PDAC (Klein 2013). Families with a history of PDAC are of considerable interest as it is presumed that an underlying germline mutation in a gene of interest segregates in affected members, leading to this increased risk. Furthermore, the study of familial associations in other cancer types, such as CRC, has also shown to inform genetic studies in sporadic cancer (Fearnhead, Wilding et al. 2002). Therefore, by extension, a better understanding of hereditary pancreas cancer genetics has likely beneficial implications relevant to individuals and relatives affected by disease and potentially sporadic cases (Klein 2013).

1.5.1 GERMLINE RISK PREDISPOSITION TO PDAC

Numerous studies have investigated the presence of germline variants, mutations and polymorphisms that may have a putative role in PDAC. Traditional linkage studies, although commonly used in family based gene identification studies, have been met with considerable challenges secondary to genetic locus heterogeneity and the rapidly fatal nature of PDAC precluding the collection of blood/tissue from affected members (Klein 2005). Genome wide linkage analysis using microsatellite markers in an atypical family with early onset (approximately 43 years), high penetrant (>80%) PDAC associated with pancreatic insufficiency
and diabetes demonstrated a linked region (Logarithm of Odds: 4.56) to chromosome 4q32-34 (Eberle, Pfutzer et al. 2002) reputed as the first major locus in PDAC. This locus corresponded to the Palladin gene. Subsequent independent studies, however, have been unable to replicate these early findings (Zogopoulos, Rothenmund et al. 2007, Klein, Borges et al. 2009).

Due to the scarcity and initial failures of linkage studies, several genome-wide association studies (GWAS) have been performed on unrelated sporadic PDAC cases. The first GWAS involved a comparison between 1,896 individuals with PDAC vs. 1,939 controls obtained from 12 cohort and 9 case-control studies mostly of European ancestry (Amundadottir, Kraft et al. 2009). A significant association was seen mapping to the chromosome 9q34 region involving the first intron of the ABO locus confirming previous epidemiological findings of blood type O as a protective factor in PDAC compared to non-type O (OR: 1.20, 95% CI: 1.12-1.28). This locus has subsequently been identified as significant in various case-control studies (Wolpin, Kraft et al. 2010, Iodice, Maisonneuve et al. 2010, Risch, Yu et al. 2010). Another GWAS comparing 3,851 individuals with PDAC vs. 3,934 controls from 12 cohort and 8 case-control studies adjusted for sex, age and ancestry demonstrated a statistically significant association for either an increased or decreased risk of PDAC at a non-genic region of chromosome 13q22.1 (OR: 1.21, 95% CI: 1.13-1.30), to the gene NR5A2 on chromosome 1q32.1 (OR: 0.77, 95% CI: 0.71-0.84) and the CLPTM1L-TERT locus of chromosome 5p15.33 (OR: 1.19, 95% CI: 1.11-1.27) (Petersen, Amundadottir et al. 2010). The importance of these findings, however, is still under investigation. A follow-up pathway analysis of the major GWAS data in PDAC demonstrated a significant association between PDAC and 5 pathways (and their corresponding genes): pancreatic development (NR5A2, HNF1A, HNF4G, PDX1), Helicobacter Pylori lacto/neolacto (ABO), hedgehog signaling (SHH), apoptosis (MAPK8, BCL2L11) and Th1/Th2 immune response (TGFB2, CCL18) (Li, Duell et al. 2012). These findings also provide further evidence that pathway based analysis and related studies can assist in identifying and filtering genes generated from data in single and pooled studies.

GWAS based approaches are useful for gene identification in diseases where locus heterogeneity common in complex disorders significantly hinder more traditional segregation based linkage methods. Furthermore, GWAS studies rely on the robustness of the common disease/common variant principle whereby alterations at specific genes have a small to moderate effect size but are present at a relatively high frequency (>1-5%) in the population (Hemminki, Forsti et al.
Alternatively, the common disease/rare variant principle postulates that common complex disorders arise from variants present at a relatively low frequency (<1%) however with high penetrance or effect size. Disease based on this model possessing allelic heterogeneity would not be suitable for GWAS based study designs. Credence to this new hypothesis has largely developed in recent years due to the challenge in replicating GWAS findings and the fact that GWAS data accounts for only a small proportion of the heritable changes behind complex disease (Gibson 2012). Furthermore, genetic analyses of syndromic cases have demonstrated that disease segregates in those individuals with a rare allele variant of high penetrance. Although the true genetic model for a complex disease remains controversial, genetic investigations in patients belonging to syndromic PDAC families may further explain the current missing heritability.

1.5.2 HEREDITARY PANCREATIC CANCER SYNDROMES

In addition to individuals with family histories of PDAC, 5-10% of PC patients belong to an underlying familial syndrome, which substantially increases the lifetime risk of cancer (Klein 2012). Collectively, this group is referred to as the Hereditary Pancreatic Cancer Syndromes (HPC) (Figure 4) and includes: Hereditary Non-Polyposis Colon Cancer (Lynch Syndrome), Familial Atypical Malignant Mole Melanoma (FAMMM), Hereditary Breast Ovarian Cancer (HBOC), Peutz-Jeghers Syndrome (PJS), Hereditary Pancreatitis (HP) and Familial Pancreatic Cancer (FPC).

HEREDITARY NON-POLYPOSION COLON CANCER (LYNCH SYNDROME):

Lynch Syndrome (LS) historically is the most common CRC predisposition syndrome arising in 1/500-1/1000 individuals (Shi, Hraban et al. 2009). Germline mutations in the mis-match repair (MMR) genes are the cause of LS and include MLH1, MSH2, MLH6 and PMS2 (Lynch, Lynch et al. 2009). Diagnosis is usually suspected in individuals with early onset colorectal cancer who satisfy the requirements set forth by the Bethesda or Amsterdam Criteria. These patients undergo testing for tumour microsatellite instability (MSI) and individuals with high frequency tumor microsatellite instability (MSI-H) then undergo further germline testing for MMR mutations with immunohistochemistry to pinpoint the gene, followed by gene sequencing.

Aside from the inherent CRC risk in individuals with Lynch Syndrome (LS), several extracolonic manifestations also occur with a collective lifetime frequency of approximately 10%
For example, females with germline MMR mutations are at an increased risk of developing endometrial cancer (20-60%) and regular screening commencing at the age of 30-35 years is recommended (Lynch, Lynch et al. 2009). Other extra-colonic manifestations include gastric, renal pelvis/ureteric, and rarely brain and small bowel. Patients with LS are also at increased risk of developing PDAC (OR:3.68, 95% CI: 1.45-5.88).

**FAMILIAL ATYPICAL MALIGNANT MOLE MELANOMA (FAMMM):**

FAMMM is an autosomal dominant syndrome characterized by a family history of melanoma and multiple melanocytic nevi (Eckerle Mize, Bishop et al. 2009). The underlying causative mutation is the CDKN2A/p16 gene. In previous epidemiological studies, 5-12% of patients with melanoma had an underlying hereditary basis of which 40% were accounted for by a mutation in CDKN2A. Furthermore, patients with a history of melanoma were also at increased risk of developing PDAC (OR: 47.8, 95% CI: 28.4-74.7, (Shi, Hruban et al. 2009) and subsequent genetic studies attributed this to a germline mutation in CDKN2A (Lynch, Brand et al. 2002). This association is now referred to as the FAMMM-PC carcinoma syndrome.

**HEREDITARY BREAST OVARIAN CANCER (HBOC):**

Individuals in families with HBOC report a history of multiple members affected with breast and/or ovarian cancer or the onset of breast cancer prior to 50 years of age (Shi, Hruban et al. 2009). Linkage and segregation studies have identified germline mutations of the BRCA1 and BRCA2 locus as causative in this syndrome. Individuals belonging to HBOC syndrome families carry an elevated risk of developing PDAC. Carriers of germline BRCA1 mutations have a 2.26 (95% CI: 1.26-4.06) increase whereas germline carriers of BRCA2 mutations have a 3.5 fold risk (95% CI: 1.87-6.58) of developing PDAC. Interestingly, patients without BRCA mutations in high risk breast/ovarian cancer families demonstrate a 1.30 fold risk of developing PDAC suggesting that yet to be identified mutations account for a proportion of cases (Mocci, Milne et al. 2013).

The likelihood of carrying an underlying germline BRCA mutation in patients with PDAC increases as the number of relatives affected with PDAC increases. In a study of unselected patients with PDAC, 26% of patients were classified as intermediate to high risk of belonging to a familial syndrome as based on family history. Of these patients, 10.5% carried a germline
BRCA1 or BRCA2 mutation (pathogenic non-sense or frameshift) leading to a truncated protein (Lal, Liu et al. 2000). In similar reports, a family history of >1 first-degree relative (FDR) with PDAC conferred an individual likelihood of 6-12% of harboring a germline BRCA 2 mutation (Couch, Johnson et al. 2007).

**PEUTZ-JEGHERS SYNDROME (PJS):**

PJS is a relatively uncommon autosomal dominant, high penetrance syndrome, occurring in 1/150,000 live births. It is characterized by the presence of pigmented macular lesions on the gums, oral mucosa and skin as well as multiple hamartomatous polyps of the gastro-intestinal tract, lungs and bladder (Zbuk, Eng 2007). Greater than 80% of PJS has been attributed to an underlying germline mutation of STK11/LKB1 located on chromosome 19. Patients with PJS have a significantly elevated risk of developing many cancers and previous studies have reported that 93% of individuals will develop a cancer by age 65 (Giardiello, Brensinger et al. 2000). The most common cancer sites include: small intestine (RR:520), stomach (RR: 213), colon (RR:84), esophagus (RR:57) and breast/lung/ovary (RR:>10). Indeed, individuals with PJS also have a significantly elevated risk of developing PDAC (RR:132, 95% CI: 44-261).

**HEREDITARY PANCREATITIS (HP):**

Patients with HP are characterized by repeated attacks of pancreatitis in the absence of other definable causes (e.g. gallstones, significant alcohol consumption, infections). Repeated attacks eventually result in chronic exocrine and endocrine insufficiency. HP is a rare inherited autosomal dominant disorder (prevalence: 0.3/100,000) characterized by a germline Arginine to Histidine substitution at codon 117 of the cationic trypsinogen gene (PRSS1) on chromosome 7q35 (Whitcomb, Gorry et al. 1996). Individuals with HP have a 53-fold risk (95% CI: 50-60) of developing PDAC after the age of 50 years (Shi, Hruban et al. 2009). Furthermore, a personal history of smoking acts synergistically with this elevated baseline risk of PDAC and these individuals develop cancer 20 years prior to non-smoking patients with HP.
1.6 Familial Pancreatic Cancer (FPC)

1.6.1 EPIDEMIOLOGY OF FPC

The majority of patients (>80%) with PDAC and an extensive family history do not cluster into one of the aforementioned syndromes. Mathematical modeling from genetic and epidemiological data on individuals and affected relatives without a known underlying cause for PDAC has demonstrated an autosomal dominant mode of inheritance for a rare major allele modulating the risk of pancreatic cancer (Klein, Beaty et al. 2002). This novel syndrome is referred to as Familial Pancreatic Cancer (FPC) and is defined by the presence of 2+ affected FDR with PDAC in the absence of an underlying known HPC syndrome (Figure 5). When compared to individuals in the general population, the average age of onset of patients in a FPC kindred is 6 years younger (70 years vs. 64 years) with an elevated baseline risk of 6.79 (95% CI: 4.54-9.75, p<0.001) for developing PDAC (Brune, Lau et al. 2010). Risk of disease appears to be related, not surprisingly, with the number of affected relatives with disease. Individuals in families with >3 FDRs possess a 17.02 fold risk (95% CI: 7.34-33.5) compared to families with 2 FDRs.
(standardized incidence ratio [SIR]: 3.97; 95% CI: 1.59-8.2) or a single FDR with PDAC (SIR: 6.86; 95% CI: 3.75-11.04). Furthermore, when stratifying families for early vs. late onset PDAC as defined by a relative with disease prior to 50 years of age, the early onset cohort possess a higher risk for cancer (SIR: 9.31, 95% CI: 3.42-20.28 vs. SIR: 6.34, 95% CI: 4.20-9.51; p<0.001). When examining pre-established environmental risk factors in the context of individuals in FPC families, smoking once again further elevates the risk of developing disease (OR: 3.7, 95% CI: 1.8-7.6) and this is more significant in male subjects (OR: 5.2) and those younger than 50 years (OR: 7.6) (Rulyak, Lowenfels et al. 2003).

Figure 5: A pedigree depicting a typical FPC Family

1.6.2 FPC AND THE ROLE OF MULTI-CENTER COLLABORATION

Over the last decade, considerable efforts have been placed on characterizing FPC kindreds. Although tremendous insights have been obtained in the epidemiology of FPC, ongoing uncertainties exist regarding its underlying biology and molecular changes. In light of the rapidly fatal nature of PDAC and the documented genetic heterogeneity of sporadic PDAC, it has now become obvious that further advancements in the field will be facilitated with the assistance of streamlined dedicated efforts from multiple institutional collaborations. One such consortium is the Pancreatic Cancer Genetic Epidemiology Consortium (PACGENE), which is dedicated to the
identification of susceptibility genes in FPC (Petersen, de Andrade et al. 2006). PACGENE comprises 7 centers across the United States and Canada where individuals are registered based on the modified definition of FPC (an individual with a history of >2 blood relatives with PDAC). Other cancer based consortia include the International Cancer Genome Consortium (ICGC) that has an overarching goal of performing large scale genomic studies of over 50 cancer types (International Cancer Genome Consortium, Hudson et al. 2010). As outlined by the ICGC, such global consortia have the added benefit of rapid dissemination of data sets and analytical methods, duplication and validation of consolidated efforts and the standardization of studies. Specifically in PDAC, of which a small subset of cases are from FPC kindreds, the ICGC strives to sequence, using next generation technology, the genomes of 750 cases in order to characterize both structural variation and changes at the base pair level. Furthermore, the data catalogued by these large-scale organizations will act as references for independent studies performed at various institutions serving a public domain for genomic data validation in FPC.

1.6.3 THE GENETICS OF FPC

Earlier genetic investigations in FPC focused on candidate genes whose loci had been described in sporadic PDAC. The role of germline mutations in MAP2K4, DPC4/SMAD4, ACVR1B and BRCA2 were analyzed in a cohort of 31 FPC patients representing 29 kindreds enrolled in a national family registry at an American institution (Murphy, Brune et al. 2002). Individuals were selected based on a family history of >3 affected relatives with PDAC of which >2 were FDRs. This study demonstrated that 17.2% of cases had a germline mutation in BRCA 2 whereas no mutations were seen in the other 3 genes. This is in comparison to previous reports prior to this study whereby 7% of unselected individuals with sporadic PDAC had a germline BRCA2 mutation (Goggins, Schutte et al. 1996). Furthermore, individuals with germline BRCA2 mutations had a similar age of onset of PDAC as reported for patients with sporadic disease. More recently, the prevalence of germline BRCA 2 mutations was investigated in a cohort of FPC patients of European ancestry (Slater, Langer et al. 2010). In this study, a mutational range of 2.8-11.4% was reported when germline BRCA2 sequencing data of 56 patients were combined with previous reports from this institution. Pundits of this study challenged its conclusions based on the proportion of patients who were truly high-risk. It is therefore possible that the lower prevalence of germline mutations may have been accounted for by this difference.
Nevertheless, this and previous studies have demonstrated that germline BRCA2 mutations are the most common genetic alterations seen in FPC and these mutations occur without a family or personal history of breast and ovarian cancer.

The role of BRCA1 has similarly been investigated in FPC. As previously discussed, individuals in HBOC families with an underlying germline BRCA1 mutation have an elevated risk of developing PDAC. In an investigation performed at a high volume Canadian academic center, patients with a family history of breast or ovarian cancer or who presented with an incident PDAC in the absence of an underlying family history, underwent germline BRCA testing (Al-Sukhni, Rothenmund et al. 2008). A total of 7 patients with PDAC and a family history of disease and 9 patients with PDAC and no family history of PDAC tested positive for a BRCA1 germline mutation. Subsequently analysis for LOH in the tumours of these patients demonstrated loss in 71% of cases with a family history of PDAC (5/7) and in 11% of sporadic cases (1/9). Further sequencing analysis of the retained allele demonstrated a confirmed loss of the wild-type allele in 60% of the familial cases (3/5). Thus, the conclusions from these findings were that a predisposition to PDAC is seen in patients belonging to familial kindreds. However, one limitation of this study was that it was not specific to FPC kindreds. In a recent study of 66 patients with PDAC adhering to the modified definition of FPC, full sequencing analysis of germline tissue was performed for BRCA1 mutation carrier status (Axilbund, Argani et al. 2009). Results from this study demonstrated no known mutations in BRCA1 although 3 variants of unknown significance (VUS) were documented and these patients all developed PDAC after the age of 60 years. This is the largest study to date reporting no involvement of germline BRCA1 mutations in cancer development in PDAC.

In the study by Murphy et al., no germline mutations of DPC4/SMAD4 were identified in FPC kindreds (Murphy, Brune et al. 2002). Previous studies have described a Juvenile Polyposis Syndrome (JPS) associated with inherited germline mutations of SMAD4 in approximately one third of cases (Moguelet, Plassa et al. 2004). JPS is a rare autosomal dominant disorder characterized by the presence of multiple hamartomatous polyps of the gastro-intestinal (GI) tract. Furthermore, individuals with JPS have an elevated risk of various malignancies such as colon (RR:34) and rarely pancreatic cancer, likely mediated through biallelic inactivation of SMAD4. Similarly, inactivation of CDKN2A is common in sporadic PDAC and frequently associated in patients with PDAC and either a familial or personal history of melanoma. In a
study of 18 FPC families and 5 families with at least one relative with melanoma and PDAC, no germline mutations in CDKN2A were detected in the absence of melanoma, whereas 40% of patients (2/5) with a history of both melanoma and PDAC had a germline CDKN2A mutation (Bartsch, Sina-Frey et al. 2002). In another report of 28 FPC families, 6 kindreds (21%) tested positive for an underlying CDKN2A mutation without a personal or family history of melanoma prompting the authors to recommend CDKN2A counseling for genetic testing in individuals belonging to FPC kindreds (Harinck, Kluijt et al. 2012). These results, however, are yet to be replicated and therefore, currently, germline CDKN2A mutations are not associated with PDAC in FPC kindreds in the absence of melanoma.

With improvements in computational biology and sequencing technologies, the ability to analyze whole genomes or a portion thereof have become efficient and accessible (Please refer to section 3 for a detailed discussion of NGS). These advancements have also impacted the field of FPC genetics. Exome sequencing (the high-throughput sequencing of all protein coding regions in DNA) of a single patient with PDAC belonging to a FPC kindred demonstrated over 15,000 unique germline variants not previously described in a reference database (Jones, Hruban et al. 2009). Of these variants, 64 were predicted as non-sense and thus mutation led to the insertion of a premature stop codon and ultimately a truncated protein. Filtering of these genes based on the presence of a germline non-sense variant with a second genetic event in the tumor identified 3 genes (RAGE, SERPINB12 and PALB2). The authors highlighted PALB2 as the most relevant based on biological function and from anecdotal studies describing the presence of germline RAGE and SERPINB12 mutations in healthy controls. PALB2 (also referred to as FANCN) belongs to a family of genes involved in Fanconi Anemia, which also includes BRCA2. PALB2 is involved with the homologous recombination mechanism of dsDNA repair and is a co-localizer with BRCA in the nucleus. The co-localization of PALB2 stabilizes BRCA2 and provides a synergistic effect in DNA damage repair.

A follow-up analysis in 96 FPC cases (16 patients had >1 FDR and 80 patients with >2 FDR) demonstrated truncating mutations in 3.1% (3/96) of patients. Interestingly, in each case, the truncating PALB2 mutation was different than the original variant described by exome sequencing in the index analysis. The authors of this study thus concluded that PALB2 is the second most common gene affected in FPC following BRCA2. Although this concept is currently accepted, follow-up reports have been controversial as evidenced by a Dutch study of
64 patients from 56 non-BRCA FPC kindreds who did not display any germline PALB2 mutations prompting the exclusion of this gene from routine testing at this institution (Harinck, Kluijt et al. 2012). Furthermore, an Italian study of 29 patients with PDAC and either a personal or familial history of breast/ovarian cancer demonstrated no germline PALB2 mutations (Ghiorzo, Pensotti et al. 2012). Finally, germline PALB2 variants were investigated in 81 FPC cases and a mutation frequency of 3.7% (3/81) was observed (Slater, Langer et al. 2010). Although these studies all varied in their inclusion criteria for FPC kindreds, the conclusions from these reports challenge the importance of PALB2 mutations as causal in PDAC.

Although BRCA2 and PALB2 have been the focus of most studies in FPC, the identification of ATM as a putative causative gene in FPC has garnered attention. ATM is a serine/threonine protein kinase located on chromosome 11q22.3, which is involved in cell cycle delay and DNA repair following dsDNA damage. Bi-allelic inactivation of ATM is the cause of Ataxia Telangiectasia (AT) characterized by cerebellar degeneration, immuno-deficiency and hereditary predisposition to cancer (Lavin 2008). Recently, whole genome sequencing (WGS) of 16 patients from 6 FPC kindreds and whole exome sequencing (ES) of 22 subjects from another 10 FPC kindreds was performed. Families had at least 3 FDRs affected with cancer and germline tissue was available on at least 2 of these members (Roberts, Jiao et al. 2012). Findings from this study demonstrated no mutations in known FPC or PDAC predisposition genes. Candidate germline genes were filtered based on the presence of a rare heterozygous deleterious variant (non-sense, splice-site or frameshift insertion/deletion polymorphism) demonstrating segregation among affected family members. Based on these criteria, 156 candidate genes were highlighted of which ATM was identified as the most interesting and biologically relevant. Validation with an independent cohort of 166 FPC cases and 190 spouse controls demonstrated a significant difference in mutational frequency in ATM (4/166, 2.4% VS. 0/190, 0%; p=0.046) in FPC cases vs. controls. In individuals with an extensive history of PDAC as defined by >3 affected members, 4.6% (4/87) of cases had a mutation in ATM. The analysis of tumor tissue for one FPC case with a germline non-sense variant demonstrated LOH with retention of the mutant allele. The authors therefore have concluded that a small proportion of FPC can be explained by germline mutations of the ATM gene, which may have future counseling and therapeutic implications.
Once again, these findings have been challenged by a recent Canadian study (Grant, Al-Sukhni et al. 2013) of 11 cases from 5 FPC kindreds who underwent germline ES. A novel rare variant in ATM was detected in one case whereas a novel PALB2 variant was detected in another. These variants were not detected in an independent group of 47 FPC probands, 90 in-house controls or various public databases. Interestingly, in the proband with a germline ATM variant, the variant was not present in an affected younger sibling. Sequencing analysis of a metastatic lesion in the proband demonstrated LOH at the ATM locus with retention of the wild-type allele. In the other FPC kindred with a rare PALB2 variant, once again the variant of interest was not detected in the germline tissue of an affected cousin. Further sequencing analysis of the primary tumor for the proband demonstrated LOH at the PALB2 locus, however, similar to the ATM family, retention of the wild-type allele was observed. Although there are limitations to this analysis, they do challenge the findings of ATM and PALB2 as causative genes in FPC and suggest that further evaluation is required.

Despite the numerous aforementioned efforts, the majority of genetic changes causing FPC remain unknown. Therefore, in light of the dismal prognosis of patients with PDAC coupled with the lack of effective treatments, future studies unraveling the molecular framework and biology of FPC and ultimately PDAC are desperately needed.
2 Copy Number Variation

2.1 CNV Detection Methods

Over the last decade, there has been a robust collaborative relationship between advances in human genetics, technology and computational methodology with each field propelling further insights into the other (LaFramboise 2009). Similarly, the discoveries of CNVs and particularly small-scale changes affecting a few hundred bps have also been more common with advances in detection methods. A number of platforms have emerged varying in genomic resolution capabilities and coverage of the human genome. Initial CNV detection was reliant on low resolution karyotyping and FISH based methods restricting detection to DNA segments >500kb. The identification of CNVs currently occurs with the use of DNA microarray technologies and further discussion will focus predominantly on this strategy.

DNA microarrays were initially developed for the purpose of gene expression studies or SNP genotyping. These platforms were created in a microminiaturized fashion allowing for data to be collected in parallel (Heller 2002). Microarrays consist of specific DNA probes either as synthetic oligonucleotides or large DNA fragments which are bonded to either a solid or substrate based surface. The density of an array is essentially the number of DNA probes and how well covered the probes are for the entire genome or, in customized arrays, for the region of interest. The essential principles of detection are based on complementary hybridization of probe and target DNA by hydrogen bonding (LaFramboise 2009). The affinity of hybridization is thus reliant on the extent of complementarity between the probe and target and the amount of source DNA. Probe-target hybridization is observed and quantified by the use of fluorescence labeling or chemi-luminescent techniques.

Microarray technology is divided into 2 major platforms: array comparative genomic hybridization (aCGH) and SNP microarray. The earliest CGH platforms relied on metaphase chromosomes serving as genomic probes with CNVs mapped to the physical location of the chromosome (Pinkel, Albertson 2005). These methods were initially chromosome or region specific and later extended to whole genome assays albeit at a low resolution (10Mb) (Davies, Wilson et al. 2005). CNV detection and genotyping with aCGH relies on the comparison between two samples, usually a test and reference, differentially labeled with a fluorescent marker. The signal intensity ratio is calculated between the two samples based on this fluorescent
output and is then computed into an integer based copy number state (Alkan, Coe et al. 2011). It should be mentioned that central to the analysis of aCGH is the proper characterization of reference genomes since CNVs calls are ratio based. Genomic probes on preliminary aCGH consisted of bacterial artificial chromosome (BAC) clones or long oligonucleotide sequences spaced at an average marker interval of 1.4Mb, which improved CNV detection resolution to >100kb when compared to conventional CGH methods (Davies, Wilson et al. 2005). However, one limitation of BAC arrays was the poor characterization of CNV breakpoints and thus the earliest studies grossly over-estimated the true size of CNVs and thus the gene content associated with these regions. Moreover, the synthesis of high quantity pure BAC DNA for array construction was quite challenging thus limiting the utility of this method. Current aCGH platforms rely on short oligonucleotide sequences, generally 25 nucleotides in length, tiled across the entire genome. This method allows for custom designing and high probe density arrays, which have been of great importance in clinical diagnostics and karyotyping. Furthermore, tiling aCGH is able to identify gains and losses at the gene level by virtue of its higher density, greater genomic coverage and marker overlap creating sampling redundancy (Davies, Wilson et al. 2005).

The second platform used for CNV detection once again employs short 25bp oligonucleotide markers spaced with varying density and coverage across the genome. These arrays rely on the a-priori characterization and annotation of SNPs throughout the genome and on linkage disequilibrium (LD): the finding that a group of SNPs belong to particular haplotype blocks that segregate at a high frequency more often than would be expected randomly (Takeuchi, Yanai et al. 2005). Hybridization on SNP arrays is performed on an individual sample basis and CNV signals are obtained by aggregating, normalizing and averaging clusters of probe intensities. Earliest SNP arrays relied on the arbitrary notation of an “A” allele and “B” allele representing a particular SNP in a diploid state (LaFramboise 2009). Therefore, an individual may have one of three possibilities at a particular locus (AA, AB, BB). By extension, one of the key features of SNP arrays is the ability to probe CNVs on an allele specific basis that ultimately provides a heightened sensitivity for CNV detection. A particular SNP on an array may have a number of probes varying in DNA sequences around the SNP and also shifting the SNP position on the probe. As such, probe sets include a number of perfect match (PM) and mismatch (MM) probes for each allele (PM_A, MM_A etc.) and these probes have varying level of hybridization to test
DNA. Test DNA is denatured into a single stranded form, shattered and PCR amplified prior to array hybridization. By virtue of being a PCR dependent process, one of the limitations of SNP arrays are the higher quantity of DNA required for effective data processing and the low signal to noise ratio generated by a ratio-based analysis compared to conventional aCGH platforms. This is counter-balanced, however, by the ability of SNP arrays to detect allele specific CNVs and by extension, copy neutral or gain changes (such as uniparental or acquired uniparental disomy-UPD vs. aUPD) and detection of LOH which would not be possible on a intensity based ratio analysis performed on aCGH.

CNV data is obtained from SNP arrays by log-transformed based ratios from probe intensities (Log R Ratios-LRR) and furthermore from genotyping data which facilitates allele specific analysis denoted as a B-Allele Frequency (BAF). BAF is defined as the ratio of B allele intensity in relation to the total allelic intensity. The ability to infer a copy number state from intensity and genotype based raw data requires the development of a series of computational models. These models are based on statistical theories, which hypothesize a predicted copy number state at a particular locus. Hidden Markov Models (HMM) are perhaps the most common statistical algorithms employed for CNV detection and rely on translating intensity data between neighboring SNPs in a sample and between samples into a copy number (CN) state (Scharpf, Parmigiani et al. 2008). Briefly, HMMs are a form of Gaussian analysis where the variable of interest (CN at a particular locus) is not visible. However, probe intensity (which is dependent on the CN state) is known and from the intensity of the probe just previous and subsequent to the region of interest, known as a Markov Chain, a CN state is inferred. The other major class of CNV calling algorithm is non-Markov based and relies on the principles of Circular Binary Segmentation (CBS) (Olshen, Venkatraman et al. 2004). In this model, copy states are inferred to be discrete and to occur in contiguous regions of the chromosome affecting multiple markers. Furthermore, as array data can be noisy, CBS divides or segments a chromosome into regions of equal CN accounting for noise. Although numerous studies have attempted to outline a superior CNV calling model, this is currently unclear and both models are used for the analysis of SNP microarray data (Staaf, Lindgren et al. 2008).

The main advantages of both aCGH and SNP array technologies are the ability to infer copy number profiles in a massively parallel fashion with relatively low cost. This is of tremendous importance particularly in common complex disorders presumed to follow a rare variant
hypothesis where power requirements mandate the analysis of a large quantity of samples simultaneously. As such, arrays have been commonly used in GWAS based designs whereby these advantages are maximally exploited (Alkan, Coe et al. 2011). Furthermore, fine break point mapping of current array designs to a resolution of 2kb, allows for accurate CNV detection and characterization. Microarrays, however, are restricted in their ability to accurately detect CNVs at the genome wide level at <10kb in size (McCarroll, Kuruvilla et al. 2008). Furthermore, accurate CNV mapping is dependent on appropriate annotation in available databases and proper characterization of reference genomes. Finally, by design, array based CNV detection methods are limited in their ability in detecting balanced changes, more commonly reserved for FISH based methods, small copy number gains and between tandem and interspersed duplications.

As described, tremendous progress has been made in CNV detection methods from preliminary FISH technologies and spectral karyotyping to array based methods allowing for fine mapping at the genome wide level. Future CNV detection may rely on DNA “barcoding” methodologies with the potential for detecting high-throughput balanced structural changes (Alkan, Coe et al. 2011). Moreover, absolute CN estimation via the amplification of single molecules is demonstrating potential with the use of emulsion picoliter droplet PCR. With the routine use of NGS, it is anticipated that future CNV calling methods will be based on these platforms. Although WGS allows for the possibility of CNV detection, its use, currently, is largely limited by cost, the lack of accepted statistical calling algorithms and challenges with mapping bias, quality control measures of reads and identification of duplications (Teo, Pawitan et al. 2012).

2.2 CNVs and Human Disease

CNVs are an important form of genotypic variation and their frequency would suggest that they account for a proportion of human phenotypic diversity. It is, therefore, not unreasonable to hypothesize that CNVs may also serve a role in disease susceptibility and development. Earlier studies in healthy individuals have already demonstrated that a small number of CNVs cluster in coding regions of known disease causing genes or at unstable genomic locations that commonly undergo rearrangements and translocations. Multiple CNV studies have been performed in the neuro-cognitive conditions such as Autism Spectrum Disorders (ASD). Linkage analysis of >1100 autistic families on SNP arrays demonstrated a linked CNV at chromosome 11p12-13 with the neurexin gene highlighted as a susceptibility locus of interest (Autism Genome Project
Consortium, Szatmari et al. 2007). In a follow up microarray study of 427 unrelated ASD cases, 277 unbalanced CNVs not present in >1100 controls were detected (Marshall, Noor et al. 2008). The majority of these CNVs were inherited however 7% of CNVs were de novo. 13 recurrent regions shared between cases confirmed the presence of neurexin but as well genes such as SHANK3 and ANKRD11 as ASD susceptible loci. Interestingly, few of the CNVs detected were also present in other mental retardation syndromes suggesting a common basis for cognitive impairment in these disorders.

Rare CNVs, either inherited or de novo, defined as present in the population at a frequency of <1% have also been associated with other disorders such as pediatric multiple sclerosis (McElroy, Krupp et al. 2012), schizophrenia (Van Den Bossche, Strazisar et al. 2013), attention deficit hyperactivity disorder (Jarick, Volckmar et al. 2012, Chapman, Rees et al. 2013) and Alzheimer’s disease (Chapman, Rees et al. 2013) among many others. Although most studies detailing rare CNVs have focused on neuro-psychiatric conditions, CNVs also been observed in diseases such as sporadic congenital heart disease (Soemedi, Wilson et al. 2012) and they affect the burden of disease in patients with Tetralogy of Fallot (Silversides, Lionel et al. 2012). Rare CNVs in obesity now appear to account for a relevant proportion of “missing heritability” not accounted for by common susceptibility variants from GWAS studies and may have future diagnostic potential (D'Angelo, Koiffmann 2012). Recently, a GWAS of 19,000 cases investigating 8 common complex disorders such as rheumatoid arthritis, Crohn’s disease, type 1 and 2 diabetes demonstrated that common CNVs (>1% population frequency) frequently segregate in affected individuals although these CNVs are tagged to SNPs already identified in these conditions (Wellcome Trust Case Control Consortium, Craddock et al. 2010). Furthermore, the relative paucity of common known CNVs would imply that these common CNVs would have to be shared among many common disorders. The authors argue that this is inconsistent with current findings, and therefore it is unlikely that common CNVs account for a sizeable portion of genetic changes in common disorders. These findings along with studies detailing rare CNVs provide further impetus to the common disease/rare variant hypothesis.

Visible phenotypes from common complex disorders are dependent on a complex interplay between genetic susceptibility and environmental triggers. Structural variation, including CNVs, account for a small but potentially significant portion of disease susceptibility. Although CNVs are associated with numerous common disorders, their role in “driving” disease still remains
unknown. In a study of gene expression, the fraction of expression variation attributed to SNPs and CNVs was found to be 83% and 18% respectively (McCarroll, Altshuler 2007). Further limitations in correlating CNVs to disease phenotype arise from challenges in accurate CNV measurement and discrepancy between predicted CNV status and actual DNA variation present (i.e. CNV discovery vs. CNV genotyping). Moreover, penetrance and variable expressivity of CNVs coupled with modifier effects have yet to be well characterized and complicate disease association in both traditional linkage and large-scale GWAS studies (Lee, Scherer 2010). Finally, somatic mosaicism, or the presence of cells with a variable genetic make-up in the same tissue, is common in health and disease and this further complicates the study of CNVs with particular relevance in cancer genomes (Piotrowski, Bruder et al. 2008, Lee, Scherer 2010).

2.3 CNVs and Human Cancers

The role of CNVs in both sporadic and hereditary cancer development have been investigated in various cancer subtypes. From the findings in non-oncologic diseases as described, it is presumed that rare CNVs may impact important cancer associated genes and pathways which may account for an increased predisposition to cancer in moderate to high risk families. In a study of 382 genomes representing 116 case-parent trios and unaffected siblings, de-novo CNVs of individuals with early onset testicular cancer were found in 7% of affected individuals and undetected in unaffected siblings (Stadler, Esposito et al. 2012). Interestingly, in comparison to a cohort of sporadic breast and CRC, no de-novo CNVs were detected suggesting that de-novo changes may be more relevant for early onset disease. It has been reported that over 100 highly penetrant germline mutated genes involved in hereditary cancer predisposition have been identified and furthermore these genes are inherited by a classical Mendelian pattern (Krepischi, Pearson et al. 2012). CNVs affecting these genes have also been described albeit at a much lower frequency (<5%). A study of 1903 Caucasians with prostate cancer identified 2 novel loci. The first locus is associated with disease at a non-coding enhancer element on chromosome 15q21.3 and the second maps to the MGA4TC gene on chromosome 12q21.31 (Demichelis, Setlur et al. 2012). Both of these regions have demonstrated an association with a more aggressive phenotype in prostate cancer. Similarly, in a study of non-BRCA associated early onset breast cancer, 68 germline samples of affected individuals were analyzed with whole genome CGH (Krepischi, Achatz et al. 2012). Rare CNVs were detected in 37% of affected cases however they were also detectable in 23% of controls. Furthermore, the total number of rare CNVs did not differ
between control and affected individuals although the proportion of rare CNVs was higher in breast cancer specimens. CNVs in breast cancer patients involved numerous genes previously ascribed to cancer and included a novel TSG, ST6GALNAC5, which is implicated in brain metastasis. Although, there was no significant difference in the number of rare CNVs between cases and controls, the location of rare CNVs in cases demonstrates that rare inherited CNVs do increase breast cancer susceptibility. Furthermore, the presence of rare germline CNVs has also recently been associated with prognosis and breast cancer recurrence (Sapkota, Ghosh et al. 2013). Similar findings have been detected in other cancer subtypes such as ovarian cancer (Fridley, Chalise et al. 2012) and CRC (Krepischi, Pearson et al. 2012).

2.4 CNVs and FPC

The seemingly increased susceptibility to hereditary cancer imposed by rare CNVs has led to similar investigations in FPC. Historically, PDAC has been notoriously difficult to study genetically largely due to the rapidly fatal nature of this disease, limiting both germline and somatic sample acquisition. A recent study investigated the role of rare CNVs in hereditary predisposition to FPC (Al-Sukhni, Joe et al. 2012). This PACGENE study led by our lab examined 120 germline FPC samples vs. 1,194 controls. Samples were processed on the Affymetrix 500K SNP DNA microarray and a subset of cases (n=36) and 2,357 controls were processed on the Affymetrix 6.0 microarray containing 1 million SNP probes. Similar to the aforementioned breast study by Krepischi et al, no significant difference was observed in the number of rare CNVs between cases and controls however a total of 93 unique CNVs (53 losses and 40 gains) were identified in 50 cases. These CNVs overlapped 88 genes, many of which, having a potential role in cancer.

The majority of CNV studies in FPC have focused on somatic tissue profiling of tumors. Molecular analysis of 60 PanIN lesions from individuals with a family history of PDAC, performed on SNP microarrays (Illumina 370/660K SNP), demonstrated a paucity of CN changes (Hong, Vincent et al. 2012). Furthermore, only 8 lesions had >1 somatic copy number alteration (SCNA) of which the majority were chromosomal loss and no single region of recurrent chromosomal gain or loss was detected. This indicated that SCNAs occur later in tumorigenesis and that no single TSG is responsible for cancer progression in early lesions. The largest study to date of CNVs in FPC cancer specimens has been performed using an older
method of detection, specifically microsatellite (MS) markers (Abe, Fukushima et al. 2007). Unlike microarray technologies, MS markers have low-resolution genome wide coverage, are inherently PCR dependent and are unable to detect copy neutral changes like with SNP arrays. In the study by Abe et al, DNA was extracted from neoplastic tissue with matched normal from 20 cases with PDAC and compared with 5 cases with sporadic PDAC. These cases had a family history of >1 FDR with PDAC and therefore did not satisfy the classical criteria for FPC. Genome wide probing was performed with 391-labeled MS markers. Average LOH was seen in 49.9% of informative markers ranging from 13.0% to 89.2%. Numerous chromosomal arm losses were detected in familial cases and these profiles were slightly different than sporadic profiles, although a formal test for significance was not performed. Furthermore, it appears that familial cases demonstrated a larger average loss frequency than sporadic cases. Regions of high frequency loss in familial cases corresponded to known genes in sporadic disease namely, CDKN2A/p16, p53 and SMAD4. The limitations of this study are the small case numbers and likely low power to detect a difference of small effect size, the use of a low resolution detection method and, finally, including cases that do not adhere to the accepted or modified definition of FPC. Although other studies of SCNAs in PDAC have been described (See Discussion section), further insights on genomic profiles in FPC are required.
3  Next Generation Sequencing

This section briefly describes the development and applications of NGS. Although NGS methodologies form only a small component of this thesis, no description of cancer genomics is complete without a discussion of the significance of NGS.

3.1  Sequencing and Human Genome Project

Advances in sequencing techniques and technology resulted in the first automated sequencing machine, developed by Applied Biosystems in the late 1980s, employing capillary electrophoresis (Liu, Li et al. 2012). The earliest automatic sequencing machines were capable of sequencing 96bp in a single run with a 600bp read length, to a maximum of 500kb per day.

Undoubtedly, one of the largest initiatives successfully undertaken and mediated by advances in DNA sequencing was the completion of the 13-year HGP in 2003, at an estimated cost of $2.7 billion. This project involved the generation of approximately 20,000 BAC clones serving as a physical map and facilitating sequencing of the human genome (McPherson, Marra et al. 2001). BACs were generated with comprehensive coverage of the human genome and minimum redundancy whereby each clone was identified by a characteristic restriction fragment pattern. These clones were then shotgun sequenced and reconstructed forming a physical map of the human genome. Results from the HGP were not based exclusively on a single diploid genome but rather the incorporation of haploid consensus sequences from several individuals of varying ancestry. This initiative demonstrated that the human genome constitutes >2 billion bps of which <2% codes for a transcript comprising >26,000 proteins (Venter, Adams et al. 2001). Shortly after the description of the HGP, the first diploid human genome was sequenced by Sanger technology of 32 million DNA fragments (Levy, Sutton et al. 2007). Interestingly, results from this endeavor demonstrated tremendous single genome diversity (4.1 million variants) involving 12.3Mb and highlighted the importance of non-SNP based variation to genomic diversity. The consequence of the HGP is that it has created a reference database for human sequencing forming a platform for future genetic studies in health and disease. One such follow-up initiative was the international Hap Map consortium, which described common variation at the SNP level for varying ancestral backgrounds. (See Section 2.2: CNV Detection Methods) (International HapMap Consortium 2005).
3.2 Second Generation Sequencing

Following the initial description of the human genome, major progress occurred in further characterizing human genomic diversity by technological advances in sequencing methodologies. The basic principles of NGS rely on massive parallel sequencing of clonally amplified or single cell DNA (Voelkerding, Dames et al. 2009). This is opposed to conventional Sanger technology whereby samples are sequenced in individual reactions. Furthermore, NGS is performed by repetitive cycles of polymerase mediated nucleotide extension, which allows the generation of gigabase (Gb) nucleotide output in a single run. The utility of such technology has largely surpassed conventional Sanger sequencing for large-scale genomic studies and as such, NGS has now been referred to as the “second generation” of DNA sequencing. In comparison, whereas the HGP was a multi-billion dollar endeavor, rapid advances in NGS have reduced the cost per Mb from >$5000 to <$0.10. The first genome sequenced using massively parallel technology occurred at a cost of $1 million and belonged to re-known scientist and Nobel laureate, James Watson (Wheeler, Srinivasan et al. 2008). Subsequently, the National Human Genome Research Institute outlined the goal of lowering the cost of sequencing a whole genome to $1000. Although lofty, this goal will likely be attained within the next 3-5 years although the greatest challenge will be in analyzing the enormous volume of data. Hence, it is anticipated that the $1000 genome will be associated with a >$20,000 analysis cost in addition to the challenges of applying this data towards the better understanding of human health or disease (McPherson 2009).

A complete description of the various NGS platforms is beyond the scope of this thesis although it is important to indicate that although the basic principles of library construction, sequencing and re-alignment to a reference assembly are similar between platforms, various companies employ different sequencing chemistries. This results in modifications of read length and time required for whole genome sequencing (Liu, Li et al. 2012, McPherson 2009). Large-scale WGS still remains routinely impractical due to cost and complexity of analysis. Alternatively, whole exome sequencing (WES) defined as sequencing of all the exons or protein coding regions of the genome has garnered tremendous interest. Exons account for approximately 1% of the entire human genome but constitute 85% of disease causing mutations, which have a significant impact on disease-related traits (Choi, Scholl et al. 2009). Therefore, analysis of exomes for single bp or small indel is possible on a large-scale basis. Indeed, as previously described, WES has led to the
discovery of PALB2, which is the second most common mutated germline gene in FPC. Furthermore, WES has been instrumental in the discovery of novel variants in endometrial cancer (Liang, Cheung et al. 2012), melanoma (Krauthammer, Kong et al. 2012) and familial CRC (Ku, Cooper et al. 2012) among many others.
Chapter 2  
Study Outline

2  Rationale, Aims and Hypothesis

2.1  Rationale

As outlined in the introductory section, PDAC is a rapidly fatal condition with most patients presenting with advanced metastatic disease (>80%). The median OS for patients with advanced disease who do not undergo any form of palliative systemic chemotherapy is 2-3 months (Stathis, Moore 2010). Therefore, there is urgency to better characterize and understand the molecular changes that drive carcinogenesis in the pancreas. As evidenced by other cancer subtypes and particularly the CRC model, a detailed analysis of familial syndromes may uncover novel genes which are important not only for relatives within a high risk family but also for a proportion of individuals with or at risk for sporadic disease. For example, linkage studies of patients with familial adenomatous polyposis (FAP), that represent <1% of colon cancer cases, identified a causative mutation in the adenomatous polyposis coli (APC) gene. Although this syndrome is relatively uncommon, mutations in the APC gene have also been found to occur in >90% of sporadic disease and this discovery was a direct result from findings in hereditary cases (Kinzler, Vogelstein 1996).

Sporadic PDAC, as previously described, demonstrates tremendous locus heterogeneity and therefore the majority of genetic changes at the basis of disease remain unknown. Similarly, the majority of underlying genetic changes in FPC remains unknown however kindreds demonstrate an elevated risk of developing disease likely due to a small number of causative mutations. It is anticipated, therefore, that findings in FPC will directly inform a proportion of sporadic PDAC.

2.2  Aims

The aims of this study are listed below:

1.) To characterize the somatic profile of sporadic PDAC and particularly FPC tumours.

2.) To compare somatic profiles obtained with novel microarray technology with profiles obtained through previously published lower resolution platforms.
3.) To use an integrative genomics approach using NGS of germline FPC tissue with microarray data of somatic FPC tissue to identify novel FPC TSGs.

2.3 Hypothesis

We hypothesize that consistent with previous reports in other cancer subtypes, a detailed genetic analysis of FPC cases will identify novel loci with potential broad applicability to the more prevalent sporadic PDAC. Specifically, we suspect that a causative gene(s) in FPC will have an underlying inherited germline mutation and that this same locus/loci will have a subsequent region of somatic loss (or referred to as LOH) consistent with the classical Knudson 2-hit hypothesis for TSGs (Knudson 1971) as demonstrated in Figure 6 below. Furthermore, we hypothesize that a small number of high risk TSGs with somatic LOH contribute to disease and these TSGs are more frequently and consistently loss in FPC than in the heterogeneous sporadic form.

Figure 6: Schematic Diagram of the Knudson 2-Hit Hypothesis Model (Foulkes 2008) (Reproduced with Permission, Copyright Massachusetts Medical Society)
Only 20% of patients with PDAC present with disease amenable to surgical resection and furthermore, <10% of patients with PDAC belong to a hereditary pancreatic cancer syndrome. The proportion, therefore, of FPC cases that undergo resection is small and this has largely hampered genetic studies in FPC (Abe, Fukushima et al. 2007). To that effect, analysis of formalin fixed paraffin embedded (FFPE) tissue is attractive especially for diseases like PDAC where the availability of tissue is scarce. We therefore hypothesize that microarray analysis of FFPE DNA is feasible (similar to fresh frozen tissue) with the added benefit of increasing power of detection, which will assist in identifying novel regions of genomic gains and losses.

Microarray analysis of FPC tumors has not been performed and therefore we believe that modern array technology will provide a more accurate CNV profile then previous microsatellite studies. Specifically, we hypothesize that copy neutral LOH regions are prevalent in FPC tumors and that these regions also harbor gene(s) important for tumorigenesis. These regions may be important for disease, however have been undetected due to the use of older CNV detection strategies. In addition, a subset analysis of cases within a family will demonstrate shared regions of loss on microarray analysis and these shared regions will harbor disease causing FPC gene(s) suitable for further validation and confirmatory germline testing.

Finally, we hypothesize that coupling germline NGS data with somatic microarray CNV profiles will allow effective filtering of numerous potential TSG candidates. Our mode of gene discovery will remain consistent with the common disease/rare variant hypothesis as outlined in chapter 1 (section 1.6.1). We suspect that through an integrative genomics approach combining the identification of rare germline variants with somatic LOH data, we will be able to prioritize gene(s) that act as disease causing TSGs.
Chapter 3
Study Design

3 Material and Methods

3.1 Ontario Pancreas Cancer Study and Tissue Specimens

The Ontario pancreas cancer study (OPCS) was established in April 2003 and includes a prospective database of patients diagnosed with PDAC (both sporadic and familial). The study also attempts to procure tissue samples of both primary and metastatic PDAC for the purpose of genetic analysis (Eppel, Cotterchio et al. 2007). The OPCS has the primary purpose of identifying both epidemiological and molecular causes of PDAC with the ultimate intent of developing effective screening and diagnostic protocols, biomarker development and improved therapeutics. The OPCS is the largest known database dedicated to PDAC in Canada and furthermore constitutes 1 of 7 sites under the PACGENE umbrella, as previous described (Chapter 1; section 1.7.2).

A query of the OPCS database identified 184 FPC cases with available tissue however only 133 FPC cases were physically present at SLRI. Of these 133 cases, 55 cases represented samples obtained from primary resections whereas tissue for the other 78 cases was obtained by biopsy (either CT guided or endoscopic). The remaining 51 cases with specimens were from patients outside Toronto, Ontario however, they were recruited to the OPCS after permission from the treating hospital. Demographic and epidemiological data was available for these latter cases however tissue specimens remained at the treating hospital. A request was sent by the OPCS for release of these tissue samples for the purpose of this study and 25 cases were successfully obtained. Therefore, the total number of FPC specimens available for potential analysis was 158. For comparison to FPC cases and as a positive control to ensure authenticity of microarray findings, 25 sporadic FFPE samples resected between January to December 2011 were obtained and included for microarray analysis.

3.2 Histo-pathological Confirmation

Hematoxylin and Eosin (H+E) stained slides were obtained for the 158 FPC and 25 sporadic PDAC samples. When absent, H+E slides were obtained from original FFPE tissue blocks either
available at the SLRI tissue bank or from the department of pathology at the Toronto General Hospital, Toronto, Ontario. FFPE tissue blocks were provided to the Pathology Research Processing (PRP) laboratory. H+E slides were cut to a thickness of 1-2 microns and stained according to in-house laboratory protocols.

The author along with Dr. Serra at the Department of Pathology, Toronto General Hospital, reviewed all the aforementioned cases. Main criteria of interest was the presence of cancer affecting the ductal epithelium along with a tumour cellularity >70% as determined visually by Dr. Serra (MacConaill, Campbell et al. 2009). Although the optimal tumor cellularity is undefined, most studies attempt to extract DNA from >70% tumor cellularity regions however less stringent parameters can be employed if a significant quantity of tissue is available (Dancey, Bedard et al. 2012). Furthermore, most microarray platforms and subsequent CNV calling algorithms rely on a higher tumour cellularity and generally show consistent results above 70% cellularity. Samples that failed to meet this minimum cellularity requirement were selected for subsequent laser capture microdissection (LCM) prior to DNA extraction. In total, 21 sporadic PDAC samples and 73 FPC samples satisfied the minimum cellularity requirements. A subsequent 7 FPC samples had <70% tumour cellularity and were candidates for LCM. The remaining cases not selected were excluded either due to a mis-classification of tumour histology (e.g. case initially believed to be PDAC but found to be an ampullary cancer), cases that had appropriate cellularity but no blocks or slides were available for DNA extraction, or finally, cases that failed to meet the minimum tumour cellularity with no tumour blocks available to cut tissue onto LCM slides.

Blocks for samples with >70% tumour cellularity was submitted to the PRP lab as above and blocks were cut on regular glass slides with a specification of 10-micron thickness. Tissue was left unstained and prepared following DNA/RNA extraction protocols. Similarly, sample blocks selected for LCM were cut onto the nuclease and human nucleic acid free LCM FrameSlides PET 1.4 microns (Quorum Technologies Inc., Germany) at a thickness of 10-microns once again following DNA/RNA extraction protocols. Tissue on LCM slides were stained with Hematoxylin and air dried for 24 hours.
3.3 Laser Capture Microdissection (LCM)

LCM was performed on the Molecular Machines and Industries AG (MMI) CellCut microscope. LCM was performed at the Advanced Optical Microscopy Facility (AOMF) at the Toronto General Hospital. LCM is a technique that enables the isolation of a specific subpopulation of cells (Espina, Wulfkuhle et al. 2006). In the case of PDAC, it allows the dissection of ductal epithelium specific cells while avoiding the inclusion of desmoplastic and surrounding normal tissue enabling tumour specific genetic analysis. Briefly, a blank regular glass slide is obtained and cleansed with 70% ethanol x 2. The slide and surrounding workbench is then rinsed with RNaseZap (Ambion, USA). A single LCM cap is inserted into a tube holder of the LCM microscope in an inverted fashion and the tube is placed on a magnetic holder on the microscope. Next, the PET 1.4 microns LCM FrameSlides containing tissue stained with Hematoxylin is superimposed on the cleansed regular glass slide with the LCM slide membrane facing down and inserted into the microscope slide holder. The microscope cut speed is set at <30%, laser focus at 10x, 49.0%; 20x, 51.3%; 40x, 47.6%; 60x. 50.0% and finally laser power is set at >80%. Laser dissection repeat function is set to 2-3 indicating the number of times the laser actually makes a cut in the areas highlighted. Mapping of the entire slide is performed at 4x objective with the microscope camera. While at 4x focus, laser alignment is tested to ensure that areas mapped indeed correspond to where the laser cuts. Once confirmed, regions of ductal epithelium are highlighted and the area of interest is zoomed in on with 20x and 40x objective. Region of interest is highlighted with a digital LCM marker and once appropriate, the cut function is executed. At this stage, the LCM cap contacts the tissue and once the UV laser has finished making the cut of interest, the LCM cap is lifted with the tissue of interest adherent to the cap surface. The above is repeated until the LCM cap is 25-50% covered with tissue and repeated for a total of 3-4 caps per sample.

3.4 DNA Extraction

DNA extraction of FFPE tissue was performed using the QiAmp DNA FFPE Tissue extraction kit (Qiagen, USA). The kit includes 4 commercially prepared solutions: Buffer ATL, AL, AW1 and AW2. Prior to tissue dissection, buffer ATL and AL are heated at 70°C with gentle shaking for 10 minutes. Buffer AW1 containing 19mL of solution is mixed with 25mL of 99% ethanol
and is vortexed vigorously. Similarly, buffer AW2 containing 13mL of stock solution is mixed with 30mL of 99% ethanol and vortexed vigorously.

Unstained 10 micron slides corresponding to 21 sporadic PDAC and 73 FPC samples were superimposed with their corresponding H+E slides and region of tumor with >70% cellularity were marked on 5 unstained slides per case. These corresponding regions were then micro-dissected using a sterile 15-blade scalpel and placed in a 1.5mL eppendorf tube. Enough slides were micro-dissected per case to ensure that a quarter of the 1.5mL eppendorf tube was filled with tissue. 320 microliters (μL) of deparaffinization solution (Qiagen, USA) was added to each eppendorf tube and vortexed vigorously for 30 seconds. The eppendorf tubes were then centrifuged at 14,000 rotations per minute (RPM) for 2 minutes. Samples were transferred to a thermomixer 5436 in batches of 24 eppendorf tubes and incubated at 56°C for 3 mins. Eppendorf tubes were then allowed to cool at room temperature for 5 minutes. Following this cool down period, 180μL of lysis buffer (ATL buffer) was added to each sample causing the formation of 2 separate layers: deparaffinization solution at the top and ATL buffer with tumour tissue at the bottom. Eppendorf tubes were then vortexed vigorously for 30 seconds and centrifuged at 10,000 RPMs for 1 minute. Following centrifugation, 20μL of proteinase K solution (600mAU/mL) solution was added gently to the lower clear phase containing tissue and ATL buffer and then this layer was mixed gently by pipetting up and down. The samples were then vortexed vigorously for 15 seconds and incubated at 56°C for 24 hours. Following 24-hour incubation, 10μL of proteinase K was re-added to the lower layer of the samples and re-incubated at 56°C for 1 hour. Samples were then transferred to a thermomixer set at 90°C and incubated for another 1 hour. Following incubation, samples were centrifuged for 2 minutes at 14,000 RPMs. The lower digested clear phase now containing lysed cell contents was transferred to a new 1.5mL eppendorf tube.

In the new eppendorf tube containing a clear solution with lysed cell contents, 200μL of buffer AL was added and mixed vigorously by vortexing. This was then followed by the addition of 200μL of 99% ethanol to the eppendorf tubes and vigorous mixing by vortexing. The eppendorf tubes were centrifuged at 10,000 RPMs for 1 minute. Cell lysates were transferred to the QIAamp MinElute column containing a 2mL collection tube and samples were centrifuged at 8,000 RPM for 1 minute. The elution column was then transferred to a new 2mL collection tube and the contents of the previous collection tube containing the flow-through was discarded. A
total of 500µL of mixed buffer AW1 with 99% ethanol was added to the elution column and once again centrifuged at 8,000 RPMs for 1 minute. The elution column was then transferred to a new 2mL collection tube and the previous collection tube with flow-through was discarded. This was repeated following the addition of 500µL of mixed buffer AW2 with 99% ethanol. A final centrifugation step was performed with the elution column and a new 2mL collection tube at 14,000 RPMs for 3 minutes in order to dry the column central membrane completely. The column was then transferred to a topless 1.5mL microcentrifuge tube and to the central column membrane, 50µL of buffer ATE (DNA elution buffer) was added. The column was allowed to incubate at room temperature and this was followed by centrifugation at 14,000 RPM for 1 minute. Column flow-through was collected and transferred to a labeled 500µL eppendorf tube. The above elution step was repeated one more time with buffer ATE and the solution now containing extracted DNA in the 500µL eppendorf tube was placed in a 4°C refrigerator.

The above was repeated for all matched normal samples.

3.5 Nanodrop Spectrophotometry

Initial quantification of extracted DNA from FFPE tissue was performed on the Nanodrop ND-1000 Spectrophotometer (Thermo Scientific, USA). The “nucleic acid” option is selected on the operational software. Once loaded, the pedestal of the nanodrop reader is cleaned with a kimwipe (Kimberley-Clark Professional) and 2µL of double distilled water (ddH₂O) is placed as a control. The “DNA” reading mode is selected and a blank measurement is performed. Once complete, the nanodrop pedestal is once again cleaned with a kimwipe. Next, 2µL of ATE buffer is added to the nanodrop reader and the calibration option is selected. Once the calibration function normalizes the spectrophotometer to 0ng/µL, the machine is now ready to perform sample measurements of FFPE DNA already suspended in ATE buffer. Once again, the nanodrop pedestal is cleaned with a kimwipe and 2µL of sample DNA is placed on the reader and quantified for each tumour/normal pair. Nanodrop readings include DNA concentration in ng/µL and optical density (OD) 260/280 and 260/230 ratios.

3.6 Whole Genome Amplification

Whole genome amplification (WGA) was performed on all extracted tumour/normal DNA pairs using the REPLI-g FFPE WGA kit (Qiagen, USA) following the high-yield reaction protocol.
The WGA kit comes with prepared FFPE buffer, ligation enzyme, FFPE enzyme, reaction buffer and DNA polymerase. The principle behind WGA is based on the genomiphi polymerase amplification (ϕ29) method coupled with random adaptor ligation allowing uniform genome wide amplification (Pinard, de Winter et al. 2006). DNA amounts of >100ng and ideally 300ng were used in this procedure based on the nanodrop readings obtained (as per section 3.5). Depending on DNA concentrations and amounts for each sample in the stock solution, 100-300ng of DNA was obtained and added to a 500µL eppendorf tube and brought to a final volume of 10µL by adding ATE buffer. The eppendorf tubes were placed in a thermomixer and incubated at 95°C for 5 minutes and then allowed to cool on ice for 5 minutes. In a separate 1.5mL eppendorf tube, an FFPE master mix is created on ice consisting of: 8µL of FFPE buffer; 1µL of ligation enzyme; 1µL of FFPE enzyme. The above volumes are per sample and therefore the total amount of stock solution was multiplied by the total number of samples. To the 10µL sample of extracted DNA, 10µL of FFPE master mix was added and vortexed for 15 seconds. The sample was then centrifuged at 14,000 RPM for 1 minute. The sample was then incubated at room temperature for 30 minutes. These steps constitute the adaptor ligation process and following the 30 minute incubation, the reaction is stopped by a 5 minute incubation at 95°C and a 5 minute cool down on ice. Next, a REPLI-g master mix is created consisting of: 29µL of REPLI-g Midi Reaction Buffer and 1µL of REPLI-g Midi Reaction Buffer. Once again these volumes are per sample and the total amount of mix was multiplied by the total number of samples. A total of 30µL of REPLI-g master mix was added to the denatured DNA solution now consisting of ligated adaptors. The solution was vortexed for 15 seconds and subsequently centrifuged at 14,000 RPMs for 1 minute. The sample was incubated at 30°C for 8 hours and the reaction was stopped following this time period by a 10-minute incubation at 95°C.

3.7 Qubit double stranded DNA Quantification

Double stranded DNA quantification was performed using the Qubit dsDNA broad range (BR) assay kits using the Qubit 2.0 fluorometer (Invitrogen, USA). All assays were performed at room temperature (22-28°C) while attempting to minimize temperature fluctuations. The kit contains 4 separate reagents: 1.25mL of Qubit dsDNA BR reagent (component A); 250mL of Qubit dsDNA BR buffer (component B); 5mL of Qubit dsDNA BR standard #1 (component C); 5mL of Qubit dsDNA BR standard #2 (component D). Firstly, Qubit specific 0.5mL spectrophotometer assay tubes were obtained for both samples and standards and labeled accordingly. Secondly, the Qubit
working solution was prepared by diluting the Qubit dsDNA BR reagent 1:200 in the Qubit dsDNA BR buffer. The samples were quantified in two separate batches performed on 2 separate days. DNA quantification of tumour DNA was performed on day 1 whereas DNA quantification of matched normal DNA was performed on day 2. For 80 samples, 100µL of Qubit reagent was diluted in 19.90mL of Qubit buffer and vortexed vigorously in a 50mL tube. Extreme care was taken to not expose the Qubit reagent and the working solution tube to visible light and both were covered with aluminium foil. Next, calibration of the Qubit fluorometer was performed using both provided standards. 190µL of Qubit working solution was added to an empty 0.5mL Qubit assay tube and 10µL of standard #1 was added and vortexed for 15 seconds. The solution was allowed to stand for 2 minutes. On the fluorometer working screen, the DNA mode is selected with the dsDNA BR option. The standards screen is then selected and the “run new standards” option is chosen. The tube containing the working solution and standard #1 are loaded into the fluorometer chamber and the “read” option is selected. Following 3-5 seconds, calibration of standard #1 is complete. The above procedure is repeated for standard #2 ensuring that 190µL of working solution is added to 10µL of standard #2.

Lastly, dsDNA quantification of sample DNA was performed on tumour DNA. In a labeled 0.5mL Qubit assay tube, 198µL of Qubit working solution was added and to this solution, 2µL of sample DNA was added for a total volume of 200µL. The solution was allowed to stand for 2 minutes. As previously described, the Qubit fluorometer was set to read dsDNA and the “use current standards” option was selected. The sample DNA was loaded into the Qubit reader chamber and the “read next sample” option was chosen. A QF value is given by the fluorometer and this is converted into a concentration in µg/mL by selecting the quantify DNA option on the reader and selecting 200µL as the volume of sample and working solution. DNA concentration is obtained using the formulation: sample concentration = QF value x (200/X) where X= the number of µL of sample in the assay tube. Results were then converted manually into ng/µl. The above was repeated for all tumour samples and for normal samples on day 2 using the “use current standards” option.

3.8 Affymetrix Oncoscan FFPE Express Version 2.0

Tumour and matched normal DNA samples were processed on the Oncoscan FFPE Express version 2.0 SNP microarray. Samples were shipped to Affymetrix in Santa Clara, California for
sample manipulation, processing and normalization. Following, dsDNA quantification, a minimum of 75ng of DNA were aliquoted onto a 96 well plate provided by Affymetrix. Although 75ng of DNA was the average amount required for successful DNA processing, Affymetrix recommended a total of 250ng of dsDNA and therefore all attempts were made to provide this amount. Stock DNA solution was provided where possible, however if a total of 250ng of DNA could not be achieved with stock DNA alone, then WGA DNA was substituted instead. In total, 70 tumour samples from source stock DNA were included whereas 10 WGA samples were required. In addition, 71 normal samples from source stock DNA were included whereas 9 WGA normal samples were required.

As the technical aspects of this section were performed at Affymetrix (at the time of this study, Oncoscan was not released to individual or core laboratories), we will provide a brief description of molecular inversion probe technology (MIP) which is the basis behind why Oncoscan is ideal for and geared specifically towards FFPE tissue.

MIP technology is based on a previous microarray probe technique known as padlock probes (Wang, Cottman et al. 2012). Briefly, padlock probes (Figure 10 a.) consist of a nucleic acid probe that is capable of hybridizing to a genomic target of interest. The hybridization occurs in a fashion, which allows the probe to bind to the target at two locations (one at the 5’ and 3’ end of DNA) over a 40bp span and then the two ends are joined or circularized. These circularized probes are resistant to exonuclease activity as opposed to linear probes and therefore following digestion, only hybridized circularized probes of interest remain. These probes are then amplified with the use of universal primers allowing for target region enrichment. With further development, cleavage of these probes with subsequent amplification and hybridization to an array was possible.

MIP probes follow a similar workflow. Firstly, linearized probes are added to the genomic DNA of interest in 4 separate reactions corresponding to the 4 nucleotide bases. Complementarity leads to circularization of the probe of interest. However, probe design is such that binding to target of interest and circularization leaves a gap at the hybridized region, which is then filled using a DNA polymerase with free nucleotides. DNA ligase then completes the circular joining of the probe. Gap filling and circularization is not performed for non-hybridized probes and these probes are then degraded by the addition of exonucleases and DNAases. Subsequently, probe
release is performed whereby the probe of interest, now the only remaining probe in solution, is linearized with the use of restriction enzymes. In a linearized form, universal PCR primer sites are present at the 5’ and 3’ end of the probe, which allows traditional PCR amplification to be performed. The amplified probe is now ready for hybridization onto a microarray and this is performed via a probe specific tag or barcode present on the MIP, which uniquely identifies this probe and the target region on the array. These unique tags are cleaved by restriction enzymes, released and hybridized in this case to probes that are complementary on the Oncoscan array. A pictorial description is available in Figure 10 b.). By using, MIP technology, PCR amplification and direct hybridization of source DNA is avoided which is advantageous for fragmented FFPE DNA. The end result is an increase in DNA signal in comparison to noise, an issue that is significant with FFPE DNA on other non-MIP platforms. Whereas, MIP arrays were designed initially for SNP genotyping, their function has now expanded to include CN and LOH analysis. 

The Oncoscan FFPE express version 2.0 is a 660,000 (660K) SNP microarray with focus on known cancer associated regions/genes while probing at a genome wide level at a lower resolution. Median probe spacing is every 4.2kb. Building off the original 330,000 SNP array, Oncoscan was devised in collaboration with the American Association for Cancer Research Stand Up to Cancer Initiative and includes an enhanced coverage of known oncogenes and TSGs (Wang, Cottman et al. 2012). The array provides coverage of over 200 TSGs and oncogenes with a median probe density every 0.5kb for the top 10 most “common” TSGs and 1 probe every 2kb for the 190 most “common” oncogenes.

This paragraph describes briefly the sample workflow for the Oncoscan MIP array, provided by Affymetrix Inc. Following sample shipment with a minimum of 75ng (ideally 250ng) of DNA, 2 quality control (Qc) steps are immediately performed on the samples of interest. First, DNA is quantified using PICO-green analysis to confirm dsDNA concentration values. A minimum cut off of 75ng of DNA is set however in cases where tissue is scarce, as little as 20ng of DNA can be attempted. As previously indicated, samples with 250ng were provided either as stock DNA or WGA DNA. The second Qc step involves agarose gel electrophoresis following a PCR amplification step of the DNA. Samples are compared to a known banding pattern, which correlates with a likelihood of a successful assay on the microarray. At this point, samples with a lower likelihood of hybridization success are noted and can be removed from subsequent processing. The samples are then incubated with the MIP probes and hybridized onto the
Oncoscan array. Following sample hybridization, signal normalization is performed. Briefly, a clustering algorithm is performed on an individual marker basis using tumour/normal pairs. An average cluster signal of the diploid state is generated per marker site. Next, a 5 step iterative process per tumour sample in comparison to the average cluster signal and the matched normal signal is performed and a Langmuir regression analysis is employed providing a copy number state for each sample at a particular marker. These raw copy number signals can then be converted into an interpretable copy number state using a bioinformatics software package (see section 3.9).

Figure 7: MIP technology and array hybridization (Reproduced with Permission) ([http://en.wikipedia.org/wiki/Molecular_Inversion_Probe](http://en.wikipedia.org/wiki/Molecular_Inversion_Probe))

a.) Linear MIP

![Linear MIP probe](image)

Figure 7 a.) represents a schematic diagram of an un-circularized MIP probe. The 5’ and 3’ ends both have a sequence complementary to the target region, which permits hybridization and circularization. Once exonuclease digestion of unbound probes and gap filling phase is complete, the universal primer sites are used to amplify the probe followed by tag cleavage allowing for tag hybridization to a complementary region on the Oncoscan microarray.
b.) MIP probe array workflow design

3.9 Bioinformatics

Normalized raw intensity microarray data was analyzed using Nexus Copy Number software for CNV detection and analysis version 6.1 (Darvishi 2010) provided by Biodiscovery Inc. USA. Two algorithms were used to identify regions of copy loss and gain including copy neutral LOH. The first algorithm was a HMM based algorithm, SNP-FASST2 Segmentation, which is not a cancer specific algorithm. FASST2 algorithms are different from conventional HMM models in that a CN state is not made at a specific marker however uses many states to cover possibilities such as mosaic events and then generates a CN value based on these possibilities. Furthermore, SNP-FASST2 algorithms employ a multiple of states, which allow allelic-based events to be determined. The second algorithm used was a segmentation-based algorithm designed
specifically for CNV analysis in cancer tissue. The Allele Specific Copy Number Analysis of Tumors (ASCAT) algorithm accounts for challenges encountered in cancer CNV discovery, namely normal cell contamination or admixture and tumour cell ploidy changes (Van Loo, Nordgard et al. 2010). In both these algorithms, a copy loss or gain region was defined as a LRR deflection of -0.3 and +0.3, respectively, with a high loss and high gain defined as a LRR deflection of -1.2 and +1.2 respectively. A region of LOH was observed when a minimum of 3 contiguous probes were loss, stretching a minimum of 500kb up to a maximum of 1000 contiguous SNP probes. The significance threshold was set a 5.0 x 10^{-7} and these parameters were the default settings in the Nexus software. Although the “true” definition of a CN loss and gain region is arbitrary among studies, these parameters are within the limits currently accepted for performing CNV studies.

In this study, paired analysis of tumour/normal tissue was performed on each algorithm allowing for a tumour specific CNV profile (the paired analysis enables subtraction of the germline CNV profile leaving a tumour specific CNV output). This analysis was performed for both sporadic PDAC and FPC samples. A Fisher exact test comparing sporadic and FPC profiles was performed on the Nexus software attempting to find regions of difference between these two cohorts which may signal a region corresponding to an FPC gene of interest. Although both gain and loss regions were analyzed, the focus for this thesis was predominantly for regions demonstrating CN Loss and/or copy neutral LOH regions.

Finally, stratification of higher priority CNV regions for FPC cases on both SNP FASST2 Segmentation and ASCAT was performed using the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm version 2.0 (Beroukhim, Getz et al. 2007), available directly on the Nexus software. This algorithm attempts to identify those regions, which are gained, or loss more likely then would be expected by chance using the false discovery rate method. Regions displaying gain with consistent frequency based on the significance threshold chosen and those regions showing high amplitude gain are highlighted. Similarly, region displaying loss with consistent frequency and regions with homozygous deletion are prioritized. As previously performed in our laboratory, regions of high confidence gains or loss were identified by regions shared on GISTIC between both CNV calling algorithms (SNP-FASST2 Segmentation and ASCAT) and/or regions with >20% overlap (Al-Sukhni, Joe et al. 2012).
Regions present in a single algorithm, regions shared between algorithms as above, shared LOH regions between members of the same family, novel regions not described in the previous FPC microsatellite LOH study (Abe, Fukushima et al. 2007), regions of loss occurring more frequently in FPC than sporadic PDAC identified with a Fisher test and regions in cases with underlying germline NGS data (see below) were used to filter candidate genes/regions of interest.

3.10 Germline Next Generation Sequencing and Gene Filtering

Although ES data was used in this study to filter and prioritize candidate TSGs as potential FPC causative genes, the experiments and technical aspects of NGS were not a major focus of this thesis. Analysis of the germline ES data referenced in this thesis was performed by the OICR and through collaborators at Genome Quebec and the Zogopoulos lab at the Rosalind and Morris Goodman Cancer Research Center, McGill University, Montreal, Quebec.

The following is our laboratory based ES and variant calling protocol as published by Grant, Al-Sukhni et al. 2013 (Reproduced with Permission). Mr. Robert Grant, MA (Gallinger Lab) and Dr. John D. McPherson, PhD (OICR), performed all variant calling and NGS specifics in Toronto. Germline DNA for ES in Toronto was obtained through the BSR at the SLRI using in-house organic solvent isolation or column-based purification methods. Genomic germline DNA was constituted to a concentration of 10ng/µL. Germline DNA samples for ES were captured over time through array hybridization with exon-targeted oligo probes using the Nimblegen HD2 sequence capture array (http://www.nimblegen.com/products/seqcap/arrays/index.html, Roche Nimblegen, Madison, WI), Agilent SureSelect All Exon (www.agilent.com), or Illumina TruSeq Exome Enrichment Kit, following manufacturers suggested protocols. Genomic DNA was fragmented by sonication to obtain a fragment size distribution ranging from 300 to 500 bps, verified via the Bioanalyzer 2100 DNA chip (Agilent Inc. USA). End repair and adaptor ligation were performed and followed by library construction as per manufacturer protocols. The whole genome library was quantified prior to sequence capture hybridization procedure. Post-enrichment DNA libraries of the resulting captured DNAs were sequenced on the Illumina Genome Analyzer IIx or HiSeq platform with paired-end 76 to 101 base reads following the manufacturer's protocols and using standard sequencing primers. Image analyses and base callings were performed by the Genome Analyzer Pipeline with default parameters and default
filtering. Two lanes per sample were used, or until we achieved coverage of at least 8 sequencing reads in 90% of the target.

Illumina’s latest and standard pipeline with the recommended parameters produced the raw FASTQ reads. Basic Qc metrics, including average read quality values, average base quality values, distributions of A, C, G, and T, and distributions of GC contents of all reads were performed on each lane of FASTQ reads to ensure they passed our internal Qc before we proceeded to the next phase of our analysis. Lanes that passed our Qc were then aligned to the UCSC HG19 human reference genome, including random and unknown sequences, using Novoalign (www.novocraft.com) and only reads that aligned uniquely to the reference genome were kept for further analysis. Base qualities of the aligned reads were re-calibrated based on mismatches and known single nucleotide variants (SNVs) from dbSNP132 using the Genome Analysis Toolkit (GATK). We used GATK for local realignments around the indels, and to remove cryptic or inconsistent SNVs and indels. GATK then called SNVs and indels using the parameters recommended on the GATK wiki website (www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit). From VCF files outputted by GATK, we retained SNVs with over 7 total reads, and indels with over 11 total reads where at least 30% of which supported the alternate allele. These filters provided over 90% true positives in validation studies conducted on other samples at the OICR. The remaining variants were then annotated using ANNOVAR for: the predicted effect of the variant on protein sequence; and whether the variant was present in dbSNP135, the March 2012 data release from the 1000 genomes project, or the ESP5400 data release from NHLBI GO Exome Sequencing Project; and whether the mutation was in a gene previously associated with FPC (ATM, BRCA1, BRCA2, CDKN2A, MLH1, MSH2, MSH6, PALB2, PMS2, PRSS1, and PRSS2, STK11, TP53). Mutations in FPC genes considered as potentially causative, absent from the control databases and predicted to inactivate their respective protein, or were previously associated with FPC were of interest.

As an initial method of gene filtering, cases with both somatic Oncoscan and germline ES data were analyzed for regions containing a rare germline variant in a gene of interest with the presence of a subsequent LOH region in the respective tumour. These genes were technically considered to satisfy the classical 2-hit hypothesis and were annotated as high priority genes. Genes satisfying the 2-hit model and shared between cases were of even greater importance and
highlighted as front-runners for validation studies. Another method of gene filtering involved examining all samples with germline ES data and determining genes with a rare deleterious variant present in 2+ individuals. Furthermore, genes with a rare variant present in 2+ individuals were cross-referenced with GISTIC somatic data obtained by Oncoscan. Any genes with both a highlighted GISTIC region of loss and rare variants shared between 2+ cases were annotated as high priority FPC genes for subsequent validation assays.

3.11 Sanger Sequencing of Candidate Variants

Genes highlighted as candidate tumour suppressor FPC genes by cross-referencing germline NGS data with somatic Oncoscan data underwent validation of both germline and somatic changes. Germline rare variants of candidate genes were validated using conventional Sanger sequencing. Primers for respective genes were determined by USCS genome browser (http://genome.ucsc.edu) and designed on Primer3 freeware software (http://frodo.wi.mit.edu) and confirmed with in-silico PCR on the UCSC genome browser. Genes, primer sequences and annealing temperatures are listed below:
Table 3: Candidate FPC genes PCR primer specifications (Toronto Cohort):

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant Location</th>
<th>Forward Primer Sequence</th>
<th>Reverse Primer Sequence</th>
<th>Annealing Temperature(°C)</th>
<th>Concentration of MgCl₂ (mM)</th>
<th>Amplicon Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VANGL1</td>
<td>chr1:116228631</td>
<td>AGTTCACCTCTCTGGGGGGAAGT</td>
<td>CAAGCTTTTGGGAGGCTGAAG</td>
<td>60</td>
<td>3.0</td>
<td>482</td>
</tr>
<tr>
<td>POGK</td>
<td>chr1:156819402</td>
<td>ACCTCAAGCTCTCAGGTTGCT</td>
<td>ACAGTGTACAGCAAAGGCACTG</td>
<td>60</td>
<td>2.5</td>
<td>390</td>
</tr>
<tr>
<td>C3orf72</td>
<td>chr3:138868461</td>
<td>CTGTGAGAGAGAGAGGATGCT</td>
<td>AGGGACAGTATCCCTGCTTC</td>
<td>60</td>
<td>3.0</td>
<td>374</td>
</tr>
<tr>
<td>GSDMA</td>
<td>chr17:38126784</td>
<td>CAAACTCTAGCCCTATCCAGG</td>
<td>CAAAGGAGGGCTGGAATCTAA</td>
<td>60</td>
<td>3.0</td>
<td>393</td>
</tr>
<tr>
<td>ATAF1A5L</td>
<td>chr1:47138993</td>
<td>CTTGCTGTCAACAGACTGA</td>
<td>CAGATTCAAGGAGGAATGGA</td>
<td>60</td>
<td>3.0</td>
<td>312</td>
</tr>
<tr>
<td>MAP3K6</td>
<td>chr1:27685021</td>
<td>GTCTAAGTTCGAGGAGGAGA</td>
<td>GTTCACTATGGGTGAGAGTG</td>
<td>59</td>
<td>2.5</td>
<td>441</td>
</tr>
<tr>
<td>TNC</td>
<td>chr9:117780260</td>
<td>CTTTCCCTCTGACAGAGGGCC</td>
<td>ACAGCGCTGCCTACTGCAACC</td>
<td>60</td>
<td>2.5</td>
<td>443</td>
</tr>
<tr>
<td>TNC</td>
<td>chr9:117777545</td>
<td>TGCACTGGGTTAGAAAAACA</td>
<td>CCAAAGCCAAGCCAGTTAGAA</td>
<td>60</td>
<td>2.5</td>
<td>466</td>
</tr>
</tbody>
</table>
**PCR Working Mix:**

All steps were performed on ice. PCR solutions were constituted to a final working solution of 25µL (22.5µL of PCR master mix + 2.5µL of sample DNA). Sample DNA was constituted to a concentration of 10ng/µL. Assuming a MgCl$_2$ concentration of 2.5mM in the working mix, enough volume for 5 reactions per sample were prepared which include 3x sample reactions, 1x positive control (in-house genomic DNA) and 1x negative (ddH$_2$O) water control. The volume per reaction is listed and all reagents were added following a 15 second vortex: 15.6µL of ddH$_2$O is added to a 500µL eppendorf followed by 2.5µL of 10x PCR buffer. This is then followed by the addition of 1.25µL of 50mM MgCl$_2$, 1µL of 5µM forward and reverse primer each, 1µL of 10mM dNTPs and lastly 0.15µL of 0.75units Platinum Taq polymerase (Invitrogen, USA). The solution is vortexed for 15 seconds then centrifuged at 10,000 RPMs for 30 seconds. For solutions requiring a final MgCl$_2$ concentration of 3.0mM, adjustments were made according to the following formula: $1.25\mu L/2.5mM = x/3.0mM$, where $x$ is the volume of MgCl$_2$ required to achieve a concentration of 3.0mM. The amount of MgCl$_2$ added to the working solution is directly subtracted from the amount of ddH$_2$O to ensure a final working solution volume of 22.5µL.

Following vortexing of the PCR working solution, 22.5µL of the solution is added to 2.5µL of sample DNA and pipetted up and down gently 3x to ensure mixing. The eppendorf tube is then vortexed for 5 seconds and centrifuged at 10,000 RPMs for 30 seconds.

**PCR Thermocycling:**

Samples are loaded in a 96 well Veriti Thermal Cycler (Applied Biosystems, USA) after an initial warming of the cycler to 95°C. Samples are cycled as follows: initial denaturation at 95°C x 4 minutes followed by 40 cycles of 95°C x 30 seconds, annealing at the primers respective annealing temperature as indicated for 30 seconds, extension at 72°C x 30 seconds and final extension at 72°C x 7 minutes. Following PCR amplification, samples were kept at 4°C till gel electrophoresis and amplicon cleanup were performed (See below).

**Gel Electrophoresis:**
A 2% gel was created for electrophoresis while DNA samples were amplified in the thermal cycler. 1.8g of Agarose (Invitrogen, USA) was added to 90mL of 1x TAE buffer and mixed thoroughly. The beaker containing mixed TAE buffer and agarose was then placed in the microwave and warmed at high setting for 1 minute. The beaker was removed and swirled vigorously for 1 minute and then re-placed in the microwave and warmed at high setting for another 30 seconds. Following the warming step, the beaker was examined to ensure that the agarose had completely dissolved in the TAE buffer. If any undissolved agarose was still present in solution, the beaker was re-warmed for 30 seconds in the microwave at high setting. Once all agarose was dissolved, 5µL of RedSafe Nuclei Acid Staining Solution 20,000X (Intron Biotechnology, USA) was added to the solution and the beaker was swirled until the dye was properly mixed. The solution was then poured onto a gel electrophoresis cartridge ensuring no bubbles were present during the casting process and the gel was allowed to solidify for 40-60 minutes.

Once PCR amplification was complete, 4µL of each sample was aliquoted onto a Parafilm strip and 1µL of in-house loading dye was added to each sample and mixed gently by pipetting the solution up and down. The samples are then loaded onto the gel and the gel is placed on the electrophoresis cartridge filled with 900mL of TAE buffer ensuring that there is enough solution to cover the surface of the gel. The cartridge is set at 110 volts and allowed to run for 40 minutes. The gel was then visualized under ultraviolet with the UVP transilluminator and images were captured using the BioDoc-It image capture system.

**PCR Product Clean-up and Sequencing:**

Following gel electrophoresis confirmation of amplified PCR product, amplicons were cleaned for residual primers and dNTPs using Exonuclease I and Antarctic Phosphatase (New England Biolabs Inc.). An aliquot of 8µL of PCR product were transferred to a new PCR tube with 0.90µL of ddH₂O, 0.74µL of Exonuclease I and 0.36µL of Alkaline Phosphatase. The solution is gently pipetted up and down and then vortexed for 15 seconds. Samples are then centrifuged at 8,000 RPM for 10 seconds. Samples are then placed in a PCR thermal cycler at 37°C for 30 minutes and then for 95°C for 15 minutes. Samples are allowed to cool at room temperature. A 4.5µL aliquot of cleaned-up product along with a 1µL aliquot of gene specific forward and reverse primer are transferred into separate PCR tubes and centrifuged at 8,000 RPM for 15
seconds and were then delivered to the Analytical Genetics Technology Center of Mount Sinai Hospital (Toronto, Canada) for Sanger Sequencing. PCR amplicons were sequenced in both the forward and reverse direction on the ABI 3730 analyzer (Applied Biosystems) and sequences were analyzed using Seqman Pro on the Lasergene Core Suite software (DNASTAR Inc. USA).

### 3.12 Fluorescence In-Situ Hybridization

FISH was performed at the Applied Molecular Pathology Laboratory at Princess Margaret Hospital, Toronto Ontario by Ms. Olga Ludkovski, MSc. 14 FFPE tumor sections and corresponding matched normal tissues were tested by FISH to confirm the LOH status of the following germline validated genes: ATPAF1-AS1, C3ORF72, DCLK3, GSDMA, MAP3K6 and SERPINF1.

Two FISH probes were applied per slide. Commercially available CEP probes (Abbott Molecular, Des Plaines, IL) were obtained. CEP1 and CEP3 were labeled with spectrum orange (SpO) and CEP17 was labeled with spectrum green (SpG) and used as an internal control. CEP probes were combined with in-house generated gene specific probes, which were derived from BAC clones identified from the Human Genome Browser Assembly (Feb.2009 GRch37/hg19). The following BAC clones (MAP3K6 – RP11-401H24; DCLK3 - RP11-776G6; SERPINF1 – RP11-4F24; C3ORF72 - RP11-769B22; GSDMA – RP11-387H17 and ATPAF1-AS1 – RP11-151C6) were obtained from the Centre for Applied Genomics (TCAG, Toronto). The BAC clones were labeled using SpO-dUTP or SpG-dUTP by means of a nick translation kit (Abbott Molecular, Des Plaines, IL) according to manufacturer’s protocol and verified on metaphase spreads from normal peripheral lymphocytes for correct chromosome location. A summary of BAC clones and locations are available in Table S1 (See Appendix Section).

Briefly, FFPE sections were deparaffinized in Xylene, dehydrated in Ethanol, pretreated in Citrate buffer (pH-6.8) for 50 min at 80oC followed by pepsin digestion for 19 minutes at 37oC. Slides were hybridized for 2 days at 37oC, washed and counterstained with DAPI nuclear staining dye. Tissues and cells were examined and scored at 63x magnification on an Imager M1 Zeiss microscope (Carl Zeiss Canada Limited) equipped with the appropriate filters. The JAI CV-M4+CL progressive scan monochrome camera (JAI Inc., San Jose, USA) and the MetaSystems Isis FISH Imaging software programs v5.3 (MetaSystems, Germany) were used for capturing images.
Chapter 4
Data Collection

4   Results

4.1   Sample Cohort and Microarray Profile Interpretation

The total number of samples included in this study after microarray processing are shown in Figure 8. From an initial sample size of 80 FPC cases, 2 failed the pre-array quality control metric (1 tumour sample and 1 unmatched normal sample) performed by Affymetrix. During the sample processing and subsequent data normalization stage, samples with a mean absolute percent deviation (MAPD) of >0.6 were excluded from further processing. A total of 24 samples were removed (14 tumour samples and 10 normal samples) including all WGA DNA cases. A total of 65 FPC tumour cases and 69 FPC normal cases were available for subsequent analysis. However, as a paired analysis could only be performed on matched tumour/normal pairs, the final FPC sample size available for CNV analysis was 55 cases. Conversely, all 21 sporadic PDAC tumour/normal pairs submitted to Affymetrix were successfully processed on the Oncoscan microarray and were available for a paired analysis.

Figure 8: Total FPC cases available for a paired matched analysis
A typical microarray tracing from an individual sample is shown in Figure 9.

Figure 9: Individual genome-wide Oncoscan microarray tracing for an FPC sample

As with any microarray output, a LRR tracing centered on 0 (top left panel) suggests the presence of a diploid state as would be expected in normal tissue. Further analysis of the BAF tracing (bottom left panel) demonstrates a tracing centered on 0.5 which suggests the presence of both an A and B allele as would be expected in a heterozygous state. A copy loss region is observed as a downward deflection of the LRR towards -1.0 (top center panel) with a splitting of the BAF towards 0 and 1.0 (bottom center panel) indicating the presence of a hemizygous state (only an A or B allele is present). Finally, a copy neutral region (CN-LOH) is encountered when the LRR is centered on the 0 baseline (top right panel) however analysis of the BAF demonstrates splitting as would be expected in a hemizygous loss. In this situation, although the LRR would suggest a diploid state, the BAF demonstrates a pattern consistent with a hemizygous loss indicating only one allele is present (either a retained A or B allele which has
reduplicated attenuating a negative deflection of the LRR). In a pure tumour sample, a hemizygous loss would be seen as a complete negative deflection of the LRR to -1.0 and a complete splitting of the BAF to 0 or 1.0. In Figure 9, although a downward deflection of the LRR and/or splitting of the the BAF is noticed for a hemizygous and/or CN-LOH, the magnitude is less than predicted. This is likely due to the presence of non-cancer cell admixture (i.e. normal cell contamination) which dilutes the true signal intensity from the cancer tissue.

4.2 FPC ASCAT Copy Number Profile

4.2.1 PLOIDY AND NORMAL CELL CONTAMINATION PROFILES:

A paired matched analysis of 55 FPC tumour/normal cases was initially performed on the ASCAT segmentation based algorithm. CNV outputs on ASCAT are based, as previously described, on the extent of ploidy changes and the level of normal cell contamination. From an initial sample set of 55 matched cases, only 51 cases were able to provide an ASCAT profile due to extensive non-cancer cell admixture. Similarly, a paired analysis of 21 sporadic PDAC tumour/normal cases was performed on ASCAT for which only 19 cases generated an ASCAT CNV profile. Table 4 illustrates predicted ploidy changes and level of normal cell contamination for both FPC and sporadic cases. Interestingly, although one of the main inclusion criteria for microarray analysis was tumour cellularity >70% as determined by a pancreatic pathologist, our findings demonstrate that the majority of cases did not meet this criteria when predicted on ASCAT. The average predicted percent normal contamination in the FPC samples was 61.59% with 35 cases demonstrating >50% contamination with non-neoplastic cells. As would be expected, all cases that underwent LCM (n=7) demonstrated normal cell contamination levels <30%. Similarly, the average percent normal cell contamination in the sporadic cases was 66.68% with only 2 cases demonstrating contamination levels <30% as initially predicted visually by the author and pathologist.

Our data also demonstrates that FPC and sporadic PDAC samples display varying aneuploidy between cases. In the FPC cohort, the average sample ploidy was 3.55n and similarly for the sporadic cases, average ploidy was 3.54n. In the ASCAT algorithm analysis (Van Loo, Nordgard et al. 2010), a diploid state for cases ranges between 2-2.7n. In our cohort, 23.5% (12/51) of the
FPC cases and 10.5% (2/19) of the sporadic cases demonstrated a ploidy level consistent with the ASCAT diploid definition.

Table 4: Individual sample ASCAT Ploidy and Normal Cell Contamination Predictions

<table>
<thead>
<tr>
<th>Pass/Fail</th>
<th>Sample</th>
<th>Status</th>
<th>Data Type</th>
<th>Quality</th>
<th>Ploidy</th>
<th>Contamination (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIL</td>
<td>0007065_C02</td>
<td>Unprocessed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.06E-01</td>
<td>4.35</td>
<td>2.00</td>
</tr>
<tr>
<td>FAIL</td>
<td>0007065_C04</td>
<td>Unprocessed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.10E-01</td>
<td>2.45</td>
<td>18.00</td>
</tr>
<tr>
<td>FAIL</td>
<td>0007065_F04</td>
<td>Unprocessed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.41E-01</td>
<td>2.55</td>
<td>11.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F06</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.99E-02</td>
<td>2.20</td>
<td>13.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>8.04E-02</td>
<td>2.25</td>
<td>13.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_A03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>8.78E-02</td>
<td>2.25</td>
<td>14.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_C10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.03E-01</td>
<td>2.30</td>
<td>16.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_E01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.38E-01</td>
<td>2.20</td>
<td>23.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_C01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.81E-01</td>
<td>2.30</td>
<td>32.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A04</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.01E-01</td>
<td>2.50</td>
<td>35.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_E03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>9.65E-02</td>
<td>2.20</td>
<td>39.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_G03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>7.56E-02</td>
<td>2.35</td>
<td>39.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.23E-01</td>
<td>4.80</td>
<td>43.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_B10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.30E-01</td>
<td>4.60</td>
<td>45.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_D03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.96E-02</td>
<td>2.30</td>
<td>46.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_A10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.44E-02</td>
<td>4.40</td>
<td>48.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_C07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.81E-01</td>
<td>2.95</td>
<td>59.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>9.95E-02</td>
<td>4.70</td>
<td>60.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_D07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.22E-02</td>
<td>4.65</td>
<td>65.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_G09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.70E-02</td>
<td>4.55</td>
<td>65.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.18E-01</td>
<td>3.05</td>
<td>66.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_G07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.06E-01</td>
<td>3.40</td>
<td>66.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_G01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.25E-01</td>
<td>4.30</td>
<td>68.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_C05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.21E-01</td>
<td>2.90</td>
<td>69.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A06</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.02E-01</td>
<td>4.65</td>
<td>72.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F08</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.63E-02</td>
<td>4.50</td>
<td>72.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.35E-01</td>
<td>2.95</td>
<td>73.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_H11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.06E-02</td>
<td>4.65</td>
<td>74.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>4.89E-02</td>
<td>4.55</td>
<td>74.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_B09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>7.05E-02</td>
<td>4.45</td>
<td>75.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_G05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>8.68E-02</td>
<td>3.80</td>
<td>76.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_C05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>8.55E-02</td>
<td>3.40</td>
<td>78.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.96E-01</td>
<td>4.55</td>
<td>80.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_A10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.27E-01</td>
<td>3.35</td>
<td>81.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_B02</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>7.55E-02</td>
<td>3.00</td>
<td>81.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.10E-02</td>
<td>4.55</td>
<td>81.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>8.22E-02</td>
<td>4.65</td>
<td>82.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_B10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.34E-01</td>
<td>3.60</td>
<td>82.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_E09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.93E-01</td>
<td>3.60</td>
<td>82.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_B07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.06E-01</td>
<td>3.35</td>
<td>82.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_B05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.23E-01</td>
<td>3.65</td>
<td>86.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_H07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.31E-01</td>
<td>3.45</td>
<td>87.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_D05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.77E-02</td>
<td>2.95</td>
<td>87.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_D10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.57E-01</td>
<td>3.80</td>
<td>87.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_A08</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>4.64E-02</td>
<td>4.75</td>
<td>88.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_C11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.99E-02</td>
<td>3.20</td>
<td>88.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_D01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>6.42E-02</td>
<td>3.10</td>
<td>88.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_D03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>5.17E-02</td>
<td>3.20</td>
<td>88.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007065_H09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>9.35E-02</td>
<td>4.75</td>
<td>89.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_A05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.38E-01</td>
<td>3.20</td>
<td>89.00</td>
</tr>
</tbody>
</table>

Average: 3.55 | 61.59
b.) Sporadic PDAC Samples

<table>
<thead>
<tr>
<th>Pass/Fail</th>
<th>Sample</th>
<th>Status</th>
<th>Data Type</th>
<th>Quality</th>
<th>Ploidy</th>
<th>Contamination (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIL</td>
<td>0007066_E06</td>
<td>Unprocessed</td>
<td>Affymetrix MIP ASCAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIL</td>
<td>0007066_F02</td>
<td>Unprocessed</td>
<td>Affymetrix MIP ASCAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_E05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.76E-01</td>
<td>4.45</td>
<td>80.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_E09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.54E-01</td>
<td>4.05</td>
<td>86.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_E11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>2.33E-01</td>
<td>3.20</td>
<td>74.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F04</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.08E-01</td>
<td>4.75</td>
<td>76.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.26E-01</td>
<td>2.50</td>
<td>50.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F08</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.11E-01</td>
<td>3.20</td>
<td>81.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.47E-01</td>
<td>2.95</td>
<td>77.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_F11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.52E-01</td>
<td>2.85</td>
<td>61.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_G03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.04E-01</td>
<td>2.90</td>
<td>88.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_G04</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.41E-01</td>
<td>4.75</td>
<td>84.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_G07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.01E-01</td>
<td>4.15</td>
<td>84.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_G09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.28E-01</td>
<td>2.45</td>
<td>4.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_G11</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.12E-01</td>
<td>3.40</td>
<td>81.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H01</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.51E-01</td>
<td>2.90</td>
<td>63.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H03</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.61E-01</td>
<td>4.70</td>
<td>25.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H05</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.22E-01</td>
<td>2.95</td>
<td>48.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H07</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.10E-01</td>
<td>4.70</td>
<td>66.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H09</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.34E-01</td>
<td>3.35</td>
<td>65.00</td>
</tr>
<tr>
<td>PASS</td>
<td>0007066_H10</td>
<td>Processed</td>
<td>Affymetrix MIP ASCAT</td>
<td>1.35E-01</td>
<td>3.10</td>
<td>74.00</td>
</tr>
</tbody>
</table>

**Average** 3.54 66.68

4.2.2 ASCAT FPC CNV OUTPUT TRACING:

As previously described, the output in Figure 9 demonstrates a genome-wide CNV profile for an individual FPC sample. The inclusion of all 51 FPC cases analyzed on ASCAT generates an absolute copy number profile demonstrated in Figure 10. The upper panel demonstrates a genome wide view of all CNV changes with the baseline centered on 0%. Genomic gains are represented by the blue traces demonstrating an upward deflection past the +50% level. Upon careful examination, two distinct superimposed blue traces are present. The lighter blue tracing represents the presence of copy gain as defined in the methods section as an LRR>0.3 with the darker blue region demonstrating high copy gain regions defined by a LRR>1.2. The presence of high copy gains is indicative of the underlying aneuploidy of the samples. Conversely, traces in red showing a negative deflection below 0% denote copy loss regions. Although, copy loss regions are infrequent compared to gains, careful examination demonstrates copy loss in at least one sample throughout the genome. Furthermore, only select regions demonstrate recurrent copy loss in the FPC samples. Frequent recurrent regions of loss in this tracing occur on chromosome 9p21.3 (19.61%), 17p13.1 (9.8%) and 18q21.2 (15.69%) (circled). These regions carry many genes, including the three most commonly described TSGs in PDAC, namely CDKN2A, TP53.
and SMAD4, respectively. Other regions of frequent copy loss listed in ascending chromosomal order include: 1p36.22-1p35 (9.8%), 3p25.1 (9.8%), 3p22.3-22.1 (9.8%), 6q16.1-16.3 (9.8%), 6q21-26 (9.8%), 9p24.3 (11.76%), 9p24.1-23 (15.69%), 9p21.1 (17.65%), 9p13.3-13.1 (15.69%), 17p13.3-13.2 (11.76%), 18q12.1-12.3 (15.69%), 18q12.3-21.32 (15.69%), 18q21.33 (17.65%), 18q22.2-22.3 (15.68%), Xp21.1 (17.65%), Xp11.3 (11.76%), Xp11.23 (27.45%). The bottom panel illustrates a chromosome summary view of all CNV changes in the FPC samples including those described above. Furthermore, a region of higher frequency copy loss (43.14%) is demonstrated on both the genome and chromosome summary view on chromosome Xq28 that includes the Factor VIII hemophilia A gene (boxed). A complete summary of all gains and loss regions is shown in Table S2 (See Appendix Section).

Figure 10: ASCAT CNV Output Tracing for FPC cases

a.) Genome Summary View
b.) Chromosome summary view

Figure 10 demonstrates only absolute CN changes. Figure 11 represents a genome wide allelic ratios plot both in genome summary view (upper panel) and chromosome summary view (lower panel). In the genome summary view, once again the baseline is centered on 0%. The purple tracing demonstrating an upward deflection represents allelic imbalances between tumour and normal profiles, reflective of ploidy changes, copy gains and losses. Negative deflections, as outlined in orange, shows regions of CN differences with splitting of the BAF in the absence of LRR changes. As previously described, this would be consistent with CN-LOH. Figure 11 highlights the complete genomic involvement by CN-LOH with a region affected in at least one sample. Combined with Figure 10, a loss region is present throughout the entire FPC genome with varying frequency and distribution among cases. Higher frequency loss regions are once again seen on chromosome 9p21.3 (43.14%), 17p13.1 (50.98%) and 18q21.2 (41.18%) corresponding to CDKN2A, TP53 and SMAD4. Other selected recurrent regions of CN-LOH in ascending chromosomal order include: 6q15-16.2 (29.41%), 6q16.3-21 (31.37%), 6q21-23.2 (33.33%), 6q25.1 (35.29%), 6q26 (35.29%), 6q27 (29.41%), 8p23.3 (33.33%), 8p23.2-23.1 (33.33%), 8p22 (31.37%), 8p21.3-21.2 (29.41%), 9p24.3-24.1 (47.06%), 9p23 (49.02%), 9p21.3 (43.14%), 9p21.1 (41.18%), 9p13.3-13.2 (37.25%), 17p13.3-13.1 (50.98%), 17p12 (52.94%), 17p11.2 (43.14%), 18q12.1-12.3 (35.29%), 18q21.1-21.2 (41.18%), 18q21.32 (43.14%), 18q21.33-22.22.3 (43.14%), 18q22.3-23 (41.18%), Xp11.3 (31.37%), Xp11.23 (43.14%), Xq28
(37.25%). A complete summary of genome wide allelic ratios is shown in Table S3 (See Appendix Section).

Figure 11: ASCAT Allelic Ratios CNV Output Tracing for FPC Cases

a.) Genome Summary View

b.) Chromosome Summary View
4.2.3 ASCAT SPORADIC CNV OUTPUT TRACING:

Figure 12 demonstrates a genome wide CNV output profile tracing for 19 sporadic PDAC cases on ASCAT. The top two panels show a genome and chromosome summary view for absolute CN changes. Once again copy gain and high copy gain regions are present throughout the genome re-affirming ploidy changes present in the tumour samples. A summary of copy gain regions is shown in Table S4 (See Appendix Section). Copy loss regions are scattered throughout the genome and aside from the known loci of importance in PDAC, namely CDKN2A (15.79%; CN LOH: 47.37%) and SMAD4 (15.79%; CN LOH: 68.42%) (circled), recurrent loss was observed along the X-chromosome in as many as 31.58% of samples. Interestingly, the chromosome 17p region carrying p53 did not display high frequency absolute CN loss (5.26%).

Select regions of recurrent loss listed in ascending chromosomal order include: 9p24.3-24.1 (15.79%), 9p23 (21.05%), 9p21.3-21.1 (21.05%), 9p13.3 (15.79%), 17q25.3 (15.79%), 18q12.1 (15.79%), 18q23 (21.05%), 19p13.3 (15.79%), Xp22.2 (15.79%), Xp22.2 (31.59%), Xp22.11 (26.32%), Xp21.3-21.1 (15.79%), Xp11.23 (31.59%), Xq13.1 (10.53%), Xq13.2 (26.32%), Xq13.2 (26.32%), Xq23 (15.79%), Xq26.2 (15.79%), Xq27.1 (15.79%), Xq28 (31.59%).

An examination of allelic ratios both in the genome and chromosome summary view (Figure 12: bottom two panels) demonstrates allelic imbalances representing ploidy changes, gains and losses (purple tracing) throughout the genome. Similar to FPC, genome-wide CN-LOH (orange tracing) is present and increases both the frequency and distribution of loci affected by loss. Once again, the majority of the PDAC genome is affected by loss in at least one PDAC sample consistent with FPC. Repeated regions of loss once again center on chromosomes harboring many genes and notably CDKN2A and SMAD4. When examining copy neutral changes, the TP53 locus once again demonstrates a high frequency loss (63.16%). Other selected regions of recurrent copy neutral loss in ascending chromosomal order include: 3p26.3 (36.84%), 3p25.3 (42.11%), 3p24.2-24.1 (42.11%), 3p24.1-21.31 (36.84%), 3p21.1-11.1 (42.11%), 6q14.3-27 (42.11%), 8p23.3-23.2 (42.11%), 8p23.1-22 (42.11%), 9p24.3-21.2 (47.37%), 9p13.3-13.2 (36.84%), 13q14.2 (42.11%), 13q31.3-34 (42.11%), 15q11.2 (36.84%), 15q14 (36.84%), 15q21.1-22.2 (36.84%), 17p13.2-11.2 (63.16%), 17q22-23.1 (36.84%), 17q23.2-24.1 (42.11%), 17q25.2-25.3 (57.89%), 18q11.2 (42.11%), 18q11.2-12.1 (52.63%), 18q12.1 (57.89%), 18q12.3-21.1 (68.42%), 18q21.2 (73.68%), Xp22.2-22.11 (36.84%), Xp11.23 (47.37%), Xq13.2...
(42.11%), Xq28 (47.39%). A complete summary of absolute CN loss and allelic ratios is shown in Table S4 and S5, respectively (See Appendix Section).

Figure 12: ASCAT Sporadic CNV Output Tracing

a.) Genome Wide Absolute CN Changes Summary View
b.) Absolute CN Changes Chromosome Summary View

c.) Allelic Ratios Genome Wide Summary View
4.3 FPC SNP-FASST2 Segmentation Copy Number Profile

4.3.1 FASST2 FPC CNV OUTPUT TRACING:

Unlike ASCAT, the SNP-FASST2 segmentation CNV algorithm is not designed specifically for cancer samples. Therefore predictions of tumour ploidy and percent normal cell contamination are not performed. This HMM based model does not generate an ASCAT like profile where, as was seen in the previous section, contamination can lead to a “no solution” call. Therefore, all matched tumour/normal samples that passed the Oncoscan processing stage (MAPD<0.6) were included on the FASST2 analysis. Figure 13 demonstrates the copy number profiles for the 55 matched FPC cases. As with the ASCAT tracings, the baseline values are centered on 0% with gains seen as blue positive deflections and losses as red negative deflections. What is immediately apparent between the FASST2 tracings and the ASCAT profiles is the absence of high copy gains present genome-wide, which represented tumour aneuploidy on ASCAT. Furthermore, higher frequency gain regions occur at select regions in the genome including chromosomes: 1p12-11.2 (9.09%), 1q21.1 (10.91%), 1q21.2-22 (12.73%), 1q23.1 (18.18%), 1q23.2 (20%), 1q24.1-24.2 (16.36%), 1q31.1 (14.55%), 1q31.2 (18.18%), 1q41 (16.36%), 1q43-
Consistent with the ASCAT profiles, loss regions are scattered throughout the majority of the tumour genome with higher frequency loss regions once again occurring on chromosomes 9p21.3 (36.36%), 17p13.1 (25.45%) and 18q21.2 (27.27%) involving CDKN2A, TP53 and SMAD4 (circled). Other regions of recurrent loss include chromosomes: 1p36.22-36.21 (12.73%), 1p21.3 (10.91%), 2p11.1 (10.91%), 3p25.1 (10.91%), 3p25.1-24.3 (12.73%), 3p24.1 (12.73%), 3p22.2 (14.55%), 3p14.2 (12.73%), 6p25.2 (10.91%), 6q15 (12.73%), 6q16.1 (12.73%), 6q21 (12.73%), 6q23.2 (12.73%), 6q25.1 (12.73%), 8p23.3-23.1 (16.36%), 8p22 (18.18%), 8p21.3 (20%), 9p24.3 (14.55%), 9p24.1-21.3 (18.18%), 9p21.3 (32.73%), 9p21.2 (27.27%), 9q33.2-33.3 (10.91%), 10q23.1 (10.91%), 11q23.3 (10.91%), 15q11.1-11.2 (12.73%), 17p13.3-13.2 (20%), 17p13.1 (23.64%), 17p12 (20%), 18p11.23-11.21 (10.91%), 18q11.2 (10.91%), 18q12.1 (12.73%), 18q12.1-12.3 (18.18%), 18q21.1 (20%), 18q21.33 (27.27%), 18q23 (21.82%), 21q22.11-22.3 (10.91%), Xp11.4 (12.73%), Xq24-25 (10.91%). The previous high frequency loss region on chromosome Xq28 corresponding to the factor VIII gene, called on ASCAT, does not appear to display a high frequency loss on the FASST2 algorithm (5.45% vs. 43.14% on ASCAT). A summary of copy gain and loss regions on the FASST2 algorithm is summarized in Table S6 (See Appendix Section).
Figure 13: SNP FASST2 Segmentation FPC CNV Output Tracing

a.) Genome-wide Summary View

An examination of the allelic ratios plot (Figure 14) once again demonstrates genome-wide allelic imbalances indicating the presence of copy gain and loss. As ploidy levels are not predicted on the FASST2 algorithm, the magnitude of the upward deflection is smaller than that
seen in the ASCAT algorithm. Interestingly, aside from a small region on chromosome 5p15.31-15.1, 7q33-36.1, 12p13.33-13.31, 12p13.2-11.22, 15q25.3-26.2 and 18q21.1, each occurring in a single sample, no CN-LOH regions were called on the FASST2 algorithm. Therefore interalgorithm CNV calling variation exists and can be significant in terms of CNV regions identified.

Figure 14: SNP FASST2 Segmentation FPC Allelic Ratios CNV Plot

a.) Genome Wide Summary View

b.) Chromosome Summary View
4.3.2 FASST2 SPORADIC CNV OUTPUT TRACING:

Consistent with ASCAT, a matched paired analysis of 21 sporadic PDAC cases on the FASST2 algorithm demonstrated CN gain and loss regions scattered throughout the genome with varying frequency. Recurrent regions of chromosomal gain on the FASST2 algorithm include the following regions: 1q22 (38.10%), 1q32.1 (42.86%), 3q21.3 (38.10%), 5p15.33 (38.09%), 7q22.1 (33.33%), 13q14.2 (33.33%), 14q11.2 (33.33%), 16p13.3 (33.33%), 17q11.2 (33.33%), 17q21.2 (38.10%), 17q21.31 (42.86%), 17q21.22 (38.10%), 19p13.11 (33.33%), 20q11.21 (33.33%), 20q13.13 (33.33%), 20q13.31 (38.10%). Conversely, the most frequent regions of loss once again center around known loci in PDAC including CDKN2A (38.1%), TP53 (33.33%) and SMAD4 (23.81%) (circled in the genome wide view). Higher frequency loss are also observed along chromosome 1p36.33-36.32 (28.57%), 1p36.22 (23.81%), 1p36.12 (23.81%), 2p25.1 (23.81%), 3p25.3 (23.81%), 3p21.31-21.1 (23.81%), 3p16.1 (28.57%), 5p15.33 (28.57%), 5q32 (23.81%), 6p21.31 (23.81%), 6q16.1 (23.81%), 6q16.3 (23.81%), 7p22.3 (23.81%), 8p23.3-23.2 (23.81%), 8p23.1 (28.57%), 9p24.3 (23.81%), 9p23-22.2 (23.81%), 9p21.3 (38.10%), 9q12-13 (28.57%), 9q34.2-34.3 (33.33%), 11p15.5 (23.81%), 11q12.2-12.3 (23.81%), 12p13.33-13.32 (23.81%), 16p13.3 (28.51%), 17p13.3 (28.51%), 17q21.2 (23.81%), 17q25.3 (33.33%), 18q12.1-12.2 (33.33%), 18q21.2-21.32 (38.10%), 19p13.3 (28.57%), 19q13.13-13.2 (23.81%), 19q13.32-13.33 (28.57%), 22q11.21 (28.57%), 22q11.23 (23.81%) and 22q13.1-13.32 (28.57%). An examination of allelic ratios in these sporadic cases demonstrates a small number of CN-LOH regions including: 6p21.1 (4.76%), 8p11.21-11.1 (4.76%), 8q24.21 (4.76%), 9p22.1-21.3 (4.76%) and 19q12-13.11 (4.76%). The above findings are shown in Figure 15, where the top two panels illustrate absolute CN changes and the bottom two panels show allelic ratios. A summary of all copy changes called on the FASST2 algorithm in sporadic PDAC can be found in Table S7 (See Appendix Section).
Figure 15: SNP FASST2 Segmentation Sporadic Allelic Ratios CNV Plot

a.) Genome-Wide Summary View

b.) Chromosome Summary View
4.4 Sporadic vs. FPC CNV Profile Comparison

As a first method of identifying regions of interest in FPC, a direct comparison of CNV profiles between FPC and Sporadic PDAC samples was performed using a Fisher Exact Test at the \( \alpha = 0.05 \) level, accounting for genome wide multiple hypothesis testing using the false discovery rate method. Table 5 summarizes the results obtained on the Fisher test for copy loss regions on both algorithms.
Table 5: Fisher Exact Test demonstrating significant regions of LOH difference between sporadic PDAC and FPC CNV profiles

a.) ASCAT Algorithm

<table>
<thead>
<tr>
<th>Region</th>
<th>Event</th>
<th>Genes</th>
<th>Freq. in &lt;Hereditary&gt; (%)</th>
<th>Freq. in &lt;Sporadic&gt; (%)</th>
<th>p-value</th>
<th>q-bound</th>
<th>% of CNV Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>chrX:15,366,179-15,847,887</td>
<td>CN Loss</td>
<td>11</td>
<td>2</td>
<td>31.57894737</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>17.58429697</td>
</tr>
<tr>
<td>chrX:15,847,887-16,167,262</td>
<td>CN Loss</td>
<td>2</td>
<td>6</td>
<td>31.57894737</td>
<td>0.010541064</td>
<td>0.704494455</td>
<td>0</td>
</tr>
<tr>
<td>chrX:24,381,416-24,610,170</td>
<td>CN Loss</td>
<td>37</td>
<td>2</td>
<td>31.57894737</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>100</td>
</tr>
<tr>
<td>chrX:24,610,170-25,242,925</td>
<td>CN Loss</td>
<td>3</td>
<td>0</td>
<td>63.15789474</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>0</td>
</tr>
<tr>
<td>chrX:153,277,507-153,296,363</td>
<td>CN Loss</td>
<td>3</td>
<td>0</td>
<td>63.15789474</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>40.49639372</td>
</tr>
<tr>
<td>chrX:153,296,363-153,554,661</td>
<td>CN Loss</td>
<td>8</td>
<td>0</td>
<td>31.57894737</td>
<td>0.010541064</td>
<td>0.704494455</td>
<td>57.12936221</td>
</tr>
<tr>
<td>chrX:153,554,661-153,770,540</td>
<td>CN Loss</td>
<td>3</td>
<td>0</td>
<td>63.15789474</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>45.10397028</td>
</tr>
<tr>
<td>chrX:153,770,540-154,133,139</td>
<td>CN Loss</td>
<td>27</td>
<td>4</td>
<td>31.57894737</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>49.5611904</td>
</tr>
<tr>
<td>chrX:154,194,800-155,270,560</td>
<td>CN Loss</td>
<td>32</td>
<td>4</td>
<td>31.57894737</td>
<td>0.001304098</td>
<td>0.174314464</td>
<td>45.10397028</td>
</tr>
</tbody>
</table>

b.) SNP FASST2 Segmentation Algorithm

<table>
<thead>
<tr>
<th>Region</th>
<th>Event</th>
<th>Genes</th>
<th>Freq. in &lt;Hereditary&gt; (%)</th>
<th>Freq. in &lt;Sporadic&gt; (%)</th>
<th>p-value</th>
<th>q-bound</th>
<th>% of CNV Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5:1,223,484-1,298,376</td>
<td>CN Loss</td>
<td>3</td>
<td>0</td>
<td>28.57142857</td>
<td>2.48E-04</td>
<td>0.41612861</td>
<td>25.95871388</td>
</tr>
<tr>
<td>chr9:65,629,772-66,457,628</td>
<td>CN Loss</td>
<td>0</td>
<td>0</td>
<td>28.57142857</td>
<td>2.48E-04</td>
<td>0.41612861</td>
<td>100</td>
</tr>
<tr>
<td>chr9:136,197,049-137,107,123</td>
<td>CN Loss</td>
<td>25</td>
<td>5.454545455</td>
<td>33.33333333</td>
<td>0.00352982</td>
<td>0.66657639</td>
<td>53.71969752</td>
</tr>
<tr>
<td>chr16:0,1,833,938</td>
<td>CN Loss</td>
<td>83</td>
<td>1.818181818</td>
<td>28.57142857</td>
<td>0.001418358</td>
<td>0.66657639</td>
<td>96.72829725</td>
</tr>
<tr>
<td>chr16:1,833,938-2,211,485</td>
<td>CN Loss</td>
<td>32</td>
<td>0</td>
<td>28.57142857</td>
<td>2.48E-04</td>
<td>0.41612861</td>
<td>100</td>
</tr>
<tr>
<td>chr16:3,075,016-3,133,833</td>
<td>CN Loss</td>
<td>5</td>
<td>0</td>
<td>28.57142857</td>
<td>2.48E-04</td>
<td>0.41612861</td>
<td>3.580597446</td>
</tr>
<tr>
<td>chr17:76,869,327-76,900,349</td>
<td>CN Loss</td>
<td>2</td>
<td>0</td>
<td>7.272727273</td>
<td>33.33333333</td>
<td>0.007936337</td>
<td>0.85147832</td>
</tr>
<tr>
<td>chr17:78,790,085-79,195,619</td>
<td>CN Loss</td>
<td>11</td>
<td>0</td>
<td>7.272727273</td>
<td>0.007936337</td>
<td>0.85147832</td>
<td>100</td>
</tr>
<tr>
<td>chr17:79,397,488-80,242,188</td>
<td>CN Loss</td>
<td>41</td>
<td>5.454545455</td>
<td>33.33333333</td>
<td>0.00352982</td>
<td>0.66657639</td>
<td>85.81249961</td>
</tr>
<tr>
<td>chr18:58,873,900-58,884,950</td>
<td>CN Loss</td>
<td>0</td>
<td>0</td>
<td>10.90909091</td>
<td>38.0952381</td>
<td>0.016595145</td>
<td>0.85147832</td>
</tr>
<tr>
<td>chr22:18,782,628-19,185,395</td>
<td>CN Loss</td>
<td>12</td>
<td>1.818181818</td>
<td>28.57142857</td>
<td>0.001418358</td>
<td>0.66657639</td>
<td>100</td>
</tr>
</tbody>
</table>

The above traces were generated on the Nexus CNV software package Version 6.1. Regions along with CNV events are demonstrated in the first 2 columns with the number of genes affected by the specific loss shown in the third column. The final column of the table shows the % CNV overlap with a known CNV annotated in the DGV. Assuming that FPC is caused by one or a small collection of genes represented only in a proportion of sporadic cases, an FPC CNV of interest in this analysis would be observed at a higher loss frequency than in sporadic samples. From the above data, no regions of loss satisfied this hypothesis in both the ASCAT and
FASST2 algorithm. Conversely, significant regions of CN loss occurred more frequently in sporadic cases than in FPC. A complete list of CN Loss genes is shown in Table S8 and Table S9 (See Appendix Section). From this data, although there exist a small number of regional differences between sporadic and FPC, the vast majority of CNV changes present in both sample sets occur with frequencies of no statistically significant difference.

### 4.5 Comparison with Microsatellite Marker CNV Profile

As another method of identifying novel previously undescribed regions with genomic loss, a visual comparison was made with the published somatic FPC data, obtained using MS markers (Abe, Fukushima et al. 2007). MS markers are capable of only identifying absolute CN changes and therefore all CN-LOH discovered in this study were novel. Figure 16 summarizes the CNV profiles present in the Abe et al study superimposed with genomic losses identified on Oncoscan on both the ASCAT and SNP-FASST2 Segmentation algorithms. As chromosomal resolution with MS markers are quite low, peak loss changes per chromosomal arm with Oncoscan was reported to appropriately match the MS marker data resolution. Although similar overlap is displayed at various genomic loci in both the ASCAT and FASST2 algorithm, the frequency varies between studies. Furthermore, an examination of Oncoscan specific changes called between the ASCAT and FASST2 algorithms, once again demonstrates frequency and distribution loss differences demonstrating inter-algorithmic variation.

Regions of loss previously identified with MS markers include chromosomes: 1q, 2p, 4q, 5p, 6q, 7q, 9p, 15q, 17p, 18q, 19p and Xp (Abe, Fukushima et al. 2007). Consistent with our study, higher frequency loss in the MS marker study centers on chromosome 9p, 17p and 18q. In contrast, the Oncoscan data demonstrates loss (either absolute or CN-LOH) throughout the FPC genome with novel involvement of particular chromosomes (e.g. chr:3). Both algorithms demonstrate these findings although with varying locus frequency. Furthermore, as the reported Oncoscan data represents peak regions, intra-chromosomal frequency variation between loci exist and would not be appreciated in Figure 16. In aggregate, these novel regions of loss contain hundreds of putative candidate FPC TSGs (not shown) and in the absence of an overwhelming high frequency loss region, this strategy remains ineffective in formulating a manageable non-arbitrary candidate TSG list.
Figure 16: MS marker vs. Oncoscan SNP array FPC CNV profile

a.) Comparison with loss regions called on ASCAT

b.) Comparison with loss regions called on SNP-FASST2 Segmentation
4.6 Shared LOH Regions

One strategy for CNV/gene filtering was by examining shared LOH regions in members of the same family, called on one or both CNV algorithms. In total, 18 of 55 FPC cases belonged to one of 9 families (each family having 2 members represented on Oncoscan). CNVs were examined in both relatives and if a shared loss region was present, this region was cross-referenced with the other families and the entire FPC cohort. From the 9 families, only 2 generated a CNV profile with shared loss regions and these regions were identified only on the FASST2 algorithm. The regions identified in each family were unique to that family. The pedigrees for the respective families are shown in Figure 17. The affected relatives belonging to family 545P include a father-son pair whereas the relatives in family 385P are sisters.

Figure 17: Pedigrees of families demonstrating shared LOH regions on the SNP-FASST2 Segmentation Algorithm

a.) Family 545P
b.) Family 385P

Examination of the family 545P pair demonstrates a single shared loss region spanning the chromosome 18q21.2 locus. A more detailed analysis of this region demonstrates three putative genes: ME2, ELAC1 and SMAD4. As the SMAD4 locus is a commonly loss region in PDAC, this region was not pursued in this study. Furthermore, germline SMAD4 mutations have been described however not in the context of HPC syndromes (See chapter 1; Section 1.6.3).

Microarray analysis of the family 385P siblings demonstrates three shared loss regions. The first region spans the chromosome 3p22.2-22.3 locus which includes three putative genes including: DCLK3, TRANK1 and EPM2A1P1. The second shared loss region in the siblings involves chromosome 9p21.3, which corresponds to the CDKN2A locus. As this locus has already been described in HPC syndromes, this region was excluded from further analysis. The final locus involved the chromosome 17p13.3 region including: CRK, MYO1C, PITPNA, SLC43A2, PRPF8, SERPINF1, SERPINF2 and SMYD4. Microarray tracings demonstrating shared loss regions in families 545P and 385P are illustrated in Figure 18, whereas relevant gene annotations in the family 385P shared LOH regions are outlined in Table 6.
Figure 18: Shared Loss Regions identified on the SNP-FASST2 Segmentation Algorithm

a.) Family 545P Shared Loss Region

![Diagram of Family 545P Shared Loss Region](image)

b.) Family 545P Shared Loss Region Magnified (chromosome 18q21.2)

![Diagram of Family 545P Shared Loss Region Magnified](image)

c.) Family 385P Shared Loss Regions

![Diagram of Family 385P Shared Loss Regions](image)
d.) Family 385P Shared Loss Region Magnified (Chromosome 3p22.2-22.3)

![Gene Annotations](image)

Table 6: Gene Annotations for the Shared LOH regions in Family 385P

<table>
<thead>
<tr>
<th>Gene</th>
<th>Location</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCLK3</td>
<td>Chr 3: 36753912-36781352</td>
<td>Serine/Threonine Kinase Activity; ATP Binding</td>
</tr>
<tr>
<td>TRANK1</td>
<td>Chr 3: 36868307-36986548</td>
<td>Hydrolase Activity; ATP Binding</td>
</tr>
<tr>
<td>EPM2AIP1</td>
<td>Chr 3: 37027356-37034795</td>
<td>Encodes Laforin; Mutated in an autosomal recessive form of adolescent progressive myoclonal epilepsy</td>
</tr>
</tbody>
</table>
4.7 FISH Confirmation of Shared LOH Regions

Shared LOH regions of interest in family 385P were confirmed using BAC generated FISH probes (See Table S1 in the Appendix Section). The DCLK3 locus on chromosome 3p22.2-23.3 and the SERPINF1 locus on chromosome 17p13.3 were selected to confirm predicted shared loss regions called by the SNP-FASST2 algorithm. Each region of interest was assayed on paraffin slides from both siblings and the results for the chromosome 3 region are illustrated in Figure 19 and in Figure 20 for the chromosome 17 region.

Figure 19: FISH assay for the shared LOH region on chromosome 3p22.2-23.3 using a DCKL3 specific BAC clone

a.) Peripheral Blood lymphocyte control demonstrating proper localization of BAC probes
b.) Family 385P106 (Sister 1) Normal Paraffin Slide Profile

c.) Family 385P106 Tumour Paraffin Slide
Peripheral blood lymphocyte DNA (Figure 19 a.) was used to ensure proper localization of the control CEP3 probe (Red) and DCLK3 probe (green) to the same chromosome. Normal paraffin slide (panel b.) confirms a diploid state of both the control CEP3 probe and the DCLK3 gene which is a surrogate for the chromosome 3p22.2-22.3 region. Tumour paraffin slide for sister 1 (panel c.) demonstrates multiple cells with the majority demonstrating hemizygous deletion of the DCLK3 probe. One characteristic cell has been expanded in panel c.) to demonstrate this hemizygous deletion. As can be observed, two CEP3 markers are present in the nucleus in contrast to one DCLK3 marker. Panel d.) represents the FISH image capture for sister #2. An expanded picture of a dominant clone displays an aneuploid cell with multiple CEP3 markers (5) but also relative loss of the DCLK3 marker (2) confirming a deleted state.
Figure 20: FISH assay for the shared LOH region on chromosome 17p13.3 using a SERPINF1 specific BAC clone

a.) Peripheral Blood lymphocyte control demonstrating proper localization of BAC probes

b.) Family 385P106 (Sister 1) Normal Paraffin Slide Profile
c.) Family 385P106 Tumour Paraffin Slide

Once again, control peripheral blood lymphocyte cells (panel a.) confirms the proper localization of the CEP17 centromeric probe (green) and the SERPINF1 BAC probe (red) to the chromosome 17 region. Normal paraffin tissue (panel b.) confirms the diploid state of both probes with a characteristic cell magnified confirming a double copy of each probe. FISH assay of sister 1 (panel c.) demonstrates the presence of 2 centromeric probes and only 1 SERPINF1 probe. Hemizygous loss of this locus on the slide is confirmed. Similarly, FISH assay on tumour tissue...
for sister 2 (panel d.) confirms the presence of hemizygous loss of the SERPINF1 (and thus the chromosome 17p13.3) locus on the slide. Therefore predicted changes made on the FASST2 CNV calling algorithm are consistent with the underlying tumour biology.

4.8 2-Hit Hypothesis Genes

Another major gene filtering strategy involved cross-referencing CNV data of individual cases with matched germline exome data available through ongoing initiatives in our laboratory. At the time of thesis completion, a total of 33 germline cases (27 Toronto + 6 Montreal) with exome data were available, however only 6 cases were also analyzed on Oncoscan. Candidate gene(s) of interest had to harbor a rare inactivating variant defined as: nonsense generating SNVs, frameshift indels or splice-site alterations. Rare was defined as the absence of the variant in dbSNP135 or at a frequency of <1% in the 1000 exomes project or NHLBI EVS server database (See Chapter 3; Section 3.9). Case specific germline variants were cross-referenced with matched tumour Oncoscan data and gene(s) with a rare variant(s) and a somatic LOH region were retained. Results of this filtering strategy are summarized in Table 7. A complete list of case specific CNVs can be found in Table S10 (See Appendix Section). CNV regions in each individual algorithm or shared CNVs between algorithms were included for analysis. Ideally, regions shared between both CNV algorithms were of greater importance than regions called on a single algorithm. A total of 6 genes representing 4 cases (ATPAF1-AS1, C3ORF72, GSDMA, FAM186A, POGK, VANGL1) satisfied the above criteria and each gene remained unique to a case (i.e. no gene was shared among these 6 cases). Upon examining the remaining 27 exomes, which did not have tumour Oncoscan data, another rare germline variant of ATPAF1-AS1 was discovered. Each gene was identified from CNV regions called on the FASST2 algorithm only (i.e. no LOH region with a candidate 2-hit gene was called on both algorithms or solely on the ASCAT algorithm).
4.9 Sanger Sequencing Validation of 2-Hit Genes

Candidate TSGs described in section 4.9 were confirmed in germline lymphocyte DNA with conventional Sanger sequencing. Confirmation electropherograms are illustrated in Figure 21. PCR amplification was possible in all candidate TSGs except FAM186A and VANGL1. This occurred due to a predominance of repetitive sequences near the variant of interest in FAM186A and VANGL1, preventing the generation of a region specific primer (i.e. the primer hybridized non-specifically to many regions of the genome). Germline variants were confirmed in patient specific DNA for both variants of ATPAF1-AS1 and for each single variant of C3ORF72 and GSDMA. The predicted POGK variant was found to be wild-type.
Figure 21: Sanger Sequencing Electropherograms for Candidate 2-Hit Hypothesis Genes

a.) ATPAF1-AS1 exon2;c.486_487insC;p.L162fs (Reverse Tracing-Insertion G)

b.) ATPAF1-AS1 exon2;c.577delG
c.) GSDMA 38126784; G>Deletion (Reverse Tracing- C Deletion)

![GSDMA Reverse Tracing](image)

d.) C3ORF72 exon2;c.200_209del;p.67_70del

![C3ORF72 Sanger Tracing](image)

e.) POGK exon5;c.C1676A (Wild-type Sanger Tracing)

![POGK Sanger Tracing](image)
4.10 FISH Confirmation of Somatic LOH of Putative 2-Hit TSGs

Regions of somatic LOH in candidate 2-hit TSGs were confirmed by FISH on tumour paraffin slides. Genes that had germline Sanger validation were selected for somatic FISH analysis. BAC probes for each gene were created and assayed on regular unstained slides containing paraffin tissue. Specific BAC clones used are described in chapter 3 (Section 3.11) and Table S1. Peripheral blood lymphocyte cells were used to confirm proper localization of the control centromeric probe and the region of interest (See Figure S1 in the Appendix Section). FISH results for the putative genes of interest are shown for tumour slides in Figure 22 along with normal tissue images.

Figure 22: FISH Results for candidate 2-Hit genes demonstrating somatic LOH

a.) Normal slide for case S06-21896 (corresponding to ATPAF1-AS1 region)

b.) Tumour slide for ATPAF1-AS1
c.) Normal slide for case S07-16089 (corresponding to the GSDMA region)

d.) Tumour slide for GSDMA probing
e.) Normal Slide for case 05-1679 (corresponding to the C3ORF72 region)

f.) Tumour Slide for C3ORF72 probing

Panel a.) demonstrates FISH images for the normal tissue slide of the ATPAF1-AS1 case. FISH confirms the diploid state of the non-cancer tissue as evident by two copies of the CEP1 probe (red) and two copies of the ATPAF1-AS1 probe (green). In contrast, panel b.) illustrates FISH images for the ATPAF1-AS1 locus in the pancreatic tumour tissue. The image demonstrates a cluster of cells with varying FISH patterns demonstrating clonal heterogeneity. Furthermore,
aneuploidy is observed by the presence of multiple control probes at the chromosome of interest. A characteristic tumour cell is amplified to demonstrate aneuploidy of the tumour cell but also the presence of relative CN loss at the ATPAF1 locus (only one 1 green probe present for 5 control red probes). Panel c.) represent the FISH pattern on the normal paraffin tissue slide for the case corresponding to the GSDMA locus. Once again, diploidy is demonstrated by the presence of 2 copies of the control CEP17 probe (green) and 2 copies of the GSDMA probe (red). Similarly, tumour tissue FISH patterns demonstrates diploidy which is inconsistent with the predicted call on Nexus. Panel d.) represents the normal tissue for the C3ORF72 case. A baseline diploid state is confirmed as evident by 2 copies of the CEP3 probe (red) and 2 copies of the C3ORF72 probe (green) in the majority of cells. Tumour FISH pattern at the C3ORF72 locus is inconsistent with the predicted hemizygous loss call on Nexus. This is evident by the retention of 2 green probes in the majority of cells in the image.

4.11 Somatic GISTIC Analysis and Gene Filtering

One strategy for identifying CNV regions of interest was by using the GISTIC 2.0 statistical algorithm. This model examines a particular data set for regions of interest that occur more likely than would be expected by chance alone. Through a complex set of mathematical permutations, GISTIC stratifies gain regions based on frequency of occurrence and magnitude. As such, a region with high copy gain would be ranked higher than a region with a lower copy gain assuming their frequencies are relatively similar. A loss region is stratified based on frequency of loss and the presence of homozygous vs. hemizygous deletion. Regions of interest are tested for genome-wide significance using the false discovery rate method for multiple hypothesis testing. In this study, GISTIC regions were obtained on both the ASCAT and FASST2 algorithms and shared GISTIC regions between these two algorithms were given higher priority for subsequent confirmation and validation. GISTIC regions are summarized in Table 8 including a set of CN loss regions with a minimum of 20% overlap between both calling algorithms. A complete set of GISTIC genes per called region is available in Table S11 and S12 for ASCAT and FASST2, respectively (See Appendix Section).
Table 8: Summary of GISTIC Loss Regions on Nexus 6.1

a.) GISTIC Regions based on the ASCAT Algorithm

<table>
<thead>
<tr>
<th>Region</th>
<th>Extended Region</th>
<th>Type</th>
<th>Q-Bound</th>
<th>G-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1:97,486,735-97,588,871</td>
<td>chr1:97,463,800-99,798,853</td>
<td>Loss</td>
<td>0.038525344</td>
<td>4.746199613</td>
</tr>
<tr>
<td>chr1:28,282,292-28,596,084</td>
<td>chr1:27,766,773-28,653,864</td>
<td>Loss</td>
<td>0.042179124</td>
<td>4.719470112</td>
</tr>
<tr>
<td>chr3:21,280,793-21,345,398</td>
<td>chr3:18,501,095-25,512,119</td>
<td>Loss</td>
<td>0.001390448</td>
<td>5.812390748</td>
</tr>
<tr>
<td>chr3:69,806,635-69,841,522</td>
<td>chr3:69,486,414-69,930,995</td>
<td>Loss</td>
<td>0.009005419</td>
<td>5.213976531</td>
</tr>
<tr>
<td>chr6:99,526,222-99,690,312</td>
<td>chr6:98,528,389-100,417,417</td>
<td>Loss</td>
<td>2.56E-05</td>
<td>6.974887805</td>
</tr>
<tr>
<td>chr8:6,480,174-6,602,077</td>
<td>chr8:6,233,286-6,779,799</td>
<td>Loss</td>
<td>3.94E-05</td>
<td>6.873629564</td>
</tr>
<tr>
<td>chr8:15,484,030-15,561,772</td>
<td>chr8:14,268,661-15,561,772</td>
<td>Loss</td>
<td>4.11E-05</td>
<td>6.864549182</td>
</tr>
<tr>
<td>chr17:11,507,099-11,917,647</td>
<td>chr17:10,213,929-14,294,567</td>
<td>Loss</td>
<td>3.25E-08</td>
<td>8.445168548</td>
</tr>
<tr>
<td>chr17:48,654,321-49,734,442</td>
<td>chr17:48,645,767-51,413,143</td>
<td>Loss</td>
<td>6.81E-12</td>
<td>10.27488706</td>
</tr>
<tr>
<td>chrX:154,133,139-154,194,800</td>
<td>chrX:154,133,139-154,194,800</td>
<td>Loss</td>
<td>2.22E-12</td>
<td>22.99974984</td>
</tr>
</tbody>
</table>

b.) GISTIC Regions based on the SNP-FASST2 Segmentation Algorithm

<table>
<thead>
<tr>
<th>Region</th>
<th>Extended Region</th>
<th>Type</th>
<th>Q-Bound</th>
<th>G-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1:26,948,921-27,271,359</td>
<td>chr1:24,708,348-28,879,780</td>
<td>Loss</td>
<td>3.08E-04</td>
<td>5.861550027</td>
</tr>
<tr>
<td>chr8:23,243,246-23,516,440</td>
<td>chr8:21,057,891-31,983,747</td>
<td>Loss</td>
<td>7.93E-06</td>
<td>6.698400013</td>
</tr>
<tr>
<td>chr10:89,564,384-89,675,443</td>
<td>chr10:84,708,696-91,297,263</td>
<td>Loss</td>
<td>0.005523234</td>
<td>5.048000001</td>
</tr>
<tr>
<td>chr17:7,983,481-8,096,876</td>
<td>chr17:0,22,193,626</td>
<td>Loss</td>
<td>5.84E-13</td>
<td>10.31970002</td>
</tr>
<tr>
<td>chr18:44,092,849-44,253,110</td>
<td>chr18:31,175,899-57,133,122</td>
<td>Loss</td>
<td>5.84E-13</td>
<td>10.66199999</td>
</tr>
<tr>
<td>chr19:7,217,196-7,252,904</td>
<td>chr19:0,9,516,323</td>
<td>Loss</td>
<td>3.17E-05</td>
<td>6.412950005</td>
</tr>
<tr>
<td>chr21:34,551,133-34,585,079</td>
<td>chr21:15,386,949-14,129,895</td>
<td>Loss</td>
<td>0.009256727</td>
<td>4.890199916</td>
</tr>
<tr>
<td>chrX:37,648,774-37,697,671</td>
<td>chrX:2,699,968-58,545,809</td>
<td>Loss</td>
<td>7.50E-04</td>
<td>5.613100002</td>
</tr>
<tr>
<td>chrX:154,133,959-154,158,520</td>
<td>chrX:61,944,035-155,270,560</td>
<td>Loss</td>
<td>0.040363505</td>
<td>4.429899978</td>
</tr>
</tbody>
</table>
c.) GISTIC Regions demonstrating at least 20% overlap between algorithms

Although GISTIC 2.0 provides a method of formulating a candidate CNV list and highlights regions of recurrent significance in an individual data set, the number of candidate GISTIC genes still remains unmanageable for confirmation and validation. As such, GISTIC regions (either from individual algorithms or shared between algorithms) were cross-referenced with rare germline deleterious variants identified by examining all 33 FPC exomes available. Gene(s) with a rare inactivating variant, as previously described, shared between 2+ individuals and shown to be associated with a GISTIC 2.0 region in the tumour data set were selected as candidates. Only 2 genes satisfied the above requirements and were identified only on the FASST2 algorithm: Tenascin C (TNC) and MAP3K6. None of the shared GISTIC regions on both ASCAT and FASST2 contained a gene with a rare deleterious variant in the exome data. Table 9 provides a summary of these genes including the rare variant observed in the exome data, frequency of samples showing loss in the 55 FPC somatic cases and gene function.

Table 9: Summary of GISTIC Genes with Rare Germline Inactivating Variants

<table>
<thead>
<tr>
<th>Gene</th>
<th>Location</th>
<th>Function</th>
<th>Exome Variant</th>
<th>Frequency of LOH (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNC</td>
<td>chr9: 117781853-117880536</td>
<td>Extracellular Matrix protein, cell adhesion</td>
<td>Split Site: C&gt;A</td>
<td>7.27</td>
</tr>
<tr>
<td>Tenascin C</td>
<td>chr9: 117781853-117880536</td>
<td>Extracellular Matrix protein, cell adhesion</td>
<td>Frameshift Insertion: _&gt;G</td>
<td>7.27</td>
</tr>
<tr>
<td>MAP3K6</td>
<td>chr1: 27681669-27693337</td>
<td>Serine/Threonine Kinase involved in signal transduction</td>
<td>Frameshift Insertion: _&gt;G</td>
<td>14.55</td>
</tr>
<tr>
<td>MAP3K6</td>
<td>chr1: 27681669-27693337</td>
<td>Serine/Threonine Kinase involved in signal transduction</td>
<td>Non-Sense SNV: C&gt;T</td>
<td>14.55</td>
</tr>
</tbody>
</table>
4.12 GISTIC Genes Germline and Somatic Confirmation

The aforementioned genes with shared rare germline variants in 2+ individuals and high frequency somatic loss identified by GISTIC 2.0, underwent subsequent confirmatory assays. Germline DNA was Sanger sequenced to confirm each variant called by ES and the somatic LOH region was confirmed using FISH. Figure 23 and 24 illustrates the results from Sanger sequencing and FISH respectively. In total, 7/8 cases that demonstrated a predicted CN loss at the MAP3K6 locus had paraffin slides available for FISH analysis.

Figure 23: Sanger Sequencing of Germline Tissue corresponding to GISTIC Gene variants

a.) TNC Splice Site Variant (C>A predicted; shown as wild-type in this image)

b.) TNC Frameshift Insertion (>_G predicted; shown as wild-type in this image)
c.) MAP3K6 Frameshift Insertion (\_>G predicted; shown as wild-type in this image)

Panel a.) and b.) demonstrate Sanger sequencing results for the germline TNC variants identified on ES. A predicted rare splice site variant (panel a.) and frameshift insertion indel (b.) are found to be wild-type, illustrated by the black arrows. This is inconsistent with the predicted call by ES.

Panel c.) and d.) represent Sanger sequencing results for the germline MAP3K6 variants identified on ES. The rare frameshift insertion indel (panel c.) is found to be wild-type (black arrow) therefore failing to confirm the predicted call by ES. Panel d. demonstrates the presence of a C>T deleterious non-sense variant that has been confirmed by Sanger sequencing as illustrated by the black arrow.

d.) MAP3K6 Non-Sense Variant (C>T)
Figure 24: FISH Results for Candidate GISTIC Genes with Rare Germline Variants

a.) characteristic normal paraffin slide (Sample 08-24208)

b.) Tumour Slide for case 08-24208
c.) Tumour Slide for Sample SS03-799

d.) Tumour slide for sample SP02 13681

e.) Tumour slide for sample 06-21896
f.) Tumour slide for sample TB12-G

g.) Tumour slide for sample 05-1679
h.) Tumour slide for sample MSSH 07-225

Panel a.) demonstrates a characteristic non-cancer cell FISH pattern for the MAP3K6 locus. 2 copies of the control CEP1 centromeric probe (red) and 2 copies of the MAP3K6 BAC clone (green) are observed throughout the slide. Similarly, panel b.) demonstrates a FISH pattern for the corresponding tumour slide, which fails to demonstrate hemizygous deletion. This is inconsistent with the predicted loss on GISTIC. On this slide, there continues to remain 2 copies of the MAP3K6 locus with an isolated gain of CEP3 (3 copies). Panel c.), d.), e.) illustrate a relative loss of the MAP3K6 locus. On these slides, gain/aneuploidy of the tumour cell is shown by multiple centromeric markers with a non-equivalent and reduced number of MAP3K6 probes. In panel c.), a characteristic magnified cell demonstrates the presence of approximately 8 CEP1 probes with only 2 MAP3K6 probes. Similarly, a characteristic cell in panel d.) demonstrates 4 CEP1 probes with only 2 MAP3K6 probes. Panel e.) displays 6 CEP1 probes with only a single MAP3K6 signal. Panel g.) and h.) demonstrate a FISH pattern consistent with hemizygous deletion whereby 2 control CEP1 probes are present with only 1 MAP3K6 probe. Finally, panel f.) demonstrates a FISH pattern negative for hemizygous deletion therefore failing to confirm the GISTIC predicted LOH call.
Chapter 5
Data Analysis

5 Discussion

5.1 FFPE DNA and Microarray Analysis in PDAC

The purpose of this study was to characterize the somatic copy number profile of pancreatic tumours and to outline genomic differences between FPC and sporadic PDAC. An integrative genomic strategy was applied to identify novel regions of loss, which may harbor TSGs as the basis of FPC tumourigenesis. Central to this investigation was the use of FFPE tumor DNA for somatic microarray analysis.

FFPE DNA has historically been used largely for cytological architecture studies and candidate gene based investigations, however genome-wide investigations have been met with challenges. The process of formalin fixation leads to cross-linking of amino acid residues, which prepares the tissue for various pathological studies such as immunohistochemistry. As a by-product, nucleic acid moieties are also cross-linked, shattered and fragmented, rendering downstream genomic applications difficult. Previous microarray investigations failed to generate interpretable data largely due to: challenges with DNA/RNA quantity and length for PCR amplification, subsequent PCR failure due the presence of formalin cross links, and challenges in generating specific PCR products due to non-specific primer hybridization. All these factors result in a low microarray signal to noise ratio, preventing meaningful analysis.

Our study is the first large-scale genome wide investigation of PDAC using microarrays and FFPE DNA. By utilizing the Affymetrix Oncoscan 660K MIP SNP microarray, we have successfully profiled the tumour genome of 55-matched FPC and 21 sporadic PDAC specimens. Furthermore, we have added to a small but growing body of literature describing the potential of somatic FFPE DNA for studies in cancer biology and genetics. Previous studies of FFPE DNA in pediatric atypical rhaboid/teratoma tumours and adult granulosa cell tumours of the ovary (Hasselblatt, Isken et al. 2013, Geiersbach, Jarboe et al. 2011) were able to successfully confirm known genetic changes at the basis of disease. Furthermore, these studies were able to identify novel regions of potential importance in their respective cancer type. As a proof of concept, we have demonstrated that microarray analysis of FFPE DNA is feasible in PDAC and generates
data capable of interpreting the tumour genome. Furthermore, as PDAC is an aggressively fatal
disease with only a minority of patients undergoing biopsy or surgical resection, we have
demonstrated that successful analysis of FFPE DNA with Oncoscan is capable of increasing the
power of studies in PDAC from which, novel CNV regions can be identified. As an example, the
only other known somatic study in FPC (Abe, Fukushima et al. 2007) had a sample size of 20
fresh frozen tumours. In this study, the initial sample size was 158 FPC tumours with matched
normal tissue, however, following filtering for tumour cellularity requirements (>70%) and the
availability of unstained slides or tumour blocks, 80 matched FPC samples remained. Greater
than 1/3 of our tumour samples were >5 years old. We therefore anticipate the use of FFPE FPC
tumour DNA in future large-scale consortium based efforts will identify an even greater number
of meaningful somatic changes.

5.2 Sample Characteristics and Microarray Analysis in PDAC

5.2.1 NORMAL CELL CONTAMINATION AND TUMOUR PLOIDY

This study has demonstrated the genomic tumour profile for both a cohort of FPC and sporadic
PDAC samples using 2 separate CNV calling algorithms. One of these algorithms, ASCAT, has
been developed specifically for CNV detection in cancer tissue by accounting for tumour ploidy
and normal cell contamination. Analysis with the ASCAT algorithm highlighted two major
characteristics in our cohort.

First, although concerted attempts at microdissection were made to select for tumour samples
with >70% cellularity, over half of our samples had >50% normal cell admixture as
demonstrated in Table 4. Not surprisingly, those samples that underwent LCM demonstrated a
normal cell admixture <30%, which was the intended study cut-off value. The ASCAT model is
robust at CNV calling and the initial description of this algorithm demonstrated accurate calls in
a cell line model with >80% normal cell contamination (Van Loo, Nordgard et al. 2010).

Therefore, although a limitation of this study, we do not believe that contamination compromised
the CNV calls made although there is a possibility that a higher level of false negatives may have
been observed in comparison to a more pure tumour cohort. A recent study has demonstrated that
only an approximate 30% correlation exists between pathological estimates and true tumour
purity as determined by KRAS deep sequencing (Song, Nones et al. 2012). Although tumour
cellularity estimate algorithms have been generated, such as qPure, our findings in combination with the available evidence, suggest that future studies using hypocellular cancers like PDAC should employ LCM to ensure the most accurate tumour CNV calls.

Second, the ASCAT algorithm demonstrated that both FPC and PDAC samples contain a varying level of aneuploidy from near diploid to tetraploid tumours and possibly higher DNA content as shown in Table 4. The true consequence of aneuploidy in cancer remains poorly understood although it is believed to be a hallmark of genomic instability (Gordon, Resio et al. 2012). Furthermore, a varying level of aneuploidy between cases may also suggest that FPC/PDAC (and likely other cancers) represent a collection of many clonal diseases with unique biology that appears as a phenotypically homogeneous disease (see section 5.2.2 below). In this study, we did not assess whether there were differences in terms of clinical presentation, response to treatment, severity of disease between diploid and aneuploid tumours or whether aneuploidy is a response to chemo-radiation treatment. This information would be interesting in further investigations for understanding the biology of disease. Nevertheless, our findings suggest that the two-hit model for a TSG may be too simplistic for all cases in a particular cancer and that perhaps a gene dosage or relative loss model would be more consistent with tumour biology.

5.2.2 GENETIC HETEROGENEITY

Although only one cancer specific calling algorithm was used in our study, both calling algorithms (ASCAT and SNP-FASST2 Segmentation) were able to demonstrate genome wide genetic heterogeneity and recurrent genetic loss at regions including the known TSGs in PDAC, CDKN2A/p16, TP53 and SMAD4.

Genetic heterogeneity was evident in both FPC and sporadic tumour profiles as low frequency involvement of a portion of the genome in at one least sample. When empirically comparing two samples in our cohort, no pair had an identical tumour genome profile, which mirrors our aneuploidy findings in section 5.2.1. Furthermore, when excluding regions of known loss in PDAC, no loci demonstrating loss of a putative TSG recurred at a consistently high frequency. The involvement of the entire genome merited additional filtering strategies to identify important predicted driver genes for carcinogenesis. One possible interpretation of this data has been
postulated by Yachida et al. whereby FPC and sporadic PDAC do not constitute a single disease that undergoes a consistent, homogenous set of genetic events. Disease may arise from many sub-clones within a tumour, all possessing a unique genetic signature in the background of a common genetic backbone (Yachida, Jones et al. 2010). It is therefore possible that this heterogeneity could exist not only within a PDAC tumour, but also between PDAC tumours. Thus, PDAC could be viewed as a disease defined by clonal mosaicism manifesting with a common phenotype termed clinically as PC. Previous microarray studies would also support this idea. In a study of 27 microdissected sporadic PDAC specimens (Harada, Chelala et al. 2008) analyzed on an 116,000 SNP microarray platform, high frequency loss and homozygous deletion was observed at the chromosome 9p21.3 locus harboring CDKN2A (63% of cases). Furthermore, recurrent loss was also observed in DCC (48%), SMAD4 (33%) and TP53 (26%). The remainder of genetic changes were dispersed throughout the tumour genome and did not recur with any frequency, precluding the identification of a consistent gene. The authors thus concluded that genetic changes previously known in PDAC were the most recurrent changes in their sample set.

Therefore, one possible limitation of our study is that, if PDAC is a disease of inter and intra-tumoural clonal mosaicism, the Oncoscan somatic profile solely represents a clone present in a particular tumour section block for that respective case. It is possible, therefore, that our study is not representing the most dominant clone present in a tumour (if any). This is more significant for biopsy samples where an even smaller fraction of the tumour is represented in the extracted DNA. By extension, sampling various tumour blocks of the same patient representing various geographic locations of the tumour may show a different CNV profile, re-affirming the clonal nature of disease. This will undoubtedly further complicate our understanding of the significance of these CNV changes.

Based on the aforementioned, genetic analysis of FPC and sporadic PDAC may require sorting of cases based on tumour characteristics identified at surgery or during imaging, mode and method of presentation, and by factoring in patient history including epidemiologic risk factors. This strategy of analyzing pancreatic tumour characteristics and stratifying based on a common phenotype may result in the ability to detect common genotypic changes, which account for disease causation in these cohorts. We have already discussed the possibility of comparing aneuploid with diploid tumours in section 5.2.1, which could be an initial grouping method. In a
recent study of 20 PDAC tumours analyzed on a high density 500K SNP microarray (Gutierrez, Munoz-Bellvis et al. 2011), somatic genome profiles of these tumours were correlated with relevant clinical and histo-pathological features of disease. Frequent regions of loss in these tumours included chromosomes 1p, 9p, 12q, 17p and 18q. Once again multiple CNV and LOH regions were identified in every case. However, when factoring in the relevant clinical and histological features, PDAC tumours clustered into 2 distinct groups. The first group contained small, low to intermediate grade tumours whereas the second group contained large, high-grade cancers. In addition, this strategy highlighted regions of loss containing putative TSGs occurring at a frequency greater than the common 3 known loci, in respective cohorts.

In our study, a sub-group analysis of cases within the same pedigree appears to have provided preliminary data supporting the concept of “private” mutations or genomic changes as the basis of disease in specific FPC cohorts. As illustrated in Figure 17 and 18, 2 families (out of 9 initial families) each with 2 relatives represented on the Oncoscan array, generated analyzable data. An evaluation of the tumours of 2 sisters in family 385P demonstrated three regions of shared LOH, one of these corresponded to the CDKN2A locus. Of the two other loci, shared regions were observed at chromosome 3p22.2-22.2 and 17p13.3.

The chromosome 3p LOH region has not been previously described in FPC (see section 5.3) and this shared locus contains three genes. DCLK3 belongs to a family of proteins with serine/threonine kinase and ATP binding activity. DCLK3 has been implicated with increased invasion and metastasis in pituitary tumours (Wierinckx, Auger et al. 2007). Furthermore, mutations of this gene have also been documented in the COSMIC database as part of various large-scale ES studies in prostate, CRC, melanoma and gastric cancer. DCLK3 belongs to a large family of proteins of which one, DCLK1, has recently been shown to mediate epithelial to mesenchymal transition in PDAC (Sureban, May et al. 2011). Therefore, as a member of this gene family, DCLK3 was of interest for further confirmation. The other shared region on chromosome 17p consisted of 8 genes of which one, SERPINF1, was initially selected as a gene of interest, based on its documented involvement in sporadic PDAC (Gutierrez, Munoz-Bellvis et al. 2011). This gene has also been associated with cervical cancer and multiple myeloma in the COSMIC database. SERPINF1 mediates angiogenic signaling and is degraded by matrix metalloprotease, involved in cancer metastasis. Figure 19 and 20 confirm the predicted somatic shared loss of the above regions using FISH. Although the chromosome 17p region shows
hemizygous deletion in both sisters, the chromosome 3p region shows hemizygous deletion in the tumour for sister 1 however sister 2’s tumour demonstrates aneuploidy with relative loss of the 3p locus. As normal non FFPE DNA was unavailable for these cases, confirmation of a germline “1st hit” was not performed in this investigation. Taken together, these findings support the notion of tumour aneuploidy, inter-tumoural heterogeneity and the possibility of “private” mutations among a certain subset of FPC individuals.

5.3 Comparisons of FPC with Sporadic PDAC and CN-LOH

5.3.1 MOLECULAR PROFILE OF FPC VS. SPORADIC PDAC

In a recent molecular study of PanIN II/III pancreatic lesions in FPC obtained at prophylactic pancreatectomies, whole transcriptome profiling demonstrated many up-regulated and down-regulated genes of interest compared to normal pancreas (Crnogorac-Jurcevic, Chelala et al. 2013). When comparing the molecular profiles of FPC PanINs to sporadic PanINs, up to 185 dysregulated genes were shared, representing approximately 25% overlap. This study demonstrated the molecular similarity of FPC to sporadic PDAC. These findings supported the initial data presented by Schwartz et al. whereby, through a pathological and epidemiological approach, familial and sporadic carcinoma shared overlapping pathways in CRC (Schwartz, Henson 2007).

Our investigation represents the most extensive molecular analysis of FPC with a direct comparison to sporadic PDAC. An empiric similarity of FPC and sporadic PDAC CNV profiles was evident on both the SNP-FASST2 segmentation and ASCAT algorithms. Furthermore, a Fisher Exact test, for data called on the ASCAT algorithm, demonstrated a significant difference on the X-chromosome representing approximately 133 genes. Similarly, a Fisher Exact test performed on calls made on the SNP-FASST2 segmentation algorithm demonstrated scattered regions of significant difference, as illustrated in Table 5, representing 232 genes. Although a putative candidate list could be created for genes likely possessing tumour suppressor properties, this was not performed as loss regions occurred more frequently in sporadic cases than in FPC. At the study outset, we hypothesized that causative genes in FPC could be identified when comparing FPC CNV profiles to sporadic PDAC; as sporadic disease likely possessed a greater genetic heterogeneity. Therefore, we believed there existed significant regional differences
containing higher risk FPC genes accounted for in a smaller portion of sporadic cases. Although previous studies have alluded to a similar loss profile between FPC and sporadic PDAC (see below), we anticipated that with the detection of CN-LOH regions, a consistent set of loss regions with higher frequency in FPC vs. sporadic would be appreciated. In our study, it is possible, however, that differences in both algorithms observed could have been attributed to a sample size difference between the two cohorts (55 FPC vs. 21 sporadic). Therefore, if a repeat analysis was performed with an equal number of sporadic and FPC samples, these significant regional differences would be eliminated or possibly more prevalent in the FPC cohort.

The somatic profile of FPC tumours in a previous investigation (Abe, Fukushima et al. 2007) empirically demonstrated a similar overall profile to historic sporadic data with a slightly higher frequency of genome-wide LOH in FPC. Furthermore, regional differences in the FPC group were most notable at the chromosome 2p and 19p arm when compared to sporadics, however a formal test of comparison was not performed. Nevertheless, the authors of this study concluded that regions targeted for loss in FPC overall correspond closely to those regions in sporadic PDAC. In our study, it is notable to mention that the chromosome 2p and 19p regions did not demonstrate a significant difference in loss between the FPC and sporadic cohort with a Fisher comparison. The finding of a similar somatic profile between FPC and sporadic PDAC is not entirely surprising considering our understanding of other cancer models such as CRC. The molecular framework of carcinogenesis in CRC is similar between hereditary (FAP) and sporadic cases, however the occurrence of cancer differs due to when the predisposing gene is mutated. In the case of FAP, this event (APC mutation) likely occurs in-utero either by inheritance or de-novo mutation whereas for sporadic disease, the initiating even occurs post-partum (Aretz, Uhlhaas et al. 2004). It is conceivable, therefore, that a similar model can be applied to FPC and sporadic PDAC.

One limitation of this study, and any study examining FPC, arises inherently from the simplistic definition of an FPC patient. FPC is defined clinically by the presence of 2+ FDR with PDAC who do not belong to an HPC syndrome. Determining a genetic cause to a disease clustered clinically is filled with many challenges as outlined in this study. Specifically, locus heterogeneity is likely more extensive as the study population is also quite heterogeneous. Equally as important, is the difficulty to identify potential phenocopies representing sporadic disease. In our cohort, it is possible that a subset of cases clumped into the FPC group actually
represent sporadic PDAC by virtue of being phenocopies. However these cases belong to a family where another individual has been identified with PDAC, thus satisfying the definition for FPC. Furthermore, by adhering to the PACGENE definition of FPC (2+ affected blood relatives with PDAC) as was done in this study, we increase our likelihood of including phenocopies in our FPC cohort since our criteria for inclusion to the hereditary group is relaxed. This could also account for the extensive locus heterogeneity observed and the challenge of identifying an FPC gene, as genetic changes in true FPC cases are diluted by sporadic PDAC samples.

5.3.2 CNV PROFILE IN FPC

The central aim of this investigation was to analyze FPC tumours using a novel high-density microarray platform. We suspected that such a method would unravel novel regions of loss previously unknown and undetected by more archaic methodologies. We have successfully demonstrated that detailed regions of loss can be identified using FFPE DNA on SNP microarrays. We have shown that genetic heterogeneity in FPC is extensive and, in aggregate, most of the tumour genome is affected by copy loss. Furthermore, within a region of interest, frequency variation in loss exists between adjacent loci.

Figure 16 shows a graphical representation of peak regions affected by CN loss in this study compared to the MS marker data published by Abe et al. (Abe, Fukushima et al. 2007). From these figures illustrating data from both the ASCAT and FASST2 algorithm, we notice, generally, that previous high frequency loss regions described with MS markers are also represented by high frequency loss in the Oncoscan data (specifically chr: 1p, 6q, 9p, 17p, 18q and Xp for the ASCAT algorithm and chr: 9p, 17p and 18q for the FASST2 algorithm). In addition, Figure 16 also successfully proves the hypothesis that novel regions of copy loss exist that have not been previously detected by MS markers (both absolute and copy neutral). Novel regions were observed with the highest frequency on chromosome 3p, 8p, 9q and Xq with the ASCAT algorithm and along chromosome 3p, 8p, 9q, 10q and 14q with the FASST2 algorithm. Unfortunately, the number of putative TSGs at these novel loci in aggregate was extensive and remained unmanageable to validate without additional filtering strategies.

An important finding from our data is the presence of CN-LOH, highlighted in green on Figure 16 with the ASCAT algorithm. CN-LOH accounts for a substantial portion of CNV changes in our data set and likely represents a mechanism of LOH in the previous MS marker study that
could not be detected. Our study is therefore the first to demonstrate CN-LOH in FPC and to demonstrate that CN-LOH affects both regions previously known in FPC and novel regions not discovered in the Abe study. CN-LOH in sporadic PDAC has been described and found to occur at a frequency of 40% when analyzed on a low-density array ((Harada, Chelala et al. 2008) to as high as 70% in PDAC derived cell lines (Calhoun, Hucl et al. 2006). Our dataset would suggest that employing CNV detection platforms incapable of detecting CN-LOH would miss a considerable portion of genetic changes possibly with implications in understanding disease biology.

As an example, CN-LOH has recently been shown to play an important role in sporadic CRC development in MSI tumours (Melcher, Hartmann et al. 2011). This CRC subtype was classically thought to be chromosomally stable and scarce of copy loss regions. An analysis of CN-LOH regions identified MTUS1 as a putative TSG in the aforementioned study playing a role in MLH1 inactivation. The authors of this study hypothesized that maintaining a copy neutral state of the defective locus promotes a further growth advantage to the cancer cell. Similarly, CN-LOH of the p53 locus in esophageal squamous carcinoma was found to be the major mechanism of allelic inactivation in cancers with p53 mutations (Saeki, Kitao et al. 2011). Although our study was successful in demonstrating the presence of CN-LOH, we failed to demonstrate a consistent high frequency region of CN-LOH, as hypothesized, containing a putative TSG. Nevertheless, future studies examining the functional role of CN-LOH and the possible advantageous properties for cancer to maintain a copy neutral vs. hemizygous state may shed light on the pathobiology of pancreatic tumorigenesis.

5.4 Integrative Genomics
5.4.1 2-HIT HYPOTHESIS MODEL

The recognized complexity of cancer genomics in adult solid malignancies has signaled the need for integration of many data sets in order to understand disease pathogenesis. Integrative genomics is a general description for the use of multiple gene discovery platforms including DNA sequencing, microarray technologies, epigenetic techniques and clinico-epidemiological tools as a method of understanding cancer etiology. It is believed that an integrative genomics model is effective at pooling across multiple heterogeneous datasets, which appears to be a
consistent attribute of cancer (Liu, Huang et al. 2013). Integrative genomic analysis in sporadic oral squamous cell carcinoma (OSCC) utilizing microarray platforms, methylation techniques and WES, was capable of expanding the list of known somatic events in this disease (Pickering, Zhang et al. 2013). Furthermore, 4 novel driver pathways and 2 novel genes were identified allowing for the description of a new class of OSCC. Similar findings using various integrative approaches has also been described in cervical cancer (Lando, Wilting et al. 2013), ovarian clear cell carcinoma (Sung, Choi et al. 2013) and neuroendocrine tumours of the small intestine (Banck, Kanwar et al. 2013).

Our study applied an integrative genomics approach utilizing both WES and microarray data as a filtering method for potential novel FPC TSGs, satisfying the Knudson 2-hit hypothesis model. We have successfully proven our hypothesis, that by combining germline ES with somatic microarray CNV analysis, a candidate TSG shortlist in a heterogeneous disease can be formulated. Our first method of filtering genes was by analyzing cases with both germline exome data and somatic Oncoscan data. This strategy identified 6 novel candidate TSGs (Table 7) of which only 1 (ATPAF1-AS1) was confirmed by both germline Sanger sequencing validation and somatic FISH for LOH ((Figure 21 a.) and Figure 22 b.) respectively. ATPAF1-AS1 codes for an anti-sense RNA, which interacts with ATP-AF1, important in cell membrane integrity. Little is known about the tumour suppressor properties of ATPAF1-AS1 although a missense mutation has been described in the COSMIC database in a case with endometrial cancer. Further analysis of this gene in our germline exome data set, identified another case with a rare deleterious variant which was confirmed by Sanger sequencing (Figure 21 b.). Unfortunately, this case was not represented in our Oncoscan analysis and therefore efforts are underway to examine for copy loss in the somatic tissue by FISH.

Interestingly, we did not observe any shared genes satisfying the 2-hit hypothesis between cases that had both germline WES and somatic Oncoscan data. Our initial impression was that an important gene in FPC would demonstrate a recurrent deleterious variant in the germline tissue with a consistent shared LOH at this locus in the somatic tissue. However, the number of cases with both WES and somatic array data was limited (n=6). Coupled with the extensive locus heterogeneity present, it is possible that by examining more tumour/normal pairs with our current integrative strategy, recurrence of our gene of interest will be observed. Alternatively, such a finding may further support the model that “private mutations” in a specific cohort of individuals...
is present, accounting for only a small proportion of disease (See Chapter 5, section 5.2). However, our findings do not prove the “private” mutations concept and any firm conclusions would be premature. Further investigations involving more FPC tumour/normal pairs are required.

As described in chapter 1, only one other study in sporadic PDAC has employed an integrative approach and this method was effective in confirming known genetic changes in PDAC and as well identifying novel mutations in axon guidance pathways (Biankin, Waddell et al. 2012). Up until this study, such methodologies in FPC have been challenging largely due to the paucity of resectable cases encountered each year and the difficulty in appropriately utilizing archived samples, which are largely FFPE. Our study is therefore important not only in that we have identified a novel FPC gene but have also demonstrated that integrative approaches utilizing archival somatic DNA is possible, meaningful and informative (See Chapter 5, section 5.1).

5.4.2 GISTIC FILTERING OF PUTATIVE TSGs

Another novel method of gene filtration in this investigation involved the GISTIC algorithm. GISTIC is a statistical model attempting to identify those regions and genes that are gained or lost with presumed biological significance (see Chapter 3, section 3.9). The first description of GISTIC was in a glioblastoma model, which identified novel regions in this disease. The algorithm demonstrated concordance between previous published glioblastoma studies, which individually were quite heterogeneous and discrepant in their findings (Beroukhim, Getz et al. 2007). By extension, GISTIC models have been utilized in many cancer types including intra-hepatic cholangiocarcinoma as an example. Cholangiocarcinoma is another aggressively fatal hepatobiliary cancer, and GISTIC was successfully able to identify genomic regions, which stratified disease into 2 separate classes. Each class had distinct clinical implications and prognostic sequelae (Sia, Hoshida et al. 2013).

Based on these reports, our study attempted to filter the observed FPC locus heterogeneity using the GISTIC algorithm as a method of identifying important regions of genetic loss. Our data set demonstrated 17 GISTIC regions on the FASST2 algorithm and 19 regions on the ASCAT algorithm, corresponding to 4373 and 173 genes, respectively, as summarized in Table 8, S11 and S12. Although this narrowed the list of potential candidate genes and partly mediated the genetic heterogeneity, the data obtained was still too extensive to validate without additional
stratification measures. As an additional filtering strategy, we examined our data of 33 germline FPC exome cases and cross-referenced any genes with a rare inactivating variant shared between 2+ individuals, to our GISTIC data demonstrating high frequency loss. Two such genes with TSG potential were identified from the FASST 2 algorithm only, TNC and MAP3K6. Sanger validation failed to confirm the germline variant in TNC and confirmed 1/2 variants in MAP3K6. Tumour tissue was available for 7/8 cases with a predicted loss in MAP3K6 and FISH analysis confirmed loss in 5/7 cases. Interestingly, our FISH patterns, although demonstrating involvement at the MAP3K6 locus, were not uniform among these 5 cases. 4 cases demonstrated genome wide aneuploidy with amplification of the control probe and relative loss of the MAP3K6 locus. The other remaining case was more consistent with a genome wide diploid state and hemizygous deletion of the MAP3K6 locus (See Figure 24). The biological significance, if any, of these varying FISH patterns is unclear and would require further investigation as discussed in section 5.2. The findings obtained from these two independent data sets (germline exome and somatic CNV) suggest that MAP3K6 potentially follows a 2-hit TSG model. Our study is limited, however, in that the cases with germline MAP3K6 variants did not have tumour tissue represented in the Oncoscan data set. Therefore, we could not confirm whether cases with germline variants also had LOH of MAP3K6 in their tumours. We therefore applied importance to this gene by the observation that recurrent rare germline mutations are found in FPC patients, coupled with the GISTIC identification of high frequency loss in a group of somatic FPC tissue.

MAP3K6 has a serine threonine protein kinase activity and is involved with signal transduction and ATP binding as part of the MAP kinase family. Little has been described regarding the TSG function of MAP3K6. In a previous genome wide RNA interference study targeting modulators of Vascular Endothelial Growth Factor (VEGF), MAP3K6 was found to interact and synergize with VEGF promoting angiogenesis (Eto, Miyagishi et al. 2009). As such, inhibition of MAP3K6 resulted in a downregulation of VEGF, reduced endothelial cell proliferation and, as the authors describe, decreased tumour cell growth in-vivo. The aforementioned study was not performed in a PDAC model. Integrative analysis of the gastric cancer “kinome” using NGS to identify SNVs and CNVs, also demonstrated the involvement of MAP3K6 although the functional role of this gene was not elucidated in this study (Zang, Ong et al. 2011). It is well known that PDAC is a hypovascular tumour with hypoxia potentially serving as an initiating and propagating event for disease (Guillaumond, Vasseur 2013). Furthermore, tumour hypoxia has
shown to correlate with metastasis in a PDAC murine model (Buchler, Reber et al. 2004). Therefore, it is possible that MAP3K6 inhibition in FPC/PDAC impedes angiogenesis, which, either, promotes the initial tumour formation or increases intra-tumoural hypoxia that favours tumour growth. All together, these findings point to the potential role of MAP3K6 as a putative TSG in FPC and potentially sporadic PDAC, however further investigations, including functional experiments, are required.

6 Conclusion

PDAC remains a lethal malignancy with most patients presenting at an advanced stage. It is anticipated that with improvements in screening, diagnosis and therapeutics for the major common cancers in North America (lung, breast, CRC, prostate), PDAC will become the leading cause of cancer mortality. In this thesis, we have outlined a series of genetic investigations in FPC, a hereditary form of PDAC, with the belief that novel regions of chromosomal loss containing putative TSGs exist, explaining the biology of disease.

We have successfully demonstrated that genome-wide molecular investigation of FFPE DNA in PDAC is possible. Ours is the largest known somatic series in FPC utilizing a novel high density SNP array. We have demonstrated that through this approach, novel regions of chromosomal loss are present harboring hundreds of genes. Furthermore, we have demonstrated that copy neutral LOH is present in these tumours and that utilizing platforms incapable of detecting these changes would miss a sizeable portion of genetic loss. Indeed, these regions may also harbor putative TSGs of interest. As a consequence of using a genome wide approach, we have also shown that the somatic profiles of FPC and sporadic PDAC are similar and that locus heterogeneity is extensive. This has introduced the possibility of “private” mutations accounting for a proportion of disease. As a consequence, the importance of stratifying cohorts, in order to mitigate such heterogeneity has been highlighted and supported through an analysis of tumours from 2 sisters. This investigation identified 2 loci of interest containing DCLK3 and SERPINF1 respectively. Finally, by utilizing an integrative strategy with WES and following the Knudson 2-Hit hypothesis model for TSGs, we have shown that gene filtering using this strategy is effective and have identified ATPAF1-AS1 and MAP3K6 as putative TSGs in FPC.
7 Future Directions

The extensive dataset generated in this study coupled with the success of our methodology naturally lead to a series of further investigations. At the time of completion, of this MSc work, we were in the process of analyzing another 55 FPC germline exomes totaling the number of cases to 88. By increasing the number of germline exomes available, this will also increase the number of samples with both exome and Oncoscan data from 6 (in this study) to 17. The results of these investigations may further support ATPAF1-AS1 as a putative TSG if other cases are found to possess a rare germline variant with subsequent copy loss in the tumour tissue. Furthermore, by cross-referencing our somatic GISTIC data with the new germline exome data, we may identify novel putative TSGs satisfying our 2-hit hypothesis model.

The genes of interest identified through our integrative approach (ATPAF1-AS1 and MAP3K6) have putative TSG properties although their true functional impact in cancer remains an area of future study. Additional data is required to support the predicted structural genomic findings in this study and therefore a natural extension to our investigation would be to observe the behavior of these genes in a cell line model. Numerous PDAC cell lines exist, however, not surprisingly, the mutational status of ATPAF1-AS1 and MAP3K6 has not been documented. Ideally, a cell line with wild-type copies of these genes and another cell line with mutated copies of these genes would be ideal and this could be determined by target capture NGS. RNA interference study could be performed where any RNA copies of these genes are antagonized. Alternatively, an expression vector assay with the wild-type copy of the gene could be used to ‘add’ copies to cell lines. Once either knock down or restored or overexpression is completed, cell line behavior could be observed by virtue of a: colony formation assay detailing the ability to form colonies; modified scratch assay examining the migration properties of the cell line; matrigel invasion assay which profiles the ability to invade across a synthetic basement membrane. As a subgroup analysis, the above experiments could be performed in cell lines with known PDAC causative mutations such CAPAN-1, which is a BRCA-2 deficient cell line, testing for a more aggressive phenotype when ATPAF1-AS1 or MAP3K6 are knocked down.

A central challenge in this study, which has been re-iterated many times, is the extensive locus heterogeneity present in our somatic dataset. The strategy used to sort through this heterogeneity for FPC gene discovery, will have a direct influence in the design of future investigations. First,
locus heterogeneity, as detailed in chapter 5 section 5.2, would point towards the clonal nature of FPC/PDAC. Genetic changes may occur with a low to medium frequency and only in a small proportion of disease serving as “private” mutations. Our findings of the potential involvement of DCLK3 and SERPINF1 loci in family 385P support this notion. Although, FISH studies have confirmed the somatic loss at these regions, we have yet to confirm a germline mutation. Future studies would require target capture NGS or Sanger sequencing of normal FFPE DNA of both genes in entirety, examining for a rare germline variant. As well, the functional consequences of knockdown of these loci could also be determined by the cell line studies detailed above.

Second, future experiments could build on the integrative strategy used in this study by combining somatic LOH and germline ES data with whole transcriptome and epigenetic experiments in FPC. A limitation of our study is that we only examined the coding regions of genes and therefore it is possible that changes in intronic or promoter regions, including methylation and histone events, may contribute to either the first or second hit in disease pathogenesis. Such studies have not been performed in FPC, once again due to challenges with sample collection. Novel methylation microarrays and RNA seq platforms have been described for FFPE RNA and these platforms may bypass the above challenges as we demonstrated using genomic DNA in this study. Furthermore, current online genomic databases have improved with respect to annotation of non-coding changes. Thus incorporating this data may also provide further insight into the genes highlighted in this study and bring attention to other genes that may have been ‘discarded’ in our experiments as they only partly satisfied our criteria for inclusion (i.e. only had a germline event but no somatic change or vice-versa).

Finally, our studied relied on the validity of the Knudson 2-hit hypothesis model for TSGs and did not consider a haploinsufficiency model. Haploinsufficiency of key genes have been described in other cancers such as prostate cancer (Kwabi-Addo, Giri et al. 2001, Matsuyama, Oba et al. 2007) and MLH1 haploinsufficiency has been shown to increase genomic instability in sporadic PDAC (Wang, Tsutsumi et al. 2012). Therefore, future studies incorporating our exome and CNV data testing for haploinsufficiency may characterize important genomic changes in FPC.
Bibliography


MELCHER, R., HARTMANN, E., ZOPF, W., HERTERICH, S., WILKE, P., MULLER, L., ROSLER, E., KUDLICH, T., AL-TAIE, O., ROSENWALD, A., KATZENBERGER, T., SCHOLTKA, B., SEIBOLD, S., ROGOLL, D., SCHEPPACH, W., SCHEURLLEN, M. and LUHRS, H., 2011. LOH and copy neutral LOH (cnLOH) act as alternative mechanism in


GLANOWSKI, S., GLASSER, K., GLODEK, A., GOROKHOV, M., GRAHAM, K.,
GROPMAN, B., HARRIS, M., HEIL, J., HENDERSON, S., HOOVER, J., JENNINGS, D.,
JORDAN, C., JORDAN, J., KASHA, J., KAGAN, L., KRAFT, C., LEVITSKY, A., LEWIS,
M., LIU, X., LOPEZ, J., MA, D., MAJOROS, W., MCDANIEL, J., MURPHY, S., NEWMAN,
M., NGUYEN, T., NGUYEN, N., NODELL, M., PAN, S., PECK, J., PETERSON, M., ROWE,
W., SANDERS, R., SCOTT, J., SIMPSON, M., SMITH, T., SPRAGUE, A., STOCKWELL, T.,
TURNER, R., VENTER, E., WANG, M., WEN, M., WU, D., WU, M., XIA, A., ZANDIEH, A.
291(5507), pp. 1304-1351.


WANG, L., TSUTSUMI, S., KAWAGUCHI, T., NAGASAKI, K., TATSUNO, K.,
YAMAMOTO, S., SANG, F., SONODA, K., SUGAWARA, M., SAIURA, A., HIRONO, S.,
YAMAUE, H., MIKI, Y., ISOMURA, M., TOTOKI, Y., NAGAE, G., ISAGAWA, T., UEDA,
H., MURAYAMA-HOSOKAWA, S., SHIBATA, T., SAKAMOTO, H., KANAI, Y., KANEDA,
cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and


WELLCOME TRUST CASE CONTROL CONSORTIUM, CRADDOCK, N., HURLES, M.E.,
CARDIN, N., PEARSON, R.D., PLAGNOL, V., ROBSON, S., VUKCEVIC, D., BARNES, C.,
CONRAD, D.F., GIANNOULATOU, E., HOLMES, C., MARCHINI, J.L., STIRRUPS, K.,
TOBIN, M.D., WAIN, L.V., YAU, C., AERTS, J., AHMAD, T., ANDREWS, T.D., ARBURY,
H., ATTWOOD, A., AUTON, A., BALL, S.G., BALMFORTH, A.J., BARRETT, J.C.,
BARROSO, I., BARTON, A., BENNETT, A.J., BHASKAR, S., BLASZCZYK, K., BOWES, J.,
BRAND, O.J., BRAUND, P.S., BREDIN, F., BREEN, G., BROWN, M.J., BRUCE, I.N., BULL,
J., BURREN, O.S., BURTON, J., BYRNES, J., CAESAR, S., CLEE, C.M., COFFEY, A.J.,
CONNELL, J.M., COOPER, J.D., DOMINICZAK, A.F., DOWNES, K., DRUMMOND, H.E.,
DUDAKIA, D., DUNHAM, A., EBBS, B., ECCLES, D., EDKINS, S., EDWARDS, C.,
ELLiot, A., EMERY, P., EVANS, D.M., EVANS, G., EYRE, S., FARMER, A., FERRIER,
J.N., FEUK, L., FITZGERALD, T., FLYNN, E., FORBES, A., FORTY, L., FRANKLYN, J.A.,
FREATHY, R.M., GIBBS, P., GILBERT, P., GOKUMEN, O., GORDON-SMITH, K., GRAY,
E., GREEN, E., GROVES, C.J., GROZIVA, D., GWILLIAM, R., HALL, A., HAMMOND, N.,
HARDY, M., HARRISON, P., HASSANALI, N., HEBASI, H., HINES, S., HINKS, A.,
HITMAN, G.A., HOCKING, L., HOWARD, E., HOWARD, P., HOWSON, J.M., HUGHES,
D., HUNT, S., ISAACS, J.D., JAIN, M., JEWELL, D.P., JOHNSON, T., JOLLEY, J.D.,
JONES, I.R., JONES, L.A., KIROV, G., LANGFORD, C.F., LANGO-ALLEN, H., LATHROP,
G.M., LEE, J., LEE, K.L., LEES, C., LEWIS, K., LINDGREN, C.M., MAIRURAMA-ARMER, M.,
MALLER, J., MANSFIELD, J., MARTIN, P., MASSEY, D.C., MCARDLE, W.L.,
MCGUFIN, P., MCLAY, K.E., MENTZER, A., MIMMACK, M.L., MORGAN, A.E.,
MORRIS, A.P., MOWAT, C., MYERS, S., NEWMAN, W., NIMMO, E.R., O'DONOVAN,
M.C., ONIPINLA, A., ONYIHA, I., O'VINGTON, N.R., OWEN, M.J., PALIN, K., PARNELL,


Appendices

Table S1: Summary of BAC clones used for FISH assays

Human Genome Feb 2009 Assembly (GRCh37/hg19)

<table>
<thead>
<tr>
<th>#</th>
<th>Gene/Clone ID</th>
<th>Chromosome Region</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MAP3K6</td>
<td>1p35.11</td>
<td>27681670</td>
<td>27693337</td>
</tr>
<tr>
<td></td>
<td>RPM401H24</td>
<td></td>
<td>27222109</td>
<td>28379312</td>
</tr>
<tr>
<td></td>
<td>RPM350G18</td>
<td></td>
<td>27612094</td>
<td>27803079</td>
</tr>
<tr>
<td>2</td>
<td>DCLK3</td>
<td>3p22.2</td>
<td>36753913</td>
<td>36781352</td>
</tr>
<tr>
<td></td>
<td>RPM344P18</td>
<td></td>
<td>36579353</td>
<td>36753133</td>
</tr>
<tr>
<td></td>
<td>RPM1776G6</td>
<td></td>
<td>36679158</td>
<td>38667436</td>
</tr>
<tr>
<td>3</td>
<td>SERPINF1</td>
<td>17p13.3</td>
<td>1665259</td>
<td>1680859</td>
</tr>
<tr>
<td></td>
<td>RPM114F24</td>
<td></td>
<td>1618008</td>
<td>2015783</td>
</tr>
<tr>
<td></td>
<td>RPM357G7</td>
<td></td>
<td>1934509</td>
<td>2130401</td>
</tr>
<tr>
<td>4</td>
<td>C30RF72</td>
<td>3q22.3</td>
<td>138666076</td>
<td>138672830</td>
</tr>
<tr>
<td></td>
<td>RPM1768G22</td>
<td></td>
<td>138514728</td>
<td>138703534</td>
</tr>
<tr>
<td></td>
<td>RPM15480D</td>
<td></td>
<td>138524280</td>
<td>138705376</td>
</tr>
<tr>
<td>5</td>
<td>ATPAF1-A51</td>
<td>1p33</td>
<td>47137497</td>
<td>47139256</td>
</tr>
<tr>
<td></td>
<td>RPM140P4</td>
<td></td>
<td>48992866</td>
<td>47149927</td>
</tr>
<tr>
<td></td>
<td>RPM115C8</td>
<td></td>
<td>47021730</td>
<td>47169634</td>
</tr>
<tr>
<td>6</td>
<td>GSDMA</td>
<td>17q21.1</td>
<td>38121911</td>
<td>38134019</td>
</tr>
<tr>
<td></td>
<td>RPM11387H17</td>
<td></td>
<td>37953483</td>
<td>38181328</td>
</tr>
<tr>
<td></td>
<td>RPM11881A13</td>
<td></td>
<td>38034549</td>
<td>38205024</td>
</tr>
</tbody>
</table>

The above table summarizes the BAC clones for FISH analysis covering the candidate FPC TSGs of interest. Rows highlighted in yellow indicate BACs used whereas teal rows indicate BAC clones assayed however not used for FISH due to improper chromosomal localization. All locations are based on BAC physical mapping as per the human genome Feb 2009 assembly.

Table S2: Summary of FPC Copy Gain and Loss Regions Called on ASCAT
(Table available as excel sheet on attached CD)

Table S3: Summary of FPC Allelic Loss Called on ASCAT
(Table available as excel sheet on attached CD)

Table S4: Summary of Sporadic PDAC Copy Gain and Loss Regions Called on ASCAT
(Table available as excel sheet on attached CD)
Table S5: Summary of Sporadic PDAC Allelic Loss Called on ASCAT
(Table available as excel sheet on attached CD)

Table S6: Summary of FPC Copy Gain and Loss Regions Called on SNP-FASST2 Segmentation
(Table available as excel sheet on attached CD)

Table S7: Summary of Sporadic Copy Gain and Loss Regions Called on SNP-FASST2 Segmentation
(Table available as excel sheet on attached CD)

Table S8: Summary of CN Loss Genes identified by Fisher Test on SNP-FASST2 Segmentation
(Table available as excel sheet on attached CD)

Table S9: Summary of CN Loss Genes identified by Fisher Test on ASCAT
(Table available as excel sheet on attached CD)

Table S10: CNV Regions for Cases with Germline Exome Sequencing and Oncoscan Data called on SNP-FASST2 Segmentation
(Table available as excel sheet on attached CD)

Figure S1: BAC probe localization on peripheral blood lymphocyte chromosomes

a.) Localization of ATPAF1-AS1 probe
b.) Localization of GSDMA Probe

c.) Localization of C3ORF72 Probe
The above three figures illustrate localization of the respective BAC clones for the region of interest along with their respective control probes. Localization of the control and experimental probes occur on the same chromosome for each locus of interest.

Table S11: Genes identified on GISTIC on the ASCAT Algorithm
(Table available as excel sheet on attached CD)

Table S12: Genes identified on GISTIC on the SNP-FASST2 Segmentation Algorithm
(Table available as excel sheet on attached CD)