A Gene Association Study to Identify Novel Pancreatic Cancer Susceptibility Genes

Cavin Wong

Human Genetics, McGill University, Montreal

December 2017

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

Abstract

Approximately 10% of pancreatic cancer (PAC) cases are hereditary in nature, however, only a fraction is explained by known susceptibility genes. Recent efforts using Next-Generation Sequencing (NGS) data to elucidate novel predisposition genes for PAC risk have been plagued with challenges due to the limitations of the methods in identifying a "faint signal" from a large amount of "noise" created by non-causal variants. Thus, in this dissertation, I have hypothesized that a region-based case-control gene association test, the Mixed-effects Score Test (MiST), will allow for the identification of these "faint signals" in NGS data sets. Compared to previous methods, MiST is a less biased statistical approach which compares mutation frequency across a gene or region for cases versus controls, while accounting for individual variant characteristics. For PAC cases, DNA extracted from circulating lymphocytes or saliva was used as a surrogate for germline DNA. Our case series consists of 109 exomes from high risk PAC cases (familial pancreatic cancer or young onset <50) and 289 prospectively collected PAC cases sequenced for 710 cancer-related genes. Our control series consists of 987 non-cancer cases collected locally from multiple projects. All samples were processed on the same pipeline and variants were limited to exonic and splicing variants. Prior to analysis, we used a principal component analysis to exclude genetic outliers. To increase the power of our study, we limited our association test to 449 DNA repair genes, for which 6842 rare variants (MAF<0.5%) were identified across 418 of these genes. MiST was performed for 235 genes which had >10 rare variants identified across all cases and controls; of these, 42 genes were significant prior to multiple testing correction ($p<0.05$), including the known susceptibility genes, *BRCA1, BRCA2* and *STK11.* After correction ($p<0.00021$), only two genes remained significant, *RECQL* and *SMG1*. The drop-one method was performed to determine candidate variants driving the association with PAC for the 3 known

susceptibility genes and the 2 candidate genes. Using drop-one (increase in p-value >35%), we were able to identify the known pathogenic mutations in *BRCA1* and *BRCA2*, as well as a list of candidate variants in all 5 genes. The association for *STK11* and *RECQL* were driven by variants in controls, thus, these were discarded as a protective effect is difficult to evaluate with our sample size. On the other hand, *SMG1*, our top candidate ($p=3.22x10^{-7}$), was driven by 15 variants identified across 29 cases and 1 control. To further support *SMG1* as a candidate susceptibility gene, segregation analyses were performed for two families, where samples were available. In one family, we observed segregation of the variant with 3 individuals with PAC. In summary, this dissertation demonstrates the feasibility and utility of using a region-based gene association study to identify novel susceptibility genes, as well as prioritizing variants for future functional analyses. Using these methods, we have identified a novel candidate PAC susceptibility gene, *SMG1*, which will need to be further validated in other cohorts and functional analyses.

Résumé

Environ 10% des cancers du pancréas (CP) sont héréditaires. Cependant, peu d'entre eux sont expliqués par des gènes connus de susceptibilité au cancer. Les efforts récents pour identifier de nouveaux gènes de susceptibilité pour le CP ont été généralement infructueux, probablement à cause des limitations des méthodes utilisées quant à l'identification d'un « signal faible » dans une grande quantité de « bruit » causé par des variants non causaux. Ainsi, dans cette dissertation, j'ai teste l'hypothèse qu'un test d'association cas-contrôles par région, le *Mixed-effects Score Test* (MiST), va permettre d'identifier ces « signaux faibles ». En comparaison avec d'autres méthodes, MiST est une approche statistique peu biaisée qui compare la fréquence de mutation d'un gène ou d'une région entre les cas et les contrôles, tout en tenant compte des caractéristiques individuelles des variants. Dans les cas de CP, l'ADN extrait de lymphocytes en circulation ou de salive a été utilisé pour représenter l'ADN germinal.Notre série de cas consiste en 109 cas de CP à haut risque (cancer du pancréas familial ou âge au diagnostic <50) avec séquençage complet d'exome et 289 cas de CP collectés prospectivement de deux centres hospitaliers canadiens (Montréal et Toronto) avec séquençage ciblé pour 710 gènes reliés au cancer. Notre série contrôle consiste en 987 cas indemnes de cancer collectés de multiples projets locaux. Tous les échantillons ont été traités sur la même pipeline et les variants étaient limités aux variantsexoniques et variants d'épissage. Avant de faire ce test, nous avons utilisé une analyse en composantes principales pour exclure les valeurs aberrantes. Pour augmenter la puissance de notre étude, nous avons limité notre test d'association à 449 gènes associés à la réparation de l'ADN, pour lesquels 682 variants rares (fréquence d'allèle mineur <0.5%) ont été identifiés dans 418 de ces gènes. MiST a été exécuté pour 235 gènes qui avaient >10 variants rares identifiés dans tous les cas et contrôles; de ceux-ci, 42 gènes étaient significatifs avant

correction pour tests multiples (p<0,05), incluant les gènes connus de prédisposition au cancer *BRCA1*, *BRCA2* et *STK11*. Après correction (p<0,00021), seulement deux gènes sont restés significatifs : *RECQL* et *SMG1*. La méthode « *drop-one* » a été utilisée pour déterminer les variants candidats responsables de l'association avec le CP pour les 3 gènes connus de prédisposition héréditaire et les 2 gènes candidats. Avec cette méthode, (augmentation de valeur-p>35%), nous avons pu identifier les mutations pathogènes connues dans les gènes *BRCA1* et *BRCA2*, de même qu'une liste de variants candidats dans les 5 gènes. Comme l'association pour les gènes *STK11* et *RECQL* était causée par des variants chez les contrôles, celles-ci ont été écartées comme un effet protecteur difficile à évaluer avec notre taille d'échantillon. En revanche, l'association pour *SMG1*, notre meilleur candidat (p=3.22E-7), était causée par 15 variants identifiés chez 29 cas et 1 contrôle. Pour supporter davantage *SMG1* comme gène candidat de susceptibilité au cancer, des analyses de ségrégation ont été réalisées pour deux familles pour lesquelles des échantillons étaient disponibles. Chez l'une de ces familles, nous avons observé une ségrégation du variant chez trois individus avec CP. En résumé, cette dissertation démontre la faisabilité et l'utilité d'une approche d'association génétique par région pour identifier de nouveaux gènes de susceptibilité, de même que pour prioriser les variants à investiguer pour de futures analyses fonctionnelles. En utilisant ces méthodes, nous avons identifié un nouveau gène candidat de susceptibilité au CP, *SMG1*, qui devra être validé davantage dans d'autres cohortes et par des analyses fonctionnelles.

Table of Contents

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. George Zogopoulos, for his patience, guidance, and wisdom throughout my studies. As a mentor, he has taught me important skills in research and clinical medicine, and as a friend, he has taught me many valuable life lessons. Throughout these two years, I have grown and learned more than I could have ever imagined before my graduate studies.

I would also like to extend my thanks to my thesis committee members, Dr. Celia Greenwood and Dr. Jacek Majewski, for their guidance and support throughout my studies.

In addition, a sincere thank you to my fellow graduate students and friends in the lab, Patrick Park, Alyssa Smith and Yifan Wang. Whether it was sharing laughs during those late nights at the lab or enjoying brunch the next morning, without them, my graduate experience would not have been as enjoyable. I am forever grateful for their love and support these past two years.

Furthermore, I would like to extend my gratitude to the members of our lab, Anita Hall, Adeline Cuggia, Celine Domecq, and Amanda Tanti, as well as the previous members of the lab, and our collaborators, for their constant support and guidance throughout my studies.

Finally, I would like to dedicate this thesis to both my grandmother and mother for always supporting me in pursuing my dreams and encouraging me, no matter how long and difficult the path.

Dr. George Zogopoulos provided mentorship and guidance in experimental design, interpretation of data, and the preparation of this dissertation.

Ms. Adeline Cuggia assisted with the collection of patient clinical information and the French-to-English translation of the abstract.

Dr. Jacek Majewski and Najmeh Alirezaie provided their bioinformatics expertise and data for the control cohort.

Dr. Celia Greenwood provided guidance in experimental design, statistical guidance, and interpretation of data.

Our collaborators at Toronto (**Dr. Steven Gallinger and lab)** for sample collection, experimental design, and providing patient clinical information.

I performed all experiments presented in this dissertation, including the bioinformatics and statistical analyses, unless mentioned above. I prepared the dissertation.

List of Tables and Figures

Chapter III

List of Abbreviations

Abbreviation	Definition
1000s	1000 genomes project
A-T	Ataxia Telangiectasia
ABO	Blood type locus
AJ	Ashkenazi Jewish
AP	Apurinic/apyrimidinic
BER	Base Excision Repair
CADD	Combined Annotation Dependent Depletion
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats
D1V	Drop-one variant analysis
D1W	Drop-one window analysis
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic acid
DSB	Double-strand DNA break
EUS	Endoscopic Ultrasound
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
FAMMM	Familial Atypical Multiple Mole Melanoma
FAP	Familial Adenomatous Polyposis
FC	French Canadian
FDR	First degree relative
FPC	Familial pancreatic cancer
G1	First Gap phase
G2/S	Second Gap phase and Synthesis phase
GATK	Genome Analysis Toolkit
GERP	Genomic Evolutionary Rate Profiling
GO	Gene Ontology
GWAS	Genome-wide Association Study
HBOC	Hereditary Breast and Ovarian Cancer syndrome
HBC	Hereditary Breast Cancer syndrome
HDR	Homology-directed DNA Repair
Hg19	Human Genome build 19
HP	Hereditary Pancreatitis
HSF	Human Splicing Finder
INDELs	Insertions/deletions
IPMN	Intraductal Papillary Mucinous Neoplasm
JPS	Juvenile Polyposis Syndrome

LFS	Li-Fraumeni Syndrome
LOH	Loss-of-Heterozygosity
LS	Lynch Syndrome
MAF	Minor Allele Frequency
MaxEntScan	Maximum Entropy Scan
MiST	Mixed-effects Score Test
MMR	Mismatch Repair
MRN	Mre11-Rad50-Nbs1 complex
Nab-Paclitaxel	Albumin-bound Paclitaxel
NCCN	National Comprehensive Cancer Network
NER	Nucleotide Excision Repair
NGS	Next-generation sequencing
NHEJ	Non-homologous End Joining
NMD	Nonsense-mediated mRNA decay
OPCS	Ontario Pancreas Cancer Study
PAC	Pancreatic ductal adenocarcinoma
PALLD	Palladin
PanIN	Pancreatic intraepithelial neoplasia
PARP	Poly(ADP-ribose) polymerase
PC1/2	Principal Component 1 and Principal Component 2
PCA	Principal Component Analysis
PJS	Peutz-Jeghers Syndrome
PTV	Protein truncating variant
QPCS	Quebec Pancreas Cancer Study
RefSeq	Reference Sequence database
ROC	Receiver-operator characteristic
SD	Standard deviation
shRNA	Small hairpin RNA
SIFT	Sorting Intolerant From Tolerant
SKAT	SNP-set Sequence Kernel Association Test
SKAT-O	SNP-set Sequence Kernel Association Test Optimized
SNV	Single nucleotide variant
SSB	Single-strand DNA break
TLS	Translesion Synthesis
UV	Ultraviolet
VUS	Variants of Uncertain Significance
WES	Whole-exome sequencing

Chapter I: Literature Review

1.1 Pancreatic Cancer

1.1.1 The Pancreas at a Glance

The pancreas is a glandular organ of the digestive tract which can be divided into 5 different regions: the head, uncinate process, neck, body and tail[1,2]. It is responsible for both exocrine functions and endocrine functions involved in digestion and glucose regulation[1,2]. The pancreas is located within the retroperitoneum of the abdomen, and is situated between the duodenum and spleen[1,2].

The bulk of the gland is made up of two types of exocrine cells[2]. The acinar cells are responsible for the production of digestive enzymes that are secreted into the mucinous columnar epithelial cells that make up the pancreatic duct, eventually emptying into the duodenum[2]. In addition, there are groups of endocrine cells, known as the Islet of Langerhans, found within the gland that are responsible for the production of different hormones involved in glucose regulation[2]. There are five distinct cell types that compose the pancreatic islets; the bulk is composed of the beta cells which produces insulin and amylin[2]. The other four cell types are the alpha cells (secretes insulin), delta cells (secretes somatostatin), gamma cells (secretes pancreatic polypeptide), and the epsilon cells (secretes ghrelin)[2].

Neoplasms can arise from any of the cell types found in the pancreas, with exocrine cell cancers classified as adenocarcinomas and endocrine cell cancers classified as neuroendocrine tumours[3,4].Pancreatic adenocarcinomas (PAC)are further classified histologically as pancreatic ductal adenocarcinoma (PDAC), acinar cell carcinoma, mucinous cystadenocarcinoma, and

invasive adenocarcinoma arising from interpapillary mucinous neoplasia (IPMN)[3,4]. PDAC is the most common subtype, accounting for approximately 90% of all pancreatic adenocarcinomas[5].

1.1.2 Precursor Lesions of Pancreatic Ductal Adenocarcinomas

The PanIN classification is a histopathological system developed for the classification of premalignant lesions that may develop into invasive ductal adenocarcinomas of the pancreas based on cell morphology; ranging from normal, pancreatic intraepithelial neoplasia IA (PanIN-IA), PanIN-IB, PanIN-2, and PanIN-3[4,6,7]. The differences found between each stage are noted by cellular shape, and nuclear size, shape and abnormalities[6,7]. In addition to these differences, studies have shown a difference in genetic, epigenetic and protein expression; in fact, these lesions are thought to be driven by gain-of-function mutations in *KRAS*[4,6,7]. Higher grade lesions (i.e PanIN-3) are more likely found in individuals with pancreatic cancer or chronic pancreatitis[4,6,7]. These studies suggest that the genetic evolution of lesions are gradual but linear, starting with gain-of-function mutations in *KRAS,* followed by loss-of-mutations in*CDKN2A, TP53,* and then *SMAD4*[6,7].Clinically, PanINs are thought to help aid in early detection before the development of pancreatic cancer. However, there are difficulties in detecting PanIN lesions in the pancreas as cross-sectional imaging and endoscopic ultrasound (EUS) remain ineffective, as these are microscopic lesions that may or may not become malignant[4,6-8].

The PanIN model has been challenged as it does not explain the divergence observed in the genetic evolution of pancreatic clonal lesions[9]. In fact, a recent paper by Notta*et al*, suggests that preneoplastic lesions may not accumulate mutations linearly, but it may be more rapid and simultaneous due to large genomic events, such as polyploidization and

chromothripsis[9].Polyploidization leads to increase in copy number of genes and may lead to a more rapid accumulation of mutations[9]. Similarly, chromothripsis results in the shattering of the chromosome into multiple fragments followed by random end-joining, resulting in rapid accumulation of rearrangements in the genome[9]. Thus, the transformation from preneoplastic to invasive disease may be very sudden, explaining the aggressiveness seen in pancreatic tumours[9]. In addition, they show that most ofthese genomic instability events are present early in tumorigenesis prior to metastases[9].

1.1.3Pancreatic Ductal Adenocarcinoma

PAC is one of the most lethal solid malignanciesmainly attributed to late detection and the innate chemoresistance of the tumour[8,10]. In fact, PAC patients have an estimated 5-year survival rate of only 8%, and is predicted to become the 3rd leading cause of cancer deaths in North America by 2030[8,10,11]. The current standard of care for early stage cases is resection of the tumour, which accounts for about 20% of all PACs, while the remaining 80% of PAC cases are unresectable[10]. For late stage cases, neoadjuvant/palliative chemotherapy, usually Gemcitabine, FOLFIRINOX, or Gemcitabine with albumin-bound paclitaxel (nab-paclitaxel), is given to control the disease and improve quality of life[10,12]. FOLFIRINOX is a cocktail drug containing fluorouracil (5-FU), irinotecan and oxaliplatin; a pyrimidine analog, an inhibitor of topoisomerase I, and a platinum, respectively[10,12]. Each of these drugs act by increasing genomic instability leading to apoptosis of the tumour cells[10,12]. Gemcitabine, a nucleoside analog, works similarly as FOLFIRINOX, however, it is less aggressive and less toxic[10,12]. Chemotherapy is also given in the adjuvant setting following curative-intent resection as it has been shown to increase survival. Following the CONKO trial, adjuvant chemotherapy with Gemcitabine was

established as standard of care[13]. More recently, the ESPAC-4 trial results have shown combination of gemcitabine and capecitabineshould be the new standard of care following curative intent surgery for PAC[14].

Advances in the field have been promising with the advent of next-generation sequencing, including the development of several subtype classification systems[15-18]. The first subtype classification for pancreatic cancer was developed by Waddell *et al*, describing four different subtypes based on frequency of structural variation: stable, locally rearranged, scattered, and unstable[15]. The unstable subtype is described as a tumour genome containing more than 200 structural variations, likely suggesting a defect in DNA repair[15]. In fact, they showed clustering of *BRCA1* and *BRCA2* germline mutation carriers, two genes known to play a role in homologous-directed DNA repair (HDR), in the unstable subtype[15,19]. These HDR-deficient tumours are clinically significant as studies have shown increased sensitivities to platinum-based drugs, such as oxaliplatin and cisplatin[10,20]. This is likely due to an accumulation of DNA damage created by the cross-linking of DNA by cisplatin in combination with the inefficient DNA repair in HDR-deficient tumours, leading to increased apoptosis of tumour cells[10,20].

1.2 DNA Damage Repair Pathways

The genome is constantly being replicated as cells divide in our body, giving rise to the possibility of errors in replication[19]. This may be dangerous for cell survival as mutations in key genes may result in loss of proper cell proliferation control[19]. To prevent this, several DNA repair mechanisms exist to correct for errors during replication, as well as mutational changes induced by the environment[19]. Following detection of DNA damage, the cell will undergo a

series of signal transduction resulting in cell cycle arrest to allow for the DNA repair process to occur[19].In the following section, a brief overview will be given for the three major single-stranded break (SSB) DNA repair pathways and two double-stranded break (DSB) DNA repair pathways.

1.2.1 Single-Stranded Break Repair Pathways:

1.2.1.1 Mismatch Repair (MMR)

The simplest mutational event is the insertion of an incorrect base during replication, which leads to single nucleotide changes in newly synthesized DNA[19,21]. Usually these mistakes are corrected by the innate repair function of DNA polymerase δ and ε, the main polymerases responsible for elongation during replication[19,21]. However, errors may be missed despite the initial proofreading, thus an additional proofreading checkpoint following DNA strand synthesis is required to ensure fidelity[19,21]. The MSH6-MSH2 protein complex scans the newly synthesized DNA for errors; upon detection, single-stranded breaks are induced surrounding the mismatched base by the MLH1-PMS2 complex in the presence of the DNA clamp protein, PCNA[19,21-23]. The exonuclease EXO1 will degrade the nicked DNA strand removing the segment with the mismatched base[19,21,24].DNA polymerase is recruited to re-synthesize the missing segment of DNA, completing the MMR process[19,21].

1.2.1.2 Base-Excision Repair (BER)

The BER pathway is responsible for the repair of small and simple base lesions which do not affect the structure of the DNA helix, including, oxidized bases, alkylated bases, anddeaminated bases, most commonly uracil sites[19,25]. BER is initiated by the removal of the

improper base by a corresponding DNA glycosylase, such as Uracil-DNA glycosylase (UNG)for uracil sites, resulting in the generation of an apurinic/apyrimidic (AP) site[19,25].The AP site is processed by APEX1, a DNA-AP-site lyase, followed by removal of the remaining residue and insertion of the proper nucleotide by DNA polymerase β[19,25,26]. Finally, the DNA ligase III-XRCC1 complex will seal the nick in the DNA[26].

1.2.1.3Nucleotide-Excision Repair (NER)

NER is responsible for the detection and repair of bulky base lesions that affect the DNA helix structure, such as pyrimidine dimers created by UV exposure[19,27,28]. The protein XPA is responsible for recognition of DNA lesions and initiation of protein complex formations required for NER[19,27,28]. Two helicases, XPB and XPD, are responsible for unwinding DNA, followed by excision of the segment of DNA surrounding the improper base by the XPF/ERCC1 complex and XPG[19,27,28]. Similar to MMR, DNA polymerase δ or ε will be recruited to fill in the missing segment of DNA[19,27,28].

1.2.2 Double-Stranded Break Repair Pathways:

1.2.2.1 Homology-directed DNA Repair (HDR)

The previous 3 pathways were responsible for lesions resulting in single-stranded breaks in DNA. NHEJ and HDR are the two major pathways responsible for repair of double-stranded breaks (DSB) in DNA[19,29]. HDR is the error-free pathway for DSB repair, as it utilizes the sister chromatid as a template for repair of the damaged chromosome; thus HDR is normally only active in G2/S due to the necessity of the sister chromatid following DNA replication[19,29].

ATM is recruited to the site of DSBs following detection of damageby the MRN complex composed of three proteins: Mre11, Rad50, NBS1[19,29]. Upon recruitment, ATM phosphorylates γH2AX leading to recruitment of multiple HDR proteins, including *BRCA1*[19,29].In parallel, the MRN complex cooperates with CtIP, a transcription factor, to initiate end resection of the break site, followed by end processing by the helicase BLM and two nucleases, EXO1 and DNA2[19,29]. The presence of processed ends at the DSB site leads to recruitment of RPA, which likely stabilizes the single-stranded DNA ends[19,29]. The presence of RPA at the break site leads to recruitment of PALB2 followed by BRCA2, two proteins that work in tandem to displace RPA and load RAD51, the recombinase responsible for strand invasion of the sister chromatid[19,29]. Recombination intermediates are formed and resolved indicating the completion of the error-free repair of the damaged chromosome[19,29].

1.2.2.2 Non-Homologous End Joining (NHEJ)

NHEJ is the non-conservative form of DNA repair,typically occurring outside of G2/S where sister chromatids are absent[19,30]. NHEJ consists of multiple sub-branches depending on the complexity of the DSB, and the compatibility of the two ends[19,30]. Briefly, this pathway is characterized by the presence of the Ku-70/Ku-80 heterodimer responsible for detection of DSB breaks[19,30]. Following detection, many major NHEJ players are recruited to the break site, including the XRCC4-LIG4 complex, and the DNA-PKcs complex composed of two subunits encoded by *XRCC7* and *PRKDC*[19,30]. In the presence of these major players, mismatched bases and single-stranded overhangs are processed at both ends to ensure blunt ends required for NHEJ[19,30]. This is followed by resolution of the NHEJ complex resulting in ligation of the two

ends[19,30]. Although the pathway is non-conservative, the repair of the DSB maintains genomic stability and allows for cell survival[19,30].

1.2.2.3 ATM – the Key Regulator for DSB Repair Pathway Decision

ATM is a large gene with 63 exons spanning approximately 146,000 base-pairs; the protein is composed of 3065 amino acids[31].ATMis a serine-threonine kinase involved in cell cycle checkpoint and DNA repair; it is located in the nucleus, and is recruited and activated by DNA DSBs[31-33]. ATMis an upstream molecule in these pathways and has been shown to interact with many different proteins, including NBS1, CHK2, BRCA1, 53BP1, Ct-IP, and H2AX[31,32,34].

Studies have shown that ATM is an important molecule in DNA DSB repair pathway decision between HDR and NHEJ[34,35]. In G1, 53BP1 is recruited to the site of DSB and is activated by ATM via phosphorylation[34,35]. Phosphorylated 53BP1 then recruits Rif1, leading to prevention of DNA DSB end resection and recruitment of BRCA1, both required for HDR[34,35]. As a result, the cell will activate the NHEJ pathway to repair the DNA DSBs[34,35]. Conversely, in G2 and S phase, Ct-IP is present and interacts with BRCA1 to prevent the recruitment of Rif1, thus preventing the inhibition on DSB end resection[34,35]. In fact, HDR activity is rescued in *BRCA1*-deficient cells when Rif1 is knocked out[34,35].

NHEJ has been shown to be capable of occurring independently of *ATM* expression, although loss of *ATM* does result in some impairment[36]. On the other hand, studies on the effect of *ATM* loss on HDR is slightly conflicting as some studies using expression of a HDR reporter gene showed intact HDR function in the absence of *ATM*[34]. Other studies suggest that RAD51

foci formation, an important initiation step of HDR, does occur in the absence of *ATM*, however, it is slower than normal, and the foci persist longer suggesting that initiation of HDR occurs but completion of repair is impaired[33,34].

1.3 Genetic Testing

The ability to identify individuals that have an increased risk of developing PAC, compared with the general population, may help develop effective early detection strategies. Approximately 10% of all pancreatic cases are thought to be hereditary, which is likely driven by a germline mutation in a susceptibility gene[37]. A fraction of these mutations are found in known genes, with a majority inthe hereditary breast and ovarian cancer syndrome genes:*BRCA1, BRCA2, PALB2* and *ATM*[6,38,39]. In addition, other diseases may increase the risk of developing pancreatic cancer, includinghereditary pancreatitis (*PRSS1*), Li-Fraumeni (*TP53),* Peutz-Jehgers (*STK11),* juvenile polyposis syndrome (*BMPR1A* and*SMAD4)*, familial adenomatous polyposis (APC), familial atypical multiple mole melanoma (*CDKN2A*), and Lynch syndrome (*MLH1*, *MSH2*, *MSH6*, and *PMS2*)[37]. While mutations in these genes are thought to behave in a Mendelian fashion, the genetics in cancer are likely muddledwith the possibility of mutations contributing to increased cancer risk as a complex trait. Furthermore, not all carriers of mutations in these genes will develop PAC, which is likely due to the variable penetrance of the mutation, while non-carriers in the same kindred may develop PAC that is a phenocopy. While the understanding of penetrance plays an important factor in developing screening guidelines, it is difficult to ascertain the true penetrance of mutations in these predisposition genes, especially as penetrance may differ for different cancer types. (e.g., breast cancer *versus* PAC). In addition to the hypothesis that there are additional unknown PAC predisposition genes, these latter caveats

may be underlying why hereditary PAC cannot always be attributed to these known predisposition genes[40,41].

Current screening practices do not include routine testing of all PAC cases, given the large resources needed[42]. Instead, an individual's personal medical history, as well as a family history of PAC and other related cancers are taken into consideration[42]. Unfortunately, due to differences in resources and mixed results in literature, there is no clear consensus on genetic testing referrals across centres. However, the National Comprehensive Cancer Network (NCCN) provides guidelines for genetic screening of PAC, with annual updates if needed[42]. NCCN guidelines suggest screening for individuals with a family history suggestive of any of the hereditary cancer syndromes listed above[42]. In addition, individuals with a combination of criteria including, relatives with PAC, relatives with related-cancers (breast, ovarian, prostate), young age of onset, and/or a good response to chemotherapy, may be considered for genetic testing depending on the institution[42].

1.3.1 Familial Adenomatous Polyposis (FAP)

FAP is an autosomal dominant disease caused by mutations in the adenomatous polyposis coli (*APC)* gene, a tumour suppressor that plays a role in multiple key functions including cell division, polarization, and adhesion[43,44]. These individuals are prone to developing hundreds of polyps, defined as extra tissue growth in the colon[43,44]. Typically, patients diagnosed with FAP will develop colorectal cancer by age 35-40 with a lifetime risk of almost 100%; in addition, studies have shown an increased risk for other related cancers, including PAC[43-45].The absolute lifetime risk of developing PAC for FAP patients is estimated to be around 2%, while the relative

risk is increased four-fold compared with the general population[43-45]. Given the early onset of cancer for FAP patients, screening in the early ages, as early as 10-12 years old, is suggested to help prevent colorectal and other related cancers[43-45].

1.3.2 Juvenile Polyposis Syndrome (JPS)

JPS is also an autosomal dominant disease that results in the early development of polyps in the colon[46,47]. JPS is caused by mutations in either *SMAD4* and *BMPR1A*[46,47]. *SMAD4* encodes a signal transduction protein in the transforming growth factor pathway, while *BMPR1A* encodes a serine/threonine kinase receptor[46,47]. JPS is differentiated from FAP by the histology of the polyp, as the polyps of JPS are classified as "juvenile" polyps, while FAP polyps are classified as adenomatous[46,47]. Furthermore, JPS results in a smaller quantity of polyps compared with the hundreds seen in FAP patients[46,47]. Studies estimate JPS patients have a relative risk of 34.0 for colorectal cancer compared with the general population, and an absolute lifetime risk of 38.7%[48,49]. In addition, many case studies have reportedPAC cases in JPS patients, though no risk analyses have been performed given the rarity of these cases[48,50,51].

1.3.3 Peutz-Jehgers Syndrome (PJS)

Similar to the previous two, PJS is an autosomal dominant disease characterized by gastrointestinal polyps[52,53]. This disease is caused by mutations in *STK11*, a tumour suppressor gene encoding a serine/threonine kinase that plays a role in cell polarization, and apoptosis[53]. Studies estimate the relative risk of colorectal cancer to be 13.5-80 compared with general population, and pancreatic cancer relative risk may be as high as 140-fold[52-56]. Given an overall

cumulative risk of 90% for developing cancer, screening is suggested to begin at 18 years old and continue every 3 years until age 50, where screening is suggested every 1-2 years[52-56].

1.3.4 Lynch Syndrome (LS)

Lynch syndrome is characterized by mutations in the mismatch repair (MMR) genes, including *MLH1, MSH2, MSH6,* and *PMS2*[57,58].MMR is a high-fidelity conservative repair mechanism for correction of mismatched-bases during replication and homologous recombination[21]. In the absence of MMR, mutations become more frequent in the entire genome, especially in repetitive DNA elements leading to microsatellite instability (MSI)[21,59]. Lynch syndrome is characterized by a high frequency of MSI, and is responsible for1-4% of all colorectal cases[56,58]. Lifetime risk of colorectal cancer is dependent on the specific gene mutated, with studies suggesting a 30-74% for *MLH1* and *MSH2*, 10%-22% for *MSH6*, and 15%-20% for *PMS2*

[56,58,60].Similar to the previous syndrome, age of onset is younger, with a mean age estimated between 44-61, compared with sporadic cases, mean age of 69 years[56,58,60]. Thus, early screening is also suggested for individuals with Lynch syndrome[56,60]. Lynch syndrome patients are also at an increased risk for other related cancers including pancreatic cancer, which studies have shown a relative risk of 9-11 fold compared with general population[61].

1.3.5 Familial Atypical Multiple Mole Melanoma Syndrome (FAMMM)

FAMMM is an autosomal dominant disease of the skin, characterized by a large number of benign moles (>50), and a family history of melanoma[62,63].FAMMM is associated with mutations in *CDKN2A,* a tumour suppressor gene encoding a kinase inhibitor that plays a role in

cell cycle regulation[62,63]. In addition to increased risk for melanomas, some FAMMM patients have been shown to have increased risk for PAC, with an estimated relative risk of 20 to 47-fold increase compared with general population[56,62,63]. In addition, there are some studies that have shown that FAMMM patients may have a younger age of onset for PAC[56,62,63].

1.3.6 Li Fraumeni Syndrome (LFS)

LFS is an autosomal dominant cancer syndrome due to germline mutations in *TP53,*a tumour suppressor known as the guardian of the genome[64]. LFS patients are characterized by a greatly increased risk of developing several types of cancers, including breast cancer, osteosarcomas, leukemia, and brain cancers[64]. Studies have estimated lifetime cancer risk to be 73% in males, with a 54% risk of developing cancer by age 45, and 100% in females, with a 100% risk of developing cancer by age 45[56,64]. Although pancreatic cancers are not often observed in LFS patients, there have been cases reported, and studies estimate a relative risk of 7.3-fold compared with the general population[64,65]. Early screening is suggested as these patients are highly likely to develop cancers starting at a young age[64,65].

1.3.7 Hereditary Pancreatitis (HP)

Hereditary pancreatitis is characterized by recurring inflammation in the pancreas. Most cases are due to gain-of-function mutations in *PRSS1*, leading to an increased expression of cationic trypsinogen, a digestive enzyme[66-69]. These cases are inherited in an autosomal dominant fashion, however, an estimated 20% of *PRSS1* mutation carrier cases do not present with HP[56,66-69]. Studies estimate that approximately 65-80% of HP cases can be explained by *PRSS1* mutations, with other cases likely due to mutations in other genes, including *SPINK1,* a serine

protease inhibitor,and *CFTR*, a gene encoding a protein transmembrane channel linked with cystic fibrosis[64-6669,70]. HP driven by these genes are inherited in an autosomal recessive manner[69,70]. The chronic inflammation of the pancreas has been associated with increased lesions in the pancreas, leading to increased risk of PAC[56,71,72]. Studies suggest a relative risk of 7.2-fold increase compared with general population for PAC in HP patients, although it is not known whether *PRSS1* mutations increases risk for PAC in the absence of HP[56,71,72].

1.3.8 Hereditary Breast and Ovarian Cancer Syndrome (HBOC) and Hereditary Breast Cancer Syndrome (HBC)

HBOC is characterized by an increased risk in development of breast and ovarian cancer, as well as prostate and pancreatic cancers to a lesser extent[73,74].Typically, families are screened based on family history of any of the related cancers, or a young age of onset of breast/ovarian cancer[74-76].Approximately 25% of HBOC cases are attributed to mutations in *BRCA1* and *BRCA2*, with carriers estimated to have a lifetime risk of 57-65% and 45%-55% for breast cancer, 39-44% and 11-18% for ovarian cancer, 8.6% and 20% for prostate cancer, and 1-3% and 2-7% for pancreatic cancer, respectively[74-76].

In addition, *PALB2* and *ATM* have been implicated in HBC, with a lifetime risk of 35% and 38% for developing breast cancer, respectively[77-83].Both of these genes are known to play a role in DNA repair; *PALB2* is a key player in the HDR pathway, while *ATM* is an important player in determining whether the HDR or the non-homologous end joining (NHEJ) pathway is activated in response to DNA double-stranded breaks (DSBs)[84].

PALB2 has been implicated in pancreatic cancer risk, especially in familial pancreatic cancer (FPC) cases, with a prevalence of 3-4%[85,86]. However, in unselected cases, *PALB2* mutations are rarer, estimated at <1%[85,86]. Risk estimates for pancreatic cancer are not well established for *PALB2* mutation carriers.

ATM is a relatively new implicated gene in PAC risk, with a study by Roberts *et al*suggesting a prevalence of 3.2% in FPC cases as theyobserved 19 *ATM* carriersin a cohort of 593 FPC cases[87]. In addition, another recent study of 96 incident cases recruited by Mayo Clinic, suggested a *ATM* mutation prevalence rate of 4% for unselected PAC cases[39]. However, the true prevalence rate and clinical implications of *ATM* in incidentPAC cases remains to be established in further studies[88].

1.3.9 Founder Populations

The human genome consists of approximately 3 billion nucleotides, with thousands of mutations that differentiate each of our genomes[89]. However, there are certain events in history that may decrease the genetic variety within a population, such as the creation of new coloniesresulting in a founder effect[89]. These colonies have decreased genetic variation as there are only a few unique genotypes across the small number of individuals, and breeding within only the population itself leads to a plateau in genetic variation[89]. This is compounded by genetic drift effects, the loss of genotypes by random chance as individuals die or do not reproduce, leading to decrease in genetic variation in the founder population[89]. These founder effects are best noted in certain ethnic populations including the Ashkenazi Jewish and French Canadian populations[90,91].

French Canadian (FC) Population

In the 16th century, a group of 8000-10000 French immigrants settled in the Americas, giving rise to the current French Canadian population[91,92]. A majority of this population now resides in the province of Quebec in Canada[91,92]. Studies have shown that a strong founder effect exists in this population, as multiple common haplotypes have been identified in this population[91,92]. Many of these haplotypes include mutations that confer rare diseases such as cystic fibrosis, beta-thalassemia, Tay-Sachs, phenylketonuria (PKU), and HBOC[92].

HBOC is enriched in the FC population, and this is likely due to recurring founder mutations in *BRCA1, BRCA2,* and *PALB2*[91,93]. In fact, 20 founder mutations have been identified in this population; consisting of 11 *BRCA1* mutations, 8 *BRCA2* mutations, and 1 *PALB2* mutation[91,93,94]. Interestingly, one of the *BRCA2* mutations is a missense resulting in only a single amino acid change in structure[93]. However, functional assays and association studies have shown loss of functional protein as a result of this mutation[91,93].

These 20 founder mutations are thought to underlie 6% of young onset breast cancers and 16% of ovarian cancers in the FC population[93,94]. In FC families with three or more cases of breast and/or ovarian cancer, the founder mutations are thought to account for 40% of these cancers[93,94].

As these genes are also known to increase risk for PAC, individuals of the FC population are speculated to have higher incidences of PAC. The prevalence of the FC founder mutations has not been well studied in the context of PAC. However, our recent study suggests a 5.3%

founder mutation prevalence[133]. Screening guidelines for PAC in the FC population are currently not well established across the country[42].

Ashkenazi Jewish (AJ) Population

The AJ population is thought to have arisen 600-800 years ago from a mix of ancestral European and ancestral Middle Eastern descent[90]. Studies have shown that they may have had the smallest starting population of only 350 individuals, suggesting a severe founder effect[90]. Similar to the FC population, AJ individuals are at a higher risk for many autosomal recessive diseases including, Tay-Sachs, Gaucher disease, cystic fibrosis, Type-C Fanconi Anemia, and HBOC[90].

HBOC is enriched in the AJ population, likely due to increased prevalence of *BRCA1* and *BRCA2* mutations[90]. Approximately 1/40 (2.6%) AJ individuals are estimated to carry a *BRCA1* or *BRCA2* mutation, compared with a 0.2% in the general population[90]. One *BRCA2* and two *BRCA1* founder mutations have been identified in the AJ population[90]. Of the mutations identified in AJ individuals, approximately 99% are accounted for by the three founder mutations[90].

Studies have suggested that approximately 11% of breast cancer cases and 40% of ovarian cancer cases in AJs are due to the three founder mutations in this population[90]. In addition, Spring *et al.,* suggested that approximately 10% of incident AJ PAC cases are accounted for by these founder mutations[38]. The increased prevalence rate of founder mutations

in the AJ population has led to screening guidelines suggesting reflex testing for all AJ patients with breast, ovarian, or pancreatic cancer[42].

Importance of Founder mutations in FC and AJ population for PAC

Our recent study evaluated the prevalence of mutations in these populations for PAC cases and have observed a founder mutation prevalence of 5.3% and 10% in the FC and AJ populations, respectively[133]. In addition, our study observed a survival difference between HDR-gene mutation carriers (*BRCA2, BRCA1* and *PALB2*) and non-mutation carriers, especially when carriers were treated with targeted therapy[133]. This study supports the importance of reflex testing for founder mutations in these populations, and the identification of mutation carriers for both therapeutic implications for the patient and screening implications for relatives with the mutation.

1.4 Novel Susceptibility Genes in PAC

1.4.1 *ATM*- Role in Cancer Predisposition

In addition to effect on DNA repair, ATM has also been shown to be involved in checkpoint arrest through its interaction with Chk2, and *Spoerri et al*showed that loss of this checkpoint lead to genomic instability in melanoma[95].

In terms of hereditary predisposition, *ATM* causes Ataxia-Telangiectasia (A-T), an autosomal recessive disease, requiring pathogenic mutations in both copies of the gene[96]. Individuals are characterized by a deficiency in coordinated movement, including walking, balance, hand-eye coordination, and speech impairment[96]. In addition, these individuals have a

weakened immune system, leading to development of chronic infections in the lung[96]. As the immune cells are affected by this disease, A-T individuals are more likely to develop leukemia and lymphomas. As a result, the median age of death is 22[96].

As stated previously, mutations in *ATM*have been recently suggested to be associated with predisposition of PAC and prostate cancer[87]. Prior studies also suggest a role for *ATM* in the predisposition of breast cancer[77-83].Roberts *et al* has reported a prevalence of 3.4% in FPC cases, and a study of 96 incident PAC cases from Mayo Clinic by Hu *et al*suggests a 4% frequency rate of mutations in *ATM*[39,87]. However, as the role of *ATM* is quite new in PAC the penetrance, risk, and significance are not well established yet.

Clinically, studies on lymphoid tumour cells and breast cancer cell lines have shown that *ATM*-deficiency may lead to increased sensitivity to PARP-inhibitors, suggesting a defect in DNA repair in these tumours[97]. Unfortunately, the effect is not well studied and the reason for the sensitivity is not well understood[97].

1.4.2 Novel Susceptibility Gene Discovery

As stated previously, approximately 10% of pancreatic cases are thought to be hereditary in nature, however only a fraction is explained by known predisposition genes[37]. Many attempts have been made to better understand the genetics underlying PAC using different methodology, including linkage analyses, case-control association studies and genome-wide association studies (GWAS). Through linkage analyses, studies identified PAC risk to be associated with the locus 4q32-34 on chromosome 4[98]. Following, *Palladin* (*PALLD*) was identified as the PAC

susceptibility gene within the 4q32-34 locus, which led to the identification of the p.P239S variant in *PALLD* which segregated with all affected members in a FPC family[99]. This variant was suggested to overexpress *PALLD* leading to cytoskeletal changes, however, studies suggested that this was likely a private mutation and *PALLD*did not explain a significant fraction of hereditary PAC[99,100]. Furthermore, additional studies on FPC families did not support a link between the 4q32-34 locus and PAC risk[101]. In the era of GWAS, studies suggested a link between the blood type locus (ABO) and PAC risk, with blood group O being at adecreased risk for multiple cancer types, including PAC[102].More recently, studies have attempted to identify novel susceptibility genes that may increase risk for PAC using cohorts likely to be enriched for hereditary PAC, including FPC and young-onset cases[40,41].Two recent papers by Roberts *et al* and Smith *et al* attempted to identify causal genes by applying specific filtering criteria to identify protein truncating and predicted pathogenic variants for prioritizing candidate genes[40,41].

Roberts *et al* attempted to identify novel FPC susceptibility genes using a cohort of 638 FPC patients from 593 kindred[40]. Following variant calling of germline variants, they prioritized a list of all 20049 genes by the number of protein truncating variants (PTVs) identified[40]. As a result, they presented 16 DNA repair and cancer driver genes, each with three or more PTVs identified[40]. The top gene with the most PTVs was *ATM* with PTVs identified in 19 FPC kindred, followed by *TET2* with 9*,* and *DNMT3A* with 7[40]. Other genes with three or more PTVs were *POLN, POLQ, ASXL1, BRCA2, PALB2, CDKN2A, FANCG, BUB1B, ESCO2, FANCC, FANCM, MSH4,* and *RAD54L*[40]. To provide evidence to support predisposition, they evaluated 38 families for segregation where DNA was available[40]. Of the 110 genes shown to segregate with disease, 5 were implicated in DNA repair or known cancer drivers, including *ATM* and

CDKN2A[40].However, even for PTVs identified in these two genes, segregation was not always observed, likely suggesting phenocopies in these families[40]. To further characterize these susceptibility genes, somatic second hit analyses were performed for 39 FPC tumours with whole-exome sequencing (WES)[40]. Of the 16 genes identified with multiple PTVs, somatic second hits were seen in two tumours for *FANCM*, and one tumour for *BRCA2* and *BUB1B*.However, these tumours did not carry a corresponding germline PTV[40].This study provided evidence to further support the role of a few of the known predisposition genes (*BRCA2, PALB2, ATM, CDKN2A*)[40]. In addition, they observed an increased rate of PTVs in candidate genes for FPC cases compared with controls.However, they were not able to provide further evidence to suggest a role in predisposition[40].

Smith *et al*performed a similar filter-based analysis on 109 FPC and young-onset patients[41]. In this study, FPC was defined as any two relatives with PAC compared with 2 first degree relatives (FDR) with PAC in the Roberts *et al* study[40,41]. To further increase the likelihood of identifying a novel gene, this study focused on 513 genes involved in the DNA repair pathway, as most of the known susceptibility genes are involved in DNA repair[41]. Similar to the Roberts *et al* study, PTVs were identified and evaluated through visual inspection of the sequencing files followed by confirmation by sanger sequencing; 44 PTVsvalidated across 41 novel DNA repair genes[41]. Missense mutations and in-frame insertions/deletions (INDELs) were assessed in these 41 genes using bioinformatics pathogenicity prediction tools[41]. Following the identification of these mutations, segregation analyses and loss-of-heterozygosity (LOH) analyses were performed on mutation carriers[41]. Of the 41 genes, 17 were identified as strong candidates with either PTVs identified in multiple kindred, segregation within a family, or

presence of LOH[41]. The top 3 candidate genes were *FAN1*, *NEK1*, and *RHNO1*; each with mutations found in multiple kindred, and co-segregation of the variant in at least 2 kindred[41]. In fact, a study by Rouse *et al* observed tumorigenesis following the knockout of *FAN1*, including the specific variant seen in the Smith *et al* study; supporting the role of *FAN1* as a predisposition gene[103].

The filter-based approach used in these two studies are capable of identifying strong candidate genes, as demonstrated by the identification of *PALB2* as a predisposition gene for pancreatic cancer[104]. However, these methods are limited in their power to identify more faint signals due to the large amount of background "noise"as a majority of variants are likely non-causal[105]. These faint signals may be attributed to low penetrant genes, or genes with a low frequency of mutation[105]. In addition, these methods are biased, as they are limited to subjective judgement for prioritizing variants based on the current understanding of mutations and the model defined by known predisposition genes[105].

1.4.3 Burden, Variance, and Combined Regional-Based Statistical Tests

To overcome these limitations, regional-based case-control statistical tests have been developed to further increase the likelihood of identifying causal genes[105-109]. These tests rely on collapsing all variants within a region (e.g. gene or pathways)to compare mutation frequency between cases and controls[105-109].

The simplest of these tests is the burden test, which uses all variants in a gene to calculate a burden score, and tests for a difference in means for the cases versus controls[105,110]. The

limitation of the simple burden test is that it relies on the assumption that all variants will affect risk in the same direction (deleterious or protective) and is unable to account for factors such as population stratification[108,110].

In contrast, the sequence kernel association test (SKAT) is a variance-based test that calculates scores on an individual variant level prior to aggregation for testing a difference in variance between cases and controls[109]. The SKAT is more robust than the simple burden as it allows for the possibility of both deleterious and protective variants simultaneously[109]. In addition, covariates can be included in the formula to increase power by accounting for population stratification, and minor allele frequency (MAF)[109]. However, the test is limited by a higher Type-I error rate for low sample sizes ($n<500$), and suffers from decreased power when both deleterious and protective variants are present[109].

Although the SKAT test is usually higher powered, in special cases when all the assumptions of the burden test are true, the SKAT is weaker[106]. To account for different possibilities in the disease model, unified tests that combine both the burden-based and variance-based tests were developed[106]. The two most robust are the SKAT-optimized (SKAT-O) and the newer, Mixed-Effects Score Test (MiST)[106,111]. The SKAT-O is adjusted from the SKAT to allow for increased robustness by incorporating both a test for difference in mean and variance, and to include an adjustment for lower sample size ($n>200$)[106]. This allows for increased power in situations where the true disease model is likely in between the two extremes defined by the burden test and the SKAT, which is more likely to represent a real disease model[106]. In addition, the SKAT-O has improved on many of the limitations of SKAT, including an improvement

inpower loss due to the presence of both deleterious and protective variants, and better control of Type-I errors[106].

On the other hand, MiST is a hierarchical-based model that tests for a difference in both mean and variance, while accounting for heterogeneous variant effects[111]. Similar to SKAT-O, MiST is able to account for MAF, and confounders such as population stratification[111]. In addition, due to the hierarchical nature of MiST, the test is able to account for individual variant effects such as predicted protein functionality and type of mutation[111].

A recent paper by Moutsianas et al, compared the power of 11 region-based tests, including both unified tests, SKAT-O and MiST[110]. Among all the tests, excluding extreme situations, the unified tests were more likely to identify true causal associations; and between the two, MiST was higher powered across most situations, especially when using stringent significance thresholds[110].

1.5Hypothesis and Rationale

Despite recent efforts to identify new PAC susceptibility genes using a filter-based approach focused on rare inactivating variants, success has been limited and a large fraction of hereditary PAC cases remains unexplained by mutations in known susceptibility genes[37,112]. Thus, I hypothesized that the remaining fraction is likely due to genes with very rare variants of variable penetrance. The methodology of these previous studies are limited in their power to identify susceptibility genes for this disease model as it is difficult to separate out faint "signals" from the large amount of "noise" created by non-pathogenic mutations across a gene, especially given our lack of understanding of the effect of missense variants on protein function[110,111]. To better address this disease model, I hypothesized that a case-control region-based gene association test, the Mixed-effects Score Test (MiST), would provide a less biased method with greater power for identifying candidate genes associated with increased PAC risk[111]. To further increase the likelihood of identifying a candidate susceptibility gene, we focused our study on genes in the DNA damage response pathway, as a majority of known PAC susceptibility genes play a role in DNA repair. In addition, a novel statistical test, the drop-one method, can be used to identify a list of candidate variants driving the association with PAC, which allows for prioritization of variants for functional analyses[113].

1.6 Specific Aims:

1. To demonstrate the feasibility of using MiST and the drop-one method for identifying known PAC susceptibility genes and known pathogenic mutations in these genes.
2. To identify a list of candidate genes associated with increased risk for PAC, and determine the candidate variants driving the association.

Chapter II: Methods

Sample collection:

The case series consisted of PAC cases collected from two cohorts. The high-risk series (Series A) consisted of 101FPC (defined as ≥2 PAC affected relatives) and 8 young onset (<50 years old) cases collected for a previous project. The FPC cases were considered to be inherited in an autosomal dominant inheritance pattern[114].In addition, known *BRCA1* and *BRCA2* mutation carriers were excluded from this series. The Montreal-Toronto series (Series B) consists of 289 prospectively enrolled unselected PAC cases collected from two Canadian cancer registries, the Ontario Pancreas Cancer Study (OPCS) and the Quebec Pancreas Cancer Study (QPCS)[115]. The control series consists of 987 in-house samples collected locally over time across multiple projects from patients without a personal history of cancer. DNA from circulating lymphocytes or saliva was obtained from participants for germline genetic testing. Ethics approval was obtained for each study.

Full gene sequencing:

Series A was assessed for genetic variation using whole-exome sequencing (WES); exome capture was completed using Illumina TruSeq Exome Enrichment Kit (Illumina Inc., San Diego, CA, USA; n=69), Agilent SureSelect Human All Exon V4 (Agilent, Santa Clara, CA, USA; n=14) or Roche NimbleGen SeqCap EZ kit v3.0 (Roche NimbleGen Inc., Madison, WI, USA; n=26), and then sequenced on Illumina HiSeq2000 platforms with 100 base paired-end reads

Series B was assessed using a targeted panel of 710 cancer-related genes; targeted regions were captured using Agilent SureSelect technology (Agilent Technologies, Santa Clara, California, USA). Samples were sequenced on the Illumina MiSeq platform, using a V3

sequencing cartridge (Illumina Inc., San Diego, California, USA), with 300 base paired-end reads.

Variant calling:

Raw sequencing data (FASTQ) files were obtained for both case series and processed through the same pipeline as the control series, which was provided as processed and filtered Variant Call Format (VCF) files. Burrows Wheeler Alignment was used to align reads to the reference genome (Hg19)[116]. Picard was used for converting files from the SAM format to BAM, sorting and indexing of the BAM files, and to mark duplicate reads[117]. The GATK package was used to remove duplicate reads, and for the realignment of insertions and deletions (INDELs)[117]. The samtools module, mpileup, was used for variant calling to call single nucleotide variations (SNVs) and small INDELs[117]. Variants were only considered with a depth≥3, base quality≥20, and the alternate allele present in at least ≥15% of reads[117].

Variant annotation:

Multiple databases were used through ANNOVAR to annotate all variants. The RefSeq database was used to annotate gene names, location of variation (eg. exonic), and type of mutation (eg. nonsynonymous)[118]. The Exome Aggregate Consortium (ExAC), Exome Variant Server (EVS), and 1000 genomes project (1000s) were used to determine MAF of variants in the public database[118]. The Combined Annotation Dependent Depletion (CADD) database was used to determine predictive pathogenicity scores for each variant, including the scores for Polyphen-2, Sorting Intolerant from Tolerant (SIFT), and Genomic Evolutionary Rate Profiling (GERP)[119].

Principal Component Analysis (PCA):

A principal component analysis (PCA) was performed to determine whether any individual sample in the case series was genetically diverse from the rest[120]. Only exonic variants identified in the 710 genes sequenced in the Series B with a minor allele frequency (MAF) > 5% were included in the analysis. In addition, variants that were present in only one series were excluded due to difference in coverage between the two sequencing technologies. A total of 743 variants passed these criteria. A PCA plot of principal component (PC) 1 and PC2 was used to determine which individuals were >10 standard deviations from their respective ethnic populations, and thus, should be excluded from further analyses.

Candidate gene list:

Our analysis was limited to the 710 cancer-related genes sequenced on the targeted panel for Series B. Prior to analysis, only genes defined as being involved in DNA repair were included in further analyses to increase statistical power. To define these genes, the Database for Annotation, Visualization and Integrated Discovery (DAVID) pathways, and gene ontology (GO) level 4 and level 5 were used to determine which pathways each gene played a role in[121-123]. Pathways involved included, DNA repair (HDR, NHEJ, MMR, NER, BER, TLS), DNA damage response and checkpoints, end processing of break sites, and modification of DNA repair proteins. Three hundred and seventy-eight genes were determined to be involved in DNA repair. Literature review was performed for the remaining 332 genes to determine whether a role in DNA repair was supported, of which 67 genes were chosen[84,124]. A total of 445 genes were suggested to play a role in DNA damange response and repair.

Mixed Effects Score Test (MiST):

Only exonic and splicing variants with a depth≥10 in at least one sample and a MAF<0.5% in the 445 DNA repair genes were included in the analysis (Figure 1). Candidate genes with less than 10 variants across the case-control series were removed, as the MiST analysis could not be performed. Synonymous mutations were excluded due to the large quantity across all genes and lack of understanding of the effect on protein function.

Individual characteristics were not included as clinical data was missing for the control series. Both the CADD score and type of mutation (frameshift, nonframeshift, missense, nonsense, and splicing) were included as variant characteristics to increase statistical power.

Drop-one Window and Drop-one Variant:

The drop-one method, consisting of two complementary tests, was used to identify specific variants driving the association seen with MiST[113]. The first test, the drop-one window (D1W), splits each gene into smaller windows of 30 variants with a 10 variant overlap between each window. Each window was dropped one at a time, with the p-value recalculated; an increase in p-value suggested the window likely encompassed a variant of interest. For the drop-one variant (D1V), each variant within a window of interest is dropped, with the p-value recalculated; similarly, an increase in p-value suggested the variant was likely driving the association in MiST. Any variants identified in D1V with a CADD score between 0-1.0 and only seen in one case were disregarded as they were unlikely to cause loss of function and was likely due to the fact that MiST places a larger weight for the extremes of continuous variables[111,119].

Receiver Operator Curve (ROC):

To determine a significant threshold for a p-value increase in the D1V test, we performed a ROC curve analysis for *BRCA2*, as a proof of principle, since *BRCA2* is the best-established and prevalent in PAC risk. Within the 394 cases, 9 pathogenic mutations in *BRCA2* were identified. Sensitivities and specificities were calculated for different thresholds (5%-105% increase in p-value at intervals of 10%), under the assumption that all variants, except from the known pathogenic mutations, were not associated with PAC risk.

Segregation analyses:

Segregation of variants were analysed either through available sequencing data for related individuals or through PCR and sanger-based genotyping of germline DNA extracted from circulating lymphocytes, where possible.

In Silico splicing analyses:

All missense variants were assessed for loss/creation of splice sites using two *in silico* splicing prediction algorithms: Human Splicing Finder (HSF) and MaxEntScan[125,126]. For HSF, a score > 65 is considered to be a functional splice site, and any mutation that results in loss/creation of a splice site with a score difference >10% between wild-type and mutant is predicted to affect splicing. Similarly, MaxEntScan considers a variant with a score >3 with a score difference > 20% between wild-type and mutant to be predicted to affect splicing[125,126].

Power calculation:

Power was calculated for a simple Z test for testing difference in proportions between two independent groups[127]. Using data from literature, we estimated a 3% pathogenic mutation

prevalence rate for *BRCA2* in incident PAC cases and a 0.1% prevalence rate in the general population[38,39]. Using these estimates, we calculated the sample size needed to achieve 80% power for identifying*BRCA2* with a significance level of $p<0.000111$ (corrected for multiple testing for 449 DNA repair genes) to be 426 PAC cases and 852 controls. In addition, a post-hoc power calculation was performed to determine the minimum pathogenic mutation prevalence rate that would be identified as associated with PAC risk at our corrected significance threshold ($p<0.00021$) with a power of 80%.

Chapter III: Results

A Priori power calculation for sample size

As proof of the feasibility of using MiST for our model, we powered the study for identifying *BRCA2* and genes with a similar pathogenic mutation frequency.To achieve a power of 80% for identifying *BRCA2* at a significance level of $p<0.000111$ (Bonferroni correction for 449 DNA repair genes), we required 426 PAC cases and 852 controls. However, this is likely conservative as the population in Toronto and Montreal have an increased frequency of patients with AJ or FC ancestry, which would increase the pathogenic mutation rate of *BRCA2* observed due to founder mutations[38].

All variants identified across 710 cancer-related genes

A total of 21002exonic and splicing variants were identified in the 677 genes across all 1385 samples. Of these, 8390, 11283, 477, 290, 217, and 151 variants were synonymous,missense, non-frameshift INDEL, frameshift INDEL, stop gain/stop loss and splicing, respectively.Additionally, there was 194 variants with unknown consequences identified in *PRKDC*, *UHRF1*, and *VEGFA*.

Removal of genetic outliers

Of the variants in the case series, 1703 variants had a MAF>5% and passed all criteria for the principal component analysis. The PCA plot for cases showed a separation of three distinct populations, which represented the Asian ancestry population, the European ancestry population and the Central/Southern American ancestry population (Figure 2). Individuals with Asian ancestry and Central/Southern American ancestrywere not removed as the control series was also likely multi-ethnic. However, 4 individuals were removed as 3 were of Asian ancestry more than

10 standard deviations (SD) away from the Asian population, and the fourth was of multiracial ancestry more than 10 SD from any of the other populations. Unfortunately, PCA could not be performed for the control series as the raw data was unavailable.

Rare nonsynonymous variants across candidate DNA repair genes

Only 6842 variants remained after filtering for rare exonic and splicing variants (MAF≤0.5%) in 418 genes of interest, excluding synonymous mutations. Of these, 6318, 142, 157, 131, and 92 variants were missense, non-frameshift INDEL, frameshift INDEL, stopgain/stoploss, and splicing respectively. The number of variants in each gene ranged from 1-99 variants; 183 of these genes had <10 variants across all cases and controls and were removed from the MiST analysis.

Post-hoc power calculation

We performed a post-hoc power analysis to evaluate the pathogenic mutation frequency that would be identified in our study with at least 80% power at a significance level of $p<0.00021$. For genes with a minimal pathogenic mutation frequency in controls (≤0.0001%), our study is powered to identify an association for genes with a pathogenic mutation frequency of at least 2.5% in cases. As we increase the frequency observed in controls, the pathogenic mutation frequency in cases need to increase to achieve the same power with our sample size.

MiST and Drop-one analysis

Of the 235 genes tested for an association with PAC risk, 42had a p-value <0.05, including *BRCA1, BRCA2,* and *STK11* (Table 1). There were 2 genes that were significant after

multiple testing correction by Bonferroni's method (p-value < 0.00021), *RECQL*, and *SMG1*. The known genes significant at p-value < 0.05 and the candidate genes significant after multiple testing correction were brought to the drop-one analysis stage.

Prior to drop-one, we performed a ROC curve analysis using variants identified in *BRCA2* to determine the threshold for p-value increase that would provide the highest sensitivity and specificity for the drop-one test (Figure 3). Across all samples, 96 rare unique variants were identified in *BRCA2*, which was split into 5 windows of 30 variants with a moving window of up to 20 variants. The first four windows (spanning variants 1-30, 21-50, 41-70, 61-90) had an increase in p-value and thus drop-one variant was performed on these 4 windows (Figure 4 and 5). Using the ROC analysis, we determined that a threshold of 35% increase in p-value resulted in a sensitivity of 100% (95% CI 66-100%) and a specificity of 88% (95% CI 78-94%) for identifying pathogenic variants. At the 35% threshold, 19 unique variants identified in 25 cases and 1 control were identified as driving the association with PAC risk (Table 2).

The 35% increase in p-value threshold was used for the remaining genes, *BRCA1, STK11, RECQL* and *SMG1* to determine a list of candidate variants associated with PAC risk.There were 44 unique variants identified in *BRCA1*, which was split into two windows (spanning variants 1-30, 15-44) for the D1W analysis. Both windows had an increase in p-value, thus D1V was performed which identified 7 variants, including two known pathogenic frameshift variants, in 8 cases (Figure 4 and 5, Table 2).*STK11* had only 10 unique variants identified, thus the D1W was skipped. D1V identified 2 variants, however, both variants were only observed in controls (Figure 5, Table 2).

There were 24 unique *RECQL* variants identified in 16 cases (4%) and 45 controls (4.6%), of which D1V identified 3 variants (Figure 5); 1 variant was observed in 8 cases and 1 control, and the other two variants were identified in a total of 25 controls (Table 2). In the other candidate gene, *SMG1,* there were 45 unique variants identified in 41 cases (10.3%) and 45 controls (4.6%), which was split into two windows (spanning variants 1-30 and 16-45). Both windows were significant, and D1V was performed for both windows (Figure 4 and 5). This identified 15 variants, of which 13 were missenses and 2 were splicing variants, across 29 cases and 2 control (Table 2).

Evaluation of candidate variants in BRCA1, BRCA2, and SMG1

First, we evaluated the list of candidate variants in the two known predisposition genes, *BRCA1* and *BRCA2*. Excluding the known pathogenic variants, there were 5 missenses in *BRCA1* and 8 missenses in *BRCA2.* However the 5 missenses in *BRCA1* were discarded as they had a CADD score between 0-1.0. The 8 missenses in *BRCA2* were identified in 13 cases and 1 control, and the information is summarized in Table 2. Unfortunately, we were unable to perform further validation for these variants astumour tissue was not available for these cases to determine whether there was a somatic 2nd hit, and there was also no opportunityfor segregation studies.

The list of candidate variants in *SMG1* were evaluated for somatic second hit and segregation opportunities (Table 2).Unfortunately, tumour tissue was not available for any of the cases with an identified mutation to investigate for somatic inactivation of the second allele. However, there was a segregation opportunity for two families with multiple PAC diagnoses in

each family (Figure 6). For the first family (A-78), the mutation was identified in two related individuals in our case series, the proband and the maternal aunt(Figure 6a). We were able to confirm the mutation in one of two maternal first cousins whose father had PAC. Thus, the mutation segregated with all 3 individuals with PAC on the maternal side, although, the father of the proband also had PAC. In the second family (B-105), there was a history of PAC on both the maternal (1 relative) and paternal side (3 relatives) of the family (Figure 6b). The mutation was identified in the paternal aunt in our case series, however, it did not segregate in the proband with PAC, possibly representing phenocopies in the family. Unfortunately, samples for other paternal relatives with PAC and their children were not available to be genotyped to determine whether the *SMG1* variant was driving the PAC on the paternal side of the family.

To further evaluate the consequence of the variants identified in D1V, we performed in-silico splicing prediction analyses for all missense variants for either gain or loss of a canonical splice acceptorand/or splice donor site, which may result in alternate non-functional protein transcripts (Table 1). Interestingly, the variant identified in family A-78, which segregated with 2 relatives with PAC, was predicted to create both a splice acceptor and splice donor site. In addition, the variant identified in family B-105 tested for segregation was also predicted to create a splice donor site.

Figures and Tables

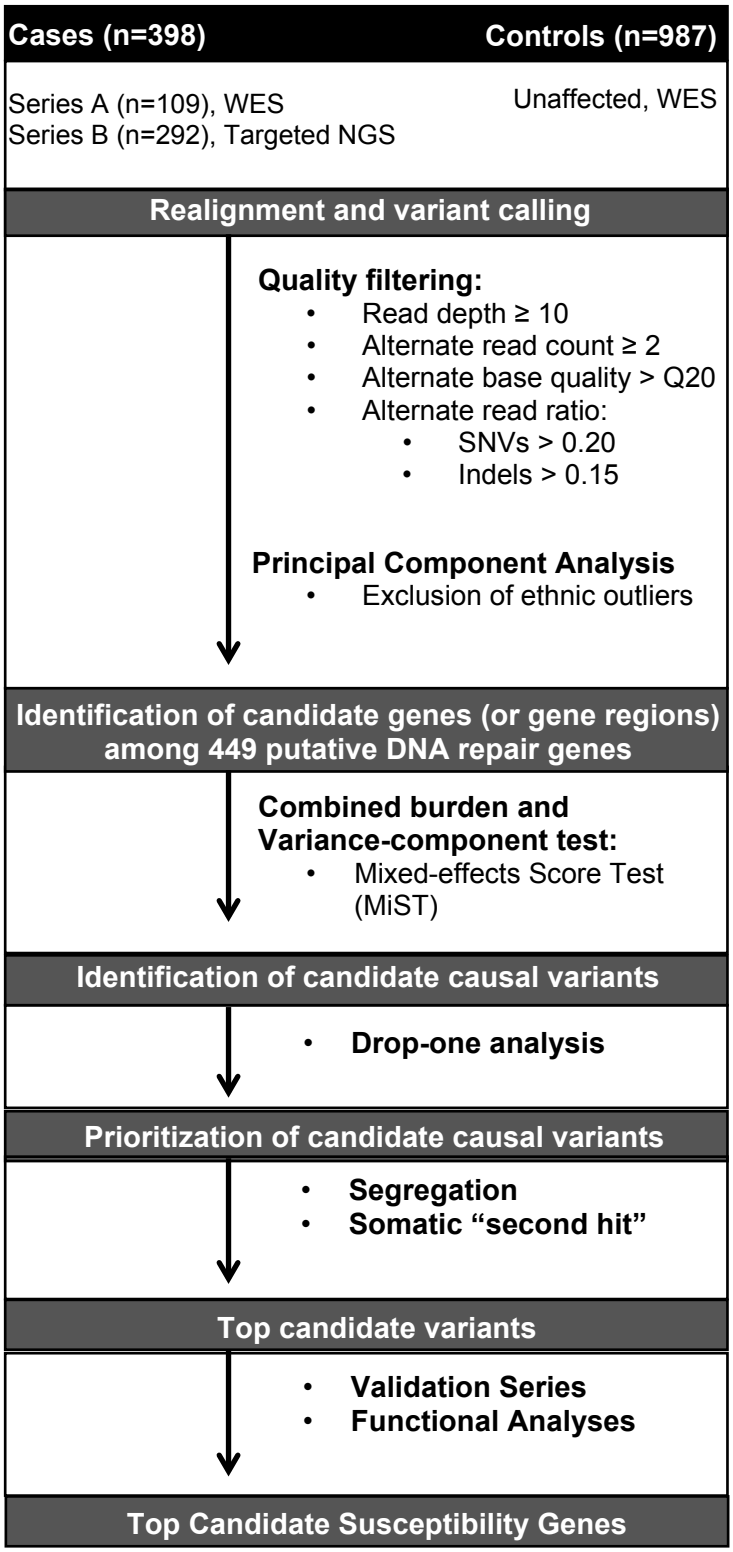


Figure 1.Schematic of gene association study design.(Above) Series A, cases at high risk for hereditary PAC; Series B, unselected prospectively collected PAC cases; WES, whole-exome sequencing; NGS, next-generation sequencing; SNVs, single nucleotide variants; Indels, insertions/deletions.

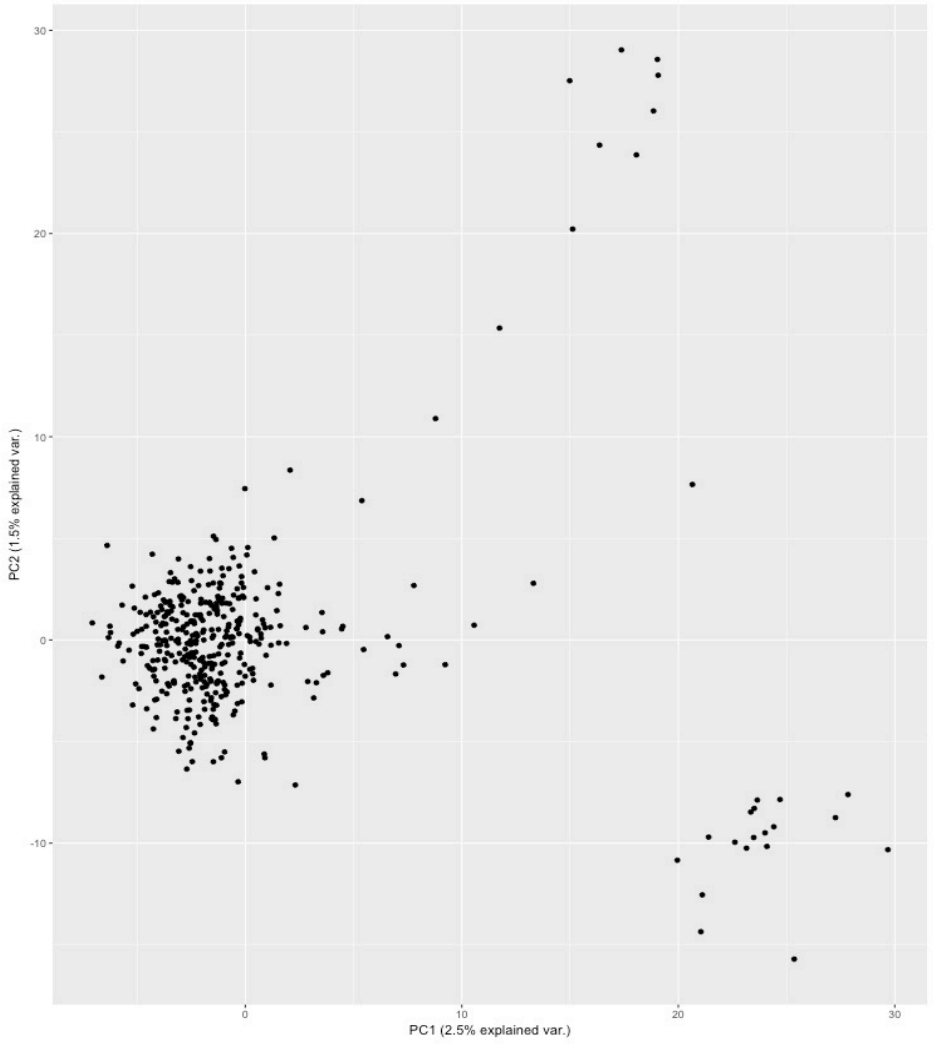


Figure 2. PCA plot for PAC cases (Series A and Series B). Using exonic and splicing variants with MAF >5% in 710 cancer-related genes.

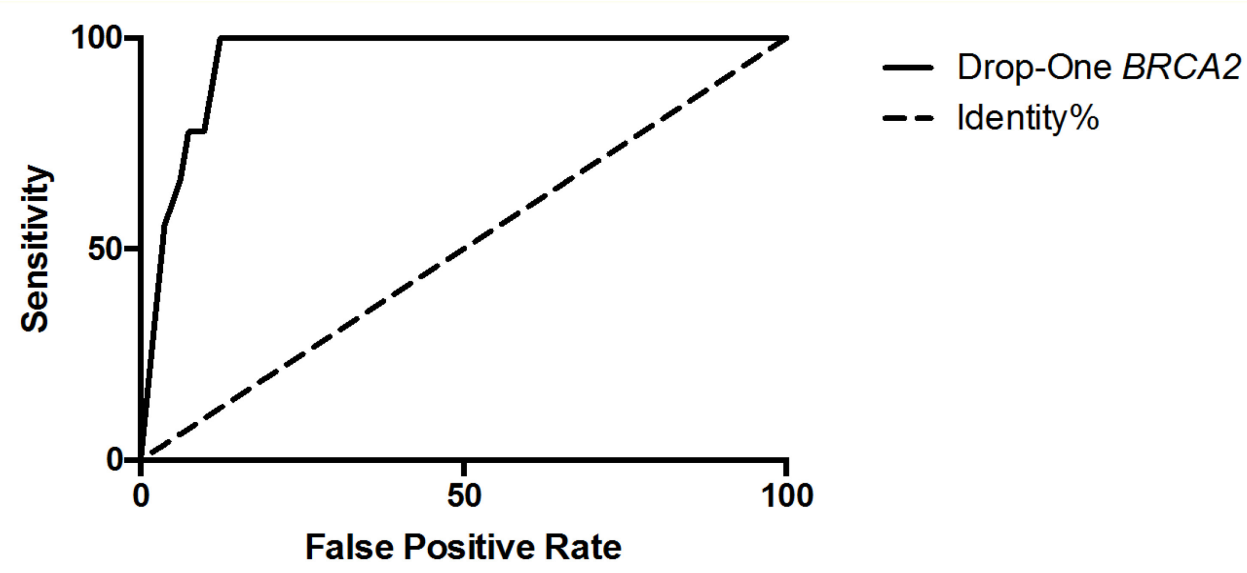


Figure 3. ROC curve for *BRCA2* drop-one analysis. Solid line represents the sensitivity and false positive rate for different p-value increase thresholds (5%-105%) for the D1V analysis for *BRCA2*. The dotted line represents the identity line for a 50/50 test.

Table 1. List of genes with a p-value ≤ 0.05 in MiST. The 42 DNA repair genes with a significant association ($p \leq 0.05$) in MiST. Bolded genes have a p-value that is significant after multiple testing correction by Bonferroni ($p \leq 0.00021$).

Gene	MiST P-value
AATF	0.020793742
ALKBH3	0.001178369
ASTE1	0.02217663
ATR	0.010081606
AXIN2	0.02377471
BAZ1B	0.007698043
BRCA1	0.029705866
BRCA2	0.002067188
BUB1	0.030246121
CDC25B	0.002881731
CDH1	0.022117929
CHEK2	0.033475147
CRB2	0.047336481
CREBBP	0.03592323
DCLRE1C	0.009149606
DDX1	0.033438119
ERCC3	0.046283948
FAM175A	0.040531643
FANCM	0.002860933
JMY	0.02609159
NEK1	0.019974732
NEK11	0.038891494
PARG	0.001433575
PARP4	0.025116917
POLE	0.031026389
POLL	0.009017405
RAD9A	0.020940274
RASSF1	0.015921096
RBM14	0.008426823
RECQL	**0.000159041**
RFWD2	0.010593987

SETD2	0.041731183
SMC5	0.040889598
SMG1	**3.22E-07**
STK11	0.038894861
TDG	0.003088256
TET1	0.011589982
USP1	0.005562616
UVRAG	0.000333677
WRN	0.018082861
XAB2	0.007911871
XPA	0.012377063

Figure 4. The –log p-value graphs for the D1W analysis for *BRCA2, BRCA1* and *SMG1*. A decrease in the –log p-value is an increase in p-value signifying the window dropped contains variants driving the association with PAC risk. Any window with an increase in p-value was analyzed by D1V for potential variants of interest. A) D1W for *BRCA2*. B) D1W for *BRCA1*. C) D1W for *SMG1*.

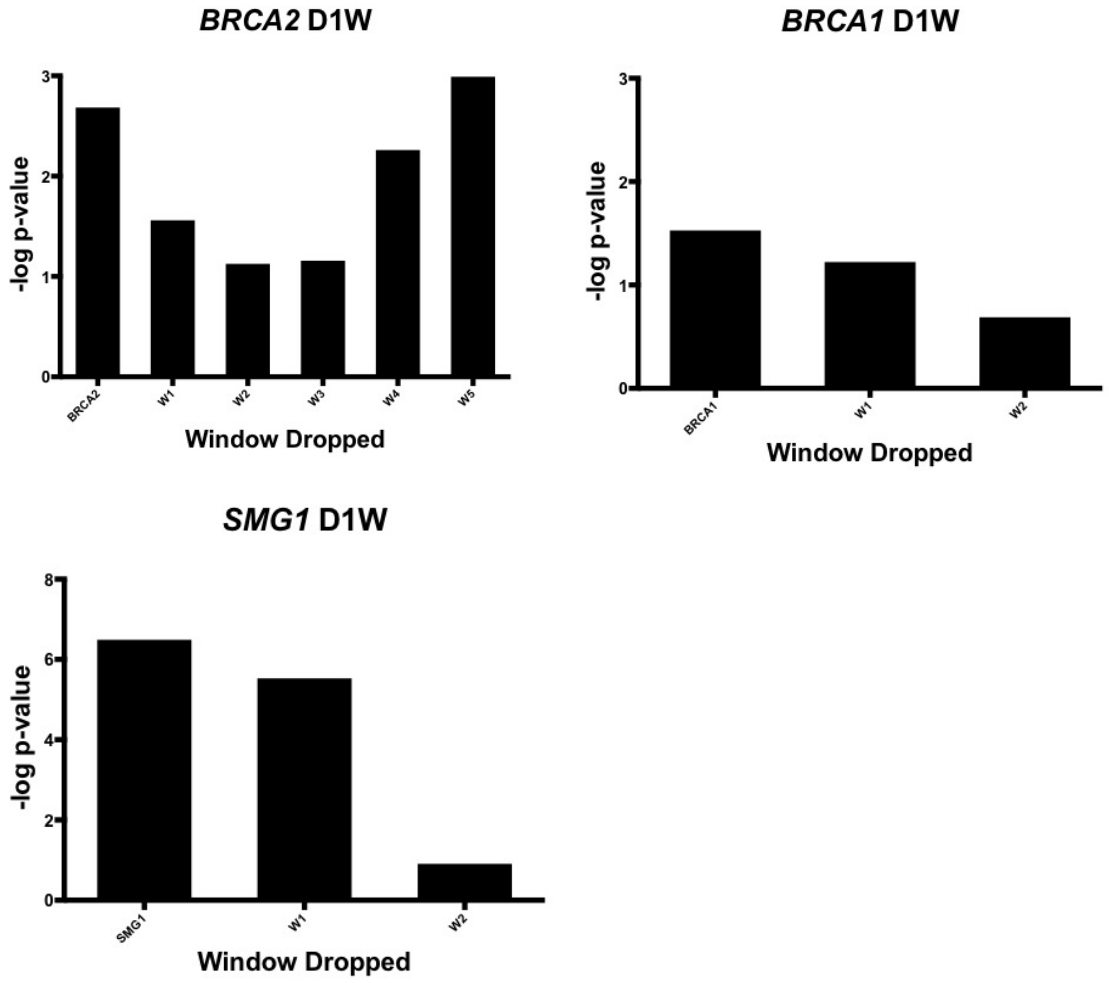
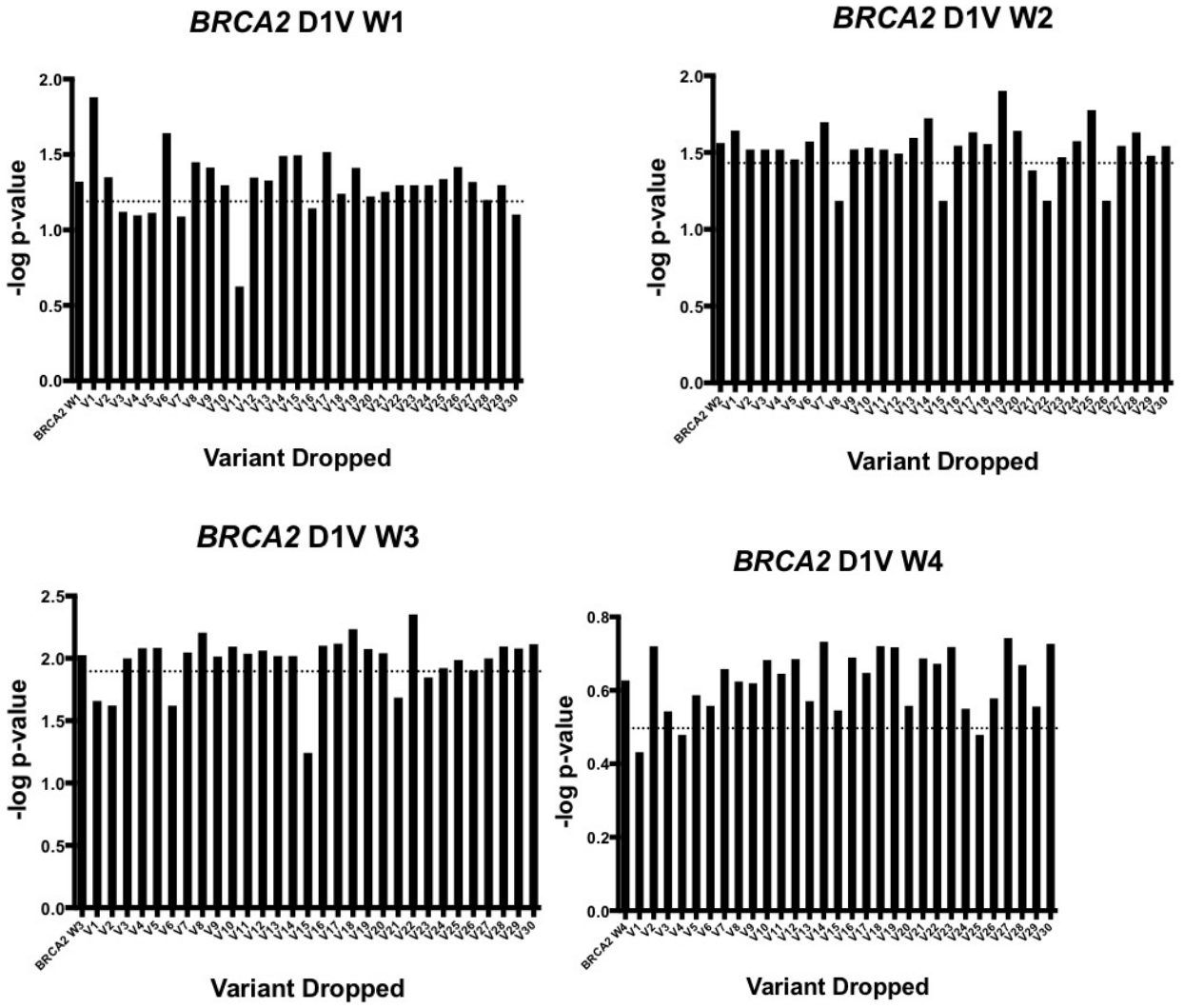
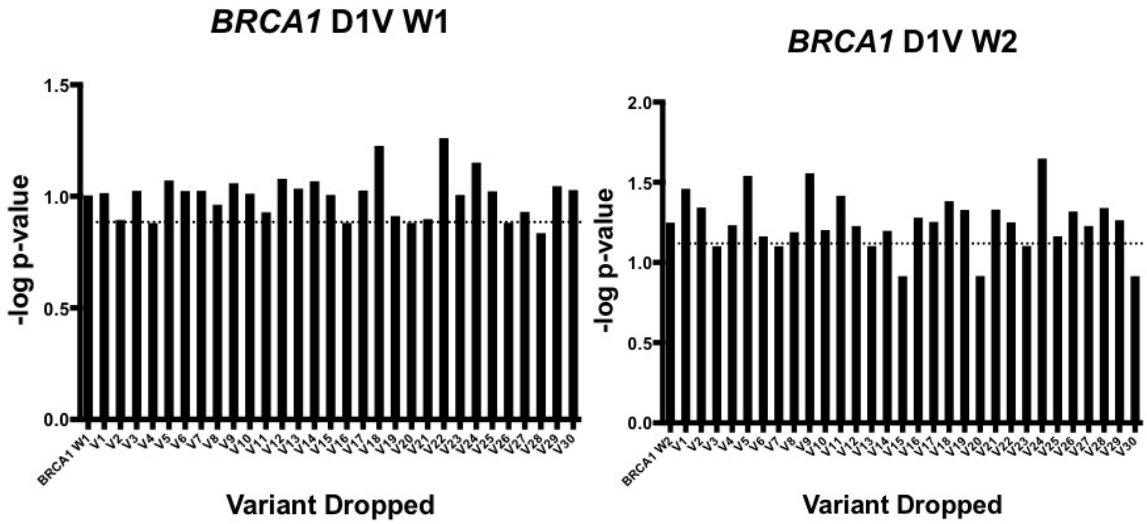
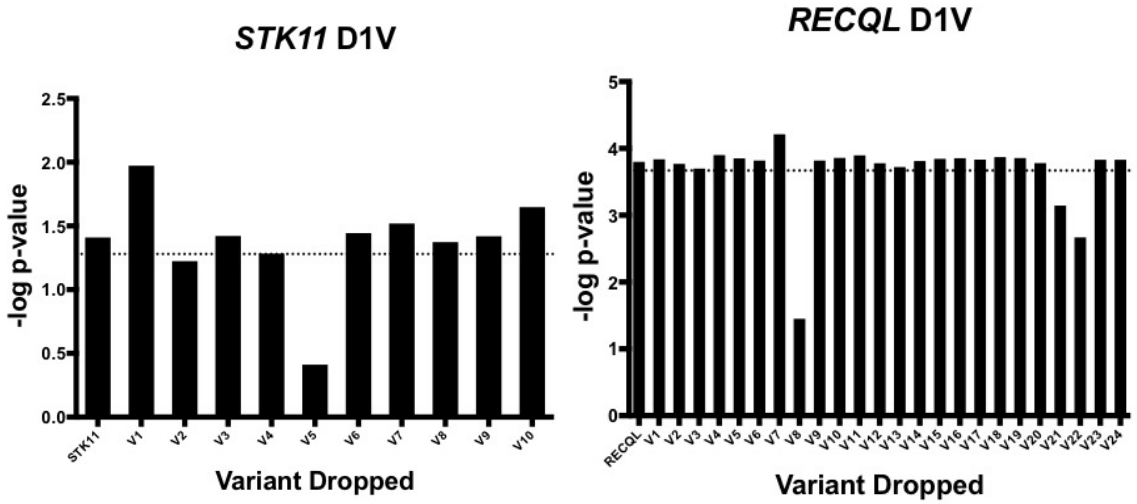
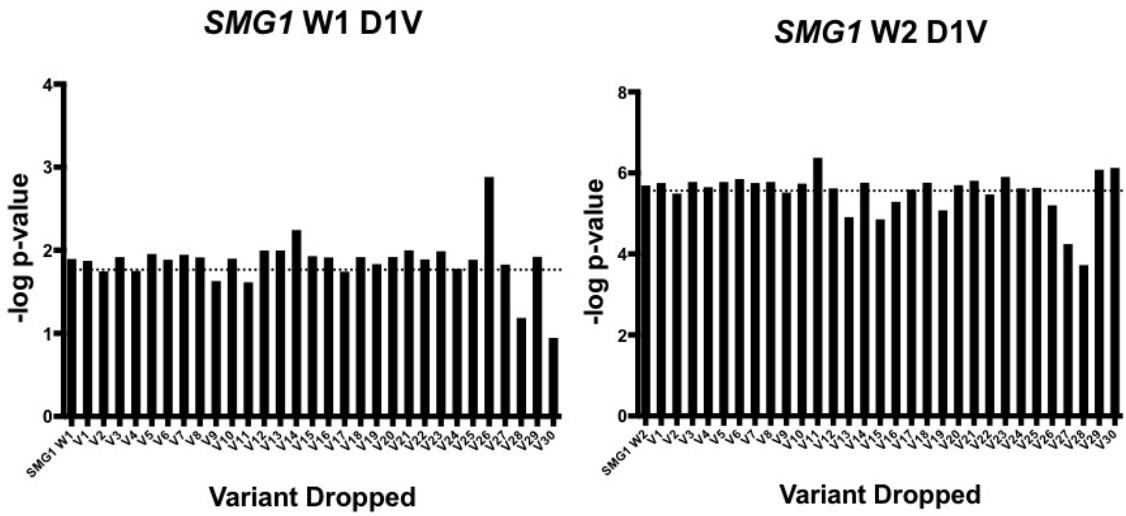


Figure 5. The –log p-value graphs for D1V analysis for each significant window for *BRCA2, BRCA1, STK11, SMG1, RECQL.* A decrease in the –log p-value is an increase in p-value signifying the variant dropped is potentially driving the association with PAC risk. The dotted line represents a decrease in –log p-value of 35% (increase in p-value of 35%) compared with the p-value of the window; this is the threshold for identifying a candidate variant. A) D1V for window 1 to window 4 of *BRCA2*. B) D1V for window 1 and window 2 of *BRCA1*. C) D1V for *STK11*. D) D1V for *RECQL* of E) D1V for window 1 and window 2 of *SMG1*.




BRCA1 D1V W1
-log p-value
Variant Dropped
BRCA1 D1V W2
-log p-value
Variant Dropped


STK11 D1V
-log p-value
Variant Dropped
RECQL D1V
-log p-value
Variant Dropped


SMG1 W1 D1V
-log p-value
Variant Dropped
SMG1 W2 D1V
-log p-value
Variant Dropped

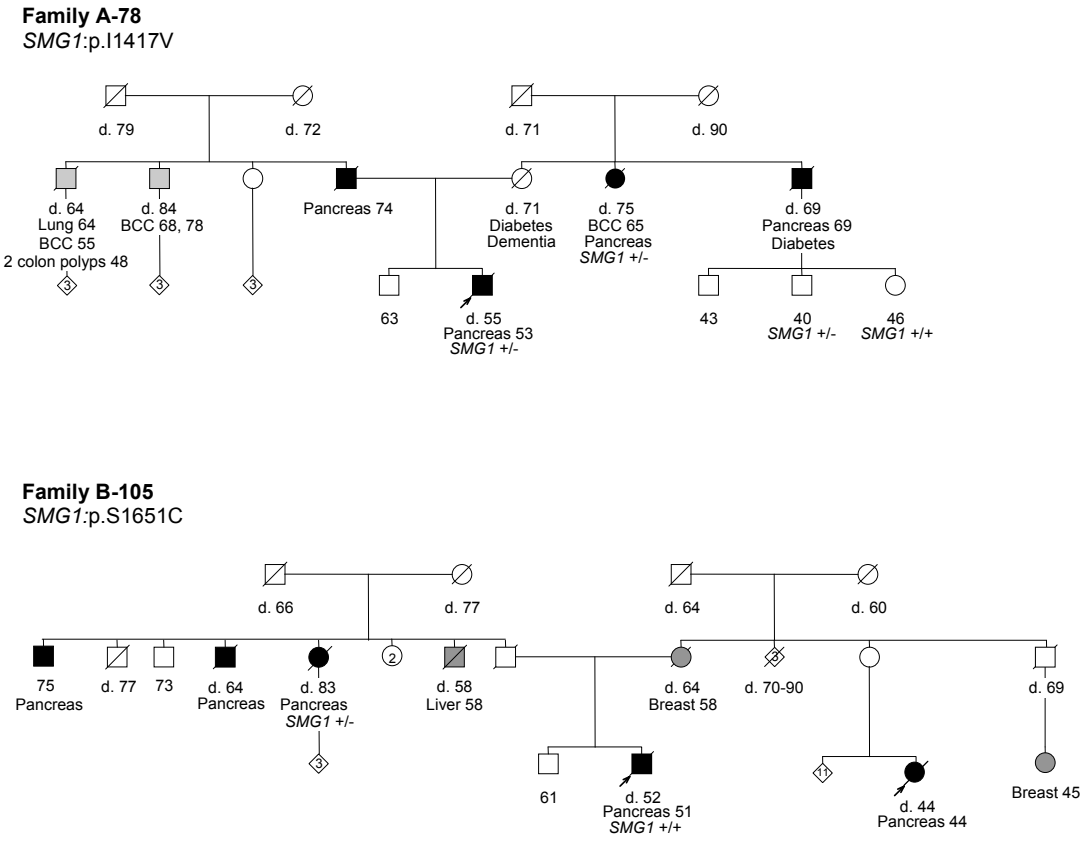


Figure 6. Pedigrees for two *SMG1* variant carriers with segregation opportunities. In family A-78, both the proband and the maternal aunt were included in our case series and were identified to be carriers of the p.I1417V *SMG1* variant. Although the maternal uncle with PAC did not have samples available for testing, we were able to infer his genotype through one of his sons, whose sample was available. In family B-105, the maternal aunt was included in our case series and was identified to be a carrier of the p.S1651C *SMG1* variant. The proband in this family was found to not be a carrier of this *SMG1* mutation.

Table 2. Summary of mutations identified in drop-one analysis for 3 known susceptibility genes and 2 candidate genes. (Below) Minor allele frequency for our case series, control series, and 3 public databases are presented, as well as the CADD score. The p-value increase observed for the drop-one variant test is presented, as well as the prediction on splicing for missense variants. MAF, minor allele frequency; CADD, combined annotation depletion dependent; EVS, Exome Variant Server; ExAC, Exome Aggregation Consortium; 1000s, 1000 genomes project; HSF, Human Splicing Finder; MES, MaxEntScan; BD, broken splice donor site; CA, creation of splice acceptor site; CD, creation of splice donor site.

Gene	Mutation	Type	Case MAF	Control MAF	CADD	EVS	ExAC	1000s	HSF/MES	p-value increase (%)
BRCA2	c.658_659delGT:p.V220fs	Frameshift	0.0013	0	24.3	.	4.9E-5	.	.	>55%
BRCA2	c.927delA:p.S309fs	Frameshift	0.0013	0	26.4	.	.	.	.	>65%
BRCA2	c.3109C>T:p.Q1037X	Stopgain	0.0013	0	37	.	.	.	.	>105%
BRCA2	c.4171G>T:p.E1391X	Stopgain	0.0013	0	41	.	.	.	.	>105%
BRCA2	c.4691dupC:p.A1564fs	Frameshift	0.0013	0	25.6	8E-5	.	.	.	>105%
BRCA2	c.5064dupA:p.E1688fs	Frameshift	0.0013	0	34	.	.	.	.	>105%
BRCA2	c.5946delT:p.S1982fs	Frameshift	0.0026	0	35	.	2.6E-4	.	.	>105%
BRCA2	c.7008-2A>T	Splicing	0.0013	0	23	.	.	.	.	>35%
BRCA2	c.8677C>T:p.Q2893X	Stopgain	0.0013	0	51	.	.	.	.	>35%
BRCA2	c.223G>C:p.A75P	Missense	0.0026	0.0005	23.8	3.1E-4	1.6E-4	.	.	>55%
BRCA2	c.631G>A:p.V211I	Missense	0.0013	0	26	.	.	.	BD	>65%
BRCA2	c.1151C>T:p.S384F	Missense	0.0039	0	23.4	1.1E-3	6.8E-4	.	.	>105%
BRCA2	c.1780A>T:p.I594L	Missense	0.0013	0	20.5	.	1.6E-5	.	.	>45%

BRCA2	c.3568C>T:p.R1190W	Missense	0.0013	0	25.3	.	1.1E-4	.	.	>65%
BRCA2	c.4585G>A:p.G1529R	Missense	0.0026	0	29.2	4.6E-4	4.2E-4	.	CA	>105%
BRCA2	c.6853A>G:p.I2285V	Missense	0.0026	0	25.5	2.3E-4	2.7E-4	.	CA	>105%
BRCA2	c.6953G>A:p.R2318Q	Missense	0.0013	0	35	.	1.6E-5	.	.	>45%
BRCA2	c.7928C>G:p.A2643G	Missense	0.0013	0	32	7.7E-5	2.5E-5	.	.	>35%
BRCA2	c.8279G>T:p.G2760V	Missense	0.0013	0	31	.	.	.	CD	>35%
BRCA1	c.2125_2126insA:p.F709fs	Frameshift	0.0013	0	22.3	.	.	.	.	>105%
BRCA1	c.66_67del:p.L22fs	Frameshift	0.0013	0	32	.	2.2E-4	.	.	>105%
BRCA1	c.3418A>G:p.S1140G	Missense	0.0013	0	0.19	0.011	0.003	.	CD	>35%
BRCA1	c.2750T>C:p.I917T	Missense	0.0013	0	0.31	.	.	.	.	>35%
BRCA1	c.2207A>C:p.E736A	Missense	0.0013	0	0.42	.	.	.	.	>35%
BRCA1	c.1487G>A:p.R496H	Missense	0.0026	0	0.004	6.1E-4	4.7E-4	.	.	>105%
BRCA1	c.1243G>A:p.V415I	Missense	0.0013	0	0.51	.	4.1E-5	.	.	>35%
STK11	c.316C>T:p.R106W	Missense	0	0.001	33	.	1.6E-5	.	.	>45%
STK11	c.1062C>G:p.F354L	Missense	0.0013	0.01	0.84	0.0038	0.0049	.	.	>105%
RECQL	c.1460A>C:p.K487T	Missense	0.01	0.0005	14.4	0.017	0.0057	.	.	>105%
RECQL	c.304G>A:p.V102I	Missense	0	0.0056	21.1	0.048	0.013	.	.	>105%
RECQL	c.207T>A:p.N69K	Missense	0	0.0061	11.8	.	9.1E-5	.	.	>105%
SMG1	c.10921A>G:p.I3641V	Missense	0.0013	0	18.52	.	.	.	.	>35%
SMG1	c.10685T>C:p.L3562P	Missense	0.0013	0	18.78	.	.	.	.	>35%
SMG1	c.9145C>T:p.L3049F	Missense	0.0013	0	12.59	.	.	.	.	>85%
SMG1	c.8665G>A:p.G2889S	Missense	0.0013	0	11.67	.	0.0013	.	CA	>85%
SMG1	c.7447G>A:p.V2483I	Missense	0.0013	0	18.38	.	.	.	.	>55%
SMG1	c.4952C>G:p.S1651C	Missense	0.0013	0	19.77	8.4E-5	2E-4	.	CD	>45%
SMG1	c.4249A>G:p.I1417V	Missense	0.0026	0	6.2	.	1.7E-5	.	CD/CA	>105%
SMG1	c.3917C>T:p.P1306L	Missense	0.0052	0.0005	8.947	0.0018	0.0017	.	.	>105%
SMG1	c.3773A>G:p.N1258S	Missense	0.0013	0	4.325	.	4.2E-5	.	CA	>105%

SMG1	c.2494A>G:p.N832D	Missense	0.0026	0	12.03	.	0.0021	.	.	>105%
SMG1	c.1835T>A:p.I612K	Missense	0.0013	0	17.25	.	0.043	.	.	>65%
SMG1	c.256+2delTC	Splicing	0.0013	0	24.1	.	.	.	.	>105%
SMG1	c.256+2delGA	Splicing	0.0039	0	24.1	.	.	.	.	>105%
SMG1	c.103G>A:p.A35T	Missense	0.0088	0.0005	21.8	0.0027	0.019	.	.	>105%

Chapter IV: Discussion

Over the past decade, PAC has remained one of the deadliest solid tumours with a 5-year survival rate of only 8%[8,10,11]. The poor prognosis is likely due to the chemoresistance of the tumours and the late presentation of the disease at diagnosis[8,10,11]. This indicates that there is a need for improvements in both therapeutic options and early screening strategies[8,10,11]. One common risk factor is hereditary mutations in key predisposition genes that increase an individual's risk of developing cancers, such as the infamous *BRCA1* and *BRCA2* for HBOC[37]. An estimated 10% of PAC cases present with familial clustering (usually defined as ≥2 PAC diagnoses in a family[114]), with only a fraction explained by known predisposition genes, of which a majority of cases are explained by mutations in the HDR pathway (*BRCA1, BRCA2, ATM,* and *PALB2*[37-39]*).* Genetic screening for patients with pathogenic mutations in these genes have not only allowed benefits for patients through precision therapy designed to target the specific defects of these tumours, but there are implications for relatives that are carriers as they may be eligible for both clinic-based and research-based screening programs to help with early detection.

Recent studies by Roberts et al and Smith et al, have tried to identify novel susceptibility genes using a filtered-based approach[40,41]. These studies demonstrated the challenges in identifying "faint signals" from a large amount of "noise" from non-causal variants[105,128].In addition, the filter-based approach is biased by our limited understanding of the consequences of mutations by focusing only on PTVs, while excluding the role of missense mutations[105]. Furthermore, a recently published paper by Grant *et al*, attempted to identify novel PAC susceptibility genes using an exome-wide case-control association study using 437 unrelated PAC cases and 1922 non-cancer controls[112]. Aside from *BRCA2*, they identified several

candidate genes with suggestive evidence. In addition, they performed a variance-based regionassociation test, the SKAT, which identified seven genes with suggestive evidence ($p<0.001$). However, no gene reached exome-wide significance ($p<2.5x10^{-6}$) for both the classic case-control association test or the variance-based region association test. The authors conclude that larger collaborative initiatives and use of novel statistical tests will be required to identify novel susceptibility genes due to the genetic heterogeneity of PAC. Coincidently, in this study, we decided to use a less biased novelstatistical method to help identify candidate genes that may increase risk for PAC[105]. To further increase the power of our model, we focused on genes involved in the DNA repair pathway, as a majority of the known PAC susceptibility genes are DNA repair genes. We used a region-based association test,MiST, that compares mutations found in a gene for cases vs controls, while factoring in individual variant characteristics, including mutation frequency, predicted pathogenicity, and type of mutation[111]. Following MiST, we used the drop-one method, consisting of two complementary tests, to identify variants of interest for evaluation[113].

MiST was performed for 235 DNA repair genes which passed all of our criteria, and of these, 42 genes were significant prior to multiple testing correction, including the known predisposition genes, *BRCA1*($p=0.03$), *BRCA2*($p=0.002$), and *STK11*($p=0.039$). However, these 3 genes were not significant after multiple testing correction ($p<0.00021$). *BRCA1* and *STK11* are likely not significant due to the fact that pathogenic mutations have been shown to be rare in these genes and our study was powered to find genes with a pathogenic mutation frequency at approximately 3% in cases and 0.3% in controls. Based on the literature, our study should be powered to identify *BRCA2* as a predisposition gene. However, it was likely insignificant after

correction in our discovery series because known *BRCA2* mutation carriers were excluded when we were collecting the 109 high-risk PAC cases for our previous project[38,39,41]. This likely lowers the expected *BRCA2* pathogenic mutation frequency in our case series and thus results in a less significant p-value.

The consequences of VUS is a recurring problem in both research and clinical genetics, as it is difficult to determine which variants may affect gene function and are causal for increased risk of disease[129]. Using the drop-one method in combination with MiST, we are able to identify a list of variants, including both PTVs and missenses, driving the association between *BRCA2* and PAC risk (Table 2), which can be used to prioritize functional experiments to determine the effect of these VUS. As we move forward, it will be important to characterize the effects of VUS as they may play a role in predisposition and have clinical implications for early detection strategies for mutation carriers[129].

In addition to the three known susceptibility genes (*BRCA1*, *BRCA2*, and *STK11*), we identified two candidate genes significant after correction, *RECQL*(0.00016) and *SMG1*(3.22×10^{-7}). However, the signal for *RECQL* was mainly driven by the two variants seen exclusively across 25 controls. As protective effects are difficult to evaluate and validate with our sample size, we decided to exclude *RECQL* variants from further evaluation.

On the other hand, the association with *SMG1* was largely driven by 19 unique variants identified across 22 cases and 1 control. To further evaluate these variants, we collected clinical data for these cases, including age of diagnosis, stage at diagnosis, related cancers in both the

proband and relatives, specimen availability, and opportunity for segregation analyses. Unfortunately, a majority of these cases were late stage cancers, and thus tumour specimens were unavailable for somatic mutation testing. There were two families for which segregation opportunities were available.

The first family was for the *SMG1* variant p.I1417V that was identified in two-related PAC cases, the proband and a maternal aunt (Figure 6a). This variant segregated in one of two children of a maternal uncle with PAC. Family members on the paternal side were not evaluated, as the mutation was already shown to segregate in the maternal aunt in our initial case-control series. Thus, this variant segregated with all 3 PACs on the maternal side of the family, although the father of the proband also had PAC, which could possibly be aphenocopy.As this variant is a missense variant, without proven pathogenicity, we evaluated the potential for splicing alteration. HSF predicted that this variant created both a canonical splice donor and splice acceptor site. To further characterize the effects of this variant, we are currently using CRISPR/Cas9 technology to create cell lines which are either homozygous or heterozygous with the other allele inactivated. Using these cell lines, we plan on evaluating the potential effects on splicing and on the function of downstream effectors of *SMG1*.

The *SMG1* variant p.S1651C was identified in the paternal aunt of the second family. The proband for this family was sequenced as part of a different project and was found to not carry this *SMG1* variant. However, this family has 5 affected individuals with PAC, of which 3 are on the paternal side, 1 on the maternal side, and the proband(Figure 6b). Thus, even though the mutation did not segregate in the proband, it is possible that the proband is a

phenocopyinheriting a risk allele from the maternal side. Unfortunately, segregation opportunities from the other two affected individuals on the paternal side and their children were unavailable.

SMG1 is a serine/threonine-protein kinase in the same protein family as *ATM*[130-132]. *SMG1* has been shown to be a part of the mRNA surveillance complex involved in the nonsense-mediated mRNA decay pathway and regulating p53 function following cell exposure to genotoxic stress[133]. In fact, there have been studies linking loss of *SMG1* function with tumorigenesis[130,131]. Gubanova et al evaluated the in-vitro effect of *SMG1*-knockdown in U2-OS cells[130].In cells with loss of *SMG1*, they observed that p53 activity was decreased following ionizing radiation (IR) compared with *SMG1*-wildtype cells, leading to proliferation of cells[130]. In addition, they evaluated the effect of *SMG1* on two cell cycle checkpoint proteins, Cdc25a and CDK2[130]. In normal cells, Cdc25a is responsible for removing the inhibitory phosphorylation of CDK2 required for transition into S-phase[130]. In response to genotoxic stress, Cdc25a degradation is induced leading to maintained inhibition of CDK2 which causes cell cycle arrest[130]. In *SMG1*-deficient cells, exposure to IR was unable to induce degradation of Cdc25a and also resulted in an upregulation of Cdc25a leading to increased activation of CDK2 allowing for cell cycle progression through the G1/S transition leading to further cell proliferation[130]. Following in vitro analyses, Gubanova et al injected HA1EB cells with either a retrovirus expressing SMG1-shRNA or an empty vector into 5 mice each[130]. They observed that mice with cells expressing the SMG1-shRNA developed subcutaneous tumours more rapidly compared to mice with normal expression of SMG1[130]. In addition, they observed that tumours in SMG1-shRNA expressing mice reached the 300 mm^3 end point more rapidly compared to mice with normal SMG1 expression ($p < 0.05$)[130].

Another study by Roberts et al, showed that *SMG1* heterozygous knockdown mice had shorter lifespans compared to wildtype mice, and they were significantly more likely to develop lung papillary adenocarcinoma and chronic inflammation in the kidney and lungs[131]. The heterozygous mice showed no difference in DNA damage response compared with wildtype. However, this may be due to the fact that one functional copy of *SMG1* is sufficient for normal DNA damage response[131]. In addition, *SMG1* homozygous knockdown mice were embryonic lethal; however, the authors suggested this was likely due to the critical role of NMD in embryonic development[131]. These studies suggest that *SMG1* has a role as a tumor suppressor, and in the absence of *SMG1*, tumours are more likely to develop.

To further provide evidence for *SMG1* as a candidate susceptibility gene for PAC, we are currently validating the association using another case-control series consisting of 532 FPC cases and 754 controls from our collaborators at John Hopkins. We will perform both a Fischer's exact test to evaluate whether there is an increase in mutation frequency in cases versus controls, as well as the same MiST analysis using the same model as our series.

In this study, we have shown that a region-based statistical approach is useful for helping to prioritize a list of genes for association with increased risk for PAC. As a proof of principle, we were able to show that MiST was able to identify *BRCA1* and *BRCA2*, although not significant after multiple-testing correction. This may be due toour study design, sample size, and/or the rarity of pathogenic mutations in *BRCA1*. In addition, using the drop-one method, we were able to identify all the known pathogenic mutations in both *BRCA1* and *BRCA2*, as well as

a list of candidate missense variants that may be pathogenic. Thus, this method can be useful in prioritizing candidate variants for further functional experiments to help characterize the effects of VUS. Our study identified one candidate gene, *SMG1*, where the association was driven by an increased frequency of mutations in cases vs controls. We are currently further validating this finding in another series and through functional assessment of a variant that segregated with 3 family members with PAC.

Although our study uses a less biased statistical approach, it is still influencedby the individual variant characteristics that we assign based on our current knowledge of variant consequences. As suggested by Grant *et al*, larger collaborative initiatives will be required, as the remaining fraction of hereditary predisposition is likely due to genes with a low frequency of pathogenic mutations, and with our current sample size, it is very difficult to identify a true positive for this model. We also included multiple individuals from the same family in our discovery series, which may bias the mutation frequency as these related individuals will increase the mutation frequency of non-causal rare variants that are shared. However, this is rationalized as we are trying to increase power for identifying a list of candidate genes in our discovery series which will be validated in our validation series.

In summary, this dissertation has demonstrated the feasibility and utility of a region-based gene association study in combination with a variant prioritization test to identify candidate genes associated with a rare disease. We identify a candidate susceptibility gene for PAC which seems to be more frequently mutated in PAC cases vs non-cancer controls. In addition, we present a list of VUS in a well-established PAC predisposition gene, *BRCA2,* that

can be used to further prioritize functional studies aimed to characterize the effect of rare missense variants on protein function and HDR-efficiency.

Chapter V: Bibliography

1. DePhilip RM. Atlas of Human Anatomy, Fourth Edition, by Frank H. Netter and edited by Jennifer K. Brueckner, et al. *Clinical Anatomy*. 2008;21(7):735-736. doi:10.1002/ca.20682.

2. Slack JM. Developmental biology of the pancreas. *Development*. 1995;121(6):1569-1580.

3. Hosoda W, Wood LD. Molecular Genetics of Pancreatic Neoplasms. *Surg Pathol Clin*. 2016;9(4):685-703. doi:10.1016/j.path.2016.05.011.

4. Distler M, Aust D, Weitz J, Pilarsky C, Grützmann R. Precursor Lesions for Sporadic Pancreatic Cancer: PanIN, IPMN, and MCN. *BioMed Research International*. 2014;2014(9):1-11. doi:10.1155/2014/474905.

5. Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. *Lancet*. 2011;378(9791):607-620. doi:10.1016/S0140-6736(10)62307-0.

6. Hruban RH, Maitra A, Kern SE, Goggins M. Precursors to pancreatic cancer. *Gastroenterol Clin North Am*. 2007;36(4):831–49–vi. doi:10.1016/j.gtc.2007.08.012.

7. Hruban RH, Maitra A, Goggins M. Update on pancreatic intraepithelial neoplasia. *Int J Clin Exp Pathol*. 2008;1(4):306-316.

8. Chari ST, Kelly K, Hollingsworth MA, et al. Early detection of sporadic pancreatic cancer: summative review. *Pancreas*. 2015;44(5):693-712. doi:10.1097/MPA.0000000000000368.

9. Notta F, Chan-Seng-Yue M, Lemire M, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*. 2016;538(7625):378-382. doi:10.1038/nature19823.

10. Wolfgang CL, Herman JM, Laheru DA, et al. Recent progress in pancreatic cancer. *CA Cancer J Clin*. 2013;63(5):318-348. doi:10.3322/caac.21190.

11. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. *Cancer Research*. 2014;74(11):2913-2921. doi:10.1158/0008-5472.CAN-14-0155.

12. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med*. 2011;364(19):1817-1825. doi:10.1056/NEJMoa1011923.

13. Oettle H, Neuhaus P, Hochhaus A, et al. Adjuvant chemotherapy with gemcitabine and long-term outcomes among patients with resected pancreatic cancer: the CONKO-001 randomized trial. *JAMA*. 2013;310(14):1473-1481. doi:10.1001/jama.2013.279201.

14. Neoptolemos JP, Palmer DH, Ghaneh P, et al. Comparison of adjuvant gemcitabine and

capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial. *Lancet*. 2017;389(10073):1011-1024. doi:10.1016/S0140-6736(16)32409-6.

15. Pajic M, Patch A-M, Chang DK, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. 2015;518(7540):495-501. doi:10.1038/nature14169.

16. Moffitt RA, Marayati R, Flate EL, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature Publishing Group*. 2015;47(10):1168-1178. doi:10.1038/ng.3398.

17. Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*. 2016;531(7592):47-52. doi:10.1038/nature16965.

18. Collisson EA, Sadanandam A, Olson P, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med*. 2011;17(4):500-503. doi:10.1038/nm.2344.

19. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nature Publishing Group*. 2014;15(9):585-598. doi:10.1038/nrg3729.

20. Dasari S, Bernard Tchounwou P. Cisplatin in cancer therapy: Molecular mechanisms of action. *European Journal of Pharmacology*. 2014;740:364-378. doi:10.1016/j.ejphar.2014.07.025.

21. Marinus MG. DNA Mismatch Repair. *EcoSal Plus*. 2012;5(1):1-62. doi:10.1128/ecosalplus.7.2.5.

22. Alani E. The Saccharomyces cerevisiae Msh2 and Msh6 proteins form a complex that specifically binds to duplex oligonucleotides containing mismatched DNA base pairs. *Mol Cell Biol*. 1996;16(10):5604-5615.

23. Wu X, Platt JL, Cascalho M. Dimerization of MLH1 and PMS2 limits nuclear localization of MutLalpha. *Mol Cell Biol*. 2003;23(9):3320-3328. doi:10.1128/MCB.23.9.3320-3328.2003.

24. Tran PT, Fey JP, Erdeniz N, Gellon L, Boiteux S, Liskay RM. A mutation in EXO1 defines separable roles in DNA mismatch repair and post-replication repair. *DNA Repair (Amst)*. 2007;6(11):1572-1583. doi:10.1016/j.dnarep.2007.05.004.

25. Krokan HE, Bjørås M. Base excision repair. *Cold Spring Harb Perspect Biol*. 2013;5(4):a012583-a012583. doi:10.1101/cshperspect.a012583.

26. Krokan HE, Nilsen H, Skorpen F, Otterlei M, Slupphaug G. Base excision repair of DNA in mammalian cells. *FEBS Lett*. 2000;476(1-2):73-77.

27. Vermeulen W, de Boer J, Citterio E, et al. Mammalian nucleotide excision repair and

syndromes. *Biochem Soc Trans*. 1997;25(1):309-315.

28. de Boer J, Hoeijmakers JH. Nucleotide excision repair and human syndromes. *Carcinogenesis*. 2000;21(3):453-460.

29. Jasin M, Rothstein R. Repair of strand breaks by homologous recombination. *Cold Spring Harb Perspect Biol*. 2013;5(11):a012740-a012740. doi:10.1101/cshperspect.a012740.

30. Lieber MR, Ma Y, Pannicke U, Schwarz K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol*. 2003;4(9):712-720. doi:10.1038/nrm1202.

31. McKinnon PJ. ATM and ataxia telangiectasia. *EMBO Rep*. 2004;5(8):772-776. doi:10.1038/sj.embor.7400210.

32. Thompson LH. Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: The molecular choreography. *Mutation Research/Reviews in Mutation Research*. 2012;751(2):158-246. doi:10.1016/j.mrrev.2012.06.002.

33. Köcher S, Rieckmann T, Rohaly G, et al. Radiation-induced double-strand breaks require ATM but not Artemis for homologous recombination during S-phase. *Nucleic Acids Research*. 2012;40(17):8336-8347. doi:10.1093/nar/gks604.

34. Bhoumik A, Takahashi S, Breitweiser W, Shiloh Y, Jones N, Ronai Z. ATM-Dependent Phosphorylation of ATF2 Is Required for the DNA Damage Response. *Molecular Cell*. 2005;18(5):577-587. doi:10.1016/j.molcel.2005.04.015.

35. Escribano-Díaz C, Orthwein A, Fradet-Turcotte A, et al. A Cell Cycle-Dependent Regulatory Circuit Composed of 53BP1-RIF1 and BRCA1-CtIP Controls DNA Repair Pathway Choice. *Molecular Cell*. 2013;49(5):872-883. doi:10.1016/j.molcel.2013.01.001.

36. Adams BR, Hawkins AJ, Povirk LF, Valerie K. ATM-independent, high-fidelity nonhomologous end joining predominates in human embryonic stem cells. *Aging (Albany NY)*. 2010;2(9):582-596. doi:10.18632/aging.100197.

37. Hruban RH, Canto MI, Goggins M, Schulick R, Klein AP. Update on familial pancreatic cancer. *Adv Surg*. 2010;44:293-311.

38. Holter S, Borgida A, Dodd A, et al. Germline BRCAMutations in a Large Clinic-Based Cohort of Patients With Pancreatic Adenocarcinoma. *JCO*. 2015;33(28):3124-3129. doi:10.1200/JCO.2014.59.7401.

39. Hu C, Hart SN, Bamlet WR, et al. Prevalence of Pathogenic Mutations in Cancer Predisposition Genes among Pancreatic Cancer Patients. *Cancer Epidemiol Biomarkers Prev*. 2016;25(1):207-211. doi:10.1158/1055-9965.EPI-15-0455.

40. Roberts NJ, Norris AL, Petersen GM, et al. Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer. *Cancer Discovery*. 2016;6(2):166-175. doi:10.1158/2159-8290.CD-15-0402.

41. Smith AL, Alirezaie N, Connor A, et al. Candidate DNA repair susceptibility genes identified by exome sequencing in high-risk pancreatic cancer. *Cancer Letters*. 2016;370(2):302-312. doi:10.1016/j.canlet.2015.10.030.

42. Robson ME, Bradbury AR, Arun B, et al. American Society of Clinical Oncology Policy Statement Update: Genetic and Genomic Testing for Cancer Susceptibility. *J Clin Oncol*. 2015;33(31):3660-3667. doi:10.1200/JCO.2015.63.0996.

43. Syngal S, Brand RE, Church JM, Giardiello FM, Hampel HL, Burt RW. ACG Clinical Guideline: Genetic Testing and Management of Hereditary Gastrointestinal Cancer Syndromes. *The American Journal of Gastroenterology*. 2015;110(2):223-262. doi:10.1038/ajg.2014.435.

44. Galiatsatos P, Foulkes WD. Familial Adenomatous Polyposis. *The American Journal of Gastroenterology*. 2006;101(2):385-398. doi:10.1111/j.1572-0241.2006.00375.x.

45. Elkharwily A, Gottlieb K. The pancreas in familial adenomatous polyposis. *JOP*. 2008;9(1):9-18.

46. Calva D, Howe JR. Hamartomatous polyposis syndromes. *Surg Clin North Am*. 2008;88(4):779–817–vii. doi:10.1016/j.suc.2008.05.002.

47. Riegert-Johnson DL, Boardman LA, Hefferon T, Roberts M, Calva D, Howe J. Juvenile Polyposis. 2009.

48. Howe JR, Mitros FA, Summers RW. The risk of gastrointestinal carcinoma in familial juvenile polyposis. *Ann Surg Oncol*. 1998;5(8):751-756.

49. Brosens LAA, van Hattem A, Hylind LM, et al. Risk of colorectal cancer in juvenile polyposis. *Gut*. 2007;56(7):965-967. doi:10.1136/gut.2006.116913.

50. Walpole IR, Cullity G. Juvenile polyposis: a case with early presentation and death attributable to adenocarcinoma of the pancreas. *Am J Med Genet*. 1989;32(1):1-8. doi:10.1002/ajmg.1320320102.

51. Stemper TJ, Kent TH, Summers RW. Juvenile polyposis and gastrointestinal carcinoma. A study of a kindred. *Ann Intern Med*. 1975;83(5):639-646.

52. Jelsig A, Qvist N, Brusgaard K, Nielsen C, Hansen T, Ousager L. Hamartomatous polyposis syndromes: A review. *Orphanet J Rare Dis*. 2014;9(1):101-110. doi:10.1186/1750-1172-9-101.

53. Jenne DE, Reimann H, Nezu J, et al. Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nat Genet*. 1998;18(1):38-43. doi:10.1038/ng0198-38.

54. McAllister AJ, Hicken NF, Latimer RG, Condon VR. Seventeen patients with Peutz-Jehgers syndrome in four generations. *The American Journal of Surgery*. 1967;114(6):839-843. doi:10.1016/0002-9610(67)90403-5.

55. Kim BG, Park HI, Kwun CS. Peutz-Jehgers syndrome: A case report. *Journal of the Korean Radiological Society*. 1974;10(2):157-161. doi:10.3348/jkrs.1974.10.2.157.

56. Syngal S, Brand RE, Church JM, Giardiello FM, Hampel HL, Burt RW. ACG Clinical Guideline: Genetic Testing and Management of Hereditary Gastrointestinal Cancer Syndromes. *The American Journal of Gastroenterology*. 2015;110(2):223-262. doi:10.1038/ajg.2014.435.

57. Marinus MG. DNA Mismatch Repair. *EcoSal Plus*. 2012;5(1):1-62. doi:10.1128/ecosalplus.7.2.5.

58. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*. 2010;56(2):167-179. doi:10.1111/j.1365-2559.2009.03392.x.

59. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*. 2010;56(2):167-179. doi:10.1111/j.1365-2559.2009.03392.x.

60. Giardiello FM, Allen JI, Axilbund JE, et al. Guidelines on Genetic Evaluation and Management of Lynch Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. July 2014:1-21. doi:10.1038/ajg.2014.186.

61. Kastrinos F, Mukherjee B, Tayob N, et al. Risk of pancreatic cancer in families with Lynch syndrome. *JAMA*. 2009;302(16):1790-1795. doi:10.1001/jama.2009.1529.

62. Lynch HT, Shaw TG. Familial atypical multiple mole melanoma (FAMMM) syndrome: history, genetics, and heterogeneity. *Familial Cancer*. 2016;15(3):487-491. doi:10.1007/s10689-016-9888-2.

63. Riegert-Johnson DL, Boardman LA, Hefferon T, et al. Familial Atypical Multiple Mole Melanoma Syndrome. 2009.

64. Malkin D. Li-Fraumeni Syndrome. *Genes & Cancer*. 2011;2(4):475-484. doi:10.1177/1947601911413466.

65. McBride KA, Ballinger ML, Killick E, et al. Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nature Publishing Group*. 2014;11(5):260-271. doi:10.1038/nrclinonc.2014.41.

66. Leading article Hereditary pancreatitis: new insights into acute and chronic pancreatitis. July 1999:1-6.

67. Lew D, Afghani E, Pandol S. Chronic Pancreatitis: Current Status and Challenges for

Prevention and Treatment. *Dig Dis Sci*. 2017;62(7):1702-1712. doi:10.1007/s10620-017-4602-2.

68. Celeste A Shelton 1 2DCW1-4. Hereditary Pancreatitis. *The Pancreapedia: Exocrine Pancreas Knowledge Base*. doi:10.3998/panc.2016.10.

69. Rosendahl J, Bödeker H, Mössner J, Teich N. Hereditary chronic pancreatitis. *Orphanet J Rare Dis*. 2007;2(1):1. doi:10.1186/1750-1172-2-1.

70. Rosendahl J, Landt O, Bernadova J, et al. CFTR, SPINK1, CTRC and PRSS1 variants in chronic pancreatitis: is the role of mutated CFTR overestimated? *Gut*. 2013;62(4):582-592. doi:10.1136/gutjnl-2011-300645.

71. Lowenfels AB, Maisonneuve P, DiMagno EP, et al. Hereditary pancreatitis and the risk of pancreatic cancer. International Hereditary Pancreatitis Study Group. *J Natl Cancer Inst*. 1997;89(6):442-446.

72. Raimondi S, Lowenfels AB, Morselli-Labate AM, Maisonneuve P, Pezzilli R. Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. *Best Pract Res Clin Gastroenterol*. 2010;24(3):349-358. doi:10.1016/j.bpg.2010.02.007.

73. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nature Publishing Group*. 2016;16(9):599-612. doi:10.1038/nrc.2016.72.

74. Claus EB, Schildkraut JM, Thompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. *Cancer*. 1996;77(11):2318-2324. doi:10.1002/(SICI)1097-0142(19960601)77:11<2318::AID-CNCR21>3.0.CO;2-Z.

75. Claus EB, Schildkraut J, Iversen ES, Berry D, Parmigiani G. Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *J Natl Cancer Inst*. 1998;90(23):1824-1829.

76. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nature Publishing Group*. 2016;16(9):599-612. doi:10.1038/nrc.2016.72.

77. Fernandes PH, Saam J, Peterson J, et al. Comprehensive sequencing of PALB2 in patients with breast cancer suggests PALB2 mutations explain a subset of hereditary breast cancer. *Cancer*. 2014;120(7):963-967. doi:10.1002/cncr.28504.

78. Pylkas K, Tommiska J, Syrjakoski K, et al. Evaluation of the role of Finnish ataxia-telangiectasia mutations in hereditary predisposition to breast cancer. *Carcinogenesis*. 2006;28(5):1040-1045. doi:10.1093/carcin/bgl237.

79. Goldgar DE, Healey S, Dowty JG, et al. Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res*. 2011;13(4):R73. doi:10.1186/bcr2919.

80. Graña B, Fachal L, Darder E, et al. Germline ATM mutational analysis in BRCA1/BRCA2 negative hereditary breast cancer families by MALDI-TOF mass spectrometry. *Breast Cancer Res Treat*. 2011;128(2):573-579. doi:10.1007/s10549-011-1462-x.

81. Olsen JH, Hahnemann JMD, Børresen-Dale A-L, et al. Breast and other cancers in 1445 blood relatives of 75 Nordic patients with ataxia telangiectasia. *Br J Cancer*. 2005;93(2):260-265. doi:10.1038/sj.bjc.6602658.

82. Marabelli M, Cheng S-C, Parmigiani G. Penetrance of ATMGene Mutations in Breast Cancer: A Meta-Analysis of Different Measures of Risk. *Genet Epidemiol*. 2016;40(5):425-431. doi:10.1002/gepi.21971.

83. Swift Michael RPJMDCCL. Breast and Other Cancers in Families with Ataxia-Telangiectasia. November 2008:1-6.

84. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nature Publishing Group*. 2014;15(9):585-598. doi:10.1038/nrg3729.

85. Hofstatter EW, Domchek SM, Miron A, et al. PALB2 mutations in familial breast and pancreatic cancer. *Familial Cancer*. 2011;10(2):225-231. doi:10.1007/s10689-011-9426-1.

86. Zhen DB, Rabe KG, Gallinger S, et al. BRCA1, BRCA2, PALB2, and CDKN2A mutations in familial pancreatic cancer: a PACGENE study. *Genet Med*. 2014;17(7):569-577. doi:10.1038/gim.2014.153.

87. Roberts NJ, Jiao Y, Yu J, et al. ATMMutations in Patients with Hereditary Pancreatic Cancer. *Cancer Discovery*. 2012;2(1):41-46. doi:10.1158/2159-8290.CD-11-0194.

88. Bakker JL, de Winter JP. A Role for ATMin Hereditary Pancreatic Cancer. *Cancer Discovery*. 2012;2(1):14-15. doi:10.1158/2159-8290.CD-11-0318.

89. Diamond JM, Rotter JI. Observing the founder effect in human evolution. *Nature*. September 1987:105-106.

90. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc Natl Acad Sci USA*. 2010;107(37):16222-16227. doi:10.1073/pnas.1004381107.

91. Tonin PN, Mes-Masson AM, Futreal PA, et al. Founder BRCA1 and BRCA2 mutations in French Canadian breast and ovarian cancer families. *The American Journal of Human Genetics*. 1998;63(5):1341-1351. doi:10.1086/302099.

92. Scriver CR. Human genetics: lessons from Quebec populations. *Annu Rev Genomics Hum Genet*. 2001;2(1):69-101. doi:10.1146/annurev.genom.2.1.69.

93. Ghadirian P, Robidoux A, Zhang P, et al. The contribution of founder mutations to

early-onset breast cancer in French-Canadian women. *Clin Genet*. 2009;76(5):421-426. doi:10.1111/j.1399-0004.2009.01277.x.

94. Tischkowitz M, Sabbaghian N, Hamel N, et al. Contribution of the PALB2 c.2323C>T [p.Q775X] Founder mutation in well-defined breast and/or ovarian cancer families and unselected ovarian cancer cases of French Canadian descent. *BMC Medical Genetics*. 2013;14(1):1-1. doi:10.1186/1471-2350-14-5.

95. Spoerri L, Brooks K, Chia K, et al. A novel ATM-dependent checkpoint defect distinct from loss of function mutation promotes genomic instability in melanoma. *Pigment Cell Melanoma Res*. 2016;29(3):329-339. doi:10.1111/pcmr.12466.

96. Rothblum-Oviatt C, Wright J, Lefton-Greif MA, McGrath-Morrow SA, Crawford TO, Lederman HM. Ataxia telangiectasia: a review. *Orphanet J Rare Dis*. 2016;11(1):159. doi:10.1186/s13023-016-0543-7.

97. Gilardini Montani MS, Prodosmo A, Stagni V, et al. ATM-depletion in breast cancer cells confers sensitivity to PARP inhibition. *J Exp Clin Cancer Res*. 2013;32(1):95. doi:10.1186/1756-9966-32-95.

98. Eberle MA, Pfützer R, Pogue-Geile KL, et al. A new susceptibility locus for autosomal dominant pancreatic cancer maps to chromosome 4q32-34. *The American Journal of Human Genetics*. 2002;70(4):1044-1048. doi:10.1086/339692.

99. Pogue-Geile KL, Chen R, Bronner MP, et al. Palladin mutation causes familial pancreatic cancer and suggests a new cancer mechanism. Tuveson D, ed. *PLoS Med*. 2006;3(12):e516. doi:10.1371/journal.pmed.0030516.

100. Zogopoulos G, Rothenmund H, Eppel A, et al. The P239S palladin variant does not account for a significant fraction of hereditary or early onset pancreas cancer. *Hum Genet*. 2007;121(5):635-637. doi:10.1007/s00439-007-0361-z.

101. Klein AP, de Andrade M, Hruban RH, et al. Linkage analysis of chromosome 4 in families with familial pancreatic cancer. *Cancer Biol Ther*. 2007;6(3):320-323.

102. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature Publishing Group*. 2009;41(9):986-990. doi:10.1038/ng.429.

103. Lachaud C, Moreno A, Marchesi F, Toth R, Blow JJ, Rouse J. Ubiquitinated Fancd2 recruits Fan1 to stalled replication forks to prevent genome instability. *Science*. 2016;351(6275):846-849. doi:10.1126/science.aad5634.

104. Hofstatter EW, Domchek SM, Miron A, et al. PALB2 mutations in familial breast and pancreatic cancer. *Familial Cancer*. 2011;10(2):225-231. doi:10.1007/s10689-011-9426-1.

105. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at

loci with multiple signals of association. *Genetics*. 2014;198(2):497-508. doi:10.1534/genetics.114.167908.

106. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91(2):224-237. doi:10.1016/j.ajhg.2012.06.007.

107. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*. 2014;95(1):5-23. doi:10.1016/j.ajhg.2014.06.009.

108. Peralta JM, Almeida M, Kent JW Jr, Blangero J. A variance component-based gene burden test. *BMC Proceedings*. 2014;8(Suppl 1):S49. doi:10.1186/1753-6561-8-S1-S49.

109. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029.

110. Moutsianas L, Agarwala V, Fuchsberger C, et al. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. Ripatti S, ed. *PLoS Genet*. 2015;11(4):e1005165–24. doi:10.1371/journal.pgen.1005165.

111. Sun J, Zheng Y, Hsu L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet Epidemiol*. 2013;37(4):334-344. doi:10.1002/gepi.21717.

112. Grant RC, Denroche RE, Borgida A, et al. Exome-Wide Association Study Of Pancreatic Cancer Risk. *Gastroenterology*. October 2017. doi:10.1053/j.gastro.2017.10.015.

113. Zheng HF, Forgetta V, Hsu YH, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*. 2015;526(7571):112-117. doi:10.1038/nature14878.

114. Klein AP, Brune KA, Petersen GM, et al. Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. *Cancer Research*. 2004;64(7):2634-2638.

115. Smith AL, Bascuñana C, Hall A, et al. Establishing a clinic-based pancreatic cancer and periampullary tumour research registry in Quebec. *Curr Oncol*. 2015;22(2):113-121. doi:10.3747/co.22.2300.

116. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.

117. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.

118. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010;38(16):e164-e164. doi:10.1093/nar/gkq603.

119. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Publishing Group*. 2014;46(3):310-315. doi:10.1038/ng.2892.

120. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847.

121. Dennis G, Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003;4(5):P3.

122. Carbon S, Ireland A, Mungall CJ, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25(2):288-289. doi:10.1093/bioinformatics/btn615.

123. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Research*. 2010;38(Database issue):D331-D335. doi:10.1093/nar/gkp1018.

124. Ruark E, Snape K, Humburg P, et al. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*. 2012;493(7432):406-410. doi:10.1038/nature11725.

125. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*. 2009;37(9):e67-e67. doi:10.1093/nar/gkp215.

126. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394. doi:10.1089/1066527041410418.

127. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39(2):175-191.

128. Epstein MP, Duncan R, Ware EB, et al. A Statistical Approach for Rare-Variant Association Testing in Affected Sibships. *The American Journal of Human Genetics*. 2015;96(4):543-554. doi:10.1016/j.ajhg.2015.01.020.

129. Cheon JY, Mozersky J, Cook-Deegan R. Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Med*. 2014;6(12):121. doi:10.1186/s13073-014-0121-3.

130. Gubanova E, Issaeva N, Gokturk C, Djureinovic T, Helleday T. SMG-1 suppresses

CDK2 and tumor growth by regulating both the p53 and Cdc25A signaling pathways. *Cell Cycle*. 2014;12(24):3770-3780. doi:10.4161/cc.26660.

131. Roberts TL, Ho U, Luff J, et al. Smg1 haploinsufficiency predisposes to tumor formation and inflammation. *Proc Natl Acad Sci USA*. 2013;110(4):E285-E294. doi:10.1073/pnas.1215696110.

132. Gehen SC, Staversky RJ, Bambara RA, Keng PC, O'Reilly MA. hSMG-1 and ATM sequentially and independently regulate the G1 checkpoint during oxidative stress. *Oncogene*. 2008;27(29):4065-4074. doi:10.1038/onc.2008.48.

133. Morita T, Yamashita A, Kashima I, Ogata K, Ishiura S, Ohno S. Distant N- and C-terminal domains are required for intrinsic kinase activity of SMG-1, a critical component of nonsense-mediated mRNA decay. *J Biol Chem*. 2007;282(11):7799-7808. doi:10.1074/jbc.M610159200.