ASSOCIATION OF DNA REPAIR GENE VARIANTS WITH PANCREATIC CANCER IN PATIENTS WITH CHRONIC PANCREATITIS

by

Siddhartha Das

B.Tech. Biotechnology, West Bengal University of Technology, India, 2009

Submitted to the Graduate Faculty of

the Human Genetics Department

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Siddhartha Das

It was defended on

July11, 2013

and approved by

Thesis Advisor: David Whitcomb, M.D., Ph.D., Professor, Department of Medicine, School of Medicine, University of Pittsburgh

Robert Ferrell, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

John R Shaffer, Ph. D., Research Assistant Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

David Whitcomb, M.D., Ph.D

ASSOCIATION OF DNA REPAIR GENE VARIANTS WITH PANCREATIC CANCER IN PATIENTS WITH CHRONIC PANCREATITIS

Siddhartha Das, MS

University of Pittsburgh, 2013

ABSTRACT

Background: Pancreatic cancer (PC) is one of the most devastating cancers with less than 5% surviving after five years of diagnosis. Most of the risk factors have non-specific odds ratio except for chronic pancreatitis (CP) which has an extremely high odds ratio as reported in the literature. CP is characterized by inappropriate activation of trypsinogen in the pancreas resulting in inflammation of the pancreas. Most of the genetic factors behind the inflammatory to cancerous progression still remain unexplained. This research study describes identification of multiple non-synonymous mutations and implicates specific DNA repair pathways in the risk of progression from CP to PC.

Methods: Whole exome sequencing was carried out in 16 CP individuals with PC and 11 individuals without PC, following which the thousand genomes project data was used to identify the rare and novel germline non-synonymous variants among 159 DNA repair genes and burden test of DNA repair pathways was used to identify the most frequently mutated DNA repair pathways.

Results: We were able to able to identify at least 30 rare and novel non-synonymous germline variants at sufficient read depth in our CP+PC cohort that warrant investigation in pancreatic cancer tumor tissues as well as larger PC patient cohorts.

Public Health Significance: The public health significance of this work lies in the fact that it provides for the first time an opportunity to genetically screen the potentially high risk pancreatic cancer patient cohorts to determine their individual risk of development of the disease and based on risk assessment, a strategy could be developed to determine if an individual needs to undergo high risk surgical procedures like total pancreatectomy to reduce their risk of developing pancreatic cancer or develop changes in their habits to reduce the possibility of the development of pancreatic cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

At the beginning of my dissertation I would like to express my sincere appreciation to all who made my research work successful. My sincere gratitude goes to my MS committee chair and mentor Dr. David C Whitcomb, for his continuous encouragement and other committee members Dr. Eleanor Feingold, Dr. Robert Ferrell and Dr. John Shaffer for their efforts in making this thesis a success. I was enriched by their wise guidance throughout my research work.

I would like to thank Dr. Michael Barmada who's preliminary analysis of the exome sequencing data was most critical in the successful completion of thesis and his unwavering patience to answer my queries from time to time was one the key ingredients behind the success of this thesis.

I would like to thank my fellow lab mates Jessica Larusch, Kimberly Stello, Nijole Pollock, Danielle Dwyer and Alexander Rowland for their untiring efforts in keeping the lab running and preparation of DNA samples for sequencing.

I would also like to thank all my friends in Pittsburgh for the wonderful time they gave me and last but not the least I am deeply grateful to my family, especially my parents without whose constant encouragement and blessings, this thesis would not have been successful.

Finally I would like to acknowledge the Wayne Fusaro Pancreatic Cancer Research Fund (Dr Whitcomb), RO1 DK061451 from the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK, Dr Whitcomb) and UL1 RR024153 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research, through the Clinical and Translational Science Institute and Competitive Medical Research Fund Genomics Pilot Program (C2GP2) funding using of the University of Pittsburgh Genomics and Proteomics Core Laboratory.

1.0 PANCREATIC CANCER

1.1 INTRODUCTION

Pancreatic cancer (PC) is a highly malignant disease of the pancreas. It is usually highly metastasized by the time it is diagnosed. Thus it is associated with extremely poor quality of life and has very low one year and five year survival rates. The National Cancer Institute considers that pancreatic cancer is estimated to account for 45220 newly diagnosed cases of cancer out of total 1283788 estimated cancer cases in 2013 in USA which makes it the least frequent kind of major cancer among all the common cancers. (National Cancer Institute, 2013) However with an estimated death number at 38460, an alarming 85% are estimated to die within a year thus putting PC among the top brackets of disease specific mortality rates among all cancers.

The most common mode of treatment is administration of gemcitabine, a nucleoside analogue that was found to have significant improvement in median overall survival as compared to fluoricil administration. However it increases survival by only 0.8 months (Di Marco M et al. 2010) and thus further genetic studies are warranted to identify suitable therapeutic targets to PC. Further subsequent phase 3 trials of gemcitabine as a single agent ranged only from 5 to 7.2 months and the combination of gemcitabine and other cytotoxic and targeted agents showed significant survival advantage as compared with gemcitabine alone but was also associated with increased toxicity. It was also concluded from this study that if response to drug was stable, then chemotherapy could be discontinued which often is known to have damaging side-effects in cancer individuals. (Conroy T et al. 2011) Hence surgery remains the only other treatment option that offers an advantage in terms of 5 year overall survival. Thus developing preventive measures against development of PC continues to be a top public health necessity.

1.2 EPIDEMIOLOGY

With an estimated 45220 new cases of PC in 2013, the incidence rate of pancreatic cancer is relatively low and would have been of little concern except for its high mortality rate. This high mortality rate places PC as the fourth or fifth most frequent cause of cancer death in most developed countries. (Ferlay J et al. 2010)

PC has a peculiar trend with respect to geographical variations in that the rates are 3 to 4 times higher in northern countries far away from the equator like Iceland, Finland or northern USA while countries close to the equator such as Egypt or India have much lower rates. Of the several reasons suggested for this phenomenon, the most common is believed to be related to age. Since the incidence of PC is strongly related to age, improving life span in the developed countries that will give to rise to increased population will give rise to increasing incidence of PC thus highlighting its public health importance. However, improvements in lifespan in the general population mean that the absolute frequency of pancreatic cancer is likely to rise in countries like China, India and other Asian regions that have large aging populations confirming its status as a disease of major public health concern. (Maisonneuve P et al. 2010)

The following tables below shows the epidemiology data by NCI's SEER Cancer Statistics Review.

Table1.1: Median age of death and diagnosis stratified by race for PC in US for 2006-2010

	All races			Whites			Blacks		
	Total	Male	Female	Total	Male	Female	Total	Male	Female
PC median age of death	73	70	75	73	71	76	69	66	72
PC median age of diagnosis	71	69	74	72	69	74	67	64	70

Table 1.2: Incidence rate of PC stratified by sex and race for PC in US for 2006-2010

Type of race	Male (per 1,00,000)	Female (per 1,00,000)
All	13.9	10.9
White	13.8	10.7
Black	17.6	14.3

Table 1.3: Mortality rate of PC stratified by sex and race for PC in US for 2006-2010

Type of race	Male (per 1,00,000)	Female (per 1,00,000)
All	12.5	9.6
White	12.5	9.4
Black	15.3	12.5

The above data displays the two grim aspects of reality with respect to PC. Table1 shows that PC is detected at a relatively late age with time from diagnosis and death being 2 years at the maximum and comparing table 2 and 3 data the incidence rate and mortality rate are very close implying current treatment measures in terms of surgery or drug is not effective for PC individuals. Thus developing effective methods for determining who is at risk of the disease and then devising methods for preventing development of those risk factors remains the most effective way to treat PC.

1.3 RISK FACTORS

Pancreatic cancer is a highly heterogeneous disorder with the exact cause still unknown which possibly explains the difficulty of its clinical management and the high mortality rates. Several environmental factors have been implicated with causative evidence mainly for tobacco use with an association between smoking and pancreatic cancer showed by most published studies. (Iodice S et al. 2008).However most of them have relatively small odds ratio for PC that are not effective enough to be targeted for an intervention thus minimizing their role as putative screening factor for determining PC risk. Table 1.4 includes a short list of factors with the risk that they are known to possess in PC.

Table 1.4 Major non-genetic risk factors of Pancreatic Cancer

	Risk factor	Risk, CI*	Reference
Environment – lifestyle			
	Smoking	1.74, CI 1.61-1.87	(Iodice S et al. 2008)
	Alcohol	possible	(Duell J et al. 2012)
Occupational			
	Chlorinated hydrocarbons	1.4-4.4	(Andreotti G et al. 2012)
	Polycyclic aromatic	1.5	(Andreotti G et al. 2012)
Diet			
	n-nitroso foods	1.27 CI 1.09-1.48	(Risch HA et al. 2012)
	saturated fat / animal fat	~1.2 – 2.0	(Sanchez GV et al. 2012)
Medical conditions			
	Pancreatitis	5.1 CI 3.5-7.3	(Raimondi S et al. 2012)
	Chronic pancreatitis	13.3 CI 6.1-28.9	(Raimondi S et al. 2012)
	Hereditary pancreatitis	69.9 CI 56.4-84.4	(Raimondi S et al. 2012)
	Allergies	0.39-0.77	(Olson SH 2012)
	Diabetes Mellitus	1.5-2.0	(Li D 2012)
	Obesity	~1.1-1.3	(Bracci PM 2012)
	ABO blood group	1.65 CI 1.30-2.09	(Risch HA et al. 2012)

Despite the environmental risk factors in most cases having non-specific odds ratios, relatively rare medical conditions such as chronic pancreatitis and hereditary pancreatitis have been associated with the highest odds ratios of risk. The primary pathway between inflammation and cancer remains an area of active research and controversy. (Whitcomb DC 2004) It is hypothesized that the mechanism behind such elevated risk of PC in pre-existing cases of pancreatitis is related to the continuous exposure of duct cells to reactive oxygen species and other toxic inflammatory factors that can cause DNA damage and thus increased cell turnover (replacement of old cells with newly generated ones from existing ones) that facilitates clonal expansion of early metaplastic cells into metastatic PC. However with no more than 5% of diagnosed cases of pancreatic cancer being explained by recurrent attacks of chronic pancreatitis (Raimondi S et al. 2010) it is unclear which are the high risk CP individuals that are most likely to progress to PC.

1.4 CONUNDRUM-WHICH CP CASES WILL PROGRESS ON TO PC?

The link between chronic inflammation and tumorigenesis has been recognized in several disorders including colorectal cancer after inflammatory bowel disease or bladder cancer after schistosomasis as well as pancreatic ductal adenocarcinoma after chronic pancreatitis. Severe inflammation during chronic inflammatory diseases exposes the tissues of the organ to the cytotoxic agents such as proinflammatory cytokines and reactive oxygen species. Their presence leads to activation of cellular protective mechanisms, such as cell death and acinar to ductal metaplasia and increased proliferation of replacement cells aimed at organ regeneration. Increased cell turnover in an environment rich in oxidative species favors accumulation of DNA damage thus increasing chances of positive selection for mutations that confer growth advantage.

Thus of all the PC risk factors, CP and more specifically HP are by far the largest risk factors in terms of relative risk. A meta-analysis by Raimondi et. al. (Raimondi S et al. 2010) published in 2010 showed that there was a statistically significant increase in the risk of pancreatic cancer risk for all the major type of pancreatitis, with summary RRs (95%CI) of 5.1 (3.5–7.3) for unspecified pancreatitis, 13.3 (6.1–28.9) for chronic pancreatitis and 69.0 (56.4–84.4) for hereditary pancreatitis.

Genetic studies addressing this issue have been few and in one of the earliest those studies by Moskovitz et. al. (Moskovitz et al. 2003) found an increased level of chromosomal abnormality in cell cultures of normal appearing, non-neoplastic pancreatic epithelia in both patients with chronic pancreatitis and patients with pancreatic cancer relative to normal donor-derived cells but there was no significant difference in chromosomal abnormalities in the ductal cells from patients with pancreatitis as compared to the cancer patients. Later Yan et. al. (Yan L et al. 2005) reported on the investigation of pancreatic juices from 146 patients with PDAC or CP or biliary tract stones and found that mutant p53 were present in a greater number of PC patients as compared to CP individuals and more recently in 2010, Baumgart et. al. (Baumgert M et al. 2010) reported that in the early PanIN of CP patients, there were heterozygous mutations of p53 and p16 along with chromosomal instability early in CP. With the vital role of the DNA repair system in maintaining the stability of the genome by preventing accumulation of mutations and thus preventing chromosomal instability, it could be argued that accumulation of mutations in DNA repair genes in a cancerous patient is interfering with their normal functioning that is preventing the repair of the DNA damage occurring in other critical cell cycle regulation genes and thus giving rise to loss of control of cell cycle regulation and a potential tumorigenic phenotype.

Although these studies have looked at specific genes to be involved in the risk of PC in CP individuals, they have confirmed the role of genetic factors in the progression of chronic pancreatitis to pancreatic cancer.

2.0 GENETICS OF PANCREATIC CANCER

Myriad of genetic pathways have been shown to be involved in causation of PC that could be involved in cell growth regulation or cell proliferation. In a landmark study by Jones et.al. (Jones S et al. 2008) where all the exons from tumor DNA of 20661 protein coding genes were sequenced from 24 advanced pancreatic adenocarcinoma patients, multiple genetic pathways including 9 DNA repair genes were found to be mutated more than once with *TP53*, *ERCC4* and *ERCC6* being most frequent genes carrying deleterious mutations in about 83% of the cases thus giving rise to the question that are critical DNA repair gene mutation vital to the PC pathogenesis process? Given the role of the DNA repair system as the guardian of the genome, it is expected that the various DNA repair mechanisms would act to repair the damaged DNA and thus prevent the increased cell turnover by preventing them from acquiring a mutator phenotype. However if there are defects in the DNA repair mechanism, they will allow the accumulation of damaged DNA and thus facilitate the transition from an inflammatory to a metastatic process by improper regulation of signal transduction, embryonic signaling and related pathways.

Thus multiple DNA repair based studies of PC has been undertaken in the past to study the functional effects of genetic variations or to look at how genetic variations in various DNA repair genes might interact with environmental factors to influence PC risk or how they might affect overall survival or overall outcome in PC cases.

2.1 REVIEW OF DNA REPAIR STUDIES IN PANCREATIC CANCER

A pubmed search of the keywords “DNA repair” and “pancreatic cancer” in the last 10 years (2003-2012) yielded 213 articles of which 19 studies have been genetic epidemiology based looking at role of DNA repair genes in PC either as genetic variations that interact with dietary intake factors to modulate risk of PC or if SNPs were associated with increased risk of PC, overall survival or simply as a potential marker of tumor response to therapeutic measures. SNP based studies have also looked at potential interaction with environmental risk factors like smoking and diabetes or simply interactions among SNPs or genes. Studies looking at multiple SNPs in high LD in a particular gene have also looked at haplotypes as potential indicator of overall survival or simply as risk factor for PC.

All the four traditional DNA repair pathways (base excision repair, nucleotide excision repair, mismatch repair and double strand break repair) have been studied in pancreatic cancer via genetic association studies. In probably the largest of such studies till date by McWilliams et al. (McWilliams RR et al. 2009), 236 tag SNPs among all NER genes were investigated in 1143 PC cases and 1097 healthy controls and *MMS19L* was the gene that appeared to be significant in both the corrected and uncorrected analysis. However this study had some major limitations in that there was a genotyping failure in 5% of the samples that could significantly affect the power and results of this study. Further there was no correction for multiple testing which could typically overestimate the significance of the SNPs tested for.
In the same year, another large case control study from Li et al. (Li D et al. 2009) at the M.D. Anderson Cancer Center investigated 9 previous clinically investigated SNPs of the genes *hOGG1*, *LIG3*, *LIG4*, *POLB*, *ATM*, *RAD54L* and *RECQL* on risk of PC and reported that the *LIG3* p.G39A and *ATM* p.D1853N SNP homozygote variant had significant interaction with diabetes on risk of PC. Although this was the first well designed study to show the possible involvement of genes involved in repair of DNA strand breaks or cellular response to DNA damage in risk of PC, it had low frequency of homozygous mutants and interaction of genotype with other risk factors thus leaving the requirement of a larger population based study to validate these findings. Further the authors pointed out that some observations could be false discovery associated with multiple testing issues plus the limited number of genes and SNPs studied restricts the usefulness of their findings. Besides, lack of studies exploring the functional significance of these SNPs also limits the significance of these findings.

Moreover, shortcomings such as those reported by Li et al. including heterogeneity of patient group which can result in different penetration of a variant as per ethnic differences can also result in over-estimation or under-estimation of the true effect of a variant such as the fact that this was the first study that showed that BER variants had an effect on PC outcome. Thus the relatively low frequency of PC and heterogeneous nature of the studies also limits the potential usage of the association study findings as screening tools or as determiners of therapeutic outcome.

Although these studies have advanced the knowledge of the involvement of DNA repair pathways in PC, the rapid progress of PC to metastasis from initial Pan-IN lesions coupled with the fact that genetic association does not imply causation, has resulted in no significant therapeutic or preventive measures based on the genetic profile of the DNA repair genes in an individual.

Furthermore pancreatic cancer in general is a complex disease with multiple potential pathways being disrupted and multiple genetic variants could be simultaneously acting to give rise to genomic instability that is characteristic of cancers. This is a critical factor not addressed by the previous genetic association studies which has attempted to implicate single variants in the causation of the disease. Besides, the extremely poor prognosis of PC where the 6 month survival rate is less than 5% as reported

by literature limits the recruitment of large patient population groups from multiple centers to address the issues of small sample size, lack of replication set, heterogeneity of patients and potential for false positive comparisons due to multiple comparisons.

In order to fulfill the deficiency left by genetic association studies, functional studies that rely on cell culture and animal models and other in vitro techniques to explain how disruption of a normal physiologic process is involved in disease causation have been undertaken and several of these have been landmark papers in their contribution to the understanding of role of DNA repair genes in PC.

In one of the very first of these studies by Crnogorac-Jarcevic (Crnogorac-Jurcevic et al. 2002) comparing pure neoplastic and normal duct cells of the pancreas using cDNA microarray, the base excision repair gene *XRCC1* was found to be significantly downregulated in PC tissues as compared to normal tissues. Further they detected reduced immunoreactivity in 75% of cancer cells for *XRCC1* as compared to a strong nuclear and cytoplasmic immunoreactivity of non-neoplastic pancreatic epithelial cells. The findings of this study were supported from another study by Mathews et. al. in 2011 (Mathews LA et al. 2011) who sought to investigate the gene expression profile of highly invasive PC cells using an in-vitro assay and found that invasive PC cells have significantly elevated levels of DNA repair genes and they hypothesized that it was perhaps due to greater accumulation of genomic changes in the highly invasive cells. However traditional cDNA microarray studies suffer from several limitations including low density compared to oligonucleotide arrays, presence of repetitive and common sequences from gene families that would be present in all cDNA from a particular family of genes thus giving rise to potential for cross hybridizations and false signals. Thus these findings need to be followed up by further expression studies in PC cell lines to add to their validity.

Genomic instability has always been found to be reported as the root cause of cancers and thus studies addressing repair potential of cancerous cell line for specific genomic defects have always gained attention. In one of the very first of the studies by Maple et. al. (Maple JT et al. 2005**)** assessing 35 long term survivors of PC ruled out defective MMR as being common in PC and as potential survival benefiter of those with sporadic cancer. However majority of patients in this subset of long-term survivors had small tumors, negative surgical margins, and adjuvant chemoradiation all of which could be potential confounding factors limiting the findings of this study. Further Nyaga et. al. (Nyagi SG et al. 2008) analyzing the effectiveness of a PC cell line to repair 8-hydroxyguanine relative to a non-malignant cell line reported that there was significant down regulation of the BER pathway gene *hOGG1* protein and mRNA level compared to the control leading to accumulation of 8-hydroxyguanine thus overwhelming the DNA repair genes leading to increased probability of deleterious mutations accumulating and hence risk of PC. Indeed one of the most vital BER genes, *XRCC1,* has repeatedly been detected as a predictor of survival in PC patients either those treated by platinum and fluoropyrimidine or as overall determiner

of survival in those with resectable pancreatic adenocarcinoma albeit with conflicting findings. Akita et. al. in 2009 (Akita H et al. 2009) attempting to identify relationship between protein expressions and clinical outcomes in PC patients who underwent surgery, including response to gemcitabine at the time of disease recurrence reported that patients with high *ERCC1* had a trend of better overall survival and in combination with *RRM1* had better overall and disease free survival but in a recent study Maithel et. al. (Maithel SK et al. 2011) from their immunohistochemistry results of 95 patients who underwent pancreaticoduodenectomy reported that high *ERCC1* was associated with both reduced recurrence free survival and overall survival after resection. Unless there are more studies addressing the role of *ERCC1* in survival of PC patients, these conflicting results limit their utility. Thus contradictory results along with the fact that functional studies in vitro cannot exactly replicate the events occurring in vivo limit the findings of functional studies as a potential for better diagnosis or treatment options for PC

Despite significant contribution to understanding the role of DNA repair defects in PC, single variant association studies have suffered from their problems including lack of replication, multiple testing and insignificant information regarding effects of single variants on PC causation that were tested using genetic association study and limited contribution of functional studies to pancreatic cancer treatment. With a hope to fulfill the deficiencies left by single variant association and functional studies, genome wide association studies that focus on genotyping multiple tag SNPs all throught the genome came to the fore.

2.2 GENOME WIDE ASSOCIATION STUDIES AND PANCREATIC CANCER

With the advent of advanced genotyping technology, genome-wide association studies that attempt to find multiple tag SNPs and possible suspect genes have become very popular for studying the genetic factors for complex diseases and multiple PC GWAS studies have been done since the inception of this technology.

The first PC GWAS study reported in 2009 by Amundadottir et. al. (Amundadottir et. al. 2009) in a Caucasian population found an increased susceptibility between a SNP in the intronic region of ABO blood group gene and PC risk suggesting that people with blood group O may have a lower risk of pancreatic cancer than those with groups A or B. However this finding has not been replicated in any other GWAS study despite being previously associated in other non-genome wide association studies. Indeed a key feature of all the genome wide association studies in PC is the lack of replication of the

findings across the studies or the lack of common genes across the studies suggesting that either ethnicity is a major factor determining PC risk or that PC is a highly complex disease with multiple genes and the risk posed by each gene for the disease depends on the population being studied. Further relatively limited knowledge regarding the role of highly significant SNPs in the non-coding regions of the genome limits the significance of the findings in terms of having a clinical relevance. Recent failure of replication studies of a Chinese and a Japanese GWAS in a European population such as those by reported by Campa et. al. (Campa et. al. 2009) indicates that fairly common variants despite high levels of significance might have arisen just by chance in GWAS studies as well as there is sufficient heterogeneity between population studied for any common gene to possess risk across any population. Further the findings of variants in non-coding region with odds ratios ranging from 1 to 2 makes it fairly difficult to identify how significant the role is of that particular SNP or gene in the disease pathogenesis or what is its exact role.
Finally one of the key issues which genome wide association studies that attempt to find common variants involved with the disease fails to address is that of missing heritability. The common variants are not able to account for all the genotypic heritable traits that combine with the environmental effects to give rise to a particular phenotype. Thus despite multiple genes being found to be associated with PC pathogenesis in the last couple of years via genome wide significant studies, the findings have not been of sufficient clinical relevance and next generation sequencing studies and more specifically exome sequencing studies that can identify rare deleterious non-synonymous variants have currently emerged as one of the predominant modes of studying cancer genomics.

2.3 NEXT GENERATION SEQUENCING STUDIES

The advances in next generation sequencing have helped to make it possible to replace laborious techniques such as positional cloning to identify genetic changes in both rare and common diseases. With the significantly large amount of data generated by next generation sequencing coupled with the computational burden of analyzing it, whole exome sequencing has taken precedence over whole genome sequencing to identify pathogenic genetic factors for various diseases. The whole exome platform consists of all the exons of a genome that are transcribed into mature mRNA. A major reason to focus on the exome as compared to the whole genome other than the cost efficiency is the fact that even though the exome constitutes only 1% of the genome, it consists of about 85% of the mutations that affect disease traits. (Bamshad MJ et al. 2011) Following is the standard workflow often adopted in exome sequencing studies:

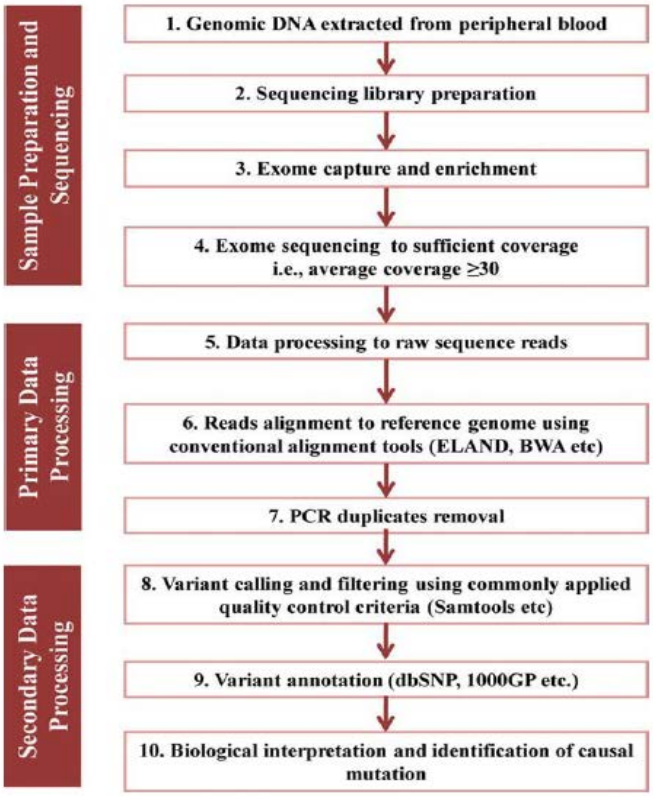

Figure 2.1: General Work flow of exome sequencing studies from isolation of DNA till identification of potential disease causing mutations. The workflow has three components: a) Sample Preparation and Sequencing, b) Primary Data Processing, c) Secondary Data Processing.
Figure directly adapted from Chee- Seng Ku et al. 2012

Although exome sequencing has been significantly successful in identifying genetic variations that might disrupt normal physiological process and hence be involved in disease causation especially rare Mendelian diseases or diseases within a family, the concept of being able to capture the coding regions of the genome have also spawned cancer genome based studies to identify the mutation profile of cancer patients. Following the increased efficiency and resolution of next generation sequencing that facilitates detection of genetic and genomic alterations such as mutations, insertions, deletions, as well chromosomal rearrangements and copy number variations, comprehensive analysis of cancer genome through these studies have significantly increased our understanding of the challenges in cancer biology, diagnosis and therapy.

2.3.1 EXOME SEQUENCING

Exome sequencing is the process of targeted resequencing of the protein coding regions of the human genome to identify disease causing mutations. Sample is prepared by using exome capturing arrays or libraries which help in isolating and enriching the DNA template to be analyzed. Following sequencing, quality estimation allows the evaluation of each analyzed base and the reads generated are mapped onto the reference genome. In order to gain first insights into library preparation and sequencing efficiency, filtering steps are required to determine the percentage of sequence reads which do not originate from protein coding regions or could not be aligned at all. Subsequently, variant detection algorithms obtain a set of genome positions where the analyzed sample differs significantly from the reference. Since these call sets contain numerous non-biologically based variations, further filtering steps are applied to increase the number of true biological variants. Exome sequencing contains the following major steps in its workflow:

2.3.1.1 Base/color calling quality assignment-In order to determine the accuracy of the base called, a certain quality score is assigned known as phred scale quality values for each base or color call. Initially introduced by the base-calling program Phred, it links error probabilities logarithmically to a base or color call. It is defined as:

$$qphred = -10 \cdot \log10\,(p)$$

where p is the estimated error probability for that call.

2.3.1.2 Alignment- Alignment can be described as the process of determining the most likely source within the genome sequence for the observed DNA sequence read, given the knowledge of which species the sequence came from.

Traditional alignment algorithms such as BLAST or BLAT do not scale well with NGS reads in terms of processing time, mapping accuracy and memory use. Thus new alignments for aligning short reads to the reference genome have been developed specifically for this purpose.

These programs generally use a strategy where heuristic techniques first identify a small subset of reference genome where the read is most likely to align and then a slower and more accurate algorithm is used to determine exact position of sequence read.

Short read aligners is divided in Burrows-Wheeler transform (BWT, (Burrows and Wheeler 1994) and hash table based algorithms. Hash table aligners index the read sequences and search through the reference genome or vice-versa. Read indexing algorithm do not need much memory but may be

inefficient for aligning small amount of reads while reference indexing methods need large memory capacities. BWT based aligners use a reversible compression algorithm to build a reference index suffix tree and then search within this suffix tree for possible alignments. The BWT index needs only a fraction for the whole genome sequence alignment as compared to hash table based methods.

In order to handle lack of accuracy in alignments, Li et. al.(Li et. al. 2008) introduced the mapping quality concept which is the measure of confidence that a read actually originated from the position it was aligned to by the mapping program. They consider a read alignment as an estimate of the true alignment and calculate the mapping quality Q_s by phred scaling the alignment's error probability P:

$$Q_s = -10*\log_{10}(P\{\text{read is wrongly mapped}\})$$

Consequently the mapping quality is set to zero for reads that map equally well to multiple positions in the genome. It is common practice to apply mapping qualities to 255 to indicate that mapping quality is not available. As paired end reads combine information of both DNA fragment sides, their mapping quality Q_P is calculated as $Q_P = Q_{s1} + Q_{s2}$. This applies only when both alignments are consistent that is the insert size and alignment directions are both correct. If the alignments do not add up, then both reads will be treated as SE regarding their mapping quality score calculations.

2.3.1.3 Variant detection- The main role of exome sequencing is to detect variants from the reference genome to determine genes associated with diseases. SNPs are determined by the comparison of an assembled consensus sequence which represents the most likely genotype based on the analyzed sequence reads with its reference genome. Simple variant detection approaches apply fixed filters based on percentage of reads containing the same non-reference base call while more advanced methods use a Bayesian approach in combination with prior genotype probabilities to infer the genotype and detect variants. Most of these methods differ in their estimated prior genotype probabilities and thus different quality indices such as base and mapping quality as poor data quality affects SNP calling accuracy. Phred scale quality scores for variant quality estimation determine the probability that a genotype call is wrong.

2.3.1.4 File Formats- Several file formats have been established for handling data in exome sequencing. Following are the most commonly encountered ones:

2.3.1.4.1 FASTQ Format- The FASTQ format is a text based file format for storing sequence read data and associate per base quality read score. It stores sequences and Phred qualities in a single file. It is concise and compact. It is closely related to the FASTA sequence file format and thus lacks an explicit

definition leading to introduction of several incompatible variants. A FASTQ file normally uses four lines per sequence. Following is its description:

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2 using ASCII codes, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

@EAS100R:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAG

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!*((((***+))%%%++)(%%%%).1***-+*))**55CCF>>>>>>CCCCCCC65

2.3.1.4.2 SAM and BAM Format- The SAM format was designed to store nucleotides in a generic way. It consists of one header section and one alignment section. The lines in the header section start with character '@', and lines in the alignment section do not. All lines are TAB delimited and each alignment has 11 mandatory fields and a variable number of optional fields. The mandatory fields are shown in the figure below.

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Figure 2.2: A SAM format file mandatory fields.
Figure directly adopted from Li et. al. 2009

They must be present but their value can be a '*'or a zero (depending on the field) if the corresponding information is unavailable. The optional fields are presented as key-value pairs in the format of TAG: TYPE: VALUE. They store extra information from the platform or aligner. For example, the

'RG' tag keeps the 'read group' information for each read. In combination with the '@RG' header lines, this tag allows each read to be labeled with metadata about its origin, sequencing center and library. The SAM format specification gives a detailed description of each field and the predefined TAGs.

The standard CIGAR description of pair wise alignment defines three operations: 'M' for match/mismatch, 'I' for insertion compared with the reference and 'D' for deletion. The extended CIGAR proposed in SAM added four more operations: 'N' for skipped bases on the reference, 'S' for soft clipping, 'H' for hard clipping and 'P' for padding. These support splicing, clipping, multi-part and padded alignments.

Further additional optional fields also allow for the documentation of less important or program specific data. Color space read information is also described in the optional fields

To improve the performance, a companion format Binary Alignment/Map (BAM) was designed, which is the binary representation of SAM and keeps exactly the same information as SAM. BAM is compressed by the BGZF library, a generic library that was developed to achieve fast random access in a zlib-compatible compressed file. Majority of the space in the BAM file is used to store the base qualities.

2.3.1.4.3 Variant Call Format (VCF)- The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project.

A VCF file is divided into a header and a body section where each header line is identified by a leading '#'. The header stores mandatory information about the file format version and body content. Optional header lines contain meta-data about annotations in the VCF body section. The most common annotations include genotype likelihoods, dbSNP membership, ancestral allele, read depth and mapping quality.

(a) **VCF example**

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

The variant call format (VCF) is a text based file format designed for storing the most prevalent types of sequence variations - including SNPs, DIPs and larger structural variants - together with rich annotations in a standardized way (Danecek et al. [2010]). It is divided into a header and a body section where each header line is identified by a leading '#'. The header stores mandatory information about the file format version and body content. Optional header lines contain meta-data about annotations in the VCF body section. Commonly used annotations include genotype likelihoods, dbSNP membership, ancestral allele, read depth, and mapping quality (Danecek et al. [2010], 1000 Genomes [2010a], 1000 Genomes [2010b]).

Figure 2.3: A VCF file format with its header.
Figure directly adapted from Danecek et. al. 2011

2.3.2 EXOME SEQUENCING AND PANCREATIC CANCER

In order to overcome the loopholes left by association and functional in vitro studies, next generation or massively parallel sequencing studies have developed that are assisted by the completion of the human genome project.

The advantage of massively parallel sequencing (MPS) technology is that they have transformed approaches to data generation enabling the sequencing of an entire human genome region of interest within a very short period of time of less than 2 weeks. These studies have facilitated the comparison of cancers of matched tumor and normal genomes that has significantly increased our understanding of cancer genome biology. They work by shearing the normal DNA into small fragments which are then

amplified and sequenced and mapped back to the reference genome following which sophisticated computational and statistical tools are used to make variant calls.

One of the applications of MPS is exome sequencing that captures the coding regions of the genome using whole genome MPS libraries that effects the nucleic acid hybridization between the genome or genomes of interest and the exome capture probes. Although this approach captures only about 1.8% of the genome, it serves as an outstanding first pass for detecting genes or variations of interest as compared to whole genome approach that generates data with excessively large computational burden of analysis or genome wide associations studies which detects variants with moderate odds ratios in potentially non-coding regions of the genome which are hard to interpret.

Genetic studies of PC taking advantage of this latest advent in sequencing technology have taken place in the last few years. In one of the very first of these studies by Jones et. al. (Jones S et al. 2008), twenty four advanced pancreatic adenocarcinoma had their exonic regions sequenced and among all the core signaling pathways found to be altered, DNA damage control was among the most common ones with 9 genes found to be altered and 83% of the tumors had at least one of the 9 genes altered. Further Wang et. al. (Wang L et al. 2012) investigating 15 pancreatic ductal adenocarcinomas and matched control for genomic changes using exome sequencing which were validated by RNA sequencing found that the most frequently mutated genes were the cell cycle gene *CDKN2A* along with DNA repair genes *TP53* and *SMAD4*. With sporadic pancreatic cancers being well known for increased rates of microsatellite instability and in keeping with this notion, the DNA mismatch repair gene *MLH1* was to found have an increased mutation rate in the cell lines with high variations in general and the *MLH1* expression level correlated with the mutation rates. In another study by Roberts et. al. (Roberts NJ et al. 2012) that utilized exome sequencing for investigating the genetic basis of familial PC (FPC) which is still yet unknown showed that potentially deleterious non-sense mutations of the *ATM* gene were segregating with pancreatic cancer disease status in two families who met the clinical criteria for FPC.

Thus next generation sequencing when applied with appropriate caution has the potentiality of identifying disease causing mutations as opposed to associated mutations from previous association studies and holds the potentiality of spurring on further functional studies to identify how disruptions in these genes are related to the PC pathogenesis.

2.3.3 THE 1000 GENOMES PROJECT: A CATALOGUE OF HUMAN GENETIC VARIATION

The International HapMap project completed in 2003 aimed to determine the common pattern of DNA sequence variation in human genome by characterizing sequence variants and their frequencies spawned

studies focusing on genome-wide search for common variants for disease risk. These studies better known as genome-wide association studies were carried out on multiple different diseases and with sufficient statistical rigor led to identification of variants that were highly significant in disease causation. However a major drawback of these studies were that although they helped identify hitherto unknown genes for common diseases, highly significant variants in non-coding regions of the genome proved difficult to interpret and only explained a fraction of the heritability of a disease. Especially faced with difficulty of interpreting the role of highly significant variants in the non-coding regions of the genome, in 2009 the first human whole exome sequencing study was reported (Ng SB et al. 2009) that was carried out on 8 HapMap individuals and 4 unrelated individuals of a rare dominantly inherited disorder known as Freeman-Sheldon syndrome. This study showed that candidate genes for monogenic disorders could be identified by exome sequencing of a small number of unrelated, affected individuals and further this strategy could be extended to diseases with more complex genetics through larger sample sizes and appropriate weighting of nonsynonymous variants by predicted functional impact.

Thus exome sequencing with its ability to target all the coding regions of the known genes in the human genome spawned the largest study of to identify variants in the human genome's coding region. This project known as the 1000 genomes project aimed to discover genotype and provide accurate haplotype information on all types of DNA polymorphisms in multiple different population groups. (Genomes Project C et al. 2010)

Briefly the 1000 genomes project consisted of three sub-projects namely the trio project, followed by the low-coverage project and the exon project. The trio project consisted of whole genome shotgun sequencing at high coverage (average 42x) of two families of the hapmap population groups YRI and CEU. The low coverage project (2-6x) was performed in 179 individuals from the YRI, CEU, CHB and JPT Hapmap population groups. The final and exon project targeted 8140 exons from 906 randomly selected genes followed by sequencing them at high coverage of greater than 50X on average in 697 individuals from 7 populations groups of Africa, Europe and East Asia.

A total of approximately 15 million SNPs, 1 million short insertions and deletions and 20000 structural variants were discovered with union across of the projects. Of these variations, approximately 55% of the SNPs and 57% of the insertion/deletions were novel having not been found in the dbSNP previously. With respect to the number of variants found to cause synonymous or non-synonymous amino acid changes, it was found that an individual differs from the reference human genome at 10000-11000 non-synonymous sites in addition to 10000-12000 synonymous sites. However the number were much lower for insertion deletions with an average of 190-210 in-frame indels and 220-250 deletions that shift the reading frame per individual.

With such a huge repertoire of human genetic variation data, the 1000 genomes project potentially serves as an efficient filtering tool for removing potentially non-deleterious mutations generated by exome sequencing data as well as identify rare variants that could be potentially risk variants for disease causation.

3.0 SHORTLIST OF GENES SELECTED FOR STUDY

With the focus solely being on the DNA repair genes, 156 DNA repair genes as curated by the UPMC dnapitt crew (https://dnapittcrew.upmc.com/db/orthologs.php) was included for analysis. Following is a list of the genes with the respective pathways they are involved in:

Table 3.1 DNA repair pathways along with genes for each pathway included for study

DNA REPAIR PATHWAY	GENES INVOLVED
BER	APEX1, APEX2, LIG3, MBD4, MPG, MUTYH, NEIL1, NEIL2, NEIL3, NTHL1, OGG1, PARG,
CHROMATIN STRUCTURE	CHAF1A, H2AFX
CONSERVED DDR	CHEK1, CHEK2, HUS1, RAD1, RAD17, RAD9A, RRM1, RRM2B, TP53, ATR, ATRIP
DIRECT REVERSAL OF DAMAGE	ALKBH3, ALKBH2, MGMT
DNA POL	MAD2L2, PCNA, POLA, POLB, POLD1, POLE, POLG, POLH, POLI, POLK, POLL, POLM,
EDITING & PERFORMING ENDONUCLEASE	EXO1, FEN1, ENDOV, SPO11, TREX1, TREX2
FANCONI ANEMIA	FANCA, FANCE, FANCB, FANCC, FANCD2, FANCF, FANCG, FANCI, FANCL, FANCM
GENES DEFECTIVE IN DISEASE	ATM, BLM, RECQL4, WRN
HOMOLOGOUS RECOMBINATION	BRCA1, BRCA2, DMC1, EME1, GEN1, MRE11A, MUS81, NBN, RAD21, RAD50,
MMR	MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1, PMS2
MODULATION OF NUCLEOTIDE POOL	DUT, NUDT1, RRM2B
NER RELATED	ERCC8, DDB1, DDB2, ERCC6, MMS19, TFF2, XAB2
NHEJ	DCLRE1C, PRKDC, SIRT1, XRCC4, XRCC5, XRCC6, LIG4
NER	CCNH, CDK7, CETN2, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, FBXL2, GTF2H1, GTF2H2,
OTHER SUSPECTED GENE	APTX, DCLRE1A, DCLRE1B, RAD52B/RDM1, RECQL, RECQL5, RPA4
RAD6 PATHWAY	RAD18, UBE2A, UBE2B, UBE2N, UBE2V2

4.0 HYPOTHESIS

Based on the previous findings of the confirmed role of DNA repair genes in PC and the success to exome sequencing till date to develop further insights into PC pathogenesis, this technology was utilized to test the following hypothesis:

1. Individuals with both germline mutations and inflammatory conditions (CP) with PC have a higher genetic risk of PC as compared to long term CP individuals who have not progressed on to PC.

2. The pancreas primarily utilizes a subset of DNA repair mechanisms and defects in these mechanisms give the chronic pancreatitis individuals an increased risk of PC as compared to the general population.

3. Rare variants that result in amino acid changes in the DNA repair genes are present in significantly greater number in CP+PC individuals as compared to the CP-PC individuals and thus contribute to greater PC risk..

5.0 PREMISE

PC is a result of prior present inherited and acquired germline mutations and modifications. Increased risk occurs with inherited heterozygous mutations with a greaterprobability for a loss of the the second allele when duct cells are continuously exposed to reactive oxygen species and other toxic inflammatory factors that can cause DNA damage. An increased injury and cell turnover in an oxidative environment favors accumulation of mutations that might give a selective growth advantage to apparently normal cells which go on to develop malignant tumors via uncontrolled proliferation. The table in appendix A shows the inherited germline mutations that have been detected in the genetic studies of inherited PC. (Solomon S et. al. 2012)

6.0 MATERIAL & METHODS

6.1 SELECTION OF STUDY SUBJECTS

A total of 16 patients who had chronic pancreatitis before pancreatic cancer and 11 individuals who had long standing chronic pancreatitis were selected for the study. These individuals were either part of the North American Pancreatitis Study 2 (NAPS2) or the Hereditary Pancreatitis study (HP) or the Pancreatic Adenocarcinoma Gene-Environment Risk (PAGER) study.

Follow-up had been completed on the subset enrolled at Pittsburgh and patients with CP who developed pancreatic cancer had been identified. PAGER represents a cohort of patients with pancreatic cancer that were ascertained using the same instruments as NAPS2 and HP

The primary criteria for selection is prospectively ascertained patients in the NAPS2, HP and PAGER studies with documented chronic pancreatitis for more than 3 years prior to the diagnosis of pancreatic cancer.

6.2 RATIONALE OF STUDY DESIGN

Next generation sequencing has made it possible to identify genetic differences among two different groups of study subjects by scanning the specific regions of the genome leading to identification of possible changes in protein sequence or other changes in regulatory mechanisms that might be involved in disease causation. Specifically massively parallel sequencing technologies that produce raw data with high quality scores and relatively high read depth can accurately identify subtle variations between two different groups of individuals without having to study very large population groups.

One of the most commonly applied massively parallel sequencing technologies is whole exome sequencing that involves sequencing of the coding regions (CDS) and untranscribed regions (UTR) of all the known genes in the genome. Highly malignant diseases like various forms of cancer have been studied using this technology and have led to identification of disease causing variants that are likely to be effective therapeutic targets for the future.

This technology has been applied here to identify variants in the coding regions of the genomic DNA of the DNA repair genes from 16 pancreatic cancer individuals who progressed on to PC from CP. The rationale for focusing on genomic DNA is the target to identify variants that give high risk of PC to CP individuals as compared to the general population. Despite the human genome consisting of more than 20000 genes, the logic behind solely focusing on the DNA repair genes is that cancer is a disease that is characterized by genomic instability and it is the role of the DNA repair system to maintain the stability of genome by removing erroneous genetic changes. Thus if there are germline mutations in the DNA repair genes such that their normal function is disrupted leading to non-correction of mutations that arise as a result of DNA damage after severe inflammation during chronic inflammatory diseases that exposes the tissues of the organ to cytotoxic agents such as proinflammatory cytokines and reactive oxygen species, then the possibility of those damages remaining unrepaired increase and this increases the instability of the genome which can undergo further spontaneous mutations that is characteristic of cancer genomes. Their presence would lead to activation of cellular protective mechanisms, such as cell death and increased proliferation aiming at organ regeneration. Increased cell turnover in an environment rich in oxidative species favors accumulation of DNA damage thus increasing chances of positive selection for mutations that confer growth advantage and hence increases the risk of cancer. Further only variant positions that were heterozygous were included in the analysis. A reason for doing that is genomic DNA was chosen for sequencing and if homozygous mutations generated by sequencing of genomic DNA were chosen for further analysis and if they happen to be in coding regions that affects the protein function, then that affect would be seen in all the cells of the individual and that particular individual could develop multiple different cancers or other genetic diseases characterized by loss of function of DNA repair genes. Due to the fact that the cancer individuals only developed PC and no other form of cancer, heterozygous germline variants would be expected to give the risk with the rationale that loss of the second allele during the inflammatory to cancer transition with the increased cell turnover in an environment rich in oxidative species would facilitate crossing of the threshold of risk needed for a CP individual to progress to PC.

A major outcome of exome sequencing studies is the identification of thousands of variants most of which are non-deleterious mutations that differ from the reference genome mainly because of natural variations and are not likely to be disruptive to normal functioning of the gene and thus unlikely to be pathogenic. Hence a filter is necessary that will help remove these non-pathogenic mutations and limit the analysis to variants that could be potentially pathogenic. The 1000 genomes project is a collaborative project aims that aims to characterize the human genome sequence for investigating relationships between genotype and phenotype. It is one of the largest repositories of human genetic variation data and thus

serves as a suitable filter to remove variants that are unlikely to be causal by setting different parameters such as allele frequency. The 1000 genome project was discussed in detail earlier.

Although whole exome sequencing data that has been filtered against 1000 genomes is useful to significantly narrow down the potentially causal variant list for a disease, there are still likely to be variants that are non-pathogenic and would be difficult to differentiate from the actual causal ones. In order to further aid in determining variants of higher risk scores, GERP or genomic evolution rate profiling scores were used to quantify amino acid variants. GERP is a framework for identifying constrained elements based on the assumptions that purifying constraints result in a relative lack of substitution events. GERP estimates evolutionary rates for individual alignment columns, and compares these inferred rates with a tree describing the neutral substitution rates relating the species under consideration. It subsequently identifies candidate constrained elements by annotating those regions that exhibit fewer than expected substitutions. Each of these elements is scored according to the magnitude of the substitution deficit, measured as "rejected substitutions" (RS). GERP scores can range from a minimum negative value of -6.18 to +12.36 with values above 2 believed to be enriched for truly constrained sites. (Cooper GM et. al. 2005)

6.3 EXPERIMENT SET UP TO TEST HYPOTHESIS

Hypothesis 1: Individuals with CP who have progressed on to PC have a greater load of rare and novel variants in their DNA repair system as compared to long standing CP patients.

In order to test the above hypothesis, the rare and novel variant count in the genes of interest was determined for all the individuals in the study and then Mann-Whitney U test with a standard type one error rate of 0.05 was used to check whether CP+PC individuals carried a greater load of rare genetic variants in the 159 genes of interest as compared to the CP-PC individuals.

Hypothesis 2: The pancreas primarily utilizes a subset of DNA repair mechanisms and defects in these mechanisms give the chronic pancreatitis individuals an increased risk of PC as compared to the general population.

In order to test the above hypothesis, the rare and novel variant count for each pathway as defined previously was determined for all the individuals in the study and Mann- Whitney U test was used to check which pathways had a greater load of rare variants in CP+PC individuals as compared to CP-PC individuals.

Hypothesis 3: Rare variants that result in amino acid changes in the DNA repair genes are at highly conserved positions and thus contribute to PC risk as compared to non-cancerous individuals

In order to test the above hypothesis, rare variants that result in amino acid changes were annotated using GERP scores and the Mann-Whitney U test was used to test the hypothesis that CP+PC individuals had a greater mean number of variants causing amino acid changes in the 159 genes of interest as compared to the CP-PC individuals.
As per the GERP score guidelines, only RS score threshold of above 2 was considered as providing evidence for a highly sensitive site.

6.4 ANNOTATION OF VCF FILE

The VCF file for each variant was annotated using the tool ANNOVAR. (Wang K et. al. 2010) Briefly the annotation files that ANNOVAR utilizes for annotating a set of variants was downloaded from the web using ANNOVAR generated shell scripts and then each VCF file was annotated against the 1000 genomes data to determine variant 1000G minor allele frequency of variants present in 1000G and to identify variants that were novel. The two files generated one with variants novel to 1000 genomes and the other with variants present in 1000 genomes were then again annotated using ANNOVAR to determine the gene, transcript, amino acid change and region of gene where variant is located for each of the heterozygous variant. Each ANNOVAR input file generated two output files after filtering against 1000G and each of these files when annotated using ANNOVAR to identify gene and amino acid changes generated two files one with the gene and region where variation is located and the other with the variants known to effect the protein sequence.

The latest version of the ANNOVAR tool that was installed at FRANK core of the Simulation and Modelling (SAM) Centre of the University of Pittsburgh was used for annotating the variants via UNIX command prompts.

The ANNOVAR specified shell script was used to download the database that ANNOVAR uses for annotation and then utilize them to annotate the variants. Following are the scripts used and the snapshot of the file generated by ANNOVAR.

In order to download the annotation file that ANNOVAR uses for annotating against the thousand genomes project, the following script is used.

annotate_variation.pl -downdb 1000g2012dec_all humandb -buildver hg19 /home/dwhitcomb/sid10

In order to download the annotation file that ANNOVAR uses for annotating variants with gene and amino acid changes, the following script is used.

annotate_variation.pl -downdb -buildver hg19 refGene /home/dwhitcomb/sid10

In order to download the annotation file for determining GERP scores, the following script is used.

annotate_variation.pl -downdb ljb_all -buildver hg19 -webfrom annovar /home/dwhitcomb/sid10

In order to convert the VCF file to an ANNOVAR input file using a sample file HET_ONLY, following script is used.

convert2annovar.pl –format vcf4 –includeinfo /home/dwhitcomb/sid10/ONLY_HET/HET_ONLY.vcf > /home/dwhitcomb/sid10/HET_ONLY.avinput (where HET_ONLY is symbolic of an input VCF file and HET_ONLY.avinput is an ANNOVAR input format file generated that is used for further annotation)

Below are the snapshot of a VCF file and the corresponding ANNOVAR input file.

```
##fileformat=VCFv4.0
##fileDate=20130429
##fileEncoding=windows-1252
##source=CLC Genomics Workbench 6.0.2 build 60287191
##referenceUrl=clc://frank.sam.pitt.edu:7777/3076010-BAAAAAAAAAAAAAP02aa923fd6cba39f--4c219
##referenceFile=unknown
#CHROM  POS       ID  REF  ALT   QUAL  FILTER  INFO
Y       10011648  .   T    G     .     .       AC=10;AF=29.412;AN=
Y       10012034  .   A    G     .     .       AC=6;AF=25.0;AN=24
Y       10028213  .   G    GGGC  .     .       AC=9;AF=75.0;AN=12
Y       13309624  .   A    G     .     .       AC=4;AF=36.364;AN=1
```

Figure 6.1: Snapshot of a VCF file

Y	10011648	10011648	T	G	Y	10011648	. T
Y	10012034	10012034	A	G	Y	10012034	. A
Y	10028213	10028213	-	GGC	Y	10028213	. G
Y	13309624	13309624	A	G	Y	13309624	. A

Figure 6.2: Snapshot of a VCF file converted to ANNOVAR format

In order to annotate an ANNOVAR file against the 1000 genomes data, using a sample file HET_ONLY, following script is used:

cat /home/dwhitcomb/sid10/HET_ONLY.avinput | annotate_variation.pl --filter --dbtype 1000g2012apr_all --buildver hg19 /home/dwhitcomb/sid10/HET_ONLY.avinput /home/dwhitcomb/sid10 (generates two files one with variants present in 1000 genomes and their corresponding MAF and the other with variants novel to 1000 genomes)

Below are the snapshots of the files generated where one contains all the variants found in the 1000 genomes project with their corresponding minor allele frequency and the other contains all the novel variants.

1000g2012	0.47	1	1E+08	1E+08	T	C	1	1E+08
1000g2012	0.07	1	1E+08	1E+08	G	A	1	1E+08
1000g2012	0.56	1	1E+08	1E+08	T	C	1	1E+08
1000g2012	0.68	1	1E+08	1E+08	G	A	1	1E+08
1000g2012	0.45	1	1E+08	1E+08	T	C	1	1E+08
1000g2012	0.4	1	1.01E+08	1.01E+08	G	A	1	1.01E+08
1000g2012	0.38	1	10093457	10093457	C	T	1	10093457
1000g2012	0.39	1	1.01E+08	1.01E+08	A	G	1	1.01E+08
1000g2012	0.48	1	1.01E+08	1.01E+08	T	C	1	1.01E+08
1000g2012	0.44	1	1.01E+08	1.01E+08	A	G	1	1.01E+08
1000g2012	0.48	1	10190884	10190884	A	T	1	10190884
1000g2012	0.59	1	1.02E+08	1.02E+08	G	A	1	1.02E+08
1000g2012	0.01	1	1025673	1025674	CA	-	1	1025672
1000g2012	0.26	1	10318652	10318652	C	G	1	10318652
1000g2012	0.37	1	10322054	10322054	G	A	1	10322054
1000g2012	0.27	1	10339210	10339211	AT	-	1	10339209
1000g2012	0.24	1	1.03E+08	1.03E+08	A	G	1	1.03E+08

Figure 6.3: Snapshot of an ANNOVAR file after annotating with thousand genomes minor allele frequency displaying variants present in the thousand genomes project

13	73330313	73330313	G	A	13	73330313	.	G
7	75616779	75616779	C	-	7	75616778	.	CC
4	17844980	17844980	G	A	4	17844980	.	G
3	75785170	75785170	-	CCA	3	75785170	.	C
22	50656096	50656096	A	-	22	50656095	.	AA
7	1.48E+08	1.48E+08	TT	-	7	1.48E+08	.	TTT
7	1.29E+08	1.29E+08	-	A	7	1.29E+08	.	C
2	96610899	96610899	T	C	2	96610899	.	T
11	9003366	9003366	T	G	11	9003366	.	T
6	52055259	52055259	C	T	6	52055259	.	C
2	1.33E+08	1.33E+08	A	T	2	1.33E+08	.	A
9	34979253	34979253	A	-	9	34979252	.	AA
11	1.03E+08	1.03E+08	C	G	11	1.03E+08	.	C

Figure 6.4: Snapshot of an ANNOVAR file after annotating with thousand genomes minor allele frequency displaying variants absent in the thousand genomes project

In order to annotate the rare variants with minor allele frequency of less than 0.01 and novel variants, the file containing the variants present in the thousand genomes project was edited in excel and only those variants that were rare were annotated with gene and amino acid changes. Following is the script used:

cat /home/dwhitcomb/sid10/HET_ONLY.avinput.hg19_ALL.sites.2012_04_dropped
|annotate_variation.pl --buildver hg19
/home/dwhitcomb/sid10/HET_ONLY.avinput.hg19_ALL.sites.2012_04_filtered /home/dwhitcomb/sid10

Following is the snapshots of the files generated by ANNOVAR.

Location of mutation	Gene	Chr		
UTR3	BORA,DIS3	13	73330313	73330313
intronic	TMEM120A	7	75616779	75616779
exonic	NCAPG	4	17844980	17844980
downstream	ZNF717	3	75785170	75785170
downstream	SELO,TUBGCP6	22	50656096	50656096
UTR3	CNTNAP2	7	1.48E+08	1.48E+08
UTR3	TSPAN33	7	1.29E+08	1.29E+08
intergenic	TRIM43(dist=345430),FAHD2CP(dist=65400)	2	96610899	96610899
UTR3	NRIP3	11	9003366	9003366
UTR3	IL17A	6	52055259	52055259
intergenic	ANKRD30BL(dist=3349),GPR39(dist=155256)	2	1.33E+08	1.33E+08
UTR3	KIAA1045	9	34979253	34979253
exonic	MMP13	11	1.03E+08	1.03E+08
exonic	MUC2	11	1093582	1093582
UTR3	TRERF1	6	42194278	42194278
UTR5	NR5A2	1	2E+08	2E+08

Figure 6.5: Snapshot of an ANNOVAR file showing genes and region in gene where variants are located.

Line no	Type of mutation	gene/transcript/nucleotide change/aa change
line3	nonsynonymous SNV	NCAPG:NM_022346:exon21:c.G2980A:p.A994T,
line13	nonsynonymous SNV	MMP13:NM_002427:exon3:c.G403C:p.E135Q,
line14	nonsynonymous SNV	MUC2:NM_002457:exon30:c.G5401C:p.A1801P,
line23	frameshift deletion	CLDN16:NM_006580:exon1:c.164delG:p.R55fs,
line24	synonymous SNV	ZNF717:NM_001128223:exon5:c.A1392G:p.S464S,
line34	frameshift deletion	ZNF717:NM_001128223:exon5:c.1422delA:p.T474fs,
line42	nonsynonymous SNV	PLXNA4:NM_020911:exon23:c.T4345C:p.F1449L,
line44	synonymous SNV	CNN2:NM_004368:exon6:c.G636A:p.T212T,CNN2:NM_201277:exor
line46	frameshift deletion	SCRIB:NM_015356:exon15:c.2021delT:p.V674fs,SCRIB:NM_182706:
line50	synonymous SNV	MUC6:NM_005961:exon31:c.C5643A:p.T1881T,
line51	nonsynonymous SNV	ZNF488:NM_153034:exon2:c.A46G:p.I16V,
line58	stopgain SNV	PPEF2:NM_006239:exon17:c.C2212T:p.Q738X,

Figure 6.6: Snapshot of an ANNOVAR file after filtering showing variants that alter the protein sequence

In order to annotate the amino acid altering variants with GERP scores, the following ANNOVAR script was used.

cat /home/dwhitcomb/sid10/HET_ONLY.avinput | annotate_variation.pl --filter --dbtype ljb_all --buildver hg19 /home/dwhitcomb/sid10/HET_ONLY.avinput /home/dwhitcomb/sid10

Following is the snapshot of the file generated by ANNOVAR:

ljb_all	0.999449	1	1.01E+08	1.01E+08	G	A	1	1.01E+08
ljb_all	0.992872	1	1.08E+08	1.08E+08	G	A	1	1.08E+08
ljb_all	0.074812	1	1.09E+08	1.09E+08	C	T	1	1.09E+08
ljb_all	0.001663	1	1.09E+08	1.09E+08	G	T	1	1.09E+08
ljb_all	0.807261	1	1.09E+08	1.09E+08	G	C	1	1.09E+08

Figure 6.7: Snapshot of variants that cause an amino acid change with their corresponding GERP score.

6.5 STATISTICAL TESTING

The Mann Whitney U test was used to test the hypothesis that CP+PC individuals have a greater mutation load of common, rare, synonymous rare and novel as well protein sequence altering synonymous and novel variants as compared to CP-PC individuals. The same test was also used to test the hypothesis that specific DNA repair pathways carry a greater load of rare and novel variants in the CP+PC individuals as compared to CP-PC individuals. A standard type one error rate of 0.05 was chosen and any p value below 0.05 was considered as significant.

7.0 RESULTS

7.1 PATIENT CHARACTERISTICS

A total of 16 CP+PC patients of which 10 were male (Mean age=56, SD=9.1, Range=49-79) and 6 who were female (Mean age=55.8, SD=13.36, Range=40-76) were selected for whole exome sequencing. Of the 10 male patients, 4 had primary phenotype of CP, with one having a secondary phenotype of hereditary pancreatitis and 4 had recurrent acute pancreatitis while the remaining two had a primary phenotype of PC. Among the 6 females, two had a primary phenotype of CP while 3 of them had PC as the primary phenotype. One of the males with CP also had a secondary phenotype of HP while a second male had a secondary phenotype of acute pancreatitis. Two females with primary phenotype of PC also had reported family history of PC.

A major difference between the CP+PC and CP-PC patients is the average age which was greater than 50 for the CP+PC patients and less than 30 for the females ($p<0.0001$) and with this being an association study, the significant age difference of the two groups of individuals could be a major confounding factor that could distort the association test findings. Thus an age adjustment of the association test results was done in all of the later comparisons.

The details of the patients are listed in the table 7.1 below.

Table 7.1 Phenotype and demographic detail of CP+PC patients included in study

Patient ID	Primary phenotype	Secondary phenotype	Age, Sex, Cancer history
HP3			45, Female, PC
HP512	CP	HP	76 , Male, PC
HP637	CP		76, Female, PC
NAPS232	RAP		Male, 66, PC & Parathyroid cancer
NAPS823	RAP		Male, 49, PC & Parathyroid adenoma
NA20	CP		Male, 60, PC
NA52	RAP		Male, 59, PC
NA63	CP		Male, 58, PC
NA437	CP		Male, 79, PC
NA1066	CP		Female, 40, PC
NA1265	RAP		Male, 57, PC
PA18	PC		64, Female, PC
PA227	PC	AP	69, Male, PC
PA884	PC	Family history of PC	60, Female, PC
PA1238	PC	Family history of PC	50, Female, PC
PA1306	PC		62, Male, PC

A total of 11 CP-PC patients of which 7 were male (Mean age=30.4, SD=17.4, Range=12-59) and 4 of whom were female (Mean age=26, SD=16.5, Range=2-39) were selected for whole exome sequencing. Of the 7 male patients, 2 of them had acute pancreatitis as primary phenotype and 5 of them had CP. Among the 4 females, two of them had acute pancreatitis while two of them had idiopathic CP.

Table 7.2 Phenotype and demographic detail of CP-PC patients included in study

Patient ID	Primary phenotype	Secondary phenotype	Age, Sex
HP470	Acute Pancreatitis (AP)		12,M
HP653	Acute Pancreatitis (AP)		34,F
HP654	Acute Pancreatitis (AP)		2,F
HP657	Acute Pancreatitis (AP)	Had cholecystectomy at 28yr	39,M
NA1135	Chronic Pancreatitis (CP) (idiopathic)	gallstones	39,F
NA1396	Chronic Pancreatitis (CP) (alcohol & hereditary)	duct obstruction	59,M
NA1499	Chronic Pancreatitis (CP) (idiopathic)	duct obstruction pancreatic divisum	15,M
NA1501	Chronic Pancreatitis (CP) (obstructive)		15,M
NA324	Chronic Pancreatitis (CP) (autoimmune and idiopathic)		32,M
NA6600	Chronic Pancreatitis (CP) (alcoholic)		41,M
NA992	Chronic Pancreatitis (CP) (idiopathic)		29,F

An ANOVA test of the comparison of the mean difference in age of the two groups of patients yielded a significant p value of less than 0.0001 which suggests that age could act as a confounder of the findings of this association study. However the fact that germline mutations have been investigated which

does not change over time eliminates age as a confounding variable. Further the potentiality remains that some of these non cancerous patients may go on to develop PC in the future biasing the test towards null hypothesis of no significant difference in age and the effects further investigated over here could possibly be the same or even larger if the comparison group was confirmed cancer free older participants.

7.2 EXOME SEQUENCING MAPPING SUMMARY REPORT

Each PC individual had an average of 48376235 reads (SD=7248196, Range=37167293-62894601) of which an average of 84% of reads (SD=0.94, Range=82.81-86.12) mapped uniquely to the reference database and an average of 16% of reads (SD=0.94, Range=13.88-17.19) failing to map to the reference genome. Table 6.3 below shows that for each CP+PC individual, more than 80% of the reads successfully mapped to the reference genome.

Table 7.3 Percentage of mapped and unmapped read for each CP+PC individual

Patient ID	Total number of reads	No of mapped reads (% of mapped reads)	No of unmapped reads (% of unmapped reads)
HP3	54008511	44947941 (83.22)	9060570 (16.78)
HP512	45629884	38513826 (84.4)	7116058 (15.6)
HP637	42664418	35644087 (83.55)	7020331 (16.45)
NA20	45937737	39293836 (85.54)	6643901 (14.46)
NA52	42370995	35088538 (82.81)	7282457 (17.19)
NA63	50234433	42092016 (83.79)	8142417 (16.21)
NA232	56855082	47638720 (83.79)	9216362 (16.21)
NA437	62894601	52546095 (83.55)	10348506 (16.45)
NA823	45073267	38553787 (85.54)	6519480 (14.46)
NA1066	53461052	44834817 (83.86)	8626235 (16.14)
NA1265	52768731	44898646 (85.09)	7870085 (14.91)
PA18	37167293	31095808 (83.66)	6071485 (16.34)
PA227	41662025	35045369 (84.12)	6616656 (15.88)
PA884	57983386	49936887 (86.12)	8046499 (13.88)
PA1238	43969114	36643806 (83.34)	7325308 (16.66)
PA1306	41339236	34710146 (83.96)	6629090 (16.04)

With regards to distribution of mapped read length, paired end reads that were 75bp long formed an average of 98.71% (SD=0.10, Range=98.44-98.78) of the total mapped reads. The table below lists the number of mapped 75bp paired end reads and the number of mapped bases covered by it as the percentage of total mapped reads.

Table 7.4 Number of 75bp paired end read and percentage of 75bp paired end reads as percentage of all mapped reads for each CP+PC individual

Patient ID	Number of 75bp paired end reads that mapped to the reference genome	Base pairs covered by 75bp paired end reads (% of total mapped reads)
HP3	44387653	3329073975 (99.3)
HP512	38035542	2852665650 (99.3)
HP637	35194982	2639623650 (99.3)
NA20	38809613	2910720975 (99.3)
NA52	34647990	2598599250 (99.3)
NA63	41563427	3117257025 (99.3)
NA232	47039672	3527975400 (99.3)
NA437	51896161	3892212075 (99.3)
NA823	38078575	2855893125 (99.3)
NA1066	44273771	3320532825 (99.3)
NA1265	44343472	3325760400 (99.3)
PA18	30702596	2302694700 (99.3)
PA227	34607919	2595593925 (99.3)
PA884	49329035	3699677625 (99.3)
PA1238	36182747	2713706025 (99.3)
PA1306	34280247	2571018525 (99.3)

7.3 EXOME SEQUENCING COVERAGE STATISTICS FOR CP+PC INDIVIDUALS

A major issue in next generation sequencing is whether the data has sufficient coverage in order to make variant calls with a high confidence. In the dataset for the 16 CP+PC patients, the average coverage for all the cancer patients for all the reads was 34X and for the mapped reads it was 31X suggesting that the

overall dataset was of fairly high quality to make variant calls. Following table lists the coverage of all reads and mapped reads for the CP+PC individuals.

Table 7.5: Average coverage of total reads and mapped reads for 16 CP+PC patients

PATIENT ID	COVERAGE OF ALL READS	COVERAGE OF MAPPED READS
HP3	41X	34X
HP512	35X	29X
HP637	32X	27X
NA20	35X	30X
NA52	32X	27X
NA63	38X	32X
NA232	43X	36X
NA437	48X	40X
NA823	34X	29X
NA1066	40X	34X
NA1265	40X	34X
PA18	28X	24X
PA227	32X	27X
PA884	44X	38X
PA1238	33X	28X
PA1306	31X	27X

7.4 EXOME SEQUENCING VARIANT CALL STATISTICS

Each of the paired end reads after mapping to the reference genome (NCBI GRCh37/hg19), produced homozygous and heterozygous variant calls with the following statistics.

Table 7.6 Number of homozygous and heterozygous variant loci for each CP+PC patient

Patient ID	No of homozygous / No of heterozygous variation loci
HP3	12313/19303
HP512	10441/18792
HP637	10419/17930
NA20	11198/18870
NA52	10098/17456
NA63	12253/20255
NA232	12794/21287
NA437	13825/23269
NA823	11429/19304
NA1066	12595/20920
NA1265	12661/20849
PA18	9295/16397
PA227	9780/16642
PA884	12285/18886
PA1238	10332/17557
PA1306	9819/17042

7.5 ANNOVAR 1000 GENOMES ANNOTATION RESULTS

Following annotation of the heterozygous variants using ANNOVAR against the 1000 genomes project detected variants, two output files were generated of which one listed the variants present in the genomes project with their corresponding 1000 genomes minor allele frequency and one listing the variants novel to 1000 genomes. An average of 23852 (SD=1684, Range=21321-27133) heterozygous variants were present in each cancer individual of which an average of 20164 (SD=1337, Range=18136-22840) variants were present in 1000 genomes. An average of 825 (SD=51, Range=747-924) variants were present in 1000 genomes at a minor allele frequency of less than 1%. The following table lists the frequency of the

number of variants found in the 1000 genomes, the number of variants that were rare as per the 1000 genomes minor allele frequency annotations and finally the variants that were novel as per 1000 genomes.

Table 7.7 Number of rare and novel variants as per thousand genomes annotation for each CP +PC patient

Patient ID	Number of heterozygous variant positions	Number of variants annotated by 1000g2012apr (as percentage of all heterozygous variants)	No of variants present in 1000 genomes that had MAF<0.01 (as percentage of all 1000 genomes variants)	Number of variants absent in 1000g2012apr (as percentage of all heterozygous variants)
HP3	25831	21929 (84.9%)	867 (3.95%)	3630 (14.1%)
HP512	24065	20404 (84.8%)	850 (4.17%)	3410 (14.1%)
HP637	23026	19397 (84.2%)	826 (4.26%)	3397 (14.8%)
NA20	23422	19872 (84.8%)	863 (4.34%)	3336 (14.2%)
NA63	24969	20709 (82.9%)	879 (4.25%)	3954 (15.8%)
NA52	22356	19022 (85.1%)	796 (4.19%)	3138 (14.9%)
NA232	25408	21522 (84.7%)	872 (4.05%)	3601 (14.2%)
NA437	27133	22840 (84.2%)	924 (4.05%)	3997 (14.7%)
NA823	24213	20364 (84.1%)	821 (4.03%)	3586 (14.8%)
NA1066	25450	21343 (83.9%)	854 (4.00%)	3795 (14.9%)
NA1265	25230	21197 (84%)	790 (3.73%)	3756 (14.9%)
PA18	21868	18741 (85.7%)	774 (4.13%)	2933 (13.4%)
PA227	21321	18136 (85.1%)	748 (4.12%)	2983 (14%)
PA884	22644	19135 (84.5%)	821 (4.29%)	3282 (14.5%)
PA1238	22646	19084 (84.3%)	773 (4.05%)	3335 (14.7%)
PA1306	22058	18940 (85.9%)	747 (3.94%)	2939 (13.3%)

7.6 ANNOVAR GENE AND AMINO ACID CHANGE ANNOTATION RESULT FOR DNA REPAIR GENES IN CP+PC INDIVIDUALS

Following the gene and amino acid annotation of the variants that were rare and absent as per reported 1000 genomes, minor allele frequency, only variants corresponding to the variants in the 159 genes of interest were chosen for further analysis. An average of 183 variants (SD=22 to the nearest integer, Range=145-219) were found in the 16 patients of which an average of 157 variants (SD= 20 to the nearest integer, Range=124-196) were common and an average of 12 variants were rare (SD=7 to the nearest integer, Range=4-29). On an average each cancer individual carried 13 novel variants (SD=4 to the nearest integer, Range=6-19) in the 159 genes. Further each cancer individual carried an average of 4 rare protein sequence altering mutation (SD=3 to the nearest integer, Range=0-8) and average of 2 novel protein sequence altering variants (SD=1.5, Range=0-5). Further every individual carried at most 4 rare synonymous variants and at the least no rare synonymous variants with the average being 1.6 (SD=1.5, Range=0-4). With regards to novel variants accounting for synonymous mutations, the average was less than one at 0.69 (SD=0.6, Range=0-2). The following table lists the number of variants that were heterozygous for all the 159 genes of interest that were either rare as per 1000 genomes reported minor allele frequency or were completely absent in the 1000 genomes variant data.

Table 7.8 Total, rare and novel nucleotide and protein sequence altering/non-altering variants detected for each CP+PC patient

Patient ID	Total number of het variant	Number of common variants with MAF>0.01 (% of total het variant)	Number of rare variants with MAF<0.01 & number of novel variants (%of total het variants)	Number of rare protein seq altering (% of rare variants) & novel variants (% of total novel variants)	Number of rare synonymous (% of rare variants) & novel synonymous variants (% of total novel variants)
HP3	211	165 (78.2)	29 (13.7) & 17 (8.1)	8 (27.6) & 3 (17.7)	4 (13.8) & 1 (5.9)
HP512	169	132 (78.1)	23 (13.6) & 14 (8.3)	8 (34.8) & 3 (21.4)	4 (17.4) & 1 (7.1)
HP637	184	150 (81.5)	21 (11.4) & 13 (8.1)	8 (38.1) & 4 (30.8)	0 & 0
NA20	201	169 (84.1)	16 (8) & 16 (7.9)	0 & 0	2 (12.5) & 0
NA52	155	124 (80)	17 (11) & 14 (9)	4 (23.5) & 1 (7.1)	0 & 0
NA63	214	189 (88.3)	13 (6.1) & 12 (5.6)	5 (38.5) & 2 (16.7)	4 (30.7) & 1 (8.3)
NA232	164	146 (89)	5 (3.1) & 13 (7.9)	2 (40) & 5 (38.5)	0 & 1 (7.7)
NA437	219	196 (89.5)	5 (2.3) & 19 (8.2)	2 (40) & 5 (26.3)	0 & 1 (5.3)
NA823	145	135 (93.1)	4 (2.8) & 6 (4.1)	1 (25) & 1 (16.6)	1 (25) & 0
NA1066	183	168 (91.8)	7 (3.8) & 8 (4.4)	2 (28.6) & 2 (25)	3 (43) & 1 (12.5)
NA1265	207	181 (87.4)	14 (6.7) & 12 (5.9)	6 (42.9) & 1 (8.3)	2 (14.3) & 1 (8.3)
PA18	189	163 (86.2)	13 (6.9) & 13 (6.9)	5 (38.5) & 2 (15.4)	0 & 1 (7.7)
PA227	167	150 (89.8)	7 (4.2) & 10 (6)	3 (42.9) & 1 (10)	2 (28.3) & 1 (10)
PA884	182	160 (87.9)	6 (3.3) & 16 (8.8)	0 (0) & 1 (6.25)	2 (33.3) & 2 (12.5)
PA1238	162	144 (88.9)	10 (6.2) & 8 (4.9)	1 (10) & 2 (25)	1 (10) & 0
PA1306	179	154 (86)	6 (3.4) & 19 (10.6)	2 (33.3) & 2 (10.5)	1 (16.7) & 0

7.7 ANNOVAR GENE AND AMINO ACID CHANGE ANNOTATION RESULT FOR DNA REPAIR GENES IN CP-PC INDIVIDUALS

An average of 54 variants (SD=12 to the nearest integer, Range=29-71) were found in the 11 patients of which an average of 50 variants (SD= 12.53 to the nearest integer, Range=25-68) were common and an average of 4 variants were rare (SD=2.71 to the nearest integer, Range=1-10). On an average each CP individual carried 2 novel variants (SD=1.3 to the nearest integer, Range=1-5) in the 159 genes. Further each pancreatitis individual carried an average of 1 rare protein sequence altering mutation (SD=0.5 to the nearest decimal, Range=0-2) and average of 0.4 novel protein sequence altering variants (SD=0.5, Range=0-1). Further every individual carried at most 4 rare or novel synonymous variants with the average being 0.9 (SD=1.3, Range=0-4). The following table lists the number of variants that were heterozygous for all the 159 genes of interest that were either rare as per 1000 genomes reported minor allele frequency or were completely absent in the 1000 genomes variant data.

Table 7.9 Total, rare and novel nucleotide and protein sequence altering/non-altering variants detected for each CP-PC patient

Patient ID	Total number of het variant	Number of common variants with MAF>0.01 (% of total het variant)	Number of rare variants with MAF<0.01 & number of novel variants (% of total het variants)	Number of rare protein seq altering (% of rare variants) & novel variants (% of total novel variants)	Number of rare synonymous (% of rare variants) & novel synonymous variants (% of total novel variants)
HP470	71	68 (95.8)	2 (2.8) & 1 (1.4)	1 (50) & 0	0 & 1 (100)
HP653	54	47 (87)	5 (9.3) & 2 (3.7)	1 (20) & 0	2 (40) & 0
HP654	50	39 (78)	6 (12) & 5 (10)	1 (16.7) & 0	2 (33.3) & 2
HP657	57	55 (96.5)	1 (1.75) & 1 (1.75)	1 (100) &	0 & 0
NA1135	29	25 (86.2)	1 (3.45) & 3 (10.34)	1 (100) & 1	1 (100) & 0
NA1396	51	50 (98)	0 & 1 (2)	0 & 0	0 & 0
NA1499	71	65 (91.6)	3 (4.2) & 3 (4.2)	1 (33.3) & 0	2 (66.6) & 0
NA1501	56	52 (92.9)	2 (3.6) & 2 (3.6)	2 (100) & 0	0 & 0
NA324	51	48 (94.1)	0 & 3 (5.9)	0 & 1(33.3)	0 & 0
NA6600	60	55 (91.7)	4 (6.7) & 1 (1.7)	1 (25) & 1	0 & 0
NA992	40	39 (97.5)	0 & 1 (2.5)	0 & 0	0 & 0

7.8 RELATIVE FREQUENCY OF PATHWAYS MOST COMMONLY DISRUPTED IN CP+PC AND CP-PC INDIVIDUALS

Multiple DNA repair mechanisms are known to be involved in cellular processes and multiple genes involved in DNA repair have been shown to be involved in causation of PC. Thus we sought to determine that is a there a specific DNA repair pathway that is more frequently mutated in CP+PC individuals as compared to CP-PC individuals. Following table shows the rare and novel variant count of the previously stated DNA repair pathways under the shortlist of genes in the 16 CP+PC and the 11 CP-PC individuals.

Table 7.10 Rare, novel, and total variant count stratified by DNA repair pathway for first 8 CP+PC individuals. [Legend: rare, novel (sum of rare and novel variants)]

Pathway wise rare/novel variant count as per 1000genomes MAF	HP 3	HP 5 1 2	HP 6 3 7	NA 2 0	NA 5 2	NA 6 3	NA 2 3 2	NA 4 3 7
BER	4,3 (7)	4,1 (5)	1,0 (1)	1,2 (3)	1,1 (2)	1,1 (2)	0,2 (2)	
Conserved DDR					6,4 (10)	0,2 (2)	1,1 (2)	
Direct Reversal of Damage			1,0 (1)				1,0 (1)	
DNA Pol	6,1 (7)	4,3 (7)		5,4 (9)	1,1 (2)	3,1 (4)	2,1 (3)	1,0 (1)
Editing and performing	1,1 (2)						0,1 (1)	0,1 (1)
Fanconi anemia	4,5 (9)		3,1 (4)	1,0 (1)	2,1 (3)	1,0 (1)	0,1 (1)	2,0 (2)
Genes def in disease	2,0 (2)			1,1 (2)		0,1 (1)	0,1 (1)	1,0 (1)
Homologous recombination	6,3 (9)	5,5 (10)	9,8 (17)	5,5 (10)	2,1 (3)	2,3 (5)	0,3 (3)	0, 12 (12)
Mismatch repair		3,0 (3)	2,0 (2)		1,1 (2)	2,0 (2)		0,1 (1)
Modulation of nucleotide pool					1,1 (2)			
NER related	1,0 (1)	1,2 (3)		0,1 (1)		1,0 (1)		0,1 (1)
NHEJ		1,0 (1)			1,1 (2)	1,0 (1)		
NER	2,2 (4)	3,2 (5)	3,2 (5)	2,2 (4)	1,3 (4)	2,2 (4)	0,2 (2)	2,3 (5)
Other suspected genes	2,2 (4)	2,1 (3)	1,2 (3)	0,1 (1)	1,0 (1)		1,1 (2)	
RAD6 pathway		1,0 (1)						0,1 (1)

T able 7.11: Rare, novel, and total variant count stratified by DNA repair pathway for remaining 8 CP +PC individuals. [Legend: rare, novel (sum of rare and novel variants)]

Pathway wise rare/novel variant count as per 1000genomes MAF	NA 8 2 3	NA 1 0 6 6	NA 1 2 6 5	PA 1 8	PA 2 2 7	PA 8 8 4	PA 1 2 3 8	PA 1 3 0 6
BER			2,1 (3)	0,2 (2)	1,3 (4)	1,3 (4)	4,1 (5)	
Conserved DDR	1,0 (1)	1,2 (3)		1,2 (3)	0,1 (1)			0,2 (2)
DNA Polymerases	1,1 (2)	2,2 (4)	2,1 (3)	2,0 (2)	3,1 (4)	2,1 (3)	4,0 (4)	0,1 (1)
Editing and perform endonuclease	0,2 (2)							
Fanconi anemia	0,1 (1)	1,0 (1)	3,1 (4)		1,0 (1)	0,4 (4)		2,0 (2)
Genes def in disease							1,1 (2)	
Homologous recombination	0,1 (1)	1,3 (4)	3,2 (5)	2,4 (6)	1,2 (3)	0,1 (1)	1,2 (3)	1,9 (10)
Mismatch repair		2,0 (2)		4,1 (5)	1,0 (1)	2,0 (2)	0,1 (1)	1,2 (3)
Modulation of nucleotide pool		0,1 (1)						
NER related				1,0 (1)		0,1 (1)		
NHEJ	1,1 (2)		1,3 (4)	0,3 (3)	0,1 (1)	1,3 (4)	0,1 (1)	0,2 (2)
NER	1,0 (1)		2,4 (6)	2,0 (2)	0,1 (1)	0,3 (3)	0,2 (2)	1,3 (4)
Other suspected genes				1,1 (2)				
RAD6 pathway					0,1 (1)			1,0 (1)
DNA protein cross-link			1,0 (1)					

Table 7.12 Rare and novel variant count Stratified by DNA repair pathway for CP-PC individuals [legend: rare, novel (sum of rare and novel variants)]

Pathway wise rare/novel variant count as per 1000genomes MAF	HP470	HP653	HP654	HP657	NA11355	NA13966	NA14999	NA15001	NA3244	NA66000	NA9922
BER	1,0 (1)			0,1 (1)			2,0 (2)	1,0 (1)		1,0 (1)	
Conserved DDR	1,0 (1)										
DNA Polymerases		2,0 (2)	2,1 (3)		0,1 (1)		0,1 (1)		0,1 (1)	2,0 (2)	0,1 (1)
Fanconi Anemia		0,2 (2)	0,3 (3)				0,1 (1)	0,1 (1)			
Homologous Recombination							0,1 (1)	1,0 (1)		0,1 (1)	
Mismatch Repair			1,0 (1)		1,1 (2)				0,1 (1)		
NER Related		3,0 (3)				0,1 (1)	1,0 (1)		0,1 (1)		
NHEJ								0,1 (1)		1,0 (1)	
Nucleotide Excision Repair			2,0 (2)								

7.9 NUMBER OF RARE NON-SYNONYMOUS PROTEIN SEQUENCE ALTERING VARIANTS DETECTED IN DNA REPAIR GENES OF 16 CP+PC INDIVIDUALS

Of the 16 CP+PC individuals, 14 patients carried at least one rare non-synonymous mutation while 15 patients carried at least one novel non-synonymous mutation. The average number of rare non-synonymous mutation in CP+PC individuals were 4 to the nearest integer (SD=2.8 Range=0-8) and an average of 2 (SD=1.5 Range=0-4) to the nearest integer of rare synonymous variants. The average number of novel non-synonymous mutations was in the PC individuals was 2 to the nearest integer (SD=1.5, Range=0-5) and the average of novel synonymous mutations was less than one at 0.7 to the nearest integer (SD=0.6, Range=0-2). The following table lists the number of rare and novel non-synonymous and synonymous variants found in each of the cancer individuals.

Table 7.13 Number of rare and novel protein sequence altering variants for each CP+PC individual

Patient ID	Number of rare/novel protein sequence altering variants	Number of rare/novel synonymous variants
HP3	5 / 3	3 / 1
HP512	4 / 4	2 / 1
HP637	3 /5	0 / 0
NA20	0 / 0	1/ 0
NA52	2 / 2	0 / 0
NA63	5 / 2	3 / 1
NA232	2 / 5	0 / 1
NA437	2 / 5	0 / 1
NA823	1 / 1	1 / 0
NA1066	2 / 2	3 / 1
NA1265	6 / 1	2 / 1
PA18	5 / 2	0 / 1
PA227	3 / 1	2 / 1
PA884	0 / 1	2 / 2
PA1238	1 / 2	1 / 0
PA1306	2 / 2	0 / 1

7.10 NUMBER OF RARE NON-SYNONYMOUS PROTEIN SEQUENCE ALTERING VARIANTS DETECTED IN DNA REPAIR GENES OF 11 CP-PC INDIVIDUALS

Of the 11 CP-PC individuals, 9 patients carried at least one rare non-synonymous mutation while 4 patients carried at least one novel non-synonymous mutation. The average number of rare non-synonymous mutation in CP-PC individuals was 1 to the nearest integer (SD=0.54 Range=0-2) and an average of 0.72 (SD=0.90 Range=0-2) to the nearest integer of rare synonymous variants. The average number of novel non-synonymous mutations in the CP individuals was 0.3 to the nearest decimal point

(SD=0.5, Range=0-1) and the average of novel synonymous mutations was less than one at 0.2 to the nearest decimal point (SD=0.6, Range=0-2). The following table lists the number of rare and novel non-synonymous and synonymous variants found in each of the chronic pancreatitis individuals.

Table 7.14 Number of rare and novel protein sequence altering variants for each CP-PC individual

Patient ID	Number of rare non-synonymous/novel non-synonymous variants	Number of rare synonymous/novel synonymous variants
HP470	1/0	1/0
HP653	1/0	2/0
HP654	1/0	2/2
HP657	1/1	0/0
NA1135	1/1	1/0
NA1396	0/0	0/0
NA1499	1/0	2/0
NA1501	2/0	0/0
NA324	1/1	0/0
NA6600	1/1	0/0
NA992	0/0	0/0

7.11 NOVEL GERMLINE PROTEIN SEQUENCE ALTERING MUTATIONS FOUND IN CP+PC INDIVIDUALS

The rare and novel germline mutations that resulted in alteration of protein sequences in the CP+PC individuals are listed in the table below along with the patient ID of the individual carrying that mutation. Using Annovar to annotate resulted in annotation of the variant position in the transcript, the transcript ID, the cDNA change, amino acid change and the type of mutation. A total of 82 germline mutations that

altered protein sequence were found in the 16 individuals of which 38 were absent from the thousand genomes project and 44 were present in the thousand genomes at a minor allele frequency of 0.01 or less. A major issue in exome sequencing as compared to traditional Sanger sequencing is that unlike Sanger sequencing, exome sequencing does not have very high fidelity rates. Thus coverage data along with allele count of a particular variant is very important in determining whether a variant is truly present or is just a sequencing artifact. Of the 38 novel protein sequence altering variants, 21 were present with an alternative allele counting for greater than 20 reads and being present at a frequency of at least 20% in those reads and of the 44 rare variants, 38 were present with an alternative allele counting for greater than 20 reads and being present at a frequency of at least 20% in those reads.

The GERP score annotation results did not report a single mutation to have a GERP score above the threshold of 2 suggesting that majority of the mutations identified might not be evolutionarily constrained and thus not highly conserved sites.

The table below lists all the protein sequence altering variants found in the 16 CP+PC individuals that were either rare as per thousand genomes minor allele frequency or were absent in thousand genomes variant data.

Table 7.15 Novel and rare germline protein sequence altering variants for HP3

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
GEN1	GEN1:NM_001130009:exon5:c.G566A p.S189N	2	17947886	0.01	AC=43;AF=51.807;AN=83
POLQ	POLQ:NM_199420:exon16:c.C4141A:p.P1381T	3	121207637	0.01	AC=24;AF=39.344;AN=61
POLN	POLN:NM_181808:exon5:c.C943T:p.P315S	4	2195009	0.01	AC=16;AF=41.026;AN=39
REV3L	REV3L:NM_002912:exon13:c.G5434C:p.D1812H,	6	111694124	0.01	AC=17;AF=35.417;AN=48
BRCA1	BRCA1:NM_007294:exon15:c.G4956A:p.M1652I BRCA1:NM_007300:exon16:c.G5019A:p.M1673I BRCA1:NM_007297:exon14:c.G4815A:p.M1605I BRCA1:NM_007298:exon14:c.G1644A:p.M548I	17	41222975	0.01	AC=31;AF=26.724;AN=116
MRE11A	MRE11A:NM_005590:exon8:c.T752G:p.I251R	11	94204833	NOVEL	AC=3;AF=25.0;AN=12
BRCA2	BRCA2:NM_000059:exon10:c.1751_1755del:p.584_585del	13	32907366-32907370	NOVEL	AC=25;AF=46.296;AN=54
FANCA	FANCA:NM_001018112:exon8:c.C775A:p.P259T		89869684	NOVEL	AC=5;AF=27.778;AN=18

Table 7.16 Novel and rare germline protein sequence altering variants for HP512

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
MSH6	MSH6:NM_000179:exon4:c.T2633C:p.V878A	2	48027755	0.01	AC=10;AF=23.256;AN=43
PMS1	PMS1:NM_000534:exon6:c.G605A:p.R202K	2	190708712	0.01	AC=9;AF=25.714;AN=35
MBD4	MBD4:NM_001276271:exon3:c.T1073C:p.I358T	3	129155414	0.01	AC=20;AF=64.516;AN=31
PMS2	PMS2:NM_000535:exon11:c.A1789T:p.T597S,	7	6026607	0.0046	AC=31;AF=40.789;AN=76
DCLRE1A	DCLRE1A:NM_001271816:exon2:c.T100G:p.S34A	10	115612842	NOVEL	AC=3;AF=23.077;AN=13
EME1	EME1:NM_152463:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=32;AF=54.237;AN=59
XRCC1	XRCC1:NM_006297:exon17:c.A1896T:p.Q632H	19	44047550	NOVEL	AC=5;AF=33.333;AN=15
GEN1	GEN1:NM_001130009:exon14:c.2518_2519del:p.840_840del	2	17962997 - 17962998	NOVEL	AC=1;AF=2.273;AN=44

Table 7.17: Novel and rare germline protein sequence altering variants for HP637

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
ALKBH2	ALKBH2:NM_001205179:exon3:c.G409A:p.V137M ALKBH2:NM_001001655:exon4:c.G608A:p.R203H	12	109526189	0.01	AC=12;AF=44.444;AN=27
BRCA2	BRCA2:NM_000059:exon10:c.A1385G:p.E462G,	13	32907000	NOVEL	AC=7;AF=33.333;AN=21
BIVM-ERCC5, ERCC5	ERCC5:NM_000123:exon4:c.G438C:p.L146F BIVM-ERCC5:NM_001204425:exon12:c.G1800C:p.L600F,	13	103506695	NOVEL	AC=4;AF=23.529;AN=17
FANCM	FANCM:NM_020937:exon9:c.1490_1491insA:p.S497fs,	14	45628392	NOVEL	AC=10;AF=31.25;AN=32
FANCM	FANCM:NM_020937:exon20:c.A5224G:p.I1742V,	14	45658449	0.0041	AC=15;AF=30.612;AN=49
MLH3	MLH3:NM_001040108:exon2:c.T2896C:p.S966P	14	75513463	0.01	AC=27;AF=47.368;AN=57
ERCC4	ERCC4:NM_005236:exon11:c.T2117C:p.I706T,	16	14041570	NOVEL	AC=35;AF=47.297;AN=74
EME1	EME1:NM_152463:exon2:c.409_410insAGC:p.K137delinsKQ EME1:NM_001166131:exon2:c.409_410insAGC:p.K137delinsKQ,	17	48452978	NOVEL	AC=19;AF=26.027;AN=73

Table 7.18 Novel and rare germline protein sequence altering variants for NA52

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
GEN1	GEN1:NM_001130009:exon14:c.C2692T:p.R898C	2	17963171	0.01	AC=27;AF=37.5;AN=72
NUDT1	NUDT1: NP_945192.1:c.505C>A; Leu169Met	7	2290601	NOVEL	AC=14;AF=35.0;AN=40
PRKDC	PRKDC: NP_001075109 c340C>T P1154S	8	48811034	NOVEL	AC=3;AF=27.273;AN=11
RECQL5	RECQL5:NM_004259:exon15:c.G1883A:p.S628N,	17	73625852	0.01	AC=5;AF=38.462;AN=13

Table 7.19 Novel and rare germline protein sequence altering variants for NA63

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
ERCC5	ERCC5:NM_000123:exon12:c.A2636G:p.N879S BIVMERCC5:NM_001204425:exon20:c.A3998G:p.N1333S,	13	103520565	0.01	AC=18;AF=58.065;AN=31
FANCM	FANCM:NM_020937:exon1:c.G171C:p.L57F,	14	45605405	0.0009	AC=15;AF=45.455;AN=33
BRCA1	BRCA1:NM_007294:exon10:c.A4039G:p.R1347G BRCA1:NM_007297:exon9:c.A3898G:p.R1300G	17	41243509	0.0009	AC=32;AF=45.07;AN=71
PMS2	PMS2:NM_000535:exon11:c.G1688T:p.R563L	7	6026708	0.01	AC=42;AF=41.176;AN=102
PRKDC	PRKDC: NP_001075109.1:p.[Pro695Ser]; PRKDC: NP_008835.5:p.[Pro695Ser]	8	48841708	0.01	AC=11;AF=35.484;AN=31
ATM	ATM:NM_000051:exon18:c.2807delT:p.L936fs,	11	108139305	NOVEL	AC=4;AF=30.769;AN=13
CETN2	CETN2:NM_004344:exon2:c.92_93insG:p.Q31fs,	X	151998216	NOVEL	AC=3;AF=30.0;AN=10

Table 7.20 Novel and rare germline protein sequence altering variants for NA232

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
POLE	POLE:NM_006231:exon8:c.G776A:p.R259H	12	133253974	0.01	AC=45;AF=40.909;AN=110
POLH	POLH:NM_006502:exon11:c.A1783G:p.M595V,	6	43581935	0.01	AC=39;AF=37.5;AN=104
GEN1	GEN1:NM_001130009:exon14:c.2516_2517del:p.839_839del	2	17962995	NOVEL	AC=1;AF=1.587;AN=63
WRN	WRN:NM_000553:exon25:c.A3101T:p.Y1034F	8	30999079	NOVEL	AC=18;AF=27.692;AN=65
RAD54B	RAD54B:NM_012415:exon14:c.T2363C:p.L788P RAD54B:NM_001205263:exon12:c.T1811C:p.L604P,	8	95390560	NOVEL	AC=6;AF=24.0;AN=25
TDG	TDG:NM_003211:exon3:c.286_287insA:p.E96fs	12	104373728	NOVEL	AC=12;AF=54.545;AN=22
EME1	EME1:NM_001166131:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=35;AF=47.297;AN=74

Table 7.21 Novel and rare germline protein sequence altering variants for NA437

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
ATM	ATM:NM_000051:exon31:c.T4709C:p.V1570A,	11	108164137	0.0005	AC=17;AF=35.417;AN=48
ERCC5	ERCC5:NM_000123:exon12:c.A2636G:p.N879S BIVM-ERCC5:NM_001204425:exon20:c.A3998G:p.N1333S	13	103520565	0.01	AC=15;AF=29.412;AN=51
SMC6	SMC6:NM_001142286:exon5:c.G308A:p.G103E	2	17919524	NOVEL	AC=4;AF=25.0AN=16
GEN1	GEN1:NM_001130009:exon14:c.2518delT:p.L840X	2	17962997	NOVEL	AC=1;AF=1.639;AN=61
GEN1	GEN1:NM_001130009:exon14:c.2515_2516del:p.839_839del	2	17962994	NOVEL	AC=1;AF=1.639;AN=61
EME1	EME1:NM_001166131:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=19;AF=21.591;AN=88
EME1	EME1:NM_001166131:exon7:c.G1306A:p.A436T	17	48456849	NOVEL	AC=32;AF=36.782;AN=87

Table 7.22 Novel and rare germline protein sequence altering variants for NA823

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
REV3L	REV3L:NM_002912:exon13:c.A4004G:p.N1335S	6	111695554	0.0005	AC=13;AF=46.429;AN=28
EME1	EME1:NM_152463:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=21;AF=28.767;AN=73

Table 7.23 Novel and rare germline protein sequence altering variants for NA1066

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
MRE11A	MRE11A:NM_005590:exon13:c.C1475A:p.A492D	11	94192599	0.0018	AC=13;AF=48.148;AN=27
FANCA	FANCA:NM_000135:exon35:c.C3427G:p.L1143V,	16	89813078	0.0009	AC=17;AF=39.535;AN=43
POLL	POLL:NM_001174084:exon9:c.G1492A:p.E498K POLL:NM_001174085:exon9:c.G1216A:p.E406K,	10	103339446	NOVEL	AC=15;AF=46.875;AN=32
POLE	POLE:NM_006231:exon12:c.1171_1173del:p.391_391delTT	12	133252037-133252039	NOVEL	AC=7;AF=31.818;AN=22

Table 7.24 Novel and rare germline protein sequence altering variants for NA1265

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
BRCA2	BRCA2:NM_000059:exon11:c.C5744T:p.T1915M,	13	32914236	0.01	AC=9;AF=50.0;AN=18
APEX1	APEX1:NM_080648:exon3:c.G153C:p.Q51H	14	20924167	0.01	AC=4;AF=36.364;AN=11
FANCI	FANCI:NM_001113378:exon18:c.C1813T:p.L605F	15	89828441	0.0018	AC=17;AF=36.17;AN=47
POLG	POLG:NM_001126131:exon16:c.A2492G:p.Y831C	15	89865073	0.0018	AC=4;AF=28.571;AN=14
GEN1	GEN1:NM_001130009:exon14:c.T2619G:p.S873R	2	17963098	0.01	AC=14;AF=29.167;AN=48
GEN1	GEN1:NM_001130009:exon14:c.C2692T:p.R898C	2	17963171	0.01	AC=37;AF=41.573;AN=89
MUTYH	MUTYH:NM_001048174:exon8:c.G566A:p.R189H MUTYH:NM_001048172:exon8:c.G569A:p.R190H MUTYH:NM_001128425:exon8:c.G650A:p.R217H MUTYH:NM_001048171:exon8:c.G608A:p.R203H	1	45798286	NOVEL	AC=5;AF=31.25;AN=16

Table 7.25 Novel and rare germline protein sequence altering variants for PA18

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
MLH3	MLH3:NM_001040108:exon2:c.G1870C:p.E624Q	14	75514489	0.01	AC=21;AF=53.846;AN=39
BRCA1	BRCA1:NM_007300:exon10:c.G3119A:p.S1040N BRCA1:NM_007297:exon9:c.G2978A:p.S993N	17	41244429	0.01	AC=18;AF=47.368;AN=38
POLQ	POLQ:NM_199420:exon16:c.C4141A:p.P1381T,	3	121207637	0.01	AC=24;AF=64.865;AN=37
MLH1	MLH1:NM_001258274:exon17:c.A1129G:p.K377E MLH1:NM_001258271:exon16:c.A1852G:p.K618E	3	37089130	0.0041	AC=23;AF=46.0;AN=50
MLH1	MLH1:NM_001258274:exon17:c.A1130C:p.K377T MLH1:NM_001258271:exon16:c.A1853C:p.K618T	3	37089131	0.0041	AC=23;AF=46.0;AN=50
RECQL5	RECQL5:NM_004259:exon16:c.C2217G:p.S739R	17	73625286	NOVEL	AC=3;AF=27.273;AN=11
EME1	EME1:NM_152463:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=13;AF=54.167;AN=24

Table 7.26 Novel and rare germline protein sequence altering variants for PA227

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Novel	Read count for variant in exome seq data
FANCM	FANCM:NM_020937:exon20:c.A5224G:p.I1742V,	14	45658449	0.0041	AC=37;AF=43.023;AN=86
MLH3	MLH3:NM_001040108:exon2:c.T2896C:p.S966P	14	75513463	0.01	AC=24;AF=34.286;AN=70
POLG	POLG:NM_001126131:exon10:c.C1760T:p.P587L	15	89868870	0.0009	AC=3;AF=30.0;AN=10
EME1	EME1:NM_152463:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=16;AF=32.0;AN=50

Table 7.27 Novel and rare germline protein sequence altering variants for PA884

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
POLI	POLI:NM_007195:exon9:c.A1255G:p.M419V	18	51818259	NOVEL	AC=6;AF=35.294;AN=17

Table 7.28 Novel and rare germline protein sequence altering variants for PA1238

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
REV3L	REV3L:NM_002912:exon13:c.G5434C:p.D1812H	6	111694124	0.01	AC=27;AF=44.262;AN=61
GTF2H4	GTF2H4:NM_001517:exon2:c.T46A:p.C16S	6	30876859	NOVEL	AC=6;AF=25.0;AN=24
BLM	BLM:NM_000057:exon7:c.A1538T:p.K513I,	15	91304141	NOVEL	AC=10;AF=28.571;AN=35

Table 7.29 Novel and rare germline protein sequence altering variants for PA1306

Gene	Transcript/cDNA/exon no/amino acid change	Chr no	Genomic co-ordinate	Thousand genome MAF/Nov	Read count for variant in exome seq data
MUS81	MUS81:NM_025128:exon10:c.C1048T:p.R350W	11	65631361	0.01	AC=16;AF=43.243;AN=37
FANCC	FANCC:NM_000136:exon7:c.A584T:p.D195V	9	97912307	0.0046	AC=9;AF=75.0;AN=12
ATR	ATR:NM_001184:exon8:c.T1878G:p.D626E	3	142277473	NOVEL	AC=59;AF=47.967;AN=123
EME1	EME1:NM_001166131:exon2:c.409_410insAGC:p.K137delinsKQ	17	48452978	NOVEL	AC=13;AF=25.0;AN=52
PRKDC	PRKDC: NP_001075109.1:p.[*2393fs]; NP_008835.5:p.[*2393fs]	8	48761817	NOVEL	AC=5;AF=29.412;AN=17

7.12 TEST OF SIGNIFICANCE OF DIFFERENCE OF FREQUENCY OF OVERALL TYPE OF MUTATION BETWEEN CP+PC INDIVIDUALS AND CP-PC INDIVIDUALS.

Following the classification of the variants as rare or common according to minor allele frequency, the Mann-Whitney U test was carried out to see if there is a difference in the common or rare or rare and novel synonymous and protein sequence altering variants between CP+PC and CP-PC individuals.

Table 7.30 Mann-Whitney results for common, rare and synonymous variants for comparing CP +PC and CP-PC individuals.

Category of mutations	Mann-Whitney p value for CP+PC vs. CP-PC individuals
Common variants at MAF>0.01 in 156 DNA repair genes	p<0.0001
Rare variants at MAF<=0.01 and novel variants in 156 DNA repair genes	p<0.0001
Synonymous rare and novel variants	p=0.1372
Protein sequence altering rare and novel variants	p=0.0003

7. 13 RARE VARIANT BURDEN TEST OF DNA REPAIR PATHWAYS

The ANOVA test of normality showed that the normality of residual assumption was violated for most of the pathway comparisons of CP+PC and CP-PC individuals. For those pathways or comparisons types failing the normality assumption, the non-parametric Mann-Whitney test was used to compare the rare variant load and rank the most frequently mutated DNA repair pathways in CP+PC individuals as compared to CP-PC individuals. The following table lists the Mann-Whitney test p values ranked from the most significant to the least significant.

Table 7.31 Mann Whitney test for comparison of rare and novel variants in CP+PC vs. CP-PC

DNA repair pathway	Mann Whitney test p	Rank as the most significant
Homologous recombination	2.14*10^-6	1
Nucleotide excision repair	5.13*10^-6	2
DNA polymerases	7.89*10^-4	3
Base excision repair	3.9*10^-3	4
Non-homologous end joining	8.6*10^-3	5
Fanconi Anemia	1.06*10^-2	6
Mismatch repair	1.11*10^-2	7
Conserved DNA damage repair	1.25*10^-2	8
Other suspected genes	1.29*10^-2	9
Genes defective in disease	2.71*10^-2	10
Editing and performing	0.1225	11
RAD6 pathway	0.1225	11
Direct reversal of damage	0.3419	13
Modulation of nucleotide pool	0.3419	13
Nucleotide excision repair	0.4762	15
Repair of DNA protein cross	0.5926	16

The Mann-Whitney test results showed that among the major DNA repair pathways, homologous recombination and nucleotide excision repair were the most frequently carriers of novel and rare germline mutations followed by the DNA polymerases and base excision repair and finally the non-homologous end joining pathway.

A major factor that may have propelled the homologous recombination pathway to be most significant is the fact that it has the second highest number of genes to be involved in DNA repair pathways at 21 and some of the genes could be highly polymorphic as well coupled with the fact that non-homologous end joining type of DNA repair pathway was also fairly significant suggesting DNA double strand break repair might be one of the most frequently mutated pathways in the germline DNA of the repair genes of PC patients as compared to non cancerous patients.

7.14 RANKING OF THE MOST FREQUENTLY MUTATED GENES IN THE DNA REPAIR PATHWAYS

A major objective of this study was to identify which genes were most frequently mutated in the CP+PC individuals that would ideally serve as potential gene to be screened in a high risk population for risk of PC. With PC being a disease with a fairly poor prognosis, identifying high risk genes could be useful for a potential genetic screening in high risk patients such as HP or CP for the future. The following table lists the most frequently mutated genes in the PC patients either as carriers of novel variants or rare variants at a MAF<0.01.

Table 7.32 List of most frequently mutated genes that were either rare or novel for each CP+PC individual

Patient ID	Most frequently mutated gene carrying rare variants	Most frequently mutated gene carrying novel variants
HP3	PARP1 (3)	PARP1, FANCM (3)
HP512	GEN1, REV3L, XRCC1 (2)	GEN1, REV3L (2)
HP637	RAD52 (7), FANCM (2)	RAD52 (6)
NA20	RAD52 (4), PAPD7 (3), RAD23B (2)	RAD52 (4)
NA52	RRM2 , TP53 (2)	GTF2H5 (2)
NA63	ERCC5 (2)	GEN1, CHEK2 (2)
NA232	ATR, POLH. MGMT, POLE, RAD52B (1)	RAD23B (2)
NA437	FANCA, POLK, FANCC, ATM, ERCC5, MNAT1 (1)	RAD52 (6)
NA823	ATR, XRCC4, REV3L, RAD52B (1)	ENDOV (2)
NA1066	MSH6, RAD17, MSH5, MRE11A, POLG, FANCA, POLM (1)	GEN1, SMC6, NUDT1, POLL, RAD9A, RRM1, POLE, RAD51B (1)
NA1265	GEN1 (2)	RAD23B (2)
PA18	MLH1 (2)	APEX1 (2), TP53 (2)
PA227	PARP1, POLS, POLM, BRCA2, FANCM, MLH3, POLG (1)	GEN1, SMC6, NUDT1, POLL, RAD9A, RRM1, POLE, RAD51B (1)
PA884	XRCC5, POLK, MSH3, MLH3,APEX1, POLD1 (1)	FANCD2, SIRT1, APEX1 (2)
PA1238	MUTYH, REV1, REV3L, POLB,WRN, MUS81, POLG, PNKP, XRCC1, APEX2 (1)	OGG1, GTF2H4, MSH5, XRCC2, PRKDC, RAD23B, RAD52, BLM (1)
PA1306	UBE2V2 (1), FANCC, MUS81, GTF2H1, MLH3, FANCA (1)	RAD52 (6)

As the table above shows, that there was no one single common gene or genes that carried rare or novel variants in all CP+PC individuals thus suggesting that depending on the DNA damage type,

multiple DNA repair pathways are involved in the repair of damaged DNA in the pancreas and statistical testing results of pathways tested before using Mann-Whitney test to find the most significantly mutated pathway need to be followed up in larger exome sequencing studies to see if this significance remains even after the patient number increases. Further resequencing based studies in larger PC cohorts to determine the genotype status of the potentially deleterious mutations detected here will help identify the significance of that gene in PC individuals who had CP.

8.0 DISCUSSION

In this study we report 59 out of a total of 82 protein sequence altering germline mutations in 16 PC patients who had CP before developing cancer that were present at an allele count of greater than 20 with coverage of at least 31. A relatively high coverage and allele count was considered due to the fact that aim of the study was to identify germline mutations which should be a moderately high coverage for a variant to be considered as a true variant and not a sequencing artifact.

With the initial hypothesis being that CP+PC had a higher genetic risk of PC as compared to non-cancerous CP individuals, a burden test in the form of Mann-Whitney U test was used to identify if the DNA repair genes carried a greater load of rare or common variants in the genes of interest in CP+PC individuals as compared to CP-PC individuals. The Mann-Whitney U test suggested that CP+PC individuals carry a greater load of both common genetic variants ($p<0.0001$) and rare genetic variants ($p=0.0021$) as well significantly higher proportion of protein sequence altering rare genetic variants ($1.4*10^{-3}$). This proves our initial hypothesis to significant extent with certain limitations including an incapability to gauge the effect of intronic and UTR variants that do not affect protein sequence. However UTR variants are known to be involved in post-transcriptional regulation of gene expression including transport of mRNA out of nucleus as well as determination of translational efficiency and subcellular localization. Thus although the CP+PC patient NA20 did not carry any protein sequence altering variants, it could be that the improper functioning of the DNA repair genes in this individual was related to the presence of the UTR variants that interfered with the regulation of expression of a vital gene thus predisposing this individual to the risk of PC. It is difficult to estimate the effect of common variants as the variant frequency above the threshold of 0.01 has significant variation with minor allele frequency ranging from 0.02-1. However with pancreatic cancer being a fairly rare disease and having a worse prognosis, fairly common variants could be ruled out as susceptible highly penetrative mutations for disease in a high risk group such as chronic pancreatitis individuals. Thus rare variants and novel variants of DNA repair pathways were the focus of further analysis.

With multiple pathways are involved in DNA repair, one of our hypotheses sought to determine if any specific DNA repair pathway/s were more frequently disrupted in the cancer individuals as compared to the non-cancer individuals. We found homologous recombination and nucleotide excision repair being the most frequently carriers of novel and rare germline mutations followed by the DNA polymerases and base excision repair and finally the non-homologous end joining pathway. A major factor contributing to this observation could be the fact that of all the DNA repair pathways, homologous recombination and

nucleotide excision repair has the highest number of genes at 21 and 23 respectively. No particular gene dominated in terms of having the highest number of rare or novel mutations or being common in all cancer individuals suggesting that with DNA repair protein forming complexes to repair damaged DNA, disruption in the sequence of any particular gene could affect the binding of the overall complex and thus interfere with repair of damaged DNA.

Non-synonymous and other protein sequence altering mutations are to known to have the potentiality to interfere with the normal function of a protein and since in the majority of the DNA repair pathways, multi protein complexes act to repair damaged DNA, disruption in normal function of one protein could interfere with the functioning of a protein complex thus interfering with the overall functioning of the DNA repair pathway. Thus it was sought to determine if there is greater proportion of novel and rare protein sequence altering variants in the DNA repair genes of the CP+PC individuals as compared to the CP-PC individuals and it was found that protein sequences altering mutations were present at a significantly greater average in CP+PC individuals as compared to CP-PC individuals. Indeed 15 out of 16 CP+PC patient carried at least one rare or novel protein sequence altering mutation at an average of 4 rare and 2 novel protein sequence altering mutations while although 9 out of 11 CP-PC patients carried at least one rare or novel germline protein sequence altering variants, the average was much lower at 1 rare and 0.4 novel protein sequence altering mutations. The Mann-Whitney U test results supported this finding with p value reported to be 0.0003 for comparison of number of protein sequence altering variants between the two groups of patients. With coverage being a major issue in exome sequencing when the non-synonymous variant comparison was done only variants with coverage of at least 20 and the non-reference allele being present in at least 4 reads, the Mann-Whitney test still reported a high p value of 0.0001 thus proving that rare and novel protein sequence altering variants are present at a greater proportion in cancer individuals and possibly loss of function of these variants give the greater risk of cancer to the CP+PC individuals as compared to the CP-PC individuals.

Among the genes that have been found to be involved in determining PC survival or prognosis or been suspected to have some role in the disease via genetic association studies, only 3 genes, namely *ERCC4*, *ATM* and *MLH1* were found to carry germline mutations in the CP+PC individuals thus demanding significant further attention. Further none of the previous mutations identified in PC mutation screening studies were present in any of the CP+PC individuals suggesting that potentially CP+PC individuals have a different genetic risk as compared to PC individuals who did not have CP.

Multiple genes have been found to be involved in PC causation via genome-wide association studies and none of the genes found to be significant at the genome-wide level in these studies were found to carry a high risk variant. This point to the conclusion that common variants in DNA repair genes are might not explain the high genetic risk that CP patients that develop PC have and helps to

rationalize and further justify the rare and novel variant approach adapted over here. Follow up study that focuses on sequencing the tumor DNA of the CP+PC individuals to determine genotype status of the mutations detected in this study would help increase the susceptibility that these mutations are the risk variants that predispose CP individuals to risk of PC. It would also be advisable to follow up these mutations in PC individuals in general by genotyping matched normal and tumor DNA to determine how significant an effect these variants have in general PC population and finally functional studies that focus on creating mutant constructs with these mutations to compare the expression level of these genes to PC cell line expression of that particular mutant construct could help deduce the effect of these mutations at the cellular level. Comparing tissue histologies of pancreas cell line with these mutations and pancreatic cancer cell line will also help identify if these mutations truly have an effect at the normal functioning of the gene at the cellular level.

A major confounding factor that reduces the significance of the results found in this study is the significant difference in mean age of the two groups of patients. However the fact that germline mutations that have been investigated here do not change over time eliminates age as a confounding variable. Further the potentiality remains that some of these non cancerous patients may go on to develop PC in the future biasing the test towards null hypothesis of no significant difference in age and the effects further investigated over here could possibly be the same or even larger if the comparison group was confirmed cancer free older participants.

Although a relatively low number of patients were included in this study, the p values for the pathway based burden test as well the burden test of difference in number of non-synonymous variants were highly significant between the two groups of patients and thus are worthy of follow up in a exome sequencing study with a greater number of patients as well as in a larger PC patient cohort to see if the potentially high risk variants identified here are a risk in a general cohort of PC patients or not. Hence the findings of this pilot study should be treated with discretion and follow up studies are highly vital to establish the conclusions derived from this study as being generally applicable to all at risk PC patient cohorts. Further exome sequencing does not take into account epigenetic factors that could affect the expression of a gene regardless of sequence differences resulting in exclusion of one major factor that controls expression of cancer genes. It also worthwhile to note that with the advent of RNA sequencing, exome data is not always found to correlate with RNA sequencing data and hence follow up RNA sequencing studies would also be highly helpful. Finally factors like microRNA that might direct transcript degradation or control gene expression in cancer cells is not accounted by exome sequencing studies coupled with fact that evidence exists for RNA editing in tumors where single nucleotide variants are introduced into transcribed RNAs that is not present in exons of genomic DNA leads to the cautionary treatment of these findings until backed up by multiple validation studies.

APPENDIX A: LIST OF GERMLINE MUTATIONS IDENTIFIED IN PREVIOUS STUDIES

Table A1: Germline mutation among 159 genes of interest identified in previous studies

Gene	Mutation
ATM	c.8266A>AT p.K2756X c.170G>GA p.W57X c.3214G>GT p.E1072X c.6095G>GA p.R2032K IVS41-1G>GT c.3801delG
BRCA1	c.514delC p. Gln172AsnfsX62 c.1687C>T p.Gln563Stop c.3756_3759delGTCT p.Ser253ArgfsX10 c.5030_5033delCTAA p.Thr1677IlefsX2 185delAG 5382insC
BRCA2	c.514delC p.Gln172AsnfsX62 c.5796_5797delTA p.His1932GlnfsX12 c.6468_6469delTC p.Glu2157IlefsX18 6174delT 6672insT 6819delTG 4075delGT R2034C G3076E 10323delCins11 IVS 16-2A>G (splice acceptor site of intron 16) IVS 15-1G>A (splice donor site of intron 15) M192T K3326X 2458insT

Table A1 continued

CDKN2A/p16	p.E27X p.L65P c.201 ACTC>CTTT (promoter) p.G67R p.R144C p.G101W p.E27X -34G>T (initiation codon) c.47T>G p.L16R c.71G>C p.R24P c.192G>C L64L c.238_251del p.R80fs c.283del p.V95fs c.318G>A p.V106V c.457G>T D153spl (affects splicing) c.324T>A p.V95E c.482G>A p.A148T c.323_324insG p.E119X

Table A1 continued

MEN1	c.304G>T p.R102S c.723 to 724 del 320 CCC to C 68 CCC to CC 179 GAG to GTG c.249-252 del c.183G>A p.W61X c.196G>T p.V66F c.482delG c.1213C>T p.Q405X c.969C>A p.Y323X c.973G>C p.A325P 210-211insAGCCC c.712delA p.K201R c.CCT>CCGG, p.55fs64aaX c.GAG>AAG, p.E26K c.AGC>AAAC p. 66fs50aaX c. CGG>CAG p.R171Q c.CTG>CCG p.L168P c.GTG>GTTG p.236 fs12aaX c.TAT>TAG p.T268X c.GCC>CC p.437 fs15aaX c.GCA>G p.510fs19aaX c.CCG>GG p.493fs65aaX
MLH1	K618A
MSH2	Q402X G322D E205Q V367I c.1046C>T p.P349L c.1147C>T p.R383X
PALB2	c.1240C>T p.R414X c.508-9delAG p.R170I,183X c.3116delA, p.N1039fs heterozygous 6.7kb deletion of exon 12 & 13 c. 172-5delTTGT

APPENDIX B: ABBREVIATIONS USED IN TEXT

CP= Chronic pancreatitis

PC=Pancreatic cancer

HP=Hereditary pancreatitis

CP+PC= Documented chronic pancreatitis prior to pancreatic cancer

CP-PC= Only chronic pancreatitis and no pancreatic cancer

RR=Relative risk

PDAC= Pancreatic ductal adenocarcinoma

PanIN= Pancreatic intraepithelial neoplasia

NER= Nucleotide excision repair

NGS=Next generation sequencing

SE= Single end read

SD=Standard deviation

NCBI GRCh37/hg19= National Center for Biotechnology Human Genome reference genome version GRCh37 /UCSC genome browser genome version hg19

AN=coverage for a particular variant locus

AC= coverage for non-reference allele at a particular variant locus

AF= frequency of non-reference allele calculated as ratio of AC and AN converted as a percentage of 100

BIBLIOGRAPHY

1. National Cancer Institute [Internet]. Bethesda: National Cancer Institute; 2010 [cited May 2013]. Available from: http://seer.cancer.gov/csr/1975_2010/results_single/sect_01_table.01.pdf

2. Di Marco M, Di Cicilia R, Macchini M, Nobili E, Vecchiarelli S, Brandi G, et al. Metastatic pancreatic cancer: is gemcitabine still the best standard treatment? (Review). Oncology reports. 2010 May;23(5):1183-92. PubMed PMID: 20372829.

3. Conroy T, Desseigne F, Ychou M, Bouche O, Guimbaud R, Becouarn Y, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. The New England journal of medicine. 2011 May 12;364(19):1817-25. PubMed PMID: 21561347.

4. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer Journal international du cancer. 2010 Dec 15;127(12):2893-917. PubMed PMID: 21351269.

5. Maisonneuve P, Lowenfels AB. Epidemiology of pancreatic cancer: an update. Digestive diseases. 2010;28(4-5):645-56. PubMed PMID: 21088417.

6. Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie. 2008 Jul;393(4):535-45. PubMed PMID: 18193270.

7. Duell EJ. Epidemiology and potential mechanisms of tobacco smoking and heavy alcohol consumption in pancreatic cancer. Molecular carcinogenesis. 2012 Jan;51(1):40-52. PubMed PMID: 22162230.

8. Andreotti G, Silverman DT. Occupational risk factors and pancreatic cancer: a review of recent findings. Molecular carcinogenesis. 2012 Jan;51(1):98-108. PubMed PMID: 22162234.

9. Risch HA. Pancreatic cancer: Helicobacter pylori colonization, N-nitrosamine exposures, and ABO blood group. Molecular carcinogenesis. 2012 Jan;51(1):109-18. PubMed PMID: 22162235.

10. Sanchez GV, Weinstein SJ, Stolzenberg-Solomon RZ. Is dietary fat, vitamin D, or folate associated with pancreatic cancer? Molecular carcinogenesis. 2012 Jan;51(1):119-27. PubMed PMID: 22162236. Pubmed Central PMCID: 3496767.

11. Raimondi S, Lowenfels AB, Morselli-Labate AM, Maisonneuve P, Pezzilli R. Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. Best practice & research Clinical gastroenterology. 2010 Jun;24(3):349-58. PubMed PMID: 20510834.

12. Olson SH. Selected medical conditions and risk of pancreatic cancer. Molecular carcinogenesis. 2012 Jan;51(1):75-97. PubMed PMID: 22162233.

13. Li D. Diabetes and pancreatic cancer. Molecular carcinogenesis. 2012 Jan;51(1):64-74. PubMed PMID: 22162232. Pubmed Central PMCID: 3238796

14. Bracci PM. Obesity and pancreatic cancer: overview of epidemiologic evidence and biologic mechanisms. Molecular carcinogenesis. 2012 Jan;51(1):53-63. PubMed PMID: 22162231. Pubmed Central PMCID: 3348117.

15. Whitcomb DC. Inflammation and Cancer V. Chronic pancreatitis and pancreatic cancer. American journal of physiology Gastrointestinal and liver physiology. 2004 Aug;287(2):G315-9. PubMed PMID: 15246966.

16. Raimondi S, Lowenfels AB, Morselli-Labate AM, Maisonneuve P, Pezzilli R. Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. Best practice & research Clinical gastroenterology. 2010 Jun;24(3):349-58. PubMed PMID: 20510834.

17. Moskovitz AH, Linford NJ, Brentnall TA, Bronner MP, Storer BE, Potter JD, et al. Chromosomal instability in pancreatic ductal cells from patients with chronic pancreatitis and pancreatic adenocarcinoma. Genes, chromosomes & cancer. 2003 Jun;37(2):201-6. PubMed PMID: 12696069.

18. Yan L, McFaul C, Howes N, Leslie J, Lancaster G, Wong T, et al. Molecular analysis to detect pancreatic ductal adenocarcinoma in high-risk groups. Gastroenterology. 2005 Jun;128(7):2124-30. PubMed PMID: 15940643.

19. Baumgart M, Werther M, Bockholt A, Scheurer M, Ruschoff J, Dietmaier W, et al. Genomic instability at both the base pair level and the chromosomal level is detectable in earliest PanIN lesions in tissues of chronic pancreatitis. Pancreas. 2010 Oct;39(7):1093-103. PubMed PMID: 20531246

20. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008 Sep 26;321(5897):1801-6. PubMed PMID: 18772397. Pubmed Central PMCID: 2848990

21. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nature genetics. 2009 Sep;41(9):986-90. PubMed PMID: 19648918. Pubmed Central PMCID: 2839871.

22. Campa D, Rizzato C, Bauer AS, Werner J, Capurso G, Costello E, et al. Lack of replication of seven pancreatic cancer susceptibility loci identified in two Asian populations. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2013 Feb;22(2):320-3. PubMed PMID: 23250936.

23. McWilliams RR, Bamlet WR, de Andrade M, Rider DN, Cunningham JM, Petersen GM. Nucleotide excision repair pathway polymorphisms and pancreatic cancer risk: evidence for role of MMS19L. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2009 Apr;18(4):1295-302. PubMed PMID: 19318433. Pubmed Central PMCID: 2710767.

24. Li D, Suzuki H, Liu B, Morris J, Liu J, Okazaki T, et al. DNA repair gene polymorphisms and risk of pancreatic cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2009 Jan 15;15(2):740-6. PubMed PMID: 19147782. Pubmed Central PMCID: 2629144.

25. Crnogorac-Jurcevic T, Efthimiou E, Nielsen T, Loader J, Terris B, Stamp G, et al. Expression profiling of microdissected pancreatic adenocarcinomas. Oncogene. 2002 Jul 4;21(29):4587-94. PubMed PMID: 12085237.

26. Mathews LA, Cabarcas SM, Hurt EM, Zhang X, Jaffee EM, Farrar WL. Increased expression of DNA repair genes in invasive human pancreatic cancer cells. Pancreas. 2011 Jul;40(5):730-9. PubMed PMID: 21633318. Pubmed Central PMCID: 3116046.

27. Maple JT, Smyrk TC, Boardman LA, Johnson RA, Thibodeau SN, Chari ST. Defective DNA mismatch repair in long-term (> or =3 years) survivors with pancreatic cancer. Pancreatology : official journal of the International Association of Pancreatology. 2005;5(2-3):220-7; discussion 7-8. PubMed PMID: 15855819.

28. Nyaga SG, Lohani A, Evans MK. Deficient repair of 8-hydroxyguanine in the BxPC-3 pancreatic cancer cell line. Biochemical and biophysical research communications. 2008 Nov 14;376(2):336-40. PubMed PMID: 18774780. Pubmed Central PMCID: 2699024.

29. Akita H, Zheng Z, Takeda Y, Kim C, Kittaka N, Kobayashi S, et al. Significance of RRM1 and ERCC1 expression in resectable pancreatic adenocarcinoma. Oncogene. 2009 Aug 13;28(32):2903-9. PubMed PMID: 19543324.

30. Maithel SK, Coban I, Kneuertz PJ, Kooby DA, El-Rayes BF, Kauh JS, et al. Differential expression of ERCC1 in pancreas adenocarcinoma: high tumor expression is associated with earlier recurrence and shortened survival after resection. Annals of surgical oncology. 2011 Sep;18(9):2699-705. PubMed PMID: 21360249.

31. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews Genetics. 2011 Nov;12(11):745-55. PubMed PMID: 21946919.

32. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Annals of neurology. 2012 Jan;71(1):5-14. PubMed PMID: 22275248.

33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. PubMed PMID: 19451168. Pubmed Central PMCID: 2705234.

34. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008 Nov;18(11):1851-8. PubMed PMID: 18714091. Pubmed Central PMCID: 2577856.

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. PubMed PMID: 19505943. Pubmed Central PMCID: 2723002.

36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8. PubMed PMID: 21653522. Pubmed Central PMCID: 3137218.

37. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Annals of neurology. 2012 Jan;71(1):5-14. PubMed PMID: 22275248.

38. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, et al. Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome research. 2012 Feb;22(2):208-19. PubMed PMID: 22156295. Pubmed Central PMCID: 3266029.

39. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, et al. Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome research. 2012 Feb;22(2):208-19. PubMed PMID: 22156295. Pubmed Central PMCID: 3266029.

40. Roberts NJ, Jiao Y, Yu J, Kopelovich L, Petersen GM, Bondy ML, et al. ATM mutations in patients with hereditary pancreatic cancer. Cancer discovery. 2012 Jan;2(1):41-6. PubMed PMID: 22585167. Pubmed Central PMCID: 3676748.

41. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10;461(7261):272-6. PubMed PMID: 19684571. Pubmed Central PMCID: 2844771.

42. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092. Pubmed Central PMCID: 3042601.

43. Solomon S, Das S, Brand R, Whitcomb DC. Inherited pancreatic cancer syndromes. Cancer journal. 2012 Nov-Dec;18(6):485-91. PubMed PMID: 23187834. Pubmed Central PMCID: 3565835.

44. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome research. 2005 Jul;15(7):901-13. PubMed PMID: 15965027. Pubmed Central PMCID: 1172034.

45. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010 Sep;38(16):e164. PubMed PMID: 20601685. Pubmed Central PMCID: 2938201.