Identification Of Copy Number Variations (CNVs) Of Epigenetic Factors Related to The Progression of Pancreatic ductal adenocarcinoma (PDAC)

Pavithra Raju

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Recommended Citation
Raju, Pavithra, "Identification Of Copy Number Variations (CNVs) Of Epigenetic Factors Related to The Progression of Pancreatic ductal adenocarcinoma (PDAC)" (2023). Master's Projects. 1321.
https://scholarworks.sjsu.edu/etd_projects/1321
Identification Of Copy Number Variations (CNVs) Of Epigenetic Factors Related to The
Progression of Pancreatic ductal adenocarcinoma (PDAC)

A project

Presented to

The Faculty of the Department of Computer Science
San José State University

In Partial Fulfilment

of the Requirements for the Degree

Master of Science

By

Pavithra Raju

December 2023
Identification Of Copy Number Variations(CNVs) Of Epigenetic Factors Related to The Progression of Pancreatic ductal adenocarcinoma(PDAC)

by

Pavithra Raju

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

December 2023

Dr. Leonard Wesley  Department of Computer Science

Dr. Philip Heller  Department of Computer Science

Dr. William Andreopoulos  Department of Computer Science
ABSTRACT

Identification Of Copy Number Variations (CNVs) Of Epigenetic Factors Related to The Progression of Pancreatic ductal adenocarcinoma (PDAC)

by Pavithra Raju

Pancreatic ductal adenocarcinoma (PDAC) is a formidable challenge in oncology due to its aggressive form and late-stage detection. PDAC is known to be influenced by various epigenetic factors like DNA methylation and histone modifications. This study focuses on copy number variations (CNVs) within epigenetic factors which for their role in early diagnosis. Thus, paving the way for identification of potential biomarkers. The epigenetic pipeline was extended based on CNVs and the CNV modified sequences extracted were compared with the wild type sequences of epigenetic PDAC genes. The epigenetic gene KCNJ11 with copy number gain of CNV id 46771406 was used and its CNV modified sequence was compared with wild type sequences of genes RPS2P40, NCR3LG1 and KCNJ11 respectively. This CNV was found to be significant in genes RPS2P40, KCNJ11 and non-significant in gene NCR3LG1.

Keywords: Pancreatic ductal adenocarcinoma, copy number variations, epigenetic factors, epigenetic pipeline, CNV modified sequence.
ACKNOWLEDGMENTS

I would like to thank Dr Leonard Wesley for giving me this opportunity to work under him and on this project to produce efficient results in PDAC early diagnosis. I would also like to thank Dr. Philip Heller and Dr. William Andreopoulos for volunteering and being a part of my master’s thesis committee.
TABLE OF CONTENTS

CHAPTERS

Introduction .................................................................................................................. 1

Background .................................................................................................................. 4

2.1 Detection with Serum Biomarkers ....................................................................... 5
2.2 Detection with Genomic Biomarkers .................................................................. 6
2.3 Circulating Tumor Cells ....................................................................................... 8
2.4 Sequencing ........................................................................................................... 9
2.5 Role of CNVs of in PDAC detection ................................................................... 10
2.6 Remaining technical Gaps .................................................................................. 11

Approach and Method ................................................................................................. 12

3.1 Data ..................................................................................................................... 13
3.1.1 Pancreatic Genes .......................................................................................... 13
3.1.2 Epigenetic PDAC genes ................................................................................ 13
3.1.3 Enrichr python module ................................................................................ 13
3.2 Epigenetic Pipeline .............................................................................................. 13
3.2.1 CNV module Development .......................................................................... 14
3.2.2 Environment .................................................................................................. 15
3.3 CNV genes considered for clinical significance ................................................ 16
Result .................................................................................................................................19

4.1 Enrichr module result .................................................................................................. 19

4.2 CNV module result ...................................................................................................... 19

4.3 Significance results .................................................................................................... 22

Discussion .......................................................................................................................... 25

Conclusion .......................................................................................................................... 27

Future Research .................................................................................................................. 28

LIST OF REFERENCES ...................................................................................................... 29
LIST OF TABLES

Table 1: Enrichr module results.........................................................................................19
LIST OF FIGURES

Figure 1: Flowchart of methods followed ................................................................. 18
Figure 2: CNV module results .................................................................................. 20
Figure 3: CNV ID - 46771406 results ..................................................................... 20
Figure 4: Significance prediction - RPS2P40 result .................................................... 22
Figure 5: Significance prediction - NCR3LG1 results ............................................... 23
Figure 6: Significance prediction - KCNJ11 results .................................................... 24
CHAPTER 1

Introduction

PDAC is the most prevalent form of pancreatic cancer [1]. It is the fourth most common cause of oncology related deaths in Unites States and further predicted to be the second cause by 2030 [1][2]. It is an aggressive form of cancer often detected at later stages with poor prognosis as it lacks or shows no prominent symptoms in its early stages [2][3][5]. It has a low five-year survival rate of less than 10% with 80% of patients diagnosed at higher stages[2][4]. Therefore, early detection is very important in treatment of PDAC patients as it is still localized [4]. Continuous research efforts are being made and explored at the genome level for early diagnosis of PDAC cancer.

One such field being explored is the role of epigenetics in early diagnosis of PDAC. Epigenetics provides information on different gene expression profiles by controlling the access of transcription machinery to target genes without modifying the DNA sequences [7]. The epigenetic modifications which play an important role in PDAC progression are aberrant DNA methylations and post translational histone modifications as they exert substantial effects on gene expression and cellular functions [8][9]. In PDAC, irregular epigenetic changes contribute to the dysregulation of critical pathways involved in cell proliferation, apoptosis and metastasis [4]. Therefore, understanding these epigenomic changes is essential for the development of potential biomarkers and for targeted therapeutic regimens.

Once the genes which undergo epigenetic changes and play an important role in PDAC have been identified. These genes are further analyzed for the identification of mutations on their DNA
sequences. These mutations can be single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertions and deletions (Indel) etc. CNVs are deletions or duplications of size varying from 100bp to 3Mb in the genome[11]. These variations correspond to changes in gene expression and is of functional importance[11]. In PDAC, CNVs have been observed in genes associated with cell cycle regulation, DNA repair and chromatin remodeling[9][10]. Exploration of CNVs in the context of epigenetic factors provides a nuanced understanding of the genomic landscape and its impact on PDAC pathogenesis.

The rationale for investigating CNVs of epigenetic factors in PDAC is multifaceted. Firstly, it addresses the gap in knowledge regarding the genomic alterations affecting key epigenetic regulators in PDAC[12]. Secondly, it offers insights into the functional consequences of these CNVs on the epigenetic landscape and downstream gene expression[12]. Thirdly, the discovery and determination of significant CNVs plays an important role in PDAC early prognosis[12]. They could be considered as potential therapeutic targets and prognostic markers to enhance patient outcomes and, improve mortality rates due to faster access of medicines and tumor specific treatments [13].

The primary objective of my project is to systematically identify and characterize CNVs of key epigenetic factors in PDAC. This study also bridges the gap between genomic alterations and epigenetic dysregulation in PDAC[13]. By elucidating the specific CNVs affecting epigenetic factors, we aim to provide a foundation for targeted therapies tailored to the individual genomic and epigenomic landscape of PDAC patients. Additionally, the identification of CNV based prognostic makers can inform clinical decision making and improve patient outcomes.

Existing epigenetic pipeline was used for the identification and characterization of CNVs. The epigenetic pipeline processes the epigenetic genes involved in PDAC and gives information
on the specific SNPs involved. This epigenetic pipeline was further extended to obtain information on specific CNVs of the epigenetic genes involved in PDAC. These extracted CNVs from the pipeline were further analyzed and explored for their functional importance and annotation. Based on the functional change and annotation, certain CNVs were determined which can be utilized as predictive indicators in PDAC progression. The entire process represents a crucial step towards unraveling the complexities of this lethal malignancy. This study holds the promise of advancing our understanding of PDAC pathogenesis, offering new avenues for therapeutic interventions, and ultimately improving the diagnosis for individuals affected by this devastating disease.
CHAPTER 2

Background

Pancreatic ductal adenocarcinoma contributes to 90% of pancreatic cancers [17]. This disease is one of the major leading cause of oncology related deaths because 80-85% of patients are prognosed at advanced stages when it is spread to the surrounding areas or organs in the body [15]. Therefore, surgical resection at this stage is not a viable option. Nonetheless, there exists a considerable amount of time span of one to two decades between the diagnosis of pancreatitis and the manifestation of a clinically evident tumor [16]. This duration can be utilized to improve the patient survival rates through early diagnosis and prompt initiation of treatment.

Detection of pancreatic cancer especially in the early stages proves to be a formidable challenge. This is primarily due to the inconspicuous nature of the symptoms. Even when symptoms do emerge, they are often subtle and easily mistaken for other illnesses. Early indicators such as weight loss, loss of appetite, abdominal pain, jaundice, heartburn, nausea, dysgeusia, bloating, etc., may not immediately raise concerns often leading to delayed diagnosis [14]. The difficulty intensifies in later stages when symptoms persist yet can be easily perplexed with other disorders. This underscores the pressing need for innovative approaches that enable early detection.

Medical imaging becomes a crucial component when pancreatic cancer is suspected, typically leading to the discovery of masses [18]. However, these cancerous masses detectable through imaging are mostly indicative of advanced stage of the disease [18]. Thus, highlighting the critical importance of pioneering methods that facilitate earlier diagnosis. Early prognosis is paramount
for improving patient outcomes as interventions become more limited at later stages. Thereby significantly reducing the survival rates.

To overcome the limitations posed by late-stage detection, one promising strategy involves exploring biomarkers associated with pancreatic cancer. However, studying these biomarkers for early prognosis face the significant hurdle of the scarcity of samples collected before a diagnosis is even made [19]. Pancreatic cancer patients’ samples contribute valuable insights in determination of potential biomarkers [19]. However, samples collected before the onset of the disease are preferred for a comprehensive understanding of biomarker dynamics [19]. Addressing this challenge is pivotal for advancing research in biomarker identification and developing reliable tools for early detection.

2.1 Detection with Serum Biomarkers

Serum biomarkers are indicators or substances found in the blood serum that can be used to measure, analyze and indicate the presence of various physiological and pathological conditions inside the body. They can be cell molecules like proteins, antigens in PDAC patients and their presence in high or low quantities can indicate a disease state. Hence, they play a crucial role in diagnosing, monitoring and predicting various health conditions.

Identification of biomarkers stands at the forefront of early pancreatic cancer detection, where carbohydrate antigen 19-9 (CA19-9) takes center stage as the primary serum biomarker. Demonstrating sensitivity between 79% to 81%, with a specificity ranging from 82% to 90%, CA19-9 has been a focus of extensive research [20][21]. O’Brien et al.’s study, utilizing serum samples from women enrolled in an ovarian cancer study but later diagnosed with pancreatic cancer, showcased CA19-9 levels were evident in pancreatic ductal adenocarcinoma (PDAC) cases
within 12 months of diagnosis, with two cases even revealing increased levels three years prior to PDAC diagnosis [19]. The study also explored the role of other proteins CA125, CEACAM1 and REG3A in early PDAC prognosis. It was found that the combination of CA 19-9 and CA 125 had high sensitivity and could be used as early prognostic tools, whereas CEACAM1 and REG3A proved to be late markers in PDAC detection [19].

Despite its prominence, CA19-9 has limitations, being detected in other organ tumors such as colorectal cancer, gastric cancer and endometrial cancer, rendering it non-specific to pancreatic cancer [27]. Approximately 5%-10% of the Caucasian population faces challenges in producing this protein resulting in false negatives in them [28]. While CA19-9 alone cannot confirm pancreatic cancer diagnosis, its correlation with the disease makes it valuable for post-diagnosis management [19]. Research suggests promise in exploring CA19-9 levels in the years leading up to the disease, emphasizing the need for additional biomarkers to enhance diagnostic accuracy [19][28].

2.2 Detection with Genomic Biomarkers

Genomic biomarkers provide information on the genetic constitution and alterations within an individual’s DNA and RNA. They play an important in disease diagnosis, prognosis and personalized medicine based on gene expression profiles. The explorations of mutations causing pancreatic cancer involves a focus on specific biomarkers in patient tissue specimens. Traditional biomarker detection methods, like tissue biopsy, involve surgical or needle biopsy sectioning of patient tissues [22]. But liquid biopsy is non-invasive and gives access to circulating tumor cells (CTCs), cell-free circulating nucleic acids (cfNAs) and exosomes [22]. CfNAs contain information about somatic mutations, miRNAs, DNA methylation and more.
The levels of circulating free DNA (cfDNA) exhibited variability among different types and stages of pancreatic cancer in a study noted by Qi et al, 2018. Analyses of cfDNA have revealed mutations in genes associated with pancreatic cancer. Zill et al, 2015 investigated 54 common genes found in PDAC patients and reported that 90% of mutations found in tumor biopsies were also detected in cfDNA, suggesting the potential for highly specific and sensitive cfDNA detection. Notably, Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS) mutations were identified in nearly 50% of PDAC patients’ plasma with minimal occurrences in healthy donors indicating the potential of KRAS mutations in circulating tumor DNA (ctDNA) as a viable biomarker [22].

MicroRNA (miRNA), a specific type of circulating free nucleic acid (cfNA) demonstrates the required specificity and sensitivity to serve as a biomarker for pancreatic cancer. They play an important role in gene expression regulation in oncology environments using tumor suppressor functions [30]. Furthermore, distinct miRNA patterns were observed in the saliva and stool samples of PDAC patients exhibited elevated levels of specific miRNA 21 compared to those of normal controls [22].

However, Qi et al, 2018 argue that reliance on a single biomarker or studying a lone biopsy sample may provide limited information due to tumor heterogeneity and genetic composition. They advocate for a more effective technique to monitor new biomarkers, considering the limited true positive and true negative rates of biomarkers like carcinoembryonic antigen (CEA) and carbohydrate antigen [22]. Liquid biopsy emerges as a promising approach due to its non-invasiveness and efficacy in detecting circulating tumor cell (CTCs) and cell free circulating nucleic acids (cfNAs) present in body fluids like blood [22]. Nevertheless, a significant drawback of the liquid biopsy technique is its dependence on body fluids rather than cancerous tumors for biomarker detection, potentially leading to misleading conclusions [22].
Integrin alpha 1 (ITGA1) has been identified as an ideal biomarker for diagnosing and treating pancreatic ductal adenocarcinoma (PDAC) in a study [24]. Elevated ITGA1 levels were observed in 42% of PDAC patient tumor issues and were found to be absent in normal pancreatic ductal epithelial cells [24]. This indicated its effective utilization in PDAC detection. In another study various genetic mutations were explored to pinpoint Kirsten rat sarcoma (KRas) as the most frequently mutated gene in PDAC cases [23]. Other mutated genes included p53, p16, SMAD4, PUC1, and SRC leading to inappropriate cell proliferation in the pancreas upon mutation [23]. P53, p16, and SMAD4 are tumor-suppressing transcription factors in the normal state that inhibit cell cycle progression [23]. However, mutations or absence of these factors resulted in uncontrolled cell proliferation. Thus, marking these transcription factors as crucial biomarkers in PDAC treatment [23].

Pseudopodium-Enriched Atypical Kinase One (PEAK1) was also determined as a novel biomarker, central signaling nexus and a recent therapeutic target in PDAC [25]. PEAK1 regulates cell migration and proliferation, playing a pivotal role in restricting cancer cell migration and growth [25]. However, a significant limitation of biomarker studies lies in their focus on a limited set of genetic materials. This potentially proves their inconsistency, given the diverse set of genetic mutations present in the tumors. Furthermore, not all biomarkers contribute to early pancreatic cancer detection, except for a few specific genes [22].

2.3 Circulating Tumor Cells

Circulating Tumor Cells (CTCs) are cancer cells that have detached from a primary tumor and entered the blood stream. These cells play a crucial role in the metastatic spread of cancer. CTCs have risen to prominence as a promising biomarker in cancer research, particularly in the realm of pancreatic cancer. In a study which involved investigating KRAS mutations within CTCs
from pancreatic cancer patients, it was found that these mutations in the KRAS gene were present in the CTCs of eleven out of twelve patients [31]. The hematopoietic cells from the same group showed no mutant KRAS sequences [31]. This study emphasized on the importance of the presence at least ten CTCs for proper detection of KRAS mutations as detection rates dropped significantly below this threshold [31].

Detecting CTCs in the blood stream poses inherent challenges due to their low concentration with an estimated 1 – 50 CTCs present in a 7.5 ml blood sample containing over a million white blood cells [31]. Precise detection requires tests with high sensitivity and specificity amidst this cellular complexity [31]. Highlighting the rarity of CTCs, certain studies have found the presence of approximately one CTC per $10^9$ hemocytes in a cancer patient’s blood sample [22]. Despite the minute fraction of just 0.01% of CTCs progressing to metastases, the potential of CTCs as valuable biomarkers in understanding cancer progression persists [31]. Ongoing research aims to develop advanced techniques for the precise detection of these elusive circulating cells.

2.4 Sequencing

DNA and RNA sequencing technologies have empowered scientists to find the differences at the genetic level between the normal and malignant tissues. Sequencing is the process in which the nucleotide arrangement of DNA and RNA sequences are found. Next generation sequencing (NGS) plays an important role in oncology research and therapies. Identification of genetic mutations for disease prognosis is one of its main applications. A study including 23 pancreatic cancer patients found that the detection rates were high for main mutated genes with an NGS success rate of 76.7% [33]. These main mutated genes were KRAS, TP53, CDKN2A and SMAD4 [33].
NGS has also been used to detect low allele frequency mutations from PDAC patient plasma samples [34]. The most found significant mutations were on genes KRAS and SMAD4 [34]. The study also revealed that circulating tumor DNA (ctDNA) serves as a prognostic biomarker as it is intricately linked to cancer progression and the extent of tumor differentiation [34]. Thus, blood-based assays are important in genetic detection of PDAC. In another study whole exome sequencing was employed for the determination of significant prognostic genes and precise therapeutics [32]. KRAS mutations were present in almost 90% of 109 patient samples collected [32]. The genetic landscape also revealed frequent alternations in major cell signaling pathways like Wnt signaling, Hedgehog signaling, DNA repair and cell cycle processes [32]. These findings reveal novel genetic diversity in PDAC by offering insights into prognostic factors and potential therapeutic targets.

2.5 Role of CNVs of in PDAC detection

In a study conducted by Sukhni et al, 2012, familial pancreatic cancer (FPC) genes were determined by utilization of genomic copy number variations (CNVs) in high-risk patients and then compared with control group. 120 FPC cases were used along with 1,194 normal control group [35]. In the results, no major changes were observed in the germline CNVs between FPC and normal cases [35]. 93 FPC-specific CNVs each unique to individual were determined in 50 cases [35]. These CNVs were found in the exons of 88 reference sequence database (RefSeq) genes which were known to play a role in PDAC expression profiles [35]. These findings suggested a potential link between specific genes affected by CNVs and genetic predisposition susceptibility to pancreatic cancer.

Another study by Willis et al, 2014 shed light on the importance of single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) by usage of genome wide study and
analysis in PDAC. This study was performed for the determination of novel CNVs as well as total CNVs for their role in PDAC patients [36]. The study used 223 cancer cases and 169 normal cases for identification of germline CNVs [36]. From the total CNV analysis it was found that these CNVs did not play a significant role in pancreatic cancer risk [36]. Further no associations between specific CNVs and pancreatic cancer risk were found [36]. Thus, this study shows that CNVs may not be a factor in the genetic origins of pancreatic cancer, but it may be due to the small sample size considered.

Huang et al, 2012 delves into the functional implications of Copy Number Variations Region 2966.1 (CNVR2966.1) and its potential connection with susceptibility to pancreatic cancer. CNVR2966.1 was identified as 10,379 bp nucleotide deletion/insertion present on chromosome 6 [37]. This study found that CNVR2966.1 functioned as an active regulator with upstream interactions with gene CDKN2B and was conformed through various gene assays [37]. Presence of higher copy numbers showed increased CDKN2B gene expression when compared with lower copy numbers between 1027 cancer patients and 1031 normal people [37]. The odds ratio also had 95% confidence interval inferring a strong link between this CNV and pancreatic cancer risk [37].

2.6 Remaining technical Gaps

Current research shows promising results in exploring CNV (Copy Number Variations) as novel biomarkers in PDAC early prognosis. The current epigenetic pipeline provides information on the single nucleotides polymorphism (SNPs) involved in PDAC. This project aims to further extend the pipeline based on copy number variations (CNVs) by developing a CNV module for the pipeline which can extract information from the NCBI dbVar database. These CNVs can be further processed in the pipeline for their significance in early detection of PDAC.
CHAPTER 3

Approach and Method

In this project we aim to identify the significant CNVs of epigenetic genes involved in early diagnosis of PDAC. Significant CNVs can control gene expression profiles and affect the downstream processes ultimately leading to functional changes. These epigenomic signatures can be used as prospective biomarkers in PDAC enabling early access to tumor specific targeted therapies for patients. As most PDAC are diagnosed at later stages, these biomarkers play a vital role in early screening and prognosis. Therefore, this project aims to bridge the gap between genomic alterations using CNVs and the epigenetic dysregulation of genes involved in PDAC progression.

The biomarker used in this project is mainly copy number variations (CNVs). The epigenetic genes involved in pancreatic cancer were obtained from the Enrichr module developed by San Jose State University Professor Leonard Wesley [40]. NCBI dbVar database was used to obtain information on the CNVs of PDAC related epigenetic genes. The epigenetic pipeline developed by San Jose State University Professor Leonard Wesley, his previous and current project students is used to extract single nucleotide polymorphism(SNP) information of the epigenetic genes in PDAC progression [39]. This epigenetic pipeline is further developed and extended to extract copy number variation (CNV) information of epigenetic genes involved in PDAC diagnosis. The CNV modified sequences are extracted and then compared with their respective wild type of DNA sequences for the identification of CNVs. Further these CNVs are analyzed for their functional significance and reported in results. These significant CNVs are useful as promising indicators in early PDAC screening and detection.
3.1 Data

3.1.1 Pancreatic Genes

Various genes were identified which were associated with normal functioning of pancreas. The commonly found genes in pancreas included INS, GCG, AMY2A, PNLIP, PRSS1, CTRC, SLC30A8, GCK, KCNJ11 and PDX1 whose functions mostly involved in maintaining blood sugar levels in humans. These genes were involved in most of the functions of the pancreas.

3.1.2 Epigenetic PDAC genes

The genes associated with normal function of pancreas were further subjected to epigenetic analysis. This was performed by providing the normal genes as input to the Enrichr python module so that the associated epigenetic genes and their epigenetic modifications were identified and provided as output.

3.1.3 Enrichr python module

This python code was developed by San Jose State University Professor Leonard Wesley. It was based on the Enrichr database which gives information on the epigenetic modifications a gene undergoes and its associated epigenomic genes [40]. Enrichr database was developed in the Ma’ayan Lab by Avi Ma’ayan and his associates at Center for Bioinformatics in Ichan School of Medicine at Mount Sinai. Epigenomic genes control the expression of genes though their modifications like DNA methylations, histone modifications, non-coding RNA, etc. This module was used to identify the epigenetic genes involved in PDAC and their respective modifications.

3.2 Epigenetic Pipeline

The epigenetic pipeline was already developed by San Jose State University Professor Leonard Wesley and his project students. It takes the epigenetic genes as input and processes them
to identify the SNPs involved in PDAC by extracting information from the NBCI dbSNP database [39]. It also has a significant protein prediction module which predicts the significant changes by taking two or more DNA or RNA sequences as input, produces their respective protein sequences and compares them [39]. If there are significant changes present that affects the protein functions, then this module predicts these changes to be significant.

The epigenetic pipeline also contains two major dictionaries. These are chromosome dictionary and gene code fasta dictionary. The chromosomes contain all the chromosome numbers as keys and their entire nucleotide sequences as values [39]. The gene code fasta dictionary contains most of the gene names as keys and their respective wild type of nucleotide sequences as values [39]. Both dictionaries are very large in size. The pipeline also has a significant protein prediction module which takes two or more DNA or RNA sequences to be compared and converts them to their respective protein sequences or retains as DNA sequences, looks for differences between them and based on the important metrics involved predicts their significance [39]. This module predicts the change in protein sequences to be significant only if their functions change.

In this project, the epigenetic pipeline was further developed and extended to obtain information about the CNVs of the epigenetic PDAC genes to be processed. The CNV module of this pipeline was developed by extracting information from the NCBI dbVar database and finally to be integrated into the main pipeline. The epigenetic PDAC genes to be processed were given as input. This was followed by running the newly developed CNV module and obtaining the CNV results. These CNV results were further analyzed for their promising role in early PDAC prognosis.

3.2.1 CNV module Development

CNV module code was developed to process genes present in the epigenetic pipeline and then extract their respective CNV information from NCBI dbVar database. This module was developed
to be ultimately integrated into the main epigenetic pipeline. For each gene processed in the pipeline, a list of CNV ids, CNV type and CNV location on the respective chromosomes were obtained. CNV location information was used to extract the CNV modified sequence from their respective chromosomes and then compared with the wild type of DNA sequences of the genes involved for their significance as a potential biomarker in PDAC early detection.

### 3.2.2 Environment

The epigenetic pipeline can be used to run on our local computer systems and Github is used for version control. Github repository is used to store all the information of the pipeline. The repository contains the data files or input files, python codes for the different modules present and integrated into the main pipeline and finally, the readme files with instructions to execute the modules in the pipeline correctly.

PyCharm version 3.2 was used as the integrated development environment for successfully running the pipeline. The pipeline was further developed and extended by creating the CNV module using the same environment. All the python libraries required to run the pipeline was downloaded using the terminal in PyCharm.

For the development of the CNV module of the pipeline Entrez module from Biopython package was imported. Biopython provides modules to access online biological databases like NCBI, retrieve data from them and perform bioinformatic analyses. Entrez module in Biopython provides Entrez utilities which allows for the interaction with NCBI database. These utilities enable to search, retrieve and link information from various biological databases maintained by NCBI. BeautifulSoup python library was used for website scraping of specific and important information from NCBI dbVar database. BeautifulSoup can also be used for parsing XML or HTML documents.
3.3 CNV genes considered for clinical significance

After obtaining the CNV information as output from CNV module of the pipeline. Only one CNV id 46771406 was considered for downstream analysis of epigenetic gene KCNJ11. This gene KCNJ11 had histone modifications and was found to have number gain as the type of CNV. This CNV type was found to be present in many genes. But only three genes were considered for assessment of their role in potential significance. These three genes considered were RPS2P40, NCR3LG1 and KCNJ11. The functions of these genes were found as follows –

1. **RPS2P40** – Ribosomal protein S2 pseudogene 40 is present in humans on chromosome 11 at location 11p15.1. This gene encodes ribosomal protein which is part of the 40s subunit and is a pseudogene. It is also known as RPS2_17_1125

2. **NCR3LG1** - Natural killer cell cytotoxicity receptor 3 ligand 1 is present in humans on chromosome 11 at location 11p15.1 with an exon count of 7. It is selectively expressed in tumor cells. The interaction of this gene with NKp30 results in natural killer (NK) cell activation and cytotoxicity. It is also known as B7H6 and DKFZp686O24166.

3. **KCNJ11** - Potassium inwardly rectifying channel subfamily J member 11 is present in humans on chromosome 11 at location 11p15.1 with an exon count of 4. The protein encoded by gene has higher affinity to allow inflow potassium ions to a cell than outside of it. This gene is controlled by G-proteins and is found associated with the sulfonylurea receptor SUR. Mutations present on this gene may lead to unregulated insulin secretions. It is also known as BIR, HHF2, PHHI, IKATP, PNDM2, TNDM3, KIR6.2 and MODY13.

The modified CNV sequence and the wildtype sequences of all the three above mentioned genes were extracted and given as input to the significant protein prediction module of the epigenetic pipeline. This module converts the respective DNA sequence into protein sequences
or retains as DNA sequences, compares them, looks for important differences and based on certain metrics predicts if the changes present on protein sequences are causing functional change followed by final output of clinical significance or non-significance.
Identification of significant CNVs as potential biomarker for early PDAC detection

Figure 1: Flowchart of methods followed
CHAPTER 4

Result

4.1 Enrichr module result

The initial set of genes considered for pancreas functioning were INS, GCG, AMY2A, PNLIP, PRSS1, CTRC, SLC30A8, GCK, KCNJ11 and PDX1. Their associated epigenetic related genes were found to be as in the Enrichr module results. Only first twelve epigenetic genes were considered for downstream analysis. These twelve genes considered were MAFB, KCNJ11, ABCC8, ARX, PDX1, HNF1B, PPARG, GATA4, SOX9, PTF1A, NKX6-1, FOXA2 as results from the Enrichr module with histone modifications as the epigenetic factors.

<table>
<thead>
<tr>
<th>rank</th>
<th>description</th>
<th>genes</th>
<th>adjusted_p</th>
<th>gene_set_library</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H3K27me3 ES-bruced mm9</td>
<td>MAFB*, KCNJ11*, ABCB2*, ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.00750943</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>2</td>
<td>H3K27me3 kidney mm9</td>
<td>TRPIL, MAFB*, ONECUT1*, ARX*, PDX1*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.00750943</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>3</td>
<td>H3K27me3 cerebellum mm9</td>
<td>TRPIL, ONECUT1*, ARX*, PDX1*, HNF1B*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.01003332</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>4</td>
<td>H3K27me3 B1 hg19</td>
<td>MAFB*, KCNJ11*, ONECUT1*, ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.02749562</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>5</td>
<td>H3K27me3 Sk-N-Sh hg19</td>
<td>MAFB*, ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.0282886</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>6</td>
<td>H3K27me3 testis mm9</td>
<td>TRPIL, ONECUT1*, PDX1*, HNF1B*, PPARG*, GATA4*, HNF1A*, SOX9*, NKX6-1*, FOXA2*</td>
<td>0.04767415</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>7</td>
<td>H3K27me3 thymus mm9</td>
<td>ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.04767415</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>8</td>
<td>H3K27me3 fibroblast of lung</td>
<td>TRPIL, ABCB2*, ONECUT1*, ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.09583219</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>9</td>
<td>H3K27me3 myocyte mm9</td>
<td>TRPIL, KCNJ11*, ABCB2*, ONECUT1*, PAX4*, HNF1B*, CTRC*</td>
<td>0.09609871</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>10</td>
<td>H3K27me3 C2c11 mm9</td>
<td>TRPIL, ABCB2*, ONECUT1*, ARX*, PDX1*, HNF1B*, HNF1A*</td>
<td>0.11893313</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>11</td>
<td>H3K27me3 spleen mm9</td>
<td>TRPIL, KCNJ11*, ABCB2*, ONECUT1*, PAX4*, HNF1B*, CTRC*</td>
<td>0.11893313</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>12</td>
<td>H3K27me3 bronchial epithel</td>
<td>TRPIL, ABCB2*, HNF1A*, ARX*, PDX1*, PAX4*, HNF1B*, CTRC*</td>
<td>0.14249246</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>13</td>
<td>H3K27me3 cardiac meso</td>
<td>ABCB2*, ONECUT1*, ARX*, PDX1*, HNF1B*, PPARG*, GATA4*, SOX9*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.17990835</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>14</td>
<td>H3K27me3 kidney epithelia</td>
<td>TRPIL, ABCB2*, ONECUT1*, ARX*, PDX1*, HNF1B*, GATA4*, PTF1A*, NKX6-1*, FOXA2*</td>
<td>0.18408490</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
<tr>
<td>15</td>
<td>H3K27me cerebellum mm9</td>
<td>YLF11*, MAFB*, KCNJ11*, PDX1*, HNF1B*, GATA4*, PTF1A*, NKX6-1*</td>
<td>0.2917164</td>
<td>ENCODE_Histone_Modifications_2015</td>
</tr>
</tbody>
</table>

Table 1: Enrichr module results

4.2 CNV module result

These epigenetic genes were further processed for their CNV information by giving them as input to the CNV module. The CNV module obtained information for each gene and gave the respective list of CNV ids for each gene processed. Each CNV id was further processed to obtain
information on the type of CNV, CNV location information on chromosome, variant region id, study id, object type, variant call count, sort order, organism, gene list, method used and clinical significance as the CNV module output.

Figure 2: CNV module results

For genes FOXA2 and MAFB the CNV module found no CNV data. The rest of the ten genes were processed and their completed CNV information was successfully obtained. Among them, only KCNJ11 gene’s CNV information obtained was used for further analysis. The CNV module of the epigenetic pipeline returned a lists of CNV ids for gene KCNJ11. The CNV ids were 50144542, 50057890, 46771406, 46615600, 28888893, 9592911, 9464134.

Figure 3: CNV ID - 46771406 results

The CNV id used for further analysis was 46771406 with variant region id nsv3147429 and study id nstd15 as given in output by the module. The output contained object type as variant, organism as human, variant call count as one, method used as sequencing, clinical significance
was blank and sort order was 1000230201608010000339522000003147429. The CNV type was copy number gain. The CNV chromosome location information was list of two assemblies with information on the CNV chromosome number and its start and end location on it. This CNV was found located on chromosome 11 on both genome assemblies GRCh37 and GRCh38.p12. On GRCh37 assembly the start and end position of this CNV was found to be 17096641 and 17436162 respectively. On GRCh38.p12 assembly the start and end position of this CNV was found to be 17075094 and 17414615 respectively. The gene list from the output provided the gene names and their respective ids on which this CNV is prevalent. The gene names provided in the gene list of this CNV included RPS2P40, NCR3LG1, KCNJ11, ABCC8, SNORD14A, LOC105376576, RPL29P21, RPS13, RNU6-593P, RPL34P24, SNORD14B, NUCB2, PIK3C2A.

GRCh37 assembly and its start and end positions were only used for the extraction of the CNV sequence on chromosome 11 from chromosome dictionary in the pipeline and then utilized as the modified CNV sequence for downstream analysis. The three genes RPS2P40, NCR3LG1 and KCNJ11 were used for further analysis. Wild type DNA sequences of these genes were extracted from the gene code fasta dictionary. The modified CNV sequence was then compared with the wild type of DNA sequences of these genes individually.

The modified CNV DNA sequence and the wild type of DNA sequences of each of the three genes RPS2P40, NCR3LG1 and KCNJ11 were given as input to the significant protein prediction module of the pipeline. As a result, the output gives the protein or DNA sequences of both the modified CNV and the wild type genes, compares between them and looks for any major protein or DNA sequence differences that affects its functions and then predicts if the changes present are significant or not.
4.3 Significance results

1) For gene RPS2P40 the following output was obtained from the significant protein prediction module of the pipeline -

Figure 4: Significance prediction - RPS2P40 result

Here the DNA sequences of the modified CNV and the wild type gene were used as input and output was obtained in the form of DNA sequences as well. The module predicted the copy number gain type of CNV 46771406 to be significant in this gene.

2) For gene NCR3LGI the following output was obtained from the significant protein prediction module of the pipeline -
Here the DNA sequences of the modified CNV and the wild type gene were used as input and output obtained in the form of protein sequences. The module predicted the copy number gain type of CNV 46771406 to be non-significant in this gene.

3) For gene KCNJ11 the following output was obtained from the significant protein prediction module of the pipeline -
Figure 6: Significance prediction - KCNJ11 results

Here the DNA sequences of the modified CNV and the wild type gene were used as input and output obtained in the form of protein sequences. The module predicted the copy number gain type of CNV 46771406 to be significant in this gene.
CHAPTER 5

Discussion

PDAC is a lethal form of cancer often diagnosed at later stages due to no major symptoms [2][3]. At later stages, the cancer would be easily spread to other organs resulting in higher death rates of patients [4]. Thus, early detection is key to access medicines and for better treatment. The epigenetic regulations of genes are of major importance in PDAC early prognosis leading to major biomarker discoveries.

In this project the PDAC epigenetic genes with histone modifications were identified. Then their CNV information was extracted using the extended CNV module of the pipeline. Using the CNV information, modified CNV sequences were compared with the wild type sequences of genes which had these CNVs in them, finally to determine their significance. In this study only one epigenetic gene KCNJ11 was considered and its CNV id 46771406 of type copy number gain. Three genes used for this CNV were RPS2P40, NCR3LG1 and KCNJ11 followed by their check for significance. In the results it was found that this CNV mutation was significant RPS2P40 and KCNJ11 genes, non-significant in NCR3LG1 gene.

The main aim of the project was to extend the epigenetic pipeline based on CNVs. This was achieved as we were able to successfully process each epigenetic gene and get their complete CNV information from the pipeline output. We were also able to generate modified CNV sequences based on the CNV information obtained and compare them with wild type sequences to access their significance. But we used only one type of CNV mutation, and a few genes associated with it for this project. For much more robust results and analysis multiple types of CNV mutations with multiple genes needs to be accessed. Also, a combination of SNPs and CNVs leading to
mutations to be considered in multiple genes affecting PDAC can provide a novel and an efficient biomarker in PDAC early diagnosis.
CHAPTER 6

Conclusion

This project was completed to develop and extend the current epigenetic pipeline’s ability to extract information based on CNVs from the NCBI dbVar database. The CNV module developed for the pipeline was utilized to extract and provide complete CNV information on the epigenetic PDAC genes. The complete CNV information of a specific CNV id 46771406 of CNV type copy number gain was used for its location on the chromosome and its subsequent sequence extraction as modified CNV sequence. Later the modified CNV sequence was then compared with the wild type of DNA sequences of three genes - RPS2P40, NCR3LG1 and KCNJ11 on which the CNV was found to be present. Then their respective protein and DNA sequences were produced and checked for any significant changes for potential use as a biomarker in early PDAC detection. Among the three genes evaluated, RPS2P40 and KCNJ11 were found to be significant whereas NCR3LG1 was non-significant. The significant RPS2P40 and KCNJ11 can be further analyzed for their role in PDAC progression.
CHAPTER 7

Future Research

Future work can be done by complete integration of the code developed into the pipeline in which we get the results as the modified CNV sequences, and the wild type of DNA sequences of the genes associated with these CNVs along with their significance prediction. Thus, making the CNV module of the pipeline more robust and efficient. This project has only considered significant CNVs in PDAC early diagnosis, which may not prove to be very effective. Instead considering a combination of significant changes in both CNVs and single nucleotides polymorphisms (SNPs) may improve and provide meaningful insights to be explored in PDAC prognosis. In this study only one CNV mutation is considered for its significance as an early indicator for PDAC. Further research needs to be done by considering the multiple mutations present in CNVs, SNPs and their combination to find the effects they have on early PDAC diagnosis.
LIST OF REFERENCES


24. Gharibi, A., La Kim, S., Molnar, J., Brambilla, D., Adamian, Y., Hoover, M., ... Kelber, J. A. (2017). ITGA1 is a pre-malignant biomarker that promotes therapy resistance and metastatic potential in pancreatic cancer. *Scientific Reports*, 7(1), 10060. [https://doi.org/10.1038/s41598-017-09946-z](https://doi.org/10.1038/s41598-017-09946-z)


