THE GENOMIC ALTERATION LANDSCAPE OF PANCREATIC DUCT ADENOCARCINOMA

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

David Adeleke

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Genomics Phenomics, and Bioinformatics
Option: Bioinformatics

November 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

THE GENOMIC ALTERATION LANDSCAPE OF PANCREATIC DUCT ADENOCARCINOMA

By

David Adeleke

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rick Jansen
Chair

Dr. Changhui Yan

Dr. Bong-Jin Choi

Approved:

November 30, 2022
Date

Dr. Changhui Yan
Department Chair

ABSTRACT

By 2030, PDAC is projected to be the second leading cause of cancer-related death in the US. PDAC is a multifactorial disease driven by genomic alterations. Understanding this alteration landscape will both refine the knowledge of disease etiology and enhance disease stratifications, drug design, and targeted treatment.

This study aimed to identify novel genetic alterations that are associated with pancreatic cancer biology and prognosis to further refine the genetic focus for therapy development, disease subtyping, and risk assessment in PDAC.

To this end, SNV, CNV, and clinical data for PDAC patients were downloaded from the ICGC data portal and analyzed for somatic mutations and recurrent copy number variations.

This study showed that KRAS, TP53, and TTN are not only highly mutated but also associated with poor survival in PDAC. Also, this study showed that CN-LOH TP53, KRAS, SMAD4, and RYR3 were associated with reduced risk of death from PDAC.

ACKNOWLEDGMENTS

This project was supported by the NIH, under Award Number 1P20GM109024.[1]
Words cannot express my deepest gratitude to my advisor, Dr. Rick Jansen. I will never forget your inspiring mentorship, and unwavering support. How you meticulously read every version of my thesis and revert in a twinkling of an eye still amazes me.

Special thank you to my committee members (Dr. Changhui Yan & Dr. Choi Bong-Jin) for their constructive feedback and comments on my thesis. Prior to thesis submission, Dr. Yan's computational methods class aided my understanding of the SNP component of this study while Dr. Choi's applied regression class shaped the mathematical modelling component of my ongoing Ph.D. research.

Heartfelt gratitude to my non-NDSU mentor, Dr. Adelaide Rhodes. Thank you for being a shoulder to lean on and listening ears to pour all my concerns as a foreign student in USA.
Finally, A very big thank you to our external collaborator from University of Wisconsin-Eau Claire (Dr. Gomes Rahul), and every member of our research group (Mariam Zamani, Nijhum Paul, Tanha Tabassum, and others I could not mention because of space constraint), your insightful questions and contributions during our weekly meetings are priceless.

[1] The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH

DEDICATION

This thesis is dedicated to the memory of the people who have died from any form of cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNV	Copy Number Variations.
SNV	Simple Nucleotide Variations.
PDAC	Pancreatic ductal adenocarcinoma
DNA	Deoxyribonucleic Acid
SNP	Single-Nucleotide Polymorphism
ICGC	International Cancer Genome Consortium

1. BACKGROUND

In the last 5 years, pancreatic ductal adenocarcinoma (PDAC) has remained the seventh leading cause of cancer-related deaths worldwide, with global incidence and mortality rates of 2.6% and 4.7% respectively [1].Owing to its poor prognosis, the number of new cases (495,773), and the number of deaths (466,003) are almost the same, with the highest incidence rates in Europe, Northern America, and Australia/New Zealand. By 2025, pancreatic cancer may surpass breast cancer as the third cause of cancer-related deaths in Europe and second leading cause of cancer-related death in US by 2030[2,3].

Mutations are genetic changes that become incorporated into the DNA of an organism. Mutations are either somatic or germline. In germline, mutations are passed on to offspring, while in somatic mutations occur at the cellular level in somatic tissues and occur post fertilization. Mutations often manifest in diverse forms such as single nucleotide variations or copy number variations [4]. In single nucleotide variation (SNV), there is a variation at a single nucleotide position when compared to those of the population genome (accepted human genome sequence reference). While in Copy number variation (CNV) sections of the genome are repeated and the number of repeats in the genome varies between individuals. Altogether, these two types of variations contributes to the overall genomic mutation that drive tumorigenesis [4,5].

As a key regulator of genomic and epigenomic abnormalities, several CNVs have been found among those with pancreatic cancer. For example, Zhan et al., (2021) categorized PDAC patients into four molecular subtypes based on solid tumor derived CNV value for a set of 8 genes (PALB2, RAD51C, FGF3, NF1, FGF4, PIK3CA, MAPKAP1, RICTOR) that shows statistically significant association with overall survival. Zhan et al., (2021) further showed that low copy number subtype (DNA repair deficient) demonstrated poor prognosis [6]. Similarly, Liming et al.,

(2012) who examined the functional significance of CNVR2966.1 at 6q13 , showed that the risk of pancreatic cancer observed in 1027 cases and 1031 controls was significantly associated with copy number of CNVR2966.1, (HR=1.31 95% confidence interval = 1.08–1.60; P = 0.007) for one copy genotype compared with two copies genotype[7].

In the same vein, as a key driver of tumor, studies have shown that somatic mutations of genes such as KRAS, TP53 CDKN2A, and SMAD4 not only accumulate in PDAC but also significantly associates with PDAC survival ($p < 0.05$) [6,8,9].

The identifications of the associations between genomic alterations and pancreatic may not be sufficient to reveal the complete causal relationship between genomic alterations and PDAC progression, such information could refine the current knowledge of disease process and biological pathways [10]. The reliability and generalization of the association between genomic alterations, and pancreatic cancer relies on large scale quality genomic data.

The International Cancer Genome Consortium (ICGC) provides public access to quality genomic, epigenomic, transcriptomic, and proteomic data for over 33 human cancers derived from whole tumor tissue. Thus, providing rich dataset for computational multi-omics data exploration and model building.

The aim of this research was to apply a bioinformatics approach to understand genomic alterations landscape in pancreatic cancer, and how it is associated with prognosis. We hope that this study will identify novel gene alterations that are associated with pancreatic cancer biology and prognosis, thereby providing further refined genetic focus for therapy development, disease subtyping, and risk assessment in PDAC.

To the best of our knowledge, no previous study has combined dataset from these sites, the use of datasets from three distinct study sites with possible inherent genomic variations will provide more sample size to accommodate all possible subtypes of PDAC.

1.1. Modifiable Risk Factors

Although there is no single risk factor that can be regarded as sufficient or necessary cause of pancreatic cancer, several modifiable risk factors such as smoking, drinking, diets of processed meat, high-fructose beverages, and saturated fat have been associated with increased risk of pancreatic cancer[11] [12].

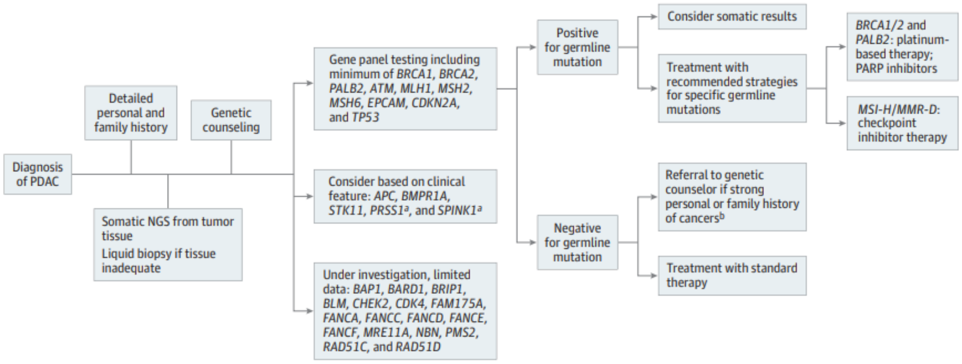
Among lifestyle risk factors, cigarette smoking shows the strongest association with pancreatic cancer, followed by daily alcohol intake in excess of 30 g per day [13]. The presence of two or more of these factors further increase the risk of PDAC, a recent study showed that the association between body mass index (BMI) and risk of pancreatic cancer increase by several-fold among obese smokers [14].

1.2. Genetic Risk Factors

In addition to the modifiable risk factors, several germline genotypes have also been associated with pancreatic cancer risks. In view of this, the National Comprehensive Cancer Network guidelines recommended that all new PDAC cases should undergo a seven gene (BRCA1/2, ATM, MLH1, MSH2, MSH6, and PMS2) panel test to determine the right line of treatment specific to the germline mutations status [15].

Similarly, an individual with familial genetic history of pancreatic cancer is often encouraged to undergo genetic counseling, especially if such individual showed higher risk factor, and negative genetic testing for the 7 gene sets. (Figure 1) [15].

The completion of the human genome and further analysis of its sequence unveils an unprecedented amount of variability in human populations. The most common forms of genetic variation in pancreatic cancer include gene copy number variations (CNV), simple nucleotide variations (SNV), chromosomal translocations and microsatellite instability. The ICGC publicly available dataset on PDAC is limited to CNV, and SNV, as such this study only focuses on these two common genomic alteration types.


Diagnosis of PDAC
Detailed personal and family history
Genetic counseling
Somatic NGS from tumor tissue
Liquid biopsy if tissue inadequate
Gene panel testing including minimum of BRCA1, BRCA2, PALB2, ATM, MLH1, MSH2, MSH6, EPCAM, CDKN2A, and TP53
Consider based on clinical feature: APC, BMPR1A, STK11, PRSS1a, and SPINK1a
Under investigation, limited data: BAP1, BARD1, BRIP1, BLM, CHEK2, CDK4, FAM175A, FANCA, FANCC, FANCD, FANCE, FANCF, MRE11A, NBN, PMS2, RAD51C, and RAD51D
Positive for germline mutation
Negative for germline mutation
Consider somatic results
Treatment with recommended strategies for specific germline mutations
BRCA1/2 and PALB2: platinum-based therapy; PARP inhibitors
MSI-H/MMR-D: checkpoint inhibitor therapy
Referral to genetic counselor if strong personal or family history of cancersb
Treatment with standard therapy

Abbreviations: MMR-D, mismatch repair-deficiency; MSI-H, microsatellite instability-high; NGS, next-generation sequencing; PARP, poly-ADP (adenosine diphosphate)-ribose polymerase.

[a] Indicates hereditary pancreatitis genes.

[b] If no prior genetic counseling.

Figure 1: Suggested Algorithm for Germline Testing for Individuals With PDAC.[15]

1.3. Chromosomal Translocation

In chromosomal translocation, a segment from one chromosome is transferred to a nonhomologous chromosome or to a new site on the same chromosome, thereby generating a novel chromosome [16]. This results in the placement of genes in new linkage relationships and generates additional chromosomes without formal nuclear recombination.

Chromosomal translocations, although very rear in PDAC, have been implicated in human cancer, particularly in hematopoietic and lymphoid tumors [17]. A recent comparative genomic hybridization study by Ghadimi et al., 1999 revealed recurring chromosomal change on chromosome arms 3q, 5p, 7p, 8q, 12p, and 20q in pancreatic carcinoma cell lines [18].

1.4. Microsatellite Instability

Microsatellites (MS), also called Short Tandem Repeats (STRs) consist of repeated sequences of 1–6 nucleotides. The distribution characteristics are different from 15 to 65 nucleotides tandem repeats of small satellite DNA, which is mainly located near the ends of chromosomes. MS are widely distributed and mostly is located near the coding region and may be located others region like intron or non-coding region[19]. Eatsride et al., (2016) analyzed 109 pancreatic cancer biopsies , and observed that 22% are MSI high, as such they have better prognosis and are likely to respond to immunotherapy [20].

1.5. Copy Number Variation

The term "copy number variation" is used to describe a form of structural variation in a genome where large segments of DNA have different numbers of copies between individual organisms of the same species. Operationally, CNV describes segments greater than 1 kilo base pairs in length but typically less than 5 megabases. CNVs are broadly classified as either the inclusion of additional copies of sequence (duplications) or losses of sequence segment (deletions). Although an organism’s CNVs are either inherited or the result of a de novo mutation, its discovery has led to some important biological implications.

First, beyond Mendel’s law of independent assortment, an individual's unit of hereditary is more than the sum of the genetic contributions of the individual's two parents, this is partly because the unequal crossover events responsible for CNVs occur during the production of sperm and eggs, children may have lost or gained additional copies of genetic information that were present in either of their parents' chromosomes [21]. Second, the association between CNV and diseases could not be fully explained by the fundamental genetic basis of human diseases [21].

1.6. Single Nucleotide Variation

A single-nucleotide polymorphism (SNP) is an “inheritable single nucleotide variation between members of species or paired chromosomes”. They were considered as common variants in general population (minor allele frequency (MAF) of somatic mutation is > 1%) [22] . On the other hand, single-nucleotide variants (SNV) are point mutations frequently found in cancer tissues but with MAF less than 1%, as such they do not qualify as SNP. Most of them are missense mutations locating at exons and causing alterations of protein’s structure/function. SNVs are abundant in cancer driver genes and cellular pathways essential for cancer progression [23].

DNA mutations are the hallmarks of cancer. Some mutations, termed drivers, give tumour cells a selective growth advantage and promote cancer development. A typical example is the inactivation of mutations of BRCA genes which lead to the activation of downstream pathways in DNA damage repair [24].

Thus, knowledge of these drivers can guide targeted therapy by targeting genes based on their genomic profile. Similarly, there are passenger mutations, which seem to be important but do not directly drive cancerous growth, as well as, other molecular changes at the RNA and protein levels [24]. These mutations all play a role in deregulating cell metabolism, stimulating cell growth and promoting metastasis, but their exact contributions are largely unclear.

The understanding of SNV has several clinical implications on the management and treatment of PDAC. For example, SNV may unveil possible racial differences in disease genomic landscape, which may influence treatment design and generalization. A Recent study by Guo et al., 2021, showed that the most frequent genomic alteration in PDAC among oriental patients of

Chinese descent was KRAS (n = 262, 86.75%), followed by TP53 (n = 171, 56.62%), GNAS (n = 90, 29.80%), RYR1 (n = 73, 24.17%), and POLE (n = 59, 19.54%). [25] This findings aligns with the work of Yachida et al., 2012 who proposed that KRAS, TP53, CDKN2A, and SMAD4 are the four major driver genes mutated in nearly 100% of PDAC patients [26]. Yachida et al., further showed the number of altered genes in PDAC was significantly correlated with both median disease free survival (p=0.008) in patients with Stage I/II disease, and median overall survival (p=0.041) [26].

Furthermore, the knowledge of SNV landscape also plays an important role in predicting the response of pancreatic cancers to DNA damaging agents. For example, Villarroe et al., 2011 showed that the inactivation BRCA2, PALB2 or FANC gene mutations confer susceptibility to cisplatin or PARP inhibitors [27].

The mutation status of specific genes can also be used to infer molecular subtypes of PDAC cancers as exemplified by Xu et al., 2021, who noted that the classical PDAC subtype can be determined by double negative (KRT81−HNF1A−, DN) status, KRT81+HNF1A− for QM-PDA, and KRT81− HNF1A+ for the exocrine-like group in patients [28,29].

Finally, gene mutation status can also be used to infer possible druggable target. Unfortunately, this KRAS mutation is in pancreatic cancer is undruggable except for a specific mutant form, G12C (Janes et al., 2018). Unfortunately, KRAS gene generally has a broad impact on the tumor microenvironment, contributing to promotion and maintenance of cancer malignancy, responses to immunotherapy, and drug delivery [25].

1.7. Genomic Mutation in PDAC

In the last two decades, considerable research has focused on the identification and explanation of molecular correlates of pancreatic carcinogenesis and pathophysiology. Under the exposure of listed risk factors, there is evidence of gradual accumulations of genetic alterations that triggers the expression and/or activation of pancreatic cancer oncogenes, and repression and/or inactivation of tumor suppressor genes coupled with the deregulation of certain signaling pathways cascaded to the onset of pancreatic cancer [30].

Considering the unprecedented dismal survival of pancreatic cancer, identification of candidate biomarkers and key signaling pathways is important for early clinical diagnosis, prevention, and tailored treatment of pancreatic cancer [31]. Increasing research evidence is showing that the integration of diverse omics data sets offers novel insights and further understanding of complex multifactorial diseases such as cancer.

Genomic aberrations due to DNA copy number variation (CNV) and the simple nucleotide variation (SNV) are known to be associated with the onset and progression of pancreatic cancer [32]. Analysis of copy-number profiles of 3131 cancer samples showed that an average tumor sample consists of 17% genome amplification and 16% deletion as compared to averages of 0.35% and less than 0.1% in the normal counterparts [33].

One possible mechanism of CNV pathogenicity is gene duplications, if the duplicated gene is not dosage sensitive, the inactive copy may escape selective pressure and silently accumulate mutations at a faster rate which may eventually affect future protein products [34].

Similarly, pharmacogenomics evidence has shown that aberrations in gene copy numbers and simple somatic mutations can cause significant impact on therapeutic targets, efficacy and adverse effects in several types of cancer [35]. These are particularly important in the treatment administration, as excess gene copies can deleteriously speed up the rate of drug metabolism, and loss of key genes in drug metabolic pathways may lead to build-up of intermediate metabolites [36].

1.8. International Cancer Genomics Consortium

The International Cancer Genome Consortium Accelerating Research in Genomic Oncology (ICGC ARGO) initiative brings together international researchers to analyze genomic and transcriptomic changes along with high-quality clinical data from over 100,000 patients [37].

The Ontario Institute for Cancer Research (OICR) operates as the Data Coordination Centre (DCC) to develop the ARGO Data Platform which manages the submission, processing, analysis, and dissemination of high-quality clinical and molecular data [37].

Participating programs submit clinical data through the ARGO Data Platform, which is also the destination for public data display and analysis [37].

2. DATA ACQUISITION AND PROCESSION

Data in the ICGC GDC can be accessed through the user-friendly web-based GDC Data Portal, which enables browsing, querying and downloading of data and metadata. In addition, the GDC provides a command-line tool for downloading large volumes of data, and an application programming interface (API) for programmatic access to GDC functionality. The dataset for this thesis used the unrestricted publicly available ICGC genomic data common website.

This study utilized datasets from three different cohorts, ICGC- AU, ICGC-US, and ICGC-CA. For each cohort, release 28 of two omics data types (SNV and CNV) were downloaded, and for each subject whose CNV and SNV have been downloaded, the corresponding clinical metadata file was also downloaded for downward clinical association analysis. All data download and downstream analysis was performed using the R statistical package version 4.2.1. The detailed codes, list of packages and versions are documented in *APPENDIX B. DATA PROCESSING*.

For each cohort, the unit record for each dataset is the donor identification number. Table 1 summarizes the number of records per cohort per sample type.

Table 1: Sample size distribution by ICGC pancreatic cancer cohort and data types

	PACA-AU	**PAAD-US**	**PACA-CA**	**Total**
Number of clinical metadata	461	185	317	963
Number of CNV records	461	185	317	963
Number of SNV records	391	177	268	836
Number of SNV records with Clinical Records	391	177	268	836
Number of CNV records with Clinical Records	461	185	317	963
Number of SNV records with CNV Records	391	177	268	836
Number of SNV records with CNV and clinical records	391	177	268	836

The result from Table 1 above showed the absolute counts of omics data per study site included in this study. Overall, there are 963 patients' records, of which only 836 have valid SNV records.

2.1. Survival Analysis

Cox PH regression model was used to compare the time-to-event between mutant and wild types. Our goal for using Cox PH model is to compare the hazard rates of individuals with mutant genes to individuals with wild type genes.

The general form of Cox PH regression models is given as follows:

$$log\big(h(t)\big) = log(h0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta pxp$$

This implies that the natural log of the hazard at time *t*, denoted by h(t)h(t), as a function of the baseline hazard, h0(t)h0(t) (the hazard for an individual where all exposure variables are 0), and multiple exposure variables x1,x2,…,xpx1,x2,…,xp.

In order to eliminate any bias average hazard ratio estimates, the weighted Coxregression was used, as proposed by Schemper et al., (2009) [38]. The coxphw package was used for the weighted coxPH survival analysis.

2.1.1. Diagnostics for the Cox Model

The Cox proportional hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data.

The fundamental assumption in the Cox model is that the hazards are proportional (PH), which means that the relative hazard remains constant over time with different predictor or covariate levels.

The following assumptions were checked for all the COX PH Models used in this study.

2.1.1.1. The Proportional Hazards Assumption.

The proportional hazards assumes that estimates β do not vary much over time. The Schoenfeld residual was used to check the proportional hazards assumption. In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption. The proportional hazard assumption is supported by a non-significant relationship between residuals and time and refuted by a significant relationship. In this study, none of the models violated this assumption.

2.1.1.2. Testing Influential Observations

To check for influential observations, *deviance residual* (symmetric transformation of the Martingale residuals) was used. The residuals for the models showed a roughly symmetrical distribution around zero.

2.1.2. Hypothesis Testing

For our COXph model, the hypothesis was tested using the wald test. The null hypothesis of the Wald test states that the coefficient βj is equal to 0.

$$Z = \frac{\hat{\beta}_j - 0}{Std.Err(\hat{\beta}_j)} \sim N(0,1)$$

If the test fails to reject the null hypothesis, this suggests that removing the variables from the model will not substantially harm the fit of that model, since a predictor with a coefficient that is very small relative to its standard error is generally not doing much to help predict the dependent variable.

2.2. Simple Nucleotide Variation Analysis

The Variant Call Format (VCF) is a generic format for storing DNA polymorphism data. It is a text file format (often compressed) and contains meta-information lines, a header line, and then data lines each containing information about a position in the genome [39].

However, with advances in Cancer Genomics, Mutation Annotation Format (MAF) is being widely accepted and used to store somatic variants detected. Mutation Annotation Format

(MAF) is a tab-delimited text file produced through the Somatic Aggregation Workflow with aggregated mutation information from variant call format (VCF) Files.

The International Cancer Genome Consortium data are stored in Simple Somatic Mutation Format (SSMF) which is similar to MAF format in its structure, however, the field names and classification of variants is different from that of MAF [40]. The " icgcSimpleMutationToMAF" function of the "MAFTOOLS" R package was used to read ICGC (SSMF) data and convert them to MAF for ease of compatibility with the suite of functions in the maftool. The corresponding clinical metadata for each study cohort was integrated into their MAF file using the "read.maf" function of the "MAFTOOLS".

The final output files were used for the downstream analysis such as "oncoplot", somatic interactions, association between mutations and survival, gene mutation disease signatures. Single nucleotide variation analysis was done according to the workflow described by Mayakonda et al., (2018) [40]. The details of the R implementation of this workflow were included in APPENDIX D. SNV DATA ANALYSIS.

2.3. Identification of Recurrent CNV in Cancer.

The most significant recurrent CNV was identified using genomic analysis of significant chromosomal aberrations (GAIA) package in R. Briefly, GAIA which is based on a conservative permutation test was used to estimate the probability distribution of the contemporary mutations expected for non-driver markers.

Genomic regions identified as significantly altered in copy number (corrected p-value < 0.001) were then annotated to report amplified and deleted genes potentially related with PDAC. The CNV identification and annotation were done in line with the workflow described by Silva et al., (2016). [41] The R code implementation of the CNV workflow, and its association with survival are documented in APPENDIX E. CNV DATA ANALYSIS.

2.4. Mutational Signatures Analysis

Somatic mutations found in cancer genomes may be as a result of natural infidelity in DNA replications, or endogenous exposure to mutagens, enzymatic alterations of DNA, or defective DNA repair mechanism. In some cancer types, a substantial proportion of somatic mutations are said to occur as a result of exposures to external mutagens, for example, tobacco smoking in lung cancers and ultraviolet light in skin cancers, or by abnormalities of DNA maintenance, for example, defective DNA mismatch repair in some colorectal cancers [42]. Alexandrov et.al (2013) argued that different mutational processes often generate different combinations of mutation types, termed ‘signatures. In this study, the mutation signature analysis was done using the method described by Alexandrov et.al (2013). This is described briefly as follows.

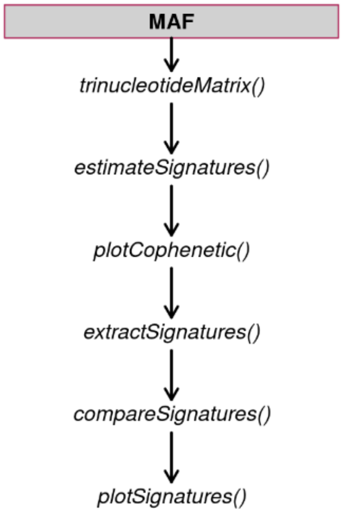


Figure 2: Signature analysis steps

I. estimateSignatures - which runs NMF on a range of values and measures the goodness of fit - in terms of Cophenetic correlation.

II. plotCophenetic - which draws an elbow plot and helps you to decide optimal number of signatures.

III. extractSignatures - uses non-negative matrix factorization to decompose the matrix into n signatures. n is chosen based on the above two steps.

IV. compareSignatures - extracted signatures from above step can be compared to known signatures from COSMIC database, and cosine similarity is calculated to identify best match.

V. plotSignatures - plots signatures

3. RESULTS

3.1. Clinical Metadata

After all datasets have been downloaded, columns with missing values for all records were excluded case wise and/or listwise, depending on the downward analysis need. Simple cross tabulations, measures of proportion and histogram were used to describe the key clinical variables. The details of this were documented in APPENDIX C. CLINICAL DATA ANALYSIS.

Table 2: Summary of Clinical metadata by study sites

Characteristics	**PAAD-US, N=185**	**PACA-AU, N=461**	**PACA-CA, N=317**
Disease status last follow-up			
Missing	111 (60%)	41 (8.9%)	105 (33%)
complete remission	48 (26%)	0 (0%)	65 (21%)
no evidence of disease	0 (0%)	0 (0%)	9 (2.8%)
partial remission	0 (0%)	0 (0%)	8 (2.5%)
progression	26 (14%)	292 (63%)	35 (11%)
relapse	0 (0%)	0 (0%)	95 (30%)
stable	0 (0%)	128 (28%)	0 (0%)
Donor relapse type			
0 (NA%)	225 (49%)	194 (61%)	
distant recurrence/metastasis	0 (NA%)	188 (41%)	83 (26%)
local recurrence	0 (NA%)	48 (10%)	22 (6.9%)
local recurrence and distant metastasis	0 (NA%)	0 (0%)	18 (5.7%)
Unknown	185	0	0
Donor Sex			
Female	83(45%)	210(46%)	118(37%)
Male	102(55%)	249(54%)	152(48%)
Donor Survival status			
Missing	0(0%)	1(0.2%)	51(16%)
Alive	119(64%)	167(36%)	75(24%)
Deceased	66(36%)	293(64%)	191(60%)

The data from table 2 showed that there are more male cases of PDAC in all the three regions than female. Furthermore, the chemotherapy status at follow-up was grossly missing the three sites, with PAAD-US site recording about 60% missing values, PACA-AU 8.9% and PACA-CA 33%. Overall, PACA-AU has more records of stable cancer outcome after chemotherapy treatment 128(28%) than other sites. However, PACA-AU tends to have more cases of progression 292 (63%) than other sites. This observation is consistent with the data on overall survival. In PAAD-US 65% are alive as at last follow-up while in PACA-AU and PACA-CA only 36% and 24% are alive respectively.

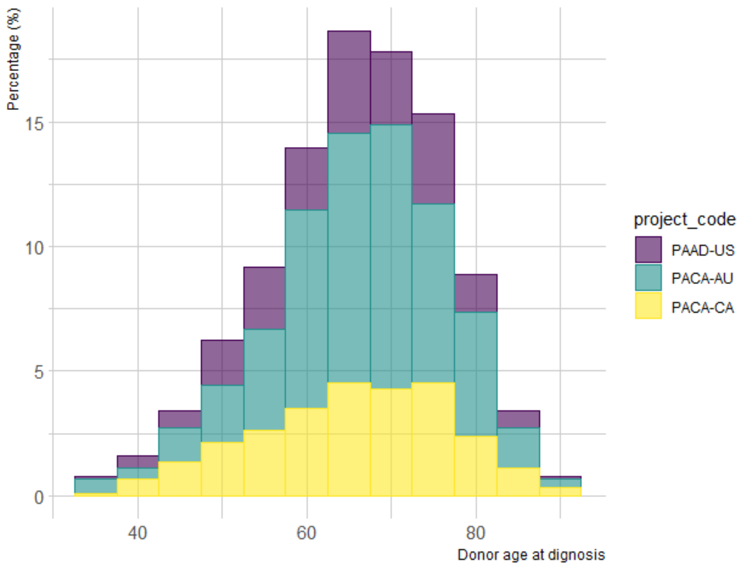


Figure 3: Age distributions of patients at diagnosis stacked by study sites

Figure 3 showed that there is consistent age distribution of population at risk of PDAC in the three sites(PACA-US, PACA-AU, and PACA-CA). In all the three sites, the median age for PDAC diagnosis ranges from 70-75 years. Overall, the median age of diagnosis age is about 71 years, and more than 70% of PDAC cases happened after age 61. This further emphasizes the late presenting attribute of PDAC.

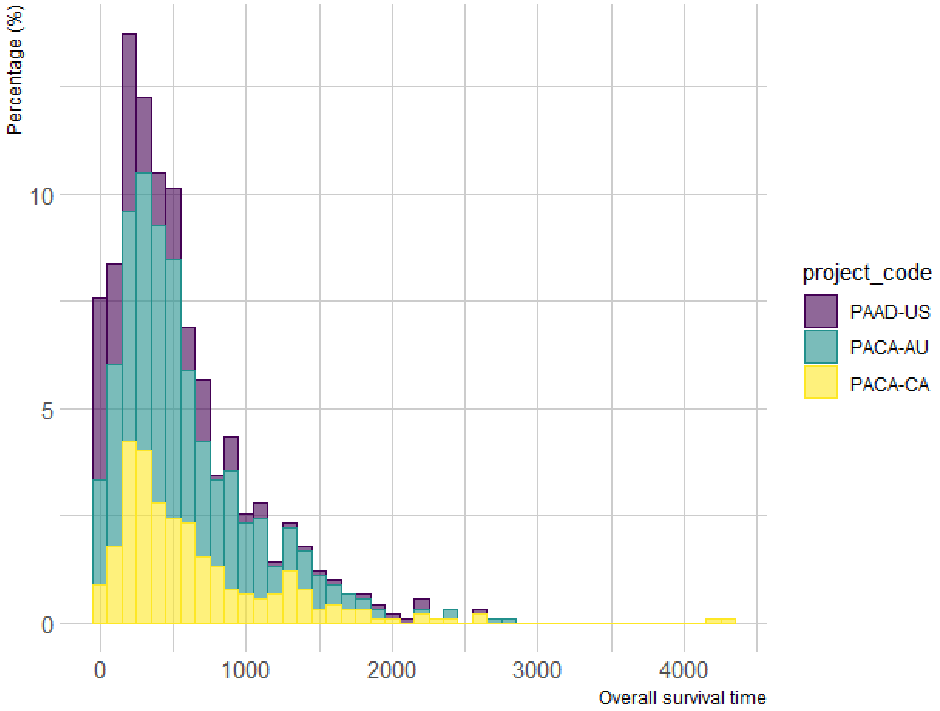


Figure 4: Overall survival time of patients stacked by study sites

The overall survival time between the three sites is relatively consistent with an overall median value of about 500 days. The overall survival pattern appeared to be skewed to the right, with a few cases that survived far above the median time. It is also worth mentioning that the interpretation of this graph may be bias if the study started at different year or patients have entered into the study at different time, unfortunately there was no data on the study start date or patient inclusion date.

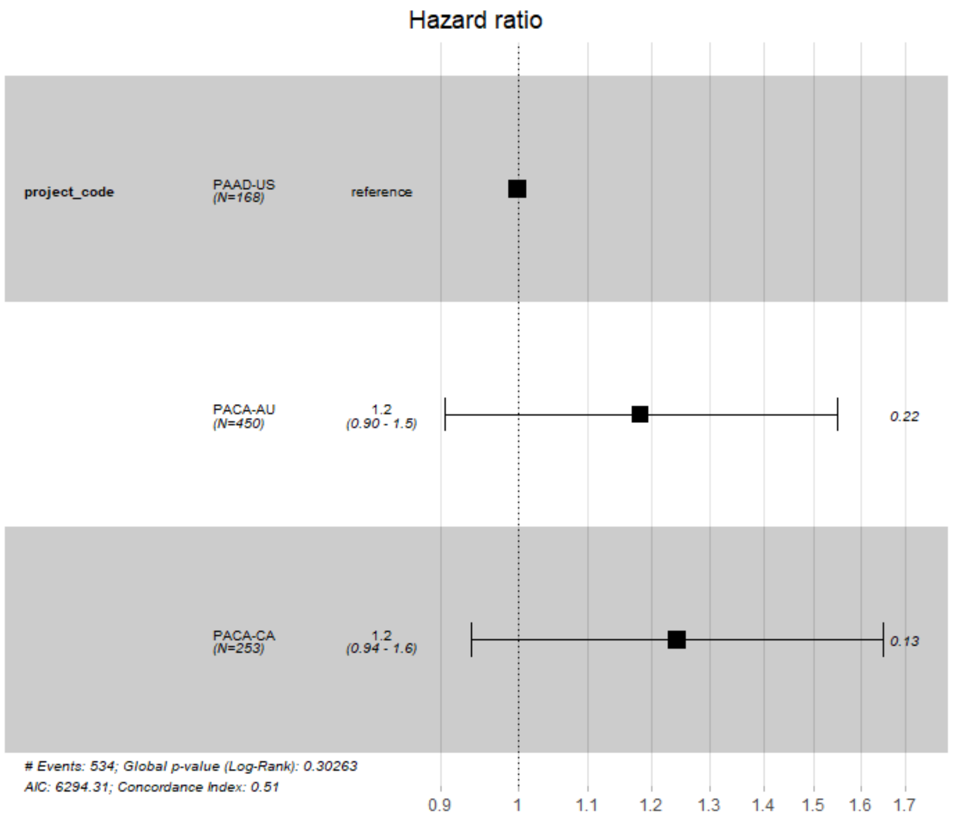


Figure 5: Forest plot showing the hazard ratio stratified study sites

In order to ascertain any differential survival pattern between the three study sites, weighted Cox-regression was used to estimate unbiased average hazard ratio estimates, and wald test was used to test the null hypothesis . Figure 5 shows that there is no statistically significant difference in the hazard ratio estimate among the three sites (weighted pvalue = 0.367, unweighted pvalue= 0.30263).

3.2. Summary of Somatic Mutations

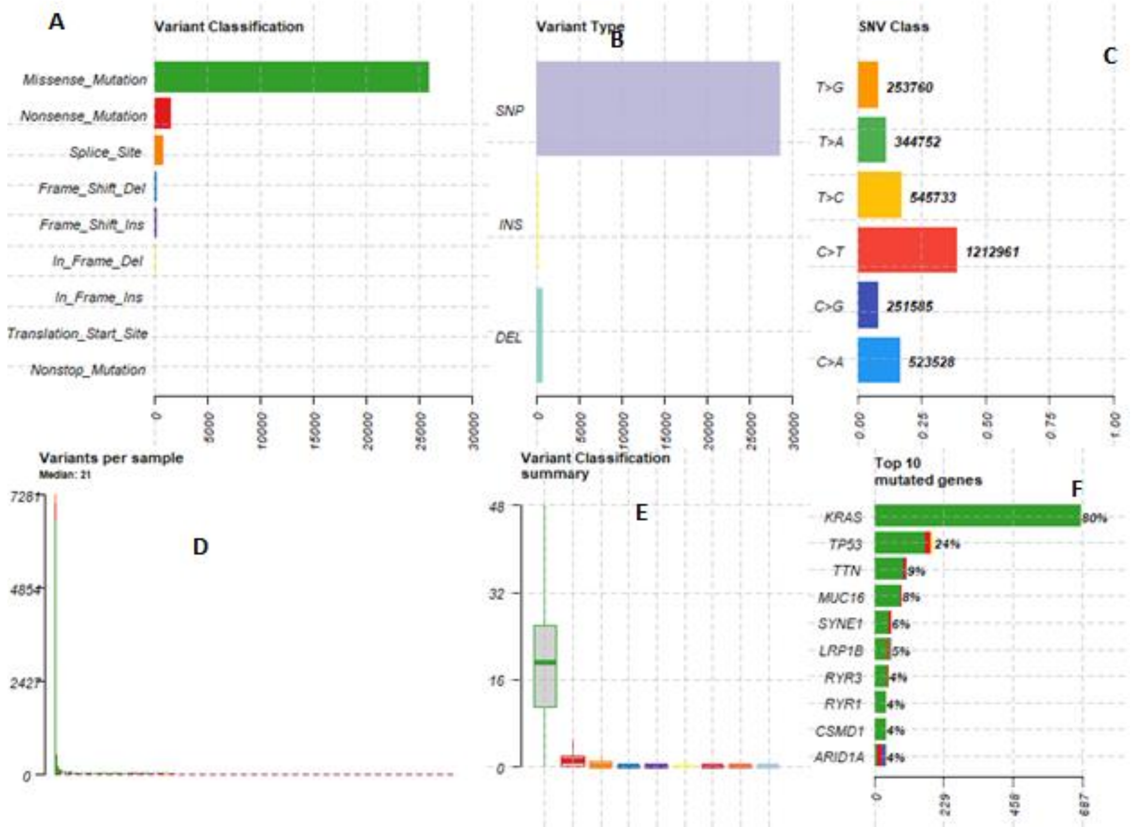


Figure 6: Summary of somatic mutations for pooled data from the three study sites.

Figure 6 above shows the summary of the somatic mutations of PDAC of pooled data from three sites. (A) Bar chart showing occurrence of variant classes, this sub-figure showed that missense mutations which accounted for over 90% is the most occurring variant classifications. (B) Bar chart showing occurrence of variant types, this sub-figure showed that single nucleotide polymorphism accounted for over 90% of the variant types. (C) Bar chart showing simple nucleotide variation classes, this sub-figure showed that the replacement of Cytosine base by Thymine and vice versa are the most occurring SNV class, this means that transversion mutations accounted for over 60% while transition mutations accounted for 40% in PDAC. (D) Bar chart showing absolute count of variant per sample, this sub-figure showed that the median number of variants is 21, this implies that tumor variant burden is minimal in PDAC. (E) Box plot showing variance classification distributions, this sub-figure further emphasized that missense mutation which is the most occurring has a mean value of 20.

(E) Bar chart showing the distribution of most mutated genes, this sub-figure showed that KRAS, TP53, TTN, MUC16, SYNE1, LRP1B, RYR3, RYR1, CSMD1, ARID1A are the most occurring mutations with KRAS gene mutated in 80% of the samples.

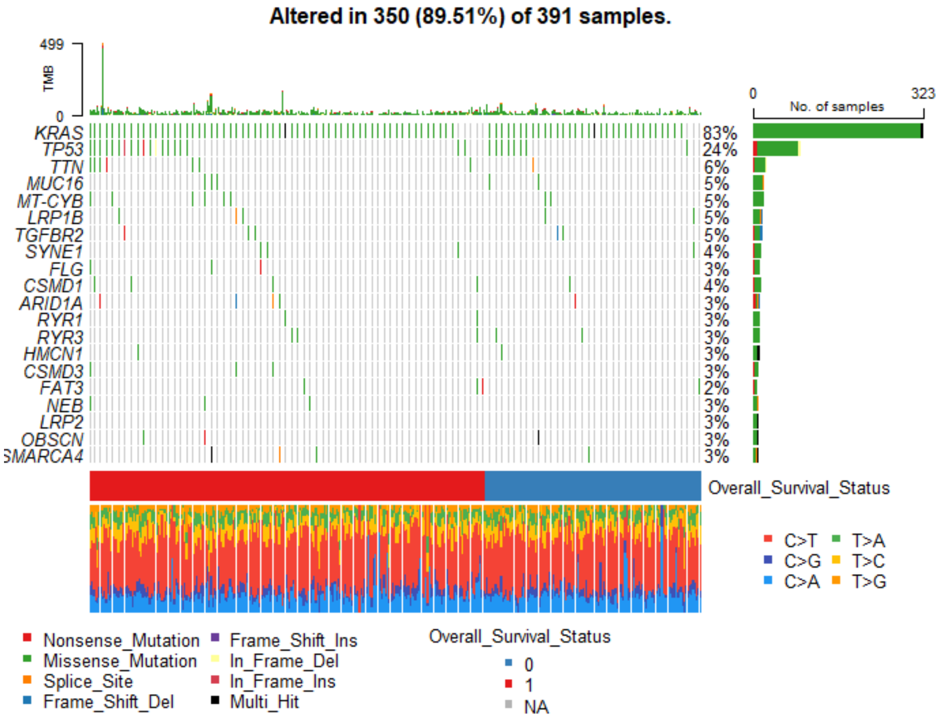


Figure 7: SNV landscape of the top 20 genes among donors in PACA-AU site.

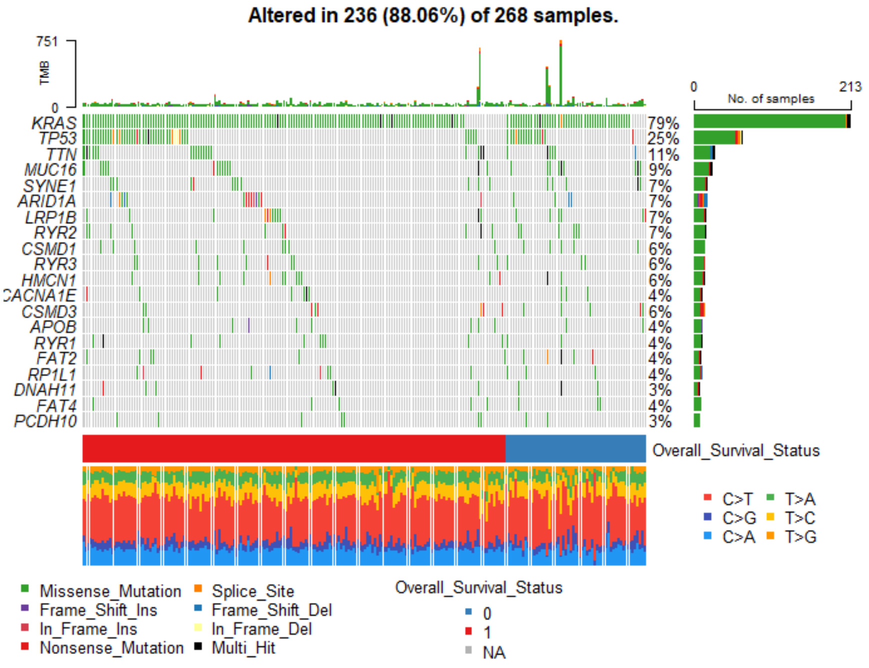


Figure 8: SNV landscape of the top 20 genes among donors in PACA-CA site.

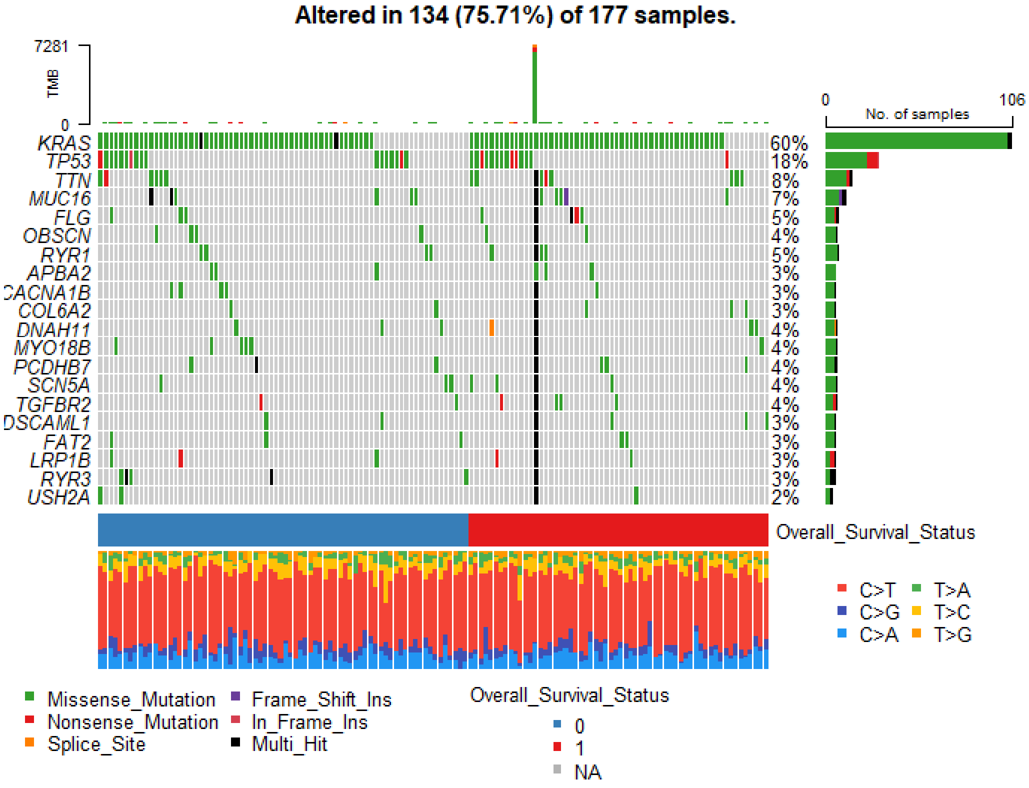


Figure 9: SNV landscape of the top 20 genes among donors in PAAD-US site.

In order to understand between sites variations in the SNV landscape, specific one oncoplot was drawn for each site (Figure 7-9) . Interestingly, there is slight variation in the top mutated genes. While KRAS, TP53, TTN, MUC16 are consistently the top 4 mutated genes, their distribution varies between sites. For example, KRAS genes are mutated in 83%, 79% , and 60% in PACA-AU, PACA-CA and PACA-US sites respectively while TP53 genes are mutated in 24%, 25% , and 18% in PACA-AU, PACA-CA and PACA-US sites respectively.

Lastly, missense mutation is the most occurring mutation type, and it spreads across the patients irrespective of their overall survival status (1= dead, 0= alive), and nucleotide base change types.

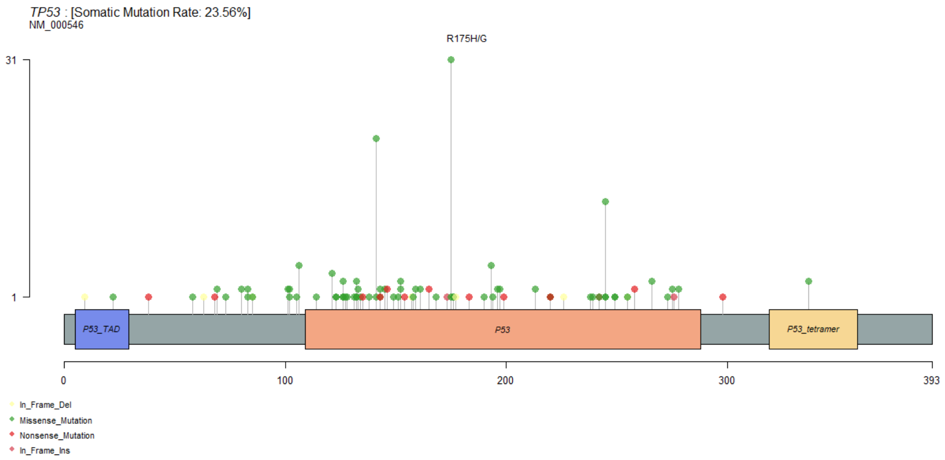


Figure 10: Lollipop Plot showing the amino change due to TP53 gene mutation

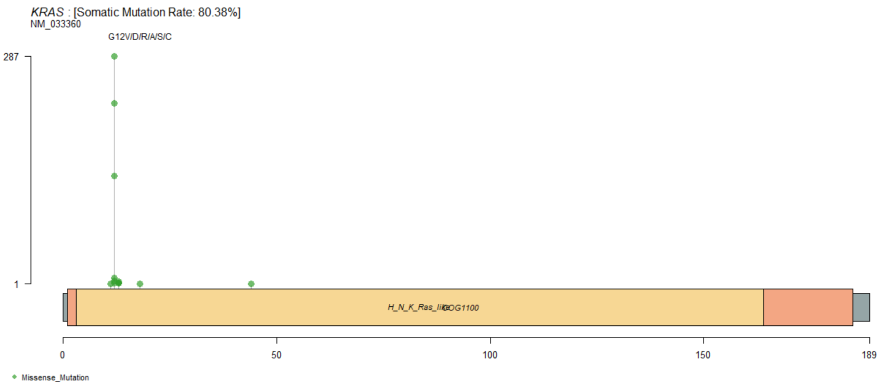


Figure 11:LollipopPlot showing the amino change due to KRAS gene mutation

Since KRAS and TP53 genes are mutated in over 80% of the samples, it is imperative to look closer into these genes and understand the loci of these mutations. As shown in figure 10 & 11, while 23.56% of the subjects had TP53 mutations, the points of this mutations spread across the protein domain. Overall, position 175 is the most occurring point of mutation in TP53 (Figure 10) and the points of mutations are spread across the P53 domain.

Unlike TP53, KRAS gene has a conserved point of mutation, with about 287 samples mutations occurring at position 12, replacing Glycine with either Glutamate, Arginine, Alanine, Serine, and Cysteine (Figure 11).

3.3. Somatic Interaction

Somatic mutations are not random, they often occur in characteristics pattern to drive or inhibit certain biological activities. In this section of the analysis, we are interested in the occurrence of pattern of the mutations. Particularly, among the top mutated genes. We would like to know if the gene mutations exhibit any co-occurrence pattern or mutual exclusiveness.

To this end, the top 15 highly mutated genes were selected and analyzed for occurrence pattern. The Fisher's exact test was used to test for occurrence association between set of genes. Figure 11 below shows the occurrence pattern of the top 15 highly mutated genes.

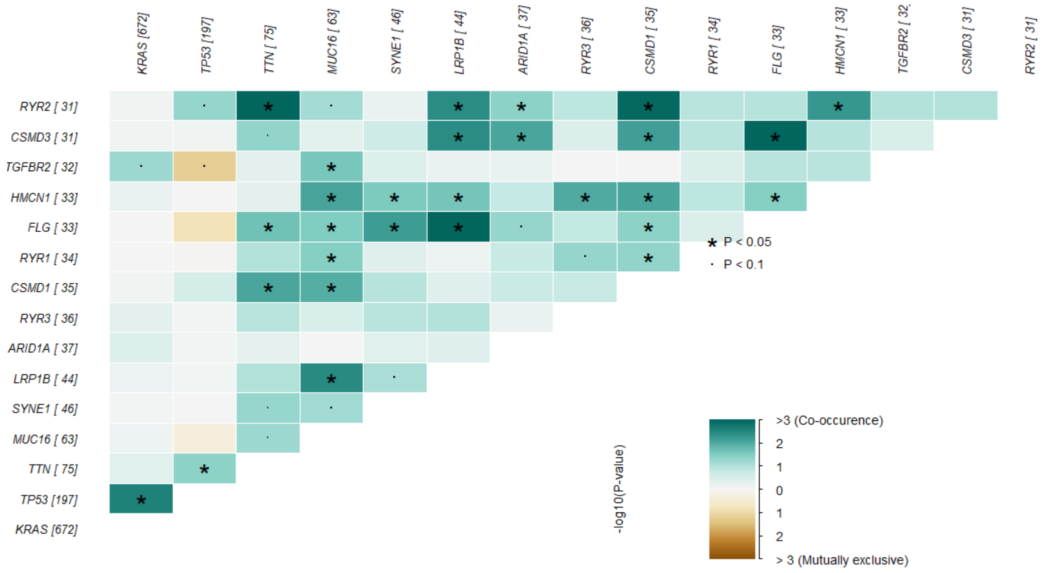


Figure 12: Somatic interaction plot showing mutually exclusive or co-occurring genes

As shown in figure 11, KRAS mutations co-occurred with TP53 mutations, while RYR2 mutation co-occurred with, TTN, MUC16, LRP1B, ARD1A, RYR, and CSMD1 and MUC16 co-occurred with TGFBR2, HMCN1, FLG, RYR1, CSMD1, RYR3, and LRP1B. In all these combinations, there are no significant mutually exclusive occurrences (Fisher's exact p< 0.05).

3.4. Detecting Cancer Driver Genes

It is unlikely that the observed millions of mutations drive PDAC, most of the variants in cancer causing genes are enriched at few specific hot spots, as such it is important to narrow down these mutations and separate driver genes from passenger genes. This is very important in understanding the disease pathway, disease subtyping, and development of candidate drug[43].

Using the oncodriveCLUST algorithm – "detect significant clustering signals across genomic regions based on a local background model derived from the simulation of mutations accounting for the composition of tri- or penta-nucleotide context substitutions" [43], top 10 driver genes were identified (KRAS, ESX1, PROX2, PSG6, TP53, SF3B1, SMAD4, RNF150, CDKN2A, METTL14) (Figure 12). Some of these driver genes (KRAS, TP53, SMAD4) are also among the top mutated genes (Figure 5).

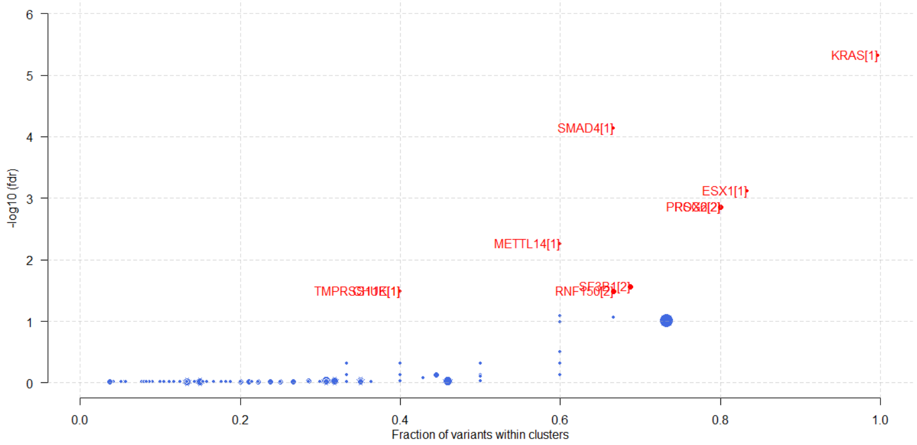


Figure 13: Top PDAC driver genes identified by oncodriveCLUST algorithm

3.5. Association between Driver Genes and Overall Survival

Having identified top mutated genes, some of which have also been confirmed to be drivers of PDAC, it is imperative that we ascertain their association with survival. Using the COXPH model ($log(h(t)) = log(h0(t)) + \beta_1 Mut\_status_1$). Prior to the COXPH, preliminary analysis was done to check if the data obeys the assumptions of COXPH, the preliminary analysis showed that the COXPH assumptions were not violated (APEENDIX D).

The overall survival of the patients with mutant genotypes was compared to those of patients with wild type genotypes for each driver gene. The genes with statistically significant associations are shown in the forest plot in figure 14. Figure 14 showed that patients with TP53 gene mutation have a 2 times risk of death from PDAC than patients with wide type TP53Similarly, patients with KRAS, CSMD1, and ARID1A mutations have an increased risk of death from PDAC than patients with wide type genotype.

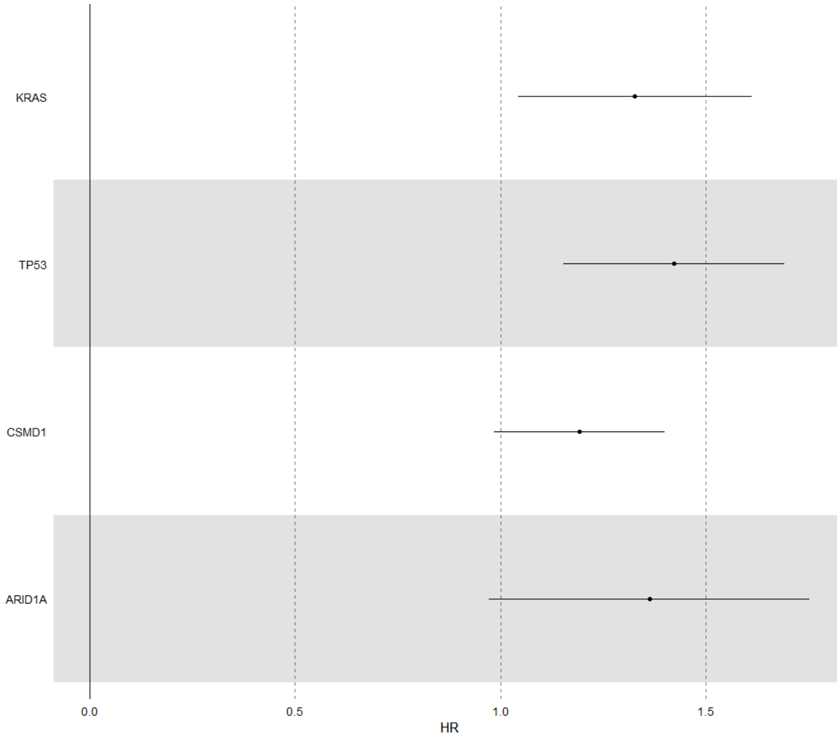


Figure 14: Hazard ratio for survival based on gene mutation status (95% CI), (p<0.05)

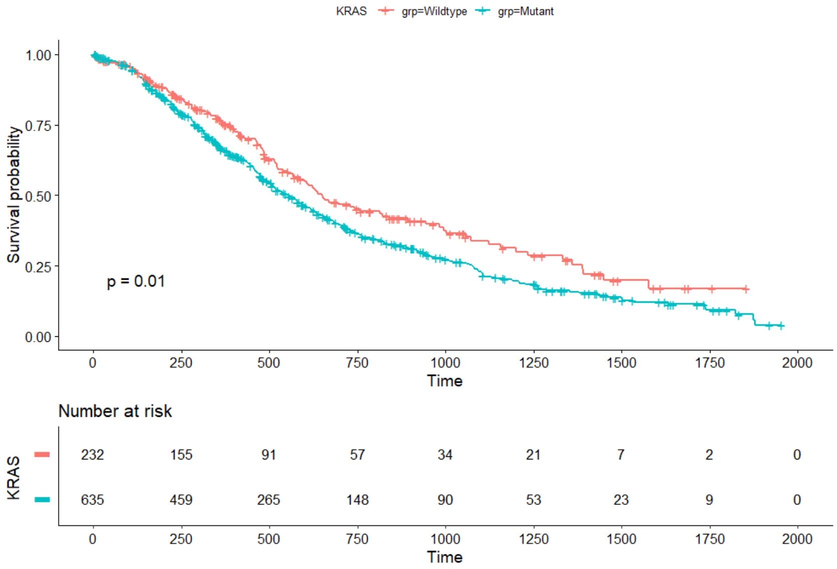


Figure 15: KM Survival Pattern based on KRAS mutation status

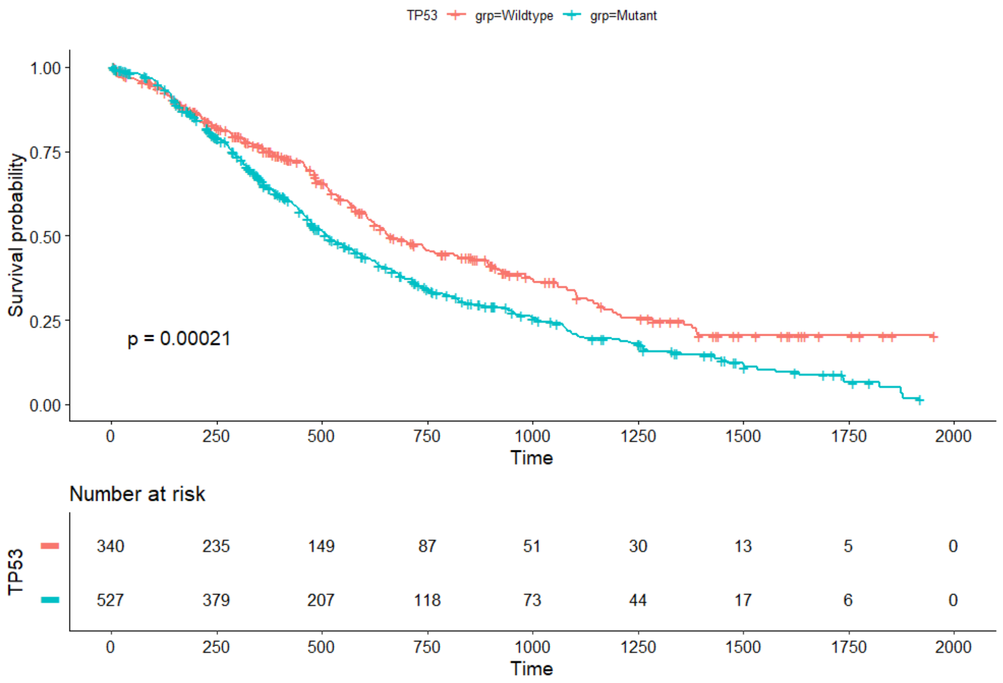


Figure 16: KM Survival Pattern based on TP53 mutation status

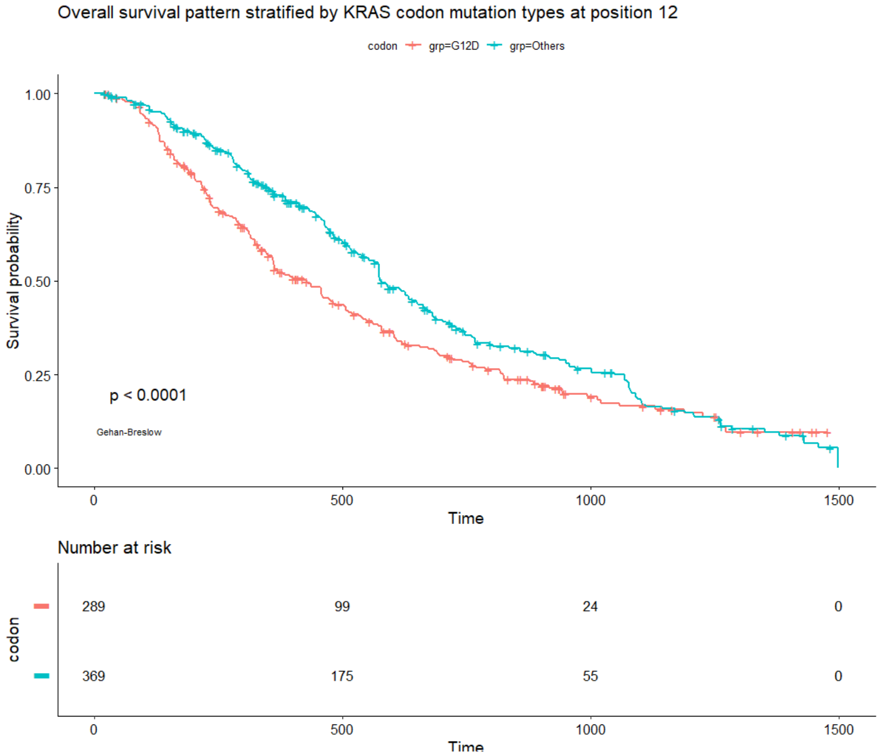


Figure 17: KM Survival Pattern based KRAS codon mutation at position 12.

3.6. Oncogenic Signaling Pathway Enrichment

Signaling Pathway describes a series of chemical reactions in which a group of molecules in a cell work together to control a cell function, such as cell division or cell death. Abnormal activation of signaling pathways is one of the mechanisms that drives cancer [44]. The identification of these pathways , and therapeutic targeting of specific regions can help keep cancer cells from growing[45].

The genomic alterations in signaling pathways that control cell cycle progression, cell growth, and apoptosis, are common hallmarks of cancer. However, the extent, mechanisms, and co-occurrence of alterations in these pathways differ between individual tumors and tumor types. Identifying these pathways could further enhance the understanding of PDAC pathophysiology and likely drug targets.

To this end, the presence of enriched oncogenic signaling pathway was checked in PDAC. Ten (10) significantly enriched signaling pathways were identified. RTK-RAS pathway is the key pathway that was enriched in about 84% of the samples, followed by TP53 and HIPPOS pathways with about 25% and 19% representations respectively (Figure 18).

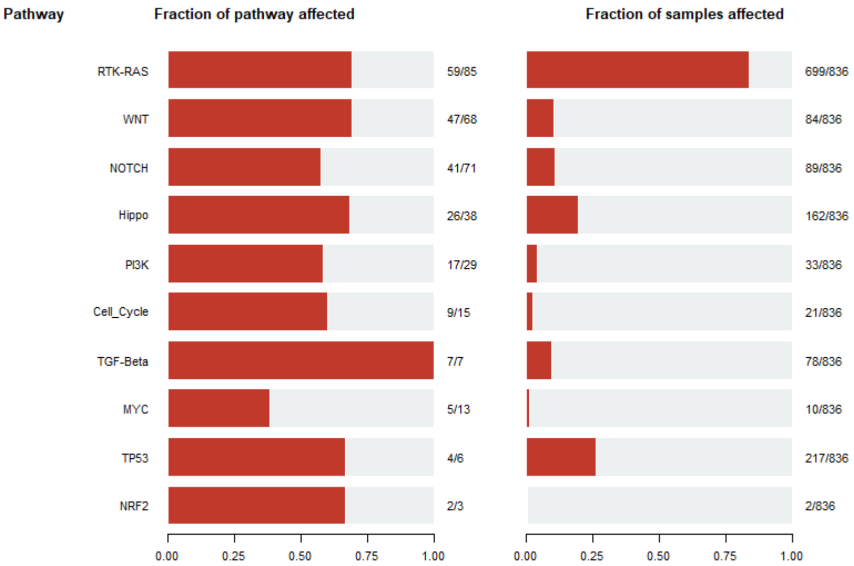


Figure 18: Enrichment of known Oncogenic Signaling Pathways.

3.7. Mutational Signatures Analysis.

Every cancer, as it progresses, leaves a signature characterized by a specific pattern of nucleotide substitutions. Such signatures can be identified and compared to curated signatures from public database. For this study, comparison will be made with the COSMIC database. The COSMIC signature database was curated by identifying signatures from the analysis of the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset and through curation of specific papers, and also updates as new data become available.

Although this signature catalogue is not exhaustive or a final set, but it is a reference set of high confidence signatures that have been curated by experts in the field [46] . For each of the identified mutational signature, the COSMIC database provides key information possible etiology and tissue distribution of each signature [46].

In this study, the signatures in the PDAC were extracted compared to validated signature using the method described by Alexandrov et.al which modeled mutational processes as a blind source separation problem [43].

Overall, the PDAC mutational signature was consistent with three existing cancer signatures from the COSMIC signatures database. Out of these three signatures, two are linked to spontaneous or enzymatic deamination of 5-methylcytosine and DNA mismatch repair while the last one has unknown etiology Figure (22).

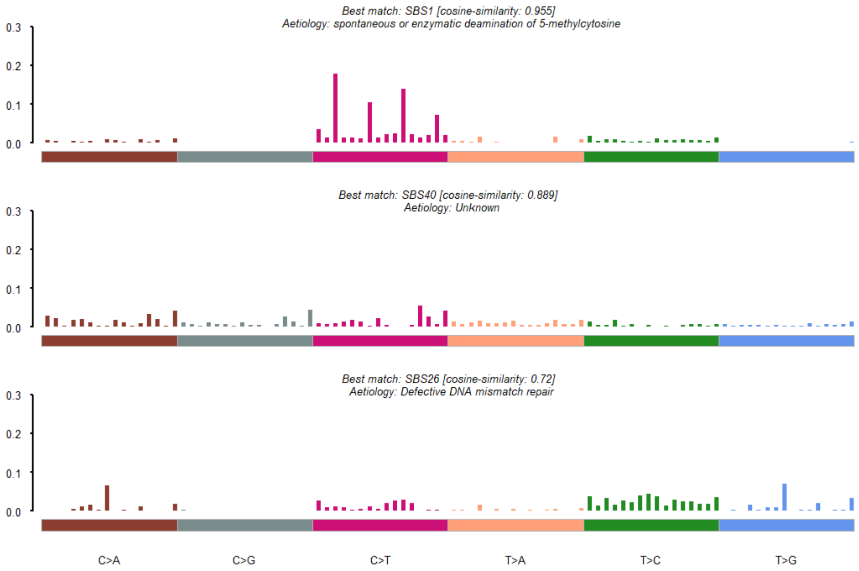


Figure 19: PDAC Signature and possible aetiology [2]

[2] SBS: means Single base substitutions

4. COPY NUMBER VARIATION ANALYSIS

In order to improve our findings on the landscape of genomic alterations, the scope of the gene CNV needs to be understood, particularly among the highly mutated genes.

The ICGC CNV data somatic copy Number workflow uses a tumor-normal pair of either SNP6 raw CEL data, or WGS data as input. The ASCAT algorithm derives allele-specific copy number segments while estimating and adjusting for tumor purity and ploidy [47] . Because there are two parental strands, the resulting, the total copy number at any locus is the sum of the copies from both parents.

For CNV analysis, only protein-coding genes were kept, and their numeric CNV values were further threshold by a noise cutoff of 0.3 as follows:

I. Genes with focal CNV values smaller than -0.3 are categorized as a "loss" (-1)

II. Genes with focal CNV values larger than 0.3 are categorized as a "gain" (+1)

III. Genes with focal CNV values between and including -0.3 and 0.3 are categorized as "neutral" .

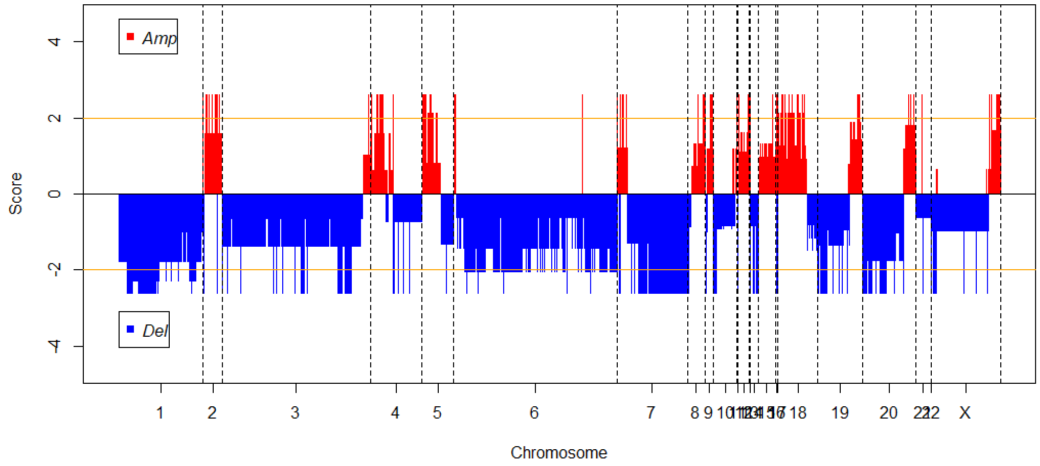


Figure 20: Recurrent amplifications and deletions identified in PDAC

The pattern of the recurrent amplifications and deletions showed that there is widespread CNV across the 23 chromosomes in PDAC While, there might be widespread of CNV across the chromosome length, it is more occurring in some regions of the chromosome than others for example, chromosome 1,3, and 6 showed increased abundance of CNV than other chromosomes. However, copy number deletion is the most common CNV (figure 20).

Interestingly, the chromosomes where the previously identified top mutated genes were located also showed significant copy number alterations. For example, chromosomes 17 which house Tp53 genes showed copy number amplification for TP53 (figure 21), while chromosome 12 which houses KRAS showed copy number amplification. Biologically, CNV arise by homologous recombination between repeated sequences (recurrent CNVs) or by non-homologous recombination mechanisms that occur throughout the genome (non-recurrent CNVs)[48], while SNV accumulate either from spontaneous errors in replication that evade the proofreading function of the DNA or from mutagen that react with parent DNA[49] . This implies that both types of genetic alterations can accumulate independently. However, their effect on the phenotype is could be synergistic [46].

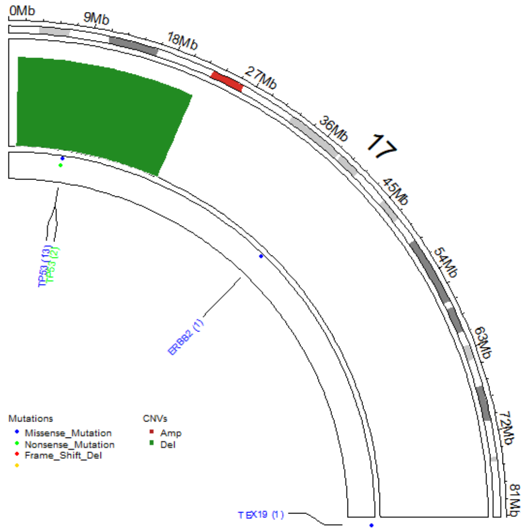


Figure 21: Chromosome 17 showing recurrent CNVs pattern and mutated genes.

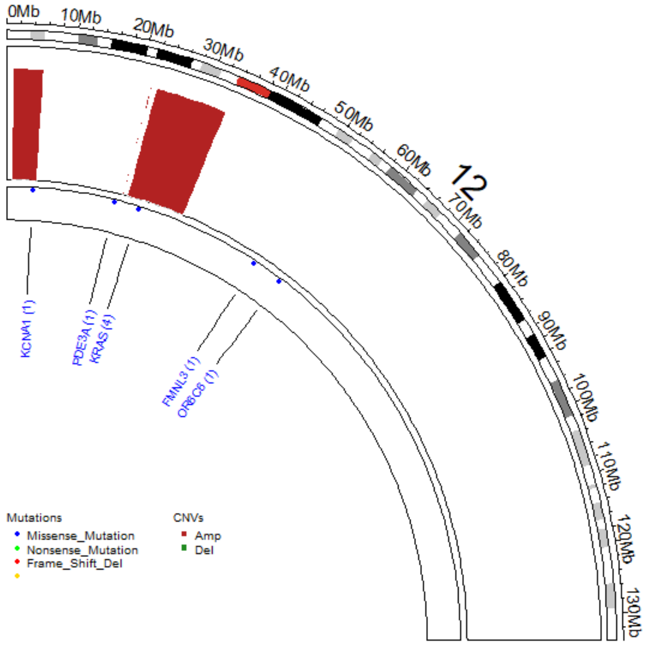


Figure 22: Chromosome 12 showing recurrent CNVs and mutated genes.

4.1. Association between CNV Types and Overall Survival

The previous section established the presence of a significant association between gene mutations and PDAC survival. In the same vein, it is important to confirm the association between CNV pattern and overall survival in PDAC. This section looked at the clinical consequence of the CNV types that were observed in PDAC.

This will help to infer the deleteriousness of the specific gene copy number type. Cox proportional hazards regression analysis was used to infer the association between the survival time of patients and copy number types for a given gene. To this end, we used COXPH to evaluate the association between the copy number type of the established driver genes and survival in PDAC.

Using the COXPH model (log(h(t))=log(h0(t))+β_1 〖CNV_status〗_1). Prior to the COXPH, preliminary analysis was done to check if the data obeys the assumptions of COXPH, the preliminary analysis showed that the COXPH assumptions were not violated (APEENDIX D). In all cases, the null hypothesis that the coefficient βj is equal to 0 was tested using wald test.

Only 5 genes (TP53, KRAS, SMAD4, RYR3, CDKN2A) showed significant association with survival in PDAC. Figure 22 below shows that LOH of these genes associated with reduced risk of death from PDAC, while KRAS and RYR3 CNV loss associated with increased risk of deaths.

CNV	Definition
LOH	LOH (short for “loss of heterozygosity”) refers to a type of mutation that results in the loss of one copy of a segment of DNA. Some may be Copy neutral loss of heterozygosity (CN-LOH) , which refers to a special case of LOH occurring without any resulting loss in copy number [50].
Gain	When the number of copies of a gene is more than two
Loss	When the number of copies of a gene is zero or one.

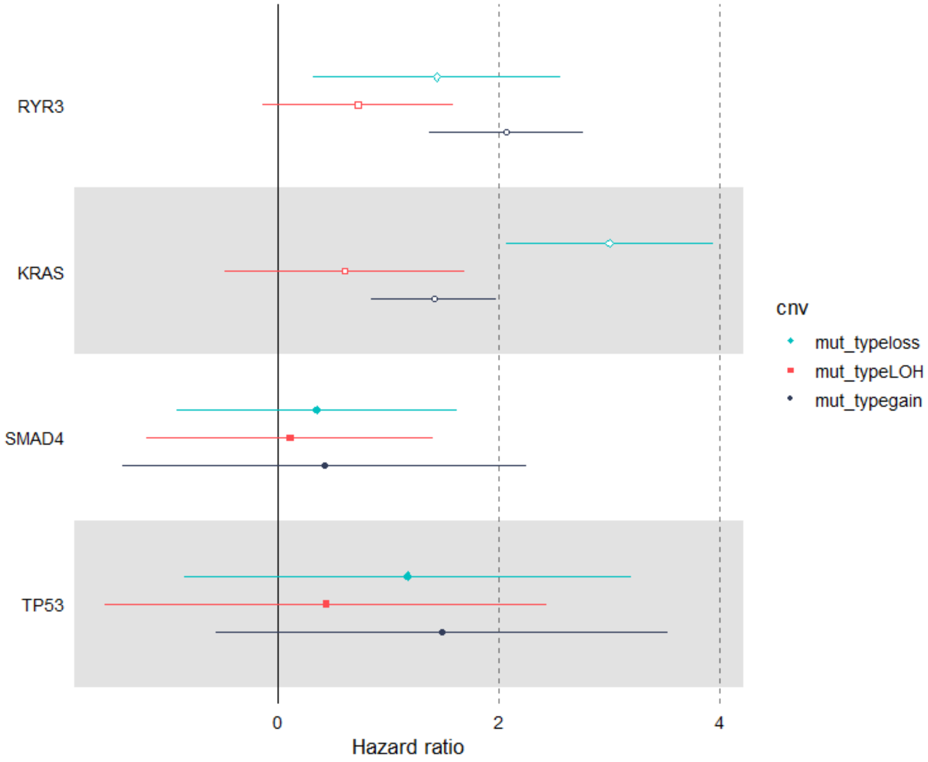


Figure 23: Hazard ratio for survival based on gene CNV alteration (95% CI)[3]

[3] Mut_typeloss = CNV Loss, Mut_typeLOH=CNV LOH, Mut_typegain= CNV gain.

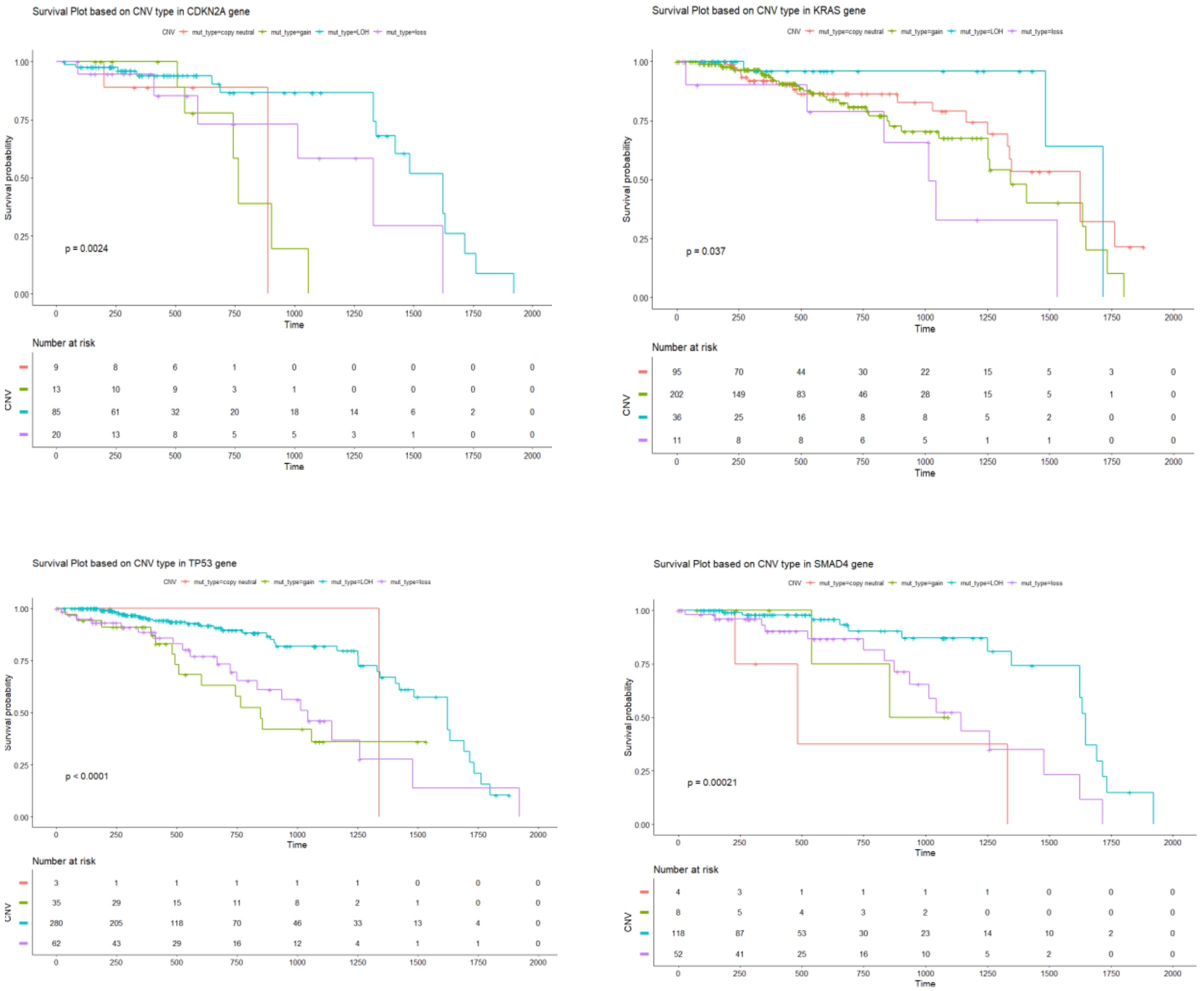


Figure 24: KM survival pattern of PDAC based on gene CNV alteration type

5. DISCUSSION OF RESULTS

Pancreatic cancer is one of the most lethal cancers worldwide, but its etiology, therapeutic modality, and prognosis are not fully understood. This study pooled publicly available ICGC data on SNV, CNV and clinical information for PDAC at PAAD-US, PACA-AU, and PACA-CA study sites. A total of 963 donor samples were included in this analysis, and the median age of diagnosis is 71 years (figure 3). This observation corroborates the late-onset nature of pancreatic cancer. This finding is supported by the work of Hongcheng et al., 2020 and Robert et al., (2016) who observed that the median age of PDAC in China and the United states is 70 and 71 years respectively [51,52].

Our observations of more male cases than female is in alignment with the global trend of PDAC, as shown in the work of Rawla et al., (2019), who stated that globally, the number of new cases of pancreatic cancer is 5.5 per 100,000 for men and 4.0 per 100,000 for women [53]. In terms of overall survival, we observed that a median overall survival time of about 500 days and only 12% lived up to 3 years, across the study sites. This is not surprising for a disease with such a poor prognosis and these findings are similar to the report of the World Health Organization (WHO), which observed that typically after diagnosis, only 9% live for 5 years [54].

This study examined the SNV landscapes in PDAC in order to identify a commonly mutated set of genes, the mutation pattern, signatures, and their impact on overall survival. The findings from this study showed that those missense mutations which accounted for over 90% of all variants class are the most occurring variant classifications in all the top mutated genes. Similarly, SNV of KRAS, TP53, TTN, MUC16, SYNE1, LRP1B, RYR3, RYR1, CSMD1, and ARID1A genes are the most occurring somatic mutations in PDAC.

The most commonly altered gene, KRAS, had mutations in 83% of samples of PACA-AU, 79% in PACA-CA, and 60% in PAAD-US . Pooling samples from all study sites, it was observed that 80.4% mutations in KRAS occurred in codon 12 in which amino acid glycine was replaced with any of aspartic acid, Arginine, Alanine, Serine and Cysteine (D,R,A,S, and C). This shows the specificity of KRAS mutation in PDAC tumorigenesis (Figure 9 & 10).

We further attempted to compare subjects' survival based on their positions of KRAS codon mutations, and we observed that patients with G12D codon mutations have increased risk of death from PDAC (HR=1.4, p<0.001). While there was no study that corroborated the increased risk of death due to codon mutation positions, the presence of varying KRAS mutation codons was supported by the work of Guo et al., (2021) who observed that in KRAS mutations of 408 Chinese PDAC patients, 117 cases were G12D, 105 cases were G12V, 27 cases were G12R, 5 cases were G12C, and 1 case was G12A [25]. Similarly, TP53 mutations were also observed in 23.6% of the samples, but the affected codons in TP53 are widespread across the P53 domain, overall, the most frequently mutated site was in codon 175 (n = 31 (Figure 11).

We observed that KRAS, TP53, TTN, MUC16, and SYNE1 are the five major driver genes mutated in nearly 100% of PDAC patients. Two of these genes (TTN, and MUC16) are absent in what was previously reported to be the top 10 mutated genes in PDAC in some other studies [8,26]. At first, we hypothesized that the difference in the top mutated genes between sites could be as result of different geographical location. However, this turned out not to be supported by our data, as the site specific onco-plots are not consistent with such hypothesis.

However, it could also be related to the stages of the cancer, since the recent WHO classification, PDAC variants were categorized into eight variants with specific histomorphological features, in which the tumor stage is one of the defining features [55]. To check this theory, we stratified samples based on tumor stage at diagnosis, and observed differential mutation landscape between stage 1 and 11. Unfortunately, there were no sufficient samples to ascertain the mutation profile for stage 3 and 4, also the classification methods are not consistent across study sites. (Appendix A).

Nevertheless, the roles of KRAS, TP53, TTN, and SYNE1 have been widely investigated[8,9], and KRAS is considered a driver gene in the initial stages of most PDAC. Since mutations that drive disease conditions are not random, we further checked for cooccurring mutations and mutually exclusive mutations. As shown in figure 13, KRAS mutations co-occurred with TP53 mutations, while RYR2 mutation co-occurred with, TTN, MUC16, LRP1B, ARD1A, RYR, and CSMD1 and MUC16 co-occurred with TGFBR2, HMCN1, FLG, RYR1, CSMD1, RYR3, and LRP1B. In all these combinations, there were not any significant mutually exclusive occurrences ($p< 0.05$).

This observation emphasized the fact that Co-occurring mutations preferentially occur in functionally related gene pairs to drive a disease condition; a typical example is the KRAS signaling pathway. To the best of my knowledge, there was no study that specifically looked at co-mutation in PDAC specifically, but other studies have looked into other cancer types such as lung cancer and multiple cancers as a whole, and observed that mutations co-occurrence is characteristic of the pathways involved. [56,57].

We next attempted to detect the PDAC driver gene using positional clustering. We observed that KRAS, ESX1, PROX2, PSG6, TP53, SF3B1, SMAD4, RNF150, CDKN2A, and METTL14 are the top driver genes identified through the positional clustering. Our observations are similar to the work of Hu et al., 2021 who explained that oncogenic mutations in PDAC driver genes such as KRAS and loss-of-function mutations in tumor suppressors, such as TP53, CDNK2A, DPC4/SMAD4, and BRCA2, are frequently observed[58]. We next explored the correlation between the top 20 mutated genes and overall survival. It was observed that patients with PEG3 gene mutation have 2 times the risk of death from PDAC than patient with wild type PEG3 while patients with FAT2 mutation have reduced risk of death from PDAC when compared to wild type. Yili et al., (2022) also observed that PEG3 gene mutations status correlates with total mutation burden and patient clinical outcomes in PDAC [59].

However, there was no previous publication on the role of the FAT2 gene and pancreatic cancer, although this gene has been linked to other cancer, for example, Li et al., (2017) showed that FAT2 to be a novel independent prognostic factor for the poor prognosis of gastric carcinoma [60].

Similarly, patients with KRAS and TP53 mutations have an increased risk of death from PDAC than patients with wild type genotype for these genes. To this end, the presence of an enriched oncogenic signaling pathway was checked in PDAC. About 10 significantly altered signaling pathways were identified (assessed for recurrent alterations within and across samples). RTK-RAS pathway is the top pathway that was enriched in about 84% of the samples, followed by TP53 and HIPPOS pathways with about 25% and 19% representations respectively (Figure 17).

The presence of the RTK-RAS pathway is not surprising, owing to the presence of KRAS and TP53 mutations which are key genes in cell development, tissue-specific cellular homeostasis, and cell differentiation,; any dysregulation in this gene will likely affect its hallmark functions [61,62]. At the same time, hippo signaling is a key regulator of organ size, tissue hemostasis, and regeneration. Dysregulation of the Hippo pathway has been recognized in a variety of human cancers, including pancreatic cancer [63]

Furthermore, in this study, the PDAC mutational signature was consistent with spontaneous or enzymatic deamination of 5-methylcytosine and DNA mismatch repair. The presence of deaminase enzymes could be an indicator of DNA damage proliferation in PDAC, whose expression could be used as a proxy indicator for early detection of PDAC [64]. Similarly, a recent study in Nature confirmed that 5-Hydroxymethylcytosine signatures in circulating cell-free DNA are diagnostic biomarkers for human cancers [65], further transcriptomics and epigenetic analysis will be required to completely understand the scope of this signature in PDAC.

In the same vein, the landscape of CNV in PDAC was also analyzed. It was observed that the pattern of the recurrent copy number amplification and deletions are widespread across the 23 chromosomes in PDAC. However, copy number deletion is the most common CNV (figure 23).

Interestingly, the chromosomes where the previously identified top mutated genes were located also showed significant copy number alterations. For example, chromosomes 17 which houses Tp53 genes showed copy number amplification for TP53 (figure 24), while chromosome 12 which houses KRAS showed copy number deletion.

Finally, this study looked at the clinical consequence of the CNV types that were observed in PDAC. Overall, we observed that copy number loss of heterozygosity of (TP53, KRAS, SMAD4, RYR3) was associated with reduced risk of death from PDAC, while KRAS and RYR3 CNV loss associated with an increased risk of deaths (Figure 22 & 23). Since in CNV-LOH, the biological system put in place mechanisms to ensure adequate expression of the genes to meet body needs through coordinated gene expression regulations. Unfortunately, the converse of such mechanism doesn’t work for copy number gain.

While other studies have associated actual copy number values independently and in association with other omics data with survival of PDAC, there was no publication that mentioned the role of CNV genes LOH on PDAC survival, however this is important to know, considering the strong positive association we observed between CNV LOH and PDAC prognosis.

5.1. Strengths

The major strength of this study is the unprecedented PDAC sample size. Such scale of genome data improves the reliability and quality of genomic data analysis.

Also, the presence of transcriptomic and epigenomcs data of similar scale for the same subjects will facilitate the continuation of this research at other omics level without worries about quality control process or bias due to sample sources.

The scope of the data spanned through different regions of the world; such data will help to further spread any genomic variations within each region, and also allow easy generalization of research findings.

5.2. Limitations

The key limitation of this study is the high number of missing values for some key clinical correlates, for example, tumor status at the last follow-up, tumor stage etc. Such missingness affects the inclusion of such variables as covariates in the survival model.

Also, the classification of tumor is not consistent between study sites, thus making it hard to compare tumor classes between sites. This limitation further affected the number of samples available to compare mutation profiles between stages of cancer.

Finally, the fact that tumor tissue samples were taken at admission, but no follow-up samples were taken, posed a challenge in assessing mutation clearance profile. As such, we could not infer residual or time point mutation profiles at follow-up.

6. CONCLUSION

This study analyzed PDAC SNV and CNV data from ICGC public database in order to further understand the landscape of genomic alterations in PDAC and add to the existing knowledge through the discovery of new gene alteration patterns. This study has identified interesting patterns, particularly the key driver genes of PDAC, affected biochemical pathways. As expected for multifactorial diseases such as PDAC, knowledge of genomic alteration is not enough to provide full scale understanding of the disease; however, analyzing genomic data at the genomic level can generate insightful research questions and hypotheses.

On this note, this study has generated a number of important questions and hypotheses that will require further analysis at other genomic levels. For example, this study has identified specific codons in the mutated genes; it will be important to understand how the mutation codon position affects prognosis. Similarly, this study has shown that over 80% of the PDAC cases have KRAS mutations, but only 25% of these KRAS mutated samples have TP53 mutations; it will also be interesting to understand the effect of this differential genomic alteration pattern on overall progression and prognosis, that is $KRAS^+$ $TP53^+$ versus $KRAS^+$ $TP53^-$ group.

This study further observed that SNV mutation landscapes vary with the stages of PDAC; it will be helpful to understand the association between these differences and disease subtypes. Furthermore, this study has shown that CNV LOH of driver genes is associated with prognosis; it will be interesting to understand the biology behind this observation. Lastly, it will be more helpful to integrate the knowledge from SNV and CNV into a single point value that can be used to infer the prognosis of PDAC.

REFERENCES

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA. Cancer J. Clin.* **71**, 7–33 (2021).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
3. Ferlay, J., Partensky, C. & Bray, F. More deaths from pancreatic cancer than breast cancer in the EU by 2017. *Acta Oncol. Stockh. Swed.* **55**, 1158–1160 (2016).
4. Miles, B. & Tadi, P. *Genetics, Somatic Mutation.* (StatPearls, 2022).
5. Bertram, J. S. The molecular biology of cancer. *Mol. Aspects Med.* **21**, 167–223 (2000).
6. Zhan, Q. *et al.* Identification of copy number variation-driven molecular subtypes informative for prognosis and treatment in pancreatic adenocarcinoma of a Chinese cohort. *EBioMedicine* **74**, 103716 (2021).
7. Huang, L. *et al.* Copy number variation at 6q13 functions as a long-range regulator and is associated with pancreatic cancer risk. *Carcinogenesis* **33**, 94–100 (2012).
8. Cicenas, J. *et al.* KRAS, TP53, CDKN2A, SMAD4, BRCA1, and BRCA2 Mutations in Pancreatic Cancer. *Cancers* **9**, (2017).
9. Masetti, M. *et al.* Long-term survivors of pancreatic adenocarcinoma show low rates of genetic alterations in KRAS, TP53 and SMAD4. *Cancer Biomark. Sect. Dis. Markers* **21**, 323–334 (2018).
10. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

11. Capasso, M. *et al.* Epidemiology and risk factors of pancreatic cancer. *Acta Bio-Medica Atenei Parm.* **89**, 141–146 (2018).

12. Sung, H., Siegel, R. L., Rosenberg, P. S. & Jemal, A. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. *Lancet Public Health* **4**, e137–e147 (2019).

13. Bosetti, C. *et al.* Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **23**, 1880–1888 (2012).

14. Johansen, D. *et al.* Metabolic factors and the risk of pancreatic cancer: a prospective analysis of almost 580,000 men and women in the Metabolic Syndrome and Cancer Project. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **19**, 2307–2317 (2010).

15. Rainone, M., Singh, I., Salo-Mullen, E. E., Stadler, Z. K. & O'Reilly, E. M. An Emerging Paradigm for Germline Testing in Pancreatic Ductal Adenocarcinoma and Immediate Implications for Clinical Practice: A Review. *JAMA Oncol.* **6**, 764–771 (2020).

16. O'Connor, C. Human chromosome translocations and cancer. *Nature Education* https://www.nature.com/scitable/topicpage/human-chromosome-translocations-and-cancer-23487/ (2018).

17. Nambiar, M., Kari, V. & Raghavan, S. C. Chromosomal translocations in cancer. *Biochim. Biophys. Acta* **1786**, 139–152 (2008).

18. Ghadimi, B. M. *et al.* Specific chromosomal aberrations and amplification of the AIB1 nuclear receptor coactivator gene in pancreatic carcinomas. *Am. J. Pathol.* **154**, 525–536 (1999).

19. Li, K., Luo, H., Huang, L., Luo, H. & Zhu, X. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int.* **20**, 16 (2020).

20. Eatrides, J. M. *et al.* Microsatellite instability in pancreatic cancer. *J. Clin. Oncol.* **34**, e15753–e15753 (2016).

21. Mneimneh, S. Crossing over...Markov meets Mendel. *PLoS Comput. Biol.* **8**, 1–12 (2012).

22. Sharp, A. J., Cheng, Z. & Eichler, E. E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).

23. He, Q. *et al.* Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data. *Am. J. Cancer Res.* **4**, 394–410 (2014).

24. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

25. Guo, S. *et al.* The Landscape of Genetic Alterations Stratified Prognosis in Oriental Pancreatic Cancer Patients. *Front. Oncol.* **11**, 717989 (2021).

26. Yachida, S. *et al.* Clinical significance of the genetic landscape of pancreatic cancer and implications for identification of potential long-term survivors. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18**, 6339–6347 (2012).

27. Villarroel, M. C. *et al.* Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer. *Mol. Cancer Ther.* **10**, 3–8 (2011).

28. Xu, Z. *et al.* Clinical Impact of Molecular Subtyping of Pancreatic Cancer. *Front. Cell Dev. Biol.* **9**, 743908 (2021).

29. Scott, A. J. & Wilkinson, J. C. HNF1A, KRT81, and CYP3A5: three more straws on the back of pancreatic cancer? *Transl. Cancer Res.* **5**, S253–S256 (2016).

30. Iovanna, J., Mallmann, M. C., Gonçalves, A., Turrini, O. & Dagorn, J.-C. Current knowledge on pancreatic cancer. *Front. Oncol.* **2**, 6 (2012).

31. Turanli, B., Yildirim, E., Gulfidan, G., Arga, K. Y. & Sinha, R. Current State of "Omics" Biomarkers in Pancreatic Cancer. *J. Pers. Med.* **11**, (2021).

32. Kong, L. *et al.* Multi-omics analysis based on integrated genomics, epigenomics and transcriptomics in pancreatic cancer. *Epigenomics* **12**, 507–524 (2020).

33. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

34. Keightley, P. D. Rates and fitness consequences of new mutations in humans. *Genetics* **190**, 295–304 (2012).

35. Wheeler, H. E., Maitland, M. L., Dolan, M. E., Cox, N. J. & Ratain, M. J. Cancer pharmacogenomics: strategies and challenges. *Nat. Rev. Genet.* **14**, 23–34 (2013).

36. He, Y., Hoskins, J. M. & McLeod, H. L. Copy number variants in pharmacogenetic genes. *Trends Mol. Med.* **17**, 244–251 (2011).

37. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).

38. Wang, H., Chen, D., Pan, Q. & Hueman, M. T. Using Weighted Differences in Hazards as Effect Sizes for Survival Data. *J. Stat. Theory Pract.* **16**, (2022).

39. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

40. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).

41. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **5**, 1542 (2016).

42. Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.* **2**, 54 (2010).

43. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinforma. Oxf. Engl.* **29**, 2238–2244 (2013).

44. Nair, A., Chauhan, P., Saha, B. & Kubatzky, K. F. Conceptual Evolution of Cell Signaling. *Int. J. Mol. Sci.* **20**, (2019).

45. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e10 (2018).

46. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

47. Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr. Protoc. Bioinforma.* **56**, 15.9.1-15.9.17 (2016).

48. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).

49. Brown, T. A. *Genomes. Mutation, Repair and Recombination.* (Oxford: Wiley-Liss;, 2002).

50. Kryh, H. *et al.* Comprehensive SNP array study of frequently used neuroblastoma cell lines; copy neutral loss of heterozygosity is common in the cell lines but uncommon in primary tumors. *BMC Genomics* **12**, 443 (2011).

51. McWilliams, R. R. *et al.* Risk Factors for Early-Onset and Very-Early-Onset Pancreatic Adenocarcinoma: A Pancreatic Cancer Case-Control Consortium (PanC4) Analysis. *Pancreas* **45**, 311–316 (2016).

52. Wang, H. *et al.* Survival of pancreatic cancer patients is negatively correlated with age at diagnosis: a population-based retrospective study. *Sci. Rep.* **10**, 7048 (2020).

53. Rawla, P., Sunkara, T. & Gaduputi, V. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World J. Oncol.* **10**, 10–27 (2019).

54. McGuire, S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. *Adv. Nutr. Bethesda Md* **7**, 418–419 (2016).

55. Bazzichetto, C. *et al.* Morphologic and Molecular Landscape of Pancreatic Cancer Variants as the Basis of New Therapeutic Strategies for Precision Oncology. *Int. J. Mol. Sci.* **21**, (2020).

56. Cui, Q. A network of cancer genes with co-occurring and anti-co-occurring mutations. *PloS One* **5**, e13180 (2010).

57. Jiang, L. *et al.* Comprehensive Analysis of Co-Mutations Identifies Cooperating Mechanisms of Tumorigenesis. *Cancers* **14**, (2022).

58. Hu, H.-F. *et al.* Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications. *Acta Pharmacol. Sin.* **42**, 1725–1741 (2021).

59. Huang, Y., Liu, J. & Zhu, X. Mutations in lysine methyltransferase 2C and PEG3 are associated with tumor mutation burden, prognosis, and antitumor immunity in pancreatic adenocarcinoma patients. *Digit. Health* **8**, 20552076221133700 (2022).

60. Li, L. *et al.* FAT2 is a novel independent prognostic factor for the poor prognosis of gastric carcinoma. *Int. J. Clin. Exp. Pathol.* **10**, 11603–11609 (2017).

61. Pudewell, S., Wittich, C., Kazemein Jasemi, N. S., Bazgir, F. & Ahmadian, M. R. Accessory proteins of the RAS-MAPK pathway: moving from the side line to the front line. *Commun. Biol.* **4**, 696 (2021).

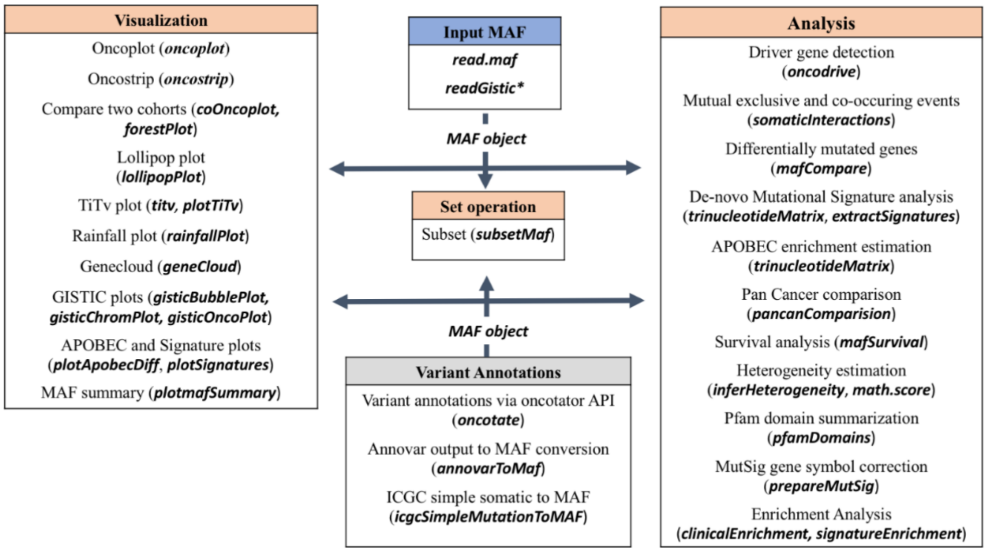
62. Pettazzoni, P. *et al.* Genetic events that limit the efficacy of MEK and RTK inhibitor therapies in a mouse model of KRAS-driven pancreatic cancer. *Cancer Res.* **75**, 1091–1101 (2015).

63. Ansari, D. *et al.* The Hippo Signaling Pathway in Pancreatic Cancer. *Anticancer Res.* **39**, 3317–3321 (2019).

64. Guler, G. D. *et al.* Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nat. Commun.* **11**, 5270 (2020).

65. Li, W. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res.* **27**, 1243–1257 (2017).

APPENDIX A. OVERVIEW OF THE MAFTOOLS PACKAGE


Visualization
Oncoplot (oncoplot)
Oncostrip (oncostrip)
Compare two cohorts (coOncoplot, forestPlot)
Lollipop plot (lollipopPlot)
TiTv plot (titv, plotTiTv)
Rainfall plot (rainfallPlot)
Genecloud (geneCloud)
GISTIC plots (gisticBubblePlot, gisticChromPlot, gisticOncoPlot)
APOBEC and Signature plots (plotApobecDiff, plotSignatures)
MAF summary (plotmafSummary)
Input MAF
read.maf
readGistic*
MAF object
Set operation
Subset (subsetMaf)
MAF object
Variant Annotations
Variant annotations via oncotator API (oncotate)
Annovar output to MAF conversion (annovarToMaf)
ICGC simple somatic to MAF (icgcSimpleMutationToMAF)
Analysis
Driver gene detection (oncodrive)
Mutual exclusive and co-occuring events (somaticInteractions)
Differentially mutated genes (mafCompare)
De-novo Mutational Signature analysis (trinucleotideMatrix, extractSignatures)
APOBEC enrichment estimation (trinucleotideMatrix)
Pan Cancer comparison (pancanComparision)
Survival analysis (mafSurvival)
Heterogeneity estimation (inferHeterogeneity, math.score)
Pfam domain summarization (pfamDomains)
MutSig gene symbol correction (prepareMutSig)
Enrichment Analysis (clinicalEnrichment, signatureEnrichment)

APPENDIX B. DATA PROCESSING

```
// RScript source code
## Data acquisition
baseurl <- "https://dcc.icgc.org/api/v1/download?fn=/release_28/Projects/"
url_US_cnv<- paste0(baseurl,"PAAD-US/copy_number_somatic_mutation.PAAD-US.tsv.gz")
url_AU_cnv<- paste0(baseurl,"PACA-AU/copy_number_somatic_mutation.PACA-AU.tsv.gz")
url_CA_cnv<- paste0(baseurl,"PACA-CA/copy_number_somatic_mutation.PACA-CA.tsv.gz")

for (i in c(url_US_cnv,url_AU_cnv,url_CA_cnv))

{
   if (!file.exists( substring(i,78,120))) {
    download.file(i,method="auto",destfile=substring(i,78,120), mode="ab")
   }

  gunzip(substring(i,78,117), destname = gsub("[.]gz$", "",substring(i,78,120)),
overwrite = TRUE, remove = FALSE)

  assign(substring(i,70,109),read.table (substring(i,78,117),sep = "\t", header = T))

}
ICGC_cnv <- rbind (`copy_number_somatic_mutation.PAAD-
US.tsv`,`copy_number_somatic_mutation.PACA-AU.tsv`,`copy_number_somatic_mutation.PACA-
CA.tsv`)

 CNV_data <-ICGC_cnv [,c(1:20)]

 rm(`copy_number_somatic_mutation.PAAD-US.tsv`,`copy_number_somatic_mutation.PACA-
AU.tsv`,`copy_number_somatic_mutation.PACA-CA.tsv`)

 saveRDS(CNV_data, file = "cnv.rds")

#### Get Mutation file

## Mutation

baseurl <- "https://dcc.icgc.org/api/v1/download?fn=/release_28/Projects/"
url_US_MUT<- paste0(baseurl,"PAAD-US/simple_somatic_mutation.open.PAAD-US.tsv.gz")
url_AU_MUT<- paste0(baseurl,"PACA-AU/simple_somatic_mutation.open.PACA-AU.tsv.gz")
url_CA_MUT<- paste0(baseurl,"PACA-CA/simple_somatic_mutation.open.PACA-CA.tsv.gz")

for (i in c(url_US_MUT,url_AU_MUT,url_CA_MUT))

{

   destfile= substring(i,70,112)


 if (!file.exists( substring(i,70,112))) {

    download.file(i,method="auto",destfile=substring(i,70,112), mode="ab")
   }
  assign(substring(i,70,105),icgcSimpleMutationToMAF(icgc = substring(i,70,112),
addHugoSymbol = TRUE))
 }
```

```
// RScript source code

snv.us=`simple_somatic_mutation.open.PAAD-US`
snv.au=`simple_somatic_mutation.open.PACA-AU`
snv.ca= `simple_somatic_mutation.open.PACA-CA`
snv.all= rbind(snv.us,snv.au,snv.ca)

rm(snv.us,snv.au,snv.ca)
rm(`simple_somatic_mutation.open.PAAD-US`,`simple_somatic_mutation.open.PACA-
AU`,`simple_somatic_mutation.open.PACA-CA`)

saveRDS(snv.all, file = "snv.all.rds")

#ICGC_mut<- ICGC_mut[,c(1,5:7,11:13,22:23,31:34)]

# clinical
baseurl <- "https://dcc.icgc.org/api/v1/download?fn=/release_28/Projects/"
url_US_clin<- paste0(baseurl,"PAAD-US/donor.PAAD-US.tsv.gz")
url_AU_clin<- paste0(baseurl,"PACA-AU/donor.PACA-AU.tsv.gz")
url_CA_clin<- paste0(baseurl,"PACA-CA/donor.PACA-CA.tsv.gz")

for (i in c(url_US_clin,url_AU_clin,url_CA_clin)){

   if (!file.exists( substring(i,70,112))) {
    download.file(i,method="auto",destfile=substring(i,70,90), mode="ab")
   }

if (!file.exists( substring(i,70,90))) {

  gunzip(substring(i,70,90), destname = gsub("[.]gz$", "",substring(i,70,90)),
overwrite = TRUE, remove = TRUE)
  #read data
}
   assign(substring(i,70,82),read.table (substring(i,70,86),sep = "\t", header = T))
  }

ICGC_clin <- rbind (`donor.PAAD-US`,`donor.PACA-AU`,`donor.PACA-CA`)

ICGC_clin$Overall_Survival_Status <- 0
ICGC_clin$Overall_Survival_Status[which(ICGC_clin$donor_vital_status == "deceased")]
<- 1
ICGC_clin$time <- ICGC_clin$donor_survival_time
ICGC_clin$time[is.na(ICGC_clin$donor_survival_time)] <-
ICGC_clin$donor_interval_of_last_followup[is.na(ICGC_clin$donor_survival_time)]

ICGC_clin<-ICGC_clin[,-c(3,16,20)]

saveRDS(ICGC_clin, file = "ICGC_clin.rds")
```

APPENDIX C. CLINICAL DATA ANALYSIS

```
// RScript source code
### Descriptives
id.s= unique(snv.us$icgc_donor_id)
id.c= unique(cnv.us$icgc_donor_id)
id.l= unique(clin.us$icgc_donor_id)
length(id.l)
length(id.c)
length(id.s)
length(intersect(id.l, id.s))
length(intersect(id.l, id.c))
length(intersect(id.c, id.s))
length(intersect(intersect(id.l, id.s), id.c))
c.all=  clin.all %>% select(project_code,disease_status_last_followup,
donor_relapse_type, donor_age_at_diagnosis, donor_sex, donor_vital_status)

t=c.all %>%
  gtsummary::tbl_summary(
    by = project_code
  ) %>%
  gtsummary::bold_labels() %>%
  gtsummary::as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = ""
    )

# plot
hist(clin.all %>% filter(project_code == "PACA-AU") %>%
pull(donor_age_at_diagnosis), breaks=10, xlim=c(0,90), col=rgb(1,0,0,0.5), xlab="Age
at diganosis",  ylab="Count", main="distribution of age by region")
hist(clin.all %>% filter(project_code == "PACA-CA") %>%
pull(donor_age_at_diagnosis), breaks=10, xlim=c(0,90), col=rgb(0,0,1,0.5), add=T)
hist(clin.all %>% filter(project_code == "PAAD-US") %>%
pull(donor_age_at_diagnosis), breaks=10, xlim=c(0,90), col=rgb(1,0,1,0.5), add=T)

# Add legend
legend("topright", legend=c("PACA-AU","PACA-CA","PAAD-US" ), col=c(rgb(1,0,0,0.5),
     rgb(0,0,1,0.5),rgb(1,0,1,0.5)), pt.cex=2, pch=15 )

pp <- plot(ggplot(clin.all, aes(x = donor_age_at_diagnosis,  color=project_code,
fill=project_code)) +
  geom_histogram(aes(y = (..count..)/sum(..count..)*100), alpha=0.6, binwidth = 5))
+
scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    ) +
    xlab("donor age at dignosis") +
    ylab(" (%)") +
    facet_wrap(~project_code)
```

```
ppp <- plot(ggplot(clin.all, aes(x = donor_age_at_diagnosis,  color=project_code,
fill=project_code)) +
  geom_histogram(aes(y = (..count..)/sum(..count..)*100), alpha=0.6, binwidth = 5))
+
scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    ) +
    xlab("Donor age at dignosis") +
    ylab(" Percentage (%)")
p <- clin.all %>%
  ggplot( aes(x=donor_age_at_diagnosis, color=project_code, fill=project_code)) +
    geom_histogram(alpha=0.6, binwidth = 5) +
    scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    ) +
    xlab("donor age at dignosis") +
    ylab(" (%)") +
    facet_wrap(~project_code)
pq <- clin.all %>%
  ggplot( aes(x=time, color=project_code, fill=project_code)) +
    geom_histogram(alpha=0.6, binwidth = 5) +
    scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    ) +
    xlab("Overall survival time") +
    ylab("Assigned Probability (%)") +
    facet_wrap(~project_code)
ppq <-suppressWarnings(plot(ggplot(clin.all, aes(x = time,  color=project_code,
fill=project_code)) +
  geom_histogram(aes(y = (..count..)/sum(..count..)*100), alpha=0.6, binwidth =
100)) +
scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 10)
    ) +
    xlab("Overall survival time") +
    ylab(" Percentage (%)"))
```

```
os <- clin.all %>%
  ggplot( aes(x=Overall_Survival_Status, color=project_code, fill=project_code)) +
    geom_histogram(alpha=0.6, binwidth = 5) +
    scale_fill_viridis(discrete=TRUE) +
    scale_color_viridis(discrete=TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    ) +
    xlab("Overall survival status") +
    ylab("Assigned Probability (%)") +
    facet_wrap(~project_code)


hist_info= hist(clin.all %>% filter(project_code == "PACA-AU") %>%
pull(donor_age_at_diagnosis), breaks=12, plot = FALSE )

hist_info$density <- hist_info$counts /    # Compute density values
  sum(hist_info$counts) * 100
plot(hist_info, freq = FALSE, col=rgb(1,0,0,0.5), xlab="Age at diganosis",
ylab="Proportion (%)", main="Distribution of PDAC patients age at diagnosis by study
site")


hist_infob= hist(clin.all %>% filter(project_code == "PACA-CA") %>%
pull(donor_age_at_diagnosis), breaks=8, plot = FALSE )

hist_infob$density <- hist_infob$counts /    # Compute density values
  sum(hist_infob$counts) * 100

plot(hist_infob, freq = FALSE,  xlim=c(0,90), col=rgb(0,0,1,0.5), add=T)

hist_infoc= hist(clin.all %>% filter(project_code == "PAAD-US") %>%
pull(donor_age_at_diagnosis), breaks=6, plot = FALSE )

hist_infoc$density <- hist_infoc$counts /    # Compute density values
  sum(hist_infoc$counts) * 100

#plot(hist_infoc, freq = FALSE, xlim=c(0,90), col=rgb((0,1,1,0.5), add=T) )

# Add legend
legend("topright", legend=c("PACA-AU","PACA-CA","PAAD-US" ), col=c(rgb(1,0,0,0.5),

rgb(0,0,1,0.5),rgb(0,1,1,0.5)), pt.cex=2, pch=15 )

sfit <- survfit(Surv(time, Overall_Survival_Status)~project_code, data=
clin.all[clin.all$time >= 10 & clin.all$time <= 2000,])
```

```
ggsurvplot(sfit, conf.int=FALSE, pval=TRUE, risk.table=TRUE,
           #legend.labs=c("Male", "Female"),
           legend.title="Position of Mutation",
           #palette=c("dodgerblue2", "orchid2"),
           title=("Overall survival pattern stratified by project site"),
           risk.table.height=.3)

prosit=ggsurvplot(
   sfit,                     # survfit object with calculated statistics.
   risk.table = TRUE,       # show risk table.
   pval = TRUE,             # show p-value of log-rank test.
   conf.int = FALSE,         # show confidence intervals for
                            # point estimaes of survival curves.
   xlim = c(0,2000),        # present narrower X axis, but not affect
                            # survival estimates.
   break.time.by = 250,    # break X axis in time intervals by 500.
   #ggtheme = theme_RTCGA(), # customize plot and risk table with a theme.
 risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE, # show bars instead of names in text annotations
                            # in legend of risk table
 title=("Overall survival pattern stratified by project site")
 )
```

APPENDIX D. SNV DATA ANALYSIS

```
/ RScript source code
snv.all =
readRDS("C:/Users/17018/Documents/NDSU/ExpScore/pub/msthesis/snv.all.rds")
  # Filtering mutations in gliomas
 EA_pathways <- TCGAbiolinks:::listEA_pathways
pdac_pathways <- EA_pathways[grep("pancreatic", tolower(EA_pathways$Pathway)),]

 pdac_signaling_genes <- unlist(strsplit(as.character(pdac_pathways$Molecules),","))

snv.all <-snv.all[snv.all$Hugo_Symbol %in% pdac_signaling_genes,]


snv=  snv.all %>% select("Hugo_Symbol", "Variant_Classification",  "Variant_Type",
"Reference_Allele" ,    "Tumor_Seq_Allele1", "Tumor_Seq_Allele2",
"consequence_type", "project_code" , "aa_mutation")


s.tv= snv  %>% select(Variant_Type, project_code) %>%
  gtsummary::tbl_summary(
    by = project_code
  ) %>%
  gtsummary::bold_labels() %>%
  gtsummary::as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = ""
    )

clin.all=readRDS("C:/Users/17018/Documents/NDSU/ExpScore/pub/msthesis/ICGC_clin.rds"
)

colnames(clin.all)[1]= "Tumor_Sample_Barcode"
colnames(snv.all)[16]= "Sample_Barcode"
colnames(snv.all)[22]= "Tumor_Sample_Barcode"
clin.all$time = as.numeric(clin.all$time)
clin.all = clin.all[clin.all$time >=5,]

clin.all =clin.all[!is.na(clin.all$time),]
#library("GenVisR")
clin.all$time = as.numeric(clin.all$time)
laml = read.maf(maf = snv.all, clinicalData = clin.all)
usmf =   read.maf(maf = snv.all[snv.all$project_code == "PAAD-US",], clinicalData =
clin.all[clin.all$project_code == "PAAD-US",])
aumf=   read.maf(maf = snv.all[snv.all$project_code == "PACA-AU",], clinicalData =
clin.all[clin.all$project_code == "PACA-AU",])
camf=   read.maf(maf = snv.all[snv.all$project_code == "PACA-CA",], clinicalData =
clin.all[clin.all$project_code == "PACA-CA",])

plotmafSummary(maf = camf, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE,
titvRaw = FALSE)
```

```
#oncoplot for top ten mutated genes.
oncoplot(maf = laml, top = 10)

fabcolors = RColorBrewer::brewer.pal(n = 3,name = 'Spectral')
names(fabcolors) = c("PAAD-US", "PACA-CA", "PACA-AU")
fabcolors = list(project_code = fabcolors)

oncoplot(  maf = laml, top=20 ,  clinicalFeatures = 'project_code', sortByAnnotation
= TRUE, annotationColor =fabcolors, draw_titv = TRUE, pathways = "auto")


vc_cols = RColorBrewer::brewer.pal(n = 8, name = 'Paired')

names(vc_cols) = c(
  'Frame_Shift_Del',
  'Missense_Mutation',
  'Nonsense_Mutation',
  'Multi_Hit',
  'Frame_Shift_Ins',
  'In_Frame_Ins',
  'Splice_Site',
  'In_Frame_Del')

oncoplot(maf = usmf,
         top = 20,draw_titv = TRUE)

waterfall(laml, mainRecurCutoff = 0.06)

oncoplot(
  maf = usmf,
  draw_titv = TRUE,
  #pathways = pathways,
  clinicalFeatures = c('Overall_Survival_Status'),
  sortByAnnotation = TRUE,
 # additionalFeature = c("Tumor_Seq_Allele2", "C"),
  #leftBarData = aml_genes_vaf,
  leftBarLims = c(0, 100)
  #rightBarData = laml.mutsig[,.(gene, q)],
 )

oncoplot(
  maf = camf,
  draw_titv = TRUE,
  #pathways = pathways,
  clinicalFeatures = c( 'Overall_Survival_Status'),
  sortByAnnotation = TRUE,
 # additionalFeature = c("Tumor_Seq_Allele2", "C"),
  #leftBarData = aml_genes_vaf,
  leftBarLims = c(0, 100)
  #rightBarData = laml.mutsig[,.(gene, q)],
 )
```

```
oncoplot(
  maf = aumf,
  draw_titv = TRUE,
  #pathways = pathways,
  clinicalFeatures = c('Overall_Survival_Status'),
  sortByAnnotation = TRUE,
 # additionalFeature = c("Tumor_Seq_Allele2", "C"),
  #leftBarData = aml_genes_vaf,
    leftBarLims = c(0, 100)
  #rightBarData = laml.mutsig[,.(gene, q)],
 )
somaticInteractions(maf = laml, top = 15, pvalue = c(0.05, 0.1))

somaticInteractions(maf = usmf, top = 15, pvalue = c(0.05, 0.1))

somaticInteractions(maf = aumf, top = 15, pvalue = c(0.05, 0.1))

somaticInteractions(maf = camf, top = 15, pvalue = c(0.05, 0.1))

sv= c("Nonsense_Mutation", "Frame_Shift_Ins", "Frame_Shift_Del",
"Translation_Start_Site", "Splice_Site", "Nonstop_Mutation", "In_Frame_Ins",
"In_Frame_Del", "Missense_Mutation", "5'Flank", "3'Flank", "5'UTR", "3'UTR", "RNA",
"Intron", "IGR", "Silent")

 x=  snv.all %>% select(Tumor_Sample_Barcode, Hugo_Symbol,Variant_Classification )
%>% filter(Variant_Classification %in% sv ) %>% as.data.frame()

prog_geneset = survGroup(maf = laml, top = 50, geneSetSize = 1, time = "time",
Status = "Overall_Survival_Status", verbose = FALSE)
prog_geneset= prog_geneset[prog_geneset$P_value <= 0.1,]
usg=c("KRAS", "SCN5A", "FAT4", "PREX1")
asg=c("BAI3", "FAT4", "COL6A5", "LRP2", "DCHS1","FLG")
cag= c("MUC19", "ZNF536", "PAPPA2")
mafSurvival(maf = laml, genes = 'KRAS', time = "time", Status =
"Overall_Survival_Status", isTCGA = TRUE)

mafSurvGroup(maf = laml, geneSet = c("KRAS", "TP53", "TTN"), time = "time", Status =
"Overall_Survival_Status")

laml.sig = oncodrive(maf = laml, AACol = 'aa_mutation', minMut = 2, pvalMethod =
'zscore')

##signature

library("BSgenome.Hsapiens.UCSC.hg19", quietly = TRUE)
laml.tnm = trinucleotideMatrix(maf = laml, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")
aumf.tnm = trinucleotideMatrix(maf = aumf, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

camf.tnm = trinucleotideMatrix(maf = camf, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

usmf.tnm = trinucleotideMatrix(maf = usmf, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = laml, pVal = 0.2)
plotApobecDiff(tnm = aumf.tnm, maf = aumf, pVal = 0.2)
```

```
library('NMF')
library('pheatmap')

##########ALL#################

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)

#Compate against original 30 signatures
laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")

#Compate against updated version3 60 signatures
laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE,
main = "cosine similarity against validated signatures")

maftools::plotSignatures(nmfRes = laml.sig, title_size = 1.2, sig_db = "SBS")

####################AU####################################################
aumf.sign = estimateSignatures(mat = aumf.tnm, nTry = 6)

aumf.sig = extractSignatures(mat = aumf.tnm, n = 3)

#Compate against original 30 signatures
aumf.og30.cosm = compareSignatures(nmfRes = aumf.sig, sig_db = "legacy")

#Compate against updated version3 60 signatures
aumf.v3.cosm = compareSignatures(nmfRes = aumf.sig, sig_db = "SBS")

pheatmap::pheatmap(mat = aumf.og30.cosm$cosine_similarities, cluster_rows = FALSE,
main = "cosine similarity against validated signatures")

maftools::plotSignatures(nmfRes = aumf.sig, title_size = 1.2, sig_db = "SBS")

camf.sign = estimateSignatures(mat = camf.tnm, nTry = 6)

camf.sig = extractSignatures(mat = camf.tnm, n = 3)

#Compate against original 30 signatures
camf.og30.cosm = compareSignatures(nmfRes = camf.sig, sig_db = "legacy")

#Compate against updated version3 60 signatures
camf.v3.cosm = compareSignatures(nmfRes = camf.sig, sig_db = "SBS")
pheatmap::pheatmap(mat = camf.og30.cosm$cosine_similarities, cluster_rows = FALSE,
main = "cosine similarity against validated signatures")
maftools::plotSignatures(nmfRes = camf.sig, title_size = 1.2, sig_db = "SBS")

usmf.sign = estimateSignatures(mat = usmf.tnm, nTry = 6)
usmf.sig = extractSignatures(mat = usmf.tnm, n = 3)
#Compate against original 30 signatures
usmf.og30.cosm = compareSignatures(nmfRes = usmf.sig, sig_db = "legacy")
#Compate against updated version3 60 signatures
usmf.v3.cosm = compareSignatures(nmfRes = usmf.sig, sig_db = "SBS")
pheatmap::pheatmap(mat = usmf.og30.cosm$cosine_similarities, cluster_rows = FALSE,
main = "cosine similarity against validated signatures")
maftools::plotSignatures(nmfRes = usmf.sig, title_size = 1.2, sig_db = "SBS")
```

APPENDIX E. CNV DATA ANALYSIS

```
// RScript source code
clin.all=readRDS("C:/Users/17018/Documents/NDSU/ExpScore/pub/msthesis/ICGC_clin.rds"
)
cnv.all=readRDS("C:/Users/17018/Documents/NDSU/ExpScore/pub/msthesis/cnv.rds")

cnv= cnv.all[,c(1,2,12,13,14),]; rm(cnv.all)
library(GenomicRanges)
# Get gene information from GENCODE using biomart
genes <- TCGAbiolinks:::get.GRCh.bioMart(genome = "hg19")
genes <- genes[genes$external_gene_name != "" & genes$chromosome_name %in%
c(1:22,"X","Y"),]
genes[genes$chromosome_name == "X", "chromosome_name"] <- 23
genes[genes$chromosome_name == "Y", "chromosome_name"] <- 24
genes$chromosome_name <- sapply(genes$chromosome_name,as.integer)
genes <- genes[order(genes$start_position),]
genes <- genes[order(genes$chromosome_name),]
genes <- genes[,c("external_gene_name", "chromosome_name",
"start_position","end_position")]
colnames(genes) <- c("GeneSymbol","Chr","Start","End")
genes_GR <- makeGRangesFromDataFrame(genes,keep.extra.columns = TRUE)

cnv= cnv[,c(3,4,5,1,2)]
colnames(cnv) <- c("Chr","Start","End","sid", "proj")

cnv[cnv$Chr == "X", "Chr"] <- 23
cnv[cnv$Chr == "Y", "Chr"] <- 24

sCNV_GR <- makeGRangesFromDataFrame(cnv,keep.extra.columns = TRUE)
hits <- findOverlaps(genes_GR, sCNV_GR, type = "within")
sCNV_ann <- cbind(cnv[subjectHits(hits),],genes[queryHits(hits),])
sCNV_ann <- sCNV_ann [,c(1,2,3,4,5,6)]
sCNV_ann <- sCNV_ann  %>% distinct()

sCNV_ann <- sCNV_ann %>% mutate(n=nchar(GeneSymbol)) %>% filter(n <=5)

############I tried with ICGC , but failed
mycnv= cnv.all[,c(1,2,12,13,14,9, 10, 11, 18,19)]
mycnv$segmean= log2(mycnv$copy_number/ 2)

mycnv=mycnv[mycnv$project_code == "PACA-CA",][,-2]

mycnv$probe=(mycnv$end_probe_id - mycnv$start_probe_id) +1

mycnv= mycnv[,c(1:4,11,10)]

colnames(mycnv)= c("Sample", "Chromosome","Start","End","Num_Probes","Segment_Mean")

pdac.nocnv <-mycnv;rm(mycnv)
pdac.nocnv=pdac.nocnv[pdac.nocnv$Segment_Mean !="-Inf",]
pdac.nocnv=pdac.nocnv[pdac.nocnv$Segment_Mean !="NaN",]
Tquery.pdac.nocnv <- GDCquery(project = "TCGA-PAAD",
                    data.category = "Copy number variation",
                    data.type = "Copy number segmentation",
                    legacy = TRUE,
                    file.type = "nocnv_hg19.seg",
                    sample.type = c("Primary Tumor"))
GDCdownload(Tquery.pdac.nocnv, files.per.chunk = 100)
pt.nocnv <- GDCprepare(Tquery.pdac.nocnv, )
pdac.nocnv <- pt.nocnv;rm(pt.nocnv)
# pdac.nocnv <- pdac.nocnv[,-1]
# pdac.nocnv <- pdac.nocnv[,c(6,1,2,3,4,5)]

rm(Tquery.pdac.nocnv)

# Add label (0 for loss, 1 for gain)
cnvMatrix <- cbind(pdac.nocnv,Label=NA)
cnvMatrix[cnvMatrix[,"Segment_Mean"] < -0.3,"Label"] <- 0
cnvMatrix[cnvMatrix[,"Segment_Mean"] > 0.3,"Label"] <- 1
cnvMatrix <- cnvMatrix[!is.na(cnvMatrix$Label),]
#head(cnvMatrix)
# Remove "Segment_Mean" and change col.names
cnvMatrix <- cnvMatrix[,-c(6)]
colnames(cnvMatrix) <- c("Sample.Name", "Chromosome", "Start", "End",
"Num.of.Markers","Aberration")
head(cnvMatrix)
# Substitute Chromosomes "X" and "Y" with "23" and "24"
cnvMatrix[cnvMatrix$Chromosome == "X","Chromosome"] <- 23
```