An Investigation of Point Mutations Discovers Novel Genes and Their Corresponding Motifs in Pancreatic Cancer

Amin Ghareyazi[1], Amir Mohseni[1], Hamed Dashti[1], Abdollah Dehzangi*[2,3], Amin Beheshti[4], Hamid R. Rabiee*[1], Hamid Alinejad-Rokny* [5,6,7]

[1] Bioinformatics and Computational Biology Lab, Sharif University of Technology, Tehran, IR
[2] Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[3] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA.
[4] Department of Computing, Macquarie University, Sydney, NSW 2109, AU.
[5] BioMedical Machine Learning Lab (BML), The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, 2052, AU.
[6] Core Member of UNSW Data Science Hub, The University of New South Wales (UNSW Sydney), Sydney, 2052, AU.
[7] Health Data Analytics Program Leader, AI-enabled Processes (AIP) Research Centre, Macquarie University, Sydney, 2109, AU.
* Correspondence: hamid.alinejad@mq.edu.au; Tel: +61 2 9385 3911; rabiee@sharif.edu; i.dehzangi@rutgers.edu

Abstract: It has now known that at least 10% of samples with pancreatic cancers (PC) contain a causative mutation in the known susceptibility genes, suggesting the importance of identifying cancer-associated genes that carry the causative mutations in high-risk individuals for early detection of PC. In this study, we develop a statistical pipeline using a new concept, called gene-motif, that utilizes both mutated genes and mutational processes to identify 4,211 3-nucleotide PC-associated gene-motifs within 203 significantly mutated genes in PC. Using these gene-motifs as distinguishable features for pancreatic cancer subtyping results in identifying five PC subtypes with distinguishable phenotypes and genotypes. Our comprehensive biological characterization reveals that these PC subtypes are associated with different molecular mechanisms including unique cancer related signaling pathways, in which for most of the subtypes targeted treatment options are currently available. Some of the pathways we identified in all five PC subtypes, including cell cycle and the Axon guidance pathway are frequently seen and mutated in cancer. We also identified Protein kinase C, EGFR (epidermal growth factor receptor) signaling pathway and P53 signaling pathways as potential targets for treatment of the PC subtypes. Altogether, our results uncover the importance of considering both the mutation type and mutated genes in the identification of cancer subtypes and biomarkers.

Keywords: Pancreatic cancer, Cancer subtype identification, Somatic point mutations, Genotype and phenotype characterization, Therapeutic targets, Personalized medicine.

1. Introduction

Pancreatic cancer (PC) is the third leading cause of death among all cancers, with the lowest survival rate of 9% [1]. PC is predicted to become the second leading fatal cancer [2]. Besides, the advancement achieved in increasing survival time for Lung and Pancreatic cancers has been slow compared to other types of cancers [1]. PC can be categorized into different subtypes based on specifications of mutations, molecular profile, and histopathological characteristics. Such subtypes can have different mechanisms and different responses to treatments [3]. Therefore, identifying subtypes can lead to the identification of unique biomarkers, more effective treatment approaches, and also directly contributing to personalized medicine. Identification of subtypes for breast [4] and lung [5] cancers has led to finding new effective treatments, and better-targeted drugs. Moreover, determining subtypes can potentially play a vital role in increasing prognostic accuracy for pancreatic cancer.

During the last decade, a wide range of studies has been conducted to identify corresponding pancreatic cancer subtypes with a special focus on gene expression profiles as features [6]. In 2011, Collisson *et al*. proposed a combined analysis to tackle the limitations of the number of tumor samples for PC subtype identification [7]. They used combined analysis of transcriptional profiles of primary Pancreatic Ductal Adenocarcinoma (PDAC is an exocrine type of pancreatic cancer) from several studies, along with the human and mouse PDAC cell lines. By using gene expression, They identified three subtypes and 62-gene signatures for PC [7]. In 2015, Moffit *et al*. expanded the Collison *et al*. work by adding stromal classifications [8]. They also employed the global gene expression analysis with RNA sequencing validation and proposed two subtypes for each stroma-specific and tumor-specific group. Remarkably, they reported an overlap between one of their identified

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

tumor-specific subtypes called "classical" and the Collisson *et al*. classical subtype [8]. Both of these studies were served as the basic foundation of the Bailey *et al*. research [9]. They proposed an integrated genomic analysis by using deep-exome and whole-genome with gene copy number analysis, along with RNA-seq validation. They identified four subtypes, namely, squamous, pancreatic progenitor, immunogenic, and *Aberrantly Differentiated Endocrine Exocrine (ADEX)* for pancreatic cancer. Furthermore, they specified several gene-based categories according to similarities among their pathways [9]. In another study, Sivakumar *et al*. used expression profiles of 204 ICGC and 149 TCGA samples to tackle this problem [10]. Using a network-based and community detection method, they identify three main subtypes for PC. In their study, the focus was the activity and characteristics of the *KRAS* gene in PC. In one of the latest works on PC subtyping, Puelo *et al*. used gene expression of 309 resected primary PDAC and identified five different subtypes based on features of cancer cells and the tumor microenvironment [11].

A mentioned earlier, pancreatic subtype identification by using the gene expression data, is widely popular. However, gene expression is tissue and time-specific. It means that the gene expression of tissue can vary at different time points. Moreover, gene expressions of different tissues are different at a single time point [12]. Hence, relying on gene expression for cancer subtype identification might not provide a general and reliable result. On the other hand, somatic mutations, as important players in cancer development and disease progression, are less affected by factors that can influence gene expression [13].

Recently, Kuijjer *et al*. used somatic point mutations for identifying mutational diversities in pan-cancer to find new types of cancer among all cancers [14]. They classified patients with similar mutation profiles into subgroups by applying biological pathways [14]. In another pan-cancer research, Kuipers *et al*. proposed a method for finding subgroups of cancer-based on interactions of mutations [15]. In the field of pancreatic cancer subtype identification, Waddell *et al*. provided a pipeline for analysis of the pattern of structural variations (including copy number variations, somatic and germline mutations) in 100 PDAC samples [16]. They identified four main subtypes and named them as "stable", "locally rearranged", "scattered" and "unstable". They have not included any samples from the exocrine type (a rare type of PC) in their study.

In 2013, Alexandrov *et al.* published a paper and showed that there are 78 mutational signatures in cancers, most of them associated with a specific molecular mechanism to uncover the causality behind somatic point mutations across the genome [17]. The proposed concept provided the importance of motifs in the analysis of somatic point mutations in cancer genomics. To the best of our knowledge, nobody has used the context of mutations in highly mutated genes for cancer subtype identification. As we discussed above, multiple groups identified 3-5 PC subtypes, however, they did not consider the underlying mutational context to cluster affected patients. In this study, we perform an integrative analysis using "gene-motif" information extracted from somatic mutations to tackle this problem. We hypothesize that accurate PC subtypes identification depends on both mutations and their corresponding motifs as well as the respective mutated genes. Therefore, we proposed a feature called "gene-motif" to accurately identify subtypes in pancreatic cancer. We conduct our integrative analysis on the dataset from ICGC consortia consisting of 774 samples with PC. This dataset is by far larger than those used in the previous studies which demonstrate the comprehensiveness of this study, and generality of our findings. To build our model, we first identify candidate gene-motifs as our features to cluster the PC samples. Such features are selected based on the empirical distribution of the number of mutations in gene-motifs. After the candidate gene-motifs are identified, we use a model-based clustering approach for clustering the PC samples to identify the subtypes. We have identified five subtypes with distinguishable relations between candidate genes, phenotype, and genotype characteristics of PC subtypes. We have also identified subtype-specific mutational signatures and compared them with the latest COSMIC (34) mutational signatures to investigate the molecular mechanisms behind mutations in each subtype. We have also investigated the mutational load in coding genes to identify subtype-specific genes. Our gene ontology and pathway analyses also demonstrate common and subtype-specific terms. We next analyzed RNA-Seq gene expression data of PC samples and investigated the difference of gene expression between the identified subtypes. We also conducted a complete survival analysis and studied the effects of histopathological information on survival time prediction. An overview of the analysis pipeline used in this study is demonstrated in **Figure 1**. Our proposed model and its related codes are publicly available online at: https://github.com/bcb-sut/Pancreatic-Cancer-Subtype-Identification.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

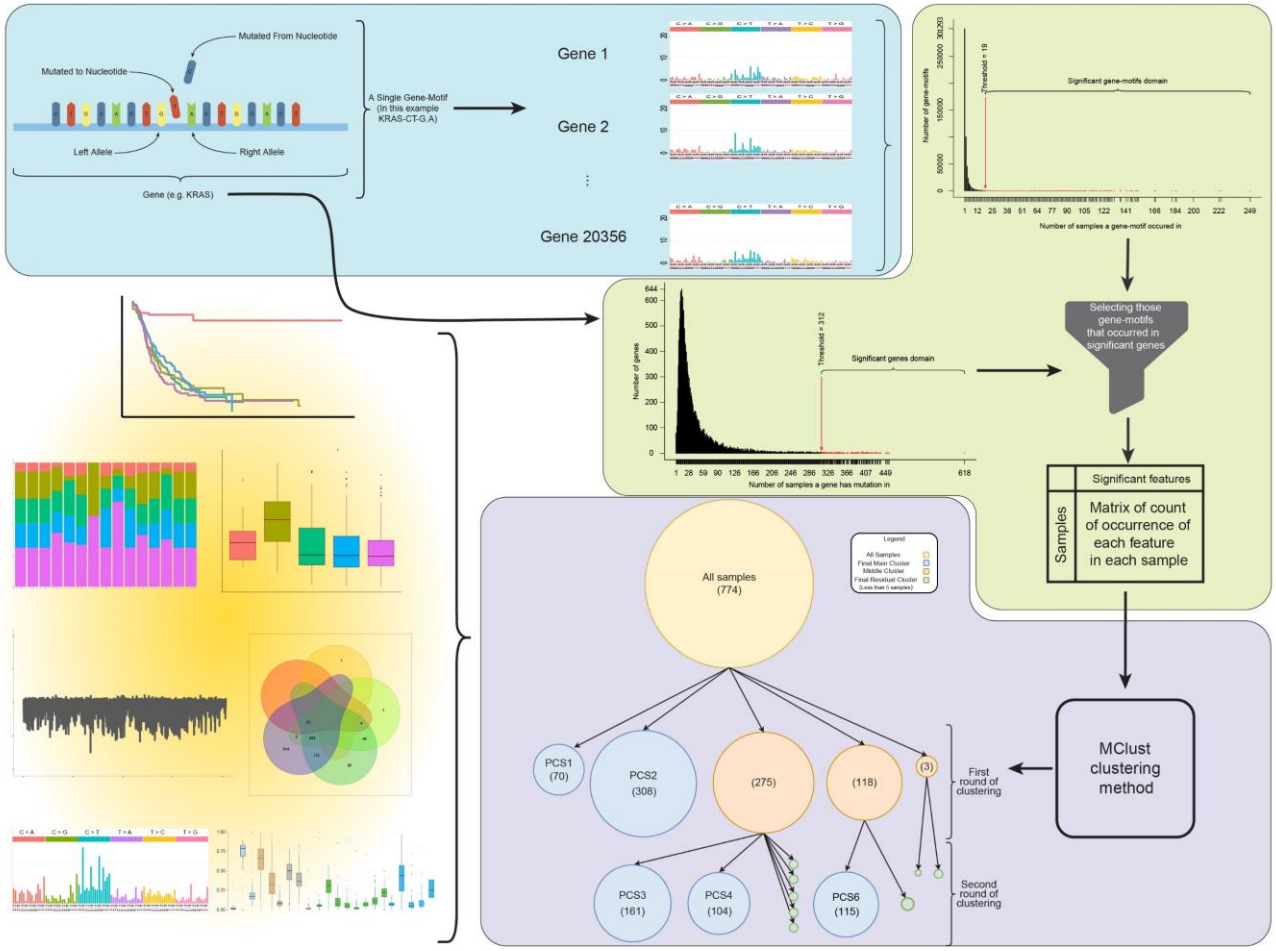


Figure 1. The workflow of Pancreatic Cancer Subtype Identification and Clustering tree**.** In top left an overall view of the 3-mer motif and the gene-motif concept is illustrated. At first, we construct features named gene-motifs based on the 3-mer motif and the gene that motif has occurred in. These features were constructed for all samples and in all of their protein-coding genes. In top right Feature selection process is illustrated. We calculated the number of samples each gene-motif has occurred in, and based on their distributions, we found the most frequent (and hence significant) gene-motifs. We also found the most frequent mutated genes or significantly mutated genes to filter out those gene-motifs that have not occurred in significant genes. This leads to significant features for clustering. In bottom left Clustering process and tree is drawn. After constructing a matrix of occurrence for each feature in each sample (each cell indicates whether a feature has occurred in a sample or not) the Mclust algorithm was employed to cluster samples into subtypes. after two rounds of clustering, 5 main subtypes have revealed themselves. Finally comprehensive genotype and phenotype characteristic study was performed to find differences and/or commonality in subtypes (bottom left). This includes gene association, mutational signature, deep mutational profile investigation, finding DEGs, survival analysis, etc.

2. Materials and Methods

2.1. Data

Simple somatic mutation data for all pancreatic cancer projects from ICGC were gathered from https://dcc.icgc.org/ in November 2017. This Dataset includes information of 17,284,164 simple somatic mutations of 827 samples. RNA-seq gene expression data of 534 PC samples were also available from the ICGC data.

2.2. Data cleaning & filtration

We collected all the information about the position of simple somatic mutation (SSM) from the ICGC dataset which includes chromosomal position, start and end position, gene ID, transcript ID, reference genome allele, and mutated allele. We also extracted the information such as consequence type and project code from ICGC for analysis. Only simple somatic point mutations were considered for further analyses.

To find the 3-nucleotide motifs, we used BSgenome.Hsapiens.UCSC.hg19, GenomicRnages [18], and SomaticSignatures [19] packages in the R programming language. 3-nucleotide motifs and their respective genes were concatenated as the clustering features. For the analysis of gene expression data, samples that had the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

same sample ID from five subtypes were used. Finally, the RNA-seq data from 307 samples remained for the downstream analysis. Consequently, PCS1 contains 22 samples and PCS2, PCS3, PCS4 and PCS5 contains 10, 108, 76, 91 samples, respectively. Some genes in this dataset had more than one value for some samples, and we used the mean value from them. Protein coding genes that had NA values were removed from the analysis.

2.3. Feature selection

We used gene-motifs [20], as the feature for our analyses. For each gene, we counted the number of mutated samples. Moreover, for each gene-motif, we counted the number of samples in which that gene-motif has occurred. Afterward, we used the empirical distribution of these counts to assign the probability of occurrence to each gene and gene-motif. In this step, genes and gene-motifs with probability <= 0.01 were considered to be significant genes and gene-motifs. For our dataset, the genes with mutations in at least 312 samples, constitute about 0.01 of all protein-coding genes. Also, those gene motifs that have been occurred in at least 19 samples, are about 0.01 of all possible gene-motifs. As a result, 203 genes and 5704 gene-motifs were above their respective criteria and considered to be significant gene and gene-motif, respectively. All the significant gene and gene-motifs are provided in **Supplementary Table S1** and **Supplementary Table S2.** Finally, the significant gene-motifs that their respective genes were members of the significant gene set, was selected as the clustering features. A matrix consists of all the 4211 candidate gene-motifs (**Supplementary Table S3**) for all samples, were used for clustering.

2.4. Clustering method

In this study, we used the Mclust method [21] to cluster samples. To identify the optimum number of clusters, Mclust utilizes a Gaussian mixture model with the Bayesian Information Criterion (BIC) that is the best criterion for finding optimal number of clusters in mixture models [22, 23]. We used the implementation of Mclust in the CRAN package. Mclust does not make any assumption on the parameters of distribution function for features, and it does not make any assumption on the number of clusters. These properties make it a suitable clustering method. Mclust was applied recursively until no meaningful new cluster was generated. A cluster is assumed to be meaningful if it contains at least 1% of the total number of samples (at least 7 samples). Hence, clusters with less than this threshold were outliers. To illustrate the segregation and differentiation between clusters we performed a PCA analysis. First two principal components of our data demonstrates that clusters are separated well (**Supplementary Figure S1).**
The results of our clustering method are shown in **Figure 1** On the first round of clustering, 5 clusters with a size of 70, 308, 275, 118, and 3 samples were found. On the second round of clustering, clusters with 70 and 308 samples did not break into smaller clusters. Hence, these two were considered as the main subtypes and were named PCS1 and PCS2. On the other hand, the cluster with 275 samples, split into 2 clusters with 161 (called PCS3), and 104 (called PCS4) samples, and 5 other clusters with 2 samples. The cluster with 118 samples that were found in the first round was divided into two clusters with 115 (called PCSS5) and 3 samples. Other small clusters with less than 7 samples were considered as outliers.

2.5. Differential analysis

We used the differential analysis to investigate differences in rates of samples with mutation, in the protein-coding gene. We counted the number of samples with a mutation in each gene, for all 5 subtypes. The rates of each gene in each cluster were deducted from its rate in other subtypes. The same process was performed on clustering features.

2.6. Mutational signature analysis

We used the CANCERSIGN package in R [24] to calculate the mutational signatures of pancreatic cancer, and the level of exposures of each sample to each signature. This tool implements the Non-negative Matrix Factorization (NMF) method to find patterns of 3-nucleotide motifs among samples. The signatures have been extracted for all pancreatic cancer samples as well as each subtype, individually. The input to CANCERSIGN is a matrix of samples in rows, and features (including chromosome, mutation position, reference allele, and mutated-to allele) in columns. The analysis was performed with the number of signatures ranging from 1 to 15, and the maximum bootstrap iterations for each step was set to 780. The cosine distance was used to compare the signatures. The evaluation plot of deciphering 3-mer mutational signatures is provided in **Supplementary Figure S2**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

2.7. Motif analysis

Each mutation and its context (left and right alleles of a mutated position), and the substituted nucleic acid in that position, constructs a 3-nucleotide motif. There are 96 combinations of 3-nucleotide motifs. Patterns of these 3-nucleotide motifs can provide important information about the molecular mechanism in biological [25-27]. The relative frequency of motifs was calculated cumulatively for subtypes (**Supplementary Figure S3**), and common associated genes (**Supplementary Table S4**). Motif rates of outlier clusters are provided in **Supplementary Figure S4**. Tests for the piqued motifs in common associated genes were done by utilizing the Fisher exact test. For example, we counted the number of samples that had the motif TA-A.A and also were in PCS1 (or PCS3) for the gene NRG1. We tested the relationship between these two dichotomous variables by Fisher's exact test.

2.8. Transcript type analysis

Each mutation can affect one or more transcripts of the gene. The differences in subtypes indicate different effects on the organisms. To investigate these rates, the relative frequency of samples in each subtype with a mutation in each transcript, for all protein-coding genes were calculated.

2.9. Gene association

The association of protein-coding genes to each subtype was done by utilizing Fisher's exact test. This test was applied to identify mutated genes as the potential biomarker for each subtype. To identify such association using Fisher's exact test we used a 2×2 contingency matrix. This matrix contains information relating to the number of samples in all possible combinations of two variables. These variables are: 1) being categorized as a member of a certain subtype or not, and 2) having at least one mutation in a given gene or not. This test was used for all genes in all subtypes. To find a significant threshold for p-values, a permutation test was conducted. To do this, first, a table of the number of mutated samples for each gene was randomly generated, such that their total number over all the genes remains the same. This table was created for all subtypes. Second, Fisher's exact test was conducted as described above on all genes and for all subtypes. Third, these steps were repeated 10,000 times. Fourth, for each gene, 10,000 p-values are generated. We considered the p-value of the lowest 0.05 percent of these numbers as the significance threshold. For the final step, we chose the genes that were mutated at least in 50% of samples of their respective subtype and considered them as associated genes to subtypes (**Supplementary Table S5**). A Venn diagram of common associated genes in subtypes is provided in **Supplementary Figure S5**.

2.10. Gene expression analysis

Raw read count of 19104 protein-coding genes from 307 samples was gathered in a matrix. The DESeq2 package and its guideline were used for finding differentially expressed genes (DEGs) between the groups [28]. Genes with a P-value of less than 0.05 were considered as significantly differentially expressed genes. First, significant DEGs of PCS1 were compared to all other subtypes. This was also done for other subtypes. Second, in five-set of DEGs, unique genes and common genes were distinguished, as shown in the Venn diagram of **Supplementary Figure S6**. Those genes that are only in the respective set of each subtype, are considered as uniquely differentially expressed genes (UDEGs).

2.11. Gene ontology and pathway

Gene ontology and pathway analyses were performed by using the Enrichr online tool (https://amp.pharm.mssm.edu/Enrichr/). Associated genes to each subtype were used as input to this tool. For the p-value adjustment, the Benjamini-Hochberg method was employed. Only ontologies with FDR < 0.05 were considered.

2.12. Gender and project code analysis

Project codes of the ICGC database contain information related to the types of pancreatic cancer and the region where the data is gathered. We can also retrieve the gender of donors in the meta-data of donors in this database. Here, this information was used to investigate the possible relation between subtypes, their living location, and their gender. We used genders and project codes in each subtype. Our samples were either male or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

female, and belong to 4 project codes, namely Pancreatic Cancer Ductal adenocarcinoma from Australia (PACA-AU), Pancreatic Cancer from Canada (PACA-CA), Pancreatic Cancer Endocrine neoplasms from Australia (PAEN-AU), and Pancreatic Endocrine neoplasms from Italy (PAEN-IT). Frequencies of genders and project codes are provided in **Supplementary Table S6** and **Supplementary Table S7**. We used the frequency of samples of each project in each subtype to investigate if any meaningful relationship between this information and subtypes can be observed.

2.13. Literature search for non-coding interacting genes

Our literature searches to identify cancer-associated genes were focused on human studies and English language publications available in the PubMed, Scopus, and Web of Science. We also used data and text mining techniques to extract additional related studies **[29-35]**. A decision tree approach and a knowledge-based filtering system technique have been also used to categorize the texts from the literatures search **[33, 36]**. The search terms included "noncoding RNA" or "lncRNAs" or "genes name + cancer". "BC" or "breast carcinoma" and "breast neoplasm".

2.14. Survival analysis

We used the donor survival time as the overall survival time for each donor, and vital status was used for the Kaplan-Maier method to estimate the overall survival of each subtype. To conduct survival analysis, we discarded the data for donors that had missing or NA values for survival time (or survival time with zero days) or vital status.

Here, we also report the mean and median of overall survival for each subtype and their 95% confidence interval in **Supplementary Table S8**. We also studied differences of overall survival between subtypes by using the log-rank test, Breslow test [37], and Taron-Ware test [38]. Pairs of subtypes with a p-value of less than 0.05 were considered to be unequal in terms of their survival curves. Results are provided in **Supplementary Table S9.**

We also used the Cox proportional hazards model to evaluate the prognosis power of subtype indicators for survival prediction [15]. We applied several models with adjustment for age at diagnosis, tumor stage, tumor grades, and subtype indicator variables to survival data. As a result, tumor stage categories were aggregated into 5 categories (stage I to IV and stage X; indicating samples with unknown staging status). We also assembled grade categories in 7 levels, namely, (1) well differentiated, (2) moderately differentiated, (3) poorly differentiated, (4) undifferentiated, (5) NET well differentiated, (6) NET moderately differentiated, and (7) NET poorly differentiated. Samples without information or false values (e.g., 0 or NA) were removed, leaving 625 samples for analysis.

A full model with adjustment for all variables is presented in **Supplementary Table S10**. We also report the p-values of the coefficient and likelihood ratio test of comparison with the null model (a model without predictor) for each case. We also test the effect of each variable on survival prediction in the complete model using the likelihood ratio test. A likelihood ratio test was conducted to compare the reduced model and complete model to assess the effect of the removed variable in the complete model. To do this, variables were removed from the complete model, one at a time, to generate the reduced model. A likelihood ratio test comparing the complete model and the reduced model was employed to evaluate the effect of the removed variables on the survival time. P-values of these tests are reported in **Supplementary Table S11**.

2.15. Comparing the overall mutation rate in subtypes

We conducted an independent t-test to explore the statistical differences between subtypes, in terms of the relative frequency of samples with mutations in significant genes, and all protein-coding genes. We also investigated the rate of mutation in significant gene-motifs, and significant features (**Supplementary Table S12**). After calculating the rate of samples that had a mutation in each protein-coding gene for all subtypes, these rates were tested in pairs, to compare means of mutation rates in subtypes. The same process was also carried out for significant genes, significant gene-motifs, and significant features.

3. Results

3.1. Background model to identify subtypes in pancreatic cancer

In this study, we used somatic point mutations in pancreatic cancer patients from the ICGC dataset. This dataset contains information on mutations of PC samples, based on the whole genome sequencing technology.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

These samples are collected from three different regions including Australia, Canada, and Italy. 57% of these samples are collected from male and 43% from female donors. After the data cleaning process, (*see methods section*) information of 774 samples were selected for further analysis.

Since somatic mutations are one of the important factors in the progression and development of cancers [13], they were the main point of attention for integrative analysis of PC subtype identification. There are different mutational processes in living organisms causing mutations across the genome. These molecular mechanisms may cause mutations based on the adjacent bases of a locus [39]. This means that mutational mechanisms may act differently based on neighboring positions of a locus. According to [40], there are 96 possible mutation types in 3-nucleotide motifs (the original nucleic acid, the nucleic acid which it has mutated to, and it's right and left alleles). Rates of each 3-nucleotide motif may vary in genes. To count each 3-nucleotide motif in each gene separately, we constructed a new feature called "gene-motif". For example, KRAS-CT-A.G refers to the number of CT-A.G 3-nucleotide motif in the KRAS gene. "CT-G.A" stands for a mutation point that its reference nucleobase is C, mutated to T, its right nucleobase is A, and its left nucleobase is G. The main reason for studying this type of variation is that by depending solely on genes we will overlook the information stored on mutations themselves. An overview of the gene-motif context is shown in **Figure 1**.

In the first step, we used empirical distributions to select 4,211 significant gene-motifs as the features for clustering (*see methods section*). A full list of clustering features is provided in **Supplementary Table S3**. We then adopted the *Mclust* algorithm [21] (*see methods section*) and identified five main clusters in pancreatic cancer individuals, which we refer to them as PCS1, PCS2, PCS3, PCS4, and PCS5. A PCA analysis also confirmed that clusters are well identified and separated (**Supplementary Figure 1**). The clustering tree for the process of finding these subtypes is available in **Figure 1**. As the figure shows, there are 70, 308, 161, 104, and 115 samples for subtypes PCS1, PCS2, PCS3, PCS4, and PCS5, respectively.

In the following sections, genotype and phenotype characteristics of these five PC subtypes are analyzed and discussed in detail to represent their unique and differentiative properties.

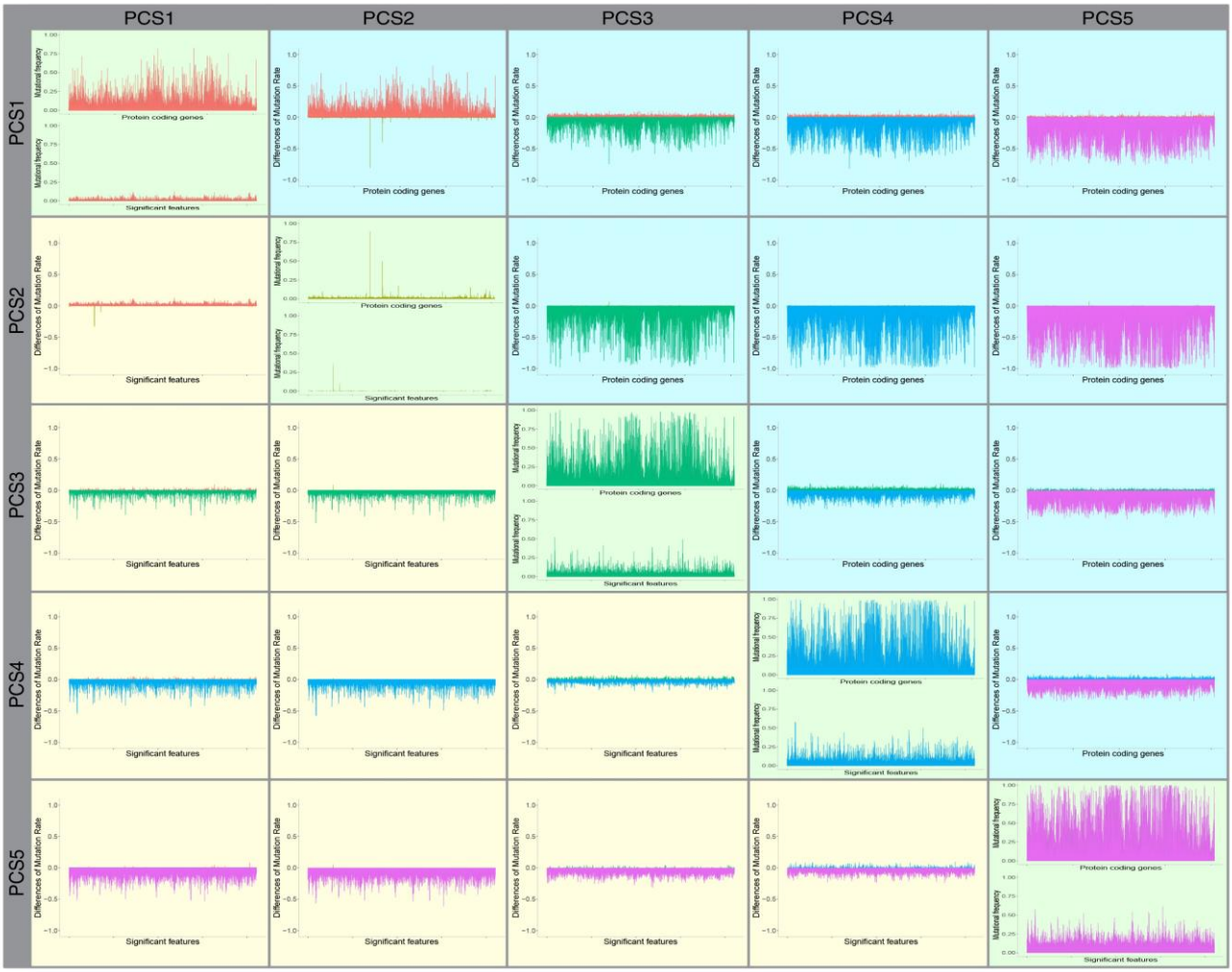


Figure 2. The mutation rate in subtypes and their differences. Bar plots in green tiles exhibit mutation frequency in protein-coding genes and significant features. Yellow tiles include bar plots of differences of mutation rate for significant features and blue tiles include differences in mutation rate in protein-coding genes. For example, the bar plot in the tile which is in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

PCS2 column and PCS2 row represents the difference of mutation rate in protein-coding genes in PCS2 and PCS1. The color of bars in differential bar plots represents the subtype with the higher mutation rate. For instance, if a bar in differential, the bar plot is red, and PCS1 has the higher mutation rate in that comparison (the same color as bars in the bar plot of mutation rate, in that subtype).

3.2. Relative frequency of mutations in the PC subtypes

We investigated the differences in the mutational rate of all protein-coding genes and also significantly mutated genes to determine the mutational specification of our identified subtypes, and explore if the PC subtypes are different concerning their mutational load. As shown in **Figure 2,** the level of mutational load in all coding genes and significantly mutated genes within the identified subtypes are different. This difference is much clear in subtypes PCS1, PCS2, and PCS3, while PCS4 and PCS5 have almost similar patterns. To investigate the mutational difference between the subtypes, we performed a differential mutation analysis (*see method section*) to explore the difference between the frequency of mutations in each gene and gene-motifs in each pair of subtypes. As shown in **Figure 2**, the difference between PC samples in subtypes PCS4 and PCS5 is more significant, indicating they are correctly grouped in two different subtypes. The figure also shows that the mutational frequency of the candidate gene-motifs in the subtypes are different, suggesting the different mutational mechanisms among the identified subtypes. We also conducted an independent sample t-test and determined when the difference between subtypes in the differential mutation analysis is more evident (*see method section*). We performed this analysis in three different scenarios by using all protein-coding genes, significant genes, and significant gene-motifs. Our results demonstrate that all subtypes are significantly different in terms of their mutational load in protein-coding genes and the significant gene-motifs (**Supplementary Table S12**).

3.3. Biological characterization of each subtype

We next investigate different aspects of the biological characterization of each subtype. Here, we search for unique genes, motifs, and transcripts in each subtype to investigate the differences of subtypes in each experiment. The 3-nucleotide motifs of each mutation also construct for further mutational signature analysis. Furthermore, pathways and GOs of associated genes will analyze to discover molecular and functional characteristics of each subtype. We also examine molecular data that was available for a subset of the pancreatic cancer samples. Lastly, we investigate the difference between survival cures of PC subtypes.

3.4. Motif rates and signatures analysis

There are 96 different types of mutations concerning their 3-nucleotide motifs in DNA. The occurrence rate of these motifs in subtypes is related to specific molecular mechanisms behind mutations in cancers [40]. To study the 3-nucleotide motifs rates in PC subtypes, the relative frequency of each 3-nucleotide motif among samples in each subtype has been calculated and plotted in **Supplementary Figure S3.**

In our gene association analysis, we identified several genes that were significantly mutated in multiple subtypes. Here, we explore if the mutations in these genes have different motif preferences in each subtype. To do this, we investigated the mutational load in 3-nucleotide motifs in the highly mutated genes that were associated to multiple subtypes. As shown in **Figure 3,** even though some genes are associated with more than one subtype, the mutations within these genes are enriched in different 3-nucleotide motifs. We have shown two oncogenes which were associated with more than one subtype, in **Figure 3**. As demonstrated in this figure, *PTPRD* has significantly different motifs when the rate of each motif in subtypes are being considered. CT-A.G, CT-C.G, and CT-G.G were occurred more frequently in PCS3, compared to PCS1. In *ROBO2* gene, four motifs (CG-T.A, CT-A.T, TA-A.T, and TA-T.A) were occurred more frequently in PCS4, compared to PCS3. We also found 21 different motifs in PCS4 and PCS5. A full list of genes that significantly mutated in multiple subtypes, but in different 3-nucleotide motifs, is provided in **Supplementary Table S4.**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

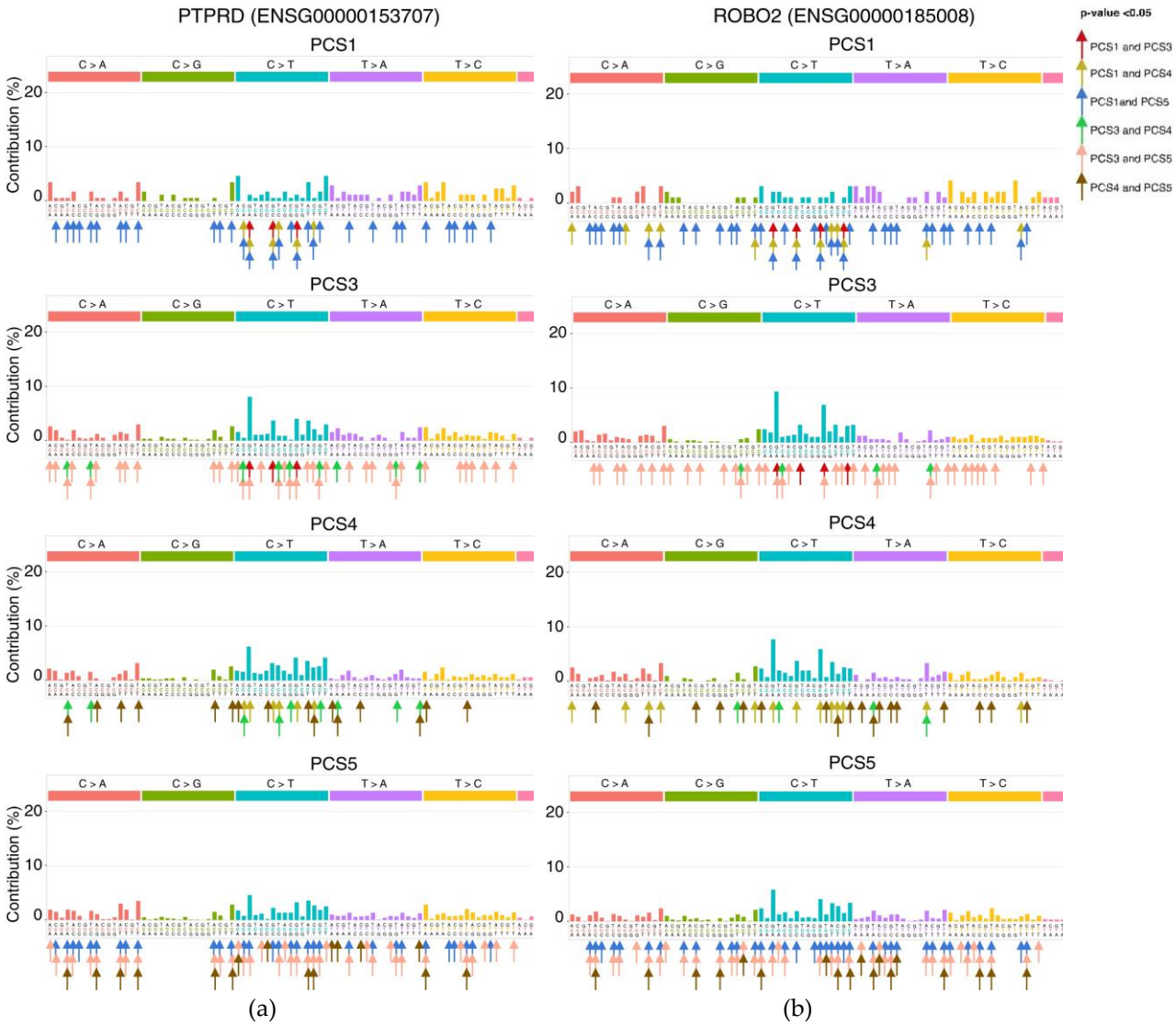
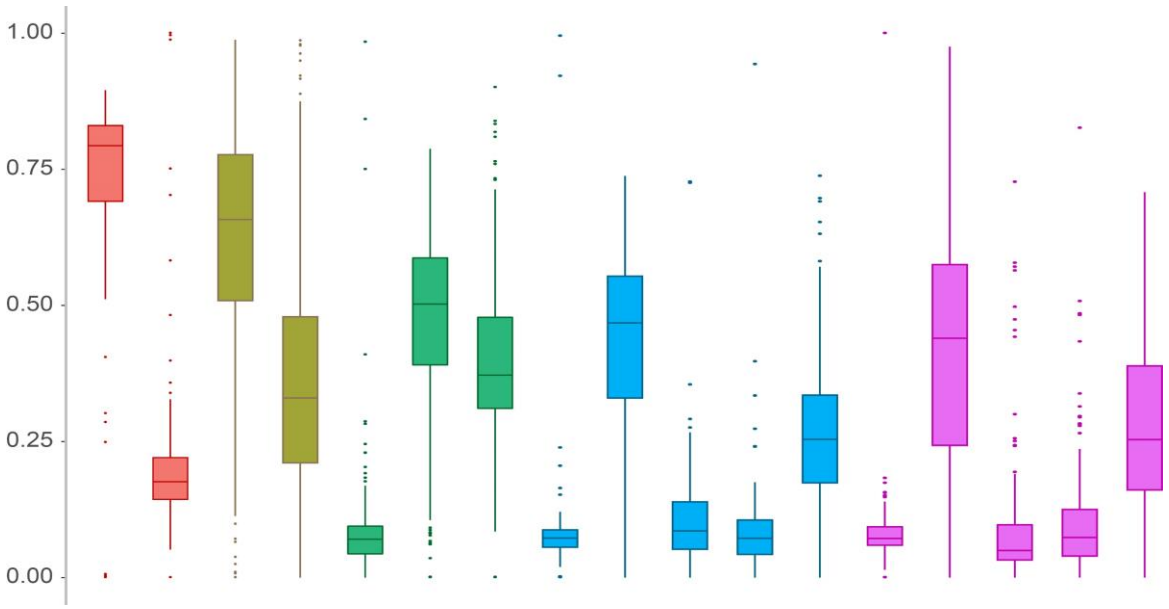


Figure 3. Significantly different motifs in common associated genes. 426 genes are associated with more than two subtypes. But they have mutated in different motifs. (a) PTPRD and (b) ROBO2 are two oncogenes that are good examples of this phenomenon. Although they are associated with 4 subtypes, as it is evident in their bar plot of motif rates, there are multiple differently mutated motifs when rates in subtypes are compared. Each arrow represents a significant difference in the rate of occurrence of the motif that is pointed to, and the color of the arrow indicates the comparison that motif was significant in. The p-values can be found in **Supplementary Table S4**.

We also investigated the mutational signatures in each PC subtype, separately [40]. Alexandrov *et al.* studied mutational signatures to find molecular mechanisms concerning the occurrence of each signature [41]. As it was discussed in [41], different signatures can be interpreted as different molecular mechanisms of mutations. Here, we used CANCERSIGN tool [24] to identify mutational signatures in the identified PC subtypes. Patterns of extracted mutational signatures are provided in **Supplementary Figure S7** (considering to Alexnadrov *et al*. signatures profile, we have excluded unknown and artifact signatures from our analysis). The importance and commonality of signatures in each subtype are shown in terms of boxplots of levels of exposures of samples in **Figure 4a**. We also calculated the angular similarity between identified signatures in each subtype and the signatures reported by Alexandrov *et al.*[17, 41]. 12 signatures in our study had Angular similarity more than 70% with Alexandrov*'s signatures*. SBS1, a spontaneous deamination of 5-methylcytosine was presented in all the subtypes (signature 3 of PCS1 with 72% similarity, signature 1 of PCS2 with 81% similarity, signature 2 of PCS3 with 79% similarity, signature 2 of PCS4 with 87% similarity, and signature 2 of PCS5 with 71% similarity). This signature is potentially associated with the most active mutational molecular mechanism in PC and is related to spontaneous or enzymatic deamination of DNA in which the failure in its detection causes fixation of T substitution for C, before the DNA replication **(Figure 4b)**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

(a)



(b)

	PCS1		PCS2		PCS3			PCS4					PCS5					
COSMIC Signatures	Signature 1	Signature 2	Signature 1	Signature 2	Signature 1	Signature 2	Signature 3	Signature 1	Signature 2	Signature 3	Signature 4	Signature 5	Signature 1	Signature 2	Signature 3	Signature 4	Signature 5	Mutational process
SBS1	0.04	0.72	0.81	0.07	0.03	0.79	0.10	0.02	0.79	0.11	0.17	0.15	0.01	0.71	0.10	0.15	0.06	Spontaneous deamination of 5-methylcytosine
SBS2	0.22	0.03	0.10	0.16	0.65	0.08	0.10	0.01	0.11	0.67	0.14	0.05	0.02	0.15	0.07	0.57	0.04	APOBEC activity
SBS3	0.65	0.17	0.13	0.67	0.22	0.17	0.68	0.20	0.17	0.25	0.36	0.74	0.21	0.21	0.22	0.23	0.81	HR deficiency
SBS4	0.34	0.11	0.08	0.41	0.10	0.12	0.39	0.44	0.09	0.12	0.14	0.36	0.45	0.11	0.07	0.10	0.38	Tobacco smoking
SBS5	0.75	0.29	0.24	0.63	0.26	0.28	0.74	0.16	0.29	0.33	0.43	0.67	0.16	0.36	0.23	0.28	0.60	Aging / Tobacco smoking / NER deficiency
SBS6	0.18	0.62	0.63	0.23	0.08	0.61	0.25	0.03	0.62	0.15	0.23	0.26	0.03	0.63	0.11	0.17	0.16	MMR deficiency
SBS7a	0.29	0.07	0.14	0.23	0.46	0.12	0.21	0.02	0.14	0.50	0.18	0.14	0.03	0.19	0.07	0.43	0.11	UV light exposure
SBS7b	0.34	0.09	0.14	0.30	0.20	0.13	0.33	0.03	0.15	0.26	0.18	0.26	0.03	0.20	0.07	0.21	0.19	UV light exposure
SBS7c	0.21	0.04	0.04	0.17	0.07	0.08	0.21	0.03	0.08	0.09	0.19	0.24	0.03	0.10	0.15	0.07	0.25	UV light exposure
SBS7d	0.23	0.08	0.04	0.21	0.06	0.08	0.22	0.03	0.08	0.09	0.15	0.20	0.03	0.11	0.09	0.07	0.20	UV light exposure
SBS8	0.40	0.10	0.10	0.38	0.09	0.14	0.45	0.35	0.12	0.13	0.19	0.46	0.36	0.15	0.11	0.09	0.48	HR deficiency / NER deficiency
SBS9	0.37	0.15	0.10	0.36	0.09	0.16	0.40	0.16	0.15	0.13	0.57	0.36	0.15	0.18	0.49	0.11	0.37	Polymerase eta somatic hypermutation
SBS10a	0.13	0.04	0.05	0.11	0.04	0.09	0.14	0.56	0.04	0.05	0.08	0.04	0.55	0.06	0.04	0.06	0.04	POLE exonuclease domain mutation
SBS10b	0.13	0.18	0.21	0.15	0.13	0.22	0.12	0.09	0.21	0.18	0.17	0.09	0.09	0.22	0.11	0.16	0.06	POLE exonuclease domain mutation
SBS10c	0.30	0.15	0.12	0.30	0.08	0.19	0.34	0.59	0.14	0.12	0.28	0.24	0.58	0.17	0.19	0.10	0.23	Defective POLD1 proofreading
SBS10d	0.18	0.04	0.06	0.15	0.07	0.10	0.19	0.68	0.05	0.07	0.11	0.08	0.68	0.06	0.05	0.07	0.08	Defective POLD1 proofreading
SBS11	0.34	0.09	0.12	0.32	0.19	0.11	0.34	0.02	0.14	0.25	0.19	0.25	0.03	0.19	0.07	0.20	0.20	Temozolomide chemotherapy / MMR deficiency + temozolomide
SBS13	0.16	0.01	0.03	0.22	0.35	0.03	0.08	0.10	0.03	0.25	0.04	0.20	0.10	0.02	0.01	0.39	0.24	APOBEC activity
SBS14	0.17	0.15	0.08	0.27	0.04	0.12	0.22	0.41	0.09	0.05	0.10	0.18	0.41	0.11	0.05	0.05	0.18	MMR deficiency + POLE mutation
SBS15	0.15	0.49	0.48	0.21	0.08	0.47	0.21	0.08	0.46	0.14	0.21	0.23	0.08	0.48	0.10	0.15	0.14	MMR deficiency
SBS17a	0.10	0.09	0.02	0.15	0.02	0.03	0.12	0.01	0.03	0.03	0.17	0.14	0.01	0.03	0.23	0.02	0.14	Damage by ROS
SBS17b	0.04	0.04	0.01	0.14	0.01	0.03	0.07	0.01	0.01	0.01	0.52	0.03	0.01	0.01	0.70	0.01	0.04	Damage by ROS / 5FU chemotherapy
SBS18	0.31	0.14	0.15	0.32	0.10	0.19	0.34	0.73	0.14	0.15	0.16	0.24	0.74	0.16	0.08	0.12	0.24	Damage by ROS
SBS20	0.21	0.21	0.10	0.36	0.05	0.13	0.27	0.16	0.13	0.07	0.15	0.27	0.17	0.16	0.07	0.06	0.25	MMR deficiency + POLD1 mutation
SBS21	0.19	0.17	0.05	0.23	0.03	0.08	0.22	0.03	0.08	0.05	0.15	0.20	0.03	0.11	0.08	0.03	0.20	MMR deficiency
SBS22	0.17	0.02	0.02	0.14	0.03	0.05	0.18	0.02	0.06	0.04	0.09	0.22	0.02	0.07	0.06	0.03	0.24	Aristolochic acid exposure
SBS24	0.28	0.14	0.13	0.34	0.08	0.15	0.34	0.32	0.13	0.11	0.16	0.28	0.33	0.15	0.08	0.09	0.29	Aflatoxin exposure
SBS26	0.32	0.24	0.07	0.33	0.05	0.12	0.33	0.04	0.12	0.08	0.26	0.32	0.03	0.15	0.18	0.06	0.31	MMR deficiency
SBS28	0.09	0.03	0.02	0.10	0.02	0.05	0.11	0.02	0.03	0.03	0.41	0.07	0.01	0.04	0.38	0.03	0.09	POLE exonuclease domain mutation
SBS29	0.26	0.16	0.17	0.32	0.08	0.19	0.32	0.46	0.16	0.11	0.13	0.26	0.47	0.17	0.06	0.09	0.27	Tobacco chewing
SBS30	0.42	0.12	0.17	0.35	0.33	0.16	0.40	0.03	0.20	0.37	0.25	0.29	0.04	0.25	0.09	0.32	0.22	BER deficiency
SBS31	0.39	0.10	0.12	0.34	0.11	0.12	0.42	0.08	0.14	0.15	0.19	0.41	0.09	0.17	0.10	0.12	0.32	Platinum chemotherapy
SBS32	0.37	0.17	0.20	0.32	0.18	0.21	0.40	0.07	0.23	0.25	0.22	0.31	0.08	0.27	0.09	0.20	0.27	Azathioprine exposure
SBS35	0.33	0.10	0.08	0.40	0.11	0.10	0.38	0.20	0.09	0.12	0.15	0.42	0.21	0.12	0.10	0.11	0.40	Platinum chemotherapy
SBS36	0.25	0.08	0.09	0.25	0.10	0.14	0.27	0.96	0.08	0.11	0.12	0.16	0.96	0.09	0.05	0.11	0.16	BER deficiency
SBS38	0.14	0.08	0.03	0.22	0.04	0.06	0.16	0.23	0.04	0.05	0.06	0.16	0.23	0.05	0.04	0.04	0.17	UV light exposure (indirect effect)
SBS42	0.38	0.15	0.11	0.43	0.10	0.13	0.43	0.17	0.13	0.14	0.22	0.35	0.18	0.19	0.08	0.11	0.32	Haloalkanes exposure
SBS44	0.34	0.31	0.19	0.43	0.07	0.24	0.42	0.12	0.24	0.11	0.26	0.38	0.12	0.29	0.12	0.09	0.33	MMR deficiency
SBS45	0.15	0.09	0.04	0.26	0.05	0.07	0.18	0.35	0.04	0.04	0.06	0.17	0.35	0.05	0.03	0.05	0.18	8-oxo-guanine introduced during sequencing
SBS54	0.18	0.22	0.07	0.25	0.03	0.09	0.20	0.04	0.09	0.05	0.14	0.19	0.03	0.11	0.09	0.03	0.19	Germline variants contamination
SBS84	0.31	0.12	0.10	0.32	0.08	0.11	0.36	0.03	0.13	0.12	0.18	0.31	0.04	0.17	0.06	0.09	0.28	AID activity
SBS85	0.36	0.07	0.05	0.25	0.09	0.12	0.34	0.03	0.13	0.14	0.24	0.35	0.03	0.17	0.14	0.10	0.34	AID activity
SBS87	0.14	0.49	0.54	0.17	0.06	0.52	0.18	0.05	0.53	0.13	0.18	0.21	0.05	0.50	0.11	0.16	0.15	Thiopurine chemotherapy
SBS88	0.34	0.08	0.08	0.17	0.04	0.13	0.31	0.05	0.13	0.08	0.29	0.31	0.04	0.16	0.18	0.05	0.29	Colibactin exposure
SBS90	0.12	0.02	0.02	0.09	0.02	0.08	0.13	0.01	0.08	0.04	0.12	0.17	0.01	0.10	0.09	0.02	0.17	Duocarmycin exposure
SBS92	0.64	0.22	0.17	0.53	0.24	0.21	0.62	0.17	0.22	0.28	0.31	0.60	0.17	0.28	0.14	0.26	0.52	Tobacco smoking

Figure 4. Signature Analysis. (a) Exposure of samples to signatures. Exposure of each sample to each signature indicates the engagement level of a sample. For example, samples of PCS5 are more exposed to signature 2 of this subtype. This indicates that the molecular mechanism associated with this signature has potentially more affected samples of this

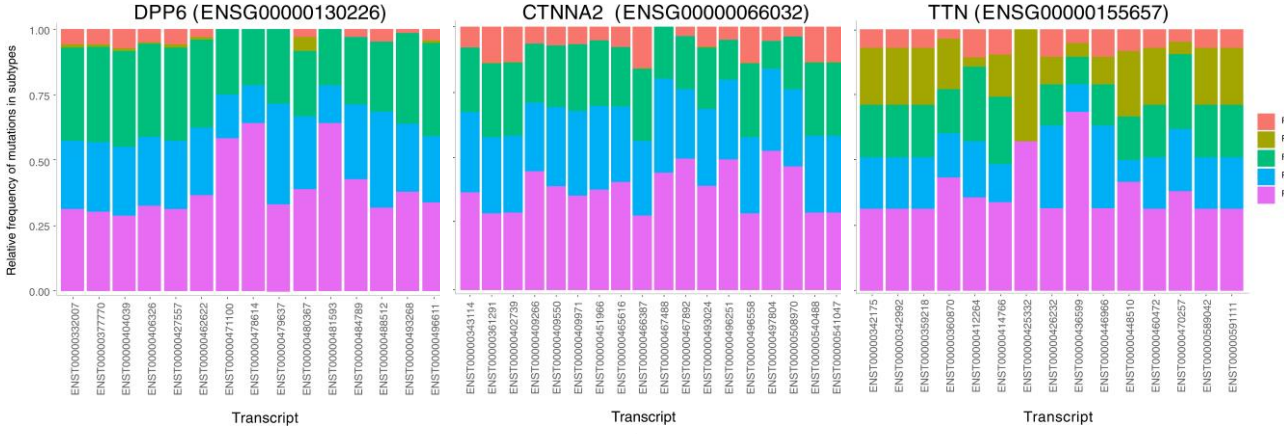
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

subtype. b. Comparing deciphered signatures to COSMIC signatures. This comparison can lead to revealing associated molecular mechanisms causing PC subtype signatures. Each cell of this heatmap indicates a level of similarity.

SBS3 is presented in PCS4 and PCS5 (with similarity rate 74% and 81% with PCS4 and PCS5, respectively). This is a defective homologous recombination-based DNA damage repair. SBS3 in Pancreatic cancer is related to responders to platinum therapy. Our clinical investigation for these two subtypes reveled that most of the patients in these subtypes were under platinum therapy. Our analysis also showed that SBS5 was presented in PCS1 and PCS3 with similarity rate more 75% and 74% to PCS1 and PCS3, respectively. This signature is associated to tobacco smoking. Interestingly, we found genes *PDE4D* and HECW1 are the highly mutated genes in PCS1 and PCS3, respectively. Mutation in these genes are known to be associated with smoking behavior[42, 43]. SBS17b is only presented in PCS5 (with similarity rate 70%). This signature is possibly associated to fluorouracil (5FU) chemotherapy treatment. Interestingly, we found out that at least 29% of patients in this subtype were under chemotherapy treatment. SBS18 and SBS36 are other Alexandrov's signature that are highly associated with subtypes PCS4 and PCS5, suggesting these two subtypes are also under pressure of DNA damage due to reactive oxygen species or somatic *MUTYH* mutations.

3.5. The mutational rate in transcripts

Mutations in genes can affect their transcripts and consequently their corresponding proteins based on their respective transcripts. To investigate the effect of mutations concerning transcripts in pancreatic cancer subtypes, we calculated the difference between our identified subtypes concerning the mutational load in different transcripts of the coding genes. Our analyses showed that for many of the candidate protein-coding genes, the mutations occurred in specific transcripts of the genes. To this end, the somatic point mutations were enriched in different transcripts of the genes. For instance, although *CTNNA2* has a 100% mutation rate in both PCS3 and PCS5, their mutational patterns were different from their transcripts (**Figure 5**). In PCS3, 52% of samples mutated in transcript ENST00000493024, while 88% of samples in PCS5 have mutated in the same transcript. In the *TTN* gene, PCS2 samples had more mutation compared to the *CTNNA2* gene. Interestingly, in various transcripts of this gene, the rate of mutation in subtypes are different. Another interesting point is that in transcript ENST00000425332 only PCS5 and PCS2 had mutations, while the other three subtypes had no mutation at all. In 4 out of 15 transcripts of the *DPP6* gene, only PCS3, PCS4 and PCS5 had the mutation and the other two subtypes had none. This analysis revealed the importance of *De Novo* somatic mutational load in transcripts, which can be used in a better understanding of the underlying mechanisms.



(a) (b) (c)

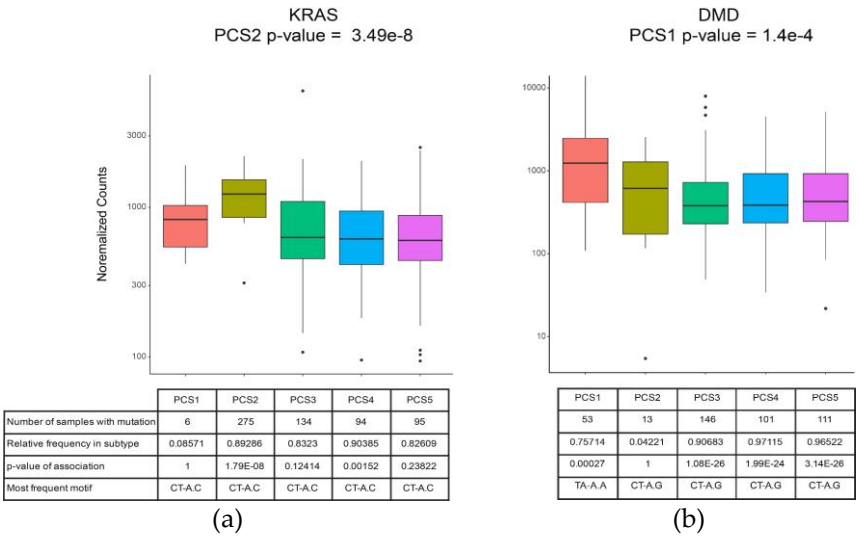
Figure 5. Rate of Mutated Transcripts in Subtypes. Some subtypes tend to mutate more in some transcripts of a gene. This can lead to different outcomes in subtypes.

3.6. Gene expression analysis and finding differentially expressed genes

To study the impact of *De Novo* mutations in the PCS subtypes on the expression level of coding genes, RNA-seq data of 307 samples were analyzed. This analysis was conducted by using the DEseq2 package [28] and its defined workflow (*see methods section*). The Gene expression level of each gene in PC subtypes was compared to the level of gene expression in other subtypes. As a result, we identified 303 uniquely differentially expressed genes (UDEG) in PCS1, 2,427 UDEG genes in PCS2, 267 UDEG genes in PCS3, 136, and 940 UDEG genes in PCS4, and PCS5, respectively. For example, *KRAS* was differentially expressed in PCS2 only; interestingly, it is the only gene that was significantly mutated to this subtype. Another example is *DMD* gene in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

PCS1. This gene has a tumor suppression activity, and alterations in the expression of this gene in pancreatic tumors have been discussed in [44]. Boxplot of expression of these two genes in the five subtypes is shown in **Figure 6**. A list of UDEGs for each subtype is provided in **Supplementary Table S13**.



	PCS1	PCS2	PCS3	PCS4	PCS5
Number of samples with mutation	6	275	134	94	95
Relative frequency in subtype	0.08571	0.89286	0.8323	0.90385	0.82609
p-value of association	1	1.79E-08	0.12414	0.00152	0.23822
Most frequent motif	CT-A.C	CT-A.C	CT-A.C	CT-A.C	CT-A.C

(a)

PCS1	PCS2	PCS3	PCS4	PCS5
53	13	146	101	111
0.75714	0.04221	0.90683	0.97115	0.96522
0.00027	1	1.08E-26	1.99E-24	3.14E-26
TA-A.A	CT-A.G	CT-A.G	CT-A.G	CT-A.G

(b)

Figure 6. Expression boxplots, mutation, and motif. Expression levels of (a) KRAS and (b) DMD in 5 subtypes are illustrated in form of boxplots. Some information about mutations in the genome is also provided in tables under each boxplot to represent the potential association of mutation (and their types) on expression levels.

3.7. Gene ontology and pathway analyses

We then performed gene ontology (GO) and gene pathway analyses to investigate whether candidate genes in PCS subtypes are significantly associated with any specific term [45]. We employed gene set enrichment analysis for all associated genes (see methods section) for each subtype. There are some common GO terms in two or more subtypes and some unique terms for each subtype as listed in **Supplementary Table S14**. For example, "regulation of protein kinase C signaling (GO:0090036)" is only associated with PCS3, "DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest (GO:0006977)" and "G1 DNA damage checkpoint (GO:0044783)" are unique to PCS4. On the contrary, the "epidermal growth factor receptor signaling pathway (GO:0007173)" is an example of common gene ontologies enriched in subtypes PCS2 and PSC5. Some terms like "regulation of cell proliferation (GO:0042127)" are common in all subtypes. Interestingly, a large number of GO terms across subtypes are related to the nervous system and axons, and there is a known relation between pancreatic nerve alterations and pancreatic cancer. These alterations in size and density of nerves have also some effects on non-neural pancreatic cells [46]. A table of all GO terms related to genes of each subtype is provided in **Supplementary Table S14**.

Our pathway analysis also revealed known pathways that are related to pancreatic cancer. For example, the cell adhesion molecules (CAMs) pathway that is enriched in PCS3, PCS4, and PCS5, is a well-known pathway in PC development [47-49]. The Axon guidance pathway was also enriched in all subtypes. Mutations and other genomic variations were seen in genes of this pathway for PC samples, and have been shown to have a critical role in the PC mechanism [50]. Pathways related to associated genes of each subtype are provided in **Supplementary Table S15**.

3.8. Clinical report and survival analysis of the identified subtypes

To study and understand the characteristics of each pancreatic cancer subtype, we also examined clinical data and phenotypic information including age, gender, and location of the samples. Besides, we studied the impact of tumor information including stage and grade on the survival of patients of each subtype, to gain more insight on their effect on patient's health status. To this end, the frequency of samples for each project in each subtype was counted to investigate if any meaningful relationships between this information and subtypes can be observed. The frequencies of gender and project are provided in **Supplementary Table S6** and **Supplementary Table S7** which illustrate distinguishable patterns among our identified subtypes.

As demonstrated in **Supplementary Tables S6** and **S7,** about 70% of all endocrine samples are in PCS1. Pancreatic cancer endocrine neoplasm is a rare type of pancreatic cancer that occurs in less than 1 per 100,000 persons per year in the general population [51]. Other remaining ductal samples (9 out of 70 samples) in this subtype may have similar molecular functionalities to those endocrine samples. Also, men are twice as likely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

as women in this subtype. For PCS2, all samples (except one sample) are of ductal adenocarcinoma type. Pancreatic ductal adenocarcinoma is the most common and more lethal type of pancreatic cancer that is more resistant to drugs and existing treatments [52]. For this subtype, about 75% of samples were gathered from Australia, meaning that this subtype is more likely to happen in this country. Besides, PCS2 is more common among men (about 58% are men and 48% are women). The majority of samples in PCS3 have a ductal adenocarcinoma type, and one-tenth of this subtype has an endocrine tumor. It is also observed that PCS4 and PCS5 have approximately the same proportion of ductal and endocrine types. This may indicate that these two subtypes are based on a common functionality between ductal and exocrine types. Moreover, 60% of samples of PCS4 are women, and 66% of samples in PCS5 are men.

We next conducted a survival analysis to estimate the overall survival of each subtype. Kaplan-Meier curves of each subtype are shown in **Figure 7**. The overall survival of PCS1 could not be estimated by its median due to their large amount of censorship. PCS1 mostly contains samples with an endocrine type that is the least lethal type of pancreatic cancer. Hence, donors with this type of cancer may skip follow-ups due to their good health conditions which in turn cause censorship. However, the mean overall survival time of the endocrine is 12.5 years.

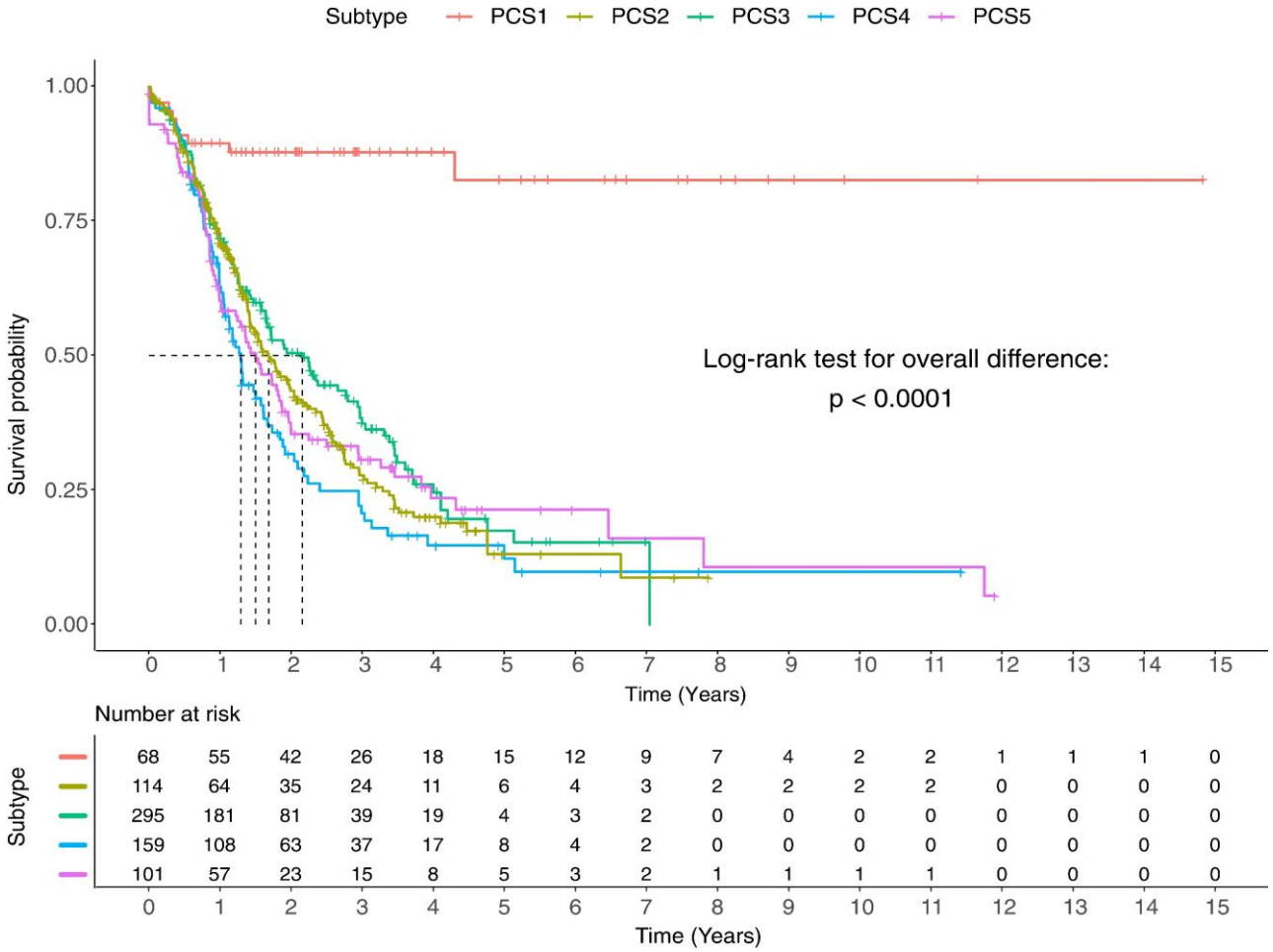


Figure 7. Survival Curves. Survival curves of the five subtypes reveal that PCS1 has the longest survival time, and PCS4 has the shortest. For detailed values see **Supplementary Table S8,** and for pairwise comparison, survival curves see **Supplementary Table S9**.

Among other subtypes, PCS3 has the longest overall survival time (25.8 months) which is followed by PCS2 (20.2 months), PCS5 (18 months), and PCS4 (15.5 months) (**Supplementary Table S8**). The log-rank test also shows that the overall survivals of all subtypes are statistically different (P-value of 0.0001). We also made pairwise comparisons between subtypes with multiple testing methods (*see methods section*). Among all subtypes, PCS1 has a significantly different survival curve from others, based on all the testing methods. The survival curve of PCS4 is different from PCS1, PCS2, and PCS3 while indicating some similarities to PCS5. PCS2, PCS3, and PCS5 have similar survival curves (**Supplementary Table S9**).

Many factors affect the survival of a pancreatic cancer patient. Among them, several clinical and histopathological information can help to improve the assessment of a patient's health condition and survival time. We

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

employed the Cox proportional hazard models to predict the survival time of subtypes by incorporating information such as age, gender, tumor grade, tumor stage, and our subtype indicators [15]. Their power of improving prediction accuracy was also measured by likelihood ratio tests. Here, we first conducted the Cox proportional hazard model with each variable, separately. These single-variable models show the power of improvement of survival prediction only based on the variables in the model (**Supplementary Table S10**). Our results demonstrate that subtype indicators are highly significant in predicting survival (P-value < 2.2e-16). All other variables also performed well in the single variable models (P-value < 1e-4). There is only one category in the stage variable (stage III), and one category in the grade variable (Undifferentiated) that had non-significant effects in their respective models (P-value > 0.05).

We then tested the performance of all variables in a complete model (**Supplementary Table S10**). The effects of each variable in a full model are tested by removing them one by one from the complete model and comparing the accuracy of the reduced models with the full model. All variables except age (P-value = 0.088) are shown to have significant effect in prediction accuracy of the complete model (P-value < 1e-4) (**Supplementary Table S11**). The results demonstrate that all of these properties (e.g., staging, grading, etc.) are important in the survival time prediction.

4. Discussion

Despite the large numbers of whole genome and exome sequencing data that are available for cancers, in particular Pancreatic cancer, there are still some ambiguities on the vast number of mutations, their types, and causes, leading to significant challenges in identifying mutational subtypes in Pancreatic cancer. However, different mutation rates of samples may shed some light on different molecular mechanisms behind mutations among the groups of patients. Mutation is the hallmark of cancer genome, and many studies have reported cancer subtyping based on the type of frequently mutated driver genes [14, 40, 53], or the proportion of mutational processes [40], however, none of these existing cancer subtyping methods consider these features simultaneously. In other words, the sequence context of somatic point mutations in driver genes have not been taken into consideration in cancer subtyping and biomarker discovery. Here, we integrated these two features (frequently mutated genes and sequence context of mutated sites) and implemented a bioinformatics pipeline for Pancreatic cancer subtyping using 774 pancreatic cancer samples from ICGC consortia. We found 4,211 significantly mutated gene-motifs in the pancreatic cancer samples and used them as the features for clustering, resulting in 5 subtypes among the pancreatic cancer samples (PCS1 to PCS5). PCS1 is potentially the subtype known as *ADEX* that consists of many samples with the endocrine neoplasm type of PC. ADEX tumours are shown to be involved in the upregulation of genes that regulate networks involved in *KRAS* activation, exocrine (*NR5A2* and *RBPJL*), and endocrine differentiation (*NEUROD1* and *NKX2-2*) [9]. *PTPRD* (Protein Tyrosine Phosphatase, Receptor Type D) has been mutated in 81.4% of the patients in this subtype (57 samples out of 70). It is shown that *the PTPRD* gene is a tumor suppressor and mutation/downregulation in this gene was observed in multiple cancers including lung and glioblastoma multiforme (a fatal form of brain cancer) [54]. *PTPRD* acts as a regulator for *STAT3*, which is shown to be activated in colon tumors and cell lines. Mutations in PTPRD restrict its ability to regulate STAT3, which promotes cancer progression [55]. PCS2 which is the largest subtype with 308 samples, is a *KRAS* addicted subtype as *KRAS* was mutated in more than 89.28% of samples in this subtype. Previous subtyping studies demonstrated that the *RAS* family are highly mutated genes in the lung, colorectal and pancreatic cancers [56]. Moreover, pancreatic ductal adenocarcinoma (PDAC) was reported as the most RAS-addicted among all cancers, which impacts cell differentiation, proliferation, and apoptosis [56]. The clinical trials using small molecule inhibitors targeting KRAS, resulted in promising anti-tumour effect for KRAS-mediated subtypes in pancreatic cancer [57].

The other 3 subtypes have a higher rate of mutation compared to PCS1 and PCS2, we there can called these three subtypes as hypermutated subtype. PCS3 samples were highly mutated in *SLIT2* and *ROBO1*. Bailey *et al*. in the previous Pancreatic cancer studies demonstrated that ROBO/SLIT signaling pathway play contradictory and anti-angiogenic roles in tumorigenesis, endometriosis and renal ischemia-reperfusion injury [9, 58-61]. Therefore, the ROBO/SLIT signaling pathway might be a promising target in pancreatic cancer therapy. *TP53* was the highly mutated gene in PCS4. Previous PC subtyping by Bailey *et al*. demonstrated that Squamous tumours are enriched for *TP53* mutations which interacts with ASCOM complex constituents *MLL2* and *MLL3*, and upregulation of the *TP63ΔN* transcriptional network [9]. As like as PCS3, PCS5's samples are also highly mutated in many genes including ROBO1 and SLIT2 demonstrating the contribution of ROBO/SLIT signaling pathway in tumorigenesis of PCS5's samples. However, mutations in PSC5's samples were also enriched in ROBO2, which is a stroma suppressor gene in the pancreas and acts via TGF-β signaling [62], which may suggest that both ROBO/SLIT and TGF-β signaling pathways play in tumorigenesis of PCS5. Previous

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

study on pancreatitis and PDAC mouse models showed that Robo2 can act as a stroma suppressor gene by restraining myofibroblast activation and T-cell infiltration [62].

Our pathway analysis also revealed cell cycle and The Axon guidance pathways the most common pathways in all PC subtypes identified in this study. Interestingly, The Axon guidance pathway was previously observed in murine Sleeping Beauty transposon-mediated somatic mutagenesis models of pancreatic cancer. In addition to common pathways, some subtype-specific pathways were also seen. For example, we identified Protein kinase C signaling pathway, EGFR (epidermal growth factor receptor) signaling pathway and p53 signaling pathway and as potential targets for treatment of the PSC1, PSC2, and PSC4 subtypes, respectively. It is worth mentioning that targeted treatment options are available for some of the pathways observed in our subtypes. For example, Cell adhesion molecules (CAMs) are glycoproteins expressed on the surface of cell membranes that act as oncogenes or tumor suppressors in signal transduction; they also act as tumor progression and metastasis regulators [47-49]. We identified CAMs as potential therapeutic targets in PCS3, PCS4, and PCS5 subtypes. By considering somatic mutations in all of the genes associated with our PC subtype-specific mutation of signaling pathways, we might be able to find additional cancer patients who could benefit from targeted treatment options. On the one hand, the pathways identified in our analysis are mutated in a large number of PC patients, and on the other hand, targeted treatment options are currently available for most of these pathways. We therefore believe that these treatment options are worthy for further investigation to develop better therapeutic targets.

Our analysis also revealed subtype-specific mutated genes which may be the main cause of functionality among each subtype. Although there are some genes that significantly mutated in multiple subtypes however, these genes are mutated in different motifs, indicating the context of the mutation are different in these genes in each subtype. This is also true for non-associated genes. For example, about 30% of samples in PCS5 had TTG-to-TGT mutations in *LRP1b* gene while PCS1 and PCS2 had no mutation in this gene-motif, and only 7% of samples in the other two subtypes had mutation in this gene-motif. *PTPRD* in another example that significantly mutated in PCS4 and PCS5 subtypes, however, the mutations were enriched in different motifs in each subtype (**Figure 3**). This would suggest that, rather than only investigating mutation in well-known oncogenes, we should consider the context of the mutations within driver genes (frequently mutated genes) to accurately identify cancer subtypes as well as targeted treatment biomarkers. By identifying subtype-specific gene-motif profiles, in addition to subtype-specific targeted therapeutics, we may obtain a clearer picture of the molecular mechanisms that cause a high rate of mutations (and consequently a high number of associated genes), in subtypes.

Our mutational signature analysis in the identified subtypes also revealed some new and subtype-specific signatures in addition to the well-known COSMIC signatures in the identified subtypes in this study; these signatures may be utilized to find the molecular mechanisms that are responsible in these subtypes (**Figures 4 and S2**). These molecular mechanisms systematically make changes across the genome, and hence they can leave a trace of their activity that corresponds to a different rate of motifs. We also found some signatures that are common among all subtypes, but with different exposures. Although the etiology of many COSMIC signatures is still unknown, some of them contain critical information. For instance, signature 1 of PCS4 and PCS5 are similar to SBS10a of COSMIC, and it is known that samples with this signature are hypermutator. The driver of some signatures such as signature 6 of PCS4 (similar to SBS31 of COSMIC) is chemotherapy with platinum drugs, and the driver of those that are similar to SBS3 and SBS6 of COSMIC, are DNA repair mechanism deficiency. The combination of these molecular mechanisms and their effect becomes dominant and drives cancer to different subtypes.

It is now well known that molecular mechanisms underlying the mutational processes can cause mutation across genome, blindly, because of their shape and structure. But different rates of gene-motifs may point to different accessibility of molecular mechanisms to genome in different genes. This can possibly be a result of different epigenetic factors in genes. However, this couldn't be investigated in this study due to lack of epigenetic data but can be a lead for future works.

With genomic medicine emerging as a routine part of the health system, tumors mutational profiling will help to better understanding of the underlying genetic causes of cancers. The current treatments options are usually based on assessing single gene mutations. Our study and the proposed pipeline to identify PC subtype associated genes and the context of the mutations within these genes (either by identifying gene-motifs or mutation signatures) could help more precise diagnoses by assigning patients to available therapeutic targets or ongoing clinical trials targeting specific mutations; and identifying subtype-specific pathways that might be useful treatment targets for therapeutic intervention.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2021 doi:10.20944/preprints202108.0368.v1

The gene expression analysis also revealed genes that are differentially expressed in the subtypes. This may be due to the centrality of associated genes or the genes they affect in the pathways or regulatory (co-expression) network. In other words, expression levels of some of our identified subtypes are only driven by mutations, while some others such as PCS2 and PCS5, are only influenced by mutations besides other factors. To verify this claim, we extracted downstream neighbors of associated genes in pathways of each subtype, up to 4 levels. We discovered that 16, 65, 27, 19, and 166 of UDEGs of PCS1 to PCS5 are among the neighbor genes, respectively (**Supplementary Table S16**). Interestingly, "Pathways in cancer" that been observed for PCS5 has contains 30 PCS5 associated genes 14 UDEGs in PCS5 (see the number of associated genes and UDEGs in each pathway in **Supplementary Table S17**). The "RAS signaling pathway" in PCS2 has also the largest number of UDEGs (equal to 20). Interestingly, *KRAS* gene was the only associated gene to PCS2 and has probably a strong effect on the expression alteration.

Our investigations of clinical information, available for a subset of the patients, revealed an association between the survival time of PC patients and histopathological factors such as grading and staging. For example, PCS1 has the longest survival time, and its curve is differentiated compared to the other subtypes (**Figure 7**). This is because most PCS1 samples were from the endocrine type of PC that has lower lethality. More investigations on the centers that have collected the samples demonstrate that the PCS2 samples mainly came from Australia, and the PCS5 samples from Canada (60%) (**Supplementary Table S6**). There is a possibility that some molecular mechanisms associated with the mutational signatures are influenced or driven by ethnicity or geographical variables. There were also some biases towards genders in some subtypes (**Supplementary Table S7**), in which 60% of samples in PCS1 are male, and about 60% of samples in PCS4 are female.

5. Conclusions

High-throughput sequencing has provided many improvements in finding the key mutations and molecular events by providing a high number of samples. This will lead to accurate classification of patients based on their mutational profiles, and consequently, and better clinical decisions on their treatment. In this manuscript, we provided a list of subtype-specific gene-motifs which can be useful in better understanding the underlying genetic causes of pancreatic cancer, by exploiting the context of the mutations in the driver genes. Considering the genes with significant mutations rate in PC, and the contexts of the mutations in the genes can provide a more effective and personalized treatment for pancreatic cancer. We showed that our proposed pipeline helps discovering mutational patterns associated with cancer related pathways, clinical phenotypes, and potential therapeutic target options for cancer-specific subtypes, as well as mutational patterns that are observed across multiple Pancreatic cancer types. Our proposed model and its related codes are publicly available online at: https://github.com/bcb-sut/Pancreatic-Cancer-Subtype-Identification.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Supplementary Figure S1: Scatterplot of samples in first two Principal Component dimensions. Supplementary Figure S2: Evaluation plots. Evaluation plots for deciding the number of mutational signatures. Supplementary Figure S3: Motif rate - main subtypes. Rate of occurrence of 96 types of 3-mer motifs in 5 subtypes of PC. Supplementary Figure S4: Motif rate - outlier subtypes. Rate of occurrence of 96 types of 3-mer motifs in 9 outlier clusters. Supplementary Figure S5: Venn diagram of associated genes. Venn diagram of associated genes of PC subtypes (For more information on gene association study see the Methods section). Supplementary Figure S6: Venn diagram of DEGs. Venn diagram of differentially expressed genes of each subtype. Expression levels of each subtype are compared to all other four subtypes, and DEGs are inferred. Genes that are only in one subtype are considered as UDEGs (uniquely Differentially Expressed Genes) of that subtype. Supplementary Figure S7: Signatures of pancreatic cancer subtypes. Signatures of PC subtypes extracted by the CANCERSIGN tool. These are prevalent patterns of 3-mer motif of mutations among samples of each subtype. Supplementary Table S1: Significant genes. Supplementary Table S2: Significant gene-motifs. Supplementary Table S3. Significant features. Supplementary Table S4: Significantly different motifs in common associated genes. Supplementary Table S5: All final associated genes. Supplementary Table S6: Project frequencies. Supplementary Table S7: Gender frequency. Supplementary Table S8: Overall survival time estimation. Supplementary Table S9: Survival pairwise comparison. Supplementary Table S10: Survival cox regression. Supplementary Table S11: Survival cox regression likelihood ratio test. Supplementary Table S12: T-test results. Supplementary Table S13: Unique differentially expressed genes. Supplementary Table S14: Gene ontologies (GO). Supplementary Table S15: pathways analysis. Supplementary Table S16: Number of UDEGs among up to fourth-order downstream neighbors of associated genes in pathways. Supplementary Table S17: The number of associated genes and UDEGs in their neighborhood in pathways.

Author Contributions: HAR designed the study; AG, HAR, AM, AD, AB, and HRR wrote and edited the manuscript. AM, AM, HD carried out the analyses including the statistical analyses, candidate gene, gene-motif identification, text mining, gene prioritization, and gene ontology, mutational signature analysis (with help from MB). HRR helped with the statistical

analyses. AG and AM generated all figures and tables. All authors have read and approved the final version of the manuscript.

Funding: HAR was supported by the UNSW Scientia Fellowship Program and the UNSW Graduate School of Biomedical Engineering. HRR and AG were supported by the IR National Science Foundation (INSF) Grant No. 96006077.

Data Availability Statement: The source code and a sample dataset are available as a supplementary file.

Conflicts of Interest: The authors declare no competing financial and non-financial interests.

References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA: a cancer journal for clinicians, 2019. **69**(1): p. 7-34.
2. Rahib, L., et al., *Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States.* Cancer research, 2014. **74**(11): p. 2913-2921.
3. Collisson, E.A., et al., *Molecular subtypes of pancreatic cancer.* Nature Reviews Gastroenterology & Hepatology, 2019: p. 1.
4. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* New England Journal of Medicine, 2001. **344**(11): p. 783-792.
5. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib.* New England Journal of Medicine, 2004. **350**(21): p. 2129-2139.
6. Garcea, G., et al., *Molecular prognostic markers in pancreatic cancer: a systematic review.* European journal of cancer, 2005. **41**(15): p. 2213-2236.
7. Collisson, E.A., et al., *Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy.* Nature medicine, 2011. **17**(4): p. 500.
8. Moffitt, R.A., et al., *Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma.* Nature genetics, 2015. **47**(10): p. 1168.
9. Bailey, P., et al., *Genomic analyses identify molecular subtypes of pancreatic cancer.* Nature, 2016. **531**(7592): p. 47.
10. Sivakumar, S., et al., *Master regulators of oncogenic KRAS response in pancreatic cancer: an integrative network biology analysis.* PLoS medicine, 2017. **14**(1): p. e1002223.
11. Puleo, F., et al., *Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features.* Gastroenterology, 2018. **155**(6): p. 1999-2013. e3.
12. Androulakis, I., E. Yang, and R. Almon, *Analysis of time-series gene expression data: methods, challenges, and opportunities.* Annu. Rev. Biomed. Eng., 2007. **9**: p. 205-228.
13. Tate, J.G., et al., *COSMIC: the catalogue of somatic mutations in cancer.* Nucleic acids research, 2018. **47**(D1): p. D941-D947.
14. Kuijjer, M.L., et al., *Cancer subtype identification using somatic mutation data.* British journal of cancer, 2018. **118**(11): p. 1492.
15. Kuipers, J., et al., *Mutational interactions define novel cancer subgroups.* Nature communications, 2018. **9**(1): p. 4353.
16. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495.
17. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. **578**(7793): p. 94-101.
18. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS computational biology, 2013. **9**(8): p. e1003118.
19. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants.* Bioinformatics, 2015. **31**(22): p. 3673-3675.
20. Dashti, H., et al., *Integrative analysis of mutated genes and mutational processes reveals seven colorectal cancer subtypes.* bioRxiv, 2020.
21. Scrucca, L., et al., *mclust 5: clustering, classification and density estimation using Gaussian finite mixture models.* The R journal, 2016. **8**(1): p. 289.

22. Fraley, C. and A.E. Raftery, *Model-based methods of classification: using the mclust software in chemometrics.* Journal of Statistical Software, 2007. **18**(6): p. 1-13.

23. Fraley, C. and A.E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis.* The computer journal, 1998. **41**(8): p. 578-588.

24. Bayati, M., et al., *CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes.* bioRxiv, 2019: p. 424960.

25. Alexandrov, L.B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... & ICGC MMML-Seq Consortium, *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-421.

26. Alinejad-Rokny, H., et al., *Source of CpG depletion in the HIV-1 genome.* Molecular Biology and Evolution, 2016. **33**(12): p. 3205-3212.

27. Ebrahimi, D., M.P. Davenport, *Insights into the motif preference of APOBEC3 enzymes.* PLoS One, 2014. **9**(1): p. e87679.

28. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.

29. Javanmard, R., JeddiSaravi, K., *Proposed a new method for rules extraction using artificial neural network and artificial immune system in cancer diagnosis.* Journal of Bionanoscience, 2013. **7**(6): p. 665-672.

30. Rad, M.P., Pourshaikh, R., *Conceptual Information Retrieval in Cross-Language Searches.* Research Journal of Applied Sciences, Engineering and Technology, 2012. **4**(12): p. 1714-1720.

31. Parvin, H., Parvin, S., *Divide and conquer classification.* Australian Journal of Basic and Applied Sciences, 2011. **5**(12): p. 2446-2452.

32. Hasanzadeh, E.,et al., *Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm.* International Journal of Physical Sciences, 2012. **7**(1): p. 16.

33. Esmaeili, L., Minaei-Bidgoli, B., Nasiri, M., *Hybrid recommender system for joining virtual communities.* Research Journal of Applied Sciences, Engineering and Technology, 2012. **4**(5): p. 500-509.

34. Parvin, H., et al. *A Novel Classifier Ensemble Method Based on Class Weightening in Huge Dataset*. 2011. Berlin, Heidelberg: Springer Berlin Heidelberg.

35. Alinejad-Rokny, H., Anwar, F., Waters, S. A., Davenport, M. P., & Ebrahimi, D., *Source of CpG depletion in the HIV-1 genome.* Molecular biology and evolution, 2016. **33**(12): p. 3205-3212.

36. Parvin, H., MirnabiBaboli, M., *Proposing a classifier ensemble framework based on classifier selection and decision tree.* Engineering Applications of Artificial Intelligence, 2015. **37**: p. 34-42.

37. Woolson, R.F., *Rank tests and a one-sample logrank test for comparing observed survival data to a standard population.* Biometrics, 1981: p. 687-696.

38. Karadeniz, P.G. and I. Ercan, *Examining tests for comparing survival curves with right censored data.* Stat Transit, 2017. **18**(2): p. 311-28.

39. Zhu, Y., et al., *Statistical methods for identifying sequence motifs affecting point mutations.* Genetics, 2017. **205**(2): p. 843-856.

40. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415.

41. Alexandrov, L., et al., *The repertoire of mutational signatures in human cancer.* BioRxiv, 2018: p. 322859.

42. Zuo, H., et al., *Cigarette smoke up‐regulates PDE3 and PDE4 to decrease cAMP in airway cells.* British journal of pharmacology, 2018. **175**(14): p. 2988-3006.

43. Park, S.L., et al., *Mercapturic acids derived from the toxicants acrolein and crotonaldehyde in the urine of cigarette smokers from five ethnic groups with differing risks for lung cancer.* PloS one, 2015. **10**(6): p. e0124841.

44. Luce, L.N., et al., *Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations.* Oncotarget, 2017. **8**(1): p. 145.

45. Consortium, G.O., *Expansion of the Gene Ontology knowledgebase and resources.* Nucleic acids research, 2016. **45**(D1): p. D331-D338.

46. Demir, I.E., H. Friess, and G.O. Ceyhan, *Neural plasticity in pancreatitis and pancreatic cancer.* Nature reviews Gastroenterology & hepatology, 2015. **12**(11): p. 649.

47. Moh, M.C. and S. Shen, *The roles of cell adhesion molecules in tumor suppression and cell migration: a new paradox.* Cell adhesion & migration, 2009. **3**(4): p. 334-336.

48. Bassagañas, S., et al., *Pancreatic cancer cell glycosylation regulates cell adhesion and invasion through the modulation of α2β1 integrin and E-cadherin function.* PloS one, 2014. **9**(5): p. e98595.

49. Farahani, E., et al., *Cell adhesion molecules and their relation to (cancer) cell stemness.* Carcinogenesis, 2014. **35**(4): p. 747-759.

50. Biankin, A.V., et al., *Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes.* Nature, 2012. **491**(7424): p. 399.

51. Halfdanarson, T.R., et al., *Pancreatic endocrine neoplasms: epidemiology and prognosis of pancreatic endocrine tumors.* Endocrine-related cancer, 2008. **15**(2): p. 409.

52. Ilic, M. and I. Ilic, *Epidemiology of pancreatic cancer.* World journal of gastroenterology, 2016. **22**(44): p. 9694.

53. Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., Lander, E.S., Van Allen, E.M. and Sunyaev, S.R., *Identification of cancer driver genes based on nucleotide context.* Nature Genetics, 2020. **52**(2): p. 208-218.

54. Veeriah, S., et al., *The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers.* Proceedings of the National Academy of Sciences, 2009. **106**(23): p. 9435-9440.

55. Funato, K., et al., *Tyrosine phosphatase PTPRD suppresses colon cancer cell migration in coordination with CD44.* Experimental and therapeutic medicine, 2011. **2**(3): p. 457-463.

56. Waters, A.M. and C.J. Der, *KRAS: the critical driver and therapeutic target for pancreatic cancer.* Cold Spring Harbor perspectives in medicine, 2018. **8**(9): p. a031435.

57. Canon, J., et al., *The clinical KRAS (G12C) inhibitor AMG 510 drives anti-tumour immunity.* Nature, 2019. **575**(7781): p. 217-223.

58. Guo, S.-W., et al., *Slit2 overexpression results in increased microvessel density and lesion size in mice with induced endometriosis.* Reproductive Sciences, 2013. **20**(3): p. 285-298.

59. Rama, N., et al., *Slit2 signaling through Robo1 and Robo2 is required for retinal neovascularization.* Nature medicine, 2015. **21**(5): p. 483-491.

60. Li, S., et al., *Slit2 promotes angiogenic activity via the Robo1-VEGFR2-ERK1/2 pathway in both in vivo and in vitro studies.* Investigative ophthalmology & visual science, 2015. **56**(9): p. 5210-5217.

61. Chaturvedi, S., et al., *Slit2 prevents neutrophil recruitment and renal ischemia-reperfusion injury.* Journal of the American Society of Nephrology, 2013. **24**(8): p. 1274-1287.

62. Pinho, A.V., et al., *ROBO2 is a stroma suppressor gene in the pancreas and acts via TGF-β signalling.* Nature communications, 2018. **9**(1): p. 1-14.