Article

Pan-cancer integrative analysis of whole-genome *De novo* somatic point mutations reveals 17 cancer types

Amirreza Kazemi[1,2], Amin Ghareyazi[1], Kmia Hamidh[1], Hamed Dashti[1], Maedeh Sadat Tahaei[1], Hamid R. Rabiee[1 *], Hamid Alinejad-Rokny [3,4,5], Abdollah Dehzangi[6, 7, *]

[1] Bioinformatics and Computational Biology Lab, Department of Computer Engineering, Sharif University of Technology, Tehran, 11365, IR
[2] Department of Computer Engineering, Simon Fraser University, Burnaby, BC, 1S6, CA
[3] BioMedical Machine Learning Lab (BML), The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, AU
[4] UNSW Data Science Hub, The University of New South Wales (UNSW Sydney), Sydney, NSW, 2052, AU
[5] AI-enabled Processes (AIP) Research Centre, Macquarie University, Sydney, 2109, AU
[6] Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[7] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA

* Correspondence: rabiee@sharif.edu and i.dehzangi@rutgers.edu; Tel: +1 (856) 225-6699

Abstract: The advent of high throughput sequencing has enabled researchers to systematically evaluate the genetic variations in cancer, resulting in identifying many cancer-associated genes. Although cancers in the same tissue are widely categorized in the same group, they demonstrate many differences concerning their mutational profiles. Hence there is no "silver bullet" for the treatment of a cancer type. This reveals the importance of developing a pipeline to identify cancer-associated genes accurately and re-classify patients with similar mutational profiles. Classification of cancer patients with similar mutational profiles may help discover subtypes of cancer patients who might benefit from specific treatment types. In this study, we propose a new machine learning pipeline to identify protein-coding genes mutated in a significant portion of samples to identify cancer subtypes. We applied our pipeline to 12270 samples collected from the International Cancer Genome Consortium (ICGC), covering 19 cancer types. Here we identified 17 different cancer subtypes. Comprehensive phenotypic and genotypic analysis indicates distinguishable properties, including unique cancer-related signaling pathways, in which, for most of them, targeted treatment options are currently available. This new subtyping approach offers a novel opportunity for cancer drug development based on the mutational profile of patients. We also comprehensive study the causes of mutations among samples in each subtype by mining the mutational signatures, which provides important insight into their active molecular mechanisms. Some of the pathways we identified in most subtypes, including the cell cycle and the Axon guidance pathways, are frequently observed in cancer disease. Interestingly, we also identified several mutated genes and different rates of mutation in multiple cancer subtypes. In addition, our study on "gene-motif" suggests the importance of considering both the context of the mutations and mutational processes in identifying cancer-associated genes. The source codes for our proposed clustering pipeline and analysis are publicly available at: https://github.com/bcb-sut/Pan-Cancer.

Keywords: Pan-Cancer; somatic point mutations; cancer subtyping; biomarker discovery; driver genes; personalized medicine; health data analytics.

1. Introduction

Cancer is a heterogeneous disease characterized by the progression of molecular changes that can develop in different tissues. Many tumors within a tissue have different molecular mechanisms. Moreover, some tumors across multiple tissues appeared to have similar biological mechanisms (1-4). Different histology, mutation profiles, or ex-

pression profiles can distinguish tumors into several subtypes, enabling us to classify patients into subgroups with similar clinical characteristics or medical diagnoses better than cancer types. Pan-cancer classification is a relatively new approach to understanding the origin and cause of all cancer types. Although cancer types have many differences, we consider them as a single disease. Hence, by subtyping this disease, we can have a better insight into its origins and causes. New studies in this field provide us with promising findings (2). As a result of recent progressive experiments on the large-scale genome, cancer subtype identification has been performed in multiple cancer types based on expression data (5, 6), copy number (7), DNA methylation data (8, 9), or integration of different omics type (10, 11). For instance, (12) employed three different similarity kernels on three types of profile data (gene expression, miRNA expression, and isoform expression data) for five cancer types from TCGA and then aggregated computed similarities by using the Similarity Kernel Fusion (SKF) for tumor subtyping. In (10), the authors used a hierarchically stacked autoencoder (called HI-SAE) on the gene expression and transcriptome alternative splicing profiles to learn new data representations. Then, based on the newly learned data representations, they classified breast cancer patients from TCGA.

Transcriptional profiling of samples has multiple issues, including the effect of invasive sampling and its impact on expression profiles and noise in collected data. In contrast, mutational profiles are more robust to these problems (13). However, a limited number of studies have tried to perform identification based on the somatic point mutations instead of the expression data. Somatic mutation is closely related to cancer due to its essential role in cancer progression. Since mutational processes or genes involved can be linked to different molecular mechanisms driving tumor progression and clustering tumors, this data type can be very informative and compelling. However, existing sparseness (many samples have only a limited number of mutations) and heterogeneity (two tumors rarely share the same mutations) in mutation data bring new challenges. Some studies have addressed the sparseness issue by using gene interaction networks as prior knowledge. For instance, (14) applied an algorithm called NetNorM on raw somatic mutation data and constructed more amenable data by employing Pathway Commons (a dataset containing gene network information). They removed non-essential mutations for high-mutated tumors to creating normalized data and added missing mutations for less-mutated tumors. (15) also used different gene interaction networks to construct network smoothed mutational data by propagating driver mutated gene into its neighborhood in the genes network. This approach may identify sub-networks around a highly connected or mutated gene while other genes in that sub-network have not mutated significantly. Other studies, such as (16) proposed modifying heats (score) to genes in the network to reduce the diffusion of genes like *TP53.*

However, using a gene interaction network as prior knowledge to de-sparsify mutational profiles may be irrelevant for some cancer types. Moreover, it ignores possible indirect interactions between genes that are not captured in gene networks. The study in (17) resolved the data sparseness challenge by developing a de-sparsification method that summarizes somatic mutations in genes into pathway-level mutation scores. Then, they used the binomial distance to cluster pathway mutation scores. Although this method helps identify the mutational patterns associated with clinical phenotypes, they just focused on the previously cancer-associated genes (18, 19) to find pathway scores. As a result, this method is not the most suitable approach for cancer subtyping because already known genes may not fit the best model for mutational profiles. In other words, it does not consider the essential unknown genes that might play a significant role in developing cancer. This study addresses this issue by finding the best-fitted distributions for each cancer type's mutational profiles, which enabled us to identify the significantly mutated genes in each cancer by defining a threshold. We believe that our approach can identify biologically important genes beyond the set of previously cancer-associated genes for more accurate subtyping. Also, the pan-cancer study of the heterogeneity problem of mutational profiles can be improved because more significant num-

 doi:10.20944/preprints202111.0266.v1

bers of tumors are under investigation, and mutational subtypes amongst cancer types can be identified.

To the best of our knowledge, mutational processes do not have the same effect on genes. This has never been adequately explored for cancer subtype identification in previous studies. By studying mutation rate among samples and mutational signatures in subtypes, we demonstrate that mutational processes do not have the same effect in different cancer types. While, in our new classes of cancers, we show that this effect is homogenous among samples. This provides better insight for researchers and clinicians to understand the origin of a patient's cancer and develop new treatments. In this study, based on the idea of pan-cancer and the advantages and helpful insights of somatic mutations, we studied mutational profiles available from International Cancer Genome Consortium (ICGC), which contains tumors from 19 cancer types. We then made use of hg19 annotation to annotate mutations in each gene of this dataset. We explored a wide range of statistical distributions for each cancer type to model mutational profiles and identify significantly mutated genes for each cancer type. Then, we performed a hierarchical clustering model on somatic mutations in these genes by aggregating identified candidate genes of each cancer type. Our clustering approach is based on the Gaussian Mixture Model (GMM), which outperforms other techniques such as K-means. This method chooses the best number of clusters by evaluating various metrics. We started by performing model-based clustering on all tumors, and two significant sub-groups were divided. We iteratively repeated this process on each sub-group until they reached our defined threshold, and at the end, we identified 17 subtypes. We further provide a comprehensive analysis including mutational load, gene association, mutational signature, gene ontology, pathway enrichment, and survival analysis for each subtype. These experiments help us to indicate that different distinguishable molecular mechanisms exist in each identified subtype. The source codes for our proposed clustering pipeline and analysis are publicly available at: https://github.com/bcb-sut/Pan-Cancer.

2. Materials and Methods

After cleaning the data, we performed a distribution-based analysis of genes and samples in which mutations occurred in this study. We fitted the distribution for each cancer type and identified which genes are significantly mutated. We clustered all samples in all 19 cancer types and determined 17 cancer subtypes. Next, we comprehensively studied phenotypic and genotypic characteristics of each subtype to investigate differences and commonalities among different cancer subtypes. This includes: "Gene and Gene-motif association as a biomarker of each subtype", "Mutational load of genes for each subtype", "Mutational signature analysis", "Gene ontology and pathway analysis", and "Clinical report and survival analysis". Throughout this paper, the "cancer type" term indicates traditional cancer types, which were identified by the tissue of origin and histopathology-based classification. While the "Cancer subtype" term is used to indicate our newly proposed classes of cancers. In the following sections, we discuss our experiments and analyze the results.

ICGC Dataset

We used the International Cancer Genome Consortium (ICGC) dataset, which contains 19 types of cancers. In this study, we focus on somatic point mutations. We combined available data of each cancer type and then built the somatic mutation profile of 12270 samples, in which 48.5% are female and 51.5% male. To determine which genes were mutated, we used the FANTOMCAT robust gene list. As a result, we identified 20,345 protein-coding and 7,114 long non-coding regions mutated among all our samples. Then we annotated the genes with somatic mutation for all samples.

Statistical pipeline to identify significant genes

Machine learning and statistical approaches have been widely used to identify cancer biomarkers (20-23). In this study, we used the Cullen-Frey graph (24) to find the best-fitted distribution for each cancer. Here we focused only on coding genes and identified

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

significantly mutated ones in each cancer type. We first counted the number of samples that had a mutation in each gene and then used different distribution models to identify significantly mutated genes. Among different distributions, negative binomial demonstrated the best fit to our data. We have also experimentally investigated different distributions which among them, negative binomial distribution fitted the best to our data. We next used each cancer type's best-fitted distribution (Figure 1.b) to identify significant genes. We then calculated the *p*-value for each gene in all cancers using the following formula:

$$\mathbf{P(x > k) = 1 - \sum_{i=r}^{k} P(x = i) = 1 - \sum_{i=r}^{k} \binom{i-1}{r-1} p^r q^{k-r}} \quad (1)$$

This is the probability of samples having more than *k* mutations in a given gene, where *p* is the probability that a sample has a mutation in a given gene and *q* is the complementary probability of having a mutation in a gene (not having a mutation in a given gene or *1-p*).

Comparing the significance of obtained genes in each cancer is a challenging task. Still, if we select the mutated genes in a more significant portion of samples of each cancer type, we can get the genes primarily associated with cancer types. Therefore, genes located in the 0.001 right tail of the distribution (in other words, with a *p*-value less than 0.001) of each cancer type were selected to avoid unwanted redundancies. These 684 extracted genes are our features for the clustering step. For the rest of this paper, we refer to these genes as "Significant Genes".

Clustering method

We used model-based clustering to identify subtypes. Since we did not consider any assumption on several subtypes, we preferred a non-parametric method. Model-based clustering is one of the density-based and non-parametric unsupervised machine learning methods for clustering. Another reason to apply model-based clustering was due to sample independence and its number of mutations. Hence, we anticipated that candidate genes' mutational load follows Gaussian distribution due to the central limit theorem if subtypes are precisely identified. Mclust is an available package in R, which we used to apply model-based clustering. Mclust fits each cluster's best Gaussian Mixture models and utilizes the Bayesian Information Criterion (BIC) metric to find an optimal number of clusters (25, 26). Here, we hierarchically used Mclust with three levels of clustering. At first, we clustered all samples into two groups. After that, each group was clustered into their subgroup or subclasses. This process continued until no new group was found in identified clusters or the algorithm returns a big cluster with more than 95% samples of the parent cluster and the rest to some small residual clusters. As shown in Figure 2.a, 12270 samples are clustered into two clusters with 9318 and 2952 samples in each. After that, each of these two clusters was given to the clustering algorithms, and the results are illustrated in the second level of clustering of Figure 2.a. This hierarchical process was employed to make sure clusters and subtypes are homogenous, and no heterogeneity could be found among samples of each cluster.

Mutational load analysis

We performed mutational load analysis on protein-coding genes and the feature genes (candidate genes) for each subtype separately. Mutational load of gene 'g' in subtype 'C' is the number of samples in subtype C that mutated in gene g, divided by the total number of samples in subtype C.

Mutational signature analysis

A mutational signature is a fingerprint for a molecular mechanism that is causing mutation across the genome. Molecular mechanisms are blind to what location they are causing the mutation. Therefore, to identify the molecular mechanism of the mutational

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

signature, we have to consider all mutations in the whole genome (except mitochondria). Here we used CANCERSIGN to identify mutational signatures that are represented in our cancer samples (27). The process of finding mutational signatures involves a computational method named Non-negative Matrix Factorization (NMF). Since this method is an unsupervised machine learning method (just like clustering) and the number of molecular mechanisms (hence mutational signatures) that are active among input samples is unknown, we have to run this algorithm multiple times to test multiple possibilities. Each time we assume that the number of signatures in the samples is N. We then change N each time in the range of 2 to 15. After calculating all the possibilities, results are tested in the evaluation plots provided in Supplementary Figure 1. With the elbow rule, we can decide which N is more accurate and hence which N is optimal. The complete procedures for selecting the optimal number of clusters are provided in the CANCERSIGN tool paper [23].

Gene and gene-motif rates analyses

We used Fisher's exact test to identify coding genes that significantly mutated in each subtype. Fisher's exact test is done by computing a contingency table for each pair (gene, subtype). The contingency table consists of the number of samples in the subtype with a mutation in the gene, the number of samples in the subtype that had no mutation in the gene, the number of samples from other subtypes with a mutation in the gene, and the number of samples from different subtypes that had no mutation in the gene. The results are shown in Supplementary file S1. We performed the same analysis for gene-motif to identify significantly mutated gene-motifs in each subtype. The results for the top 100 significant coding genes and top 100 significant gene-motifs for each subtype are shown in Supplementary tables S4 and S5.

Consequence type of mutations

The consequence type of mutations is available in the ICGC dataset. For each mutation, there may be multiple consequence types. We counted the consequences of each subtype's significant genes and then calculated the frequency of consequence types for each subtype.

Gene ontology analysis and gene pathway analysis on the significantly mutated coding genes

We used the gene ontology analysis tool, enrichr (28), to observe the over-representation of gene ontology and pathways associated with each subtype's top 100 significant genes separately. We used default value for adjusted p-value in enrichr (FDR < 0.05,). Gene ontology covers three domains, namely, biological process, cellular component, and molecular function. The complete list of enriched gene ontology and pathway is provided in Supplementary Table S6.

Clinical information

We downloaded clinical data for samples from ICGC (http://cancer.digitalslidearchive.net). Metadata files containing information about donors and their respective samples have been used to analyze gender and region. For each sample, we used the clinical data of the donor that the sample belonged. For gender analysis, we found the gender for each donor. But for ethnicity analysis, we used the project-code feature in ICGC metadata and extracted the region part from it to find the region that the sample was sequenced

Survival analysis

Like the Clinical report section, after obtaining the clinical data, specifically survival data, we filtered the patients that all its samples belonged to a specific subtype. We used the Kaplan-Meier method to conduct survival curves for all subtypes. We used "survival"(29) and "survminer" (30) R packages to perform Kaplan-Meier curves and obtain the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

significance of survival prediction for subtypes. A Long-rank test was also applied to obtain the *p*-value for survival analysis.

3. Results and Discussion

We performed a background model to extract significant coding genes to be able to distinguish cancer patients. Single-base mutational profile of samples obtained from the International Cancer Genome Consortium (ICGC) dataset. We clustered 12,270 samples across 19 cancer types into new subtypes using model-based clustering by considering extracted genes. Finally, we performed comprehensive biological analyses on our identified subtypes to investigate each subtype's biological characterization and gain new insights into cancer subtyping.

Pre-analysis

This study focuses on somatic point mutations from the ICGC dataset, which contains 19 tissue cancers data. We used 12,270 cancer samples available in the dataset in which 48.5% are female, and 51.5% are male across different projects, including READ-US, COAD-US, COCA-CN, etc. We used the Ensemble gene annotation dataset (31) to identify somatic point mutations in coding genes. This dataset contains 20,345 protein-coding genes. We excluded all non-single-base mutations (e.g., insertions, deletions) from our analyses. To identify new subtypes, we only considered coding mutations.

Background model to identify candidate genes

To cluster samples based on their mutational profiles, we first need to identify candidate genes that significantly mutated in each cancer. To do this, we first determined for each cancer, best-fitted distribution to identify genes that significantly mutated in each cancer. We used Cullen and Frey's graph (*see method*) with 500-fold bootstrapping to find the best-fitted distribution (Figure 1.a). Although we examined various distributions, a negative binomial was the best-fitted distribution for all cancer types (Figure 1.b). Next, we used the Cramer-Von Mises test to confirm the perceived distributions. We considered each cancer type's perceived distribution to detect candidate genes and then calculated the mutational load's *p*-value for each gene separately. We then used a threshold of 0.001 on the *p*-value to determine candidate genes of each cancer type. We then gathered all candidate genes from all cancer types and identified 684 genes significantly mutated in at least one cancer. A complete list of candidates (features) genes and their *p*-value is provided (Supplementary Table S1). The mutational load of feature genes in each cancer type is shown in (Supplementary figure 2). According to this figure, some genes are significantly mutated in multiple cancers, which aligns with the idea presented by pan-cancer research. For instance, *TP53* and *KRAS* are examples of genes among the significant genes of many cancers such as Breast, Brain, and Ovarian. TP53 single-base substitution, known as the primary type of alterations, is associated with cellular proteins' inactivation and leads to many cancers (32). As the figure shows, pancreatic and prostate cancers are the most mutated cancers in the candidate genes (43.8% and 44.3% of the candidate genes are mutated in pancreatic and prostate cancers, respectively). Esophagus and nervous system cancers are the less mutated cancers in the candidate genes (only 1 and 2 genes out of 684 candidate genes are mutated in the esophagus and nervous system cancers, respectively).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

Figure 1.Best-ftted distribution for each cancer type

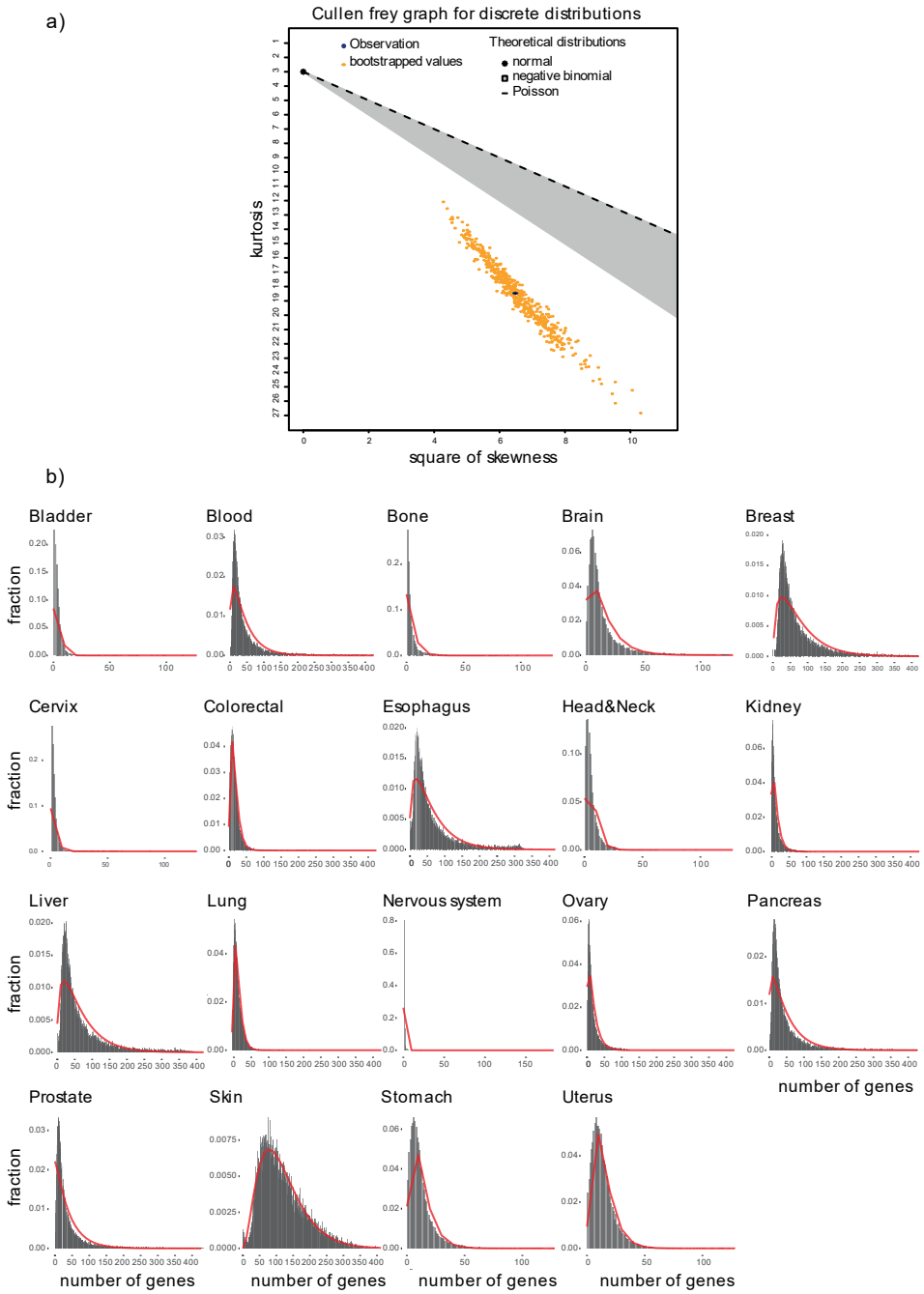


Figure 1. (a) Best-fitted distribution to discover feature genes. Cullen Frey method was applied to identify the best distribution fitting for mutational data of cancer types. The figure shows the Cullen-Frey graph for discrete distributions. (b) The distribution of mutational profiles in each cancer. For all cancer types, we fitted a negative binomial to their mutational data. Each plot shows empirical mutation data in a specific cancer type, and the red line shows a negative binomial distribution fitted to the cancer type. The X-axis indicates the number of mutations in each gene, and Y-axis shows the fraction of samples in the specific cancer type.

Model-based clustering to detect new subtypes

Having significantly mutated genes identified, we then used these genes as the features for our multi-level clustering approach to identifying cancer subtypes. For this purpose, we used the Mclust package implemented in R (see method section). We pre-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

ferred this method over other clustering approaches because it builds a robust model to identify clusters without random initialization. Unlike Mclust, classic clustering algorithms such as k-means need to be randomly initialized and sensitive to initialization. Besides, clustering methods such as dbscan (33) and hdbscan (34) require the user to optimally specify the optimal number of clusters. Hence, we applied the method based on Gaussian mixture models to overcome these issues and cluster our samples based on their inherent. As shown in (Figure 2.a) at the first level of clustering, the algorithm breaks down all samples into 2 clusters (Cluster 1 with 9318 and Cluster 2 with 2952 samples, respectively). Custer1 was divided into eight sub-clusters (from Cluster 1.1 to Cluster 1.8), and Cluster 2 was split into two sub-clusters (Cluster 2.1 and Cluster 2.2). Finally, at the third level, only Cluster 1.5 was divided into eight sub-clusters. We used a threshold of 95% to stop breaking down a cluster meaning that if we observed that applying the algorithm to a cluster would bring us a new sub-cluster with more than 95% of samples in, clustering is not valid. The cluster would be identified as a new subtype. We used this threshold because of some possible outliers in clusters. When a cluster was divided into multiple sub-clusters and a sub-cluster with more than 95% of the father's samples, we believed that all other sub-clusters could be outliers, so the cluster should not be divided. Eventually, we obtained 17 clusters as our new identified subtypes (17 subtypes from C1 to C17). Supplementary Table S2 shows all samples identified in each subtype.

We then investigated the contribution of samples from different cancers in our identified subtypes. Figure 2.b shows this contribution relatively in a bar plot, and Supplementary figure 3 shows the contribution specifically in a heat map. A table of the number of the contribution of samples from different cancers in our identified subtypes is also provided in Supplementary Table S3. As we can see in both figures, most of the subtypes consist of various cancer types. For example, subtype C7 and subtype C12 contain all cancer types, but subtype C4 and subtype C8 are mainly composed of Head & Neck Cancer with 78% and 82% of their samples. Also, many subtypes mostly contain samples of 2 to 4 cancer types. For instance, Subtype C2 mainly consists of Prostate (48.2%), Blood (17.5%), and Breast (16.8%) cancers which are about 82.5% of all samples in this subtype. Subtype C3 consists of Blood (68.1%) and Lung (24.5%) types which are about 92% of all samples in this subtype, and subtype C9 also contains samples from Bladder (53.3%), Kidney (26.7%), and Cervix (11.1%) types which are about 91% of samples. Moreover, many cancers are primarily scattered in 3 or 4 subtypes. For instance, more than 95% of the Prostate cancer samples are grouped in C1(19.8), C2 (27.9%), C7 (20.6%), and C16 (20%). The esophagus samples are primarily in C16 (26.7%) and C17(30.6%), and more than 75% of Ovary samples are in C16 (44.9%) and C12 (31.4%).

In the following sections, we present the analysis results to demonstrate the biological characterization of identified subtypes. We hence illustrate the effectiveness of our new subtyping approach over traditional cancer type classification approaches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

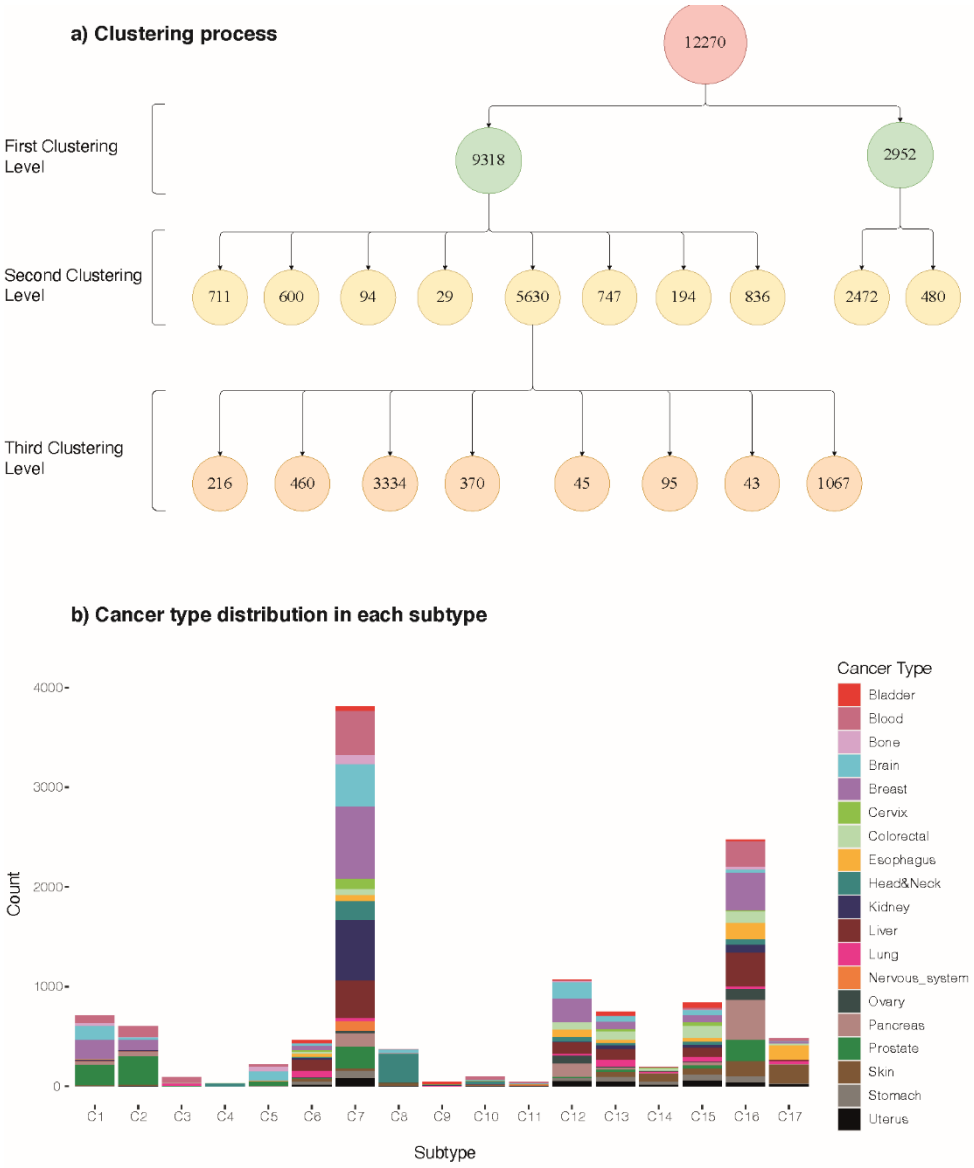


Figure 2. (a) The clustering tree shows the process performed by the Model-based method. In the first level of clustering, all 12,270 samples were divided into two sub-groups. In the second round of clustering, the first sub-group with 9,318 samples was divided into eight sub-groups (1.1 to 1.8), and the second sub-group was split into two new sub-groups (2.1 and 2.2). And finally, the third level of clustering sub-group 1.5 with 5,630 samples was divided into eight sub-groups (1.5.1 to 1.5.8). (b) Distribution of all samples in identified subtypes. Each color corresponds to a cancer type. The X-axis shows subtypes, and Y-axis indicates the number of samples. Subtype C7 is the most populated subtype and comprises many samples from all cancer types (Kidney and Breast are observed in the C7 subtype more than other cancer types). Subtype C16, the next most populated, comprises samples from all cancer types (Pancreas, Liver, and Breast cancers are observed in C16 more than other cancer types).

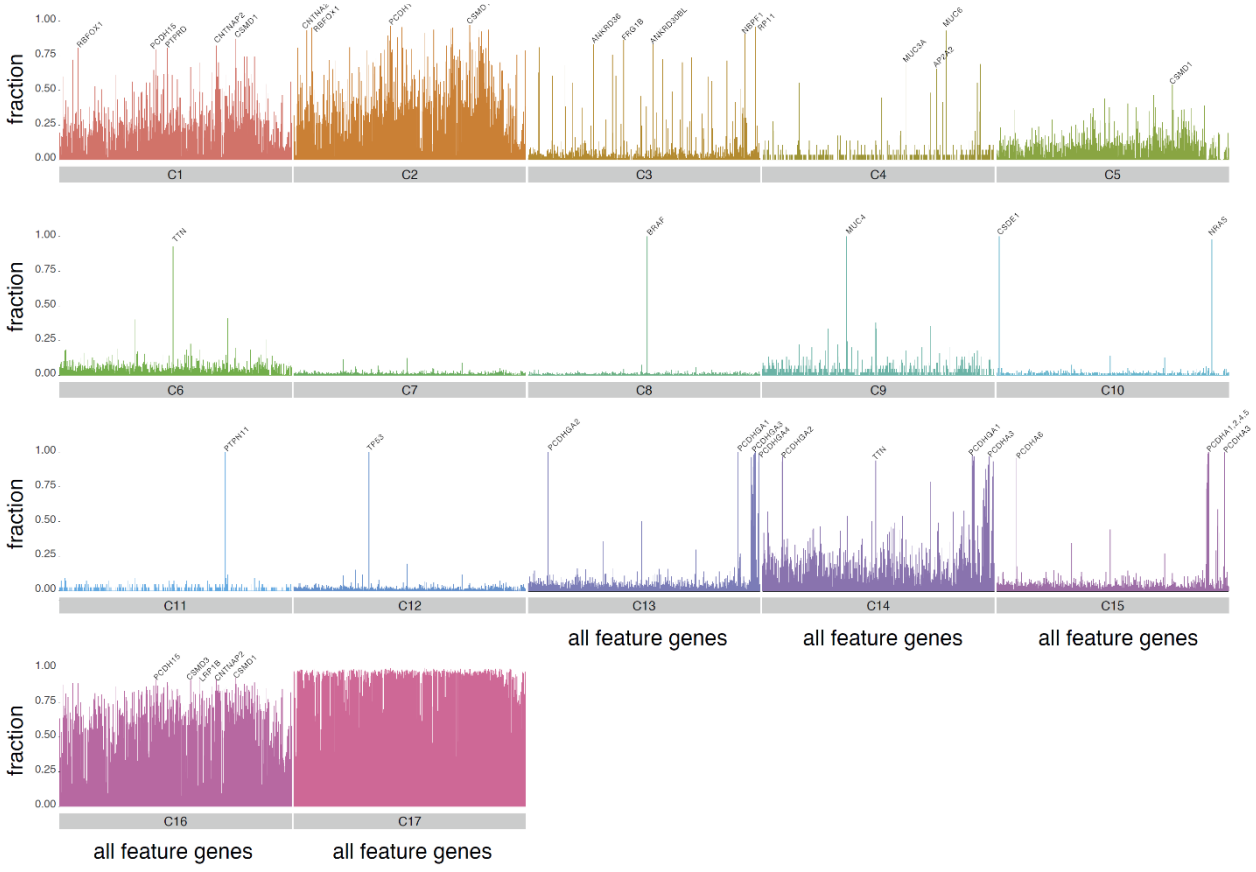
Mutational load of genes for each subtype

We study the mutational load of candidate genes and all protein-coding genes in our subtypes in this analysis. To compute the mutational load of gene 'g' in subtype 'C' we counted the number of samples in subtype 'C' which have a mutation in gene 'g' and then normalized it by dividing it into the number of all samples of subtype 'C'. As Figure 3 shows, the mutational load distribution of candidate genes in the identified subtypes is different, demonstrating correct pan-cancer clustering of the samples. As the figure shows (Figure 3), subtypes C16 and C17 are hyper-mutated subtypes (5 genes with at least one mutation in 90% of samples in C16 and 276 genes with at least one mutation in 95% of samples in C17) which can be a reason that samples of these two sub-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

types were separated from others at the first level of clustering. Notably, 92.6% of samples in C6 had a mutation in *TTN*; all samples in C8 had a mutation in *BRAF*; all C9 samples had a mutation in *MUC4*; 100% and 97.9 % of samples in C10 had a mutation in *CSDE1* and *NRAS*, respectively; all samples in C11 had a mutation in *PTPN11*; all samples in C12 had a mutation in *TP53*; all samples in C13 had a mutation in *PCDHGA1*, *PCDHGA2*, and *PCDHGA3*. Moreover, 98.6%, 96.2%, 99.1 and 98.1% of samples in this subtype had a mutation in *PCDHGA4*, *PCDHGA5*, *PCDHGB1*, and *PCDHGB2*, respectively (gamma Protocaderins family highly mutated in this subtype); 96.9% of samples in C14 had a mutation in PCDHGA2,3,4 and PCDHA1,2,3. Moreover, 97.4% of samples in this subtype had a mutation in *PCDHGA1*; 99.7% of samples in C15 had a mutation in PCDHA1,2,3. Moreover, 99.7%, 96.1%, and 98.8% of samples in this subtype had a mutation in *PCDHA4*, *PCDHA5*, *PCDHA6*, respectively (alpha Protocadherins family highly mutated in this subtype).

However, for some subtypes, we found similar patterns. For instance, *CSMD1* and *RBFOX1* are highly mutated in both subtypes C1 and C2 (*CSMD1* and *RBFOX1* are mutated in 80.5% and 87% of samples in C1, respectively, and 95.1% and 96.8% of samples in C2, respectively). Other examples are *PCDHGA1* and *PCDHGA2* that is mutated in almost all samples of C13 and C14. To understand the difference between similar subtypes (C1 and C2; C13 and C14), we plotted the fraction of samples that had at least three mutations in each candidate (feature) gene (Supplementary figure 4). As the figure shows, tumor samples in C2 and C14 have higher mutations than subtypes C1 and C13, respectively. In addition, we observed that *CSMD1* is mutated in 74% of the samples in C2. In comparison, only 42% of samples in C1 are mutated within this gene, meaning that the difference between C1 and C2 originated from the different mutation rates in significant common genes. Supplementary Figure 4 also shows *PCDHGA1* mutated in 6% and 43 % of C13 and C14 subtypes, respectively, demonstrating the effect of mutation numbers in distinguishing these two subtypes. Another example for C13 and C14 is *PCDHGA2* which mutated in 5% and 37% of C13 and C14. Results demonstrate that common genes have more mutations in C14 than C13.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

Figure 3. Each graph for each subtype illustrates the portion of samples in a subtype that has a mutation in each of the 684 genes. In other words, each bar indicates the number of samples that have a mutation in a gene among 684 genes, divided by the total number of samples in that subtype. The taller a bar, the more important that gene is for that subtype because more samples had a mutation in that specific gene.

Gene and Gene-motif association as a biomarker of each subtype

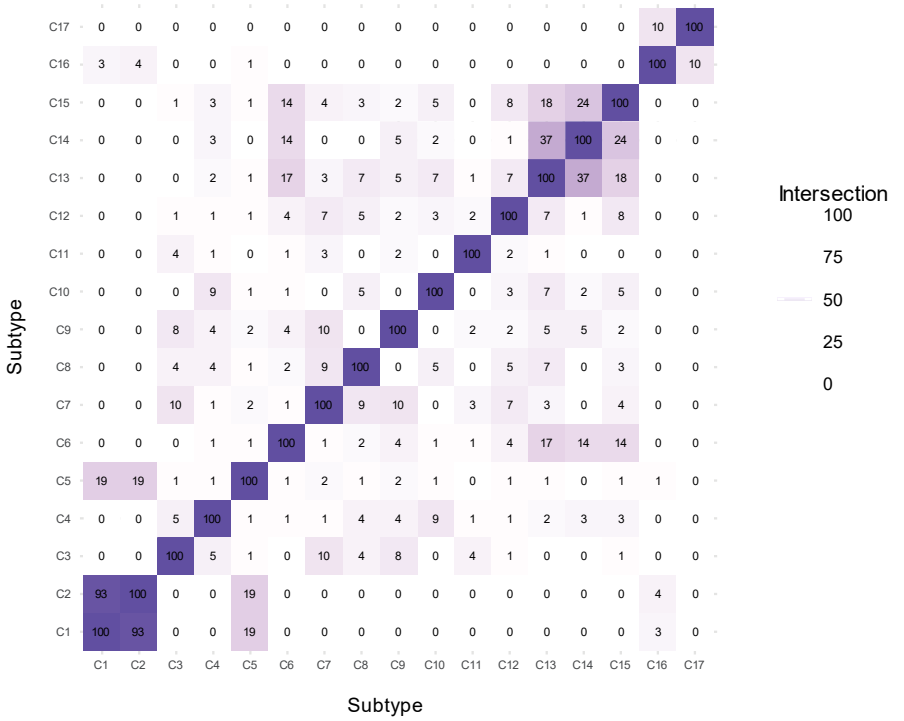
We then investigated the top 100 highly mutated genes in each subtype (Supplementary Table S4). We then asked how many of the top 100 highly mutated genes are common between every two subtypes (Figure 4.a). This figure shows that many pairs of subtypes have a few common genes, while some pair with numerous common genes. For example, subtypes C1 and C2 have 93 significant common genes out of 100 in both subtypes. While subtypes C13 and C14 have 34 common genes in their top 100 significant genes. Interestingly, there is no common gene between their top 100 important genes for many of the subtypes.

It has been recently shown in (35) that gene-motifs are the primary source of disease-related variations in cancer. Gene-motifs refer to the 3-nucleotide sequence mutated within a gene, i.e., NXN-to-NYN (where reference nucleotide X mutated to Y, and N: A, C, G, or T). There are 96 combinations of mutations within 3-nucleotide motifs. For example, *MUC16, LRRC4C,* and *IL1RAPL1* are examples of genes that appeared as significant genes within the top 100 important genes of different subtypes. We investigated each subtype's mutations in tri-nucleotide motifs to show the motif preferences of mutations in each of these genes (Supplementary figure 5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

Figure 4

a)Gene intersection between each two subtypes



b)Gene-motif intersection between each two subtypes

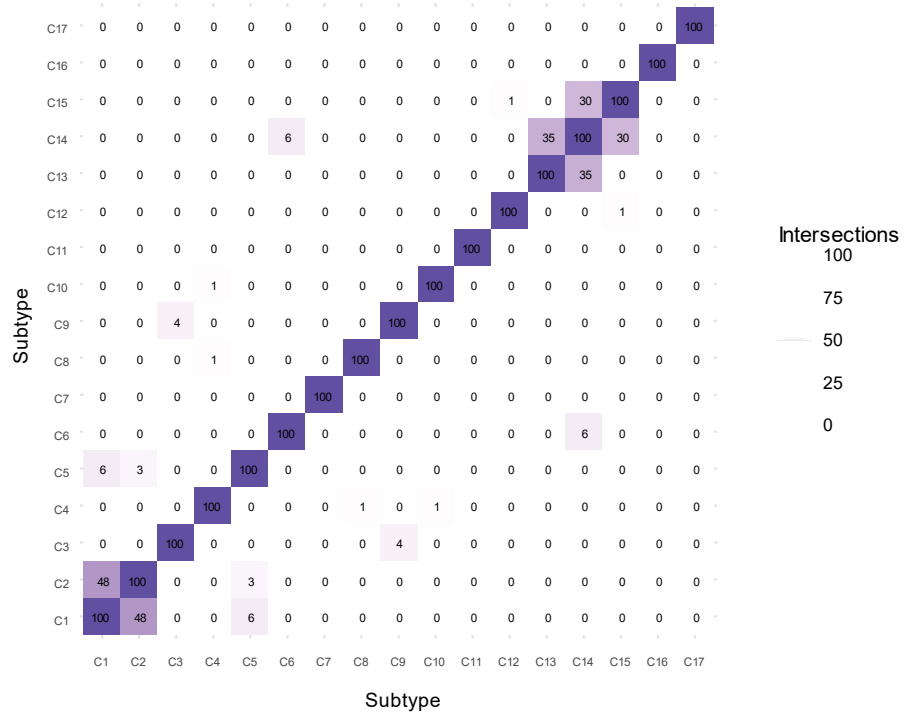


Figure 4. (a) Each cell corresponds to the number of genes in the top 100 significant genes among two subtypes. Many subtypes have very few genes in common with others except C1 and C2 that have 93 significant common genes out of 100, or C13 and C14 that have 37 significant common genes out of 100, or C13 and C15 that have 24 significant common genes of 100. (b) Each cell corresponds to the number of gene-motifs in the top100 significant gene-motifs of every two subtypes. Compared to Figure 12, interestingly, almost all subtypes have a fewer gene-motif in common than genes. For example, in subtypes C1 and C2, which have 93 genes in common, only 48 gene-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

motifs are in common. The only exception is C14 and C15, which have 30 gene-motifs in common, while these two subtypes also have 24 genes in common.

Interestingly, our result (Supplementary Figure 5.a) indicates that *IL1RAPL1* mutations in C2 samples within T>A and T>C occurred more among C1 samples, and C5 samples were mutated in a smaller number of motifs compared to C1 and C2. According to Supplementary figures 5.b and 5.c, the same results are observed for *LRRC4C* and *MUC16*.

Many shared genes between different pairs of subtypes (e.g., C1 and C2) led us to investigate the mutational loads within 3-mer motifs in the top 100 important genes, separately. We used Fisher exact test (*method section*) to do this, and we identified significantly mutated motifs within the top 100 significant genes in each subtype, separately. Considering the top 100 gene-motifs for each subtype, we identified common gene-motifs between every two subtypes as shown in Figure 4.b. Interestingly, this analysis showed a more evident difference between the identified subtypes. Compared to Figure 4.a all pairs have less common significant gene-motifs than significant common genes (except C14 and C15, which have 30 common gene-motifs). There is no common gene-motif between most of the paired subtypes (Figure 4.b). Importantly, subtypes C1 and C2, with 93 significant common genes within their top 100 most mutated genes, have only 48 common gene-motifs (within their top 100 gene-motifs), showing different molecular mechanisms within these subtypes. The complete lists of the top 100 significant gene-motifs for each subtype is provided in Supplementary Table S5.

Mutational signature analysis

We also investigated mutational signatures in our identified subtypes. A mutational signature is a fingerprint for a molecular mechanism that is causing mutation across the genome. Molecular mechanisms are blind to what location they are causing the mutation. Therefore, to identify the molecular mechanism of the mutational signature, we have to consider all mutations in the whole genome (except mitochondria). We applied the CANCERSIGN tool (27) on complete mutational profiles of each subtype separately and found 121 signatures. We then compared our identified signatures with 67 signatures identified in COSMIC (27). We calculated the angular similarity between our identified signatures and COSMIC signatures to extract each signature's biological information and their associated subtypes. A Heatmap of similarities between our signatures and Alexandrov signatures is shown in Figure 5. Hierarchical clustering enables us to find similar signatures. Hierarchical clustering enables the finding of similar signatures beside each other. As shown in the figure, COSMIC's signature 1, whose number of mutations correlates with the individual's age, is significantly correlated with many signatures in our identified subtypes, including C1.S1, C2.S1, and C12.S1. COSMIC's signature 1 is shown to be highly associated with breast cancer. Interestingly C1, C2 and C12 contain many breasts cancer samples (27.4%, 16.8%, 22%, respectively, as shown in Supplementary Figure 3). Also, COSMIC's signature two, attributed to the activity of the AID/APOBEC family of cytidine deaminases, is highly observed in the nervous system and is significantly correlated with signatures in C8 (C8.S7 and C8.S1), which consists of nervous system cancer (77.8%). Similarly, COSMIC's signature 4 is also associated with smoking and highly observed in lung cancer, is highly correlated with two signatures in C6 (C6.S4, C6.S5), a subtype that consisted of lung cancer patients (14.3%) and 20.4% of lung samples are in C6. Similarly, COSMIC's signature 5 is associated with Skin cancer and is also correlated with a signature of C17 (C17.S8), which consists of skin cancer patients (38.5%). Furthermore, COSMIC's signature 6 is associated with defective DNA mismatch repair and is correlated with signatures from various subtypes, including C7.S4, C8.S3, C6.S1, C15.S1. Also, COSMIC's signatures 7 and 8 are related to ultraviolet light and Skin cancer. These two signatures are highly correlated with signatures of C14 and C17 (C14.S11 and C17.S10), which consists of skin cancer (40.7% and 38.5%, respectively). Similarly, characteristics of COSMIC's signature 16 is yet unknown but has been observed in liver cancer tumors and are highly correlated with C7.S8 (9.9% of samples in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

this subtype are liver and 32.1% of liver samples are in C7) and C16.S5 (13.6% of samples in this subtype are liver). COSMIC's signature 22 is also highly observed in Eso-AdenoCA cancer. This signature is correlated with C17.S1 (30.2% samples in this subtype are Esophagus). Finally, COSMIC's signature 34 is observed in samples from individuals with a tobacco chewing habit and found in oral and liver cancer. This signature is highly correlated with C6.S5 and C7.S1. 9.3% and 32.1% of liver samples are in C6 and C7, respectively.

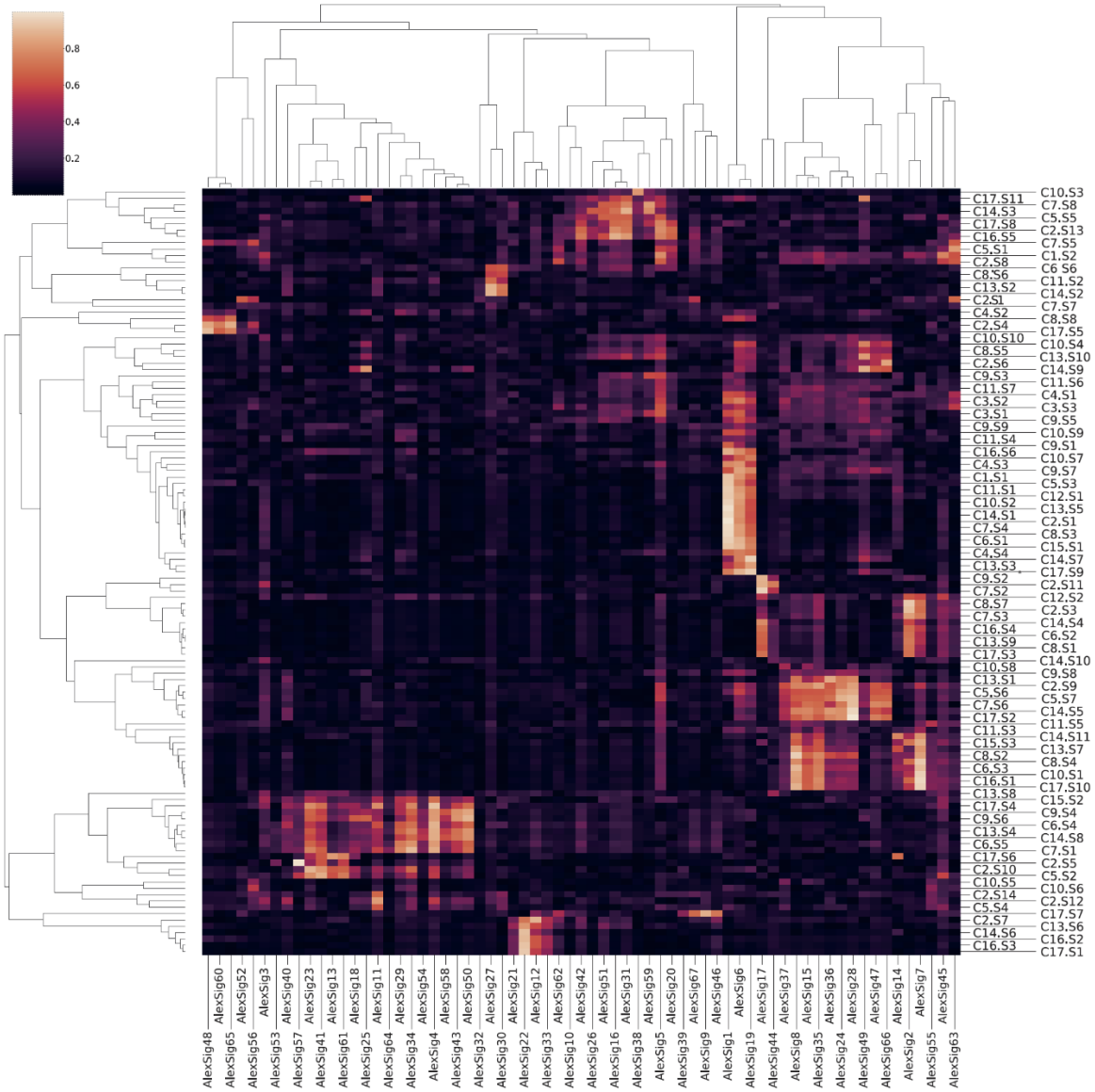


Figure 5. Cluster heatmap between our 121 signatures identified in our study and 67 COSMIC's signatures. Identified signature 'j' from subtype 'I' is shown by Ci.Sj. In this figure, brighter cells correspond with a significant amount of correlation and similarity.

The figure also shows that some of the signatures are presented in multiple subtypes. However, most of the signatures identified in each subtype are specific, indicating that samples within subtypes have the same mutational process. The exact amount of correlation between identified signatures and COSMIC's (Alexandrov's) and correlation between each two identified signatures are provided in (Supplementary Table S7). Molecular mechanism respective to each mutational signature of COSMIC is also provided in this table.

Gene ontology and pathway analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

We next investigate whether each subtype's top 100 significant genes are associated with any gene ontology (GO) or gene pathway terms (36, 37). To do this, we used the *enrichr* (38) package in R (*see method*) for gene ontology and pathway terms analyses. Gene ontology covers three main domains, namely, biological process, molecular function, and cellular component. We considered all these domains and only retained enriched terms with FDR < 0.05. In general, we identified at least one GO term for 10 subtypes out of 17 subtypes (Figure 6). Most GO terms are uniquely enriched in one subtype, while others are enriched in multiple subtypes. For example, the "integral component of plasma membrane" is associated with five subtypes (C2, C13, C14, C15, and C17), and "nervous system development" is associated with another five subtypes (C1, C2, C13, C14, and C15). Conversely, the "bitter taste receptor activity," "MHC class II protein complex", "anterograde trans−synaptic signaling", "actin−myosin filament sliding", and "anion channel activity" are examples of terms that are uniquely associated with C4, C5, C13, C14, and C17, respectively. Moreover, associated terms in C1 and C2 are almost the same, and only three terms associated uniquely in one of them ("axolemma" and "integral component of plasma membrane" only associated with C2 and "integral component of lumenal side of endoplasmic reticulum membrane" only associated with C1).

Figure 6. Gene ontology analysis of identified subtypes. For ten subtypes, we found enriched ontologies. The X-axis shows Gene-ontologies (a different color shows three collections of gene ontologies). The Y-axis shows subtypes, and the darkness of each cell corresponds to the *p*-value for enriched ontology. Many ontologies have significantly enriched for many subtypes, while there is a unique enriched gene ontology.

Clinical report and survival analysis

We also examined clinical data, such as gender and region (where the data were collected), available for a subset of the ICGC data. We identified several interesting results in the gender distribution of subtypes. For instance, C4 and C8, which mainly contains nervous system samples, are female-biased (69% of samples in these subtypes are female), and C2 (48.2% of samples are prostate cancer), C5 (84.3% of samples are prostate, blood, or brain cancers) and C17 (68.7% of samples are skin or esophagus cancers) are male-biased (Figure 7.a). The geographical distribution of identified subtypes is shown in Figure 7.b.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

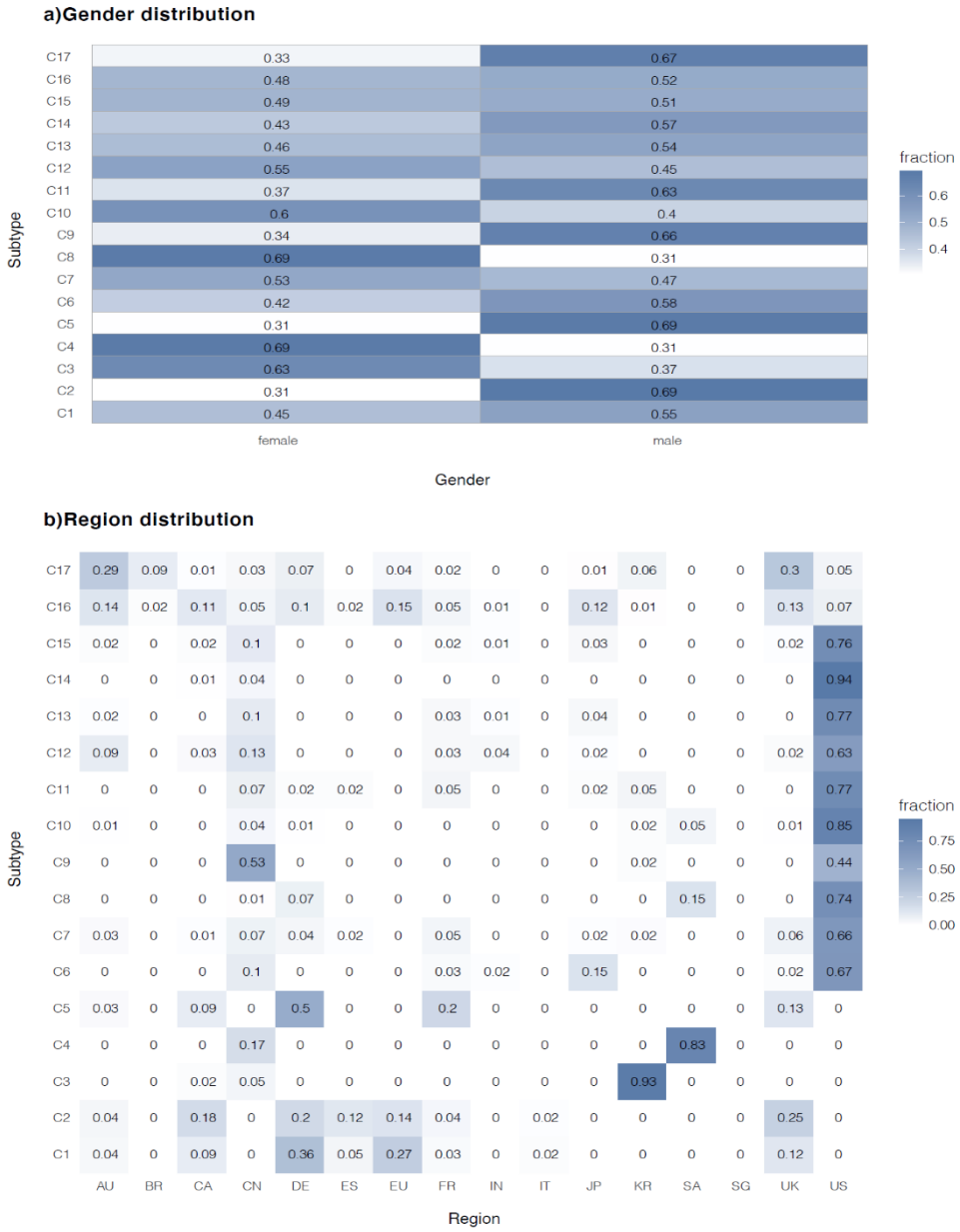


Figure 7. (a) Gender analysis of identified subtypes. The darkness of each cell corresponds to the fraction of samples that are male or female. Some subtypes (C3, C4, C8, and C10) are female-biased, while other subtypes such as C2, C5, C9, C11, and C17 are male-biased. (b) Region distribution analysis of identified subtypes. The darkness of each cell corresponds to the fraction of samples that come from a specific region. Many subtypes are mainly from the US due to many samples from this country in the ICGC dataset, while C3 samples are mainly from Korea and C4 samples are mainly from Saudi Arabia.

We also used molecular data available for a subset of the ICGC dataset to demonstrate the difference between our identified subtypes regarding their survival curve. We begin with excluding samples of patients that were placed in different subtypes. To estimate survival probability over time, we used Kaplan-Meier (39) method and created survival curves for each subtype (Figure 8). As shown in this figure, the difference between subtypes is not only demonstrated by the *p*-value, but it is also observable in the plot itself that the survival times of identified subtypes are different from each other. Since the data we use to cluster samples is entirely based on the somatic mutation data without any clinical information, this survival plot and *p*-value explicate influential biological signals. As shown in Figure 8, more than 75% of patients in C1 (significantly mutated in CSMD1/CNTNAP2) have a good survival length of 10 years. C2 and C5 (significantly mutated in DPP10/PTPRD and DMD, respectively) are also subtypes with a high

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2021 doi:10.20944/preprints202111.0266.v1

chance of survival (survival of 13 years for more than 50% of their patients). However, patients in PCDHGA-driven subtype (C13), patients in PCDHA/PCDHGA-driven subtype (C14), and patients in PCDHA-driven subtype (C15) have the most unfortunate results since only 25% of patients of these subtypes have an overall survival of only five years. Moreover, NBPF/USP17-driven subtype (C3) and CSDE1/NRAS-driven subtype (C10) have the worst survival time (all patients in these two subtypes have a survival length of fewer than six years). Our results suggest that the Protocadherin family, USP17 family, NBPF family, NRAS, and CSDE1 substantially affect survival time.

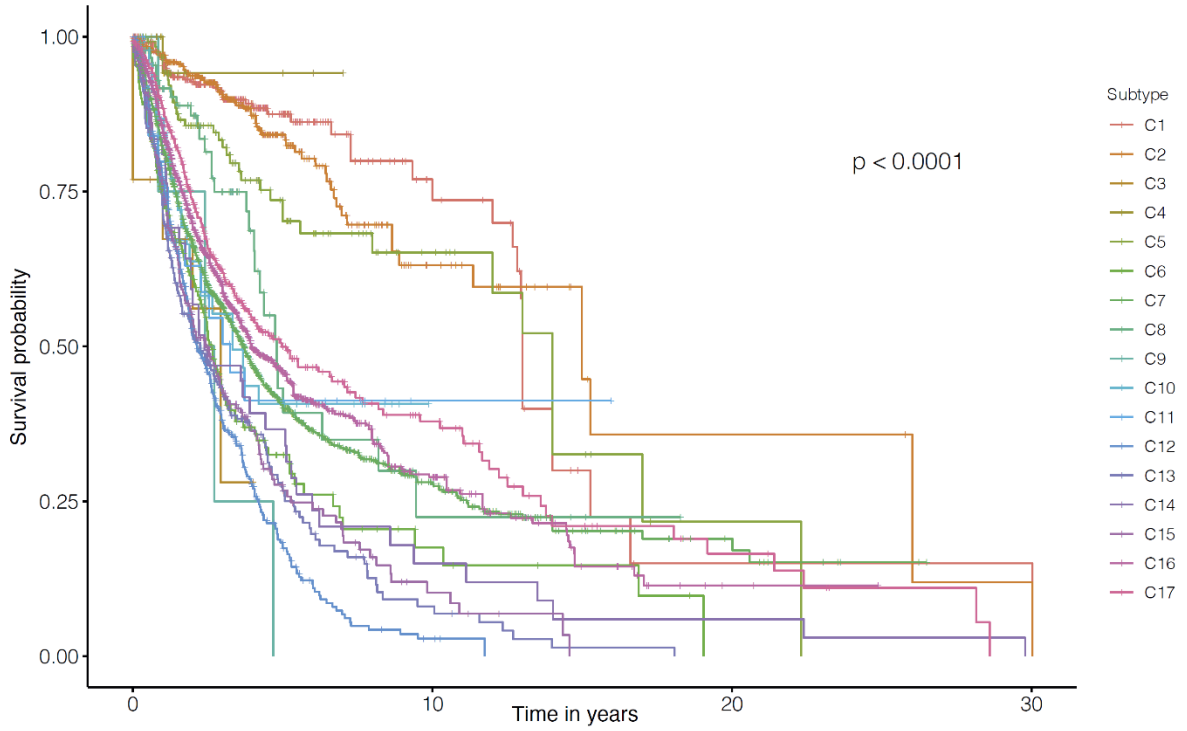


Figure 8. Survival analysis of identified subtypes. Each curve corresponds to a survival curve of a specific subtype. The X-axis shows time in years, and Y-axis shows a fraction of the survived samples. The survival curve demonstrates different survival times of different subtypes. Subtypes such as C3, C9, C12, and C15 have the worst survival, and samples in subtypes such as C1 and C2 have a higher chance of survival.

4. Conclusions

High-throughput sequencing has provided many improvements in finding the key mutations and molecular events by delivering a high number of samples. This will lead to accurate classification of patients based on their mutational profiles, and consequently, better clinical decisions on their treatment. This manuscript used machine learning techniques to provide a new clustering of cancer samples based on their mutational profiles. This can be useful in better understanding the underlying genetic causes of cancers by exploiting the context of the mutations in the driver genes in each subtype. We showed that considering both mutation rates in genes and the contexts of the mutations might be a more effective way to understand the molecular mechanism in cancer genomes. We showed that our proposed pipeline helps discover mutational patterns associated with cancer-related pathways, clinical phenotypes, and cancer subtypes. The source codes for our proposed clustering pipeline and analysis are publicly available at: https://github.com/bcb-sut/Pan-Cancer.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: HAR designed the study. HAR, AK, AD wrote the manuscript. AK and KH performed the analyses including candidate gene and gene-motif identification, clustering, hi-

erarchical models, text mining, and biological characterization (mutational signature analysis gene ontology, survival analysis) with help from HD, AD, MB, MT, HRR, and AD and under HAR supervision. AK generated all figures and tables. HAR, AK, HRR, JB, AD, and NL edited the manuscript; MB analyzed mutational signatures. AK generated all figures and tables. All authors have read and approved the final version of the paper.

Funding: NA.

Data Availability Statement The source codes used in this Study for clustering and analysis of subtypes are provided in https://github.com/bcb-sut/Pan-Cancer.

Acknowledgments: HAR is supported by a UNSW Scientia Program Fellowship. Analysis was made possible with the BioMedical Machine Learning Lab and Telethon Kids Bioinformatics Server's computational resources with funding from the Australian Government and the Government of Western Australia. HRR is supported by IRN National Science Foundation (INSF) Grant No. 96006077.

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

References

1. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929-44.

2. Chen F, Wendl MC, Wyczalkowski MA, Bailey MH, Li Y, Ding L. Moving pan-cancer studies from basic research toward the clinic. Nature Cancer. 2021:1-12.

3. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291-304. e6.

4. Kim H, Kim Y-M. Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. Scientific reports. 2018;8(1):1-14.

5. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609-15.

6. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164-9.

7. Wong G, Leckie C, Kowalczyk A. FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number. Bioinformatics. 2012;28(2):151-9.

8. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17(5):510-22.

9. Zhang W, Feng H, Wu H, Zheng X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics. 2017;33(17):2651-7.

10. Guo Y, Shang X, Li Z. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing. 2019;324:20-30.

11. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell. 2010;17(1):98-110.

12. Jiang L, Xiao Y, Ding Y, Tang J, Guo F. Discovering Cancer Subtypes via an Accurate Fusion Strategy on Multiple Profile Data. Frontiers in Genetics. 2019;10(20).

13. Lin VT, Yang ES. The pros and cons of incorporating transcriptomics in the age of precision oncology. JNCI: Journal of the National Cancer Institute. 2019;111(10):1016-22.

14. Le Morvan M, Zinovyev A, Vert J-P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. PLOS Computational Biology. 2017;13(6):e1005573.

15. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nature Methods. 2013;10(11):1108-15.

16. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics. 2015;47(2):106-14.

17. Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J. Cancer subtype identification using somatic mutation data. British Journal of Cancer. 2018;118(11):1492-501.

18. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research. 2015;43(D1):D805-D11.

19. Ostlund G, Lindskog M, Sonnhammer EL. Network-based Identification of novel cancer genes. Mol Cell Proteomics. 2010;9(4):648-55.

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

21. 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):665-72.

22. Parvin H, Parvin, S. Divide and conquer classification. Australian Journal of Basic and Applied Sciences. 2011;5(12):2446-52.

23. Shamshirband S, Fathi, M., Dehzangi, A., Chronopoulos, A.T. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. Journal of Biomedical Informatics. 2020;113:103627.

24. Cullen AC, Frey HC, Frey CH. Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs: Springer Science & Business Media; 1999.

25. Fraley C, Raftery A. Model-based methods of classification: using the mclust software in chemometrics. Journal of Statistical Software. 2007;18(1):1-13.

26. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal. 1998;41(8):578-88.

27. Bayati M, Rabiee HR, Mehrbod M, Vafaee F, Ebrahimi D, Forrest AR, et al. CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes. Scientific reports. 2020;10(1):1-11.

28. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research. 2016;44(W1):W90-W7.

29. Therneau T. A Package for Survival Analysis in R. R package version 3.2-7. 2020. 2021.

30. Kassambara A, Kosinski M, Biecek P, Fabian S. Package 'survminer'. Drawing Survival Curves using 'ggplot2'(R package version 03 1). 2017.

31. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database. 2016;2016.

32. Tommasino M, Accardi R, Caldeira S, Dong W, Malanchi I, Smet A, et al. The role of TP53 in Cervical carcinogenesis. Hum Mutat. 2003;21(3):307-12.

33. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; Portland, Oregon: AAAI Press; 1996. p. 226–31.

34. Astels LMaJHaS. hdbscan: Hierarchical density based clustering. The Open Journal. 2017;2:205.

35. Ghareyazi A, Mohseni A, Dashti H, Beheshti A, Dehzangi A, Rabiee HR, et al. Whole-Genome Analysis of De Novo Somatic Point Mutations Reveals Novel Mutational Biomarkers in Pancreatic Cancer. Cancers. 2021;13(17):4376.

36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

37. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833-d9.

38. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research. 2016;44(W1):W90-W7.

39. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ (Clinical research ed). 1998;317(7172):1572-.