Yale School of Public Health
BIS 634: Computational Methods for Informatics
FINAL PROJECT REPORT

Name: Nilay Bhatt
Major: Health Informatics

CANCER TYPE PREDICTOR

NOTE: Please access complete codes, .ipynb files, raw and filtered data csvs and html files to run this analysis on your own via my GitHub: https://github.com/itsjustnilay/BIS_634-Cancer-Predictor

INTRODUCTION

Cancer is an aggregation of multiple disorders of genetic or epigenetic nature which may occur through many processes [1][2]. The Central Dogma, delineating the flow of genetic information from DNA to RNA to proteins, serves as a foundational principle in molecular biology. Its significance lies in unraveling the intricacies of genetic processes. In the context of cancer, genetic mutations act as disruptors, perturbing the normal flow of information within the Central Dogma. Specifically, cancer is marked by mutations in genes that regulate growth, causing aberrant cellular proliferation. Notably, disruptions in DNA processes, a consequence of these mutations, significantly contribute to the pathogenesis of cancer. Understanding the Central Dogma is therefore imperative for comprehending the genetic underpinnings of cancer, as deviations from this fundamental process emerge as key contributors to oncogenic transformations.

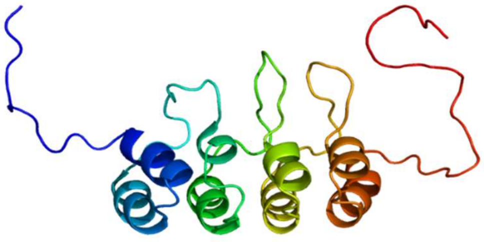


Figure 1: CDKN2A PDB structure

Among causes of progression of cancer, a few of them include improper gene expression, metabolic, genetic and epigenetic aberrations, dysfunction at cellular level, improper cell cycle progression or signal transduction[3][4][5]. Resting cells have a stringent machinery in place in the form of regulatory proteins for managing the cell cycle and they do so by controlling various checkpoints. Several cyclins and cyclin-dependant kinases, or CDKs regulate the progression of the mammalian cell cycle from G1 to mitosis. There also exist a family of CDK inhibitor proteins, which play the function of inactivating the CDKs and act as tumor suppressors [1][6]. Genes known as tumor suppressors prevent the growth of tumors and cell division and encode these proteins [2][7]. These genes are frequently deleted, silenced or inactivated in tumors, eliminating any inhibitors of cell growth, and promoting the uncontrolled growth of tumour cells, or tumorigenesis. Unlimited proliferative capacity, self-sufficient growth signalling, and resistance to anti-proliferative and apoptotic stimuli are characteristics shared by tumor cells [1][8].

p16^{INK4a} or Cyclin-dependent kinase inhibitor 2A protein is a tumor suppressor protein which gets silenced during tumorigenesis in multiple cancers. It is encoded by a gene called CDKN2A (Cyclin-dependent kinase inhibitor 2A, also called multiple tumor suppressor 1 or MST1) which is located on chromosome 9, band p21.3, a stretch of DNA known in oncogenomics as a hotspot for deleterious germ-line mutations and substitutions leading to familial cutaneous melanoma among other cancers [9][10]. In addition to melanomas, mutations in CDKN2A have also been recorded in instances of other cancers, like pancreatic adenocarcinoma, gastric lymphoma, head & neck squamous cell carcinoma, prostate cancer, gastric and colorectal cancer, among many others [11][12][13][14][15]. CDKN2A gene (8.5 kb full length) contains two introns and three exons. It codes for two tumor-suppressor proteins, p16^{INK4a} and p14^{ARF}, which are coded by alternatively spliced transcripts of the first exon [16]. Both of these proteins work in conjunction with a cascade of other proteins to maintain proper functioning of the cell cycle in transitioning from G1 to S phase. The p16^{INK4a} or CDKN2A protein consists of 156 amino acids with a molecular weight of 16 kDa and is a negative regulator of the cell cycle. In the presence of stress conditions, like DNA damage or oncogenic signals, p16^{INK4a} is expressed and it stops improper cell division. If overexpressed, it may also lead to cell senescence [17][18].

In this project I have leveraged this information and importance of this gene to understand cancer associated by building a cancer predictor along with doing other analyses which will be described under other sections. The complete runnable python code is also attached along with this report as a separate file called “complete code.pdf”.

This project essentially involves applying the computational steps that we have discussed in the class to understand cancer association with CDKN2A. In brief, I have found a dataset, standardized it, and provided a web interface for analyzing the data.

DATASET and DATA

Describe the dataset and why is this data interesting?

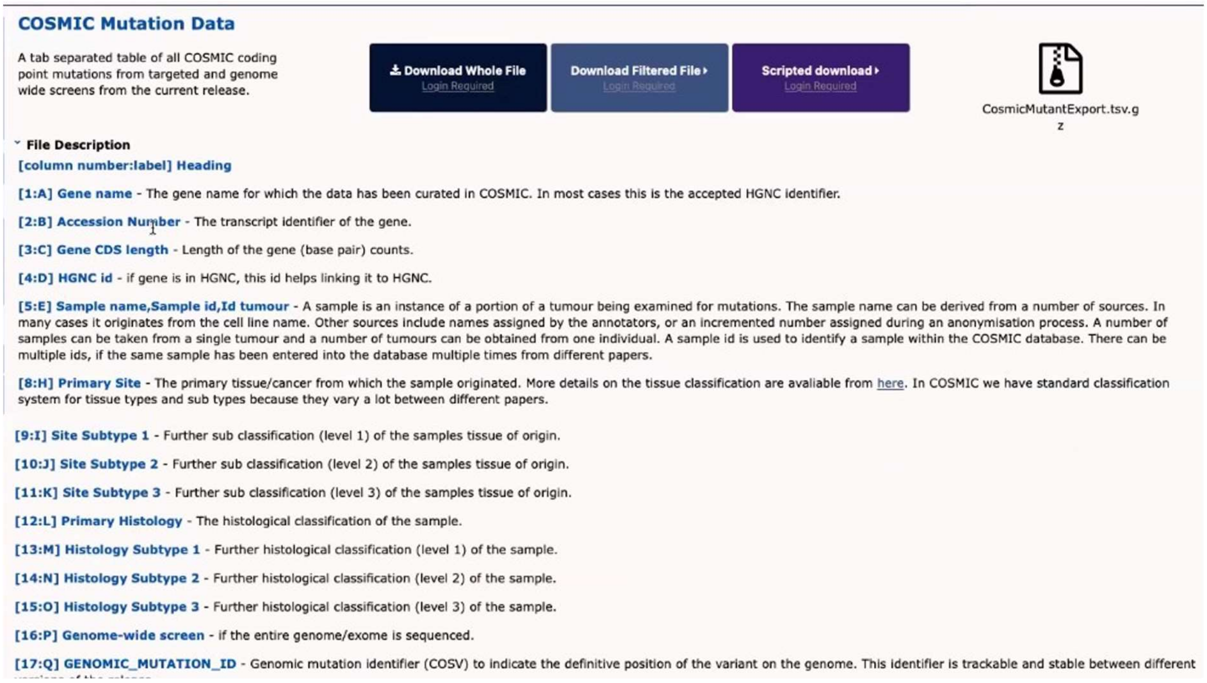
The CDKN2A gene data was taken from the COSMIC database (the Catalogue Of Somatic Mutations In Cancer), which is an online database comprising of almost 6 million coding mutations across 1.4 million tumor samples, curated from over 26000 publications making it one of the most extensive and thorough databases for somatic mutation study in cancer [19]. It can be accessed via https://cancer.sanger.ac.uk/.

The dataset provides a comprehensive overview of somatic mutations in the CDKN2A gene across various cancer types. With 3716 entries spanning 37 columns, it captures a diverse landscape of genetic alterations, including point mutations, insertions, deletions, and silent substitutions. It also captures information on a genomic, proteomic and a histopathological level along with clinical information of the samples collected, as well as genomic coordinates of the mutation. The CDKN2A gene is a known tumor suppressor associated with multiple cancer types, and it exhibits a range of mutations that potentially disrupt its function. This dataset is particularly intriguing as it unveils the molecular intricacies of CDKN2A across different tissues, shedding light on the diverse genetic alterations contributing to cancer development. Understanding these variations can offer valuable insights into the specific mechanisms underlying tumorigenesis, aiding in the development of targeted therapeutic interventions and personalized treatment strategies.

Explain how you acquired it (e.g. via an API, file download, etc). Discuss the FAIRness of the data provider. Include: Was the data well-annotated with metadata? Was the license clear?

The data was acquired via a file download. A user has to create a non-commercial account using their institutional id in order to acquire the data as a .tsv.gz file under a Non-commercial licence agreement which can be unzipped to obtain a CSV file. Non-Commercial license means that the data is not primarily intended for or directed towards commercial advantage or monetary compensation. In addition to this, citation is required also [19].

The data in this scientific report adheres to the principles of FAIRness (Findable, Accessible, Interoperable, and Reusable) by incorporating unique identifiers for each somatic mutation entry, ensuring effective search and retrieval. The metadata is well-described, facilitating comprehensive understanding, and the dataset is made available through open or controlled access, prioritizing privacy considerations. Standard protocols are implemented for seamless data retrieval, and the information is structured in standardized formats, promoting interoperability. The inclusion of APIs and exchange formats further enhances accessibility and usability. Clear licensing and usage policies are articulated to provide transparency and legal clarity. Specifically, a non-commercial license is stipulated for non-commercial use, with due consideration for citation requirements. This approach not only upholds the FAIR principles but also fosters responsible and ethical data sharing within the scientific community, promoting transparency and collaboration.



COSMIC Mutation Data

A tab separated table of all COSMIC coding point mutations from targeted and genome wide screens from the current release.

File Description

[column number:label] Heading

[1:A] Gene name - The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.

[2:B] Accession Number - The transcript identifier of the gene.

[3:C] Gene CDS length - Length of the gene (base pair) counts.

[4:D] HGNC id - if gene is in HGNC, this id helps linking it to HGNC.

[5:E] Sample name,Sample id,Id tumour - A sample is an instance of a portion of a tumour being examined for mutations. The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymisation process. A number of samples can be taken from a single tumour and a number of tumours can be obtained from one individual. A sample id is used to identify a sample within the COSMIC database. There can be multiple ids, if the same sample has been entered into the database multiple times from different papers.

[8:H] Primary Site - The primary tissue/cancer from which the sample originated. More details on the tissue classification are avaliable from here. In COSMIC we have standard classification system for tissue types and sub types because they vary a lot between different papers.

[9:I] Site Subtype 1 - Further sub classification (level 1) of the samples tissue of origin.

[10:J] Site Subtype 2 - Further sub classification (level 2) of the samples tissue of origin.

[11:K] Site Subtype 3 - Further sub classification (level 3) of the samples tissue of origin.

[12:L] Primary Histology - The histological classification of the sample.

[13:M] Histology Subtype 1 - Further histological classification (level 1) of the sample.

[14:N] Histology Subtype 2 - Further histological classification (level 2) of the sample.

[15:O] Histology Subtype 3 - Further histological classification (level 3) of the sample.

[16:P] Genome-wide screen - if the entire genome/exome is sequenced.

[17:Q] GENOMIC_MUTATION_ID - Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different

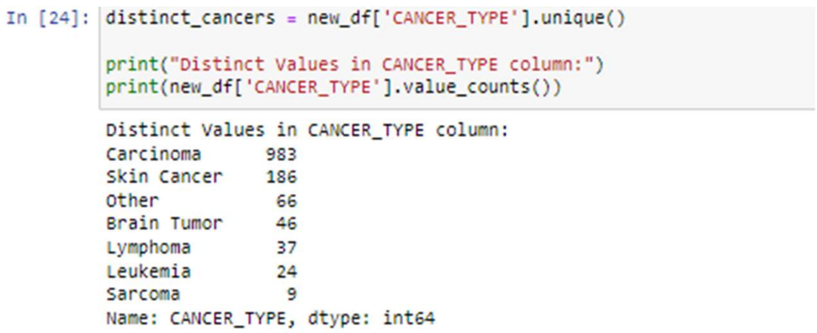
Figure 2: **Some features included in the dataset acquired.**

Describe any data cleaning or other preprocessing. E.g., If some data was missing, how did you handle it?

Extensive data cleaning was done before the dataset was analyzed. Initially the data included 3716 entries. Firstly, the dataset was filtered to include only missense substitution mutation entries. This reduces the number of rows to 1386 entries. Now, the data was filtered to remove mutation points where there is no genomic reference coordinate. This is the data, if missing, should be removed. We find that all entries have a genomic reference. This is good because all our records have known substitution information. There is no missing data of significance to be dealt with. Then the data is filtered to remove mutations pertaining to insertion, duplication, inversion, and deletion because our focus is on missense substitution mutations, which reduces our records to 1351 from the initial

3716. Now that the data has been filtered as much as it was possible, it needs to be rearranged to optimize the analyses. Firstly, data is rearranged to get genomic level features. This includes the gene name, accession number, length of the gene CDNA, HSVSG (genomic coordinate), HGVSC (the HGVS coding sequence name), and the mutation information. New columns are created using information extracted from other columns to encode information regarding the base allele, mutation event and mutant allele. The proteomic level features are built. We create a dictionary with single and three letter codes for amino acids and use it to map and build columns including the wild type (amino acid that got substituted) and mutant amino acids (the amino acid that substituted it) in both, 1-letter, and 3-letter formats. Other relevant columns like HGVSP (the HGVS protein sequence name), mutation event and codon position are included.

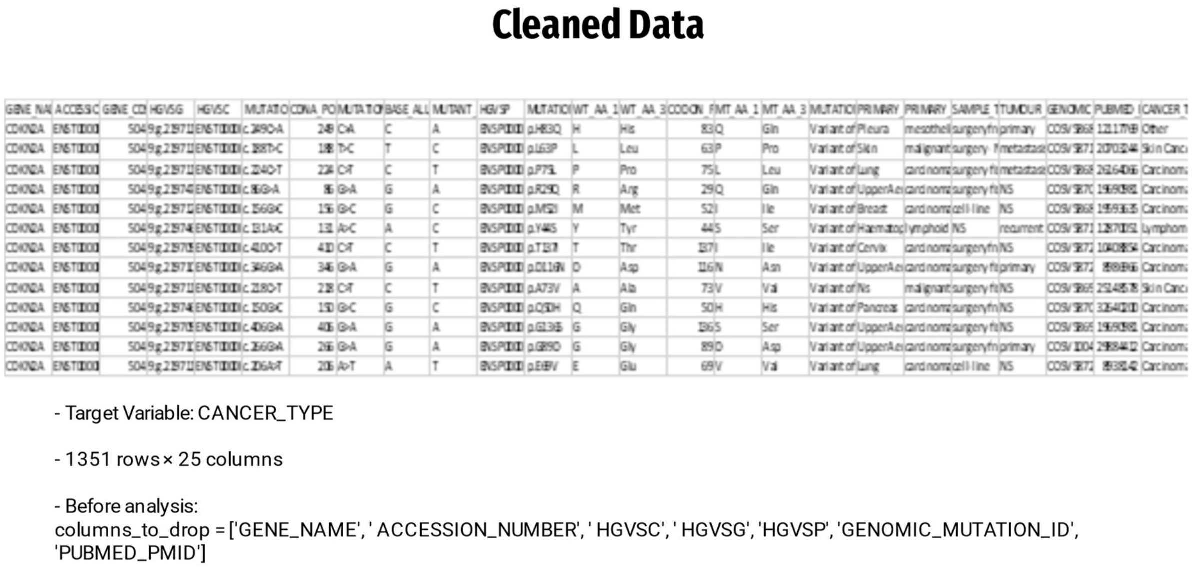
Ultimately, site, histopathology and tissue-specific features are added. The target variable is encoded in a new column called "CANCER_TYPE" which includes distinct values from "PRIMARY_HISTOLOGY column. We observe that 35 distinct cancer types are observed. Based on the counts of these, we group these 35 cancer types in seven distinct categories for optimizing our classification for the cancer type predictor.



```
In [24]: distinct_cancers = new_df['CANCER_TYPE'].unique()

         print("Distinct Values in CANCER_TYPE column:")
         print(new_df['CANCER_TYPE'].value_counts())

         Distinct Values in CANCER_TYPE column:
         Carcinoma      983
         Skin Cancer    186
         Other           66
         Brain Tumor     46
         Lymphoma        37
         Leukemia        24
         Sarcoma          9
         Name: CANCER_TYPE, dtype: int64
```

Figure 3: **Classes of CANCER_TYPE**

This is what our final dataframe looks like:



Cleaned Data

GENE_NA	ACCESSIC	GENE_CD	HGVSG	HGVSC	MUTATIO	CDNA_PO	MUTATIO	BASE_ALL	MUTANT	HGVSP	MUTATIO	WT_AA_1	WT_AA_3	CODON_F	MT_AA_1	MT_AA_3	MUTATIO	PRIMARY	PRIMARY	SAMPLE_T	TUMOUR	GENOMIC	PUBMED_	CANCER_T
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.249C>A	249	C>A	C	A	ENSP0000	p.H83Q	H	His	83	Q	Gln	Variant of	Pleura	mesotheli	surgery fn	primary	COSV5868	12117769	Other
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.188T>C	188	T>C	T	C	ENSP0000	p.L63P	L	Leu	63	P	Pro	Variant of	Skin	malignant	surgery - f	metastasis	COSV5871	20703244	Skin Canc
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.224C>T	224	C>T	C	T	ENSP0000	p.P75L	P	Pro	75	L	Leu	Variant of	Lung	carcinom	surgery fi	metastasis	COSV5868	26164066	Carcinom
CDKN2A	ENST0000	504	9:g.21974	ENST0000	c.86G>A	86	G>A	G	A	ENSP0000	p.R29Q	R	Arg	29	Q	Gln	Variant of	UpperAe	carcinom	surgery fi	NS	COSV5870	19690981	Carcinom
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.156G>C	156	G>C	G	C	ENSP0000	p.M52I	M	Met	52	I	Ile	Variant of	Breast	carcinom	cell line	NS	COSV5868	19593635	Carcinom
CDKN2A	ENST0000	504	9:g.21974	ENST0000	c.131A>C	131	A>C	A	C	ENSP0000	p.Y44S	Y	Tyr	44	S	Ser	Variant of	Haematop	lymphoid	NS	recurrent	COSV5871	12870051	Lymphom
CDKN2A	ENST0000	504	9:g.21970	ENST0000	c.410C>T	410	C>T	C	T	ENSP0000	p.T137I	T	Thr	137	I	Ile	Variant of	Cervix	carcinom	surgery fn	NS	COSV5872	10408854	Carcinom
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.346G>A	346	G>A	G	A	ENSP0000	p.D116N	D	Asp	116	N	Asn	Variant of	UpperAe	carcinom	surgery fi	primary	COSV5872	8898966	Carcinom
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.218C>T	218	C>T	C	T	ENSP0000	p.A73V	A	Ala	73	V	Val	Variant of	NS	malignant	surgery fi	NS	COSV5865	25148578	Skin Canc
CDKN2A	ENST0000	504	9:g.21974	ENST0000	c.150G>C	150	G>C	G	C	ENSP0000	p.Q50H	Q	Gln	50	H	His	Variant of	Pancreas	carcinom	surgery fn	NS	COSV5870	32540000	Carcinom
CDKN2A	ENST0000	504	9:g.21970	ENST0000	c.406G>A	406	G>A	G	A	ENSP0000	p.G136S	G	Gly	136	S	Ser	Variant of	UpperAe	carcinom	surgery fi	NS	COSV5865	19690981	Carcinom
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.266G>A	266	G>A	G	A	ENSP0000	p.G89D	G	Gly	89	D	Asp	Variant of	UpperAe	carcinom	surgery fn	primary	COSV1004	29884412	Carcinom
CDKN2A	ENST0000	504	9:g.21971	ENST0000	c.206A>T	206	A>T	A	T	ENSP0000	p.E69V	E	Glu	69	V	Val	Variant of	Lung	carcinom	cell line	NS	COSV5872	8938142	Carcinom

- Target Variable: CANCER_TYPE

- 1351 rows × 25 columns

- Before analysis:
columns_to_drop = ['GENE_NAME', ' ACCESSION_NUMBER', ' HGVSC', ' HGVSG', 'HGVSP', 'GENOMIC_MUTATION_ID', 'PUBMED_PMID']

Figure 4: **Cleaned dataframe**

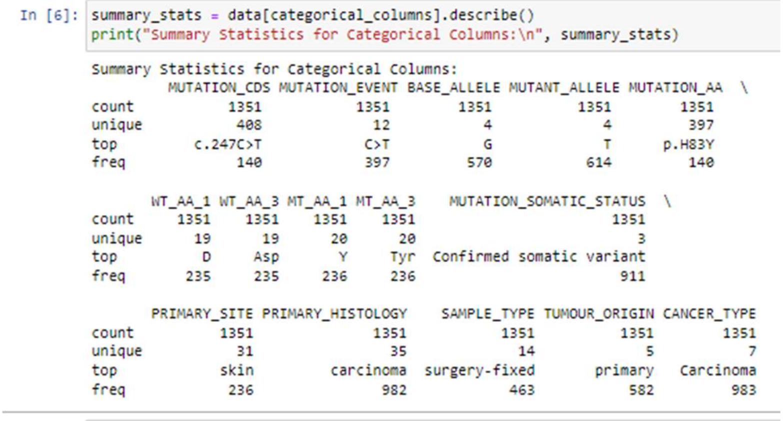
These columns mentioned to be dropped in Figure 9 are dropped before analysis, not from the dataframe. They are dropped because they do provide any useful information for classification purposes.

DATA VISUALIZATION

The data gives us a look at the genetic changes happening in the CDKN2A gene across different types of cancer. Summary statistics were performed. We see that the mutation c.247C>T is the most common, appearing 140 times, and c.341C>T follows with 74 occurrences. This diversity in genetic alterations highlights the complexity of changes in this gene.

Looking at the building blocks of DNA, the base alleles, G and T, stand out as the most frequent. Moving to the protein level, we see various alterations like p.H83Y, p.P114L, and p.D84N. These changes might have important functions. The dataset also tells us about the status of these mutations, where they occur, how tissues are classified, the types of samples, where the tumors start, and the specific types of cancer involved. This comprehensive information gives us a detailed picture of how the CDKN2A gene is behaving in different cancer situations, providing valuable insights.

Most noticeably, The dataset includes 1351 somatic mutations, with 408 unique mutations. The most frequent mutation, as mentioned before, is c.247C>T (140 occurrences). Predominant mutation characteristics include C>T changes (397 occurrences), G as the top base allele (570 occurrences), and p.H83Y as the most common amino acid change (140 occurrences). Confirmed somatic variants account for 911 mutations. Across 31 primary sites, carcinoma is the leading histology (982 occurrences), and surgery-fixed samples are predominant (463 occurrences), with skin being the most common primary site. There are not any significant outliers because most data features are categorical and those that are numerical are codon positions that lay within biological constraints.



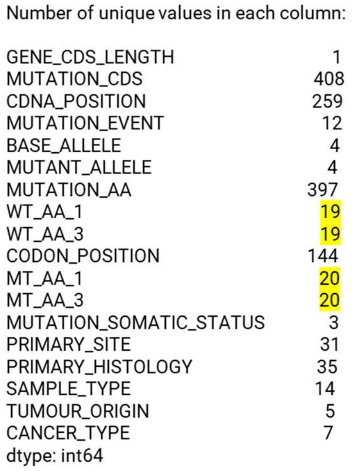
```
In [6]: summary_stats = data[categorical_columns].describe()
        print("Summary Statistics for Categorical Columns:\n", summary_stats)

Summary Statistics for Categorical Columns:
        MUTATION_CDS MUTATION_EVENT BASE_ALLELE MUTANT_ALLELE MUTATION_AA  \
count          1351           1351        1351          1351        1351
unique          408             12           4             4         397
top         c.247C>T           C>T           G             T      p.H83Y
freq            140            397         570           614         140

       WT_AA_1 WT_AA_3 MT_AA_1 MT_AA_3    MUTATION_SOMATIC_STATUS  \
count     1351    1351    1351    1351                       1351
unique      19      19      20      20                          3
top          D     Asp       Y     Tyr  Confirmed somatic variant
freq       235     235     236     236                        911

       PRIMARY_SITE PRIMARY_HISTOLOGY   SAMPLE_TYPE TUMOUR_ORIGIN CANCER_TYPE
count          1351              1351          1351          1351        1351
unique           31                35            14             5           7
top            skin         carcinoma surgery-fixed       primary   Carcinoma
freq            236               982           463           582         983
```

Figure 5: **Summary Statistics**

Discuss any ways in which summary statistics on your data might be misleading. (e.g., are they skewed by outliers, etc.?)

There is a misleading facet in summary statistics (see figure 6). WT_AA_1 has 19 distinct counts and WT_AA_3 has 20. This is despite me providing a correct key for adding information on one and three letter codes.

This is misleading. There are no other cases where the summary statistics on the data can be misleading.



```
Number of unique values in each column:

GENE_CDS_LENGTH              1
MUTATION_CDS               408
CDNA_POSITION              259
MUTATION_EVENT              12
BASE_ALLELE                  4
MUTANT_ALLELE                4
MUTATION_AA                397
WT_AA_1                     19
WT_AA_3                     19
CODON_POSITION             144
MT_AA_1                     20
MT_AA_3                     20
MUTATION_SOMATIC_STATUS      3
PRIMARY_SITE                31
PRIMARY_HISTOLOGY           35
SAMPLE_TYPE                 14
TUMOUR_ORIGIN                5
CANCER_TYPE                  7
dtype: int64
```

Figure 6: **Number of unique values per feature**

Other Data Visualizations

The count plots for primary site (i.e., organ where the mutation was found), mutation event, sample type, tumour origin and cancer types can be found in Figure 7. They validate the information mentioned above.

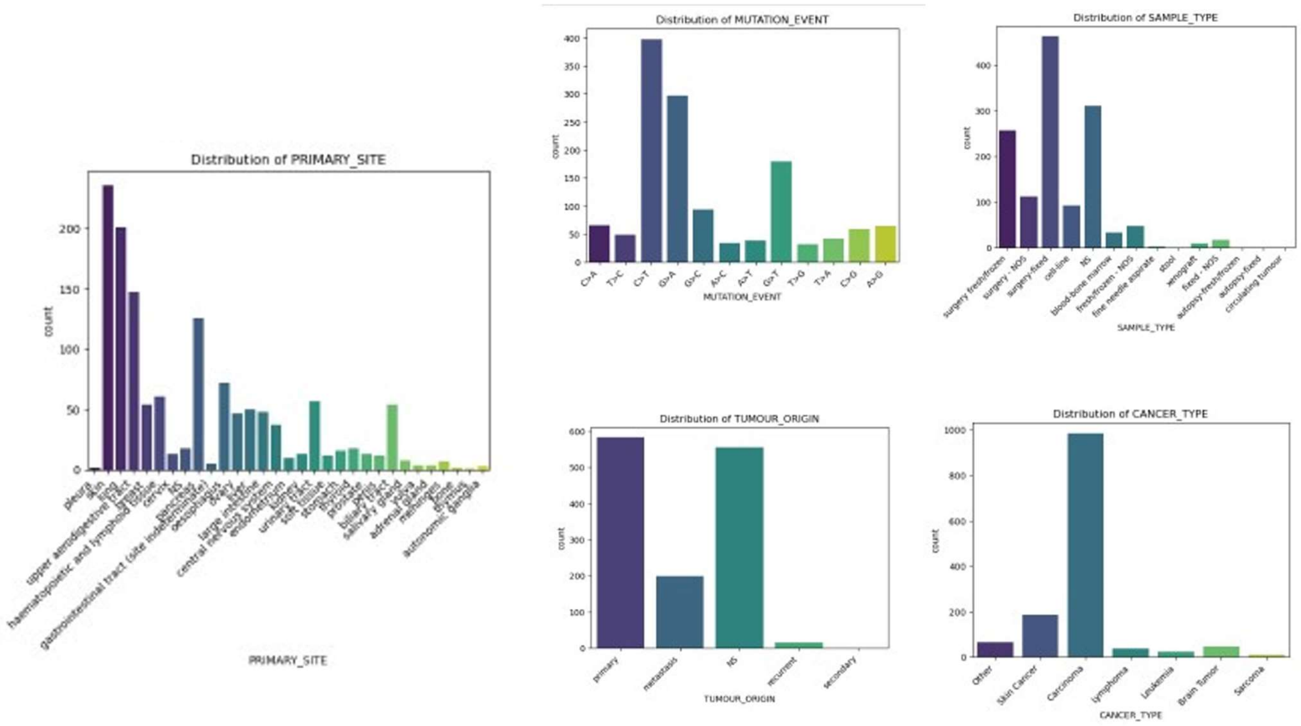
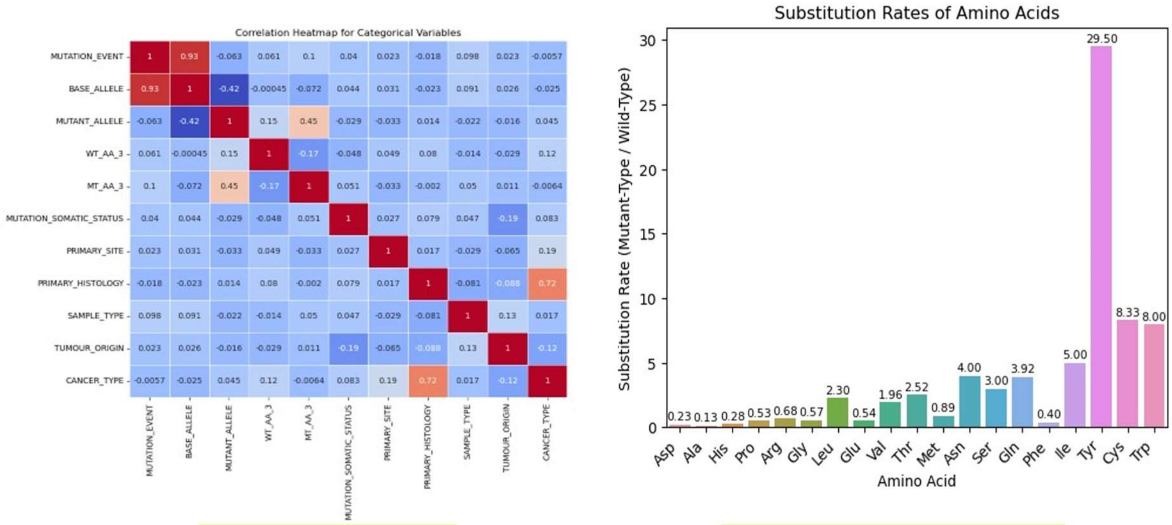


***Figure 7*: The count plots for primary site, mutation event, sample type, tumour origin and cancer types**

I have also performed other analyses, like the Pearson Correlation matrix heatmap (Figure 8a) and substitution rates for each amino acid (Figure 8b).



a. CorrelationHeatmap

b. SubstitutionRates of Amino acids

***Figure 8:* a) Pearson Correlation Matrix b) Substitution rate per amino acid**

The correlation matrix for categorical variables provides insights into potential relationships among different attributes within the dataset. Notably, the strong positive correlation between mutation event and base allele (0.93) suggests a significant association between the mutation type and the base allele involved. Additionally, the negative correlation between mutant allele and base allele (-0.42) indicates an inverse relationship, implying that certain mutations tend to occur more frequently with specific base alleles. The moderate positive correlation between wild type amino acids and mutant amino acid (0.45) suggests some consistency in the amino acid alterations in the wild-type and mutant states. These correlations provide valuable insights for further exploration and may guide more targeted analyses in understanding the intricate relationships among genetic and clinical attributes in the context of cancer-related mutations.

The graph for substitution rates for each amino acid shows that Tyrosine is substituted maximally (29%) and Alanine is the least substituted (13%). This analysis provides insights into the conservation, variability, and functional significance of different amino acids within a protein, aiding in the understanding of evolutionary constraints and adaptive changes in protein sequences.

Figure 9 shows Substitution matrix for amino acids. It is displaying the counts of amino acid substitutions between wild-type (WT) and mutant (MT) amino acids in a protein. Each cell denotes the number of occurrences for a specific substitution event. For instance, the most frequent substitution is observed between histidine in the WT and tyrosine in the MT (146). The total counts at the bottom of the matrix highlight the overall diversity and complexity of amino acid substitutions within the protein. Analyzing this matrix can offer valuable insights into the preferential amino acid changes and potential functional implications in the evolutionary or pathological context of the protein.

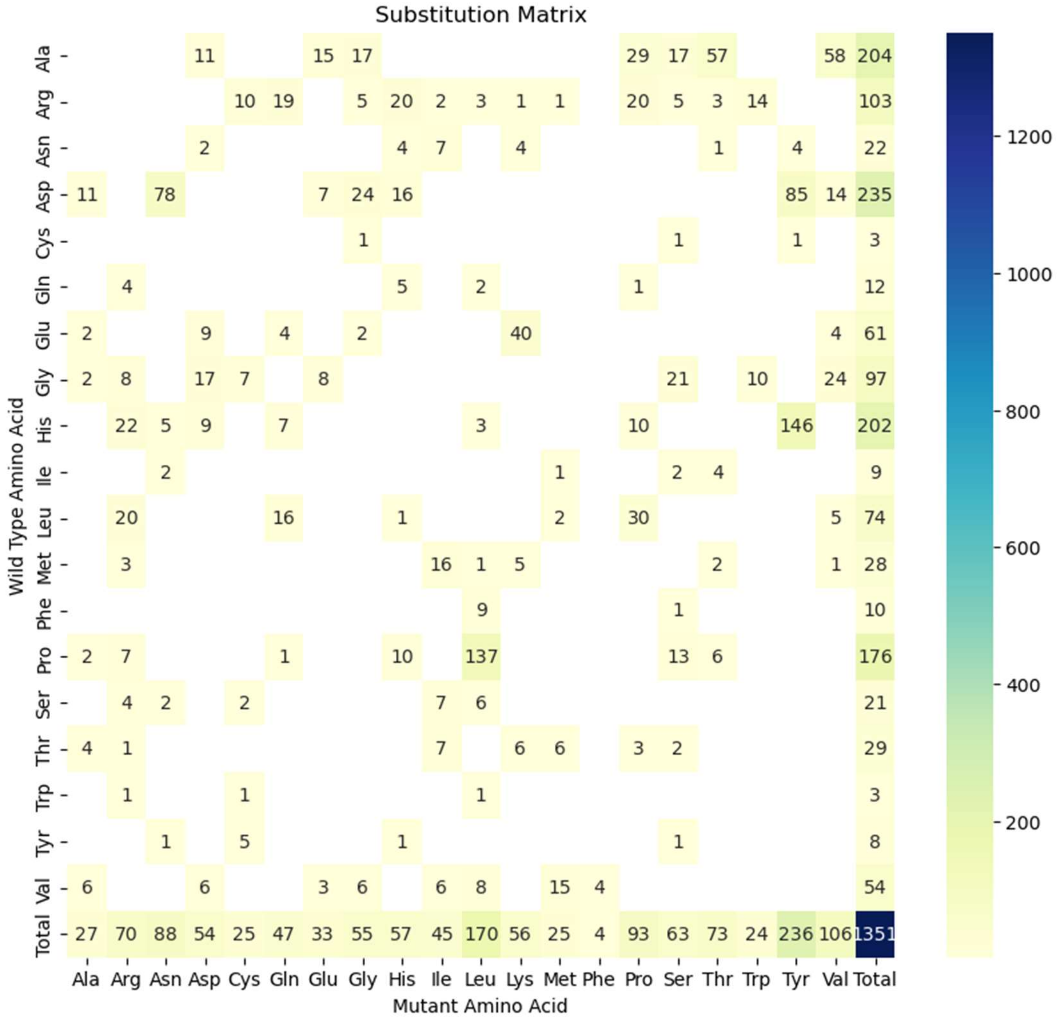


Figure 9: **Substitution matrix for amino acids**

Figure 10 shows cancer types as per primary sites they are found in. Skin carcinoma and lung carcinoma are notably frequent, with 62 and 200 occurrences, respectively, emphasizing their significant impact. Additionally, it reveals diversity in cancer types associated with specific organs, such as brain tumors in the central nervous system and leukemia in haematopoietic and lymphoid tissues. The presence of certain cancers, such as those in the adrenal gland, autonomic ganglia, and soft tissues, is relatively limited. This analysis helps in identification of organ-specific cancer patterns.

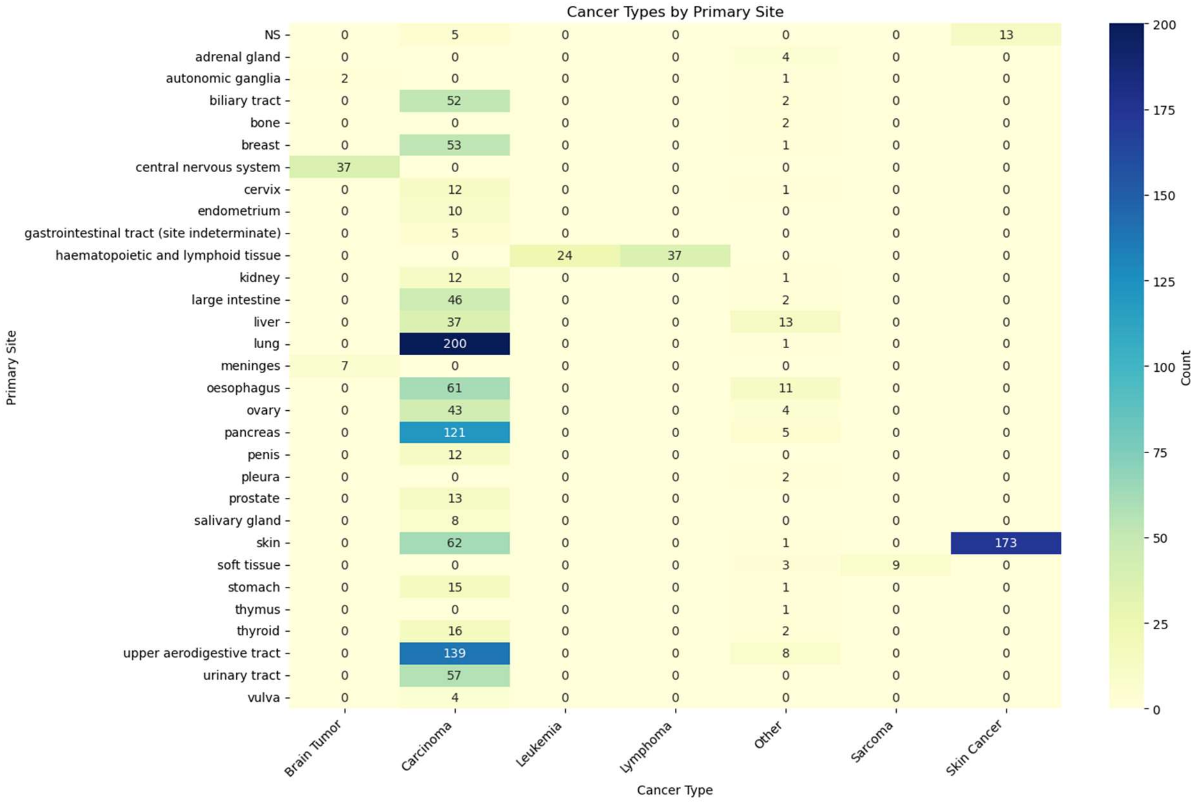


***Figure 10*: Cancer Type Matrix by Primary Site**

DATA ANALYSES

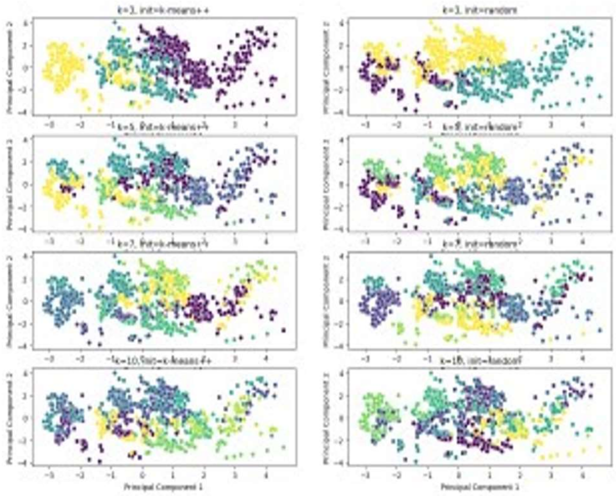
I used two techniques or analyses to understand this data. One of them is unsupervised, k-means clustering and other is supervised, XGBoost predictor model.

The rationale behind using k-means clustering for the classification of this dataset to predict cancer type lies in its ability to identify inherent patterns and groupings within the data based on similarity in feature space. K-means clustering is an unsupervised machine learning algorithm that partitions the dataset into k clusters, with each cluster representing a group of data points that share similarities. In our context for cancer type prediction, the features used for clustering might represent various genetic or molecular characteristics associated with different types of cancer. By applying k-means clustering, the algorithm then would aim to find natural clusters within the dataset, where data points within the same cluster are more similar to each other than to those in other clusters. The assumption is that these clusters might correspond to different cancer types or subtypes. My reasoning for doing this is to make use of genomic, proteomic, and histopathological data to do predictive modeling using this approach to classify correct cancer class. While my approach is an oversimplification, one hopes that these sorts of analysis are useful in cases of cancers which are especially rare. Unfortunately, not enough data is available for rare tumours and cancers, so we go on with our 7 classes here.

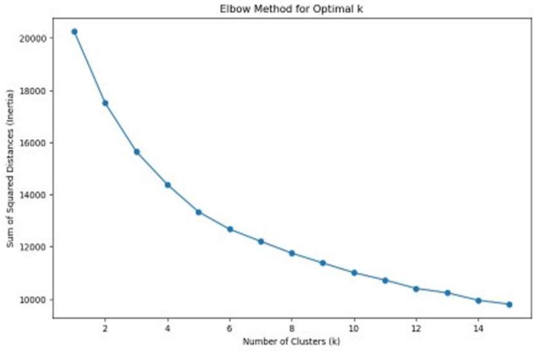
I use two initialization methods, kmeans++ and random to observe clustering effect on our data for k=3,5,7 and 10. Then I use the Elbow method to determine an optimal k value. Finally, I perform k-means clustering and PCA for k=7, my optimal value and analyze it to obtain Adjusted Rand Index: 0.05504803139599864 and a Silhouette Score: 0.1375302277714877. The Adjusted Rand Index (ARI) of 0.055 indicates a weak agreement between the true clustering structure and the clusters identified by the algorithm. A positive ARI suggests some similarity, but the low value suggests that the clustering results may not align well with the actual groups. The Silhouette Score of 0.138 indicates a fair level of cohesion and separation among the clusters, suggesting moderate quality in the clustering assignments.

So, for validation, I used Elbow method here and my surprise? The number of clusters recommended by the elbow is same as number of classes we originally split cancer types into, seven!

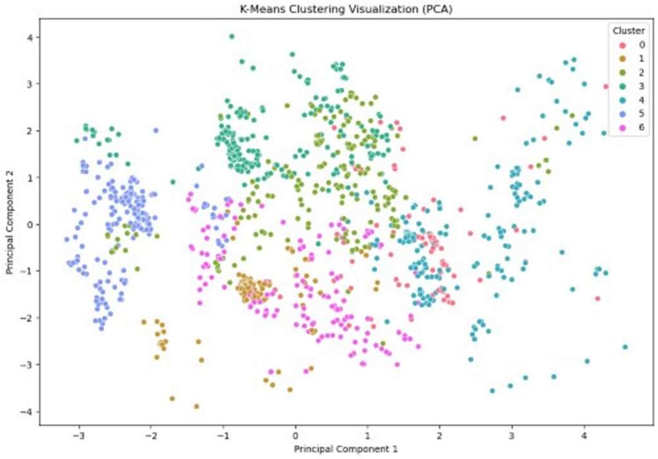
K-means clustering



K-means using different initialization methods and different k values.



Finding optimal k using the elbow method.

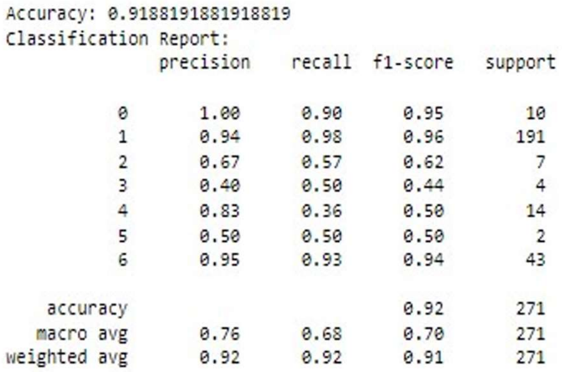


K-means clustering with k=7.

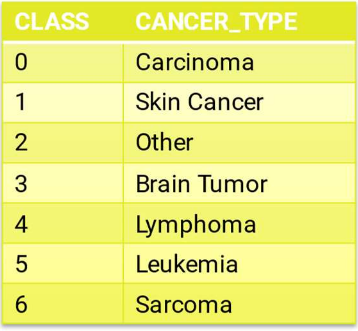
***Figure 11*: K-means clustering results.**

The other analysis I run is XGBoost for predictive modelling. XGBoost is chosen for classification in this dataset due to its effectiveness in handling complex relationships within the data and its ability to handle a large number of features. It is an ensemble learning algorithm that combines the strength of multiple decision trees, making it robust and capable of capturing non-linear patterns and interactions in the genetic data. Its ability to handle imbalanced datasets, feature importance analysis, and robustness against overfitting also contribute towards reasoning behind using it. XGBoost is well-suited for predicting cancer types in this dataset, providing accurate and interpretable results.

Figure 12 shows results of using XGBoost. The data was split into 80-20 for training and test sets and categorical variables were label encoded. The hyperparameters used were library defaults. The model demonstrates an overall accuracy of 91.88%, suggesting its effectiveness in predicting cancer types. The precision and recall values for each class vary, indicating differences in the model's ability to correctly identify instances of specific cancer types. Notably, the high precision and recall for Class 0 and Class 1 imply accurate predictions for these classes, while the lower values for Classes 2, 3, and 4 suggest challenges in distinguishing these cancer types. The weighted average F1-score of 91% underscores the model's balanced performance across classes. In a biological context, the model's ability to differentiate between cancer types is crucial for personalized treatment strategies, and the analysis of misclassifications could provide insights into genetic similarities or complexities among certain cancer subtypes.



```
Accuracy: 0.9188191881918819
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.90      0.95        10
           1       0.94      0.98      0.96       191
           2       0.67      0.57      0.62         7
           3       0.40      0.50      0.44         4
           4       0.83      0.36      0.50        14
           5       0.50      0.50      0.50         2
           6       0.95      0.93      0.94        43

    accuracy                           0.92       271
   macro avg       0.76      0.68      0.70       271
weighted avg       0.92      0.92      0.91       271
```



CLASS	CANCER_TYPE
0	Carcinoma
1	Skin Cancer
2	Other
3	Brain Tumor
4	Lymphoma
5	Leukemia
6	Sarcoma

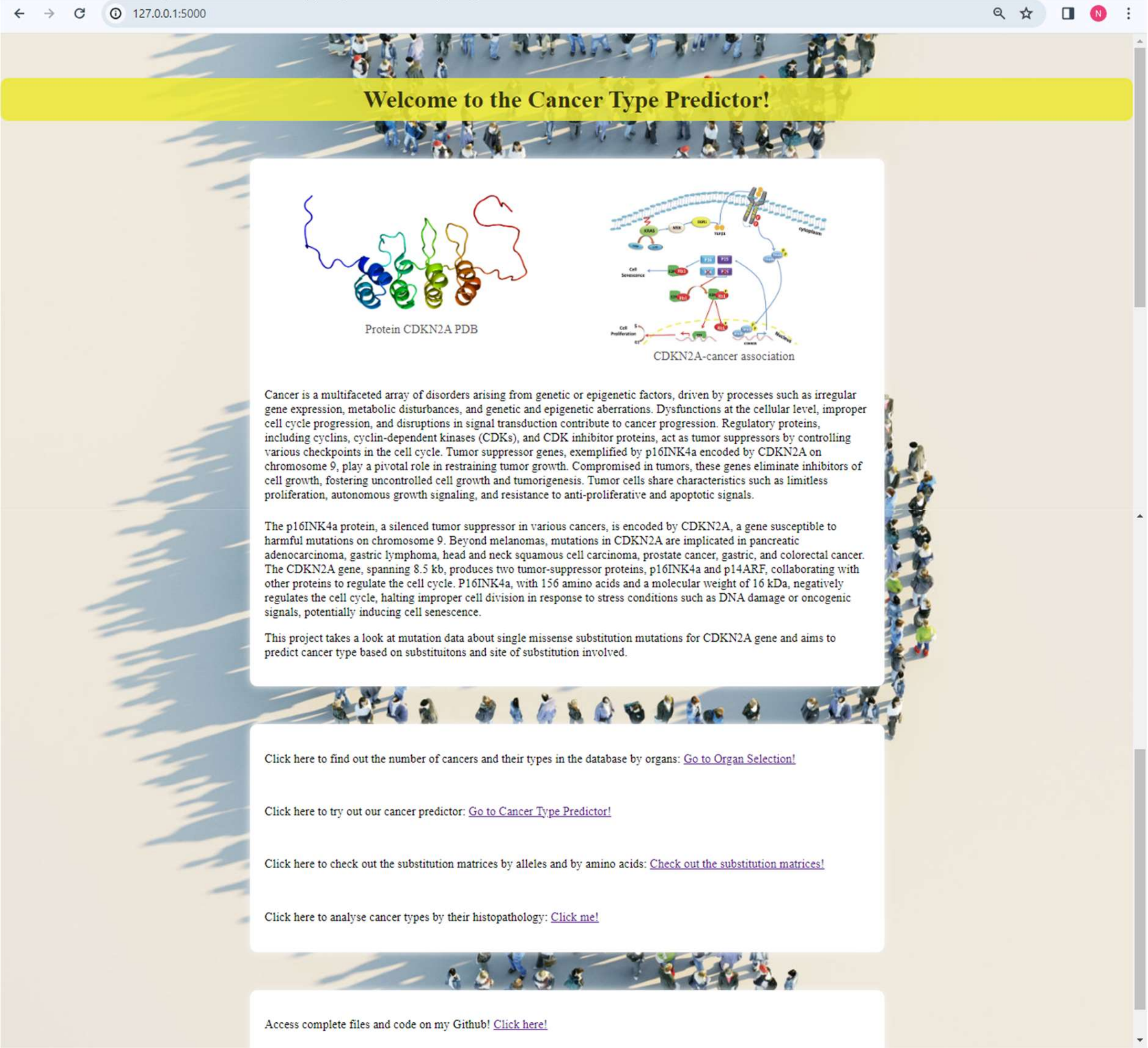
Figure 12: **XGBoost performance**

SERVER API & WEB FRONT-END

These analyses served as a playground for developing a server API and a web front-end framework using Flask. The code for all html templates and the flask implementation can be found on the GitHub under templates folder and app.py file respectively.

The web server shows selected visualizations and involves user interactivity in letting user select analyses they want to see. It also features a cancer type predictor which lets users input their alleles (wild type and mutant) and select a primary site in order to find out the probability of all possible cancer types for that location within the substitution constraints.

The following is the webpage homepage:



Welcome to the Cancer Type Predictor!

Protein CDKN2A PDB

CDKN2A-cancer association

Cancer is a multifaceted array of disorders arising from genetic or epigenetic factors, driven by processes such as irregular gene expression, metabolic disturbances, and genetic and epigenetic aberrations. Dysfunctions at the cellular level, improper cell cycle progression, and disruptions in signal transduction contribute to cancer progression. Regulatory proteins, including cyclins, cyclin-dependent kinases (CDKs), and CDK inhibitor proteins, act as tumor suppressors by controlling various checkpoints in the cell cycle. Tumor suppressor genes, exemplified by p16INK4a encoded by CDKN2A on chromosome 9, play a pivotal role in restraining tumor growth. Compromised in tumors, these genes eliminate inhibitors of cell growth, fostering uncontrolled cell growth and tumorigenesis. Tumor cells share characteristics such as limitless proliferation, autonomous growth signaling, and resistance to anti-proliferative and apoptotic signals.

The p16INK4a protein, a silenced tumor suppressor in various cancers, is encoded by CDKN2A, a gene susceptible to harmful mutations on chromosome 9. Beyond melanomas, mutations in CDKN2A are implicated in pancreatic adenocarcinoma, gastric lymphoma, head and neck squamous cell carcinoma, prostate cancer, gastric, and colorectal cancer. The CDKN2A gene, spanning 8.5 kb, produces two tumor-suppressor proteins, p16INK4a and p14ARF, collaborating with other proteins to regulate the cell cycle. P16INK4a, with 156 amino acids and a molecular weight of 16 kDa, negatively regulates the cell cycle, halting improper cell division in response to stress conditions such as DNA damage or oncogenic signals, potentially inducing cell senescence.

This project takes a look at mutation data about single missense substitution mutations for CDKN2A gene and aims to predict cancer type based on substituitons and site of substitution involved.

Click here to find out the number of cancers and their types in the database by organs: Go to Organ Selection!

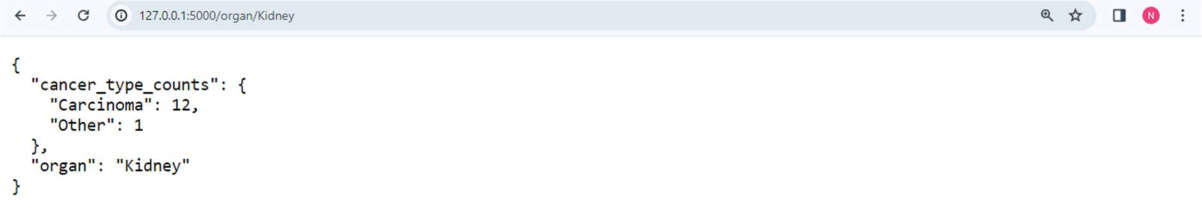
Click here to try out our cancer predictor: Go to Cancer Type Predictor!

Click here to check out the substitution matrices by alleles and by amino acids: Check out the substitution matrices!

Click here to analyse cancer types by their histopathology: Click me!

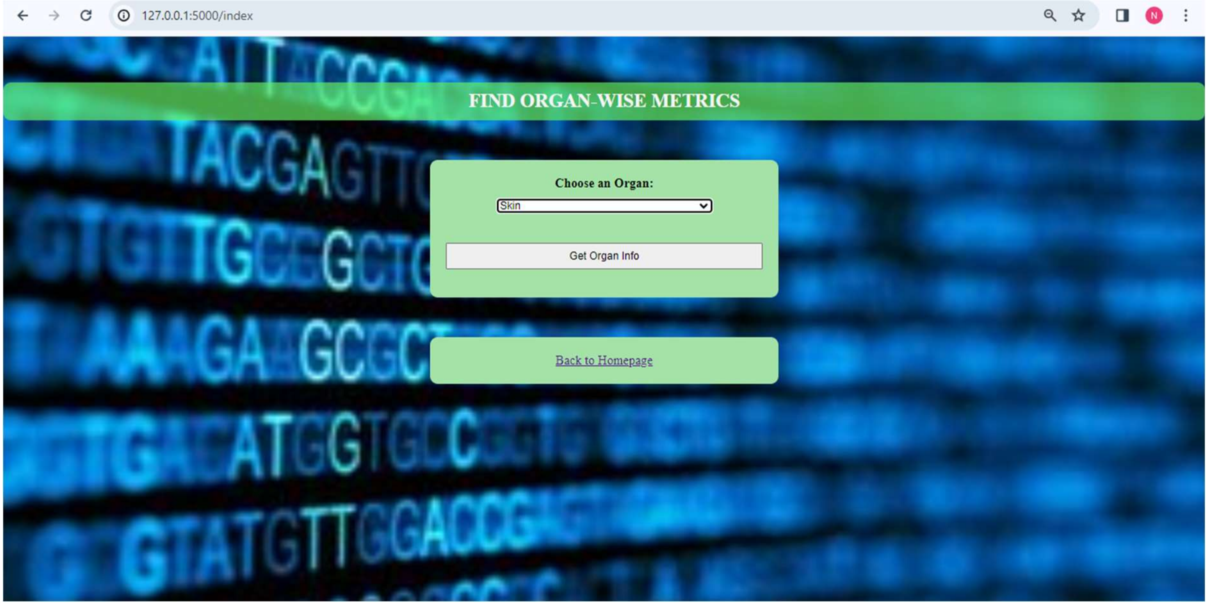
Access complete files and code on my Github! Click here!

I have also implemented a route which lets user view organ specific cancer counts via an API:



```
{
  "cancer_type_counts": {
    "Carcinoma": 12,
    "Other": 1
  },
  "organ": "Kidney"
}
```

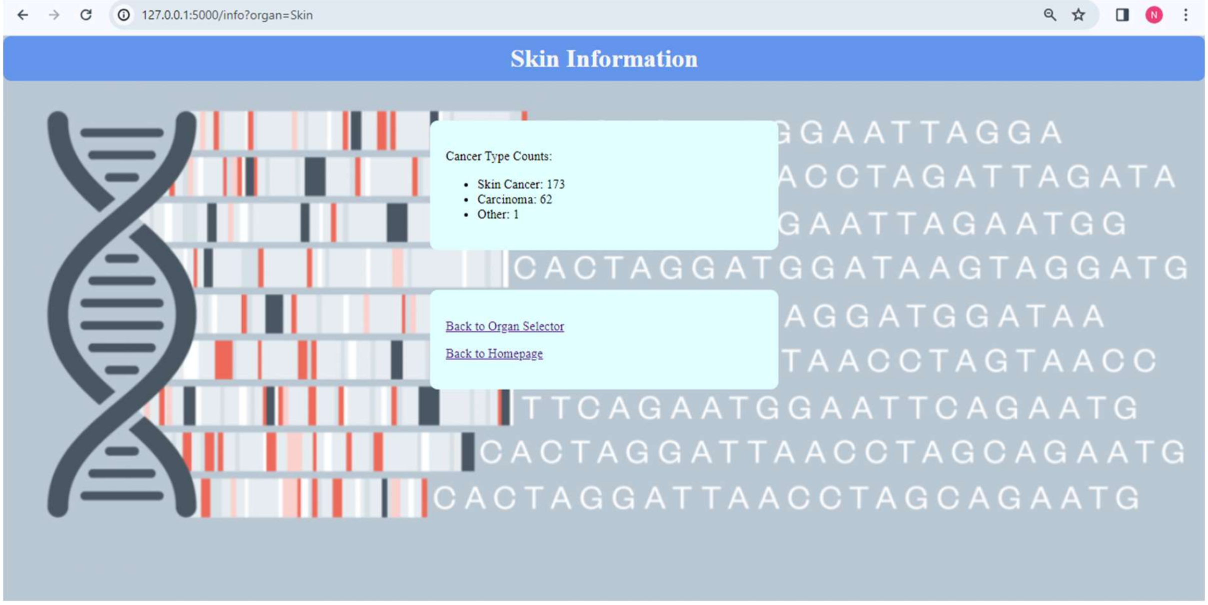
If we go to the Organ Selector from the hyperlink on the homepage, we encounter a form where we can select an organ from the drop-down list and click Get Organ Info button to find out specific cancer counts for that organ. Understandably, this functionality is doing the exact same thing as what I showed you right above this, but it’s using a GET query.
Also to be noted is the fact that all the information and results generated using all the functionalities and interactive elements of this website are very specific to CDKN2A, and specifically for missense substitution mutations which are somatic.



127.0.0.1:5000/index

FIND ORGAN-WISE METRICS

Choose an Organ: Skin



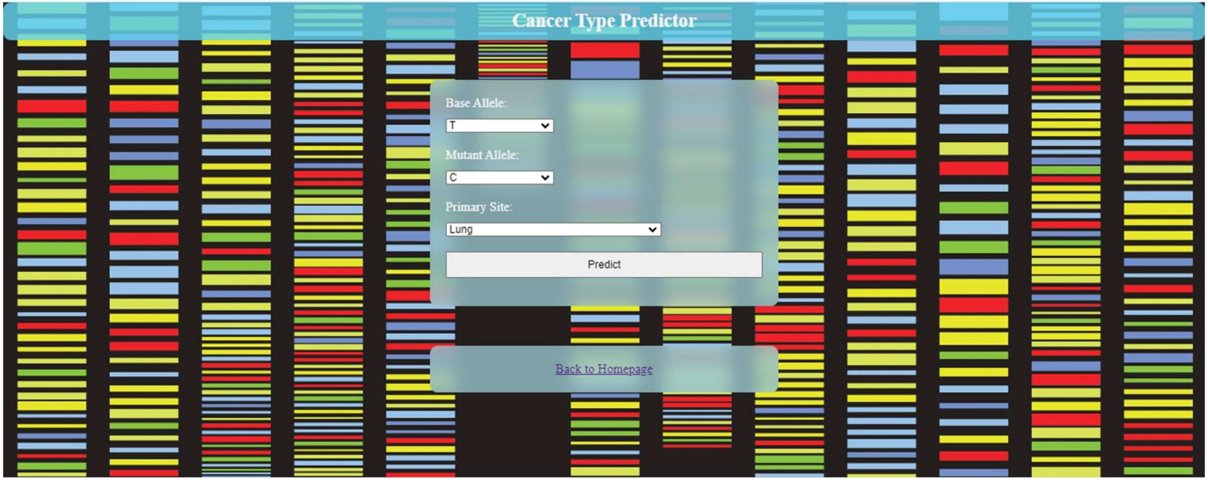
127.0.0.1:5000/info?organ=Skin

Skin Information

Cancer Type Counts:

- Skin Cancer: 173
- Carcinoma: 62
- Other: 1

Going to the cancer type predictor on homepage, it lets user enter base, mutant allele and site of cancer to predict cancer type outcomes.

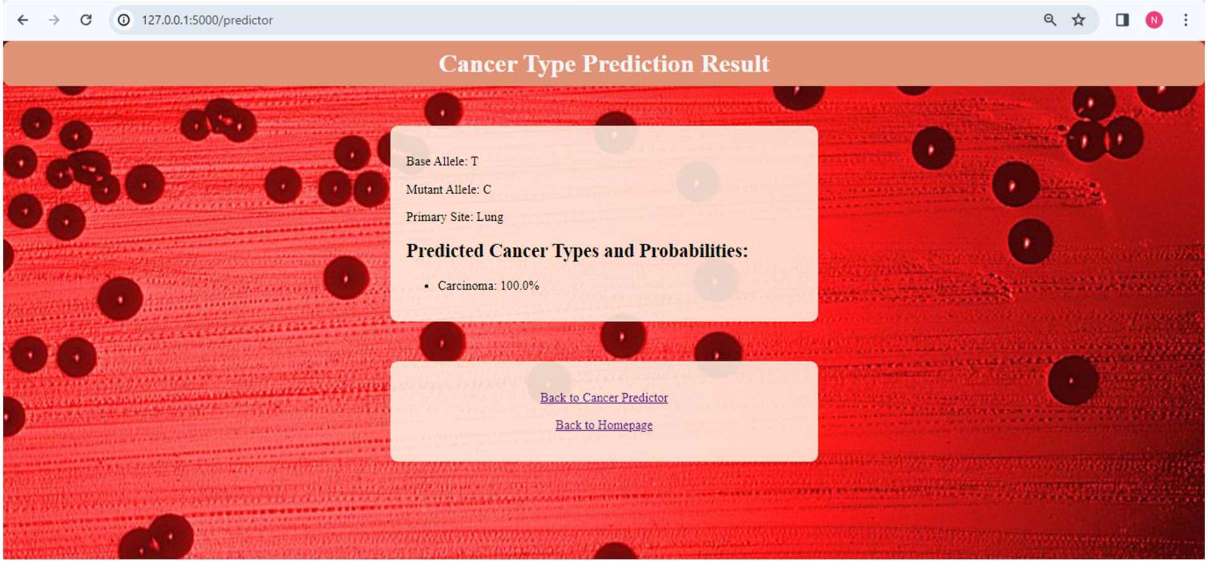


Cancer Type Predictor

Base Allele: T

Mutant Allele: C

Primary Site: Lung



Cancer Type Prediction Result

Base Allele: T

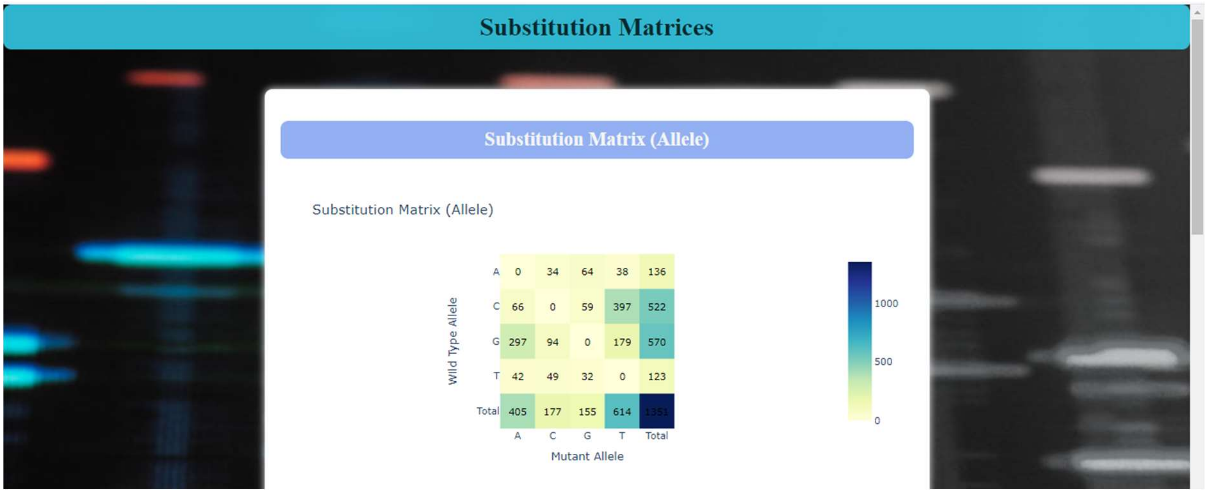
Mutant Allele: C

Primary Site: Lung

Predicted Cancer Types and Probabilities:

- Carcinoma: 100.0%

Finally, the next two options on homepage, substitution matrices and histopathology visualizations generate plots for nucleotide and amino acid substitution and cancer type count per tumour origin respectively.

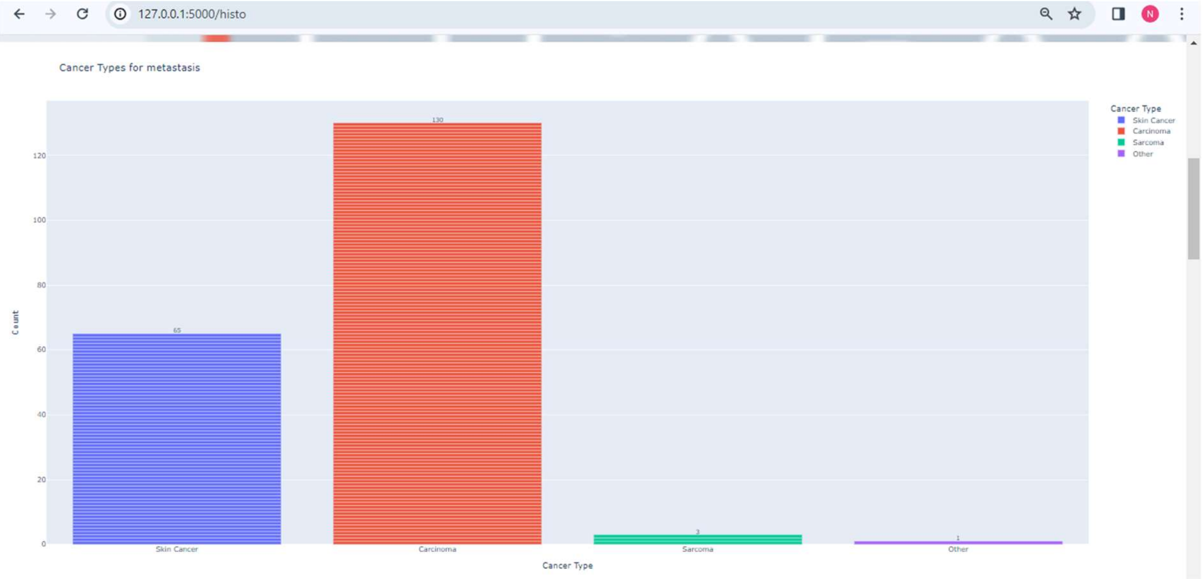


Substitution Matrices

Substitution Matrix (Allele)

Substitution Matrix (Allele)

Wild Type Allele \ Mutant Allele	A	C	G	T	Total
A	0	34	64	38	136
C	66	0	59	397	522
G	297	94	0	179	570
T	42	49	32	0	123
Total	405	177	155	614	1351



SCOPE AND LIMITATIONS

Scope:

This system when scaled can be used for:

- Personalized Treatment Strategies: To empower tailored medical interventions by predicting cancer types based on individual genomic profiles, integrating information on base and mutant alleles or base and mutant amino acids.

- Early Detection and Prognosis: To facilitate early cancer diagnosis and prognosis, enabling timely interventions and personalized treatment plans for improved patient outcomes.

- Precision Oncology: Enabling alignment with precision medicine principles, leveraging genetic data to refine cancer diagnoses, and optimizing the selection of targeted therapies.

- Integrated Data Analysis: To utilize bioinformatics tools to integrate diverse data sources, including genomic information and tumor location, providing a comprehensive understanding of cancer development.

- Public Health Impact: Can help support public health initiatives through a systematic and data-driven approach to cancer prediction, contributing to epidemiological research and informing preventive strategies.

Limitations:

- This system is not very easily scalable.
- It includes very specific data for a very specific type of gene for a very specific type of mutation. It needs to be more generalizable.
- Unsupervised methods are not as successful as supervised methods like XGBoost.
- More data is needed to make the system more efficient and robust, especially for rare cancers.

- **REFERENCES**

1. Zhao R, Choi BY, Lee M-H, Bode AM, Dong Z **(2016).** Implications of Genetic and Epigenetic Alterations of CDKN2A (p16INK4a) in Cancer. *EBioMedicine*, 8:30-39. https://doi.org/10.1016/j.ebiom.2016.04.017.

2. Tallen G., Riabowol K. **(2014)**. Keep-ING balance: tumor suppression by epigenetic regulation. *FEBS Lett*., 588:2728-2742

3. AlAjmi MF, Khan S, Choudhury A, Mohammad T, Noor S, Hussain A, Lu W, Eapen MS, Chimankar V, Hansbro PM, Sohal SS, Elasbali AM and Hassan MI. (**2021**). Impact of Deleterious Mutations on Structure, Function and Stability of Serum/ Glucocorticoid Regulated Kinase 1: A Gene to Diseases Correlation. *Front. Mol. Biosci*. 8:780284. doi: 10.3389/fmolb.2021.780284

4. Lu Y., Chan Y.T. Tan H.Y., Li S., Wang N., and Feng Y. (**2020**). Epigenetic Regulation in Human Cancer: the Potential Role of Epi-Drug in Cancer Therapy. *Mol. Cancer* 19, 79–16. doi:10.1186/s12943-020-01197-3

5. Sekido Y. (**2010**). Genomic Abnormalities and Signal Transduction Dysregulation in Malignant Mesothelioma Cells. *Cancer Sci*. 101, 1–6. doi:10.1111/j.1349-7006.2009.01336.x

6. Nabel EG. **(2002)**. CDKs and CKIs: molecular targets for tissue remodeling. *Nat. Rev. Drug Discov.* 1:587-598

7. Collins K, Jacks T, Pavletich NP **(1997)**. The cell cycle and cancer. *Proc. Natl. Acad. Sci. U. S. A.* 94:2776-2778

8. Hanahan D, Weinberg RA **(2011).** Hallmarks of cancer: the next generation. *Cell*, 144:646-674. https://doi.org/10.1016/j.cell.2011.02.013.

9. Aoude LG, Wadt KA, Pritchard AL, Hayward NK **(2015)**. Genetics of familial melanoma: 20 years after CDKN2A. *Pigment Cell & Melanoma Research*. 28 (2): 148–60. doi:10.1111/pcmr.12333

10. Hayward NK **(2003)**. Genetics of melanoma predisposition. *Oncogene*. 22 (20): 3053–62. doi:10.1038/sj.onc.1206445.

11. Jiao L, Zhu J, Hassan MM, Evans DB, Abbruzzese JL, Li D **(2007).** K-ras mutation and p16 and preproenkephalin promoter hypermethylation in plasma DNA of pancreatic cancer patients: in relation to cigarette smoking. *Pancreas*. 34(1):55–62. doi:10.1097/01.mpa.0000246665.68869.d4.

12. Huang Q, Su X, Ai L, Li M, Fan CY, Weiss LM **(2007).** Promoter hypermethylation of multiple genes in gastric lymphoma. *Leukemia & Lymphoma*. 48(10):1988–96. doi:10.1080/10428190701573224.

13. El-Naggar AK, Lai S, Clayman G, Lee JK, Luna MA, Goepfert H, Batsakis JG **(1997).** Methylation, a major mechanism of p16/CDKN2 gene inactivation in head and neck squamous carcinoma. *The American Journal of Pathology*. 151(6):1767–74.

14. Ameri A, Alidoosti A, Hosseini SY, Parvin M, Emranpour MH, Taslimi F, Salehi E, Fadavip P **(2011).** Prognostic Value of Promoter Hypermethylation of Retinoic Acid Receptor Beta (RARB) and CDKN2 (p16/MTS1) in Prostate Cancer. *Chinese Journal of Cancer Research = Chung-Kuo Yen Cheng Yen Chiu*. 23(4):306–11.

15. Rajendran P, Dashwood WM, Li L, Kang Y, Kim E, Johnson G, Fischer KA, Löhr CV, Williams DE, Ho E, Yamamoto M, Lieberman DA, Dashwood RH **(2015).** Nrf2 status affects tumor growth, HDAC3 gene promoter associations, and the response to sulforaphane in the colon. *Clinical Epigenetics*. 7:102.

16. Al-Kaabi A, van Bockel LW, Pothen AJ, and Willems SM. **(2014).** p16INK4A and p14ARF Gene Promoter Hypermethylation as Prognostic Biomarker in Oral and Oropharyngeal Squamous Cell Carcinoma: A Review. *Disease Markers.* https://doi.org/10.1155/2014/260549.

17. Takahashi A, Ohtani N, Yamakoshi K, Iida S, Tahara H, Nakayama K, Nakayama KI, Ide T, Saya H, Hara E **(2006).** Mitogenic signalling and the p16INK4a-Rb pathway cooperate to enforce irreversible cellular senescence. *Nature Cell Biology*. 8(11): 1291–7. doi:10.1038/ncb1491.

18. Witkiewicz AK, Knudsen KE, Dicker AP, Knudsen ES **(2011).** The meaning of p16(ink4a) expression in tumors: functional significance, clinical associations and future developments. *Cell Cycle*. 10(15): 2497–503. doi:10.4161/cc.10.15.16776

19. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, et al. **(2019).** COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. 47(D1): D941–D947. https://doi.org/10.1093/nar/gky1015.