**CPT\_S-475/575 Data Science Final Project Progress Report**

**Pei-Yau Weng 11766251, Denise Tanumihardja 11698042**

**Abstract-** Genetic mutations cause negative effects on lives in most cases. In humans, genetic mutations cause cancers that generate unhealthy tissue that disrupts physiological functions. Thus, locating key genes to find out the cause of cancer is an important issue in biology and medicine. In this paper, we propose a method to locate and evaluate the genes identified by our method. First, we indicate the related background of cancer gene networks and our motivation. Second, we address problems our method wants to solve. Third, we explain the workflow and algorithm of our method. Fourth, we show how our method can identify key genes, and these genes are compared to the biological authoritative database, Human Protein Atlas (HPA), to support the performance of our method. Fifth, we discuss the relationship between each target cancer and their key genes our method identifies and HPA database evaluates, and we also refer to some biological research that approves our results. Sixth, our method is compared to current methods to show the differences and the advantage of our method. Finally, we make a conclusion and indicate the importance and contribution of our method.

1. **Introduction**

Genetic mutations gives life negative attributes, such as breaking physiological functions, corrupting normal tissues, and immune system disorder. Cancers are especially the worst case of mutations for human body, so many medical and biological research are trying to figure out which group of genes cause cancers/tumors. However, the cause of cancers/tumors is sophisticated since the different combination of genetic mutations could cause different kinds of cancers/tumors. All scientists from different fields, geneticists, biologists, and data scientists, have cooperated for over 10 years to build the database of cancer gene and gene network to solve this problem. They use gene sequencing screen methods, medical analysis, computational algorithms, and recent machine learning techniques to find out the solution. As development of computers improved, scientists can collect and analyze more complicated information and utilize the method of machine learning and data science to filter out key information. As we mention, the mechanism and the cause of driving cancers/tumors are from the combination of genetic mutations. They precisely are cause by the series of malfunctions of genes and their produce, the protein generation system. Therefore, it is necessary to build a network of protein molecules to explore the pattern of human cancer. Inspired by this paper**[40]**, we built a sequential similarity network (SSN) to find the strongly related genes from donor samples, and then use these genes to construct protein-protein interaction network (PPIN) and filter out the key genes for a target cancer. In this paper, our method identifies the group of key genes and constructs their PPIN, to show that said key genes have a significant impact on the cancers/tumors and the proteins synthesized via gene translation, and how they interact with each other.

1. **Problem**

Many studies currently use Deep learning methods to identify key genes for certain cancers. Most of them build a classifier model**[55]** to recognize key genes in a DNA sequence. Whereas, as we mentioned previously, the cause of driving cancers/tumors are from the combination of genetic mutations and the disorder of protein interaction. The problem is that quantifying accuracy of the recognition is hard to apply on cancer gene identification because it only considers the number of recognized genes matched labels, but the driver of a cancer/tumor in fact is the interaction of a group of mutated genes and their translated proteins. In most medical cases, a part of genetic mutation is able to cause a cancer and there is not a standard quantification of the genetic mutation to cause a cancer, so only using a classifier model trained by quantized labels is not an appropriate method. Because of this problem and challenge, we use SNN to find the similar DNA sequences and strongly related genes in the dataset, and then construct PPIN by these genes to identify the key genes for a certain cancer.

1. **Models/Algorithms**

Our proposed method consists of four computational steps and one evaluation. 1) The Directed Weighted All Nearest Neighbors (DiWANN) network**[40]**, to construct an SSN. In the network, each node symbolizes a donor and his information including a cancer, mutated genes, and a DNA sequence 2) For a small SNN, we search the nodes that have the maximumber of the betweenness centrality and the maximal outgoing degree. For a complex SNN, we use the Girvan-Newman (GN) algorithm to split a network into several communities and search the nodes in the largest community and the node with the maximal outgoing degree. 3) We use the Deep First Search method to find all neighborhood nodes, and then collect mutated genes from each node (donor). The mutated genes we collect here are called strongly related genes for a certain cancer. 4) We construct PPIN by using StringDB**[56]** in R to identify the key genes among the strongly related genes. Finally, we evaluate our result by the Human Protein Atlas (HPA) database to show if these identified key genes have high expressions in a certain cancer. We describe the four computational steps and the evaluation by HPA database.

1. **DiWANN network**

The DiWANN network**[40]** is a new kind of SSN that is constructed by the edit distance algorithm**[57]** and is without setting thresholds. It produces an distance matrix by inputting a set of n sequences, only calculate the distances in the half of the upper diagonal matrix and construct a network by the matrix.

1. **Maximal Outgoing Degree, Betweenness Centrality & GN Algorithm**

Constructing SSN by DiWANN is the first step of our method to make connections between the closest DNA sequences. We subsequently find the center nodes that have maximal outgoing center degree, which means that the center node has the largest number of neighbors. The equation of the center degree is given:

(1)

In addition, if we only find the neighbors of the center nodes, the small number of samples cause a bias and is not enough to build a complete PPIN. Thus, we also consider the betweenness centrality of each node and select the nodes that have the maximal betweenness. The equation of the betweenness centrality is given:

(2)

For some SSNs that have divergent structures, we use GN algorithm**[58]** to split an SSN into several communities and select all nodes in the largest community.

1. **DFS**

After finding out all center nodes and maximal betweenness nodes, we let these nodes as the roots and use DFS method to search their neighbors that have outgoing directed edge from the roots.

1. **StringDB-based PPIN**

We now find out all neighbors in SSN and subsequently extract their tumor genes corresponding to the neighborhood nodes (donor ID) in the dataset. The StringDB-based PPIN is a function from the R library, StringDB. We input collected genes, which we call strongly related genes, generated from previous three steps to this function, and it produces a PPIN by the StringDB database. Finally, we reorganize the PPIN to find the key genes.

1. **HPA Evaluation**

HPA database is from an authoritative website**[1]** in biology. It provides an R library, HPAanalyze, to let users easily input genes and assign a cancer or a set of human tissues, and HPAanalyze will give the magnitude of expressions for each gene corresponding a cancer and a tissue.

1. **The pseudocode of our method**

|  |
| --- |
| **Algorithm:** D is the processed dataset for a certain cancer, is total DNA sequences in D, is the total number of nodes (donor IDs), M is minEditDistanceMatrix, is the total nodes in SSN. B is the betweenness values of all nodes, C is the degree values of all nodes, is the total number of maximal betweenness nodes, is the total number of maximal degree node (center nodes). is the total number of nodes in the largest community. is the total number of neighborhood nodes, is the genes corresponding to the , and is the total key genes. The procedure is to make M, and the procedure is to construct an SSN by M. is a result of the evaluation (a graphical data) |
| 1. Extract and from D 2. M 3. SSN 4. B by eq 2. 5. C by eq 1. 6. **if** () > 0.3 **then:**   N(   1. **else:**     N(   1. PPIN //process in R 2. PPIN //process in R 3. //process in R |

1. **Implementation/Analysis**

In this section, we describe the details of our experiment including the dataset we use, our hypothesis, and its proof by illustrations. Finally, we list the overall workflow of our method and explain the evaluation by HPAanalyze. In the next section, we show the graphical result of our method.

1. **The dataset for our analysis**

The dataset is from ICGC**[59]**. There are three csv files, donor\_art\_seq\_468, icgc\_468\_gene\_tissue, and icgc\_data\_468 in the dataset, and it includes 15 cancers: breast cancer, prostate cancer, esophagus cancer, stomach cancer, head cancer, biliary cancer, skin cancer, pancreatic cancer, lymph cancer, liver cancer, ovary cancer, kidney cancer, myeloid cancer, CNS cancer, and bone cancer.

* **donor\_art\_seq\_468.csv**: The sequence here is a normalized artificial sequence created by combining the mutations for all 28 genes.
* **icgc\_468\_gene\_tissue.csv**: It provides information on the types of mutated genes and the types of cancer each donor has.
* **Icgc\_data\_468.csv**: It lists all the mutated genes in this dataset. Each mutated gene provides the mutated position, variant type, reference sequence, the corresponding chromosome, and the donor.

In our experiment, we combine the data of the CNS, myeloid, and bone cancer together for analysis since each of them drive tumors in the same tissues. We selected ten cancers excluding the esophagus cancer and lymph cancer since HPA database does not have the data about the esophagus cancer and using only a few samples of the lymph cancer in the dataset could cause biases.

1. **Our Hypothesis**

Our hypothesis includes three aspects. 1) We assume that the center degree nodes and maximal betweenness nodes in DiWANN network have nodes shared genetic mutations among all nodes. 2) GN algorithm applied on divergent SSNs has a better performance to find strongly related nodes than finding maximal betweenness nodes by eq 2. 3) We assume that we can use the DFS method to find the neighbors, and that we can input them to construct a worthful PPIN. 4) The organized PPIN more easily extracts the key genes than the original PPIN.

In order to prove our hypothesis, we later describe how DiWANN network construct connections between the closest nodes (DNA sequences and donor ID) so that the center degree nodes and maximal betweenness nodes shared genetic mutations among all nodes, which indicates that the GN algorithm performs better on the network by having a divergent structure. This explains why we would use DFS to find neighbors, and how we can find the key genes by a reorganized PPIN.

1. **How DiWANN method construct an SSN**

The algorithm of DiWANN network includes two parts, making a distance matrix and constructing a directed graph by the matrix. For making distance matrix, there are four steps. First, it initializes each entry to infinity and then computes all values of edit distances in the first row as the baseline of the pruning optimization. Second, it subsequently calculates distance row by row and records the minimal and maximal distance of each row. Third, it sets up the lower and upper bound based on the triangle inequality and the minimum and maximum in each row for pruning. Finally, it records the edit distance that is either smaller than the lowest maximum in all rows or smaller than the minimum of the current row. For a directed graph construction, each column name is the start node, and each row name is the end node. It selects the minimal value corresponding to each row and column to be the weight of a directed edge and the end node connects back to the start node if the weight is the minimum in the whole row. We take a simple example to illustrate the graph construction. The illustration is as follows:

A picture containing diagram

Description automatically generated

Fig 1. The illustration of the graph construction of DiWANN

A and B have bidirected edge because the column B to the row A is 2 and the minimal value is 2 in the whole row A. B and C have bidirected edge because the same reason as the edge between A bad B. C and A have a one directed edge because the column C to the row A is 3, but the value 3 is not the minimum in the row A. Therefore, DiWANN network make connections between the closest nodes so that the center degree nodes and maximal betweenness nodes shared genetic mutations among all nodes

1. **Apply GN algorithm on the divergent network**

The reason why we apply GN algorithm is because the neighbors of the center nodes and the nodes having maximal betweenness are less than 30 percent of all nodes in the network. As we mentioned in section 3.2, an insufficient number of samples causes bias and so it cannot build a meaningful PPIN. Taking an SSN of the skin cancer in our dataset as an example, the following Fig 2 shows the comparison of using betweenness and GN algorithm.

|  |  |
| --- | --- |
| Chart, radar chart  Description automatically generated | Chart, radar chart  Description automatically generated |
| **Finding neighbors (red nodes) by the betweenness** | **Finding neighbors (red nodes) by the GN algorithm** |
| **Chart, bubble chart  Description automatically generated** |  |
| **PPIN without using GN algorithm** | **PPIN with using GN algorithm** |

Fig 2. The comparison of the betweenness and GN algorithm

In Fig 2, the neighbors of the center nodes and the nodes having maxiaml betweeenness are much less than the neighbors found by center nodes and GN algorithm. When we input the collected genes from the neighbors, PPIN is not feasible to extract key genes if GN algorithm is not applied to find the neighbors. Therefore, using GN algorithm to find the neighbors in the divergent network is necessary.

1. **Why using DFS to find neighbors**

DiWANN network has made connections between the closest nodes (DNA sequences and donor ID), while some of nodes do not provide strongly related information for a certain cancer to later construct a PPIN. We think the nodes that have outgoing directed connections are strongly related to their neighbors. Thus, we use the DFS method to find the neighbors and the complexity of SSN can be reduced to efficiently construct PPIN later. On the other hand, if we directly select the nodes close to the the center nodes and the maxiaml betweeenness nodes, the number of sampled nodes is insufficient, which is the same reason as why we use GN algorithm. The following figure shows that the number of sampled nodes is not enough to construct a complete PPIN if not use DFS.

|  |  |
| --- | --- |
| Chart, radar chart  Description automatically generated | Chart, radar chart  Description automatically generated |
| **Finding neighbors (red nodes) without using DFS** | **Finding neighbors (red nodes) by DFS** |
| **Diagram  Description automatically generated** | **A picture containing transport  Description automatically generated** |
| **PPIN without DFS** | **PPIN with DFS** |

Fig 3. The comparison of finding neighbors with and without DFS

In Fig. 3 we take skin cancer as an example, SMARCA4, CTNNB1, CASP8, and PIK3CA are missing, while all of them are the key genes and have high expression in the skin cancer (we will show those genes are important for the skin cancer in section 5). Therefore, Finding the neighbors by DFS is necessary for our method when especially handling complicated SSNs.

1. **How to find the key genes by a reorganized PPIN**

The edge in the PPIN includes seven parts, experimental data, data from biological paper mining, data from authoritative databases, gene adjacency, gene fusion, and gene co-expression. The function has a scoring mechanism to give a score (weight) to an edge (the full score is a thousand), so the more edges and more score between two nodes (genes) are, the stronger relation two nodes have. However, StringDB-based PPIN looks too complex and cannot be directly analyzed, so we reorganized the PPIN, and mark the strongly related edges, where scores greater than 850 are red lines, and label the strong related edge with the score. The following figure illustrates this computational step.

Diagram

Description automatically generated

Fig 3. The illustration of PPIN reorganization

It is clear to research those genes, which have high score and red edge, in the reorganized PPIN, and we name those genes as “key genes.” In Fig 3, there are 9 key genes, smarca4, grin2a, nras, nf1, fbxw7, atm, egfr, tp53, and keap1. In the evaluation phase, we input those key genes to HPAanalyze, to evaluate our method.

1. **Evaluation by HPA database**

We use R library, HPAanalyze, to evaluate the key genes obtained by our method. It has databases of 20 cancers. We can input the key genes and assign a specific cancer to its function, hpaVis, and it shows the patient's proportions of the genetic expression. If HPAanalyze does not provide the data of the cancer we want to analyze, we can use the function, hpaVisTissue. We input the key genes and desired human tissue/cell to this function, and it shows the genetic expression in the desired tissue/cell.

1. **The overall workflow of our method**
2. Combine donor\_art\_seq\_468.csv file and icgc\_468\_gene\_tissue.csv files by donator ID and separate the artificial normalized sequences and their donors by different kind of cancers.
3. Input the group of the artificial sequences attributed to a certain cancer into the DiWANN method, label each node in the network with the corresponding donor ID, and then find both the node with maximal degree and the node with highest betweenness centrality.
4. Using the DFS method, search neighborhood nodes that have bi-directed edges with these two points and their neighbors, and subsequently record their labels, which are the donor IDs.
5. If the number of neighbors is less than 30 percent of the total, we use the GN algorithm to find the nodes in the largest community and return to step 2.
6. Sort by donor IDs, collect all key tumor genes and count the tally for each. Then store them as a csv file (each file stores for each kind of cancer)
7. Import the R library, stringDB, and input the key genes from a csv file to construct a PPIN for a certain cancer.
8. A PPIN stores the combined score on the edge between any two proteins. The higher the combined score is, the more highly the two proteins interact.
9. Reorganize PPI and Mark the edges whose scores are greater than 850 (full score is 1000) as red, and the others as blue.
10. Extract the genes that have red edge as the key genes
11. Import the R library, HPAanalyze, and input the key gens to analyze and evaluate each one that has significant expression in human tissue/cell or for certain cancers.
12. **Results and Discussion**

There are 10 cancers we have analyzed. We discuss the breast cancer, liver cancer, stomach cancer, pancreatic cancer, and skin cancer. The results of the other 5 cancers, the prostate cancer, head cancer, biliary cancer, ovary cancer, and CNS & bone & myeloid cancer, are shown in Appendix. For each cancer, we showed the graphical result of the SSN, where all neighbors are marked as red nodes, the strongly related genes that collect by all neighbors, the original PPIN from stingDB, the reorganized PPIN, and the evaluation of the genetic expression by HPAanalyze. To support our results and evaluations, we dicuss the similarities and differences in our output against what is already known about these cancer types and their relationships between each key gene associated with it.

* 1. **Breast Cancer (Breast Adenocarcinoma)**

The graphical result

|  |  |
| --- | --- |
| Chart, radar chart, polygon  Description automatically generated  **SSN** | Table  Description automatically generated  **The strongly related genes** |
| A picture containing bubble chart  Description automatically generated  **The original PPIN** | Diagram  Description automatically generated  **Reorganized PPIN** |

The evaluation by HPAanalyze

A picture containing background pattern

Description automatically generated

Our method identified TP53, RB1, PTEN, PIK3CA, and EGFR as the key genes responsible for breast cancer. According to the Human Protein Atlas, in regard to the breasts itself (specifically the adipocytes, glandular and myoepithelial cells), these key genes have a range of expressions; TP53 is not expressed at all in either cell types; RB1 is partially highly expressed in glandular cells and low expressed in adipocytes and myoepithelial cells; PTEN is not expressed in adipocytes but has low expression in glandular and myoepithelial cells; PIK3CA is partially highly expressed in glandular and myoepithelial cells and low expressed in adipocytes; EGFR has low expression in either cell types.**[1]** In terms of cancer:

* Mutation (inactivation) in the TP53 gene is a factor in breast cancer development as it is considered as a tumor suppressor gene which not only synthesizes P53 protein that is in charge of the cell cycle, checkpoint control, repair of DNA, and apoptosis, but also the loss of heterozygosity or allelic imbalance in sporadic breast cancer**[17]**
* Mutation (inactivation) in the RB1 gene is a factor in breast cancer development as it is considered as a tumor suppressor gene and it increases the translation of mitochondrial proteins as well as oxidative phosphorylation, which cancer stem cells use as an energy source instead of glycolysis**[13]**
* Mutation (low expression) of the PTEN gene is a factor in breast cancer development as it is considered as a tumor suppressor gene and the main mechanism may be due to PTEN promoter methylation**[11]**
* Mutation (amplification) in the PIK3CA gene is a factor in breast cancer development as it encodes a catalytic subunit of PI3K and its mutation (clustering within the helical and kinase domain) found highly frequently in breast cancers**[9]** as well as the activation of said subunit, resulting in increased lipid kinase activity and AKT activation**[18]**
* Mutation (overexpression) of the EGFR gene is only partially a factor in breast cancer development as it activates several major signaling pathways, resulting in several biological functions, and there is evidence that there are significant interactions with other receptor tyrosine kinases**[5]**.
  1. **Liver Cancer (Hepatocellular Carcinoma)**

Our method identified GRIN2A, CDKN2A, KRAS, PIK3CA, TP53, CTNNB1, NF1, and NRAS as the key genes responsible for liver cancer. According to the Human Protein Atlas, in regard to the liver itself (specifically the cholangiocytes and hepatocytes), these key genes have a range of expressions; CDKN2A is not expressed hepatocytes but is low expressed in cholangiocytes; PIK3CA is partially medium expressed in hepatocytes and low expressed in cholangiocytes; CTNNB1 is highly expressed in both cell types; NF1 is not expressed in cholangiocytes but highly expressed in hepatocytes; GRIN2A, KRAS, TP53, and NRAS are not expressed at all in either cell types.**[1]** In terms of cancer:

* Mutation () in the GRIN2A gene is a factor in liver cancer development as \_**[27]**
* Mutations (deletion) in the CDKN2A gene is a factor in liver cancer development as it is a part of the cell cycle signaling pathway and is considered as a tumor suppressor gene which promotes cell cycle arrest in G1 and G2 phases, as well as suppresses MDM2**[4]**
* Mutations (activation) in the KRAS gene is a factor in liver cancer development as it exerts distinct cellular and molecular responses dependent on cell types, and both its activation along with HBV X protein expression, promoted the initiation and progression of the cancer**[28]**
* Mutations (frameshift) in the PIK3CA gene is a factor in liver cancer development as it encodes a catalytic subunit of PI3K, and this mutation results in the change in the last C-terminal amino acid of the gene and creates three additional amino acids**[29]**
* Mutation (somatic mutation) in the TP53 gene is a factor in liver cancer development as it is considered as a tumor suppressor gene and it activates the WNT signal transduction pathway as well as activation of AFB1, which induces transversions in the gene codon alongside HBV to cause a mutations in P53 protein**[30]**
* Mutations (activation) in the CTNNB1 gene is a factor in liver cancer development as it is a part of the WNT signaling pathway and controls cell adhesion, growth and differentiation**[3]**
* Mutations (knockdown) in the NF1 gene is a factor in liver cancer development as it is involved in targetable pathways and leads to the formation of cancer cells with constitutive activation of MAPK/ERK signaling**[31]**
* Mutations (overexpression) in the NRAS gene is a factor in liver cancer development as it is involved in regulating cell division[32], and it along with the P53 protein have important roles in the carcinogenesis and maintenance of the cancer**[33]**.
  1. **Stomach Cancer (Stomach Adenocarcinoma)**

The graphical result

|  |  |
| --- | --- |
| Chart, diagram, bubble chart  Description automatically generated  **SSN** | Table  Description automatically generated with medium confidenceTable  Description automatically generated  **The strongly related genes** |
| A picture containing transport  Description automatically generated  **The original PPIN** | Chart  Description automatically generated  **Reorganized PPIN** |

The evaluation by HPAanalyze

Chart, bar chart

Description automatically generated

Our method identified GRIN2A, SMARCA4, KRAS, PIK3CA, RB1, TP53, EGFR, ATM, and BRAF as the key genes responsible for stomach cancer. According to the Human Protein Atlas, in regard to the stomach itself (specifically the glandular cells), these key genes have a range of expressions; SMARCA4 is highly expressed in the glandular cells; KRAS has medium to low expression in the glandular cells; PIK3CA and RB1 has medium expression in the glandular cells; EGFR is low expressed in the glandular cells; ATM is highly expressed in the glandular cells; BRAF has high to medium expression in the glandular cells; GRIN2A and TP53 are not expressed at all.**[1]** In terms of cancer:

* Mutation (aberrant expression) in the GRIN2A gene is a factor in stomach cancer development as it encodes for a protein subunit of the NMDA receptor, and potentially a relation to lower CLDN18.2 expression resulting in cancer**[34]**
* Mutation (aberrant expression) in the SMARCA4 gene is a factor in stomach cancer development as it is a subunit of the SWI/SNF complex, as such its mutation induces EMT markers as well as promoting tumor progression and lymph node metastasis**[35]**
* Mutation (activation) in the KRAS gene is a factor in stomach cancer development as it has an important role in regulating normal signal transduction as well as binding to GTP and GDT with high affinity, alongside being part of the MAP kinase signaling pathway, and it results in higher BRAF kinase and ERK1/2 phosphorylation activities**[36]**
* Mutation (up-regulation) in the PIK3CA gene is a factor in stomach cancer development as it encodes a catalytic subunit of PI3K, and induces putative cell cycle-regulated and mapped genes to various functions including diverse cellular processes, cell cycle regulation, cell proliferation and DNA replication**[10]**
* Mutation (loss of function) in the RB1 gene is a factor in stomach cancer development as it is considered as a tumor suppressor gene as well as a regulator to control G1/S transition during cell cycle progression via E2F transcription factor family protein interaction**[14]**, and when it is phosphorylated, transcription factors are released, inducing G1 to S phase in the cell, and resulting in the excessive proliferation of the cell**[37]**
* Mutation (loss of heterozygosity) in the TP53 gene is a factor in stomach cancer development as it is considered as a tumor suppressor gene which not only synthesizes P53 protein that is in charge of the cell cycle, checkpoint control, repair of DNA, and apoptosis, but also loss of function of P35 protein resulting in cancer**[38]**
* Mutation (overexpression) in the EGFR gene is a factor in stomach cancer development as it plays an important role in the regulation of cell growth and proliferation and when mutated, activates downstream molecules abnormally which in turn increases cancer cell proliferation and decreases apoptosis**[6]**
* Mutation (deletion) in the ATM gene is a factor of stomach cancer as it plays a role in cell cycle control as well as cancer development, and truncates encoded ATM proteins**[39]**
* Mutation (various) in the BRAF gene is a factor in stomach cancer development as it encodes serine/threonine kinases which are regulated by binding to RAS, as well as exciting the MAP kinase signaling pathway, and it results in higher BRAF kinase and ERK1/2 phosphorylation activities**[36]**.
  1. **Pancreatic Cancer (Pancreatic Adenocarcinoma)**

The graphical result

|  |  |
| --- | --- |
| Chart, radar chart  Description automatically generated  **SSN** | Table  Description automatically generatedTable  Description automatically generated  **The strongly related genes** |
| Chart, diagram  Description automatically generated  **The original PPIN** | Chart, diagram  Description automatically generated  **Reorganized PPIN** |

The evaluation by HPAanalyze

**Chart, bar chart

Description automatically generated**

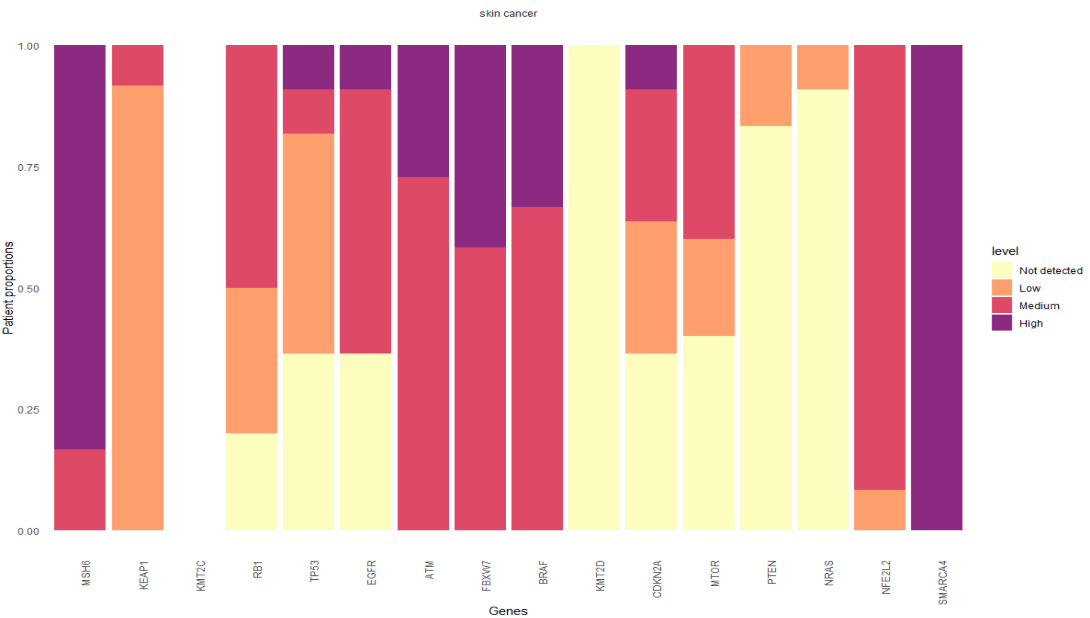
Our method identified CTNNB1, CASP8, NF1 and SMARCA4 as the key genes responsible for head and neck cancer. According to the Human Protein Atlas, in regard to the ovaries itself (specifically the exocrine glandular and pancreatic endocrine cells), these key genes have a range of expression; CTNNB1 is highly expressed in the exocrine glandular cells but has medium expression in the pancreatic endocrine cells; CASP8 is medium expressed in both the exocrine glandular and pancreatic endocrine cells; NF1 is highly expressed in the exocrine glandular cells; SMARCA4 is highly expressed in both the exocrine glandular and pancreatic endocrine cells**[1]**. In terms of cancer:

* Mutation (overexpression) in the CTNNB1 gene is a factor in pancreatic cancer development as it produces proteins found in the junctions between adjacent cells**[53]**, and the MIR454 gene upregulates CTNNB1, resulting in the TSPAN1 gene activation, which proliferates cancer.**[54]**
* Mutation (deletion) in the CASP8 gene is a factor in pancreatic cancer development as it has a role in triggering apoptosis in cells, and decreases RNA expression in lymphocytes as well as decreasing genetic activity and apoptosis of T lymphocytes.**[46]**
* Mutation (deletion) in the NF1 gene is a factor in pancreatic cancer development as it is considered as a tumor suppressor gene, and mutations result in a loss of growth control of the cells as well as enhanced cellular proliferation.**[45]**
* Mutation (deletion) in the SMARCA4 gene is a factor in pancreatic cancer development as it is a subunit of the SWI/SNF complex, and the mutation results in altered structures of the biochemical subunits of the complex.**[47]**
  1. **Skin Cancer (Melanoma)**

The graphical result

|  |  |
| --- | --- |
| Chart, radar chart  Description automatically generated  **SSN** | Table  Description automatically generatedTable  Description automatically generated  **The strongly related genes** |
| A picture containing transport  Description automatically generated  **The original PPIN** | A picture containing text, sky, map  Description automatically generated  **Reorganized PPIN** |

The evaluation by HPAanalyze

****

Our method identified MSH6, ATM, FBXW7, BRAF, and SMARCA4 as the key genes responsible for head and neck cancer. According to the Human Protein Atlas, in regard to the ovaries itself (specifically the cells in the basal, corneal, glandular and spinous layers, endothelial cells, fibrohistiocytic cells, fibroblasts, keratinocytes, langerhans, lymphocytes, melanocytes, epidermal cells, eccrine glands, endothelial cells, hair follicles, sebaceous cells, and vascular mural cells), these key genes have a range of expression; MSH6 is highly expressed in the basal layer cells, eccrine glands, endothelial cells, fibrohistiocytic cells, hair follicles, lymphocytes, melanocytes, sebaceous glands, and vascular mural cells, but medium expressed in the spinous layer and langerhans cells, and low expressed in the granular layer; ATM is medium expressed in the keratinocytes but low expressed in the fibroblasts and langerhans; FBXW7 is highly expressed in the basal layer cells, endothelial cells, fibrohistiocytic and langerhans cells, lymphocytes and melanocytes, but has medium expression in the granular and spinous layer cells; BRAF is medium expressed in the keratinocytes, langerhans and melanocytes; SMARCA4 is highly expressed in the keratinocytes and langerhans but medium expressed in the fibroblasts and melanocytes**[1]**. In terms of cancer:

* Mutation (overexpression) in the MSH6 gene is a factor in skin cancer development as it has a role in DNA repair response and pathways, and mutation results in the protection and repair of tumors.**[48]**
* Mutation (loss of function) in the ATM gene is a factor in skin cancer development as it has a role in maintaining genomic integrity, as well as has a central role in the repair of DNA double-strand breaks, and the mutation results in the inability to stabilize cell cycle arrest.**[49]**
* Mutation (inactivation) in the FBXW7 gene is a factor in skin cancer development as it has an important role in managing cell division, growth and differentiation, as well as a tumor suppressor gene, and the mutation results in the activation of NOTCH1, and in turn resulted in tumor angiogenesis promotion.**[50]**
* Mutation (amplification) in the BRAF gene is a factor in skin cancer development as it has a role in apoptosis prevention as well as the progression of the cell cycle, and as such, mutation results in the maintenance of the tumor.**[51]**
* Mutation (inactivation) in the SMARCA4 gene is a factor in skin cancer development as it is a subunit of the SWI/SNF complex which regulates gene expression, and as such mutations result in the loss of protein expression leading to cancer.**[52]**
  1. **Our hypothesis VS the result**

We briefly discussed our hypothesis and showed some simple tests to prove it in section 4. Most of the cancer analysis supported our hypothesis, and our method has good performance to find the key genes in these cancers. However, For the stomach cancer and CNS & bone & myeloid cancer (see in Appendix), our method makes all nodes be neighbors. It happens because the betweenness values of all nodes are zero, so our method identifies all nodes as the maximal betweenness nodes. Although this problem happens, it does not impact the key gene extraction in the next step. It implies that the donor samples of these two cancers in our dataset is not enough to construct a usable SSN and make a bias when searching neighbors. To solve this problem, we will adopt a larger dataset in the future or use the other state-of-the-art method to reduce the bias from small number of samples.

1. **Related Work**

There are two significant researches for cancer gene networks by using SSNs. One is MUFFIN (Mutations For Functional Impact on Network Neighbours), which accounts for not only individual genes, but neighbouring genes as well in functional networks, to prioritize cancer genes. This is done by selecting genes with more than just neighbour, to prevent false positives. Then, there are two ways of using the mutation information of direct neighbours: Use the information of all the direct neighbours in the network, or just use the information of the direct neighbour with the highest number of mutations.**[41]** The other is WGCNA (Weighted Gene Co-expression Network Analysis), which was used to predict phenotypical characteristics of Endometrial Cancers via their cancer hub gene signatures. This is done by converting gene expression profiles into connection weights, visualized as a topology overlap measure (TOM), which are then used with the topological overlap dissimilarity measure (1-TOM) as a distance measure, to define modules using the hierarchical cluster method with a heigh limit. Then, cancer hub genes which contribute to the growth of ECs are determined, using WGCNA via calculating its scaled connectivity (K) and genes significance (GS). Next, is linking the hubs to the EC phenotypes, by using elastic-net analysis.**[42]**

Compared to the two methods above, we utilized the DiWANN network where each node is a sequence and is connected via an edge to the next closest node(s)**[40]**. This reduces the complexity of the network, so we can efficiently construct an SSN and obtain the neighbors. Moreover, by introducing a PPIN in our methods and translating an SSN to it using the neighbors, as the cause of driving cancers/tumors are from the combination of genetic mutations and the disorder of protein interaction. More scientists use protein molecule networks to explore the causes of diseases because the molecule networks can explain the diseases at the systematic level. The PPIN is a kind of protein molecule network, so we can clarify the driven of cancers/tumors at the systemic level and precisely extract the key genes for a certain cancer.

1. **Conclusion**

Our method uses the DiWANN network to efficiently construct a SNN and find the neighbors among it, and PPIN lets us precisely locate the key genes. In the HPAanalyze evaluation, the key genes our method extract have high and medium expression in most of cancer we analyze, while our method has bad performances in a few cases. Therefore, we will further analyze the reason for those bad performances. This problem might be from the dataset or from the method of finding neighbors in SSN. In the future, we will test our method with a larger dataset and use a state-of-the-art method to find neighbors.

**References**

**[1]** “The Human Protein Atlas.” The Human Protein Atlas, [www.proteinatlas.org/](http://www.proteinatlas.org/).

**[2]** “Head and Neck Cancers.” *National Cancer Institute*, [www.cancer.gov/types/head-and-neck/head-neck-fact-sheet](http://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet).

**[3]** Khemlina, Galina, et al. “The Biology of Hepatocellular Carcinoma: Implications for Genomic and Immune Therapies.” *Molecular Cancer*, vol. 16, no. 1, 2017, doi:10.1186/s12943-017-0712-x.

**[4]** Cabanillas, Rubén, et al. “Novel Germline*cdkn2a*Mutation Associated with Head and Neck Squamous Cell Carcinomas and Melanomas.” *Head & Neck*, vol. 35, no. 3, 2011, doi:10.1002/hed.21911.

**[5]** Masuda, Hiroko, et al. “Role of Epidermal Growth Factor Receptor in Breast Cancer.” *Breast Cancer Research and Treatment*, vol. 136, no. 2, 2012, pp. 331–345., doi:10.1007/s10549-012-2289-9.

**[6]** Liu, Zhimin, et al. “Epidermal Growth Factor Receptor Mutation in Gastric Cancer.” *Pathology*, vol. 43, no. 3, 2011, pp. 234–238., doi:10.1097/pat.0b013e328344e61b.

**[7]** Yoo, Nam Jin, et al. “Somatic Mutations of the KEAP1 Gene in Common Solid Cancers.” *Histopathology*, vol. 60, no. 6, 2012, pp. 943–952., doi:10.1111/j.1365-2559.2012.04178.x.

**[8]** Churi, Chaitanya R., et al. “Mutation Profiling in Cholangiocarcinoma: Prognostic and Therapeutic Implications.” *PLoS ONE*, vol. 9, no. 12, 2014, doi:10.1371/journal.pone.0115383.

**[9]** Zardavas, Dimitrios, et al. “PIK3CA Mutations in Breast Cancer: Reconciling Findings from Preclinical and Clinical Data.” *Breast Cancer Research*, vol. 16, no. 1, 2014, doi:10.1186/bcr3605.

**[10]** Li, Vivian Sze, et al. “Mutations of pik3cain Gastric Adenocarcinoma.” *BMC Cancer*, vol. 5, no. 1, 2005, doi:10.1186/1471-2407-5-29.

**[11]** ZHANG, HONG-YAN, et al. “PTEN Mutation, Methylation and Expression in Breast Cancer Patients.” *Oncology Letters*, vol. 6, no. 1, 2013, pp. 161–168., doi:10.3892/ol.2013.1331.

**[12]** Ikenoue, Tsuneo, et al. “A Novel Mouse Model of Intrahepatic Cholangiocarcinoma Induced by Liver-Specific KRAS Activation and PTEN Deletion.” *Scientific Reports*, vol. 6, no. 1, 2016, doi:10.1038/srep23899.

**[13]** Yao, Yiran, et al. “Novel Insights into RB1 Mutation.” *Cancer Letters*, vol. 547, 2022, p. 215870., doi:10.1016/j.canlet.2022.215870.

**[14]** Deng, Yujie, et al. “Mir-215 Modulates Gastric Cancer Cell Proliferation by Targeting RB1.” *Cancer Letters*, vol. 342, no. 1, 2014, pp. 27–35., doi:10.1016/j.canlet.2013.08.033. 17

**[15]** Choi, Michael, et al. “ATM Mutations in Cancer: Therapeutic Implications.” *Molecular Cancer Therapeutics*, vol. 15, no. 8, 2016, pp. 1781–1791., doi:10.1158/1535-7163.mct-15-0945.

**[16]** Borg, A., et al. “High Frequency of Multiple Melanomas and Breast and Pancreas Carcinomas in CDKN2A Mutation-Positive Melanoma Families.” *JNCI Journal of the National Cancer Institute*, vol. 92, no. 15, 2000, pp. 1260–1266., doi:10.1093/jnci/92.15.1260.

**[17]** Patocs, Attila, et al. “Breast-Cancer Stromal Cells with*tp53*Mutations and Nodal Metastases.” *New England Journal of Medicine*, vol. 357, no. 25, 2007, pp. 2543–2551., doi:10.1056/nejmoa071825.

**[18]** Kalinsky, Kevin, et al. “PIK3CA Mutation Associates with Improved Outcome in Breast Cancer.” *Clinical Cancer Research*, vol. 15, no. 16, 2009, pp. 5049–5059., doi:10.1158/1078-0432.ccr-09-0632.

**[19]** Ecke, Thorsten H., et al. “TP53 gene mutations in prostate cancer progression.” Anticancer research 30.5 (2010): 1579-1586.

**[20]** Rodriguez, Antonio, and Mark A. Rubin. “The Genomic Landscape of Prostate Cancer Brain Metastases.” 2020, doi:10.26226/morressier.5f69edb69b74b699bf38c5d9.

**[21]** Lian, Jianpo, et al. “Histone Methyltransferase KMT2C Plays an Oncogenic Role in Prostate Cancer.” *Journal of Cancer Research and Clinical Oncology*, vol. 148, no. 7, 2022, pp. 1627–1640., doi:10.1007/s00432-022-03968-5.

**[22]** Boysen, Gunther, et al. “SPOP Mutation Leads to Genomic Instability in Prostate Cancer.” *ELife*, vol. 4, 2015, doi:10.7554/elife.09207.

**[23]** Qiu, Wanglong, et al. “*pik3ca* Mutations in Head and Neck Squamous Cell Carcinoma.” *Clinical Cancer Research*, vol. 12, no. 5, 2006, pp. 1441–1446., doi:10.1158/1078-0432.ccr-05-2173. 13

**[24]** Zhou, Ge, et al. “*tp53*Mutations in Head and Neck Squamous Cell Carcinoma and Their Impact on Disease Progression and Treatment Response.” *Journal of Cellular Biochemistry*, vol. 117, no. 12, 2016, pp. 2682–2692., doi:10.1002/jcb.25592.

**[25]** ---

**[26]** Hill, Margaret A., et al. “*Kras* and *tp53* Mutations Cause Cholangiocyte- and Hepatocyte-Derived Cholangiocarcinoma.” *Cancer Research*, vol. 78, no. 16, 2018, pp. 4445–4451., doi:10.1158/0008-5472.can-17-1123.

**[27]** ---

**[28]** Ye, H, et al. “Synergistic Function of KRAS Mutation and HBx in Initiation and Progression of Hepatocellular Carcinoma in Mice.” *Oncogene*, vol. 33, no. 43, 2013, pp. 5133–5138., doi:10.1038/onc.2013.468.

**[29]** Lee, Jong Woo, et al. “PIK3CA Gene Is Frequently Mutated in Breast Carcinomas and Hepatocellular Carcinomas.” *Oncogene*, vol. 24, no. 8, 2004, pp. 1477–1480., doi:10.1038/sj.onc.1208304.

**[30]** Hussain, S P, et al. “TP53 Mutations and Hepatocellular Carcinoma: Insights into the Etiology and Pathogenesis of Liver Cancer.” *Oncogene*, vol. 26, no. 15, 2007, pp. 2166–2176., doi:10.1038/sj.onc.1210279.

**[31]** Moon, Hyuk, and Simon Weonsang Ro. “MAPK/ERK Signaling Pathway in Hepatocellular Carcinoma.” *Cancers*, vol. 13, no. 12, 2021, p. 3026., doi:10.3390/cancers13123026.

**[32]** “NRAS Gene: Medlineplus Genetics.” *MedlinePlus*, U.S. National Library of Medicine, medlineplus.gov/genetics/gene/nras/.

**[33]** Luo, Dan. “Analysis of n-*Ras*Gene Mutation and*p53*Gene Expression in Human Hepatocellular Carcinomas\*.” *World Journal of Gastroenterology*, vol. 4, no. 2, 1998, p. 97., doi:10.3748/wjg.v4.i2.97.

**[34]** Xu, Bo, et al. “Highly Expressed claudin18.2 as a Potential Therapeutic Target in Advanced Gastric Signet-Ring Cell Carcinoma (SRCC).” *Journal of Gastrointestinal Oncology*, vol. 11, no. 6, 2020, pp. 1431–1439., doi:10.21037/jgo-20-344.

**[35]** Sasaki, Taisuke, et al. “Tumor Progression by Epithelial-Mesenchymal Transition in arid1a- and SMARCA4-Aberrant Solid-Type Poorly Differentiated Gastric Adenocarcinoma.” *Virchows Archiv*, vol. 480, no. 5, 2022, pp. 1063–1075., doi:10.1007/s00428-021-03261-9.

**[36]** Lee, S., Lee, J., Soung, Y. et al. BRAF and KRAS mutations in stomach cancer. Oncogene 22, 6942–6945 (2003). https://doi.org/10.1038/sj.onc.1206749

**[37]** He, Xiu-Sheng. “Expression of*p16*Gene and RB Protein in Gastric Carcinoma and Their Clinicopathological Significance.” *World Journal of Gastroenterology*, vol. 11, no. 15, 2005, p. 2218., doi:10.3748/wjg.v11.i15.2218.

**[38]** Fenoglio-Preiser, C.M., et al. “TP53 And Gastric Carcinoma: A Review.” *Human Mutation*, vol. 21, no. 3, 2003, pp. 258–270., doi:10.1002/humu.10180.

**[39]** Huang, Dong-Sheng, et al. “Prevalence of Deleterious *Atm* Germline Mutations in Gastric Cancer Patients.” *Oncotarget*, vol. 6, no. 38, 2015, pp. 40953–40958., doi:10.18632/oncotarget.5944.

**[40]** Catanese, Helen N., et al. “A Nearest-Neighbors Network Model for Sequence Data Reveals New Insight into Genotype Distribution of a Pathogen.” *BMC Bioinformatics*, vol. 19, no. 1, 2018, doi:10.1186/s12859-018-2453-2.

**[41]** Cho, Ara, et al. “Muffinn: Cancer Gene Discovery via Network Analysis of Somatic Mutation Data.” Genome Biology, vol. 17, no. 1, 2016, doi:10.1186/s13059-016-0989-x.

**[42]** Chou, Wei-Chun, et al. “Visual Gene-Network Analysis Reveals the Cancer Gene Co-Expression in Human Endometrial Cancer.” *BMC Genomics*, vol. 15, no. 1, 2014, doi:10.1186/1471-2164-15-300.

**[43]** Gemignani, Mary L, et al. “Role of KRAS and BRAF Gene Mutations in Mucinous Ovarian Carcinoma.” *Gynecologic Oncology*, vol. 90, no. 2, 2003, pp. 378–381., doi:10.1016/s0090-8258(03)00264-6.

**[44]** Ben J. Milner, Lindsey A. Allan, Diana M. Eccles, Henry C. Kitchener, Robert C. F. Leonard, Kevin F. Kelly, David E. Parkin, Neva E. Haites; p53 Mutation Is a Common Genetic Event in Ovarian Carcinoma1. Cancer Res 1 May 1993; 53 (9): 2128–2132.

**[45]** Philpott, Charlotte, et al. “The NF1 Somatic Mutational Landscape in Sporadic Human Cancers.” *Human Genomics*, vol. 11, no. 1, 2017, doi:10.1186/s40246-017-0109-3.

**[46]** Yang, Ming, et al. “Functional Variants in Cell Death Pathway Genes and Risk of Pancreatic Cancer.” Clinical Cancer Research, vol. 14, no. 10, 2008, pp. 3230–3236., doi:10.1158/1078-0432.ccr-08-0177.

**[47]** Shain, A. Hunter, et al. “Convergent Structural Alterations Define Switch/Sucrose Nonfermentable (SWI/SNF) Chromatin Remodeler as a Central Tumor Suppressive Complex in Pancreatic Cancer.” Proceedings of the National Academy of Sciences, vol. 109, no. 5, 2011, doi:10.1073/pnas.1114817109.

**[48]** Alvino, Ester, et al. “High Expression of the Mismatch Repair Protein MSH6 Is Associated with Poor Patient Survival in Melanoma.” American Journal of Clinical Pathology, vol. 142, no. 1, 2014, pp. 121–132., doi:10.1309/ajcpcx2d9yulbblg.

**[49]** Spoerri, Loredana, et al. “A Novel ATM-Dependent Checkpoint Defect Distinct from Loss of Function Mutation Promotes Genomic Instability in Melanoma.” Pigment Cell &amp; Melanoma Research, vol. 29, no. 3, 2016, pp. 329–339., doi:10.1111/pcmr.12466.

**[50]** Aydin, Iraz T., et al. “FBXW7 Mutations in Melanoma and a New Therapeutic Paradigm.” JNCI: Journal of the National Cancer Institute, vol. 106, no. 6, 2014, doi:10.1093/jnci/dju107.

**[51]** Rajiv Kumar, Sabrina Angelini, Kamila Czene, Ilari Sauroja, Marjo Hahka-Kemppinen, Seppo Pyrhönen, Kari Hemminki; BRAF Mutations in Metastatic Melanoma: A Possible Association with Clinical Outcome. Clin Cancer Res 15 August 2003; 9 (9): 3362–3368.

**[52]** Conlon, Niamh, et al. “Loss of SMARCA4 Expression Is Both Sensitive and Specific for the Diagnosis of Small Cell Carcinoma of Ovary, Hypercalcemic Type.” American Journal of Surgical Pathology, vol. 40, no. 3, 2016, pp. 395–403., doi:10.1097/pas.0000000000000558.

**[53]** “CTNNB1 Gene: Medlineplus Genetics.” *MedlinePlus*, U.S. National Library of Medicine, medlineplus.gov/genetics/gene/ctnnb1/.

**[54]** Zhou, Cefan, et al. “TSPAN1 Promotes Autophagy Flux and Mediates Cooperation between Wnt-CTNNB1 Signaling and Autophagy via the mir454-fam83a-TSPAN1 Axis in Pancreatic Cancer.” Autophagy, vol. 17, no. 10, 2020, pp. 3175–3195., doi:10.1080/15548627.2020.1826689.

**[55]** DANAEE, PADIDEH, et al. “A Deep Learning Approach for Cancer Detection and Relevant Gene Identification.” Biocomputing 2017, 2016, doi:10.1142/9789813207813\_0022.

**[56]** “STRINGdb.” Bioconductor, www.bioconductor.org/packages/release/bioc/html/STRINGdb.html.

**[57]** “The Levenshtein Distance Algorithm.” Educative, www.educative.io/answers/the-levenshtein-distance-algorithm.

**[58]** chiang, Jeffery. “Girvan–Newman - the Clustering Technique in Network Analysis.” Medium, Analytics Vidhya, 16 Apr. 2022, medium.com/analytics-vidhya/girvan-newman-the-clustering-technique-in-network-analysis-27fe6d665c92.

**[59]** International Cancer Genome Consortium. dcc.icgc.org/.

**Appendix**

**Prostate Cancer (Prostate Adenocarcinoma)**

Our method identified TP53, NF1, KEAP1, KMT2C, and SPOP as the key genes responsible for prostate cancer. According to the Human Protein Atlas, in regard to the prostate itself (specifically the glandular cells), these key genes have a range of expressions; TP53 is not expressed at all in either cell types; NF1 and KEAP1 has medium expression in the glandular cells; Unfortunately, it is unknown as to the level of expression of KMT2C and SPOP.**[1]** In terms of cancer:

* Mutation (overexpression) in the TP53 gene is a factor in prostate cancer development as it is considered as a tumor suppressor gene which not only synthesizes P53 protein that is in charge of the cell cycle, checkpoint control, repair of DNA, and apoptosis, but also mutations in the exon 7 and 8 of the gene are another factor of cancer progression**[19]**
* Mutation (inactivation) of the NF1 gene is a factor in prostate cancer development as it is involved in targetable pathways and it is possible for downstream activation of druggable pathways RAS/RAF/MEK1-2/ERK1-2 and PI3K/AKT/mTOR1/2, respectively**[20]**
* Mutations (inactivation) in the KEAP1 gene is a factor in prostate cancer development as it is considered as a tumor suppressor gene, and it is possible that cancer cells are protected from oxidative insults via an increase in cytoprotective proteins**[7]**
* Mutation (overexpression) of the KMT2C gene is a factor in prostate cancer development as it is an epigenetic modifier that catalyzes monomethylation of H3K4 in the gene enhancer regions, and it is linked to the CAMs signaling pathway as well as the MAPK/ERK signaling pathway, both of which can result in cancer development**[21]**
* Mutation (loss of function) of the SPOP gene is a factor in prostate cancer development as it helps to accurately repair damaged DNA, as such the mutation of it changes the repair process resulting in genomic instability because broken chromosomes are repaired incorrectly, moreover, it also sensitizes DNA damaging therapeutic agents**[22]**.

**Head and Neck Cancer (Head and Neck Squamous Cell Carcinoma)**

Our method identified CDKN2A, PIK3CA, and TP53 as the key genes responsible for head and neck cancer. According to the Human Protein Atlas, in regard to the head itself, specifically the Nasopharynx (ciliated cells and basal cells) and Salivary Gland (glandular cells)**[2]**, these key genes have a range of expressions; CDKN2A is low expressed in the salivary glands (glandular cells), but not expressed in the nasopharynx; PIK3CA is highly expressed in the salivary gland (glandular cells) and partially high in the nasopharynx (low in basal cells and high in the cell body of ciliated cells); TP53 is not expressed at all in either body parts and cell types.**[1]** In terms of cancer:

* Mutations (inactivation) in the CDKN2A gene is a factor in head and neck cancer development as its role is of a tumor suppressor gene and produces a negative regulator of cyclin-dependent kinases, and a stabilizer for p53, another tumor suppressor gene**[4]**
* Mutation (gene amplification, overexpression and small mutations) in the PIK3CA gene is a factor in head and neck cancer development as it encodes a catalytic subunit of PI3K, and it activates the gene’s oncogenic properties**[23]**
* Mutations (missense) in the TP53 gene is a factor in head and neck cancer development as it is considered as a tumor suppressor gene, and so two types of mutant P53 protein, one present in the amino acids and bind to responsive element in the DNA, and one altering its structure to remove its DNA binding ability, are synthesized**[24]**.

**Biliary Cancer (Cholangiocarcinoma)**

Our method identified GRIN2A, KRAS, PTEN, and TP53 as the key genes responsible for biliary cancer. According to the Human Protein Atlas, in regard to the gallbladder itself (specifically the glandular cells), these key genes have a range of expression; KRAS is highly expressed in the glandular cells; PTEN is low expressed in the glandular cells; GRIN2A and TP53 are not expressed at all.**[1]** In terms of cancer:

* Mutation () in the GRIN2A gene is a factor in biliary cancer development \_**[25]**
* Combined mutation in the KRAS (activation) and PTEN (deletion) genes are factors in biliary cancer development as the latter antagonizes the activity of PI3K, and with its deletion and PTEN inactivation, excessive PI3K signaling pathway activation occurs**[12]**, with KRAS being a more frequent mutation and PTEN a more common one**[8]**
* Mutation (loss) in the TP53 gene is a factor in biliary cancer development as it is considered as a tumor suppressor gene and enhanced the reprogramming of cholangiocytes from hepatocytes[26]. Additionally, it is a more common mutation**[8]**.

**Ovarian Cancer (Ovarian Adenocarcinoma)**

Our method identified KRAS, TP53, ATM and NF1 as the key genes responsible for head and neck cancer. According to the Human Protein Atlas, in regard to the ovaries itself (specifically the follicle cells and ovarian stroma cells), these key genes have a range of expression; KRAS and TP53 are not expressed at all in either cell types; ATM and NF1 are low expressed in the ovarian stroma cells **[1]**. In terms of cancer:

* Mutation (activation) in the KRAS gene is a factor in ovarian cancer development as it acts as a relay switch which from the cell surface, transduces various growth signals towards the nucleus, and with its activation, makes it growth stimulatory. **[43]**
* Mutation (aberration) in the TP53 gene is a factor in ovarian cancer development as it is considered as a tumor suppressor gene, and the synthesis of P53 protein, which when mutated can cause cancer, however it is not known how this occurs. **[44]**
* Mutation (deletion) in the ATM gene is a factor in ovarian cancer development as it maintains genomic integrity, as well as has a central role in the repair of DNA double-strand breaks, and increases instability of the genome as repair mechanisms increasingly depend on cancer cells for replication stress after cell division. **[15]**
* Mutation (deletion) in the NF1 gene is a factor in ovarian cancer development as it is considered as a tumor suppressor gene, and mutations result in a loss of growth control of the cells as well as enhanced cellular proliferation. **[45]**

**CNS&Bone&Myeloid Cancer**

The graphical result

|  |  |
| --- | --- |
| Diagram  Description automatically generated  **SSN** | Table  Description automatically generatedTable  Description automatically generated  **The strongly related genes** |
| Diagram  Description automatically generated  **The original PPIN** | Chart, diagram  Description automatically generated  **Reorganized PPIN** |

The evaluation by HPAanalyze

**Chart, bar chart

Description automatically generated**