Universiteit Leiden
The Netherlands

Opleiding Bioinformatica

Exploring hallmark connections through pathway enrichment in a cancer gene PPI-network.

Karl Freeke

Supervisors:
Dr. K.J. Wolstencroft

BACHELOR THESIS
Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/02/2021

Abstract

The goal of the research was to create a protein-protein interaction (PPI) network from the COSMIC Cancer Gene Census and to find a connection between the different hallmarks in the Census and overrepresented pathways in a PPI network. The expression of hallmarks is a direct result of somatic mutations in DNA. Understanding the concepts of the cancer hallmarks and connections between them at a data level, from a perspective of functional involvement of hallmark annotated genes in different or equivalent pathways, could help us infer new hallmark annotations from existing data and may illuminate connections between hallmarks. Therefore multiple sources of pathway data were overlayed onto the network.

Data on cancer gene prevalence in enriched pathways per hallmark was generated. Integrating this data into the PPI network and comparing this data mutually per hallmark enabled comparison between genes and how deeply they are involved in pathways that are connected to each hallmark.

It was found that a collection of highly connected genes seem to be involved in the expression of multiple if not all hallmarks. Separate hallmarks may not be expressed through entirely separate pathways within a cell. The expression of all different hallmarks is to some extent interconnected through a subset of genes, which are significantly enriched in different pathways connected to different hallmarks.

By analysing the (lack of) overlap of gene prevalence of genes annotated with different hallmarks, in combination with the generated pathway scores, it was attempted to infer some predictions for new hallmark annotations.

Another challenge was to create a clear and intuitive visualization of the data in the network. A pathway enriched PPI Cytoscape network was produced along with the network analysis results and some Python code.

Contents

1 Introduction

1.1 The Hallmarks of Cancer

It can be stated that cancer consists of abnormal cells which divide in an uncontrolled way. This however is an oversimplified rendering of many complex processes. In order to summarise the complexities of carcinogenesis and the multi-factorial aetiology of cancer in general, a collection of six hallmarks were proposed D. Hanahan and R. Weinberg[7]. In 2011, four more hallmarks were added to the list to establish a total of ten hallmarks [6]. Each hallmark can be interpreted as a principle by which cancer cells prevail abnormally. Twenty years later, this study has become a true landmark. Before this publication the field of cancer research was collection of individual findings that were not connected to any underlying principles. From this publication on, many cancer research findings have been placed in the context of the hallmarks concept. The hallmarks of cancer are now a widely accepted concept within the field of cancer research.

The list of ten hallmarks include: proliferative signaling, suppression of growth, escaping immunic response to cancer, cell replicative immortality, tumour promoting inflammation, invasion and metastasis, angiogenesis, genome instability and mutations, escaping programmed cell death, change of cellular energetics.
These hallmarks are shared in common by practically all cancer cells. Figure 1 shows a visualization and overview of the hallmarks which are all summarised on the next page.

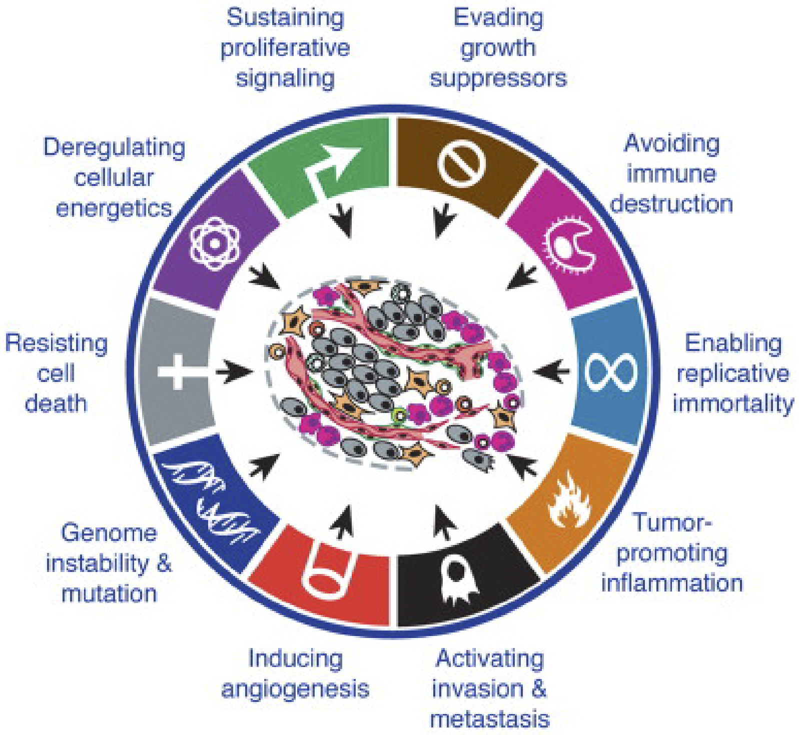

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1: A visual representation of the 10 hallmarks [7].

Sustaining proliferative signaling
When a cell is not dependent on external growth signals, but has self-sufficiency in providing growth signals instead.

Evading growth suppressors
Tumour suppressor genes provide anti-growth signals to prevent cells from growing and replicating uncontrollably. Cancer cells can be insensitive to these inhibitory signals.

Escaping immunic response to cancer
The immune system is capable of inducing apoptosis of damaged cells. Cancer cells however, are unaffected by the immune system. Cancer cells often exist in an equilibrium between uncontrolled proliferation and immunic inhibition of growth [12] [16]. During this period of equilibrium a cancer cell can either adapt sufficiently, or later succumb to the T-cells of the immune system.

Cell replicative immortality
Healthy cells have a natural maximum limit to their number of divisions, determined by telomere length. This limit is called the Hayflick limit [22]. As a result of overexpression, cancer cells have an increased activity of telomerase [23]. Telomerase adds a telomere repeat sequence to the 3' end of the telomere. This enables cancer cells to replicate unlimitedly without dying.

Tumour promoting inflammation
This hallmark has also been characterized as an enabling hallmark. There is a large body of evidence indicating that chronic inflammation can be a starting point for cells to become cancerous [20]. Inflamed tissue produces chemokines and cytokines among other components which can enhance carcinogenesis [20]. This can contribute to enabling more hallmark capabilities in a cell. Cancer cells can use inflammatory components for their own proliferation [10].

Invasion and metastasis
Cancer cells are capable of spreading through the body and form secondary tumours in different parts of the body by invading healthy tissue.

Angiogenesis
Cancer cells are capable of inducing the growth of new blood vessels. These new blood vessels will then enable the growth of more cancer cells.

Genome instability and mutations
This hallmark has also been characterized as an enabling hallmark. Genomic instability in cancer cells can result in random mutations including chromosomal translocations, deletions, duplications, and inversions. In rare occasions, these mutations can result in cells acquiring hallmark capabilities.

Escaping programmed cell death
When the DNA of a cell has been damaged, the tumor-suppressor protein p53 accumulates in order to promote the apoptosis of the cell [1]. This is a form of programmed cell death. There are different mechanisms that allow cancer cells to avoid apoptosis capable of bypassing this self-destruction mechanism. Genome instability can result in a mutation in the P53 gene with an non-coding gene as result. P53 expression is often reduced or even absent in cancer cells [25].

Change of cellular energetics
A cancer cell is capable of changing its metabolic system. A healthy cell usually generated its energy from cellular respiration with breaking down glucose as its main source of energy. Cancer cells however often rely on other forms of deregulated cellular energetics. Most cancer cells depend on aerobic glycolysis instead, also known as the Warburg effect [27]. Some cancer cells show increased consumption of and dependence on glutamine, which is also a promoting factor of cell proliferation [4].

1.2 COSMIC Cancer Gene Census

The COSMIC (Catalogue Of Somatic Mutations In Cancer) Cancer Gene Census[24] database contains a list of genetic mutations which are considered to have a causal relation to carcinogenesis. The list currently contains 723 different genes, manually curated from over 26 000 peer reviewed scientific publications. In order to add a general description of the function of genes, the curators of this database are continuously trying to add hallmark annotations by reviewing experimental evidence of functional involvement. The Cancer Gene Census is an ongoing effort and updates may be released at any time.

The Census is divided in Tier 1 genes and Tier 2 genes. Tier 1 genes have documented evidence that shows activity relevant to cancer. Tier 2 genes have a strong indication of playing a role in carcinogenesis, but with a smaller body of evidence. Most of these genes are connected to one or multiple hallmarks which describe their role in carcinogenesis based on high-confidence scientific publications. The database also mentions if the gene-product either promotes or suppresses a process related to cancer. All genes in the Census have been causally implicated in cancer, meaning that most genes should probably be linked to at least one hallmark. The fact that not all Census genes have been annotated with a hallmark will be one of the points of interest in this thesis. All current hallmark annotations have been assigned after careful manual curation of scientific papers, which requires great effort. Through network analysis, it may be possible to speed up this process by inferring new hallmark predictions from previously assigned hallmarks.

1.3 Protein-protein Interaction Networks

Proteins are responsible for a wide range of functions within a cell. Often a single protein does not have any function within a cell until the protein interacts with one or more other proteins. These interactions are physical contacts between two or more proteins as a result of biophysical

conditions.
In a protein-protein interaction network (PPI network), gene products are represented by nodes and (possible) interactions are represented by edges. Interactions often have a evidence-score, describing the probability of the interaction. PPI networks are by definition highly interconnected and have a small diameter. This means that the distance from any given node to any other node is usually not bigger than six steps. This is also known as a small-world network. It must be mentioned that all data representing protein-protein interactions is to some degree noisy, as these molecular interactions can not simply be represented by a binary 1 or 0. Therefore PPI networks are always undirected weighted graphs.

The tendency for two proteins to form an interaction with each other is called the binding affinity. Because of gene mutations, the binding site of proteins may be altered in such a way that interactions that could happen before are no longer possible. In other words, the binding affinity of the resulting gene products may be affected by these genetic mutations.
It can be very useful to study these PPI networks in order to discover which proteins are most prominently involved in certain biological pathways. Protein function can be determined through experiments, but can also be predicted with computational approaches involving PPI networks [18]. PPI networks can also help bring to light new pathways and protein complexes within a cell.

Protein-protein-interaction data can be produced by both computational and experimental methods. The most used experimental methods are the yeast two-hybrid protein-fragment complementation assay and affinity purification - mass spectrometry. Computational prediction of PPI's is usually based on existing experimental PPI data. However, machine learning methods have proven that de novo PPI prediction, without the use of previous experimental data, can be a legitimate approach for finding new PPI's. [8]

1.4 Pathway Enrichment

A biological pathway can be described as a sequence of interactions or chemical reactions between genes, proteins, or other molecules within a cell. These pathways will often result in a new cellular state. A new molecular product such as a new protein could be produced, or a yet existing molecule could be modified as a result of the chemical interactions within a pathway. Biological pathways can either occur within a single cell, but they can also be part of a larger intercellular mechanism. Pathways are mostly discovered through laboratory studies.
There are different kind of biological pathways, including metabolic pathways, gene regulation pathways, and signal transduction pathways. It is expected to find different kinds of pathways to be related to our Cosmic Cancer Gene data set, as metabolic processes, gene regulation, and signal transduction can all play a role in the development of cancer [28].

The approach of pathway enrichment analysis is not only useful for cancer research, it can help us understand which exact processes in the cell are responsible for the development of diseases in general. Identifying which step of a a pathway compromises the healthy state of a cell can help us find drug targets for the development of new treatments [2]. Genetic alterations in signaling

pathways involved in cell growth, apoptosis, or the cell-division cycle have often been topics of research, with the goal of discovering new drug targets [9].
Pathway enrichment results will usually be in the form of a list of pathways that are overrepresented in the queried list of genes. Each returned pathway has a P-value which describes how likely that pathway could be connected to the gives list of genes by chance. A p-value of 0.05 is usually considered as the threshold by which the enrichment of a pathway can be considered statistically significant. Each pathway entry will contain a list of gene products which are involved in said pathway. Such enrichment can result in functionally redundant pathways, which means that trimming of some pathway data may be required. One of the goals for this thesis was to summarize our gene list into a collection of overrepresented biological pathways. This data can then be used for further analysis.

1.5 Network Generation and Analysis Tools

1.5.1 Cytoscape

All network analysis, network visualization, and data integration was done in the software environment of the Cytoscape application [15] [21]. Cytoscape is an open source, Java based application for Windows/Linux/MacOS which can be used to visualise and analyse molecular interaction networks. Cytoscape can be used in combination with third party tools which can be downloaded from the Cytoscape App Store. All work was done in Cytoscape version 3.8.2.

1.5.2 STRINGapp and PPI networks

The STRINGapp plugin was used to generate the PPI network from the Cosmic Cancer Census Genes list. STRINGapp [17] is a Cytoscape application and serves to incorporate data from the STRING database [29] in order to retrieve and visualise PPI networks. STRINGapp was chosen over other similar applications for its database which is curated from five different sources. Furthermore, the implementation of a 'score cutoff value' can help reduce the amount of noisy data when generating the PPI network.
The STRING database is a protein-protein interaction database which uses both experimentally inferred and computationally predicted data. All interactions are annotated with a confidence score which quantifies their reliability.

Interactions in the STRING database are derived from the following five sources:

- Genomic Context Predictions
- High-throughput Lab Experiments
- (Conserved) Co-Expression
- Automated Textmining
- Previous Knowledge in databases (IntAct, BioGRID)

When running any query, a confidence cutoff score can be configured. A higher confidence score threshold results in a more specific network with fewer nodes and edges with a higher confidence score, whereas a lower confidence score threshold allows more nodes and edges with lower confidence scores to be generated.
The evidence for each interaction is divided in seven components:

- fusion evidence
- neighborhood evidence
- cooccurrence evidence
- experimental evidence
- textmining evidence
- database evidence
- coexpression evidence.

1.5.3 g:Profiler and Enrichment map

g:Profiler's [19] g:GOSt is a web based application which can discover statistically significantly enriched biological processes (such a pathways) from a list of genes. g:Profiler uses a broad array of different data sources which a user can choose from. The supported data sources for pathway enrichment are Reactome, KEGG, and Wikipathways. The typical g:Profiler enrichment analysis result is a list of enriched pathways. The statistical significance for each pathway is included as a P-value. Furthermore, g:Profiler's [19] g:Convert was used, which can convert pathway term id's to a list of involved genes.
EnrichmentMap [14] is a Cytoscape plugin which allows analysis and visualization of enrichment data. EnrichmentMap can produce a network from enrichment data and calculates the overlap between different enriched terms. In a Cytoscape network generated by EnrichmentMap each node represents an enriched pathway. Each edge represents an overlap of genes between two different enriched pathways. The generated Cytoscape network helps to filter and visualize the data generated by g:Profiler. In this thesis. EnrichmentMap was only used for data integration purposes.

1.5.4 Omics Visualizer

Omics Visualizer [13] is a Cytoscape plugin which allows for visual representation of data in the shape of pie charts and donut charts on nodes.

1.6 Data Resources

Protein-protein interaction data source:

STRING database v11.0 [26]

Multiple pathway data sources were used for the pathway enrichment. Data from the following databases was used:

KEGG, version: FTP Release 2020-09-07

Reactome, version: annotations: BioMart, classes from version 2020-10-12

WikiPathways, version: 2020-10-10

1.7 Research Question

The goal of this thesis was to add structure to the generated PPI network of cancer genes from the COSMIC Cancer Gene Census by integrating biological pathway data and to try to infer possible predictions for hallmark annotations from this newly introduced structure. This involved analysis of over-represented pathways within the PPI network. Multiple sources of pathway data were overlayed on the network.

Finding a connection between the different hallmarks in the Census and over represented pathways in the PPI network may help us understand to what extent the distribution of hallmarks depends on biological mechanisms. It may be possible to connect unannotated genes to a hallmark by enriching the network with pathway data and then applying different methods of network analysis. Another challenge was to create a clear and intuitive visualization of the network.

The work done for this thesis is centered around the following research question:

> *Can we identify cancer hallmark structures by integrating pathway data in a PPI-network?*

Enriching the PPI network with pathway data will allow us to comprehend how the hallmarks are distributed over different overrepresented pathways that are involved in cancer. It will be interesting to investigate whether or not the actual distribution of hallmarks meets the hypothetical expectations. For example, it would be expected that genes annotated with the 'Angiogenesis' hallmark would be overrepresented in certain signal transduction pathways connected to angiogenesis. Another example is that it would be expected to see the 'escaping programmed cell death' hallmark overrepresented in pathways that are involved with regulation of TP53 activity.

2 Methods

A summarizing workflow diagram can be found in figure 3.

2.1 Network generation

Using the curated list of genes from the COSMIC Cancer Gene Census we can create a PPI network which can be interpreted as a subset of the protein interactome of all cancers. To generate the PPI network the unranked gene list from the Cosmic Cancer Gene Census was queried using the 'STRING protein query' function. The goal was to generate a network with a justifiable trade-off between retrieving high confidence interactions with low selectivity and lower confidence interaction with high sensitivity. Default cutoff value is 0.5. The goal was to minimize the amount of noisy interactions without eliminating too much data needed for analysis. For this thesis, a lower coverage with higher confidence interactions and less false positives is desired. Multiple PPI networks were generated with different confidence cutoff scores.

Below a table can be seen with different score cut-off value resulting in different network sizes and interconnectivity between nodes.

Confidence cutoff	Number of interactions
0.4	14884
0.5	9948
0.6	7062
0.65	6197
0.7	5375

The fusion evidence score for interactions is derived from fused proteins in other species. We may want to avoid interactions with a substantial fusion evidence score as we are only interested in interactions that occur in Homo Sapiens. It was found that the amount of fusion evidence annotated in the PPI network for any confidence cutoff was negligible. Therefore the amount of fusion evidence was not taken into account for selecting the confidence cutoff score.

A higher than default cutoff score of 0.65 eliminates 8687 interactions compared to when a confidence score of 0.4 is used. Going up to 0.7 removes another 822 interactions which would be 14.2% of the total set of interactions. A further small step up in cutoff score would result in a substantially smaller data set without adding much validity to the data Therefore a middle ground was found at 0.65. This cutoff score resulted in 50 unconnected nodes, as can be seen at the bottom of figure 6.

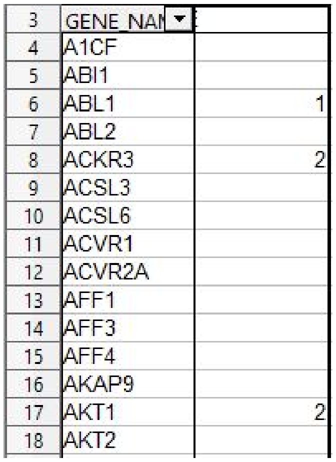
The generated PPI network contains 717 nodes, whereas the Census contains 723 entries. This is because the following 6 genes could not be mapped to any matching gene product by the STRINGapp application:

- HMGN2P46
- IGK
- MRTFA
- TENT5C
- TRB
- TRD

Furthermore, multiple possible identifiers were found for 6 other genes. By considering the known synonyms and genomic locations, the correct gene product identifiers were selected.

2.2 Adding Hallmark annotations to the PPI network

The *Cancer_Gene_Census_Hallmarks_Of_Cancer.csv* file was downloaded from the COSMIC website. Ten new columns were created of which each column contains the name of one of the ten hallmarks. The *HALLMARK* column was then split per hallmark and added to the new 10 separate hallmarks columns. As the file contains multiple entries per gene (individual entries for individual hallmarks annotated to the same gene), all separate hallmark annotations had to be mapped to a single gene. This was done by using the pivot tables functionality of OpenOffice Calc. In the pivot table setup, *'GENE_NAME'* was added to the *'Rows'* field, and any of the separated hallmark columns was added to the *'Data'* field. The result was then sent to a new worksheet. A small snippet of the resulting file can be seen in figure 2 below.



3	GENE_NAM	
4	A1CF	
5	ABI1	
6	ABL1	1
7	ABL2	
8	ACKR3	2
9	ACSL3	
10	ACSL6	
11	ACVR1	
12	ACVR2A	
13	AFF1	
14	AFF3	
15	AFF4	
16	AKAP9	
17	AKT1	2
18	AKT2	

Figure 2: Pivot Table Result.

Above result consists of a column of all COSMIC cancer genes without duplicates, and another column with a number representing how often that hallmark annotation occurred in the *Cancer_Gene_Census_Hallmarks_Of_Cancer.csv* for a given gene. All number entries in the second column were then changed to **True** and the full column was then copied over to the final .csv file with all other hallmark annotations. All remaining empty fields were set to **False**. This file was then imported to the STRING PPI network in Cytoscape, with the gene name as key/shared column. This data integration enabled boolean filtering on hallmark characteristics in Cytoscape. This process was repeated for each of the hallmarks, and also for Tumour Suppressor Gene, Fusion Gene, and Oncogene annotations. The resulting file can be found in the the GitLab repository. This file was then imported into the generated PPI network in Cytoscape, essentially adding all hallmark annotations to the network.

2.3 Elemental Network Topology and Hallmark Metrics Analysis

The first step was to look at some of the basic network topology. Using a native Cytoscape application called NetworkAnalyser [3], elemental network metrics were analysed.
To get a general sense of the connectivity of the network, the closeness centrality, betweenness centrality, and clustering coefficient, and other metrics were calculated by using the NetworkAnalyzer application. This was done for each individual node from which full network averages were calculated.

The **betweenness centrality** is a measure that reflects the amount connections between other nodes that flow through a given node, and is calculated as follows:

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

Where g_{jk} = the number of geodesics (shortest paths) connecting jk, and $g_{jk}(n_i)$ = the number that node i is on.

The **closeness centrality** is a measure that reflects how close a node is to all other reachable nodes in the network, and is calculated as follows:

$$CC(i) = \frac{N-1}{\sum_j d(i,j)}$$

where
$i \neq j$,
d_{ij} is the length of the shortest path between nodes i and j in the network,
N is the number of nodes.

The **clustering coefficient** describes how close a node and its neighbours are to forming a clique. A clique is a maximally connected subgraph, e.g. a subgraph in which each node has an edge to all other nodes. The clustering coefficient is calculated as follows:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

2.3.1 Global Network Metrics

The **network density** value describes the normalized average number of neighbors, therefore being a value between 0 and 1.

The **characteristic path length** value describes the average shortest path length and is therefore the average expected distance between two random nodes in the network.

The **network centralization** value describes how central the most central node is compared the to the centrality of all other nodes. In other words, the proportion of connectivity between central nodes and lesser connected nodes.

2.3.2 Network Hubs

To find out which genes play a most prominent role, the hubs of the network were identified by looking at the centrality analysis results. The hubs of a network were identified by a combination of betweenness, closeness, and degree, i.e. the nodes with the most connections to other nodes. Top ten lists were determined for these three metrics. All unique genes from those lists were then identified as the hubs of the network, resulting in a set of 14 nodes.

2.4 Pathway Enrichment

2.4.1 Data Curation

The next step was to summarize the cancer gene list into a collection of over represented biological pathways.
For each hallmark, the list of genes annotated with that hallmark was queried in g:Profiler's g:GOSt application for functional profiling. All but the following settings were left on default:

Organism: Homo Sapiens

Data sources: No electronic GO annotations, KEGG, Reactome, WikiPathways

For some gene identifiers, multiple Ensembl GeneID's (ENSG) were found by g:Profiler, which allows an option to select the genes with the most GO annotations. Using this option results in an incorrect selection of multiple gene identifiers. Therefore, these gene identifiers were compared to the gene identifiers in the COSMIC Cancer Gene Census to ensure the correct genes were selected. The query results were then downloaded as .CSV and .GEM files for each of the hallmarks (**gene list query**). The query URLs can be found in 2.5.1.

Using the EnrichmentMap application within Cytoscape was used to generate pathway networks from the g:Profiler results .GEM files. This was done for all ten files for each hallmark. All settings were left on default. These EnrichmentMap networks were only generated as a means of adding the full g:Profiler results to the Cytoscape network file and increasing the completeness of data in the network file. No further research was done on these EnrichmentMap networks.

The third column in the .CSV files contains the Pathways term id´s. This full column (minus the column name) was copied into g:Profiler's g:Convert application in order to retrieve all genes that are involved in these pathways (**pathway query**). The query URLs can be found in 2.5.1. The results were downloaded as .CSV files. The third column of these .CSV files contains the occurences for all genes in all retrieved pathways.

2.4.2 Data Integration

A small Python script was written to help integrate the g:Profiler pathway data into the STRINGapp PPI network. The code of the Python program can be found in the appendix 6.1, as well as in the GitLab repository. This program calculates the occurrences of per gene from a file containing a list of genes. These occurrences are then sorted from high to low. After this the number of occurrences are normalized to a value between 0 and 1. The output of the program is a text file with list of the gene names and normalized scores of their occurrences, separated by a comma. The program also plots this data in a bar graph.

Using the following pipeline, the full list of occurring genes was used as input for the prevalence.py script for calculating normalized prevalence scores and generating the prevalence graphs:

```
cut -d ',' -f3 Angiogenesis.csv | cut -c 2-  | sed 's/.$//' | Python3 prevalence.py
```

This process was repeated for all ten hallmark g:Profiler results. The resulting text files with normalized prevalence scores were then imported into Cytoscape.

2.4.3 Pathways In Cancer (KEGG:5200)

The 'Pathways in cancer' (KEGG:5200) is one of the most enriched pathways in all g:Profiler queries. Comparing the included genes of this pathway to our network could possibly bring forth new insights. Therefore this pathway was overlayed onto the network to compare with the PPI network and to see if there are any anomalies. In order to integrate this data, a column was added to include which genes are involved in the KEGG pathway 'Pathways In Cancer' (hsa05200/KEGG:5200).

2.5 Network Visualization

The colours that are automatically added to all nodes by the STRINGapp application on generation of the network are arbitrary. Therefore all node colours were changed to black with a white font for a more clear visualization. For all network visualizations, the node size is mapped to the degree of the node. The hubs as determined in 3.1 were visualized with a red border. The resulting visualization of the network can be seen in figure 7.

To get an overview of all genes that have a significant prevalence in the pathways connected to one or more of the ten hallmarks, a subnetwork was created using a node filter in Cytoscape. This filter selects all genes that have a pathway prevalence score >0.5 for at least one hallmark.

A visualization of the resulting subnetwork can be seen in figure 10 in the results section. From here on, this subnetwork will be referred to as the 'pathway score >0.5' subnetwork. An image of this subnetwork can be seen in 10. Omics Visualizer [13] was then used to create the split donut and pie chart visualizations of the hallmark annotations and pathway scores. This was done for the complete network as well as multiple subnetworks.

Adding the pathway scores to the network enabled visualization of these scores corresponding to the visualization of the hallmarks. The pathway scores were visualized in a split donut chart corresponding to the legend in figure 8. The pathway scores are mapped to a colour gradient from white to red, where white resembles the lowest possible score and becomes more red as the score increases. This enabled easy comparison of hallmark annotation and associated pathway scores. The resulting visualization can be seen in figure 11.
The Compound Spring Embedder (CoSE) [5] layout was applied to all networks. CoSE is an algorithm for force-directed graph drawing which results in a layout in which the number of crossing edges is minimized and where all edges are of approximately equal length. This results in a visually pleasing layout in which connected nodes are placed closely together.

2.5.1 Query URLs

The queries that were ran and their results can be accessed through the following links:

Angiogenesis:
Gene list query: https://biit.cs.ut.ee/gplink/l/_Z9C3hqoT-
Pathway query: https://biit.cs.ut.ee/gplink/l/AZlhzsT8RL

Cell Replicative Immortality:
Gene list query: https://biit.cs.ut.ee/gplink/l/wsj82QVlSi
Pathway query: https://biit.cs.ut.ee/gplink/l/AZlhzsT8RL

Change of Cellular Energetics:
Gene list query: https://biit.cs.ut.ee/gplink/l/DHUmKedMTS
Pathway query: https://biit.cs.ut.ee/gplink/l/L80R7JL0T0

Escaping Immune Response:
Gene list query: https://biit.cs.ut.ee/gplink/l/uPBcgd6ER_
Pathway query: https://biit.cs.ut.ee/gplink/l/nZoKGvHVQV

Escaping Programmed Cell Death:
Gene list query: https://biit.cs.ut.ee/gplink/l/wk0rbfi1Qw
Pathway query: https://biit.cs.ut.ee/gplink/l/yfwIK3egS6

Genome Instability and Mutations:
Gene list query: https://biit.cs.ut.ee/gplink/l/vNllutJ_RA
Pathway query: https://biit.cs.ut.ee/gplink/l/QRJj6CRVRY

Invasion and Metastasis:
Gene list query: https://biit.cs.ut.ee/gplink/l/0Ut1518XTh
Pathway query: https://biit.cs.ut.ee/gplink/l/z2PLcCNOQX

Proliferative Signaling:
Gene list query: https://biit.cs.ut.ee/gplink/l/VFDInpBoS4
Pathway query: https://biit.cs.ut.ee/gplink/l/Ct_XekRMR7

Suppression of Growth:
Gene list query: https://biit.cs.ut.ee/gplink/l/mZSttq9mTx
Pathway query: https://biit.cs.ut.ee/gplink/l/9Gk2V3aGS5

Tumour Promoting Inflammation:
Gene list query: https://biit.cs.ut.ee/gplink/l/Da-v6WC_Ss
Pathway query: https://biit.cs.ut.ee/gplink/l/ZUnEyQM6S0

2.6 Workflow Diagram

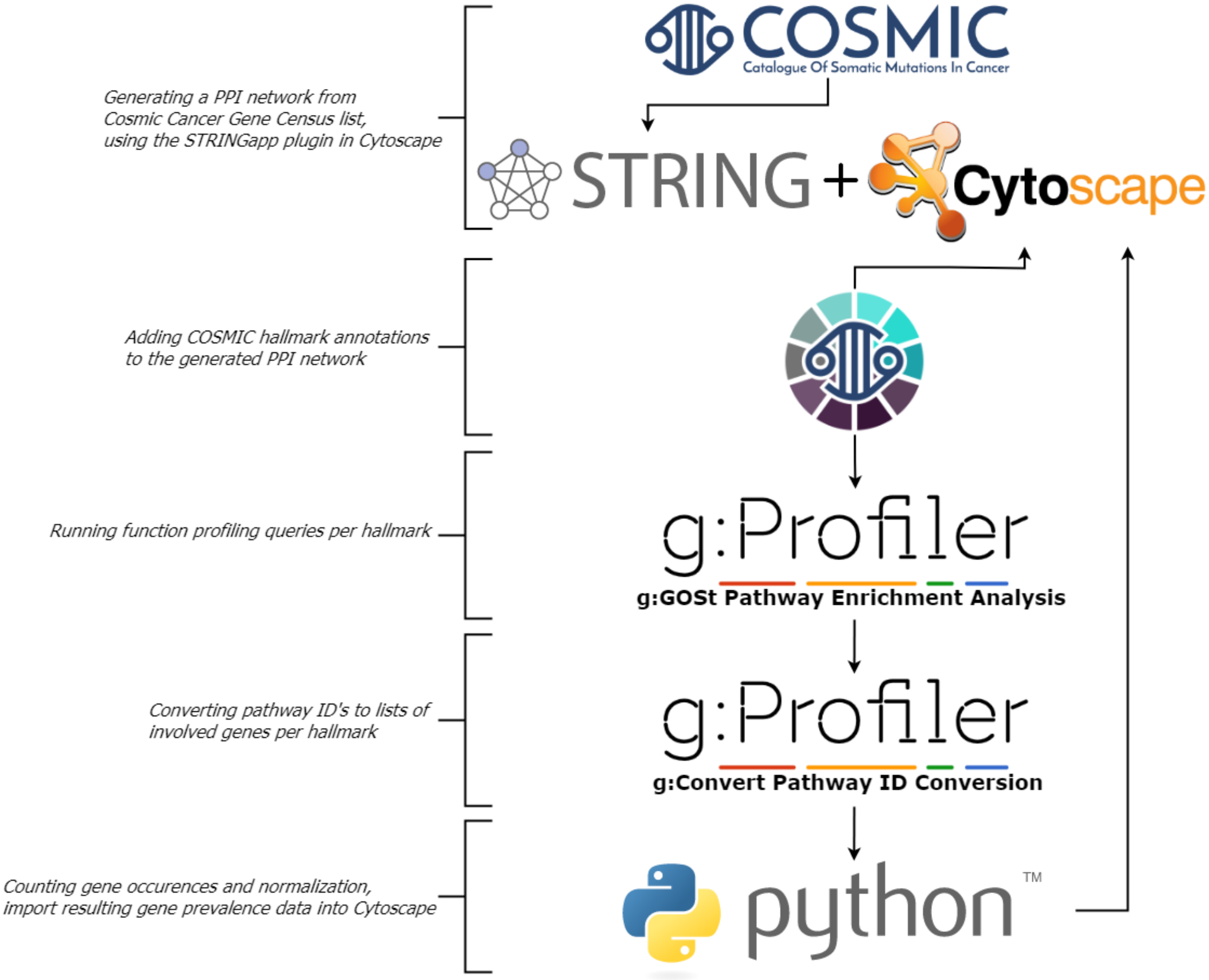

Generating a PPI network from
Cosmic Cancer Gene Census list,
using the STRINGapp plugin in Cytoscape
COSMIC
Catalogue Of Somatic Mutations In Cancer
STRING + Cytoscape
Adding COSMIC hallmark annotations
to the generated PPI network
Running function profiling queries per hallmark
g:Profiler
g:GOSt Pathway Enrichment Analysis
Converting pathway ID's to lists of
involved genes per hallmark
g:Profiler
g:Convert Pathway ID Conversion
Counting gene occurences and normalization,
import resulting gene prevalence data into Cytoscape
python
TM

Figure 3: Workflow diagram of network generation, data curation, and data integration.

3 Results

3.1 Network Topology and Hallmark Metrics

For the analysis of basic network parameters, the Cytoscape plugin 'NetworkAnalyser' was used.

The 10 nodes with the highest degree can be found in the table below. These 10 genes are connected to 66 hallmark annotations, which indicates that a lot of research in the context of cancer has already been done on these genes.

	Gene Name	Degree
1.	TP53	187
2.	AKT1	135
3.	HRAS	119
4.	EP300	118
5.	MYC	115
6.	STAT3	111
7.	PIK3CA	107
8.	SRC	101
9.	MAPK1	98
10.	KRAS	97

The top 10 genes with highest **betweenness** can be found in the table below, rounded to three decimals.

	Gene Name	Betweenness
1.	TP53	0.148
2.	EP300	0.053
3.	AKT1	0.047
4.	MYC	0.037
5.	CTNNB1	0.036
6.	STAT3	0.032
7.	MAPK1	0.031
8.	EGFR	0.030
9.	HRAS	0.030
10.	JUN	0.027

The top 10 genes with highest **closeness** can be found in the table below, rounded to three decimals.

	Gene Name	Closeness
1.	TP53	0.545
2.	AKT1	0.505
3.	MYC	0.502
4.	HRAS	0.496
5.	EP300	0.494
6.	CTNNB1	0.488
7.	PTEN	0.486
8.	STAT3	0.486
9.	KRAS	0.484
10.	EGFR	0.482

After removing duplicates from the above three tables, the hubs of the network were established as follows:

- **AKT1**
- **CTNNB1**
- **EGFR**
- **EP300**
- **HRAS**
- **JUN**
- **KRAS**
- **MAPK1**
- **MYC**
- **PIK3CA**
- **PTEN**
- **SRC**
- **STAT3**
- **TP53**

Of these 14 hubs, 3 have not been annotated with any hallmarks (**JUN, SRC, STAT3**). Therefore the focus for new hallmark predictions will be on these three genes.

The betweenness values and closeness values of all nodes were plotted against each other in a scatter plot as seen in figure 4.

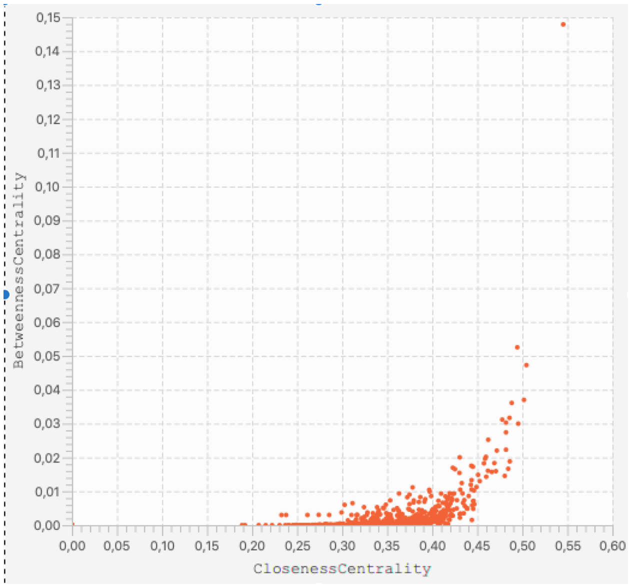


Figure 4: Scatterplot of all nodes Betweenness values and Closeness values.

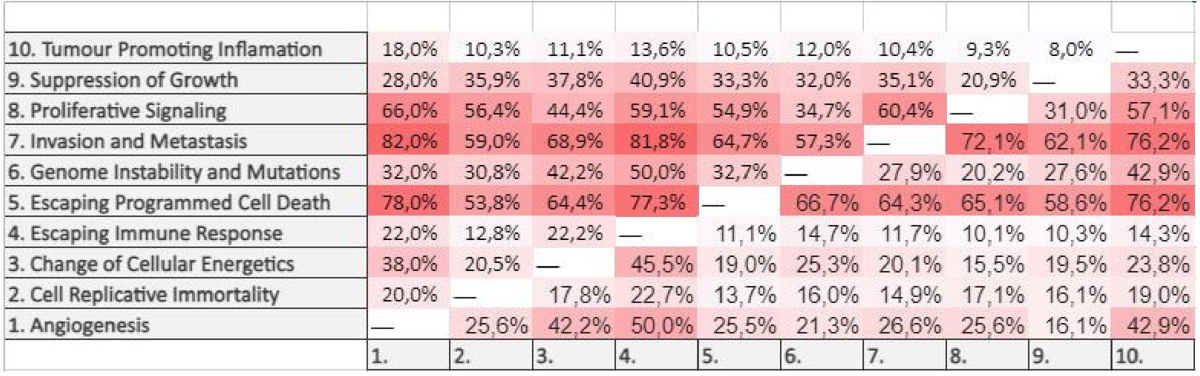
When fitting a linear regression model using the least squares method, this does not result in an accurate model. There appears to be some positive exponential correlation between these two measures. The same was found when plotting the degree against closeness, and when plotting degree against betweenness. This indicates these three measures are closely related and can all tell us something about the hubs of the network. This can also be concluded from the fact that there exists substantial overlap between the top 10 genes of these measures.

The table below shows an overview of the distribution of all hallmark annotations in both the full network and the 'pathway score >0.5' subnetwork.

Hallmark	**# of annotations**	**% of total**	**% in subnetwork**
Angiogenesis	50/717	6.97%	33.33%
Cell Replicative Immortality	39/717	5.44%	23.81%
Change of Cellular Energetics	45/717	6.28%	28.57%
Escaping Immune Response	22/717	3.07%	21.43%
Escaping Programmed Cell Death	153/717	21.34%	61.90%
Genome Instability and Mutations	75/717	10.46%	33.33%
Invasion and Metastasis	155/717	21.62%	64.29%
Proliferative Signaling	129/717	17.99%	42.86%
Suppression of Growth	87/717	12.13%	30.95%
Tumour Promoting Inflammation	21/717	2.93%	9.52%

Genes can be annotated with multiple hallmarks. A total of 272 genes has been annotated with one or more hallmarks. There is a total of 50 genes which have no connected edges, meaning they are not connected to the network. These 50 genes (7% of all genes) which are not connected to the network only contain 1,68% of all hallmark annotations within the network.This means that relatively not much hallmark annotation data is 'lost'.

Figure 5 shows an overview of overlapping annotations between all hallmarks. The bottom row should be read first, e.g. 20.0% of genes annotated with the 'Angiogenesis' hallmark are also annotated with the 'Cell Replicative Immortality' hallmark. The first thing to notice is the high percentage of overlap between all hallmarks and the 'Escaping Programmed Cell Death' and 'Invasion and Metastasis' hallmarks. These are also the most annotated hallmarks, it can be concluded that these are the most dominant hallmarks in the network.



10. Tumour Promoting Inflamation	18,0%	10,3%	11,1%	13,6%	10,5%	12,0%	10,4%	9,3%	8,0%	—
9. Suppression of Growth	28,0%	35,9%	37,8%	40,9%	33,3%	32,0%	35,1%	20,9%	—	33,3%
8. Proliferative Signaling	66,0%	56,4%	44,4%	59,1%	54,9%	34,7%	60,4%	—	31,0%	57,1%
7. Invasion and Metastasis	82,0%	59,0%	68,9%	81,8%	64,7%	57,3%	—	72,1%	62,1%	76,2%
6. Genome Instability and Mutations	32,0%	30,8%	42,2%	50,0%	32,7%	—	27,9%	20,2%	27,6%	42,9%
5. Escaping Programmed Cell Death	78,0%	53,8%	64,4%	77,3%	—	66,7%	64,3%	65,1%	58,6%	76,2%
4. Escaping Immune Response	22,0%	12,8%	22,2%	—	11,1%	14,7%	11,7%	10,1%	10,3%	14,3%
3. Change of Cellular Energetics	38,0%	20,5%	—	45,5%	19,0%	25,3%	20,1%	15,5%	19,5%	23,8%
2. Cell Replicative Immortality	20,0%	—	17,8%	22,7%	13,7%	16,0%	14,9%	17,1%	16,1%	19,0%
1. Angiogenesis	—	25,6%	42,2%	50,0%	25,5%	21,3%	26,6%	25,6%	16,1%	42,9%
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.

Figure 5: Hallmarks overlap heatmap.

3.2 Generated Network and Visualization

A snapshot of the PPI network as generated by STRINGapp can be seen in figure 6. The first observation is that there exist 50 nodes without outgoing edge and are therefore unconnected to the central network. The network shows a periphery where nodes have a lower degree compared to the center of the network which is more dense. The main network properties can be found in 3.3.

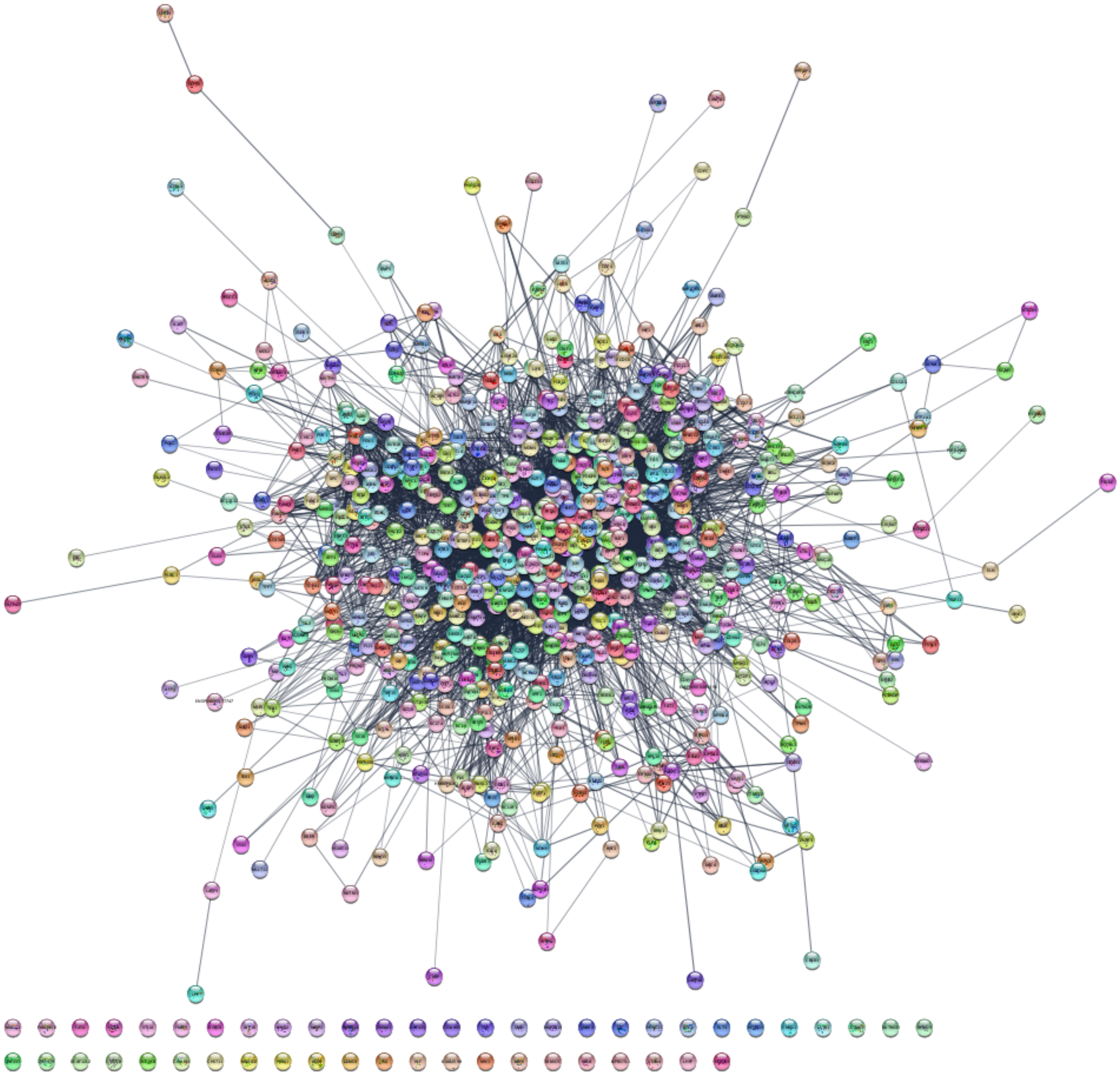


Figure 6: Unedited PPI network as produced by STRINGapp.

Figure 7 shows the hubs of the network with a red border. The hubs were determined by betweenness, closeness, and degree and are therefore located centrally in the network. This slightly zoomed in image also captures the many edges and high connectivity of the network.

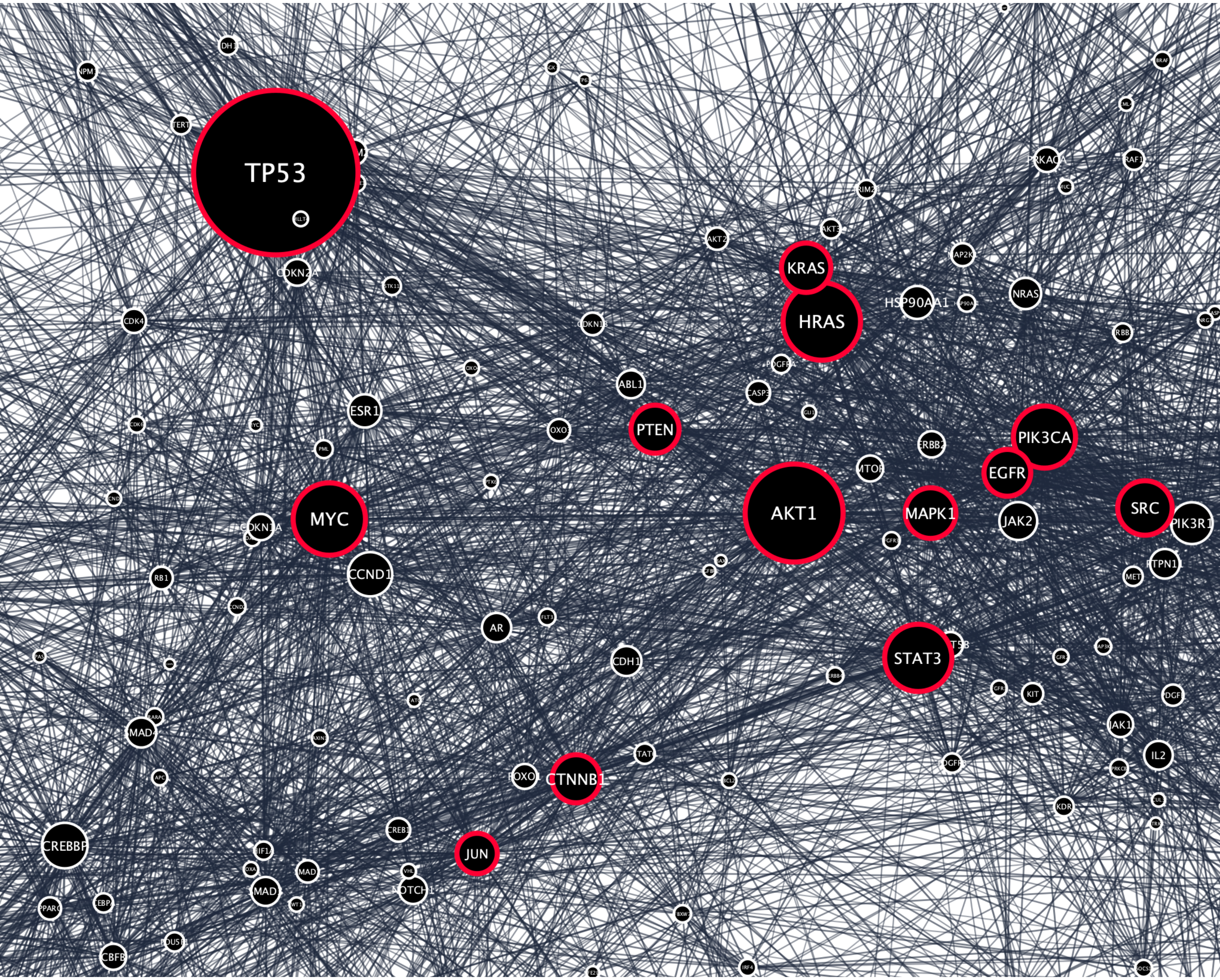


Figure 7: The PPI network as visualized in Cytoscape. The hubs as determined by centrality analysis in 3.1 are visualized with a red border.

Next, using Omics Visualizer plugin [13], the hallmark annotations were visualized in a pie chart within each node. A red slice represents a hallmark annotation, corresponding to the legend in figure 8. The resulting visualization can be seen in figure 8 on the next page.

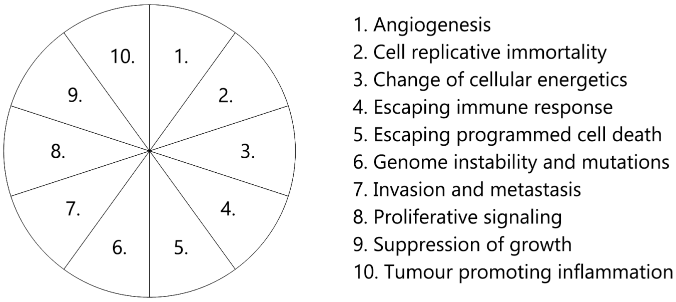

10.
1.
9.
2.
8.
3.
7.
4.
6.
5.
1. Angiogenesis
2. Cell replicative immortality
3. Change of cellular energetics
4. Escaping immune response
5. Escaping programmed cell death
6. Genome instability and mutations
7. Invasion and metastasis
8. Proliferative signaling
9. Suppression of growth
10. Tumour promoting inflammation

Figure 8: Hallmark annotation and pathway score distribution in the pie charts/donut charts as visualized on the nodes in figure 9.

Figure 9 shows the hallmark annotations visualized on the nodes. The are noteworthy differences in annotations between the hub genes. There exist three hub genes without any hallmarks annotations (JUN, SRC, STAT3). It would be very unlikely for hubs to not be involved in the expression of any hallmark. It seems that some annotations are missing for these central genes. The other hubs that have been annotated with at least one hallmark have an average number of 6.27 (69/11) hallmark annotations, whereas all non-hub genes have an average number of 2.85 (775/272) hallmark annotations. Thus, assuming that JUN, SRC, STAT3 are indeed missing one or multiple hallmark annotations, hub genes are structurally annotated with more hallmarks than non-hub genes. This could be because hubs in general are involved in a high number of hallmarks. Bias could also be part of the explanation for this discrepancy, as hub genes may have been more intensively researched. This would mean that there is naturally a larger body of evidence to be found for possible hallmark annotations for hub genes.

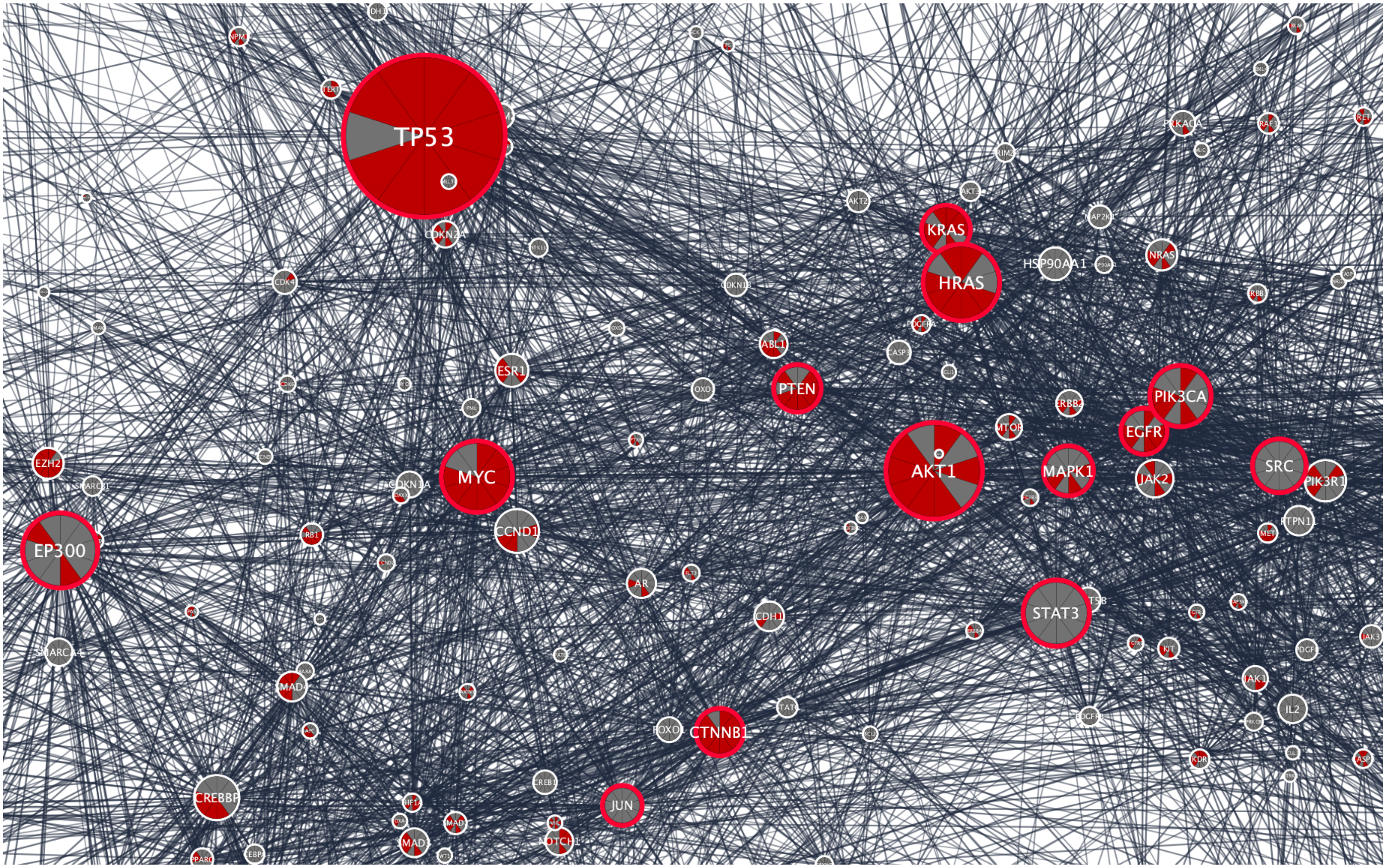


Figure 9: Hallmark annotations visualized. A red pie segment represents a hallmark annotation corresponding to the pie chart in figure 8.

After the pathway scores were added to the network, a subnetwork was created by using a node filter which selects all genes that have a pathway prevalence score >0.5 for at least one hallmark. This results in an overview of genes that are significantly involved in pathways that are connected to at least one of the ten hallmarks. The resulting subnetwork can be seen in figure 10 below. The interpretation of this subnetwork can be found in 3.3.

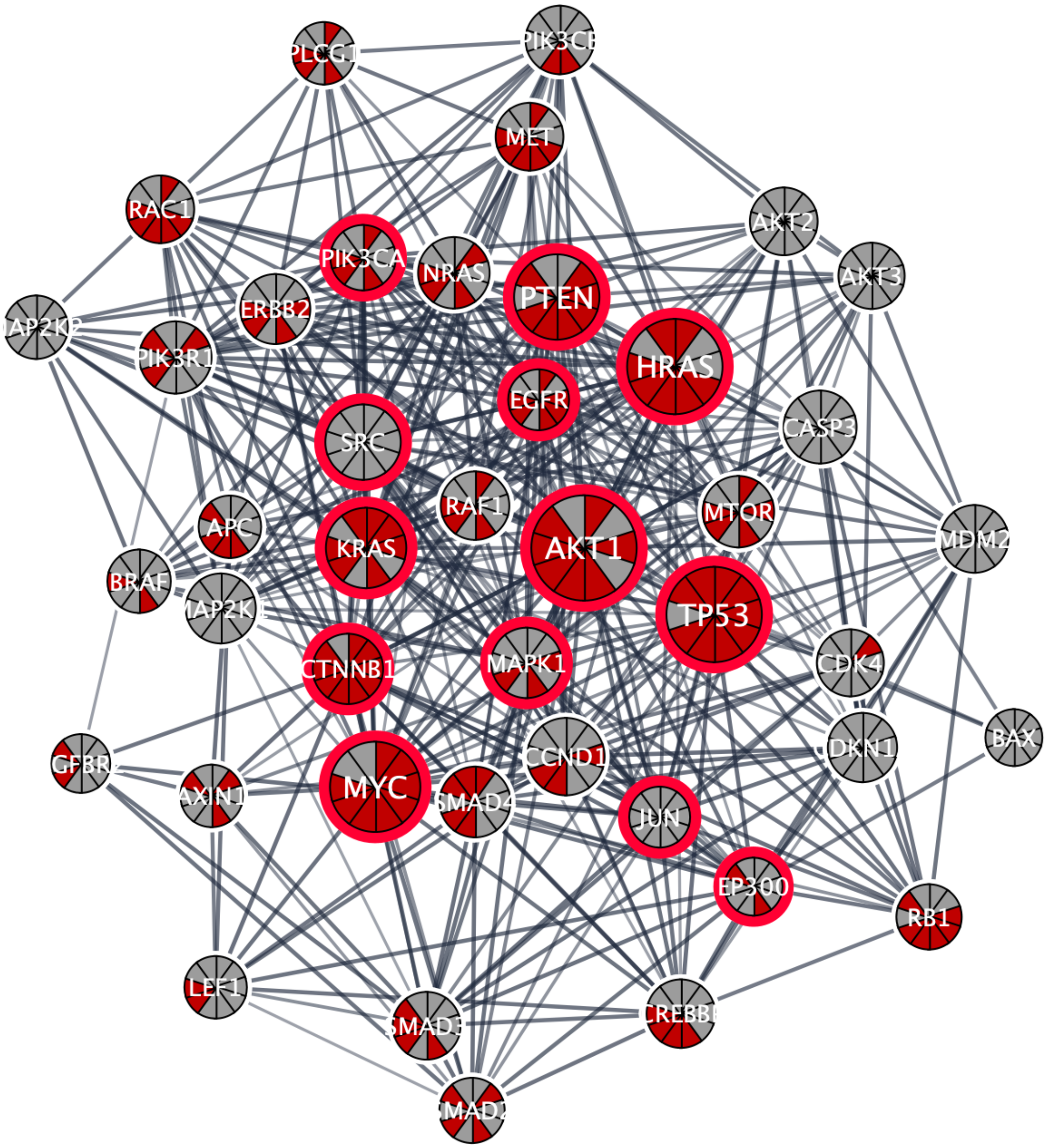


Figure 10: Subnetwork of genes which have a pathway score of >0.5 for at least one of the hallmarks.

3.3 Pathway Enrichment Results

The network statistics resulting from NetworkAnalyser are compared in the table below. The subnetwork consists of genes that have a pathway score >0.50 for any hallmark.

	Full Network	Subnetwork: Pathway score >0.50
Network Size	717	42
Network Diameter	8	3
Network Radius	4	2
Clustering Coefficient	0,414	0,698
Network Density	0,028	0,469
Characteristic Path Length	2,885	1,544
Network Centralization	0,254	0,378
Average degree	18,582	19,238

From these statistics it can be concluded that the subnetwork containing only genes with at least one pathway score >0.5 forms a denser and more strongly interconnected network than the full PPI network. This indicates that there exists substantial evidence for interactions between genes annotated with different hallmarks. The table below also shows that the genes that were identified as hubs in the full PPI network, are mostly still the most connecting genes in the subnetwork as well. This shows that the hub genes are an important bridging factor for these interactions between genes annotated with different hallmarks, and that the hub genes themselves are connected to genes with different hallmark annotations.

Top 15 Genes by Degree in Subnetwork		
Gene Name	Degree	Hub in full PPI network?
AKT1	34	Yes
HRAS	32	Yes
TP53	32	Yes
MYC	31	Yes
PTEN	30	Yes
KRAS	29	Yes
SRC	28	Yes
CTNNB1	27	Yes
MAPK1	27	Yes
CCND1	26	No
PIK3CA	26	Yes
CASP3	23	No
EGFR	23	Yes
JUN	23	Yes

Figure 11 shows the collection of genes which have a pathway score >0.5 for at least one gene, with added pathway score visualization as described in 2.5. There are multiple genes (AKT2, AKT3, MAP2K1, MAP2K2) which have considerably raised pathway scores but have not been annotated with any hallmark. There are also genes with annotations for hallmarks for which these genes have a relatively low pathway score (LEF1, APC, AXIN1). This means that a hallmark annotation does not always correspond with a high pathway score for that hallmarks. Vice versa, a low pathway score for a certain hallmark does not necessarily mean that no annotation for that hallmark exists.

Furthermore it is worth noting that many hub genes have significantly elevated pathway scores for multiple or even all hallmarks. This could indicate that these genes play a central role in the expression of multiple or even all hallmarks, in which case these genes would be primary catalysts for carcinogenesis in general.

The gene prevalence graphs, which can be found in the appendix 6.2, show which genes are the most prevalent in pathways connected to a certain hallmark. The graphs show the top 30 genes in terms of prevalence in overrepresented pathways queried from the subset of COSMIC genes annotated with a certain hallmark. There is a difference in composition and placement of the top 30 genes per hallmark. This indicates that different genes can play a more or lesser prominent role in the expression of certain hallmarks than in other hallmarks. More notable are the considerable similarities. Many genes are in the top 30 prevalence for multiple or even all hallmarks. These include some of the hub genes (AKT1, HRAS, TP53, PIK3CA, MAPK1, KRAS, PIK3R1), but also some non-hub genes (RAF1, NRAS, MTOR, RAC1 PIK3CB, PIK3R1, AKT2, AKT3, MAP2K2, MAP2K1). This could mean that the latter are maybe not hub genes from a centrality analysis point of view, but are hub-genes from a hallmark point of view as they have strong presence in pathways connected to different hallmarks.

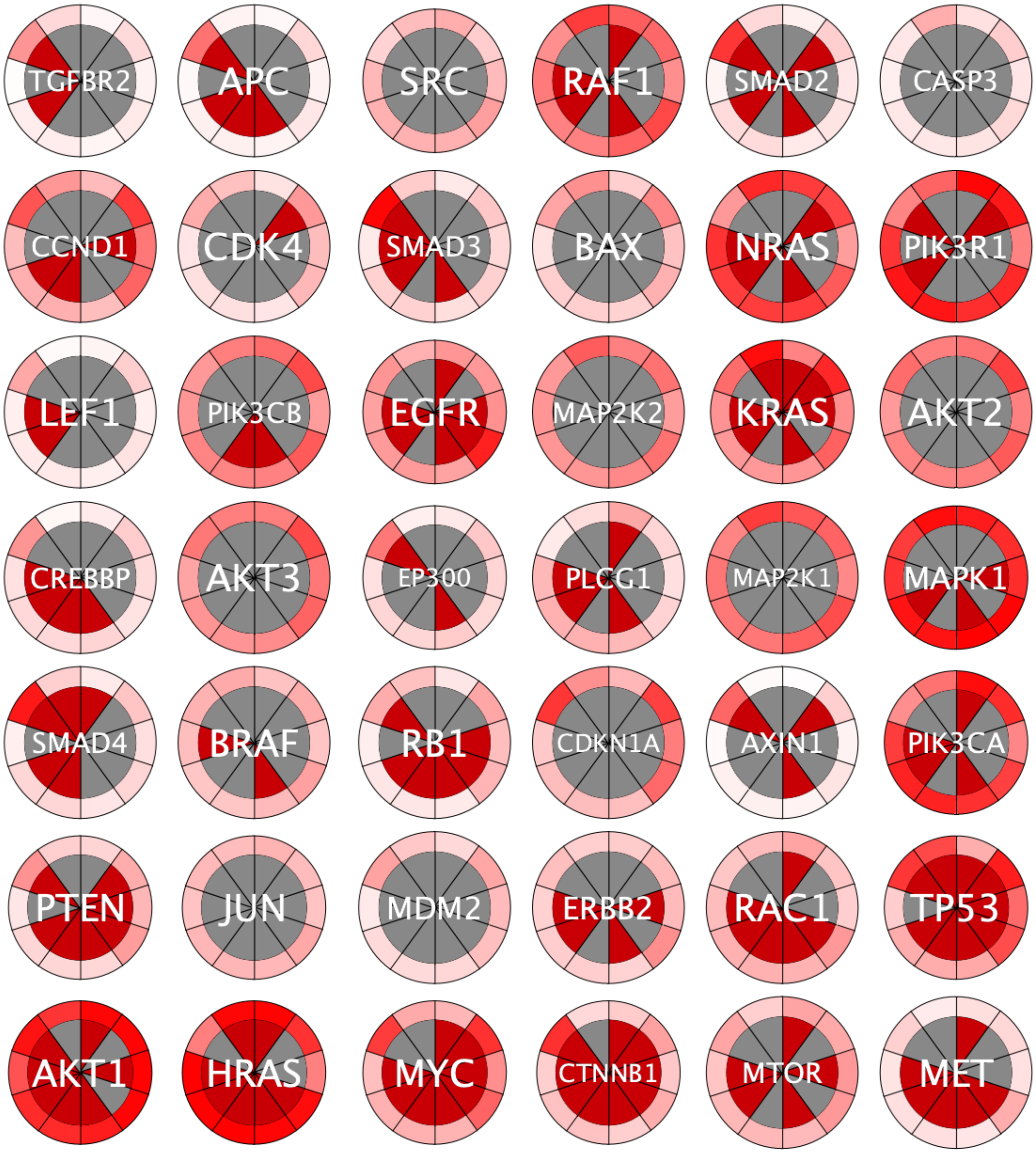


Figure 11: The same set of genes as shown in figure 10, with a visual representation of the hallmark annotations and pathway scores.

Another interest was to see if there are any genes with a substantially high pathway score for a certain hallmark, that are not annotated with this hallmark. Per hallmark, it was derived which genes that are not annotated with a certain hallmark, do have a pathway score >0.75 for that hallmark. This threshold was chosen in order to identify genes with a strong indication of involvement for some hallmark. The list can be seen below:

Angiogenesis	
Gene Name	Pathway Score
MAP2K1	0.7829
MAPK1	0.9314
NRAS	0.8514
PIK3R1	0.9829

Genome Instability and Mutation	
Gene Name	Pathway Score
MAPK1	0.9522
PIK3R1	0.9474
PIK3CA	0.9139
NRAS	0.8469

Cell Replicative Immortality	
Gene Name	Pathway Score
AKT1	0.9877
AKT2	0.8395
AKT3	0.8148
CCND1	0.8519
CDKN1A	0.8395
HRAS	0.8765
MAP2K1	0.7531
MAPK1	0.8889
PIK3CA	0.8889
PIK3CB	0.8148

Suppression of Growth	
Gene Name	Pathway Score
MAPK1	0.8507
CCND1	0.791
CDKN1A	0.8657
MYC	0.8507

Escaping Immune Response	
Gene Name	Pathway Score
AKT1	0.9886
MAPK1	0.9886
AKT3	0.75
CCND1	0.75
AKT2	0.7727
PIK3R1	0.8977
PIK3CA	0.8409
MAP2K1	0.8068
KRAS	0.7614
RAF1	0.8182
NRAS	0.7727
PIK3CB	0.7727

Escaping Programmed Cell Death	
Gene Name	Pathway Score
PIK3R1	0.8837
MAP2K1	0.7581

Proliferative Signaling	
Gene Name	Pathway Score
MAPK1	0.8714
PIK3R1	0.8667

Tumour Promoting Inflammation	
Gene Name	Pathway Score
AKT1	0.8667
MAPK1	0.9667
PIK3R1	0.75
MAP2K1	0.85
RAF1	0.85
MAP2K2	0.7667
NRAS	0.9

Invasion and Metastasis:
None

Change of Cellular Energetics:
None

On 45 occasions a gene has a pathway score >0.75 for a certain hallmark, without being annotated with that hallmark. These tables could be leading for future hallmark annotation predictions.

3.4 Unannotated Genes in Subnetwork

The following genes are included in the 'pathway score >0.5' subnetwork, but have no single hallmark annotation:

- **AKT2**
- **AKT3**
- **BAX**
- **CASP3**
- **CDKN1A**
- **JUN**
- **MAP2K1**
- **MAP2K2**
- **MDM2**
- **SRC**

It would be somewhat unexpected to see 10 unannotated genes in this subnetwork, considering their significant involvement in pathways for at least one hallmark. Most of these genes have remarkably high pathway scores for multiple hallmarks, most notable being AKT2, AKT3, MAP2K1, MAP2K2, CDKN1A. It seems therefore that a lot of hallmarks annotations are missing.

3.5 Pathways In Cancer

Out of the 272 genes that have been annotated with one or more hallmarks, 187 genes are not connected to KEGG:5200. In other words, only 85/272 annotated genes occur in the pathway. This reveals that this pathway, as a general gene and pathway collection for cancer, is not complete. This could be because not all hallmarks are (equally) represented in the pathway. On the other hand, 13 of the 14 hubs are included in the pathway, indicating that these hubs seem to escape this hallmark selectivity found in Pathways In Cancer. Combining this with the observation of high pathway scores for many hallmarks, and with the plethora of hallmark annotations for multiple hubs, it seems that the hubs play a more central role amid the pathways involved in different hallmarks.

3.6 Hallmark predictions for unannotated hubs.

3.6.1 The SRC gene

The SRC gene is one of the hub genes which does not have any hallmarks annotated. Considering its betweenness, closeness, degree, and the fact that is is part of the 'pathway score >0.5 subnetwork', it is highly unlikely for this gene to not be involved in the expression of any of the hallmarks. The SRC gene has been brought in connection with cancer through its involvement in signal conduction [11].

The SRC gene has the following pathway scores:

SRC pathway scores	
Angiogenesis	**0.5029**
Cell Replicative Immortality	0.358
Change of Cellular Energetics	0.437
Escaping Immune Response	**0.5682**
Escaping Programmed Cell Death	0.4884
Genome Instability and Mutation	0.4833
Invasion and Metastasis	0.4833
Proliferative Signaling	0.4571
Suppression of Growth	0.3284
Tumour Promoting Inflammation	0.4333

In figure 12, the 'Pathways in cancer' (KEGG:5200) has been visualized in the 'pathway score >0,5' subnetwork. The first observation is that the SRC gene is the only gene that is not involved in the 'Pathways in cancer' pathway. Furthermore, the SRC gene has not been annotated with any hallmarks. There exists edges between SRC and 28 of 41 other genes in the subnetwork, showing high interconnectivity in the subnetwork. The highest pathway scores for SRC are Angiogenesis (0.5029) and Escaping Immune Response (0.5682). The prevalence of annotations of these two hallmarks in genes connected to the SRC gene was examined:

- 13/28 (46.4%) genes connected to SRC have been annotated with the Angiogenesis hallmark, whereas only 6.97% of genes in the full network have been annotated with this hallmark
- 8/28 (28.6%) genes connected to SRC have been annotated with the Escaping Immune Response hallmark, whereas only 3.07% of genes in the full network have been annotated with this hallmark

The SRC gene product exhibits significant interaction with other gene products annotated with these two hallmarks. In the subnetwork, 14/42 (33.33%) genes have been annotated with Angiogenesis, and 9/42 (21.43%) have been annotated with Escaping Immune Response. This means that the SRC gene is connected to 13/14 genes annotated with Angiogenesis and 8/9 genes annotated with Escaping Immune Response in the subnetwork. It can be concluded that these hallmarks are strongly enriched in genes connected to the SRC gene.

The pathway scores and considerably enriched connections to genes with these two hallmarks provide multiple arguments for the involvement of SRC in these hallmarks. These observations can be interpreted as a indication that the SRC gene is involved in the expression of these two hallmark characteristics, which will be the final prediction.

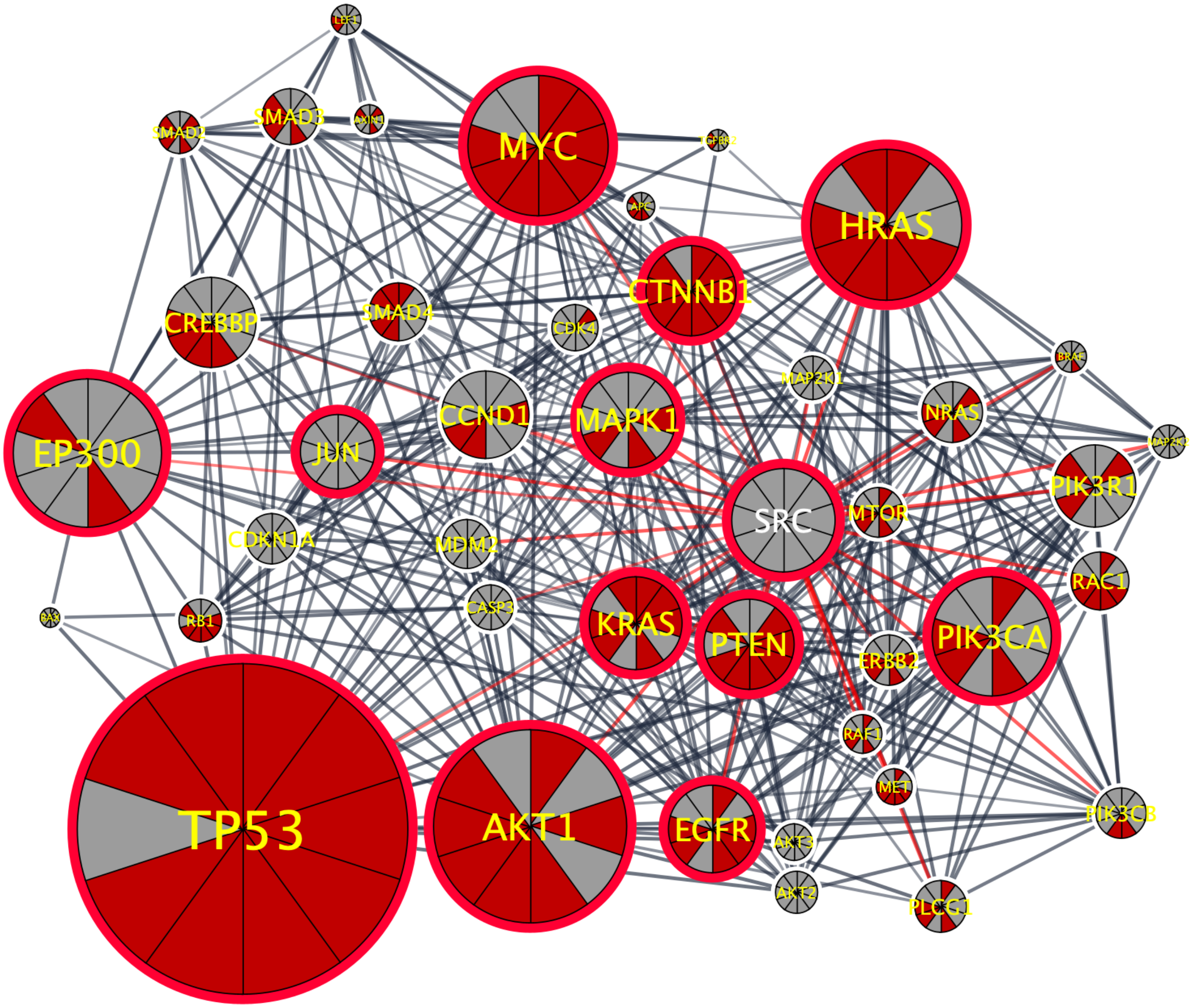


Figure 12: The 'pathway score >0.5' subnetwork. Genes involved in the 'Pathways In Cancer' pathway (KEGG:5200) are visualized with a yellow label. Others with a white label. All edges connected to the SRC gene are highlighted in red.

3.6.2 The STAT3 gene

The 'pathway score >0.5' subnetwork contains 13 of the 14 hubs. The STAT3 gene is the only hub gene which is not included in the 'pathway score >0.5' subnetwork. STAT3 does however have pathway scores for multiple pathways which approach the threshold score of 0.5 as seen in the table below. The pathway scores for Escaping Immune Response and Tumour Promoting Inflammation are very close to the threshold score of 0,5. This shows that the STAT3 gene was only just not included in the subnetwork, because of the decision for the cutoff value.

STAT3 pathway scores	
Angiogenesis	0.4171
Cell Replicative Immortality	0.4074
Change of Cellular Energetics	0.3926
Escaping Immune Response	**0.4773**
Escaping Programmed Cell Death	0.4
Genome Instability and Mutation	0.3876
Invasion and Metastasis	0.3876
Proliferative Signaling	0.3875
Suppression of Growth	0.3881
Tumour Promoting Inflammation	**0.4833**

STAT3 is involved in the 'Pathways In Cancer' pathway (KEGG:5200), but has no hallmark annotations. However, with a degree of 111 in the full PPI-network, it is the 6th most connected node. Of those 111 genes connected to STAT3, 59 have been annotated with some hallmark (53.15%), whereas 272/717 (37.94%) of genes in the full network have been annotated with one or multiple hallmarks. This shows that the STAT3 gene is significantly more connected with genes annotated with at least one hallmark. This could be an indicator that the STAT3 is also involved in the expression of certain hallmarks and that possibly STAT3 is lacking one or multiple hallmark annotations.

Looking at the genes connected to STAT3, 10/111 (9%) have been annotated with Tumour Promoting Inflammation, whereas only 21/717 (2.93%) of genes in the full network are annotated with this hallmark. Furthermore, 15/111 (13.5%) of the genes connected to STAT3 are annotated with Escaping Immune Response, whereas only 22/717 (3.07%) of genes in the full network are annotated with this hallmark. In other words, the STAT3 genes is connected to 10/21 genes annotated with Tumour Promoting Inflammation, and to 15/22 genes annotated with Escaping Immune Response. This shows that the STAT3 gene is considerably more connected with genes annotated with these two hallmarks. Combining these observations, it is predicted that this hallmark should be annotated with the Tumour Promoting Inflammation and Escaping Immune Response hallmarks.

3.6.3 The JUN gene

The JUN gene has the following pathway scores:

JUN pathway scores	
Angiogenesis	0.4686
Cell Replicative Immortality	0.4321
Change of Cellular Energetics	0.4
Escaping Immune Response	**0.5227**
Escaping Programmed Cell Death	0.4558
Genome Instability and Mutation	0.4593
Invasion and Metastasis	0.4593
Proliferative Signaling	0.4
Suppression of Growth	0.4478
Tumour Promoting Inflammation	0.4333

For JUN, there is one pathway score that is somewhat higher than all others. The highest pathway score is 0.5227 for Escaping Immune Response. All other pathway scores are within the narrow range of 0.4 and 0.4686, indicating that the JUN gene might play a more general role in carcinogenesis. The JUN gene is connected to 7 out of 9 genes that are annotated with the Escaping Immune Response hallmark in the subnetwork.

All 23 nodes connected to JUN are also connected to TP53 and MYC. Furthermore, 21/23 genes are connected to AKT1, and 20/23 to CCND1. TP53, MYC, AKT1, and CCND1 are all genes with raised pathway scores for all hallmarks. This makes it difficult to assign a plausible hallmark prediction for Escaping Immune Response, as JUN interacts with many other genes with differing hallmark annotations, and has lesser substantial pathway score anomalies.

4 Conclusions

There is a collection of genes with a raised pathway score for multiple or even all hallmarks. This could indicate that these genes play a central role in the expression of multiple or even all hallmarks, and thus a central role in carcinogenesis in general. This collection of genes with high pathway scores for multiple hallmarks includes many of the hubs of the network that were identified through centrality analysis. Some genes with high pathway scores for multiple hallmarks were not identified as a hub through centrality analysis. Nonetheless, these genes could still be considered hubs from a cancer induction point of view.

On average, the hubs are annotated with significantly more hallmarks than non hub-genes (6.27 versus 2.85). Furthermore, the subnetwork containing only genes with at least one pathway score>0.5 forms a denser and more strongly interconnected network than the full PPI network. This indicates that there exists significant evidence for interactions between genes annotated with different hallmarks, mainly for the hubs of the network.

High pathway scores for certain unannotated genes with a high connectivity are an indication for missing hallmark annotations, for which some predictions have been made. The different g:Profiler Pathway Enrichment queries resulted in different scores per hallmark, confirming that genes can be more strongly involved in some hallmarks than in others. On the other hand, the top 30 gene prevalence graphs per hallmark also show many recurring genes, including many of the hubs. This could mean that separate hallmarks may not be expressed through entirely separate pathways within a cell. The expression of all different hallmarks in conjunction, with a cancerous cell as a result, may to some extent be connected through a limited collection of highly interconnected genes.

Additionally, some hallmark predictions were done for unannotated hub genes. Three of the identified hub genes have not been annotated with any hallmark. Considering their centrality statistics indicating a high connectivity, and considering the pathway scores, it would be very unlikely for these hubs to not be involved in the expression of any hallmark. Based on their pathway scores and the distribution of hallmarks of neighbouring genes, a few hallmark predictions were done. This analysis resulted in the following predictions:

- SRC: Angiogenesis + Escaping Immune Response
- STAT3: Tumour Promoting Inflammation + Escaping Immune Response

Summarizing, a collection of highly connected genes seem to be involved in the expression of multiple if not all hallmarks. Therefore a distinction should be made between genes that are involved in the expression of a few hallmarks and genes that are involved in the expression of an abundance of hallmarks. Separate hallmarks may not be expressed through entirely separate pathways within a cell. The expression of all different hallmarks is to some extent interconnected through a subset of genes, which are significantly enriched in different pathways connected to different hallmarks.

5 Discussion and Further Research

Hallmark Annotation Bias

Only hallmark annotated genes were used for retrieval of pathway data. This means that any conclusions and predictions were indirectly inferred from previously assigned annotations. All inference is based on known hallmarks annotations. Known hallmark annotations may favor certain genes or hallmarks, and this could create a bias. Because only 'known' data and annotations were used, this may reinforce the narrative of findings that have already been done.

Pathway Score Bias

For the calculation of the pathway scores, each gene occurrence in a pathway for a hallmark was weighed equally. There are however very general pathways such as 'Pathways In Cancer', and more specific pathways such as 'MAPK signaling pathway'. Some of these more specific pathways may be better describers of the functional expression of a certain hallmark. Gene occurrences in pathways that are more strongly correlated to a hallmark should maybe therefore be weighed higher in the calculation of pathway scores. This however introduces a problem of arbitrariness. It would be hard to decide which pathways should be considered more representative of the expression of a certain hallmark, and it would be even harder to then assign scores to these different pathways. A possible method to be more selective in pathway enrichment would be to only use pathways that are proven to have some connection to the hallmark.
Furthermore, each pathway has an assigned P-value describing the significance of the enriched pathway in the queried gene list. Gene occurences in pathways with a lower P-value should maybe be valued higher compared to gene occurences in pathways with a higher P-value.

Analysis of Hub Genes

The definition of a hub gene from a hallmark point of view should maybe be expanded. There is a collection of genes that have raised pathway scores for multiple or all hallmark, which were not identified as hub genes. However, if these genes are involved in the expression of all hallmarks, it could mean that they play a central role in the expression of all these hallmarks and therefore a central role in carcinogenesis in general. Therefore these genes could be considered hubs, not from a network topology point of view, but from a cancer induction point of view. Another possibility is that more centrality measures (Stress / Clustering Coefficient) should have been considered for the identification for hubs. This could have possibly led to a greater number of hubs and may have included more genes that have raised pathway scores for all hallmarks.

Future Research

Future research could consist of trying a method for weighing gene occurrences in different pathways. This may produce more accurate pathway scores. Another possibility would be to do a more exact analysis of which hallmarks are expressed through which pathways. Being able to better distinguish which combinations of pathways are responsible for the expression of which hallmark could also help us calculate more accurate pathway scores, by having a more specific pathway selection per hallmark.

It could also be interesting to revisit this thesis at a later time, when more hallmark annotations have been added to the Cosmic Cancer Gene Census. Comparing these new hallmark annotations to the predictions in this thesis could also either validate or disprove the method of using pathway scores for predicting new hallmarks annotations. Genes that have a high pathway score for a certain hallmarks but are currently unannotated with that hallmark would also be of particular interest.

References

[1] H. Yanai et al. A. Takaoka, S. Hayakawa. Integration of interferon-α/β signaling to p53 responses in tumour suppression and antiviral defence. *Nature*, 424:516–523, 2003.

[2] Boyer S Russell RB. Apic G, Ignjatovic T. Illuminating drug discovery with biological pathways. *FEBS Letters*, 579(8):1872–1877, 2005.

[3] Ramírez F. Schelhorn S.E. Lengauer T. Albrecht M. Assenov, Y. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 2008.

[4] Lukey M. J. Cerione R. A. & Locasale J. W. Cluntun, A. A. Glutamine metabolism in cancer: Understanding the heterogeneity. *Trends in cancer*, 3:169–180, 2017.

[5] Giral E. Cetintas A. Civril A. Demir E. Dogrusoz, U. A layout algorithm for undirected compound graphs. *Information Sciences*, 179(7):980–994, 2009.

[6] Robert A. Weinberg Douglas Hanahan. The hallmarks of cancer. *Cell*, 100 (1):57–70, 2000.

[7] Robert A. Weinberg Douglas Hanahan. Hallmarks of cancer: The next generation. *Cell*, 144 S:646–674, 2011.

[8] Heath LS. Eid FE, ElHefnawi M. Denovo: virus-host sequence-based protein-protein interaction prediction. *Bioinformatics*, 32(8):1144–50, 2016.

[9] Wang H. Gaur U. Little P. J. Xu J. Zheng W. Farhan, M. Foxo signaling pathways as therapeutic targets in cancer. *International journal of biological sciences*, 13(7):815–827, 2017.

[10] Antonio Sica Cecilia Garlanda Alberto Mantovani Francesco Colotta, Paola Allavena. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis*, 30(7):1073–1081, 2009.

[11] Yeatman T. Irby, R. Role of src expression and activation in human cancer. *Oncogene*, 19:5636–5642, 2000.

[12] Tanabe K. Kim R, Emi M. Cancer immunoediting from immune surveillance to immune escape. *Immunology*, 121(1):1–14, 2007.

[13] Morris JH Jensen LJ. Legeay M, Doncheva NT. Visualize omics data on networks with omics visualizer, a cytoscape app. *F1000Research*, 9:157, 2020.

[14] Stueker O Emili A Bader GD. Merico D, Isserlin R. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, 5:11, 2010.

[15] Johannes Ruscheinski Peng-Liang Wang Trey Ideker Michael E. Smoot, Keiichiro Ono. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27:431–432, 2011.

[16] Gubin M. M. Schreiber R. D.-& Smyth M. J. Mittal, D. New insights into cancer immunoediting and its three component phases–elimination, equilibrium and escape. *Current opinion in immunology*, 27:16–25, 2014.

[17] Jan Gorodkin Nadezhda T. Doncheva, John H. Morris and Lars J. Jensen. Cytoscape stringapp: Network analysis and visualization of proteomics data. *Journal of Proteome Research*, 18:623–632, 2019.

[18] Clark W. Oron T. et al. Radivojac, P. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10:221–227, 2013.

[19] Kuzmin I Arak T-Adler P Peterson H Vilo J. Raudvere U, Kolberg L. g:profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Research*, 47, W1:191–198, 2019.

[20] Weitzman SA. Shacter E. Chronic inflammation and cancer. *Oncology*, 16(2):217–26, 2002.

[21] Ozier O et al. Shannon P, Markiel A. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[22] Wright W. Shay, J. Hayflick, his limit, and cellular ageing. *Nature Reviews Molecular Cell Biology*, 1:72–76, 2000.

[23] Bacchetti S. Shay JW. A survey of telomerase activity in human cancer. *Science*, 324:1029–1033, 2009.

[24] Bamford S. Cole C.G. et al. Sondka, Z. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*, 18:696–705, 2018.

[25] Khoury M. P. & Bourdon J. C. Surget, S. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *OncoTargets and therapy*, 7:57–68, 2013.

[26] Lyon D et al. Szklarczyk D, Gable AL. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):607–613, 2019.

[27] Cantley L. C. & Thompson C. B. Vander Heiden, M. G. Understanding the warburg effect: the metabolic requirements of cell proliferation. *Eur J Cancer*, Apr;33(5):787–91, 1997.

[28] Kinzler K. Vogelstein, B. Cancer genes and the pathways they control. *Nature Medicine*, 10:789–799, 2004.

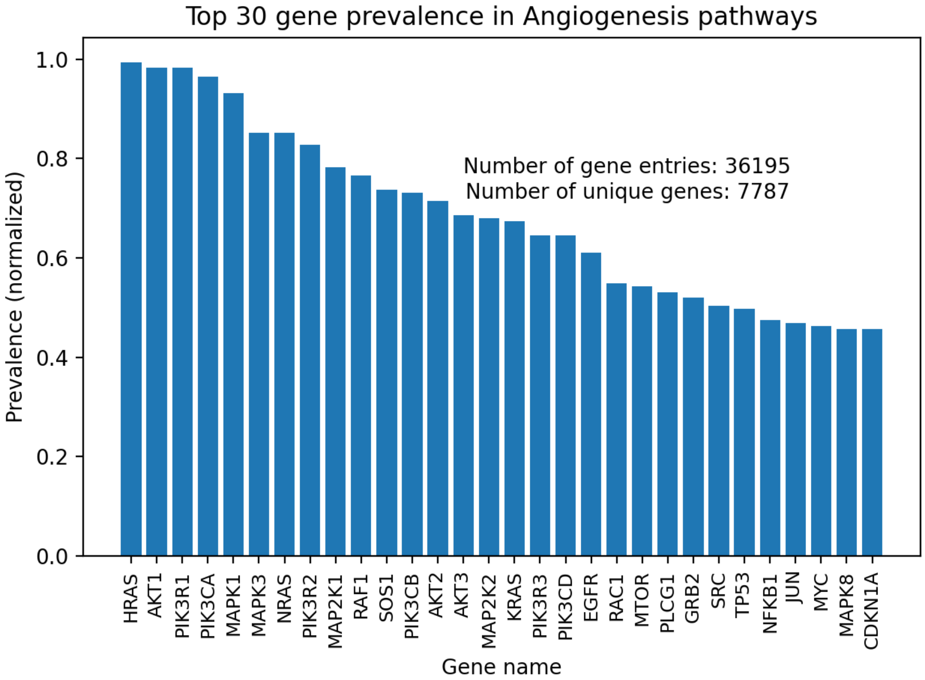
[29] Jaeggi D et al. von Mering C, Huynen M. String: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261, 2003.

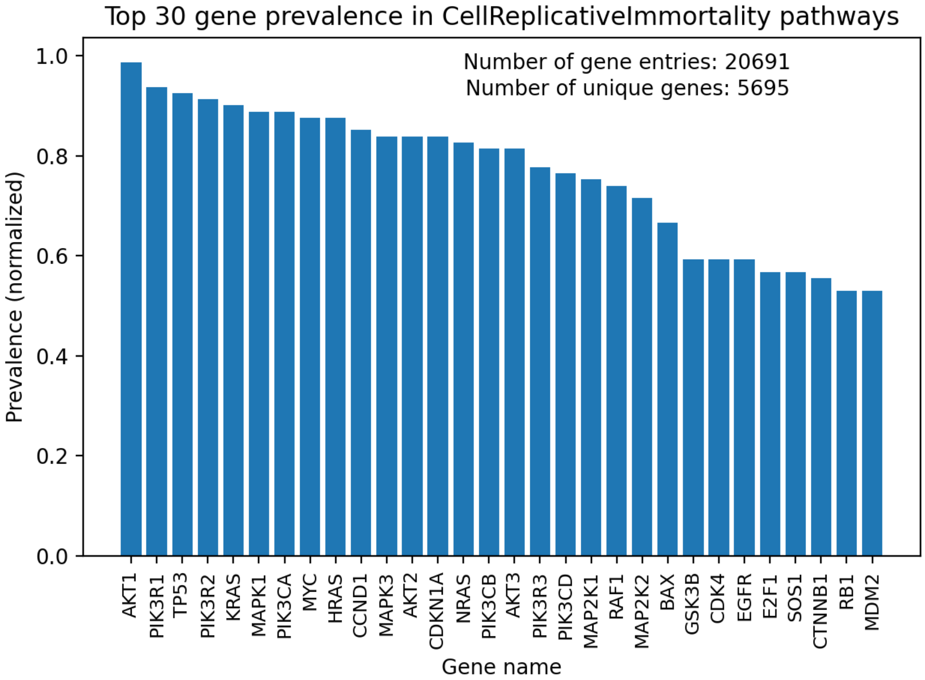
6 Appendix

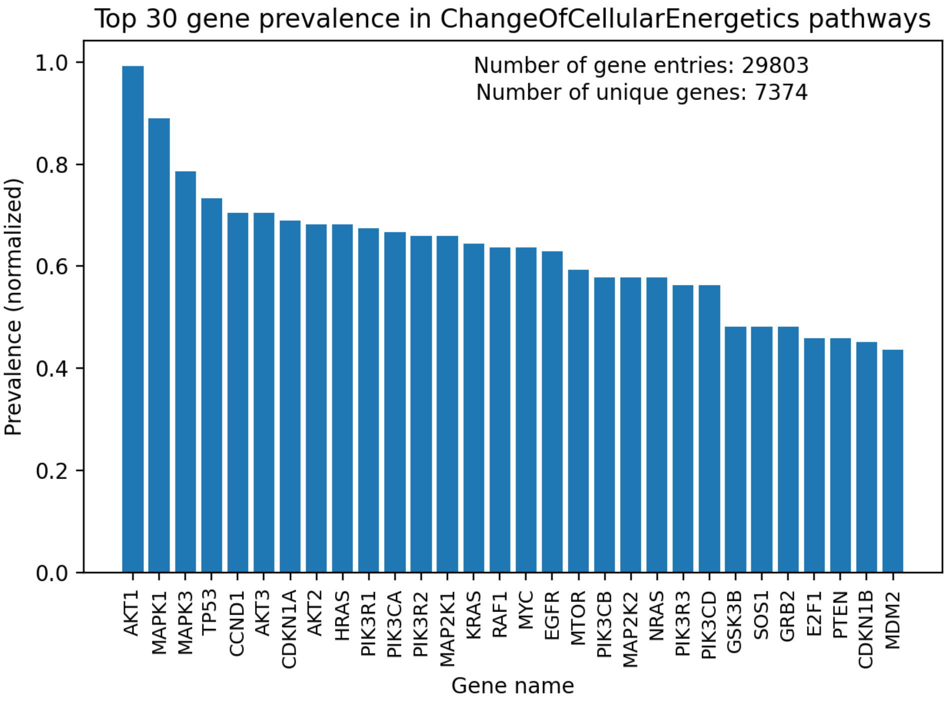
6.1 Python code for generating pathway scores.

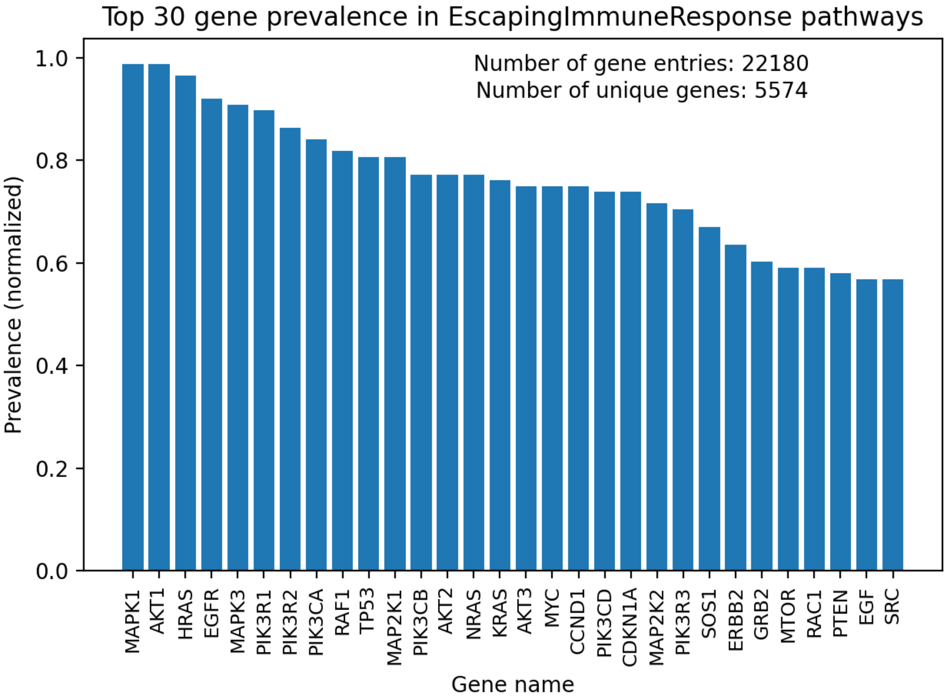
```python
#!/usr/bin/python

import sys
from collections import Counter
import matplotlib.pyplot as plt

def main():
    genelist = sys.stdin.read()
    # with open ("Angiogenesis.txt", "r") as myfile:
    #     genelist = myfile.read().replace("\n", ",")
        # genelist = genelist.replace("][", ",").replace(
        #     "[","").replace("]", "").replace(’"’,"")

    genelist = genelist.split(’\n’)
    totalNrOfGenes = str(len(genelist))
    numberOfUniqueGenes = str(len(set(genelist)))
    countedArray = Counter(genelist)

    x = countedArray.most_common()
    listedx = list(x) # use list because tuples are immutable
    arraySize = len(listedx)

    # calculate range of occurence data for normalization step
    maxValue = listedx[0]
    maxValue = maxValue[1] +1 # -1 to prevent ’1’ values after normalization
    minValue = listedx[arraySize-1]
    minValue = minValue[1] -1 # +1 to prevent ’0’ values after normalization

    # normalized formula: (x-min(x))/(max(x)-min(x)), rounded to 4 decimals

    f = open("TumourPromotingInflammation_Normalized.txt", "a")
    f.write("display name, TumourPromotingInflammation_Pathway_Score" + ’\n’)
    for gene in range(len(listedx)):
        normalized = round((listedx[gene][1]-minValue)/(maxValue-minValue), 4)
        listedx[gene] = (listedx[gene][0], normalized)
        testje = ’,’.join(map(str, listedx[gene]))
        f.write(testje + ’\n’)
    f.close()

    listedx = listedx[:30] # only use top 30 most prevalent genes for graph ge

    plt.bar(*zip(*listedx))
    plt.text(0.65,0.95,"Number of gene entries: " + totalNrOfGenes,
        horizontalalignment='center',verticalalignment='center',
        transform = plt.gca().transAxes)
    plt.text(0.65,0.9,"Number of unique genes: " + numberOfUniqueGenes,
        horizontalalignment='center',verticalalignment='center',
        transform = plt.gca().transAxes)
    plt.xticks(rotation=90,ha='center', fontsize=9)
    plt.title('Top 30 gene prevalence in TumourPromotingInflammation pathways')
    plt.xlabel('Gene name')
    plt.ylabel('Prevalence (normalized)')
    plt.tight_layout()
    plt.show()

if __name__ == "__main__":
    main()
```

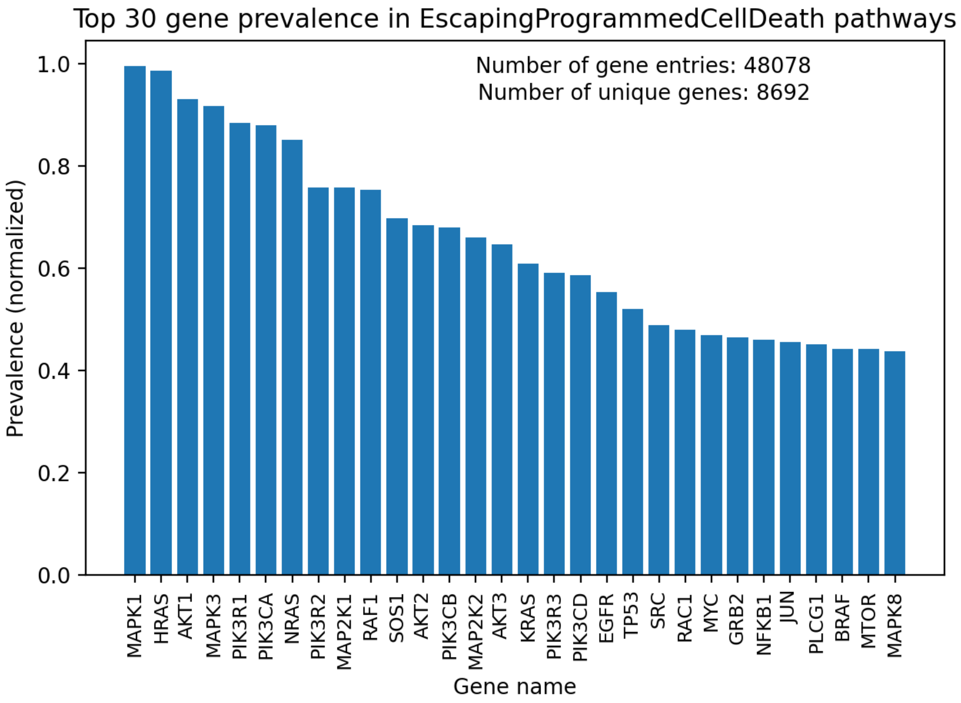
6.2 Gene prevalence graphs

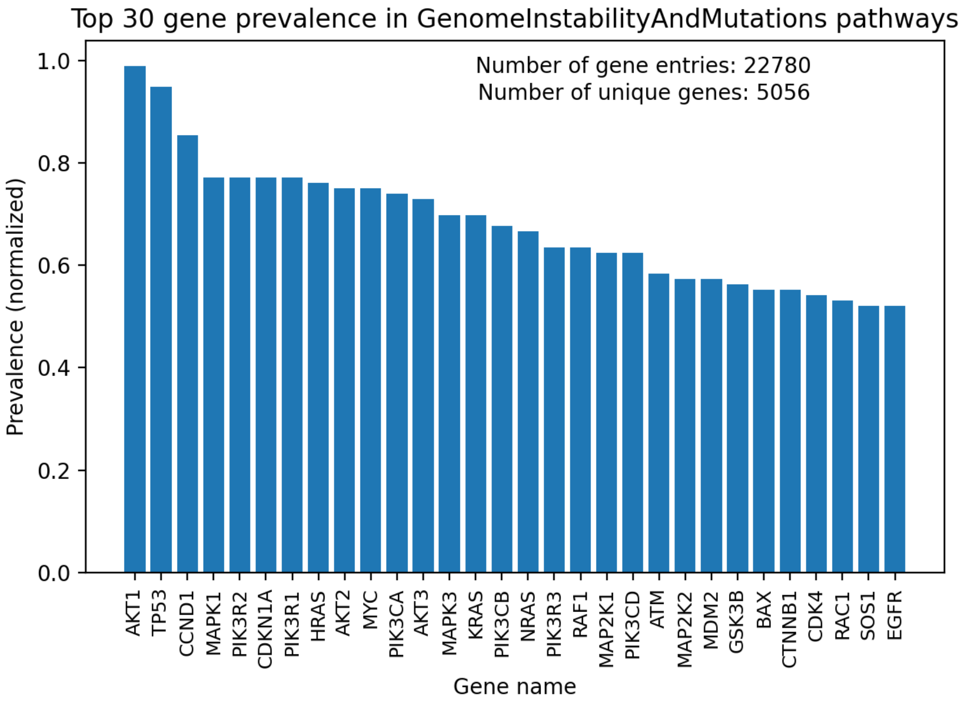


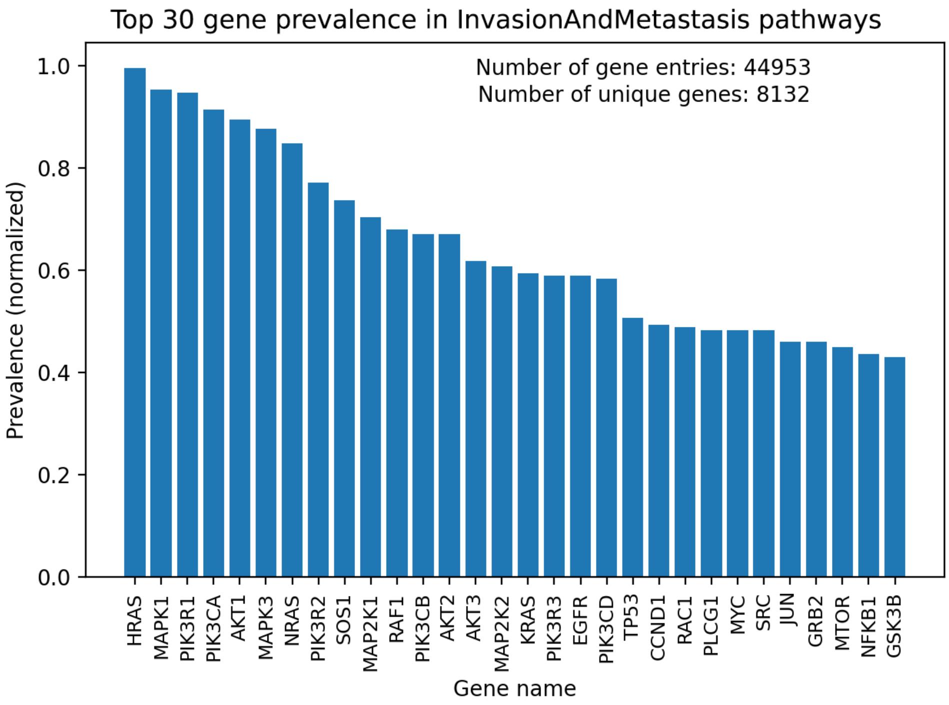


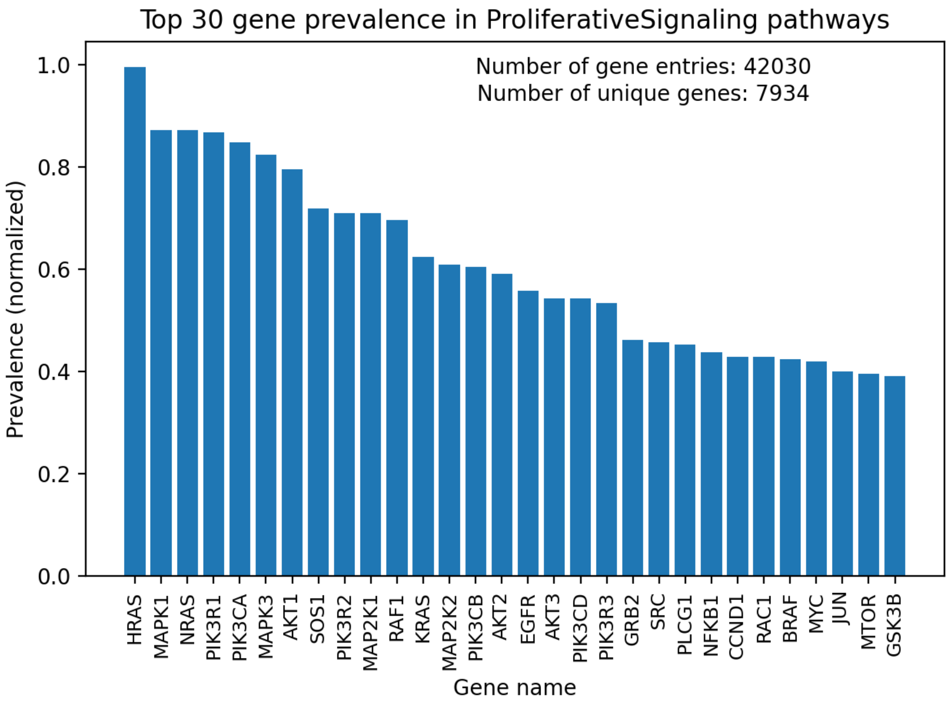

Top 30 gene prevalence in ChangeOfCellularEnergetics pathways
Number of gene entries: 29803
Number of unique genes: 7374
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
AKT1
MAPK1
MAPK3
TP53
CCND1
AKT3
CDKN1A
AKT2
HRAS
PIK3R1
PIK3CA
PIK3R2
MAP2K1
KRAS
RAF1
MYC
EGFR
MTOR
PIK3CB
MAP2K2
NRAS
PIK3R3
PIK3CD
GSK3B
SOS1
GRB2
E2F1
PTEN
CDKN1B
MDM2
Gene name

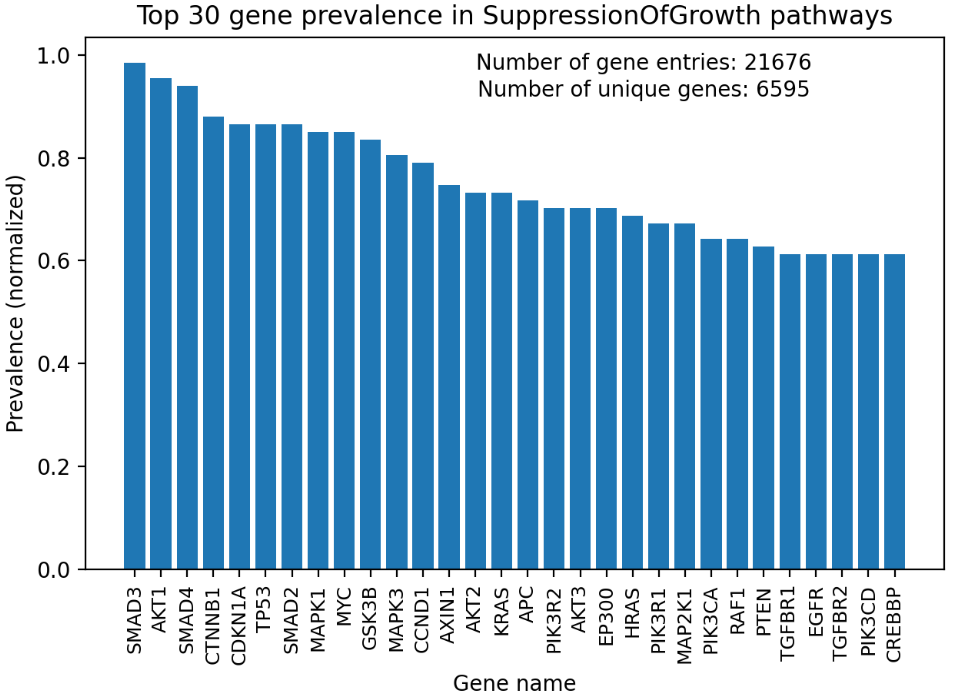

Top 30 gene prevalence in EscapingImmuneResponse pathways
Number of gene entries: 22180
Number of unique genes: 5574
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
MAPK1
AKT1
HRAS
EGFR
MAPK3
PIK3R1
PIK3R2
PIK3CA
RAF1
TP53
MAP2K1
PIK3CB
AKT2
NRAS
KRAS
AKT3
MYC
CCND1
PIK3CD
CDKN1A
MAP2K2
PIK3R3
SOS1
ERBB2
GRB2
MTOR
RAC1
PTEN
EGF
SRC
Gene name

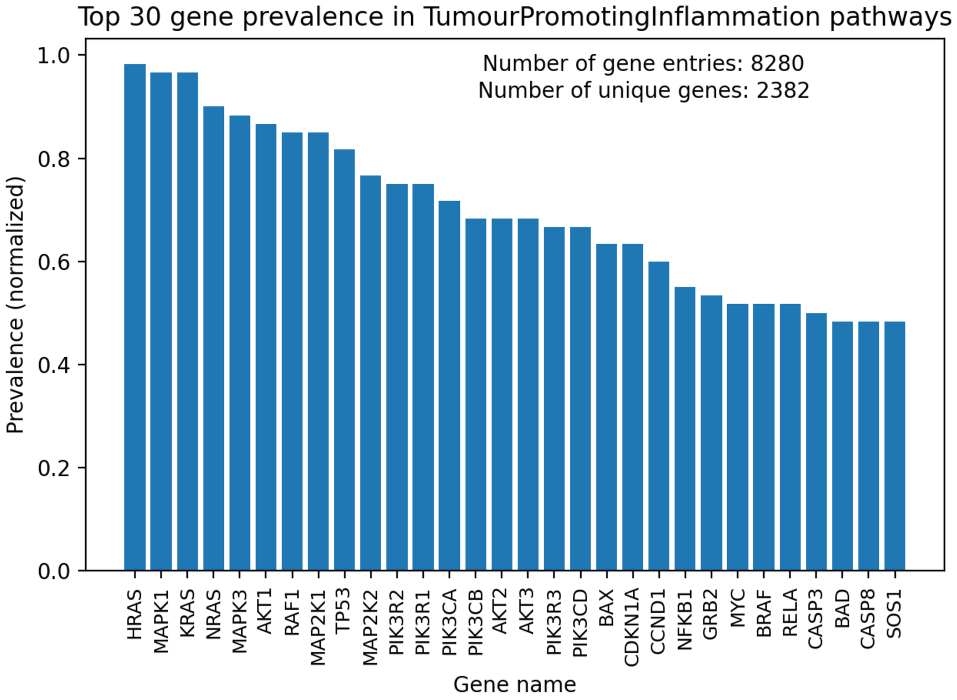

Top 30 gene prevalence in EscapingProgrammedCellDeath pathways
Number of gene entries: 48078
Number of unique genes: 8692
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
MAPK1
HRAS
AKT1
MAPK3
PIK3R1
PIK3CA
NRAS
PIK3R2
MAP2K1
RAF1
SOS1
AKT2
PIK3CB
MAP2K2
AKT3
KRAS
PIK3R3
PIK3CD
EGFR
TP53
SRC
RAC1
MYC
GRB2
NFKB1
JUN
PLCG1
BRAF
MTOR
MAPK8
Gene name


Top 30 gene prevalence in GenomeInstabilityAndMutations pathways
Number of gene entries: 22780
Number of unique genes: 5056
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
AKT1
TP53
CCND1
MAPK1
PIK3R2
CDKN1A
PIK3R1
HRAS
AKT2
MYC
PIK3CA
AKT3
MAPK3
KRAS
PIK3CB
NRAS
PIK3R3
RAF1
MAP2K1
PIK3CD
ATM
MAP2K2
MDM2
GSK3B
BAX
CTNNB1
CDK4
RAC1
SOS1
EGFR
Gene name


Top 30 gene prevalence in InvasionAndMetastasis pathways
Number of gene entries: 44953
Number of unique genes: 8132
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
HRAS
MAPK1
PIK3R1
PIK3CA
AKT1
MAPK3
NRAS
PIK3R2
SOS1
MAP2K1
RAF1
PIK3CB
AKT2
AKT3
MAP2K2
KRAS
PIK3R3
EGFR
PIK3CD
TP53
CCND1
RAC1
PLCG1
MYC
SRC
JUN
GRB2
MTOR
NFKB1
GSK3B
Gene name


Top 30 gene prevalence in ProliferativeSignaling pathways
Number of gene entries: 42030
Number of unique genes: 7934
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
HRAS
MAPK1
NRAS
PIK3R1
PIK3CA
MAPK3
AKT1
SOS1
PIK3R2
MAP2K1
RAF1
KRAS
MAP2K2
PIK3CB
AKT2
EGFR
AKT3
PIK3CD
PIK3R3
GRB2
SRC
PLCG1
NFKB1
CCND1
RAC1
BRAF
MYC
JUN
MTOR
GSK3B
Gene name


Top 30 gene prevalence in SuppressionOfGrowth pathways
Number of gene entries: 21676
Number of unique genes: 6595
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
SMAD3
AKT1
SMAD4
CTNNB1
CDKN1A
TP53
SMAD2
MAPK1
MYC
GSK3B
MAPK3
CCND1
AXIN1
AKT2
KRAS
APC
PIK3R2
AKT3
EP300
HRAS
PIK3R1
MAP2K1
PIK3CA
RAF1
PTEN
TGFBR1
EGFR
TGFBR2
PIK3CD
CREBBP
Gene name


Top 30 gene prevalence in TumourPromotingInflammation pathways
Number of gene entries: 8280
Number of unique genes: 2382
Prevalence (normalized)
1.0
0.8
0.6
0.4
0.2
0.0
HRAS
MAPK1
KRAS
NRAS
MAPK3
AKT1
RAF1
MAP2K1
TP53
MAP2K2
PIK3R2
PIK3R1
PIK3CA
PIK3CB
AKT2
AKT3
PIK3R3
PIK3CD
BAX
CDKN1A
CCND1
NFKB1
GRB2
MYC
BRAF
RELA
CASP3
BAD
CASP8
SOS1
Gene name