HELLENIC REPUBLIC

National and Kapodistrian University of Athens

EST. 1837

Network Graphs of Cancer Mutations

Stamatis Choudalakis

Thesis Project
Department of Mathematics
Master of Science in Applied Mathematics

Supervisors
Prof. Stratis, National and Kapodistrian University of Athens
Dr. Dikaios, Academy of Athens
Dr. Kastis, Academy of Athens

2021-2022

Supervisors

- Ioannis Stratis, Professor in National and Kapodistrian University of Athens, Department of Mathematics
- Nikos Dikaios, Researcher C, Academy of Athens
- George Kastis, Researcher A, Academy of Athens

Acknowledgements

I would like acknowledge and give my warmest thanks to my supervisors for their support and their advices throughout this year, as well as for giving me the opportunity to work on a very crucial and complex subject, cancer.

I would like to give my appreciation to prof. Stratis who made this project possible and recommended me in the Academy of Athens.

I would also like to thank Dr. Dikaios who improved my computational skills at a great extend, and enhanced my learning skills and methodologies through his guidance.

I would finally like to thank Dr. Kastis who improved my english academic writing and helped me not to leave anything misinterpreted in the present project.

Contents

1 Introduction

Cancer is a disease that causes millions of deaths annually. Cancer goes back from the ancient times; at some point it was considered to be contagious [1]. We now know that cancers develops due to the accumulation of genetic and epigenetic alterations over somatic cells (i.e cells that don't pass to offsprings) [2].

Genetic mutations (or alterations) alter the DNA sequence of the cell and can be caused by a number of things, such as tobacco and other substances, called carcinogens [3], or even caused by viruses like HPV [4] and Epstein-Barr virus [5].

Mutations can be categorized according to the magnitude of the DNA sequence they alter. If a mutation affects a small amount of the DNA sequence, then the mutation is called point-wise. If the mutation is of a bigger scale, the mutation is called chromosomal.

Mutations can initiate or confer to cancer, by altering the function of proteins. Proteins are molecules essential for the function and structure of the organism and they are coded by parts of the DNA called genes. A number of genes is involved in a procedure called cell cycle (or cell division), where the cell replicates itself. Therefore, tumorigenesis can be caused by cells that continuously undergo the cell cycle, when they shouldn't, due to mutations that affect the functionality of proteins involved during the cell cycle.

However, not all mutations are able to confer to cancer. In order for the mutated cells to divide uncontrollably, several defense mechanisms of cell division must be bypassed. For example, it must evade cell destruction (apoptosis), sustain growth-promoting and avoid growth-suppressing signals [6]. If one or more of the aforementioned properties are present within a cell through a mutation of a gene, then the gene is called driver gene. Driver genes play a natural role over cancer initiation and/or progression. Genes that don't confer to cancer initiation and/or progression, through mutations, are called passengers [7].

The search of driver genes is crucial to the development of therapeutic plans and to understand cancer, in general [8, 9, 10]. Genomic analysis, to identify driver genes, can be performed either with respect to the organ, where the cancer initiated (primary site) or to the specific tissue of the organ (cell line). Current methods of driver gene identification rely, in general, on finding specific driver genes or cancer driver modules. The tools used for the identification task can be a number of molecular data such as gene expression data and protein-protein interactions networks.

Until 2002, more than 100 oncogenes have been discovered [2]. Next generation sequencing has made able the launch of projects such as International Cancer Genome Consortium (ICGC) [11] and The Cancer Genome Atlas (TCGA) [12] in an effort to systematically catalogue somatic mutations [7]. Research over

these and other datasets has been carried out in order to distinguish driver from passenger events (see Identification of cancer drivers).

We must note that epigenetic mutations can also initiate and/or progress cancer. Epigenetic mutations, such as DNA methylation [13, 14], don't alter the DNA sequence affected. Apart from epigenetic mutations, genes that don't code for proteins (non-coding genes) can also be implicated in cancer tumorigenesis [2].

Here, Network Graph Theory is exploited for the search of cancer drivers, combining mathematical concepts of graph theory and clustering as a means to recover modules containing genes and cancer patients.

Networks arise everywhere in real world. From people [15] to medicine [16] and chemistry [17], a variety of networks can be described. Any set of objects and interactions between them can naturally be expressed as a graph.

Graphs are a useful tool to describe data. A graph is defined by its nodes (or vertices) and edges (or linkes), while the meaning of each graph depends on the nature of the study. For example, the structure of a molecule can be represented as a graph that contains the atoms (nodes) and lines (edges) between them if the atoms bond [18]. Another example is the map of a country where cities (nodes) are connected with each other through roads (edges).

A number of variables can be defined in networks and graphs alike, such as the density and the diameter of the network. Node-specific variables, like the centralities, can also be defined. These measures can be used as a means to understand the topology of the networks and also create computer-generated graphs similar to the graphs of study.

One of the problems of interest over a network is to reveal its underlying structure, if such exist. Clustering analysis is one of the branches of machine learning that offers techniques to solve the issue. Clustering over networks is based on the idea that a set of nodes who share more links among them than with the rest of the nodes, present a kind of independence with respect to the rest of the network. These nodes can form a cluster (also called community, module or group).

The definition of a group (or community, cluster) though is vague, leading to various criteria upon when two nodes must belong to a certain cluster [19]. Given that degree of freedom, a number of algorithms have been proposed to address the problem. They can either exploit the overall structure of the network or reside in the dynamics of the network.

One of the most famous tools in clustering is an index called modularity (see Modularity). The core of the index is the comparison of every edge of the network with a corresponding null model. The closer it is to one, the better the partition of the network that the index corresponds to. In that way, algorithms based on modularity are meant to maximize it. Apart from modularity, statistics can also be used to recover clusters. Currently, only one algorithm is based solely on statistics, called OSLOM (see OSLOM). OSLOM and methods based

on modularity have in common the involvement of null models. On the other hand, algorithms based on the dynamics of the network, don't make use of null models. These algorithms can be based, for example, in the flow of the network, which is defined by the edges of the network. Here, INFOMAP is an algorithm that makes use of flow dynamics and is further analyzed in INFOMAP.

In the present project, Chapter 2 contains the basics of molecular biology and cancer, in order for the reader to understand cancer's initiation and progression. Chapter 3 is divided in two major parts. The first part, contains the fundamentals of Graph Theory, while the second part contains the fundamentals of Network Theory and Clustering over networks. Examples of algorithms and the mathematics behind them are also present in the same chapter. Chapter 4, 5 and 6 present the implementation of clustering in a gene-patient network. The network is built using TCGA somatic mutations data for only primary samples, as a mean to recover cancer driver modules. Specifically, Chapter 4 contains the methods used, Chapter 5 contains the results and their analysis, while Chapter 6 is a discussion regarding the results, along with some conclusions.

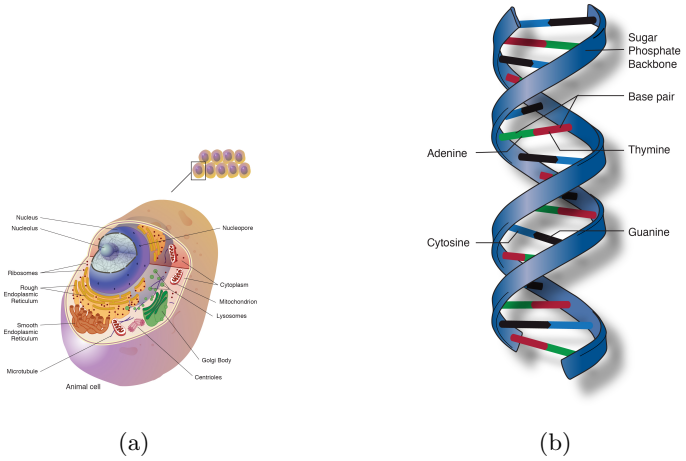
The graphs of Chapter 3 in Graph Theory, are made using [20]. Graphs of Chapter 3 in Network Theory, were made using Gephi [21] and python's library Networkx [22]. The code used for this project is fully written in Python [23], while third party applications, such as MODULAR [24], OSLOM2 [25], and the python-integrated version of INFOMAP [26], were also utilized. It must also be pointed out that the mathematics and definitions over Graph and Network Theory are mainly derived from [27, 28, 29].

2 Molecular Biology and Cancer

2.1 DNA

Every living organism consists of smaller units, the cells, which are also referred to as the "building blocks of life". Each cell contains a variety of organelles, one of which is the nucleus (see Figure 1a). Inside the nucleus, the information for the structure, functioning and reproduction of the being is stored in the form of DNA. The same structure is observed in every living organism including some viruses [30]. Apart from the DNA in the nucleus, there exist the mitochondrial DNA (mtDNA). [31].

DNA, or else, Deoxyribonucleic acid is a molecule in the form of a double helix, with two strands. Each strand is composed of a polynucleotide chain. This chain is formed by 4 specific bases, namely, Adenine (A), Thymine (T), Cytosine (C), Guanine (G). Adenine pairs with Thymine, Cytosine pairs with Guanine and all these base pairs along with a sugar-phosphate backbone, ultimately form the DNA (see Figure 1b) [32]. DNA sequences are called genes (see Figure 2a) and genes packed together accompanied by some proteins form the chromosomes (see Figure 2b). Chromosomes can be seen only through the process of cell division. They consist of two identical parts, the chromatids, that intersect at the centromere. Finally, the protection of the terminal regions of the chromosome, reside in the telomeres [33].

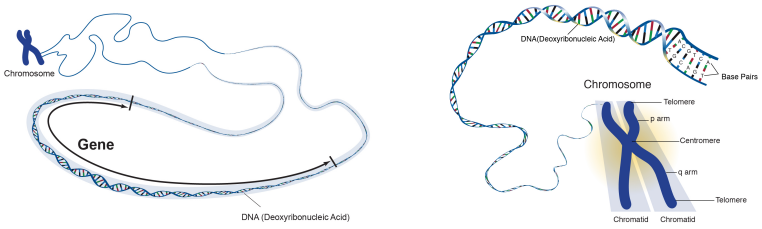


(a) (b)

Figure 1: Illustrations of (a) an animal cell and (b) the double helix form of DNA. Credits to the National Human Genome Research Institute.

2.1.1 Genes

Genes, can be classified in two basic categories, the functional and the non-functional genes (also called "junk-DNA"). The functional genes will be segmented in two more groups. The protein-encoding and the non-coding genes.



(a) (b)

Figure 2: Illustrations of (a) genes packing up to create chromosomes and (b) the structure of the chromosomes. Credits to the National Human Genome Research Institute.

Protein-Encoding Regions Protein-encoding genes, via two processes, make up proteins, which are essential for the structure, function, and regulation of the body's tissues and organs. During the first process, called transcription, a molecule, called mRNA, is formed. Its aim is to transfer the genetic information of the gene outside of the nucleus through the cytoplasm where it will be bounded by organelles, called ribosomes, in order to initiate translation, where the protein will be formed (see Figure 4). Not all of the functional protein-encoding gene information will be carried out by the mRNA. Exons, parts of the gene, will be the ones to be translated. Exons, in the genetic sequence of the gene, are separated by intervening sections, called introns, through a procedure, called splicing (see Figure 3) [31]. In addition, the mRNA is read in triplets, called codons, and there exist a "start" and a "stop" codon.

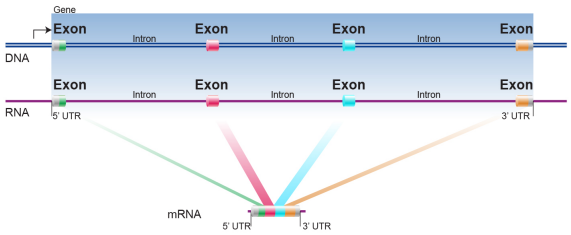


Figure 3: Illustration of Splicing. Credits to National Human Genome Research Institute.

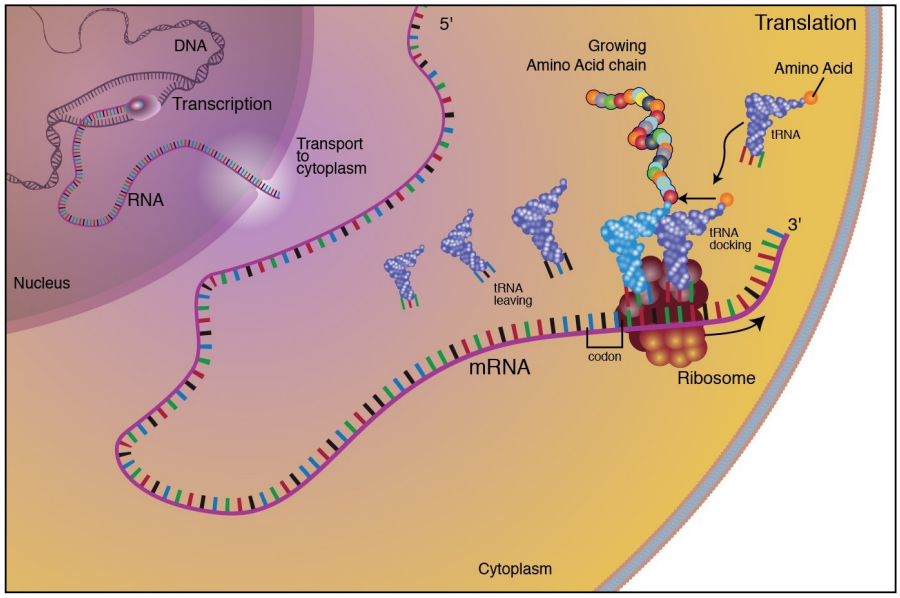


Figure 4: Illustration of Translation. Credits to National Human Genome Research Institute.

Non-Coding Regions

Non-coding regions are the parts of the DNA sequence that don't code for proteins. These regions can provide with RNA units such as the tRNAs and rRNAs, which are essential parts of the translation process. Some regions can also serve as binding sites for proteins, that regulate a gene's expression. These regions are namely: promoters, enhancers, silencers and insulators. Other non-coding regions can be part of the structure of the chromosome, such as the telomeres, which play a crucial role on the integrity of the chromosome [34].

The non-coding RNA subgroups are multiple, as seen in Figure 5. We note that these regions were previously thought as non-functional DNA, and there is still "dark matter genome" yet to be fully understood [35].

2.2 Mutations

A mutation (also called alteration) is an event where the DNA sequence of a cell is altered. A number of reasons may cause mutations. Exposure to UV radiation, tobacco and errors in the replication of the DNA are some of them. If the cell afflicted is a germ cell, meaning that the DNA of the cell can pass to the offspring, the mutation is called germline. In all other cases, the mutation is called somatic. A mutation can also be point-wise or chromosomal, depending on the magnitude of the nucleotides it affects. A point-wise mutation affects a small amount of bases, while a chromosomal mutation affects either the structure

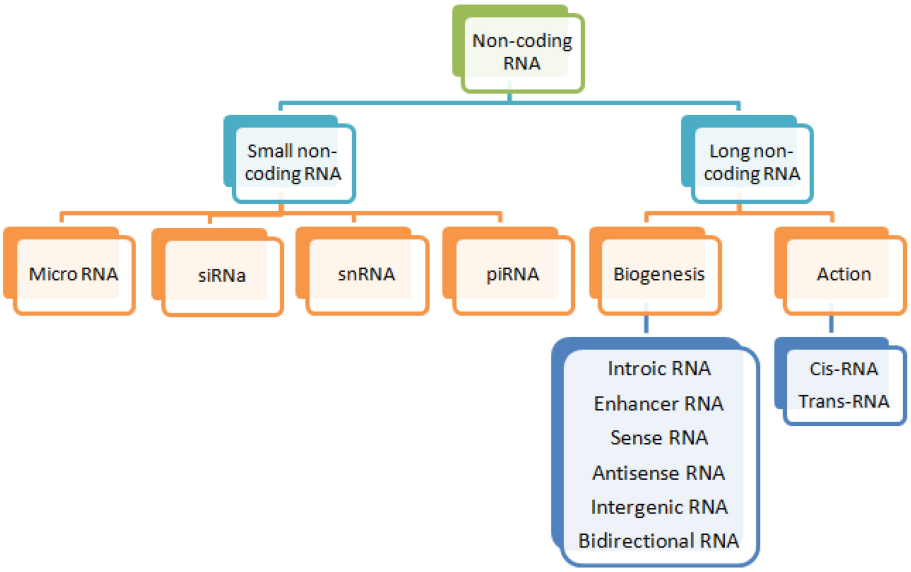


Figure 5: Classification of Non-Coding RNA. Credits to [36].

of a chromosome or the number of chromosomes [37, 38, 39].

Base pair substitutions refer to a change from a base pair to another. When this happens, a codon is altered. The altered codon can either code for the same amino-acid as the initial (silent mutation), or can code for a different (missense mutation) one. If a different amino-acid is coded, the effect on the overall functionality of the protein may (conservable mutation) or may not (non-conservable mutation) change. A nonsense mutation is a result of a codon conversion to a stop codon, which will produce a smaller polypeptide than the normal one, usually causing the protein to be non-functional. Examples of such mutations can be seen in Figure 6.

Due to a frameshift mutation a base pair may be added (insertion) or deleted (deletion). Frameshift mutations may result in the formation of new stop codons, reducing the size of the protein, often making the final product not functional. Examples of such mutations can be seen in Figure 7.

Chromosomal mutations can be of structural or of numerical nature. Structural chromosomal mutations affect the arrangement of the chromosomes and can either occur to a single chromosome (Intrachromosomal) or to a combination of them (Interchromosomal). In the first case, a segment of the chromosome may be deleted (Deletion), duplicated (Duplication), inverted 180 angles (Inversion) or moved from one location to another (Translocation), in which case there is no loss of genetic information. When the translocation occurs in a single chromosome it is called non-reciprocal. If the translocation involves two

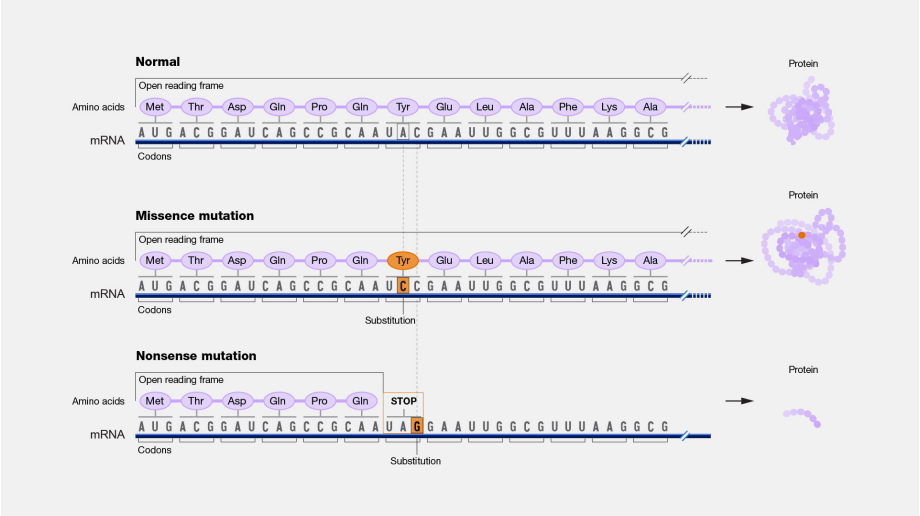


Figure 6: Examples of Point Mutations. This figure illustrates how the final protein is afflicted with respect to the normal protein (top) after a missense mutation (center) or after a nonsense mutation (bottom). Credits to the National Human Genome Research Institute.

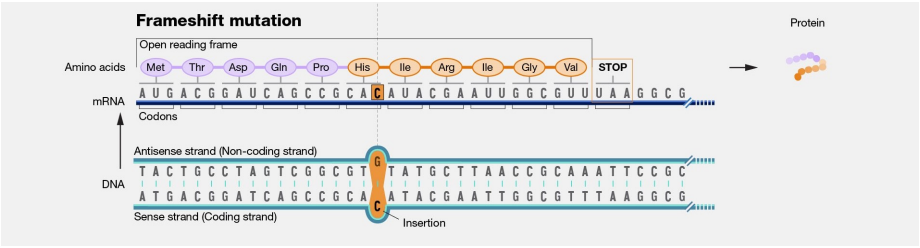


Figure 7: Example of Frameshift Mutation (Insertion) and how it affects the final protein with respect to the normal one of Figure 6. Credits to National Human Genome Research Institute. (Figure is cropped)

chromosomes and parts of the chromosomes are exchanged, the translocation is called reciprocal. Examples of such mutations can be seen in Figure 8.

Copy Number Alteration (CNA) refers to the mutation where part of the DNA sequence appears more (or less) times than it should. Various examples of CNAs can be viewed in Figure 9.

As far as the numerical chromosomal mutations are concerned, Aneuploidy is the mutational event where a chromosome appears more (or less) times than it should. Figure 10 is an example of trisomy 21, where chromosome 21 appears three times rather than two. [40].

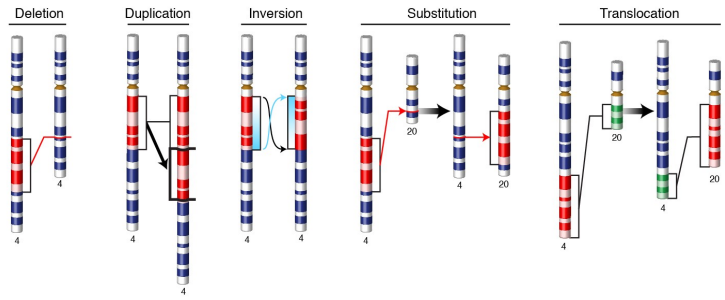


Figure 8: Mutations affecting the structure of the chromosome. From left to right: Deletion, Duplication, Inversion, Substitution, Reciprocal Translocation. Credits to the National Human Genome Research Institute.

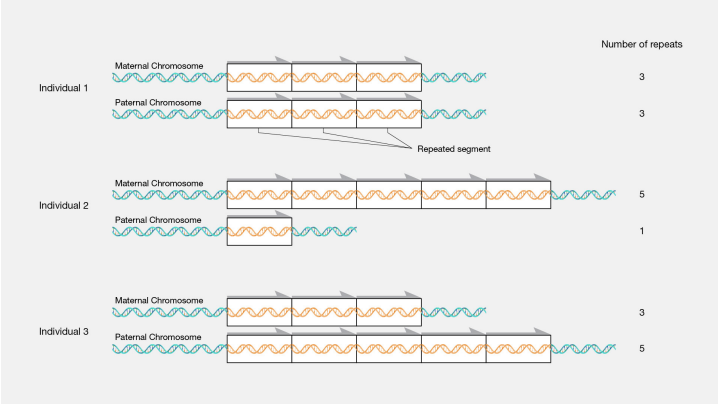


Figure 9: Illustration of Copy Number Alterations. In this figure, a variety of cases where a genetic sequence appears more than is illustrated. Credits to the National Human Genome Research Institute.

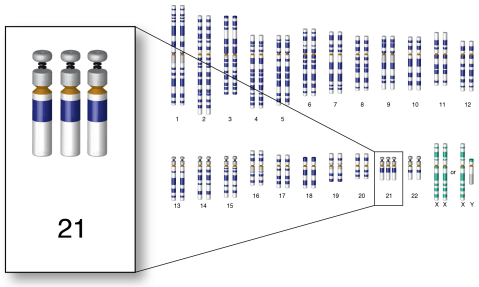


Figure 10: Example of Aneuploidy case, trisomy 21, where chromosome 21 appears 3 times.

2.3 Cell Cycle

Now that the basic elements of genetics such as the DNA, it's structure and its potential alterations, are presented, we recall that cancer is a disease mainly caused by accumulation of mutations over somatic cells that lead the cell to divide uncontrollably. Examining the cell division procedure is, consequently, an important element to understand the disease.

A cell may, or may not divide. Some types of cells divide rapidly, and in these cases, the daughter cells (i.e. the cells that occur after cell division) may immediately undergo another round of cell division. For instance, many cell types in an early embryo divide rapidly, and so do cells in a tumor. Others, like neurons, cells that conduct signals, or liver cells that store carbohydrates, are not actively preparing to divide. The decision whether the cell enters the cycle or not, depends on extracellular signals (see Figure 11). These signals may be amino-acids, lipids etc. and they are detected by specific receptors of the cell. Receptors are chemical structures composed of proteins [40, 41].

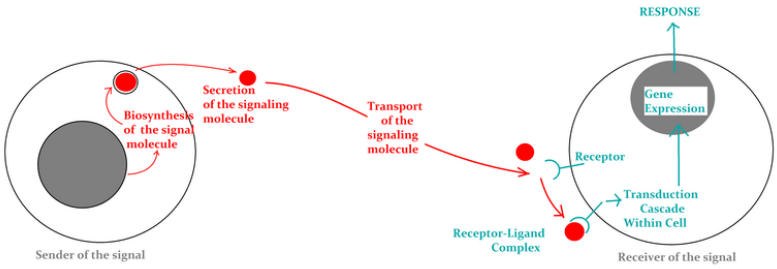


Figure 11: Illustration of cell-to-cell communication. The "sender" cell (to the left) produces molecules that are binded by the "receiver" cell (to the right), causing the gene expression of the "receiver" cell. Taken from [42].

The stages (see Figure 12) of the procedure and some examples of the mechanisms in each stage, are presented in Table 1.

G1 phase: • the cell grows physically larger • organelles are copied	S phase: • the cell synthesizes a complete copy of the DNA
G2 phase: • the cell grows more • makes proteins and organelles • begins to reorganize its contents	M phase: •the cell divides its copied DNA and cytoplasm

Table 1: Phases of the cell cycle and examples of the mechanisms taking place in each phase.

The G1, S and G2 are also called Gap 1, Synthesize and Gap 2 phases. The interphase consists of the three of them, while the M is called the Mitotic phase. Between these events there are 3 major checkpoints to maintain the normal outcome of the process

1. G1/S checkpoint:
   - Cell size
   - Nutrients
   - Growth factors
   - DNA damage
2. G2/M checkpoint:
   - DNA damage
   - DNA replication completeness
3. M/G1 (Spindle) checkpoint:
   - Chromosome attachment to spindle at the opposite poles.

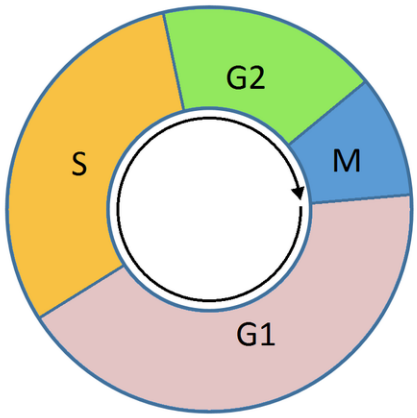


Figure 12: The cell cycle. Taken from [43]. (Figure is cropped)

2.4 Cancer Initiation and Progression

Mutations are capable of the onset of cancer. They are able to alter or even eliminate the gene's function. These events may lead to unrestrained cell proliferation and, eventually, tumorigenesis. Two very common examples are proto-oncogenes and tumor-suppressor genes. Proto-oncogenes promote growth of the cell and cell division. On the other hand, tumor-suppressor genes slow down cell division, repair DNA damages etc. Mutations of the first category may cause false signaling on cells that normally shouldn't divide. On the other hand, intrachromosomal deletion, for example, may cause loss of function (LOF) of the tumor-suppressor genes which does not allow the process to end when a problem occurs [40].

Not all mutations, though, are able to make a cell cancerous. There are mutations that are beneficial to humans, such as those that make us more protected against certain diseases, while there are others that are aloof on the overall health of the organism. All of the above mutations are called passengers, while the rest are called drivers. The number of drivers is significantly less than the number of passengers and the total number of drivers required for abnormal cell proliferation varies with cancer type [2, 7].

Cells have to harbor certain properties in order to be considered cancerous because they would have to avoid several checkpoints as seen in Section Cell

Cycle. These properties are called the "Hallmarks of Cancer" [6] and are namely the following:

- Self-sufficiency in growth signals
- Insensitivity to anti-growth signals
- Evasion of programmed cell death (Apoptosis)
- Limitless replicative potential
- Sustained Angiogenesis
- Metastasis

Therefore, the most likely scenario is that a driver mutation occurs, which is the one that initiates cancer. From thereafter, the cell seems to acquire a "mutator phenotype", making it prone to mutations. This can also be viewed as an evolutionary process in which the cell acquires the rest of the properties it needs in order to survive and evade other tissues and organs [7].

2.5 Cancer on cellular level

Observing cancer on cellular level under the microscope, we notice two things. First, the amount of cells in the tissue is larger than normal, which is called hyperplasia. Second, the shape of the cells may vary from cell to cell, which is called dysplasia. See Figure 13 [44].

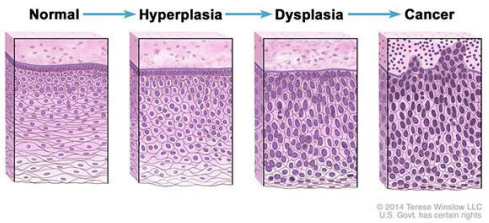


Figure 13: Different types of cells from normal to cancerous. Credits to the National Cancer Institute.

Moreover, cancer cells via cell-to-cell interactions may be able to turn healthy neighbor cells into cancerous ones. This can happen because cancer cells seem to produce more exosomes than normal cells. Exosomes are vesicles with specialized function, such as intercellular signaling. Exposure to the cancer exosomes can alter gene expression in the normal cells, causing them to be cancerous as well [45].

2.6 Gene Expression

The expression of a gene is sometimes regulated by another gene. Therefore, mutations of one gene may affect the outcome of other genes called downstream genes. Genes, like TP53, may code for proteins called transcription factors, whose use is to bind to DNA sequences (enhancer or promoter sequences) and

therefore manage the expression of the downstream genes. Thus, genes that do not carry DNA alterations can have an oncogenic or tumor suppressing role [46].

2.7 DNA Methylation

Genetic as well as epigenetic alterations may play a crucial role over tumorigenesis. Such variation is the DNA methylation in which, methyl groups are added to DNA (in cytosine or adenine). Similar to gene expression, hypermethylation can lead to over-expression of the genes as hypomethylation, which can lead to transcriptional silency. These events can enhance the oncogenic role of the gene affected, or can suppress it's function. Dnmt genes (a family of genes) are the ones responsible for the catalysis of DNA methylation [13, 14].

2.8 SMGs, Pathways and Therapeutics

Over the years, different methods have been used in order to better understand and, of course, confront cancer. Cancer genomics exploit mutational data to come across Significant Mutated Genes (SMGs). Studies have been conducted not only to find specific genes, but also to find the factors associated with cancer (e.g. tobacco), the type of mutations, the number of the drivers, etc. One of the elements, essential for therapeutic methods, seems to be the pursuit of specific pathways in which SMGs take place. An example of the TP53 pathway is illustrated in Figure 14. We must note that the TP53 pathway in Figure 14 is altered due to mutation over MDM2 gene of a breast cancer patient labeled as "TCGA-3C-AAAU".

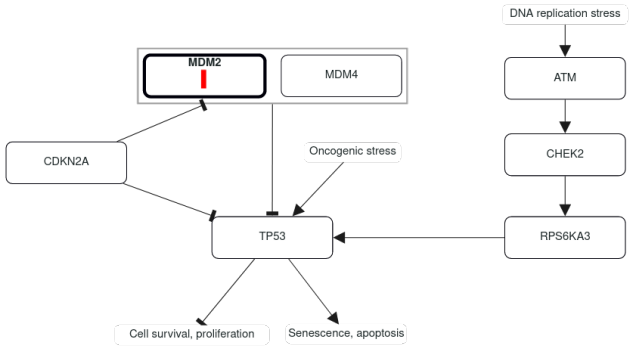


Figure 14: TP53 pathway of patient TCGA-3C-AAAU. The red color in the MDM2 gene of the pathway indicates that the sample, derived from the TCGA-3C-AAAU patient, carries a mutation in MDM2. Taken from cBioPortal.org.

2.9 Primary Site over Cell Line

Various mutations that affect the same target cell result in different tumour types, causing intertumoral heterogeneity between different cancer types. Mutational differences between cancers are not only limited in the disease site. Certain cancer types may be divided in subtypes, according to the genetic alterations of definite cells in the tissue that account as cell of origin [47]. In breast cancer, for example, levels of HER2 gene expression lead to the taxonomy of the cancer in four molecular subtypes [48]. Diverse subgroups are observed even within the subtype itself [49]. Intratumoral heterogeneity is also present, since only a fraction of the tumour's cells will be mutated. Other reasons causing intratumoral heterogeneity inlcude the contamination by normal cells [50].

Data of somatic mutations in tumour samples is, therefore, very sparse. This high heterogeneity makes difficult the task of recognizing low frequency mutations, as it reduces the levels of the desired signal from driver mutations. The presence of both driver and passenger mutations in the genome further reduces signal to noise ratio [48]. In addition, the mutational profile of patients with tumours from the same cancer subtype rarely have the same alterations [51]. Studies combining somatic mutations along with other data have revealed subtypes that, in general, include samples of various cancers per cohort [48, 52]. As pointed out in [52], cell-of-origin may not fully determine tumor classification, but even so, influence it. Other studies have also focused on subtype identification with a priori knowledge of the primary site [53, 54, 55, 56]. Moreover, knowledge of primary site is essential for a suitable therapeutic plan, especially in metastatic patients where the tissue of origin remains unknown, as $2-4\%$ of cancers are characterized as "Cancers of Unknown Primary" [57, 58]. The applications of analysis of somatic alterations with respect to the disease site also extend to a rapid diagnosis and treatment through identification of disease-specific mutations found in cell-free circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) [59] through blood or urine sequencing [60].

2.10 Identification of cancer drivers

Computational tools have been developed which separate driver from passenger events and can generally be divided in three categories (as listed here [61]): single cancer driver classification tools, cancer driver module and personalized classification tools. Thus, single genes can be correlated to cancer types as well as groups of them. The functional impact or the structural consequence of the mutations [62, 63, 64], gene expression data [65] and protein-to-protein interaction networks [66] are some of the few genomic and molecular data that algorithms can utilize in order to assess the significance of mutational events.

3 Network Graph Theory

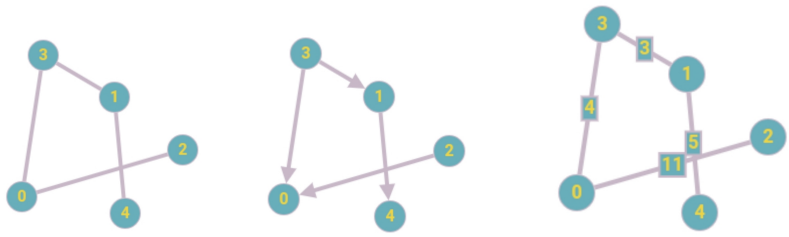
3.1 Graph Theory

3.1.1 Definition

A graph $G = \{E, V\}$ is a set of two sets, the edges E and the nodes V. The edges connect two nodes and two nodes connected together are called endpoints. The set of Neighbors $N(u)$ of a node u contains all the nodes linked with it. Edges can be directed or undirected. Undirected edges usually reveal a correlation of the two nodes, while directed edges indicate a one-way connection of an endpoint to another. Similarly, graphs are called undirected or directed. Directions aside, the two sets of a graph may be accompanied by a third, ordered one, W, which assigns a weight to each edge. Its physical meaning depends on the context of the graph. For example, in a graph of cities (nodes) and connecting roads between the cities (edges), the weight can indicate the distance between the cities. The graph is then defined as $G = \{E, V, W\}$. Examples of graphs can be seen in Figure 15.

Furthermore, an edge may connect more than two nodes with each other, in which case the graph is called hypergraph. Many edges can also exist between two endpoints, forming a multigraph. Lastly, the graph may contain self-loops.

We define a subgraph $G' = \{E', V'\}$ where $V' \subseteq V$, $E' \subseteq E$ and we say $G' \subseteq G$. Graph G is called supergraph.

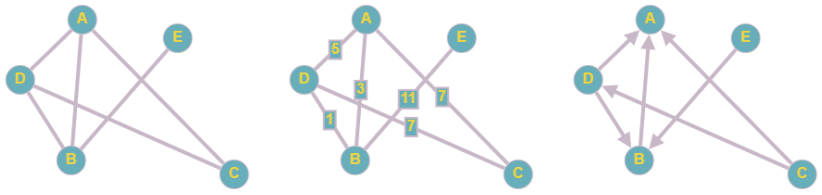


(a) (b) (c)

Figure 15: Examples of (a) undirected graph , (b) a directed graph and (c) a weighted graph.

3.1.2 Adjacency Matrix

In every graph we can assign an adjacency matrix A. Each element a_{ij} of A indicates whether node i is connected with node j. If i is connected to j, then $a_{ij} = 1$. Moreover, if the graph is weighted, then $a_{ij} = w_{ij}$ where w_{ij} is the weight of the edge between the nodes. If the nodes aren't linked, $a_{ij} = 0$. In the undirected case, the matrix is symmetric because $a_{ij} = a_{ji}$, which is not necessarily true in a directed graph. Also, the diagonal has zero values to every element if no self-loops exist. Tables 2a, 2b and 2c present the adjacency matrices of the graphs presented in Figs. 16a, 16b, and 16c respectively.



(a) (b) (c)

Figure 16: Examples of (a) an undirected graph, (b) a weighted graph and (c) of a directed graph.

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	1	1
C	1	0	0	1	0
D	1	1	1	0	0
E	0	1	0	0	0

(a)

	A	B	C	D	E
A	0	3	7	5	0
B	3	0	0	1	11
C	7	0	0	7	0
D	5	1	7	0	0
E	0	11	0	0	0

(b)

	A	B	C	D	E
A	0	0	0	0	0
B	1	0	0	0	0
C	1	0	0	1	0
D	1	1	0	0	0
E	0	1	0	0	0

(c)

Table 2: Adjacency matrices of (a) graph of Figure 16a , (b) graph of Figure 16b and (c) graph of Figure 16c.

3.1.3 Link List

Apart from the adjacency matrix, another representation, which will be exploited here, is the link list as mentioned in Infomap Online. This link-list consists of three columns: the "Source", the "Target" and the "Weight". If the network is unweighted, every element of the "Weight" column is equal to one. If the network is undirected it makes no difference whether a node is the "Source" or the "Target" except for specific cases, such as the exploitation of the bipartiness of the network in algorithms (see Classes). An example of a link list is presented in Table 3.

Source	Target	Weight
A	B	3
A	C	7
A	D	5
B	D	1
B	E	11
C	D	7

Table 3: An example of a link-list of the graph in Figure 16b.

3.1.4 Isomorphism

Let G_1 and G_2 be two graphs. If there exist a one-one correspondence (bijection) ϕ which maps one edge of a graph to exactly one edge of the other, then the graphs are called isomorphic. Specifically, if $e \in E_1$ then $\phi(e) \in E_2$. This means that a graph can be visualized in many ways by changing the position of a node in multiple ways. An example of isomorphic graphs is in Figure 17.

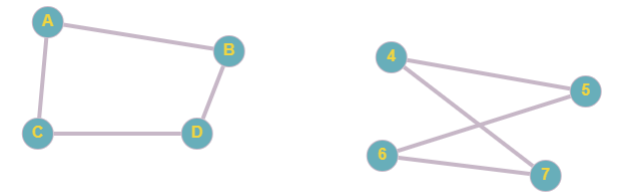


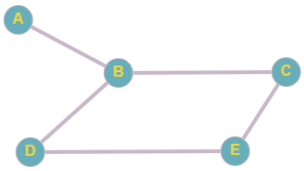
Figure 17: Two isomorphic graphs, where edges {A,B}, {A,C}, {C,D} and {D,B} correspond to edges {4,5}, {4,7}, {7,6} and {6,5} accordingly.

3.2 Definitions

A walk is defined as an ordered sequence of nodes where each node is linked with its previous and next. If nodes appear only once, then it's called a path and if edges appear only once it's called a trail. If the first and the last node are the same, it's called a cycle. If the graph contains a cycle it's called cyclic and if not, it's called acyclic. Moreover, if for every two nodes in an undirected graph, there exists a walk that connects those nodes, the graph is connected and if the graph is directed it's called strongly connected. Examples of the aforementioned definitions are presented in Figure 18

- walk: $\{A, B, D, E, C\}$
- path: $\{B, D, E, C\}$
- trail: $\{A, B, C, E\}$
- cycle: $\{B, D, E, C, B\}$

(a)



(b)

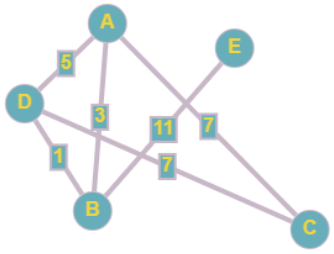
Figure 18: Examples of node sequences for the graph presented in (b).

3.2.1 Distance Matrix

Distance matrix is the matrix whose elements represent the minimum distance between the nodes. For example, if the (i, j) element equals 5 then the length of the shortest path that connects i to j is 5. If the graph is weighted, then the distance will be the sum of the weights of the edges of the shortest path. An example of distance matrix is presented in Figure 19.

	A	B	C	D	E
A	0	3	7	4	14
B	3	0	8	1	11
C	7	8	0	7	19
D	4	1	7	0	12
E	14	11	19	12	0

(a)



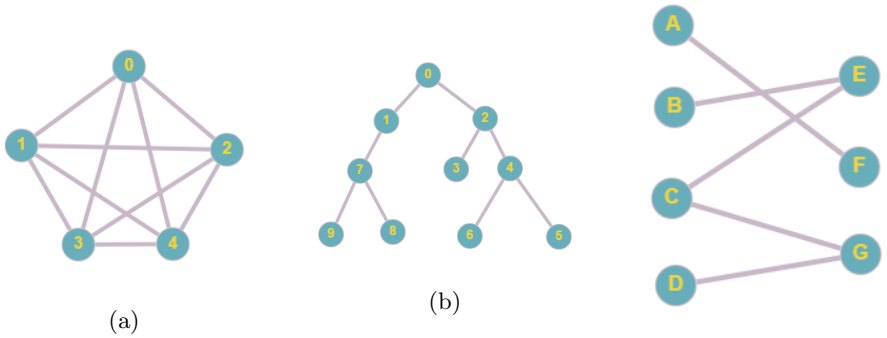
(b)

Figure 19: A weighted graph (b) and its distance matrix (a).

3.2.2 Classes

There are various graphs of certain form that can be categorized as explained below:

- If a randomly-chosen node connects with any other node in the graph the graph is called complete (see Figure 20a). By definition, exactly $\frac{N(N-1)}{2}$ edges exist within the graph. A complete subgraph is called a clique.
- If a graph is acyclic, it's called a forest and a connected forest is called a tree (see Figure 20b).
- If the nodes can be divided into two sets where there exist no edge between nodes of the same set, then the graph is called bipartite (see Figure 20c).



(a) (b) (c)

Figure 20: Examples of graphs that belong to a certain class. (a) is a complete graph. (b) is a tree. (c) is a bipartite graph.

3.3 Network Theory

3.3.1 Clustering Analysis

Clustering Analysis (or data segmentation) is a branch of Machine Learning which aims to find a partition that is most appropriate to depict the network in clusters (also called groups, modules or communities). The network clustering problem is an ill-defined problem as there is no specific way to define if two nodes of the network should belong in the same module. In that way, algorithms cannot be easily compared with each other. The general idea, though, is that the more similar the elements of the network are, based on a similarity (or dissimilarity) measure, the more likely it is that they will belong to the same cluster in the final partition. What we present below is the formulation of the above definitions.

Mathematically defining clustering Let $X = \{x_1, x_2, ..., x_m\}$ be the set of the m nodes of the network. The information over the data set may be provided in two ways.

First, as in many Machine Learning applications, a matrix of size $m \times n$ is utilized, where n is the number of features for each of the m objects. Then, each row will correspond to the embedded representation of the nodes, i.e $x_i = (x_{i_1}, x_{i_2}, ..., x_{i_n})$. The vector is also called feature vector of x_i or image of x_i.
Apart from the embedded representation of the nodes, an $m \times m$ matrix may carry out the information of the network which will be based on the relation of the nodes. Specifically, let S be the matrix that makes for that relational representation, then, each element s_{ij} of S is indicative of the similarity or dissimilarity (based on the structure of S) of the objects $x_i, \ldots, x_j$.

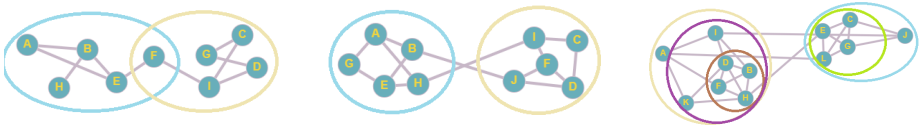
The clusters of X, $C = \{C_1, C_2, \ldots, C_k\}$, where $k < m$, formed usually have the following properties:

- Each cluster must not be empty: $C_i \neq \varnothing$.
- Each object must be contained in exactly one cluster: $C_{j_1} \cap C_{j_2} = \varnothing, j_1 \neq j_2$.
- Each object must belong to a cluster: $\cup_{j=1}^{k} C_j = X$.

Some of the properties can be overlooked. For example, we shall see later that OSLOM (see OSLOM) may not assign nodes that don't fulfill certain properties, in a specific module. OSLOM results may also contain overlapping modules (i.e modules that share nodes). INFOMAP (see INFOMAP) may also lead in overlapping modules. UPGMA (see UPGMA), on the other hand, maintains all three properties.

It becomes clear that the community detection problem has many degrees of freedom, that vary from the choice of the measure of similarity as well as the nature of the clusters.

Types of Clusters In general, there are three families of clusters: the overlapping clusters, the non-overlapping clusters and the hierarchical clusters. In the overlapping clusters, there exist nodes that belong to more than one cluster. The clustering, then, is characterized as soft, and the clusters are called covers (see Figure 21a). On the other hand, non-overlapping clusters are called partitions, where no node belongs in more than one cluster (see Figure 21b). The clustering in this case is called hard. Finally, hierarchical clusters occur when there are clusters inside other clusters (see 21c). A method can find hierarchical clusters either through a "bottom-up" or a "top-down" search approach, which defines the method as agglomerative or divisive.



(a) (b) (c)

Figure 21: Examples of graph clustering. (a) is soft clustering. (b) is hard clustering. (c) is hierarchical clustering.

Variables In this part of the theory, more variables will be defined, similar to Graph Theory's definitions (see Definitions). Figures 22, 24, 25 and 26 illustrate a network with adrenocortical carcinoma patients and genes as nodes. Each edge has a patient and a gene as endpoints and its physical meaning is that

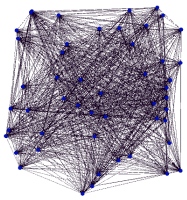
this patient carries a mutation over this gene. The graph of the network is, by definition, bipartite.

Degree We define the degree $deg(u)$ of the node u as the number of the edges that contain u as an endpoint. We can also define the number $d(G) := \frac{1}{|V|}\sum_{v\in V} d(v)$ as the average degree of the graph G. The degree of a node is further explained in paragraph Degree Centrality.

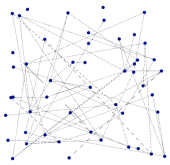
Diameter-Shortest Path We denote the distance $\delta(i,j)$ to be the shortest path from node i to node j and if such path doesn't exist, then we write $\delta(i,j) = \infty$. Diameter is the maximum of the minimum distances and we write, $D = max\delta_{min}(i,j)$, where $\delta_{min}(i,j)$ is the minimum distance between i and j. The average path length is also defined as, $\delta = \frac{2}{N(N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{min}(i,j)$. An example of the shortest path between two nodes is given in Figure 22.

Figure 22: The highlighted pathway is the shortest path of the graph from gene CDKN2A to gene GATA1 and is the following:
CDNK2A → TCGA-OR-A5JM → CTNBB1 → TCGA-OR-A5K9 → SETD2 → TCGA-OR-A5JA → GATA1.

Density The density of the graph is defined as the ratio of the edges in the network over the number of all the possible edges that can take place in the graph. Mathematically, the density is written as $D = \frac{2E}{N(N-1)}$. An example of a dense graph is given in Figure 23a, and an example of a sparse graph is given in Figure 23b.



(a)



(b)

Figure 23: Examples of: (a) a dense graph with 50 nodes, 997 edges and density of 0.814, and (b) A sparse graph with 50 nodes, 57 edges and density of 0.047.

Degree Centrality Degree Centrality shows how "important" a node is via its number of interactions. The higher the degree centrality of the node u, the higher the number of edges that have node u as an endpoint. Specifically, if the graph is undirected, then the degree centrality is equal to the degree $C_{deg}(u)$ as mentioned in Degree. If the graph is directed, $C_{deg}(u)$ can be separated in $C_{deg_{out}}(u)$ and $deg_{in}(u)$ where, $C_{deg_{out}}(u)$ is the number of edges that start from u and $C_{deg_{in}}(u)$ is the number of edges that end up to u. We can also define $C_{deg}(u) = C_{deg_{out}}(u) + deg_{in}(u)$. An example of nodes with high degree centrality is given in Figure 24 and Table 4.

Node	Degree
TCGA-OR-A5KB	18
TP53	17
TCGA-PK-A5HB	16
CTNNB1	14

Table 4: The 4 nodes with the highest degree centrality of graph in Figure 24.

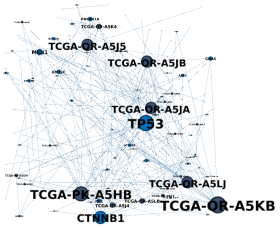


Figure 24: Acc network as described in Variables. The bigger the size of the node, the higher it's degree centrality.

Betweeness Centrality Betweeness centrality is a measure that indicates how much a certain node serves as a "bridge" between other nodes. We denote all the possible paths between i,j nodes,as σ_{ij} and $\sigma_{ij}(w)$ those that pass from node w. Then, $C_b(w) = \sum_{(i,j)\in V(w)} \frac{\sigma_{ij}(w)}{\sigma_{ij}}$, where $V(w)$ is the ordered pair of all the (i,j) elements of $V(G) \times V(G)$ such that i,j are all distinct. $C_b(w)$ is the betweeness centrality of node w. An example of nodes with high betweeness centrality is given in Figure 25 and Table 5.

Node	Betweeness Centrality
TP53	3730.96
TCGA-OR-A5KB	2106.94
TCGA-OR-A5J5	1733.75
TCGA-OR-A5JA	1705.30

Table 5: The 4 nodes with the highest betweeness centrality of graph in Figure 25.

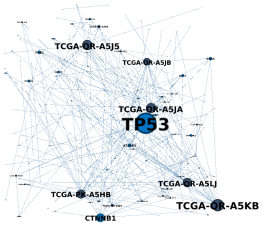


Figure 25: Acc network as described in Variables. The bigger the size of the node, the higher it's betweeness centrality.

Closeness Centrality Closeness centrality of the node u is the variable corresponsive of how close node u is with respect to the rest of the nodes in the network. It is defined as $C_{clo}(i) = \frac{1}{\sum_{j\in V}^{|V|} dist(i,j)}$, where $dist(i,j)$ is the shortest path from i to j. It can also be defined as $C_{clo}(i) = \frac{N-1}{\sum_{j\in V}^{|V|} dist(i,j)}$ where N is the total amount of nodes in the network. An example of nodes with high closeness centrality is given in Figure 26 and Table 6.

Node	Closeness Centrality
HIST1H3B	1.0
TCGA-OR-A5J1	1.0
JAK1	1.0
TCGA-OR-A5LC	1.0

Table 6: The 4 nodes with the highest closeness centrality of graph in Figure 26.

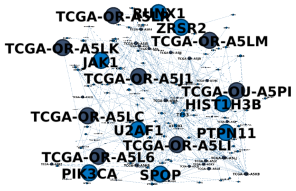
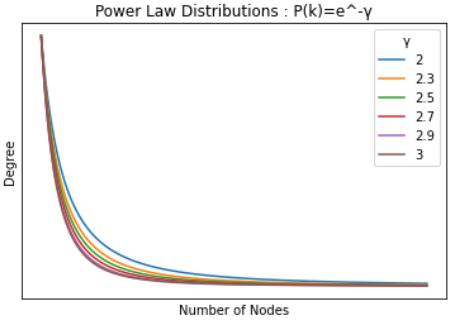


Figure 26: Acc network as described in Variables. The bigger the size of the node, the higher it's closeness centrality.

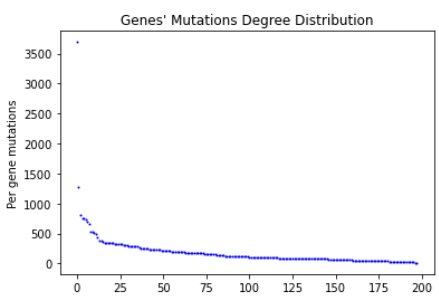
3.3.2 Network Topology

The topology of networks may vary from network to network depending on the nature of the network. Network topology can be of great importance, especially in community detection.

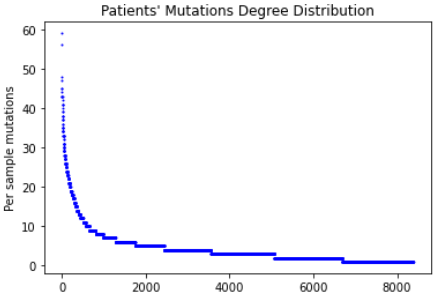
Scale-Free Networks The underlying structure of scale-free networks remains the same even if the network grows in size. They exhibit a power law distribution $P(k) \sim k^{-\gamma}$, where γ is called degree exponent and varies between 2 and 3 (see Figure 27a). In scale-free networks, a small number of nodes with really high degree centrality can be found. A famous example of the real world is the World Wide Web (WWW), where only a fraction of websites (nodes) have most of the visits (edges) with respect to the total amount of websites. Below, Figure 27b and Figure 27c represent the gene and sample degree distributions of the curated TCGA cancer network which will later be explained in Methods.



(a)



(b)



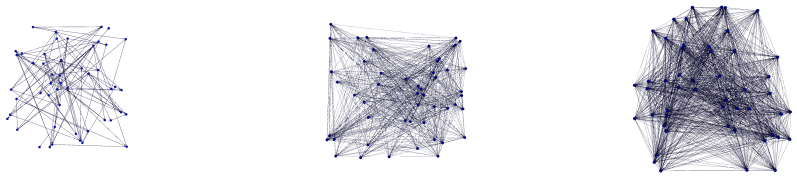
(c)

Figure 27: Power-law distribution ($e^{-\gamma}$) for various exponentials in (a). Degree distribution of (b) gene nodes and (c) patient nodes of the cancer gene-patient network that will be presented in Methods.

3.3.3 Random Graphs

In order to study networks, random graphs can be created. They are computer-generated models, which are built to describe the global structure of networks and they can be based on the degree distribution.

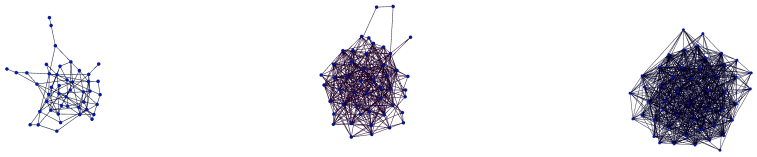
Erdos-Renyi The Erdos-Renyi model refers to a network of nodes that have the same probability of forming a connection with other nodes. It was the first model to be introduced. In this model, a node will link with any other node with a known probability p. Thus, given that the network has n nodes, the degree distribution of each node is binomial with parameters $(n-1, p)$. Let $z = (n-1)p$. Then $p = \frac{z}{n-1}$, and p can be approximated by $\frac{z}{n}$ for large n. Through this notation, the degree distribution is Poisson with parameter z and therefore, $P(deg(u) = k) = e^{-z}\frac{z^{-k}}{k!}$ [67].
For p=1, all of the $\frac{n(n-1)}{2}$ possible edges will be formed and the network will be represented as a complete graph. The smaller the probability p is,the less cohesive the graph will be (see 28) [68].



(a) $p = 0.1$ (b) $p = 0.5$ (c) $p = 0.9$

Figure 28: Erdos-Renyi Graphs for 50 nodes, created in the networkx python's library, for: (a) $p = 0.1$, (b) $p = 0.5$ and (c) $p = 0.9$.

Chung-Lu However, we expect, in some networks, that the degree of the nodes will not likely be the same for every node (see Scale-Free Networks). The Chung-Lu model is a model built based on an expected degree sequence. The degree sequence is denoted by a vector $w = (w_1, w_2, ..., w_n)^T$ which each w_i element is representative of the degree of node i. Then the probability of i being connected to j is proportional to w_i and w_j. Specifically, $P_{ij} = \frac{w_i w_j}{\sum_{k=1}^{n} w_k}$ [69].



(a) max degree = 10
#edges = 117

(b) max degree = 50
#edges = 551

(c) max degree = 100
#edges = 1316

Figure 29: Chung-Lu random graphs of 50 nodes created in the networkx python's library with random degree sequence where the maximum degree of the nodes was: (a) 10, (b) 50 and (c) 100.

3.3.4 Validation

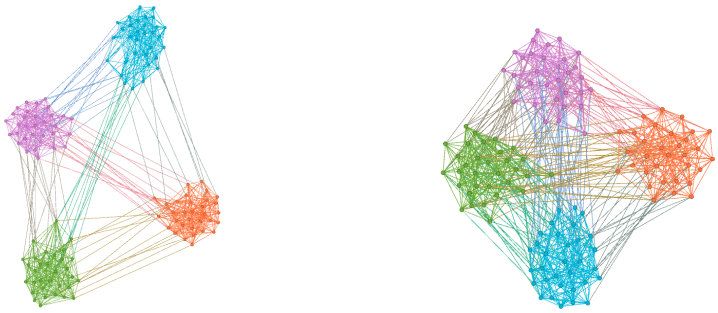
We recall that there is not a specific trait that the nodes of a network must have in order to form a clusters. The general notion, though, is that the nodes of a cluster are more connected with each other, rather than with the rest of the network. This leads to several approximations of the community detection issue and, occasionally, several partitions for the same network. Nevertheless, there are, in general, two ways of evaluating a graph clustering algorithm. Implementing the algorithm over Artificial Benchmarks and over Metadata.

Artificial Benchmarks Artificial Benchmarks are computer-generated models with known clusters. Two of the most famous Artificial Benchmarks are the Girvan-Newman Benchmark and the LFR benchmark.

Girvan-Newman Benchmark Girvan-Newman artificial benchmark consists of 128 nodes divided in 4 groups of 32 nodes each. Expected internal and external degrees are defined as $< k_{in} >= p_{in} n_c$ and $< k_{out} >= p_{out} n_c (q-1)$, where p_{in} and p_{out} are the probabilities of edges being formed within a cluster and between clusters, accordingly. n_c indicates the size of the clusters and q the number of clusters. In Girvan-Newman's case, $n_c = 32$ and $q = 4$. If $< k_{out} >$ is smaller than 8, the groups are well defined and the algorithms should recover them (see Figure 30).

While this benchmark is the most famous one, it doesn't correspond to real-world cases because the nodes of the benchmark have more or less the same degree and this is something not expected, for example, in scale-free networks (see Scale-Free Networks) [19].

LFR Benchmark A variant of GN benchmark, introduced by Lancichinetti-Fortunato-Radicchi (LFR), maintains the heterogeneity observed in networks. This artificial network is built as follows: The degrees of the nodes and the community sizes follow a power law distribution with degree exponent γ and β



(a) $p_{out} = 1/(32 * 3)$ (b) $p_{out} = 1/32$

Figure 30: Girvan - Newman benchmarks for 128 nodes and probability of an edge being formed between two nodes that belong to different clusters (p_{out}) of: (a) $p_{out} = 1/(32 * 3)$ and (b) $p_{out} = 1/32$.

accordingly. The degrees of the nodes are restrained between k_{min} and k_{max}. The community sizes are restrained between s_{min} and s_{max}. These extremes ensure the average degree of the nodes is $< k >$ while for the whole communities the following inequalities take place:

$$s_{min} > k_{min} \text{ and } s_{max} > k_{max}. \tag{1}$$

Both inequalities of 1 ensure that every node will be included in a cluster (see Figure 31). Moreover, the mixing parameter μ_C of the subgraph C is defined as $\mu_C = \frac{k_C^{ext}}{k_C}$ where k_C^{ext} is the sum of the external connections of C with other nodes of the network. Each node will share a fraction $1 - \mu$ with nodes within the same community and μ with the rest of the nodes. The desired μ can be chosen so as to be approximated by the ratio of the internal and external degrees of each node [19].

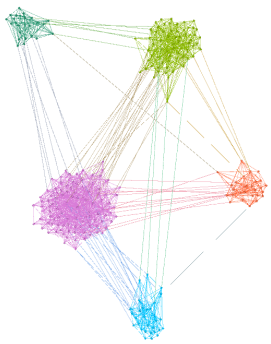


Figure 31: LFR Benchmark with 250 nodes where the power law exponent of the degree distribution equals to 3.

Metadata Metadata consist of real-world networks whose structure is known. That means that we know how the network is divided in clusters and therefore, metadata can be used for algorithmical validation [19].

The dataset of Zackary karate club is one of the most famous examples. It was created through a real-case scenario where the instructor and the owner of the karate club had a conflict and the club was divided in two groups. The network is defined if we consider each person of the club as a node and the interactions between them as edges. In Figure 32, the two large nodes refer to the instructor (blue node) and the owner (purple node), while the color of each node is correspondive of the side each member took. This particular partition was recovered by OSLOM2 implementation, which we shall describe later in 3.4.2.

However, it is posed that metadata might not be a better indicative than the artificial benchmarks due to various reasons. Some of them are human errors when handling the data and/or the irrelevance of the network's structure with the structure of the metadata [70]. Zackary's dataset was download by Konect [71].

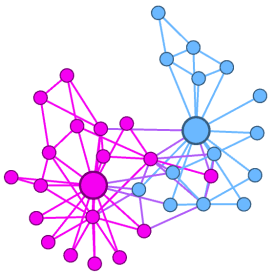


Figure 32: Partition found by OSLOM2 for Zackary's karate club metadata. The large blue node corresponds to instructor of the club while the large purple node to its president.

3.3.5 Methods

Community detection algorithms can be, in general, separated in three main categories. Those that rely on the structure of the network, the ones that use statistics to acquire clusters and, finally, those that are based on the dynamics of the network.

Modularity Random graphs can be used as a baseline for comparison with real cases, serving as null models. Various null models, like the ones described in Random Graphs, may be used in order to compare the edges of the network with the edges of the null models. This is the general idea behind the index

called modularity. Modularity is formally defined as:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j), \tag{2}$$

where m is the number of edges of the network, A_{ij} is the adjacency matrix of the network, P_{ij} is the adjacency matrix of the null model, and δ is the Kronecker delta for C_i,C_j communities. Lastly, the matrix $B_{ij} = A_{ij} - P_{ij}$ is called Modularity matrix. Defining $S = [s_1|s_2|...|s_c]$ where c is the number of clusters, S is a $n \times c$ matrix, where n is the number of the nodes of the network. Each element s_{ij} states that the node i is assigned to the cluster j. Maintaining the 2nd property of the clusters' definition (i.e each object must be contained in exactly one cluster) we derive that S is orthogonal and the equation 2 can now be rewritten as:

$$Q = \frac{1}{2m}TrS^T BS. \tag{3}$$

Equation 3 is an alternative way to compute the modularity index.

Modularity has also been extended to other forms, more suitable to the natural structure of the network [19, 72, 73]. For example, in a bipartite network (see Classes), edges between nodes of the same set of the network may appear in the null model, which will have no physical meaning. Thus, we would like to exclude these nodes and bring Q in a specific form, suitable for bipartite networks. That is exactly what Barber's modularity Q_B does. The matrices A,P are rewritten in a block diagonal form and the modularity matrix then becomes:

$$\tilde{B} = \begin{pmatrix} \mathbb{O}_{pxp} & \tilde{B_{pxq}} \\ (\tilde{B}^T)_{qxp} & \mathbb{O}_{qxq} \end{pmatrix}.$$

One can then compute Barber's modularity index from the formula deriving from 2:

$$Q = \frac{1}{m}\sum_{i=1}^{p}\sum_{j=1}^{q}\tilde{B}_{ij}\delta(g_i, h_j). \tag{4}$$

The p, q quantities refer to the number of nodes of the two bipartite sets respectively and $g_i \in 1, 2, ...c$ refers to the cluster where node i belongs. Finally. we note that $h_j = g_{j+p}$.

In any form of the modularity index, the higher the value of the index is, the more likely that the clusters recovered are "non-random". Specifically, the values of the index range from $[-0.5, 1]$, with the following physical meaning:

$$Q = \begin{cases} < 0, \text{partition found is worst than a random network,} \\ = 0, \text{network is random, one partition constitutes it,} \\ \approx 1, \text{correct split of modules.} \end{cases} \tag{5}$$

Modularity is an NP-hard problem, meaning that there is no algorithm solving the problem in polynomial time [74, 19]. Thus, every maximization of modularity will be correspondive of a local maxima. Furthermore, between two values

of the index, the higher one may not always correspond to the best partition [75]. Therefore, searching for a partition occuring through the maximization of modularity, might be a naive approach. Moreover modularity also involves a null model. It has been shown that modularity itself cannot be representative of the final partition's quality, in terms of partitioning, as high modularity may also occur due to random fluctuations in random networks [76]. Finally, we must note that small scale clusters may not appear in the final partition, through modularity maximization, because the modularity index is dependent on the number m of the edges of the network. This problem is also known as the resolution limit [75, 19].

Several of these issues are solved through ensemble and consensus clustering (further explained in Ensemble and Consensus Clustering). For example, in order to find out if the partition does appear to have a modular structure, the modularity index of the network must be compared to the modularity index of a null model of the same size. Furthermore, consensus clustering can be implemented to combine several solutions that correspond to local maxima to get a more stable solution. Finally, to overcome the network's size dependence, Zhang and Moore treated modularity as a Hamiltonian and combining modularity optimization along with an hierarchical algorithm, they recovered lower level partitions [77].

Dynamics Community detection algorithms may also rely on dynamic processes in order to split the network in modules. One example of such an algorithm is Infomap, which is mentioned later in this project (see INFOMAP). Infomap is based in flow dynamics (i.e the dynamics of a random walker) and the minization of an index called Map Equation. Another example of an algorithm based in dynamics is spin dynamics, which resides in spin-spin interactions and the minimization of a hamiltonian.

3.3.6 Ensemble and Consensus Clustering

In the next session four algorithms will be fully presented. Three of them are based on the structure or the statistics of the network (MODULAR, UPGMA, OSLOM2) and one on its flow (INFOMAP). Simulated annealing is the default algorithmical choice for bipartite networks of MODULAR [24]. This method derives from statistical mechanics. UPGMA [78] is established by the distance between the nodes. OSLOM2 [25] is the only algorithm so far based on statistical significance. INFOMAP [26] is based on random walks and minimum description length statistics.

We shall see later that all but one (UPGMA) out of the four are of stochastic nature. This means that the final partition is biased by the initial seed, to some extent. This lead us to adopt the notion of consensus clustering, where one algorithm is used more than one times. The final partitions are combined, forming an object, such as a new "consensus" adjacency matrix, which will, ultimately, lead to a more stable result [79]. Furthermore the notion of ensem-

ble clustering will be adopted. Ensemble clustering, like ensemble learning in machine learning, can be acquired by combining two or more algorithms which divide the network with different criteria and offer a more stable solution [29].

3.4 Algorithms

3.4.1 MODULAR

MODULAR is a software for community detection based on modularity optimization. In unipartite networks, Q modularity index of Newman and Girvan is used, while in bipartite networks there is a choice between the Q and Q_B (Barber's modularity index). Clustering can occur in five available options [24]:

1. Fast Greedy Algorithm (FG)
2. Simulated Annealing Algorithm (SA)
3. Spectral Partitioning (SP)
4. Hybrid of SA,SP
5. Hybrid of SA-FG

In addition, MODULAR has two preinstalled choices regarding the null models:

1. Erdos-Renyi Model: $\begin{cases} P(i,j) = \frac{E}{RC} & \text{if } R \neq C, \\ P(i,j) = \frac{E}{R(R-1)} & \text{if } R = C. \end{cases}$
2. Null Model 2: $P(i,j) = \frac{1}{2}(\frac{k_{i\in R}}{C} + \frac{k_{j\in C}}{R})$.

The numbers R, C indicate the number of the nodes in each set of the bipartite network. If the network in unipartite, then $R = C$. Also, $k_{i\in R}$ and $k_{i\in C}$ indicate the number of edges of node i in R and in C correspondingly.

Fast Greedy The Fast Greedy algorithm at first, as proposed by Clauset, Newman, and Moore, computes the modularity as if every single node is a cluster itself. For every pair of clusters, the modularity that will occur if two clusters merge, is calculated [80].

$$\Delta Q^{\mathcal{C}}_{c_i,c_j} = Q(G, \mathcal{C} - c_i - c_j + (c_i \cup c_j)) - Q(G, \mathcal{C}). \tag{6}$$

The maximum of those ΔQ is chosen and the algorithm continues until no merge can increase the modularity.
Wakita and Tsurumi, improved the computational efficiency of the algorithm by adding a ratio between the pairs c_i, c_j. Specifically, the $min(\frac{|c_i|}{|c_j|}, \frac{|c_j|}{|c_i|})$ is calculated and the pair (c_i, c_j) that returns the $max\{\Delta Q^{\mathcal{C}}_{c_i,c_j} \cdot \text{ratio}(c_i, c_j)\}$ is joined [81].

Simulated Annealing Simulated Annealing is a stochastic optimization technique. Here the objective is to maximize the modularity and for that reason, the cost is set as $C = -M$, where M is the modularity index. The algorithm randomly exchanges nodes, splits or merges modules and then it computes the modularity. If the result is greater than before, then, this result will be the current solution. If not, this result will be accepted with a probability based on both the previous and the new modularity. Specifically:

$$p = \begin{cases} 1 & \text{if } C_f \leqslant C_i, \\ exp\{-\frac{C_f - Ci}{T}\} & \text{if } C_f > C_i. \end{cases} \tag{7}$$

where C_f is the cost after the update and C_i is the cost before the update. The parameter T is called temperature, which gradually decreases. As the temperature decreases, the probability to accept a worse solution as itself, also decreases. In that way, the algorithm avoids getting trapped in a local maxima. The process continues until a given number of iterations is reached, or when the temperature exceeds a certain threshold [82].

Spectral Partitioning Spectral Partitioning exploits the eigenvectors of matrices correspondive of the network, such as the modularity matrix, the Laplacian, the adjacency matrix, etc. Specifically, let λ_1 be the highest positive eigenvector of the matrix and u_1 the eigenvector, corresponding to λ_1. Then, we define s as the index vector through it elements s_i as follows:

$$s_i = \begin{cases} +1 & \text{if } u_i^{(1)} \geqslant 0, \\ -1 & \text{if } u_i^{(1)} < 0. \end{cases} \tag{8}$$

The network is now divided in two groups according to the sign of the s vector, as described in equation 8. The sequence is repeated until no submatrix has a positive eigenvalue. This process usually results in more than two groups [83].

3.4.2 OSLOM

OSLOM is the first method capable of detecting communities via statistical significance. It compares the network with a random one given the statistics described in the next paragraph. Through it's function, OSLOM is capable of handling directed graphs, weighted graphs, reveal hierarchies, overlapping clusters and community dynamics, presenting a great flexibility over the network's structure [25].

Statistics of OSLOM Let C be a subgraph of the original graph G and let $i \notin C$. We define k_i^{in},k_i^{out} as the neighbors of i in C and in $G \backslash C$ accordingly. Similarly the degree m_C can be separated in m_C^{in},m_C^{out}, while the internal degree of $G \backslash [C \cup i]$ is set as M^*. Then the probability (along with a normalization

factor A) of vertex i having k_i^{in} neighbors is:

$$p(k_i^{in}|i,C,G) = A\frac{2^{-k_i^{in}}}{k_i^{in}!k_i^{out}!(m_C^{out}-k_i^{in})!(M*/2)!}. \quad (9)$$

The cumulative probability $r(k_i^{in}) = \sum_{j=k_i^{in}}^{k_i} p(j|i,C,G)$ of the vertex i to have k_i^{in} or more edges inside C is estimated. On each vertex i, r_i is drawn in a random way from the interval $[r(k_i^{in}), r(k_i^{in}+1)]$, in order to compare vertices with various degrees. Note that introducing r, the stochastic element of the process is revealed. The variable r, will then be used to determine if a topological relation between the external vertex i and C exists*. That leads to the computation of the order statistic distributions of r. Given that $r \sim U(0,1)$, the cumulative distribution of $r_1 = min\{r\}$ in the null model is given by

$$\Omega_1(r) = P(r_1 < r) = 1-(1-r)^{N-n_c}. \quad (10)$$

In general for rank q:

$$\Omega_q(r) = p(r_1 < r) = \sum_{i=q}^{N-n_c} \binom{N-n_c}{i} x^i(1-x)^{N-n_c-i}. \quad (11)$$

Ω_q is an indicator of the compatibility of the external vertices to the null model. Defining $c_m = min_q\{\Omega(r_q)\}$ among all the neighbors of C, the cumulative distribution $P(c_m < x) = \phi(x, N-n_c)$ is calculated and we will call $\phi(x, N-n_c)$ as the score of the cluster C.

*Note: Vertices belonging to the same cluster tend to be connected and so, the statistics cannot be calculated.

Single Cluster Analysis Single Cluster Analysis is a two-step method used to "Clean-up" a given cluster via the aforementioned score ϕ. In order for OSLOM to implement Single Cluster Analysis, a certain threshold P is given as input.

1. For each vertex connected to C, $\Omega_1(r)$ is calculated and, if $\phi = \phi(\Omega_1(r), N-n_c) < P$ then the vertex is added in C. If $\phi > P$ we look for the second best,third best, etc until for some q, $\phi < P$. In that case, all the q vertices are included in C. Otherwise ($\phi < P$ for no vertices) C remains the same. In any case, C is altered to C' (in the last case $C = C'$).
2. For each vertex i of C, r_i is calculated with respect to $C'\backslash\{i\}$. The vertex with the highest r_i is chosen to be excluded from the cluster, in order to perform the first step for $C'\backslash\{i\}$. If i is significant, it's being added again in C'. If not, the procedure repeats for the next worse vertex. Finally a cluster C is formed, that will contain only significant vertices.

Due to the stochastic element of the methodology (bootstrap of r), the Single Cluster Analysis is performed several times. If, for a given module C, there exists a non empty subgraph more than half of the repeats, it's considered to be significant. The nodes that C will contain will have to appear more than half of the times, with respect to the times C was not empty: $(\frac{\#vertex\_appears}{\#cluster\_non\_empty}) > 0.5$

Full Network Analysis OSLOM performs clustering in the following manner:

1. OSLOM randomly chooses single vertices and considers them to be clusters, it adds the q most significant neighbors of each single-vertex cluster, where q is taken from an arbitary distribution (default option for q is a power law distribution with exponent -3).
2. It performs Single Cluster Analysis in every cluster. This is repeated several times, resulting in covers of the network with modules that may or may not overlap. This procedure stops when similar covers are found again and again.
3. To choose between C_k clusters or their union C_u, the clusters are "cleaned-up" within C_u, resulting in $C_k^{'}$. Then, if $|\cup_i C_i^{'}| > P_2|C_u|$ (default option of P_2 is for 0.7), C_u is discarded.
4. OSLOM has found minimal clusters (with no significant internal form) and converts them to supervertices, where the weights of the superedges linking them are calculated based on the number of links between the initial groups. The steps above are repeated in the supernetwork to reveal the hierarchical structure of the network.

For graphs with a large amount of nodes, OSLOM can be used at a second stage, reading an initial partition recovered by a quicker algorithm.

It must be noted that OSLOM2 (used in this project) is significantly faster than the original OSLOM algorithm. That is because, for finding modules it exploits the nearest neighbor algorithms, similar to the Luvain method, as noted here [84]. From now on, OSLOM2 rather than OSLOM will be referenced.

3.4.3 UPGMA

UPGMA, or else Unweighted Pair Group Method with Arithmetic Mean, is a simple agglomerative hierarchical clustering method. The algorithm is based on the average dissimilarity $d(G,H) = \frac{1}{N_G N_H}\sum_{i\in G}\sum_{i'\in H} d_{ii'}$ between the two clusters G, H, exploiting the distance matrix D_{ij} [78]
More specifically, UPGMA works as follows:

1. Find the smallest element of the distance matrix.
2. Merge the two nodes in one cluster and calculate branches.
3. Update the matrix .

4. Repeat until two clusters are left (or else perform (#nodes - 1) iterations).

An illustration is given in Table 7 along with the final dendrogram in Figure 33 of the result:

	A	B	C	D	E
A		2	4	6	4
B			8	4	8
C				10	4
D					6
E					

(a) Minimum value: (A,B).
Elements update:
$(D(A,C) + D(B,C))/2 = 6$
$(D(A,D) + D(B,D))/2 = 5$
$(D(A,E) + D(B,E))/2 = 4$

	(A,B)	C	D	E
(A,B)		6	5	4
C			10	4
D				6
E				

(b) Minimum value: (C,E).
Elements update:
$(D(C,(A,B)) + D(E,(A,B)))/2 = 5$
$(D(C,D) + D(E,D)/2 = 8$

	(A,B)	(C,E)	D
(A,B)		5	5
(C,E)			8
D			

(c) Minimum value: ((A,B),D).
Elements update:
$(D((A,B),(C,E)) + D(D,(C,E)))/2 = 7.5$

	((A,B),D)	(C,E)
((A,B),D)		7.5
(C,E)		

(d) Final iteration. Algorithm stops resulting in the following:
(((A,B),D),(C,E)).

Table 7: UPGMA implementation over the distance matrix (a). Note that this matrix is correspondive of an undirected graph with no self-loops. Thus, the matrix is symmetrical, the entries of the diagonal are 0, and the upper diagonal part of the matrix is the only one needed. Below each table, the minimum value of each matrix is mentioned (highlighted in red) and the values of the updated matrix are calculated (highlighted in green).

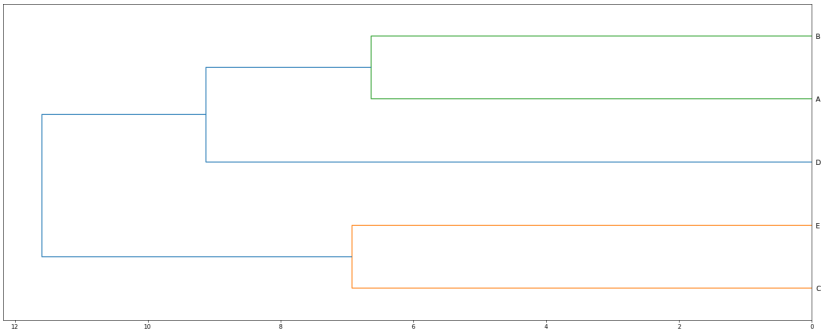


Figure 33: The resulting dendrogram of UPGMA implementation over the distance matrix 7a.

3.4.4 INFOMAP

Minimum description length statistics (MDL) provide the guidelines to find a solution to the community detection problem through the flow of the information in the network. To represent that flow, INFOMAP exploits random walks and represents their trace with a compressed message, combining information theory along with stochastic processes. The heart of the software is to be able to describe the locations of random walks as efficient as possible, depicting the underlying structure of the network. Specifically, the map equation is the one that handles the duality of the detection of the modular nature of the network and the minimization of the description length of the steps of a random walker.

Coding Structure In order to capture the movements of the random walker, a codeword is in each node. This codeword is used to describe the location of the node. Specifically, Huffman codes are used [85], assigning short or long codewords to each of them according to the average visit frequency of an infinite length random walk. Short codewords correspond to common locations and long ones correspond to more rare locations. A codebook consists of all the codeworks.
Pursuing the concept that modules appear to have a certain level of autonomy, with respect to the rest of the network, due to high intramodular links, the random walker will tend to stay longer in a group before exiting. Thus, rather than a single codebook for the whole network, several codebooks are created, enabling the reusing of short codewords. Consequently the overall description length required is reduced. To do so, an index codebook is created in order to define when the walker exits a module in order to enter another.
As far as data compression is concerned, the Shannon's source coding theorems will provide a lower bound for the average description length of a codeword [86]:

$$L(X) = H(X) = -\sum p_i log_2(p_i). \tag{12}$$

Equation 12 is the entropy of the random variable X with n states and frequencies p_i. Code lengths are measured in bits and thus log_2 is used. Later, X will be replaced by the distribution where the p_i will be the frequencies of the visits of the nodes.

The Map Equation

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i). \tag{13}$$

The first term of equation 13 corresponds to the average number of bits used to describe inter-modular movements. The second term is indicative of the intra-modular movements.

Individually:
given that the network is partitioned in m modules, the probability of the random walker to switch modules (per-step probability) is:

$$q_{\curvearrowright} = \sum_{j=1}^{m} q_{\curvearrowright}, \tag{14}$$

while the following equation represents the amount of time the codebook of module i is used. It consists of the sum of the steady state distribution for all α nodes within the modules plus the probability it exits the module:

$$p_{\circlearrowright} = q_{\curvearrowright} + \sum_{\alpha \in i} p_{\alpha}. \tag{15}$$

The other two quantities presented in the map equation are the entropies of the index and module i codebook accordingly:

$$H(\mathcal{Q}) = \sum_{i=1}^{m} \frac{q_{\curvearrowright}}{\sum_{j=1}^{m} q_{\curvearrowright}} log(\frac{q_{\curvearrowright}}{\sum_{j=1}^{m} q_{\curvearrowright}}), \tag{16}$$

$$\begin{aligned} H(\mathcal{P}^i) = & \frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_{\beta}} log(\frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_{\beta}}) + \\ & \sum_{\alpha \in i} \frac{p_{\alpha}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_{\beta}} log(\frac{p_{\alpha}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_{\beta}}). \end{aligned} \tag{17}$$

Infomap is implemented in C++ and it firstly uses a greedy approach to minimize the map equation. The steady state visit frequency for each node is calculated, each node consists of a unique module and the exit probabilities are calculated. Then, similar to MODULAR's Fast Greedy algorithm, the two modules that, when combined in one group decrease the map equation more efficiently, are merged. The process continues until module combinations result in higher than the previous state description length. After that, simulated annealing is used to further reduce the result [26].

4 Methods

4.1 Data Preprocessing

In order to build a gene-patient network, data of TCGA-PanCancer Atlas project were downloaded from cBioPortal [87, 88] referring to 29 cancer types as listed in Table 8. Only primary samples and somatic mutations were exploited out of the data. Also, if a sample had over 3000 mutations (with respect to every gene of the data) and thus, being a hypermutator, then this sample was excluded. The 198 genes included derived from [89]. The only mutations taken into consideration were missense, nonsense, nonstop, frameshifts, in frame deletions and insertions, and mutations occuring in splice sites.

When the initial data were curated, they were exported to another folder, which was the one used to build the network. The data-cleanup procedure of preprocessing is also described in Algorithm 1.

Algorithm 1: Data Preprocessing

```
Input: path_raw = Path where raw cBioPortal data is stored
path_fixed = path where curated files will be
for folder in path_raw do
    df = mutation file
    df = remove hypermutators(threshold = 3000)
    df = keep genes(df)
    df = keep certain mutations(df)
    df = remove metastatic samples(df)
    df = ["Gene","Patient","Mutation"] form
    export df to path_fixed
```

The adjacency matrix $\{a_{ij}\}$ of the network will be binary, where "1" will indicate the existance of at least one of the aforementioned mutations in gene i of patient j. "0", on the other hand, will be indicative of the absence of such mutations. The size of the biadjacency matrix was 198×8386 (198 genes and 8386 patients).

Due to the large size of the network, a smaller subnetwork containing the 6 cancer types with the most patients in the curated dataset was used to derive preliminary results of the methodology over clustering. The size of the biadjacency matrix of the subnetwork was 198×3402 (198 genes and 3402 patients), which is significantly smaller than the 198×8386 network.

4.2 Clustering Analysis

Ensemble and Consensus Clustering MODULAR's simulated annealing and INFOMAP were the algorithms used to obtain an initial partition. Both algorithms are implemented based on the ensemble and consensus clustering approaches as referred in Ensemble and Consensus Clustering. As far as the

consensus part of the methods, Monti's algorithm was implemented [90]. That is, for every partition h recovered by either of the algorithms, we define the matrix M^h and the (i, j) pair of the matrix M^h was calculated as follows:

$$M^h(i,j) = \begin{cases} 1 & \text{if items i and j belong to the same cluster,} \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Then, every matrix M^h was used with the final M matrix being defined as:

$$M = \frac{\sum_h M^h(i,j)}{\text{total amount of partitions.}} \tag{19}$$

INFOMAP-OSLOM2 pipeline INFOMAP uses the link list of the network and runs in matter of a few seconds but the number of clusters varies each time due to the stochasticity of the algorithm [26]. This fact justifies the use of consensus clustering, where the consensus matrix M_I was obtained (through Monti's consensus clustering) after 200 INFOMAP runs. The options of the INFOMAP were set to get a "two-level" partition over an "undirected" graph (unipartite version of INFOMAP). M_I was weighted with elements varying between 0 and 1. All elements whose values were below 0.4 were zeroed and the final matrix was converted into an unweighted (binary) matrix M_I^*. INFOMAP was run again once over the link list of matrix M_I^*, with the same options as the ones of the 200 runs to get the partition C_{info}

OSLOM2 was implemented in the partition C_{info} and excluded the non statistical significant communities that had a P-value lower than 0.05, which resulted in nodes being left unassigned to clusters and the final partition C^*_{info}. To obtain a full network partition, OSLOM2 provides a result where no node is left unassigned with the cost of bypassing the P-value threshold. This result won't be used here, due to low amount of MODULAR results and the low amount of threshold used to define a significant cluster with respect to each cancer type. Nevertheless, the contribution of statistics in the result gives grounds for OSLOM2 implementation and thus, ensemble clustering. The INFOMAP-OSLOM2 methodology is presented in Algorithm 2.

The ensemble approach of INFOMAP-OSLOM2 was compared to the INFOMAP approach by implementing both approaches in the subnetwork.

MODULAR-UPGMA-OSLOM2 pipeline As far as MODULAR is concerned, we recall that simulated annealing is the default method used for community detection over bipartite networks, which is computationally expensive. For a single outcome over the 29-types network, one day (more or less) was required get one partition of the network and the modularities for the two null models. For the subnetwork, the amount of time required was significantly less, at about 3 hours each. All the results (45 partitions for the network and 100 for the subnetwork) were exploited through consensus clustering (Monti's consensus clustering) to get the consensus matrix M^*_M.

Then, in order to overcome the resolution limit of modularity as an index of reference (see Ensemble and Consensus Clustering), an hierarchical clustering method was utilized. Consensus clustering matrix M_M^* represents a similarity matrix of the network (like an adjacency matrix [29]) and thus, $1 - M_M^*$ will serve as a dissimilarity matrix [90] (like a distance matrix [29]). Note that M_M^* is weighted. In that manner, UPGMA was implemented over $1 - M_M^*$. The dendogram produced had several branches. For each branch, nodes below that branch would be colored in a specific color (indicative of the module they belong via UPGMA implementation), while nodes above that branch would be colored with the same color, "grey" for example. Each "grey" node would consist of a single-node module. Nodes below the branch that are of the same color would belong in the same module. For every possible branch, the Barber's modularity index (with respect to the original network) was calculated. The branch, whose correspondive partition maximized the index was considered as the optimal branch. Let C_{mod}^* be the partition that corresponds to the optimal branch.

Finally, OSLOM2 was implemented over the C_{mod} partition, just like in the INFOMAP consensus pipeline to get the final partition C_{mod}^*. The MODULAR-UPGMA-OSLOM2 methodology is presented in Algorithm 3.

Algorithm 2: INFOMAP + OSLOM2

```
Input: link_list = Link list of network
Output: Partition of infomap - consensus + OSLOM2
G = Multigraph()
options_infomap = two level - undirected
options_oslom = lowest level partition - p_value < 0.05
for times in range(200) do
    result = run_infomap (link_list)
    G = add_to_consensus (G, result)
G = G/200
G = G[G > 0.4]
consensus_link_list = link list of adjacency matrix of G
result_consensus = run_infomap(consensus_link_list, options_infomap)
result_infomap_oslom = run_oslom(result_consensus, options_oslom)
return result_infomap_oslom
```

Algorithm 3: MODULAR + UPGMA + OSLOM2

```
Input: adj = Adjacency matrix of network
Output: Partition of MODULAR + UPGMA + OSLOM2
options_oslom = lowest level partition - p_value < 0.05
G = Multigraph()
for times in range(200) do
    result = run_modular (adj,options = default)
    G = add_to_consensus (G,result)
G = G/200
distance_matrix = 1 - G
Z = linkage(distance_matrix, method = 'average')
branches = possible distances of nodes
upgma_all = []
for branch in branches do
    dn = dendrogram(cut_point = branch)
    result_temp = result based on leaves' colors
    modularity_upgma = barber_modularity(result_temp)
    upgma_all.append(branch,modularity)
optimal_branch = branch that maximized modularity
dn_final = dendrogram(cut_point = optimal_branch)
result_modular_upgma = result based on leaves' colors
result_modular_upgma_oslom = run_oslom(result_upgma)
return result_modular_upgma_oslom
```

4.3 Modular Structure

In order to ensure that both the network (29 cancer types) and the subnetwork (6 cancer types) had modular structures and clustering could indeed by implemented, data of MODULAR runs were used. Specifically, we recall that in each run of the simulated annealing algorithm of MODULAR, the algorithm was also used on the two null models that MODULAR provides. After all runs, the modularities of the network (or the subnetwork) and the modularites of the two null models consisted of three separate vectors. These vectors were used to perform Welch's t-test:

$$t = \frac{\tilde{X_1} - \tilde{X_2}}{\sqrt{\sigma^2_{\tilde{X_1}} + \sigma^2_{\tilde{X_2}}}}. \tag{20}$$

The terms $\tilde{X}$ and $\sigma^2_{\tilde{X}}$ are the mean and variance of the modularities. $\tilde{X_1}$ for example is the mean of the vector containing the modularities of the network. The test was performed twice for each network; the first to compare the modularities of the network (or the subnetwork) to the modularities of the first null model and the second to compare them to the modularities of the second null

model. Tables 9 and 18 in the Results section contain the resulting p values.

4.4 Biological Analysis

Given a partition C^* being either C^*_{info} or C^*_{mod}, every cluster C in C^* consisted of patients and nodes. Here, a 10-patient threshold was adopted in the following manner:

- For every cluster C.
- For every cancer type x.
- If $|x| \geqslant 10$.
- For every gene y of C.
- Assess if gene y is driver for cancer type x through current literature.

This analysis lead into the "Cancer-Gene association" two-column Tables (see Tables 17, 25). Each line of the "Cancer-Gene association" Table has (a) the cancer type in the first column and (b) the genes of the modules in which there were more than (or exactly) 10 patients of that cancer type. Each of the genes on the right was looked up through current literature to verify if it is, indeed, a driver gene for that cancer type. cBioPortal [87, 88] provided with charts that were also employed in order to further analyze the results, regarding the subtype (or other traits).

4.5 Cancer types

Table 8: Cancer Types Abbreviations as listed in TCGA [91].

acc	Adrenocortical carcinoma
blca	Bladder Urothelial Carcinoma
brca	Breast invasive carcinoma
cesc	Cervical squamous cell carcinoma and endocervical adenocarcinoma
coadread	Colorectal adenocarcinoma
esca	Esophageal carcinoma
gbm	Glioblastoma multiforme
hnsc	Head and Neck squamous cell carcinoma
kich	Kidney Chromophobe
kirc	Kidney renal clear cell carcinoma
kirp	Kidney renal papillary cell carcinoma
laml	Acute Myeloid Leukemia
lgg	Brain Lower Grade Glioma
lihc	Liver hepatocellular carcinoma
luad	Lung adenocarcinoma
lusc	Lung squamous cell carcinoma
meso	Mesothelioma
ov	Ovarian serous cystadenocarcinoma
paad	Pancreatic adenocarcinoma
pcpg	Pheochromocytoma and Paraganglioma
prad	Prostate adenocarcinoma
sarc	Sarcoma
skcm	Skin Cutaneous Melanoma
stad	Stomach adenocarcinoma
tgct	Testicular Germ Cell Tumors
thca	Thyroid carcinoma
thym	Thymoma
ucec	Uterine Corpus Endometrial Carcinoma
ucs	Uterine Carcinosarcoma

5 Results

5.1 Six cancer types subnetwork

Modular Structure of the Subnetwork

MODULAR was run 100 times for the subnetwork that contained 6 cancer types. Via the 100 runs, the mean modularity of the subnetwork was compared to each of the null models of MODULAR (see Modular Structure). The histogram in Figure 34 along with Table 9, are indicative of the modular structure of the subnetwork.

	Null Model 1	Null Model 2
Subnetwork	1.36e-135	5.37e-148

Table 9: p-value to asess statistical difference of the mean modularity of the subnetwork partitions and the null models.

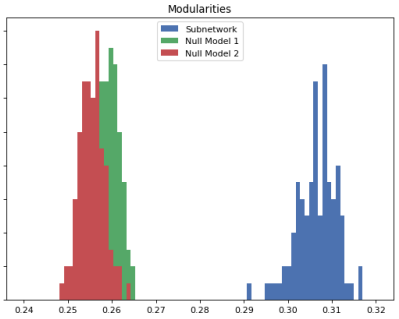


Figure 34: Comparing Barber's modularity of the subnetwork partitions to the null models created through MODULAR.

After assessing the modular structure of the subnetwork, both clustering analysis pipelines were implemented (see Clustering Analysis). After the implementation, the percentage that corresponds to 10 patients for each cancer type (with respect to the total amount of patients of the cancer type) was calculated (see Biological Analysis). The calculated percentages are presented in Table 10.

Cancer Type	10-patients Percentage
brca	1.07%
coadread	2.03%
gbm	2.93%
hnsc	2.04%
lgg	2.02%
luad	1.87%
ucec	2.15%

Table 10: Percentages correspondive to 10 patient with respect to the total amount of patients (per cancer).

INFOMAP Results

		Modules											
		0	1	2	3	4	5	6	7	8	9	10	11
Cancer Types	brca	55.59	0.86	9.46	0.86	7.2	6.88	3.33	3.01	2.47	2.47	2.47	1.4
	coadread	94.5	0.61	0.2	0.0	0.41	0.0	0.61	0.81	0.61	0.0	0.0	0.2
	hnsc	93.66	0.61	0.61	0.61	0.2	0.0	0.2	0.0	0.0	0.2	0.2	0.41
	lgg	9.94	74.24	0.0	6.49	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.61
	luad	85.39	1.12	0.0	7.68	0.37	0.19	0.0	0.37	0.56	0.19	0.19	0.37
	ucec	93.98	0.0	0.22	0.0	0.0	0.0	0.0	0.0	0.22	0.43	0.0	0.22

(a)

		Modules										
		12	13	14	15	16	17	18	19	20	21	22
Cancer Types	brca	0.65	0.0	0.43	0.22	0.75	0.22	0.11	0.75	0.43	0.32	0.11
	coadread	0.0	0.0	0.2	1.22	0.2	0.2	0.0	0.0	0.2	0.0	0.0
	hnsc	0.2	0.0	2.04	0.0	0.2	0.41	0.2	0.0	0.2	0.0	0.0
	lgg	0.2	4.06	0.0	1.42	0.41	0.0	0.61	0.61	0.41	0.61	0.0
	luad	0.0	0.0	0.37	0.37	0.37	1.12	0.94	0.19	0.19	0.0	0.0
	ucec	2.58	0.22	0.0	0.22	0.43	0.65	0.43	0.22	0.22	0.0	0.0

(b)

Table 11: Modules recovered after consensus clustering over 200 INFOMAP partitions. The cells of Tables (a) and (b) refers to the percentage of the patients of the cancer type of the row contained in the module of the column.

Module	Genes
0	172 genes
1	IDH1, ATRX, CDKN2C, CIC, FUBP1
2	GATA3, CBFB
3	EGFR
4	MAP3K1
5	CDH1
6	RUNX1
7	MAP2K4
8	TBX3
9	AKT1
10	FOXA1

Module	Genes
11	SF3B1
12	SPOP
13	IDH2
14	CYLD
15	PTPN11
16	CDKN1B
17	U2AF1
18	MAX
19	DDX5
20	PRKAR1A
21	H3F3A
22	N/A

Table 12: Genes of modules as recovered by INFOMAP consensus clustering in Table 11.

INFOMAP - OSLOM2 Results

		Modules					
		0	1	2	3	4	5
Cancer Types	brca	0.32	6.67	3.76	1.08	2.69	0.0
	coadread	0.41	0.0	0.0	0.0	0.2	0.0
	hnsc	0.41	0.0	0.2	0.0	0.41	0.0
	lgg	63.49	0.0	0.0	0.0	0.41	4.06
	luad	0.19	0.0	0.0	0.0	0.19	0.0
	ucec	0.0	0.0	0.0	0.0	0.22	0.0

Table 13: Modules recovered after ensemble clustering of 200 INFOMAP partitions (consensus clustering) and OSLOM2 implementation. The cells of the table refers to the percentage of the patients of the cancer type of the row, contained in the module of the column.

Module Infomap -Oslom2	Genes	Module Infomap Blend
0	ATRX,CDKN2C,CIC, FUBP1,IDH1,TP53	1+0
1	CBFB,GATA3,PIK3CA	2+0
2	MAP3K1,PIK3CA	4+0
3	CDH1,RUNX1	5+6
4	FOXA1	10
5	IDH2	13

Table 14: Genes of modules as recovered by the ensemble clustering approach of INFOMAP (consensus clustering) and OSLOM2. OSLOM2 exchanges nodes from one module to another. Each element of the "Module Infomap Blend" column refers to the exchanging of genes with respect to Table 12. For example, the third module in this table contains the genes CDH1 and RUNX1 of modules 5 and 6 of INFOMAP.

Comparing the results from Tables 11,12,13,14 we can make two remarks, regarding the efficency of INFOMAP-OSLOM2 ensemble clustering approach over INFOMAP approach.

Firstly, module 0 of Table 11 is the module that contains 172 genes (out of 198), as well as the highest percentage of patients for all but one cancer type (lower grade glioma). Module 0 appears most likely due to the partial modular structure of the network [92, 76], and it is not considered statistically significant for OSLOM2 to include it in the final partition. However, besides module 0, the number of clusters is decreased, leading to more robust results, as many modules (module 15 to module 22) didn't have more than 10 patients for any of the cancer types of the subnetwork. Thus, these modules not only didn't

suffice for the 10 patient threshold as mentioned in Biological Analysis, but also weren't statistically significant through the statistics of OSLOM2.

Secondly, we recall that OSLOM2 can exchange nodes from one cluster to another by implementing "Single Cluster Analysis" locally in every cluster (see OSLOM). This leads in clustering of several patients and genes together. As far as the genes are concerned, OSLOM2 managed to move genes from the statistically insignificant module (module 0 of Table 11) in statistically significant modules. Specifically, in the third column, named "Module Infomap Blend", this becomes clear, as 3 out of the 6 clusters of the final partition, contain genes of the statistically insignificant group of the partition of INFOMAP.

MODULAR-UPGMA-OSLOM2 Results

		Modules											
		0	1	2	3	4	5	6	7	8	9	10	11
Cancer Types	brca	0.75	0.32	0.22	0.0	6.67	2.8	1.29	18.06	0.75	0.0	0.54	0.32
	coadread	0.2	0.41	0.81	0.2	0.2	0.61	53.97	1.22	0.2	0.2	1.02	0.2
	hnsc	0.61	0.41	0.61	20.45	0.0	0.41	2.66	1.43	1.02	0.0	15.95	1.02
	lgg	6.49	0.0	0.0	0.41	0.0	0.0	0.2	0.41	47.26	4.26	7.3	0.0
	luad	7.3	0.94	0.75	3.18	0.0	0.19	3.18	0.19	0.75	0.0	2.43	12.55
	ucec	0.0	38.06	11.61	0.22	0.0	1.51	4.95	0.65	0.22	0.0	0.65	0.22

Table 15: Modules recovered after ensemble clustering of 100 MODULAR partitions (consensus clustering), UPGMA implementation and OSLOM2 implementation. The cells of the table refer to the percentage of the patients of the cancer type of the row, contained in the module of the column.

Module	Genes
0	EGFR
1	ACVR1,ARID1A,CCND1, CTNNB1,FGFR2,NFE2L2, PIK3R1, PTEN
2	PPP2R1A
3	CDKN2A
4	MAP3K1
5	AKT1

Module	Genes
6	APC,FBXW7,KRAS,NRAS, SMAD4,TCF7L2,TP53
7	CDH1,FOXA1,GATA3, PIK3CA
8	ATRX, IDH1, TP53
9	IDH2
10	NOTCH1
11	STK11

Table 16: Genes of modules as recovered by the ensemble clustering approach of MODULAR (consensus clustering), UPGMA and OSLOM2 in Table 15.

Cancer - Gene association

brca	MAP3K1,CDH1,FOXA1,GATA3,PIK3CA,AKT1,RUNX1,CBFB
coadread	APC, FBXW7,KRAS,NRAS,SMAD4,TCF7L2,TP53
hnsc	CDKN2A,APC,FBXW7,KRAS,NRAS,SMAD4,TCF7L2,TP53
lgg	ATRX, IDH1, TP53, IDH2, EGFR, NOTCH1, FUBP1,CIC,CDKN2C
luad	EGFR,CDKN2A,APC,FBXW7,KRAS,NRAS,SMAD4,TCF7L2,TP53
ucec	ACVR1,ARID1A,CCND1,CTNNB1,FGFR2,NFE2L2,PIK3R1,PTEN, PPP2R1A,APC,FBXW7,KRAS,NRAS,SMAD4,TCF7L2,TP53

Table 17: Genes linked with more than 10 patients through clustering and colors correspond to pipelines recovering them.
Blue is for MODULAR-UPGMA-OSLOM2.
Green is for INFOMAP-OSLOM2.
Red is for both.

The blue highlighted cells of Tables 13 and 15, along with Tables 14 and 16, were used to create the Cancer-Gene association table, which is Table 17.

TCGA studies for all 6 cancer types can validate that most of the genes linked to certain cancer types (as viewed in the Table 17), are considered driver genes for the correspondive cancer type [93, 94, 95, 96, 97, 98]. Thus, in order to validate the accuracy of the outcome, we will proceed in the network analysis of the 29 cancer types and compare the results to current literature.

5.2 Twenty-nine cancer types network

Modular Structure of the Network

MODULAR was run 45 times for the network, that contained 29 cancer types. Via the 45 results, the mean modularity of the subnetwork was compared to each of the null models of MODULAR (see Modular Structure). The histogram in Figure 35 along with Table 18, are indicative of the modular structure of the network.

	Null Model 1	Null Model 2
Network	8.32e-59	1.22e-65

Table 18: p-value to asess statistical difference of the mean modularity of the network partitions and the null models.

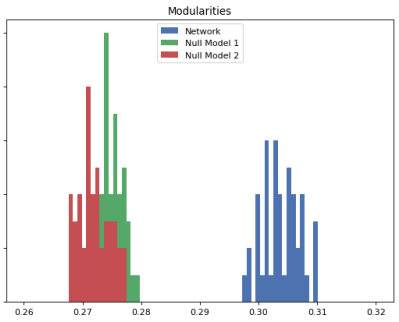


Figure 35: Comparing Barber's modularity of the network partitions to the null models created through MODULAR.

After assessing the modular structure of the subnetwork, both clustering analysis pipelines were implemented (see Clustering Analysis). After the implementation, the percentage that corresponds to 10 patients for each cancer type was calculated (see Biological Analysis). The percentages are presented in Table 19. Furthermore, a numerical table considering the patients and the mutations per cancer type, is presented in Table 20.

Cancer Type	10-patients Percentage
acc	16.94%
blca	2.51%
brca	1.07%
cesc	3.90%
coadread	2.03%
esca	5.64%
gbm	2.93%
hnsc	2.04%
kich	26.31%
kirc	3.16%

(a)

Cancer Type	10-patients Percentage
kirp	4.67%
laml	5.52%
lgg	2.02%
lihc	2.97%
luad	1.87%
lusc	2.18%
meso	15.15%
ov	2.50%
paad	6.53%

(b)

Cancer Type	10-patients Percentage
pcpg	12.19%
prad	3.34%
sarc	5.74%
skcm	13.69%
stad	2.45%
tgct	15.15%
thca	2.59%
thym	20.0%
ucec	2.15%
ucs	18.18%

(c)

Table 19: Percentages correspondive to 10 patient with respect to the total amount of patients (per cancer).

Cancer	Patients	Missense Mutation	Nonsense Mutation	Nonstop Mutation	Frame Shift Del	Frame Shift Ins	In Frame Del	In Frame Ins	Splice Site	Total
acc	59	117	32	0	26	5	5	0	8	193
blca	398	2159	533	4	165	77	23	2	127	3090
brca	930	1795	322	2	316	234	66	8	149	2892
cesc	256	905	197	1	48	13	7	2	38	1211
coadread	491	2516	541	0	567	186	44	5	102	3961
esca	177	553	88	0	71	32	13	7	33	797
gbm	341	760	113	1	95	28	22	4	54	1077
hnsc	489	1609	379	0	154	74	30	3	120	2369
kich	38	50	8	0	9	1	1	0	4	73
kirc	316	448	138	3	190	49	11	3	67	909
kirp	214	396	46	0	67	20	4	1	20	554
laml	181	294	50	0	31	88	4	37	34	538
lgg	493	1093	140	0	217	68	57	2	65	1642
lihc	336	748	97	1	88	36	21	5	59	1055
luad	534	2555	367	0	139	46	48	9	156	3320
lusc	458	2062	349	3	165	46	29	3	150	2807
meso	66	64	28	1	19	4	4	0	12	132
ov	399	798	111	1	92	78	24	1	69	1174
paad	153	319	58	1	43	22	11	2	21	477
pcpg	82	66	4	0	11	4	2	1	3	91
prad	299	386	42	1	75	30	20	0	19	573
sarc	174	309	46	0	57	11	5	0	32	460
skcm	73	500	59	0	7	6	2	1	18	593
stad	407	1901	237	0	534	132	50	4	94	2952
tgct	66	73	4	0	7	2	3	0	1	90
thca	386	479	27	0	11	3	3	0	4	527
thym	50	71	9	0	10	1	0	1	0	92
ucec	465	2777	405	3	698	198	128	21	126	4356
ucs	55	183	18	0	16	8	4	1	8	238
Total	8386	25986	4448	22	3928	1502	641	123	1593	38243

Table 20: Numerical table considering the number of patients per cancer and the number of mutations per cancer type and in total.

MODULAR-UPGMA-OSLOM2 Results

Cancer Types	Modules																					
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
acc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.03	1.69	3.39	0.0	0.0	0.0	0.0	0.0	0.0	1.69	1.69	1.69	0.0	1.69
blca	0.25	0.5	0.25	0.75	0.5	0.75	0.0	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.25	0.25	0.25	0.0	0.0	0.0	0.5	0.75
brca	0.22	0.0	0.11	0.32	0.22	0.43	0.32	2.37	0.11	1.08	0.86	0.43	0.54	0.0	0.0	11.72	0.0	0.11	12.37	0.11	0.65	10.43
cesc	0.0	0.39	1.56	1.17	1.95	2.73	4.69	1.95	0.0	3.12	0.39	0.78	0.78	0.0	0.78	0.78	0.0	0.0	0.39	0.39	1.56	0.0
coadread	0.0	0.41	36.46	1.63	0.81	8.35	0.2	1.43	0.81	1.22	0.0	1.63	0.2	0.0	0.0	0.41	1.02	0.0	0.61	0.41	0.2	1.22
esca	0.0	0.0	0.56	0.56	1.69	5.08	1.13	1.69	1.69	0.0	0.56	2.26	0.0	0.0	0.56	1.13	0.56	0.56	0.0	2.26	0.0	2.26
gbm	6.45	0.0	0.0	1.17	0.29	0.0	0.0	9.68	0.0	1.76	16.42	1.47	0.0	0.0	0.29	0.59	0.29	0.0	0.29	0.59	0.0	1.47
hnsc	0.2	0.0	0.2	0.41	0.61	2.45	1.02	1.02	0.41	1.23	0.41	0.61	1.84	0.0	0.41	0.61	0.2	0.2	0.2	0.61	0.41	0.41
kich	0.0	0.0	0.0	2.63	0.0	0.0	0.0	0.0	0.0	2.63	0.0	0.0	0.0	0.0	2.63	0.0	0.0	0.0	0.0	0.0	0.0	0.0
kirc	0.0	0.0	0.0	0.32	0.0	0.0	0.0	0.32	0.32	13.61	0.32	0.0	1.58	0.0	51.27	0.32	0.0	0.0	0.0	0.95	6.96	0.95
kirp	0.0	0.0	2.34	0.47	1.4	0.93	0.93	0.93	0.47	7.01	0.0	0.0	2.8	0.0	2.34	0.47	1.4	0.0	0.47	1.87	7.94	1.4
laml	0.0	0.0	1.1	48.62	0.0	0.0	0.0	0.0	0.0	0.55	0.55	4.42	27.07	0.0	0.0	0.0	0.0	0.0	7.18	8.84	0.0	0.0
lgg	46.25	0.0	0.0	0.41	0.0	0.0	0.0	3.65	0.61	1.62	5.68	1.01	2.03	0.0	0.2	0.0	0.41	0.0	0.2	19.68	0.0	0.61
lihc	0.3	0.6	0.0	0.89	0.89	1.19	0.0	1.19	19.94	3.57	0.6	2.38	0.89	0.0	0.89	0.0	0.0	0.3	0.3	0.89	5.65	1.19
luad	0.37	0.0	2.62	0.75	0.75	2.43	12.36	0.19	1.31	3.18	6.37	0.56	2.81	0.0	0.0	0.37	0.94	0.0	0.19	0.37	0.56	0.56
lusc	0.44	0.0	0.66	0.87	1.31	1.75	0.87	0.66	0.22	1.31	0.66	2.4	1.09	0.0	0.22	0.44	0.22	0.0	0.44	0.0	0.44	0.44
meso	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.12	0.0	1.52	3.03	0.0	4.55	0.0	1.52	0.0	0.0	0.0	48.48	0.0
ov	2.26	0.0	0.5	1.75	1.25	0.75	0.0	0.25	0.0	2.01	0.75	2.01	1.0	0.0	0.25	0.0	0.25	0.0	0.75	1.0	1.0	1.0
paad	0.0	0.0	26.14	0.0	0.65	23.53	1.96	0.0	1.31	0.65	0.0	0.0	1.96	0.0	1.96	0.0	0.0	0.0	0.0	0.65	0.65	0.0
pcpg	0.0	0.0	1.22	0.0	0.0	0.0	0.0	0.0	0.0	2.44	0.0	1.22	1.22	21.95	3.66	0.0	1.22	0.0	0.0	0.0	0.0	0.0
prad	0.0	0.67	0.33	0.67	0.0	1.34	0.0	0.0	3.34	2.34	0.67	0.33	0.33	1.0	0.67	1.0	1.0	13.38	0.33	0.33	0.33	8.7
sarc	6.9	0.0	0.57	0.0	0.57	0.57	0.0	0.0	0.0	0.57	0.0	1.15	1.72	0.0	0.57	0.57	0.0	0.0	1.15	0.0	0.57	0.0
skcm	0.0	1.37	0.0	4.11	0.0	0.0	0.0	2.74	1.37	0.0	0.0	5.48	0.0	0.0	0.0	0.0	6.85	0.0	0.0	0.0	1.37	0.0
stad	0.49	0.0	2.95	0.74	0.49	6.14	0.74	0.74	2.21	1.23	0.0	0.98	0.0	0.0	0.49	0.98	0.0	0.0	3.44	0.49	0.74	0.25
tgct	0.0	0.0	9.09	7.58	0.0	0.0	0.0	0.0	0.0	0.0	1.52	30.3	0.0	0.0	1.52	1.52	0.0	0.0	1.52	1.52	1.52	3.03
thca	0.0	0.0	0.78	9.84	0.52	0.0	0.26	0.52	0.26	0.52	0.0	0.26	1.3	3.37	0.0	1.3	64.51	0.26	0.0	0.0	0.26	0.52
thym	0.0	0.0	2.0	4.0	2.0	0.0	0.0	4.0	2.0	0.0	0.0	2.0	0.0	8.0	4.0	2.0	0.0	0.0	0.0	0.0	4.0	2.0
ucec	0.0	13.55	0.0	0.22	11.61	0.22	0.22	10.75	6.45	0.43	0.0	1.08	0.22	0.0	0.43	0.65	0.0	0.86	0.0	0.22	0.22	0.0
ucs	0.0	0.0	0.0	0.0	27.27	1.82	0.0	9.09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.82	0.0	0.0	0.0	1.82

Table 21: Modules recovered after ensemble clustering of 45 MODULAR partitions (consensus clustering), UPGMA implementation and OSLOM2 implementation. The cells of the table refer to the percentage of the patients of the cancer type of the row, contained in the module of the column.

Module	Genes
0	ATRX, IDH1, TP53
1	CTNNB1, PTEN
2	APC, KRAS, SMAD4
3	FLT3, NPM1, NRAS
4	PPP2R1A
5	SMAD4
6	STK11

(a)

Module	Genes
7	PIK3R1
8	CTNNB1
9	SETD2
10	EGFR
11	KIT
12	DNMT3A
13	HRAS
14	PBRM1, VHL

(b)

Module	Genes
15	AKT1, GATA3
16	BRAF
17	SPOP
18	CDH1, RUNX1
19	CIC, IDH2
20	BAP1, NF2
21	FOXA1, MAP3K1

(c)

Table 22: Genes of modules as recovered by the ensemble clustering approach of MODULAR (consensus clustering), UPGMA and OSLOM2 in Table 21.

INFOMAP-OSLOM2 Results

		Modules										
		0	1	2	3	4	5	6	7	8	9	10
Cancer Types	acc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.69
	blca	0.5	1.01	0.0	0.0	0.5	3.27	0.5	0.0	0.0	0.0	2.51
	brca	0.22	0.43	0.0	0.11	13.12	0.43	0.43	0.0	1.08	0.0	3.44
	cesc	0.0	0.78	0.0	0.0	0.0	0.39	0.78	0.0	0.0	0.0	0.39
	coadread	0.41	1.83	0.0	0.0	0.61	0.41	1.63	0.0	0.0	0.0	0.2
	esca	1.69	0.56	0.0	0.0	0.56	0.0	2.26	0.0	0.0	0.0	1.13
	gbm	6.16	1.76	0.0	6.45	0.29	0.0	1.47	0.0	0.0	0.0	0.88
	hnsc	0.41	0.61	0.2	0.0	1.43	5.73	0.61	0.0	0.0	0.0	0.82
	kirc	0.0	0.32	0.0	0.0	0.32	0.0	0.0	0.32	0.0	32.59	0.0
	kirp	0.0	1.87	0.0	0.0	0.47	0.0	0.0	0.93	0.0	1.4	0.47
	laml	0.0	0.0	20.44	0.0	0.0	0.0	4.42	0.0	0.0	0.0	0.0
	lgg	61.87	0.41	0.0	1.42	0.41	0.0	1.01	0.0	0.0	0.0	0.81
	lihc	0.0	0.0	0.0	0.0	0.3	0.0	2.38	0.0	0.0	0.0	0.3
	luad	0.19	3.37	0.0	0.0	0.75	0.56	0.56	0.0	0.0	0.0	0.19
	lusc	0.44	0.66	0.0	0.0	1.31	1.09	2.4	0.0	0.0	0.0	1.09
	meso	0.0	1.52	0.0	0.0	0.0	0.0	1.52	9.09	0.0	0.0	0.0
	ov	2.26	0.25	0.0	0.0	0.75	0.0	2.01	0.25	0.0	0.0	0.25
	paad	0.65	0.65	0.0	0.0	0.65	0.0	0.0	0.0	0.0	0.0	1.31
	pcpg	0.0	1.22	0.0	0.0	0.0	21.95	1.22	0.0	0.0	0.0	0.0
	prad	0.0	2.01	0.0	0.0	0.0	1.34	0.33	0.0	0.0	0.0	25.42
	sarc	6.9	0.0	0.0	0.0	0.57	0.0	1.15	0.0	0.0	0.0	0.0
	skcm	0.0	24.66	0.0	0.0	0.0	1.37	5.48	0.0	0.0	0.0	0.0
	stad	0.49	0.25	0.0	0.0	1.23	0.0	0.98	0.0	0.0	0.25	0.0
	tgct	0.0	0.0	0.0	0.0	1.52	0.0	30.3	0.0	0.0	0.0	3.03
	thca	0.0	73.83	0.0	0.0	0.0	4.15	0.26	0.0	0.0	0.0	0.26
	thym	0.0	0.0	0.0	0.0	0.0	20.0	2.0	0.0	0.0	0.0	0.0
	ucec	0.0	0.22	0.0	0.0	0.43	0.22	1.08	0.0	0.0	0.0	6.02
	ucs	0.0	0.0	0.0	0.0	1.82	0.0	0.0	0.0	0.0	0.0	7.27

Table 23: Modules recovered after ensemble clustering of 200 INFOMAP partitions (consensus clustering) and OSLOM2 implementation. The cells of the table refer to the percentage of the patients of the cancer type of the row, contained in the module of the column.

Module	Genes
0	ATRX, CIC, FUBP1, IDH1, TP53
1	BRAF
2	DNMT3A, FLT3, NPM1
3	EGFR, PTEN
4	GATA3
5	HRAS

(a)

Module	Genes
6	KIT
7	BAP1, NF2
8	CDH1, RUNX1
9	BAP1,PBRM1, SETD2, VHL
10	FOXA1, SPOP

(b)

Table 24: Genes of modules as recovered by the ensemble clustering approach of INFOMAP (consensus clustering) and OSLOM2 in Table 23.

Cancer - Gene association

acc	CTNNB1
blca	HRAS, FOXA1, SPOP
brca	CDH1, RUNX1, GATA3, FOXA1, AKT1, MAP3K1, SETD2, PIK3R1, SPOP
cesc	STK11
coadread	APC, KRAS, SMAD4
esca	N/A
gbm	EGFR, ATRX, IDH1, TP53, PIKER1, PTEN, CIC, FUBP1
hnsc	SMAD4, HRAS
kich	N/A
kirc	PBRM1, VHL, BAP1, SETD2, NF2
kirp	BAP1, NF2, SETD2
laml	FLT3, NPM1, DNMT3A, RUNX1, ,IDH2 CIC, NRAS, CDH1
lgg	ATRX, IDH1, TP53, CIC, IDH2, EGFR, PIK3R1, DNMT3A, FUBP1
lihc	CTNNB1, BAP1, NF2, SETD2
luad	STK11, EGFR, SETD2, DNMT3A, APC, KRAS, SMAD4, BRAF
lusc	KIT
meso	BAP1, NF2
ov	N/A
paad	APC, KRAS, SMAD4
pcpg	HRAS
prad	SPOP, FOXA1, MAP3K1, CTNNB1
sarc	ATRX, IDH1, TP53, CIC, FUBP1
skcm	BRAF
stad	CDH1, RUNX1, APC, KRAS, SMAD4
tgct	KIT
thca	BRAF, HRAS, NRAS, FLT3, NPM1
thym	HRAS
ucec	PPP2R1A, PIK3R1, CTNNB1, PTEN, FOXA1, SPOP
ucs	PPP2R1A

Table 25: Genes linked with more than 10 patients through clustering and colors correspond to pipelines recovering them. Blue is for MODULAR-UPGMA-OSLOM2. Green is for INFOMAP-OSLOM2. Red is for both.

The blue highlighted cells of the Tables 21 and 23, along with the tables 22 and 24, were used to create the Cancer-Gene association table, which is Table 25. Then, every gene of the Table was compared with respect to current literature, to verify if it is indeed a driver gene for the correspondive cancer type.

Furthermore, the patients corresponding to each of the highlighted blue highlighted cells of Tables 21 and 23, were manually used to create cBioPortal "virtual studies". Hyperlinks for each of the blue highlighted cells are available in Table 26.

Adrenocortical Carcinoma - acc

CTNNB1 is a know driver gene of adrenocortical carcinoma [99, 100]. CTNNB1 was found to be mutated in every patient of the module, while no other gene (all genes of the dataset included) was found mutated simultaneously in more than 3 (out of 13) patients. .

Bladder Urothelial Carcinoma - blca

Regardless of the high number of patients of blca (398) in the network, only INFOMAP-OSLOM2 pipeline was able to assign a relatively small amount of patients in statistically significant groups. The genes associated with blca patients were HRAS, FOXA1 and SPOP. All three can be linked to prostate cancer [101, 102].

The first module accounts for 13 patients and contains the HRAS gene. Every patient of the module (13 out of 13 patients) carries HRAS mutations.

The second module contains the genes FOXA1 and SPOP. None of the 10 patients of the module carry mutations in both genes simultaneously. FOXA1 has been marked as a candidate regulator gene [101]. 3 out of 10 patients of the FOXA1-SPOP module carried FOXA1 mutations.

SPOP is also found to be a frequently mutated gene of prostate cancers [102]. 3 out of 10 patients of the FOXA1-SPOP module carried SPOP mutations

Notably, no patient of the FOXA1-SPOP module carried mutations in both FOXA1 and SPOP simultaneously

Breast invasive carcinoma - brca

CDH1 and RUNX1 are two genes that form a module, along with brca patients, in through INFOMAP-OSLOM2 and MODULAR-UPGMA-OSLOM2 approach. Both genes are associated with the luminal subtype of breast cancer. [93, 103, 104]. This fact is further supported by the resulting groups of the clustering. Through INFOMAP-OSLOM2 pipeline, 9 out of 10 patients of the cohort are of luminal A subtype and 9 out of 10 patients were of Breast Invasive Lobular Carcinoma (ILC). Through MODULAR-UPGMA-OSLOM2 pipeline 98 out of 115 were of luminal A subtype and 8 out of 115 were of luminal B subtype. Most patients (80 out of 115) were of ILC.

GATA3 was also recovered by both pipelines. It is a gene, which is considered non-significant as far as the ILC is concerned, but considered significant with respect to Breast Invasive Ductal Carcinoma (IDC) [93] and IDC luminal subtypes. Through INFOMAP-OSLOM2 pipeline 94 out of 122 patients were of IDC, while 76 and 37 out of 122 patients were of luminal A and B subtype accordingly. Through MODULAR-UPGMA-OSLOM2 pipeline, 82 out of 109 patients were of IDC and 75 and 27 out of 109 patients were of luminal A and B subtype accordingly. It must be noted that GATA3 was along AKT1 gene through the MODULAR-UPGMA-OSLOM2 approach, but many of the

patients were the same in both approaches. That is because the 122 and 109 patients recovered through the pipelines were actually 138 different patients. AKT1 is also a gene linked to IDC and the luminal subtypes.

FOXA1 is also a gene recovered by both pipelines. It is a gene associated with ILC [93]. Most patients of the cohorts though are of IDC. Specifically, through the INFOMAP-OSLOM2 pipeline, 16 out of 32 patients were of IDC and 11 out of 32 were of ILC. Thus, these two specific types of breast cancer account for 27 out of the 32 patients of the module in total. 23 out of the 27 of them carry FOXA1 mutations. Also, 24 and 4 out of the 32 patients are of luminal A and B subtype, respectively. Similarly, through the MODULAR-UPGMA-OSLOM2 pipeline, 73 out of the 97 were of IDC and 12 out of the 97 were of ILC, account for a total of 85 out of 97 patients. 16 out of the 85 of them carried FOXA1 mutations. Also, 81 and 8 out of the 97 patients were of luminal A and B subtypes, respectively. MAP3K1 was also assigned in this module (of MODULAR-UPGMA-OSLOM2 with FOXA1 present). MAP3K1 is also a gene implicated in breast cancer [93, 105].

SPOP, recovered only by INFOMAP-OSLOM2 pipeline (along with FOXA1), can have either a tumor-suppresing or onogenic role over breast cancer initiation and progression [106].

PIK3R1 is another significantly mutated gene in breast cancer, and specifically, in IDC [93]. 17 out of the 22 patients of the module recovered by the MODULAR-UPGMA-OSLOM2 pipeline, were of IDC.

Finally, SETD2, recovered by MODULAR-UPGMA-OSLOM2 pipeline, may have a tumour suppressor role in breast cancer [107, 108].

Cervical squamous cell carcinoma and endocervical adenocarcinoma - cesc

12 patients were assgined by MODULAR-UPGMA-OSLOM2 pipeline with STK11 gene which is associated with poor outcomes of cervical cancer patients [109].

Colorectal adenocarcinoma - coadread

APC, SMAD4 and KRAS are known cancer genes to colon and rectum cancers [94]. They were all associated through MODULAR-UPGMA-OSLOM2 pipeline.

Worse outcome is linked with loss of SMAD4 [110].

Either gain or loss of function of APC can lead to colorectal cancer initiation [111].

Glioblastoma multiforme - gbm

Glioblastoma represents grade IV gliomas. Some genes, mutated in wild-type IDH lower-grade gliomas are common between the two cancer types (lgg and gbm) [96]. Those genes recovered here are PTEN, EGFR and TP53. All three are common between lgg and gbm.

EGFR is a driver gene of glioblastoma. EGFR mutations are indicative of poor survival outcome [112]. PTEN is also a driver of gbm [113] and so is TP53 [114]. MODULAR-UPGMA-OSLOM2 pipeline assigned EGFR with gbm patients. 46 out of 56 patients of the module are deceased. INFOMAP-OSLOM2 pipeline assigned EGFR along with PTEN, with gbm patients. 19 out of the 22 patients of the module are deceased.

Patients carrying IDH1 mutations are usually younger than those with EGFR ones and have better survival prognosis [115]. Both pipelines assigned IDH1 with gbm patients. Through the MODULAR-UPGMA-OSLOM2 pipeline, IDH1 was along with ATRX and TP53, while 13 out of 22 patients are not deceased. Through the INFOMAP-OSLOM2 pipeline, IDH1 was along with ATRX, TP53, CIC and FUBP1, while 13 out of 21 patients are not deceased. It must be noted that all of the 21 patients of the module recovered by INFOMAP-OSLOM2 pipeline, also belong in the MODULAR-UPGMA-OSLOM2 module.

ATRX mutations appear mostly in young adults and pediatric glioblastoma [116, 117, 118]. 1 patient was below 22 years old in any of the two approaches (MODULAR-UPGMA-OSLOM2, INFOMAP-OSLOM2).

PIKER1 is also a significantly mutated gene for the cancer type [119].

As far as CIC and FUBP1 (associated through INFOMAP-OSLOM2), only 1 out of 21 patients carried CIC mutations. No patient out of the 21 carried FUBP1 mutations. Thus, existance of both of the genes will be addressed to lgg patients (see lgg).

Head and Neck squamous cell carcinoma - hnsc

Head and neck squamous cell carcinoma is one of the cancers associated with the virus HPV(Human Pappiloma Virus). HRAS and SMAD4 are genes found to be mutated in patients that weren't infected with HPV [95].

This fact is further supported for HRAS by the INFOMAP-OSLOM2 pipeline regarding hnsc patients, as 26 out of 28 patients of the cohort weren't infected with HPV.

As far as SMAD4 is concerned, through MODULAR-UPGMA-OSLOM2 pipeline, 10 out of the 12 patients of the cohort weren't infected with HPV.

Kidney renal clear cell carcinoma - kirc

PBRM1, VHL, BAP1 and SETD2 are the four most commonly mutated genes found in kidney clear cell carcinoma patients [120]. All 4 genes were retrieved by both methodolgies. Note that BAP1, SETD2 and PBRM1 are all chromatin remodelling genes [120] while they are close with VHL, as all three of them belong to chromosome 3 [121, 122].

VHL is a gene commonly found in clear cell carcinoma patients [120, 123]. VHL has a critical role in the cell's normal response to hypoxia [124].

PBRM1 is linked with better survival prognostics, while BAP1 and SETD2 mutations are indicative of poor clinical outcome [121, 125].

NF2 is also a gene associated with kidney clear cell carcinoma [126, 127]. It was recovered through the MODULAR-UPGMA-OSLOM2 pipeline and was along with BAP1.

Kidney renal papillary cell carcinoma - kirp

BAP1, NF2 and SETD2 are all found to be significantly mutated in kidney papillary cell carcinoma patients [128, 129, 127]. They were all recovered by the MODULAR-UPGMA-OSLOM2 pipeline.

Acute Myeloid Leukemia - laml

FLT3, NPM1, DNMT3A, RUNX1 and IDH2 genes were associated with acute myeloid leykemia patients and all of them are considered significantly mutated. It's suggested that all of 5 of them may have impact analogous to fusion genes [130].

NRAS is also a gene found to be mutated in acute myeloid leukemia patients as well as other hematologic malignancies [131] and may indicate a poor outcome [132].

Through INFOMAP-OSLOM2 pipeline, the module containing the genes FLT3, NPM1 and DNMT3A, consisted of 37 laml patients. All of them are also assigned in the module of the MODULAR-UPGMA-OSLOM2 pipeline, that contained the genes FLT3, NPM1 and NRAS. 25 out of the 37 patients of INFOMAP-OSLOM2 were also contained in the module of the MODULAR-UPGMA-OSLOM2 pipeline, that contained the gene DNMT3A. Furthermore, between the two modules of MODULAR-UPGMA-OSLOM2 pipeline (FLT3-NPM1- NRAS and DNMT3A), 33 patients belonged in both modules.

CDH1 mutations are absent in the patients of the corrsepondive module (associated MODULAR-UPGMA-OSLOM2). Thus, existance of the gene will be addressed to brca (see brca) and stad (see stad) patients.

In the same manner, only 1 patient carried CIC mutations (associated through MODULAR-UPGMA-OSLOM2) and existance of the gene will be addressed to lgg patients (see lgg).

Brain Lower Grade Glioma - lgg

IDH1, IDH2, CIC, FUBP1, TP53 and ATRX are genes recovered by the clustering pipelines. IDH mutations and no codeletion of 1p/19q is linked with TP53 and ATRX mutations. IDH mutations and codeletion of 1p/19q is linked with CIC and FUBP1 mutations [96].

The INFOMAP-OSLOM2 pipeline linked lgg patients to one module, which contained the genes ATRX, CIC, FUBP1, IDH1, and TP53. The existance of both genes that are linked with and without codeletion of 1p/19q justifies the

high number of both types of patients. 218 out of 305 patients of the module didn't have codeletion of 1p/19q, while 86 out of 305 patients had codeletion of 1p/19q.

Similarly, through MODULAR-UPGMA-OSLOM2 pipeline, 224 out of the 228 patient of the ATRX- IDH1- TP53 module didn't have codeletion of 1p/9q. 220 out of the 228 patients also belonged in the INFOMAP-OSLOM2 module.

The module of MODULAR-UPGMA-OSLOM2, which contained the genes CIC and IDH2, consisted of 90 out of 97 patients that had 1p/19q codeletion. 67 out of 97 patients also belonged in the INFOMAP-OSLOM2 module.

PIK3R1 is also one of the genes linked to codeletion of 1p/19q [96] and 9 out of 18 patients of the cohort do have codeletion of 1p/19q.

EGFR gene is linked to wild type IDH lower grade glioma [96]. 27 out of 28 patients of the cohort were of wild type IDH subtype.

DNMT3A (associated through MODULAR-UPGMA-OSLOM2) could not be considered as a driver gene for lgg. However, all patients (10 patients) of the cohort carried DNMT3A mutations.

Liver hepatocellular carcinoma - lihc

CTNNB1 and BAP1 are found to be significantly mutated in liver hepatocellular carcinomas [133, 134]. Both genes were recovered by the MODULAR-UPGMA-OSLOM2 pipeline through two separate modules. The first module contained only the CTNNB1 gene, while the second contained BAP1 along with NF2.

Homogenous deletions of NF2 have been associated with hepatocellular carcinoma [135]. Nevertheless, only 4 out of 19 patients of the cohort harbored NF2 mutations.

As far as SETD2 gene is concerned, there were no evidence to propose that mutations over SETD2 could initiate or promote tumorigenesis in lihc. However, a study concluded that downregulation of SETD2 by a non-coding RNA, named HOTAIR, could lead to tumorigenesis [136]. 12 out of 12 patients of the MODULAR-UPGMA-OSLOM2 cohort carried at least one SETD2 mutation.

Lung adenocarcinoma - luad

MODULAR-UPGMA-OSLOM2 pipeline was able to link 7 genes that with more than 10 patients of lung adenocarcinoma per module.

STK11 is one of the genes considered significantly mutated in lung adenocarcinoma [97], having a tumor suppressor role [137, 138].

SETD2 is a significant mutated gene [97],also having a tumor suppressor role [139].

Expression of DNMT3A is linked with cancer progression but not initiation, also proposing a tumor suppressor role [140].

EGFR gene is found to be commonly mutated in women with lung adenocarcinoma [97] and here, 22 out of the 34 patients were female. EGFR mutations are also found more frequently in non-smokers [141].

KRAS is also a gene commonly mutated in this cancer type [97] while there are no notable contrasts among smokers and non-smokers [142].

APC are infrequent in lung carcinomas but do exist in some patients [143]. Only 4 out of 14 patients of the module carried at least one mutation over APC.

SMAD4 may confer to tumor progression followed by KRAS or APC mutations in a number of cancers [144]. SMAD4 is linked to luad patients through two separate modules. The first module contains the genes APC, KRAS and SMAD4, where 3 out of 14 patients carry SMAD4 mutations. 2 out of the 3 of the patients carrying SMAD4 mutations were also assinged in the second module that contains only the gene SMAD4. All the patients of the SMDA4 (along with no other gene) module harbored SMAD4 mutations.

BRAF is the only gene that was recovered by INFOMAP-OSLOM2 methodology and it is marked as a potential driver of luad [97].

Lung squamous cell carcinoma - lusc

KIT is a gene found to be play a significant role in squamous cell lung cancers thrugh somatic copy number alterations [145]. Other literature report overexpression of KIT to be present in both large cell neuroendocrine carcinoma and small cell lung cancers [146, 147, 148].

Mesothelioma - meso

Almost half of the mesothelioma patients of the network (32 out of 66) were assigned to the group with BAP1 and NF2 genes present through the MODULAR-UPGMA-OSLOM2 pipeline. Mutations over both genes are observed in mesothelioma patients [149, 150, 151, 152, 153].

NF2 signaling disruption seems to be necessary in cancer initiation of mesothelioma [153], while BAP1 homozygous deletion is linked mostly with epithelioid and biphasic subtypes [152]. 19 out of 32 patients were of epithelioid subtype and 8 out of 32 patients were of biphasic subtype.

Pancreatic adenocarcinoma - paad

SMAD4 and KRAS are known driver genes of panceratic adenocarcinoma [154, 155, 156]. SMAD4 is recovered in two separate modules. One module, where SMAD4 is along with APC and KRAS (40 patients), and one module, where SMAD4 is the only gene of the cohort (36 patients). Thirty-three patients existed in both modules.

APC mutations are absent from patients of the APC-KRAS-SMAD4 cohort. Thus, existance of the gene will be adressed to stad (see stad), luad (see luad) and coadread (see coadread) patients.

Pheochromocytoma and Paraganglioma - pcpg

HRAS has been previously linked to pcpg cases [157, 158, 159]. The exact same module was recovered by either of the pipelines. (INFOMAP-OSLOM2, MODULAR-UPGMA-OSLOM2). Note that, while every patient of the cohort carried HRAS mutations, no other gene (with respect to all genes of the raw TCGA data) was found to be mutated in more than two patients simultaneously.

Prostate adenocarcinoma - prad

FOXA1 and SPOP are found to be recurrently mutated in prostate cancer patients [160, 161].

INFOMAP-OSLOM2 pipeline assigned FOXA1 and SPOP in one module along with prostate adenocarcinoma patients. 5 out of 76 patients of the module carried mutations in both genes. MODULAR-UPGMA-OSLOM2 pipeline, on the other hand, recovered two modules, regarding FOXA1 and SPOP. The first module contains only SPOP and every patient (40 out of 40 patients) of the module carried SPOP mutations. The second module contained FOXA1 along with MAP3K1. Interestingly, 5 out of 26 of the patients of the FOXA1-MAP3K1 module carried SPOP mutations and 4 out 5 of them, also belonged in the FOXA1-SPOP module that INFOMAP-OSLOM2 retrieved and all 4 of them carried both SPOP and FOXA1 mutations.

MAP3K1 deletions can be found in several cancer patients [162]. In the FOXA1-MAP3K1 module, only 2 out of 26 patients carried MAP3K1 mutations.

CTNNB1 is one of the genes found to be less frequently mutated in prostate adenocarcinoma [160]. 10 out of 10 patients of the CTNNB1 module, carried CTNNB1 mutations.

Sarcoma - sarc

Here [163], only three genes were recovered as significantly mutated and two of them, ATRX and TP53 are linked by both methodologies and the same patients were recovered each time.

ATRX and TP53 can be found mutated across many tumor types of sarcoma [164, 165]. Five different tumor types co-exist in the module (5 of Leiomyosarcoma (LMS), 3 of Pleomorphic MFH (Undifferentiated Pleomorphic Sarcoma), 2 of Myxofibrosarcoma, 1 of Dedifferentiated Liposarcoma and 1 of Pleomorphic Sarcoma (UPS) (Undifferentiated)).

Only 1 out of 12 harbored IDH1 mutations. Thus, existance of the gene in the module will be addressed to lgg (see lgg) and gbm (see gbm) patients.

Furthermore, no mutations over CIC and FUBP1 (associated through INFOMAP-OSLOM2) genes appear in the patients of the modules. Thus, existance of the genes will be addressed to lgg patients (see lgg).

Skin Cutaneous Melanoma - skcm

BRAF is a known driver gene of skin cutaneous melanoma [166, 167], while it is estimated that ≈ 50% of cutaneous melanoma cases carry BRAF mutations [168]. Moreover, targeted therapy addressing to BRAF gene (BRAF-inhibitors) has shown advantages with respect to the overall surival of the patients [168, 169].

INFOMAP-OSLOM2 pipeline recovered a module with BRAF gene and 18 skcm patients where 18 out of 18 carried BRAF mutations.

Stomach adenocarcinoma - stad

SMAD4 [170, 171, 172], APC, KRAS [170, 173, 174, 175] and CDH1 [170] are genes associated with stomach adenocarcinoma and they were recovered by the MODULAR-UPGMA-OSLOM2 pipeline.

SMAD4 appears in two separate modules, one in a module along with APC and KRAS and 12 stad patients, and one in a module with no other gene and 25 stad patients. 9 patients belonged in both modules. Furthermore, these 9 patients were of the chromosomal instability subtype. The rest of the 3 patients of the APC - KRAS - SMAD4 module belonged in different subtypes. 19 out of 25 patients of the SMAD4 cohort were of the chromosomal instability subtype as well.

CDH1 mutations are associated with genomically stable stomach adenocarcinoma subtype [170] and 10 out of 14 patients are of that subtype.

No patient harbored mutations over RUNX1 (associated through MODULAR-UPGMA-OSLOM2). Thus, existance of the gene will be addressed to brca (see brca) and laml (see laml) patients

Testicular Germ Cell Tumors - tgct

KIT is one of the three genes that were considered significantly mutated in testicular germ cell tumors here [176]. Both community detection pipelines accurately associated KIT with tgct patients. In fact, the modules recovered were exactly the same. All 20 out of 20 patients carried KIT mutations. No other gene (with respect to all genes of the raw TCGA data) was found to be mutated in more than two patients, simultaneously. KIT mutations are mostly present in seminomas [176, 177, 178]. 19 out of 20 patients were of the seminoma subtype.

Thyroid carcinoma - thca

BRAF, NRAS and HRAS are considered as driver genes for thyroid papillary carcinoma [179, 180].

RAS mutations are associated with the follicular thyroid cancers [181]. 25 out of the 38 patients of the module containing NRAS are of this tumor type. This module was recovered by the MODULAR-UPGMA-OSLOM2 pipeline. HRAS was associated through both pipelines and HRAS was the only gene of the

module in any of the approaches. 7 out of the 13 patients of the HRAS module recovered by the MODULAR-UPGMA-OSLOM2 approach and 10 out of the 16 patients of the HRAS module recovered by the INFOMAP-OSLOM2 approach, were of follicular thyroid cancer tumor type as well. We must note that all the 13 patients of the MODULAR-UPGMA-OSLOM2 approach were contained in the corresponding INFOMAP-OSLOM2 module.

On the other most of the patients associated with BRAF were of casual thyroid cancer tumor type. BRAF was associated through both pipelines and BRAF was the only gene of the module in any of the approaches. 205 out of the 249 patients of the BRAF module recovered by the MODULAR-UPGMA-OSLOM2 approach and 233 out of the 285 patients of the BRAF module recovered by the INFOMAP-OSLOM2 approach, were of this tumor type. We must note that all the 249 patients of the MODULAR-UPGMA-OSLOM2 approach were contained in the corresponding INFOMAP-OSLOM2 module.

No patient harbored mutations over FTL3 or NPM1 (associated through MODULAR-UPGMA-OSLOM2). Thus, existance of the genes will be addressed to laml patients (see laml)

Thymoma - thym

HRAS mutations are associated with thymoma [182, 183].

Uterine Corpus Endometrial Carcinoma - ucec

PPP2R1A, PIK3R1, CTNNB1, PTEN and SPOP are among the significantly mutated genes of uterine corpus endometrial carcinoma [98, 184]

PPP2R1A mutations are linked to serous endometrial carcinomas [98, 184]. 38 out of 54 patients of the module containing PPP2R1A were of serous endometrial adenocarcinoma tumor type. This module was recovered by the MODULAR-UPGMA-OSLOM2 pipeline.

PTEN and CTTNB1 mutations are linked to endometrioid endometrial tumor type [98, 184]. All of the 63 patients of the module containing PTEN and CTNNB1 were of endometrioid endometrial adenocarcinoma tumor type. This module was recovered by the MODULAR-UPGMA-OSLOM2 pipeline.

PIK3R1 mutations are also linked to endometrioid endometrial tumor type [98, 184]. 40 out of 40 patients of the module containing PIK3R1 were of endometrioid endometrial adenocarcinoma tumor type. This module was recovered by the MODULAR-UPGMA-OSLOM2 pipeline.

The role of SPOP in uterine sarcomas remains controversial [185], but it is a gene commonly mutated in serous endometrial cancers [186]. 15 out of 28 patients of the module containing SPOP were of endometrioid endometrial adenocarcinoma tumor type and 13 out of 28 patients were of serous endometrial adenocarcinoma tumor type. This module was recovered by the INFOMAP-OSLOM2 pipeline.

Only 1 out of the 28 patients of the module harbored FOXA1 mutations (associated through INFOMAP-OSLOM2). Thus, existance of the gene will be addressed to blca (see blca), brca (see brca) and prad (see prad) patients

Uterine Carcinosarcoma - ucs

PPP2R1A is a gene found to be recurrently mutated in uterine carcinoma patients [187, 188]. All of the patients of the cohort (15 patients) carried PPP2R1A mutations.

It must be noted that although 13 out of the 15 patients carried TP53 mutations, MODULAR-UPGMA-OSLOM2 pipeline assigned no patients in module 0 where TP53 belongs. This propably occurred due to lack of mutations of the ucs patients over the ATRX, IDH1 genes (that coexist with TP53 in module 0). TP53 is also found to be recurrently mutated in uterine carcinoma patients.

Table 26: Hyperlinks to cBioPortal Virtual Studies, each containing patients of the corresponding cancer types and modules.

Cancer Type	MODULAR-UPGMA-OSLOM2	INFOMAP-OSLOM2
acc	CTNNB1	
blca		HRAS,FOXA1,SPOP
brca	PIK3R1,SETD2,AKT1,GATA3,CHD1,RUNX1,FOXA1,MAP3K1	GATA3,CDH1,RUNX1,FOXA1,SPOP
cesc	STK11	
coadread	APC,KRAS,SMAD4,SMAD4	
gbm	ATRX,IDH1,TP53,PIK3R1,EGFR	ATRX,CIC,FUBP1,IDH1,TP53, EGFR,PTEN
hnsc	SMAD4	HRAS
kirc	SETD2,PBRM1,VHL,BAP1,NF2	BAP1,PBRM1,SETD2,VHL
kirp	SETD2,BAP1,NF2	
laml	FLT3,NPM1,NRAS,DNMT3A,CDH1,RUNX1,CIC,IDH2	DNMT3A,FLT3,NPM1
lgg	ATRX,IDH1,TP53,PIK3R1,EGFR,DNMT3A,CIC,IDH2	ATRX,CIC,FUBP1,IDH1,TP53
lihc	CTNNB1,SETD2,BAP1,NF2	
luad	APC,KRAS,SMAD4,SMAD4,STK11,SETD2,EGFR,DNMT3A	BRAF
lusc	KIT	KIT
meso	BAP1,NF2	
paad	APC,KRAS,SMAD4,SMAD4	
pcpg	HRAS	HRAS
prad	CTNNB1,SPOP,FOXA1,MAP3K1	FOXA1,SPOP
sarc	ATRX,IDH1,TP53	ATRX,CIC,FUBP1,IDH1,TP53
skcm		BRAF
stad	APC,KRAS,SMAD4,SMAD4,CDH1,RUNX1	
tgct	KIT	KIT
thca	FLT3,NPM1,NRAS,HRAS,BRAF	BRAF,HRAS
thym		HRAS
ucec	CTNNB1,PTEN,PPP2R1A,PIK3R1,CTNNB1	FOXA1,SPOP
ucs	PPP2R1A	

6 Discussion / Conclusion

Comparison to previous work J.Iranzo, Inigo Martincorena and Eugene V.Koonin were the ones that created a gene-patient network and implemented MODULAR-UPGMA-OSLOM and INFOMAP-OSLOM pipelines as described in Methods [92]. The main reasons for which differences in results occur are the following:

- MODULAR was run 200 times - Here it was run 45 times
- Color and Rectal cancer were treated separately - Here they both account for colorectal adenocarcinoma (coadread).
- Data used were from the year 2016 - Here they date back to 2018.
- OSLOM and INFOMAP were slower - Here both softwares run very fast.
- INFOMAP's bipartite version was used - Here the network was treated as unipartite because the biparite one may create structural artifacts.

The highest importance is attached to the first bullet, due to high computational cost of MODULAR which leads to expected differences. To compensate for them, the union of the two approaches is used and an extended analysis exploiting cBioPortal charts took place for every cancer type when more than 10 patients of the cancer belonged in a module, as described in Methods.

10-patients threshold The lower bound for patient nodes to be considered modular was 10, which proved effective even for types in clusters where the absolute number was exactly the boundary (blca, brca, lgg, prad, thym). We recall the need for recovering significantly mutated genes for primary sites over cell lines but, nevertheless, the analysis revealed several groups that were not only related to the cancer type but was also related to certain tumor or molecular subtypes and are consistent with TCGA publications and other literature. Through that concept, almost all but three cancer types (esophageal carcinoma, kidney chromophobe, oovarian serous cystadenocarcinoma) failed to exceed the threshold in any of the modules. Number 10 was chosen as threshold given that there exist studies containing $5 - 15$ samples each.

Subnetwork Both methodologies were firstly applied in a smaller network (including 6 cancer types) where more runs of MODULAR could be utilized. Ensemble clustering of INFOMAP-OSLOM2 was also justified (INFOMAP vs INFOMAP-OSLOM2), as both the number of modules was decreased and the large pseudomodule of INFOMAP (see 5.1) was removed. 100 runs of MODULAR combined with UPGMA resulted in a better gene-cancer association. For example in the analysis made by Iranzo et. al [92], the module containing colon and rectal adenocarcinoma patients contained the genes APC, KRAS, TP53, FBXW7, SMAD4 and TCF7L2. All of them are assigned in the subnetwork (along with NRAS), while through the network, only APC, KRAS and

SMAD4 were recovered. Thus, the exploitation of more MODULAR partitions is paramount for robust results.

Per cancer accuracy In order to make use of all the available results, only the OSLOM2 statistical significant modules appear here ($p\_value < 0.05$), while a cancer-gene association through the modules was evaluated. To do that, each gene that belonged with more than patients of a specific type was looked up to literature to determine whether there is indeed a connection. The reason that for some genes this couldn't be confirmed, is patients of other cancer types assigned in the same module.

Union effectiveness Comparing the two approaches, MODULAR-UPGMA-OSLOM2 and INFOMAP-OSLOM2, we conclude through Table 25 that both provide meaningful biological evidence (see Table 27)

	Nodes Assigned	Modular Nodes
Genes	16.7%	16.7%
Patients	36.3%	27.8%

	Nodes Assigned	Modular Nodes
Genes	11.6%	11.6%
Patients	17.5%	14.4%

	Nodes Assigned	Modular Nodes
Genes	17.2%	17.2%
Patients	39.1%	30.4%

Table 27: Nodes assigned and nodes patients considered modular through the 10-patient threshold (patients of the blue highlighted cells of the tables).
Left is for MODULAR-UPGMA-OSLOM2.
Center is for INFOMAP-OSLOM2.
Right is for the union.

7 References

[1] *Understanding what cancer is: Ancient Times to present.* URL: https://www.cancer.org/cancer/cancer-basics/history-of-cancer/what-is-cancer.html.

[2] Daphne W Bell. "Our changing view of the genomic landscape of cancer". In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 220.2 (2010), pp. 231–243.

[3] Véronique Bouvard et al. "A review of human carcinogens–Part B: biological agents." In: *The Lancet. Oncology* 10.4 (2009), pp. 321–322.

[4] NFE2L2 HLA-B. "Integrated genomic and molecular characterization of cervical cancer". In: *Nature* 543 (2017), p. 16.

[5] Adam J Bass et al. "Comprehensive molecular characterization of gastric adenocarcinoma". In: *Nature* 513.7517 (2014), p. 202.

[6] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer". In: *cell* 100.1 (2000), pp. 57–70.

[7] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (2009), pp. 719–724.

[8] Irene M Ghobrial, Thomas E Witzig, and Alex A Adjei. "Targeting apoptosis pathways in cancer therapy". In: *CA: a cancer journal for clinicians* 55.3 (2005), pp. 178–194.

[9] Liqun Yang et al. "Targeting cancer stem cell pathways for cancer therapy". In: *Signal transduction and targeted therapy* 5.1 (2020), pp. 1–35.

[10] Thomas Helleday et al. "DNA repair pathways as targets for cancer therapy". In: *Nature Reviews Cancer* 8.3 (2008), pp. 193–204.

[11] International Cancer Genome Consortium et al. "International network of cancer genome projects". In: *Nature* 464.7291 (2010), p. 993.

[12] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary oncology* 19.1A (2015), A68.

[13] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[14] Michał W Luczak and Paweł P Jagodziński. "The role of DNA methylation in cancer development." In: *Folia histochemica et cytobiologica* 44.3 (2006), pp. 143–154.

[15] B Martínez-López, AM Perez, and JM Sánchez-Vizcaíno. "Social network analysis. Review of general concepts and use in preventive veterinary medicine". In: *Transboundary and emerging diseases* 56.4 (2009), pp. 109–120.

[16] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[17] Kristof T Schütt et al. "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions". In: *Nature communications* 10.1 (2019), pp. 1–10.

[18] Subhash C Basak, Gerald J Niemi, and Gilman D Veith. "Predicting properties of molecules using graph invariants". In: *Journal of Mathematical Chemistry* 7.1 (1991), pp. 243–272.

[19] Santo Fortunato and Darko Hric. "Community detection in networks: A user guide". In: *Physics Reports* 659 (2016). Community detection in networks: A user guide, pp. 1–44. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2016.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0370157316302964.

[20] *Graph Online*. URL: https://graphonline.ru/en/.

[21] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: (2009). URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[23] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[24] Flavia Maria Darcie Marquitti et al. *MODULAR: Software for the Autonomous Computation of Modularity in Large Network Sets*. 2013. arXiv: 1304.2917 [q-bio.QM].

[25] Andrea Lancichinetti et al. "Finding statistically significant communities in networks". In: *PloS one* 6.4 (2011), e18961.

[26] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. "The map equation". In: *The European Physical Journal Special Topics* 178.1 (2009), pp. 13–23.

[27] Georgios A Pavlopoulos et al. "Using graph theory to analyze biological networks". In: *BioData mining* 4.1 (2011), pp. 1–27.

[28] Sabine Landau et al. *Cluster analysis*. John Wiley & Sons, 2011.

[29] Sławomir T Wierzchoń and Mieczysław A Kłopotek. *Modern algorithms of cluster analysis*. Vol. 34. Springer, 2018.

[30] Terence A Brown. "The human genome". In: *Genomes. 2nd edition*. Wiley-Liss, 2002.

[31] Lawrence Hunter. "Molecular biology for computer scientists". In: *Artificial intelligence and molecular biology* 177 (1993), pp. 1–46.

[32] Bruce Alberts et al. "The structure and function of DNA". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[33] Masood A Shammas. "Telomeres, lifestyle, cancer, and aging". In: *Current opinion in clinical nutrition and metabolic care* 14.1 (2011), p. 28.

[34] *What is noncoding DNA?: Medlineplus Genetics*. Jan. 2021. URL: https://medlineplus.gov/genetics/understanding/basics/noncodingdna/#:~:text=Noncoding%5C%20DNA%5C%20does%5C%20not%5C%20provide,the%5C%20control%5C%20of%5C%20gene%5C%20activity..

[35] Mark Kowarsky et al. "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA". In: *Proceedings of the National Academy of Sciences* 114.36 (2017), pp. 9623–9628.

[36] Chhavi Choudhary et al. "Long non-coding RNAs in insects". In: *Animals* 11.4 (2021), p. 1118.

[37] *Mutation*. URL: https://www.genome.gov/genetics-glossary/Mutation#:~:text=Mutations%5C%20can%5C%20result%5C%20from%5C%20DNA,and%5C%20are%5C%20not%5C%20passed%5C%20on..

[38] By: BD Editors, By: and BD Editors. *Missense mutation*. Aug. 2018. URL: https://biologydictionary.net/missense-mutation/.

[39] Dr.Mark. *Cell Division, Genetics, and Molecular Biology Cell Division, Genetics, and Molecular Biology*. URL: https://www.pdfdrive.com/cell-division-genetics-and-molecular-biology-cell-division-genetics-and-molecular-biology-e22406140.html.

[40] Eldra Solomon, Linda Berg, and Diana W Martin. *Biology*. Cengage Learning, 2010.

[41] GM Cooper. "The Cell: A Molecular Approach 2nd edition Boston University". In: *Sunderland (MA): Sinauer Associates.[Google Scholar]* (2000).

[42] *Outline of Cell cell communication*. URL: https://commons.wikimedia.org/wiki/File:Outline_of_Cell_cell_communication.png.

[43] *Cell cycle simple*. URL: https://commons.wikimedia.org/wiki/File:Cell_cycle_simple.png.

[44] *NCI Dictionary of Cancer terms*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/hyperplasia.

[45] Jie Dai et al. "Exosomes: Key players in cancer and potential therapeutic strategy". In: *Signal Transduction and Targeted Therapy* 5.1 (2020), pp. 1–10.

[46] Ruth Sager. "Expression genetics in cancer: shifting the focus from DNA to RNA". In: *Proceedings of the National Academy of Sciences* 94.3 (1997), pp. 952–955.

[47] Jane E Visvader. "Cells of origin in cancer". In: *Nature* 469.7330 (2011), pp. 314–322.

[48] Marieke Lydia Kuijjer et al. "Cancer subtype identification using somatic mutation data". In: *British journal of cancer* 118.11 (2018), pp. 1492–1501.

[49] Dong-Yu Wang et al. "Molecular stratification within triple-negative breast cancer subtypes". In: *Scientific reports* 9.1 (2019), pp. 1–10.

[50] Fatemeh Dorri et al. "Somatic mutation detection and classification through probabilistic integration of clonal population information". In: *Communications biology* 2.1 (2019), pp. 1–10.

[51] Matan Hofree et al. "Network-based stratification of tumor mutations". In: *Nature methods* 10.11 (2013), pp. 1108–1115.

[52] Katherine A Hoadley et al. "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer". In: *Cell* 173.2 (2018), pp. 291–304.

[53] Wei Zhang, Jianzhu Ma, and Trey Ideker. "Classifying tumors by supervised network propagation". In: *Bioinformatics* 34.13 (2018), pp. i484–i493.

[54] Narjes Rohani and Changiz Eslahchi. "Classifying breast cancer molecular subtypes by using deep clustering approach". In: *Frontiers in genetics* 11 (2020), p. 1108.

[55] Suleyman Vural, Xiaosheng Wang, and Chittibabu Guda. "Classification of breast cancer patients using somatic mutation profiles and machine learning approaches". In: *BMC systems biology* 10.3 (2016), pp. 263–276.

[56] Felipe De Sousa E Melo et al. "Cancer heterogeneity—a multifaceted view". In: *EMBO reports* 14.8 (2013), pp. 686–695.

[57] Shai Rosenwald et al. "Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin". In: *Modern Pathology* 23.6 (2010), pp. 814–823.

[58] F Anthony Greco. "Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management". In: *Current treatment options in oncology* 14.4 (2013), pp. 634–642.

[59] Chetan Bettegowda et al. "Detection of circulating tumor DNA in early- and late-stage human malignancies". In: *Science translational medicine* 6.224 (2014), 224ra24–224ra24.

[60] Andrea Marion Marquard et al. "TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen". In: *BMC medical genomics* 8.1 (2015), pp. 1–13.

[61] Vu Viet Hoang Pham et al. "Computational methods for cancer driver discovery: A survey". In: *Theranostics* 11.11 (2021), p. 5553.

[62] Michael S Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457 (2013), pp. 214–218.

[63] Abel Gonzalez-Perez and Nuria Lopez-Bigas. "Functional impact bias reveals cancer drivers". In: *Nucleic acids research* 40.21 (2012), e169–e169.

[64] Jüri Reimand and Gary D Bader. "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers". In: *Molecular systems biology* 9.1 (2013), p. 637.

[65] Abel Gonzalez-Perez et al. "IntOGen-mutations identifies cancer drivers across tumor types". In: *Nature methods* 10.11 (2013), pp. 1081–1082.

[66] Arunachalam Vinayagam et al. "Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets". In: *Proceedings of the National Academy of Sciences* 113.18 (2016), pp. 4976–4981.

[67] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. "Random graphs with arbitrary degree distributions and their applications". In: *Physical review E* 64.2 (2001), p. 026118.

[68] Erdös Paul and Rényi Alfréd. "On random graphs I". In: *Publicationes Mathematicae (Debrecen)* 6 (1959), pp. 290–297.

[69] Dario Fasino, Arianna Tonetto, and Francesco Tudisco. "Generating large scale-free networks with the Chung–Lu random graph model". In: *Networks* 78.2 (2021), pp. 174–187.

[70] Leto Peel, Daniel B Larremore, and Aaron Clauset. "The ground truth about metadata and community detection in networks". In: *Science advances* 3.5 (2017), e1602548.

[71] Jérôme Kunegis. "Konect: the koblenz network collection". In: *Proceedings of the 22nd international conference on world wide web*. 2013, pp. 1343–1350.

[72] Mel MacMahon and Diego Garlaschelli. "Community detection for correlation matrices". In: *arXiv preprint arXiv:1311.1924* (2013).

[73] Vincent A Traag and Jeroen Bruggeman. "Community detection in networks with positive and negative links". In: *Physical Review E* 80.3 (2009), p. 036115.

[74] Ulrik Brandes et al. "On modularity clustering". In: *IEEE transactions on knowledge and data engineering* 20.2 (2007), pp. 172–188.

[75] Santo Fortunato and Marc Barthelemy. "Resolution limit in community detection". In: *Proceedings of the national academy of sciences* 104.1 (2007), pp. 36–41.

[76] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. "Modularity from fluctuations in random graphs and complex networks". In: *Physical Review E* 70.2 (2004), p. 025101.

[77] Pan Zhang and Cristopher Moore. "Scalable detection of statistically significant communities and hierarchies, using message passing for modularity". In: *Proceedings of the National Academy of Sciences* 111.51 (2014), pp. 18144–18149.

[78] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[79] Andrea Lancichinetti and Santo Fortunato. "Consensus clustering in complex networks". In: *Scientific reports* 2.1 (2012), pp. 1–7.

[80] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Physical review E* 70.6 (2004), p. 066111.

[81] Ken Wakita and Toshiyuki Tsurumi. "Finding community structure in mega-scale social networks". In: *Proceedings of the 16th international conference on World Wide Web - WWW '07* (2007). DOI: 10.1145/1242572.1242805. URL: http://dx.doi.org/10.1145/1242572.1242805.

[82] Roger Guimerà and Luís Amaral. "Functional Cartography of Complex Metabolic Networks". In: *Nature* 23 (Jan. 2005), pp. 22–231.

[83] Mark EJ Newman. "Finding community structure in networks using the eigenvectors of matrices". In: *Physical review E* 74.3 (2006), p. 036104.

[84] *OSLOM*. URL: http://www.oslom.org/software.htm.

[85] David A Huffman. "A method for the construction of minimum-redundancy codes". In: *Proceedings of the IRE* 40.9 (1952), pp. 1098–1101.

[86] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[87] Ethan Cerami et al. "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data". In: *Cancer discovery* 2.5 (2012), pp. 401–404.

[88] Jianjiong Gao et al. "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal". In: *Science signaling* 6.269 (2013), pl1–pl1.

[89] Iñigo Martincorena and Peter J Campbell. "Somatic mutation in cancer and normal cells". In: *Science* 349.6255 (2015), pp. 1483–1489.

[90] Stefano Monti et al. "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data". In: *Machine learning* 52.1 (2003), pp. 91–118.

[91] *TCGA study abbreviations*. URL: https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations.

[92] Jaime Iranzo, Iñigo Martincorena, and Eugene V Koonin. "Cancer-mutation network and the number and specificity of driver mutations". In: *Proceedings of the National Academy of Sciences* 115.26 (2018), E6010–E6019.

[93] Giovanni Ciriello et al. "Comprehensive molecular portraits of invasive lobular breast cancer". In: *Cell* 163.2 (2015), pp. 506–519.

[94] Cancer Genome Atlas Network et al. "Comprehensive molecular characterization of human colon and rectal cancer". In: *Nature* 487.7407 (2012), p. 330.

[95] Cancer Genome Atlas Network et al. "Comprehensive genomic characterization of head and neck squamous cell carcinomas". In: *Nature* 517.7536 (2015), p. 576.

[96] Cancer Genome Atlas Research Network. "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas". In: *New England Journal of Medicine* 372.26 (2015), pp. 2481–2498.

[97] Cancer Genome Atlas Research Network et al. "Comprehensive molecular profiling of lung adenocarcinoma". In: *Nature* 511.7511 (2014), p. 543.

[98] Douglas A Levine. "Integrated genomic characterization of endometrial carcinoma". In: *Nature* 497.7447 (2013), pp. 67–73.

[99] Siyuan Zheng et al. "Comprehensive pan-genomic characterization of adrenocortical carcinoma". In: *Cancer cell* 29.5 (2016), pp. 723–736.

[100] Rajani Maharjan et al. "Comprehensive analysis of CTNNB1 in adrenocortical carcinomas: Identification of novel mutations and correlation to survival". In: *Scientific reports* 8.1 (2018), pp. 1–10.

[101] A Gordon Robertson et al. "Comprehensive molecular characterization of muscle-invasive bladder cancer". In: *Cell* 171.3 (2017), pp. 540–556.

[102] Lin-Gao Ju et al. "SPOP suppresses prostate cancer through regulation of CYCLIN E1 stability". In: *Cell Death & Differentiation* 26.6 (2019), pp. 1156–1168.

[103] Jose Mercado-Matos, Asia N Matthew-Onabanjo, and Leslie M Shaw. "RUNX1 and breast cancer". In: *Oncotarget* 8.23 (2017), p. 36934.

[104] Maaike PA van Bragt et al. "RUNX1, a transcription factor mutated in breast cancer, controls the fate of ER-positive mammary luminal cells". In: *Elife* 3 (2014), e03881.

[105] Cheukfai Li et al. "Spectrum of MAP3K1 mutations in breast cancer is luminal subtype-predominant and related to prognosis". In: *Oncology Letters* 23.2 (2022), pp. 1–12.

[106] Yizuo Song et al. "The emerging role of SPOP protein in tumorigenesis and cancer therapy". In: *Molecular Cancer* 19.1 (2020), pp. 1–13.

[107] RF Newbold and K Mokbel. "Evidence for a tumour suppressor function of SETD2 in human breast cancer: a new hypothesis". In: *Anticancer research* 30.9 (2010), pp. 3309–3311.

[108] Wail Al Sarakbi et al. "The mRNA expression of SETD2 in human breast cancer: correlation with clinico-pathological parameters". In: *BMC cancer* 9.1 (2009), pp. 1–7.

[109] Sou Hirose et al. "Genomic alterations in STK11 can predict clinical outcomes in cervical cancer patients". In: *Gynecologic Oncology* 156.1 (2020), pp. 203–210.

[110] Bryan C Szeglin et al. "A SMAD4-modulated gene profile predicts disease-free survival in stage II and III colorectal cancer". In: *Cancer Reports* 5.1 (2022), e1423.

[111] Lu Zhang and Jerry W Shay. "Multiple roles of APC and its therapeutic implications in colorectal cancer". In: *JNCI: Journal of the National Cancer Institute* 109.8 (2017).

[112] Eskil Eskilsson et al. "EGFR heterogeneity and implications for therapeutic intervention in glioblastoma". In: *Neuro-oncology* 20.6 (2018), pp. 743–752.

[113] Steven I Wang et al. "Somatic mutations of PTEN in glioblastoma multiforme". In: *Cancer research* 57.19 (1997), pp. 4183–4186.

[114] Ying Zhang et al. "The p53 pathway in glioblastoma". In: *Cancers* 10.9 (2018), p. 297.

[115] Sumihito Nobusawa et al. "IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas". In: *Clinical Cancer Research* 15.19 (2009), pp. 6002–6007.

[116] Carl Koschmann, Pedro R Lowenstein, and Maria G Castro. "ATRX mutations and glioblastoma: impaired DNA damage repair, alternative lengthening of telomeres, and genetic instability". In: *Molecular & cellular oncology* 3.3 (2016), e1167158.

[117] Yuchen Jiao et al. "Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas". In: *Oncotarget* 3.7 (2012), p. 709.

[118] Jeremy Schwartzentruber et al. "Driver mutations in histone H3. 3 and chromatin remodelling genes in paediatric glioblastoma". In: *Nature* 482.7384 (2012), pp. 226–231.

[119] Cancer Genome Atlas Research Network Tissue source sites: Duke University Medical School McLendon Roger 1 Friedman Allan 2 Bigner Darrell 1 et al. "Comprehensive genomic characterization defines hu-

man glioblastoma genes and core pathways". In: *Nature* 455.7216 (2008), pp. 1061–1068.

[120] Cancer Genome Atlas Research Network et al. "Comprehensive molecular characterization of clear cell renal cell carcinoma". In: *Nature* 499.7456 (2013), p. 43.

[121] Francesco Piva et al. "BAP1, PBRM1 and SETD2 in clear-cell renal cell carcinoma: molecular diagnostics and possible targets for personalized therapies". In: *Expert review of molecular diagnostics* 15.9 (2015), pp. 1201–1210.

[122] Svenja Bihr et al. "Expression and mutation patterns of PBRM1, BAP1 and SETD2 mirror specific evolutionary subtypes in clear cell renal cell carcinoma". In: *Neoplasia* 21.2 (2019), pp. 247–256.

[123] Chuan Shen and William G Kaelin Jr. "The VHL/HIF axis in clear cell renal carcinoma". In: *Seminars in cancer biology*. Vol. 23. 1. Elsevier. 2013, pp. 18–25.

[124] Peter E Clark. "The role of VHL in clear-cell renal cell carcinoma and its relation to targeted therapy". In: *Kidney international* 76.9 (2009), pp. 939–945.

[125] Lucy Gossage et al. "Clinical and pathological impact of VHL, PBRM1, BAP1, SETD2, KDM6A, and JARID1c in clear cell renal cell carcinoma". In: *Genes, Chromosomes and Cancer* 53.1 (2014), pp. 38–51.

[126] Carole Sourbier et al. "Targeting loss of the Hippo signaling pathway in NF2-deficient papillary kidney cancers". In: *Oncotarget* 9.12 (2018), p. 10723.

[127] Pedram Argani et al. "Biphasic hyalinizing psammomatous renal cell carcinoma (BHP RCC): a distinctive neoplasm associated with somatic NF2 mutations". In: *The American journal of surgical pathology* 44.7 (2020), p. 901.

[128] Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of papillary renal-cell carcinoma". In: *New England Journal of Medicine* 374.2 (2016), pp. 135–145.

[129] Michal Kovac et al. "Recurrent chromosomal gains and heterogeneous driver mutations characterise papillary renal cancer evolution". In: *Nature communications* 6.1 (2015), pp. 1–11.

[130] Cancer Genome Atlas Research Network. "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia". In: *New England Journal of Medicine* 368.22 (2013), pp. 2059–2074.

[131] Atsushi Nonami et al. "Identification of novel therapeutic targets in acute leukemias with NRAS mutations using a pharmacologic approach". In: *Blood, The Journal of the American Society of Hematology* 125.20 (2015), pp. 3133–3143.

[132] Shujuan Wang et al. "Mutational spectrum and prognosis in NRAS-mutated acute myeloid leukemia". In: *Scientific Reports* 10.1 (2020), pp. 1–9.

[133] Adrian Ally et al. “Comprehensive and integrative genomic characterization of hepatocellular carcinoma”. In: *Cell* 169.7 (2017), pp. 1327–1341.

[134] Kornelius Schulze et al. “Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets”. In: *Nature genetics* 47.5 (2015), pp. 505–511.

[135] Pascal Pineau et al. “Homozygous deletion scanning in hepatobiliary tumor cell lines reveals alternative pathways for liver carcinogenesis”. In: *Hepatology* 37.4 (2003), pp. 852–861.

[136] Haiyan Li et al. “LncRNA HOTAIR promotes human liver cancer stem cell malignant growth through downregulation of SETD2”. In: *Oncotarget* 6.29 (2015), p. 27847.

[137] Montserrat Sanchez-Cespedes et al. “Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung”. In: *Cancer research* 62.13 (2002), pp. 3659–3662.

[138] Zhenqing Li et al. “Relevance of STK11 mutations regarding immune cell infiltration, drug sensitivity, and cellular processes in lung adenocarcinoma”. In: *Frontiers in Oncology* 10 (2020), p. 580027.

[139] Xin Yang et al. “Methyltransferase SETD2 inhibits tumor growth and metastasis via STAT1–IL-8 signaling-mediated epithelial–mesenchymal transition in lung adenocarcinoma”. In: *Cancer science* 113.4 (2022), p. 1195.

[140] Qing Gao et al. “Deletion of the de novo DNA methyltransferase Dnmt3a promotes lung tumor progression”. In: *Proceedings of the National Academy of Sciences* 108.44 (2011), pp. 18061–18066.

[141] William Pao et al. “EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib”. In: *Proceedings of the National Academy of Sciences* 101.36 (2004), pp. 13306–13311.

[142] Gregory J Riely et al. “Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma”. In: *Clinical cancer research* 14.18 (2008), pp. 5731–5734.

[143] Hiroko Ohgaki et al. “APC mutations are infrequent but present in human lung cancer”. In: *Cancer letters* 207.2 (2004), pp. 197–203.

[144] Ming Zhao, Lopa Mishra, and Chu-Xia Deng. “The role of TGF-β/SMAD4 signaling in cancer”. In: *International journal of biological sciences* 14.2 (2018), p. 111.

[145] Cancer Genome Atlas Research Network et al. “Comprehensive genomic characterization of squamous cell lung cancers”. In: *Nature* 489.7417 (2012), p. 519.

[146] Giulio Rossi et al. “Kit expression in small cell carcinomas of the lung: effects of chemotherapy”. In: *Modern pathology* 16.10 (2003), pp. 1041–1047.

[147] William D Travis. “Update on small cell carcinoma and its differentiation from squamous cell carcinoma and other non-small cell carcinomas”. In: *Modern Pathology* 25.1 (2012), S18–S30.

[148] Kazuhiro Araki et al. “Frequent overexpression of the c-kit protein in large cell neuroendocrine carcinoma of the lung”. In: *Lung cancer* 40.2 (2003), pp. 173–180.

[149] Julija Hmeljak et al. “Integrative Molecular Characterization of Malignant Pleural MesotheliomaIntegrative Molecular Characterization of Mesothelioma”. In: *Cancer discovery* 8.12 (2018), pp. 1548–1565.

[150] Guangwu Guo et al. “Whole-exome sequencing reveals frequent genetic alterations in BAP1, NF2, CDKN2A, and CUL1 in malignant pleural mesothelioma”. In: *Cancer research* 75.2 (2015), pp. 264–269.

[151] Matthew Bott et al. “The nuclear deubiquitinase BAP1 is commonly inactivated by somatic mutations and 3p21. 1 losses in malignant pleural mesothelioma”. In: *Nature genetics* 43.7 (2011), pp. 668–672.

[152] Marta Cigognetti et al. “BAP1 (BRCA1-associated protein 1) is a highly specific marker for differentiating mesothelioma from reactive mesothelial proliferations”. In: *Modern Pathology* 28.8 (2015), pp. 1043–1057.

[153] Claudio Thurneysen et al. “Functional inactivation of NF2/merlin in human mesothelioma”. In: *Lung cancer* 64.2 (2009), pp. 140–147.

[154] Benjamin J Raphael et al. “Integrated genomic characterization of pancreatic ductal adenocarcinoma”. In: *Cancer cell* 32.2 (2017), pp. 185–203.

[155] Shirin Hafezi, Maha Saber-Ayad, and Wael M Abdel-Rahman. “Highlights on the Role of KRAS Mutations in Reshaping the Microenvironment of Pancreatic Adenocarcinoma”. In: *International Journal of Molecular Sciences* 22.19 (2021), p. 10219.

[156] Metin Tascilar et al. “The SMAD4 protein and prognosis of pancreatic ductal adenocarcinoma”. In: *Clinical Cancer Research* 7.12 (2001), pp. 4115–4121.

[157] Lauren Fishbein et al. “Comprehensive molecular characterization of pheochromocytoma and paraganglioma”. In: *Cancer cell* 31.2 (2017), pp. 181–193.

[158] Adam Stenman et al. “HRAS mutation prevalence and associated expression patterns in pheochromocytoma”. In: *Genes, Chromosomes and Cancer* 55.5 (2016), pp. 452–459.

[159] Joakim Crona et al. “Somatic mutations in H-RAS in sporadic pheochromocytoma and paraganglioma identified by exome sequencing”. In: *The Journal of Clinical Endocrinology & Metabolism* 98.7 (2013), E1266–E1271.

[160] Adam Abeshouse et al. “The molecular taxonomy of primary prostate cancer”. In: *Cell* 163.4 (2015), pp. 1011–1025.

[161] Christopher E Barbieri et al. “Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer”. In: *Nature genetics* 44.6 (2012), pp. 685–689.

[162] Trang T Pham, Steven P Angus, and Gary L Johnson. “MAP3K1: genomic alterations in cancer and function in promoting cell survival or apoptosis”. In: *Genes & cancer* 4.11-12 (2013), pp. 419–426.

[163] Alexander J Lazar et al. “Comprehensive and integrated genomic characterization of adult soft tissue sarcomas”. In: *Cell* 171.4 (2017), pp. 950–965.

[164] Jau-Yu Liau et al. “Comprehensive screening of alternative lengthening of telomeres phenotype and loss of ATRX expression in sarcomas”. In: *Modern Pathology* 28.12 (2015), pp. 1545–1554.

[165] Elizabeth Thoenen, Amanda Curl, and Tomoo Iwakuma. “TP53 in bone and soft tissue sarcomas”. In: *Pharmacology & therapeutics* 202 (2019), pp. 149–164.

[166] Rehan Akbani et al. “Genomic classification of cutaneous melanoma”. In: *Cell* 161.7 (2015), pp. 1681–1696.

[167] Francisco Sanchez-Vega et al. “Oncogenic signaling pathways in the cancer genome atlas”. In: *Cell* 173.2 (2018), pp. 321–337.

[168] Enrica Teresa Tanda et al. “Current state of target treatment in BRAF mutated melanoma”. In: *Frontiers in Molecular Biosciences* 7 (2020), p. 154.

[169] Jacek Mackiewicz and Andrzej Mackiewicz. “BRAF and MEK inhibitors in the era of immunotherapy in melanoma patients”. In: *Contemporary Oncology* 22.1A (2018), p. 68.

[170] Cancer Genome Atlas Research Network et al. “Comprehensive molecular characterization of gastric adenocarcinoma”. In: *Nature* 513.7517 (2014), p. 202.

[171] Li-Hui Wang et al. “Inactivation of SMAD4 tumor suppressor gene during gastric carcinoma progression”. In: *Clinical cancer research* 13.1 (2007), pp. 102–110.

[172] Steven M Powell et al. “Inactivation of Smad4 in gastric carcinomas”. In: *Cancer research* 57.19 (1997), pp. 4221–4224.

[173] Akira Horii et al. “The APC gene, responsible for familial adenomatous polyposis, is mutated in human gastric cancer”. In: *Cancer research* 52.11 (1992), pp. 3231–3233.

[174] Lindsay C Hewitt et al. “KRAS status is related to histological phenotype in gastric cancer: results from a large multicentre study”. In: *Gastric Cancer* 22.6 (2019), pp. 1193–1203.

[175] Lindsay C Hewitt et al. “KRAS, BRAF and gastric cancer”. In: *Transl Gastrointest Cancer* 4.6 (2015), pp. 429–47.

[176] Hui Shen et al. “Integrated molecular characterization of testicular germ cell tumors”. In: *Cell reports* 23.11 (2018), pp. 3392–3406.

[177] Kevin Litchfield et al. “Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours”. In: *Nature communications* 6.1 (2015), pp. 1–8.

[178] Qingsheng Tian et al. “Activating c-kit gene mutations in human germ cell tumors”. In: *The American journal of pathology* 154.6 (1999), pp. 1643–1647.

[179] Cancer Genome Atlas Research Network et al. “Integrated genomic characterization of papillary thyroid carcinoma”. In: *Cell* 159.3 (2014), pp. 676–690.

[180] MBRAF Xing. "BRAF mutation in thyroid cancer". In: *Endocrine-related cancer* 12.2 (2005), pp. 245–262.

[181] Hans-Juergen Schulten et al. "Comprehensive survey of HRAS, KRAS, and NRAS mutations in proliferative thyroid lesions from an ethnically diverse population". In: *Anticancer research* 33.11 (2013), pp. 4779–4784.

[182] Milan Radovich et al. "The integrated genomic landscape of thymic epithelial tumors". In: *Cancer cell* 33.2 (2018), pp. 244–258.

[183] Song Xu et al. "Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial Tumors". In: *Frontiers in oncology* 11 (2021), p. 2594.

[184] Melissa K McConechy et al. "Use of mutation profiles to refine the classification of endometrial carcinomas". In: *The Journal of pathology* 228.1 (2012), pp. 20–30.

[185] P Zhang et al. "Endometrial cancer-associated mutants of SPOP are defective in regulating estrogen receptor-α protein turnover". In: *Cell death & disease* 6.3 (2015), e1687–e1687.

[186] Xian-Miao Li et al. "Novel insights into the SPOP E3 ubiquitin ligase: From the regulation of molecular mechanisms to tumorigenesis". In: *Biomedicine & Pharmacotherapy* 149 (2022), p. 112882.

[187] Andrew D Cherniack et al. "Integrated molecular characterization of uterine carcinosarcoma". In: *Cancer cell* 31.3 (2017), pp. 411–423.

[188] Dorien Haesen et al. "Recurrent PPP2R1A mutations in uterine cancer act through a dominant-negative mechanism to promote malignant cell growth". In: *Cancer research* 76.19 (2016), pp. 5719–5731.