Patters of Intra-Clustering Analysis Reveal Hidden Oncogenic Relations

Stamatis Choudalakis[1,2], George Kastis[1], Nikolaos Dikaios[1]

December, 2023

- **Affiliations:** [1]Mathematics Research Center, Academy of Athens, 11527 Athens, Greece
[2]Medical School of Athens, National and Kapodistrian University of Athens, 11527 Athens, Greece

- **Classification:** Biological Sciences - Medical Sciences
- **Keywords:** Cancer, Oncogenes, Graph Theory, Single Point Mutations

Abstract

The identification of genes that can initiate and progress carcinogenesis through mutations is a challenging issue due to the sparsity of mutations and the high mutational heterogeneity between tumors of the same cancer type. While current methods mostly rely on finding recurrently mutated genes, this work aims to recover infrequently mutated genes that can contribute to cancer, by exploiting clustering analysis. To that end, we constructed a Network Graph of 8303 patients and 198 genes from single-point-mutation data, retrieved from the The Cancer Genome Atlas (TCGA), to find patient-gene groups with the parallel use of two separate methodologies: (a) one based on Barber's modularity index, and (b) one based on network dynamics. A systematic analysis was applied over 23 statistically meaningful groups of 2037 patients spanning 22 cancer types, derived from both methodologies, using the Fisher's exact test followed by the Benjamini-Hochberg false discovery rate method. This procedure recovered 32 known significantly mutated genes and 4 putative driver genes that are not identified by standard driver-gene identification algorithms used in the TCGA consortium analysis. This study has also revealed 14 statistically significant patterns not reported in the current literature, resulting in a total of 18 hidden oncogenic relations. Notably, among the results, skin cutaneous melanoma is related to IL7R. However, current literature suggests that only IL7R overexpression is linked to skin cancers. To the best of our knowledge, this is the first reported pan-cancer intra-clustering analysis.

Significance

Driver genes can initiate and progress carcinogenesis through mutations. Here, we aim to recover infrequently mutated driver genes by constructing a Network Graph of 8303 patients and 198 genes from single-point mutation data to obtain patient-gene groups with the parallel use of two separate methodologies. Robust statistical analysis over the resulting modules recovered 18 hidden oncogenic relations: 4 putative driver genes and 14 statistically significant mutational patterns. To the best of our knowledge, this is the first reported pan-cancer intra-clustering analysis.

1 Introduction

Cases of Unknown Primary (CUP) refer to aggressive cancers that have metastized to organs, and the origin of the primary tumor is unknown [1, 2]. These cases represent approximately $3 - 5\%$ of all cancer types, and they generally have low survival rates [1, 2]. Knowledge of the primary cancer site in CUP cases is essential for the use of site-specific treatments and the development suitable therapeutic plans [1, 2, 3]. To determine the primary cancer site, physical exams and a number of lab tests and scans, such as the PET-CT scan, are conducted [4]. Despite the abundance of such procedures , there are cases where the site of origin cannot determined, even after autopsy [5].

Recent studies have focused on the use of molecular data to define the primary cancer site in cases of unknown primary [2, 3]. Systematic analysis of gene expression, DNA methylation, mRNA expression, miRNA expression, and protein expression data have revealed distinct subtypes across cancer types [6]. However, these

subtypes are not cancer-specific in a sense that samples from various cancer types might coexist in a subtype. Furthermore, this molecular data can be affected by factors other than cancer [7]. On the other hand, somatic mutations are proved to be a reliable and stable source of data [7]. For that reason, tumor prediction models mostly rely on genomic abberations such as point mutations, mutation rates, and copy number alterations [8, 9, 10, 11].

Nonetheless, only a small fraction of mutations may be able to confer to cancer [12]. Mutations that initiate and/or promote tumorigenesis are called driver mutations, while the genes that harbor such mutations are called driver genes. The rest of the mutations are called passenger mutations [12]. Usually, a driver gene will acquire one or more of the "hallmarks of cancer," including evasion of cell destruction (apoptosis), sustainability of growth promotion, and avoidance of growth-suppressing signals, as these criteria will most likely lead to uncontrollable division and eventually carcinogenesis [13].

Statistically, the term driver gene often coincides with the term "significantly mutated gene," which means that most drivers are the genes found to be mutated in most patients of the same cancer type [14, 15]. This is the reason driver gene identification methods, such as MuSiC [16] and MutSigCV [17], mostly rely on the number of mutations as a metric to distinguish driver from passenger mutations. However, even if a significantly mutated gene is most likely to be a driver gene, the reverse hypothesis is not always true [18].This remark increases the overall complexity of cancer.

Somatic mutations can be studied based on the cell-of-origin or based on the primary cancer site. Various mutations that affect the same target cell result in different tumor types, causing intertumoral heterogeneity between different cancer types. Mutational differences between cancers are not only limited to the disease site. Certain cancer types may be divided into subtypes, according to the genetic alterations of cells in the tissue that account as cell of origin [19]. Diverse subgroups are observed even within the subtype itself [20]. Intratumoral heterogeneity is also present, as only a fraction of the tumor's cells will be mutated. Other reasons causing intratumoral heterogeneity include the contamination by normal cells [21].

Consequently, data of somatic mutations in tumor samples is very sparse [22]. This high heterogeneity makes it difficult to recognize 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 [23]. In addition, the mutational profile of patients with tumors from the same cancer subtype rarely has the same alterations [24]. Studies combining somatic mutations along with other data have revealed subtypes that, in general, include samples of various cancers per cohort [6, 23]. As pointed out in [6], the cell-of-origin may not fully determine tumor classification, but even so, it influences it. Therefore, other studies have also focused on subtype identification with a priori knowledge of the primary site [7, 25].

The applications of the analysis of somatic alterations with respect to the disease site also extend to a rapid diagnosis and treatment through the identification of disease-specific mutations found in cell-free circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) [26] through blood or urine sequencing [10].

To address driver gene identification and other cancer-genomic-related research such as cancer prognosis, numerous studies focus on the development of network-graph-based methods [24, 27, 28, 29]. One reason to explain this scientific trend is that biological interactions can be interpreted as a network. The structural elements required to build a network are: (a) the set of elements-of-study (nodes) and (b) the relation(s) between these elements (edges). To that end, any pairwise interaction can be expressed as a network. Furthermore, a network can also be presented as a graph, allowing the use of graph-theoretic tools to provide insights into the network.

In this project, we seek to recover driver genes that promote tumorigenesis through infrequent or rare driver mutations under a network-graph embedding. Given that the number of patients of a specific cancer type that carry such mutations is expected to be very small, we focus on finding groups of patients with similar genetic traits or genetic patterns over a suitable network, regardless of the group size. Clustering provides the most appropriate workspace of this task, as it collects the most similar nodes based on their relation, where the size of the resulting group is not necessarily fixed.

To assess the potential of clustering in the identification of infrequently mutated genes, we constructed a Network Graph using data of high-impact somatic-point mutations over a set of 198 known driver genes and 8303 patients. Subsequently, in order to enhance accuracy, we employed two cutting-edge clustering analysis pipelines on this network, each varying in nature, and conducted a rigorous statistical analysis over the final partitions, to recover hidden oncogenic relations. We further exploited the pattern-driven mechanism of clustering to assess if any significant patterns of co-occurrence or of mutual exclusivity occured within the clusters. The clustering analysis pipelines and the statistical analysis are carefully designed in order to avoid known methodological mistakes in clustering biological data [30].

2 Results

2.1 Implementation of the clustering pipelines

An initial estimation of the modular structure of the network was assessed with respect to the two null models provided by MODULAR. The dissociation between the modularities in of the null models to the network was apparent (Figure 1C) and the p-value using Welch's t-test was close to zero in each case. Additionally, we observed that 75% of the patient nodes in the network had a degree of less or equal to 5. This indicates that, for most patient nodes, only a small amount of information contributes to the overall network magnitude.

Despite the network's sparsity, the clustering pipelines were implemented as described in subsection 4.5, and approximately a third of the nodes were assigned to meaningful clusters. Specifically, the MODULAR pipeline recovered a total of 1833 (22%) patients and 57 (29%) genes in statistically significant modules, while the Infomap pipeline recovered a total of 2141 (26%) patients and 48 (24%) genes in statistically significant modules. In summary, 2719 (33%) patients and 64 (33%) genes were assigned to statistically significant modules through the union of the two final partitions. This was the only part of the project where patients from every cancer type are considered in order to establish the modular structure of the network.

A threshold filtering for each pipeline, as described in section 4.6, was applied due to the presence of patients spanning multiple cancer types in the modules. This procedure filtered out the most prevalent cancer types in each module. Tables 1 and 2 summarize the results of the MODULAR and the Infomap pipelines, referring to 19 and 21 cancer types, respectively. The total number of distinct cancer types presented in the final partitions is 22. Detailed tables of the node assignments in clusters, containing every cancer type, are presented in the Supplementary File.

It should be noted that Module "10" of the Infomap pipeline did not contain a significant number of patients in any cancer type and was therefore excluded from further investigation.

The full names of the 22 cancer types are presented in the Results section along with their TCGA abbreviations in the parentheses. The 8 cancer types not present in the final results are: Adrenocortical carcinoma (acc), Cervical squamous cell carcinoma and endocervical adenocarcinoma (cesc), Esophageal carcinoma (esca), Kidney Chromophobe (kich), Liver hepatocellular carcinoma (lihc), Lung squamous cell carcinoma (lusc), Ovarian serous cystadenocarcinoma (ov) and Stomach adenocarcinoma (stad).

Table 1: Modules recovered after ensemble clustering of 200 MODULAR partitions (consensus clustering), UPGMA and OSLOM2 implementation.

Module	Cancer Types	Genes mutated in > 30% tumors from the module	Genes mutated in < 30% tumors from the module
0	hnsc: 6.57%, pcpg: 23.17%, skcm: 8.45%, thca: 59.07%, thym: 8.00%	BRAF	CASP8, HLA-A, HRAS, IL7R, NOTCH1, RAC1
1	kirp: 5.14%, kirc: 41.14%, meso: 13.64%	PBRM1, SETD2, VHL	BAP1, KDM5C, NF2
2	gbm: 8.21%	EGFR, PTEN	
3	laml: 40.33%	DNMT3A, FLT3, NPM1	CEBPA, IDH2, NRAS, PTPN11, TET2, WT1
4	ucec: 34.92%, ucs: 7.27%	ARID1A, CTNNB1, PIK3CA, PIK3R1, PTEN	ACVR1, CCND1, FGFR2, MED12, NSD1
5	coad: 43.46%, paad: 37.25%, read: 63.41%, ucs: 9.09%	APC, KRAS, TP53	AMER1, FBXW7, SMAD4, TCF7L2
6	gbm: 6.45%, lgg: 66.53%, sarc: 6.94%	ATRX, IDH1, TP53	CIC, FUBP1, NOTCH1
7	blca: 7.56%, brca: 19.57%	CDH1, PIK3CA	AKT1, CBFB, FGFR3, FOXA1, GATA3, KDM6A, MAP2K4, MAP3K1, RUNX1, SPOP, STAG2, TBX3

2.2 Module analysis

Out of a total of 30 cancer types included in the network, 22 cancer types were significantly represented (as described in section 4.6) in modules recovered by either of the two clustering analysis pipelines. These

Table 2: Modules recovered after ensemble clustering of 200 Infomap partitions (consensus clustering) and OSLOM2 implementation.

Module	Cancer Types	Genes mutated in > 30% tumors from the module	Genes mutated in < 30% tumors from the module
0	ucec: 26.03%	ARID1A, CTNNB1, PIK3R1, PTEN	ACVR1, BCOR, CCND1, CUX1, FGFR2, JAK1, MED12, NSD1, PTCH1, RNF43
1	coad: 36.27%, luad: 9.55%, paad: 24.84%, read: 43.09%	APC, KRAS, SMAD4	AMER1, STK11, TCF7L2
2	gbm: 5.57%, lgg: 60.45%	ATRX, IDH1, TP53	CIC, FUBP1
3	skcm: 29.58%, thca: 73.83%	BRAF	
4	blca: 4.03%	FGFR3, KDM6A, STAG2	
5	kirp: 7.01%, kirc: 13.61%, meso: 12.12%	SETD2	
6	brca: 13.12%	GATA3	
7	hnsc: 5.75%, pcpg: 21.95%, thca: 4.15%, thym: 20.00%	HRAS	
8	laml: 4.42%, skcm: 5.63%, tgct: 30.30%	KIT	
9	meso: 9.09%	BAP1, NF2	
10	N/A	CDH1, RUNX1	
11	laml: 6.63%	IDH2	CIC
12	kirc: 30.06%	PBRM1, VHL	BAP1, SETD2
13	laml: 25.97%	DNMT3A, FLT3, NPM1	PTPN11, TET2
14	coad: 4.25%, laml: 7.18%, read: 10.57%, skcm: 14.08%, tgct: 7.58%, thca: 10.10%, thym: 6.00%	NRAS	
15	prad: 25.42%, ucec: 6.07%, ucs: 7.27%	FOXA1, SPOP	

cancer types, along with the genes that co-exist in the same module, were extensively analyzed to evaluate the performance of the pipelines in recovering driver genes (see Figure 2). The results are presented below:

- **Bladder Urothelial Carcinoma (blca):** Fourteen genes were clustered with 30 bladder urothelial carcinoma patients (7.56% of bladder urothelial carcinoma patients) in two modules. STAG2, KDM6A and FGFR3 were recovered as statistically significant in at least one module. All three genes were also reported as statistically significant in the corresponding TCGA publication, along PIK3CA and FOXA1, which were also clustered with bladder urothelial patients: [31].

- **Breast invasive carcinoma (brca):** Fourteen genes were clustered with 255 breast invasive carcinoma patients (27.42% of breast invasive carcinoma patients) in two modules. MAP3K1, GATA3 and CDH1 were recovered as statistically significant in at least one module. All three genes were also reported as statistically significant in the corresponding TCGA publication, along with AKT1, CBFB, FOXA1, PIK3CA, MAP2K4, RUNX1 and TBX3, which were also clustered with breast invasive carcinoma patients [32, 33]. Four statistically significant patterns of mutual exclusion were also observed within at least one module: AKT1-PIK3CA [32], GATA3-PIK3CA [34], CDH1-MAP3K1 [35] and CDH1-GATA3 [36].

- **Colon Adenocarcinoma (coad):** Nine genes were clustered with 158 colon adenocarcinoma patients (51.63% of colon adenocarcinoma patients) in three modules. APC and AMER1 were recovered as statistically significant in at least one module. APC was also reported as statistically significant in the corresponding TCGA publication, along with KRAS, TP53, SMAD4, FBXW7, TCF7L2 and NRAS, which were also clustered with colon adenocarcinoma patients [37]. However, there is evidence in the literature to support the role of AMER1 as a driver gene in both primary and metastatic colorectal adenocarcinoma [38, 39]. It must be noted that here, AMER1 is not detected as significantly mutated for rectal adenocarcinoma, but only for colon adenocarcinoma.

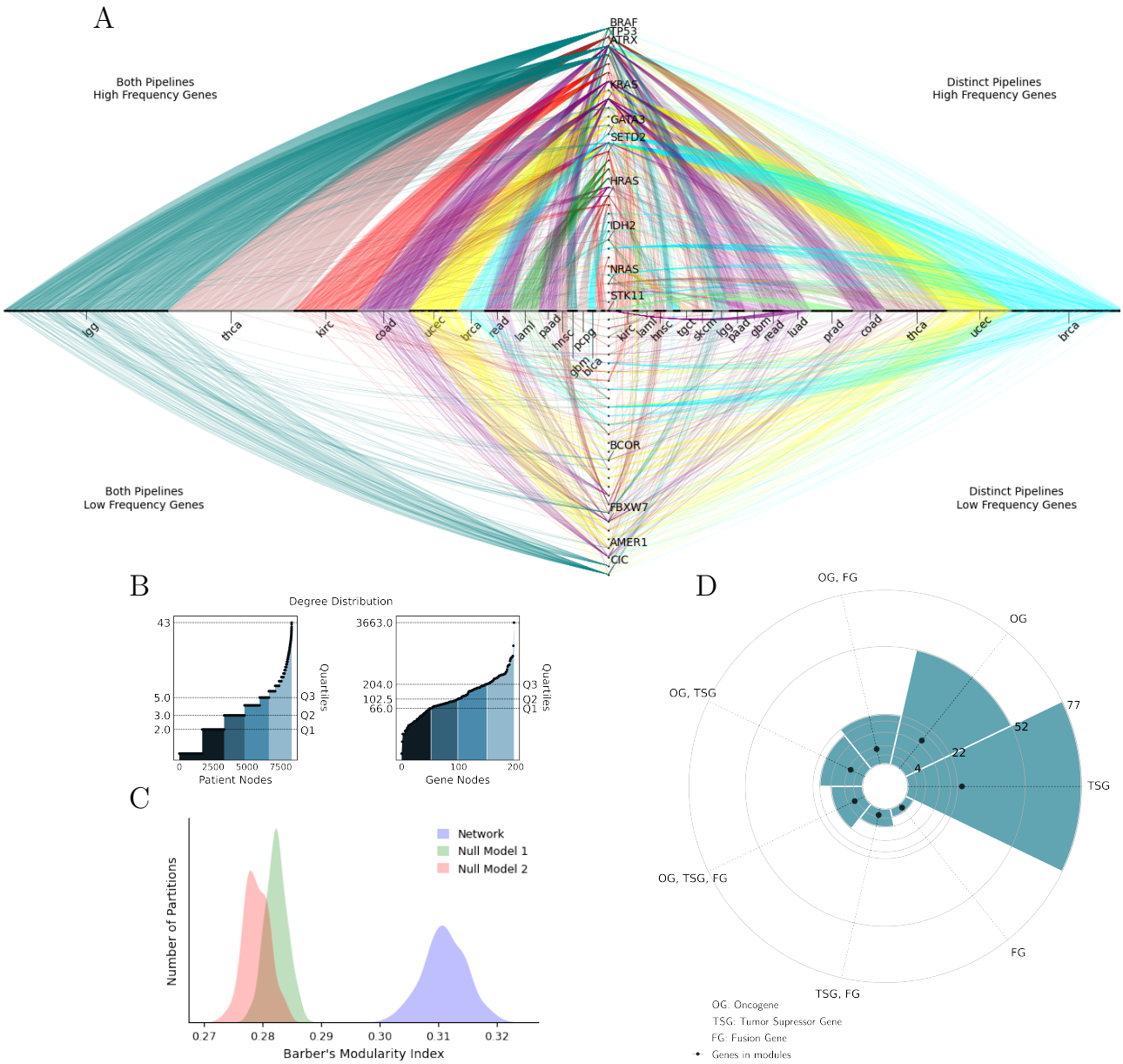

Figure 1: (A) A view of the network in unified clusters (See Supplementary File for details on the unified clusters). The x-axis contains only patients that are clustered from any pipeline. If a patient is clustered from both pipelines, the corresponding node is presented at the left side of the axis, on the other hand if a patient is clustered from a single pipeline, the node is presented at the right side of the axis. The y-axis contains the nodes that are assigned in modules. If a gene is mutated in a percentage higher than 30%, in any of the obtained modules, then the corresponding gene node is presented at the upper side of the axis. On the other hand if the gene node is mutated in a percentage lower than 30%, the gene node is presented in the lower side of the axis. (B) The degree distribution of the patient nodes (left) and of the gene nodes (right). The different colors of the shaded area are defined by the known quartiles. Nodes of the same shade have a degree lower than the degree that corresponds to the 25%, 50%, 75% and 100% of the degrees. For example, the darkest blue area corresponds to the patients with a degree of less or equal than 2 ($=Q_1$).(C) The modular structure of the network is compared to the two null models provided by the MODULAR software. The dissociation of the modularities of the null models and the network is apparent. (D) The cancer-gene classification of the 198 genes of the dataset presented as a polar barplot. The black dot corresponds to the number of genes of the corresponding classification assigned in clusters. The gene classification is described at the legend below the plot.

- **Glioblastoma (gbm):** Eight genes were clustered with 50 glioblastoma patients (14.66% of glioblastoma patients) in three modules. TP53 and ATRX were recovered as statistically significant in at least one module. TP53 was also reported as statistically significant in the corresponding TCGA publication, along with EGFR and PTEN, which were also clustered with glioblastoma patients [40]. Brennan *et al.* recovered ATRX as significantly mutated [41].

- **Head and Neck squamous cell carcinoma (hnsc):** Seven genes were clustered with 41 head and

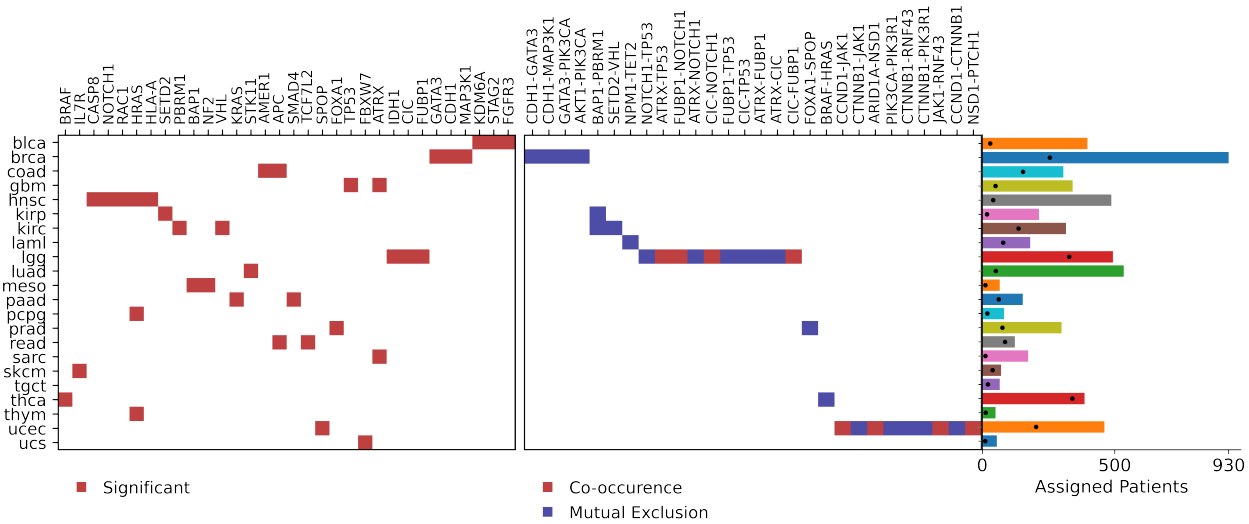

Figure 2: This figure summarizes the significant relations and the number of clustered patients per cancer type. A significant relation is referred to as significant if the adjusted p-value from the Benjamini-Hochberg false disovery rate is lower than 0.05. At the left heatmap, the significant cancer-gene relations are marked in red. At the center, the significant gene-to-gene relations are marked in a color. If the color is red, the significant pattern is a pattern of co-occurrence, and if the color is blue, the significant pattern is a pattern of mutual exclusion. At the right, the length of each bar of the bar plot is equivalent to the number of patients of the curated dataset of the cancer type corresponding to this bar. Furthermore, the black dot indicates the number of patients being clustered in the corresponding cancer type.

neck squamous cell carcinoma patients (8.42% of head and neck squamous cell carcinoma patients) in two modules. HRAS, NOTCH1, HLA-A, CASP8 and RAC1 were recovered as statistically significant in at least one module. Apart from RAC1, the remaining four genes were also reported as statistically significant in the corresponding TCGA publication [42]. RAC1 mutations are known to cause resistance to therapies [43]. Furthermore, Ngan *et al.* also recovered RAC1 somatic mutations related to head and neck squamous cell carcinoma patients [44].

- **Kidney renal clear cell carcinoma (kirc):** Six genes were clustered with 137 kidney renal clear cell carcinoma patients (43.35% of kidney renal clear cell carcinoma patients) in three modules. VHL and PBRM1 were recovered as statistically significant in at least one module. Both of them are also reported as statistically significant in the corresponding TCGA publication, along with SETD2, BAP1 and KDM5C, which were also clustered with kidney renal clear cell carcinoma patients [45]. Two statistically significant patterns of mutual exclusion were also observed within at least one module: SETD2-VHL and BAP1-PBRM1 [46].

- **Kidney renal papillary cell carcinoma (kirp):** Six genes were clustered with 18 kidney renal papillary cell carcinoma patients (8.41% of kidney renal papillary cell carcinoma patients) in two modules. SETD2 was recovered as statistically significant in at least one module. SETD2 was also reported as statistically significant in the corresponding TCGA publication, along with PBRM1, BAP1 and NF2, which were also clustered with kidney renal papillary cell carcinoma patients: [47]. One statistically significant pattern of mutual exclusion was also observed within at least one module: BAP1-PBRM1 [47].

- **Acute Myeloid Leukemia (laml):** Eleven genes were clustered with 90 acute myeloid leukemia patients (49.72% of acute myeloid leukemia patients) in five modules. None of these genes were been recovered as statistically significant in any of the modules. However, DNMT3A, FLT3, NPM1, CEBPA, IDH2, NRAS, PTPN11, TET2, WT1 and KIT were reported as driver genes in the corresponding TCGA publication [48]. One statistically significant pattern of mutual exclusion was also observed within at least one module: NPM1-TET2.

- **Lower Grade Glioma (lgg):** Six genes were clustered with 328 lower grade glioma patients (66.53% of lower grade glioma patients) in two modules. FUBP1, CIC and IDH1 were recovered as statistically significant in at least one module. All three genes were also reported as statistically significant in the corresponding TCGA publication, along with TP53, ATRX and NOTCH1, which were also clustered with lower grade glioma patients [49]. Six statistically significant patterns of mutual exclusion were also

observed within at least one module: ATRX-FUBP1 [50], FUBP1-TP53 [50], ATRX-CIC [50], CIC-TP53 [50], ATRX-NOTCH1, NOTCH1-TP53. Four statistically significant patterns of co-occurrence were also observed within at least one module: ATRX-TP53 [50], CIC-FUBP1 [51], FUBP1-NOTCH1 and CIC-NOTCH1.

- **Lung Adenocarcinoma (luad):** Six genes were clustered with 51 lung adenocarcinoma patients (9.55% of lung adenocarcinoma patients) in one module. STK11 was recovered as statistically significant in the module. STK11 was also reported as statistically significant in the corresponding TCGA publication, along with KRAS that was also clustered with kidney lung adenocarcinoma patients [52].
- **Mesothelioma (meso):** Six genes were clustered with 12 mesothelioma patients (18.18% of mesothelioma patients) in three modules. NF2 and BAP1 were recovered as statistically significant in at least one module. Both genes were also reported as statistically significant in the corresponding TCGA publication, along with SETD2 that was also clustered with mesothelioma patients [53].
- **Pancreatic Adenocarcinoma (paad):** Eight genes were clustered with 62 pancreatic adenocarcinoma patients (40.52% of pancreatic adenocarcinoma patients) in two modules. KRAS and SMAD4 were recovered as statistically significant in at least one module. Both genes were also reported as statistically significant in the corresponding TCGA publication, along with TP53 that was also clustered with pancreatic adenocarcinoma patients [54].
- **Pheochromocytoma and Paraganglioma (pcpg):** Seven genes were clustered with 19 pheonochromocytoma and paraganglioma patients (23.17% of pheonochromocytoma and paraganglioma patients) in two, almost identical, modules. HRAS was recovered as statistically significant in at least one module. HRAS was also reported as statistically significant in the corresponding TCGA publication [55].
- **Prostate Adenocarcinoma (prad):** Two genes were clustered with 76 prostate adenocarcinoma patients (25.42% of prostate adenocarcinoma patients) in one module. FOXA1 was recovered as statistically significant in the module. FOXA1 was also reported as statistically significant in the corresponding TCGA publication, along with SPOP that was also clustered with prostate adenocarcinoma patients [56]. One statistically significant patterns of mutual exclusion was also observed within at least one module: FOXA1-SPOP [56].
- **Rectal Adenocarcinoma (read):** Nine genes were clustered with 90 rectal adenocarcinoma patients (73.17% of rectal adenocarcinoma patients) in three modules. APC and TCF7L2 were recovered as statistically significant in at least one module. Both genes were also reported as statistically significant in the corresponding TCGA publication, along with KRAS, TP53, SMAD4, FBXW7 and NRAS, which were also clustered with rectal adenocarcinoma patients: [37].
- **Sarcoma (sarc):** Six genes were clustered with 12 sarcoma patients (6.94% of sarcoma patients) in one module. ATRX was recovered as statistically significant in the module. ATRX was also reported as statistically significant in the corresponding TCGA publication along with TP53 that was also clustered with sarcoma patients [57].
- **Skin Cutaneous Melanoma (skcm):** Nine genes were clustered with 39 skin cutaneous melanoma patients (54.93% of skin cutaneous melanoma patients) in four modules. IL7R was recovered as statistically significant in at least one module. IL7R was not reported as statistically significant in the corresponding TCGA publication, unlike BRAF, RAC1 and NRAS which were clustered with skin cutaneous melanoma patients [58]. IL7R overexpression is associated in the literature with metastasis and poor prognosis in melanomas [59, 60].
- **Testicular Germ Cell Tumors (tgct):** Two genes were clustered with 23 testicular germ cell tumor patients (34.85% of testicular germ cell tumor patients) in two modules. None of these genes were recovered as statistically significant in any of the modules. However, NRAS and KIT were reported as driver genes in the corresponding TCGA publication [61].
- **Thyroid cancer (thca):** Eight genes were clustered with 340 thyroid cancer patients (88.08% of thyroid cancer patients) in four modules. BRAF was recovered as statistically significant in at least one module. BRAF was also reported as statistically significant in the corresponding TCGA publication, along HRAS and NRAS, which were also clustered with thyroid cancer patients: [62]. One statistically significant patterns of mutual exclusion was also observed within at least one module: BRAF-HRAS [62].
- **Thymoma(thym):** Eight genes were clustered with 13 thymoma patients (26% of thymoma patients) in three modules. HRAS was recovered as statistically significant in at least one module. HRAS was also reported as statistically significant in the corresponding TCGA publication, along with NRAS that was also clustered with thymoma patients [63].

- **Uterine Corpus Endometrial Carcinoma (ucec):** Seventeen genes were clustered with 211 uterine corpus endometrial carcinoma patients (45.77% of uterine corpus endometrial carcinoma patients) in three modules. SPOP was recovered as statistically significant in at least one module. SPOP was also reported as statistically significant in the corresponding TCGA publication, along with ARID1A, CTNNB1, PIK3CA, PIK3R1, PTEN, CCND1, FGFR2, BCOR and RNF43, which were also clustered with uterine corpus endometrial carcinoma patients [64]. Five statistically significant patterns of mutual exclusion were also observed within at least one module: PIK3CA-PIK3R1 [64], CCND1-CTNNB1, CTNNB1-PIK3R1, CTNNB1-RNF43 and CTNNB1-JAK1. Four statistically significant patterns of co-occurrence were also observed within at least one module: ARID1A-NSD1, NSD1-PTCH1, JAK1-RNF43, CCND1-JAK1.

- **Uterine Carcinosarcoma (ucs):** Twelve genes were clustered with 12 uterine corpus endometrial carcinoma patients (21.82% of uterine carcinosarcoma patients) in three modules. FBXW7 was recovered as statistically significant in at least one module. FBXW7 was also reported as statistically significant in the corresponding TCGA publication along with ARID1A, PIK3CA, PIK3R1, PTEN, KRAS, TP53 and SPOP, which were also clustered with uterine carcinosarcoma patients [65].

3 Discussion

In this research, our primary emphasis was on pinpointing hidden oncogenic relations across 30 different cancer types, using data sourced from TCGA. Our methodology involved employing rigorous statistical analysis on gene-patient clusters derived from two clustering pipelines. What distinguishes our approach from current methods is the integration of statistics within smaller cohorts, a departure from the traditional analysis of complete curated datasets. This shift allows clustering analysis to reveal low-frequency driver genes by grouping patients with comparable genetic traits, regardless of cohort size.

The key finding of our study is the identification of 4 mutated genes that were overlook in the TCGA consortium analysis, thereby suggesting their putative roles as driver genes. Specifically, AMER1 in colon adenocarcinoma, ATRX in glioblastoma, and RAC1 in head and neck squamous cell carcinoma emerged as statistically significant, corroborating findings from other studies incorporating somatic mutation data [38, 44, 66]. In contrast, our work introduces IL7R as a potential driver gene for skin cutaneous melanoma, despite the absence of prior literature support. Notably, while IL7R overexpression is observed in skin cancers, no work to date designates it as a driver gene [59, 60].

The second key finding was that known significantly mutated genes may not always be recovered, even when co-existing with patients of the corresponding cancer type. This discrepancy is observed across all instances except for pheochromocytoma and paraganglioma, and head and neck squamous cell carcinoma, where every clustered gene had been previously confirmed as significantly mutated. In other cases, our approach exhibits a maximum overlap of three genes, suggesting its limited suitability for detecting frequently mutated genes.

The third key finding was our method's ability in detecting 15 known significant patterns of co-occurrence and mutual exclusivity, aligning with TCGA consortium analysis and other studies. Impressively, we identify 14 statistically significant patterns previously unreported in the current literature. This suggests the ability of clustering to group patients with similar or contrasting genetic alterations, extended to discrepancies among patients of the same cancer type within different clusters. For example, the differentiation of EGFR-PTEN mutations and IDH1-TP53-ATRX mutations in glioblastoma was depicted here and was consistent to the current literature [27, 41]. Notably, the impact of these novel patterns remains unknown and falls beyond the scope of this study.

A more technical insight arises from a comparison between our method and the work of J. Iranzo *et al.* [27], who conducted the most comparable study. The lower number of modules recovered by the MODULAR pipeline in our study, likely due to an increase in patients numbers, led us to hypotherize that Infomap could detect nested relations within the MODULAR baseline partition, due to its ability to overcome the well-known resolution limit of modularity-based algorithms. Indeed, Infomap increased the overall number of assigned patients from 1527 to 2037, providing meaningful results for lung adenocarcinoma, prostate adenocarcinoma and testicular germ cell tumors, which faced limitations in the MODULAR partition. Detailed analysis of discrepancies between the two pipelines is presented in the Supplementary File.

Overall, the use of both clustering analysis pipelines has uncovered clusters containing reported and non-reported statistically significant relations. Further research on the non-reported relations is imperative to validate or invalidate their oncogenic role. In conclusion, our study advocates for the incorporation if intra-clustering statistical analysis in cancer genomics, emphasizing the potential insights gained by examining smaller patient cohorts.

3.1 Statistical Considerations

The statistical analysis employed in this study is based on the Fisher's exact test. Seven clusters from the Infomap pipeline and two clusters from the MODULAR pipeline were found to predominantly contain a single cancer type. These clusters exhibit enrichment in cancer-specific driver relations, with only 7 out of the 42 cancer-gene combinations not reported as significant in the current literature. Even when the Fisher's exact test failed to recover these relations by default, the pattern detection statistical analysis proved efficient, revealing significant patterns within these clusters. In all cases except one (NSD1-PTCH1 co-occurrences in uterine corpus endometrial carcinoma), at least one of the two genes implicated in the pattern is a known significantly mutated gene.

4 Materials and Methods

4.1 Data Acquisition

In order to construct a gene-patient network, data from the TCGA-PanCancer Atlas project were downloaded from cBioPortal [67, 68], encompassing the thirty cancer types detailed in the results. Two of these cancer types include colon adenocarcinoma and rectal adenocarcinoma. Patients of both cancer types were included in a single dataset (colorectal adenocarcinoma), along with mucinous adenocarcinoma of the colon and rectum patients. Patients of this mixed cancer type were excluded from the network.

4.2 Data Preprocessing

To minimize noise, hypermutated samples (those with over 3000 mutations across all genes) were excluded from the dataset. After this filtering, mutations in 198 known cancer driver genes, as derived from [69], were included in the curated dataset. Only high-impact mutations, including missense, nonsense, nonstop, frameshifts, in-frame deletions and insertions, and mutations occurring in splice sites [27, 70], were considered. Metastatic samples were also removed to enhance data homogeneity.

4.3 Network Creation

The curated dataset was utilized to construct a network representing 30 cancer types, based solely on somatic point mutations. Each edge in the network represents a mutation in the patient on one endpoint, harbored in the gene of the other endpoint. In that context, the network is bipartite as existence of patient-to-patient or gene-to-gene edges in this network will have no physical meaning. The bipartite setting of the network allows for utilization of tools specifically designed for this topology [71, 72].

Various tools and metrics are available for the characterization of the networks. One such tool, applicable to bipartite networks, is the biadjacency matrix. This matrix is of size $m \times n$, where m and n represent the number of nodes in the two sets that constitute the network. In this context, m denotes the number of genes, while n corresponds to the number of patients. Another metric employed is the network density. To calculate the density of a bipartite network, the number of the edges E in the network is compared to the number of edges in a fully connected bipartite network with an equivalent number of nodes, distributed across the two sets that constitute the network. The maximum number of edges is given by the combinations between the two disjoint sets, i.e., $m \times n$. The density in bipartite networks can be defined mathematically as: $D = E/(m \times n)$.

In this context, the biadjacency matrix $\{a_{ij}\}$ of the network is binary, with "1" indicating the presence of at least one of the aforementioned mutations in gene i of patient j, while "0" denotes the absence of such mutations. The size of the biadjacency matrix of the network was 198×8303 (198 genes and 8303 patients) containing 34994 edges. This yields a density of approximately 0.0213, leading to the deduction that the network is highly sparse.

4.4 Methods Overview

The methodology employed to unveil the underlying structure of the network involved four algorithms: Simulated Annealing (SA), Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Infomap, and Order Statistics Local Optimization Method 2 (OSLOM2).

SA was implemented through the MODULAR software [71], which provides algorithms that aim to maximize the modularity index. In bipartite networks, the Barber's modularity index [72] is commonly utilized, serving as the reference index of this study. A higher index value, approaching 1, signifies a more favorable corresponding partition. Since modularity maximization is an NP-hard problem, SA approximates the global maximum. The resulting partition corresponds to the local maximum of the modularity index found by SA.

UPGMA, an agglomerative hierarchical clustering method, based on the average dissimilarity index [73], was implemented using the "scipy.cluster.hierarchy.linkage" function of package "scipy" in Python, with the "method" parameter set to "average" [74].

Infomap [75], relying on network dynamics and minimum description length statistics (MDL), leverages random walks to cluster the network based on walk trajectories. For a given a cluster, the algorithm calculates the probability of a random walker remaining within the cluster and the probability of the random walker exiting the cluster. This information can be encapsulated in the "Map Equation", and the resulting partition is the one that minimizes this equation.

OSLOM2 [76], a statistics-centric algorithm, aims to recover statistically meaningful modules. OSLOM2 employs a method called "Single Cluster Analysis" (SCA), which for a given $P-value$ examines the statistical significance of each cluster locally, by iteratively adding and removing nodes. In order to recover a partition of the network, SCA is implemented multiple times. OSLOM2 also assesses if the union of two or more clusters is more statistically significant than the clusters individually, resulting in a partition with statistically significant clusters and potentially with nodes that do not belong in any cluster (homeless nodes).

4.5 Clustering Pipelines

4.5.1 Infomap-OSLOM2 pipeline

Infomap was utilized to establish an initial partition. Given the stochastic nature of the algorithm, Infomap underwent implementation 200 times and the outcomes were combined using Monti's consensus clustering algorithm (see section 4.5.3). Infomap was set to recover a "two-level" partition over an "undirected" graph using the "skip-adjust-bipartite-flow" option. The consensus matrix was weighted with elements varying between 0 and 1. Any elements with a value below 0.4 [27] were zeroed, returning an unweighted (binary) final matrix M_I^*. Infomap was implemented again once over the matrix M_I^*, with the same parameters used in the 200 executions to get the partition C_{info}.

OSLOM2 was then implemented in the partition C_{info} excluding clusters that lacked statistical significance with a P-value lower than 0.05 [27]. This led to the final partition C_{info}^*. The term 'Infomap pipeline' will be used to denote this entire pipeline.

4.5.2 MODULAR-UPGMA-OSLOM2 pipeline

The SA algorithm of MODULAR was employed to obtain an initial partition. Given the stochastic nature of the algorithm, MODULAR underwent 200 implementations and the outcomes were combined using Monti's consensus clustering algorithm (see section 4.5.3). To address the resolution limit of modularity-based methods, a hierarchical clustering approach was utilized [77]. This limit stipulates that the resultant partition from such methods depends on the size of the network, potentially impeding the identification of small-scale modules. The application of a hierarchical algorithm to a dissimilarity matrix of the network mitigates this challenge. The consensus clustering matrix M_M^* is a similarity matrix [78] of the network, and therefore, $1-M_M^*$ will is a dissimilarity matrix (please note that M_M^* is weighted) [79]. UPGMA was then applied over $1-M_M^*$, generating a dendogram with several branches. For each branch, nodes below that branch were assigned in modules via the UPGMA, while where the rest of the nodes remained unassigned. Each unassigned node would consist of a single-node module. For every possible branch, the Barber's modularity index (with respect to the original network) was calculated. The branch, whose corresponding partition maximized the index was considered as the optimal branch. This partition will be referred to as C_{mod}.

OSLOM2 was then implemented in the partition C_{mod} partition, mirroring the Infomap pipeline, to yield the final partition C_{mod}^*. The term 'MODULAR pipeline' will be used to denote this entire pipeline.

4.5.3 Ensemble and Consensus Clustering

Monti's consensus clustering algorithm was employed to integrate multiple individual outcomes [79]. For each partition h recovered by any of the algorithms, the matrix M^h was established. The pair (i,j) in the matrix M^h is set to "1" if nodes i and j belonged to the same cluster in the partition h; otherwise the pair (i,j) was set to "0". The final consensus matrix, denoted as $\sum_h M^h/H$, was then defined, where H represents the total number of partitions obtained by the algorithm.

4.6 Threshold Filtering

The final C_{mod}^* and C_{info}^* partitions contained modules with patient nodes from several cancer types. To enhance the robustness of the results, a threshold of 5% [27] per cancer type was maintained as the minimum percentage regarded significant for the MODULAR pipeline outcomes. It's worth noting that Infomap is less affected by the resolution limit compared to modularity-based methods [80]. Consequently, a lower threshold of

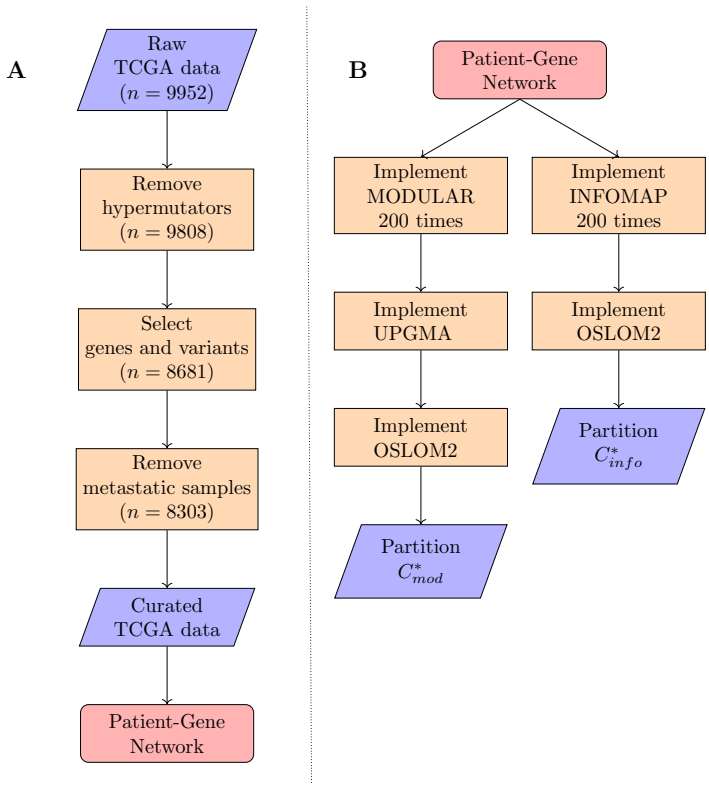

Figure 3: Two flow diagrams representing: (A) the data preprocessing as described in section 4.2 and (B) the implementation of the clustering analysis pipelines as described in section 4. In flow chart (A) the number n indicates the number of patients after the corresponding procedure. The flow chart (B) is separated in two columns, indicating that the pipelines where implemented independently and in parallel.

4% was selected for the Infomap pipeline, expecting the detection of smaller-sized modules than those identified by the Modular pipeline, if such modules exist.

4.7 Statistical Analysis

Following the threshold filtering of the results described in Section 4.6, each module appearing in the final results underwent further analysis. More specifically, two separate statistical pipelines were executed for each module recovered by any of the pipelines: (i) to evaluate the significance of an intra-cluster cancer-gene relation and (ii) to assess the significance of a gene-to-gene pattern for any of the cancer types within the module.

For case (i), right-tailed Fisher's exact tests were employed to compare the mutational load of a gene among patients of a specific cancer type with respect to the rest of the patients from other cancer types within the module. For case (ii), two-tailed Fisher's exact tests were applied by counting the co-existing mutations in patients of the same cancer type within the cluster.

More specifically, let C be a module with patients of the cancer types $c_i, i \in 1, 2, \ldots m$ and genes $g_j, j \in 1, 2, \ldots n$. Then, for the cases (i) and (ii), $m \times n$ and $m \times \binom{n}{2}$ p-values were obtained for every cluster. To minimize the number of false positives (i.e. the number of statistically significant cancer-gene relations that are not truly significant), we implemented the false discovery rate method of Benjamini and Hochberg [81] to adjust the p-values in both cases. We set the false discovery rate threshold to be 0.05.

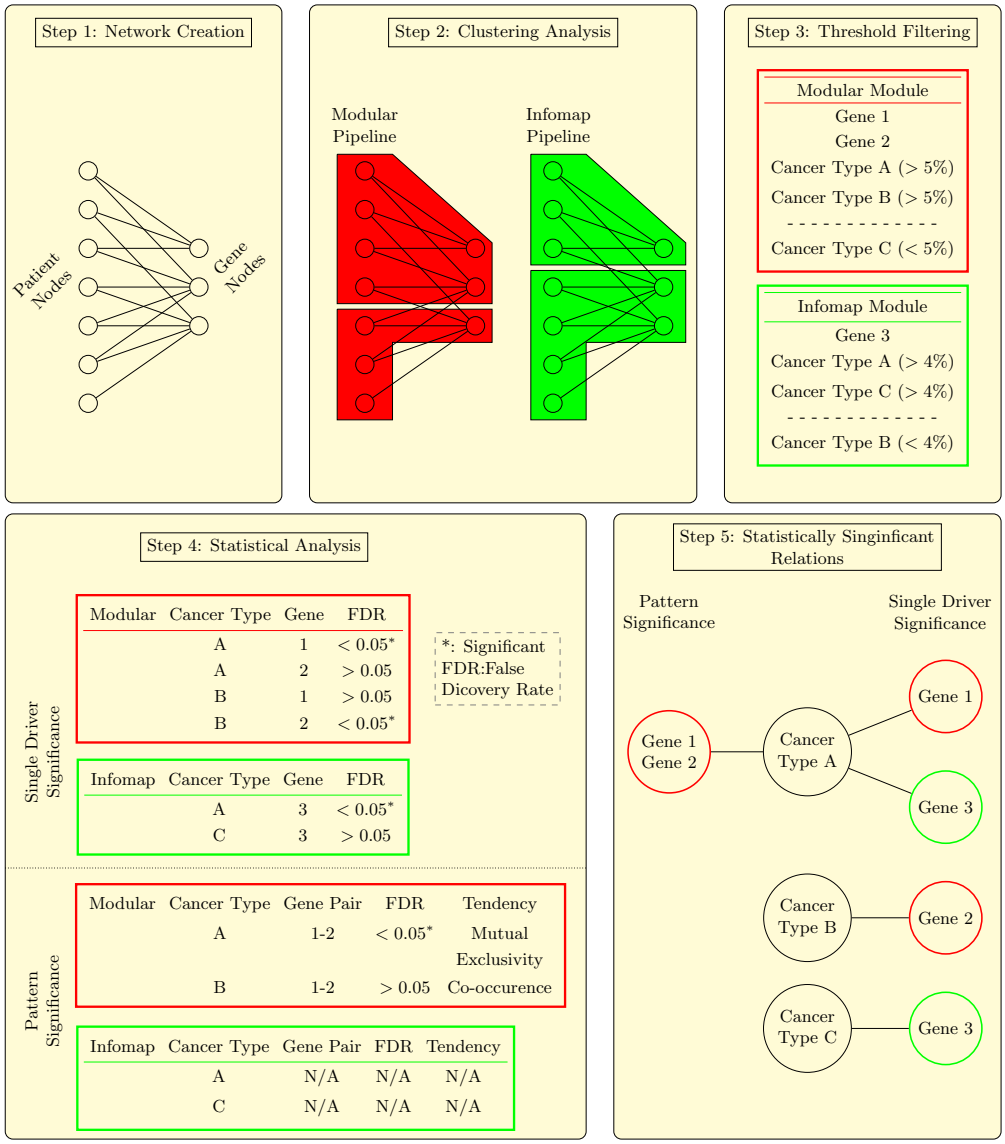

Figure 4: A schematic overview of the proposed method for a network with patients from cancer types A, B and C, and genes 1, 2 and 3. Step 1: The network was created from somatic-point-mutation data and is, by default, bipartite. Step 2: We use the two clustering analysis pipelines to obtain two different partitions of the network in clusters. Step 3: We filtered each cluster with respect to the amount of patients per cancer type, in order to provide robust results per cancer type. The threshold of each pipeline depends on the underlying mechanism of the initial algorithm used to obtain the final partition. Step 4: We performed Fisher's exact test on each cluster for every cancer-gene combination in the cluster and for every gene-to-gene pattern for every cancer type within the cluster to obtain the p-values. The Benjamini-Hochberg false discovery rate method was employed on each cluster to obtain the adjusted p-values. A relation was considered significant if the adjusted p-value was lower than 0.05. The cases of the single gene significance and pattern significance were treated separately. Step 5: Visualization of the significant relations per cancer type.

Appendix

Modular Structure

In order to ensure that the network had modular structure and clustering could indeed by implemented, data of MODULAR executions were used. In every execution of the SA algorithm of MODULAR, the algorithm was also used over the two null models that MODULAR provides. After every execution, the modularities of the network and the modularites of the two null models were stored in three separate vectors. These vectors were used to perform Welch's t-test (Eq. 1) to compare the network modularities with both the first and the second null model.

$$t = \frac{\tilde{X}_1 - \tilde{X}_2}{\sqrt{\sigma^2_{\tilde{X}_1} + \sigma^2_{\tilde{X}_2}}}. \tag{1}$$

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.

Methods Overview

SA was implemented through the software MODULAR [71], which provides algorithms that aim to maximize the modularity index. A higher index value, approaching 1, signifies a more favorable corresponding partition. In bipartite networks, the Barber's modularity index is often used [72]. In general, modularity can be mathematically calculated by equation 2:

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

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 clusters. Equation 2 reveals that algorithms based on modularity maximization, rely on simple edges between the nodes of the network.

UPGMA is a simple agglomerative hierarchical clustering method. The algorithm is based on the average dissimilarity:

$$d(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}, \tag{3}$$

between the two clusters G, H, exploiting the distance matrix D_{ij} [73]. The terms N_G, N_H correspond to the number of the nodes of the clusters G, H accordingly. The distance matrix has a size of D_{ij} is a $(N_G + N_H) \times (N_G + N_H)$, with elements $d_{ii'}$ which correspond to the minimum distance of node i to $i^{'}$

Infomap [75] is an algorithm based on network dynamics. Minimum description length statistics (MDL) can 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 main task of the software is to be able to describe the locations of random walks as efficiently as possible, depicting the underlying structure of the network. More 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. Mathematically, the map equation is expressed as:

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i). \tag{4}$$

The first term of equation 4 corresponds to the average number of bits used to describe inter-modular movements, as $q_{\curvearrowright} = \sum_{j=1}^{m} q_{\curvearrowright}$ is the probability of the random walker to switch modules (per-step probability). The second term is indicative of the intra-modular movements, as $p_{\circlearrowright} = q_{\curvearrowright} + \sum_{\alpha \in i} p_{\alpha}.$, represents the number of times the codebook of module i was used. The $p_{\circlearrowright}$ consists of the sum of the steady state distribution for all α nodes within the modules and the probability of exiting the module. The quantity $H(X) = -\sum p_i log_2(p_i)$ is the entropy of the random variable X with n states and p_i frequencies. A logarithm of base 2 (log_2) is employed since the code lengths are measured in bits and are therefore binary.

OSLOM2 [76] is an algorithm based solely in statistics. OSLOM2 is able to read an initial partition and implement a method called "Single Cluster Analysis" (SCA). Given equationa $P-value$, this procedure examines locally the statistical significance of each of the clusters of the networequationk, by adding and removing nodes from the rest of the network. In order to recover a partition of the network, SCA is implemented multiple

times. OSLOM2 also verifies if the union of two (or more) clusters is more statistically significant than the clusters themselves (separately). This full network analysis of OSLOM2 results in a partition with statistically significant clusters and, possibly, nodes that do not belong in any cluster (homeless nodes).

Acknowledgments

The authors would like to thank Dr. Jaime Iranzo for his guidance and advice on terms of network and pipeline construction. His thorough explanation on both technical and biological matters had a crucial role on the completion of this work.

Contact

Stamatis Choudalakis: st_xoud@hotmail.com
George Kastis: gkastis@Academyofathens.gr
Nikolaos Dikaios: ndikaios@Academyofathens.gr

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Code and Data Availability

The code will be published in Github along with the curated data. Raw data can be downloaded from cBioPortal.org.

References

[1] George Pentheroudakis, Vassilios Golfinopoulos, and Nicholas Pavlidis. "Switching benchmarks in cancer of unknown primary: from autopsy to microarray". In: *European journal of cancer* 43.14 (2007), pp. 2026–2036.

[2] 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.

[3] F Anthony Greco et al. "Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction". In: *The oncologist* 15.5 (2010), pp. 500–506.

[4] F Losa et al. "SEOM clinical guideline on unknown primary cancer (2017)". In: *Clinical and Translational Oncology* 20 (2018), pp. 89–96.

[5] Nabeel Al-Brahim et al. "The value of postmortem examination in cases of metastasis of unknown origin—20-year retrospective data from a tertiary care center". In: *Annals of diagnostic pathology* 9.2 (2005), pp. 77–80.

[6] 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.

[7] 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.

[8] Yuchen Yuan et al. "DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations". In: *BMC bioinformatics* 17.17 (2016), pp. 243–256.

[9] Kanggeun Lee et al. "CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network". In: *Scientific reports* 9.1 (2019), pp. 1–9.

[10] 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.

[11] Yukun Chen et al. "Classification of cancer primary sites using machine learning and somatic mutations". In: *BioMed research international* 2015 (2015).

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

[13] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[14] Yun-Yun Tang et al. "Identification of driver genes based on gene mutational effects and network centrality". In: *BMC bioinformatics* 22.3 (2021), pp. 1–16.

[15] Siân Jones et al. "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses". In: *science* 321.5897 (2008), pp. 1801–1806.

[16] Nathan D Dees et al. "MuSiC: identifying mutational significance in cancer genomes". In: *Genome research* 22.8 (2012), pp. 1589–1598.

[17] 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.

[18] Giovanni Parmigiani et al. "Design and analysis issues in genome-wide somatic mutation studies of cancer". In: *Genomics* 93.1 (2009), pp. 17–21.

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

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

[21] 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.

[22] Nikolaos Dikaios. "Sparse-Input Neural Networks to Differentiate 32 Primary Cancer Types on the Basis of Somatic Point Mutations". In: *Onco* 2.2 (2022), pp. 56–68.

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

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

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

[26] 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.

[27] 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.

[28] Marine Le Morvan, Andrei Zinovyev, and Jean-Philippe Vert. "NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis". In: *PLoS computational biology* 13.6 (2017), e1005573.

[29] Christof Winter et al. "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes". In: *PLoS computational biology* 8.5 (2012), e1002511.

[30] Tom Ronan, Zhijie Qi, and Kristen M Naegle. "Avoiding common pitfalls when clustering biological data". In: *Science signaling* 9.432 (2016), re6–re6.

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

[32] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. "Comprehensive molecular portraits of human breast tumours". In: *Nature* 490.7418 (2012), pp. 61–70.

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

[34] Yi-Zhou Jiang et al. "GATA3 mutations define a unique subtype of luminal-like breast cancer with improved survival". In: *Cancer* 120.9 (2014), pp. 1329–1337.

[35] Yulan Deng et al. "Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability". In: *Briefings in Bioinformatics* 20.1 (2019), pp. 254–266.

[36] Xing Hua et al. "MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations". In: *The American Journal of Human Genetics* 98.3 (2016), pp. 442–455.

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

[38] Rebeca Sanz-Pamplona et al. "Exome sequencing reveals AMER1 as a frequently mutated gene in colorectal cancer". In: *Clinical cancer research* 21.20 (2015), pp. 4709–4718.

[39] Siqin Lei et al. "AMER1 deficiency promotes the distant metastasis of colorectal cancer by inhibiting SLC7A11-and FTL-mediated ferroptosis". In: *Cell Reports* 42.9 (2023).

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

[41] Cameron W Brennan et al. "The somatic genomic landscape of glioblastoma". In: *Cell* 155.2 (2013), pp. 462–477.

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

[43] Pradip De et al. "Active RAC1 promotes tumorigenic phenotypes and therapy resistance in solid tumors". In: *Cancers* 12.6 (2020), p. 1541.

[44] Hoi Lam Ngan et al. "RAC1 genomic aberrations as predictive biomarkers for head and neck squamous cell carcinoma (HNSCC)". In: *Cancer Research* 79.13_Supplement (2019), pp. 4033–4033.

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

[46] James Brugarolas. "Molecular genetics of clear-cell renal cell carcinoma". In: *Journal of clinical oncology* 32.18 (2014), p. 1968.

[47] 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.

[48] 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.

[49] 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.

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

[51] Felix Sahm et al. "CIC and FUBP1 mutations in oligodendrogliomas, oligoastrocytomas and astrocytomas". In: *Acta neuropathologica* 123 (2012), pp. 853–860.

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

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

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

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

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

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

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

[59] Chuan Zhang et al. "Pivotal factors associated with the immunosuppressive tumor microenvironment and melanoma metastasis". In: *Cancer medicine* 10.14 (2021), pp. 4710–4720.

[60] João T Barata, Scott K Durum, and Benedict Seddon. "Flip the coin: IL-7 and IL-7R in health and disease". In: *Nature immunology* 20.12 (2019), pp. 1584–1593.

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

[62] Nishant Agrawal et al. "Integrated genomic characterization of papillary thyroid carcinoma". In: *Cell* 159.3 (2014), pp. 676–690.

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

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

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

[66] Carl Koschmann et al. "ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma". In: *Science translational medicine* 8.328 (2016), 328ra28–328ra28.

[67] 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.

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

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

[70] Collin J Tokheim et al. "Evaluating the evaluation of cancer driver genes". In: *Proceedings of the National Academy of Sciences* 113.50 (2016), pp. 14330–14335.

[71] 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]`.

[72] Michael J Barber. "Modularity and community detection in bipartite networks". In: *Physical Review E* 76.6 (2007), p. 066102.

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

[74] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

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

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

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

[79] 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.

[80] Tatsuro Kawamoto and Martin Rosvall. "Estimating the resolution limit of the map equation in community detection". In: *Physical Review E* 91.1 (2015), p. 012809.

[81] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.