BiGPICC: a graph-based approach to identifying carcinogenic gene combinations from mutation data

Vladyslav Oles[1,*], Sajal Dash[1], and Ramu Anandakrishnan[2]

[1]Oak Ridge National Laboratory, National Center for Computational Sciences, Oak Ridge, 37830, USA
[2]Virginia Tech, Edward Via College of Osteopathic Medicine, Blacksburg, 24060, USA
[*]vladyslav.oles@gmail.com

ABSTRACT

Genome data from cancer patients encapsulates explicit and implicit relationships between the presence of a gene mutation and cancer occurrence in a patient. Different types of cancer in human are thought to be caused by combinations of two to nine gene mutations. Identifying these combinations through traditional exhaustive search requires the amount of computation that scales exponentially with the combination size and in most cases is intractable even for even cutting-edge supercomputers. We propose a parameter-free heuristic approach that leverages the intrinsic topology of gene-patient mutations to identify carcinogenic combinations. The biological relevance of the identified combinations is measured by using them to predict the presence of tumor in previously unseen samples. The resulting classifiers for 16 cancer types perform on par with exhaustive search results, and score the average of 80.1% sensitivity and 91.6% specificity for the best choice of hit range per cancer type. Our approach is able to find higher-hit carcinogenic combinations targeting which would take years of computations using exhaustive search.

1 Introduction

Multi-hit theory of carcinogenesis states that it takes combinations of gene mutations to initiate carcinogenesis in humans[1]. Clinical studies and a body of mathematical models established that the size of such combinations (hits) can range from 2 to 9 depending on the cancer types[2–6]. However, most computational efforts to find carcinogenic mutations focus on finding individual "driver mutations" responsible for carcinogenesis[7–10]. While these mostly mutational frequency- and signature-based methods identify driver mutations associated with increased risk of cancer, these mutations by themselves cannot cause cancer.

Cancer-screening methods based on genetic predisposition rely heavily on identifying driver mutations. However, some people with a genetic predisposition may never get the cancer, while the others get the cancer as they grow older and accumulated more mutations. For example, women under the age of 20 with the BRCA1 mutation are very unlikely to get breast cancer; moreover, 28% of all women with BRCA1 mutation never get the cancer[11]. Similar statistics can be observed for Li Fraumeni syndrome[12–15]. These observations provide strong evidence that cancer in a patient is caused not by an individual driver gene mutation but rather by a combination of gradually accumulated gene mutations. Though there are several other factors responsible for cancer such as epigenetic modifications, tumor environment, and adaptive evolution, carcinogenesis is primarily a result of genetic mutations.

Different combinations of gene mutations can cause cancer of the same type but with different etiologies and pathologies representing different subtypes. To design and develop individualized and precision drugs for treating an individual with cancer, we need to identify the combinations of gene mutations in that patient. Since current computational approaches mostly look for driver gene mutations, they come short in the context of precision drug discovery. The human genome is comprised of $\sim$20,000 genes and each gene can host dozens of mutations. Finding the carcinogenic combinations of size h requires examining $\binom{|\mathscr{G}|}{h}$ combinations even with the simplistic assumption of one mutation per gene (here, $\mathscr{G}$ is the set of genes in human genome). For $h > 4$, computationally evaluating such a large number of combinations can take years. In this work, we propose an approach that can efficiently find carcinogenic gene combinations even for large values of h.

1.1 Exhaustive search methods

A recent body of work[16–18] explored all possible gene combinations to identify those responsible for cancer based on tumor and normal samples from cancer patients. They mapped the task to a weighted set cover problem, where each combination is associated with a set of tumor samples in which it is jointly mutated and assigned weight based on its classification accuracy when used for differentiating between tumor and normal samples. They implemented a heuristic approximating the minimum weight set cover that requires $O\left(\binom{|\mathscr{G}|}{h}\right)$ time for identifying a single combination of length h. For $|\mathscr{G}| \approx 20{,}000$, the size of

human genome, and $h \geq 5$ it can take days to years on one of the fastest supercomputers in the world, Summit.

Klein et al.[19] developed another computational method for inferring combinations of gene mutations that cooperatively drive tumor formation in individual patients. Their parameter-heavy approach combines exhaustive search and heuristics when navigating the combinatorial space for the human genome, and can identify carcinogenic combinations of length h for $h \leq 6$.

1.2 Graph algorithms in cancer genomics

Graphs are combinatorial structures that model pairwise relationships between objects, e.g. protein-protein interactions (PPI) or mutations of genes in humans. In recent years, a number of studies have relied on graph-based analyses to identify combinations of cancer drivers[20–25]. A typical approach is to enrich a molecular network of choice using gene-sample mutation data, then analyze the resulting graph, usually by means of network propagation, to identify clusters of carcinogenic genes. For example, Leiserson et al.[21] use a heat diffusion algorithm on PPI networks to find their mutated subnetworks, based on the gene mutation frequencies extracted from data. The identified communities (or modules) are hypothesized to represent carcinogenic gene combinations, and many of the corresponding proteins have a documented role in development of cancer. In a similar vein, Leung et al.[22] employ a greedy search to identify co-mutated communities in gene networks. Cho et al.[20] rely on gene neighborhoods in functional networks in combination with mutation information.

Unlike the above studies, Iranzo et al.[26] directly use the graphical structure of mutation data to examine community structure in cancer drivers. The approach of this work is similar to ours in the idea of leveraging the topology of gene-sample mutations to learn about group interactions of carcinogenic genes, although their methodology is different. Another distinction is that Iranzo et al. study the $\sim$1% of the human genome that has known associations with cancer, while our work aims to discover novel carcinogenic combinations by analyzing the full genome.

1.3 Our contribution

Here we present BiGPICC (Bipartite Graph Partitioning for Identifying Carcinogenic Combinations) — a parameter-free approach to finding multi-hit combinations based on the topology of mutation data. Analyzing mutation data is potentially more insightful for cancer genomics than using molecular networks, as the latter are simplified abstractions of gene or protein interactions. To the best of our knowledge, this is one of the first network-based algorithms working directly with gene-sample mutations.

We formulate the search in terms of a community detection problem on a bipartite graph representation of the data, and design and implement an algorithm for solving it. Our numerical experiments on Summit supercomputer for 16 cancer types demonstrate that it identifies combinations of comparable biological relevance to state-of-the-art results. At the same time, our approach is capable of efficiently identifying relevant (5+)-hit combinations, unavailable to the existing methods due to high computational cost. An additional advantage of out method is that it does not require any manual tuning due to the absence of parameters representing some form of domain knowledge, which means that it is readily available to work with a broad variety of datasets.

2 Methods

Our approach for identifying carcinogenic combinations from data on gene mutations in tumor and normal tissues is outlined in Figure 1. Viewing the data as a bipartite graph, we iteratively partition it using community detection to find candidate gene combinations whose mutations tend to occur in tumor tissues, then filter out those with frequent mutations in normal samples. The final set of carcinogenic combinations is identified by selecting a minimum set cover of the tumor samples from the filtered candidate pool. To assess the relevance of the identified combinations, we use them to differentiate between tumor and normal samples on previously unseen data. A detailed description of these steps is provided below.

2.1 Graph formulation

Let $\mathscr{G}$ and $\mathscr{S}$ denote the sets of genes and samples in the data, respectively. The input data contains information about whether a mutation of gene g was observed in sample s for every gene-sample pair $(g,s) \in \mathscr{G} \times \mathscr{S}$. Specifically, the data is stored in a binary $|\mathscr{G}| \times |\mathscr{S}|$ matrix $\mathbf{D}$ whose entries are 1 if mutation of the corresponding gene was observed in the corresponding sample, and 0 otherwise. This information can be equivalently represented as an unweighted bipartite graph on the vertices of two distinct classes $\mathscr{G}$ and $\mathscr{S}$, where an edge connects some $g \in \mathscr{G}$ and $s \in \mathscr{S}$ if and only if the corresponding mutation has been observed. We denote such a graph G, and notice that the adjacency matrix of G is the symmetric block matrix $\begin{bmatrix} \mathbf{0} & \mathbf{D} \\ \mathbf{D}^T & \mathbf{0} \end{bmatrix}$. The presence of both normal and tumor samples in the data means that the vertex class $\mathscr{S}$ is itself partitioned into two subclasses, denoted $\mathscr{S}^{\text{norm}}$ and $\mathscr{S}^{\text{tum}}$ respectively, with opposing significance to the problem. To isolate the information about gene mutations in tumor tissues, we consider the graph $G^{\text{tum}} := G[\mathscr{G} \sqcup \mathscr{S}^{\text{tum}}]$, the induced subgraph of G on the vertices $\mathscr{G} \sqcup \mathscr{S}^{\text{tum}}$.

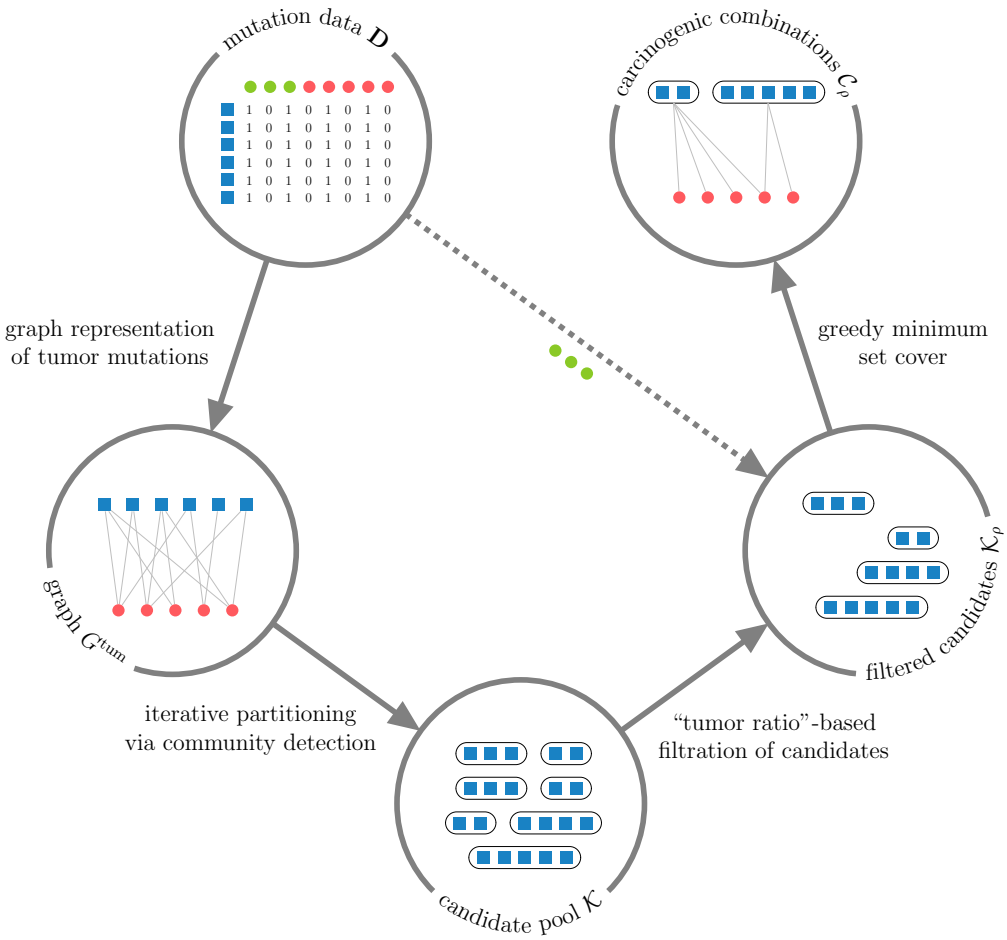

Figure 1. Outline of BiGPICC approach. Its four stages, shown as solid arrows in counterclockwise order starting from the upper left, are described respectively in Sections 2.1 through 2.4.

Recall that a *community* is a subset of graph vertices that are more densely connected with one another than with the rest of the graph, according to some metric. Any community C in G (or G^{tum}) consists of the gene component $C_{\mathscr{G}} := C \cap \mathscr{G}$ — a combination of genes whose mutations tend to occur together, and the sample component $C_{\mathscr{S}} := C \cap \mathscr{S}$ (or $C \cap \mathscr{S}^{\text{tum}}$, respectively) — the samples in which these genes co-mutate the most. Because bipartite communities of multiple vertices must contain vertices from both classes to allow for internal connections, both $C_{\mathscr{G}}$ and $C_{\mathscr{S}}$ are non-empty whenever $|C| > 1$.

For some gene $g \in \mathscr{G}$, let $M(g) \subset \mathscr{S}$ denote the samples in which a mutation of g has been observed. Furthermore, for a gene combination $C_{\mathscr{G}}$, denote the samples in which these genes are jointly mutated as $M(C_{\mathscr{G}}) := \bigcap_{g \in C_{\mathscr{G}}} M(g)$. Analogously, we define the tumor samples with a mutation of g as $M^{\text{tum}}(g) := M(g) \cap \mathscr{S}^{\text{tum}}$, and with a joint mutation of $C_{\mathscr{G}}$ — as $M^{\text{tum}}(C_{\mathscr{G}}) := M(C_{\mathscr{G}}) \cap \mathscr{S}^{\text{tum}}$. If $C_{\mathscr{G}}$ is carcinogenic, its joint mutation is thought to explain some number of tumor samples from the data, so $C_{\mathscr{G}}$ must be fully connected to the non-empty set $M^{\text{tum}}(C_{\mathscr{G}})$ in the graph representation of data. For any unrelated gene $g' \in \mathscr{G} \setminus C_{\mathscr{G}}$, the connectivity between genes $C_{\mathscr{G}} \cup \{g'\}$ and samples $M^{\text{tum}}(C_{\mathscr{G}})$ is expected to be less dense, because a mutation of g' is unlikely to appear in every sample from $M^{\text{tum}}(C_{\mathscr{G}})$ by chance unless it is mutated in most of $\mathscr{S}^{\text{tum}}$.

It follows that in the absence of highly mutable genes any carcinogenic gene combination $C_{\mathscr{G}}$ is expected to correspond to a community $C_{\mathscr{G}} \sqcup M^{\text{tum}}(C_{\mathscr{G}})$ in G^{tum}. At the same time, the converse is not necessarily true — because carcinogenic combinations are assumed to jointly mutate in predominantely tumor samples, not every community in G^{tum} has its gene component as a carcinogenic combination. For example, if tumor samples are but a small fraction of all the samples with a joint mutation of $C_{\mathscr{G}}$ (i.e. $|M^{\text{tum}}(C_{\mathscr{G}})| \ll |M(C_{\mathscr{G}})|$), then $C_{\mathscr{G}}$ is unlikely a carcinogenic combination even despite $C_{\mathscr{G}} \sqcup M^{\text{tum}}(C_{\mathscr{G}})$ being a community in G^{tum}. Therefore, the task of finding carcinogenic combinations of gene mutations can be cast as a problem of identifying communities with some desired degree of tumor prevalence in their sample components.

2.2 Community detection

We use the *Constant Potts Model* (CPM) to formally define the notion of community structure in a graph. CPM was proposed as an alternative to the commonly-chosen *modularity* approach that does not suffer from the issue of inconsistent communities across different scales[27]. For the case of G^{tum}, an unweighted bipartite graph on the vertex classes $\mathscr{G}$ and $\mathscr{S}^{\text{tum}}$, CPM formalizes its partition into disjoint communities $\mathscr{P}$ (i.e. $\bigsqcup_{C\in\mathscr{P}} C = \mathscr{G} \sqcup \mathscr{S}^{\text{tum}}$) as desirable if it maximizes the partition quality

$$\sum_{C\in\mathscr{P}} (e(C) - \gamma|C_{\mathscr{G}}||C_{\mathscr{S}}|), \tag{$*$}$$

where $e(C)$ is the total number of internal edges in community C (that is, between $C_{\mathscr{G}} = C \cap \mathscr{G}$ and $C_{\mathscr{S}} = C \cap \mathscr{S}^{\text{tum}}$), and $\gamma \in [0,1]$ is the so-called resolution parameter. This parameter can be viewed as the density of internal connectivity required from a set of vertices C for it to qualify as a community — or, more specifically, for C to positively contribute to the partition quality ($*$). Notice that whenever gene component or sample component of C are empty, its contribution to ($*$) is zero.

To identify communities within a graph, we rely on the Leiden algorithm[28], chosen for its guarantees on the community connectivity and speed. In particular, the Leiden algorithm converges to a partition in which all subsets of all communities are guaranteed to be locally optimally assigned. The Leiden algorithm starts by viewing every vertex as a separate community, and then alternates between moving nodes between communities, refining the partition, and aggregating communities into single vertices to reduce the graph size, until no further improvement to the chosen partition quality (in our case, ($*$)) can be made.

Because the possible number of carcinogenesis drivers in a combination is known or hypothesized for many cancer types, we are interested in finding communities whose gene component size $h := |C_{\mathscr{G}}|$ is within a certain range $l \le h \le u$ obtained from the literature. To enforce this size constraint, we iteratively refine the identified communities using the Leiden algorithm until the size of their gene component does not exceed u, the maximum possible number of carcinogenesis drivers. Specifically, if a community C in G^{tum} has gene component $C_{\mathscr{G}}$ that is too large, we refine it by first extending its sample component to be $\bigcup_{g\in C_{\mathscr{G}}} M^{\text{tum}}(g)$ to include all relevant tumor samples, then applying the Leiden algorithm to the resulting subgraph $G^{\text{tum}}[C_{\mathscr{G}} \sqcup \bigcup_{g\in C_{\mathscr{G}}} M^{\text{tum}}(g)]$ to partition $C_{\mathscr{G}}$ based on their mutations. Every time the Leiden algorithm is run on a (sub)graph induced by genes $C_{\mathscr{G}}$ and samples $C_{\mathscr{S}}$, the CPM resolution parameter is set to the connectivity density of this (sub)graph, $\gamma = \frac{e(C_{\mathscr{G}} \sqcup C_{\mathscr{S}})}{|C_{\mathscr{G}}||C_{\mathscr{S}}|}$, requiring the communities to be internally connected at least as densely as the (sub)graph itself. After filtering out the the communities whose gene component is too small from the results, the gene components of the remaining communities are considered candidates for carcinogenic combinations.

Procedure 1 Iterative Partitioning

Input: graph G^{tum} on vertices $\mathscr{G} \sqcup \mathscr{S}^{\text{tum}}$, carcinogenic combination size bounds l and u
Output: candidates for carcinogenic combinations $\mathscr{K}^{\bullet}$

```
1:  𝒦• ← ∅
2:  𝒬 ← {𝒢}                                  ▷ initialize queue of gene combinations to partition
3:  for C_𝒢 ∈ 𝒬 do                            ▷ process next gene combination C_𝒢
4:      𝒬 ← 𝒬 \ {C_𝒢}
5:      C ← C_𝒢 ⊔ ⋃_{g∈C_𝒢} M^tum(g)          ▷ include all connected samples
6:      𝒫 ← Leiden(G^tum[C])                  ▷ partition C into smaller communities
7:      for C' ∈ 𝒫 do
8:          if C'_𝒢 ⊊ C_𝒢 then                ▷ ensure C_𝒢 was reduced by the partitioning
9:              if |C'_𝒢| > u then
10:                 𝒬 ← 𝒬 ∪ {C'_𝒢}            ▷ set C'_𝒢 for partitioning
11:             else if l ≤ |C'_𝒢| then
12:                 𝒦• ← 𝒦• ∪ {C'_𝒢}          ▷ consider C'_𝒢 a candidate combination
```

The above approach to partitioning G^{tum} to identify candidate gene combinations is described formally in Procedure 1. Because the Leiden algorithm is randomized, every call of Procedure 1 may yield a different set of candidate gene combinations. To increase the likelihood of carcinogenic combinations to appear among the results, we perform multiple iterative partitioning passes via Procedure 1 and combine the obtained sets of candidate combinations $\mathscr{K}^{\bullet}$ into the joint candidate pool $\mathscr{K}$.

The iterative partitioning passes are independent of each other and can be performed in distributed fashion. We run them in parallel on the Summit supercomputer, relying on the implementation of Leiden algorithm from the Python library `leidenalg`.

2.3 Filtration of candidates

For a candidate combination $C_{\mathscr{G}} \in \mathscr{K}$, define its *tumor ratio* as the share of tumor samples among all those in which $C_{\mathscr{G}}$ is jointly mutated, $r(C_{\mathscr{G}}) := \frac{|M^{\text{tum}}(C_{\mathscr{G}})|}{|M(C_{\mathscr{G}})|} \in [0,1]$. We impose a threshold $\rho \in [0,1]$ on the tumor ratio of candidate combinations, so that only "sufficiently carcinogenic" candidates $\mathscr{K}_\rho := \{C_{\mathscr{G}} \in \mathscr{K} : r(C_{\mathscr{G}}) \geq \rho\}$ (notice that $\mathscr{K}_0 = \mathscr{K}$) are considered when selecting the final set of carcinogenic combinations from the pool. For example, the choice of $\rho = 1$ implies considering only candidate combinations with no joint mutations in normal samples. This particular choice however is likely to result in implausible carcinogenic combinations due to the possibility of non-mutagenic drivers behind carcinogenesis and potential inaccuracies of genomic data.

2.4 Minimum set cover

For a reasonable choice of ρ, the set of tumor samples explained by the "sufficiently carcinogenic" candidates, $\mathscr{S}_\rho^{\text{tum}} := \bigcup_{C_{\mathscr{G}} \in \mathscr{K}_\rho} M^{\text{tum}}(C_{\mathscr{G}})$, is expected to coincide with $\mathscr{S}^{\text{tum}}$ assuming that every tumor there is caused by a joint mutation of some gene combination from $\mathscr{G}$ whose size is in the hypothesized range.

Given the threshold ρ, we select carcinogenic combinations as a subset of $\mathscr{K}_\rho$ whose joint mutations explain $\mathscr{S}_\rho^{\text{tum}}$ most concisely, by constructing a *minimum set cover* of $\mathscr{S}_\rho^{\text{tum}}$ using sets of tumor samples from $\{M^{\text{tum}}(C_{\mathscr{G}}) : C_{\mathscr{G}} \in \mathscr{K}_\rho\}$. Specifically, we employ a greedy heuristic that iteratively chooses a candidate combination from $\mathscr{K}_\rho$ to cover the most of yet unexplained samples in $\mathscr{S}_\rho^{\text{tum}}$ on each step (see Procedure 2). The size of the resulting set of carcinogenic combinations $\mathscr{C}_\rho$ is guaranteed to approximate the size of the true solution within the factor of $\ln m$, where $m := \max\{|M^{\text{tum}}(C_{\mathscr{G}})| : C_{\mathscr{G}} \in \mathscr{K}_\rho\}$ is the biggest number of tumor samples explained by a "sufficiently carcinogenic" candidate[29].

Procedure 2 Greedy Minimum Cover

Input: samples $\mathscr{S}_\rho^{\text{tum}}$ to cover, "sufficiently carcinogenic" candidate combinations $\mathscr{K}_\rho$
Output: carcinogenic combinations $\mathscr{C}_\rho$

1: $\mathscr{C}_\rho \leftarrow \emptyset$
2: **while** $\mathscr{S}_\rho^{\text{tum}} \neq \emptyset$ **do**
3: $\quad C_{\mathscr{G}} \leftarrow \arg\max_{C'_{\mathscr{G}} \in \mathscr{K}_\rho} |M^{\text{tum}}(C'_{\mathscr{G}}) \cap \mathscr{S}_\rho^{\text{tum}}|$ ▷ *find $C_{\mathscr{G}}$ covering most unexplained samples*
4: $\quad \mathscr{C}_\rho \leftarrow \mathscr{C}_\rho \cup \{C_{\mathscr{G}}\}$
5: $\quad \mathscr{S}_\rho^{\text{tum}} \leftarrow \mathscr{S}_\rho^{\text{tum}} \setminus M^{\text{tum}}(C_{\mathscr{G}})$

2.5 Algorithmic complexity

The time complexity of Procedure 1 is defined by the number and complexity of the calls to the Leiden algorithm it makes. Let $n_0, n_1, \ldots$ denote the numbers of genes in the inputs to these calls in chronological order. In particular, $n_0 = |\mathscr{G}|$ as the first Leiden algorithm call takes the whole graph G^{tum} as the input. For any $k > 0$, the input genes of the k-th call are a strict subset of the input genes of some previous call and therefore $n_k \leq n_j - 1$ for some $j < k$. Also, if every $i < k$ satisfies $n_i \leq |\mathscr{G}| - i$ then $i = k$ also does, as $n_k \leq n_j - 1 \leq |\mathscr{G}| - j - 1 \leq |\mathscr{G}| - k$. Because the condition is satisfied for $k = 1$, we obtain by induction that $n_i \leq |\mathscr{G}| - i$ for every i. It follows that the number of Leiden algorithm calls made by Procedure 1 is $O(|\mathscr{G}|)$, and because the i-th call takes $O(|\mathscr{G}| - i)$ genes and $O(|\mathscr{S}|)$ samples, its time complexity is $\sum_i^{|\mathscr{G}|} T_{\text{Ldn}}(|\mathscr{G}| - i + |\mathscr{S}|)$, where $T_{\text{Ldn}}(n)$ is the runtime of the Leiden algorithm on a graph with n vertices.

Because partitioning of $\mathscr{G}$ into combinations of size $\geq l$ produces at most $\frac{|\mathscr{G}|}{l}$ candidate combinations, their total number after calling Procedure 1 p times is bounded by $|\mathscr{K}| = O(p|\mathscr{G}|l^{-1})$. Procedure 2 enumerates the remaining candidate combinations on each step, and terminates after at most $|\mathscr{S}^{\text{tum}}|$ steps. If a step covers only one new tumor sample, updating the count of yet unexplained samples covered by some candidate combination takes $O(1)$ time. Viewing each step covering several new tumor samples as multiple single-sample steps then bounds the time complexity of Procedure 2 by $O(|\mathscr{S}^{\text{tum}}||\mathscr{K}|) = O(p|\mathscr{S}^{\text{tum}}||\mathscr{G}|l^{-1})$.

The resulting time complexity of BiGPICC pipeline is $p\sum_{i=1}^{|\mathscr{G}|} T_{\text{Ldn}}(|\mathscr{G}| - i + |\mathscr{S}|) + O(p|\mathscr{S}^{\text{tum}}||\mathscr{G}|l^{-1})$, or $\sum_{i=1}^{|\mathscr{G}|} T_{\text{Ldn}}(|\mathscr{G}| - i + |\mathscr{S}|) + O(p|\mathscr{S}^{\text{tum}}||\mathscr{G}|l^{-1})$ if the p calls to Procedure 1 are done in parallel. While it was proved that a CPM-optimal partition of an arbitrary n-vertex graph is reachable in $O(n)$ steps of the Leiden algorithm, no upper bound on the number of steps, and therefore on $T_{\text{Ldn}}(n)$, is known. However, the algorithm was empirically shown to run in near-$O(n)$ time on a variety of real-world and generated graphs of up to 10^7 vertices[28]. If assuming linear runtime of the Leiden algorithm, the BiGPICC time complexity becomes $O(\frac{1}{2}|\mathscr{G}|^2 + |\mathscr{G}||\mathscr{S}| + p|\mathscr{S}^{\text{tum}}||\mathscr{G}|l^{-1})$.

Abbreviation	Cancer type	h, # hits	$\lvert\mathscr{G}\rvert$, # genes	$\lvert\mathscr{S}\rvert$, # samples	# mutations
BLCA	Bladder urothelial carcinoma	4, 6-7, 7	18,511	699	519,225
BRCA	Breast Invasive arcinoma	3, 2-3, 6-7	19,411	1,242	837,207
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	5, 6-7	18,813	605	459,351
COAD	Colon adenocarcinoma	3, 4, 5, 6	19,487	716	665,932
GBM	Glioblastoma multiforme	2, 3	18,920	662	852,723
HNSC	Head and neck squamous cell carcinoma	4, 5	18,707	801	635,410
KIRP	Kidney renal papillary cell carcinoma	8	18,117	559	429,005
LGG	Brain lower grade glioma	3, 6	17,770	810	495,887
LIHC	Liver hepatocellular carcinoma	8	17,592	643	438,879
LUAD	Lung adenocarcinoma	3, 6-7	18,783	740	629,351
LUSC	Lung squamous cell carcinoma	5	18,454	636	488,441
OV	Ovarian serous cystadenocarcinoma	2, 6-7	18,814	658	671,888
PRAD	Prostate adenocarcinoma	6-7, 7	17,274	752	476,115
SARC	Sarcoma	5, 8	17,687	550	379,824
THCA	Thyroid carcinoma	5, 6	18,264	752	524,992
UCEC	Uterine corpus endometrial carcinoma	2, 6-7	19,889	826	911,903

Table 1. Hit ranges required for carcinogenesis and data parameters for the 16 cancer types.

The mutation data is represented as a $|\mathscr{G}| \times |\mathscr{S}|$ adjacency matrix. Candidate combinations identified in the community detection step are a superset of all BiGPICC outputs and span the total of $O(p|\mathscr{G}|)$ non-unique vertices. Therefore, the method's memory complexity is $O(|\mathscr{G}||\mathscr{S}| + p|\mathscr{G}|)$.

2.6 Learning threshold for the tumor ratio

To choose the best value of ρ, we frame the BiGPICC pipeline as a classification problem by using the selected carcinogenic combinations $\mathscr{C}_\rho$ to predict whether previously unseen samples are normal or tumor. Namely, a sample is classified as tumor if it has a joint mutation of any $C_\mathscr{G} \in \mathscr{C}_\rho$, and as normal otherwise. Notice that an increase in the value of ρ improves the precision score of the classifier while also growing the size of $\mathscr{C}_\rho$ (as the choice of $C_\mathscr{G} \in \mathscr{K}_\rho$ at each step of Procedure 2 is narrowed). Because the number of gene combinations in $\mathscr{C}_\rho$ corresponds to classifier's complexity and therefore its propensity to overfit, $1-\rho$ can be viewed as the amount of regularization in the training, a trade-off between the "carcenogenicity" and the generalizability of learned combinations.

Before BiGPICC can access the data, we remove 25% of the samples to serve as a test dataset $\mathscr{S}_{\text{test}}$ for the final model. On the remaining 75% samples $\mathscr{S} \setminus \mathscr{S}_{\text{test}}$, we employ 4-fold cross-validation to find the optimal value of hyperparameter ρ. Specifically, $\mathscr{S} \setminus \mathscr{S}_{\text{test}}$ is partitioned into equal parts $\mathscr{S}^{(k)}_{\text{fold}}, k = 1,2,3,4$ and in 4 separate scenarios the pipeline is run using the samples $(\mathscr{S} \setminus \mathscr{S}_{\text{test}}) \setminus \mathscr{S}^{(k)}_{\text{fold}}$ to produce $\mathscr{C}_\rho$ for each value of ρ ranging from 0 to 1 with the increment of 0.01. The values of ρ are then assessed by averaging the performance of $\mathscr{C}_\rho$-based binary classifier on $\mathscr{S}^{(k)}_{\text{fold}}$ across the 4 scenarios. We use the *Matthews correlation coefficient* (MCC, also called the *phi coefficient*) — which takes values between -1 and 1 and is considered more informative than the commonly chosen F1 and accuracy scores[30] — as a single performance metric. Because constructing $\mathscr{C}_\rho$ and evaluating its performance for individual values of ρ can be done in parallel, the search for best ρ does not introduce a significant computational overhead. The value of ρ delivering highest mean MCC is then used to train the final classifier (by re-running BiGPICC pipeline) on the samples $\mathscr{S} \setminus \mathscr{S}_{\text{test}}$ and report its performance on $\mathscr{S}_{\text{test}}$.

3 Results

3.1 Experimental setup

Our somatic mutation data was collected in mutation annotation format (MAF) from The Cancer Genome Atlas (TCGA) for 16 cancer types using Mutect2 software. We identify a set of 331 matched blood-derived normal samples from all cancer types. We use the Variant Effect Predictor (VEP) to determine the location (intron, exon, UTR) and effect of these variants (synonymous, non-synonymous, missense, nonsense), and only consider protein-altering variants (non-synonymous, nonsense, and insertion/deletions in exons).

We apply BiGPICC to 16 datasets for different cancer types in 33 numerical experiments, each exploring a combination size h or its range for a cancer type according to existing literature. Each of the 16 datasets contains the same set of 331 normal

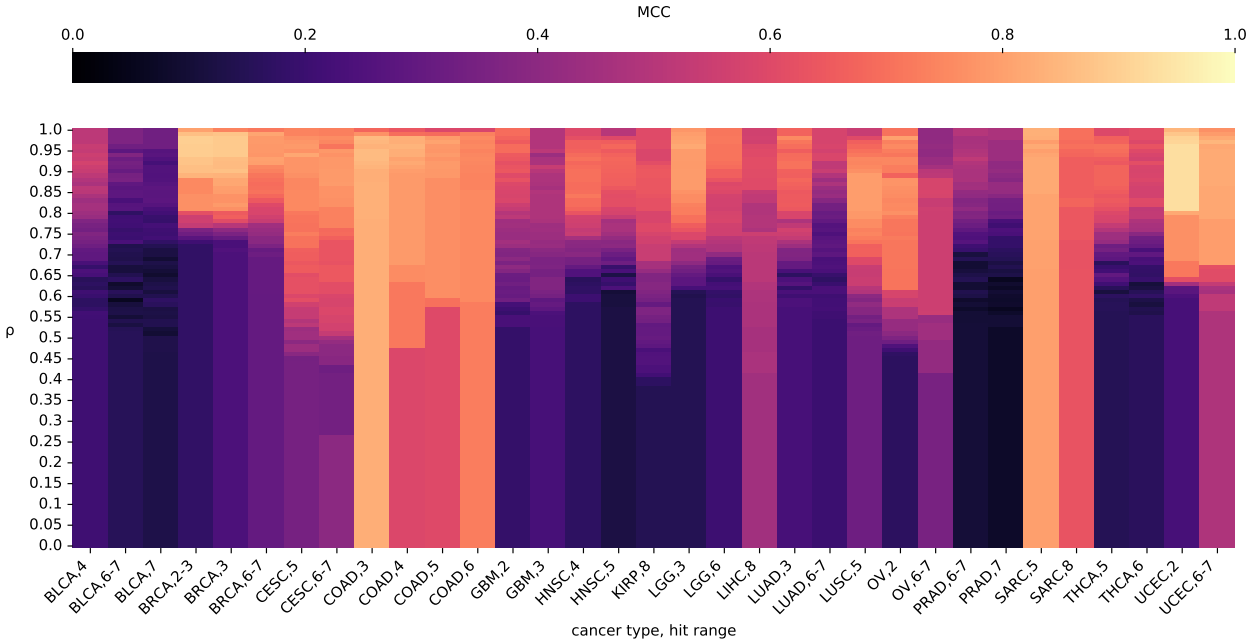

Figure 2. MCC fold-average of the MCC in cross-validation runs for the full range of ρ values. The range of MCC is truncated to $[0,1]$ due to the absence of negative values in our results.

tissue samples. Table 1 shows the hypothesized ranges of h and the dataset parameters for each cancer type. For each cancer type and multi-hit range, we run the community detection step (Procedure 1) $p = 10{,}000$ times to increase the probability of including all carcinogenic combinations to the candidate pool.

3.2 Classification performance

In each numerical experiment we build a classifier based on the identified gene combinations of desired size to differentiate between tumor and normal samples. We use the MCC, specificity, sensitivity, and F1 metrics to highlight various aspects of our classifiers. Each of the numerical experiments performs the 4-fold cross-validation runs and chooses the best value of ρ based on the highest mean MCC across the 4 folds. We demonstrate the effect of the value of ρ on the MCC in individual numerical experiments in Figure 2. After using the chosen value of ρ to train the final classifier on $\mathscr{G} \sqcup (\mathscr{S} \setminus \mathscr{S}_{\text{valid}})$, we report its combination count and performance metrics on the previously unseen test samples $\mathscr{S}_{\text{test}}$ in Table 2. The spread of classification performance within the same cancer type is likely attributed to the biological relevancy of the chosen hit range in individual experiments. The variation in performance across different cancer types can be explained by the data availability in each case and the representativity of available dataset for the high-dimensional search space. In particular, poor performance of PRAD and BLCA classifiers is consistent with the previous results based on the same data[16,19].

3.3 Comparison against exhaustive search results

The authors of Al Hajri et al.[17] identified carcinogenic combinations of length 2, 3, and 4 for various cancer types through an exhaustive search (ES) of all possible gene combinations of given length. Similarly to the minimum set cover step of BiGPICC, they iteratively add carcinogenic gene combinations until their resulting set explains all tumor samples in the data. Each step of ES approach selects the combination $C_{\mathscr{G}}$ maximizing $\alpha n_{\text{TP}} + n_{\text{TN}}$, where n_{TP} is the number of currently unexplained tumor samples in which $C_{\mathscr{G}}$ is jointly mutated and n_{TN} is the number of normal samples in which it is not. The parameter α represents the importance of correctly identifying normal samples relative to the analogous importance for tumor samples, and was fixed to $\alpha = 0.1$ in all ES experiments.

Figure 3 shows the comparison between the classification performance of carcinogenic combinations identified by BiGPICC for $h < 5$ and the metrics reported in Al Hajri et al. The two methods have comparable performance except for GBM cancer where BiGPICC performs significantly worse than ES. In seven out of ten cases, BiGPICC outperforms ES in either specificity or sensitivity, and in two cases — in both the metrics. A possible explanation for the heuristics-based BiGPICC outperforming an exhaustive search method may be a suboptimal choice of constant $\alpha = 0.1$ in the latter. In addition, the train-test splits used in our runs are not the same as in Al Hajri et al. If the amount of samples in the dataset is insufficient, a particular choice of the split may significantly impact the resulting performance, e.g. if the training samples $\mathscr{S} \setminus \mathscr{S}_{\text{test}}$ fail to capture all relevant

Cancer type	Hit range	$\lvert\mathscr{C}_\rho\rvert$, # combinations	MCC	Specificity	Sensitivity	F1
BLCA	4	24	**0.555**	0.855	0.696	0.762
	6-7	43	0.550	0.976	0.522	0.676
	7	71	0.336	0.916	0.370	0.511
BRCA	2-3	5	**0.900**	0.892	0.987	0.974
	3	9	**0.900**	0.892	0.987	0.974
	6-7	34	0.749	0.916	0.882	0.922
CESC	5	12	0.826	0.916	0.912	0.905
	6-7	13	**0.868**	0.976	0.882	0.923
COAD	3	14	**0.911**	0.976	0.938	0.957
	4	18	0.890	0.964	0.938	0.952
	5	24	0.827	0.964	0.865	0.912
	6	33	0.747	0.916	0.833	0.874
GBM	2	22	0.447	0.759	0.687	0.712
	3	26	**0.462**	0.795	0.663	0.710
HNSC	4	12	0.734	0.855	0.881	0.889
	5	18	**0.736**	0.867	0.873	0.888
KIRP	8	30	**0.663**	0.952	0.667	0.768
LGG	3	11	**0.847**	0.916	0.933	0.937
	6	27	0.687	0.916	0.783	0.851
LIHC	8	59	**0.542**	0.952	0.538	0.677
LUAD	3	18	**0.777**	0.928	0.853	0.892
	6-7	31	0.525	0.904	0.608	0.721
LUSC	5	10	**0.723**	0.867	0.855	0.855
OV	2	15	**0.790**	0.916	0.873	0.890
	6-7	47	0.404	0.928	0.418	0.559
PRAD	6-7	45	0.436	0.831	0.600	0.692
	7	62	**0.487**	0.928	0.533	0.671
SARC	5	10	**0.840**	1.000	0.800	0.889
	8	19	0.626	0.988	0.545	0.698
THCA	5	26	**0.671**	0.904	0.771	0.835
	6	39	0.499	0.904	0.581	0.701
UCEC	2	3	**0.889**	0.928	0.960	0.956
	6-7	15	0.818	0.880	0.935	0.928

Table 2. Classification performance of identified combinations on $\mathscr{S}_{\text{test}}$. Best MCC values per cancer type are in bold.

gene interactions. In particular, our second attempt of the random train-test split of GBM data resulted in significantly higher MCC scores on $\mathscr{S}_{\text{test}}$ — 0.791 for GBM, 2 and 0.714 for GBM, 3, with the new specificity for GBM, 2 outperforming its ES counterpart.

3.4 Runtime performance

The 4-fold cross-validation BiGPICC runs on 75% of all samples for tuning the parameter ρ were conducted using 280 Summit nodes each and took on average 26 minutes across all experiments, ranging from 11 minutes (SARC, 8) to 1 hour 29 minutes (UCEC, 2). Running the pipeline on the full dataset (without the validation samples) to measure the resulting classification performance was done using 70 nodes and took 23 minutes on average, ranging from 8 minutes (SARC, 8) to 1 hour 33 minutes (UCEC, 6-7). Adding the two mean runtimes gives that identifying carcinogenic combinations on the full dataset from scratch takes BiGPICC on average 146 node-hours on Summit.

For comparison, the ES method from Dash et al.[16] requires 124 Summit node-hours to identify 4-hit combinations for BRCA cancer type based on 75% of the available samples. Assuming ideal scaling of the method, the time required to identify 5-, 6-, and 7-hit combinations using all 4,600 nodes of Summit supercomputer would be 4.4 days, 39 years, and 107,000 years, respectively. In contrast, the BiGPICC runtimes are not noticeably affected by the combination size h and instead depend on the size and topology of the input graph determined by the cancer type. It follows that BiGPICC is orders of magnitude faster than ES for any $h > 4$.

While ES relies on GPUs, BiGPICC only uses CPU cores. Per Summit node, BiGPICC uses 42 CPU cores, while ES uses $6 \times 5{,}120 = 30{,}720$ GPU cores. Though CPU and GPU cores are not directly comparable, BiGPICC is using roughly 2% of

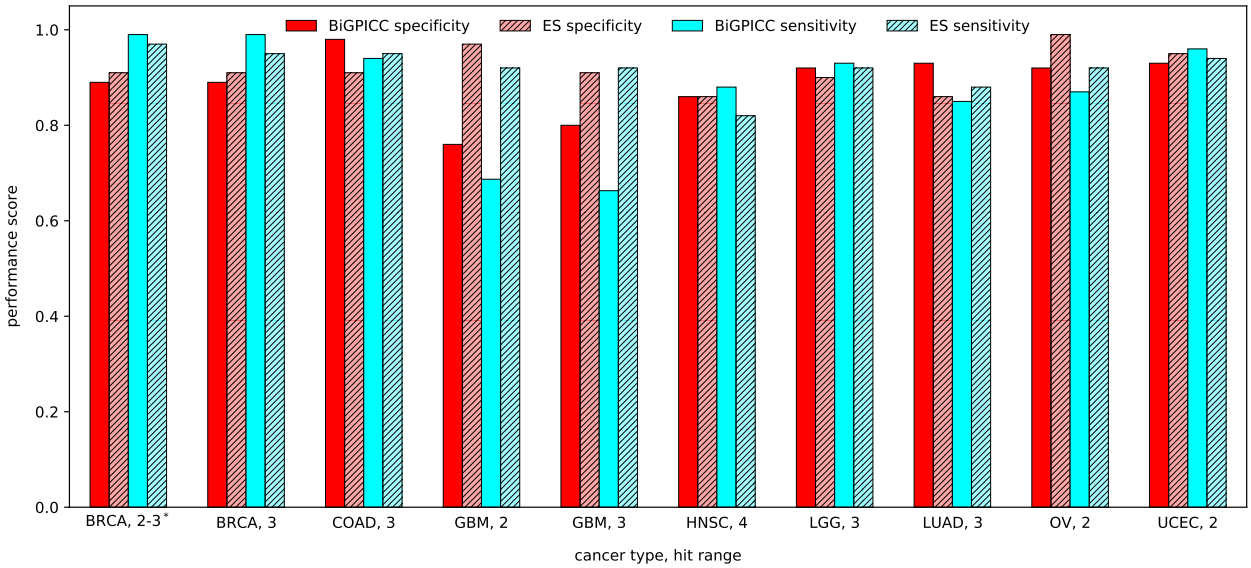

Figure 3. Comparison between classification performances of BiGPICC and Exhaustive Search (ES), rounded to 2 decimal places. *BiGPICC runs for BRCA, 2-3 result in at most one 3-hit combination per fold and are compared against the ES results for BRCA, 2 as the best matching option.

each compute node and thus can be moved to a less expensive CPU-only cluster to achieve the same runtime.

4 Conclusions and discussion

We proposed a community detection-based approach for identifying carcinogenic combinations and showed that the biological relevance of its results is comparable to that for state-of-the-art. At the same time, BiGPICC enables discovery of (5+)-hit combinations intractable for exhaustive search methods even on most modern supercomputers.

In all cancer types considered, biological relevance of the identified combinations, measured as classification performance, tends to drop as the number of hits increases. An identical trend in the exhaustive search results from Al Hajri et al.[17] suggests that, aside from biological reasons, the issue may lie with the so-called curse of dimensionality in both the machine learning and combinatorial contexts. An increase in the number of hits means fewer samples with joint mutation of a gene combination exist in the data while the search space of possible combinations grows exponentially. Therefore, the number of samples required for finding carcinogenic combinations grows with the number of hits, while our runs use the same dataset for every multi-hit range.

The classification performance of BiGPICC exhibits a significantly higher variability in sensitivity, the percentage of correctly classified tumor samples, than in specificity, the analogous percentage for normal samples. The sensitivity and specificity scores across the experiments are distributed as 0.763 ± 0.175 (mean $\pm$ SD) and 0.911 ± 0.054, respectively. We attribute this trend to the implicit control exhibited over the specificity score by the parameter ρ, whose learned value is consistently high in the experiments (0.938 ± 0.051). Let n_{TP}, n_{FN}, n_{TN}, and n_{FP} denote respectively the number of correctly predicted tumor samples, incorrectly predicted tumor samples, correctly predicted normal samples, and incorrectly predicted normal samples in the training dataset with the total number of samples $n = n_{\mathrm{TP}} + n_{\mathrm{FN}} + n_{\mathrm{TN}} + n_{\mathrm{FP}}$. Parameter ρ ensures that the combinations selected for the classifier have sufficient tumor ratio, which translates into controlling its precision $\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FP}}}$. Assuming for simplicity no overlap between the samples in which the selected combinations are jointly mutated, the relationship is given by $\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FP}}} \geq \rho$ (in general, the right hand side can be both larger or smaller depending on the tumor ratio in the overlapping samples). It follows that the number of false positives is limited by the number of tumor samples in the dataset, $n_{\mathrm{FP}} \leq \frac{1-\rho}{\rho} n_{\mathrm{TP}} \leq \frac{1-\rho}{\rho}(n_{\mathrm{TP}} + n_{\mathrm{FN}}) = \frac{1-\rho}{\rho} rn$, where $r = \frac{n_{\mathrm{TP}}+n_{\mathrm{FN}}}{n}$ is the ratio of tumor samples in the data. Because n_{FP} is the only factor negatively affecting the specificity $\frac{n_{\mathrm{TN}}}{n_{\mathrm{TN}}+n_{\mathrm{FP}}}$, the latter is bounded from below by $\frac{n_{\mathrm{TN}}}{n_{\mathrm{TN}}+n_{\mathrm{FP}}} = 1 - \frac{n_{\mathrm{FP}}}{(1-r)n} \geq 1 - \frac{(1-\rho)r}{(1-r)\rho} \xrightarrow[\rho \to 1]{} 1$. In particular, the average value of $1 - \frac{(1-\rho)r}{(1-r)\rho}$ across the experiments is 0.865 (if r is calculated using the full datasets; it is expected to be the same along the training-validation split). At the same time, the pipeline does not exhibit control over n_{FN}, the adverse factor for the sensitivity $\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FN}}}$, which is comprised by the tumor samples left unexplained by "sufficiently carcingonenic" combinations.

Unlike most other approaches, BiGPICC does not require manually balancing the importance of tumor and normal samples

in the data. Instead, its hyperparameter ρ which implicitly takes on this role is learned from the data using cross-validation. This renders our approach parameter-free, alleviating the need to manually tune it on a case-by-case basis.

5 Limitations and future work

BiGPICC does not offer a dedicated mechanism to differentiate between the driver mutations causing carcinogenesis and the passenger mutations that do not contribute to cancer formation. It means that some of the highly mutable genes among the identified ones can be mutated in tumor samples by chance and therefore lack biological relevance. In particular, each of our datasets contains between 41 and 67 genes with mutations in over 50% of the samples. This can lead to many candidate combinations with near-identical explanatory power, which can be behind the observed phenomenon of the similar classification performance exhibited by significantly different sets of carcinogenic combinations. Increasing the number of samples in the data would mitigate the issue of driver-passenger distinction by amplifying the signal of carcinogenic pathways against the backdrop of noise from chance mutations. Under the assumption that highly mutable genes are unlikely to drive carcinogenesis, another approach would be to remove genes with mutation frequency above some domain-informed threshold from the analysis.

Another limitation of the method is that it is run on the gene-patient data encoding all possible mutations of a gene as a binary variable, thus ignoring the variability in mutations of individual genes. However, the BiGPICC pipeline could similarly be applied to mutation-level data for the identified carcinogenic gene combinations. Given that the identified genes typically constitute less than 1% of the original gene pool, we expect the mutation-level data to be of similar size and tractability to the datasets used in this study.

Assuming that the sufficient amount of data is available, an additional use of our pipeline is testing out the competing ranges of multi-hit theories by comparing the performance of learned carcinogenic combinations of the corresponding size. The ICGC database[31] contains roughly twice as many samples as our datasets for individual cancer types, and may be a prospective dataset to try this approach on.

References

1. Sutherl, J. V. & Bailar III, J. C. The multihit model of carcinogenesis: etiologic implications for colon cancer. *J. Chronic Dis.* **37**, 465–480 (1984).
2. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci.* **112**, 118–123 (2015).
3. Zhang, X. & Simon, R. Estimating the number of rate limiting genomic changes for human breast cancer. *Breast cancer research treatment* **91**, 121–124 (2005).
4. Little, M. & Wright, E. A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Math. biosciences* **183**, 111–134 (2003).
5. Ashley, D. The two" hit" and multiple" hit" theories of carcinogenesis. *Br. journal cancer* **23**, 313 (1969).
6. Anandakrishnan, R., Varghese, R. T., Kinney, N. A. & Garner, H. R. Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLoS computational biology* **15**, e1006881 (2019).
7. Tian, R., Basu, M. & Capriotti, E. Contrastrank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics* **30**, 572–578 (2014).
8. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2242 (2013).
9. Dees, N. D. *et al.* Music: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589–1598 (2012).
10. Kumar, R. D., Swamidass, S. J. & Bose, R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nat Genet.* **48**, 1288–1294 (2016).
11. Kuchenbaecker, K. B. *et al.* Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* **317**, 2402–2416 (2017).
12. Mai, P. *et al.* Li-Fraumeni syndrome: report of a clinical research workshop and creation of a research consortium. *Cancer Genet.* **205**, 479–87 (2012).
13. Pantziarka, P. Primed for cancer: Li Fraumeni Syndrome and the pre-cancerous niche. *Ecancermedicalscience* **9**, 541 (2015).

14. Guha, T. & Malkin, D. Inherited TP53 mutations and the Li-Fraumeni syndrome. *Cold Spring Harb Perspect Med* **7**, a026187 (2017).

15. Amadou, A., Waddington Achatz, M. & Hainaut, P. Revisiting tumor patterns and penetrance in germline TP53 mutation carriers: temporal phases of Li-Fraumeni syndrome. *Curr Opin Oncol* **30**, 23–29 (2018).

16. Dash, S. *et al.* Differentiating between cancer and normal tissue samples using multi-hit combinations of genetic mutations. *Sci. reports* **9**, 1–13 (2019).

17. Al Hajri, Q., Dash, S., Feng, W.-c., Garner, H. R. & Anandakrishnan, R. Identifying multi-hit carcinogenic gene combinations: Scaling up a weighted set cover algorithm using compressed binary matrix representation on a gpu. *Sci. reports* **10**, 1–18 (2020).

18. Dash, S., Al-Hajri, Q., Feng, W.-c., Garner, H. R. & Anandakrishnan, R. Scaling out a combinatorial algorithm for discovering carcinogenic gene combinations to thousands of gpus. In *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 837–846 (IEEE, 2021).

19. Klein, M. I. *et al.* Identifying modules of cooperating cancer drivers. *Mol. systems biology* **17**, e9810 (2021).

20. Cho, A. *et al.* Muffinn: cancer gene discovery via network analysis of somatic mutation data. *Genome biology* **17**, 1–16 (2016).

21. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. genetics* **47**, 106–114 (2015).

22. Leung, A., Bader, G. D. & Reimand, J. Hypermodules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics* **30**, 2230–2232 (2014).

23. Ruffalo, M., Koyutürk, M. & Sharan, R. Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS computational biology* **11**, e1004595 (2015).

24. Drake, J. M. *et al.* Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell* **166**, 1041–1054 (2016).

25. Dimitrakopoulos, C. *et al.* Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).

26. Iranzo, J., Martincorena, I. & Koonin, E. V. Cancer-mutation network and the number and specificity of driver mutations. *Proc. Natl. Acad. Sci.* **115**, E6010–E6019 (2018).

27. Traag, V. A., Van Dooren, P. & Nesterov, Y. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E* **84**, 016114 (2011).

28. Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Sci. reports* **9**, 1–12 (2019).

29. Slavík, P. A tight analysis of the greedy algorithm for set cover. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 435–441 (ACM, 1996).

30. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1–13 (2020).

31. Zhang, J. *et al.* The international cancer genome consortium data portal. *Nat. biotechnology* **37**, 367–369 (2019).

Acknowledgements

This work was supported by the resources of the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at ORNL, which is managed by UT Battelle, LLC for the U.S. DOE (under the contract No. DE-AC05-00OR22725).

Author contributions statement

R.A. preprocessed the data, S.D. conceived the idea, V.O. designed and implemented the algorithm, V.O. and S.D. ran the experiments and analyzed the results. All authors reviewed the manuscript.

Additional information

Our code and the datasets for two cancer types are available at https://code.ornl.gov/vo0/bigpicc. The datasets for other cancer types are available upon request.

The authors declare no competing interests.